COMPUTATIONAL ANALYSIS OF HUMAN ADENOVIRUS EVOLUTION AND DEVELOPMENT OF BIOINFORMATICS TOOLS

by

Elizabeth Liu
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, Dissertation Director

_________________________________ Dr. Jason Kinser, Committee Member

_________________________________ Dr. Dmitri Klimov, Committee Member

_________________________________ Dr. Andrea Weeks, Committee Member

_________________________________ Dr. James D. Willett, Director, School of Systems Biology

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

_________________________________ Dr. Peggy Agouris, Dean, College of Science

Date: ___________________________ Fall Semester 2015
George Mason University
Fairfax, VA
Computational Analysis of Human Adenovirus Evolution and Development of Bioinformatics Tools

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

By

Elizabeth B. Liu
Master of Science
George Mason University, 2006
Bachelor of Science
Virginia Polytechnic Institute & State University, 2005

Director: Donald Seto, Professor
School of System Biology

Fall Semester 2015
George Mason University
Fairfax, VA
DEDICATION

This is dedicated to my loving parents Pingping Fan & Weiguo Ge and to my grandparents Qinghua Yang & Rensheng Fan, who have supported me through all the years of pursuing my degrees.
ACKNOWLEDGEMENTS

I would like to thank the many friends, relatives, and supporters who have made this happen. My loving parents and grandparents supported me throughout the years. Dr. Donald Seto, and the members of my committee Dr. Jason Kinser, Dr. Dmitri Klimov and Dr. Andrea Weeks for their invaluable help and input. I would also like to thank members of Dr. Seto’s lab (Shoaleh Dehghan, Jason Seto and Michael Walsh) for their valuable support and help.

For the HAdV16 analysis, I would like to thank Dr. Clark Tibbetts, James M. Clark and Dr. Anjan Purkayastha for initial discussions; and Drs. David Metzgar, Morris Jones and James Chodosh for continuing discussions. During the collection of this genome data set and preliminary analyses, DS (2002-2004), AP (2003-2005), JMC (2004) and CT (2001-2005) were affiliated with the HQ USAF Surgeon General Office, Directorate of Modernization (SGR) and the Epidemic Outbreak Surveillance (EOS) Program, 5201 Leesburg Pike, Suite 1401, Falls Church, VA 22041. Portions of this work were funded, during these time periods, by a grant from the U.S. Army Medical Research and Material Command (USAMRMC) (DAMD17-03-2-0089) and additional support was through the EOS Project, funded by HQ USAF Surgeon General Office, Directorate of Modernization (SGR) and the Defense Threat Reduction Agency (DTRA). Therefore, it is noted that the opinions and assertions contained herein are the private ones of the authors and are not to be construed as official or reflecting the views of the U.S. Department of Defense. I would also like to thank George Mason University’s Provost for their Ph.D. thesis completion grant.
TABLE OF CONTENTS

List of
Tables.................................................................................................................. vii
List of Figures.......................................................................................................... viii
List of Appendices................................................................................................... ix
Abstract.................................................................................................................. x
Chapter 1 - Introduction ....................................................................................... 1
Chapter 2 - Computational analysis of Human Adenovirus Type 16 ................. 6
Abstract.................................................................................................................. 6
Introduction............................................................................................................. 6
Material & methods ............................................................................................ 9
Results................................................................................................................... 9
  Genome sequence analysis.................................................................................. 9
  Comparative genome analysis.......................................................................... 11
  Hexon recombination....................................................................................... 14
  Phylogeny analysis......................................................................................... 17
Discussion............................................................................................................ 21
  Recombination, molecular evolution and new serotype.................................. 21
Conclusion.......................................................................................................... 24
Chapter 3 - Computational analysis of Human Adenovirus Type 21 ............... 26
Abstract............................................................................................................... 26
Introduction........................................................................................................ 26
Material & methods .......................................................................................... 27
Results............................................................................................................... 28
  Genome sequence analysis............................................................................ 28
  Comparative whole genome analysis............................................................ 28
  Hexon recombination.................................................................................... 31
  Phylogeny analysis....................................................................................... 35
Discussion........................................................................................................ 38
Conclusion......................................................................................................... 38
Chapter 4 - Human Adenovirus Type 58 .......................................................... 40
Introduction....................................................................................................... 40
Chapter 5 - Human Adenovirus Type 59 .......................................................... 52
Introduction....................................................................................................... 52
Chapter 6 - Bioinformatics Tools Development to Enhance Viral Genome Analysis ... 62
Introduction...................................................................................................... 62
Discussion........................................................................................................ 64
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1 – Taxonomy of Human Adenovirus</td>
<td>4</td>
</tr>
<tr>
<td>Table 2 – Percent Identity of HAdV-B16</td>
<td>13</td>
</tr>
<tr>
<td>Table 3 – Percent Identity of HAdV-B21</td>
<td>31</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1.</td>
<td>Structure of Human Adenovirus</td>
<td>3</td>
</tr>
<tr>
<td>Figure 2.</td>
<td>HAdV-B16 whole genome mapping</td>
<td>10</td>
</tr>
<tr>
<td>Figure 3.</td>
<td>zPicture analysis of HAdV-B16</td>
<td>12</td>
</tr>
<tr>
<td>Figure 4.</td>
<td>Whole genome recombination analysis HAdV-B16</td>
<td>15</td>
</tr>
<tr>
<td>Figure 5.</td>
<td>Hexon recombination analysis of HAdV-B16</td>
<td>16</td>
</tr>
<tr>
<td>Figure 6.</td>
<td>Whole genome phylogenetic analysis of HAdV-B16</td>
<td>18</td>
</tr>
<tr>
<td>Figure 7.</td>
<td>Hexon genome phylogenetic analysis of HAdV-B16</td>
<td>19</td>
</tr>
<tr>
<td>Figure 8.</td>
<td>Hexon halves phylogenetic analysis of HAdV-B16</td>
<td>20</td>
</tr>
<tr>
<td>Figure 9.</td>
<td>Fiber knobs phylogenetic analysis of HAdV-B16</td>
<td>21</td>
</tr>
<tr>
<td>Figure 10.</td>
<td>zPicture analysis of HAdV-B21</td>
<td>30</td>
</tr>
<tr>
<td>Figure 11.</td>
<td>Whole genome recombination analysis HAdV-B21</td>
<td>33</td>
</tr>
<tr>
<td>Figure 12.</td>
<td>Hexon recombination analysis of HAdV-B21</td>
<td>34</td>
</tr>
<tr>
<td>Figure 13.</td>
<td>Whole genome phylogenetic analysis of HAdV-B21</td>
<td>35</td>
</tr>
<tr>
<td>Figure 14.</td>
<td>Hexon genome phylogenetic analysis of HAdV-B21</td>
<td>36</td>
</tr>
<tr>
<td>Figure 15.</td>
<td>Hexon halves phylogenetic analysis of HAdV-B21</td>
<td>37</td>
</tr>
<tr>
<td>Figure 16.</td>
<td>Fiber knobs phylogenetic analysis of HAdV-B21</td>
<td>38</td>
</tr>
<tr>
<td>Figure 17.</td>
<td>DrawBar input screen</td>
<td>66</td>
</tr>
<tr>
<td>Figure 18.</td>
<td>DrawBar result screen</td>
<td>67</td>
</tr>
<tr>
<td>Figure 19.</td>
<td>GeneMap input screen</td>
<td>70</td>
</tr>
<tr>
<td>Figure 20.</td>
<td>GeneMap result screen 1</td>
<td>71</td>
</tr>
<tr>
<td>Figure 21.</td>
<td>GeneMap result screen 2</td>
<td>72</td>
</tr>
<tr>
<td>Figure 22.</td>
<td>HAdV working group website serotyping tool</td>
<td>74</td>
</tr>
</tbody>
</table>
## LIST OF APPENDICES

<table>
<thead>
<tr>
<th>Appendix</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appendix 1. DrawBar source code</td>
<td>77</td>
</tr>
<tr>
<td>Appendix 2. GeneMap source code</td>
<td>82</td>
</tr>
</tbody>
</table>
Computational analysis of human adenovirus evolution and development of bioinformatics tools

Elizabeth Liu, Ph.D.

George Mason University, 2015

Dissertation Director: Dr. Donald Seto

Human adenoviruses (HAdVs) may be highly contagious and may be human pathogens that can cause a wide range of illnesses including respiratory, gastrointestinal and ocular infections. Individuals with immune deficiency are especially prone to such infections, leading to fatalities. Even though adenoviruses continue to cause concerns due to notable mortality and morbidity in human populations, their existence may also provide a possible benefit for patients with a broader range of illnesses as well. In recent studies, for example, adenoviruses have been used in gene therapy and vaccine vector development. Recently, bioinformatics and genomics are both high-resolution approaches and resources available for studying the adenovirus for such purposes. One observation is that genome recombination is a driving force in the molecular evolution of human adenoviruses. This has implications for their use as vectors, particular across host species. Computational analysis of this event can provide a better understanding of the role recombination plays in the human adenovirus evolution and pathology, which may later
provide for a rational design of vaccines and for gene delivery vector development. Custom developed bioinformatics tools will also help to facilitate the process of data mining and analysis, and its presentation. In the course of this project, the genomes of three respiratory and gastrointestinal pathogens, HAdV-B16, B21, D58 and D59, were analyzed using bioinformatics tools to understand their origins and evolution as pathogens, in particular, changes in their genomes. To facilitate this, several tools were developed to assist this genome analysis. The analysis of both HAdV-B16 and HAdV-B21, archived 1950s prototypes, provided examples of “then novel and emergent” HAdVs arising as the result of genome recombination events with simian adenoviruses, across host species. Recently emergent HAdV-D58 and HAdV-D59 are novel pathogens that are characterized by genome sequencing and analysis. Their results have shown also that recombination plays an important role in their molecular evolution.
CHAPTER 1 - INTRODUCTION

Introduction

The first two human adenoviruses (HAdVs) were isolated around the same time in 1953: 1) from a child’s adenoid tissue as a non-specific human respiratory infectious agent and 2) from an U.S. Army recruit that presented with respiratory disease (Rowe et al., 1953; Hilleman et al., 1954). Since the first isolation of the virus, numerous members of the *Adenoviridae* family have been identified and characterized. It is likely all vertebrates are affected by adenoviruses (AdVs), including but not limited to fish, frogs, snakes, birds, canines, and primates, for example, chimpanzee and human. Human adenoviruses are grouped under the genus *Mastadenovirus* (mammalian) (Fenner et al., 1993). Within this, there are seven species (A-G) of HAdVs recognized, with more than 70 types that are identified based on immunochemistry, originally, and homologies of nucleic acid sequences and hexon and fiber protein sequences, as well as biological and genomic properties recently (Lion et al., 2014). The different types are associated with different tissue tropism, such as the upper and lower respiratory tracts, urinary and digestive tracts and eyes, allowing for characteristic diseases in these tissues and organs.

AdVs are non-enveloped icosahedral viruses comprising a nucleocapsid and a linear double strand DNA with a genome size average of 30 kb that encodes about 30
proteins (Rowe et al., 1953). The icosahedral capsid contains 12 vertices and seven
surface proteins (Figure 1) It has a unique spike-like fiber protein, associated with each
penton base of the capsid, which is the cell recognition domain, enabling the attachment
of the virus to the host cell. The major outer proteins of the capsid are 240 hexon (protein
I), 12 penton base at the vertices (protein II), and 12 protruding trimeric fibers (protein IV)
attached to each of the penton base. These outer capsid proteins define individual virus
types, and include the recognition site for host's immune system interaction. They are
also very useful for serotyping e.g., antibody assay, which was the main method for the
classification of AdVs. Recently, computational analyses of whole genome sequence and
the individual protein sequences have proved to be a faster, more complete and more
efficient process for typing AdVs (Jones et al., 2007).
Figure 1 - Structure of a human adenovirus.  
The hexon proteins provide structural stability while penton base and fiber proteins are responsible for host recognition and virus penetration. Adapted from Waye, M.M.Y.; Sing, C.W. Anti-Viral Drugs for Human Adenoviruses. Pharmaceuticals 2010, 3, 3343-3354.

Known diseases caused by HAdVs include but are not limited to, gastroenteritis, acute febrile pharyngitis, pharyngoconjunctival fever, acute respiratory disease, pneumonia, keratoconjunctivitis, pertussis like syndrome, acute hemorrhagic cystitis, meningoencephalitis and hepatitis (Jones et al., 2007). The HAdV-A species has also been shown to initiate sarcoma development in certain rodents (Ogawa K. 1989). Respiratory diseases are mainly due to species HAdV-B, E, and C. Ocular diseases are caused by HAdV-B, E, and D species. Gastroenteritis is due to the HAdV-F serotypes 40
and 41, although recent studies have shown HAdV-D serotype causes also gastroenteritis (Liu et al., 2011; Liu et al., 2012). HAdV-E only has one serotype, 4; it is primarily responsible for acute respiratory disease but may cause ocular disease at times. HAdV-4 is also one of the two HAdVs that has a vaccine developed against it (Jones et al., 2007), indicating its importance as a respiratory pathogen.

The classification of AdVs is complex. As mentioned earlier, HAdVs are classified under the genus Mastadenovirus. Currently there are more than 70 accepted HAdV types (unpublished observation) based on genomics (HAdV-1 to HAdV-70), which are categorized in seven species (HAdV-A to G) as shown in Table 1.

Table 1 - Taxonomy of Human Adenoviruses.
Using genomics and bioinformatics approaches, all of the recognized types of human adenoviruses are parsed into species. These conform to original observation based on serology, sequence comparisons and biological attributes.

<table>
<thead>
<tr>
<th>Species</th>
<th>Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>12, 18, 31, 61</td>
</tr>
<tr>
<td>B</td>
<td>3, 7, 11, 14, 16, 21, 34, 35, 50, 55, 66, 68</td>
</tr>
<tr>
<td>-B1</td>
<td>3, 7, 16, 21, 50, 66, 68</td>
</tr>
<tr>
<td>-B2</td>
<td>11, 14, 34, 35, 55</td>
</tr>
<tr>
<td>C</td>
<td>1, 2, 5, 6, 57</td>
</tr>
<tr>
<td>D</td>
<td>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</td>
</tr>
<tr>
<td>E</td>
<td>4</td>
</tr>
<tr>
<td>F</td>
<td>40, 41</td>
</tr>
<tr>
<td>G</td>
<td>52</td>
</tr>
</tbody>
</table>

In the past, HAdV serotyping and species classification were defined by reactivity of the outer coat proteins to discriminating antibodies. Additional biological properties were used as well, e.g., oncogenic potential and hemagglutination properties (Lion et al.,
However, these are lacking since they only probe a small view of the virus, for example the antibody epitope DNA sequence for the hexon protein is approximately 2.6% of the entire genome. Another drawback is that they are also expensive and time consuming to perform. Recent studies using DNA sequencing and bioinformatics methods, including phylogenetic analysis and amino acid sequence analysis, have provided a suitable alternative that is cost-effective, quantitative and much less time-consuming. Thus, bioinformatics has become a preferred and reliable method for demonstrating how those viruses are related through molecular evolution by using primary sequence data (Seto et al., 2009).

Currently, bioinformatics tools are lacking for mining viral genome data. Several computational tools were developed to meet this need and were applied to the study of four human adenovirus pathogens. The main objective of this thesis is to examine the genomes and determine the mechanisms of the molecular evolution of three respiratory and one gastrointestinal human pathogens, HAdV-16, 21, 58 and 59. This is accomplished by studying the changes in their genomes in order to understand the genesis of emergent and novel human viral pathogens.
CHAPTER 2 – COMPUTATIONAL ANALYSIS OF HUMAN ADENOVIRUS
Type 16

Abstract

Molecular evolution of human adenoviruses (HAdVs) is driven by recombination. The computational analysis of HAdV-B16, a subspecies B1 member, provides evidence for recombination between subspecies B2 genomes within the second half portion of the hexon gene, previously unreported, and HAdV-E4, species E, within the first half portion, an interspecies recombination event previously not known. As HAdV-B16 is a candidate human gene transfer vector and HAdV-E4 is an important human pathogen, understanding the role recombination plays in the evolution and pathoepidemiology of HAdV has applications in the rational design of vaccines and for gene delivery vector development.

Introduction

Human adenoviruses (HAdVs) have been characterized using available assays since the 1950s (Hilleman et al., 1954; Rowe et al., 1953). As pathogens, they are of interest because they may occur in highly contagious outbreaks infecting 100s (Binn et al., 2007; Engelmann et al., 2006; Ishiko et al., 2008) and can also cause a wide range of
diseases, including respiratory, ocular, gastrointestinal and metabolic (Echavarria et al., 2009). There are 70 different types partitioned into six (A-F) species based on biology, immunochemistry and recombinant DNA methodologies, with genomics recently providing additional prototypes for defining a new species G (Echavarria et al., 2009; Ishiko et al., 2008; Jones et al., 2007; Walsh et al., 2009).

