Phylogenomics and comparative genomics: The case of coevolution between the swallowtail butterflies (Papilionidae) and their host plants

GAS seminar
November 30th, 2021

Rémi Allio

Summary

PART 1 - Phylogenomics and comparative genomics in myrmecophagous mammals

PART 2 - The case of coevolution between the swallowtail butterflies (Papilionidae) and their host plants

Summary

PART 1 - Phylogenomics and comparative genomics in myrmecophagous mammals

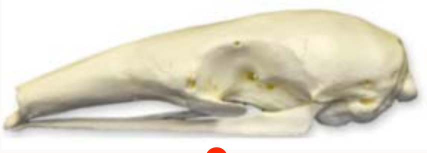
1 − Genome wide approach

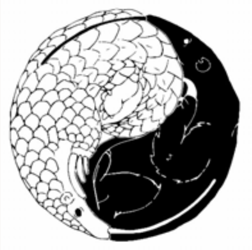
2 − Candidate gene family: Chitinases

Understanding the complex interplay between morphology, genome, and microbiome

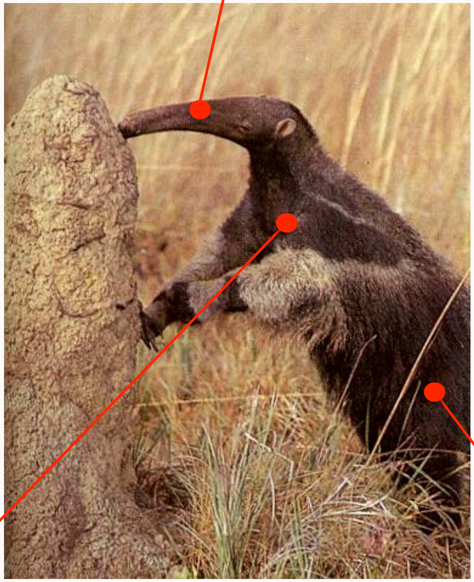
Consolidator Grant
2016-2022

Morphology



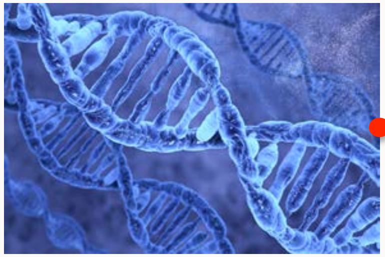


Development



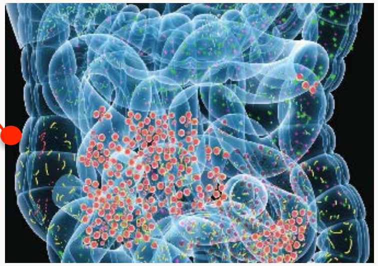
Gut anatomy

Genome

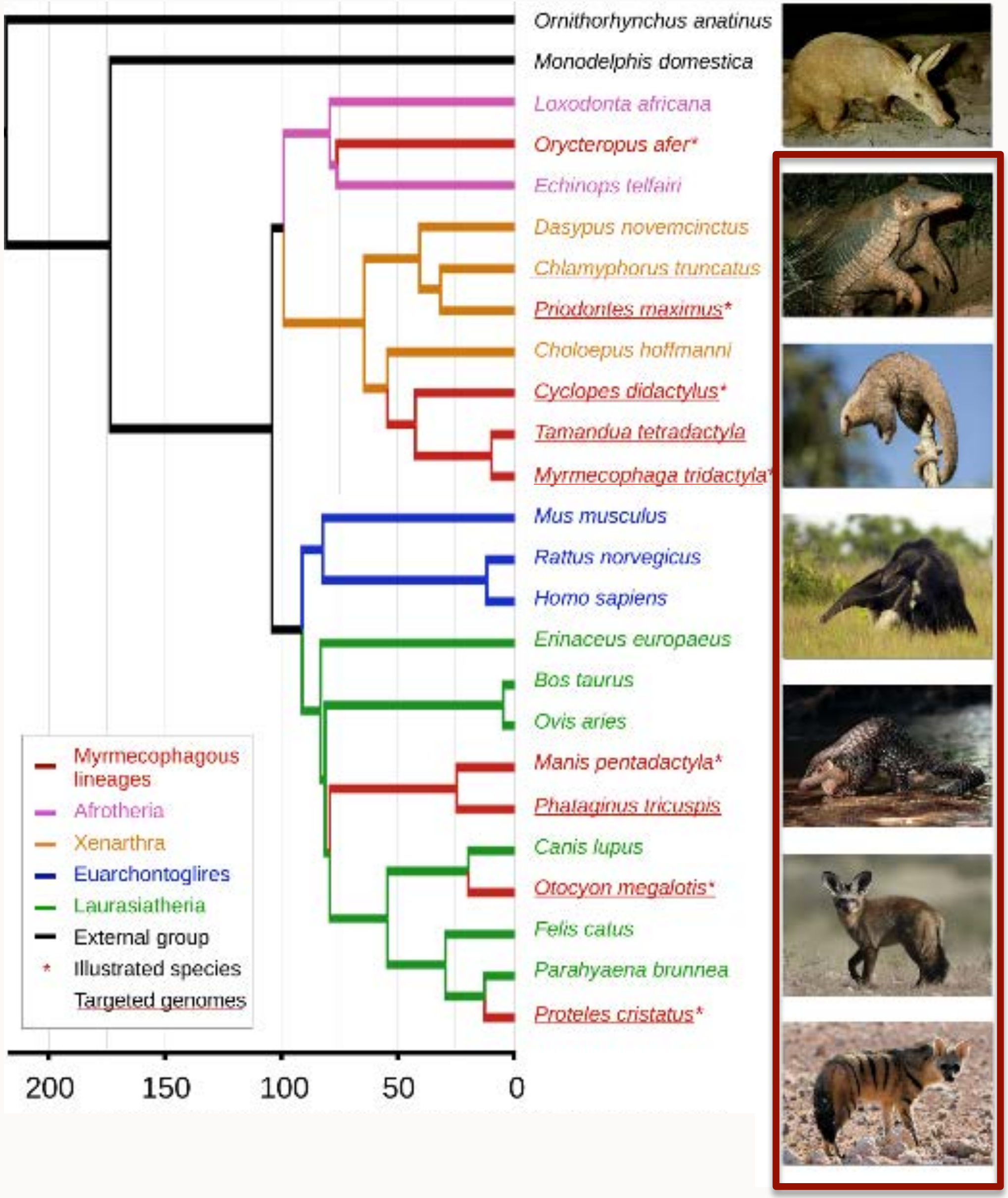


Host immunity

Microbiome



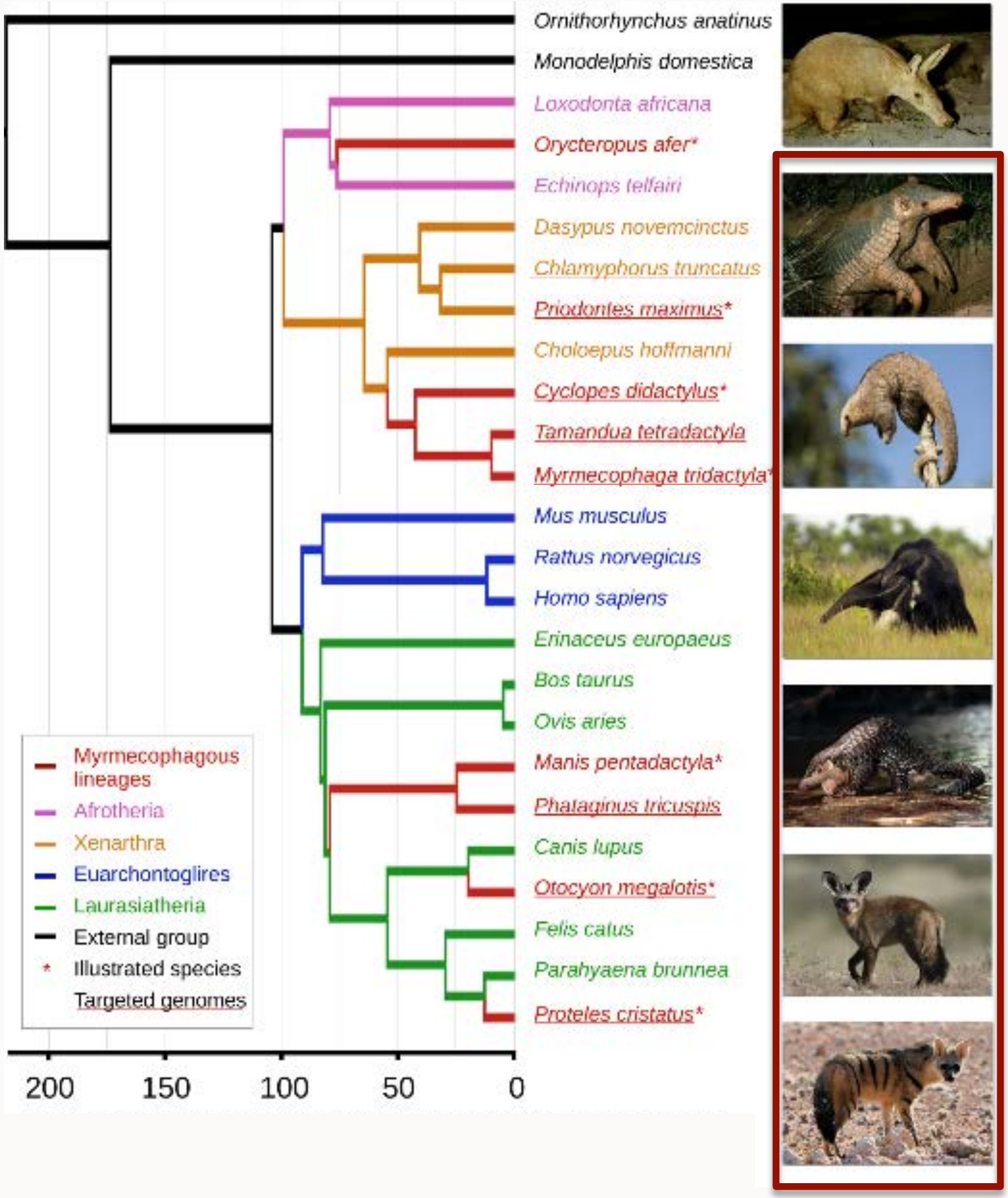
Genome assemblies



9 genomes

Genome size 2-5 Gb

Genome assemblies

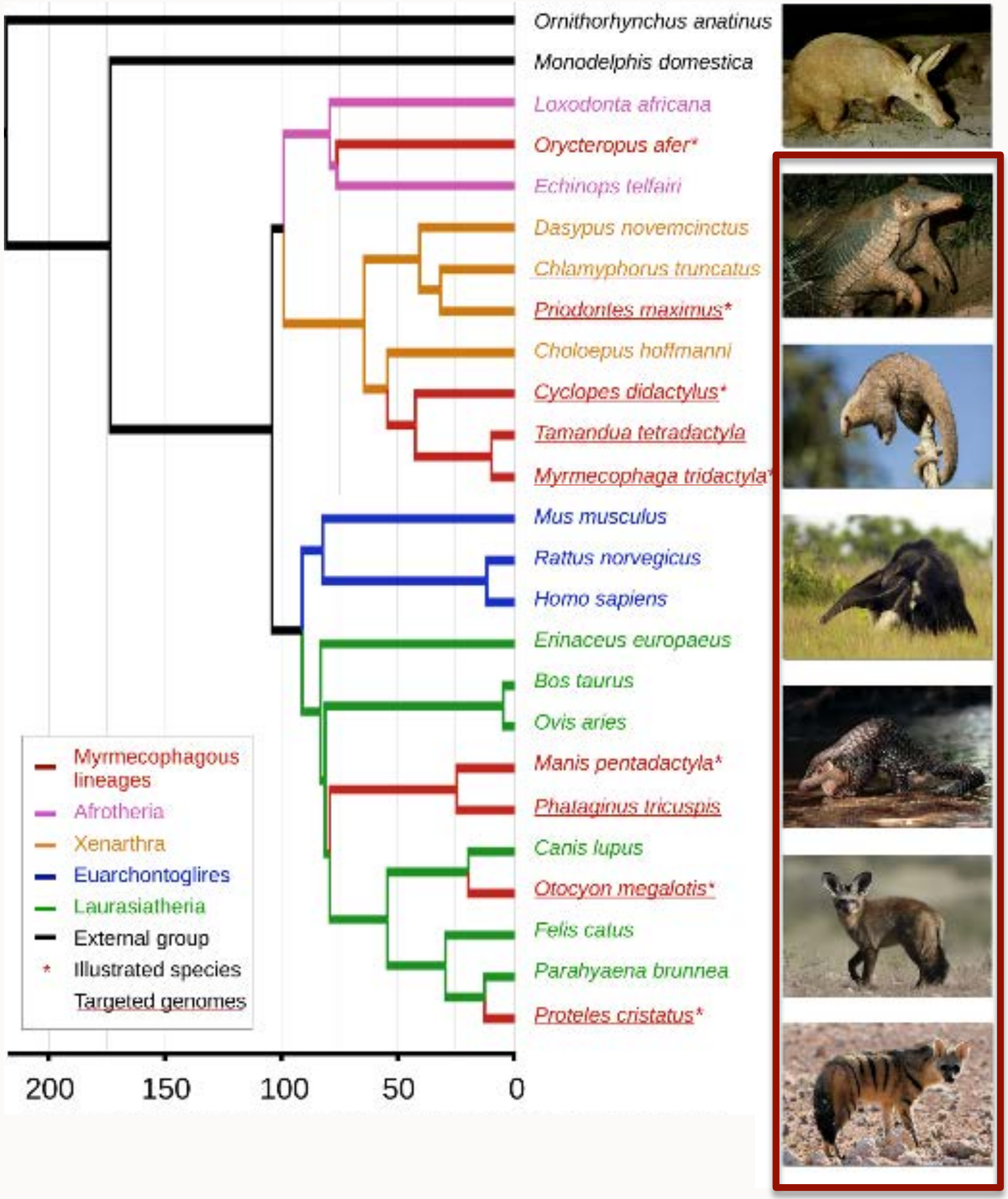


9 genomes
Genome size 2-5 Gb

Short reads sequencing
(Illumina)
+
Long reads sequencing
(PacBio)

Hybrid Assembly

Genome assemblies



9 genomes
Genome size 2-5 Gb

↓

Short reads sequencing
(Illumina)
+
Long reads sequencing
~~(PacBio)~~
(Minion)

↓

Hybrid Assembly

Genome from road kills

Marie-Ka
Tilak

Roadkill samples

Sampling

DNA extraction

Library construction

Genome from road kills

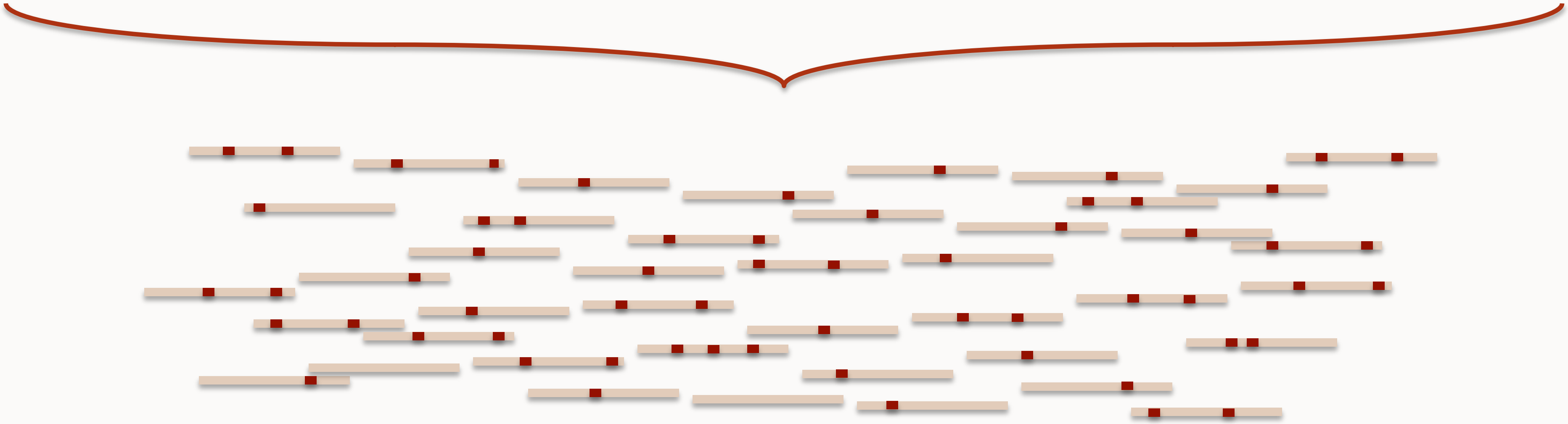
Marie-Ka
Tilak

Roadkill samples

Sampling

DNA extraction

Library construction



Many **short** reads with **many errors**…

Genome from road kills

An optimized protocol

Marie-Ka
Tilak

Roadkill samples

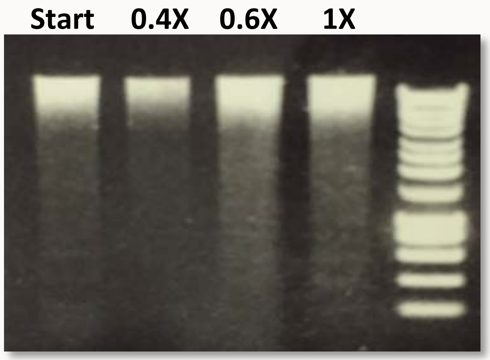
Sampling

DNA extraction

Library construction

Two keys

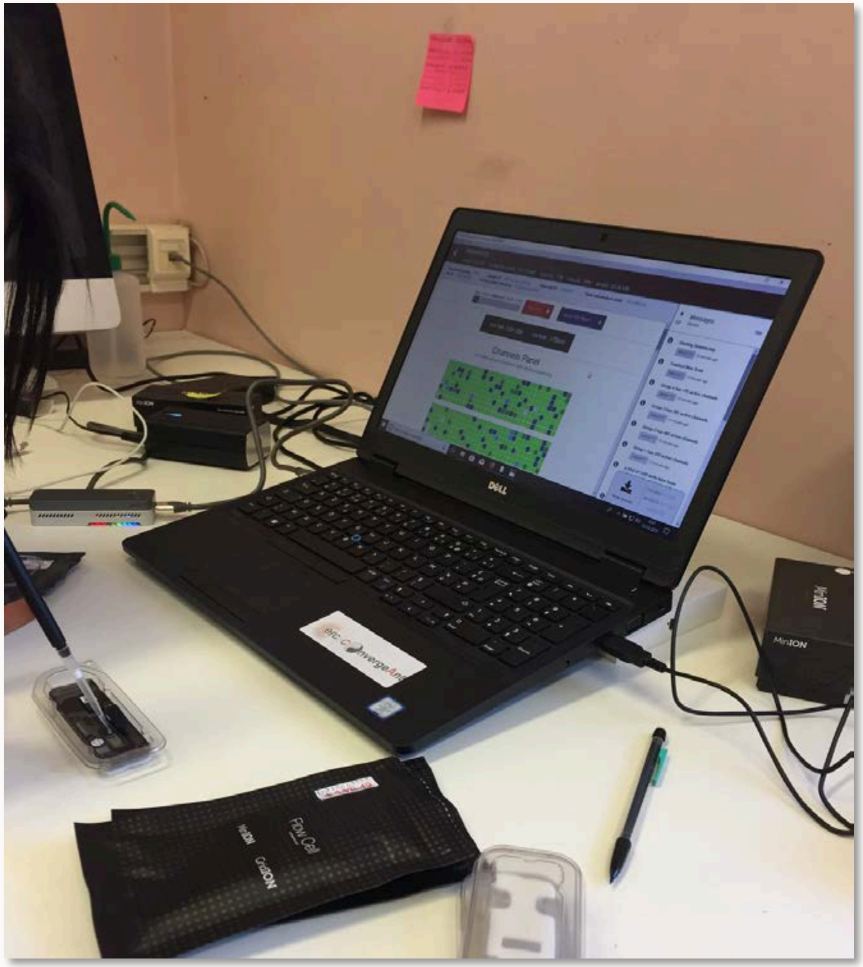
Tissues cleaning
DNA extraction check



DNA size selection
AMPure beads

It works!

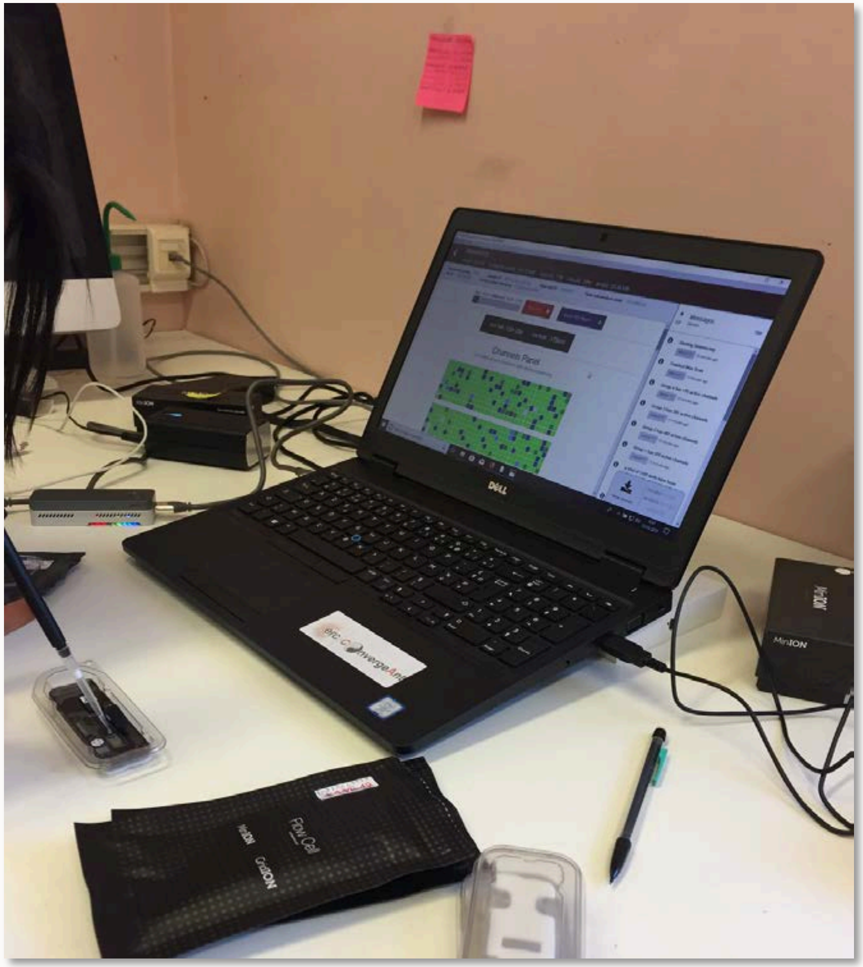
Marie-Ka
Tilak



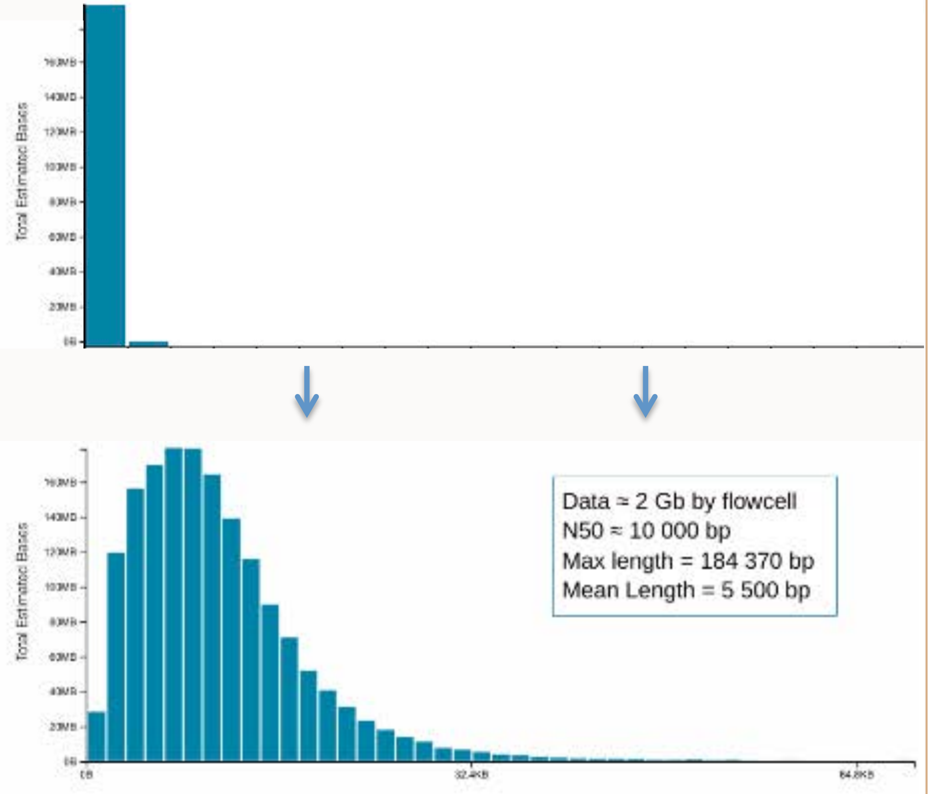
Read length distribution

It works!

Marie-Ka
Tilak



Read length distribution



Data ≈ 2 Gb by flowcell
N50 ≈ 10 000 bp
Max length = 184 370 bp
Mean Length = 5 500 bp

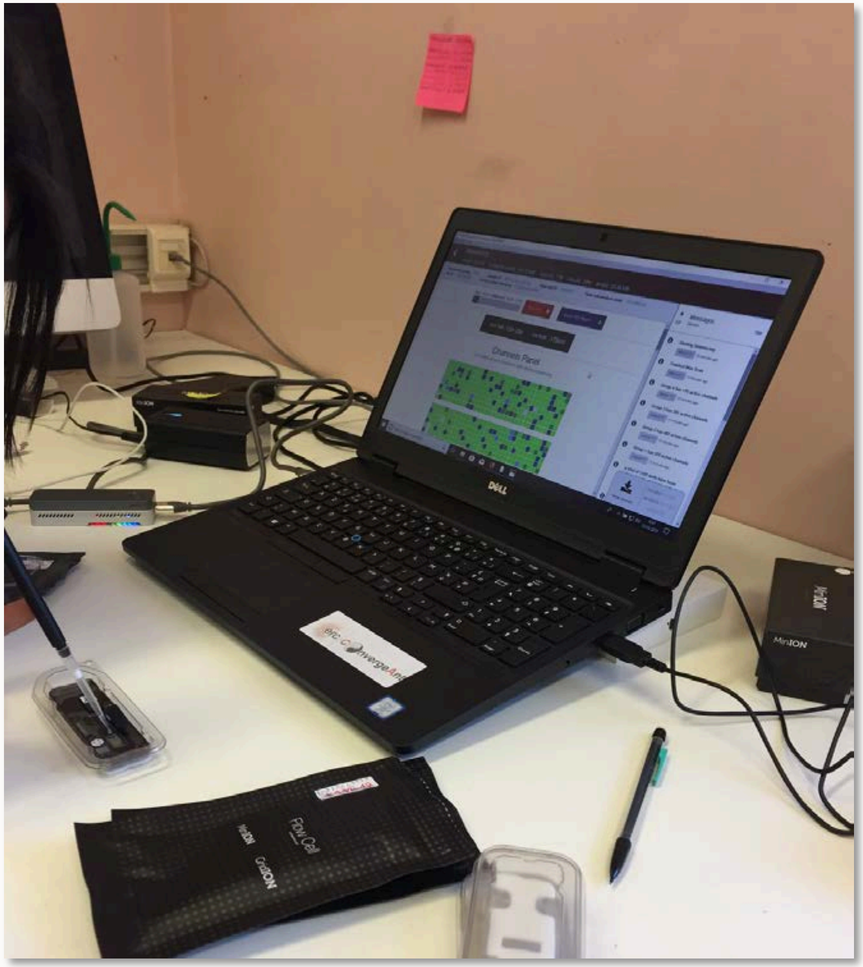
An optimized protocol for sequencing mammalian roadkill tissues with Oxford Nanopore Technology (ONT) V.2

Marie-Ka Tilak[1], Rémi Allio[1], Frédéric Delsuc[1]

[1] Institut des Sciences de l'Evolution de Montpellier (ISEM), CNRS, IRD, EPHE, Université de Montpellier, Montpellier, France

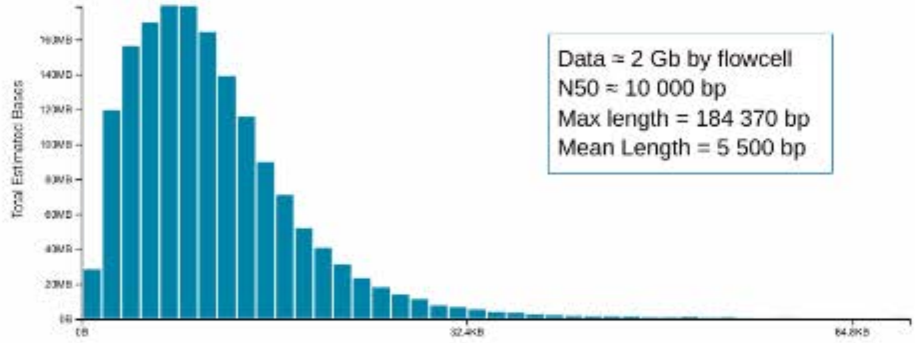
It works!

