Dating events within gene trees: speciations, duplications, losses

PhD defense

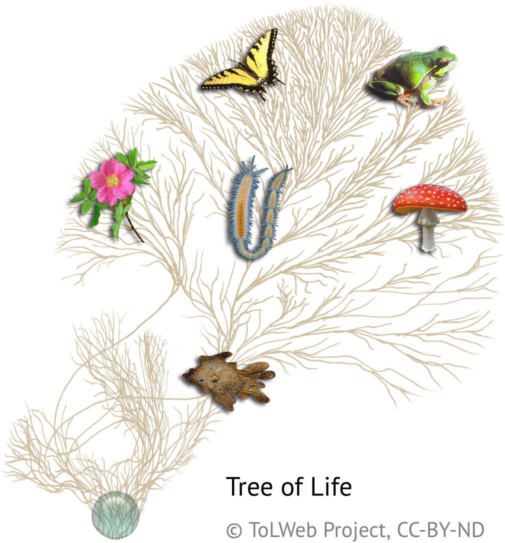
Guillaume Louvel

Supervision: Dr. Hugues Roest Crollius

September 7, 2020

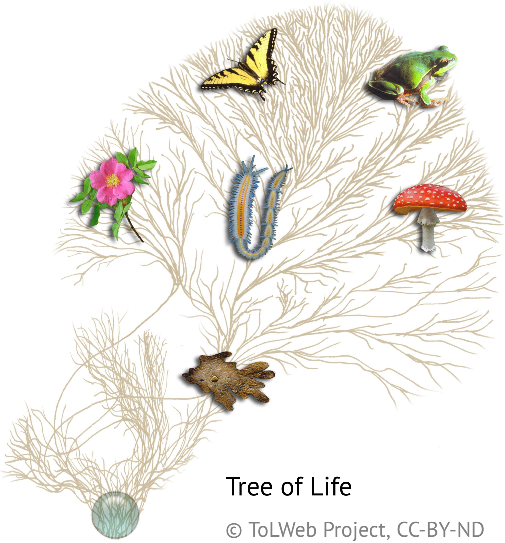
Institut de Biologie de l'École normale supérieure, Paris

Introduction



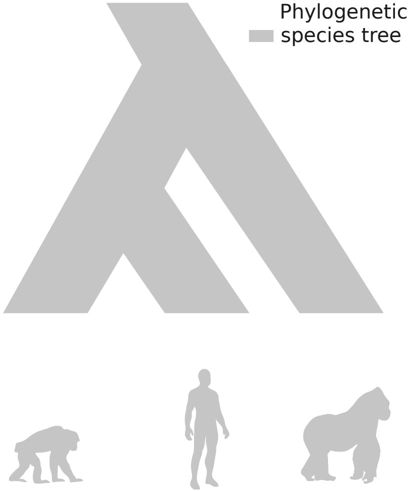
Tree of Life

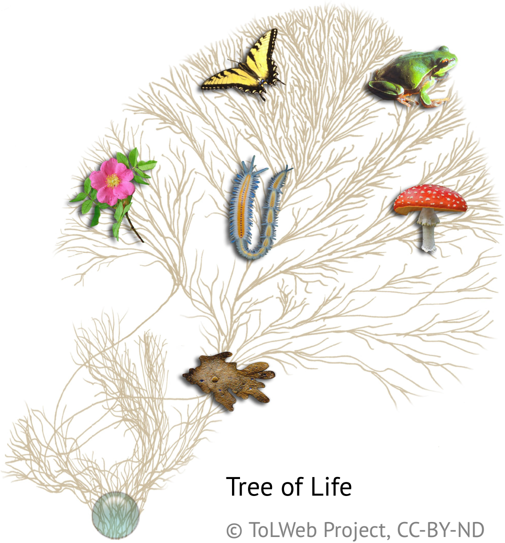
© ToLWeb Project, CC-BY-ND



Tree of Life

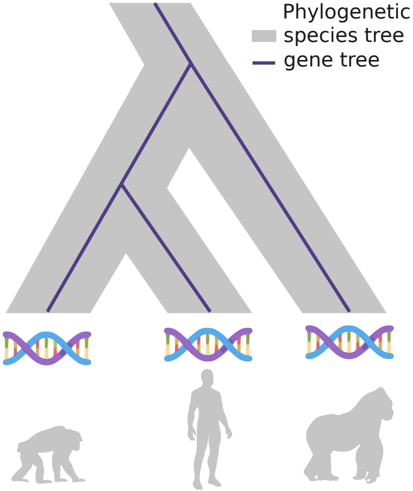
© ToLWeb Project, CC-BY-ND

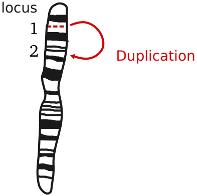




Tree of Life

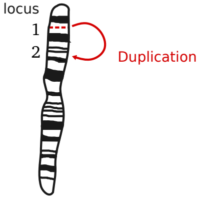
© ToLWeb Project, CC-BY-ND



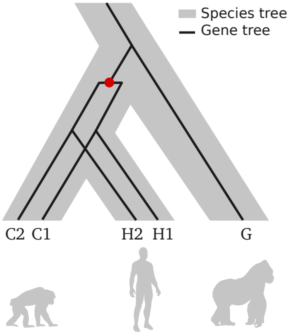


Ancestral chromosome

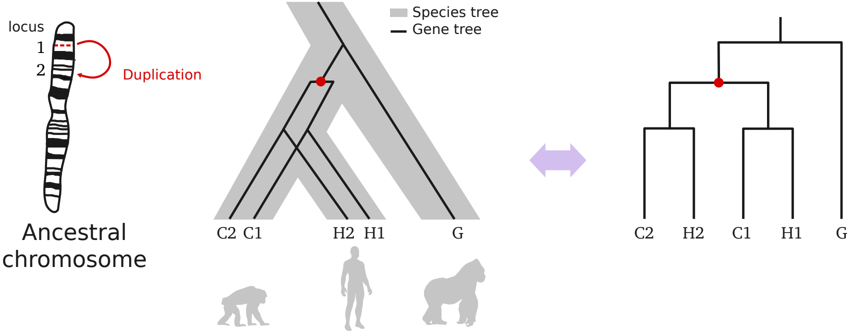
Gene tree ≠ species tree



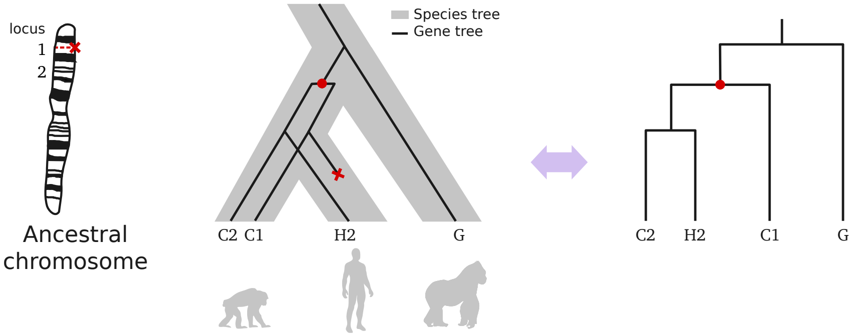
Ancestral chromosome



Gene tree ≠ species tree



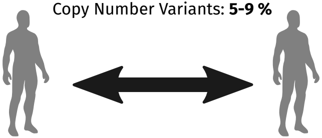
Gene tree ≠ species tree



Topics

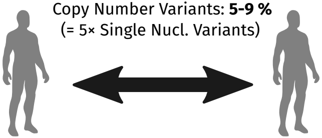
I. The accuracy of dating gene tree events
II. Finding gene trees with correlated losses

Gene duplications: a continuous mutation process



Copy Number Variants: **5-9 %**

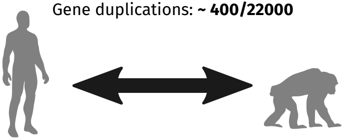
Gene duplications: a continuous mutation process



Copy Number Variants: **5-9 %**
(= 5× Single Nucl. Variants)

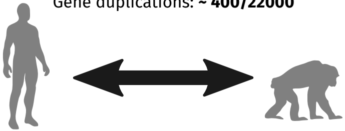
Gene duplications: a continuous mutation process

Gene duplications: **~ 400/22000**



Gene duplications: a continuous mutation process

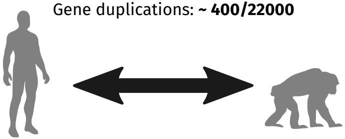
Gene duplications: **~ 400/22000**



Fate of duplicate genes

- loss, pseudogeneisation
- selection for increased dosage
- neofunctionalisation
- subfunctionalisation

Gene duplications: a continuous mutation process



Gene duplications: ~ **400/22000**

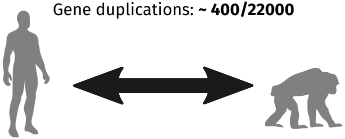
Fate of duplicate genes

- loss, pseudogeneisation
- selection for increased dosage
- neofunctionalisation
- subfunctionalisation

- Material for adaptation
- Source of genetic reproductive incompatibilities

Gene duplications: a continuous mutation process

Gene duplications: ~ **400/22000**

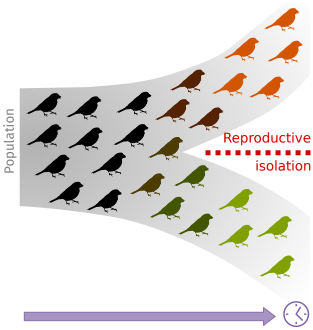


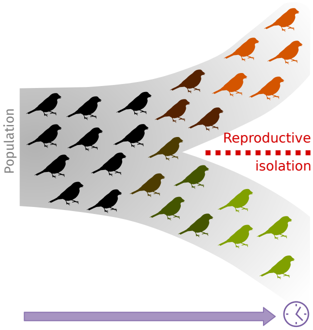
Fate of duplicate genes

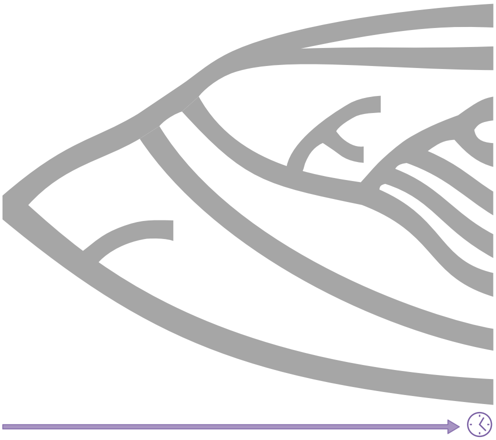
- loss, pseudogeneisation
- selection for increased dosage
- neofunctionalisation
- subfunctionalisation

- Material for adaptation
- Source of genetic reproductive incompatibilities

? Does duplication foster species diversification?

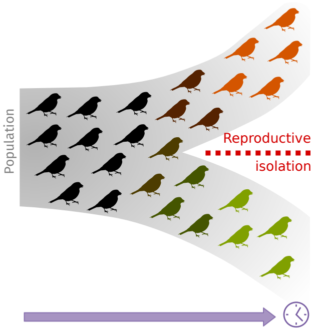

Population
Reproductive
isolation

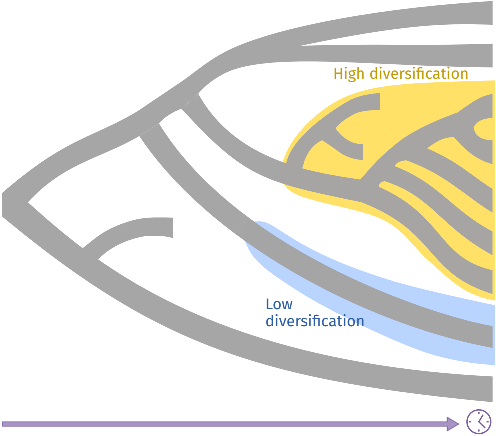




Diversification $=$

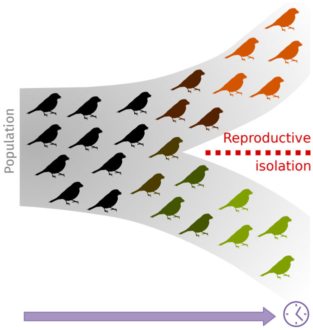
speciation $-$ extinction

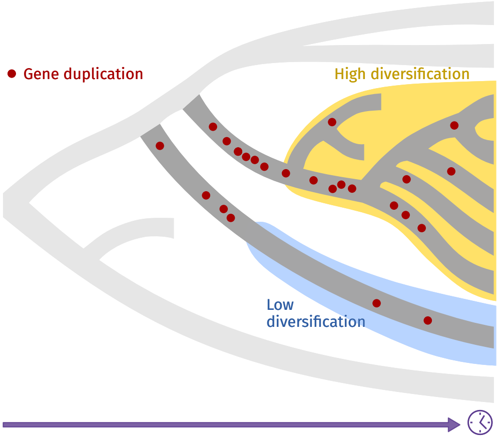




Diversification $=$

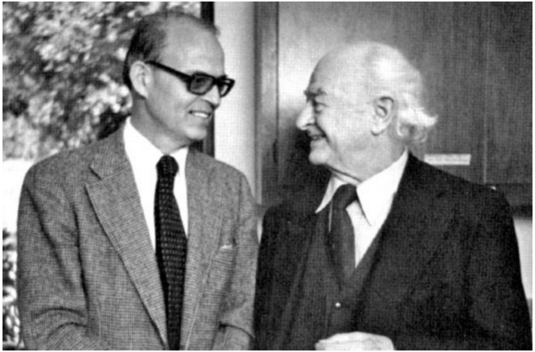
speciation $-$ extinction



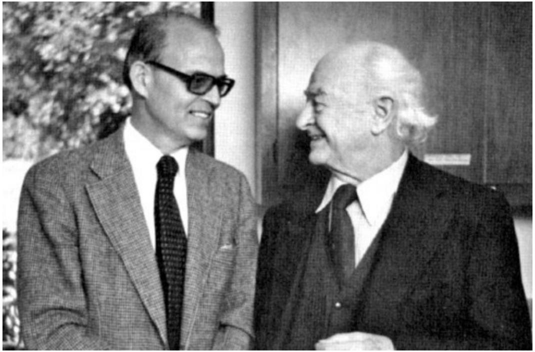


Diversification = speciation − extinction

? What is the dynamic of gene duplication?



Zuckerkandl & Pauling, early 80’s (Linus Pauling Institute Newsletter)



Zuckerkandl & Pauling, early 80’s (Linus Pauling Institute Newsletter)

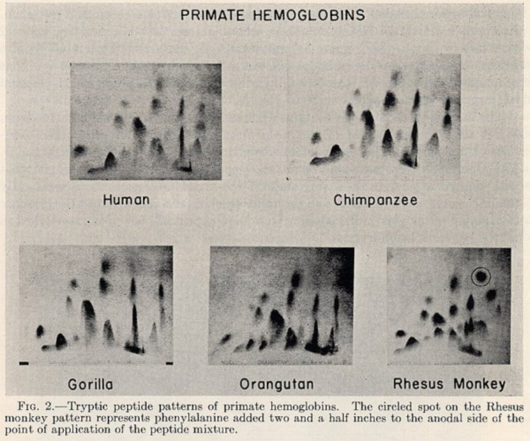
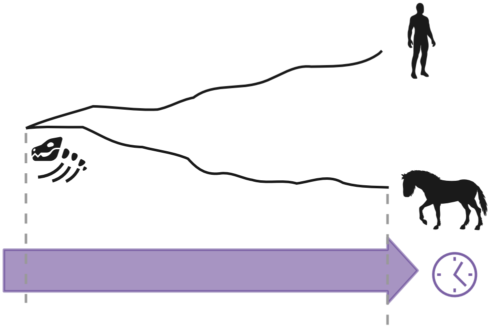


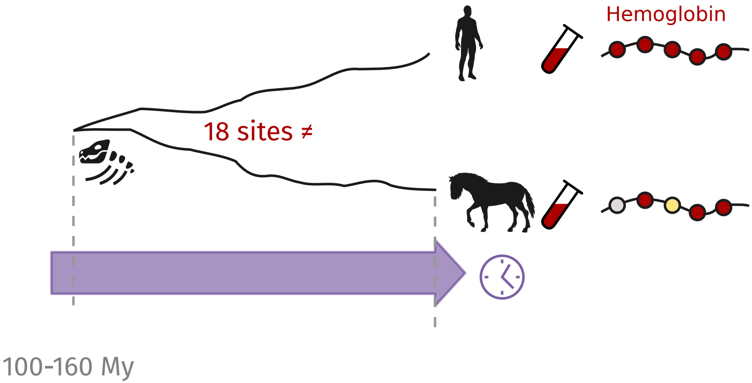
Fig. 2.—Tryptic peptide patterns of primate hemoglobins. The circled spot on the Rhesus monkey pattern represents phenylalanine added two and a half inches to the anodal side of the point of application of the peptide mixture.

Figure 2 from Zuckerkandl, Jones, et al. (1960).

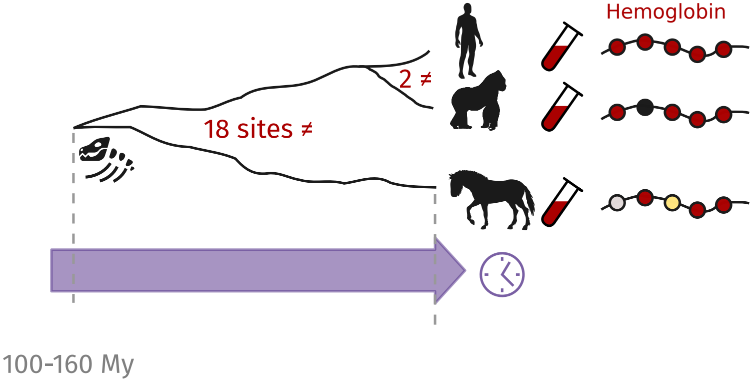


100-160 My

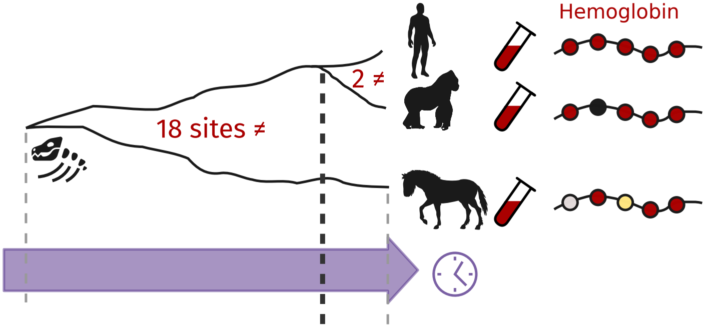
The date computed in Zuckerkandl and Pauling (1962)



The date computed in Zuckerkandl and Pauling (1962)



The date computed in Zuckerkandl and Pauling (1962)



100-160 My

$$\frac{2 \times 130}{18} = 14.5 \text{ My}$$

The date computed in Zuckerkandl and Pauling (1962)

From strict to relaxed clocks

Observations

- Between genes variation[1]

[1] Wolfe et al, 1989

[2] Wu et al, 1985; Britten et al, 1986; Pagel et al, 2006

Observations

- Between genes variation[1]
- Between taxon variation[2]

[1] Wolfe et al, 1989

[2] Wu et al, 1985; Britten et al, 1986; Pagel et al, 2006

Observations

- Between genes variation[1]
- Between taxon variation[2]
- True for DNA as for proteins.

