# The B-55’s: I am livin’ on Channel B!

## Abstract

*\*Human adenovirus B-55\** (HAdV-B55) is a re-emergent pathogen that threatens dense populations worldwide.

## Introduction

### Significance

*Human adenovirus* is a model organism with significant biotechnology applications. It is also a diverse pathogen with global distribution, causing sporadic outbreaks, originally and usually affecting military populations. Consequently, the U.S. military vaccinates recruits against types E4 and B7. Recent outbreaks of different types in civilian populations are on the rise. B55 is no exception. Data acquired from these outbreaks has yielded a data set suitable for molecular clock analysis. This research characterizes the molecular evolution of B55, testing different clock models and calculating the substitution rate and time to most recent common ancestor (TMRCA). This data helps public health officials make informed decision with respect to biosurveillance efforts and vaccine development for the general population. In addition, these methods are applicable to other emerging types.

### History

In 1953, an epidemic occurred at the Fort Leonard Wood U.S. Army installation in Missouri. A patient presented with pneumonia-like symptoms and provided a throat wash sample that contained the first viral isolate, initially called “adenoid degeneration agent” and later adenovirus 1,2. Subsequent outbreaks resulted in the discovery of numerous other types, including HAdV-B55.

The first B55 samples originated during a 20 year period starting in 1965 at the 302nd Hospital in Beijing, China 3. In March 1969, the first military outbreak of B55 occurred at the Sant Climent Sescebes camp of Alt Empordà, Girona, Catalonia, Spain 4. The virus continues spreading throughout the world, primarily affecting civilian populations. Figure 1 plots date ranges for outbreaks associated with surveillance studies 3–18 and GenBank accessions with collection date and country metadata.

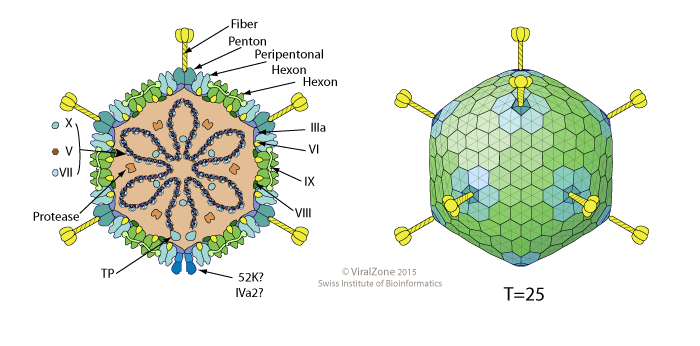
#### Figure 1. Surveillance Study Outbreak Ranges and GenBank Accession Collection Dates



### Biology

The \**Adenoviridae\** family includes genera that infect a wide range of hosts and cell types. The \**Mastadenovirus\** genus includes species that infect mammalian hosts. They are globally distributed, causing sporadic outbreaks in densely populated regions and close-living quarters. Carriers may expose others via aerosol or fecal-oral transmission, possibly asymptomatically 19. Symptoms range from acute respiratory disease to organ failure, depending on the viral species and host immune strength 19. Accordingly, individuals with developing or weakened immune systems account for most outbreak deaths. The *Adenoviridae* are class I, linear, double-stranded DNA viruses 20. The nonenveloped icosahedral nucleocapsid consists of hexon and penton capsomers forming the faces and vertexes respectively. The coxsackie adenovirus receptor recognizes the fiber knob while the penton RGD motif induces structural changes to gain entry into the host 21.

#### Figure 2. Adenovirus Virion Structure



B55 is a re-emergent respiratory pathogen with a B14 genomic backbone and a recombinant hexon partially derived from B11. A new typing scheme that includes genomic analysis corrected it’s previous misidentification as B11a due to limitations associated with serological assays with respect to recombination effects 22.

