BIOINFORMATIC APPLICATIONS FOR VIROLOGY RESEARCH

LEE WAH HENG CHARLIE

(B.Sc Computing (Hons.), NUS)

A THESIS SUBMITTED

FOR THE DEGREE OF PHD OF COMPUTING
SCHOOL OF COMPUTING
NATIONAL UNIVERSITY OF SINGAPORE

2010

ACKNOWLEDGEMENTS

The research presented in this thesis has been carried out at the School of Computing at the National University of Singapore. The main setting of this research work has been at the Genome Institute of Singapore, who provided me with the Ph.D scholarship. I would like to express my gratitude to Professor Edison Tak-Bun Liu, M.D. Executive Director of the Genome Institute of Singapore, for the facilities, funding and the help provided during my work. I am most grateful to my supervisor, Associate Professor Ken Sung Wing-Kin, Ph.D, School of Computing, National University of Singapore, for the invaluable advice that helped me overcome the challenges I've faced during the course of the Ph.D. I would also like to thank him for his understanding and support that allowed me to work and pursue my Ph.D concurrently without conflict.

I would like to acknowledge the support, advice and assistance rendered from my colleagues in the Genome Institute of Singapore. In particular, I would like to thank Dr. Christopher Wong, Chief Scientific Officer for Biomarker Development, and Dr. Martin Hibberd, Ph.D, Senior Group Leader, Assoc Director, Infectious Diseases, for their guidance and expert advice on my research. Special thanks go to my closest colleagues in the Computational and Mathematical Biology group and the Biomarker Flagship, all of whom made working in the Genome Institute of Singapore a wonderful experience.

Lastly, I would like to thank the evaluators and oral panel, namely Associate Professor Donald Seto, Ph.D, Microbial Genomics and Diversity, Bioinformatics, George Mason University, Associate Professor David Hsu, Ph.D, School of Computing, National University of Singapore, Professor Wong Lim Soon, Ph.D, Head of Computer Science, School of Computing,

National University of Singapore, and Associate Professor Anthony Tung, Ph.D, School of Computing, National University of Singapore, for their valuable comments and report on my thesis.

TABLE OF CONTENTS

SUMMARY

The primary objective of this dissertation is to address a number of key challenges and issues in the detection, resequencing and evolutionary analysis of viruses. Using novel ideas to improve upon existing approaches, it aims to develop better technologies and bioinformatics tools that would have a greater impact on clinical decision-making.

Amplification of viral genomes is a necessary first step of diagnosis and sequence analysis. The thesis explores the pitfalls of using specific primers for amplification and proposes to use random-tagged primers, particularly for amplification of unknown viruses. Although it is theoretically possible for random-tagged primers to bind to any sequence, the blind use of such primers without careful design does not guarantee genome-wide amplification of the virus. In the second chapter, the thesis introduces a model to predict amplification efficiency of random-tagged primers and developed an algorithm, LOMA, to design random-tagged primers with optimal amplification efficiency. Experiments show that the random-tagged primers generated by LOMA can amplify up to 90% of the genomes of the target viruses.

In the third chapter, the thesis argues the advantages of using DNA microarrays for diagnostics over traditional PCR methods. To increase the sensitivity and specificity of microarray diagnostics, the thesis makes use of random-tagged primers for amplification and proposes an algorithm (PDA) that analyzes the distribution of probe signal intensities of in-silico recognition signatures probe sets of each virus based on a novel weighted Kullback-Leibler divergence that is sensitive to the tail of the distribution. Validation experiments show that PDA is able to accurately detect and identify co-infections of multiple viruses, as well as unknown viruses initially missed by PCR tests.

In the fourth chapter, the thesis demonstrates the feasibility of using resequencing microarrays as a large scale bio-surveillance tool. In the wake of the 2009 H1N1 influenza pandemic, a novel resequencing kit that is capable of interrogating all eight segments of the H1N1 2009 influenza, with accommodation for mutation hotspots, was developed. The accompanying base-calling software EvolSTAR is a new method that utilizes neighbourhood hybridization intensity profiles and substitution bias of probes on the microarray for mutation confirmation and recovery of ambiguous base queries. Validation experiments show that EvolSTAR can achieve a much higher accuracy and call rate than existing competing methods.

The fifth chapter discusses the role that recombination plays in the emergence of novel or more virulent strains of viral pathogens. Understanding the mechanisms of viral evolution will aid in the development of better anti-viral drugs, vaccines, as well as diagnostics and surveillance tools. The thesis presents an algorithm (RB-Finder) that uses a more informative distance metric that overcomes the inaccuracies of methods that uses base-by-base comparisons. Experiments show that RB-Finder is able to achieve accuracies comparable to the most accurate phylogeny-based methods but within a much shorter time. In addition, RB-Finder is able to distinguish regions of high mutation rates from recombination breakpoints.

In summary, the thesis has contributed several technologies and novel methods that have significantly improved existing bioinformatics approaches in virology research.

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 The Threat of Human Viruses

Viruses are one of the main classes of microscopic agents which cause infectious disease in humans. They are made up of genetic material known as DNA or RNA and require a host to survive. By means of proteins on its outermost surface, a virus can recognize and attach itself to the appropriate host cells. The virus then multiplies by tricking the healthy cell to duplicate the viral nucleic acid as well as enzymes needed by the virus for enveloping and coat protein formation, killing or altering the functions of the cell in the process. New viruses are released via lysis of the host cells and begin infecting other cells, causing disease.

Viruses have the potential to spread rapidly in a locality or even worldwide and infect a large proportion of the human population. One of the most widespread disease affecting humans is influenza, commonly referred to as the flu, caused by RNA viruses of the family Orthomyxoviridae. They affect millions of people and result in the deaths of hundreds of thousands worldwide annually. However, there is another virus, the human immunodeficiency virus (HIV) that causes acquired immunodeficiency syndrome (AIDS) in humans and it has become one of the most serious health threat to the human population over the last few decades. Believed to have originated from non-human primates [1], HIV targets the T cells of the immune system, severely weakening it. As the immune system weakens, common organisms such as bacteria and viruses become fatal as the body can no longer defend against them. Since its discovery in 1981, AIDS has infected an estimated 33.4 million people and killed more than 25

million people [2]. Besides causing diseases, viruses are also responsible for certain cancers. Recently, a new virus, Merkel cell polyomavirus, has been identified to cause cancer in humans [3]. This adds to the list of five other cancer-linked viruses: papillomaviruses [4], human T-lymphotropic virus Type I [5], hepatitis B virus [6], Epstein-Barr virus [7] and Kaposi's Sarcoma Herpesvirus [8]. While there is much circumstantial evidence that viruses can cause certain types of cancer, it is still unclear exactly how viruses trigger cancer. It is most likely that tumors are formed when viruses integrated their genetic material with that of the host cells [9].

To reduce disease mortality and risk of certain cancers in humans, early detection of viral infections is vital. As such, there has been continual development of virological tests to provide fast, accurate and cost-effective diagnosis. So far, these virological tests have proven to be essential for the management of viral infections and administration of treatment.

The genetic arms race between viruses and host cells is never-ending. As host cells produce stronger immune responses to counteract the invading viruses, viruses evolve to enhance their ability to infect. Mechanisms of viral evolution include point mutations, genome rearrangements, as well as recombination and translocation events that may result in gene acquisition, gene creation and gene deletion [10]. As viruses evolve, they may become new variants or novel viruses with unpredictable virulence. In some cases, these new viruses become so virulent that they resulted in pandemics with high levels of mortality. For example, influenza viruses evolve into new strains almost yearly via mutations or re-assortment of their genes with other flu viruses. One such new strain of influenza A virus of subtype H1N1 is responsible for the 1918 Spanish flu pandemic, the worst pandemic in history that resulted in the deaths of over 50 million people [11]. Other notable influenza pandemics include the 1957 Asian flu pandemic, 1968 Hong Kong flu pandemic and most recently, the 2009 H1N1 pandemic. Besides influenza,

other novel or mutant viruses can also cause serious viral outbreaks. In 2003, a novel coronavirus emerged from the Guangdong province in China and caused the Severe Acute Respiratory Syndrome (SARS) outbreak that claimed the lives of 774 individuals in 37 countries around the world in a matter of weeks [12]. Based on historical data since 1901, studies have shown that on average, more than two new species of human virus are reported every year [13]. By fitting a statistical model to these data, it is predicted that 10 to 40 new species of virus will be discovered by 2020 [14]. As such, the health threat that these new viruses will present cannot be overlooked.

Early detection and continual biosurveillance of viruses, as well as understanding their evolution, are the solutions for preventing viral pandemics and controlling emerging infectious diseases. Over the years, a myriad of technology and methods have been developed to detect, obtain and analyze the genetic information of viruses to understand their virulence and evolution. This thesis presents new tools and methods that improve upon existing approaches.

1.2 Diagnostic and Detection of Viruses

In viral diagnosis, a virus is detected directly via detection of its proteins or nucleic acids, or indirectly via an immunological response to the virus. Generally, each method presents a different set of pros and cons with respect to sensitivity, efficiency and feasibility. Depending on the circumstances at which the viral sample was collected, virus type and concentration, certain methods may be more effective than others. As such, timely, accurate and sensitive detection of viruses is still difficult today.

One widely-used method is to “grow” the virus in cell cultures [15]. When the virus infects a host cell, the host cell may undergo changes such as cell rounding, disorientation, swelling, shrinking, or death. These cytopathic effects may be a defense mechanism used by the host cell against the virus or induced by the virus to enhance its survivability and reproduction. Since different viruses produce different cytopathic effects on different cells, the identity of the virus may be deduced from the cytopathic effects observed. Although large concentrations of viral products can be obtained in a successful culture, the process is labour-intensive and may potentially take up to 4 weeks to complete. Contamination is always a risk and sensitivity is often poor as it depends largely on the compatibility between the live virus and the cell lines chosen. Furthermore, a number of viruses such as hepatitis B, parvovirus and papillomavirus will not grow in cell cultures.

Another widely-used approach to detect viruses is to detect the presence of antigens or antibodies in bodily fluids. Such antibodies are typically produced in response to an infection and can be detected using techniques such as direct or immune fluorescence assays or enzyme immuno assays (ELISA). However, the effectiveness varies with different viruses. For example, ELISA has proven to be a highly sensitive test for HIV [16]. On the other hand, such an approach would not be useful in the diagnosis of certain viruses such as respiratory viruses, enteroviruses and diarrhoeal viruses because antibodies are produced only after the onset of clinical disease [17].

A direct method to detect viruses is to view them using electron microscopes. This method requires a purified or high concentration of virus that can be obtained directly from clinical sample. An experienced technician then discerns the virus by its physical structure features. Viruses such as poxviruses and herpesviruses can be easily identified using this

technology [18]. The high cost in equipment and maintenance and low sensitivity are the major drawbacks of this method.

In recent years, molecular methods based on the detection of the viral genome have been touted as the future direction of viral diagnosis [19]. One of the most commonly used methods for virus detection is polymerase chain reaction (PCR) [20]. PCR is used to amplify a single or few copies of the target nucleic sequence many folds using synthetic oligonucleotides flanking the target nucleic sequence, generating thousands to millions of copies of the target sequence. Detection of sequence product of the PCR assay may be achieved by gel electrophoresis. Although PCR is highly sensitive (may detect down to one viral genome per sample volume) and fast, the selection of suitably specific oligonucleotides (primers) that flank the target nucleic sequence may be difficult. Furthermore, sequence information of the suspected viruses must also be known in order to select the primers for amplification. Consequently, PCR cannot be used to detect novel viruses. A promising technology that has began to establish itself as an important diagnostic tool is the microarray or DNA chip [21]. A microarray consists of thousands, even millions, of fluorescence-labeled nucleic acid probes that bind (hybridize) with high specificity to complementary sequences of nucleic acid. By analyzing the microarray data, the virus present in the sample is easily identified since only probes that are complementary to the sequences of the virus will show high levels of fluorescence. Using microarrays, it is now possible to detect a large number of viruses at one time by designing and including specific probes complementary to sequences of all viruses of interest. Thus, compared to traditional methods, microarrays are far less reliant on clinical prediction of the infectious source for diagnosis [22]. Unfortunately, microarrays are susceptible to non-specific hybridization noise and its sensitivity is usually not higher than traditional detection methods such as cell cultures, antibody-based detection and

PCR. In order to make microarrays a sensitive and reliable viral detection tool, there is much interest in developing new sample preparation and hybridization protocols, as well as analysis methods to overcome the shortcomings of microarrays.

1.3 Resequencing of Viruses

Sequencing of viral genomes is historically performed using standard dye termination technologies. In dye-terminator (capillary) sequencing, negatively-charged DNA fragments are labeled with fluorescent dyes and applied with a high voltage to make them move through capillaries filled with polymer toward a positively-charged electrode in the sequencing machine. A laser beam is then shone on the DNA fragments just before they reach the positive electrode. The laser beam causes the dyes on the fragments to fluoresce. These fluorescence signals are detected by an optical device and converted to digital data. Since each dye emits light at different wavelengths, the four different bases (A, C, G, T) can be detected and distinguished [23]. Although capillary sequencing produces highly accurate base calls, it is slow and costly. The first 15-40 bases and 700-900 bases thereafter of the sequences generated tend to be poor quality.

The drawbacks of traditional capillary sequencing have motivated the development of high-throughput sequencing technologies that are capable of producing millions of sequences by parallelization of the sequencing process [24]. This dramatically lowers the cost of DNA sequencing with respect to the amount of throughput, prompting many researchers to utilize high-throughput sequencing technologies such as 454 sequencing [25], sequencing by oligonucleotide ligation and detection (SOLiD) [26], and Solexa sequencing [27] for a variety of genome projects. High-throughput sequencing technologies are best suited to provide deep

sequencing data of a few samples. In our experience with the 454 system, much of the amplified material is still human (as the bulk of the patient sample material is human RNA with very little influenza RNA), requiring very deep sequencing to obtain a complete flu genome sequence, with one compartment of a run not yielding sufficient viral information. Furthermore, assembly of the sequence fragments is required before any analysis can be done. Any abnormalities or gaps in the assembly would then require additional runs of 454, incurring more cost and time. Hence, they may be ill-suited for use in viral outbreaks, where the impetus is to obtain complete genome sequences from as many infected individuals as possible to monitor them for potential mutations or recombinations that might affect drug resistance or virulence.

Contrary to the above mentioned sequencing technologies, sequencing by hybridization is a novel non-enzymatic method in which a solution of target DNA sequence is fluorescently labeled and hybridized to a microarray containing short known sequences (probes) [28]. A combinatorial method is then used to reconstruct the DNA sequence from probes with strong hybridization signals that bind to the target sequence [29]. Since the introduction of sequencing by hybridization in the late 1980s, continual research and advancement over the years have alleviated some of the limitations of resequencing arrays, such as cost, accuracy and high-throughput processing. Currently, a single multiplexed resequencing array is capable of generating viral genomic sequences from multiple infected individuals. This translates to dramatic savings in cost, labour and time taken to continually obtain relatively high quality viral genomic sequences from infected individuals needed for evolutionary surveillance and studies in viral outbreaks. One major concern of using microarrays for resequencing is their susceptibility to non-specific hybridizations that may result in inaccurate sequence calls. As such, more research is needed to develop methods to analyze microarray data that may be noisy, so as to

achieve accuracies comparable to the “gold standard” traditional capillary sequencing techniques.

1.4 Recombination Detection in Viruses

Recombination is an important evolutionary mechanism for the continual survival of viruses. In addition to the ability to change its own genetic material rapidly via mutation, recombination allows viruses to add the capacity to exchange genetic material with one another, and to acquire genes from their hosts. Such capability enables viruses to (1) remove deleterious genes from their genomes and (2) create and spread advantages traits in an efficient manner [30]. A number of studies have shown that viruses do benefit from the effects of recombination. For example, a study on Sindbis viruses demonstrated that weaker strains of certain viruses can recombine to form stronger, more infectious strains [31]. Some viruses such as the bovine viral diarrhoea virus can also generate new variations by borrowing genetic material from their hosts [32]. Viruses with high recombination rates are a serious threat to human health. Antiviral drugs or vaccines for some viruses, such as influenza viruses, that target certain proteins have to be updated every few years to ensure they do not become less effective or even obsolete once these viruses undergo mutation or recombination [33]. For some viruses with very high evolutionary rates such as HIV, drugs or vaccines development have proven to be extremely difficult and may even be impossible without significant breakthrough in understanding virus evolution mechanisms [34].

The identification of the locations of the recombination events is the first step to an accurate phylogenetic analysis, which gives us important clues on the origins, pathogenicity and

treatments of viruses. This led to the development of numerous sequence analysis methods and phylogenetic techniques that have proven to be effective and accurate for detecting and characterizing recombination events among viruses. Traditionally, an accurate and reliable multiple sequence alignment of a given set of sequences is an essential starting point for many tools that analyze evolution, including phylogenetics and recombination detection [35]. The primary focus of a multiple sequence alignment is to identify, within several related sequences, regions that are highly conserved in identity or similarity, and therefore probably have functional and/or structural significance. Conversely, sequences that share a common ancestor but have since diverged may have clusters of mismatches and gaps that indicate the time since they diverged from one another. As such, a multiple sequence alignment of the query sequences is the *de facto* input for recombination detection tools.

The main goal of recombination detection tools is to find recombination breakpoints, the exact locations where a recombination event occurs in a sequence. Generally, the detection of breakpoints depends on the strength of a recombination event, which is affected by factors such as the mutation rate and the time at which the recombination event took place [36]. The unpredictable conditions at which recombination events occur make the task of finding breakpoints difficult. For recombination events which make little changes to the sequence, the detection of breakpoints may even be impossible [37]. Thus, in order to pinpoint the precise locations where recombination events may have taken place, regardless of when they happened or frequency, and to test their correctness, it is vital to develop sensitive and accurate methods for detecting recombination.

1.5 Literature Survey

1.5.1 Diagnostic and Detection of Viruses

DNA microarrays have become an essential tool in clinical diagnostics. In recent years, microarrays have been used to detect and subtype a multitude of human viruses such as herpesviruses [38], respiratory viruses [39], human rotaviruses [40], papillomaviruses [41], orthopoxviruses [42], hepatitis [43], cytomegalovirus [44], and the influenza virus [45, 46]. Comprehensive pan-viral detection microarrays have also aided in the discovery of novel viruses [46, 47].

Though virus detection by microarray is a young field, a number of different platforms and approaches have been described, each with important attributes. For example, the array described by Wang et. al. [46] is based on probes designed to recognize the most conserved viral domains, facilitating the detection of a taxonomic fingerprint that provides powerful clues to viral identity with minimal probe usage. Lin et. al. [39], on the other hand, described a probe-dense resequencing array capable of detecting a smaller set of predefined pathogens, but with higher detection specificity, including the ability to discern highly related subtypes. Thus, it is important to devise a probe design strategy that allows us to detect viruses in accordance to the intended use of the detection array.

The accuracy of virus detection depends largely on the algorithms used to analyze microarray data. Simple algorithms usually determine whether a particular virus is present or not depending on the relative abundance of probes with high hybridization intensities [45, 47, 48]. The pathogen predictions made by these simple algorithms may be affected by systematic and cross-hybridization errors frequently experienced by microarrays. Unfortunately, few other

algorithms exist. One such algorithm that has been reported and validated, E-predict [46], matches hybridization signatures with predicted pathogen signatures derived from the theoretical free energy of hybridization for each microarray probe. E-predict is less geared towards identifying and distinguishing specific pathogen strains, and aimed more at elucidating the best possible candidates as supported by the available probes. As such, this approach is particularly advantageous in situations where the sequence of the virus is not fully known. Another approach [49] uses tiling resequencing microarrays to obtain sequence fragments from consecutive (≥ 3) high confidence base calls. BLAST is then used to match these sequence fragments to a public database of viruses for diagnosis. However, due to the high number of probes required for tiling the viral genomes, this approach can only detect a limited number of viruses.

1.5.2 Resequencing of Viruses

Resequencing microarrays offer a low-cost, efficient and high-throughput solution to obtain whole-genome primary sequences of viruses for practical large-scale bio-surveillance and epidemiology studies. For example, resequencing microarrays have been used to generate complete sequences of the severe acute respiratory syndrome (SARS) coronavirus [50, 51] and poxviruses [52]. Recently, resequencing microarrays have also been used successfully to generate primary sequences for highly dangerous biothreat agents such as filoviruses of the Ebola Zaire group, or the Machupo and Lassa arenaviruses [53].

Base-calling is a critical step in analyzing resequencing microarrays. There are two main base-calling software commonly used in the reported studies, namely the ABACUS algorithm [54] and Nimblescan PBC algorithm [50]. Both algorithms employ a gain-of-signal approach

[55] based on relative hybridization to allele-specific probes complementary to each of the four possible nucleotides at interrogated nucleotide position for base-calling. A probability based on the degree of differentiation of hybridization intensities among the querying probes is then computed for each base call. A base call is of high confidence if they exceed a pre-defined significance or probability threshold. This approach is statistically sound but is susceptible to a myriad of factors such as mutations, random noise, probe quality, sample quality and experimental conditions. These artifacts can cause poor hybridization performance, resulting in ambiguous and sometimes false-positive calls.

Efforts have been made to improve the call rates and accuracies of existing base-calling algorithms. For example, Zhan and Kulp used sequences of the probe and target to predict probe intensities in resequencing microarrays [56]. By accounting for probes may be noisy, they were able to achieve a higher call rate and accuracy than ABACUS. Another approach proposed by Pandya et al. involves a post-processing strategy to filter low confidence base calls made by ABACUS that reside in problematic regions such as highly mutative and repeat regions [57]. Although the number of false positive base calls was reduced, the call rate may suffer if too stringent filters are used. Most recently, Zheng et al. identified “dips” in hybridization intensities of probes near/at mutation sites of variant samples when compared against non-variant control samples [58]. As such, they were able to identify variations in the sequence more accurately.

1.5.3 Recombination Detection in Viruses

Over the years, a wide range of tools to detect recombination have been developed using many different strategies. Most of these tools adopted strategies that detected changes in phylogeny or

distance metrics at different parts of the input sequences as an indication of the presence of recombination. There are also tools that use probabilistic methods, substitution models or cost models to infer recombination. Comparative studies on the performances of different recombination detection methods have also been carried out [59, 60]. Regardless of the strategies they used, recombination detection tools employ (as an initial step) a multiple sequence alignment of all input sequences. A multiple sequence alignment tries to align all sequences in a given query set. If two sequences in an alignment are descendents of a common ancestor, mismatches can be interpreted as point mutations and gaps as insertions/deletions introduced in one or both lineages in the time since they diverged from one another. The absence of substitutions, or the presence of only very conservative substitutions in a particular region of the sequence, suggest that this region is structural or functional importance. Multiple alignments are often used in identifying conserved sequence regions across a group of sequences hypothesized to be evolutionarily related.

We focus on a particular class of methods which uses a sliding window to detect recombination. Sliding window approaches are preferred over those that use a global reference tree because they can localize breakpoints more accurately and thus detect weak recombination events in the presence of strong recombination events [61]. Typically, recombination is detected by comparing and noting significant differences in distance measures (distance-based) or in phylogenetic trees' topologies (phylogeny-based) computed for the alignments in adjacent windows.

Distance-based window methods such as PhyPro [36] and DSS [62] are fast and fairly accurate. PhyPro computes, for every sliding window, a test statistic known as the minimum distance vector correlation using only non-conserved sites of an alignment. Then, they estimated

the p-value for the null hypothesis of no recombination by permuting the alignment 1000 times and counting the number of times the original minimum distance vector correlation was smaller than the minimum distance vector correlation for each permutation. In DSS, a length-500 window is slid along an alignment of DNA sequences. For each window, two distance matrices (one for each half of the window) are calculated according to some Markov model of nucleotide substitution. A sum-of-squares statistic is then computed for each distance matrix. Since each distance matrix encodes the phylogenetic relationships in its corresponding window, the presence of recombination will result in a big difference in the two sum-of-squares statistics for the two matrices. Unfortunately, distance-based methods tend to suffer from information loss when estimating recombination. Phylogenetic information is lost when only pair-wise distance comparisons are made using conventional distance metrics that measures only the global homology between two sequences.

On the other hand, phylogeny-based window methods such as PDM [63], Pruned-PDM [64] and RECOMP [65] try to generate the most likely phylogenetic trees for the alignments enclosed in neighboring windows and compare them to estimate recombination more accurately. The PDM and Pruned-PDM methods focused on estimating the topology changes based on a likelihood score. To reduce the uncertainty of tree estimation from short sequence alignments enclosed by the sliding window, they used a distribution of trees instead of a single tree as reference. Although accurate, their methods are very slow due to the need for Markov chain Monte Carlo simulations and hence intractable for larger datasets. RECOMP was then developed to provide a faster means of detecting recombination. Using a sliding window, a set of trees is generated for each window based on a maximum parsimony heuristic. Recombination is determined by comparing four different measures such as the Robinson-Foulds distance [66] of

sets of trees in adjacent windows. An accuracy comparable to Pruned-PDM is claimed but the interpretation of the four measures as an indication of recombination is sometimes not straight-forward and may even be ambiguous at times.

Regardless of distance-based or phylogeny-based methods, a sliding window approach has a major concern: the selection of window length. In previous works, the window length is usually arbitrary chosen within the range of 200–500. However, window length affects the sensitivity and accuracy of window-based methods to detect recombination. Recent works have shown that their results are most accurate when the given window length is approximately the recombinant subsequence length [62, 64]. If the length of the recombinant is not known in advance, an algorithm using different window lengths may produce vastly different analysis results on the same dataset. Furthermore, there may be problems in detecting recombinant regions shorter than the given window length due to the noise caused by the original sequence on either side of the recombinant subsequence included in the window.

Recently, a method to detect recombination without the use of sliding windows was introduced. Recco [67] uses a model of cost minimization and dynamic programming to detect recombination breakpoints. The basic model is to construct each sequence in the alignment in turn, from the other sequences in the alignment using only the mutation and recombination operators such as insertion and deletion. The minimum cost solution identifies the best recombination breakpoints and also the parental sequences. The performance of Recco is enhanced by a succeeding sensitivity analysis that provides an intuitive visualization of the solution. A major limitation of this method is that a recombinant sequence may not be detected if there is another similar sequence in the alignment. Consequently, the user needs to manually remove the closest sequence to the putative recombinant sequence iteratively.

1.6 Overview of Thesis

The rest of the thesis contains details on the detection and resequencing of viruses, and how their genomes are analyzed. Through the use of microarray technology, we are able to detect and resequence viruses in a cost-effective, efficient and high-throughput manner. However, more research is required to address accuracy and sensitivity issues faced by existing microarray data analysis methods. Downstream interpretation of sequence information also poses interesting challenges. An accurate interpretation may provide valuable insights on viral evolution and aid the development of viral treatments and vaccines.

In Chapter 2, we describe how PCR techniques are used to amplify DNA fragments of a virus to a magnitude required for a successful microarray hybridization. We also present the difficulties and limitations of various PCR amplification techniques on viruses. The main novel contribution of this chapter is a way to predict how well a random primer can amplify a given set of viruses. We then describe a fast algorithm to design better primers for amplification, along with wet-lab validation results. Implications of amplification efficiency on microarray probe selection and quality of data are also discussed.

In Chapter 3, we describe how microarrays are used to detect viruses and study the factors that affect detection accuracy. We report the results of a systematic investigation of the complex relationships between viral amplification efficiency, hybridization signal output, target-probe annealing specificity, and reproducibility of viral detection using a custom designed microarray platform. The novel contributions are a methodology for the *in silico* prediction of viral "signatures" and a statistics-based virus detection algorithm that can identify sequence-characterized and co-infecting viruses with low false positive rate.

In Chapter 4, we explore the capabilities of microarrays to generate whole-genome primary sequences of viruses for large-scale evolutionary biosurveillance. In response to the most recent H1N1 pandemic, we developed a resequencing kit that is capable of interrogating all eight segments of the H1N1 2009 influenza A virus genome and its variants. The novel contribution is a base-calling software (EvolSTAR) that introduces new methods that utilizes neighbourhood hybridization intensity profiles and substitution bias of probes on the microarray for mutation confirmation and recovery of ambiguous base queries. We demonstrate that EvolSTAR is highly accurate and has high call rates with a pilot study of 15 patient samples.

In Chapter 5, we describe various techniques to detect recombination events from genomic sequences of viruses. Recombination detection is important for a better understanding of viral evolution, more accurate genotyping and advancements in drug and vaccine developments. The main contribution of this chapter is a fast and accurate distance-based sliding-window method to detect recombination in a multiple sequence alignment. Using synthetic and biological datasets, we show that our method is more accurate than existing phylogeny-based methods. We also discuss how our method has potential use in other related applications such as genotyping.

In Chapter 6, we discuss the practical implications of this thesis in the field of evolutionary research and clinical decision making. We then present and summarize the main contributions made in this thesis. Lastly, we describe some of the future work stemming from our research.

1.7 Publications

This thesis is based on the following published material:

- Wong CW, Lee WH, Leong WY, Soh SW, Kartasasmita CB, Simoes EA, Hibberd ML, Sung WK, Miller LD. Optimization and clinical validation of a pathogen detection microarray. *Genome Biol.*, 2007, 8(5): R93.

- Lee WH, WK Sung. RB-Finder: An Improved Distance-based Sliding Window Method to Detect Recombination Breakpoints. In *RECOMB*, 2007.

- Lee WH, Sung WK. RB-finder: an improved distance-based sliding window method to detect recombination breakpoints. *J Comput Biol*., 2008, 15(7): 881-98.

- Lee WH, Wong CW, Leong WY, Miller LD, Sung WK. LOMA: a fast method to generate efficient tagged-random primers despite amplification bias of random PCR on pathogens. *BMC Bioinformatics*, 2008, 9: 368.

- Lee WH, Koh CW, Chan YS, Aw PPK, Loh KH, Han BL, Thien PL, Nai GYW, ML Hibberd, CW Wong, WK Sung. Large Scale Evolutionary Surveillance of the 2009 H1N1 Influenza A Virus Using Resequencing Arrays. *Nucleic Acids Research*, 2010.

- Lee VJ, Yap J, Cook AR, Chen MI, Tay J, Tan BH, Loh JP, Chew SW, Koh WH, Lin R, Cui L, Lee WH, Sung WK, Wong CW, Hibberd ML, Kang WL, Seet B, Tambyah PA. Oseltamivir ring prophylaxis for containment of Influenza A (H1N1-2009) outbreaks. *New England Journal of Medicine*, 2010.

Chapter 2

AMPLIFICATION OF VIRAL GENOMES

2.1 Polymerase Chain Reaction and its Limitations on Viral Genome Amplification

The Polymerase Chain Reaction (PCR) is a laboratory technique that "amplifies" a particular DNA sequence, generating millions of copies of it in the process. PCR uses specifically designed primers that are complementary to the sequence to be amplified. The primers provide a starting point for the extension of the DNA by a DNA polymerase (usually Taq or Pfu polymerase). Amplification is carried out in cycles. First, the DNA sample is heated up to separate the double strands. The sample is cooled slowly, allowing the primers to bind. Then, the sample is incubated at 72°C so that the DNA polymerase can extend the primers, creating a long complementary strand of DNA. As PCR progresses, the DNA generated is itself used as a template for replication, setting in motion a chain reaction in which the DNA template is exponentially amplified [68]. Furthermore, PCR requires only a minute amount of DNA sample for amplification. This extremely high efficiency and sensitivity has made PCR an essential tool for many applications such as forensic analysis [69], genome sequencing [70] and cancer diagnostics [71].

Another important application of PCR is the amplification and subsequent detection and identification of viruses for disease diagnosis [72, 73]. PCR-based virus detection is highly sensitive and accurate, with the capability to detect viruses soon after infection and even before

the onset of disease. In practice, there are several factors that may cause PCR-based detection to fail. PCR needs to use primers that are specific to the targeted sequence for a successfully amplification. These primers are typically designed from sequence information stored in public databases. However, as some viruses mutate or recombine, their sequence information may become inaccurate. Moreover, sequence information for novel viruses such as severe acute respiratory syndrome (SARS) and H1N1 2009 influenza virus will not be available until much later. Even if the virus is not novel, a clinical prediction of the infectious source would have to be made before PCR can be conducted [22].

2.2 Tagged-random Primer Amplification

The limitations of PCR in amplifying novel, unknown or highly mutative viruses have led to the development of an alternative amplification strategy known as tagged-random primer amplification (T-PCR). Unlike PCR which requires the use of specific primers chosen from target sequences, T-PCR uses a tagged-random primer consisting of a constant 17 bp at the 5'-end known as the 5' tag and a random oligomer (unknown base N) of length 9-15 at the 3'-end which could theoretically bind to any DNA sequence [74]. However, many viruses, such as influenza viruses and retroviruses, are composed of RNA rather than DNA. As a result, tagged-random primers cannot be used directly to amplify such viruses. To amplify RNA viruses, a variant of PCR known as reverse transcription polymerase chain reaction (RT-PCR) is needed. In RT-PCR, RNA is first reverse transcribed into cDNA using the enzyme reverse transcriptase. In theory, the random oligomer now binds indiscriminately to the nucleic acids template. The resulting cDNA is then amplified using the 17-mer 5' tag to generate PCR products (Figure 1).

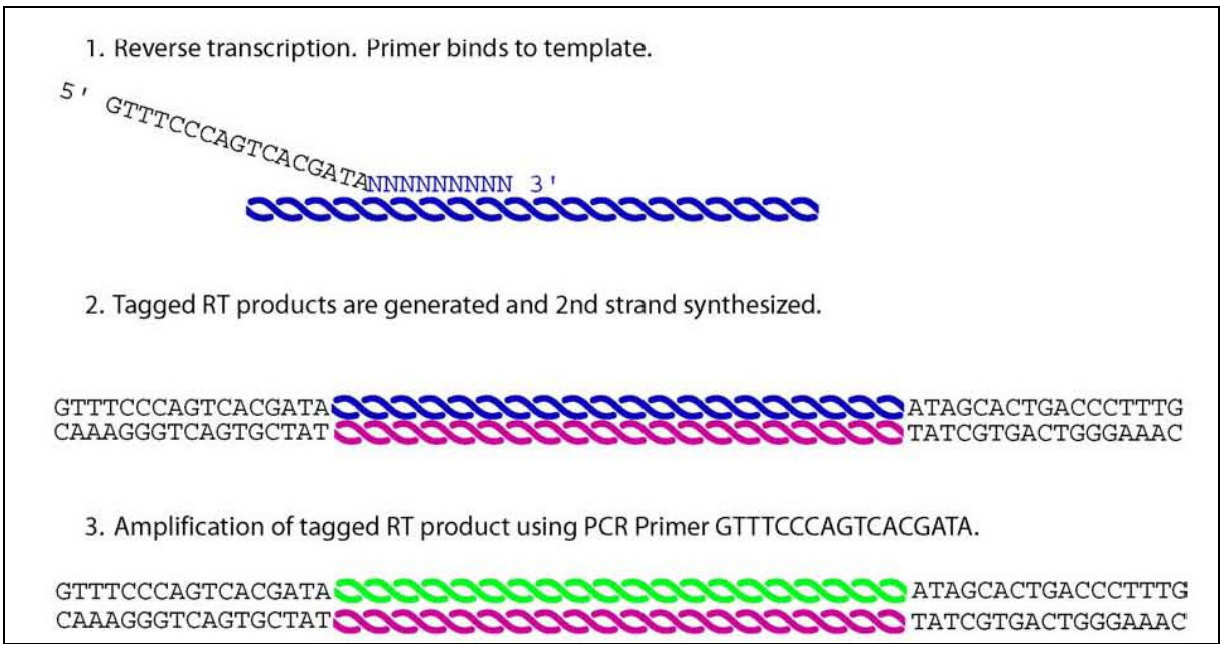

Figure 1: RT-PCR binding process of tagged-random primers on a RNA viral sequence followed by PCR.

