# COMPUTATIONAL IDENTIFICATION OF VIRAL/BACTERIAL EPITOPES FOR TYPE AND STRAIN CHARACTERIZATION, LEADING TO THE DEVELOPMENT OF VACCINES AGAINST HUMAN ADENOVIRUS, HIV, AND STAPHYLOCOCCUS AUREUS

by

Kalpana Dommaraju

A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Bioinformatics and Computational Biology

Committee:

_____ Dr. Donald Seto, Committee Chair

_____ Dr. Iosif Vaisman, Committee Member

_____ Dr. Patrick Gillevet, Committee Member

_____ Dr. Iosif Vaisman, Director, School of Systems Biology

_____ Dr. Donna M. Fox, Associate Dean, Office of Student Affairs & Special Programs, College of Science

_____ Dr. Miralles-Wilhelm, Dean, College of Science

Date: _____ Spring Semester 2021
George Mason University
Fairfax, VA

# Computational Identification of Viral/Bacterial Epitopes for Type and Strain Characterization, Leading to the Development of Vaccines against Human Adenovirus, HIV, and Staphylococcus aureus

A Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

by

Kalpana Dommaraju
Master of Science
George Mason University, 2006
Bachelor of Science
A.N.G.R.A. University, India 2001

Director: Donald Seto, Professor
Microbial Genomics and Diversity Bioinformatics

Spring Semester 2021
George Mason University
Fairfax, VA


Copyright 2021 Kalpana Dommaraju
All Rights Reserved

## DEDICATION

This work is dedicated to my loving parents, Subbamma and Obularaju, my husband Rayudu and my kids, Hasini, Aditya and my brother Rajesh.

## ACKNOWLEDGMENTS

I would like to express my special thanks and appreciation to my dissertation director, Dr. Donald Seto; you have been the greatest support for me. I would like to thank you for encouraging my research for all these years. Your advice is the greatest help in my research and career as well.

I would like to thank my dissertation committee members, Dr. Iosif Vaisman and Dr. Patrick Gillevet, for their suggestions, support, and guidance to the right direction.

Special thanks to Kimberly Harris and Diane St. Germain for their administrative support, and Chris Ryan for system support.

I sincerely appreciate the encouragement and support given by my family and friends, Shoaleh Dehghan, June King, Dakshaini, Geeta, Dinesh, Siddu, Anju, Prakash, Deepthi, Rani, Sirisha, Sridhar, Naresh, Chandana and Kumar.

This dream of mine would not be completed without the love and support of my parents and my family. Thank you, Hasini and Aditya, for travelling with me in this journey. Your unconditional love and encouragement brought me here. Thanks to my family, Rajesh and Jahnavi, Lakshit, Parnika, Subbamma, Venkata Raju, and my grandparents.

I would like to thank God for giving me strength and giving me the best friends, family, and mentors who helped me throughout this process.

**TABLE OF CONTENTS**

## LIST OF TABLES

## LIST OF FIGURES

## LIST OF ABBREVIATIONS

AA ..... Amino Acid
AdV ..... Adenovirus
AIDS ..... Acquired Immune Deficiency Syndrome
BLAST ..... Basic Local Alignment Search Tool
CAR ..... Coxsackie and Adenovirus Receptor
CDC ..... Centers for Disease Control and Prevention
DNA ..... Deoxyribonucleic Acid
EKC ..... Epidemic Keratoconjunctivitis
HAI ..... Hospital Associated Infection
HAdV ..... Human Adeno Virus
HIV ..... Human Immunodeficiency Virus
HVR ..... Hyper Variable Region
IV ..... Human Immunodeficiency Virus
IBA ..... Identity Based Analysis
ITR ..... Inverted Terminal Repeats
MEGA ..... Molecular Evolutionary Genetics Analysis
MHRP ..... Military HIV Research Program
MLEE ..... Multilocus Enzyme Electrophoresis
MLST ..... Multi-Locus Sequence Typing
MRSA ..... Methicillin Resistant *S. aureus*
MSA ..... Multiple Sequence Alignment
NA ..... Nucleic Acid
NIAID ..... National Institute of Allergy and Infectious Diseases
NCBI ..... National Center for Biotechnology Information
ORF ..... Open Reading Frame
PBA ..... Phylogenetic Based Analysis
PD ..... Percent Divergence
PDS ..... Pairwise Divergent Score
PI ..... Percent Identity
PHF ..... Penton base, Hexon, and Fiber
RNA ..... Ribonucleic Acid
SBA ..... Similarity Based Analysis
SBT ..... Sequence-Based Typing
SSOP ..... Sequence-Specific Oligonucleotide Probe
VA ..... Virus-Associated

VNTR.....Variable Number Tandem Repeats

## ABSTRACT

COMPUTATIONAL IDENTIFICATION OF VIRAL/BACTERIAL EPITOPES FOR TYPE AND STRAIN CHARACTERIZATION, LEADING TO THE DEVELOPMENT OF VACCINES AGAINST HUMAN ADENOVIRUS, HIV, AND STAPHYLOCOCCUS AUREUS

Kalpana Dommaraju, Ph.D.

George Mason University, 2021

Dissertation Director: Dr. Donald Seto, Professor

Identifying newly isolated adenovirus is important to the understanding of pathogens and molecular evolution. With the advancement of sequencing technology, virus sequences are accumulated. New genotypes emerge due to recombination in capsid proteins, penton base, hexon, and fiber. A rapid identification of newly emerging genotypes became necessary for current HAdV scientists, and a software tool was developed to accurately identify and characterize the genotypes based on recombination. The genotyping tool characterizes newly emergent adenoviruses by identifying the type-specific epitopes and complementing them with the referencing database. The genotyping outcomes are based on BLAST results, phylogenetic trees, and sequence identity with the pre-existing epitopes. The performance of the genotyping was tested on 52 human

adenovirus genotype sequences, resulting in 100% sensitivity and specificity. Genotyping enables quick and accurate detection of novel strains.

The modest protection afforded by the RV144 vaccine offers an opportunity to evaluate its mechanisms of protection. Differences between HIV-1 breakthrough viruses from vaccine and placebo recipients can be attributed to the RV144 vaccine, as this was a randomized and double-blinded trial. CD8 and CD4 T-cell epitope repertoires were predicted in HIV-1 proteomes from 110 RV144 participants. Predicted Gag epitope repertoires were smaller in vaccine recipients than in placebo recipients ($p = 0.019$). After comparing participant-derived epitopes to corresponding epitopes in the RV144 vaccine, the proportion of epitopes that could be matched differed depending on the protein conservation (only 36% of epitopes in Env vs 84%–91% in Gag/Pol/Nef for CD8 predicted epitopes) or on vaccine insert subtype (55% against CRF01_AE vs 7% against subtype B). To compare predicted epitopes to the vaccine, we analyzed predicted binding affinity and evolutionary distance measurements. Comparisons between the vaccine and placebo arm did not reveal robust evidence for a T-cell driven sieve effect, although some differences were noted in Env-V2 (0.022#p-value#0.231). The paucity of CD8 T-cell responses identified following RV144 vaccination, with no evidence for V2 specificity, considered together both with the association of decreased infection risk in RV 144 participants with V-specific antibody responses and a V2 sieve effect, led us to hypothesize that this sieve effect was not T-cell-specific. Overall, our results did not reveal a strong differential impact of vaccine-induced T-cell responses among breakthrough infections in RV144 participants.

*Staphylococcus aureus* is the leading cause of skin and soft tissue infections worldwide. Strain identification is critical for understanding the epidemiology of this pathogen. The highly variable *spa* gene provides a sensitive method for distinguishing *S. aureus* isolates. The *spa* gene product is a surface antigen with Ig-binding activity that is involved in immune evasion. *Spa* includes a variable Xr region that has multiple repeats, usually 8 amino acids in length. Repeats within an Xr region may have different amino acid sequences. Over 19,817 variants of the *spa* gene are known to date. Combined with other methods such as multilocus sequence typing, *spa* typing can provide high resolution strain identification. This is an open-source *spa* typing program that: requires no helper applications; it can be integrated into bioinformatics pipelines and can analyze complete draft genome sequences.

# CHAPTER 1

# INTRODUCTION TO HUMAN ADENOVIRUS

Human adenoviruses (HAdVs) were first isolated from a child's adenoid tissue in 1953 (Rowe et al., 1953) as well as from an U. S. Army recruit as a respiratory infection (Hilleman et al., 1954). Since the identification of adenoviruses (AdVs), numerous members of the *Adenoviridae* family have been characterized across different host species. Most likely, all vertebrates are affected by AdVs, including fish, frogs, snakes, birds, canines, and primates (for example, chimpanzee and human) (Russell & Benkö, 1999). Human adenoviruses can cause respiratory diseases, gastroenteritis, acute febrile pharyngitis, pharyngoconjunctival fever, acute respiratory disease, pneumonia, keratoconjunctivitis, pertussis-like syndrome, acute hemorrhagic cystitis, meningoencephalitis, and hepatitis (Jones et al., 2007). Species HAdV-B, HAdV-E, and HAdV-C are associated with respiratory diseases. Species HAdV-B, HAdV-E, and HAdV-D are responsible for ocular diseases. Serotypes HAdV-F40 and HAdV-F41 cause gastroenteritis, although recent studies have shown the HAdV-D genotype also causes gastroenteritis (Liu et al., 2011; Liu et al., 2012). HAdV-E4 is primarily responsible for acute respiratory disease, but also causes ocular disease. HAdV-E4 is also one of the two HAdVs for which a vaccine has been developed (Jones et al., 2007), reflecting its importance as a human respiratory pathogen (Table 1).

**Table 1**

*Taxonomy of Human Adenoviruses and Their Cell Tropism*

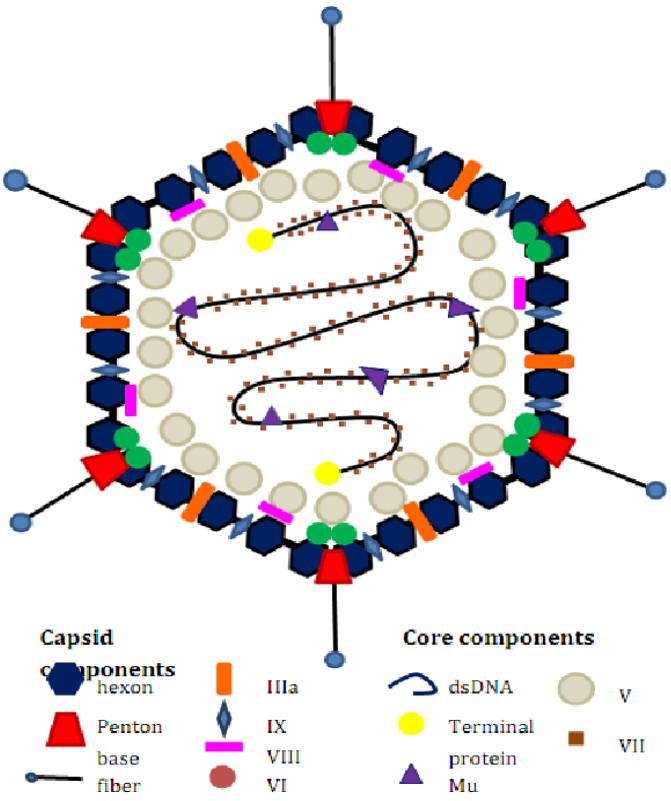
| Species | Genotypes | Site of Infection |
|---|---|---|
| A | 12, 18, 31, 61 | Gastrointestinal tract |
| B | 3, 7, 11, 14, 16, 21, 34, 35, 50, 55, 66, 68, 76, 77, 78, 79 | Respiratory tract and/or ocular tract, Urogenital system |
| C | 1, 2, 5, 6, 57, 89 | Respiratory tract |
| D | 8, 9, 10, 13, 15, 17, 19, 20, 22, 23, 24, 25, 26, 27, 28, 29, 30,32, 33, 36, 37, 38, 39, 42, 43, 44, 45, 46, 47, 48, 49, 51, 53, 54, 56, 58, 59, 60, 62, 63, 64, 65, 67, 69, 70, 71, 72, 73, 74, 75, 80, 81,82, 83, 84, 85, 86, 87, 88, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103 | Ocular tract (Keratoconjunctivitis), Gastrointestinal tract |
| E | 4 | Respiratory tract, ocular tract |
| F | 40, 41 | Gastrointestinal tract |
| G | 52 | Gastrointestinal tract |

Note: Human adenovirus genotypes 1 to 103 are classified based on genomic and bioinformatic approaches.

## 1.1 Morphology

The adenoviruses are non-enveloped icosahedral viruses comprising a linear double-stranded DNA genome surrounded by a protein capsid. The shell is 70–100 nm in diameter and is made up of 252 capsomeres, out of which 240 homotrimeric hexons make up the faces and 12 penton bases on the edges that have protruding trimerized fiber proteins that include a knob region (Figure 1) (Waye & Sing, 2010). The knob is

responsible for cell recognition and entry. Hexon proteins contain type-specific serum neutralization epitopes, and the fiber knob contains hemagglutination epitopes (Hierholzer, 1992). These proteins define individual virus genotypes. The capsid core is made of two major proteins (polypeptide V and polypeptide VII) and a minor arginine-rich protein. A 55 kDa protein is covalently attached to the 5'ends of the DNA.




Note: The penton base and fiber proteins function to recognize and allow penetration into virus while hexon maintains capsid structural stability. Adapted from Mary Miu Yee Waye and Chor Wing Sing (2010), "Anti-Viral Drugs for Human Adenoviruses," *Pharmaceuticals 3*(10), 3343–3354.

**Figure 1: Structure of a Human Adenovirus**

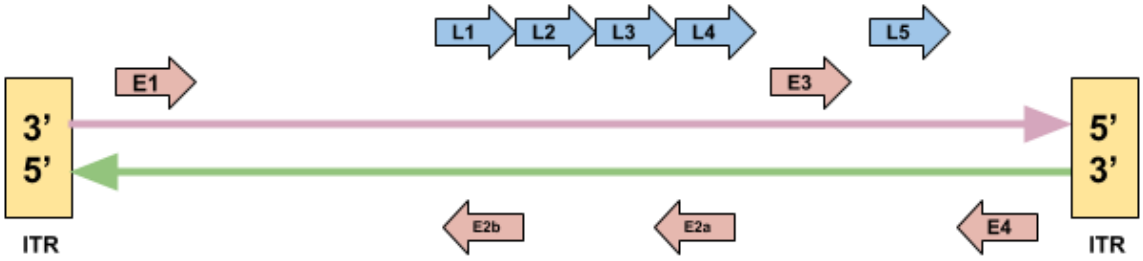
## 1.2 Life Cycle

An adenovirus attaches to the cell surface protein receptor through its fiber knob to a Coxsackie and Adenovirus Receptor (CAR) on a host T-cell. Attachment is followed by interaction of the penton base with cellular integrin that promotes receptor-mediated internalization. The virus is uncoated due to high pH within the host T-cell and the DNA is transported into the nucleus. Early transcription genes undergo translation to produce 20 early proteins, which induce the host T-cell to create favorable conditions for further viral replication. The viral encoded protein at the 5'end initiates the DNA synthesis. Synthesized DNA is spliced into 18 mRNA fragments and are transported to cytoplasm. Translation processes result in viral structural proteins that are transported into the nucleus, where morphogenesis occurs. Viral DNA is packed into capsid and released from the host T-cell by budding. The virus is transmitted from host to host through aerosol droplets, faeco-oral route, contaminated fingers, and contaminated fomites.

## 1.3 Genome

The human adenovirus (HAdV) genome is a linear double-stranded DNA molecule of approximately 36 kb, containing flanking ~100 bp inverted terminal repeats (ITR) (Figure 1 and Figure 2). These sequences contain the origins of viral replication. ITRs are followed by the viral packaging sequences at the left end of the genome, which are used in DNA encapsulation. The terminal proteins help in the initiation of viral DNA replication. Viral capsid is composed of three major proteins (II, III, and IV) and five minor proteins (IIIa, IVa2, VI, VIII, and IX) (see Figure 1). Proteins V, VII, and X condense the DNA and mediate core and capsid interactions. Protease is required for the

maturation of the assembled infectious virus. The adenovirus genome encodes ~40 proteins, which are classified as early or late expressing. Early proteins are expressed before DNA replication, and late proteins are expressed after replication. The genome contains early transcription units (E1-E4) and late regions (L1-L5), shown in Figure 2. The adenovirus major late promoter directs the synthesis of late pre-mRNA that id alternatively spliced to generate late mRNAs to produce late proteins. Two additional small late transcripts produce virus-associated (VA) RNAs (RNAI and RNAII). RNAI plays a role to stimulate the early and late genes, such as E3 and hexon (Svensson & Akusjärvi, 1984). Due to the controlled progression of gene expression, adenoviruses are good vectors for gene therapy.




Note: The genome is organized into sets of early proteins (E1, E2a, E2b, E3, and E4), late proteins (L1, L2, L3, L4, and L5), and inverted terminal repeats (ITR) at the ends.

**Figure 2: Transcription map of the human adenovirus genome**

## 1.4 Classification

Human adenoviruses are classified under the genus Mastadenovirus. There are more than 103 HAdV genotypes (HAdV Working Group, 2019) accepted, based on

genomic data and characterization, and are classified into species HAdV-A to HAdV-G (Table 2). Traditionally, HAdVs were classified based on haemagglutination and serum neutralization, cell culture, and pathogenicity.

**Table 2**

*GC% Content of HAdVs Species A to G*

| HAdV | Min-GC% | Max-GC% |
|---|---|---|
| HAdV-A | 46.37 | 46.52 |
| HAdV-B | 48.78 | 51.28 |
| HAdV-C | 55.19 | 55.36 |
| HAdV-D | 54.59 | 57.48 |
| HAdV-E | 57.67 | 57.67 |
| HAdV-F | 50.95 | 51.22 |
| HAdV-G | 55.11 | 55.11 |

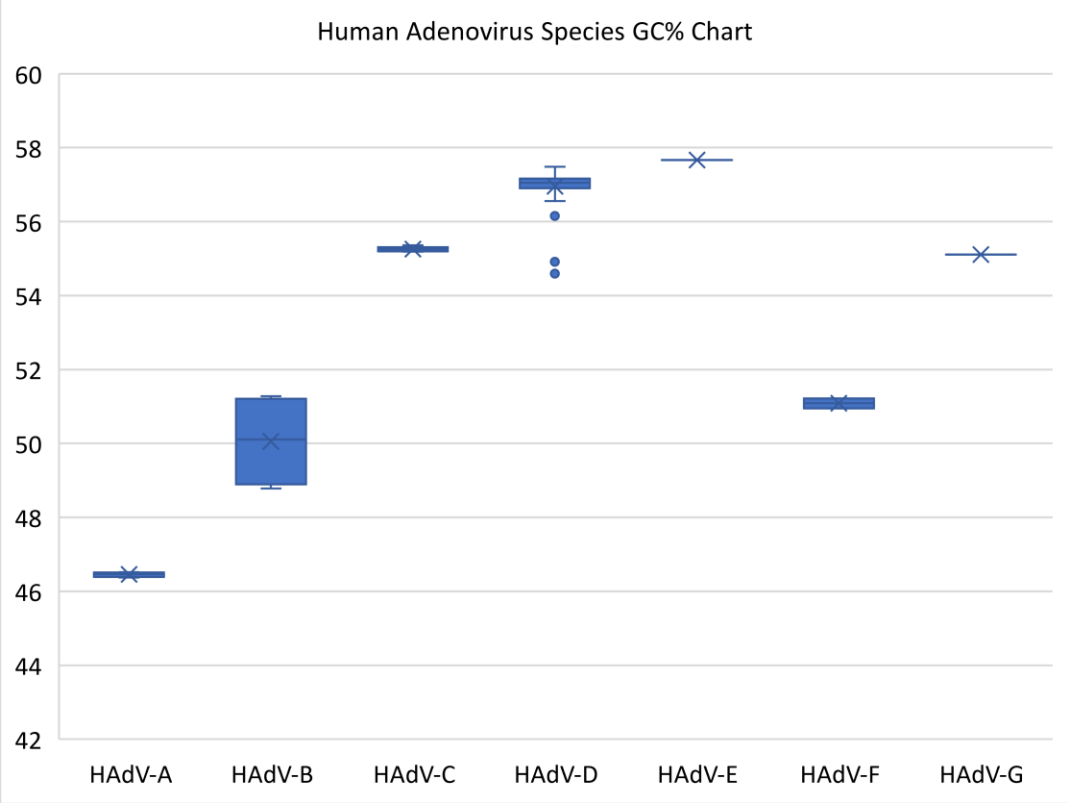
Note: Species A has lowest GC% content while species D and E have the highest.

Serological tests recognize the hexon ε (epsilon) epitope for virus neutralization, and haemagglutination assays identify the *γ* (gamma) epitope on the fiber (Rosen et al., 1960). These two serological assays must be run simultaneously to achieve acceptable HAdV serotyping. Serological assays are difficult and tedious to perform, requiring a complete set of viruses, antisera, and various red blood cells (Crawford-Miksza & Schnurr, 1994; Shenk et al., 2003), all of which are not readily available to the research community, and may have lot-to-lot variability. For typing novel adenoviruses, exact titrations of all adenovirus genotypes and their antisera with the novel adenovirus are

required, which is not practical. Also, these traditional tests analyze only two very small parts of the genome, so their results may be inaccurate, particularly in the identification of recombinant HAdVs (Wigand et al., 1982. This was proven to be the case in a neonate pneumonia case reported in France, as the virus also caused highly contagious keratoconjunctivitis in the health care providers (Henquell et al., 2009). Another mischaracterization by using the serum neutralization assay was demonstrated by the analysis of an emergent respiratory pathogen, HAdV-B55 (Walsh et al., 2010). This was shown to be a recombinant of HAdV-B11 (a renal pathogen) and HAdV-B14 (a respiratory pathogen), and reported as "HAdV-B11a," with questions of tropisms. Genome analysis showed approximately 97% of the genome was from a parental HAdV-B14. Other regions of the genome contribute to virus pathology. For example, recent research indicated that genomic regions such as E3, E4, penton base, and fiber showed better correlation of virus biology and pathology (for example in viral tropism and virulence) than hexon (Robinson et al., 2013; Walsh et al., 2009).

An enormous amount of adenovirus data has been generated from genome sequencing. As it is impossible to study these sequences base by base, broad overviews were examined within these strains. Quantitative measures include GC content (Figure 3) and genome percent identity to support species classification and characterization. Phylogenetic analysis is the most common way to examine human adenovirus identities and relationships (Walsh et al., 2009). Significant biological and pathogenic changes (Walsh et al., 2009), along with genome-based paradigms and algorithms help aid in the

analysis and characterization of the whole genome (Seto et al., 2011), specifically, the penton base, hexon variable loop regions, and fiber knob proteins.




Note: Graphical representation of the GC% range of HAdV species A to G. GC% is one of the species defining criteria for human adenovirus. Since there are overlaps in between C and G, B and F, other computational tools are needed to accurately identify the species.

**Figure 3: HAdVs Species A to G GC% Range Chart**

## 1.5 Recombination

Organisms, through their genomes, undergo changes through different mechanisms, including sequence recombination, insertion, and deletions. These provide

for evolution and may contribute to fitness and adaptation. Recombination is a mechanism that human adenoviruses commonly utilize (Kajon et al., 2010; C. M. Robinson et al., 2011b; O. J. Robinson et al., 2009, 2013b; Yang et al., 2009). In particular, human adenoviruses have been observed to contain many instances of recombination among the penton base, hexon, and fiber genes (Grodzicker et al., 1974; Mautner et al., 1984; Williams et al., 1975). The recombination results in inserting a pathogenic gene into a nonpathogenic genotype (C. M. Robinson, 2013; O. J. Robinson et al., 2011). One example of this recombination-driven evolution is HAdV-B55. The hexon in HAdV-B55 is identical to that from the renal pathogen HAdV-B11, and it is serologically typed as such. The hemagglutination reaction, however, recognizes it as a respiratory pathogen, HAdV-B14. Serologically, it would appear in hosts as a renal pathogen, but would cause respiratory diseases. Genome analysis shows this novel respiratory pathogen was a recombinant of HAdV-B14 and HAdV-B11 (Liu et al., 2014; Walsh et al., 2010; Yang et al., 2009). Another example of the importance of recombinants is HAdV-D53. HAdV-D53 was classified originally as HAdV-D22 when it first appeared. HAdV-D22 is a nonpathogenic strain, whereas HAdV-D53 is a highly contagious epidemic keratoconjunctivitis (EKC) agent. HAdV-D53 has undergone multiple recombinations (Engelmann et al., 2006). HAdV-D53 is considered a novel strain based on the genomic analysis (Walsh et al., 2009). As noted by the pathogenic differences, it is insufficient to name HAdV-D53 as a variant of HAdV-D22, due to its serological profile. Another similar event occurred in HAdV-D64. It was classified originally as serotype HAdV-D19 by serology. Like HAdV-D22, HAdV-D19 is a non-

pathogenic virus. However, HAdV-D64 is a highly contagious EKC agent, resulting from the recombination among HAdV-D19p, HAdV-D22, and HAdV-D37. HAdV-D19 and HAdV-D64 have 100% sequence identity. HAdV-D19 had a lack of association with the keratitis and is not infectious for corneal epithelial cells. HAdV-D64 was later reclassified due to its unique pathogenicity (Zhou et al., 2012).

### 1.6 Zoonosis

Nonhuman primates are a potential source for emerging transmission of zoonotic disease to humans (Wolfe et al., 2007; Gillespie et al., 2008). Human and nonhuman primates share similar physiological and genetic properties. This close phylogenetic proximity is favored for interspecies transmissions of pathogens between human and nonhuman primates (Leendertz et al., 2006). Jones et al. (2008) have estimated that 75% of the infectious diseases with severe implications for public health has a zoonotic origin. Diseases such as HIV and Ebola hemorrhagic fever have been linked with different primate species (Palmarini et al., 2007). A possible horizontal transmission event was also detected in human adenovirus between nonhuman primates and humans (Wevers et al., 2011).

Studying and understanding zoonosis could provide important information for possible contagious diseases in humans and help set up surveillance for epidemic outbreaks (Gillespie et al., 2008). Zoonosis may also have an interesting implication in the use of simian adenovirus as a vector for human gene therapy and vaccines. Simian adenovirus proteins and human immune responses apparently have no cross-reactivity (Roy et al., 2004). This implies that simian adenovirus-based vectors are safe for humans

to use. However, if zoonosis occurs, there may be nonhuman simian adenoviruses in the human population, which complicates cross-reactivity.

### *1.6.1 HAdV-E4 - Zoonosis Example*

HAdV-E4 is a major human respiratory pathogen with implications for epidemic outbreaks globally. HAdV-E4 and HAdV-B7 are responsible for about 60%–80% of the acute respiratory disease cases in U. S. military training camps. They are the only HAdVs for which vaccines have been developed and applied. Interestingly, HAdV-E4 is the sole HAdV member in the species E. This species contains several chimpanzee adenoviruses that are highly similar to one another. For example, it shares a 97% sequence similarity with SAdV-E26, which is much higher than with any HAdVs. That is, genome analysis of HAdV-E4 showed a closer phylogenetic relationship with the chimpanzee adenoviruses rather than with human adenoviruses (Dehghan et al., 2013). This suggests the possibility of HAdV-E4 arising through zoonosis from chimpanzee to human through a series of adaptations.

## 1.7 Human Adenovirus Working Group

The Human Adenovirus (HAdV) Working Group at George Mason University is a collaboration of adenovirus researchers with the mission to refine and standardize the typing criteria. The new genome sequence information and the detailed working group contacts are posted on the HAdV Working Group website (http://hadvwg.gmu.edu/). The current parameters and criteria used to identify novel sequences can be located on the *Submit "candidate" HAdV* page (http://hadvwg.gmu.edu/index.php/submit-candidate-hadv/). All novel candidate strains need to be submitted to the correct email addresses

(provided on the *Submit* page) for review. Once the approval of the candidate strain has been passed unanimously in favor, a letter will be sent to the requester and the new strain will be announced on the *Home* page (Table 1). The human adenovirus working group helps standardize the process of identifying and acknowledging novel adenoviruses. The grouping and typing criteria are part of an ongoing project that helps develop and refine adenovirus identification.

Originally, human adenoviruses were of 52 different serotypes (Shenk et al., 2001), which were characterized by traditional serological methods. Types are now characterized by genomic analysis and are called genotypes, including the 52 serotypes. There are two current criteria used for identifying a novel type. The first criterion is that to be considered a novel type, the hexon or fiber gene must be a new "serological" type, or the sequence presumably contributing to the serology must be "new" or different. A more recent criterion is that the candidate type must be a new "recombinant" of hexon, penton base, or fiber genes. The typing of novel candidates also includes differences in the penton base (Oostrum et al., 1985; Biere & Schweiger, 2010; Ebner et al., 2005a, 2005b; Gall et al., 1998). The penton base is an important marker; modifying penton base RGD or HVR1 sequences will change the viruses' pathogenicity and cell type recognition. Therefore, the penton base gene is an important consideration used to identifying a novel adenovirus type.

### 1.8 Proposal

Advancements in DNA sequencing technology have allowed human adenovirus genomes to be readily sequenced and used to develop a standardized and consistent

bioinformatics approach to identify, characterize, and genotype newly emergent pathogens through sequence analysis of the whole genome along with the major capsid proteins (hexon, penton base, and fiber). These capsid proteins contain useful defining-specific variable regions appropriate for typing. For example, the hexon protein comprises highly variable regions HVRs1-6 in loop 1 and HVR 7 in loop 2 (Crawford-Miksza & Schnurr, 1996) followed by conserved region in its sequence. The penton base has two highly variable regions, HVR1 and RGD. Similarly, fiber has a highly conserved shaft, but highly variable knob regions. Considering these variable regions, a three-step bioinformatics analysis is proposed and designed for a thorough and exact characterization, typing, and classification of human adenovirus. The flow of genotyping consists of Similarity Based Analysis (SBA), Percent Identity Based Analysis (IBA), and Phylogenetic Based Analysis (PBA). Currently there are 103 genotypes recognized: 4 HAdV-A, 16 HAdV-B, 6 HAdV-C, >71 HAdV-D, 1 HAdV-E, 2 HAdV-F, and 1 HAdV-G (Table 1).

# CHAPTER 2

# GENOTYPING TOOL FOR CHARACTERIZING HUMAN ADENOVIRUS

## 2.1 Similarity Based Analysis (SBA)

Identifying newly isolated adenovirus is important to the understanding of pathogens and molecular evolution. With the advent of viruses being sequenced and accumulated, a software tool is developed for the rapid identification of types based on the epitopes of hexon, penton base, and fiber. Genotyping identification tool is a semi-automatic system used to type and characterize newly emergent adenoviruses by identifying the type-specific epitopes and complementing with the referencing database. The Genotyping outcomes are based on the BLAST results, phylogenetic trees, and sequence identity with the pre-existing epitopes. The performance of the Genotyping was tested on 53 human adenovirus genotype sequences, resulting in 100% sensitivity and specificity. Genotyping enables quick and accurate detection of novel strains.

### *2.1.1 Introduction*

Various genotyping servers employ sequence SBA to characterize the virus. Examples include HepSEQ (Gnaneshan et al., 2007) for Hepatitis B virus; BioAfrica for Human Immunodeficiency Virus (HIV); and HIV STAR (Myers et al., 2005). This is a paradigm change from serology-based typing methods to more exact genome sequence-based typing. The human adenovirus genome is approximately 35kb in length and

encodes for major proteins that are used for genotyping: penton base, hexon, and fiber. The genotyping of this virus is based on sequence similarities and variations in major capsid proteins as recommended (Seto et al., 2011).

### *2.1.2 Material and Methods*

#### 2.1.2.1 Compilation of Reference Dataset.

##### *2.1.2.1.1 Sequence Extraction.* Genbank

(https://www.ncbi.nlm.nih.gov/nucleotide/) has 361 entries for "human adenovirus genomes" currently. Genotypes 1 to 52 are considered as a reference dataset, as originally characterized serotypes. In this project, a pipeline has been developed to automate the sequence extraction from Genbank. FASTA format whole genomes for species A, B, C, D, E, F, and G are extracted from Genbank though the automated application (Note: Accession numbers are listed in a text file. Application reads the accession numbers from an input text file and creates a URL to query GenBank and download the files in GenBank and FASTA format.

Figure 4). Coding regions/CDs of capsid proteins penton base, hexon, and fiber nucleic acid (NA) sequences and amino acid (AA) FASTA format sequences are downloaded and saved to local folders.

```
public static void getAccNum(String ipFile, String gbOPDir, String fastaOPDir){
		try{
			FileReader fr = new FileReader(ipFile);
			BufferedReader br = new BufferedReader(fr);
			String s1 = "";
			while((s1 = br.readLine()) != null){
				String accNum = s1.trim();
				System.out.println("Accession: " + accNum);
				String urlGB = "https://eutils.ncbi.nlm.nih.gov/entrez/eu-
				tils/efetch.fcgi?db=nuccore&amp;id=" + accNum.trim() +
				"&amp;rettype=gb&amp;retmode=text";
				String urlFasta = "https://eutils.ncbi.nlm.nih.gov/en-
				trez/eutils/efetch.fcgi?db=nuccore&amp;id=" + ac-
				cNum.trim() + "&amp;rettype=fasta&amp;retmode=text";
				System.out.println(urlGB);
				String fileDirGB = gbOPDir + accNum + ".gb";
				String fileDirFasta = fastaOPDir + accNum + ".fasta";
				download(urlGB , fileDirGB);
				download(urlFasta , fileDirFasta);}
			br.close();  }
		catch (IOException e){
			System.out.println(e.getMessage());}}
public static void download(String address, String localFileName) {
	OutputStream out = null;URLConnection conn = null;InputStream  in = null;
	try{
		URL url = new URL(address);
		out = new BufferedOutputStream(new FileOutputStream(localFileName));
		conn = url.openConnection();
		in = conn.getInputStream();
		byte[] buffer = new byte[1024];
		int numRead;
		while ((numRead = in.read(buffer)) != -1){
			out.write(buffer, 0, numRead);}}
		catch (Exception exception){
			exception.printStackTrace();}
		finally{
			try{
				if (in != null)
					in.close();
				if (out != null)
					out.close(); }
			catch (IOException ioe){}}}}
```

Note: Accession numbers are listed in a text file. Application reads the accession numbers from an input text file and creates a URL to query GenBank and download the files in GenBank and FASTA format.

