8-8-2017

Analysis of NGS Data from Immune Response and Viral Samples

Ekaterina Gerasimov

ANALYSIS OF NGS DATA FROM IMMUNE RESPONSE AND VIRAL SAMPLES

by

EKATERINA GERASIMOV

Under the Direction of Alexander Zelikovsky, PhD

ABSTRACT

This thesis is devoted to designing and applying advanced algorithmical and statistical tools for analysis of NGS data related to cancer and infection diseases. NGS data under investigation are obtained either from host samples or viral variants. Recently, random peptide phage display libraries (RPPDL) were applied to studies of host's antibody response to different diseases. We study human antibody response to breast cancer and mouse antibody response to Lyme disease by sequencing of the whole antibody repertoire profiles which are represented by RPPDL. Alternatively, instead of sequencing immune response NGS can be applied directly to a viral population within an infected host. Specifically, we analyze the following RNA viruses: the human immunodeficiency virus (HIV) and the infectious bronchitis virus (IBV). Sequencing of RNA viruses is challenging because there are many variants inside population due to high mutation rate.

Our results show that NGS helps to understand RNA viruses and explore their interaction with infected hosts. NGS also helps to analyze immune response to different diseases, trace changing of immune response at different disease stages.

INDEX WORDS: Next generation sequencing, RNA virus, Viral population reconstruction, Error correction of NGS reads, Assembly of NGS reads, Random peptide phage display libraries, Breast cancer, Lyme disease

ANALYSIS OF NGS DATA FROM IMMUNE RESPONSE AND VIRAL SAMPLES

by

EKATERINA GERASIMOV

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

2017

ANALYSIS OF NGS DATA FROM IMMUNE RESPONSE AND VIRAL SAMPLES

by

EKATERINA GERASIMOV

Committee Chair: Alexander Zelikovsky

Committee: Robert Harrison
Pavel Skums
Yurij Ionov
Artem Rogovskyy

Electronic Version Approved:

Office of Graduate Studies
College of Arts and Sciences
Georgia State University
August 2017

DEDICATION

To my husband Sergey and son Misha, for their support, patience and love.

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Alex Zelikovsky for all of the opportunities I was given to conduct my research and further my dissertation at GSU. Without your guidance, enthusiasm and persistent help this dissertation would not have been possible. It has been a period of intense learning for me, not only in the scientific arena, but also on a personal level.

I would like to thank Dr. Yurij Ionov and Dr. Artem Rogovskyy for the opportunity to participate in their research and cooperate with them. It was a great experience for me. Special thanks to the rest of my committee members Dr. Robert Harrison and Dr. Pavel Skums for helping to guide this research and provide exceptional feedback. I am grateful for all the help that Dr. Raj Sunderraman has shown me over the years. Special thanks to the GSU Molecular Basis of Disease Fellowship for financial support.

Special thanks to my colleagues Ludmila Perelygina and Min-hsin Chen from my internship at CDC for their wonderful collaboration. You supported me greatly and were always willing to help me. I would particularly like to single out my supervisor Ludmila Perelygina, I want to thank you for your excellent cooperation and for all of the opportunities I was given.

Thanks to all of my friends in the department, especially those who collaborated with me on projects. We were not only able to support each other by deliberating over our problems and findings, but also happily by talking about things other than just our papers. Igor Mandric, Sasha Artyomenko, Sergey Knyazev, Andrii Melnyk, Pelin Icer, Olga Glebova, Seth Sims... I will miss you, guys. Special thanks to the former GSU students Serghei Mangul and Bassam Tork for the opportunity to help them with their researches.

Finally, special recognition goes out to my family for their support throughout my life. My parents and siblings, my husband and his parents have always be the first to offer help.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

- NGS - Next Generation Sequencing
- RPPDL - Random Peptide Phage Display Libraries
- safe-SeqS - Safe-Sequencing System
- CAST - Clustering Affinity Search Technique
- CMIM - CAST-based Motif Identification Method
- PPV - Predictive Positive Value

PART 1

INTRODUCTION

Next generation sequencing (NGS) refers to such modern high-throughput sequencing technologies as Illumina sequencing, Roche 454 sequencing, Ion torrent and SOLiD sequencing. Technological revolution allowed to sequence DNA and RNA much more quicker and cheaper than previously used Sanger sequencing. Millions of reads now can be processed in parallel. Table 1.1 provides specifications of different sequencing technologies. NGS also needs significantly less DNA and is more reliable than Sanger sequencing. However, NGS reads are much shorter which requires assembly the vast amount of sequences that are generated. They also prone to high error rates and need additional error correction.

This work is concentrated on analysis of NGS data related to cancer and infection diseases. In this case NGS data represent either antibody response to a disease or a viral population within the infected host. Both types of the data have their own drawbacks and require the development of special methods for analysis.

Table 1.1 Comparison of different sequencing technologies.

manufacturer	avg read length	avg read count	avg error rate	paired reads
Sanger	1000bp	96	0.001%	No
Roche/454	450bp	1M	1.0%	Yes
Illumina	150bp	500M	0.1%	Yes
Ion Torrent	300bp	6M	1.0%	Yes
SOLiD	50bp	1400M	0.1%	Yes

1.1 Analysis of viral NGS data

RNA viruses, in which the genetic information is stored in the form of RNA, are the most abundant group of subcellular parasites infecting animals, plants, and bacteria.

They cause such severe diseases as hepatitis, viral encephalitis, yellow fever, and AIDS. Distinctive features of RNA virus replication include high mutation rates, high yields, and short replication times [2]. As a result, RNA viruses replicate as a dynamic and complex mutant group, called viral quasispecies.

NGS presents a novel opportunity for understanding RNA viruses, particularly, their evolution, drug resistance and immune escape. It allows to analyze a great number of viral variants from infected patients offering a deep coverage of genomic data [3]. Even low frequency variants may have a great interest, for example, affect virulence [4]. Although NGS provides cost-effective access to high-throughput data, inferring the viral genetic diversity of a mixed sample from sequencing data is still a challenging task. The reasons lie in difficulties of sample preparation and sequencing errors, short read length and, in particular, an incomplete a priori knowledge of existing viruses and their diversity [5]. Estimating viral genetic diversity and reconstructing the individual haplotype sequences relies on both error correction and read assembly [5].

1.1.1 Error correction of viral NGS data

NGS provides reads with very deep coverage. However, the relatively high error rate limits the ability to analyze population diversity directly. The error correction problem involves identifying and correcting errors in reads introduced during sequencing. The main purpose of error correction of viral NGS data is to discriminate between artifacts and actual sequences. This task is complicated by the fact that quasispecies contains close variants and some of them may have low frequency.

Computer-assisted studies of viruses is a neglected area of research. The attention of bioinformaticians to this challenging field is far from sufficient. There is a lack of error correction methods for viral NGS data. In fact, there are only a couple of existing tools - KEC [6] and ShoRAH [7].

KEC stands for K-mer-based error correction. This algorithm is for viral amplicons based on k-mers. It starts from calculating of k-mers and their frequencies. Then it de-

termines the count threshold which distinguish "correct" or high count k-mers from low count k-mers contained errors. Finally, it finds error regions of the reads where all k-mers have low count and corrects the errors in those regions [6].

ShoRAH stands for Short Reads Assembly into Haplotypes. It is a genome assembly tool which include also its own error correction. It corrects sequencing errors by clustering all reads that overlap the same region of the genome of length approximately equal to the read length. The consensus sequence of each cluster represents the original haplotype from which the erroneous reads are obtained [7].

1.1.2 Assembly of viral NGS data

NGS does not directly sequence viral genomes describing a viral population. Instead, it produces relatively short reads that should be assembled into population variants. The problem formulation is following: given a collection of NGS reads from a viral sample, variants and their relative frequencies inside the sample need to be reconstructed. The problem is challenging because variants are similar to each other, there are overlapped regions. The decision should be made which read belongs to which variant. Due to the great diversity and the high mutation rates, there is a lack of a suitable reference genome which is a major hindrance for many viral quasispecies assembly approaches. A brief description of some popular viral quasispecies assembly tools are given below.

The first publicly available tool for reconstructing a viral quasispecies was ShoRAH [7]. The analysis with ShoRAH includes four major steps: alignment; 2) error correction; 3) haplotype reconstruction and 4) frequency estimation. The reconstruction method employs a parsimony principle and computes a minimal set of haplotypes that explains all reads in the dataset. Once the putative variants have been assembled, their relative abundances are computed using an Expectation-Maximization (EM) algorithm. ShoRAH can be run on NGS shotgun data or single amplicon data [7].

Another assembly tool is the Viral Spectrum Assembler, or ViSpA [8]. It reconstructs quasispecies by constructing a weighted read-overlap graph. Similarly to ShoRAH,

ViSpA utilizes an EM algorithm for estimating variant frequencies. Originally it was designed for NGS shotgun data, but in theory it could be applied to amplicon-based data as well.

QuRe [9] is a tool for viral Quasispecies Reconstruction, specifically developed to analyze long read (>100 bp) NGS data such as Roche 454 Life Sciences. It performs alignments of sequence fragments against a reference genome, finds an optimal division of the genome into sliding windows based on coverage and diversity and attempts to reconstruct all the individual sequences of the viral quasispecies-along with their prevalence-using a heuristic algorithm, which matches multinomial distributions of distinct viral variants overlapping across the genome division [9].

QuasiRecomb [4] is another assembly tool. It uses a hidden Markov model (HMM) to generate viral populations, i.e., haplotype distributions, and their probing by means of NGS. In this model, the haplotypes are originating from a small number of generating sequences via recombination, described as switch of state in the HMM that selects from which sequence the haplotype derives, and mutation, described by position-specific probability tables for the generating sequences. It is suitable for amplicon and shotgun sequencing projects [4].

SAVAGE [10] is a new computational tool for reconstructing individual haplotypes of intrahost virus strains without the need for a high-quality reference genome. It constructs an overlap graph directly from the patient sample reads. It is the first approach for de novo assembly of viral quasispecies based on overlap graphs.

1.2 Phage display data processing for immune response analysis

Instead of direct sequencing of viral population within an infected host, a disease can be explored by sequencing of the host's immune response. The immune system protects the body from possibly harmful substances by recognizing and responding to antigens. Antigens are substances (usually proteins) on the surface of cells, viruses, fungi, or bacteria. Serum antibodies are valuable source of information on the health state of an organ-

ism [11]. Recently, random peptide phage display libraries (RPPDL) together with NGS were applied to studies of antibody response to different diseases.

1.2.1 Phage display

Phage display is a laboratory technique for the study of protein-protein, protein-peptide, and protein-DNA interactions. It uses viruses that infect bacteria (bacteriophage or, simply, phage) to connect proteins with the genetic information encoded them. Phage display was proposed by George P. Smith in 1985 [12]. The genome of a bacteriophage contains the gene that encodes the phage coat protein. In this technique, a gene encoding a protein (or peptide) of interest is inserted into a virus coat protein gene, causing the phage to "display" the protein on its outside. The pool of phages with different proteins of interest forms a library. A library can be screened against other proteins, peptides or DNA sequences, in order to detect interaction between the displayed protein and those other molecules. The process is called *in vitro* screening. The phage display library is mixed with the desired target DNA sequences or proteins. After some time allowing the phage to bind the mix is washed. All phage-displaying proteins that are not interacted are washed away. Attached phage may be eluted and amplified in bacterial hosts. The procedure can be repeated several times. Then the DNA withing the phage is sequenced using NGS.

Since an antibody recognizes not the whole antigen but 4-7 critical amino acids within the antigenic determinant (epitope), the whole serum antibody repertoire profile can be represented by RPPDL (Figure 1.1). In RPPDL, a DNA sequence encoding some random peptide is inserted into a phage coat gene which leads to displaying of this peptide on the phage outside. Ph.D.™-7 Phage Display Peptide Library Kit is an example of such library. It is a library of almost all possible 7-mer peptides and commercially available from New England Biolabs (NEB). After mixing the library with a blood serum, DNA withing the phage interacted with antibody is sequenced and translated into peptide. Obtained peptides represent a mimotope profile of a serum sample - mimic of all

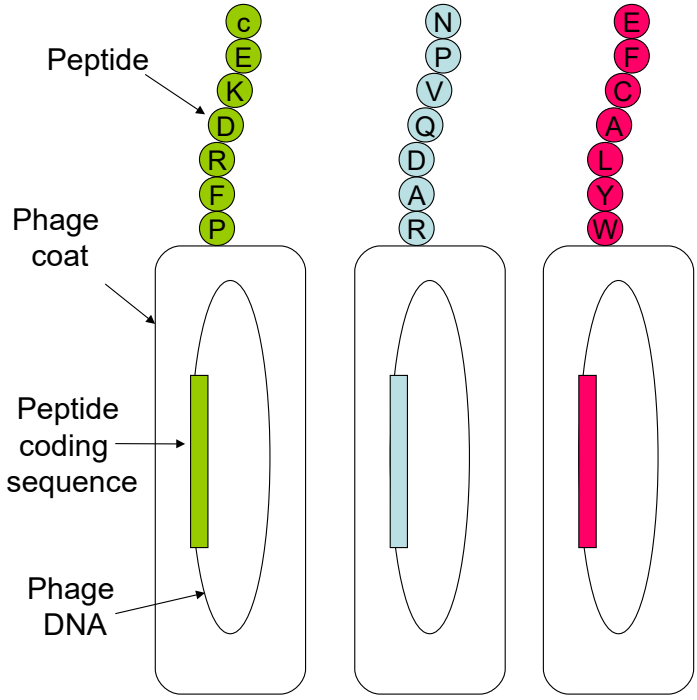

Figure 1.1 Schematic representation of random peptide phage display library

epitopes recognized by all antibodies contained in the serum. Mimotope profiles of different samples can be studied and compared with each other.

RPPDL together with the NGS technology make possible to identify all the epitopes recognized by all antibodies contained in the blood serum using one run of the sequencing machine. However, this approach has one drawback. The obtained peptide profiles can consist of millions of sequences. A significant fraction of these sequences is not related to the repertoires of antibody specificities, but produced by nonspecific binding and preferential amplification in bacteria. The presence of these ”parasitic” sequences produces a lot of noise and complicates the analysis of immune response. Considering that the affinity selected sequences can be clustered into the groups of similar sequences with shared consensus motifs, while the parasitic sequences are usually represented by single

copies. Motifs are stable to parasites.

1.2.2 Mimotope motifs

A group of peptides with a common sequence pattern forms a motif. An alignment of peptides forming a motif contains conserved positions. A motif can be represented as a position weight matrix which summarizes the amino acid preferences for each column of the alignment. Each value in this matrix corresponds to frequency of an amino acid in current position. Alternatively, in [13] it was proposed to represent a given alignment by a log-odds weight matrix. A log-odds weight matrix is calculated as $log_2(p_{A,j}/q_A)$, where $p_{A,j}$ is the frequency of amino acid A at position j, and q_A is the background frequency of A. The quality of a motif can be measured by information content (IC) or Kullback-Leibler divergence (KLD). The IC says how different a given position weight matrix is from a uniform distribution. IC can be calculated as $IC = log_2 20 + \sum_{A,j} p_{A,j} * log_2(p_{A,j})$. Often, it is more useful to calculate the IC with the background letter frequencies of the sequences you are studying rather than assuming equal probabilities of each amino acid. Thus, IC corresponds to the KLD or relative entropy. The equation in this case becomes: $KLD = -\sum_{A,j} p_{A,j} * log_2(p_{A,j}/q_A)$. The KLD allows measuring the information gain of an observed amino acid distribution compared with a background distribution (the frequency of each amino acid in random protein sequences).

One of the examples of mimotope motifs' application is the early detection of cancer. Circulating auto-antibodies produced by the patient's own immune system after exposure to cancer proteins are promising bio-markers for the early detection of cancer. In this case, motifs represent cancer bio-markers [14]. Another example is finding epitopes of Borrelia burgdorferi bacteria [15]. This bacteria causes Lyme disease - the most prevalent tick-borne illness in North America and Europe. Lyme disease is problematic because early diagnosis is easily missed due to flu-like symptoms, which only transiently appear in humans during an early stage of disease [16]. When missed and, therefore left untreated, Lyme disease becomes chronic, presenting itself as skin lesions, arthritis, cardi-

tis, and occasionally followed by nervous system involvement [17]. No preventable or therapeutic vaccine is currently available. In this case, mimotope motifs correspond to epitopes.

1.2.3 Finding mimotope motifs

The are two main approaches for finding mimotope motifs. First one is based on clustering, second is based on BLAST alignment.

If we consider a motif as a cluster formed by peptides with the center represented by a consensus sequence then construction of a motif corresponds to a difficult clustering problem with many closely located centers. The radius of a cluster may exceed the distance from one cluster to another one. Recently, several software tools based on clustering were proposed for finding mimotope motifs. MUltiple Specificity Identifier (MUSI) [18] relies on mixture model optimization. The algorithm behind MUSI is based on the idea of fitting several linear probabilistic models (in this case, position weight matrices) to a set of sequences to optimally describe the different specificity classes. It uses a Maximum Likelihood approach with Dirichlet priors for fitting. GibbsCluster is another tool proposed in [13] and based on Gibbs sampling. Both tools try to be versatile and allow for peptide data from any biological source to be processed, but may require some prior data knowledge, such as the number of clusters to identify. They perform well on smaller datasets of up to thousands of sequences, but not million of sequences. Finally, Hammock [19] is a hidden Markov model-based peptide clustering algorithm. Hammock uses profile HMMs for presise computational representation of sequence motifs and is based on the idea of progressive cluster growth. Unlike two other tools, it is able to process very large datasets, to identify multiple distinct motifs within one dataset and no prior data knowledge is required.

A method for finding mimotope motifs based on protein-protein BLAST alignment was proposed in [11]. The idea of the method is following. BLAST aligns sample peptides to interesting proteins (for example, all proteins of the organism). The goal is to

retrieve proteins that have multiple matches to different peptides. The probability for a protein to have multiple matches to different peptides due to a chance is proportional to the length of the protein [11]. So, each protein receives a score equaled to a number of mapped peptides divided by the length of the protein. The proteins with the high scores are considered as antigen candidates and positions on those proteins where the peptides are mapped are epitope candidates. A group of peptides aligned to the same location on a protein forms a mimotope motif.

1.3 RNA-seq data processing for immune response analysis

RNA-Seq is an approach to transcriptome profiling using NGS technologies. The transcriptome is the complete set of transcripts in a cell, and their quantity, for a specific developmental stage or physiological condition. RNA-Seq is the first sequencing-based method that allows the entire transcriptome to be surveyed in a very high-throughput and quantitative manner [20]. Depending on whether there is a disease in an organism or not, the transcriptional activity of genes involved in immune response is different. One particularly powerful advantage of RNA-Seq is that it can capture this difference. Although, like other high-throughput sequencing technologies, RNA-Seq requires development of special methods for data analysis.

Many variations of RNA-seq protocols and analyses have been published, but there is still no optimal pipeline. In general, for a typical RNA-seq experiment, the major analysis steps are quality control, read alignment with and without a reference genome, obtaining metrics for gene and transcript expression, and approaches for detecting differential gene expression [21]. Quality control for the raw reads involves the analysis of sequence quality, GC content, the presence of adaptors, overrepresented k-mers and duplicated reads in order to detect sequencing errors, PCR artifacts or contaminations. Next, the analysis involves the mapping of the reads onto the reference genome or transcriptome to infer which transcripts are expressed. In case the organism does not have a sequenced genome, the analysis first assembles reads into longer contigs and then treats these contigs as the

expressed transcriptome to which reads are mapped back again for quantification. Then the estimation of gene and transcript expression is done usually based on the number of reads that map to each transcript sequence. Finally, the differential expression analysis compares gene expression values among samples. Recent independent comparison studies have demonstrated that the choice of the method (or even the version of a software package) can markedly affect the outcome of the analysis and that no single method is likely to perform favorably for all datasets [21].

1.4 Contributions

The main contributions are:

- The-state-of-the-art error correction programs were compared on HIV NGS data.
- The reconstructed IBV quasispecies were validated by comparing them with IBV variants sequenced by Sanger.
- A novel method for motif identification in mimotope profile (CMIM) based on CAST clustering [22]. The statistical analysis is performed to distinguish between cancer patients and controls and find cancer specific motifs.
- A novel method for motif identification in mimotope profile based on k-means clustering.
- Application of the BLAST alignment method for finding motifs for epitopes to Lyme disease data. The antibody response to known antigens and VlsE protein responsible for persistence of the infection between days 28 and 70 of the disease has been compared.
- Application of RNA-seq data analysis tools for measuring human immune response to wild and vaccine rubella virus.

1.5 Future work

We plan to apply pathways analysis to RNA-seq rubella data.

1.6 Roadmap

The rest of the dissertation is organized as follows. Chapter 2 presents analysis of viral NGS data. First, the-state-of-the-art error correction methods are benchmarked on HIV NGS data. Then different methods for viral quasispecies reconstruction are compared. The reconstructed IBV quasispecies were validated by comparing them with IBV variants sequenced by Sanger. In Chapter 3 we analyze NGS data representing immune response to breast cancer and Lyme disease though mimotope profiles. In the first part, two novel methods are proposed for identification of motifs in mimotope profiles based on CAST and k-means clustering. In the second part, motif finding method based on BLAST alignment was applied to detect any significant changes in antibody repertoires on early and late stages of Lyme disease. In chapter 4 we applied RNA-seq data analysis pipeline for for measuring human immune response to wild and vaccine rubella virus. Our plans for future work are discussed in Chapter 5.

1.7 Scientific products

1.7.1 Publications

Book Chapters

1. S. Mangul, N. Wu, **E. Nenastyeva**, N. Mancuso, A. Zelikovsky, R. Sun, E. Eskin "Applications of High-Fidelity Sequencing Protocol to RNA Viruses" Computational Methods for Next Generation Sequencing Data Analysis (2016), John Wiley & Sons, Inc., Hoboken, NJ, USA. doi: 10.1002/9781119272182.ch4.

2. B. Tork, **E.Nenastyeva**, A. Artyomenko, N. Mancuso, M. I. Khan, R. O'Neill, I. Mandoiu and A. Zelikovsky "Reconstruction of Infectious Bronchitis Virus Qua-

sispecies from NGS Data" Computational Methods for Next Generation Sequencing Data Analysis (2016), John Wiley & Sons, Inc., Hoboken, NJ, USA. doi: 10.1002/9781119272182.ch17

Refereed Journal Papers

1. **E. Gerasimov**, A. Zelikovsky, I.I. Mandoiu, Y. Ionov "Identification of cancer-specific motifs in mimotope profiles of serum antibody repertoire" BMC bioinformatics, 18(Suppl 8):244, 2017. DOI: 10.1186/s12859-017-1661-5.

2. A. Rogovskyy, D. Gillis, Y. Ionov, **E. Gerasimov** and A. Zelikovsky "Antibody response to Lyme disease spirochetes in the context of VlsE-mediated immune evasion" Infection and Immunity (2016): IAI-00890.

Conference Abstracts

1. **E. Nenastyeva**, A. Zelikovsky, I.I. Mandoiu, Y. Ionov "Identification of cancer-specific motifs in mimotope profiles of serum antibody repertoire" Proceedings of 2015 IEEE 5th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS). DOI: 10.1109/ICCABS.2015.7344724

1.7.2 Presentations

1. **E. Nenastyeva**, Y. Ionov, I. Mandoiu, A. Zelikovsky "Short Abstract: Using Random Peptide Phage Display Libraries for early Breast cancer detection" Oral presentation at International Symposium on Bioinformatics Research and Applications (ISBRA), 2013.

2. **E. Nenastyeva**, A. Zelikovsky, I.I. Mandoiu, Y. Ionov "Identification of cancer-specific motifs in mimotope profiles of serum antibody repertoire" Poster at International Symposium on Bioinformatics Research and Applications (ISBRA), 2015.

3. **E. Gerasimov**, A. Zelikovsky "Motif finding in large peptide data sets" Poster at GSU Molecular Basis of Disease Fellows Retreat, 2017.

