Viral Quasispecies Reconstruction Using Next Generation Sequencing Reads

Bassam A. Tork
Georgia State University

Follow this and additional works at: https://scholarworks.gsu.edu/cs_diss

Recommended Citation
Tork, Bassam A., "Viral Quasispecies Reconstruction Using Next Generation Sequencing Reads."
Dissertation, Georgia State University, 2013.
doi: https://doi.org/10.57709/4313753

This Dissertation is brought to you for free and open access by the Department of Computer Science at ScholarWorks @ Georgia State University. It has been accepted for inclusion in Computer Science Dissertations by an authorized administrator of ScholarWorks @ Georgia State University. For more information, please contact scholarworks@gsu.edu.
The genomic diversity of viral quasispecies is a subject of great interest, especially for chronic infections. Characterization of viral diversity can be addressed by high-throughput sequencing technology (454 Life Sciences, Illumina, SOLiD, Ion Torrent, etc.). Standard assembly software was originally designed for single genome assembly and cannot be used to assemble and estimate the frequency of closely related quasispecies sequences.

This work focuses on parsimonious and maximum likelihood models for assembling viral
quasispecies and estimating their frequencies from 454 sequencing data. Our methods have been applied to several RNA viruses (HCV, IBV) as well as DNA viruses (HBV), genotyped using 454 Life Sciences amplicon and shotgun methods.

INDEX WORDS: RNA viruses, Viral quasispecies, Genome assembly
VIRAL QUASISPECIES RECONSTRUCTION USING NEXT GENERATION
SEQUENCING READS

by

BASSAM TORK

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2013
VIRAL QUASISPECIES RECONSTRUCTION USING NEXT GENERATION SEQUENCING READS

by

BASSAM TORK

Committee Chair: Dr. Alexander Zelikovsky

Committee: Dr. Yury Khudyakov
Dr. Robert Harrison
Dr. Rajshekhar Sunderraman

Electronic Version Approved:

Office of Graduate Studies
College of Arts and Sciences
Georgia State University
August 2013
DEDICATION

To my parents Abdel Rahman Tork and Mariam Tork.
ACKNOWLEDGEMENTS

I would especially like to thank my advisor Dr. Alexander Zelikovsky for his encouragement, advisement and constant support over the years under his supervision at Georgia State University. This dissertation could not have been done without his guidance, patience and motivation. I have learned more about computer science and conducting research from him than virtually everyone else. I would especially like to thank my committee members: Dr. Yury Khudyakov, Dr. Robert Harrison, and Dr. Rajshekhar Sunderraman. Their critique of my work with no doubt strengthened and improved it. Special thanks to Professor Ion Mandoiu and Professors I.Khan and O’Neill of University of Connecticut for their valuable advisement.

Special thanks to Dr. Yi Pan Chairman of the Computer Science Department for his encouragement, advisement and support over the years of my PhD study.

Special thanks to Dr. Rajshekhar Sunderraman for his encouragement, advisement and support over the years of my PhD study.

Special thanks to Computer Science Department of Georgia State University faculty, staff and students.

Special thanks to the GSU Molecular Basis of Disease Initiative (MBD).

Special thanks go out to my dear friends Nick, Hamed, Serghei, Adrian, Sasha, Blanche, Olga, and Ekaterina.

Finally I am grateful to my parents, my brothers, my sisters and all my friends for their support and encouragement. Special thanks go from my heart to my nice family and daughters whom I miss a lot.
# TABLE OF CONTENTS

**ACKNOWLEDGEMENTS** .............................................. v

**LIST OF TABLES** ..................................................... ix

**LIST OF FIGURES** ..................................................... xi

**LIST OF ABBREVIATIONS** ........................................ xiv

**PART 1  INTRODUCTION** ........................................... 1

1.1 Viral Quasispecies .................................................. 1
1.2 Challenges of Quasispecies Reconstruction From NGS Reads ... 2
1.3 Viral Quasispecies Spectrum Reconstruction Problem .......... 2
1.4 Existing Approaches For Sequencing Viral Quasispecies ....... 3
1.5 Main Contributions ................................................... 4
1.6 Future Work ............................................................ 5
1.7 Roadmap of the Rest of the Dissertation ......................... 5
1.8 Publications ........................................................... 6
1.9 Oral Presentations ..................................................... 7
1.10 Book Chapters .......................................................... 7
1.11 Posters ................................................................. 7

**PART 2  STATE-OF-THE-ART IN VIRAL QUASISPECIES RECONSTRUCTION** . 9

2.1 Introduction ......................................................... 9
2.2 NGS Technologies ...................................................... 10
2.3 Existing Tools ........................................................ 17
    2.3.1 ShoRAH ........................................................... 18
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.3.2</td>
<td>QuRe</td>
</tr>
<tr>
<td>2.3.3</td>
<td>Error Correction</td>
</tr>
</tbody>
</table>

**PART 3** QUASISPECIES SPECTRUM RECONSTRUCTION FROM SHOTGUN READS 21

<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>ViSpA</td>
</tr>
<tr>
<td>3.1.1</td>
<td>Read Error Correction</td>
</tr>
<tr>
<td>3.1.2</td>
<td>Read Alignment</td>
</tr>
<tr>
<td>3.1.3</td>
<td>Processing of 454 Reads</td>
</tr>
<tr>
<td>3.1.4</td>
<td>Read Graph</td>
</tr>
<tr>
<td>3.1.5</td>
<td>Candidate path selection.</td>
</tr>
<tr>
<td>3.1.6</td>
<td>Estimation of candidate quasispecies sequence frequencies.</td>
</tr>
<tr>
<td>3.2</td>
<td>Fixing Erroneous Homopolymer Indels via Protein Alignment</td>
</tr>
</tbody>
</table>

**PART 4** QUASISPECIES SPECTRUM RECONSTRUCTION FROM AMPLICON READS 27

<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Theoretical Analysis of Quasispecies Assembly Problem for Amplicon Reads</td>
</tr>
<tr>
<td>4.1.1</td>
<td>Quasispecies Assembly in the Error-Free, Ideal-Frequency Model Problem</td>
</tr>
<tr>
<td>4.1.2</td>
<td>Reduction of Skewed-Frequency Model to Ideal-Frequency Model</td>
</tr>
<tr>
<td>4.2</td>
<td>Viral Quasispecies Reconstruction Algorithms From Amplicon Reads</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Read Graph</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Maximum Bandwidth Path</td>
</tr>
<tr>
<td>4.2.3</td>
<td>Maximum Frequency Path.</td>
</tr>
<tr>
<td>4.2.4</td>
<td>Greedy Algorithm for Resolving Forks.</td>
</tr>
<tr>
<td>4.3</td>
<td>Data Sets</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Simulated Reads</td>
</tr>
<tr>
<td>4.3.2</td>
<td>HBV Data</td>
</tr>
</tbody>
</table>
4.4 Validation Methods .............................................. 34
  4.4.1 Quasispecies Assembling Validation ...................... 35
  4.4.2 Validation of Frequency Distribution .................... 35
4.5 Experimental Results .......................................... 35
  4.5.1 Simulated Reads ........................................... 35
  4.5.2 Real Reads .................................................. 42

PART 5  RECONSTRUCTION OF IBV QUASISPECIES ................. 44
  5.1 Introduction .................................................... 44
  5.2 Methods ........................................................ 45
    5.2.1 Compared Methods ........................................ 46
  5.3 Data Sets and Golden Standards ............................. 47
  5.4 Validation ...................................................... 48
    5.4.1 Validation of 454 Reads .................................. 48
    5.4.2 Validation of Correction Methods ....................... 48
    5.4.3 Comparison Measures ..................................... 49
  5.5 Experimental Results ......................................... 50
    5.5.1 Results of 454 Reads Validation .......................... 50
    5.5.2 Results of Correction Methods Validation ............... 53
    5.5.3 Results of Reconstructed Quasispecies .................. 53

PART 6  CONCLUSIONS AND FUTURE WORK .......................... 63

REFERENCES ......................................................... 65
**LIST OF TABLES**

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Average sensitivity/PPV of ViSpA and Maximum Bandwidth algorithms for shotgun and amplicon error-free read data sets simulated from ten and twenty quasispecies [1, 2].</td>
<td>42</td>
</tr>
<tr>
<td>4.2</td>
<td>Total and distinct reads of amplicons per patient [1, 2].</td>
<td>43</td>
</tr>
<tr>
<td>5.1</td>
<td>Pairwise edit distance between the 10 Sanger clones [3].</td>
<td>54</td>
</tr>
<tr>
<td>5.2</td>
<td>Edit distance between collapsed Sanger clones and reconstructed quasispecies using parameters 1, 2, 5 (the number of mismatches between sub-reads and super-reads, the number of mismatches between two overlapped reads, and mutation rate respectively), threshold 0.005 on KEC corrected reads using ViSpA [3].</td>
<td>55</td>
</tr>
<tr>
<td>5.3</td>
<td>Edit distance between collapsed Sanger clones and reconstructed quasispecies using parameters 1, 2, 5 (the number of mismatches between sub-reads and super-reads, the number of mismatches between two overlapped reads, and mutation rate respectively), threshold 0.005 on uncorrected reads using ViSpA [3].</td>
<td>56</td>
</tr>
<tr>
<td>5.4</td>
<td>Edit distance between collapsed Sanger clones and reconstructed quasispecies using ShoRAH assembler with default parameters [3].</td>
<td>58</td>
</tr>
<tr>
<td>5.5</td>
<td>Average distance to clones (ADC) for the reconstructed quasispecies using different methods [3].</td>
<td>60</td>
</tr>
<tr>
<td>5.6</td>
<td>Average prediction error (APE) for the reconstructed quasispecies using different methods [3].</td>
<td>60</td>
</tr>
</tbody>
</table>
Table 5.7  Average prediction error (APE) and Average distance to clones (ADC) of reconstructed quasispecies for different methods [3].
LIST OF FIGURES

Figure 2.1 454 Sequencing Method [4]. ........................................... 12
Figure 2.2 Illumina Sequencing Method [5]. ................................. 14
Figure 2.3 SOLiD Sequencing Method [6]. .................................. 16
Figure 2.4 Ion Torrent Sequencing Method [7]. ............................ 17
Figure 3.1 Algorithmic approach to solve QSA problem. ............... 22
Figure 3.2 protein sequence alignment (CLUSTAL W) of the quasispecies sequence (0.0498579823867_ID20_*-1) with the reference (contig20_*-3) before systematic error fixing. ................................. 25
Figure 3.3 protein sequence alignment (CLUSTAL W) of the quasispecies sequence (0.0498579823867_ID20_*-1) with the reference (contig20_*-3) after systematic error fixing. ................................. 26
Figure 4.1 The case of two distinct reads for both amplicons [1,2]. .... 30
Figure 4.2 Sensitivity for different amplicon window shifts from uniform distribution [1,2] ................................................................. 37
Figure 4.3 Sensitivity for different amplicon window shifts from skewed distribution [1,2] ................................................................. 37
Figure 4.4 Sensitivity for different amplicon window shifts from geometric distribution [1,2] ................................................................. 38
Figure 4.5 PPV for different amplicon window shifts from uniform distribution [1,2] ................................................................. 38
Figure 4.6 PPV for different amplicon window shifts from skewed distribution \([1, 2]\) 

Figure 4.7 PPV for different amplicon window shifts from geometric distribution \([1, 2]\) 

Figure 4.8 Jensen-Shannon Divergence for different amplicon window shifts from uniform distribution \([1, 2]\) 

Figure 4.9 Jensen-Shannon Divergence for different amplicon window shifts from skewed distribution \([1, 2]\) 

Figure 4.10 Jensen-Shannon Divergence for different amplicon window shifts from geometric distribution \([1, 2]\) 

Figure 5.1 Viral quasispecies reconstruction pipeline \([3]\) 

Figure 5.2 Schematic representation of calibration, validation experiments based on cloned quasispecies \([3]\). 

