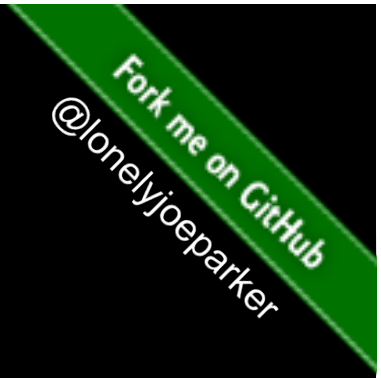


Why Aren't We Benchmarking Bioinformatics?

Joe Parker

Early Career Research Fellow (Phylogenomics)
Department of Biodiversity Informatics and Spatial Analysis
Royal Botanic Gardens, Kew
Richmond TW9 3AB
joe.parker@kew.org

Kew
ROYAL BOTANIC GARDENS

Outline

- Introduction
- Brief history of Bioinformatics
- Benchmarking in Bioinformatics
- Case study 1: Typical benchmarking across environments
- Case study 2: Mean-variance relationship for repeated measures
- Conclusions: implication for statistical genomics

A (very) brief history of bioinformatics

QUANTITATIVE PHYLETICS AND THE EVOLUTION OF ANURANS

ARNOLD G. KLUGE AND JAMES S. FARRIS

Abstract

In the quantitative phyletic approach to evolutionary taxonomy, quantitative methods are used for inferring evolutionary relationships. The methods are chosen both for their operationism and for their connection to evolutionary theory and the goals of evolutionary taxonomy. As an example of this approach, a detailed analysis of a set of anuran characters is presented and taxonomic conclusions based on those characters are drawn. The methods and conclusions of the quantitative phyletic analysis are compared and contrasted with the methods of previous workers in the field of anuran classification.

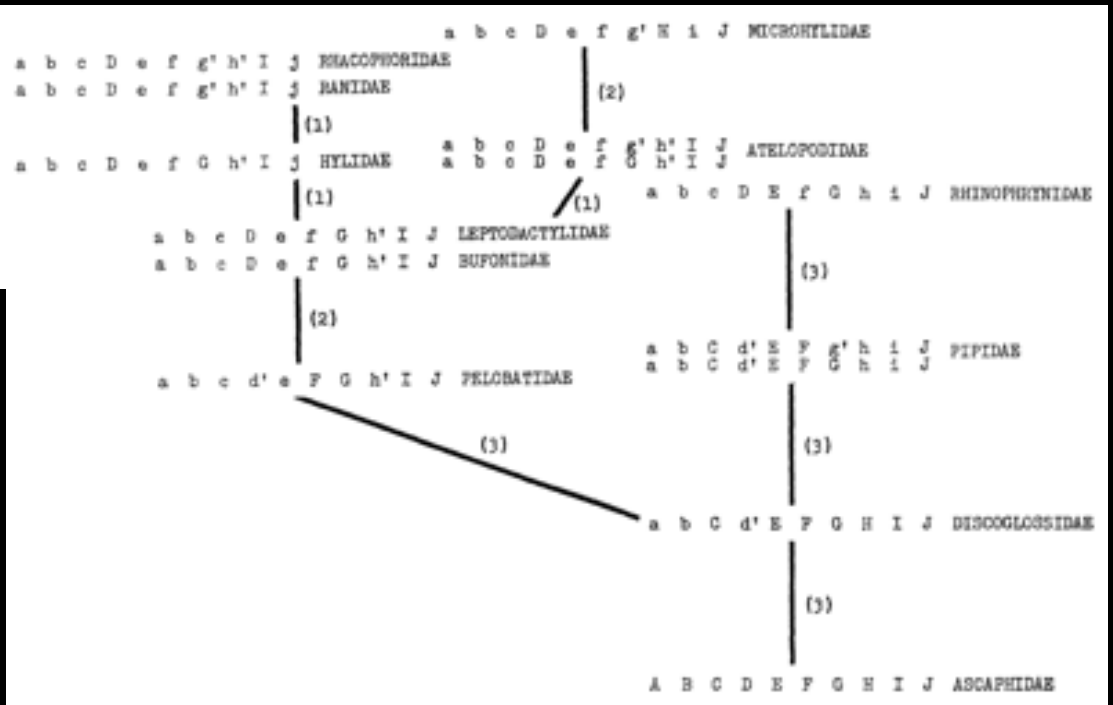


FIG. 3.—A maximum parsimony phylogeny of anuran families constructed according to the procedure of Wagner (1961). The 10 characters are those used by Inger (1967), except that all character states exhibited by the Ascaphidae are treated as primitive; see page 8 for further description of data modification. Compared to Fig. 1, this phylogeny is shorter by one evolutionary step, 19, and it includes one less homoplasy, d, and one extra reversal, D.