Marie-Ka
Tilak



Aardwolf (*Proteles cristatus*)
16 flow cells
27.5 Gb (~11x)

Scaffolds: 5,669
N50: 1.3 Mb



Data ≈ 2 Gb by flowcell
N50 ≈ 10 000 bp
Max length = 184 370 bp
Mean Length = 5 500 bp

An optimized protocol for sequencing mammalian roadkill tissues with Oxford Nanopore Technology (ONT) V.2

Marie-Ka Tilak[1], Rémi Allio[1], Frédéric Delsuc[1]

[1] Institut des Sciences de l'Evolution de Montpellier (ISEM), CNRS, IRD, EPHE, Université de Montpellier, Montpellier, France

All genomes

Marie-Ka
Tilak

Amandine
Magdeleine

Southern tamandua (*Tamandua tetradactyla*)
8 flow cells
51.3 Gb (~16x)

Scaffolds: 4,309
N50: 3.1 Mb

Giant anteater (*Myrmecophaga tridactyla*)
16 flow cells
50.3 Gb (~17x)

Scaffolds: 6,911
N50: 1.5 Mb

Silky anteater (*Cyclopes didactylus*)
21 flow cells
49.7 Gb (~15x)

Scaffolds: 14,869
N50: 795 Kb

Giant pangolin (*Smutsia gigantea*)
4 flow cells
29.1 Gb (~13x)

Scaffolds: 24,429
N50: 227 Kb

Pink fairy armadillo (*Chlamyphorus tr.*)
21 flow cells
32.9 Gb (~10x)

Scaffolds: 36,203
N50: 247 Kb

Bat-eared fox (*Otocyon megalotis*)
15 flow cells
33.0 Gb (~13x)

Scaffolds: 11,081
N50: 728 Kb

Giant armadillo (*Priodontes maximus*)
15 flow cells
52.6Gb (~15x)

Scaffolds: 51,157
N50: 185 Kb

Pale-throated sloth (*Bradypus tridactylus*)
15 flow cells
41.7 Gb (~12x)

Scaffolds: 15,237
N50: 684 Kb

All genomes

Marie-Ka
Tilak

Amandine
Magdeleine

Southern tamandua (*Tamandua tetradactyla*)
8 flow cells
51.3 Gb (~16x)

Scaffolds: 4,309
N50: 3.1 Mb

Pink fairy armadillo (*Chlamyphorus tr.*)
21 flow cells
32.9 Gb (~10x)

Scaffolds: 36,203
N50: 247 Kb

Giant anteater (*Myrmecophaga tridactyla*)
16 flow cells
50.3 Gb (~17x)

Scaffolds: 6,911
N50: 1.5 Mb

Bat-eared fox (*Otocyon megalotis*)
15 flow cells
33.0 Gb (~13x)

Scaffolds: 11,081
N50: 728 Kb

Silky anteater (*Cyclopes didactylus*)
21 flow cells
49.7 Gb (~15x)

Scaffolds: 14,869
N50: 795 Kb

Giant armadillo (*Priodontes maximus*)
15 flow cells
52.6Gb (~15x)

Scaffolds: 51,157
N50: 185 Kb

Giant pangolin (*Smutsia gigantea*)
4 flow cells
29.1 Gb (~13x)

Scaffolds: 24,429
N50: 227 Kb

Pale-throated sloth (*Bradypus tridactylus*)
15 flow cells
41.7 Gb (~12x)

Scaffolds: 15,237
N50: 684 Kb

All genomes

Marie-Ka
Tilak

Amandine
Magdeleine

Southern tamandua (*Tamandua tetradactyla*)
8 flow cells
51.3 Gb (~16x)

Scaffolds: 4,309
N50: 3.1 Mb

Giant anteater (*Myrmecophaga tridactyla*)
16 flow cells
50.3 Gb (~17x)

Scaffolds: 6,911
N50: 1.5 Mb

Silky anteater (*Cyclopes didactylus*)
21 flow cells
49.7 Gb (~15x)

Scaffolds: 14,869
N50: 795 Kb

Giant pangolin (*Smutsia gigantea*)
4 flow cells
29.1 Gb (~13x)

Scaffolds: 24,429
N50: 227 Kb

Pink fairy armadillo (*Chlamyphorus tr.*)
21 flow cells
32.9 Gb (~10x)

Scaffolds: 36,203
N50: 247 Kb

Bat-eared fox (*Otocyon megalotis*)
15 flow cells
33.0 Gb (~13x)

Scaffolds: 11,081
N50: 728 Kb

Giant armadillo (*Priodontes maximus*)
15 flow cells
52.6Gb (~15x)

Scaffolds: 51,157
N50: 185 Kb

Pale-throated sloth (*Bradypus tridactylus*)
15 flow cells
41.7 Gb (~12x)

Scaffolds: 15,237
N50: 684 Kb

All genomes

Marie-Ka
Tilak

Amandine
Magdeleine

Southern tamandua (*Tamandua tetradactyla*)
8 flow cells
51.3 Gb (~16x)

Scaffolds: 4,309
N50: 3.1 Mb

Giant anteater (*Myrmecophaga tridactyla*)
16 flow cells
50.3 Gb (~17x)

Scaffolds: 6,911
N50: 1.5 Mb

Silky anteater (*Cyclopes didactylus*)
21 flow cells
49.7 Gb (~15x)

Scaffolds: 14,869
N50: 795 Kb

Giant pangolin (*Smutsia gigantea*)
4 flow cells
29.1 Gb (~13x)

Scaffolds: 24,429
N50: 227 Kb

Pink fairy armadillo (*Chlamyphorus tr.*)
21 flow cells
32.9 Gb (~10x)

Scaffolds: 36,203
N50: 247 Kb

Bat-eared fox (*Otocyon megalotis*)
15 flow cells
33.0 Gb (~13x)

Scaffolds: 11,081
N50: 728 Kb

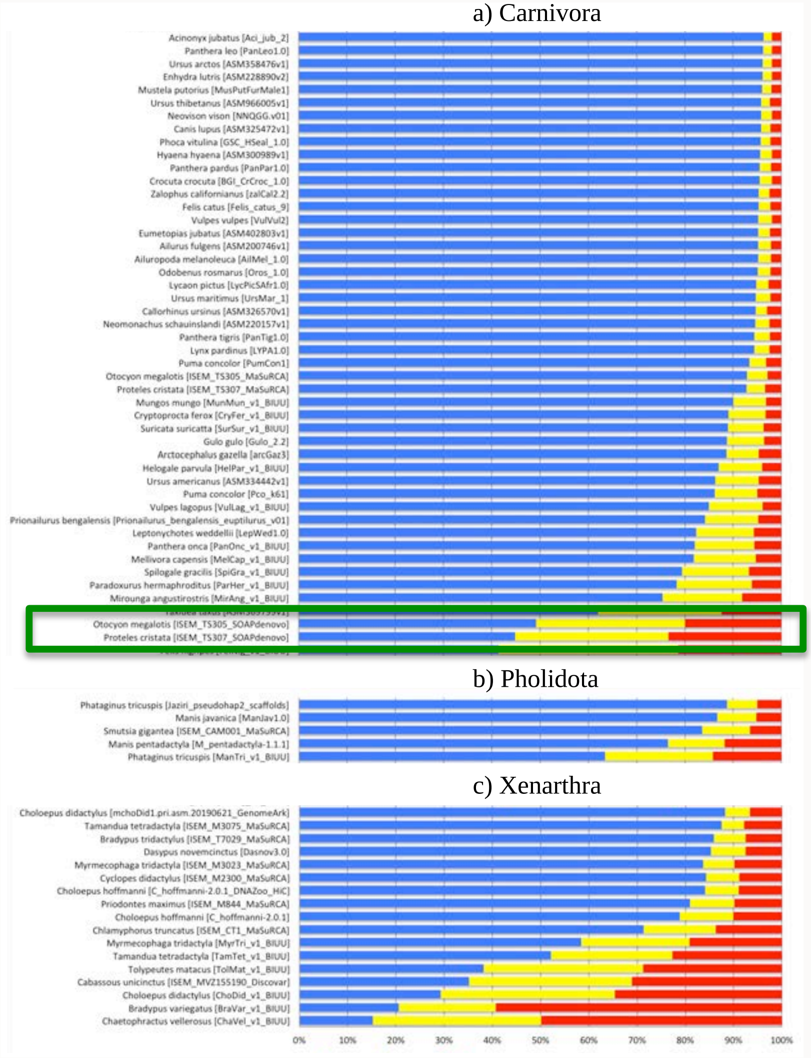
Giant armadillo (*Priodontes maximus*)
15 flow cells
52.6Gb (~15x)

Scaffolds: 51,157
N50: 185 Kb

Pale-throated sloth (*Bradypus tridactylus*)
15 flow cells
41.7 Gb (~12x)

Scaffolds: 15,237
N50: 684 Kb

BUSCO scores



🟦 Complete genes 🟨 Fragmented genes 🟥 Missing genes

BUSCO scores

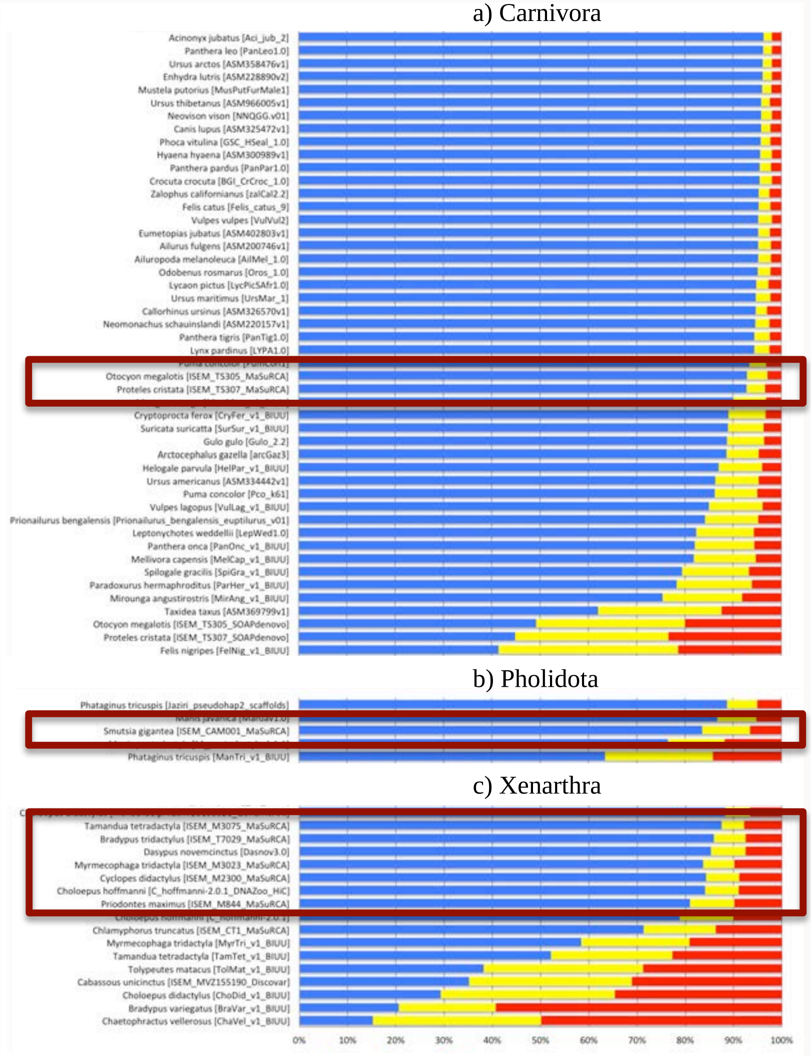
a) Carnivora

b) Pholidota

c) Xenarthra

🟦 Complete genes 🟨 Fragmented genes 🟥 Missing genes

BUSCO scores


a) Carnivora

b) Pholidota

c) Xenarthra

🟦 Complete genes 🟨 Fragmented genes 🟥 Missing genes

What are these genomes for?

What are these genomes for?

1- Species delineation analyses

What are these genomes for?

1- Species delineation analyses

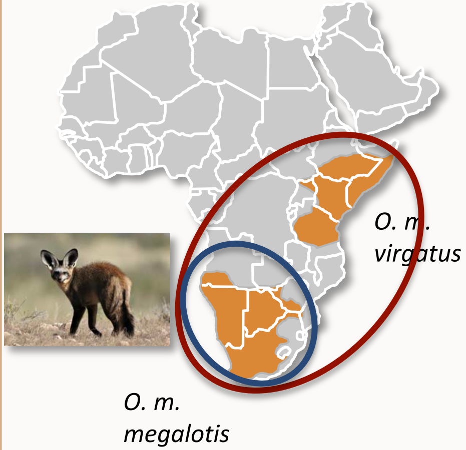
In Carnivora

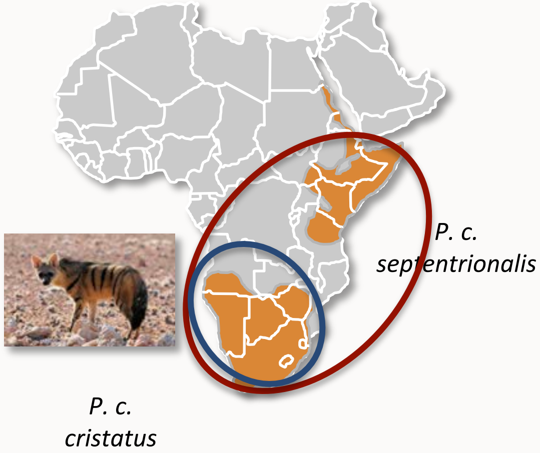
Koehler & Richardson (1990) *Mam. Spe.*; Clark (2005) *Mam. Spe.*

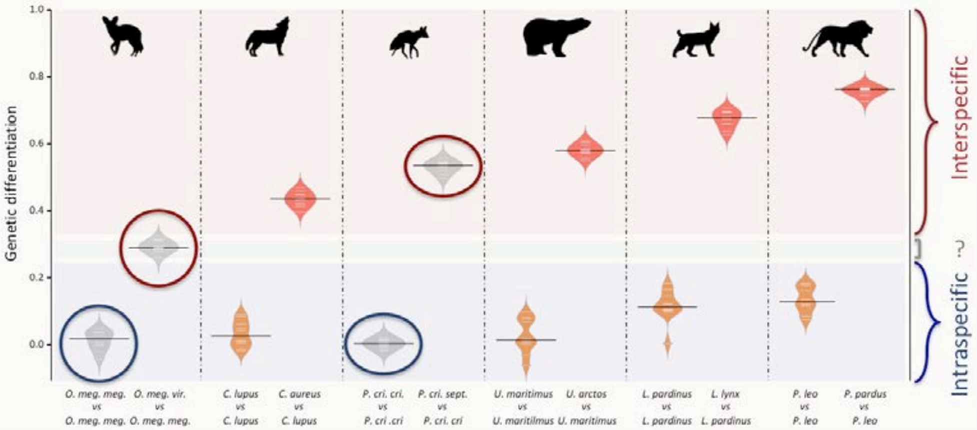
What are these genomes for?

1- Species delineation analyses

In Carnivora





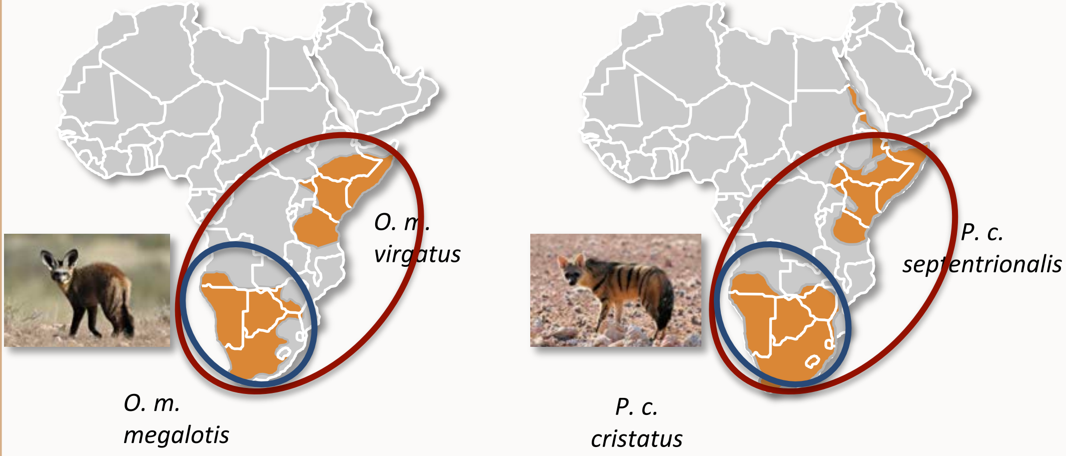


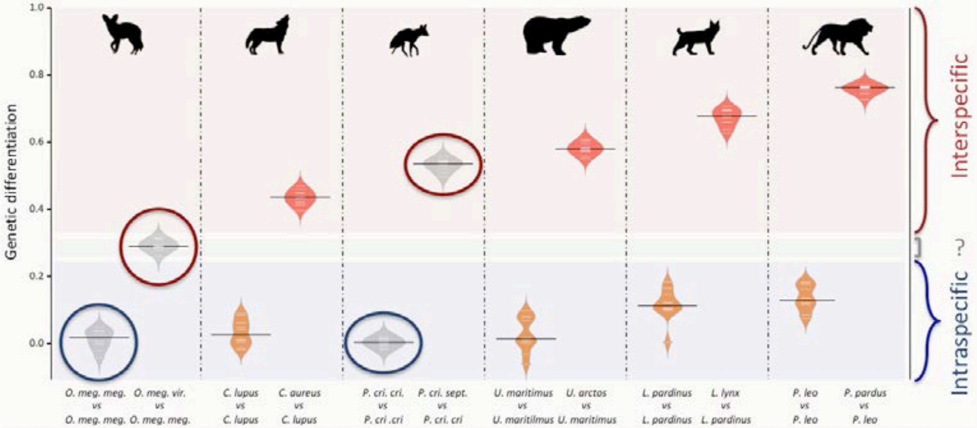
Allio et al. 2021 *eLife*

What are these genomes for?

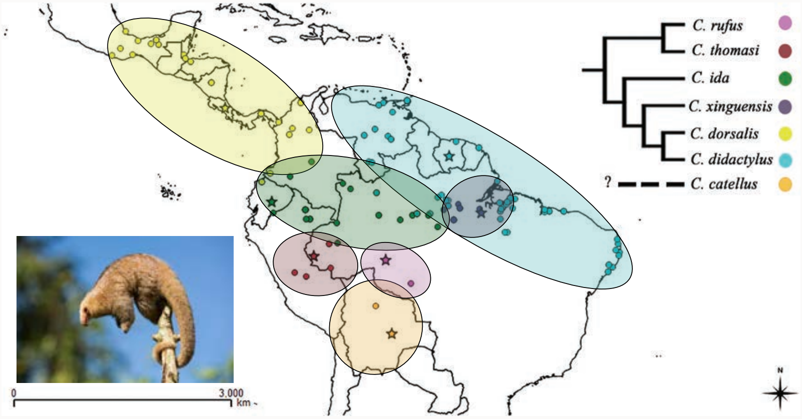
1- Species delineation analyses

In Carnivora





In Xenarthra



In progress!

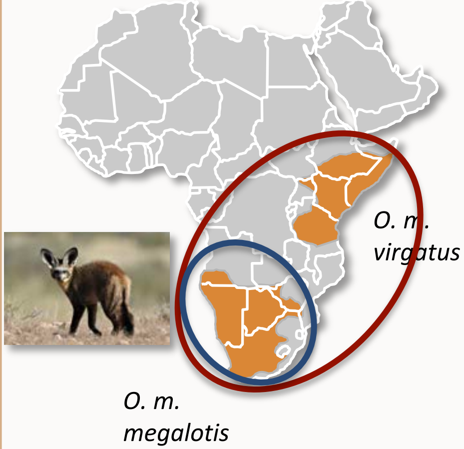
Mathilde Barthe

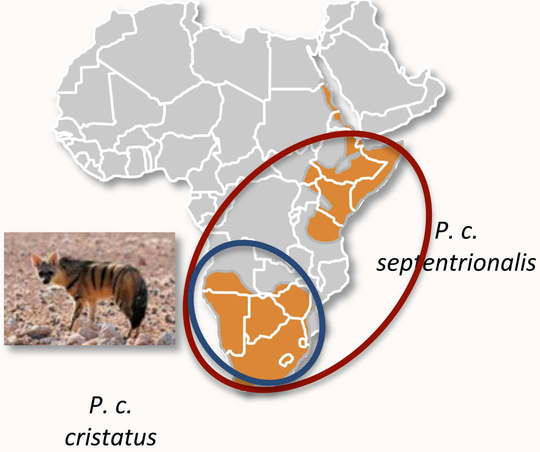
Allio et al. 2021 *eLife; Miranda et al. (2017) Zool. J. Linn. Soc.*

What are these genomes for?

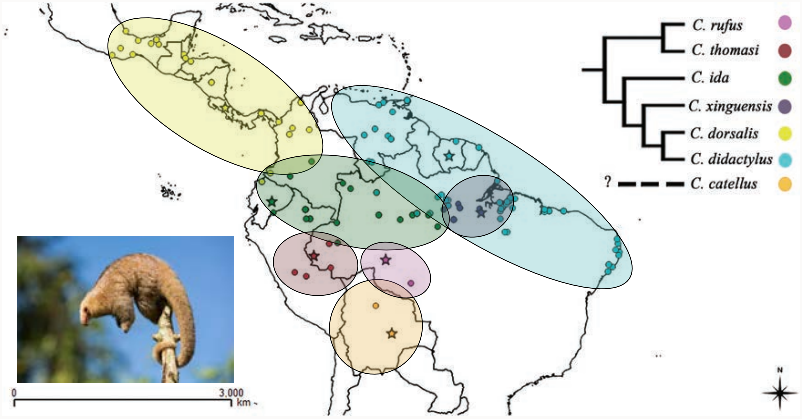
1- Species delineation analyses

In Carnivora





In Xenarthra

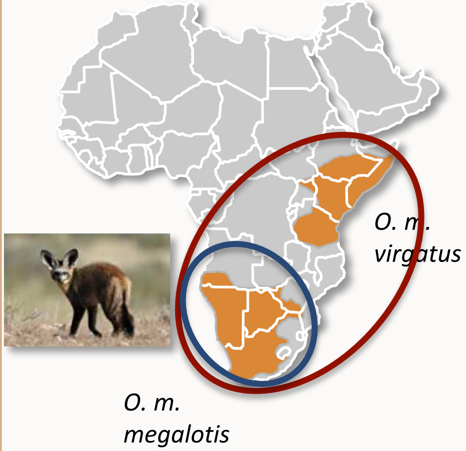


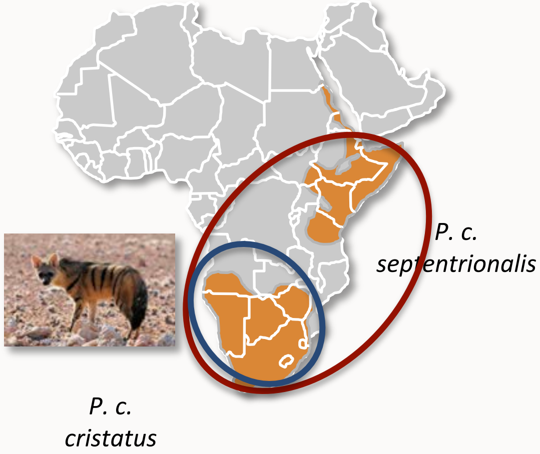
2- Genome-wide signatures of convergent evolution?

What are these genomes for?

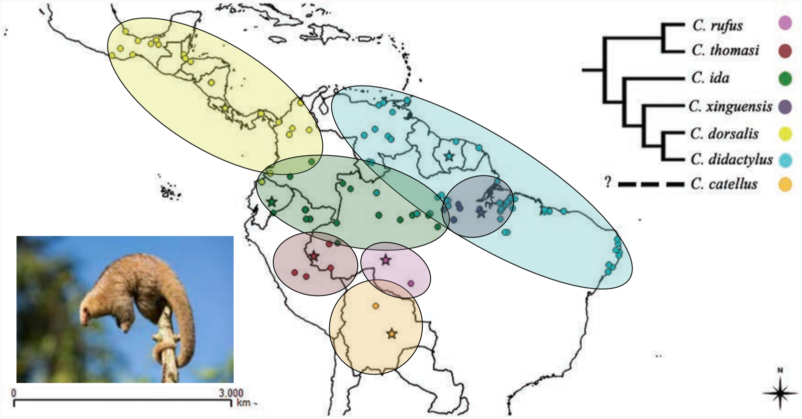
1- Species delineation analyses

In Carnivora



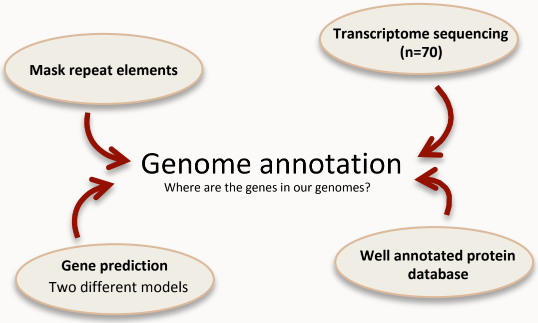


In Xenarthra



2- Genome-wide signatures of convergent evolution?

Genome annotation is done
Dataset construction in progress…



Mask repeat elements

Transcriptome sequencing (n=70)

Genome annotation
Where are the genes in our genomes?

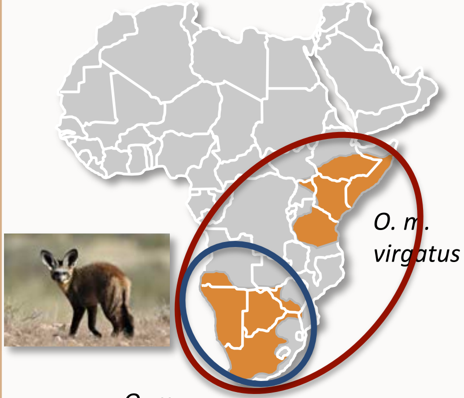
Gene prediction
Two different models

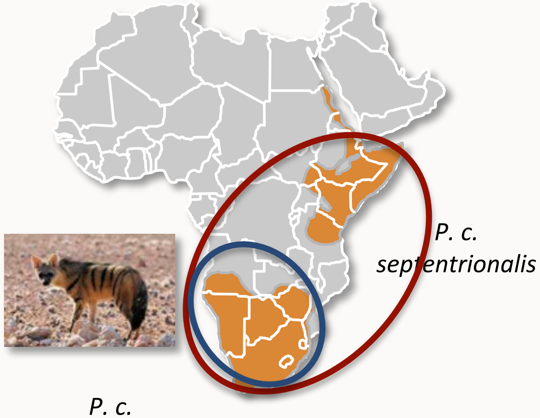
Well annotated protein database

What are these genomes for?