Figure 5.3 Uncorrected reads validation. 

Figure 5.4 SAET corrected reads validation. 

Figure 5.5 KEC corrected reads validation. 

Figure 5.6 ShoRAH corrected reads validation. 

Figure 5.7 The distribution of 454 IBV reads categories (edit distance to the closest Sanger clone) for different correction methods. 

Figure 5.8 Phylogenetic tree over collapsed Sanger clones and collapsed reconstructed quasispecies inferred from the first method with parameters 1_2_5 on KEC corrected reads using ViSpA \([3]\).
Figure 5.9  Phylogenetic tree over collapsed Sanger clones and collapsed reconstructed quasispecies inferred from the second method with parameters $1.25$ on uncorrected reads using ViSpA [3].

Figure 5.10  Evaluation Diagram for Average Prediction Error(APE) and Average Distance to Clones(ADC) values for different methods [3].

59

62
LIST OF ABBREVIATIONS

• NGS - Next Generation Sequencing

• HTS - High Throughput Sequencing
1.1 Viral Quasispecies

Many viruses (e.g., hepatitis C virus (HCV), influenza) encode their genome in RNA. RNA viruses tend to have much higher mutation rates during replication compared to DNA-encoded viruses [8] due to the more error-prone RNA polymerase. The mutation rate of RNA viruses can be as high as 1 mutation per 1,000-100,000 bases per replication cycle [9]. These mutations tend to be well tolerated and passed down to viral descendants, producing a family of closely related variants called quasispecies [10–14], Quasispecies is a family of co-existing related variants of the original viral genome.

Quasispecies mutations may make some sequence more competitive for infecting cells than other quasispecies sequences, which would result in that sequence becoming more frequent in the organism. It may, also, occur that the frequency of all quasispecies sequences in the organism shifts dramatically.

Researchers have observed that the quasispecies sequences in two organisms infected simultaneously might drift so rapidly apart that within a few years it would be impossible to infer that they were infected from the same source [10, 12].

Even though DNA viruses, such as HBV, have much lower mutation rates, they also exist in the form of quasispecies [15]. Great diversity of viral sequences in an infected individual can increase viral resistance to existing drug therapies, thus impeding the utility of vaccines [16].

Reconstructing genomic diversity of viral quasispecies is of significant importance for this reason and can additionally be applied in designing vaccines [17, 18] and drugs [19, 20] curing patients from particular viral species in vivo. By knowing particular quasispecies sequences Skums et al. [21] could estimate outcome of interferon treatment for HCV infected
host, therefore improving cost effectiveness and reduce patient hardship.

1.2 Challenges of Quasispecies Reconstruction From NGS Reads

Below we describe two main challenges facing quasispecies reconstruction from next generation sequencing (NGS). The first challenge is related to NGS technology limitations, causing genotyping errors. There are two types of genotyping errors: mismatches and indels, i.e., insertions into reads and deletions from reads with respect to reference sequence (or consensus). For example 454 Life Sciences erroneously sequences on average 1 in 1000bp [4]. A large amount of these mistakes are indels [22, 23]. The reason is incorrect resolution of homopolymers, i.e., substrings repeating a single base (for example, AAAAAA).

The second challenge in assembling quasispecies is caused by conserved regions longer than NGS reads. As a result it is difficult to match high variable regions flanking long lower variable region [24–28].

1.3 Viral Quasispecies Spectrum Reconstruction Problem

Quasispecies assembly addresses the problem of reconstructing the sampled quasispecies’ genome mainly with the consultation of previously resolved sequence. Since multiple quasispecies have indistinguishable common segments that may be longer than a read, one cannot guarantee to find even the exact number of quasispecies sequences (although it is possible to find a lower bound for the number of sequences). Although cross-validation of proposed techniques can tell if their quality are of practical interest, it is important to formally formulate the problem of interest.

**Quasispecies Spectrum Reconstruction (QSR) Problem**: given a set of 454 Life Sciences reads taken from a quasispecies population of unknown size and unknown distribution, reconstruct the quasispecies spectrum (i.e., the set of sequences and the relative frequency of each sequence) [1, 2, 25–28].

The QSR problem is related to several problems, like haplotype assembly ([29–35]), *de novo* assembly ([36–45]), and metagenomics. QSR problem can be reduced to a DNA
fragment assembly if all quasispecies sequences are placed one-by-one in a long sequence and common segments are treated as repeats.

QSR problem can be viewed also as an analysis of genomes in a mixed environment that searches for all different species in a given DNA sample, known as metagenomics, where the difference between QSR problem and the analysis of metagenomic data is in the diversity, where quasispecies are much closer to each other at the nucleotide level than genomes of different species. The main factors that determine complexity of the QSR problem, are the diversity of quasispecies (this results in easier partitioning of reads among individual quasispecies) and the mean read length (longer reads results in more assembling accuracy).

1.4 Existing Approaches For Sequencing Viral Quasispecies

Several assembly softwares have been produced mainly for de novo genome assembly, in which unique genome reconstruction is done without the consultation to a previously resolved genome sequence. In Quasispecies Spectrum Reconstruction(QSR), mapping based approaches are preferred to de novo assembly, since viral genomes do not contain repeats and reference genomes are easy to obtain for viruses of interest. Such approaches were adopted in the pioneering works on the QSR problem [24, 27, 46]. Eriksson et al. [24] introduced a multi-step approach consisting of sequencing error correction via multiple alignment and clustering using a statistical testing procedure, haplotype reconstruction via chain decomposition, and haplotype frequency estimation via expectation-maximization, with validation on HIV data. Westbrooks et al. [27], introduced haplotype reconstruction via transitive reduction taking into consideration overlap probability estimation, in addition to network flows application, with validation on simulated error-free HCV data. Astrovskaya et al. [25] introduced quasispecies reconstruction via error correction, overlap probability estimation, transitive reduction of the read graph, and shortest paths, with application to both simulated and real HIV and HCV data. The QSR software tool ShoRAH was developed [46] and applied to HIV data [47]. Shorah has a significant advantage over other tools on shotgun reads with genotyping errors due to its advanced error correction algorithm. It uses a
generative probabilistic model where it assigns observed reads to unobserved haplotypes in
the presence of sequencing errors, and solves this clustering via a Bayesian fashion using the
Dirichlet process mixture to define a prior distribution on the unknown number of haplotypes
in the mixture [46].

Another QSR software tool, Qure, was developed to reconstruct quasispecies spectrum
from amplicon reads rather than shotgun reads [48, 49]. Prosperi et al.[49] proposed a com-
binatorial method for QSR using NGS amplicon reads. Their method was developed and
applied to HIV and HBV data, using an algorithm based on multinomial distributions. The
algorithm reconstructs quasispecies by coupling reads that have consistent overlaps with
similar frequencies across the various amplicons.

1.5 Main Contributions

The main contributions are as follows:

• An improved method for reconstruction of quasispecies from shotgun reads including:
  
  – developing of the novel method for correction of erroneous homopolymer indels
    via protein alignment.
  
  – combining various error correction and alignment algorithms with ViSpA [25, 26]
    software.
  

• Novel method for reconstruction of quasispecies from amplicon reads including:  
  
  – novel theoretical analysis of QSR problem for amplicon reads including the fol-
    lowing models:
    
    * maximum parsimony
    
    * maximum likelihood, and

1The methods for reconstruction of viral quasispecies from amplicon reads have been developed in col-
  laboration with PhD. student Nicholas Mancuso.
* minimum entropy
  – several novel methods for assembling quasispecies from amplicon reads including:
    * maximum frequency
    * maximum bandwidth
    * greedy fork resolution
  – validation of the proposed methods on simulated data without sequencing errors
  – experimental validation of the reconstructed quasispecies sequences

• Reconstruction of IBV Quasispecies including:
  – different methods with different parameter settings
  – validation of 454 IBV reads
  – validation of correction methods
  – validation of reconstructed quasispecies

1.6 Future Work

Future work includes possible combining of shotgun and amplicon data, as well as incorporation of paired reads.

1.7 Roadmap of the Rest of the Dissertation

The remainder of the dissertation is devoted to the viral quasispecies assembly problem. Chapter 2 introduces the problem and 454 pyrosequencing, the technology used to collect the viral quasispecies sequence data. Chapter 3 introduces novel method for correction of erroneous homopolymer indels via protein alignment of reconstructed quasispecies from existing tools (e.g. ViSpA [25, 26]). Chapter 4 describes theoretical and algorithmic techniques using read graph model for quasispecies reconstruction from amplicon reads. Chapter 5 describes how to improve the predictive power of quasispecies inference problem by using
different methods with different parameter settings and parameter calibration. Chapter 6 describes our conclusions and future work.

1.8 Publications

(the major contributors are in bold font)


∗: contributed equally

1.9 **Oral Presentations**

*(the presentors are in bold font)*


1.10 **Book Chapters**

*(the major contributors are in bold font)*


1.11 **Posters**

*(the major contributors are in bold font)*


PART 2

STATE-OF-THE-ART IN VIRAL QUASISPECIES RECONSTRUCTION

2.1 Introduction

Quasispecies present big challenges towards drug and immune system resistance, which leads to the continuous evolution of viruses in humans and animals. Because of the lack of proofreading and repair mechanisms, quasispecies genomes mutate at average of $10^{-3}$ substitutions per nucleotide per replication cycle (s/n/c) [3, 9, 25–28, 49, 51–53]. Since mutations are tolerated, many RNA viruses exist as a set of closely related sequences called quasispecies rather than populations of similar clones. Genomic heterogeneity of viruses has many implications, contributing to virulence, drug and vaccince resistant variants [3, 16, 18–20, 25–28, 53]. Biological implications of quasispecies diversity remain poorly understood, because of the difficulty of identifying quasispecies sequences and their frequencies. One useful way is identifying the consensus sequence of the quasispecies, which can be done by Sanger sequencing of RT-PCR amplicons. Unfortunately, the frequencies of distinct variants can not be inferred easily [3, 53, 54]. When original sequence is not known, cloning followed by Sanger sequencing of individual variants is considered the gold standard in quasispecies analysis. Unfortunately, this process takes long time to be done, costs a lot, and the number of sequenced clones is limited, resulting in the recovery of a small fraction of the full quasispecies variants [3, 53, 54].

