Quantifying Uncertainty in Evolutionary Analyses

Alexandros Stamatakis

ERA Chair, Institute of Computer Science, Foundation for Research and Technology - Hellas
Research Group Leader, Heidelberg Institute for Theoretical Studies
Full Professor, Dept. of Informatics, Karlsruhe Institute of Technology

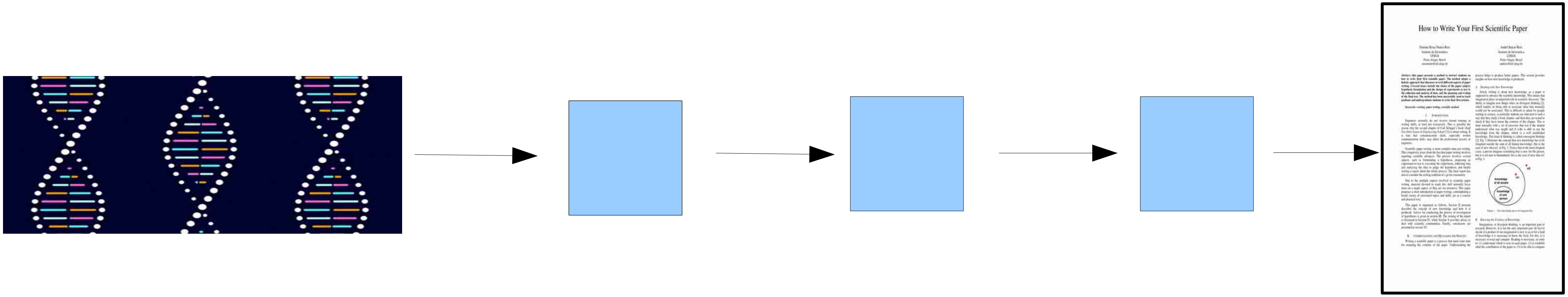
www.biocomp.gr (Crete lab)

www.exelixis-lab.org (Heidelberg lab)

Outline

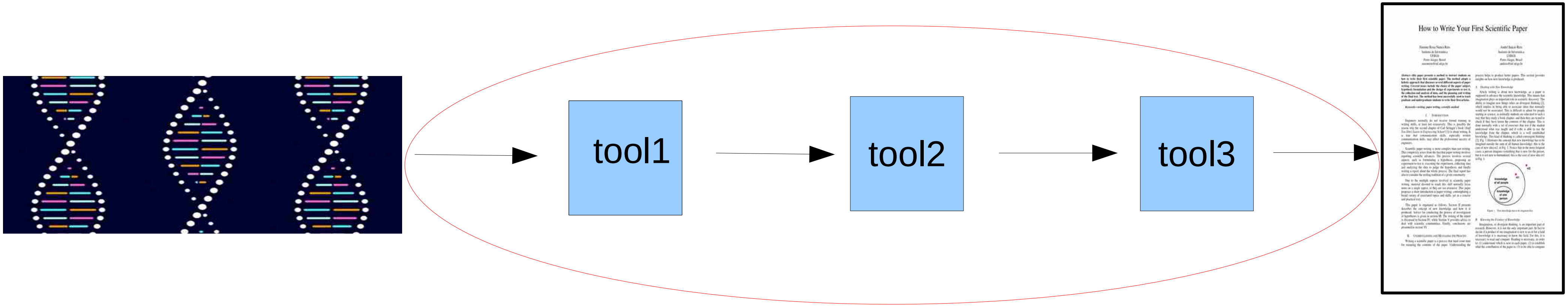
- **Our Approach to Bioinformatics**
- Introduction to Phylogenetic Inference
- Sources of Uncertainty
- Phylogenetic Difficulty
- Other stuff we are working on

Bioinformatics



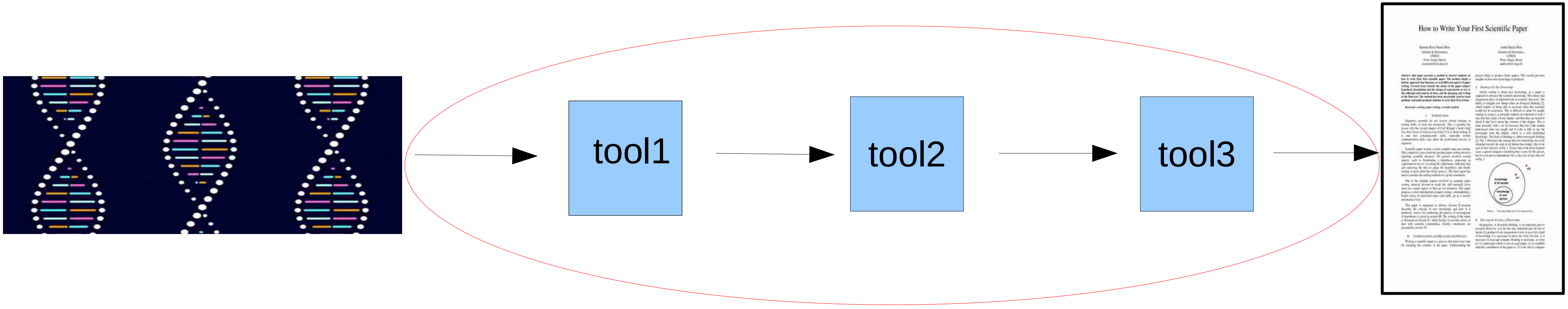
How to Write Your First Scientific Paper

Bioinformatics

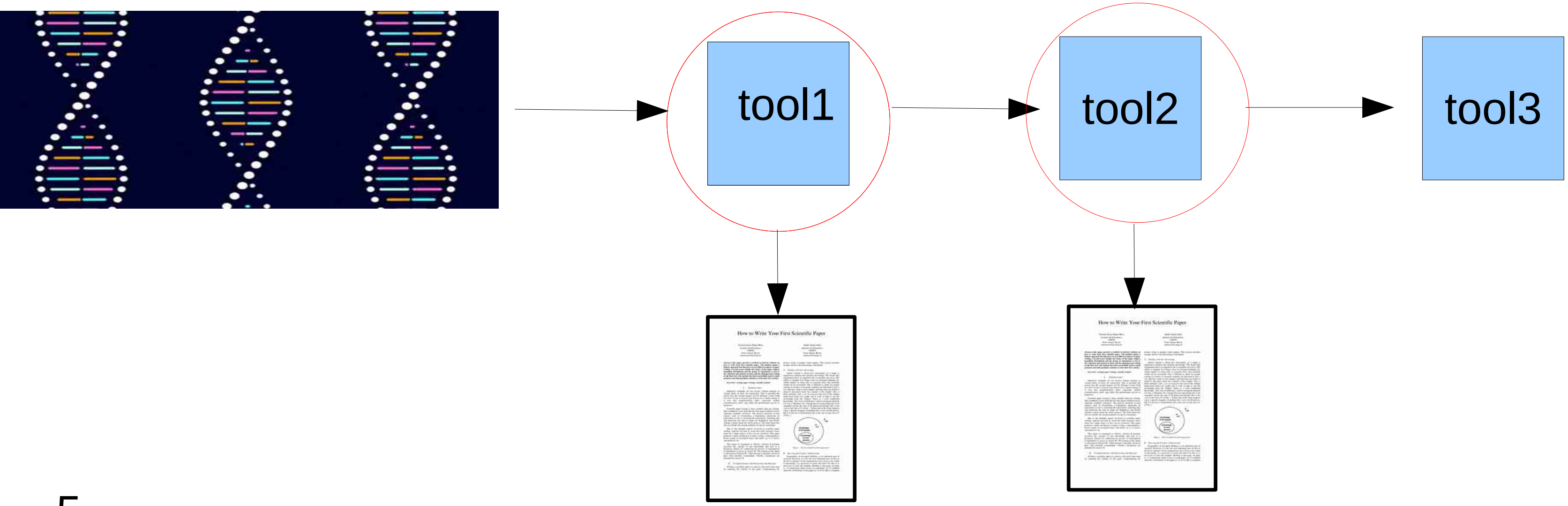


Data-centric: pipeline building

Bioinformatics



Data-centric: pipeline building



Method-centric: tool building

Our Approach

- Focus on *core* tool, model, algorithm, and method development
- Method development better fits the research interests of a computer scientist
- **Goal:** Enable Research in Evolutionary Biology

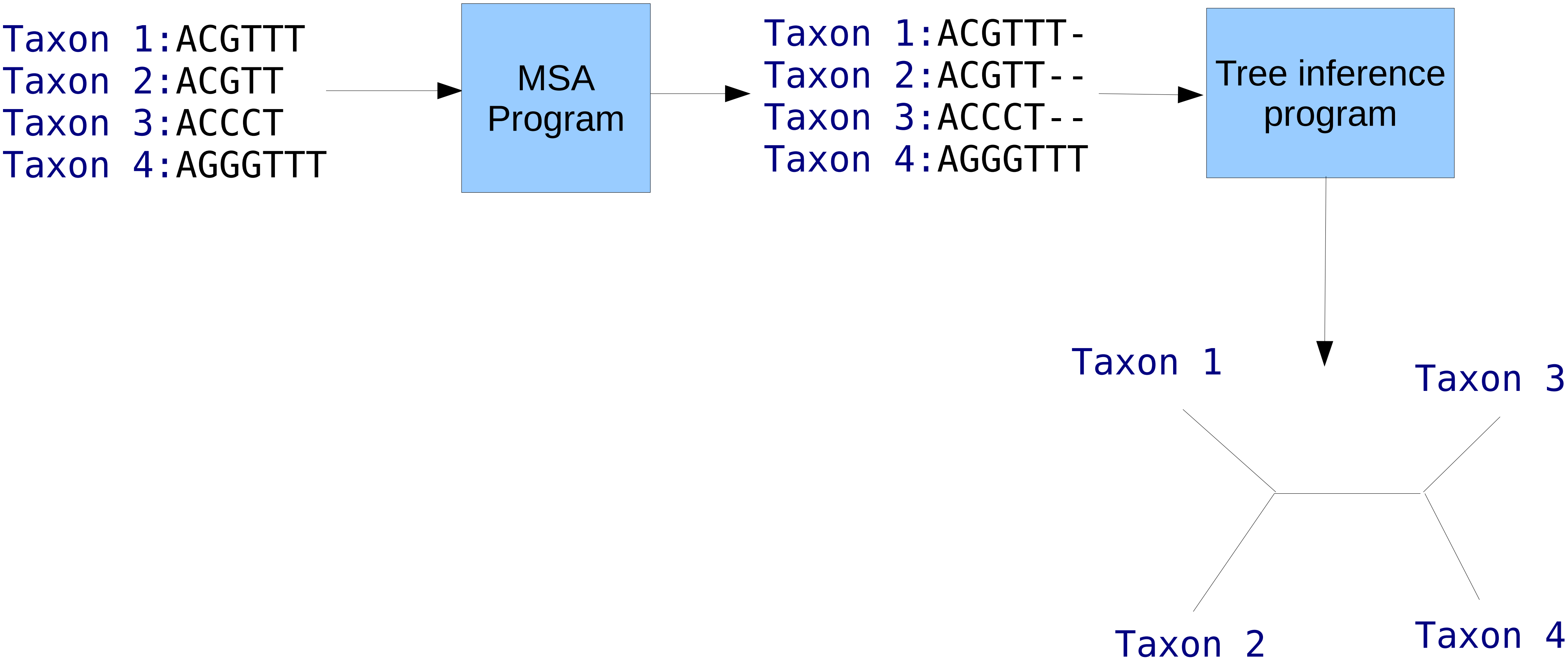
Our Approach

- Focus on *core* tool, model, algorithm, and method development
- Method development better fits the research interests of a computer scientist
- **Goal:** Enable Research in Evolutionary Biology
- Nonetheless, we often conduct data centric research in side projects

Outline

- Our Approach to Bioinformatics
- **Introduction to Phylogenetic Inference**
- Sources of Uncertainty
- Phylogenetic Difficulty
- Other stuff we are working on

Tree Inference Pipeline



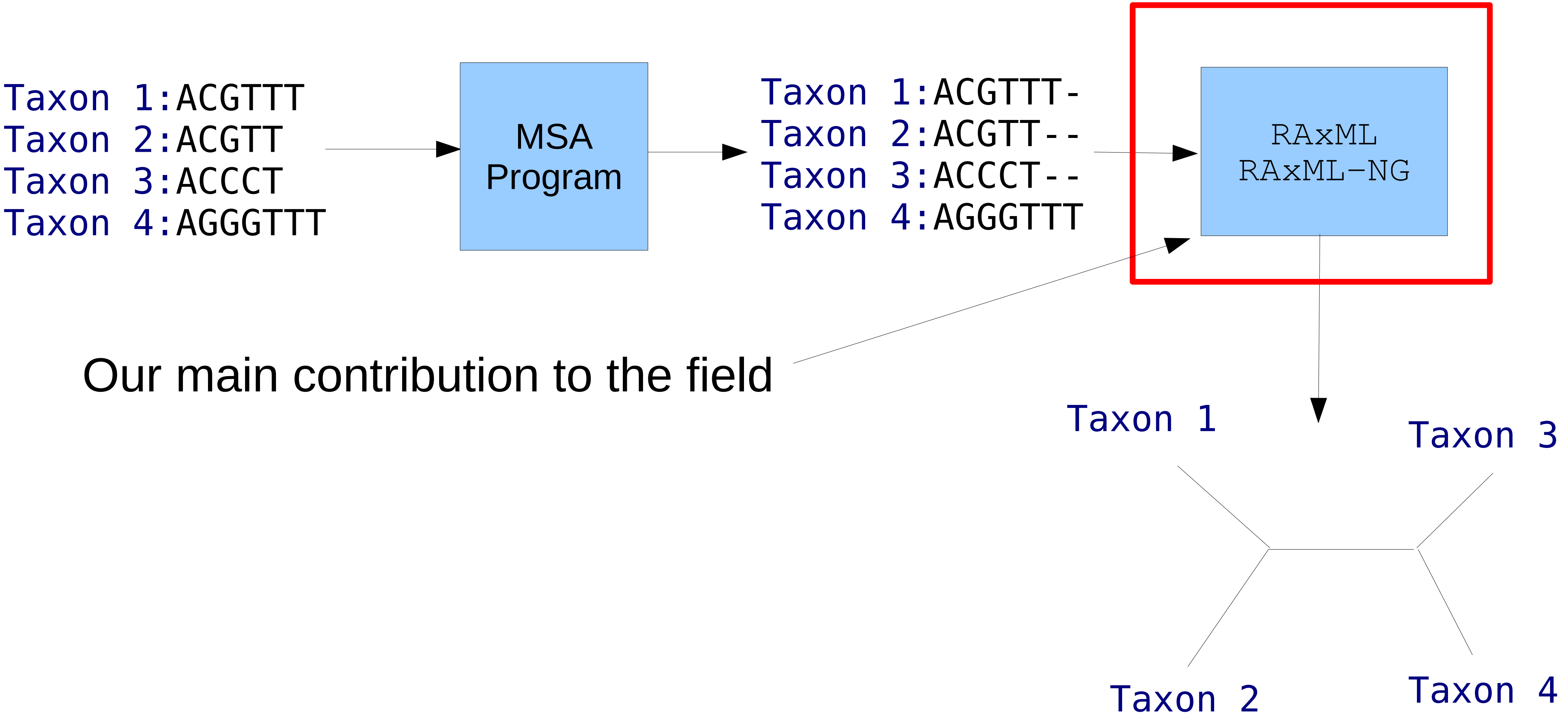
Taxon 1:ACGTTT
Taxon 2:ACGTT
Taxon 3:ACCCT
Taxon 4:AGGGTTT

MSA Program

Taxon 1:ACGTTT-
Taxon 2:ACGTT--
Taxon 3:ACCCT--
Taxon 4:AGGGTTT

Tree inference program

Tree Inference Pipeline



Taxon 1:ACGTTT
Taxon 2:ACGTT
Taxon 3:ACCCT
Taxon 4:AGGGTTT

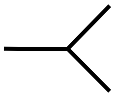
MSA Program

Taxon 1:ACGTTT-
Taxon 2:ACGTT--
Taxon 3:ACCCT--
Taxon 4:AGGGTTT

RAxML
RAxML-NG

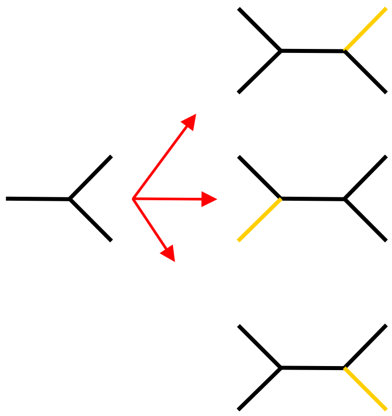
Our main contribution to the field

The number of trees



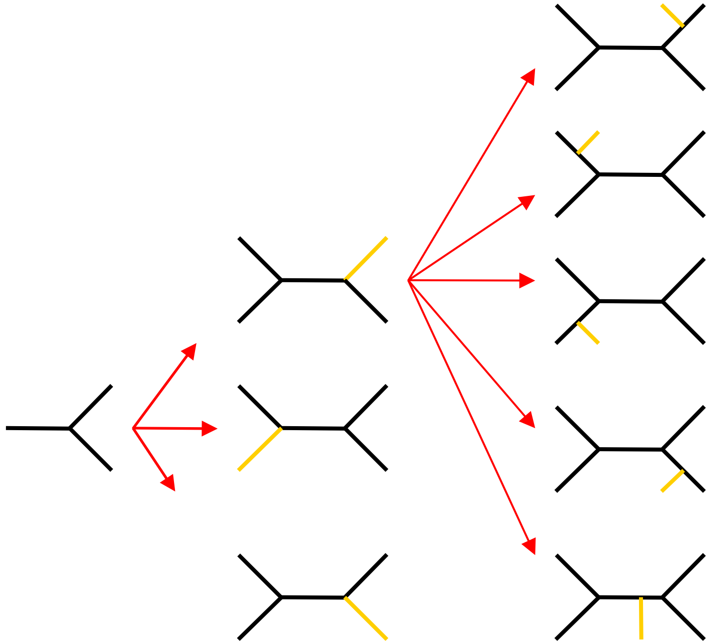
3 taxa → *1 tree*

The number of trees



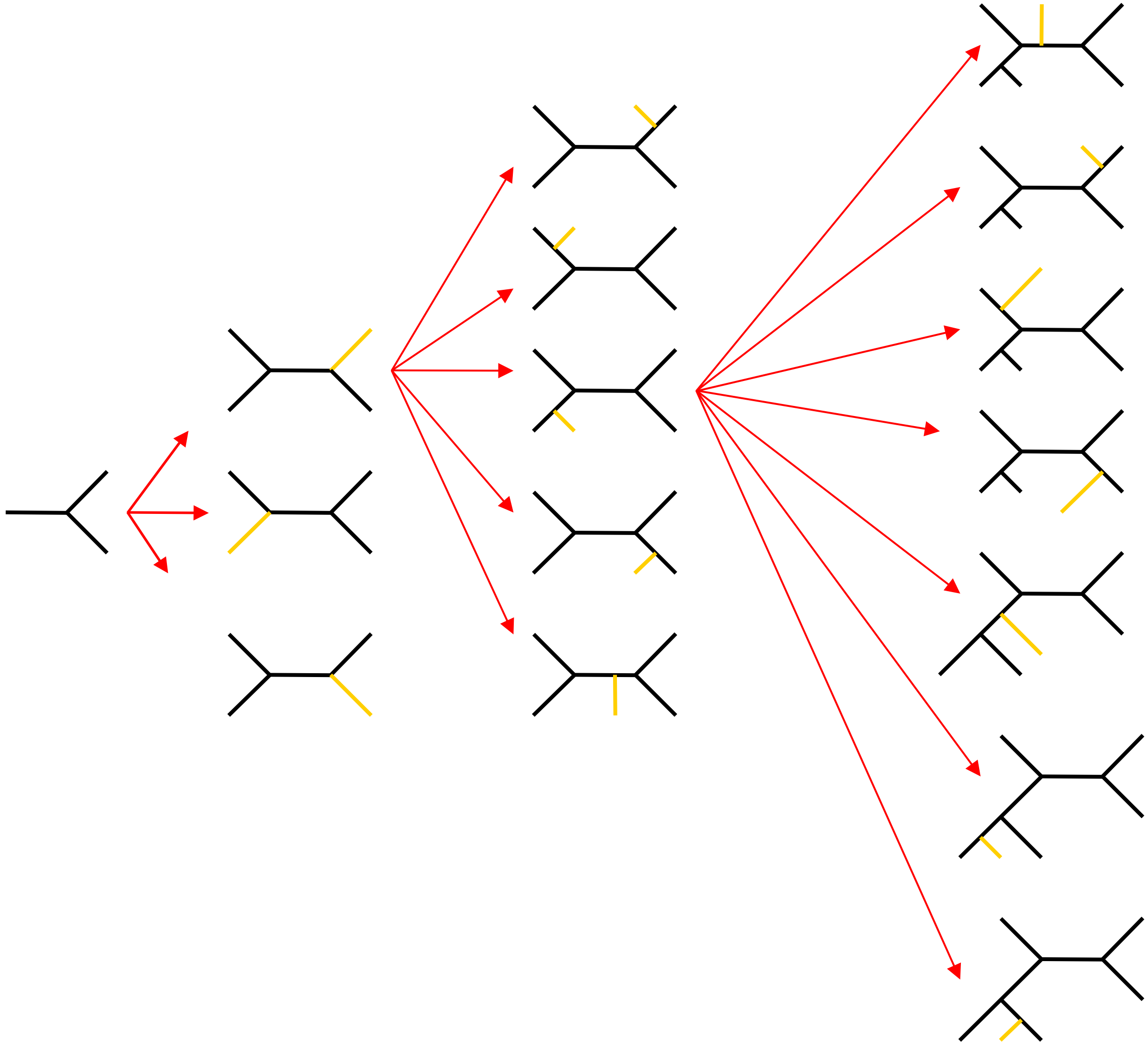
4 taxa → *3 trees*

The number of trees



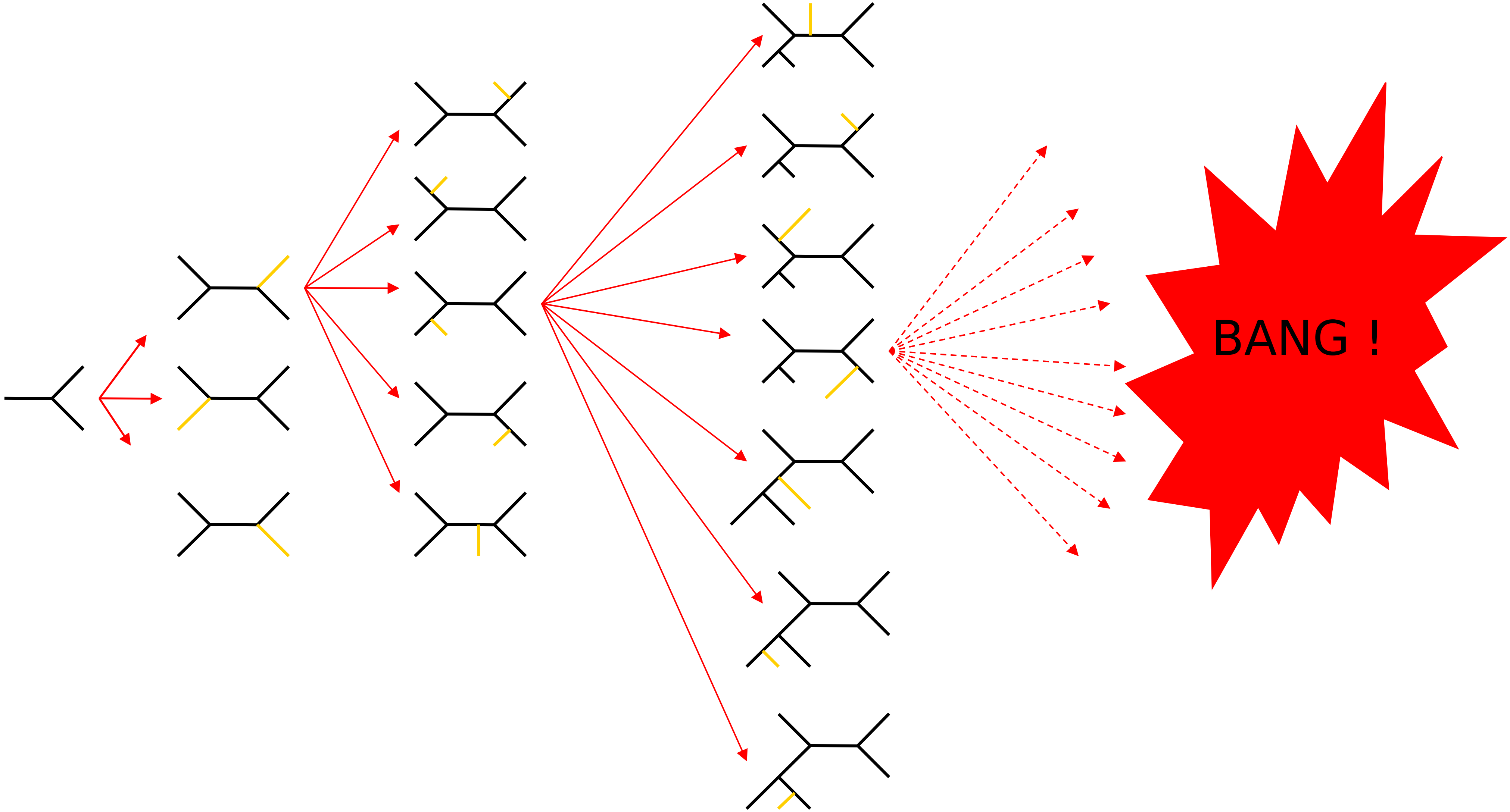
5 taxa → *15 trees*

The number of trees



6 taxa → *105 trees*

The number of trees explodes!



possible trees with 2000 taxa

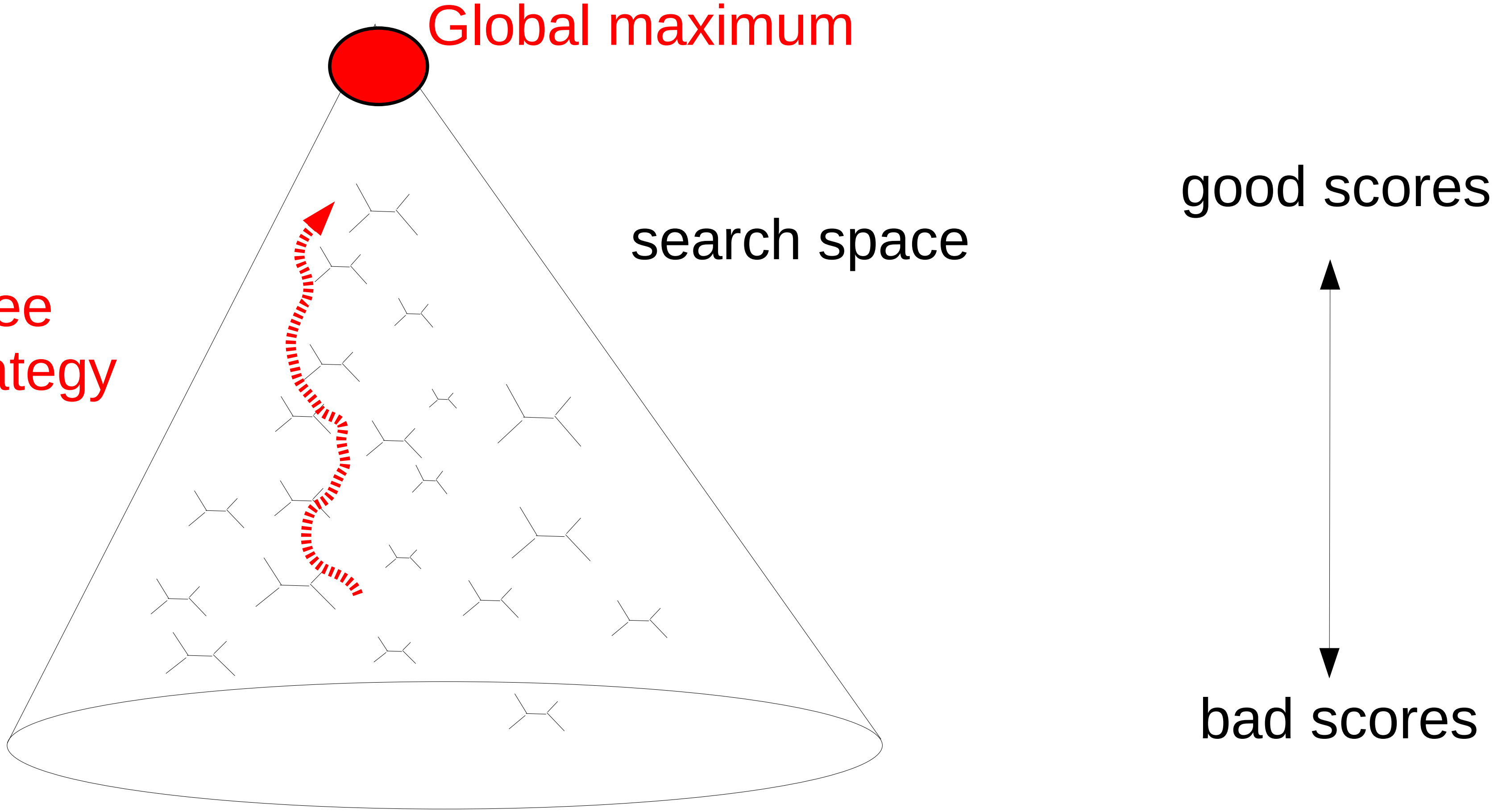
```
stamatak@exelixis:~/Desktop/GIT/TreeCounter$ ./treeCounter -n 2000


GNU GPL tree number calculator released June 2011 by Alexandros Stamatakis

Number of unrooted binary trees for 2000 taxa: 30049638174211656151632910065681814981377232074237013089504954043012636525258308210827685996688247000464352735214265634288295
8915023446000631493969130632970436056184861877465482277991223536809233455563199910834597693126756525012899867433187752811401960991631522367030609121735709762379847705467667
7795324797182614385273338226727784250737252849916669687584403510579587020686505817687044666318123742901021438506432471360934491667021135969756940300666252646479269124551031
4942366195542824118277625114848758254581227914289801132648902674033761294712745767036267579086843169660718609847941818865957214557044744572288661729053583520744253688123124
0106613156948861960941195646736200342575241335277575085829161096422575727699767991408283343210161327401652830993803904592327690690035972919709940739349563486203899010742687
2822975974655377102257672676842858011877224950106218117340523208265397342962227352536590515865631383272031119841987467599738646318290320383252308597997992216101227215780805
2481458312068440167606239306009711616729715504728487799634337531348994230372437347879131989085953764070134849446113877572576952408702461720107874297380462275052545706689372
3194182064407068918840038705902897721975164544959758216621306205064617761099485663734168183584989329076993382067801052437284614924034229611551826097782286191926720712951895
8936009959130974233072316382518428110330571017441156884305131865877544376308500311451110723837039707465182232040406154708273078629957549331031275208616700660791298014262230
0565123522718063819509335872651728623589020520016144361756075654286471422126613004434807084067501589247673166341539540575074474994909831496473031080411401891849735912811228
3787740498848340562102420566424463860093899650857429619472690543015281237526510965815284699797036792171112903556809818079169587951614159281049528179855847292534447864424435 9
9808531537204796814969465991768614533701051985928577157482455943377242369582576242663016946320482495182255939287403177623433881048604630975191556923871167513095213415098816
7154643078623526062377864068386804246902527491139319276802611515990582603886733172930713673903403618637463980605764836474670274446727880885337074254421922726677747003329403
3201038288035112689026255183096791948358678929370163768175304820633894387149793115235369822962511116307148294599211620803302684762013335690441089668145436150905155877581167
9770012563912151116237444170497371704604029481104114822286466131918821997571383368352072526055202769823974613218495249264897050790398360256255606289852288839561357874156576
6488999260873286612630642543260248979229113560071640573984516375245243376943755857384725545564397599604255914640112221144755235573176239973057747183956531217416532295986675
9012941161239240722093250369673124884491553759210650656015416720774159236240868667675348286512964888739059707578802473393463470848159011639772797747480417316268700916728735
6121642268468160683198959801260376485615312781611689587215123123308760063473381097253118423339640390937378395066835578735307886358646400563299499490631187424029092779272693
3003224453775957972248734568915114585570783850541681667667425811301958063621907500790295031088209097271748136436989473971079932777700676301730617566538739726037777173008441
3439405123669055544932486165082539957795036326704947844293498853172797348177779714656717515117887639643406933245807634611073421432819504990968087402739768891470451747205 5543
8969396668742601477241894693212902453317334188286773194653544113302100866575081713240342647580489218623663461607955943720516395154096949806486242309796947211169665961580041
7883992232264628498942352276926391124360767508997671789683459337890742346354557193530665615379900277916265636361861974048590938234062235459769733137213659343717585590664439
6461328300113672601934068706442339489199215304385281541659630118549423634863524585746642834609062916279564926584723003608555598989761916293248140094592489899468468862322586
0170551468905649837276003927487069550463788819741699429049102853180410776516007263384721638903782270013843599599730265727198804324664312359758555217196960513921010226596324
7887830977405333131551762978152880718652603176327264828094499370456258099938053058497699570080289379808014902901005293898472279947167804821689424159118284257696464786505731
5325178302336307298251692210346584265894447464916123854689718507968172909139032182834111184821384767728316548653212317382004131990510518967022201887049585687180509590730360
6930402937216038968917605587676955382318093705826257083898387409098468656634271397500013291835105943321729879825243707508272087959859437157667660155782699660343197752623308
8989962587800628009560944416932377944955441033696586261556256010669390303203878970983673786087056641433585106111658314520424513208508589994932364831689671194951671619567622
7070906973889588855579562466641536561723549301807394004760529801721771391686788000277851966173070061284517307582503735643102065112443730825229625040453160590741343881872563
4779138306605909318802522310085340176840261401539616989192075147108033757708849740141834599753972059878682064879116064969858177601153972058498222698907181349432691801821173
3188063653910893689811714891357456680542807485170175858266639633570189354449832669762835092657922201746372190273119641751489944010079636876017826747107019945473218887832742
6088966724371574713420600093704251309893630537459784279980403132989417266492290425730958368534416215640557290282066224003863237526380910233269897838860423759625601567975262
6950798639868104294832333160267216555178120899264677804935741326387137408423885546538336158643451305439624281397279559725995110706314305992615495622958320232708057681156690
4895866105220300573725298472118747827136713666058669271094875563974858489475910819727033878284439864486743456200958161930314727345961900499318424337975243662489363321244850
5971992523668529249305346252764137853413208943128901523738092556045987090912766662329678703328882059134949580074074473143388800724532321747309659741967114441453127132790205
1010047671014350638857953478447255389801541923317027519896180635152682543173193832925891931530164130548972311128666465492971930479296432829556719092881692091042334122007454
2420499008725850462080511048758830594959903111887366685094148821725734576355233964038481318213167408359006916400053262258184783765067804451177717328658189899215358309447765
350341796875

Approximately 3.00 times 10^6328
```

