Bioinformatics Application Note:

CONTEXT – A Phylogenomics Dataset Browser

Joe Parker[1,2] and Stephen J. Rossiter[2]

1. Kitson Consulting, Bristol, UK; Present address: Jodrell Laboratory, Royal Botanic Gardens, Kew, UK

2. School of Biological and Chemical Sciences, Queen Mary University of London, UK.

Word count: 1,035

Corresponding Author:
Joe Parker
Jodrell Laboratory,
Royal Botanic Gardens, Kew,
TW9 3DS, UK
Tel. +44 20-8332-5063
Fax +44 20-8332-5197
joe.parker@kew.org

Project email: joe+CONTEXT@kitson-consulting.co.uk

Abstract

Summary. Quality control (QC) in large phylogenomic datasets is a key requirement for reliable and reproducible research in evolution, adaptation, speciation and taxonomy. CONTEXT is a browser for high-throughput visualisation and comparative QC of phylogenomic datasets, consisting of a Java API and an executable binary jarfile with graphical user interface (GUI). The tool allows users to rapidly and easily visualise thousands of multiple sequence alignments and hundreds of phylogenies using a GUI to identify outliers which could affect downstream analyses. CONTEXT calculates a variety of downstream statistics on alignments and phylogenies including entropy, informativeness, imbalance, signal:noise and size.

Motivation. Comparative genomics studies have become increasingly common, but these analyses are sensitive to the quality and heterogeneity of input datasets (multiple sequence analyses and phylogenies). Currently few tools exist to readily compute descriptive statistics, or to visualise large numbers of input datasets. CONTEXT is a phylogenomics dataset browser which facilitates these analyses in a lightweight application. It allows any user to rapidly visualise, inspect, score, and sort input datasets to identify outlying datasets which may need additional processing, filtering, or masking from further analyses.

Results. The application has been successfully implemented on a variety of infrastructures. A variety of common input data formats including FASTA, Phylip/PAML, Nexus, and Newick conventions are automatically read and parsed.

Availability and implementation. The API is implemented in native Java code, available online at https://github.com/lonelyjoeparker/qmul-genome-convergence-pipeline. The executable binary can be downloaded at https://github.com/lonelyjoeparker/qmul-genome-convergence-pipeline/tree/master/trunk/bin. The project page is at https://github.com/lonelyjoeparker/qmul-genome-convergence-pipeline/blob/master/CONTEXT.md

Contact. joe.parker@kew.org

Introduction

In the last decade, ‘phylogenomic’ analyses of sequence data (typically, phylogenetic analyses comprising genome-scale data, e.g. thousands of coding sequences) have become increasingly common. They have provided a powerful means to highlight evolutionary insights in animals (Tsagkogeorga *et al.*, 2013; McCormack *et al.*, 2013), plants (Zhao *et al.*, 2016), fungi (Dentinger *et al.*, 2016) and prokaryotes (Jungblut *et al.,* 2016). As well as pure phylogenetic inference, the greatly expanded data these analyses drive permit detection of subtle but key genomic effects such as adaptive molecular convergence (Parker *et al.*, 2013) and incomplete lineage sorting (Scornavacca & Galtier, 2016), among others. The trend towards using genomic big data to answer evolutionary questions will surely only accelerate.

However, the multiplicity of sequencing methods, annotation and orthology protocols, multiple sequence aligners, and phylogeny inference tools available mean that comparing the inputs to downstream analyses (typically, multiple sequence alignments and phylogenies) is essential. In particular, between-workflow differences can lead to widely differing results, even for identical data (Blackburne & Whelan, 2013). For this reason, robust quality control procedures should be applied to dataset curation and phylogeny inference – with repeatability an essential criterion. Unfortunately, many labs' workflows have yet to meet this demand; typically only limited comparative QC analysis is performed on sets of alignments and phylogenies. In part, this is because, although certain descriptive statistics are available, they are not routinely calculated by upstream (genomics) packages such as assemblers or annotation tools. In many cases, inspection by eye remains the primary means of error detction – and even then, often only used *post hoc* to troubleshoot outliers in downstream analyses. This situation is clearly untenable in the phylogenomic era.

