Detection of Engineered *E. coli* Reference Genomes within NCBI Using Molecular Clock Analysis

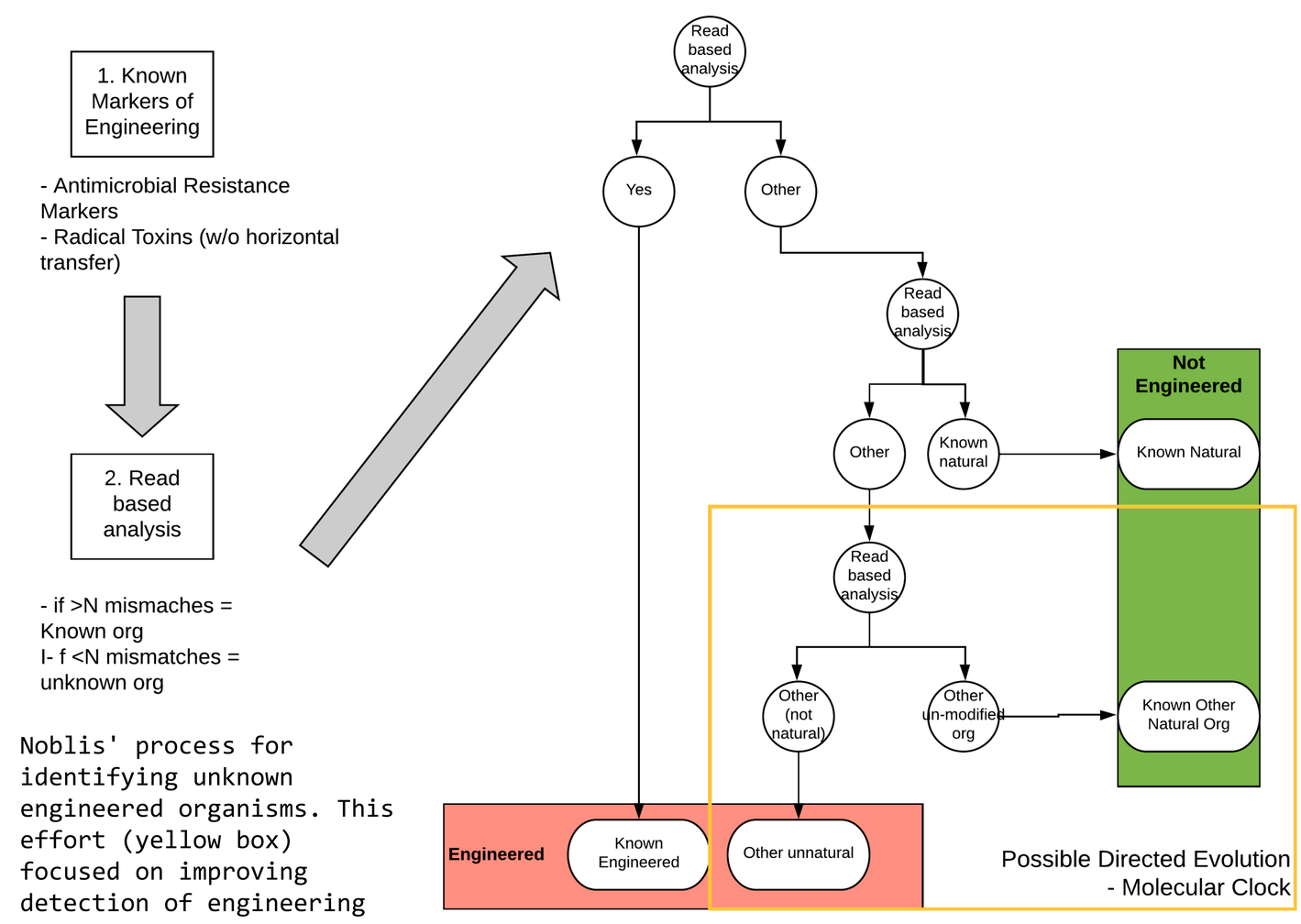
Sterling Thomas, PhD; Tyler Barrus; Daniel Antonio Negrón, PhD-c

noblis®
For the best of reasons

[Introduction]

Noblis, Inc. is developing a method to detect artificial sequences. The following approach employs molecular clock analysis to build a model based on sequence evolution. The hypothesis is that sequences with significantly different mutation rates may contain engineering. This work aims to quickly identify candidates for subsequent detection of potential engineering, using the MAFFT, Gubbins, and BactDating programs to enable scalable, recombination-aware inference of phylogeny and clock-like behavior. These preliminary results indicate the presence of a clock signal and outliers across DNA Pol II sequences of *E. coli*. Planned work includes testing using altered sequences and integration into the directed evolution machine learning program.