PART 2

ANALYSIS OF VIRAL NGS DATA

RNA virus has RNA as its genetic material. Notable examples include HIV, HCV and influenza. It is difficult to make effective vaccines against RNA viruses because they have high mutation rates compared to DNA viruses. Thus, mutation rates vary between 10^{-4} and 10^{-6} per nucleotide [23]. Distinctive features of RNA virus replication include high mutation rates, high yields, and short replication times [2]. As a result, RNA viruses replicate as a dynamic and complex mutant group, called viral quasispecies. A quasispecies refers to a population of genetically related viruses that are closely distributed around a consensus sequence [23].

RNA viruses have high genomic diversity within an infected host. It effects many clinically important phenotypic traits such as escape from vaccine-induced immunity, virulence, and response to antiviral therapies [24]. Sequencing technologies must be sensitive enough in order to accurately characterize an intra-host RNA virus population virus population [25, 26]. NGS technologies offer deep coverage of genomic data in the form of millions of sequencing reads allowing to capture rare variants [27]. But the full picture of viral diversity in a population remains undiscovered due to errors produced by sequencing platforms. The presence of sequencing errors makes it difficult to distinguish between variants and sequencing errors. Additionally, low viral population variability (i.e., pairs of individual viral genomes that have small genetic distance) and the presence of individual variants having low abundance complicates accessing viral diversity and assembling full-length viral variants [28].

Current sequencing technologies use different underlying chemistry and offer trade-offs among throughput, read length and cost [27]. Although the sequencing platforms can potentially detect point-mutations, error rates may result in false positive SNV calls

or wrong genome variant sequences. Current assembly methods [8, 7, 9, 29, 30] are not able to differentiate true biological mutations from sequencing artifacts, thus significantly limiting the possibility of a method to assemble the underlying viral population. Computational error correction approaches are able to partially correct the sequencing error but they are not well suited for mixed viral samples and may lead to filtering out true biological mutations. As the result the low abundant variants remain undiscovered. Additional difficulty is the genomic architecture of viruses. Long conserved regions shared across viral population introduce ambiguity in the assembly process, because they may have multiple cross-overs. In contrast to repeats in genome assembly conserved regions may be phased based on relative abundances of viral variants. Estimating viral genetic diversity and reconstructing the individual haplotype sequences relies on both error correction and read assembly [5].

2.1 Error correction of viral NGS data

2.1.1 Introduction

NGS provides reads with very deep coverage. However, the relatively high error rate limits the ability to analyze population diversity directly. The main purpose of error correction of viral NGS data is to discriminate between artifacts and actual sequences. This task is complicated by the fact that quasispecies contains close variants and some of them may have low frequency.

In this work, different error correction tool were compared. As far as not many research groups focusing on viruses, general methods developed for human sequencing data and metagenomics data were also considered. For evaluation a high-fidelity sequencing protocol was used. It eliminated errors from sequencing data. The protocol, known as the Safe-Sequencing System (”safe-SeqS”), has been proposed in [31] and applied to detect rare somatic mutations, but its application on detecting rare viral mutations has been neglected.

2.1.2 Materials and methods for error correction

We used an Illumina HiSeq HIV paired-end dataset. A special library preparation protocol based on safe-SeqS that eliminates sequencing errors during the de-multiplexing step was applied. In [28] the authors proposed the application of safe-SeqS on detecting rare viral mutations. The protocol involved two steps. The first step was the assignment of a unique barcode sequence to each DNA fragment. The second step was the amplification of each tagged fragment. After sequencing the reads belonging to the original fragment were clustered based on the barcode. Then consensus sequence of each cluster was calculated with condition that the most popular nucleotide at a position had frequency at least 0.9, otherwise the whole barcode cluster was disregard. Thus every sequenced position of the fragment would have multiple independent evidence supporting highly accurate consensus read. We deleted sequence fragments of 13bp represented barcodes from consensus reads and considered the reads as ground truth (further referred to as barcode-corrected reads). We also deleted barcode sequences from the original reads and considered the reads as test data for error correction methods (further referred as original reads).The length of barcode-corrected and original reads after barcode deletion became 87bp.

Computer-assisted studies of viruses is a neglected area of research. The attention of bioinformaticians to this challenging field is far from sufficient. There is a lack of error correction methods for viral NGS data. For this reason together with error correction methods developed for viruses we considered methods developed for human sequencing data and metagenomics data. We ran 5 error correction tools: BLESS V1.02 [32], Fiona V0.2 [33], Pollux V1.0.2 [34], K-mer-based error correction (KEC) [6] and Short Reads Assembly into Haplotypes (ShoRAH) [7]. All tools were downloaded and installed on our server. Brief descriptions of tools with parameters we used are given in this section.

BLESS stands for BLoom-filter-based Error correction Solution for high-throughput Sequencing reads and belongs to the k-mer spectrum-based method. It uses a single minimum-sized Bloom filter [35], and is also able to tolerate a high false-positive rate, al-

lowing to correct errors with a reduce memory usage comparing to other methods. BLESS can extend reads to correct errors at the end of reads as accurately as other parts of the reads. Another feature of BLESS is ability to handle repeats in genome and long k-mers which leads to high accuracy. BLESS accepts single-end or paired-end reads as input. Another parameter is k-mer length. We chose it equal to 20 based on empirical formula provided by the tool's authors.

Fiona developed for different organisms with short or long genomes. It uses an efficient implementation of suffix arrays to detect read overlaps with different seed lengths in parallel. The developers of Fiona claims that it is an accurate error-correction method able to handle substitution, insertion and deletion errors equally well and can be applied to any sequencing technology. It is known that Illumina technology produces mostly substitution errors [36]. We ran a special version of Fiona for Illumina reads that corrects indels but only considers hamming pairwise read distance. In fact, Fiona works only on single-end reads, so we mixed all pared-end reads into one data set. The only parameter should be set was the length of the genome. We used the length of Gag/Pol 3.4 Kb HIV region.

Pollux is a tool that corrects errors introduced by Illumina, Ion Torrent, and Roche 454 sequencing technologies and can be applied to single- or mixed-genome data. The method approaches the problem of error correction using k-mers. During error correction Pollux scans reads and divides them into k-mers of length 31. It counts the number of occurrences of each observed k-mer. Then it scans reads again and generates a k-mer depth profile for each read, using this information to correct the k-mer profile [34]. One of the reason we tested Pollux because it could be applied to mixed genomes. We ran Pollux on our data with default parameters.

KEC is error correction algorithm for viral amplicons based on k-mers. It starts from calculating of k-mers and their frequencies. Then it determines the count threshold which distinguish "correct" or high count k-mers from low count k-mers contained errors. Finally, it finds error regions of the reads where all k-mers have low count and corrects the

errors in those regions [6]. KEC needs k-mer length as a parameter. We decided to use default 25. Running KEC on original reads, we faced some difficulties. First, KEC does not work on paired-end reads, so we mixed all reads together and considered them as single. Second, KEC could not handle all reads. We considered only reads assigned to barcode clusters with the consensus (true read) mapped to Gag/Pol 3.4 Kb HIV region. The reads were sorted according the position of their cluster consensus in Gag/Pol. Then the reads were divided into groups of approximately 5 million reads. KEC was run on each of those groups separately.

ShoRAH was developed to correct sequencing errors and to estimate the population structure of a heterogeneous sample. It corrects sequencing errors by clustering all reads that overlap the same region of the genome of length approximately equal to the read length. The consensus sequence of each cluster represents the original haplotype from which the erroneous reads are obtained /citeshorah. ShoRAH should be run on aligned reads. So, we used the result of BWA-MEM mapping of original reads on Gag/Pol 3.4 Kb HIV region with default parameters.

2.1.3 Results

Datasets

We benchmark the performance of error correction tools on a 3.4 kb region of HIV spanning the Gag/Pol genes. The total number of original paired-end reads was 106,667,644. The original reads having identical 13 bp barcodes were grouped together in so-called barcode clusters. As a result of safe-SeqS protocol 3,052,919 clusters left others were disregarded. A consensus sequence was identified for each cluster and considered as a ground true read. In Figure 2.2 the size of a barcode cluster shows the number of original reads assigned to it. About 52% of all clusters consisted of 2 reads; about 31% of clusters consisted of 3, 4 or 5 reads; 4% - 6..10 reads; 5.5% - 11..20 reads; 4% - 21..30 reads; 1% - 31..40 reads and others were less than 0.2%. Thus, 0.12% of clusters consisted of more than 1000 reads and 0.02% consisted of more than 2000 reads. Maximum barcode

size was 3178 reads.

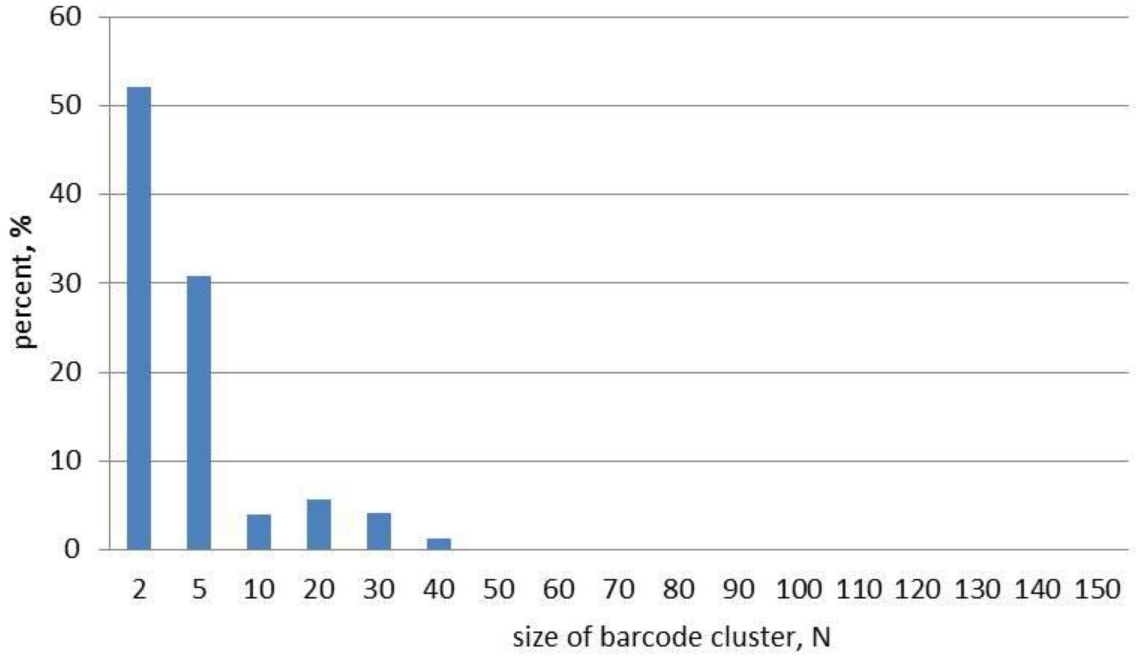

Figure 2.1 Percentage of barcode clusters with size N.

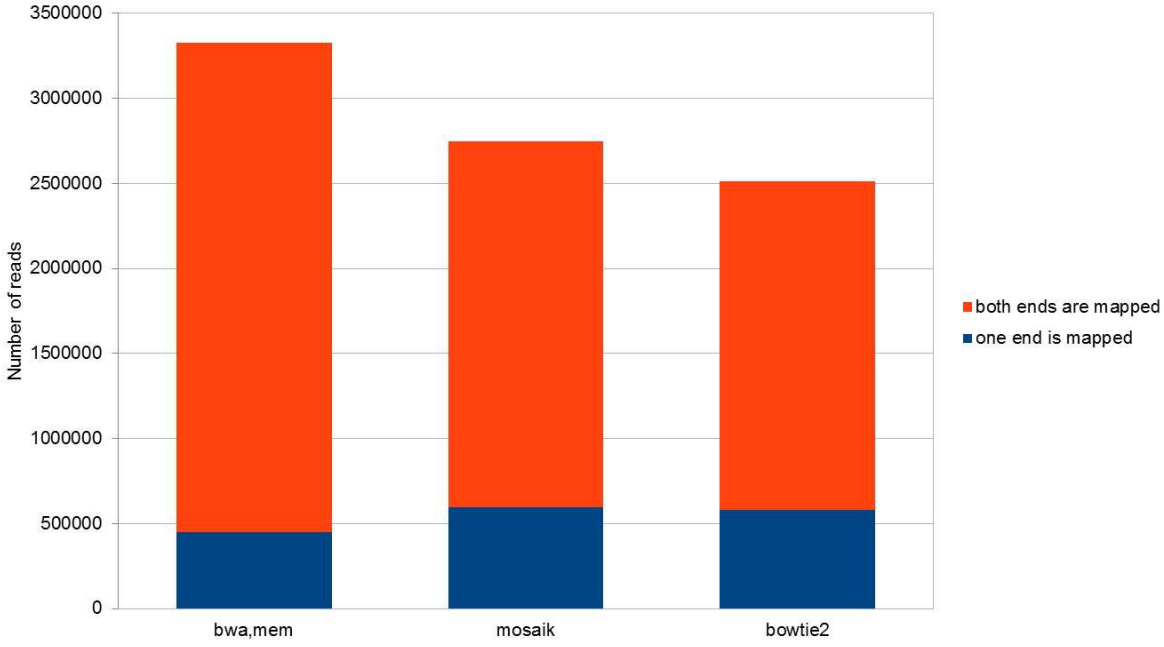

Figure 2.2 The number of mapped barcode-corrected reads.

All 6,105,838 forward and reverse parts of barcode-corrected reads (or 3,052,919 pairs) were mapped to Gag/Pol 3.4 Kb HIV region using BWA-MEM with default parameters, Mosaik with default parameters and Bowtie2 with the following parameters: very-sensitive, -N 1, -a (see Figure 2.2). BWA-MEM mapped 3,327,717 reads (54.5%), among them 2,880,818 reads where both ends were mapped (or 1,440,409 pairs), 446,899 reads where just one end was mapped. Mosaik mapped 2,748,821 reads (45%), among them ,2155,756 reads where both ends were mapped (or 1,077,878 pairs), 593,065 reads where just one end was mapped. Bowtie2 mapped 2,510,061 reads (41%), among them 1,929,344 reads where both ends were mapped (or 964,672 pairs); 580,717 reads where just one end was mapped.

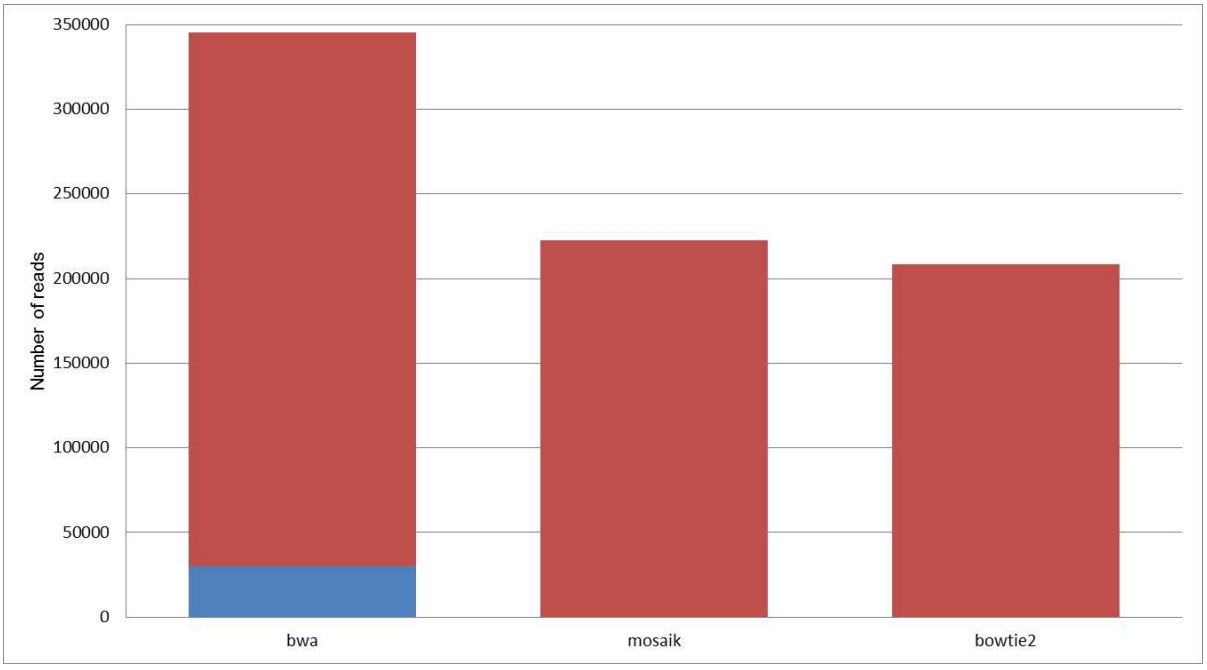

Figure 2.3 Overlapped barcode-corrected reads which are agree or do not agree in overlap. Red color - agree, blue - do not agree.

In Figure 2.3, according BWA-MEM 345,100 paired-end barcode-corrected reads have overlap between forward and reverse parts. Among them 315,327 reads are agree in overlap (both ends are equal each other) and 29,773 do not agree. According Mosaik 222,760 reads have overlap, among them 222,765 agree and 4 do not agree. According Bowtie2

208,237 reads have overlap, among them 208,174 agree and 63 do not agree.

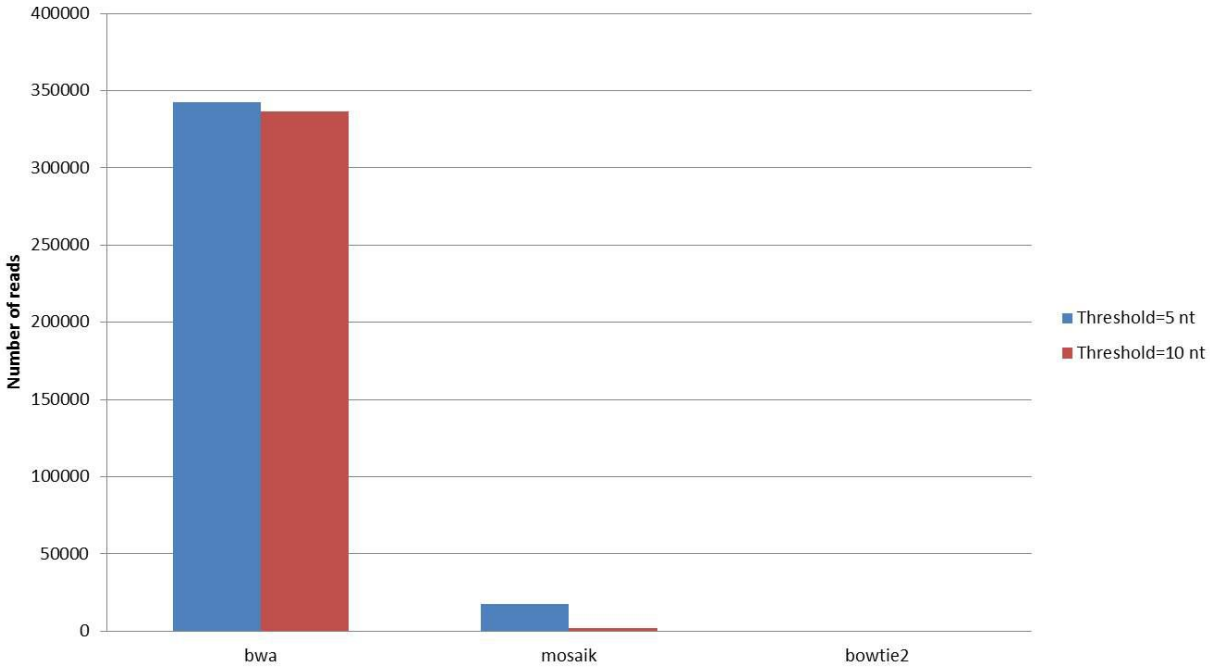

Figure 2.4 Soft-clipped barcode-corrected reads.

In Figure 2.4, BWA-MEM produced 342,761 soft-clippings with the length 5nt or greater, and 336,269 soft-clippings with the length 10nt or greater. Mosaik produced 17,498 soft-clippings with the length equal or greater than 5nt, and 1,942 soft-clippings with the length equal or greater than 10nt. Bowtie2 did not produce any soft-clipping.

In Figures 2.5, 2.6, 2.7 the fragment length distribution of barcode-corrected reads mapped by different aligners are represented.

In Figures 2.8, 2.9, 2.10 the number of SNPs in barcode-corrected reads according different aligners is shown. The first chart represents the number of SNPs in all reads with thresholds at least 1 SNP per read, at least 5 SNPs per read and at least 10 SNPs per read. The second chart shows SNPs in overlapped reads, the third shows SNPs in not overlapped reads. It may be noticed that BWA produces mostly alignments with more than 10 SNPs per read.

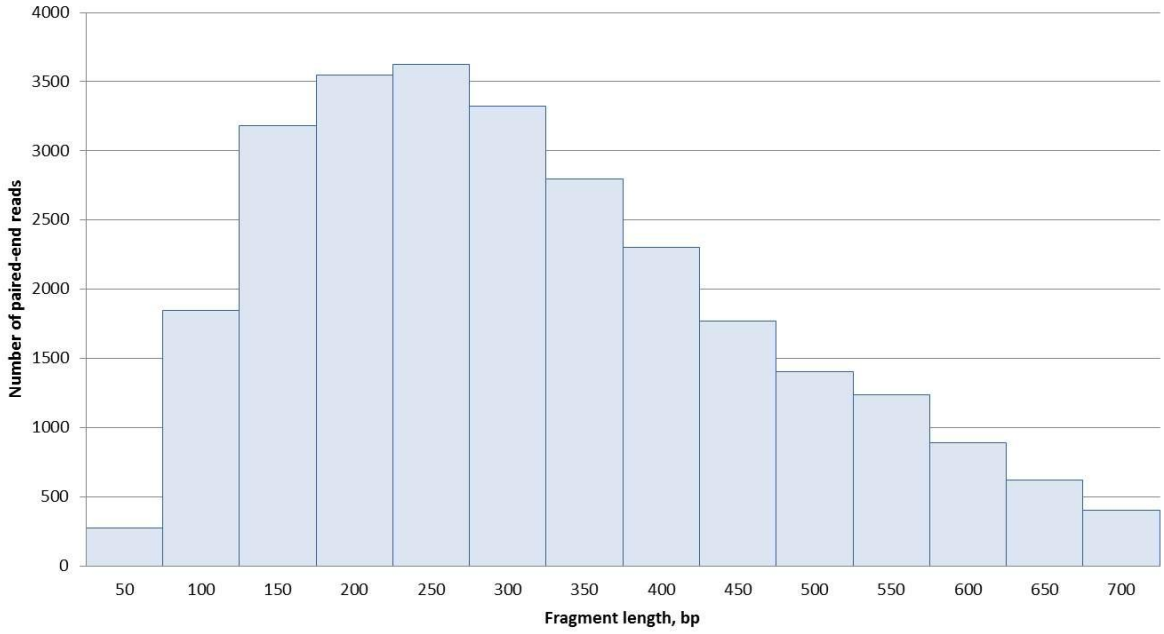

Figure 2.5 Fragment length distribution of barcode-corrected reads mapped by BWA-MEM. Results are shown for the length from 0 bp to 700bp. The smallest fragment length was 2bp, the biggest fragment length was 3440bp.