CONTEXT (for 'COmparative Nucleotides and Trees EXploration Tool') is designed to improve this situation. To improve the objectivity by which alignments and phylogenies are selected for further analyses, various helpful statistics are rapidly calculated to aid dataset QC. Their summaries are also calculated at the dataset level, and simple visual plots can be used to identify outliers. Meanwhile, the graphical user interface provides a fast method by which thousands of alignments and phylogenies can be loaded and inspected by eye if the user wishes.

Features and implementation

The API elements contain resources for phylogenomics such as input/output and parsing utilities; trimming, pruning and validation methods for alignments and phylogenies; statistics for evaluating alignments, phylogenies, likelihood fits and dN/dS values; UI elements including two main GUI platforms; post-processing including linear regression and descriptive parametric statistics on large distributions of small floating-point numbers. Implemented alignment statistics include (for nucleotide and amino-acid sequences): number of taxa; length (excluding and including gaps and invariant sites); longest sequence without gaps; mean sitewise entropy (after Shannon (1948)) and longest run of non-zero entropy. Phylogeny statistics (after Felsenstein, 2004) include number of tips; tree and root length and height; imbalance (Colless, 1982); cherry-count (Steel & McKenzie, 2001); signal:noise (aka internal branch lengths:total tree length Phillips & Penny, 2001) and external:internal branch length ratio.

Evaluation

Example usage statistics shown in Table 1.

	OS	**Arch**	**CPU type, clock GHz**	**cores**	**RAM Gb**	**HDD Gb**
pandanus	Ubuntu 16.04 LTS / Biolinux	i686	Xeon E5620 @ 2.4	4	33	1000 @ ATA 7200rpm

2	Ubuntu 16.04 LTS / MATE	ARM	ARMv7 @ 0.9	1	1	8 @ SD
toshiba	Windows 7	x64	Core i7 @ 2.4	4	3	128 @ SSD
Tower	Ubuntu 16.04.02 LTS	X64	8x Intel(R) Xeon(R) CPU E3-1245 v5 @ 3.50GHz	8	65	1.0Tb SSD, 4.0Tb HDD
MBP	Mac OSX 10.9.5	x64	Core i7 @ 2.2	4	8 (1333MHz DDR3)	250 @ SSD
EC2 m4.10xlarge	Ubuntu 16.04 LTS	x64	Xeon E5-2670 @ 2.5	16	122	320 @ SSD
EC2	Ubuntu !6.04 LTS	x64	Xeon E5-2680 @ 2.8	32	60	2x320 @ SSD

Table 1

Roadmap and versioning

CONTEXT is currently supplied at **Version 0.8.4 prerelease.**

Acknowledgements

This work has been funded by BBSRC at QMUL, specifically the methodological innovations for convergence detection methods correctly controlling for false positives (essential in genomic datasets) and a core API to implement these and facilitate handling genomic sequence data, carried out principally by Dr. Parker (with input from Prof. Rossiter (PI), Drs. James Cotton & Elia Stupka (Co-I) and Dr. Tsagkogeorga (PDRA)) under BBSRC # BB/H017178/1.

Figures / data / tables

Table 1: Example system resource usage. The RAM usage (in megabytes) and average load time (initialisation to completion) of the CONTEXT under a variety of test computer architectures and input datasets. Benchmarking scripts and analyses are available online at https://github.com/lonelyjoeparker/qmul-genome-convergence-pipeline/benchmarking