The number and frequency of different variants in a quasispecies is affected by host-specific factors like immune responses and by virus-specific factors such as the mutation rate and generation time [3, 10, 12, 25–28, 53]. Although such viruses have high mutation rate in infected hosts, they generate low genomic diversity, which makes quasispecies assembly a hard problem [3, 25–28, 53].
2.2 NGS Technologies

Recent advances in Next Generation Sequencing (NGS) technologies such as the Roche 454 [4], Illumina [5], SOLiD [6], and Ion Torrent [7] have led to massively higher throughput compared to Sanger sequencing [4–7, 55]. Each of previously mentioned platforms is capable of producing millions or billions of base pairs per run with low cost [3–7, 25–28, 53]. NGS is a promising approach to genomics research [3, 25–28, 53] such as individual genome re-sequencing [26, 56, 57], metagenomics [58, 59], DNA methylation [60], histone modifications [61], paleogenomics [62, 63], transcription analysis [64, 65], and small non-coding RNAs [66], NGS is also emerging as a key technology for quasispecies analysis, it can identify low frequency variants due to the high coverage it provides [4–7, 53, 67–71]. Realizing the potential of NGS technologies requires the development of novel methods, developing such methods is challenging due to the short read lengths which make it difficult to match high variable regions flanking long lower variable region [24, 26, 28, 53] and error rates. Tools developed for Sanger reads work poorly when applied to NGS data in terms of quality and runtime [3, 25–28, 53]. Methods designed to reconstruct diploid haplotypes from shotgun genome sequencing data [3, 29–32, 34, 35, 53, 72] are not suitable to reconstruct the set of quasispecies sequences and estimate their frequencies [3, 25–28, 53]. Methods developed for de novo genome assembly from NGS data [3, 25–28, 39, 42, 53, 73, 74] do not perform well when the sequenced sample contains a large number of closely related quasispecies variants.

Many Next Generation Sequencing technologies have been realised as a commercial product (e.g., the Illumina HiSeq Systems/Illumina (Solexa) sequencing, the SOLiD Systems/SOLiD sequencing, 454 Genome Sequencers/454 Pyrosequencing, Ion Personal Genome Machine Sequencer/Ion semiconductor sequencing) [75]. Below we will give a brief description of each sequencing technology.

1. **454 Pyrosequencing.** 454 Life Sciences firm which is owned by Roche Diagnostics, developed a massively-parallel pyrosequencing system for DNA sequencing[4]. The system provides relatively low-cost, high-throughput sequencing with a decent-sized
read length. What is interesting in using 454 technology is that the technology has been delivering longer and longer reads.

Whenever we get longer read length, it would make quasispecies reconstruction less challenging, because the mean length between high variable regions in viral quasispecies is still large that a read may not cover enough SNPs to be able to differentiate between sequences with better accuracy.

454 shotgun sequencing works as follows [4, 25–28, 76, 77](see Figure.2.1):

(a) The DNA or RNA are randomly sheared, each fragment is about 300-800 basepair.

(b) Adapter sequences are added to the 3’ and 5’ ends.

(c) The DNA fragments are attached to Sepharose beads, which are with oligonucleotides complementary to the adapters.

(d) The beads are subjected to emulsion PCR, which amplifies the DNA templates from a single copy to approximately 10 million copies on each bead.

(e) Subsequently, the template-carrying beads are deposited into open wells arranged along one face of a 6060 mm2 fibre-optic slide (picotiter plate). The wells fit only a single bead and each plate contains approximately two million wells.

(f) Reagents are supplied to the picotiter plate for sequential rounds of sequencing by synthesis using a modification of the pyrosequencing method (Ronaghi et al. 1996).

(g) Nucleic acids are supplied to the picotiter plate in a known order, the binding of an acid to a complimentary strand, is accompanied by the release of pyrophosphate, which is converted to the emission of light by a reaction including ATP sylfurylase, luciferase and luciferin. The chemoluminescent signal is captured by a sensitive CCD camera. The location and intensity of the signal determines which base pair is located in which position on which read. 454 system works using the following procedure:
The 454 system can provide the deep sequencing coverage required for viral RNA sequencing [4, 25–28, 77]. It can sequence 400-600 million high-confidence nucleobases in a single 10 hour run, producing reads with an average length of 400 bp [77].

The 454 system is now most frequently used for amplicon studies where a defined region of DNA is sequenced in many samples, or very deeply in one sample.

The difference between shotgun and amplicon reads is that amplicon reads have fixed beginnings and almost have the same length, while shotgun reads beginnings are random, and reads lengths sometimes vary too much.

2. **Illumina (Solexa) Sequencing.** The sequencing method of Illumina is based on
reversible dye-terminators technology, and engineered polymerases, that it developed internally. The basic procedure in an Illumina dye sequencing reaction is as follows [5, 78](see Figure 2.2).

(a) At the beginning, DNA molecules are attached to primers on a slide and amplified so that local colonies, initially coined "clusters", are formed.

(b) The reversible terminate bases (A(adenine), C(cytosine), G(guanine), and T(thymine)) are added, each one is labeled with a different color and attached with a blocking group.

(c) The four bases then compete for binding sites on the template DNA to be sequenced and non-incorporated molecules are washed away.

(d) After each synthesis a laser is applied resulting in the removal of the 3 terminal blocking group and the probe

(e) A fluorescent color specific to each of the four bases is then visible, allowing for sequencing the bases.

The above process is repeated until the full DNA molecule is sequenced [5, 78].
3. SOLiD Sequencing.

SOLiD (Sequencing by Oligonucleotide Ligation Detection) is comprised of the following steps [6, 78] (see Figure 2.3).

(a) Library Preparation. The SOLiD system supports sample preparation for fragment or mate-paired libraries with insert sizes ranging from 60-110bp and 600-6kbp, and libraries with a range of insert sizes are created (fragment or Mate-paired library).

(b) Emulsion PCR and Bead Enrichment Beads. In this step clonal bead populations are generated in water-in-oil microreactors containing P1-coupled beads, PCR components, and primers. Fragment or mate-paired libraries are amplified approximately 30,000 times. A polymerase extends from the P1 adapter after the template anneals to the P1-coupled beads, and a complementary sequence is extended off bead surface followed by template dissociation.

(c) Bead Deposition. The templated beads are 3 modified. And beads deposited on a glass slide in a random array. Different deposition chambers can segment a
slide into one, four or eight chambers, and consequently different samples can be loaded in the same sequence run.

(d) Sequencing by Ligation and Data Analysis. After loading onto the Analyzer, primers hybridize to the adapter sequence within the library template. A set of four fluorescently labeled di-base probes competes to ligate to the sequencing primer. Specificity of the di-base probe is achieved by interrogating every 1st and 2nd base in each ligation reaction. Multiple cycles of ligation, detection and cleavage are performed, with the number of cycles determining the eventual read length. Following a series of ligation cycles, the extension product is removed and the template is reset with a primer complementary to the n-1 position for a second round of ligation cycles. Five rounds of primer resets are completed for each sequence tag.
4. **Ion Semiconductor Sequencing.** Ion Torrent uses simple sequencing chemistry. It does not use enzymatic cascade, or fluorescence, or chemiluminescence, or optics, or light. The Chip is the Machine. Ion Torrent pairs proprietary semiconductor technology with a simple sequencing chemistry that is based on a well-characterized biochemical process. The sequencing process consists of the following steps [7, 78](see Figure 2.4):

(a) In nature, when a nucleotide is incorporated into a strand of DNA by a polymerase, a hydrogen ion is released as a byproduct.

(b) Ion Torrent uses a high-density array of micro-machined wells to perform this biochemical process in a massively parallel way. Each well holds a different DNA template. Beneath the wells is an ion-sensitive layer and beneath that a propri-
etary Ion sensor.

(c) If a nucleotide, is added to a DNA template and is then incorporated into a strand of DNA, a hydrogen ion will be released. The charge from that ion will change the pH of the solution, which can be detected by the proprietary ion sensor. The sequencer will call the base, going directly from chemical information to digital information.

(d) The Ion Personal Genome Machine (PGM) sequencer then sequentially floods the chip with one nucleotide after another. If the next nucleotide that floods the chip is not a match, no voltage change will be recorded and no base will be called.

(e) If there are two identical bases on the DNA strand, the voltage will be double, and the chip will record two identical bases called.

Although of low cost of Ion Torrent technology, some of its disadvantages is its relatively low throughput, error-prone at homopolymers, and washing can lead to errors.

![Figure 2.4 Ion Torrent Sequencing Method [7].](image)

2.3 Existing Tools

Achieving high sequencing depth for viral samples is very inexpensive, owing to the short length of viral genomes. Mapping based approaches to QSR are naturally preferred
to *de novo* assembly since reference genomes are available (or easy to obtain) for viruses of interest, and viral genomes do not contain repeats. Thus, such approaches were mainly adopted in the QSR problem.

The QSR problem has been first addressed directly by [24], and [27]. Both approaches build a read graph, where vertices correspond to reads, edges represent overlaps with agreement between reads, and possible quasispecies sequences are paths from the leftmost to the rightmost vertices in the graph. Eriksson et al. [24] proposed a multi-step approach consisting of genotyping error correction via local clustering, haplotype reconstruction via chain decomposition, and haplotype frequency estimation via expectation maximization (EM) method with validation on HIV-1 data. Their method is similar to Westbrooks et al.[27], in that they build a read graph and use expectation maximization to compute the probability of sequences. But differs in the sense that the approach outlined in [24] doesn’t have any notion of edge costs to favor one solution over another, or the notion of transitively reduced read graph.

Zagordi et al. [46] improved error correction of the method by applying Dirichlet process mixture model for probabilistic local clustering. Each position is covered by 3 consecutive windows and final decision follows a majority rule. The method is implemented in assembling software ShoRAH and validated in HIV-1 population [47]

Prosperi et al.[49] proposed a combinatorial method for QSR using NGS amplicon reads. Their method was developed and applied to HIV and HBV data, using an algorithm based on multinomial distributions. Out of the previous mentioned methods, the following softwares/tools are publicly available:

2.3.1 ShoRAH

ShoRAH [46, 47, 50] goes through the following steps, and outputs the quasispecies spectrum (i.e quasispecies sequences and their relative frequencies):

- **Alignment.** 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.

- **Error correction (local haplotype reconstruction).** While ViSpA uses independent error correction programs, ShoRAH have 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.

- **Global haplotype reconstruction.** This step is similar to assembly of candidate quasispecies sequences in ViSpA.

- **Frequency estimation.** In this step, ShoRAH estimates the frequency of each candidate sequence.