Genomics and bioinformatics are high-resolution approaches and resources that are available for studying HAdVs at the primary nucleotide sequence level. Computational analysis of the genome data is providing insights into the molecular evolution of HAdVs, with the three novel “types” (HAdV-G52, D53 and B55) characterized and christened using sequence analysis (Jones et al., 2007; Walsh et al., 2009a; Walsh Seto et al., 2009) rather than serological characterization. Two of these are the results of genome recombination (Walsh et al., 2009a; Walsh et al., 2009b), providing support for the hypothesis that recombination is a major driving force for novel prototypes (Crawford-Miksza et al., 1996). This was also recently suggested by the limited analyses of sixteen species C field isolates by serology (Lukashev et al., 2008). A re-examination of archived prototypes by genomics provides additional support that novel HAdV arise as the result of recombination. An example is a penton base recombination characterized in HAdV-D22 (Robinson et al., 2009).

This project describes two partial hexon recombinations uncovered during the computational analysis of HAdV-B16. One is as an acceptor and the other is as a donor. The donated sequence was from HAdV-E4 (Dehghan et al., 2013), representing an interspecies event. Previous hypotheses and limited laboratory data indicated
interspecies recombination does not occur for the HAdV genomes (Lukashev et al., 2008; Wadell et al., 1980; Williams et al., 1975).

Species HAdV-B is subdivided into two subspecies; using molecular biological techniques and taking into account their biology as well as proteome differences (Echavarria et al., 2009; Wadell et al., 1984; Wadell et al., 1980). Members of B1 (HAdV-B3, B7, B16, B21 and B50) are human acute respiratory disease (ARD) pathogens, with two exceptions, HAdV-B50 (De Jong et al., 1999) and HAdV-B16. HAdV-B16 was isolated originally from conjunctival scrapings in 1955 and recognized as a new serotype (Bell et al., 1959; Hierholzer et al., 1991, Pereira et al., 1963). It was therefore associated with ocular disease (Bell et al., 1959; Feng et al., 1959) originally, but subsequently recognized as a respiratory pathogen as well, causing pharyngoconjunctival fever (Echavarria et al., 2009), pneumonia (Morgan et al., 1984) and other respiratory disease (Metzgar et al., 2005). HAdV-B16 appears to be either an underreported or uncommonly encountered HAdV (D’Ambrosio et al., 1982; Metzgar et al., 2005; Morgan et al., 1984) as there are few mentions in the literature. Members of subspecies B2, with the exception of HAdV-B14 (Louie et al., 2008; Van Der Veen et al., 1957) and HAdV-B55 (formerly “HAdV-B11a”) (Walsh et al., 2010; Yang et al., 2009; Zhu et al., 2009), are not associated with respiratory disease (Echavarria et al., 2009).

Aside from its pathogenicity, HAdV-B16 is of interest as a vector candidate for gene delivery in gene therapy protocols (Skog et al., 2007). It is reported to infect human low-passage brain tumor cells as well as cancer stem cells, unlike HAdV-C5, giving it an important advantage.
**Materials and Methods**

HAdV-B16 is archived at the American Type Culture Collection (ATCC; Manassas, VA) as “VR-17”, strain ch. 79. This was obtained and processed using a protocol described for similar HAdV sequencing projects (Purkayastha et al., 2005; Seto et al., 2009), with virus growth in A-549 cells and DNA purification outsourced to Virapur, LLC. (San Diego, CA). Commonwealth Biotechnologies, Inc. (Richmond, VA) sequenced the genome, using the Sanger chemistry with the DYEnamic ET Terminator Cycle Sequencing kit (Amersham Biosciences; Piscataway, NJ); sequence ladders were resolved on an ABI Prism 377 DNA Sequencer (Applied Biosystems; Foster City, CA); and assembled it using DNA Sequencher (GeneCodes, Inc.; Ann Arbor, MI). A minimum of three-fold sequences, covering both directions, with overall five-fold coverage, was supplemented with PCR-driven amplification and re-sequencing of areas that were found to be ambiguous upon sequence assembly and genome annotation.

**Results**

**Genome sequence analysis**

HAdV-B16 has a genome of 35,522 nucleotides with a GC content of 51%, consistent with subspecies B1 (51%) and differing from species B2 (49%) and other HAdV species; GC% is a species-defining criterion. Its genome is approximately 94% identical to the HAdV-B1 members and 82% to HAdV-B2 members, with lower
identities to other serotypes and species (the next highest is HAdV-E4 at 72%). It contains two VA-RNAs, as well as other genome and proteome attributes, that fit it into subspecies B1 rather than B2 (Wadell et al., 1980; Wadell et al., 1984). Mapping of HAdV-B16 proteins is shown in **Figure 2**.

**Figure 2** - HAdV-B16 whole genome mapping. Whole genome protein mapping of HAdV-B16 with GeneMap software on binf.gmu.e.du/eliul/genemap/
**Comparative genome analysis**

Initial comparative sequence analysis was performed using zPicture ([http://zpicture.dcode.org/](http://zpicture.dcode.org/)), a blastz algorithm based dynamic alignment visualization tool, to align the HadV-B16 hexon genome against each of the eleven sequences from the respiratory pathogens of species E, subspecies of B1 and B2 members. The zPicture genome results showed sequence divergence across the genomes for the B1 genomes, including within the hexon sequence. However, the B2 genomes showed sequence conservation at the distal portion of the hexon gene, with similarity levels higher than the B1 genomes (93% vs. 88%). Interestingly, HAdV-E4 showed a low similarity at the distal portion, 81%, but a much higher level 97%, at the proximal end. The zPicture sequence comparisons of the hexon sequences are shown in **Figure 3**.

Further analysis of HAdV-B16 included using the EMBOSS NEEDLE pairwise sequence alignment tool (EMBL-EBI 2015) to calculate percent identity to each of the eleven genomes to provide a more detailed relationship between each of the major protein genes. The findings are consistent with zPicture results. The hexon sequences showed that the proximal HAdV-B16 sequence that is highly similar (94.6%) to its HAdV-E4 counterpart, comprising approximately 900 nucleotides that represent 31.9% of the HAdV-B16 hexon gene and 2.53% of the whole genome. This is in line with similar hexon recombinants reported in the literature for HAdV-D53 and B55 (Walsh et al., 2009; Walsh et al., 2010).
Figure 3 - zPicture Analysis of HAdV-B16.
The HAdV-B16 hexon gene was aligned, using zPicture, against subspecies B1 (HAdV-B3, B7, B21, B50 and SAdV-B21), subspecies B2 (HAdV-B11, B14, B34, B35, and B55), and HAdV-E4. The x-axis ranges from nucleotide 1 to 2800. Numbers along the y-axis represent the percent identity from 50% to 100%. Hypervariable regions L1 and L2 are indicated at the top of the alignments, for reference and are approximate locations. GenBank accession numbers: HAdV-B16 (AY601636), HAdV-B3 (AY599836), HAdV-B7 (AY594255), HAdV-B21 (AY601633), HAdV-B50 (AY737798), SAdV-B21 (AC_000010), HAdV-B11 (AC_000015), HAdV-B14 (AY803294), HAdV-B34 (AY737797), HAdV-B35 (AC_000019), HAdV-B55 (FJ643676) and HAdV-E4 (AY594253).
In contrast, the distal sequences of HAdV-B16 showed 92.5% identity with HAdV-E4, where the zPicture alignments showed that the distal portion of the HAdV-B16 hexon as having high similarities to the subspecies B2 sequences, for example percent identity results showed 98% identity with HAdV-B11, HAdV-B34 and HAdV-B35; 97% with rest of the HAdV-B species, see Figure 3. This region represents approximately 34.0% of the hexon and 2.70% of the genome (Table 2). Also noted is that HAdV-B16 has the highest percent identity of 95% with HAdV-B3 and B7 in the penton base gene, while the rest of percent identities of the reference genomes are in the low to mid 80%. The fiberknob region of HAdV B-16 also shows the highest percent identity to Simian Adenovirus (SAdV) 35.1 and 35.2 of 97.4% and 97.9% respectively.

Table 2 - Percent Identity of HAdV-B16.

<table>
<thead>
<tr>
<th>Species</th>
<th>Types</th>
<th>Penton base</th>
<th>Hexon</th>
<th>Hexon Proximal</th>
<th>Hexon Distal</th>
<th>Fiber knob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>HAdV-B3</td>
<td>95.0</td>
<td>86.8</td>
<td>75.4</td>
<td>97.4</td>
<td>61.8</td>
</tr>
<tr>
<td>Human</td>
<td>HAdV-B7</td>
<td>95.0</td>
<td>86.8</td>
<td>75.2</td>
<td>97.4</td>
<td>52.3</td>
</tr>
<tr>
<td>Human</td>
<td>HAdV-B21</td>
<td>84.8</td>
<td>85.2</td>
<td>72.4</td>
<td>97.2</td>
<td>52.2</td>
</tr>
<tr>
<td>Simian</td>
<td>SAdV-B21</td>
<td>84.0</td>
<td>85.0</td>
<td>71.8</td>
<td>97.6</td>
<td>79.4</td>
</tr>
<tr>
<td>Human</td>
<td>HAdV-B50</td>
<td>84.7</td>
<td>86.1</td>
<td>74.2</td>
<td>97.2</td>
<td>51.9</td>
</tr>
<tr>
<td>Human</td>
<td>HAdV-B11</td>
<td>84.2</td>
<td>85.5</td>
<td>72.1</td>
<td>98.0</td>
<td>51.7</td>
</tr>
<tr>
<td>Human</td>
<td>HAdV-B14</td>
<td>85.1</td>
<td>85.4</td>
<td>72.2</td>
<td>97.8</td>
<td>50.9</td>
</tr>
<tr>
<td>Human</td>
<td>HAdV-B34</td>
<td>85.9</td>
<td>85.2</td>
<td>71.6</td>
<td>98.0</td>
<td>51.9</td>
</tr>
<tr>
<td>Human</td>
<td>HAdV-B35</td>
<td>84.1</td>
<td>85.3</td>
<td>71.8</td>
<td>98.0</td>
<td>51.9</td>
</tr>
<tr>
<td>Simian</td>
<td>SAdV-B35.1</td>
<td>84.5</td>
<td>84.7</td>
<td>71.7</td>
<td>97.0</td>
<td>97.4</td>
</tr>
<tr>
<td>Simian</td>
<td>SAdV-B35.2</td>
<td>84.5</td>
<td>84.7</td>
<td>71.7</td>
<td>97.0</td>
<td>97.9</td>
</tr>
<tr>
<td>Human</td>
<td>HAdV-B55</td>
<td>85.4</td>
<td>85.4</td>
<td>72.0</td>
<td>97.8</td>
<td>50.6</td>
</tr>
<tr>
<td>Human</td>
<td>HAdV-E4</td>
<td>80.2</td>
<td>93.5</td>
<td>94.6</td>
<td>92.5</td>
<td>29.7</td>
</tr>
<tr>
<td>Simian</td>
<td>SAdV-27.1</td>
<td>84.9</td>
<td>85.2</td>
<td>74.1</td>
<td>96.0</td>
<td>55.7</td>
</tr>
</tbody>
</table>
**Hexon recombination**

To explore the zPicture and percent identity results in greater detail, and to examine its origins, the HAdV-B16 genome was examined using software to detect sequence recombination, SimPlot 3.5.1 ([http://sray.med.som.jhmi.edu/SCSoftware/SimPlot](http://sray.med.som.jhmi.edu/SCSoftware/SimPlot)). SimPlot calculates and plots the percent identity of the query sequence to a panel of reference sequences in a sliding window, which move across the alignment of steps (Lole et al., 1999). This was approached initially by using whole genome and hexon sequences from all of the sequenced HAdVs, and then narrowing the eventual query set to the B1, B2 and E genomes. Bootscan analysis, an option of SimPlot, revealed a high degree of similarity of the proximal portion of the HAdV-B16 hexon with HAdV-E4 (Figure 4A). SimPlot analysis of the distal portion of the HAdV-B16 hexon indicated conservation to the B2 species collectively (Figure 4B). Iterations of recombination analyses using each of the B2 genomes separately showed each contributed equally to the proximal recombination as shown with the Bootscan analyses (Figure 5A). Together, these provide evidence for a recombination event with an ancestral species B2 genome at the distal portion of hexon (Figure 5B), unlike the two other HAdV recombinations described recently involving the proximal portions of the hexon gene (Walsh et al., 2009; Walsh et al., 2010). Bootscan analysis is an option of SimPlot, with adjustable window, step and repeat sizes; it repeatedly generates bootscan phylogenetic trees using random halves of the sequence within a given window. The reference sequence that has been clade with the query sequence the most number of times will be represented on the top of the plot.
Figure 4 - Whole genome recombination analysis of HAdV-16.
A) Whole genome Bootscan analysis of HAdV-B16 with representative HAdV genomes, showing sequence similarity to HAdV-E4 at the hexon sequence, with all four subspecies B2 genomes contributing equally and diluting out the similarity. (window size 1000bp, step size 200bp, repeat 100) B) Whole genome SimPlot analysis of HAdV-B16 with representative HAdV genomes. (window size 1000bp, step size 200bp, repeat 100).
Figure 5 - Whole genome recombination analysis of HAdV-16.

A) Hexon Bootscan analysis of HAdV-B16 with representative HAdVs. This is a composite, using HAdV-B14 as a representative of the subspecies B2. Additional iterations with each B2 shows the identical pattern, and inclusion of all B2 members “competed” out the high similarity. (window size 200bp, step size 20bp)

B) Hexon SimPlot analysis of HAdV-B16 with representative HAdVs, showing the subspecies B2 sequence contributions. (window size 200bp, step size 20bp) GenBank accession numbers, in addition to ones noted earlier, are as follows: HAdV-A12 AC_000005, HAdV-C5 AC_000008, HAdV-D9 AJ854486, HAdV-D53 FJ169625, HAdV-F40 NC_001454, HAdV-G52 DQ923122.
The second recombination event, involving the proximal portion of the hexon, suggests HAdV-B16 contributed to the evolution of HAdV-E4 (Dehghan et al., 2013), or vice versa. This is hypothesized as a host adaptation event as HAdV-E4 is the only human HAdV of species E with the rest being chimpanzee adenoviruses (SAAdV-E22 to E25). The virus was reported as originating from a zoonotic event (Purkayastha et al., 2005).

Given the whole genome recombination analysis, without clear similarities to other genomes, it is likely the HAdV-B16 genome is an “ancient” sequence that has accumulated enough nucleotide changes to show divergence from other genomes as shown in the recombination analysis. The distal B2 recombination is likely a “recent” event, as it retains a high level of identity to the B2 sequences, e.g., without subsequent accumulated nucleotide changes. The “donation” of the proximal sequence to HAdV-E4 is likely recent as well, from the HAdV-E4 genome perspective (Dehghan et al., 2013).

**Phylogeny analysis**

Molecular Evolutionary Genetics Analysis (MEGA) 4.0.2 (Tamura et al., 2007) was used for phylogenetic analysis using bootstrap-confirmed neighbor-joining trees of the HAdV genomes, hexon gene and its parts, penton base gene and fiber gene, allowing a detailed examination of HAdV evolution and providing an additional view of these recombination events (data not shown). **Figure 6** displays a portion comprehensive whole genome phylogeny analysis tree. Shown in this phylogenetic snapshot are the B1 and B2 genomes forming a subclade together, as species B, but branching separately as
subspecies. As expected, HAdV-B16 subclades with species B1 members. HAdV-E4 is also shown in a subclade separated from species B and the other HAdV species. This clade includes other members of species E, the chimpanzee adenoviruses (Purkayastha et al., 2005; Purkayastha et al., 2005) (data not shown).

Figure 6 - Whole genome phylogenetic analysis of HAdV-B16. The phylogenetic tree was constructed from aligned sequences using MEGA, via the neighbor-joining methods and a bootstrap test of phylogeny. Bootstrap values shown at the branching points indicate the percentages of 1000 replications produce the clade. A Bootstrap value of 70 and above is considered to be robust.
Phylogeny analysis of the hexon genes (Figure 7) shows HAdV-B16 in the same clade as, but branching away from, HAdV-B3 and B7. It has a similarity to the B2 subclade and the chimpanzee SAdV-B21 and HAdV-B21 subclade. Dividing the hexon sequence into proximal and distal subsequences, (Figure 8A and 8B), defined in the zPicture analysis, the proximal portion subclades with HAdV-E4. The distal portion of hexon branches with B2 subclade. Both hexon and fiber phylogeny trees presented the same results as the whole genome tree, the HAdV-B16 genes subclade with B1 members and away from both species E and subspecies B2 (data not shown). These results are consistent with and support the findings of the zPicture and the Bootscan analyses.

