Introduction to Model Selection

2022 Woods Hole Molecular Evolution Workshop

David Swofford

Florida Museum of Natural History
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: (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.

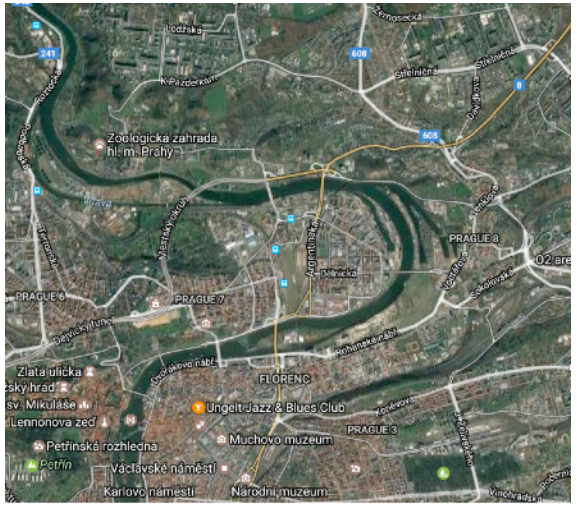
(Peterson poll)

Jordan Peterson

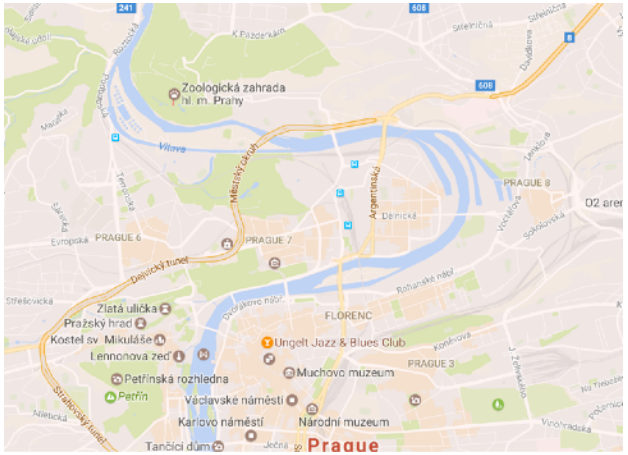
Jordan Bernt Peterson is a Canadian clinical psychologist, YouTube personality, author, and a professor emeritus at the University of Toronto. Peterson began to receive widespread attention as a public intellectual in the late 2010s for his views on cultural and political issues, often described as conservative.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

Peterson on models

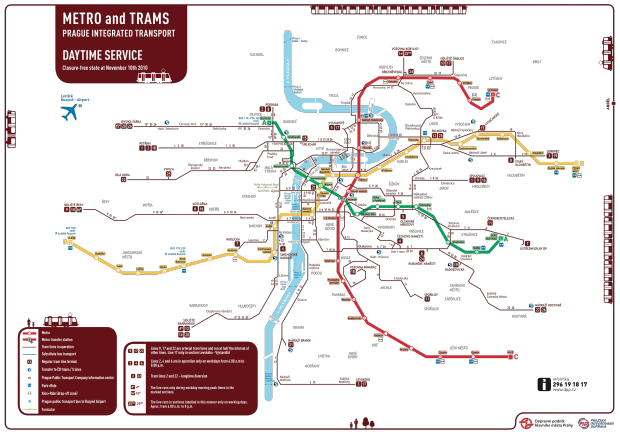
Which is more useful?



"Reality"



Detailed map



Detailed public transportation

Simplified metro

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)

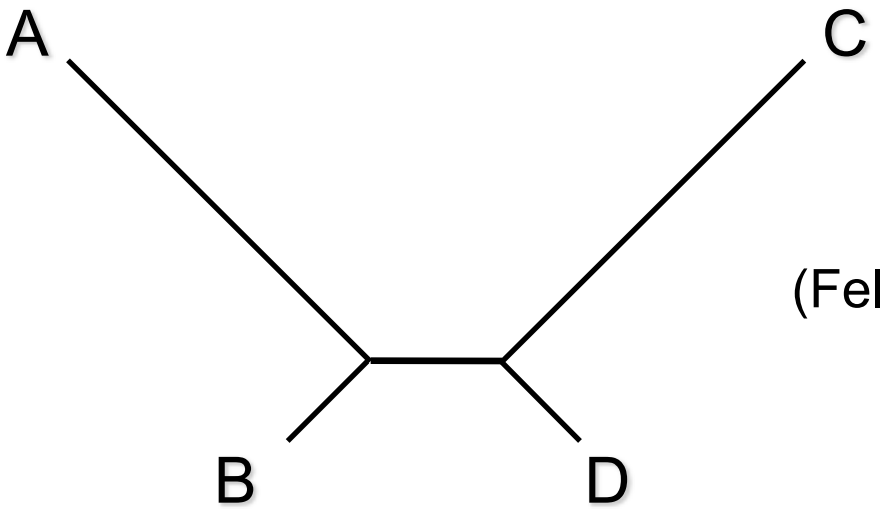
A model is a simplification or approximation of reality and hence will not reflect all of reality ... While a model can never be “truth,” a model might be ranked from very useful, to useful, to somewhat useful to, finally, essentially useless. (Burnham and Anderson, 2002)

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?

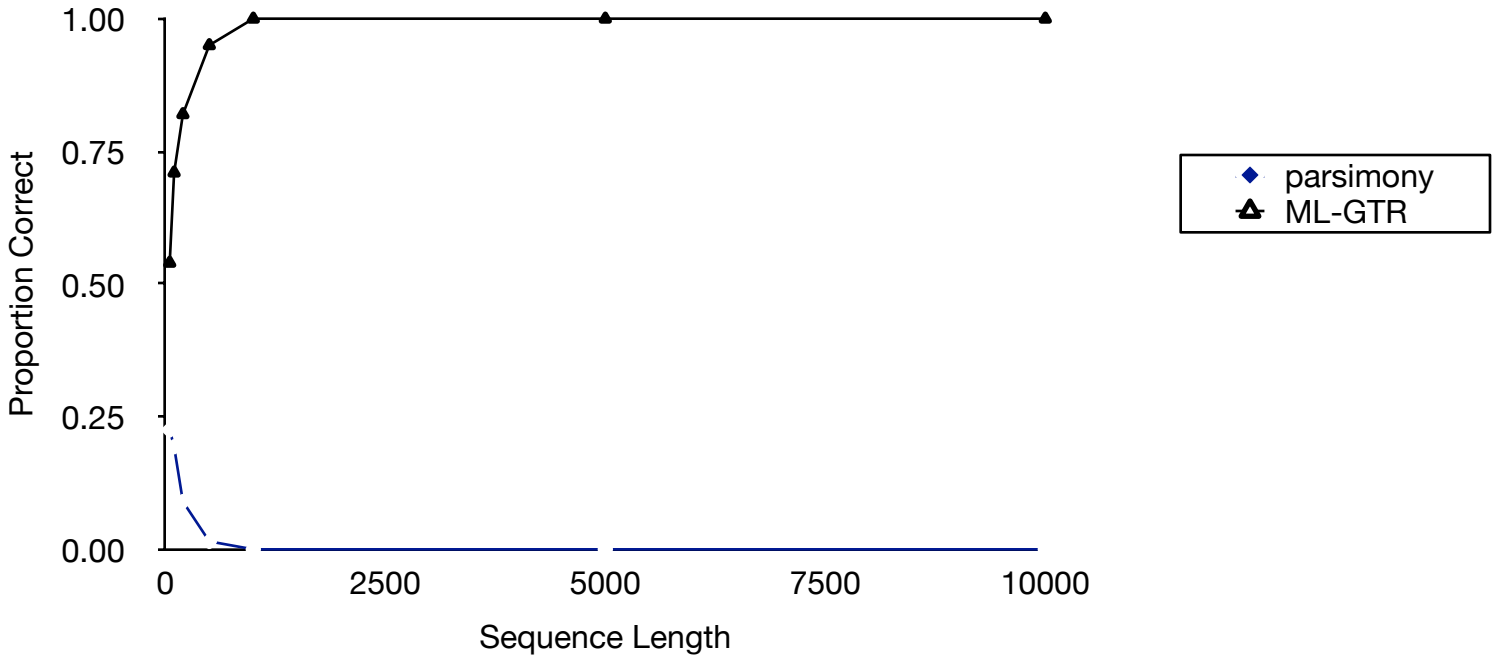
Model-based methods including ML and Bayesian inference (typically) make a *consistent* estimate of the phylogeny (estimate converges to true tree as number of sites increases toward infinity)

... 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.