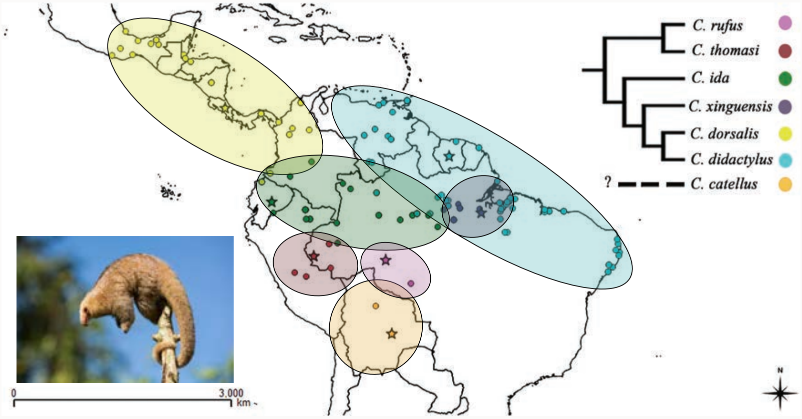
1- Species delineation analyses

In Carnivora





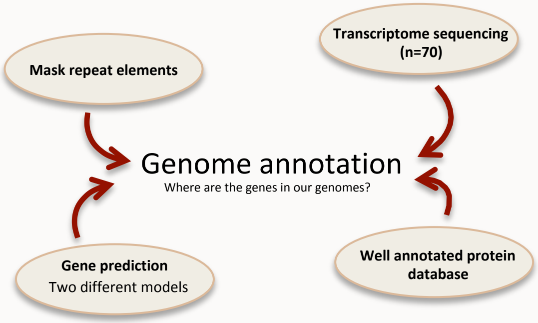
In Xenarthra



2- Genome-wide signatures of convergent evolution?

Genome annotation is done
Dataset construction in progress…

└→ **Candidate gene family: Chitinases**


Mask repeat elements

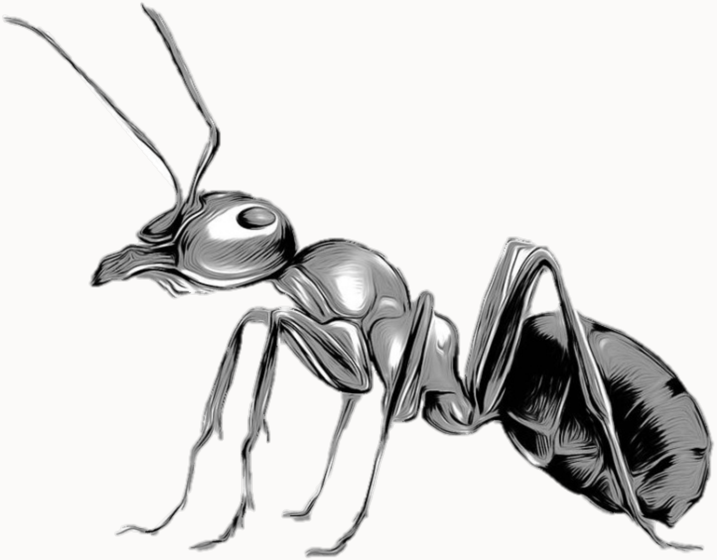
Transcriptome sequencing (n=70)

Genome annotation
Where are the genes in our genomes?

Gene prediction
Two different models

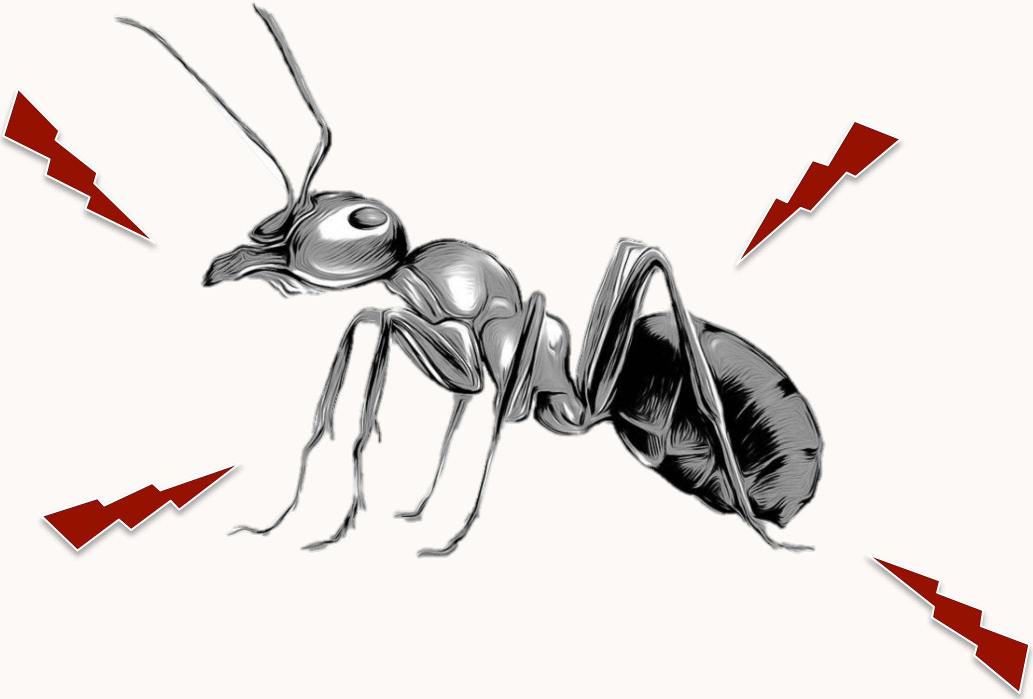
Well annotated protein database

Why the chitinase gene family?



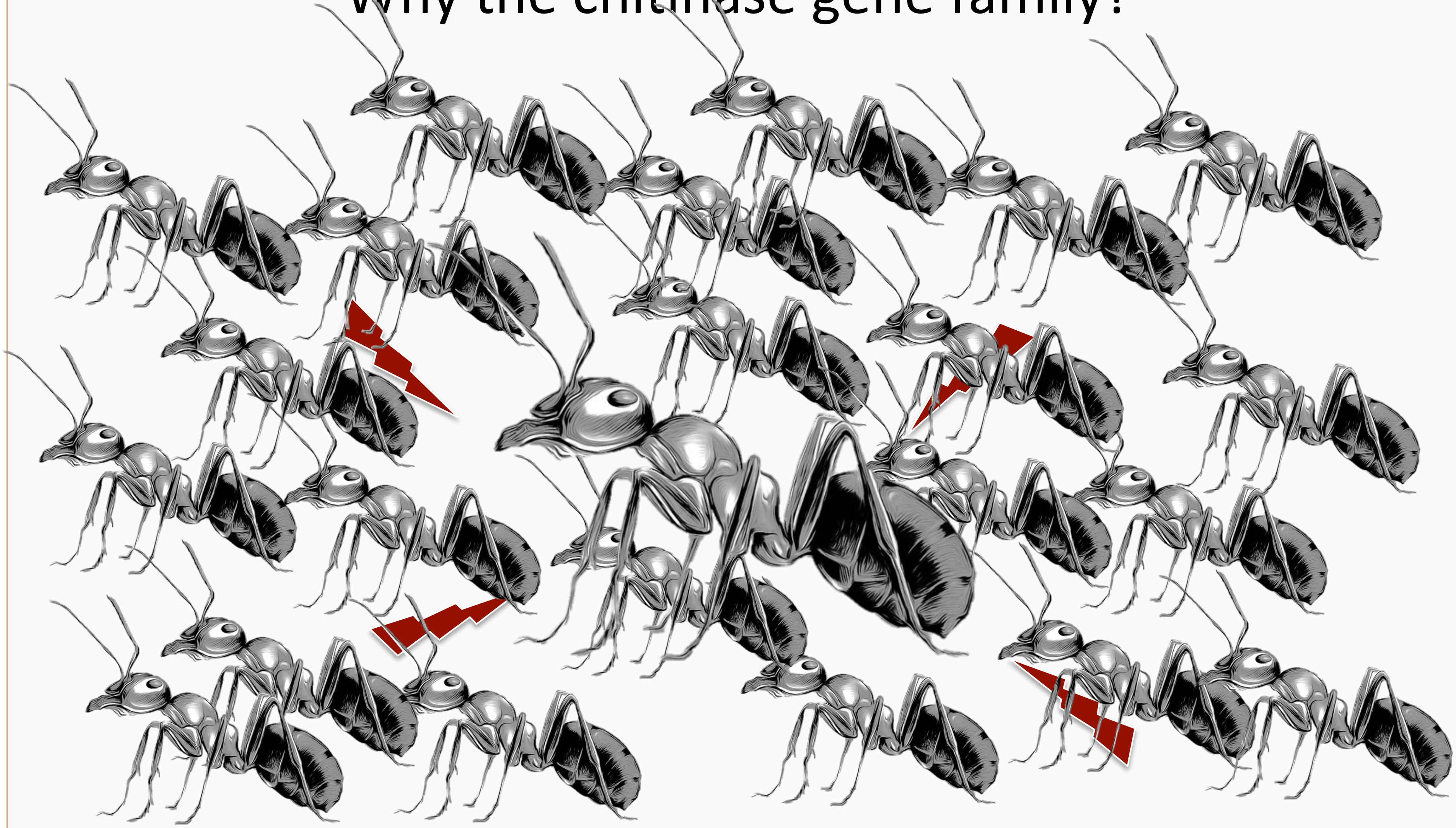
Chitin = insects’ skeleton

Why the chitinase gene family?



Chitin = insects’ skeleton

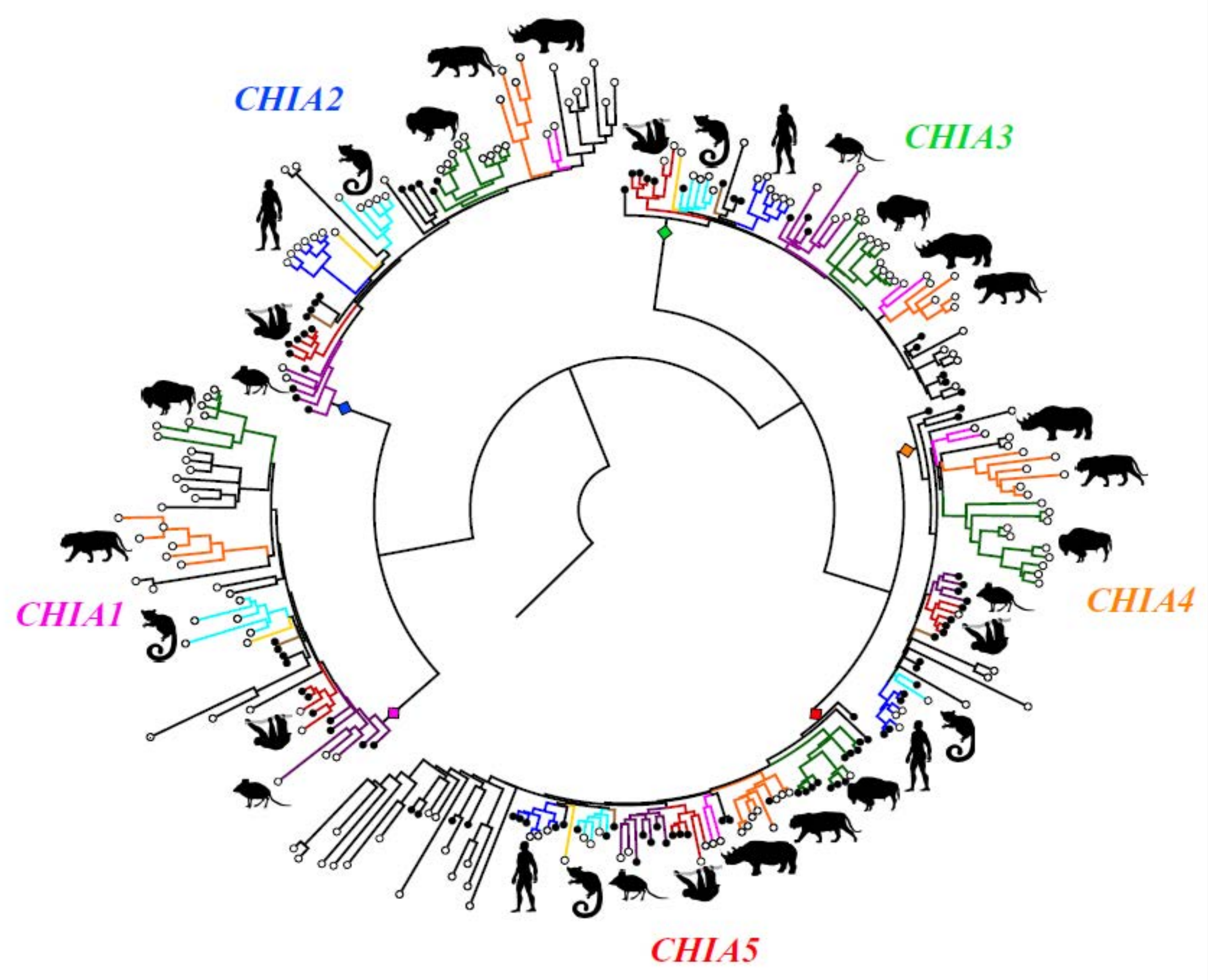
Why the chitinase gene family?



Chitin = insects’ skeleton

The chitinase family

0-5 functional genes



Emerling *et al.* (2018) Science advances

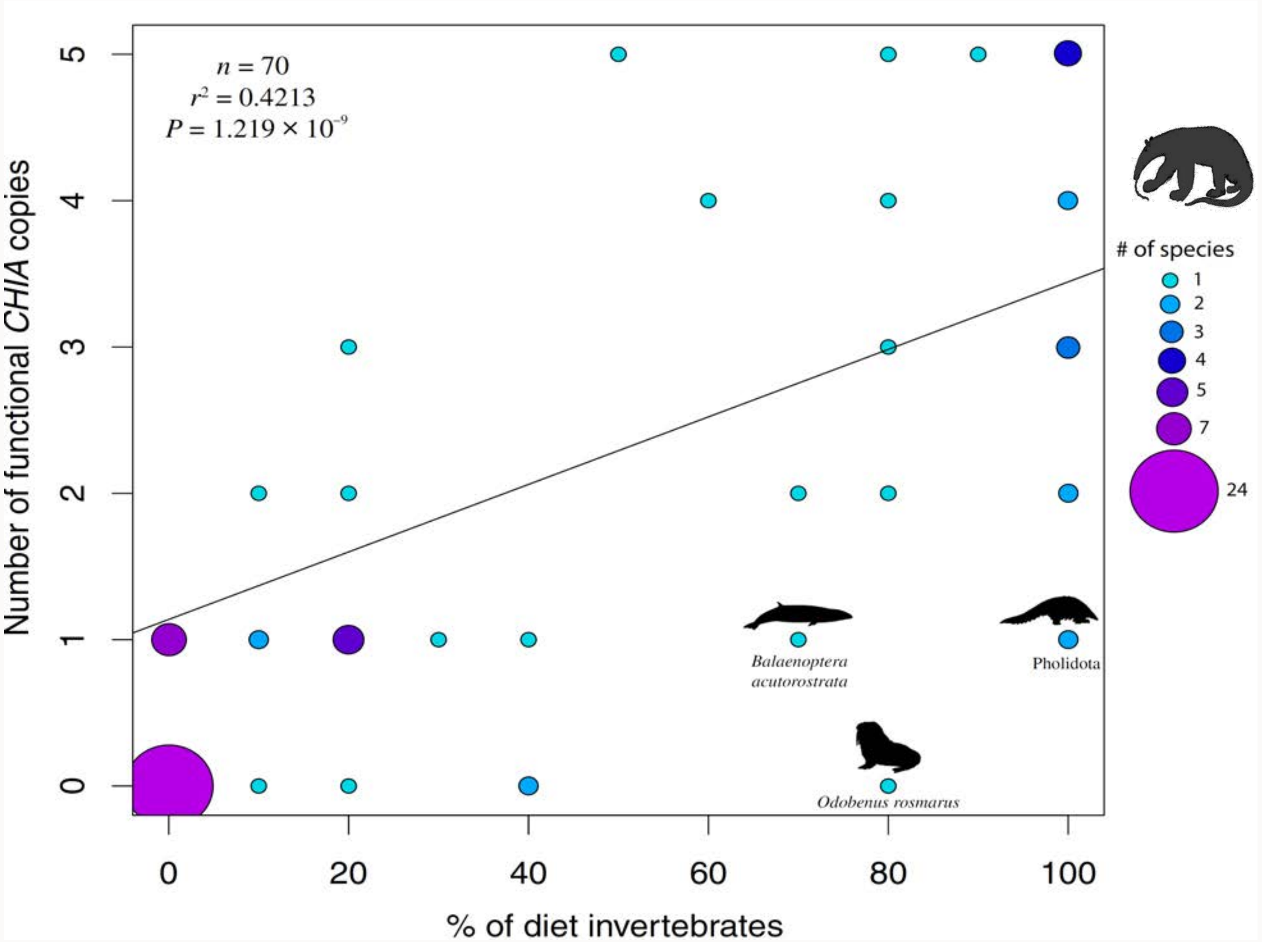
The chitinase family

0-5 functional genes

↗ % of diet invertebrates

=

↗ number of functional genes



Emerling *et al.* (2018) Science advances

The chitinase family

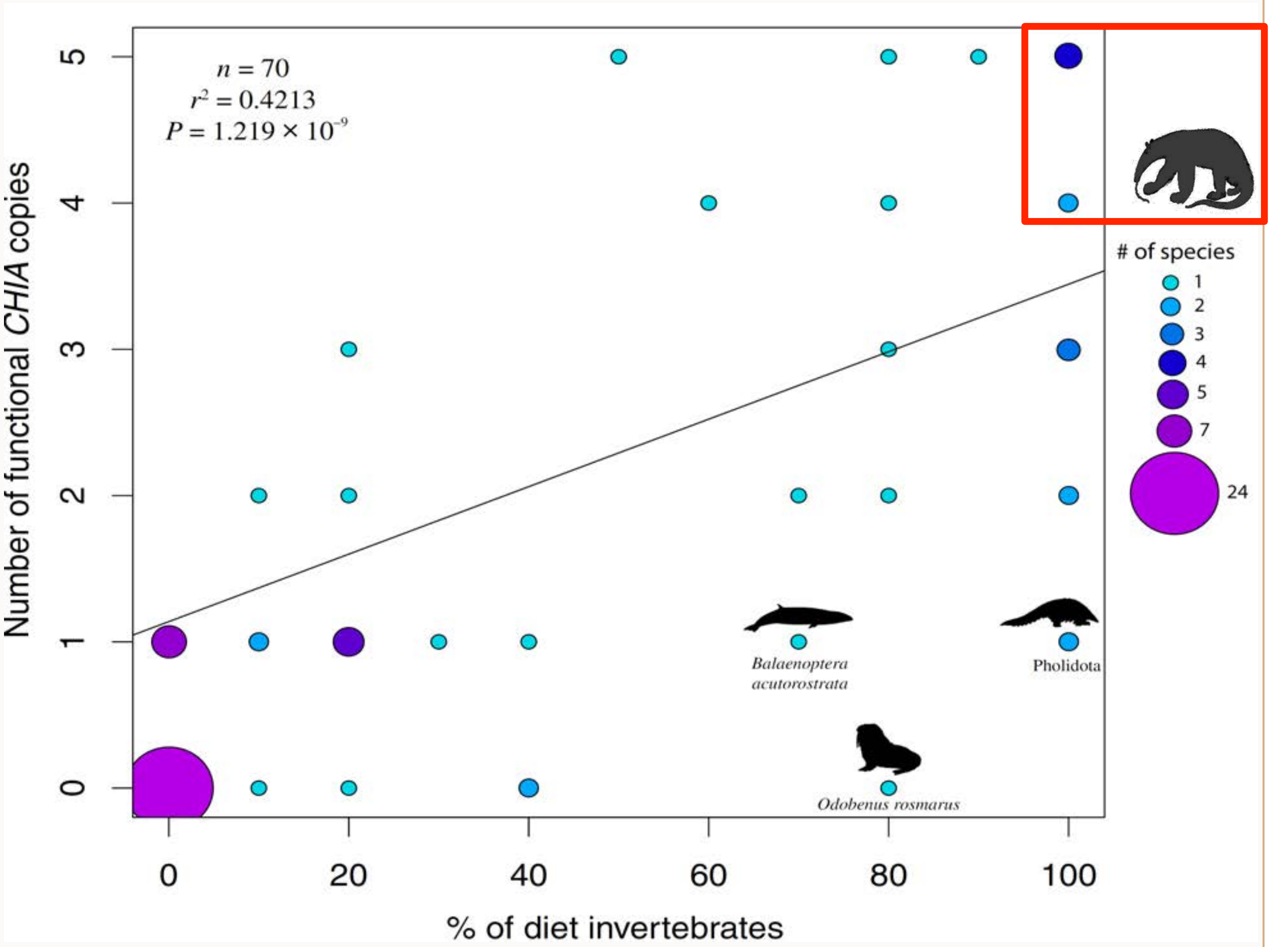
0-5 functional genes

↗ % of diet invertebrates

=

↗ number of functional genes

Myrmecophagous mammals have several copies!



Emerling *et al.* (2018) Science advances

The chitinase family

0-5 functional genes

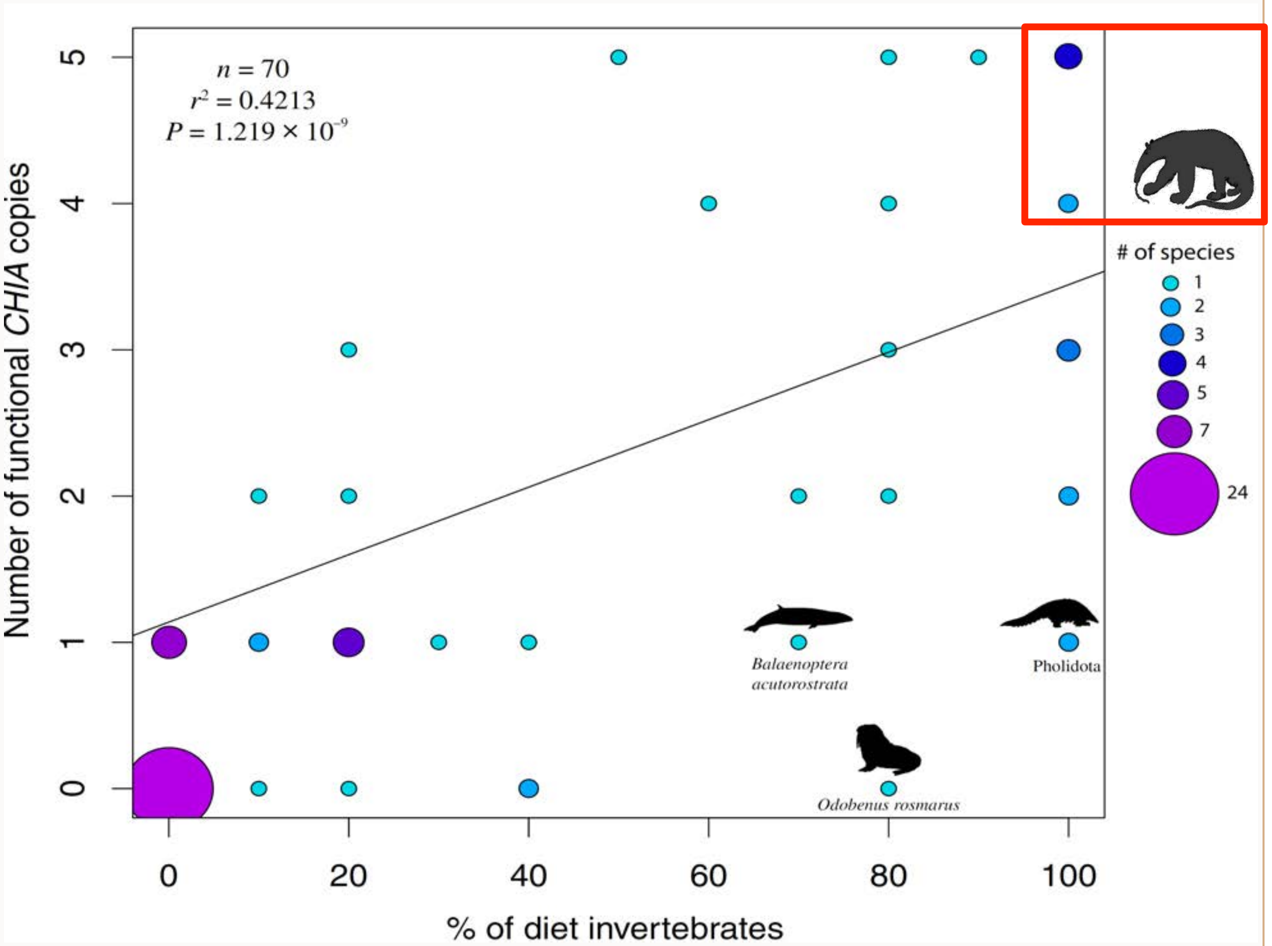
↗ % of diet invertebrates

=

↗ number of functional genes

Myrmecophagous mammals
have several copies!