### 2.3.2 QuRe

QuRe [48, 49] reconstructs quasispecies by coupling reads that have consistent overlaps and considering reads that have similar frequencies across the various amplicons. Each amplicon consists of its distinct reads associated with its frequency (i.e., count). After sorting each amplicon in descending order according to read frequency, the algorithm chooses a guide distribution. The decision to select a guide distribution can be made either by random selection or by a simple chi-squared test. It then selects the most frequent read within the distribution and goes left or right reconstructing sub-quasispecies. It then selects the neighbor closest in frequency that has a consistent overlap until it reaches the first end. In the same way, it expands the sub-quasispecies to the other end. The frequency of the guide read is then subtracted from the selected reads’ frequencies. The reconstructed quasispecies has a count equal to the initial read count of the guide amplicon. The algorithm continues reconstructing the next quasispecies by choosing another guide distribution among amplicons.
after having updated the reads’ frequencies. This process is repeated until the last read from
the guide distribution has been evaluated, or until a distribution has all zero-frequencies.

2.3.3 Error Correction

Next Generation Sequencing (NGS) technologies produce large sets of short reads that
may contain errors. These sequencing errors make quasispecies assembly challenging. Error
correction aims to reduce the error rate prior assembly[79].

Some methods use a reference for error correction, as in ShoRAH program [46]. Zagordi
et al [46] develop a generative probabilistic model for assigning observed reads to unobserved
haplotypes in the presence of sequencing errors. This clustering problem is solved in a
Bayesian fashion using the Dirichlet process mixture to define a prior distribution on the
unknown number of haplotypes in the mixture. The method devise a Gibbs sampler for
sampling from the joint posterior distribution of haplotype sequences, assignment of reads
to haplotypes, and error rate of the sequencing process, to obtain estimates of the local
haplotype structure of the population. The method is evaluated on simulated data and on
experimental deep sequencing data obtained from HIV samples.

Other methods don’t use a reference for error correction. Skums et al [80] develop new
efficient error correction algorithm (KEC program), for detecting and removing sequencing
errors from reads (prior to sequence assembly). For each input read, the set of all length k
substrings (k-mers) are calculated. The read is evaluated based on the frequency with which
each k-mer occurs in the complete data set (k-count). For each read, k-mers are clustered
using the variable-bandwidth mean-shift algorithm. Based on the k-count of the cluster
center, clusters are classified as error regions or non-error regions.

It is preferred to use methods that don’t use reference, like the last two methods that
uses k-mer calculations, because corrected reads using reference could be biased.
PART 3

QUASISPECIES SPECTRUM RECONSTRUCTION FROM SHOTGUN READS

3.1 ViSpA

ViSpA [25–28], an existing tool, consists of two main stages. First, it solves Quasispecies Sequences Assembly (QSA) problem by aligning 454 Life Sciences reads, constructing read graph and assembling candidate quasispecies sequences corresponding to maximum minimum edge cost paths in the reads graph. Then it solves Quasispecies Frequencies Estimation (QFR) problem by EM-based algorithm that finds maximum likelihood frequency estimates of candidate quasispecies sequences based on the given 454 LifeScience reads.

The flow for solving the QSA problem (see Figure. 3.1) includes the following steps:

- read error correction.
- alignment of 454 Life Sciences reads either versus reference or iteratively versus consensus.
- preprocessing of aligned reads to deal with indels and missing values.
- constructing a read graph to represent all possible haplotype sequences.
- assembling candidate sequences.
- frequency estimation of haplotype sequences.
3.1.1 Read Error Correction

This step is necessary since the reads produced by 454 Life Sciences system are prone to errors, it can erroneously sequence one bp per 1000bp [4], and it is important to distinguish reading errors from (rare) viral variants. The error rate is strongly related to the presence and size of homopolymers [81], i.e., genome regions, consisting of consecutive repetition of a single base (for example, TTTTTT). For more flexibility, ViSpA uses independent programs (Kec [80], Saet [82]) to do error correction before assembly.

3.1.2 Read Alignment

In this step, ViSpA uses an independent alignment program to map reads against a reference viral sequence. Originally SEHEMEHL aligner [83] was used. Unfortunately SEGEMEHL aligner did not perform well and users complained. Therefore it was replaced by
Mosaik [84] aligner, it can be replaced by other aligners. After performing multiple sequence alignment of reads, the reference and reads will be extended. The symbol $I$ in reads means that there is insertion $I$ in the reference in that position, and $D$ in reads means that there is a missing nucleotide in that position with respect to the reference.

3.1.3 Processing of 454 Reads

As mentioned in the previous step, ViSpA uses placeholders $I$ and $D$ to reference insertions and deletions of aligned reads.

Deletion supported by a single read is replaced either with the base pair in all other reads in the same position, or with $N$ (unknow). All insertions supported by a single read are removed from consideration.

3.1.4 Read Graph

In the read graph each vertex corresponds to a read and each directed edge connects a consecutive pair of reads which can possibly come from the same quasispecies sequence, i.e., if the left read overlaps with the right read and they (almost fully) agree in the overlapped positions. Then each quasispecies sequence would correspond to a path from the leftmost to rightmost vertex. ViSpA differentiates between two types of reads, super-reads and sub-reads, Where the overlap between the two reads completely subsumes one of the reads(subread). The read graph consists only of super-reads.

3.1.5 Candidate path selection.

To generate a set of paths rich enough to explain observed reads, ViSPA uses what so called max-bandwidth paths.

3.1.6 Estimation of candidate quasispecies sequence frequencies.

In this step, ViSpA uses Expectation Maximaization algorithm to estimate the frequency of each candidate sequence. Where it maps all reads (super-reads and sub-reads) to corresponding candidate quasispecies via bipartite graph and run EM.
### 3.2 Fixing Erroneous Homopolymer Indels via Protein Alignment

A well known drawback of 454 pyrosequencing system is the erroneous indels in homopolymers. Quasispecies sequences may partition the reads between themselves and some of them can easily have majority of wrong reads. We have encountered this problem when checking the proteins coded by quasispecies sequences, many had unexpected stop-codons. Fortunately, such errors can be directly addressed if they happen in the coding region. We suggest to repeatedly apply the following procedure:

1. find the frame with the start-codon, and align the amino-acid decoded sequence with the one for the reference sequence.

2. in the region immediately preceding stop-codon find the position $X$ when the protein alignment becomes very poor.

3. in the nucleotide quasispecies sequence the position $X$ should correspond to a homopolymer which should be either extended or reduced by 1 (the correct fix will make the protein alignment almost perfect).

Figure 3.2 shows the protein alignment (CLUSTAL W alignment [85, 86]) of a quasispecies constructed by ViSpA with a reference. There are 2 stop codons at the end part of the quasispecies sequence. In the region preceding the stop codon the protein alignment becomes very poor, go to the left until alignment agree (around 100 base pair from stop codon), go to the closest homopolymer (a single repeating nucleotide), you will see the amino acids (PP) which are the homopolymers (CCC CCA) at nucleotide position (5074), after deleting nucleotide (C) at position 5078, the quasispecies is aligned with consensus as shown in Figure 3.3.
Figure 3.2 protein sequence alignment (CLUSTAL W) of the quasispecies sequence (0.0498579823867_ID20_*1) with the reference (contig20_*3) before systematic error fixing.
Figure 3.3 protein sequence alignment (CLUSTAL W) of the quasispecies sequence (0.0498579823867_ID20_*._1) with the reference (contig20_*._3) after systematic error fixing.
Quasispecies Spectrum Reconstruction Problem (QSR) can be formulated as follows. Given a set of NGS 454 amplicon reads generated from a viral sample, reconstruct the quasispecies sequences, and estimate the relative frequency of each sequence in the sample.

4.1 Theoretical Analysis of Quasispecies Assembly Problem for Amplicon Reads

Our model, consists of $K$ overlapping windows with predefined positions within the genome, called amplicons [1, 2]. Each amplicon $W_1, \ldots, W_K$, is assumed to have the same length equal to the length of a single read and is sequenced to the same depth $D$, i.e., covered with $D$ reads [1, 2]. In our analysis, we distinguish between two error models.

- Error-free Model: assumes that all reads don’t have genotyping errors, or corrected, using some genotyping error correction methods.

- Error-prone Model: reads contain genotyping errors.

We also distinguish two frequency models.

- Ideal-frequency Model: assumes that the reads distribution in each amplicon is identical to the true distribution of quasispecies.

- Skewed-frequency Model: in each amplicon, the quasispecies distribution is represented differently from the true distribution, this is considered more realistic than the previous model.
In the next section, we address the QSR problem in the error-free, ideal-frequency model, and then show how we can adjust frequencies to reduce the skewed-frequency model to the ideal-frequency model. Note that shotgun reads sequencing is more prone to genotyping errors than amplicon reads sequencing. On the other hand, amplicon reads sequencing is more prone to frequency skewness.

The methods used to reconstruct quasispecies from shotgun reads (e.g., ViSpA), heavily rely on the uniform distribution of reads starting positions with respect to the reference, which allows more accurate estimate of the probability of two overlapping reads coming from the same quasispecies.

In case of quasispecies reconstruction from amplicon reads, it is necessary to rely on parsimonious considerations, rather than the distribution of reads starting positions with respect to the reference, because this information is not available, and we have only fixed amplicons starting positions. Further, we discuss several optimization formulations. The following formulation is standard, but difficult to solve [1, 2].

**Most Parsimonious Spectrum.** The most parsimonious spectrum requires the minimum number of distinct quasispecies that explains the observed reads. The problem is NP-hard by simple reduction from SUBSET SUM [1, 2]. But, we expect a small number of distinct reads in each amplicon, i.e. small number of quasispecies. This allows us to solve the problem in practical time. The sets of overlapping amplicons are partitioned into subsets with the same parts in the intersection of these amplicons. The number of distinct reads in the overlap is at most the minimum of the number of distinct reads in the previous and next amplicons. In next sections, we will discuss several models and corresponding optimization formulations [1, 2].

4.1.1 Quasispecies Assembly in the Error-Free, Ideal-Frequency Model Problem

The input data can be viewed as a $K$-staged read graph $G = (V = V_1 \cup \cdots \cup V_K, E)$ [1, 2], where

- each vertex $v$ in $V_i$ corresponds to a distinct read in the $i$-th amplicon $A_i$ and has a
count $c(v)$, where the total counts is equal to amplicon depth
\[ \sum_{v \in V_i} c(v) = D. \]

- each edge $(u, v)$ connects two reads from consecutive amplicons $A_i$ and $A_{i+1}$ which have consistent overlap.

To get the set of reconstructed quasispecies $Q = \{ q_j \}$, we need to solve the following problem [1, 2]:

**find all $u - v$-paths. where $u \in V_1$, $v \in V_K$, each with the frequency $f_j$ such that for each vertex $v \in V$**

\[ \sum_{v \in q_j} f_j = c(v) \quad (4.1) \]

If we are given $K$ amplicons $A_1, \ldots, A_K$ sequenced to the depth $D$, we need to assemble the most likely full-genome quasispecies and find their frequency distribution.

