# The B-55’s

## Abstract

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

## Introduction

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. Symptoms range from acute respiratory disease to organ failure, depending on the viral species and host immune strength. Accordingly, individuals with developing or weakened immune systems account for most outbreak deaths.

## 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 (Kans, 2019). 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 (Romiti and Cooper, 2011). 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 (Wickham, 2017). Another R script relied on the \*\*lubridate\*\* package to automatically convert the date field values into a consistent ISO-8601 format (Spinu et al., 2018). This series of commands generated a sorted tab-separated file, mapping accessions to collection dates.

### Sequence Extraction

The full genome extraction method involved a simple 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 (Katoh, 2002). 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 (Nguyen et al., 2015). 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 (Kalyaanamoorthy et al., 2017). 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 (Anisimova et al., 2011) and branch support (Hoang et al., 2018). 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

Testing the strict molecular clock hypothesis on the genome, fiber, and hexon nucleotide sequences of each species required the TempEst (Rambaut et al., 2016) and BEAST (Drummond and Bouckaert, 2015). The programs evaluated the clock signal and estimated model parameters respectively. A filtering step removed sequences lacking the collection\_date metadata.

The TempEst program tests the strength of the strict molecular clock hypothesis for a given phylogeny. It plots the taxon date against the root-to-tip patristic distance and fits 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. Model parameters are only useful for data exploration since the variables are dependent. Development of 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 (Paradis et al., 2019) and adephylo (Jombart et al., 2017) packages.

The BEAUti program is a graphical tool that outputs BEAST model parameters as XML files.

## References

* Anisimova, M., Gil, M., Dufayard, J.-F., Dessimoz, C., and Gascuel, O. (2011). Survey of Branch Support Methods Demonstrates Accuracy, Power, and Robustness of Fast Likelihood-based Approximation Schemes. Syst. Biol. *60*, 685–699.
* Drummond, A.J., and Bouckaert, R.R. (2015). Bayesian Evolutionary Analysis with BEAST (Cambridge: Cambridge University Press).
* Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., and Vinh, L.S. (2018). UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol. Biol. Evol. *35*, 518–522.
* Jombart, T., Dray, S., and Bilgrau, A.E. (2017). adephylo: Exploratory Analyses for the Phylogenetic Comparative Method.
* Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., and Jermiin, L.S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods *14*, 587–589.
* Kans, J. (2019). Entrez Direct: E-utilities on the UNIX Command Line (National Center for Biotechnology Information (US)).
* Katoh, K. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. *30*, 3059–3066.
* Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol. Biol. Evol. *32*, 268–274.
* Paradis, E., Blomberg, S., Bolker, B., Brown, J., Claude, J., Cuong, H.S., Desper, R., Didier, G., Durand, B., Dutheil, J., et al. (2019). ape: Analyses of Phylogenetics and Evolution.
* Rambaut, A., Lam, T.T., Max Carvalho, L., and Pybus, O.G. (2016). Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol. *2*.
* Romiti, M., and Cooper, P. (2011). Search Field Descriptions for Sequence Database (National Center for Biotechnology Information (US)).
* Spinu, V., Grolemund, G., and Wickham, H. (2018). lubridate: Make Dealing with Dates a Little Easier.
* Wickham, H. (2017). tidyverse: Easily Install and Load the “Tidyverse.”