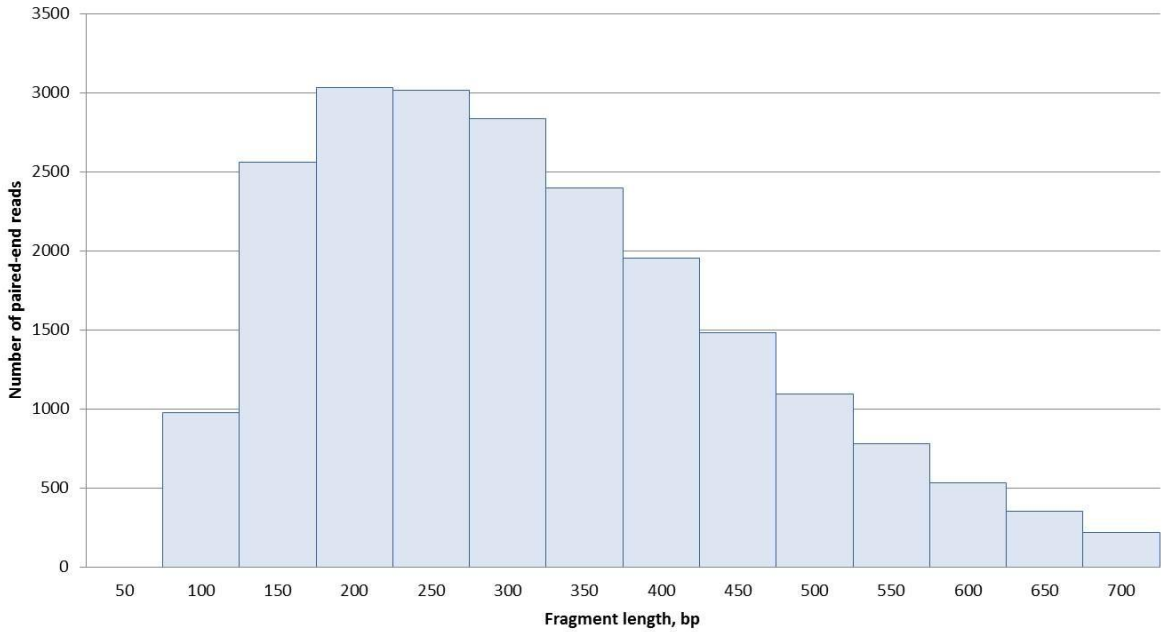

Figure 2.6 Fragment length distribution of barcode-corrected reads mapped by Mosaik. Results are shown for the length from 0 bp to 700bp. The smallest fragment length was 73bp, the biggest fragment length was 3397bp.

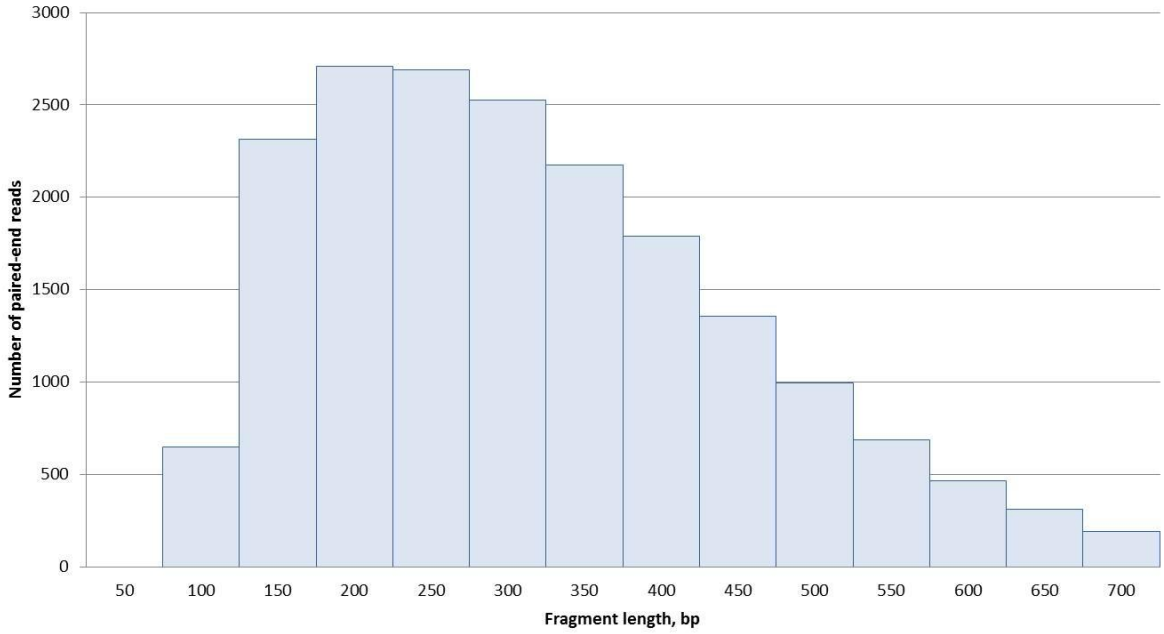

Figure 2.7 Fragment length distribution of barcode-corrected reads mapped by Bowtie2. Results are shown for the length from 0bp to 700bp. The smallest fragment length was 77bp, the biggest fragment length was 3398bp.

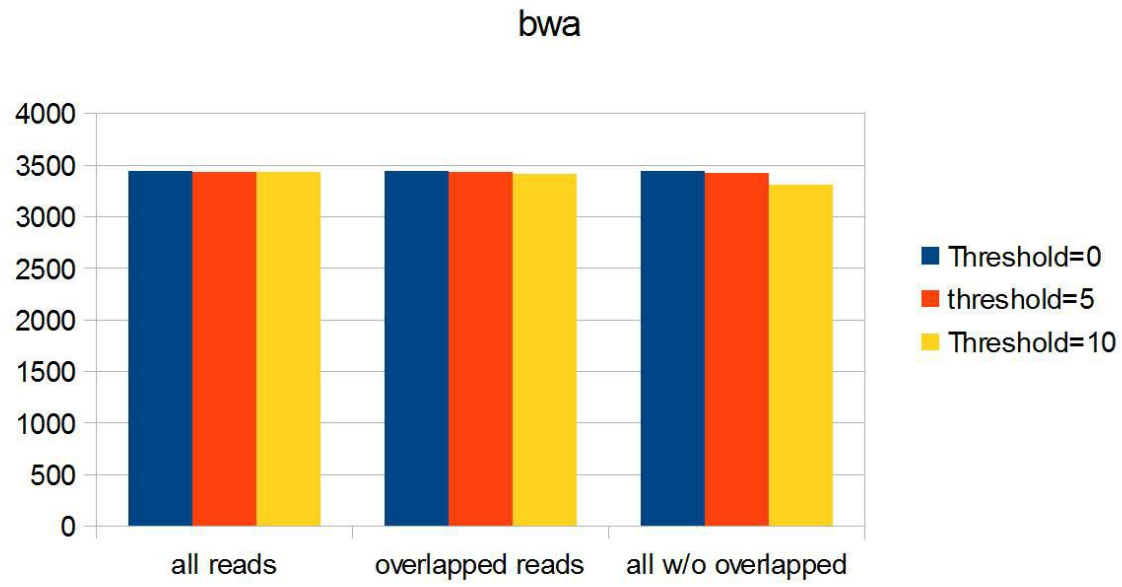

Figure 2.8 Number of SNPs in barcode-corrected reads according BWA-MEM

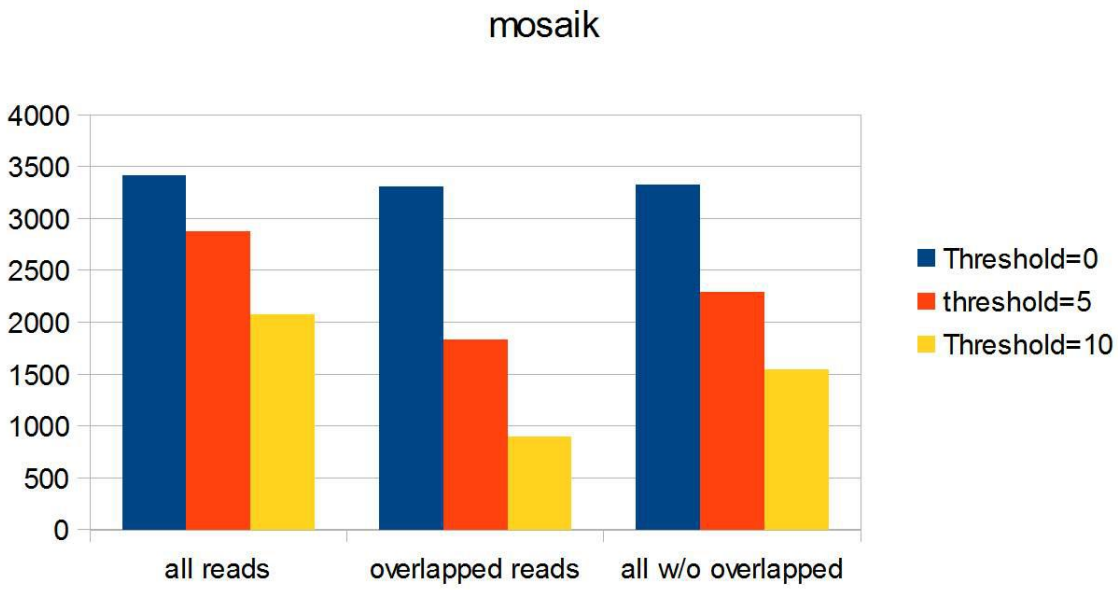

Figure 2.9 Number of SNPs in barcode-corrected reads according Mosaik

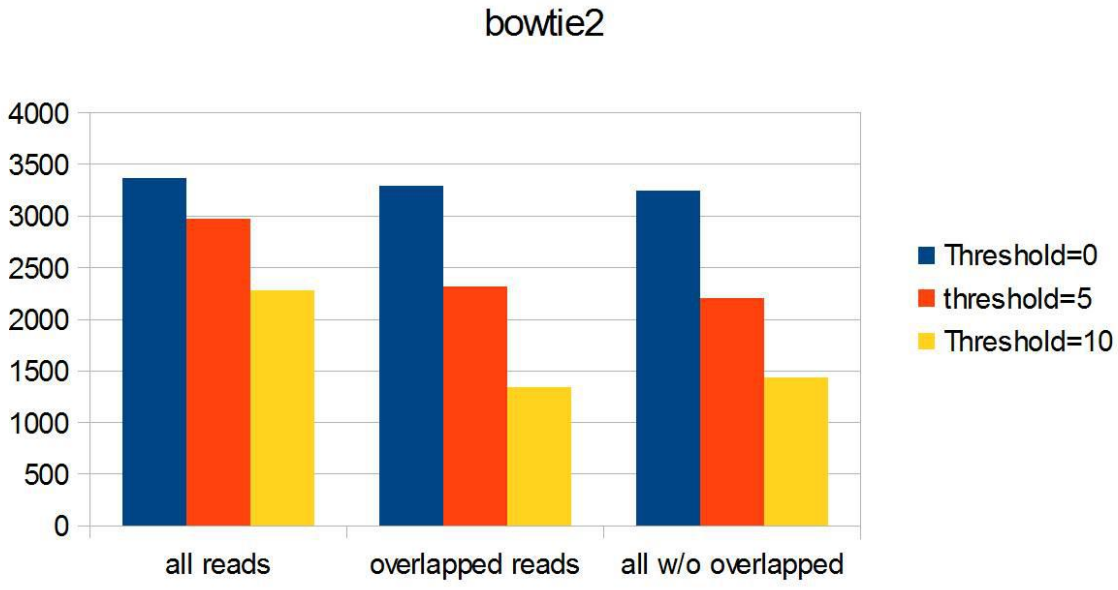

Figure 2.10 Number of SNPs in barcode-corrected reads according Bowtie2

Results of error correction

We evaluated the performance of error correction tools BLESS, Fiona, Pollux, KEC, and ShoRAH on correcting original reads to their true sequence. The consensus sequence of the barcoded reads is considered as its true sequence and compared to the error corrected read from each tool.

We compared the length of the error corrected reads from each tool. A majority of reads generated from a tool are of length 87bps, which is the length of DNA segment sequenced in each read. Nevertheless, Fiona, Pollux, and KEC also generate error corrected reads that are greater than 87 bps, indicating that these methods may have introduced new bases in these reads during error correction. BLESS corrects 92.1% of the reads to 87 bps, while none of the reads have lengths greater than 87 bps, indicating that it only trims the erroneous reads during error correction and does not adds additional bases to the reads (Figures 2.11, 2.12, 2.13, 2.14, 2.15).

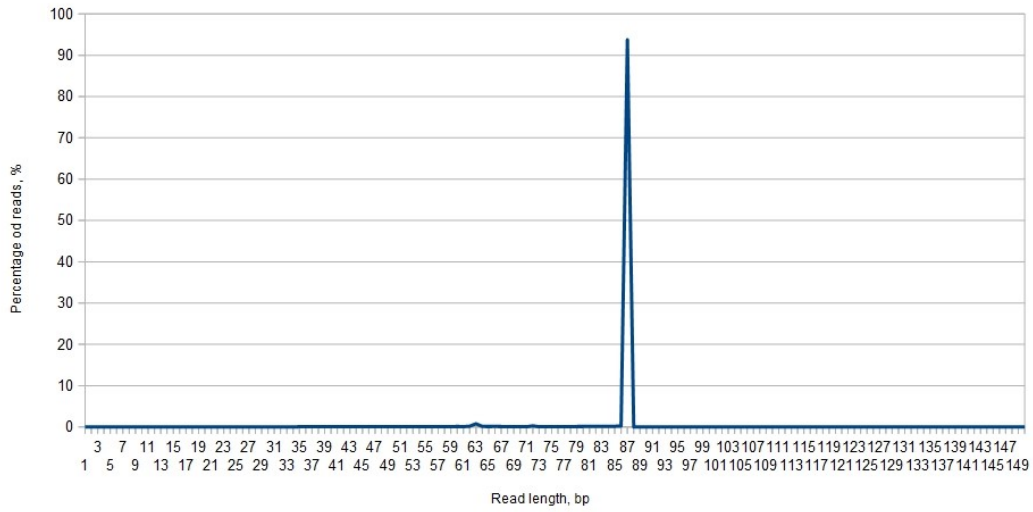

Figure 2.11 BLESS corrected reads length distribution.

Figures 2.16, 2.17, 2.18 represent the cumulative number of reads corrected by different tools with different edit distance to the corresponding true reads divided by the total number of original reads. BLESS was able to accurately correct 90% of original reads to their consensus sequences. 95% of reads were corrected with no more than 1 mismatch.

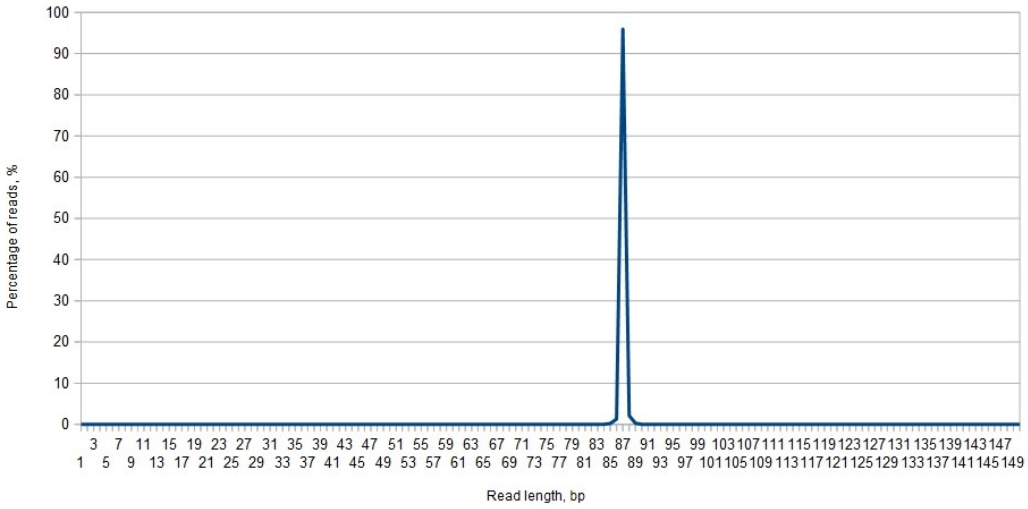

Figure 2.12 Fiona corrected reads length distribution.

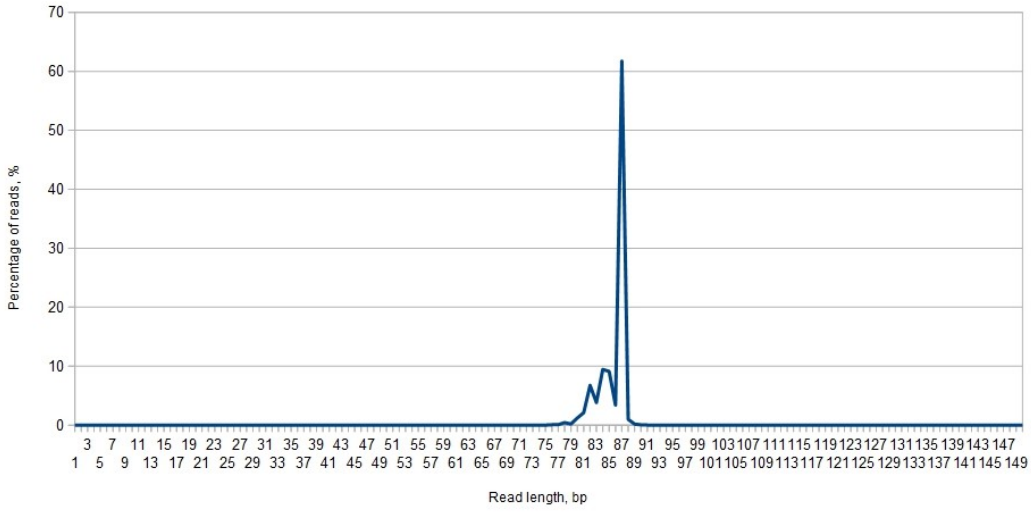

Figure 2.13 Pollux corrected reads length distribution.

Fiona did not correct right any reverse original read. Fiona over corrected even those original reads that were right. Totally Fiona was able to accurately correct 46% of original reads, if count reverse and forward together. Pollux was able to accurately correct 70% of original reads. 95% of reads contained no more than 3 mismatches.

ShoRAH returns a non-redundant set of reads that can be used as input for its viral reconstruction algorithm, thus this comparison cannot be performed for it. Running

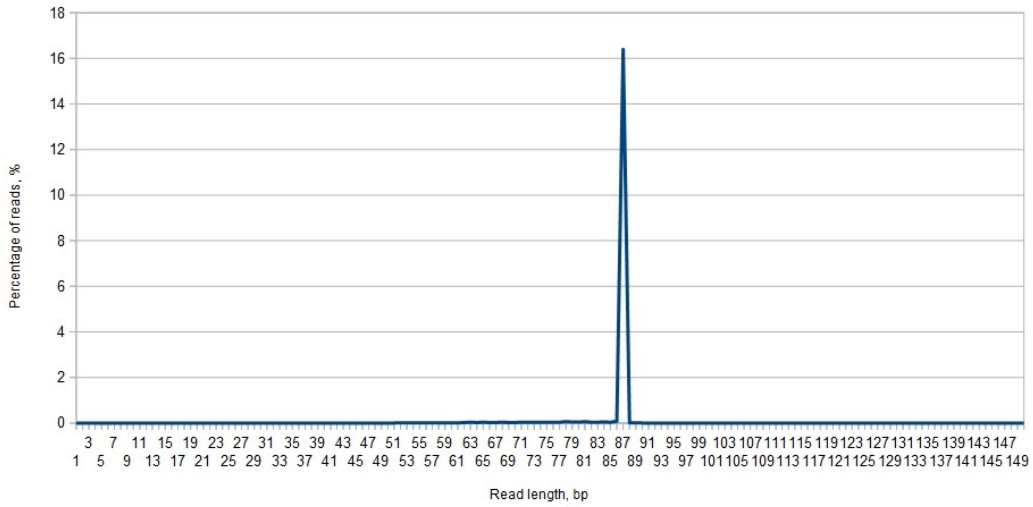

Figure 2.14 KEC corrected reads length distribution.

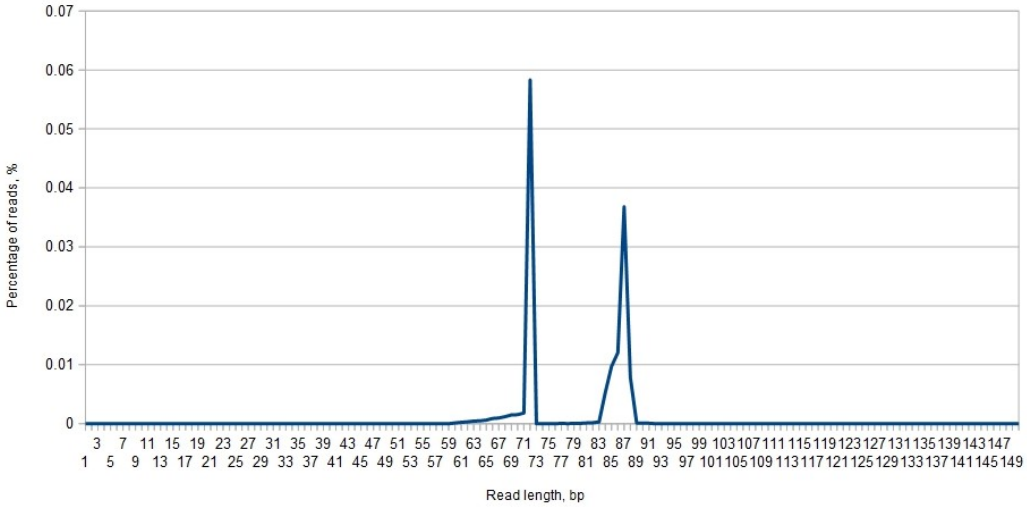

Figure 2.15 ShoRAH corrected reads length distribution.

KEC error correction tool on original reads we faced some difficulties. First, KEC does not work on paired-end reads, so we mixed all reads together and considered them as single. Second, KEC could not handle all reads. We considered only reads assigned to barcode clusters with consensus (true read) mapped to Gag/Pol 3.4 Kb HIV region. The number of such reads was 37,624,122. The reads were sorted according the position of their cluster consensus in Gag/Pol. Then the reads were divided into groups of approxi-

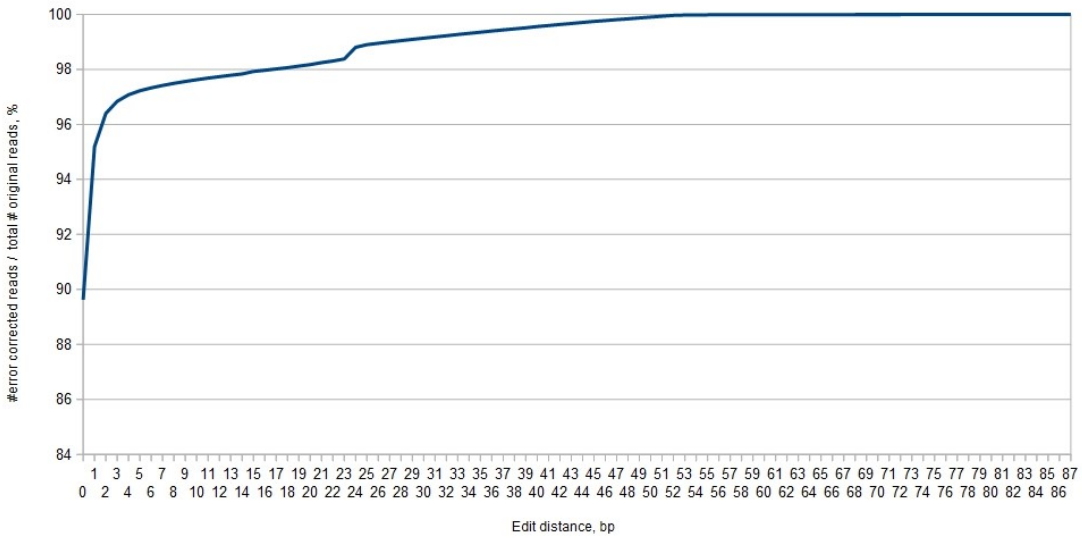

Figure 2.16 Cumulative number of reads corrected by BLESS with different edit distance to the corresponding true reads divided by the total number of original reads.