### Prior Work

Previous adenovirus research has included molecular clock analysis to characterize outbreak samples. A common protocol involves a preliminary pseudo-statistical test of the strict molecular clock hypothesis followed by Markov chain Monte Carlo (MCMC) simulation to estimate model parameters and unknown uncertainties. Tracing each MCMC run evaluates convergence and determines whether the estimates sample size (ESS) of each parameter of interest is acceptable. A model selection step then compares runs with different clock hypotheses, coalescent models, and substitution models. The result is an estimation of model parameters with confidence intervals, such as divergence times and substitution rates.

Similar protocols appeared in recent adenovirus research. One paper analyzed B55 isolates between June 2009 and January 1012 in Chongqing, China 10 and another paper analyzed E4 isolates obtained between 1953 and 2015 in the United States and Japan 23. However, both papers don’t account for the time-dependent rate phenomenon (TDRP), where mutation rates obscure substitution rates in short-term studies 24. The latter paper also didn’t include confidence intervals around the node age estimates in the chronogram.

### Proposed Work

This research aims to characterize the molecular evolution of B55 using all publicly available sequence data. This includes comparison of the clock models and substitution rates of the genome, polymerase, and surface proteins. The hypothesis is that the fiber, hexon, and penton genomes evolve faster than the genome and polymerase in order to sample a larger sequence space to enter the host. This may account for the wide variety of hosts and cell types. A robust investigation must account for the TDRP. This is critical since available sequence data only goes back to 1969. Recent papers applied a methodology using host fossil-derived node calibration to offset the effect 25,26. Such an approach is applicable to *Mastadenovirus*.

#### Preliminary Work

So far, this research has completed investigations into the molecular clock regarding the genome, hexon, and fiber of B55. This includes root-to-tip regression analyses, Bayesian MCMC simulation, and chronogram rendering of maximum credibility clades for each set of sequences. Different runs tested several clock and coalescent models. The model selection step determined the best model for each by calculating the Bayes factor. A GitHub repository maintains the set of programs developed to coordinate pipelines and normalize data for each experiment.

## Methods

### Data

A Bash script coordinated the retrieval of nucleotide sequences, generating a local BLAST database. The Entrez Direct E-utilities scripts provided a command-line interface to query and retrieve data from the federated set of National Center for Biotechnology Information (NCBI) databases 27. The script generated the following query to download the sequences from the Nucleotide database.

```sql

(txid714978[PORG] OR txid343463[PORG]) AND biomol\_genomic[PROP] NOT gbdiv\_pat[PROP] NOT gbdiv\_syn[PROP]

```

Each search value has an associated keyword field enclosed within square brackets. The PORG and PROP fields correspond to primary organism and property respectively. The first two PORG terms use NCBI Taxonomy identifiers to limit the search to HAdV-55 and HAdV-11a respectively. The first PROP term sets the molecule type while the next two exclude GenBank divisions 28. Specifically, these terms limit the search to genomic DNA while excluding patents and synthetic constructs respectively. The \*\*esearch\*\* and \*\*efetch\* programs executed the query and retrieved the FASTA-formatted sequence data respectively. The script piped the results directly into the \*\*makeblastdb\*\* program to generate a BLAST database.

### Metadata

The same Bash script also coordinated the retrieval and normalization of sequence metadata to create a database of “collection\_date” values for subsequent molecular clock calibration and analysis. The script retrieved the set of sequence accessions via \*\*blastdbcmd\*\* and piped them into the \*\*esummary\*\* to download the JSON-formatted metadata from Nucleotide. The \*\*jq\*\* program executed the following query to transform the records into a tab-separated file.

```jq

.result | del(.uids) | map([.accessionversion, .subtype, .subname] | @tsv) | .[]

```

Each record contains a “subtype” and “subname” property. They are pipe-delimited string values that correspond to keys and values respectively. Together, the key-value pairs represent sequence metadata contained within the zeroth sequence feature of a GenBank file. An R script used \*\*tidyverse\*\* functions to split these property values and bind the resulting key-value pairs into new columns 29. Another R script relied on the \*\*lubridate\*\* package to automatically convert the date field values into a consistent ISO-8601 format 30. This series of commands generated a sorted tab-separated file, mapping accessions to collection dates.

