Summer 8-12-2014

Algorithms for Viral Population Analysis

Nicholas Mancuso

ALGORITHMS FOR VIRAL POPULATION ANALYSIS

by

NICHOLAS MANCUSO

Under the Direction of Dr. Alexander Zelikovsky

ABSTRACT

The genetic structure of an intra-host viral population has an effect on many clinically important phenotypic traits such as escape from vaccine induced immunity, virulence, and response to antiviral therapies. Next-generation sequencing provides read-coverage sufficient for genomic reconstruction of a heterogeneous, yet highly similar, viral population; and more specifically, for the detection of rare variants. Admittedly, while depth is less of an issue for modern sequencers, the short length of generated reads complicates viral population assembly. This task is worsened by the presence of both random and systematic sequencing

errors in huge amounts of data. In this dissertation I present completed work for reconstructing a viral population given next-generation sequencing data. Several algorithms are described for solving this problem under the error-free amplicon (or sliding-window) model. In order for these methods to handle actual real-world data, an error-correction method is proposed. A formal derivation of its likelihood model along with optimization steps for an EM algorithm are presented. Although these methods perform well, they cannot take into account paired-end sequencing data. In order to address this, a new method is detailed that works under the error-free paired-end case along with maximum a-posteriori estimation of the model parameters.

INDEX WORDS: Algorithm, Viral population reconstruction, Variant quantification, Assembly, Read overlap graph, Network flows, Integer programming, Quadratic programming, Expectation maximization, Maximum likelihood, Maximum a-posteriori

ALGORITHMS FOR VIRAL POPULATION ANALYSIS

by

NICHOLAS MANCUSO

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

2014

ALGORITHMS FOR VIRAL POPULATION ANALYSIS

by

NICHOLAS MANCUSO

Committee Chair: Alexander Zelikovsky

Committee: Yury Khudyakov

Robert Harrison

Yi Pan

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

August 2014

DEDICATION

To Ellie, for her patience and understanding; to my parents and step parents, for their support and advice; and to my brother, Zack, and sister, Frankie, for being the best siblings someone could possibly have. I love you all dearly and would be most certainly not be who I am without all of you.

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Alex Zelikovsky. Without his constant encouragement and guidance I would most certainly not have completed much of anything. It was through our (often heated) discussions that I came away with better understanding of our research. I could not imagine having selected a different advisor for my Ph.D program. I am truly indebted to him.

I would also like to thank Dr. Yury Khudyakov and Dr. Pavel Skums for their patience and excellent discussions regarding RNA viruses and viral quasi-species. A special thanks to the rest of my committee members Dr. Robert Harrison and Dr. Yi Pan. I am grateful for all the help that Dr. Raj Sunderraman has shown me over the years.

I would also like to thank Dr. King for all his advice and supervision over the years for GSU's student chapter of the ACM. I am truly thankful for him sharing his wisdom throughout my (long) stay as a GSU student.

Special thanks to all of my friends in the department who commiserated with me over coffee, drinks, pizza, or more often, research papers: Adrian Caciula, Bassam Tork, Blanche Temate, Debraj De, Guoliang Liu, Igor Mandric, Katia Nenastyeva, Lei Zhang, Marco Valero, Mingyuan Yan, Olga Glebova, Peisheng Wu, Sasha Artyomenko, Serghei Mangul, and Zhiyi Wang. I will never forget the countless days and nights meeting in breakrooms or coffee shops to forget, at least for a little while, that we have work we should be doing.

I am grateful for all my friends who allowed me to humor them with explanations of what I do over beers, drinks, food, and laughs: Alan Steadman, Danny Echavarria, Darius Soodmand, Justin Wagner, Mario Segarra, Meg Barreto, Sarah Green, and Spencer Anderson. Our countless nights at the Earl trained my phone to inform me every Saturday the time it would take to drive there—despite me not ever having planned to go.

I cannt thank my family enough for their unrelenting support throughout my life. My parents, step-parents, and siblings have always been the first to offer help, advice, guidance,

and laughter. I would not be the person I am today without them, in particular, my brother Zack. His strength and courage to leave Atlanta for the bush in Ghana for two and a half years motivated me to not give up. I will always hold dear the late nights at Matilda's drinking palm wine and Fanta during my visit (maybe not so much the long treks back). I will also always remember the crazy golf-cart rides with my sister Frankie during Inman Park festival who is now old enough to have her own crazy golf-cart rides.

Lastly I would like to thank my wife Ellie. I could not have finished without her endless support and care.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS . . . . . **v**

LIST OF TABLES . . . . . **x**

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

LIST OF ABBREVIATIONS . . . . . **xv**

PART 1 INTRODUCTION **1**

1.1 RNA Viruses and Viral Quasispecies 1

1.2 Sequencing Technologies 1

1.3 Viral Quasispecies Reconstruction Problem and Challenges . . . 2

1.4 Previous work 2

1.5 Contributions 4

1.6 Paper Roadmap 5

1.7 Publications 5

PART 2 METHODS FOR QUASISPECIES RECONSTRUCTION FROM AMPLICON READS **8**

2.1 Introduction and Contributions 8

2.2 Model 8

2.2.1 Likelihood and Entropy Minimization 8

2.2.2 Quasispecies Assembly in the Error-Free, Ideal-Frequency Model Problem 10

2.2.3 Reduction of Skewed-Frequency Model to Ideal-Frequency Model 16

2.3 Experiment Setup and Validation Metrics 17

2.3.1 Datasets and Experiment Design 17

2.3.2 Validation Metrics 18
2.4 Local Fork Resolution-based Methods for Reconstruction 19
2.4.1 Greedy Algorithm for Fork Resolution 19
2.4.2 Minimum Forest Fork Resolution 19
2.5 Results 20
2.6 Flow-based Methods for Quasispecies Reconstruction from Amplicon Reads 22
2.6.1 Maximum-Bandwidth Algorithm 22
2.6.2 Maximum Frequency Path 23
2.6.3 Multi-commodity Flow Algorithm 23
2.6.4 Results 24
2.7 Extension to Shotgun Reads 25

PART 3 CORRECTING SEQUENCING ERRORS IN QUASISPECIES RECONSTRUCTION **28**
3.1 Introduction and Contributions 28
3.2 Error Correction by kGEM 28
3.2.1 Threshold Determining 31
3.2.2 Model Selection 32
3.3 VirA: Viral Assembler 32
3.4 Datasets and Experiment Design 33
3.5 Results 33

PART 4 VIRAL QUASISPECIES RECONSTRUCTION FROM PAIRED-END READS **35**
4.1 Introduction and Contributions 35
4.2 Methods 36
4.2.1 Overview 36
4.2.2 Error correction 39

4.2.3 Consensus construction 39
4.2.4 Read mapping 39
4.2.5 Viral population assembly 40
4.2.6 Viral population quantification 41
4.3 Results 43
4.3.1 Performance of VGA on simulated data 43
4.3.2 Performance of existing viral assemblers on simulated consensus error-corrected reads 47
4.3.3 Performance of VGA on real HIV data 48
4.4 Discussion 49

PART 5 DISCUSSION AND FUTURE WORK 54

REFERENCES 55

LIST OF TABLES

Table 1.1 Next-generation sequencers and properties of the produced reads[1]. 2

Table 1.2 Quasispecies reconstruction/inference tools and supported features. All main tools currently support both local and global reconstruction. All tools with the exception of QuasiRecomb utilize some form of a read-graph. 4

LIST OF FIGURES

Figure 2.1 The case of two distinct reads for both amplicons 13

Figure 2.2 Adding forks to the original $s - t$ connected read graph. 17

Figure 2.3 Sensitivity results on simulated HCV population over error-free data. The results are partitioned over each population distrubtion. . . . 21

Figure 2.4 Positive predictive value results on simulated HCV population over error-free data. The results are partitioned over each population distrubtion. 21

Figure 2.5 Jensen-Shannon divergence results on simulated HCV population over error-free data. The results are partitioned over each population distrubtion. 22

Figure 2.6 Sensitivity results on simulated HCV population over error-free data for flow-based algoirthms. The results are partitioned over each population distrubtion. 25

Figure 2.7 Positive predictive value results on simulated HCV population over error-free data for flow-based algoirthms. The results are partitioned over each population distrubtion. 26

Figure 2.8 Jensen-Shannon divergence results on simulated HCV population over error-free data for flow-based algoirthms. The results are partitioned over each population distrubtion. 26

Figure 3.1 Results obtained from simulated amplicon reads HCV data using sensitivity (weighted portion of true variants found). We relax the case of requiring an exact match and allow for hamming distance. . . . 33

Figure 3.2 Results obtained from simulated shotgun reads HCV data using sensitivity (weighted portion of true variants found). We relax the case of requiring an exact match and allow for hamming distance. 34

Figure 4.1 Overview of high-fidelity sequencing protocol. (a) DNA material from a viral population is cleaved into sequence fragments using any suitable restriction enzyme. (b) Individual barcode sequences are attached to the fragments. Each tagged fragment is amplified by the polymerase chain reaction (PCR). (c) Amplified fragments are then sequenced. (d) Reads are grouped according to the fragment of origin based on their individual barcode sequence. An error-correction protocol is applied for every read group, correcting the sequencing errors inside the group and producing corrected consensus reads. (e) Error-corrected reads are mapped to the population consensus. (f) SNVs are detected and assembled into individual viral genomes. The ordinary protocol lacks steps (B) and (D). 38

Figure 4.2 Overview of VGA. (a) The algorithm takes as input paired-end reads that have been mapped to the population consensus. (b) The first step in the assembly is to determine pairs of conflicting reads that share different SNVs in the overlapping region. Pairs of conflicting reads are connected in the "conflict graph". Each read has a node in the graph, and an edge is placed between each pair of conflicting reads. (c) The graph is colored into a minimal set of colors to distinguish between genome variants in the population. Colors of the graph correspond to independent sets of non-conflicting reads that are assembled into genome variants. In this example, the conflict graph can be minimally colored with four colors (red, green, violet and turquoise), each representing individual viral genomes. (d) Reads of the same color are then assembled into individual viral genomes. Only fully-covered viral genomes are reported. (e) Reads are assigned to assembled viral genomes. Read may be shared across two or more viral genomes. VGA infers relative abundances of viral genomes using the expectation-maximization algorithm. (f) Long conserved regions are detected and phased based on expression profiles. In this example turquoise and red viral genome share a long conserved region(colored in black). There is no direct evidence how the viral sub-genomes across the conserved region should be connected. In this example 4 possible phasing are valid. VGA use the expression information of every sub-genome to resolve ambiguous phasing. 51

Figure 4.3 Genomic architecture of 44 real HCV viral genomes from 1739-bp long fragment of E1E2 region. Length of longest common region shared between any two viral genomes is represented by color. 52

Figure 4.4 Accuracy of population size prediction. Up to 200 viral genomes were generated from the Gag/Pol 3.4 Kb HIV region. The population diversity is 5% - 10%. Variant abundances follow uniform (A) and power-law (B). Highly-accurate 100x2 bp paired-end reads were simulated from HIV population. 52

Figure 4.5 Assembly accuracy estimation. Up to 200 viral genomes were generated from the Gag/Pol 3.4 Kb HIV region. The population diversity is 3% - 20%. Variant abundances follow uniform (A) and power-law (B). Consensus error-corrected 2x100bp paired-end reads were simulated from HIV population. 52

Figure 4.6 Assembly accuracy estimation. Consensus error-corrected paired-end reads of various lengths were simulated from a mixture of 10 real viral clones from 1.3kb-long HIV-1 region. Assembly accuracy as measured by sensitivity(A) and precision(B). Results are for 50,000 reads, no improvement was observed when increasing the of number of reads. 53

Figure 4.7 Assembly accuracy estimation. Up to 200 recombinant viral genomes were generated from the from 1.3kb-long HIV-1 region. Variant abundances follow power-law and uniform. Consensus error-corrected 2x100bp paired-end reads were simulated from HIV population. . 53

LIST OF ABBREVIATIONS

- NGS - Next Generation Sequencing
- MLE - Maximum Likelihood Estimate
- MAP - Maximum A-Posteriori
- EM - Expectation Maximization

PART 1

INTRODUCTION

1.1 RNA Viruses and Viral Quasispecies

RNA viruses, as the name implies, encode their genome in RNA rather than DNA. Notable examples include human immunodeficiency virus (HIV), hepatitis C (HCV), and influenza. RNA viruses display exceptionally high mutation rates in comparison to DNA-based counterparts. Indeed, mutation rates varying between 10^{-4} and 10^{-6} per nucleotide have been observed. In both experimental and natural infections, a viral particle, or virion, upon infecting a cell may produce hundreds to thousands of progeny; thus, generating many mutant strains into the population. In addition to mutations, RNA viruses have been known to exhibit recombinant variants within the population. Cells can become co-infected by different viral strains and consequently "cross over" genomes during replication. This process may be repeated with the newly produced recombinant variants, further driving the heterogeneity of the population. This population of closely related strains is known as a quasispecies.

The replicative dynamics ensure the virus can efficiently adapt to environmental changes within an infected host. This mutational robustness is the root cause of difficulty for therapeutic treatments. Therefore, accurately determining the viral population structure (i.e., individual genomes) is of great utility for both treatment and understanding of viral quasispecies.

1.2 Sequencing Technologies

As a result of the rapid decrease in sequencing cost, it is now possible to directly inspect a viral population. This massive amount of sequence data is generated using two different processes. The first process is shotgun sequencing, whereby long genomic regions are randomly sheared by sonication and subsequently sequenced. The second is multiple

Table 1.1 Next-generation sequencers and properties of the produced reads[1].

Manufacturer	Avg Read Length	Avg Read Count	Avg Error Rate	Paired/Mate Reads
Roche/454	450bp	1M	1.0%	Yes (with kit)
Ion Torrent	200bp	60M	2.8%	Yes
Illumina	150bp	10	1.0%	Yes
Pacific Bio	8500bp	45k	12.0%	No

amplicon sequencing, which is based on PCR amplification of a set of overlapping genomic regions using sequence-specific primers. Multiple regions may be sequenced in a single run by coupling the sequence-specific primers with "tags" or unique identification sequences. Table 1.1 gives a breakdown of various companies' products and their respective specifications.

1.3 Viral Quasispecies Reconstruction Problem and Challenges

Due to the limitations of current sequencing technologies, entire viral genomes describing a viral population cannot be accurately generated. Next-generation sequencers are capable of produces massive amounts of genomic information, but are limited to short reads; therefore, these genome snippits must be assembled.

Quasispecies Spectrum Reconstruction (QSR) Problem. *Given a collection of (shotgun or amplicon) next-generation sequencing reads generated from a viral sample, reconstruct the quasispecies spectrum, i.e., the set of sequences and the relative frequency of each sequence in the sample population.*

This task is challenging for the following reasons: (i) differentiating rare variants from random sequencing errors and correcting systematic errors; (ii) deciding if multiple reads with a concordant overlap belong to the same variant; (iii) and designing scalable software to handle ever-increasing volumes of read-data.