[1] Wolfe et al, 1989

[2] Wu et al, 1985; Britten et al, 1986; Pagel et al, 2006

Observations

- Between genes variation[1]
- Between taxon variation[2]
- True for DNA as for proteins.

[1] Wolfe et al, 1989

[2] Wu et al, 1985; Britten et al, 1986; Pagel et al, 2006

Observations

- Between genes variation[1]
- Between taxon variation[2]
- True for DNA as for proteins.

No reason to be constant

- stochastic variations around a mean
- Amino-acids are subject to selection
- Different taxa have different generation times/population sizes...

[1] Wolfe et al, 1989

[2] Wu et al, 1985; Britten et al, 1986; Pagel et al, 2006

Observations

- Between genes variation[1]
- Between taxon variation[2]
- True for DNA as for proteins.

No reason to be constant

- stochastic variations around a mean
- Amino-acids are subject to selection
- Different taxa have different generation times/population sizes...

$$\underbrace{\textbf{Hetero}}_{\text{non constant}}\underbrace{\textbf{tachy}}_{\text{rate}}$$

[1] Wolfe et al, 1989

[2] Wu et al, 1985; Britten et al, 1986; Pagel et al, 2006

Modern molecular clock models

- account for site variation of substitution rates,

Modern molecular clock models

- account for site variation of substitution rates,
- branch rate relaxation,

Modern molecular clock models

- account for site variation of substitution rates,
- branch rate relaxation,
- models fossils placement probabilities.

Modern molecular clock models

- account for site variation of substitution rates,
- branch rate relaxation,
- models fossils placement probabilities.
- Algorithms of inference:

Method	Mean-Path -Length	Least squares	Markov Chain Monte Carlo (MCMC)
Speed	++	++	– –
Probabilities	✗	✗	✓✓

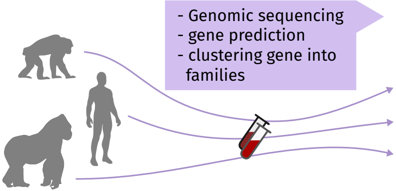
Modern molecular clock models

- account for site variation of substitution rates,
- branch rate relaxation,
- models fossils placement probabilities.
- Algorithms of inference:

Method	Mean-Path -Length	Least squares	Penalised likelihood	Markov Chain Monte Carlo (MCMC)
Speed	++	++	+	– –
Probabilities	✗	✗	✓	✓✓

Sources of uncertainty on dating

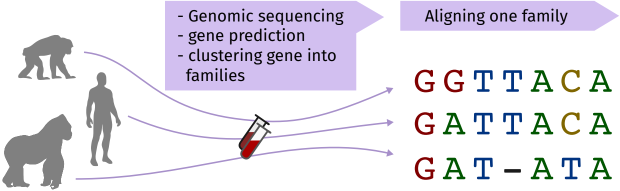
From molecular phylogeny



- Genomic sequencing
- gene prediction
- clustering gene into families

Sources of uncertainty on dating

From molecular phylogeny



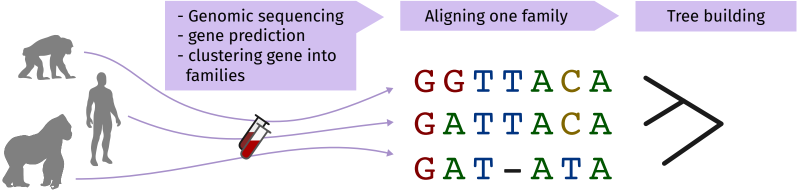
- Genomic sequencing
- gene prediction
- clustering gene into families

Aligning one family

```
G G T T A C A
G A T T A C A
G A T – A T A
```

Sources of uncertainty on dating

From molecular phylogeny



- Genomic sequencing
- gene prediction
- clustering gene into families

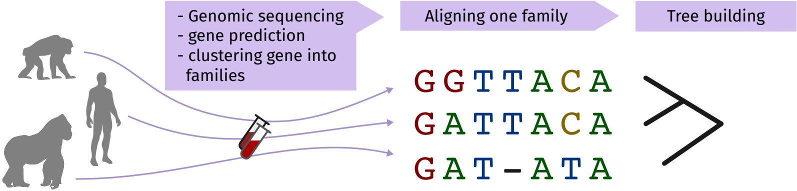
Aligning one family

Tree building

```
G G T T A C A
G A T T A C A
G A T - A T A
```

Sources of uncertainty on dating

From molecular phylogeny



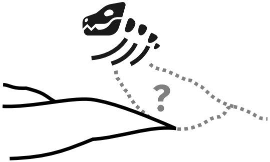
- Genomic sequencing
- gene prediction
- clustering gene into families

Aligning one family

Tree building

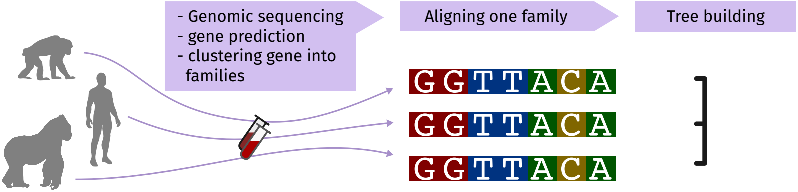
```
G G T T A C A
G A T T A C A
G A T - A T A
```

From fossils



Sources of uncertainty on dating

From molecular phylogeny



- Genomic sequencing
- gene prediction
- clustering gene into families

Aligning one family

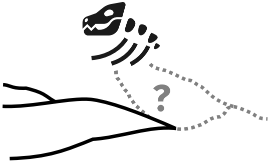
Tree building

GGTTACA

GGTTACA

GGTTACA

From fossils



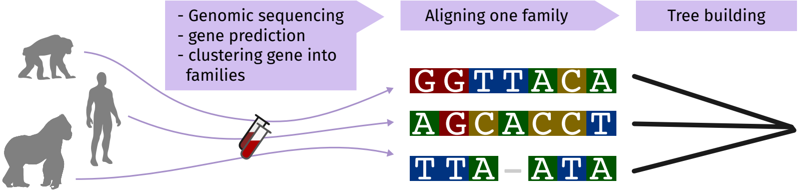
From rates

too low → large variance

too high → saturation

Sources of uncertainty on dating

From molecular phylogeny



- Genomic sequencing
- gene prediction
- clustering gene into families

Aligning one family

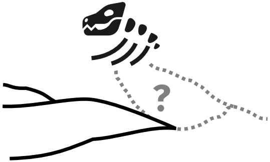
Tree building

GGTTACA

AGCACCT

TTA–ATA

From fossils



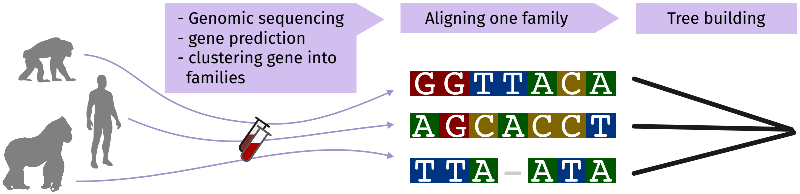
From rates

too low → large variance

too high → saturation

Sources of uncertainty on dating

From molecular phylogeny

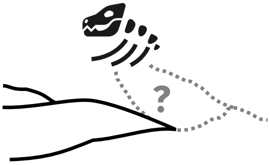


- Genomic sequencing
- gene prediction
- clustering gene into families

Aligning one family

Tree building

From fossils



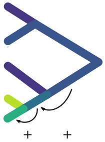
From rates

too low → large variance

too high → saturation

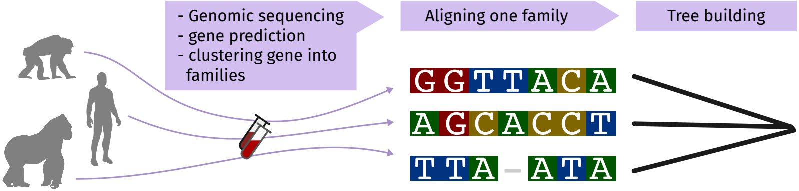
Mode of rate variation

autocorrelated



Sources of uncertainty on dating

From molecular phylogeny



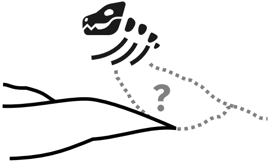
- Genomic sequencing
- gene prediction
- clustering gene into families

Aligning one family

Tree building

GGTTACA
AGCACCT
TTA–ATA

From fossils



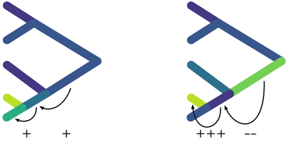
From rates

too low → large variance

too high → saturation

Mode of rate variation

autocorrelated VS uncorrelated



Standard strategy: *concatenate* genes

Standard strategy: *concatenate* genes

?

But can we date *short* sequences?

Standard strategy: *concatenate* genes

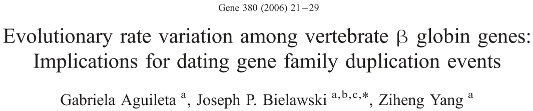
? But can we date *short* sequences?

- viruses
- transposons
- miRNAs
- genes families (duplications)

Standard strategy: *concatenate* genes

? But can we date *short* sequences?

- viruses
- transposons
- miRNAs
- genes families (duplications)



Gene 380 (2006) 21 – 29

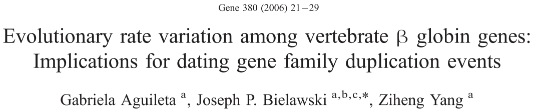
Evolutionary rate variation among vertebrate β globin genes: Implications for dating gene family duplication events

Gabriela Aguileta [a], Joseph P. Bielawski [a,b,c,*], Ziheng Yang [a]

Standard strategy: *concatenate* genes

? But can we date *short* sequences?

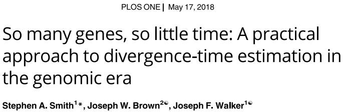
- viruses
- transposons
- miRNAs
- genes families (duplications)



Gene 380 (2006) 21 – 29

Evolutionary rate variation among vertebrate β globin genes: Implications for dating gene family duplication events

Gabriela Aguileta [a], Joseph P. Bielawski [a,b,c,*], Ziheng Yang [a]



PLOS ONE | May 17, 2018

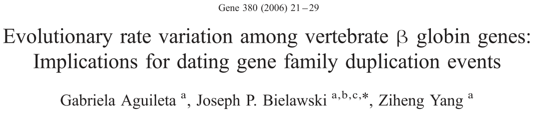
So many genes, so little time: A practical approach to divergence-time estimation in the genomic era

Stephen A. Smith[1*], Joseph W. Brown[2], Joseph F. Walker[1]

Standard strategy: *concatenate* genes

? But can we date *short* sequences?

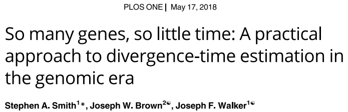
- viruses
- transposons
- miRNAs
- genes families (duplications)



Gene 380 (2006) 21 – 29

Evolutionary rate variation among vertebrate β globin genes: Implications for dating gene family duplication events

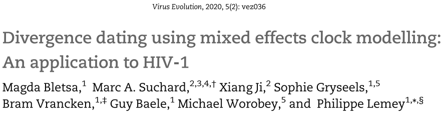
Gabriela Aguileta [a], Joseph P. Bielawski [a,b,c,*], Ziheng Yang [a]



PLOS ONE | May 17, 2018

So many genes, so little time: A practical approach to divergence-time estimation in the genomic era

Stephen A. Smith[1]*, Joseph W. Brown[2☯], Joseph F. Walker[1☯]



Virus Evolution, 2020, 5(2): vez036

Divergence dating using mixed effects clock modelling: An application to HIV-1

Magda Bletsa,[1] Marc A. Suchard,[2,3,4,†] Xiang Ji,[2] Sophie Gryseels,[1,5] Bram Vrancken,[1,‡] Guy Baele,[1] Michael Worobey,[5] and Philippe Lemey[1,*,§]

Can we date gene divergences with confidence?

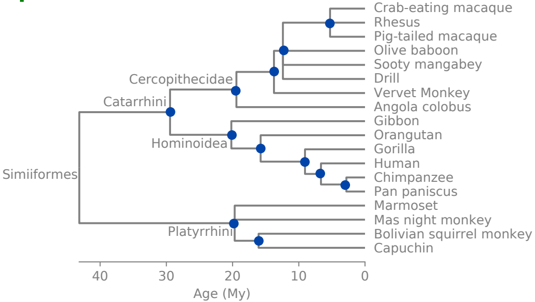
Can we date gene divergences with confidence?

speciations, duplications

The accuracy of dating gene tree events

Testing the dating accuracy with control data

Speciations: reference dates

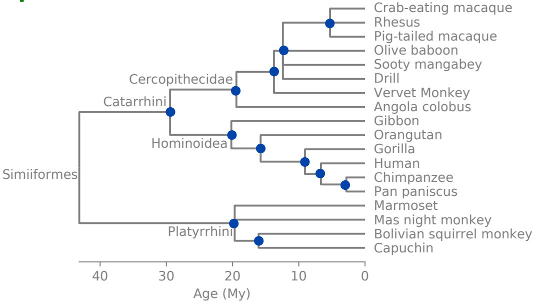


TIMETREE
THE TIMESCALE of LIFE

dos Reis et al. (2018)

Testing the dating accuracy with control data

Speciations: reference dates



TIMETREE
THE TIMESCALE *of* LIFE

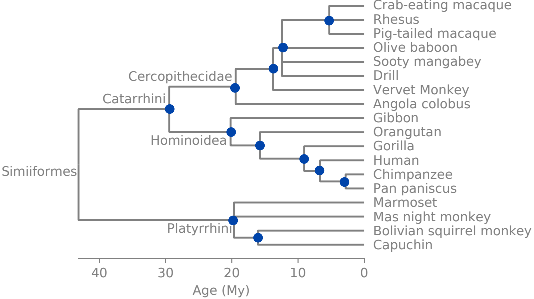
dos Reis et al. (2018)

Test data

EnsEMBL Compara Vertebrata v93

Testing the dating accuracy with control data

Speciations: reference dates



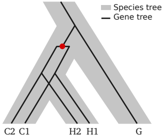
TIMETREE
THE TIMESCALE *of* LIFE

dos Reis et al. (2018)

Test data

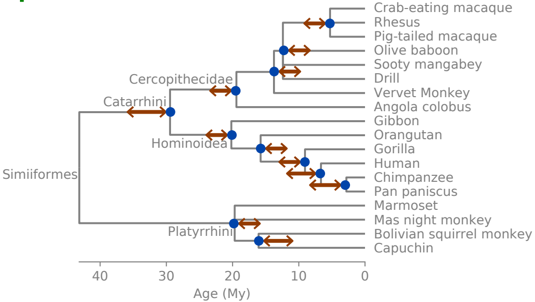
EnsEMBL Compara Vertebrata v93

- 24,562 gene trees & alignments



Testing the dating accuracy with control data

Speciations: reference dates



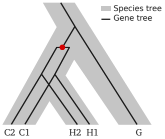
TIMETREE
THE TIMESCALE *of* LIFE

dos Reis et al. (2018)

Test data

EnsEMBL Compara Vertebrata v93

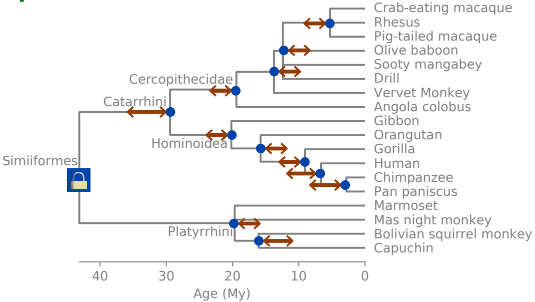
- 24,562 gene trees & alignments



- 5,235 Without duplication/loss.

Testing the dating accuracy with control data

Speciations: reference dates



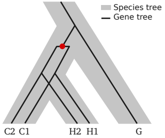
TIMETREE
THE TIMESCALE *of* LIFE

dos Reis et al. (2018)

Test data

EnsEMBL Compara Vertebrata v93

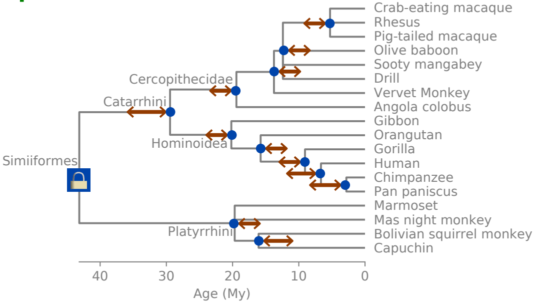
- 24,562 gene trees & alignments



- 5,235 Without duplication/loss.

Testing the dating accuracy with control data

Speciations: reference dates



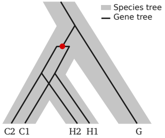
TIMETREE
THE TIMESCALE of LIFE

dos Reis et al. (2018)

Test data

EnsEMBL Compara Vertebrata v93

- 24,562 gene trees & alignments



- 5,235 Without duplication/loss.

? Accuracy of dating events in gene trees?