Problem Complexity



Global maximum

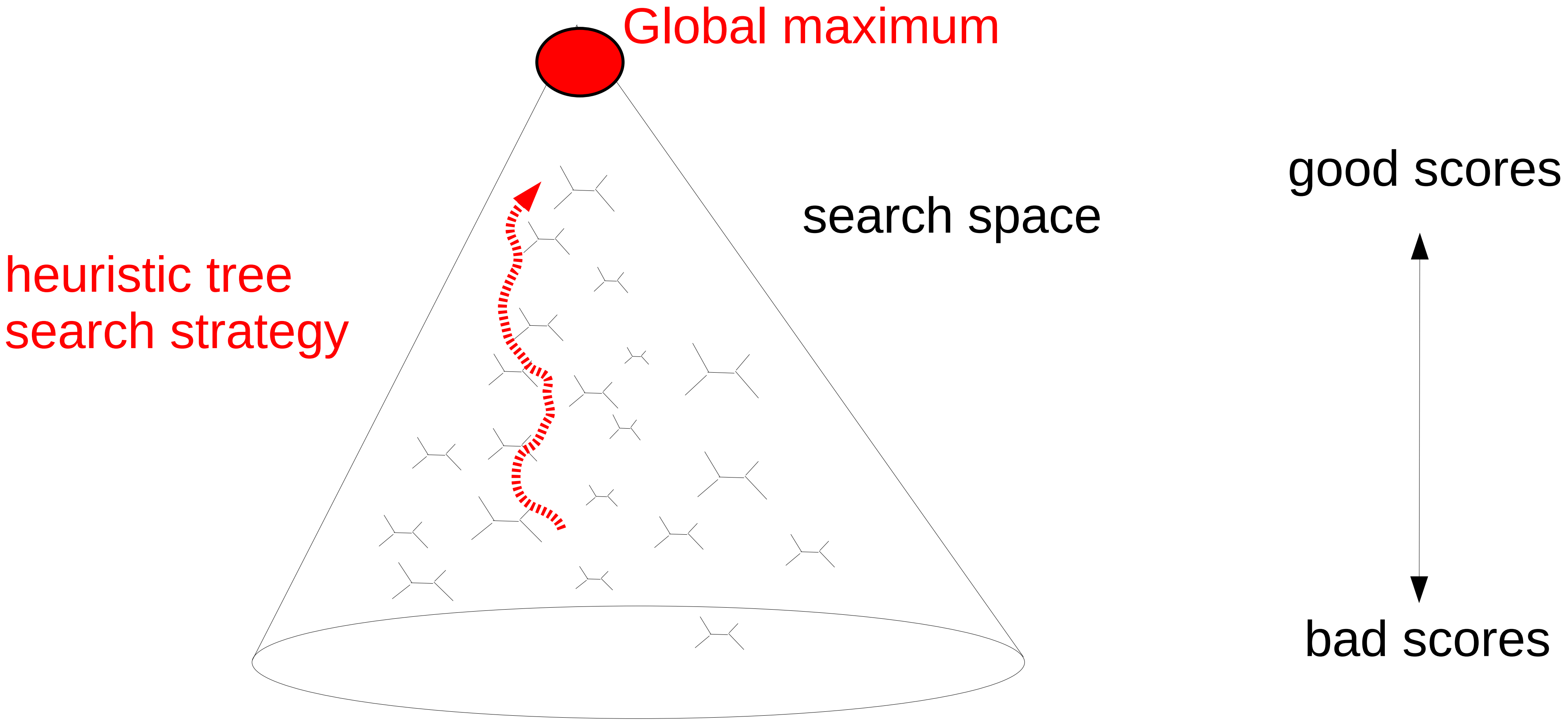
search space

heuristic tree
search strategy

good scores

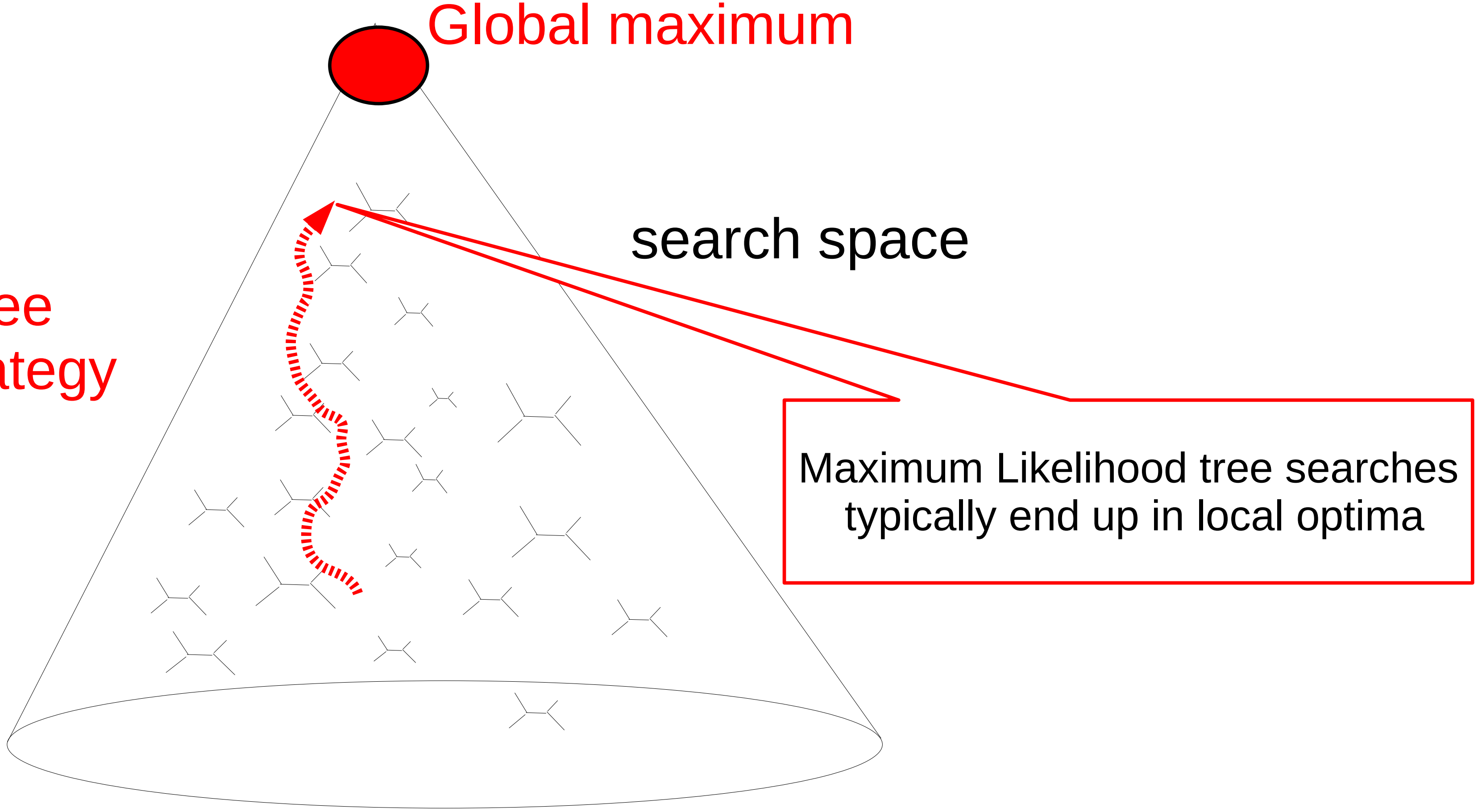
bad scores

Problem Complexity



Finding the best tree under Maximum Likelihood is **NP-hard**!

Problem Complexity



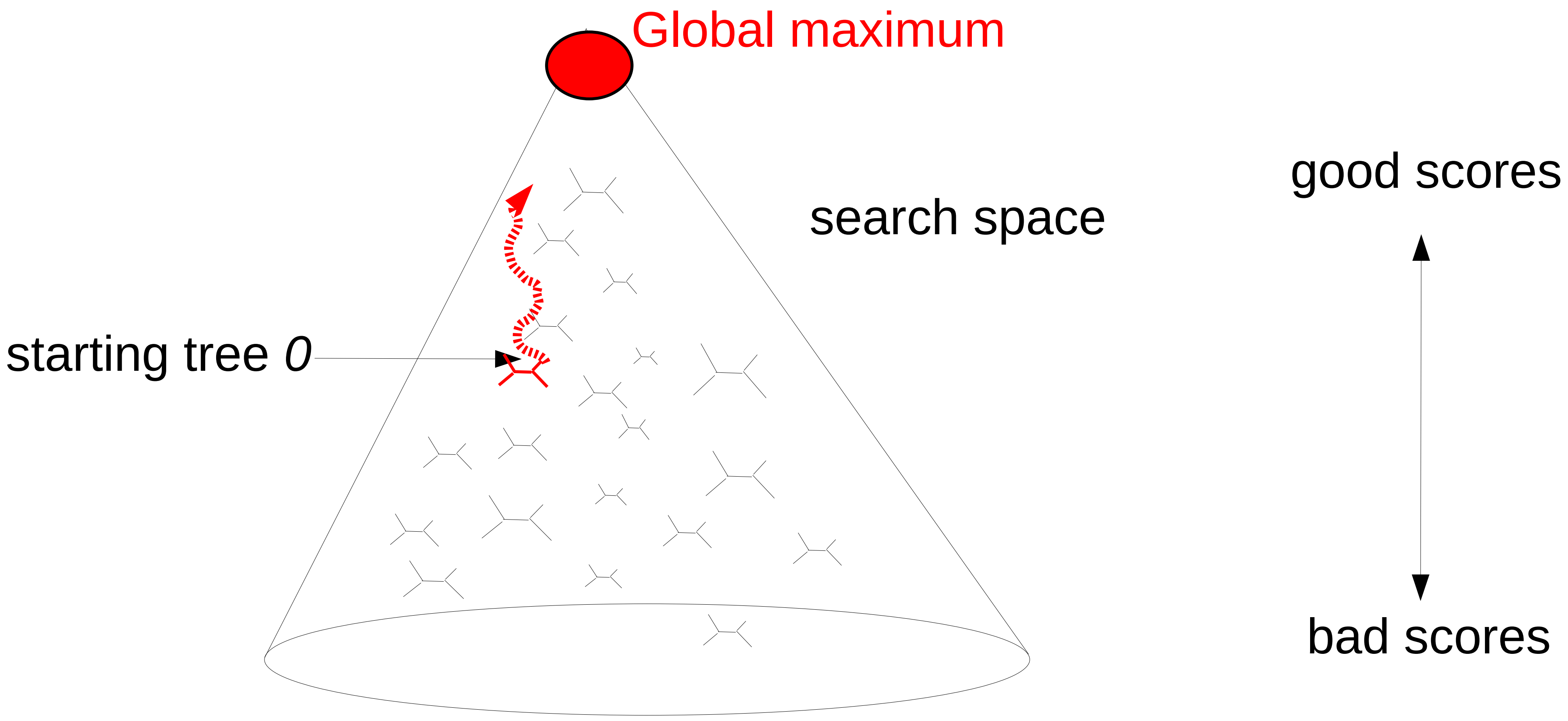
Global maximum

search space

heuristic tree search strategy

Maximum Likelihood tree searches typically end up in local optima

Starting Trees



Global maximum

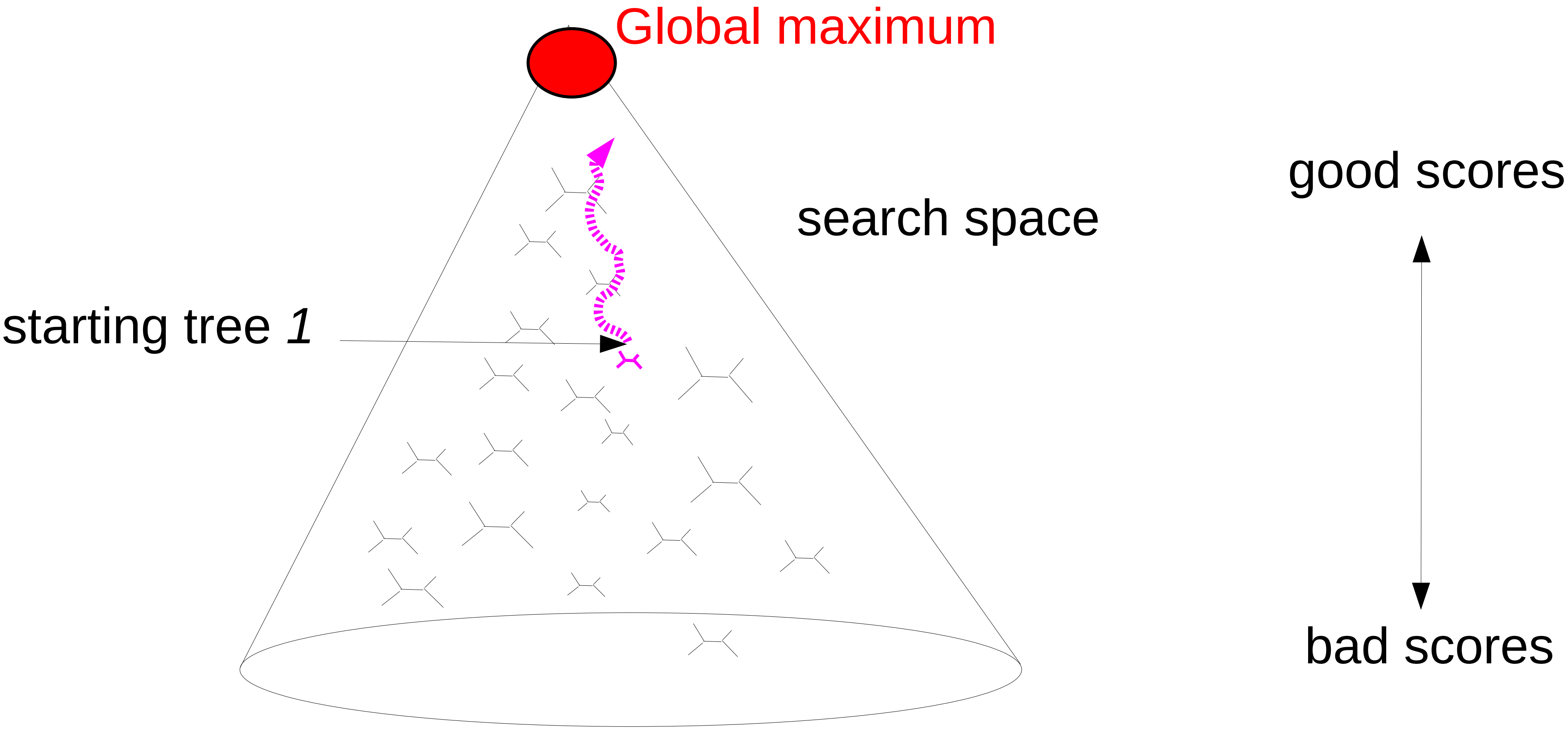
search space

good scores

bad scores

starting tree *0*

Starting Trees



Global maximum

search space

starting tree *1*

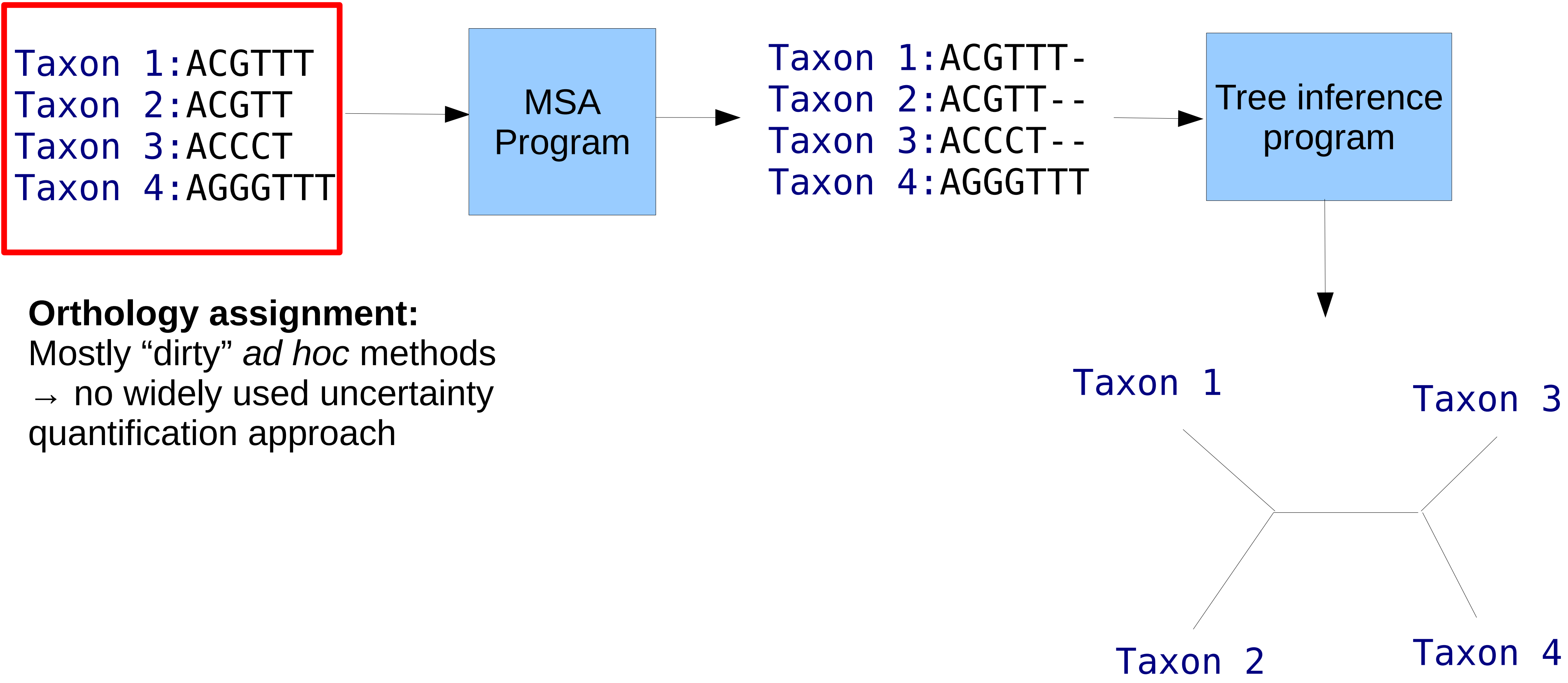
good scores

bad scores

Outline

- Our Approach to Bioinformatics
- Introduction to Phylogenetic Inference
- **Sources of Uncertainty**
- Phylogenetic Difficulty
- Other stuff we are working on

Tree Inference Pipeline



```
Taxon 1:ACGTTT
Taxon 2:ACGTT
Taxon 3:ACCCT
Taxon 4:AGGGTTT
```

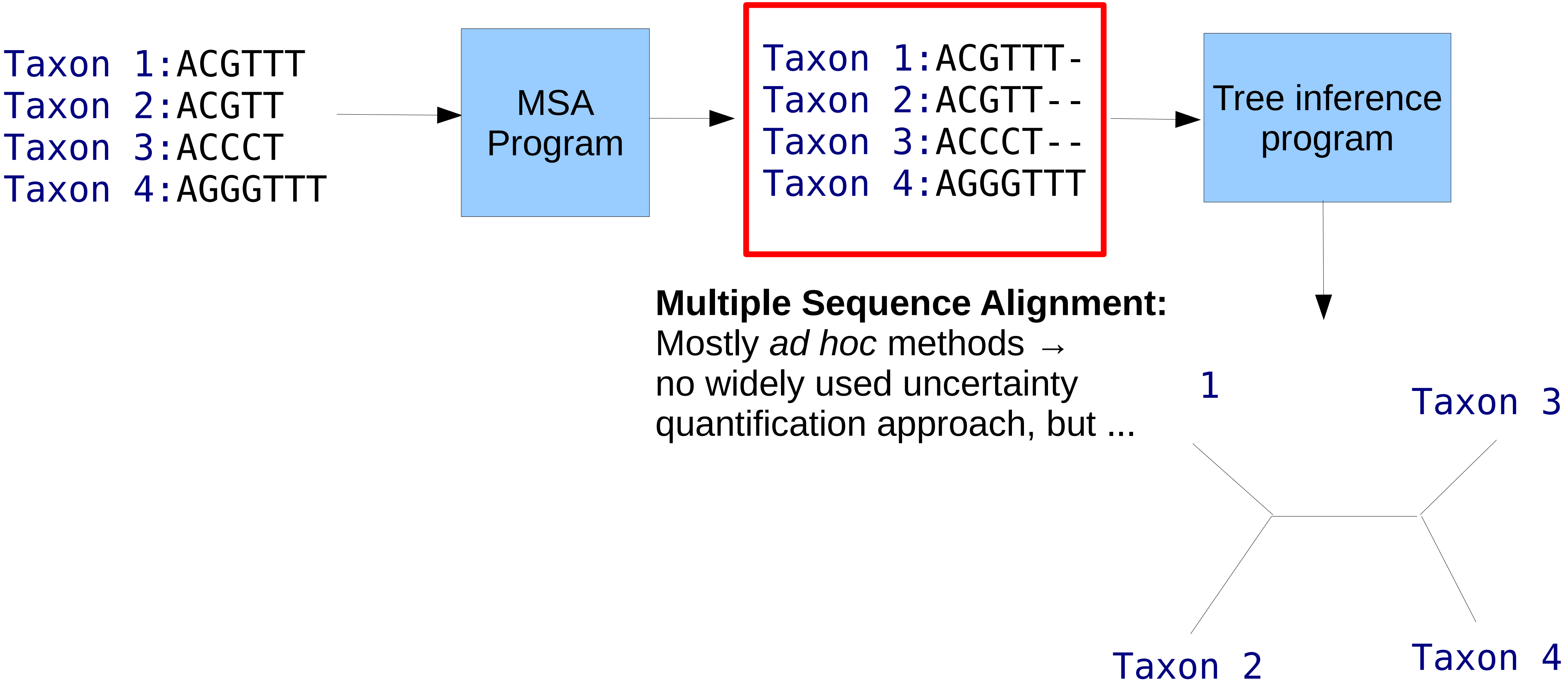
MSA Program

```
Taxon 1:ACGTTT-
Taxon 2:ACGTT--
Taxon 3:ACCCT--
Taxon 4:AGGGTTT
```

Tree inference program

Orthology assignment:
Mostly “dirty” *ad hoc* methods → no widely used uncertainty quantification approach

Tree Inference Pipeline



Taxon 1:ACGTTT
Taxon 2:ACGTT
Taxon 3:ACCCT
Taxon 4:AGGGTTT

MSA Program

Taxon 1:ACGTTT-
Taxon 2:ACGTT--
Taxon 3:ACCCT--
Taxon 4:AGGGTTT

Tree inference program

Multiple Sequence Alignment:
Mostly *ad hoc* methods →
no widely used uncertainty
quantification approach, but ...

Muscle5

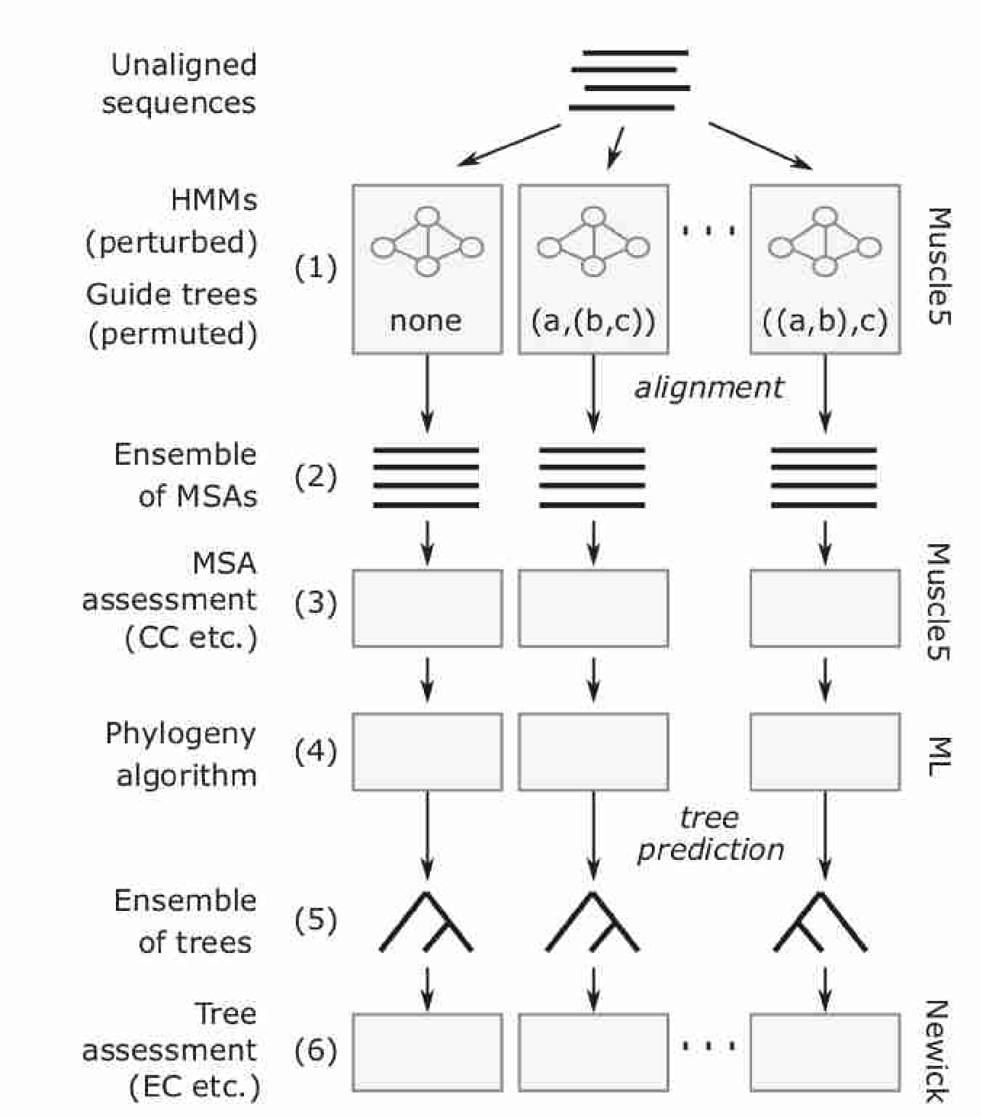
Article | Open Access | Published: 15 November 2022

Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny

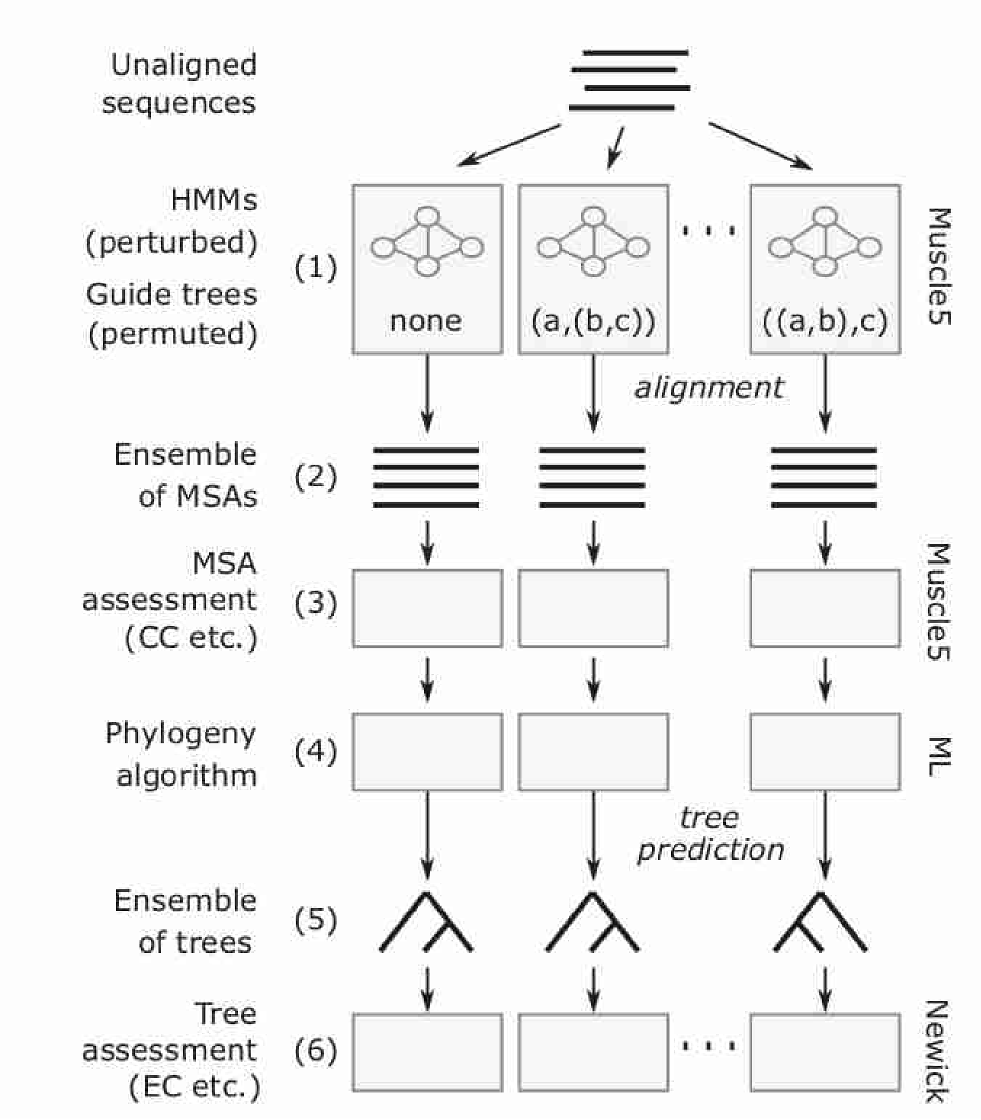
Robert C. Edgar

Nature Communications **13**, Article number: 6968 (2022) | Cite this article

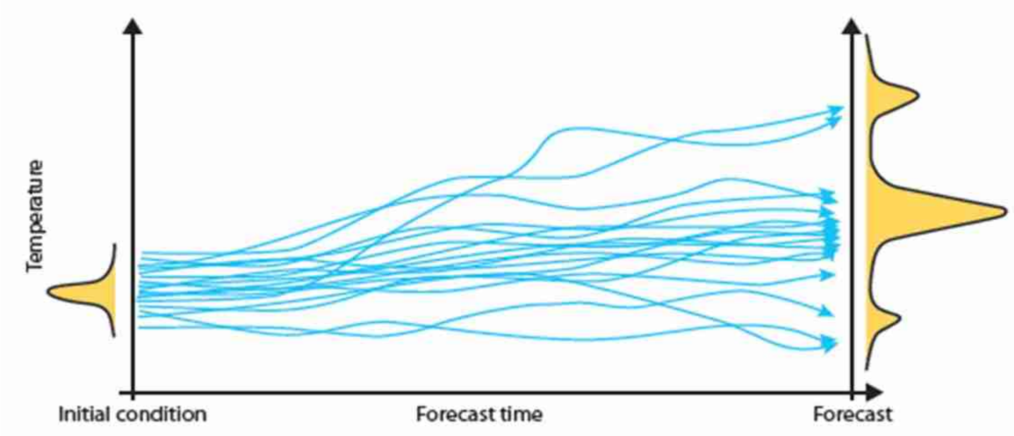
Muscle5


Unaligned sequences
HMMs (perturbed)
Guide trees (permuted)
(1)
none
(a,(b,c))
((a,b),c)
Muscle5
alignment
Ensemble of MSAs
(2)
MSA assessment (CC etc.)
(3)
Muscle5
Phylogeny algorithm
(4)
ML
tree prediction
Ensemble of trees
(5)
Tree assessment (EC etc.)
(6)
Newick

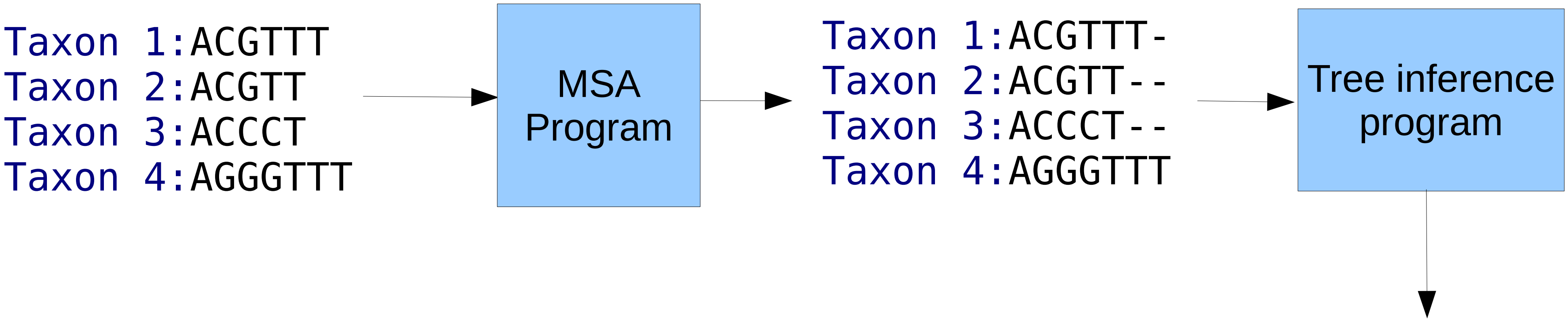
Muscle5



Temperature Ensemble Forecast



Tree Inference Pipeline

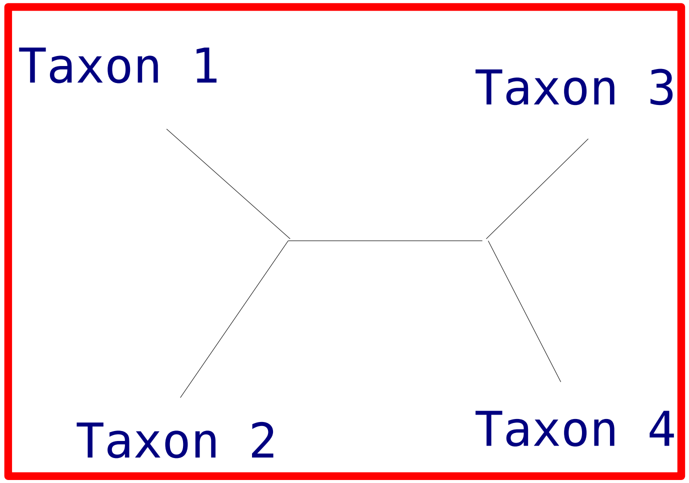


Taxon 1:ACGTTT
Taxon 2:ACGTT
Taxon 3:ACCCT
Taxon 4:AGGGTTT

MSA Program

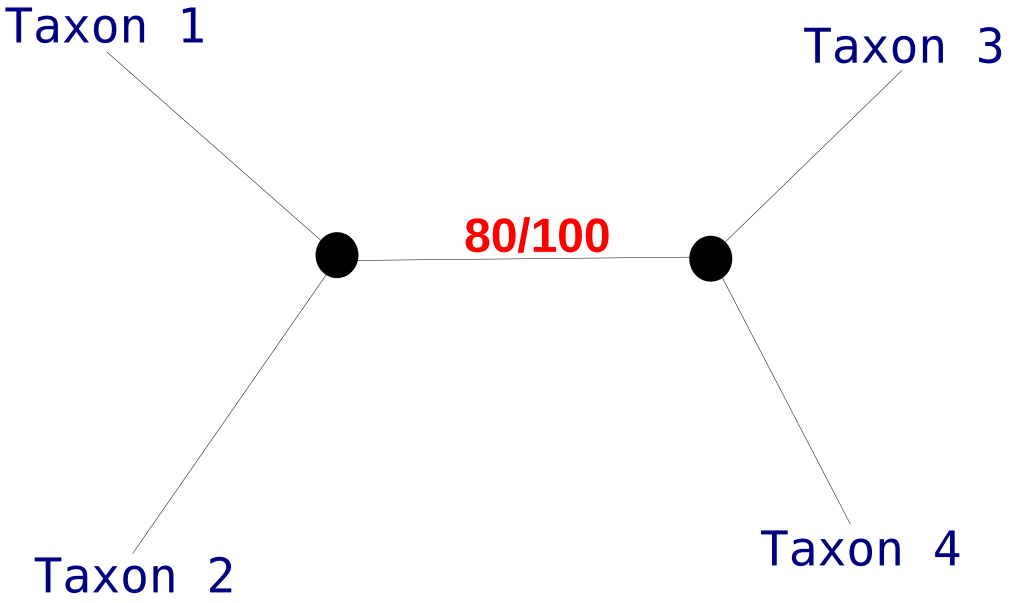
Taxon 1:ACGTTT-
Taxon 2:ACGTT--
Taxon 3:ACCCT--
Taxon 4:AGGGTTT

Tree inference program



Phylogenetic Inference:
A long history of explicit uncertainty models
Bootstrap Methods for Maximum Likelihood
Posterior Probabilities for Bayesian Inference using MCMC

A Tree with Support Values



Sources of Uncertainty

1) Orthology Assignment
2) Multiple Sequence Alignment
3) Tree Inference
4) BUT

Software Issues

- Bugs & Software Quality
- Numerical Instability
- Reproducibility
- We re-designed & optimized numerous tools – the *Next Generation* (NG) tools series
    - `RAxML-NG`
    - `ModelTest-NG`
    - `EPA-NG`
    - `Lagrange-NG`

Sources of Uncertainty

1) Orthology Assignment
2) Multiple Sequence Alignment
3) Tree Inference
4) Software issues
5) BUT

Propagating Uncertainty

- Assume
  - *10* alternative orthology assignments
  - *10 x 10* alternative MSAs
  - *10 x 10 x 10* alternative trees

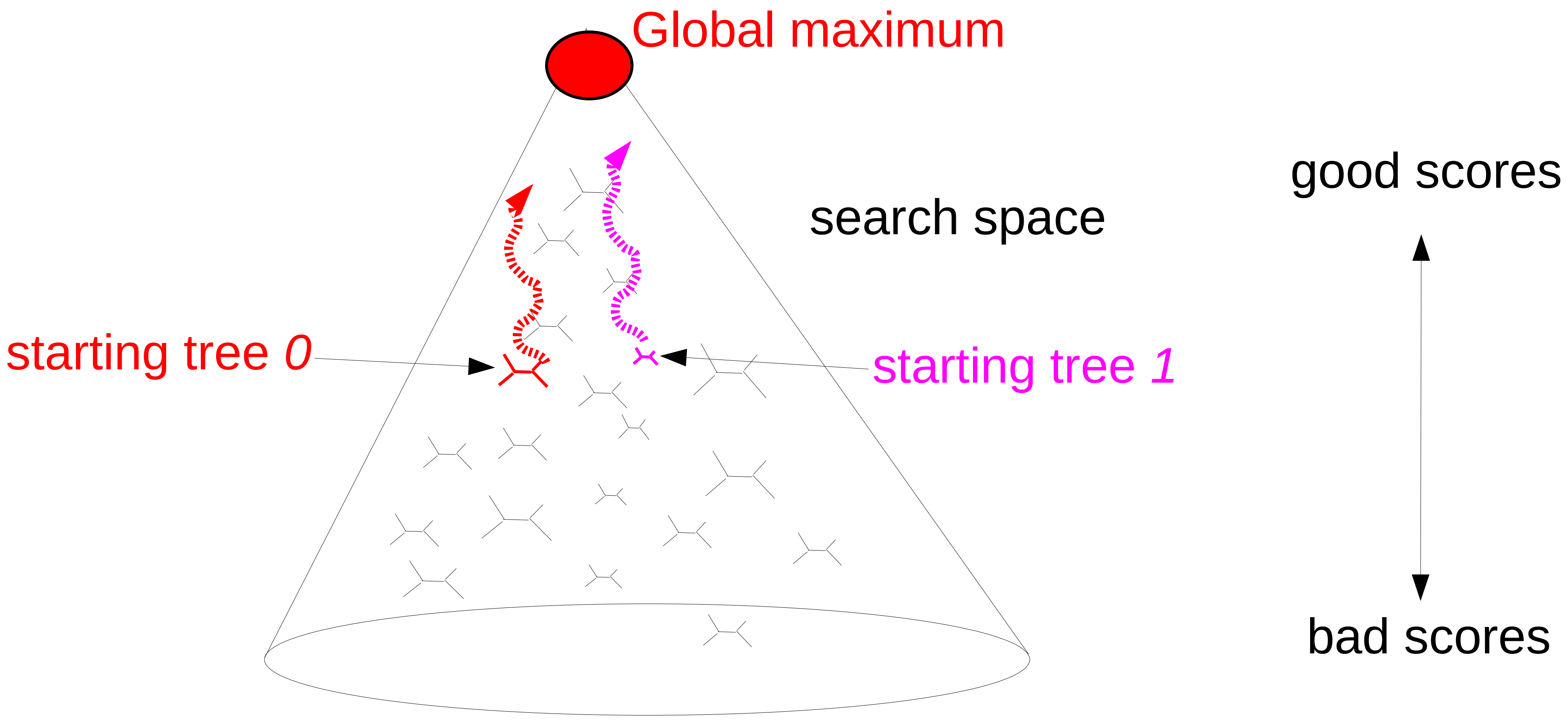
  → exponential explosion with increasing pipeline length

  → intelligent ways to explore parameter space in pipelines needed

Outline

- Our Approach to Bioinformatics
- Introduction to Phylogenetic Inference
- Sources of Uncertainty
- **Phylogenetic Difficulty**
- Other stuff we are working on

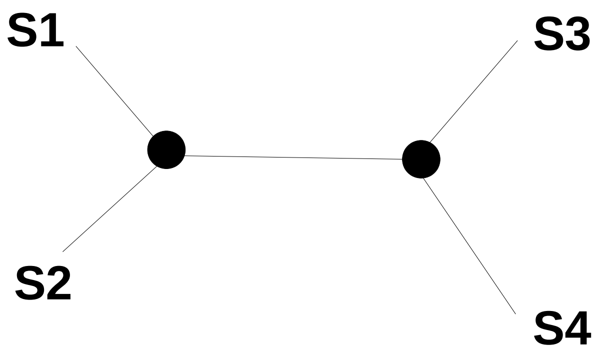
Can we predict how difficult a phylogenetic analysis will be?



