Introduction to Model Selection

2023 Woods Hole Molecular Evolution Workshop

David Swofford

Florida Museum of Natural History
University of Florida
davidswofford@ufl.edu

What is a (statistical) model?

Daniel L. Hartl, 2000:

A **model** is an intentional simplification of a complex situation designed to eliminate extraneous detail in order to focus attention on the essentials of the situation.

Wikipedia 27 May 2022:

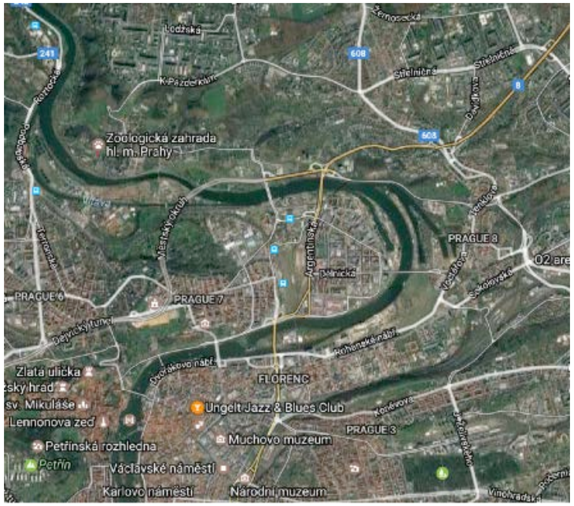
A **statistical model** is a mathematical model that embodies a set of statistical assumptions concerning the generation of sample data (and similar data from a larger population). A statistical model represents, often in considerably idealized form, the data-generating process.

Jordan Peterson

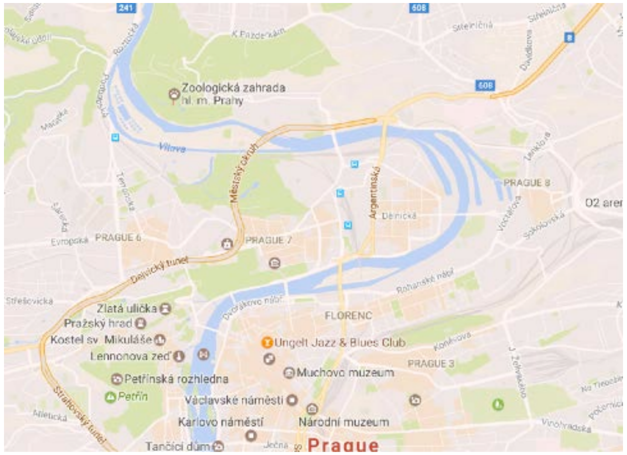
Jordan Bernt Peterson is a Canadian clinical psychologist, YouTube personality, author, and a professor emeritus at the University of Toronto. Peterson began to receive widespread attention as a public intellectual in the late 2010s for his views on cultural and political issues, often described as conservative. Wikipedia

Peterson on models

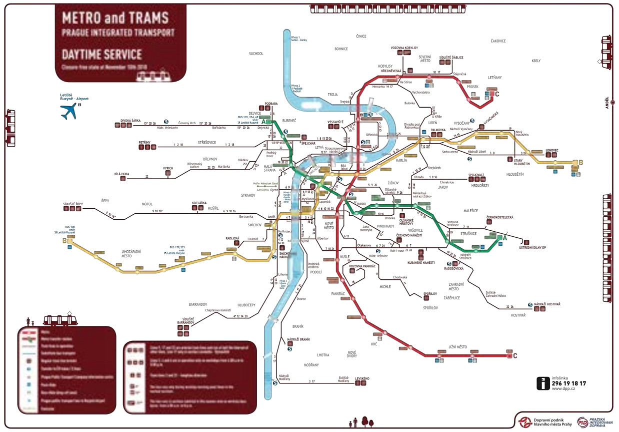
Which is more useful?



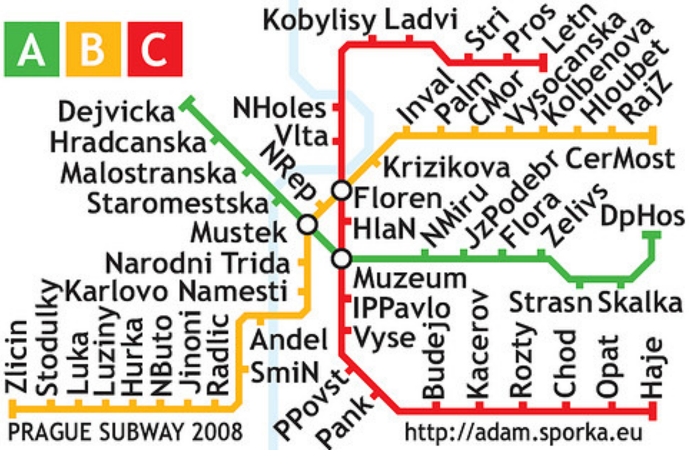
"Reality"



Detailed map



Detailed public transportation



Simplified metro

Concept credit: Paul Lewis

Models don’t need to reflect reality

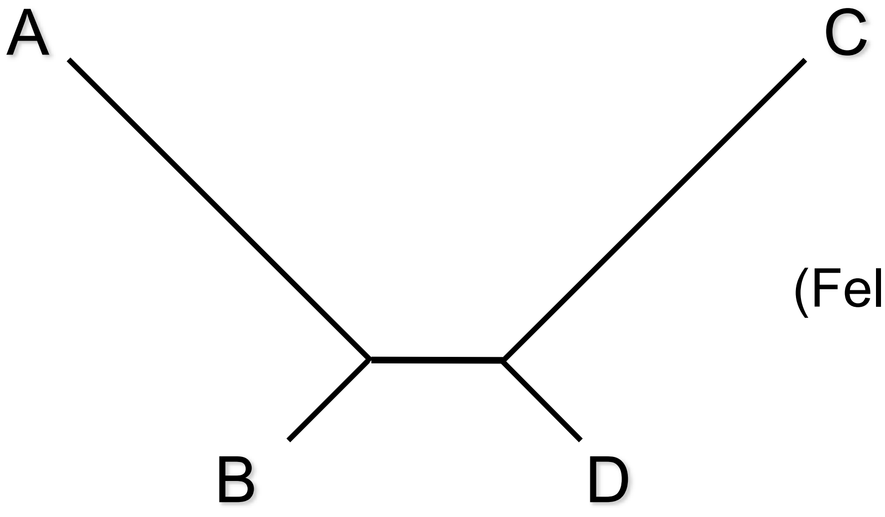
"The most that can be expected from any model is that it can supply a useful approximation to reality: **All models are wrong; some models are useful"**. (George E. P. Box, 1987)

Model selection is a process of seeking the least inadequate model from a predefined set, all of which may be grossly inadequate as a representation of reality. (J. J. Welch, 2006)

Why do models matter?

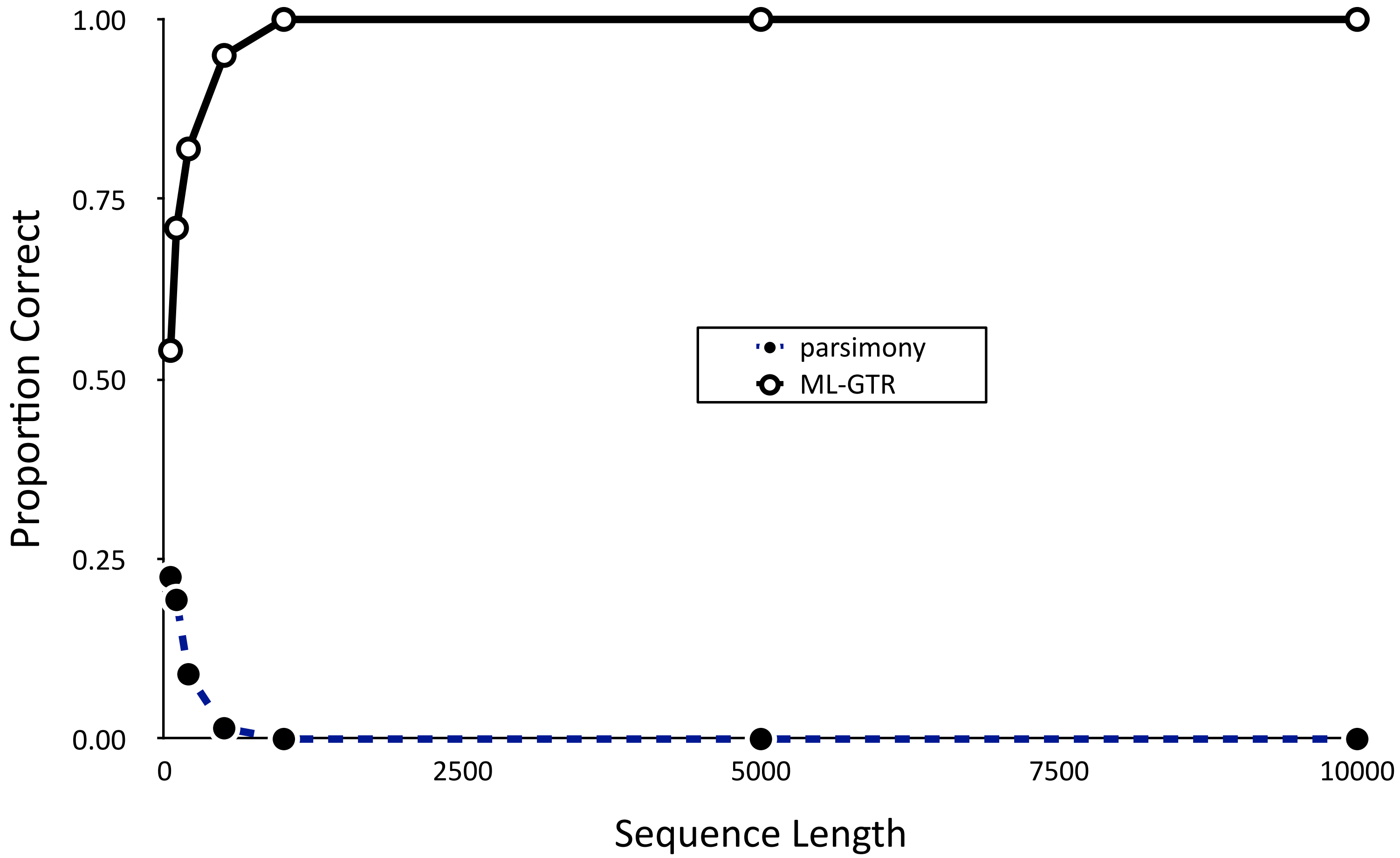
Model-based methods including ML and Bayesian inference (typically) make a *consistent* estimate of the phylogeny (estimate converges to true tree as number of sites increases toward infinity)

... even when you're in the “Felsenstein Zone”



(Felsenstein, 1978)

In the Felsenstein Zone



Simulation model = GTR

Why do models matter (continued)?