![Hexon phylogenetic analysis of HAdV-B16.](image)

**Figure 7** - Hexon phylogenetic analysis of HAdV-B16. Bootstrap neighbor-joining trees hexon gene sequence relationships, with the species B members and representatives of the other HAdV species for reference.
The phylogenetic analysis of HAdV-B16 fiber portion especially the fiber knob region appears to be grouped closest to the simian adenoviruses, SAdV-B35.1, SAdV-B35.2 and then to SAdV-B21, away from rest of the HAdV-Bs, the closest human adenovirus is HAdV-B3, while rest of the HAdV-Bs were branched separately (Figure 9).
Figure 9 - Fiber knob phylogenetic analysis of HAdV-B16. Phylogenetic tree was constructed from aligned sequences using MEGA, via the neighbor-joining method and a bootstrap test of phylogeny.

Discussion

Recombination, molecular evolution and new serotype

As double-stranded DNA viruses, HAdV genomes are relatively stable with minor nucleotide changes, such as base substitution and insertion/deletions. This was documented in genome sequence comparisons between prototype and vaccine strains of HAdV-E4 and B7 (Purkayastha et al., 2005), as well as with a recent HAdV-B7 field
strain and its prototype (Seto et al., 2010). A similar result was found for longer time-
spans, e.g., across minimum of fifty years of circulation, for five HAdV-B3 genomes
(Mahadevan et al., 2010), and forty-two years, shown for several strains of HAdV-B7,
which were assayed for antigenic differences, presumably reflecting strain variations, by
serum neutralization tests and hexon sequencing (Crawford-Miksza et al., 1999); this
contrasted with the HAdV-E4 genomes noted in the same study.

HAdV genomes also undergo relatively large-scale changes, as recombination,
which is noted as antigenic shifts. As the hexon is a target for neutralizing antisera
(Toogood et al., 1992), it would be expected that changes in this epitope would result in
altered serological response. In the past, this signified a novel HAdV serotype. A
striking example is demonstrated recently for the newly recognized HAdV-B55
(misnamed as “HAdV-B11, QS” and “HAdV-B11a” (Zhu et al., 2009)). This is a
recently reanalyzed HAdV identified as the recent re-emergent pathogen responsible for a
highly contagious ARD outbreak (Walsh et al., 2010; Yang et al., 2009; Zhu et al., 2009)
in which a recombination of the proximal portion of the hexon results in a change in
serum neutralization patterns. The genome “chassis” of HAdV-B55 is predominantly
HAdV-B14 (97.4%), with the partial hexon sequence of HAdV-B11 (2.6%) allowing it to
serotype as HAdV-B11, a renal and urinary tract virus (Li et al., 1999; Numazaki et al.,
1968). The genome recombination resolves the cell tropism riddle of “B55” having the
cell tropic characteristics of B14 while serotyping as “B11” (Li et al., 1991; Mei et al.,
1998). The original naming of HAdV-B55 as “HAdV-B11a” by serology was also
inconsistent with its biology and pathology as a respiratory pathogen, which archetype HAdV-11 was not.

In this study, HAdV-B16 is revealed to present a unique recombination event. To date, the only two hexon recombinants characterized in genomic detail involve a transfer of the proximal sequence (Walsh et al., 2009, Walsh et al., 2010), in contrast to this HAdV-B16 recombination. The distal portion of an ancestral B2 HAdV genome is shuffled, resulting in a “then-new” HAdV that warranted the recognition of a new serotype in the 1950s (Bell et al., 1959; Hierholzer et al., 1991; Pereira et al., 1963). This study also shows that HAdV-B16 contributed a proximal portion of its hexon gene to HAdV-E4, an interspecies event (the second such noted), perhaps allowing that virus to adapt and optimize to a human host following a zoonotic infection from chimpanzees to humans (Purkayastha et al., 2005; Purkayastha et al., 2005). In addition, this genome analysis explains earlier reports, and resolves a riddle, that of the “bilateral cross-neutralization observed between Ad16 of subgenus B and Ad4 of subgenus E” reported in the literature twenty-five years ago (Hierholzer et al., 1991).

HAdV genome recombination has been reported in the literature (Williams et al., 1975). Using molecular typing and/or serological techniques to characterize new variants and more virulent strains, such as HAdV-B7h (Kajon et al., 1996), as well as “intertypic” or “intermediate” strains (Boursnell et al., 1981; Engelmann et al., 2006; Hierholzer et al., 1988; Hierholzer et al., 1976; Ishiko et al., 2008), recombinants have been suggested. Recombination was noted as a frequent event during the analyses of sixteen species C field isolates (Lukashev et al., 2008) and noted in a survey of novel AIDS-associated
HAdVs (Crawford-Miksa et al., 1996). Recombinant HAdV genomes have also been generated in vitro (Boursnell et al., 1981; Mautner et al., 1984). Given all these observations, recombination was hypothesized as a driving force for the molecular evolution of new HAdV serotypes (Crawford-Miksa et al., 1996). Recent reports using genomics have reconfirmed this, e.g., identified, characterized and christened HAdV-D53 and HAdV-B55 as novel recombinant HAdV, with partial proximal hexon transfers (Walsh et al., 2009; Walsh et al., 2010). Re-analysis of HAdV-22 by genomics and computational methods shown this is also a recombinant, albeit at the penton base gene (Robinson et al., 2009). Genomics has extended and refined these observations, for example, disproving a hypothesis with limited molecular typing data for sixteen species C field isolates that interspecies recombination does not occur for “available complete genome sequences of AdB, AdC and AdD species” (Lukashev et al., 2008).

It should be noted a recombination does not automatically define nor necessitate the recognition of a new “type”. An example of this is described for two recently isolated and characterized field strains of HAdV-E4 (Dehghan et al., 2013).

**Conclusion**

Genomic and bioinformatics comparisons of the HAdV-B16 genome to other HAdV genomes identified two exclusive and partial hexon gene recombination events, one as an interspecies donator to HAdV-E4 and the other as an acceptor of an ancestral subspecies B2 sequence. The former event explains the observed serological cross-
reaction, noted in the literature, with HAdV-E4. This is a unique snapshot of the molecular evolution of HAdV, and represents an exception to the previous hypotheses and observations that HAdV genomes did not shuffle sequences across species.
CHAPTER 3 – COMPUTATIONAL ANALYSIS OF HUMAN ADENOVIRUS TYPE 21

Abstract

One driving force of human adenoviruses (HAdVs) molecular evolution is genome recombination resulting in emergent and novel pathogens in the past and the present. HAdV-B21 is a human respiratory pathogen that is a recombinant containing a large genomic sequence, including the major capsid proteins penton base and hexon, that has near sequence identity with genomic sequences identified from both chimpanzee (SAdV-B35.1) and bonobo (SAdV-B35.2) AdVs. Understanding the role that recombination plays in adenovirus evolution and pathoepidemiology is important in vaccine development, along with their long-term effectiveness and in the development of gene therapy vectors, using SAdVs as an alternative to viruses with pre-existing immunity in humans. Genome recombination provides the realization that non-human simian species are reservoirs for potentially highly contagious and deadly human adenoviral pathogens through zoonosis.

Introduction
This study of the detailed analysis of HAdV-B21 is a genomic examination of an archived prototype, *circa* 1950s. The purpose of this genomic analysis is very similar to the analysis of HAdV-B16 discussed in Chapter 2, which is to support additionally that novel HAdVs may arise from genome recombination. HAdV-B21 also belongs to subspecies B1, whose members are usually responsible for respiratory diseases. There has been a report that shows an increase in the incidence of fatal adenovirus infections (Rowe et al., 1953). A number of those severe disease cases have been linked to HAdV-B21 (Lahm et al., 2010). The exact etiology for this unexpected high mortality remains unknown; the referenced case reports a patient with severe pneumonia resulting in hemophagocytic lymphohistocytosis (HLH) with acute respiratory distress syndrome and rapid progressive multi-organ dysfunction syndrome. It was proposed that an association between HAdV-B21 and HLH may, at least in part, explain the recent observed increase in incidence of fatal adenoviral infection (Lahm et al., 2010). To explore whether genome recombination is involved in the molecular evolution of HAdV-B21 in general, this study reexamines the original isolate using genomic and bioinformatics.

**Material & Method**

HAdV-B21 (AV-1645) was purchased from the American Type Culture Collection (ATCC; Manassas, VA). This virus was processed using protocols described for similar HAdV sequencing projects (Lauer et al., 2004; Purkayastha et al., 2005a; Seto
et al., 2009) by Virapur, LLC. (San Diego, CA); these include growth in A-549 cells and subsequent DNA purification.

Genome sequencing was outsourced to Commonwealth Biotechnologies, Inc. (Richmond, VA), applying the Sanger chemistry with the DYEnamic ET Terminator Cycle Sequencing kit (Amersham Biosciences; Piscataway, NJ); ladders were resolved on an ABI Prism 377 DNA Sequencer (Applied Biosystems; Foster City, CA); and assembled using DNA Sequencher (GeneCodes, Inc.; Ann Arbor, MI). Across the genome, an average of five-fold sequencing and a minimum of three-fold coverage, and both directions. Re-sequencing of areas that were found to be questionable upon sequence assembly and genome annotation was PCR-driven amplification and sequencing. Quality control included genome annotation, with comparisons to earlier types 1, 4, and 7 genome data, including the prototype and vaccine (Lauer et al., 2004; Purkayastha et al., 2005a; Purkayastha et al., 2005b). Annotation was performed using the Genome Annotation Transfer Utility (GATU) software tool (Tcherepanov et al., 2006), and recorded and visualized using Artemis, a genome viewer (http://www.sanger.ac.uk/resources/software/artemis/) (Rutherford et al., 2003).

Whole genome sequences used in this analysis are listed here along with their accession number: HAdV-B21 (AY601633), HAdV-B16 (AY601636), HAdV-B3 (AY599836), HAdV-B7 (AY594255), HAdV-B50 (AY737798), SAdV-B21 (AC_000010), HAdV-B11 (AC_000015), HAdV-B14 (AY803294), HAdV-B34 (AY737797), HAdV-B35 (AC_000019), SAdV-B35.1 (FJ025912), SAdV-B35.2 (FJ025910), HAdV-B55 (FJ643676) and HAdV-E4 (AY594253).
Results

**Genome sequence analysis**

HAdV-B21 has a genome size of 35,382 nucleotides with a GC content of 51%, consistent with subspecies B1 (51%) and differing from B2 (49%) and other HAdV species; GC% is a species criterion. Its genome is approximately 95% and 83% identical to the HAdV-B1 and HAdV-B2 members respectively, with much lower identities to other species and serotypes, next highest is HAdV-E4 at 74%. The virus fits into subspecies B1 instead of B2 as it contains two VA-RNAs, as well as other genome and proteome attributes as noted by earlier reports (Wadell et al., 1980; Wadell et al., 1984).

**Comparative whole genome analysis**

Comparative whole genome sequence analysis was performed using zPicture. This dynamic alignment tool aligns the HAdV-B21 genome against genomes of the B1 and B2 members, all HAdVs, along with SAdV-B21, SAdV-B35.1 and SAdV-B35.2, which are from non-human simian hosts. The zPicture whole genome results showed sequence divergence across the genomes for the B1 genomes and the SAdV genomes (Figure 10). The HAdV-B21 genome is almost identical to HAdV-B50 genome, except for the divergence in the proximal hexon region. Rather, the proximal hexon region shows high sequence conservation with SAdV-B35.1 and SAdV-B35.2 instead.
The HAdV-B16 hexon gene was aligned, using zPicture, against HAdV-B50, B16, B7 and SAdV-B21, B35.1 and B35.2. The x-axis ranges from nucleotide 1 to 36000. Numbers along the y-axis represent the percent identity from 50% to 100%. Protein regions are indicated at the top of the alignments, for reference and are approximate locations.

Figure 10 - zPicture Analysis of HAdV-B21.
The HAdV-B16 hexon gene was aligned, using zPicture, against HAdV-B50, B16, B7 and SAdV-B21, B35.1 and B35.2. The x-axis ranges from nucleotide 1 to 36000. Numbers along the y-axis represent the percent identity from 50% to 100%. Protein regions are indicated at the top of the alignments, for reference and are approximate locations.

Each surface protein of HAdV-B21 was further analyzed by calculating its percent identity to the rest of the reference genomes mentioned above. These results are shown in Table 3. HAdV-B21 penton base region has a percent identity of 99.3 to HAdV-B50, SAdV-B35.1 and SAdV-B35.2. The hexon region has the highest percent identity of 98.0 and 98.1 to SAdV-B35.1 and B35.2 respectively. Upon closer
examination, the proximal portion of HAdV-B21 only has 84.7% identity of HAdV-B50 and 99.3% identity to SAdV-B35.1 and B35.2. This region comprises approximately 900 nucleotides that represent 31.9% of the HAdV-B21 hexon and 2.5% of the total genome. In contrast the distal hexon region is 100% identical to HAdV-B50, and 99.8% identity to SAdV-B35.1 and B35.2. The fiber knob region is well-conserved with HAdV-B50; it has a percent identity of 99.7% and diverges away from the SAdVs numbers.

**Table 3** - Percent identity of HAdV-B21. Percent identities of the nucleotide coding sequences of selected HAdV-B21 coding regions to homologous sequences from viruses in species HAdV-B, HAdV-E and SAdV-B.

<table>
<thead>
<tr>
<th>Species</th>
<th>Types</th>
<th>Penton base</th>
<th>Hexon</th>
<th>Hexon Proximal</th>
<th>Hexon Distal</th>
<th>Fiber knob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>HAdV-B3</td>
<td>85.8</td>
<td>85.5</td>
<td>74.9</td>
<td>95.6</td>
<td>57.0</td>
</tr>
<tr>
<td>Human</td>
<td>HAdV-B7</td>
<td>85.6</td>
<td>85.9</td>
<td>75.9</td>
<td>95.4</td>
<td>59.7</td>
</tr>
<tr>
<td>Human</td>
<td>HAdV-B16</td>
<td>84.8</td>
<td>85.2</td>
<td>72.4</td>
<td>97.2</td>
<td>52.2</td>
</tr>
<tr>
<td>Simian</td>
<td>SAdV-B21</td>
<td>97.7</td>
<td>91.3</td>
<td>83.1</td>
<td>99.0</td>
<td>56.7</td>
</tr>
<tr>
<td>Human</td>
<td>HAdV-B50</td>
<td>99.3</td>
<td>92.6</td>
<td>84.7</td>
<td>100.0</td>
<td>99.7</td>
</tr>
<tr>
<td>Human</td>
<td>HAdV-B11</td>
<td>92.2</td>
<td>93.5</td>
<td>89.1</td>
<td>97.6</td>
<td>59.4</td>
</tr>
<tr>
<td>Human</td>
<td>HAdV-B14</td>
<td>92.0</td>
<td>90.2</td>
<td>82.5</td>
<td>97.4</td>
<td>60.0</td>
</tr>
<tr>
<td>Human</td>
<td>HAdV-B34</td>
<td>92.5</td>
<td>90.3</td>
<td>82.4</td>
<td>97.6</td>
<td>95.2</td>
</tr>
<tr>
<td>Human</td>
<td>HAdV-B35</td>
<td>92.3</td>
<td>91.3</td>
<td>84.6</td>
<td>97.6</td>
<td>94.8</td>
</tr>
<tr>
<td>Simian</td>
<td>SAdV-B35.1</td>
<td>99.3</td>
<td>98.0</td>
<td>96.1</td>
<td>99.8</td>
<td>53.4</td>
</tr>
<tr>
<td>Simian</td>
<td>SAdV-B35.2</td>
<td>99.3</td>
<td>98.1</td>
<td>96.3</td>
<td>99.8</td>
<td>53.1</td>
</tr>
<tr>
<td>Human</td>
<td>HAdV-B55</td>
<td>92.2</td>
<td>92.4</td>
<td>87.1</td>
<td>97.4</td>
<td>60.0</td>
</tr>
<tr>
<td>Human</td>
<td>HAdV-E4</td>
<td>83.9</td>
<td>83.3</td>
<td>72.9</td>
<td>92.9</td>
<td>26.0</td>
</tr>
</tbody>
</table>

**Hexon recombination**
To further explore the picture and percent identity results in greater detail, and to examine HAdV-B21’s origins, a sequence recombination detection software SimPlot was used to calculate and plots the percent identities of the HAdV-B21 to a panel of reference
sequences in a sliding window that moves across in steps of the alignment in order to detect recombination (Figure 11A and 11B). A closer examination of the hexon region is done by SimPlot and Bootscan analyses, they displayed a high degree similarity of the proximal half of HAdV-B21 with SAdV-B35.1 and B35.2; distal half of the HAdV-B16 indicated conservation to HAdV-B50 (Figure 12A and 12B). Due to the near-identical percent similarity between SAdV-B35.1 and B35.2 that will compete against each other in Bootscan graph and obscuring the true result, SAdV-B35.1 was excluded and SAdV-B35.2 was used in this figure to represent the recombination event.