except …



Emerling *et al.* (2018) Science advances

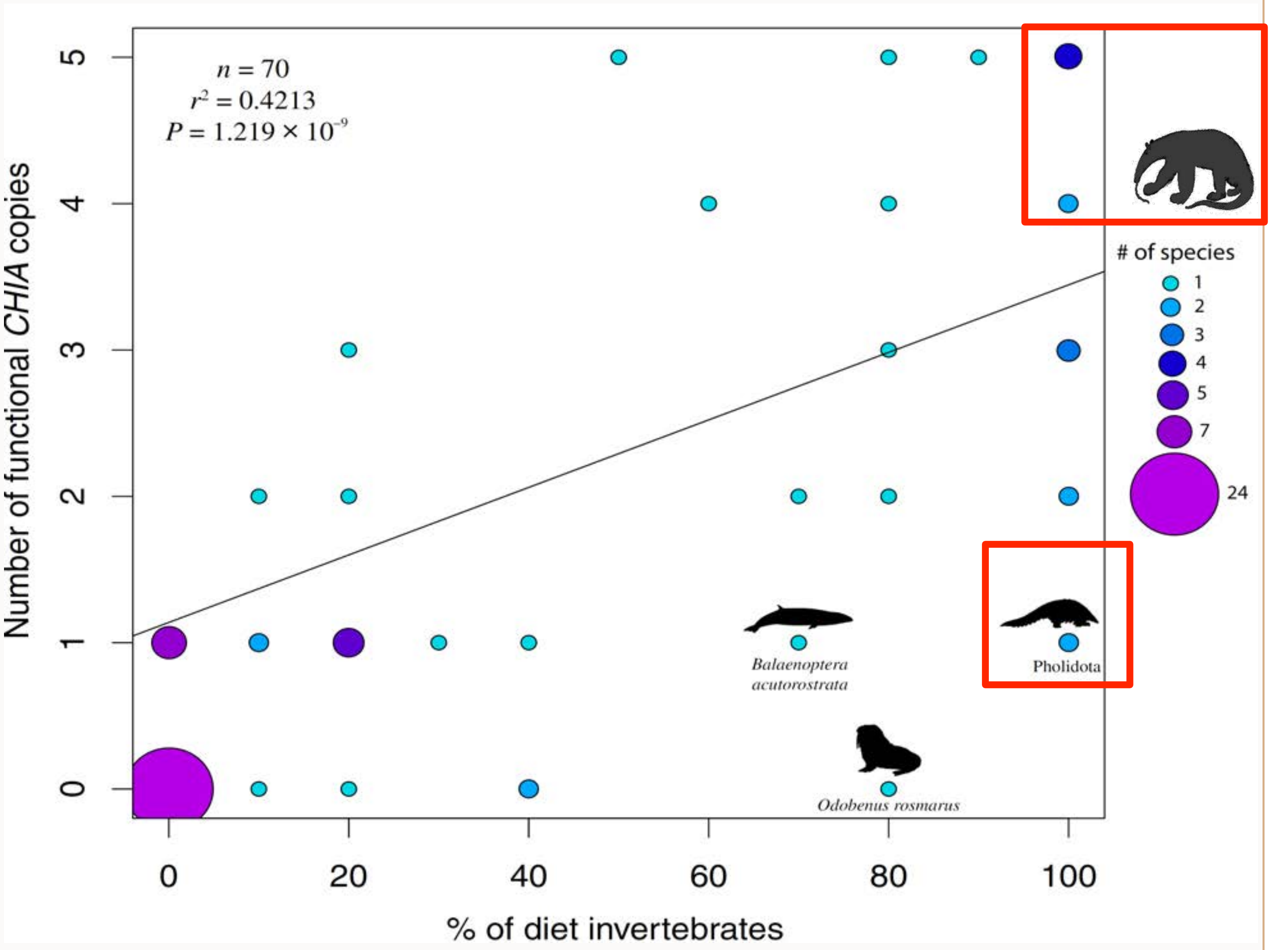
The chitinase family

0-5 functional genes

↗ % of diet invertebrates
=
↗ number of functional genes

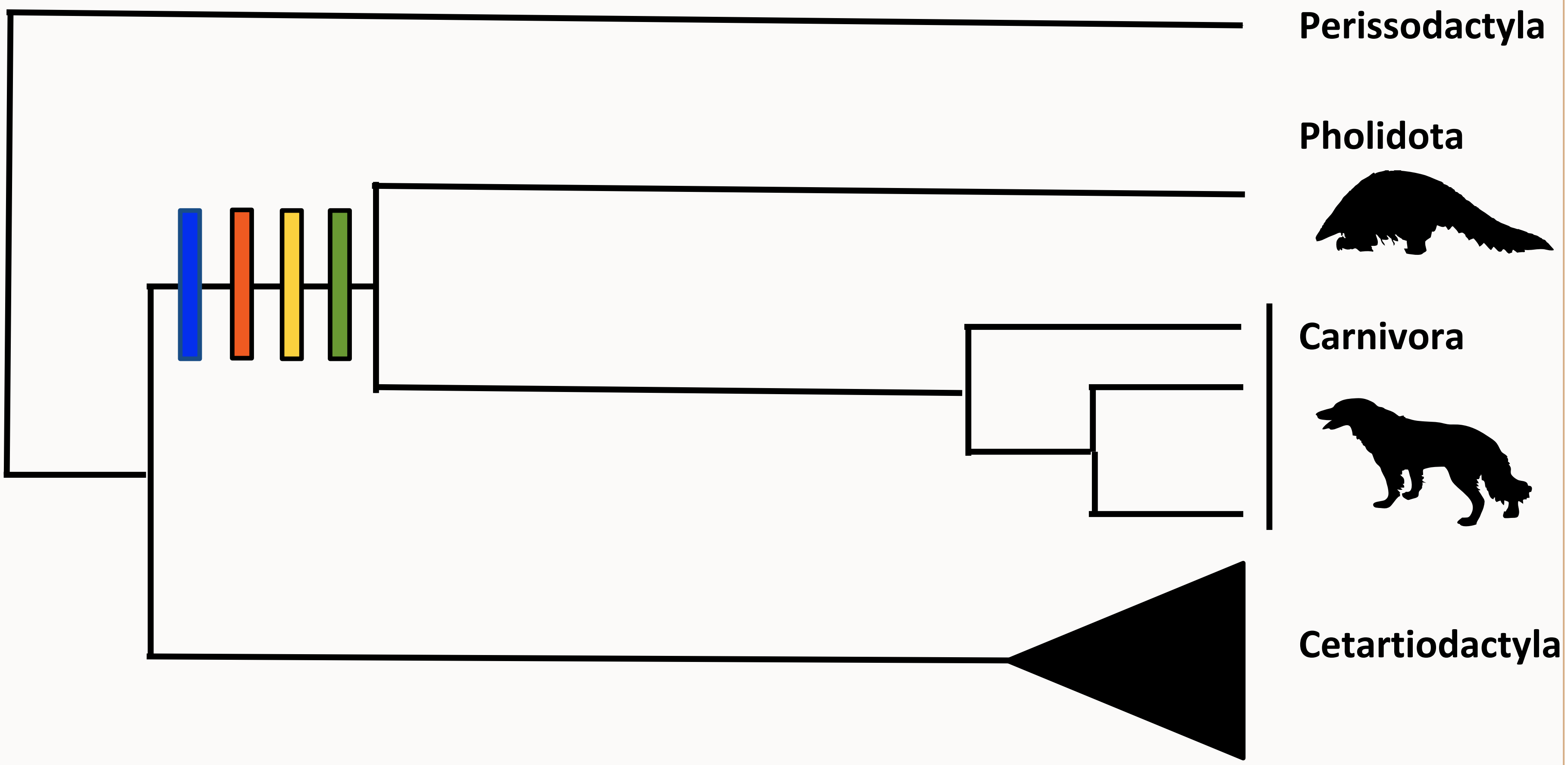
Myrmecophagous mammals have several copies!

except …
Manis javanica!



Emerling *et al.* (2018) Science advances

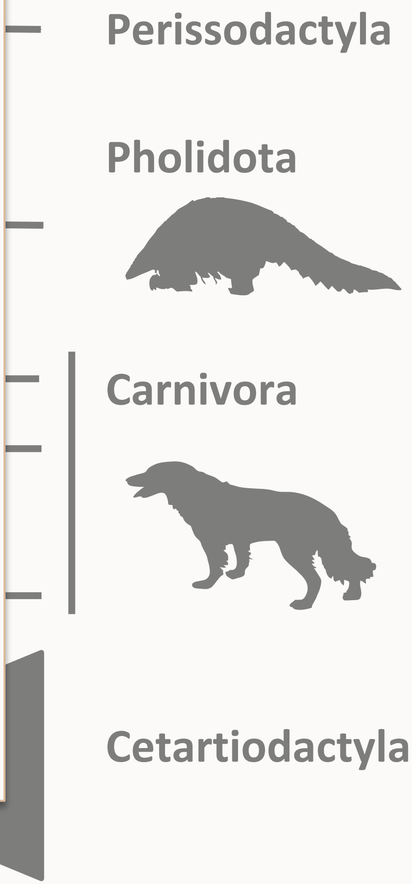
The chitinase family



Emerling *et al.* (2018) Science advances

The chitinase family

How does the pangolin digests chitin with only one functional gene?

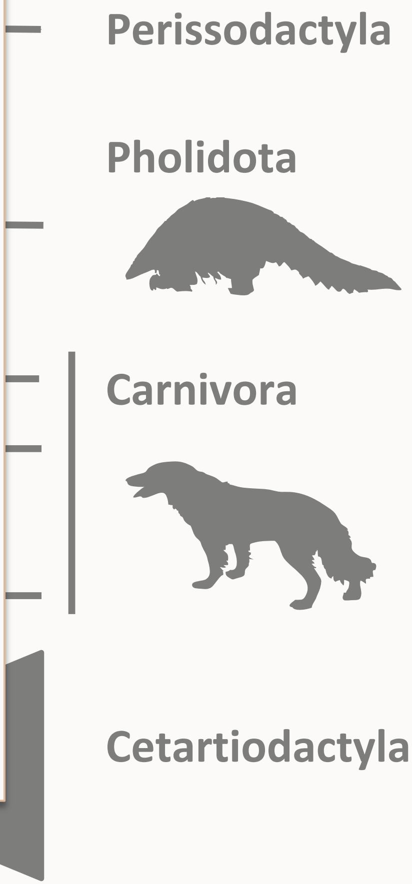


Emerling *et al.* (2018) Science advances

The chitinase family

How does the pangolin digests chitin with only one functional gene?

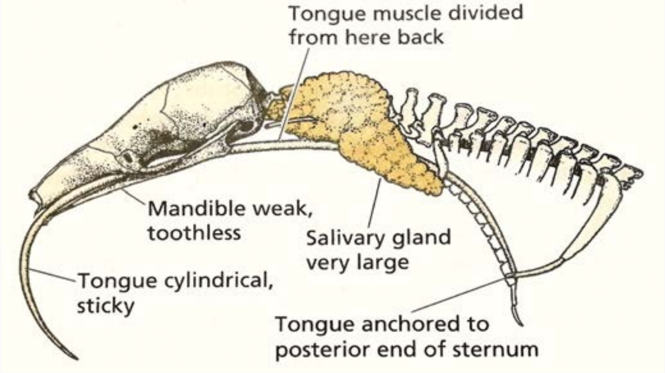
Does he over express this gene?



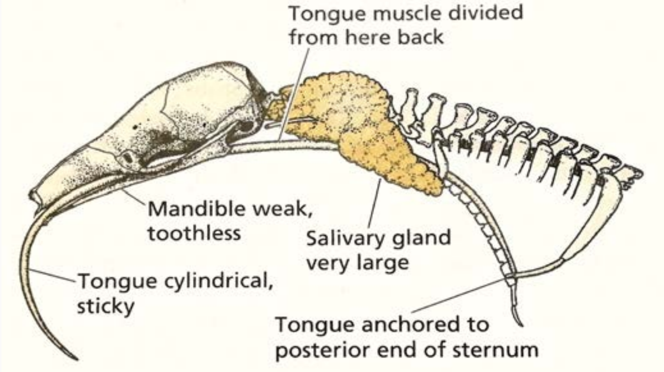
Emerling *et al.* (2018) Science advances

Gene expression in salivary glands

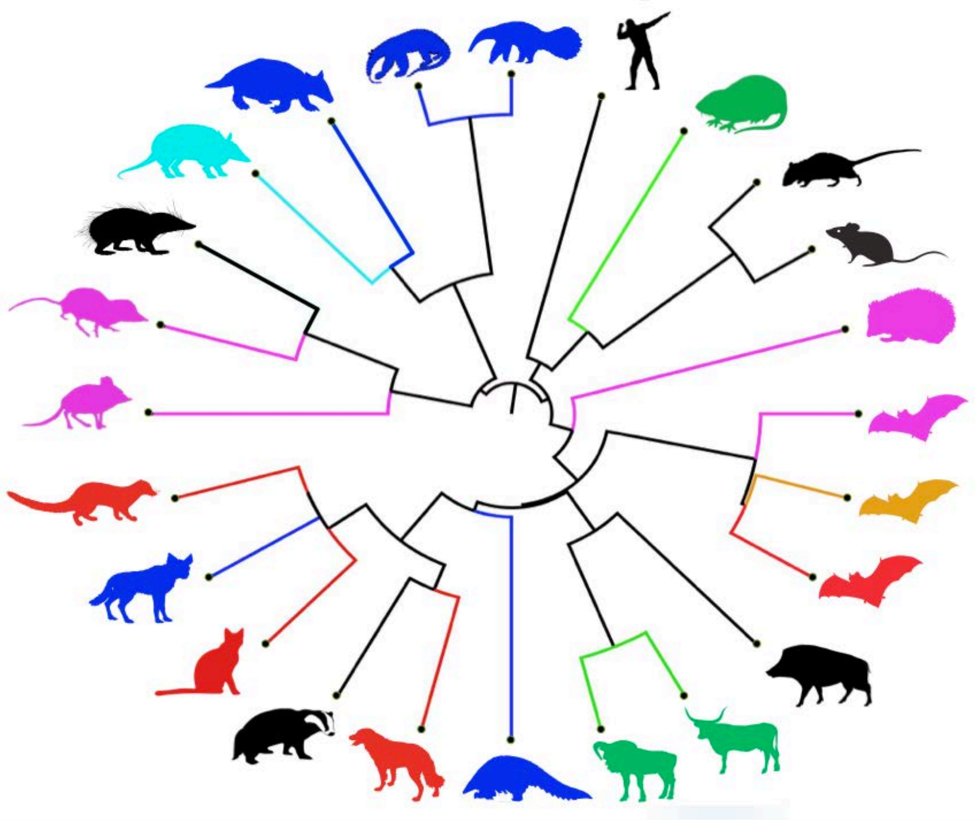
Gene expression in salivary glands



Gene expression in salivary glands



- 28 salivary gland transcriptomes
(24 species, 5 myrmecophagous)



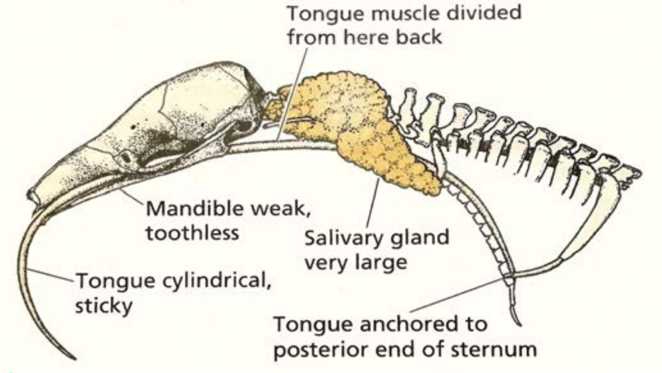
Myrmecophagous
90 -100%
75 -89%
Insectivore
Carnivore
Herbivore
Omnivore
Frugivore

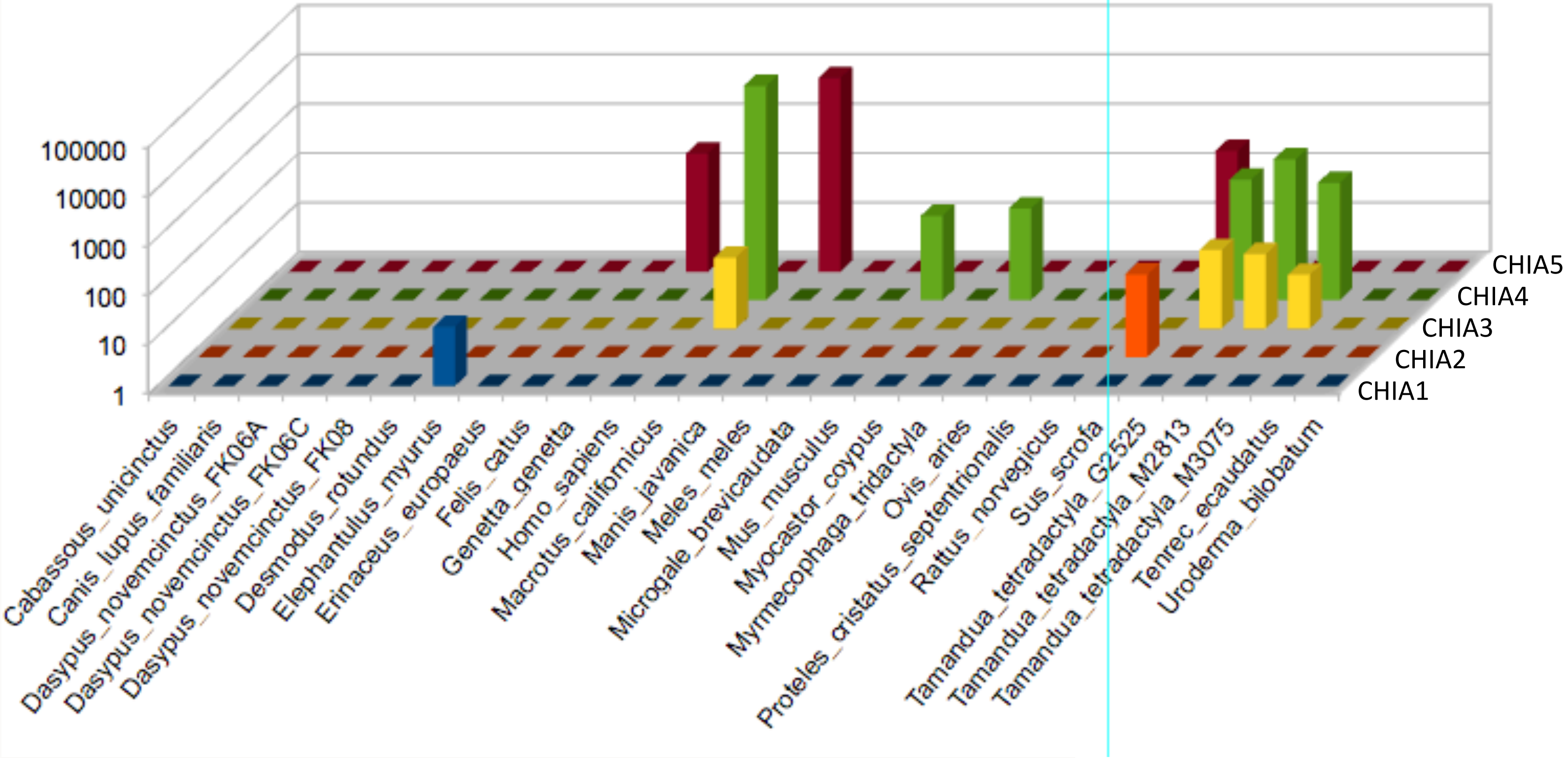
Dave Lutgen (2019) *Master thesis*

Gene expression in salivary glands

- 28 salivary gland transcriptomes

(24 species, 5 myrmecophagous)



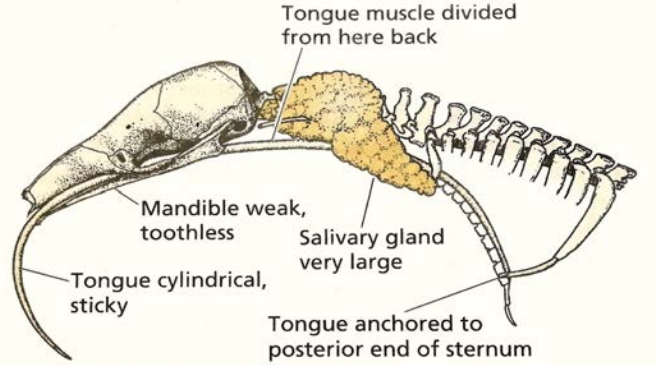


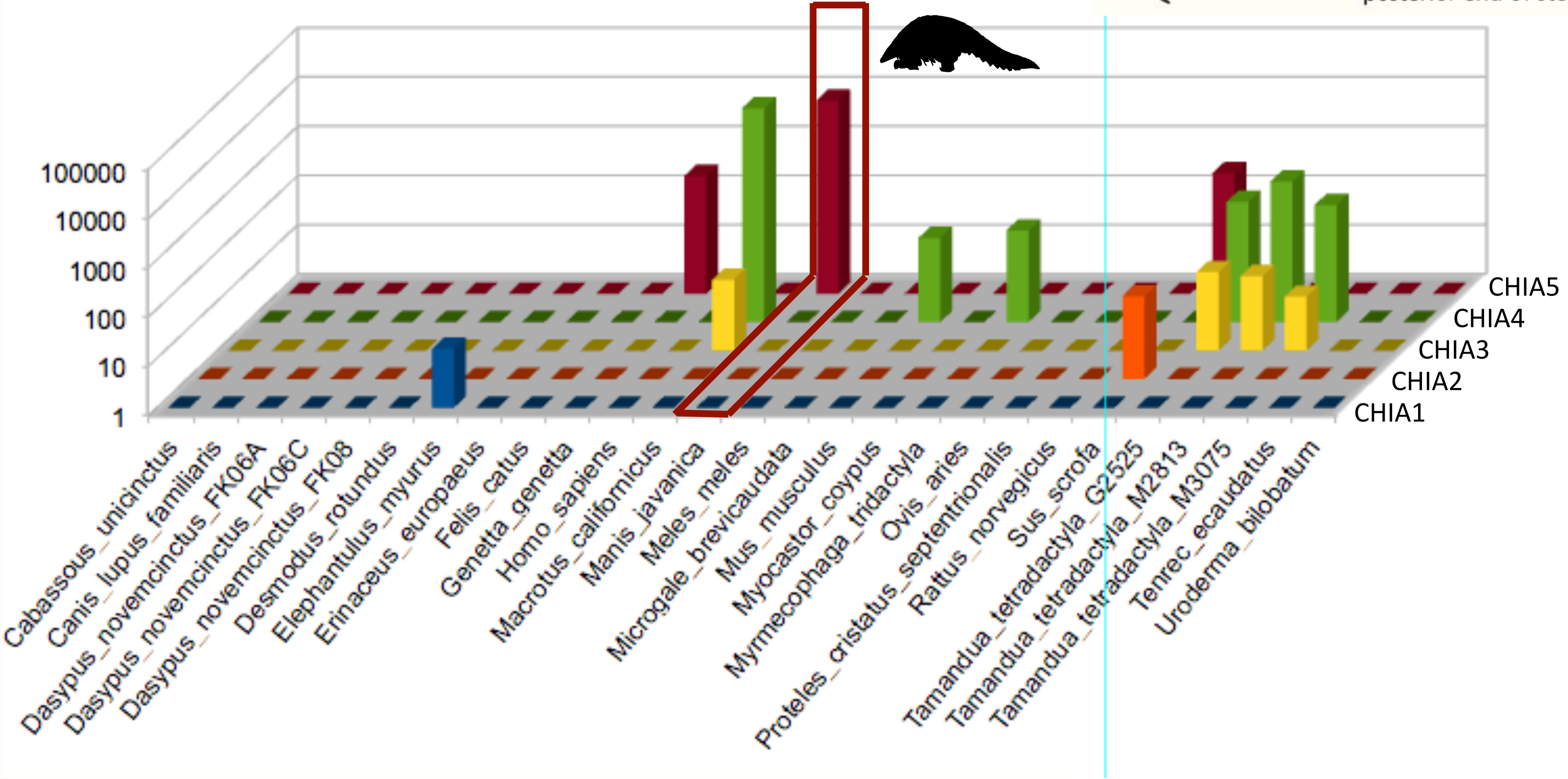
Dave Lutgen (2019) *Master thesis*

Gene expression in salivary glands

- 28 salivary gland transcriptomes

(24 species, 5 myrmecophagous)

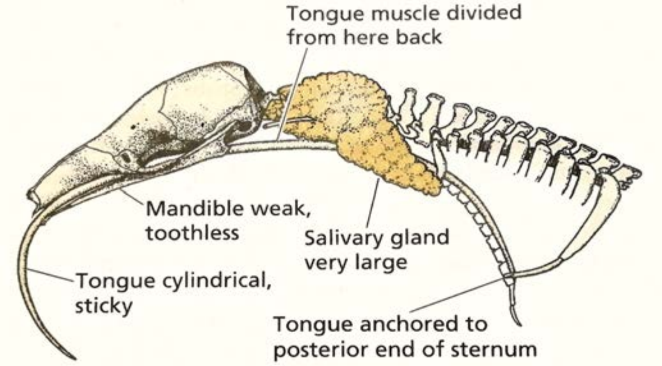


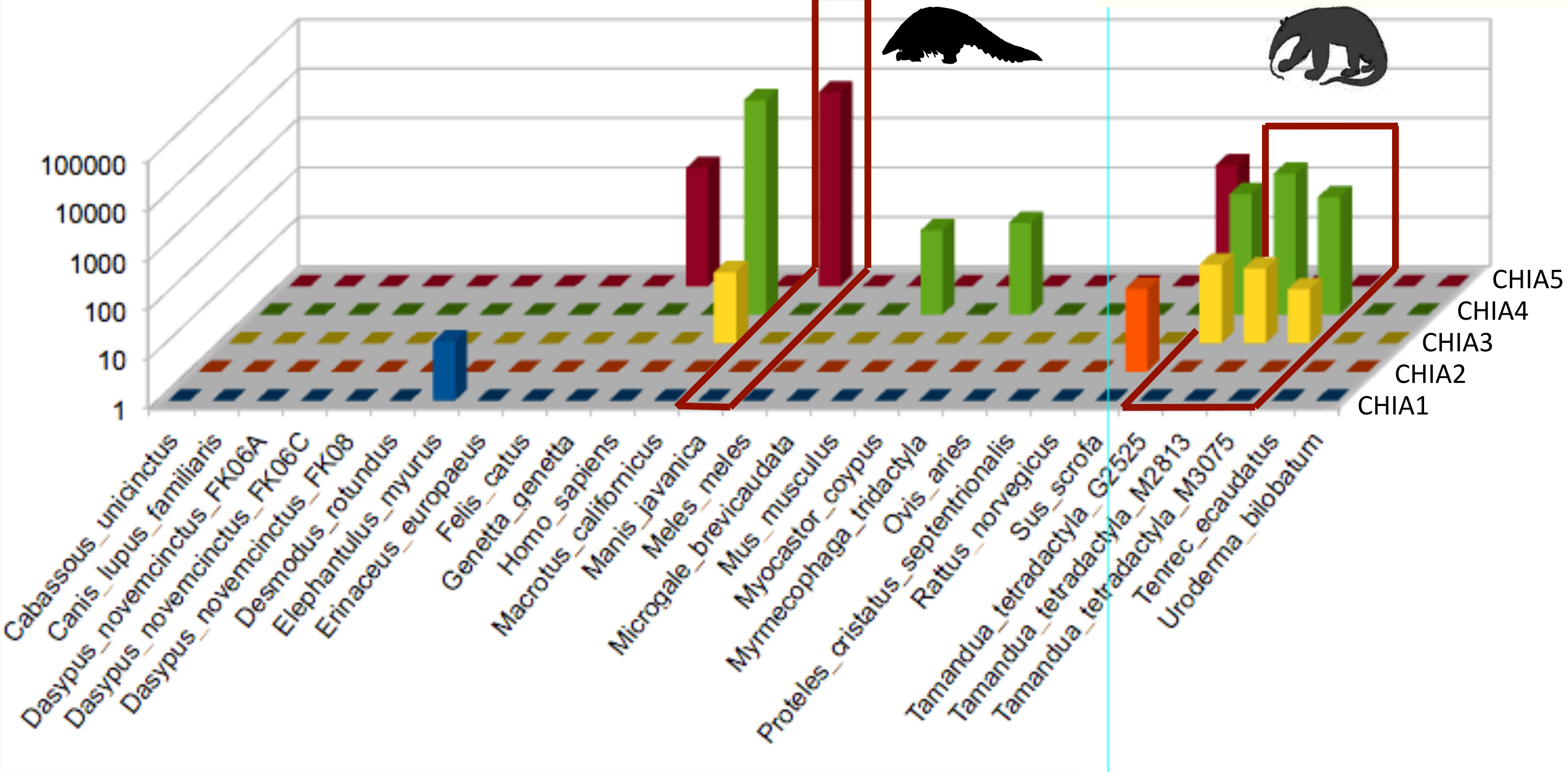


Dave Lutgen (2019) *Master thesis*

Gene expression in salivary glands

- 28 salivary gland transcriptomes
(24 species, 5 myrmecophagous)