In practice, T-PCR does not guarantee the genome-wide amplification of the virus in the sample. The exact locations where the random primers bind could be influenced by the presence of intra-primer secondary structure formation (ie the 5'-end tag forms a dimer or hairpin with the 3'-end random oligomer) or melting temperatures [75]. This differential binding of the tagged-random primers to different parts of the viral genome creates an amplification bias that prevents certain regions from being amplified, resulting in incomplete amplification of the target virus. Several studies involving microarray experiments using T-PCR that have reported the lack of hybridization signals in tiling probes that reside in certain regions of the viruses in their samples, most likely attributed to incomplete T-PCR amplification [76, 77].

2.3 Amplification Efficiency Model of the RT-PCR Process

The binding affinity of a primer pair to the target genome impacts RT-PCR efficiency. In the case of using tagged-random primers, the quality of the RT-PCR product depends on how well a tagged-random primer instantiation pair binds to the target genome. Here, we termed a particular configuration of the given tagged-random primer as a tagged-random primer instance. For example, (5'-GTT TCC CAG TCA CGA TA *TTTTAAAAG*-3') and (5'-GTT TCC CAG TCA CGA TA *CATCATCAT*-3') are instantiations of the tagged-random primer (5'-GTT TCC CAG TCA CGA TA *NNNNNNNNN*-3'). Some instantiations of the tagged-random primer can bind better to the target genome than others. The identification of such tagged-random primer instantiations and where they bind to the target genome gives us an indication of how likely a particular region of the target genome will be amplified. Using this approach, we proposed an amplification efficiency model which computes an Amplification Efficiency Score (AES) for every position of a target genome.

As a concrete example for our modeling, we use a tagged-random primer that has a fixed 17-mer header and a variable 9-mer tail of the form (5'-GTT TCC CAG TCA CGA TAN NNN NNN NN-3'). This tagged-random primer is commonly used in virus detection experiments [78, 79, 80]. Let v_a be the actual virus in the sample. To get a RT-PCR product in a region between positions i and j of v_a, we require (1) a forward primer binding to position i, (2) a reverse primer binding to position j and (3) $\lambda_l \leq |i - j| \leq \lambda_u$ where λ_l and λ_u are the lower and upper bounds of the desired PCR product length respectively (Figure 2).

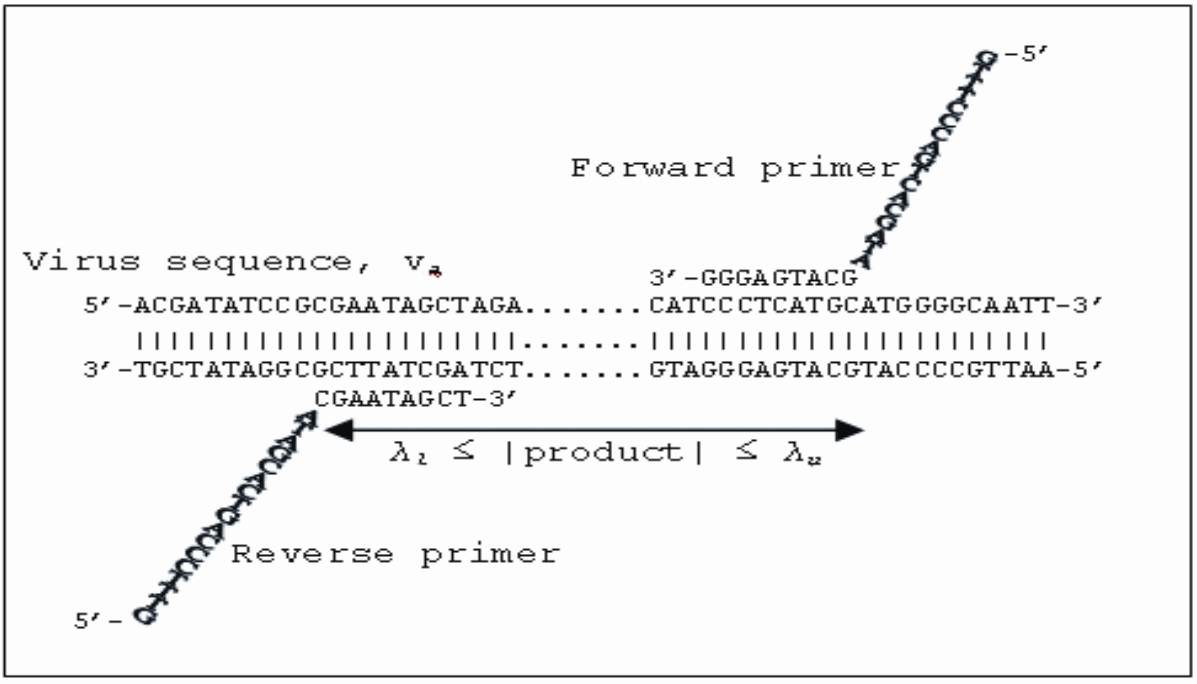

Figure 2: Amplification efficiency model of the RT-PCR binding process of an instantiation of a tagged-random primer on a RNA virus v_a.

Consider a pair of forward and reverse tagged-random primer instantiations where the forward primer is at a particular position i of v_a, the reverse primer is at position j of v_a and $|i - j|$ is the product length. Let $P^f(i)$ and $P^r(j)$ be the probability that the forward primer can bind to position i and the probability that the reverse primer can bind to position j respectively. For simplicity, we assume that a random (forward or reverse) primer instantiation can bind to a particular position i of v_a only if 9-mer of the instantiated nonamers of the random primer is a reverse complement of the length-9 substring at position i of v_a (Note that other binding criteria such as 75% similarity rule or nearest-neighbour binding free energy [81] can be used as well). Thus, all other instantiations of the given tagged-random primer whose instantiated nonamers is not a subsequence of v_a, do not contribute to the amplification process and thus omitted from the computation of AES. We compute $P^f(i)$ and $P^r(j)$ based on well-established primer design criteria

[82]. The idea is that $P^f(i)$ and $P^r(j)$ will be small if the forward primer or reverse primer forms self-dimers or has extreme melting temperatures, or form a significant primer-dimer with each other. Another consideration is that if the fixed 17 basepairs 5-end tag of the given random primer is similar to v_a, it may lead to mispriming and thus results in a lower $P^f(i)$ and $P^r(j)$ for all i and j.

It is difficult to assess the exact extent of influence of primer-dimers and melting temperatures on amplification. Hence, we estimate $P^f(i)$ and $P^r(j)$ using a simple model:

1. A primer cannot bind to the sequence efficiently if it folds onto itself. A primer is a self-dimer if it forms a 3'-end or internal hairpin with three or more bases. Thus, $P^f(i) = 0$ if the forward primer at i forms a self-dimer. Similarly, $P^r(j) = 0$ if the reverse primer at j forms a self-dimer.
2. The RT-PCR process is performed at a certain temperature, normally 55 ^{o}C to 60^{o}C. If the melting temperature of a primer is not at this ideal temperature, then the primer may not bind to the sequence. Hence, we model this observation by decreasing $P^f(i)$ and $P^r(j)$ proportionally to the difference in the melting temperature of the forward primer and reverse primer to the ideal experimental temperature respectively. Specifically, $P^f(i) = 1 - (|\text{Tm(forward primer)} - TM|/TM)$ and $P^r(j) = 1 - (|\text{Tm(reverse primer)} - TM|/TM)$ where TM is the ideal experimental temperature and Tm(x) is the melting temperature of a primer x. We compute Tm(x) = 81.5 + 16.6 (log M) + 0.41 (% G+C) - 0.72 (% formamide) where M is the monovalent salt molarity, and (% G+C) the percentage of guanine plus cytosine residues in DNA [83].
3. To avoid mispriming, if the 17bp fixed tag of the tagged-random primer has more than 75% similarity to any subsequence of the target genome, we discard this random primer.

That is, $P^f(i) = 0$ if the forward primer at i has a fixed tag with more than 75% similarity to any subsequence of the target genome. Similarly, $P^r(j) = 0$ if the reverse primer at j has a fixed tag with more than 75% similarity to any subsequence of the target genome.

Based on our model, the probability that a pair of random primer instantiations give a good quality PCR product from position i to j on v_a is $P^f(i) \times P^r(j)$. Due to the abundance of random primer instantiations used in a RT-PCR process, it is likely that all pairs of random primer instantiations that can effectively bind to v_a will contribute a PCR product. Thus, for a valid forward primer at position i, we must compute the above probabilities for a range of positions j at which a valid reverse primer exists, ie $\lambda_l \leq |i - j| \leq \lambda_u$. Thus, an Amplification Efficiency Score, AES_x, for every position x of v_a can be computed by considering the combined effect of all forward and reverse primer-pairs that amplifies it:

$$AES_x = \sum_{i=x-\lambda_u}^{x} \left\{ P^f(i) \times \sum_{j=\max(x+1,i+\lambda_l)}^{i+\lambda_u} P^r(j) \right\}$$

Once we compute the AES for all positions of v_a, we plot the AES against the genomic positions of v_a. This generates a graph which indicates the regions in v_a predicted to be amplified efficiently by the given tagged-random primer (represented by peaks) and regions that do not (represented by troughs). To validate the algorithm, we conducted two parallel experiments using two microarrays, each consisting of 1948 probes tiled across the RSV B genome. Two biologically identical samples of RSV B are amplified using two different tagged-random primers:

1. Primer A1 (5'-GTT TCC CAG TCA CGA TA NNNNNNNNN-3'): A commercially available tagged-random primer.

2. Primer A2 (5'-GAT GAG GGA AGA TGG GG NNNNNNNNN-3'): Primer with highest AES among 10000 randomly generated tags.

Next, we ranked the hybridization signal intensities for all 1,948 probes tiled across the RSV B genome in each microarray experiment and compared them to the AES values of the tagged-random primer used for amplification (Figure 3). We observed that high AES significantly correlates to probe hybridization signal intensity above the detection threshold (P=2.2x10-16; Fisher's Exact Test). In another set of experiments involving a patient sample positive for metapneumovirus (hMPV), the probes tiled across the hMPV genome showed a similar result, P=1.3x10-9. Repeatedly, we observed that higher AES correlated with greater probe detection, with, on average, >70% detection for probes in the top 20% AES.

Our model allows us to predict how successful the amplification on a target viral genome will be given a particular tagged-random primer. An ideal tagged-random primer would generate high AES values uniformly across the whole target genome. This quantification of the efficiency of amplification of a tagged-random primer on a target genome in the form of AES also enables us to compare the effectiveness of different tagged-random primers if they are to be used to amplify the genome. For example, tagged-random primer r_1 is predicted to work better than tagged-random primer r_2 if the average AES of r_1 across a target genome is higher than that of r_2. An important application of our AES model is in the design of probes for a virus-detection microarray. Probes should be chosen in regions in the target viral genomes that can be amplified efficiently by the tagged-random primer used. Conversely, we should omit probes from regions

in the target genomes which are predicted not to amplify efficiently since we cannot tell if these probes did not hybridize due to the absence of the target viruses in the sample or just that the amplification by the random primers failed.

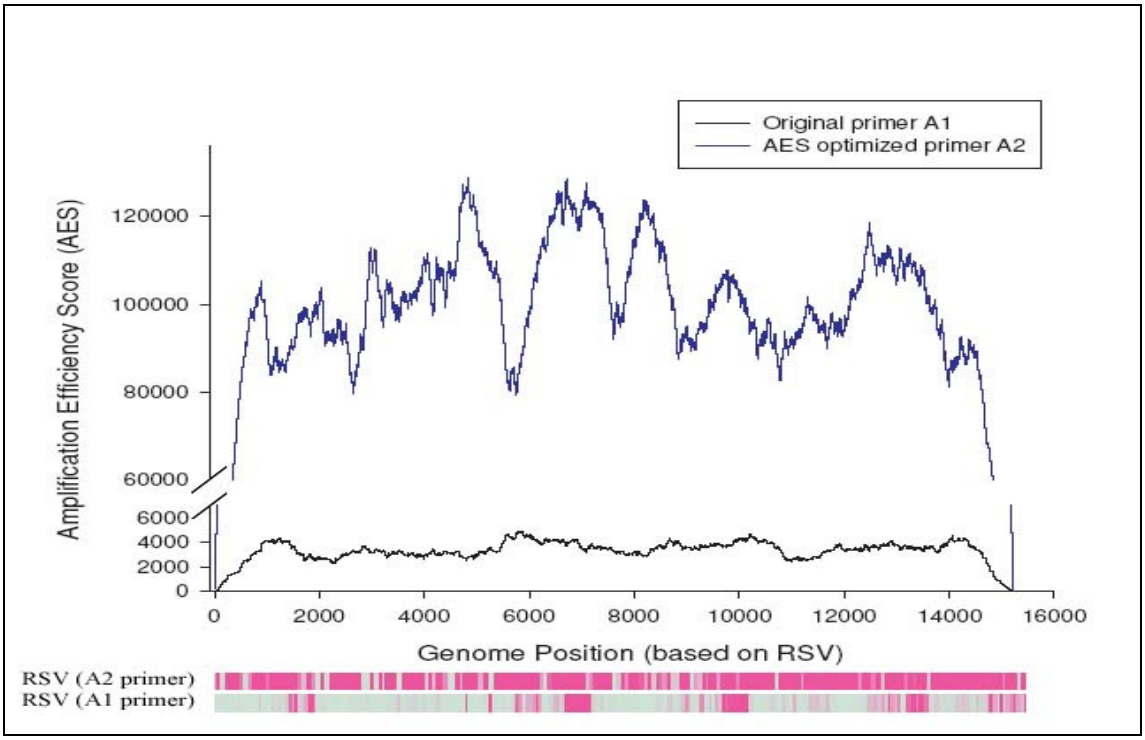

Figure 3: Correlation of probe hybridization signals with AES of tagged-random primers A1 and A2 in a RSV B sample.

2.4 Generating the Tag of a Tagged-Random Primer using LOMA

Amplification failure may occur if there are many regions of the target genome where the tagged-random primer cannot bind. As such, using any available commercial tagged-random primer or a tagged-random primer that was used in other publications may not guarantee a successful amplification on a target genome. Although the computation of AES allowed us to

compare the amplification efficiency of different tagged-random primers on a target genome, it would be most useful if we know the tagged-random primer that binds to the given genome optimally.

The best way to obtain the most efficient tagged-random primer to amplify a target genome is to compute the AES graph for all possible combinations of the 17-bp 5'end tag and choose the tag that has the highest average AES with the target genome. This is impractical as this would require 4^{17} runs of the AES computation algorithm. A naïve approach would be to randomly generate a large number of tags (eg. 10000) and choose the one that has the highest average AES with the target genome. Other similar randomization approaches could also be used to improve the chances of getting a more efficient tag to amplify the target genome. However, these approaches are still slow especially when we need to choose an efficient tagged-random primer for multiple genomes.

We propose LOMA (Least Occurrence Merging Algorithm), a more deterministic and faster algorithm to generate an efficient tag for a target genome v_a. The idea is to use a "divide and conquer" strategy to generate n-bp tags by concatenating m shorter k-mers where $m = n/k$. Recall that the 5'end tag of the tagged-random primer should be not similar to v_a to avoid mispriming. To form such a tag, the constituent k-mers should also be dissimilar to v_a. Based on this criterion, we compute the number of occurrences with more than 75% similarity in v_a for each of the 4^k k-mers. Then, we sort the k-mers based on their occurrence count in v_a in ascending order. Tags are generated using the top ranking k-mers whose number of occurrences in v_a is lower than some threshold T. Ideally, we want to generate tags using only k-mers with no occurrence in v_a, ie $T = 0$.

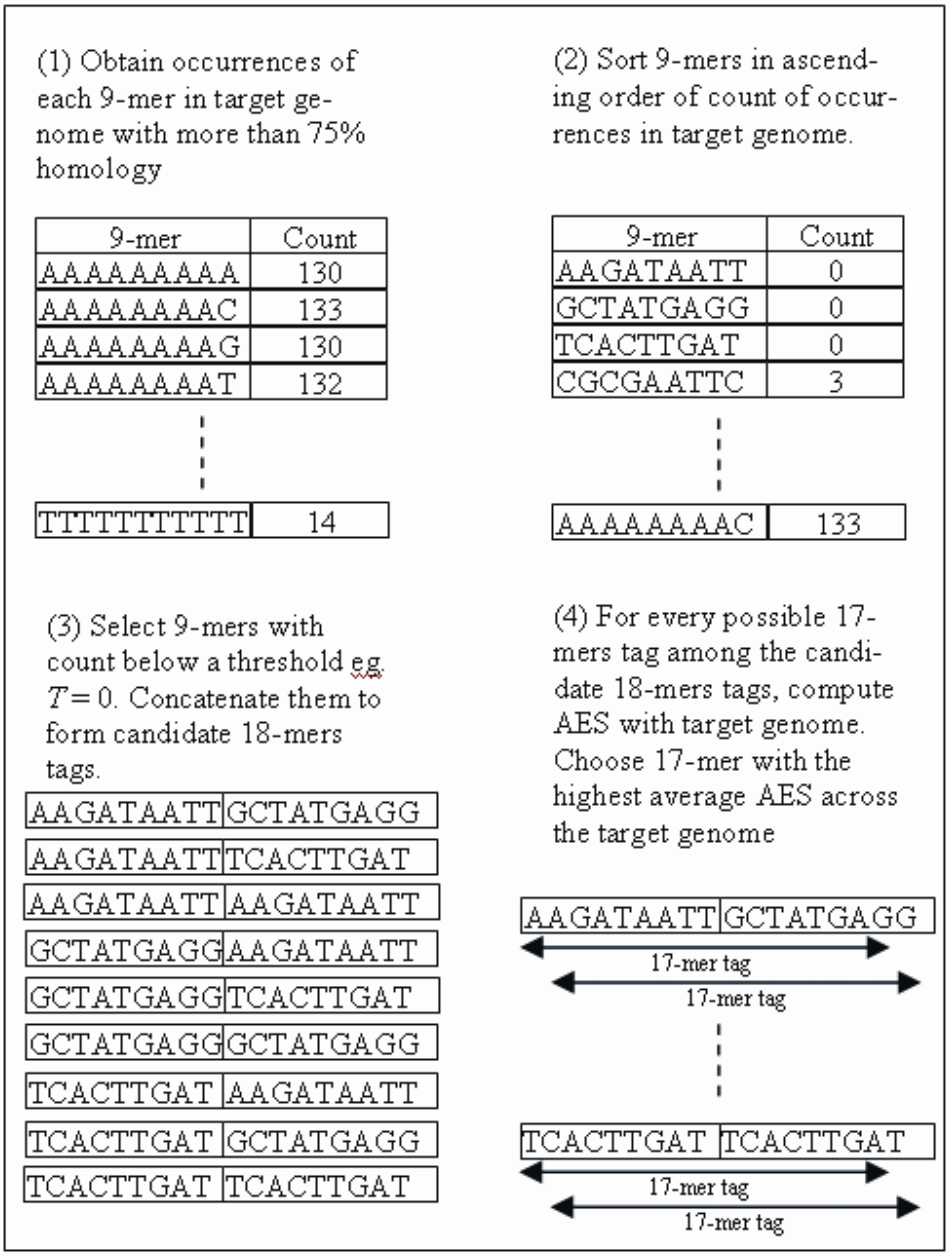

Figure 4: Flowchart of LOMA with $n = 17$, $k = 9$ and $T = 0$.

Suppose x k-mers have occurrences in v_a less than T. We generate a tag by concatenating any m of the x k-mers. This results in x^m possible tags. Since x is small and m is typically 2 or 3, the total number of tags generated is much less compared to a brute force or randomized

approach. Furthermore, the x^m tags generated by our method are guaranteed to be dissimilar to v_a. Thus, all there is left to do is to compute the AES with v_a of each of the x^m tags and choose the one with the highest average AES across v_a. Figure 4 shows the flowchart of our algorithm.

Unlike randomized approaches, LOMA is easily extended to generate an efficient tag for multiple genomes. Specifically, given a set of genomes V, we need only to modify step one of the algorithm to compute the number of occurrences with more than 75% similarity in every genome in V for each of the 4^k k-mers. Once candidate random-tags are generated, we compute their AES with each of the genomes in V and choose the one with the highest average AES for all the genomes in V.

2.5 Experimental Evaluation of LOMA

We describe experiments to test the hypothesis that different tagged-random primers have different amplification efficiencies and to assess the effectiveness of our algorithm to generate a good tagged-random primer. In our experiments, we use eight human nasopharyngeal aspirate patient samples obtained from children under 4 years of age with lower respiratory tract infections. Using real-time PCR with specific primers, we confirmed that five samples contain human respiratory syncytial virus (RSV) while the remaining three samples contain human metapneumovirus (HMPV).

Three tagged-random primers are then used to amplify the eight samples:

1. Primer A1 (5'-GTT TCC CAG TCA CGA TA NNNNNNNNN-3'): A commercially available tagged-random primer.

2. Primer A2 (5'-GAT GAG GGA AGA TGG GG NNNNNNNNN-3'): Primer with highest AES among 10000 randomly generated tags.

3. Primer A3 (5'-TAG GTC GGT CGG TAG GT NNNNNNNNN-3'): Primer generated using our proposed algorithm LOMA.

Subsequently, the samples are hybridized onto our virus detection chip. Since our virus detection chip contains tiling 40-mer probes of both RSV and HMPV, the number and distribution of the probes with high signal intensities would give a good indication of the amount of PCR products generated across the target genome by a tagged-random primer. We expect that a tagged-random primer with desirable amplification efficiency that generates sufficient PCR products uniformly across the whole target genome would result in high signal intensity probes distributed evenly across the whole genome.

We present the first set of experiments involving the amplification of five RSV patient samples by the three tagged-random primers A1, A2 and A3. In each experiment involving a particular pair of RSV patient sample and random-tagged primer, hybridization signal intensities for the 1948 probes tiled across the 15225 bp RSV genome were compared to their corresponding AES along the genome. When using primer A1, we obtained AES with values less than 5000 with an average of 3300. However, when primers A2 and A3 are used, the AES averages are 110000 and 140000, respectively. This dramatic increase in predicted amplification efficiency gave an indication that in theory, our designed tagged-random primers A2 and particularly A3 perform much better than A1.

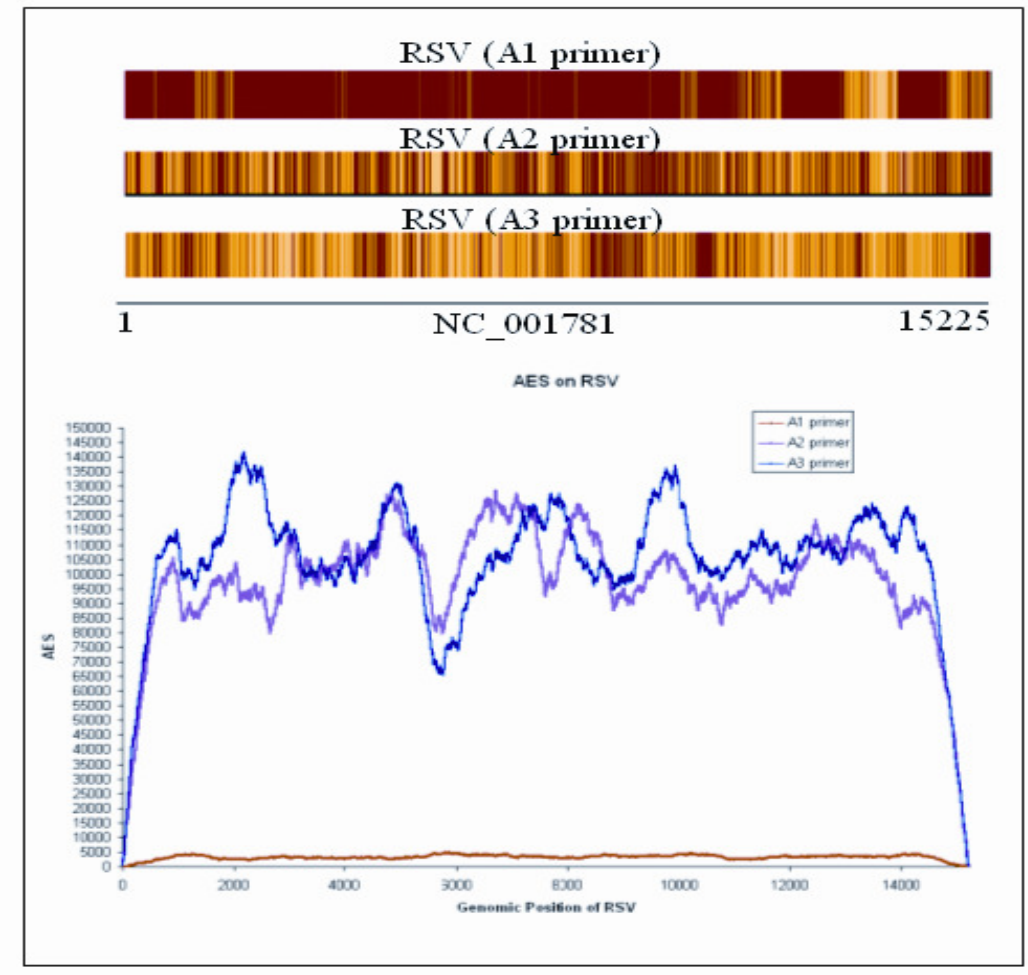

Figure 5: Application of AES on a RSV sample. An RSV patient sample was amplified separately using primer A1, primer A2 and primer A3. Hybridization signals of probes after amplification by each primer are shown as a heatmap. The probes that have detectable signals above threshold are shown in orange/yellow in the corresponding heatmaps. The graph below the heatmaps shows our AES prediction for the three primers: A1 (orange line), primer A2 (pink line) and primer A3 (dark blue line). Our AES predictions closely matches the actual hybridization results, ie primer A3 performs slightly better than primer A2 but both A3 and A2 performs significantly better than A1 on RSV.

Recall that probes in regions of high AES are expected to be least affected by a poor amplification and thus have the correct high hybridization signals if the virus is present in the sample. For all the experiments, we observed that high AES significantly correlates to probe hybridization signal intensity above the detection threshold with a p-value of 2.2 × 10-16 using

the Fisher's exact test. About 80% of the probes with high signal intensities (≥ mean + 3 standard deviation) have high AES values. We also observed that primers A2 and A3 showed a tremendous improvement in overall PCR efficiency in amplifying RSV over primer A1. This increase in PCR efficiency resulted in increased hybridization of DNA to the probes and is reflected in the uniformly higher signal intensities observed using primer A2 and A3. This is illustrated in Figure 5. Further analysis of the RSV experiments revealed that only 20% to 30% of the 1948 RSV probes had signal intensities above detection threshold when primer A1 was used. By contrast, the use of primer A2 resulted in 60% to 71% of probes with signal intensities above detection threshold. Primer A3 fared slightly better than primer A2, resulting in more than 70% of the probes having signal intensities above detection threshold.

We conducted another set of experiments to verify that the observations made involving RSV and the three tagged-random primers are not isolated observations and that they can be replicated in other genomes as well. Following the experimental procedure used in the previous set of experiments, three patient samples containing HMPV are subjected to amplification by primers A1, A2 and A3. Similarly, in each experiment involving a particular pair of HMPV patient sample and tagged-random primer, hybridization signal intensities for the 1705 probes tiled across the 13335bp HMPV genome were compared to their corresponding AES along the genome. Figure 6 shows the heatmaps and AES plots of the HMPV genome when amplified by primers A1, A2 and A3. The results are similar to that of the first set of experiments on RSV. In the three samples, Primer A1 performs worse on HMPV than RSV, causing only < 8% of the 1705 probes to be detected above threshold. Primers A2 and A3 performed much better than primer A1, causing >80% and > 88% of the probes to be detected above threshold respectively.

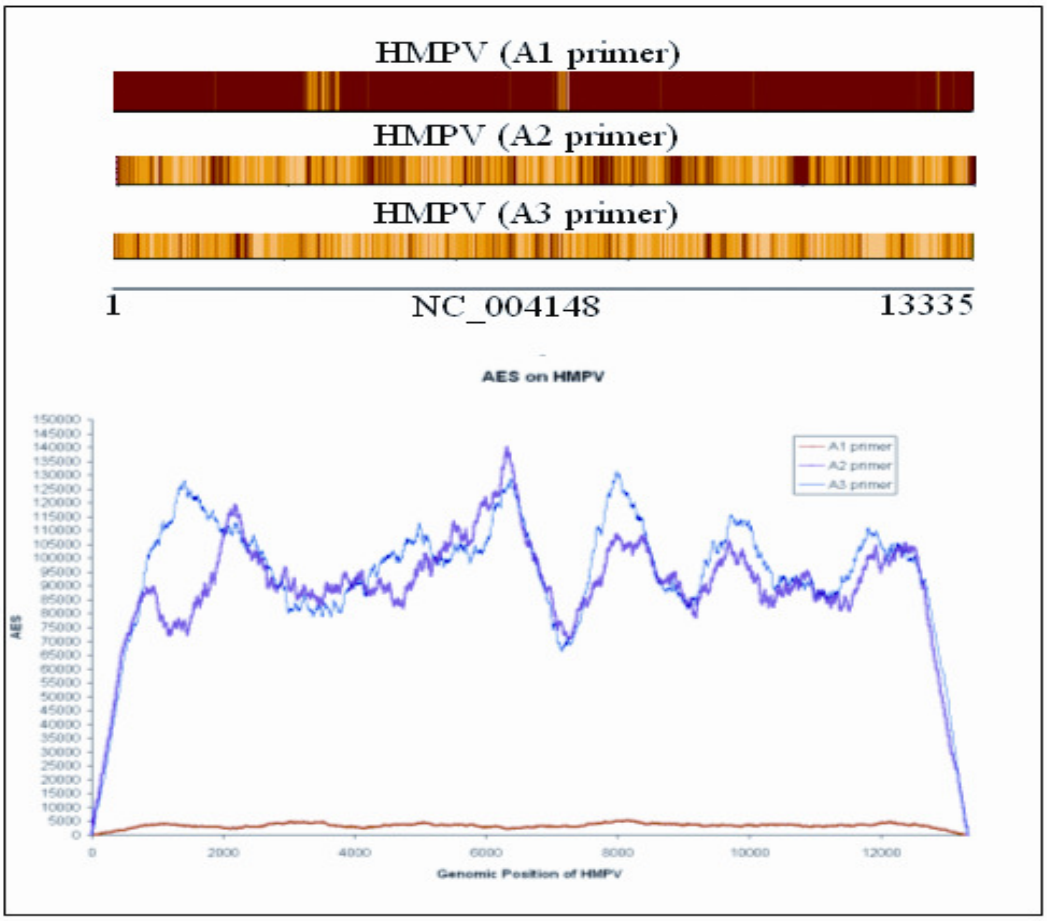

Figure 6: Application of AES on a HMPV sample. An HMPV patient sample was amplified separately using primer A1, primer A2 and primer A3. Hybridization signals of probes after amplification by each primer are shown as a heatmap. The probes that have detectable signals above threshold are shown in orange/yellow in the corresponding heatmaps. The graph below the heatmaps shows our AES prediction for the three primers: A1 (orange line), primer A2 (pink line) and primer A3 (dark blue line). Our AES predictions closely matches the actual hybridization results, ie primer A3 performs slightly better than primer A2 but both A3 and A2 performs significantly better than A1 on HMPV.

Our experiments have shown that the commonly used primer A1 amplify RSV and HMPV poorly. Further analysis reveals that many instances of the primer A1 that are supposed to bind to RSV and HMPV form self-dimers and hence unable to amplify the genome efficiently. On the other hand, primers A2 and A3 amplified RSV and HMPV efficiently. However, compared to primer A2, primer A3 was generated in a much shorter time by LOMA and performs just as well, if not better.

2.6 Multiplexing Tagged-random Primers

LOMA generates random-tagged primers that are capable of amplifying their target genomes efficiently with coverage of more than 70% up to 90%. We explore the possibility of using multiple tagged-random primers to achieve a more complete amplification of the target genome.

In our experiments, we observed that one tagged-random- primer may amplify a particular region of a target genome more efficiently than another tagged-random primer. For example in Figure 6 at genomic positions 1500–1900 of HMPV, the heatmap shows that primer A3 performs much better than primer A2. On the other hand, on the same genome at positions 2000–2200, the heatmap shows that primer A2 performs better than primer A3. This suggests that it is possible to design multiple collaborating tagged-random primers to amplify a target genome. The idea is to design additional tagged-random primers that have regions with high AES covering the regions with low AES of existing random-tagged primers. This is shown in Figure 7.

Although this approach is highly viable as suggested by our experimental results, achieving a successful multiplexing of multiple random-tagged primers in the laboratory may not be as straight-forward as the traditional multiplexing of specific primers [84]. Recall that a tagged-random primer consists of random oligomers that could theoretically bind to all possible sequences. Using two or more tagged-random primers simultaneously in a PCR amplification reaction may result in the formation of primer-dimers among all instances of the tagged-random primers and cause the amplification to fail. To ensure higher success of multiplexing tagged-random primers, an alternative solution is to perform the PCR reaction with the first tagged-random primer, then perform another PCR reaction with the second tagged-random primer and

so on. In other words, we multiplex *n* tagged-random primers by performing *n* PCR reactions in series. This will avoid the problem of primer-dimers when multiplexing tagged-random primers.

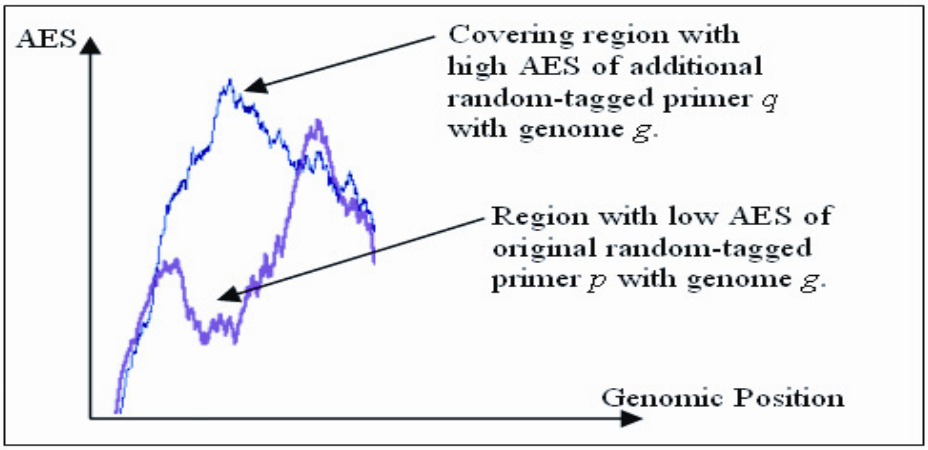

Figure 7: Design of multiple random-tagged primers to amplify a target genome *g*. Original tagged-random primer *p* has a region with low AES on *g*. We design additional tagged-random primer *q* such that it has high AES in that region.

2.7 Conclusion

New generation genomic and diagnostic applications require the amplification of a wide range of known viruses and potentially novel viruses as an initial step. As it is not cost-effective to design specific primers for all known viruses and quite impossible to design specific primers for yet to be known viruses, tagged-random primer amplification is preferred over primer-specific amplification. However, genome-wide amplification bias of tagged-random primers is a serious yet commonly overlooked problem.

In this chapter, we described a model to predict the amplification efficiency of a random-tagged primer given a target genome(s). The AES provided us with a measurement that we can use to compare the amplification efficiency of different tagged-random primers on the target

genome. This paved the way for the development of LOMA, a fast and effective tagged-random primer generator. Through experiments, we have shown that the random-tagged primer generated by LOMA performs significantly better than a commonly used random-tagged primer on different genomes. Furthermore, LOMA is able to generate good tagged-random primers much faster than randomized approaches.

Unlike specific primers that are almost always selected from the target genome under stringent primer design criteria [85], people tend to use tagged-random primers without checking their suitability with the target genome. This is a serious oversight that may cause inaccuracies in downstream work such as microarray analysis. Our research has shown that the blind use of a tagged-random primer in a PCR reaction on a virus sample may not lead to a successful amplification. Thus, the design of tagged-random primers is an important consideration when performing PCR and should be a common practice when using tagged-random primers.