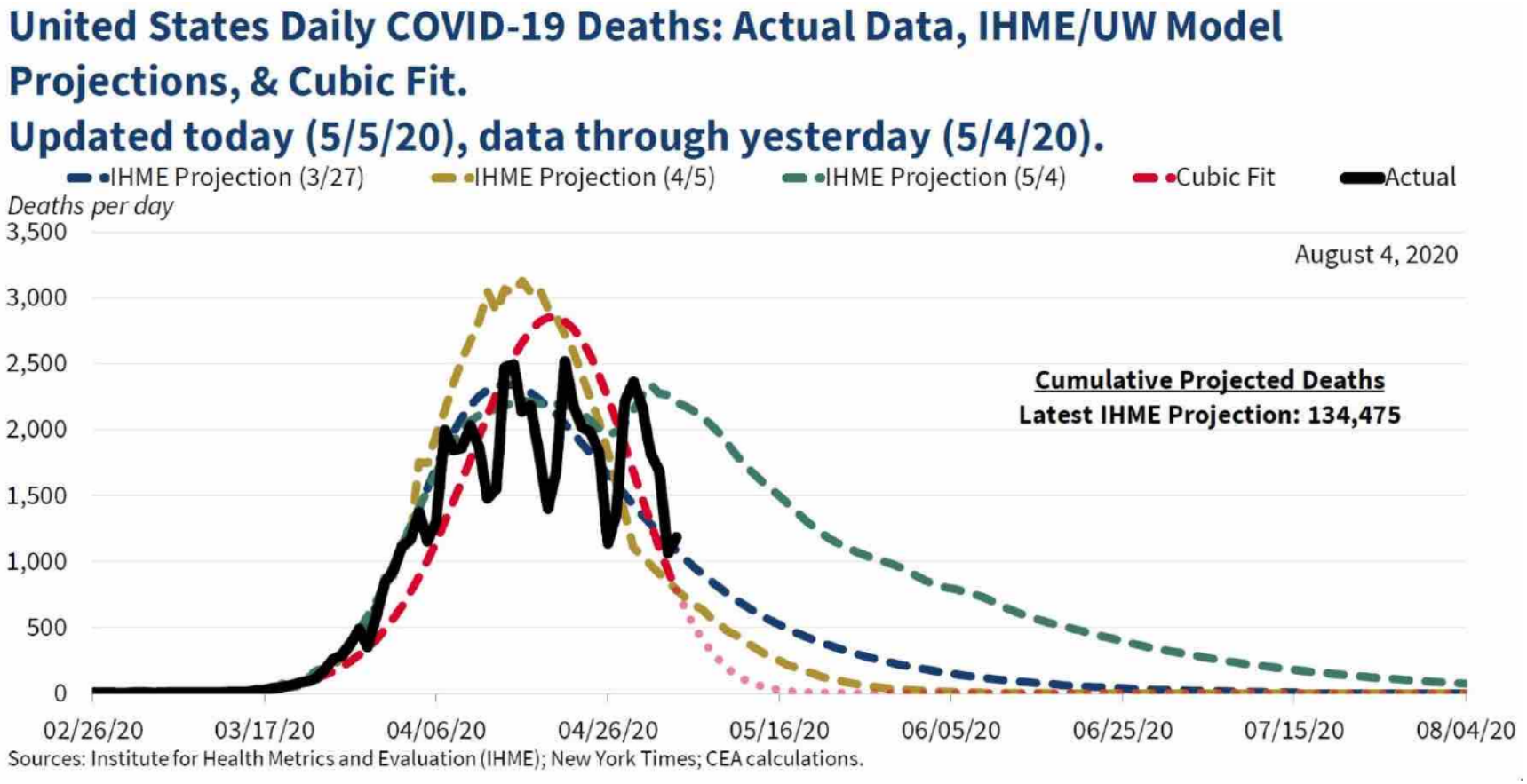
By Matthew Yglesias | @mattyglesias | matt@vox.com | May 8, 2020, 11:00am EDT