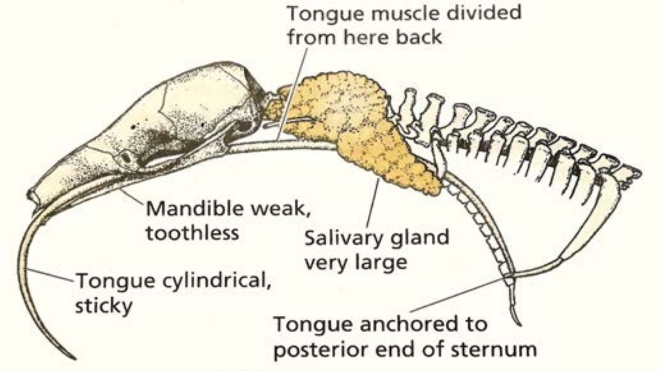


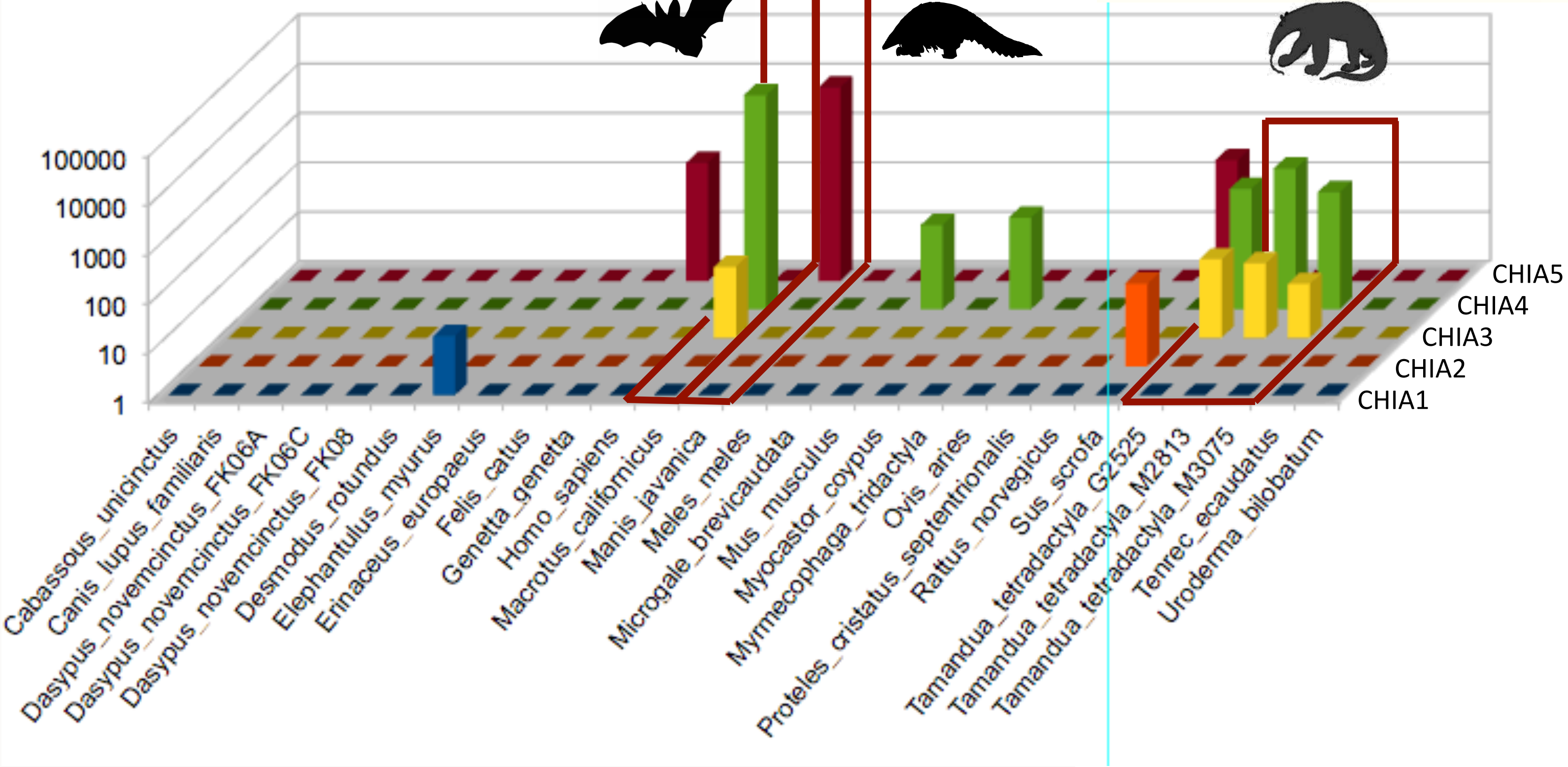


Dave Lutgen (2019) *Master thesis*

Gene expression in salivary glands

- 28 salivary gland transcriptomes
(24 species, 5 myrmecophagous)

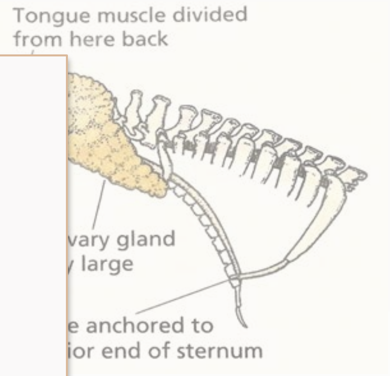




Dave Lutgen (2019) *Master thesis*

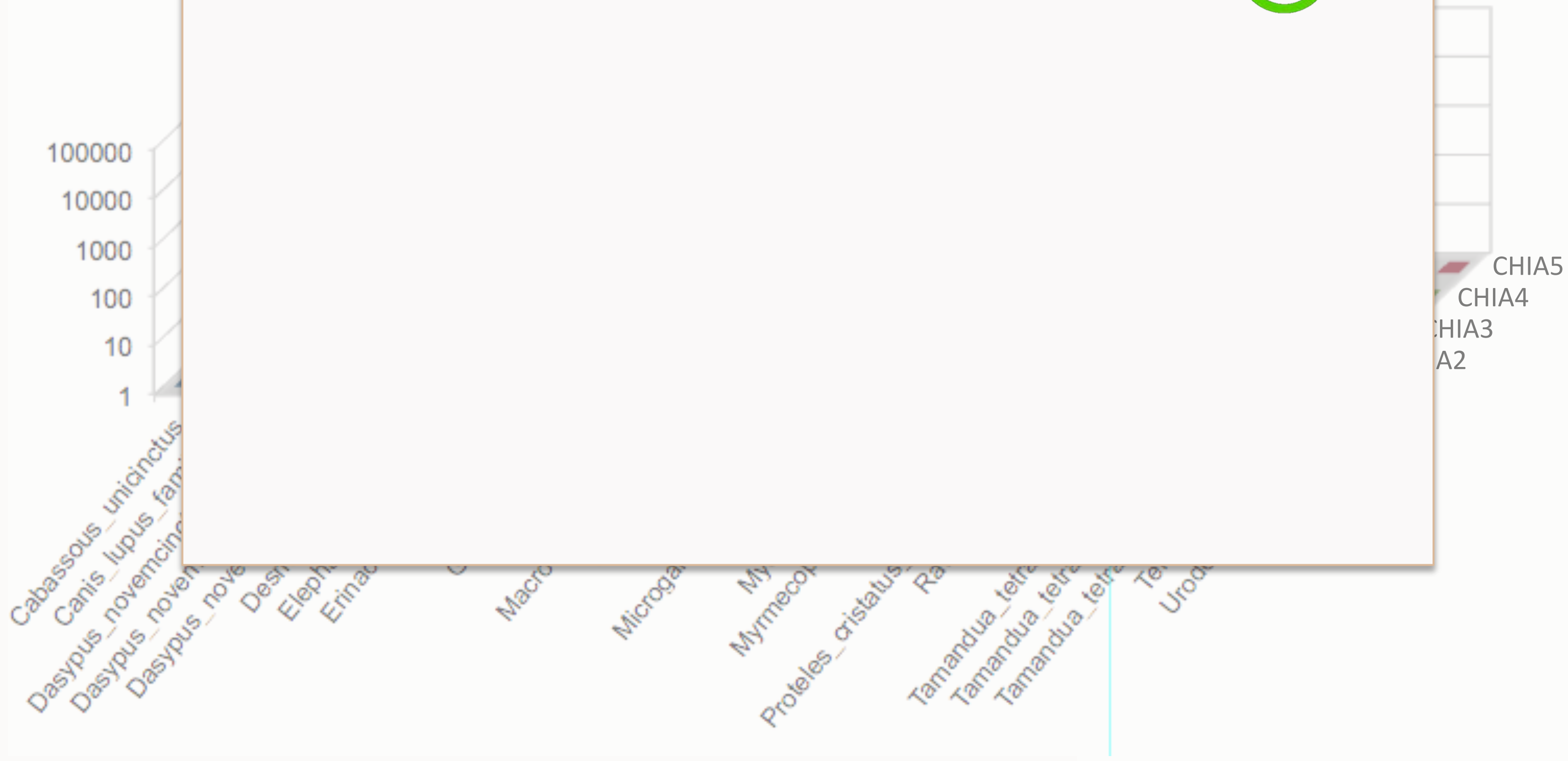
Gene expression in salivary glands

- 28 salivary gland transcriptomes



How does the pangolin digests chitin with only one functional gene?

➔ Overexpression of CHIA5 in salivary glands ✓



Dave Lutgen (2019) *Master thesis*

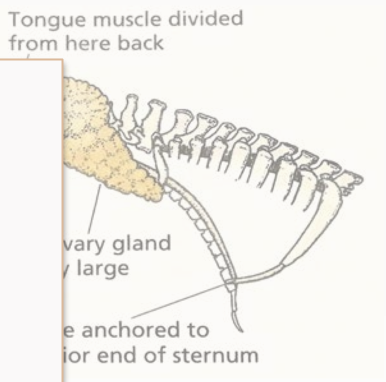
Gene expression in salivary glands

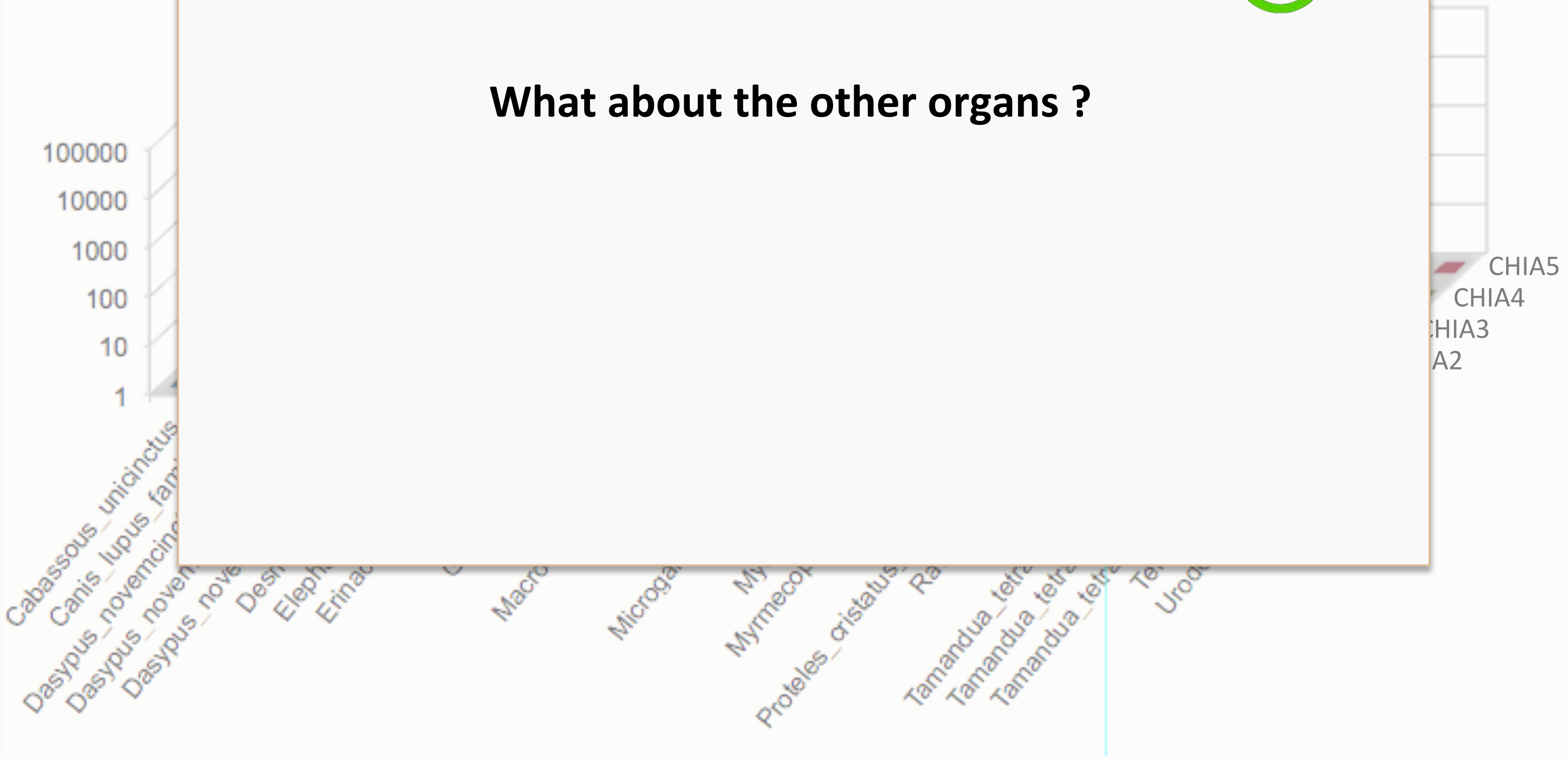
- 28 salivary gland transcriptomes

How does the pangolin digests chitin with only one functional gene?

➔ Overexpression of CHIA5 in salivary glands ✓

What about the other organs ?





Dave Lutgen (2019) *Master thesis*

Gene expression in salivary glands

- 28 salivary gland transcriptomes

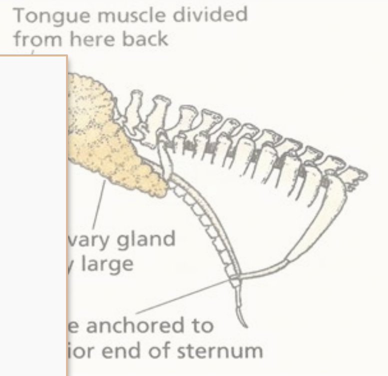
How does the pangolin digests chitin with only one functional gene?

➔ Overexpression of CHIA5 in salivary glands ✓

What about the other organs ?

New dataset with
3 myrmecophagous mammals
and several organs

Sophie Teullet (2020) *Master thesis*



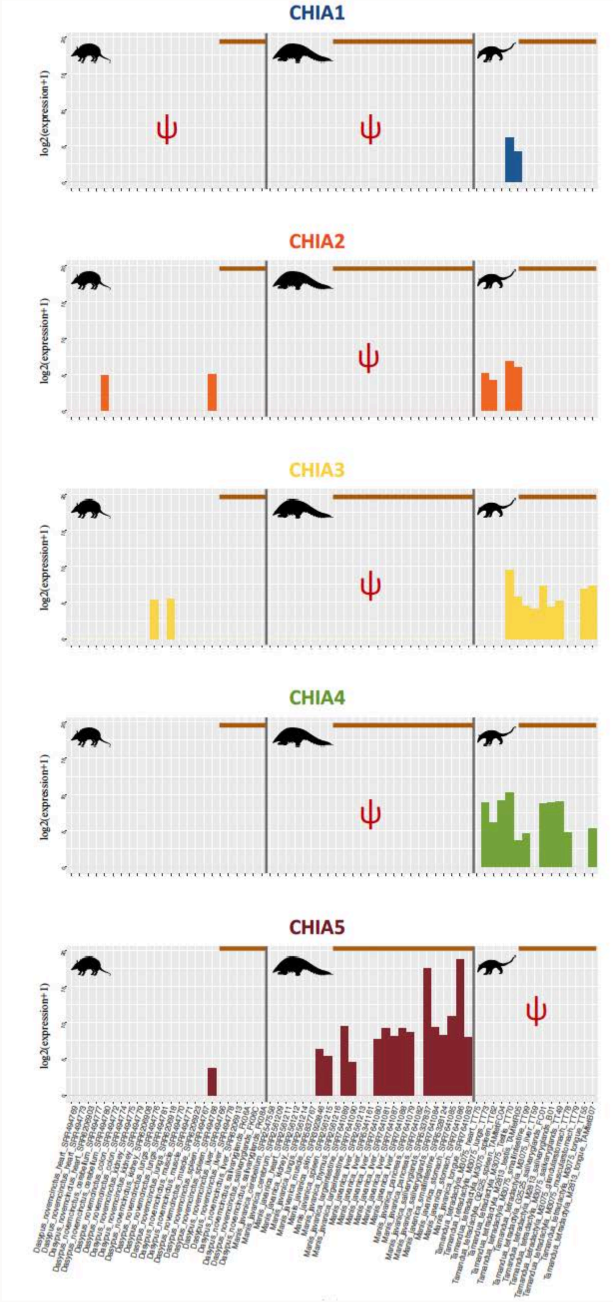
Dave Lutgen (2019) *Master thesis*

Comparative transcriptomics across different organs

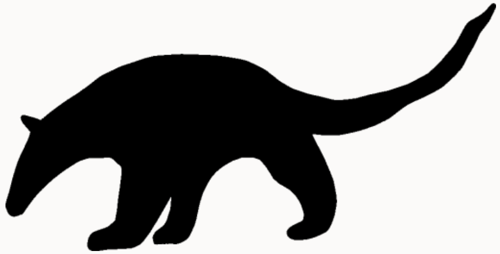
Digestive organs

ψ Pseudogenes

Manis javanica



Tamandua tetradactyla



CHIA1
CHIA2 ψ
CHIA3
CHIA4

CHIA5 ✓

Many digestive organs:
Salivary glands + stomach:
CHIA5 +++

CHIA1
CHIA2 ✓
CHIA3
CHIA4

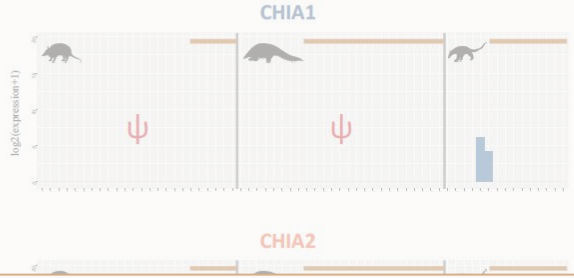
CHIA5 ψ

Salivary glands, tongue, liver, intestine: ***CHIA3* and *CHIA4***

Allio*, Teullet*, Lutgen* *et al.* (in prep.)

Comparative transcriptomics across different organs

Digestive organs



ψ Pseudogenes

Manis javanica

CHIA1
CHIA2 ψ *CHIA5* ✓
CHIA3
CHIA4

Many digestive organs:
Salivary glands + stomac
CHIA5 +++

Conclusion

Chitinase gene family

✓ Over expression of CHIA5 in pangolins

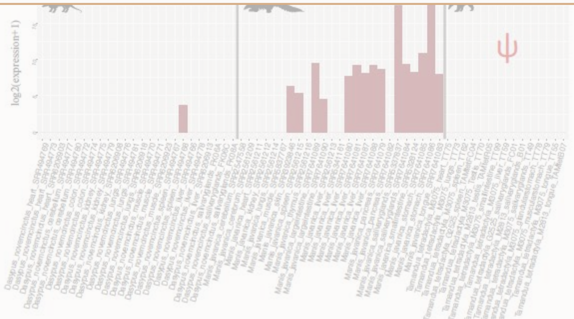
✓ Impact of historical contingency

✓ Different strategy

Tamandua tetradactyla

A1
A2 ✓ *CHIA5* ψ
A3
A4

ary glands, tongue, liver,
stine: ***CHIA3* and *CHIA4***



Allio*, Teullet*, Lutgen* *et al.* (in prep.)

Summary

PART 1 - Phylogenomics and comparative genomics in myrmecophagous mammals

Take Home Message

Roadkill tissues are suitable for long DNA sequencing

✓

High quality genomes for study convergence

✓

Species delineation possible with few individuals

✓

Impact of historical contingency

⚠

Summary

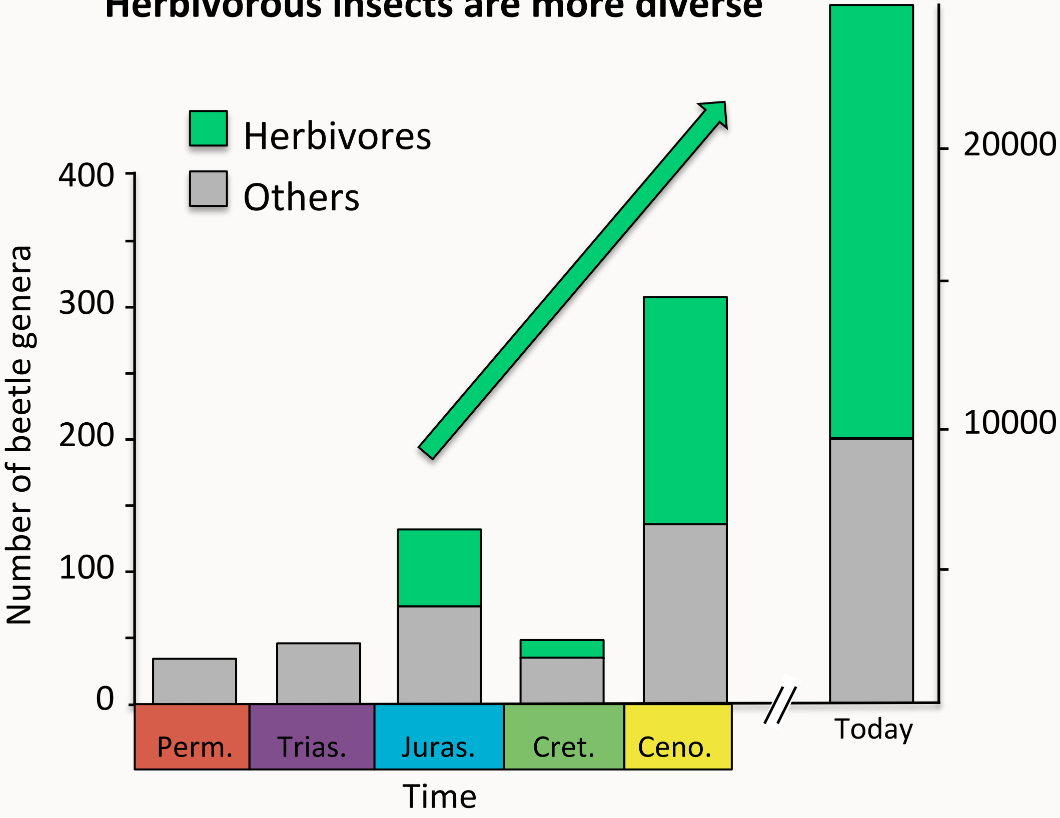
PART 1 - Phylogenomics and comparative genomics in myrmecophagous mammals

PART 2 - The case of coevolution between the swallowtail butterflies (Papilionidae) and their host plants

CONTEXT

- 1 million insect species
- 300,000 plant species

Herbivorous insects are more diverse

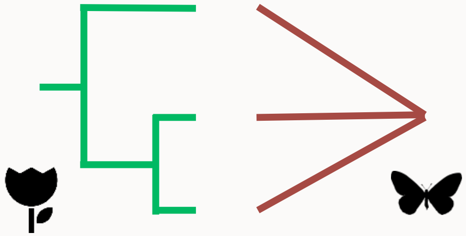


What is the role of interaction in diversification?

Farrell 1998 – Science; Grimaldi & Engel 2005; Janz *et al.* 2006 – BMC Evol. Biol., Nyman *et al.* 2012 – Ecol. Lett., Hardy & Otto 2014 – Proc. B

CONTEXT

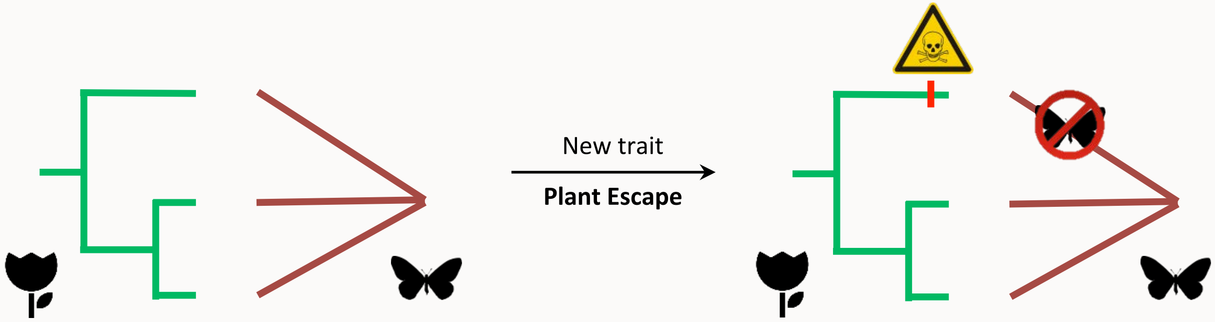
“Escape and Radiate” hypothesis



Ehrlich & Raven 1964 – Evolution, Thompson 1989 – TREE; Janz 2011 – Annu. Rev. Ecol. Evol. Syst.

CONTEXT

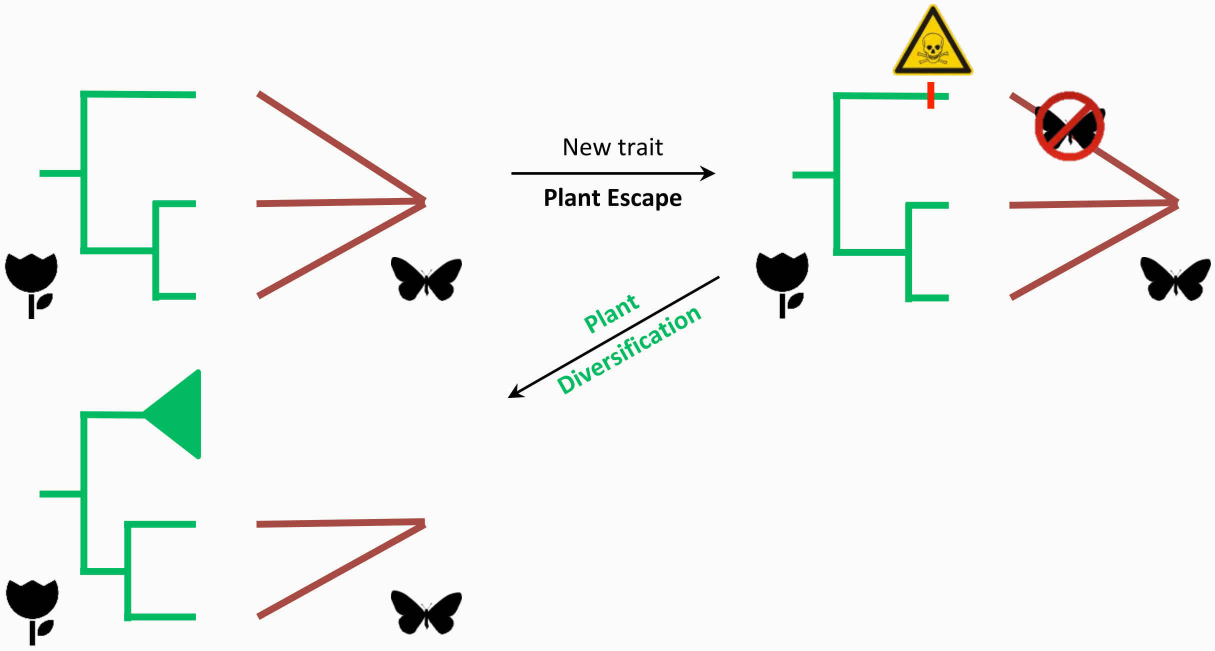
“Escape and Radiate” hypothesis



Ehrlich & Raven 1964 – Evolution, Thompson 1989 – TREE; Janz 2011 – Annu. Rev. Ecol. Evol. Syst.

CONTEXT

“Escape and Radiate” hypothesis



New trait

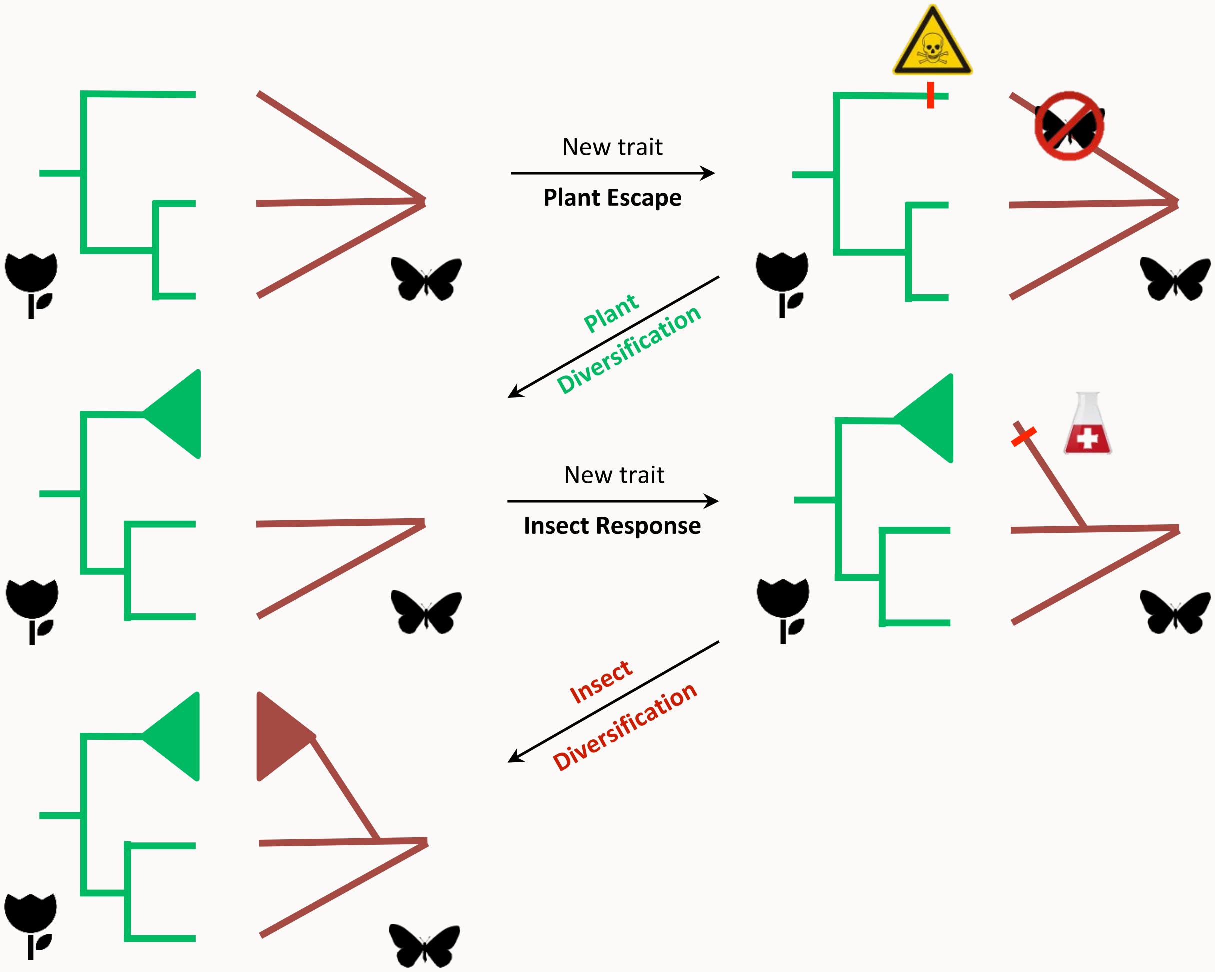
Plant Escape

Plant Diversification

Ehrlich & Raven 1964 – Evolution, Thompson 1989 – TREE; Janz 2011 – Annu. Rev. Ecol. Evol. Syst.