Given the whole genome recombination analysis and the hexon recombination analysis, it is likely that HAdV-B21 is an ancient genome sequence that has accumulated nucleotides changes across species and shown divergence from other genomes. The proximal SAdV-B35 recombination contains a high level of identity to HAdV-B16 which can be a recent event. The zoonotic transfer is consistent with the ones observed in previous studies, with the direction of the zoonosis undetermined (Dehghan et al., 2013).
Figure 11 - Whole genome recombination analysis of HAdV-21.

A) Whole genome Bootscan analysis of HAdV-B21 with representative HAdV genomes, showing sequence similarity to HAdV-B50 at the hexon sequence. (window size 1000bp, step size 200bp, repeat 100)

B) Whole genome SimPlot analysis of HAdV-B21 with representative HAdV genomes, shows high similarity of SAdV-B35.1 and B35.2 at the distal region of hexon (window size 1000bp, step size 200bp, repeat 100).
Figure 12 - Whole genome recombination analysis of HAdV-21.  
A) Hexon Bootscan analysis of HAdV-B21 with representative HAdVs. This is a composite, using HAdV-B35.2 as a representative of the SAdV-B35. Additional iterations with each SAdV-B35.1 shows the identical pattern, and inclusion of all SAdV-B35 members “competed” out the high similarity. (window size 200bp, step size 20bp)  
B) Hexon SimPlot analysis of HAdV-B21 with representative HAdVs, shows the SAdV-B35 sequence contributions. (window size 200bp, step size 20bp)
Phylogenetic analysis

MEGA was used for phylogenetic analysis with bootstrap confirmed neighbor joining trees of the HAdV genomes, penton base, hexon and its parts, and the fiber knob region. This provides an additional view of the recombination events and allowing a closer examination of the HAdV evolution. Whole genome phylogenetic analysis shows B1 and B2 genomes forming a subclade together as part of the species B but also branching separately with SAdV-Bs in between (Figure 13).

![Figure 13](image-url) - Whole genome phylogenetic analysis of HAdV-B21.
The phylogenetic tree was constructed from aligned sequences using MEGA, via the neighbor-joining methods and a bootstrap test of phylogeny. Bootstrap values shown at the branching points indicate the percentages of 1000 replications produce the clade. A Bootstrap value of 70 and above is considered to be robust.
Phylogenic analysis of the hexon gene (Figure 14) shows HAdV-B21 in the same clade but branching away from, as rest of the B1s. It has higher similarity to SAdV-B35.1, B35.2 subclades and SAdV-B21, HAdV-B50 subclades. For detailed analysis, the hexon sequence is then divided into the proximal and distal subsequences (Figure 15A and 15B) defined in zPicture. This shows that the proximal portion subclades with SAdV-B35.1 and B35.2 while the distal portion subclades with HAdV-B50. These results are consistent with and support the findings from the zPicture and SimPlot Bootscan analyses.

**Figure 14** - Hexon phylogenetic analysis of HAdV-B21.
Bootstrap neighbor-joining trees hexon gene sequence relationships, with the species B members and representatives of the other HAdV species for reference.
Figure 15 - Hexon halves phylogenetic analysis of HAdV-B21. 
A) Proximal portion of Hexon sequence phylogenetic relationships. B) Distal portion of Hexon sequence phylogenetic relationships. Phylogenetic trees were generated from aligned sequences using MEGA, via the neighbor-joining method and a bootstrap test of phylogeny.

The phylogenetic analysis of the fiber knob region shows HAdV-B21 clade closest to HAdV-B50 then to HAdV-B34 and B35. It branches away from rest of the Simian Bs and away from rest of the B1s and B2s. (Figure 16).
Discussion

With the applications of bioinformatics methods and genomic technology available today, we understand the relationship of infectious diseases and microbial pathogens in greater detail (Relman et al., 2011). Potentially emergent pathogens may be predicted using genome sequence data, as noted for a HAdV (Robinson et al., 2013). The recently availability of a large quantity of genomic data also provides valuable information to understanding the molecular evolution mechanism for emergent and novel pathogens (Robinson et al., 2009; Walsh et al., 2009).
In this study, HAdV-B21 presents a unique recombination event between human and simian adenoviruses. Although the recombination direction cannot be determined, DNA sequences are transferred laterally between human and simian viruses. These viruses from zoononic transfers may become optimized and adapted through evolution to the new host (Dehghan et al., 2013). HAdVs are important biomedical tools as vectors for gene and epitope delivery (Darr et al., 2009; Stone et al., 2006). Simian adenoviruses are increasingly being considered as alternative to its human strains due to seroprevalence concerns. There recent attentions to SAdVs as potential gene vectors have brought more genomes sequences into the data set (Roy et al., 2004), allowing for more comprehensive analyses and deeper insights, such as reported here for HAdV-B21.

**Conclusion**

Bioinformatics and genomic comparisons of HAdV-B21 genome to other HAdV genomes revealed two recombination events in the penton base and hexon genes with both chimpanzee (SAdV-B35.1) and bonobo (SAdV-B35.2) AdVs. Comprehending the role that recombination plays in adenovirus evolution is important in gene therapy vector development and in the development of vaccines. Similar to the previous chapter, this also represents another exception to the previous observations and hypotheses that HAdV genome did not exchange sequences across species.
CHAPTER 4 – HUMAN ADENOVIRUS TYPE 58

Introduction

Human adenovirus type 58 (HAdV-D58) is the first novel gastrointestinal adenoviral pathogen identified since the 1990s (De Jong et al., 1999). It was isolated from the stool sample of an AIDS patient who presented with severe chronic diarrhea. The virus contains a novel hexon gene coding sequence as well as a novel recombinant fiber gene. Serological analysis also demonstrated that HAdV-D58 has a different neutralization profile than all other HAdV characterized to date since the 1990s. The characterization of this gastrointestinal pathogen has public health significance, especially in Argentina and South America where it was first isolated, as well as other “developing” regions.
Genetic Analysis of a Novel Human Adenovirus with a Serologically Unique Hexon and a Recombinant Fiber Gene

Elizabeth B. Liu1, Leonardo Ferreyra2, Stephen L. Fischer3, Jorge V. Pavan4, Silvia V. Nates5, Nelan Ryan6, Damaris Tirado6, David W. Dyer6, James Chodoby6, Donald Seto7, Morris S. Jones8

1 Department of Bioinformatics and Computational Biology and Department of Systems Biology, Georgia Mason University, Manassas, Virginia, United States of America, 2Virology Institute, School of Medical Sciences, National University of Cordoba, Cordoba, Argentina, 3Hawaii Hospital Camp Pendleton, Camp Pendleton, California, United States of America, 4Clinical Investigation Facility, Dand Grant USAF Medical Center, Tavita AFB, Fairdale, California, United States of America, 5Department of Microbiology and Immunology, University of Oklahoma Health Sciences Center, Oklahoma City, Oklahoma, United States of America, 6Hawaii Laboratory, Massachusetts Eye and Ear Infirmary, Department of Ophthalmology, Harvard Medical School, Boston, Massachusetts, United States of America, 7Yale and Rickard Disease Laboratory, California Department of Public Health, Richmond, California, United States of America

Abstract

In February of 1996 a human adenovirus (formerly known as Ad-Cor-95-4188) was isolated from the stool of an AIDS patient who presented with severe chronic diarrhea. To characterize this apparently novel pathogen of potential public health significance, the complete genome of this adenovirus was sequenced to elucidate its origin. Bioinformatic and phylogenetic analyses of this genome demonstrate that this virus, hereafter referred to as HAdV-D58, contains a novel hexon gene as well as a recombinant fiber gene. In addition, serological analysis demonstrated that HAdV-D58 has a different neutralization profile than all previously characterized HAdVs. Bootstrap analysis of the HAdV-D58 fiber gene strongly suggests one recombination event.

Introduction

Human adenoviruses (HAdVs) were first isolated in 1953 from pediatric adenoid tissue and from a military basic trainee as respiratory pathogens [1] [2]. Since then, 56 types have been isolated and characterized [3, 4, 5, 6, 7]. Currently, there are 56 HAdVs in the genus Mastadenovirus in the family Adenoviridae, that are organized into seven species (A–G) [1, 4, 7, 8]. Individual HAdV types were originally differentiated based on immunocological or serological methods, but more recently, genomics and bioinformatic analyses have supplanted serology [8]. Members of each species are highly similar at the nucleotide level, and do not appear to recombine readily with members of another species. Species grouping also reflect a tendency for specific human diseases: for example, many HAdVs within species HAdV-D cause epidemic keratoconjunctivitis [9], whereas HAdVs in species HAdV-B are known to cause respiratory infections [10].

Currently there are three human adenoviruses (HAdV-F40, HAdV-F41 and HAdV-G25) that are associated with gastroenteritis [7, 8]. Gastroenteritis is associated with an estimated 5,000 deaths per year in United States [11]. It is likely that the virological agents of gastroenteritis include yet-to-be-identified pathogenic strains.

In this report we examined an adenovirus isolated from the stool of an AIDS patient who presented with severe chronic diarrhea. Based upon whole genomic and bioinformatic analysis, this virus appears to belong to species HAdV-D, with the proposed name HAdV-D58.

Results

Microbiological Investigation

In February 1996 an adenovirus was isolated from the stool of a 31-year-old AIDS patient who presented with severe chronic diarrhea and was subsequently hospitalized. Cryptosporidium parvum and Giardia lamblia were also found in the fecal matter of the patient; therefore, the clinical symptoms cannot be exclusively linked with the adenovirus infection.

Amplification and sequencing of the novel adenovirus

Partial sequencing of HAdV-D58, previously published as the Ad-Cor-95-4188 strain [12], via improved semen neutralization, demonstrated that portions of the hexon and fiber genes resembled HAdV-D33 and HAdV-D29, respectively [12]. This suggested that this novel HAdV isolated from an AIDS patient originated at least in part by recombination. To elucidate the genetic characteristics of HAdV-D58, the entire genome has been sequenced and analyzed.
Figure 1. Genome organization of HAdV-D58. Genome is represented by a central black horizontal line marked at 3400 bp intervals. Protein encoding regions are shown as arrowheads indicating transcriptional orientation. Forward arrows (above the horizontal black line) denote coding regions in the 5’ to 3’ direction and arrows pointing to the left (below the horizontal black line) denote coding regions in the 3’ to 5’ direction. Spliced genes are indicated by V-shaped lines.

doI:10.1371/journal.pone.0023491.g001

Physical features of the novel adenovirus genome

The genome length of HAdV-D58 is 35,218 base pairs (Fig. 1), with a base composition of 22.6% A, 20.3% T, 26.6% G, 28.4% C and with a GC content of 57.0%. The GC content is consistent with members of species *Homo sapiens* D (HAdV-D) (57.0% mean).

The organization of the 36 open reading frames (ORFs) that were annotated had a genome organization similar to other adenoviruses (Fig. 1). The inverted terminal repeat (ITR) sequences for HAdV-D58 were determined to be 160 bp in length. Within species HAdV-D, HAdV-D58 has a genome percent identity ranging from

Figure 2. SimPlot analysis of HAdV-D genomes to HAdV-D58. HAdV-D58 was compared to all fully sequenced HAdV genomes in species HAdV-D with SimPlot software. The arrows on the black line demarcate the approximate positions of the DNA polymerase, penton base, hexon, E3 encoding region, and fiber coding regions in the HAdV-D58 genome. Arrows pointing to the right are encoded in the 5’ to 3’ direction and arrows pointing towards the left are encoded in the 3’ to 5’ direction. The E3 box represents eight open reading frames.

doI:10.1371/journal.pone.0024491.g002
a low of 90.72% (HAV-D30, phylogenetic distance of 0.071) to 93.97% (HAV-D49, phylogenetic distance of 0.034).

Genomic recombination analysis

Comparison of HADV-D38 with the full-length genomes of viruses in species HADV-D using SimPlot analysis revealed significant sequence divergence in the hexon, E3, and fiber coding sequences (Fig. 2).

Genetic analysis of the novel adenovirus hexon coding sequences

Analysis of the HADV-D38 genome via pairwise comparison suggested that the hexon coding sequence was unlike any other known human adenovirus hexon sequence (Fig. 3). To determine if the hexon gene was novel, we performed SimPlot analysis using all hexon loop 1 (L1) and loop 2 (L2) coding sequences in species HADV-D. L1 and L2 contain the epitope [6] determinant, which contain the epitopes for serum neutralization [13]. SimPlot analysis confirmed that the hexon gene of HADV-D38 is unique compared with all other hexon genes in species HADV-D (Fig. 3). In terms of nucleotide identity, the L1 and L2 of HADV-D33 were most similar to HADV-D5 with 98.4 and 98.8% nucleotide identity, respectively (Table S1). No substantial evidence of recombination in the hexon coding sequence was revealed.

Analysis of the E3 genes

In the E3 region 19K, RIDa, RIDβ, and 14.7K are the only genes that have been investigated. The function of the E3/19K protein is to prevent human MHC class I molecules from being transported to the cell surface [14]. Specifically, amino acids W22, M87, and W56 were shown to be important for HLA-A1 modulation [14]. A second function of E3/19K is to inhibit NK cells from recognizing HADV-infected cells by sequestering MHC-I chain-related proteins A and B (MICA/B) [15]. The 14.7K protein product inhibits the internalization of TNF receptor 1 [16]. The RIDa and RIDβ proteins down-modulate the apoptosis receptor Fas/Apo-1 [17].

Bowscan analysis strongly suggests that there was a recombination event in the middle of the open reading frames of 19K and CR1-β (Fig. 4). These recombination events did not disrupt any of the E3 open reading frames in the HADV-D38 genome. Analysis of the 19K open reading frame in HADV-D38 demonstrated that amino acids W22, M87, and W56 were present (data not shown). The percent identities of the HADV-D38 19K, RIDa, RIDβ, and 14.7K open reading frames were 96%, 98.9, 98.4, and 97 percent identical to the homologous open reading frames of E3 coding sequences for HADV-D69/19K, HADV-D46/RIDa, HADV-D15/RIDβ, and HADV-D15-14.7K, respectively (Table S3).
Fiber recombination analysis

To determine whether or not there was recombination in the fiber gene of HAdV-D58, we performed Bootscan and SimPlot analysis using the fiber sequences is GenBank. Our results suggested the fiber gene of HAdV-D58 contains two recombination sites (Fig. 5). The first was in the middle of the shaft coding sequence and the second was in the shaft/knob boundary. The possible recombination at the shaft knob boundary is tenous since it is not possible to differentiate between HAdV-D25 and HAdV-D59 at this junction as evidenced by SimPlot analysis (Fig. 5B).

Phylogenetic analysis

Detailed phylogenetic analysis of completely sequenced HAdV genomes and selected coding sequences, performed with nucleotide data, confirmed that HAdV-D58 was a novel adenovirus (Figs. 6–8). The tree topology of HAdV-D58 was different depending on the protein analyzed. The whole genome sequence of HAdV-D58 was closest to HAdV-D59 (Fig. 6A). Using sequences available in GenBank, along with unpublished sequences, the penton base of HAdV-D58 grouped with HAdV-D8 "Triom", which is a prototype genome (Fig. 6B). Hexon loops 1 (L1) and 2 (L2) both clustered to HAdV-D33 (Fig. 7A and 7B), which was similar to results reported by Ferrer et al [11]. The fiber knob was tightly linked to HAdV-D25 (Fig. 8).

Viral neutralization

Since bioinformatic analysis showed that HAdV-D58 is genetically similar to previously typed HAdVs, correlating this data to its serum neutralization profile is important. Only antiserum to HAdV-D25, at a dilution of 1:32, was able to neutralize HAdV-D25 (Table 1). In contrast, antiserum to HAdV-D29 neutralized HAdV-D29 at 1:512 (Table 1). These results demonstrated that HAdV-D58 has a unique neutralization profile.

Serum neutralization vs. Phylogenetic analysis

A previous study proved that when the nucleotide identity of L2 in the hexon differs by 22.5%, a new HAdV type is suspected [13]. To provide a correlation between serum neutralization data, molecular typing (i.e., inferred serum neutralization), and phylogenetics data for the determination of a new HAdV type, the hexon L2 sequence of the proposed novel HAdV-D58 was compared against the L2 sequences of HAdV-33, -D9, and -D8 (the closest phylogenetic relatives of the HAdV-D58) [12]. The difference in percent nucleotide identity between L2 of HAdV-D58, -D9, and -D8 was 10.18, 20.73, and 25.99 percent, respectively. Thus, using the L2 sequencing criteria established by Madisch et al also demonstrates that HAdV-D58 is a new type.

Discussion

In the past, human adenoviruses were characterized primarily based on their serological profile and hemagglutination properties [18]. Today classification methods used for novel adenoviruses has been expanded to include whole genome sequencing and bioinformatic analysis [18]. We used whole genome sequencing, bioinformatic analysis, and serology to irrefutably demonstrate that HAdV-D58 is a novel human adenovirus type.