Kluge & Farris (1969) *Syst. Zool.* **18:**1-32

A (very) brief history of bioinformatics

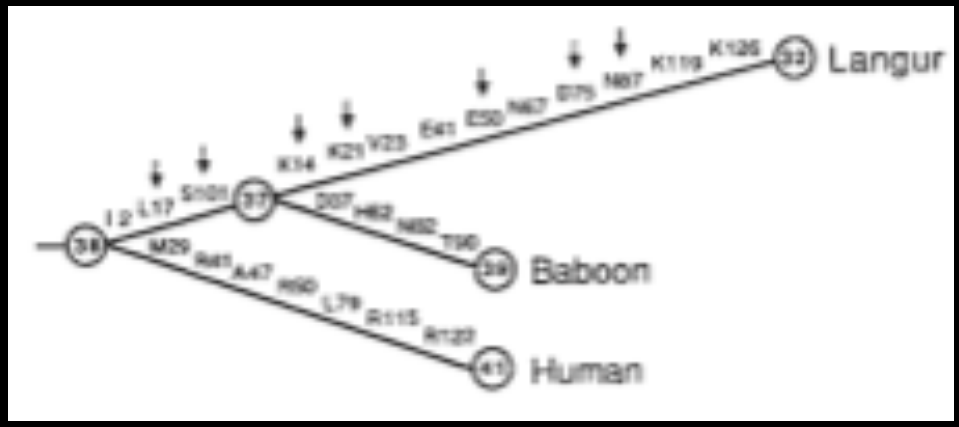
Adaptive evolution in the stomach lysozymes of foregut fermenters

Caro-Beth Stewart*†, James W. Schilling‡ & Allan C. Wilson*

* Department of Biochemistry, University of California, Berkeley, California 94720, USA
‡ California Biotechnology, Inc., 2450 Bayshore Parkway, Mountain View, California 94043, USA

The convergent evolution of a fermentative foregut in two groups of mammals offers an opportunity to study adaptive evolution at the protein level. The appearance of this mode of digestion has been accompanied by the recruitment of lysozyme as a bacteriolytic enzyme in the stomach both in the ruminants (for example the cow) and later in the colobine monkeys (for example the langur). The stomach lysozymes of these two g chemical and catalytic properties that functioning in the stomach fluid[1,2]. To shared properties, we sequenced lang compared it to other lysozymes of kno suggests that, after foregut fermentat langur lysozyme gained sequence si lysozyme and evolved two times faste lysozymes. This rapid evolution, cou sequence convergence upon cow stoma that positive darwinian selection has driven about 50% of the evolution of langur stomach lysozyme.

The majority of evolutionary changes revealed by comparative studies of proteins and nucleic acids appears to fit the neutral theory of molecular evolution; that is, they could have become fixed by random drift of selectively neutral or nearly neutral mutations rather than by positive darwinian selection[3]. For this



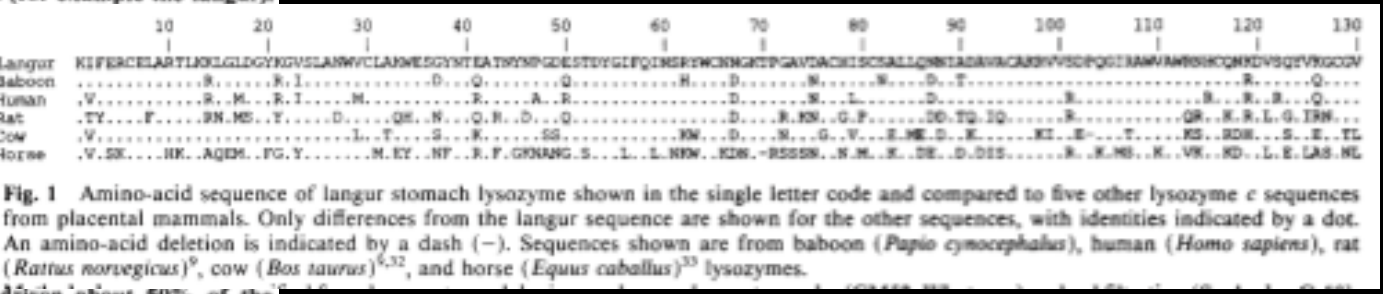


Fig. 1 Amino-acid sequence of langur stomach lysozyme shown in the single letter code and compared to five other lysozyme *c* sequences from placental mammals. Only differences from the langur sequence are shown for the other sequences, with identities indicated by a dot. An amino-acid deletion is indicated by a dash (−). Sequences shown are from baboon (*Papio cynocephalus*), human (*Homo sapiens*), rat (*Rattus norvegicus*)[9], cow (*Bos taurus*)[6,32], and horse (*Equus caballus*)[33] lysozymes.

Stewart *et al.* (1987) *Nature* **330:40**1-404

A (very) brief history of bioinformatics

ARTICLE

doi:10.1038/nature11247

An integrated encyclopedia of DNA elements in the human genome

The ENCODE Project Consortium*

The human genome encodes the blueprint of life, but the function of the vast majority of its nearly three billion bases is unknown. The Encyclopedia of DNA Elements (ENCODE) project has systematically mapped regions of transcription, transcription factor association, chromatin structure and histone modification. These data enabled us to assign biochemical functions for 80% of the genome, in particular outside of the well-studied protein-coding regions. Many discovered candidate regulatory elements are physically associated with one another and with expressed genes, providing new insights into the mechanisms of gene regulation. The newly identified elements also show a statistical correspondence to sequence variants linked to human disease, and can thereby guide interpretation of this variation. Overall, the project provides new insights into the organization and regulation of our genes and genome, and is an expansive resource of functional annotations for biomedical research.