Test case	Mac OSX 10.9.5, 2.2GHz core i7, 8Gb 1333MHz DDR3 RAM, 250 Gb SSD.	Tower Ubuntu 16.04.02, 8x Xeon E3-1245v5 3.5GHz, 65Gb RAM, 1.0Tb SSD/4.0Tb HDD	Windows 7, 2.0GHz core i7 CPU, 8Gb RAM, 500Gb SSD
692 Nucleotide alignments, 7 taxa, 78-2766nt	*6.8*	*5.0*	
2,326 Nucleotide alignments, 14-22 taxa, 450-11136nt	*21.7*	*19.0*	
3,656 Nucleotide alignments, 10 taxa, 99-6366nt	*25.1*	*23.0*	

Table 1

Table 2: Load time benchmarking. The RAM usage (in megabytes) and median load time (initialisation to completion) of the CONTEXT for scaled task sizes, on two representative systems; a multipurpose Linux bioinformatics workstation in general use mode (other applications active simultaneously) and an Apple laptop running in dedicated mode (no other user tasks active). Values shown wall-clock time in seconds to initialise CONTEXT, load the specified dataset and calculate statistics, and exit (median values of five measurements).

System	Tower Ubuntu 16.04.02, 8x Xeon E3-1245v5 3.5GHz, 65Gb RAM, 1.0Tb SSD/4.0Tb HDD			Mac OSX 10.9.5, 2.2GHz core i7, 8Gb 1333 DDR3 RAM, 250 Gb SSD.		
Taxa à **Alignments** \/	**4**	**22**	**200**	**4**	**22**	**20**
10	2.18	1.62		3.39	3.30	
50	2.05	2.28		3.71	3.74	
100	3.46	3.14		3.94	4.16	
500	6.67	10.84		5.92	7.09	
1000	10.30	16.92		8.92	10.99	
2000	15.95			16.72		
3000	21.61			24.23		

Table 2

Figure 1: CONTEXT schematic. The schematic logic flow of the phylogenomic dataset browser is shown with descriptions of key analysis steps, in flow diagram format.

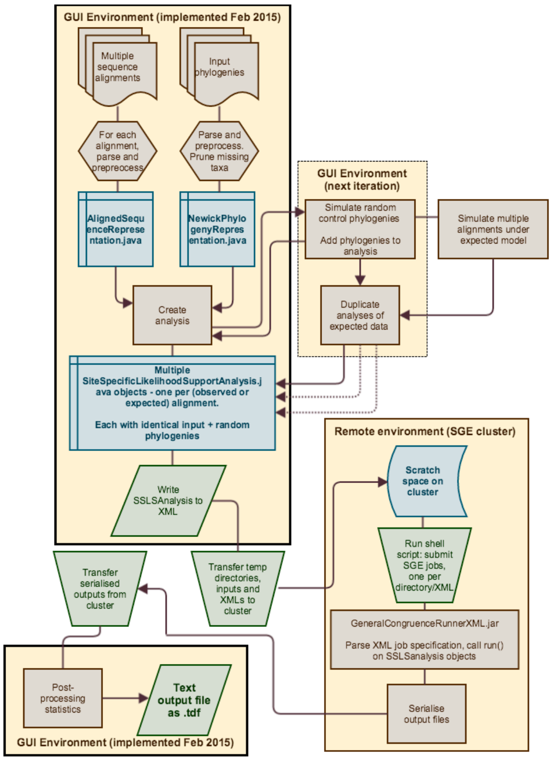
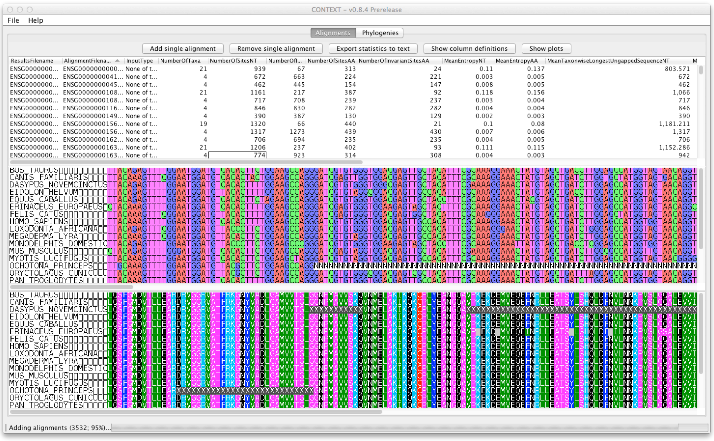