[Methods]

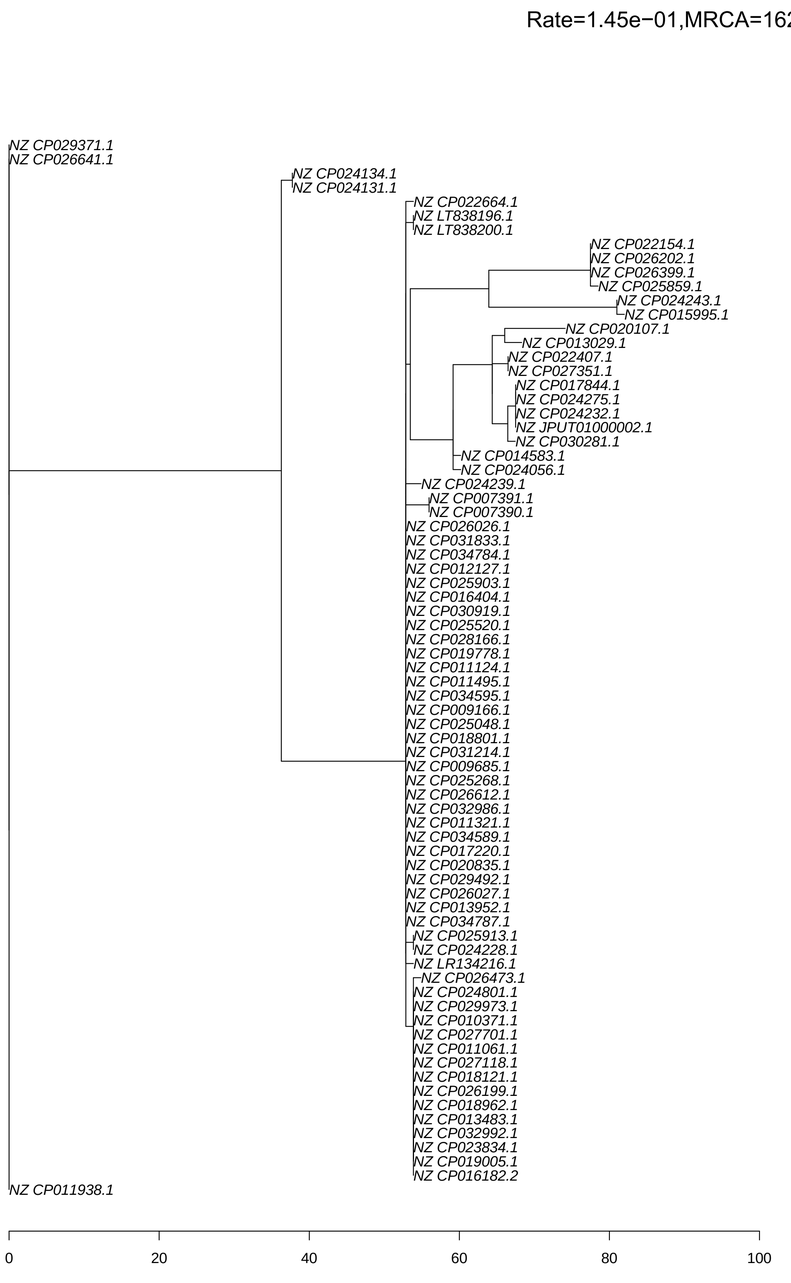


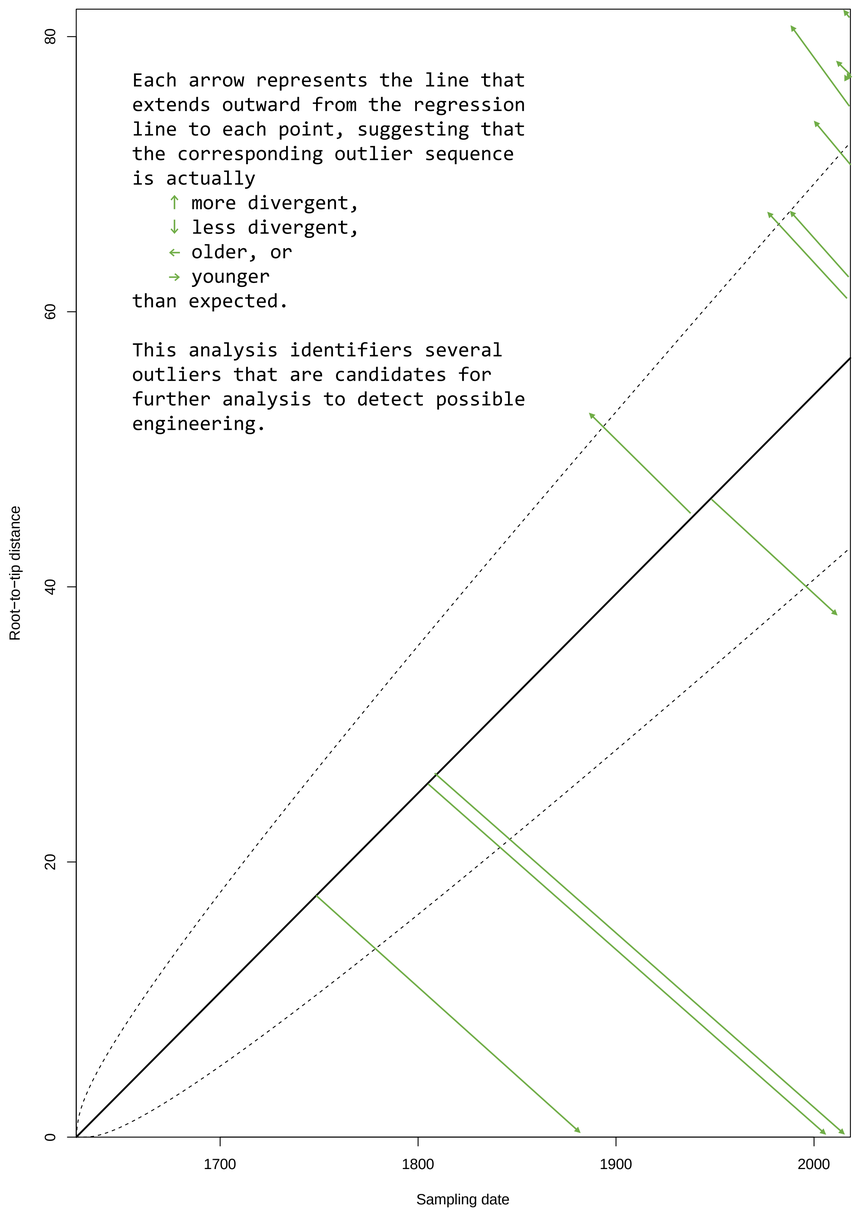
Noblis' process for identifying unknown engineered organisms. This effort (yellow box) focused on improving detection of engineering via directed evolution or lab adaptation.