**Figure 4. Java Code to Download GenBank Sequences**

HAdV 1-52 genotypes with accession numbers: HAdV-C1 (AF534906), HAdV-C2 (AC_000007), HAdV-B3 (AY599834), HAdV-4 (AY594253), HAdV-C5 (AC_000008),

HAdV-C6 (FJ349096), HAdV-B7 (AY594255), HAdV-D8 (AB448767), HAdV-D9 (AJ854486), HAdV-D10 (JN226746), HAdV-D11 (AY163756), HAdV-A12 (AC_000005), HAdV-D13 (JN226747), HAdV-B14 (AY803294), HAdV-D15 (AB562586), HAdV-B16 (AY601636), HAdV-D17 (AC_000006), HAdV-A18 (GU191019), HAdV-D19 (AB448771), HAdV-D20 (JN226749), HAdV-B21 (AY601633), HAdV-D22 (FJ404771), HAdV-D23 (JN226750), HAdV-D24 (JN226751), HAdV-D25 (JN226752), HAdV-D26 (EF153474), HAdV-D27 (JN226753), HAdV-D28 (FJ824826), HAdV-D29 (JN226754), HAdV-D30 (JN226755), HAdV-A31 (AM749299), HAdV-D32 (JN226756), HAdV-D33 (JN226758), HAdV-B34 (AY737797), HAdV-B35 (AY271307), HAdV-D36 (GQ384080), HAdV-D37 (DQ900900), HAdV-D38 (JN226759), HAdV-D39 (JN226760), HAdV-F40 (NC_001454), HAdV-F41 (DQ315364), HAdV-D42 (JN226761), HAdV-D43 (JN226762), HAdV-D44 (JN226763), HAdV-D45 (JN226764), HAdV-D46 (AY875648), HAdV-D47 (JN226757), HAdV-D48 (EF153473), HAdV-D49 (DQ393829), HAdV-B50 (AY737798), HAdV-D51 (JN226765), HAdV-G52 (DQ923122).

***2.1.2.1.2 Sequence Alignment.*** The downloaded whole genome FASTA files are combined to create a multi-FASTA file for further analysis. Similarly, penton base, hexon, and fiber nucleotide and amino acid sequence multi-FASTA files are generated to perform the multiple sequence alignments. The ClustalW algorithm in Molecular Evolutionary Genetics Analysis (MEGA) version 7.0.14 (Kumar et al., 2016) is used for sequence alignment and phylogenetic analysis.

BLOSUM matrix is used for nucleotide and PAM matrix is used for amino acid sequence alignments. For pairwise alignment, the gap opening penalty is 10 and gap extension penalty is 0.1, and for multiple sequence alignment the gap opening penalty is 10 and gap extension penalty is 0.2.

***2.1.2.1.3 Variable Regions of PHF.*** A multiple sequence alignment of hexon sequences is used to determine the variable and conserved portions of the sequences based on the primers as listed in Table 1 (see Figure 5). This is useful for characterizing and distinguishing between different adenoviruses.

**Table 1. Hexon primers**

Hyper-variable regions of hexon, Loop1 (L1), and Loop2 (L2) are determined using start and stop primers based on the multiple sequence alignment of HAdV1-52 Genbank sequences.

1a) Forward primers PSFKPY/PTFKPY positions range from 111 to 121 and reverse primers FIGLMY/FVGLMY/FVGLLY position ranges from 299 to 354 are used to extract L1.

1b) Hyper-variable region L2 is between forward primer LMLDAL/LLLDSL/LLLDSI position from 338-384 and the reverse primers AMEINL/AMEINI from 423 to 490.

| Species | L1_Start Primer | Start Position | L1_End Primer | End Position |
|---|---|---|---|---|
| A | PSFKPY | 111 | FIGLMY | 299-306 |
| B1/B2 | PSFKPY | 111 | FIGLMY / FVGLMY | 316-334 |
| C | PTFKPY | 111 | FIGLMY | 333-354 |
| D | PSFKPY | 111-121 | FVGLMY / FVGLLY | 319-341 |
| E | PSFKPY | 111 | FVGLMY | 316 |
| F | PSFKPY | 111 | FVGLMY | 306-309 |
| G | PSFKPY | 111 | FVGLMY | 301 |

| Species | L2_Start Primer | Start Position | L2_End Primer | End Position |
|---|---|---|---|---|
| A | LMLDAL | 338-348 | AMEINL | 423-428 |
| B1/B2 | LLLDSL | 355-373 | AMEINI / AMEINL | 440-458 |
| C | LLLDSI | 372-384 | AMEINL / ALEINL | 458-490 |
| D | LLLDSI | 355-379 | AMEINL | 447-468 |
| E | LLLDSI | 355 | AMEINI | 442 |
| F | LMLDAL | 354-348 | AMEINL | 429-431 |
| G | LMLDAL | 340 | AMEINL | 425 |

*Source*: Madisch et al., Table 1 from Phylogenetic analysis of the main neutralization and hemagglutination determinants of all human adenovirus prototypes as a basis for molecular classification and taxonomy, 2005.

**Figure 5: Table 1 (1a and 1b) Hexon Primers**

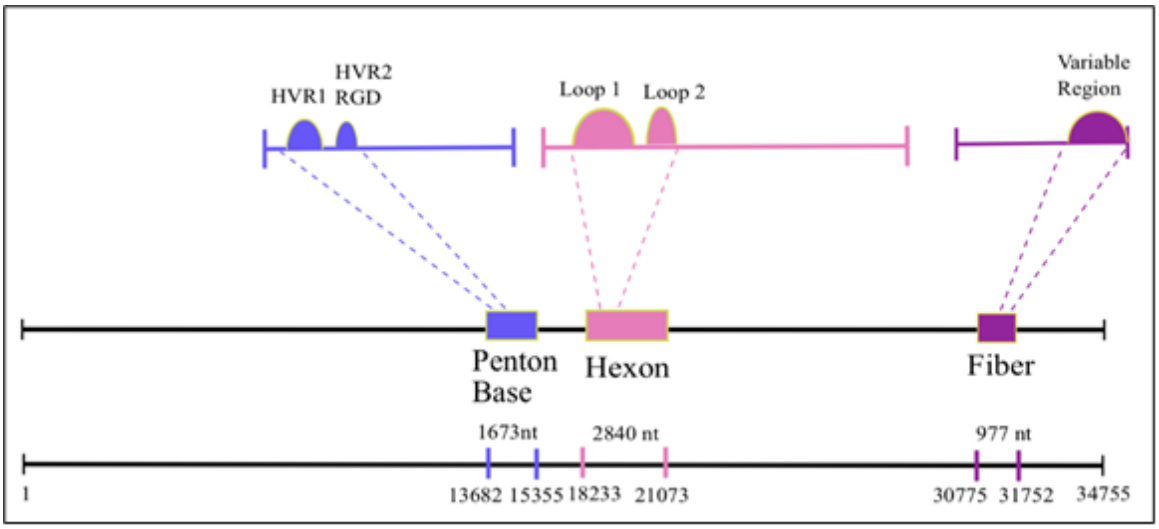
Hexon has two highly variable regions, Loop 1 and Loop 2 (L1, L2) (Notes: Epitopes are labeled based on their location in the genome. The HAdV hexon protein includes variable loops, L1, and L2; the penton base includes variable regions, hypervariable region1 (HVR1), and RGD; and the fiber includes the variable knob region.

Figure 6), followed by a highly conserved region (Con). With respect to Madisch et al.'s (2005) analysis for HAdV subtyping, the following unique primers specific to each subtype were derived from the multiple sequence alignment (Madisch et al., 2005):

- L1 is the region between amino acid sequences PSFKPY (start primer) and FIGLMY (stop primer) (Table 1a in Figure 5); and
- L2 is the region between the amino acid sequences LLLD/LLMD (start primer) and AMEI (stop primer) (Table 1b in Figure 5).

The fiber knob region spans positions 14–22 in the multiple sequence alignment for all types (see Note: The application has "Input," "prog," and "refdb" folders, a Java class file, and a "RUN_THIS.cmd" file.

Figure 9) and is variable and useful for distinguishing type.




Notes: Epitopes are labeled based on their location in the genome. The HAdV hexon protein includes variable loops, L1, and L2; the penton base includes variable regions, hypervariable region1 (HVR1), and RGD; and the fiber includes the variable knob region.

**Figure 6: Genome Organization of Adenovirus Capsid Proteins: Penton Base, Hexon, and Fiber**

**Table 2. Fiber knob domain primers.**

Fiber knob region start-primers observed from position 14-22 range for all the types. Species B with "YPYEDES", species C with "YPYDTET", species E with "YPYDADN", species F with "YPYEHYN", species G with "YPYDPPH", whereas species A and D have 3 unique start primers and to the entire length of fiber.

| Species | Fiberknob Start Primer | Start Position |
|---|---|---|
| A | YPFDPFD / YPFNPSD / YPFDPYD | 22 |
| B1 | YPYEDES | 15 |
| B2 | YPYEDES | 15 |
| C | YPYDTET | 15 |
| D | YPYGYAR / YPYDYAR / YPYGHR | 15 |
| E | YPYDADN | 16 |
| F | YPYEHYN | 14 |
| G | YPYDPPH | 14 |

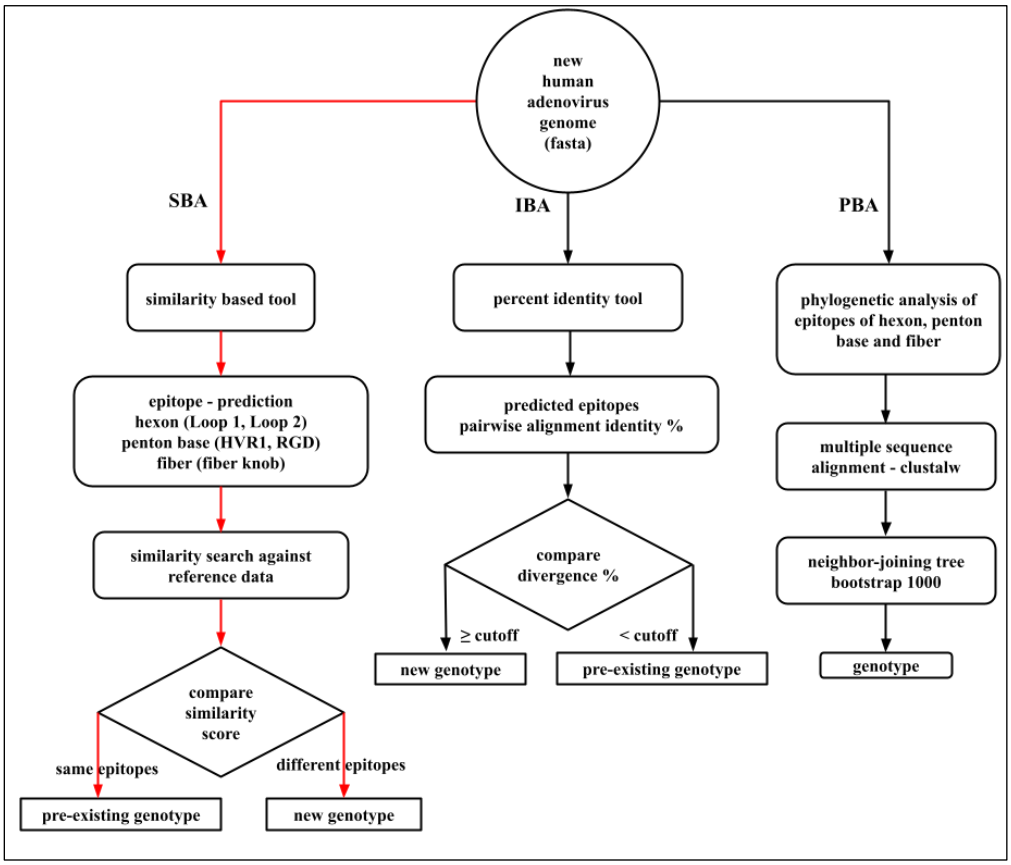
*Source*: Madisch et al., Table 2 from Phylogenetic analysis of the main neutralization and hemagglutination determinants of all human adenovirus prototypes as a basis for molecular classification and taxonomy, 2005.

**Figure 7: Table 2 Fiber Knob Domain Primers**

#### 2.1.2.2 Similarity Based Analysis – Bioinformatics Application.

As per the HAdV Working Group recommendations, the genotyping is carried out using major capsid proteins: penton base, hexon, and fiber knob. Sequence similarity analysis is performed using the National Center for Biotechnology Information (NCBI)'s Basic Local Alignment Search Tool (BLAST). A bioinformatics pipeline (Notes: Unknown HAdV FASTA sequence is BLAST-searched against reference databases of the penton base gene, hexon gene, fiber gene, and variable sequence regions of penton base (HVR1 and RGD), hexon (L1 and L2), and fiber (knob region). A BLAST search of each epitope sequence against a given reference dataset allows prediction of the virus genotype.

Figure 8) is developed to execute a local BLAST against in-house compiled reference genotypes. Running the similarity searches locally can save time and provide the flexibility of size, volume, database version, and Internet restrictions. The NCBI provides command line tools to run BLAST+ (U.S. National Library of Medicine, NCBI, accessible at https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download). The reference dataset's whole genome, penton base, hexon, and fiber nucleotide and amino acid sequences are used to generate the individual databases wholegenome_db, penton base_db, hexon_db, and fiber knob_db.




Notes: Unknown HAdV FASTA sequence is BLAST-searched against reference databases of the penton base gene, hexon gene, fiber gene, and variable sequence regions of penton base (HVR1 and RGD), hexon (L1 and L2), and fiber (knob region). A BLAST search of each epitope sequence against a given reference dataset allows prediction of the virus genotype.

**Figure 8: Similarity-Based Analysis (SBA) for Characterizing HAdV**

The FASTA format unknown genome/new genome of HAdV is given as input to

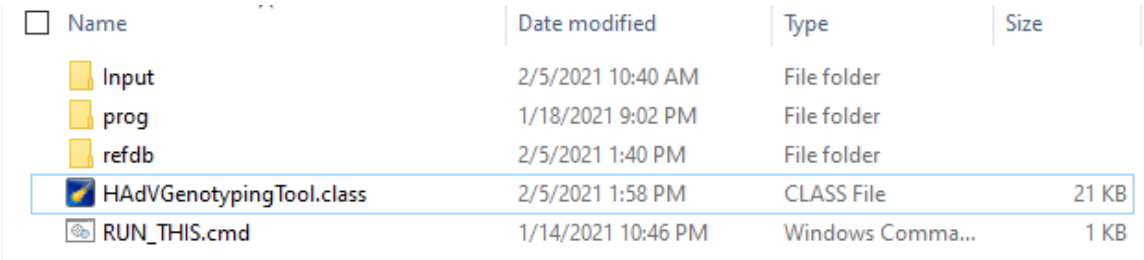
the pipeline (Note: The application has “Input,” “prog,” and “refdb” folders, a Java class file, and a “RUN_THIS.cmd” file.

Figure 9); a local BLAST is performed against all the reference databases (wholegenome_db, penton base_db, hexon_db, and fiber knob_db, L1_db, L2_db, Con_db, HVR1_db, RGD_db, fiberknob_db) and outputs the results in tab delimited format (Table 3 and Table 4).

The main purpose of the application is to characterize unknown or new human adenovirus genome to determine three capsid proteins: penton base, hexon, and fiber. BLAST is performed against the whole genome reference database, hexon reference database, penton base database, and fiber database, and the epitope databases of L1, L2 and Con, HVR1, RGD, and fiber knob. BLAST results are parsed based on highest hits—which has the highest length match and e-value and bit score. Based on the coordinates provided in the results, penton base, hexon, and fiber coding sequences are extracted for new genome and saved locally. The highest hit genotype is assigned to the individual protein to determine penton base, hexon, and fiber (PHF). These newly extracted sequences are used as input for a second iteration of BLAST against a variable region's reference database. Penton base variable region HVR1 and RGD, hexon variable regions L1, L2 and Con, and fiber variable region fiber knob's genotype values are extracted from BLAST results, and sequence files are saved locally for further bioinformatics analysis—identity based and phylogenetic analysis—to confirm the PHF assignment.

The SBA tool is a standalone application developed using Java version 1.8. It is used to characterize and classify newly sequenced human adenovirus genome. The application directory structure contains "Input", "prog", "refdb" directories, a Java class file, and "RUN_THIS.cmd" files (Note: The application has "Input," "prog," and "refdb" folders, a Java class file, and a "RUN_THIS.cmd" file.

Figure 9). A FASTA format (“>“with a header followed by nucleotide sequence) whole genome sequence file (query genome) is copied into “Input.” The “prog” directory contains BLAST executables, and “refdb” has all reference databases. Double-click the “RUN_THIS.cmd” to run the program in command line automatically. The query sequence is BLAST searched against reference datasets to determine epitopes. Based on highest similarity score of 90% and above and E-value, each epitope sequence and refined BLAST results are written into an autogenerated “Output” folder. The output contains BLAST results for individual proteins (penton base, hexon, and fiber) and epitopes (HVR1, RGD, L1, L2, and fiber knob). If both the variable regions HVR1 and RGD belong to genotype HAdV-C1, then the query genome’s penton base will be assigned as HAdV-C1 genotype; otherwise, HVR1 belongs to HAdV-C1 and RGD belongs to HAdV-C2, and the penton base is a new genotype. Similarly, for hexon, if L1, L2 and conserved BLAST results predict the same genotype (HAdV-C1), then the assignment is HAdV-C1; otherwise, the query genome’s hexon is assigned to a new genotype. If fiber knob BLAST results predict HAdV-C6, then the assigned fiber is HAdV-C6. The final output file is “AssignedPHF_Type.txt,” which is a tab delimited file with final PHF (P1H1F6) assignment (Table 3 and Table 4).




Note: The application has “Input,” “prog,” and “refdb” folders, a Java class file, and a “RUN_THIS.cmd” file.

**Figure 9: Similarity-Based Analysis Standalone Tool for Characterizing HAdV**

### *2.1.3 Results*

Characterizations of new human adenovirus are referred by distinct recombination in the PHF open reading frames (ORFs) (Seto et al., 2011). The highest similar reference genotype is assigned for each ORF to determine new genotype. The similarity-based genotyping tool has been tested on genotypes 1–52 (Table 3). These are unique genotypes, and BLAST hits identify them successfully as original genotypes with highest percent similarity score.

**Table 3**

*Similarity Based Tool Predicted Genotype Results for HAdV Genotypes 1 to 52*

| Query | Penton base | Hexon | Fiber | Assigned PHF |
|---|---|---|---|---|
| HAdV-C1 | 1 | 1 | 1 | P1/H1/F1 |
| HAdV-C2 | 2 | 2 | 2 | P2/H2/F2 |
| HAdV-B3 | 3 | 3 | 3 | P3/H3/F3 |
| HAdV-E4 | 4 | 4 | 4 | P4/H4/F4 |
| HAdV-C5 | 5 | 5 | 5 | P5/H5/F5 |
| HAdV-C6 | 6 | 6 | 6 | P6/H6/F6 |

| Query | Penton base | Hexon | Fiber | Assigned PHF |
|---|---|---|---|---|
| HAdV-B7 | 7 | 7 | 7 | P7/H7/F7 |
| HAdV-D8 | 8 | 8 | 8 | P8/H8/F8 |
| HAdV-D9 | 9 | 9 | 9 | P9/H9/F9 |
| HAdV-D10 | 10 | 10 | 10 | P10/H10/F10 |
| HAdV-B11 | 11 | 11 | 11 | P11/H11/F11 |
| HAdV-A12 | 12 | 12 | 12 | P12/H12/F12 |
| HAdV-D13 | 13 | 13 | 13 | P13/H13/F13 |
| HAdV-B14 | 14 | 14 | 14 | P14/H14/F14 |
| HAdV-D15 | 15 | 15 | 15 | P15/H15/F15 |
| HAdV-B16 | 16 | 16 | 16 | P16/H16/F16 |
| HAdV-D17 | 17 | 17 | 17 | P17/H17/F17 |
| HAdV-A18 | 18 | 18 | 18 | P18/H18/F18 |
| HAdV-D19 | 19 | 19 | 19 | P19/H19/F19 |
| HAdV-D20 | 20 | 20 | 20 | P20/H20/F20 |
| HAdV-B21 | 21 | 21 | 21 | P21/H21/F21 |
| HAdV-D22 | 22 | 22 | 22 | P42/H22/F22 |
| HAdV-D23 | 23 | 23 | 23 | P23/H23/F23 |
| HAdV-D24 | 24 | 24 | 24 | P27/H24/F24 |
| HAdV-D25 | 25 | 25 | 25 | P25/H25/F25 |
| HAdV-D26 | 26 | 26 | 26 | P26/H26/F26 |
| HAdV-D27 | 27 | 27 | 27 | P27/H27/F27 |
| HAdV-D28 | 28 | 28 | 28 | P28/H28/F28 |
| HAdV-D29 | 29 | 29 | 29 | P29/H29/F29 |
| HAdV-D30 | 30 | 30 | 30 | P30/H30/F30 |
| HAdV-A31 | 31 | 31 | 31 | P31/H31/F31 |
| HAdV-D32 | 32 | 32 | 32 | P32/H32/F32 |
| HAdV-D33 | 33 | 33 | 33 | P33/H33/F33 |

| Query | Penton base | Hexon | Fiber | Assigned PHF |
|---|---|---|---|---|
| HAdV-B34 | 34 | 34 | 34 | P34/H34/F34 |
| HAdV-B35 | 35 | 35 | 35 | P35/H35/F35 |
| HAdV-D36 | 36 | 36 | 36 | P36/H36/F36 |
| HAdV-D37 | 37 | 37 | 37 | P37/H37/F37 |
| HAdV-D38 | 38 | 38 | 38 | P38/H38/F38 |
| HAdV-D39 | 39 | 39 | 39 | P39/H39/F39 |
| HAdV-F40 | 40 | 40 | 40 | P40/H40/F40 |
| HAdV-F41 | 41 | 41 | 41 | P41/H41/F41 |
| HAdV-D42 | 42 | 42 | 42 | P42/H42/F42 |
| HAdV-D43 | 43 | 43 | 43 | P43/H43/F43 |
| HAdV-D44 | 44 | 44 | 44 | P44/H44/F44 |
| HAdV-D45 | 45 | 45 | 45 | P45/H45/F45 |
| HAdV-D46 | 46 | 46 | 46 | P46/H46/F46 |
| HAdV-D47 | 47 | 47 | 47 | P47/H47/F47 |
| HAdV-D48 | 48 | 48 | 48 | P48/H48/F48 |
| HAdV-D49 | 49 | 49 | 49 | P49/H49/F49 |
| HAdV-B50 | 50 | 50 | 50 | P50/H50/F50 |
| HAdV-D51 | 51 | 51 | 51 | P51/H51/F51 |
| HAdV-G52 | 52 | 52 | 52 | P52/H52/F52 |

Note: Results file contains tab delimited output of query genome name, assigned penton base, hexon, and fiber genotype based on similarity scores.

The application is further tested for recombinant genotypes 53 to 103 (Table 4). Accession numbers: HAdV-D53 (FJ169625), HAdV-D54 (AB333801), HAdV-B55 (FJ643676), HAdV-D56 (HM770721), HAdV-C57 (HQ003817), HAdV-D58 (HQ883276), HAdV-D59 (JF799911), HAdV-D60 (HQ007053), HAdV-A61 (JF964962),

HAdV-D62 (JN162671), HAdV-D63 (JN935766), HAdV-D64 (EF121005), HAdV-D65 (AP012285), HAdV-B66 (JN860676), HAdV-D67 (AP012302), HAdV-B68 (JN860678), HAdV-D69 (JN226748), HAdV-D70 (KP641339), HAdV-D71 (KF268207), HAdV-D72 (KF268335), HAdV-D73 (KY618676), HAdV-D74 (KY618677), HAdV-D75 (KY618678), HAdV-B76 (KF633445), HAdV-B77 (KF268328), HAdV-B78 (KT970440), HAdV-B79 (LC177352), HAdV-D81 (AB765926.1), HAdV-D82 (LC066535.1), HAdV-D83 (KX827426.1), HAdV-D84 (MF416150), HAdV-D85 (LC314153), HAdV-D86 (KX868297), HAdV-D87 (MF476841), HAdV-D88 (MF476842), HAdV-C89 (MH121097), HAdV-D91 (KF268208), HAdV-D92 (KF268325), HAdV-D93 (KF268334), HAdV-D94 (KF268201), HAdV-D95 (KF268206), HAdV-D96 (KF268327), HAdV-D97 (KF268320), HAdV-D98 (KF268332), HAdV-D99 (KF268211), HAdV-D100 (KF268330), HAdV-D101 (KF268324), HAdV-D102 (KF268312), HAdV-D103 (KF268322)

**Table 4**

*Similarity Based Tool Predicted Genotype Results for HAdV Genotypes 53to 103*

| Query | Penton base | Hexon | Fiber | Assigned P/H/F |
|---|---|---|---|---|
| HAdV-D53 | 37 | 22 | 8 | P37/H22/F8 |
| HAdV-D54 | 54 | 54 | 8 | P54/H54/F8 |
| HAdV-B55 | 14 | 11 | 14 | P14/H11/F14 |
| HAdV-D56 | 9 | 15* | 9 | P9/H15/F9 |
| HAdV-C57 | 1 | 57 | 6 | P1/H57/F6 |
| HAdV-D58 | 58 | 58 | 29 | P58/H358F29 |
| HAdV-D59 | 59 | 25 | 9 | P59/H25/F9 |

| Query | Penton base | Hexon | Fiber | Assigned P/H/F |
|---|---|---|---|---|
| HAdV-D60 | 60 | 20 | 60 | P60/H20/F60 |
| HAdV-A61 | 31 | 61 | 31 | P31/H61/F31 |
| HAdV-D62 | 62 | 62 | 62 | P62/H62/F62 |
| HAdV-D63 | 30 | 30 | 29 | P30/H30/F29 |
| HAdV-D64 | 22 | 19 | 37 | P22/H19/F37 |
| HAdV-D65 | 65 | 10 | 9 | P65/H10/F9 |
| HAdV-B66 | 7 | 7 | 3 | P7/H7/F3 |
| HAdV-D67 | 67 | 9 | 25 | P67/H9/F25 |
| HAdV-B68 | 16 | 3 | 16 | P16/H3/F16 |
| HAdV-D69 | 69 | 15* | 42 | P69/H15/F42 |
| HAdV-D70 | 70 | 70 | 29 | P70/H70/F29 |
| HAdV-D71 | 9 | 20 | 71 | P9/H20/F71 |
| HAdV-D72 | 72 | 30 | 72 | P72/H30/F72 |
| HAdV-D73 | 73 | 73 | 27 | P73/H73/F27 |
| HAdV-D74 | 74 | 74 | 51 | P74/H74/F51 |
| HAdV-D75 | 75 | 26 | 29 | P75/H26/F29 |
| HAdV-B76 | 21 | 21 | 16 | P21/H21/F16 |
| HAdV-B77 | 35 | 34 | 7 | P35/H34/F7 |
| HAdV-B78 | 11 | 11 | 7 | P11/H11/F7 |
| HAdV-B79 | 11 | 34 | 11 | P11/H34/F11 |
| HAdV-D81 | 49 | 48 | 22 | P49/H48/F22 |
| HAdV-D82 | 82 | 15* | 37 | P82/H29/F37 |
| HAdV-D83 | 83 | 9 | 15 | P83/H9/F15 |
| HAdV-D84 | 43 | 17 | 84 | P43/H17/F84 |
| HAdV-D85 | 37 | 19 | 8 | P37/H19/F8 |
| HAdV-D86 | 9 | 25 | 25 | P9/H25/F25 |
| HAdV-D87 | 9 | 15* | 25 | P9/H15/F25 |

| Query | Penton base | Hexon | Fiber | Assigned P/H/F |
|---|---|---|---|---|
| HAdV-D88 | 88 | 15* | 9 | P88/H15/F9 |
| HAdV-C89 | 89 | 2 | 2 | P89/H2/F2 |
| HAdV-D91 | 37 | 37 | 17 | P37/H37/F17 |
| HAdV-D92 | 92 | 32 | 27 | P92/H32/F27 |
| HAdV-D93 | 28 | 37 | 38 | P28/H37/F38 |
| HAdV-D94 | 33 | 15* | 9 | P33/H15/F9 |
| HAdV-D95 | 95 | 9 | 15 | P95/H9/F15 |
| HAdV-D96 | 23 | 32 | 62 | P23/H32/F62 |
| HAdV-D97 | 97 | 28 | 22 | P97/H28/F22 |
| HAdV-D98 | 98 | 46 | 9 | P98/H46/F9 |
| HAdV-D99 | 9 | 46 | 39 | P9/H46/F39 |
| HAdV-D100 | 100 | 17 | 30 | P100/H17/F30 |
| HAdV-D101 | 101 | 37 | 45 | P101/H37/F45 |
| HAdV-D102 | 102 | 38 | 30 | P102/H38/F30 |
| HAdV-D103 | 103 | 33 | 30 | P103/H33/F30 |

Note: Results are query genome, assigned penton base, hexon, and fiber based on similarity scores. Genotypes 15* and 29 are considered single neutralization genotypes of same species.

#### 2.1.3.1 SBA-NEW Genotypes.

***2.1.3.1.1 HAdV-B54.*** HAdV-B54 was first reported from a nosocomial outbreak in Japan. Initial serological methods typed this as HAdV-8. Genomic analysis of penton base, hexon, and fiber showed similarity score in BLAST search <95% with HAdV-D45, <95% with HAdV-D32, and 97% with HAdV-D8, respectively. Considering a 95% cutoff

score, this resulted in the penton base as new, hexon as new, and fiber as HAdV-D8, P54/H54/F8.

***2.1.3.1.2*** ***HAdV-B55.*** HAdV-B55 is an acute respiratory pathogen completely characterized in China in 2006. This was first identified in 1969 in Spain as HAdV-B11a. The whole genome has a 98% similarity to HAdV-B14, whereas protein level analysis was penton base, 99% with HAdV-B14; hexon, 98% with HAdV-B11; and fiber, 99% with HAdV-B14. Hence, the predicted genotype was P14/H11/F14.

***2.1.3.1.3*** ***HAdV-D56.*** Analysis found penton base, 99% with HAdV-D10; hexon, 98% with -D15 and
-D29; and fiber, 99% with D9. The SBA prediction is P10/H15/F9.

***2.1.3.1.4*** ***HAdV-A61.*** HAdV-A61 has the highest similarity score with -A31, for penton base and fiber at 99% and 98%, respectively, whereas hexon does not have a significant match with the cutoff; hence, the application predicts and assigns P31/H61/F31.

***2.1.3.1.5*** ***HAdV-D72.*** BLAST results predict significant similarity score of <95% cutoff for penton base and fiber with -D43 and -D44,

respectively. Hexon is 98% similar with -D30. Hence, the program predicts the sequence's genotype as P72/H30/F72.

***2.1.3.1.6*** ***HAdV-D85.*** The application parsed penton base of -D37 with 100% similarity, hexon of -D9 with 98%, and fiber of -D8 with 100%, and assigned P37/H9/F8.

***2.1.3.1.7*** ***HAdV-D92.*** There is no significant match found for penton base in the BLAST results within the cutoff. Hexon is -D32 with >98%, and fiber is -D27 with 99%. The predicted genotype is P92/H32/F27.

***2.1.3.1.8*** ***HAdV-D101.*** For HAdV-D101, a significant hit was identified for penton base above cutoff of -D45 with 96% similarity. Further variable region BLAST analysis predicted HVR1 and RGD with different genotypes (recombination); hence, penton base is identified as a new one. Hexon and fiber have hits of >95% with -D37 and -D45, respectively. The final assignment is P101/H37/F45.

***2.1.3.1.9*** ***HAdV-D103.*** Due to the recombination of HVRs, penton is considered as new genotype -D103. A significant hit of -D25 with >98% similarity assigns hexon H25, whereas for fiber, there are two

significant matches found within the given criteria, -D30 and -D49. The assignment is P103/H25/F30.

## 2.2 Percent Identity Based Analysis (IBA) Human Adenovirus

### *2.2.1 Introduction*

Identifying human adenovirus is essential for understanding the epidemiology of outbreaks and the pathways of molecular evolution of adenoviral genotypes. However, genomic sequencing reveals that regions may undergo genome recombination as an evolutionary mechanism (Dehghan et al., 2013; Walsh et al., 2010); ones that include the epitope regions of the three major capsid proteins may be used to characterize the emergent potential pathogen. Many examples of these novel pathogens that arose through recombination among the penton base, hexon, and fiber genes have been reported (Grodzicker et al., 1974; Mautner et al., 1975; Williams et al., 1975). Capsid proteins such as hexon and fiber contain various epitope sites and were used as a target for detection (Norrby, 1969; Rux & Burnett, 2000). Historically, serological methods utilizing specific antibodies were used to detect the epitope markers. The hexon epitope region folds into two distinct loops with hypervariable amino acid sequences—Loop 1 (L1) and Loop 2 (L2)—and was known as the **ε** (epsilon) fragment (Crawford-Miksza & Schnurr, 1996). The fiber knob contains the variable region $\gamma$ (gamma) epitope and is also variable with respect to amino acid sequence. The penton base variable regions are located on the hypervariable region 1 (HVR1) and the RGD loop (HVR2) (Zubieta et al., 2005) and have been shown to elicit an antibody response.

The novel genotypes arise from either a newly determined capsid protein, through mutations in the variable sequence regions, or predetermined capsid proteins with a different genome recombination pattern. Characterizing these epitope regions gives more insight into recognizing genotypes of novel HAdV pathogens. Based on sequence analysis of epitopes HVR1, RGD, ε (L1 and L2), and *γ*, a molecular-based classification method is developed to identify new genotypes. Analyzing the sequence identity comparing a new genomic region with existing genotypes will determine the genotype of the region. Pairwise sequence alignment is used to identify regions of similarity, which is useful in indicating the functional, structural, and evolutionary relationships between two genomic sequences. Pairwise alignment is done using different methods, such as dot-matrix, dynamic programming, etc. Software tools such as Emboss (https://www.ebi.ac.uk/Tools/psa/emboss_needle/) are used to determine the identity score between the sequences. Determining individual percent identity scores for each pair and analyzing the data is time consuming; hence, the process is automated for all proteins and epitopes.