GUI Environment (implemented Feb 2015)
Multiple sequence alignments
Input phylogenies
For each alignment, parse and preprocess
Parse and preprocess. Prune missing taxa
AlignedSequenceRepresentation.java
NewickPhylogenyRepresentation.java
Create analysis
Multiple SiteSpecificLikelihoodSupportAnalysis.java objects - one per (observed or expected) alignment.
Each with identical input + random phylogenies
Write SSLSAnalysis to XML
GUI Environment (next iteration)
Simulate random control phylogenies
Add phylogenies to analysis
Simulate multiple alignments under expected model
Duplicate analyses of expected data
Remote environment (SGE cluster)
Scratch space on cluster
Run shell script: submit SGE jobs, one per directory/XML
GeneralCongruenceRunnerXML.jar
Parse XML job specification, call run() on SSLSanalysis objects
Serialise output files
Transfer temp directories, inputs and XMLs to cluster
Transfer serialised outputs from cluster
Post-processing statistics
Text output file as .tdf
GUI Environment (implemented Feb 2015)

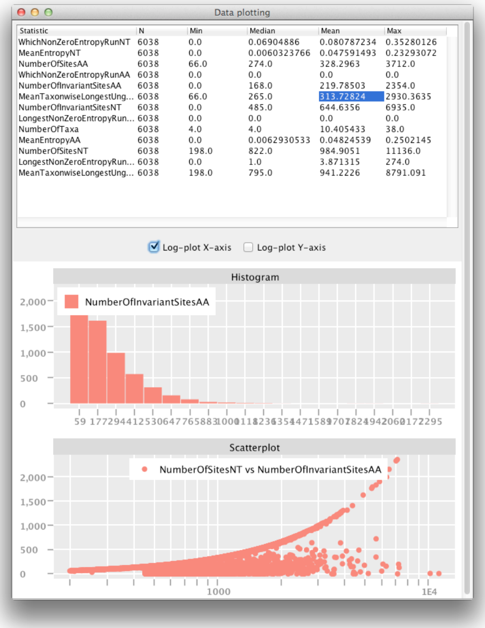
Figure 2: Screenshots showing visualisation of example datasets: (a) The alignment browser, showing 692 multiple sequence alignments together with statistics; (b) The phylogeny browser, showing phylogenies and statistics; (c) The plotting display, showing summary statistics, histogram and bivariate data (logged, in this case)



ResultsFilename	AlignmentFilena...	InputType	NumberOfTaxa	NumberOfSitesNT	NumberOfI...	NumberOfSitesAA	NumberOfInvariantSitesAA	MeanEntropyNT	MeanEntropyAA	MeanTaxonwiseLongestUngappedSequenceNT	M
ENSG0000000...	ENSG0000000000...	None of t...	21	939	67	313	24	0.11	0.137	803.571	
ENSG0000000...	ENSG0000000041...	None of t...	4	672	663	224	221	0.003	0.005	672	
ENSG0000000...	ENSG0000000045...	None of t...	4	462	445	154	147	0.008	0.005	462	
ENSG0000000...	ENSG0000000108...	None of t...	21	1161	217	387	92	0.118	0.156	1,066	
ENSG0000000...	ENSG0000000108...	None of t...	4	717	708	239	237	0.003	0.004	717	
ENSG0000000...	ENSG0000000116...	None of t...	4	846	830	282	282	0.004	0.004	846	
ENSG0000000...	ENSG0000000149...	None of t...	4	390	387	130	129	0.002	0.003	390	
ENSG0000000...	ENSG0000000156...	None of t...	19	1320	66	440	21	0.1	0.08	1,181.211	
ENSG0000000...	ENSG0000000156...	None of t...	4	1317	1273	439	430	0.007	0.006	1,317	
ENSG0000000...	ENSG0000000162...	None of t...	4	706	694	235	235	0.004	0.005	706	
ENSG0000000...	ENSG0000000163...	None of t...	21	1206	237	402	93	0.111	0.115	1,152.286	
ENSG0000000...	ENSG0000000163...	None of t...	4	774	923	314	308	0.004	0.003	942	