- Parsimony is inconsistent in the Felsenstein zone (and other scenarios)
- Likelihood is consistent in any “zone” (when certain requirements are met)

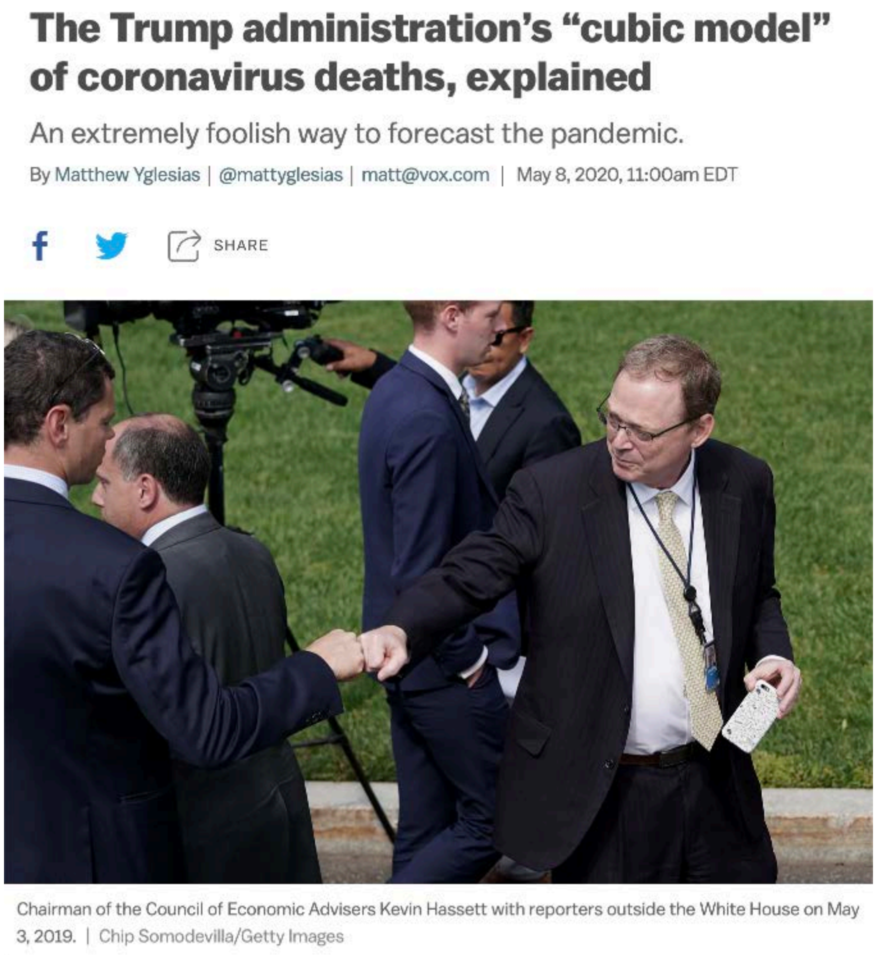
But this guarantee requires that the model be specified correctly!

Likelihood can also be inconsistent if the model is oversimplified

- Real data always evolve according to processes more complex than any computationally feasible model would permit, so we have to choose “good” rather than “correct” models

What is a “good” model?

Parsimony in statistics represents a tradeoff between bias and variance as a function of the dimension of the model. A good model is a balance between under- and over-fitting. (Burnham and Anderson, 1998)



The Trump administration’s “cubic model” of coronavirus deaths, explained

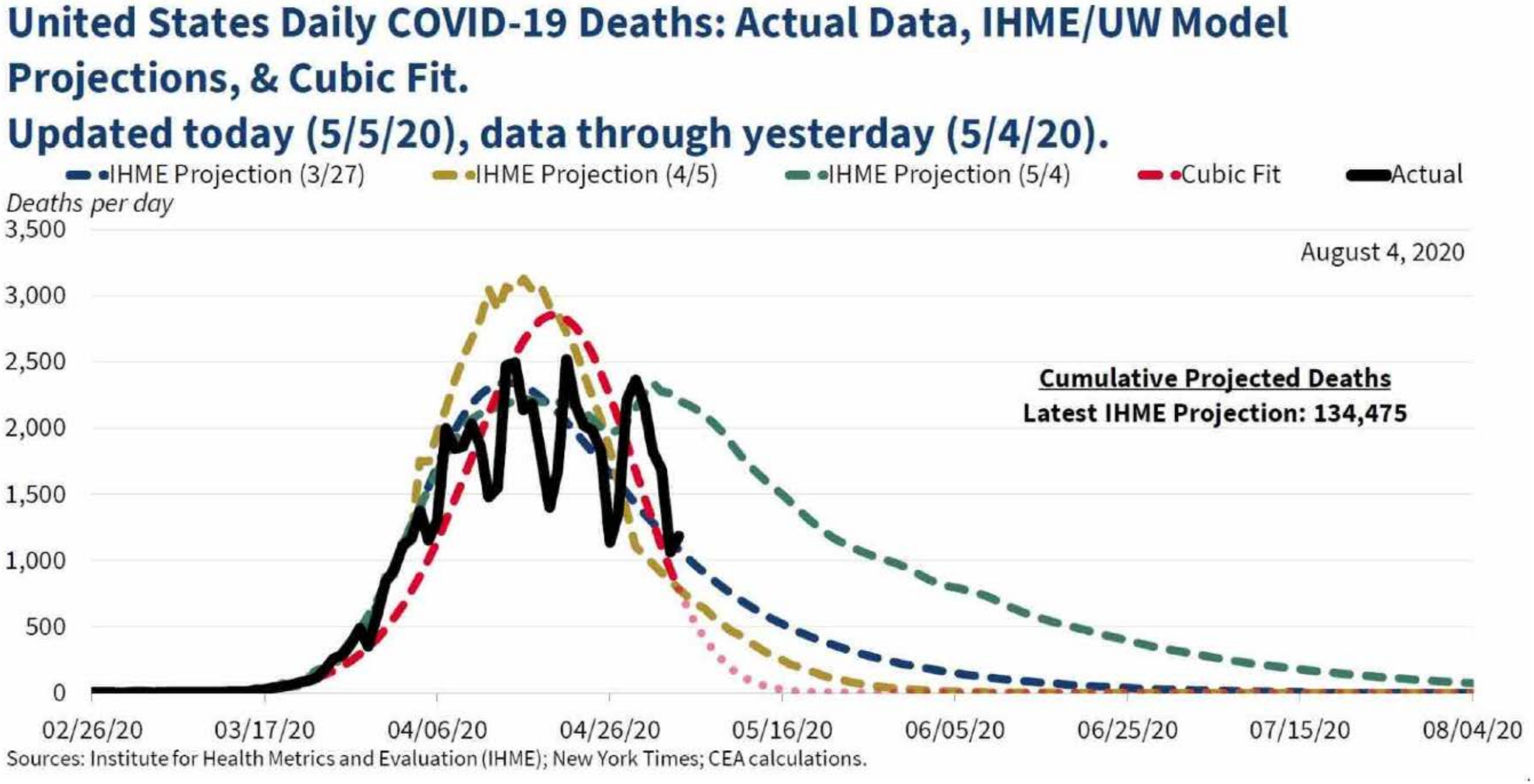
An extremely foolish way to forecast the pandemic.

By Matthew Yglesias | @mattyglesias | matt@vox.com | May 8, 2020, 11:00am EDT

Chairman of the Council of Economic Advisers Kevin Hassett with reporters outside the White House on May 3, 2019. | Chip Somodevilla/Getty Images

https://www.vox.com/2020/5/8/21250641/kevin-hassett-cubic-model-smoothing

Using curve fitting to predict COVID-19 deaths



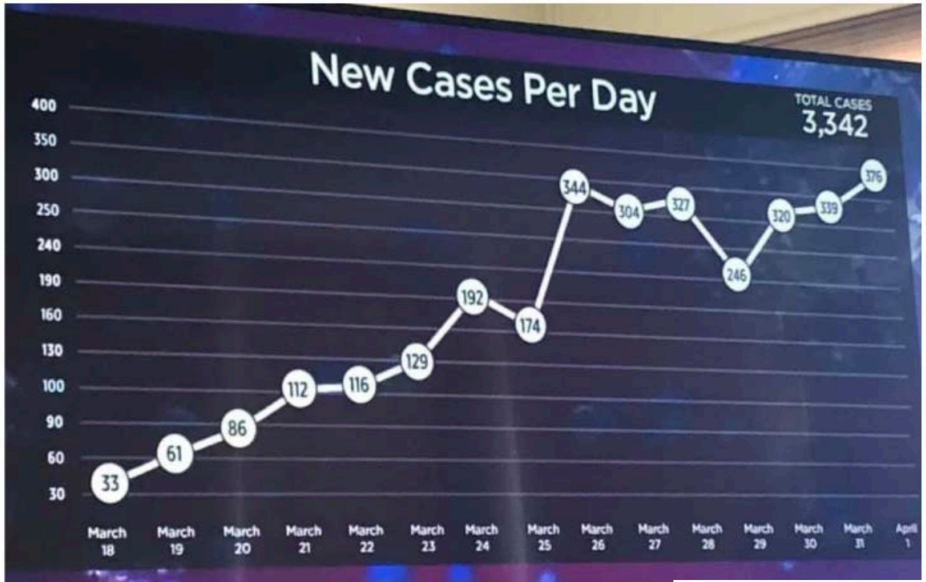
United States Daily COVID-19 Deaths: Actual Data, IHME/UW Model Projections, & Cubic Fit.
Updated today (5/5/20), data through yesterday (5/4/20).

Sources: Institute for Health Metrics and Evaluation (IHME); New York Times; CEA calculations.

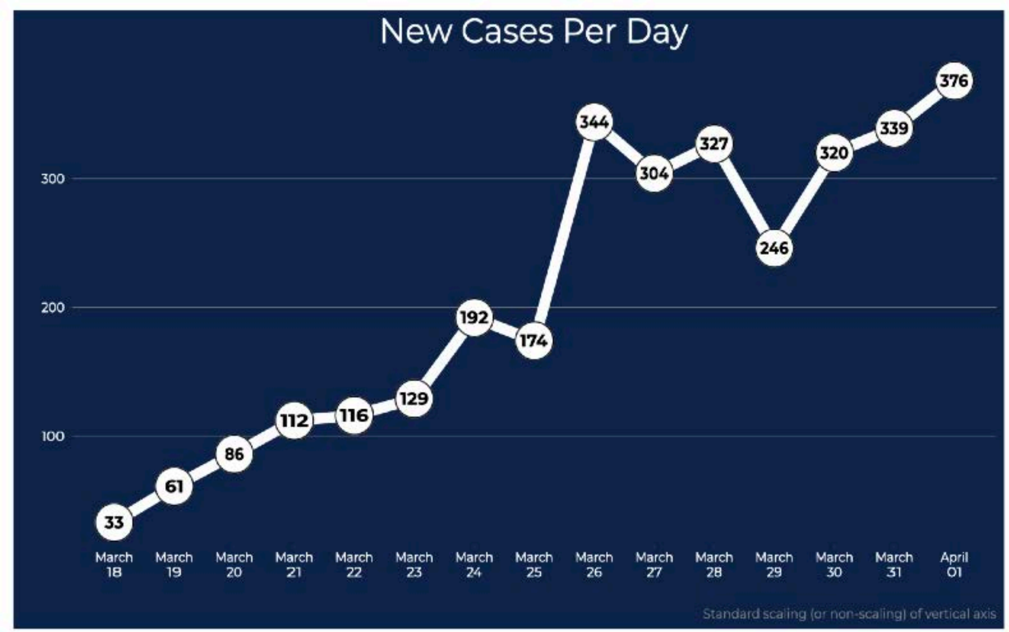
From "free range statistics" blog

(Peter Ellis)