The serological and genomic characteristics of HAdV-D58 are unique. Specifically, the whole gene of HAdV-D58 was genetically dissimilar to all known HAdV hexon genes (Fig. 5). Furthermore,
Computational Analysis of a Novel Human Adenovirus

Figure 5. Computational analysis of the fiber regions. (A) Bootrace and (B) SimPlot analysis of the fiber region of HAdV-D38 compared to fully sequenced E3 and fiber regions from species HAdV-3.

doi:10.1371/journal.pone.0034495.g005

this was corroborated by neutralization data that demonstrated both 16- and 64-fold differences with antisera from HAdV-D29 and HAdV-D33, respectively (Table 1).

We found that the fiber gene of HAdV-D38 contains at least one recombination event and possibly a second (Fig 5). The second possible recombination site is located at the shaft/knob boundary. It is not currently possible to determine if recombination happened at the shaft/knob boundary (Fig. 5B). A prior study described a recombination hotspot in the fiber gene of species HAdV-D at the shaft/knob boundary [30]. However, our Bootstrap analysis on the same fiber coding sequences listed in Darr et al. [30], did not reveal evidence of recombination (Fig. 9). This result was also corroborated independently (personal communication Jason Seto). The analysis describing recombination in the fiber protein of HAdV-D37 and HAdV-D30 utilized consensus sequences for two of four alignments [30]. The problem with this analysis is that consensus sequences do not exist in nature and could induce artificial data when introduced into recombination analysis. Furthermore, the only way to re-create the supposed recombination events (proposed by Darr et al.) [30] that created HAdV-D38 was by combining sequences HAdV-D39-FM2010501 and HAdV-D23-FM2010506, which are 100% identical (Table 2), with the X' sequences of HAdV-D20-AJ11444 and HAdV-D23-AJ11446 (see Materials and Methods, respectively (Table 5). We were also able to recompute...
Figure 6. Phylogenetic analysis of whole genome, penton base, and E3 CR1-β in HAdV-D58. Phylogenetic analysis is based on the nucleic acid sequence of (A) whole genomes, (B) penton base, and (C) CR1-β. Phylogenetic trees were constructed from aligned sequences using MEGA via the neighbor-joining method and a bootstrap test of phylogeny. Bootstrap values shown at the branching points indicate the percentages of 1000 replications that produced the clad. doi:10.1371/journal.pone.0034491.g006

the proposed recombination event that created HAdV-D25 when we combined the sequences of HAdV-D25-FM2 H5142 and HAdV-DNS-FM2 H5143, which are also 100% identical (Table 2), with the 5 sequences of HAdV-D25-A11 H46 and HAdV-D25-A11 H44 (see Materials and Methods), respectively (Table 2). When this data is considered together, we find no consens evidence that the shuf/ knob junction is a hotspot for recombination.

For HAdVs, the number of E3 ORFs ranges between 6 and 9 [7,21]. HAdVs in species HAdV-D and HAdV-G contain the 49K/CR1-β ORF [7,21]. Interestingly, Bovman analysis suggests that the E3 region of HAdV-D58 was created by recombination with HAdV-D25 (Fig. 4). However, analysis of all species in the HAdV-D38E3 regions demonstrates that recombination hot spots do not exist in this part of the genome for species HAdV-D (data not shown). Thus, it is difficult to speculate what advantage there is for a seemingly random recombination in the E3 region.
A Hexon Loop1

B Hexon Loop2

Figure 7. Phylogenetic analysis of HAdV-D58 hexon loops 1 and 2. Analysis of HAdV-D58 hexon L1 and L2 is based on the nucleic acid sequence of (A) hexon L1, and (B) hexon L2. Phylogenetic trees were constructed from aligned sequences using MEGA via the neighbor-joining methods and a bootstrap test of phylogeny. Bootstrap values shown at the branching points indicate the percentages of 1000 replications produced the clad. doi:10.1371/journal.pone.0024491.g007

Materials and Methods

Ethics Statement

The work reported herein was performed under United States Air Force Surgeon General-approved Clinical Investigation No. FDG0900/0238, by the Institutional Review Board at the David Grant USAF Medical Center. Informed Consent was not required, because we did not use clinical samples.

Viruses, cells and neutralisation test

The isolation of HAdV-D58 (previously known as Ad-Cor-96-467) was previously described [12]. In brief, the stool sample was inoculated into Hep-2 cells and subcultured in Eagle's MEM supplemented with 10% of fetal bovine serum (FBS), penicillin (200 U/ml), L-glutamine (2 mM), Gentamicin (1 μg/ml), and streptomycin (200 μg/ml). HAdV-D58 was investigated serologi-
Fiber knob

Table 1. Serum neutralization of HAdV-DS8 with hyper immune serum.

<table>
<thead>
<tr>
<th>Antiserum</th>
<th>HAdV-DS8</th>
<th>HAdV-D29</th>
<th>HAdV-D33</th>
</tr>
</thead>
<tbody>
<tr>
<td>wAdV-D8</td>
<td>&lt;8</td>
<td>&lt;8</td>
<td>&lt;8</td>
</tr>
<tr>
<td>wAdV-D9</td>
<td>&lt;8</td>
<td>&lt;8</td>
<td>&lt;8</td>
</tr>
<tr>
<td>wAdV-D10</td>
<td>&lt;8</td>
<td>&lt;8</td>
<td>&lt;8</td>
</tr>
<tr>
<td>wAdV-D11</td>
<td>&lt;8</td>
<td>&lt;8</td>
<td>&lt;8</td>
</tr>
<tr>
<td>wAdV-D13</td>
<td>&lt;8</td>
<td>&lt;8</td>
<td>&lt;8</td>
</tr>
<tr>
<td>wAdV-D15</td>
<td>&lt;8</td>
<td>&lt;8</td>
<td>&lt;8</td>
</tr>
<tr>
<td>wAdV-D17</td>
<td>&lt;8</td>
<td>&lt;8</td>
<td>&lt;8</td>
</tr>
<tr>
<td>wAdV-D29</td>
<td>32</td>
<td>S12</td>
<td>S12</td>
</tr>
<tr>
<td>wAdV-D33</td>
<td>&lt;8</td>
<td>&lt;8</td>
<td>&lt;8</td>
</tr>
<tr>
<td>wAdV-D43</td>
<td>&lt;8</td>
<td>&lt;8</td>
<td>8</td>
</tr>
<tr>
<td>wAdV-D44</td>
<td>8</td>
<td>&lt;8</td>
<td>&lt;8</td>
</tr>
<tr>
<td>wAdV-D45</td>
<td>&lt;8</td>
<td>&lt;8</td>
<td>&lt;8</td>
</tr>
<tr>
<td>wAdV-D46</td>
<td>8</td>
<td>&lt;8</td>
<td>&lt;8</td>
</tr>
</tbody>
</table>

specific antiserum were inactivated at 30°C for 30 min and serially diluted twofold, 50 μl per well with four replicate wells per dilution. A working dilution of virus (HAdV-DS8) containing 100 TCID50 in 50 μl was added to each well, and the plates were incubated at 37°C in 5% CO₂ for 1 h. During the incubation period, Hep-2 cells were trypsinized and resuspended at 5x10⁶ cells per ml. After the incubation, 100 μl of cell suspension was added to each well. The centers of each well were mixed, and the plates were incubated at 37°C in 5% CO₂ for 6 days. After 6 days, the medium was removed and cells were stained with crystal violet solution (1.46 g crystal violet, 50 ml ethanol, 300 ml formamide, 650 ml distilled water). The neutralization titer was calculated as the maximum dilution of antisem that completely inhibited viral growth as evidenced by the lack of cytopathic effect.

Nomenclature

This virus was named HAdV-DS8 because the number 57 was already taken in GenBank (HQ003817). For more adenovirus nomenclature, see http://hadrwgimhu.edu.

Nucleotide sequence accession numbers

The HAdV-DS8 genome and annotation have been deposited in GenBank prior to manuscript submission: accession number HG003817. The following HAdV genomes (GenBank accession numbers) were used for comparative analysis: HAdV-D18 (AB45878), HAdV-D9 (AJ54466), HAdV-D9C (EF121005), HAdV-D22 (EF40671), HAdV-D36 (EF395747), HAdV-D38 (EF394884), HAdV-D36 (EF394884), HAdV-D37 (EF39599), HAdV-D46 (AY753568), HAdV-D48 (EF135737), HAdV-D9 (DQ699389), HAdV-D33 (DQ699389), HAdV-D34 (DQ699389), and HAdV-D36 (DQ699389).

Amplification of the HAdV-DS8 genome

To amplify regions of HAdV-DS8 flanking the sequences previously described by Ferreire et al. [12], we designed primers based on conserved adenovirus sequences in species HAdV-D. All amplicons were then sequenced using primer walking. The genome was assembled using SeqMan, which is an assembly program inside of the Lasergene software suite.

Figure 8. Phylogenetic analysis of the fiber coding sequence in HAdV-DS8. Analysis of HAdV-DS8 is based on the nucleic acid sequence of the fiber knob. Phylogenetic trees were constructed from aligned sequences using MEGA, via the neighbor-joining methods and a bootstrap test of phylogeny. Bootstrap values shown at the branching points indicate the percentage of 1000 replications produced the clad.

doi:10.1371/journal.pone.0024491.g008
Figure 9. Bootscan analysis of selected fiber genes in species HAdV-D. (A) HAdV-D47, (B) D26, (C) D20, and (D) D30. This figure is a corrected repeat of Figure 2 in Darr et al [28].

doi:10.1371/journal.pone.0051491.g009

Nucleic Acid Isolation
HAdV-D38 particles were separated from Hep-2 cells by ultracentrifugation. Genomic DNA was acquired from viral particles using AxyPrep Genomic DNA Extraction Kit (BioCline Corporation). Finally, the viral DNA was resuspended in deionized water and stored at ~20°C until use.

Bioinformatics
The available genomes from species HAdV-D were aligned using the clustalW [23] alignment method which is available through a web interface at http://www.ebi.ac.uk/Tools/clustalw2/index.html. The default parameters for gap open penalty and gap extension penalty were used. Hoxon coding sequences used for analysis were HAdV-D8 (AB443677), HAdV-D9 (AB545996), HAdV-D10 (AB859360), HAdV-D15 (DQ496161), HAdV-D16 (AB500691), HAdV-D19 (AB428774), HAdV-D20 (AB50101.1), HAdV-D22 (FJ199077), HAdV-D24 (AB50105.1), HAdV-D25 (AB50106.1), HAdV-D26 (EF135374), HAdV-D27 (AB501081.1), HAdV-D28 (FJ823056), HAdV-D29 (AB501087), HAdV-D30 (AB501111.1), HAdV-D32 (AB501115.1).

Table 2. Comparison of the nucleotide sequences used by Darr et al to show recombination events in the fiber knob junction.

<table>
<thead>
<tr>
<th>Sequences compared</th>
<th>Nucleotide identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>HAdV-D20 vs. HAdV-D22</td>
<td>100%</td>
</tr>
<tr>
<td>HAdV-D20 vs. HAdV-D23</td>
<td>100%</td>
</tr>
<tr>
<td>HAdV-D20 vs. HAdV-D24</td>
<td>100%</td>
</tr>
<tr>
<td>HAdV-D20 vs. HAdV-D25</td>
<td>100%</td>
</tr>
<tr>
<td>HAdV-D20 vs. HAdV-D26</td>
<td>100%</td>
</tr>
<tr>
<td>HAdV-D20 vs. HAdV-D27</td>
<td>100%</td>
</tr>
<tr>
<td>HAdV-D20 vs. HAdV-D28</td>
<td>100%</td>
</tr>
<tr>
<td>HAdV-D20 vs. HAdV-D29</td>
<td>100%</td>
</tr>
<tr>
<td>HAdV-D20 vs. HAdV-D30</td>
<td>100%</td>
</tr>
<tr>
<td>HAdV-D20 vs. HAdV-D31</td>
<td>100%</td>
</tr>
<tr>
<td>HAdV-D20 vs. HAdV-D32</td>
<td>100%</td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pone.0051491.g009
Phylogenetic analysis of HAdV-D58

Nucleotide alignment of whole genome, penton, I1 and I2 of hecmon, hexon, fiber and fiber knob, ORF2 and ORF3 regions were performed using the multiple alignment software based on a "Fast Fourier Transform" algorithm (MAFFT, http://www.ebi.ac.uk/Tools/mafft/index.html). Phylogenetic distances and trees were generated from these aligned sequences by the Molecular Evolutionary Genetic Analysis software (MEGA v.6.1; http://www.megasoftware.net/). Distances were obtained with pairwise distance calculation using maximum composite likelihood model. Subsequent phylogenetic trees were obtained using bootstrap sets of phylogeny of 1000 replicates with neighbor-joining method featured in the program.

Supporting Information

Table S1 Percent identities of the nucleotide coding sequences of loop1 (L1) and loop2 (L2) HAdV-D58 coding regions to homologous sequences from other viruses in species HAdV-D.

(PDF)

Table S2 Percent identities of the nucleotide coding sequences of selected E3 HAdV-D58 coding sequences and their homologs.

(PDF)

Acknowledgments

We thank Carl Gibbons for help in sequencing this virus. The view expressed in this material are those of the authors, and do not reflect the official policy or position of the U.S. Government, the Departments of Defense, or the Department of the Air Force.

Author Contributions

Gastered and designed the experiments: DS MJ LF. Performed the experiments: NKR EBL SVJ JF. Analyzed the data: SLF MJ DS EBL JC DW. Contributed reagents/materials/analysis tools: MSJ DSJC DW. Wrote the paper: MSJ DSJC DWL.

References

CHAPTER 5 – HUMAN ADENOVIRUS TYPE 59

Introduction

Human adenovirus type 59 (HAdV-D59) is a novel human respiratory pathogen that was isolated from an AIDS patient’s bronchoalveolar lavage (BAL) biopsy sample. The patient presented with a fever, cough, tachycardia and expiratory wheezes. Using bioinformatics and genomic analysis, the genome sequence was examined for insight into its molecular origins. The results suggest that HAdV-D59 is an emergent pathogen with a molecular evolution pathway that includes multiple recombination events in the penton base, hexon and fiber genes. These were apparently transferred from parental genomes HAdV-D19C, HAdV-D25 and HAdV-D56 respectively. Furthermore, serological analysis shows a neutralization profile for the hexon epitope that is similar to but not identical to HAdV-D25, suggesting a divergent but common ancestor.
Computational and Serologic Analysis of Novel and Known Viruses in Species Human Adenovirus D in Which Serology and Genomics Do Not Correlate

Elizabeth B. Liu9*, Debra A. Wadford9, Jason Seto1, Maria Vu1, Nolan Ryan Hudson1, Lisa Thrasher1, Sarah Torres2, David W. Dyer1, James Chodosh1, Donald Seto1, Morris S. Jones1

1 School of Systems Biology, George Mason University, Manassas, Virginia, United States of America, 2 Weill and Rockefeller Disease Laboratory, California Department of Public Health, Richmond, California, United States of America, 3 Clinical Investigation Facility, David Grant USAF Medical Center, Travis AFB, Fairfield, California, United States of America, 4 Department of Microbiology and Immunology, University of Oklahoma Health Sciences Center, Oklahoma City, Oklahoma, United States of America, 5 Harvard Laboratory, Massachusetts Eye and Ear Infirmary, Department of Ophthalmology, Harvard Medical School, Boston, Massachusetts, United States of America

Abstract

In November of 2007 a human adenovirus HA-DV was isolated from a bronchoalveolar lavage (BAL) sample recovered from a biopsy of an AIDS patient who presented with fever, cough, tachycardia, and respiratory wheezes. To better understand the isolated virus, the genome was sequenced and analyzed using bioinformatic and phylogenetic analysis. The results suggest that this novel virus, which is provisionally named HA-DV059, may have been created from multiple recombination events. Specifically, the penton, hexon, and fiber genes have high nucleotide identity to HA-DV109, HA-DV295, and HA-DV56, respectively. Serological results demonstrated that HA-DV059 has a neutralization profile that is similar yet not identical to that of HA-DV025. Furthermore, we observed a two-fold difference between the ability of HA-DV109 and HA-DV295 to be neutralized by reciprocal antisera indicating that the two hexon proteins may be more similar in epitopic conformation than previously assumed. In contrast, hexon loops 1 and 2 of HA-DV109 and HA-DV295 share 79.13 and 92.56 percent nucleotide identity, respectively. These data suggest that serology and genomics do not always correlate.

Materials and Methods

Ethics Statement

The work reported herein was performed under United States Air Force Surgeon General approved clinical investigation No. FFG20040063H, by the Institutional Review Board at the David Grant USAF Medical Center. Informed Consent was not required, because we did not use clinical samples.