We can solve the $K$-staged amplicons assembly problem, by solving the two-staged case, the solution to which is then used to “stitch” together all $K$ stages. So, let us assume that there are only two stages, $V_1$ and $V_2$, thus implying that the read graph is bipartite. Then we need to look for the existence of the feasible solution exists this problem.

The problem can be reduced to linear equations consistency. Let $f_e$ be the frequency of the quasispecies $e$ corresponding to the edge $e = (u, v)$. Then for each vertex we write the following constraint (4.1) to obtain the following system of linear equations [1, 2]:

\[ \forall v \in V_1 \cup V_2 \quad \sum_{e \text{ incident to } v} f_e = c(v) \]

(4.2)

The system of equations (4.2) is consistent if and only if the corresponding two-stage assembly problem is feasible.

From maximum likelihood point, we can assume that any edge (i.e., quasispecies) is equally probable. So, it is possible to assign non-zero frequencies to all possible quasispecies. The parsimonious approach requires that most likely solutions should contain the minimum
number of quasispecies satisfying the linear system (4.2).

Let us make the case of two distinct reads for both amplicons. Assume that $|V_1| = |V_2| = 2$, $A$ and $B$ are distinct reads in the first amplicon and $C$ and $D$ are in the second (see Figure 4.1). Furthermore, let all four possible combinations of reads be consistent in their overlap $[1, 2]$.

![Figure 4.1 The case of two distinct reads for both amplicons $[1, 2]$.](image)

Looking on the left side, we can see four reads. We represent them as a graph on the right side of Figure 4.1. Let us assume that $d \leq b \leq a \leq c$. If $a = c$, then $b = d$ and we can have the minimum possible number of two non-zero edge frequencies. If $a \neq c$, then the four constraints have rank 3 and there should be 3 edges with non-zero frequency. There are two possibilities for three non-zero frequency edges $[1, 2]$:

- $AC = a$, $AD = 0$, $BC = c - a$, and $BD = d$

- $AC = a - d$, $AD = d$, $BC = b$, and $BD = 0$

The first case is more probable if $a > b$ and are equally probable if $a = b$. Therefore maximum likelihood does not seem as a good approach. So we should expect that maximizing parsimony is better.
Unfortunately the parsimony is an integral objective, i.e., does not addresses the frequencies of quasispecies. The fractional relaxation of parsimony is Shannon entropy (later referred as entropy) and maximizing parsimony is equivalent to minimizing entropy. The entropy is defined as [1, 2].

\[ -\sum_{i=1}^{n} p(x_i) \log p(x_i) \] (4.3)

The more quasispecies we have, the higher the entropy value. Therefore, finding a solution with the minimum amount of quasispecies will reduce the entropy of the solution, which corresponds to the idea of a most parsimonious solution [1, 2].

Given read graph with frequencies on reads, find the set of quasispecies (paths) with frequencies explaining all reads and their frequencies and having the minimum entropy.

The read graph for two amplicons consists of a set of disjoint bi-cliques each corresponding to a distinct overlap-part, so their number cannot exceed minimum of the number of distinct reads reads in the first and the second amplicon.

4.1.2 Reduction of Skewed-Frequency Model to Ideal-Frequency Model

This problem can be reduced to the case in which all frequencies of the reads are ideal. First, we formulate the problem of balancing forks in the graph, which is equivalent to estimating ideal frequencies of the reads.

A fork \( f = (L, R) \), where \( L \) and \( R \) are the sets of reads from the left and the right amplicons, respectively, is balanced if the total frequencies of \( L \) and \( R \) are equal to each other. It is obvious that any ideal-frequency amplicon-read spectrum has all its forks balanced. The opposite is also true, i.e., any amplicon-read spectrum with balanced forks has a feasible quasispecies spectrum which has ideal frequencies.

Given a spectrum of amplicon reads \( R = \{r_i\} \) with observed frequencies \( O = \{o_i\} \) and forks \( f_j = (L_f, R_f) \) find ideal frequencies \( F = \{f_i\} \) such that all forks are balanced.

**Least Squares Approach via Quadratic Program.** We model the fork balancing problem as a quadratic program (QP) with linear constraints (equation 4.4). The approach looks
for the minimum squared adjustment of observed read frequencies, \( X = \{x_i\} \). We can generalize this method by allowing weights to be used in the objective. The justification for allowing weights is to normalize adjustment values. We expect that a read with high frequency may be adjusted to a greater extent than a read with a very small frequency. Typical weight values for reads are \( w_i = \frac{1}{c_i} \), where \( c_i \) is the frequency of read \( i \) [1, 2].

\[
\text{Minimize : } \sum w_i x_i^2 \quad (4.4)
\]

\[
\text{Subject to : } \sum_{i \in L_f} (x_i + o_i) = \sum_{i \in R_f} (x_i + o_i), \quad \forall f \in \text{Forks} \quad (4.5)
\]

\[
x_i + o_i \geq 0 \quad (4.6)
\]

4.2 Viral Quasispecies Reconstruction Algorithms From Amplicon Reads

4.2.1 Read Graph

We begin with the definition of the reads graph, since it is the first step of our solution to the viral quasispecies assembly problem[1, 2, 27, 87]. The reads graph is a data structure that describes all possible assemblies of the reads into quasispecies. Edges and vertices have cost.

Building a \( K \)-staged read graph from \( K \) amplicons is straightforward. For each read in an amplicon \( A_i \), check for consistent reads in amplicon \( A_{i+1} \). If a consistent read exists, add both reads as vertices to the graph with a directed edge joining them. Repeat this for \( K \) amplicons \( A_i, \ldots, A_K \). Add a single source \( S \) connected with all reads in the first amplicon and single sink \( T \) with edges from all reads in the last amplicon.

In case we wish to adjust the frequencies. Then this may be done by either solving the QP (eq. 4.4) or by simple scaling.

To scale the frequencies of each complete bipartite subgraph in the directed graph \( G \), find the sum of the frequencies in the first and second partitions. If they do not equal one another, scale the values in the right partition to frequencies in the left. This may be
completed for the entire graph in a breadth first manner from the source to the sink.

The graph will be composed of bipartite cliques (bi-clique), with each clique representing a consistent overlap in the reads. The graph is then transformed by adding a fork vertex for each bi-clique. Given a bi-clique $K_{n,m}$, add an edge from $n$ vertices in the first partition to the fork vertex and an edge from the fork vertex to the $m$ in the second partition. This modification reduces the amount of $n \cdot m$ edges from each original bi-clique to a “forked” subgraph with $n + m$ edges. This is repeated for all bi-cliques in the graph [1, 2].

4.2.2 Maximum Bandwidth Path

Maximum bandwidth path is a path in a graph $G$ which maximizes the minimum edge capacity path from a source $s$ to a sink $t$. After finding such path, all edges weights are reduced by that value (maximum minimum edge capacity path value). The process is repeated until no connection exists between source and sink [1, 2].

4.2.3 Maximum Frequency Path.

Maximum frequency path, is a path in a graph $G$ from source $s$ to sink $t$, with maximum capacity. After finding such path, all its edges are removed, the we repeat the process until ther is no connection between the source and sink [1, 2].

4.2.4 Greedy Algorithm for Resolving Forks.

A fork in a graph $G$, is a vertex with at least two in-edges and two out-edges. We choose a path in a greedy way. Assume we have $L$ vertices before the fork, and $R$ vertices after, then we choose the maximum weighted vertices from $L$ & $R$ (using a priority queue) and create a new edge between them after subtracting their weights by the minimum of both. The process is repeated until no forks are left [1, 2].
4.3 Data Sets

4.3.1 Simulated Reads

In order to perform cross-validation on the assembly method, we simulate reads data from 1739-bp long fragment from the E1E2 region of 44 HCV sequences [88] where sequence frequencies are generated according to some specific distribution. In our simulation experiments, we use a uniform, geometric, and skewed distribution to create sample quasispecies populations with different number of randomly selected above-mentioned quasispecies sequences.

The reads are simulated without sequencing errors. The length of a read follows a normal distribution with mean value of 320 bp and variance 10bp. The starting position follows the uniform distribution within the set of amplicons starting positions. The total amount of reads varies from 5K,20K, to 100K [1, 2].

4.3.2 HBV Data

We used real HBV amplicon reads sampled from two patients [1] to reconstruct quasispecies using the previous methods. All reads were first aligned to a reference genome using MOSAIK Aligner [84], then cleaned by cutting the remainders of primers from both ends. Reads were corrected using the method described in [80] separately for each amplicon. The corrected reads were then realigned to the reference genome to facilitate in identifying distinct reads within amplicons.

4.4 Validation Methods

In our experiments, validation were done on simulated reads without errors given the real quasispecies from which the reads were generated. We use two measures of quality given a solution to the QSR problem. By measuring a solution both in terms of correctly assembled quasispecies, as well as how close their respective frequencies are to the actual quasispecies population.
4.4.1 Quasispecies Assembling Validation

The quality of assembling is measured by sensitivity which is analogous to recall rate, and positive predictive value (PPV) which is analogous to precision. These measures are defined as [1, 2, 89]:

- sensitivity/recall rate: the portion of real quasispecies that have been correctly inferred.
- PPV/precision: the portion of real quasispecies that have been correctly inferred among all assembled quasispecies.

4.4.2 Validation of Frequency Distribution

In the case of simulated reads, we use Jensen-Shannon Divergence (JSD). Jensen-Shannon divergence is a quasi-metric that measures the divergence or distance between two statistical distributions. It differs from the Kullback-Leibler divergence due to the addition of a midpoint. It is defined as [1, 2]:

\[
\text{JSD}(P||Q) = \frac{1}{2} D_{KL}(P||M) + \frac{1}{2} D_{KL}(Q||M)
\]

where

\[
D_{KL}(P||Q) = \sum_{i=1}^{n} P(i) \log \frac{P(i)}{Q(i)}
\]

4.5 Experimental Results

4.5.1 Simulated Reads

Ten different data sets were used to generate quasispecies following a skewed, uniform, and geometric distribution with different shift sizes (overlap between amplicons). The shown results are the mean value of the ten instances per configuration, on 100K reads and window length 300bp [1, 2].

Figure 4.2 shows that maximum bandwidth outperforms other methods (maximum
frequency, greedy, and guide distribution) in terms of sensitivity for quasispecies with uniform
distribution for shift values 50, 100, 150, 200. The quality of all three methods degrades when
the shift position increases, since shorter overlap produces more dense bicliques. Figures 4.3,
4.4 show the results of the skewed and geometric distributions respectively where maximum
bandwidth dominates other methods for shift values 50, 100, 150, greedy method performs
slightly better than maximum bandwidth for shift values 200,250 (note that sensitivity of
value 0.9 means that 9 quasispecies are correctly reconstructed out of 10 ). Figures 4.5
- 4.7 shows the positive predicted values, where maximum frequency path competes other
methods since it covers many reads in each iteration.