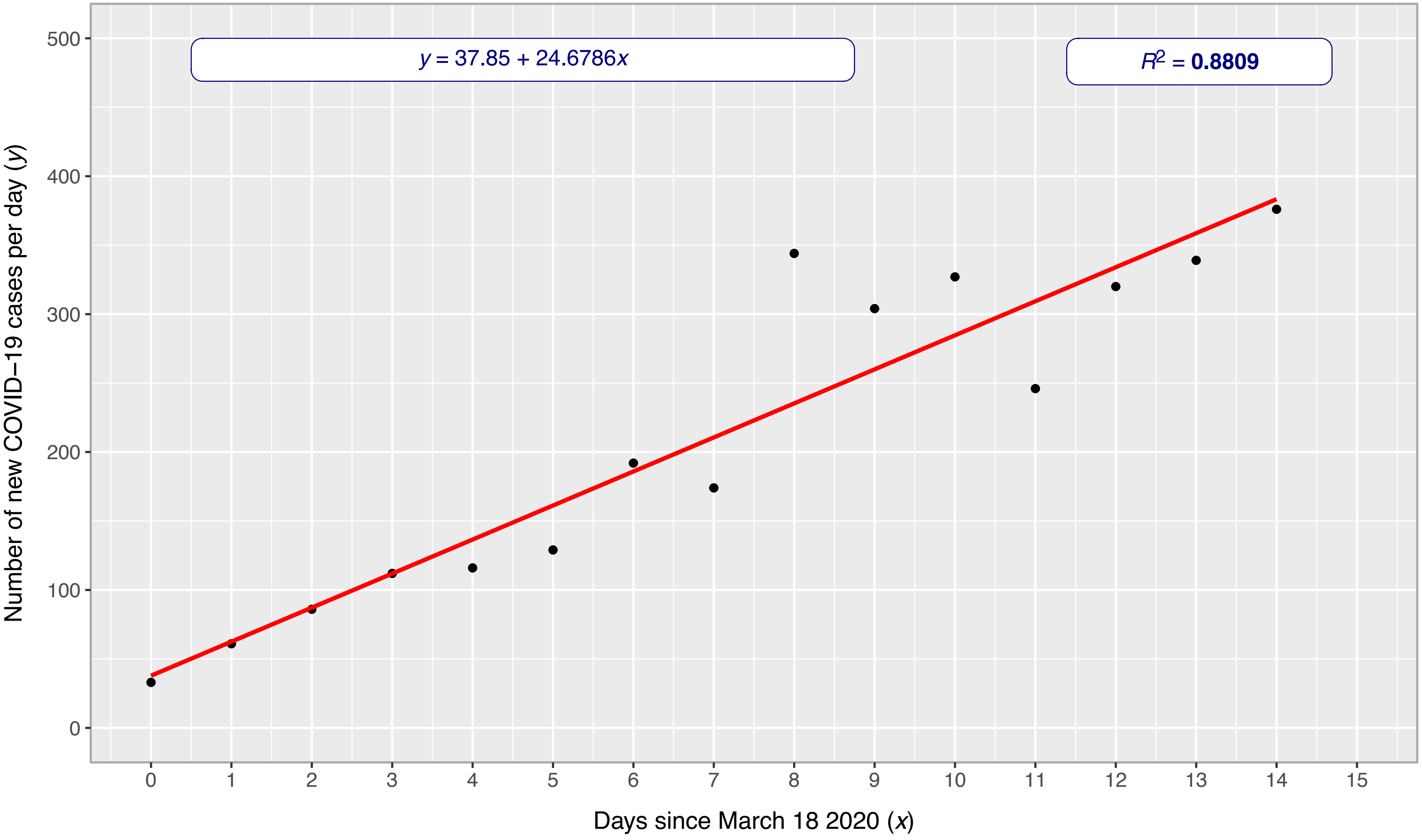
“It’s so bad it’s funny. This is clearly incompetence not malevolence. But it’s a serious degree of incompetence.”



