

Rapid Genome Analyses of Emergent Human Adenovirus 14a Causing 2006-7 Febrile Respiratory Illness (FRI) Outbreaks in the US via High Throughput “Next-Generation” Pyrosequencing Technique.

Huo-Shu H. Houg^{1)*}, Heping Gong¹⁾, Kathleen Verratti¹⁾, Lisa Lott, Leonard N. Binn¹⁾, Robert A. Kuschner¹⁾, David Metzgar²⁾, Kevin L. Russell²⁾, Adriana Kajon³⁾, Kuei-Hsiang Lin⁴⁾ and Julia A. Lynch¹⁾. Division of Viral Diseases, Walter Reed Army Institute of Research¹⁾, Silver Spring, MD; Naval Health Research Center²⁾, San Diego, CA; Lovelace Respiratory Research Institute³⁾, Albuquerque, NM. Division of Virology, Kaohsiung Medical University Hospital, Taiwan⁴⁾.

ABSTRACT

During 2006-7, Ad14a was identified during a series of FRI outbreaks across the US, involving at least ten documented pneumonia fatalities. Leveraging sequence data from the prototype strain Ad14p (GenBank # AY803294), the full genome sequence of Ad14a was determined using the conventional, and very labor-intensive, Sanger sequencing method. The same genome was analyzed using Pyrosequencing, an emerging alternative genome sequencing technology offering much higher efficiency. This direct shotgun approach relies on random sequencing of small DNA fragments using adaptor sequences, rather than independent amplification of separate fragments using pre-determined pathogen-specific sequences. This new sequencing strategy is therefore ideally suited for the rapid sequencing of hitherto uncharacterized human pathogens. The Roche 454 FLX system was used to sequence and assemble multiple Ad14a viruses from recent US outbreaks, as well as closely related Ad11a isolates causing non-US ARD infections since the 1970s. The US Ad14a strain significantly diverges from the

prototypical Eurasian strain, Ad14p, and shares greater than 98% genomic homology with Ad11a. Two genome types of Ad11, Ad11p and Ad11a display different tissue tropisms, causing renal and upper respiratory infections respectively. Ad14a and Ad11a share almost identical Fiber genes, which are known to be responsible for the adenoviruses' organ tropism, and both cause ARD infections. Both also share highly homologous Hexon genes, except for a 400 base pair (bps) region that allows these two viruses to be distinctly differentiated from each other based on serological cross reactivity. The origin of the emergent Ad14a could be related to recombination events that have shuffled the tissue tropism and antigen loci of ancestral Ad11 and Ad14 strains. High throughput sequencing is a powerful tool for rapid analysis of emerging pathogens, and can be used to generate comparative data offering information regarding the genome-wide relationship of those pathogens with well-characterized relatives.

INTRODUCTION

Human adenoviruses cause respiratory infections with symptoms ranging from, commonly, febrile respiratory illness (FRI) to, more rarely, pneumonia and death. In civilian populations human adenovirus infections occur sporadically in local or national epidemics. Epidemics are often associated with the emergence of new variants (genome types) of otherwise common serotypes, such as Ad4, or the recent emergence of rare serotypes. Adenoviruses of three species,

AdB, AdC and AdE, are frequently associated with respiratory disease in mostly young and healthy population. AdC serotypes include Ad1, Ad2, Ad5 and Ad6, and are endemic among children and young adults. Pre-existing antibodies against AdC are extremely prevalent among general population. Thus, many adults are immune to these serotypes. The sole AdE serotype, Ad4, and several other AdB serotypes including Ad3, Ad7 and Ad21, are commonly associated with epidemic outbreaks of FRI and pneumonia in healthy adults and children throughout the world. Two closely related AdB serotypes, Ad11a and Ad14, had until recently only been identified in association with respiratory disease epidemics in rare (though severe) outbreaks in Eurasia and southeastern Asia region. Prior to 2006-7 North America outbreaks, Ad14 had never been isolated from any case in North America.