The quality of the quasispecies frequencies was measured using Jensen-Shannon Divergence.
Maximum bandwidth and greedy fork resolution methods outperform maximum
frequency and GD on average. All four methods benefit from shorter shifts (see figures 4.8
- 4.10) [1, 2].
Figure 4.2 Sensitivity for different amplicon window shifts from uniform distribution [1, 2]

Figure 4.3 Sensitivity for different amplicon window shifts from skewed distribution [1, 2]
Figure 4.4 Sensitivity for different amplicon window shifts from geometric distribution [1, 2]

Figure 4.5 PPV for different amplicon window shifts from uniform distribution [1, 2]
Figure 4.6 PPV for different amplicon window shifts from skewed distribution \([1, 2]\)

Figure 4.7 PPV for different amplicon window shifts from geometric distribution \([1, 2]\)
Figure 4.8 Jensen-Shannon Divergence for different amplicon window shifts from uniform distribution [1, 2]

Figure 4.9 Jensen-Shannon Divergence for different amplicon window shifts from skewed distribution [1, 2]
Comparison With ViSpA. Table 4.1 compares the sensitivity/PPV of the maximum bandwidth on amplicon reads with ViSpA [25] on shotgun reads. Maximum bandwidth outperforms ViSpA in all cases except for twenty quasispecies and uniform distribution, on the other hand ViSpA outperforms maximum bandwidth using uniform distribution at 100k reads, its performance degrades quickly as read count goes down [1, 2].
4.5.2 Real Reads

Table 4.2 shows the starting and ending positions of amplicons along with the total number of reads and total number of distinct reads. Due to the lack of reads in several amplicons, reconstruction of quasispecies was focused over a subset of a continuous region. In particular, the highly variable region of patient one (positions 2309 through 2748) and a conserved region of patient two (positions 1109 through 1642) were chosen. In patient one, maximum bandwidth, maximum frequency, and greedy fork resolution methods produced three, two, and eleven quasispecies respectively. In patient two, both maximum bandwidth and greedy fork resolution reconstructed three quasispecies while maximum frequency reconstructed one [1, 2].
Table 4.2  Total and distinct reads of amplicons per patient [1, 2].

<table>
<thead>
<tr>
<th>File</th>
<th>Patient 1</th>
<th></th>
<th>Patient 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># Reads</td>
<td># Distinct Reads</td>
<td># Reads</td>
<td># Distinct Reads</td>
</tr>
<tr>
<td>amp33-233</td>
<td>2130</td>
<td>1</td>
<td>233</td>
<td>119</td>
</tr>
<tr>
<td>amp122-287</td>
<td>115</td>
<td>35</td>
<td>20</td>
<td>6</td>
</tr>
<tr>
<td>amp267-476</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>amp438-623</td>
<td>0</td>
<td>0</td>
<td>27</td>
<td>1</td>
</tr>
<tr>
<td>amp602-798</td>
<td>593</td>
<td>1</td>
<td>89</td>
<td>1</td>
</tr>
<tr>
<td>amp711-899</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>amp800-991</td>
<td>396</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>amp962-1130</td>
<td>437</td>
<td>1</td>
<td>213</td>
<td>1</td>
</tr>
<tr>
<td>amp1109-1273</td>
<td>384</td>
<td>1</td>
<td>319</td>
<td>2</td>
</tr>
<tr>
<td>amp1207-1403</td>
<td>1552</td>
<td>1</td>
<td>1099</td>
<td>2</td>
</tr>
<tr>
<td>amp1263-1463</td>
<td>2370</td>
<td>1</td>
<td>647</td>
<td>2</td>
</tr>
<tr>
<td>amp1445-1642</td>
<td>1260</td>
<td>1</td>
<td>200</td>
<td>2</td>
</tr>
<tr>
<td>amp1604-1759</td>
<td>2487</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>amp1751-1923</td>
<td>4624</td>
<td>11</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>amp1871-2044</td>
<td>390</td>
<td>1</td>
<td>938</td>
<td>2</td>
</tr>
<tr>
<td>amp2070-2236</td>
<td>838</td>
<td>1</td>
<td>652</td>
<td>2</td>
</tr>
<tr>
<td>amp2223-2388</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>amp2309-2508</td>
<td>21</td>
<td>3</td>
<td>30</td>
<td>29</td>
</tr>
<tr>
<td>amp2424-2629</td>
<td>5033</td>
<td>26</td>
<td>270</td>
<td>1</td>
</tr>
<tr>
<td>amp2576-2748</td>
<td>147</td>
<td>30</td>
<td>1081</td>
<td>1</td>
</tr>
<tr>
<td>amp2715-2913</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>amp2752-2945</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>amp2938-3133</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
PART 5

RECONSTRUCTION OF IBV QUASISPECIES

5.1 Introduction

RNA viruses are causing a significant burden on the health and productivity of animals since the evolution of RNA viruses within infected animals is coupled with frequent transmissions between them [3, 53]. Poultry farms are also exposed to such viral infections especially when there are thousands of birds living together in one farm and these infections could cause a significant economic loss for the farm [3, 53, 90].

In the U.S., Infectious Bronchitis Virus (IBV) is considered to be one of the biggest cause of economic loss in poultry farms, it infects the domestic fowl [91], causing respiratory disease and other infections. IBV has constant population growth and is distributed worldwide [92]. Chicken may get infected many times during their lifetime, IBV infection reduces growth and drops down egg production [90], young chicken may die from IBV infection, attenuated live vaccines are commonly used [93] but protection work for short time and chickens have to be revaccinated many times [3, 53, 90].

Many serotypes of IBV can co-circulate in a certain region, creating conditions for recombination between strains [3, 53, 90, 94]. As IBV quasispecies exist in infected poultry [95], they exist also in attenuated live vaccines [93]. Attenuated live vaccines may undergo virulent following vaccination [53, 90, 93, 96].

Viral quasispecies sequences reconstruction by analyzing NGS data will contribute to understanding the roles and interactions of infected animal and viral quasispecies for improving animal health, and productivity [3, 53]. In our work, we propose a computational pipeline for IBV quasispecies reconstruction consisting of 3 steps: (1) Read Error Correction (2) Read Alignment and (3) Reconstruction of Viral Quasispecies [3, 53]. We vary different parameter values of the pipeline, i.e parameter tuning, to get better experimental results
for the reconstructed quasispecies compared to sanger clones which we use as a ground truth.

5.2 Methods

Many tools developed for Sanger reads do not work well or have impractical runtimes when applied to NGS data [3, 53]. Newly developed algorithms for de novo genome assembly from NGS data work well for reconstruction of haploid genomes, but work poorly when the sequenced sample contains a large number of closely related sequences as in the case of viral quasispecies. To address these shortcomings we introduce a computational pipeline for accurate reconstruction of viral quasispecies sequences and estimate their frequencies from High Throughput Sequencing (HTS) data, where it incorporates different NGS reads error correction methods, aligners, and genome assemblers (ViSpA [25], and ShoRAH [46]), using different tuning parameters. We apply experiments on IBV 454 shotgun reads [3, 53], collected from commercial poultry farms. For method validation, we use IBV sanger clones (as ground truth) [3, 53].

The proposed computational pipeline for quasispecies reconstruction, consists of the following stages (see Figure 5.1) [3]:

1. **Read Error Correction.** This step is necessary since the reads produced by 454 Life Sciences system are prone to errors and it is important to distinguish reading errors from (rare) viral variants. As mentioned before, 454 Life Sciences can erroneously sequence one base pair per 1000 base pair [4]. The error rate is strongly related to the presence and size of homopolymers [81], i.e., genome regions, consisting of consecutive repetition of a single base (for example, TTTTTT).

We use KEC, SAET, and ShoRAH programs to do error correction prior to assembly. These error correction algorithms involve clustering of reads. While ShoRAH clusters the reads in Bayesian fashion using the Dirichlet process mixture [46], and KEC clusters reads based on kmers [80], SAET uses reads quality scores for error correction [82].
2. **Read Alignment.** In this step, we use independent alignment program to map reads against a reference viral sequence\[84\], this aligner can be easily replaced with another one.

3. **Reconstruction of Viral Quasispecies.** In this step, we reconstruct quasispecies from aligned reads, and estimate their relative frequencies. In this step, we use two assembly programs ViSpA \[25\], and ShoRAH \[46\]. ViSpA and ShoRAH go through different steps for quasispecies reconstruction, and frequency estimation. For more details, refer back to 2.3.1 and 3.1.

![Figure 5.1 Viral quasispecies reconstruction pipeline [3]](image)

5.2.1 **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 [3].

5.3 Data Sets and Golden Standards

Reads samples were collected from IBV infected chickens, and quasispecies variants were sequenced using life sciences 454 shotgun sequencing, followed by Sanger sequencing of individual variants. Ten Sanger clones (c1, ..., c10) were used for validation. These clones are considered as the golden standards or the ground truth for parameter calibration and comparison between different methods (see Figure 5.2) [3].

Figure 5.2 Schematic representation of calibration, validation experiments based on cloned quasispecies [3].
5.4 Validation

The validation methods in this section will be different from those defined previously in section 4.4 for simulated reads. In this case, we don’t know the real or original quasispecies sequences, the only ground truth we have are the Sanger clones.

This section will discuss three validation methods, validation of error correction methods, validation of IBV reads, and validation of the different methods used to reconstruct quasispecies using different parameter settings.

5.4.1 Validation of 454 Reads

In order to validate 454 IBV reads, we compare them with the ground truth data, i.e Sanger clones. We use 10 Sanger clones to validate reads generated from the same clone pool as follows. For each one of the 10 clones, we compute the edit distance to each IBV read, then we categorize the reads according to edit distance value for the corresponding clone (0, 1, ..., 15 edit distance classes), we call each class a mismatch. After that, we compute the average coverage of each mismatch class to its corresponding clone.

5.4.2 Validation of Correction Methods

We validate each correction method as follow. We pick a read from the set of corrected reads using particular correction method, then we calculate the edit distance of the read to each clone from the set of Sanger clones (gold standards), and choose the minimum edit distance to some clone. Then we consider that read belongs to that edit distance category or class. We repeat the process for all other corrected reads.

Then we count the number of reads per category (i.e minimum edit distance). By this method, we can check the number of corrected reads per edit distance category for each correction method. The correction method with more minimum edit distance reads, in average, is better.
5.4.3 Comparison Measures

Before defining different methods (with different parameter settings) validation, we need to define the following parameters:

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c_i$</td>
<td>Sanger clone $i$, $1 \leq i \leq 10$;</td>
</tr>
<tr>
<td>$q_j$</td>
<td>reconstructed quasispecies $j$;</td>
</tr>
<tr>
<td>$f_{c_i}$</td>
<td>frequency of Sanger clone $i$;</td>
</tr>
<tr>
<td>$f_{q_j}$</td>
<td>frequency of reconstructed quasispecies $j$;</td>
</tr>
<tr>
<td>$m_i$</td>
<td>how far is clone $i$ from closest reconstructed quasispecies;</td>
</tr>
<tr>
<td>$m_j$</td>
<td>how far is reconstructed quasispecies $j$ from closest clone;</td>
</tr>
</tbody>
</table>

$m_i$ 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 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.
5.5 Experimental Results

5.5.1 Results of 454 Reads Validation

Using Moasik Aligner, we map the 454 IBV reads to each of the 10 Sanger clones, we get the edit distance for each 454 read(or mismatch), and categorize reads according to 0,1,2...,15 mismatches to corresponding Sanger clone, for each category. We compute the average coverage corresponding to the Sanger clone, by summing up the coverage count of each position in the clone sequence, then dividing by its length.