1.4 Previous work

The first publicly available tool for reconstructing a viral quasispecies was ShoRAH [2]. It combined the clustering work initialy described in [3] along with a path cover described in

[4]. Once the putative variants have been assembled, their relative abundances are computed using an Expectation-Maximization (EM) algorithm. While ShoRAH manages to successfully capture the underlying population, it tends to vastly overestimate the true number of variants, thus skewing accuracy. ShoRAH is available as a python program that can be run for various forms a sequencing data: NGS shotgun data or single amplicon data.

The Viral Spectrum Assembler, or ViSpA[5], is another tool that reconstructs a quasispecies by constructing a weighted read-overlap graph. The edge-weights in the graph represent the probability of the overlap occurring between two sequencing reads. In order to reduce the number of edges in the graph, a transitive closure is computed. This technique removes "sub-reads" that add no further information to the population structure. ViSpA repeatedly finds maximum-weight (i.e., high-probability) paths to cover the reads until the graph is saturated. Similarly to ShoRAH, ViSpA utilizes an EM algorithm for estimating variant frequencies. ViSpA is available as a java-based tool. While it was designed for NGS shotgun data, in theory it could be applied to amplicon-based data as well.

QuRe [6] is a tool to reconstruct a viral quasispecies based on earlier work described in [7]. Initially, the software aligns reads to a reference sequence, which are then corrected via a Poisson process as described in [8]. Afterwards, reads are partitioned into sliding windows using a randomized approach. Putative partitions are scored based on coverage, overlapping sequence divergence, and internal window sequence divergence. A read-overlap graph is then constructed, and paths are found by a hueristic. Once enough paths have been found, the final results are then clustered based on error-rate parameters. QuRe is implemented in java and runs on a variety of platforms.

QuasiRecomb is a software that employs a generative model to estimate the underlying viral population[9]. It explicitly incorporates recombination into the model by utilizing "generator" sequences. Each generator sequence is modeled by a hidden Markov model. Recombination hotspots may occurr by "jumping" from one model to another at any given point in the sequence space. The parameters to the model are estimated using an EM (Baum-Welch) algorithm. Once the parameters have been found, the posterior is then sampled to

Table 1.2 Quasispecies reconstruction/inference tools and supported features. All main tools currently support both local and global reconstruction. All tools with the exception of QuasiRecomb utilize some form of a read-graph.

Tool	Reconstruction	Paired-End Support	Model
ShoRAH	Both	No	Read-Graph
ViSpA	Global	No	Read-Graph
QuRe	Both	No	Read-Graph
QuasiRecomb	Both	Yes	HMM
VirA	Both	No	Read-Graph
VGA	Global	Required	Conflict Read-Graph

estimate the population and respective abundances. QuasiRecomb is implemented in java and targets single-amplicon data.

1.5 Contributions

We present novel "fork-resolution" algorithms for viral quasispecies reconstruction as well as "flow"-based algorithms. Fork-resolution algorithms focus on solving local assembly problems within the read-graph model. They are typically quite fast in practice, but are limited in their accuracy. This inherent problem is a result of focusing only on small, local assemblies. Flow-based algorithms take a more global approach to assembling the viral population. After the read graph has been constructed, paths are found via network flows. We have previously published work utilizing single-path network flows in addition to multi-commodity flows. These algorithms and respective data-structures are implemented in a Python framework called **BIOA**. The software is open source and is available for free.

Additionally, parameter estimation and model selection were contributed to the tool kGEM. This software performs local reconstruction (short targeted region) of a viral population and can be used for error correction. In order for kGEM to perform certain clustering, a feasible error threshold must be determined. This is done using a simple p-value analysis under Bonferonni adjustment. Finally, model selection is evaluated under both Akaike and Bayesian information criteria.

It is in conjunction with kGEM and the aforementioned algorithms that constitute

our software **V**iral **A**ssembler, or VirA. VirA first aligns the data using the InDelFixer alignment tool and the random algorithm of QuRe to find good "virtual" amplicons over NGS data. Once the partitions have been found, kGEM then locally reconstructs haplotypes by correcting errors. Finally a read-graph is built over the local haplotypes and one of the flow-based algorithms is run. VirA is implemented in Python, Java, and Scala. It runs on multiple platforms and scales well as the number of reads grows.

1.6 Paper Roadmap

The paper is organized as follows. Chapter 2 discusses the model, methods, experiment setup, and results for viral population reconstruction from error-free amplicon reads. Fork-resolution methods are first described followed by the flow-based counterparts. Chapter 3 describes how to handle sequencing errors by the kGEM method. Our software VirA is then described along with experiment setup and results. Chapter 4 presents the work completed on paired-end data. Finally the paper outlines ongoing and future work in section 5.

1.7 Publications

Book Chapters

1. I. Astrovskaya, **N. Mancuso**, B. Tork, S. Mangul, A. Artyomenko, P. Skums, L. Ganova-Raeva, I. Măndoiu, and A. Zelikovsky "Inferring Viral Quasispecies Spectra from Shotgun and Amplicon 454 Pyrosequencing Reads" Genome Analysis: Current Procedures and Applications, 2013.

Journal Papers

1. Alexander Artyomenko, **Nicholas Mancuso**, Pavel Skums, Ion Mandoiu, Alex Zelikovsky, and Yury Khudyakov. "An EM-based Algorithm for Reconstructing a Viral Population from Single-Amplicon NGS Data". BMC Genomics (Submitted).

2. Pavel Skums, **Nicholas Mancuso**, Alexander Artyomenko, Bassam Tork, Ion Măndoiu, Yury Khudyakov, and Alex Zelikovsky. "Reconstruction of Viral Population Structure from Next-Generation Sequencing Data Using Multicommodity Flows", BMC Bioinformatics 2013, 14(Suppl 9):S2

3. **Nicholas Mancuso**, Bassam Tork, Pavel Skums, Lilia Ganova-Raeva, Ion Măndoiu, and Alex Zelikovsky, "Reconstructing Viral Quasispecies from NGS Amplicon Reads" In Silico Biology, An International Journal on Computational Molecular Biology, Volume 11, 5. pp 237-249. 2012.

Conference / Workshop Papers

1. Serghei Mangul, Nicholas Wu, **Nicholas Mancuso**, Alex Zelikovsky, Ren Sun, and Eleazar Eskin, "Accurate HIV population assembly from ultra-deep sequencing data". ISMB 2014.

2. Alexander Artyomenko, **Nicholas Mancuso**, Pavel Skums, Ion Măndoiu, Alex Zelikovsky. "kGEM: An Expectation Maximization Error Correction Algorithm for Next Generation Sequencing of Amplicon-based Data", 9th International Symposium on Bioinformatics Research and Applications. (Short Abstract), 2013.

3. **Nicholas Mancuso**, Bassam Tork, Pavel Skums, Ion Măndoiu and Alex Zelikovsky "Multi-Commodity Flow Methods for Quasispecies Spectrum Reconstruction Given Amplicon Reads", 8th International Symposium on Bioinformatics Research and Applications (Short Abstract), 2012.

4. **N. Mancuso**, B. Tork, P. Skums, L. Ganova-Raeva, I.I. Măndoiu, A. Zelikovsky "Workshop: A Maximum Likelihood Method for Quasispecies Spectrum Assembly" Proc. 2nd Workshop on Computational Advances for Next Generation Sequencing (CANGS 2012)

5. Marco Valero, Mingsen Xu, **Nicholas A Mancuso**, Wen-Zhan Song, and Raheem

Beyah. "EDR2: A Sink Failure Resilient Approach for WSNs." IEEE International Conference on Communications (ICC), pp. 616-621, 2012.

6. **N. Mancuso**, B. Tork, I.I. Măndoiu and A. Zelikovsky and P. Skums "Viral Quasispecies Reconstruction from Amplicon 454 Pyrosequencing Reads", Proc. 1st Workshop on Computational Advances in Molecular Epidemiology, pp. 94-101, 2011

Posters

1. Serghei Mangul, Nicholas Wu, **Nicholas Mancuso**, Alex Zelikovsky, Ren Sun, Eleazar Eskin, "Inferring HIV Quasispecies from Paired-End Reads". RECOMB, April 2013.

2. **Nicholas Mancuso**, Bassam Tork, Pavel Skums, Ion Măndoiu and Alex Zelikovsky, "Poster: Quasispecies Spectrum Reconstruction using Multi-commodity Flows" RECOMB-Seq, April 2012.

3. S. Mangul, A. Caciula, **N. Mancuso**, I. Măndoiu and A. Zelikovsky, "An Integer Programming Approach to Novel Transcript Reconstruction from Paired-End RNA-Seq Reads", Poster at 16th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2012), Barcelona, Spain

PART 2

METHODS FOR QUASISPECIES RECONSTRUCTION FROM AMPLICON READS

2.1 Introduction and Contributions

In this section I present the amplicon (overlapping window) model for quasispecies reconstruction. This model is selected for its simplicity and ease of analysis. Surprisingly this model is still NP-hard to compute optimal solutions for (which we prove). Along with the analysis, algorithms for reconstruction are proposed. These are divided along local "fork-resolution" methods and slightly more global "flow"-based methods. All algorithms are compared and validated against simulated hepatitis C (HCV) data. In addition to reconstruction algorithms, a "graph-balancing" algorithm is presented. This method makes a minimal amount of changes to the graph in order for all read-counts to be balanced. This method vastly improves the local fork resolution methods, but seems to have little affect on flow methods. However, this could be investigated further under more realistic read generation scenarios.

2.2 Model

2.2.1 Likelihood and Entropy Minimization

The amplicon-based quasispecies assembly covers the full virus genome with the set of K overlapping segments with predefined positions within the genome, called amplicons. Each amplicon $A_1, \ldots, A_K$ has a predefined length and is sequenced to the same depth D, i.e., covered with D reads. We distinguish two error models.

- The error-free model assumes that all reads are typing error-free or, equivalently, have been previously cleaned from typing errors.

- The error-prone model allows some typing errors and additionally these errors should be detected and fixed.

We also distinguish two frequency models.

- The ideal-frequency model assumes that in each amplicon's distribution of reads is identical and equal to the true distribution of quasispecies.
- The more realistic skewed-frequency model assumes that in each amplicon the quasispecies are represented differently from the true distribution.

In the next subsection we address the QSR problem in the ideal-frequency model and then show we can adjust frequencies to reduce the skewed-frequency model to the ideal-frequency model.

The main goal is to reconstruct the genome-length quasispecies from amplicon data consisting of $K \times D$ reads. The secondary goal is to optimize the amplicon-based assembly parameters K, D and amplicon positions in order to maximize the quality (sensitivity and specificity) of assembly. We also compare the amplicon-based and the shotgun sequencing approaches to quasispecies assembly. It is important to note that shotgun sequencing is more prone to typing errors but less prone to frequency skewing than amplicon based sequencing. Moreover, the methods for reconstruction of quasispecies from shotgun reads (e.g., ViSpA) heavily rely on the uniform distribution. This allows a more accurate estimate of the probability of two overlapping reads coming from the same quasispecies. For reconstruction from amplicon reads, this information is not available and therefore it is necessary to rely on parsimonious considerations. Further, we discuss several related optimization formulations. The following formulation is standard, but difficult to solve.

Most Parsimonious Spectrum. The most parsimonious spectrum requires the minimum number of distinct quasispecies that explains the observed reads. This formulation is NP-hard by simple reduction from SUBSET SUM, which is defined as: given a (multi) set S of integers and count c, does there exist a subset $S' \subseteq S$ whose sum is c[10]?

Claim 1. *QSR Problem is* NP*-hard.*

Proof. Given an instance of SUBSET SUM, $S = \{x_1, \ldots, x_k\}$ and count c where $\sum_{i=1}^{k} x_i = n$, let $A_1 = \{r_1, \ldots, r_k\}$ be reads with counts $\{x_1, \ldots, x_k\}$ and $A_2 = \{r_{k+1}, r_{k+2}\}$ be reads with counts $\{c, n-c\}$ such that the overlap is consistent between any two reads $r_i \in A_1, r_j \in A_2$. A minimal solution to the QSR problem will cover exactly k reads in A_1. Taking the $q \leq k$ reads and respective counts from A_1 that correspond to quasispecies containing read r_{k+1} we find the solution to SUBSET SUM. If any read $r_i, 0 \leq i \leq k$ is split between r_{k+1}, r_{k+2} there must exist at least $k+1$ quasispecies (if all reads are covered) and will not be a minimal solution. □

However, a small number of different reads is expected within the same amplicon (small number of quasispecies), allowing us to solve it exactly in practical time. The sets of overlapping (adjacent) 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 reads in the left and the right amplicons. Further, we discuss several models and corresponding optimization formulations.

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

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

- each vertex v in V_i corresponds to a distinct read in the i-th amplicon A_i and has a count $c(v)$ (i.e., unique read v is repeated in the read sample $c(v)$ times) $\sum_{v \in V_i} c(v) = D$.
- each edge (u, v) connects two reads from consecutive amplicons A_i and A_{i+1} which agree in the overlap region

The solution can be viewed as the set $Q = \{q_j\}$ of $u - v$-paths, $u \in V_1$, $v \in V_K$, each with the frequency f_j such that for each vertex $v \in V$,

$$\sum_{v \in q_j} f_j = c(v) \tag{2.1}$$

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

Rather than solving the K-staged assembly problem, we first focus on the two-staged case, the solution to which is then used to "stitch" together all K stages. We therefore, assume that there are only two stages, V_1 and V_2, thus implying that the read graph is bipartite. A natural question is whether a feasible solution exists for this problem.

The problem can be reduced to consistency of linear equations. 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 (2.1) to obtain the following system of linear equations:

$$\forall v \in V_1 \cup V_2 \sum_{e \text{ incident to } v} f_e = c(v) \tag{2.2}$$

The system of equations (2.2) is consistent if and only if the corresponding two-stage assembly problem is feasible. Denote by $G(c)$ the bipartite graph obtained from G by splitting each vertex v into $c(v)$ copies. It is clear that the following claim is true.

Claim 2. *The system (2.2) is consistent if and only if $G(c)$ has a perfect matching.*

The feasibility of the two-stage assembly problem could be checked using any classical perfect matching algorithm. We assume that (2.2) is consistent.

Theorem 1. *Any solution of (2.2) with the minimal number of edges is a spanning forest of G.*

Proof. We assume that G is connected. (2.2) could be written as $If = c$ or $I_1 f_1 + \cdots + I_m f_m = c$, where I is the incidence matrix of G, I_j is the jth column of I.

Since G is bipartite, $\text{rank}(I) = n - 1$ [11]. Let the first $n - 1$ columns and rows of I be the basis columns and rows. The columns of the incidence matrix corresponding to the even cycle are linearly dependent. Since G does not contain odd cycles, the edges $e_1, \ldots, e_{n-1}$ contain no cycles, and therefore form the spanning tree of G. Set free variables $f_n, \ldots, f_m$ equal to 0. Then (2.2) is equivalent to

$$I'f' = c', \tag{2.3}$$

where I' is the sub-matrix of I restricted to the first $n-1$ columns and rows, $f' = (f_1, \ldots, f_{n-1})$, $c' = (c_1, \ldots, c_{n-1})$.