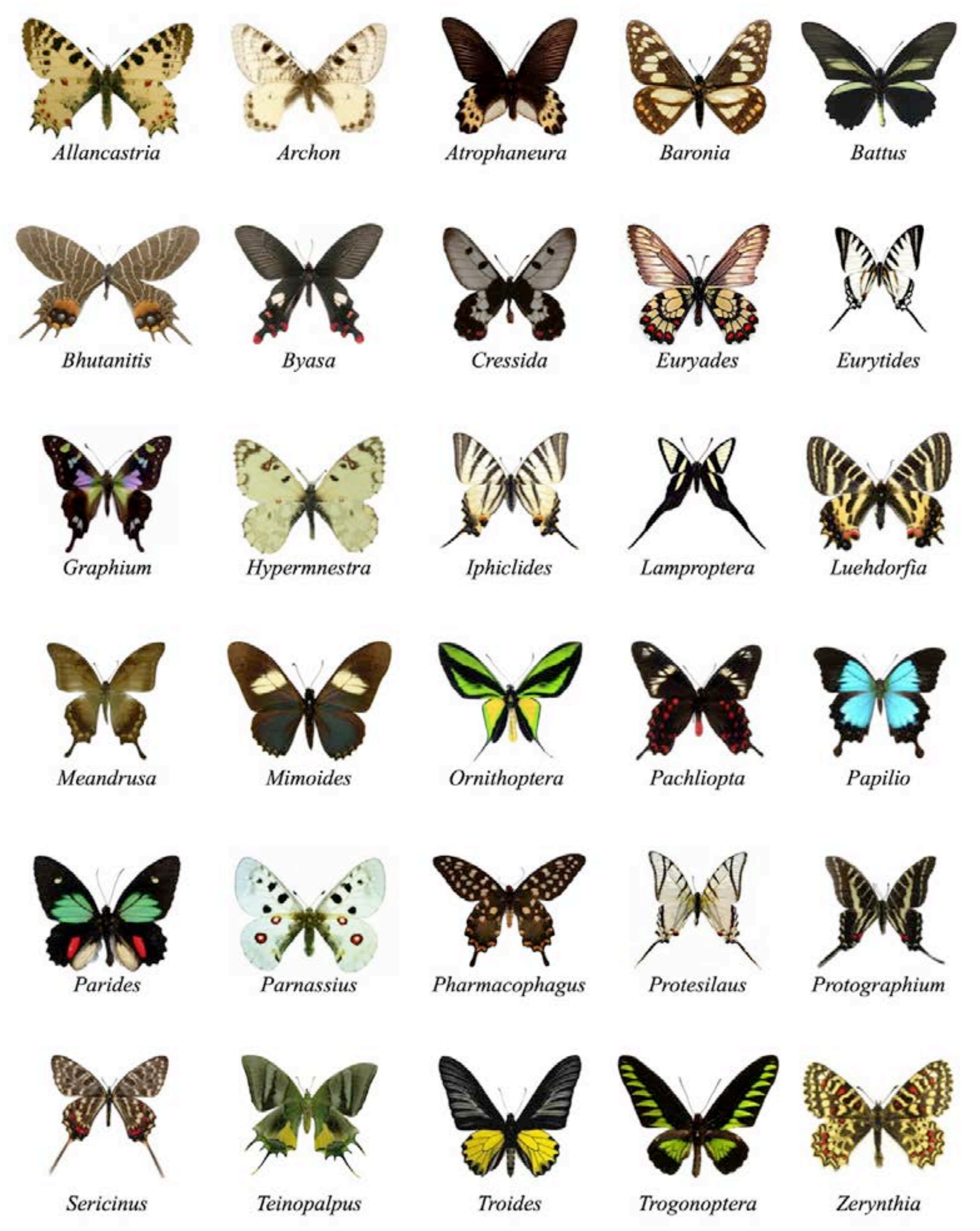
CONTEXT

“Escape and Radiate” hypothesis



Ehrlich & Raven 1964 – Evolution, Thompson 1989 – TREE; Janz 2011 – Annu. Rev. Ecol. Evol. Syst.

The swallowtail butterflies (Papilionidae)



- Well known family
- High morphological diversity
- 32 genera in 570+ species
- Feed on 12 families of flowering plants


Aristolochiaceae


Rutaceae …

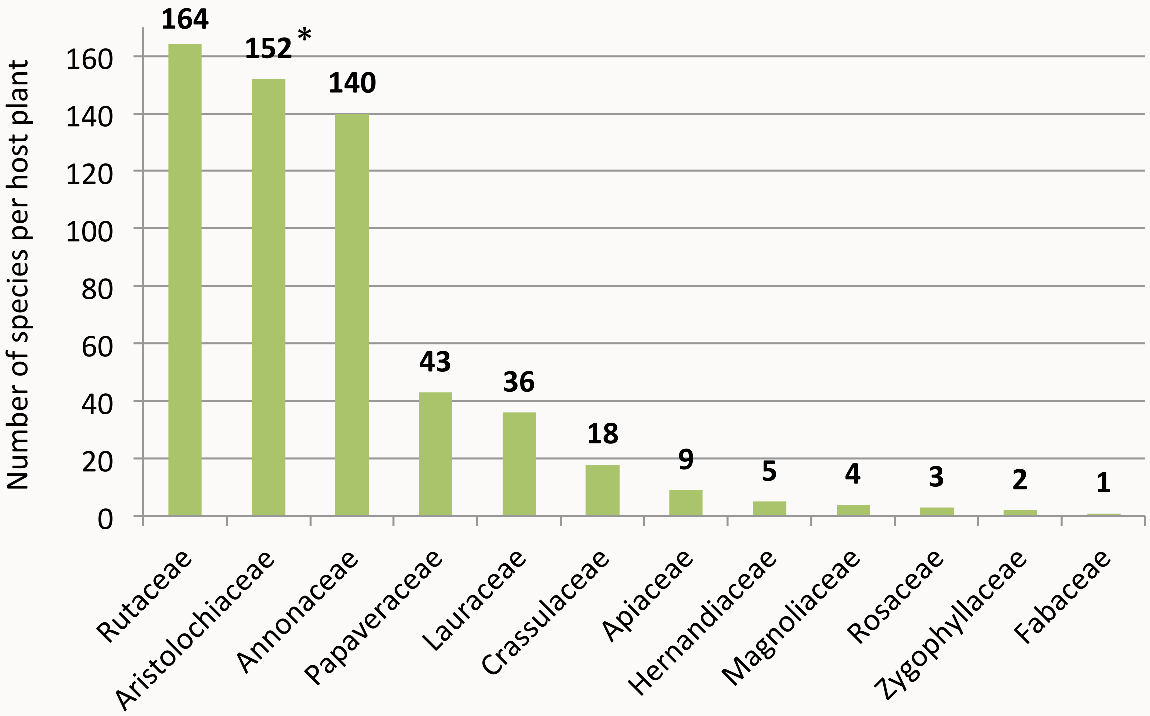

Annonaceae


Apiaceae

- Many species are monophagous
  = specialists

Swallowtail butterflies and their host plants

Highly **unbalanced associations** between papilionid diversity and host plants



* Three groups **are specialized** on **Aristolochiaceae** (highly toxic plants, Aristolochic acid)

Sequestration of toxic compounds ⟶ **Defense** against predators

Hypothesis of **coevolution** between Papilionidae and their host plants

Swallowtails on:

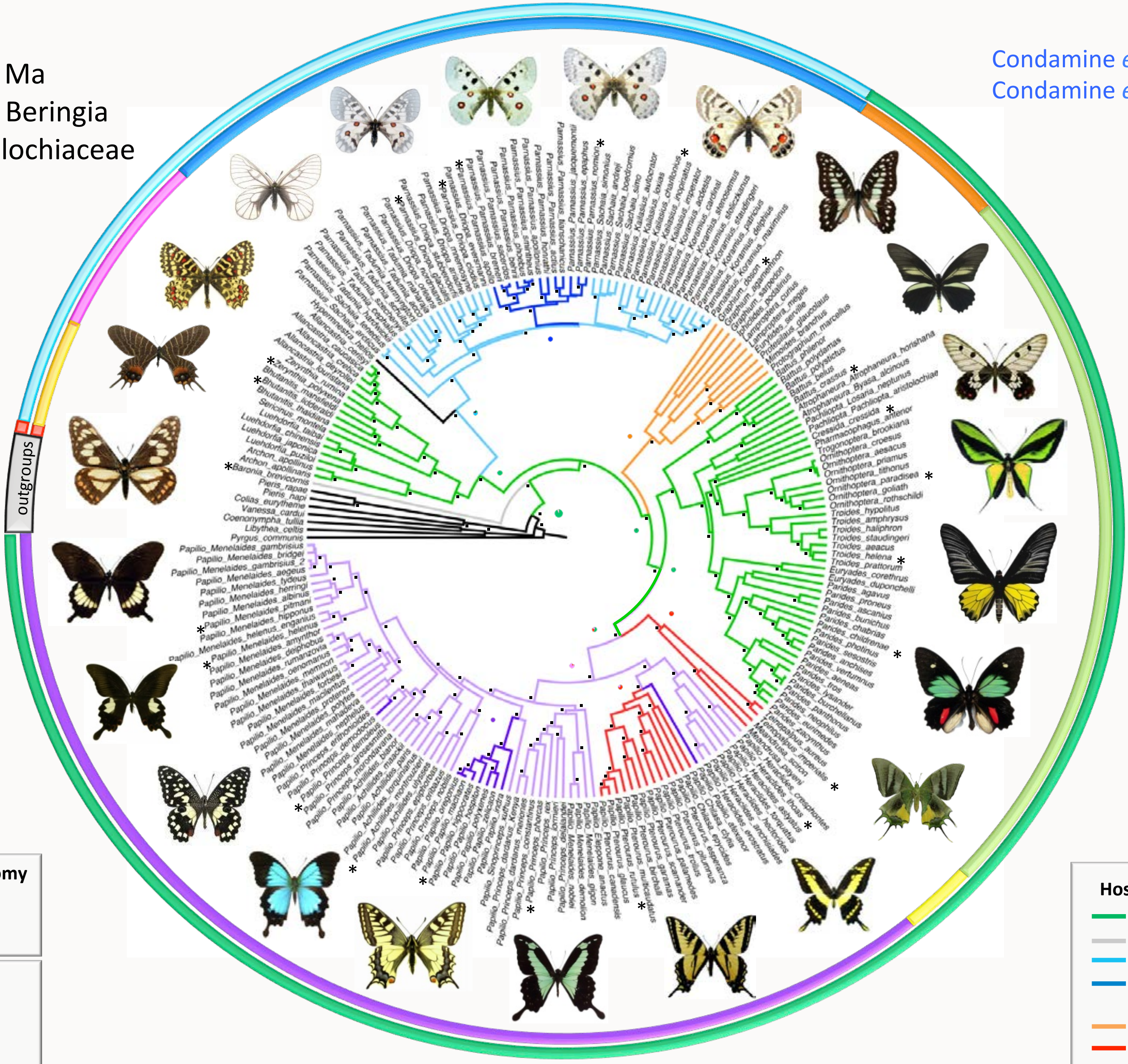
Rutaceae

Annonaceae

Aristolochiaceae

Papilionidae:
- appeared 52 Ma
- originated in Beringia
- fed on Aristolochiaceae

Condamine *et al.* 2012 – Ecol. Lett.
Condamine *et al.* 2013 – J. Biogeo.



Swallowtail taxonomy

- 🟩 Papilioninae
- 🟦 Parnassiinae
- 🟥 Baroniinae

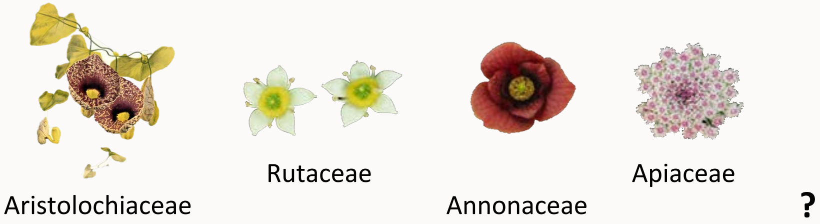
- 🟪 Papilionini
- 🟨 Teinopalpini
- 🟩 Troidini
- 🟧 Leptocircini
- 🟦 Parnassiini
- 🟪 Zerynthiini
- 🟨 Luehdorfiini

Host plant association

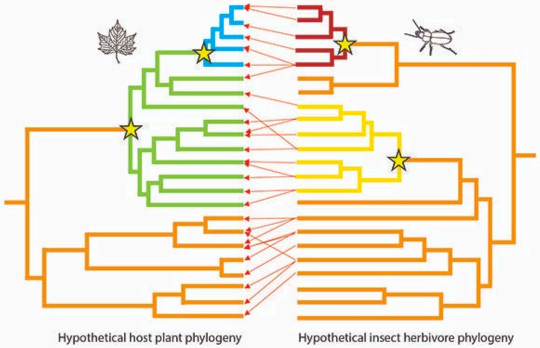
- 🟩 Aristolochiaceae
- ⬜ Fabaceae (*Acacia*)
- 🟦 Papaveraceae
- 🟦 Crassulaceae + Saxifragaceae
- 🟧 Annonaceae
- 🟥 Lauraceae
- 🟪 Rutaceae
- 🟪 Apiaceae
- ⬛ Zygophyllaceae

Main questions

1. ***What are the ancestral host plants of Papilionidae?***



2. ***Temporal congruence?***



3. ***Biogeographic congruence?***

4. ***Signatures of genomic adaptations and on speciation rates?***

Main questions

1. ***What are the ancestral host plants of Papilionidae?***

Aristolochiaceae

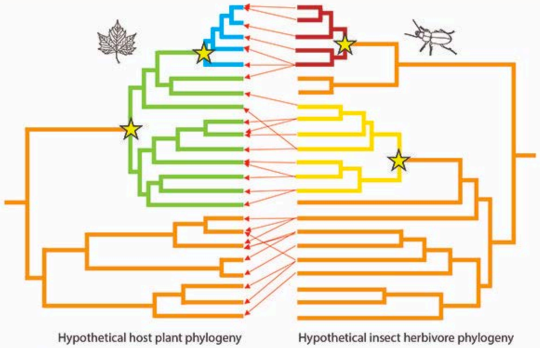
Rutaceae

Annonaceae

Apiaceae

?

2. ***Temporal congruence?***

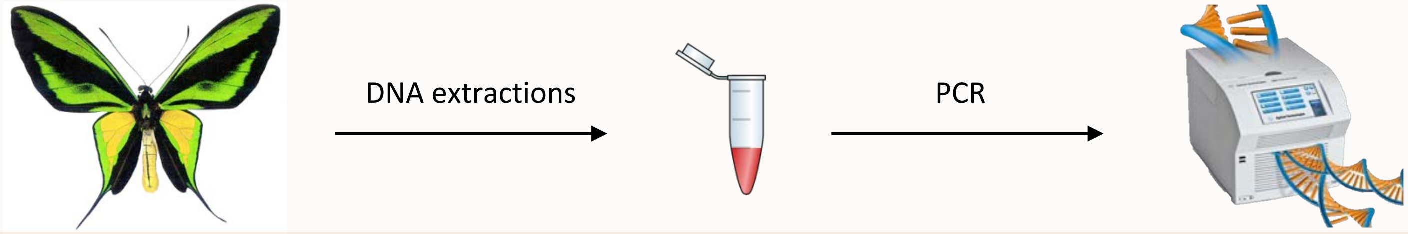


3. ***Biogeographic congruence?***

4. ***Signatures of genomic adaptations and on speciation rates?***

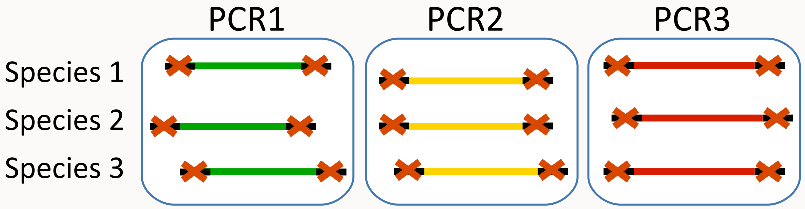
Inferring the most comprehensive and robust phylogeny with Sanger data

Sanger sequencing for **408 species (72% of the diversity)**

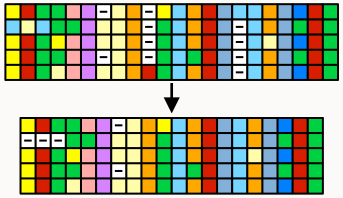


Sequence cleaning

Alignments



Geneious

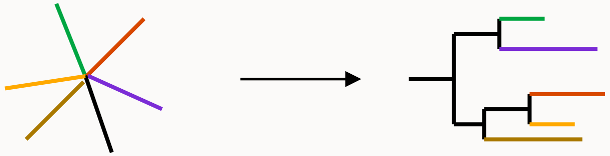


MAFFT

Mesquite

Phylogenetic inferences

Dataset: - *408 species (+ 20 outgroups)*
- *7 genes ; 5 873 nucleotides*



IQ-TREE

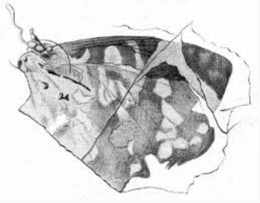
MrBayes

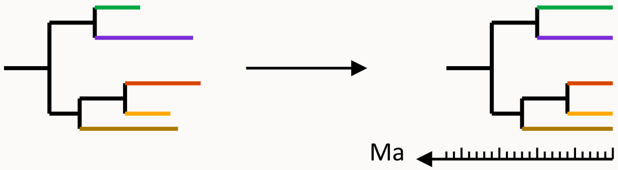
Molecular dating

4 fossils

4 ingroup fossils

Outgroups excluded





BEAST

Allio *et al.* 2021 – Nat. Comm.

Inferring the most comprehensive and robust phylogeny with Sanger data

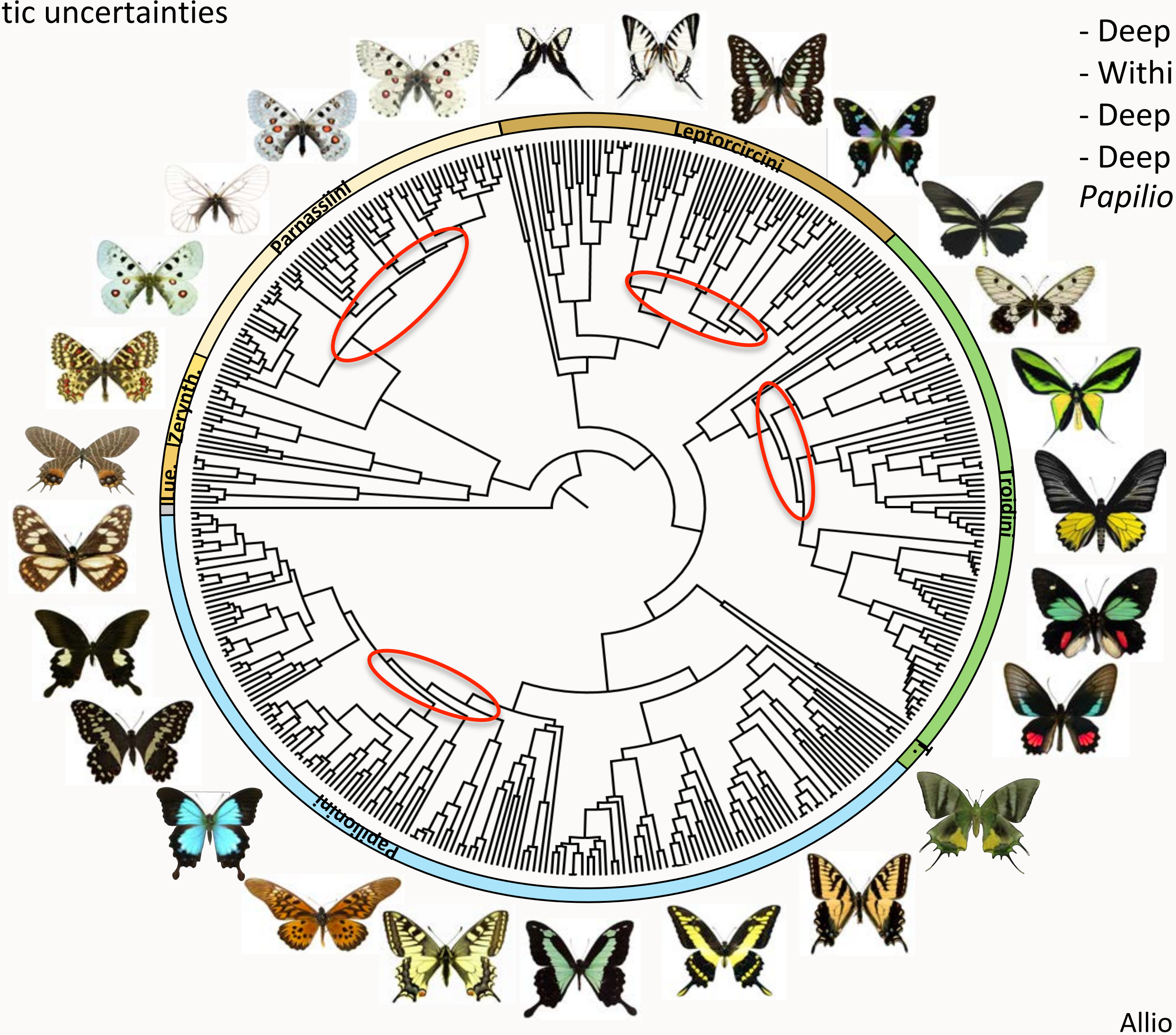
MrBayes: 100 M. MCMC generations

◯ Phylogenetic uncertainties

85% of nodes had PP ≥ 0.95

Not well resolved in rapid radiations such as:

- Deep nodes in *Parnassius*
- Within *Graphium*
- Deep nodes in Troidina
- Deep nodes in Old World *Papilio*



Allio *et al.* 2021 – Nat. Comm.

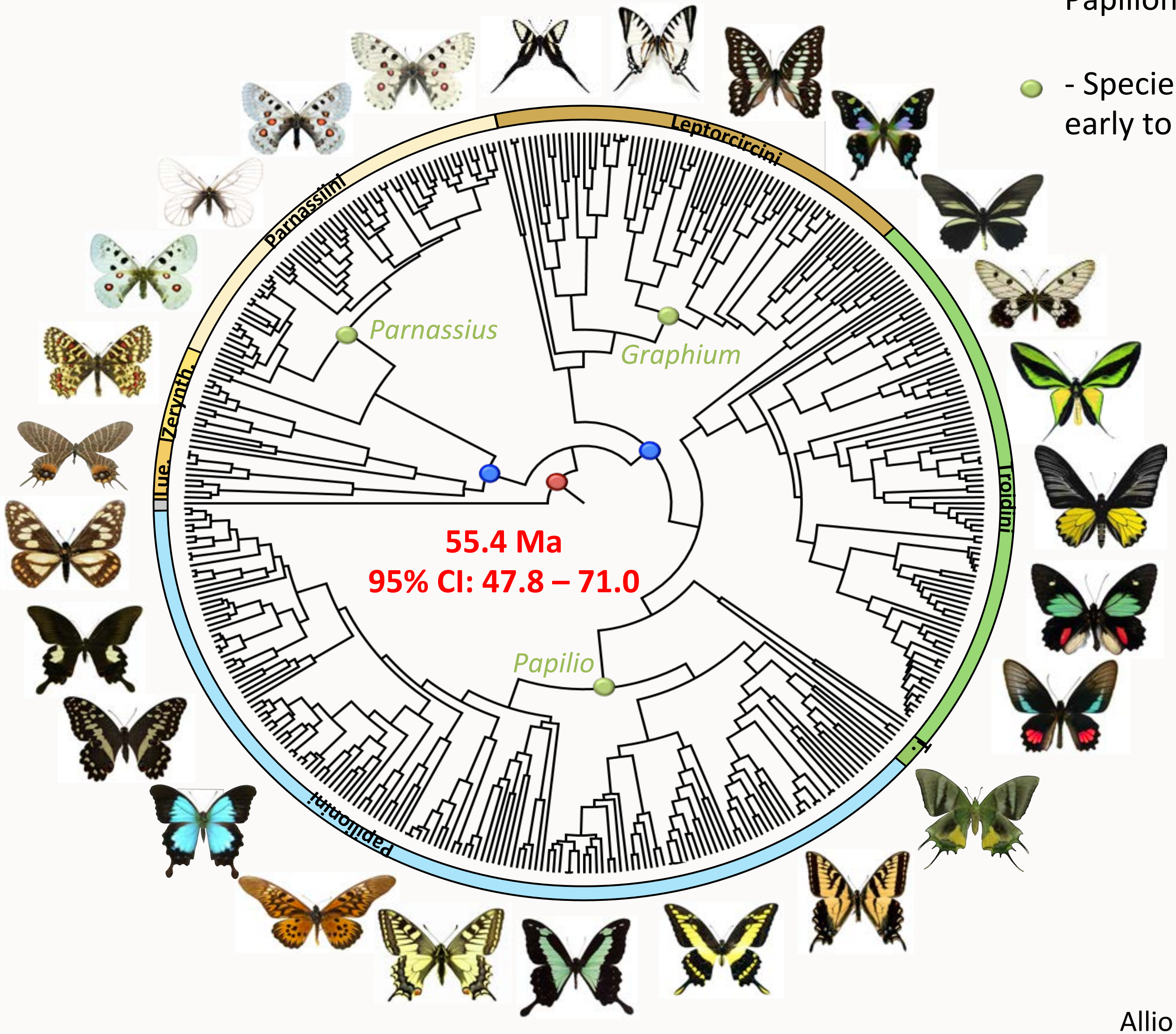
Inferring the most comprehensive and robust phylogeny with Sanger data

BEAST: fix topology / 200 M. MCMC generations

● - Origin in the early Eocene

● - Parnassiinae and Papilioninae are mid-Eocene

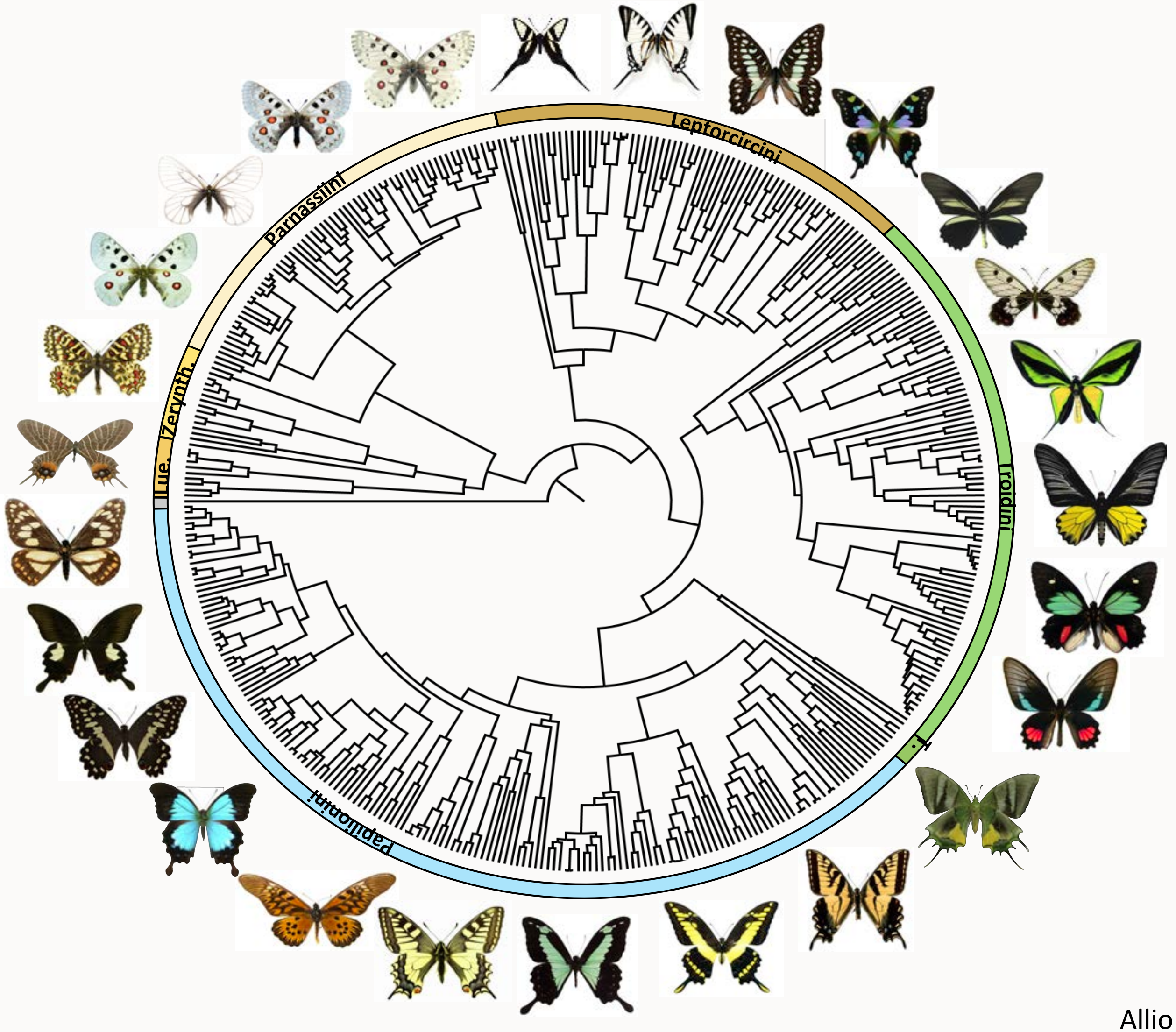
● - Species-rich genera are early to mid-Miocene



55.4 Ma
95% CI: 47.8 – 71.0

Allio *et al.* 2021 – Nat. Comm.