The goal of this study is to develop a method to automate the process of determining percent identity scores between the pairs for HVR1, RGD, L1, L2, and fiber knob epitopes of capsid proteins, for the purposes of identifying and characterizing human adenovirus genotypes, including novel, emergent types.

### *2.2.2 Materials and Methods*

***Multiple Sequence Alignment***

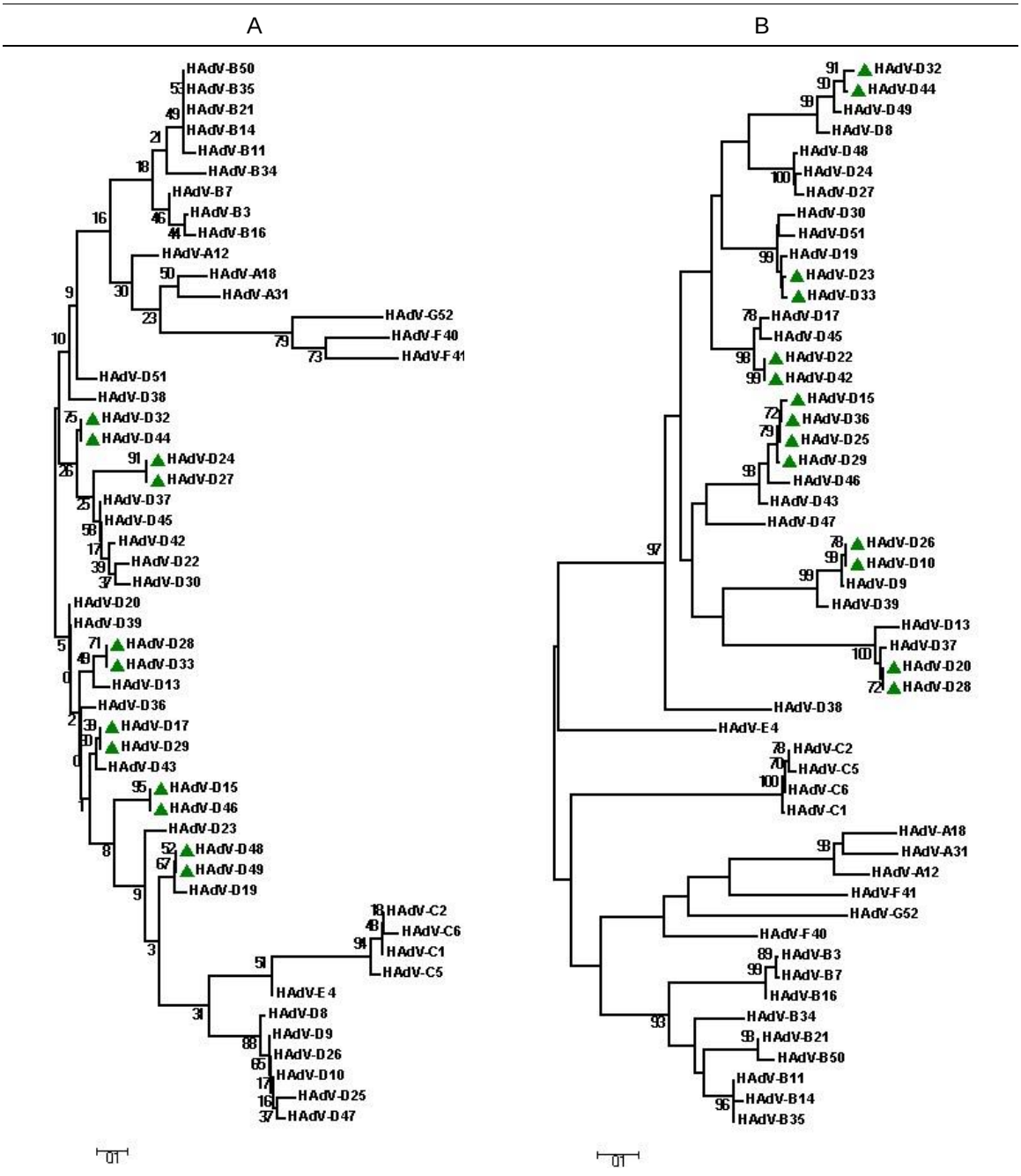
Multiple sequence alignment (MSA) is created using the ClustalW algorithm in Molecular Evolutionary Genetics Analysis (MEGA) version 7.0.14 (Saitou & Nei, 1987; Tamura et al., 2004, Kumar et al., 2016). Whole genome, penton base, hexon, and fiber nucleotide sequence alignments, as well as amino acid sequences from the latter three, are performed with gap opening and extension penalties of 10, 0.5, BLOSUM62 and PAM250 matrices, respectively.

***Phylogenetic Tree Building***

The phylogenetic trees were constructed using MEGA software. The evolutionary distances were computed using the Maximum Composite Likelihood method (Tamura et al., 2007) and are in the units of the number of base substitutions per site. The distances from each sequence to all others were calculated and stored in a matrix. The trees were constructed from the distance matrix using a specific tree-building algorithm Neighbor-Joining method (Saitou et al., 1987). The analysis involved 52 nucleotide sequences. All positions containing gaps and missing data were eliminated. The phylogenetic trees are constructed with 1000 bootstrap iteration. The scale at the bottom of the tree (Column A in Figure 10) represents 0.1 nucleotides per site in the multiple sequence alignment.

Percent identity (PI) is the percentage of identical matches between the two nucleotide sequences over the aligned region, including the gaps. PI scores were determined by aligning two sequences with Needleman-Wunsch global alignment (http://www.bioinformatics.nl/cgi-bin/emboss/needleall) under default parameters, gap

opening penalty 10.0, gap extension penalty 0.5. Percent divergence (PD) is calculated by doing a 100% identity score. The pairwise divergence scores (PDS) are sorted for each genotype from lowest to highest divergence score along with genotype. The results of sorted PD for different species are shown in Table 5.




Notes: (A) Phylogenetic tree of HVR1 nucleotide sequences for HAdV 1 through 52 genotypes. (B) Phylogenetic tree of RGD nucleotide sequences for HAdV 1 through 52 genotypes. These are constructed from CLUSTALW aligned MSA and neighbor-joining methods using MEGA. 0.05 nucleotides per site in the alignment. The highlighted (green solid triangle) pairs are closely related HAdV genotypes.

**Figure 10: Phylogenetic Analysis of Human Adenovirus HVR1, RGD Nucleotide Sequences**

**Table 5**

*Pairwise Sequence Alignment*

| HAdV | First Closest to | PDS | Second Closest to | PDS |
|---|---|---|---|---|
| HAdV-A31 | HAdV-A18 | 16.1 | HAdV-G52 | 30.9 |
| HAdV-B7 | HAdV-B3 | 10.1 | HAdV-B21 | 29.7 |
| HAdV-B11 | HAdV-B35 | 15.1 | HAdV-B21 | 19.7 |
| HAdV-C1 | HAdV-C2 | 27.2 | HAdV-C6 | 30.9 |
| HAdV-D29 | HAdV-D15 | 0.0 | HAdV-D51 | 22.1 |
| HAdV-D39 | HAdV-D43 | 5.8 | HAdV-D36 | 20.6 |

**Note:** Pairwise alignment score sorted results of Loop 1 (L1) of the hexon gene. Listed are the first and second divergence scores of HAdV-A31, -B7, -B11, -C1, -D29, -D39. The closest possible divergence score for HAdV-D39 is -D43, with a pairwise divergence score (PDS) of 5.8.

### *Sequence Divergence Determination as a Basis for HAdV Genotyping*

### *Penton Base (HVR1 and RGD)*

Determining the sequence divergence cutoff is performed by observing phylogenetic tree relationships of the variable regions HVR1 and RGD of the penton base (Figure 10). The steps include identifying the closely related pairs within the species and clade. The percent divergence scores for these identified pairs are in ascending order to determine the least possible score between the closest pair to establish the criteria or cutoff that will be used to define a new type. For HVR1, divergence scores are 0.0% for HAdV-D48 and -D49; HAdV-D27 and -D24; HAdV-D32 and -D44; and HAdV-D15 and -D46. HAdV-D29 and -D17 and HAdV-D28 and -D33 have the same score, 1.6% (Column A of Table 6). The HAdV-D28 and -D33 pair is considered because of its

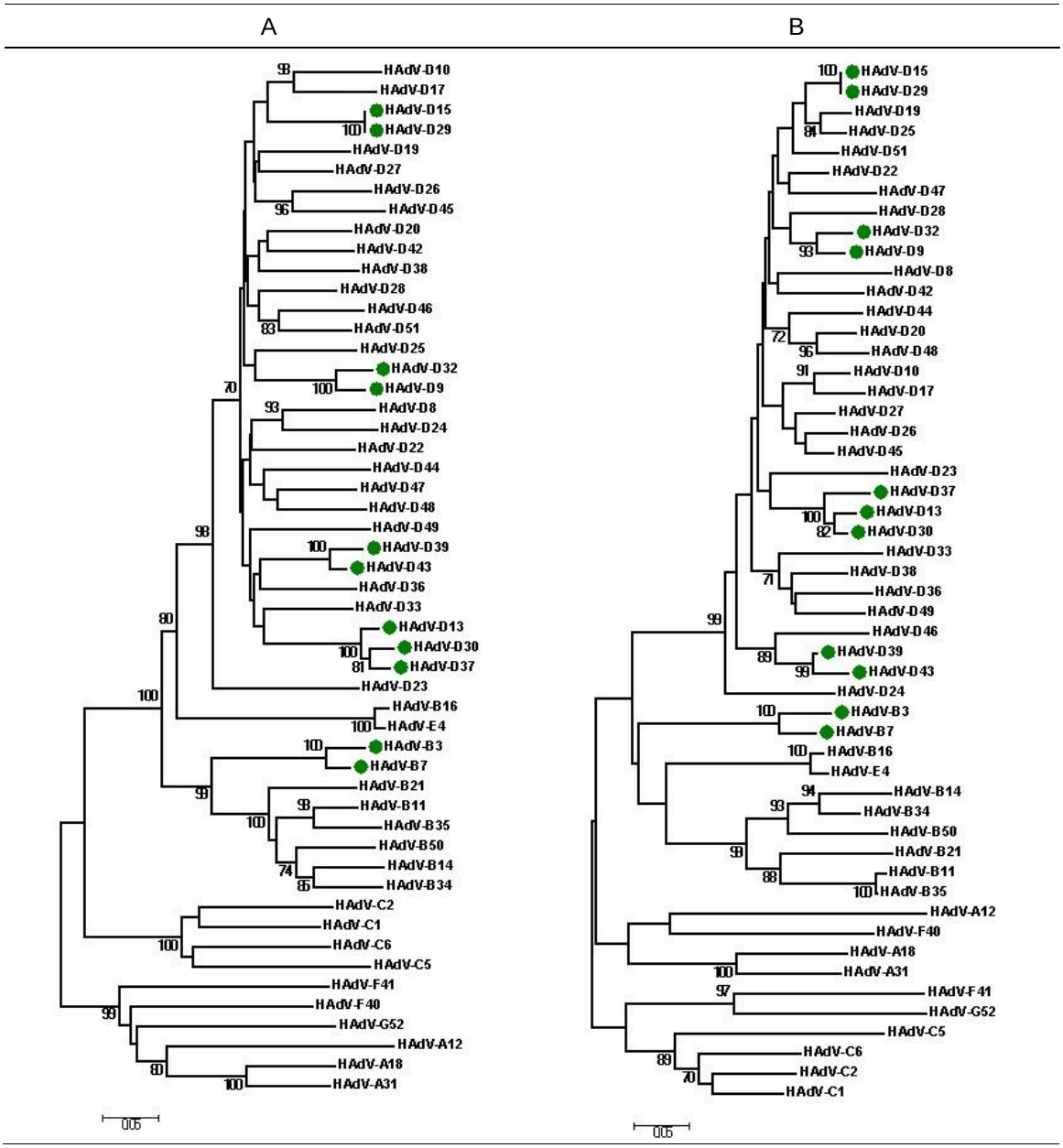
greater bootstrap value in the tree. Similarly, for RGD, the pair HAdV-D32 and -D44 with 1.5% divergence is considered for genotyping cutoff (Column B of Table 6).

**Table 6**

*HAdV Divergence Score to Determine Genotype for Penton Base*

| A | | | B | | |
|---|---|---|---|---|---|
| ID1 | ID2 | HVR1 % Divergence | ID1 | ID2 | RGD % Divergence |
| HAdV-D28 | HAdV-D33 | 1.6 | HAdV-D32 | HAdV-D44 | 1.5 |
| HAdV-D29 | HAdV-D17 | 1.6 | HAdV-D15 | HAdV-D25 | 0.5 |
| HAdV-D15 | HAdV-D46 | 0.0 | HAdV-D23 | HAdV-D33 | 0.1 |
| HAdV-D32 | HAdV-D44 | 0.0 | HAdV-D25 | HAdV-D36 | 0.0 |
| HAdV-D27 | HAdV-D24 | 0.0 | HAdV-D26 | HAdV-D10 | 0.0 |
| HAdV-D48 | HAdV-D49 | 0.0 | HAdV-D28 | HAdV-D20 | 0.0 |

Note: (A) The divergence scores for HVR1 nucleotide sequences for closest pairs from the phylogenetic tree are listed to derive the genotype cutoff. For species D, HAdV-D28 and HAdV-D33 have the lowest divergence score to differentiate genotypes within the species. For species B, HAdV-B3 and HAdV-B7 were identified with the lowest divergence score. (B) Derived divergence scores for RGD nucleotide sequences for closest pairs from the phylogenetic tree are listed to derive the genotype cutoff. The cutoff is 1.5 between HAdV-D32 and HAdV-D44.




Notes: The hyper variable loops are useful for determining new genotypes. (A) Phylogenetic tree of human adenovirus genotypes from 1 to 52 of L1 nucleotide sequences. (B) Phylogenetic tree of human adenovirus genotypes from 1 to 52 of L2 nucleotide sequences. These trees are constructed from CLUSTALW aligned MSA and neighbor-joining methods using MEGA. The scale in the graph 0.05 indicates the number of nucleotides per site in the alignment. The highlighted (green solid dots) pairs are closely related HAdV genotypes.

**Figure 11: Phylogenetic Analysis of Human Adenovirus Hexon Gene's Variable Loops, Loop 1 and Loop 2**

**Table 7**

*Genotype Determination Using HAdV Hexon Loops Divergence Score*

| ID1 | ID2 | L1 % Divergence | L2 % Divergence |
|---|---|---|---|
| HAdV-D15 | HAdV-D29 | 0.0 | 0.4 |
| HAdV-D39 | HAdV-D43 | 5.8 | 3.2 |
| HAdV-D30 | HAdV-D37 | 6.0 | 7.3 |
| HAdV-D13 | HAdV-D37 | 6.5 | 7.3 |
| HAdV-D9 | HAdV-D32 | 9.2 | 7.4 |
| HAdV-B3 | HAdV-B7 | 10.1 | 10.1 |

Notes: Listed the s divergence scores of L1 and L2 nucleotide sequences to derive the genotype cutoff. Species D, HAdV-D39 and HAdV-D43, has the lowest divergence score to differentiate genotypes within the species. For species B, HAdV-B3 and HAdV-B7 were identified with lowest divergence score.

### *Hexon (L1 and L2)*

Based on the phylogenetic trees of the hexon highly variable sequence regions (epitopes) Loops L1 and L2, closely related/pairs were identified, and the percent divergence for the pairs were determined and highlighted with solid green dots in Figure 11. The first closest pair in L1 is HAdV-D15 and HAdV-D29, which has a genetic divergence of 0.0% (Table 5) at the nucleotide sequence level, and they are highly similar with one amino acid difference (Madisch et al., 2005). The second best closely related pair within species D is HAdV-D39 and HAdV-43 (Table 5 and Table 7), with 5.8% divergence for L1 and 3.2% for L2. This pair is clearly discerned by no cross-neutralization according to Heim analysis (Madisch et al., 2005). Nucleotide sequence divergence >5.8% for L1 and >3.2% for L2 compared to the next homologous sequence

in the reference data set is the criterion for a new genotype. For species B, HAdV-B3 and HAdV-B7 with 10.1% score determines the cutoff for genotyping for both L1 and L2.

***Fiber (Fiber Knob)***

A nucleotide sequence-based phylogenetic tree of 52 genotypes is analyzed to determine the criteria cutoff for fiber knob typing. The closest pairs, HAdV-D44 and HAdV-D48, HAdV-D32 and HAdV-D33, HAdV-D34 and HAdV-D35, HAdV-D24 and HAdV-D46, HAdV-D13 and HAdV-D38, HAdV-D19 and HAdV-D37, HAdV-D30 and HAdV-D49, have highly similar fiber knob with >99% sequence similarity. HAdV-D28 and HAdV-D43, HAdV-D8 and HAdV-D9, and HAdV-D20 and HAdV-D47 (Table 8 and Figure 12) suggest that they share similar cross-neutralization and hemagglutination-inhibition (Madisch et al., 2005). The least possible divergence score between the pairs HAdV-D22 and HAdV-D42 is identified with a 9.6% score to determine a new genotype.

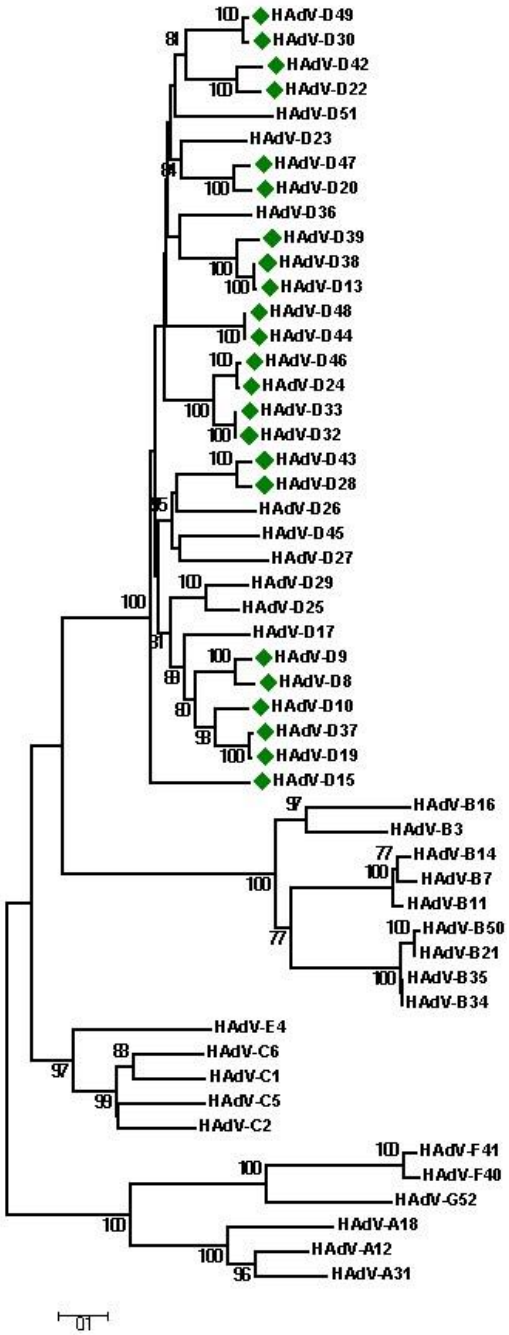
**Table 8**

*HAdV Divergence Score to Determine Genotype*

| ID1 | ID2 | Fiber knob % Divergence |
|---|---|---|
| HAdV-D22 | HAdV-D42 | 9.6 |
| HAdV-D39 | HAdV-D38 | 9.3 |
| HAdV-D20 | HAdV-D47 | 8.0 |
| HAdV-D8 | HAdV-D9 | 7.4 |
| HAdV-F40 | HAdV-F41 | 7.1 |
| HAdV-B7 | HAdV-B11 | 5.8 |
| HAdV-B11 | HAdV-B14 | 5.7 |
| HAdV-D28 | HAdV-D43 | 5.0 |

| ID1 | ID2 | Fiber knob % Divergence |
|---|---|---|
| HAdV-D30 | HAdV-D49 | 1.5 |
| HAdV-D19 | HAdV-D37 | 0.9 |
| HAdV-D13 | HAdV-D38 | 0.7 |
| HAdV-D24 | HAdV-D46 | 0.5 |
| HAdV-B21 | HAdV-B50 | 0.4 |
| HAdV-B34 | HAdV-B35 | 0.4 |
| HAdV-D32 | HAdV-D33 | 0.1 |
| HAdV-D44 | HAdV-D48 | 0.1 |

Note: Listed are divergence scores of fiber knob nucleotide sequences to derive the genotype. For species D, HAdV-D22 and HAdV-D42 have the lowest divergence score to differentiate genotypes within the species. For species B, HAdV-B7 and HAdV-B11 were identified with lowest divergence score of 5.8%.



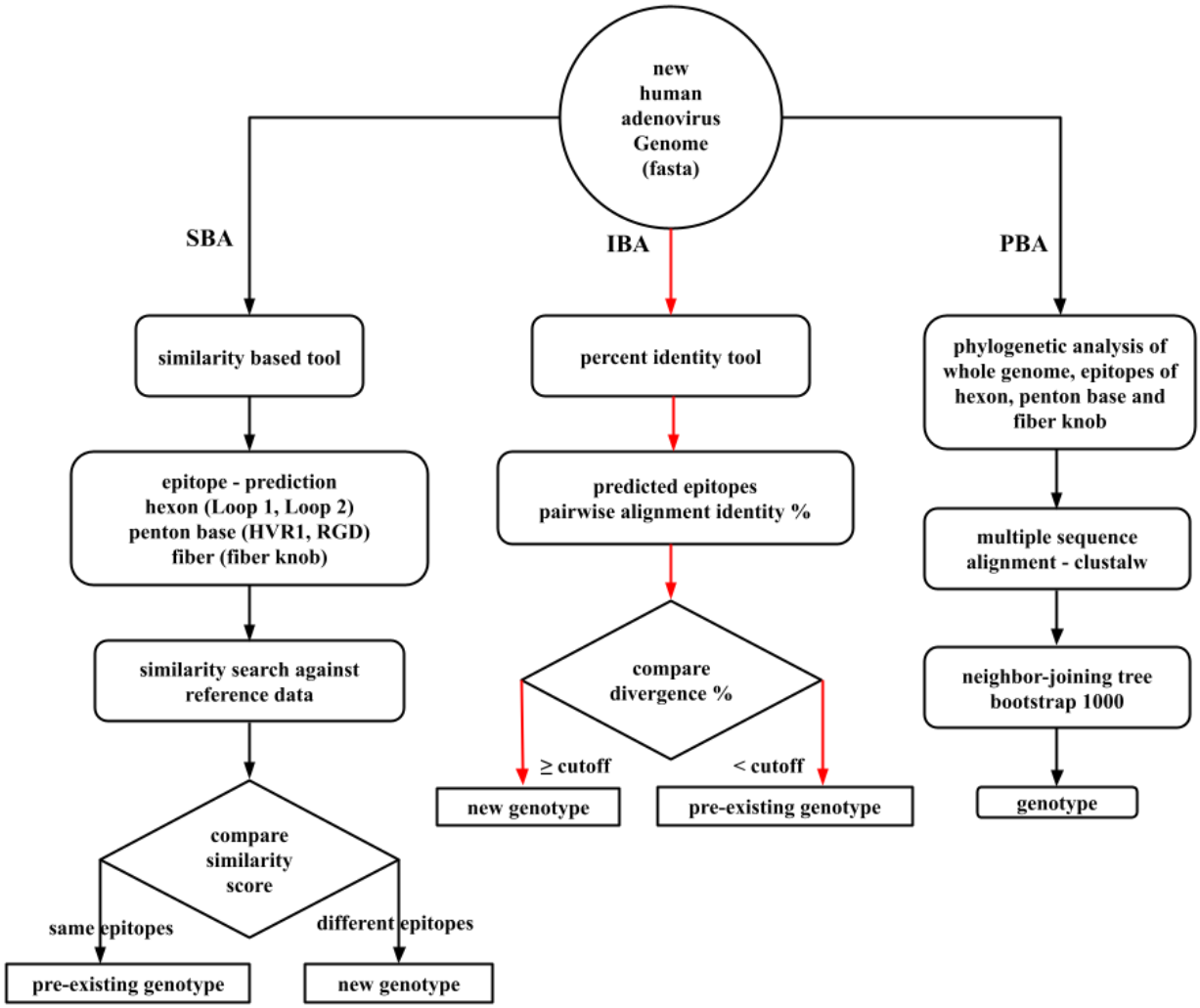

Note: This fiber knob phylogenetic tree is constructed from CLUSTALW aligned MSAs and neighbor-joining methods using MEGA. The bar shows a scale of 0.1 nucleotides per site in the alignment. Phylogenetic distance-based pairs are highlighted in green solid diamonds.

**Figure 12: Phylogenetic Tree of Fiber Knobs from 52 HAdV Genotypes**

### *2.2.3 Percent Identity-Based Analysis (IBA) – Bioinformatics Application*

Percent identity-based analysis (IBA) was applied to characterize each hypervariable region: HVR1, RGD, L1, L2, and fiber knob. For a given new genome, variable sequences are determined as described in SBA in Section 2.1. The HVR1 sequence is aligned with individual sequences in the data set comprising HAdV 1–52 genotypes. Pairwise sequence alignment is performed to determine the percent identity scores. Scores are parsed from the results and used to calculate percent divergence score. This iterative process is automated to do an alignment, and to parse the scores to generate the results. Percent divergence scores are sorted for individual epitopes. The divergence score is compared with the cutoff; if it is greater than the cutoff, it is a new type. For example, if a new genome presents a L1 divergence score that is 2.8% with HAdV-D23, the L1 is assigned as HAdV-D23 for new genome, since the cutoff is 5.8%.

The "HAdVGenotypingTool" application performs SBA as a first step (Figure 13), and the epitope sequences are saved. Newly extracted epitope sequences are used as input, pairwise alignment is performed for the reference data set of sequences from HAdV-1 through 52. Parsed identity scores are saved for each epitope and based on the cutoff value, the application determines whether the new epitope is a pre-existing or new type. The results are saved to an output file for reference. Characterization of new human adenoviruses are referred by distinct recombination in the PH) open reading frames (Seto et al., 2011) and epitope level recombination.




Notes: The epitopes predicted and extracted based on BLAST analysis of unknown HAdV FASTA sequences are used to determine the pairwise identity/divergence with the reference data set. Based on closely related pairs in the phylogenetic tree, the divergence score is determined as a cutoff percentage to determine a new genotype.Compare the score to determine the genotype based on the cutoff for each epitope.

**Figure 13: Percent Identity Based Analysis (IBA) for Characterizing Human Adenovirus Capsid Proteins**

### *2.2.4 Results*

**HAdV-D70.** Whole genome SBA predicted that HAdV-D70 is highly similar to HAdV-D27 with 96.5% BLAST identity score. Nucleotide similarity analysis identified

the predicted penton base as that of HAdV-D38 with 97.8%, with the hexon similar to those from HAdV-D30 and HAdV-D37 with a score of 97.0%, and the fiber is identical to that from HAdV-D29 with a 99.9% score. These data indicate hexon is a recombination of -D30 and -D37. Epitope level analysis predicted penton base HVR1 has an identity of 100.0% with HAdV-D37. RGD is HAdV-D38 with a 98.9% score, indicating that penton base is recombination origin, so is noted as "P70". Predicted data show that L1 is 100% identical with HAdV-D37, while L2 is a recombination of HAdV-D30 and HAdV-D37 with a 100% score. Since hexon is apparently a recombination of HAdV-D30 and HAdV-D37, it is considered novel, and the assignment is H70. The fiber knob is highly similar to HAdV-D29, with 95.8% identity; therefore, it is F29. Therefore, the final name assignment is P70/H70/F29.

**HAdV-D100.** SBA predicted that HAdV-D100 has HAdV-D48 penton base with 96.7% identity, hexon HAdV-D17 with 98.5% identity. The fiber is 90% similar with HAdV-D30, and D49. The epitope level analysis showed HVR1 with 100.0% identity with HAdV-D38 and 97% identity with HAdV-D48 for RGD, proving that penton base is a recombination of HAdV-D38 and HAdV-D48. The hexon gene's loop regions L1 and L2 showed 99% similarity to HAdV-D17. The fiber knob score with HAdV-D49 and HAdV-D30 is 97.0%. HadV-D30 and -D49 cross react at hemagglutination inhibition with high sequence similarity (Madisch et al., 2005). IBA predicted a 0.0% divergence with HAdV-D38 for HVR1, HAdV-D48 for RGD. The precent divergence for L1 and L2 is <0.7, which is below a cutoff of 5.8% and 3.2% respectively with HAdV-D17. The

fiber knob region showed 1.02% and 1.2% divergence scores with HAdV-D49 and HAdV-D30, respectively. Thus, the final assigned genotype is P100/H17/F30.

## 2.3 Phylogeny Based Analysis (PBA)

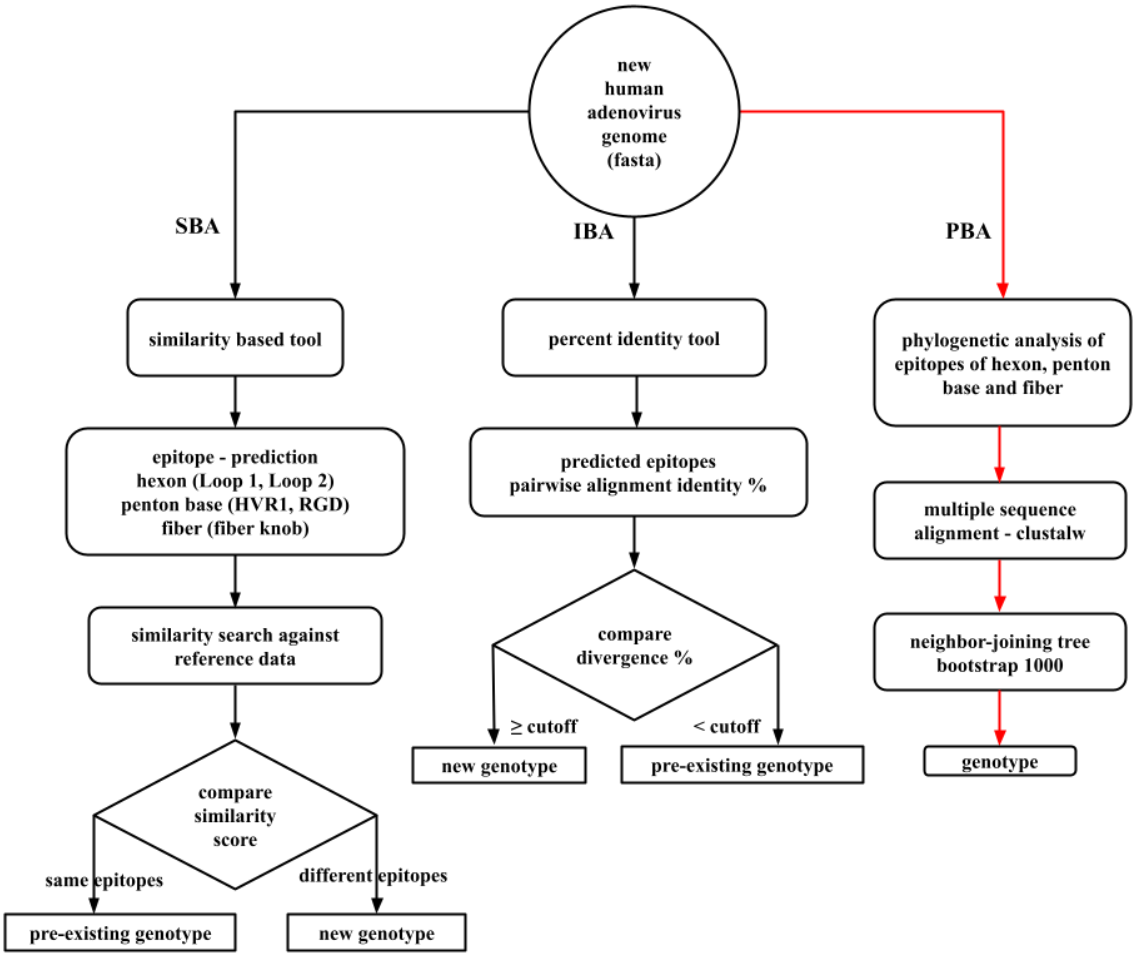
### *2.3.1 Background*

The correct identification of adenovirus isolated from human diseases had been complicated, in some instances, by contradictory typing using the then "state-of-the-art" serological methods. These protocols and reagents only allowed for the assay of two epitopes, derived from the major viral capsid, hexon, and fiber (PHF) (Gahery-Segard et al., 1998; Madisch et al., 2005). In particular, one serological method, hemagglutination, targeted the fiber protein and was limited by reagents and inconclusive reactions. In the DNA sequencing era, particularly the genome sequencing era, a proposal to use whole genome data to distinguish, identify, and type human adenoviruses was proposed in 2011 (Seto et al., 2011). This followed two novel successful applications identifying, characterizing, and typing two novel adenoviral pathogens, HAdV-53 and -55 (Walsh et al., 2009, 2010). The genome data provided for sequence comparisons as well as phylogenetic determinations. These methods were also used by other investigators in the adenovirus research community (Ishiko et al., 2008; Kaneko et al., 2011; Aoki et al., 2008; Zhou, 2012; Gonzalez, 2014; C. M. Robinson, 2011). The new types are referred to as "genotypes" due to their genomic characterization and differences. In particular, the current accepted algorithm is to characterize the three major capsid proteins, the penton base (P), hexon (H), and fiber (F) proteins. The distinct recombination in penton base, hexon, and fiber (PHF) open reading frame (ORF) denominated new genotypes in human

adenoviruses. The closest reference genotype clustered in the corresponding phylogenetic tree is assigned to each ORF (Figure 15–Figure 22).

### *2.3.2 Materials and Methods*

Phylogenetic analysis is performed using MEGA version 7.0.14 (Saitou et al., 1987; Tamura et al., 2004; Kumar, 2016). The distance-based phylogenetic trees are constructed based on the nucleotide sequences of major capsid proteins and epitopes using the neighbor-joining method. Bootstrap values were estimated with 1000 replications for reliability of individual nodes in the tree (Saitou et al., 1987; Tamura et al., 2004; Kumar, 2016), with a bootstrap value of 70 considered robust or "meaningful." Multiple sequence alignments (MSAs) are generated using a ClustalW algorithm.