As (2.2) is consistent, (2.3) is also consistent. Moreover, since G is bipartite, I is totally unimodular [11]. This implies I' is also totally unimodular with $\text{rank}(I') = n-1$. Therefore (2.3) has a positive integer solution. The non-zero edges of this solution form the spanning forest of G.

This proves that (2.2) has a spanning forest solution. Let f be a solution of (2.2) containing cycles. Consider a connected component of f, which has cycles. Find the acyclic solution of (2.2) restricted to that component. This solution has a lesser number of edges. Repeating this process for all connected components with cycles, we obtain the acyclic solution for (2.2). □

Theorem 2. *If there is a bipartite cycle between consecutive amplicons, then there is a 4-gamete rule violation, i.e., there should exist an additional (to existing in amplicons) recombination.*

Proof. Let B be bi-cycle (a bipartite cycle) with $2k$ vertices (with k vertices in the left amplicon L and with k vertices in the right amplicon R). For the sake of simplicity, assume that the alphabet has two characters. Then k different haplotypes in L can be distinguished in $k-1$ positions (the same for R). There are exactly $2k$ different edges in B. Assuming no 4-gamete rule violation, we obtain $2k$ haplotypes distinguished in $2k-2$ positions. This implies a contradiction. □

A simple maximum likelihood approach assumes that any edge (i.e., quasispecies) is equally probable. As a result, it assigns non-zero frequencies to all possible quasispecies. A more meaningful parsimonious approach requires that most likely solutions should contain the minimum number of quasispecies satisfying the linear system (2.2).

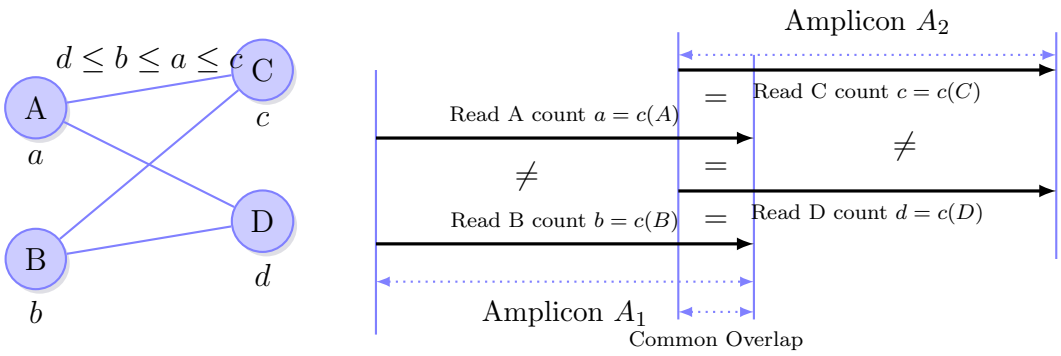

Figure 2.1 The case of two distinct reads for both amplicons

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 Fig. 2.1). Furthermore, let all four possible combinations of reads be consistent in their overlap.

On the left there are four reads and on the right there is the corresponding graph. Without loss of generality, 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 three and there should be three edges with non-zero frequency. There are two possibilities for three non-zero frequency edges:

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

The first case is more probable if $a > b$ and is equally probable if $a = b$. A simple maximum likelihood does not appear to be an efficient approach given these assumptions. We now formalize a more sophisticated model to compute maximum likelihood. We define a "fork" $F = (L, R)$ to be a bi-clique with partitions L and R, where $|L| = l, |R| = r$. Given a model, we would like to be able to compute the probability of observing the reads in L and R. As reads in both partitions overlap, no single instance of a quasispecies is capable of producing both its read in L and its read in R. Based on this reasoning, we assume that the total number of quasispecies that is capable of producing all observed reads is $c(L) + c(R)$.

For simplicity we may assume that each quasispecies containing overlap of F produces exactly one read. In other words, if a quasispecies does not "emit" a read then we consider it nonexistent. Any possible resolution of X_F of the fork F consists of $c(L) + c(R)$ instances formed by lr distinct quasispecies $q_{11}, \ldots, q_{lr}$ with counts $x_{11}, \ldots, x_{lr}$. We consider a resolution X_F to be feasible if $X_F = \{x_{ij}\}$ where $x_{ij} = x_{ij}^L + x_{ij}^R$ such that

$$\begin{aligned} \sum_{j}^{r} x_{ij}^L &= c_i, && \forall i \in L \\ \sum_{i}^{l} x_{ij}^R &= d_i, && \forall j \in R. \end{aligned} \tag{2.4}$$

Following the model (eq. 2.4) we estimate the probability $P(X_F)$ that X_F produces the observed reads L and R as follows. We can see there are in total $T_F = 2^{c(L)+c(R)}$ ways to produce reads from $c(L) + c(R)$ quasispecies restricted to two amplicons. We need to count the number of ways to produce the observed read counts, $C(X_F)$, in order to find $P(X_F) = \frac{C(X_F)}{T_F}$.

We make one more assumption that the resolution X_F is a forest (see theorem 1). For a given leaf i and count c_i we can see that there exists a single quasispecies q_{ij} with count $x_{ij} \geq c_i$ that is capable of emitting all copies of the i-th read. The number of possible assignments of quasispecies to produce this read is $\binom{x_{ij}}{c_i}$. Let F' be the fork obtained from F by removing quasispecies q_{ij} and decreasing the count of the *stem* j-th read by $x_{ij} - c_i$. We now have,

$$\begin{aligned} C(X_F) &= C(X_{F'}) \cdot \binom{x_{ij}}{x_{ij}^L} \\ &= C(X_{F'}) \cdot \frac{x_{ij}!}{x_{ij}^L! \cdot x_{ij}^R!} \end{aligned} \tag{2.5}$$

The formula (2.5) defines the recursive computation of the likelihood given a forest resolution X_F for the fork $F = (L, R)$. The computation can be done in leaf-removal order i.e., $v_1, \ldots, v_{l+r-k}$, where $k = k(X_F)$ is the number of connected components of X_F. Let the remaining count of a leaf v_i be c_i, the count of the covering quasispecies q_{ij} be x_{ij} and the count of the j-th stem's read be c_j. The definition is as follows,

$$L(X_F|F = (L, R)) = P(X_F) \tag{2.6}$$

$$= \frac{C(X_F)}{T_F} \tag{2.7}$$

$$= \frac{\prod_{i=1}^{l+r-k} \binom{x_i}{c_i}}{2^{c(L)+c(R)}}. \tag{2.8}$$

Therefore, maximizing the log-likelihood is equivalent to maximizing,

$$\log C(X_F) = \sum_{i=1}^{l+r-k} \log \binom{x_i}{c_i}.$$

Following [12], we notice that maximizing likelihood is equivalent to minimizing Shannon Entropy, which is defined as follows,

$$-\sum_{i=1}^{n} p(x_i) \log p(x_i).$$

The more quasispecies we have, the higher the entropy value. Therefore, finding a solution with the minimum amount of quasispecies reduces the entropy of the solution, which corresponds to the idea of a most parsimonious solution.

Minimum Entropy Quasispecies Assembly Problem. *Given a read graph with frequencies on reads, find the set of quasispecies (paths) and respective frequencies that explains all reads and frequencies, such that the resulting set has minimum entropy.*

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

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

We would like to reduce this problem 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. We then suggest two approaches.

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.

Fork Balancing Problem. *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 2.9). 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.

$$\textbf{Minimize} : \sum w_i x_i^2 \tag{2.9}$$

$$\textbf{Subject to} : \sum_{i \in L_f} (x_i + o_i) = \sum_{i \in R_f} (x_i + o_i), \qquad \forall f \in \text{ Forks} \tag{2.10}$$

$$x_i + o_i \geq 0 \tag{2.11}$$

Graph Data Structure. Building a K-staged read graph from K amplicons is straightforward. For each distinct read in amplicon A_i, search for consistent distinct reads in amplicon

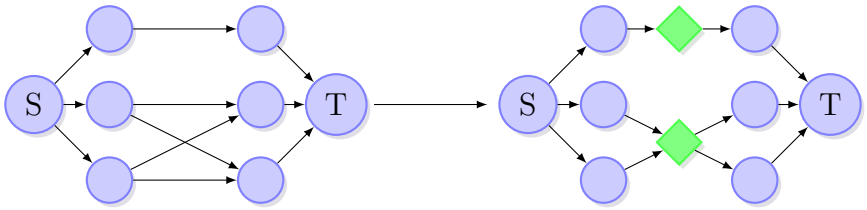

Figure 2.2 Adding forks to the original $s - t$ connected read graph.

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.

Before the final structural modification is made in the graph G, we may wish to adjust the frequencies. This may be done by either solving the QP (eq. 2.9) 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 (figure 2.2).

2.3 Experiment Setup and Validation Metrics

2.3.1 Datasets and Experiment Design

In order to perform cross-validation on the assembly method, we simulate reads data from a 1739bp long fragment from the E1E2 region of 44 HCV sequences [13] when 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. Reads are simulated without sequencing errors. The length of a read follows a normal distribution with mean value of 320bp and variance 10bp. The starting position follows a uniform distribution. The total amount of reads varies from 5K, 20K, to 100K.

2.3.2 Validation Metrics

Sensitivity is defined as $\frac{|TruePositives|}{|TruePositives|+|FalseNegatives|}$ and is a measure of a test's positive identifications. We utilize this metric as a measure of the quality assembled quasispecies. If a candidate quasispecies has an exact match (in terms of its sequence) in the simulated quasispecies, we consider it a true positive match. Thus sensitivity is a function of the number of correctly assembled quasispecies out of the population. Positive predicted value (PPV) is the ratio of correctly identified quasispecies over assembled quasispecies. It is defined as $\frac{|TruePositives|}{|TruePositives|+|FalsePositives|}$. It is useful for cross-validating a given method, by highlighting the amount of incorrectly assembled quasispecies.

Jensen-Shannon divergence is a quasi-metric that measures the divergence (i.e., distance) between two statistical distributions. It differs from the Kullback-Leibler divergence due to the addition of a midpoint. This allows for the quasi-metric to be continuous over all values of P and Q. As the size of our reconstructed spectrum may differ from the size of the "true" spectrum, we can still gain insight into how close the two are. It is defined as:

$$\mathrm{JSD}(P||Q) = \frac{1}{2}\mathrm{D}_{KL}(P||M) + \frac{1}{2}\mathrm{D}_{KL}(Q||M)$$

where

$$\mathrm{D}_{KL}(P||Q) = \sum_{i=1}^{n} P(i) \log \frac{P(i)}{Q(i)}$$

and $M = \frac{1}{2}(P + Q)$. By taking as assembled quasispecies spectrum and computing the divergence from the true (simulated) spectrum, we can measure how close our predicted

frequency distribution is to the true spectrum.

2.4 Local Fork Resolution-based Methods for Reconstruction

2.4.1 Greedy Algorithm for Fork Resolution

Each fork vertex v (a vertex with at least two in-edges and two out-edges) should be resolved, i.e., a resolution (acyclic graph connecting inputs with outputs) should replace v. We greedily resolve forks in the graph by placing left and right vertices into priority queues L and R. Extract the maximum from L and from R and create a new edge (max_l, max_r). The frequency of the new edge is $\min(max_l, max_r)$. Decrease the key (frequency) of both vertices and reinsert the larger maximum frequency; otherwise, both vertices are exhausted and no longer considered for further resolution. When either L or R are empty, the round is completed, and the fork vertex is removed. This is done recursively until no forks are left in the graph. It may be the case that read vertices (not just overlap vertices) become forks after a round and are eligible for resolution.

2.4.2 Minimum Forest Fork Resolution

Rather than rely on a greedy heuristic, we also employed an integer linear program (ILP) that minimizes absolute value difference of observed and assigned read counts, such that all reads belong to a tree. The resulting solution will be a forest over a bi-clique. Given a bipartite graph $G = (L, R, E)$ assume without loss of generality that $|L| \leq |R|$. The ILP

is given as,

$$\begin{aligned}
\textbf{Minimize: } & \sum_{i \in L} y_i & \\
\textbf{Subject to: } & \sum_{i}^{L} x_{ij} = 1 & \forall j \in R \\
& \sum_{j}^{R} r_j \cdot x_{ij} - l_i \leq y_i & \forall i \in L \\
& l_i - \sum_{j}^{R} r_j \cdot x_{ij} \leq y_i & \forall i \in L.
\end{aligned}$$

The intuition is that leaves in the forest assign "all" their reads to a single parent. Internal nodes "balance" out the counts of its children counts. The optimal solution will be the forest that is most balanced in terms of assigned reads.

2.5 Results

Experiment results clearly indicate the ability of the greedy fork-resolution algorithm to reconstruction the population accurately, as shown in figures (2.3-2.5). This is particularly, noticeably under the skewed and geometric distributed populations. Interestingly, results for the minimum forest solution were much worse than the greedy, despite being an exact solution to the ILP. This is most likely a result of the ILP minimizing total deviation, and not directly minimizing entropy. However, both algorithms outperform the guide distribution method (GD).

When validating the quality of assigned abundances, the greedy method performs exceedingly well given both skewed and geometric distributed populations. In fact, all methods tend to perform much better on average under non-uniform distributions. Unlike the uniformly distributed population, as the abundances of variants under the skewed and geometric are quite divergent, it is easier to assign the neighboring reads to the correct variant.

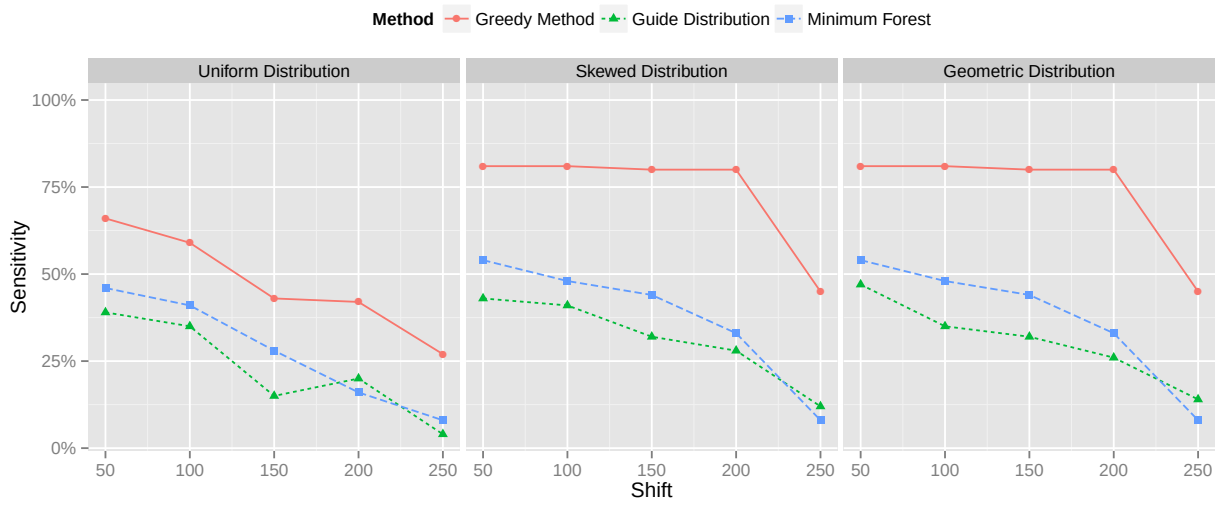

Figure 2.3 Sensitivity results on simulated HCV population over error-free data. The results are partitioned over each population distrubtion.