Phylogenetic Inference

MSA

S1 ACGTT
S2 ACCGG
S3 TGGAG
S4 GGCTT

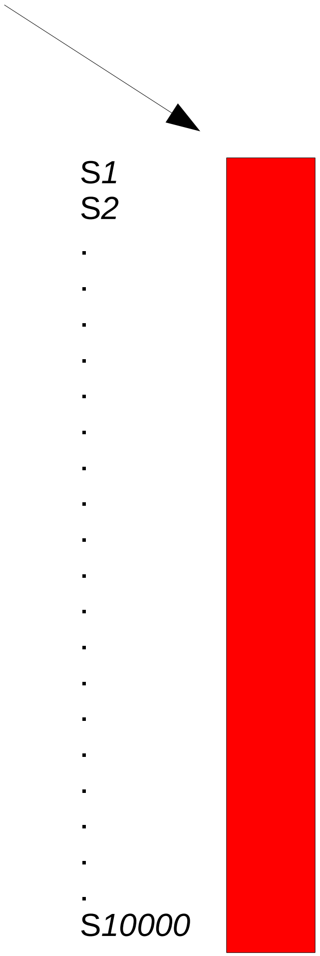


The difficulty of inferring a tree depends on the shape of the multiple sequence alignment

Dataset Shapes

This?

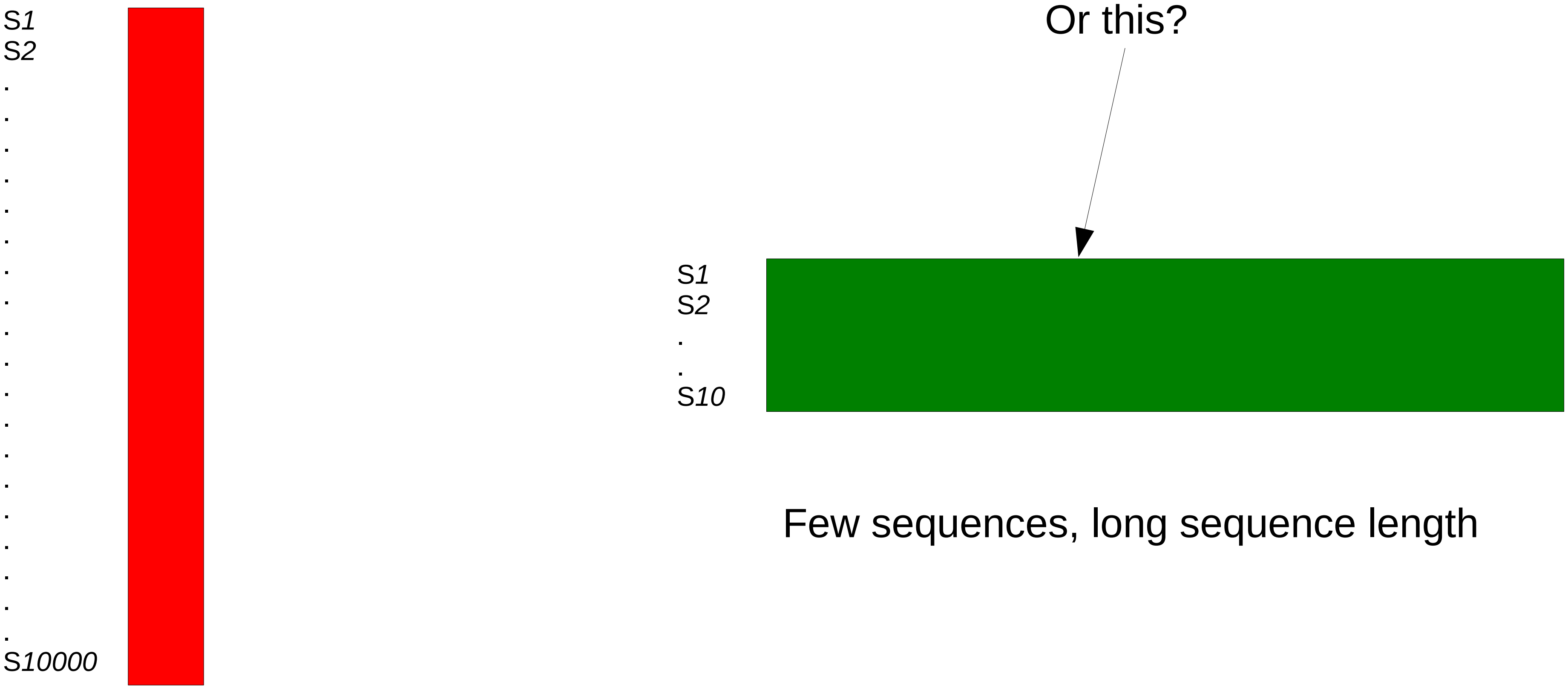
Which data is more difficult to analyze?



Thousands of sequences, short sequence length

Dataset Shapes

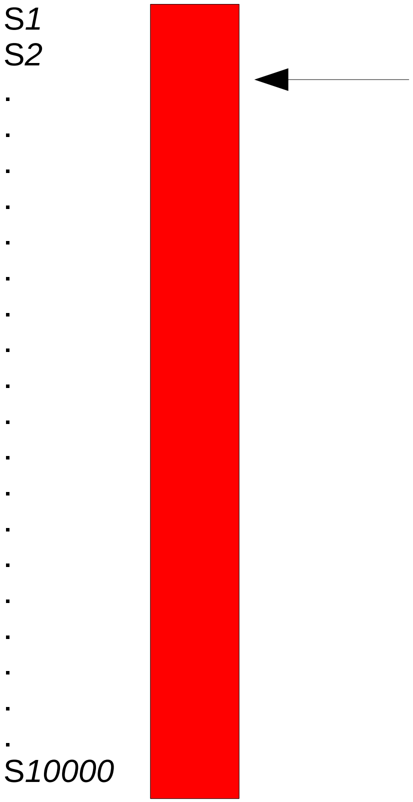
Which data is more difficult to analyze?



Or this?

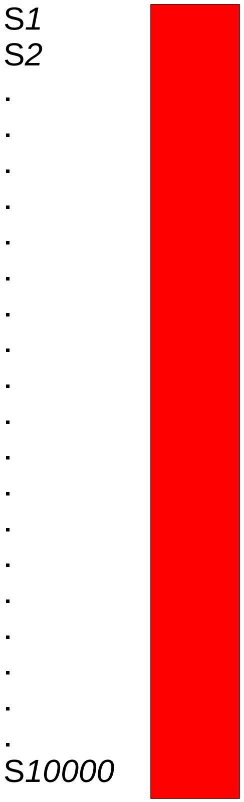
Few sequences, long sequence length

Dataset Shapes



Intuitively it is this dataset here, as it contains much less information for telling apart more sequences

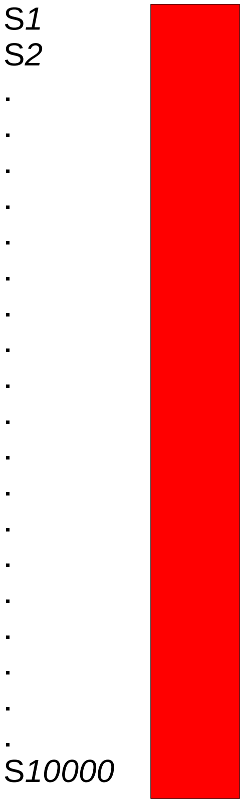
Dataset Shapes



Intuitively it is this dataset here, as it contains much less information for telling apart more sequences

SARS-CoV-2 is such a difficult dataset; it even exhibits some additional difficulties:

Dataset Shapes

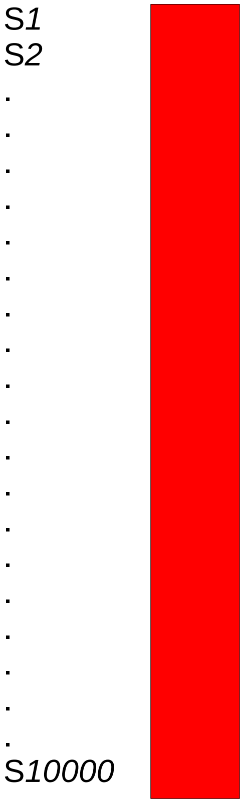


Intuitively it is this dataset here, as it contains much less information for telling apart more sequences

SARS-CoV-2 is such a difficult dataset; it even exhibits some additional difficulties:

1. Due to the low mutation rate (rate at which nucleotides change) sequences are very similar to each other

Dataset Shapes

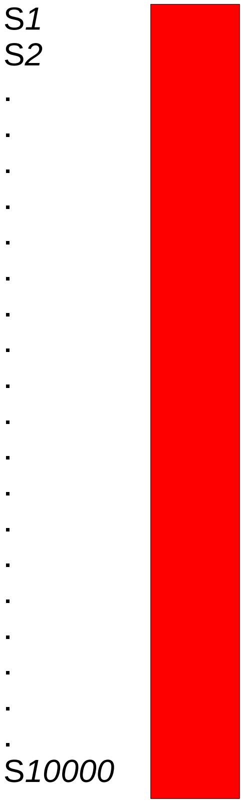


Intuitively it is this dataset here, as it contains much less information for telling apart more sequences

SARS-CoV-2 is such a difficult dataset; it even exhibits some additional difficulties:

1. Due to the low mutation rate (rate at which nucleotides change) sequences are very similar to each other
2. The genome is ≈ *30,000* nucleotides long, but the sequences differ in only *1500-2000* positions → highly similar

Dataset Shapes



Intuitively it is this dataset here, as it contains much less information for telling apart more sequences

SARS-CoV-2 is such a difficult dataset; it even exhibits some additional difficulties:

1. Due to the low mutation rate (rate at which nucleotides change) sequences are very similar to each other
2. The genome is ≈ *30,000* nucleotides long, but the sequences differ in only *1500-2000* positions → highly similar
3. The input sequences are **not from distinct species**!

Consequences

- SARS-CoV-2 data
  - Extremely hard to infer a reliable tree
  - Numerical issues with tree inference tools because the sequences are so closely related
  - Post-analyzing the tree (e.g., determining the root, identifying virus sub-classes) appears to not be feasible using computational tools

Consequences

- SARS-CoV-2 data
  - **Extremely hard to infer a reliable tree**
  - Numerical issues with tree inference tools because the sequences are so closely related
  - Post-analyzing the tree (e.g., determining the root, identifying virus sub-classes) appears to not be feasible using computational tools

For details, see: Benoit Morel, Pierre Barbera, Lucas Czech, Ben Bettisworth, Lukas Hübner, Sarah Lutteropp, Dora Serdari, Evangelia-Georgia Kostaki, Ioannis Mamais, Alexey Kozlov, Pavlos Pavlidis, Dimitrios Paraskevis, Alexandros Stamatakis. **"Phylogenetic analysis of SARS-CoV-2 data is difficult"**, *Molecular Biology and Evolution* 2021

Phylogenetic Inference

- Assembled *4* distinct input datasets
- Per input dataset → executed *100* tree searches
- As we use likelihood models, we determined the trees that are **not statistically significantly different** from each other per set of *100* trees

Results

- For all input datasets about *70* out of *100* trees are not significantly different from each other with respect to their likelihood scores

Results

- For all input datasets about *70* out of *100* trees are not significantly different from each other with respect to their likelihood scores
- But, their pair-wise topological differences (difference in tree shapes) amount on average to **70%** !

Results

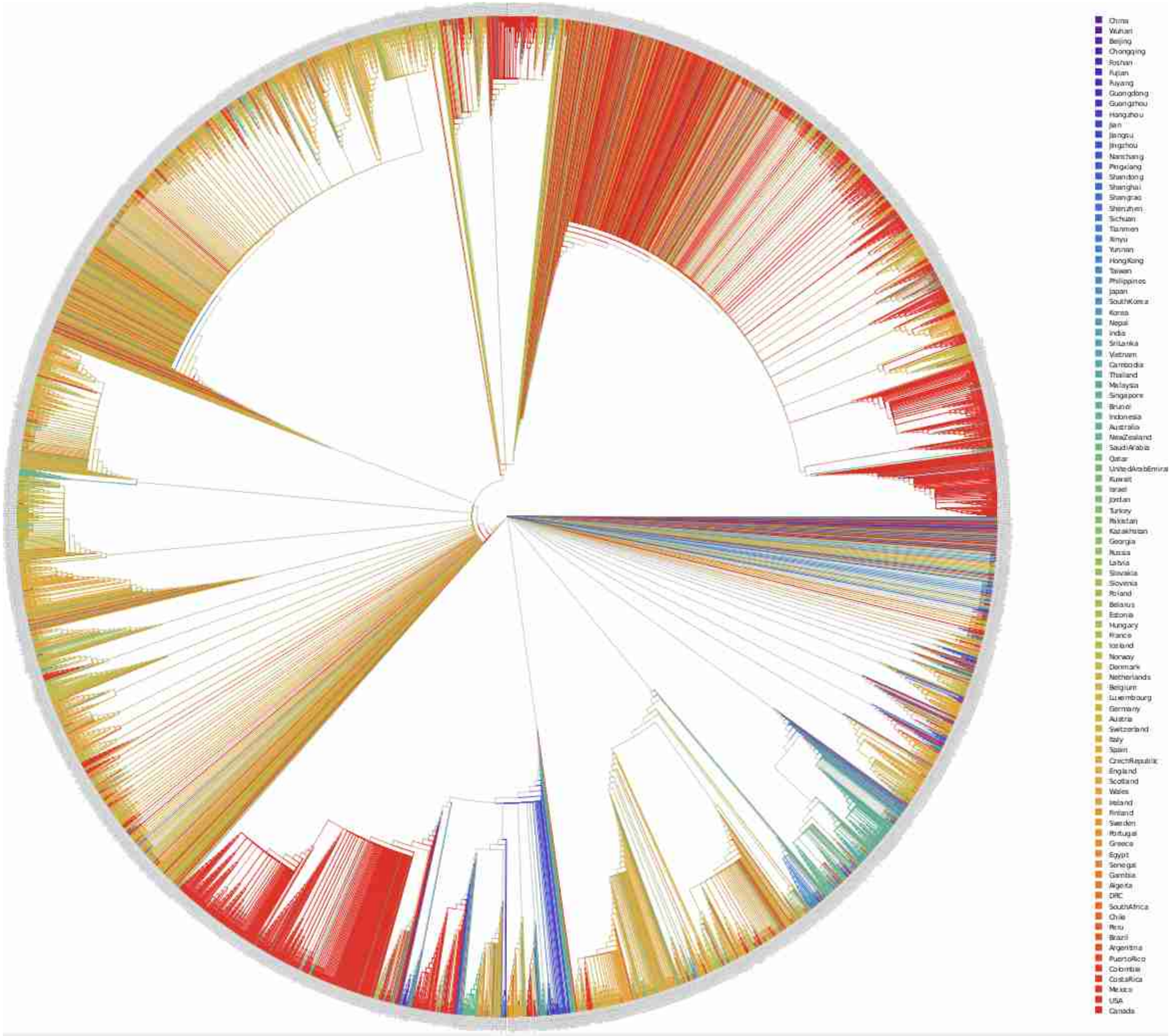
- For all input datasets about *70* out of *100* trees are not significantly different from each other with respect to their likelihood scores
- But, their pair-wise topological differences (difference in tree shapes) amount on average to **70%** !

→ extremely weak signal

→ don't draw conclusions from a single tree!

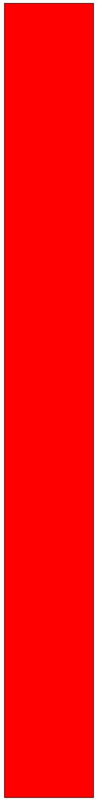
→ try to summarize the trees via summary statistics!

Summarized Trees



SARS-CoV-2 consensus tree colored by country

Difficulty of an MSA



S1
S2
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
S10000

difficult

S1
S2
.
.
S10

easy

Difficulty of an MSA

This is very hand-wavy → can we quantify & predict this

Difficulty Prediction

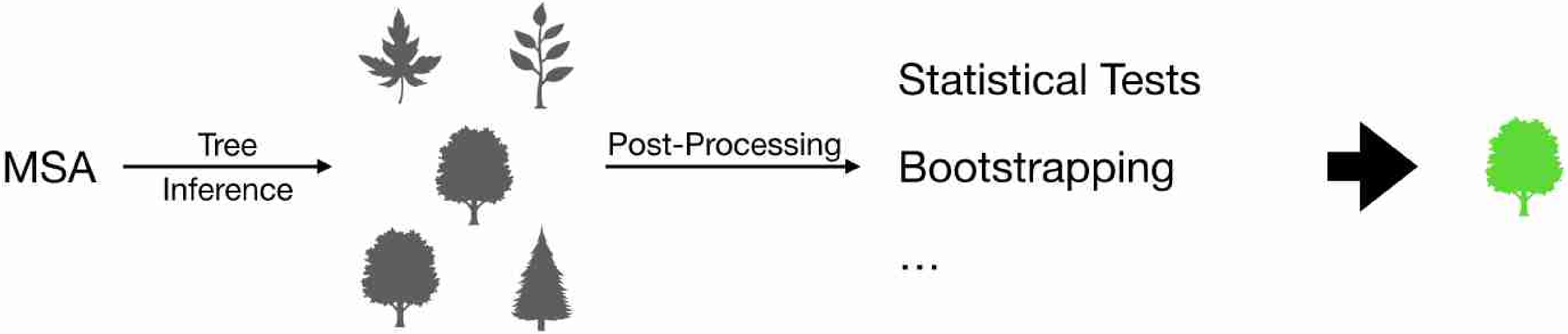
JOURNAL ARTICLE

From Easy to Hopeless—Predicting the Difficulty of Phylogenetic Analyses

Julia Haag ✉, Dimitri Höhler, Ben Bettisworth, Alexandros Stamatakis

Molecular Biology and Evolution, Volume 39, Issue 12, December 2022, msac254,
https://doi.org/10.1093/molbev/msac254
Published: 17 November 2022

Easy



MSA

Tree Inference

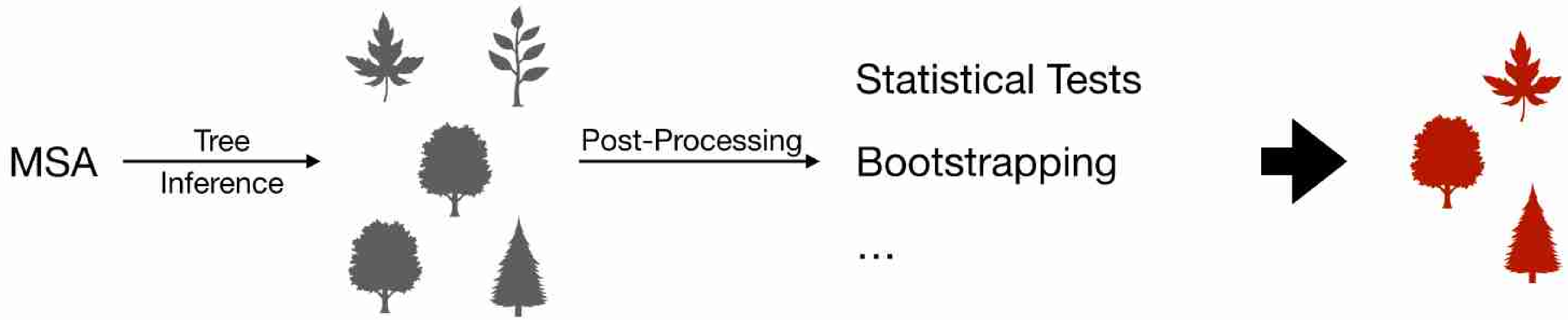
Post-Processing

Statistical Tests

Bootstrapping

…

Difficult



MSA

Tree

Inference

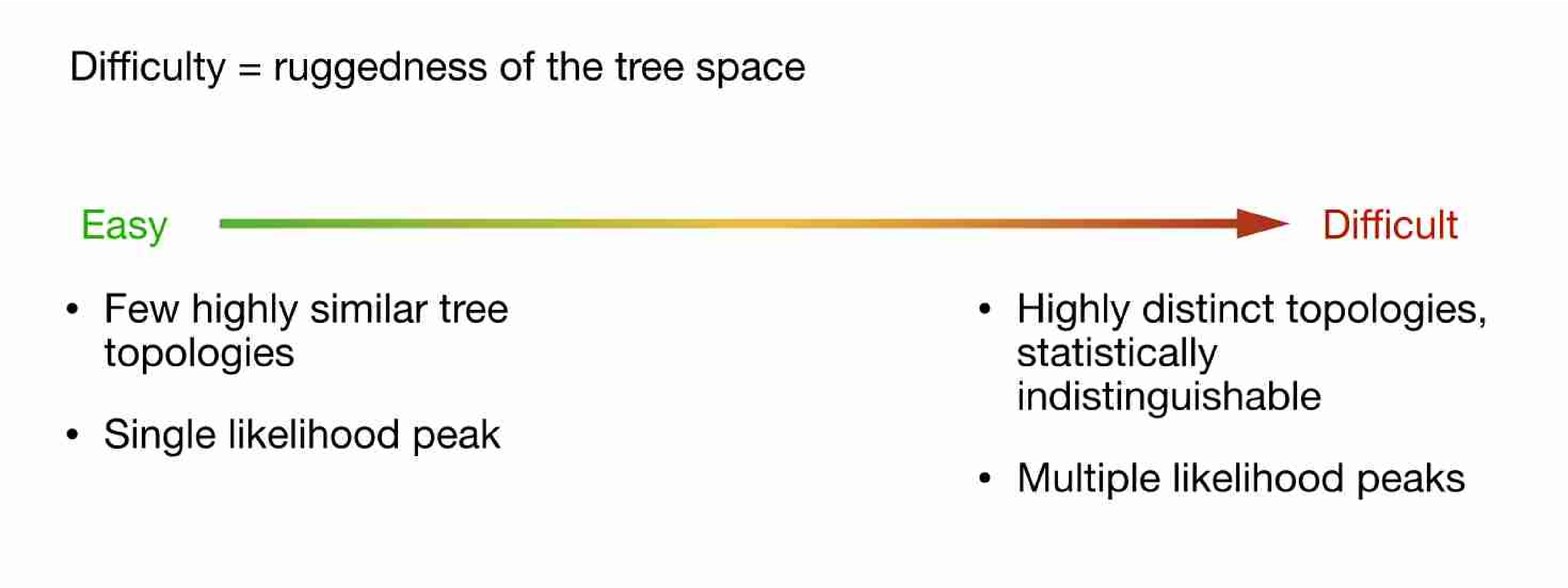
Post-Processing

Statistical Tests

Bootstrapping

...

What does Difficulty mean?



Difficulty = ruggedness of the tree space

Easy → Difficult

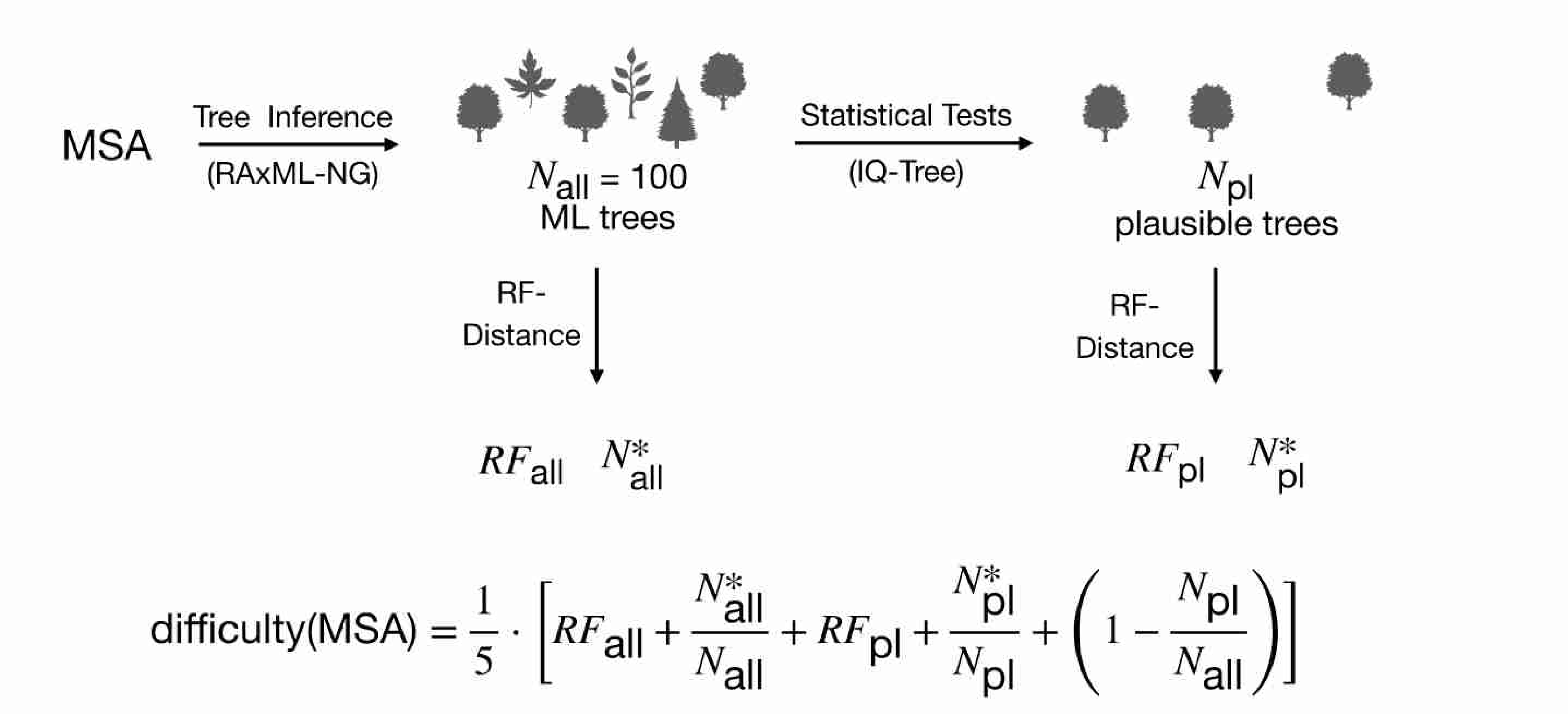
- Few highly similar tree topologies
- Single likelihood peak

- Highly distinct topologies, statistically indistinguishable
- Multiple likelihood peaks

Predicting Difficulty with `Pythia`

- `Pythia` = Boosted Tree Regressor
- Supervised Regression Task
    - Predict difficulty between 0.0 (easy) and 1 (difficult)
    - Ground truth difficulty as training target based on 100 distinct Maximum Likelihood tree inferences
- Trained on 4K empirical MSAs
    - Mean absolute error: 2.5%

Definition of Difficulty

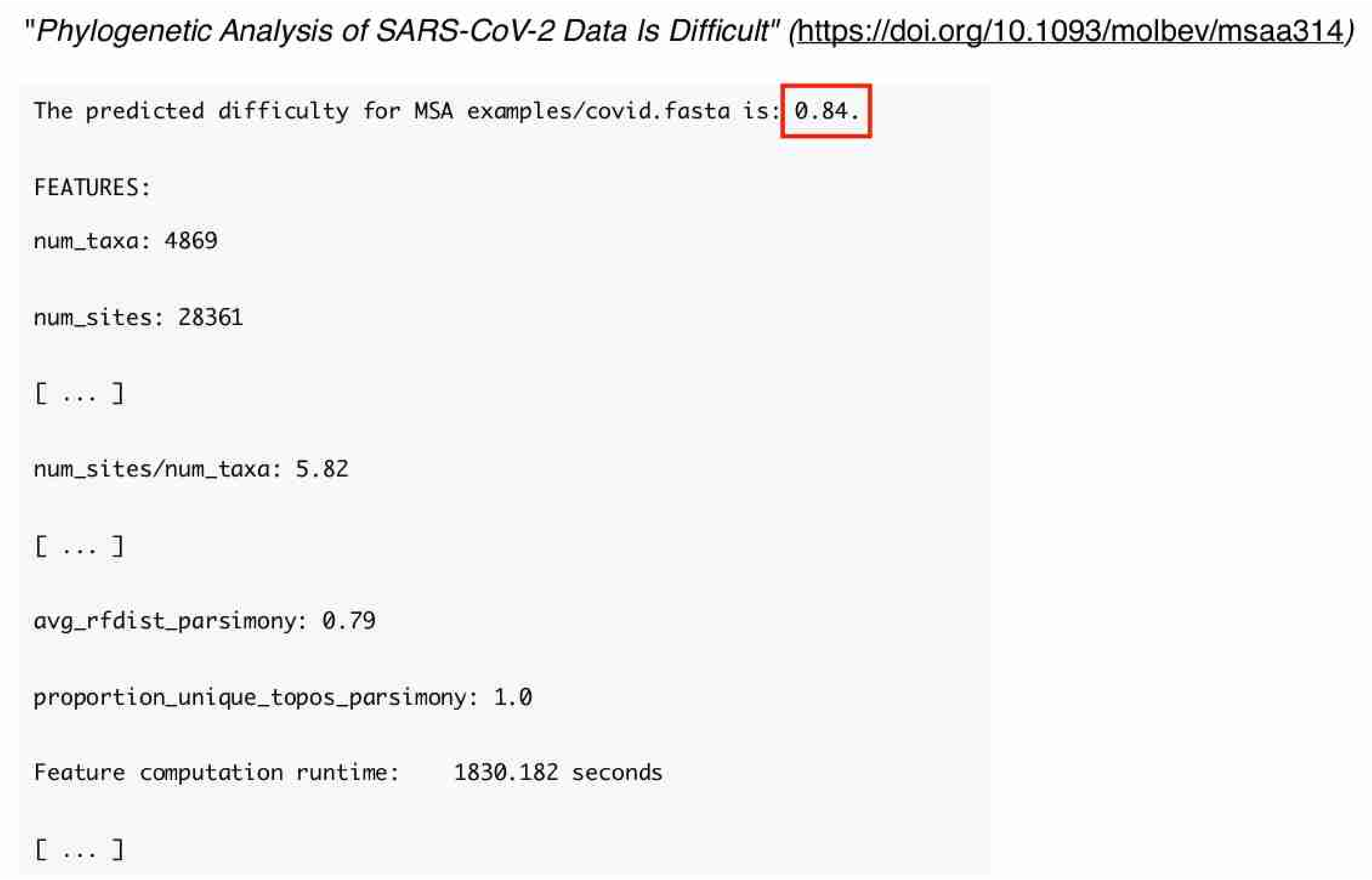


$$\text{difficulty(MSA)} = \frac{1}{5} \cdot \left[RF_{\text{all}} + \frac{N^*_{\text{all}}}{N_{\text{all}}} + RF_{\text{pl}} + \frac{N^*_{\text{pl}}}{N_{\text{pl}}} + \left(1 - \frac{N_{\text{pl}}}{N_{\text{all}}}\right) \right]$$

Prediction Features

- Eight Features
  - 4 MSA attributes
    - sites-over-taxa, patterns-over-taxa, **%** gaps, **%** invariant sites
  - 2 MSA information metrics
    - Shannon entropy, Bollback multinomial test statistic
  - 2 Parsimony-tree-based features
    - Infer 100 parsimony trees
      - → average RF-Distance
      - → **%** unique topologies

SARS-CoV-2 Example



"*Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult*" (https://doi.org/10.1093/molbev/msaa314)

```
The predicted difficulty for MSA examples/covid.fasta is: 0.84.

FEATURES:
num_taxa: 4869

num_sites: 28361

[ ... ]

num_sites/num_taxa: 5.82

[ ... ]

avg_rfdist_parsimony: 0.79

proportion_unique_topos_parsimony: 1.0

Feature computation runtime:    1830.182 seconds

[ ... ]
```

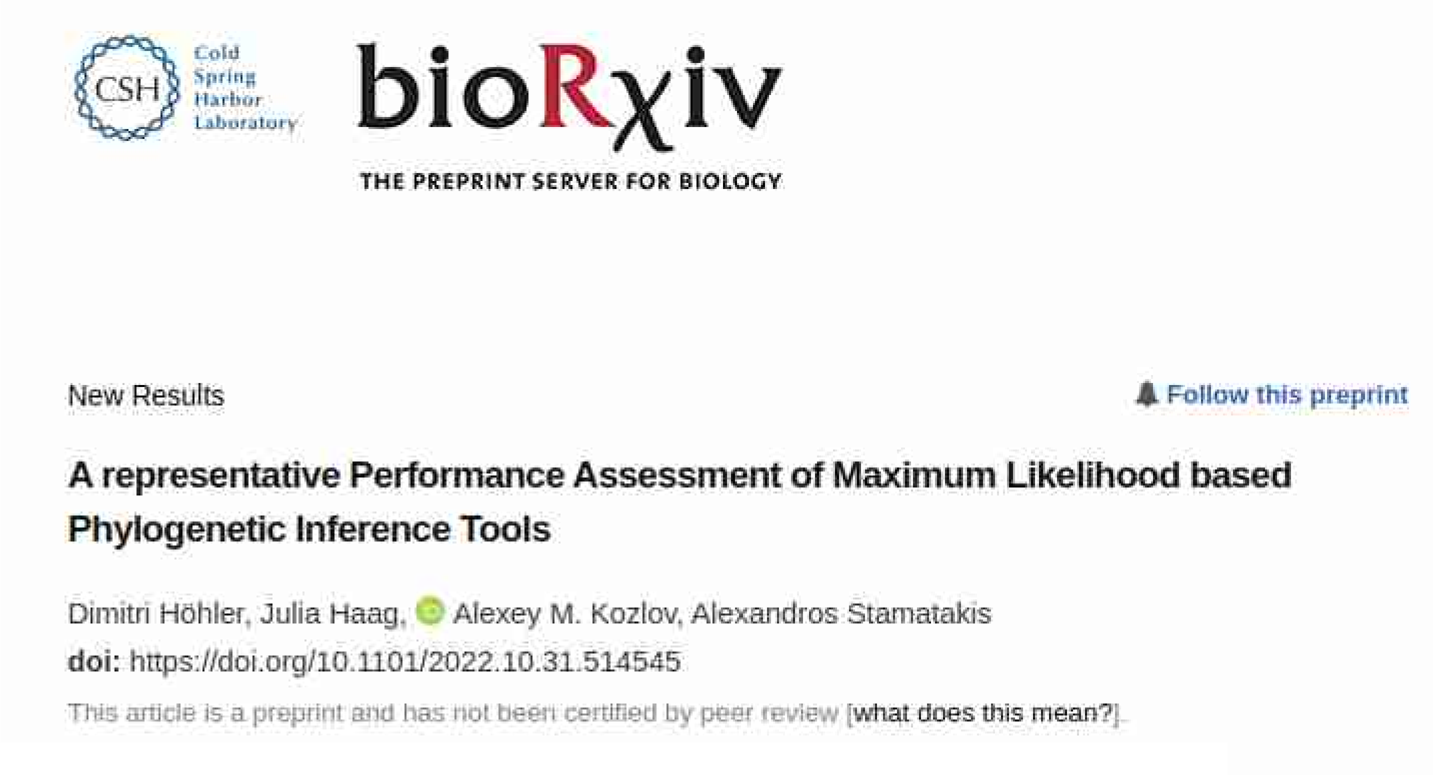
Using `Pythia`

- **Prior** to tree inference
  - → determine analysis & post-analysis setup
  - → adjust/modify MSA
  - → explore data filtering & assembly strategies
  - → adjust user expectations about data

`Pythia` developments

- New release (yesterday)
  - Trained on 12K datasets
    - 11,108 DNA MSAs
    - 979 Protein MSAs
    - 460 Morphological MSAs
  - Two new features
  - The new predictor shows an improved accuracy:
    - Mean absolute error: 0.07 (previously 0.09)
    - Mean absolute percentage error: 1.7% (previously 2.5%)
- Follow-up projects
  - A Simulation study
  - Deploy to inform tree search heuristics

Simulation Study



New Results

A representative Performance Assessment of Maximum Likelihood based Phylogenetic Inference Tools

Dimitri Höhler, Julia Haag, Alexey M. Kozlov, Alexandros Stamatakis

doi: https://doi.org/10.1101/2022.10.31.514545

This article is a preprint and has not been certified by peer review [what does this mean?].