Note: An Unknown genome is analyzed at the hypervariable regions, derived from, and designated penton base (HVR1 and RGD), hexon (L1 and L2), and fiber (knob). The phylogenetic analysis and subsequent trees show putative evolutionary relationships, using known genotypes as reference.

**Figure 14: Flow Chart for the Phylogenetic Based Method for Characterizing the Three Capsid Proteins of HAdV**

### *Phylogenetic Tree Building*

The phylogenetic trees were constructed using MEGA software. The evolutionary distances were computed using the Maximum Composite Likelihood method (Tamura et

al., 2007) and are in units of the number of base substitutions per site. The distances from each sequence to all others were calculated and stored in a matrix. The trees were constructed from the distance matrix using a specific tree-building algorithm Neighbor-Joining method (Saitou et al., 1987). The analysis involved 52 nucleotide sequences. All positions containing gaps and missing data were eliminated. The phylogenetic trees are constructed with 1000 bootstrap iterations. The scale at the bottom of the tree (Figure 15 through Figure 22) represents 0.1 nucleotides per site in the MSA.

### *2.3.3 Results*

MEGA generated neighbor-joining (NJ) phylogenetic trees of the HAdV whole genomes and the capsid genes along with their discriminatory hypervariable regions, penton base (HVR1 and RGD loops), hexon (L1 and L2), and fiber (knob region) with the nucleotide (Figure 15 through Figure 22). The phylogenetic tree of the hexon genes (Figure 15) shows closely related type pairs that subclade with each other: HAdV-D15 and HAdV-D29, B11 and B35, B3 and B7, D13 and D37, and D39 and D43. Phylogenetic analysis of hexon's hypervariable regions, L1 (Figure 16A) and L2 (Figure 16B) confirm the close clades as well.

### *2.3.4 Discussion*

**HAdV-D53.** The HAdV-D53 SBA results indicate that its penton base is identical with HAdV-D37, its hexon is 98% similar to HAdV-D22, and its fiber is 100% identical with HAdV-D8. Therefore, the predicted genotype is P37/H22/F8. Of the penton base, the HVR1 and RGD sequences showed 0.0% divergence to HAdV-D37, indicating that the penton base is identical with HAdV-D37. With regards to the hexon, the

hypervariable regions L1, L2, predicts 0.78% and 0.0% divergence score with HAdV-D22The fiber knob shows 0.0% divergence with HAdV-D8, which indicates it is identical to HAdV-D8. Therefore, the predicted genotype is P37/H22/F8. The phylogenetic tree of penton base clustered with HAdV-D37 with 100% bootstrap value (Figure 15A). Hexon, L1, and L2 nucleotide phylogenetic tree shows the HAdV-D53 clustering with HAdV-D22 with 100% (Figure 18A, Figure 19A, and Figure 20A) bootstrap values. Fiber and fiber knob trees showed a 100% boot scan (Figure 21A and Figure 22A) value with HAdV-D8. Since all three results are in concurrence, these indicate that HAdV-D53 is a novel genotype that is designated P37/H22/F8 in accordance with accepted nomenclature practice.

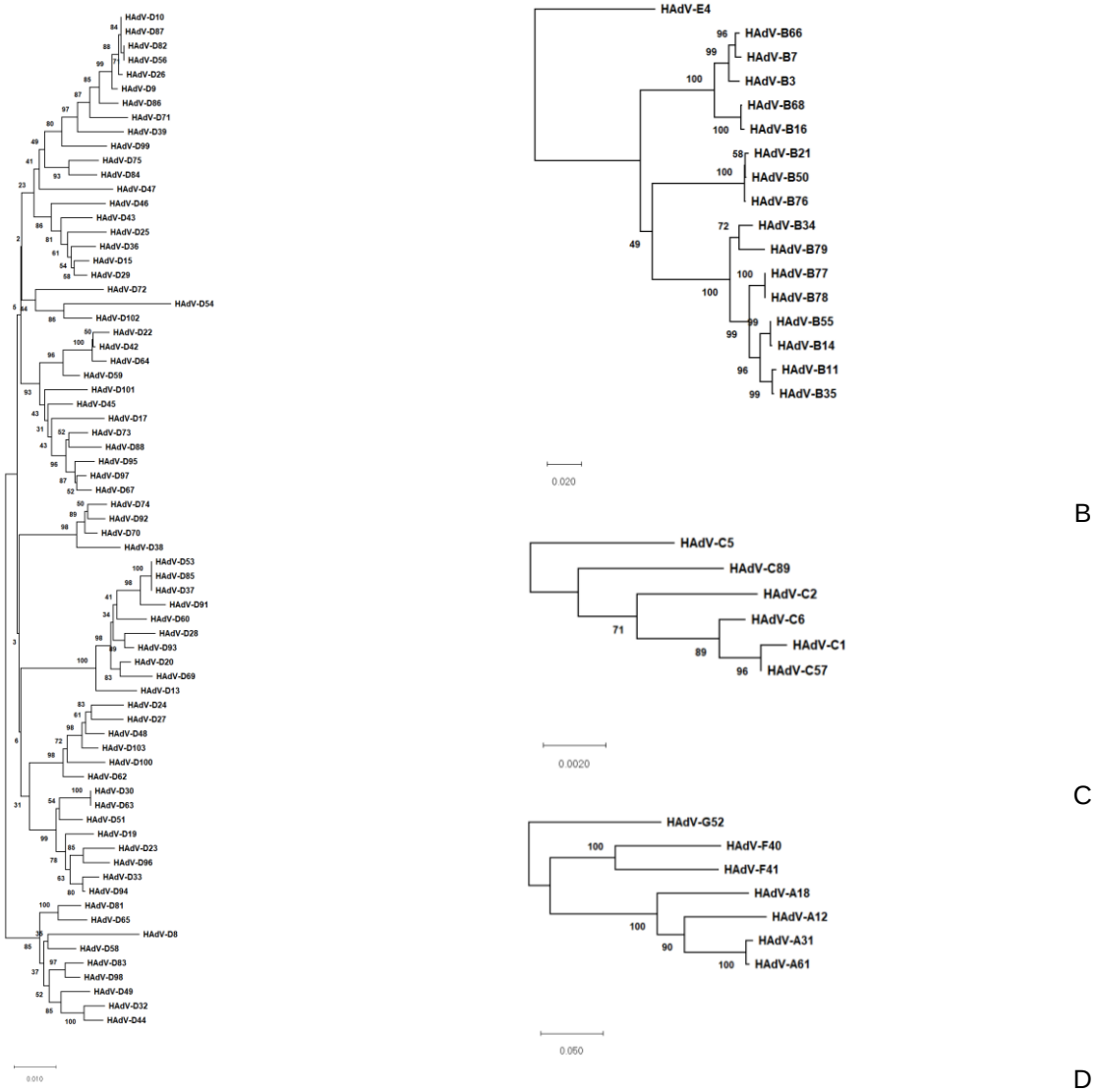
**HAdV-D54.** The BLAST analysis indicates that its penton base is similar with HAdV-D45 with 93.6% score, whereas HVR1 show equivocal identities with HAdV-D25 and HAdV-D10 with 90.0%. The phylogenetic tree of penton base and HVR1 (Figure 15A and Figure 16A) provide the evidence for the blast results. The hexon gene is closest to HAdV-D32 with 91.2% identity, whereas variable loops L1, L2 are closest to HAdV-D32 and HAdV-D9 with 81.6% and 87.7% identities, indicating that HAdV-D54 has a new hexon. Phylogenetic analysis proves there is no significant closely related single genotype for hexon, L1, and L2 (Figure 18A, Figure 19A, and Figure 20A). The fiber gene is closest with HAdV-D8 with 96.7% identity. The phylogenetic tree (Figure 21A) shows HAdV-D54 is closely branched with HAdV-D53 and HAdV-D8 with 100% boot scan value, proving that HAdV-D54 has a recombination penton base and hexon gene and HAdV-D8 fiber gene; hence, the assigned genotype is P54/H54/F8.

**HAdV-B55.** SBA shows that HAdV-B55 has a penton base of HAdV-B14 with 99.6% identity, HAdV-B11 hexon with 98%, and HAdV-B14 fiber with 99.5% identity scores (Walsh et al., 2010). Epitope-level analysis proves the results that HAdV-B55 is a recombination genotype of HAdV-B11 and HAdV -B14. HAdV-B55 is very closely clustered with HAdV-B14 for penton base, HVR1, and RGD (Figure 15B, Figure 16B and Figure 17B) with a significant boot scan value (>80%). Phylogenetic analysis of fiber and fiber knob (Figure 21C and Figure 22C) concur with the SBA and IBA results, proving that HAdV-B55 genotype is P14/H11/F14.

**HAdV-D56.** BLAST analysis of HAdV-D56 predicts that penton base is similar to HAdV-D10, hexon is similar to HAdV-D15, and fiber is similar to HAdV-D9 with 99.8%, 98.7%, and 99.5% identities, respectively. Epitope-level analysis indicates that HVR1 has 0.0% divergence with HAdV-D9 and HAdV-D10. The RGD region of HAdV-D56 showed 0.0% divergence with HAdV-D9; hence, penton base is a recombination of HAdV-D9 and HAdV-D10, proving it is new genotype. The penton base phylogenetic tree shows HAdV-D56 is closely related with HAdV-D10 with a significant boot scan value (Figure 15A, Figure 16A and Figure 17A). L1 is equivocally identical to HAdV-D29 and HAdV-D15 with 0.7% divergence. Similarly, the L2 showed 0.0% divergence with HAdV-D29. The phylogenetic trees of L1 and L2 concur the predictions shown in Figure 19A and Figure 20A. The fiber knob showed 6.9% divergence with HAdV-D9, phylogenetic tree (Figure 22A), HAdV-D56 is branched with HAdV-D9. In conclusion, HAdV-D56 assigned genotype is P56/H15/F29.

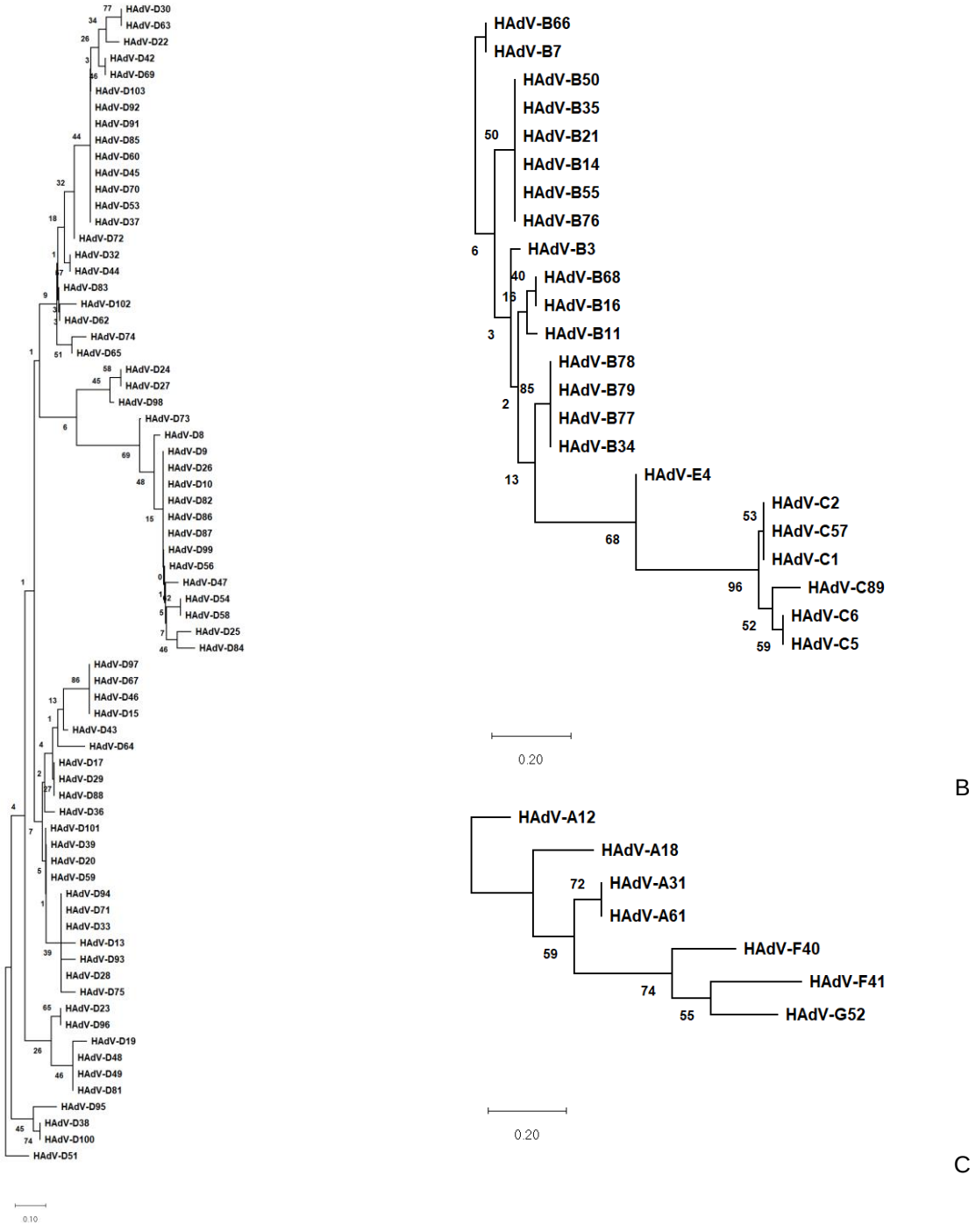
**HAdV-C57.** The SBA indicate that HAdV-C57 penton base is highly similar to HAdV-C1 with 99.8%, hexon is only 89% similar to HAdV-C6, and fiber is again 99.6% similar to HAdV-C6. A more detailed analysis of hexon gene's hypervariable region L1 shows the closest match found in the data set is with HAdV-C6 with only 77.7% identity, and for L2, the closest match found is HAdV-C2 with 92% identity, indicating that hexon is a recombination of HAdV-C6 and HAdV-C2. The phylogenetic tree of L1 (Figure 19C) is clustered with HAdV-C6 in the same clade but distantly aligned. The L2 tree (Figure 20B) HAdV-D57 is clustered with the HAdV-C2 clade, proving that hexon is new. Fiber and fiber knob BLAST analysis shows that HAdV-C57 is highly similar to HAdV-C6 with 99.6% identity. Figure 21B and Figure 22B show the evidence that fiber and fiber knob are closely clustered with HAdV-C6. HAdV-C57 has P1/H57/F6.

**HAdV-D58.** The BLAST search of HAdV-D58 against the reference data set found the closest match with HAdV-D49 with 97.0%, HAdV-D33 with 94.8%, and HAdV-D29 with 98.2% identity for penton base, hexon, and fiber, respectively. Further epitope-level analysis identified HVR1 is equivocally identical with HAdV-D10 and HAdV-D9 with 98.33%. Whereas RGD is 94.6% similar with HAdV HAdV-D44, proving that penton base is a recombinant. Analysis of hexon gene's L1, L2 and constant region showed 88.7% similarity to HAdV-D33, 89.7% with HAdV-D33, and 98.5% with HAdV-D38, respectively, which shows that hexon is a recombinant. Fiber and fiber knob analysis found that it is closest to HAdV-D29 with >98.2% identity. Therefore the assigned genotype for HAdV-D58 is P58/H58/F29.



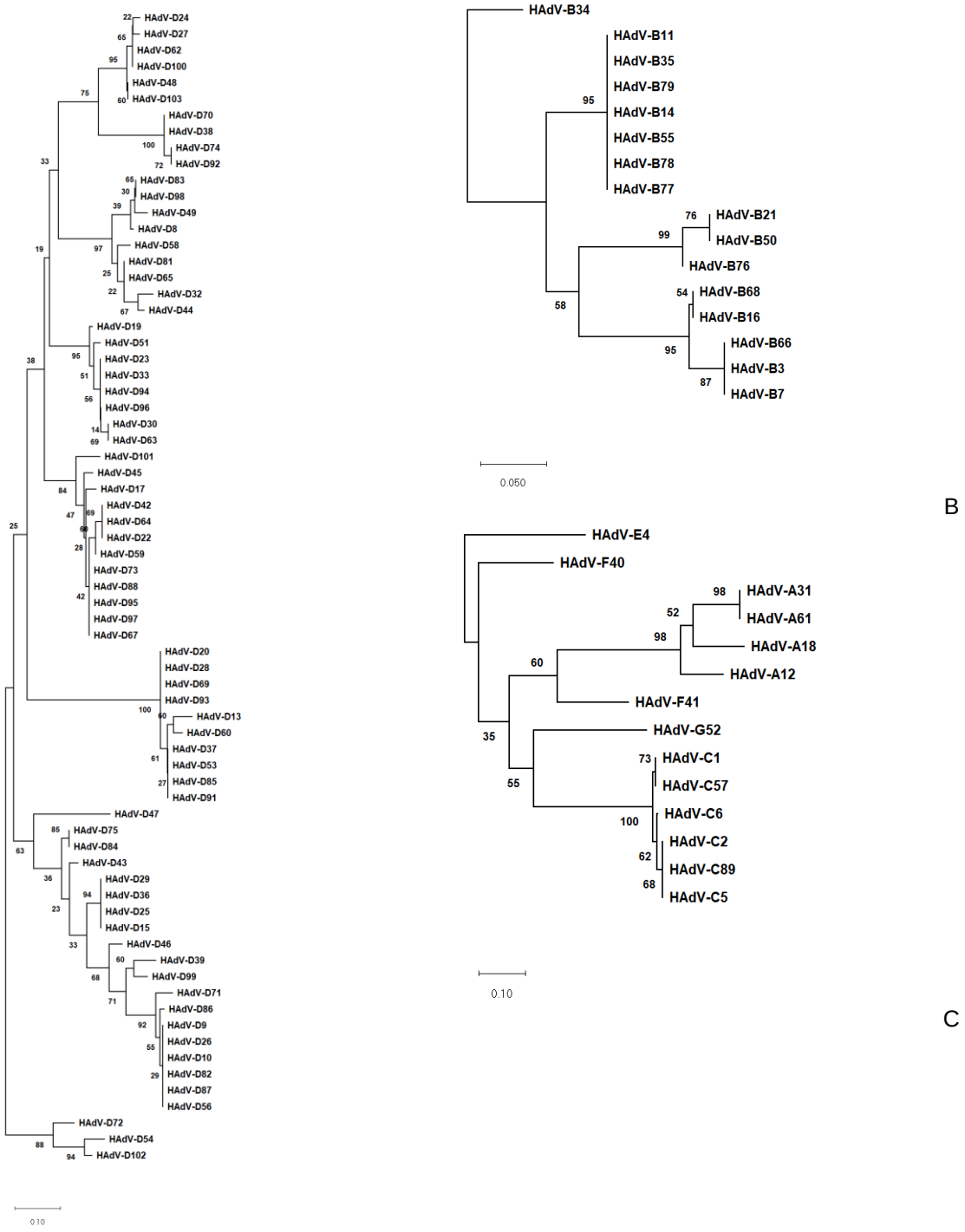

Note: (A) species D; (B) species B and E; (C) species C; (D) species A, F, and G.

**Figure 15: Phylogenetic Tree of Human Adenovirus Penton Base Nucleotide Sequence Constructed by Neighbor-Joining Method with 1000 Bootstrap Replicates**



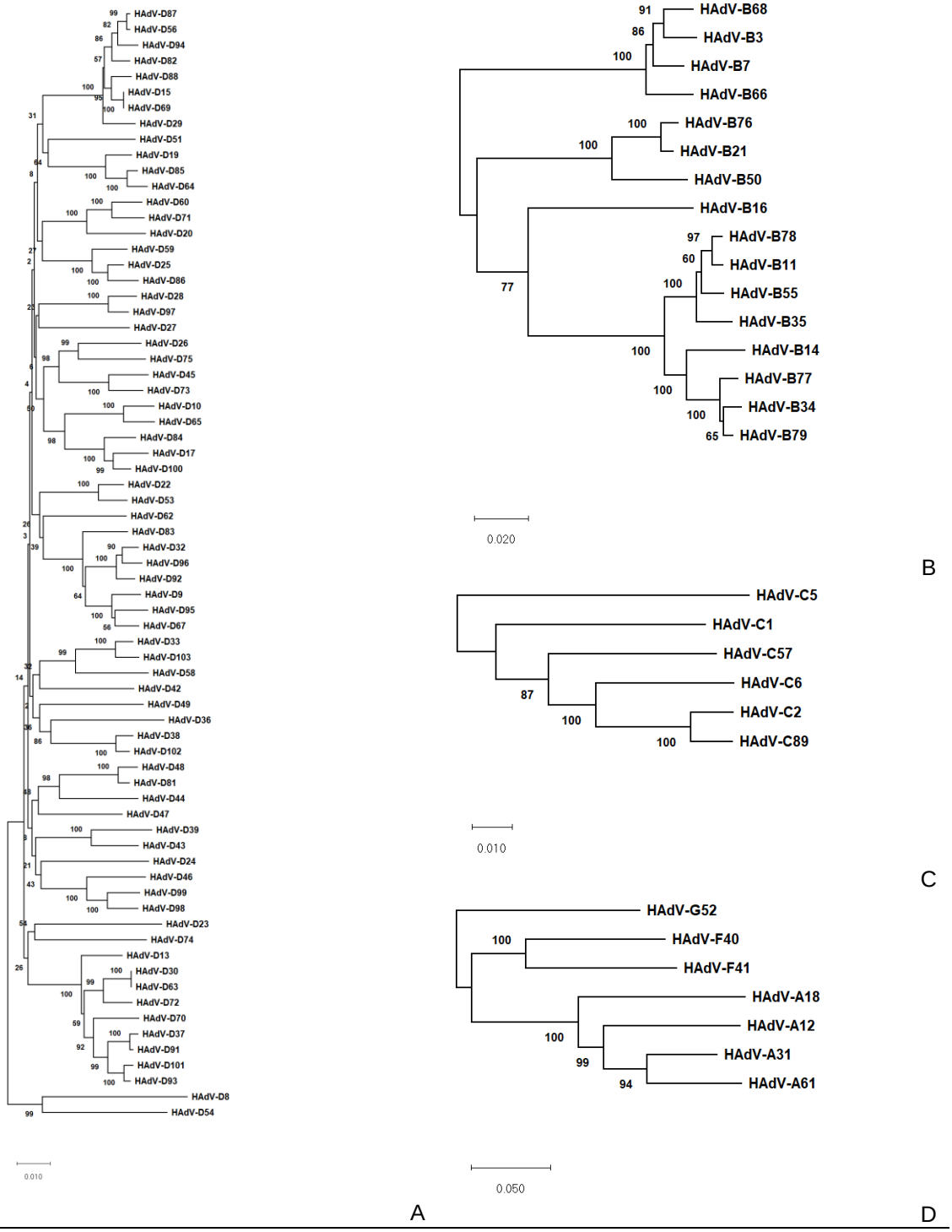

Note: (A) species D; (B) species B, E, and C; (C) species A, F, and G.

**Figure 16: Phylogenetic Tree of Human Adenovirus HVR1 of Penton Base Nucleotide Sequence Constructed by Neighbor-Joining Method with 1000 Bootstrap Replicates**



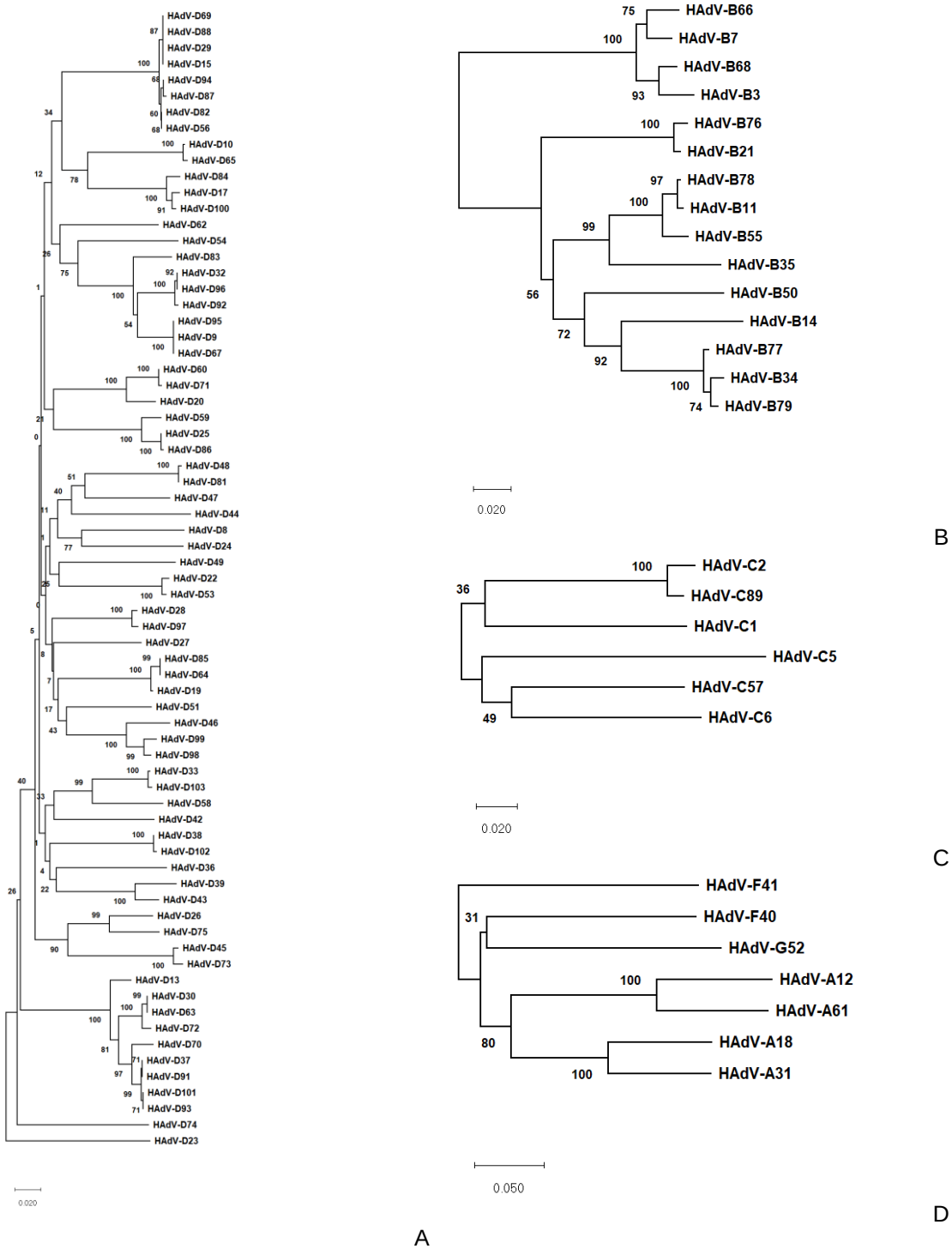

Note: (A) species D; (B) species B; (C) species A, C, E, F, and G.

**Figure 17: Phylogenetic Tree of Human Adenovirus RGD of Penton Base Nucleotide Sequence Constructed by Neighbor-Joining Method with 1000 Bootstrap Replicates**



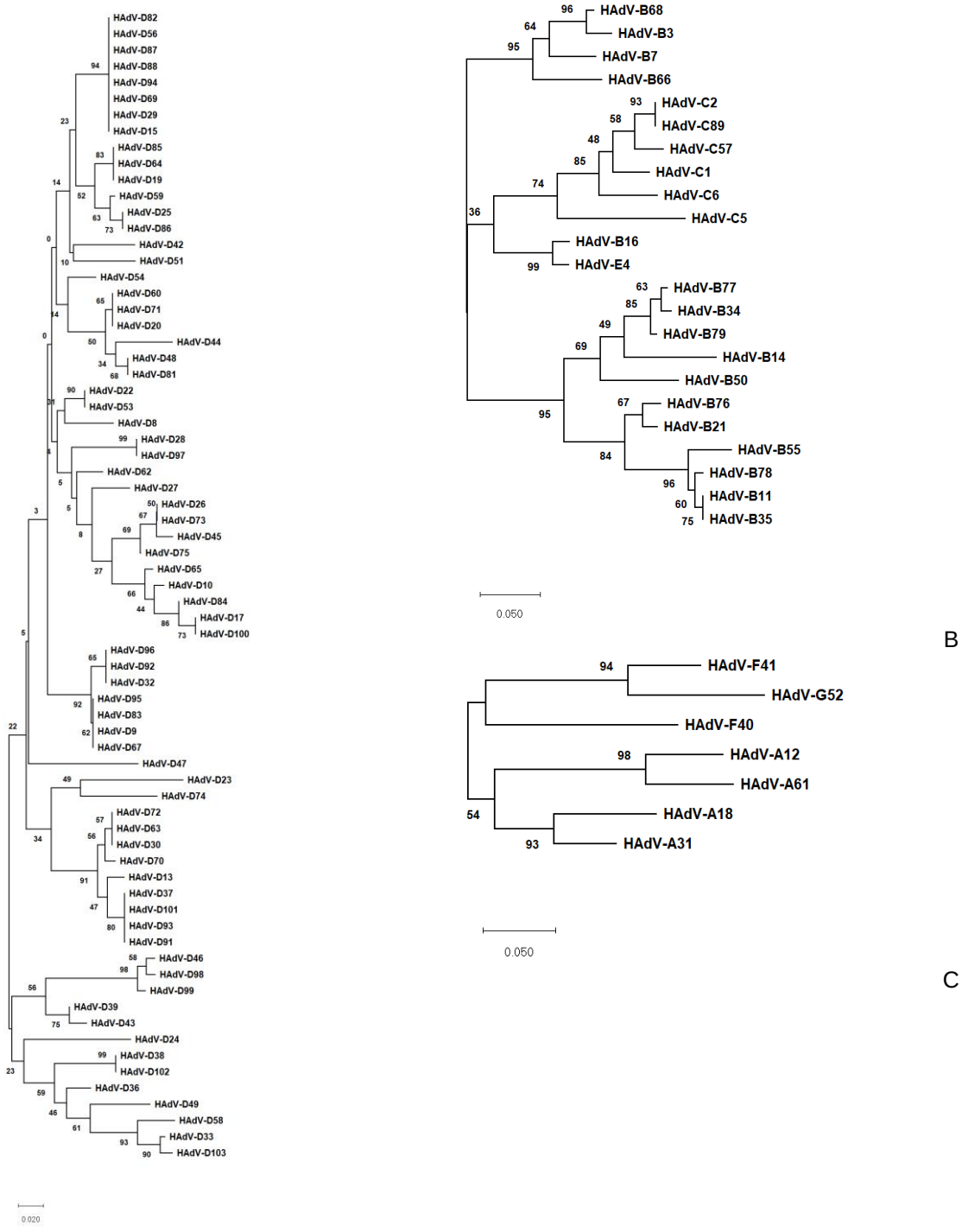

Note: (A) species D; (B) species B; (C) species C; (D) species A, F, and G.

**Figure 18: Phylogenetic Tree of Human Adenovirus Hexon Nucleotide Sequence Constructed by Neighbor-Joining Method with 1000 Bootstrap Replicates**



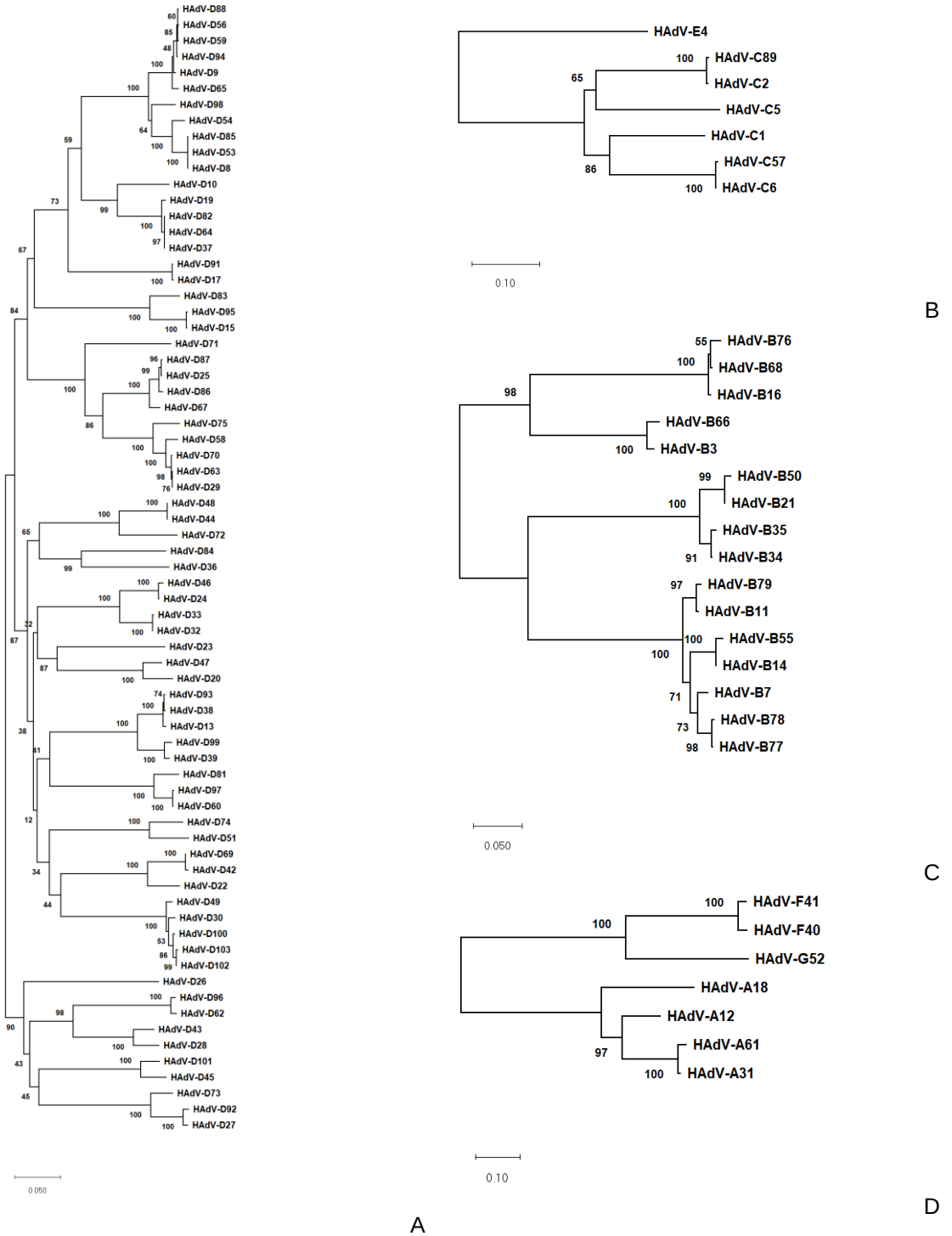

Note: (A) species D; (B) species B; (C) species C; (D) species A, F, and G.