After the previous procedures, we got the average coverage for uncorrected reads (Figure 5.3), reads corrected by SAET (Figure 5.4), reads corrected by KEC (Figure 5.5), and reads corrected by ShoRAH (Figure 5.6).

From the figures, we can see that clone 42H5_A12_01 (the suffix 01 means that the frequency of this Sanger clone is 1%) is an outlier since its peak coverage is high, around 9 mismatches. We can also see another outlier 42V1H3_G01_04 which have slight coverage through all mismatch values.

As we see from the figures, and by assuming that Sanger clones are the ground truth, we get the highest average read coverage when mismatches are less than 5 for uncorrected and SAET corrected reads, and less than 3 mismatches for KEC and ShoRAH corrected reads.
Figure 5.3 Uncorrected reads validation.

Figure 5.4 SAET corrected reads validation.
Figure 5.5  KEC corrected reads validation.

Figure 5.6  ShoRAH corrected reads validation.
5.5.2 Results of Correction Methods Validation

To validate the error correction methods in our experiments. We map each of the 454 IBV reads to all of the 10 clones using MOSAIK Aligner, and we get the minimum edit distance for each 454 IBV read to some Sanger clone. Figure 5.7 shows the number of reads per category (i.e minimum edit distance to some clone) for all used correction methods.

From Figure 5.7 we see that, for each minimum edit distance category, ShoRAH outputs more reads than KEC, which comes in the second place. Then SAET comes in the third place, we notice that SAET is too conservative, that it covers nearly the same number of reads as uncorrected per minimum edit distance category.

![Minimum Edit Distance Counts between Reads and Sanger Clones](image)

Figure 5.7 The distribution of 454 IBV reads categories(edit distance to the closest Sanger clone) for different correction methods.

5.5.3 Results of Reconstructed Quasispecies

Table 5.1 shows the the pairwise edit distance values for the 10 Sanger clones. The suffix number of the clone id is the clone frequency, e.g 42E9_A08_20 means that the frequency of the clone 42E9_A08 is 20%.

We measure the pairwise edit distance for the 10 Sanger clones as follow. We do pair-
wise alignment for the Sanger clones using clustalW, then cut the sticking out ends, and get overlapped region for first clone to others, and compute the edit distance (Levenshtein Distance) between each pair. The process is repeated for the second, the third, until the last clone [3].

From the table we can see that the clone id 42V1H3_G01_04 is an outlier, since its edit distance to other clones is large.

Table 5.1 Pairwise edit distance between the 10 Sanger clones [3].

<table>
<thead>
<tr>
<th>Clone ID</th>
<th>42V0.H08_20</th>
<th>42V1H7_G02_20</th>
<th>42V3.C07_15</th>
<th>42V4.F04_15</th>
<th>42V10_B08_10</th>
<th>42V6_E11_10</th>
<th>42V1B10_B04_04</th>
<th>42V1B13_G01_04</th>
<th>42V1H5_A12_01</th>
<th>42V1C6_B08_01</th>
</tr>
</thead>
<tbody>
<tr>
<td>42V0.H08_20</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>42</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>42V1H7_G02_20</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>41</td>
<td>6</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>42V3.C07_15</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>6</td>
<td>44</td>
<td>9</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>42V4.F04_15</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>4</td>
<td>42</td>
<td>9</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>42V10_B08_10</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>42</td>
<td>8</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>42V6_E11_10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>40</td>
<td>0</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>42V1B10_B04_04</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>45</td>
<td>11</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>42V1B13_G01_04</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>41</td>
<td>41</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>42V1H5_A12_01</td>
<td></td>
<td>0</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>42V1C6_B08_01</td>
<td></td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In order to validate the reconstruction of quasispecies using different methods. We first calculate the pairwise edit distance between Sanger clones and reconstructed quasispecies. We measure the distances as follow. We do pairwise alignment between Sanger clones and reconstructed quasispecies using clustalW, then cut sticking out ends, and get overlapped region between each clone and each quasispecies, the reason behind that, is to avoid calculating edit distance for uncovered fragments (all overlapped regions happened to be more than 500 bp long, which is close to Sanger clones average length (=550 base pairs)).

As a result of the previous step, we get several groups of identical overlapped regions of reconstructed quasispecies, so we collapse them, and their frequencies are summed up (the suffix of the quasispecies sequences ID’s represent the collapsed quasispecies abundance, for
example, the frequency of quasispecies id 42E9_A08_40 is 40%.

Identical Sanger clone fragments are also collapsed (for example, the 2 most frequent clones 42E9_A08_20 and 42V1H7_C02_20 are collapsed to 42E9_A08_40). Then we calculate edit distance (Levenshtein distance) for every overlapped region [3].

Tables 5.2 and 5.3 show distance values between Sanger clones and reconstructed quasispecies for two different methods. In the first method, ViSpA assembler is used, the number of mismatches between superreads and subreads (denoted by n) is 1, the number of mismatches in the overlap between two reads (denoted by m) is 2, the mutation rate (denoted by t) is 5, we considered only reconstructed quasispecies whose frequencies are above threshold (=0.005 in both methods), and IBV reads are corrected using KEC program. In the second method, we use the same setting values as the first one, but for uncorrected reads [3].

Table 5.2 Edit distance between collapsed Sanger clones and reconstructed quasispecies using parameters 1_2_5 (the number of mismatches between sub-reads and super-reads, the number of mismatches between two overlapped reads, and mutation rate respectively), threshold 0.005 on KEC corrected reads using ViSpA [3].

<table>
<thead>
<tr>
<th>Quasispecies</th>
<th>Clones</th>
<th>42E9_A08_40</th>
<th>42E9_100_15</th>
<th>42E9_101_15</th>
<th>42E9_102_15</th>
<th>42E9_103_15</th>
<th>42E9_104_15</th>
<th>42E9_105_15</th>
<th>42E9_106_15</th>
<th>42E9_107_15</th>
<th>42E9_108_15</th>
<th>42E9_109_15</th>
</tr>
</thead>
<tbody>
<tr>
<td>v_564666</td>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>7</td>
<td>41</td>
<td></td>
</tr>
<tr>
<td>v2_121207</td>
<td></td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>8</td>
<td>40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>v6_019895</td>
<td></td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>8</td>
<td>42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>v2_04649</td>
<td></td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>9</td>
<td>41</td>
<td></td>
<td></td>
</tr>
<tr>
<td>v12_02085</td>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>7</td>
<td>42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>v_105370</td>
<td></td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>9</td>
<td>43</td>
<td></td>
</tr>
</tbody>
</table>
Table 5.3 Edit distance between collapsed sanger clones and reconstructed quasispecies using parameters 1.2.5 (the number of mismatches between sub-reads and super-reads, the number of mismatches between two overlapped reads, and mutation rate respectively), threshold 0.005 on uncorrected reads using ViSpA [3].

<table>
<thead>
<tr>
<th>Quasispecies</th>
<th>Clones</th>
</tr>
</thead>
<tbody>
<tr>
<td>v_EJ585</td>
<td>42E5_000, 40</td>
</tr>
<tr>
<td>v_161877</td>
<td>1</td>
</tr>
<tr>
<td>v14_017446</td>
<td>1</td>
</tr>
<tr>
<td>v12_023825</td>
<td>2</td>
</tr>
<tr>
<td>v_033957</td>
<td>1</td>
</tr>
<tr>
<td>v1_080035</td>
<td>2</td>
</tr>
</tbody>
</table>

As we see from Table 5.2, the first method is able to recall or reconstruct 66% of the quasispecies with 0 edit distance with the most frequent Sanger clones (40%), 12% of the quasispecies with 0 edit distance with the first 10% frequent Sanger clone, 3% of the quasispecies with 1 edit distance with the first 15% frequent Sanger clone, 4% of the quasispecies with 2 edit distance with the second 15% frequent Sanger clone, and 2% of the quasispecies with 4 edit distance with the second 10% frequent Sanger clone.

Table 5.3 shows that the second method is able to recall or reconstruct 47% of the quasispecies with 0 edit distance with the most frequent Sanger clones (40%), 16% of the quasispecies with 0 edit distance with the first 10% frequent Sanger clone, 1% of the quasispecies with 1 edit distance with one of the first 15% frequent Sanger clone, 2% of the quasispecies with 2 edit distance with the second 15% frequent Sanger clone, and 3% of the quasispecies with 4 edit distance with the second 10% frequent Sanger clone.

Table 5.4 shows distance values between Sanger clones and reconstructed quasispecies for a third method, using ShoRAH program with default parameters.

From Table 5.4, we can see that the third method is able to recall or reconstruct 12% of the quasispecies with 0 edit distance with the most frequent Sanger clones (40%), 6% of the quasispecies with 0 edit distance with the first 10% frequent Sanger clone, 1% of
the quasispecies with 1 edit distance with the first 15% frequent Sanger clone, 2 distinct quasispecies sequences with 1.2% frequency (0.00565083+0.00688508 ) with 2 edit distance with the second 15% frequency Sanger clone, 5 distinct quasispecies sequences with 4.9% frequency (0.0171438+0.00973762+0.00879847+0.0072816+0.00618452) with 4 edit distance with the second 10% frequent Sanger clone, 2 distinct quasispecies sequences with 3.9% frequency (0.0155707+0.0242772) with 5 edit distance with the first 4% frequent Sanger clone, 3 distinct quasispecies sequences with 7.4% frequency (0.015908+0.0526278+0.00578028) with the 1% frequent Sanger clone.

From Tables 5.2, 5.3, 5.4, we see that the three methods are able to reconstruct or recall 80% (0.4+0.1+0.15+0.15) of the Sanger clones with 2 edit distance or less, and the 15% (0.1+0.01+0.04) of the Sanger clones with 4 or 5 edit distance.

From Tables 5.2, 5.3, 5.4, we also, see that the first and the second methods outperform the third one in terms of reconstructed quasispecies frequencies, the third method reconstructs more false negatives than others, which means that the first two methods are better in terms of precision.
Table 5.4 Edit distance between collapsed Sanger clones and reconstructed quasispecies using ShoRAH assembler with default parameters [3].