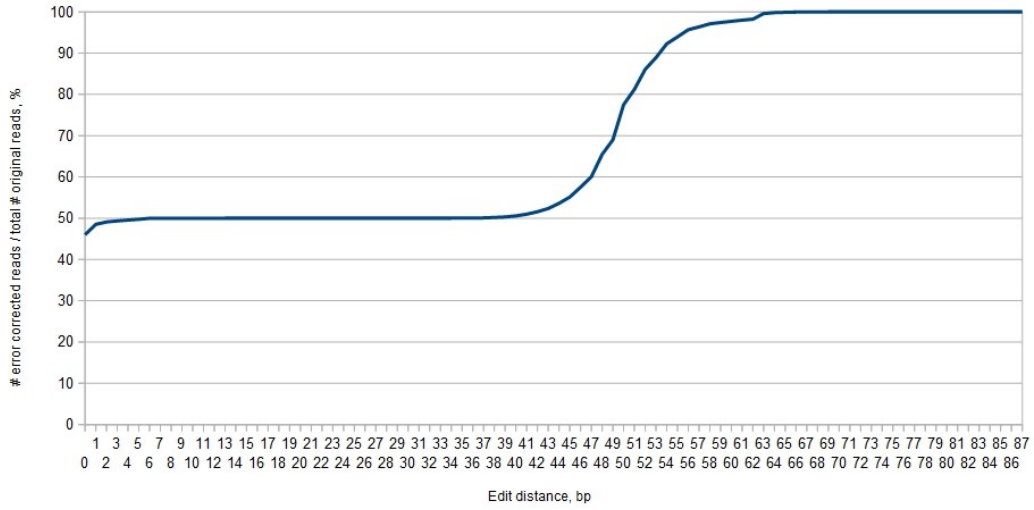

Figure 2.17 Cumulative number of reads corrected by Fiona with different edit distance to the corresponding true reads divided by the total number of original reads.

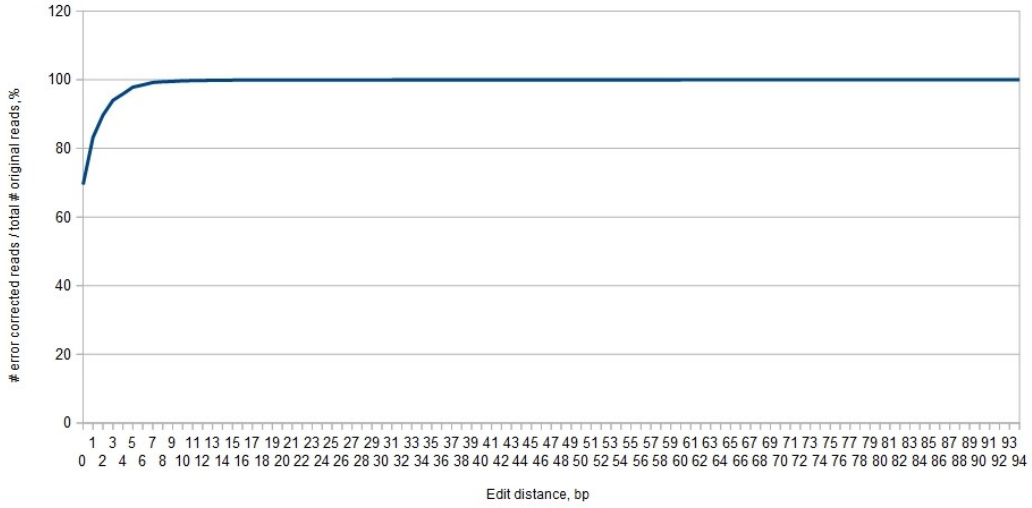

Figure 2.18 Cumulative number of reads corrected by Pollux with different edit distance to the corresponding true reads divided by the total number of original reads.

mately 5 million reads. Totally we obtained 7 such groups. KEC was run on each of those groups separately. Finally all corrected reads were mixed together (37,041,305 reads). KEC deleted some reads during error correction. After error correction most of the reads (94.4%) did not change their size. KEC removed the name of original reads which made read-by-read comparison with barcode-corrected reads impossible. Instead, we used another method for evaluation. The total number of true barcode-corrected reads mapped to Gag/Pol 3.4 Kb HIV region was 2,748,821 among them 132,816 reads were distinct. The total number of original reads belonging to barcode clusters with mapped consensus true reads was 37,624,122, among them 2,032,694 were distinct. KEC produced 37,041,05 corrected reads decreasing the original number of reads, the number of distinct reads was 604,591. Finally, BLESS output contained 37,624,122 reads with 1,907,275 distinct reads. To evaluate the tools we considered only distinct reads. Among 604,591 distinct KEC reads only 52,632 reads were equaled to the true reads which gives sensitivity of 0.396 and PPV of 0.087. Among 1,907,275 distinct BLESS reads only 95,015 reads were equaled to the true reads which gives sensitivity of 0.715 and ppv of 0.05.

2.2 Assembly of viral NGS data

2.2.1 Introduction

RNA viruses are causing a significant burden on the health and productivity of agriculturally important animals since the rapid evolution of RNA viruses within infected hosts is coupled in this context with frequent transmissions between animals, due to the typically high animal density in production environments. Poultry farms are particularly susceptible to viral infections, which cause significant economic losses worldwide in terms of impaired growth, reduced egg production and quality, and even mortality.

In the U.S. where infections with virulent strains of Newcastle disease and highly pathogenic avian influenza are not common, the infectious bronchitis virus (IBV) is considered the biggest single cause of economic loss. It infects the domestic fowl [37]. Besides causing respiratory disease it may also replicate at many non-respiratory epithelial surfaces. First described in the U.S. in 1930, IBV has undergone exponential population growth [38] and is now distributed worldwide, with dozens of circulating serotypes that may differ by as much as 25% in the amino acid sequence of the hypervariable region of the spike glycoprotein.

As serotypes cross-protect poorly, chicken may get infected multiple times during their lifetime. IBV infection of broilers retards growth while in layers it drops down egg production. Young chicken may die directly from IBV infection, but a greater number die due to secondary bacterial infections. Vaccination, most commonly with live attenuated vaccines, is broadly used to control IBV disease but protection is short-lived and layers have to be revaccinated multiple times during their lifespan, sometimes with different serotypes.

Several IBV serotypes can co-circulate in a region, creating conditions for recombination between strains and resulting in complex evolution and epidemiology [39]. Cloning and sequencing has been used to show that IBV variants exist in infected poultry [40] as well as in several widely used commercial live attenuated vaccines [41]. Furthermore,

McKinley et al.[41] has shown that attenuated live vaccines undergo in vivo selection following vaccination and can infect contact-exposed chicken. Coupled with significant field persistence of Arkansas-type vaccine viruses [42], this reinforces the evidence that infectious IBV strains can emerge from the evolution of attenuated live vaccines in commercial flocks [43].

NGS is emerging as a key technology for quasispecies analysis. NGS does not directly sequence viral genomes describing a viral population. Instead, it produces relatively short reads that should be assembled into population variants. Initial studies have focused on using NGS to identify extremely low frequency drug-resistant variants in human patients chronically infected with HIV [44, 45, 46, 47, 48]. Just as PCR and automated Sanger sequencing revolutionized molecular biology, NGS is expected to transform research on phylodynamics of RNA viruses [49]. However, NGS analysis is challenging due to the huge amount of data on one hand, and to the short read lengths and high error rates on another. Another challenge is that variants are close to each other, there are overlapping regions. There should be decision which read belongs to which variant.

Many tools developed for Sanger reads do not work at all or have impractical runtimes when applied to NGS data. Even newly developed algorithms for de novo genome assembly from NGS data are tuned for reconstruction of haploid genomes, and work poorly when the sequenced sample contains a large number of closely related sequences, as is the case in viral quasispecies. To address these shortcomings we evaluated reconstruction flows for accurate reconstruction of viral quasispecies sequences and estimate their frequencies from NGS data where it incorporates different error correction methods, aligners, and genome assemblers (ViSpA [8], and ShoRAH [7]), using different tuning parameters. It should be noted some other recent genome reconstruction tools such as QuasiRecomb [4], BIOA:VirA [30], QuRe [9] and PredictHaplo [50]. In the current work we considered only ShoRAH and ViSpA. We applied experiments on IBV 454 shotgun reads, collected from commercial poultry farms. For method validation we use IBV sanger clones as ground truth.

2.2.2 Methods for viral quasispecies reconstruction

The proposed reconstruction flows consist of the following stages (see Figure 2.19):

1. **Read error correction.** This step is necessary since the reads produced by 454 Life Sciences system are prone to errors and it is important to distinguish reading errors from rare viral variants. As mentioned before, 454 Life Sciences can erroneously sequence one base pair per 1000 base pair . The error rate is strongly related to the presence and size of homopolymers [51], i.e. genome regions, consisting of consecutive repetition of a single base (for example, TTTTTT).

   We use KEC, SAET and ShoRAH programs to do error correction prior to assembly. These error correction algorithms involve clustering of reads. While ShoRAH clusters the reads in Bayesian fashion using the Dirichlet process mixture [52], and KEC clusters reads based on kmers [53], SAET uses reads quality scores for error correction.

2. **Read alignment.** In this step, we use independent alignment program to map reads against a reference viral sequence[54], this aligner can be easily replaced with another one.

3. **Reconstruction of viral quasispecies.** In this step, we use two assembly programs ViSpA [8] and ShoRAH [7] to reconstruct variants from aligned reads and estimate their relative frequencies.

ViSpA [8] executes the following steps and outputs the quasispecies spectrum (i.e. variant sequences and their relative frequencies):

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

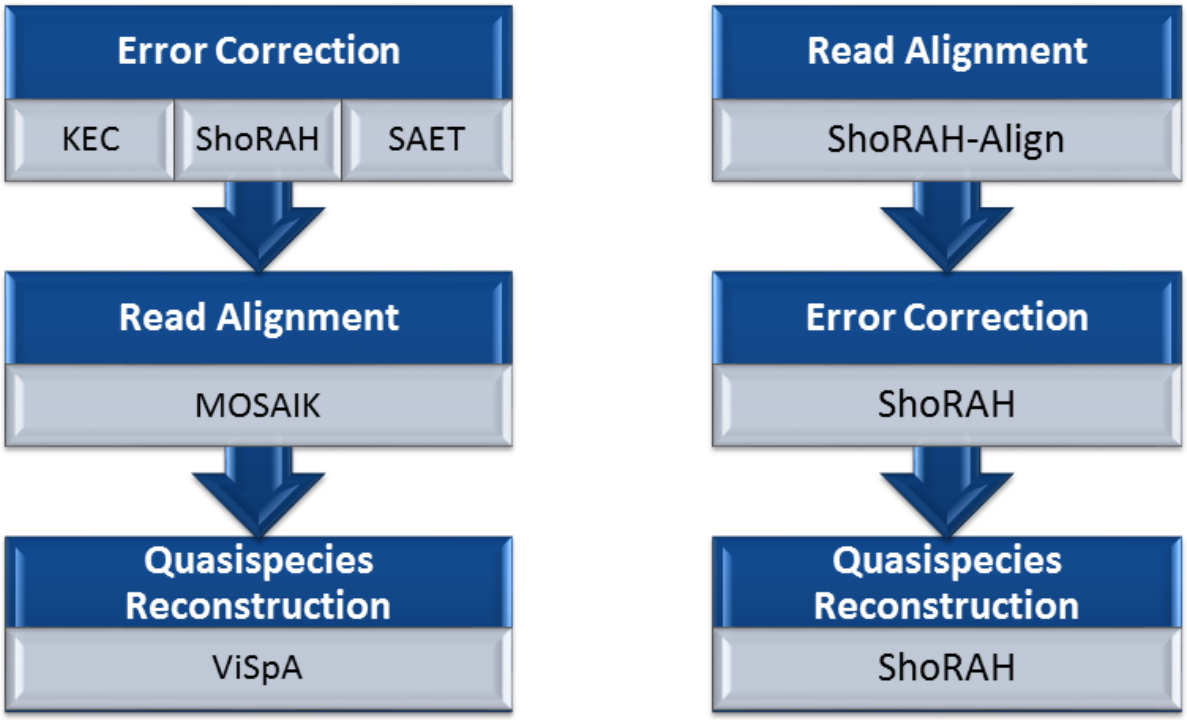

Figure 2.19 Evaluated reconstruction flows. The flows consists of 3 steps: (1) Read error correction (2) Read alignment and (3) Reconstruction of viral quasispecies.

– **Construct the read graph.** In the read graph each vertex corresponds to a read and each directed edge connects two overlapping reads. ViSpA differentiates between two types of reads, super-read and sub-read which is a substring of the super-read. The read graph consists only of super-reads.

– **Assemble candidate quasispecies sequences.** Each candidate variant corresponds to a path in the read graph. ViSpA uses what is so called max-bandwidth paths for assembly.

– **Estimate frequency of haplotype sequences.** In this step, ViSpA uses Expectation Maximaization algorithm to estimate the frequency of each reconstructed sequence using both super-reads and sub-reads.

ShoRAH [7] executes the following steps and outputs the quasispecies spectrum:

– **Align reads.** The first step for ShoRAH is producing a Multiple Sequence Align-

ment (MSA) of reads. We used the version 0.5 of the ShoRAH. That version had its own aligner which was used in the study . It aligns all reads to the reference and from the set of pairwise alignments it builds a MSA.

- **Correct Reads from genotyping errors (Local Haplotype Reconstruction).** While ViSpA uses independent error correction programs, ShoRAH uses its own error correction method. Sequencing errors are corrected by a Bayesian inference algorithm which estimates the quality of the reconstruction, although only the maximum likelihood estimate is passed on to subsequent steps. ShoRAH implements a specific probabilistic clustering method based on the Dirichlet process mixture for correcting technical errors in deep sequencing reads and for highlighting the biological variation in a genetically heterogeneous sample.
- **Reconstruct global haplotype.** This step is similar to assembly of candidate quasispecies sequences in ViSpA.
- **Estimate frequency.** In this step, ShoRAH estimates the frequency of each candidate sequence.

We varied different parameter values. The following tuning parameters are used for ViSpA quasispecies reconstruction:

- n : number of mismatches between super-reads and sub-reads
- m : number of mismatches in the overlap between two super-reads
- t : mutation rate

For ShoRAH we used the default parameters for quasispecies reconstruction.

2.2.3 Results

Data Sets

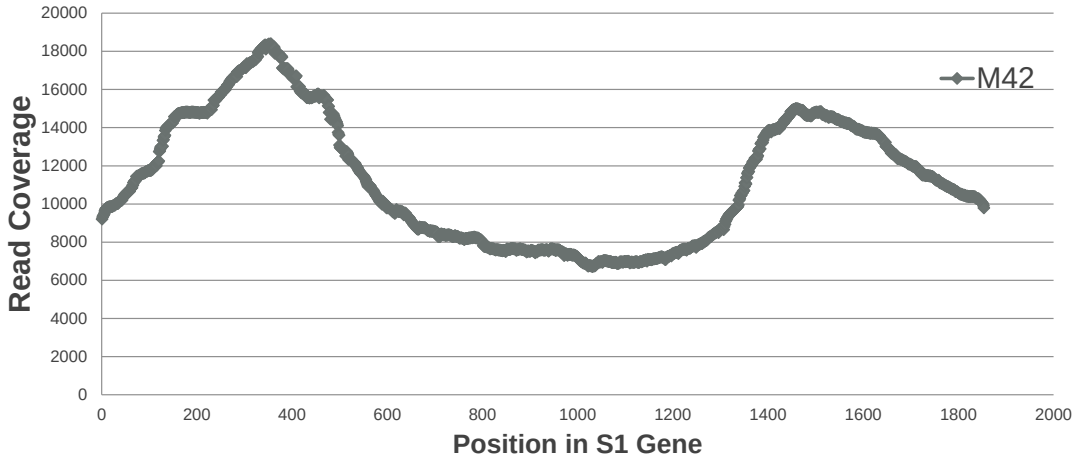

Figure 2.20 Read coverage. Number of reads covered every position in S1 gene.

Read samples were collected from IBV infected chickens, and quasispecies variants were sequenced using life sciences 454 shotgun sequencing, followed by Sanger sequencing of individual variants. Reads coverage profile is represented on Figure 2.20. The initial number of reads was 21,040. After SAET error correction the number of reads did not change. It decreased to 19,439 after ShoRAH correction and to 17,122 after KEC correction. Ten Sanger clones (true variants) were used for validation. These clones are considered as the golden standards or the ground truth for parameter calibration and comparison between different methods (see Figure 2.21).

Tuning, comparison and validation of methods for quasispecies reconstruction

We measured the pairwise edit distance for the 10 Sanger clones as follow. We ran pairwise alignment for the Sanger clones using ClustalW, then cut the sticking out ends, and computed the pairwise edit distance (Levenshtein Distance) for all clones in the overlapped region (see Table 2.1).

To validate the reconstruction of the quasispecies using different methods, we com-

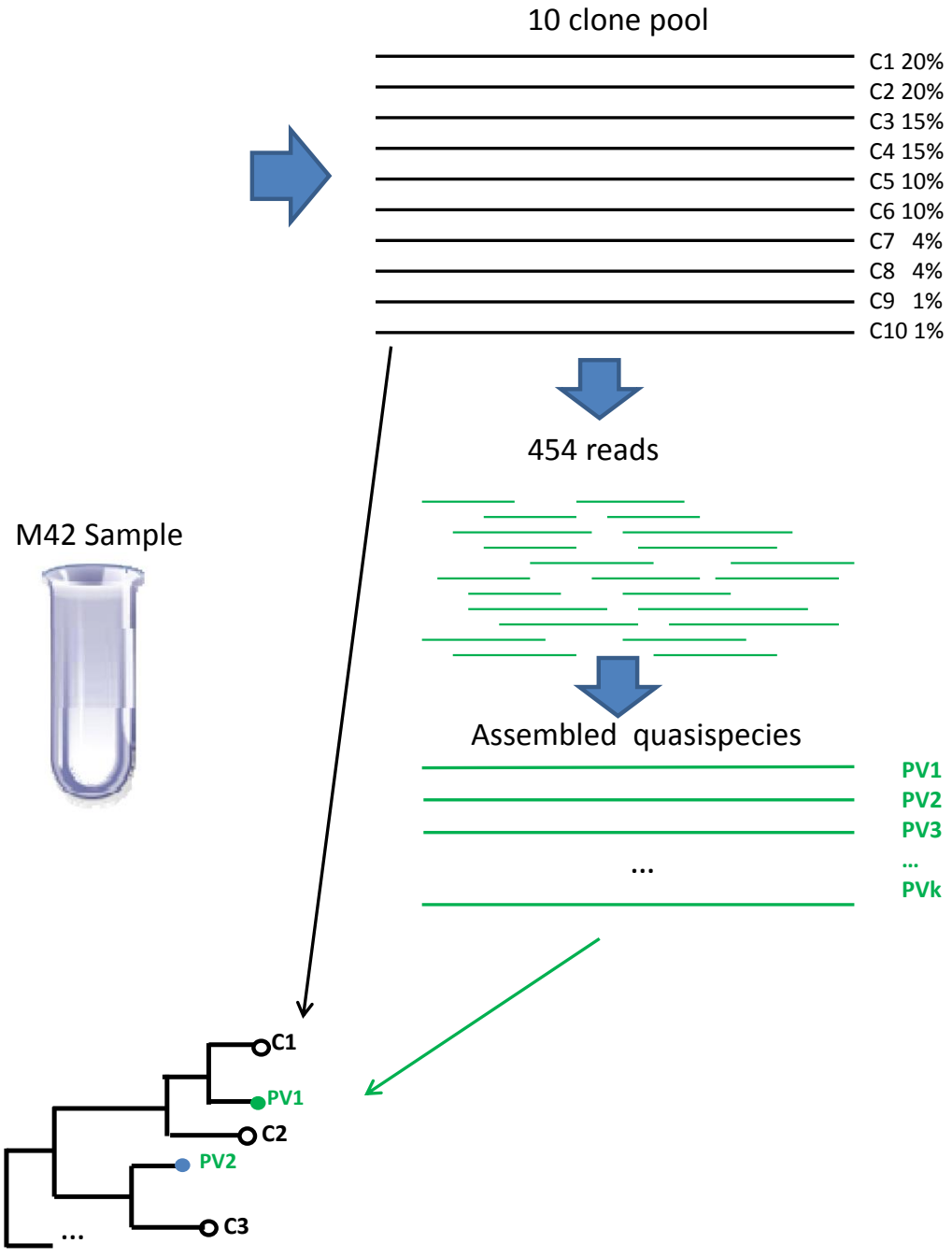

Figure 2.21 Schematic representation of calibration, validation experiments based on Sanger clones

Table 2.1 Pairwise edit distance between the 10 Sanger clones. The suffix number of the clone id is the clone frequency, e.g 42E9_A08_20 means that the frequency of the clone 42E9_A08 is 20%

Sanger clones	**42E9_A08_20**	**42V1H7_C02_20**	**42E3_C07_15**	**42A6_F04_15**	**42E10_B08_10**	**42H1_E11_10**	**42V1B10_B04_04**	**42V1H3_G01_04**	**42H5_A12_01**	**42V1C6_B08_01**
42E9_A08_20	0	0	1	1	1	1	3	3	2	3
42V1H7_C02_20		0	1	1	1	1	3	3	2	3
42E3_C07_15			0	2	2	2	4	4	3	4
42A6_F04_15				0	0	2	2	2	3	2
42E10_B08_10					0	2	2	2	3	2
42H1_E11_10						0	3	2	1	2
42V1B10_B04_04							0	2	4	3
42V1H3_G01_04								0	3	2
42H5_A12_01									0	3
42V1C6_B08_01										0

puted the pairwise distances between Sanger clones and reconstructed variants. We measured the distances as follows. We computed pairwise alignment between Sanger clones and reconstructed variants using ClustalW, then cut sticking out ends, and got overlapped region between each clone and each variant, the reason behind that, was to avoid calculating edit distance for uncovered fragments (all overlapped regions happened to be more than 500 bp long, which is close to Sanger clones average lenght (=550 base pairs)).

As a result of the previous step, we got several groups of identical overlapped regions of reconstructed variants and collapsed them, and summed up their frequencies (the suffix of the variant sequences ID's represent the collapsed variant abundance, for example, the frequency of the variant 42E9_A08_40 is 40%).

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

Table 2.2 Edit distance between collapsed Sanger clones and ViSpA reconstructed variants using parameters 1,2,5 (number of mismatches between sub-reads and super-reads, number of mismatches between two overlapped reads and mutation rate respectively), threshold=0.005 on KEC corrected reads

	Sanger clones							
Frequencies of reconstructed variants	**42E9_A08_40**	**42A6_F04_25**	**42E3_C07_15**	**42H1_E11_10**	**42V1B10_B04_04**	**42V1H3_G01_04**	**42H5_A12_01**	**42V1C6_B08_01**
0.6647	**0**	1	1	1	3	3	2	3
0.1212	1	**0**	2	2	2	2	3	2
0.0399	1	2	**0**	2	4	4	3	4
0.1054	2	3	3	3	5	5	4	5
0.0461	2	1	3	3	3	3	4	3
0.0228	1	2	2	2	4	4	3	4

Table 2.2 shows edit distance values between distinct Sanger clones and reconstructed variants for one of the dominating methods using ViSpA where the number of mismatches between super-reads and sub-reads (n) is 1, the number of mismatches in the overlapping region between two reads (m) is 2 and the mutation rate (t) is 5. We considered only reconstructed variants with frequencies more than 0.005 (threshold=0.005). Reads are corrected using KEC program.