Adding alignments (3532; 95%)...

a)

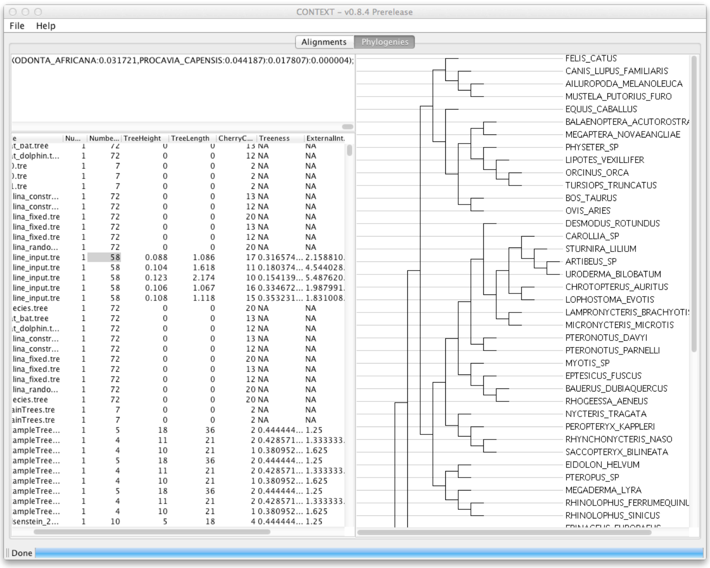


Data plotting

Statistic	N	Min	Median	Mean	Max
WhichNonZeroEntropyRunNT	6038	0.0	0.06904886	0.080787234	0.35280126
MeanEntropyNT	6038	0.0	0.0060323766	0.047591493	0.23293072
NumberOfSitesAA	6038	66.0	274.0	328.2963	3712.0
WhichNonZeroEntropyRunAA	6038	0.0	0.0	0.0	0.0
NumberOfInvariantSitesAA	6038	0.0	168.0	219.78503	2354.0
MeanTaxonwiseLongestUng...	6038	66.0	265.0	313.72824	2930.3635
NumberOfInvariantSitesNT	6038	0.0	485.0	644.6356	6935.0
LongestNonZeroEntropyRun...	6038	0.0	0.0	0.0	0.0
NumberOfTaxa	6038	4.0	4.0	10.405433	38.0
MeanEntropyAA	6038	0.0	0.0062930533	0.04824539	0.2502145
NumberOfSitesNT	6038	198.0	822.0	984.9051	11136.0
LongestNonZeroEntropyRun...	6038	0.0	1.0	3.871315	274.0
MeanTaxonwiseLongestUng...	6038	198.0	795.0	941.2226	8791.091

☑ Log-plot X-axis ☐ Log-plot Y-axis

b)