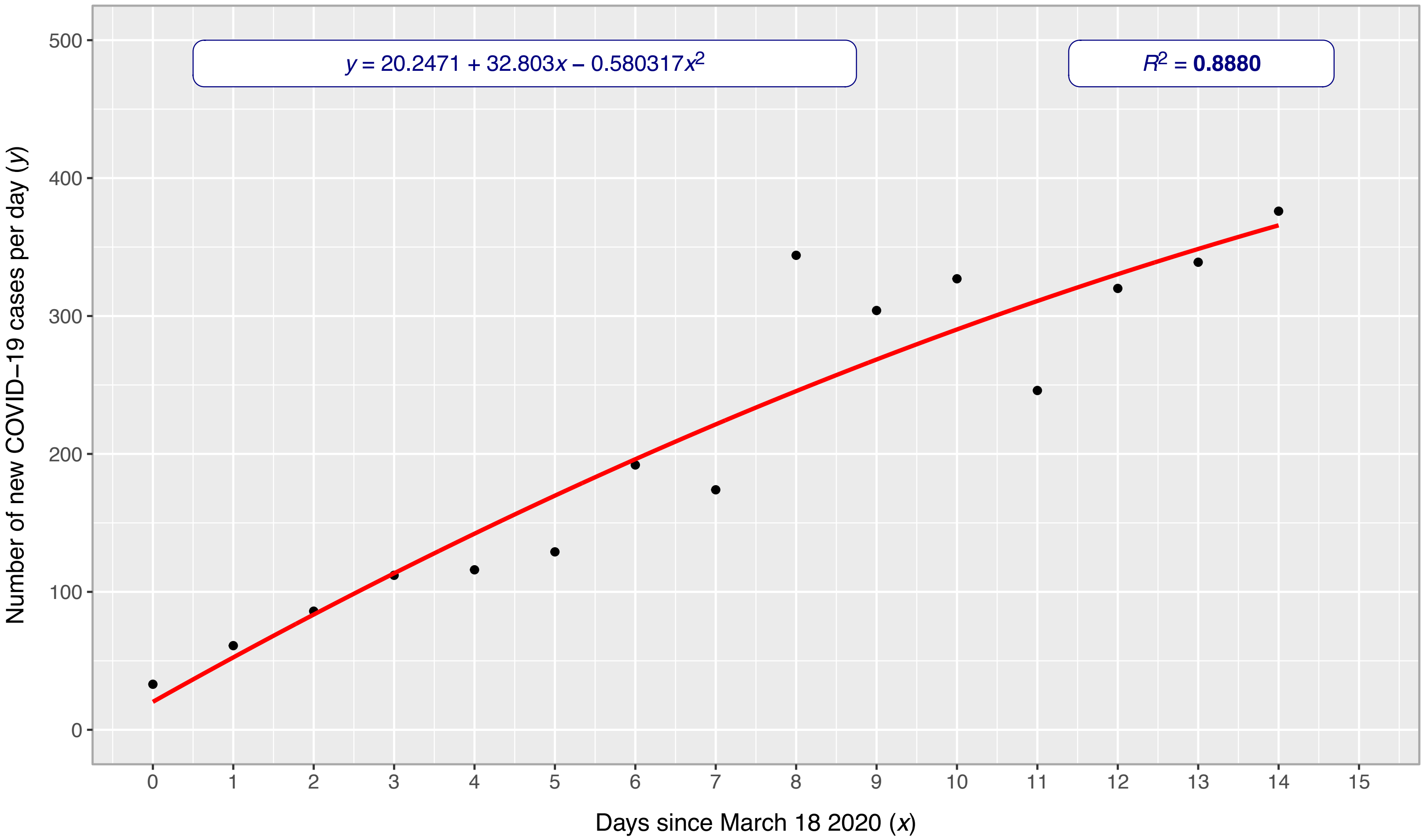
After fix-up



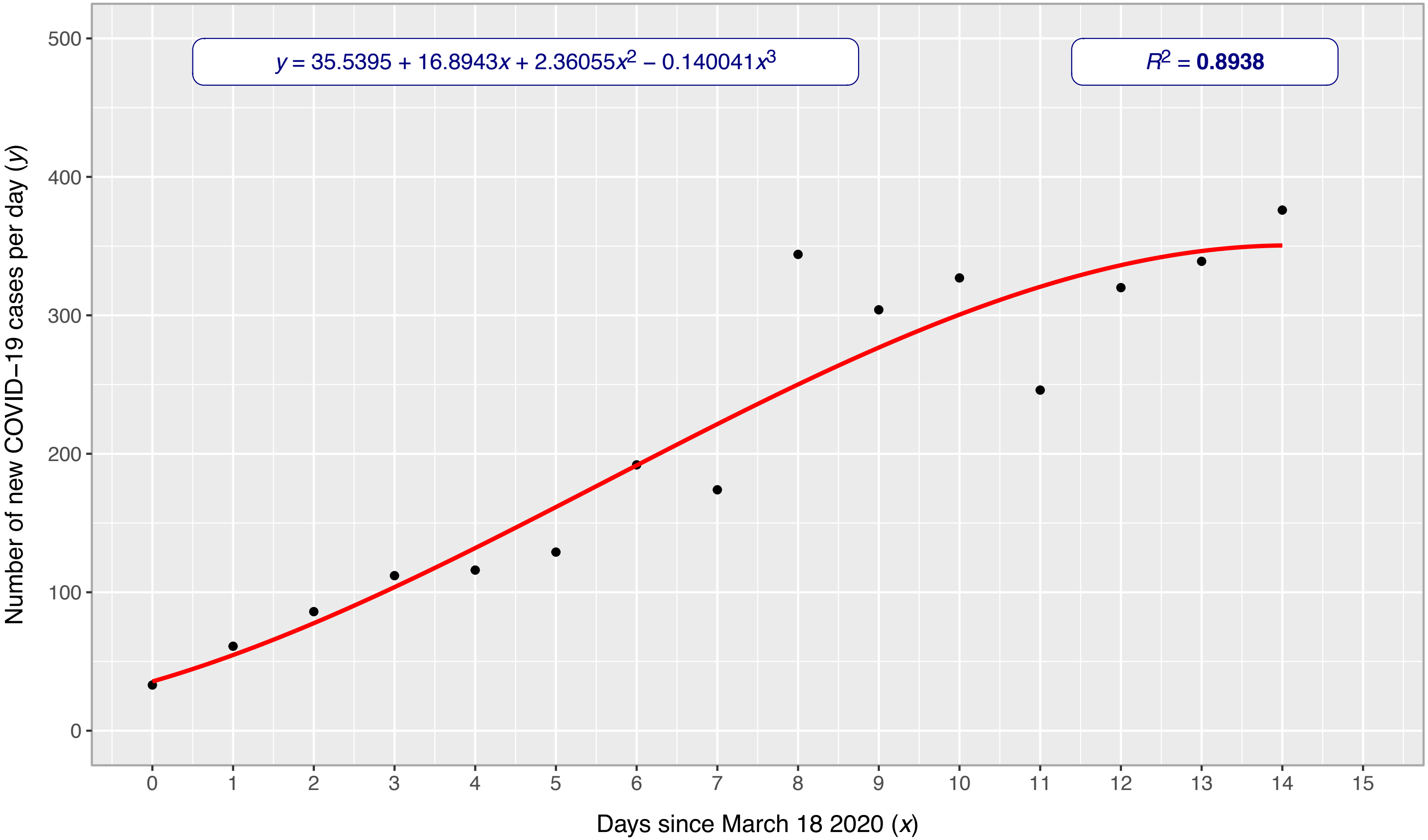
Simple linear regression



Quadratic model

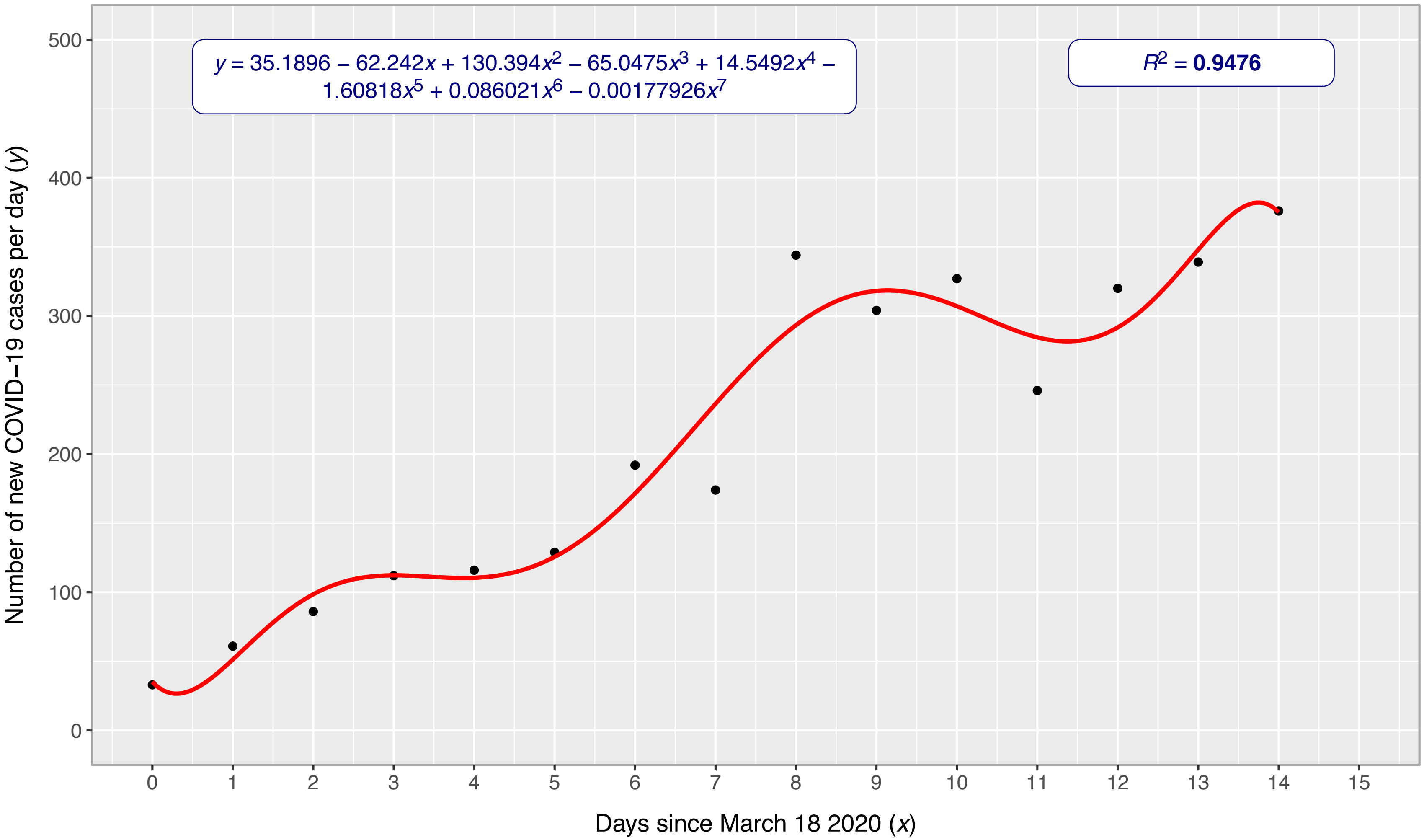


Cubic model

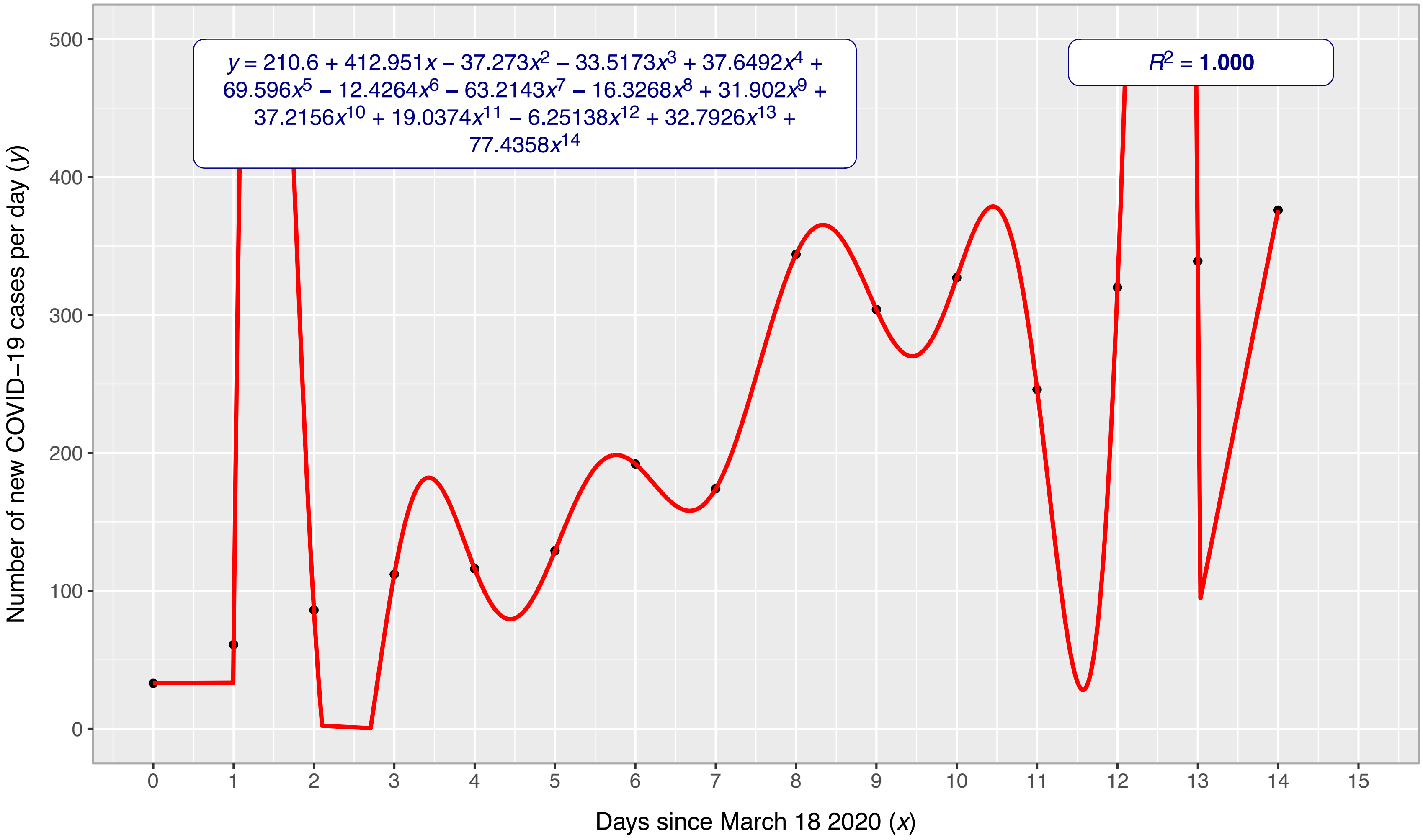


$y = 35.5395 + 16.8943x + 2.36055x^2 - 0.140041x^3$
R^2 = **0.8938**
Number of new COVID−19 cases per day (*y*)
500
400
300
200
100
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Days since March 18 2020 (*x*)

7th order polynomial



14th order polynomial

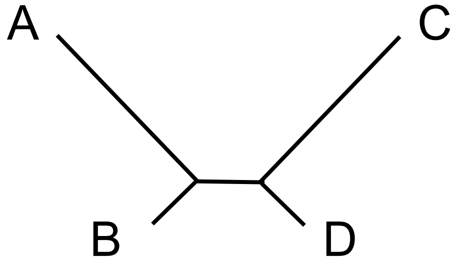


Why models don't have to be perfect

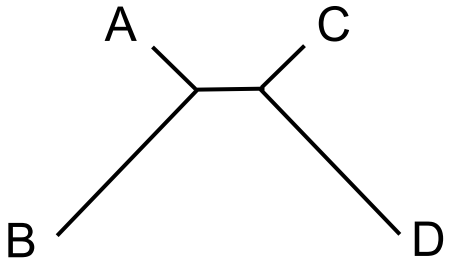
Assertion: In most situations, phylogenetic inference is relatively robust to model misspecification, *as long as critical factors influencing sequence evolution are accommodated*

Caveat: There are some kinds of model misspecification that are very difficult to overcome (e.g., “heterotachy”)

E.g.:



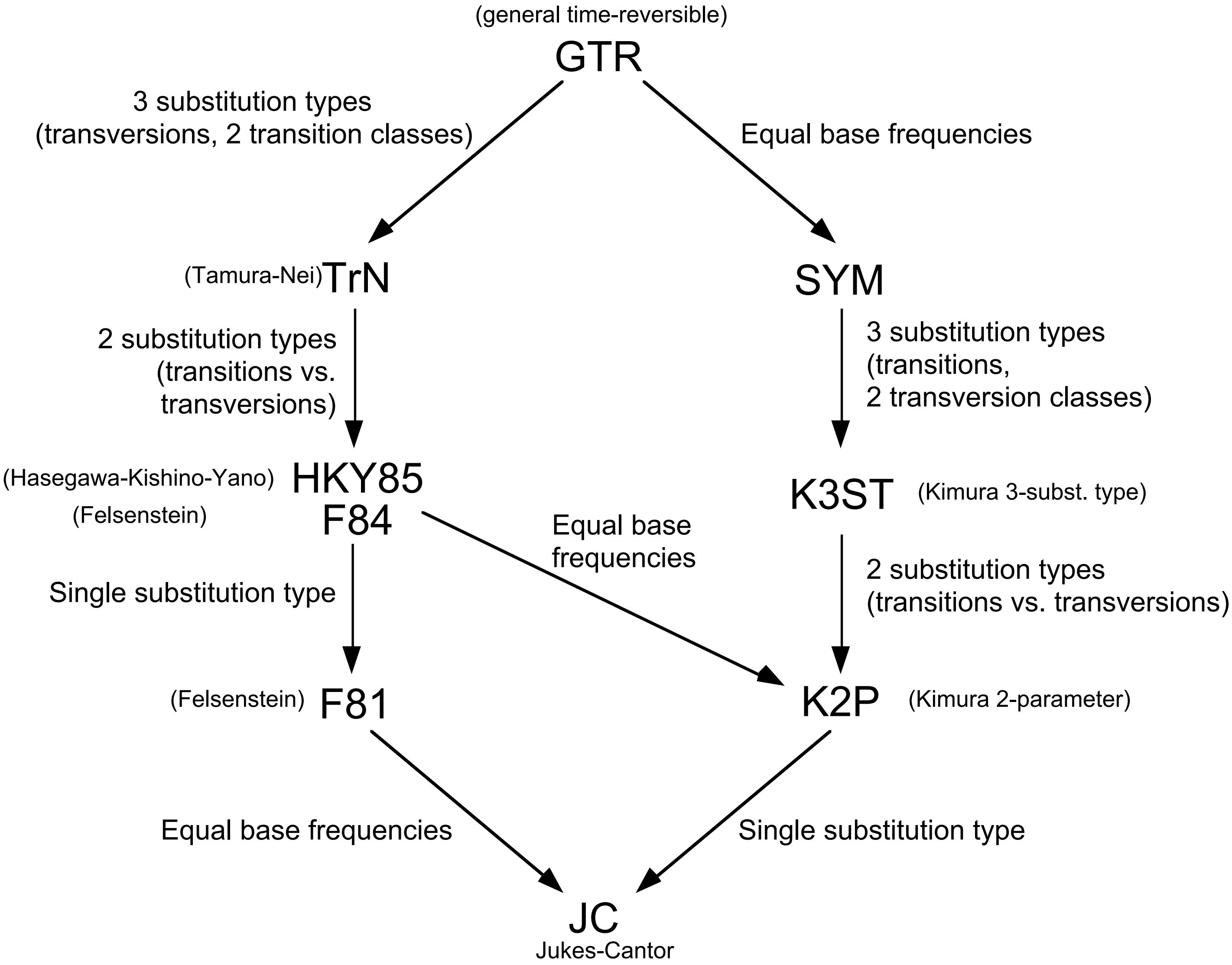
Half of sites



Other half

Likelihood can be consistent in Felsenstein zone, but will be inconsistent if a single set of branch lengths are assumed when there are actually two sets of branch lengths (Chang 1996) (“heterotachy”)

GTR Family of Reversible DNA Substitution Models



(general time-reversible)
GTR

3 substitution types
(transversions, 2 transition classes)

Equal base frequencies

(Tamura-Nei) TrN

SYM

2 substitution types
(transitions vs.
transversions)

3 substitution types
(transitions,
2 transversion classes)

(Hasegawa-Kishino-Yano) HKY85
(Felsenstein) F84

K3ST (Kimura 3-subst. type)

Equal base
frequencies

Single substitution type

2 substitution types
(transitions vs. transversions)