LOMA is available at http://www.comp.nus.edu.sg/~bioinfo/AES_LOMA/

Chapter 3

VIRUS DETECTION AND IDENTIFICATION

3.1 DNA Microarrays in Virus Detection

Developed in the 1990s, DNA microarrays detect and identify viruses through hybridization of their DNA against millions of oligomers (known as probes). Since each probe is designed to hybridize only to its intended sequence, we can easily determine if a particular virus is present in the sample by analyzing its probes (Figure 8). Hence by analyzing different sets of probes, microarrays have the ability to detect multiple viruses in a single experiment, co-infections and novel virus infections. Currently, viral detection by microarrays can rapidly decrease laboratory turnaround times so that results can be available within 2–6 hours. Future developments may see this reduced even further; and through the development of point-of-care devices, perhaps enable the clinician to make the diagnosis directly at the bed-side [86]. This would no doubt reduce morbidity and mortality, for example, through the earlier implementation of appropriate antimicrobial treatment.

While pathogen microarrays and their utility in discovering emerging infectious diseases such as SARS have been described, technical problems related to accuracy and sensitivity of the assay prevent their routine use in patient care [87, 88, 89, 90]. For microarrays to become a standard diagnostic tool, the following questions must be addressed: (1) What are the factors that influence probe design and performance? (2) How is a pathogen “signature” measured and detected? (3) What is the specificity and sensitivity of an optimized detection platform? (4) Can detection algorithms distinguish co-infecting pathogens and closely related viral strains [91, 92]?

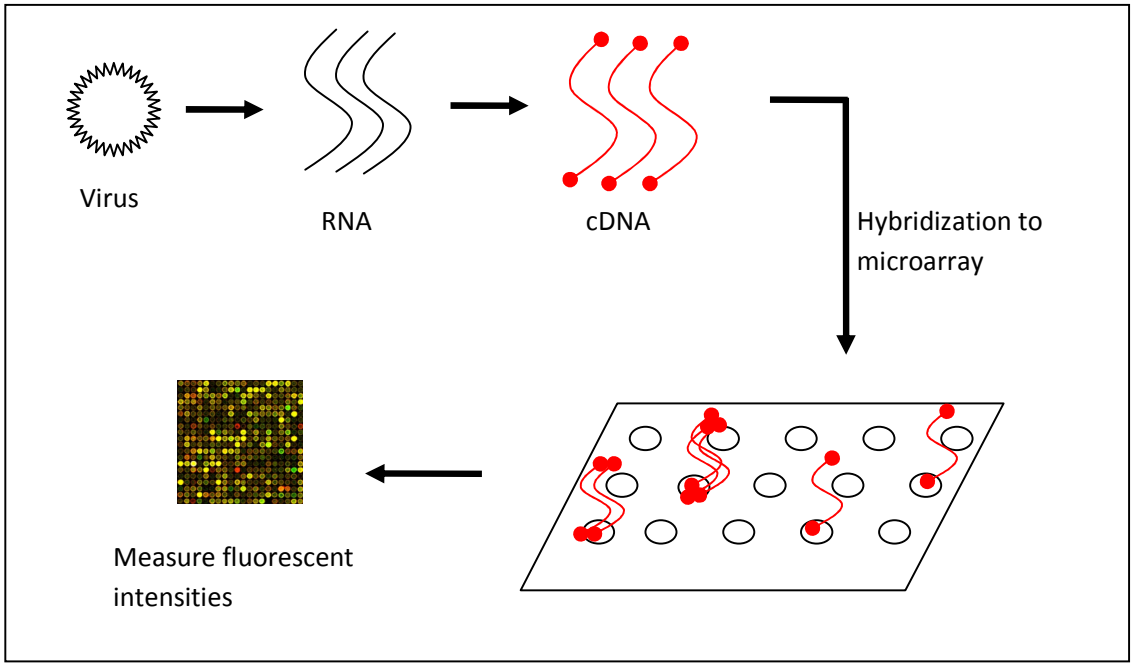

Figure 8: Microarray hybridization process. Viral RNA is first reversed transcribed to cDNA. The cDNA is then amplified, fragmented, end-labeled with biotin and applied to the microarray. The cDNA binds to complementary probes on the microarray, a process known as hybridization. A special scanner is then used to measure the fluorescent areas on the microarray. Probes that successfully bind to the cDNA fragments will generate a very bright fluorescent area, thus indicating what viruses are present in the sample.

Noisy signals caused by cross-hybridization artifacts present a major obstacle to the interpretation of microarray data, particularly for the identification of rare virus sequences present in a complex mixture of nucleic acids [93]. For example, in clinical specimens, contaminating nucleic acid sequences such as those derived from the host tissue, will cross-hybridize with virus-specific microarray probes above some threshold of sequence complementarities. This can result in false-positive signals that lead to erroneous conclusions. Similarly, the virus sequence, in addition to binding its specific probes, may cross-hybridize with other non-target probes (i.e., designed to detect other viruses). This latter phenomenon, though seemingly problematic, could provide useful information for virus identification to the extent

that such cross-hybridization can be accurately predicted. With various metrics to assess annealing potential and sequence specificity, microarray probes have traditionally been designed to ensure maximal specific hybridization (to a known target) with minimal cross-hybridization (to non-specific sequences). However, in practice we have found that many probes, though designed using optimal *in silico* parameters, do not perform according to expectations for reasons that are unclear.

3.2 Design of Virus Recognition Probe Sets

The accuracy and sensitivity of a pathogen detection microarray depend on its composition of probes. The practicality of a microarray as a low-cost pathogen detection tool also places a restriction on the number of probes that can be included on the microarray. Hence, the selection of a minimal number of "good" probes that can detect and identify a set of viruses of interest, to be synthesized onto the microarray is vital. The main challenge in probe design involves selecting probes that bind only to their intended cDNA and not to others (cross-hybridizations). In addition, proper probe design must also take into account other factors such secondary structure formation, CG-content and melting temperatures of probes that may cause hybridization errors [94].

3.2.1 Empirical Determination of Cross-Hybridization Thresholds of Probes

To systematically investigate the dynamics of array-based pathogen detection, we created an oligonucleotide array using Nimblegen array synthesis technology [95]. The array was designed

to detect up to 35 RNA viruses using 40-mer probes tiled at an average 8-base resolution across the full length of each genome. Together with 7 replicates for each viral probe, and control sequences for array synthesis and hybridization, the array contained a total of 390,482 probes. Initially, we studied virus samples purified from cell lines, reverse-transcribed and PCR-amplified with virus-specific primers (instead of random primers). This allowed us to study array hybridization dynamics in a controlled fashion, without the complexity of cross-hybridization from human RNA and random annealing dynamics which occur with random primers. We then applied our findings to clinical samples amplified using random primers.

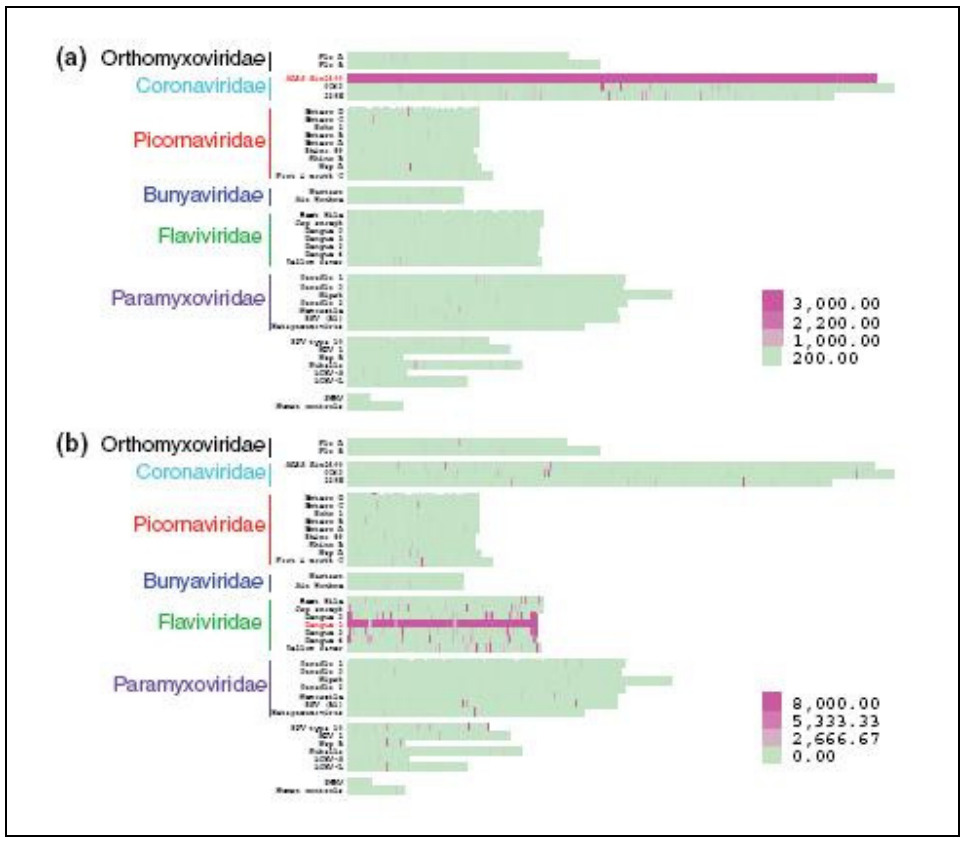

Figure 9: Heatmap of microarray probe signal intensities. Cells corresponding to probes are aligned in genomic order and colored according to the signal intensity-color scales shown. Hybridization signatures corresponding to SARS Sin850 (a) and Dengue I Hawaiian isolate (b) are shown.

SARS coronavirus and Dengue serotype 1 genomic cDNA were amplified in entirety (as confirmed by sequencing), labeled with Cy3 and hybridized separately on microarrays. The SARS sample hybridized well to the SARS tiling probes, with all 3,805 SARS-specific probes displaying fluorescent (Cy3) signal well above the detection threshold (determined by probe signal intensities > 2 standard deviations (SD) above the mean array signal intensity; Figure 9a). Cross-hybridization with other pathogen probe sets was minimal, observed only for other members of Coronaviridae and a few species of Picornaviridae and Paramyxoviridae, consistent with the observation that SARS shares little sequence homology with other known viruses [96]. The hybridization pattern of Dengue 1, on the other hand, was more complex (Figure 9b). First, we observed that hybridization to the Dengue 1 probe set was partially incomplete (i.e., regions absent of signal) due to sequence polymorphisms. The Dengue 1 sample hybridized on the array was cultured from a 1944 Hawaiian isolate, whereas the array probe set was based on the sequence of a Singaporean strain S275/90, isolated in 1990 [97]. Sequencing the entire genomes of these 2 isolates revealed that the array probes which failed to hybridize each contained at least 3 mismatches (within a 15-base stretch) to the sample sequence. Second, we observed that cross-hybridization occurred to some degree with almost all viral probe sets present on the array, particularly with probes of other Flaviviridae members, consistent with the fact that the 4 Dengue serotypes share 60-70% homology. To understand the relationship between hybridization signal output and annealing specificity, we first compared all probe sequences to each viral genome using 2 measures of similarity: probe hamming distance (HD) and maximum contiguous match (MCM). HD measures the overall similarity distance of two sequences, with low scores for similar sequences [98, 99]. MCM measures the number of consecutive bases which are exact matches, with high scores for similar sequences [99, 100].

We calculated the HD and MCM scores for every probe relative to the Hawaiian Dengue 1 isolate and observed that these scores were negatively correlated (HD) and positively correlated (MCM) to probe signal intensity (Figure 10). All probes on the array with high similarity to the Hawaiian Dengue I genome, i.e. H $\leq$ 2 (n = 942) or MCM $\geq$ 27 (n = 627), hybridized with median signal intensity 3 SD above detection threshold. Although 98% of probes were detectable at the low HD range from 0-4, or high MCM range from 18-40, median probe signal intensity decreased at every increment of sequence distance (Figure 10). Median signal intensity dropped off sharply to background levels at HD = 7 and MCM = 15, with 43% and 46% detectable probes, respectively. The majority of probes (> 96%, n > 51,000) had HD scores between 8-21 and/or MCM scores between 0-15, of which only 1.23% and 1.57% respectively, were detectable.

At the optimal similarity thresholds HD $\leq$ 4 and MCM $\geq$ 18, > 98% of probes could be detected with median signal intensity 2 SD above detection threshold, whereas adjusting the similarity threshold down 1 step to HD $\leq$ 5 and MCM $\geq$ 17 would result in only ~ 85% probe detection and median signal intensity ~ 1.2 SD above detection threshold (Figure 10). Using these optimal HD and MCM thresholds to guard against cross-hybridization, we binned all probes into specific "recognition signature probe sets" (i.e., r-signatures) most likely to specifically detect a given pathogen, and we defined r-signatures for each of the 35 pathogen genomes represented on the array. Each pathogen's r-signature comprised tiling probes derived from its genome sequence (HD = 0, MCM = 40), as well as cross-hybridizing probes derived from other pathogens (HD $\leq$ 4, MCM $\geq$ 18). According to these criteria, a given probe could belong to multiple different r-signatures, thereby maximizing probe-level evidence for pathogen detection.

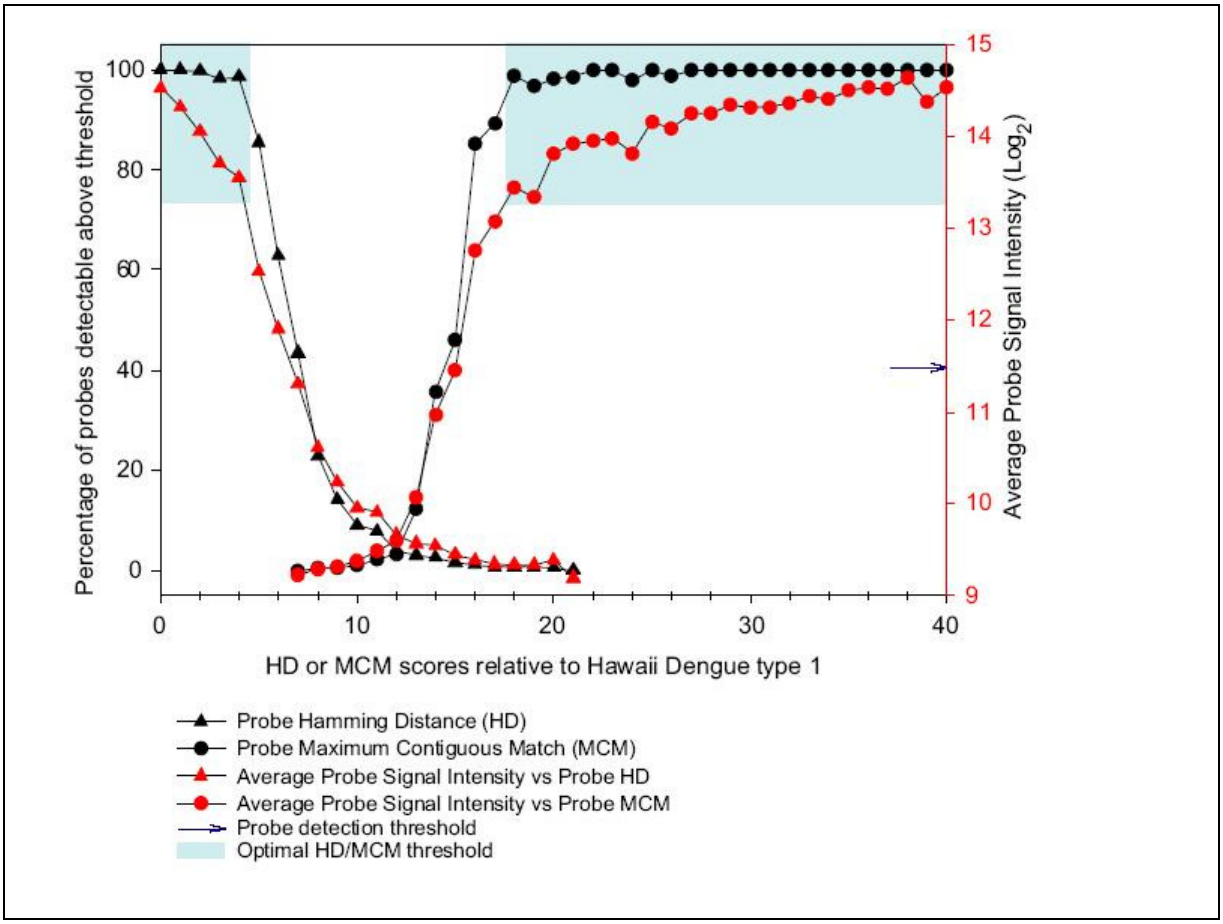

Figure 10: Relationship between probe Hamming Distance (HD), probe Maximum Contiguous Match (MCM) and probe signal intensity. Average probe signal intensity and percentage of detectable probes (signal intensity > mean + 2 SD) decreases as HD increases and MCM decreases. The optimal cross-hybridization thresholds HD ≤ 4 or MCM ≥ 18, where > 98% of probes can be detected, is shaded in blue.

We next considered other non-specific hybridization phenomena that could affect performance of our r-signature probes. For example, we observed a linear relationship between probe signal and %GC content. Consistent with previous observations, we found that probes < 40% GC hybridized with diminished signal intensities, while probes > 60% GC content showed higher signal intensities [50, 101]. Thus, we censored probes with GC < 40% or > 60% from the r-signatures, despite optimal HD or MCM values. Furthermore, as cross-hybridization with

human sequences could also confound results, we compared all probes to the human genome assembly (build 17) by BLAST using a word size of 15 [102]. Probes with expectation value of 100 were also censored.

While the ideal pathogen r-signature would be one where all probes would hybridize to the target sequence at detectable levels, polymorphic variation between the probes (derived from a consensus sequence) and the actual target would be expected to impede the performance of the r-signature probes at some level. To test this hypothesis, we compared the ratios of detectable to undetectable probes across all r-signatures in the context of the hybridization involving the Hawaiian Dengue 1 isolate. Although the Dengue 1 sequence used to derive the Dengue 1 r-signature was ~ 5% different from the Hawaiian isolate, the detectable probe ratio of the Dengue 1 specific probes was 151/152 (99%), 12 times higher then that for the nearest Dengue serotype signature, suggesting that moderate polymorphic variation is quite tolerable, allowing, in this case, for discernment of the correct pathogen.

3.2.2 Genome-wide Amplification Bias and its Implications on Viral Detection

Random priming amplification, rather than primer-specific amplification is preferred for identifying unknown pathogens in clinical specimens. However, in initial experiments using random priming amplification to identify known pathogens, we frequently observed incomplete hybridization of the pathogen genome marked by interspersed genomic regions not detected by the probes. An example of a hybridization heatmap from a microarray experiment involving the amplification of respiratory syncytial virus (RSV) B using a commercially available tagged-

random primer (5'-GTT TCC CAG TCA CGA TAN NNN NNN-3') is shown in Figure 11a. In preliminary analyses, sequence polymorphisms, probe GC content and genome secondary structure failed to explain why most probes did not light up, suggesting that it might be due to a PCR-based amplification bias stemming from differential abilities of the random primers to bind to the viral genome at the reverse transcription (RT) step. Hence, we used LOMA (described in chapter 2) to design an alternative primer (5-TAG GTC GGT GGG TAG GTN NNN NNN-3') that had a much better AES (that takes into account dimer formation and melting temperatures [77, 103]) than the original primer. The microarray experiment was then repeated using the primer designed by LOMA for amplification of RSV B (Figure 11b). From Figure 11, it is clear that amplification of RSV B using the primer designed by LOMA resulted in a dramatic increase in the number of lighted-up probes as opposed to using the original primer. A similar observation was made when the two primers were used to amplify another virus (hMPV).

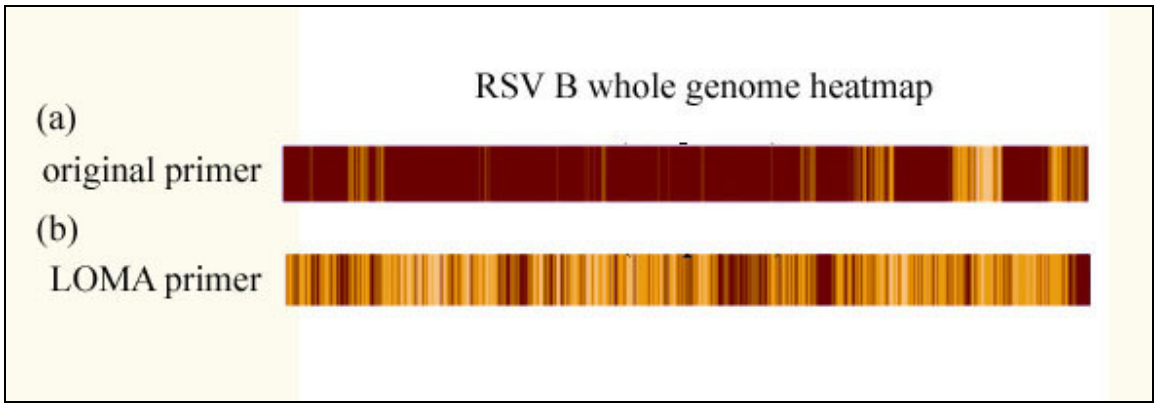

Figure 11: Heatmap of probe signal intensity for a RSV B sample following random RT-PCR by original primer and LOMA designed primer. Red regions correspond to probes that did not have signal intensities above threshold. As probe signal intensity increases, the heatmap changes from red to orange to yellow to white. LOMA primer performs significantly better than the original primer as there are less red regions indicating that most probes had high signal intensities.

The discovery and validation of amplification bias when using tagged-random primers have serious implications on viral detection via microarrays. We have shown that PCR amplification bias could cause entire regions of probes, regardless of probe homology, specificity and sensitivity properties, to fail to hybridize to its target sequences. Hence, proper probe design is no longer sufficient for an accurate interpretation of the microarray data. The tagged-random primer used for amplification must also be chosen with care. By ensuring the chosen tagged-random primer has uniformly high AES over whole genome sequences of the target viruses, we can reduce the chances of probe signal inaccuracies and ultimately improve viral detection accuracy and confidence.

3.3 PDA – A Statistical-Based Algorithm for Virus Detection

We observed that while the signal intensities for all pathogen r-signatures approximate a normal distribution (Figure 12a). We reasoned that analysis of the tails of the signal intensity distributions for each r-signature might better enable not only the identification of an infecting pathogen, but also the presence of co-infecting pathogens in the same sample. Thus, we devised a robust statistics-based pathogen detection algorithm (PDA), which analyzes the distribution of probe signal intensities relative to the *in silico* r-signatures. The PDA software comprises 2 parts: (1) Evaluation of signal intensity of probes in each pathogen r-signature using a modified Kullback-Leibler Divergence (KL), and (2) statistical analysis of modified KL scores using the Anderson-Darling test.

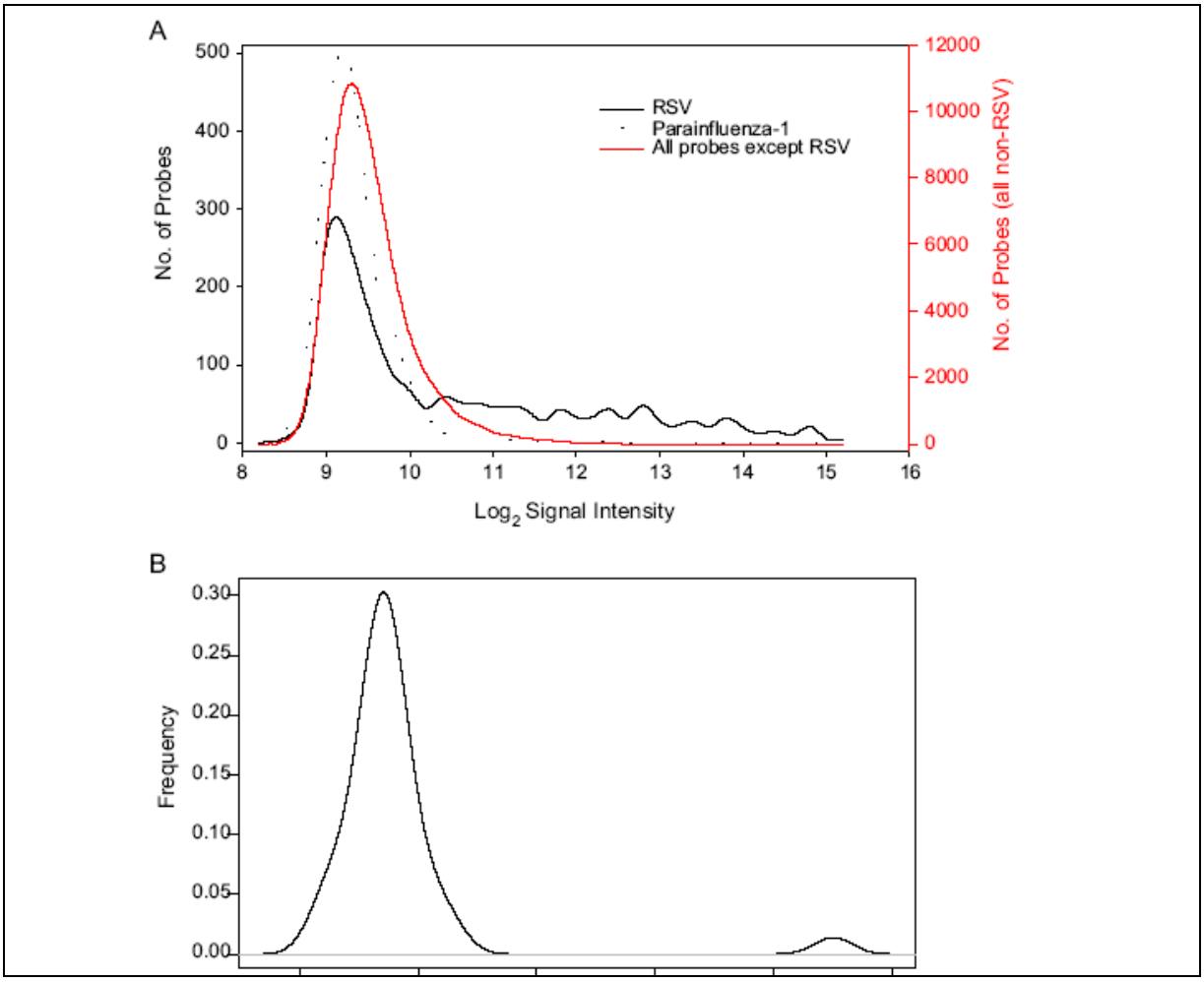

Figure 12: Distribution of probe signal intensities and WKL scores. RNA isolated from a RSV-infected patient was hybridized onto the array. (A) Distribution of probe signal intensities of all 53555 probes (red) and r-signature probes for an absent pathogen, eg. parainfluenza-1 (dotted line) show a normal distribution. The distribution of signal intensity for RSV r-signature probes are positively skewed, with higher signal intensities in the tail of the distribution. (B) Distribution frequency of WKL scores for the 35 pathogen r-signatures with majority ranging between -5 and 3. A non-normal WKL score distribution is observed ($P<0.05$ by Anderson Darling test). The presence of a pathogen is indicated by a non-normal distribution caused by outlier WKL=17, corresponding to RSV. Excluding the RSV r-signature WKL score results in a normal distribution. From this computation, we conclude that RSV is present in the hybridized sample.

Consider the virus v_a. Let P_a be the set of probes of a virus v_a and $\overline{P_a} = P - P_a$. Let $[r_{low}, r_{high}]$ be the signal intensity range. We partition it into c bins

$[r_{low} + j(\frac{r_{high} - r_{low}}{c}), r_{low} + (j+1)(\frac{r_{high} - r_{low}}{c})]$ for j = 0, 1, …, c-1. The unmodified Kullback-Leibler divergence can be computed by $KL(P_a \mid \overline{P_a}) = \sum_{j=0}^{c-1} f_a(j) \log(\frac{f_a(j)}{f_{\bar{a}}(j)})$ where $f_a(j) = \frac{n_a^j}{\sum_{h=0}^{c-1} n_a^h}$ is the fraction of probes in P_a found in bin b_j, $f_{\bar{a}}(j) = \frac{n_{\bar{a}}^j}{\sum_{h=0}^{c-1} n_{\bar{a}}^h}$ is the fraction of probes in $\overline{P_a}$ found in bin b_j, n_a^j and $n_{\bar{a}}^j$ are the number of probes in P_a and $\overline{P_a}$ in the bin b_j respectively.

To compare the signal difference of the tail of the probability distribution, we set $r_{low} = \overline{\mu_a}$, the mean signal intensity of the probes in $\overline{P_a}$, and r_{high} = maximum signal intensity. We set the default number of bins, c = 20. Since the original KL cannot reliably determine differences in the tails of a probability distribution, and is highly dependent on the number of probes per genome and the size of each signal intensity bin, we incorporated the Anderson-Darling statistic to give more weight to the tails of each distribution. By using a cumulative distribution function instead of the original probability distribution, the p-value generated is independent of the binning criteria, eliminating errors which occur if a particular signal intensity bin is empty [104, 105]. We call our modified KL divergence the Weighted Kullback-Leibler divergence (WKL):

$$WKL(P_a \mid \overline{P_a}) = \sum_{j=0}^{k-1} \left[w(j) \frac{Q_a(j) \log(\frac{Q_a(j)}{Q_{\bar{a}}(j)})}{\sqrt{Q_{\bar{a}}(j)\left[1 - Q_{\bar{a}}(j)\right]}} \right]$$

where $Q_a(j)$ is the cumulative distribution function of the signal intensities of the probes in P_a found in bin b_j; $Q_{\bar{a}}(j)$ is the cumulative distribution function of the signal intensities of the probes in $\overline{P_a}$ found in bin b_j. R-signatures representing absent pathogens should have normal signal intensity distributions and thus relatively low WKL scores, whereas those representing present pathogens should have high, statistically significant outlying WKL scores (Figure 12b).

Next, we claim that the distribution of WKL scores of all viruses $v_a \in V$ is approximately normal if there is no virus present in a sample. We empirically verify if our claim is correct by a bootstrapping process: Let n be the number of viruses in V. For each virus $v_k \in V$ where $k = 1, \ldots, n$, we choose $|v_k|$ probe signal intensities from a real dataset randomly with replacement to form a "perturbed" signal intensity distribution of v_k. Such distribution can mimic the situation where virus v_k is not present in the sample. Thereafter, n WKL scores are generated for the set of n viruses. Next, we check if the n WKL scores follow a normal distribution by the Anderson-Darling test for normality at 95% confidence interval. The bootstrap is repeated 100,000 times. The distribution is found to be normal in more than 99% of the time. (NB: since there are 35 viral genomes represented on our microarray, n =35)

Based on the above discussion, we can test if a sample contains virus(es) by making the following null and alternative hypothesis:

H_0: The distribution of WKL scores is normal, i.e. viruses are not present in the sample.

H_1: The distribution of WKL scores is not normal, i.e. at least 1 virus is present in the sample.

We proceed to apply the Anderson-Darling test for normality on the distribution of WKL scores to reject H_0 with 95% confidence interval. If $p < 0.05$, the WKL distribution is considered not

normal, implying that the pathogen with outlying WKL scores is present. Upon identification of a pathogen, that pathogen's WKL score is left out, and a separate Anderson-Darling test is performed to test for the presence of co-infecting pathogens. In this manner, the procedure is iteratively applied until only normal distributions remain (i.e., $p > 0.05$). The PDA algorithm is extremely fast, capable of making a diagnosis from a hybridized microarray in less than 10 seconds. Figure 13 shows the pseudo-code for our virus-detection algorithm.

Given a pathogen microarray data D with virus set V and probe set P,

Let $V_{\text{present}} = \Phi$

Let D_{WKL} be the set of $WKL(P_v \,\|\, P_{\bar{v}})$ for all $v \in V$;

1. Determine normality of D_{WKL} with Anderson Darling test for normality. If D_{WKL} is a normal distribution with significance level 0.05, return V_{present}. Else, go to step 2.
2. Find the virus v_a with the highest $WKL(P_a \,\|\, P_{\bar{a}})$ from D_{WKL}.
   Let $V_{\text{present}} = V_{\text{present}} \cup \{ v_a \}$; $D_{WKL} = D_{WKL} - \{ WKL(P_a \,\|\, P_{\bar{a}}) \}$; Go to step 1.
3. Remove detected r-signature and verify that WKL distribution is normal.
4. If distribution is not normal, go back to step 2 to find co-infecting pathogen.

Figure 13: Analysis framework of pathogen detection microarray data.

3.4 Microarray Performance on Clinical Specimens

To assess the clinical utility of the pathogen prediction platform, we analyzed 36 nasal wash specimens according to the workflow illustrated in Figure 14. These specimens were obtained from children under 4 years of age with lower respiratory tract infections (LRTI) of which 14 were hospitalized for severe disease and 22 with ambulatory LRTI. The clinical diagnosis of these patients was bronchiolitis or pneumonia. All 36 specimens had been previously analyzed

for the presence of hMPV, RSV A and B using real-time PCR. Twenty-one specimens tested positive for one or more viruses, while 15 were PCR-negative for all three. All specimens were analyzed by microarray in a blinded fashion (Table 1).

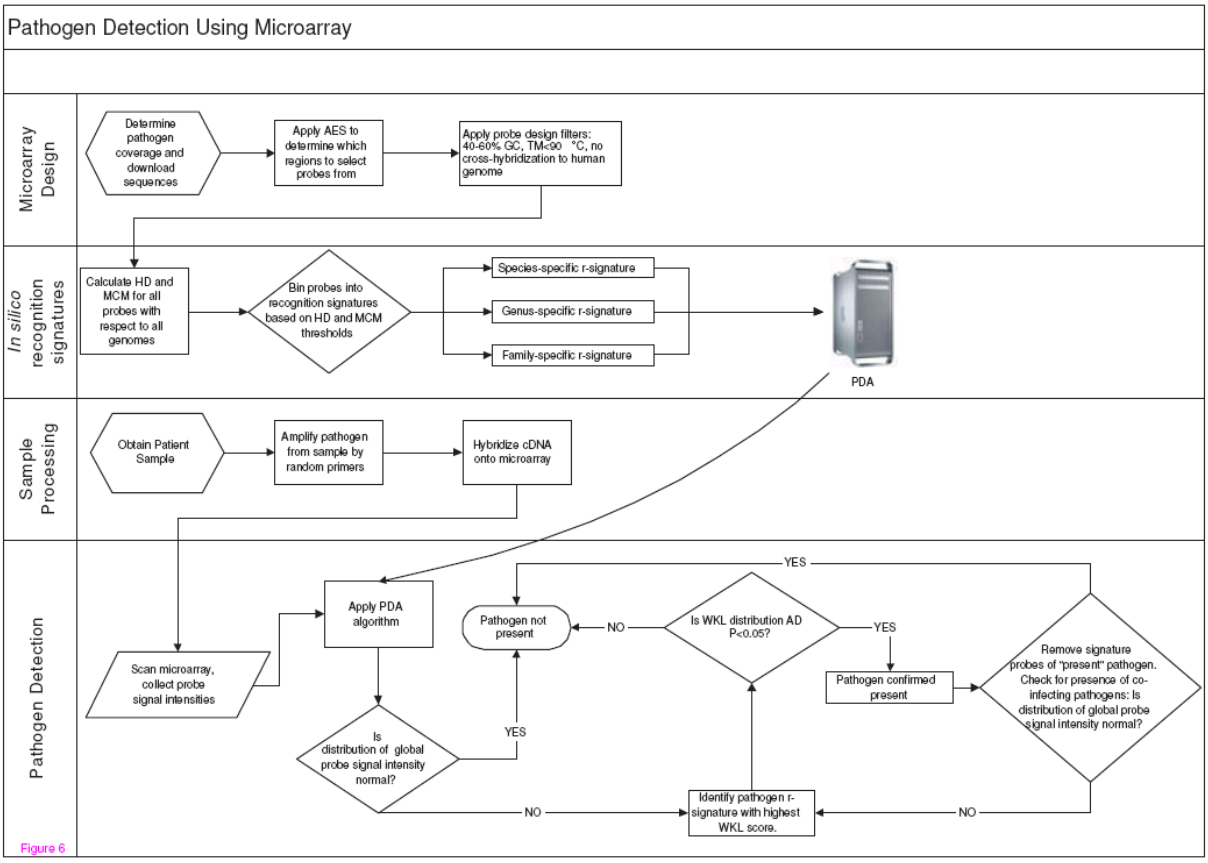

Figure 14: Schema of pathogen detection process.