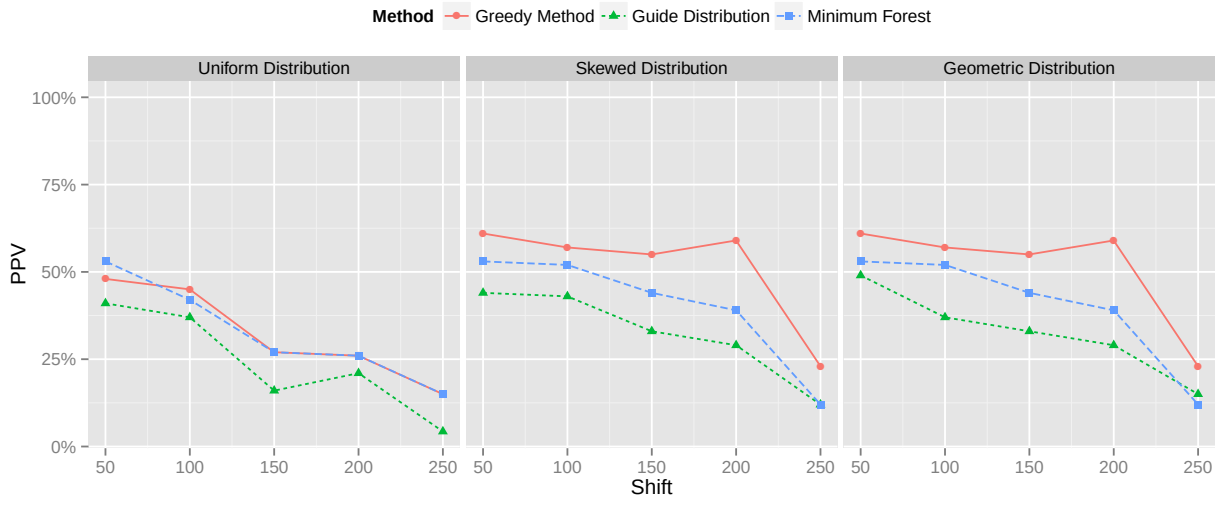

Figure 2.4 Positive predictive value results on simulated HCV population over error-free data. The results are partitioned over each population distrubtion.

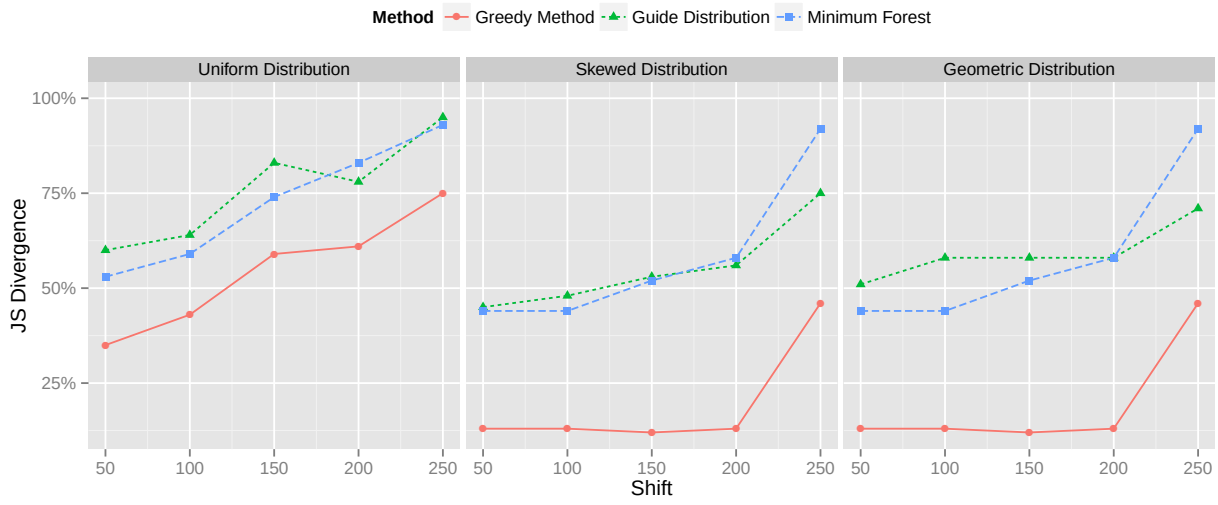

Figure 2.5 Jensen-Shannon divergence results on simulated HCV population over error-free data. The results are partitioned over each population distrubtion.

2.6 Flow-based Methods for Quasispecies Reconstruction from Amplicon Reads

2.6.1 Maximum-Bandwidth Algorithm

A maximum bandwidth path in a graph G is a path from a source S to a sink T, such that the bandwidth corresponds to the minimum capacity (frequency) edge in the path. One method to resolve the QSR problem is to continually find maximum bandwidth paths through the read graph. Repeat the following until S and T are disconnected (cannot exist an edge with frequency/count of zero or less than some number > 0 for real valued edges): Find an $S - T$ max-bandwidth path P with bandwidth m. Subtract bandwidth m from G-edges of P. This will exhaust at least one edge in the path, forcing the next iteration to choose a new path. Frequencies of the quasispecies are normalized from a count (maximum bandwidth) to a percentage simply by dividing a quasispecies count over the sum of all counts.

2.6.2 Maximum Frequency Path

A maximum frequency path in a graph G is a path from a source S to a sink T, such that the total frequency of the path is maximized. If frequencies are integer values, a maximum frequency path is equivalent to the longest path of the graph G. While LONGEST PATH is NP-hard in the general case, it is solvable in polynomial time on acyclic graphs using dynamic programming[10]. If the graph G is a directed acyclic graph (as is our case), it may be solved in linear time by using a topological ordering of vertices combined with dynamic programming for computing path lengths. Maximum frequency paths may be used to resolve the QSR problem in a fashion similar to the maximum bandwidth path method. In other words, until there exists no path from the source S to the sink T, find the maximum frequency path P with total frequency f. Set the corresponding frequency in G to 0 for each edge in the path P. Repeat until no non-zero frequency paths exist. Path frequencies are then normalized in the same manner as maximum bandwidth path. The concept is inspired by the greedy approximation algorithm for MIN ENTROPY SET COVER from [12]. The $OPT + 3$ approximation is achieved by greedily selecting a set that covers the most elements from the universe. After removing the selection from the universe, the algorithm repeats on the remaining sets until nothing is left to be covered. In a similar fashion, our algorithm covers the largest possible amount of reads in the graph, removes them, and then recurses.

2.6.3 Multi-commodity Flow Algorithm

While the previous flow-based algorithms utilize more global information then the local fork-resolution methods, solutions are still found iteratively. This may result in non-optimal solutions with respect to parsimony. This shortcoming can be addressed by modelling the problem as a multi-commodity flow. Given parameter k, the optimal solution to the following mixed integer program corresponds to k paths in the graph that cover all reads. The ILP is defined as the following,

$$\textbf{Minimize:} \quad \sum_{\substack{0 \le i \le k \\ (s,u) \in E}} f^i_{s,u}$$

$$\begin{aligned}
\textbf{Subject to:} \quad & \sum_{i}^{k} g^i_{u,v} \ge c_{u,v} && \forall (u,v) \in E \\
& \sum_{u \in \text{ pred}(v)} g^i_{u,v} = \sum_{u \in \text{ succ}(v)} g^i_{v,u} && \forall v \in V, i = 1 \ldots k \\
& \sum_{u \in \text{ succ}(v)} f^i_{v,u} = 1 && \forall v \in V, i = 1 \ldots k \\
& f^i_{u,v} \ge g^i_{v,u} && \forall (u,v) \in E, i = 1 \ldots k \\
& f^i_{u,v} \in \{0,1\} && \forall (u,v) \in E, i = 1 \ldots k \\
& g^i_{u,v} \in [0,1] && \forall (u,v) \in E, i = 1 \ldots k
\end{aligned}$$

where f variables represent the binary-flow over an edge, and g variables represent the fractional amount of flow. Together, these variables indicate which edges to select and what percentage of reads will be assigned to a given flow. Without the binary variable, it would not be possible to constrain the fractional flow to single paths. A consequence of this is that this ILP cannot be solved in polynomial time in the worst case. However, in practice it is found to perform quite well.

2.6.4 Results

Maximum bandwidth outperforms the maximum frequency method as well as guide distribution method in terms of sensitivity for quasispecies with uniform distribution (see figure 2.6). It correctly assembles seven out of ten quasispecies on average over the data set when given 100K reads and window length 300bp. As shift position increases (less overlap between windows) the quality of all four methods degrades. A shorter overlap would produce bi-cliques of larger size in the subgraph, which increases the complexity of the problem. Under the skewed distribution of quasispecies frequencies, both maximum bandwidth and

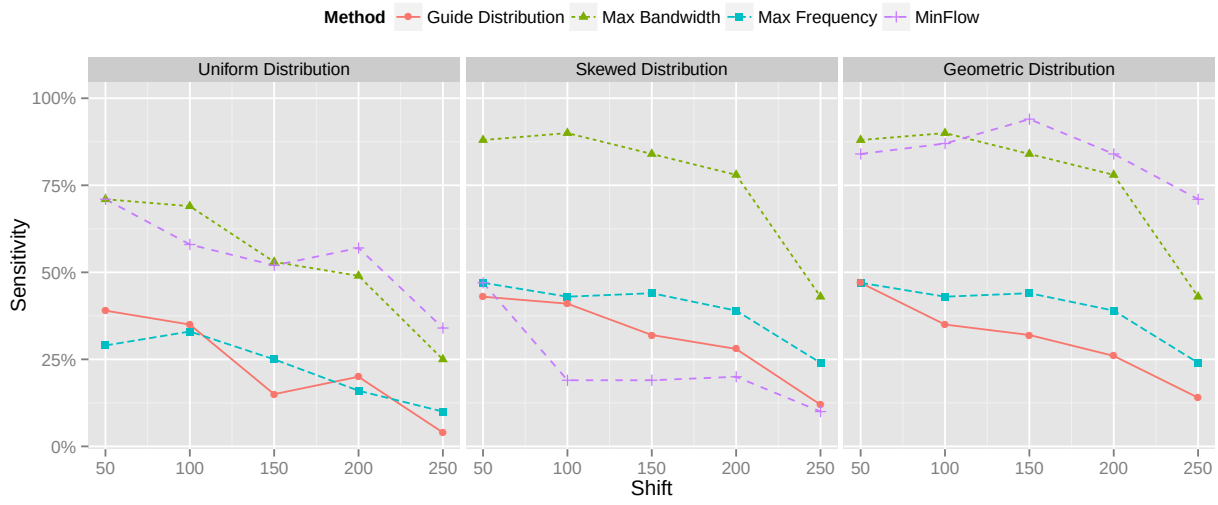

Figure 2.6 Sensitivity results on simulated HCV population over error-free data for flow-based algoirthms. The results are partitioned over each population distrubtion.

the multi-commodity method assemble between eight and nine out of ten correct variants of the population, while GD and maximum frequency produce four out of ten (see figure 2.6). Similarly under the geometric distribution, maximum bandwidth and the greedy resolution method correctly produce approximately nine and eight out of ten respectively, with GD and maximum frequency reconstructing five out of ten (see figure 2.6). Interestingly, positive predicted values for maximum frequency path tend to be more competitive with the other methods despite its typically lower sensitivity. This is most likely the result of maximum frequency paths quickly covering many reads in the graph each iteration (see figure 2.7). Similarly for maximum bandwidth, which exhausts the graph much faster than greedy methods.

2.7 Extension to Shotgun Reads

So far, we have focused on solving the reconstruction problem given amplicon reads. Still, shotgun sequencing offers a viable alternative approach. The benefit of shotgun sequencing is that it does not require complicated primer design of targeted regions. In order to solve the quasispecies reconstruction problem from shotgun reads, we provide a reduction

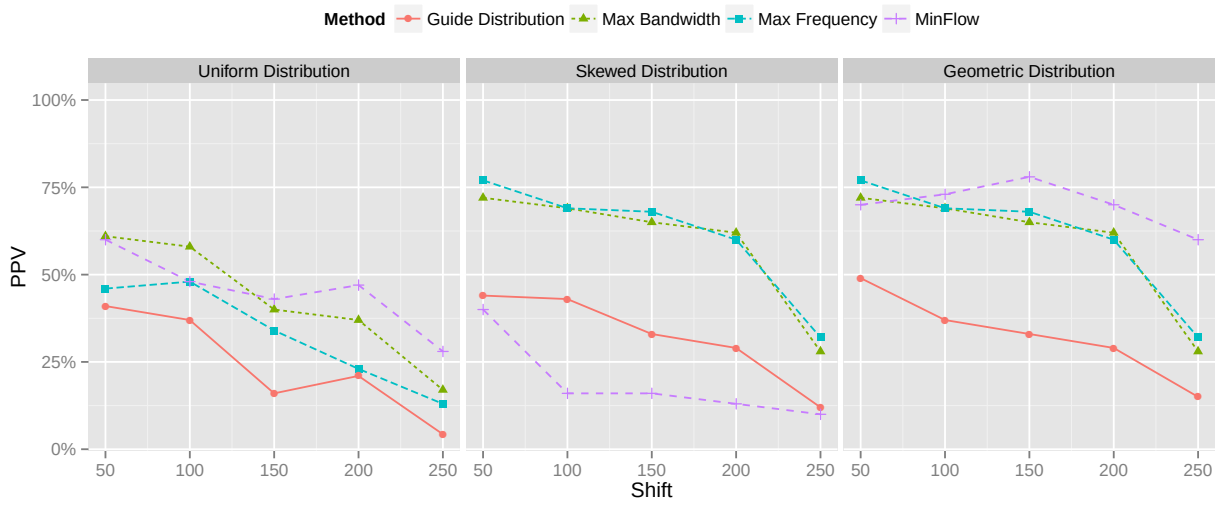

Figure 2.7 Positive predictive value results on simulated HCV population over error-free data for flow-based algoirthms. The results are partitioned over each population distrubtion.