The human genome sequence provides the underlying code for human biology. Despite intensive study, especially in identifying protein-coding genes, our understanding of the genome is far from complete, particularly with regard to non-coding RNAs, alternatively spliced transcripts and regulatory sequences. Systematic analyses of transcripts and regulatory information are essential for the identification of genes and regulatory

ENCODE
Encyclopedia of DNA Elements
nature.com/encode

95% of the genome lies within 8 kilobases (kb) of a DNA–protein interaction (as assayed by bound ChIP-seq motifs or DNase I footprints), and 99% is within 1.7 kb of at least one of the biochemical events measured by ENCODE.

• Primate-specific elements as well as elements without detectable mammalian constraint show, in aggregate, evidence of negative selection; thus, some of them are expected to be functional.

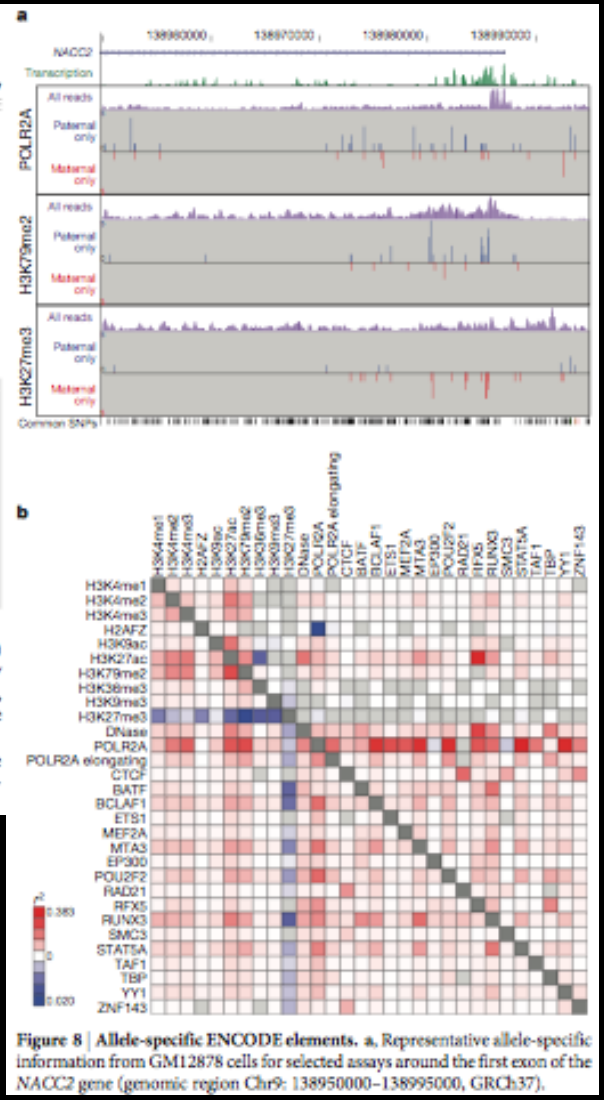


Figure 8 | Allele-specific ENCODE elements. a, Representative allele-specific information from GM12878 cells for selected assays around the first exon of the *NACC2* gene (genomic region Chr9: 138950000–138995000, GRCh37).

ENCODE Consortium (2012) *Nature* **489:**57-74

A (very) brief history of bioinformatics

QUANTITATIVE PHYLETICS AND THE EVOLUTION OF ANURANS

ARNOLD G. KLUGE AND JAMES S. FARRIS

Abstract

In the quantitative phyletic approach to evolutionary taxonomy, quantitative methods are used for inferring evolutionary relationships. The methods are chosen both for their operationism and for their connection to evolutionary theory and the goals of evolutionary taxonomy. As an example of this approach, a detailed analysis of a set of anuran characters is presented and taxonomic conclusions based on those characters are drawn. The methods and conclusions of the quantitative phyletic analysis are compared and contrasted with the methods of previous workers in the field of anuran classification.