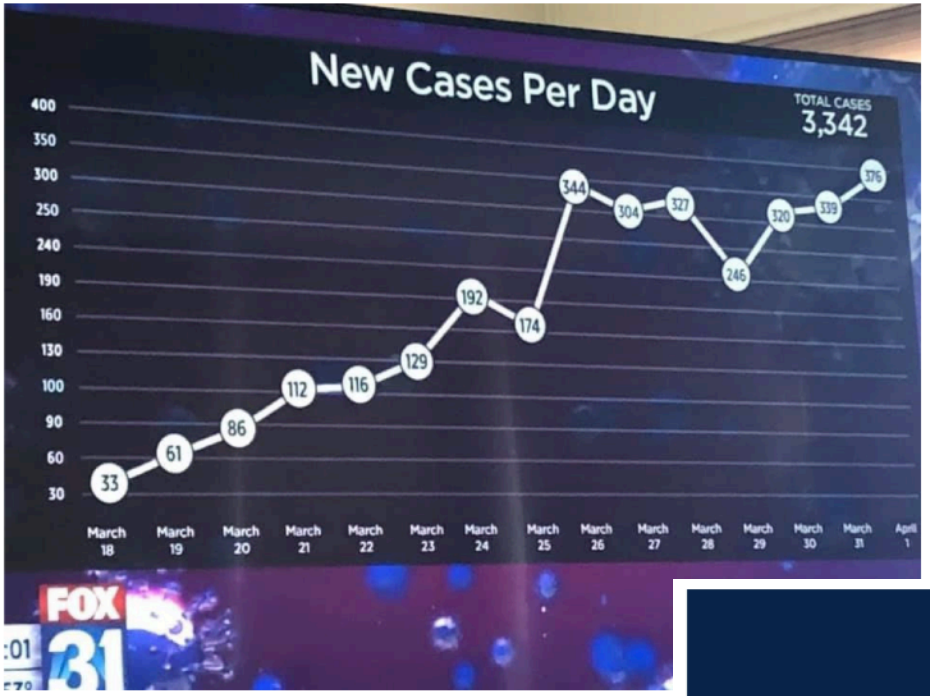
SHARE

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



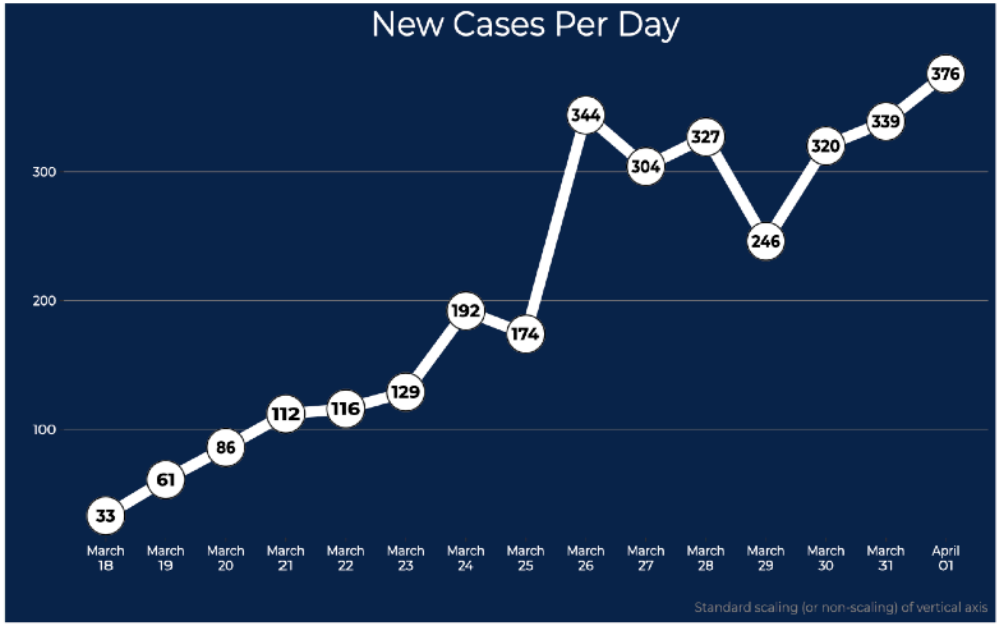


From "free range statistics" blog

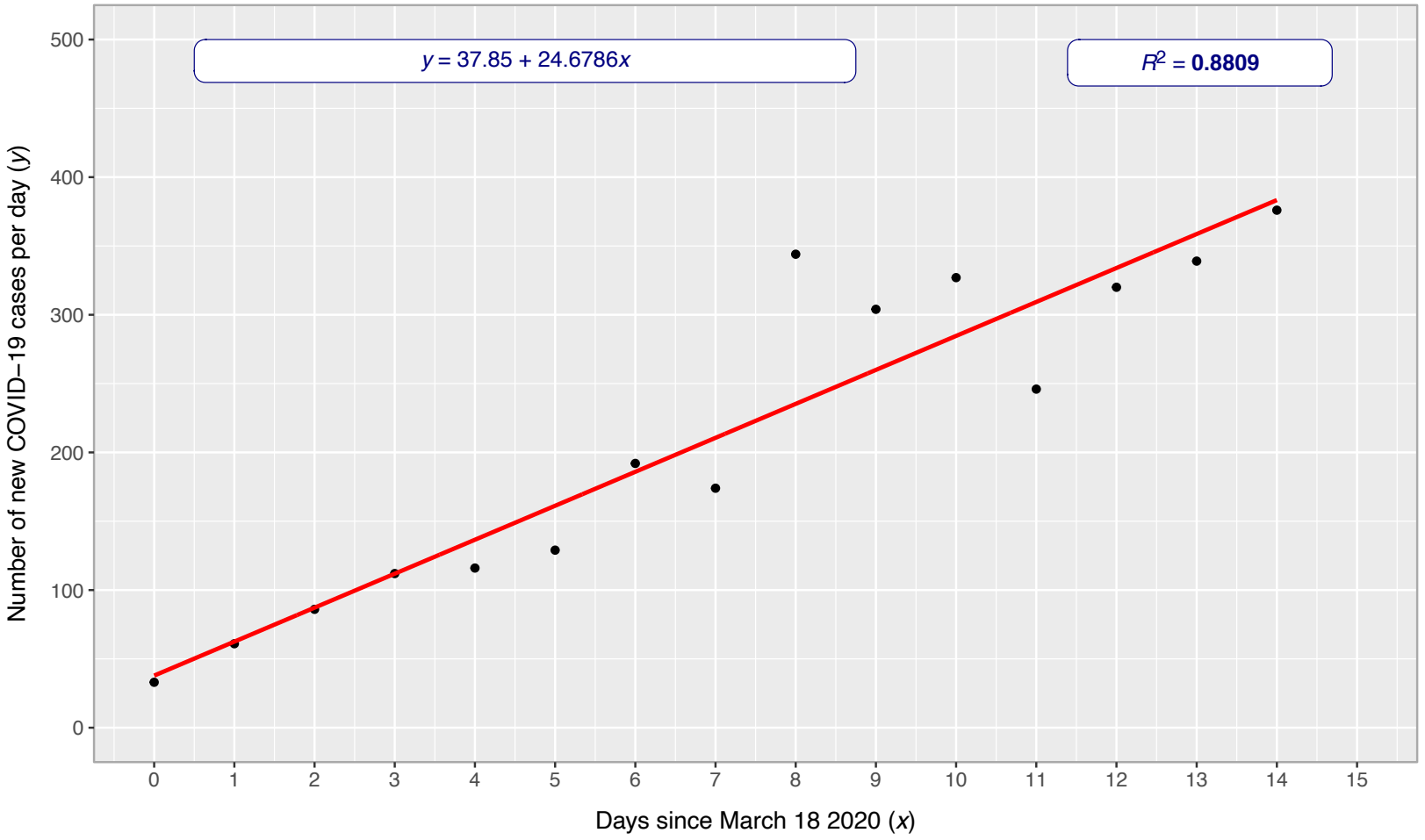
(Peter Ellis)

"It's so bad it's funny. This is clearly incompetence not malevolence. But it's a serious degree of incompetence."