My steps to molecular dating

Process alignments

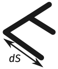
```
GGTTACA
GCTTACA
GAT-ATA
```

My steps to molecular dating

Process alignments

```
GGTTACA
GCTTACA
GAT-ATA
```

Infer synonymous substitutions

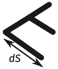


My steps to molecular dating

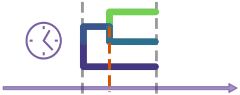
Process alignments

GGTTACA
GCTTACA
GAT–ATA

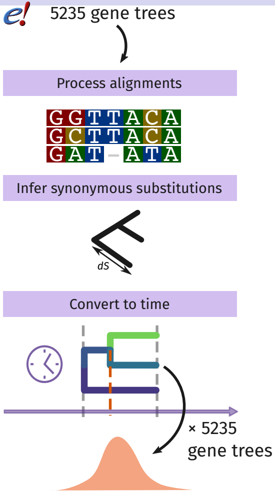
Infer synonymous substitutions



Convert to time



My steps to molecular dating



5235 gene trees

Process alignments

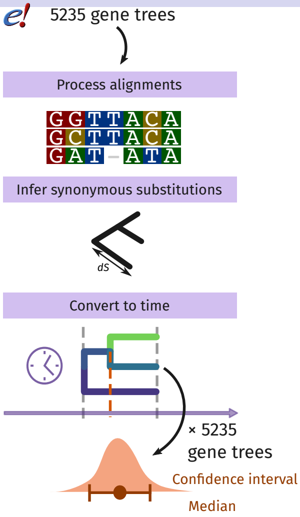
GGTTACA
GCTTACA
GAT-ATA

Infer synonymous substitutions

Convert to time

× 5235 gene trees

My steps to molecular dating



5235 gene trees

Process alignments

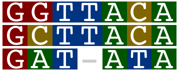
GGTTACA
GCTTACA
GAT-ATA

Infer synonymous substitutions

Convert to time

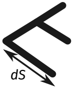
7 dating procedures

Process alignments

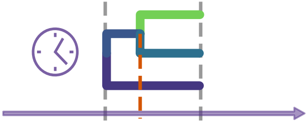


```
GGTTACA
GCTTACA
GAT-ATA
```

Infer synonymous substitutions



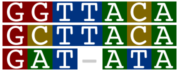
Convert to time



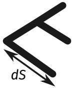
1

7 dating procedures

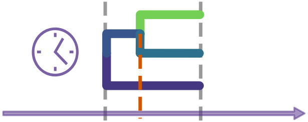
Process alignments



Infer synonymous substitutions



Convert to time

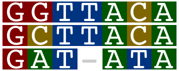


original	Ensembl
cleaned	HmmCleaner[1]: alignment segment filtering.
realigned	FSA: Fast Statistical Aligner[2]: conservative

[1] Di Franco et al. (2019) [2] Bradley et al. (2009)

7 dating procedures

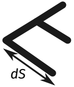
Process alignments



GGTTACA
GCTTACA
GAT–ATA

original	`Ensembl`
cleaned	`HmmCleaner`[1]: alignment segment filtering.
realigned	FSA: Fast Statistical Aligner[2]: conservative

Infer synonymous substitutions

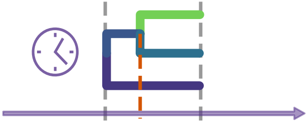


`Codeml`[3]

Measure synonymous substitutions (*dS*).

- *site* model (gamma)
- *branch* model ("free-ratio").

Convert to time



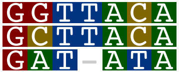
Mean-Path-Length (MPL)[4]

(implemented)

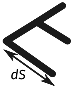
[1] Di Franco et al. (2019) [2] Bradley et al. (2009) [3] Yang (2007)

7 dating procedures

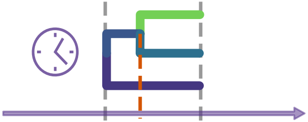
Process alignments



Infer synonymous substitutions



Convert to time



original	Ensembl
cleaned	HmmCleaner[1]: alignment segment filtering.
realigned	FSA: Fast Statistical Aligner[2]: conservative

Codeml[3]

Measure synonymous substitutions (*dS*).

- *site* model (gamma)
- *branch* model ("free-ratio").

Mean-Path-Length (MPL)[4]

(implemented)

(1) *original, siteMPL*

(2) *original, branchMPL*

(3) *cleaned, branchMPL*

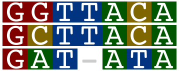
(4) *FSA, branchMPL*

(5) *FSA+cleaned, branchMPL*

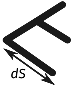
[1] Di Franco et al. (2019) [2] Bradley et al. (2009) [3] Yang (2007) [4] Britton, Oxelman, et al. (2002) and Britton, Anderson, et al. (2007)

7 dating procedures

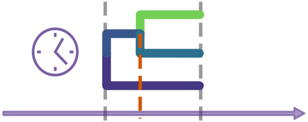
Process alignments



Infer synonymous substitutions



Convert to time



original	`Ensembl`
cleaned	`HmmCleaner`[1]: alignment segment filtering.
realigned	FSA: Fast Statistical Aligner[2]: conservative

`Codeml`[3]

Measure synonymous substitutions (*dS*).

- *site* model (gamma)
- *branch* model ("free-ratio").

Mean-Path-Length (MPL)[4]

(implemented)

`Beast2`[5]

Bayesian (MCMC) simultaneous fit.

(1) *original, siteMPL*

(2) *original, branchMPL*

(3) *cleaned, branchMPL*

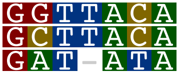
(4) *FSA, branchMPL*

(5) *FSA+cleaned, branchMPL*

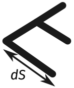
[1] Di Franco et al. (2019) [2] Bradley et al. (2009) [3] Yang (2007) [4] Britton, Oxelman, et al. (2002) and Britton, Anderson, et al. (2007) [5] Bouckaert et al. (2019)

7 dating procedures

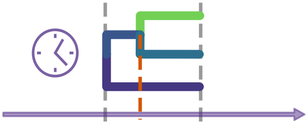
Process alignments



Infer synonymous substitutions



Convert to time



original	`Ensembl`
cleaned	`HmmCleaner`[1]: alignment segment filtering.
realigned	FSA: Fast Statistical Aligner[2]: conservative

`Codeml`[3]

Measure synonymous substitutions (*dS*).

- *site* model (gamma)
- *branch* model ("free-ratio").

Mean-Path-Length (MPL)[4]

(implemented)

`Beast2`[5]

Bayesian (MCMC) simultaneous fit.

(1) *original, siteMPL*

(2) *original, branchMPL*

(3) *cleaned, branchMPL*

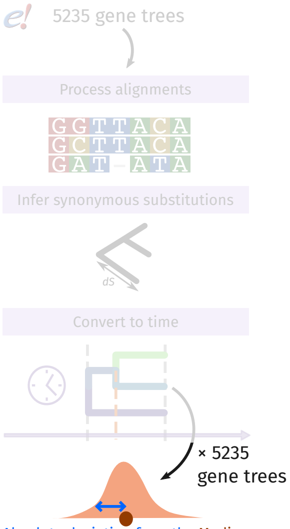
(4) *FSA, branchMPL*

(5) *FSA+cleaned, branchMPL*

(6) *FSA, Beast*

(7) *FSA+cleaned, Beast.*

[1] Di Franco et al. (2019) [2] Bradley et al. (2009) [3] Yang (2007) [4] Britton, Oxelman, et al. (2002) and Britton, Anderson, et al. (2007) [5] Bouckaert et al. (2019)



5235 gene trees

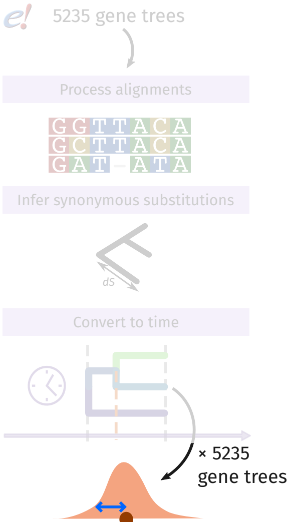
Process alignments

Infer synonymous substitutions

Convert to time

× 5235 gene trees

Absolute deviation from the Median



Absolute deviation from the Median

Dispersion:

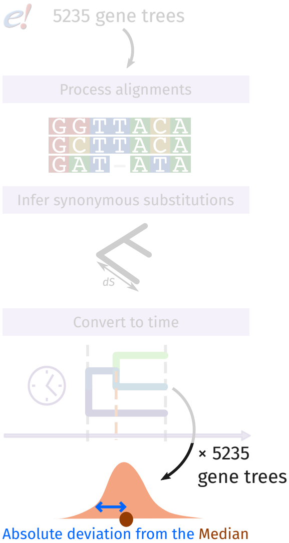
Mean

Absolute

Deviation

from the Median

$$\frac{1}{n}\sum_{i}^{n} |x_i - \mathrm{med}(X)|$$



5235 gene trees

Process alignments

Infer synonymous substitutions

Convert to time

× 5235 gene trees

Absolute deviation from the Median

Dispersion:

Mean

Absolute

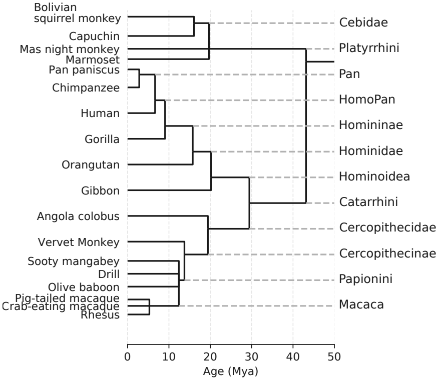
Deviation

from the Median

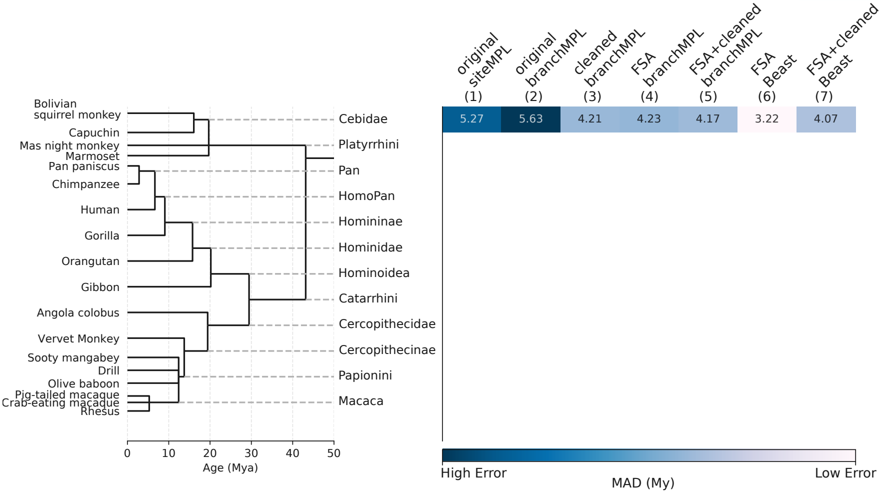
$$\frac{1}{n}\sum_{i}^{n} |x_i - \mathrm{med}(X)|$$

More robust than standard deviation

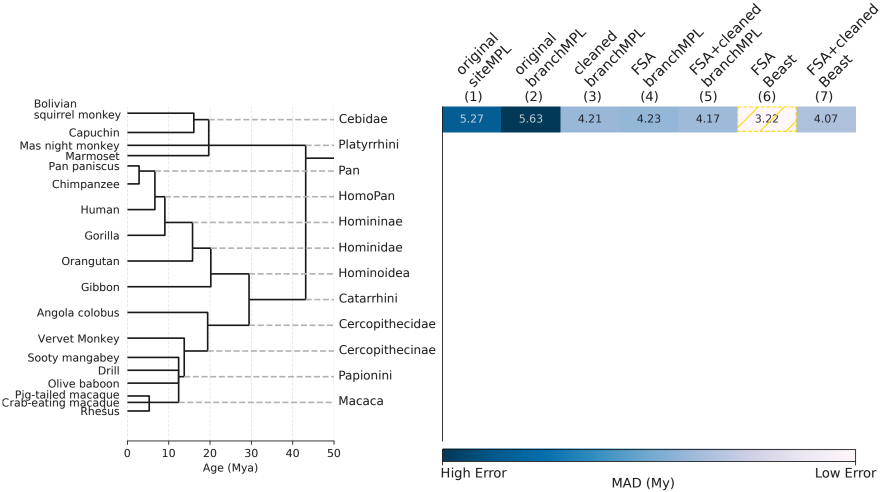
More sensitive than quantile intervals.


Bolivian squirrel monkey
Capuchin
Mas night monkey
Marmoset
Pan paniscus
Chimpanzee
Human
Gorilla
Orangutan
Gibbon
Angola colobus
Vervet Monkey
Sooty mangabey
Drill
Olive baboon
Pig-tailed macaque
Crab-eating macaque
Rhesus
Cebidae
Platyrrhini
Pan
HomoPan
Homininae
Hominidae
Hominoidea
Catarrhini
Cercopithecidae
Cercopithecinae
Papionini
Macaca
0
10
20
30
40
50
Age (Mya)

I. Evaluating the dating confidence

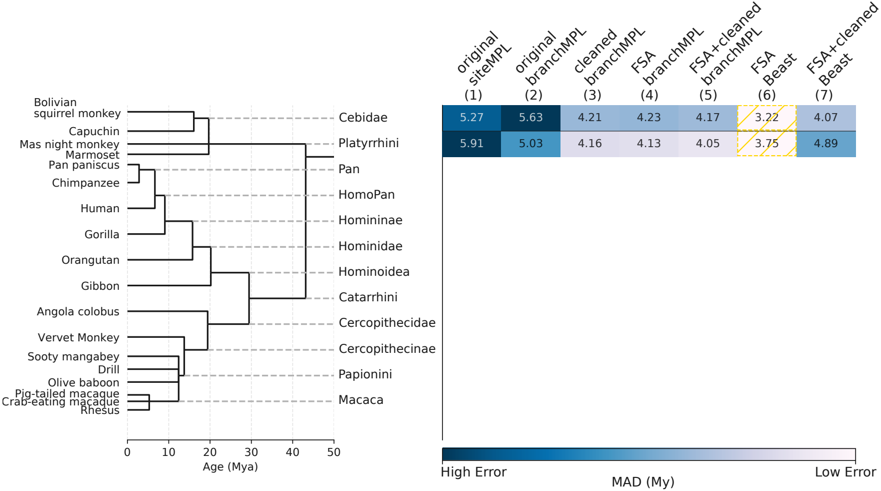


I. Evaluating the dating confidence



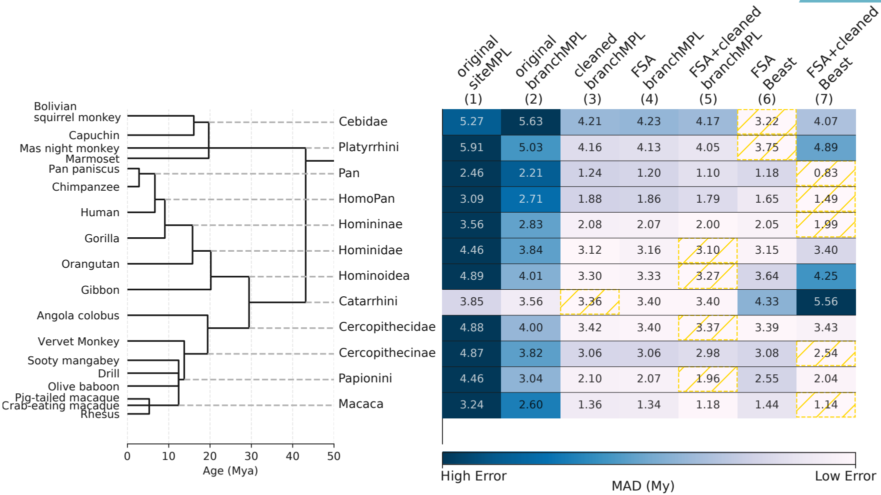
original siteMPL (1)	original branchMPL (2)	cleaned branchMPL (3)	FSA branchMPL (4)	FSA+cleaned branchMPL (5)	FSA Beast (6)	FSA+cleaned Beast (7)
5.27	5.63	4.21	4.23	4.17	3.22	4.07

I. Evaluating the dating confidence



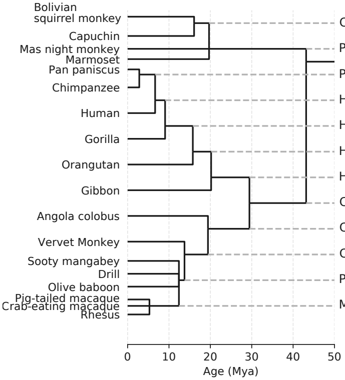
	original siteMPL (1)	original branchMPL (2)	cleaned branchMPL (3)	FSA branchMPL (4)	FSA+cleaned branchMPL (5)	FSA Beast (6)	FSA+cleaned Beast (7)
Cebidae	5.27	5.63	4.21	4.23	4.17	3.22	4.07
Platyrrhini	5.91	5.03	4.16	4.13	4.05	3.75	4.89

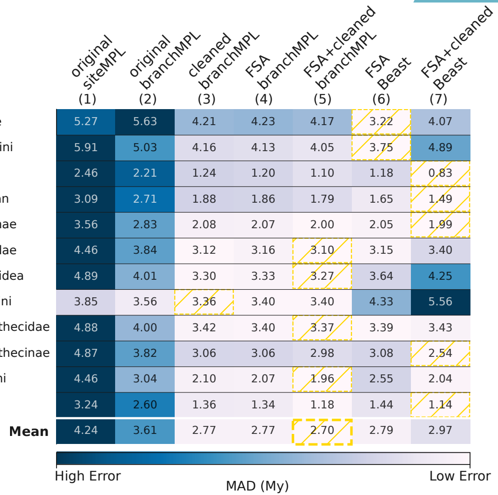
I. Evaluating the dating confidence



	original siteMPL (1)	original branchMPL (2)	cleaned branchMPL (3)	FSA branchMPL (4)	FSA+cleaned branchMPL (5)	FSA Beast (6)	FSA+cleaned Beast (7)
Cebidae	5.27	5.63	4.21	4.23	4.17	3.22	4.07
Platyrrhini	5.91	5.03	4.16	4.13	4.05	3.75	4.89
Pan	2.46	2.21	1.24	1.20	1.10	1.18	0.83
HomoPan	3.09	2.71	1.88	1.86	1.79	1.65	1.49
Homininae	3.56	2.83	2.08	2.07	2.00	2.05	1.99
Hominidae	4.46	3.84	3.12	3.16	3.10	3.15	3.40
Hominoidea	4.89	4.01	3.30	3.33	3.27	3.64	4.25
Catarrhini	3.85	3.56	3.36	3.40	3.40	4.33	5.56
Cercopithecidae	4.88	4.00	3.42	3.40	3.37	3.39	3.43
Cercopithecinae	4.87	3.82	3.06	3.06	2.98	3.08	2.54
Papionini	4.46	3.04	2.10	2.07	1.96	2.55	2.04
Macaca	3.24	2.60	1.36	1.34	1.18	1.44	1.14