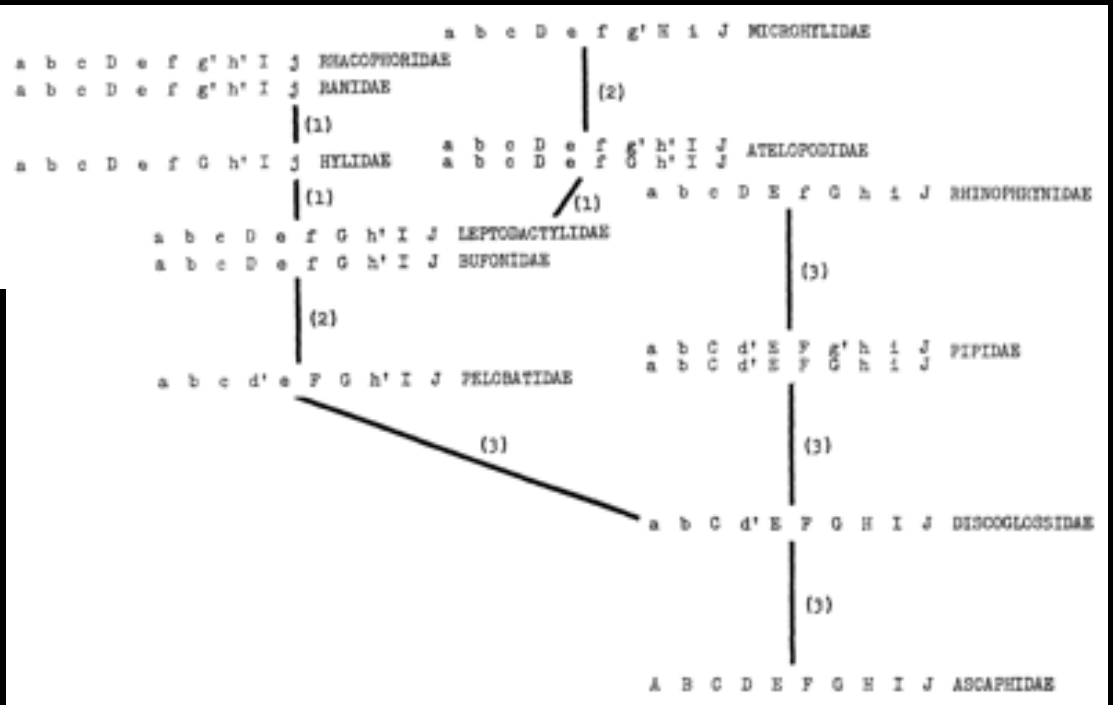


FIG. 3.—A maximum parsimony phylogeny of anuran families constructed according to the procedure of Wagner (1961). The 10 characters are those used by Inger (1967), except that all character states exhibited by the Ascaphidae are treated as primitive; see page 8 for further description of data modification. Compared to Fig. 1, this phylogeny is shorter by one evolutionary step, 19, and it includes one less homoplasy, d, and one extra reversal, D.

Kluge & Farris (1969) *Syst. Zool.* **18:**1-32

Benchmarking to biologists

- Benchmarking as a comparative process
- i.e. ‘which software’s best?’ / ‘which platform’
- Benchmarking application logic / profiling unknown
- Environments / runtimes generally either assumed to be identical, or else loosely categorised into ‘laptops vs clusters’

Case Study 1

aka

‘Which program’s the best?’

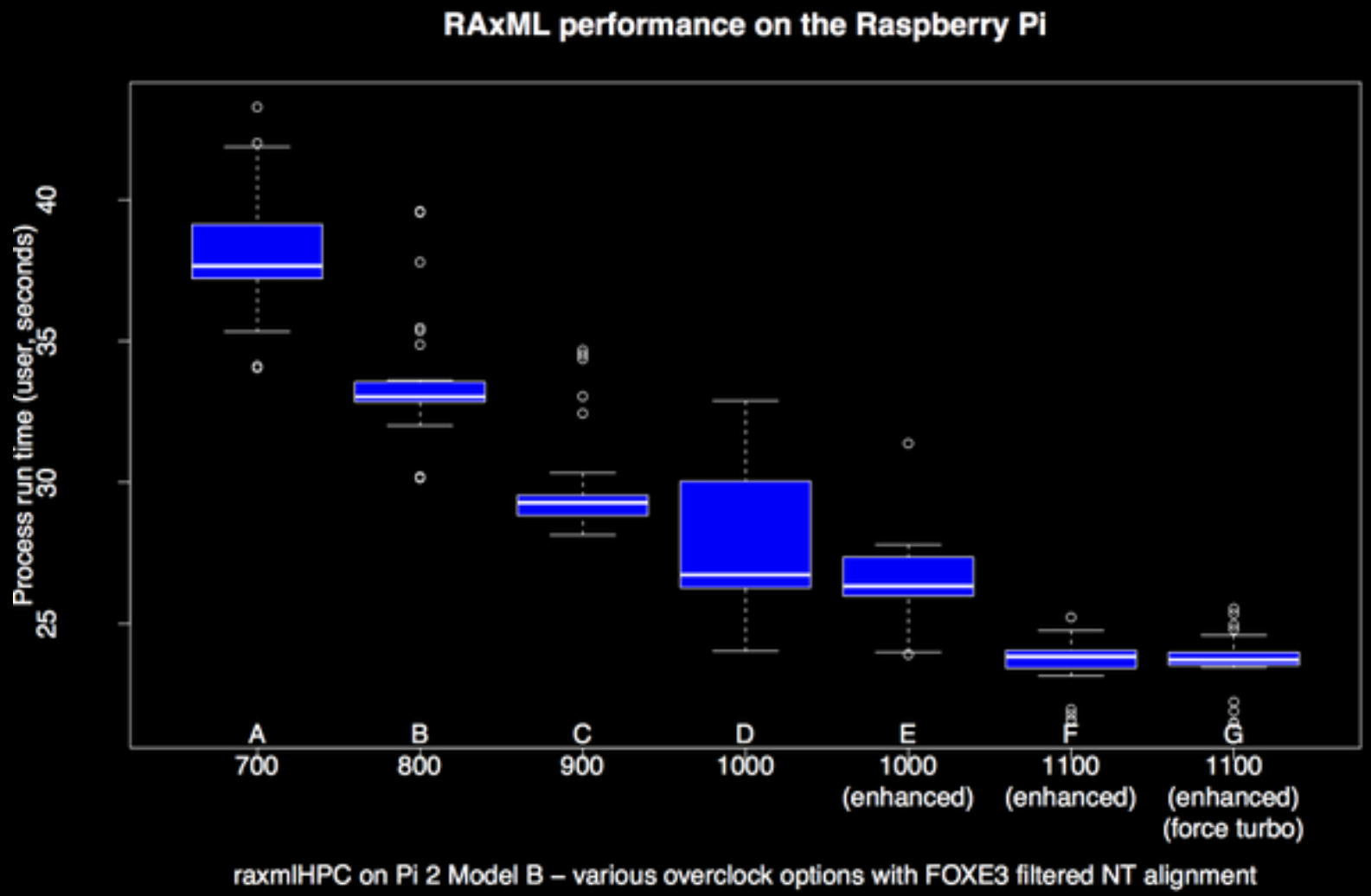
Bioinformatics environments are very heterogeneous