Patient ID	Array	WKL	P-value	PDA genus diagnosis	PCR diagnosis	PCR Ct value	Virus copy no.
111	35915			ND	ND		
122	35887	20.87	2.47×10^{-29}	pneumovirus	pneumovirus	24.8	5.0×10^{4}
133	71180	22.33	6.93×10^{-62}	pneumovirus	pneumovirus	25.1	4.0×10^{4}
165	66691	16.95	3.49×10^{-4}	pneumovirus	pneumovirus	27.9	3.9×10^{3}
185*	66696			ND	ND		
254	70935	25.02	2.87×10^{-39}	pneumovirus	pneumovirus	22	5.4×10^{5}
261*	66697			ND	ND		
283*	63781	23.99 14.07	2.28×10^{-25} 4.66×10^{-11}	pneumovirus HRV	HRV	28.3	6.1×10^{4}
312*	66701			ND	pneumovirus†	33.7	44
321*	71006			ND	pneumovirus†	31.1	340
324*	35259	20.61	3.55×10^{-94}	pneumovirus	pneumovirus	21.4	3.0×10^{6}
331*	66698			ND	HRV	31.7	3.6×10^{3}
337	71192	21.73 8.3	3.49×10^{-14} 1.92×10^{-4}	pneumovirus HRV	pneumovirus HRV	26.2 29.1	1.1×10^{5} 3.1×10^{4}
355	35662	18.00	2.97×10^{-40}	pneumovirus	pneumovirus	20.3	6.7×10^{6}
368*	66702			ND	ND		
374	66695			ND	pneumovirus	34.1	500
378	70933	13.82	7.77×10^{-17}	pneumovirus	pneumovirus	23.9	5.4×10^{5}
393*	71189	25.41	1.15×10^{-18}	HRV	HRV	30.2	2.1×10^{5}
412	35890	19.66	2.42×10^{-49}	pneumovirus	pneumovirus	23.5	6.9×10^{5}
414	71025	49.91	1.18×10^{-65}	pneumovirus†	pneumovirus† HRV	22.3 33	3.9×10^{5} 2.6×10^{3}
461	66699			ND	ND		
478	71027			ND	pneumovirus†	34.8	18
483*	36053	12.17	1.47×10^{-12}	pneumovirus	pneumovirus	24.8	2.9×10^{5}
554	70997	78.55	4.59×10^{-120}	HRV	HRV	23.5	1.5×10^{6}
573	66700	38.09	6.26×10^{-22}	HRV	HRV	22.2	3.6×10^{6}
639*	71182	9.23	7.91×10^{-6}	HRV	ND		
699	71007			ND	ND		
769	73067	24.62	3.70×10^{-52}	pneumovirus	pneumovirus	25.7	2.5×10^{4}
818	70927	10.40	1.63×10^{-8}	HRV	HRV	34.2	1.2×10^{3}
832	73068	13.52 40.43	4.54×10^{-6} 1.73×10^{-36}	pneumovirus pneumovirus†	pneumovirus pneumovirus†	28.2 23.8	3.1×10^{3} 1.2×10^{5}
841	73070	22.11	6.80×10^{-50}	pneumovirus	pneumovirus HRV	20.9 35.4 29.2	4.5×10^{6} 8 3.3×10^{4}
853*	66690			ND	ND		
859	71188	72.17	1.42×10^{-128}	HRV	HRV	24.5	2.8×10^{6}
892*	68359	12.43	5.77×10^{-5}	HRV	pneumovirus HRV	34 32.3	27 4.2×10^{3}
913	71028	40.67	1.60×10^{-50}	pneumovirus†	pneumovirus†	19.1	4.7×10^{6}
924*	66703	12.79	2.56×10^{-6}	pneumovirus†	pneumovirus† pneumovirus	31.5 33.7	250 630

Table 1: Comparison of microarray and real-time PCR performance in detection of pathogen genera (HRV, pneumovirus). ND=none detected. *Hospitalized patients. †RSV A patient samples.

As RSV A full-genome sequence is not published, our array was not designed to specifically detect this virus. Thus we first assessed array performance using only results from the 16 patients diagnosed with either hMPV or RSV B by PCR (Table 2). Of this cohort, the microarray correctly detected the presence of hMPV or RSV B in 13/16 samples. This corresponds to an assay specificity of 100%, sensitivity of 76%, and diagnostic accuracy of 94%. All 4 false negative samples (patient #374, 841, 892, 924) had Ct values > 33.5, which is near the detection limit of real-time PCR, and thus perhaps beyond the range of detection by microarray.

Patient ID	Array	WKL	P-value	PDA diagnosis	PCR diagnosis	PCR Ct value	Virus copy no.
122	35887	20.87	2.47×10^{-29}	hMPV	hMPV	24.8	5.0×10^{4}
133	71180	22.33	6.93×10^{-62}	hMPV	hMPV	25.1	4.0×10^{4}
165	66691	16.95	3.49×10^{-4}	hMPV	hMPV	27.9	3.9×10^{3}
254	70935	25.02	2.87×10^{-39}	hMPV	hMPV	22	5.4×10^{5}
769	73067	24.62	3.70×10^{-52}	hMPV	hMPV	25.7	2.5×10^{4}
832	73068	13.52	4.54×10^{-6}	hMPV	hMPV	28.2	3.1×10^{3}
892*	68359			ND	hMPV	34	27
324*	35259	20.61	3.55×10^{-94}	RSV B	RSV B	21.4	3.0×10^{6}
355	35662	18.00	2.97×10^{-40}	RSV B	RSV B	20.3	6.7×10^{6}
374	66695			ND	RSV B	34.1	500
378	70933	13.82	7.77×10^{-17}	RSV B	RSV B	23.9	5.4×10^{5}
412	35890	19.66	2.42×10^{-49}	RSV B	RSV B	23.5	6.9×10^{5}
483*	36053[1]	12.17	1.47×10^{-12}	RSV B	RSV B	24.8	2.9×10^{5}
924*	66703			ND	RSV B	33.7	630
337	71192	21.73	3.49×10^{-14}	RSV B	RSV B		1.1×10^{5}
841	73070	22.66	4.21×10^{-50}	RSV B	RSV B hMPV	20.9 35.4	4.4×10^{6} 8

Table 2: Comparison of microarray and real-time PCR performance in detecting RSV B or hMPV. ND=none detected. *Hospitalized patients.

We next assessed array performance in the group of patients PCR-positive for RSV A (n=7) and PCR-negative for all tested viruses (n=15). The microarray made only 2 positive calls in this group, both for RSV B. Interestingly, both RSV B calls corresponded to high-titre RSV A

specimens by PCR (#414, 913) suggesting that certain probe sets can detect the presence of related, but unspecified, viruses. Analysis of the published RSV A partial genome sequence (923 bp, GenBank ID: AF516119) revealed that 7 probes on our microarray had 100% identity to RSV A. We created an "RSV A r-signature" comprising these 7 probes, enabling the specific detection of RSV A by microarray in 4/7 patient samples PCR-positive for RSV A (#414, 832, 913 and 924). Although the performance of this small r-signature was not as robust as the other virus r-signatures (median size: 249 probes), it suggested that it was feasible to pursue a "viral discovery" approach using r-signatures created to detect viruses at the family or genus level, that were related to those species already represented on the microarray. Specifically, we binned probes into family- or genus level r-signatures by relaxing our similarity criteria (to HD≤5 or MCM≥25) and selecting probes common to genome sequences within families and genera for the picornaviridae family, paramyxoviridae family, rhinovirus genus (HRV) and pneumovirus genus (inclusive of RSV and hMPV).

Upon re-analysis of all 36 samples, we identified the presence of pneumovirus in 17 specimens as expected (1 false positive, #283), and additionally detected the presence of HRV in 9 specimens (Table 1). As HRV was a novel discovery, we re-screened all 36 samples by PCR and found HRV in 11 specimens. All 9 HRV calls by microarray were confirmed by PCR except for 1. This finding was intriguing given that the genomic diversity of the over 100 known rhinovirus serotypes makes detection by PCR notoriously difficult [106]. As the real-time PCR primers were capable of identifying only ~70% of rhinovirus strains, it is possible that the microarray correctly detected a rhinovirus strain that PCR failed to detect. Similarly, the pneumovirus genus detected in patient #283 could not be verified by RT-PCR, possibly owing to subtle genetic variations that prevented primer annealing. Thus, the greater genomic coverage

afforded by the microarray might, in some cases, provide a more sensitive and accurate detection capability than pathogen-specific PCR.

Though the microarray identified the majority of HRV and RSV A samples using the genus-level r-signatures, the array failed to detect 3 samples positive for HRV and 3 positive for RSV A by real-time PCR. These false negatives had an average Ct value > 32, again suggesting a detection threshold close to that of real-time PCR. However, the microarray also made a number of accurate discoveries in the 30-35 Ct range, suggesting a considerable degree of detection variability in the titre range above a ~30 Ct equivalency. Notably, the microarray correctly detected the presence of co-infecting pathogens in 2 samples (#337, #832), demonstrating the unique potential of this microarray platform to reveal complex disease etiologies.

3.5 Conclusion

DNA microarrays have the potential to revolutionize clinical diagnostics through their ability to simultaneously investigate thousands of potential pathogens in order to make a diagnosis. However, questions remain regarding their sensitivity and reliability. In this work, we investigated the myriad of factors that influence microarray performance in the context of virus detection in clinical specimens, and describe an optimized platform capable of identifying individual and co-infecting viruses with high accuracy and sensitivity that brings microarray technology closer to the clinic.

Future improvements will include significant reductions in microarray manufacturing and usage costs. Multiplex microarray formats and “re-usable” arrays are developing technologies that promise to drive down these costs. Furthermore, alternative technologies such as beads [107], microfluidics [108, 109] and nanotube microarrays [110], might provide advantages in both assay cost and speed relative to traditional microarray platforms. Technology considerations aside, the advantages of a highly parallel, nucleic acid-based screening approach for detecting disease pathogens are clear. Validations in larger patient cohorts and in diverse clinical settings will be an important next step towards establishing the clinical role of pathogen detection microarrays.

Chapter 4

RESEQUENCING OF VIRAL GENOMES

4.1 Resequencing Microarrays as a Large-scale Bio-surveillance Tool

Historically, sequencing of viral genomes is performed using standard dye termination technologies. These conventional sequencing technologies produce accurate data but are too slow, costly, labour-intensive and low throughput to be practical for large-scale epidemiologic or evolutionary investigations in viral outbreaks. In recent years, next-generation sequencing technologies that can produce millions of sequences at once have emerged. Through massive parallelization of the sequencing process, technologies such as 454 are able to achieve high throughput at a progressively lower cost. However, next-generation technologies that are more suited for deep sequencing of a few samples, become less cost-effective when used to sequence a large number of samples. Oligonucleotide resequencing microarrays that are capable of identifying nucleotide sequence variants may offer a low-cost rapid solution for whole-genome sequencing of viruses [111].

In April 2009, a novel influenza virus (H1N1) emerged from Mexico and rapidly spread to the rest of the world [112]. The global infection quickly forced the World Health Organization (WHO) to declare the outbreak an pandemic. As of December 2009, more than 414000 confirmed cases and nearly 5000 deaths worldwide have been reported (http://www.who.int/en/). Nevertheless, the 2009 H1N1 pandemic did provide an unique opportunity to find out if

resequencing microarrays can be used as a practical, large-scale tool for bio-surveillance. In fact, resequencing microarrays have already been used for detecting and subtyping influenza viruses in recent years [113, 114]. However, only sequences from partial fragments of the hemagglutinin (HA) and neuraminidase (NA) genes were obtained for analysis. Hence, the full genome sequencing of a influenza A virus using microarrays remains a novelty.

To address if microarrays can be used as a practical, large scale resequencing tool, we have developed a system comprising customized sequence amplification primers, a 12-plex DNA resequencing array for 2009 influenza A H1N1 and an automated base-calling and variant analysis software (EvolSTAR). In subsequent sections, we describe the development of the various genetic analysis components, and their validation using clinical samples.

4.2 Design of Resequencing Microarrays

We generated a consensus sequence for each segment of the H1N1(2009) virus by aligning all 1715 complete and partial sequences available from the NCBI H1N1 flu resources database (http://www.ncbi.nlm.nih.gov/genomes/FLU/SwineFlu.html) as of June 11th 2009 using MAFFT [115] with high accuracy option.

Tiling probes spanning the entire genome segments on both the forward and reverse strands were created at 1 base resolution [50]. Analysis of the sequence alignments revealed that there were no deletions, insertions or recombination. However, we found 36 mutation hotspots in the alignments where mutations occurred near one another (within 20 bp). A perfect match (PM) probe residing in a mutation hotspot may contain mismatches that will have a detrimental effect

on its hybridization intensity. To avoid this problem, we designed additional PM probes that contain all possible combinations of mutations found in each mutation hotspot. Thus, if 2 mutations are found within 20bp of each other in the alignments, then we need in total 4 (2^2) PM probes to encode them. In general, 2^x PM probes are needed to completely encode a cluster of x mutations that occur within 20bp of one another in the alignments.

Furthermore, to ensure that we have accurate sequence of the drug binding pocket targeted by neuraminidase inhibitors [116] such as oseltamivir (Tamiflu®) and zanamivir (Relenza®) in the neuraminidase gene of the H1N1(2009) virus, additional probes were added. In total, the array contains 8,236 control probes and 121,928 H1N1(2009) probes, which provides 2x coverage of the entire H1N1(2009) genome, and up to 8x coverage of the regions comprising the 36 mutation hotspots and 10 drug-binding sites.

4.3 Optimization of RT-PCR Primers and Conditions

Due to the small amount of virus present in samples relative to human or cell-line total RNA, it was necessary to amplify the viral RNA through PCR. We employed a combination of sequence-specific and random PCR approaches using LOMA-optimized primers as previously described in chapter 2. The addition of random primers ensured complete genome amplification, even if mutations were present at the specific-primer binding sites. PCR conditions were optimized by conducting 5 duplicate hybridizations of the same virus sample cultured from a patient sample under different PCR conditions. The optimized method was then tested on RNA isolated directly

from nasal swabs obtained from the same patient and from virus grown in cell culture. Microarray sequences generated from these replicate experiments were compared with capillary sequencing to estimate sequencing accuracy.

4.4 Evolution Surveillance and Tracking Algorithm for Resequencing arrays

Following PCR product labeling, hybridization and scanning, signal intensities for each probe was generated using Genepix 4.0 software, and annotated using Nimblescan 2.5 software. Initially, the standard Nimblescan software which employs a gain-of-signal approach (PBC algorithm [50]), was used to determine the viral sequence. The PBC algorithm assumes that the signal intensity of the perfect match (PM) probe (which matches exactly to the sequence in the sample) will be significantly higher than that of the mismatch (MM) probes. While this approach sufficed for ~90% of base queries, we observed that the discrimination between the PM and MM signals was not clear for the remaining probes.

These ambiguous signals were caused by the presence of multiple mutations in the probe sequence, homopolymers and hybridization artifacts. We developed a novel algorithm, Evolution Surveillance and Tracking Algorithm for Resequencing arrays (EvolSTAR), to resolve this problem. EvolSTAR improves upon PBC by adding an analysis of the neighbourhood hybridization signal intensity profile (NHIP) and nucleotide substitution bias.

4.4.1 Neighbourhood Hybridization Intensity Profile

Due to the use of tiling probes in resequencing arrays, a single nucleotide mutation at a particular query base could cause a dramatic reduction in the hybridization intensities of neighbouring PM probes up to 6 bases away [58]. This effect can be measured by studying the neighbourhood hybridization intensity profile (NHIP) of each query base. We defined the NHIP of each query base as the observed pattern of hybridization intensities of its PM and MM probes and neighbouring (± 6 bases from query base) PM and MM probes. To study the effects of sequence variation (mutation) and noise on the NHIP of a query base, we sequenced RNA from H1N1(2009) patient 380 by capillary sequencing and on duplicate microarrays. We compared sequence calls generated using by Nimblescan or by capillary sequencing and compiled a list of true (correct) calls, error calls and 'N' (unknown) calls. In total, of the expected 13588 bases of the H1N1 virus (based on genome described at http://www.ncbi.nlm.nih.gov/genomes/taxg.cgi?tax=211044) the microarray called 13449 bases while capillary sequence was able to call 12832 bases.

By analyzing base calls from PBC that have been confirmed by capillary sequencing, we identified five distinct types of neighbourhood hybridization intensity profile belonging to true non-mutations (wild-type), true mutations, isolated errors/'N's, long consecutive errors/'N's, and unknown errors/'N's respectively. For each non-high confidence query base, we determine the type of its NHIP by the following criteria:

a) **True-non-mutation** – The PM probe (of both strands) of the query base must be a high confidence call (i.e., it has hybridization intensity ≥ 1.4 fold that of its MM probes). Neighbourhood PM probes are also high confidence calls. Let the mean hybridization

intensity of the three nearest PM probes to the immediate left of the mutation base (at position -1, -2 and -3), denoted as $\mu_{\{-1,-2,-3\}}$, the mean hybridization intensity of the three PM probes to the far left of the mutation base (at position -4, -5 and -6), denoted as $\mu_{\{-4,-5,-6\}}$, the mean hybridization intensity of the three nearest PM probes to the immediate right of the mutation base (at position 1, 2 and 3), denoted as $\mu_{\{1,2,3\}}$, and the mean hybridization intensity of the three PM probes to the far right of the mutation base (at position 4, 5 and 6), denoted as $\mu_{\{4,5,6\}}$. We impose that $\mu_{\{-1,-2,-3\}} \approx \mu_{\{-4,-5,-6\}}$ and $\mu_{\{1,2,3\}} \approx \mu_{\{4,5,6\}}$.

b) **True-mutation** – The PM probe (of both strands) of the query base must have hybridization intensity ≥ 1.4 fold that of its MM probes. To detect the characteristic dip, we check 4 mean hybridization intensities: the mean hybridization intensity of the three nearest PM probes to the immediate left of the mutation base (at position -1, -2 and -3), denoted as $\mu_{\{-1,-2,-3\}}$, the mean hybridization intensity of the three PM probes to the far left of the mutation base (at position -4, -5 and -6), denoted as $\mu_{\{-4,-5,-6\}}$, the mean hybridization intensity of the three nearest PM probes to the immediate right of the mutation base (at position 1, 2 and 3), denoted as $\mu_{\{1,2,3\}}$, and the mean hybridization intensity of the three PM probes to the far right of the mutation base (at position 4, 5 and 6), denoted as $\mu_{\{4,5,6\}}$. If $\mu_{\{-1,-2,-3\}} < \mu_{\{-4,-5,-6\}}$ and $\mu_{\{1,2,3\}} < \mu_{\{4,5,6\}}$, we say this is a dip pattern and the query base is likely to be mutated.

c) **Isolated error/'N'** – The PM probe (of both strands) of the query base has hybridization intensity < 1.4 fold that of its MM probes. Neighbourhood PM probes are high confidence calls.

d) **Long consecutive errors/'N's** – The PM probe (of both strands) of the query base has hybridization intensity < 1.4 fold that of its MM probes. A majority of neighbourhood PM probes are non-high confidence calls.

e) **Unknown error/'N'** – All other neighbourhood hybridization profile patterns that do not fall under the previous categories.

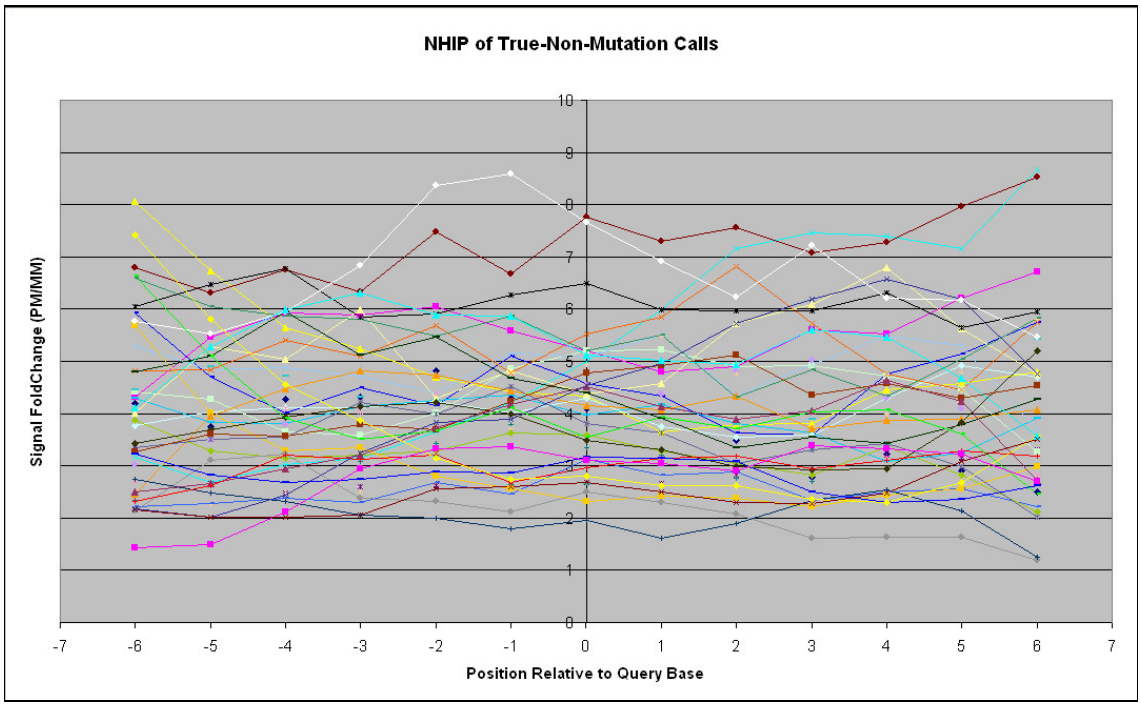

Figure 15: Observed neighborhood hybridization intensity profiles for true-non-mutation calls. A representative set of observed NHIPs for true-non-mutation calls from patient sample 380. This representative set consists of 5 true-non-mutation calls randomly selected from each segment. Each line represents the NHIP (± 6 bp from query base position) of a true-non-mutation call.

Figure 15 shows the NHIPs of a representative set of 40 randomly selected query bases that result in true-non-mutation calls (wild-type calls). We observed that in these NHIPs, the PM probe of the query base together with neighbouring PM probes, have hybridization intensities

significantly higher (> 1.4-fold) than that of their MM probes in general. We also identified 10 mutations using capillary sequencing in the patient sample. The NHIPs of these 10 true-mutation calls (Figure 16) are very different from NHIPs of wild-type calls. The presence of a mutation at the query base created a mismatch in neighbouring PM probes and caused a drop in their hybridization intensities. The closer this mutation is to the centre of a neighbouring PM probe, the bigger the drop in hybridization intensity. This results in a distinctive dip to the immediate left and right of the centre of the NHIP where the mutation is.

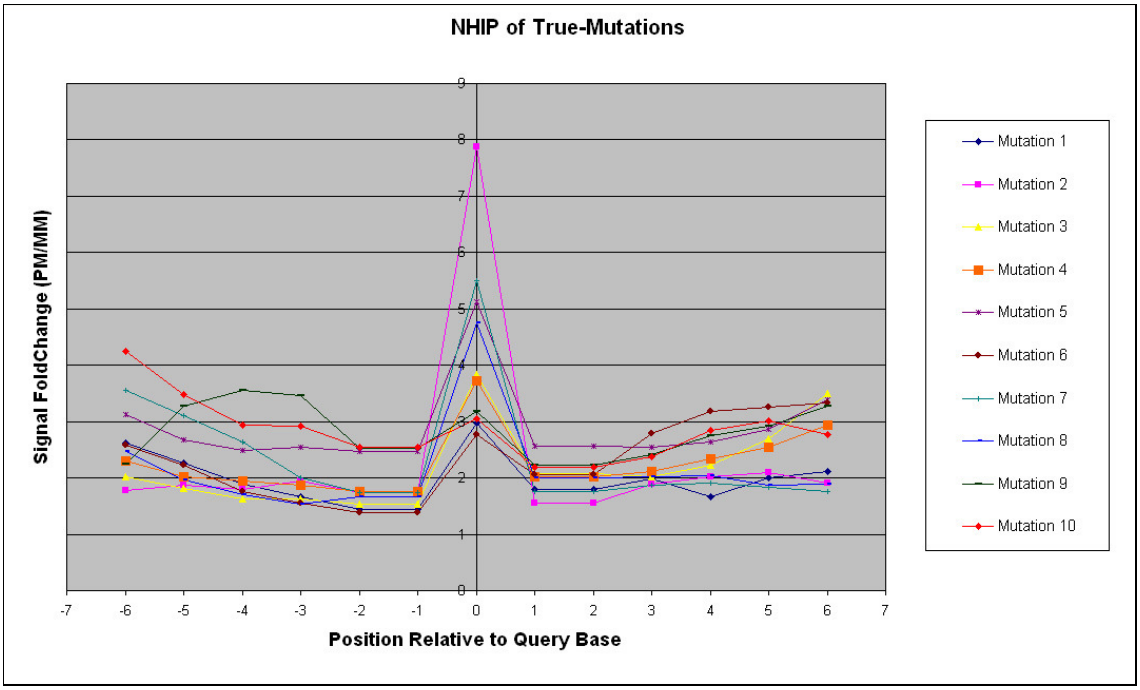

Figure 16: The observed NHIPs for all 10 identified true-mutation calls from patient sample 380.

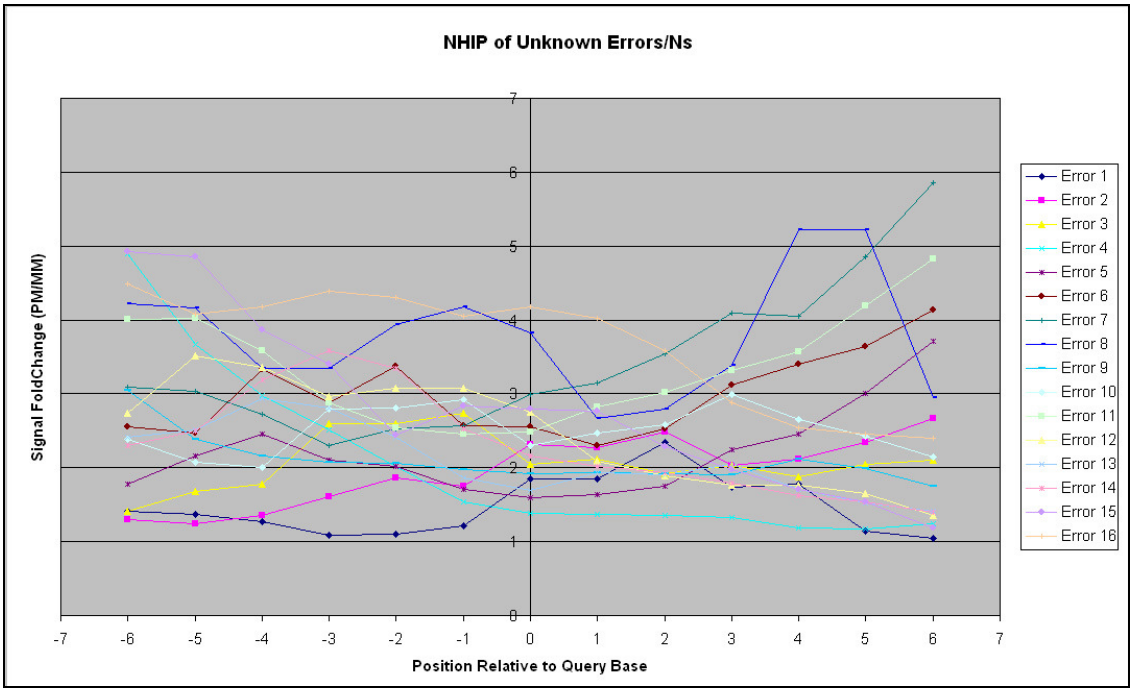

Figure 17: Observed neighborhood hybridization intensity profiles for unknown error/'N' calls. A representative set of observed NHIPs for unknown error/'N' calls from patient sample 380. This representative set consists of 2 unknown error/'N' calls randomly selected from each segment.

Unlike the NHIPs of wildtype and true-mutation calls, the NHIPs of most errors and 'N' calls appear haphazard (Figure 17). However, when we traced the locations of these errors and 'N' calls on the genome, we found that some are isolated among good calls while others are conjugated in a small locality of the genome. We investigated the NHIPs of isolated errors and 'N' calls that occurred among good calls and found that in these NHIPs, only the PM probe of the query base that is an error or 'N' call has poor hybridization differentiation with its MM probes while other PM probes have hybridization intensities significantly higher than that of their MM probes in general (Figure 18). This suggests that for such calls, only the PM and MM probes of the query base are noisy while neighbouring PM and MM probes are unaffected. In addition, we also found that long chains of consecutive error and 'N' calls (especially at the 5' and 3' end of the sample sequences) often have NHIPs where the PM probe of the query base

together with neighbouring PM probes, have poor hybridization differentiation with their MM probes (Figure 19). These error and 'N' calls usually occur at the ends of the genome segments. In summary, NHIP analysis showed that all true mutation calls had a characteristic profile (Figure 20b) that differed from wild-type sequence calls (Figure 20a). Ambiguous calls arising from different causes, such as homopolymers, isolated errors and hybridization artifacts also have profiles that are distinct from true mutation profiles (Figure 20).

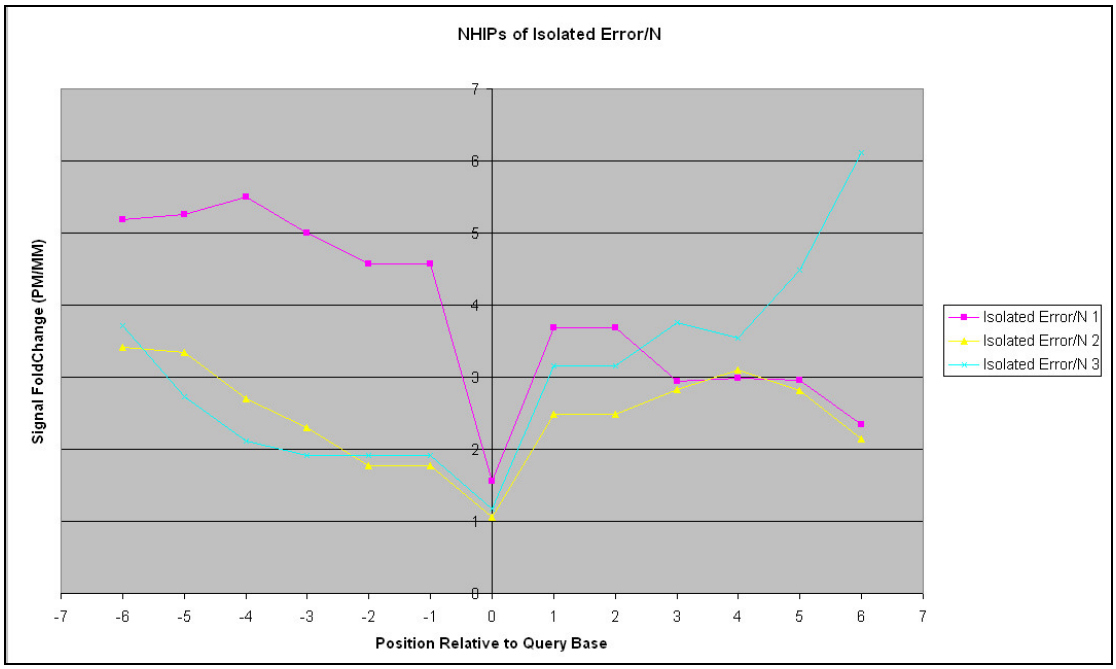

Figure 18: The observed NHIPs for all 3 identified isolated error/'N' calls from patient sample 380. These errors are flanked by true (correct) calls.

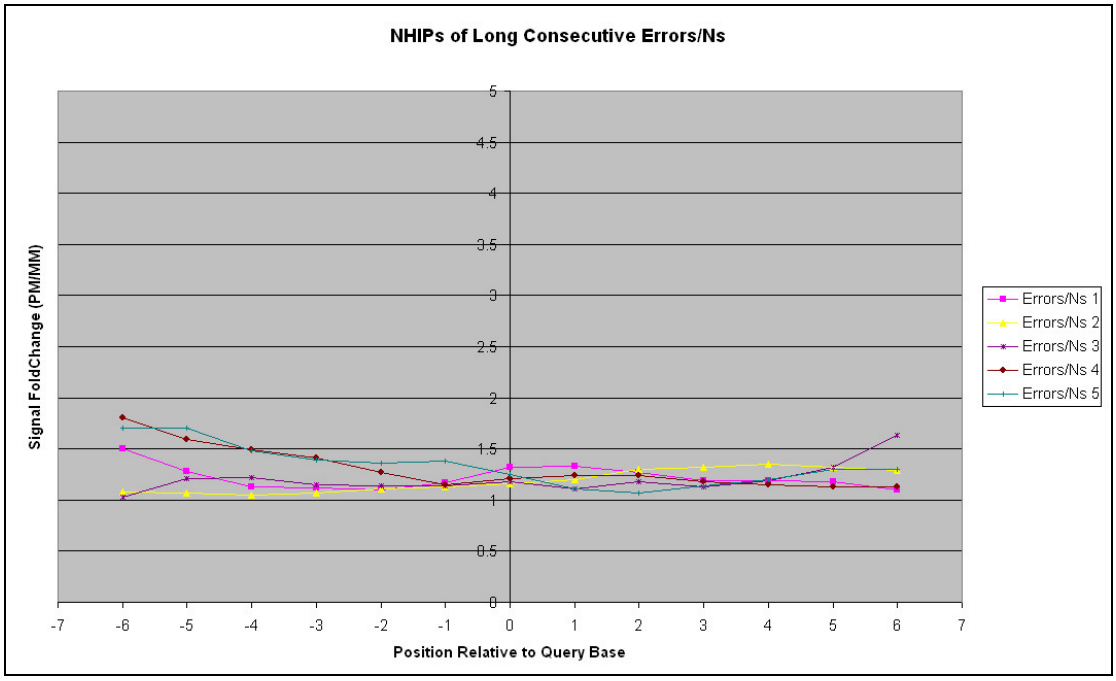

Figure 19: The observed NHIPs for 5 regions where there are long consecutive (≥ 5) error/'N' calls from patient sample 380.