(Felsenstein) F81

K2P (Kimura 2-parameter)

Equal base frequencies

Single substitution type

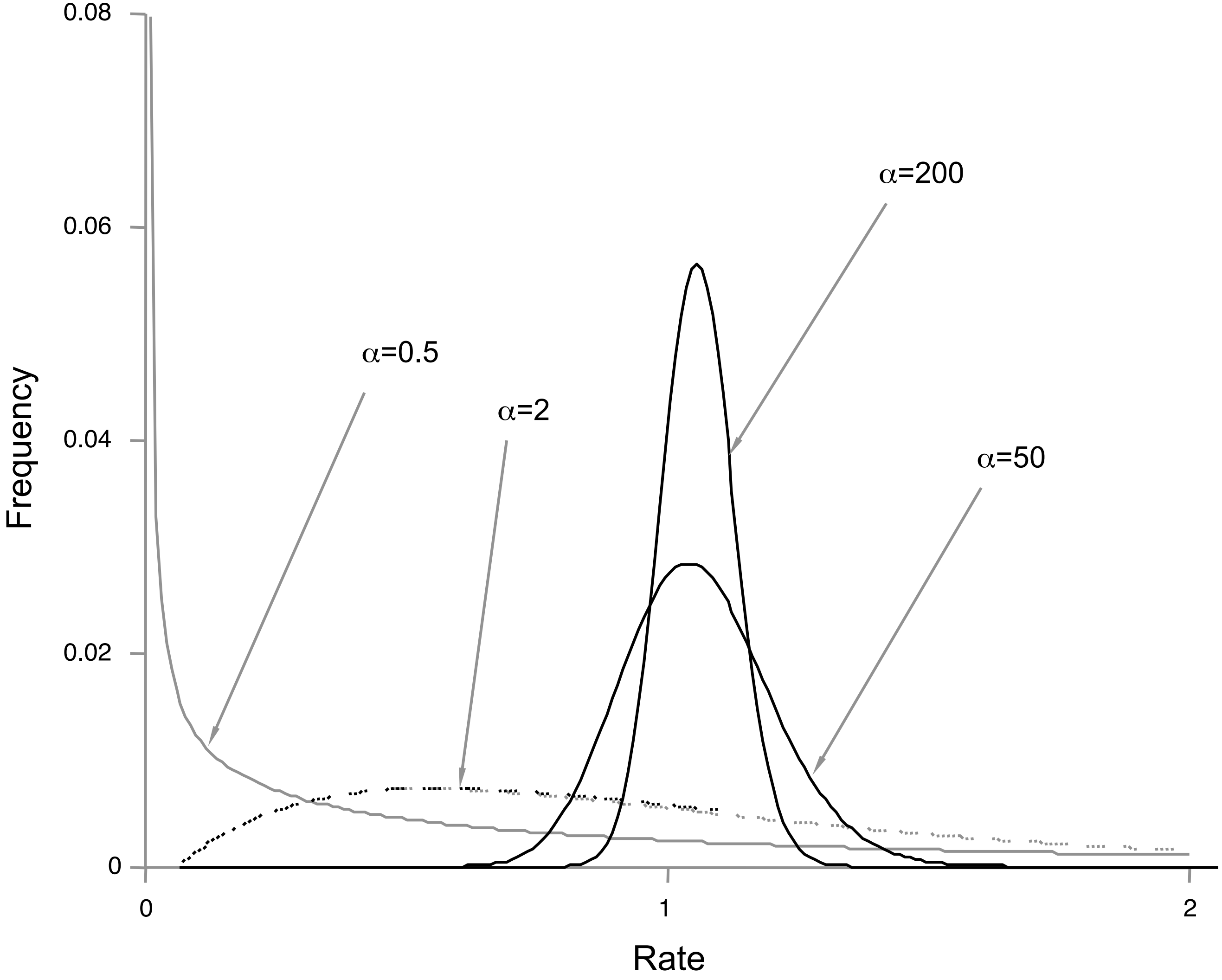
JC
Jukes-Cantor

Modeling among-site rate heterogeneity

```
Lemur AAGCTTCATAG TTGCATCATCCA …TTACATCATCCA
Homo  AAGCTTCACCG TTGCATCATCCA …TTACATCCTCAT
Pan   AAGCTTCACCG TTACGCCATCCA …TTACATCCTCAT
Goril AAGCTTCACCG TTACGCCATCCA …CCCACGGACTTA
Pongo AAGCTTCACCG TTACGCCATCCT …GCAACCACCCTC
Hylo  AAGCTTTACAG TTACATTATCCG …TGCAACCGTCCT
Maca  AAGCTTTTCCG TTACATTATCCG …CGCAACCATCCT
```

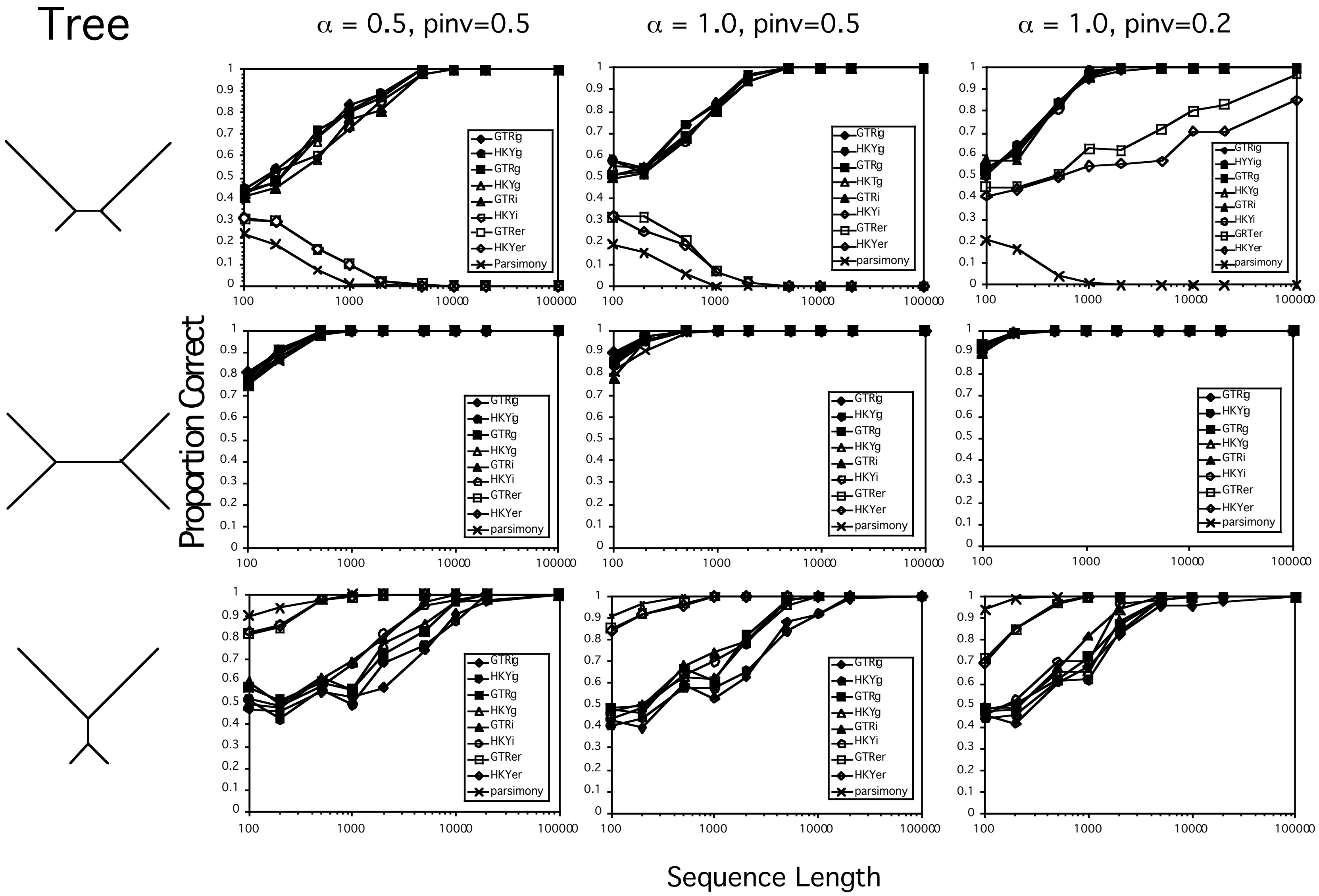
- Proportion of invariable sites
  - Some sites extremely unlikely to change due to strong functional or structural constraint (Hasegawa et al., 1985)
- Gamma-distributed rates
  - Rate variation assumed to follow a gamma distribution with shape parameter α
- Site-specific rates (another way to model ASRV)

Modeling ASRV with gamma distribution



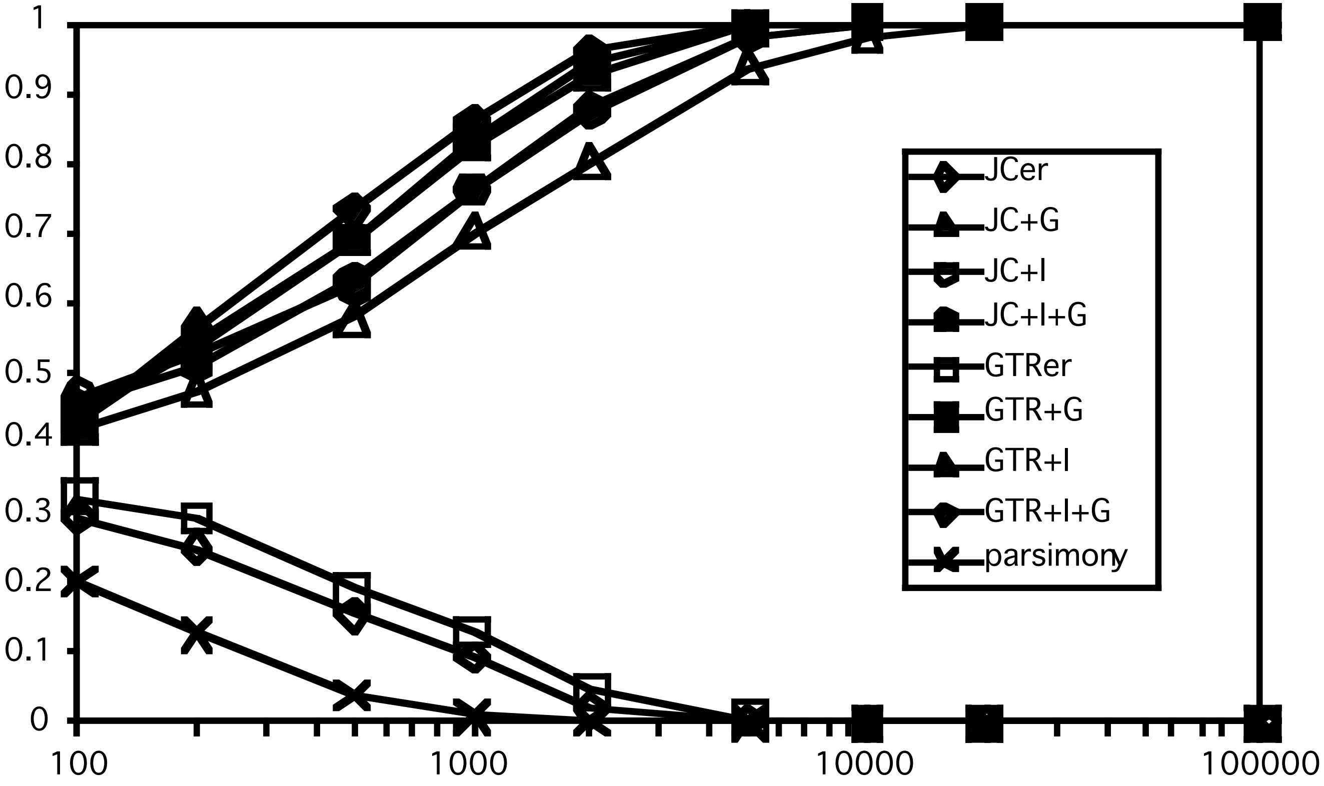
…can also include a proportion of “invariable” sites (p_{inv})