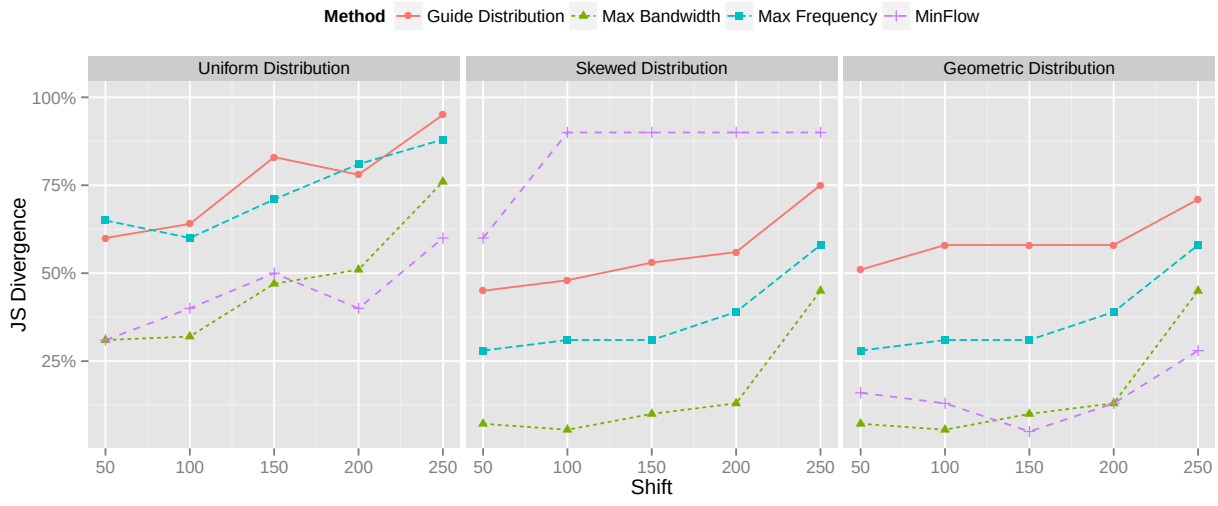

Figure 2.8 Jensen-Shannon divergence results on simulated HCV population over error-free data for flow-based algoirthms. The results are partitioned over each population distrubtion.

to the amplicon scenario. Indeed, by finding short overlapping window segments that span the region we can proceed with any of the previously described algorithms.

Given a set of shotgun reads and parameters for minimum and maximum window size, we reduce the shotgun quasispecies reconstruction problem to the amplicon case as follows. First, calculate sequencing statistics such as positional coverage and entropy. By calculating positional entropy, we have an idea on how diverse a given locus is. The rational is to partition the reads in such a way to have highly diverse overlaps. This reduces the overall complexity of the reconstruction problem. While there will be more forks, each fork will fewer possible choices. Random search is applied to intervals over the region of interest and scored based on Z-score statistics. This method was described initially in [14]. It runs quickly and finds acceptable overlaps rather than naively partitioning based on some fixed amount to shift. Results on shotgun data are presented in chapter 3.3.

PART 3

CORRECTING SEQUENCING ERRORS IN QUASISPECIES RECONSTRUCTION

3.1 Introduction and Contributions

In this section I present a formal derivation of the E and M steps for an EM algorithm to compute the maximum likelihood estimates for our generative model. This includes the mixture weights (i.e., genotype frequencies) along with the sequencing error-rates for each sequenced position. Prior to this, kGEM had not been fully formalized with some steps still needing justification. Additionally, this derivation elucidates the flexibility of the EM algorithm and how to estimate model parameters beyond simple mixing weights.

Furthermore, I describe software for reconstructing a viral population from NGS reads (shotgun or amplicon) dubbed **V**iral **A**ssembler (VirA). The software is a combination of both Java and Python. It makes use of state-of-the-art alignment software InDelFixer along with error-correction software kGEM. Once reads have been aligned, windows have been found, and reads corrected, VirA builds a read-overlap graph and computes a path-packing heuristic.

3.2 Error Correction by kGEM

The proposed kGEM algorithm attempts to find the set of variants (haplotypes) which is the most likely producing the observed set of reads R. We first describe an alignment of reads to the extended reference sequence and formally introduce the notion of a population genotype (fractional haplotype). Next we describe the iterative procedure of kGEM consisting of initialization and genotypes estimation.

The extended reference is obtained by identification of insertions using pairwise alignment of all reads (e.g. obtained using InDelFixer [15]) and multiple sequence alignment

of insertions. It may contain deletions due to insertions in reads, which we denote by the symbol `d`. We prioritize gap-minimization over mismatches minimization since we need to minimize the length of extended reference N. For each read, we assume that enumeration of alleles in it is the same as in the extended reference.

The population P could be represented as frequency distribution of alleles at each position of extended reference. Formally, given a population P, a *genotype* $G = G(P)$ of the population P is a matrix where each column corresponds to a position on the extended reference and each row corresponds to the allele $\{\mathtt{a}, \mathtt{c}, \mathtt{t}, \mathtt{g}, \mathtt{d}\}$. Each entry $f_m(e)$ is the frequency of allele e in mth position of genotype G, $\sum_e f_m(e) = 1$.

The algorithm finds a set of genotypes $\{G_1, G_2, \ldots, G_k\}$ which most adequately represents the population P. We initialize this set by selecting reads covering entire sequence space of the population P. In greedily manner we search for a read that maximizes the minimum Hamming distance to previously selected reads.

We now describe the generative model that characterizes the production of reads from genotypes. We define the probability that genotype G_i generated read r as

$$\Pr[r|G_i] = \prod_{j=1}^{|r|} (1 - 4\epsilon_j)^{\mathbb{I}[r_j = G_{ij}]} \epsilon_j^{\mathbb{I}[r_j \neq G_{ij}]}$$

where ϵ is the error-rate of the sequencing technology. The likelihood (and log-likelihood) of the population genotypes G along with their respective frequencies f is given by,

$$\mathcal{L}(G, f; R) = \prod_{r \in R} \left(\sum_{i=1}^{k} f_i \Pr[r|G_i] \right)^{n_r} \tag{3.1}$$

$$\ell(G, f; R) = \sum_{r \in R} n_r \cdot \log \left(\sum_{i=1}^{k} f_i \Pr[r|G_i] \right) \tag{3.2}$$

$$\tag{3.3}$$

where R is the set of distinct reads and n_r is the multiplicity of read r. As this function is non-convex and difficult to optimize in general, we would like to solve for the *complete-data*

log-likelihood defined as,

$$\ell_c(G, f; R) = \sum_{r \in R} \sum_{i=1}^{k} n_r \mathbb{I}[z_r = i] \log(f_i \Pr[r|G_i]) \tag{3.4}$$

where $\mathbb{I}$ is the indicator function and z_r are variables indicating which genotype generated read r. However, as we cannot possibly know a-priori values for $\mathbb{I}[z_r = i]$ we solve for the *expected* complete-data log-likelihood

$$Q(G, G^{t-1}) = \mathbb{E}\left[\sum_{r \in R} \sum_{i=1}^{k} n_r \mathbb{I}[z_r = i] \log(f_i \Pr[r|G_i])\right] \tag{3.5}$$

$$= \sum_{r \in R} \sum_{i=1}^{k} n_r \overbrace{\mathbb{E}[\mathbb{I}[z_r = i]]}^{p_{ri}} \log(f_i \Pr[r|G_i]) \tag{3.6}$$

$$= \sum_{r \in R} \sum_{i=1}^{k} n_r p_{ri} \log(f_i) + \sum_{r \in R} \sum_{i=1}^{k} n_r p_{ri} \log(\Pr[r|G_i]). \tag{3.7}$$

The MLE parameters can be found by the EM algorithm. The E-step is computed by simple Bayes' theorem,

$$p_{ri} = \frac{f_i \Pr[r|G_i^{t-1}]}{\sum_{i'} f_{i'} \Pr[r|G_{i'}^{t-1}]}. \tag{3.8}$$

The M-step for genotype frequencies f are computed by taking the derivative of Q wrt to f with the added lagrange constraint that $\sum_i f_i = 1$,

$$f_i = \frac{\sum_{r \in R} n_r p_{ri}}{|R|}. \tag{3.9}$$

Genotype error-rates ϵ (with slightly more work) are computed by

$$\frac{\partial Q}{\partial \epsilon_j} = \frac{4\overbrace{\sum_{r\in R}\sum_{i=1}^{k} n_r p_{ri} \mathbb{I}[r_j = G_{ij}])}^{N_j}}{4\epsilon_j - 1} + \frac{\sum_{r\in R}\sum_{i=1}^{k} n_r p_{ri}(1 - \mathbb{I}[r_j = G_{ij}])}{\epsilon_j} \tag{3.10}$$

$$= \frac{4N_j}{4\epsilon_j - 1} + \frac{|R| - N_j}{\epsilon_j} \tag{3.11}$$

$$\epsilon_j = \frac{|R| - N_j}{4|R|} \tag{3.12}$$

where N_j is the weighted number of matches in position j. We see that this calculation also makes intuitives sense: the error rate is the proportion of weighted mismatches to (four times) the number of reads. The reason four shows up is to account for an error being possible for each of the 3 other nucleotides (or deletion).

3.2.1 Threshold Determining

Dispite being able to infer the error-rates for the sequenced population, it would still be useful to filter results based on statistical significance. Hence, we utlize a simple test to determine the minimum possible threshold t_j for a given position. If a given genotype does not meet this criterion it is dropped from consideration. The EM algorithm estimates the sequencing error-rate, hence we can find define the probability that there exist at least q errors in position j as

$$f_j(q) = \Pr[\text{Num errors in position } j \geq q] = \sum_{\theta=q} \binom{|R|}{\theta} \epsilon_j^{\theta} (1 - \epsilon_j)^{|R|-\theta}. \tag{3.13}$$

We can determine the minimum threshold t_j given a Bonferroni corrected statistical significance α/c by computing

$$t_j = \arg\min_t (f_j(t) \leq \alpha/c)$$

where c is the number of reads. As f_j is a distribution function and therefore monotonic, t_j can be found by simple binary search over the possible threshold values.

3.2.2 Model Selection

kGEM provides a reasonable model to correct errors in a small viral genomic region. However, selecting k a-priori may be difficult. To address this issue, model selection may be performed. We investigated both Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Given k (model size) and ℓ (log-likelihood), and n (number of data points), AIC and BIC are defined as,

1. $\text{AIC} = 2k - 2\ell$
2. $\text{BIC} = k \log n - 2\ell$.

A straightforward application to kGEM may seem at first a beneficial idea; however, the "final" k is typically much less than the original k. This is a result of the algorithm's default behavior. Genotypes with frequency less than a given threshold are dropped while others may be merged. To adjust for this, the threshold value can be set to 0.

3.3 VirA: Viral Assembler

VirA takes as input the read data in fasta format, a reference, and the sequence-specific primers used for amplicon experiments. Reads are aligned to the reference using InDelFixer version 0.8[15] and then partitioned into amplicons. The individual amplicons are re-aligned to adjust for potential frame-shifts and corrected for sequencing errors using the tool kGEM[16]. All corrected amplicons are recombined for a final alignment before reconstruction. This reduces the possibility for frame-shift in overlapping segments. A read-overlap graph is then constructed for assembly. Depending on the parameters, VirA performs either the maximum bandwidth-path heuristic[17] or the exact algorithm based on a multi-commodity flow formulation[18] to reconstruct the viral population. Briefly, the maximum-bandwidth algorithm iteratively searches for the best possible single-path flow. After a path is found the read-counts are updated and the algorithm tries again. This continues until the graph reaches saturation. The multi-commodity approach simultaneously finds k flow-paths optimally covering the read-data.

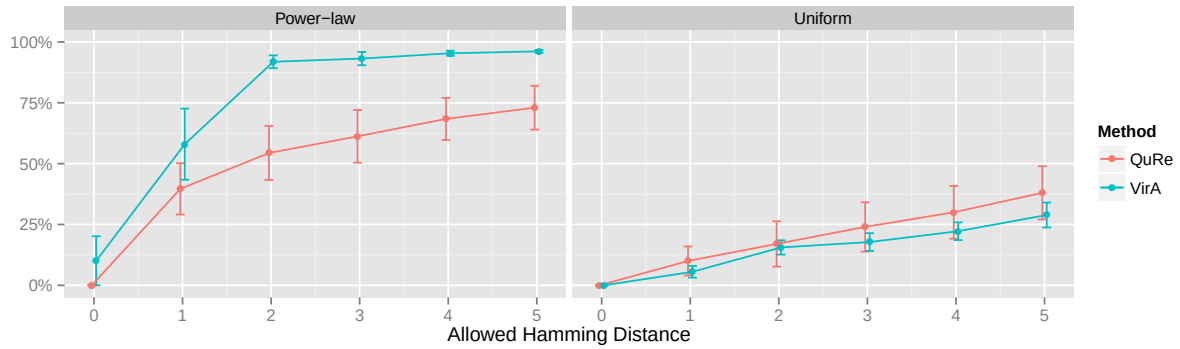

Figure 3.1 Results obtained from simulated amplicon reads HCV data using sensitivity (weighted portion of true variants found). We relax the case of requiring an exact match and allow for hamming distance.

3.4 Datasets and Experiment Design

To validate VirA we simulated 10 viral populations from a sample of 43 HCV variants targeting the E1E2 region. Each population was duplicated into two data sets with variant abundances adhering to a uniform and power-law ($\alpha = 2.0$) distribution. Primers were designed to generate between 7 to 12 amplicons covering the 1734nt long region for amplicon-based reads. Both amplicon and shotgun-based reads were generated using Grinder version 0.5 containing mutation errors (substitution & indel) uniformly at 0.1% along with homopolymer errors according to the Balzer model[19]. Weighted sensitivity and positive predictive value were used to assess the reconstructed population quality

3.5 Results

We validated VirA against QuRe using simulated HCV sequencing data. In all of the cases VirA outperformed QuRe in both sensitivity and ppv. While results for both programs are mediocre, this may be attributed to the aggressive error model used to simulate reads. By only using a homopolymer error model rather than the additional mutational errors, we expect both tools to perform much better.

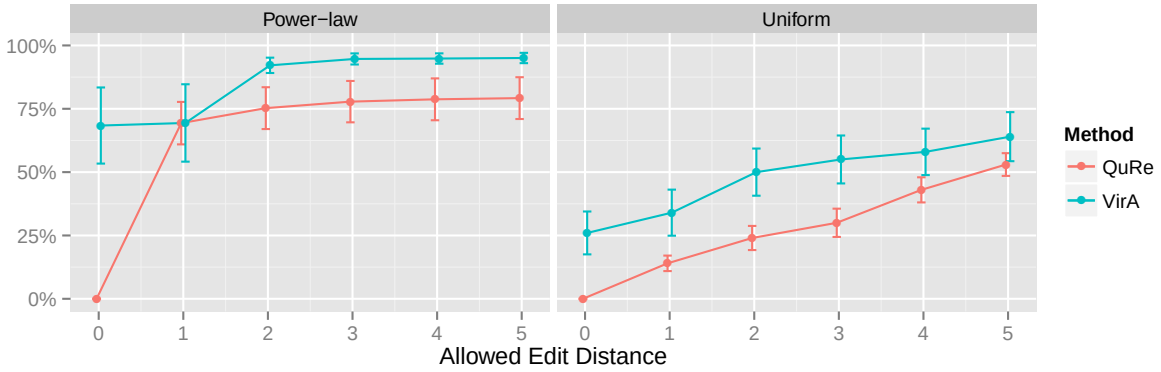

Figure 3.2 Results obtained from simulated shotgun reads HCV data using sensitivity (weighted portion of true variants found). We relax the case of requiring an exact match and allow for hamming distance.