### Sequence Extraction

The full genome extraction method involved a series of piped commands. The \*\*blastdbcmd\*\* program dumped a space-separated list of accession-length pairs. Next, \*\*awk\*\* selected accessions with lengths ≥ 34 kbp.

The gene extraction method was more involved, requiring a reference sequence. The \*\*blastdbcmd\*\* program retrieved the reference based on an accession and sequence coordinates. The \*\*blastn\*\* program used the reference to query the database using the \*megablast\* task, generating a list of subject accessions. The \*\*blastdbcmd\*\* program extracted the complete sequences for subsequent global-local alignment using \*\*glsearch36\*\*. This step guarantees complete alignment of the query to an optimal region on the subject. An \*\*awk\*\* script selected hits with sequence identity ≥ 90%.

For both methods, a series of piped \*\*awk\*\*, \*\*sort\*\*, \*\*join\*\* invocations generated \*\*sed\*\* command files for subsequent modification of the headers to include the dates based on the “collection\_date” database.

### Alignment

The \*\*mafft\*\* program performed multiple sequence alignments 31. MAFFT achieves performance gains via multithreading while maintaining accuracy via application of the Fast Fourier Transform on the sequence data to quickly identify homologous regions. Parameters included the “--auto” and “--adjustdirection” flags. The former automatically sets algorithm heuristics based on the sequence data and the latter automatically adjusts sequence direction for each entry if the reverse complement is optimal. The calling script redirected standard error into a log file to record alignment progress and heuristic selection.

### Phylogeny

The IQ-TREE program inferred phylogenies 32. The program performed a series of likelihood tests to select the optimal number of threads and sequence evolution model based on the input data. The former compared the effect of adding additional threads on efficiency and the latter exploited the ModelFinder algorithm to estimate the optimal substitution model 33. Parameters included the -alrt and -bb flags to set the number of bootstrap replicates to 1,000 for the approximate likelihood ratio test of branches 34 and branch support 35. The -bnni parameter also reduced the risk of model violations associated with ultrafast bootstrap testing via nearest neighbor interchange. The program automatically created a log file and exported the maximum likelihood tree in a variety of formats, including Newick.

### Molecular Clock

The TempEst program tests the strength of the strict molecular clock hypothesis for a given phylogeny 36. The program imports a tree file and parses the tip labels to extract the dates, plotting them against the root-to-tip patristic distance and fitting a regression line with an objective function that optimizes the correlation coefficient, R-squared value, or mean-squared residuals. This is an interactive tool that facilitates the identification of outliers that may result from incorrect collection dates, vaccine strains, or contaminated sequence data. An R script automated this process by plotting the model for the cross product of root-to-tip distance metrics and model objective functions. The script invoked the rtt and distRoot function of the ape 37 and adephylo 38 packages.

The BEAUti program is a graphical tool that outputs BEAST model parameters as XML files 39. The program imports a multiple alignment file and parses the sequence headers to extract the dates. Tip date sampling parameters included a uniform sampling distribution with an uncertainty of 10 years. Model testing from the IQ-TREE logs informed substitution model and base frequency parameter settings. Clock models included the strict clock, relaxed clock with lognormal distribution, and relaxed clock with exponential distribution 40. Tree priors included the Constant Size 41, Exponential Growth 42, and Bayesian Skyline 43 coalescent models. The MCMC parameters included a chain length of 108 with 10-3 sampling frequency. Marginal likelihood estimation included the path sampling (PS) / stepping-stone sampling method with 100 steps, chain length 106, sampling frequency 10-3, and Beta path step distribution 44,45.