- Laptop:
  - Portable
  - Very costly form-factor
  - Maté? Beer?
- Raspi:
  - Low: cost, energy **(& power)**
  - Highly portable
  - Hackable form-factor

- Clusters:
  - Not portable, setup costs
- The cloud:
  - Power closely linked to budget (as limited as)
  - Almost infinitely scalable

  -Have to have a connection to get data up there (and down!)
  - Fiddly setup

Benchmarking to biologists

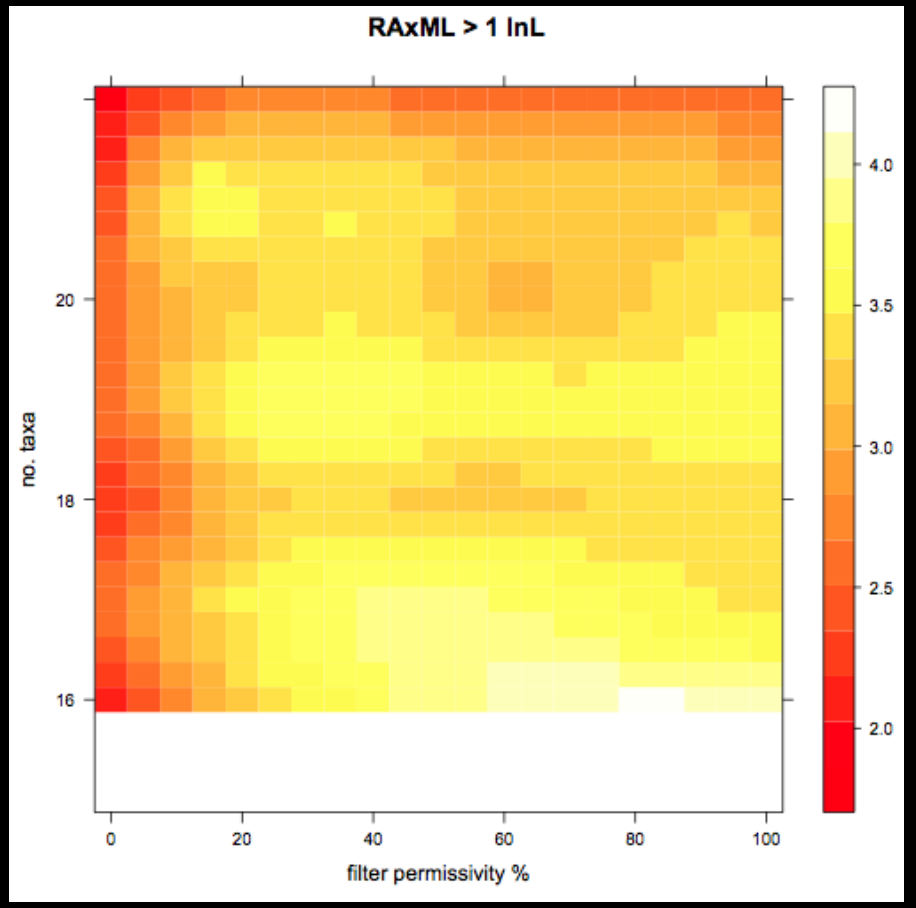


Comparison

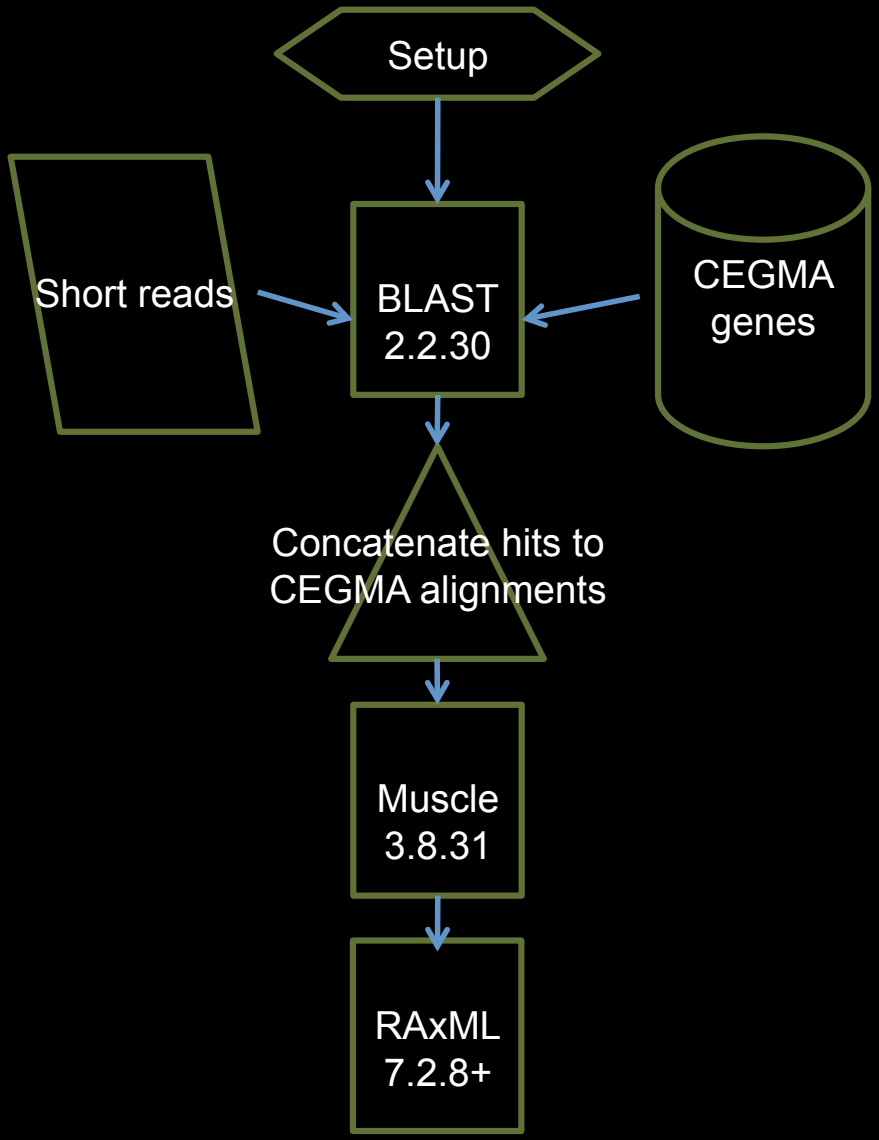
System	Arch	CPU type, clock GHz	cores	RAM Gb / MHz / type	HDD Gb
Haemodorum	i686	Xeon E5620 @ 2.4	8	33	1000 @ SATA
Raspberry Pi 2 B+	ARM	ARMv7 @ 1.0	1	1	8 @ flash card
Macbook Pro (2011)	x64	Core i7 @ 2.2	4	8	250 @ SSD
EC2 m4.10xlarge	x64	Xeon E5 @ 2.4	40	160	320 @ SSD

Reviewing / comparing new methods

- Biological problems often scale *horribly unpredictably*
- Algorithm analyses
- So empirical measurements on different problem sets to predict how problems will scale…



Workflow



Set up workflow, binaries, and reference / alignment data.
Deploy to machines.

Protein-protein blast reads (from MG-RAST repository, Bass Strait oil field) against 458 core eukaryote genes from CEGMA. Keep only top hits. Use max. num_threads available.