Accuracy as Function of Difficulty

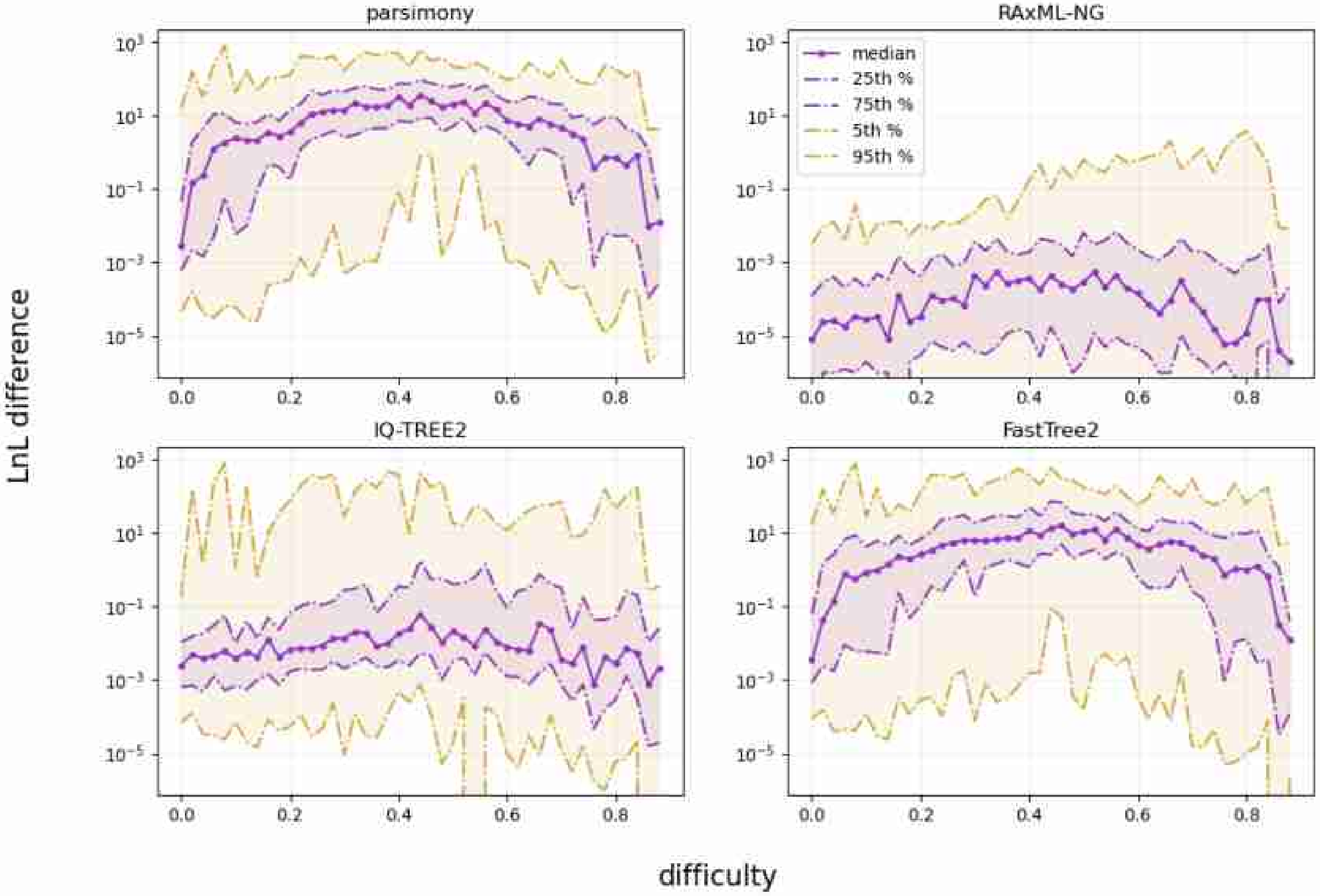


Fig. 3. Absolute log-likelihood (LnL) score differences (log scale) from the best-known ML tree on TreeBASE data.

Adaptive RAxML-NG

New Results

Adaptive RAxML-NG: Accelerating Phylogenetic inference under Maximum Likelihood using dataset difficulty

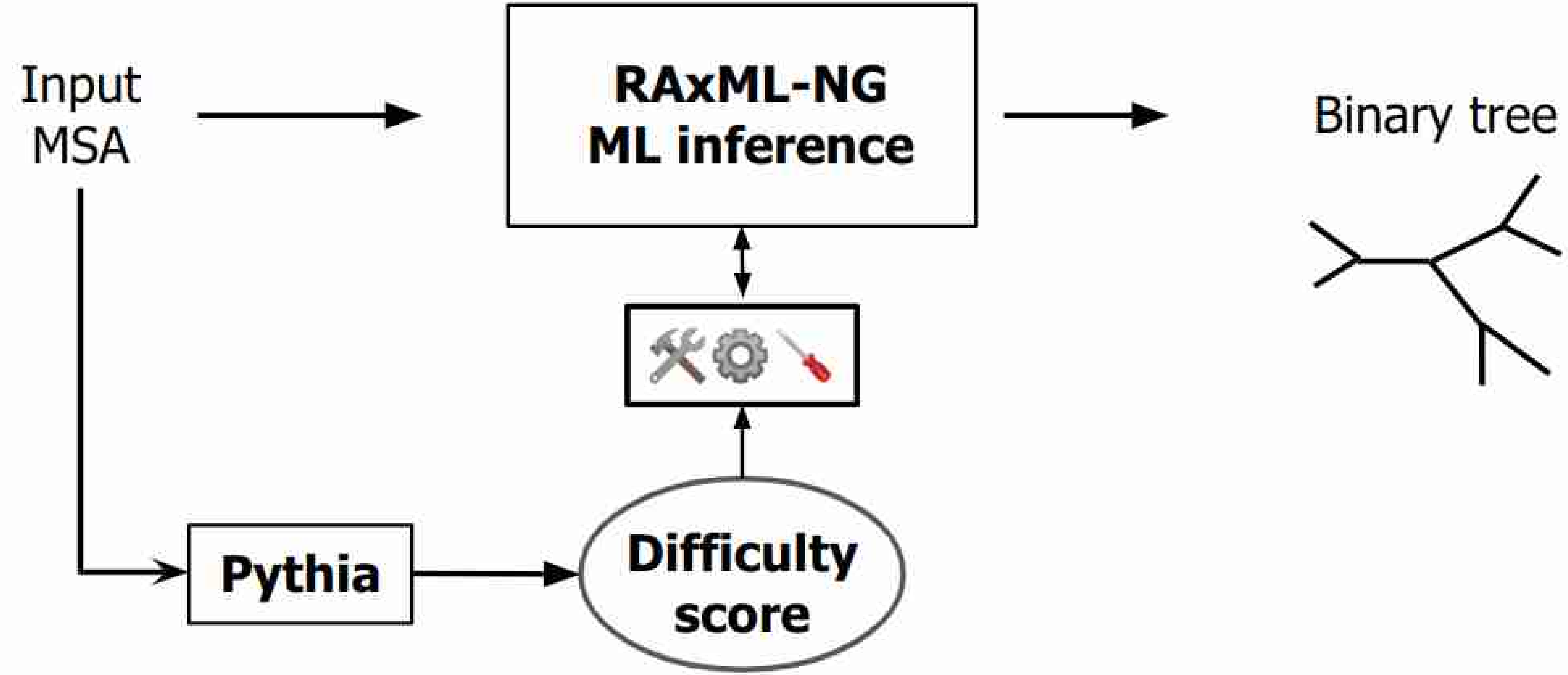
Anastasis Togkousidis, Alexey M Kozlov, Julia Haag, Dimitri Höhler, Alexandros Stamatakis

doi: https://doi.org/10.1101/2023.05.15.540873

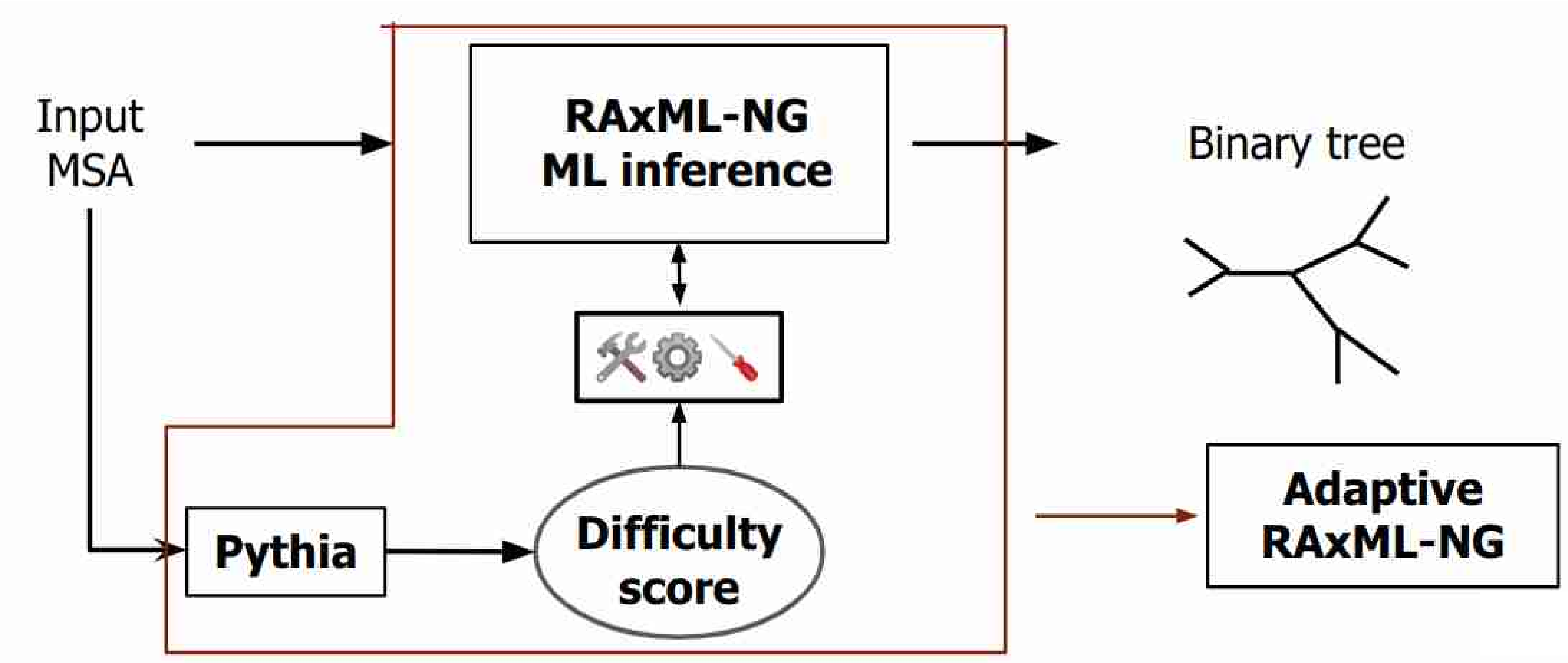
This article is a preprint and has not been certified by peer review [what does this mean?].

Preprint published two days ago - **May 18**

Adaptive `RAxML-NG`



Adaptive `RAxML-NG`



Input MSA

RAxML-NG ML inference

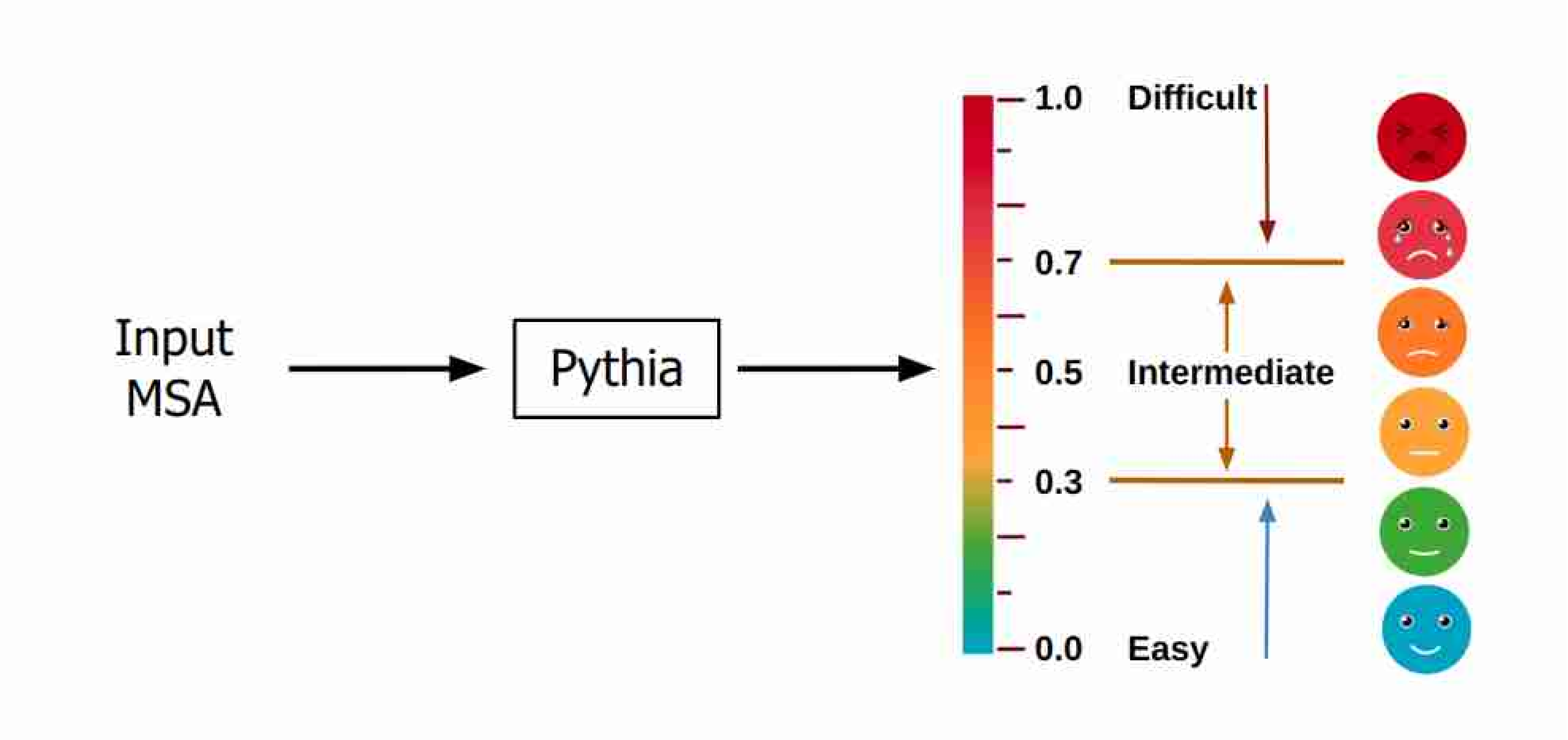
Binary tree

Pythia

Difficulty score

Adaptive RAxML-NG

Pythia



Adaptive `RAxML-NG` Heuristics

- Modify as a function of difficulty
  - the number of ML tree searches
  - the thoroughness of the search