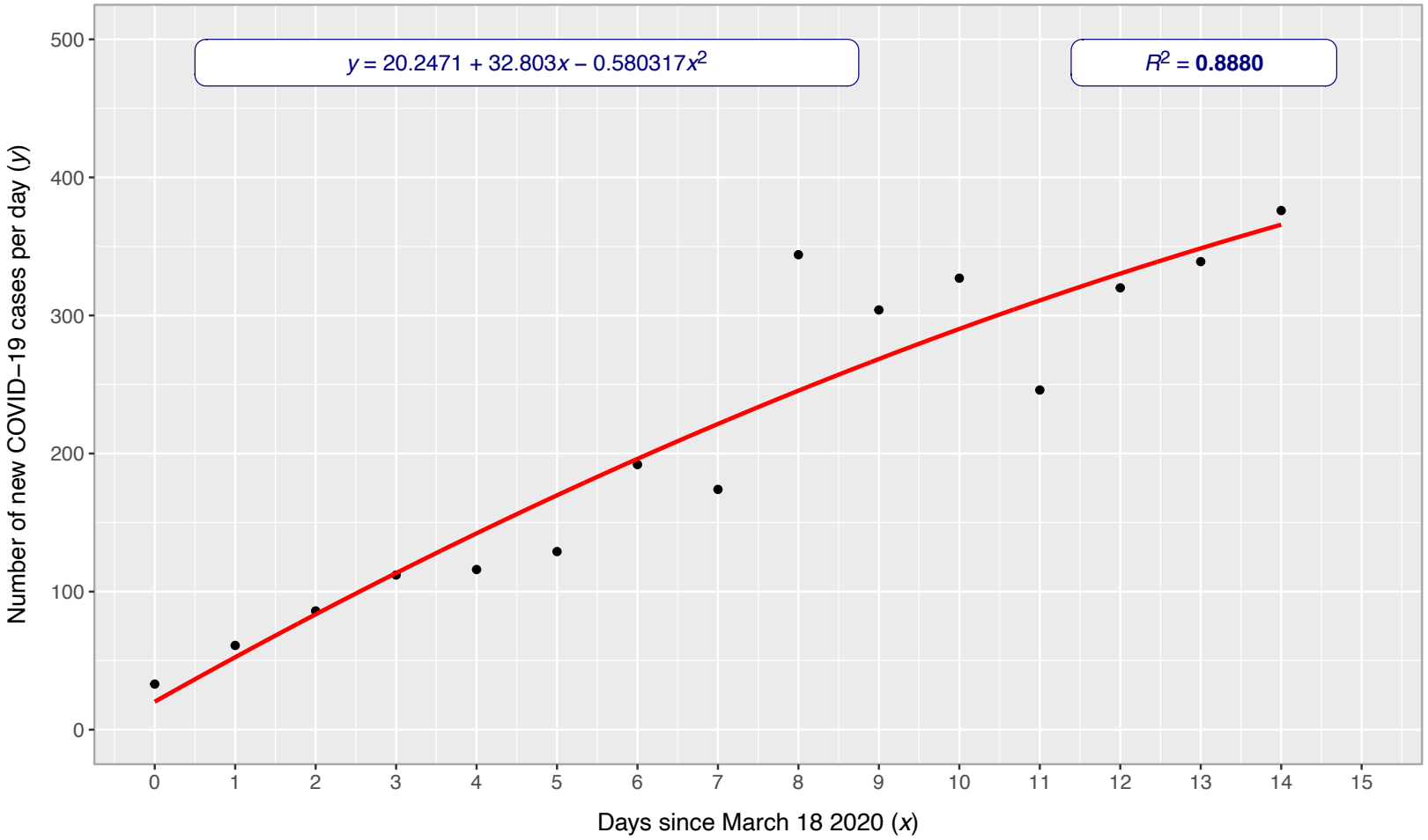
After fix-up



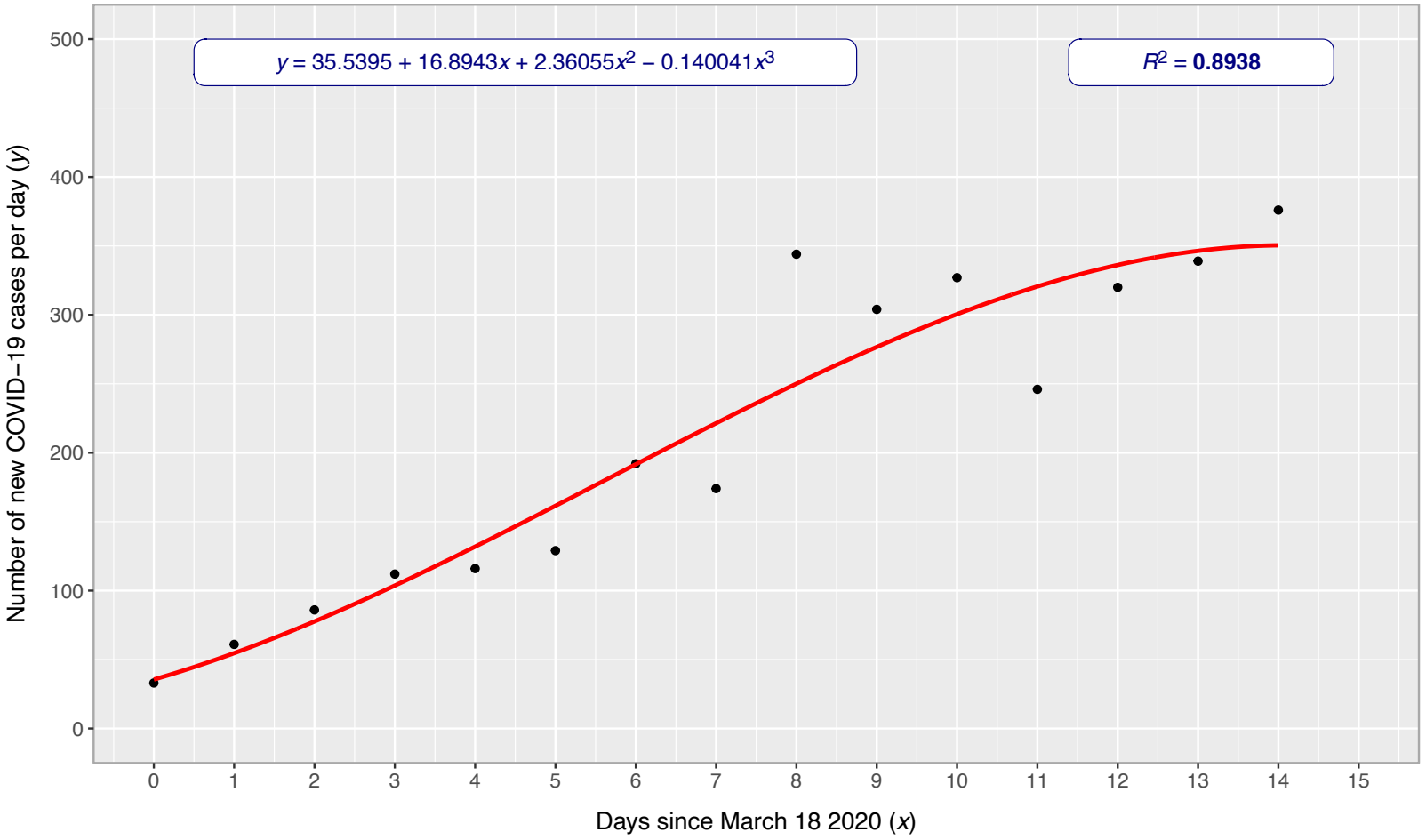
Simple linear regression



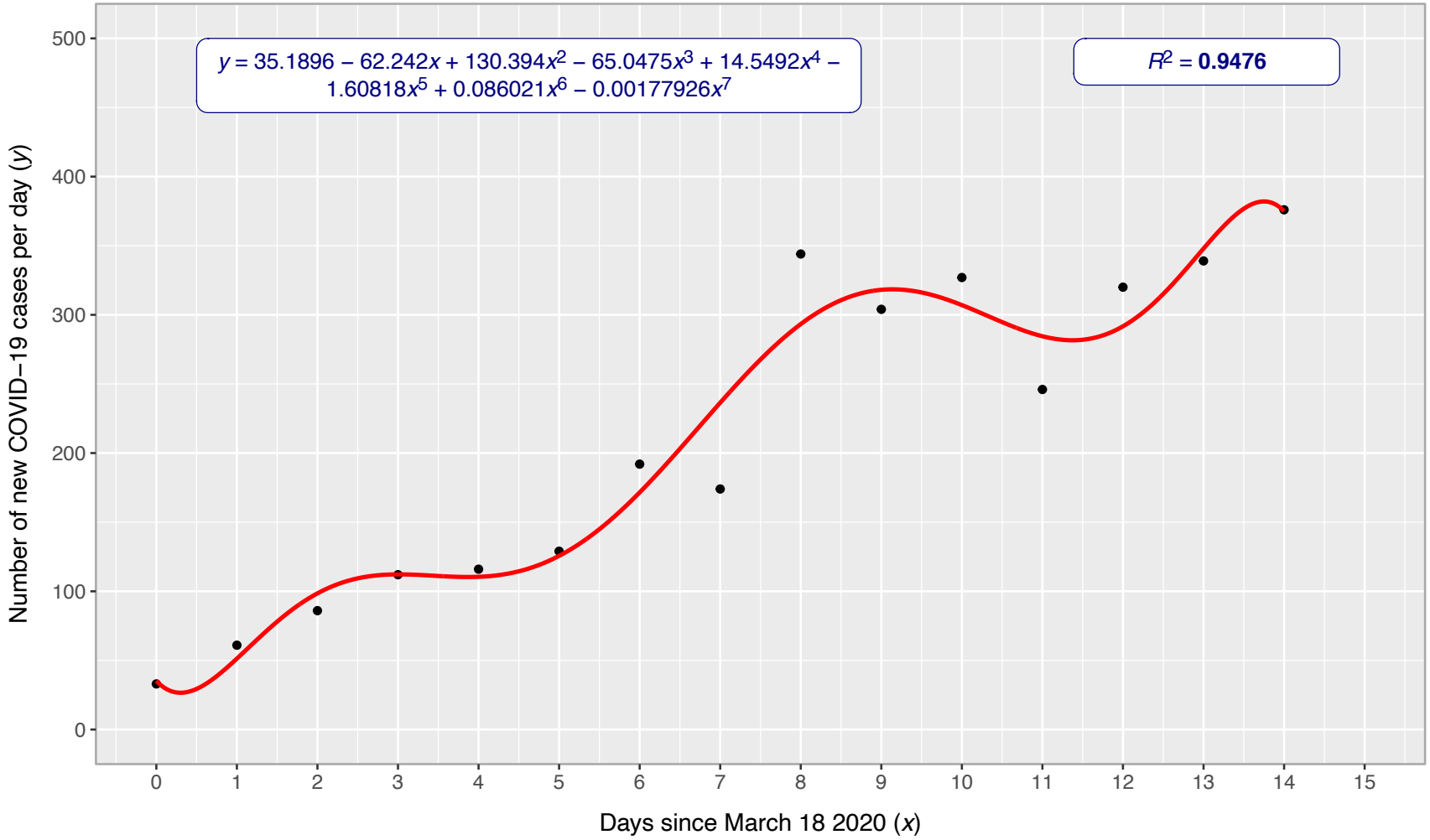
Quadratic model



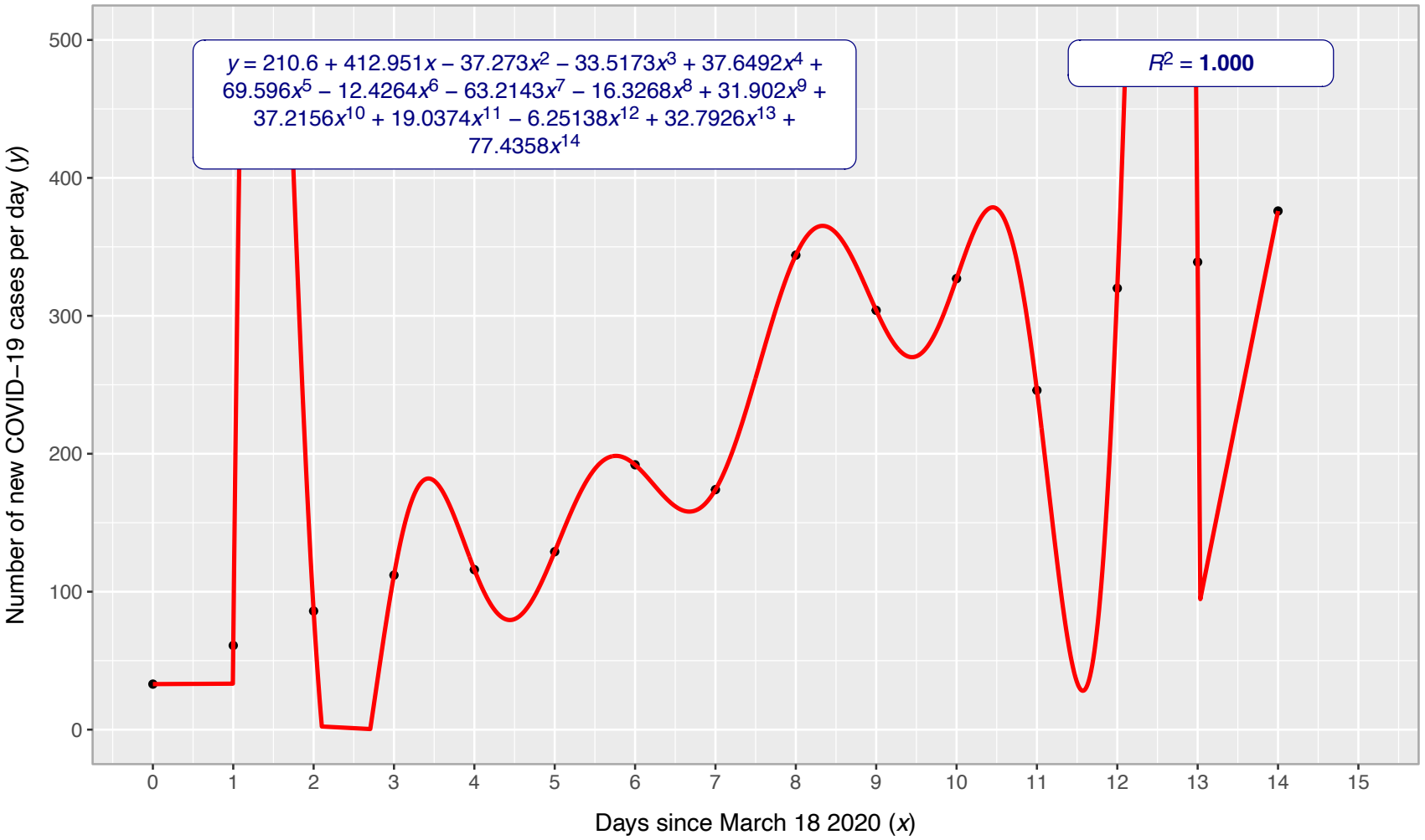
Cubic model