As we see from the Table 2.2, the method correctly reconstructed three the most frequent variants. Thus in the overlapped region, there was edit distance 0 between the most frequent Sanger clone with id 42E9_A08_40 and frequency 0.40 and the reconstructed variant with the frequency 0.6647. As well as 42A6_F04_25 clone was identical to the variant with frequency 0.1212 and 42E3_C07_15 clone was identical to the variant with frequency 0.0399. The method a little overestimated the highest frequency - obtained 0.6647 for the most popular variant instead of true 0.40. On the other hand, it underestimated the second and the third popular variants - obtained frequencies 0.1212 instead of 0.25 and 0.0399 instead of 0.15. On the whole, the validation with distinct Sanger clones

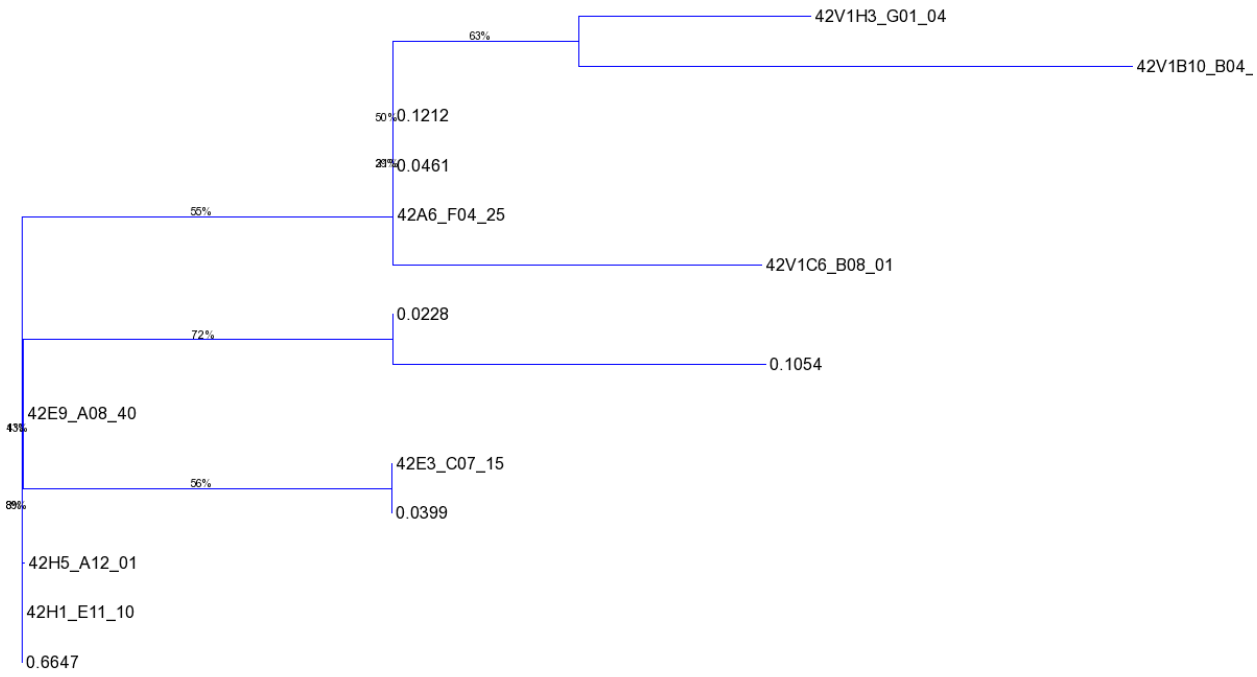

Figure 2.22 Phylogenetic tree over collapsed Sanger clones and collapsed reconstructed variants inferred from the method with parameters 1_2_5 on KEC corrected reads using ViSpA.

show that existing reconstruction methods can give adequate representation of IBV variant population.

To compare how well the reconstructed variant sequences of the previous method were compared to Sanger clones, we reconstructed the phylogenetic tree, see Figure 2.22 (it shows only the frequecies for the reconstructed variants, while a Sanger clone starts with the number 42). As we can see from the phylogenetic tree, the reconstructed variants by ViSpA are close to Sanger clones.

To compare between different methods (with different parameter settings), we use the following two measures:

- average distance to clones ADC=$\sum_i d(c_i) \cdot f(c_i)$
- average prediction error APE=$\sum_j d(q_j) \cdot f(q_j)$,

where $c_1, ..., c_{10}$ are Sanger clones, $q_1, ..., q_j$ are reconstructed variants, $f(c_i)$ and $f(q_j)$ are frequencies of ith clone and jth variant and $d(c_i)$ and $d(q_j)$ are edit distance from ith

Table 2.3 Edit distance between collapsed Sanger clones and ViSpA reconstructed variants using parameters 2, 2, 10 (the number of mismatches between sub-reads and super-reads, the number of mismatches between two overlapped reads, and mutation rate respectively), threshold=0.005 on KEC corrected reads, where 85% of reconstructed variants have perfect match with 65% of the Sanger clones.

	Sanger clones							
Frequencies of reconstructed variants	**42E9_A08_40**	**42A6_F04_25**	**42E3_C07_15**	**42H1_E11_10**	**42V1B10_B04_04**	**42V1H3_G01_04**	**42H5_A12_01**	**42V1C6_B08_01**
0.6703	**0**	1	1	1	3	3	2	3
0.1845	1	**0**	2	2	2	2	3	2
0.0054	4	3	5	5	5	5	6	5

clone to the closest reconstructed variant and edit distance from jth reconstructed variant to the closest clone.

ADC and APE are analogous to sensitivity and ppv respectively. The difference is that ADC and APE indicate better quality whenever they are closer to 0, while sensitivity and ppv indicate better quality when they are closer to 1.

In addition to the previous method. We ran many other experiments using different methods (see Tables 2.3 - 2.5 and Figures 2.23 - 2.25) and calculated the values of Average Distance to Clones (ADC) and Average Prediction Error (APE) for each method as shown in Table 2.6 and Table 2.7 respectively (ViSpA 1_2_5 means that the variants are reconstructed by ViSpA using parameter values n=1, m=2, t=5). The Table 2.5 shows that the method is able to recall or reconstruct 12.7% of the variants with 0 edit distance with the most frequent Sanger clones (40%), 6.6% of the variants with 0 edit distance with 25% frequent Sanger clones, 3.8% of the variants with 0 edit distance with 15% frequent Sanger clones, therefore the used method is able to reconstruct or recall 23% (0.127+0.066+0.038) of the variants with 0 edit distance with 80% (0.40+0.25+0.15) of Sanger clones with relatively small precision since it reconstructs more false negatives.

Table 2.4 Edit distance between collapsed Sanger clones and ViSpA reconstructed variants using parameters 1, 2, 0 (the number of mismatches between sub-reads and super-reads, the number of mismatches between two overlapped reads, and mutation rate respectively), threshold=0.005 on SAET corrected reads.

	Sanger clones							
Frequencies of reconstructed variants	**42E9_A08_40**	**42A6_F04_25**	**42E3_C07_15**	**42H1_E11_10**	**42V1B10_B04_04**	**42V1H3_G01_04**	**42H5_A12_01**	**42V1C6_B08_01**
0.1448	**0**	1	1	1	3	3	2	3
0.0714	1	**0**	2	2	2	2	3	2
0.0067	1	2	**0**	2	4	4	3	4
0.1923	1	2	2	2	4	4	1	4
0.0469	2	3	3	3	5	5	4	5
0.0061	4	5	5	5	7	7	6	7

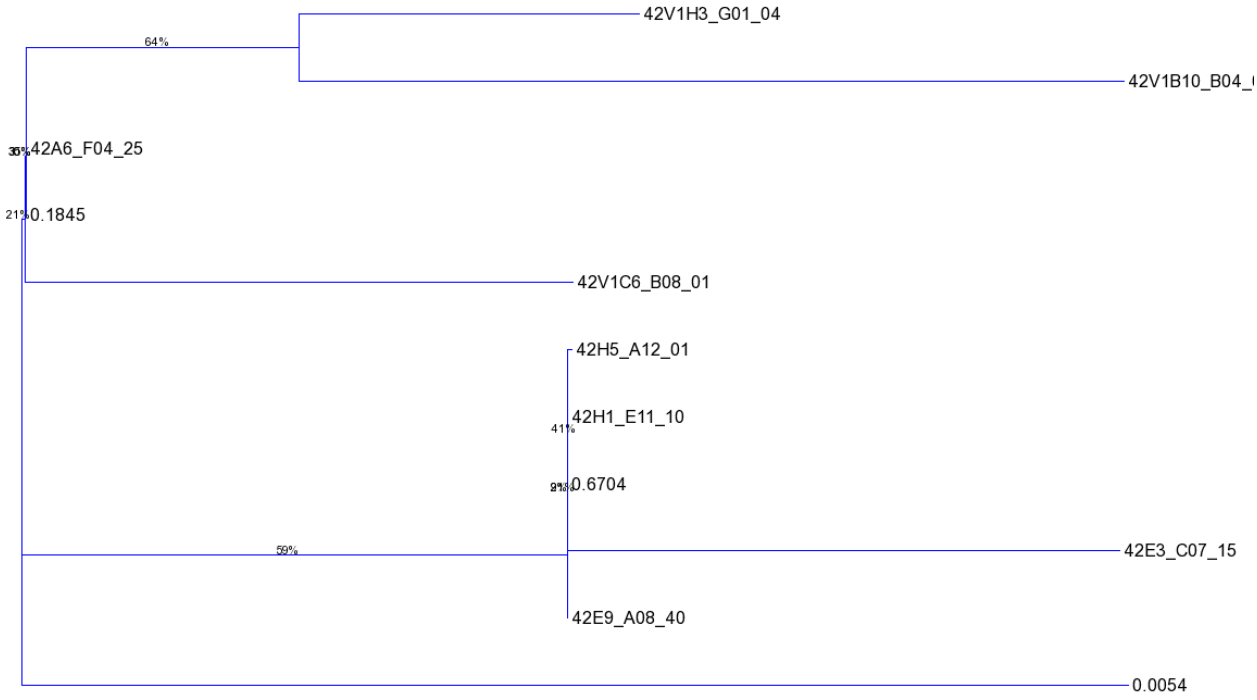

Figure 2.23 Phylogenetic tree over collapsed Sanger clones and collapsed reconstructed variants inferred from one of the dominating methods with parameters 2_2_10 on KEC corrected reads using ViSpA.

Table 2.5 Edit distance between collapsed Sanger clones and ShoRAH reconstructed variants using default parameters, threshold=0.005 on Uncorrected reads.

	Sanger clones							
Estimated variant frequencies	**42E9_A08_40**	**42A6_F04_25**	**42E3_C07_15**	**42H1_E11_10**	**42V1B10_B04_04**	**42V1H3_G01_04**	**42H5_A12_01**	**42V1C6_B08_01**
0.1277	**0**	1	1	1	3	3	2	3
0.0663	1	**0**	2	2	2	2	3	2
0.0386	1	2	**0**	2	4	4	3	4
0.0535	4	5	5	5	7	7	6	7
0.0526	2	3	3	3	5	5	4	5
0.0434	8	9	9	9	11	11	10	11
0.0289	3	4	4	4	6	6	5	6
0.0268	9	10	10	10	12	12	11	12
0.0263	5	6	6	6	8	8	7	8
0.0259	4	5	5	5	7	7	6	7
0.0243	3	2	4	4	4	4	5	4
0.0191	6	7	5	7	9	9	8	9
0.0189	5	6	4	6	8	8	7	8
0.0172	3	4	4	4	6	6	5	6
0.01718	1	2	2	2	4	4	3	4
0.0167	9	10	8	10	12	12	11	12
0.0159	2	3	1	3	5	5	4	5
0.0156	3	2	4	4	4	4	5	4
0.0127	10	9	11	11	11	11	12	11
0.0110	4	5	5	5	7	7	6	7
0.0097	1	2	2	2	4	4	3	4
0.0089	12	13	13	13	15	15	14	15
0.0088	1	2	2	2	4	4	3	4
0.0084	5	4	6	6	6	6	7	6

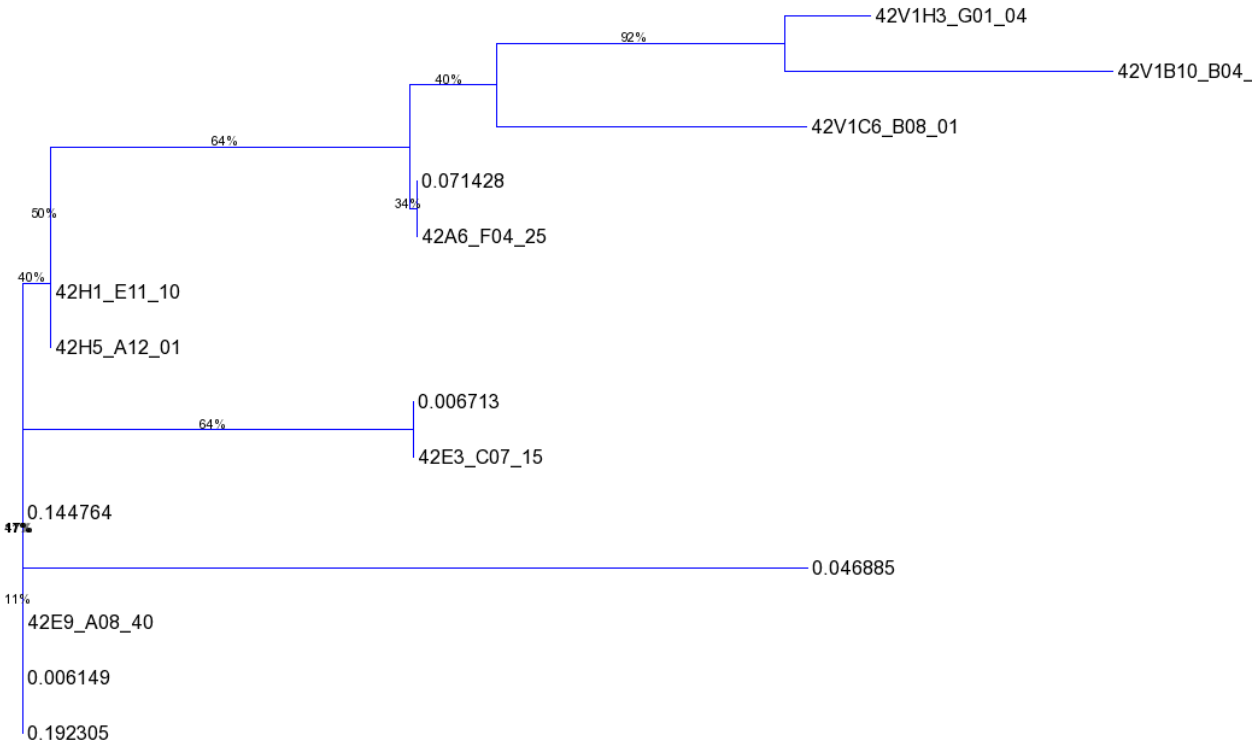

Figure 2.24 Phylogenetic tree over collapsed Sanger clones and collapsed reconstructed variants inferred from one of the dominating methods with parameters 1_2_0 on SAET corrected reads using ViSpA.

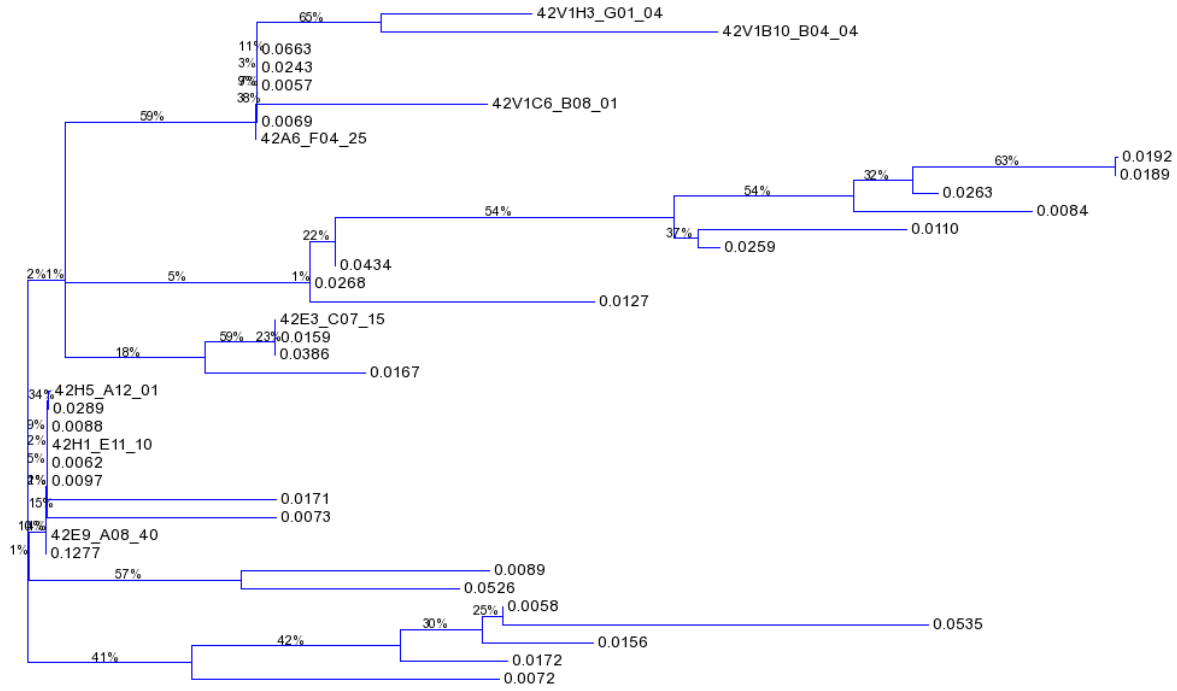

Figure 2.25 Phylogenetic tree over collapsed Sanger clones and collapsed reconstructed variants inferred from one of the methods with default parameters on Uncorrected reads using ShoRAH (close to the dominating methods).

Table 2.6 Average distance to clones (ADC) for the reconstructed variants using different methods

	Error Correction Method		
	Uncorrected	**KEC**	**SAET**
ViSpA 1_2_0	0.45	0.79	0.29
ViSpA 1_2_5	0.3	0.3	0.45
ViSpA 1_2_10	0.45	0.45	0.3
ViSpA 2_2_0	0.45	0.79	0.3
ViSpA 2_2_5	0.45	0.3	0.3
ViSpA 2_2_10	0.3	0.45	0.3
ShoRAH	0.3	0.3	0.3

We say that the method A dominates the method B if both ADC and APE values of A are at most the corresponding ADC and APE values of B. The Figure 2.26 (a pictorial diagram of Tables 2.6 and 2.7) shows that methods V125KEC (V: ViSpA assembler, 1:n, 2:m, 5:t, KEC:correction method), V2210KEC, and V120SAET dominate all other methods,

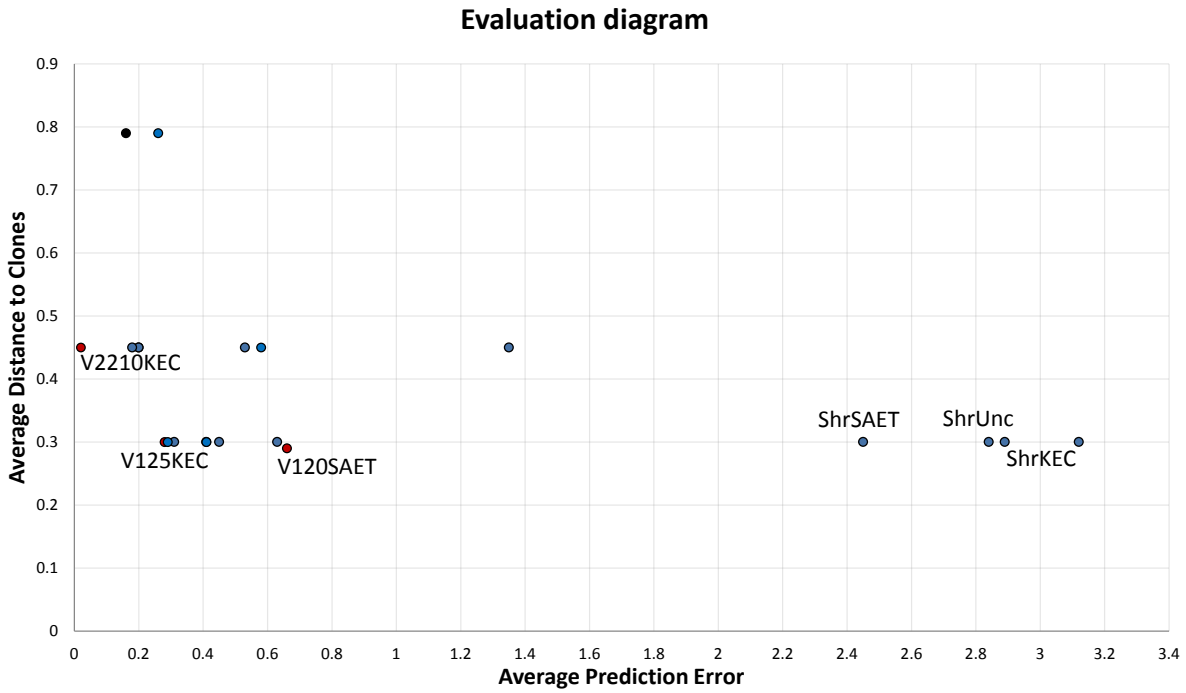

Figure 2.26 Evaluation diagram for average prediction error (APE) and average distance to clones (ADC) values for different methods. Each point corresponds to a method and the dominant solutions correspond to red points.

Table 2.7 Average prediction error (APE) for the reconstructed variants using different methods

	Error Correction Method		
	Uncorrected	**KEC**	**SAET**
ViSpA 1_2_0	0.2	0.16	0.66
ViSpA 1_2_5	0.63	0.28	0.58
ViSpA 1_2_10	0.53	0.18	0.45
ViSpA 2_2_0	1.35	0.26	3.12
ViSpA 2_2_5	0.2	0.29	0.31
ViSpA 2_2_10	0.41	0.02	0.41
ShoRAH	2.84	2.89	2.45

i.e. have the best values in terms of ADC and APE. Our results suggest that using different methods with different parameter calibration and parameter settings can improve the solution and its predictive power for the quasispecies inference problem in terms of recall and precision.

2.2.4 Conclusion

In this work, we have proposed evaluated reconstruction flows consisting of three stages, read error correction, read alignment, and quasispecies reconstruction. We vary different methods with different parameter values for accurate reconstruction of viral quasispecies sequences and estimate their frequencies from HTS data. We have proposed novel validation methods to validate our reconstructed quasispecies compared to Sanger clones as ground truth. Our experimental results show that varying different methods with different parameter calibration and parameter settings can improve the solution and predictive power of quasispecies inference problem in terms of recall and precision.

PART 3

PHAGE DISPLAY DATA PROCESSING FOR IMMUNE RESPONSE ANALYSIS

3.1 Introduction

The role of the immune system is to protect an organism against disease by responding to antigens. Antigens are substances (usually proteins) on the surface of cells, viruses, fungi, or bacteria causing disease. The immune system uses antibodies to identify antigens and neutralize disease. Serum antibodies are valuable source of information on the health state of an organism [11].

Since an antibody recognizes not the whole antigen but 4-7 critical amino acids within the epitope, the whole proteome can be represented by RPPDL which is commercially available from New England Biolabs (NEB). The RPPDL contains almost all possible 7-mer peptide sequences that can be screened against, for example, blood serum, and binding peptides will represent antibody response. The RPPDL are widely used for identification of the epitope specificity of monoclonal antibodies. The obstacle for using RPPDL as diagnostic tools was the necessity to sequence large number of individual phage DNA for identification of epitopes recognized by antibodies. The NGS technology makes possible to identify all the epitopes recognized by all antibodies contained in the human serum using one run of the sequencing machine.

Recently, RPPDL were applied to studies of antibody response to different diseases. In [55] the authors studied serum samples from patients with severe peanut allergy using phage display. The phage were selected based on their interaction with patient serum and characterized by highthroughput sequencing. The epitopes of a prominent peanut allergen, Ara h 1, in sera from patients were identified.