**Figure 19: Phylogenetic Tree of Human Adenovirus L1 of Hexon Nucleotide Sequence Constructed by Neighbor-Joining Method with 1000 Bootstrap Replicates**



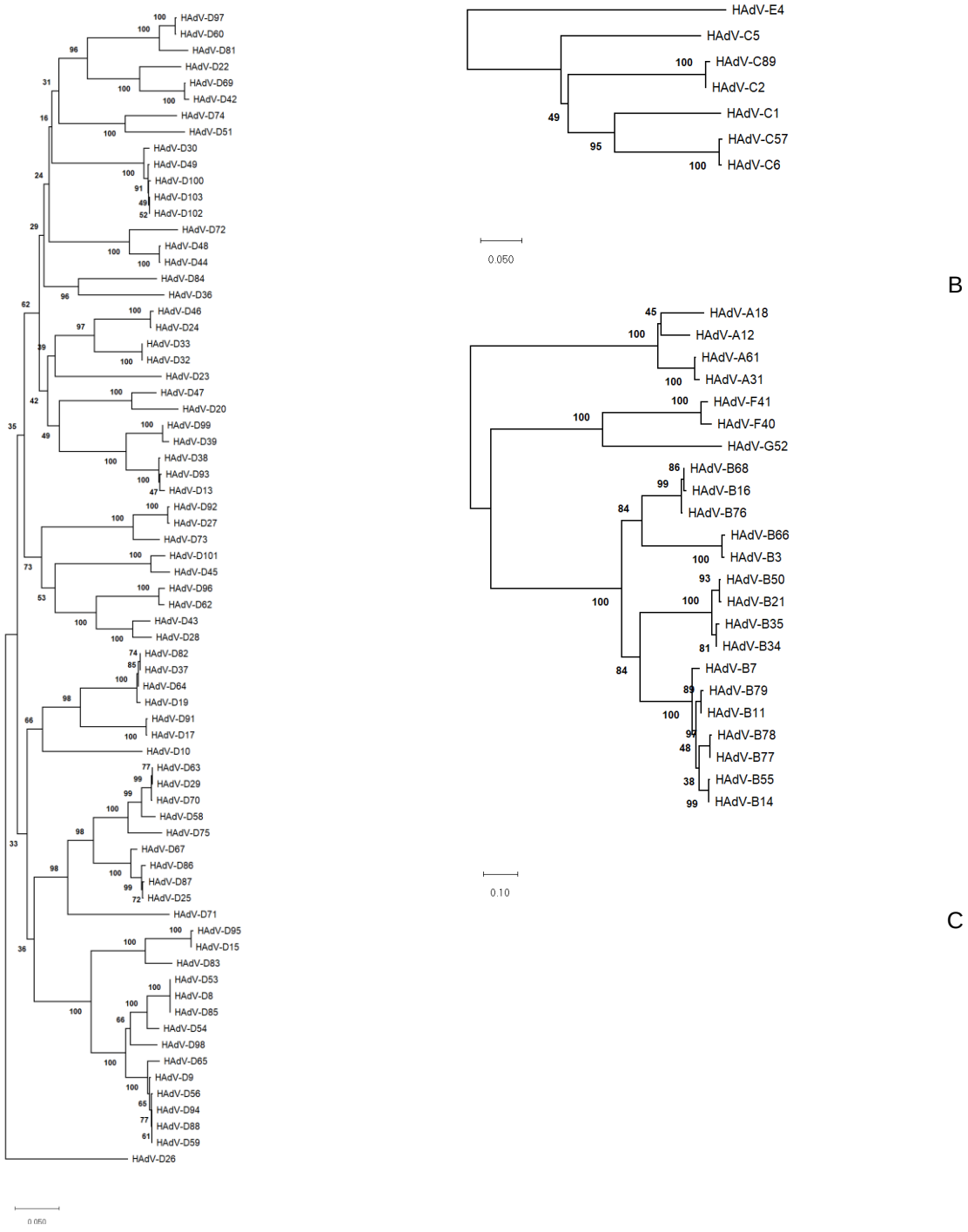

Note: (A) species D; (B) species B, C, and E; (C) species A, F, and G.

**Figure 20: Phylogenetic Tree of Human Adenovirus L2 of Hexon Nucleotide Sequence Constructed by Neighbor-Joining Method with 1000 Bootstrap Replicates**




Note: (A) species D; (B) species C and E; (C) species B; (D) species A, F, and G.

**Figure 21: Phylogenetic Tree of Human Adenovirus Fiber Nucleotide Sequence Constructed by Neighbor-Joining Method with 1000 Bootstrap Replicates**



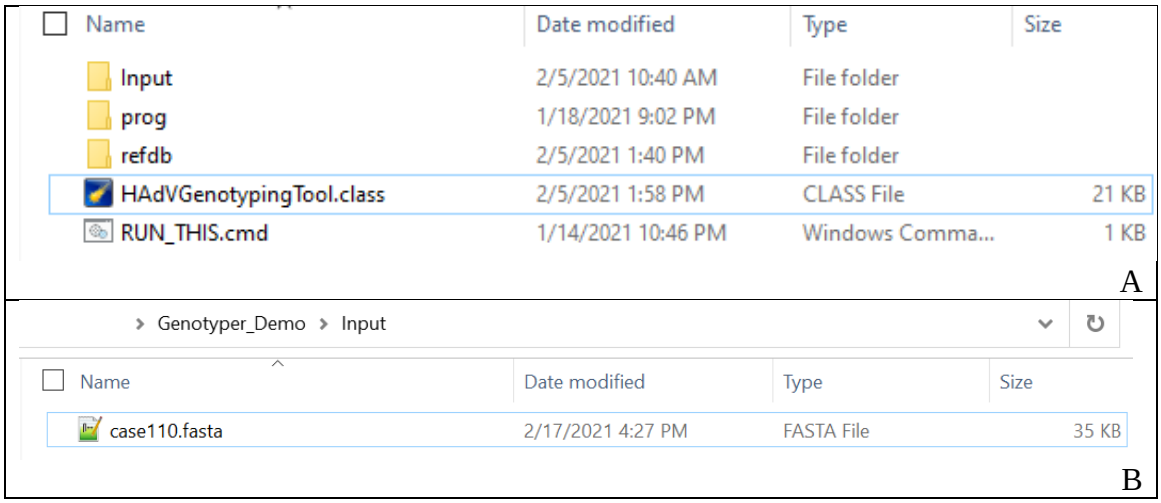

Note: (A) species D; (B) species C and E; (C) species A, B, F, and G.

**Figure 22: Phylogenetic Tree of Human Adenovirus Fiber Knob of Fiber Nucleotide Sequence Constructed by Neighbor-Joining Method with 1000 Bootstrap Replicates**

### *2.3.5 HAdV- Bioinformatics Tool for Genotyping Human Adenoviruses*

A bioinformatics pipeline (http://www.genotyper.info/Genotyper/) is developed to accurately identify and characterize human adenovirus isolates through their genomes. This is a Java-based application. The purpose of this application is to identify the genotype of a given unknown using its whole genome FASTA sequence. The HAdVGenotypingTool pipeline (Figure 23A) contains "Input", "prog", and "refdb" folders, a "HAdVGenotypingTool.class" file, and a "RUN_THIS.cmd" file. The user copies unknown a whole genome (query) FASTA format sequence into the "Input" folder. The "prog" folder contains a dependency application such as BLASTn, EMBOSS, and CLUSTAL executable files. The "refdb" folder contains reference data sets such as whole genome, penton base, hexon, fiber, HVR1, RGD, L1, L2, Con, and fiber knob gene databases of genotypes HAdV 1 through 52. The "RUN_THIS.cmd" file is double-clicked to execute the "HAdVGenotypingTool.class" file to generate the results. The user copies any unknown or query human adenovirus genome into the "Input" folder. An example sequence to illustrate this is retrieved from Genbank (accession number LC215433.1, Human adenovirus DNA, complete genome, strain: case110_20131573) (Figure 23B). The loaded FASTA sequence is processed through a Java application to run a BLAST-based analysis, which aligns to reference data sets and determines the capsid proteins, penton base, hexon, and fiber, and the epitopes HVR1, RGD, L1, L2, Con, and fiber knob. Processed results are displayed in a table format (Table 9), which contains the input file name – "case110.fasta", sequence identifier – "LC215433.1, Human adenovirus DNA, complete genome, strain: case110_20131573", sequence length – "35152", Base

count – “A:8021, G:9936, C:9943, and T:7252, GC% 56.55%, AT% 43.45%”. Identified protein genotypes are as follows: penton base, hexon, and fiber are “HAdV-D37 for all three”. Epitope-level recombination analysis provides that all HVR1, RGD, L1, L2, Con, and fiber knob are identified as “HAdV-D37”, indicating that the loaded sequence is “HAdV-D 37 genotype”. The BLAST results are displayed as text files for reference and further analysis.




Note: (A) The pipeline contains “Input”, “prog”, and “refdb” folders, a “HAdVGenotypingTool.class” file, and a “RUN_THIS.cmd” file. The query genome is copied into the “Input” folder. The “prog” folder contains dependency applications to run BLAST, EMBOSS and CLUSTAL. The “refdb” folder contains reference data sets of whole genome, penton base, hexon, fiber, HVR1, RGD, L1, L2, Con, and fiber knob gene databases of genotypes HAdV 1 through 52. The “RUN_THIS.cmd” file executes the “HAdVGenotypingTool.class” file to generate the results. (B) The “Input” folder contains a HAdV whole genome FASTA sequence “case110.fasta” (Accession number LC215433.1) copied into the “Input” folder.

**Figure 23: Bioinformatics Pipeline Folder Structure of “HAdVGenotypingTool” Tool**

**Table 9**

*Results Output from the* **HAdVGenotypingTool**

| Property | Value |
|---|---|
| Input File Name | case110.fasta |
| Seq. Identifier | LC215433.1 Human adenovirus DNA, complete genome, strain: case110_20131573 |
| Seq. Length | 35152 |
| Base Count | A: 8021 |
| | G: 9936 |
| | C: 9943 |
| | T: 7252 |
| Content | GC: 56.55% |
| | AT: 43.45% |
| Protein | P: 37 |
| | H: 37 |
| | F: 37 |
| Protein level Prediction | HAdV-D37 |
| Recombinants | hvr1: 37 |
| | rgd: 37 |
| | l1: 37 |
| | l2: 37 |
| | con: 37 |
| | fiber knob: 37 |
| Recombinant level Prediction | HAdV-D37 |
| BLAST Files | case110_penton.txt |
| | case110_hexon.txt |
| | case110_fiber.txt |

| Property | Value |
|---|---|
| | case110_pentonbase.fasta |
| | case110_hexon.fasta |
| | case110_fiber.fasta |
| | case110_hvr1.txt |
| | case110_rgd.txt |
| | case110_l1.txt |
| | case110_l2.txt |
| | case110_con.txt |
| | case110_fk.txt |
| | case110_hvr1.fasta |
| | case110_rgd.fasta |
| | case110_l1.fasta |
| | case110_l2.fasta |
| | case110_con.fasta |
| | case110_fk.fasta |
| Percent Identity Analysis Variable Region: HAdV (Divergence%): hvr1: 1.6 rgd: 1.5 l1: 5.8 l2: 3.2 con: 3.2 fk: 9.6 | hvr1: D37 (0.0), D42 (1.59) rgd: D37 (0.0) l1: D37 (0.0) l2: D37 (0.0) con: D37 (0.42), D25 (0.83), D24 (1.25) fk: D37 (0.1), D19 (0.96) |
| Alignment Files | hvr1.aln |
| | rgd.aln |
| | l1.aln |
| | l2.aln |
| | con.aln |
| | fk.aln |

| Property | Value | | |
|---|---|---|---|
| Tree Files | File | Copy To Clipboard | View |
| | hvr1.phb | Copy | View |
| | rgd.phb | Copy | View |
| | l2.phb | Copy | View |
| | con.phb | Copy | View |
| | l1.phb | Copy | View |
| | fk.phb | Copy | View |

Note: The HAdV sequence "case110.fasta" is characterized and results are displayed as follows: Input file name, sequence identifier, sequence length, base content, GC%, AT%, matched genotypes for capsid proteins and identified recombinations at epitope level and BLAST results are displayed in the results page of the application.

The pipeline creates an "Output" folder for the results, which contains a folder for BLAST results of capsid proteins and epitopes and whole genome sequence, another folder for percent identity scores, and a folder for phylogenetic analysis. The BLAST folder contains the results of nucleotide BLAST results for penton base, hexon, and fiber, and epitopes HVR1, RGD, L1, L2, Con, and fiber knob. SBA extracts corresponding nucleotide sequences of "case110" based on BLAST alignment. The extracted sequences are used for further pairwise alignment analysis to find the percent identity, and to do the multiple sequence alignment for phylogenetic analysis, which is used to construct the phylogenetic trees. Pairwise alignment scores in IBA-level analysis (Table 10) show that "case110" is HAdV-D37 genotype. The pipeline folder "PercentIdentity" has the detailed results of individual epitopes for reference.

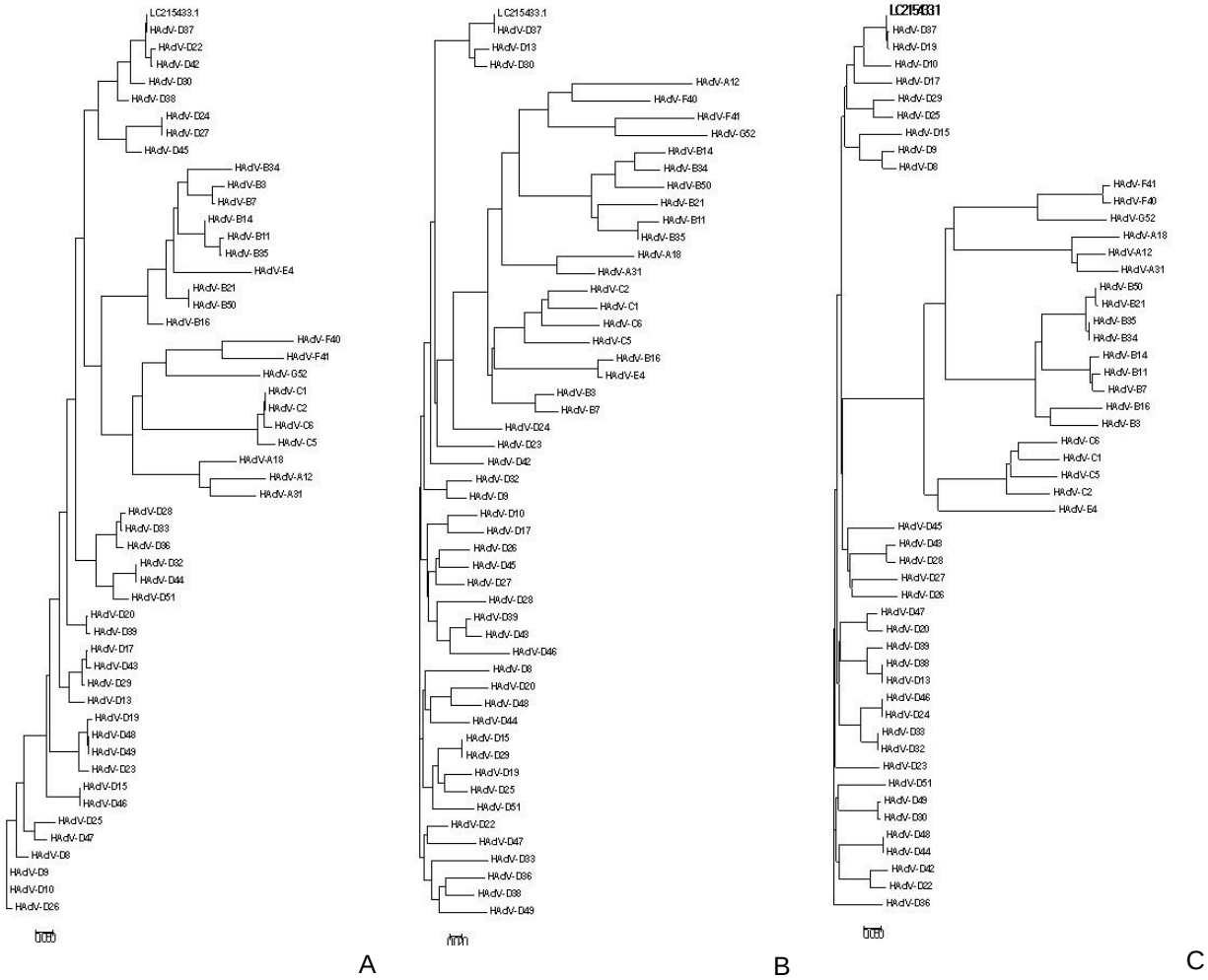
**Table 10**

*Pairwise Alignment Scores of the HAdV Epitopes, HVR1, RGD, L1, L2, Con and Fiber Knob*

| Epitope | First Closest Genotype |
|---|---|
| HVR1 | D37(0.0) |
| RGD | D37(0.0) |
| L1 | D37(0.0) |
| L2 | D37(0.0) |
| Con | D37(0.42) |
| fiber knob | D37(0.1) |

Note: These epitopes are used to determine the recombination in penton base, hexon, and fiber. Percent divergences score along with first closest genotype is listed for "case110" indicating that it belongs to HAdV-D37 genotype.

The PBA results can be accessed from the "CLWIO" folder from the pipeline. This folder contains an individual epitope's multiple sequence alignment and phylogenetic tree, which are generated by comparing the "case110" epitope sequence with a reference data set derived from the HAdV 1 to 52 genotypes. Figure 24 shows the phylogenetic trees of HVR1, L1, and fiber knob. LC215433.1 is very closely related to HADV-D37 in the phylogenetic tree of HVR1 (Figure 24A), L1 (Figure 24B), and fiber knob (Figure 24C) confirms the results.




Note: (A) HVR1 phylogenetic tree' (B) L1 phylogenetic tree; (C) Fiber knob phylogenetic tree of "case110" with HAdV genotypes 1 to 52 references.

**Figure 24: Phylogenetic Trees of "case110" with Reference Data Set**

### *2.3.6 Recombination Sites Identifier*

Genome recombination is an evolutionary process in human adenoviruses, producing new genotypes (Walsh et al., 2009). It usually occurs between strains of the same species, likely due to sequence similarity and cell tropism. Recombination rates vary across different HAdV species; for example, HAdV-D types are more likely to recombine as compared to HAdV-B types (Robinson et al., 2013). Homologous recombination is the critical mechanism for genome diversity (Szekvolgyi et al., 2010; White, 2011; Marques-Bonet & Eichler, 2009). Comparison studies of recombination and mutation rates between HAdV-Ds and HAdV-Bs showed that rho/theta is <1 for HAdV-Ds and it is >1 for HAdV-Bs, which indicates recombination is the major diversity factor within species D and mutations play an important role in species B (Robinson et al., 2013). GC-rich sites are associated with increased genomic stability and relative resistance to homologous recombination (Gruss et al., 1991).

A bioinformatics pipeline has been developed to identify the potential recombination pairs within a given query whole genome sequence. HAdV-D genomes have the highest GC content (Figure 3). HAdV-D genomes are highly conserved and possess among the highest GC content among all HAdV species. Gene-by-gene analysis for GC content revealed that regions of the HAdV-D genome most likely to undergo homologous recombination also contain abrupt reductions in GC content.

The pipeline is written in Java and is portable. A jar file is used to execute the application. A whole genome sequence FASTA file is copied to the “input” directory and the user runs the application. The genome sequence is split into 15mer sequences

(Figure 25), the base count is calculated to deduce the GC% and AT % (Figure 26). The number of G and C is divided by the total number of bases in 15mer and multiplied by 100 to get the GC%. The results are written into a tab delimited text file (

```
// this method will find the char percentage if >=61% return Y else N
public static String getCharNum(String str)
{
		String gcatInfo = "";
		String strUp = str.toUpperCase();
		char[] cArray = strUp.toCharArray();
		int freqA = 0, freqT = 0, freqG = 0,freqC = 0 ;
		for(int i = 0; i < cArray.length ; i++)
		{
			if(cArray[i]=='A')// looking for 'a' only
			{
				freqA++;
			}
			if(cArray[i]=='T')
			{
				freqT++;
			}
			if (cArray[i]=='G')
			{
				freqG++;
			}
			if (cArray[i]=='C')
			{
				freqC++;
			}
		}
		int at =  freqA + freqT;
		int gc =  freqG + freqC;
		float percentAT = (at * 100)/15f;
		float percentGC = (gc * 100)/15f;
		String atCheck = "";
		String gcCheck = "";
		if(percentAT >= 61.6)
		{
			atCheck = "Y";
		}
		else
		{
			atCheck = "N";
		}
		if(percentGC >= 61.6)
		{
			gcCheck = "Y";
		}
		else
		{
			gcCheck = "N";
		}
gcatInfo = (15 + "\t" + gc + "\t" + at + "\t" +  gcCheck + "\t" + atCheck);

		return gcatInfo;
}
```

Note: GC% and AT %. GC% and AT % are important because high GC content is associated with genome stability and resistance to recombination. For a given nucleotide sequence the method calculates the GC and AT percentages and filters if the % > 61.6% or not.

**Figure 26: A Java Method to Determine Base Count of a Given Nucleotide Sequence**

Table 11) which has sequence, length, number of GC, AT and checked status of GC% >= 61.6% and AT% >= 61.6 “Y” or “N”.

```
// this method takes wholegenome string and gives 15mers
public static String[] get15mer(String s, int chunkSize)
{
        int arraySize = (int) Math.ceil((double) s.length() / chunkSize);
        String[] returnArray = new String[arraySize];
        int index = 0;
        for(int i = 0; i < s.length(); i = i+chunkSize)
        {
                if(s.length() - i < chunkSize)
                {
                        returnArray[index++] = s.substring(i);
                }
                else
                {
                        returnArray[index++] = s.substring(i, i+chunkSize);
                }
        }
        return returnArray;
}
```

Note: This method generates 15mers from a given whole genome FASTA sequence. These are used as subunits to determine GC/AT transition zone for 30mers, 45mers and 60mers.

**Figure 25: Bioinformatics pipeline to identify recombination hot spots**

```
// this method will find the char percentage if >=61% return Y else N
public static String getCharNum(String str)
{
		String gcatInfo = "";
		String strUp = str.toUpperCase();
		char[] cArray = strUp.toCharArray();
		int freqA = 0, freqT = 0, freqG = 0,freqC = 0 ;
		for(int i = 0; i < cArray.length ; i++)
		{
			if(cArray[i]=='A')// looking for 'a' only
			{
				freqA++;
			}
			if(cArray[i]=='T')
			{
				freqT++;
			}
			if (cArray[i]=='G')
			{
				freqG++;
			}
			if (cArray[i]=='C')
			{
				freqC++;
			}
		}
		int at =  freqA + freqT;
		int gc =  freqG + freqC;
		float percentAT = (at * 100)/15f;
		float percentGC = (gc * 100)/15f;
		String atCheck = "";
		String gcCheck = "";
		if(percentAT >= 61.6)
		{
			atCheck = "Y";
		}
		else
		{
			atCheck = "N";
		}
		if(percentGC >= 61.6)
		{
			gcCheck = "Y";
		}
		else
		{
			gcCheck = "N";
		}
gcatInfo = (15 + "\t" + gc + "\t" + at + "\t" +  gcCheck + "\t" + atCheck);

		return gcatInfo;
}
```

Note: GC% and AT %. GC% and AT % are important because high GC content is associated with genome stability and resistance to recombination. For a given nucleotide sequence the method calculates the GC and AT percentages and filters if the % > 61.6% or not.

**Figure 26: A Java Method to Determine Base Count of a Given Nucleotide Sequence**

**Table 11**

*List of 15mers and their GC, AT %*

| Line Number | Sequence | Sequence Length | Number of GC | Number of AT | GC >= 61.6% | AT >= 61.6% |
|---|---|---|---|---|---|---|
| 1 | CATCATCATAATATA | 15 | 3 | 12 | N | Y |
| 2 | CCCCACAAAGTAAAC | 15 | 7 | 8 | N | N |
| 3 | AAAAGTTAATATGCA | 15 | 3 | 12 | N | Y |
| 4 | AATGAGCTTTTGAAT | 15 | 4 | 11 | N | Y |
| 5 | TTTAACGGTTTTGGG | 15 | 6 | 9 | N | N |
| 6 | GCGGAGCCAACGCTG | 15 | 11 | 4 | Y | N |
| 7 | ATTGGACGAGAAGCG | 15 | 8 | 7 | N | N |
| 8 | GTGATGCAAATAACG | 15 | 6 | 9 | N | N |
| 9 | TCACGACGCACGGCT | 15 | 10 | 5 | Y | N |
| 10 | AACGGCCGGCGCGGA | 15 | 12 | 3 | Y | N |
| 11 | GGCGTGGCCTAGGCC | 15 | 12 | 3 | Y | N |
| 12 | GGAAGCAAGTCGCGG | 15 | 10 | 5 | Y | N |
| 13 | GGCTAATGACGTATA | 15 | 6 | 9 | N | N |

Note: The input whole genome sequence is split into 15mer units. The results have tab delimited files with sequence, length of sequence, number of GCs, number of ATs, and least possible percent (9 out of 15 bases) to determine GC rich or AT rich zones.

Another method, called “getRecombPatterns” (Figure 27), identifies the consecutive sites where GC% rich is followed by AT% moderate and the transition to GC% low to AT% high. All 30mers, 45mers and 60mers are listed in a tab delimited text file (Table 12). Identifying GC-rich/AT-rich or GC-rich/GC-AT moderate/AT-rich sites of 15 bases to 60 bases are more useful in recombination analysis and prototyping of human adenovirus genomes (Robinson et al., 2013).

```
// This method finds the patterns YN NN NY and YN NY
public static void getRecombPatterns( LinkedHashMap<String, String> lhmp, String OpFile){
		String firstPLine = "";	String secondPLine = "";String thirdPLine = "";
		String fourthPLine = "";boolean bCondition1 = false;		boolean bCondition2 = false;
		boolean bCondition3 = false;String strKey, strValue;
		Set<String> set = lhmp.keySet();	Iterator<String> iter = set.iterator();
		try{
			FileWriter fw = new FileWriter(OpFile, false);
			BufferedWriter bw = new BufferedWriter(fw);
			String hdr = "LineNumber" + "\t" + "NucleotideNum"+ "\t" + "Sequence" + "\t" + "SequenceLength" +
"\t" + "NumberOfGC" + "\t" + "NumberOfAT" + "\t" + "GC >= 61.6%" + "\t" + "AT >= 61.6%";
			bw.write(hdr);bw.newLine();
			while(iter.hasNext()){
				strKey = iter.next().toString();strValue = lhmp.get(strKey).toString();
				String [] strTemp = strValue.split("\t");String sComp = strTemp[6] +  strTemp[7];
				if(sComp.equals("YN")){
					firstPLine = strValue;
					bCondition1 = true;bCondition2 = false;bCondition3 = false;}
				else if(sComp.equals("NN")){
					if(bCondition1 == true){
						if(bCondition3 == true){
						bCondition1 = false;
						bCondition2 = false;bCondition3 = false;}
						else if(bCondition2 == true){
							thirdPLine = strValue;	bCondition3 = true;}
						else{
							secondPLine = strValue;
							bCondition2 = true;}}}
				else if(sComp.equals("NY")){
					if(bCondition3 == true){
						fourthPLine = strValue;
						bw.write(firstPLine);bw.newLine();
						bw.write(secondPLine);bw.newLine();
						bw.write(thirdPLine);bw.newLine();
						bw.write(fourthPLine);bw.newLine();
						bw.write("--------------------------------\n");}
					else if(bCondition2 == true){
						thirdPLine = strValue;
						bw.write(firstPLine);bw.newLine();
						bw.write(secondPLine);bw.newLine();
						bw.write(thirdPLine);bw.newLine();
						bw.write("--------------------------------\n");}
					else if(bCondition1 == true){
						secondPLine = strValue;
						bw.write(firstPLine);bw.newLine();
						bw.write(secondPLine);bw.newLine();
						bw.write("--------------------------------\n");}

					bCondition1 = false;bCondition2 = false;bCondition3 = false;}
				else{
					bCondition1 = false;bCondition2 = false;bCondition3 = false;}}

			bw.close();}
```

Note: Patterns with 30-45 nucleotide length are parsed and written in output.

**Figure 27: Method to Identify GC-rich and AT-rich Patterns after Filtering with >= 61.6%**

**Table 12**

*List of Recombination Transition Zones. Identified Recombination GC-rich and AT-rich Sites*

| Line Number | Sequence | Sequence Length | Number of GC | Number of AT | GC >= 61.6% | AT >= 61.6% |
|---|---|---|---|---|---|---|
| 12 | GGAAGCAAGTCGCGG | 15 | 10 | 5 | Y | N |
| 13 | GGCTAATGACGTATA | 15 | 6 | 9 | N | N |

| Line Number | Sequence | Sequence Length | Number of GC | Number of AT | GC >= 61.6% | AT >= 61.6% |
|---|---|---|---|---|---|---|
| 14 | AAAAAGCGGACTTTA | 15 | 5 | 10 | N | Y |
| 37 | TGAGCTCCGCTCCCA | 15 | 10 | 5 | Y | N |
| 38 | AAGTGTGAGAAAAAT | 15 | 4 | 11 | N | Y |
| 72 | AAACGCCTCCTGCGC | 15 | 10 | 5 | Y | N |
| 73 | TCTGTGTTACATGAA | 15 | 5 | 10 | N | Y |
| 89 | CCCAGTGGCGAGAGG | 15 | 11 | 4 | Y | N |
| 90 | CGAGCAGCTGTTGAA | 15 | 8 | 7 | N | N |
| 91 | AAAATTGAGGACTTG | 15 | 5 | 10 | N | Y |
| 95 | CGCCCCAGGAACTAG | 15 | 10 | 5 | Y | N |
| 96 | GCGCAGCTGTGCTTA | 15 | 9 | 6 | N | N |
| 97 | GTCATGTGTAAATAA | 15 | 4 | 11 | N | Y |

Note: Recombination sites 30mers, 45mers and 60mers are predicted by the application.

### *2.3.7 Conclusion*

Proper identification and characterization of new human adenovirus genotypes are very important. Clinical features and whole genome analysis including the major capsid proteins, penton base, hexon, and fiber, along with detailed genomic and phylogenetic analysis of specific capsid epitopes HVR1, RGD, L1, L2, and fiber knob, provide more clarity as to possible recombination of a novel HAdV genotype. The HAdVGenotypingTool results predicted through all three methods—SBA, IBA, and

PBA—of the genotyping tool are in concurrence. As discussed, this reported bioinformatics pipeline is not limited to human adenovirus identification; it has scope for expansion to other adenoviruses as well, especially identifying and characterizing new simian adenovirus.

# CHAPTER 3

# GENOTYPING TOOL FOR *STAPHYLOCOCCUS AUREUS*

## 3.1 Abstract

*Staphylococcus aureus* is the leading cause of skin and soft tissue infections worldwide. Strain typing is critical for understanding the epidemiology in the outbreak investigation of the pathogen. The highly variable *spa* gene provides a sensitive method for distinguishing *S. aureus* isolates. The *spa* gene product is a surface antigen with Ig-binding activity that is involved in immune evasion. *Spa* includes a variable Xr region that has multiple repeats, usually 8 amino acids in length. Repeats within an Xr region may have different amino acid sequences. Over 16,500 variants of the *spa* gene are known to date. Combined with other methods such as multi-locus sequence typing (MLST), *spa* typing provides high resolution strain identification. In this paper, we developed an open-source *spa* typing program that does not requires helper applications. The program can be integrated into bioinformatics pipelines and can analyze complete draft genome sequences.

## 3.2 Introduction

*Staphylococcus aureus* is a gram-positive bacterium causing nosocomial infections (National Nosocomial Infections Surveillance System, 1999), skin and tissue infections, pneumonia, septicemia, and hospital associated infections (HAIs) in humans.

The emergence of methicillin-resistant *S. aureus* (MRSA) (Okuma et al., 2002) has become a major concern in the hospital environment and community settings due to the high mortality of the infections caused by these strains. Thus, understanding and controlling the spread of the infection is of utmost importance.

*S. aureus* is a heterogenous/polymorphic species (Fitzgerald et al., 2001). It does not undergo extensive recombination and is differentiated largely by mutations (Feil et al., 2003). Multilocus enzyme electrophoresis (MLEE) and multilocus sequence typing (MLST) have been used for typing for years (Maiden et al., 1998). However, no single technique is efficient due to different requirements and accumulation of genetic variation. To discriminate types of MRSA accurately and reliably, a single locus DNA-sequencing of the repeat region of the *Staphylococcus* protein A gene (spa) can be used. A polymorphic 24-bp (exceptions 21-bp to 30-bp) variable number tandem repeats (VNTR) between 5' and 3' (Shopsin et al., 1999; Shopsin et al., 2000). For each repeat, Xr (Figure 28) is assigned a numerical code. The numerical code pattern is used to deduce spa type (Figure 29). Spa-gene: s – signal sequence, E, D, A, B, C – IgG binding domains, X – region lacks IgG binding activity and contains repeat region (Xr) and C-Terminal region (Xc). 1095F and 1517R indicate annealing.