What are the ancestral host plants Papilionidae fed on 55 Ma?



Allio *et al.* 2021 – Nat. Comm.

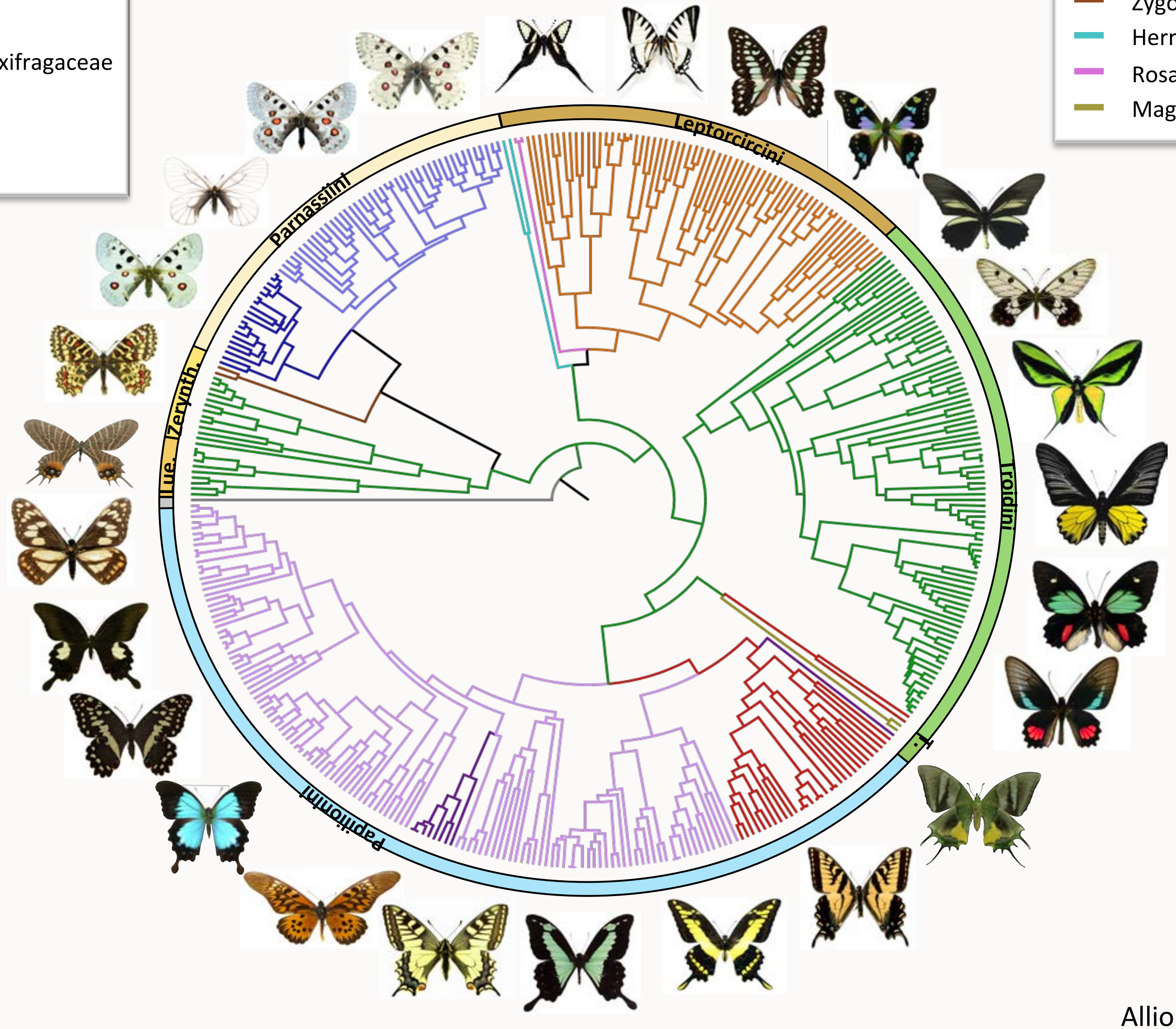
What are the ancestral host plants Papilionidae fed on 55 Ma?

Main host plants

- ▬ Aristolochiaceae
- ▬ Rutaceae
- ▬ Annonaceae
- ▬ Crassulaceae + Saxifragaceae
- ▬ Apiaceae
- ▬ Lauraceae
- ▬ Papaveraceae

Secondary host plants

- ▬ Fabaceae (*Acacia*)
- ▬ Zygophyllaceae
- ▬ Hernandiaceae
- ▬ Rosaceae
- ▬ Magnoliaceae



Allio *et al.* 2021 – Nat. Comm.

Back to the main questions

1. ***What are the ancestral host plants of Papilionidae?***

   ➡ Fed on Aristolochiaceae

2. ***Temporal congruence?***

3. ***Biogeographic congruence?***

4. ***Signatures of genomic adaptations and on speciation rates?***

Back to the main questions

1. Wh

2. Ten

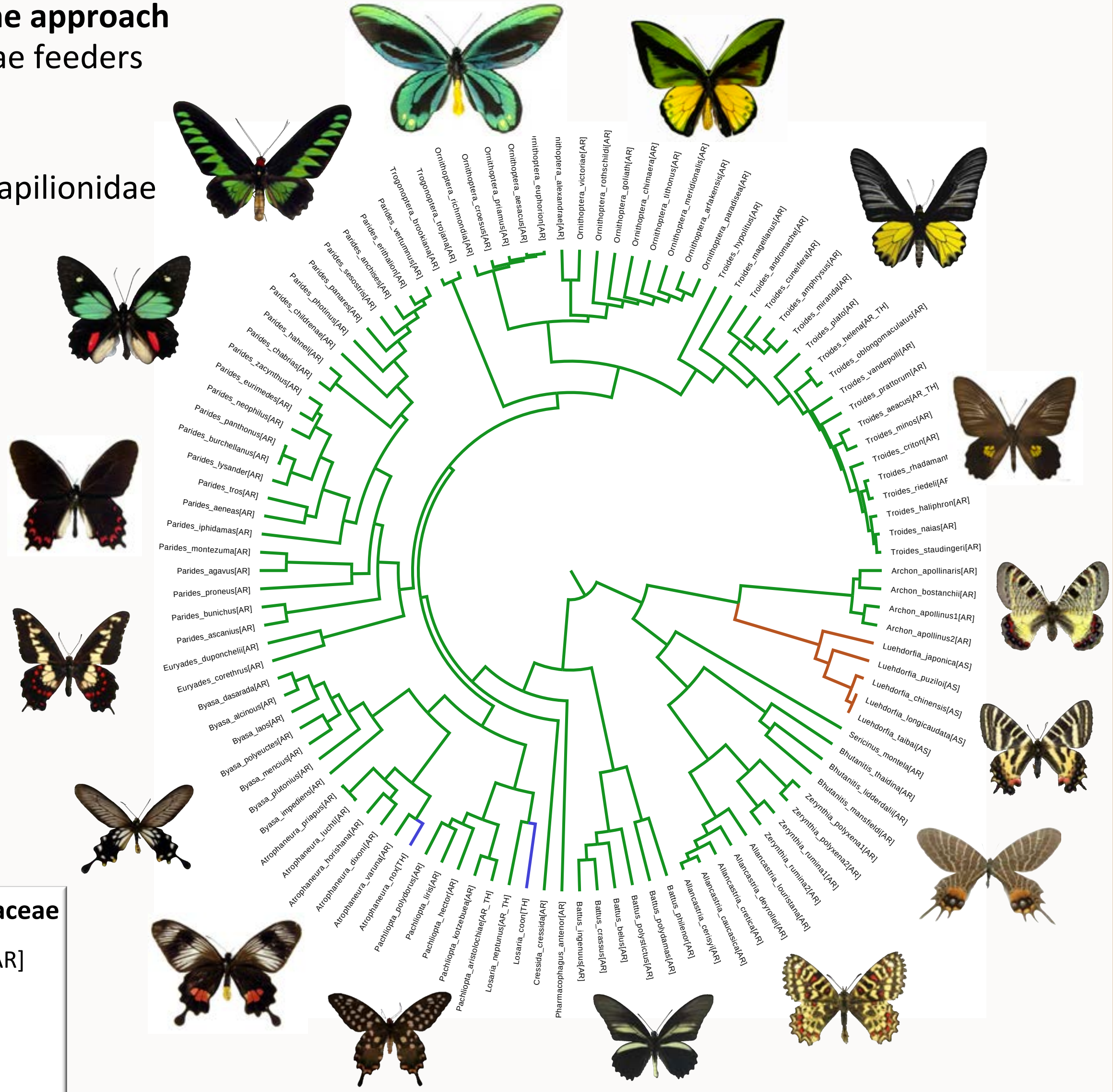
3. Bio

4. Sig

s?

Applying the same approach
to Aristolochiaceae feeders

-> Which genus Papilionidae fed on 55 Ma



Genera of Aristolochiaceae

🟩 *Aristolochia* [AR]

🟫 *Asarum* [AS]

🟦 *Thottea* [TH]

Back to the main questions

1. ***What are the ancestral host plants of Papilionidae?***

   ➡ Fed on Aristolochiaceae, very likely on the genus *Aristolochia*

2. ***Temporal congruence?***

3. ***Biogeographic congruence?***

4. ***Signatures of genomic adaptations and on speciation rates?***

Back to the main questions

1. ***What are the ancestral host plants of Papilionidae?***

   ➡ Fed on Aristolochiaceae, very likely on the genus *Aristolochia*

2. ***Temporal congruence?***

3. ***Biogeographic congruence?***

4. ***Signatures of genomic adaptations and on speciation rates?***

Inferring the most comprehensive and robust phylogeny with Sanger data

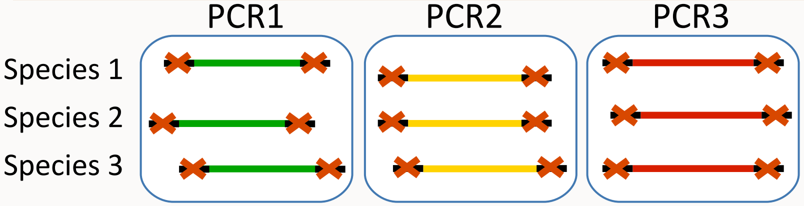
Sanger sequences for **247 species**

Collaborations → with botanists

GenBank data →

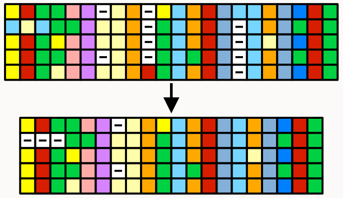
NCBI

Sequence cleaning



Geneious

Alignments

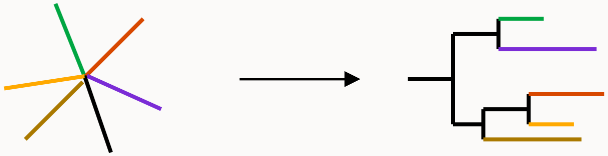


MAFFT

Mesquite

Phylogenetic inferences

Dataset: *- 247 species (+ 6 outgroups)*
- 5 genes ; 4 377 nucleotides



IQ-TREE

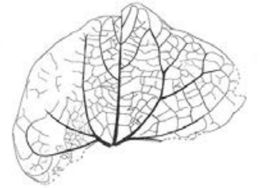
MrBayes

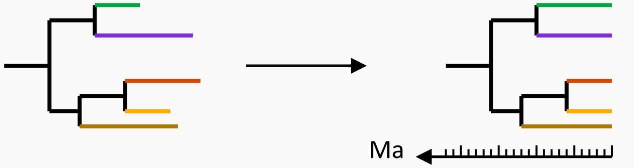
Molecular dating

4 fossils

3 ingroup fossils

1 outgroup fossil

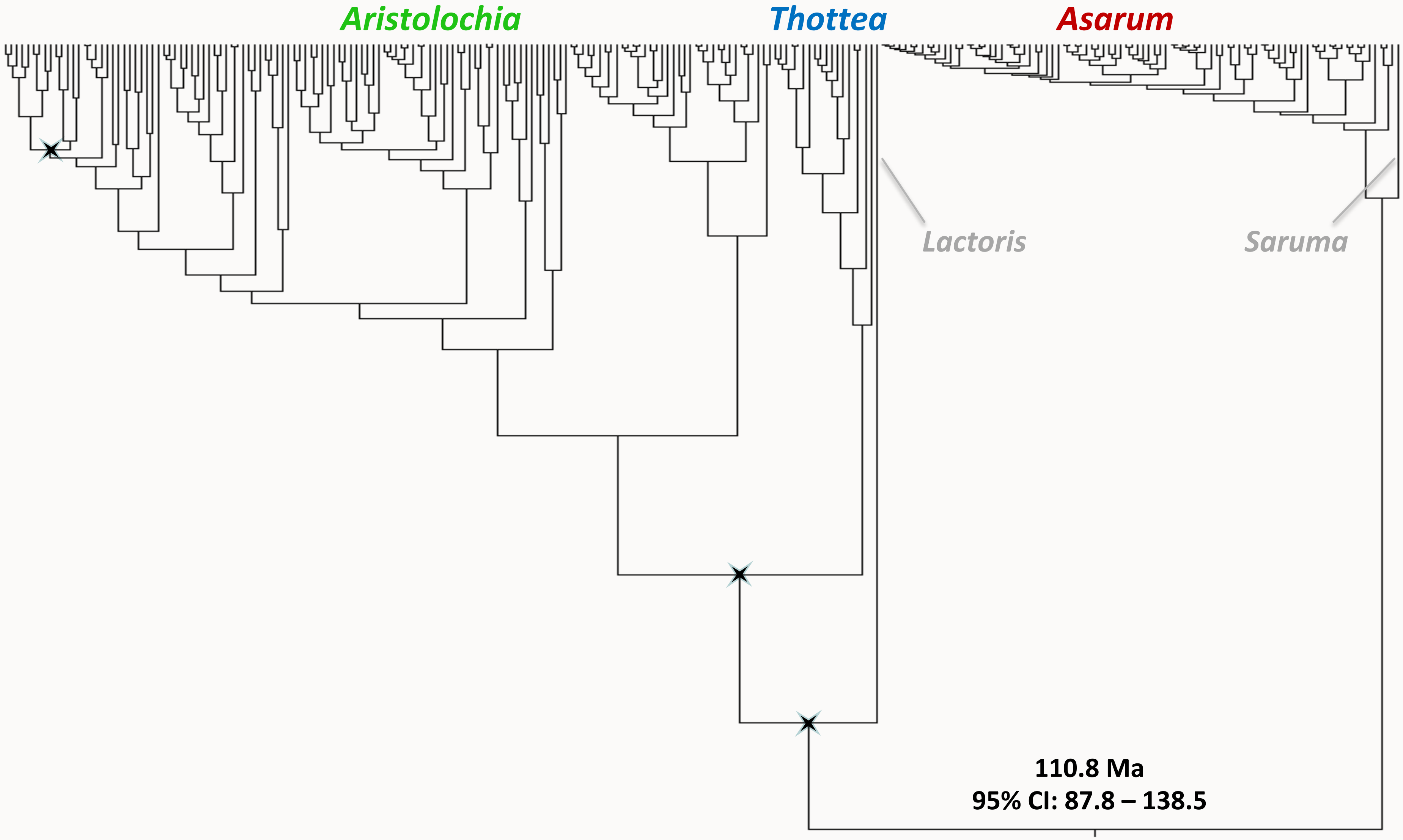




BEAST

Allio *et al.* 2021 – Nat. Comm.

Time-calibrated phylogeny of Aristolochiaceae

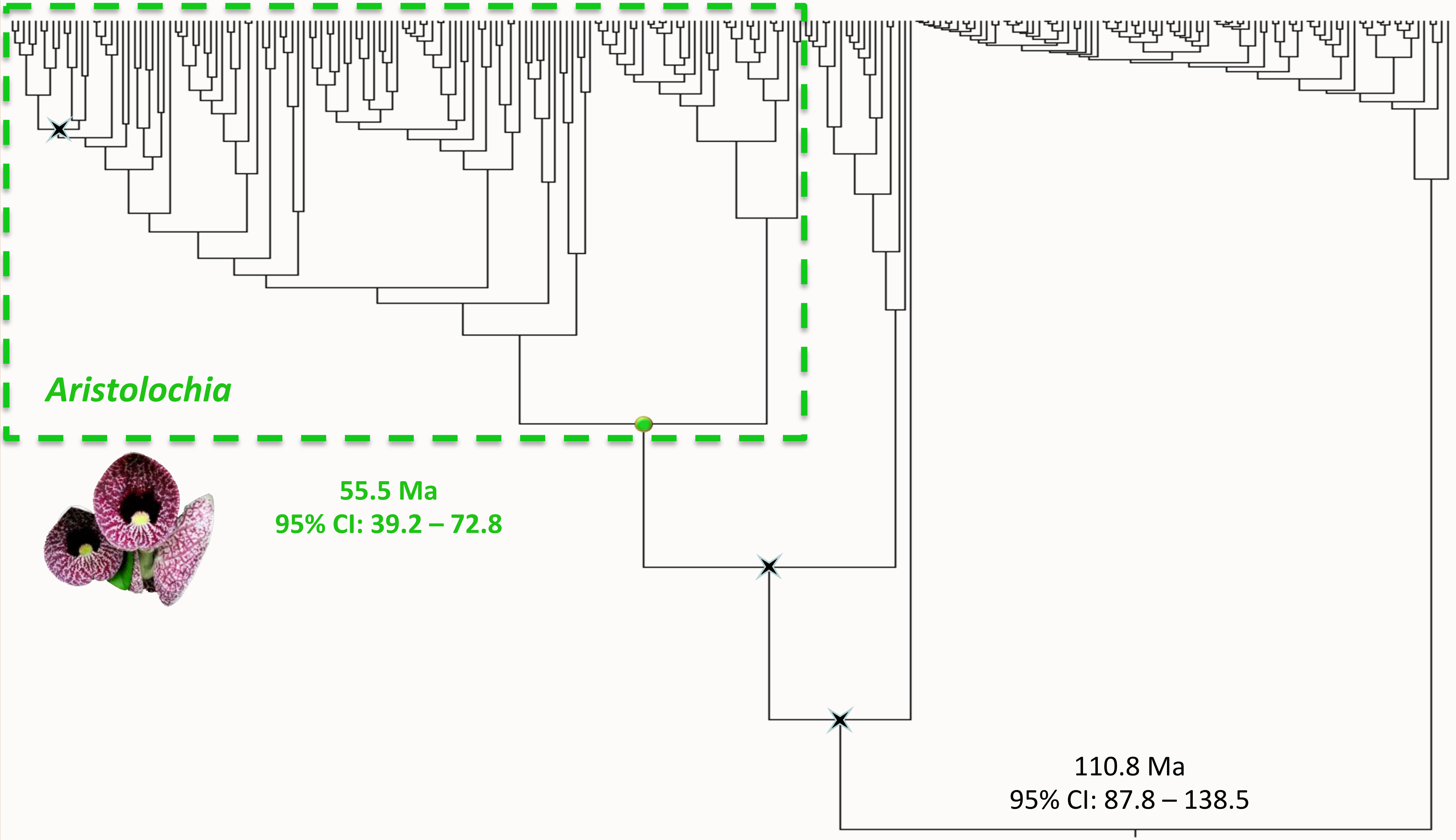


BEAST 200 M. MCMC generations

✕ = Fossil calibrations

Allio *et al.* 2021 – Nat. Comm.

Time-calibrated phylogeny of Aristolochiaceae

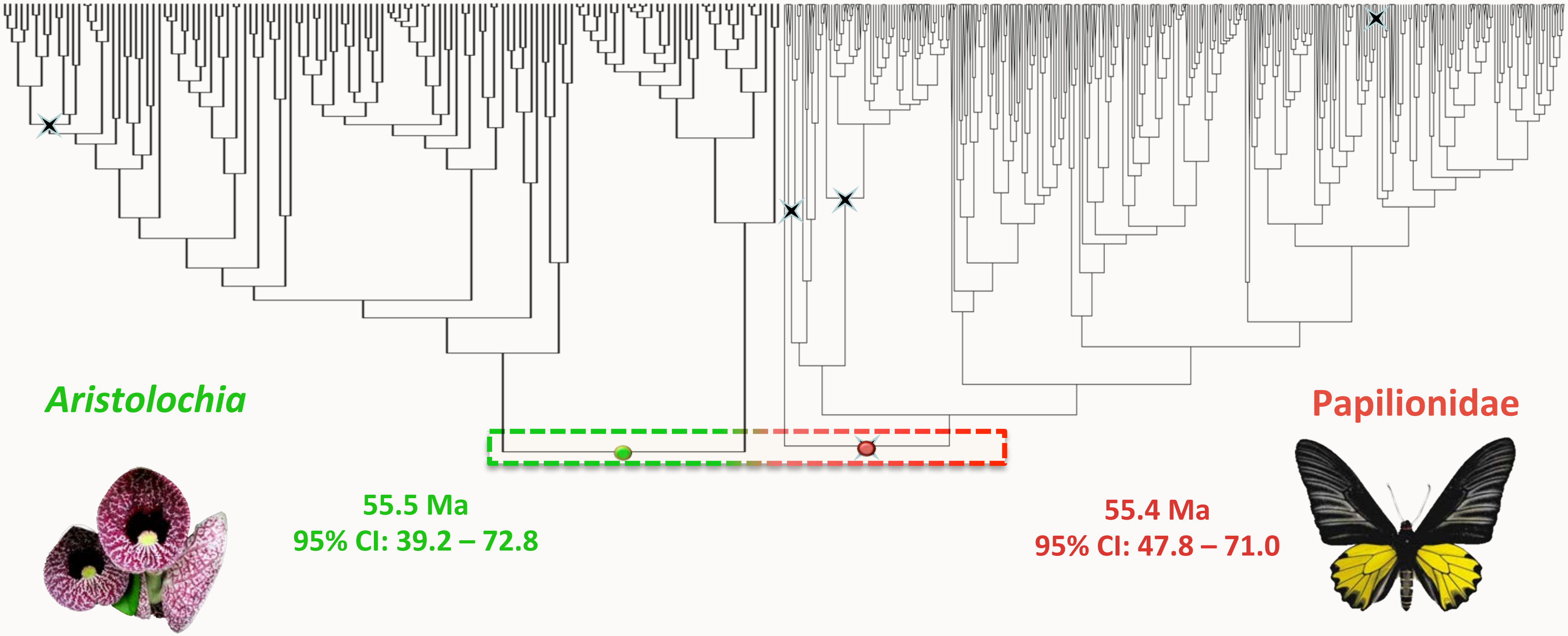


BEAST 200 M. MCMC generations

✕ = Fossil calibrations

Allio *et al.* 2021 – Nat. Comm.

Comparing herbivore and host plant origins



Aristolochia

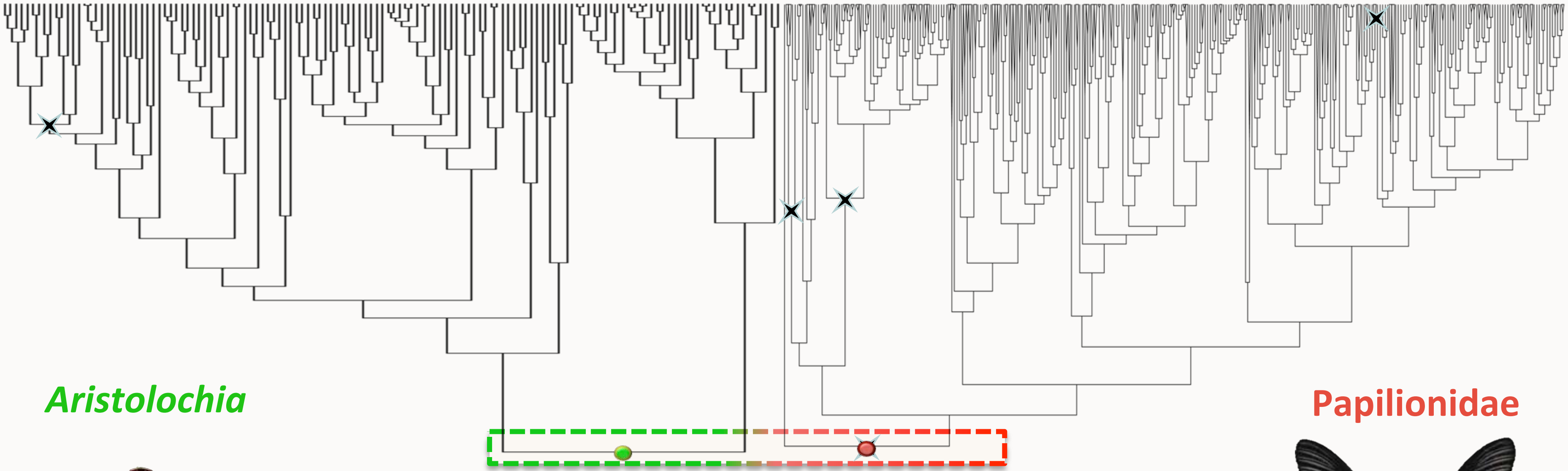
55.5 Ma
95% CI: 39.2 – 72.8

Papilionidae

55.4 Ma
95% CI: 47.8 – 71.0

Allio *et al.* 2021 – Nat. Comm.