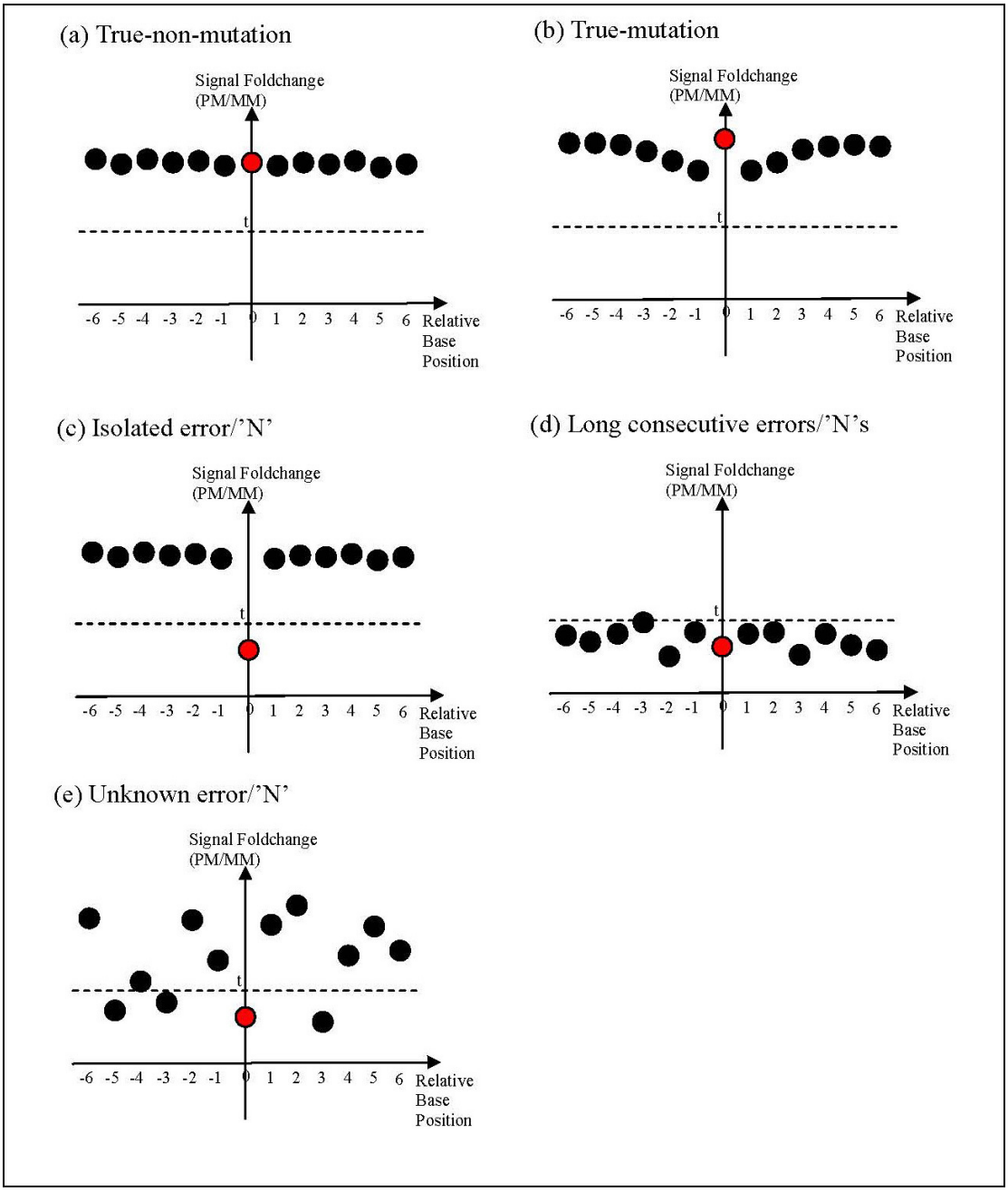

Figure 20: Summary of the characteristics of the NHIP for five types of call (true-non-mutation, true-mutation, isolated error or 'N', long chains of consecutive errors or 'N', unknown error or 'N') based on their respective observed neighbourhood hybridization intensity profiles. The PM probe (red circle) of query base is at position 0 while neighbourhood PM probes (black circles) are numbered according to their distance away from the query base. A PM probe is significantly differentiated from its MM probes if its hybridization intensity is at least t fold that of all its MM probes.

4.4.2 Nucleotide Substitution Bias

The presence of nucleotide substitution bias in Nimblegen resequencing arrays has been previously described [117]. However, this knowledge has so far been used only to improve probe design. In this paper, we propose a novel method that makes use of nucleotide substitution bias in the array to improve base-calling accuracy and call rate. The key idea is to build a likelihood model of the substitution bias among the probes of non-ambiguous calls on the array; then use this to call bases with ambiguous signals.

To build the likelihood model, we first determined the substitution bias on our platform by comparing the PM and MM probes (of both strands) of 25028 true calls made by PBC from the two replicate microarray experiments of patient sample 380 mentioned in the previous section. For each true call, we generated a hybridization intensity reduction order by ranking the PM and MM probes of a particular strand in decreasing order of hybridization intensity and recording their respective frequencies (Table 3). Table 3 shows that for each PM probe encoding, certain hybridization intensity reduction orders occur much more frequently than others. For example, if the PM probe encoding is ‘A’ (regardless of strand), then it is most likely that the hybridization intensity reduction order is ‘TGC’ or ‘GTC’. Thus, by matching the hybridization intensity reduction orders of its PM/MM probes with that in Table 3, we can compute the likelihood that the putative base call for a query base with ambiguous signals is correct.

PM probe encoding	Hybridization intensity reduction order	Forward strand	Reverse strand
		Frequency	Frequency
A	CGT	547	246
	CTG	558	237
	GCT	957	367
	GTC	2215	1407
	TCG	1049	611
	TGC	3015	2873
C	AGT	2035	2712
	ATG	1752	2400
	GAT	382	341
	GTA	159	134
	TAG	360	377
	TGA	165	129
G	ACT	1474	1043
	ATC	976	624
	CAT	1639	1534
	CTA	868	788
	TAC	594	410
	TCA	542	454
T	ACG	432	529
	AGC	562	636
	CAG	623	841
	CGA	1066	1616
	GAC	1421	1878
	GCA	1637	2841

Table 3: Hybridization intensity reduction orders found in two replicated hybridization experiments of patient sample 380. Hybridization intensity reduction orders found in 25028 true calls from two replicated hybridization experiments of patient sample 380. For each true call, for each strand, we rank the PM probe and its MM probes based on their hybridization intensities in decreasing order. We count the frequency of each hybridization intensity reduction order.

Specifically, we define that a probe encodes the base b if b is located in the centre-most position of the probe and is the base to be interrogated. For a given query base, suppose the PM probe encodes b_1 while the MM probes encode b_2, b_3 and b_4 respectively where $\{b_1, b_2, b_3,$

b_4}={A, C, G, T} and the hybridization intensity reduction order is $b_1b_2b_3b_4$. To validate if the observed PM probe encoding b_1 is indeed the true PM probe of the sample sequence, we compute the likelihood ratio of f_{obs} and f_{rand}, where f_{obs} is probability of observing the hybridization intensity reduction order $b_1b_2b_3b_4$ given that the PM probe encodes b_1 and f_{rand} is the probability of observing the hybridization intensity reduction order $b_1b_2b_3b_4$ by chance. Precisely,

$$f_{obs} = \frac{\#(b_1b_2b_3b_4)}{\#(b_1b_2b_3b_4)+\#(b_1b_2b_4b_3)+\#(b_1b_3b_2b_4)+\#(b_1b_3b_4b_2)+\#(b_1b_4b_2b_3)+\#(b_1b_4b_3b_2)}$$

and

$$f_{rand} = \frac{\#(b_1b_2)}{t} \times \frac{\#(b_2b_3)}{t} \times \frac{\#(b_3b_4)}{t}$$

where *#(wxyz)* is the number of observed hybridization intensity reduction orders from high confidence base calls and t is the total number of hybridization intensity reduction orders excluding $b_1b_2b_3b_4$ obtained from high confidence base calls. If the likelihood ratio > 2, we expect that the observed PM probe encoding b_1 is indeed the true PM probe of the sample sequence. In this way, we can recover base calls of ambiguous query bases exceeding a reasonably high likelihood threshold and achieve better accuracy and call rate than PBC.

4.4.3 Grading the Quality of the Sequence Calls

EvolSTAR employs a two-step process for base-calling (Figure 21). In the first step, each base query is scrutinized for signs of hybridization intensity abnormalities. If the gain-of-signal of the query base is strong and has no mutation, the base is called. In the second step, EvolSTAR then

tries to recover base queries that have any hybridization intensity abnormalities with two analysis methods, namely neighbourhood hybridization intensity profile analysis and nucleotide substitution bias analysis.

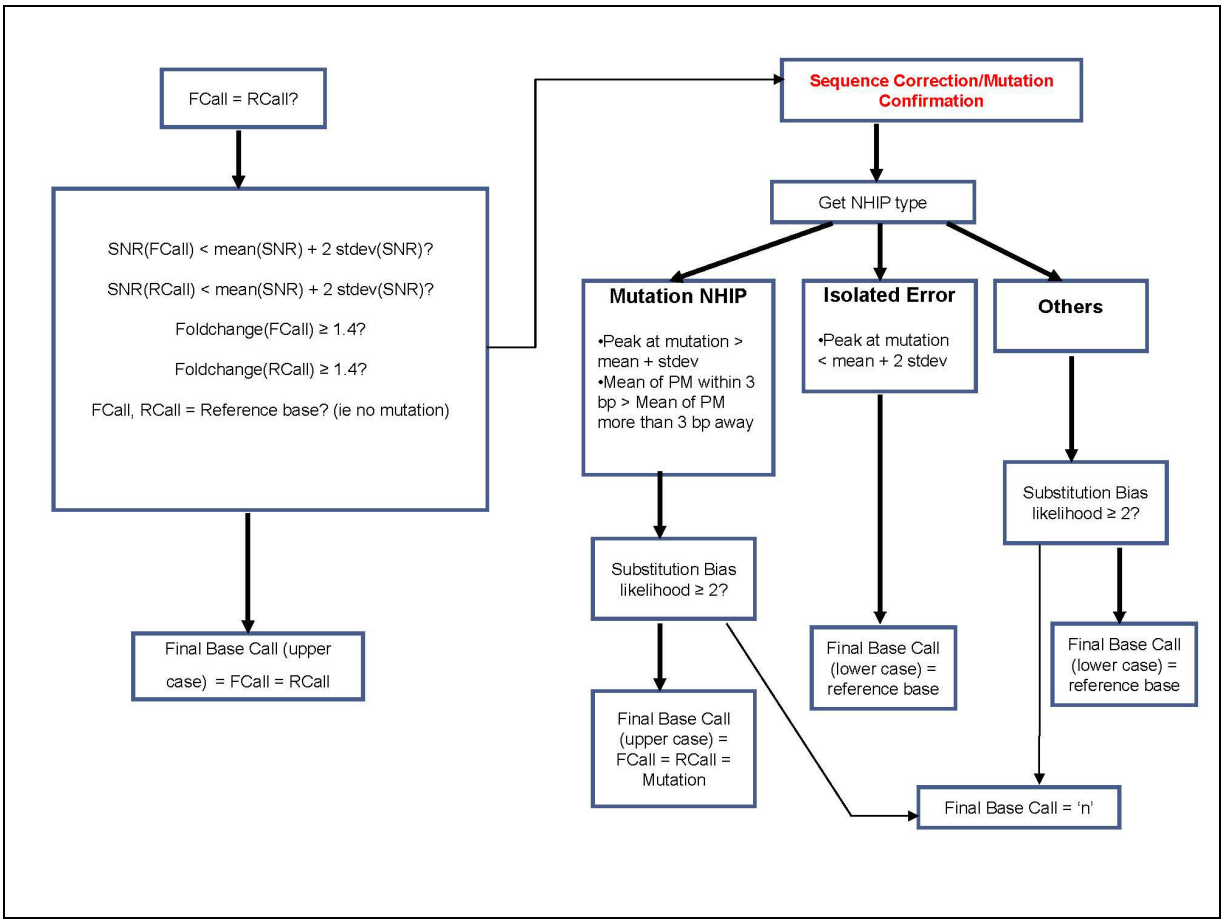

Figure 21: Flowchart of EvolSTAR. Bold arrows are "Yes" paths, while normal arrows are "No" paths. In the first step, each base query is scrutinized for signs of hybridization intensity abnormalities. Base queries with hybridization intensity abnormalities are passed to step 2 for further analysis.

Step 1: Identification of Base Queries with Ambiguity

On our array platform, the hybridization intensity of each probe is given by the mean and standard deviation of the fluorescence intensities of 9 individually scanned pixels. Hence, we

define the signal-to-noise ratio (SNR) of a probe as the ratio of the mean to the standard deviation of the intensities of the 9 pixels associated with the probe. In our experiments, we found that >95% of all probes had SNR less than T_{SNR} ($T_{SNR} = \mu_{SNR} + 2\sigma_{SNR}$ where μ_{SNR} and σ_{SNR} are the mean and standard deviation of SNR of all probes on the array). The remaining 5% of probes with SNR $\geq T_{SNR}$ are unreliable. Hence, base queries with one or more probes with SNR $\geq T_{SNR}$ are analyzed further in step 2. Furthermore, all base queries whose PM probe in the forward strand and PM probe in the reverse strand are non-complementary, or have weak PM/MM hybridization intensity differentiation (< 1.4 fold) are also passed to step 2. Lastly, we also pass all putative mutation calls to step 2 for confirmation.

Step 2: Mutation confirmation and base query recovery

A high-confidence mutation call may be a result of coincidental non-specific hybridization of the same MM probe in both strands. As such, it may be inadequate to discern true mutations based solely on differences in the hybridization intensities of PM and MM probes. From our analysis of true-mutation calls made by PBC, we have found that true mutations have a signature NHIP type as per described in Figure 20(b). Thus, query bases that result in a mutation call must have this signature NHIP. Finally, to confirm the mutation, we perform nucleotide substitution bias analysis on these query bases. For each of the query bases with NHIP of type described in Figure 20(b), we compute the likelihood ℓ that the observed PM probe (representing the mutation) is indeed the true PM probe of the sample sequence given the hybridization intensity-based ordering of its MM probes (see Method section). If $\ell > 2$, the query base results in a strong mutation call (represented by upper case base calls ‘A’, ‘C’, ‘G’ or ‘T’). If $\ell > 1$, the query base

results in a mutation call with weak support (represented by lower case base calls ‘a’, ‘c’, ‘g’ or ‘t’). Otherwise, they are re-assigned an unknown ‘N’ call.

For query bases that results in a mutation call but have NHIP of type described in Figure 20(c), they are most likely isolated errors caused by poor PM probe quality. Hence, we correct the base-calls of these query bases to their respective reference bases (but represented by lower case base calls ‘a’, ‘c’, ‘g’ or ‘t’) in the reference sequences. We also perform the same correction to non-high-confidence query bases with NHIP of type described in Figure 20(c).

We try to recover the remaining query bases that have NHIP of type described in Figure 20(d) or Figure 20(e) by analyzing the substitution bias from their PM and MM probes in the forward and reverse strands separately. Similar to how a mutation is confirmed, we compute the likelihood ℓ_f that the observed PM probe (representing the unsure base call) is indeed the true PM probe of the sample sequence given the hybridization intensity-based ordering of its MM probes in the forward strand. We also compute a similar likelihood ℓ_r for the PM probe in the reverse strand. If the PM probes in both strands are complementary and $\ell_f, \ell_r > 2$, the query base results in a strong base call (represented by upper case base calls ‘A’, ‘C’, ‘G’ or ‘T’). However, in many cases, the PM probes in both strands are not complementary due to non-specific hybridization of MM probes in one or both strands. For such query bases, we make base calls based on ℓ_f and ℓ_r: If $\ell_f > \ell_r$ and $\ell_f > 2$, a base call with weak support (represented by lower case base calls ‘a’, ‘c’, ‘g’ or ‘t’) is made from the PM probe in the forward strand. Else, if $\ell_r > \ell_f$ and $\ell_r > 2$, a base call with weak support is made from the PM probe in the reverse strand. Otherwise, they are assigned an unknown ‘N’ call.

Note that since nucleotide substitution biases may vary depending on the experimental conditions, experimental reagents or input samples, for each experiment, we obtain a set of high confidence base calls and use them to infer the hybridization intensity reduction orders for each PM probe encoding. This is then used to compute likelihood scores for base-calling non-high confidence query bases and mutation confirmation.

4.5 Performance of EvolSTAR

To validate the software, we hybridized 14 patient samples in duplicate onto the microarray. The microarrays were analyzed in parallel using Nimblescan (PBC algorithm) and EvolSTAR, and the sequences obtained were compared to Sanger capillary sequencing. We counted the number of true-non-mutation calls, true-mutation calls, error calls and ambiguous ('N') calls for both methods (Table 4). We also confirmed that the substitution bias in all 14 duplicate hybridization experiments (Table 5) were consistent with that found in Table 3. Compared with the available capillary sequences for the 14 samples, EvolSTAR had an average error rate of 0.0029% and 12 ambiguous calls per sample (346 in total). This is far superior than Nimblescan PBC, where we obtained an average error rate of 0.083% and 158 ambiguous calls per sample (4434 in total). Furthermore, EvolSTAR called all true mutations correctly. The genome coverage attained by EvolSTAR (99.02±0.82%) is also much higher than that of Nimblegen PBC (94.3±6.06%).

Sample	Program	Rep.	Total Sites Verified by Capillary	Mutations (Verified by Capillary)	True-Non-Mutation Calls	True Mutation Calls	Missed Mutations	Error Calls
129	EvolSTAR	1	4767	6	4737	6	0	0
	PBC	1	4767	6	4500	6	0	3
	EvolSTAR	2	4767	6	4737	6	0	0
	PBC	2	4767	6	4474	6	0	6
141	EvolSTAR	1	4051	6	4026	6	0	0
	PBC	1	4051	6	3832	6	0	10
	EvolSTAR	2	4051	6	4021	6	0	0
	PBC	2	4051	6	3808	6	0	4
279	EvolSTAR	1	693	2	670	2	0	0
	PBC	1	693	2	358	1	1	8
	EvolSTAR	2	693	2	682	2	0	0
	PBC	2	693	2	645	2	0	0
354	EvolSTAR	1	8950	9	8942	9	0	0
	PBC	1	8950	9	8802	9	0	1
	EvolSTAR	2	8950	9	8944	9	0	0
	PBC	2	8950	9	8851	9	0	0
380	EvolSTAR	1	12832	10	12803	10	0	0
	PBC	1	12832	10	12466	10	0	6
	EvolSTAR	2	12832	10	12816	10	0	0
	PBC	2	12832	10	12542	10	0	4
384	EvolSTAR	1	6002	6	5992	6	0	0
	PBC	1	6002	6	5888	6	0	0
	EvolSTAR	2	6002	6	5993	6	0	0
	PBC	2	6002	6	5895	6	0	1
507	EvolSTAR	1	3921	8	3913	8	0	0
	PBC	1	3921	8	3736	8	0	3
	EvolSTAR	2	3921	8	3916	8	0	0
	PBC	2	3921	8	3758	8	0	2
581	EvolSTAR	1	8574	10	8567	10	0	0
	PBC	1	8574	10	8458	10	0	2
	EvolSTAR	2	8574	10	8566	10	0	0
	PBC	2	8574	10	8461	10	0	5
582	EvolSTAR	1	3057	4	3051	4	0	0
	PBC	1	3057	4	2986	4	0	0
	EvolSTAR	2	3057	4	3053	4	0	0
	PBC	2	3057	4	3001	4	0	0
593	EvolSTAR	1	3054	3	3053	3	0	0
	PBC	1	3054	3	3007	2	1	0

	EvolSTAR	2	3054	3	3053	3	0	0
	PBC	2	3054	3	2992	2	1	0
9061364	EvolSTAR	1	5129	5	5123	5	0	0
	PBC	1	5129	5	5064	5	0	0
	EvolSTAR	2	5129	5	5122	5	0	0
	PBC	2	5129	5	5042	5	0	0
9061365	EvolSTAR	1	3000	3	2993	3	0	0
	PBC	1	3000	3	2956	3	0	1
	EvolSTAR	2	3000	3	2991	3	0	0
	PBC	2	3000	3	2941	3	0	0
9061366	EvolSTAR	1	1683	3	1683	3	0	0
	PBC	1	1683	3	1649	3	0	1
	EvolSTAR	2	1683	3	1682	3	0	1
	PBC	2	1683	3	1636	3	0	1
923	EvolSTAR	1	4373	5	4365	5	0	0
	PBC	1	4373	5	4187	5	0	1
	EvolSTAR	2	4373	5	4330	5	0	1
	PBC	2	4373	5	3738	5	0	6

Table 4: Comparison of Calls made by EvolSTAR and PBC for 14 samples. Types of calls and their frequencies generated by EvolSTAR and PBC in replicated microarray hybridizations of 14 patient samples. Partial or complete capillary sequences were generated for each sample and used to verify the calls made by EvolSTAR and PBC on each replicate. We then count the frequency of true-non-mutation, true-mutation, error and 'N' calls in each replicate.

We wondered if, and by how much, incorporating NHIP and substitution biases analysis to the PBC results would improve the performance of the PBC algorithm. We observed that more than 70% of the 65 error calls (false mutation calls) made by PBC did not have the characteristic NHIP of a true-mutation shown in Figure 20(b). The remaining 30% of the error calls had a NHIP reminiscent of a true-mutation NHIP but did not satisfy the substitution bias rule. Using NHIP and substitution biases analysis together, we were able to reduce the number of false mutation calls to only two. Most of the 4434 'N' calls made by PBC were due to conflicting base calls from the forward and reverse strand. By analyzing the NHIP and hybridization intensity reduction order of the query base in the forward and reverse strand

individually, we were able to identify the noisy strand and hence, make the base call only from the non-noisy strand. We were able to recover 92% of the 'N' calls made by PBC using this approach.

PM probe encoding	Hybridization intensity reduction order	Forward strand	Reverse strand
		Frequency	Frequency
A	CGT	2618	1030
	CTG	2347	975
	GCT	4848	1870
	GTC	12571	8889
	TCG	4417	2624
	TGC	16805	16692
C	AGT	10843	14309
	ATG	10606	14473
	GAT	1777	1567
	GTA	748	618
	TAG	2006	1784
	TGA	790	623
G	ACT	9114	7403
	ATC	5490	3647
	CAT	9369	8811
	CTA	4104	3143
	TAC	2839	1976
	TCA	2458	1790
T	ACG	1926	2080
	AGC	2489	2524
	CAG	3211	3721
	CGA	6191	8656
	GAC	7550	9533
	GCA	10713	17092

Table 5: Hybridization intensity reduction orders found in 14 hybridization experiments. Hybridization intensity reduction orders found in 135830 true calls from 14 hybridization experiments. For each true call, for each strand, we rank the PM probe and its MM probes based on their hybridization intensities in decreasing order. We count the frequency of each hybridization intensity reduction order.

Sample	Program	Rep.	Total Sites Verified by Capillary	Mutations (Verified by Capillary)	True-Non-Mutation Calls	True Mutation Calls	Missed Mutations	Error Calls
305_Nasal	EvolSTAR	1	6676	9	6601	9	0	0
	PBC	1	6676	9	6066	9	0	32
	EvolSTAR	2	6676	9	6569	9	0	1
	PBC	2	6676	9	5946	9	0	45
305_cell_cond1	EvolSTAR	1	6676	9	6667	9	0	0
	PBC	1	6676	9	6515	9	0	3
	EvolSTAR	2	6676	9	6659	9	0	0
	PBC	2	6676	9	6427	9	0	7
305_cell_cond2	EvolSTAR	1	6676	9	6652	9	0	0
	PBC	1	6676	9	6495	9	0	6
	EvolSTAR	2	6676	9	6656	9	0	0
	PBC	2	6676	9	6474	9	0	7
305_cell_cond3	EvolSTAR	1	6676	9	6664	9	0	0
	PBC	1	6676	9	6551	9	0	4
	EvolSTAR	2	6676	9	6663	9	0	0
	PBC	2	6676	9	6503	9	0	4
305_cell_cond4	EvolSTAR	1	6676	9	6658	9	0	0
	PBC	1	6676	9	6531	9	0	6
	EvolSTAR	2	6676	9	6664	9	0	0
	PBC	2	6676	9	6529	9	0	4
305_cell_cond5	EvolSTAR	1	6676	9	6660	9	0	0
	PBC	1	6676	9	6571	9	0	3
	EvolSTAR	2	6676	9	6660	9	0	0
	PBC	2	6676	9	6523	9	0	5

Table 6: Comparison of Calls made by EvolSTAR and PBC for 6 pairs of isolates belonging to patient sample 305. Types of calls and their frequencies generated by EvolSTAR and PBC in replicated microarray hybridizations of sample 305. Partial or complete capillary sequences were generated for each sample and used to verify the calls made by EvolSTAR and PBC on each replicate. We then count the frequency of true-non-mutation, true-mutation, error and 'N' calls in each replicate.

In addition, we evaluate the robustness and repeatability of EvolSTAR by employing six pairs of replicate experiments consisting of one pair nasal swab and five pairs of cell culture isolates, belonging to the same patient sample 305 (Table 6). Of the experiments, two pairs of replicates (305_nasal and 305_cell_cond1) were amplified under the same optimal experimental conditions while each of the other pairs (305_cell_cond2, 305_cell_cond3, 305_cell_cond4,

305_cell_cond5) were amplified under different sub-optimal experimental conditions (simulating experimental volatility). Compared with the available capillary sequences for sample 305, EvolSTAR had an average error rate of 0.0012% and 28 ambiguous calls per sample (338 in total). On the other hand, Nimblescan PBC obtained a relatively higher average error rate of 0.169% and 237 ambiguous calls per sample (2855 in total). Our results showed that EvolSTAR is robust and performs well when samples are prepared under sub-optimal conditions. Even for nasal swab samples that tend to have much less concentration of virus RNA than cell cultures, EvolSTAR suffered only a slight drop in performance compared to Nimblescan PBC.

In conclusion, we have shown that EvolSTAR is robust and generates sequence calls of high accuracy and reproducibility in this pilot study consisting of 40 microarray experiments. Meanwhile, efforts will be put in to continually evaluate EvolSTAR with more samples and update it on a regular basis as the H1N1 (2009) influenza virus evolves.

4.5.1 Visualization of Sequence Calls

Besides a FASTA output of the virus sequence, EvolSTAR generates a visualization map of the sequence calls using a heat map based on the percentage identity of the called sequence to the reference sequence measured at 50 bp windows (Figure 22). The map template consists of all eight segments of the 2009 influenza A(H1N1) virus and the locations of known drug binding sites (marked with green lines) on the neuraminidase gene. Locations of all mutation calls are denoted by red triangles beneath the heat map bar. Sequences that are of low coverage (< 90%) are automatically flagged, and the overall PM/MM discrimination ratio for each segment is displayed. The heat map bar allows the technician to rapidly assess the quality of the sequence

data obtained from the microarray and identify regions where PCR did not work well, or presence of potential recombination/reassortment events. Mutations, especially those in close proximity to drug binding sites, can be quickly visualized. Other details such as coverage, number of base calls successfully made, number of mutations and number of 'N' calls for each sequence call are also shown on the visualization map.

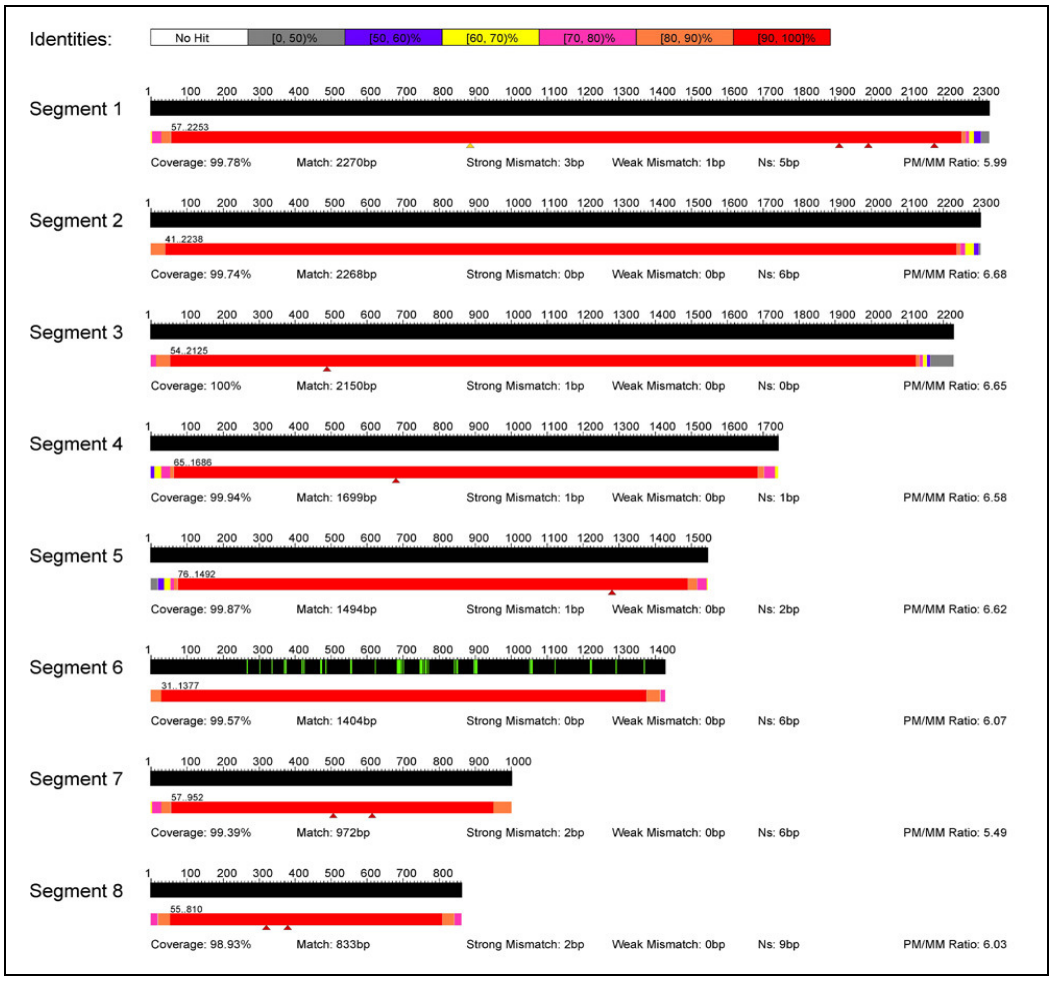

Figure 22: Visualization map of all eight segments of the 2009 influenza A(H1N1) virus and the locations of known drug binding sites (marked with green lines) on the neuraminidase (NA) gene (Segment 6). A heat map bar is used to represent the quality and coverage of its sequence calls. The locations of all mutation calls made by EvolSTAR are represented by red triangles beneath the heat map

bar. Sequences with coverage < 90% are automatically flagged as “Low Coverage”. Other details such as Coverage: percentage of base calls successfully made, Match: number of base calls that match the reference sequence ie non-mutation base calls, Strong Mismatch: number of high confidence base calls that do not match the reference sequence ie mutation base calls, Weak Mismatch: number of low confidence base calls that do not match the reference sequence ie mutation base calls and Ns: number of ‘N’ calls, for each sequence call are also shown on the visualization map.

4.6 Discussion and Conclusion

Traditional statistical and probabilistic sequence-calling techniques ascertain that a base call is of high confidence if they exceed pre-defined significance or probability thresholds. This approach works well for high confidence base calls but is inadequate to extract sufficient information from noisy base calls. It is also difficult to determine the validity of a mutation call purely based on the distribution of hybridization intensities of its PM and MM probes. In this work, we have described two new hybridization intensity analysis methods that enable us to confidently identify true mutations and recover some noisy base calls. Compared to PBC, EvolSTAR has achieved superior call rates and accuracies, especially in low concentration samples with high CT values. The robustness of the base calls enables our approach to be a practical large-scale evolutionary surveillance tool.

Although we are confident that our resequencing array can successfully generate complete sequences for the H1N1(2009) virus and its variants at the current stage, we cannot rule out the possibility of reassortments between the H1N1(2009) virus and other influenza viruses. Clearly, our resequencing array cannot fully sequence such events and will generate sequences with poor quality and coverage of the reassorted segments. To investigate the effects of a reassortment event on our array, we independently amplified segments 1, 2, 3, 5, 6 and 7 of the 2009 influenza A(H1N1) virus and segment 4 of a H3N2 influenza A virus, and hybridized them

onto our array. The visualization map of this experiment is shown in Figure 23. As expected, the sequence call for segment 4 (based on PM/MM probes from the segment 4 consensus of the 2009 influenza A(H1N1) virus) is poor in quality and coverage. However, we observed that we were able to get good base calls from region 1150-1547. This region turns out to be the only significantly similar (70% matched) region between the segment 4 consensus of the 2009 influenza A(H1N1) virus and segment 4 of a H3N2 virus (CY039087). This shows that identifying regions of high similarity between the 2009 influenza A(H1N1) virus with other influenza viruses and checking if these regions have good sequence calls may be a plausible way of detecting reassortments. The drawback of this approach is that it will fail to detect reassortment of certain segments where there are no regions of high similarity between the H1N1(2009) virus and the parental influenza virus. It is also difficult to annotate and differentiate every region that the H1N1(2009) virus and all other influenza viruses share similarity with. We propose an alternative approach to detect reassortments. By analyzing the PM/MM hybridization intensity fold-change of high confidence calls of all 8 segments, we found that the average PM/MM hybridization intensity foldchange of high confidence calls in segments 1, 2, 3, 5, 6, and 7 belonging to the 2009 influenza A(H1N1) virus is approximately 4.5 while the average PM/MM hybridization intensity fold-change of high confidence calls in segment 4 belonging to the H3N2 influenza A virus is only 1.9. The most likely reason for this huge drop in the average PM/MM hybridization intensity fold-change of high confidence calls is that the signal gained by most of the segment 4 PM probes on our array are through cross-hybridization to the segment 4 sequence of the H3N2 influenza A virus, and thus much lower than signal gained from true specific binding. Thus, by computing and comparing the average PM/MM hybridization intensity fold-change of high confidence calls in each segment, we can identify

potential reassortments in a given H1N1(2009) virus sample. Virus samples with possible reassortments can then be sequenced using capillary sequencing or customized reassortment resequencing arrays.

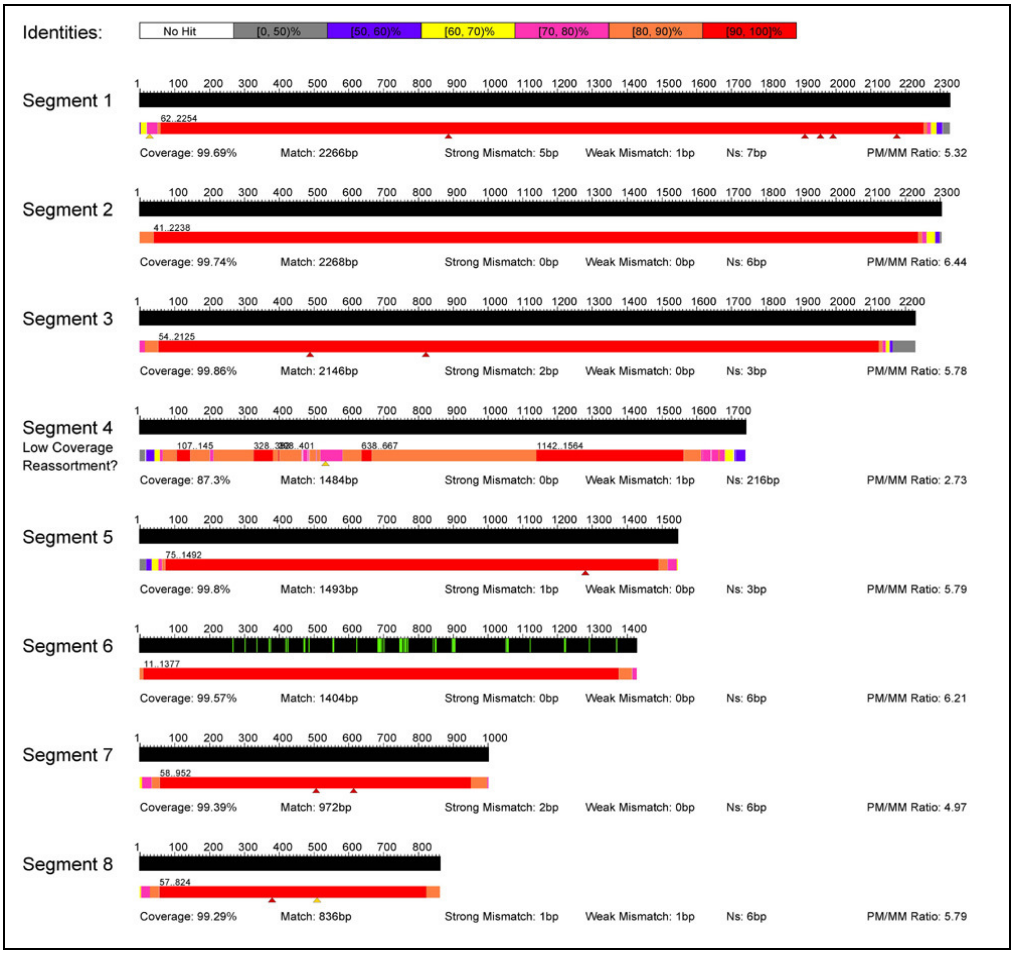

Figure 23: Visualization map of a 2009 influenza A(H1N1) virus with artificial reassortment of H3N2 segment 4. We independently amplified segments 1, 2, 3, 5, 6 and 7 of the 2009 influenza A(H1N1) virus and segment 4 of a H3N2 influenza A virus, and hybridized them onto our array. As expected, the sequence call for segment 4 (based on PM/MM probes from the segment 4 consensus of the 2009 influenza A(H1N1) virus) is poor in quality and coverage.