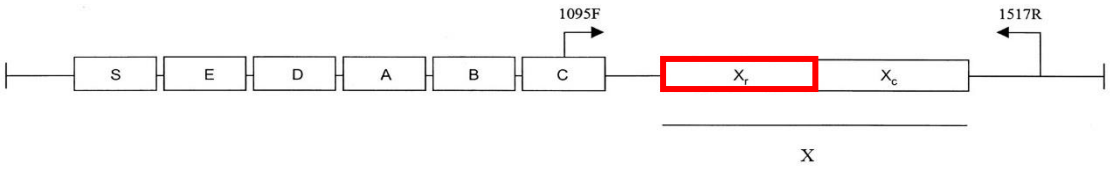



**Figure 28: Scheme of the Spa Gene with Annealing Sites**

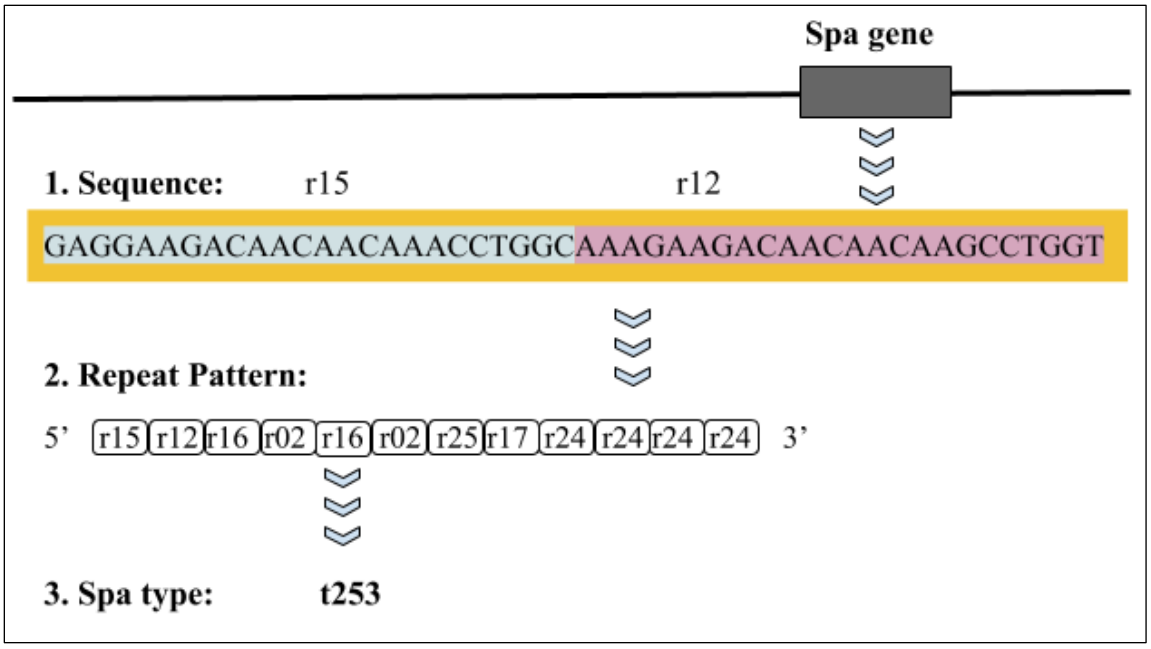



**Figure 29: Spa Typing Schema**

## 3.3 Materials and Methods

The spa sequence contains unique 24bp sequence repeat patterns flanked between 5' and 3' primers that determine the type.

### *3.3.1 Bioinformatics Pipeline – Java Application*

Based on market research that has been conducted for available spa typing applications, there are two freely available *spa* typing resources: the DNAGear program (w3.ualg.pt/~hshah/DNAGear/) and the web-based spa typing service at the Center for Genetic Epidemiology in Lyngby, Denmark (https://cge.cbs.dtu.dk/services/spatyper/). Due to lack of capabilities such as batch processing, identifying, and extracting the gene as an input to find the type, a Java-based bioinformatics application called SpaTyper has been developed (Figure 30, Figure 31, and Figure 32). Inputs for the program are Xr repeat sequences (Note: There are 816 repeat sequences available till date. Repeat sequences r1 to r715 are listed.

Figure 33 and *spa* type (Table 13 definitions (available from http://spa.ridom.de/), and a *spa* gene sequence or a set of draft genomes contig sequences in FASTA format. SpaTyper uses pattern matching (Figure 31) to identify the *spa* gene Xr region (Figure 30) and compare it to known gene variants.

```
public static String getSpa(String pattern, String spaRepeatsIP)
      {
            String spa = "";

            LinkedHashMap<String, String> lhmSpaRpts = new LinkedHashMap<String,
String>();
            lhmSpaRpts = getLhmSpas(spaRepeatsIP);

            boolean idExists = false;

            idExists = lhmSpaRpts.containsKey(pattern);
            if(idExists)
            {
                  spa = lhmSpaRpts.get(pattern);
            }
            else
            {
                  spa = "Novel_Spa";
            }
            return spa;
      }
```

Note: Identifies the repeats that have conserved 5' and 3' primers. If matched, assign the type; otherwise, it is a novel spa type.

**Figure 30: Java Application – HAdVGenotypingTool**

```
public static String patternfinder(String contigID, ArrayList<String> Al, String
rptsIP, String seqContig)
	{
		String opStr = "";
		TreeMap<Integer, String> startTm = new TreeMap<Integer, String>();
		TreeMap<Integer, String> endTm = new TreeMap<Integer, String>();
		TreeMap<Integer, String> allTm = new TreeMap<Integer, String>();

		for(int i = 0; i < Al.size(); i++)
		{
			String strTemp [] = Al.get(i).split("\t");
			int startKey = Integer.parseInt(strTemp[3]) ;
			int endKey = Integer.parseInt(strTemp[4]);
			String val = Al.get(i);
			startTm.put(startKey, val);
			endTm.put(endKey, val);
			allTm.put(Integer.parseInt(strTemp[3]), val);
		}
		opStr = getSeqStr(startTm, endTm, allTm, rptsIP, seqContig);

		return contigID + "\t" + opStr;
	}
```

Note: Identifying the 3' and 5' regions within the sequence.

**Figure 31: Pattern Finder Method**

```
public static String getCompleteSpaSeq(String rptSeqInfo, LinkedHashMap<String,
String> Lhm5, LinkedHashMap<String, String> Lhm3)
      {
            . . .
            while(iter3.hasNext())
            {
                  String key3 = iter3.next();
                  if(key3.equals("3'Missing"))
                  ...
                  else
                  {
                        String Temp [] = key3.split("\t");
                        int ThreeStart = Integer.parseInt(Temp[3]);
                        int ThreeEnd = Integer.parseInt(Temp[4]);
                        int diff3 = ThreeStart - endIndx;
                        if(diff3 >= 18 && diff3 <= 1000)
                        {
                              threeDifTm.put(diff3, key3
                        ...
            Set<String> set5 = Lhm5.keySet();
            Iterator<String> iter5 = set5.iterator();
            while(iter5.hasNext())
            {
                  String key5 = iter5.next();
                  if(key5.equals("5'Missing"))
                  . . .
            }
            if(fiveDifTm.size() > 0)
            {
                  ...
op = SpaGene5PrStartIndx + "\t" + "NA" +  ":" + prime5Str + "\t" + "NA" + "\t" +
minDif5 + "\t" + "NA";
                        . . .
                        if(threeDifTm.size() > 0)
                        {
                              . . .
                              op = SpaGene5PrStartIndx + "\t" + "NA" +  ":" +
prime5Str + "\t" + "NA" + "\t" + minDif5 + "\t" + "NA";
                        }
                  . . .
            }
            else
            {
                  if(threeDifTm.size() > 0)
                  { . .                        . . .
                        op = SpaGene5PrStartIndx  + "\t" + SpaGene3PrEndIndx +
":" + "NA" + "\t" + prime3Str + "\t" + "NA" + "\t" + minDif3;
                  }
                  else
                  {
                        op = SpaGene5PrStartIndx + "\t" + "NA" +  ":" + "NA" +
"\t" + "NA" + "\t" + "NA" + "\t" + "NA";

            return op;
      }
```

Note: Repeat patterns are flanked by conserved 5’ and conserved 3’ primers. Extracted repeats are compared to existing types to determine the novel type.

**Figure 32: Complete Spa Sequence Extraction Method**

>r01
GAGGAAGACAACAACAAGCCTAGC
>r02
AAAGAAGACAACAAAAAACCTGGC
>r03
GAGGAAGACAATAACAAACCTGGT
>r04
GAGGAAGACAATAACAAGCCTGGT
>r05
AAAGAAGACAACAAAAAGCCTGGC
…
>r711
AAAGAAGACGGCAACAAAACTGGC
>r712
GAGGAAGACAAAAACAAACCTGGT
>r713
AAAGAGGATAACAACAAACCTGGT
>r714
AAAGAAGACAACAACAAGCCCGGC
>r715
AAAGAAGACGGCAACAAACCTCGC

Note: There are 816 repeat sequences available till date. Repeat sequences r1 to r715 are listed.

**Figure 33: Repeat Sequences FASTA Sequences**

**Table 13**

*Patterns and Spa Types*

| Spa Type | Repeats Pattern | Number of Repeats |
|---|---|---|
| t0001 | 26-30-17-34-17-20-17-12-17-16 | 10 |
| t0002 | 26-23-17-34-17-20-17-12-17-16 | 10 |
| t0003 | 26-17-20-17-12-17-17-16 | 8 |
| t0004 | 09-02-16-13-13-17-34-16-34 | 9 |
| t0005 | 26-23-13-23-31-05-17-25-17-25-16-28 | 12 |
| t16213 | 08-16-02-16-02-25-17-24-02-16-02-25-17-24 | 14 |
| t16214 | 08-16-02-16-34-13-130-34-34 | 9 |
| t16215 | 26-23-17-34-17-20-17-20-17-12-17-17-16 | 13 |
| t16216 | 07-23-23-21-16-34-33-625 | 8 |
| t16213 | 08-16-02-16-02-25-17-24-02-16-02-25-17-24 | 14 |

Note: There are 19817 types available till date. Number of repeats per pattern and specific spa types are listed.

The program searches both forward and reverse strands of the DNA sequence for matching known *spa* gene repeats. The application identifies at least one of the four (ACAACAAAA, ACACCAAAA, GCAACAAAA, GCACCAAAA) available 9 bp 5' conserved sequences 250 bp upstream of the first repeat, and finds at least one of the two (TACATGTCGT, TATATGTCGT) available 10 bp 3' conserved sequences downstream 250 bp of the last repeat. If a complete Xr region (one or more repeats flanked by 5' and 3' conserved sequences) (Figure 29) is found, the application compares it to known *spa* types. A new pattern results in a novel type.

### *3.3.2 Portable Spa Typing Program Run Procedure*

Computer Requirements: Current version of JRE should be installed in the system.

Run With arguments: Program takes two input arguments from the user:

1. Sequence files Dir
2. Output directory path

Java –jar SpaTyper.jar Sequences_dir_path output_dir_path

Run Without arguments: Direct the command prompt to the folder where the SpaTyper application is downloaded. Ex: "PortableSTP" folder.

Java –jar SpaTyper.jar

Output:

Identified 3', 5', repeats and spa information files are in the output folder. Final tab delimited summary is created as "Summary.txt".

## 3.4 Results

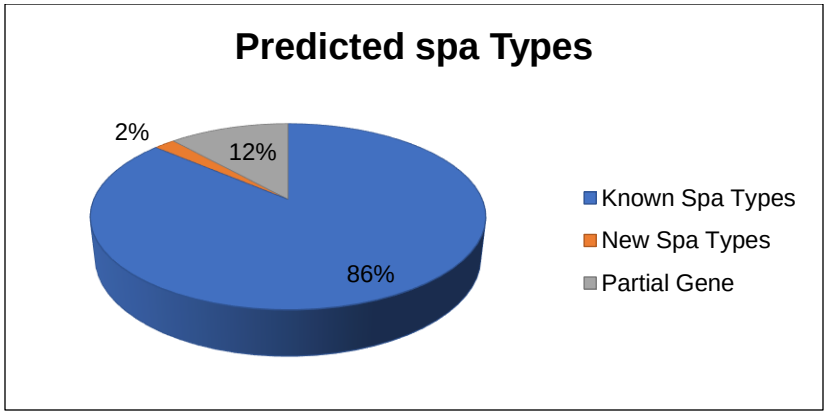
SpaTyper identifies known and novel *spa* types. It flags partial *spa* Xr regions lacking the 5' and/or 3' conserved sequence as being partial type. It detects novel repeats, identified as ~24 bp sequences adjacent to known repeats. The program writes a tab-delimited text report, which includes the pattern of repeats, the *spa* type, and the DNA sequence of the Xr repeat region (Table 14). On a standard Windows computer, SpaTyper processes a draft *S. aureus* genome in about one minute. Using the application, we typed 145 strains, of which there were 125 known *spa* types, 3 new *spa* types, and 17 partial genes (Note: The algorithm shows that out of 100% results, SpaTyper identified 86% known strain types, 2% novel strain types, and 12% partial genes.

Figure 34). The next generation sequencing technology generates split contigs. The application identifies partial genes because of split contig.

**Table 14**

*SpaTyper Results*

| Sample | Contig | Spa type | Pattern |
|---|---|---|---|
| IDXXX84 | Rev_ IDXXX84_contig00008 | t487 | 08-16-02-16-34-13-17-34-16-34-34 |
| IDXXX85 | Rev_ IDXXX85_contig00005 | t3643 | 04-17-34-17-32-23-24-24-24 |
| IDXXX86 | Rev_ IDXXX86_contig00006 | t5413 | 11-19-10-21-17-34-24-34-22-25-25 |
| IDXXX86 | Rev_ IDXXX86_contig00019 | Incomplete (3' Missing) | 11 |
| IDXXX93 | IDXXX93_contig00006 | Novel Spa | 07-23-12-34-34-12-12-34-12-12-23-02-12-23 |
| IDXXX96 | Rev_ IDXXX96_contig00065 | Incomplete (5' Missing) | 24-25 |




Note: The algorithm shows that out of 100% results, SpaTyper identified 86% known strain types, 2% novel strain types, and 12% partial genes.

**Figure 34: SpaTyper Predicted Results Venn Diagram**

The algorithm shows that out of 100% results, SpaTyper identified 86% known types, 2% novel types, and 12% partial genes.

The final tab delimited output shows the id, contig with orientation, predicted spa type and identified pattern (Table 14).

Repeats are assigned a numerical code. There are 816 repeat sequences available in Ridom SpaServer (https://spaserver.ridom.de/). Repeat sequences r1 to r715 are listed (Note: There are 816 repeat sequences available till date. Repeat sequences r1 to r715 are listed.

Figure 33).

The spa type is deduced from the order of specific repeats. There are 19817 spa types available in spa server. The number of repeats per pattern and specific spa types are listed (Table 13).

## 3.5 Conclusion

The SpaTyper (http://www.genotyper.info/Spatyper/) is used to identify the strains on clinical samples, and it worked for large sample sets. The program is open source, platform independent, and it does not need a helper application to extract a spa gene. It allows batch processing and can be easily integrated into a large pipeline. The SpaTyper identifies incomplete Xr regions, while DNAGear reports a spa type for partial Xr regions and CGE SpaTyper reports “N/A” for the spa type. SpaTyper has the advantage of reporting novel repeats, while DNAGear truncates the Xr region and CGE web service reports an “N/A” as spa type. The user does not need to provide a *spa* gene sequence in a single contig in the sense orientation as input and it allows batch processing of draft genomes.

# APPENDIX A.

# CD8 AND CD4 EPITOPE PREDICTIONS IN RV144: NO STRONG EVIDENCE OF A T-CELL DRIVEN SIEVE EFFECT IN HIV-1 BREAKTHROUGH SEQUENCES FROM TRIAL PARTICIPANTS

## Introduction

As HAdVs are not considered an especially dangerous human pathogen causing high mortality rates, epitope prediction was applied to the Human Immunodeficiency Virus (HIV) problem. In other words, a similar protocol and tool was used to develop, script, and validate a viral pathogen typing protocol based on genome sequences. In the case of HIV, a double-blind clinical vaccine trial dataset has been analyzed and published.

HIV was identified in the 1980s (CDC, 1981). It uses white blood cells called T-helper cells (CD4 cells) for its replication. The CD4 count refers to the number of T-helper cells per cubic milliliter of blood. So, if a person's CD4 count drops below 200, they are said to have Acquired Immune Deficiency Syndrome (AIDS). If patients are untreated, it may take 10–15 years to develop AIDS. According to the Centers for Disease Control and Prevention (CDC), more than 1.2 million people in the United States are living with HIV infection and 50,000 new infections occur every year. In 2013 in the United States, 1,194,039 people were diagnosed with AIDS.

There are many different strains of HIV, with HIV-1 the most common found worldwide. HIV-2 is found mainly in Western Africa, India, and Europe. HIV is an icosahedral enveloped retrovirus (120 nm sized). It carries two copies of single-stranded RNA, with a length of 9749bp (https://www.Aids.gov). Virus reverse-transcribed DNA codes for the structural proteins, Gag, Env, Polymerase, regulatory proteins Tat and Rev, and the accessory proteins Nef, Vpu, Vif, and Vpr. Epitopes on Env recognize hosT-cell receptors, which allows entry into a hosT-cell for replication (Mutch et. al., 1994). The first phase, 2–4 weeks after infection, is the acute phase, with the CD4 count dropping until the host immune system activates. The second chronic phase, lasting 10 to 12 years if not treated, results in the CD4 count dropping to below 200 MM3 and causing AIDS. Opportunistic infections such as cancers, pneumonia, and tuberculosis result in death.

Currently, there are no approved therapeutic vaccines for HIV. Thirty clinical trials are ongoing, with NIAID funding (https://niaid.nih.gov). The Military HIV Research Program (MHRP) started the RV144 clinical trial in 2003 in Thailand to test the mosaic vaccine effect using ALVAC-HIV and AIDSVAX B/E gp120 vaccines. Results showed 31% estimated protection against HIV-1 infection for vaccine recipients compared to the placebo group.

**Materials and Methods**

From RV144 participants who became HIV-1-infected during the RV144 trial (diagnosed between June 14, 2004, and February 12, 2009), near full-length HIV-1 genomes were sequenced from single RNA templates corresponding to plasma samples collected at the time of HIV-1 diagnosis (Genbank accession numbers JX446645–

JX448316). Sequences from these 110 subjects, who were infected with CRF01_AE viruses, were translated to amino acid (AAc) sequences.

The vaccine inserts sequences corresponded to the two lab strains of HIV-1 subtype B and two CRF01_AE viruses isolated in Thailand in 1990 and 1992. ALVAC-HIV canary pox prime [vCP1521] is a chimeric construct that concatenates Gag and Pro of HIV-1 subtype B (strain LAI), with gp120 of CRF01_AE (strain 92TH023) fused to a 28-AAc-long segment of the transmembrane-anchoring portion of gp41 from HIV-1-B strain LAI (HXB2 position AAc 684:711 of HXB2 gp160). The AIDSVAX B/E boost was composed of two gp120 proteins with N-terminal truncations (HIV-1 protein started at AAc 42 of HXB2 gp160): one protein was HIV-1 subtype B (strain MN) and one was HIV-1 CRF01_AE (strain CM244).

High-resolution typing of class I and II HLA was performed by DNA sequence-based typing (SBT) and by the sequence-specific oligonucleotide probe (SSOP) method. CD4 9mer and CD8 15mer epitopes were predicted for HAL and HLA-DR alleles (Hurley et al., 2006). Epitopes were predicted using the translated AAc sequences corresponding to Gag, Pol, Env, and Nef based on each subject's class I and II HLA alleles. In parallel, epitopes were predicted in the vaccine insert sequences using all of the HLA/HLA-DR alleles in the cohort of 110 RV144 subjects. For each subject, all unique HLA-peptide binding pairs were retained. Subject-specific epitopes were matched to the vaccine-derived epitopes—epitopes were matched when the subject- and vaccine-derived peptides (with the same HXB2 positions) had at least 67% AAc identity or a maximum of 3 mismatches for a 9-mer.

**Results**

Predicted Gag epitope repertoires were smaller in the vaccine than in placebo recipients ($p=0.019$). After comparing participant-derived epitopes to corresponding epitopes in the RV144 vaccine, the proportion of epitopes that could be matched differed depending on the protein conservation (only 36% of epitopes in Env vs 84%–91% in Gag/Pol/Nef for CD8 predicted epitopes) or on vaccine insert subtype (55% against CRF01_AE vs 7% against subtype B). To compare predicted epitopes to the vaccine, we analyzed predicted binding affinity and evolutionary distance measurements. Comparisons between the vaccine and placebo arm did not reveal robust evidence for a T-cell driven sieve effect, although some differences were noted in Env-V2 ($0.022 \leq p\text{-value} \leq 0.231$). The paucity of CD8 T-cell responses identified following RV144 vaccination, with no evidence for V2 specificity, considered together both with the association of decreased infection risk in RV 144 participants with V-specific antibody responses and a V2 sieve effect, led us to hypothesize that this sieve effect was not T-cell-specific. Overall, our results did not reveal a strong differential impact of vaccine-induced T-cell responses among breakthrough infections in RV144 participants.

OPEN ACCESS Freely available online

PLOS ONE

# CD8 and CD4 Epitope Predictions in RV144: No Strong Evidence of a T-Cell Driven Sieve Effect in HIV-1 Breakthrough Sequences from Trial Participants

**Kalpana Dommaraju[1,2], Gustavo Kijak[1,2], Jonathan M. Carlson[3], Brendan B. Larsen[4], Sodsai Tovanabutra[1,2], Dan E. Geraghty[5], Wenjie Deng[4], Brandon S. Maust[4], Paul T. Edlefsen[5], Eric Sanders-Buell[1,2], Silvia Ratto-Kim[1,2], Mark S. deSouza[6], Supachai Rerks-Ngarm[7], Sorachai Nitayaphan[8], Punnee Pitisuttithum[9], Jaranit Kaewkungwal[9], Robert J. O'Connell[6], Merlin L. Robb[1,2], Nelson L. Michael[1], James I. Mullins[4], Jerome H. Kim[1], Morgane Rolland[1,2]***

**1** US Military HIV Research Program, Walter Reed Army Institute of Research, Silver Spring, Maryland, United States of America, **2** Henry Jackson Foundation, Bethesda, Maryland, United States of America, **3** Microsoft Research, Los Angeles, California, United States of America, **4** Department of Microbiology, University of Washington, Seattle, Washington, United States of America, **5** Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America, **6** Armed Forces Research Institute of Medical Sciences, Bangkok, Thailand, **7** Thai Ministry of Public Health, Nonthaburi, Thailand, **8** Royal Thai Army Component, Armed Forces Research Institute of Medical Sciences, Bangkok, Thailand, **9** Faculty of Tropical Medicine, Mahidol University, Bangkok, Thailand

## Abstract

The modest protection afforded by the RV144 vaccine offers an opportunity to evaluate its mechanisms of protection. Differences between HIV-1 breakthrough viruses from vaccine and placebo recipients can be attributed to the RV144 vaccine as this was a randomized and double-blinded trial. CD8 and CD4 T cell epitope repertoires were predicted in HIV-1 proteomes from 110 RV144 participants. Predicted Gag epitope repertoires were smaller in vaccine than in placebo recipients ($p = 0.019$). After comparing participant-derived epitopes to corresponding epitopes in the RV144 vaccine, the proportion of epitopes that could be matched differed depending on the protein conservation (only 36% of epitopes in Env vs 84–91% in Gag/Pol/Nef for CD8 predicted epitopes) or on vaccine insert subtype (55% against CRF01_AE vs 7% against subtype B). To compare predicted epitopes to the vaccine, we analyzed predicted binding affinity and evolutionary distance measurements. Comparisons between the vaccine and placebo arm did not reveal robust evidence for a T cell driven sieve effect, although some differences were noted in Env-V2 ($0.022 \leq p\text{-value} \leq 0.231$). The paucity of CD8 T cell responses identified following RV144 vaccination, with no evidence for V2 specificity, considered together both with the association of decreased infection risk in RV 144 participants with V-specific antibody responses and a V2 sieve effect, lead us to hypothesize that this sieve effect was not T cell specific. Overall, our results did not reveal a strong differential impact of vaccine-induced T cell responses among breakthrough infections in RV144 participants.

**Citation:** Dommaraju K, Kijak G, Carlson JM, Larsen BB, Tovanabutra S, et al. (2014) CD8 and CD4 Epitope Predictions in RV144: No Strong Evidence of a T-Cell Driven Sieve Effect in HIV-1 Breakthrough Sequences from Trial Participants. PLoS ONE 9(10): e111334. doi:10.1371/journal.pone.0111334

**Editor:** Paul A. Goepfert, University of Alabama, United States of America

**Received** May 16, 2014; **Accepted** September 23, 2014; **Published** October 28, 2014

**Copyright:** © 2014 Dommaraju et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability:** The authors confirm that all data underlying the findings are fully available without restriction. All relevant data are within the paper and its Supporting Information files.

**Funding:** This work was supported in part by an Interagency Agreement Y1-AI-2642-12 between the U.S. Army Medical Research and Material Command and the National Institutes of Allergy and Infectious Diseases and by a cooperative agreement (W81XWH-07-2-0067) between the Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc., and the U.S. Department of Defense. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist. KD, GK, ST, ESB, SRK, MLR, NLM, JHK, and MR are employees of the US Military HIV Research Program (MHRP). The opinions expressed herein are those of the authors and should not be construed as official or representing the views of the US Department of Defense or the Department of the Army. JMC is an employee of Microsoft Research. This does not alter the authors' adherence to PLOS ONE policies on sharing data and materials.

* Email: mrolland@hivresearch.org

## Introduction

The RV144 trial showed modest efficacy in preventing HIV-1 infection with an estimated 31% reduction in HIV-1 infections in the vaccine arm (modified intent to treat population, $p = 0.04$) [1]. Among the 16,402 adult Thai participants, 125 HIV-1 infections were diagnosed following enrollment. We sequenced HIV-1 near full-length genomes from plasma samples collected at the time of HIV-1 diagnosis in 121 subjects [2]. Phylogenetic analyses showed that 110 of these infections were caused by CRF01_AE viruses.

A genetic analysis spurred by the identification of a correlate of risk of infection linked to Env-V2, focused on this Env segment and identified two signatures that distinguished viruses from vaccine and placebo recipients [2]. Viruses derived from vaccine recipients could be differentiated from those from placebo recipients at Env positions 169 and 181, which are contact sites for V2-specific antibodies including some derived from RV144 participants [3]. Genetic signatures in V2, together with the identification of binding antibodies directed against V2 as a

correlate of risk of HIV-1 infection [4], suggest that anti-V2 antibodies may have played a role in the protection conferred by the RV144 vaccine.

It is plausible that Env-V2 was not the sole viral determinant impacted by vaccine-induced immune responses. One possible path to explore is the potential role of T-cell mediated immune responses. Analyses performed on samples collected 6 months after the final immunization showed Gag or Env IFN-γ ELISpot responses in 20% of vaccinees vs 7% of placebo recipients [1]. Intracellular cytokine staining assays showed no difference between vaccine and placebo recipients for CD8 responses (Gag CD8: 7% of responders; Env CD8: 11% and 14% of responders among vaccine and placebo recipients, respectively) or for Gag CD4 responses (1% vs 0% of responders), while Env CD4 responses were significantly more frequent in vaccine than in placebo recipients (34% vs 4%, respectively) [1].

To obtain insights on the impact of T cell immunity on founder HIV-1 sequences, potential CD8 and CD4 epitopes can be predicted *in silico* based on sequence motifs matched by class I and II HLA alleles [5,6]. In this study, we performed CD8 and CD4 epitope predictions based on each subject's HLA genotype and HIV-1 proteome sequence(s) using the same methods as in the analysis of breakthrough infections in the Step/HVTN502 trial [5,6], which we have expanded to include comparisons of epitope predictions based on evolutionary distances and predicted affinity binding. We analyzed subject-derived epitope predictions as a function of the epitopes predicted in the RV144 vaccine inserts to investigate whether we could identify features distinguishing the vaccine and placebo group.

## Materials and Methods

### Ethics Statement

The RV144 clinical vaccine trial was registered with ClinicalTrials.gov and assigned the registration number NCT00223080 (Supporting File S1). The protocol was approved by the ethics committees of the Ministry of Public Health, the Royal Thai Army, Mahidol University, and the Human Subjects Research Review Board of the U.S. Army Medical Research and Materiel Command. It was also independently approved by the World Health Organization and the Joint United Nations Program on HIV/AIDS and by the AIDS Vaccine Research Working Group of the National Institute of Allergy and Infectious Diseases at the National Institutes of Health.

Written informed consent was obtained from all volunteers, who were required to pass a written test of understanding. The consent procedure was approved by the Ethics committees and IRBs listed above.

The RV144 trial was double-blinded and randomized, enrolled 16,402 participants and took place in Thailand between October 2003 and September 2009; the results of the trial were reported by Rerks-Ngarm and colleagues [1], and further details on immune correlates of risk of infection were reported by Haynes and colleagues [4].

### HIV-1 sequence data

For RV144 participants who became HIV-1-infected during the RV144 trial (diagnosed between 14 June 2004 and 12 February 2009), HIV-1 near full-length genomes were sequenced from single RNA templates corresponding to plasma samples collected at the time of HIV-1 diagnosis (GenBank accession numbers JX446645–JX448316). Sequences from the 110 subjects who were infected with CRF01_AE viruses were translated to amino acid (AA) sequences, using only sequences with open reading frames.

Vaccine insert sequences corresponded to two lab strains of HIV-1 subtype B and two CRF01_AE viruses isolated in Thailand in 1990 and 1992. The ALVAC-HIV canarypox prime [vCP1521] is a chimeric construct that concatenates *gag* and *pro* of HIV-1 subtype B (strain LAI) with gp120 of CRF01_AE (strain 92TH023) fused to a 28-AA-long segment of the transmembrane-anchoring portion of gp41 HIV-1-B strain LAI (HXB2 position AA 684:711 of HXB2 gp160). The AIDSVAX B/E boost was composed of two gp120 proteins with N-terminal truncations (HIV-1 protein started at AA42 of HXB2 gp160): one protein was HIV-1 subtype B (strain MN) and one was HIV-1 CRF01_AE (strain CM244).

### HLA genotyping

High-resolution typing of class I and II HLA was performed by DNA sequence-based typing (SBT) and by the sequence-specific oligonucleotide probe (SSOP) method, with concordant results. Class I SBT was carried out by PCR amplification and subsequent dye terminator nucleotide sequencing of exons 2 and 3, with ambiguous types being resolved to four digits using the dbMHC SBT interpretation interface [7] (http://www.ncbi.nlm.nih.gov/projects/gv/mhc/). Class II SBT were genotyped in the CLIA/ASHI accredited lab of William Hildebrand at the University of Oklahoma Health Sciences Center using in-house PCR and sequencing methodologies. The entirety of exon 2 was DNA sequenced for all class II loci with additional exons DNA sequenced for DQB1 (exon 3), DQA1 (exon 3), and DPB1 (exons 3 & 4). DNA sequence analysis and HLA allele assignment were performed with Assign-SBT v3.5.1 software (Conexio Genomics). The HLA database for allele assignment was updated with IMGT release 3.0.0 May 5th 2010. Any ambiguous types that remained following DNA Sequence Based Typing were resolved to 4-digits using the PEL-FREEZ UNITRAY SSP, Life Technologies. SSOP was conducted using the LABType SSO Class I HD system (One Lambda, Canoga Park, CA), which is based on Luminex xMAP technology, and results were interpreted using the accompanying HLA Fusion 2.0.0 software. HLA types are reported according to the IMGT/HLA nomenclature (version 3.7.0, http://www.ebi.ac.uk/imgt/hla/ambig.html).

### $CD4^+$ and $CD8^+$ T cell epitope predictions

CTL epitope predictions were done with the NetMHCpan 2.4 Server, which predicts binding of peptides to each subject's HLA genotype using artificial neural networks [8] (http://www.cbs.dtu.dk/services/NetMHCpan/). Predictions were done for 9-mers, because it is the favored length for binding (predictions for other lengths are made from approximations based on 9-mers, and are thus less accurate). CD4 epitope predictions were done with the NetMHCIIpan 2.1 Server, which predicts binding of peptides to each subject's MHC class II HLA-DR alleles using artificial neural networks [9] (http://www.cbs.dtu.dk/services/NetMHCIIpan/). Predicted epitopes were 15-mers, and unique epitopes were retained based on the core 8-mer peptide sequence: between several overlapping peptides with an identical core peptide, the peptide that had the strongest predicted binding affinity was retained.

Predictions are given with IC50 values (in nM), with a threshold of 50 nM for an epitope to be considered a strong binder (SB); weak binders (WB) have a predicted IC50 between 50 and 500 nM.

Epitopes were predicted on the translated AA sequences corresponding to Gag, Pol, Env and Nef based on each subject's class I and II HLA alleles. In parallel, epitopes were predicted in the vaccine insert sequences using all the HLA/HLA-DR alleles in

the cohort of 110 RV144 subjects. For each subject, all unique HLA-peptide binding pairs were retained, leading to some peptides being counted multiple times for each HLA to which it was predicted to bind.

Subject-specific epitopes were matched to the vaccine-derived epitopes – epitopes were matched when the subject- and vaccine-derived peptides (with the same HXB2 positions) had at least 67% AA identity, or a maximum of 3 mismatches for a 9-mer; we consider that 9-mers with 4 or more mismatches cannot be aligned with confidence and our rationale for not aligning such 9-mers is that the sequences are too distant for the peptide to be recognized by a vaccine-elicited response [10]. Matched epitope predictions were analyzed based on the respective binding affinity values of the matched epitopes, and on the evolutionary distance calculated between the matched epitopes. For each matched subject-derived and vaccine-derived epitope pair, the binding affinity value was calculated as the ratio of the binding affinity for the subject-derived epitope to the binding affinity of the vaccine-derived epitope. The evolutionary distance was calculated between the subject-derived and vaccine-derived epitope based on the HIV-specific matrix (HIV-between-10%) developed by Nickle and colleagues [11].

For each subject, summary distances were computed based on matched pairs of predicted epitopes (if there were no predicted epitope or no matched pair, then the distance could not be defined and the subject's information was not used). Wilcoxon rank sum tests (equivalently, the Mann-Whitney test) with exact 2-sided p-values were used to test for a different distribution in summary measures between the vaccine and placebo groups.

### Phylogenetic dependency networks

We used phylogenetic dependency networks, a statistical model of evolution that simultaneously takes into account HIV-1 AA co-variation, linkage disequilibrium among HLA alleles, and shared ancestry in the HIV-1 phylogeny to identify the primary source of selection pressure acting on each HIV codon [12]. For each gene, a maximum likelihood phylogenetic tree was constructed and a model of conditional adaptation was created for the vaccine status and for every HLA gene, amino acid position and state. The null hypothesis is that the observations depend on the phylogenetic tree structure; then, adaptation due to each variable is modeled along the tree by an additive process. Results were adjusted for multiple comparisons, using q values of $\leq 0.2$ with an associated p-value threshold of $\leq 0.05$ (implying a false-positive proportion of 20% among identified associations).