I. Evaluating the dating confidence

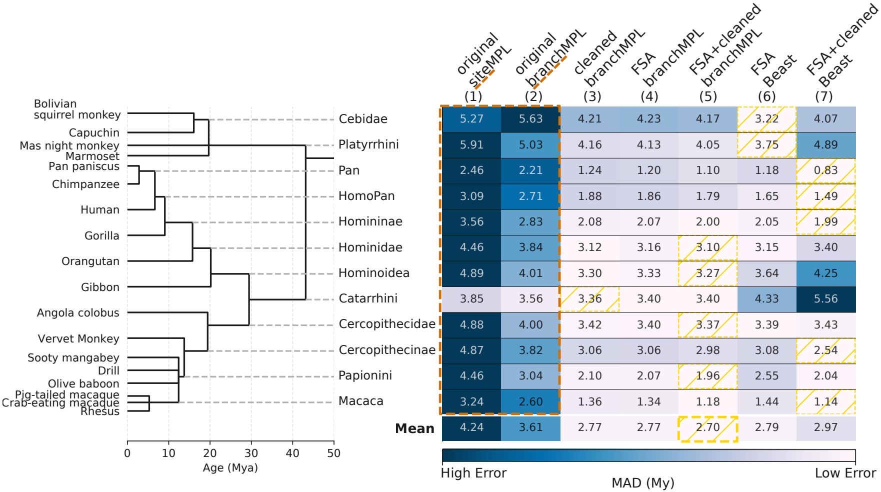




	original siteMPL (1)	original branchMPL (2)	cleaned branchMPL (3)	FSA branchMPL (4)	FSA+cleaned branchMPL (5)	FSA Beast (6)	FSA+cleaned Beast (7)
Cebidae	5.27	5.63	4.21	4.23	4.17	3.22	4.07
Platyrrhini	5.91	5.03	4.16	4.13	4.05	3.75	4.89
Pan	2.46	2.21	1.24	1.20	1.10	1.18	0.83
HomoPan	3.09	2.71	1.88	1.86	1.79	1.65	1.49
Homininae	3.56	2.83	2.08	2.07	2.00	2.05	1.99
Hominidae	4.46	3.84	3.12	3.16	3.10	3.15	3.40
Hominoidea	4.89	4.01	3.30	3.33	3.27	3.64	4.25
Catarrhini	3.85	3.56	3.36	3.40	3.40	4.33	5.56
Cercopithecidae	4.88	4.00	3.42	3.40	3.37	3.39	3.43
Cercopithecinae	4.87	3.82	3.06	3.06	2.98	3.08	2.54
Papionini	4.46	3.04	2.10	2.07	1.96	2.55	2.04
Macaca	3.24	2.60	1.36	1.34	1.18	1.44	1.14
Mean	4.24	3.61	2.77	2.77	2.70	2.79	2.97

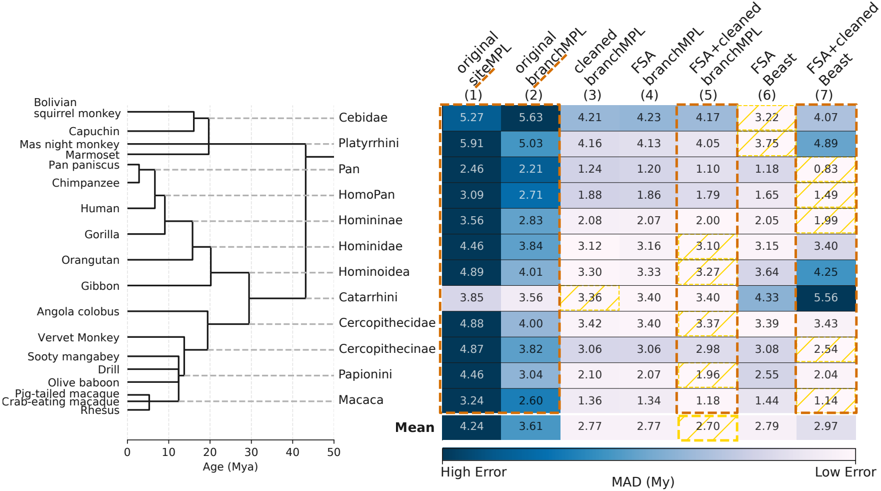
High Error — MAD (My) — Low Error

I. Evaluating the dating confidence

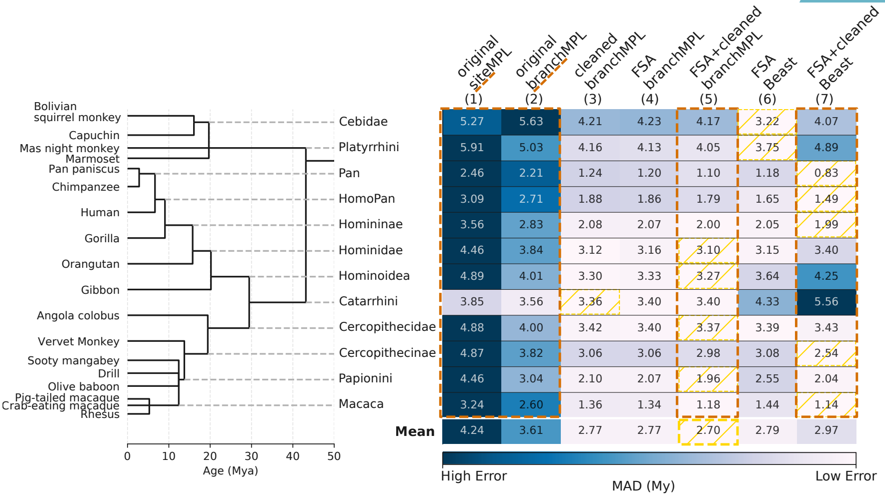


	original siteMPL (1)	original branchMPL (2)	cleaned branchMPL (3)	FSA branchMPL (4)	FSA+cleaned branchMPL (5)	FSA Beast (6)	FSA+cleaned Beast (7)
Cebidae	5.27	5.63	4.21	4.23	4.17	3.22	4.07
Platyrrhini	5.91	5.03	4.16	4.13	4.05	3.75	4.89
Pan	2.46	2.21	1.24	1.20	1.10	1.18	0.83
HomoPan	3.09	2.71	1.88	1.86	1.79	1.65	1.49
Homininae	3.56	2.83	2.08	2.07	2.00	2.05	1.99
Hominidae	4.46	3.84	3.12	3.16	3.10	3.15	3.40
Hominoidea	4.89	4.01	3.30	3.33	3.27	3.64	4.25
Catarrhini	3.85	3.56	3.36	3.40	3.40	4.33	5.56
Cercopithecidae	4.88	4.00	3.42	3.40	3.37	3.39	3.43
Cercopithecinae	4.87	3.82	3.06	3.06	2.98	3.08	2.54
Papionini	4.46	3.04	2.10	2.07	1.96	2.55	2.04
Macaca	3.24	2.60	1.36	1.34	1.18	1.44	1.14
Mean	4.24	3.61	2.77	2.77	2.70	2.79	2.97

I. Evaluating the dating confidence



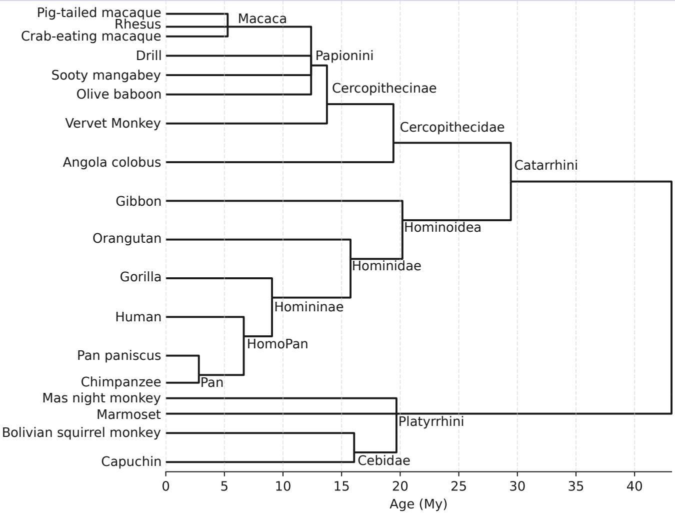
	original siteMPL (1)	original branchMPL (2)	cleaned branchMPL (3)	FSA branchMPL (4)	FSA+cleaned branchMPL (5)	FSA Beast (6)	FSA+cleaned Beast (7)
Cebidae	5.27	5.63	4.21	4.23	4.17	3.22	4.07
Platyrrhini	5.91	5.03	4.16	4.13	4.05	3.75	4.89
Pan	2.46	2.21	1.24	1.20	1.10	1.18	0.83
HomoPan	3.09	2.71	1.88	1.86	1.79	1.65	1.49
Homininae	3.56	2.83	2.08	2.07	2.00	2.05	1.99
Hominidae	4.46	3.84	3.12	3.16	3.10	3.15	3.40
Hominoidea	4.89	4.01	3.30	3.33	3.27	3.64	4.25
Catarrhini	3.85	3.56	3.36	3.40	3.40	4.33	5.56
Cercopithecidae	4.88	4.00	3.42	3.40	3.37	3.39	3.43
Cercopithecinae	4.87	3.82	3.06	3.06	2.98	3.08	2.54
Papionini	4.46	3.04	2.10	2.07	1.96	2.55	2.04
Macaca	3.24	2.60	1.36	1.34	1.18	1.44	1.14
Mean	4.24	3.61	2.77	2.77	2.70	2.79	2.97



	original siteMPL (1)	original branchMPL (2)	cleaned branchMPL (3)	FSA branchMPL (4)	FSA+cleaned branchMPL (5)	FSA Beast (6)	FSA+cleaned Beast (7)
Cebidae	5.27	5.63	4.21	4.23	4.17	3.22	4.07
Platyrrhini	5.91	5.03	4.16	4.13	4.05	3.75	4.89
Pan	2.46	2.21	1.24	1.20	1.10	1.18	0.83
HomoPan	3.09	2.71	1.88	1.86	1.79	1.65	1.49
Homininae	3.56	2.83	2.08	2.07	2.00	2.05	1.99
Hominidae	4.46	3.84	3.12	3.16	3.10	3.15	3.40
Hominoidea	4.89	4.01	3.30	3.33	3.27	3.64	4.25
Catarrhini	3.85	3.56	3.36	3.40	3.40	4.33	5.56
Cercopithecidae	4.88	4.00	3.42	3.40	3.37	3.39	3.43
Cercopithecinae	4.87	3.82	3.06	3.06	2.98	3.08	2.54
Papionini	4.46	3.04	2.10	2.07	1.96	2.55	2.04
Macaca	3.24	2.60	1.36	1.34	1.18	1.44	1.14
Mean	4.24	3.61	2.77	2.77	2.70	2.79	2.97

A branch model and better alignments give lower dispersions

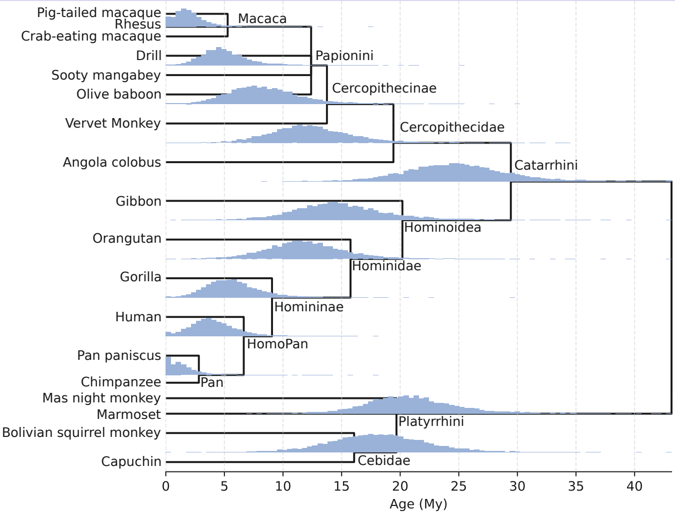
Result: dates from 5235 gene trees



Procedure

FSA+cleaned, branchMPL

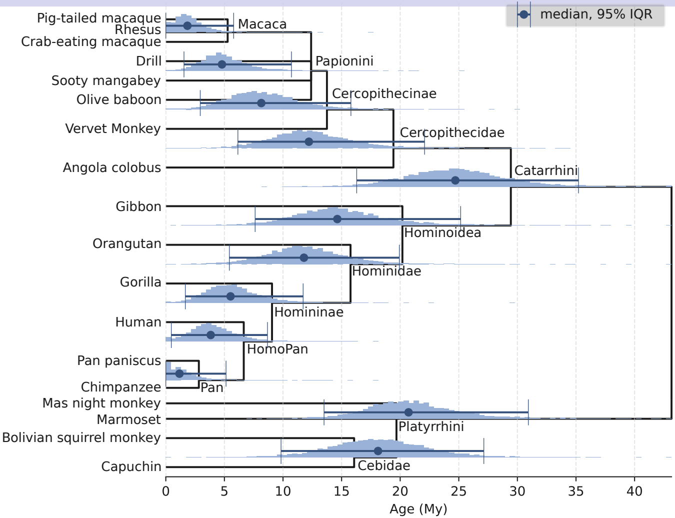
Result: dates from 5235 gene trees



Procedure

FSA+cleaned, branchMPL

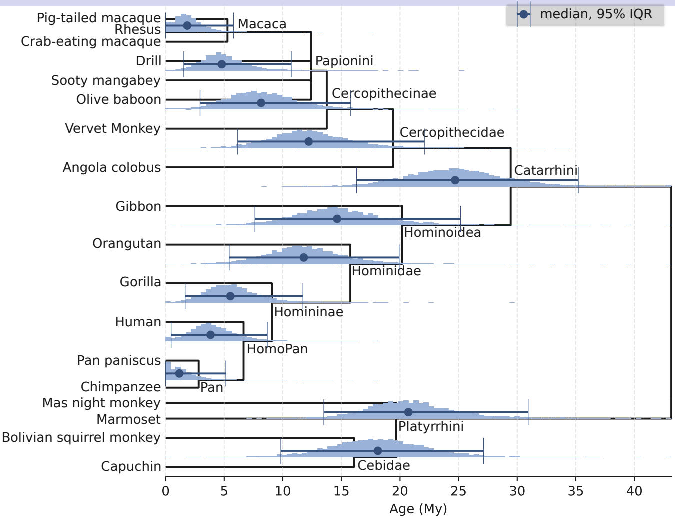
Result: dates from 5235 gene trees



Procedure

FSA+cleaned, branchMPL

Result: dates from 5235 gene trees

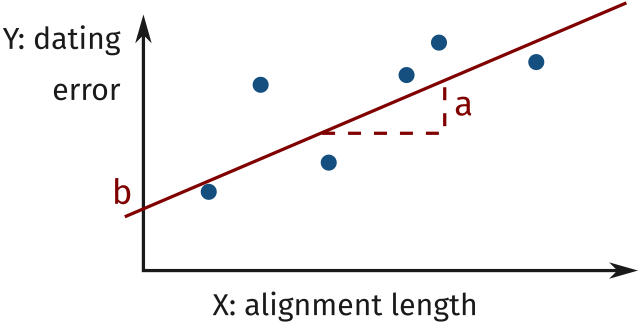


Procedure

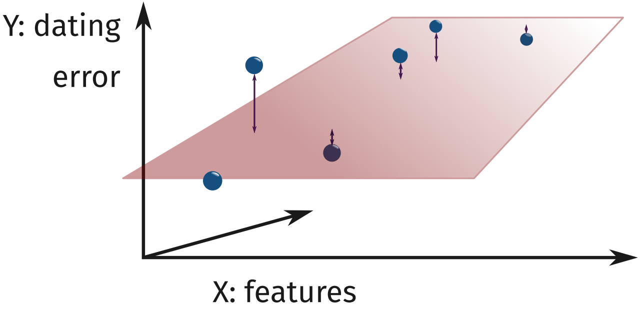
FSA+cleaned, branchMPL

? What makes a tree accurate?

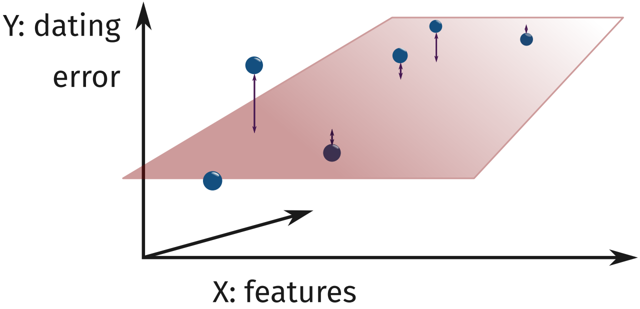
Regression: error VS features of gene trees



Regression: error VS features of gene trees



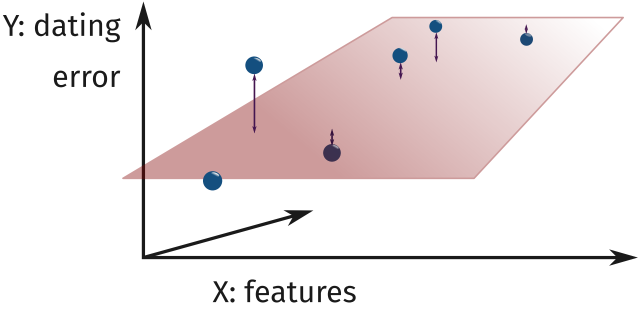
Regression: error VS features of gene trees



60 gene tree features

- **Alignment** features: length, gaps, %GC, entropy...
- **Substitution** features: dS rate, transition/transversion ratio...
- **Cleaning** statistics: proportion removed by `HmmCleaner`, ...