BEAUti exported separate XML files representing the cross product of clock and coalescent models while maintaining all other settings and parameters constant. The \*\*beast\*\* program ran the MCMC simulation for each file. The \*\*treeannotator\*\* program subsequently calculated the maximum credibility clade using a 10% state burn-in and minimum posterior probability limit of 50%. A script submitted each job to a high-performance computing cluster queue, requesting 32 processors.

## Results

### Substitution Model Selection

ModelFinder program of IQ-TREE generated recommended the substitution models listed in Table 1 for each sequence set according to different information criterions. This study used the Bayesian information criterion (BIC). Each BEAST run incorporated the corresponding site model.

### Table 1. Substitution Model Selection

|  |  |  |  |
| --- | --- | --- | --- |
| Sequence | BIC | AIC | AICc |
| Genome | K2P | K3Pu+F+I | K3Pu+F+I |
| Hexon | HKY+F | TN+F | TN+F |
| Fiber | F81+F | TPM3u+F | TPM3u+F |

### Clock Signal Analysis

The roottotip function of the BactDating R package performed the root-to-tip analysis. Figure 3 shows the estimated paramters with a rooted phylogram and regression for the genome sequences.



### Clock/Coalescent Model Selection

An R script computed the Bayes factor for each BEAST run based on the maximum likelihood estimation log. The best model is the one with the largest Bayes factor. Table 2 lists the results based on the path sampling method. In the table, str, rel, exp, con, and sky stand for strict, relaxed, exponential, constant, and Bayesian skyline.

#### Table 2. Path Sampling Bayes Factor Results

|  |  |  |  |
| --- | --- | --- | --- |
| lab | clock.coalescent | PS | PS (BF) |
| Genome | rel-exp.con | -50223.707 | 5.71213766 |
| rel-exp.exp | -50224.133 | 5.28619062 |
| rel-exp.sky | -50224.825 | 4.59395436 |
| str.sky | -50226.987 | 2.43267613 |
| str.exp | -50227.305 | 2.11472655 |
| str.con | -50229.419 | 0.00000000 |
| rel-log.con | -50230.947 | -1.52797305 |
| rel-log.exp | -50231.132 | -1.71305549 |
| rel-log.sky | -50232.355 | -2.93529225 |
| Hexon | str.con | -4075.134 | 0.00000000 |
| rel-exp.con | -4075.987 | -0.85273886 |
| rel-exp.sky | -4076.176 | -1.04191964 |
| rel-log.con | -4076.734 | -1.60004075 |
| str.sky | -4078.266 | -3.13208188 |
| rel-log.sky | -4079.408 | -4.27381500 |
| rel-log.exp | -4081.664 | -6.52990852 |
| rel-exp.exp | -4081.903 | -6.76917055 |
| str.exp | -4091.258 | -16.12425015 |
| Fiber | str.con | -1449.184 | 0.00000000 |
| rel-log.sky | -1449.245 | -0.06148823 |
| rel-exp.sky | -1449.547 | -0.36369448 |
| rel-exp.con | -1449.901 | -0.71693255 |
| str.sky | -1450.028 | -0.84390220 |
| rel-log.con | -1450.115 | -0.93144787 |
| rel-log.exp | -1456.579 | -7.39543549 |
| str.exp | -1458.150 | -8.96648035 |
| rel-exp.exp | -1458.746 | -9.56237700 |

### Chronogram

…

## Discussion

…

### Timeline

…

## References

1. Hilleman, M. R. & Werner, J. H. Recovery of New Agent from Patients with Acute Respiratory Illness. *Proc. Soc. Exp. Biol. Med.* **85**, 183–188 (1954).

2. Rowe, W. P., Huebner, R. J., Gilmore, L. K., Parrott, R. H. & Ward, T. G. Isolation of a Cytopathogenic Agent from Human Adenoids Undergoing Spontaneous Degeneration in Tissue Culture. *Proc. Soc. Exp. Biol. Med.* **84**, 570–573 (1953).