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

7th order polynomial



14th order polynomial

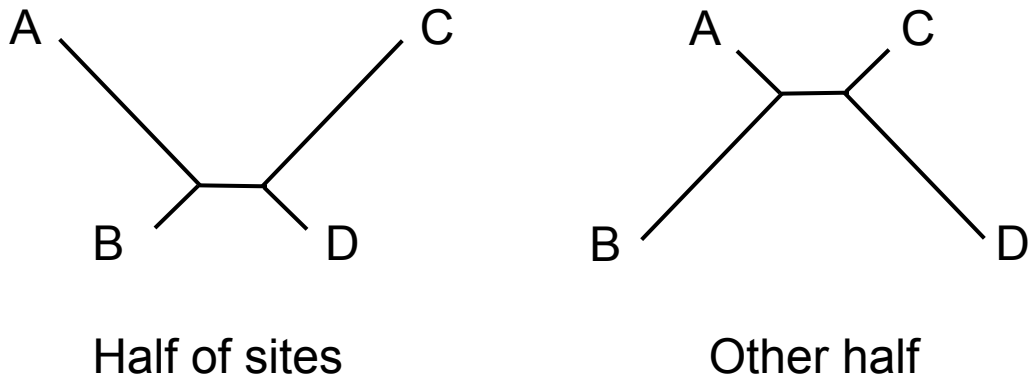


Why models don't have to be perfect

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.:

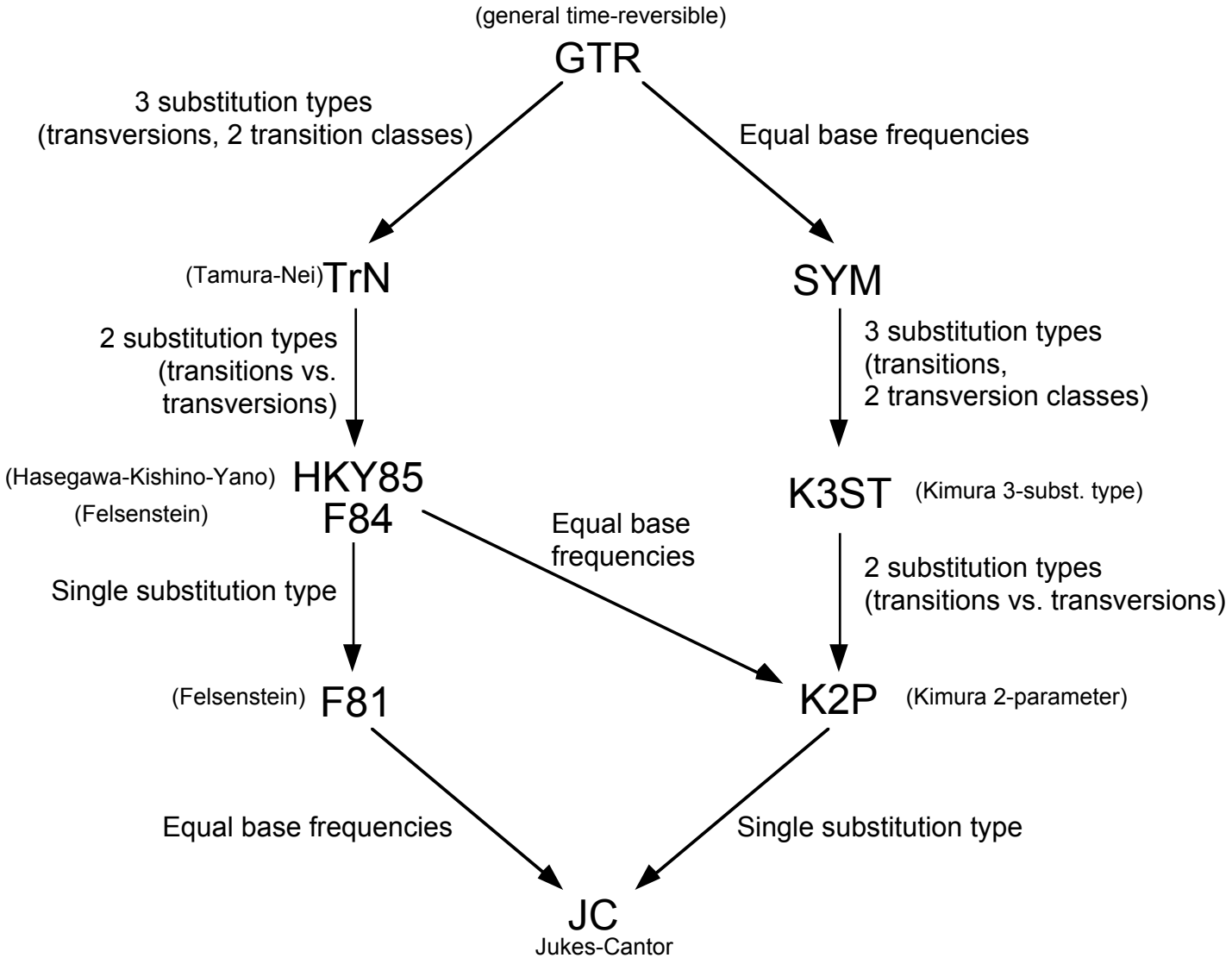


Half of sites

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

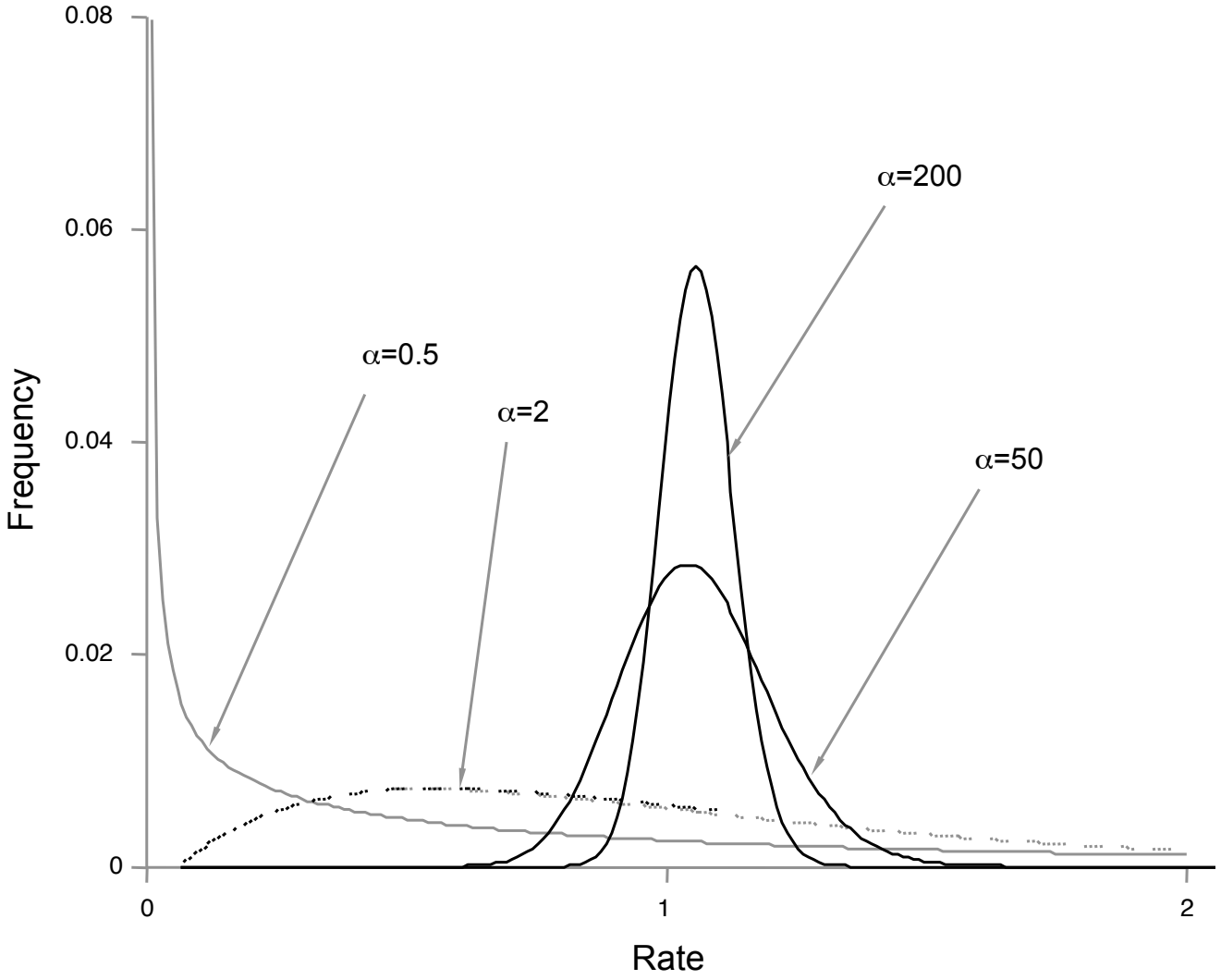


Among site rate heterogeneity

```
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
```

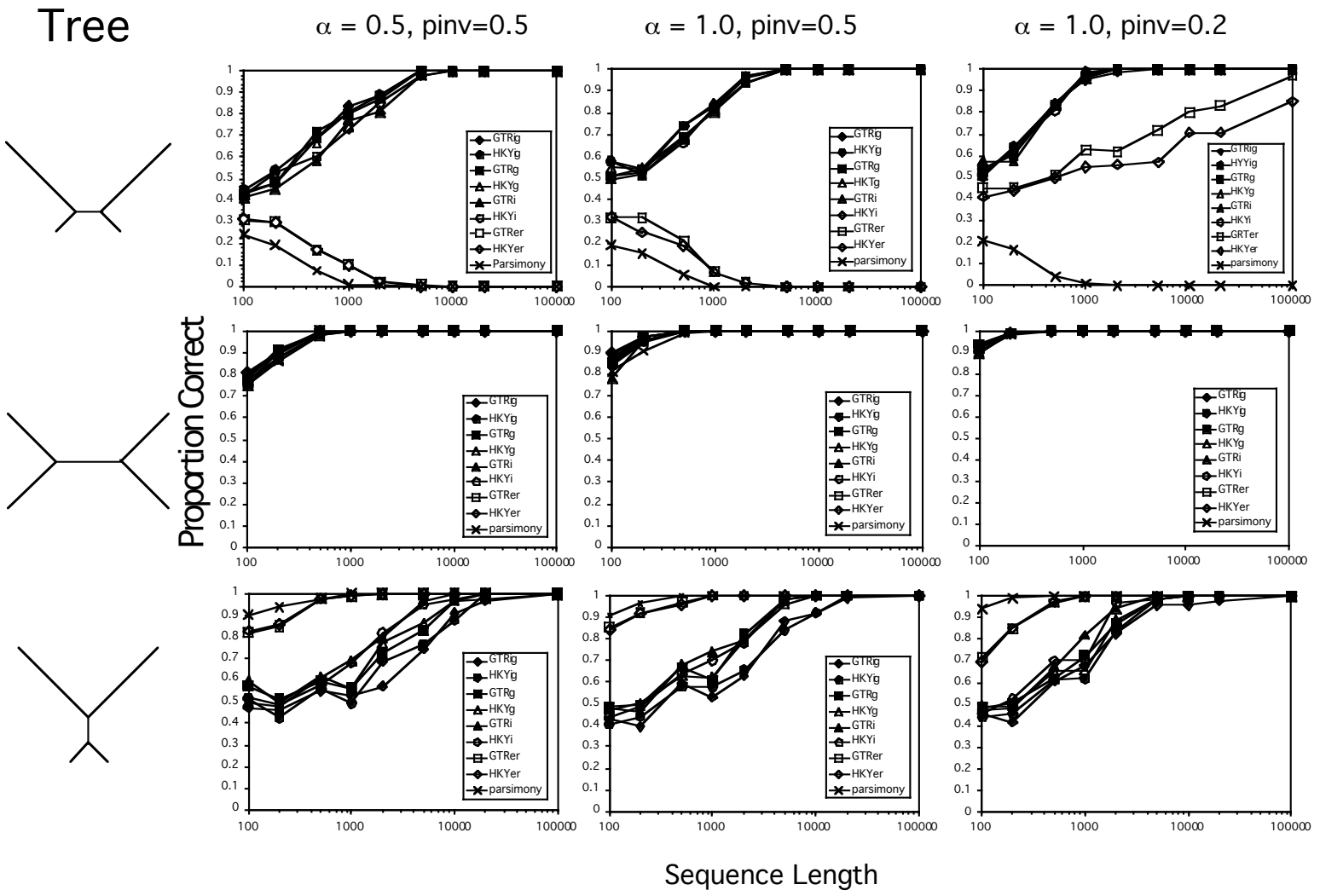
- Proportion of invariable sites
  - Some sites extremely unlikely to change due to strong functional or structural constraint (Hasegawa et al., 1985)
- Gamma-distributed rates
  - Rate variation assumed to follow a gamma distribution with shape parameter α
- Site-specific rates (another way to model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

Modeling ASRV with gamma distribution



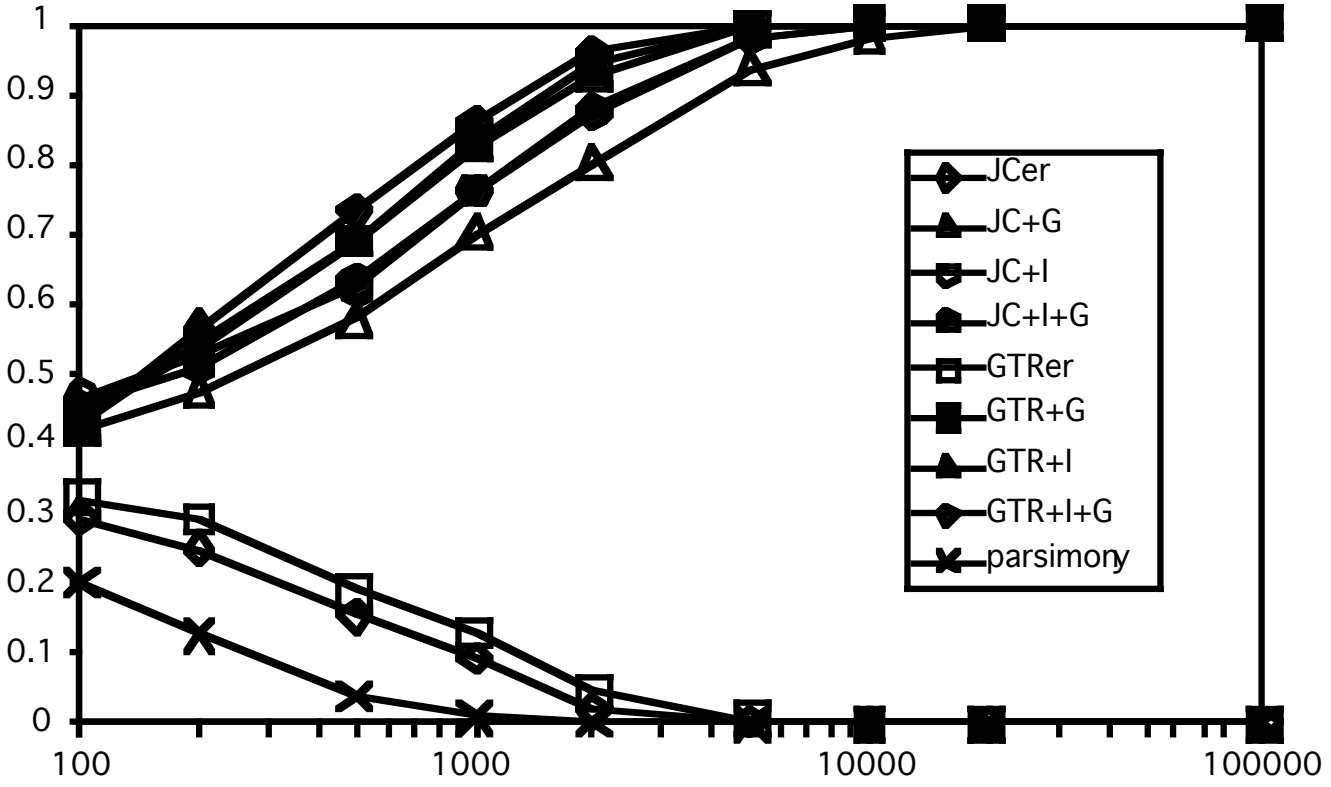
…can also include a proportion of “invariable” sites (p_{inv})