During 2006 and 2007, Ad14 was identified in a series of outbreaks across the United States, in association with widespread FRI and at least ten documented pneumonia fatalities. This phenomenon was simultaneously tracked in both civilian populations, by the CDC and local public health agencies, and in military recruits, by US Department of Defense public health agencies. Detailed comparisons of collected strains, including sequencing of fiber and hexon genes and whole-genome restriction analysis (genome typing), revealed that all of these events were caused by a single, apparently homogeneous strain. The identified strain was significantly diverged from the prototypical Eurasian strain, Ad14p identified in the mid 1900s. Full genome

sequencing of multiple Ad14a isolates, including both a fatal pneumonia isolate from a severe outbreak and an isolate from a mild outbreak that had little effect on local adenoviral illness rates, revealed only two genetic polymorphisms between the two strains, one of which was a synonymous base mutation in the fiber gene (HS Houg, unpublished data). This high degree of homogeneity (clonality) did not offer a simple way, such as pathogen or genotype specific PCR to track different lineages of the emerging viruses that have significant impact on US military personnel. Scientists at Walter Reed Army Institute of Research illustrated the utilities of recently developed high throughput “Next-Generation” Pyrosequencing Technique for rapid genome analyses of emergent human Adenovirus 14a Causing 2006-7 Febrile Respiratory Illness (FRI) Outbreaks in the US. “Next-Generation” Pyrosequencing Technique offers several advantages for full viral genomic studies. It eliminates the requirement for the target specific PCR/sequencing primers during PCR amplification and sequencing processes. And it also extends the US DoD’s capacity in identifying and detecting the future emergent pathogens with no known reference sequence available.

MATERIALS & METHODS

Sample collection. All Ad14 isolates of 2006-7 outbreaks from basic training facilities, aside from Lackland, were collected, identified, and analyzed as part of the Naval Health Research Center (NHRC)'s ongoing population-based FRI

surveillance program. Ad14s of Lackland origin were comprised of NHRC surveillance samples and samples collected from severely ill patients at Wilford Hall Medical Center located at Lackland. Those samples and samples from advanced Air Force training centers were sent to the Air Force Institute of Occupational Health (AFIOH) for diagnostic viral culture and were provided to NHRC as de-identified isolates. All samples collected by NHRC were provided for research use under informed consent and internal review board-approved human use protocols (Protocol # NHRC.2005.0017).

NHRC samples were collected as oropharyngeal (throat) swabs in VTM (Remel, Lenexa, KS), immediately frozen in either -80 freezers or in dry ice, and transported on dry ice to NHRC under CAP-accredited collection and transport protocols. AFIOH samples were collected as throat swabs in VTM, cultured in A549 cells and transported to NHRC as above. All of the above samples were tested at NHRC for Ad14 (see PCR and sequencing methods below) as raw specimens, then subsequently cultured in A549 cells (Diagnostic Hybrids, Athens, OH) and stored frozen as infected tissue culture fluid (isolated virus). Sequencing work on these samples was performed on the resulting isolates. Following PCR identification as Ad14, chosen samples were extracted and aliquotted at NHRC and transported frozen on dry ice to Walter Reed Army Institute of Research (WRAIR) for sequence analysis.

Ad14-specific PCR amplifications and Conventional Sanger-DNA sequencing of Ad14s. PCR and Sequence

analysis was accomplished, using the methods detailed in the next paragraph, at the CLIA/CLIP accredited virology facility at Walter Reed Army Institute of Research. PCR primer pairs designed and used to generate overlapped 1-2 kilobase (kb) amplicons to cover the entire genomes of various Ad14 isolates were derived from Ad14 deWit prototype sequence (GenBank Accession AY803294). All PCR products were sequenced in both directions by using forward or reverse PCR primers corresponding to each individual PCR product. All clean and verified readable sequences were used to assemble full Ad14 genome sequence via using Sequencer program (Gene Codes Co., Ann Arbor, MI).