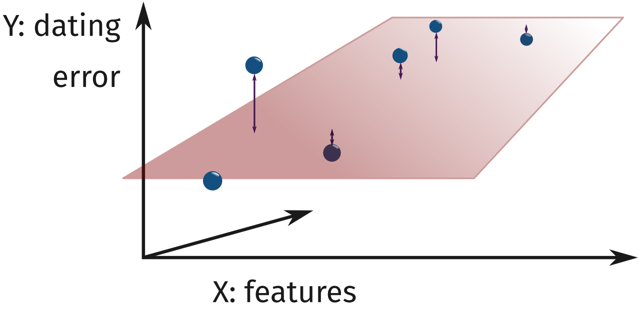
Regression: error VS features of gene trees



60 gene tree features

- **Alignment** features: length, gaps, %GC, entropy...
- **Substitution** features: dS rate, transition/transversion ratio...
- **Cleaning** statistics: proportion removed by `HmmCleaner`, ...

Regression: error VS features of gene trees



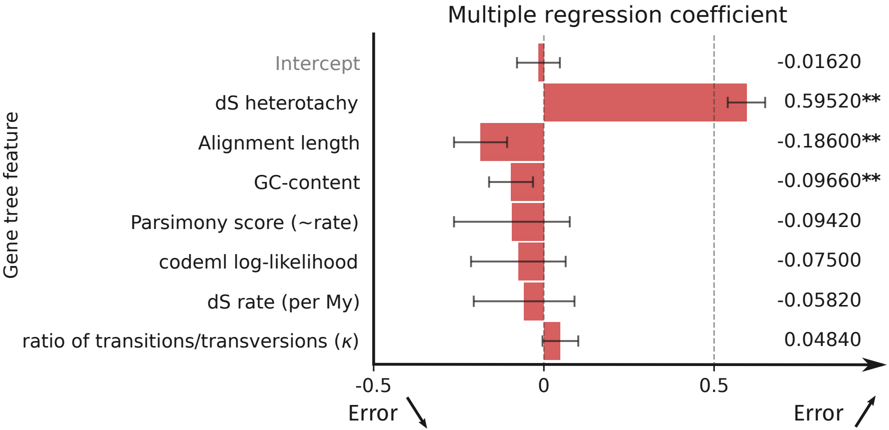
60 gene tree features

- **Alignment** features: length, gaps, %GC, entropy...
- **Substitution** features: dS rate, transition/transversion ratio...
- **Cleaning** statistics: proportion removed by `HmmCleaner`, ...

Dimension reduction

Lasso regression

Rate variation and alignment length influence the dating accuracy



95% confidence intervals

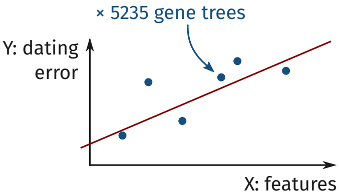
** p-value < 0.01

The controled dates were obtained for constrained trees…

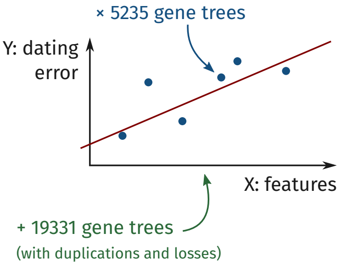
The controled dates were obtained for constrained trees...

? but how accurate are dates in trees with duplications or losses?

Gene trees with loss/duplication are less accurate



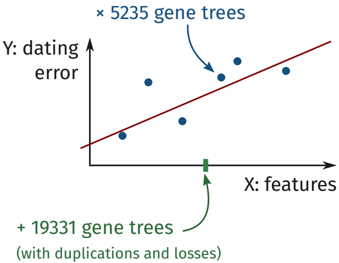
Gene trees with loss/duplication are less accurate



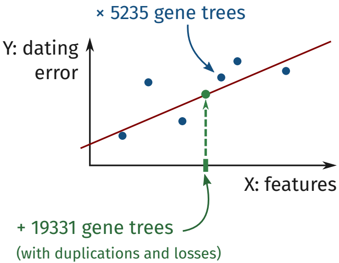
× 5235 gene trees

+ 19331 gene trees
(with duplications and losses)

Gene trees with loss/duplication are less accurate



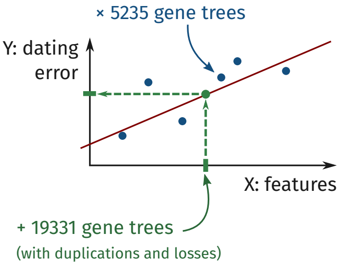
Gene trees with loss/duplication are less accurate



× 5235 gene trees

+ 19331 gene trees
(with duplications and losses)

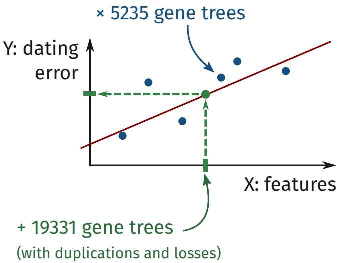
Gene trees with loss/duplication are less accurate



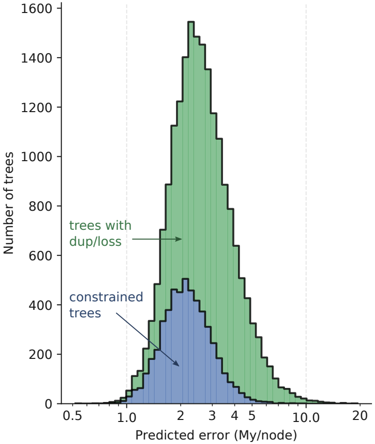
× 5235 gene trees

+ 19331 gene trees
(with duplications and losses)

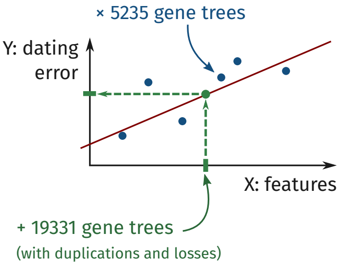
Gene trees with loss/duplication are less accurate


× 5235 gene trees

+ 19331 gene trees
(with duplications and losses)

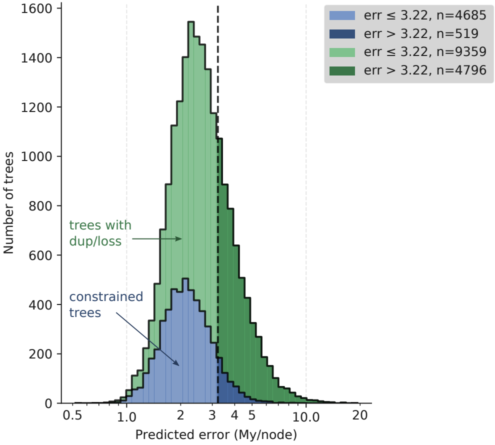


Gene trees with loss/duplication are less accurate

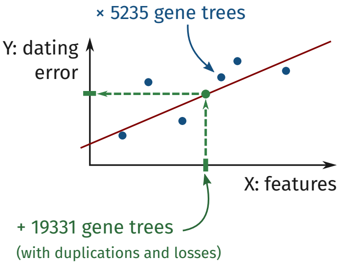


× 5235 gene trees

+ 19331 gene trees
(with duplications and losses)

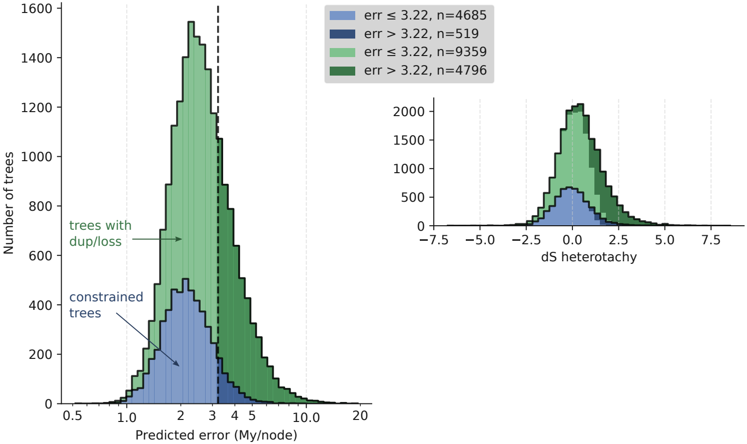


Gene trees with loss/duplication are less accurate

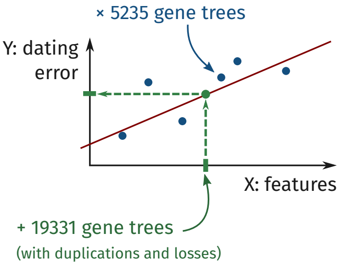


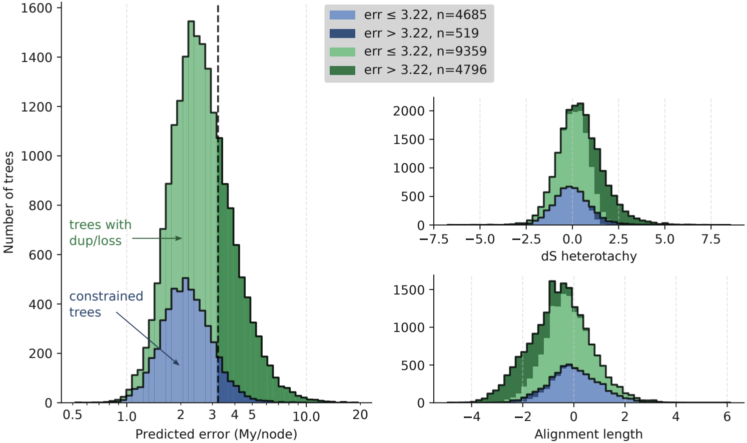
× 5235 gene trees

+ 19331 gene trees
(with duplications and losses)



Gene trees with loss/duplication are less accurate





Summary

Desirable features for confident dating

- dS heterotachy $< 1.7 \times 10^{-3}$ subst/codon/My

Summary

Desirable features for confident dating

- dS heterotachy $< 1.7 \times 10^{-3}$ subst/codon/My
- alignment lengths > 1289 bp.

Summary

Desirable features for confident dating

- dS heterotachy $< 1.7 \times 10^{-3}$ subst/codon/My
- alignment lengths > 1289 bp.

Summary

Desirable features for confident dating

- dS heterotachy $< 1.7 \times 10^{-3}$ subst/codon/My
- alignment lengths > 1289 bp.

“Factors influencing the accuracy in dating single gene trees”.
Louvel, G and H. Roest Crollius
(submitted + bioRχiv)

- **Real genome-wide data**

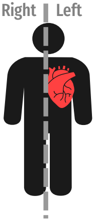
- **Real genome-wide data**
- **Noisy rates at the scale of the gene**

- **Real genome-wide data**
- **Noisy rates at the scale of the gene**
- **Genes with duplications/loss are inaccurate**

- **Real genome-wide data**
- **Noisy rates at the scale of the gene**
- **Genes with duplications/loss are inaccurate**
- **Molecular clock dating on short sequences (virus, genes): challenging.**

Finding gene trees with correlated losses

Jeanne Amiel[1], Christopher Gordon[1], Bruno Reversade[2]

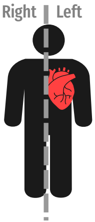


Genetic laterality disorders

- *situs inversus*
- heart & visceral organ defects

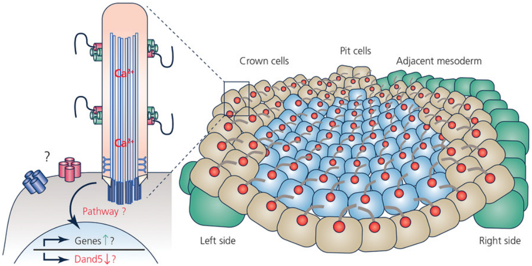
[1] Institut Imagine, Paris
[2] Institute of Medical Biology, Singapore

Jeanne Amiel[1], Christopher Gordon[1], Bruno Reversade[2]



Genetic laterality disorders

- *situs inversus*
- heart & visceral organ defects



The Journal of Physiology

Embryonic Left-Right Organizer (from Tajhya & Delling, 2019)

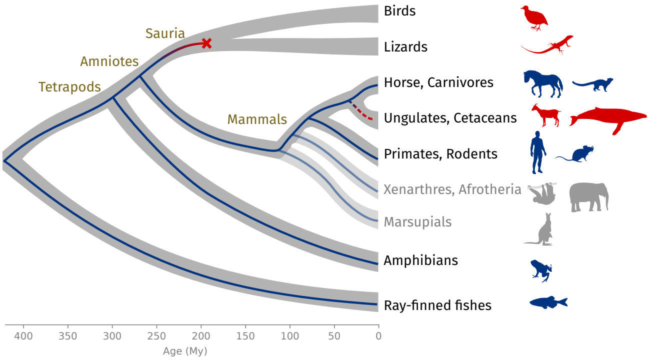
[1] Institut Imagine, Paris
[2] Institute of Medical Biology, Singapore

- The lateralisation mechanism is partially understood
- It is medically relevant (1/10000 births have defects)

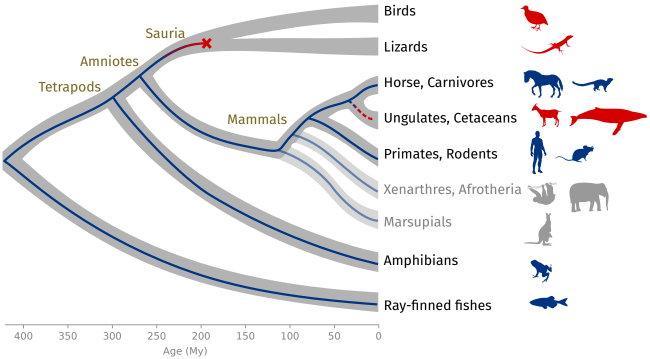
- The lateralisation mechanism is partially understood
- It is medically relevant (1/10000 births have defects)

Genes: MMP21, PKD1L1, DAND5, ...

2 independent losses of motile cilia in amniotes.

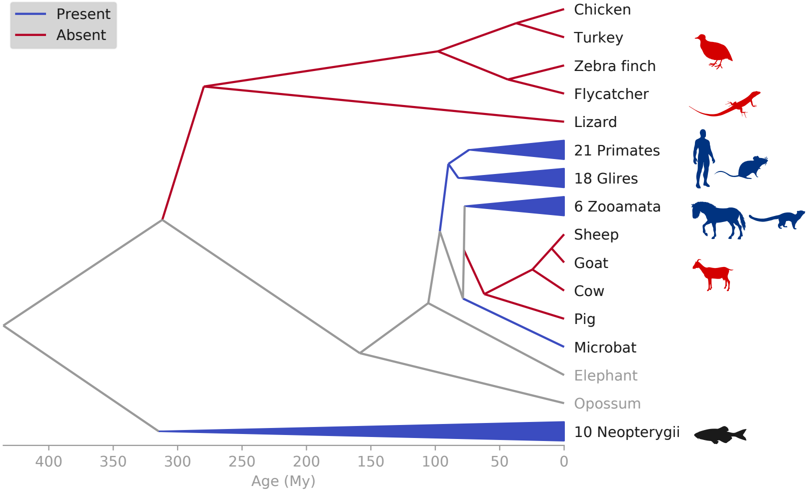


2 independent losses of motile cilia in amniotes. MMP21, PKD1L1, DAND5 genes.

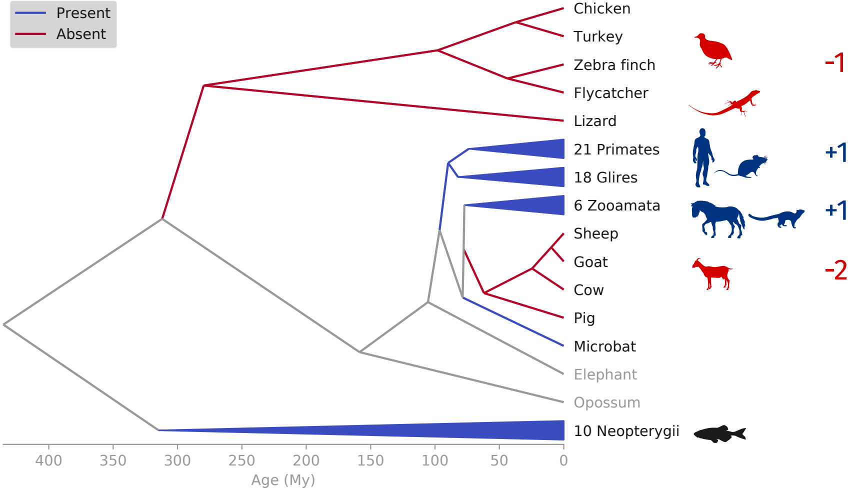


? Can we find unknown lateralisation genes based on their phylogeny?