So far, the sequence diversity of H1N1 2009 influenza virus has been rather limited. From our analysis, it would be possible to resequence all the published isolates using this resequencing approach. However, as antigenic drift is expected to occur, it is likely that the resequencing array would need to be updated at least annually. Updating the array requires only bioinformatics input, and does not require any other additional manufacturing costs. Thus, this combination of sample amplification primers, low-cost multiplex array and robust interpretation software allows sustainable, rapid, large-scale biosurveillance of the influenza H1N1(2009) virus.

Chapter 5

RECOMBINATION DETECTION IN VIRAL GENOMES

5.1 Sources of Inaccuracy of Distance-based Window Methods

Accurate phylogenetic analysis gives us important clues on the origins, pathogenicity and treatments of viruses and bacteria. The first crucial step in carrying out an accurate phylogenetic analysis is the identification of recombination breakpoints. A breakpoint is defined as the location where a recombination event occurs in a sequence.

A hugely popular technique is to slide a window along a sequence alignment and look for differences in distance metrics or phylogenetic trees within each window. More specifically, a length-w "sliding window" is defined as a window enclosing the alignments from positions i to $i + w$. Suppose b is a breakpoint, we would expect the immediate left and right neighboring windows $[b - w \ldots b - 1]$, $[b \ldots b + w]$ of b to enclose alignments that are significantly different. Hence, to find the exact positions of the breakpoints (if any), the sliding window approach performs an exhaustive search by sliding a length-w window across the alignment and for every possible position i ($w \leq i \leq N - w$), it compares the alignments enclosed by neighboring windows $[i - w \ldots i - 1]$ and $[i \ldots i + w]$. There are two main ways to compare alignments, namely by distance measures (distance-based) or phylogenetic trees' topologies (phylogeny-based). Although phylogeny-based measures are widely accepted as more accurate than distance-based measures, distance-based measures are often much faster and more scalable for large datasets. However, for distance-based window methods to be as reliable as phylogeny-based methods, their accuracy issues must be addressed.

There are two sources of inaccuracy when using distance-based window methods to detect recombination. Firstly, the use of conventional distance metrics, such as hamming-distance and edit-distance that measure overall homology results in phylogenetic information loss. Secondly, recombination detection is too sensitive on the choice of window length. Recent works have shown that their results are most accurate when the given window length is approximately the recombinant subsequence length [62, 64]. If the length of the recombinant is not known in advance, an algorithm using different window lengths may produce vastly different analysis results on the same dataset. Furthermore, there may be problems in detecting recombinant regions shorter than the given window length due to the noise caused by the original sequence on either side of the recombinant subsequence included in the window (Figure 24).

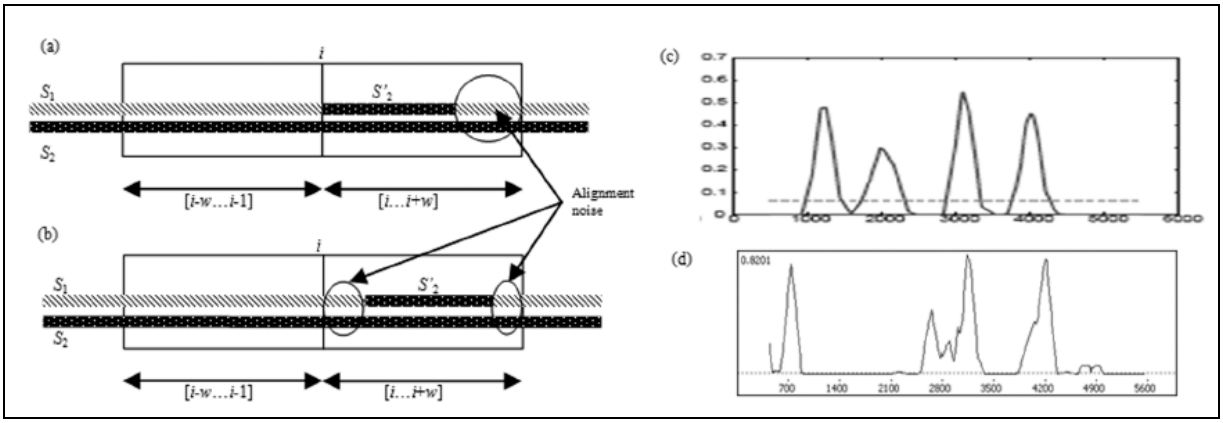

Figure 24: Window length sensitivity problem when window length w is longer than the recombinant subsequence (S'_2) length. (a) and (b) show the introduction of alignment noise into the computation of any distance measure or phylogenetic tree of the alignment in $[i…i+w]$. (c) shows the result of running Pruned-PDM on a synthetic dataset SD3 with breakpoints at site 1000, 2000, 3000 and 4000 using a length-500 window. (d) shows the result of running Pruned-PDM on the same dataset using a length-600 window. Here, the breakpoints are inaccurately detected at sites 700, 2600, 3200 and 4200.

To improve the information content of conventional distance measures and to reduce the impact of different window lengths on recombination detection, we employed three techniques:

1. Instead of using base-by-base comparison, the similarity of an alignment is computed using the number of shared (*k,m*)-mers, that is, length-*k* alignments with at most *m* mismatches. This measure takes into account the different mutation rates along the alignment by varying its mismatch threshold *m* automatically. This avoids the effect of random point mutations which causes inaccuracy in distance measures using base-by-base comparison.

2. Given a window instance, we use a weighting strategy that assigns heavier weights to positions nearer the putative breakpoint and lighter weights to positions further away from the putative breakpoint. This reduces the effect of alignment noise as seen in Figure 24(a)(b) when computing the similarity score.

3. Use contiguous chains of (*k,m*)-mers to form Contigs. Contigs have even distribution of mismatches and thus, are better estimations of long common subsequences in an alignment.

5.2 Using (k, m)-mers as the Basic Unit of Similarity Measurement

Conventional distance measures such as hamming-distance and edit-distance that perform base-by-base comparisons are susceptible to noise caused by random mutations. Furthermore, they compute only overall sequence homology and omit important details about the distribution of mismatches and the distribution of contiguous matches that may provide further indication of recombination. A possible solution is to use shared (*k,m*)-mers as the basic unit of similarity measurement in place of base-by-base comparisons.

Definition 1. Let A be a length-n alignment of two sequences S_1, S_2. Let $A[x\ldots\ x+k\text{-}1]$ be a length-k sub-alignment from position x to position $(x + k\text{-}1)$ of A. $A[x\ldots x+k\text{-}1]$ is a shared (k,m)-mer iff $A[x\ldots x+k\text{-}1]$ has less than m mismatches. This is shown in Figure 25.

Essentially, by counting the number of shared (k,m)-mers, we can identify homologous regions of two sequences with an underlying rate of random mutation. Here, the selection of values for k and m is vital as k determines the specificity of homologous regions found and m estimates the underlying random mutation rate. Note that k should be small but at least of length $(\log_2 n + 1)$ to ensure specificity. Clearly, the selection of a global value for m is not feasible because the underlying rate of random mutations varies across the alignment. Hence, a localized value of m must be chosen for each sub-alignment enclosed by a window instance. We describe a heuristic to automatically determine m for a window instance: Given a length-w window $[i\ldots i+w]$ with parameters k and m, let K_{m} and K'_{m} be the number of shared (k,m)-*mer*s and the number of non-shared (k,m)-mers respectively (note that $K_m + K'_m = w - k + 1$). Starting with $m = 0$, we iteratively increase m until $K_m > K'_m$. Denote $m_i = \min\{m \mid K_m > K'_m\}$. At this point of time, a majority of (k,m)-mers in the window are shared (k,m)-mers having less than m_i random mutations. This becomes a reasonable estimate of the underlying random mutation rate of the alignment enclosed by $[i\ldots i+w]$. Hence, we use m_i as the mismatch threshold in $[i\ldots i+w]$.

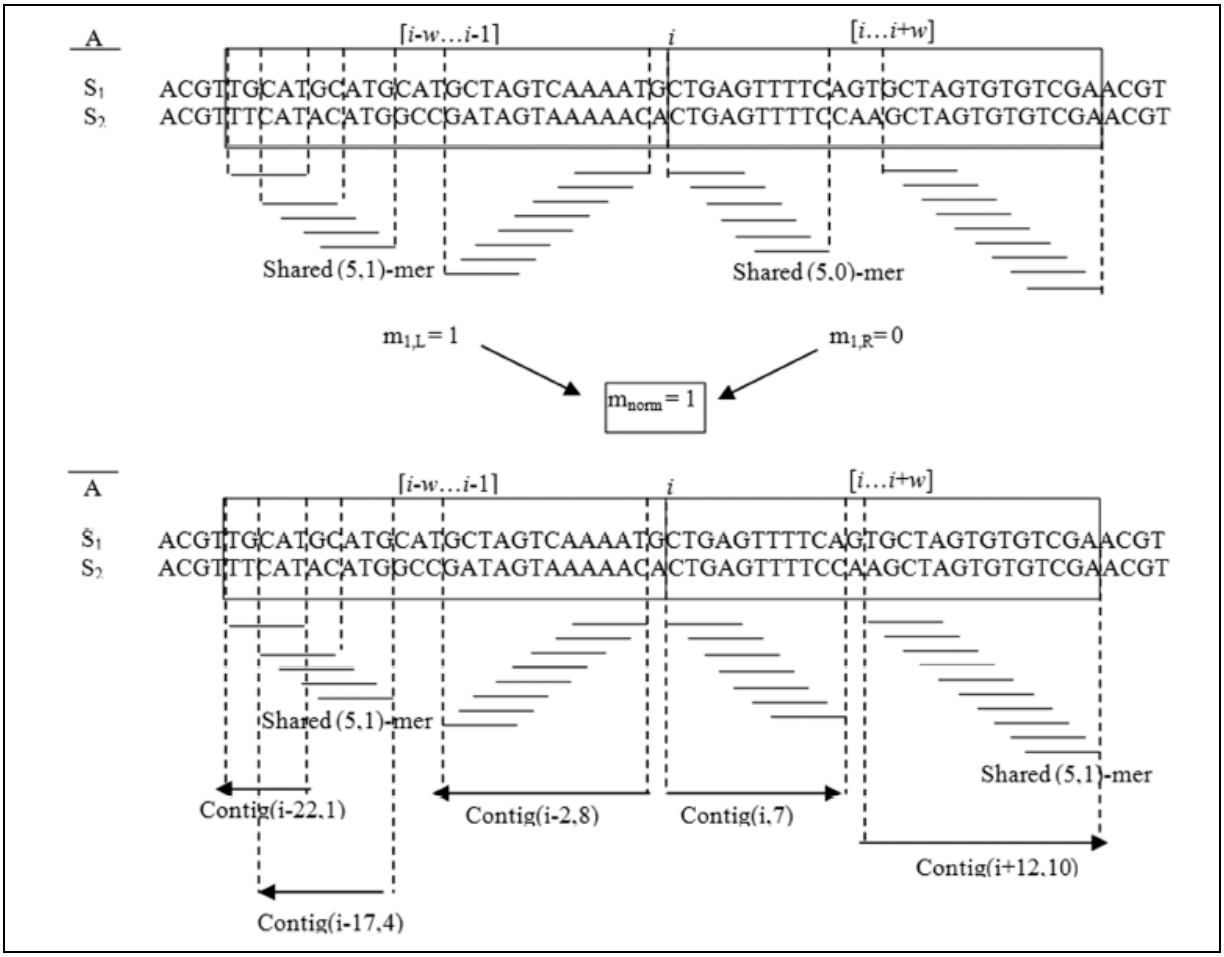

Figure 25: A diagram showing how shared (k, m)-mers are defined in each window of a putative breakpoint i of an alignment A of two sequences S_1, S_2. Top: Shared (k, m)-mers for [i-w…i-1] and [i…i+w] using mismatch thresholds $m_{i,L} = 1$ and $m_{i,R} = 0$ respectively. Bottom: After normalization ($m_i = 1$), the number of shared (5,1)-mers in each window reflects more accurately the relative sequence homology. Hence, $kmd_{i,L} = 13$ and $kmd_{i,R} = 17$. Note that consecutive shared (k, m)-mers form a Contig which we will elaborate in Section 5.1.2.

To detect recombination in an alignment of two sequences S_1 and S_2, we compute and compare a similarity score based on the shared (k,m)-mers of the two neighboring windows [i-w…i-1] and [i…i+w] of a putative breakpoint i. Let $m_{i,L}$ and $m_{i,R}$ be the mismatch thresholds of [i-w…i-1] and [i…i+w] respectively. Clearly, we cannot compare any similarity score of [i-w…i-1] and [i…i+w] when $m_{i,L} \neq m_{i,R}$. Thus, we need to normalize $m_{i,L}$ and $m_{i,R}$ so that shared (k,m)-

mers in [i-w…i-1] and [i…i+w] are defined based on a single mismatch threshold under the assumption of a common random mutation rate. Note that if i is a breakpoint or has high mutation rates, then the sequence homologies in the alignment enclosed by one window would be quite different from that enclosed in the other window (ie $m_{i,L} \neq m_{i,R}$). We exploit this observation and use the normalized mismatch threshold at i, $m_i = \max(m_{i,L}, m_{i,R})$ to define shared (k,m)-mers in both [i-w…i-1] and [i…i+w] (Figure 25). In this way, the window with the lower mismatch threshold will have many more shared (k,m)-mers than the other window with the higher mismatch threshold. Thus, any irregularities at i such as recombination and high mutation rates would be shown as a huge discrepancy in the number of shared (k,m)-mers in [i-w…i-1] and [i…i+w]. Formally, we denote the number of shared (k,m)-mers between S_1 and S_2 in each neighboring window of a putative breakpoint i as the km-distance (*kmd*):

For the left window [i-w…i-1], $kmd_{i,L}(S_1, S_2) = |P_{i,L}|$

For the right window [i…i+w], $kmd_{i,R}(S_1, S_2) = |P_{i,R}|$

where $P_{i,L} = \{j \mid i\text{-}w \leq j \leq i\text{-1-k}$ and $A[j\ldots j+k\text{-}1]$ is a shared (k, m)-mer$\}$ and $P_{i,L} = \{j \mid i \leq j \leq i+w\text{-}k$ and $A[j\ldots j+k\text{-}1]$ is a shared (k, m)-mer$\}$. Recombination is then inferred by some metric computed based on the magnitude of difference between $kmd_{i,L}$ and $kmd_{i,R}$. This is further elaborated in Section 5.5.

5.3 Contigs as a Better Estimation of Long Common Subsequences

The previous section championed the use of shared (k,m)-mers over base-by-base comparisons when measuring homology between two sequences. However, they are too short to truly represent the degree of homology between two sequences. A better indication of homology

between two sequences would be the number and length distributions of long common subsequences. Now, the question is how do we obtain long common subsequences of two sequences enclosed by a window with only short shared (*k,m*)-mers? Note that a length-*L* common subsequence of S_1 and S_2 is a tiling of $(L - k + 1)$ consecutive shared (*k,m*)-mers when $k \leq L$. In this paper, we define a common subsequence of S_1, S_2 as a chain of consecutive shared (*k,m*)-mers, which is known as a Contig.

Definition 2. A Contig is a length-*L* common subsequence of two sequences S_1 and S_2 formed by a chain of consecutive shared (*k,m*)-mers shared by S_1 and S_2. It has two parameters, namely the starting position *p* and the member size *s*. Here, *p* refers to the position of the Contig nearest to a putative breakpoint *i* and $s = L - k + 1$, that is, the number of consecutive shared (*k,m*)-mers chained to form the Contig. Thus a Contig can be written as Contig(*p*, *s*). (See Figure 25)

It is easy to see that any length-*L*' sub-Contig of a length-*L* Contig where $k \leq L' \leq L$ is guaranteed to have less than (*m*/*k* * *L*') mismatches. In addition, Contigs have an even distribution of mismatches. On the contrary, long common subsequences may have concentrations of mismatches in localized regions, despite passing the overall mismatch threshold. This creates a dilemma of whether to split a long common subsequence into shorter ones at regions where there are many mismatches. Our definition of Contigs avoids this problem and thus is more reflective of the localized similarity between two sequences.

This leads to the assessment of a position *i* being a true breakpoint by two criteria:

1. If *i* is a breakpoint, the Contig length distributions in [*i*-*w*…*i*-1] and [*i*…*i*+*w*] will be significantly different due to the distinct difference in sequence homology of the

alignment enclosed by each window. Hence, we can assess if position i is a breakpoint by perform a Kolmogorov-Smirnov test [118] on the Contig length distributions in [i-w…i-1] and [i…i+w] at 99% confidence interval.

2. There should be a significant difference between similarity scores computed for the alignment in [i-w…i-1] and [i…i+w].

Note that in the previous section, the km-distances, $kmd_{i,L}$ and $kmd_{i,R}$, were used as the similarity scores for the alignment enclosed by [i-w…i-1] and [i…i+w]. Next section describes a weighting strategy to improve the km-distance to incorporate the concept of Contigs. Similarly, we elaborate on the metric to detect recombination in Section 5.5.

5.4 Breakpoint Specific Positional Weighted Distance Measure

Depicted in Figure 24(a)(b), the alignment noise affect the detection of breakpoint. We solve the issue by assigning weights to all positions enclosed by [i-w…i-1] and [i…i+w] with respect to a putative breakpoint i. More specifically, we assign heavier weights to positions in [i-w…i-1] and [i…i+w] near to the putative breakpoint i while lighter weights to positions in [i-w…i-1] and [i…i+w] that are further away from the putative breakpoint. We justify our proposed weighting strategy to solve the window length sensitivity problem based on Figure 24(a)(b): In Figure 24(a) where i is a true breakpoint, positions that are furthest away from i are most likely to contribute to alignment noise if the window length is too large. Thus, by assigning these positions the lightest weights when computing the km-distance for each window, alignment noise is reduced. In Figure 24(b) where i is not a true breakpoint, the alignment near i in [i-w…i-

1] and [i…i+w] will not experience a sudden significant change. Since positions near i are assigned heavier weights, the km-distance in [i-w…i-1] and [i…i+w] will most likely not have a sudden significant change too.

Clearly, a good implementation of our weighting strategy requires a suitable function or a suitable family of functions that assigns a weight to a position in neighboring windows [i-w…i-1], [i…i+w] based on its absolute distance from a putative breakpoint i. Let S_1 and S_2 be two aligned length-n sequences with a putative breakpoint at position i with neighboring windows [i-w…i-1], [i…i+w]. Let x be the relative distance of a position j in [i-w…i-1], [i…i+w] from the breakpoint i, that is $x = j$-i, and $-w \leq x \leq w$. Next, we define a positive weight function $F_i : X \rightarrow \Re^+$ where $X = \{x \mid -w \leq x \leq w\}$. $F_i(x)$ satisfies the properties that (i) $F_i(x) = F_i(-x)$ and (ii) $F_i(x)$ is decreasing when $|x|$ increases. For simplicity, we set $F_i(0) = 1$ and $F_i(w) = F_i(-w) = 0$. A family of functions satisfies the above properties is as follows:

$$F_i(x) = \frac{w^k - |x|^k}{w^k}$$

Note that k controls the decreasing rate of $F_i(x)$. In our application, we need moderate decreasing rate and hence we set k=2 by default. In this case, $F_i(x)$ is in fact a reverse parabola shape.

Next, we describe how this weighting strategy incorporates the notion of Contigs in the km-distance. The idea is to assign a weight to each shared (k, m)-mer based on which Contig it belongs to. In this way, shared (k, m)-mers belonging to the more informative Contigs (closer to the putative breakpoint) are assigned heavier weights than those belonging to Contigs that are more prone to alignment noise (further away from the putative breakpoint). More specifically, our weighting strategy assigns each Contig(p, s) and its member shared (k, m)-mers a weight

$F_i(|i\text{-}p|)$. Since Contigs do not overlap in a window, the weight assigned to each Contig and its member shared (k, m)-mers is unique. Consequently, given Contig(p_1, s_1) and Contig(p_2, s_2) with assigned weights $F_i(|i\text{-}p_1|)$ and $F_i(|i\text{-}p_2|)$ respectively, $F_i(|i\text{-}p_1|) > F_i(|i\text{-}p_2|)$ iff $|i\text{-}p_1| < |i\text{-}p_2|$.

Thus, given a putative breakpoint i and the two neighboring windows [i-w…i-1] and [i…i+w], we compute the improved Breakpoint specific positional Weighted Contig-Alignment (*BWCA*) score for the alignment of S_1 and S_2 in each neighboring window of a putative breakpoint i:

For window [i-w…i-1], $$BWCA_{i,L}(S_1, S_2) = \sum_{p_j \in P_{i,L}; s_j \in S_{i,L}} F_i(i - p_j) \times s_j$$

For window [i…+-w], $$BWCA_{i,R}(S_1, S_2) = \sum_{p_j \in P_{i,R}; s_j \in S_{i,R}} F_i(p_j - i) \times s_j$$

where $C_{i,L} = \{\text{Contig}(p_j, s_j) \mid i\text{-}w \leq p_j \leq i - 1\}$ and $C_{i,R} = \{\text{Contig}(p_j, s_j) \mid i \leq p_j \leq i + w\}$ are the sets of Contigs in [i-w…i-1] and [i…i+w] respectively; $S_{i,L} = \{s_j \mid \text{Contig}(p_j, s_j) \in C_{i,L}\}$ and $S_{i,R} = \{s_j \mid \text{Contig}(p_j, s_j) \in C_{i,R}\}$ are the corresponding set of member sizes of the Contigs in $C_{i,L}$ and $C_{i,R}$; $P_{i,L} = \{p_j \mid \text{Contig}(p_j, s_j) \in C_{i,L}\}$ and $P_{i,R} = \{p_j \mid \text{Contig}(p_j, s_j) \in C_{i,R}\}$ are the corresponding set of starting positions of the Contigs in $C_{i,L}$ and $C_{i,R}$.

If the difference between $BWCA_{i,L}$ and $BWCA_{i,R}$ is big, position i is expected to be a recombination breakpoint. Section 5.5 elaborates how recombination is inferred based on this difference.

5.5 The RB-Finder Algorithm to Detect Recombination

We have presented three techniques to address the main criticisms (accuracy and efficiency) of distance-based window methods to detect recombination among two aligned sequences. Next, we empirically investigate the effectiveness of our *km*-distance and *BWCA* scores to detect recombination as opposed to using a conventional distance measure such as the Kronecker delta function. Our simulations show that in real datasets where recombination events are more complex and harder to detect, our highly sensitive *BWCA* score stands a better chance of detecting the breakpoints than other distance measures.

We make use of the *BWCA* score and propose the RB-Finder algorithm to detect recombination in a multiple sequence alignment. Given a length-n alignment of M sequences, the idea is to move a length-w sliding window along the alignment and, for each position i, computes a Recombination Detection Score (RDS_i) based on the highly sensitive *BWCA* score and two key observations to differentiate recombination and high mutation rates. At the real breakpoint i, two concurrent observations are prevalent: (1) there exists two sequences S_α, S_β in the alignment M such that the *BWCA* score increases significantly and suddenly across i, ie $BWCA_{i,L}(S_\alpha, S_\beta) - BWCA_{i,R}(S_\alpha, S_\beta) << 0$ and (2) there exists yet another sequence S_γ ($S_\gamma \neq S_\beta$) such that the *BWCA* score decreases significantly and suddenly across i, ie $BWCA_{i,L}(S_\alpha, S_\gamma) - BWCA_{i,R}(S_\alpha, S_\gamma) >> 0$. The first observation is made when there is a transfer of genetic sequence from S_β to S_α at i resulting in a sudden increase in homology between S_α and S_β. The second observation is that after the recombination event at i, S_α is no longer as homologous to some sequence S_γ as compared to prior the recombination event. From a phylogeny point of view, the two observations are in effect looking for the most divergent branch between the phylogenetic tree in [i-w…i-1] and the phylogenetic tree in [i…i+w]. This is shown in Figure 26.

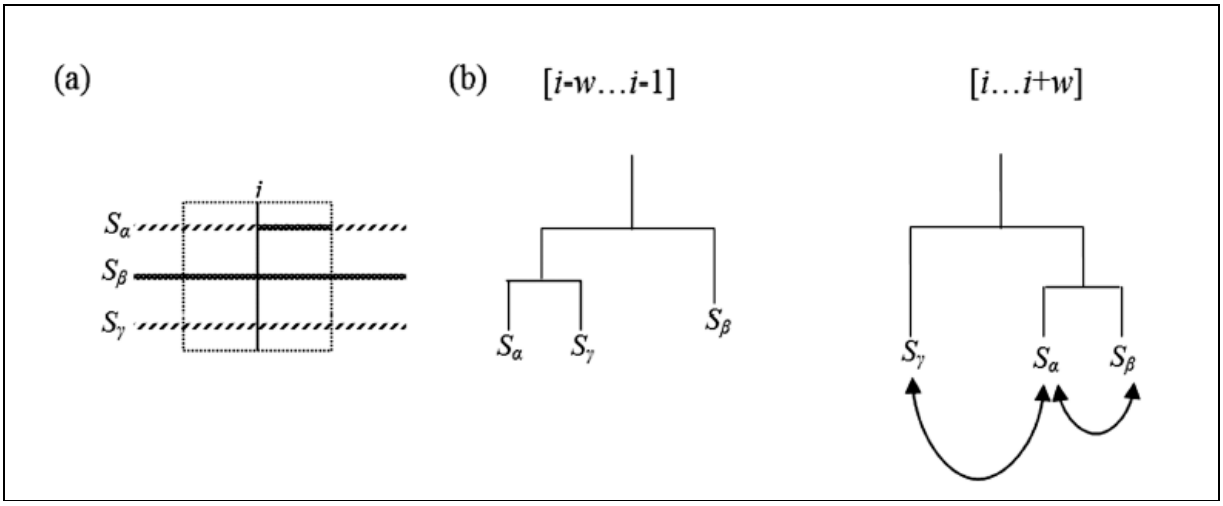

Figure 26: Search for most divergent branch in phylogenetic tree across adjacent windows. (a) shows that S_α is a recombinant of S_β at position i to (i+w). (b) shows the phylogenetic trees in [i-w…i-1] and [i…i+w] respectively. In [i…i+w], S_α is phylogenetically closer to S_β, than in [i-w…i-1], resulting in $BWCA_{i,L}(S_\alpha, S_\beta)$ - $BWCA_{i,R}(S_\alpha, S_\beta) << 0$. Concurrently, S_α is phylogenetically further away from S_γ, resulting in $BWCA_{i,L}(S_\alpha, S_\gamma)$ - $BWCA_{i,R}(S_\alpha, S_\gamma) >> 0$. S_α is the most divergent branch between the phylogenetic tree in [i-w…i-1] and [i…i+w] and thus detected to most likely contain recombination.

At every putative breakpoint i, we examine each of the M sequences for the two observations. Specifically, for each sequence $S_\alpha \in M$ at i, we find Contigs with each of the other sequences S_β ($S_\beta \in M$ and $S_\alpha \neq S_\beta$) in neighboring windows [i-w…i-1] and [i…i+w]. The KS-test is then used to filter sequences whose distributions of contig lengths with S_α in both windows are not significantly different at 99% confidence interval. We then compute the $BWCA$ scores of S_α with sequences that pass the KS-test in neighboring windows [i-w…i-1] and [i…i+w]. We obtain the most significant increase in $BWCA$ score across [i-w…i-1] and [i…i+w] for S_α, that is, $BWCA_{i,incr}(S_\alpha) = \max\{BWCA_{i,R}(S_\alpha, S_\beta) - BWCA_{i,L}(S_\alpha, S_\beta) \mid S_\beta \in M \text{ and } S_\alpha \neq S_\beta\}$. Similarly, we also obtain the most significant decrease in $BWCA$ score across [i-w…i-1] and [i…i+w] for S_α, $BWCA_{i,decr}(S_\alpha) = \max\{BWCA_{i,L}(S_\alpha, S_\gamma) - BWCA_{i,R}(S_\alpha, S_\gamma) \mid S_\gamma \in M \text{ and } S_\alpha \neq S_\gamma \neq S_\beta\}$.

Subsequently, we compute $RDS_i(S_\alpha)$, the recombination detection score for S_α accordingly as shown in Table 7:

Scenario	Observations Present	$RDS_i(S_\alpha)$ =	Reason
1	(1) and (2)	$BWCA_{i,incr}(S_\alpha) * BWCA_{i,decr}(S_\alpha)$	Recombination
2	(1)	0	Homologous regions not caused by recombination, most probably conserved regions
3	(2)	$- BWCA_{i,decr}(S_\alpha)$	High rates of mutation
4	None	0	No significant change in homology

Table 7: The formula to compute the *RDS* for a sequence based on the 2 observations.

Note that scenario 1 produces a distinctly high $RDS_i(S_\alpha)$ that indicates S_α has a recombination event at position *i*. On the other hand, scenario 3 produces a negative $RDS_i(S_\alpha)$ to clearly indicate high mutation rates. The other scenarios are deemed uninteresting in recombination detection and assigned $RDS_i(S_\alpha) = 0$.

Finally, we select the highest *RDS* among the *M* sequences to representing the *RDS* for breakpoint *i*:

$$RDS_i = \max(RDS_i(S_\alpha))$$

It is easy to see that if $RDS_i(S_\alpha) < 0$ for all $S_\alpha \in M$, then $RDS_i < 0$. This would indicate that the region around *i* suffers from high mutation rates. Conversely, a high RDS_i would mean that *i* is

most likely a true recombination breakpoint. We present the pseudo-code for the RB-Finder algorithm below in Figure 27.

```
RB-Finder{
  Given an length-N alignment of M sequences, length of
  (k,m)-mer k, and sliding window length w,

  for each position i from 1 to (N - w){
     Declare neighboring windows [i-w…i-1] and [i…i+w] ;
     for each sequence Sα in M {
       m = 0; // m to define a shared (k,m)-mer;
       Do {
          for each sequence Sα' in M {
            Obtain set of Contigs in [i-w…i-1], Pi,L
            Obtain set of Contigs in [i…i+w], Pi,R
            If(KS(Pi,L, Pi,R) <= 0.01) {
               Compute BWCAi,L (Sα , Sα');
               Compute BWCAi,R (Sα , Sα');
            }
          }
          m++;
          Compute BWCA'i,L of unshared(k,m)-mer;
          Compute BWCA'i,R of unshared(k,m)-mer;
       } while(BWCAi,L (Sα , Sα') > BWCA'i,L &&
               BWCAi,R (Sα , Sα') > BWCA'i,R )
       Compute BWCAi,incr(Sα);
       Compute BWCAi,decr(Sα);
       Compute RDSi (Sα );
     }
     Compute RDSi ;
   }
}
```

Figure 27: The pseudo-code for RB-Finder algorithm.

5.6 Evaluation of the RB-Finder Algorithm

Evaluation of our recombination detection algorithm is carried out by applying our algorithm to three synthetic and one biological datasets used in two previous papers [64, 65]. The three synthetic datasets (SD1, SD2 and SD3) each contains a 5500-bp alignment of eight sequences

(S_1, S_2, …., S_8) whose evolution was simulated with the Kimura model [119]. For SD1 and SD2, two recombination events were simulated: an ancient event affecting the region between sites 1000 and 1500, and a recent event affecting the region between sites 2500 and 3000. To test whether the detection method can successfully differentiate between recombination and rate variation, a mutational hotspot between sites 4000 and 4500 was introduced. The average branch length of the underlying phylogenetic trees for SD1 and SD2 are 0.1 and 0.01 respectively. The third synthetic dataset SD3 contains an ancient event affecting the region between sites 1000 and 2000, and a recent event between sites 3000 and 4000. The branch lengths of the tree were drawn from a uniform distribution on the interval [0.003, 0.005]. SD3 was deliberately created by Husmeier et al. [64] to thwart previous algorithms cited in their paper. The biological dataset used in our experiment is a length-3049 gap-removed ClustalW [120] alignment of 10 Hepatitis B virus sequences. It consists of two recombinant strains (D0329 and X68292) and eight non-recombinant strains (V00866, M57663, D00330, M54923, X01587, D00630, M32138 and L27106).

We ran our algorithm using (k,m)-mers with k=20 while m is automatically determined depending on the point mutation rate, and two window lengths 500 and 600. Note that the optimal window length to detect recombination in three of the four datasets (SD1, SD2 and Hep B) is 500 since all the recombination events that happened in these three datasets are of span 500 nucleotides. Thus, a window length of 600 would generate alignment noise and decrease accuracy of recombination detection in the three datasets. We shall see from the following results that the effects of alignment noise on recombination detection using our algorithm were minimal. We compared our results with that from Pruned-PDM since it has the highest accuracy. In addition, we also ran our datasets using Recco (default parameters with 1000 iterations), a newly

developed windowless method, to investigate the effectiveness of non-sliding-window methods in detecting recombination in our setting.

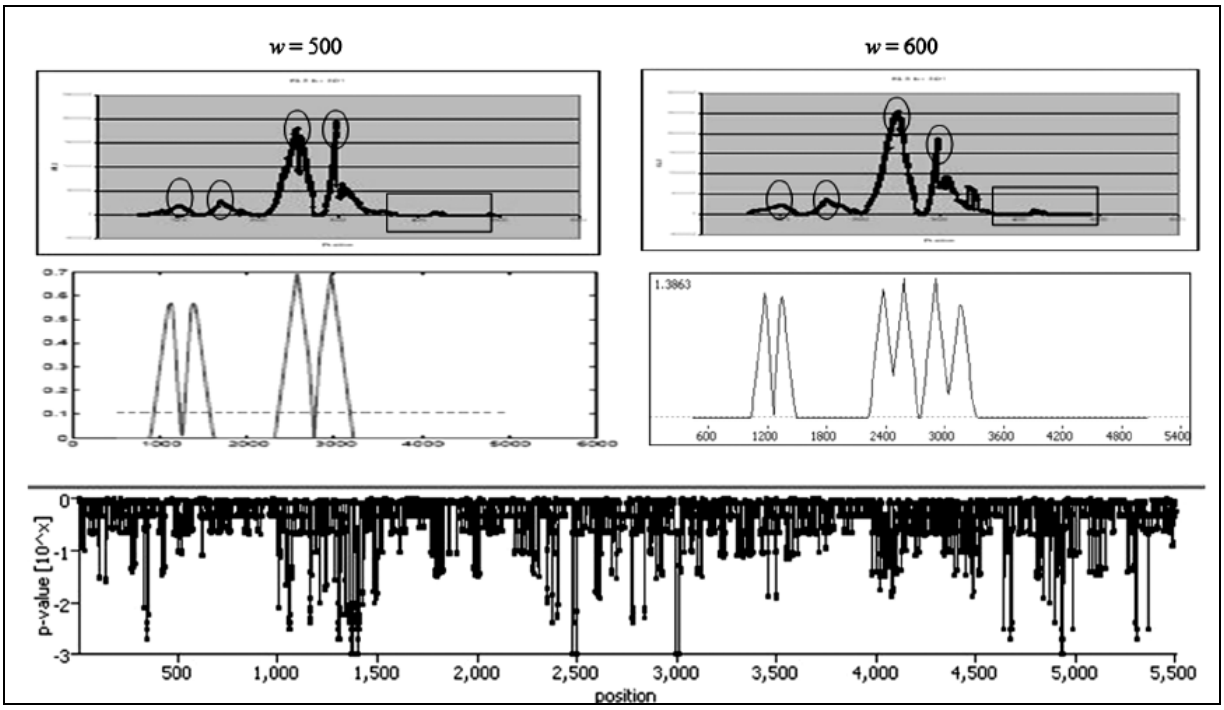

Figure 28: Recombination analysis results on SD1. Top: Results of our algorithm on SD1 with window lengths of 500 and 600. The circles highlight the recombination breakpoints at position 1000, 1500, 2500 and 3000 respectively. The rectangles highlight the high mutation regions which we detected. Middle: Results of Pruned-PDM on SD1 with window lengths of 500 and 600. When w= 600, recombination breakpoints at position 1000 and 1500 were detected at 1100 and 1400 instead. There are also multiple peaks around position 2500 and 3000. Bottom: Results of Recco showed that breakpoints at 2500 and 3000 were detected correctly. The breakpoint at 1500 was detected at position 1400. Breakpoint 1000 was not detected and the high mutation region around position 5000 was wrongly detected as a breakpoint.