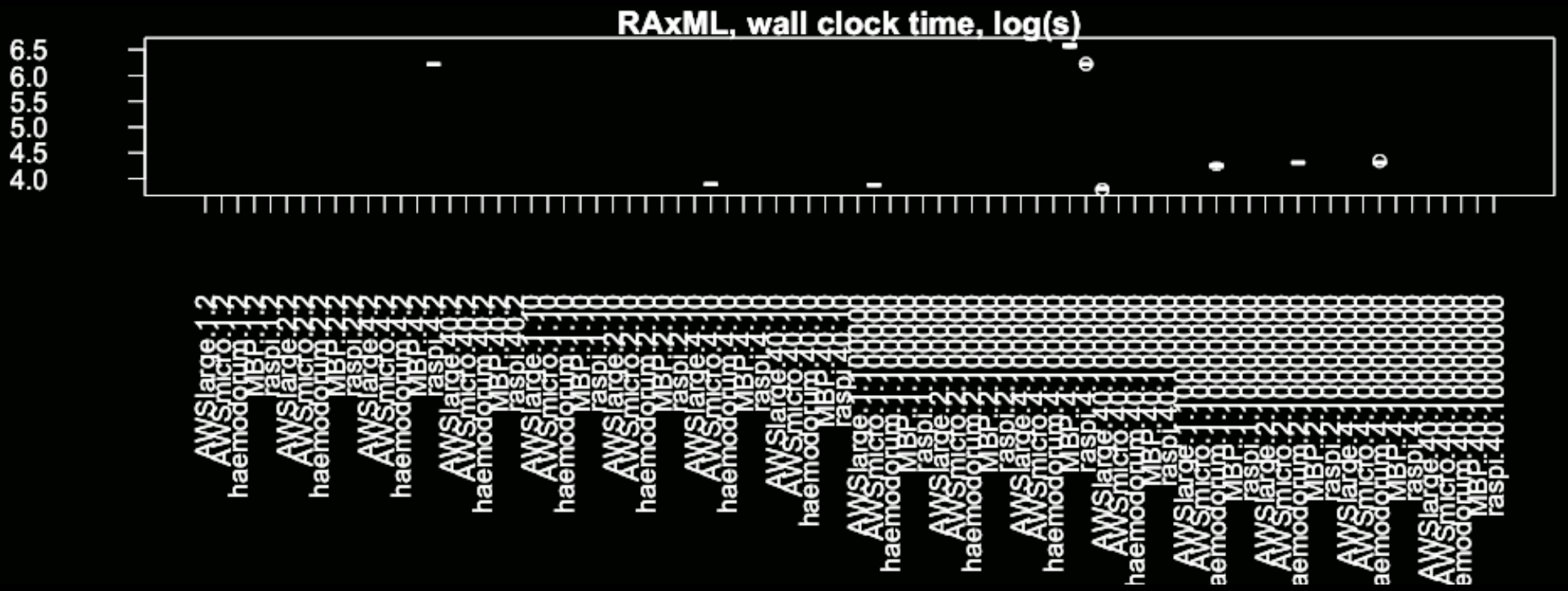
Append top hit sequences to CEGMA alignments.

For each:

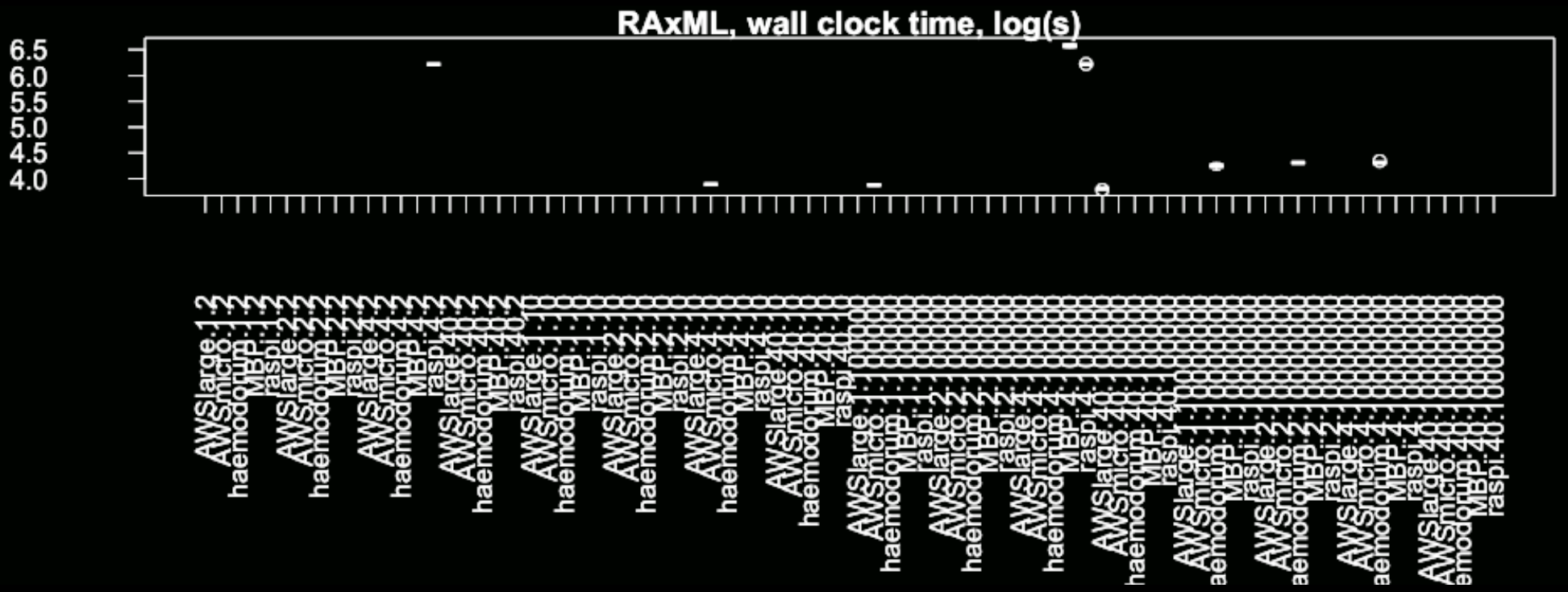
- Align in MUSCLE using default parameters
- Infer *de novo* phylogeny in RAxML under Dayhoff, random starting tree and max. PTHREADS.