PART 4

VIRAL QUASISPECIES RECONSTRUCTION FROM PAIRED-END READS

4.1 Introduction and Contributions

We propose a method to overcome these limitations by coupling a high-fidelity sequencing protocol [20] with an accurate method, referred to as **V**iral **G**enome **A**ssembler (VGA), to assemble a heterogeneous viral population. High-fidelity sequencing protocol, known as Safe-SeqS, has been applied to detect rare somatic mutations, but its application on detecting rare viral mutations has been neglected. Similar to Safe-SeqS we apply a special library preparation technique that eliminates sequencing errors during the de-multiplexing step. The proposed protocol attaches individual barcode sequences during the library preparation step for every fragment, then amplifies each tagged fragment. Reads are clustered according to the original fragment based on the attached barcode. An error-correction protocol is then applied for every read group resulting in a method that corrects errors inside the group and produces a corrected consensus read. Highly accurate data in combination with deep coverage allows for accurate estimation of the underlying diversity of a viral population. Importantly, the low per-base sequencing cost of the Illumina platform makes it realistic to greatly increase coverage in order to detect ultra-rare variants. Our sequencing protocol introduces novel challenges for virus assembly and we develop a novel assembly approach for reconstructing and estimating the frequency of a large number of closely related viral variants. Our method does not rely on the availability of a reference genome. This makes our method applicable to newly emerged viruses in which genome sequences are unknown.

The ability to discover rare viral variants makes our tool applicable for monitoring and quantifying an HIV population structure in order to dissect its evolutionary landscape and study genomic interaction. In particular, our approach allows for the discovery of rare mutations and variants which are of particular interest because of their potential influence

on drug resistance and treatment failure [8, 21, 22].

4.2 Methods

4.2.1 Overview

Advances in next generation sequencing and the ability to generate deep coverage data in the form of millions of reads provide exceptional resolution for studying the underlying genetic diversity of complex viral populations. However, errors produced by most sequencing protocols complicate distinguishing between true biological mutations and technical artifacts that confound detection of rare mutations and rare individual genome variants. A common approach is to use post-sequencing error correction techniques able to partially correct the sequencing errors. In contrast to clonal samples the post-sequencing error correction methods are not well suited for mixed viral samples and may lead to filtering out true biological mutations. For this reason, current viral assembly methods are able to detect only highly abundant single nucleotide variant (SNV), thus limiting the discovery of rare viral genomes.

Additional difficulty arises from the genomic architectures of viruses. Long common regions shared across viral population (known as conserved regions) introduce ambiguity in the assembly process. Conserved regions may be due low-diversity population or due to recombination with multiple cross-overs. In contrast to repeats in genome assembly conserved regions may be phased based on relative abundances of viral variants. Low-diversity viral populations in which all pairs of individual genomes within a viral population have a small genetic distance from each other may represent additional challenges for the assembly procedure.

We apply a high-fidelity sequencing protocol to study viral population structure (Figure 4.1). This protocol is able to eliminate errors from sequencing data by attaching individual barcodes during the library preparation step. After the fragments are sequenced, the barcodes identify clusters of reads that originated from the same fragment, thus facilitating error correction. Given that many reads are required to sequence each fragment, we are trading

off an increase in sequence coverage for a reduction in error rate. Prior to assembly, we utilize the *de novo* consensus reconstruction tool, Vicuna [23], to produce a linear consensus directly from the sequence data. This approach offers more flexibility for samples that do not have "close" reference sequences available. Traditional assembly methods [24–26] aim to reconstruct a linear consensus sequence and are not well-suited for assembling a large number of highly similar but distinct viral genomes. We instead take our ideas from haplotype assembly methods [27, 28], which aim to reconstruct two closely related haplotypes. However, these methods are not applicable for assembly of a large (*a priori* unknown) number of individual genomes. Many existing viral assemblers estimate local population diversity and are not well suited for assembling full-length quasi-species variants spanning the entire viral genome. Available genome-wide assemblers able to reconstruct full-length quasi-species variants are originally designed for low throughput and are impractical for high throughput technologies containing millions of sequencing reads.

We introduce a viral population assembly method (Figure 4.2) working on highly accurate sequencing data able to detect rare variants and tolerate conserved regions shared across the population. Our method is coupled with post-assembly procedures able to detect and resolve ambiguity raised from long conserved regions using expression profiles.(Figure 4.2.F). After a consensus has been reconstructed directly from the sequence data, our method detects SNVs from the aligned sequencing reads. Read overlapping is used to link individual SNVs and distinguish between genome variants in the population. The viral population is condensed in a conflict graph built from aligned sequencing data. Two reads are originated from different viral genomes if they share different SNVs in the overlapping region. Viral variants are identified from the graph as independent sets of non-conflicting reads. Non-continuous coverage of rare viral variants may limit assembly capacities, indicating that increase in coverage is required to increase the assembly accuracy. Frequencies of identified variants are then estimated using an expectation-maximization algorithm. Compared with existing approaches, we are able to detect rare population variants while achieving high assembly accuracy.

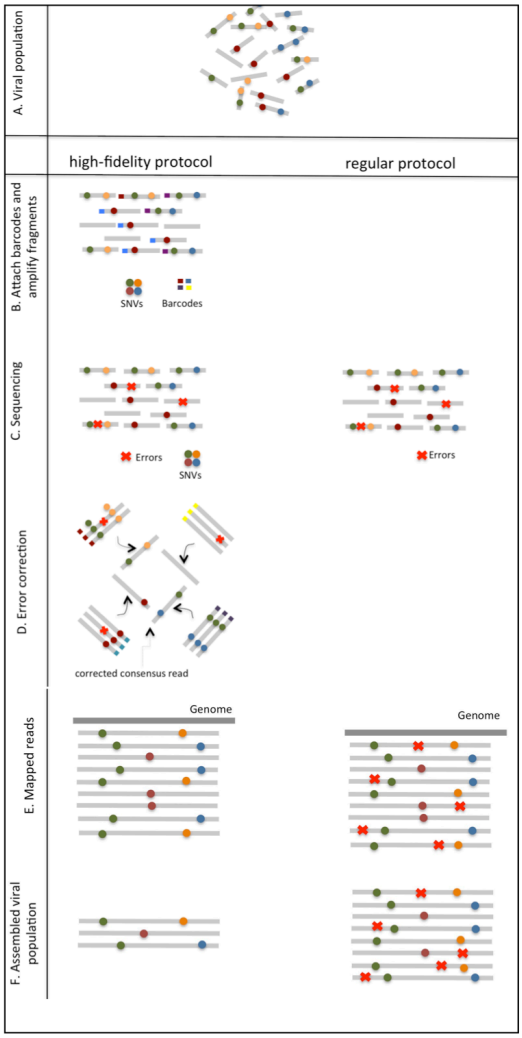

Figure 4.1 Overview of high-fidelity sequencing protocol. (a) DNA material from a viral population is cleaved into sequence fragments using any suitable restriction enzyme. (b) Individual barcode sequences are attached to the fragments. Each tagged fragment is amplified by the polymerase chain reaction (PCR). (c) Amplified fragments are then sequenced. (d) Reads are grouped according to the fragment of origin based on their individual barcode sequence. An error-correction protocol is applied for every read group, correcting the sequencing errors inside the group and producing corrected consensus reads. (e) Error-corrected reads are mapped to the population consensus. (f) SNVs are detected and assembled into individual viral genomes. The ordinary protocol lacks steps (B) and (D).

4.2.2 Error correction

The proposed sequencing is able to eliminate errors from sequencing data and produce highly accurate read sequences. It uses a high-fidelity sequencing protocol which attaches individual barcodes during the library preparation step. The barcodes are then used to identify reads originated from the same fragment, allowing to access multiple sequencing data of the same fragment. It follows that every sequenced position of the fragment would have multiple independent evidence, suitably promoting highly accurate consensus reads. By applying an error-correction procedure of the protocol, we are able to address both sequencing and PCR errors which leads to high assembly accuracy.

4.2.3 Consensus construction

We build a consensus from paired-end reads using Vicuna [23]. Our sequencing method should not contain any particularly low coverage region allowing reconstruction of population consensus for viral sample. In the event that Vicuna produces multiple contigs rather than a complete consensus, we use BLAST to merge contigs. We require 50nt overlap to merge any pair of contigs. In the next step, the population consensus is used as a reference genome to map reads. Building the reference genome from actual sequencing data rather than using an annotated genome provides us with an accurate and unique mapping.

4.2.4 Read mapping

As with many viral population analyses, the first step of VGA is to map the reads. We map reads onto the de novo consensus using InDelFixer [15] with default parameters. False read alignments are filtered out using fragment length distribution inferred from the mapping data. Assuming that the fragment length follows a normal distribution, we only keep reads with fragment length within three standard deviations from the mean. In total 1.2% of reads have been filtered out vs expected .3% according to the three-sigma rule.

4.2.5 Viral population assembly

The combination of deep coverage with high accuracy provides an unprecedented opportunity for estimating genomic diversity in a viral population. The viral population assembly starts with determining pairs of mapped reads conflicting with each other in the overlapping region. Following [29], we construct the conflict graph $G = (V, E)$ with vertices corresponding paired-end reads, i.e., $V = R$, and edges connecting conflicting pairs of paired-end reads.

Obviously, any true viral genome corresponds to a maximal independent set in the conflict graph (i.e., a maximal set of pairwise nonadjacent vertices), although, not every maximal independent set necessarily corresponds to a true viral genome. We adopt a parsimonious approach requiring to cover the conflict graph with the minimum number of maximal independent sets. This problem is equivalent to MIN-GRAPH-COLORING which is NP-hard. There exists many heuristics for solving this problem (see, e.g., [30, 31]) based on greedy selection of a maximal independent set. Unfortunately, our attempts to build even a single viral genome failed since it is difficult to arrange paired-end reads into a connected single path. Indeed, a greedy algorithm runs out of any possible extension after just a few steps while concatenating paired-end reads from left-to-right.

Instead, we apply an alternative "top-down" approach of recursive graph partitioning along the maximum cut (Max Cut) which has been previously successfully applied for human haplotyping [32]. Given a graph $G = (V, E)$, the Max Cut problem asks for partitioning of the vertices into two components $V = V_1 \cup V_2$ maximizing the total number of edges which have one endpoint in V_1 and the other in V_2. The Max-Cut problem though NP-hard is well approximated by a simple 0.5-approximation algorithm which randomly assigns vertex to one of the two components [33]. Our Max-Cut heuristic starts with alternatively assigning left-to-right sorted mapped reads to two components and then repeatedly moves one vertex at a time from one component to another, improving the solution at each step, until no more improvements of this type can be made.

Our coloring heuristic recursively partitions the conflict graph until each component becomes independent. If reads of a given color completely cover the consensus genome, then

the resulted sequence is accepted as the next viral genome. Otherwise, if assembled genome contains gaps, we add non-conflicting reads from other color classes in left-to-right order in attempt to fill the gaps. If all SNV positions are covered, then a newly reconstructed viral genome is added to the set $\mathcal{VG}$. Finally, the genomes whose gaps cannot be filled with the above procedure are dropped.

Algorithm 1 VGA assembly algorithm

Input : Set of reads R aligned to the consensus genome
Build conflict graph G=(V,E) from set R
Recursively color G into color classes $\mathcal{C}$ using Max-Cut
Initialize the set of complete viral genomes $\mathcal{VG} \leftarrow \emptyset$
for each color class $c_i \in \mathcal{C}$ **do**
  Compute maximal independent set in $G = (V, E)$ containing c_i
  Assemble reads in c_i into viral genome g_i
  if g_i covers all positions in the consensus genome **then**
    $\mathcal{VG} \leftarrow \mathcal{VG} \cup \{g_i\}$
  end if
end for
Output: Set of complete viral genomes $\mathcal{VG}$

4.2.6 Viral population quantification

In the final step of the workflow an expectation-maximization algorithm is used to infer the relative abundances of assembled viral quasi-species similar to what is described in [34]. We extend the previous EM and likelihood formulation to incorporate a prior probability for the viral population and compute the *maximum a-posteriori* estimate, rather than the MLE.

Let H be a random variable over the set of viral variant genomes $\mathcal{H} = \mathcal{VG}$ and let R be a random variable over the set of reads $\mathcal{R}$. Let $p[\mathcal{H}] \sim \text{Dir}(\alpha_1, \ldots, \alpha_{|\mathcal{H}|})$ be the prior probability of observing a given set of variants and denote $p_h = \Pr[H = h]$ to be the probability of observing a particular variant h. The probability of observing read $r \in \mathcal{R}$ is

given by marginalizing over all variants

$$\Pr[R = r] = \sum_{h \in \mathcal{H}} \Pr[R = r | H = h] \cdot p_h,$$

where

$$\Pr[R = r | H = h] = \begin{cases} 1/K_h & \text{if } r \text{ is consistent with } h \\ 0 & \text{otherwise} \end{cases}.$$

and K_h is the number of reads consistent with h. We can now define the log-posterior as

$$\log \Pr[\mathcal{H} | \mathcal{R}] = \sum_{r \in \mathcal{R}} n_r \cdot \log \Pr[R = r] + \sum_{h \in \mathcal{H}} (\alpha_h - 1) \log p_h - C_{\mathcal{R}},$$

where $C_{\mathcal{R}}$ is a constant and n_r is the number of reads r. As this function is non-convex and difficult to globally optimize for, we solve the easier problem of maximizing its lower-bound,

$$\sum_{r \in \mathcal{R}} \sum_{h \in \mathcal{H}} n_{rh} \cdot \log \left(\Pr[R = r | H = h] \cdot p_h \right) + \sum_{h \in \mathcal{H}} (\alpha_h - 1) \log p_h.$$

where n_{rh} is the expected number of reads r generated by variant h. The EM algorithm computes this by

$$n_{rh} = n_r \cdot \frac{p_h \cdot \Pr[R = r | H = h]}{\Pr[R = r]},$$

and subsequently maximizes the log-posterior with the MAP estimate given by

$$\hat{p_h} = \left(\alpha_h - 1 + \sum_{r \in \mathcal{R}} n_{rh} \right) \Big/ \left(|\mathcal{R}| + \sum_{h \in \mathcal{H}} \alpha_h - |\mathcal{H}| \right).$$

We see that by setting a uniform prior, i.e., $\alpha_h = 1$ for all $h \in \mathcal{H}$, we obtain the original MLE formulation.

4.3 Results

4.3.1 Performance of VGA on simulated data

Since the ground truth is unknown for sequenced viral populations, simulations present a standardized way to assess the performance of viral assembly tools. The proposed high-fidelity protocol allows to correct sequencing errors, thus giving access to highly-accurate sequencing data. Post-sequencing error correction techniques are available for reads obtained by regular protocol offering the possibility to partially correct sequencing errors trading off for real biological mutations. Grinder [19] is used to generate reads from both the high-fidelity and regular sequencing protocol. Reads are generated from both real and synthetic viral variants with different sequencing parameters and viral expression profiles. Grinder is a state-of-the-art sequencing read simulator able to produce shotgun sequencing data from a viral population with different expression profiles. We mapped the simulated paired-end reads onto the consensus using Mosaik. The consensus was constructed using Vicuna [23], a *de novo* assembly tool able to produce a linear consensus from deep paired-end sequencing data (see Section 4.2.3 for details).