Performance of ML when its model is violated

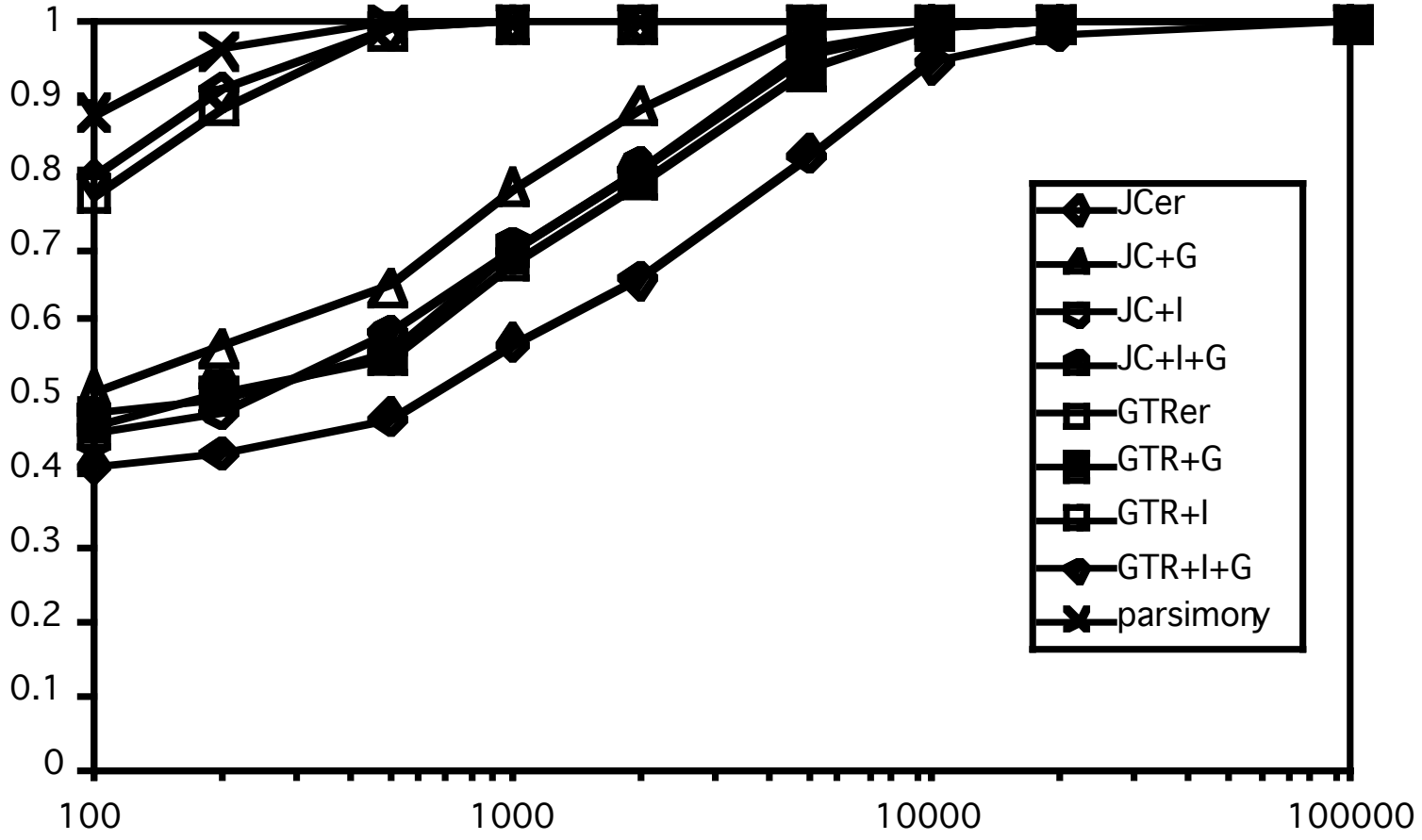


“MODERATE”–Felsenstein zone

$\alpha = 1.0, p_{inv}=0.5$



“MODERATE”–Inverse-Felsenstein zone



Likelihood ratio tests

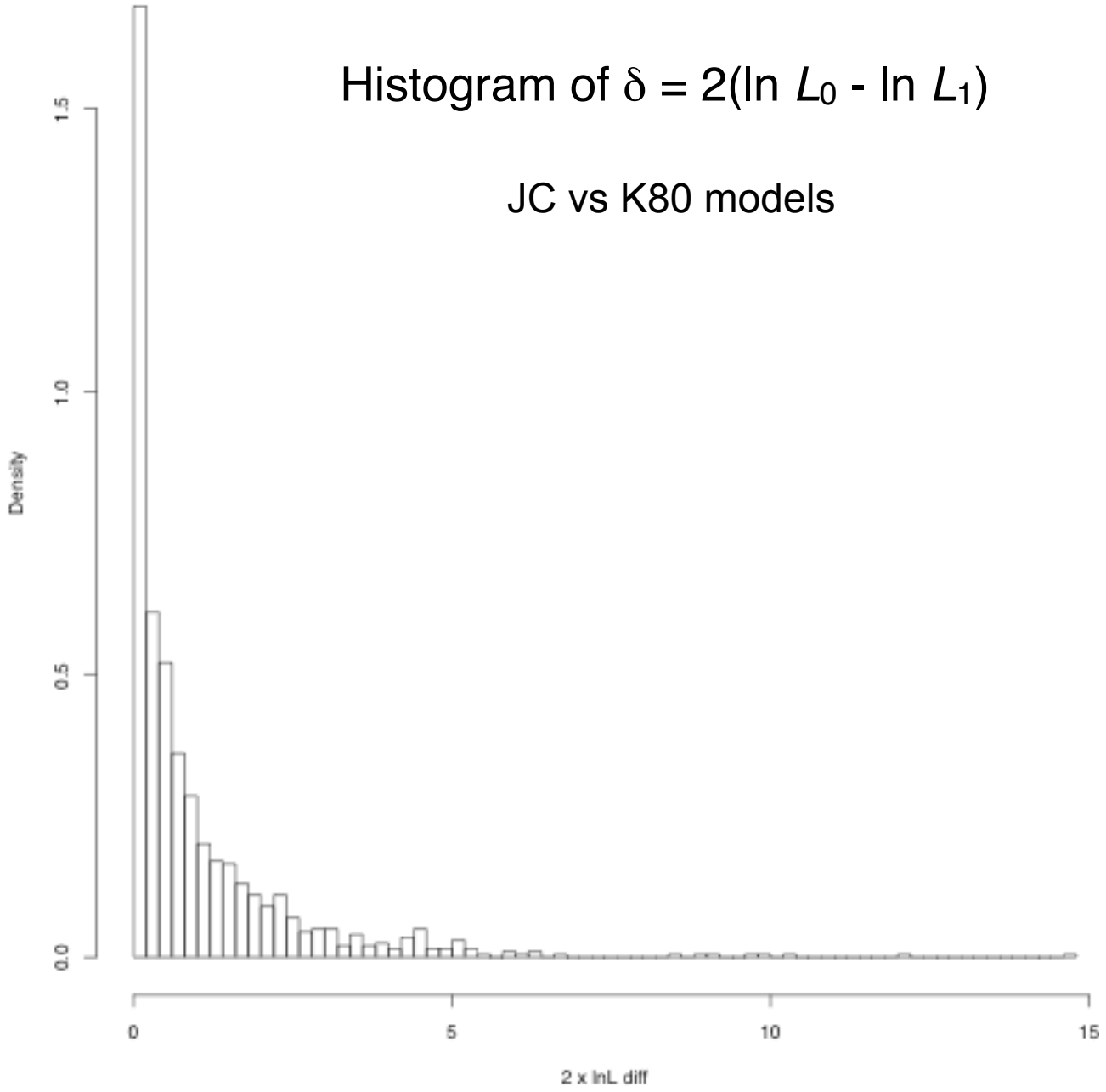
- Calculate "delta" statistic

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

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

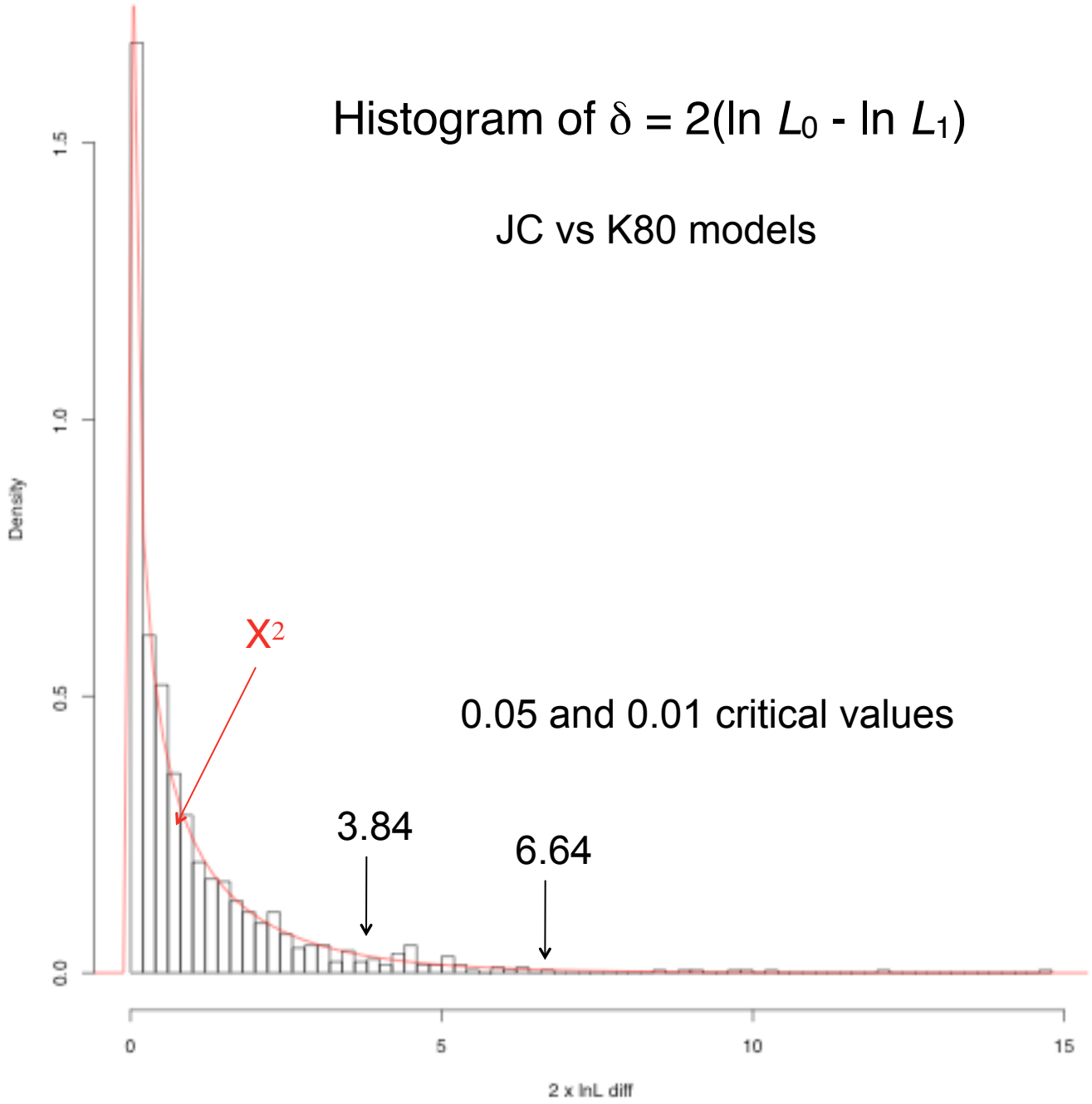
Histogram of $\delta = 2(\ln L_0 - \ln L_1)$

JC vs K80 models



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

JC vs K80 models