### Association between viral loads and HLA alleles

We analyzed the effect of the HLA (2 digit) genotype on viral loads measured at the time of HIV-1 diagnosis (VL corresponding to the plasma sample sequenced).

L1-regularized linear regression analyses were used to test which features predict VL for each sample. A bootstrap procedure (sample with replacement 1,000 times) was used to estimate a bootstrap support frequency for a given predictor and its effect size on viral loads.

## Results

### Gag epitope repertoires from vaccine recipients were smaller than those of placebo recipients

We performed CD8 and CD4 epitope predictions based on each subject's genotype and HIV-1 genomic sequences derived at the time of HIV-1 diagnosis. To avoid the misidentification of effects that would be due to a phylogenetic difference, we only included the 110 subjects infected with CRF01_AE viruses – see Flowchart on Figure 1. For these 44 vaccine and 66 placebo recipients, the vaccine efficacy was estimated at 34% (95% C.I. = 7.8%, 54.7%) (compared to 31% in the full mITT cohort [1]), thus allowing vaccine/placebo investigations of the impact of the vaccine. Epitope predictions from two subjects in a linked HIV-1 transmission were included since, despite being infected with nearly identical viruses, both subjects had different HLA genotypes and thus non-overlapping predicted epitope repertoires (Figure 2A).

The number of predicted epitopes for each subject depends on both the genotype of the individual and the HIV-1 sequence they were infected with. There was no difference in the distribution of HLA types between the vaccine and placebo groups; then we looked at the epitopes predicted for these HLA alleles. When the vaccine and placebo groups were compared, there was no difference in the number of predicted CD8 epitopes in Pol, Env and Nef: Median number of epitopes in Pro: $n = 5$ (vaccine) and $n = 6$ (placebo), $p = 0.464$; in RT-IN: $n = 116$ (vaccine) and $n = 117.5$ (placebo), $p = 0.707$; in Env: $n = 64$ (vaccine) and $n = 66$ (placebo), $p = 0.238$; in Nef: $n = 25$ (vaccine) and $n = 26.5$ (placebo), $p = 0.408$ (Table 1). In contrast, there were significantly fewer Gag epitopes predicted in sequences from vaccine recipients ($n = 45$) than in those from placebo recipients ($n = 51.5$), $p = 0.019$ (Table 1).

### Env epitopes from RV144 participants poorly matched vaccine-derived epitopes

We can hypothesize that, due to vaccine-engendered escape mutations, fewer epitopes may have a corresponding matched epitope in the vaccine insert sequence in sequences from vaccine recipients than in those from placebo recipients. For each subject, the list of predicted autologous class I and class II epitopes was compared to corresponding epitopes predicted in the vaccine insert strain to identify pairs of matched subject+vaccine epitopes (Schematic representation for class I epitopes in Figure 2B). Considering each vaccine insert sequence separately, we found no difference between vaccine and placebo recipients in the ratio of subject-derived epitopes that could be matched to vaccine-derived epitopes for any of the HIV-1 proteins: $p \geq 0.314$ vs CM244, $p \geq 0.214$ vs LAI (MN for Env), $p \geq 0.540$ vs 92TH023. Since the RV144 vaccine was composed of proteins of different subtype (Subtype B and CRF01_AE; while all subjects evaluated were infected with CRF01_AE) and of proteins that are relatively variable (Env-gp120) or conserved (Gag/Pro), we tested if these factors affected the ability to match predicted autologous epitopes to the predicted epitopes in the inserts. We found that the ratio of matched epitopes differed depending on the protein and on the vaccine reference considered but not on the vaccine/placebo status. The proportion of matched epitopes is much higher in Pol, Gag, or Nef (73 to 91%) than in Env, for which only about a third (36%) of predicted epitopes in subject-derived sequences could be matched to epitopes identified in the vaccine inserts (Table 2). In addition, there were subtype-specific differences, with significantly more subject-derived epitopes matched against the CM244 (also CRF01) than against the MN (or LAI; both subtype B) vaccine insert. For example, an average of 55% of Env predicted epitopes were matched against CM244 compared to 7% matched against MN ($p < 0.0001$) whether vaccine or placebo recipients are considered. The limited number of matched epitopes is due to the high diversity between Env sequences, hence the large distance between specific strains, which is amplified when attempting cross-subtype epitope matching.

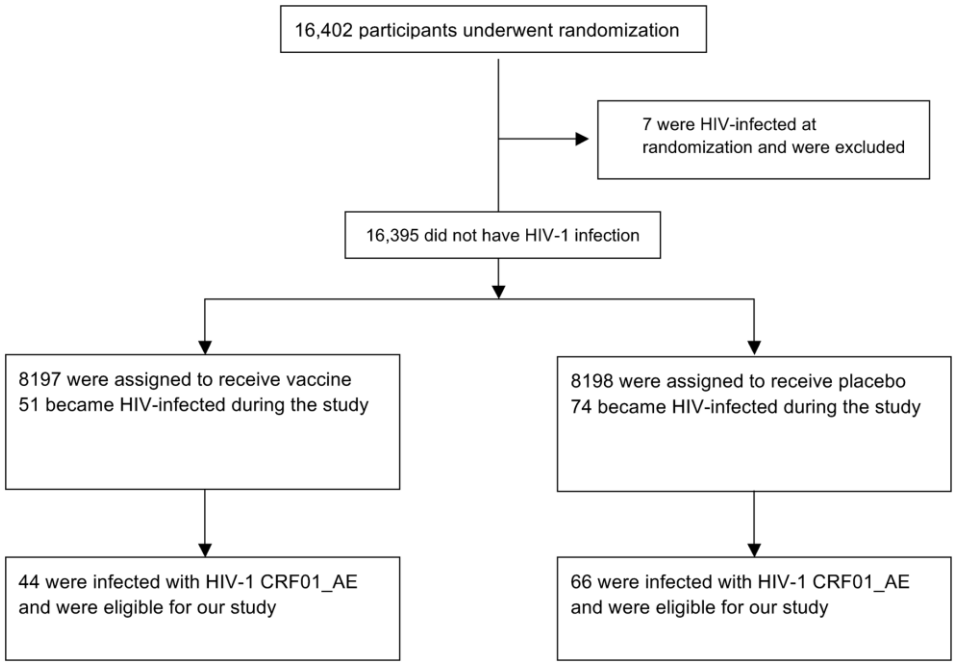



**Figure 1. Flowchart diagram of HIV-1 breakthrough infections in RV144.**
doi:10.1371/journal.pone.0111334.g001

## No vaccine/placebo distinction in Gag, Pro, Gp120 epitope comparisons against the vaccine inserts

Under the hypothesis that vaccine-induced immune responses could lead to escape mutations in sequences from vaccinees, we used epitopes predicted based on each subject's genotype to test whether epitope changes relative to the vaccine inserts differed between the vaccine and placebo groups. Because the vaccine was multivalent, we compared the epitopes predicted in RV144

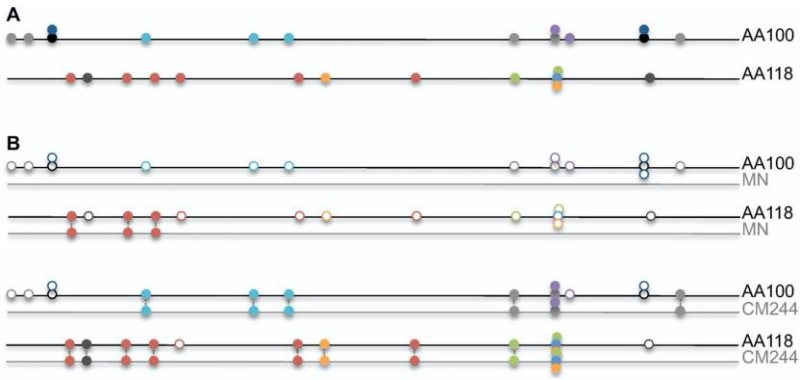



**Figure 2. Schematic representation of epitope predictions in RV144 HIV-1 breakthrough sequences, and their comparison to RV144 vaccine inserts.** A. Each line represents the Env-gp120 sequence from a subject and each circle a CD8 epitope prediction (different colors for different HLA alleles). The figure represents epitopes predicted based on each subject's HLA class I genotype for two subjects who were infected with a nearly identical virus (AA100: HLA-A*02:03, HLA-A*24:10, HLA-B*18:01, HLA-B*18:02, HLA-C*07:04; AA118: HLA-A*11:01, HLA-A*24:07, HLA-B*44:03, HLA-C*01:02, HLA-C*07:01). B. Epitope repertoires from a given subject are compared to the epitope predictions for the vaccine insert sequences (CM244 and MN) based on that subject's HLA class I genotype. Empty circles represent epitopes predicted in the sequence from a subject that could not be matched to a corresponding epitope prediction based on the vaccine insert sequence and the subject's HLA class I genotype. More subject-derived epitopes were matched against the vaccine insert CM244 than against MN; both subjects were infected by a CRF01-AE virus like CM244, while MN is a subtype B virus.
doi:10.1371/journal.pone.0111334.g002

**Table 1.** Number of CD8 epitopes predicted for each subject depending on his HLA type.

| | Gag | | Pro | | Env | |
|---|---|---|---|---|---|---|
| | **Placebo** | **Vaccine** | **Placebo** | **Vaccine** | **Placebo** | **Vaccine** |
| Nr of subjects | 66 | 44 | 65 | 43 | 66 | 44 |
| Median | 51.5 | 45 | 6 | 5 | 66 | 64 |
| Mean | 52.71 | 46.11 | 6.39 | 6.21 | 69.68 | 64.75 |
| p-value | 0.019 | | 0.464 | | 0.238 | |
| | **RT-IN** | | **Nef** | | | |
| | **Placebo** | **Vaccine** | **Placebo** | **Vaccine** | | |
| Nr of subjects | 66 | 44 | 66 | 44 | | |
| Median | 117.5 | 116 | 26.5 | 25 | | |
| Mean | 116.30 | 113.80 | 26.42 | 25.30 | | |
| p-value | 0.707 | | 0.408 | | | |

Predictions are given for proteins that corresponded to the vaccine inserts (Gag, Pro, Env), and for proteins that were not part of the vaccine (RT-IN, and Nef).
doi:10.1371/journal.pone.0111334.t001

participants' sequences to the subtype B and CRF01_AE reference strains included in the vaccine (CRF01_AE: CM244 and 92TH023; subtype B: LAI and MN). We tested both differences in binding affinities (by comparing predicted IC50 for the epitopes in the vaccinee's sequence and in the vaccine insert) and in evolutionary distances (i.e., protein distances between RV144 participants' epitopes and vaccine-derived epitopes were calculated using an HIV-specific substitution model). We looked for evidence of a difference between vaccine and placebo group by analyzing predictions in proteins that were part of the vaccine (Gag, Pro, Gp120) and used as control the epitope predictions derived from RT-IN and Nef (to test the hypothesis that there would be no difference between the vaccine and placebo group outside of the vaccine insert).

We first looked at predicted CD8 and CD4 epitopes defined as both weak and strong binders. We found no difference between the vaccine and placebo group either when we focused on binding affinities or evolutionary distances for the different proteins and references considered (data not shown). Second, we focused on the subset of epitopes identified as strong binders (i.e., with a predicted IC50≤50 nM). Again, there was no concordant evidence of a difference between the vaccine and placebo group, although, for CD8 epitope predictions, there was a trend suggesting a difference between the vaccine and placebo groups in Pro ($p = 0.080$) and gp120 ($p = 0.065$) when binding affinities of predicted epitopes were considered (Table 3). In addition, when the analysis was limited to V2, some results were significant ($0.022 \leq$ p-value$\leq 0.231$; see below). Comparisons of epitope predictions in RT-IN and Nef verified that there was also no distinction between the vaccine and placebo group for proteins not included in the vaccine (p-values≥0.16 for both CD4 and CD8 predictions).

## Env-V2 and Env-V3-specific comparisons of epitope predictions

Given the results of the RV144 correlates of risk study [4], a V2-specifc analysis was performed. While analyses of CD8 epitope predictions that included both strong and weak binders showed that only the comparison of V2-binding affinity measures against CM244 was significant ($p = 0.017$), the results were more consistent when we focused on the subset of epitopes identified as strong binders (i.e., with a predicted IC50≤50 nM) (Table 3). Data were not available against the subtype B vaccine boost (MN) because subject-derived epitopes were too divergent to be matched to the vaccine strain. For predicted strong CD8 binders, there was a significant difference between the vaccine and placebo group in the V2 region of Env when binding affinities were considered. Comparisons were significant against the CRF01_AE strains (CM244 $p = 0.022$; 92TH023 $p = 0.047$). When evolutionary distances were considered for CD8 epitopes, there was only a trend suggesting a difference between the vaccine and placebo group against the CRF01_AE strains (CM244 $p = 0.058$; 92TH023 $p = 0.231$).

When CD4 epitope predictions were considered, there were weak trends suggesting a greater number of strong binders in the

**Table 2.** Subject-specific epitopes matched to vaccine-derived epitopes.

| Matched epitopes | Gag | Pol | Env | Nef |
|---|---|---|---|---|
| CD8 | 4,629/5,509 | 12,198/13,368 | 5,545/15,520 | 2,389/2,860 |
| CD8 (%) | 84% | 91% | 36% | 84% |
| CD4 | 11,177/13,419 | 25,273/29,921 | 13,109/36,738 | 3,006/4,119 |
| CD4 (%) | 83% | 84% | 36% | 73% |

Epitopes were considered matched when the subject- and vaccine-derived peptides had at least 67% AA identity. Number and percentages of matched epitopes are given for all of the 110 subjects in the cohort; there was no difference between the vaccine and placebo groups.
doi:10.1371/journal.pone.0111334.t002

**Table 3.** CD8 and CD4 epitopes predicted to be strong binders matched against different vaccine inserts.

| | | | | | | |
|---|---|---|---|---|---|---|
| **CD8 epitope predictions - Binding affinity** | | | | | | |
| **Gag** | **P CM244 (66)** | **V CM244 (44)** | **P LAI (66)** | **V LAI (44)** | | |
| Mean | 0.996 | 0.965 | 0.979 | 0.979 | | |
| p-value | 0.266 | | 0.813 | | | |
| **Pro** | **P CM244 (31)** | **V CM244 (20)** | **P LAI (31)** | **V LAI (20)** | | |
| Mean | 0.24 | 0.338 | 0.426 | 0.354 | | |
| p-value | 0.08 | | 0.333 | | | |
| **gp120** | **P CM244 (66)** | **V CM244 (44)** | **P MN (66)** | **V MN (44)** | **P 92TH (66)** | **V 92TH (44)** |
| Mean | 0.95 | 0.997 | 0.623 | 0.676 | 0.826 | 0.864 |
| p-value | 0.065 | | 0.468 | | 0.448 | |
| **V2** | **P CM244 (58)** | **V CM244 (35)** | **P MN (0)** | **V MN (0)** | **P 92TH (58)** | **V 92TH (35)** |
| Mean | 0.957 | 1.123 | n.a. | n.a. | 0.017 | 0.114 |
| p-value | 0.022 | | n.a. | | 0.047 | |
| **V3** | **P CM244 (13)** | **V CM244 (9)** | **P MN (0)** | **V MN (0)** | **P 92TH (0)** | **V 92TH (0)** |
| Mean | 0.903 | 0.548 | n.a. | n.a. | n.a. | n.a. |
| p-value | 0.108 | | n.a. | | n.a. | |
| **RT-In** | **P CM244 (66)** | **V CM244 (44)** | **P LAI (66)** | **V LAI (44)** | | |
| Mean | 1 | 1 | 0.439 | 0.641 | | |
| p-value | n.a. | | 0.378 | | | |
| **Nef** | **P CM244 (66)** | **V CM244 (44)** | **P LAI (66)** | **V LAI (44)** | | |
| Mean | 0.98 | 0.9 | 0.669 | 0.635 | | |
| p-value | 0.162 | | 0.958 | | | |
| **CD8 epitope predictions - Evolutionary distance** | | | | | | |
| **Gag** | **P CM244 (66)** | **V CM244 (44)** | **P LAI (66)** | **V LAI (44)** | | |
| Mean | 0.033 | 0.051 | 0.142 | 0.139 | | |
| p-value | 0.351 | | 0.998 | | | |
| **Pro** | **P CM244 (31)** | **V CM244 (20)** | **P LAI (31)** | **V LAI (20)** | | |
| Mean | 0.124 | 0.14 | 0.152 | 0.113 | | |
| p-value | 0.851 | | 0.361 | | | |
| **gp120** | **P CM244 (66)** | **V CM244 (44)** | **P MN (66)** | **V MN (44)** | **P 92TH (66)** | **V 92TH (44)** |
| Mean | 0.056 | 0.064 | 0.005 | 0.005 | 0.014 | 0.016 |
| p-value | 0.403 | | 0.356 | | 0.859 | |
| **V2** | **P CM244 (58)** | **V CM244 (35)** | **P MN (0)** | **V MN (0)** | **P 92TH (58)** | **V 92TH (35)** |
| Mean | 0.095 | 0.101 | n.a. | n.a. | 0.049 | 0.025 |
| p-value | 0.058 | | n.a. | | 0.231 | |
| **V3** | **P CM244 (13)** | **V CM244 (9)** | **P MN (0)** | **V MN (0)** | **P 92TH (0)** | **V 92TH (0)** |
| Mean | 0.21 | 0.148 | n.a. | n.a. | n.a. | n.a. |
| p-value | 0.227 | | n.a. | | n.a. | |
| **RT-In** | **P CM244 (66)** | **V CM244 (44)** | **P LAI (66)** | **V LAI (44)** | | |
| Mean | 0.002 | 0.002 | 0.038 | 0.03 | | |
| p-value | 0.792 | | 0.616 | | | |
| **Nef** | **P CM244 (64)** | **V CM244 (44)** | **P LAI (64)** | **V LAI (44)** | | |
| Mean | 0.081 | 0.087 | 0.281 | 0.247 | | |
| p-value | 0.803 | | 0.458 | | | |
| **CD4 epitope predictions - Binding affinity** | | | | | | |
| **Gag** | **P CM244 (62)** | **V CM244 (43)** | **P LAI (62)** | **V LAI (43)** | | |
| Mean | 0.991 | 0.985 | 1.002 | 0.948 | | |
| p-value | 0.688 | | 0.296 | | | |

**Table 3.** Cont.

| **CD8 epitope predictions - Binding affinity** | | | | | | |
|---|---|---|---|---|---|---|
| **Pro** | **P CM244 (40)** | **V CM244 (22)** | **P LAI (40)** | **V LAI (22)** | | |
| Mean | 0.816 | 0.804 | 0.897 | 0.971 | | |
| p-value | 0.794 | | 0.662 | | | |
| **gp120** | **P CM244 (59)** | **V CM244 (43)** | **P MN (59)** | **V MN (43)** | **P 92TH (59)** | **V 92TH (43)** |
| Mean | 0.86 | 0.847 | 0.548 | 0.374 | 0.528 | 0.4 |
| p-value | 0.989 | | 0.127 | | 0.336 | |
| **V2** | **P CM244 (45)** | **V CM244 (22)** | **P MN (45)** | **V MN (22)** | **P 92TH (45)** | **V 92TH (22)** |
| Mean | 1.001 | 1.14 | n.a. | n.a. | 0 | 0.045 |
| p-value | 0.177 | | n.a. | | 0.162 | |
| **V3** | **P CM244 (15)** | **V CM244 (4)** | **P MN (15)** | **V MN (4)** | **P 92TH (15)** | **V 92TH (4)** |
| Mean | 0.314 | 0.259 | n.a. | n.a. | n.a. | n.a. |
| p-value | 0.881 | | n.a. | | n.a. | |
| **RT-In** | **P CM244 (62)** | **V CM244 (43)** | **P LAI (62)** | **V LAI (43)** | | |
| Mean | 0.999 | 1 | 0.689 | 0.6 | | |
| p-value | 0.416 | | 0.403 | | | |
| **Nef** | **P CM244 (54)** | **V CM244 (35)** | **P LAI (54)** | **V LAI (35)** | | |
| Mean | 1.001 | 1.044 | 0.28 | 0.201 | | |
| p-value | 0.202 | | 0.186 | | | |
| **CD4 epitope predictions - Evolutionary distance** | | | | | | |
| **Gag** | **P CM244 (61)** | **V CM244 (43)** | **P LAI (61)** | **V LAI (43)** | | |
| Mean | 0.021 | 0.025 | 0.112 | 0.089 | | |
| p-value | 0.806 | | 0.04 | | | |
| **Pro** | **P CM244 (40)** | **V CM244 (22)** | **P LAI (40)** | **V LAI (22)** | | |
| Mean | 0.111 | 0.076 | 0.121 | 0.095 | | |
| p-value | 0.299 | | 0.732 | | | |
| **gp120** | **P CM244 (59)** | **V CM244 (43)** | **P MN (59)** | **V MN (43)** | **P 92TH (59)** | **V 92TH (43)** |
| Mean | 0.097 | 0.083 | 0.025 | 0.015 | 0.046 | 0.035 |
| p-value | 0.108 | | 0.12 | | 0.195 | |
| **V2** | **P CM244 (45)** | **V CM244 (22)** | **P MN (0)** | **V MN (0)** | **P 92TH (45)** | **V 92TH (22)** |
| Mean | 0.168 | 0.089 | n.a. | n.a. | 0.041 | 0.047 |
| p-value | 0.01 | | n.a. | | 0.602 | |
| **V3** | **P CM244 (15)** | **V CM244 (4)** | **P MN (0)** | **V MN (0)** | **P 92TH (15)** | **V 92TH (4)** |
| Mean | 0.235 | 0.2 | n.a. | n.a. | 0.043 | 0.176 |
| p-value | 0.96 | | n.a. | | 0.272 | |
| **RT-In** | **P CM244 (62)** | **V CM244 (43)** | **P LAI (62)** | **V LAI (43)** | | |
| Mean | 0.006 | 0.002 | 0.035 | 0.035 | | |
| p-value | 0.332 | | 0.65 | | | |
| **Nef** | **P CM244 (54)** | **V CM244 (35)** | **P LAI (54)** | **V LAI (35)** | | |
| Mean | 0.107 | 0.105 | 0.347 | 0.405 | | |
| p-value | 0.956 | | 0.296 | | | |

Epitopes were predicted in all HIV-1 proteome sequences derived from RV144 breakthrough infections. The epitopes were matched against epitopes derived from the RV144 vaccine inserts of subtype B (MN, LAI) or CRF01_AE (CM244, 92TH023); two epitope characteristics were used to compare epitopes from the breakthrough to the vaccine: the predicted binding affinity for each epitope and the protein distance between the epitope sequences. One summary measure was computed for each protein and each subject, and comparisons were done between the vaccine (V) and placebo (P) groups (the number of vaccine and placebo recipients included in each group is in parenthesis) with Mann-Whitney tests for proteins corresponding to those included in the RV144 vaccine insert (Gag, Pro, gp120 (including V2)) and those not part of the RV144 vaccine (RT-IN, Nef).

doi:10.1371/journal.pone.0111334.t003

V2 of vaccinees compared to the placebo group (CM244 $p = 0.177$; 92TH023 $p = 0.162$), and significantly greater evolutionary distances in placebo recipients against CM244 ($p = 0.010$), but not against 92TH023 ($p = 0.602$).

Given that high responses to V3 have been associated with a lower risk of HIV-1 infection among vaccinees who had low gp120-specific plasma IgA: (OR = 0.49, p = 0.007), we also performed a V3-specific analysis for CD8 and CD4 epitopes. There was no significant distinction found between the vaccine and placebo groups using the tests described above ($p>0.108$), noting that there were no epitopes in V3 that had enough similarity with MN for the corresponding epitopes to be matched (i.e., less than 2/3 of the residues of a peptide matched).

Parallels between V2 and V3 were notable: i) both V2 and V3 were a hotspot of RV144-induced antibody responses [13]; ii) V2- and V3-specific binding antibodies were identified as correlates of risk of infection in RV144; and iii) they harbored signature sites of RV144 vaccination. Statistical analyses that interrogated individual sites in Env from RV144 participants identified two sites in V2 (sites 169 and 181, [2]) and one in V3 (site 317). Thus, we looked specifically at predicted epitopes in V2 and V3 that spanned the signature sites. Figure 3 shows the overlap between CD8 and CD4 predicted epitopes in V2 and V3: we note that because multiple alleles can restrict the same peptide the number of epitope predictions starting at a specific location can surpass the number of subjects in the cohort. In addition, there are epitopes spanning the signature sites in both V2 and V3; however, no distinction was detected between the vaccine and placebo group (Table S1 in File S2).

### No allele-specific effect in vaccine/placebo epitope comparisons against the vaccine inserts

By analyzing Env sequence data as a function of each subject's genotype on a site-by-site basis with a phylogenetically-corrected method, five DRB1 and one class I HLA alleles (DRB1*03, DRB1*04, DRB1*07, DRB1*11, DRB1*15, A*68) were associated with specific AA polymorphisms in Env. However, the limited number of subjects with a given allele (ranging from four with DRB1*11 to 33 with DRB1*15) did not allow us to perform meaningful vaccine/placebo comparisons restricted to carriers of these specific alleles.

Next, when $\log_{10}$ viral loads (measured at the time of sampling for viral genome sequencing, corresponding to early infection as the last negative visit happened six months prior diagnosis) were analyzed as a function of Env sequence together with each subject's genotype, HLA alleles HLA-A*11 and HLA-B*46 were associated with higher viral loads, although the effect sizes were small (HLA-A*11: weight = 0.23, bootstrap support = 0.87; HLA-B*46: weight = 0.12, bootstrap support = 0.70). We therefore examined epitope metrics for carriers of HLA-A*11 (n = 59), HLA-B*46 (n = 40), as well as for carriers of HLA-A*02 as this is another frequent allele in the RV144 cohort (n = 53). We repeated the analysis described above by taking into account separately the predicted epitopes restricted by HLA-A*02, HLA-A*11, and HLA-B*46; including strong and weak binders (focusing only on strong binders reduced the subset of predicted epitopes to a number too small for adequate comparisons).

For these allele-specific comparisons, there was no difference between the vaccine and placebo groups whether evolutionary distances or binding affinity were considered – the smallest p-values were p = 0.206 for evolutionary distances (gp120 vs CM244 for HLA-A*02), and p = 0.123 for binding affinity measures (V2 vs CM244 for HLA-B*46) (Table S2 in file S2). There was a p-value = 0.045 for Nef epitope predictions restricted by HLA-A*11; this result should probably not be viewed as a significant vaccine/placebo distinction given the high number of tests performed, that Nef was not part of the vaccine, and that it is a comparison against the cross-subtype reference (subtype B) while all subjects analyzed here were infected with CRF01_AE viruses.

### No relationship between epitopic distances and the duration of HIV-1 infection

We looked at the relationship between the evolutionary distance (defined above) and the mean diversity in each subject, the latter of which can be used as a measure of the age of the infection. The hypothesis is that if the epitope distances track intra-host diversity it could be interpreted as a sign of intra-host evolution, i.e., a post-infection effect. We found no relationship between the epitopic evolutionary distance and the mean diversity in each subject: For V2 CD8 predictions: Spearman correlation coefficient $\rho = -0.113$ ($p = 0.283$); for V2 CD4 predictions: Spearman correlation coefficient $\rho = 0.011$ ($p = 0.929$). There was also no relationship between the epitopic evolutionary distance and the number of days since the last negative visit in each subject (V2 CD8 predictions: Spearman correlation coefficient $\rho = -0.188$ ($p = 0.071$); V2 CD4 predictions: Spearman correlation coefficient $\rho = -0.075$ ($p = 0.544$)).

## Discussion

Here we analyzed epitope predictions derived from HIV-1 genome sequences corresponding to 110 CRF01_AE breakthrough infections in the RV144 trial, including 44 vaccine and 66 placebo recipients. We tested for evidence of a distinction between the vaccine and placebo groups and found evidence potentially suggestive of a weak T cell driven sieve effect among breakthrough viruses as vaccine/placebo comparisons showed a) some evidence of a signal converging on the Env-V2 segment, and b) smaller Gag epitope repertoires in vaccine recipients compared to placebo recipients.

One indication that the V2 signal, although weak, may be genuine is the fact that differences between the vaccine and placebo group were only seen when comparisons were made against CRF01_AE strains (CM244, 92TH023). No difference was seen when subject-derived sequences were compared to epitopes derived from the subtype B MN strain; this appears logical as the MN boost protein would seem unlikely to have elicited a substantial number of cross-reactive T cell responses toward infecting CRF01_AE viruses, which at the epitope level often differed from MN by 4–5 residues out of nine in a CTL epitope. The lack of V2 signal against MN as a reference could be expected as cross-reactivity decreases drastically with more than 2 mutations out of 9 residues in a CTL epitope [10] - a typical instance when epitopes are compared between subtype B and CRF01_AE. An additional factor that may explain the identification of a vaccine/placebo distinction in V2 is the fact that 60% of the CD4 responses detected in a subset of RV144 vaccine recipients were directed against V2 [14], although, paradoxically, these responses were not identified post-infection, suggesting that antigen-specific T cells could possibly have been preferentially infected and deleted [15]. Importantly, it is possible that the signal detected in V2 was the consequence of an antibody-mediated effect, as binding antibodies targeting V2 were associated with a decreased risk of HIV-1 infection in the RV144 trial. As such, two vaccine-associated signatures in V2 [2] that were linked to vaccine-derived binding antibodies [3] were located within predicted CD8/4 epitopes. Overall, it is difficult to hypothesize that T cell driven immune responses played an important role in the protection associated with the RV144 vaccine because few RV144 subjects mounted CD8/CD4 responses following RV144 vaccination [1]. However, we cannot discount that T cell responses may have played a role in

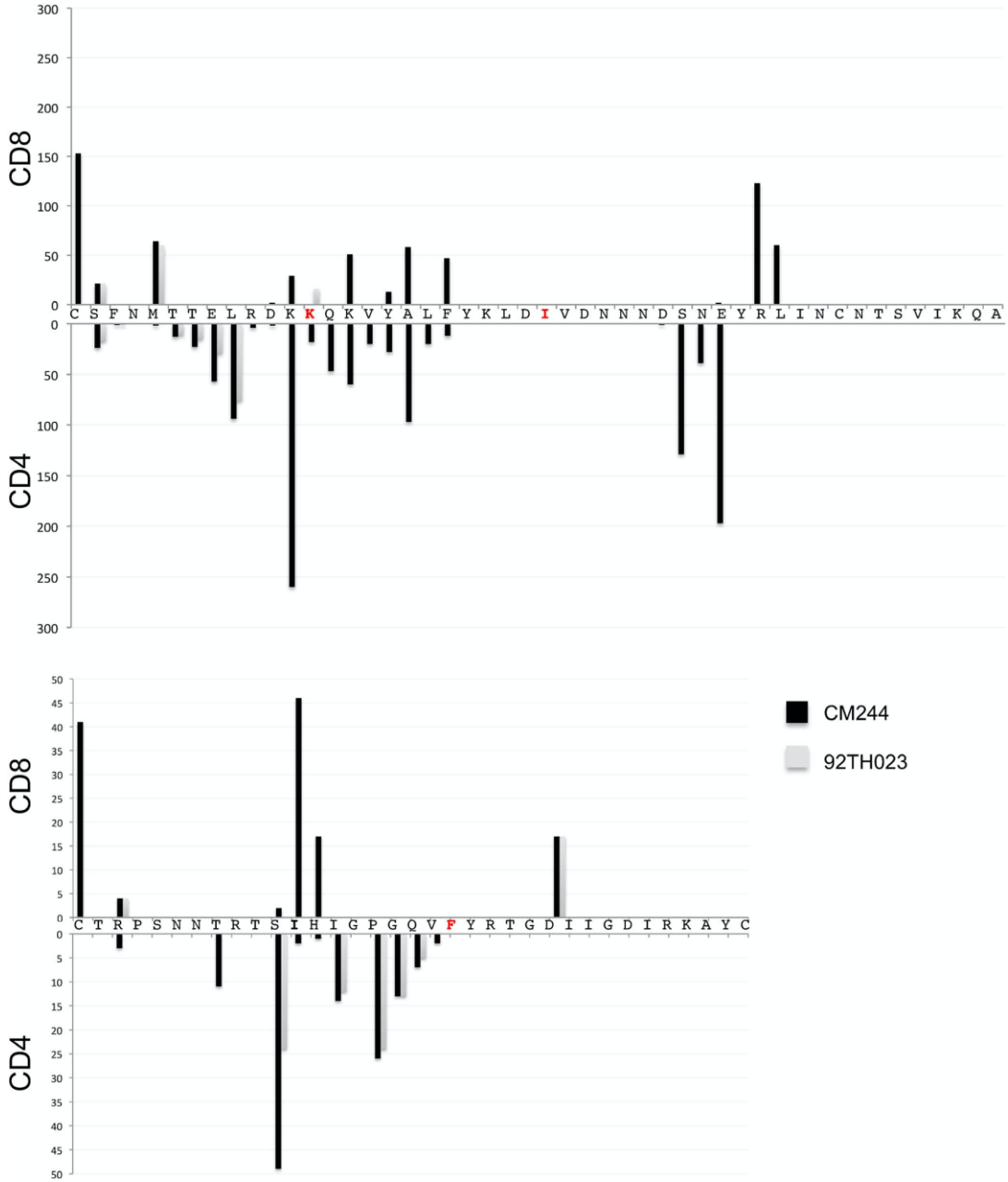



**Figure 3. Overlap between predicted CD8 and CD4 epitopes in Env-V2 and -V3 in sequences from HIV-1-infected RV144 participants.** The x-axis corresponds to the V2 and V3 sequence, and the y-axis corresponds to the number of predicted epitopes starting at each position. The epitope predictions correspond to all unique HLA/peptide combinations, hence the number of epitope predictions starting at a specific location can surpass the number of subjects in the cohort because a given peptide can be predicted as an epitope for multiple HLA alleles. The amino acids in red correspond to sites that were identified as genetic signatures that distinguished breakthrough sequences from vaccine and placebo recipients.
doi:10.1371/journal.pone.0111334.g003

a more restricted way as the limited sensitivity of the T cell assays employed in the RV144 trial means that some responses may not have been detected.