We use sensitivity and positive predictive value (PPV) to evaluate the quality of viral genomes assembled by VGA. We consider fully assembled viral genome without errors. Sensitivity is defined as the portion of assembled quasi-species that match true quasi-species, i.e., $Sensitivity = TP/(TP+FN)$. Positive predictive value is defined as the portion of true sequences among assembled sequences, i.e., $PPV = TP/(TP+FP)$. Additionally, we evaluate ability of our method to estimate population size (i.e. number of viral genomes in the population). Accuracy of population size prediction is defined as a ratio between estimated and true population sizes. Finally, we use Jensen-Shannon divergence (JSD) to measure the accuracy of frequency estimation. Given two probability distributions, JSD measures the "distance" between them, or in other words, the quality of approximation of one probability distribution by the other distribution. It is defined as the Kullback-Leibler divergence from distributions P and Q to their mixture. Formally, the Jensen-Shannon divergence between

true distribution P and approximation distribution Q is given by the formula

$$\mathrm{JSD}(P||Q) = \frac{1}{2}\mathrm{D}_{KL}(P||M) + \frac{1}{2}\mathrm{D}_{KL}(Q||M)$$

where Kullback-Leibler divergence D_{KL} is

$$\mathrm{D}_{KL}(P||Q) = \sum_{i=1}^{n} P(i) \log \frac{P(i)}{Q(i)}$$

and $M = \frac{1}{2}(P + Q)$. The motivation for using JSD is a consequence of KL divergence being undefined when assembly methods fail to reconstruct some variant i, hence forcing $Q(i)$ to be 0. JSD averts this by measuring the distance to the mixture, which contains all true and called variants (TP and FP).

Our first simulated study compares the assembly accuracy across different virus species. We focus on effect of read-length and throughput on assembly quality for different types of viruses. Paired-end reads of various length corresponding to high-fidelity and regular sequencing protocols are simulated from HIV and HCV populations assuming uniform and power-law distributions. A power-law distribution (i.e., frequency of an individual viral genome is a power of the previous one) corresponds to a population with several dominant variants and many rare variants. The uniform distribution has equal frequencies for all viral genomes. HCV population is presented by 1739-bp long fragment from the E1E2 region of 44 real HCV sequences. HIV population consist of 10 real intra-host viral variants mixture from 1.3Kb-long HIV-1 region, which included pol protease and part of the pol reverse transcriptase [35].

The genomic architecture across virus species was investigated and its influence on assembly accuracy was studied. HCV virus exhibits more complex genomic architecture with lower population diversity and longer conserved regions (Figure 4.3) than HIV. Conserved regions were present in both viruses, although only HCV contains conserved regions longer then 450bp. Conserved regions longer then the average fragment length (450bp) may introduce ambiguity in the assembly process due to a lack of direct evidence of sub-genomes

phasing across the conserved region. We performed simulated sequencing experiment where the average fragment amplification rate is 5, resulting in a five time decrease in throughput due to the consensus error correction performed by the high-fidelity sequencing protocol. Also the simulation experiments were adjusted to simulate a non-uniform amplification rate. Non-uniform amplification rate results in discarding fragments with insufficient amplification rate (less then 3). From real studies it is known that around 10% of fragments are amplified less the three times. Sequencing errors produced by the regular protocol limited the ability of VGA to accurately assemble a viral population. All assembled variants contained large number of mismatches, additionally VGA significantly overestimated population size.

As expected, short read lengths dramatically inhibit reconstruction, which is evidenced by VGA failing to produce any full-length genomes when given 2x36bp reads (Figure 4.6). Since common regions for distinct HCV viral genomes are significantly longer than for HIV, it is not surprising that performance of VGA is worse on HCV data – for 3M 2x150bp reads simulated from 44 1739bp-long viral genomes, sensitivity is 50%, PPV is 80% . Results on HCV data confirm that the lower mutation rate and presence of conserved regions have a negative impact on the ability to accurately reconstruct individual viral genomes. Surprisingly, increasing the read length for HIV from 100bp to 150bp yields no benefits for reconstruction accuracy suggesting that 100bp read length is enough to distinguish between HIV viral variants with high mutation rate. Although further experiments are needed to determine optimum read length, our simulations suggest that 2x100bp is recommended for small HIV viral populations and 2x150bp is recommended for medium HCV population with complex genomic architecture.

We separately analyzed the ability of our method to estimate the viral population size (i.e. number of genomic variants present in the population). Non-continuous coverage limits the ability of the method to assemble full length viral variants To evaluate the accuracy of population size estimation we compared the true population size known from simulated data with estimated results. Continuous coverage of each individual viral genome present in the sample has a strong impact on quality of population assembly. The probability of non-

continuous coverage increases dramatically for viral genomes with low abundance. Thus, the presence of coverage gaps for rare variants introduce additional challenges in the assembly process, making rare genomes unreachable by assembly tools. The number of problematic genomes can be reduced by increasing sequencing depth; however, it does not guarantee complete elimination. While complete assembly of all such genomes is unrealistic, it is still possible to estimate the number of viral genomes present in the sample (population size). The number of independent sets reported by VGA provides us with an accurate population size estimation. Intuitively, predicting the population size of a large viral population with many rare variants is more difficult than predicting for uniformly distributed or small populations (Figure 4.4). The predicted population size may serve as an indication of insufficient coverage to detect the full viral diversity present in the sample.

Deep coverage is a key for accurate estimation of underlying viral diversity. One such platform capable of offering millions of sequencing reads is Illumina HiSeq. The relatively short length of the produced reads is compensated for by sequencing the same fragment from both ends; therefore, producing coupled reads separated by a "gap", known as paired-end read. To our knowledge, VGA is the first method scalable to millions of short paired-end sequencing reads able to produce full-length viral variants spanning the entire viral genome. We explore the influence of sequencing depth on the reconstruction accuracy for varying population structures (uniform and power-law distributions of viral genomes within the population). HIV-1 is known to have greater genetic variability than any other known virus[36]. The diversity among viral genomes in an HIV population can vary from 3% to 20% depending on regions [37, 38]. Heterogeneous viral samples was prepared by generating viral populations from the Gag/Pol 3.4 Kb HIV region. We simulated variant abundances adhering to either a uniform or power-law distribution. Not surprisingly, our simulations suggest that increased sequencing depth has a direct positive effect on the discovery of rare variants and improves the overall assembly accuracy. Figure 4.5 shows the effect of coverage and population size on assembly for reads of length 100bp. Throughout all experiments, VGA maintained a PPV value of 100%.

In addition to point mutations, genetic recombination facilitates rapid evolution and production of diverse HIV genomes. Indeed, co-infected cells may produce recombinant viral progeny at levels lower than mutation rates in an intra-patient environment [39]. Hence, simulated datasets must account for both possible phenomena when determining the quality of assembly. We utilize a simulation model able to integrate both point mutations and recombination in the generated viral population depending on the amount of diversity required. A mixture of 10 real intra-host viral variants from 1.3kb-long HIV-1 form the basis population. In addition to point mutations, our simulation model implicitly produces recombinant genomes by first constructing the genotype (i.e., sequence of SNVs) for the population. A random walk is performed over this genotype as specified number of times. Any cross-over that occurs represents a new recombination between the "left" and "right" original genomes. Recombinations are implicitly produced and no control is imposed over number and length of the recombination. This model produces highly recombinant data on average, posing challenges for assembly and can be used to assess assembly quality. Simulation model incorporate mutation into the process by selecting a position and nucleotide-swap uniformly at random. Simulation results (Figure 4.7) suggest that our method can accurate assemble viral population in presence of recombinations and point mutations, maintaining PPV of 100%.

Finally, we evaluate population quantification accuracy, i.e. the accuracy of our method in prediction abundances of the assembled variants. Taking the results from VGA on 10 real HIV clones with 50,000 reads and 2x100bp, the Jensen-Shannon Divergence was 2.93e-05 for the Power-law and 0.001 for the Uniform-based populations. This already small measure only decreases as the size of the input grows.

4.3.2 Performance of existing viral assemblers on simulated consensus error-corrected reads

We have evaluated the performance of ShoRAH [40] and QuasiRecomb [9] for simulated consensus error-corrected read data.

ShoRAH disregards pairing information of reads but it is scalable enough to handle up to 1M reads. ShoRAH fails to produce full-length viral genome but reliably spans 98% of the consensus genome. It reasonably estimates the number of different viral genome but even the most accurate ShoRAH-assembled viral genome differs from the closest true 1.3Kb-long viral genome in 5 nucleotides.

QuasiRecomb is designed to handle paired-end read data and manages to produce full-length viral genomes. Unfortunately it can reliably process no more than 100K reads. Also the number of assembled distinct viral genomes 10 to 200 times more than the number of true distinct viral genomes. The most accurate QuasiRecomb-assembled viral genome still differs from the closest true 1.3Kb-long viral genome in 4 nucleotides.

Unfortunately we could not compare our method with QColors [29] assembly algorithm, which employs a similar conflict graph to represent viral population. A CSP solver is used by QColors for coloring the graph which may limit its scalability to high-throughput datasets consisting of millions of sequencing reads. Currently, QColors is not publicly available[1].

4.3.3 Performance of VGA on real HIV data

To further test the ability of VGA to accurately assemble a diverse natural occurring population and predict variant abundance levels we used an Illumina HiSeq HIV dataset, which consisted of 15M 2×100bp paired-end reads with attached barcodes. Next, the high-fidelity sequencing protocol able to eliminate sequencing errors was applied resulting in 3.2M consensus error-corrected reads (further referred to as reads). The reads were then used to build *de novo* population consensus using Vicuna 1.3 [23]. When run on our real data, Vicuna produced four contigs of average length 1195bp. Each contig was then run through BLAST to check for overlaps. Once overlaps were found, the contigs were assembled into a final consensus of length 4337bp.

[1]Upon querying for information on obtaining QColors, the authors were informed that the original software was tightly coupled for the analyses done in its original manuscript, and is not currently available for general use.

Validation of de novo consensus. A de novo assembled consensus was compared against reference-based consensus. To produce reference-based consensus we iteratively map reads onto the HIV reference (Gag/Pol 3.4 Kb HIV region) using InDelFixer. InDelFixer iteratively changes the reference genome based on the mapping of the current iteration. Also, we used InDelFixer in single iteration mode to map reads onto the constructed de novo consensus. De novo consensus is longer (4337bp) than the reference-based consensus (3440bp) and contained two regions with extremely peaked coverage compared to the surrounding regions. Both regions were considered to be the result of technical artifacts and removed from further consideration. After removing both regions, the length of new de novo consensus becomes 3452bp. We also filtered reads that belonged to regions with extreme coverage. Finally we compared the number of reads mapped to the reference-based consensus vs the de novo consensus. A larger amount of reads mapped to the assembled consensus, thereby highlighting the advantage of de novo procedure for consensus construction over a reference-based.

From the de novo consensus, VGA assembled 32 full-length viral genomes which differ from each other in 2,145 SNVs. Among known HIV sequences, Gag/Pol is the closest to the de novo consensus. Each of the 32 full-length viral genomes do not contain stop codons inside two known coding regions of Gag/Pol of length 1520bp and 1820bp, respectively. Alternatively, when VGA is applied to all 15M original uncorrected reads, 57 distinct viral genomes are assembled among which 36 contain stop codons in the two coding regions. This shows that a regular sequencing protocol is unsuitable for viral genome reconstruction.

4.4 Discussion

We have presented VGA, an accurate method for viral population assembly from ultra-deep sequencing data. The proposed algorithm is coupled with a high-fidelity sequencing protocol able to eliminate errors from sequencing data. Deep coverage in combination with highly accurate data allows our method to accurately estimate the underlying diversity of a viral population. In particular, it makes possible to distinguish true biological mutations from

sequencing errors, facilitating assembly of rare individual genomes. Our method condenses the viral population into a conflict graph built from aligned reads. To distinguish between viral variants, the conflict graph is colored into a minimal set of colors. Each color represents individual viral genomes composed from the set of non-conflicting reads. An Expectation-Maximization algorithm was used to estimate relative abundance frequencies of assembled viral genomes.

To our knowledge, our method is the first viral assembly method which scales to millions of paired-end sequencing reads. Experiments on both real and synthetic HIV datasets generated with various sequencing parameters and distribution assumptions suggest that VGA is able to assemble diverse viral population from millions of paired-end reads. The ability of our method to maintain 100% assembly accuracy makes it suitable for clinical applications. In addition, the constant increase of sequencing depth offered by high-throughput technologies provide us with unprecedented resolution promising to increase number of discovered ultra-rare viral variants in the population.

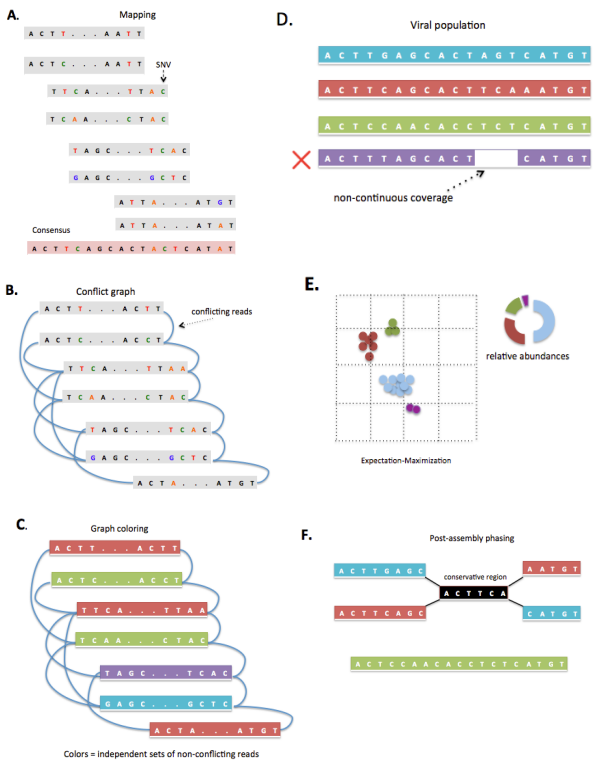

Figure 4.2 Overview of VGA. (a) The algorithm takes as input paired-end reads that have been mapped to the population consensus. (b) The first step in the assembly is to determine pairs of conflicting reads that share different SNVs in the overlapping region. Pairs of conflicting reads are connected in the "conflict graph". Each read has a node in the graph, and an edge is placed between each pair of conflicting reads. (c) The graph is colored into a minimal set of colors to distinguish between genome variants in the population. Colors of the graph correspond to independent sets of non-conflicting reads that are assembled into genome variants. In this example, the conflict graph can be minimally colored with four colors (red, green, violet and turquoise), each representing individual viral genomes. (d) Reads of the same color are then assembled into individual viral genomes. Only fully-covered viral genomes are reported. (e) Reads are assigned to assembled viral genomes. Read may be shared across two or more viral genomes. VGA infers relative abundances of viral genomes using the expectation-maximization algorithm. (f) Long conserved regions are detected and phased based on expression profiles. In this example turquoise and red viral genome share a long conserved region(colored in black). There is no direct evidence how the viral sub-genomes across the conserved region should be connected. In this example 4 possible phasing are valid. VGA use the expression information of every sub-genome to resolve ambiguous phasing.