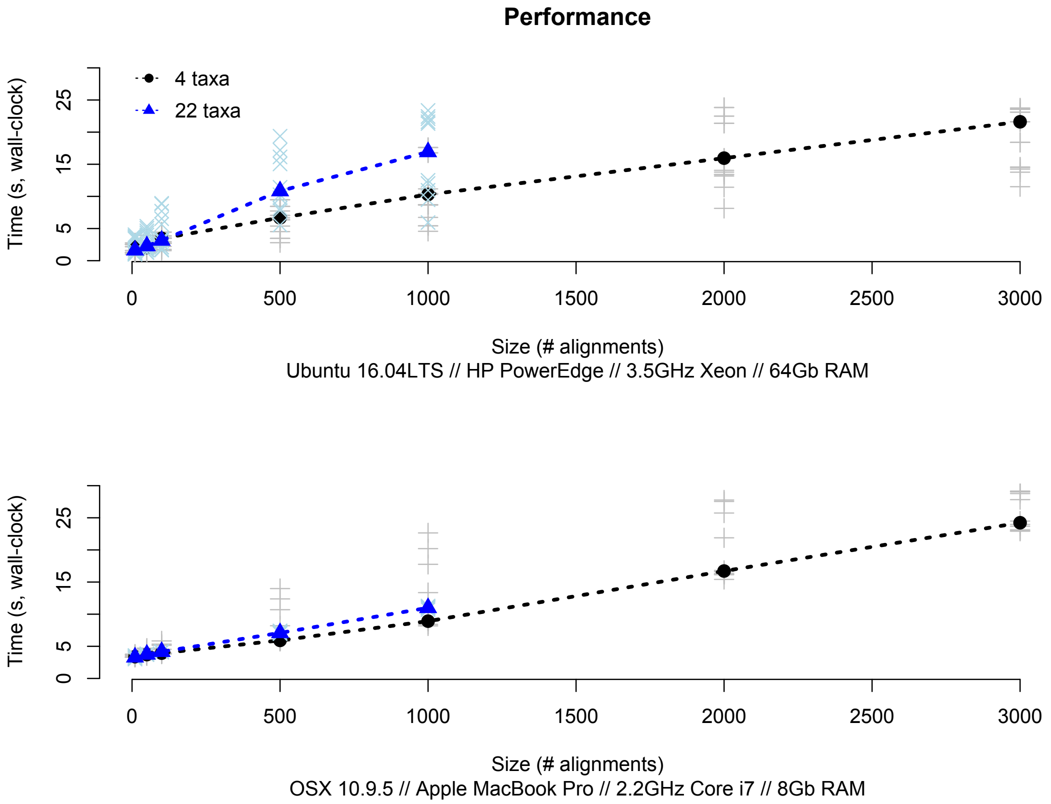
CONTEXT – v0.8.4 Prerelease

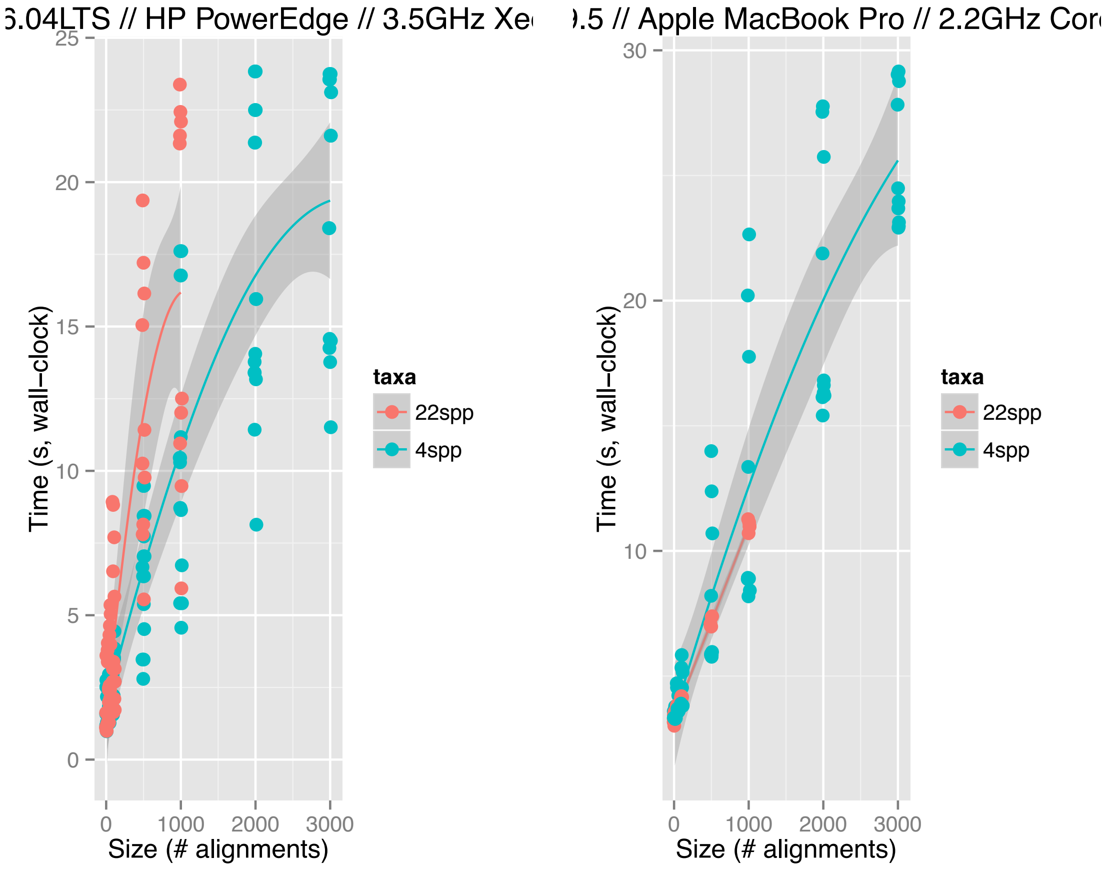
(ODONTA_AFRICANA:0.031721,PROCAVIA_CAPENSIS:0.044187):0.017807):0.000004);