Performance of ML when its model is violated

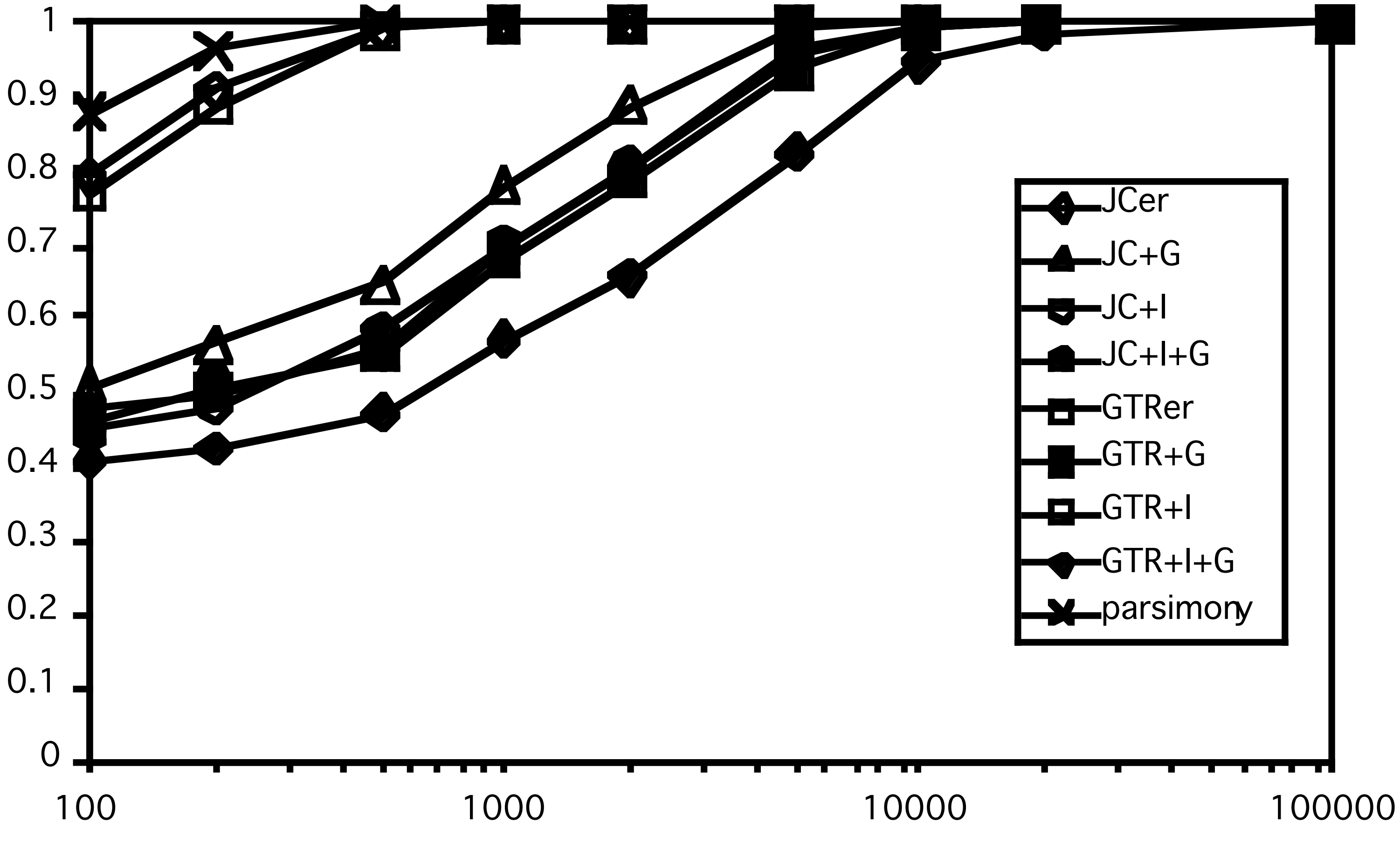


“MODERATE”–Felsenstein zone

α = 1.0, p_{inv}=0.5



“MODERATE”–Inverse-Felsenstein zone



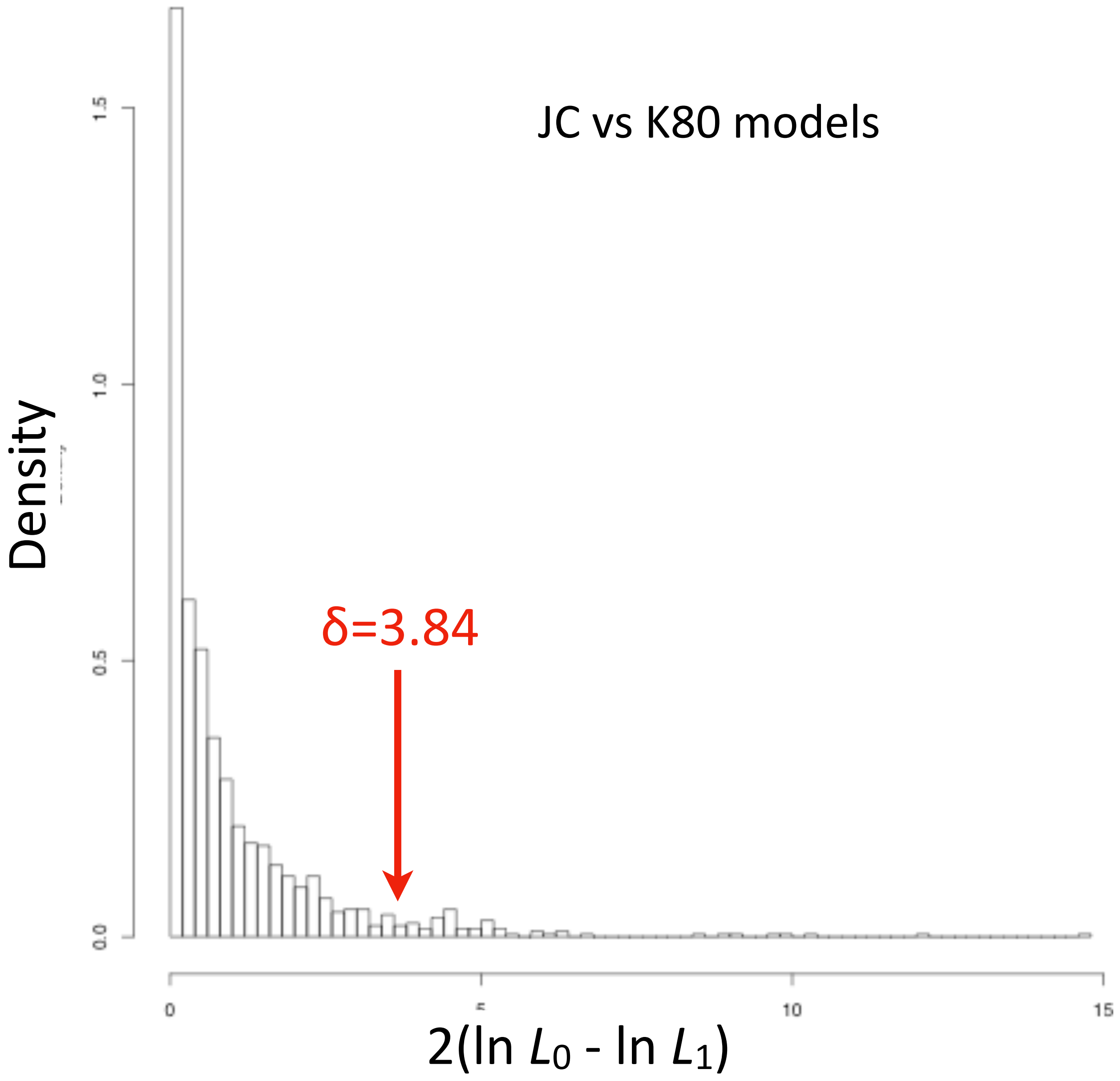
Likelihood ratio tests

- Calculate "delta" statistic

$$\delta = 2(\ln L_1 - \ln L_0)$$

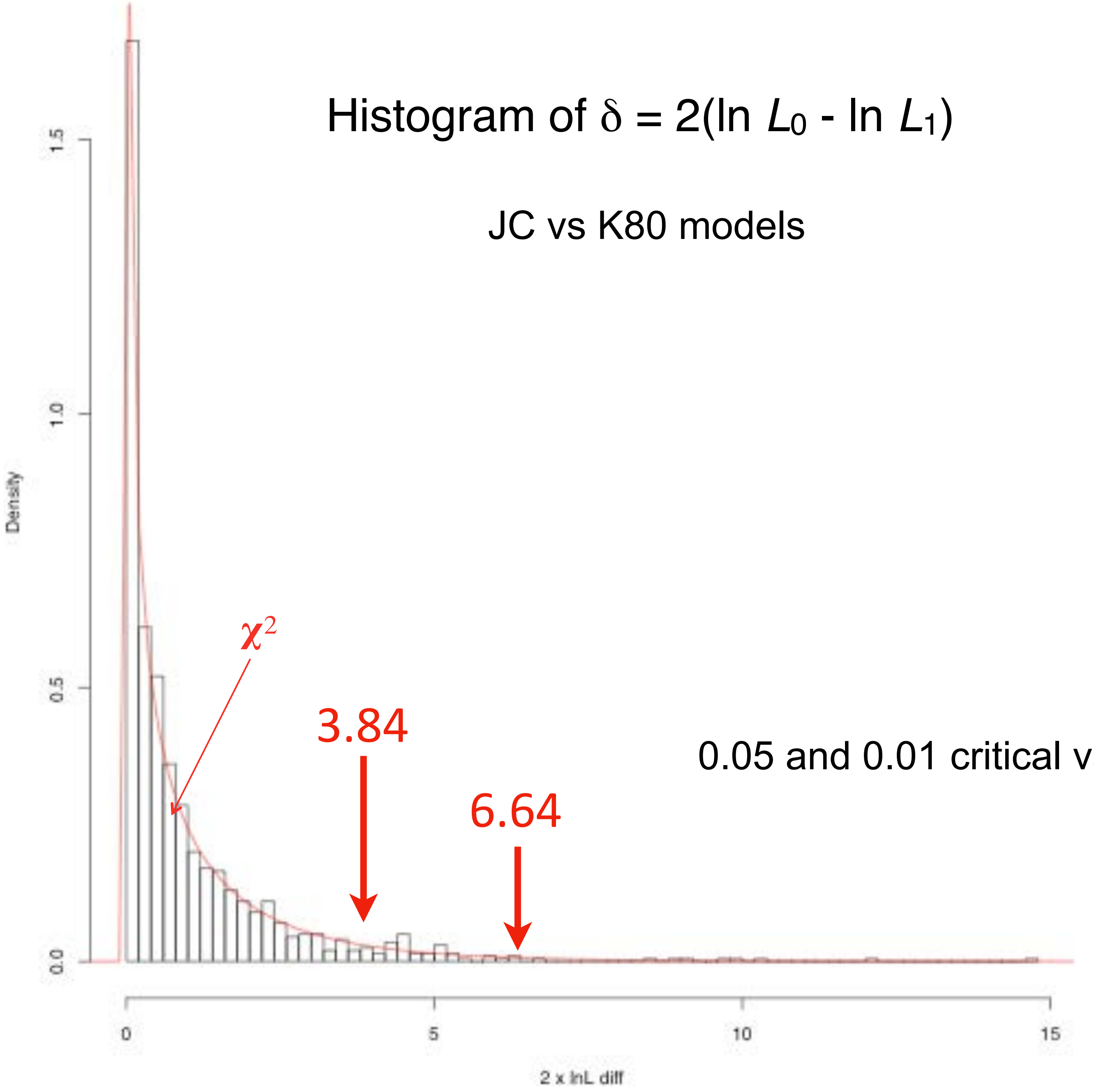
If model 0 is nested within model 1, δ is distributed as a χ^2 random variable with degrees-of-freedom equal to the difference in number of free parameters

Histogram of $\delta = 2(\ln L_0 - \ln L_1)$



JC vs K80 models

$\delta=3.84$



Histogram of $\delta = 2(\ln L_0 - \ln L_1)$

JC vs K80 models

χ^2

3.84

6.64

0.05 and 0.01 critical values

Model selection criteria

- Akaike information criterion (AIC)

$$AIC_i = -2\ln L_i + 2k$$

where *k* is the number of free parameters estimated

- AICc (corrected AIC)

$$AIC_c = AIC + \frac{(2k(k+1))}{(n-k-1)}$$

- Bayesian information criterion (BIC)

$$BIC_i = -2\ln L_i + k\ln n$$

where *k* is the number of free parameters estimated and *n* is the "sample size" (typically number of sites)

AIC(c) vs. BIC

- BIC performs well when true model is contained in model set, and among a set of simple models, AIC often selects a more complex model than the truth (indeed, AIC is formally statistically inconsistent)
- But in phylogenetics, no model is as complex as the truth, and the true model will never be contained in the model set.
- BIC often chooses models that seem *too* simple, however.