- And introduce an additional tree search mechanism

Test Data & Setup

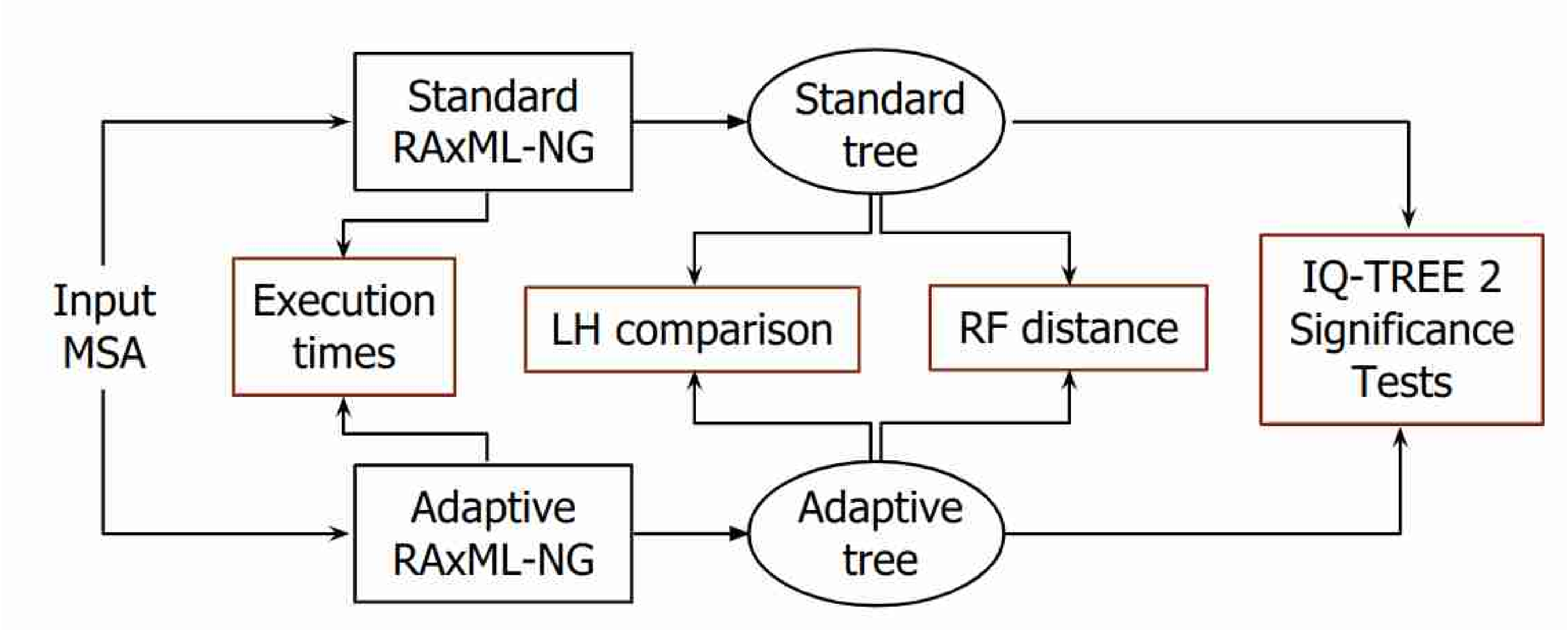
- 10K empirical MSAs from `TreeBase`

  → 9192 MSAs after filtering

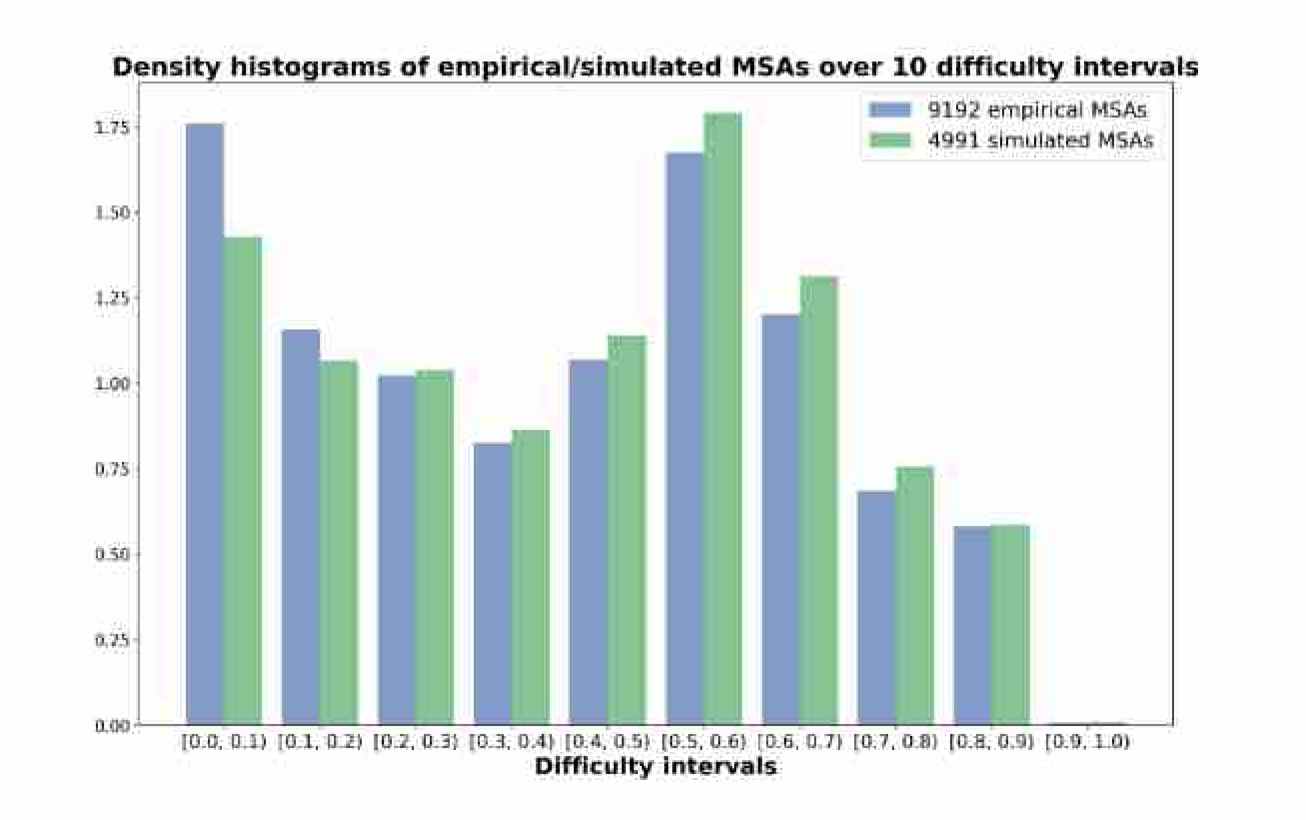
Test Data & Setup

- 10K empirical MSAs from `TreeBase`

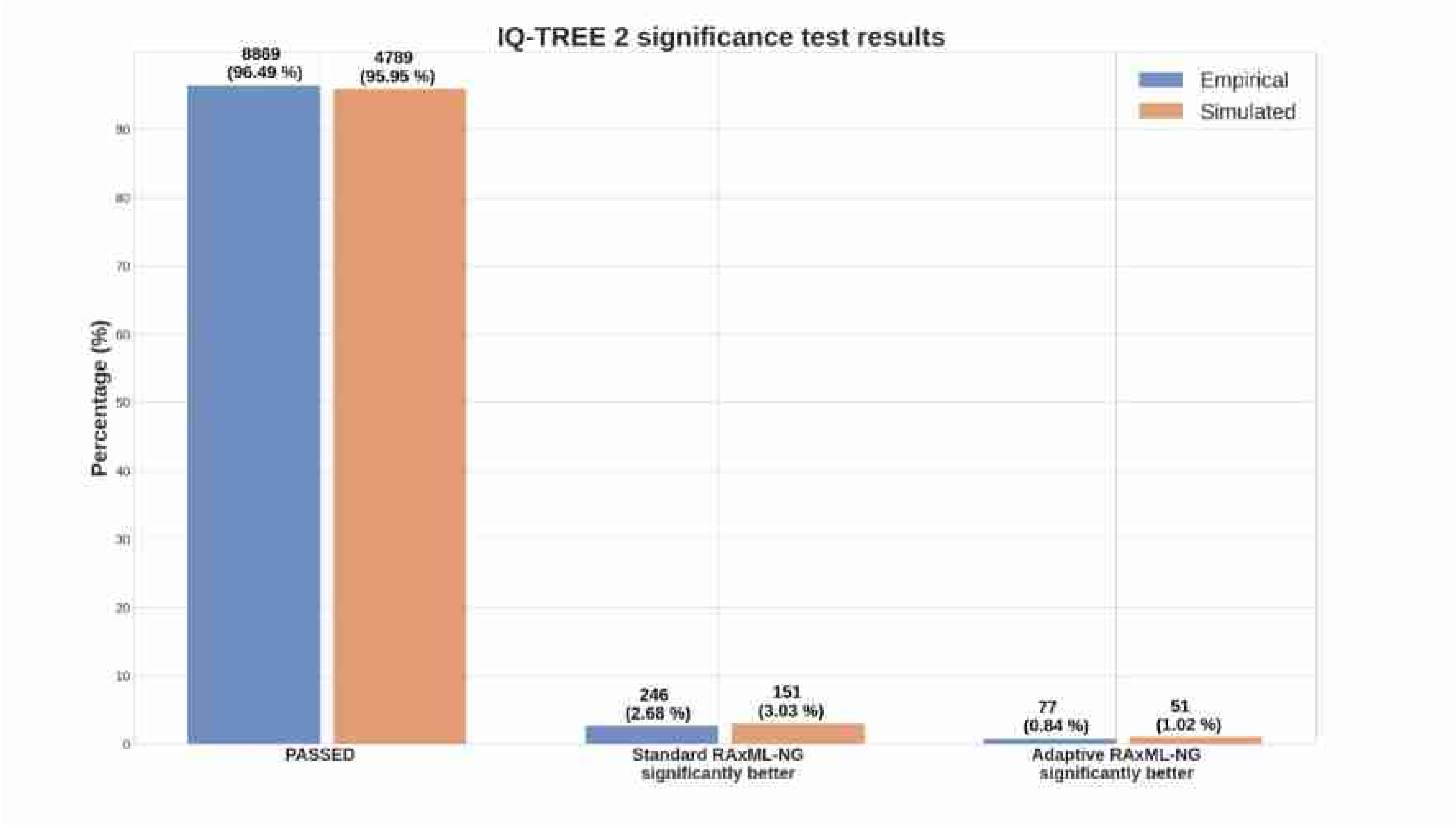
  → 9192 MSAs after filtering



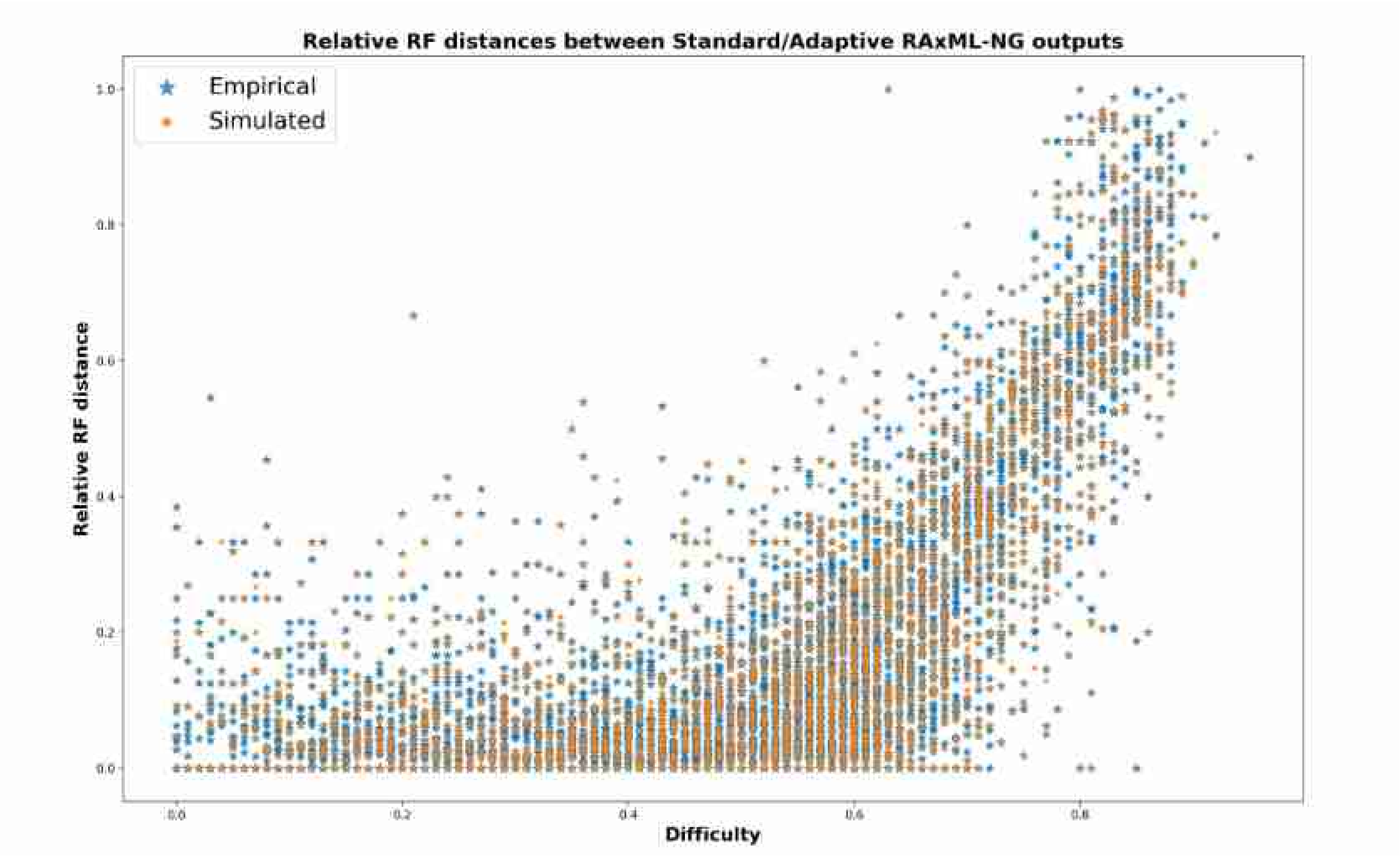
Difficulty Score Distribution



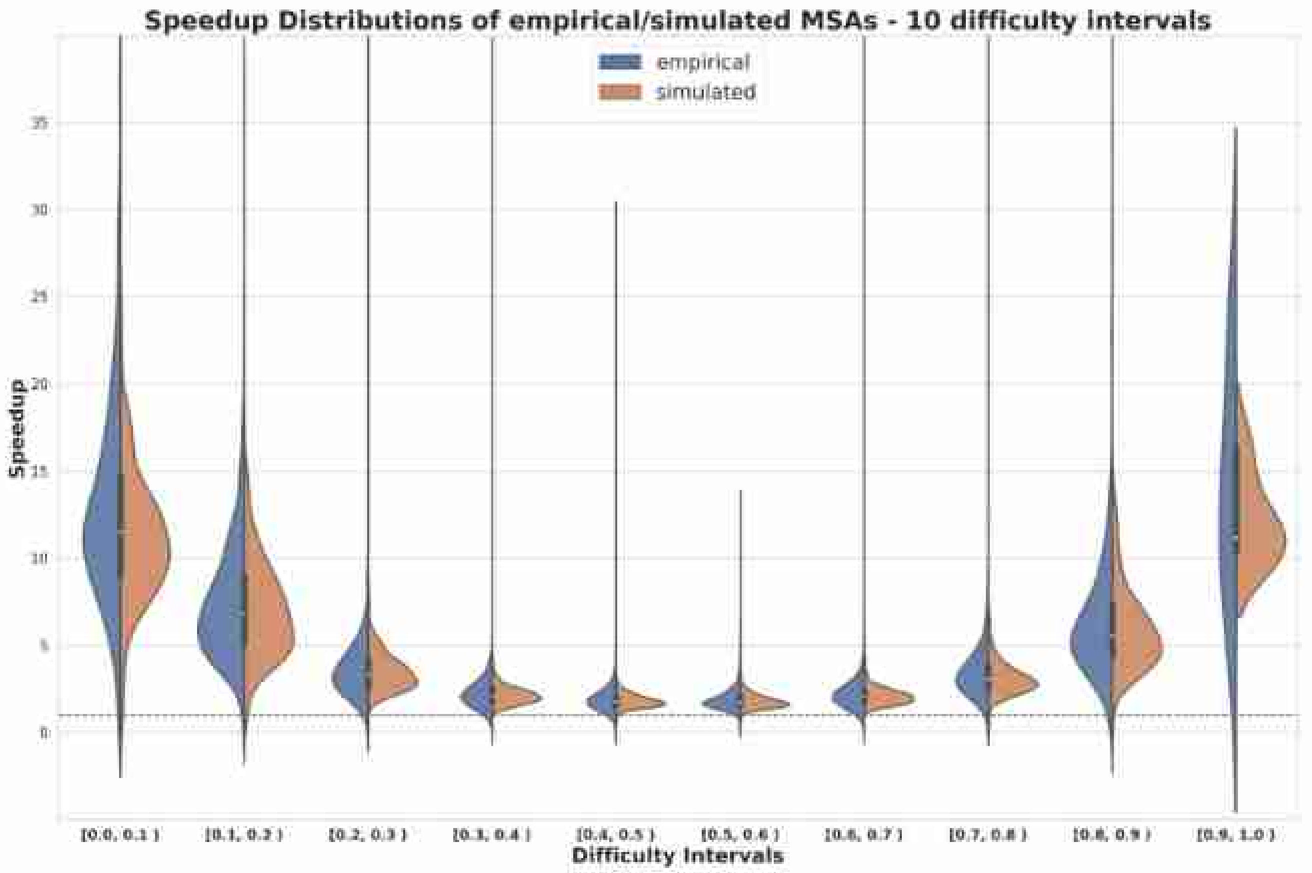
Significance Tests



Distances between trees

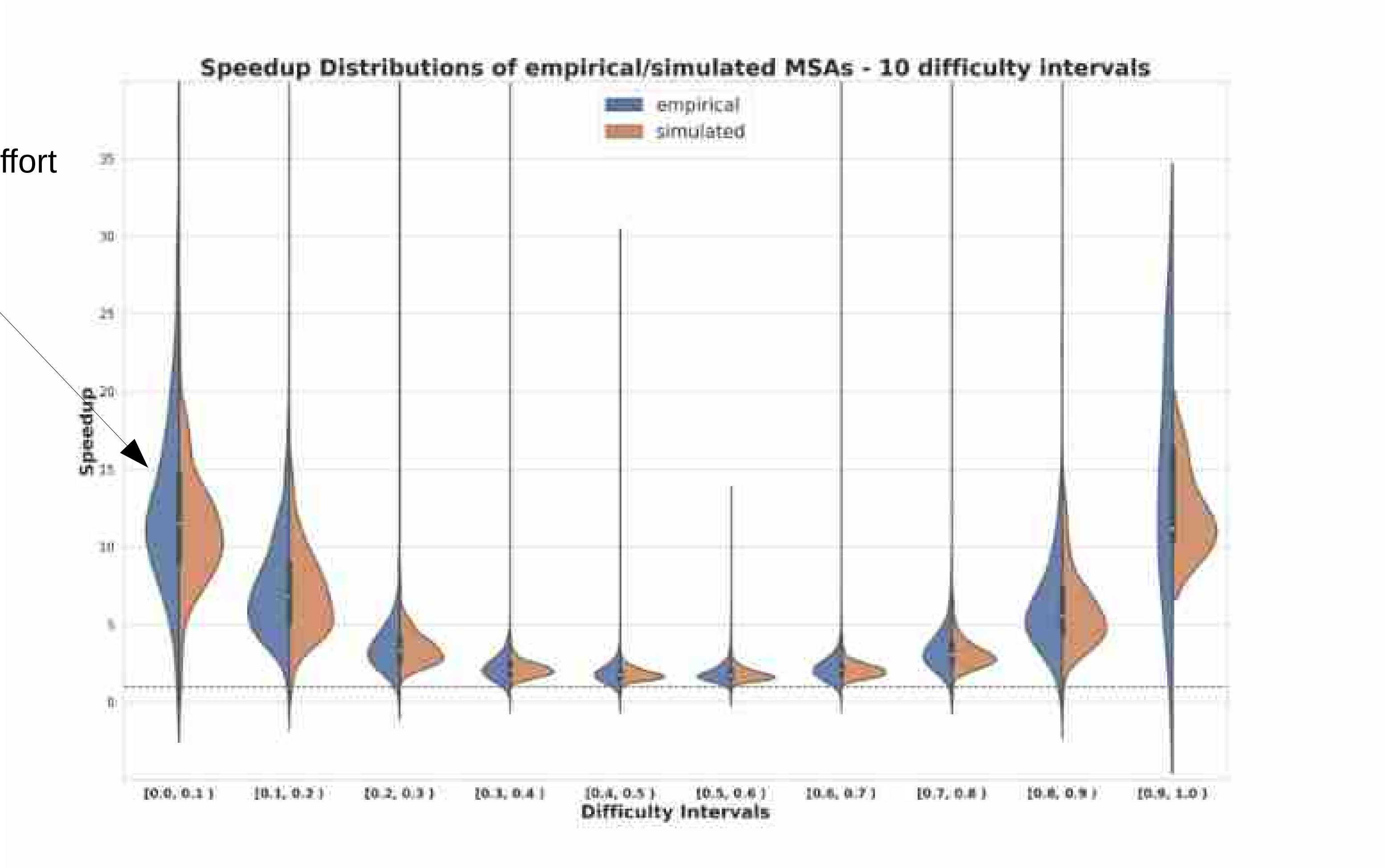


Speedups

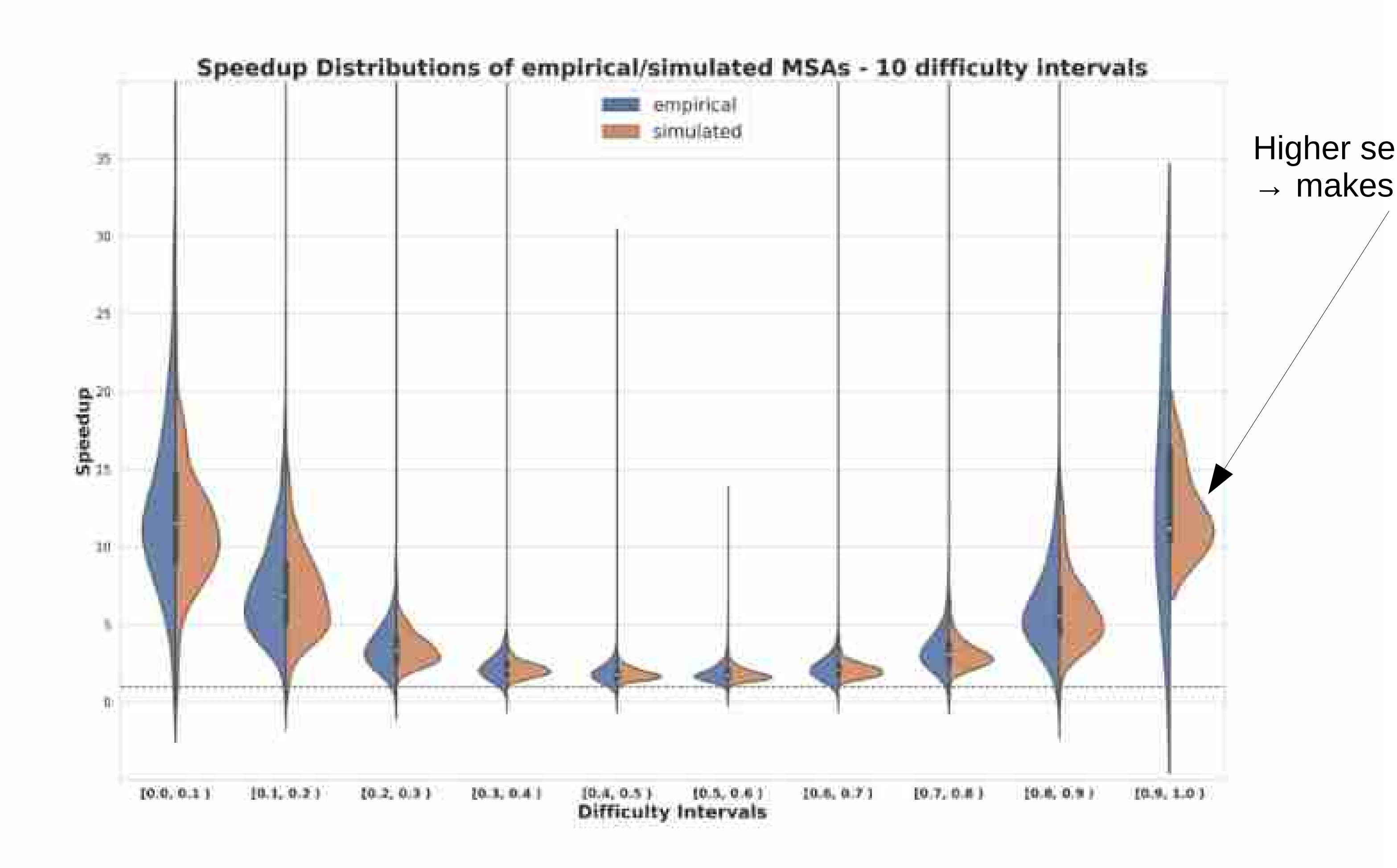

Speedup Distributions of empirical/simulated MSAs - 10 difficulty intervals
empirical
simulated
Speedup
35
30
25
20
15
10
5
0
[0.0, 0.1)
[0.1, 0.2)
[0.2, 0.3)
[0.3, 0.4)
[0.4, 0.5)
[0.5, 0.6)
[0.6, 0.7)
[0.7, 0.8)
[0.8, 0.9)
[0.9, 1.0)
Difficulty Intervals

Speedups

Higher search effort
→ not required

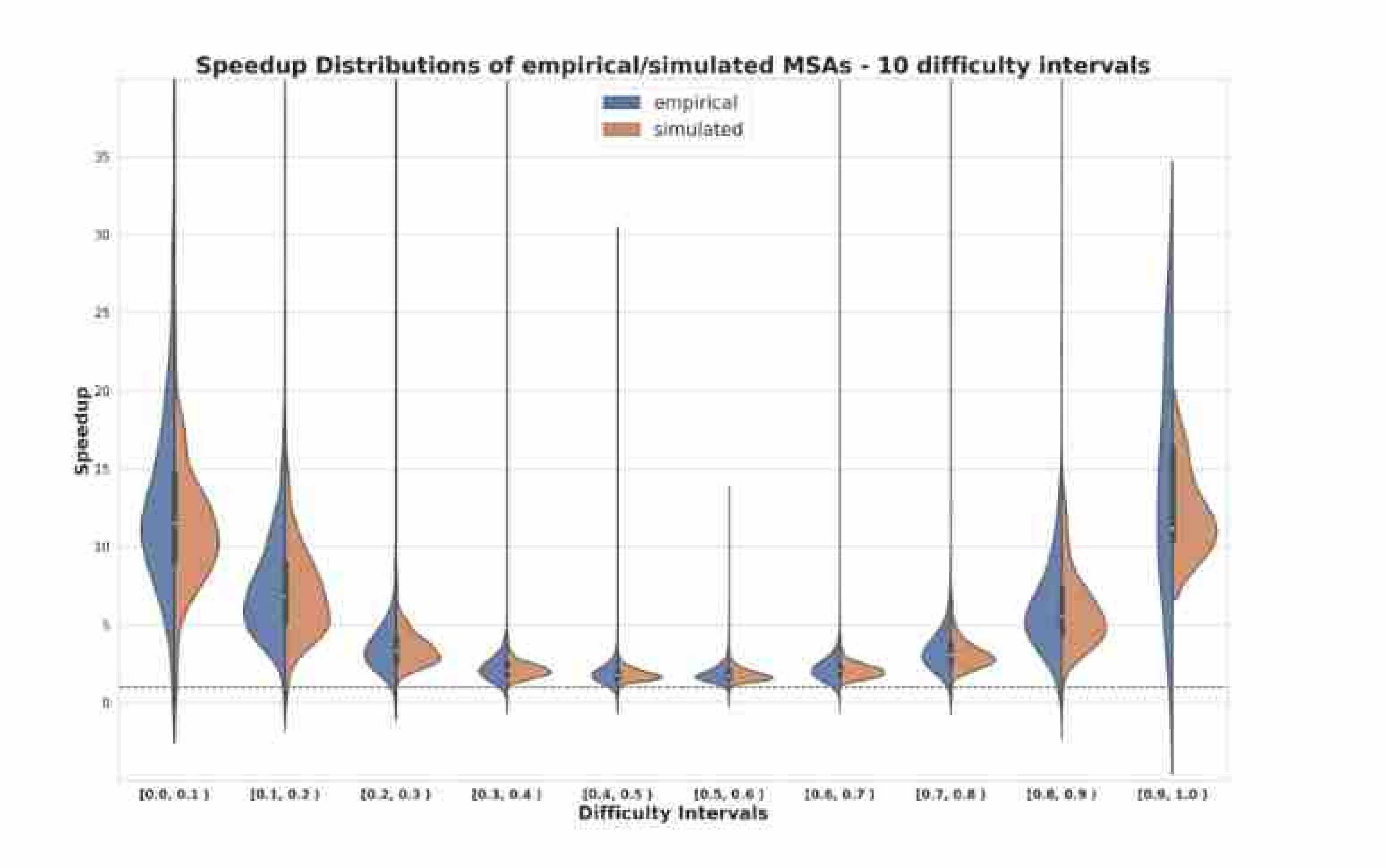


Speedups



Higher search effort
→ makes no sense

Speedups

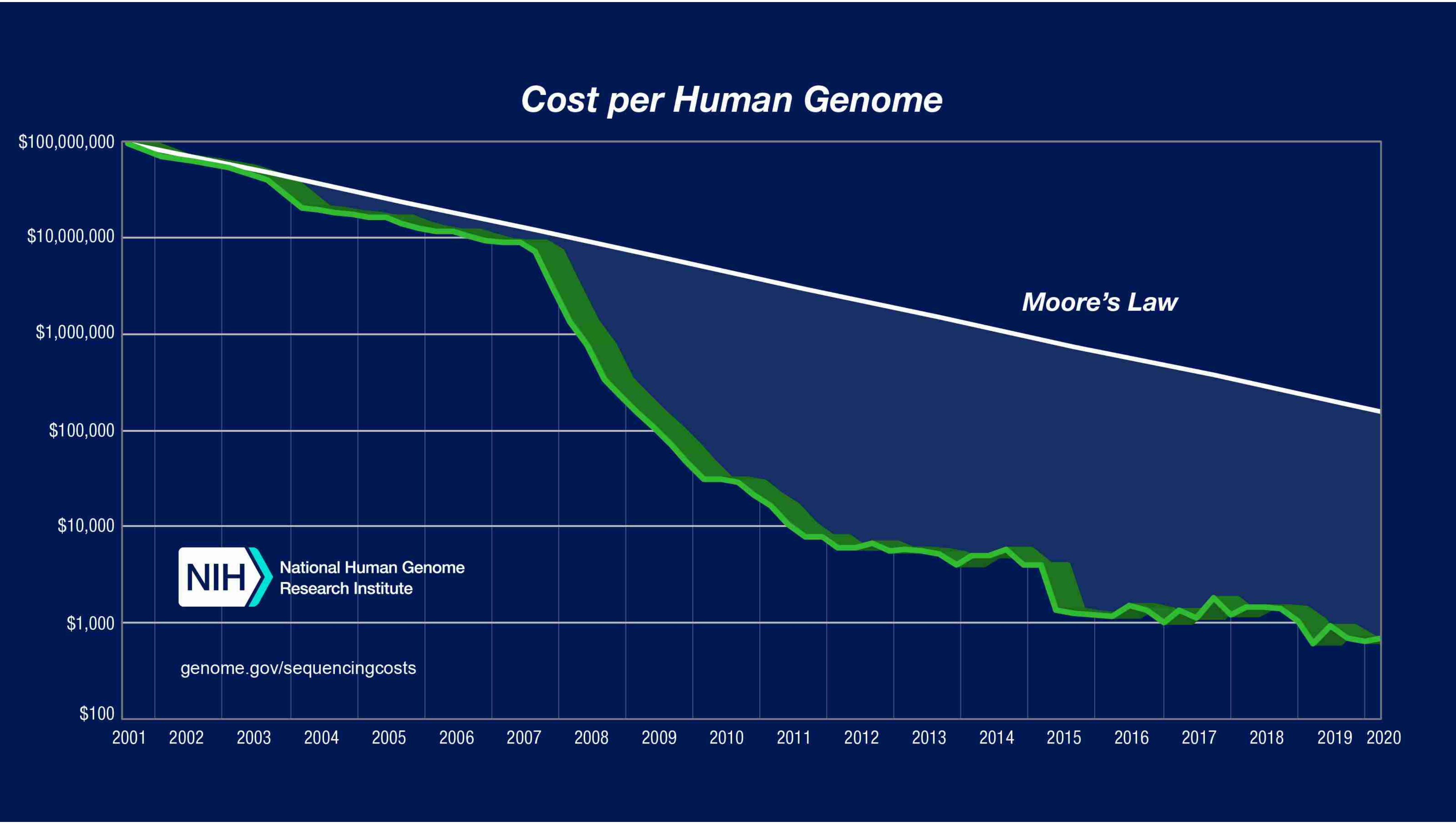


Overall accumulated speedup: approx. 3 on empirical data

Outline

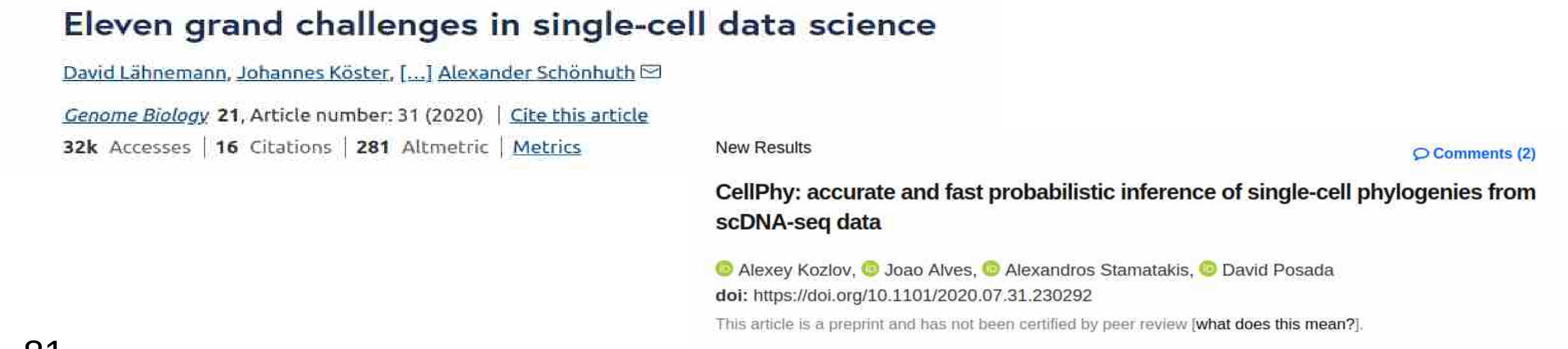
- Our Approach to Bioinformatics
- Introduction to Phylogenetic Inference
- Sources of Uncertainty
- Phylogenetic Difficulty
- **Other stuff we are working on**

Scalability



Single Cell Evolution

- Reconstructing the evolution, e.g., of cancer cells in a single patient is challenging
  - Noisy data
  - Erroneous data
  - Little signal
  - Few & simplistic models



Eleven grand challenges in single-cell data science

David Lähnemann, Johannes Köster, [...] Alexander Schönhuth

Genome Biology **21**, Article number: 31 (2020) | Cite this article

32k Accesses | **16** Citations | **281** Altmetric | Metrics

New Results

Comments (2)

CellPhy: accurate and fast probabilistic inference of single-cell phylogenies from scDNA-seq data

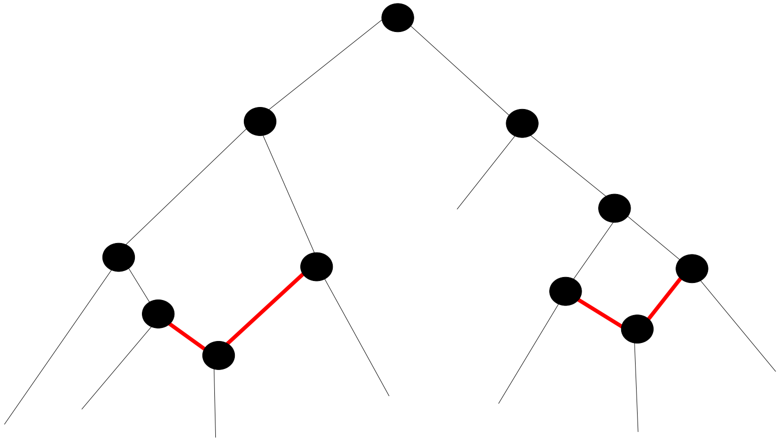
Alexey Kozlov, Joao Alves, Alexandros Stamatakis, David Posada

doi: https://doi.org/10.1101/2020.07.31.230292

This article is a preprint and has not been certified by peer review [what does this mean?].

Phylogenetic Networks

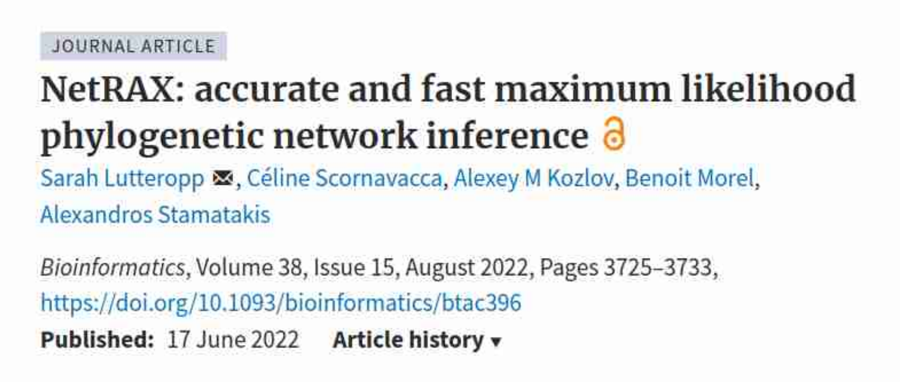
- Evolution does not need to occur in a tree-like manner due to recombination events
- We can model this via so-called phylogenetic networks



Phylogenetic Networks

- Evolution does not need to occur in a tree-like manner due to recombination events
- We can model this via so-called phylogenetic networks
- The likelihood of such a network is substantially more difficult to compute than on a tree

→ computational challenges



JOURNAL ARTICLE

NetRAX: accurate and fast maximum likelihood phylogenetic network inference

Sarah Lutteropp ✉, Céline Scornavacca, Alexey M Kozlov, Benoit Morel, Alexandros Stamatakis

Bioinformatics, Volume 38, Issue 15, August 2022, Pages 3725–3733, https://doi.org/10.1093/bioinformatics/btac396

Published: 17 June 2022 **Article history** ▾

Gene Tree Species Tree Reconciliation

- There are other phenomena that complicate evolution
    - Gene loss
    - Gene transfer
    - Gene duplication

      → gene tree ≠ species tree
- Infer & correct trees under a joint likelihood model comprising the phylogenetic likelihood and a reconciliation likelihood model

GeneRax

- First full and efficient Maximum Likelihood implementation to infer gene family trees using a given rooted species tree under a joint phylogenetic & reconciliation likelihood model

GeneRax: A Tool for Species-Tree-Aware Maximum Likelihood-Based Gene Family Tree Inference under Gene Duplication, Transfer, and Loss

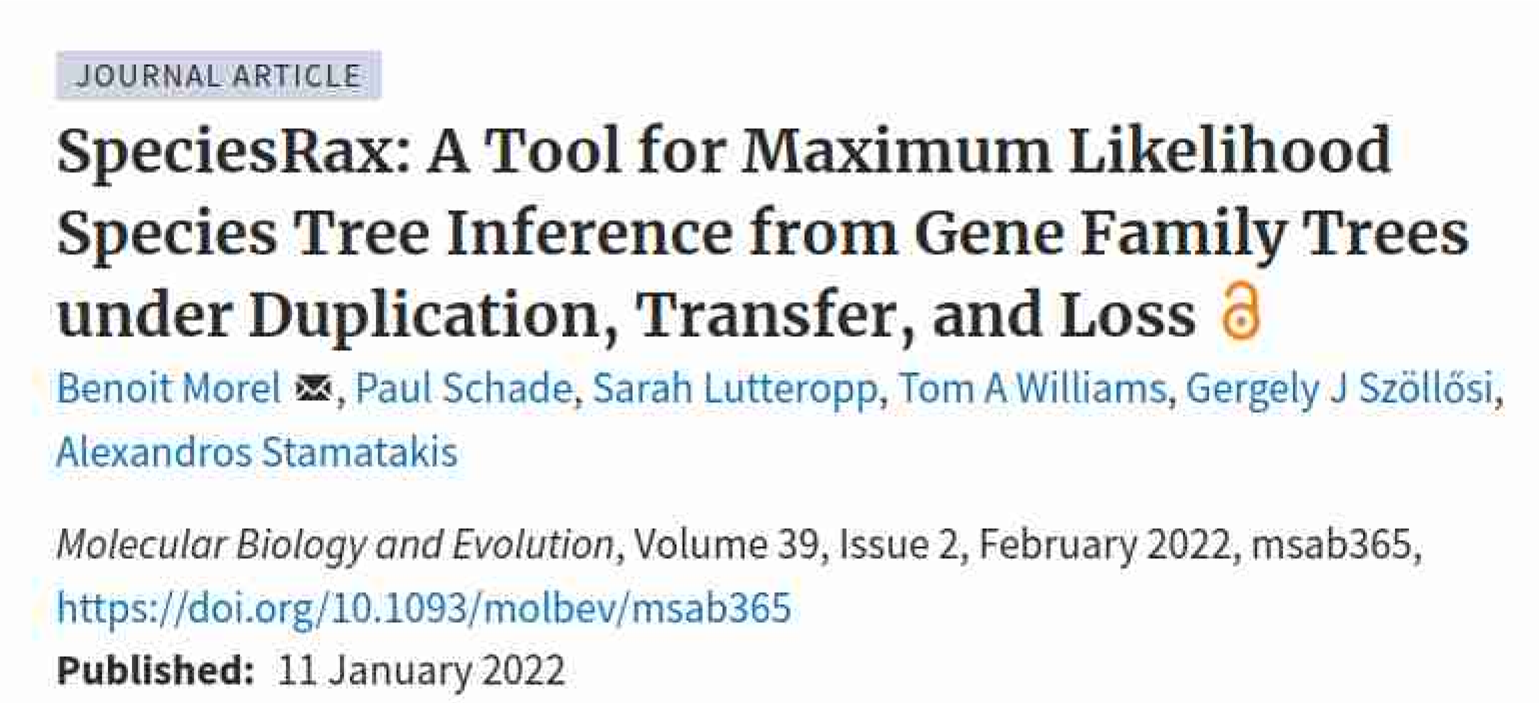
Benoit Morel, Alexey M Kozlov, Alexandros Stamatakis, Gergely J Szöllősi

Molecular Biology and Evolution, Volume 37, Issue 9, September 2020, Pages 2763–2774, https://doi.org/10.1093/molbev/msaa141

Published: 05 June 2020

SpeciesRax

- **Goal:** Simultaneously infer the gene family trees **and** the species tree under a joint phylogenetic/reconciliation likelihood model



JOURNAL ARTICLE

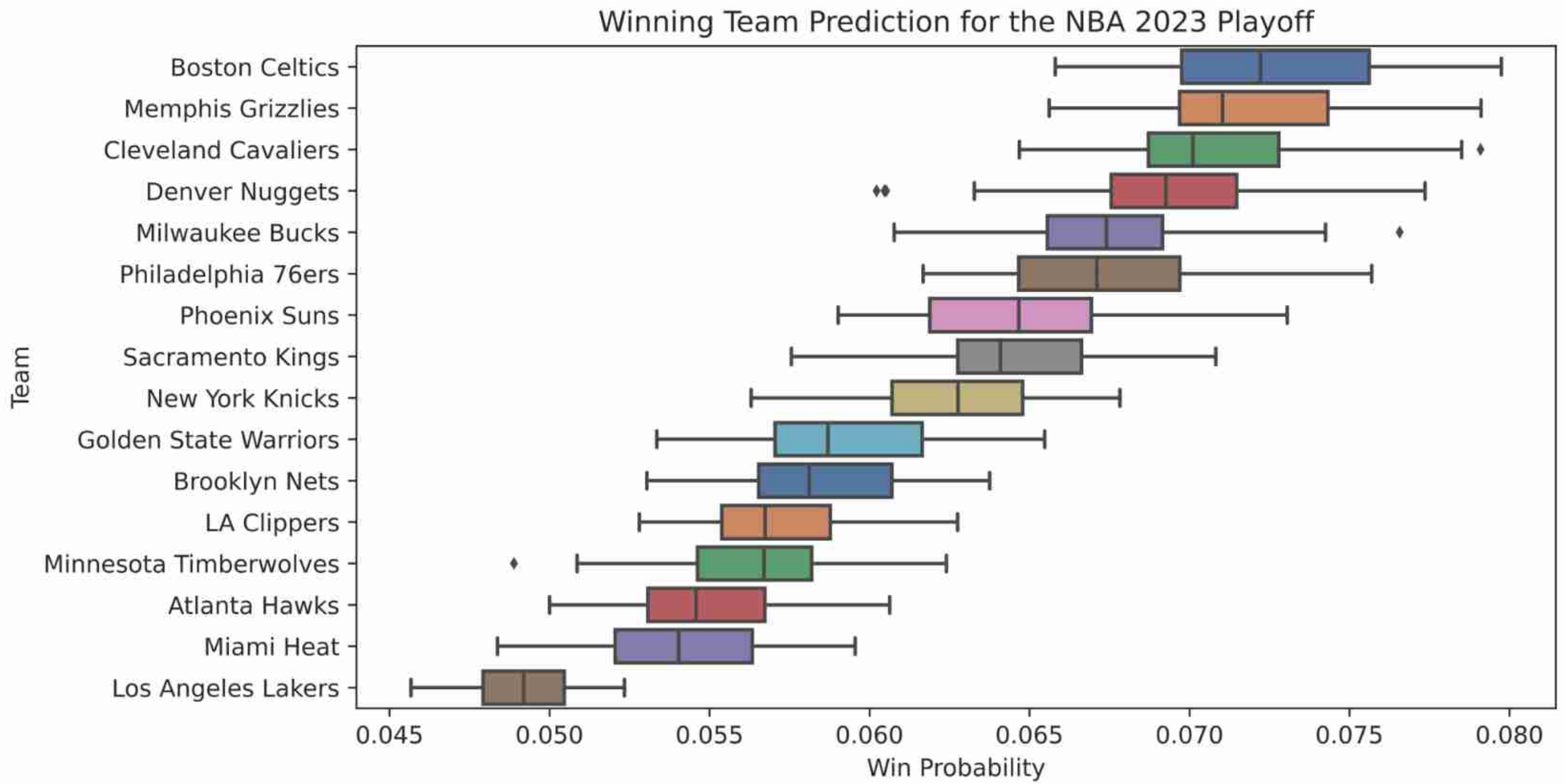
SpeciesRax: A Tool for Maximum Likelihood Species Tree Inference from Gene Family Trees under Duplication, Transfer, and Loss

Benoit Morel, Paul Schade, Sarah Lutteropp, Tom A Williams, Gergely J Szöllősi, Alexandros Stamatakis

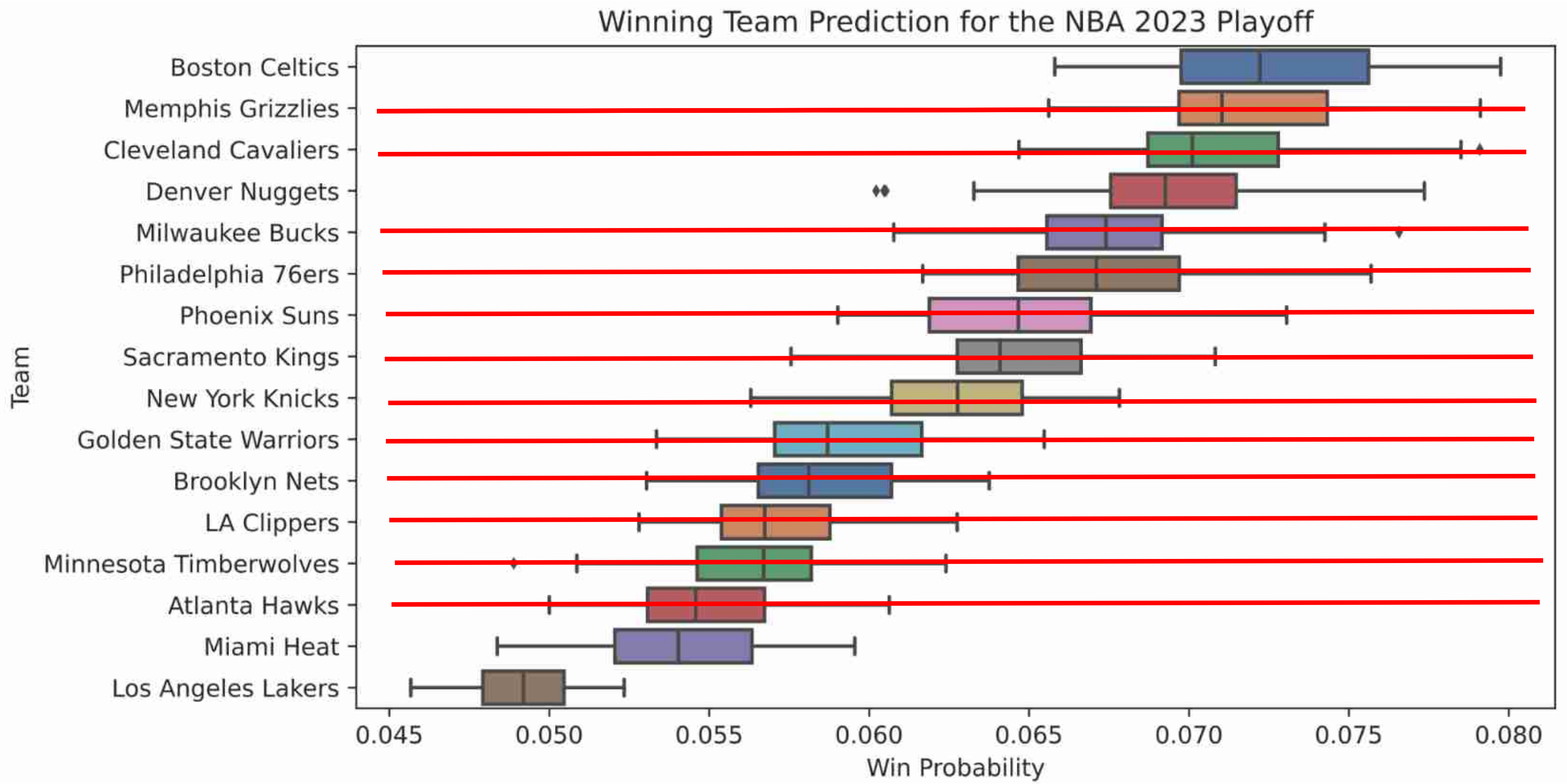
Molecular Biology and Evolution, Volume 39, Issue 2, February 2022, msab365, https://doi.org/10.1093/molbev/msab365

Published: 11 January 2022

Tournament Prediction



Tournament Prediction



Software Quality Assessment

- `SoftWipe` tool for automatic scientific software quality assessment (`C` and `C++`)

Article | Open Access | Published: 11 May 2021

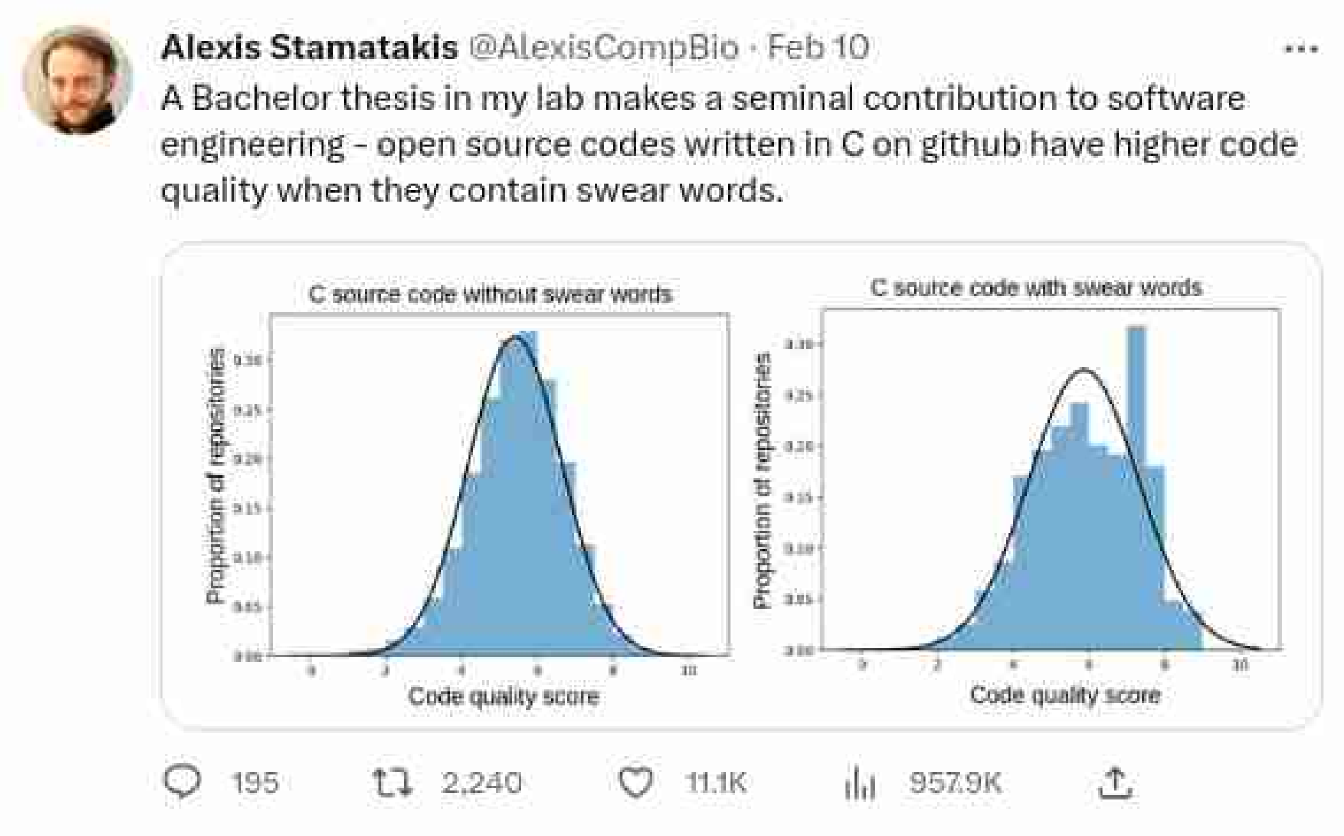
The SoftWipe tool and benchmark for assessing coding standards adherence of scientific software