3. Li, Q., Hambraeus, J. & Wadell, G. Genetic Relationship between Thirteen Genome Types of Adenovirus 11, 34, and 35 with Different Tropisms. *Intervirology* **32**, 338–350 (1991).

4. Hierholzer, J. C., Pumarola, A., Rodriguez-Torres, A. & Beltran, M. OCCURRENCE OF RESPIRATORY ILLNESS DUE TO AN ATYPICAL STRAIN OF ADENOVIRUS TYPE 11 DURING A LARGE OUTBREAK IN SPANISH MILITARY RECRUITS. *Am. J. Epidemiol.* **99**, 434–442 (1974).

5. Kajon, A. E. *et al.* Molecular epidemiology of adenovirus acute lower respiratory infections of children in the South Cone of South America (1991–1994). *J. Med. Virol.* **48**, 151–156 (1996).

6. Chmielewicz, B. *et al.* Respiratory disease caused by a species B2 Adenovirus in a military camp in Turkey. *J. Med. Virol.* **77**, 232–237 (2005).

7. Kajon, A. E. *et al.* Outbreak of Febrile Respiratory Illness Associated with Adenovirus 11a Infection in a Singapore Military Training Camp. *J. Clin. Microbiol.* **48**, 1438–1441 (2010).

8. Zhu, Z. *et al.* Outbreak of Acute Respiratory Disease in China Caused by B2 Species of Adenovirus Type 11. *J. Clin. Microbiol.* **47**, 697–703 (2009).

9. Kajon, A. E. *et al.* Molecular and serological characterization of species B2 adenovirus strains isolated from children hospitalized with acute respiratory disease in Buenos Aires, Argentina. *J. Clin. Virol. Off. Publ. Pan Am. Soc. Clin. Virol.* **58**, 4–10 (2013).

10. Lu, Q.-B. *et al.* Epidemiology of human adenovirus and molecular characterization of human adenovirus 55 in China, 2009–2012. *Influenza Other Respir. Viruses* **8**, 302–308 (2014).

11. Cao, B. *et al.* Emergence of Community-Acquired Adenovirus Type 55 as a Cause of Community-Onset Pneumonia. *Chest* **145**, 79–86 (2014).

12. Jing, S. *et al.* Household Transmission of Human Adenovirus Type 55 in Case of Fatal Acute Respiratory Disease - Volume 25, Number 9—September 2019 - Emerging Infectious Diseases journal - CDC. doi:10.3201/eid2509.181937.

13. Salama, M. *et al.* Outbreak of adenovirus type 55 infection in Israel. *J. Clin. Virol.* **78**, 31–35 (2016).

14. Li, X. *et al.* An outbreak of acute respiratory disease in China caused by human adenovirus type B55 in a physical training facility. *Int. J. Infect. Dis.* **28**, 117–122 (2014).

15. Lafolie, J. *et al.* Severe Pneumonia Associated with Adenovirus Type 55 Infection, France, 2014. *Emerg. Infect. Dis.* **22**, 2012–2014 (2016).

16. Sun, B. *et al.* Emergent severe acute respiratory distress syndrome caused by adenovirus type 55 in immunocompetent adults in 2013: a prospective observational study. *Crit. Care* **18**, 456 (2014).

17. Yi, L. *et al.* A cluster of adenovirus type B55 infection in a neurosurgical inpatient department of a general hospital in Guangdong, China. *Influenza Other Respir. Viruses* **11**, 328–336 (2017).

18. Yoo, H. *et al.* Febrile Respiratory Illness Associated with Human Adenovirus Type 55 in South Korea Military, 2014–2016. *Emerg. Infect. Dis.* **23**, 1016–1020 (2017).

19. Lynch III, J. P. & Kajon, A. E. Adenovirus: Epidemiology, Global Spread of Novel Serotypes, and Advances in Treatment and Prevention. *Semin. Respir. Crit. Care Med.* **37**, 586–602 (2016).