200 µl aliquots of Ad14 samples were extracted using the Invitrogen Charge-Switch DNA extraction kit (Invitrogen Inc., CA) per the manufacturer's instructions, and eluted into 200 µl of buffer. 100 µl PCR amplification reactions consisted of 2mM MgCl₂, 0.6mM dNTP (1.5mM each A, C, T and G), 200 nM each primer (see previous paragraph), 2.5 units of Platinum Taq Polymerase (Invitrogen, Carlsbad, CA), and 1ul of extracted sample in 1X ABI Buffer II (Applied Biosystems, Foster City, CA). Thermal cycling was carried out on an ABI9700 platform (Applied Biosystems) using the following parameters: Initial activation for 2min at 94°C, then 35 cycles of: 20sec at 94°C, 20sec at 53°C, and 2min at 72°C. Final extension was for 7min at 72°C. PCR cleanup was performed using the Qiagen PCR Cleanup Kit (Qiagen) per the manufacturer's instructions. Sequencing reactions were set up per the manufacturer's instructions using the ABI Big Dye

Terminator Kit (manual version 3.2, Applied Biosystems), and run on an ABI9700 platform. Reaction products were analyzed on an ABI3130XL automated sequencer (Applied Biosystems) per the manufacturer's instructions. Resulting data was then edited and aligned using Mac Sequencer software (Gene Codes Inc, Ann Arbor, MI).

Full Ad14 genome sequencing via “Next-Generation” pyrosequencing sequencing. GS DNA Library Preparation Kit (Roche 454 Life Science, Branford, CT) was used to process Ad14 DNA sample, 1-2 µg per virus into a library of single-stranded template DNA fragments (sstDNA). Such a library can then be used as input to the GS emPCR amplification (Shotgun), and further sequenced in the Genome Sequencer System, developed by 454 Life Sciences Corporation. The sequencing reads obtained from Roche 454 FLX sequencers of each individual virus were used assemble into a single contig representing as a full and complete Ad14 genome sequence.

Bioinformatic analysis of assembled Ad14 genomes. Full Ad14 genome sequences obtained from this study derived from both conventional Sanger-sequencing and Pyrosequencing were compared using Clustal W alignment program (DNASstart, WI).

Results & Discussions

Assembling full Ad14 genomic sequences of 2006-7 US outbreaks via Ad14-specific PCR amplifications followed by conventional Sanger DNA sequencing method. We

demonstrated the utilities of the methods applied to various full Ad14 genomic sequencing derived from the recent 2006-7 North America outbreaks, including a paired Ad14s isolated from severe (Ad14 LL1986T) and mild (Ad14 LL303600) ARD infected patients from Lackland, Texas. Another Ad14 isolated from mild ARD patient of San Diego origin, Ad14 NHRC 22039 was also employed to obtain the full genome sequence in this study. At present, almost all of GenBank Ad viral DNA data were derived from either PCR amplicons or recombinant cloned DNA fragments via conventional Sanger DNA sequencing method. We decided to sequence and assemble full Ad14 genome sequence derived from Ad14 isolated from various military basic training camps including Ad14 that caused either severe or mild ARD infections. This would allow scientists to understand molecular epidemiology and possible molecular pathogenesises of the recent Ad14s infections of US origin as compared to the prototype Ad14p de Wit infections in 1950s. As described in Materials & Methods, multiple Ad14 PCR primer pairs (derived from Ad14 deWit prototype sequence, GenBank Accession AY803294) were used to generate overlapped 1-2 kilobase (kb) amplicons to cover the entire genomes of the interested Ad14 strains. All PCR products were sequenced in both directions by using forward or reverse PCR primers corresponding to each individual PCR product. All clean and verified readable sequences were used to assemble full Ad14 genome sequence via using Sequencer program (Gene Codes Co., Ann Arbor, MI). In order to finish any particular Ad14 strain, it regularly took more than

two months of efforts after testing more than 100s Ad14 specific primer pairs through PCR amplifications followed by Sanger DNA sequencing of PCR products. It's not straightforward in sequencing high G/C content regions of human Ad14s that cause failures in PCR amplifications or Sanger BigDye terminator reactions. Thus, the actual time of completing full Ad14 genome becomes unpredictable, i.e., greater than two months in searching for workable PCR/sequencing primers to fill the sequencing gaps.