<table>
<thead>
<tr>
<th>Quasispecies</th>
<th>Clones</th>
</tr>
</thead>
<tbody>
<tr>
<td>HAP_0.12788831</td>
<td>0</td>
</tr>
<tr>
<td>HAP_0.0629981</td>
<td>1</td>
</tr>
<tr>
<td>HAP_0.0818072</td>
<td>1</td>
</tr>
<tr>
<td>HAP_0.0055063</td>
<td>2</td>
</tr>
<tr>
<td>HAP_0.00888508</td>
<td>2</td>
</tr>
<tr>
<td>HAP_0.0875438</td>
<td>1</td>
</tr>
<tr>
<td>HAP_0.08198591</td>
<td>1</td>
</tr>
<tr>
<td>HAP_0.0005947</td>
<td>1</td>
</tr>
<tr>
<td>HAP_0.0072846</td>
<td>1</td>
</tr>
<tr>
<td>HAP_0.0008452</td>
<td>1</td>
</tr>
<tr>
<td>HAP_0.0057069</td>
<td>1</td>
</tr>
<tr>
<td>HAP_0.0042722</td>
<td>1</td>
</tr>
<tr>
<td>HAP_0.0059003</td>
<td>1</td>
</tr>
<tr>
<td>HAP_0.026262</td>
<td>1</td>
</tr>
<tr>
<td>HAP_0.0057003</td>
<td>2</td>
</tr>
<tr>
<td>HAP_0.0035362</td>
<td>4</td>
</tr>
<tr>
<td>HAP_0.0013501</td>
<td>8</td>
</tr>
<tr>
<td>HAP_0.0008998</td>
<td>2</td>
</tr>
<tr>
<td>HAP_0.0058712</td>
<td>3</td>
</tr>
<tr>
<td>HAP_0.0018907</td>
<td>9</td>
</tr>
<tr>
<td>HAP_0.0056912</td>
<td>5</td>
</tr>
<tr>
<td>HAP_0.0019597</td>
<td>6</td>
</tr>
<tr>
<td>HAP_0.001958</td>
<td>6</td>
</tr>
<tr>
<td>HAP_0.0018582</td>
<td>8</td>
</tr>
<tr>
<td>HAP_0.0037056</td>
<td>2</td>
</tr>
<tr>
<td>HAP_0.0053489</td>
<td>9</td>
</tr>
<tr>
<td>HAP_0.0036064</td>
<td>10</td>
</tr>
<tr>
<td>HAP_0.0019098</td>
<td>4</td>
</tr>
<tr>
<td>HAP_0.0008408</td>
<td>12</td>
</tr>
<tr>
<td>HAP_0.00081059</td>
<td>5</td>
</tr>
</tbody>
</table>

We reconstruct the phylogenetic tree for the first two methods, to compare how well the reconstructed quasispecies sequences are, compared to Sanger clones, see Figures 5.8 and 5.9. The reconstructed quasispecies starts with letter "v", while Sanger clones start with number"42". As we can see from the phylogenetic trees, the reconstructed quasispecies are close to Sanger clones [3].
Figure 5.8  Phylogenetic tree over collapsed Sanger clones and collapsed reconstructed quasispecies inferred from the first method with parameters 1, 2, 5 on KEC corrected reads using ViSpA [3].

Figure 5.9  Phylogenetic tree over collapsed Sanger clones and collapsed reconstructed quasispecies inferred from the second method with parameters 1, 2, 5 on uncorrected reads using ViSpA [3].
To compare between different methods (with different parameter settings), we use the two measurements, Average Distance to Clones (ADC), and Average Prediction Error (APE), as described in section 5.4.3.

In addition to the previous 3 methods. We ran many other tests using different methods, and calculated the values of Average Distance to Clones (ADC) (we disregarded one of the clones(42V1H3_G01_04) in ADC calculation, since it is an outlier), and Average Prediction Error (APE) for each method as shown in Tables 5.5 and 5.6 (ViSpA 1_2_5 means that the quasispecies are reconstructed by ViSpA using parameter values n=1, m=2, t=5) [3].

Table 5.5 Average distance to clones (ADC) for the reconstructed quasispecies using different methods [3].

<table>
<thead>
<tr>
<th>Error Correction Method</th>
<th>ShoRAH</th>
<th>KEC</th>
<th>SAET</th>
<th>uncorrected</th>
</tr>
</thead>
<tbody>
<tr>
<td>ViSpA 1_2_0</td>
<td>1.31</td>
<td>1.31</td>
<td>0.83</td>
<td>1</td>
</tr>
<tr>
<td>ViSpA 1_2_5</td>
<td>1.16</td>
<td>0.84</td>
<td>1</td>
<td>0.84</td>
</tr>
<tr>
<td>ViSpA 1_2_10</td>
<td>1.31</td>
<td>1</td>
<td>0.84</td>
<td>1</td>
</tr>
<tr>
<td>ViSpA 2_2_0</td>
<td>1.31</td>
<td>1.31</td>
<td>0.84</td>
<td>1</td>
</tr>
<tr>
<td>ViSpA 2_2_5</td>
<td>1.16</td>
<td>0.84</td>
<td>0.84</td>
<td>1</td>
</tr>
<tr>
<td>ViSpA 2_2_10</td>
<td>1.31</td>
<td>1</td>
<td>0.84</td>
<td>0.84</td>
</tr>
<tr>
<td>ShoRAH</td>
<td>0.84</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.6 Average prediction error (APE) for the reconstructed quasispecies using different methods [3].

<table>
<thead>
<tr>
<th>Error Correction Method</th>
<th>ShoRAH</th>
<th>KEC</th>
<th>SAET</th>
<th>uncorrected</th>
</tr>
</thead>
<tbody>
<tr>
<td>ViSpA 1_2_0</td>
<td>0.19</td>
<td>0.25</td>
<td>0.68</td>
<td>0.2</td>
</tr>
<tr>
<td>ViSpA 1_2_5</td>
<td>0.29</td>
<td>0.319</td>
<td>0.58</td>
<td>0.65</td>
</tr>
<tr>
<td>ViSpA 1_2_10</td>
<td>0.11</td>
<td>0.18</td>
<td>0.49</td>
<td>0.53</td>
</tr>
<tr>
<td>ViSpA 2_2_0</td>
<td>0.85</td>
<td>0.27</td>
<td>3.58</td>
<td>1.49</td>
</tr>
<tr>
<td>ViSpA 2_2_5</td>
<td>0.77</td>
<td>0.32</td>
<td>0.63</td>
<td>0.2</td>
</tr>
<tr>
<td>ViSpA 2_2_10</td>
<td>0.02</td>
<td>0.02</td>
<td>0.48</td>
<td>0.43</td>
</tr>
<tr>
<td>ShoRAH</td>
<td>3.13</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.7 summarizes tables 5.5 and 5.6, it shows both ADC and APE values for each
method.

Table 5.7  Average prediction error (APE) and Average distance to clones (ADC) of reconstructed quasispecies for different methods [3].

<table>
<thead>
<tr>
<th>Method</th>
<th>ADC</th>
<th>APE</th>
</tr>
</thead>
<tbody>
<tr>
<td>V120SHR</td>
<td>1.31</td>
<td>0.19</td>
</tr>
<tr>
<td>V125SHR</td>
<td>1.16</td>
<td>0.29</td>
</tr>
<tr>
<td>V1210SHR</td>
<td>1.31</td>
<td>0.11</td>
</tr>
<tr>
<td>V220SHR</td>
<td>1.31</td>
<td>0.85</td>
</tr>
<tr>
<td>V225SHR</td>
<td>1.16</td>
<td>0.77</td>
</tr>
<tr>
<td>V2210SHR</td>
<td>1.31</td>
<td>0.02</td>
</tr>
<tr>
<td>V120KEC</td>
<td>1.31</td>
<td>0.25</td>
</tr>
<tr>
<td>V125KEC</td>
<td>0.84</td>
<td>0.319</td>
</tr>
<tr>
<td>V1210KEC</td>
<td>1</td>
<td>0.18</td>
</tr>
<tr>
<td>V220KEC</td>
<td>1.31</td>
<td>0.27</td>
</tr>
<tr>
<td>V225KEC</td>
<td>0.84</td>
<td>0.32</td>
</tr>
<tr>
<td>V2210KEC</td>
<td>1</td>
<td>0.02</td>
</tr>
<tr>
<td>V120SAET</td>
<td>0.83</td>
<td>0.68</td>
</tr>
<tr>
<td>V125SAET</td>
<td>1</td>
<td>0.58</td>
</tr>
<tr>
<td>V1210SAET</td>
<td>0.84</td>
<td>0.49</td>
</tr>
<tr>
<td>V220SAET</td>
<td>0.84</td>
<td>3,58</td>
</tr>
<tr>
<td>V225SAET</td>
<td>0.84</td>
<td>0.63</td>
</tr>
<tr>
<td>V2210SAET</td>
<td>0.84</td>
<td>0.48</td>
</tr>
<tr>
<td>V120Unc</td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td>V125Unc</td>
<td>0.84</td>
<td>0.65</td>
</tr>
<tr>
<td>V1210Unc</td>
<td>1</td>
<td>0.53</td>
</tr>
<tr>
<td>V220Unc</td>
<td>1</td>
<td>1.49</td>
</tr>
<tr>
<td>V225Unc</td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td>LV2210Unc</td>
<td>0.84</td>
<td>0.43</td>
</tr>
<tr>
<td>ShoRAH</td>
<td>0.84</td>
<td>3.13</td>
</tr>
</tbody>
</table>

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. 5.10 (a pictorial diagram of Table 5.7), 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 and precision.
Figure 5.10 Evaluation Diagram for Average Prediction Error (APE) and Average Distance to Clones (ADC) values for different methods [3].
PART 6

CONCLUSIONS AND FUTURE WORK

NGS is emerging as a key technology for quasispecies analysis, since it allows sampling a much larger fraction of the quasispecies. The potential of NGS technologies requires the development of novel analysis methods. Analysis is challenging due to the huge amount of data generated by NGS technologies and the short read lengths and high error rates. In this dissertation we develop a novel method for correction of erroneous homopolymer indels via protein alignment.

We propose several novel methods for solving the quasispecies inference problem from amplicon reads, reducing the skewed frequency model to the ideal-frequency model, maximum bandwidth and greedy fork resolution reconstruction methods outperform on average the algorithm of Prosperi et al. for synthetic error-free reads, in terms of sensitivity, ppv, and divergence. Our experimental results in IBV quasispecies reconstruction, show how 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 divergence.

ViSpA and other published tools like ShoRAH, assemble quasispecies sequences fast, as long as number of reads are in terms of tens or hundreds of thousands. But when the number of reads become in the level of millions, these tools become slow, we propose speeding up ViSpA via parallelizing some steps, especially, the step of classification of reads to super-reads and sub-reads, and assembling of sequences from the candidate paths.

Currently, we are exploring modification of ViSpA to make it applicable for Ion Torrent reads, for example we are exploring speeding up the step of assembly from the candidate paths in a naive way, by restricting the overhang distance less than a threshold to create an edge between two overlapping reads. Additionally, we are exploring adding to ViSpA
automatic correction of erroneous homopolymer indels.

Future research includes possible combining long and short read technologies, incorporation of paired reads, study of quasispecies persistence and evolution in layer flocks following administration of modified live IBV vaccine, and optimization of vaccination strategies.
REFERENCES