e	Nu...	Numbe...	TreeHeight	TreeLength	CherryC...	Treeness	ExternalInt.
t_bat.tree	1	72	0	0	13	NA	NA
t_dolphin.t...	1	72	0	0	12	NA	NA
).tre	1	7	0	0	2	NA	NA
).tre	1	7	0	0	2	NA	NA
l.tre	1	7	0	0	2	NA	NA
lina_constr...	1	72	0	0	13	NA	NA
lina_constr...	1	72	0	0	12	NA	NA
lina_fixed.tre	1	72	0	0	20	NA	NA
lina_fixed.tre	1	72	0	0	13	NA	NA
lina_fixed.tre	1	72	0	0	12	NA	NA
lina_rando...	1	72	0	0	20	NA	NA
line_input.tre	1	58	0.088	1.086	17	0.316574...	2.158810.
line_input.tre	1	58	0.104	1.618	11	0.180374...	4.544028.
line_input.tre	1	58	0.123	2.174	10	0.154139...	5.487620.
line_input.tre	1	58	0.106	1.067	16	0.334672...	1.987991.
line_input.tre	1	58	0.108	1.118	15	0.353231...	1.831008.
ecies.tree	1	72	0	0	20	NA	NA
t_bat.tree	1	72	0	0	13	NA	NA
t_dolphin.t...	1	72	0	0	12	NA	NA
lina_constr...	1	72	0	0	13	NA	NA
lina_constr...	1	72	0	0	12	NA	NA
lina_fixed.tre	1	72	0	0	20	NA	NA
lina_fixed.tre	1	72	0	0	13	NA	NA
lina_fixed.tre	1	72	0	0	12	NA	NA
lina_rando...	1	72	0	0	20	NA	NA
ecies.tree	1	72	0	0	20	NA	NA
ainTrees.tre	1	7	0	0	2	NA	NA
ainTrees.tre	1	7	0	0	2	NA	NA
ampleTree...	1	5	18	36	2	0.444444...	1.25
ampleTree...	1	4	11	21	2	0.428571...	1.333333.
ampleTree...	1	4	10	21	1	0.380952...	1.625
ampleTree...	1	5	18	36	2	0.444444...	1.25
ampleTree...	1	4	11	21	2	0.428571...	1.333333.
ampleTree...	1	4	10	21	1	0.380952...	1.625
ampleTree...	1	5	18	36	2	0.444444...	1.25
ampleTree...	1	4	11	21	2	0.428571...	1.333333.
ampleTree...	1	4	10	21	1	0.380952...	1.625
senstein_2...	1	10	5	18	4	0.444444...	1.25

c)

Figure 3: Performance of CONTEXT in preload task, 50-3000 alignments: (a) Load times of the Linux workstation in mixed-use mode. 4-taxa alignment tasks shown in black dots, 22-taxa load times as blue triangles. Measurements shown in light points with medians (n=5) shown as solid-point; (b) Load times of the Apple laptop in dedicated mode. Labels as above.