Output and parse times.

Results - BLASTP



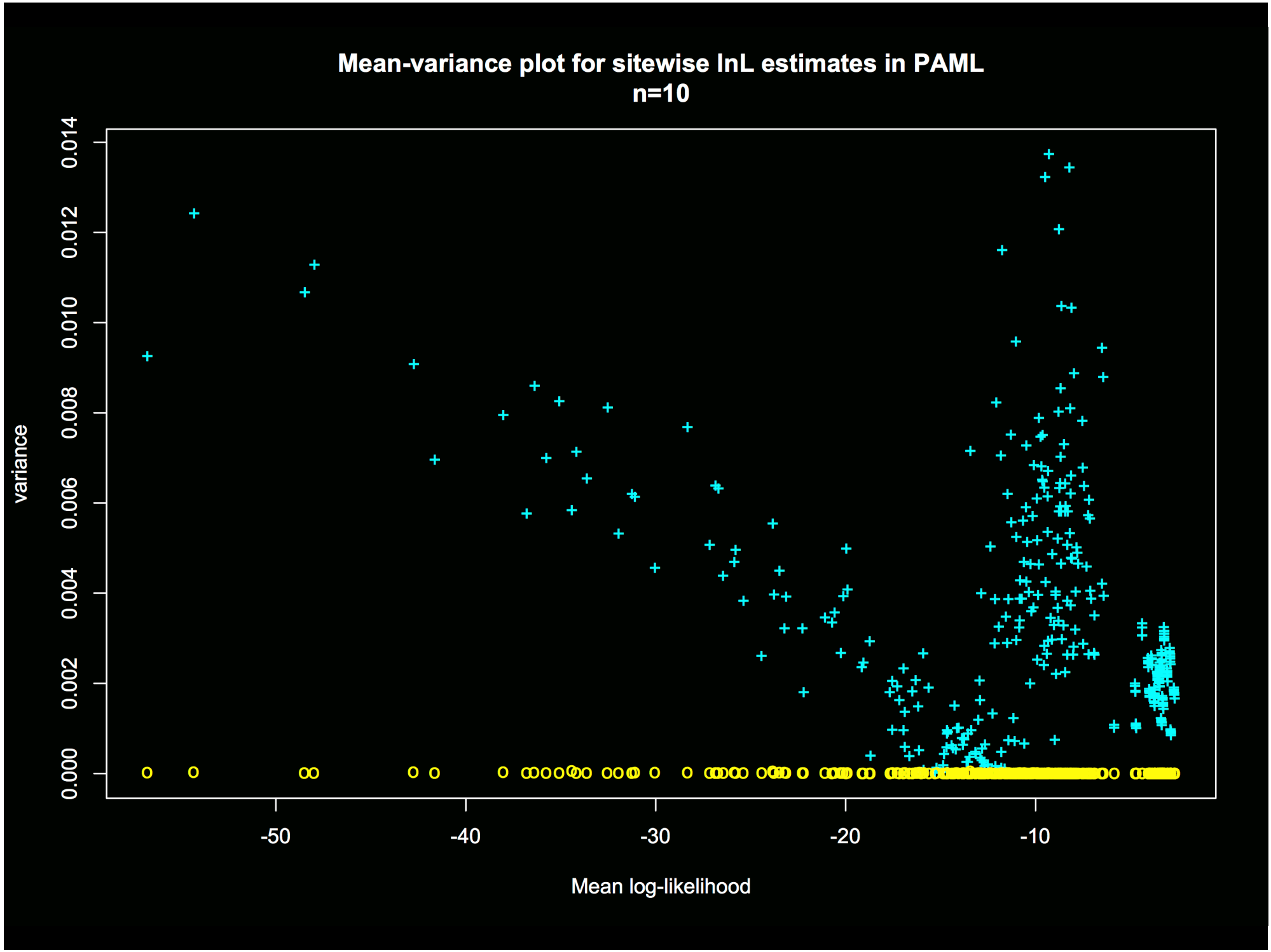
Results - RAxML



Case Study 2

aka

‘What the hell’s a *random seed*?’


Mean-variance plot for sitewise lnL estimates in PAML
n=10
variance
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0.014
-50
-40
-30
-20
-10
Mean log-likelihood

Properly benchmarking workflows

- Ignoring limiting-steps analyses
  - in many workflows might actually be data cleaning / parsing / transformation
- Or (most common error) inefficiently iterating
- Or even disk I/O!

Workflow benchmarking very rare

- Many bioinformatics workflows / pipelines limiting at odd steps, parsing etc
- http://beast.bio.ed.ac.uk/benchmarks
- Many e.g. bioinformatics papers
- More harm than good?

Conclusion

- Biologists and error
- Current practice
- Help!
- Interesting challenges too…

Thanks:

RBG Kew, BI&SA, Mike Chester