Phylogenetic presence score



Phylogenetic presence score



Score	Human gene	Description
1.948	MMP21	matrix metallopeptidase 21
1.897	CDC42SE2	CDC42 small effector 2
1.807	ADIPOR1	adiponectin receptor 1
1.756	DAND5	DAN domain BMP antagonist family member 5
1.748	SLC25A18	solute carrier family 25 member 18
1.671	TNFRSF14	TNF receptor superfamily member 14
1.628	TRIM60;TRIM75P	tripartite motif containing 60; 75, pseudogene
1.410	PKD1L1	polycystin 1 like 1, transient receptor potential channel interacting
1.346	RPL41	ribosomal protein L41
1.307	ARL6IP1	ADP ribosylation factor like GTPase 6 interacting protein 1
1.282	MS4A4A;MS4A4E	membrane spanning 4-domains A4A; A4E
1.243	HS3ST3A1	heparan sulfate-glucosamine 3-sulfotransferase 3A1
1.230	C1orf127	chromosome 1 open reading frame 127
1.230	SMIM22	small integral membrane protein 22
1.205	L3MBTL4	L3MBTL4, histone methyl-lysine binding protein
1.2	PPP2R2D	protein phosphatase 2 regulatory subunit Bdelta
1.182	ADGRG4	adhesion G protein-coupled receptor G4
1.141	PSKH2	protein serine kinase H2
1.141	AC067968.1; ZNF155, 221-225, 230, 234, 284	Zinc fingers proteins
1.128	LMLN2	leishmanolysin like peptidase 2
1.102		
1.102	AC022167.5	lipopolysaccharide-induced tumor necrosis factor-alpha factor-like
1.082	CHIT1	chitinase 1
1.076	RPL39	ribosomal protein L39
1.076	KIAA1586	KIAA1586

Score	Human gene	Description
1.948	MMP21	matrix metallopeptidase 21
1.897	CDC42SE2	CDC42 small effector 2
1.807	ADIPOR1	adiponectin receptor 1
1.756	DAND5	DAN domain BMP antagonist family member 5
1.748	SLC25A18	solute carrier family 25 member 18
1.671	TNFRSF14	TNF receptor superfamily member 14
1.628	TRIM60;TRIM75P	tripartite motif containing 60; 75, pseudogene
1.410	PKD1L1	polycystin 1 like 1, transient receptor potential channel interacting
1.346	RPL41	ribosomal protein L41
1.307	ARL6IP1	ADP ribosylation factor like GTPase 6 interacting protein 1
1.282	MS4A4A;MS4A4E	membrane spanning 4-domains A4A; A4E
1.243	HS3ST3A1	heparan sulfate-glucosamine 3-sulfotransferase 3A1
1.230	C1orf127	chromosome 1 open reading frame 127
1.230	SMIM22	small integral membrane protein 22
1.205	L3MBTL4	L3MBTL4, histone methyl-lysine binding protein
1.2	PPP2R2D	protein phosphatase 2 regulatory subunit Bdelta
1.182	ADGRG4	adhesion G protein-coupled receptor G4
1.141	PSKH2	protein serine kinase H2
1.141	AC067968.1; ZNF155, 221-225, 230, 234, 284	Zinc fingers proteins
1.128	LMLN2	leishmanolysin like peptidase 2
1.102		
1.102	AC022167.5	lipopolysaccharide-induced tumor necrosis factor-alpha factor-like
1.082	CHIT1	chitinase 1
1.076	RPL39	ribosomal protein L39
1.076	KIAA1586	KIAA1586

Score	Human gene	Description
1.948	MMP21	matrix metallopeptidase 21
1.897	CDC42SE2	CDC42 small effector 2
1.807	ADIPOR1	adiponectin receptor 1
1.756	DAND5	DAN domain BMP antagonist family member 5
1.748	SLC25A18	solute carrier family 25 member 18
1.671	TNFRSF14	TNF receptor superfamily member 14
1.628	TRIM60;TRIM75P	tripartite motif containing 60; 75, pseudogene
1.410	PKD1L1	polycystin 1 like 1, transient receptor potential channel interacting
1.346	RPL41	ribosomal protein L41
1.307	ARL6IP1	ADP ribosylation factor like GTPase 6 interacting protein 1
1.282	MS4A4A;MS4A4E	membrane spanning 4-domains A4A; A4E
1.243	HS3ST3A1	heparan sulfate-glucosamine 3-sulfotransferase 3A1
1.230	C1orf127	chromosome 1 open reading frame 127
1.230	SMIM22	small integral membrane protein 22
1.205	L3MBTL4	L3MBTL4, histone methyl-lysine binding protein
1.2	PPP2R2D	protein phosphatase 2 regulatory subunit Bdelta
1.182	ADGRG4	adhesion G protein-coupled receptor G4
1.141	PSKH2	protein serine kinase H2
1.141	AC067968.1; ZNF155, 221-225, 230, 234, 284	Zinc fingers proteins
1.128	LMLN2	leishmanolysin like peptidase 2
1.102		
1.102	AC022167.5	lipopolysaccharide-induced tumor necrosis factor-alpha factor-like
1.082	CHIT1	chitinase 1
1.076	RPL39	ribosomal protein L39
1.076	KIAA1586	KIAA1586

With signal peptide

Score	Human gene	Description
1.948	MMP21	matrix metallopeptidase 21
1.897	CDC42SE2	CDC42 small effector 2
1.807	ADIPOR1	adiponectin receptor 1
1.756	DAND5	DAN domain BMP antagonist family member 5
1.748	SLC25A18	solute carrier family 25 member 18
1.671	TNFRSF14	TNF receptor superfamily member 14
1.628	TRIM60;TRIM75P	tripartite motif containing 60; 75, pseudogene
1.410	PKD1L1	polycystin 1 like 1, transient receptor potential channel interacting
1.346	RPL41	ribosomal protein L41
1.307	ARL6IP1	ADP ribosylation factor like GTPase 6 interacting protein 1
1.282	MS4A4A;MS4A4E	membrane spanning 4-domains A4A; A4E
1.243	HS3ST3A1	heparan sulfate-glucosamine 3-sulfotransferase 3A1
1.230	C1orf127	chromosome 1 open reading frame 127
1.230	SMIM22	small integral membrane protein 22
1.205	L3MBTL4	L3MBTL4, histone methyl-lysine binding protein
1.2	PPP2R2D	protein phosphatase 2 regulatory subunit Bdelta
1.182	ADGRG4	adhesion G protein-coupled receptor G4
1.141	PSKH2	protein serine kinase H2
1.141	AC067968.1; ZNF155, 221-225, 230, 234, 284	Zinc fingers proteins
1.128	LMLN2	leishmanolysin like peptidase 2
1.102		
1.102	AC022167.5	lipopolysaccharide-induced tumor necrosis factor-alpha factor-like
1.082	CHIT1	chitinase 1
1.076	RPL39	ribosomal protein L39
1.076	KIAA1586	KIAA1586

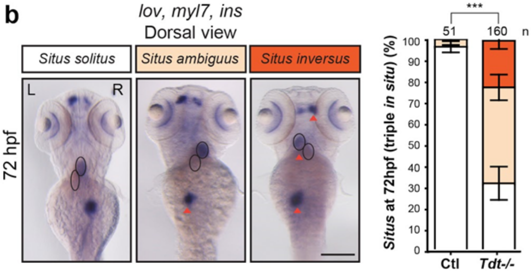
With signal peptide

Score	Human gene	Description
1.948	MMP21	matrix metallopeptidase 21
1.897	CDC42SE2	CDC42 small effector 2
1.807	ADIPOR1	adiponectin receptor 1
1.756	DAND5	DAN domain BMP antagonist family member 5
1.748	SLC25A18	solute carrier family 25 member 18
1.671	TNFRSF14	TNF receptor superfamily member 14
1.628	TRIM60;TRIM75P	tripartite motif containing 60; 75, pseudogene
1.410	PKD1L1	polycystin 1 like 1, transient receptor potential channel interacting
1.346	RPL41	ribosomal protein L41
1.307	ARL6IP1	ADP ribosylation factor like GTPase 6 interacting protein 1
1.282	MS4A4A;MS4A4E	membrane spanning 4-domains A4A; A4E
1.243	HS3ST3A1	heparan sulfate-glucosamine 3-sulfotransferase 3A1
1.230	C1orf127	chromosome 1 open reading frame 127
1.230	SMIM22	small integral membrane protein 22
1.205	L3MBTL4	L3MBTL4, histone methyl-lysine binding protein
1.2	PPP2R2D	protein phosphatase 2 regulatory subunit Bdelta
1.182	ADGRG4	adhesion G protein-coupled receptor G4
1.141	PSKH2	protein serine kinase H2
1.141	AC067968.1; ZNF155, 221-225, 230, 234, 284	Zinc fingers proteins
1.128	LMLN2	leishmanolysin like peptidase 2
1.102		
1.102	AC022167.5	lipopolysaccharide-induced tumor necrosis factor-alpha factor-like
1.082	CHIT1	chitinase 1
1.076	RPL39	ribosomal protein L39
1.076	KIAA1586	KIAA1586

With signal peptide

Score	Human gene	Description
1.948	MMP21	matrix metallopeptidase 21
1.897	CDC42SE2	CDC42 small effector 2
1.807	ADIPOR1	adiponectin receptor 1
1.756	DAND5	DAN domain BMP antagonist family member 5
1.748	SLC25A18	solute carrier family 25 member 18
1.671	TNFRSF14	TNF receptor superfamily member 14
1.628	TRIM60;TRIM75P	tripartite motif containing 60; 75, pseudogene
1.410	PKD1L1	polycystin 1 like 1, transient receptor potential channel interacting
1.346	RPL41	ribosomal protein L41
1.307	ARL6IP1	ADP ribosylation factor like GTPase 6 interacting protein 1
1.282	MS4A4A;MS4A4E	membrane spanning 4-domains A4A; A4E
1.243	HS3ST3A1	heparan sulfate-glucosamine 3-sulfotransferase 3A1
1.230	C1orf127 = **TDT**	chromosome 1 open reading frame 127
1.230	SMIM22	small integral membrane protein 22
1.205	L3MBTL4	L3MBTL4, histone methyl-lysine binding protein
1.2	PPP2R2D	protein phosphatase 2 regulatory subunit Bdelta
1.182	ADGRG4	adhesion G protein-coupled receptor G4
1.141	PSKH2	protein serine kinase H2
1.141	AC067968.1; ZNF155, 221-225, 230, 234, 284	Zinc fingers proteins
1.128	LMLN2 = **ALED**	leishmanolysin like peptidase 2
1.102		
1.102	AC022167.5	lipopolysaccharide-induced tumor necrosis factor-alpha factor-like
1.082	CHIT1	chitinase 1
1.076	RPL39	ribosomal protein L39
1.076	KIAA1586	KIAA1586

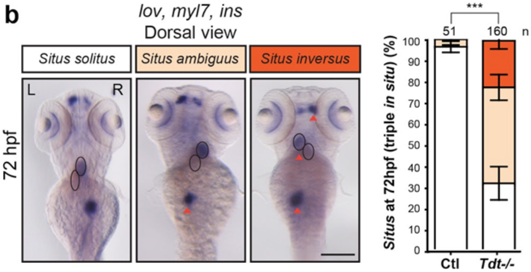
Experimental validation of TDT and ALED.



“A functional operon delineates an extracellular pathway that controls Left-Right patterning only in animals with a ciliated organizer”.

Szenker-Ravi, E et al. (submitted)

Experimental validation of TDT and ALED.



Screened 2 millions Conserved Non-Coding Elements.

"A functional operon delineates an extracellular pathway that controls Left-Right patterning only in animals with a ciliated organizer".

Szenker-Ravi, E et al. (submitted)

Conclusions & perspectives

Correlating duplications with diversification: suggestions

- Estimating *rates* does not necessarily requires dates (Birth-Death model)

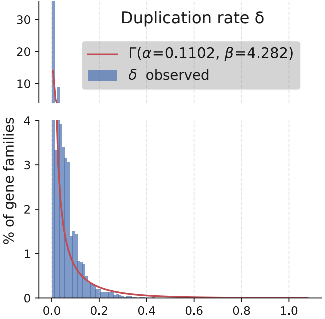
Correlating duplications with diversification: suggestions

- Estimating *rates* does not necessarily requires dates (Birth-Death model)

Correlating duplications with diversification: suggestions

- Estimating *rates* does not necessarily requires dates (Birth-Death model)

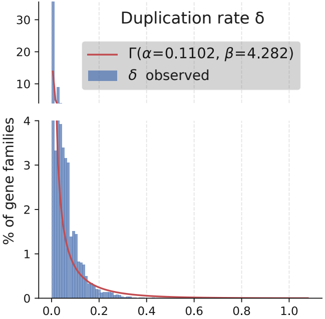
Family-wise duplication rates



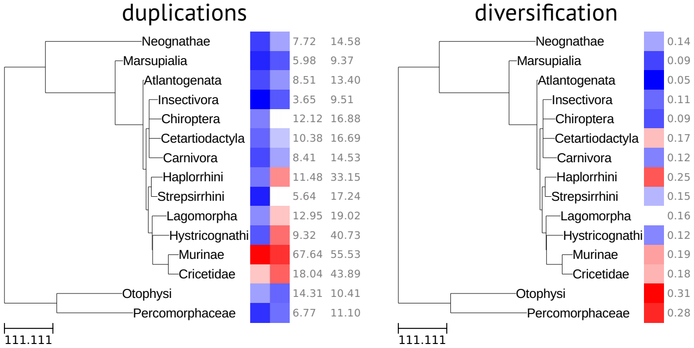
Correlating duplications with diversification: suggestions

- Estimating *rates* does not necessarily requires dates (Birth-Death model)

Family-wise duplication rates



Phylogeny-aware correlation



Perspectives

- “clocks without rocks”[1]

[1] Tiley et al, 2020

Perspectives

- “clocks without rocks”[1]
- flood of new genomes (Genome10K vertebrates)

[1] Tiley et al, 2020

Perspectives

- “clocks without rocks”[1]
- flood of new genomes (Genome10K vertebrates)

[1] Tiley et al, 2020

Perspectives

- “clocks without rocks”[1]
- flood of new genomes (Genome10K vertebrates)

Are needed:

- Phylogenetic Comparative Methods

[1] Tiley et al, 2020

Perspectives

- "clocks without rocks"[1]
- flood of new genomes (Genome10K vertebrates)

Are needed:

- Phylogenetic Comparative Methods
- Genomic information complementary to sequences.

[1] Tiley et al, 2020

Hugues Roest Crollius

Dyogen team
Alexandra Louis
Camille Berthelot
Yves Clément
Lambert Moyon
Élise Parey
François Giudicelli
Nga Thi-Thuy Nguyen
Gosia, Axelle,
Franklin

IT platform
Pierre Vincens
Bilel, Catherine, Nolwenn,
Phi-Phong

PhD committee & advice
Gilles Fischer
Nicolas Lartillot
Hélène Morlon

4th year funding
Fondation pour la
Recherche Médicale
FRM FDT201904008392

Supplementary

Desirable features for confident dating

From 5235 constrained trees, 24566 total.

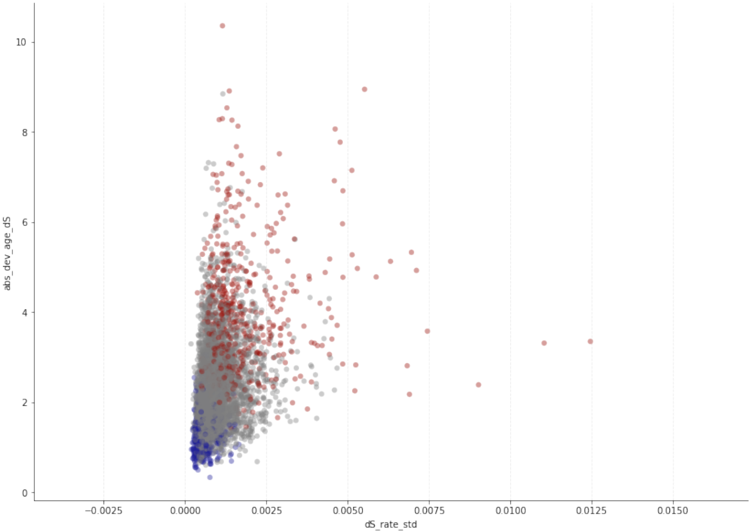
Number of trees

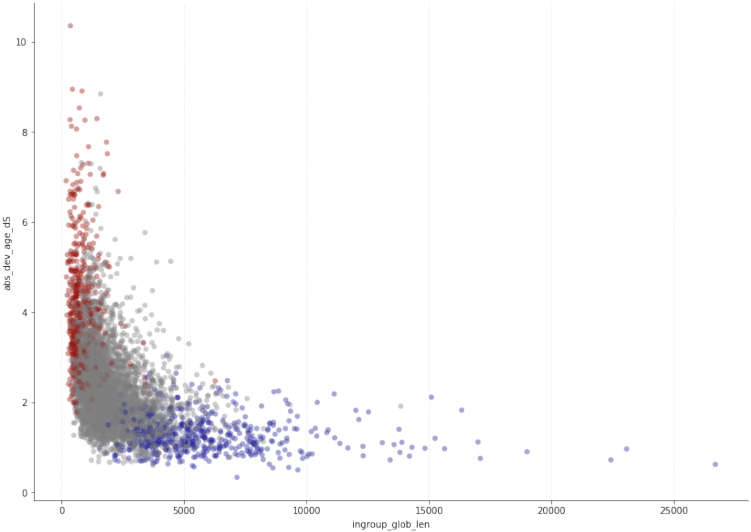
- alignment length < 1289
- dS heterotachy > 1.7×10^{-3}:

Alternative threshold

50% of tested trees are in the 10th decile of the training set dating error.