In this report, an adenovirus isolated from a bronchoalveolar lavage (BAL) sample that was biopsied from an AIDS patient who presented with fever, cough, tachycardia and respiratory wheezes is examined using genomics and bioinformatics. Based on the whole genome analysis and supported by limited serological data, this adenovirus belongs to species HA-DV-D, and is a 'novel' virus, to be given the name of HA-DV059.
and sera raised against reference stock adenoviruses types 1–49 from the collection maintained by the Viral and Rickettsial Disease Laboratory of the California Department of Public Health, Richmond, CA. Reference viruses were originally obtained from the reporting investigators: the Reference Reagents Branch, National Institutes of Health; the Respiratory Viral Disease Unit, Centers for Disease Control and Prevention; or the American Type Culture Collection. Stock virus cultures were passaged in A549 cells (American Type Culture Collection, Rockville, MD), the cells were disrupted by vortexing, and cell-free supernatant fluid was then frozen at −70°C.

Equal volumes of diluted virus and immune serum were mixed and incubated for one hour in 5% CO2 at 37°C. Thereafter A549 cells were added, mixed, and incubated at 37°C in 5% CO2 for 7 days. Each assay contained a back titration of the virus used. Living cells were distinguished from dead cells by measuring the amount of Pitter’s Neutral Red [15] present as indicated by absorbance at 530 nm using a microplate spectrophotometer (Bio-Tek Instruments, Winooski, VT). Virus neutralization titers were determined by equaling cell death to virus growth (no virus neutralization). Neutralization was plotted as a percentage of cell control absorbance, to determine endpoint virus and serum titers. Three independent experiments were run yielding similar results.

Nucleotide sequence accession numbers

The HAdV-D59 genome sequence and its annotation are deposited in GenBank and retrievable as accession number JF739911. In addition, the following HAdV genomes (GenBank accession numbers) were used for comparative computational analyses: HAdV-D1 (AB442747), HAdV-D9 (ABJ54968), HAdV-D15 (AB23596), HAdV-D17 (AF108165), HAdV-D19C (EF121063), HAdV-D22 (FJ992771), HAdV-D25 (unpublished), HAdV-D26 (EF152876), HAdV-D28 (FJ924826), HAdV-D29 (GG000000), HAdV-D37 (DQ980090), HAdV-D46 (AY175648), HAdV-D48 (EF13573), HAdV-D49 (DQ893201), HAdV-D59 (AB330801), HAdV-D66 (AF770721), and HAdV-D88 (HQ838276). Fiber genes from species HAdV-D genomes were aligned using ClustalW [16]. For this analysis, the default gap opening and gap extension penalties were 15.0 and 6.66.

Amplification and DNA sequencing of the HAdV-D59 genome

To amplify regions of HAdV-D59 using the polymerase chain reaction (PCR) protocol, conserved adenovirus sequences in species HAdV-D were used to design primers. All amplicons were then sequenced on an ABI 3130xl using a primer walking strategy. The HAdV-D59 genome was sequenced to 8-fold coverage following PCR amplification, with both strands represented.

Bioinformatics

The HAdV-D59 genome was compared against a select number of viral genomes from the HAdV-D group based on its GC content, which is indicative of HAdV species. The selection of which genomes was based on initial overall high nucleotide identity to HAdV-D59. The data presented are final iterations of analyses that initially included all of the sequenced genomes in species HAdV-D.

Recombination analysis

Whole genome sequences of HAdV-D59 and number of species HAdV-D were first aligned with kalign [http://www.ebi.ac.uk/Tools/msa/kalign/] for a broad perspective of the genome. SimPlot [17] was then used to construct a Bandeau analysis of the aligned sequences. The window size and step size were set to 100 and 200 respectively.

Following this, to provide a detailed close inspection of recombination events, the penton base gene, hexon gene, E5 coding region and fiber gene from the HAdV-D59 genomes were aligned to their counterparts using ClustalW [16]. This was also followed by recombination analysis using SimPlot with the window size and step size set to 250 and 50, respectively.

Figure 1. Genome organization of HAdV-D59. The HAdV-D59 genome is represented by a black horizontal line marked at 54bp intervals. Protein encoding regions are shown as arrows indicating transcriptional orientation above and below the genome. Spliced genes are indicated by V-shaped lines.
Figure 2. Comparative genomic analysis. (A) Pairwise nucleotide comparison of selected HADV-D genomes to HADV-D59 using PICTture. The arrows above the x-axis demarcate the positions of penton base, hexon, L1 region, and fiber coding sequences in the genome of HADV-D59. The y-axis notes the percent identity. HADV-D9, HADV-D22, HADV-D19C, HADV-D25, HADV-D28, HADV-D36, and HADV-D56 were used for comparison to HADV-D59 because they share high nucleotide identity to the aforementioned virus in different sections of their genome. (B) Whole genome phylogenetic analysis. The phylogenetic tree was constructed from aligned sequences using MEGA via the neighbor-joining methods and a bootstrap test of phylogeny. Bootstrap values shown at the branching points indicate the percentages of 1000 replications produced the clad. A bootstrap value of 70 and above is considered to be robust. doi:10.1371/journal.pone.0033217.g002

Percent Identity
Whole genome, penton base gene, hexon gene, E3 coding region and fiber gene nucleotide sequences of HADV-D56, along with members of species HADV-D9, were aligned using kalign; these were then compared to each other based on percent identity values calculated with Chimera [18].

Phylogenetic analysis of HADV-D59
Sequence alignments for phylogenetic analysis were generated using the kalign method noted earlier. Phylogenetic trees were constructed from these aligned sequences using Molecular Genetic Analysis Software (MEGA 4.1; http://www.megasoftware.net), via neighbor-joining methods and bootstrap test of phylogeny with replicates set to 1000.

Results
Clinical investigation
In November 2007, an AIDS patient was admitted to San Francisco General Hospital, presenting with fever. The patient also complained of a cough productive of yellow sputum and blood. Clinical examination revealed a body temperature of 101°F, tachycardia and expiratory wheezes. During the hospital stay, a CT scan displayed results suggestive of a cavitary lung lesion. This prompted a bronchoalveolar lavage (BAL) for a diagnostic specimen (via bronchoscopy). A virus was cultured from the BAL sample and sent to the California Department of Public Health (CDPH) for further analysis and identification. No other pathogens were isolated from this patient.

The virus was propagated at the Viral and Rickettsial Disease Laboratory at the California Department of Public Health, and was identified as an unknown adenovirus by serum neutralization assay. Initial sequence analysis of amplicons derived from the hexon and the fiber genes revealed similarity to gene sequences from HADV-D25 and HADV-D56, respectively. The possibility that this virus might represent a novel, recombinant pathogen prompted whole genome analysis in order to characterize this isolate more thoroughly.

Amplification, sequencing, and genetic characteristics of the novel adenovirus
To elucidate the genetic characteristics of this pathogen (HADV-D59), we sequenced and analyzed the entire genome. The genome length of HADV-D59 is 35,072 base pairs (Figure 1), with a base composition of 22.4% A, 20.4% T, 28.5% G, 28.7% C. The GC content of 57.2% is consistent with the values found for members of species HADV-D (mean of 57%). The organization of the 36 open reading frames (ORFs) that were annotated had a genome organization similar to other mastadenoviruses (Fig 1). The inverted terminal repeat (IR) sequences for HADV-D59 were determined to be 151 bp in length. Within species HADV-D, HADV-D59 has a genome percent identity ranging from a low of 92.06% (HADV-D8) to 96.50% (HADV-D9).

Genome Analysis
Since initial DNA sequencing suggested evidence of recombination, we analyzed the HADV-D59 genome using PICTture, a dynamic blast alignment visualization program designed for comparative analysis (http://pictrure.ddcords.org/). Comparisons of the HADV-D59 genome with the whole genome sequences of HADV-D9, -D22, -D19C, -D25, -D28, -D36, -D56 and -D58 were performed. Consistent with other previously reported viruses in species HADV-D [4,18], HADV-D59 shows heterogeneity in the penton base, hexon, E3 and fiber coding sequences (Fig S1). Pairwise alignment suggested that these regions had highest nucleotide identities with sequences from HADV-D9, HADV-D25, HADV-D28, and HADV-D56 (Table 1). Comprehensive phylogenetic analysis of whole genome HADV-D59 were performed. Using sequences available in GenBank as well as the unpublished sequence of HADV-D25, the whole genome phylogenetic tree analysis resulted in a clade that includes HADV-D59, HADV-D9, and HADV-D36 with a high confidence bootstrap value of 97 (Fig 2B).

Penton Base Gene Analysis
Recently it was shown that two coding sequences for the external hypervariable loops in the penton base gene contain hotspots for recombination in species HADV-D [20]. Analysis of the primary amino acid sequences in species HADV-D showed that the most similar loop sequences (nucleotides 200-600) to that of HADV-D59 were HADV-D28 and HADV-D36 (80.3%). In addition, the most similar RGD loop (nucleotides 650-1120) to HADV-D59 was HADV-D28 with 100% amino acid identity. Bootscan analysis [17] with penton base sequences from species HADV-D confirmed the aforementioned relationships. Phylogenetic analysis of the HADV-D59 penton base hypervariable loop 1 also confirmed that it is a close relative of HADV-D28 and HADV-D36 with a robust bootstrap value of 91 (Fig 2B). Phylogenetic analysis also demonstrated that the HADV-D59

Figure 3. Computational analysis of the penton base gene. (A) Bootscan analysis of the HAV-D39 penton base gene with fully sequenced penton gene sequences of species HAV-D [8]. This has also been demonstrated in other species of HAVs [1,5].

To determine whether or not the HAV-D59 hexon coding region was either novel or the result of a recombination event, SimPlot analysis was performed. Similar to Mafidis et al. [12], the RD3 loop was conserved with the inclusion that HAADV-D59 and HAADV-D22 with a bootstrap value of 92 (Fig. 3C).

Hexon Gene Analysis

A previous study showed that recombination in HAVs can occur within the hexon gene of viruses in species HAV-D [8].

The nucleotide percent identity to other hexon coding sequences in species HAV-D was determined (Table 1). HAV-D59 loop sequences 1, 2, 4, 6, and L2 were 91.8%, 85.3%, 91.9%, and 80.4% identity, respectively, to those of L1 and L2 of HAADV-D25, respectively (Table 1). Phylogenetic analysis demonstrated that L1 of HAADV-D59 and L2 of HAADV-D25 are distant related to the following HAV-D pairs which were shown to be distinct via serology [21]: HAADV-D9 and HAADV-D35; HAADV-D79 and HAADV-D11; HAADV-D15 and HAADV-D30; HAADV-D92 and HAADV-D30; HAADV-D9 and HAADV-D32, as well as HAADV-D5 and HAADV-D25 (Fig. 4B). Furthermore, the same phenomenon was observed for L5 with the exception of HAADV-D45/D26 pair (Fig. 4C).

The difference in nucleotide identity between HAADV-D59 and HAADV-D25 (the nearest phylogenetic relative to HAADV-D59) in the L1 and L2 domains is greater than 2.5% (3.2% and 4.8% identity, respectively). Mafidis et al. stated that percent nucleotide identity differences greater than 3.4% and 7.5% in L1 and L2, respectively, strongly suggests the identification of a novel HAV-D [22]. Therefore, the percent of nucleotide identity differences in L1 and L2 of HAADV-D9 further suggest that the aforementioned virus is novel.

E3 Genome Region Analysis

SimPlot-based tools suggest that a large portion of the E3 transcription region (gene encoding for the CR1) and CR1-like region (RD2) in HAADV-D59 may have originated from a recombination event between either HAADV-D6 or HAADV-D9 and another yet to be described HAADV (Fig. 5). Interestingly, the SimPlot results suggest that a recombination event took place within the open reading frame (ORF) of the CR1 gene (Fig. 5A). We also examined whether E3 genes in species HAV-D and did not detect common recombination loci (data not shown).

Phylogenetic analyses of two genes (CR1/CR1 and CR1 genes) in the E3 region demonstrate that the coding regions for CR1/CR1 and CR1/CR1 in HAADV-D59 are closely related to those of HAADV-D9 and HAADV-D35 (Fig. 5B, 5C).

Fiber Gene Analysis

SimPlot analysis of the HAADV-D59 fiber was performed on fiber sequences extracted from HAADV. The results suggested that the fiber gene of HAADV-D59 is nearly identical with sequences from both HAADV-D6 and HAADV-D9 (Fig. 5A). Furthermore, the fiber of HAADV-D59 had 99.39% and 99.81% nucleotide identity with HAADV-D9 and HAADV-D35 (Table 1), respectively. Furthermore, phylogenetic analysis of the fiber genesthe species HAADV-D9 confirms that the fiber of HAADV-D59 was closest in sequence to the corresponding sequences in HAADV-D35 and HAADV-D9 (Fig. 5B).

Serum Neutralization

HAADV-D59 was neutralized by both HAADV-D25 and HAADV-D15 antiserum, yet not by HAADV-D9 antiserum (Table 2). Interestingly, HAADV-D25 antiserum showed at least a two-fold higher neutralization titer to HAADV-D59 (greater than 1:1000) than did anti cognate antigen HAADV-D25 (1:1000) (Table 2). These results were not consistent with recent recombination events in the hexon of HAADV-D39 (Fig. 4A).

Discussion

Our results demonstrated that HAADV-D59 antiserum was more effective at neutralizing HAADV-D59 than HAADV-D25 (Table 2). Since L1 and L2 promulge from the surface of HAADV-D25 [23], it is not surprising that there is a difference in the ability of HAADV-D25 antiserum to neutralize the different viruses. One possibility for the differences in neutralization may be that the few differences in the primary amino acid structure present the HAADV-D59 hexon three-dimensional structure in such a way that the neutralizing epitopes are enhanced, thus making the virus easier to neutralize. Interestingly, we also observed a two-fold difference between the ability of HAADV-D15 and HAADV-D25 to be neutralized by reciprocal antiserum. This contradicts one study that showed antiserum to HAADV-D15 and HAADV-D59 did not cross-react in reciprocal neutralization experiments [21]. However, our data are consistent with the original characterization of HAADV-D15 and HAADV-D25 (previously called BP-1) [24], thus we conclude that HAADV-D15 and HAADV-D25 are separate serotypes according to the traditional methods used for differentiating serotypes. These results also demonstrate that using neutralization as a criterion to type novel adenoviruses is complicated by non-standard serology methods and serotypes that may vary interlaboratory variability of neutralization results. In contrast, using genomics as a method for typing HAVs is consistent regardless of which laboratory generates the results.

Even though HAADV-D15 and HAADV-D59 showed only a two-fold difference in serum neutralization, they were recognized as different serotypes by Roen et al., because they had different fiber protein [21]. HAADV-D15 and HAADV-D25 share 79.3% and 92.56 percent nucleotide identity in L1 and L2, respectively. Thus, bioinformatic analysis demonstrates that they are actually different types using the criteria established by Mafidis et al. which states that the nucleotide identity of L2 must differ by greater than 2.5 percent to type a novel virus [22]. Furthermore, pairwise nucleotide comparison of the hexon coding sequences for HAADV-D15 and HAADV-D25 show homologically distinct they are (Fig. 6). Neutralization assays measure the overall effect of various antibodies that bind to multiple epitopes and may yield...
Figure 4. Computational analysis of the hexon gene. (A) SimPlot analysis of the HAdV-D25 hexon coding sequence. L1 and L2 correspond to loops 1 and 2 of the hexon gene, which comprise the determinant of virus serum neutralization. (B) Phylogenetic analysis of the hexon L1 sequences in species HAdV-D. (C) Phylogenetic analysis of the hexon L2 sequences in species HAdV-D. A phylogenetic tree was generated along with...
Computational and Serologic Analysis of Adenovirus

representatives from the other species. Phylogenetic trees were generated from aligned sequences using MEGA, via the neighbor-joining method and a bootstrap test of phylogeny.
doi:10.1371/journal.pone.0033212.g004

Figure 5. Computational analysis of the E3 and fiber regions. (A) SimPlt analysis of the E3 and fiber region of HAdV-059 compared to fully sequenced E3 and fiber regions from species HAdV-D. The arrows over the bootcans demarcate the approximate positions of the E3 coding sequences. Phylogenetic analysis of (B) HAdV-059 CR1β (C) HAdV-059 CR1γ, and (D) fiber loci. The phylogenetic trees were generated from aligned sequences using MEGA, via the neighbor-joining method and a bootstrap test of phylogeny.
doi:10.1371/journal.pone.0033212.g005
variable interlaboratory result due to non-standard methods and reagents. Genomic analyses measure genetic differences in the genome that have the potential to affect the pathogenicity of the virus and can be independently verified by most laboratories. The contrasting serology and genomics results for HAdV-D15 and HAdV-D25 demonstrate that these two methods do not always yield concordant results.