Our results were generally, but not necessarily, in agreement between the CD8 and CD4 epitope predictions. It has to be noted that CD4 epitopes are less well defined than CD8 epitopes in the context of HIV-1 natural infection, hence *in silico* predictions of CD4 epitopes are less accurate than those for CD8 epitopes. It is possible that the absence of information and algorithms for the prediction of epitopes in HLA class II alleles common in Thailand may underestimate epitope enumeration and bias this analysis. To overcome the uncertainty in epitope predictions, additional analyses were focused on a few V2 epitopes that had been precisely characterized in some RV144 participants ([16]; and personal communication from Mark deSouza and Silvia Ratto-Kim). For example, we specifically looked at predicted epitopes corresponding to the peptide that showed the highest recognition in a study of responses from 25 RV144 vaccine recipients following vaccination: peptide VHALFYKLDIVPIED (AA 172–186) [16]. However, these epitope comparisons focused on the experimental data showed no distinction between vaccine and placebo groups, while this may be due in part to the small number of subjects, it also illustrates the limited congruence between *in silico* predictions and mapped CTL responses, as previously noted [6].

The second signal we identified was the significantly ($p = 0.019$) smaller size of Gag epitope repertoires among vaccine recipients compared to placebo recipients. One interpretation is that there could be more escape mutations in Gag sequences from vaccinees (due to vaccine-induced responses), and that these mutations prevented certain sequence motifs from being recognized as epitopes. Another interpretation is that vaccine efficacy might depend on the number of Gag epitopes in the circulating viruses, implying that viruses with canonical epitopes were more likely to be blocked from establishing infection. Some evidence of a vaccine/placebo distinction in Gag is interesting as Gag is typically a preferential target of T cell responses compared to the V2 region of Env. Indeed, a T cell driven sieve effect may have been expected to be uncovered in an immunodominant region of the proteome, as was shown in a study of the Step trial [6]. Analyses of epitope repertoires from breakthrough infections in the Step trial showed a distinction between the vaccine and placebo group that was largely driven by an effect in Gag. Several factors allowed for the detection of a T cell based signal in Gag in the Step trial: i) most subjects mounted T cell responses following vaccination; ii) there are immunodominant responses in Gag; iii) a Gag immunodominant response (SL9) was restricted by a common allele in the Step cohort (HLA-A*02); iv) Gag is a relatively conserved protein allowing for an easier identification of genetic signals than a more variable protein (such as Env); v) there was no env immunogen to potentially shift immunodominance patterns in vaccinees.

Lastly, our results differ from those of Gartland and colleagues [17], who reported evidence of greater predicted HLA binding escape for an HLA A*02 peptide in vaccine versus placebo recipients, and greater vaccine efficacy in A*02-positive participants than in A*02-negative participants (VE = 54% versus 3%, $P = 0.05$). We note that a comprehensive test of all HLA alleles in the RV144 cohort failed to show that A*02 modified vaccine efficacy. The differences between our studies could also be due to the fact that the MN-derived V2 peptide linked to the findings by Gartland and colleagues was not included as a predicted epitope in our analysis, because we only considered as a predicted epitope the peptides that had a predicted binding affinity <500 nm (different methods can be used to analyze sequences without conditioning on the presence of potential epitopes [18]). Besides, few MN-derived epitopes were included in our cross-proteome analysis since subtype B- (MN vaccine insert) and CRF01_AE- (breakthrough viruses) derived epitopes were poorly matched; the MN peptide highlighted in [17] showed three to six AA differences with corresponding breakthrough-derived peptides, suggesting a limited probability of cross-reactive CTL responses with the breakthrough peptides (knowing that the MN component of the RV144 vaccine was a protein boost, which would not typically be expected to elicit CTL responses).

In conclusion, while there is some evidence that T cell driven immune responses may have been associated with genetic changes in HIV-1 breakthrough viruses from vaccinees, the fact that most results are not strongly corroborated across our proteome-wide epitope comparisons is consistent with weak CD8-driven cytotoxic T cell responses in RV144. Hence, in the absence of evidence for strong CD8+ T cell epitopic signals across the RV144 vaccine inserts, the most conservative interpretation of our findings for V2 epitope predictions would suggest a relationship to V2-specific binding antibody responses previously identified as a correlate of risk in RV144 [4].

## Supporting Information

**File S1 File S1 corresponds to the RV144 protocol (v3_7, from July 27, 2008).**
(PDF)

**File S2 File S2 includes tables S1 and S2.** Table S1 corresponds to epitope predictions limited to the Env-V2/V3 region and Table S2 to predictions for specific HLA alleles.
(XLSX)

## Acknowledgments

Disclaimer: Data correspond to some breakthrough infections in the RV144 vaccine efficacy trial (ClinicalTrials.gov registration number NCT00223080); results of the RV144 trial were published by Rerks-Ngarm and colleagues [1].

We thank Drs. Peter B. Gilbert, Alexandra Schuetz, Rasmi Thomas for helpful discussions and comments on the manuscript.

## Author Contributions

Conceived and designed the experiments: MR. Performed the experiments: KD JMC. Analyzed the data: KD MR JMC. Contributed reagents/materials/analysis tools: GK JMC BBL ST ESB DEG WD BSM PTE SRK MdS SRN SN PP JK. Contributed to the writing of the manuscript: MR KD JMC BBL ROC MLR NLM JIM JHK.

## References

1. Rerks-Ngarm S, Pitisuttithum P, Nitayaphan S, Kaewkungwal J, Chiu J, et al. (2009) Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand. N Engl J Med 361: 2209–2220.
2. Rolland M, Edlefsen PT, Larsen BB, Tovanabutra S, Sanders-Buell E, et al. (2012) Increased HIV-1 vaccine efficacy against viruses with genetic signatures in Env V2. Nature.
3. Liao HX, Bonsignori M, Alam SM, McLellan JS, Tomaras GD, et al. (2013) Vaccine induction of antibodies against a structurally heterogeneous site of immune pressure within HIV-1 envelope protein variable regions 1 and 2. Immunity 38: 176–186.
4. Haynes BF, Gilbert PB, McElrath MJ, Zolla-Pazner S, Tomaras GD, et al. (2012) Immune-correlates analysis of an HIV-1 vaccine efficacy trial. N Engl J Med 366: 1275–1286.

5. Rolland M, Heckerman D, Deng W, Rousseau C, Coovadia H, et al. (2008) Broad and Gag-biased HIV-1 epitope repertoires are associated with lower viral loads. PLoS ONE 3: e1424.
6. Rolland M, Tovanabutra S, deCamp AC, Frahm N, Gilbert PB, et al. (2011) Genetic impact of vaccination on breakthrough HIV-1 sequences from the STEP trial. Nat Med 17: 366–371.
7. Hurley CK, Mack SJ, Mickelson E, Marsh S, Tilanus MGJ, et al. (2006) HLA typing and informatics. in immunobiology of the human MHC, in J A Hansen (ed), 13th International Histocompatibility Workshop protocols IHWG Press, Seattle, WA: 179–352.
8. Nielsen M, Lundegaard C, Blicher T, Lamberth K, Harndahl M, et al. (2007) NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. PLoS ONE 2: e796.
9. Nielsen M, Justesen S, Lund O, Lundegaard C, Buus S (2010) NetMHCIIpan-2.0 - Improved pan-specific HLA-DR predictions using a novel concurrent alignment and weight optimization training procedure. Immunome Res 6: 9.
10. Rolland M, Frahm N, Nickle DC, Jojic N, Deng W, et al. (2011) Increased breadth and depth of cytotoxic T lymphocytes responses against HIV-1-B Nef by inclusion of epitope variant sequences. PLoS ONE 6: e17969.
11. Nickle DC, Heath L, Jensen MA, Gilbert PB, Mullins JI, et al. (2007) HIV-specific probabilistic models of protein evolution. PLoS One 2: e503.
12. Carlson JM, Brumme ZL, Rousseau CM, Brumme CJ, Matthews P, et al. (2008) Phylogenetic dependency networks: inferring patterns of CTL escape and codon covariation in HIV-1 Gag. PLoS Comput Biol 4: e1000225.
13. Karasavvas N, Billings E, Rao M, Williams C, Zolla-Pazner S, et al. (2012) The Thai Phase III HIV Type 1 Vaccine trial (RV144) regimen induces antibodies that target conserved regions within the V2 loop of gp120. AIDS Res Hum Retroviruses 28: 1444–1457.
14. de Souza MS, Ratto-Kim S, Chuenarom W, Schuetz A, Chantakulkij S, et al. (2012) The Thai phase III trial (RV144) vaccine regimen induces T cell responses that preferentially target epitopes within the V2 region of HIV-1 envelope. J Immunol 188: 5166–5176.
15. Douek DC, Betts MR, Hill BJ, Little SJ, Lempicki R, et al. (2001) Evidence for increased T cell turnover and decreased thymic output in HIV infection. J Immunol 167: 6663–6668.
16. de Souza MS, Ratto-Kim S, Chuenarom W, Schuetz A, Chantakulkij S, et al. (2012) The Thai phase III trial (RV144) vaccine regimen induces T cell responses that preferentially target epitopes within the V2 region of HIV-1 envelope. Journal of immunology 188: 5166–5176.
17. Gartland AJ, Li S, McNevin J, Tomaras GD, Gottardo R, et al. (2014) Analysis of HLA A*02 Association with Vaccine Efficacy in the RV144 HIV-1 Vaccine Trial. Journal of virology 88: 8242–8255.
18. Edlefsen PT, Gilbert PB, Rolland M (2013) Sieve analysis in HIV-1 vaccine efficacy trials. Curr Opin HIV AIDS 8: 432–436.

## REFERENCES


Al-Tam, F., Brunel, A. S., Bouzinbi, N., Corne, P., Bañuls, A-L., & Shahbazkia, H. R. (2012). DNAGear – a free software for *spa* type identification in *Staphylococcus aureus*. *BMC Research Notes, 5*, 642. https://doi.org/10.1186/1756-0500-5-642

Bartels, M. D., Petersen, A., Worning, P., Nielsen, J. B., Larner-Svensson, H., Johansen, H. K., Andersen, L. P., Jarløv, J. O., Boye, K., Larsen, A. R., Westh, H. (2014). Comparing whole-genome sequencing with Sanger sequencing for *spa* typing of methicillin-resistant *Staphylococcus aureus*. *Journal of Clinical Microbiology, 52*(12), 4305–4308. https://doi.org/10.1128/JCM.01979-14

Biere, B., & Schweiger, B. (2010) Human adenoviruses in respiratory infections: Sequencing of the hexon hypervariable region reveals high sequence variability. *Journal of Clinical Virology, 14*(4), 366–71. https://doi,org/10.1016/j.jcv.2010.01.005

Center for Genetic Epidemiology, Lyngby, Denmark. https://cge.cbs.dtu.dk/services/spatyper/

CDC 1981. Kaposi's sarcoma and Pneumocystis pneumonia among homosexual men—New York City and California. *MMWR Morb Mortal Wkly Rep* 30: 305–308

Crawford-Miksza, L. K., & Schnurr, D. P. (1994). Quantitative colorimetric microneutralization assay for characterization of adenoviruses. *Journal of Clinical*

*Microbiology*, *32*(9), 2331–2334. https://doi.org/10.1128/JCM.32.9.2331-2334.1994

Crawford-Miksza, L. K., & Schnurr, D. P. (1996). Adenovirus serotype evolution is driven by illegitimate recombination in the hypervariable regions of the hexon protein. *Virology*, *224*(2), 357–367.

Davison, A. J. (2003). Genetic content and evolution of adenoviruses. *Journal of General Virology*, *84*(11), 2895–2908. https://doi.org/10.1099/vir.0.19497-0

Dehghan, S., Seto, J., Jones, M. S., Dyer, D. W., Chodosh, J., & Seto, D. (2013). Simian adenovirus type 35 has a recombinant genome comprising human and simian adenovirus sequences, which predicts its potential emergence as a human respiratory pathogen. *Virology*, *447*(1–2), 265–273. https://doi.org/10.1016/j.virol.2013.09.009

Ebner, K., Pinsker, W. & Lion, T. (2005a). Comparative sequence analysis of the hexon gene in the entire spectrum of human adenovirus serotypes: Phylogenetic, taxonomic, and clinical implications. *Journal of Virology*, *79*(20), 12635–42. https://doi.org/10.1128/JVI.79.20.12635-12642.2005

Ebner, K., Suda, M., Watzinger, F., & Lion, T. (2005b). Molecular detection and quantitative analysis of the entire spectrum of human adenoviruses by a two-reaction real-time PCR assay. *Journal of Clinical Virology*, *43*(7), 3049–53. https://doi.org/10.1128/JCM.43.7.3049-3053.2005

Engelmann, I., Madisch, I., Pommer, H., & Heim, A. (2006). An outbreak of epidemic keratoconjunctivitis caused by a new intermediate adenovirus 22/H8 identified by

molecular typing. *Clinical Infectious Diseases, 43*(7), e64–6. https://doi.org/10.1086/507533

Feil, E. J., Cooper, J. E., Grundmann, H., Robinson, D. A., Enright, M. C., Berendt, T., Peacock, S. J., Smith, J. M., Murphy, M., Spratt, B. G., Moore, C. E., & Day, N. P. (2003). How clonal is *Staphylococcus aureus*? *Journal of Bacteriology, 185*(11), 3307–3316. https://doi.org/10.1128/JB.185.11.3307-3316.2003

Fenner, F. J., Bachmann, P. A., Gibbs, E. P. J., Murphy, F. A., Studdert, M. J., & White, D. O. (1993). *Veterinary Virology* (2nd ed.). Academic Press, Inc.

Fitzgerald, J. R., Sturdevant, D. E., Mackie, S. M., Gill, S. R., & Musser, J. M. (2001). Evolutionary genomics of *Staphylococcus aureus*: Insights into the origin of methicillin-resistant strains and the toxic shock syndrome epidemic. *Proceedings of the National Academy of Sciences of the United States of America, 98*(15), 8821–8826. https://doi.org/10.1073/pnas.161098098

Gall, J. G., Crystal, R. G., & Falck-Pedersen, E. (1998). Construction and characterization of hexon-chimeric adenoviruses: Specification of adenovirus serotype. *Journal of Virology, 72*(12), 10260–64. https://doi.org/10.1128/JVI.72.12.10260-10264.1998

Gillespie, T. R., Nunn, C. L., & Leendertz, F. H. (2008). Integrative approaches to the study of primate infectious disease: Implications for biodiversity conservation and global health. *American Journal of Physical Anthropology, Suppl.,47*, 53–69. https://doi.org/10.1002/ajpa.20949

Gnaneshan, S., Ijaz, S., Moran, J., Ramsay, M., & Green, J. (2007). HepSEQ:

International public health repository for hepatitis B. *Nucleic Acids Research, 35*(Issue suppl_1), D367–70. https://doi.org/10.1093/nar/gkl874

Grodzicker, T., Anderson, C., Sharp, P. A., & Sambrook, J. (1974). Conditional lethal mutants of adenovirus 2-simian virus 40 hybrids. I. Host range mutants of Ad2+ND1. *Journal of Virology, 13*(6), 1237–44. https://doi.org/10.1128/JVI.13.6.1237-1244.1974

Gruss, A., Moretto, V., Ehrlich, S. D., Duwat, P., & Dabert, P. (1991). GC-rich DNA sequences block homologous recombination in vitro. *Journal of Biological Chemistry, 266*(11), 6667–9.

HAdV Working Group. (2019). *Home*. George Mason University. http://hadvwg.gmu.edu/

Henquell, C., Bœuf, B., Mirand, A., Bacher, C., Traore, O., Déchelotte, P., Labbé, A., Bailly, J-L., & Peigue-Lafeuille, H. (2009). Fatal adenovirus infection in a neonate and transmission to health-care workers. *Journal of Clinical Virology, 45*(4), 345–348. https://doi.org/10.1016/j.jcv.2009.04.019

Hierholzer, John C. (1992). Adenoviruses—A spectrum of human diseases. *Clinical Microbiology Newsletter, 14*(15), 113–117. https://doi.org/10.1016/0196-4399(92)90010-7

Hilleman, M. R., & Werner, J. H. (1954). Recovery of new agent from patients with acute respiratory illness. *Proceedings of the Society for Experimental Biology and Medicine, 85*(1), 183–8. https://doi.org/10.3181/00379727-85-20825

Hurley, C. K., Wagner, J. E., Setterholm, M. I., & Confer, D. L. (2006). Advances in

HLA: Practical implications for selecting adult donors and cord blood units. *Biology of Blood and Marrow Transplantation, 12*(1), 28–33. https://doi.org/10.1016/j.bbmt.2005.10.005

Ishiko, H., & Aoki, K. (2009). Spread of epidemic keratoconjunctivitis due to a novel serotype of human adenovirus in Japan. *Journal of Clinical Microbiology, 47*(8), 2678–2679. https://doi.org/10.1128/JCM.r00313-09

Jones, K. E., Patel, N., Levy, M., Storeygard, A., Balk, D., Gittleman, J. L., & Daszak, P. (2008). Global trends in emerging infectious diseases. *Nature,451,* 990–94. https://doi.org/10.1038/nature06536

Jones, M. S., 2nd, Harrach, B., Ganac, R. D., Gozum, M. M., Dela Cruz, W. P., Riedel, B., Pan, C., Delwart, E. L., & Schnurr D. P. (2007). New adenovirus species found in a patient presenting with gastroenteritis. *Journal of Virology, 81*(11), 5978–84. https://doi.org/10.1128/JVI.02650-06

Kajon A. E., Dickson, L. M., Metzgar, D., Houng, H-S., Lee, V., & Tan, B-H. (2010). Outbreak of febrile respiratory illness associated with adenovirus 11a infection in a Singapore military training camp. *Journal of Clinical Microbiology, 48*(4), 1438–1441. https://doi.org/10.1128/JCM.01928-09

Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution, 33*(7), 1870–74. https://doi.org/10.1093/molbev/msw054

Leendertz, F. H., Yumlu, S., Pauli, G., Boesch, C., Couacy-Hymann, E., Vigilant, L., Junglen, S., Schenk, S., & Ellerbrok, H. (2006). A new Bacillus anthracis found in

wild chimpanzees and a gorilla from West and Central Africa. *PLoS Pathogens, 2*(6), e64. https://doi.org/10.1371/journal.ppat.0020064

Lion, T., Baumgartinger, R., Watzinger, F., Matthes-Martin, S., Suda, M., Preuner, S., Futterknecht, B., Lawitschka, A., Peters, C., Potschger, U., & Gadner, Helmut. (2003). Molecular monitoring of adenovirus in peripheral blood after allogeneic bone marrow transplantation permits early diagnosis of disseminated disease. *Blood, 102*(3), 1114–1120. https://doi.org/10.1182/blood-2002-07-2152

Liu, E. B., Ferreyra, L., Fischer, S. L., Pavan, J. V., Nates, S. V., Hudson, N. R., Tirado, D., Dyer, D. W., Chodosh, J., Seto, D., & Jones, M. S. (2011) Genetic analysis of a novel human adenovirus with a serologically unique hexon and a recombinant fiber gene. *PLoS One, 6*(9), e24491. https://doi.org/10.1371/journal.pone.0024491

Liu, E. B., Wadford, D. A., Seto, J., Vu, M., Hudson, N. R., Thrasher, L., Torres, S., Dyer, D. W., Chodosh, J., Seto, D., & Jones, M. S. (2012) Computational and serologic analysis of novel and known viruses in species human adenovirus D in which serology and genomics do not correlate. *PLoS One, 7*(3), e33212. https://doi.org/10.1371/journal.pone.0033212

Liu, J., Nian, Q-G., Zhang, Y., Xu, L-J., Hu, Y., Li, J., Deng, Y-Q., Zhu, S-Y., Wu, X-Y., Qin, E-D., Jiang, T., & Qin, C-F. (2014). In vitro characterization of human adenovirus type 55 in comparison with its parental adenoviruses, types 11 and 14. *PLoS One, 9*(6), e100665. https://doi.org/10.1371/journal.pone.0100665

Madisch, I., Harste, G., Pommer, H., & Heim, A. (2005). Phylogenetic analysis of the main neutralization and hemagglutination determinants of all human adenovirus

prototypes as a basis for molecular classification and taxonomy. *Journal of Virology, 79*(24), 15265–15276. https://doi.org/10.1128/JVI.79.24.15265-15276.2005

Madisch, I., Hofmayer, S., Moritz, C., Grintzalis, A., Hainmueller, J., Pring-Akerblom, P., & Heim, A. (2007). Phylogenetic analysis and structural predictions of human adenovirus penton proteins as a basis for tissue-specific adenovirus vector design. *Journal of Virology,* 81(15), 8270–8281. https://doi.org/10.1128/JVI.00048-07

Madisch, I., Wölfel, R., Harste, G., Pommer, H., & Heim, A. (2006). Molecular identification of adenovirus sequences: A rapid scheme for early typing of human adenoviruses in diagnostic samples of immunocompetent and immunodeficient patients. *Journal of Medical Virology*, *78*(9), 1210–1217. https://doi.org/10.1002/jmv.20683

Maiden, M. C., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D. A., Feavers, I. M., Achtman, M., & Spratt, B. G. (1998). Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences of the United States of America, 95*(6), 3140–3145. https://doi.org/10.1073/pnas.95.6.3140

Marques-Bonet, T., & Eichler, E. E. (2009). The evolution of human segmental duplications and the core duplicon hypothesis. *Cold Spring Harbor Symposia on Quantitative Biology, 74*, 355–62. https://doi.org/10.1101/sqb.2009.74.011

Mautner, V., & Mackay, N. (1984). Recombination in adenovirus: Analysis of crossover

sites in intertypic overlap recombinants. *Virology, 139*(1), 43–52. https://doi.org/10.1016/0042-6822(84)90328-3

Mautner, V., Williams, J., Sambrook, J., Sharp, P. A., & Grodzicker, T. (1975). The location of the genes coding for hexon and fiber proteins in adenovirus DNA. *Cell, 5*(), 93–99. https://doi.org/10.1016/0092-8674(75)90097-5

Mutch, D., Underwood, J., Geysen, M., & Rodda, S. (1994). Comprehensive T-cell epitope mapping of HIV-1 env antigens reveals many areas recognized by HIV-1-seropositive and by low-risk HIV-1-seronegative individuals. *Journal of Acquired Immune Deficiency Syndromes, 7*(9), 879–890.

Myers, R. E., Gale, C. V., Harrison, A., Takeuchi, Y., & Kellam, P. (2005). A statistical model for HIV-1 sequence classification using the subtype analyser (STAR). *Bioinformatics, 21*(17), 3535–40. https://doi.org/10.1093/bioinformatics/bti569.

National Nosocomial Infections Surveillance System. (1999). *National Nosocomial Infections Surveillance (NNIS) System report, data summary from January 1990-May 1999, issued June 1999*. *American Journal of Infection Control, 27*(6), 520–532. https://doi.org/10.1016/s0196-6553(99)70031-3

Nemerow, G. R., Pache, L., Reddy, V., & Stewart, P. L. (2009). Insights into adenovirus hosT-cell interactions from structural studies. *Virology, 384*(2),380–388.

Norrby, E. (1969). The structural and functional diversity of adenovirus capsid components. *Journal of General Biology*, *5*(2), 221–236. https://doi.org/10.1099/0022-1317-5-2-221

Okuma, K., Iwakawa, K., Turnidge, J. D., Grubb, W. B., Bell, J. M., O'Brien, F. G.,

Coombs, G. W., Pearman, J. W., Tenover, F. C., Kapi, M., Tiensasitorn, C., Ito, T., & Hiramatsu K. (2002). Dissemination of new methicillin-resistant *Staphylococcus aureus* clones in the community. *Journal of Clinical Microbiology,*.40:4289-4294. https://doi.org/10.1128/JCM.40.11.4289-4294.2002

Palmarini, M. (2007). A veterinary twist on pathogen biology. *PLoS Pathogens, 3*(2), e12. https://doi.org/10.1371/journal.ppat.0030012

Robinson, C. M., Seto, D., Jones, M. S., Dyer, D. W., & Chodosh, J. (2011a). Molecular evolution of human species D adenoviruses. *Infection, Genetics and Evolution, 11*(6), 1208–1217. https://doi.org/10.1016/j.meegid.2011.04.031

Robinson, C. M., Singh, G., Henquell, C., Walsh, M. P., Peigue-Lafeuille, H., Seto, D., Jones, M. S., Dyer, D. W., & Chodosh, J. (2011b). Computational analysis and identification of an emergent human adenovirus pathogen implicated in a respiratory fatality. *Virology, 409*(2), 141–147. https://doi.org/10.1016/j.virol.2010.10.020

Robinson, C. M., Singh, G., Lee, J. Y., Dehghan, S., Rajaiya, J., Liu, E. B., Yousuf, M. A., Betensky, R. A., Jones, M. S., Dyer, D. W., Seto, D., & James Chodosh. (2013). Molecular evolution of human adenoviruses. *Science Reports 3*(1812). https://doi.org/10.1038/srep01812

Robinson, O. J., Krimpsky, M., & Grillon, C. (2013a). The impact of induced anxiety on response inhibition. Front. *Human Neuroscience, 7*, 69. https://doi.org/10.3389/fnhum.2013.00069

Robinson, O. J., Letkiewicz, A. M., Overstreet, C., Ernst, M., & Grillon, C. (2011). The

effect of induced anxiety on cognition: Threat of shock enhances aversive processing in healthy individuals. *Cognitive, Affective & Behavioral Neuroscience, 11*(2), 217–227. https://doi.org/10.3758/s13415-011-0030-5

Robinson, O. J., Overstreet, C., Charney, D. S., Vytal, K., & Grillon, C. (2013b). Stress increases aversive prediction-error signal in the ventral striatum. *Proceedings of the National Academy of Sciences of the United States of America, 110*(10), 4129–4133. https://doi.org/10.1073/pnas.1213923110

Robinson, O. J., Overstreet, C., Letkiewicz, A., & Grillon, C. (2012). Depressed mood enhances anxiety to unpredictable threat. *Psychological Medicine, 42*(7), 1397–1407. https://doi.org/10.1017/S0033291711002583

Robinson, O. J., and Sahakian, B. J. (2009). A double dissociation in the roles of serotonin and mood in healthy subjects. *Biological Psychiatry, 65*(1), 89–92. https://doi.org/10.1016/j.biopsych.2008.10.001

Rosen, L. (1960). A hemagglutination-inhibition technique for typing adenoviruses. *American Journal of Hygiene, 71*, 120–128. https://doi.org/10.1093/oxfordjournals.aje.a120085

Rowe, W. P., Huebner, R. J., Gilmore, L. K., Parrott, R. H., & Ward. T. G. (1953). Isolation of a cytopathogenic agent from human adenoids undergoing spontaneous degeneration in tissue culture. *Experimental Biology and Medicine, 84*(3), 570–573.

Roy, S., Gao, G., Clawson, D. S., Vandenberghe, L. H., Farina, S. F., & Wilson, J. M. (2004). Complete nucleotide sequences and genome organization of four

chimpanzee adenoviruses. *Virology*, *324*(2), 361–372. https://doi.org/10.1016/j.virol.2004.03.047

Russell, W. C., & Benkö, M. (1999). Adenoviruses (*Adenoviridae*): Animal viruses. In R. G. Webster and A. Granoff (eds.), *Encyclopedia of virology* (pp. 14–21). Academic Press.

Rux, J. J., & Burnett, R. M. (2000). Type-specific epitope locations revealed by X-ray crystallographic study of adenovirus type 5 hexon. *Molecular Therapy*, *1*(1), 18–30. https://doi.org/10.1006/mthe.1999.0001

Saitou, N., & Nei, M. (1987). The neighbor-joining method—a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, *4*(4), 406–425. https://doi.org/10.1093/oxfordjournals.molbev.a040454

Seto, D., Chodosh, J., Brister, J. R., & Jones, M. S. (2011). Using the whole-genome sequence to characterize and name human adenoviruses. *Journal of Virology*, *85*(11), 5701–5702. https://doi.org/10.1128/JVI.00354-11

Sharp, P. (1994). Split genes and RNA splicing. *Cell*, *77*(6), 805–815. https://doi.org/10.1016/0092-8674(94)90130-9

Shenk, T. E. (2001). Adenoviridae: The viruses and their replication. In D. M. Knipe and P. M. Howley (eds.), *Fields' Virology* (4th ed., vol. 2, pp. 2265–2300). Lippincott/Williams & Wilkins.

Shopsin, B., Gomez, M., Montgomery, S. O., Smith, D. H., Waddington, M., Dodge, D., E., Bost, D. A., Riehman, M., Naidich, S., & Kreiswirth B. N. (1999). Evaluation of protein A gene polymorphic region DNA sequencing for typing of

*Staphylococcus aureus* strains. *Journal of Clinical Microbiology, 37*(11), 3556–3563. https://doi.org/10.1128/JCM.37.11.3556-3563.1999

Shopsin, B., Gomez, M., Waddington, M., Riehman, M., & Kreiswirth, B. N. (2000). Use of coagulase gene (*coa*) repeat region nucleotide sequences for typing of methicillin-resistant *Staphylococcus aureus* strains. *Journal of Clinical Microbiology*, *38*(9), 3453–3456. https://doi.org/10.1128/JCM.38.9.3453-3456.2000

Stillman, Bruce W. (1983). The replication of adenovirus DNA with purified proteins. *Cell*, *35*(1), 7–9. https://doi.org/10.1016/0092-8674(83)90201-5

Svensson, C., & Akusjärvi, G. (1984). Adenvirus VA RNAI: A positive regulator of mRNA translation. *Molecular and Cellular Biology*, *4*(4), 736–42. https://doi.org/10.1128/mcb.4.4.736

Tamura, K., Nei, M., & Kumar, S. (2004). Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proceedings of the National Academy of Sciences of the United States of America, 101*(30), 11030–11035. https://doi.org/10.1073/pnas.0404206101

van Oostrum, J., & Burnett, R. M. (1985). Molecular composition of the adenovirus type 2 virion. *Journal of Virology, 56*(2), 439–448. https://doi.org/10.1128/JVI.56.2.439-448.1985

Walls, Tony, Shankar, A. G., & Shingadia, D. (2003). Adenovirus: An increasingly important pathogen in paediatric bone marrow transplant patients. *The Lancet Infectious Diseases*, *3*(2), 79–86. https://doi.org/10.1016/s1473-3099(03)00515-2

Walsh, M. P., Chintakuntlawar, A., Robinson, C. M., Madisch, I., Harrach, B., Hudson, N. R., Schnurr, D., Heim, A., Chodosh, J., Seto, D., & Jones, M. S. (2009). Evidence of molecular evolution driven by recombination events influencing tropism in a novel human adenovirus that causes epidemic keratoconjunctivitis. *PLoS One, 4*(6), e5635. https://doi.org/10.1371/journal.pone.0005635

Walsh, M. P., Seto, J., Jones, M. S., Chodosh, J., Xu, W., & Seto, D. (2010). Computational analysis identifies human adenovirus type 55 as a re-emergent acute respiratory disease pathogen. *Journal of Clinical Microbiology, 48*(3), 991–993. https://doi.org/10.1128/JCM.01694-09

Waye, Mary Miu Yee, & Sing, Chor Wing. (2010). Anti-viral drugs for human adenoviruses. *Pharmaceuticals, 3*(10), 3343–3354. https://doi.org/10.3390/ph3103343

Wevers, D., Metzger, S., Babweteera, F., Bieberbach, M., Boesch, C., Cameron, K., Couacy-Hymann, E., Cranfield, M., Gray, M., Harris, L. A., Head, J., Jeffery, K., Knauf, S., Lankester, F., Leendertz, S. A., Lonsdorf, E., Mugisha, L., Nitsche, A., Reed, P., Robbins, M., Travis, D. A., Zommers, Z., Leendertz, F. H., & Ehlers, B. (2011). Novel adenoviruses in wild primates: A high level of genetic diversity and evidence of zoonotic transmissions. *Journal of Virology, 85*(20), 10774–10784. https://doi.org/10.1128/JVI.00810-11

White, M. F. (2011). Homologous recombination in the archaea: The means justify the ends. *Biochemical Society Transactions 39*(1), 15–9. https://doi.org/10.1042/BST0390015

Wigand, R., Keller, D., & Werling, I. (1982). Immunological relationship among human adenoviruses of subgenus D. *Archives of Virology, 72*(3), 199–209. https://doi.org/10.1007/BF01348965

Williams, J., Grodzicker, T., Sharp, P., & Sambrook J. (1975). Adenovirus recombination: Physical mapping of crossover events. *Cell, 4*(2), 113–19. https://doi.org/10.1016/0092-8674(75)90117-8

Wolfe, N. D., Dunavan, C. P., & Diamond, J. (2007). Origins of major human infectious diseases. *Nature, 447*, 279–283. https://doi.org/10.1038/nature05775

Yang, Z., Zhu, Z., Tang, L., Wang, L., Tan, X., Yu, P., Zhang, Y., Tian, X., Wang, J., Zhang, Y., Li, D., & Xu, W. (2009). Genomic analyses of recombinant adenovirus type 11a in China. *Journal of Clinical Microbiology*, *47*(10), 3082–3090. https://doi.org/10.1128/JCM.00282-09

Zhou, X., Robinson, C. M., Rajaiya, J., Dehghan, S., Seto, D., Jones, M. S., Dyer, D. W., & Chodosh, J. (2012). Analysis of human adenovirus type 19 associated with epidemic keratoconjunctivitis and its reclassification as adenovirus type 64. *Investigative Ophthalmology & Visual Science, 53*, 2804–2811. https://doi.org/10.1167/iovs.12-9656

Zhu, Z., Zhang, Y., Xu, S., Yu, P., Tian, X., Wang, L., Liu, Z., Tang, L., Mao, N., Ji, Y., Li, C., Yang, Z., Wang, S., Wang, J., Li, D., & Xu, W. (2008). Outbreak of acute respiratory disease in China caused by b2 species of adenovirus type 11. *Journal of Clinical Microbiology*, 47(3), 697–703. https://doi.org/10.1128/JCM.01769-08

Zubieta, C., Schoehn, G., Chroboczek, J., & Cusack, S. (2005). The structure of the

human adenovirus 2 penton. *Molecular Cell*, *17*(1), 121–135. https://doi.org/10.1016/j.molcel.2004.11.041

## BIOGRAPHY

Kalpana Dommaraju was born and brought up in Andhra Pradesh State in India. She received her Bachelor of Science in Horticulture in 2001. Ms. Dommaraju moved to Virginia and earned her Master of Science in Bioinformatics from George Mason University in 2006. She is now pursuing her Doctorate in Philosophy from George Mason University, in Bioinformatics and Computation Biology.