Assembling various full Ad14 genomic sequences of 2006-7 US outbreaks via “Next-Generation” pyrosequencing sequencing. We illustrated and confirmed the utilities of “Next Generation” Pyrosequencing could be used to generate massive quantities, greater than 15 Mbs human Ad14 sequences per virus from each DNA sequencing experiment. Up to 8 different viruses could be fully sequenced per pyrosequencing run. It was proven that 15 Mbs sequencing data per virus provide ample depth of full genome coverage, i.e., up to 4-500 times coverage for the entire 34,768 bps Ad14 genomes of Lackland origin, such as, Ad 14 LL1986T and LL303600.

Accurate & Compatible/Identical full Ad14 genomic sequences of 2006-7 US outbreaks derived from conventional Sanger DNA sequencing and “Next-Generation” pyrosequencing sequencing. In this study, we have adapted both the conventional Sanger sequencing methodology as well as “Next Generation” Pyrosequencing

technique to assemble compatible/identical full genome sequences of human adenoviruses, Ad14s of 2006-7 outbreaks. We illustrated that “Next-Generation” pyrosequencing sequencing technology can be used to replace labor intensive Sanger DNA sequencing method to generate accurate full Ad14 genome sequences as compared to the reference Sanger DNA sequencing Ad14 sequences. Most of all, “Next-Generation” pyrosequencing sequencing offers tremendous time saving, i.e., multiple Ad14s up to 8 different strains could be sequenced and assembled in less than 5 working days. In addition to Ad14s, the Roche 454 FLX system was used to sequence and assemble another closely related Ad11a isolates causing non-US ARD infections (mostly in Southeastern Asia) since the 1970s. It was shown that the US Ad14a strain significantly diverged from the prototypical Eurasian strain, Ad14p, and shares greater than 98% genomic homology with Ad11a. Two genome types of Ad11, Ad11p and Ad11a display different tissue tropisms, causing renal and upper respiratory infections respectively. Ad14a and Ad11a share almost identical Fiber genes, which are known to be responsible for the adenoviruses' organ tropism, and both cause ARD infections. Both also share highly homologous Hexon genes, except for a 400 base pair (bps) region that allows these two viruses to be distinctly differentiated from each other based on serological cross reactivity. The origin of the emergent Ad14a could be related to recombination events that have shuffled the tissue tropism and antigen loci of ancestral Ad11 and Ad14 strains. High throughput sequencing is a powerful tool for rapid

analysis of emerging pathogens, and can be used to generate comparative data offering information regarding the genome-wide relationship of those pathogens with well-characterized relatives. This lately developed Next-Generation” pyrosequencing technology will be an invaluable tool to quickly uncover and study potential future emergent infectious diseases by assembling full and accurate pathogen genomes without any previously known literature or reference sequences available.