For SD1 and SD2, when we ran our algorithm with window lengths 500 and 600, we detected a pair of weak recombination breakpoints at sites 1000 and 1500, and a pair of strong recombination breakpoints at sites 2500 and 3000. This is shown in Figure 28 and Figure 29 respectively. Note that a majority of the scores at sites 4000 to 5000 are less than 0 which is an

indication of high mutation rates. Thus, our algorithm also has a nice property of detection high mutation rates that may be mistaken for recombination. Pruned-PDM on the other hand produced inaccurate breakpoint positions with window size 600 for both datasets. We also ran Recco on the two datasets. For SD1, Recco could only detect the weak recombination breakpoint at position 1500 and the strong recombination breakpoints at position 2500 and 3000. Breakpoint at position 1000 was not detected and the mutation hotspot around position 5000 was wrongly detected as a breakpoint. For SD2, Recco could only detect the two strong recombination breakpoints. SD3 is a difficult dataset to analyze because there are only subtle differences in the alignment. Despite the very low rate of evolution in SD3, our algorithm detected recombination breakpoints at sites 1000, 2000, 3000 and 4000. In addition, our algorithm also detected a mutational hotspot around site 5000. The results for SD3 using window lengths 500 and 600 are shown in Figure 30. This time, Pruned-PDM not only produced inaccurate breakpoints, but also failed to detect the breakpoint at position 2000. Recco did not perform well on this dataset and could only detect breakpoints at positions 3000 and 4000.

For the Hepatitis B dataset, we detected three breakpoints around the sites 600, 1700 and 2200 for window lengths 500 and 600. In addition, we also detected an unreported recombination breakpoint around site 1000, which can be found by DSS and Pruned-PDM only when they use a window length of 300. The three breakpoints around the sites 600, 1700 and 2200 were also found by Recco. However, Recco also detected several other breakpoints which were previously unreported. This is shown in Figure 31.

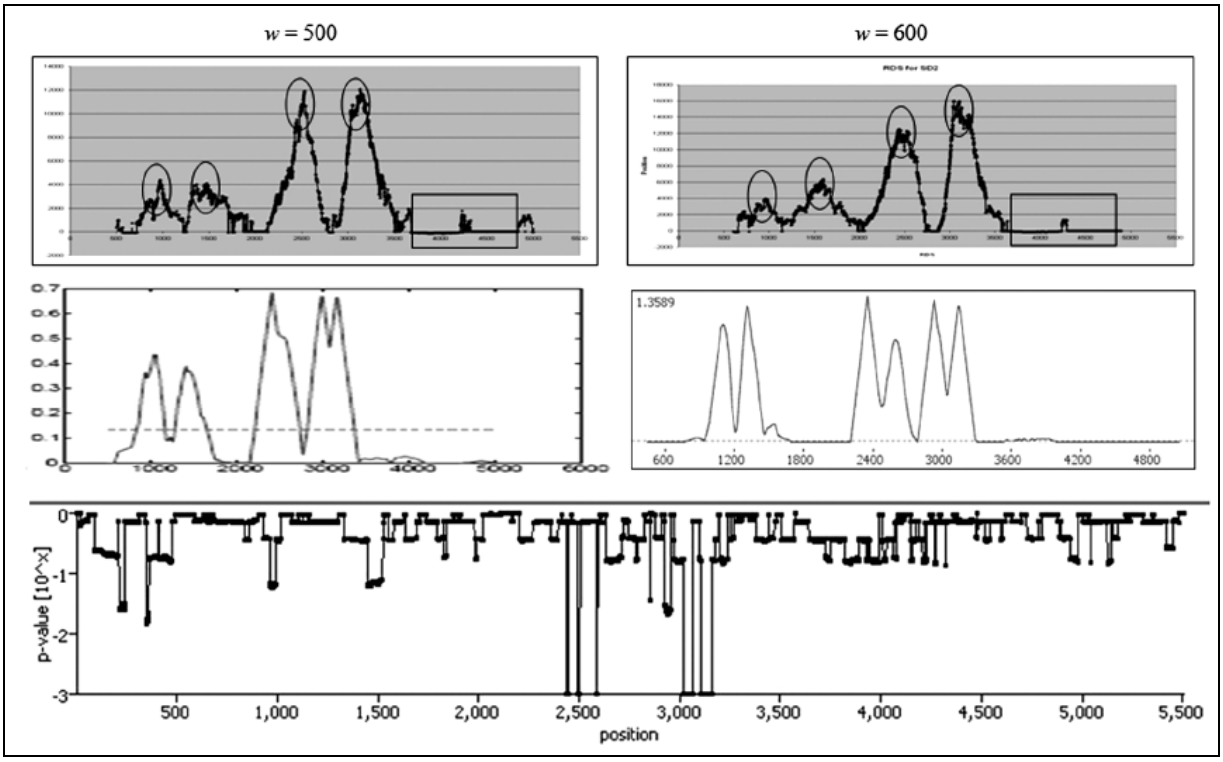

Figure 29: Recombination analysis results on SD2. Top: Results of our algorithm on SD2 with window lengths of 500 and 600. The circles highlight the recombination breakpoints at position 1000, 1500, 2500 and 3000 respectively. The rectangles highlight the high mutation regions which we detected. Middle: Results of Pruned-PDM on SD2 with window lengths of 500 and 600. When w = 600, the breakpoints are inaccurately detected at sites 1100, 1300, 2400 and 3000. Bottom: Results of Recco showed that except the strong recombination breakpoints at 2500 and 3000, the rest were not detected.

Our results using the four datasets are consistent with the DSS, Pruned-PDM and RECOMP methods. Through the use of our proposed weighting strategy and two key recombination-identifying observations, our algorithm is able to compute recombination breakpoints of a given alignment in a matter of minutes. Compared to Pruned-PDM which takes hours to analyze the same dataset, our algorithm is very much faster and achieves similarly accurate results. Unlike RECOMP which gives the user a choice of four graphs to decipher the recombination breakpoints, our algorithm generates only a single graph and thus prevents ambiguity of deciding which graph best represents the correct recombination breakpoints. An

example of this ambiguity is shown in their results for the SD3 dataset [65]. In their analysis, three of four graphs wrongly indicated that there is a recombination breakpoint at position 5000. Hence, it is difficult for the user to correctly infer that the recombination breakpoint detected at position 5000 by the three graphs is incorrect based on the remaining graph. In our analysis, we correctly identify the region around position 5000 as a mutational hotspot. This example also illustrates another advantage that our algorithm has over previous methods. Previous algorithms cannot differentiate normal regions from mutation hotspots. A useful feature of our algorithm is that it produces a $RDS < 0$ when the region has a high mutation rate. This immediately provides more biological information about the sequences to aid experimental studies.

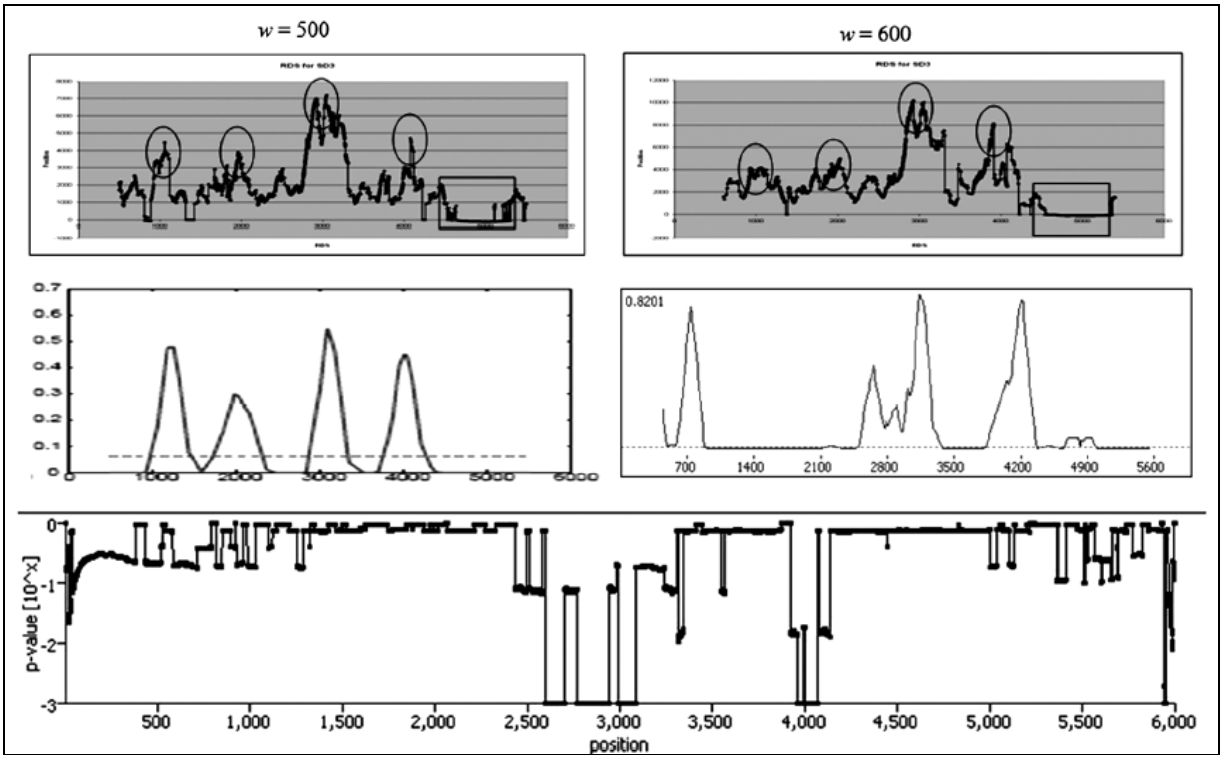

Figure 30: Recombination analysis results on SD3. Top: Results of our algorithm on SD3 with window lengths of 500 and 600. The circles highlight the recombination breakpoints at position 1000, 2000, 3000 and 4000 respectively. The rectangles highlight the high mutation regions which we detected. Middle: Results of Pruned-PDM on SD3 with window lengths of 500 and 600. When w = 600, the breakpoint at

position 2000 was undetected. Other breakpoints are also inaccurate. Bottom: Results of Recco showed that only breakpoints at 3000 and 4000 were detected.

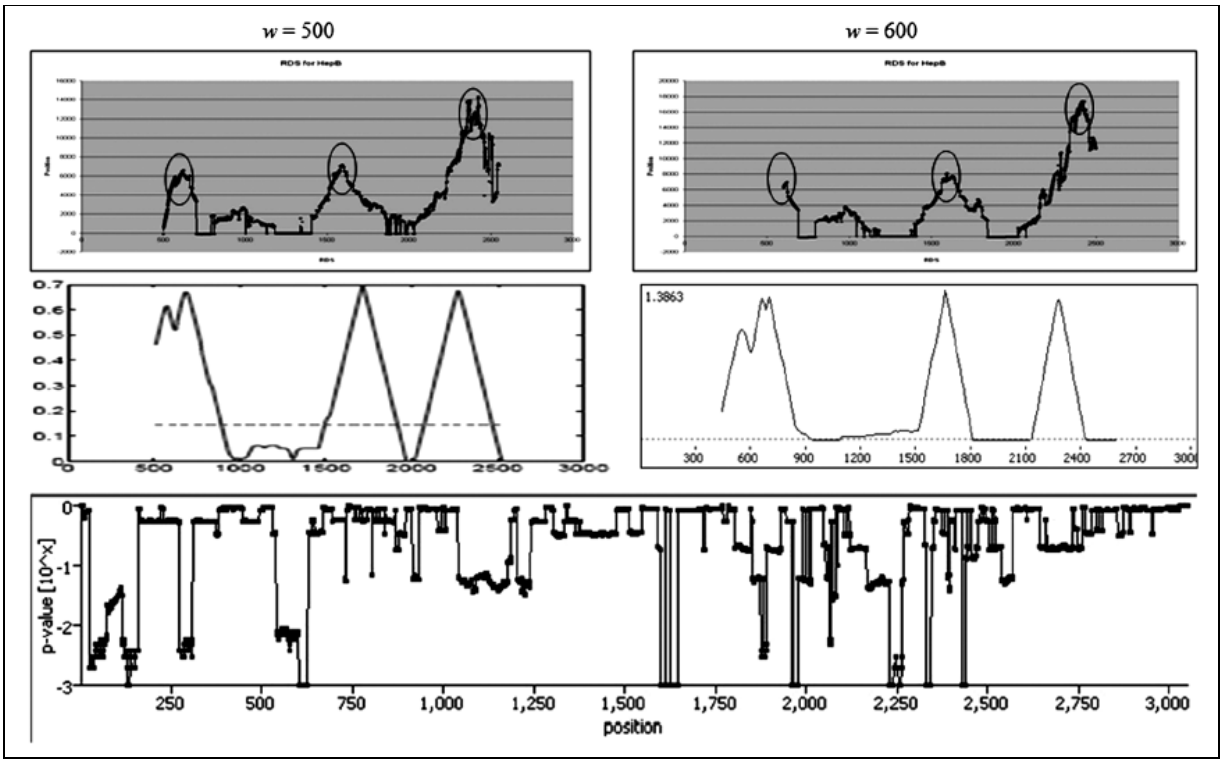

Figure 31: Recombination analysis results on Hep B. Top: Results of our algorithm on Hep B with window lengths of 500 and 600. The circles highlight the recombination breakpoints at position 600, 1700 and 2200 respectively. There is an unreported recombination breakpoint at position 1000. Middle: Results of Pruned-PDM on SD3 with window lengths of 500 and 600. The breakpoints are also detected at the same positions. Bottom: Results of Recco showed that breakpoints at 600, 1700 and 2200 were detected. However, there were several unreported breakpoints detected.

From the experiments, it is interesting to see that the non-window method Recco did not perform as well as the sliding window methods presented, with respect to the four datasets. Recco uses a cost model which tries to reconstruct the recombinant sequence using other sequences in the alignment. In regions where there is weak recombination, this reconstruction is often thwarted by noise and thus may be extremely difficult. In such cases, sliding window

methods which are able to isolate small regions for separate detection analysis may be a better alternative.

5.7 Selection of Optimal Shared (*k*, *m*)-mers

Our algorithm uses shared (*k*,*m*)-mers to avoid the effect of random point mutations in an alignment which improves the accuracy and sensitivity of detecting recombination. Since the underlying mutation rate *m* is automatically determined, the selection of the parameter *k* used to define a shared (*k*,*m*)-mer is crucial in achieving optimal performance for our method. When $k \approx 1$, our algorithm is in fact using hamming distance and hence suffers from noise caused by random point mutations. On the other hand, if *k* is too long, it becomes harder to find shared (*k*,*m*)-mers in the alignment. Consequently, our algorithm loses sensitivity in detecting recombination breakpoints.

To determine the range of *k* values where our method works optimally, we ran RBFinder on dataset SD1 and the more difficult dataset SD2 with different values of *k* and window size of 500. From Figure 32, we see that our algorithm performs optimally when *k* is between 10 and 20. When $k = 1$, our algorithm detects all the recombination breakpoints in SD1 accurately. However, the same *k* did not yield good results when applied on SD2, the dataset with more random point mutations. This clearly demonstrates the susceptibility of conventional distance metrics such as hamming distance when used in recombination detection. On the other hand, when *k* gets large, our method is less sensitive to subtle local alignment changes. Thus, this results in distortions in regions of the graph where the breakpoints are.

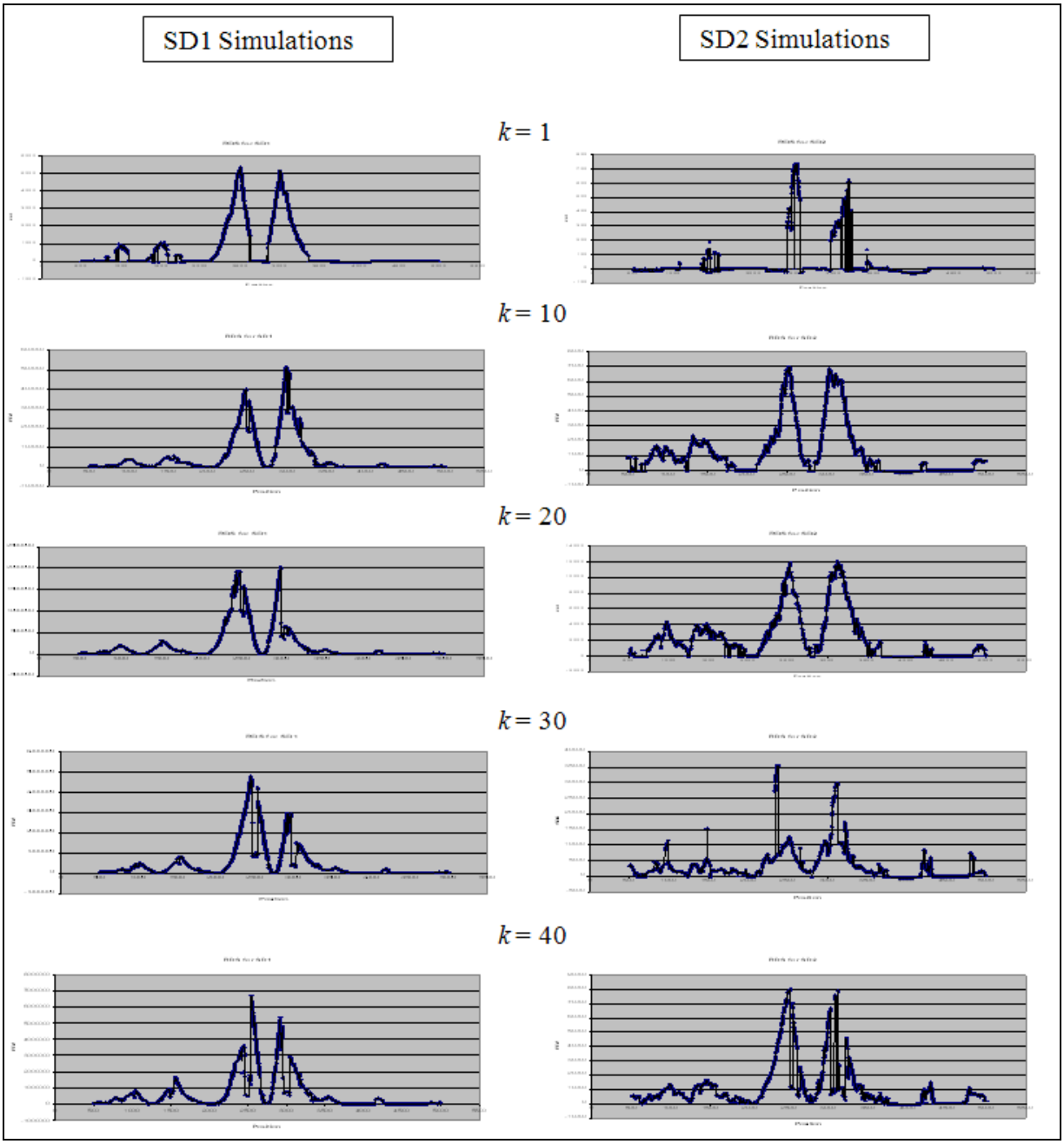

Figure 32: Results of running RBFinder on dataset SD1 and the more difficult dataset SD2 with different values of k for defining shared (k,m)-mers and window size of 500. Our algorithm performs optimally when $k = 10$ and $k = 20$.

5.8 Analysis of Circulating Recombinant Forms of HIV-1

Acquired Immune Deficiency Syndrome (AIDS) is a worldwide epidemic caused by a virus known as human immunodeficiency virus (HIV). Efforts to develop HIV vaccines and medicine have been thwarted with difficulties due to the fast mutation and recombination rates of HIV [121] There are two types of HIV, namely HIV-1 and HIV-2. HIV-1, which is responsible for most human infections, consists of three major groups: M, N and O. The most common group M of HIV-1 is further characterized into 9 subtypes (A, B, C, D, F, H, J and K). The ease at which HIV-1 subtypes recombine has also resulted in numerous circulating recombinant forms (CRFs) of HIV-1 [122]. These CRFs pose more difficulties to finding a cure. Thus, the analysis of existing CRFs and the identification of new CRFs will be vital to the efforts against HIV.

The HIV database (http://www.hiv.lanl.gov/content/hiv-db/mainpage.html) contains a list of all existing CRFs to date. Each CRF is classified according to its recombinant subtypes. Eg CRF02_AG is a recombinant of subtype A and subtype G. In addition, a graphical representation of each CRF in terms of where the recombination occurs at gene level is also available (http://hiv-web.lanl.gov/content/hiv-db/CRFs/CRFs.html). From the database, we downloaded the sequences of 13 CRFs and 35 reference sequences of the 9 subtypes. Experiments are performed in the following manner:

1. For each CRF, download its sequence and several reference sequences of its parent subtypes. Align with ClustalW with default parameters.
2. Run RB-Finder on each alignment. Gaps are not removed because they appear in almost all regions of the alignments. Instead, we modified our program so that a length-w sliding

window is allowed to have $0.1w$ gaps. This is to accommodate insertions and deletions in our computation of recombination breakpoints. To prevent inaccuracies caused by incompletely assemblies, sequences with more than $0.1w$ gaps in a particular window will not be considered.

3. Identify breakpoints of each CRF and compare with the graphical representation of the corresponding CRF. In addition, identify the reference sequences of its parent subtypes that contributed to the recombination.

We summarize our findings in Table 8.

CRF	Reference Strain	Assigned Subtypes	Number of breakpoints in graphical representation	Number of breakpoints detected by RB-Finder	Comments
CRF02_AG	L39106	A,G	9	8	• Breakpoint at LTR not detected because reference sequences at LTR is incomplete • AF286238 belonging to subtype A did not contribute to recombination
CRF03_AB	AF193276	A,B	2	2	• AF286238 belonging to subtype A did not contribute to recombination
CRF05_DF	AF193253	D,F	9	9	• AY371158 belonging to subtype F2 did not contribute to recombination
CRF10_CD	AF289548	C,D	9	9	
CRF12_BF	AF385936	B,F	10	7	• Breakpoint at LTR not detected because reference sequences at LTR is incomplete • 2 short recombinant regions (≈ 100 bp) resulted in 1 breakpoint each instead of 2
CRF14_BG	AF423756	B,G	2	2	
CRF16_A2D	AF457060	A2,D	3	2	• Breakpoint at LTR not detected because reference sequences at LTR is incomplete • None of the reference sequences belonging to subtype A1 contributed to recombination
CRF20_BG	AY900577	B,G	6	4	• 2 short recombinant regions (≈ 100 bp) resulted in 1 breakpoint each

					instead of 2
CRF21_A2D	AF457051	A2,D	7	7	• 3 of 4 of the reference sequences belonging to subtype A1 (AF069670, AF004885, AF484509) contributed to recombination
CRF23_BG	AY900571	B,G	6	5	• 1 short recombinant region (≈ 150 bp) resulted in 1 breakpoint each instead of 2
CRF24_BG	AY900574	B,G	6	4	• 2 short recombinant regions (≈ 100 bp) resulted in 1 breakpoint each instead of 2
CRF28_BF	DQ85873	B,F	2	2	• None of the reference sequences belonging to subtype F2 contributed to recombination
CRF29_BF	DQ85876	B,F	4	4	• None of the reference sequences belonging to subtype F2 contributed to recombination

Table 8: Results of running RB-Finder on 13 CRFs.

RB-Finder finds almost all breakpoints indicated by the literature, except that when the recombination region is too short, RB-Finder reports only one breakpoint for that region instead of two breakpoints. Furthermore, using RB-Finder to identify reference sequences of parent subtypes that contributed to recombination events in the CRFs yielded some interesting findings. Firstly, RB-Finder is able to determine that none of the reference sequences belonging to subtype F2 contributed to recombination events in CRF28_BF and CRF29_BF. Subsequently, we found that although some CRFs are labeled as a recombinant of two subtypes, not all reference sequences of the parent subtypes are involved in the recombination events (Eg CRF02_AG, CRF03_AB and CRF05_DF). On the other hand, we also found that some reference sequences not in the reported parent subtype of a CRF may be involved in some of its recombination events (Eg CRF21_A2D may be a recombinant of subtype A1 as well). This suggests that further analysis may be needed to more accurately classify CRFs. Based on the results from RB-Finder,

we propose a putative phylogenetic network of the 35 reference sequences belonging to the 9 subtypes and the 6 CRFs which had irregularities with their subtyping in Figure 33.

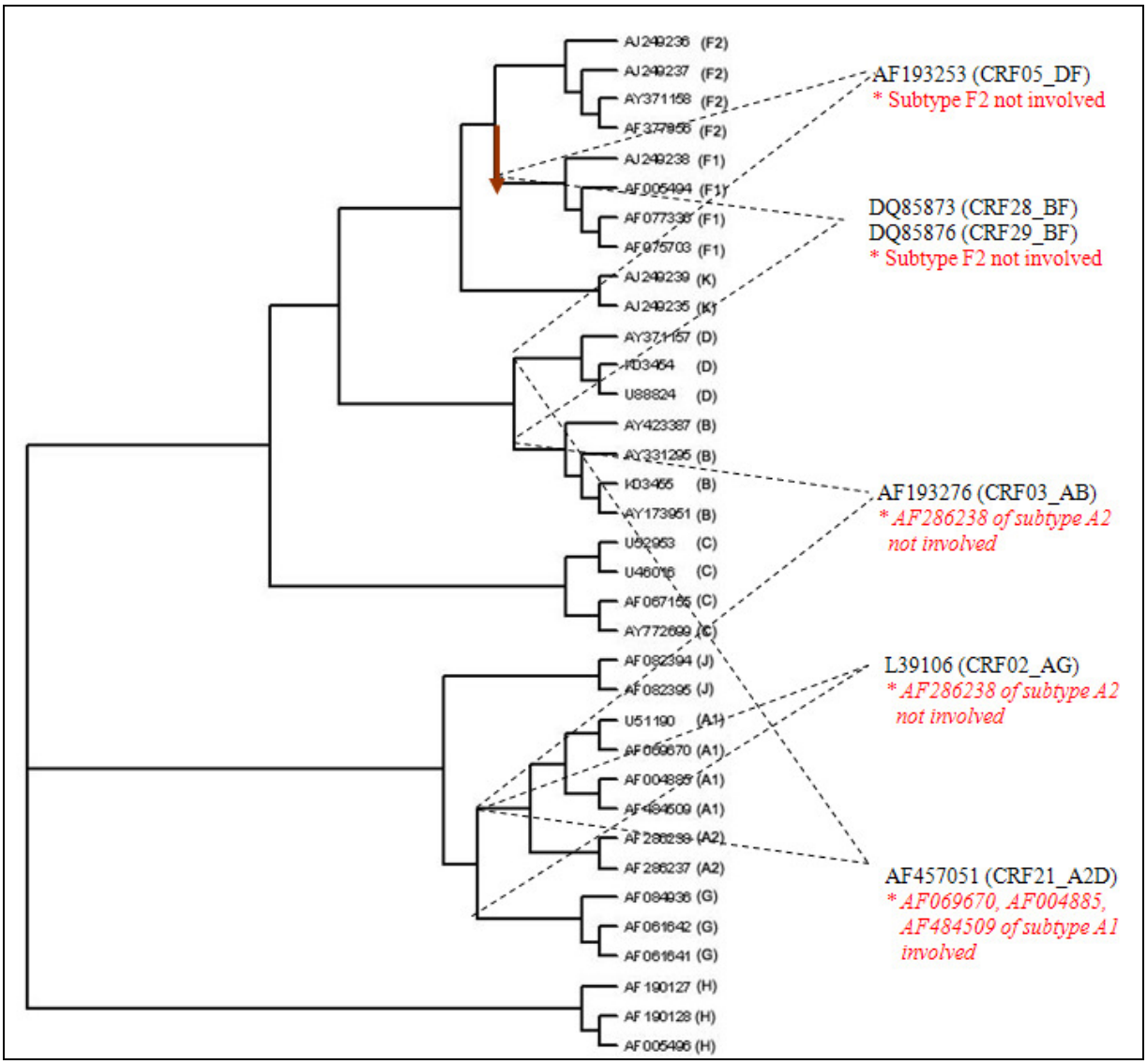

Figure 33: The proposed putative phylogenetic network of 35 reference sequences of the 9 HIV type M subtypes and the 6 CRFs which had irregularities with their subtyping. In the diagram, irregularities are presented in red, italic font under their respective CRFs. In addition, the red arrow indicates our proposed change in genotyping of CRF05_DF, CRF28_BF and CRF29_BF not to include the subtype F2. The other 7 CRFs which had no irregularities with their subtyping are not shown.

Chapter 6

DISCUSSION AND CONCLUSION

6.1 Practical Implications

The current genomic age of biology, powered by high-throughput sequencing technologies, has brought about an explosion of genomic data. As a result, the number of complete genomes of viruses made available in public databases has increased dramatically in recent years. Despite this new wealth of viral genome sequence data, these data has been largely used for evolution and epidemiology studies, with little direct clinical application [123]. Therefore, one of the main goals of this thesis is to develop technologies and bioinformatics tools that have a greater impact on clinical decision-making. For example, our research in tagged-random primer amplification has enabled the design of better tagged-random primers that can uniformly and efficiently amplify a large number of known viruses, as well as novel and unknown viruses that specific primers cannot amplify. Although random primers are generally not as sensitive as specific primers, it is likely that well-designed random primers may have sensitivities comparable to that of specific primers.

In this thesis, we have used microarray technologies to develop two applications of high clinical relevance. The first application, a pathogen detection chip, is aimed at providing clinicians with a fast and accurate diagnosis using carefully selected oligonucleotide probes from hundreds of reported human viruses in a single experiment. This will eliminate unnecessary lab tests based on educated guesses from the patient's symptoms and significantly reduce the cost and turnaround time of a medical diagnosis. Furthermore, the pathogen chip with its

accompanying statistical software would be able to detect and distinguish a majority of virus variants so that the correct treatment can be administered decisively. The second application, a resequencing microarray (H1N1 2009), is a cost-effective and viable tool for large-scale biosurveillance in the H1N1 2009 pandemic. The accompanying base-calling software EvolSTAR has also greatly improved the accuracy and base-calling rate of existing base-calling algorithms. Most importantly, our study has highlighted the feasibility of using resequencing microarrays for high-throughput full genome sequencing of viruses. In our application, resequencing microarrays are relatively low-cost, costing only a 10^{th} that of a 454 run, and equivalent to that of a traditional capillary sequencing run. However, through multiplexing, our system can generate full genomes of 24 different H1N1(2009) samples in 30 hours. In comparison, capillary sequencing and next-generation technologies such as 454 may obtain full genomes of only one or two different samples in the same time-frame.

Last but not least, we studied the problem of detecting recombination in a set of aligned virus sequences. We have identified and addressed the major causes of inaccuracies pertaining to sliding window distance-based methods for detecting recombination. The resulting distance-based algorithm RBFinder is able to achieve accuracies comparable to phylogeny-based methods, but requires a much shorter analysis time. This would enable recombination detection analysis to be performed on large datasets, which would otherwise be impractical using phylogeny-based methods.

6.2 Conclusion

Viruses have the ability to cause devastating pandemics that may result in high morbidity and mortality rates. Thus it is imperative that any viral outbreaks be detected and contained as fast as possible. The next step would then be to obtain viral genome sequence information that will give us important clues on its lineage, epidemiology, drug resistance and possible vaccine development. These factors motivated us to develop a number of bioinformatics tools for viral detection, resequencing and evolutionary analysis. A summary of the work accomplished in this thesis is as follows:

- In Chapter 2, we studied how the primer efficiency and biases inherent in random PCR amplification affect accuracies in PCR-based detection methods. We describe a model that predicts the amplification efficiency of a given tagged-random primer on a target viral genome. The prediction allows us to filter false-negative probes of the genome that lie in regions of poor random PCR amplification and improves the accuracy of pathogen detection. Subsequently, we propose LOMA, an algorithm to generate random primers that have good amplification efficiency. Wet-lab validation showed that the generated random primers improve the amplification efficiency significantly. The blind use of a random primer with attached universal tag in a PCR reaction on a pathogen sample may not lead to a successful amplification. Thus, the design of tagged-random primers is an important consideration when performing PCR. This work has been published in BMC Bioinformatics 2008 [124].

- In Chapter 3, we investigated the potential of DNA microarrays as "genomic sensors" in clinical diagnostics. Biases inherent in random PCR-amplification, cross-hybridization

effects, and inadequate microarray analysis, however, limit detection sensitivity and specificity. We studied the relationships between viral amplification efficiency, hybridization signal, and target-probe annealing specificity using a customized microarray platform. Novel features of this platform include the development of a robust algorithm that accurately predicts PCR bias during DNA amplification and can be used to improve PCR primer design, as well as a powerful statistical concept for inferring pathogen identity from probe recognition signatures. Compared to real-time PCR, the microarray platform identified pathogens with 94% accuracy (76% sensitivity and 100% specificity) in a panel of 36 patient specimens. Our findings show that microarrays can be used for the robust and accurate diagnosis of pathogens, and further substantiate the use of microarray technology in clinical diagnostics. This work has been published in Genome Biology 2007 [125].

- In Chapter 4, we developed and field-tested a resequencing kit that is capable of interrogating all eight segments of the 2009 influenza A(H1N1) virus genome and its variants, with added focus on critical regions such as drug-binding sites, structural components and mutation hotspots. The accompanying base-calling software (EvolSTAR) introduces novel methods that utilize neighbourhood hybridization intensity profiles and substitution bias of probes on the microarray for mutation confirmation and recovery of ambiguous base queries. Our results demonstrate that EvolSTAR is highly accurate and has a much improved call rate. The high throughput and short turn-around time from sample to sequence and analysis results (30 hours for 24 samples) makes this kit an efficient large-scale evolutionary biosurveillance tool. This work has been

published in Nucleic Acids Research 2010 [126]. An application of our tool has also been accepted for publication in New England Journal of Medicine 2010 [127].

- In Chapter 5, we developed RB-Finder, a distance-based window method for finding recombination breakpoints in a set of alignments. By avoiding the computationally expensive and complicated comparisons of phylogenetic trees of phylogeny-based methods, our algorithm is faster and thus more scalable to analyze big datasets. Moreover, we minimize information loss experienced by conventional distance-based methods by introducing a new distance metric that takes into consideration important details such as the distribution of mismatches, the number of consecutive matches and the locations of common subsequences in an alignment. To improve the accuracy, we propose using a weighting strategy that assigns different weights to positions enclosed in a window with respect to a putative breakpoint. The idea is to lessen the contribution of less important regions of a window when computing a distance measure for the alignment. Subsequently, we applied our weighting strategy to our new distance metric and describe a fast, simple and intuitive algorithm to detect recombination. Experimental results using both simulated and real datasets show that the efficiency and accuracy of RB-Finder are better than that of most existing methods. In addition, we present an application of RB-Finder in genotyping by analyzing a set of 13 HIV recombinant sequences. In our analysis, we detected almost all reported breakpoints of the 13 sequences and made several novel findings regarding their genotypes. Specifically, we found irregularities in the genotyping of six sequences which may trigger new considerations when assigning genotypes. This work has been published in RECOMB

2007 [128]. The extended version of this work is also published in a special issue of Journal of Computational Biology 2008 [129].