We used RPPDL to study antibody response to breast cancer and Lyme disease. The approach for the representation of serum antibody repertoire by RPPDL is outlined in the

flowchart in Figure 3.1. Ph.D-7 phage displayed library of 7-mer random peptides was mixed with the serum and incubated overnight. Phage bound to antibodies was isolated using protein-G beads and eluted from the beads using low pH buffer. The eluted phage was amplified by propagation in E.coli and the amplified library was incubated with the same serum. The phage bound to antibodies was isolated using protein-G beads and the phage DNA was PCR amplified with the primers flanking the peptide coding insert. The library of peptide coding inserts was sequenced by NGS and the DNA sequences were translated into the peptide sequences. Obtained peptides represent a mimotope profile of a serum sample - mimic of all epitopes recognized by all antibodies contained in the serum. Mimotope profiles of different samples can be studied and compared with each other.

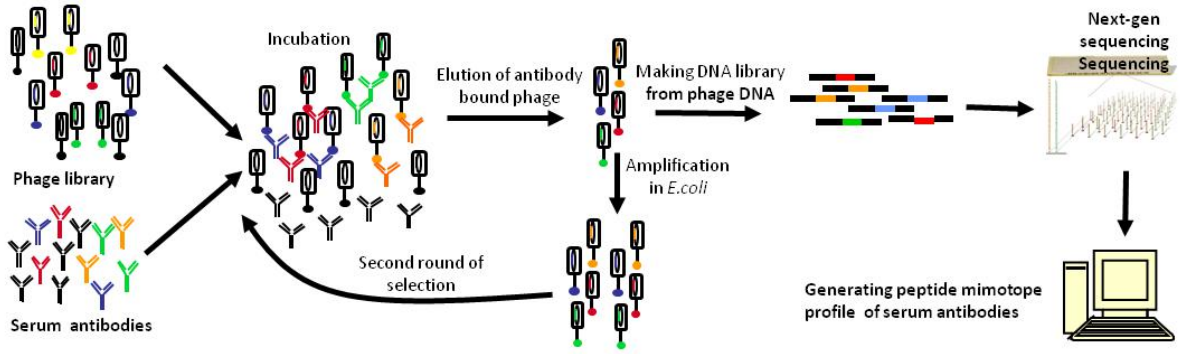

Figure 3.1 A scheme for generating mimotope profiles of serum antibody repertoire.

Generating peptide profiles. Twenty μl of serum and 10 μl of the enriched library were diluted in 200 μl of the Tris Buffered Saline (TBST) buffer containing 0.1% Tween 20 and 1% BSA and incubated overnight at room temperature. The phages bound to antibodies were isolated using low pH buffer as described above for the enrichment of the library and the phage DNA was isolated using phenol-chloroform extraction and ethanol precipitation. The 21 nt long DNA fragments coding for random peptides were PCR-amplified using primers containing a sequence for annealing to the Illumina flow cell, the sequence complementary to the Illumina sequencing primer and the 4 nt barcode

sequence for multiplexing. The PCR-amplified DNA library was purified on agarose ge-multiplexed and sequenced by 50 cycle HiSeq 2500 platform.

The sequences were de-multiplexed to determine its source sample. The 21- base nucleotides were extracted between base position 29 and 49 and translated to 7-amino-acid peptide using the first frame. Any peptide containing stop codon was discarded.

3.2 Methods for finding mimotope motifs

The profiles generated by NGS following several iterative round of affinity selection and amplification in bacteria can consist of millions of peptide sequences. A significant fraction of these sequences is not related to the repertoires of antibody specificities, but produced by nonspecific binding and preferential amplification in bacteria. The presence of high amounts of these unspecific, quickly growing ”parasitic” sequences can complicate the analysis of serum antibody specificities. The affinity selected sequences can be clustered into groups of similar sequences with shared consensus motifs, while the parasitic sequences are usually represented by single copies.

There are two main approaches for finding mimotope motifs. First one is based on clustering, second is based on BLAST alignment. Further in this chapter, we describe two proposed methods for motif finding based on CAST clustering and based on k-means clustering and apply these methods to real data. We also apply BLAST alignment method to Lyme disease NGS data to find epitopes and analyze immune response during the different stages of the disease disease.

3.3 Mimotope motif finding based on CAST clustering

We define motif as a group of peptides having common sequence pattern. If we consider a motif as a cluster formed by peptides with the center represented by a consensus sequence then construction of a motif corresponds to a difficult clustering problem with many closely located centers. The radius of a cluster may exceed the distance from one

cluster to another one. To solve the problem we modified CAST clustering algorithm (Clustering Affinity Search Technique) [22]. We did not know in advance how many motifs should be found in each sample. Other words, we did not know the number of clusters. For this reason we used CAST. It does not assume a given number of clusters and an initial spatial structure of them, but determines cluster number and structure based on the data.

The input of CAST consists of a similarity matrix to store the distances of all of the peptides and an similarity threshold. We defined the *similarity* of two sequences of equal length as the number of positions where the corresponding symbols are equal. We also consider the shifts of sequences relative to each other where it is necessary. For example, if we have two peptide sequences MLPHWAS and LPHWASK we need to shift them on one position relative to each other to get common overlap LPHWAS. In this example the similarity will be equal 6. Since the minimal length of a peptide sequence that can mimic the epitope recognized by antibody is usually in the range from 4 to 7 amino acids, we assigned similarity threshold equal 4. So any two peptides in a motif should have approximately 4 common amino acids (diameter of a motif). As well as no more than three shifts between peptides to the right or left sides were allowed.

The Algorithm 1 describes the CAST-based motif identification method (CMIM). On every iteration of the algorithm two peptides with the highest similarity were chosen as the initial center of a cluster. Next the process of adding and removing of peptides from the cluster was performed while the similarity between every pair of petides in a final set were not less than the threshold. During that step initially assigned central peptides could be removed. A measure of similarity between a peptide and all other peptides in a cluster was called *affinity*. Obtained cluster was saved removing its peptides from further consideration as initial centers. Then the procedure was repeated to find remaining motifs. Unlike CAST our algorithm allows intersection between clusters. As result some consensus sequences of motifs could be too close to each other. So the obtained clusters were collapsed if they had more than 50% common peptides. The last step was to align all

Algorithm 1 CAST-based motif identification (CMIM)

Input: Set of peptides P, similarity matrix D, threshold θ
Set of seed peptides $S \leftarrow P$
while $S \neq \emptyset$ **do**
  Cluster set $M \leftarrow \{s_1, s_2\}$, s_1, s_2 - the two most similar peptides in S
  Set of petides outside the cluster $R \leftarrow P \setminus M$
  $affinity(p) \leftarrow D(p, s_1) + D(p, s_2)$, for all $p \in P$
  while there is any change in M **do**
    while $\exists r \in R$ s.t. $affinity(r)/|M| \geq \theta$ **do**
      $M \leftarrow M \cup \{r'\}$, $r' \in R$ - peptide with the highest affinity
      $affinity(p) \leftarrow affinity(p) + D(p, r')$, for all $p \in P$ - update affinity of all peptides
    end while
    while $\exists m \in M$ s.t. $affinity(m)/(|M| - 1) < \theta$ **do**
      $M \leftarrow C \setminus \{m'\}$, $m' \in M$ - peptide with the lowest affinity
      $affinity(p) \leftarrow affinity(p) - D(p, m')$, for all $p \in P$ - update affinity of all peptides
    end while
  end while
  $S \leftarrow S \setminus M$
  Add M to set of clusters M
end while
for any pair $\{M', M''\} \in M$ **do**
  if $(|M' \cap M''|/|M'| > 0.5)$ or $(|M' \cap M''|/|M''| > 0.5)$ **then**
    Collapse M' and M''
  end if
end for
for any $M \in M$ **do**
  align peptides in M
  calculate entropy in every position i of aligned M
  find consensus K for 7-mer window with the min entropy
end for
Output: Set of motifs M, represented by clusters M_i and consensus sequences K_i

peptides in the cluster and compute entropy in every position. Seven positions with the smallest cumulative entropy (the most conserved part) were chosen, and the consensus amino acid sequence was found. The output of the algorithm was a set of finding motifs in a serum sample, each represented by a cluster and its consensus 7-mer sequence. To compute consensus sequence for a motif we aligned peptide sequences in its cluster and calculated entropy in every position of the cluster. Then we chose seven positions win-

dow with the minimum total entropy and identified consensus as the order of the most frequent amino acids found at each chosen position.

3.4 Application of CAST-based motif finding to breast cancer NGS data

3.4.1 Motivation

Cancer cells starts out as normal body cells, but they begin to grow out of control because of an abnormal gene expression. The immune system plays a major role in limiting the development of these abnormalities. There are multiple lines of evidence that the immune system elicits a detectable humoral immune response to changes in antigen profiles caused by growing cancer cells [56, 57, 58, 59, 60]. Circulating autoantibodies produced by the patient's own immune system after exposure to cancer proteins are promising biomarkers for the early detection of cancer. An advantage of autoantibodies in cancer detection is their production in large quantities, despite the presence of a relatively small amount of the corresponding antigen. It has been demonstrated, that panels of antibody reactivities can be used for detecting cancer with high sensitivity and specificity [61].

The current methods of analysis of antitumor humoral immune response, such as SEREX, SERPA, antigen microarrays, or ELISA are designed to detect high- affinity/high-titer IgG or IgM antibodies. However, the immune system can react to alterations in local antigenic compositions caused by growing tumors by pro- ducing a variety of low-affinity/low-titer antibodies. There is a need to develop more sensitive method. Recently the authors tested whether immunosignatures correspond to clinical classifications of disease using samples from people with brain tumors. The immunosignaturing platform distinguished not only brain cancer from controls, but also pathologically important features about the tumor including type and grade [62]. These results clearly demonstrate that random peptide arrays can be applied to profiling serum antibody repertoires for detection of cancer. The important advantage of using peptide arrays instead of antigen arrays is that peptides can mimic not only the protein epitopes but also the carbohydrate

epitopes that represent an essential part of the repertoire of cancer-associated autoantibodies.

3.4.2 Library enrichment

Studying immune response to breast cancer we decided to decrease the number of peptides in profiles generated by NGS without losing cancer-specific sequences.For generation of profiles we used the phage library enriched by panning on the pool of cancer sera. The approach for generating mimotope profiles of serum antibody repertoire is outlined in the flowchart in Figure 3.2. We extended a standard approach from Figure 3.1 by enriching the library by cancer specific peptides. First, the initial RPPDL was mixed with pooled cancer serum. After three rounds of incubation, elution and amplification a new library was obtained enriched by peptides specific for cancer. The second step was directly generating of mimotope profiles and NGS using the enriched library instead of the initial random library. Pooled serum from eight stage 0 breast cancer patients were used for enrichment of the library.

Generating peptide profiles with enrichment. The enrichment was performed as follows. Twenty μl of pooled serum and 10 μl of the Ph.D.7 random peptide library (NEB) were diluted in 200 μl of the Tris Buffered Saline (TBST) buffer containing 0.1% Tween 20 and 1% BSA and incubated overnight at room temperature. The phages bound to antibodies were isolated by adding 20 μl of protein G agarose beads (Santa Cruz) to the phage antibody mixture and incubating for 1 hour. To eliminate the unbound phage the mixture with beads was transferred to the well of 96-well MultiScreen-Mesh Filter plate (Millipore) containing 20 μm pore size nylon mesh at the bottom. The unbound phage was removed by applying vacuum to the outside of the nylon mesh using micropipette tip. The beads were washed 4 times by adding to the well 100 μl of TBST buffer and removing the liquid by applying vacuum to the outside of the nylon mesh using micropipette tip. The phage bound to the antibodies was eluted by adding to the beads of

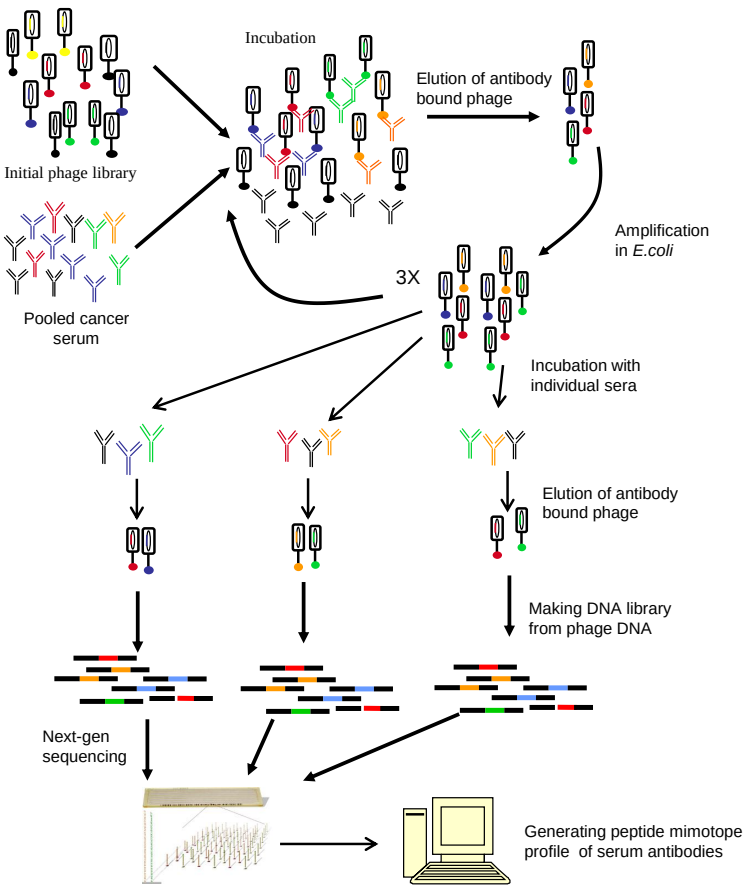

Figure 3.2 A scheme for generating mimotope profiles of serum antibody repertoire. The first step of the experiment is library enrichment, the second step is directly generating of mimotope profiles and NGS.

100 µl of 100 mM Tris-glycine buffer pH 2.2 followed by neutralization using 20 µl 1 M Tris buffer pH 9.1. The eluted phages were amplified in bacteria by infecting 3 ml of an early log-phase culture . The amplified phages were isolated by precipitating phage with $^1/_6$ volume of 20% PEG, 05.M NaCl precipitation buffer. The cycle of incubation-bound phage isolation-amplification was repeated two more times and the isolated after the 3rd amplification library was used for analyzing antibody repertoires.

3.4.3 Results

Data set. We analyzed the profiles generated for the 15 serum samples of the stage 0 and 1 breast cancer patients and for the 15 serum samples of the healthy donors. For each

serum sample the experiment was performed separately using the same enriched library on all samples. In average, for the experimental condition selected, the total number of distinct peptide sequences generated in one sample was 18,450, and standard deviation σ was 6205. The average count value (expression) of a sample was 407,335(σ = 252,393).

After applying CMIM separately to every sample, we obtained in average 3,000(1,073) motifs per a control sample and 3,490(1,315) motifs per a case sample. The average size of a motif in a case was 7.1(1.8) peptides, in a control it was 6.8(1.3) peptides. Every sample contained significant amount of large motifs. Thus, the average number of motifs consisting of 20 and more peptides was 154(71) and 131(53) for cases and controls respectively.

Motif validation. To validate found motifs we generated pseudo mimotope profiles using two strategies. The first strategy was random permutation of amino acids in a sample peptides. As result, we received 30 samples consisting of random 7-mer peptides. We ran our motif search method on the samples and obtained about 6,639(1,967) motifs with the average size 4.2(0.7). Although, the largest motif among all samples contained only 17 peptides. More than 95% of motifs in all samples had size no more than 4 peptides.The obtained motifs were significantly different from those found in real serum samples. This result proves the amino-acid order is meaningful in mimotope motifs found by CMIM.

The second strategy was random selection of peptides from existing samples and generating random samples. We collapse all original serum samples together assigning count value to each peptide. The more abundant and popular a peptide was among samples the more probable it would be selected to a new random sample. We generated 30 samples with 20K peptides each. We also applied motif search method to the random samples. In average we obtained 3,890(34) motifs with the size of 5.71(0.04) peptides. To compare the group of random samples with the group of real serum samples we applied Kruskal-Wallis test [63]. This non-parametric method determines whether samples originate from the same distribution. The result p-value was $7.5 * 10^{-5}$ rejecting the null hy-

pothesis that the population medians of both groups were equal. Thus, the single sample motifs are significantly different from motifs in peptides drawn from multiple samples.

Cancer-specific motifs. The cancer-specific motifs were defined as motifs significantly prevalent in cases. We compared motifs based on their consensus 7-mers. If two samples shared any consensus sequence, we considered they shared the corresponding motif. A motif was associated with cancer if probability of its appearance in cases against controls by chance was less than 0.05. We calculated the probability of all possible combinations of 15 cases and 15 controls and chose the most discriminating. As result, we received the following case-control significant combinations with probability less 0.05: 4-0 (a motif should appeared in 4 cases and 0 controls), 5-0, 6-0,...,15-0,6-1,...,15-1,8-2,...15-2,9-3,...15-3,10-4,...,15-4,11-5,...15-5,12-6,...,15-6,13-7,...,15-7,14-8,...,15-8,...,15-11. We also found the combinations with probability less than 0.04, 0.03, 0.02 and 0.01. There were 67 cancer specific motifs with probability of case-control appearance less than 0.05, 27 motifs with probability less than 0.04, 24 motifs with probability less than 0.03, 10 and 4 motifs with probability less than 0.02 and 0.01 respectively.

To validate obtained motifs we applied permutation test. We tested, at 5% significance level, whether the number of observed motifs can be obtained by chance. The test proceeded as follows. Cases and controls were randomly swapped, so some cases were considered as controls while controls were considered as cases. Totally 10K random permutations were performed. For every permutation the number of motifs with significant case-control appearance was count. The one-sided p-value of the test was calculated as the proportion of permutations where the number of significant motifs was greater or equal to observed number (see Table 3.1). As far as all p-values were greater than 0.05 we can not reject the hypothesis that the number of observed motifs could be obtained by chance.

The number of expected and observed motifs as well as False Discovery Rate (FDR) [64] adjustment are also shown in Table 3.1. Notice that the number of observed motifs

Table 3.1 Statistics for case-specific motifs. The number of observed motifs with expected number, FDR and p-value of the permutation test.

probability	observed	expected	FDR	p-value of the permutation test
<0.05	67	51.9	0.77	0.15
<0.04	27	20.5	0.76	0.21
<0.03	24	16.6	0.69	0.15
<0.02	10	8.1	0.81	0.32
<0.01	4	4.2	1.06	0.52

with probability of case-control appearance less than 0.01 equals to 4 which is less than expected number 4.2. That gives FDR greater than 1. Despite the fact that no motif is statistically significant, we can see that their number is still larger than expected.

3.4.4 Conclusion

In current work we identified cancer-specific motifs by analyzing peptide profiles of serum samples from cancer patients and from healthy donors. These profiles were generated using a phage DNA sequencing following single selection without amplification on the serum samples with the library enriched by the cycles of affinity selection-amplification using a pool of serum samples from additional cancer patients.

A novel motif identification method based on CAST clustering (CMIM) was proposed. We found that for any real serum sample the number of peptides per a motif is significantly greater comparing with pseudo epitope repertoire consisting of a randomly permuted peptides. Also the single sample motifs are shown to be significantly different from motifs in peptides drawn from multiple samples.

Running on case-control data CMIM identified cancer-specific motifs. Although no motif is statistically significant after permutation test, the number of found motifs is larger than expected and may therefore contain useful cancer markers.

3.5 Mimotope motif finding based on k-means clustering

Another clustering method is k-means. It requires less computing power than CAST but the number of clusters (motifs among peptides) should be predefined. For this reason, as an initial step, we propose to build a grid of peptides which pairwise similarity far from each other, less than some similarity threshold. The size of the grid is input parameter for k-means. All clusters are built around grid peptides (see Algorithm 2).

Algorithm 2 K-means-based motif identification

Input: Set of peptides P, similarity matrix D, threshold θ
1: Select a set of seeds S **from the set of peptides** $P = \{p_i, i = 0, \ldots |P|.\}$
$S \leftarrow \{p_0\}$
for $i = 1, \ldots, |P|$ do if for each $s \in S$ similarity of p_i and s is less than T_{seeds}, then $S \leftarrow S \cup p_i$
2: For each seed peptide s_i, initialize cluster $C = s_i$ and compute the profile $M(C)$ associated with C.
repeat
3: For each peptide $p_i \in P$ find the closest motif M and assign p_i to M's cluster.
4: For each cluster C do remove all peptides, add all peptides assigned to C, and update profile $M(C)$.
until no changes in clusters

3.6 Application of mimotope motif finding based on BLAST alignment to Lyme disease

3.6.1 Motivation

Lyme disease (LD), the most prevalent tick-borne illness in North America and Europe, is caused by spirochetes in the genus Borrelia. Borrelia burgdorferi, the principal human pathogen in the United States, is responsible for approximately 300,000 LD cases per year [65]. LD is problematic because early diagnosis is easily missed due to flu-like symptoms, which only transiently appear in humans during an early stage of disease. No preventable or therapeutic vaccine for humans is currently available.

The long-term survival of B. burgdorferi spirochetes in the mammalian host is achieved though the B. burgdorferi antigenic variation system [66]. This elaborate system, first identified on a 28-kb linear plasmid (lp28-1) of the B. burgdorferi B31 strain, is composed of a vlsE expression site and 15 noncoding silent cassettes. As a result of segmental conversion from the cassettes into the vlsE gene, variants of the VlsE (variable major protein-like sequence expressed) surface lipoprotein are generated [67, 68]. The vls-mediated variation of VlsE is absolutely required for persistence in mice as murine antibody clears the vls-deficient B. burgdorferi clone (ΔvlsE strain) or the B. burgdorferi clone with nonswitchable VlsE (sVlsE, for static VlsE) [69, 70, 71, 72, 73, 74]. Besides VlsE, however, B. burgdorferi expresses numerous other surface (lipo)proteins that, in contrast to VlsE, are invariant [75]. Antibody developed to non-VlsE surface antigens can protect mice from B. burgdorferi infection when variable VlsE is absent [74].

The central hypothesis that we tested was whether the protective efficacy of the antibody response to B. burgdorferi non-VlsE surface antigens declines as B. burgdorferi infection progresses in mice. The approach involving RPPDL and NGS was undertaken to test the hypothesis and to compare specificities of serum antibody developed during the early and late stages of B. burgdorferi infection. Although RPPDL have been widely used for mapping epitopes [76, 77, 78], the tool was previously applied for epitope discovery of B. burgdorferi proteins by only two studies [79, 80]. Specifically, we apply statistical analysis to detect any significant difference between mimotope profiles of mice at day-28 and day-70 of infection. To solve the problem of parasitic sequences we performed additional analysis based on BLAST clustering where considered only peptides mapped with high similarity to surface lipoproteins of Borrelia burgdorferi that are antigens and to vlsE protein that is responsible for persistence of the infection in a host.