Adrian Zapletal, Dimitri Höhler, Carsten Sinz & Alexandros Stamatakis

Scientific Reports **11**, Article number: 10015 (2021) | Cite this article

4270 Accesses | **1** Citations | **115** Altmetric | Metrics

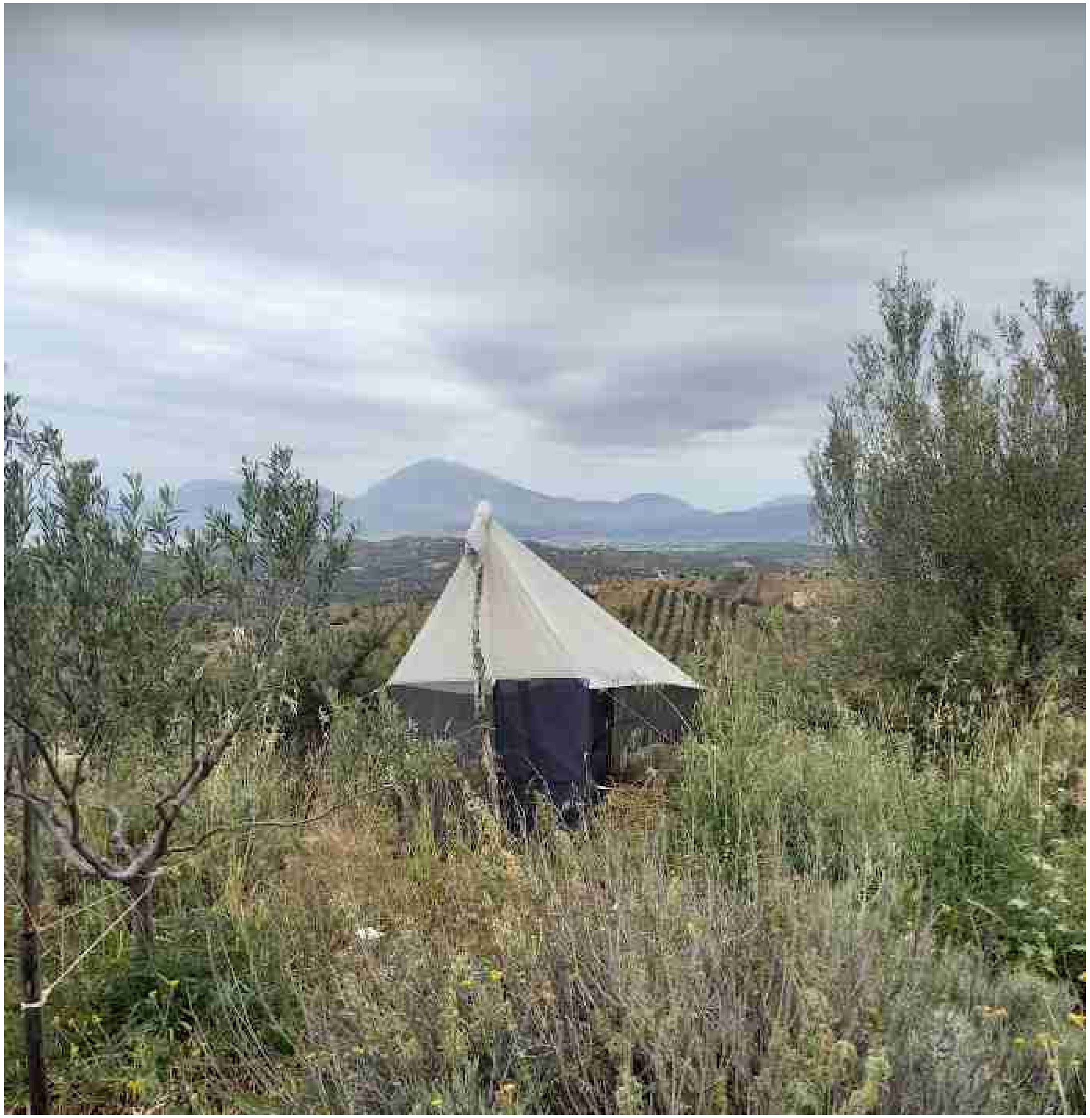
Empirical Software Engineering with `SoftWipe`



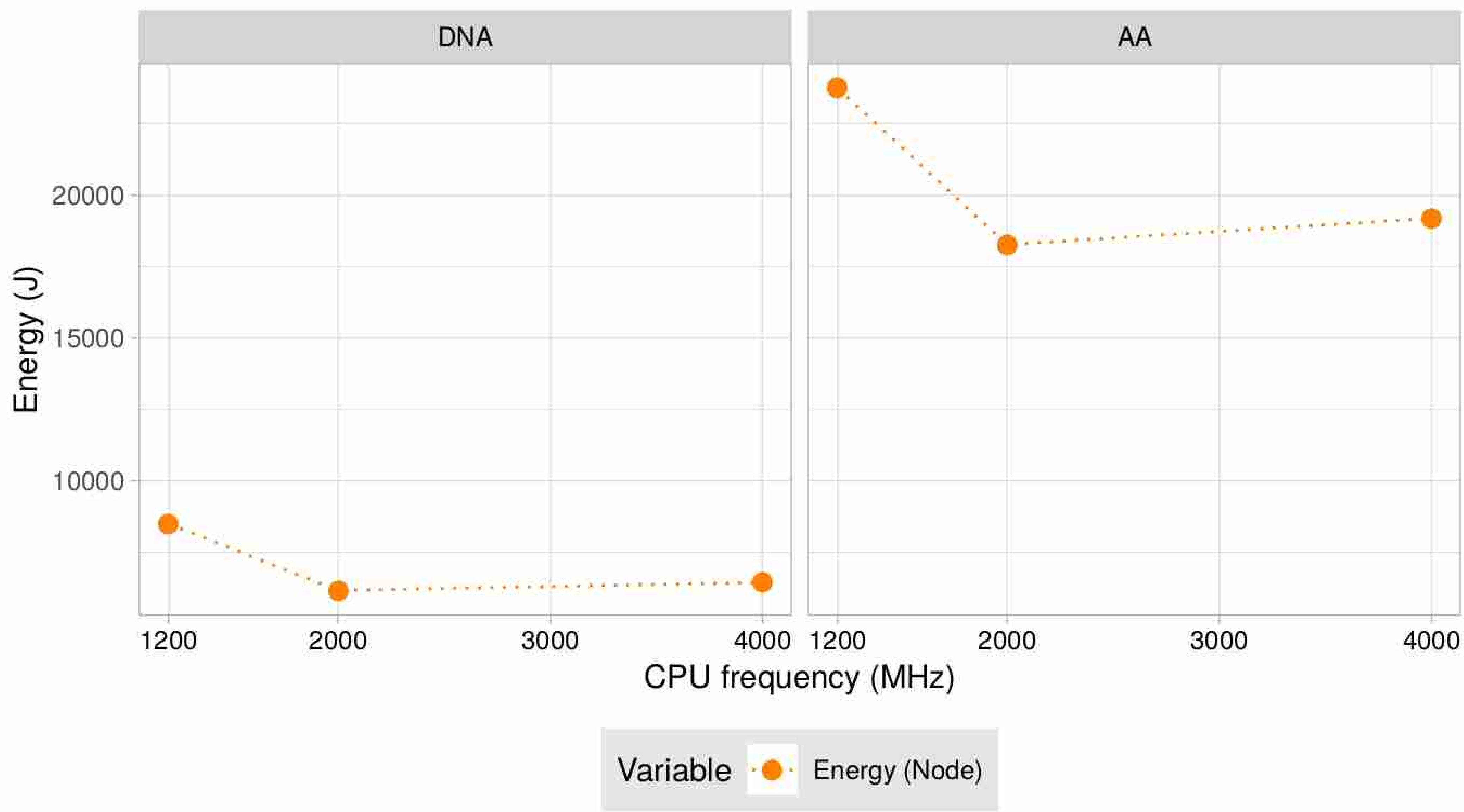
Alexis Stamatakis @AlexisCompBio · Feb 10

A Bachelor thesis in my lab makes a seminal contribution to software engineering – open source codes written in C on github have higher code quality when they contain swear words.

Biological Field Work

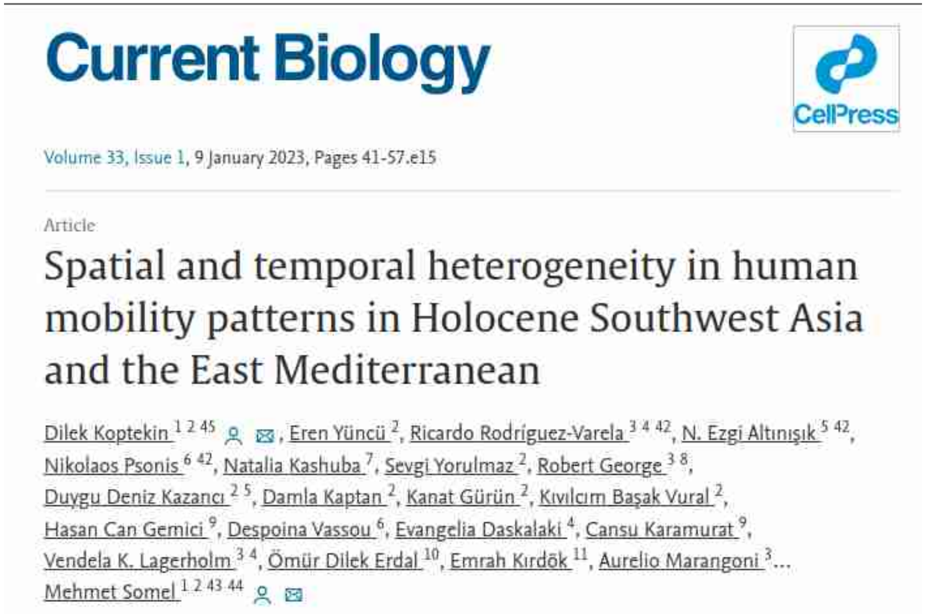


Energy Efficiency



Ancient DNA

- Better tools for ancient DNA analyses
- Classic aDNA data analyses



Current Biology

CellPress

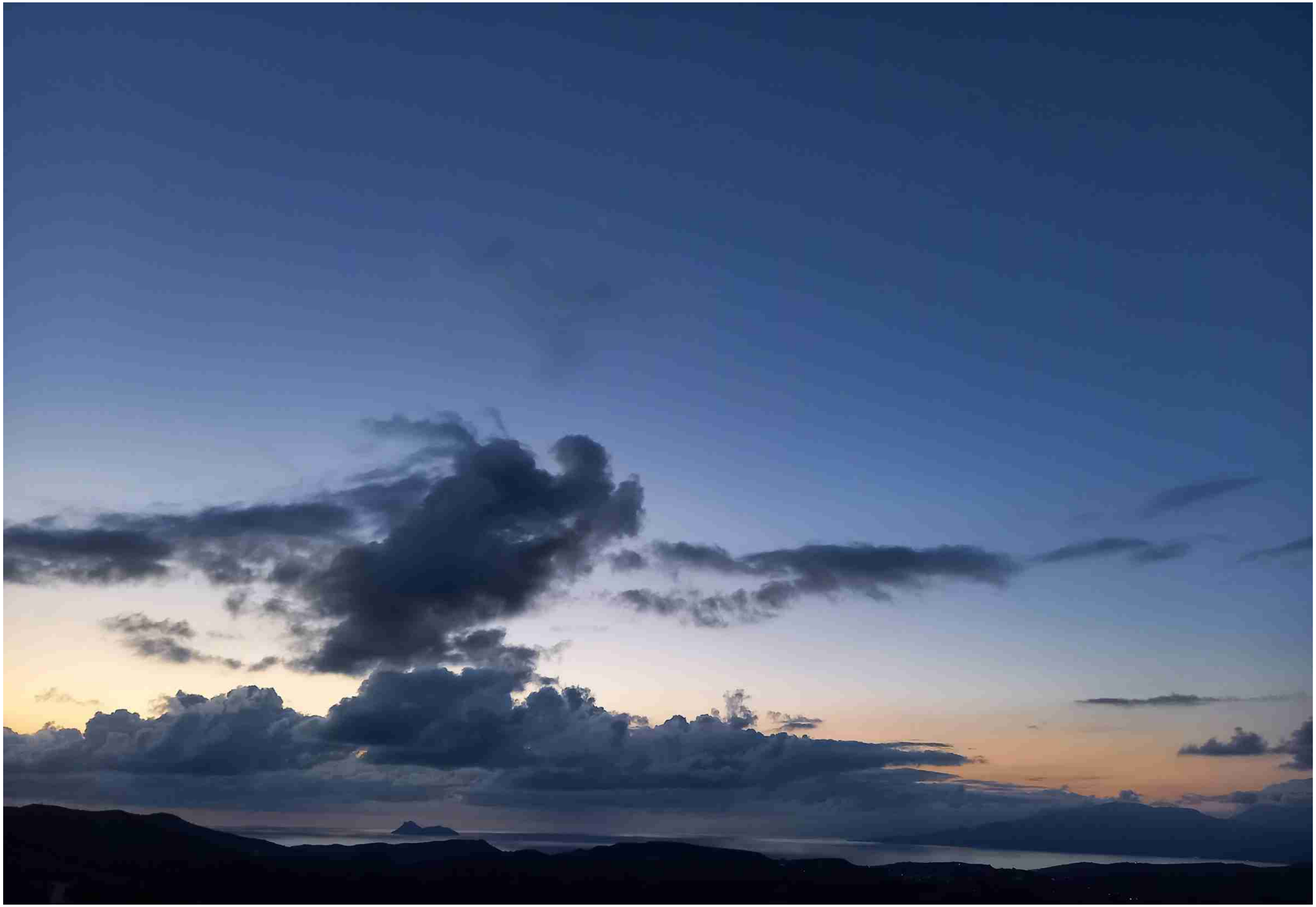
Volume 33, Issue 1, 9 January 2023, Pages 41-57.e15

Article

Spatial and temporal heterogeneity in human mobility patterns in Holocene Southwest Asia and the East Mediterranean

Dilek Koptekin[1, 2, 45], Eren Yüncü[2], Ricardo Rodríguez-Varela[3, 4, 42], N. Ezgi Altınışık[5, 42],
Nikolaos Psonis[6, 42], Natalia Kashuba[7], Sevgi Yorulmaz[2], Robert George[3, 8],
Duygu Deniz Kazancı[2, 5], Damla Kaptan[2], Kanat Gürün[2], Kıvılcım Başak Vural[2],
Hasan Can Gemici[9], Despoina Vassou[6], Evangelia Daskalaki[4], Cansu Karamurat[9],
Vendela K. Lagerholm[3, 4], Ömür Dilek Erdal[10], Emrah Kırdök[11], Aurelio Marangoni[3]…
Mehmet Somel[1, 2, 43, 44]

Thank you for your attention



Pipeline Complexity

Project Complexity

the good old days

Sequence

```
T1 ACGT
T2 ACC
T3 ACGG
T4 AAGC
```

single gene &
few species

Project Complexity

the good old days

Sequence → Align

```
T1 ACGT  T1 ACGT
T2 ACC   T2 ACC-
T3 ACGG  T3 ACGG
T4 AAGC  T4 AAGC
```

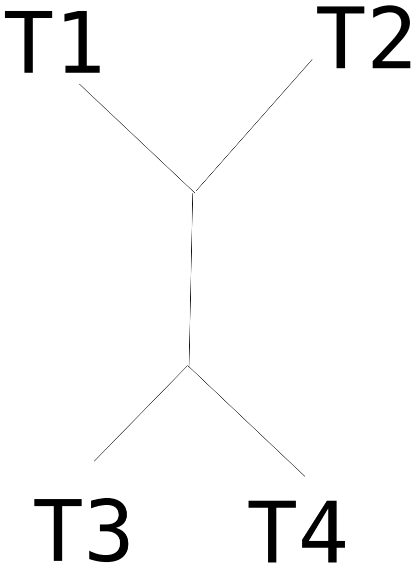
Project Complexity

the good old days

Sequence → Align → Infer Tree

```
T1 ACGT
T2 ACC
T3 ACGG
T4 AAGC
```

```
T1 ACGT
T2 ACC-
T3 ACGG
T4 AAGC
```



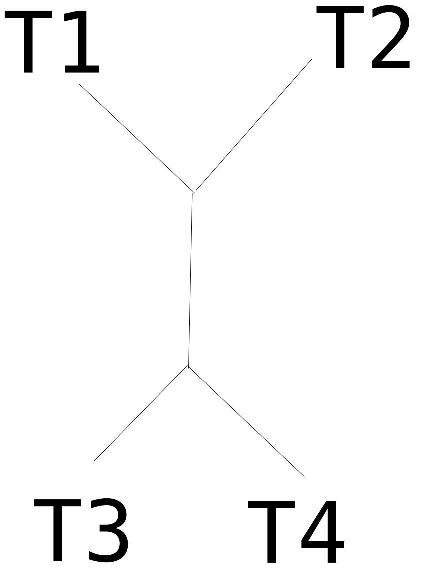
Project Complexity

the good old days

Sequence → Align → Infer Tree → Publish

```
T1 ACGT
T2 ACC
T3 ACGG
T4 AAGC
```

```
T1 ACGT
T2 ACC-
T3 ACGG
T4 AAGC
```



Project Complexity Today

All aTwitter over an Internet study p. XXX

The extragalactic background's uneven glow pp. XXX & XXX

A cellular target for human norovirus pp. XXX & XXX

Science

$10
7 NOVEMBER 2014
sciencemag.org

AAAS

Insect phylogeny resolved

Molecular insights into insect origins and evolution

p. XXX

Does DNA act as a telephone line? p. 1284

Mutations enhancing leukemia development pp. 1291 & 1373

A stable gold support gives sharper resolution p. 1377

Science

$10
12 DECEMBER 2014
sciencemag.org

AAAS

SPECIAL ISSUE

Avian genomes

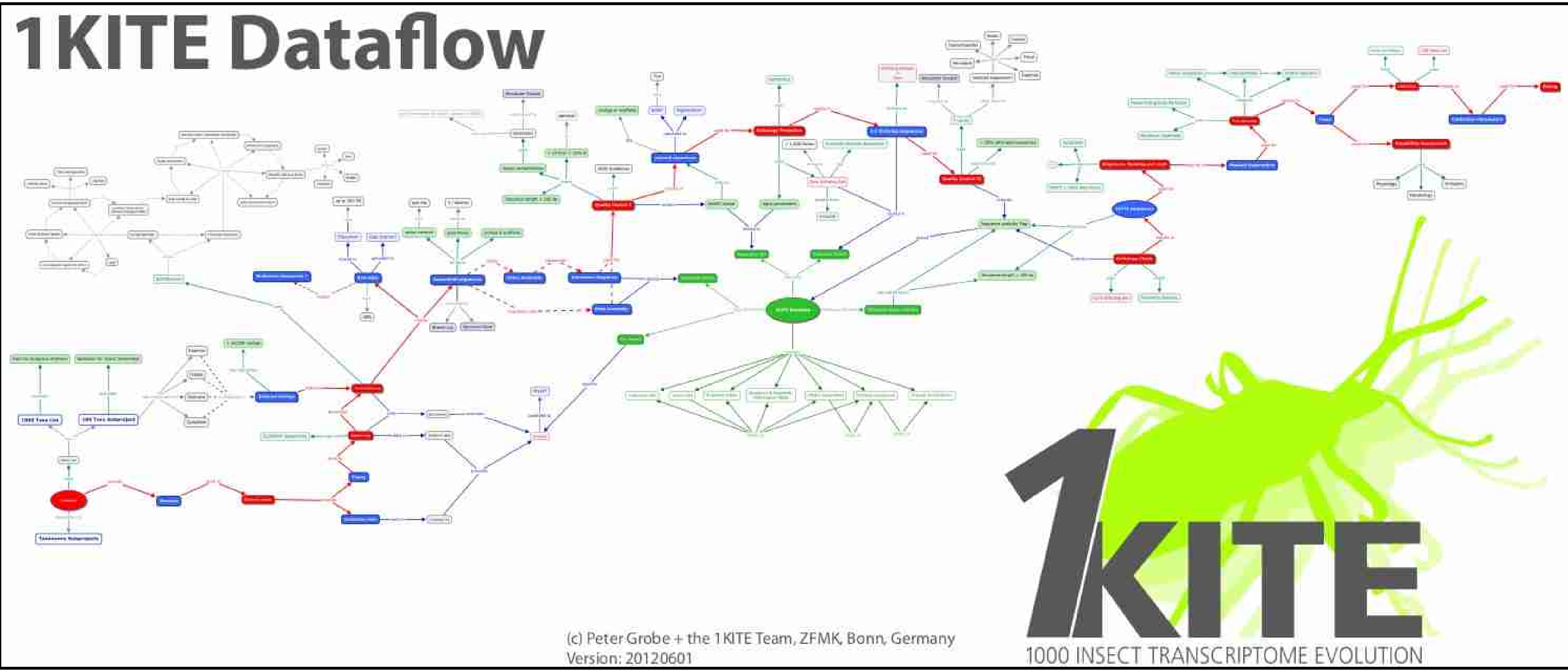
Sequencing across the bird species tree

p. 1308

150 insect transcriptomes

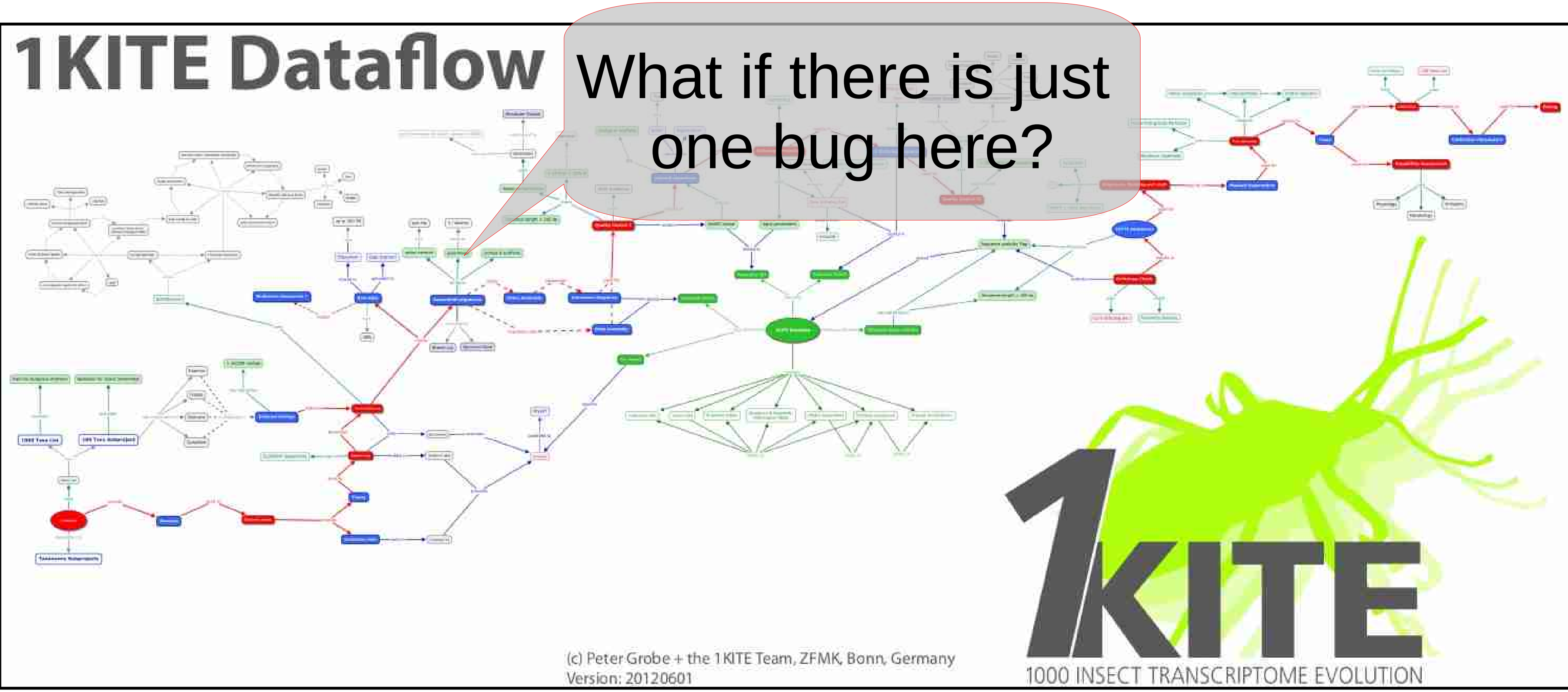
50 bird genomes

Project Complexity Today



(c) Peter Grobe + the 1KITE Team, ZFMK, Bonn, Germany
Version: 20120601

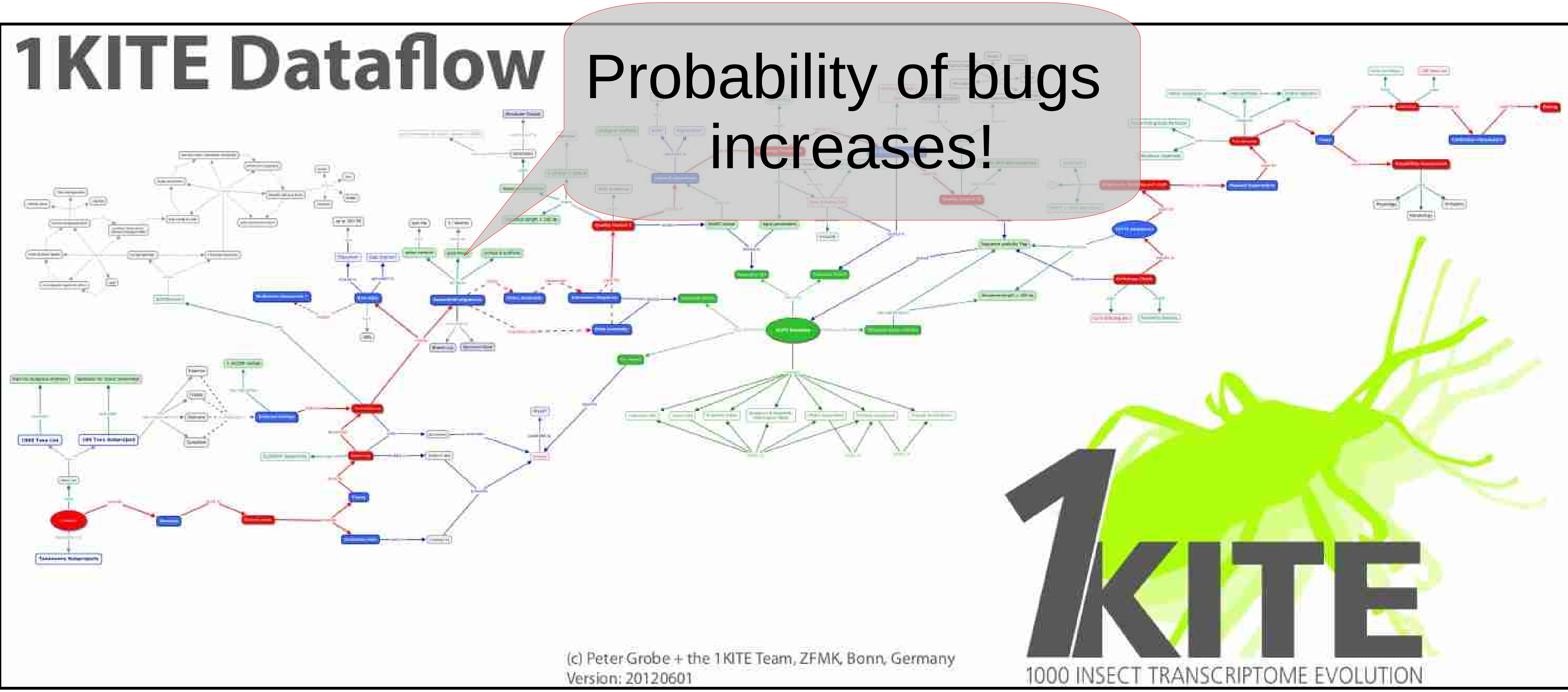
Project Complexity Today



(c) Peter Grobe + the 1KITE Team, ZFMK, Bonn, Germany
Version: 20120601

What if there is just one bug here?

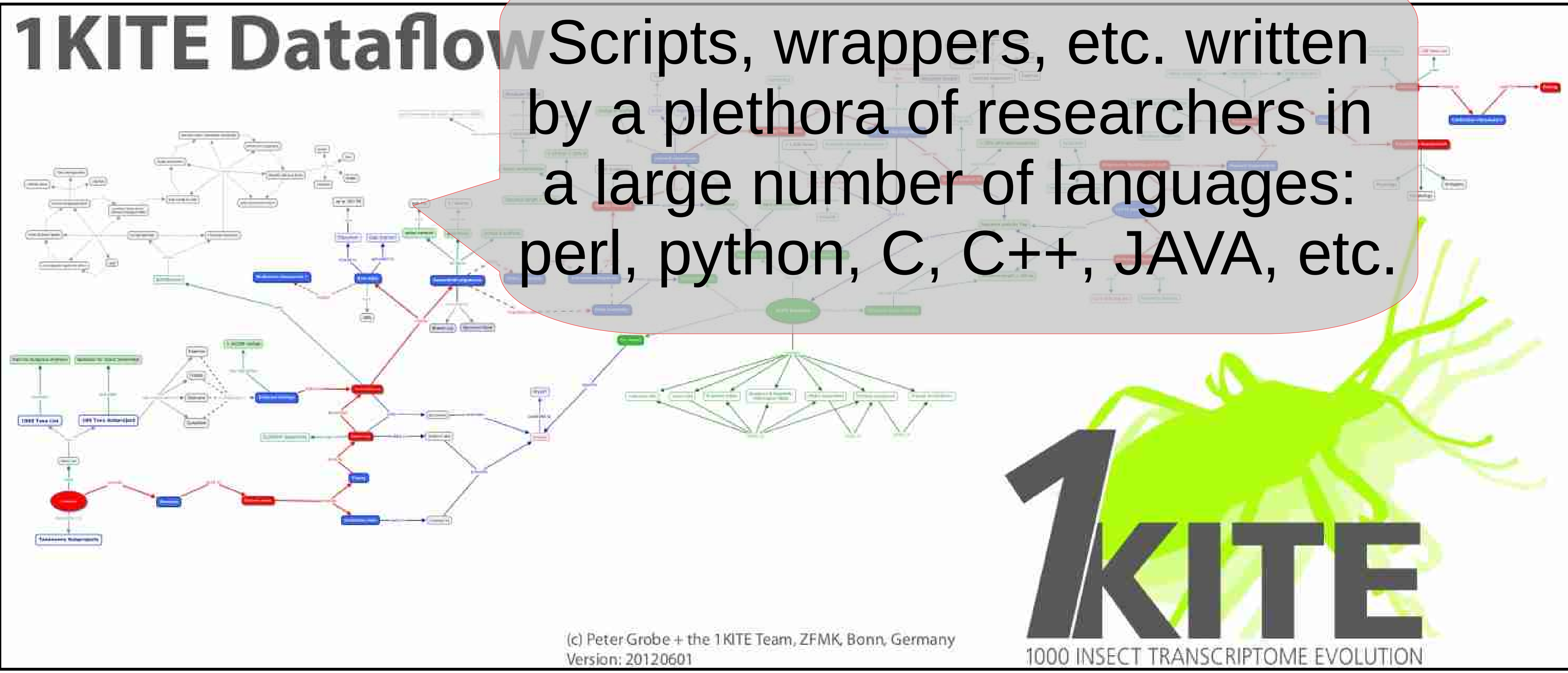
Project Complexity Today



(c) Peter Grobe + the 1KITE Team, ZFMK, Bonn, Germany
Version: 20120601

Probability of bugs increases!

Project Complexity Today

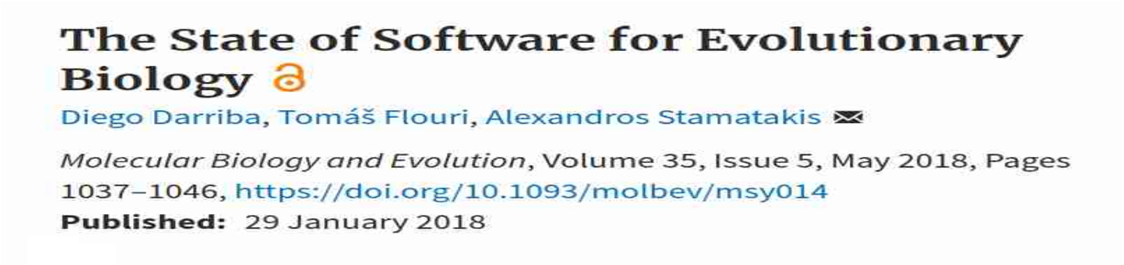


(c) Peter Grobe + the 1KITE Team, ZFMK, Bonn, Germany
Version: 20120601

Scripts, wrappers, etc. written by a plethora of researchers in a large number of languages: perl, python, C, C++, JAVA, etc.

The 'crappy' software project

- Analyzed *15* widely-used evolutionary biology tools ≈ *65,000* citations
- Analyses performed
  - Compiled with `gcc` and `clang` with all warnings enabled
  - Memory check with `valgrind`
  - Checked if assertions are used via `assert()`
  - Analyzed degree of code duplication
- **Caution:** “bad” quality does not induce that a tool is faulty, but the probability of it being faulty is higher!