Performance
4 taxa
22 taxa
Time (s, wall-clock)
0
5
15
25
0
500
1000
1500
2000
2500
3000
Size (# alignments)
Ubuntu 16.04LTS // HP PowerEdge // 3.5GHz Xeon // 64Gb RAM
Time (s, wall-clock)
0
5
15
25
0
500
1000
1500
2000
2500
3000
Size (# alignments)
OSX 10.9.5 // Apple MacBook Pro // 2.2GHz Core i7 // 8Gb RAM


6.04LTS // HP PowerEdge // 3.5GHz Xe
Time (s, wall–clock)
0
5
10
15
20
25
0
1000
2000
3000
Size (# alignments)
taxa
22spp
4spp
).5 // Apple MacBook Pro // 2.2GHz Cor
Time (s, wall–clock)
10
20
30
0
1000
2000
3000
Size (# alignments)
taxa
22spp
4spp

References

Blackburne, B.P. & Whelan, S. (2013) Class of multiple sequence alignment algorithm affects genomic analysis. *Mol. Biol. Evol.* 30(**3**):642-53.

Dentinger, B.T.M., Gaya, E., O'Brien, H., Suz., L.M., Lachlan, R., Díaz-Valderrama, J.R., Koch, R.A. & Aime, C. (2016) Tales from the crypt: genome mining from fungarium specimens improves resolution of the mushroom tree of life. *Bot. J. Linn. Soc* **117**(1):11–32. DOI: 10.1111/bij.12553

Zhao, L., Li, X. , Zhang, N., Zhang S-D., Yi, T-S., Ma, H., Guo, Z-H. & Li, D-Z. (2016) Phylogenomic analyses of large-scale nuclear genes provide new insights into the evolutionary relationships within the rosids. *Mol. Phylogenet. Evol.* **105**:166–176.

Parker, J.*, Tsagkogeorga, G.*, Cotton, J.A., Liu, Y., Provero, P., Stupka, E. & Rossiter, S.J. (2013) Genome-wide signatures of convergent evolution in echolocating mammals. *Nature* **502**(7470)**:**2280231. doi:10.1038/nature12511. *These authors contributed equally to this article.

Tsagkogeorga, G.*, Parker, J.*, Stupka, E., Cotton, J.A., & Rossiter, S.J. (2013) Phylogenomic analyses elucidate evolutionary relationships of the bats (Chiroptera) *Curr. Biol.* **23**(22):2262-2267.

Scornavacca, C. & Galtier, N. (2016) Incomplete Lineage Sorting in Mammalian Phylogenomics. *Syst. Biol.* Xxxxx.

McCormack JE, Harvey MG, Faircloth BC, Crawford NG, Glenn TC, Brumfield RT. (2013) A phylogeny of birds based on over 1,500 loci collected by target enrichment and high-throughput sequencing. *PLoS One.* **8**(1):e54848. doi: 10.1371/journal.pone.0054848.

Anne D Jungblut, Ian Hawes, Tyler J Mackey, Megan Krusor, Peter T Doran, Dawn Y Sumner, Jonathan A Eisen, Colin Hillman, Alexander K Goroncy (2016) Microbial mat communities along an oxygen gradient in a perennially ice-covered Antarctic lake *Appl. Env. Microbiol.* **82**(2):620-630

Shannon (1948)

Felsenstein (2004)

Colless, (1982)

Steel & McKenzie (2001)

Phillips & Penny (2001)