20. Baltimore, D. Expression of animal virus genomes. *Bacteriol. Rev.* **35**, 235–241 (1971).

21. Pettersson, U. Encounters with adenovirus. *Ups. J. Med. Sci.* 1–11 (2019) doi:10.1080/03009734.2019.1613698.

22. Walsh, M. P. *et al.* Computational Analysis Identifies Human Adenovirus Type 55 as a Re-Emergent Acute Respiratory Disease Pathogen. *J. Clin. Microbiol.* **48**, 991–993 (2010).

23. Gonzalez, G. *et al.* Genomic characterization of human adenovirus type 4 strains isolated worldwide since 1953 identifies two separable phylogroups evolving at different rates from their most recent common ancestor. *Virology* **538**, 11–23 (2019).

24. Ho, S. Y. W., Phillips, M. J., Cooper, A. & Drummond, A. J. Time Dependency of Molecular Rate Estimates and Systematic Overestimation of Recent Divergence Times. *Mol. Biol. Evol.* **22**, 1561–1568 (2005).

25. Aiewsakun, P. & Katzourakis, A. Time dependency of foamy virus evolutionary rate estimates. *BMC Evol. Biol.* **15**, 119 (2015).

26. Aiewsakun, P. & Katzourakis, A. Marine origin of retroviruses in the early Palaeozoic Era. *Nat. Commun.* **8**, 1–12 (2017).

27. Kans, J. *Entrez Direct: E-utilities on the UNIX Command Line*. (National Center for Biotechnology Information (US), 2019).

28. Romiti, M. & Cooper, P. *Search Field Descriptions for Sequence Database*. (National Center for Biotechnology Information (US), 2011).

29. Wickham, H. *tidyverse: Easily Install and Load the ‘Tidyverse’*. (2017).

30. Spinu, V., Grolemund, G. & Wickham, H. *lubridate: Make Dealing with Dates a Little Easier*. (2018).

31. Katoh, K. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).

32. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

33. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).

34. Anisimova, M., Gil, M., Dufayard, J.-F., Dessimoz, C. & Gascuel, O. Survey of Branch Support Methods Demonstrates Accuracy, Power, and Robustness of Fast Likelihood-based Approximation Schemes. *Syst. Biol.* **60**, 685–699 (2011).

35. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).

36. Rambaut, A., Lam, T. T., Max Carvalho, L. & Pybus, O. G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* **2**, (2016).

37. Paradis, E. *et al.* *ape: Analyses of Phylogenetics and Evolution*. (2019).

38. Jombart, T., Dray, S. & Bilgrau, A. E. *adephylo: Exploratory Analyses for the Phylogenetic Comparative Method*. (2017).

39. Drummond, A. J. & Bouckaert, R. R. *Bayesian Evolutionary Analysis with BEAST*. (Cambridge University Press, 2015). doi:10.1017/CBO9781139095112.

40. Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. Relaxed Phylogenetics and Dating with Confidence. *PLOS Biol.* **4**, e88 (2006).

41. Kingman, J. F. C. The coalescent. *Stoch. Process. Their Appl.* **13**, 235–248 (1982).

42. Griffiths, R. C. & Tavare, S. Sampling Theory for Neutral Alleles in a Varying Environment. *Philos. Trans. Biol. Sci.* **344**, 403–410 (1994).

43. Drummond, A. J., Rambaut, A., Shapiro, B. & Pybus, O. G. Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences. *Mol. Biol. Evol.* **22**, 1185–1192 (2005).

44. Baele, G. *et al.* Improving the Accuracy of Demographic and Molecular Clock Model Comparison While Accommodating Phylogenetic Uncertainty. *Mol. Biol. Evol.* **29**, 2157–2167 (2012).

45. Baele, G., Li, W. L. S., Drummond, A. J., Suchard, M. A. & Lemey, P. Accurate Model Selection of Relaxed Molecular Clocks in Bayesian Phylogenetics. *Mol. Biol. Evol.* **30**, 239–243 (2013).