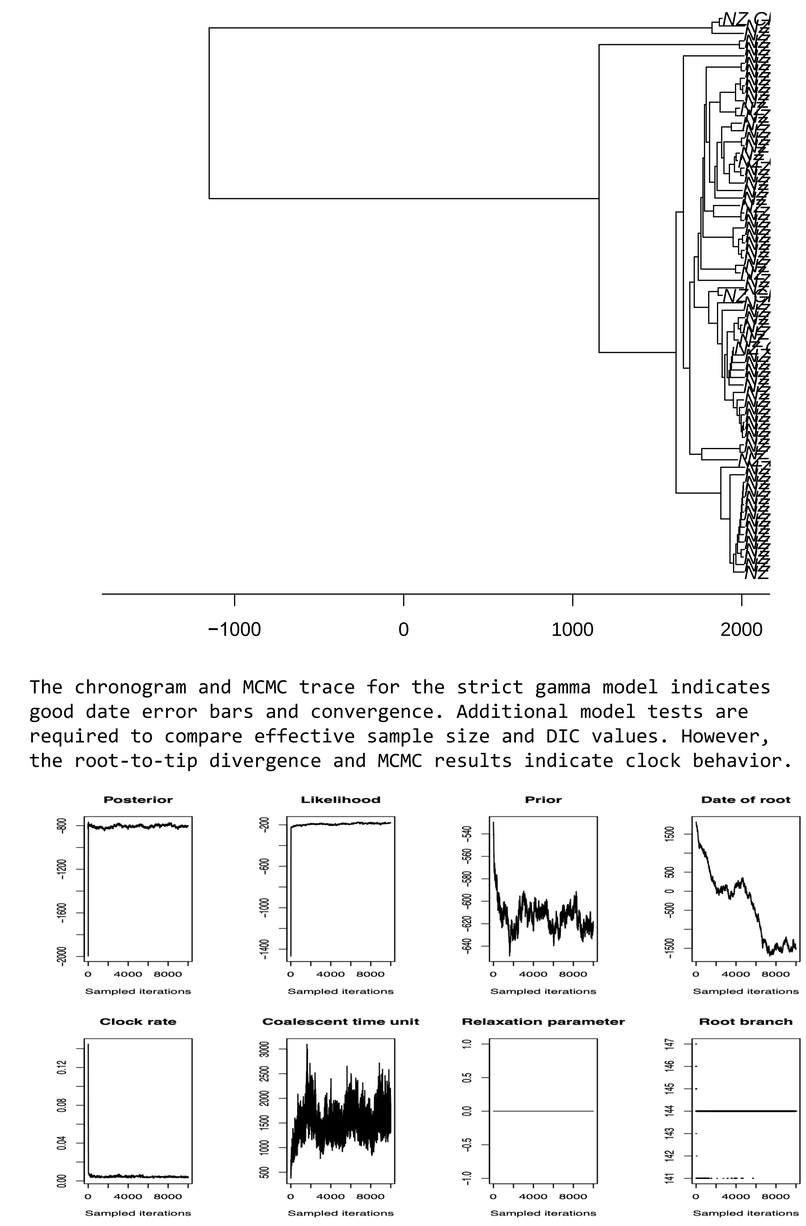
Molecular Clock

1. Download GenBank files
2. Extract collection date metadata and normalize to consistent decimal format
3. Perform multiple alignment using MAFFT
4. Infer recombination-aware maximum-likelihood phylogeny with Gubbins
5. Analyze root-to-tip correlation of observed divergence and collection date
6. Infer molecular clock using BactDating
7. Evaluate MCMC simulation
8. Identify outlier taxa to evaluate for possible engineering

[Results]







The chronogram and MCMC trace for the strict gamma model indicates good date error bars and convergence. Additional model tests are required to compare effective sample size and DIC values. However, the root-to-tip divergence and MCMC results indicate clock behavior.

[References]

- Katoh, K. “MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform.” *Nucleic Acids Research* 30, no. 14 (July 15, 2002): 3059–66. https://doi.org/10.1093/nar/gkf436.
- Croucher, Nicholas J., Andrew J. Page, Thomas R. Connor, Aidan J. Delaney, Jacqueline A. Keane, Stephen D. Bentley, Julian Parkhill, and Simon R. Harris. “Rapid Phylogenetic Analysis of Large Samples of Recombinant Bacterial Whole Genome Sequences Using Gubbins.” *Nucleic Acids Research* 43, no. 3 (February 18, 2015): e15–e15. https://doi.org/10.1093/nar/gku1196.
- R Core Team, R: A Language and Environment for Statistical Computing.
- Didelot, Xavier, Nicholas J Croucher, Stephen D Bentley, Simon R Harris, and Daniel J Wilson. “Bayesian Inference of Ancestral Dates on Bacterial Phylogenetic Trees.” Nucleic Acids Research 46, no. 22 (December 14, 2018): e134–e134. https://doi.org/10.1093/nar/gky783.
- Rambaut, Andrew, Tommy T. Lam, Luiz Max Carvalho, and Oliver G. Pybus. “Exploring the Temporal Structure of Heterochronous Sequences Using TempEst (Formerly Path-O-Gen).” Virus Evolution 2, no. 1 (April 9, 2016). https://doi.org/10.1093/ve/vew007.
- Clark, Karen, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and Eric W. Sayers. “GenBank.” Nucleic Acids Research 44, no. D1 (January 4, 2016): D67–72. https://doi.org/10.1093/nar/gkv1276.

[Contact]

https://noblis.org/bioportal/workwithus.html