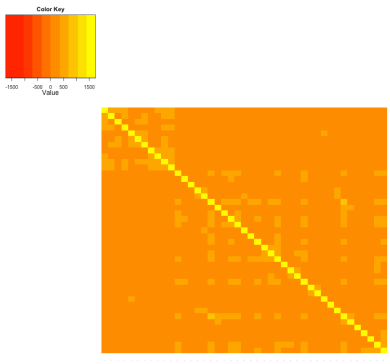

Figure 4.3 Genomic architecture of 44 real HCV viral genomes from 1739-bp long fragment of E1E2 region. Length of longest common region shared between any two viral genomes is represented by color.

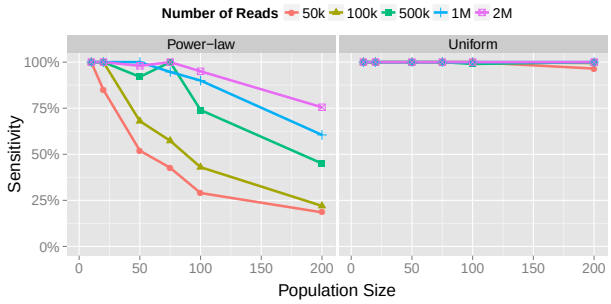

Figure 4.4 Accuracy of population size prediction. Up to 200 viral genomes were generated from the Gag/Pol 3.4 Kb HIV region. The population diversity is 5% - 10%. Variant abundances follow uniform (A) and power-law (B). Highly-accurate 100x2 bp paired-end reads were simulated from HIV population.

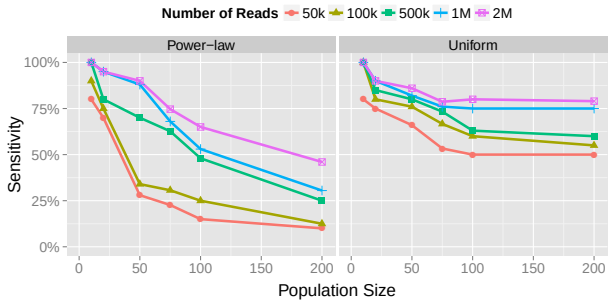

Figure 4.5 Assembly accuracy estimation. Up to 200 viral genomes were generated from the Gag/Pol 3.4 Kb HIV region. The population diversity is 3% - 20%. Variant abundances follow uniform (A) and power-law (B). Consensus error-corrected 2x100bp paired-end reads were simulated from HIV population.

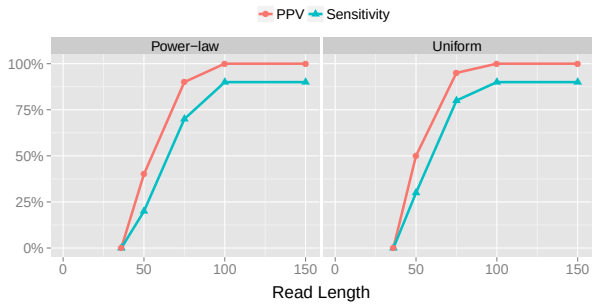

Figure 4.6 Assembly accuracy estimation. Consensus error-corrected paired-end reads of various lengths were simulated from a mixture of 10 real viral clones from 1.3kb-long HIV-1 region. Assembly accuracy as measured by sensitivity(A) and precision(B). Results are for 50,000 reads, no improvement was observed when increasing the of number of reads.

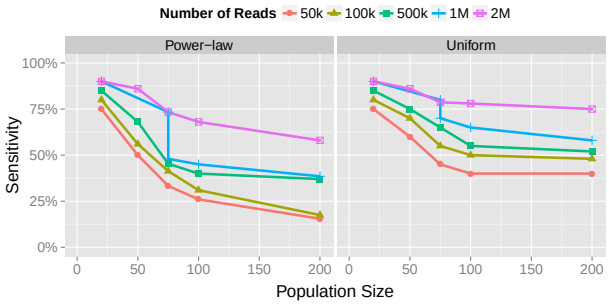

Figure 4.7 Assembly accuracy estimation. Up to 200 recombinant viral genomes were generated from the from 1.3kb-long HIV-1 region. Variant abundances follow power-law and uniform. Consensus error-corrected 2x100bp paired-end reads were simulated from HIV population.

PART 5

DISCUSSION AND FUTURE WORK

The ability of viral quasispecies to rapidly adapt to environmental changes in the form of mutation and recombination facilitates the overall survivability. With the advent of next-generation sequencing, viral populations can be analyzed directly. Accurately estimating the population structure from the sequenced reads is a difficult problem. Unquestionably, as sequencing technologies create ever-increasing amount of data, reconstruction and assembly algorithms will need to scale to the potentially massive population size. Clearly, applying these results to deeper biological studies such as evolution or fitness landscapes is the next step.

REFERENCES

[1] D. I. Lou, J. A. Hussmann, R. M. McBee, A. Acevedo, R. Andino, W. H. Press, and S. L. Sawyer, "High-throughput dna sequencing errors are reduced by orders of magnitude using circle sequencing," *Proceedings of the National Academy of Sciences*, vol. 110, no. 49, pp. 19 872–19 877, 2013.

[2] O. Zagordi, A. Bhattacharya, N. Eriksson, and N. Beerenwinkel, "Shorah: estimating the genetic diversity of a mixed sample from next-generation sequencing data," *BMC Bioinformatics*, vol. 12, no. 1, p. 119, 2011. [Online]. Available: http://www.biomedcentral.com/1471-2105/12/119

[3] O. Zagordi, L. Geyrhofer, V. Roth, and N. Beerenwinkel, "Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction," *Journal of computational biology*, vol. 17, no. 3, pp. 417–428, 2010.

[4] N. Eriksson, L. Pachter, Y. Mitsuya, S.-Y. Rhee, C. Wang, B. Gharizadeh, M. Ronaghi, R. W. Shafer, and N. Beerenwinkel, "Viral population estimation using pyrosequencing," *PLoS Computational Biology*, vol. 4, no. 5, p. e1000074, 2008.

[5] I. Astrovskaya, B. Tork, S. Mangul, K. Westbrooks, I. Măndoiu, P. Balfe, and A. Zelikovsky, "Inferring viral quasispecies spectra from 454 pyrosequencing reads," *BMC Bioinformatics*, vol. 12, no. Suppl 6, p. S1, 2011. [Online]. Available: http://www.biomedcentral.com/1471-2105/12/S6/S1

[6] M. C. F. Prosperi and M. Salemi, "Qure: software for viral quasispecies reconstruction from next-generation sequencing data," *Bioinformatics*, vol. 28, no. 1, pp. 132–133, 2012. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/28/1/132.abstract

[7] Prosperi MC, Prosperi L, Bruselles A, Abbate I, Rozera G, Vincenti D, Solmone MC,

Capobianchi MR, Ulivi G, "Combinatorial Analysis and Algorithms for Quasispecies Reconstruction using Next-Generation Sequencing," *BMC Bioinformatics*, vol. 12, 2011.

[8] C. Wang, Y. Mitsuya, B. Gharizadeh, M. Ronaghi, and R. W. Shafer, "Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance," *Genome research*, vol. 17, no. 8, pp. 1195–1201, 2007.

[9] Armin Töpfer, Osvaldo Zagordi, Sandhya Prabhakaran, Volker Roth, Eran Halperin, and Niko Beerenwinkel., "Probabilistic inference of viral quasispecies subject to recombination," in *Research in Computational Molecular Biology*, ser. Lecture Notes in Computer Science, B. Chor, Ed. Springer Berlin Heidelberg, 2012, vol. 7262, pp. 342–354. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-29627-7_36

[10] M. Garey and D. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness.* W.H. Freeman and Co., 1979.

[11] Graham R.,Grötschel M., Lovasz L, *Handbook of combinatorics.* The MIT Press, January 1996 1996.

[12] E. Halperin and R. M. Karp, "The minimum-entropy set cover problem," *Theoretical computer science*, vol. 348, no. 2, pp. 240–250, 2005.

[13] T. von Hahn, J. Yoon, H. Alter, C. Rice, B. Rehermann, P. Balfe, and J. McKeating, "Hepatitis c virus continuously escapes from neutralizing antibody and t-cell responses during chronic infection in vivo," *Gastroenterology*, vol. 132, pp. 667–678, 2007.

[14] M. C. Prosperi and M. Salemi, "QuRe: software for viral quasispecies reconstruction from next-generation sequencing data," *Bioinformatics*, vol. 28, no. 1, pp. 132–133, 2012.

[15] A. Töpfer, "Indelfixer." [Online]. Available: http://www.bsse.ethz.ch/cbg/software/InDelFixer

[16] Alexander Artyomenko, Nicholas Mancuso, Pavel Skums, Ion Măndoiu, Alex Zelikovsky, "kGEM: An Expectation Maximization Error Correction Algorithm for Next Generation Sequencing of Amplicon-based Data," in *Proc. International Symposium Bioinformatics Research and Applications*, 2013.

[17] N. Mancuso, B. Tork, P. Skums, L. Ganova-Raeva, I. Măndoiu, and A. Zelikovsky, "Reconstructing viral quasispecies from ngs amplicon reads." *In Silico Biology*, vol. 11, no. 5-6, pp. 237 – 249, 2011.

[18] P. Skums, N. Mancuso, A. Artyomenko, B. Tork, I. Măndoiu, Y. Khudyakov, and A. Zelikovsky, "Reconstruction of viral population structure from next-generation sequencing data using multicommodity flows," *BMC Bioinformatics*, vol. 14, no. Suppl 9, p. S2, 2013. [Online]. Available: http://www.biomedcentral.com/1471-2105/14/S9/S2

[19] F. E. Angly, D. Willner, F. Rohwer, P. Hugenholtz, and G. W. Tyson, "Grinder: a versatile amplicon and shotgun sequence simulator," *Nucleic Acids Research*, vol. 40, no. 12, p. e94, 2012. [Online]. Available: http://nar.oxfordjournals.org/content/40/12/e94.abstract

[20] I. Kinde, J. Wu, N. Papadopoulos, K. W. Kinzler, and B. Vogelstein, "Detection and quantification of rare mutations with massively parallel sequencing," *Proceedings of the National Academy of Sciences*, vol. 108, no. 23, pp. 9530–9535, 2011.

[21] J. Liu, M. D. Miller, R. M. Danovich, N. Vandergrift, F. Cai, C. B. Hicks, D. J. Hazuda, and F. Gao, "Analysis of low-frequency mutations associated with drug resistance to raltegravir before antiretroviral treatment," *Antimicrobial agents and chemotherapy*, vol. 55, no. 3, pp. 1114–1119, 2011.

[22] S. Palmer, V. Boltz, F. Maldarelli, M. Kearney, E. K. Halvas, D. Rock, J. Falloon, R. T. Davey Jr, R. L. Dewar, J. A. Metcalf *et al.*, "Selection and persistence of non-nucleoside

reverse transcriptase inhibitor-resistant HIV-1 in patients starting and stopping non-nucleoside therapy," *Aids*, vol. 20, no. 5, pp. 701–710, 2006.

[23] X. Yang, P. Charlebois, S. Gnerre, M. G. Coole, N. J. Lennon, J. Z. Levin, J. Qu, E. M. Ryan, M. C. Zody, and M. R. Henn, "De novo assembly of highly diverse viral populations," *BMC genomics*, vol. 13, no. 1, p. 475, 2012.

[24] S. Gnerre, I. MacCallum, D. Przybylski, F. J. Ribeiro, J. N. Burton, B. J. Walker, T. Sharpe, G. Hall, T. P. Shea, S. Sykes *et al.*, "High-quality draft assemblies of mammalian genomes from massively parallel sequence data," *Proceedings of the National Academy of Sciences*, vol. 108, no. 4, pp. 1513–1518, 2011.

[25] D. R. Zerbino and E. Birney, "Velvet: algorithms for de novo short read assembly using de Bruijn graphs," *Genome research*, vol. 18, no. 5, pp. 821–829, 2008.

[26] R. Luo, B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan, Y. Liu *et al.*, "SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler," *GigaScience*, vol. 1, no. 1, p. 18, 2012.

[27] W.-Y. Y. Yang, F. Hormozdiari, Z. Wang, D. He, B. Pasaniuc, and E. Eskin, "Leveraging Multi-SNP Reads from Sequencing Data for Haplotype Inference." *Bioinformatics*, 2013. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btt386

[28] V. Bansal and V. Bafna, "HapCUT: an efficient and accurate algorithm for the haplotype assembly problem," *Bioinformatics*, vol. 24, no. 16, pp. i153–i159, 2008.

[29] A. Huang, R. Kantor, A. DeLong, L. Schreier, and S. Istrail, "QColors: An algorithm for conservative viral quasispecies reconstruction from short and non-contiguous next generation sequencing reads," *In silico biology*, vol. 11, no. 5, pp. 193–201, 2011.

[30] M. Kubale, *Graph colorings*. American Mathematical Soc., 2004, vol. 352.

[31] D. S. Johnson and M. A. Trick, *Cliques, coloring, and satisfiability: second DIMACS implementation challenge, October 11-13, 1993*. American Mathematical Soc., 1996, vol. 26.

[32] J. Duitama, G. K. McEwen, T. Huebsch, S. Palczewski, S. Schulz, K. Verstrepen, E.-K. Suk, and M. R. Hoehe, "Fosmid-based whole genome haplotyping of a hapmap trio child: evaluation of single individual haplotyping techniques," *Nucleic acids research*, vol. 40, no. 5, pp. 2041–2053, 2012.

[33] M. Mitzenmacher and E. Upfal, *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge University Press, 2005.

[34] N. Eriksson, L. Pachter, Y. Mitsuya, S.-Y. Rhee, C. Wang, B. Gharizadeh, M. Ronaghi, R. W. Shafer, and N. Beerenwinkel, "Viral Population Estimation Using Pyrosequencing," *PLoS Comput Biol*, vol. 4, no. 5, p. e1000074, 05 2008. [Online]. Available: http://dx.plos.org/10.1371%2Fjournal.pcbi.1000074

[35] O. Zagordi, L. Geyrhofer, V. Roth, and N. Beerenwinkel, "Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction," *Journal of computational biology*, vol. 17, no. 3, pp. 417–428, 2010.

[36] T. Ndungu and R. A. Weiss, "On HIV diversity," *AIDS*, vol. 26, no. 10, pp. 1255–1260, 2012.

[37] L. P. Martins, N. Chenciner, and S. Wain-hobson, "Complex intrapatient sequence variation in the V1 and V2 hypervariable regions of the HIV-1 gp120 envelope sequence," *Virology*, vol. 191, no. 2, pp. 837–845, 1992.

[38] F. K. Yoshimura, K. Diem, G. H. Learn, S. Riddell, and L. Corey, "Intrapatient sequence variation of the gag gene of human immunodeficiency virus type 1 plasma virions." *Journal of virology*, vol. 70, no. 12, pp. 8879–8887, 1996.

[39] R. A. Neher and T. Leitner, "Recombination rate and selection strength in hiv intra-patient evolution," *PLoS computational biology*, vol. 6, no. 1, p. e1000660, 2010.

[40] O. Zagordi, A. Bhattacharya, N. Eriksson, and N. Beerenwinkel, "ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data," *BMC bioinformatics*, vol. 12, no. 1, p. 119, 2011.