References

1. **Altschul, S.F., W. Gish, W. Miller, E. W. Myers, D. J. Lipman.** 1990. Basic local alignment search tool. *J. Mol. Biol.* **215**:403–10.
2. **Baum, S.** 2005. Adenovirus, p. 1835-1840. *In* G. L. Mandell, J. E. Bennett, and R. Dolin (ed.), *Principles and practice of infectious diseases*, 6th ed., vol. 2, Churchill Livingstone, Philadelphia, Pa.
3. **Bruj, J., J. Farnik, and V. Sedmidubsky.** 1966. Epidemic of acute respiratory disease due to adenovirus type 14. *Cesk. Epidemiol. Mikrobiol. Imunol. (Czech).* **15**:165-171.
4. **Centers for Disease Control and Prevention.** 2007. Acute respiratory disease associated with adenovirus serotype 14—four states, 2006-2007. *MMWR.* **56(45)**:1181-1184.
5. **Davies K.** 2005. Fantastic 454: Biotech Unveils Rapid Genome Sequencing Platform. *Bio-IT World.*
6. **Goldberg, S.M., et al.** 2006. A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. *Proc. Natl. Acad. Sci. USA* **103**:11240-11245.
7. **Kajon, A. E., and G. Waddell.** 1996. Sequence analysis of the E3 region and fiber gene of human adenovirus genome type 7h. *Virology.* **215**:190-196.
8. **Kajon, A. E., J. M. Moseley, D. Metzgar, H. S. Houg, A. Wadleigh, M. A. Ryan, and K. L. Russell.** 2007. Molecular epidemiology of adenovirus type 4 infections in US military recruits in the postvaccination era (1997-2003). *J. Infect. Dis.* **196(1)**:67-75.
9. **Lin, B., Z. Wang, G. J. Vora, J. A. Thornton, J. M. Schnur, D. C. Thach, K. M. Blaney, A. G. Ligler, A. P. Malanoski, J. Santiago, E. A. Walter, B. K. Agan, D. Metzgar, D. Seto, L. T. Daum, R. Kruzelock, R. K. Rowley, E. H. Hanson, C. Tibbetts, and D. A. Stenger.** 2006. Broad-spectrum respiratory tract pathogen identification using resequencing DNA microarrays. *Genome Res.* **16(4)**:527-535.
10. **Lin, K. H., Y. C. Lin, H. L. Lin, G. M. Ke, C. J. Chiang, K. P. Hwang, P. Y. Chu, J. H. Lin, D. P. Lin, and H. Y. Chen.** 2004. A two decade survey of respiratory adenovirus in Taiwan: the reemergence of adenovirus types 7 and 14. *J. Med. Virol.* **73**:274-279.
11. **Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LS, Berka J, Braverman MS, Chen Y-J, Chen Z, Dewell SD, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH,**

- Irzyk GP, Jando SC, Alenquer MLI, Jarvie TJ, Jirage KB, Kim J-B, Knight JR, Lanza JP, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Weiner MP, Yu P, Begley RF and Rothberg JM.** 2005. Genome sequencing in microfabricated high-density picolitre reactors . *Nature* 437.
12. **Metzgar, D., and C. Wills.** 2000. Evidence for the adaptive evolution of mutation rates. *Cell*. **101**:581-584.
13. **Metzgar, D., D. Field, R. Haubrich, and C. Wills.** 1998. Sequence analysis of a compound coding-region microsatellite in *Candida albicans* resolves homoplasies and provides a high-resolution tool for genotyping. *FEMS Immunol. Med. Microbiol.* **20**:103-109.
14. **Nakamura, Y., M. Leppert, P. O'Connell, R. Wolff, T. Holm, M. Culver, C. Martin, E. Fujimoto, M. Hoff, E. Kumlin, and R. White.** 1987. Variable number tandem repeat (VNTR) markers for human gene mapping. *Science*. **235**:1616–1622.
15. **Noda, M., T. Yoshida, T. Sakeguchi, Y. Ikeda, K. Yamaoka, and T. Ogino.** 2002. Molecular and epidemiological analysis of human adenovirus type 7 strains isolated from the 1995 nationwide outbreak in Japan. *J. Clin. Microbiol.* **40**:140-145.
16. **Potter, C. W., and W. I. H. Shedden.** 1963. The distribution of adenovirus antibodies in normal children. *J. Hyg. (London)* **61**:155-160.
17. **Rogers Y-H. and Venter JC.** Massively parallel sequencing. *Nature* 437, 326-327 (2005).
18. **Rubin, B.A.** 1993. Clinical picture and epidemiology of adenovirus infections. *Acta Microbiol. Hung.* **40**:303-323.