The State of Software for Evolutionary Biology

Diego Darriba, Tomáš Flouri, Alexandros Stamatakis

Molecular Biology and Evolution, Volume 35, Issue 5, May 2018, Pages 1037–1046, https://doi.org/10.1093/molbev/msy014

Published: 29 January 2018

SoftWipe

- Discussion with Science Journalist - *“Can this process be automated?”*
- Development of SoftWipe - An automated tool and benchmark for **relative** quality ranking of scientific software
- Ranking of *51* open source tools written in `C` or `C++` from a wide range of research areas
  - Astrophysics
  - Computer Science
  - Bioinformatics

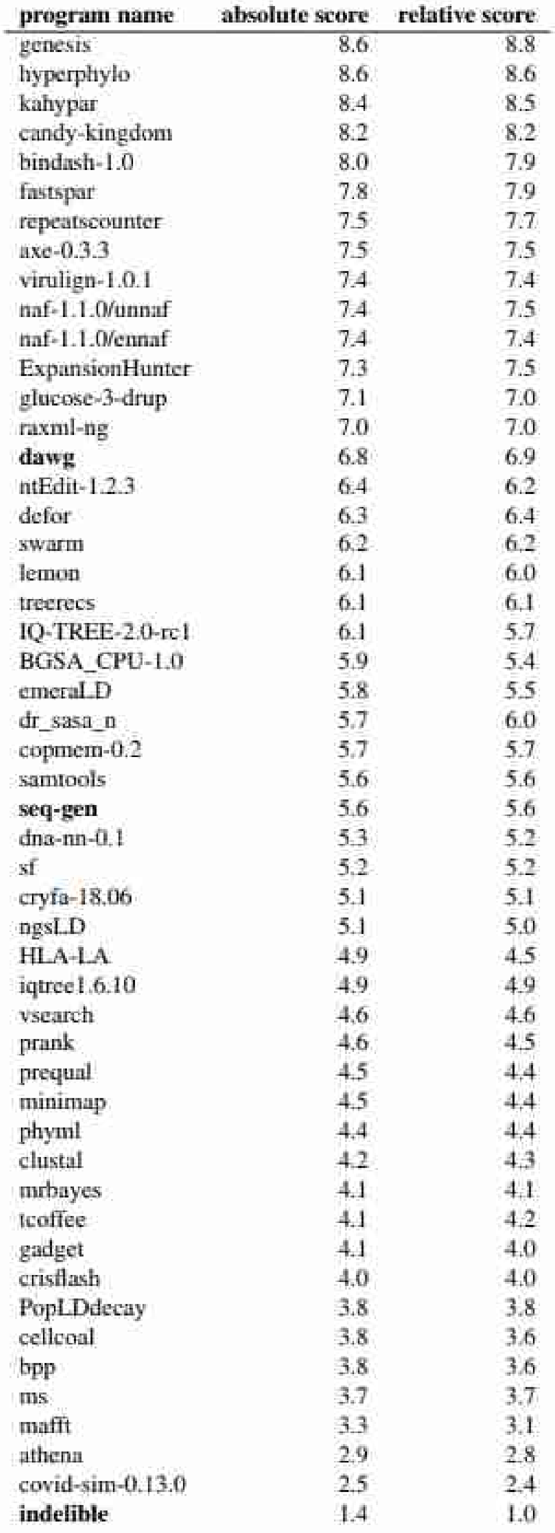
New Results

SoftWipe - a tool and benchmark to assess scientific software quality

Adrian Zapletal, Dimitri Hoehler, Carsten Sinz, Alexandros Stamatakis

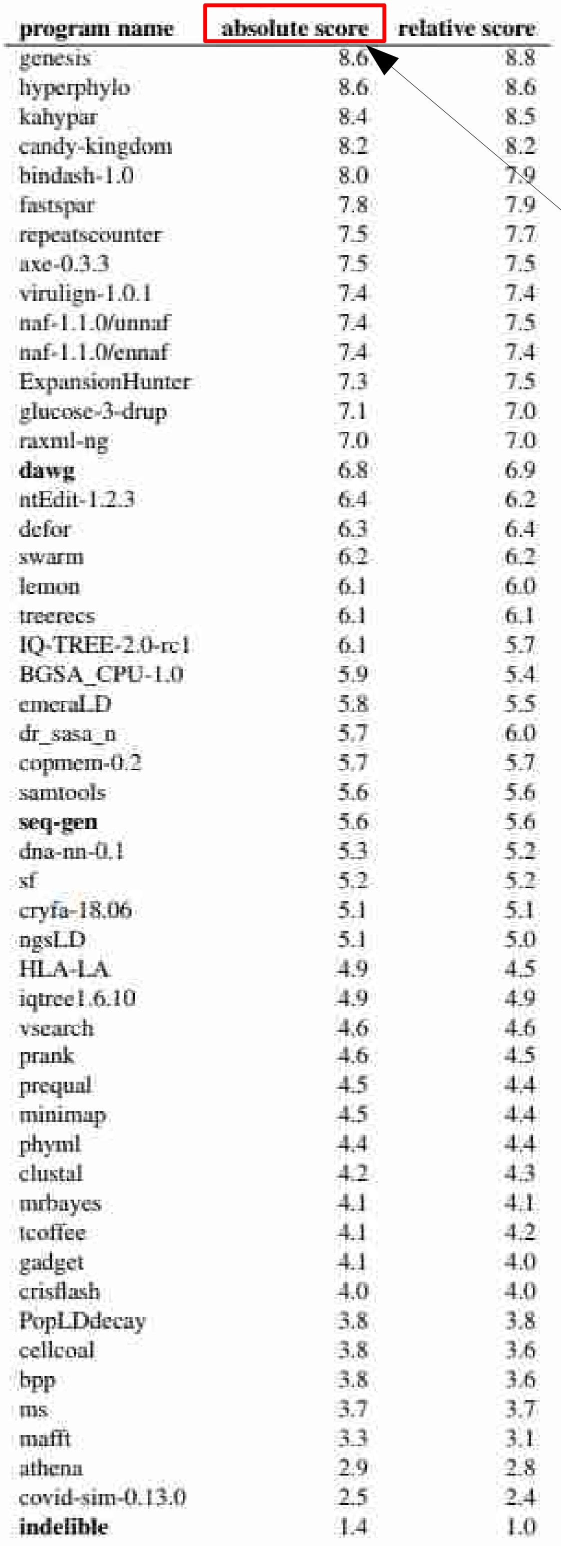
doi: https://doi.org/10.1101/2020.10.07.330621

SoftWipe Benchmark



program name	absolute score	relative score
genesis	8.6	8.8
hyperphylo	8.6	8.6
kahypar	8.4	8.5
candy-kingdom	8.2	8.2
bindash-1.0	8.0	7.9
fastspar	7.8	7.9
repeatscounter	7.5	7.7
axe-0.3.3	7.5	7.5
virulign-1.0.1	7.4	7.4
naf-1.1.0/unnaf	7.4	7.5
naf-1.1.0/ennaf	7.4	7.4
ExpansionHunter	7.3	7.5
glucose-3-drup	7.1	7.0
raxml-ng	7.0	7.0
dawg	6.8	6.9
ntEdit-1.2.3	6.4	6.2
defor	6.3	6.4
swarm	6.2	6.2
lemon	6.1	6.0
treerecs	6.1	6.1
IQ-TREE-2.0-rc1	6.1	5.7
BGSA_CPU-1.0	5.9	5.4
emeraLD	5.8	5.5
dr_sasa_n	5.7	6.0
copmem-0.2	5.7	5.7
samtools	5.6	5.6
seq-gen	5.6	5.6
dna-nn-0.1	5.3	5.2
sf	5.2	5.2
cryfa-18.06	5.1	5.1
ngsLD	5.1	5.0
HLA-LA	4.9	4.5
iqtree1.6.10	4.9	4.9
vsearch	4.6	4.6
prank	4.6	4.5
prequal	4.5	4.4
minimap	4.5	4.4
phyml	4.4	4.4
clustal	4.2	4.3
mrbayes	4.1	4.1
tcoffee	4.1	4.2
gadget	4.1	4.0
crisflash	4.0	4.0
PopLDdecay	3.8	3.8
cellcoal	3.8	3.6
bpp	3.8	3.6
ms	3.7	3.7
mafft	3.3	3.1
athena	2.9	2.8
covid-sim-0.13.0	2.5	2.4
indelible	1.4	1.0

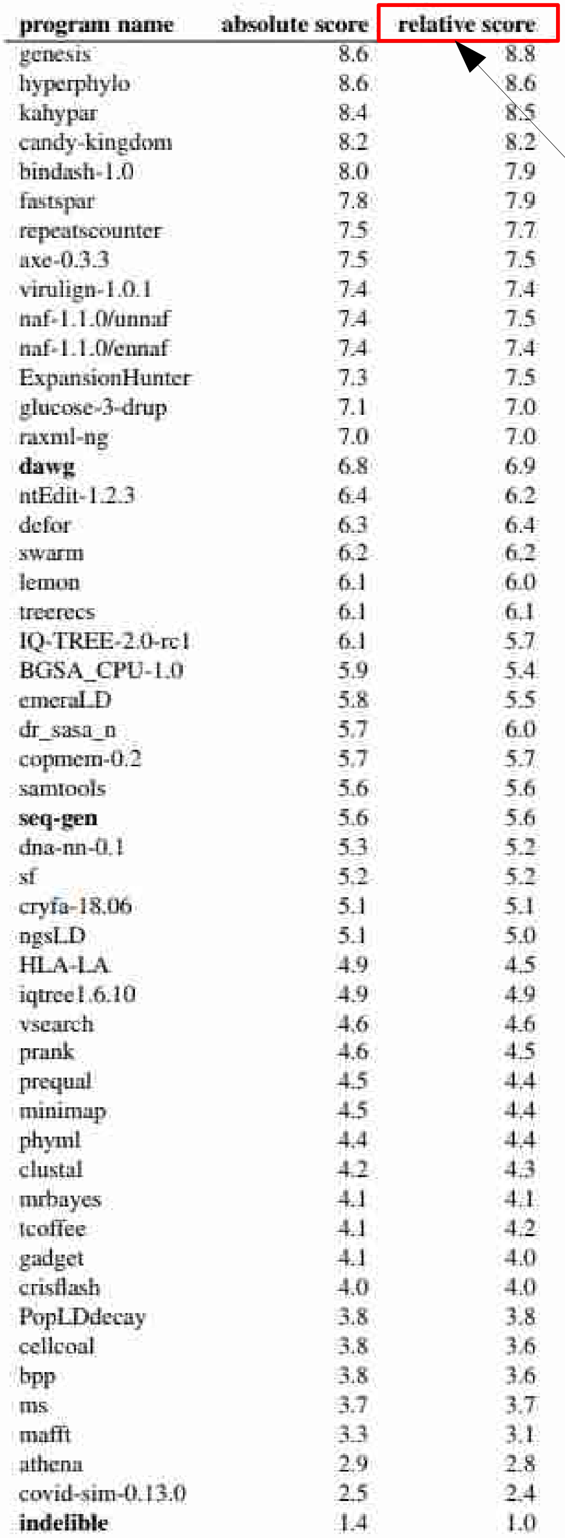
SoftWipe Benchmark



program name	absolute score	relative score
genesis	8.6	8.8
hyperphylo	8.6	8.6
kahypar	8.4	8.5
candy-kingdom	8.2	8.2
bindash-1.0	8.0	7.9
fastspar	7.8	7.9
repeatscounter	7.5	7.7
axe-0.3.3	7.5	7.5
virulign-1.0.1	7.4	7.4
naf-1.1.0/unnaf	7.4	7.5
naf-1.1.0/ennaf	7.4	7.4
ExpansionHunter	7.3	7.5
glucose-3-drup	7.1	7.0
raxml-ng	7.0	7.0
dawg	6.8	6.9
ntEdit-1.2.3	6.4	6.2
defor	6.3	6.4
swarm	6.2	6.2
lemon	6.1	6.0
treerecs	6.1	6.1
IQ-TREE-2.0-rc1	6.1	5.7
BGSA_CPU-1.0	5.9	5.4
emeraLD	5.8	5.5
dr_sasa_n	5.7	6.0
copmem-0.2	5.7	5.7
samtools	5.6	5.6
seq-gen	5.6	5.6
dna-nn-0.1	5.3	5.2
sf	5.2	5.2
cryfa-18.06	5.1	5.1
ngsLD	5.1	5.0
HLA-LA	4.9	4.5
iqtree1.6.10	4.9	4.9
vsearch	4.6	4.6
prank	4.6	4.5
prequal	4.5	4.4
minimap	4.5	4.4
phyml	4.4	4.4
clustal	4.2	4.3
mrbayes	4.1	4.1
tcoffee	4.1	4.2
gadget	4.1	4.0
crisflash	4.0	4.0
PopLDdecay	3.8	3.8
cellcoal	3.8	3.6
bpp	3.8	3.6
ms	3.7	3.7
mafft	3.3	3.1
athena	2.9	2.8
covid-sim-0.13.0	2.5	2.4
indelible	1.4	1.0

Does not change over time as more tools are added → can easily be referenced

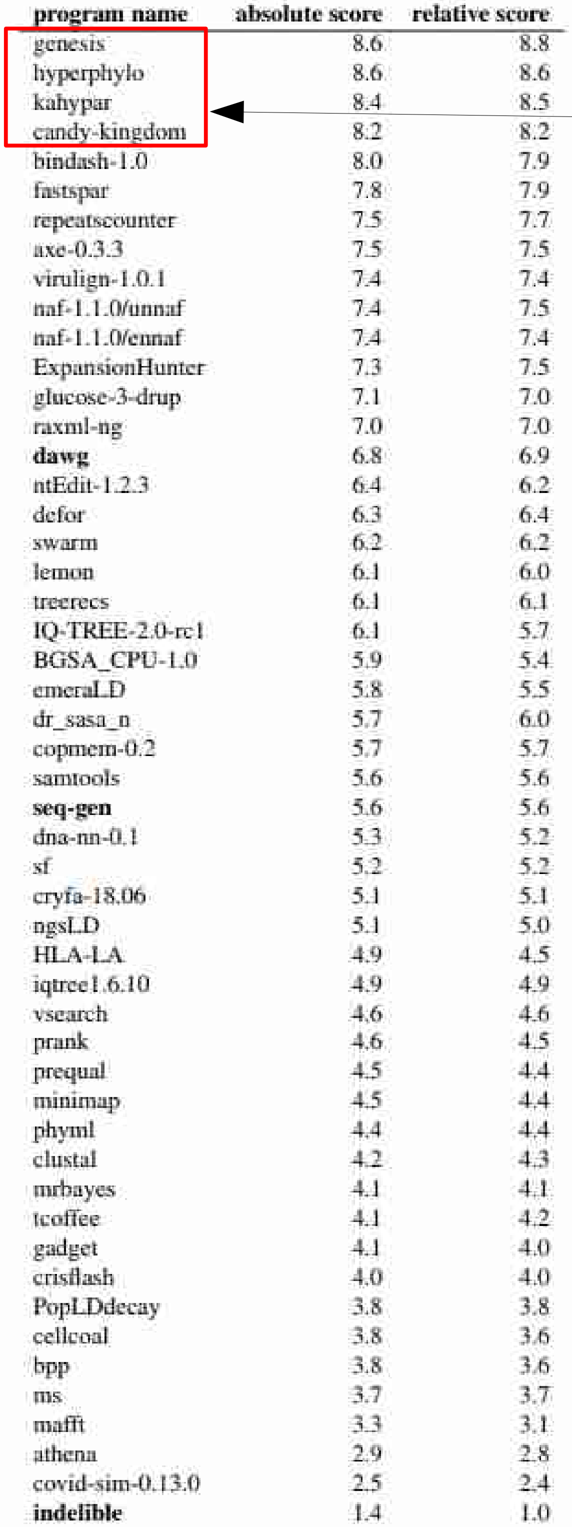
SoftWipe Benchmark



program name	**absolute score**	**relative score**
genesis	8.6	8.8
hyperphylo	8.6	8.6
kahypar	8.4	8.5
candy-kingdom	8.2	8.2
bindash-1.0	8.0	7.9
fastspar	7.8	7.9
repeatscounter	7.5	7.7
axe-0.3.3	7.5	7.5
virulign-1.0.1	7.4	7.4
naf-1.1.0/unnaf	7.4	7.5
naf-1.1.0/ennaf	7.4	7.4
ExpansionHunter	7.3	7.5
glucose-3-drup	7.1	7.0
raxml-ng	7.0	7.0
dawg	6.8	6.9
ntEdit-1.2.3	6.4	6.2
defor	6.3	6.4
swarm	6.2	6.2
lemon	6.1	6.0
treerecs	6.1	6.1
IQ-TREE-2.0-rc1	6.1	5.7
BGSA_CPU-1.0	5.9	5.4
emeraLD	5.8	5.5
dr_sasa_n	5.7	6.0
copmem-0.2	5.7	5.7
samtools	5.6	5.6
seq-gen	5.6	5.6
dna-nn-0.1	5.3	5.2
sf	5.2	5.2
cryfa-18.06	5.1	5.1
ngsLD	5.1	5.0
HLA-LA	4.9	4.5
iqtree1.6.10	4.9	4.9
vsearch	4.6	4.6
prank	4.6	4.5
prequal	4.5	4.4
minimap	4.5	4.4
phyml	4.4	4.4
clustal	4.2	4.3
mrbayes	4.1	4.1
tcoffee	4.1	4.2
gadget	4.1	4.0
crisflash	4.0	4.0
PopLDdecay	3.8	3.8
cellcoal	3.8	3.6
bpp	3.8	3.6
ms	3.7	3.7
mafft	3.3	3.1
athena	2.9	2.8
covid-sim-0.13.0	2.5	2.4
indelible	1.4	1.0

Does change over time as more tools are added →
Difficult to be referenced

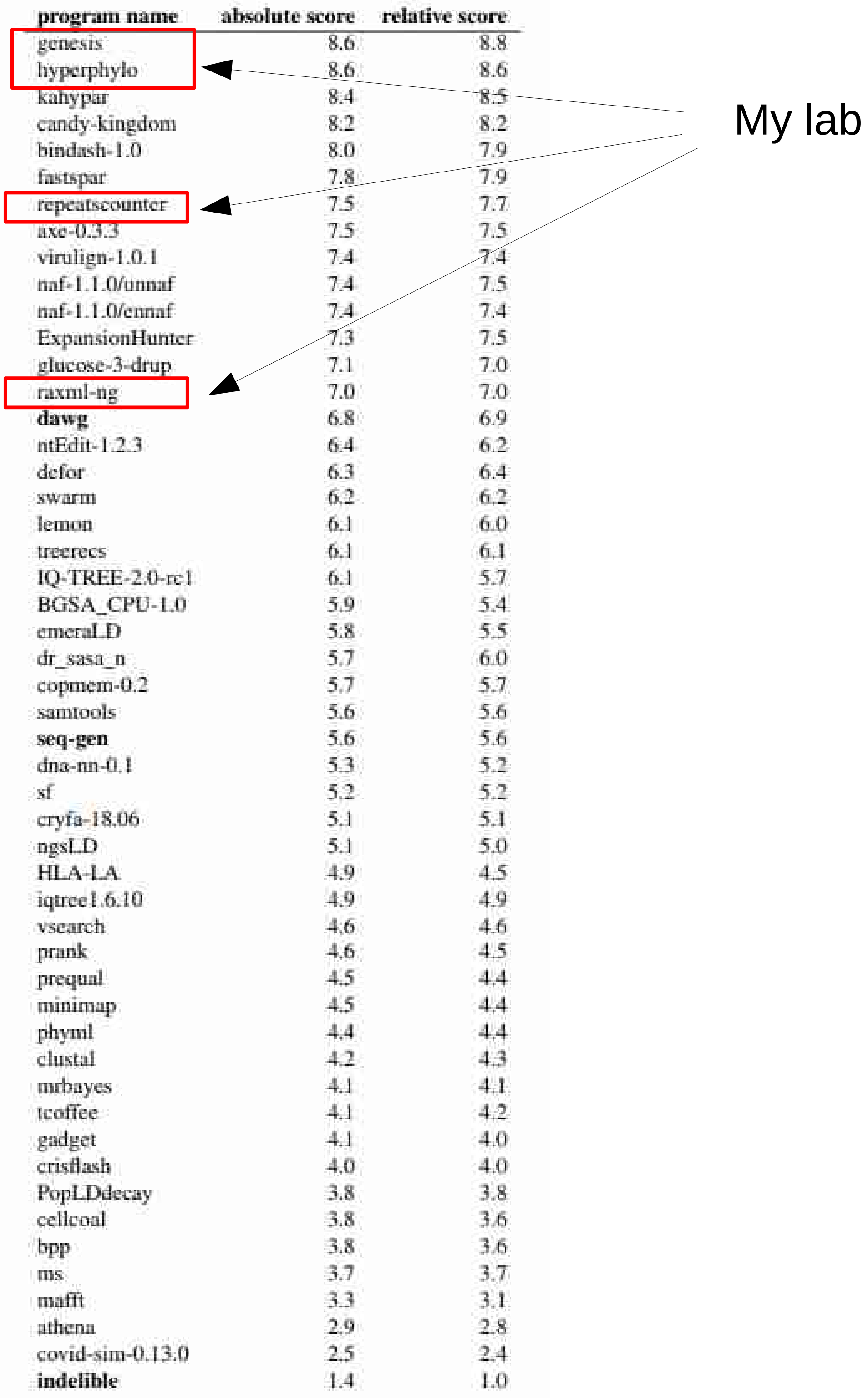
SoftWipe Benchmark



program name	**absolute score**	**relative score**
genesis	8.6	8.8
hyperphylo	8.6	8.6
kahypar	8.4	8.5
candy-kingdom	8.2	8.2
bindash-1.0	8.0	7.9
fastspar	7.8	7.9
repeatscounter	7.5	7.7
axe-0.3.3	7.5	7.5
virulign-1.0.1	7.4	7.4
naf-1.1.0/unnaf	7.4	7.5
naf-1.1.0/ennaf	7.4	7.4
ExpansionHunter	7.3	7.5
glucose-3-drup	7.1	7.0
raxml-ng	7.0	7.0
dawg	6.8	6.9
ntEdit-1.2.3	6.4	6.2
defor	6.3	6.4
swarm	6.2	6.2
lemon	6.1	6.0
treerecs	6.1	6.1
IQ-TREE-2.0-rc1	6.1	5.7
BGSA_CPU-1.0	5.9	5.4
emeraLD	5.8	5.5
dr_sasa_n	5.7	6.0
copmem-0.2	5.7	5.7
samtools	5.6	5.6
seq-gen	5.6	5.6
dna-nn-0.1	5.3	5.2
sf	5.2	5.2
cryfa-18.06	5.1	5.1
ngsLD	5.1	5.0
HLA-LA	4.9	4.5
iqtree1.6.10	4.9	4.9
vsearch	4.6	4.6
prank	4.6	4.5
prequal	4.5	4.4
minimap	4.5	4.4
phyml	4.4	4.4
clustal	4.2	4.3
mrbayes	4.1	4.1
tcoffee	4.1	4.2
gadget	4.1	4.0
crisflash	4.0	4.0
PopLDdecay	3.8	3.8
cellcoal	3.8	3.6
bpp	3.8	3.6
ms	3.7	3.7
mafft	3.3	3.1
athena	2.9	2.8
covid-sim-0.13.0	2.5	2.4
indelible	1.4	1.0

Written by computer scientists

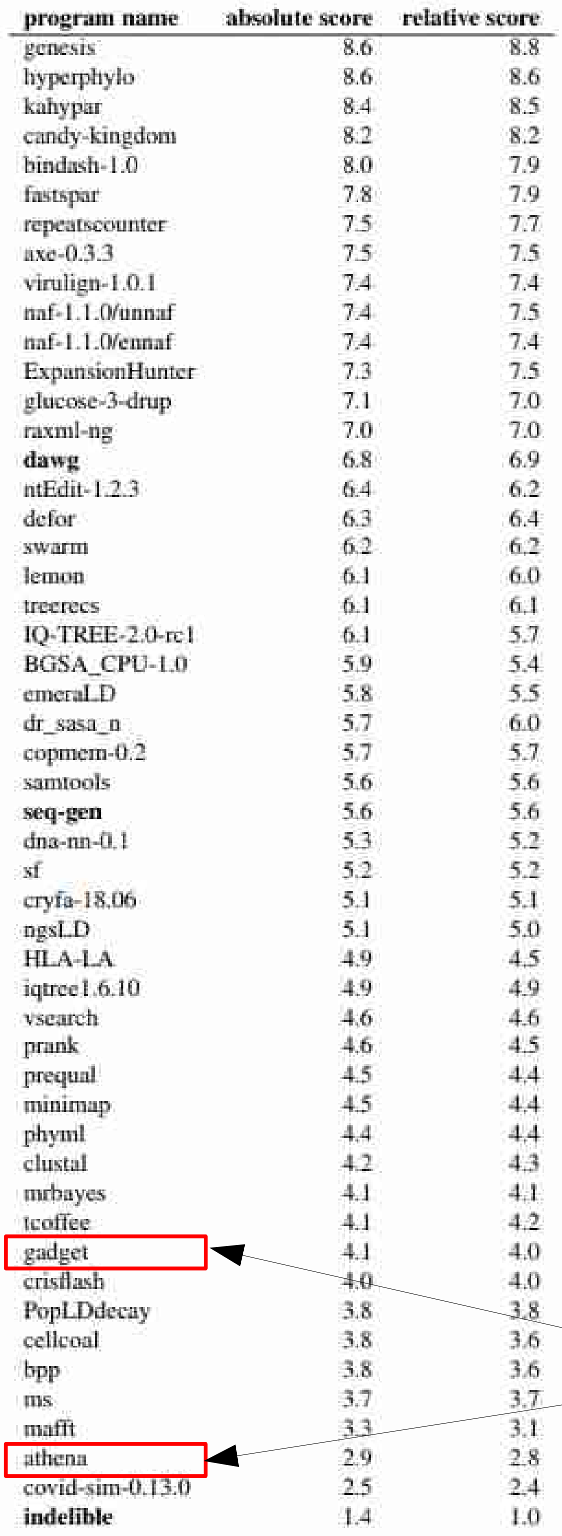
SoftWipe Benchmark



My lab

program name	absolute score	relative score
genesis	8.6	8.8
hyperphylo	8.6	8.6
kahypar	8.4	8.5
candy-kingdom	8.2	8.2
bindash-1.0	8.0	7.9
fastspar	7.8	7.9
repeatscounter	7.5	7.7
axe-0.3.3	7.5	7.5
virulign-1.0.1	7.4	7.4
naf-1.1.0/unnaf	7.4	7.5
naf-1.1.0/ennaf	7.4	7.4
ExpansionHunter	7.3	7.5
glucose-3-drup	7.1	7.0
raxml-ng	7.0	7.0
dawg	6.8	6.9
ntEdit-1.2.3	6.4	6.2
defor	6.3	6.4
swarm	6.2	6.2
lemon	6.1	6.0
treerecs	6.1	6.1
IQ-TREE-2.0-rc1	6.1	5.7
BGSA_CPU-1.0	5.9	5.4
emeraLD	5.8	5.5
dr_sasa_n	5.7	6.0
copmem-0.2	5.7	5.7
samtools	5.6	5.6
seq-gen	5.6	5.6
dna-nn-0.1	5.3	5.2
sf	5.2	5.2
cryfa-18.06	5.1	5.1
ngsLD	5.1	5.0
HLA-LA	4.9	4.5
iqtree1.6.10	4.9	4.9
vsearch	4.6	4.6
prank	4.6	4.5
prequal	4.5	4.4
minimap	4.5	4.4
phyml	4.4	4.4
clustal	4.2	4.3
mrbayes	4.1	4.1
tcoffee	4.1	4.2
gadget	4.1	4.0
crisflash	4.0	4.0
PopLDdecay	3.8	3.8
cellcoal	3.8	3.6
bpp	3.8	3.6
ms	3.7	3.7
mafft	3.3	3.1
athena	2.9	2.8
covid-sim-0.13.0	2.5	2.4
indelible	1.4	1.0

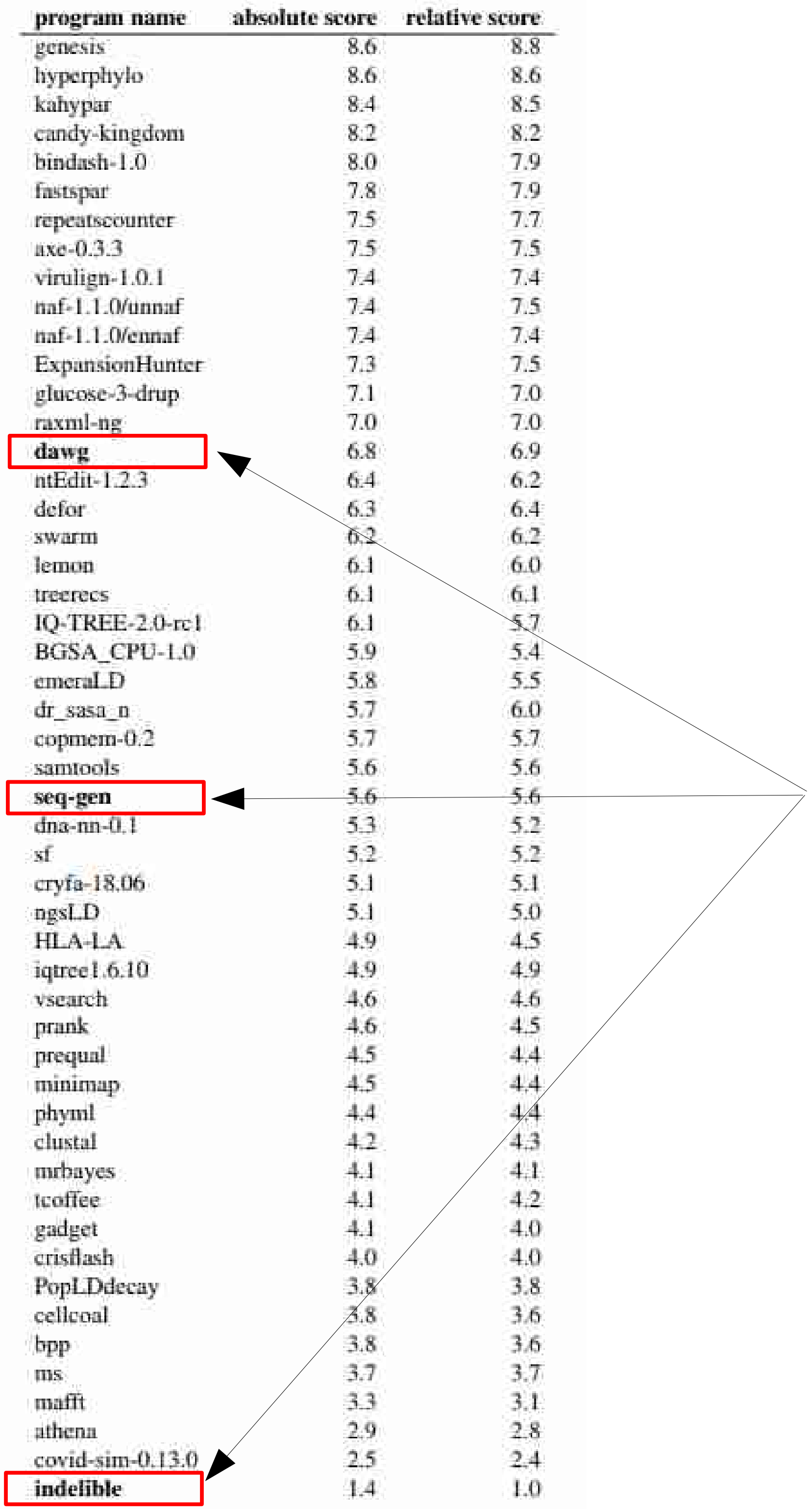
SoftWipe Benchmark



program name	absolute score	relative score
genesis	8.6	8.8
hyperphylo	8.6	8.6
kahypar	8.4	8.5
candy-kingdom	8.2	8.2
bindash-1.0	8.0	7.9
fastspar	7.8	7.9
repeatscounter	7.5	7.7
axe-0.3.3	7.5	7.5
virulign-1.0.1	7.4	7.4
naf-1.1.0/unnaf	7.4	7.5
naf-1.1.0/ennaf	7.4	7.4
ExpansionHunter	7.3	7.5
glucose-3-drup	7.1	7.0
raxml-ng	7.0	7.0
dawg	6.8	6.9
ntEdit-1.2.3	6.4	6.2
defor	6.3	6.4
swarm	6.2	6.2
lemon	6.1	6.0
treerecs	6.1	6.1
IQ-TREE-2.0-rc1	6.1	5.7
BGSA_CPU-1.0	5.9	5.4
emeraLD	5.8	5.5
dr_sasa_n	5.7	6.0
copmem-0.2	5.7	5.7
samtools	5.6	5.6
seq-gen	5.6	5.6
dna-nn-0.1	5.3	5.2
sf	5.2	5.2
cryfa-18.06	5.1	5.1
ngsLD	5.1	5.0
HLA-LA	4.9	4.5
iqtree1.6.10	4.9	4.9
vsearch	4.6	4.6
prank	4.6	4.5
prequal	4.5	4.4
minimap	4.5	4.4
phyml	4.4	4.4
clustal	4.2	4.3
mrbayes	4.1	4.1
tcoffee	4.1	4.2
gadget	4.1	4.0
crisflash	4.0	4.0
PopLDdecay	3.8	3.8
cellcoal	3.8	3.6
bpp	3.8	3.6
ms	3.7	3.7
mafft	3.3	3.1
athena	2.9	2.8
covid-sim-0.13.0	2.5	2.4
indelible	1.4	1.0

Astrophysics

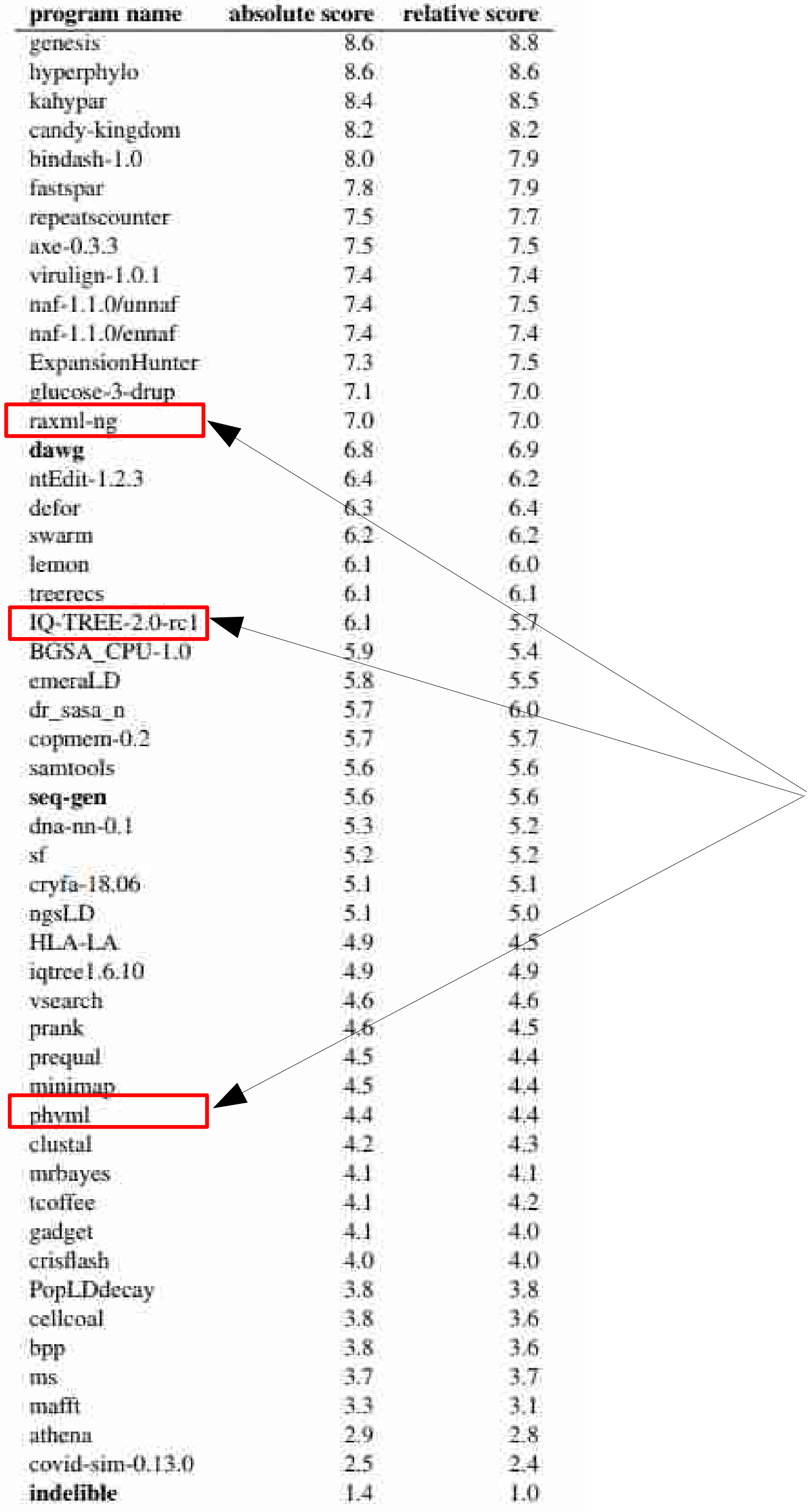
SoftWipe Benchmark



program name	absolute score	relative score
genesis	8.6	8.8
hyperphylo	8.6	8.6
kahypar	8.4	8.5
candy-kingdom	8.2	8.2
bindash-1.0	8.0	7.9
fastspar	7.8	7.9
repeatscounter	7.5	7.7
axe-0.3.3	7.5	7.5
virulign-1.0.1	7.4	7.4
naf-1.1.0/unnaf	7.4	7.5
naf-1.1.0/ennaf	7.4	7.4
ExpansionHunter	7.3	7.5
glucose-3-drup	7.1	7.0
raxml-ng	7.0	7.0
dawg	6.8	6.9
ntEdit-1.2.3	6.4	6.2
defor	6.3	6.4
swarm	6.2	6.2
lemon	6.1	6.0
treerecs	6.1	6.1
IQ-TREE-2.0-rc1	6.1	5.7
BGSA_CPU-1.0	5.9	5.4
emeraLD	5.8	5.5
dr_sasa_n	5.7	6.0
copmem-0.2	5.7	5.7
samtools	5.6	5.6
seq-gen	5.6	5.6
dna-nn-0.1	5.3	5.2
sf	5.2	5.2
cryfa-18.06	5.1	5.1
ngsLD	5.1	5.0
HLA-LA	4.9	4.5
iqtree1.6.10	4.9	4.9
vsearch	4.6	4.6
prank	4.6	4.5
prequal	4.5	4.4
minimap	4.5	4.4
phyml	4.4	4.4
clustal	4.2	4.3
mrbayes	4.1	4.1
tcoffee	4.1	4.2
gadget	4.1	4.0
crisflash	4.0	4.0
PopLDdecay	3.8	3.8
cellcoal	3.8	3.6
bpp	3.8	3.6
ms	3.7	3.7
mafft	3.3	3.1
athena	2.9	2.8
covid-sim-0.13.0	2.5	2.4
indelible	1.4	1.0