- alignment length: 822 bp (622 trees)
- dS heterotachy: 2.23×10^{-3} (241 trees)

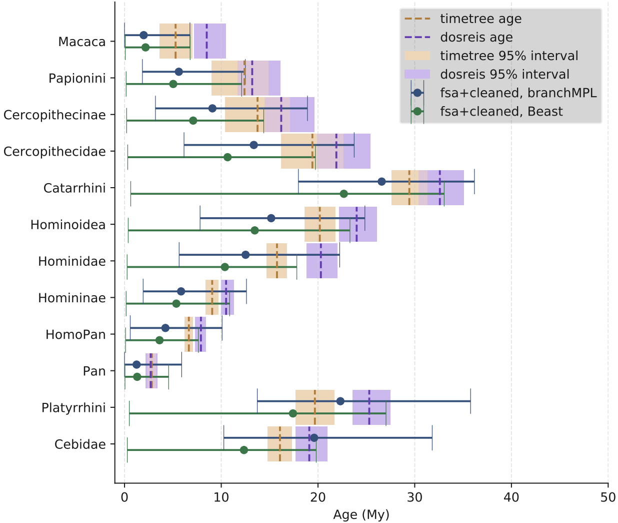

abs_dev_age_dS
0
2
4
6
8
10
−0.0025
0.0000
0.0025
0.0050
0.0075
0.0100
0.0125
0.0150
dS_rate_std


abs_dev_age_dS
0
2
4
6
8
10
ingroup_glob_len
0
5000
10000
15000
20000
25000

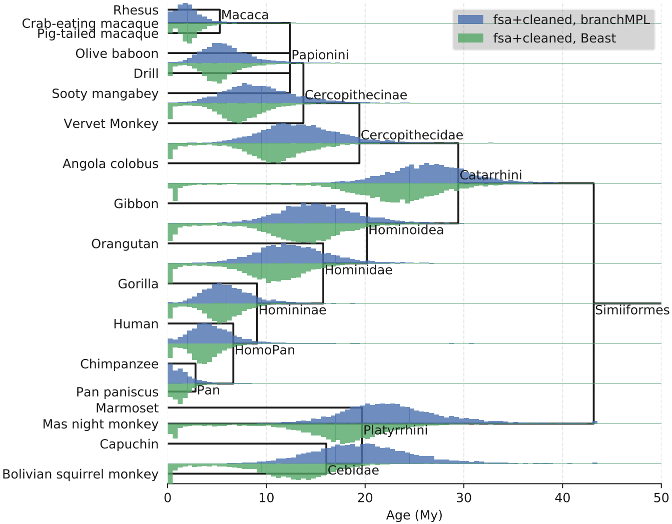
Desirable features for confident dating – table

	Error (My)	dS heterotachy (10^{-3} subst/codon/My)	Alignment length (nucleotides)	Mean dS rate (10^{-3} subst/codon/My)	Mean GC (%)
10% lowest predicted accuracy	$3,74 \pm 1,42$	$1,7 \pm 1,1$	1289 ± 702	$2,02 \pm 1,07$	$52,0 \pm 8,5$
10% highest predicted accuracy	$1,28 \pm 0,39$	$0,46 \pm 0,26$	7031 ± 3866	$2,02 \pm 0,95$	$55,7 \pm 7,4$

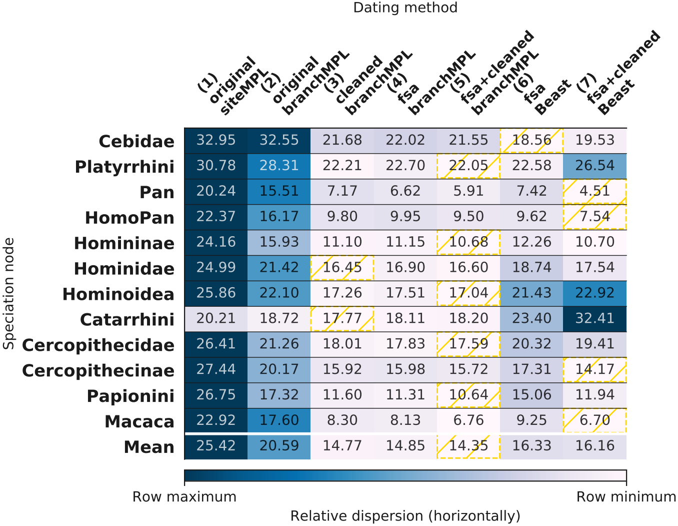
Estimated VS reference ages



Age distributions

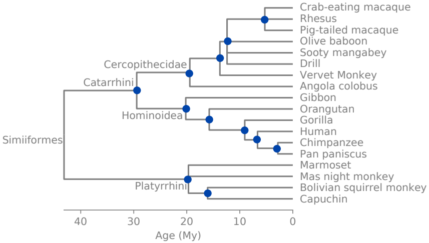


IQR95 by procedure



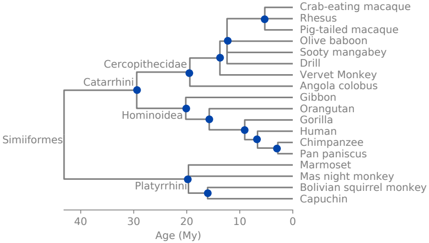
	(1) original siteMPL	(2) original branchMPL	(3) cleaned branchMPL	(4) fsa branchMPL	(5) fsa+cleaned branchMPL	(6) fsa Beast	(7) fsa+cleaned Beast
Cebidae	32.95	32.55	21.68	22.02	21.55	18.56	19.53
Platyrrhini	30.78	28.31	22.21	22.70	22.05	22.58	26.54
Pan	20.24	15.51	7.17	6.62	5.91	7.42	4.51
HomoPan	22.37	16.17	9.80	9.95	9.50	9.62	7.54
Homininae	24.16	15.93	11.10	11.15	10.68	12.26	10.70
Hominidae	24.99	21.42	16.45	16.90	16.60	18.74	17.54
Hominoidea	25.86	22.10	17.26	17.51	17.04	21.43	22.92
Catarrhini	20.21	18.72	17.77	18.11	18.20	23.40	32.41
Cercopithecidae	26.41	21.26	18.01	17.83	17.59	20.32	19.41
Cercopithecinae	27.44	20.17	15.92	15.98	15.72	17.31	14.17
Papionini	26.75	17.32	11.60	11.31	10.64	15.06	11.94
Macaca	22.92	17.60	8.30	8.13	6.76	9.25	6.70
Mean	25.42	20.59	14.77	14.85	14.35	16.33	16.16

Regressing features specific to each speciation



Considering each speciation independently: 12 regressions.

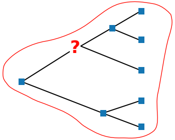
Regressing features specific to each speciation



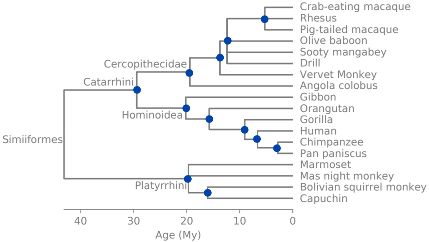
Considering each speciation independently:

12 regressions.

New X variable: local heterotachy

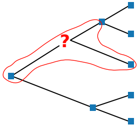


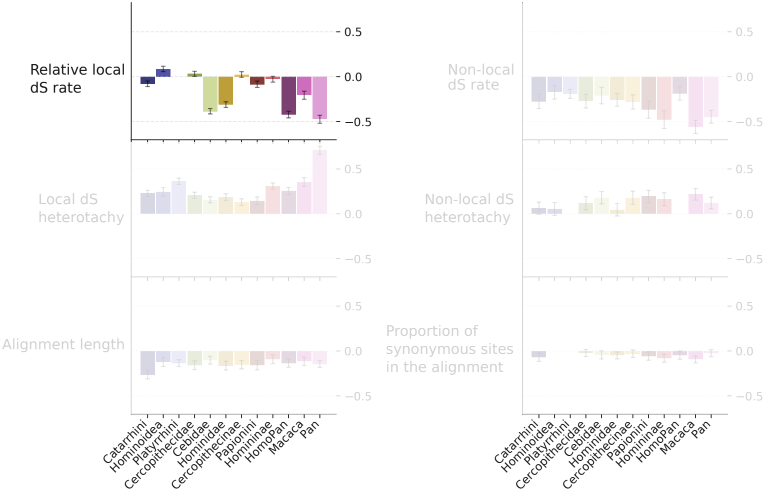
Regressing features specific to each speciation



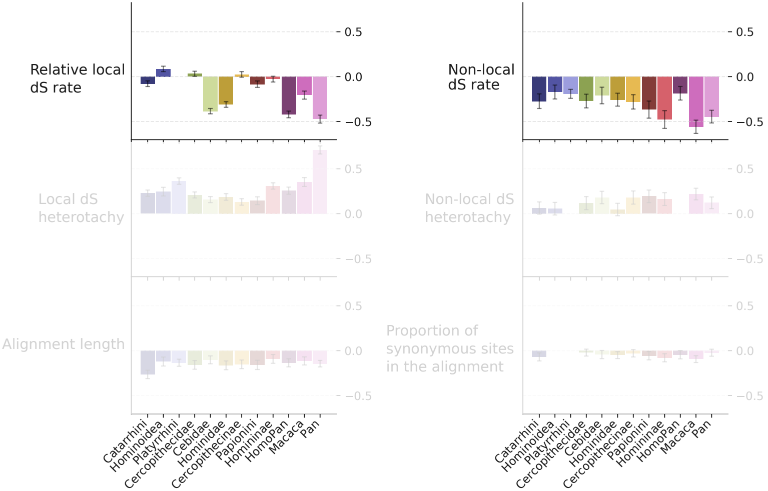
Considering each speciation independently: 12 regressions.

New X variable: local heterotachy



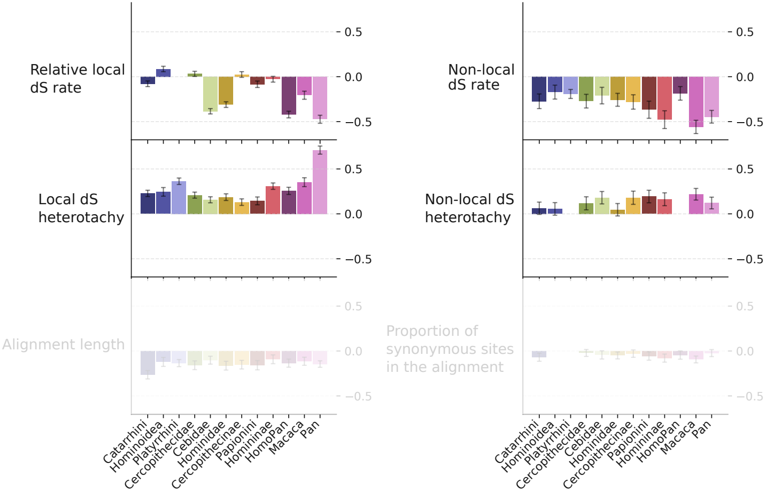


⊖ Local rate:
variable impact



⊖ Local rate: variable impact

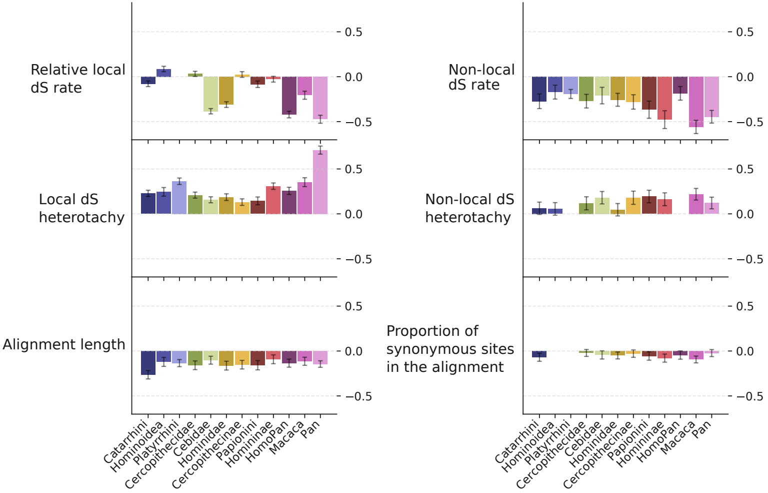
⊖ Global rate: consistent impact



⊖ Local rate: variable impact

⊖ Global rate: consistent impact

⊕ Heterotachy: more impact locally

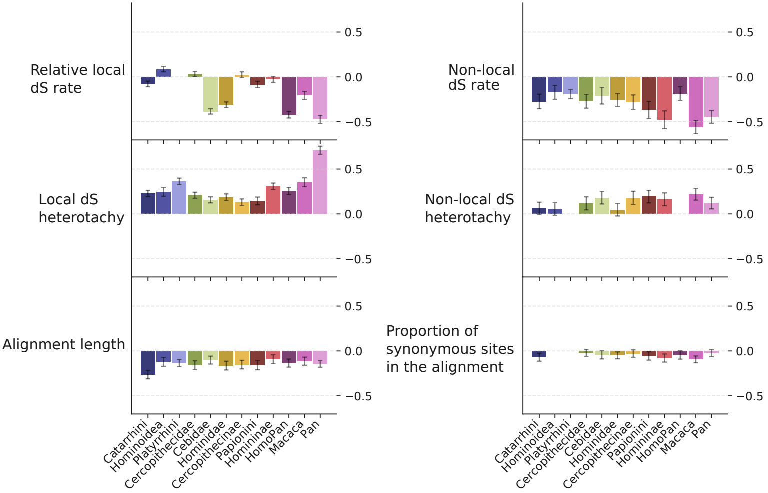


⊖ Local rate: variable impact

⊖ Global rate: consistent impact

⊕ Heterotachy: more impact locally

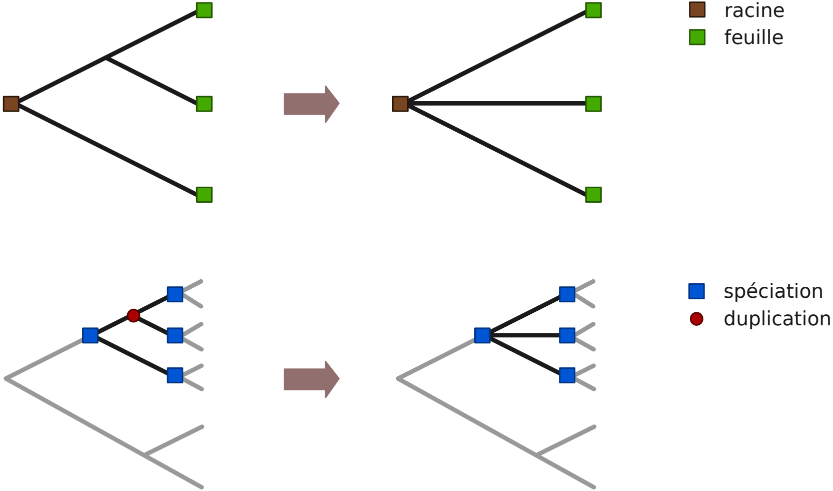
⊖ Alignment length



- ⊖ Local rate: variable impact
- ⊖ Global rate: consistent impact
- ⊕ Heterotachy: more impact locally
- ⊖ Alignment length

- No obvious trend by age

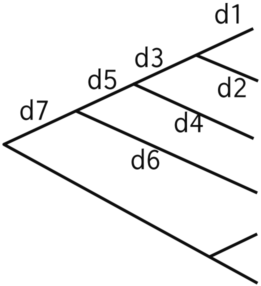
Approximate rate & heterotachy



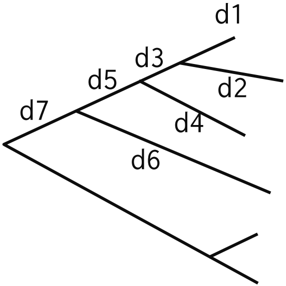
🟫 racine
🟩 feuille

🟦 spéciation
🔴 duplication

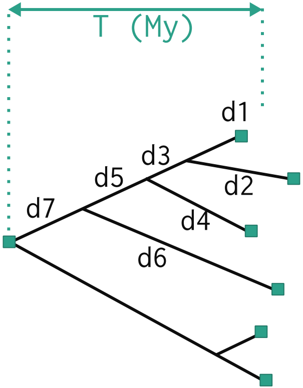
Mean-Path-Length (MPL) algorithm (similar to UPGMA)



Mean-Path-Length (MPL) algorithm (similar to UPGMA)

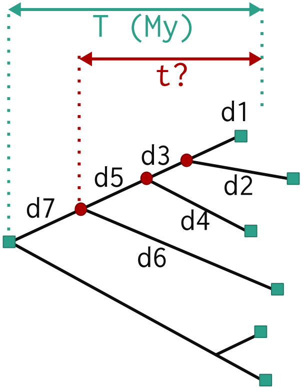


Mean-Path-Length (MPL) algorithm (similar to UPGMA)



🟩 Calibration

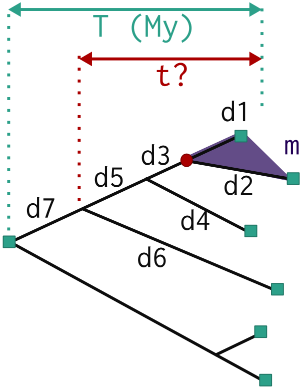
Mean-Path-Length (MPL) algorithm (similar to UPGMA)



🟩 Calibration

● Estimated

Mean-Path-Length (MPL) algorithm (similar to UPGMA)

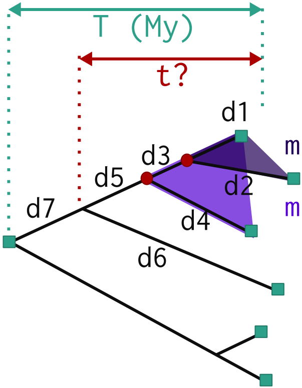


$m1 = (d1 + d2)/2$

🟩 Calibration

● Estimated

Mean-Path-Length (MPL) algorithm (similar to UPGMA)



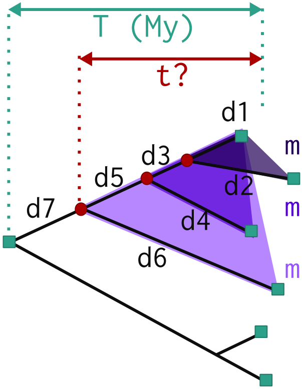
🟩 Calibration

● Estimated

m1 = (d1 + d2)/2

m2 = (2*(m1+d3) + d4)/3

Mean-Path-Length (MPL) algorithm (similar to UPGMA)

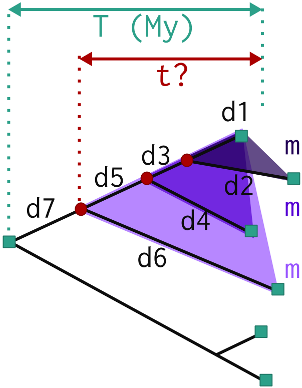


m1 = (d1 + d2)/2

m2 = (2*(m1+d3) + d4)/3

m3 = (3*(m2+d5) + d6)/4

Mean-Path-Length (MPL) algorithm (similar to UPGMA)