3.6.2 Generating serum antibody repertoire profiles using RPPDL

The approach for generating mimotope profiles of serum antibody repertoire is outlined in the flowchart in Figure 3.1. Twenty microliters of mouse serum and 10 µl of

the Ph.D.-7 random peptide library were diluted in 200 μl of Tris-buffered saline (TBST) buffer containing 0.1% Tween 20 and 1% bovine serum albumin (BSA) and then incubated overnight at room temperature. The phages bound to antibodies were isolated by applying 20 μl of protein G-agarose beads (Santa Cruz Biotechnology, Inc., TX, USA) to the phage-antibody mixture for 1 h. To eliminate unbound phages, the mixture with beads was transferred to a 96-well MultiScreen-Mesh filter plate (EMD Millipore, MA, USA) containing a 20-μm-pore-size nylon mesh on the bottom. Unbound phages were removed by applying a vacuum to the outside of the nylon mesh. The beads were washed four times with 100 μl of TBST buffer per well. Antibody-bound phages were eluted with 100 μl of 100 mM Tris-glycine buffer (pH 2.2). Then the buffer was replaced with 20 μl of 1 M Tris buffer (pH 9.1). The eluted phages were amplified by infecting bacteria according to the manufacturer's instructions. Amplified phages were subjected to two additional rounds of biopanning. Antibody-bound phages were isolated using protein G-agarose beads. DNA was isolated via phenol-chloroform extraction and ethanol precipitation. The 21-nucleotide (nt)-long DNA fragments coding random peptides were then PCR amplified using the following forward and reverse primers, respectively:

5′ - AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCT TCCGATCT(INDEX)TGGTACCTTTCTATTCTCACTCT - 3′ and

5′ - CAAGCAGAAGAGGGCATACGAGCTCTTCCGATCTAACAG TTTCGGCCGA ACCTCCACC - 3′.

The INDEX in the sequence of the forward primer indicates a 6-nt barcode, which allows sequencing multiple libraries using a single line of the Illumina flow cell. For each mouse serum, a distinct forward primer with unique index sequence was used. The multiplexed PCR-amplified DNA library was then purified on agarose gel and sequenced using an Illumina HiSeq 2500 platform.

As a result of sequencing, a total of about 116 million DNA reads were obtained. Reads were demultiplexed based on the barcodes. Each read contained a unique index sequence of 6-nt in length and a 21-nt sequence coding a random peptide: 5′-

(INDEX)GTGGTACCTTTCTATTCTCACTCT(21-nt sequence)G - 3'. Then the 21-nt sequences were extracted from each read between positions 30 and 50 and translated to 7-mer peptides in the first frame. Peptides that contained stop codons were not included in the analysis. The average number of all peptides per serum sample was approximately 1×10^7. The number of distinct peptides identified was approximately 1.4×10^5 per sample. The data was then analyzed via the Python programming language.

3.6.3 Method for comparison of mimotope profiles on days 28 and 70 of B. burgdorferi infection

The strength of association between a peptide and a serum sample was measured as follows. A peptide, P, was associated with day 28 serum if $X(P)$, the lowest frequency of P among day 28 serum samples, was higher than $Y(P)$, the highest frequency of P among day 70 serum samples. The strength of association was then measured by the size of the gap: $X(P) - Y(P)$. Similarly, a peptide was associated with day 70 serum samples if its smallest frequency among day 70 serum samples was higher than the highest frequency among day 28 serum samples. The statistical significance of the difference between the number of peptides associated with the day 28 serum and the number of peptides associated with the day 70 serum was measured using a permutation test. The permutation test was used because of the comparatively small number of samples (five serum samples per each time point). For each of 252 possible permutations, the difference between the numbers of associated peptides was found and compared with the actual difference.

3.6.4 BLAST alignment of peptides to VlsE and other B. burgdorferi proteins

As far as significant fraction of peptide sequences in a profile is noise, the additional analysis was performed where only peptides mapped to Borrelia's surface lipoproteins and vlsE were considered. For each serum sample, all peptides were mapped to VlsE of B. burgdorferi 297 (297-VlsE) (GenBank accession number AB041949.1) or B31 (B31-VlsE) (GenBank accession number AAC45733.1) strains using blastp with an identity threshold

of 4 (i.e., only alignments with at least four exact amino acid matches were taken into account). For each VlsE position X, a peptide with the amino acid matched to position X and with K different VlsE matches contributed its frequency divided by K to the coverage of X. Overall, the coverage of X, $C(X)$, was computed as the sum of contributions of all peptides matched to X. Similarly, all peptides were mapped to other B. burgdorferi surface proteins of the two B. burgdorferi strains, decorin-binding proteins A (DbpA), DbpB, and P35.

3.6.5 Results

The data analyses revealed that only 366 and 14 peptides, respectively, were associated with the day 28 and day 70 serum samples. The difference between the two numbers was statistically significant, with a p-value estimated via a permutation test being below 0.25%. However, when a multiple testing correction was applied, no significant difference in antibody repertoires was identified between the two time points of B. burgdorferi infection.

In order to compare anti-VlsE antibody responses between days 28 and 70 p.i., for each samples all peptides were mapped to the strain 297 VlsE (297-VlsE) or B31-VlsE via blastp analysis (Figure 3.3). The epitope mapping against the linear B31-VlsE structure showed no significant difference between the reactivities of day 28 and day 70 antibodies. Predicted cross-reactivity of anti-297-VlsE antibody to linear B31-VlsE is partially consistent with the anti-VlsE antibody reactivity of LD patients [81]. Microarray-based epitope mapping demonstrated that IgG antibody of patients with chronic Lyme disease were mainly reactive to six peptides of B31-VlsE. The immunodominant epitopes were located within two invariable domains (VlsE residues 21 to 31, 61, 96, and 336 and 343) and one variable domain (VlsE residues 196 and 271 to 291). Consistently, reactivity of anti-297 antibody was predicted to invariant region 6 (IR6). IR6 has been shown to be highly immunogenic in humans, monkeys, and mice [82, 83, 84]. In addition to IR6, C3H mice may also develop a strong antibody response to IR2 and IR4 [84]. However, reactivity to these

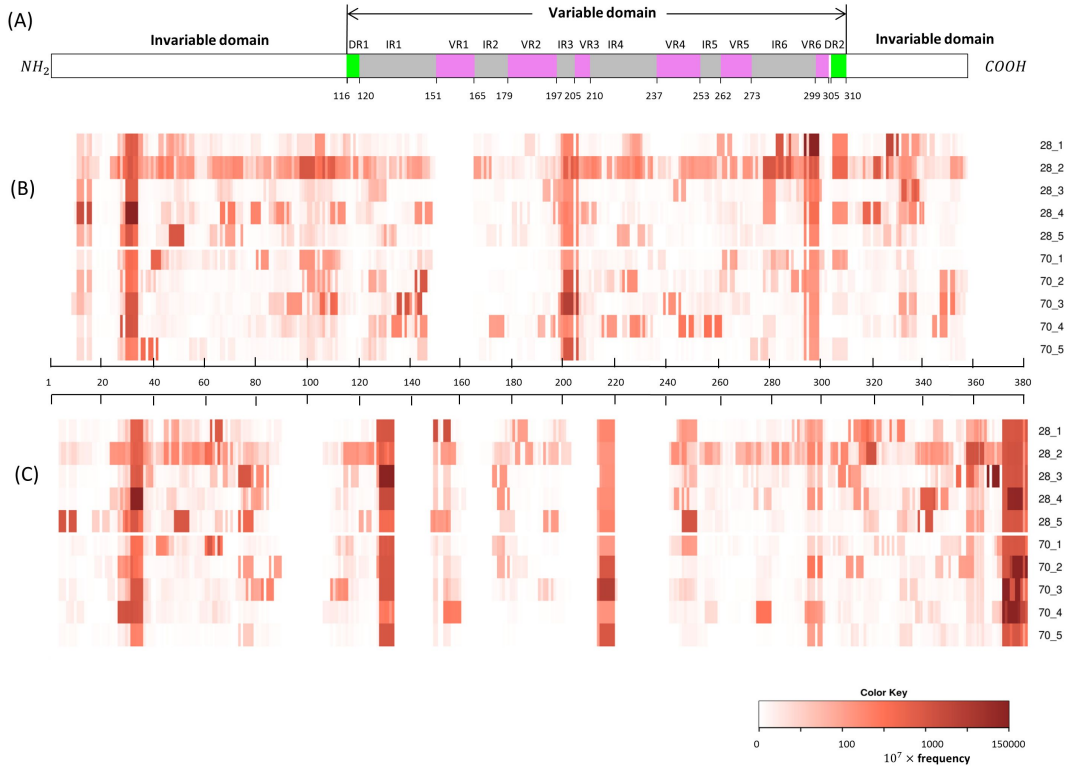

Figure 3.3 Epitope mapping of VlsE. Primary structure of B31-VlsE illustrating two direct repeats (DR1 and DR2; green) that demarcate one variable domain and two invariable domains. Shown are also six invariable (IR; gray) and variable (VR; pink) regions [1] (A). Heat maps were generated from predicted reactivity of anti-297 antibody to the primary structure of B31-VlsE (B) and 297-VlsE (C). Anti-297 sera were harvested from B. burgdorferi persistently infected C3H mice at days 28 and 70 postinfection (five animals per time point). The linear B31-VlsE structure is scaled to the B31-VlsE heat map.

conserved regions was not pronounced in the present study (Figure 3.3).

Interestingly, IR2 and IR4 are not antigenic in humans and monkeys [84]. The analysis also predicted strong reactivity of anti-297 antibody to the C-terminal invariable domain (amino acids 372 to 380) within the primary structure of 297-VlsE as opposed to that of B31-VlsE. Murine antibodies were consistently developed against the C-terminal region during the entire infection period as reactivity was detected in all mouse sera taken at day 28 and day 70 p.i. This fully supports the previous findings that the C-terminal invariable domain is highly immunodominant [85] but shows limited antigenic conser-

vation among B. burgdorferi strains [86]. It was previously shown that the C-terminal domain sequences of 11 B. burgdorferi strains were not identical to the C-terminal sequence of the B31 strain [86]. Likewise, due to a high degree of divergence with 46% identity and 53% similarity between B31-VlsE and 297-VlsE [66], strong antibody reactivity was predicted only to IR1 of 297-VlsE. Finally, no significant difference between the reactivities of day 28 and day 70 antibodies to the primary structures of translated cassettes vls2 to vls16 (vls2-vls16) [67] was identified.

Similar to computational mapping of VlsE epitopes, identified peptides were also mapped to decorin-binding protein A (DbpA), decorin-binding protein (DbpB), and P35. These B. burgdorferi surface proteins were shown to be immunogenic and afforded protection in mice against B. burgdorferi infection [87, 88, 89]. Consistently, no significant difference in antibody reactivities to contiguous epitopes of these immunogenic proteins was detected between day 28 and day 70 serum samples (Figures 3.4, 3.5).

3.6.6 Conclusion

In this study, mimotope profiles derived from day 28 and 70 mice serum samples were compared to each other. Since mimotopes are peptides that may mimic both continuous and discontinuous antigens of a different nature (e.g., proteins, polysaccharides, and lipids), the global mimotope comparison considered a variety of B. burgdorferi epitopes, including lipid and carbohydrate epitopes. However, the comparison revealed no major difference in antibody repertoires between the two time points of infection. Similarly, no significant changes were identified between the two antibody repertoires when identified mimotopes were mapped against primary structures of VlsE, DbpA, DbpB, and P35. Given, however, that many antibody epitopes of native proteins are discontinuous, the mapping inherently underestimated repertoires of antibody developed to conformational epitopes of these B. burgdorferi surface proteins. Moreover, the BLAST-based analyses of antibody repertoires missed mimotopes that mimicked lipid and carbohydrate epitopes.

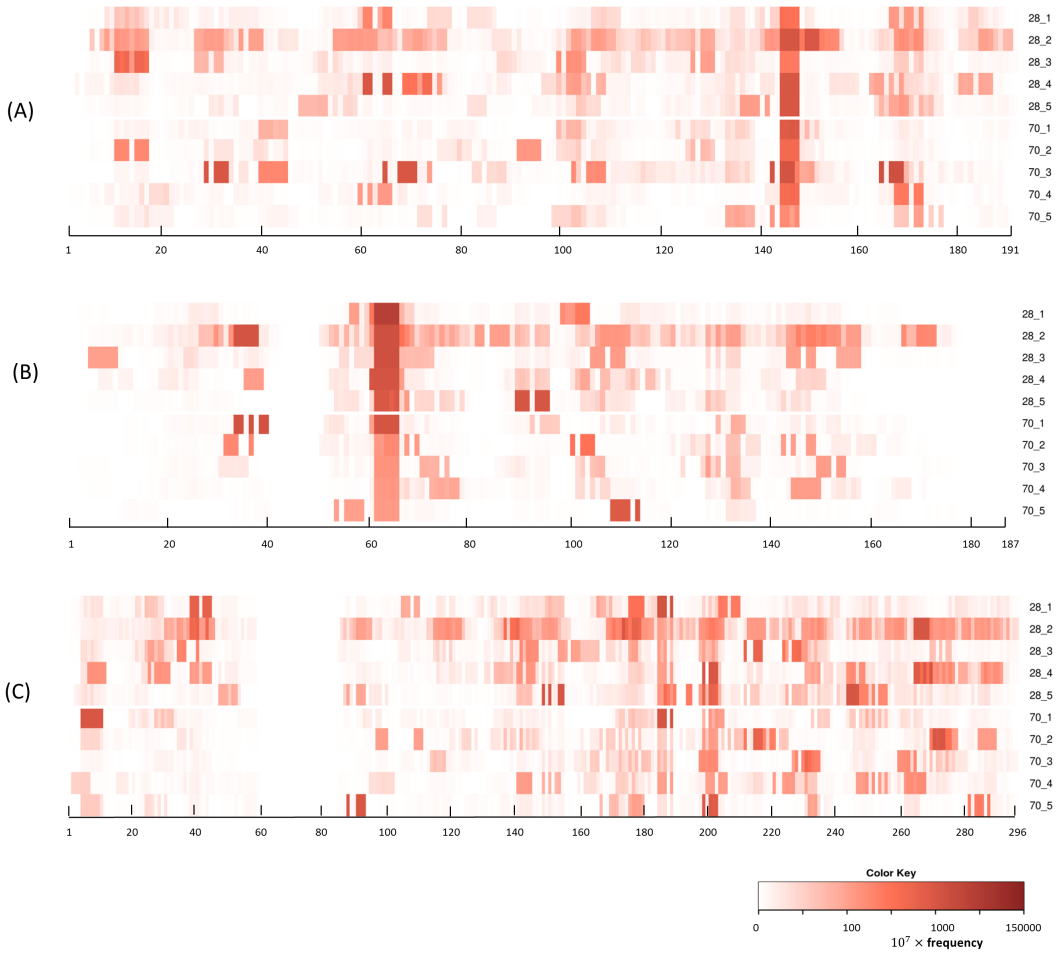

Figure 3.4 Epitope mapping of non-VlsE surface proteins of Borrelia burgdorferi B31. Heat maps generated from in silico prediction of anti-297 antibody reactivity to the primary structure of decorin-binding protein A (DbpA), decorin-binding protein B (DbpB), and P35 (panels A, B, and C, respectively). Anti-297 sera were harvested from Bb-infected C3H mice at day 28 and 70 post infection (5 animals per time point).

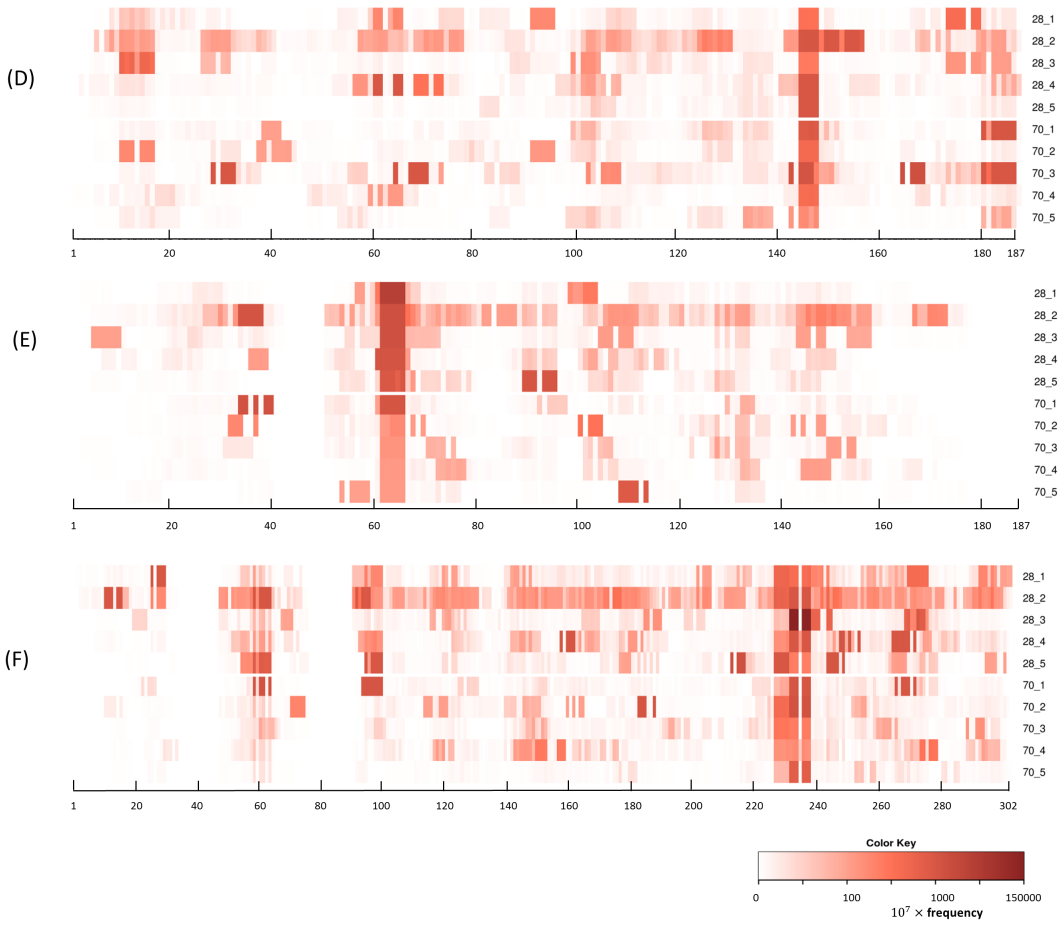

Figure 3.5 Epitope mapping of non-VlsE surface proteins of Borrelia burgdorferi 297. Heat maps generated from in silico prediction of anti-297 antibody reactivity to the primary structure of decorin-binding protein A (DbpA), decorin-binding protein B (DbpB), and P35 (panels D, E, and F, respectively). Anti-297 sera were harvested from Bb-infected C3H mice at day 28 and 70 post infection (5 animals per time point).

PART 4

RNA-SEQ DATA PROCESSING FOR IMMUNE RESPONSE ANALYSIS

4.1 Introduction

RNA-Seq is an approach to transcriptome profiling using NGS technologies. The transcriptome is the complete set of transcripts in a cell, and their quantity, for a specific developmental stage or physiological condition. RNA-Seq is the first sequencing-based method that allows the entire transcriptome to be surveyed in a very high-throughput and quantitative manner [20]. Depending on whether there is a disease in an organism or not, the transcriptional activity of genes involved in immune response is different. One particularly powerful advantage of RNA-Seq is that it can capture this difference. Although, like other high-throughput sequencing technologies, RNA-Seq requires development of special methods for data analysis. Many variations of RNA-seq protocols and analyses have been published, but there is still no optimal pipeline.

In this work, we developed a pipeline for whole transcriptomic analysis of well controlled systems of rubella infected cells. Rubella virus is a small enveloped virus with a positive single-strand RNA genome of 9.7 kb from the family *Togaviridae*, genus *Rubivirus*. Rubella is an important human pathogen because of its ability to cross placenta and establish persistent infection in a fetus causing miscarriage, stillbirth or multiple birth defects called congenital rubella syndrome. This syndrome is only associated with rubella infections during the first trimester of pregnancy. There is a vaccine from rubella. Live attenuated vaccine RA27/3, which is currently used by the majority of the countries, induces life-long immunity and highly effective in preventing post-natal rubella infection and congenital rubella syndrome. Using RNA-seq technology, we compared innate antiviral responses induced by wild type rubella virus isolates and RA27/3 vaccine strain in primary cultures of human umbilical vein endothelial cells (HUVEC).

4.2 Library preparation and sequencing

RNA extracted from infected cells are assessed using Bioanalyzer (Agilent Technologies). The quantity of total RNA is also measured using Qubit fluorometer (Thermo Fisher). Each sample is spiked with Control RNA developed by External RNA Controls Consortium (ERCC; Thermo Fisher) prior to ribosomal RNA depletion (NEB).

Strand-specific library is prepared using the NEBNext@ UltraTM Directional RNA Library Prep Kit for Illumina (NEB). Quantity and quality of each library is assessed by Bioanlyzer, Qubit and real-time PCR (NEBNext@ Library Quant Kit for Illumina). The library is sequenced using Illumiina MiSeq. As result, we obtain paired-end raw reads.

4.3 Differential gene expression analysis of RNA-seq data

We proposed the following pipeline for whole transcriptomic analysis for rubella virus infected cells:

1. Quality control of paired-end reads using Sickle [90]. Sickle is a windowed adaptive trimming tool for FASTQ files using quality. It uses sliding windows along with quality and length thresholds to determine when quality is sufficiently low to trim the 3′-end of reads and also determines when the quality is sufficiently high enough to trim the 5′-end of reads. It will also discard reads based upon the length threshold.

2. Trimming adapters using Cutadapt [91]. The tool finds and removes adapter sequences, primers, poly-A tails and other types of unwanted sequence from high-throughput sequencing reads.

3. Mapping trimmed reads to human genome using TopHat2 [92]. The tool is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra-high throughput short read aligner Bowtie2 [93], and then analyzes the mapping results to splice junctions between exons.

4. Counting reads aligned to human genes using htseq-count tool [94]. Given a file with aligned sequencing reads and a list of genomic features - genes, a task was to count how many reads map to each feature. In the case of RNA-Seq, each gene is considered as the union of all its exons.

5. Finding biological replicates outliers. The read counts for all replicates should be well correlated with each other. Although, there is no specific threshold for correlation. Pearson correlation can be used in this case. A replicate that is different from other replicates should be disregarded.

6. Finding genes deferentially expressed in samples under different conditions. EdgeR [95] tool for finding differentially expressed genes are applied. EdgeR is a differential expression analysis of RNA-seq expression profiles with biological replication. It implements a range of statistical methodology based on the negative binomial distributions, including empirical Bayes estimation, exact tests, generalized linear models and quasi-likelihood tests. EdgeR requires biological replicates. For experiments without replicates DEGseq [96] tool can be used.

4.4 Results

In this work we compare cells in different conditions: uninfected HUVEC cells, rubella wild virus (DeZhou strain) infected cells, rubella vaccine virus (RA27/3 strain) infected cells. All RNAs were obtained from cells harvested 2, 7 and 14 days after infection. The goal was to trace the development of infection over time, compare one condition in different days, and compare uninfected cells with infected cells and wild rubella infected cells with vaccine infected cells.

For every condition in each time point we had 6 biological replicates. Outliers in all groups of replicates were found using pairwise Pearson correlation. We noticed that each group of replicated were divided into two highly correlated (>0.9) groups. The correlation between those two groups was usually less that 0.85. Table shows correlation be-

tween replicates for wild rubella infected cells on day 7. Bold font represents correlation greater than 0.9. Replicates 1,2 and 3 form one highly correlated group, and replicates 4,5 and 6 form another group. Differential gene expression analysis was performed between conditions separately for each highly correlated replicate group.

Table 4.1 Pairwise Pearson correlation between biological replicates for wild rubella infected cells on day 7. Bolt text represents correlation greater than 0.9

Biological replicates	**replicate_1**	**replicate_2**	**replicate_3**	**replicate_4**	**replicate_5**	**replicate_6**
replicate_1	**1**	**0.975**	**0.957**	0.675	0.785	0.719
replicate_2	**0.975**	**1**	**0.959**	0.600	0.725	0.652
replicate_3	**0.957**	**0.959**	**1**	0.739	0.842	0.788
replicate_4	0.675	0.600	0.739	**1**	**0.966**	**0.979**
replicate_5	0.785	0.725	0.842	**0.966**	**1**	**0.992**
replicate_6	0.719	0.652	0.788	**0.979**	**0.992**	**1**

As results, there were identified 33 differentially expressed (DE) genes in wild rubella infected cells between days 2 and 7, and 13 DE genes between days 2 and 14. 5 DE genes were identified for vaccine infected cells between days 2 and 7. 180 DE genes were identified between uninfected cells and wild rubella infected cells on day 2. 152 DE genes were identified between uninfected cells and vaccine rubella infected cells on day 2. 115 DE genes were identified between uninfected cells and wild rubella infected cells on day 7. 153 DE genes were identified between uninfected cells and vaccine rubella infected cells on day 7. No significant difference were detected between uninfected cells and infected cells on day 14 and there was no difference between vaccine and wild type rubella on any time point.