Tools with highly similar functionality

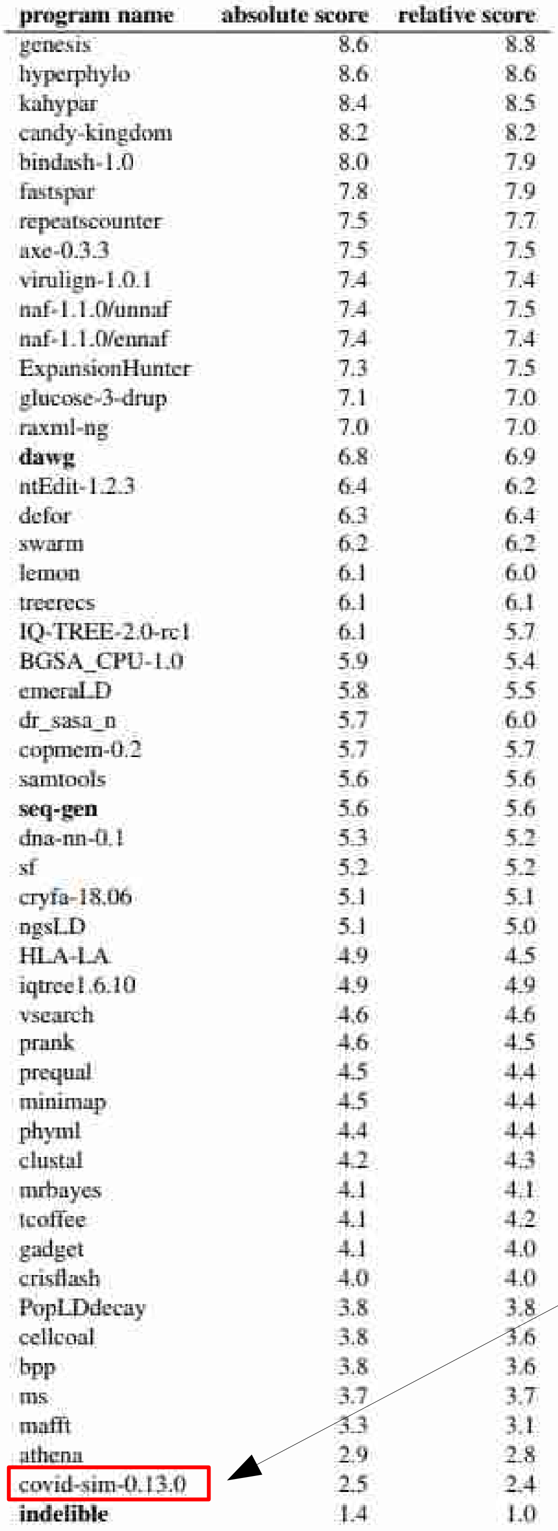
SoftWipe Benchmark



program name	absolute score	relative score
genesis	8.6	8.8
hyperphylo	8.6	8.6
kahypar	8.4	8.5
candy-kingdom	8.2	8.2
bindash-1.0	8.0	7.9
fastspar	7.8	7.9
repeatscounter	7.5	7.7
axe-0.3.3	7.5	7.5
virulign-1.0.1	7.4	7.4
naf-1.1.0/unnaf	7.4	7.5
naf-1.1.0/ennaf	7.4	7.4
ExpansionHunter	7.3	7.5
glucose-3-drup	7.1	7.0
raxml-ng	7.0	7.0
dawg	6.8	6.9
ntEdit-1.2.3	6.4	6.2
defor	6.3	6.4
swarm	6.2	6.2
lemon	6.1	6.0
treerecs	6.1	6.1
IQ-TREE-2.0-rc1	6.1	5.7
BGSA_CPU-1.0	5.9	5.4
emeraLD	5.8	5.5
dr_sasa_n	5.7	6.0
copmem-0.2	5.7	5.7
samtools	5.6	5.6
seq-gen	5.6	5.6
dna-nn-0.1	5.3	5.2
sf	5.2	5.2
cryfa-18.06	5.1	5.1
ngsLD	5.1	5.0
HLA-LA	4.9	4.5
iqtree1.6.10	4.9	4.9
vsearch	4.6	4.6
prank	4.6	4.5
prequal	4.5	4.4
minimap	4.5	4.4
phyml	4.4	4.4
clustal	4.2	4.3
mrbayes	4.1	4.1
tcoffee	4.1	4.2
gadget	4.1	4.0
crisflash	4.0	4.0
PopLDdecay	3.8	3.8
cellcoal	3.8	3.6
bpp	3.8	3.6
ms	3.7	3.7
mafft	3.3	3.1
athena	2.9	2.8
covid-sim-0.13.0	2.5	2.4
indelible	1.4	1.0

Tools with highly similar functionality

SoftWipe Benchmark



program name	absolute score	relative score
genesis	8.6	8.8
hyperphylo	8.6	8.6
kahypar	8.4	8.5
candy-kingdom	8.2	8.2
bindash-1.0	8.0	7.9
fastspar	7.8	7.9
repeatscounter	7.5	7.7
axe-0.3.3	7.5	7.5
virulign-1.0.1	7.4	7.4
naf-1.1.0/unnaf	7.4	7.5
naf-1.1.0/ennaf	7.4	7.4
ExpansionHunter	7.3	7.5
glucose-3-drup	7.1	7.0
raxml-ng	7.0	7.0
dawg	6.8	6.9
ntEdit-1.2.3	6.4	6.2
defor	6.3	6.4
swarm	6.2	6.2
lemon	6.1	6.0
treerecs	6.1	6.1
IQ-TREE-2.0-rc1	6.1	5.7
BGSA_CPU-1.0	5.9	5.4
emeraLD	5.8	5.5
dr_sasa_n	5.7	6.0
copmem-0.2	5.7	5.7
samtools	5.6	5.6
seq-gen	5.6	5.6
dna-nn-0.1	5.3	5.2
sf	5.2	5.2
cryfa-18.06	5.1	5.1
ngsLD	5.1	5.0
HLA-LA	4.9	4.5
iqtree1.6.10	4.9	4.9
vsearch	4.6	4.6
prank	4.6	4.5
prequal	4.5	4.4
minimap	4.5	4.4
phyml	4.4	4.4
clustal	4.2	4.3
mrbayes	4.1	4.1
tcoffee	4.1	4.2
gadget	4.1	4.0
crisflash	4.0	4.0
PopLDdecay	3.8	3.8
cellcoal	3.8	3.6
bpp	3.8	3.6
ms	3.7	3.7
mafft	3.3	3.1
athena	2.9	2.8
covid-sim-0.13.0	2.5	2.4
indelible	1.4	1.0

Coding that led to lockdown was 'totally unreliable' and a 'buggy mess', say experts

The code, written by Professor Neil Ferguson and his team at Imperial College London, was impossible to read, scientists claim

Covid simulation tool

SoftWipe in Practice

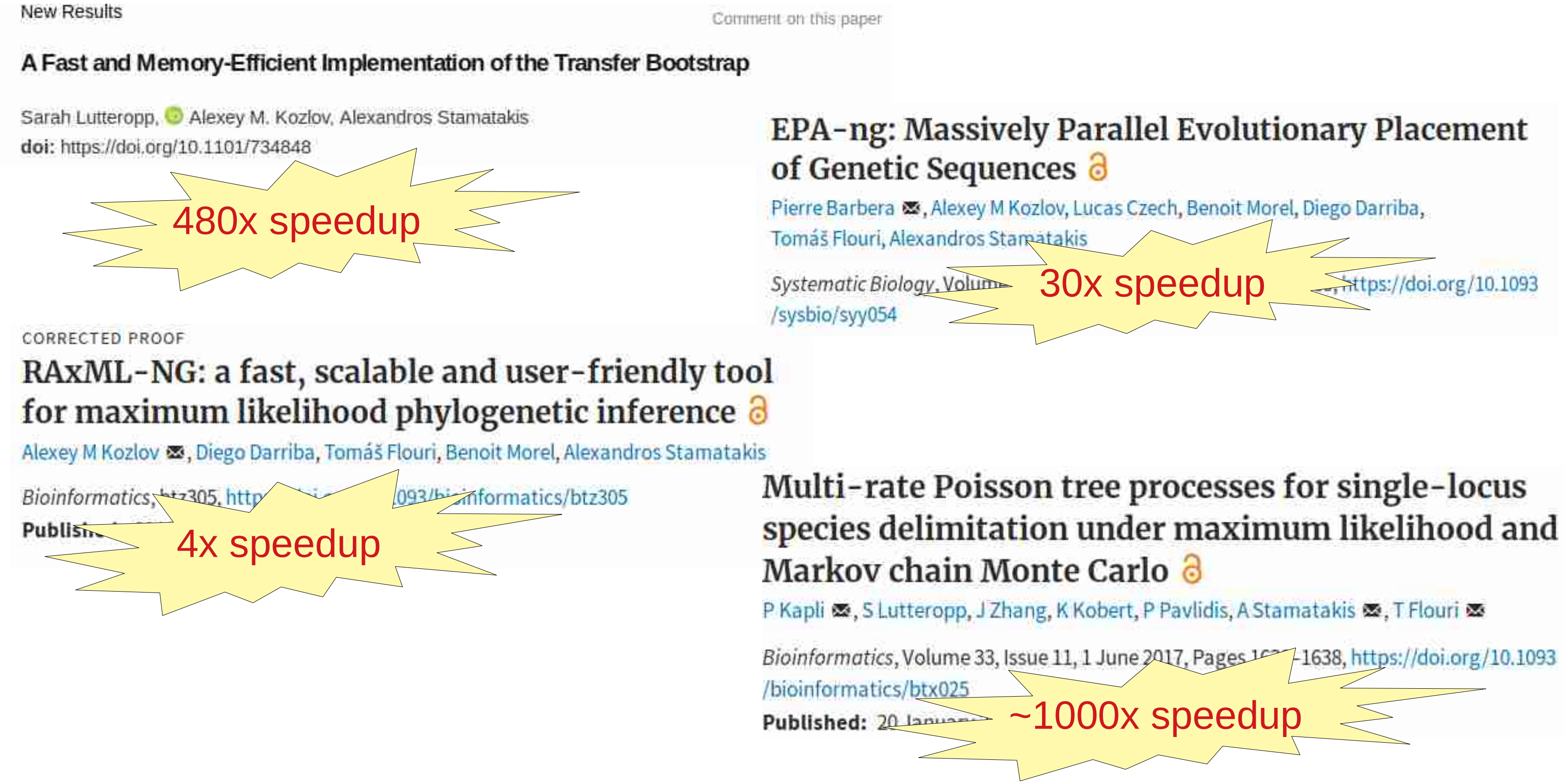
- Leads to healthy competition among lab members → everyone wants to write the cleanest code
- Used by researchers inside and outside of the lab during the development process → potential bugs identified and avoided
- Used as teaching tool in programming practicals
- SoftWipe score already used by us and others in Bioinformatics software paper submissions
- **Vision:** Establish software quality indicators as a necessary prerequisite for software paper submissions

Software Quality and Maintainability

- The Next Generation (**`-NG`**) projects:
  - Re-design, re-factoring, from scratch re-implementation of flagship tools to ensure maintainability, sustainability, and extensibility & increase scalability/performance
  - `ModelTest-NG – model testing of evolutionary models for phylogenetic inference`
  - `RAxML-NG` – phylogenetic inference
  - `EPA-NG` – phylogenetic placement of environmental reads

Energy Efficiency

Innovations



New Results

Comment on this paper

A Fast and Memory-Efficient Implementation of the Transfer Bootstrap

Sarah Lutteropp, Alexey M. Kozlov, Alexandros Stamatakis

doi: https://doi.org/10.1101/734848

480x speedup

EPA-ng: Massively Parallel Evolutionary Placement of Genetic Sequences

Pierre Barbera ✉, Alexey M Kozlov, Lucas Czech, Benoit Morel, Diego Darriba, Tomáš Flouri, Alexandros Stamatakis

Systematic Biology, Volum https://doi.org/10.1093/sysbio/syy054

30x speedup

CORRECTED PROOF

RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference

Alexey M Kozlov ✉, Diego Darriba, Tomáš Flouri, Benoit Morel, Alexandros Stamatakis

Bioinformatics, btz305, http 093/bioinformatics/btz305

Publis

4x speedup

Multi-rate Poisson tree processes for single-locus species delimitation under maximum likelihood and Markov chain Monte Carlo

P Kapli ✉, S Lutteropp, J Zhang, K Kobert, P Pavlidis, A Stamatakis ✉, T Flouri ✉

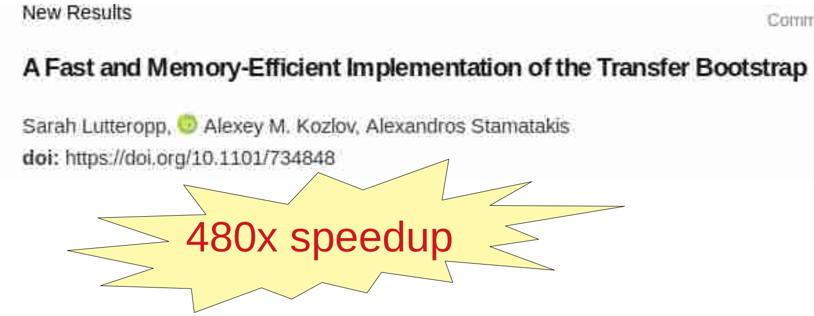
Bioinformatics, Volume 33, Issue 11, 1 June 2017, Pages -1638, https://doi.org/10.1093/bioinformatics/btx025

Published: 20

~1000x speedup

Oh, wow, this will help save a lot of energy!

The Jevons Paradox



New Results

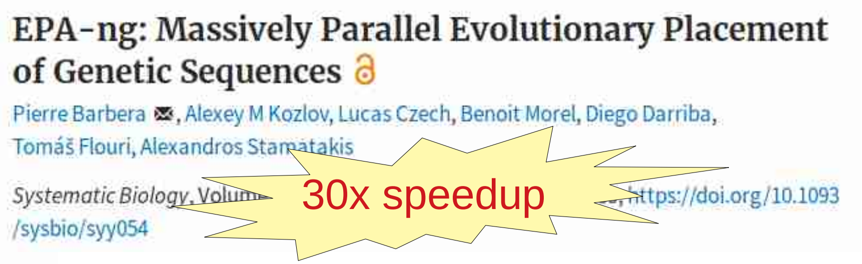
Comment on this paper

A Fast and Memory-Efficient Implementation of the Transfer Bootstrap

Sarah Lutteropp, Alexey M. Kozlov, Alexandros Stamatakis

doi: https://doi.org/10.1101/734848

480x speedup

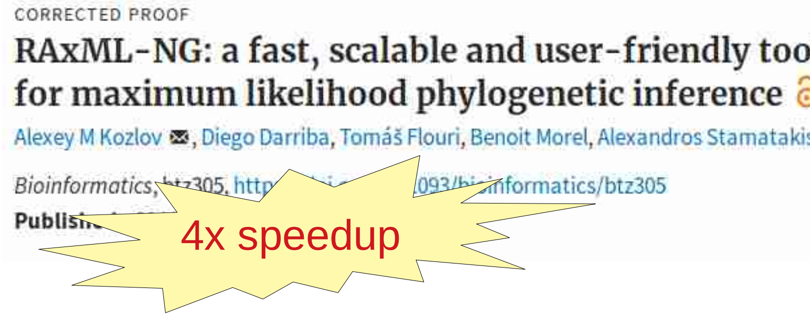


EPA-ng: Massively Parallel Evolutionary Placement of Genetic Sequences

Pierre Barbera, Alexey M Kozlov, Lucas Czech, Benoit Morel, Diego Darriba, Tomáš Flouri, Alexandros Stamatakis

Systematic Biology, Volum … https://doi.org/10.1093/sysbio/syy054

30x speedup



CORRECTED PROOF

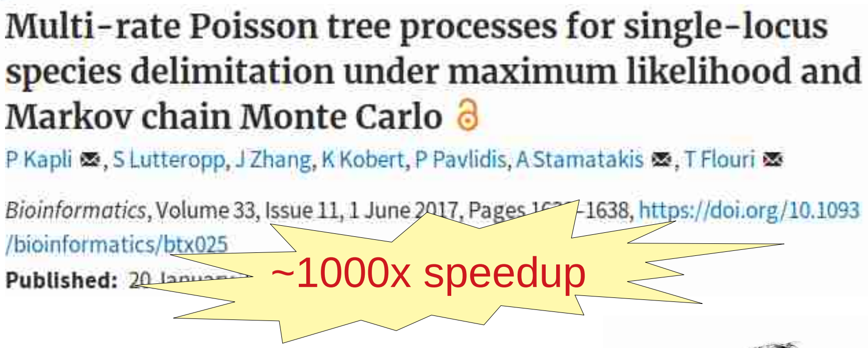
RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference

Alexey M Kozlov, Diego Darriba, Tomáš Flouri, Benoit Morel, Alexandros Stamatakis

Bioinformatics, btz305, http … 093/bioinformatics/btz305

Publis

4x speedup



Multi-rate Poisson tree processes for single-locus species delimitation under maximum likelihood and Markov chain Monte Carlo

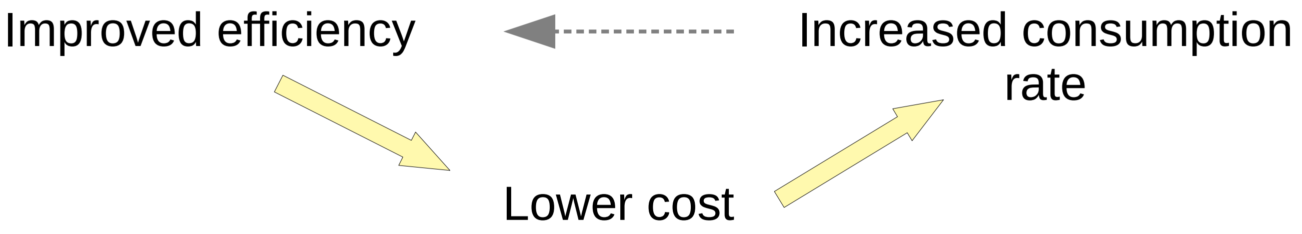
P Kapli, S Lutteropp, J Zhang, K Kobert, P Pavlidis, A Stamatakis, T Flouri

Bioinformatics, Volume 33, Issue 11, 1 June 2017, Pages 16 … -1638, https://doi.org/10.1093/bioinformatics/btx025

Published: 20 Januar

~1000x speedup

W. S. Jevons „The Coal Question“ (1865)

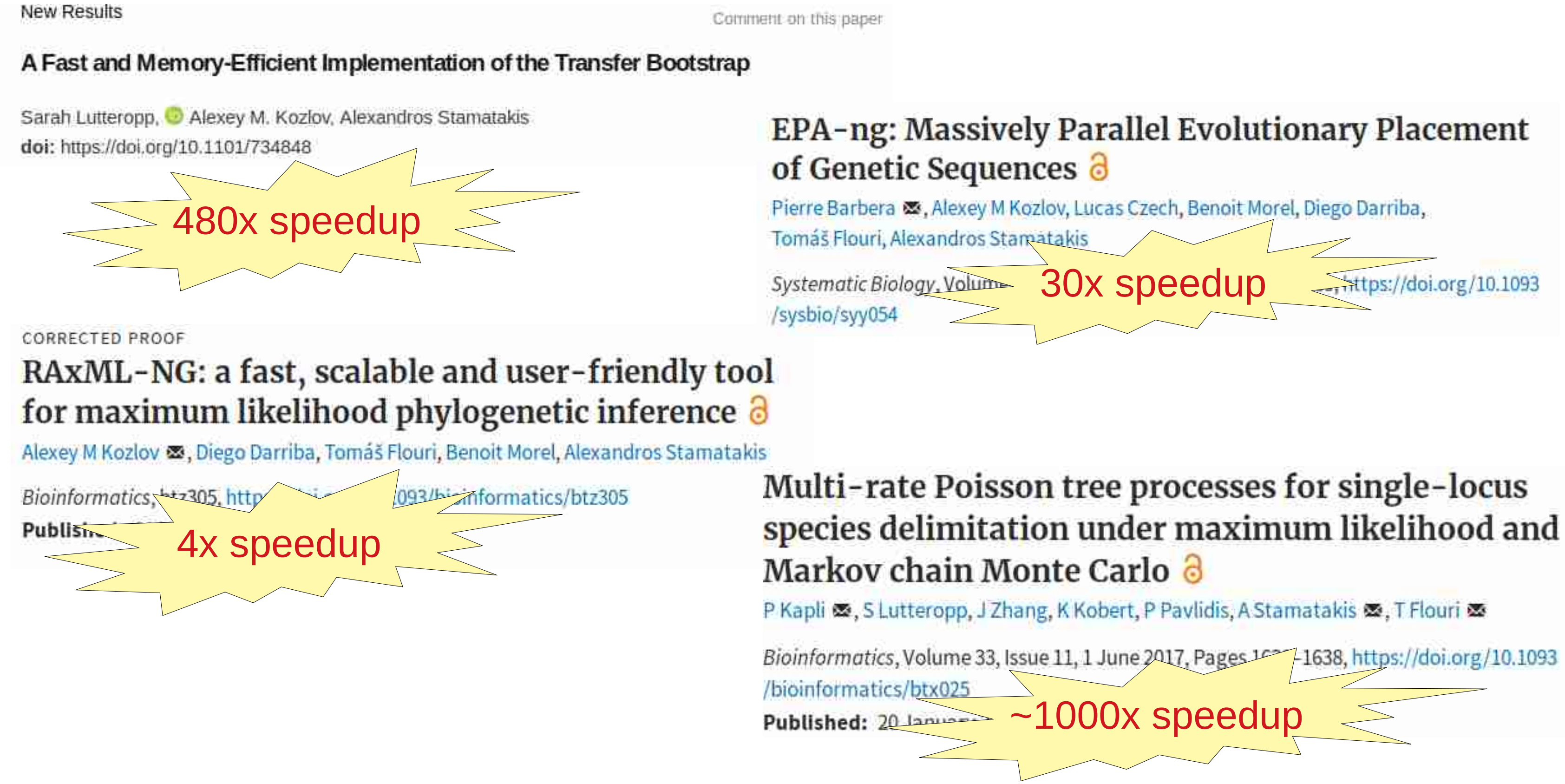


Improved efficiency

Increased consumption rate

Lower cost

The Jevons Paradox



New Results

Comment on this paper

A Fast and Memory-Efficient Implementation of the Transfer Bootstrap

Sarah Lutteropp, Alexey M. Kozlov, Alexandros Stamatakis

doi: https://doi.org/10.1101/734848

480x speedup

EPA-ng: Massively Parallel Evolutionary Placement of Genetic Sequences

Pierre Barbera, Alexey M Kozlov, Lucas Czech, Benoit Morel, Diego Darriba, Tomáš Flouri, Alexandros Stamatakis

Systematic Biology, Volum… https://doi.org/10.1093/sysbio/syy054

30x speedup

CORRECTED PROOF

RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference

Alexey M Kozlov, Diego Darriba, Tomáš Flouri, Benoit Morel, Alexandros Stamatakis

Bioinformatics, btz305, http… 093/bioinformatics/btz305

Publis…

4x speedup

Multi-rate Poisson tree processes for single-locus species delimitation under maximum likelihood and Markov chain Monte Carlo

P Kapli, S Lutteropp, J Zhang, K Kobert, P Pavlidis, A Stamatakis, T Flouri

Bioinformatics, Volume 33, Issue 11, 1 June 2017, Pages 16…–1638, https://doi.org/10.1093/bioinformatics/btx025

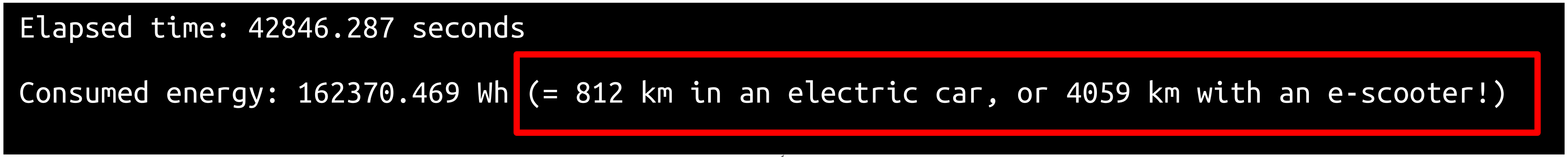
Published: 20 Januar…

~1000x speedup

We need an alternative solution!

Energy monitoring: RAxML-NG

- New in RAxML-NG v1.0: energy usage report

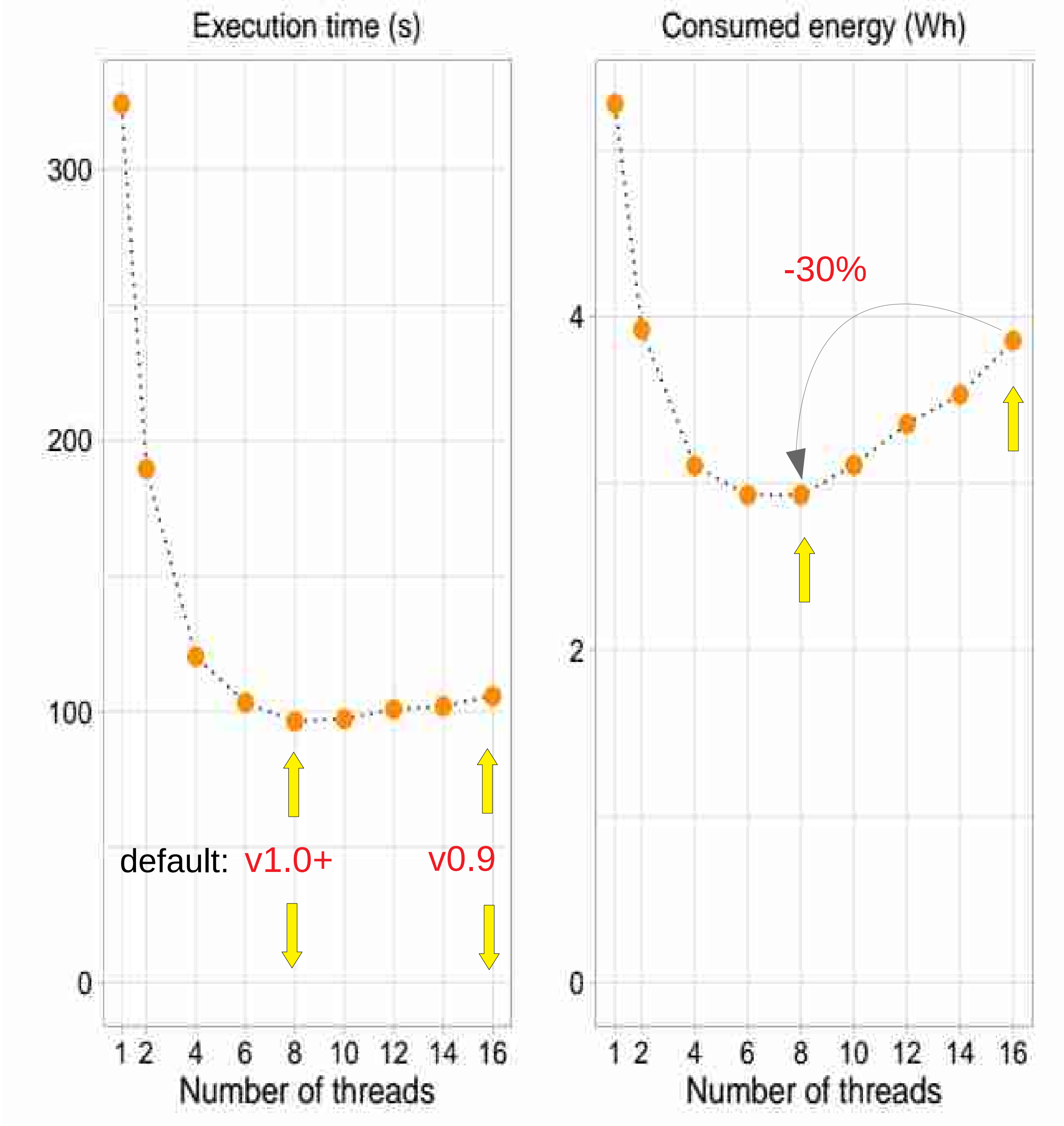


```
Elapsed time: 42846.287 seconds

Consumed energy: 162370.469 Wh (= 812 km in an electric car, or 4059 km with an e-scooter!)
```

Single tree search (96 nodes x 12h):
>160 kWh

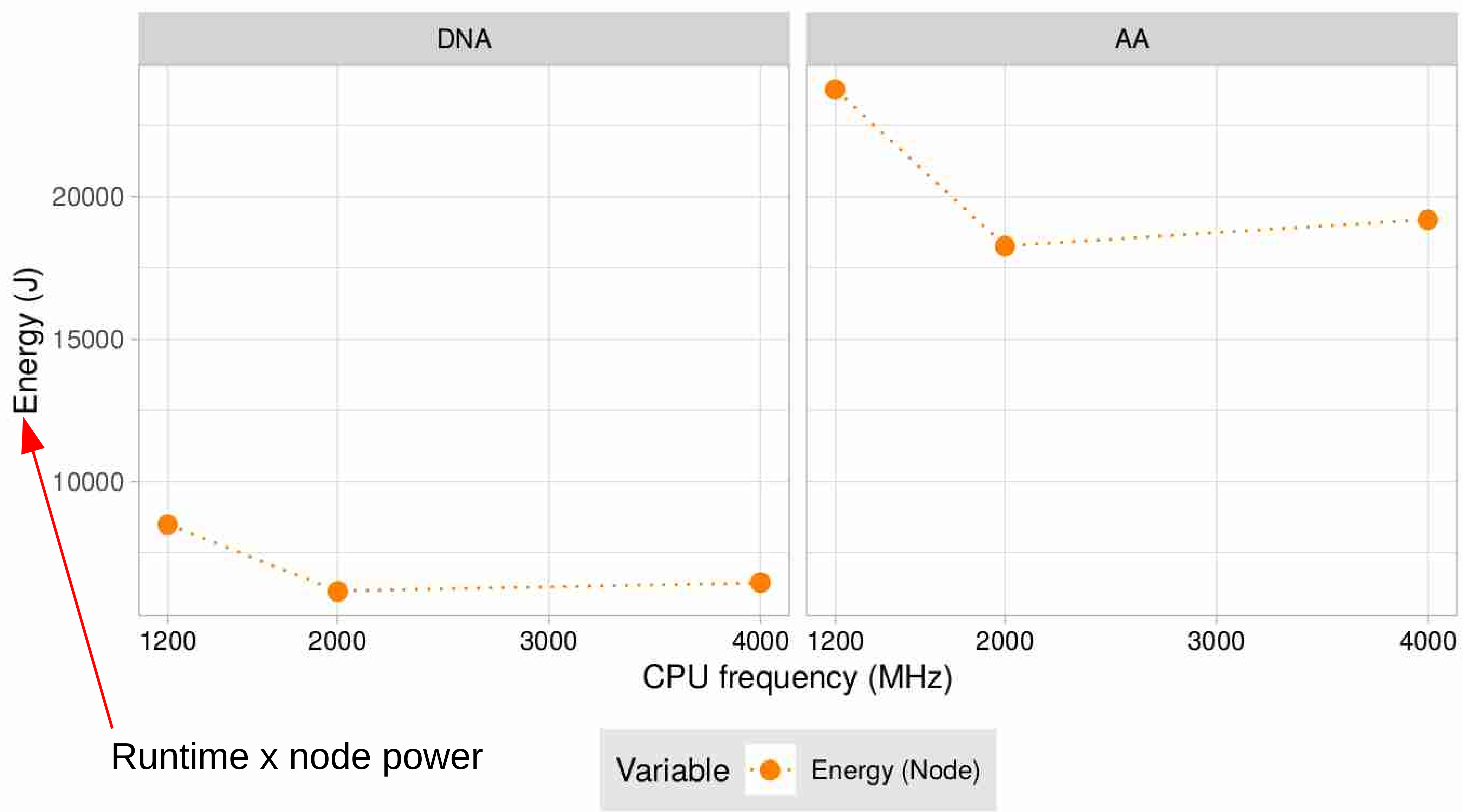
Energy Saving Mode in `RAxML-NG` v1.0



-30%

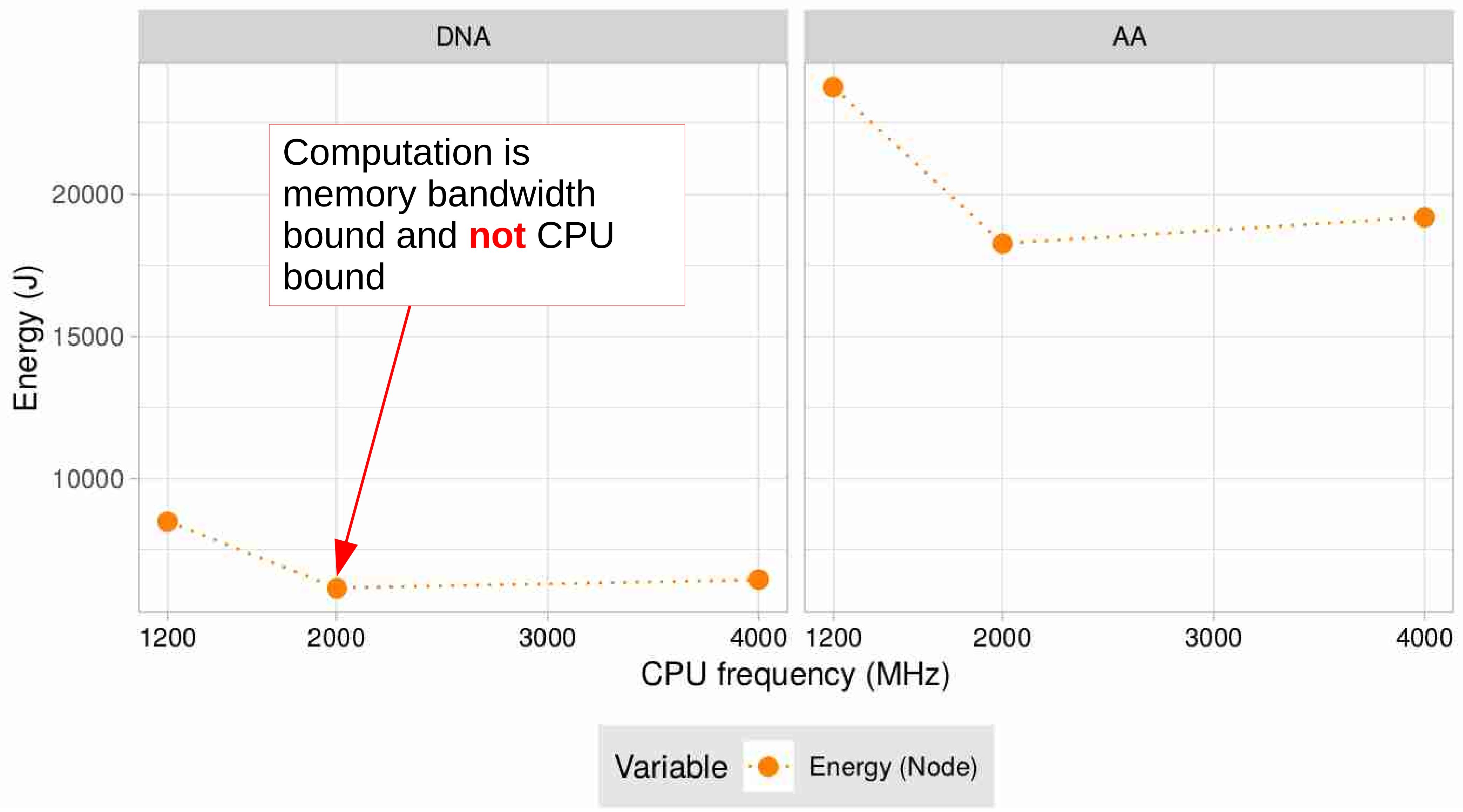
default: v1.0+ v0.9

Phylogenetic Inference: Energy as a function of CPU clock frequency



Runtime x node power

Phylogenetic Inference: Energy as a function of CPU clock frequency



Computation is memory bandwidth bound and **not** CPU bound