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 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.

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:

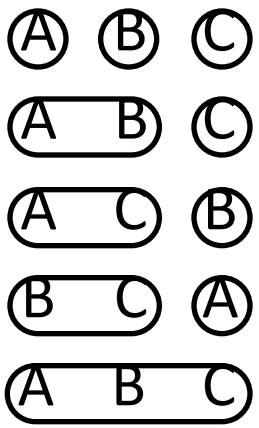
- Gene A: HKY+G, π = (0.26, 0.24, 0.23, 0.27), κ=1.1, α=3.0
- Gene B: GTR, π = (0.25,0.24,0.25,0.26), (a,b,c,d,e)=(1.1, 1.2, 0.9, 1.1, 0.95)
- Gene C: JC+I (pinv=0.05)

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* test versions

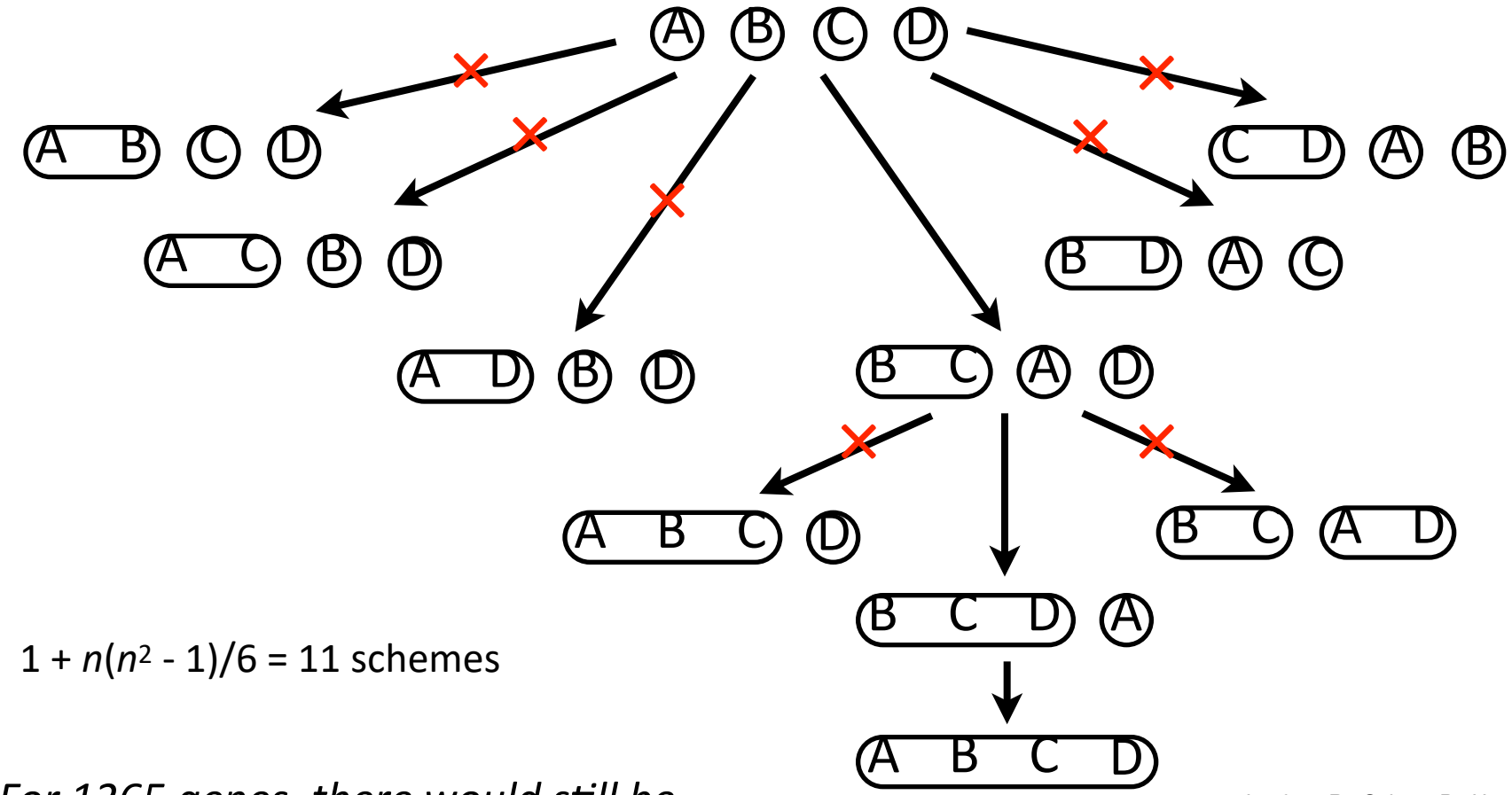
How many partitionings?

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