SimPlot analysis of the HAdV-D59 L1 and L2 regions demonstrates high nucleotide identity between the hexons of HAdV-D25 and HAdV-D59 and suggests that they may be derived from a yet undiscovered common ancestor. If L1 and L2 of HAdV-D59 are from a common ancestor, the recombination event may be ancient, as evidenced by 3.52 and 4.11 percent nucleotide differences in the L1 and L2 sequences, respectively. If the recombination events were recent, SimPlot analysis would illustrate nearly 100% nucleotide identity, which was shown for HAdV-D33 and HAdV-D35 [4,5,19]. With a distant past recombination event the hexons from HAdV-D33 and HAdV-D35 would have mutated over many replication cycles, after the initial recombination, to result in the variation we detected.

Multiple studies have shown that HAdVs in species HAdV-D recombine with one another in the penton base and hexon genes [4,20]. In this paper, we demonstrate that recombination may have occurred in the E3 region of HAdV-D59; however, after examining all of the sequenced E3 genes in species HAdV-D, we found that there was not a predictable pattern of recombination (data not shown). Viruses in species HAdV-D show variability in their cell tropism ranging from growth in ocular tissues to gastrointestinal and/or respiratory tissues [4,25,26]. Given that the fiber knobs is an important determinant of cell tropism, it may be concluded that recombination is an important molecular evolutionary pathway for the diversity observed within species HAdV-D.

The section of the HAdV-D39, -D56, and -D9 genomes that encodes for GRI, GRI, RBD2, RBD1, 14.7K, and fiber show high nucleotide identity (Fig. 5a). From the nucleotide data, it is impossible to tell whether or not the 3' end of HAdV-D39 came from HAdV-D36 or HAdV-D9. Although HAdV-D9 was discovered in 1987 [25], there has been no disease associated with this virus. In contrast, prior serological evidence suggests that HAdV-D36, an ocular and respiratory pathogen (with hexon and fiber coding sequences similar to HAdV-D15 and HAdV-D9, respectively) [12,26] has been implicated in human disease as early as 1960 [27] and at other points in time as well [27,28,29,30,31,32]. HAdV-D59 may have also existed prior to our current description yet have gone undetected during the same time periods. Thus, it is impossible to say with absolute certainty that all of the aforementioned viruses existed first and/or whether they evolved from a common ancestor. Future genomic analysis of known and unknown adenoviruses is needed to elucidate further the evolutionary history of HAdVs.

Acknowledgments

The views expressed in this material are those of the authors, and do not reflect the official policy or position of the U.S. Government, the Department of Defense, or the Department of the Air Force. We thank Drs. Walsh and Zaidle-Delgoff for advice, critical discussion and preliminary analysis.

Author Contributions

Conceived and designed the experiments: MSJ DS DAW. Performed the experiments: EBL JS MV NRH LT ST. Analyzed the data: EBL DAW JS DWD JC DS MSJ. Contributed reagents/materials/analysis tools: DWD JC DS MSJ. Wrote the paper: EBL DAW DWD JC DS MSJ.

References


CHAPTER 6 – BIOINFORMATICS TOOLS DEVELOPMENT TO ENHANCE VIRAL GENOME ANALYSIS

Introduction

Currently there are limited numbers of efficient and accurate computational software tools for the bioinformatics analysis of small genomes such as viruses, including human adenoviruses specifically. Data presentation is an important tool necessary to convey complex genomic information. Many of the current genome data presentation methods include manually-drawn figures that are created based on an estimation of genome location. To provide better visualization of the data, two tools were developed during the course of adenoviral genome investigations described in the preceding chapters. The two tools are as follows: 1) DrawBar (http://binf.gmu.edu/eliu1/drawbar/) and 2) GeneMap (http://binf.gmu.edu/eliu1/genemap/). DrawBar is a web application that generates a visual representation to portray accurately gene locations. The result is a valuable addition to the recombination analysis and presentation of SimPlot and Bootscan data. GeneMap was developed to display accurately and efficiently the locations of any proteins sequences encoded in a genome. This enables “landmarks” to be placed along the nucleotide sequence as references, for example to mark recombination event.

Finally, as a way to convey “state-of-the-art” HAdV information to the research community and the public, a resource website was created to house human adenovirus
data and HAdV typing and identification information provided by the Human Adenovirus Working Group and international human adenovirus research community (hadvwg.gmu.edu). This resource was suggested by a virology taxonomy expert/representative from the National Library of Medicine (National institutes of Health). It is an effort to provide standard for typing and identification, and to coordinate the process of naming novel human adenoviruses prior to GeneBank deposit. Furthermore, its goal is to reduce confusion and conflicts and competing deposits of GeneBank data.

Discussion

**DrawBar**

In scientific research papers and publications, figures are powerful and important tools to present data in a visual manner that allows reader to understand the easier large quantity of complex data. In this sense, the efficient and accurate labeling in a figure is equally important for the in successful presentation of data. This is especially important in genome analysis. Currently, most of the sequence analysis tools such as SimPlot, Bootscan and zPicture, provide figures in which the protein location labeling is done manually. This can be time consuming and often inaccurate, as the placements of protein locations were hand-drawn on the figures based on approximation. As an alternative, DrawBar was developed. It is a web application written in PHP, a general-purpose
scripting language that is suited especially to web development (PHP, 2015). DrawBar provides accurate placements of protein location in a pixel to base pair ratio.

DrawBar requires a few specific parameters in order to generate an accurate labeling for genome sequence data figures. There are the pixel length of the graph that requires labeling, base pair size of the displayed genome or protein and the specific base pair position of each target gene’s labeling region. With these, the DrawBar program then calculates the percentage of each gene position to the total number of base pairs that the figure displayed, then translates this calculated percentage to pixel positions in relations to the total pixel length of the figure. In the example provided in Figure 17, protein labeling is required for a whole genome SimPlot graph that has a width of 600 pixels. The target genome has a size of 35000 base pairs, and the specific locations of E1A, E1B, polymerase, penton base, hexon and fiber proteins are required for reference. The required format for DrawBar is that the gene name has to be a single string without any spacing. A single blank space separates the gene name and the gene location. The gene location’s start and end positions must be a single string numeric that is separated by a hyphen, and each gene name and location must start on a new line in the input box.
**Figure 17 -** DrawBar input screen. DrawBar ([http://binf.gmu.edu/eliul/drawbar/](http://binf.gmu.edu/eliul/drawbar/)) input page. Parameters include pixel length of the graph that requires labeling, base pair size of the displayed genome or protein and the specific base pair position of each target gene’s labeling region.
The Results page provides a legend bar that can be placed on top of the figure, with accurately calculated gene locations. Calculated gene locations, in pixel units, are displayed on the result page for debugging and accuracy check purposes (Figure 18). The user can right click on the resulting legend bar and save the image, which then can be pasted on to the SimPlot figure. The resulting legend bar serves as an accurate placement reference, and the user can then add any additional styling to the legend bar if needed.

Figure 18 - DrawBar results screen. Resulting legend bar image contains accurately calculated gene locations and serves as an placement reference.
There are five php files required for DrawBar in order to run on a server, index.php; index.display.php; header.php; footer.php; image.php and map.php. First, index.php file contains code that initializes the application and handles the order of operation. The Index.display.php file contains code that handles the display the input screen content, and initializes the input variables for later processing. Header.php and footer.php contains code that displays content in the header and footer of the webpage; in this case, application title and footer note information. Image.php and map.php handle the actual calculation of the base pair to the pixel ratio along with the actual drawing of the image that is displayed in the result page. All of the source codes for DrawBar are included in **Appendix 1**.

**GeneMap**

GeneMap is a web application developed using PHP (www.php.org). Its structure and concept is very similar to that of the DrawBar tool, providing a map that presents the locations of all of the important proteins encoded in a genome. This is usually drawn by hand or through graphic software, again by estimation. The location of each protein is based on approximation its sequence location in relations to other proteins. This process can be tedious and extremely inaccurate since the location of the proteins can be overlapping, and also may include reverse or complement strands and spliced genes. GeneMap was developed to automate this process and to present an accurate portrayal of each proteins location. It is used for data presentation in publication, and also can serve as an accurate protein placement reference for user to draw on if additional styling of the
image is needed. This tool has been used in publications to enhance the visual display of adenovirus sequence data and to support novel findings (Seto et al., 2010)

GeneMap has three input fields that take specific required parameters in order to generate an accurate protein placement map in a genome. The first required field is a map name that will be displayed on top of the resulting map. A second require field is the base pair length of the target genome. The last input box is for specific base pair position of each target gene’s mapping region with special syntax that can be used to indicate if reverse complement strand or splice gene is present. This GeneMap program then calculates the percentage of each gene position to the total number of base pair the figure displayed, translating this calculated percentage to pixel position in relations to the total pixel length of the figure, then accurately plot the protein locations on the final map image. This is very similar to the DrawBar design, but GeneMaps requires additional features that handle the placements of each protein on the map when there is a potential of overlapping. GeneMap calculates if there is another protein within the vicinity of the current protein; if so, it draws the new protein either above or below the existing protein to avoid overlapping. Also the placement and presentation of reverse or complement strands and splice genes are different from the regular forward strand proteins. In the example provided in Figure 19, a genome mapping figure is required for HAdV-B16. The target genome has a size of 35522 base pairs, with specific locations for all of the annotated proteins noted. The required format for GeneMap is that the gene name has to be a single string without any spacing. A single blank space separates the gene name and the gene location. The gene location start and end positions must be a single string
numeric that is separated by a hyphen. Each gene name and location must start on a new line in the input box. In addition, “, c” is added in the end of the protein location to indicate if a strand is reverse or complement. And in the case of a spliced gene, the two coding or exon locations are separated by a comma.

Figure 19 - GeneMap input screen.
GeneMap (http://binf.gmu.edu/eliu1/genemap/) input page. Parameters include title, pixel length of the map, base pair size of the displayed genome and the specific base pair position of each target gene’s labeling region.
The Results page provides final image of map, with accurately calculated and displayed gene locations. The calculated gene locations, in pixel units, are displayed on the Results page for debugging and accuracy check purposes (Figure 20).

Figure 20 - GeneMap result screen 1.
The calculated gene locations, in pixel units, are displayed on the Results page for debugging and accuracy check purposes.
The final image indicate all forward strands on the top half of the map and the reverse strands on the bottom half. Spliced genes have a green connector indicator (Figure 21). The user can then “right click” on the resulting image to save for later use. The resulting gene map can also serve as an accurate placement reference, and the user can then draw on any additional styling to the legend bar if needed.

Figure 21 - GeneMap result screen 2.
Result image indicate all forward strands on the top half of the map and the reverse strands on the bottom half.
There are five php files required for GeneMap to run on a server: index.php; index.display.php; header.php; footer.php; image.php and map.php. These files provide the following functions: the index.php file is used to initialize the application and handles the order of operation; the index.display.php file processes the input screen content display, and initializes the input variables for later processing; the Header.php and footer.php files include code that displays the content in the header and footer section of the webpage; and the remaining two files, similar to DrawBar, image.php and map.php files handle the actual calculation of the base pair to pixel ratio and creates the drawing of the image that is displayed in the result page. In GeneMap, these two files also have additional functions that handle the calculation and differentiate the placements for the spliced genes, forward and reverse or complement genes. Those functions also provide placement calculation to avoid overlapping genes on the final image. All of the source codes for GeneMap are included in Appendix 2.

**Human Adenovirus Working Group Reference Website**

The Human Adenovirus Working Group has been formed as collaboration between several adenoviral researchers and a viral genome representative of the National Center for Biotechnology Information (NIH) in an effort to coordinate and standardize the process of assigning names to novel HAdVs based on genome data. The website was built with WordPress (WordPress, 2015), and is a basic web content hosing site. It provides background information regarding adenovirus typing, criteria for a new HAdV type and name, instructions regarding how to submit a candidate HAdV and a serotyping
tool. The serotyping tool was constructed using a DataTables JavaScript library based plugin that is freely available in the WordPress plugin library. The resulting tool displays all potential types corresponding to the query serotype entered by a user (Figure 22).

Figure 22 - HAdV Working group website serotyping tool. Serotyping tool in HAdV Working group website (hadvwg.gmu.edu) displays the result table from the search feature.
CHAPTER 7 – FUTURE DIRECTIONS

Future Directions

There are many follow-up questions remaining to be answered following this study of the molecular evolution and characterization of four human adenoviruses pathogens. Since the directionality of the cross species recombination cannot be determined, additional studies may provide insight into these recombination events. This includes additional genome analysis with a larger data set in order to obtain a more complete and clearer picture of human adenovirus evolution, and directionality of horizontal gene transfer.

Standardized typing with genomics and bioinformatics protocols can also facilitate accurate identification of human adenoviruses. This will reduce long-standing confusion and conflicts in the naming and typing of novel and archived human adenoviruses. In addition to standardizing typing of human adenoviruses, standardized annotation of the genome is also necessary.

Currently the process of sequence annotation and isolation is repetitive and labor intensive. Also, genome annotation of human adenovirus can vary among research groups. This can provide inconsistent results in applying various bioinformatics tools such as sequence comparisons and percent identity calculations. To standardize the
annotation protocol for human adenovirus, either a manual protocol that can provide consistent results should be developed or an automated algorithm and tool should be developed. Both would help facilitate the process to provide a systematic standard that allows future analyses and understanding of any human adenovirus genome.

Additional research can be directed to software development for the graphical representation of sequence analysis. Current software such as SimPlot is extremely helpful and sufficient in graphic data representation, but are not without its caveats. As demonstrated in the HAdV-B16 and HAdV-B21 analyses and discussions, when there is more than one genome with high similarity to each other, the Bootscan results became unreadable as the highly similar sequences interfere with each other. Researchers have to eliminate the competing sequences manually in order to provide a clear presentation of the targeted data. It would be much more efficient for the Bootscan software to handle such competitive similarities and to display accurate results without the need to eliminate manually a number of sequences.

Software such as GeneMap can also be improved by having higher resolution and better quality of the resulting map. As PHP has limitations in its graphical outputs, other languages such as Java or Objective C can be considered in order to improve on the final results and to provide better data presentation.

All of the methods and tools mentioned in this dissertation have already generated vast amount of valuable and quality data. As the study of human adenoviruses, and other viruses, progress, these methods will definitely be improved and advanced.
APPENDICES

Appendix 1 DrawBar code

```php
<?
//index.php by Elizabeth Liu for Draw Bar graphical output

require_once("header.php");
require_once("index.display.php");
//extract ($_GET);
$task = $_GET['task'];

switch ($task){
    case process:
        DISPLAY_Index::ProcessData();
        DISPLAY_Index::OutputData();
        break;
    default:
        DISPLAY_Index::DefaultView();
        break;
}
require_once("footer.php");
?>
```
// index.display.php by Elizabeth Liu

class DISPLAY_Index{

    // original index page with 2 forms
    // form one is gene map title
    // form two is input box for the gene names and length
    // two buttons, one clear one submit

    function DefaultView(){

        $barlength=$_REQUEST['barlength'];
        $genomesize=$_REQUEST['genomesize'];
        $data=$_REQUEST['proteinlocations'];
        $lines=split("\n", $data);
        $i=0;

        foreach($lines as $nextLine){
            // split input
            $test=preg_split("/[\n\t\s,-]+", $nextLine);
        }
    }

    function ProcessData(){

    }
}
//remove blank space at end of array
// store gene name in a separate array
if(is_numeric($test[1])==false){
    break;
}
$genename[$i]=$test[0];
//remove genename and leave $test just as gene locations
array_splice($test, 0, 1);

echo $genename[$i];
echo "<br>";

foreach($test as $item){
    echo $item;
    echo "\t";
}
// sort the array
//sort($test);
$genePosition[$i]=$test;
    echo "<br>";
    $i+=1;
}
echo "<br>";
echo $barlength;
echo "<br>";
echo $genomesize;
echo "<br>";
$_SESSION['barlength']=$barlength;
$_SESSION['genomesize']=$genomesize;
$_SESSION['names']=$genename;
$_SESSION['genePosition']=$genePosition;

$lengthratio = 1;
if($genomesize>100&$barlength!=0){
    $lengthratio=round($genomesize/$barlength);
}
echo "length ratio is: ";
echo $lengthratio;
echo "<br>";

}
function OutputData(){
    // echo "output result here"
    echo "<br>";
    echo "<img src=image.php>";
}
}?>
session_start();
header("Content-type: image/png");

//====================================================
// DRAW BACKGROUND IMAGE
//====================================================
//Calculate image size base on bar length. height of
//the image is 1/50 for now
$imgh=0.05*$_SESSION['barlength'];
$imgl=$_SESSION['barlength'];
$img=@imagecreate($imgl,$imgh)
or die("cannot initialize new GD image stream");
$background_color=imagecolorallocate($im, 255,255,255);
$black=imagecolorallocate($im, 0,0,0);

//====================================================
// DRAW BAR
//====================================================
// draw a straight line with 2 red stoppers
$linethickness=0.004*$_SESSION['barlength'];
imagesetthickness($im,$linethickness);
imageline($im,0,round($imgh/1.5),$imgl,round($imgh/1.5),$black);