PART 5

FUTURE WORK

We plan to apply pathways analysis to RNA-seq rubella data.

REFERENCES

[1] C. Eicken, V. Sharma, T. Klabunde, M. B. Lawrenz, J. M. Hardham, S. J. Norris, and J. C. Sacchettini, "Crystal structure of lyme disease variable surface antigen vlse of borrelia burgdorferi," *Journal of Biological Chemistry*, vol. 277, no. 24, pp. 21 691–21 696, 2002.

[2] E. Domingo and J. Holland, "Rna virus mutations and fitness for survival," *Annual Reviews in Microbiology*, vol. 51, no. 1, pp. 151–178, 1997.

[3] M. L. Metzker, "Sequencing technologiesâĂŤthe next generation," *Nature reviews genetics*, vol. 11, no. 1, pp. 31–46, 2010.

[4] A. Töpfer, O. Zagordi, S. Prabhakaran, V. Roth, E. Halperin, and N. Beerenwinkel, "Probabilistic inference of viral quasispecies subject to recombination," *Journal of Computational Biology*, vol. 20, no. 2, pp. 113–123, 2013.

[5] M. Marz, N. Beerenwinkel, C. Drosten, M. Fricke, D. Frishman, I. L. Hofacker, D. Hoffmann, M. Middendorf, T. Rattei, P. F. Stadler *et al.*, "Challenges in rna virus bioinformatics," *Bioinformatics*, p. btu105, 2014.

[6] P. Skums, Z. Dimitrova, D. S. Campo, G. Vaughan, L. Rossi, J. C. Forbi, J. Yokosawa, A. Zelikovsky, and Y. Khudyakov, "Efficient error correction for next-generation sequencing of viral amplicons," *BMC bioinformatics*, vol. 13, no. 10, p. 1, 2012.

[7] 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.

[8] 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.

[9] 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.

[10] J. A. Baaijens, A. Z. El Aabidine, E. Rivals, and A. Schönhuth, "De novo assembly of viral quasispecies using overlap graphs," *Genome Research*, vol. 27, no. 5, pp. 835–848, 2017.

[11] X. Liu, Q. Hu, S. Liu, L. J. Tallo, L. Sadzewicz, C. A. Schettine, M. Nikiforov, E. N. Klyushnenkova, and Y. Ionov, "Serum antibody repertoire profiling using in silico antigen screen," *PloS one*, vol. 8, no. 6, p. e67181, 2013.

[12] G. P. Smith, "Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface," *Science*, vol. 228, no. 4705, pp. 1315–1317, 1985.

[13] M. Andreatta, O. Lund, and M. Nielsen, "Simultaneous alignment and clustering of peptide data using a gibbs sampling approach," *Bioinformatics*, vol. 29, no. 1, pp. 8–14, 2012.

[14] E. Gerasimov, A. Zelikovsky, I. Măndoiu, and Y. Ionov, "Identification of cancer-specific motifs in mimotope profiles of serum antibody repertoire," *BMC bioinformatics*, vol. 18, no. 8, p. 244, 2017.

[15] A. S. Rogovskyy, D. C. Gillis, Y. Ionov, E. Gerasimov, and A. Zelikovsky, "Antibody response to lyme disease spirochetes in the context of vlse-mediated immune evasion," *Infection and immunity*, vol. 85, no. 1, pp. e00 890–16, 2017.

[16] R. L. Bratton, J. W. Whiteside, M. J. Hovan, R. L. Engle, and F. D. Edwards, "Diagnosis and treatment of lyme disease," in *Mayo Clinic Proceedings*, vol. 83, no. 5. Elsevier, 2008, pp. 566–571.

[17] A. R. Marques, "Lyme disease: a review," *Current allergy and asthma reports*, vol. 10, no. 1, pp. 13–20, 2010.

[18] T. Kim, M. S. Tyndel, H. Huang, S. S. Sidhu, G. D. Bader, D. Gfeller, and P. M. Kim, "Musi: an integrated system for identifying multiple specificity from very large peptide or nucleic acid data sets," *Nucleic acids research*, vol. 40, no. 6, pp. e47–e47, 2011.

[19] A. Krejci, T. R. Hupp, M. Lexa, B. Vojtesek, and P. Muller, "Hammock: a hidden markov model-based peptide clustering algorithm to identify protein-interaction consensus motifs in large datasets," *Bioinformatics*, vol. 32, no. 1, pp. 9–16, 2015.

[20] Z. Wang, M. Gerstein, and M. Snyder, "Rna-seq: a revolutionary tool for transcriptomics," *Nature reviews genetics*, vol. 10, no. 1, pp. 57–63, 2009.

[21] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang *et al.*, "A survey of best practices for rna-seq data analysis," *Genome biology*, vol. 17, no. 1, p. 13, 2016.

[22] A. Ben-Dor, R. Shamir, and Z. Yakhini, "Clustering gene expression patterns," *Journal of computational biology*, vol. 6, no. 3-4, pp. 281–297, 1999.

[23] A. S. Lauring, J. Frydman, and R. Andino, "The role of mutational robustness in rna virus evolution," *Nature Reviews Microbiology*, vol. 11, no. 5, pp. 327–336, 2013.

[24] A. S. Lauring and R. Andino, "Quasispecies theory and the behavior of RNA viruses," *PLoS pathogens*, vol. 6, no. 7, p. e1001005, 2010.

[25] A. M. Tsibris, B. Korber, R. Arnaout, C. Russ, C.-C. Lo, T. Leitner, B. Gaschen, J. Theiler, R. Paredes, Z. Su *et al.*, "Quantitative deep sequencing reveals dynamic HIV-1 escape and large population shifts during CCR5 antagonist therapy in vivo," *PloS one*, vol. 4, no. 5, p. e5683, 2009.

[26] M. R. Henn, C. L. Boutwell, P. Charlebois, N. J. Lennon, K. A. Power, A. R. Macalalad, A. M. Berlin, C. M. Malboeuf, E. M. Ryan, S. Gnerre *et al.*, "Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection," *PLoS pathogens*, vol. 8, no. 3, p. e1002529, 2012.

[27] M. L. Metzker, "Sequencing technologiesâĂŤthe next generation," *Nature Reviews Genetics*, vol. 11, no. 1, pp. 31–46, 2009.

[28] S. Mangul, N. C. Wu, N. Mancuso, A. Zelikovsky, R. Sun, and E. Eskin, "Accurate viral population assembly from ultra-deep sequencing data," *Bioinformatics*, vol. 30, no. 12, pp. i329–i337, 2014.

[29] O. Zagordi, A. Töpfer, S. Prabhakaran, V. Roth, E. Halperin, and N. Beerenwinkel, "Probabilistic inference of viral quasispecies subject to recombination," in *Research in Computational Molecular Biology*. Springer, 2012, pp. 342–354.

[30] 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, pp. 237–249, 2011.

[31] 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.

[32] Y. Heo, X.-L. Wu, D. Chen, J. Ma, and W.-M. Hwu, "Bless: bloom filter-based error correction solution for high-throughput sequencing reads," *Bioinformatics*, vol. 30, no. 10, pp. 1354–1362, 2014.

[33] M. H. Schulz, D. Weese, M. Holtgrewe, V. Dimitrova, S. Niu, K. Reinert, and H. Richard, "Fiona: a parallel and automatic strategy for read error correction," *Bioinformatics*, vol. 30, no. 17, pp. i356–i363, 2014.

[34] E. Marinier, D. G. Brown, and B. J. McConkey, "Pollux: platform independent error correction of single and mixed genomes," *BMC bioinformatics*, vol. 16, no. 1, p. 1, 2015.

[35] B. H. Bloom, "Space/time trade-offs in hash coding with allowable errors," *Communications of the ACM*, vol. 13, no. 7, pp. 422–426, 1970.

[36] A. E. Minoche, J. C. Dohm, H. Himmelbauer *et al.*, "Evaluation of genomic high-throughput sequencing data generated on illumina hiseq and genome analyzer systems," *Genome Biol*, vol. 12, no. 11, p. R112, 2011.

[37] D. Cavanagh, "Coronavirus avian infectious bronchitis virus," *Veterinary research*, vol. 38, no. 2, pp. 281–297, 2007.

[38] D. Vijaykrishna, G. Smith, J. Zhang, J. Peiris, H. Chen, and Y. Guan, "Evolutionary insights into the ecology of coronaviruses," *Journal of virology*, vol. 81, no. 8, pp. 4012–4020, 2007.

[39] S. Liu, X. Zhang, Y. Wang, C. Li, Z. Han, Y. Shao, H. Li, and X. Kong, "Molecular characterization and pathogenicity of infectious bronchitis coronaviruses: complicated evolution and epidemiology in china caused by cocirculation of multiple types of infectious bronchitis coronaviruses," *Intervirology*, vol. 52, no. 4, pp. 223–234, 2009.

[40] M. Jackwood, D. Hilt, and S. Callison, "Detection of infectious bronchitis virus by real-time reverse transcriptase-polymerase chain reaction and identification of a quasispecies in the beaudette strain," *Avian diseases*, vol. 47, no. 3, pp. 718–724, 2003.

[41] E. T. McKinley, D. A. Hilt, and M. W. Jackwood, "Avian coronavirus infectious bronchitis attenuated live vaccines undergo selection of subpopulations and mutations following vaccination," *Vaccine*, vol. 26, no. 10, pp. 1274–1284, 2008.

[42] M. W. Jackwood, D. A. Hilt, A. W. McCall, C. N. Polizzi, E. T. McKinley, and S. M. Williams, "Infectious bronchitis virus field vaccination coverage and persistence of arkansas-type viruses in commercial broilers," *Avian diseases*, vol. 53, no. 2, pp. 175–183, 2009.

[43] W. Nix, D. Troeber, B. Kingham, C. Keeler Jr, and J. Gelb Jr, "Emergence of subtype strains of the arkansas serotype of infectious bronchitis virus in delmarva broiler chickens," *Avian diseases*, pp. 568–581, 2000.

[44] J. Archer, M. S. Braverman, B. E. Taillon, B. Desany, I. James, P. R. Harrigan, M. Lewis, and D. L. Robertson, "Detection of low-frequency pretherapy chemokine (cxc motif) receptor 4-using hiv-1 with ultra-deep pyrosequencing," *AIDS (London, England)*, vol. 23, no. 10, p. 1209, 2009.

[45] C. Hoffmann, N. Minkah, J. Leipzig, G. Wang, M. Q. Arens, P. Tebas, and F. D. Bushman, "Dna bar coding and pyrosequencing to identify rare hiv drug resistance mutations," *Nucleic acids research*, vol. 35, no. 13, p. e91, 2007.

[46] J. Simons, M. Egholm, J. Lanza, G. Turenchalk, B. Desany, M. Ronan, J. Knight, L. Du, J. Leamon, J. Rothberg *et al.*, "Ultra-deep sequencing of hiv from drug resistant patients," in *Antiviral Therapy*, vol. 10, no. 4. INT MEDICAL PRESS LTD 2-4 IDOL LANE, LONDON EC3R 5DD, ENGLAND, 2005, pp. S157–S157.

[47] A. Tsibris, C. Russ, W. Lee, R. Paredes, R. Arnaout, T. Honan, P. Cahill, C. Nusbaum, and D. Kuritzkes, "Detection and quantification of minority hiv-1 env v3 loop sequences by ultra-deep sequencing: preliminary results," in *ANTIVIRAL THERAPY*, vol. 11, no. 5. INT MEDICAL PRESS LTD 2-4 IDOL LANE, LONDON EC3R 5DD, ENGLAND, 2006, pp. S74–S74.

[48] 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.

[49] E. C. Holmes and B. T. Grenfell, "Discovering the phylodynamics of rna viruses," *PLoS Comput Biol*, vol. 5, no. 10, p. e1000505, 2009.

[50] S. Prabhakaran, M. Rey, O. Zagordi, N. Beerenwinkel, and V. Roth, "Hiv-haplotype inference using a constraint-based dirichlet process mixture model," in *Machine Learning in Computational Biology (MLCB) NIPS Workshop*, 2010, pp. 1–4.

[51] W. Brockman, P. Alvarez, S. Young, M. Garber, G. Giannoukos, W. L. Lee, C. Russ,

E. S. Lander, C. Nusbaum, and D. B. Jaffe, "Quality scores and snp detection in sequencing-by-synthesis systems," *Genome research*, vol. 18, no. 5, pp. 763–770, 2008.

[52] 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.

[53] P. Skums, Z. Dimitrova, D. S. Campo, G. Vaughan, L. Rossi, J. C. Forbi, J. Yokosawa, A. Zelikovsky, and Y. Khudyakov, "Efficient error correction for next-generation sequencing of viral amplicons," *BMC bioinformatics*, vol. 13, no. 10, p. 1, 2012.

[54] W.-P. Lee, M. P. Stromberg, A. Ward, C. Stewart, E. P. Garrison, and G. T. Marth, "Mosaik: a hash-based algorithm for accurate next-generation sequencing short-read mapping," *PloS one*, vol. 9, no. 3, p. e90581, 2014.

[55] A. Christiansen, J. V. Kringelum, C. S. Hansen, K. L. Bøgh, E. Sullivan, J. Patel, N. M. Rigby, T. Eiwegger, Z. Szépfalusi, F. De Masi *et al.*, "High-throughput sequencing enhanced phage display enables the identification of patient-specific epitope motifs in serum," *Scientific reports*, vol. 5, 2015.

[56] G. P. Dunn, A. T. Bruce, H. Ikeda, L. J. Old, and R. D. Schreiber, "Cancer immunoediting: from immunosurveillance to tumor escape," *Nature immunology*, vol. 3, no. 11, pp. 991–998, 2002.

[57] G. P. Dunn, L. J. Old, and R. D. Schreiber, "The immunobiology of cancer immunosurveillance and immunoediting," *Immunity*, vol. 21, no. 2, pp. 137–148, 2004.

[58] S. Gnjatic, E. Ritter, M. W. Büchler, N. A. Giese, B. Brors, C. Frei, A. Murray, N. Halama, I. Zörnig, Y.-T. Chen *et al.*, "Seromic profiling of ovarian and pancreatic cancer," *Proceedings of the National Academy of Sciences*, vol. 107, no. 11, pp. 5088–5093, 2010.

[59] M. Ho, R. Hassan, J. Zhang, Q.-c. Wang, M. Onda, T. Bera, and I. Pastan, "Humoral

immune response to mesothelin in mesothelioma and ovarian cancer patients," *Clinical Cancer Research*, vol. 11, no. 10, pp. 3814–3820, 2005.

[60] A. Sreekumar, B. Laxman, D. R. Rhodes, S. Bhagavathula, J. Harwood, D. Giacherio, D. Ghosh, M. G. Sanda, M. A. Rubin, and A. M. Chinnaiyan, "Humoral immune response to α-methylacyl-coa racemase and prostate cancer," *Journal of the National Cancer Institute*, vol. 96, no. 11, pp. 834–843, 2004.

[61] L. Zhong, S. P. Coe, A. J. Stromberg, N. H. Khattar, J. R. Jett, and E. A. Hirschowitz, "Profiling tumor-associated antibodies for early detection of non-small cell lung cancer," *Journal of Thoracic Oncology*, vol. 1, no. 6, pp. 513–519, 2006.

[62] A. K. Hughes, Z. Cichacz, A. Scheck, S. W. Coons, S. A. Johnston, and P. Stafford, "Immunosignaturing can detect products from molecular markers in brain cancer," *PloS one*, vol. 7, no. 7, p. e40201, 2012.

[63] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American statistical Association*, vol. 47, no. 260, pp. 583–621, 1952.

[64] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 289–300, 1995.

[65] A. F. Hinckley, N. P. Connally, J. I. Meek, B. J. Johnson, M. M. Kemperman, K. A. Feldman, J. L. White, and P. S. Mead, "Lyme disease testing by large commercial laboratories in the united states," *Clinical Infectious Diseases*, p. ciu397, 2014.

[66] S. J. Norris, "The vls antigenic variation systems of lyme disease borrelia: eluding host immunity through both random, segmental gene conversion and framework heterogeneity," *Microbiology spectrum*, vol. 2, no. 6, 2014.

[67] J.-R. Zhang, J. M. Hardham, A. G. Barbour, and S. J. Norris, "Antigenic variation

in lyme disease borreliae by promiscuous recombination of vmp-like sequence cassettes," *Cell*, vol. 89, no. 2, pp. 275–285, 1997.

[68] S. J. Norris, "Antigenic variation with a twist–the borrelia story," *Molecular microbiology*, vol. 60, no. 6, pp. 1319–1322, 2006.

[69] J. E. Purser and S. J. Norris, "Correlation between plasmid content and infectivity in borrelia burgdorferi," *Proceedings of the National Academy of Sciences*, vol. 97, no. 25, pp. 13 865–13 870, 2000.

[70] M. Labandeira-Rey, J. Seshu, and J. T. Skare, "The absence of linear plasmid 25 or 28-1 of borrelia burgdorferi dramatically alters the kinetics of experimental infection via distinct mechanisms," *Infection and immunity*, vol. 71, no. 8, pp. 4608–4613, 2003.

[71] R. Iyer, O. Kalu, J. Purser, S. Norris, B. Stevenson, and I. Schwartz, "Linear and circular plasmid content in borrelia burgdorferi clinical isolates," *Infection and immunity*, vol. 71, no. 7, pp. 3699–3706, 2003.

[72] M. B. Lawrenz, R. M. Wooten, and S. J. Norris, "Effects of vlse complementation on the infectivity of borrelia burgdorferi lacking the linear plasmid lp28-1," *Infection and immunity*, vol. 72, no. 11, pp. 6577–6585, 2004.

[73] T. Bankhead and G. Chaconas, "The role of vlse antigenic variation in the lyme disease spirochete: persistence through a mechanism that differs from other pathogens," *Molecular microbiology*, vol. 65, no. 6, pp. 1547–1558, 2007.

[74] A. S. Rogovskyy and T. Bankhead, "Variable vlse is critical for host reinfection by the lyme disease spirochete," *PloS one*, vol. 8, no. 4, p. e61226, 2013.

[75] M. R. Kenedy, T. R. Lenhart, and D. R. Akins, "The role of borrelia burgdorferi outer surface proteins," *FEMS Immunology & Medical Microbiology*, vol. 66, no. 1, pp. 1–19, 2012.

[76] S. E. Cwirla, E. A. Peters, R. W. Barrett, and W. J. Dower, "Peptides on phage: a vast library of peptides for identifying ligands." *Proceedings of the National Academy of Sciences*, vol. 87, no. 16, pp. 6378–6382, 1990.

[77] J. K. Scott and G. P. Smith, "Searching for peptide ligands with an epitope library," *Science*, vol. 249, no. 4967, pp. 386–390, 1990.

[78] A. Ryvkin, H. Ashkenazy, L. Smelyanski, G. Kaplan, O. Penn, Y. Weiss-Ottolenghi, E. Privman, P. B. Ngam, J. E. Woodward, G. D. May *et al.*, "Deep panning: steps towards probing the igome," *PLoS One*, vol. 7, no. 8, p. e41469, 2012.

[79] G. A. Kouzmitcheva, V. A. Petrenko, and G. P. Smith, "Identifying diagnostic peptides for lyme disease through epitope discovery," *Clinical and diagnostic laboratory immunology*, vol. 8, no. 1, pp. 150–160, 2001.

[80] C. V. Hamby, M. Llibre, S. Utpat, and G. P. Wormser, "Use of peptide library screening to detect a previously unknown linear diagnostic epitope: proof of principle by use of lyme disease sera," *Clinical and diagnostic laboratory immunology*, vol. 12, no. 7, pp. 801–807, 2005.

[81] A. Chandra, N. Latov, G. P. Wormser, A. R. Marques, and A. Alaedini, "Epitope mapping of antibodies to vlse protein of borrelia burgdorferi in post-lyme disease syndrome," *Clinical immunology*, vol. 141, no. 1, pp. 103–110, 2011.

[82] F. T. Liang, A. L. Alvarez, Y. Gu, J. M. Nowling, R. Ramamoorthy, and M. T. Philipp, "An immunodominant conserved region within the variable domain of vlse, the variable surface antigen of borrelia burgdorferi," *The Journal of Immunology*, vol. 163, no. 10, pp. 5566–5573, 1999.

[83] F. T. Liang and M. T. Philipp, "Epitope mapping of the immunodominant invariable region of borrelia burgdorferi vlse in three host species," *Infection and immunity*, vol. 68, no. 4, pp. 2349–2352, 2000.

[84] ——, "Analysis of antibody response to invariable regions of vlse, the variable surface antigen of borrelia burgdorferi," *Infection and immunity*, vol. 67, no. 12, pp. 6702–6706, 1999.

[85] M. T. Philipp, L. C. Bowers, P. T. Fawcett, M. B. Jacobs, F. T. Liang, A. R. Marques, P. D. Mitchell, J. E. Purcell, M. S. Ratterree, and R. K. Straubinger, "Antibody response to ir6, a conserved immunodominant region of the vlse lipoprotein, wanes rapidly after antibiotic treatment of borrelia burgdorferi infection in experimental animals and in humans," *Journal of Infectious Diseases*, vol. 184, no. 7, pp. 870–878, 2001.

[86] F. T. Liang, L. C. Bowers, and M. T. Philipp, "C-terminal invariable domain of vlse is immunodominant but its antigenicity is scarcely conserved among strains of lyme disease spirochetes," *Infection and immunity*, vol. 69, no. 5, pp. 3224–3231, 2001.

[87] E. Fikrig, S. W. Barthold, W. Sun, W. Feng, S. R. Telford, and R. A. Flavell, "Borrelia burgdorferi p35 and p37 proteins, expressed in vivo, elicit protective immunity," *Immunity*, vol. 6, no. 5, pp. 531–539, 1997.

[88] K. E. Hagman, P. Lahdenne, T. G. Popova, S. F. Porcella, D. R. Akins, J. D. Radolf, and M. V. Norgard, "Decorin-binding protein of borrelia burgdorferi is encoded within a two-gene operon and is protective in the murine model of lyme borreliosis," *Infection and immunity*, vol. 66, no. 6, pp. 2674–2683, 1998.

[89] M. S. Hanson, D. R. Cassatt, B. P. Guo, N. K. Patel, M. P. McCarthy, D. W. Dorward, and M. Höök, "Active and passive immunity against borrelia burgdorferi decorin binding protein a (dbpa) protects against infection," *Infection and immunity*, vol. 66, no. 5, pp. 2143–2153, 1998.

[90] N. Joshi and J. Fass, "Sickle: A sliding-window, adaptive, quality-based trimming tool for fastq files (version 1.33)[software]," *Available at httpslégitbubcoménajoshiéchkleQOll*, 2011.

[91] M. Martin, "Cutadapt removes adapter sequences from high-throughput sequencing reads," *EMBnet. journal*, vol. 17, no. 1, pp. pp–10, 2011.

[92] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg, "Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions," *Genome biology*, vol. 14, no. 4, p. 1, 2013.

[93] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with bowtie 2," *Nature methods*, vol. 9, no. 4, pp. 357–359, 2012.

[94] S. Anders, P. T. Pyl, and W. Huber, "Htseq–a python framework to work with high-throughput sequencing data," *Bioinformatics*, p. btu638, 2014.

[95] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edger: a bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, 2010.

[96] L. Wang, Z. Feng, X. Wang, X. Wang, and X. Zhang, "Degseq: an r package for identifying differentially expressed genes from rna-seq data," *Bioinformatics*, vol. 26, no. 1, pp. 136–138, 2009.