Opinion: Studies that evaluate the performance of model selection criteria based on their ability to choose the "true" model from a set of competing models are fundamentally flawed.

Partitioned Models

Many authors have emphasized the importance of modeling heterogeneity among genes or other subsets of the data appropriately

> “...data partitioning is more an art than a science, and it should rely on our knowledge of the biological system...”
>
> Yang and Rannala (2012; *Nature Rev. Genet*. 13:303-314)

Ways to partition based on biological criteria

- By gene
- By codon
- By gene/codon combination
- Stems vs. loops (probably not advisable—e.g., Simon et al., 2006)
- Coding vs. noncoding

Naive partitioning

- Run ModelTest/JModelTest; estimate a model (from the GTR+I+G family) separately for each gene/subset
- Perform an ML/Bayesian analysis, assigning the chosen models to each gene (with unlinked parameters)

Too many parameters! 1-10 parameters for each gene; amount of data available to estimate each parameter does not increase

Over-Partitioning

Consider the following (contrived) example:

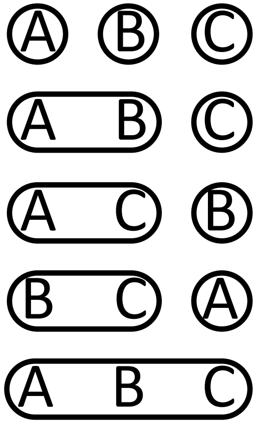
- Gene A: HKY+G, π = (0.26, 0.24, 0.23, 0.27), κ=1.1, α=3.0
- Gene B: GTR, π = (0.25,0.24,0.25,0.26), (a,b,c,d,e)=(1.1, 1.2, 0.9, 1.1, 0.95)
- Gene C: JC+I (pinv=0.05)

These are all GTR models that are not far from the Jukes-Cantor model, but they all have different names

Better to estimate one GTR model (even with 5+3+1+1=10 parameters, estimated from all data) than 3 separate models with 2+5+1=8 parameters (but only one gene's worth of data for each model)

How to find optimal partitionings?

Consider a data
set with 3 genes,
A, B, and C:



For each partitioning scheme, evaluate some set of models from the GTR+I+G (e.g., 56 models) according to AIC or BIC

Choose a combination of partitioning scheme and model for subsequent partitioned-model analyses

Rob Lanfear's **PartitionFinder** (http://www.robertlanfear.com/partitionfinder/)
automates this process; method now also available in PAUP*

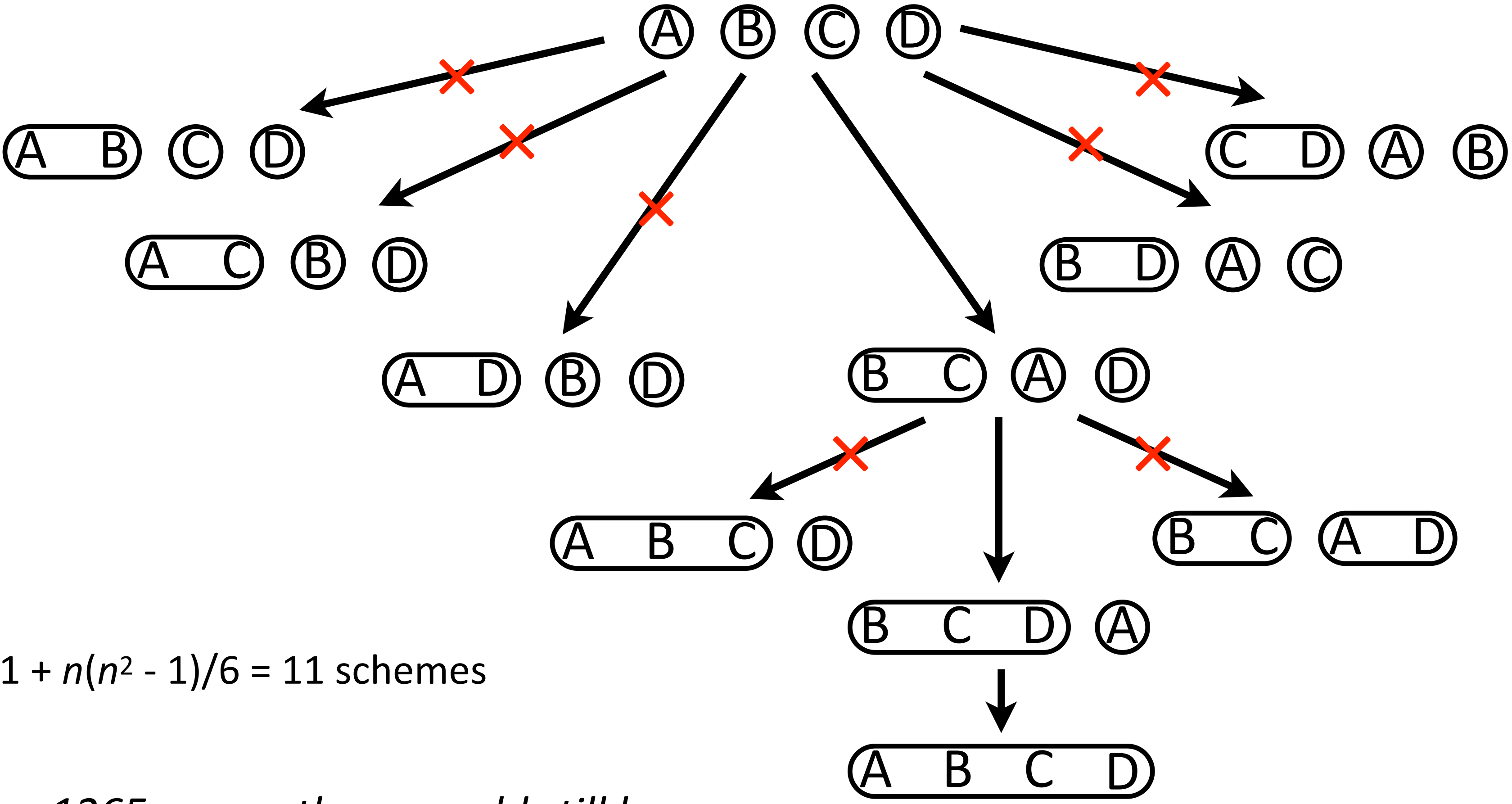
How many partitionings?

In general, the number of partitionings on *n* subsets is a “Bell number”

N	Bell number
2	2
3	5
4	15
5	52
6	203
7	877
12	4×10^{6}
60	9.8×10^{59}

Obviously, there are too many partitioning schemes to evaluate them all for more than a few subsets.

Greedy algorithm when there are too many partitionings



$1 + n(n^2 - 1)/6 = 11$ schemes

For 1265 genes, there would still be 337,380,561 schemes to evaluate!

Lanfear, R., Calcott, B., Ho, S. Y. W., & Guindon, S. (2012). Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular Biology and Evolution*, *29*(6), 1695–1701

How to partition thousands of genes (or other subsets)?

Cluster analysis

- Li, Lu, and Orti (2008)

  Estimate model parameters on a shared model; similar subsets will have similar parameter estimates and will cluster together.

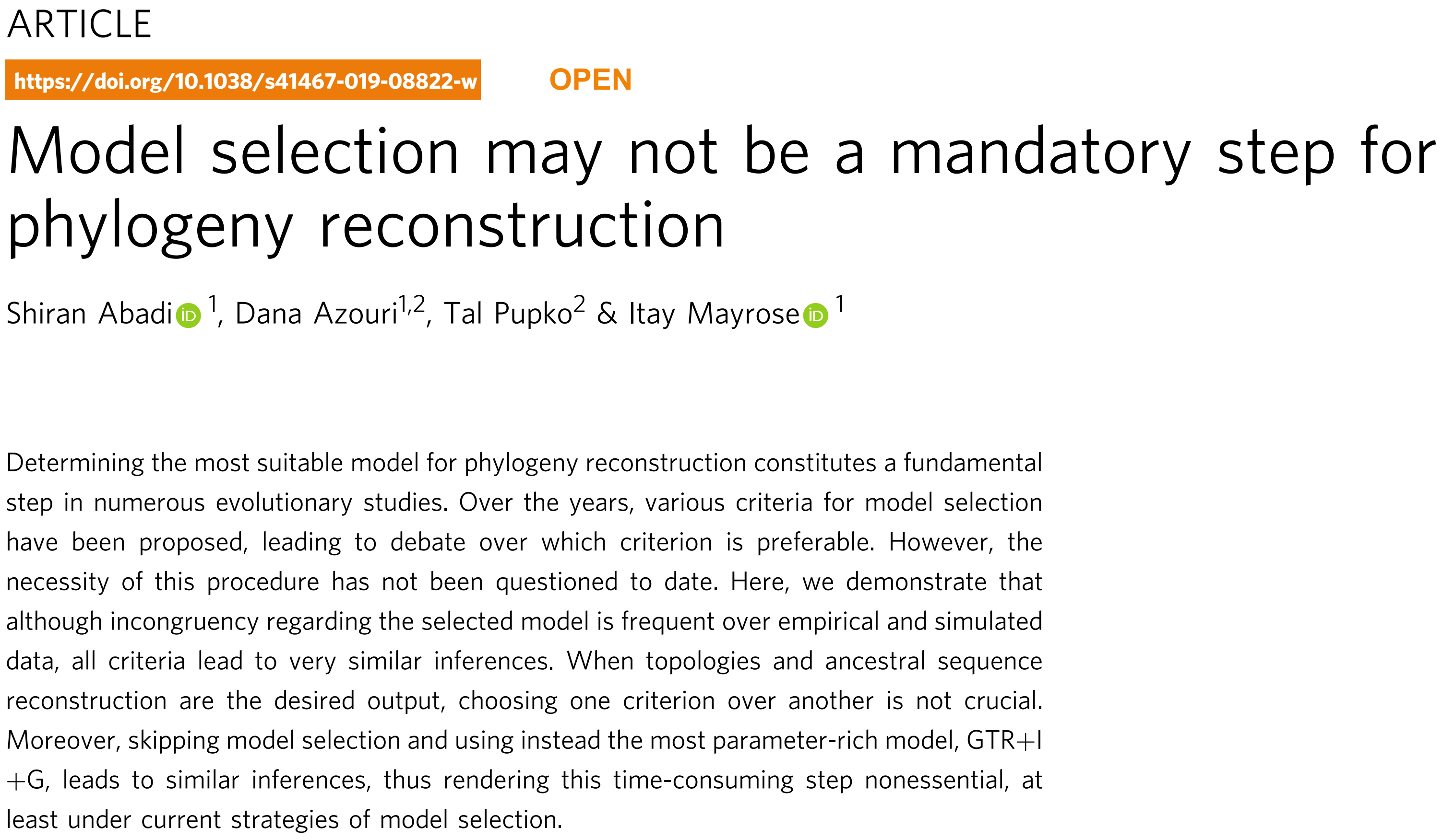
  Problem? Similar models (in the sense of predicting similar site pattern frequencies), can have different parameter MLEs. Must use same model for all subsets.