🟩 Calibration

● Estimated

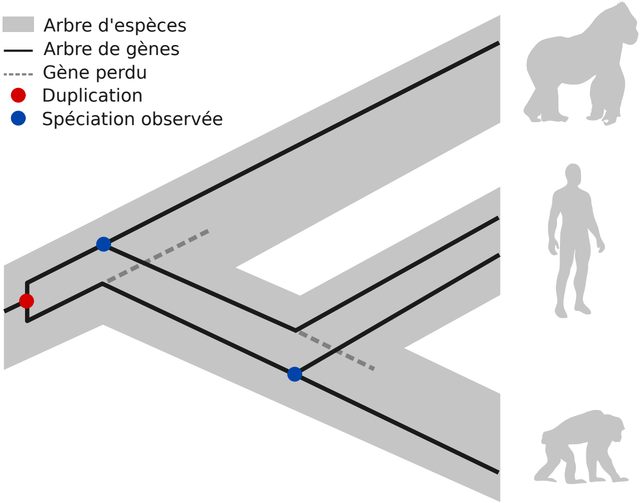
$m1 = (d1 + d2)/2$

$m2 = (2*(m1+d3) + d4)/3$

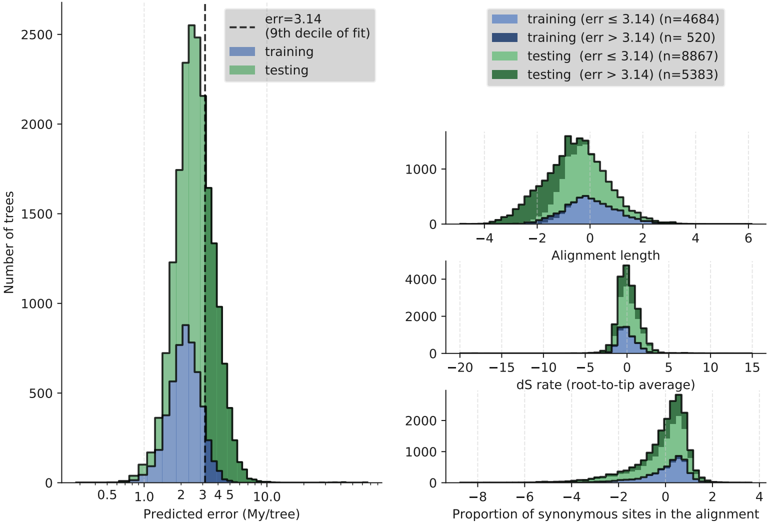
$m3 = (3*(m2+d5) + d6)/4$

$t = T \times m3/(m3+d7)$

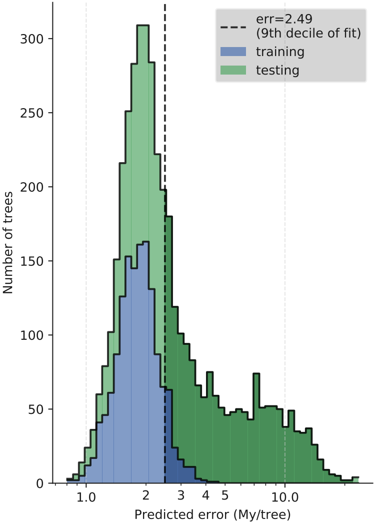
MPL with internal calibrations

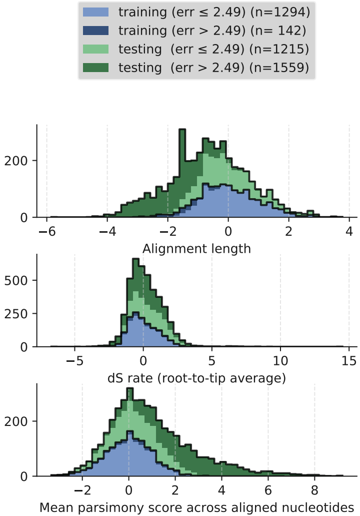


Prediction: root-to-tip approx, keep unwanted

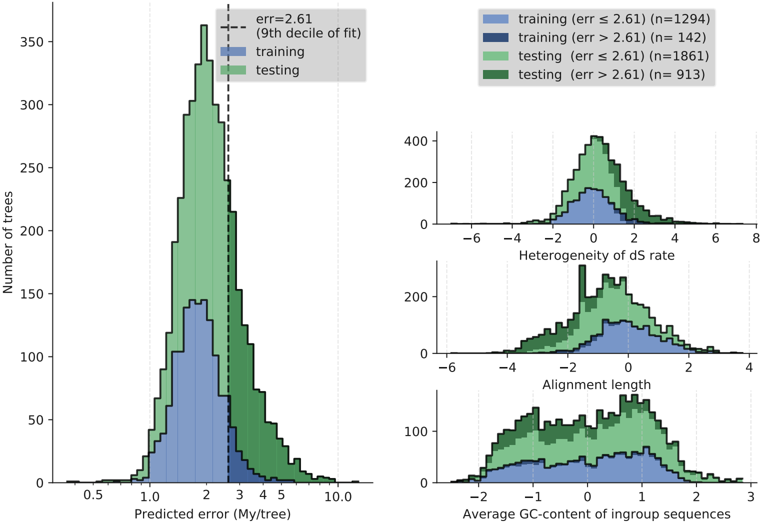


Prediction: root-to-tip approx

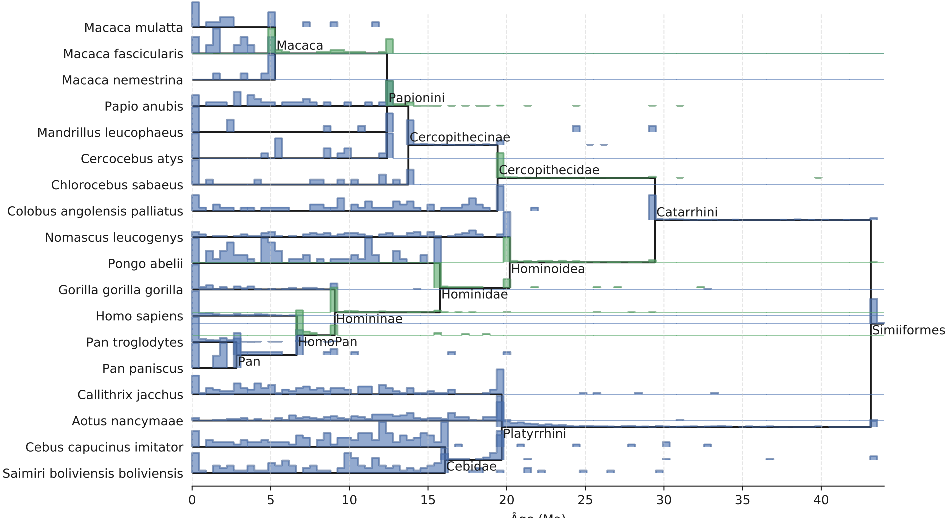




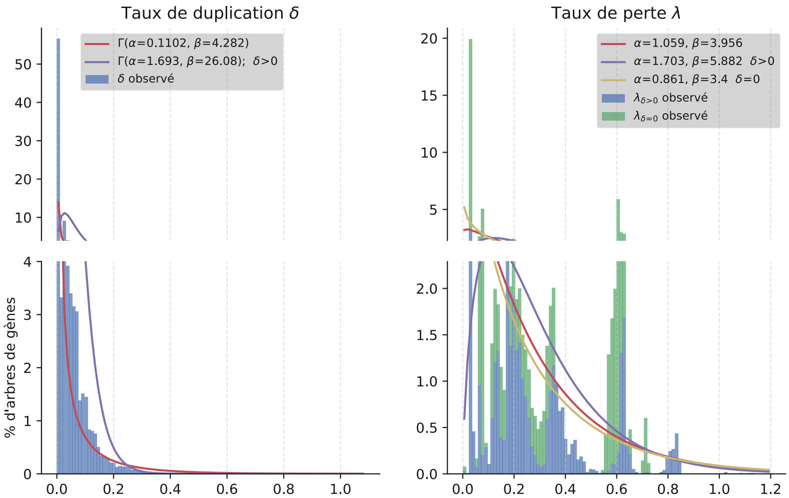
Prediction: spe-to-spe approx



Dated duplications

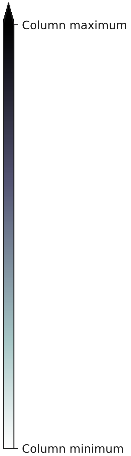


Duplication rates by family



Distribution of the family-wise duplication rate

	δ	$\delta > 0$	λ	$\lambda_{\delta>0}$	$\lambda_{\delta=0}$
Beta prime	0.1859	0.0533	0.5878	1.5192	0.8092
Chi	0.1832	0.1756	0.5753	1.4657	0.7949
Chi²	0.1859	0.0842	0.5877	1.5192	0.8092
Exponentielle	0.3270	0.0853	0.5996	1.3368	0.8027
Exponentielle puissance	0.2046	0.1684	0.5708	1.4015	0.7917
Log-logistique (Fisk)	0.2126	0.0527	0.6413	1.6800	0.8609
Cauchy repliée	0.2759	0.0993	0.6636	1.6801	0.8835
Normale repliée	0.7073	0.1748	0.5421	1.3965	0.8127
Gompertz	0.3270	0.0897	0.5582	1.3834	0.7979
Gamma	0.1860	0.0842	0.5877	1.5192	0.8092
Gamma généralisée	0.1902	0.0609	0.6181	1.4959	inf
Gamma inverse	0.2332	0.2538	1.3521	1.6917	1.3176
Gaussienne inverse	0.1178	0.1720	2.2396	2.3991	2.0919
Pareto	0.4207	0.9280	inf	inf	inf
Pareto généralisée	0.2418	0.0878	0.5532	1.3426	inf
Weibull	0.2109	0.0966	0.5820	1.5007	0.8110
Weibull exponentielle	0.1918	0.0583	0.6149	1.4947	inf
Fréchet	0.2119	0.2396	1.0256	1.3763	1.0619

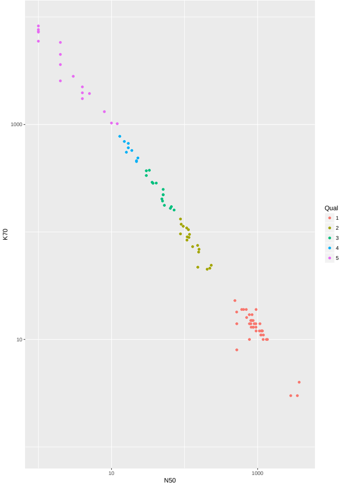


Residual plots of the global regression

Genome assembly quality

N50 *size* of scaffold such that 50% genes are in larger scaffolds

K70 *number* of largest scaffolds containing $>$ 70% genes.



Remove aberrant branch lengths from the forest

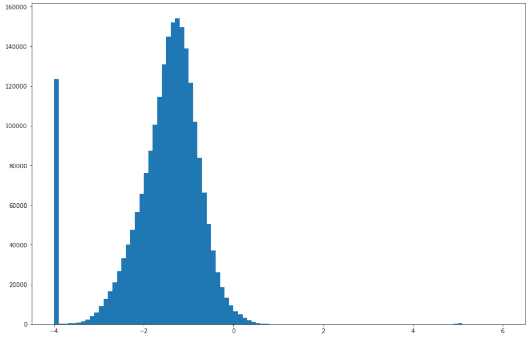


Figure 5: Distribution of log(branch lengths) in *all* gene trees

Remove aberrant branch lengths from the forest

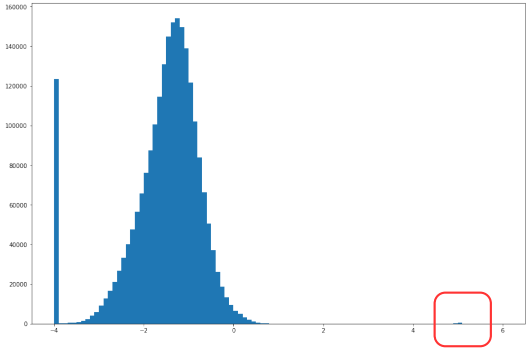


Figure 5: Distribution of log(branch lengths) in *all* gene trees

Remove aberrant branch lengths from the forest

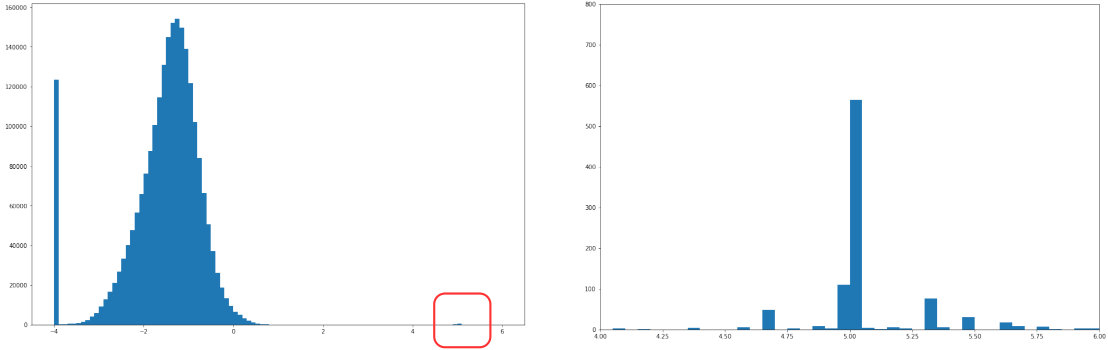
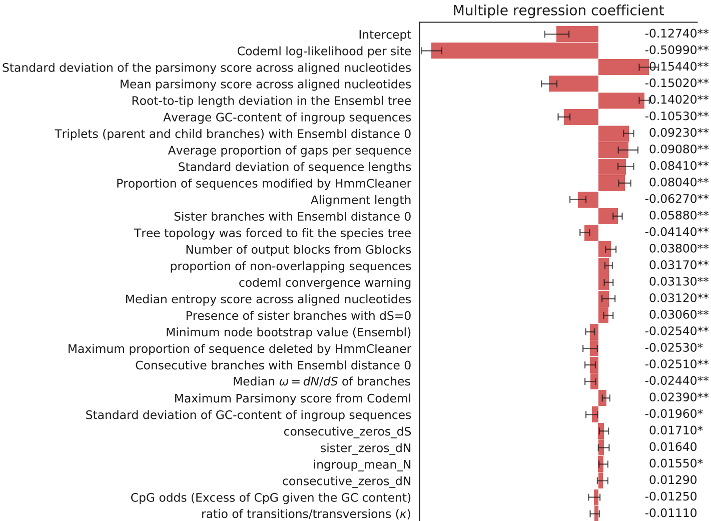
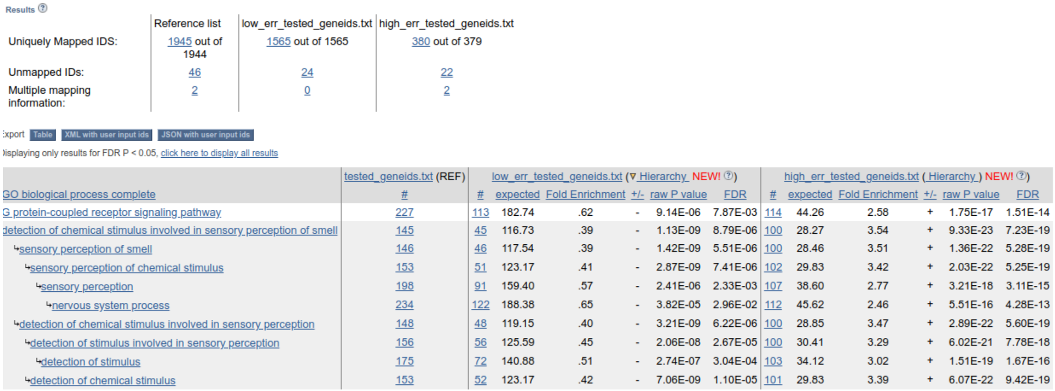


Figure 5: Distribution of log(branch lengths) in *all* gene trees

Features correlated with the duplication rate



GO terms comparison of high error trees VS low error trees



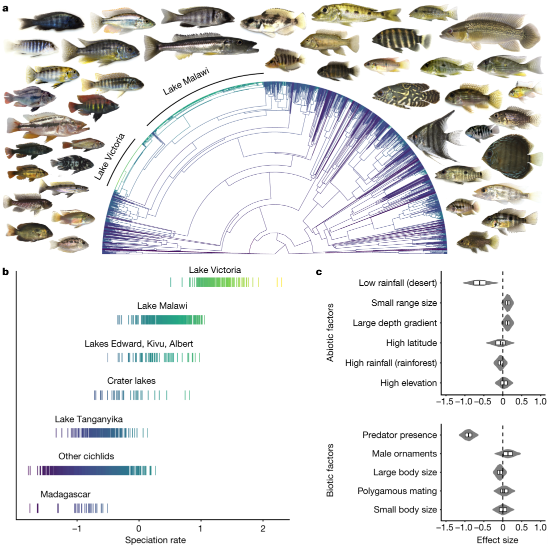
Results ?

	Reference list	low_err_tested_geneids.txt	high_err_tested_geneids.txt
Uniquely Mapped IDS:	1945 out of 1944	1565 out of 1565	380 out of 379
Unmapped IDs:	46	24	22
Multiple mapping information:	2	0	2

xport Table XML with user input ids JSON with user input ids

isplaying only results for FDR P < 0.05, click here to display all results

	tested_geneids.txt (REF)	low_err_tested_geneids.txt (▽ Hierarchy NEW! ?)						high_err_tested_geneids.txt (Hierarchy) NEW! ?)					
GO biological process complete	#	#	expected	Fold Enrichment	+/-	raw P value	FDR	#	expected	Fold Enrichment	+/-	raw P value	FDR
G protein-coupled receptor signaling pathway	227	113	182.74	.62	-	9.14E-06	7.87E-03	114	44.26	2.58	+	1.75E-17	1.51E-14
detection of chemical stimulus involved in sensory perception of smell	145	45	116.73	.39	-	1.13E-09	8.79E-06	100	28.27	3.54	+	9.33E-23	7.23E-19
↳sensory perception of smell	146	46	117.54	.39	-	1.42E-09	5.51E-06	100	28.46	3.51	+	1.36E-22	5.28E-19
↳sensory perception of chemical stimulus	153	51	123.17	.41	-	2.87E-09	7.41E-06	102	29.83	3.42	+	2.03E-22	5.25E-19
↳sensory perception	198	91	159.40	.57	-	2.41E-06	2.33E-03	107	38.60	2.77	+	3.21E-18	3.11E-15
↳nervous system process	234	122	188.38	.65	-	3.82E-05	2.96E-02	112	45.62	2.46	+	5.51E-16	4.28E-13
↳detection of chemical stimulus involved in sensory perception	148	48	119.15	.40	-	3.21E-09	6.22E-06	100	28.85	3.47	+	2.89E-22	5.60E-19
↳detection of stimulus involved in sensory perception	156	56	125.59	.45	-	2.06E-08	2.67E-05	100	30.41	3.29	+	6.02E-21	7.78E-18
↳detection of stimulus	175	72	140.88	.51	-	2.74E-07	3.04E-04	103	34.12	3.02	+	1.51E-19	1.67E-16
↳detection of chemical stimulus	153	52	123.17	.42	-	7.06E-09	1.10E-05	101	29.83	3.39	+	6.07E-22	9.42E-19



McGee et al, 2020