//====================================================
// Draw Actual Genes
//====================================================
// Draw Actual Genes
//====================================================
//draw actual genes
$pxlratio=round($_SESSION['genomesize']/$_SESSION['barlength']);
$blockh=4*$linethickness;
$blockyvalb=round($imgh/1.5)-$blockh/2;
$blockyvale=round($imgh/1.5)+$blockh/2;
$txth=$blockyvalb-15;
for ($m=0; $m<sizeof($_SESSION['names']); $m++){
    $beg=round($_SESSION['genePosition'][$m][0]/$pxlratio);
    $end=round($_SESSION['genePosition'][$m][1]/$pxlratio);
    imagefilledrectangle($im, $beg, $blockyvalb, $end, $blockyvale, $black);
    imagestring($im, 2, $beg, $txth, $_SESSION['names'][$m], $black);
}
imagejpeg($im);
imagedestroy($im);
Header.php

```php
session_start();
//header.php by Elizabeth Liu

?>
<html>
<head>
<meta http-equiv="Pragma" content="no-cache">
<meta http-equiv="Cache-Control" content="no-cache">
<meta http-equiv="Expires" content="Sat, 01 Dec 2001 00:00:00 GMT">
<title>DrawBar-Elizabeth Liu</title>
</head>
<body>
<center>
<table border=2px width=800px bgcolor="#f2f5fc" cellpadding=20 cellspacing=0>
<tr><td><h1>DrawBar</h1></td></tr>
</td><td colspan=2>
<table width=100%
<tr><td>
To save the above image:
For PC User: Right-click, Save Image / Copy Image.
For Mac User: Click on image and drag to desktop or desired folder.
This is a beta version, if any issue please contact: <a href="mailto:eliu1@gmu.edu?subject=Troubleshooting drawbar">eliu1@gmu.edu</a></td></tr>
</table>
</td></tr></table>
</body>
</html>

Footer.php

```
Appendix 2 GeneMap code

```php
//index.php by Elizabeth Liu for gene map graphical output

require_once("header.php");
require_once("index.display.php");
//extract ($_GET);
$task = $_GET['task'];

switch ($task){
    case process:
        DISPLAY_Index::ProcessData();
        DISPLAY_Index::OutputData();
        break;

    default:
        DISPLAY_Index::DefaultView();
        break;
}
require_once("footer.php");
?>
```
class DISPLAY_Index{

    // original index page with 2 forms
    // form one is genemap title
    // form two is input box for the gene names and length
    // two buttons, one clear one submit

    function DefaultView(){
        ?>
        <form method="post" action="index.php?task=\$=process?">
            Map Name: <br>
            <input type="text" name="mapname"><br><br>
            Genome Length: <br>
            <input type="text" name="maxlength"><br><br>
            Example of Gene and Location entry format: <br>
            Gene1 1-10<br>
            Gene2 20-50, 55-80 (for spliced genes)<br>
            Gene3 100-160, c (c indicates opposite strand coding) <br>
            Gene4 100-130, 145-160, c (complement and opposite strand coding)<br><br>
            Gene names need to be one entity, ie no spaces:<br>
            Gene_1 or Second_Gene_1 <br>
            NOT Gene 1 and NOT Second Gene 1<br>
            <textarea name="geneinfo" rows="20" cols="80"></textarea><br>
            <input type="reset" value="Reset"/>
            <input type="submit" value="Submit"/>
        </form>
        <?
    }

    function ProcessData(){
        $mapname=$_REQUEST['mapname'];
        $maxlength=$_REQUEST['maxlength'];
        $data=$_REQUEST['geneinfo'];
        $lines=split("\n", $data);
        $i=0;
        $max=$maxlength;
        $min=0;
        $reverse=\$array_fill(0, 500, 0);
        $fwd=0;
        $rev=0;

        foreach($lines as $nextLine){
            // split input
            $test=preg_split("/\[\n\t\s,\-]+/", $nextLine);
            // remove blank space at end of array
            // store gene name in a separate array
            if(is_numeric($test[1]) == false){
            }
        }
    }
}
break;
}
$genename[$i]=$test[0];
//remove genename to calc max and min
array_splice($test, 0,1);
for($a=0; $a < sizeof($test); $a++){
    if(is_numeric($test[$a])=== false){
        if($test[$a]=='c'){
            $reverse[$i]=1;
        }
        array_splice($test, $a);
    }
}

foreach($test as $item){
    echo $item;
    echo "\t";
}

// sort the array
sort($test);

//reverse
if($reverse[$i]==1){
    $comp[$rev]=$test;
    $rname[$rev]=$genename[$i];
    echo "\t";
    echo "reverse";
    $rev+=1;
}

//forward array
else{
    $forward[$fwd]=$test;
    $fname[$fwd]=$genename[$i];
    echo "\t";
    echo "foward";
    $fwd+=1;
}

$genePosition[$i]=$test;
    echo "<br>";

    //  if($max<max($test)){
    //      $max=max($test);
    //  }
    //  if($min>min($test)){
    //      $min=min($test);
    //  }

$i+=1;
}

echo "max is : ";
echo $max;
echo " min is : ";
echo $min;
echo "<br>
$SESSION['minMaxFwd'][0]=$min;
$SESSION['minMaxFwd'][1]=$max;
$SESSION['mapName']=$mapname;
$SESSION['names']=$genename;
$SESSION['genePosition']=$genePosition;
$SESSION['reverse']=$reverse;
$SESSION['comp']=$comp;
$SESSION['forward']=$forward;
$SESSION['fname']=$fname;
$SESSION['rname']=$rname;
$ratio = 1;
if($max>100){
    $ratio=round(($max-$min)/100);
}
echo "ratio is: ";
echo $ratio;
echo "<br>
$yval=290;
for($i=0; $i<sizeof($genename); $i++){
    if(($i==0) && ($genename[$i] == $genename[$i-1])){
        $yval-=5;
    }
    for ($k=0; $k<sizeof($genePosition[$i])-1; $k+=2) {
        $end=intval((((genePosition[$i][$k+1]-$min)/$ratio)*7+50;
        $beg=intval(((genePosition[$i][$k]-$min)/$ratio)*7+50;
        echo "pixel beg and ends are";
        echo $beg;
        echo "$ ";
        echo $end;
        echo "<br>
    }
}
}
function OutputData(){
    // echo "output result here";
    echo "<br>
    echo "<img src=image.php>";
}
?>
<?

// image.php by Elizabeth Liu
session_start();
header("Content-type: image/png");
$im=@imagecreate(1600,1200)
    or die("cannot initialize new GD image stream");
$background_color=imagecolorallocate($im, 255,255,255);
$text_color=imagecolorallocate($im, 233, 14, 91);
$black=imagecolorallocate($im, 0, 0, 0);
$red=imagecolorallocate($im, 255, 0, 0);
$green=imagecolorallocate($im, 0, 255, 0);
imagesetthickness($im, 2);
imagestring($im, 5, 300, 10, $_SESSION['mapName'], $text_color);

//===================================================
// DRAW RULER
//===================================================
// draw a straight line with 2 red stoppers
imageline($im, 50, 300, 750, 300, $black);
imageline($im, 50, 295, 50, 305, $red);
imageline($im, 750, 295, 750, 305, $red);
$rulecount=1;
$numcount=0;
$tickcount=1;
imagestring($im, 2, 50, 306, "0", $text_color);
imagestring($im, 2, 750, 306, "100%", $text_color);
for ($r=57; $r<750; $r+=7){
    // at five, output taller tick mark
    if($rulecount==5){
        imageline($im, $r, 300, $r, 305, $black);
        $rulecount=0;
        $numcount+=1;
    }
    else{
        imageline($im, $r, 300, $r, 303, $black);
    }
    if($numcount==2){
        imagestring($im, 2, $r, 306, $tickcount, $text_color);
        $numcount=0;
    }
    $rulecount+=1;
    $tickcount+=1;
}

//===================================================
// Calculate Porportion
//===================================================
// by default ratio is 1, but not true in most cases
$ratio = 1;
if($_SESSION['minMaxFwd'][1]>100){
    $ratio=round($_SESSION['minMaxFwd'][1]-$_SESSION['minMaxFwd'][0])/100;
}
imagestring($im, 2, 5, 585, "Each Unit on the Map is Equivalent to ",$ratio, $black);
imagestring($im, 2, 235, 585, $ratio, $black);
?>
```php
imagestring($im, 2, 270, 585, "basepair", $text_color);
//===---------------------------------------------------------------
//=== Draw Actual Genes, Fwd and Comp strands
//===---------------------------------------------------------------
$yval=290;
$ryval=330;
//===---------------------------------------------------------------
//=== for comp strand
//===---------------------------------------------------------------
for($i=0; $i<sizeof($_SESSION['rname']); $i++){
    // if the names is the same as previous one, then probably very
close together, move up yval
    if((($i!=0) & ($_SESSION['rname'][$i]==$_SESSION['rname'][$i-1]))){
        $ryval+=15;
    }
    else{ // if different name reset to 290
        $ryval=330;
        if($i==0){
            // calc new beg pixel
            // if spliced
            $newbeg=round((($_SESSION['comp'][0]-$_SESSION['minMaxFwd'][0])/$ratio)*7+50;
            for($r=0; $r<sizeof($rrecord); $r++){
                // now check record see if too close or not
                if($rrecord[$r][2]==$ryval){
                    $namelen=strlen($_SESSION['rname'][$r]);
                    $namelen*=10;
                    if(($newbeg<$rrecord[$r][0]+$namelen)||($newbeg<$rrecord[$r][1]+7))
                        $ryval+=15;
                }
            }
        }
    }
}
for ($k=0; $k<sizeof($_SESSION['comp'][0])-1; $k+=2){
    // if almost out of frame, reset
    if($ryval>=585){
        $ryval=330;
    }
    $obeg=round((($_SESSION['comp'][$k]-$_SESSION['minMaxFwd'][0])/$ratio)*7+50;
    $end=round((($_SESSION['comp'][$k]+$k+1)-$_SESSION['minMaxFwd'][0])/$ratio)*7+50;
    $beg=round((($_SESSION['comp'][$k]+$k)-$_SESSION['minMaxFwd'][0])/$ratio)*7+50;
    imageline($im, $beg, $ryval, $end, $ryval, $black);
    // arrow
    imageline($im, $beg, $ryval, $beg+3, $ryval-3, $black);
    imageline($im, $beg, $ryval, $beg+3, $ryval+3, $black);
    imagestring($im, 2, $obeg, $ryval-13, $_SESSION['rname'][$i], $text_color);
```
if (is_numeric($_SESSION['comp'][$i][3])) {
    // draw V for spliced gene, calc midpoint
    $sbeg = round(($_SESSION['minMaxFwd'][0] -
                  $_SESSION['minMaxFwd'][0]/$ratio)*7+50;
    $send = round(($_SESSION['minMaxFwd'][2] -
                  $_SESSION['minMaxFwd'][2]/$ratio)*7+50;
    $mid = round(($send + $sbeg)/2);
    imageline($im, $sbeg, $ryval, $mid, $ryval+5, $green);
    imageline($im, $send, $ryval, $mid, $ryval+5, $green);
}

// for Forward strand
for ($l = 0; $l < sizeof($_SESSION['fname']); $l++) {
    // if names as same as previous one, move up, likely to be close
    if (($l != 0) && ($_SESSION['fname'][$l] ==
                   $_SESSION['fname'][$l-1])){
        $yval-=15;
    } else { // if different name reset to 290
        $yval=290;
        if ($l != 0){
            // calc new beg pixel
            $newbeg = round(($_SESSION['forward'][$l][0] -
                             $_SESSION['minMaxFwd'][0]/$ratio)*7+50;
            for ($f = 0; $f < sizeof($frecord); $f++){
                // now check record see if too close or not
                if ($frecord[$f][2] == $yval){
                    $namelen = strlen($_SESSION['fname'][$f]);
                    $namelen*=8;
                    if (($newbeg < $frecord[$f][0] + $namelen) || ($newbeg < $frecord[$f][1]+7)){
                        $yval-=15;
                    }
                }
            }
        }
    }
}

for ($m = 0; $m < sizeof($_SESSION['forward'][$l]); $m++) {
    if ($yval<=15){
        $yval=290;
    }
    $obeg = round(($_SESSION['forward'][$l][0] -
                   $_SESSION['minMaxFwd'][0]/$ratio)*7+50;
    $end = round(($_SESSION['forward'][$l][m] +
                  $_SESSION['minMaxFwd'][0]/$ratio)*7+50;
    $beg = round(($_SESSION['forward'][$l][m] -
                  $_SESSION['minMaxFwd'][0]/$ratio)*7+50;
    imageline($im, $beg, $yval, $end, $yval, $black);}
// arrow
imageline($im, $end, $yval, $end-3, $yval-3, $black);
imageline($im, $end, $yval, $end-3, $yval+3, $black);
imagestring($im, 2, $obeg, $yval-13, $_SESSION['fname'][$l],
$text_color);
}
$frecord[$l][0]=$obeg;
$frecord[$l][1]=$end;
$frecord[$l][2]=$yval;
if(sizeof($_SESSION['forward'][$l])==4){
  //draw V
  $sbeg=round(($SESSION['forward'][$l][1]-
  $_SESSION['minMaxFwd'][0])*$ratio)*7+50;
  $send=round(($SESSION['forward'][$l][2]-
  $_SESSION['minMaxFwd'][0])*$ratio)*7+50;
  $mid=round((($send+$sbeg)/2);
  imageline($im, $sbeg, $yval, $mid, $yval+5, $green);
  imageline($im, $send, $yval, $mid, $yval+5, $green);
}
imagejpeg($im);
imagedestroy($im);
<?

// map.php by Elizabeth Liu

echo "image here shown";
echo "<br>";

$mapname=$_REQUEST['mapname'];
$data=$_REQUEST['geneinfo'];

$lines=explode("\n", $data);

$i=0;
$max=0;
$min=1000000;

foreach($lines as $nextLine){
    // echo $nextLine;
    //problem, adding an additional space at end of the array still questionable
    //problem temporarily fixed
    $test=preg_split("/\[\n\s,-]+/", $nextLine);
    //remove blank space at end of array
    $n=sizeof($test);
    // if last item in array is numeric, then keep, if not then splice
    if(is_numeric($test[$n-1])===false){
        array_splice($test, -1);
    }
    //store gene name at separate array
    $genename[$i]=$test[0];
    //remove genename to calc max and min
    array_splice($test, 0, 1);
    // if next item is not numeric, then break
    if(is_numeric($test[0])===false){
        break;
    }
    if($max<max($test)){
        $max=max($test);
    }
    if($min<min($test)){
        $min=min($test);
    }

    $array[i]=$test;
    /*
    echo "number of genes: ";
    echo sizeof($array[i])/2;
    echo "<br>";*/
    $i++;
}

echo "max is : ";
echo $max;
echo "min is : ";
echo $min;
echo "<br>";
?>
Header.php

```php
session_start();
//header.php by Elizabeth Liu
?
<html>
<head>
<meta http-equiv="Pragma" content="no-cache">
<meta http-equiv="Cache-Control" content="no-cache">
<meta http-equiv="Expires" content="Sat, 01 Dec 2001 00:00:00 GMT">
<title>GeneMap-Elizabeth Liu</title>
</head>
<body>
<center>
<table border=2px width=800px bgcolor="#f2f5fc" cellpadding=20 cellspacing=0>
<tr><td><h1>GeneMapBeta</h1></td></tr>
</td></tr></table>
</body>
</html>
```

Footer.php

```php

<?
// footer.php by Elizabeth Liu

</td></tr>
<tr><td colspan=2>
<table width=100%>
<tr><td>To save the above image:<br>
For PC User: Right-click, Save Image / Copy Image. <br>
For Mac User: Click on image and drag to desktop or desired folder. <br>
This is a beta version, if any issue please contact: <a href="mailto:eliu1@gmu.edu?subject=Troubleshooting genemap">eliu1@gmu.edu</a></td></tr>
</table>
</td></tr>
</td></tr></table>
</body>
</html>
```
REFERENCES


Dehghan et al., 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, 265(2013).


strains that represent new candidate serotypes Ad50 and Ad51 of species B1 and D, respectively. J Clin Microbiol 37:3940-5.


EMBL-EBI 2015 http://www.ebi.ac.uk/Tools/psa/emboss_needle/nucleotide.html


Fenner, Frank J.; Gibbs, E. Paul J.; Murphy, Frederick A.; Rott, Rudolph; Studdert, Michael J.; White, David O. (1993). Veterinary Virology (2nd ed.). Academic Press, Inc.


PHP https://secure.php.net/ 2015


Elizabeth Liu grew up in Beijing and Virginia. She attended the Virginia Polytechnic Institute & State University, where she received her Bachelor of Science in 2005. She went on to receive her Master of Science from George Mason University in 2006. She then received her Doctorate in Science from George Mason University in 2015.