- Frandsen et al. (2015); Lanfear et al. (2016): PartitionFinder2)

  Hierarchical (or non-hierarchical k-means) clustering using same idea as Li et al. (very efficient implementation)

Is model selection really needed?

Abadi et al. (2019, *Nature Communications)*:



ARTICLE

https://doi.org/10.1038/s41467-019-08822-w OPEN

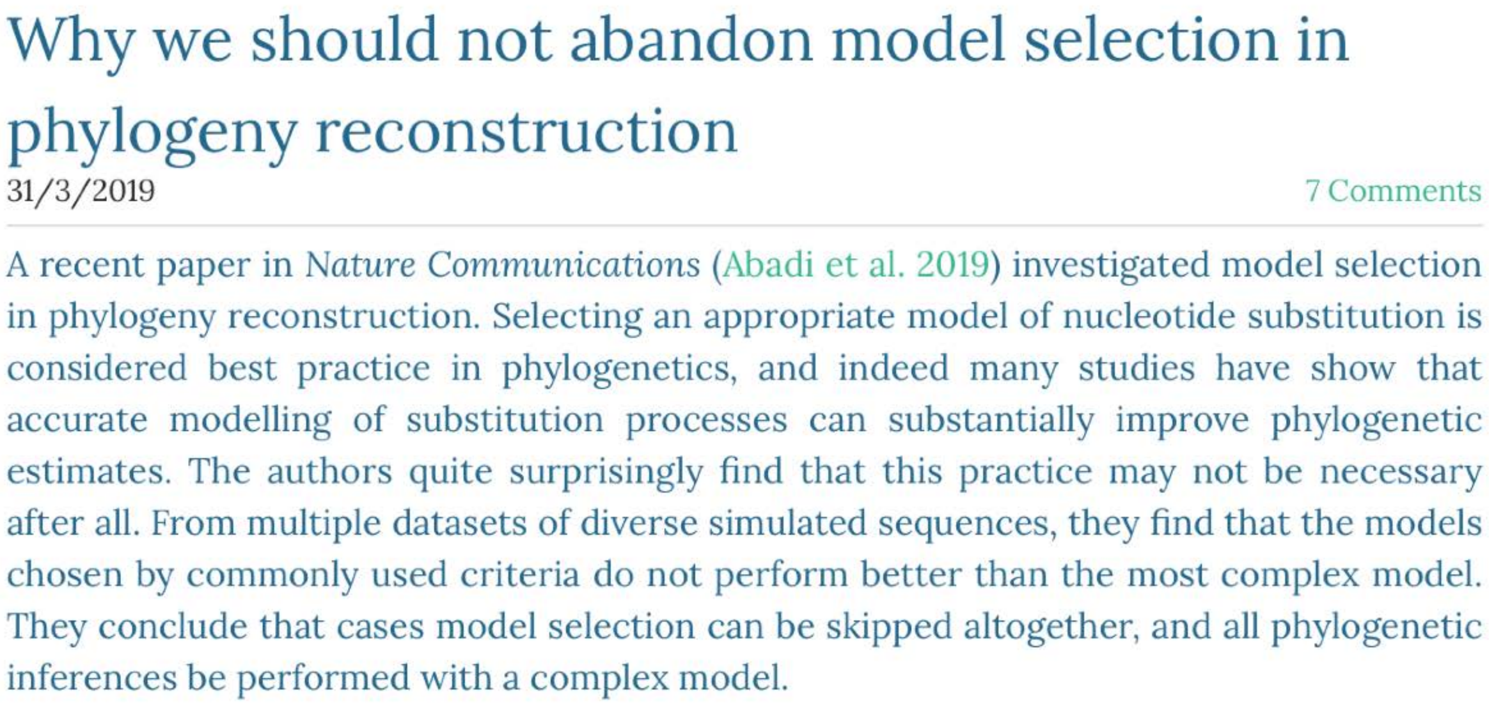
Model selection may not be a mandatory step for phylogeny reconstruction

Shiran Abadi[1], Dana Azouri[1,2], Tal Pupko[2] & Itay Mayrose[1]

Determining the most suitable model for phylogeny reconstruction constitutes a fundamental step in numerous evolutionary studies. Over the years, various criteria for model selection have been proposed, leading to debate over which criterion is preferable. However, the necessity of this procedure has not been questioned to date. Here, we demonstrate that although incongruency regarding the selected model is frequent over empirical and simulated data, all criteria lead to very similar inferences. When topologies and ancestral sequence reconstruction are the desired output, choosing one criterion over another is not crucial. Moreover, skipping model selection and using instead the most parameter-rich model, GTR+I +G, leads to similar inferences, thus rendering this time-consuming step nonessential, at least under current strategies of model selection.

Should model selection be abandoned?

Michael Gerth



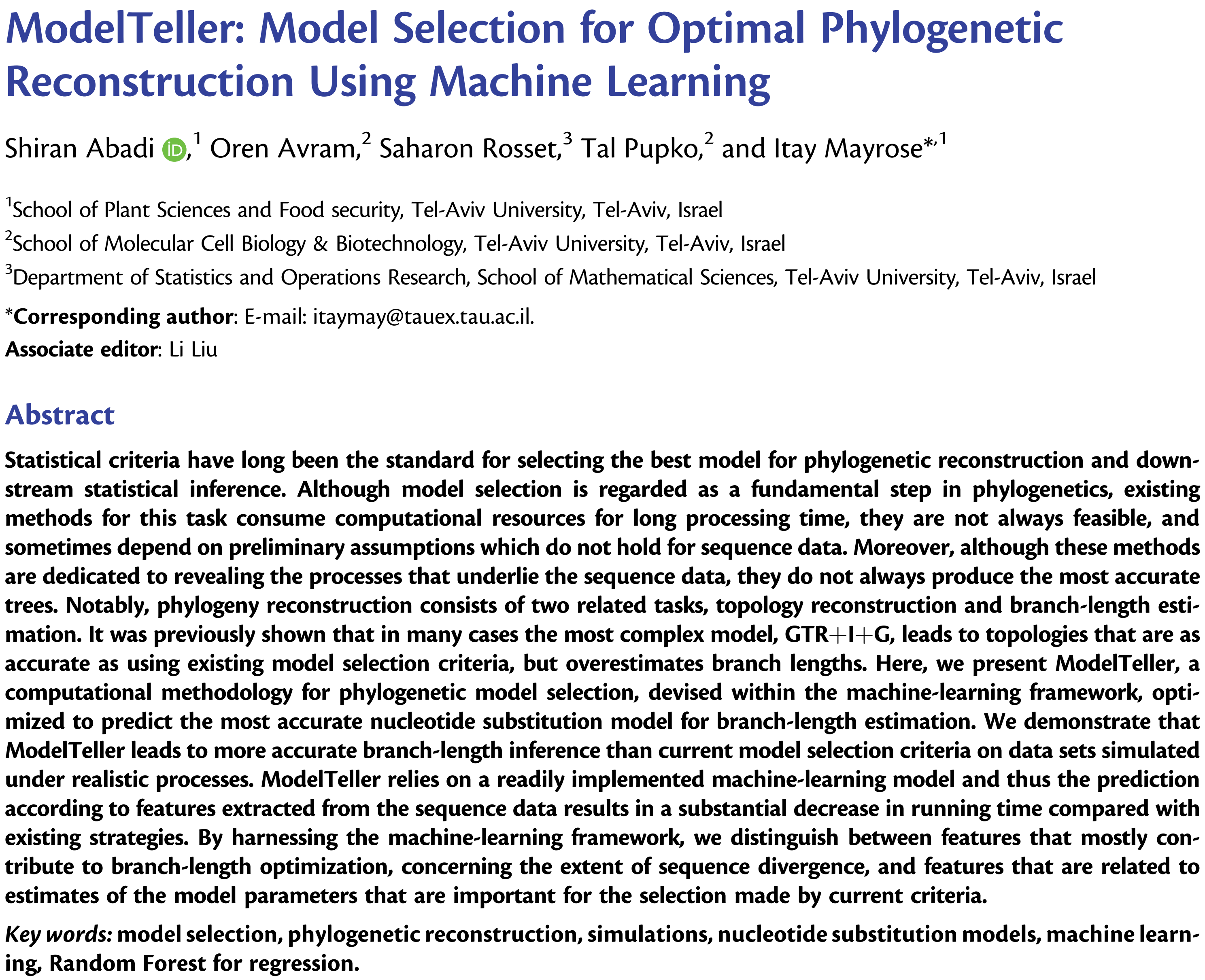
Why we should not abandon model selection in phylogeny reconstruction

31/3/2019

7 Comments

A recent paper in *Nature Communications* (Abadi et al. 2019) investigated model selection in phylogeny reconstruction. Selecting an appropriate model of nucleotide substitution is considered best practice in phylogenetics, and indeed many studies have show that accurate modelling of substitution processes can substantially improve phylogenetic estimates. The authors quite surprisingly find that this practice may not be necessary after all. From multiple datasets of diverse simulated sequences, they find that the models chosen by commonly used criteria do not perform better than the most complex model. They conclude that cases model selection can be skipped altogether, and all phylogenetic inferences be performed with a complex model.

https://www.michaelgerth.net/news--blog/why-we-should-not-abandon-model-selection-in-phylogeny-reconstruction



ModelTeller: Model Selection for Optimal Phylogenetic Reconstruction Using Machine Learning

Shiran Abadi,[1] Oren Avram,[2] Saharon Rosset,[3] Tal Pupko,[2] and Itay Mayrose*,[1]

[1]School of Plant Sciences and Food security, Tel-Aviv University, Tel-Aviv, Israel

[2]School of Molecular Cell Biology & Biotechnology, Tel-Aviv University, Tel-Aviv, Israel

[3]Department of Statistics and Operations Research, School of Mathematical Sciences, Tel-Aviv University, Tel-Aviv, Israel

***Corresponding author**: E-mail: itaymay@tauex.tau.ac.il.

Associate editor: Li Liu

Abstract

Statistical criteria have long been the standard for selecting the best model for phylogenetic reconstruction and downstream statistical inference. Although model selection is regarded as a fundamental step in phylogenetics, existing methods for this task consume computational resources for long processing time, they are not always feasible, and sometimes depend on preliminary assumptions which do not hold for sequence data. Moreover, although these methods are dedicated to revealing the processes that underlie the sequence data, they do not always produce the most accurate trees. Notably, phylogeny reconstruction consists of two related tasks, topology reconstruction and branch-length estimation. It was previously shown that in many cases the most complex model, GTR+I+G, leads to topologies that are as accurate as using existing model selection criteria, but overestimates branch lengths. Here, we present ModelTeller, a computational methodology for phylogenetic model selection, devised within the machine-learning framework, optimized to predict the most accurate nucleotide substitution model for branch-length estimation. We demonstrate that ModelTeller leads to more accurate branch-length inference than current model selection criteria on data sets simulated under realistic processes. ModelTeller relies on a readily implemented machine-learning model and thus the prediction according to features extracted from the sequence data results in a substantial decrease in running time compared with existing strategies. By harnessing the machine-learning framework, we distinguish between features that mostly contribute to branch-length optimization, concerning the extent of sequence divergence, and features that are related to estimates of the model parameters that are important for the selection made by current criteria.

Key words: **model selection, phylogenetic reconstruction, simulations, nucleotide substitution models, machine learning, Random Forest for regression.**

Abadi et al. (2020, MBE)