6.3 Future Research Work

The extensive work presented in this thesis on viral detection, resequencing and evolutionary analysis has established a large set of paths for future work. We briefly describe some of our new discoveries and research developments below.

6.3.1 Tagged-Random Primer Design with Background Amplification Avoidance

In Chapter 2, we discussed how tagged-random primers can theoretically amplify any sequence and hence have a higher chance of successfully amplifying novel, unknown and highly mutative/recombinative viruses than specific primers in PCR-based detection methods. However, this also means that tagged-random primers can amplify the background human genome and cause undesirable cross-hybridization noise. To minimize the background cross-hybridization noise, the tagged-random primer should be designed such that it has (1) high AES (amplifies well) with the target viral genomes and (2) has low AES (amplifies poorly) with the human genome. The challenge here is that it may be difficult to find candidates that amplifies the entire human genome poorly. One feasible solution is to only consider regions in the human genome where the viral probes on the microarray can cross-hybridize to. Our simulations show that we need to consider only one-third of the human genome if such an approach is used. This research

direction may pave the way for tagged-random primers with improved specificity and sensitivity for viral amplification, as well as background amplification avoidance capabilities.

6.3.2 New Generation Pathogen Chip

In Chapter 3, the pathogen chip consists of probes selected from the refSeq (as indicated by NCBI) of a target set of virus genomes. However, this was found to be not representative of a virus genome. There are often multiple genomes of a particular virus because various sequencing centres deposit complete and partial sequences of the same virus genome into GenBank.

Using the most current NT and RefSeq database to date (Feb 2010) in GenBank, we downloaded all complete genomes, complete sequences and complete cds and genes of 153 human viruses of interest. In total, 25177 sequences were obtained for the 153 viruses. A homology analysis of the sequences of individual viruses revealed the following:

- For most of the 153 viruses, there are multiple complete genomes. The most recent submission to GenBank is assigned as the reference genome for the virus.
- There are SNPs among different genomes of the same virus. They may be mutations or sequencing errors.
- It is expected that genomes belonging to different strains of a particular virus may differ greatly. However, we found that genomes belonging to the same virus without any previously reported strains may also differ greatly. This may be unreported strains of the virus residing in different parts of the world. It can also be that the same virus underwent independent evolutions.

The findings have a serious implication on probe selection for the pathogen chip. Probes selected based on a reference genome for a particular virus have a high chance of failing. These probes may detect one strain of a virus but not the other.

To avoid the above problem, for each virus, we cluster all the sequences available based on homology. Note that this clustering is blind to strain typing. We cluster sequences that have at least 98% homology with one another (performed by BLAST). This will ensure that we identify all subgroups of a virus based on homology. We then align all sequences in each subgroup and obtain the consensus sequence. The consensus sequences of the subgroups of a virus will then be collectively taken to represent the virus. In each alignment, we can also identify suspicious regions where the bases are not consistent. These regions may be mutagenic regions or prone to sequencing or base-calling errors. Thus, probes will not be selected from such regions to minimize the risk of failure. Subgrouping the 25177 sequences belonging to 153 viruses resulted in 7790 subgroups. 5677 subgroups are singletons while the remaining 2113 subgroups had more than one sequence each. In total, 19500 sequences were able to be clustered.

Due to the number of subgroups that each virus has, probes that cover the most subgroups are preferred for selection. In addition, probes that are unique to a subgroup are also important as they may help to differentiate the subgroups if needed. The location of the probes may also be important. Consecutive probes lighting up may be an indication that the fragment that the probes reside in is present. This may help in identifying recombinations or detecting the virus even though the full genome was not fully amplified.

6.3.3 Pan-Influenza Resequencing Microarray

In Chapter 4, we have presented a novel approach to obtain the full genome sequence (all 8 segments) of an influenza virus (H1N1 2009) using resequencing microarrays. However, as influenza viruses continuously evolve through mutations and reassortments, we need to periodically update (at least annually) our resequencing microarrays with sequence information from new variants. One solution is to create a pan-influenza resequencing microarray consisting of full genome sequence information from all recent (e.g. 2005 and newer) human/avian/mammalian influenza virus strains. Unfortunately, a naïve 1bp resolution tiling approach would be infeasible for such a resequencing microarray. Instead, a minimal set of probes that covers all target sequences would have to be selected and put on the array. A sophisticated assembly algorithm would then be needed to reconstruct high confidence sequence calls from the selected probes. A major advantage of this pan-influenza resequencing microarray is that it would have a much longer shelf life than existing influenza resequencing microarrays.

6.3.4 Recombination Detection Without Multiple Sequence Alignment

In Chapter 5, we note that most recombination detection methods require a multiple alignment of the input sequences as an initial step. This is problematic in several aspects. First and foremost, obtaining the multiple sequence alignment of a set of sequences is slow. The exact solution to the multiple sequence alignment problem is NP-complete [130]. Specifically, it takes $O(n^k)$ time to align k sequences of length n. Hence, heuristic and approximation approaches that seek a multiple alignment that maximizes some alignment score are adopted. One of the most popular algorithm for multiple sequence alignment is CLUSTALW. It uses a strategy known as

progressive alignment and runs in $O(k^2n^2)$ time. Another multiple sequence alignment program, T-Coffee [131], is more accurate than CLUSTALW but is also much slower. Later, MUSCLE [132] was developed that outperforms both CLUSTALW and T-Coffee in terms of speed and accuracy. However, the inability to scale of current multiple sequence alignment algorithms presents an obstacle for recombination analysis involving a large number of sequences.

Secondly, the multiple sequence alignment of sequences may be inaccurate. Misalignments can occur due to various reasons such as using suboptimal scoring parameters. The accuracy of the alignment may be even more compromised when analyzing highly evolving microorganisms. These microorganisms exhibit high levels of intragenic recombination following horizontal gene transfer events, resulting in mosaic-like sequences. Unless these sequences have high homology with one another in similar genomic regions, multiple sequence alignments may be difficult or impossible to attain. Consequently, important recombination events may be lost in the erroneous alignment. Furthermore, multiple alignments of sequences of different lengths must necessarily add gaps, which often lead to loss of information and gap scoring artifacts, which in turn distort the distance computations and phylogeny constructions. There is also a risk that these gaps, which are almost always ignored by recombination detection algorithms, contain information about recombination events involving duplications and transpositions. In many cases, multiple sequence alignments have to be manually refined to improve their accuracy.

To overcome the problem of using multiple sequence alignments, BLAST-Miner [133] was developed. It is a BLAST-based method that finds short segments of highly similar sequences among the sequences and uses them to identify sequence duplications, insertions, and rearrangements. One major disadvantage is that this approach cannot detect recombination

events among highly homologous sequences. Another drawback is that the authors only gave a graphical representation of all possible recombination events identified but failed to quantify them. Nonetheless, this approach provides an insight of how we can avoid using multiple sequence alignments to detect recombination.

REFERENCES

1. Smith RD. Responding to global infectious disease outbreaks: Lessons from SARS on the role of risk perception, communication and management. *Journal of Social Science and Medicine*, 2006, 63: 3113-3123.

2. Worobey M, Gemmel M, Teuwen DE, Haselkorn T, Kunstman K, Bunce M, Muyembe JJ, Kabongo JM, Kalengayi RM, Van Marck E, Gilbert MT, Wolinsky SM. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature*, 2008, 455(7213): 661–664.

3. Feng H, Shuda M, Chang Y, Moore P. Clonal Integration of a Polyomavirus in Human Merkel Cell Carcinoma. *Science*, 2008, 319(5866): 1096-1100.

4. Walboomers JM, Jacobs MV, Manos MM. Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. *J. Pathol*., 1999, 189(1): 12–19.

5. Michael B, Nair AM, Hiraragi H, Shen L, Feuer G, Boris-Lawrie K, Lairmore MD. Human T lymphotropic virus type-1 p30II alters cellular gene expression to selectively enhance signaling pathways that activate T lymphocytes. *Retrovirology*, 2004, 1: 39.

6. Chang M. Hepatitis B virus infection. *Seminars in fetal & neonatal medicine*, 2007, 12(3): 160–167.

7. Pattle SB, Farrell PJ. The role of Epstein-Barr virus in cancer. *Expert Opin Biol Ther.*, 2006, 6(11): 1193-1205.

8. Chang Y, Cesarman E, Pessin MS, Lee F, Culpepper J, Knowles DM, Moore PS. Identification of herpesvirus-like DNA sequences in AIDS-associated Kaposi's sarcoma. *Science*, 1994, 266(5192): 1865-1869

9. Brower V. Infections linked to prostate cancer. *Nature Medicine*, 2009, 15: 1098.

10. Witzany G. Natural Genome Editing Competences of Viruses. *Act. Biotheor.*, 2006, 54: 235-253.

11. Taubenberger J, Morens D. 1918 Influenza: the mother of all pandemics. *Emerg Infect Dis.*, 2006, 12(1): 15–22.

12. UNAIDS.AIDS epidemic update, 2009 http://www.unaids.org/en/KnowledgeCentre/HIVData/EpiUpdate/EpiUpdArchive/2009/default.asp

13. Woolhouse MEJ, Gaunt E. Ecological origins of novel human pathogens. *Crit. Rev. Microbiol.*, 2007, 33: 1–12.

14. Woolhouse MEJ, Howey R, Gaunt E, Reilly L, Chase-Topping M, Savill N. Temporal trends in the discovery of human viruses. *Proc. R. Soc. B*, 2008, 275(1647): 2111-2115.

15. Leland DS, Ginocchio CC. Role of cell culture for virus detection in the age of technology. *Clinical Microbiology Reviews (American Society for Microbiology)*, 2007, 20(1): 49-78.

16. Metcalf JA, Davey RT, Lane HC. Acquired Immunodeficiency Syndrome: Serologic and Virologic Tests. In DEVITA VT, CURRAN J, HELLMAN S, et al. *AIDS: Etiology,*

Diagnosis, Treatment and Prevention. 4th Edition. Philadelphia: Lippincott-Raven, 1997, 177-196.

17. Washington JA. Principles of Diagnosis: Serodiagnosis. in: *Baron's Medical Microbiology* (Baron S et al., eds.) (4th ed.). Univ of Texas Medical Branch, 1996.

18. Nii S, Morgan C, Rose HM. Electron microscopy of herpes simplex virus. II. Sequence of development. *J Virol.*, 1968, 2(5): 517–536.

19. Henrickson KJ. Advances in the laboratory diagnosis of viral respiratory disease. *The Pediatric Infectious Disease Journal*, 2004, 23(1): S6-S10.

20. Steininger C, Kundi M, Aberle SW, Aberle JH, Popow-Kraupp T. Effectiveness of Reverse Transcription-PCR, Virus Isolation, and Enzyme-Linked Immunosorbent Assay for Diagnosis of Influenza A Virus Infection in Different Age Groups. *Journal of Clinical Microbiology*, 2002, 40(6): 2051-2056.

21. Liu QJ, Bai YF, Ge QY, Zhou SX, Wen T, Lu ZH. Microarray-in-a-Tube for Detection of Multiple Viruses. *Clin Chem.*, 2007, 53(2): 188-94.

22. Zhang T, Breitbart M, Lee WH, Run JQ, Wei CL, Soh SW, Hibberd ML, Liu ET, Rohwer F, Ruan Y. RNA viral community in human feces: prevalence of plant pathogenic viruses. *PLoS Biol.*, 2006, 4(1): e3.

23. Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SB, Hood LE. Fluorescence detection in automated DNA sequence analysis. *Nature*, 1986, 321(6071): 674–679.

24. Hall N. Advanced sequencing technologies and their wider impact in microbiology. *J. Exp. Biol.*, 2007, 210(9): 1518–1525.

25. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM. The complete genome of an individual by massively parallel DNA sequencing. *Nature*, 2008, 452(7189): 872-876.

26. Gupta PK. Ultrafast and low-cost DNA sequencing methods for applied genomics research. *Proceedings of the National Academy of Sciences India Section B-biological Sciences*, 2008, 78: 91-102.

27. Lazarevic V, Whiteson K, Huse S, Hernandez D, Farinelli L, Osterås M, Schrenzel J, François P. Metagenomic study of the oral microbiota by Illumina high-throughput sequencing. *J Microbiol Methods*, 2009, 79(3): 266-271.

28. Hanna GJ, Johnson VA, Kuritzkes DR, Richman DD, Martinez-Picado J, Sutton L, Hazelwood JD, D'Aquila RT. Comparison of sequencing by hybridization and cycle sequencing for genotyping of human immunodeficiency virus type 1 reverse transcriptase. *J. Clin. Microbiol.*, 2000, 38(7): 2715–2721.

29. Preparata FP, Upfal E. Sequencing-by-hybridization at the information-theory bound: an optimal algorithm. *In Proceedings of the Fourth Annual International Conference on Computational Molecular Biology*, 2000, 245-253.

30. Hurst LD, Peck JR. Recent advances in understanding of the evolution and maintenance of sex. *Trends Ecol. Evol.*, 1996, 11:46-52.

31. Raju R, Subramaniam SV, Hajjou M. Genesis of Sindbis virus by in vivo recombination of nonreplicative RNA precursors. *Journal of Virology*, 1995, 69: 7391-7401.

32. Meyers G, Tautz N, Becher P, Thiel HJ, Kümmerer BM. Recovery of Cytopathogenic and Noncytopathogenic Bovine Viral Diarrhea Viruses from cDNA Constructs. *J Virol.*, 1996, 70(12): 8606-8613.

33. Muñoz ET, Deem MW. Epitope analysis for influenza vaccine design. *Vaccine*, 2005, 23(9): 1144-1148.

34. Watkins DI. Basic HIV Vaccine Development. *Top HIV Med.*, 2008, 16 (1): 7–8.

35. Gotoh O. Multiple sequence alignment: algorithms and applications. *Adv. Biophys.*, 1999, 36: 159-206.

36. Weiller GF. Phylogenetic profiles: a graphical method for detecting genetic recombination in homologous sequences. *Mol. Biol. Evol.*, 1998, 15: 326-335.

37. Myers SR, Griffiths RC. Bounds on the minimum number of recombination events in a sample history. *Genetics*, 2003, 163: 375-394.

38. Striebel HM, Birch-Hirschfeld E, Egerer R, Földes-Papp Z. Virus diagnostics on microarrays. *Curr Pharm Biotechnol.*, 2003, 4(6): 401-415.

39. Lin B, Wang Z, Vora GJ, Thornton JA, Schnur JM, Thach DC, Blaney KM, Ligler AG, Malanoski AP, Santiago J, Walter EA, Agan BK, Metzgar D, Seto D, Daum LT,

Kruzelock R, Rowley RK, Hanson EH, Tibbetts C, Stenger DA. Broad-spectrum respiratory tract pathogen identification using resequencing DNA microarrays. *Genome Res.*, 2006, 16(4): 527-535.

40. Hollingshead D, Lewis DA, Mirnics K. Platform influence on DNA microarray data in postmortem brain research. *Neurobiol. Dis.*, 2005, 18: 649-655.

41. Klaassen CH, Prinsen CF, de Valk HA, Horrevorts AM, Jeunink MA, Thunnissen FB. DNA microarray format for detection and subtyping of human papillomavirus. *J Clin Microbiol.*, 2004, 42(5): 2152-2160.

42. Laassri M, Chizhikov V, Mikheev M, Shchelkunov S, Chumakov K. Detection and discrimination of orthopoxviruses using microarrays of immobilized oligonucleotides. *J Virol Methods*, 2003, 112(1-2): 67-78.

43. Hsia CC, Chizhikov VE, Yang AX, Selvapandiyan A, Hewlett I, Duncan R, Puri RK, Nakhasi HL, Kaplan GG. Microarray multiplex assay for the simultaneous detection and discrimination of hepatitis B, hepatitis C, and human immunodeficiency type-1 viruses in human blood samples. *Biochem Biophys Res Commun.*, 2007, 356(4): 1017-1023.

44. Bresnahan WA, Shenk T. A subset of viral transcripts packaged within human cytomegalovirus particles. *Science*, 2000. 288: 2373-2376.

45. Han X, Lin X, Liu B, Hou Y, Huang J, Wu S, Liu J, Mei L, Jia G, Zhu Q. Simultaneously subtyping of all influenza A viruses using DNA microarrays. *J Virol Methods*, 2008, 152(1-2): 117-121.

46. Urisman A, Fischer KF, Chiu CY, Kistler AL, Beck S, Wang D, DeRisi JL: EPredict: a computational strategy for species identification based on observed DNA microarray hybridization patterns. *Genome Biol.*, 2005, 6: R78.

47. Conejero-Goldberg C, Wang E, Yi C, Goldberg TE, Jones-Brando L, Marincola FM, Webster MJ, Torrey EF. Infectious pathogen detection arrays: viral detection in cell lines and postmortem brain tissue. *Biotechniques*, 2005, 39(5): 741-751.

48. Chou CC, Lee TT, Chen CH, Hsiao HY, Lin YL, Ho MS, Yang PC, Peck K. Design of microarray probes for virus identification and detection of emerging viruses at the genus level. *BMC Bioinformatics*, 2006, 7: 232.

49. Metzgar D, Myers CA, Russell KL, Faix D, Blair PJ, Brown J, Vo S, Swayne DE, Thomas C, Stenger DA, Lin B, Malanoski AP, Wang Z, Blaney KM, Long NC, Schnur JM, Saad MD, Borsuk LA, Lichanska AM, Lorence MC, Weslowski B, Schafer KO, Tibbetts C. Single Assay for Simultaneous Detection and Differential Identification of Human and Avian Influenza Virus Types, Subtypes, and Emergent Variants. *PLoS ONE*, 2010, 5(2): e8995.

50. Wong CW, Albert TJ, Vega VB, Norton JE, Cutler DJ, Richmond TA, Stanton LW, Liu ET, Miller LD. Tracking the evolution of the SARS coronavirus using high-throughput, high-density resequencing arrays. *Genome Res.*, 2004, 14(3): 398-405.

51. Sulaiman IM, Liu X, Frace M, Sulaiman N, Olsen-Rasmussen M, Neuhaus E, Rota P, Wohlhueter RM. Evaluation of Affymetrix severe acute respiratory syndrome resequencing GeneChips in characterization of the genomes of two strains of coronavirus infecting humans. *Appl. Environ. Microbiol.*, 2006, 72: 207-211.

52. Sulaiman IM, Tang K, Osborne J, Sammons S, Wohlhueter RM. GeneChip Resequencing of the Smallpox Virus Genome Can Identify Novel Strains: a Biodefense Application. *Journal of Clinical Microbiology*, 2007, 45(2):. 358-363.

53. Leski TA, Lin BC, Malanoski AP, Wang Z, Long NC, Meador CE, Barrows B, Ibrahim S, Hardick JP, Aitichou M, Schnur JM, Tibbetts C, Stenger DA. Testing and Validation of High Density Resequencing Microarray for Broad Range Biothreat Agents Detection. *PLoS One*, 2009, 4(8): e6569.

54. Cutler DJ, Zwick ME, Carrasquillo MM, Yohn CT, Tobin KP, Kashuk C, Mathews DJ, Shah NA, Eichler EE, Warrington JA, Chakravarti A. High-throughput variation detection and genotyping using microarrays. *Genome Res.*, 2001, 11: 1913–1925.

55. Hacia JG. Resequencing and mutational analysis using oligonucleotide microarrays. *Nat Genet.*, 1999, 21(1): 42-47.

56. Zhan YP, Kulp D. Model-P: a basecalling method for resequencing microarrays of diploid samples. *Bioinformatics*, 2005, 21(2): 182-189.

57. Pandya GA, Holmes MH, Sunkara S, Sparks A, Bai Y, Verratti K, Saeed K, Venepally P, Jarrahi B, Fleischmann RD, Peterson SN. A bioinformatic filter for improved base-call accuracy and polymorphism detection using the Affymetrix GeneChip® whole-genome resequencing platform. *Nucleic Acids Res.*, 2007, 35(21): e148.

58. Zheng J, Moorhead M, Weng L, Siddiqui F, Carlton VE, Ireland JS, Lee L, Peterson J, Wilkins J, Lin S, Kan Z, Seshagiri S, Davis RW, Faham M. High-throughput, high-accuracy array-based resequencing. *Proc Natl Acad Sci USA.*, 2009, 106(16): 6712-6717.

59. Posada D. Evaluation of methods for detecting recombination from DNA sequences: empirical data. *Mol Biol Evol.*, 2002, 19(5): 708-717.

60. Posada D, Crandall KA. The effect of recombination on the accuracy of phylogeny estimation. *J Mol Evol*., 2002, 54(3): 396-402.

61. Schierup MH, Hein J. Consequences of recombination on traditional phylogenetic analysis. *Genetics*, 2000, 156(2): 879-891.

62. McGuire G, Wright F. TOPAL 2.0: improved detection of mosaic sequences within multiple alignments. Bioinformatics, 2000, 16(2): 130-134.

63. Husmeier D, Wright F. Probabilistic divergence measures for detecting interspecies recombination. *Bioinformatics*, 2001, 17(1): S123-S131.

64. Husmeier D, Wright F, Milne I. Detecting interspecific recombination with a pruned probabilistic divergence measure. Bioinformatics, 2005, 21(9): 1797-1806.

65. Ruths D, Nakhleh L. RECOMP: A parsimony-based method for detecting recombination. *4th Asia Pacific Bioinformatics Conference*, 2006, 59-68.

66. Robinson DF, Foulds LR. Comparison of phylogenetic trees. *Mathematical Biosciences*, 1981, 53(1-2): 131-147.

67. Maydt J, Lengauer T. Recco: recombination analysis using cost optimization. Bioinformatics, 2006, 22(9): 1064-1071.

68. Bartlett JM, Stirling D. A Short History of the Polymerase Chain Reaction. *Methods Mol Biol.*, 2003, 226: 3-6.

69. Chow-Shaffer E, Sina B, Hawley WA, De Benedictis J, Scott TW. (2000) Laboratory and Field Evaluation of Polymerase Chain Reaction-Based Forensic DNA Profiling for Use in Identification of Human Blood Meal Sources of Aedes aegypti (Diptera: Culicidae). *Journal of Medical Entomology*, 2000, 37(4): 492–502.

70. Krause J, Dear PH, Pollack JL, Slatkin M, Spriggs H, Barnes I, Lister AM, Ebersberger I, Pääbo S, Hofreiter M. Multiplex amplification of the mammoth mitochondrial genome and the evolution of Elephantidae. *Nature*, 2006, 439(7077): 724-727.

71. Ståhlberg A, Zoric N, Aman P, Kubista M. Quantitative real-time PCR for cancer detection: the lymphoma case. *Expert Rev Mol Diagn.*, 2005, 5(2): 221-230.

72. Umlauft F, Wong DT, Oefner PJ, Underhill PA, Cheung RC, Wright TL, Kolykhalov AA, Gruenewald K, Greenberg HB. Hepatitis C virus detection by single-round PCR specific for the terminal 3' noncoding region. *J Clin Microbiol.*, 1996, 34(10): 2552–2558.

73. Lakeman FD, Whitley RJ. Diagnosis of herpes simplex virus encephalitis: application of polymerase chain reaction to CSF from brain-biopsied patients and correlation with disease. *J. Infect. Dis.*, 1995, 171: 857-863.

74. Grothues D, Smith CL, Cantor CR: PCR Amplification of Megabase DNA with Tagged Random Primers (T-PCR). *Nucleic Acids Research*, 1993, 21: 1321–1322

75. Raghunathan A, Ferguson Jr HR, Bornarth CJ, Song W, Driscoll M, Lasken RS. Genomic DNA Amplification from a Single Bacterium. *Applied and Environmental Microbiology*, 2005, 71(6): 3342–3347

76. Quan PL, Palacios G, Jabado OJ, Conlan S, Hirschberg DL, Pozo F, Jack PJM, Cisterna D, Renwick N, Hui J, Drysdale A, Amos-Ritchie R, Baumeister E, Savy V, Lager KM, Richt JA, Boyle DB, Garcia-Sastre A, Casas I, Perez-Brena P, Briese T, Lipkin WI. Detection of Respiratory Viruses and Subtype Identification of Influenza A Viruses by GreeneChipResp Oligonucleotide Microarray. *J. Clin. Microbiol*, 2007, 45(8): 2359–2364

77. Nguyen HK, Southern EM: Minimizing the secondary structure of DNA targets by incorporation of a modified deoxynucleotide: implications for nucleic acid analysis by hybridization. *Nucleic Acids Res.*, 2000, 28: 3904–3909

78. Hu A, Colella M, Tam JS, Rappaport R, Cheng SM. Simultaneous detection, subgrouping, and quantitation of respiratory syncytial virus A and B by real-time PCR. *J. Clin Microbiol.*, 2003, 41: 149–154

79. Contoli M, Message SD, Laza-Stanca V, Edwards MR, Wark PA, Bartlett NW, Kebadze T, Mallia P, Stanciu LA, Parker HL, Slater L, Lewis-Antes A, Kon OM, Holgate ST, Davies DE, Kotenko SV, Papi A, Johnston SL. Role of deficient type III interferon-lambda production in asthma exacerbations. *Nat Med.*, 2006, 12(9): 1023–1026

80. Moës E, Vijgen L, Keyaerts E, Zlateva K, Li S, Maes P, Pyrc K, Berkhout B, van der Hoek L, Van Ranst M. A novel panacoronavirus RT-PCR assay: frequent detection of human coronavirus NL63 in children hospitalized with respiratory tract infections in Belgium. *BMC Infectious Diseases*, 2006, 5: 6

81. SantaLucia JJ, Allawi HT, Seneviratne PA: Improved Nearest-Neighbor Parameters for Predicting DNA Duplex Stability. *Biochemistry*, 1996, 35: 3555-3562

82. Burpo FJ. A critical review of PCR primer design algorithms and crosshybridization case study. *Biochemistry*, 2001, 218.

83. McConaughy BL, Laird CD, McCarthy BJ. Nucleic acid reassociation in formamide. *Biochemistry*, 1969, 8: 3289-3295.

84. Broude NE, Driscoll K, Cantor CR. High-Level Multiplex DNA Amplification. *Antisense and Nucleic Acid Drug Development*, 2001, 11(5): 327-332.

85. Simmler H, Singpiel H, Männer R. Real-time Primer Design for DNA chips. *Proceedings Parallel and Distributed Processing Symposium*, 2003, 153b.

86. Robertson BH, Nicholson JK. New microbiology tools for public health and their implications. *Annu Rev Public Health*, 2005, 26: 281-302.

87. Wang D, Coscoy L, Zylberberg M, Avila PC, Boushey HA, Ganem D, DeRisi JL. Microarray-based detection and genotyping of viral pathogens. *PNAS*, 2002, 99: 15687-15692.

88. Wang D, Urisman A, Liu YT, Springer M, Ksiazek TG, Erdman DD, Mardis ER, Hickenbotham M, Magrini V, Eldred J, Latreille JP, Wilson RK, Ganem D, DeRisi JL. Viral discovery and sequence recovery using DNA microarrays. *PLoS Biol.*, 2003, 1(2): E2.

89. Hong BX, Jiang LF, Hu YS, Fang DY, Guo HY. Application of oligonucleotide array technology for the rapid detection of pathogenic bacteria of foodborne infections. *J Microbiol Methods*, 2004, 58: 403-411.

90. Sergeev N, Distler M, Courtney S, Al-Khaldi SF, Volokhov D, Chizhikov V, Rasooly A. Multipathogen oligonucleotide microarray for environmental and biodefense applications. *Biosens Bioelectron*, 2004, 20: 684-698.

91. Bodrossy L, Sessitsch A. Oligonucleotide microarrays in microbial diagnostics. *Curr Opin Microbiol.*, 2004, 7: 245-254.

92. Vora GJ, Meador CE, Stenger DA, Andreadis JD. Nucleic acid amplification strategies for DNA microarray-based pathogen detection. *Appl Environ Microbiol*., 2004, 70: 3047-3054.

93. Loy A, Bodrossy L. Highly parallel microbial diagnostics using oligonucleotide microarrays. *Clin. Chim. Acta*, 2006, 363, 106–119.

94. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol.*, 1996, 14(13): 1675-80.

95. Nuwaysir EF, Huang W, Albert TJ, Singh J, Nuwaysir K, Pitas A, Richmond T, Gorski T, Berg JP, Ballin J, McCormick M, Norton J, Pollock T, Sumwalt T, Butcher L, Porter D, Molla M, Hall C, Blattner F, Sussman MR, Wallace RL, Cerrina F, Green RD. Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. *Genome Res.*, 2002, 12(11): 1749-1755.

96. Ksiazek TG, Erdman D, Goldsmith CS, Zaki SR, Peret T, Emery S, Tong S, Urbani C, Comer JA, Lim W, Rollin PE, Dowell SF, Ling AE, Humphrey CD, Shieh WJ, Guarner J, Paddock CD, Rota P, Fields B, DeRisi J, Yang JY, Cox N, Hughes JM, LeDuc JW,

Bellini WJ, Anderson LJ; SARS Working Group. A novel coronavirus associated with severe acute respiratory syndrome. *N Engl J Med.*, 2003, 348: 1953-1966.

97. Fu J, Tan BH, Yap EH, Chan YC, Tan YH. Full-length cDNA sequence of dengue type 1 virus (Singapore strain S275/90). *Virology*, 1992, 188: 953-958.

98. Hamming RW. Error Detecting and Error Correcting Codes. *Bell Syst Tech J*, 1950, 29: 147-160.

99. Sung WK, Lee WH. Fast and Accurate Probe Selection Algorithm for Large Genomes. *IEEE Computer Society Bioinformatics Conference*, 2003, 65.

100. Kane MD, Jatkoe TA, Stumpf CR, Lu J, Thomas JD, Madore SJ. Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res.*, 2000, 28(22): 4552-4557.

101. Maskos U, Southern EM. A study of oligonucleotide reassociation using large arrays of oligonucleotides synthesised on a glass support. *Nucleic Acids Res.*, 1993, 21(20): 4663-4669.

102. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 1997, 25(17): 3389-3402.

103. Ratushna VG, Weller JW, Gibas CJ. Secondary structure in the target as a confounding factor in synthetic oligomer microarray design. *BMC Genomics*, 2005, 6: 31.

104. Kullback S, Leibler RA. On information and sufficiency. *Ann Math Statist*, 1951, 22: 79-86

105. Anderson TW, Darling DA. Asymptotic theory of certain goodness of fit criteria based on stochastic processes. *Ann Math Statist*, 1952, 23: 192-212.

106. Deffernez C, Wunderli W, Thomas Y, Yerly S, Perrin L, Kaiser L. Amplicon Sequencing and Improved Detection of Human Rhinovirus in Respiratory Samples. *J Clin Microbiol*, 2004, 42: 3212-3218.

107. Steemers FJ, Gunderson KL. Illumina, Inc. *Pharmacogenomics*, 2005, 6: 777-782.

108. Kessler N, Ferraris O, Palmer K, Marsh W, Steel A. Use of the DNA Flow-Thru Chip, a Three-Dimensional Biochip, for Typing and Subtyping of Influenza Viruses. *J Clin Microbiol*, 2004, 42: 2173-2185.

109. McGlennen RC. Miniaturization Technologies for Molecular Diagnostics. *Clin Chem*., 2001, 47: 393-402.

110. Koehne JE, Chen H, Cassell AM, Ye Q, Han J, Meyyappan M, Li J. Miniaturized Multiplex Label-Free Electronic Chip for Rapid Nucleic Acid Analysis Based on Carbon Nanotube Nanoelectrode Arrays. *Clin Chem*., 2004, 50: 1886-1893.

111. ington JA, Shah NA, Chen X, Janis M, Liu C, Kondapalli S, Reyes V, Savage MP, Zhang Z, Watts R, DeGuzman M, Berno A, Snyder J, Baid J. New developments in high-throughput resequencing and variation detection using high density microarrays. *Hum Mutat.*, 2002, 19: 402–409.

112. Kingsford C, Nagarajan N, Salzberg SL. 2009 Swine-Origin Influenza A (H1N1) Resembles Previous Influenza Isolates. *PLoS ONE*, 2009, 4(7): e6402.

113. Wang Z, Daum LT, Vora GJ, Metzgar D, Walter EA, Canas LC, Malanoski AP, Lin B, Stenger DA. Identifying Influenza Viruses with Resequencing Microarrays. *Emerg. Infect. Dis.*, 2006, 12: 638–646.

114. Lin B, Malanoski AP, Wang Z, Blaney KM, Long NC, Meador CE, Metzgar D, Myers CA, Yingst SL, Monteville MR, Saad MD, Schnur JM, Tibbetts C, Stenger DA. Universal Detection and Identification of Avian Influenza Virus by Use of Resequencing Microarrays. *Journal of Clinical Microbiology*, 2009, 47(4): 988-993.

115. Toh K. Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics*, 2008, 9: 286-298.

116. Maurer-Stroh S, Ma J, Lee RT, Sirota FL, Eisenhaber F. Mapping the sequence mutations of the 2009 H1N1 influenza A virus neuraminidase relative to drug and antibody binding sites. *Biol Direct.*, 2009, 4: 18.

117. Seringhaus M, Rozowsky J, Royce T, Nagalakshmi U, Jee J, Snyder M, Gerstein M. Mismatch oligonucleotides in human and yeast: guidelines for probe design on tiling microarrays. *BMC Genomics*, 2008, 9: 635.

118. Flannery WH, Teukolsky SA, Vertterling WT. Kolmogorov-Smirnov test. *Numerical Recipes in FORTRAN: The Art of Scientific Computing, 2nd ed.*, 1992, 617–620.

119. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*, 1980, 16: 111–120.

120. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 1994, 22(22): 4673-4680.

121. Sturmer M, Doerr W, Preiser W. Variety of interpretation systems for human immunodeficiency virus type 1 genotyping: Confirmatory information or additional confusion? *Curr. Drug Targets Infect. Disord.*, 2003, 3: 373–382.

122. Leitner T, Escanilla D, Marquina S, Wahlberg J, Broström C, Hansson HB, Uhlén M, Albert J. Biological and molecular characterization of subtype D, G, and A/D recombinant HIV-1 transmissions in Sweden. *Virology*, 1995, 209: 136–146.

123. Holmes EC. RNA virus genomics: a world of possibilities. *J Clin Invest*., 2009, 119(9): 2488–2495.

124. Lee WH, Wong CW, Leong WY, Miller LD, Sung WK. LOMA: a fast method to generate efficient tagged-random primers despite amplification bias of random PCR on pathogens. *BMC Bioinformatics*, 2008, 9: 368.

125. Wong CW, Lee WH, Leong WY, Soh SW, Kartasasmita CB, Simoes EA, Hibberd ML, Sung WK, Miller LD. Optimization and clinical validation of a pathogen detection microarray. *Genome Biol.*, 2007, 8(5): R93.

126. Lee WH, Koh CW, Chan YS, Aw PPK, Loh KH, Han BL, Thien PL, Nai GYW, ML Hibberd, CW Wong, WK Sung. Large Scale Evolutionary Surveillance of the 2009 H1N1 Influenza A Virus Using Resequencing Arrays. *Nucleic Acids Research*, 2010.

127. Lee VJ, Yap J, Cook AR, Chen MI, Tay J, Tan BH, Loh JP, Chew SW, Koh WH, Lin R, Cui L, Lee WH, Sung WK, Wong CW, Hibberd ML, Kang WL, Seet B, Tambyah PA. Oseltamivir ring prophylaxis for containment of Influenza A (H1N1-2009) outbreaks. *New England Journal of Medicine*, 2010.

128. Lee WH, WK Sung. RB-Finder: An Improved Distance-based Sliding Window Method to Detect Recombination Breakpoints. *RECOMB*, 2007.

129. Lee WH, Sung WK. RB-finder: an improved distance-based sliding window method to detect recombination breakpoints. *J Comput Biol*., 2008, 15(7): 881-898.

130. Wang L, Jiang T. On the complexity of multiple sequence alignment. *J Comput Biol*., 1994, 1(4): 337-348.

131. Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, 2000, 302(1): 205-217.

132. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 2004, 32(5): 1792-1797.

133. Wertz JE, McGregor KF, Bessen DE. Detecting Key Structural Features within Highly Recombined Genes. *PLoS Comput Biol*., 2007, 3(1): e14.