Comparing herbivore and host plant origins

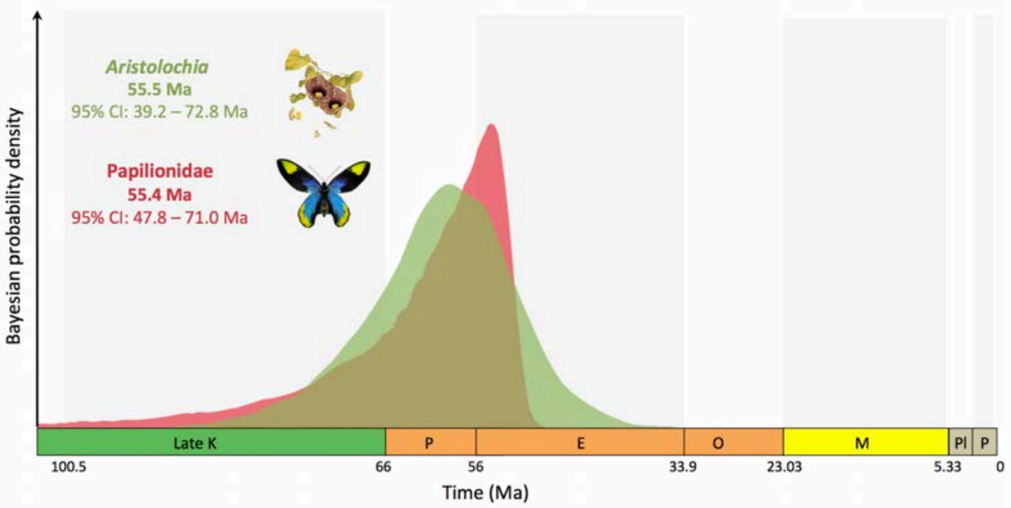


Aristolochia

55.5 Ma
95% CI: 39.2 – 72.8

Papilionidae

55.4 Ma
95% CI: 47.8 – 71.0



Allio *et al.* 2021 – Nat. Comm.

Back to the main questions

1. ***What are the ancestral host plants of Papilionidae?***

   ➡ Fed on Aristolochiaceae, very likely on the genus *Aristolochia*

2. ***Temporal congruence?***

   ➡ Both Papilionidae and *Aristolochia* originated ≈55 Ma

3. ***Biogeographic congruence?***

4. ***Signatures of genomic adaptations and on speciation rates?***

Back to the main questions

1. ***What are the ancestral host plants of Papilionidae?***

   ➡ Fed on Aristolochiaceae, very likely on the genus *Aristolochia*

2. ***Temporal congruence?***

   ➡ Both Papilionidae and *Aristolochia* originated ≈55 Ma

3. ***Biogeographic congruence?***

4. ***Signatures of genomic adaptations and on speciation rates?***

Inferring the biogeographic history of both groups

Distribution data for 408 spp. of Papilionidae and 247 spp. of Aristolochiaceae

12 regions

Time-stratified geographic model

Present

Miocene

Oligocene

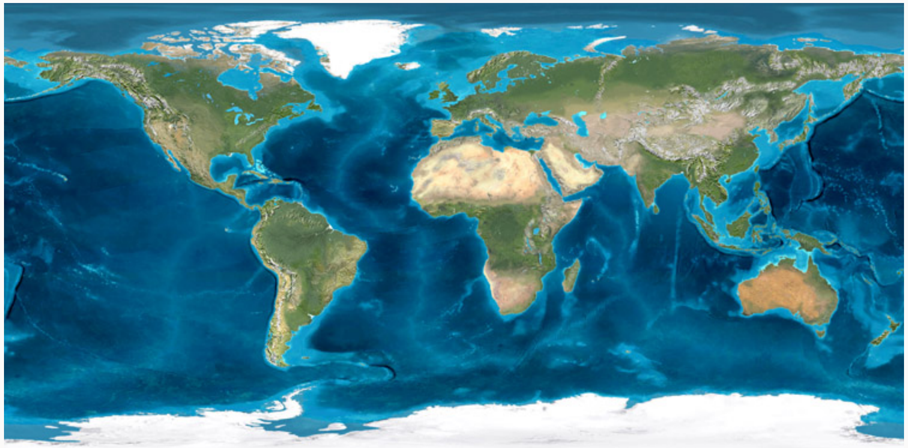
Eocene

Cretaceous-Tertiary boundary

Late Createaceous

Middle Createaceous

Early Createaceous

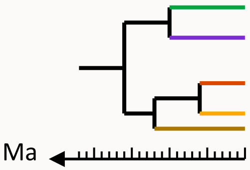


Adjacency matrices

and

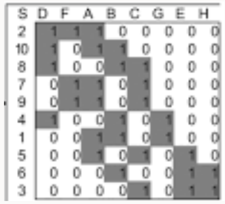
Dispersal matrices

Biogeographic (DEC) analyses



+

+

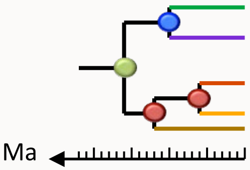


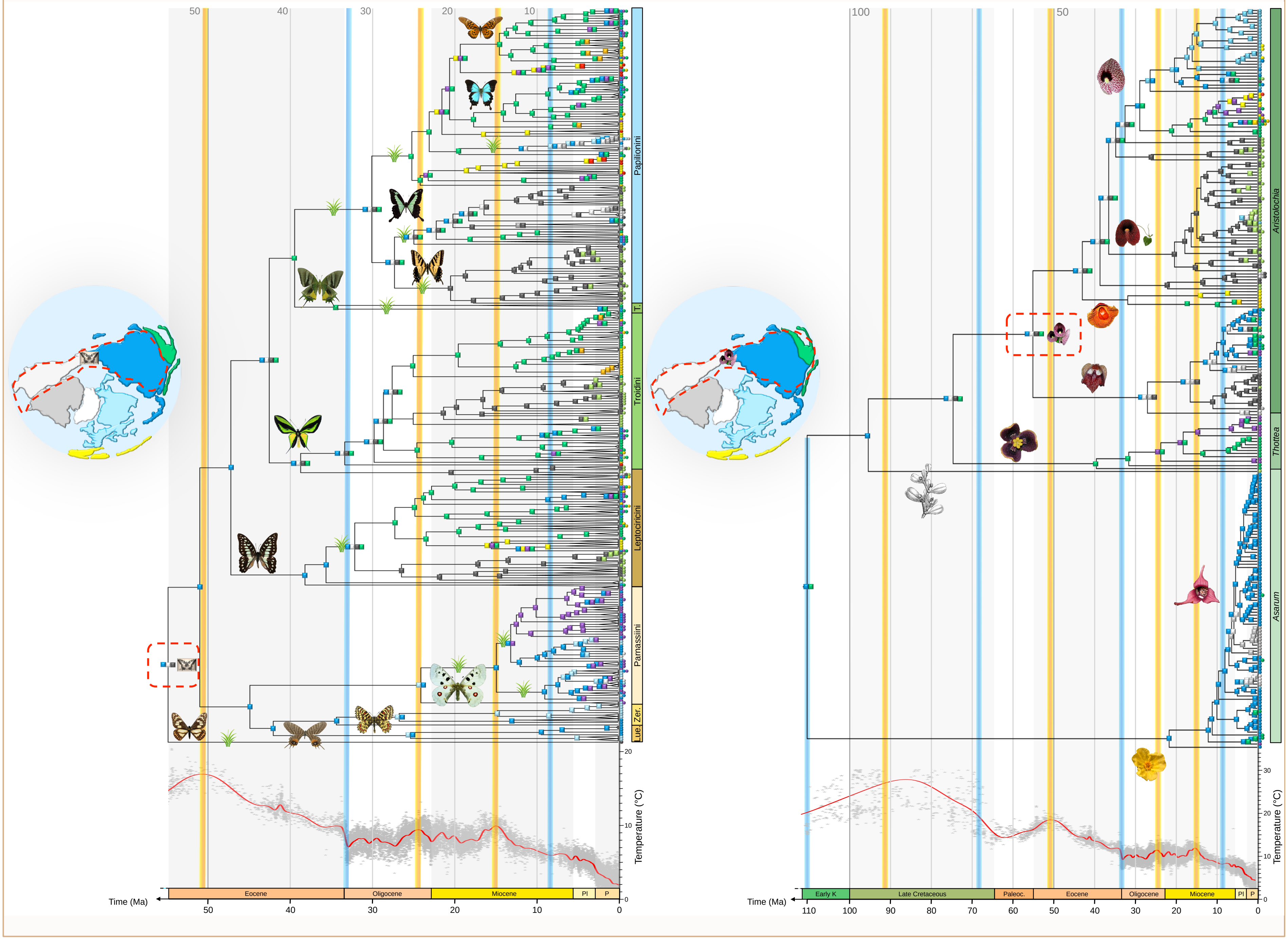
S	D	F	A	B	C	G	E	H
2	1	1	1	0	0	0	0	0
10	1	0	1	1	0	0	0	0
8	1	0	0	1	1	0	0	0
7	0	1	1	0	1	0	0	0
9	0	1	1	0	1	0	0	0
4	1	0	0	1	0	1	0	0
1	0	0	1	1	0	1	0	0
5	0	0	1	0	1	0	1	0
6	0	0	0	1	0	0	1	1
3	0	0	0	0	1	0	1	1

DEC

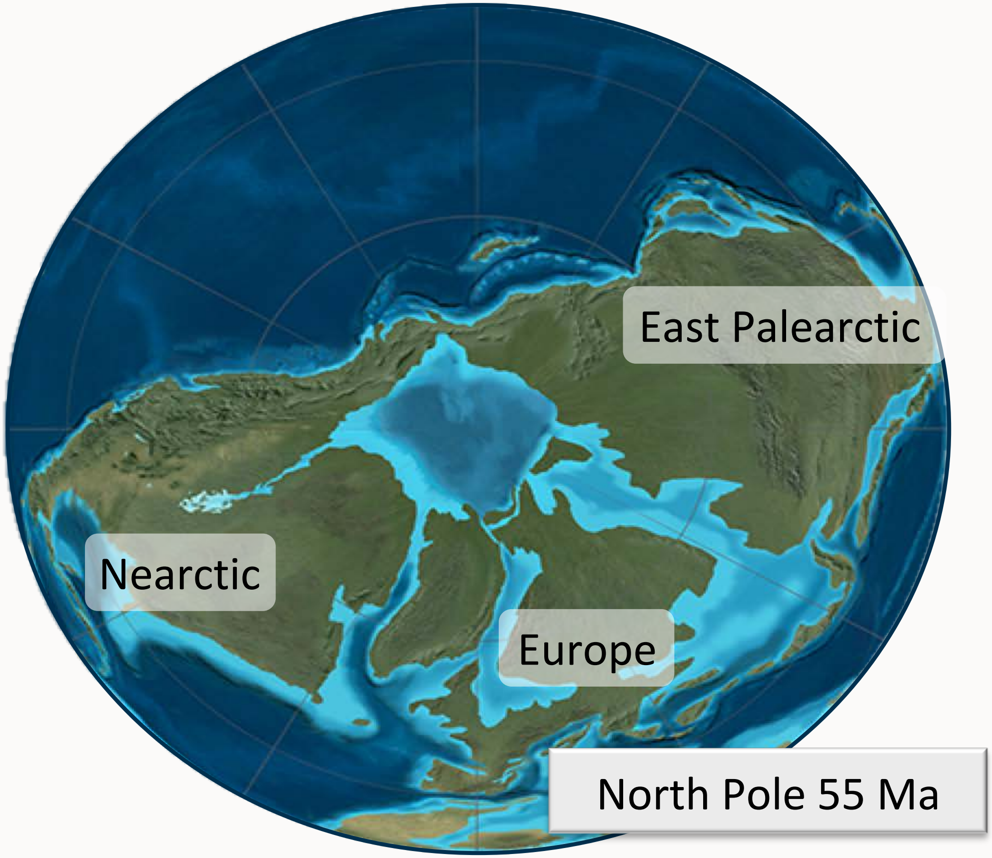
→

models



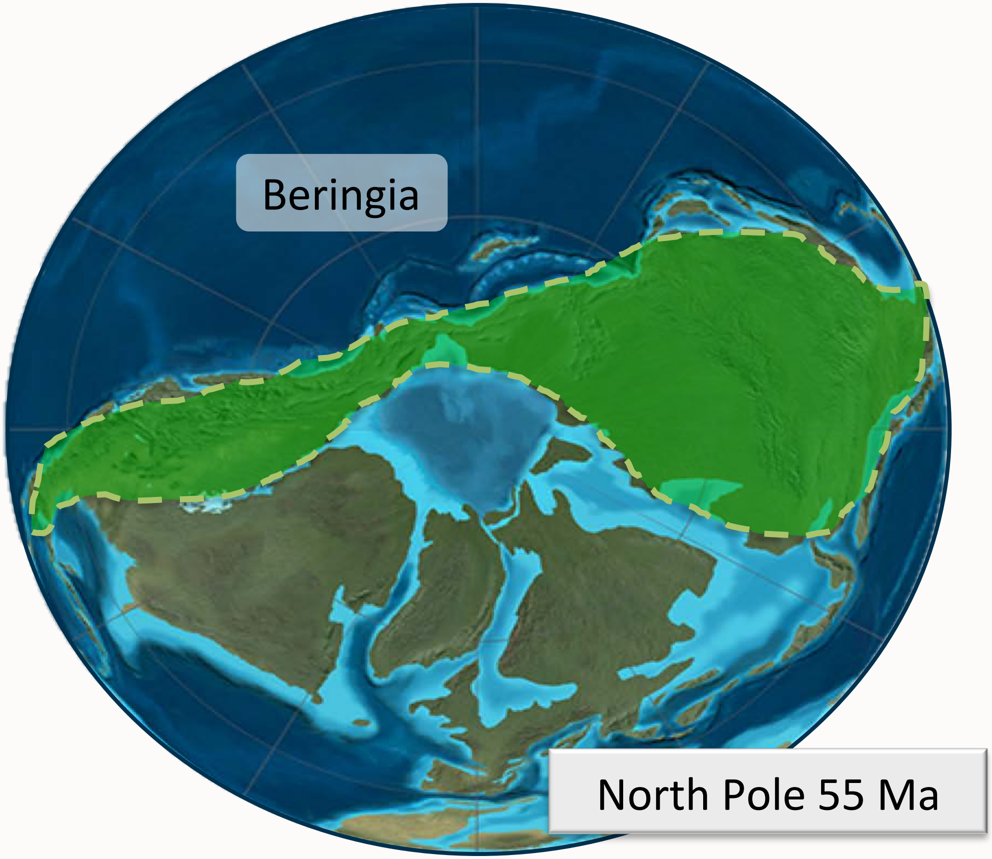

50
40
30
20
10
Papilionini
T.
Troidini
Leptocircini
Parnassiini
Zer.
Lue.
20
10
0
Temperature (°C)
Time (Ma)
Eocene
Oligocene
Miocene
Pl
P
50
40
30
20
10
0
100
50
Aristolochia
Thottea
Asarum
30
20
10
0
Temperature (°C)
Time (Ma)
Early K
Late Cretaceous
Paleoc.
Eocene
Oligocene
Miocene
Pl
P
110
100
90
80
70
60
50
40
30
20
10
0

Both groups appeared at the same time and same place



Allio *et al.* 2021 – Nat. Comm.

Both groups appeared at the same time and same place

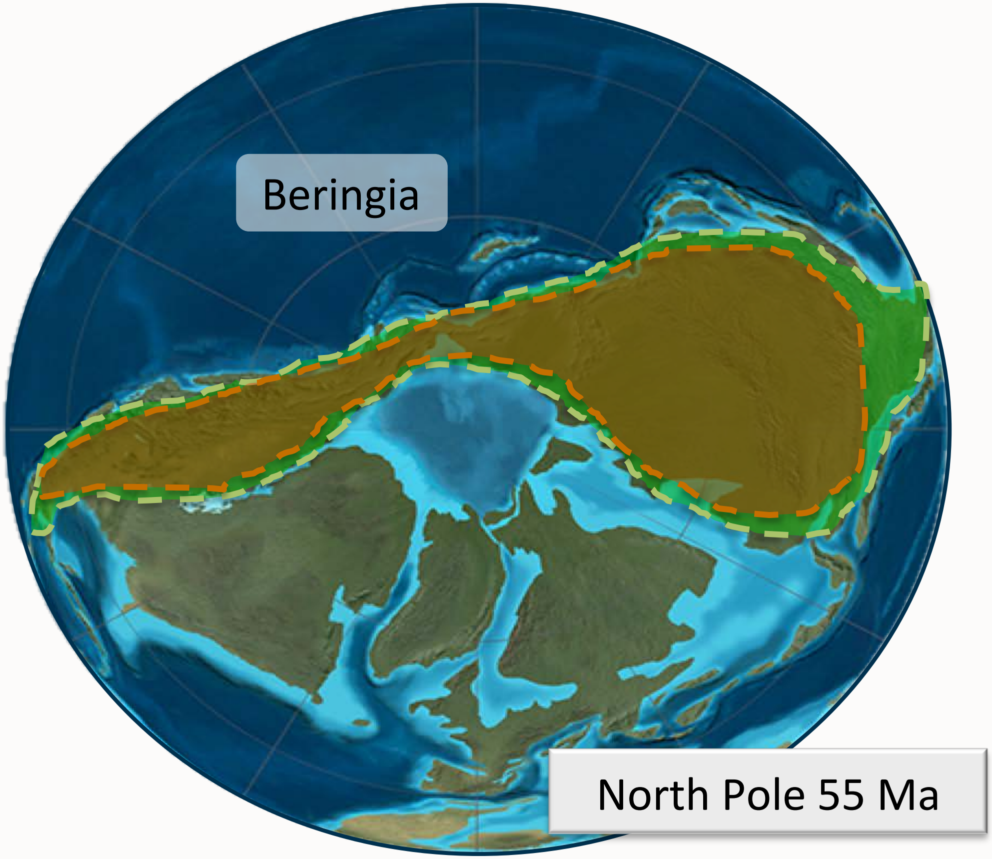


Ancestral area of *Aristolochia*

Central America + West Nearctic + East Palearctic + North Indonesia

Allio *et al.* 2021 – Nat. Comm.

Both groups appeared at the same time and same place



Ancestral area of *Aristolochia*

Central America + West Nearctic + East Palearctic + North Indonesia

Ancestral area of Papilionidae

Central America + West Nearctic + East Palearctic

Allio *et al.* 2021 – Nat. Comm.

Back to the main questions

1. ***What are the ancestral host plants of Papilionidae?***

   ➡ Fed on Aristolochiaceae, very likely on the genus *Aristolochia*

2. ***Temporal congruence?***

   ➡ Both Papilionidae and *Aristolochia* originated ≈55 Ma

3. ***Biogeographic congruence?***

   ➡ Both Papilionidae and *Aristolochia* originated ≈55 Ma in Beringia

4. ***Signatures of genomic adaptations and on speciation rates?***

Back to the main questions

1. ***What are the ancestral host plants of Papilionidae?***

   ➡ Fed on Aristolochiaceae, very likely on the genus *Aristolochia*

2. ***Temporal congruence?***

   ➡ Both Papilionidae and *Aristolochia* originated ≈55 Ma

3. ***Biogeographic congruence?***

   ➡ Both Papilionidae and *Aristolochia* originated ≈55 Ma in Beringia

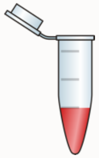
4. ***Signatures of genomic adaptations and on speciation rates?***

Genomic dataset

Shotgun sequencing for **41 species**

DNA extractions

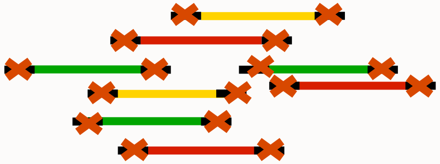
Library preparations



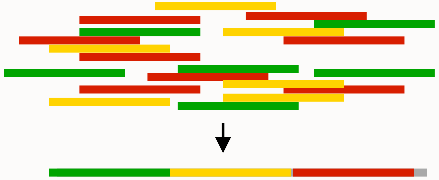
Illumina NextSeq

Reads cleaning

Shotgun assembly

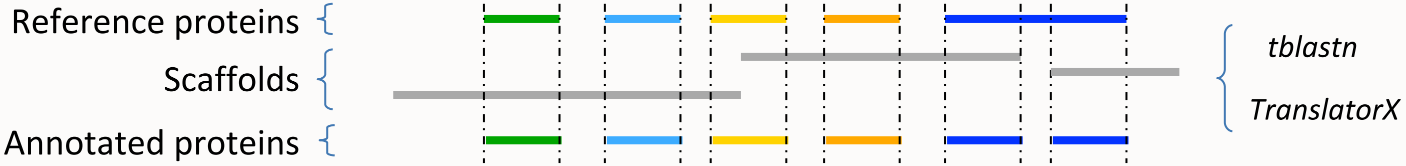


Trimmomatic



SOAP-denovo

BLAST Alignment Consensus



Reference proteins

Scaffolds

Annotated proteins

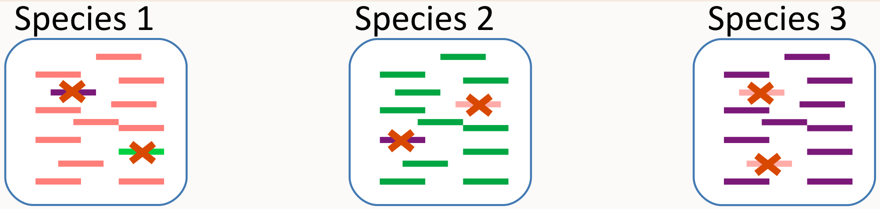
tblastn

TranslatorX

Allio *et al.* 2020 – Syst. Biol.

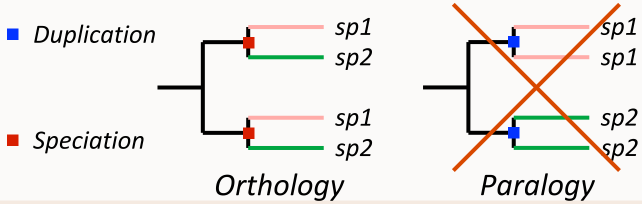
Inferring a first phylogenomic backbone with NGS data

Check for cross-contamination



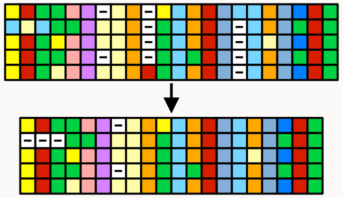
CroCo

Detections of orthologous genes



OrthoFinder

Alignment cleaning



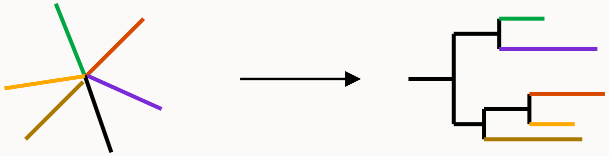
HMMcleaner

TrimAl

Phylogenetic inferences

Dataset 1: *760 genes ; 288 446 AA*
12% missing data

Dataset 2: 6621 *genes ; 1 656 028 AA*
32% missing data



IQ-TREE

PhyloBayes

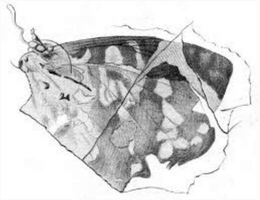
Supertree approach

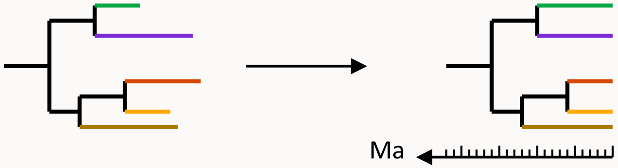
Molecular dating

5 fossils

3 ingroup

2 outgroup





PhyloBayes

Allio *et al.* 2020 – Syst. Biol.

Inferring a first phylogenomic backbone with NGS data

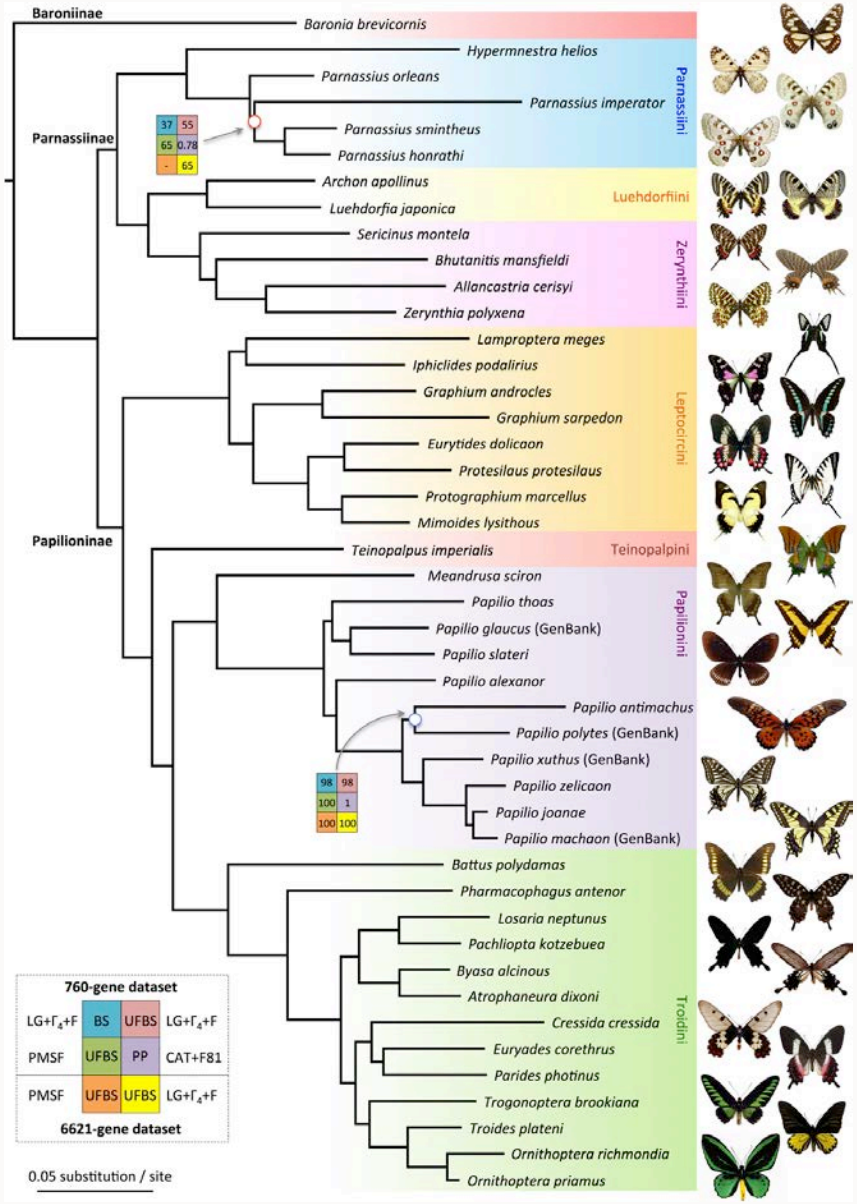
IQ-TREE many tests with:
- 100 thorough bootstraps
- 1000 ultrafast bootstraps
- Protein models: LG+Γ4+F / C50 / PMSF

PhyloBayes CAT model

All nodes but one had BV = 100 or PP = 1

All subfamilies, tribes, genera are monophyletic

Low-quality libraries are well placed (but long branches)



Taxon sampling = 45 Papilionidae + 16 outgroups

Allio *et al.* 2020 – Syst. Biol.

Inferring a first phylogenomic backbone with NGS data

IQ-TREE many tests with:
- 100 thorough bootstraps
- 1000 ultrafast bootstraps
- Protein models: LG+Γ4+F / C50 / PMSF

PhyloBayes CAT model

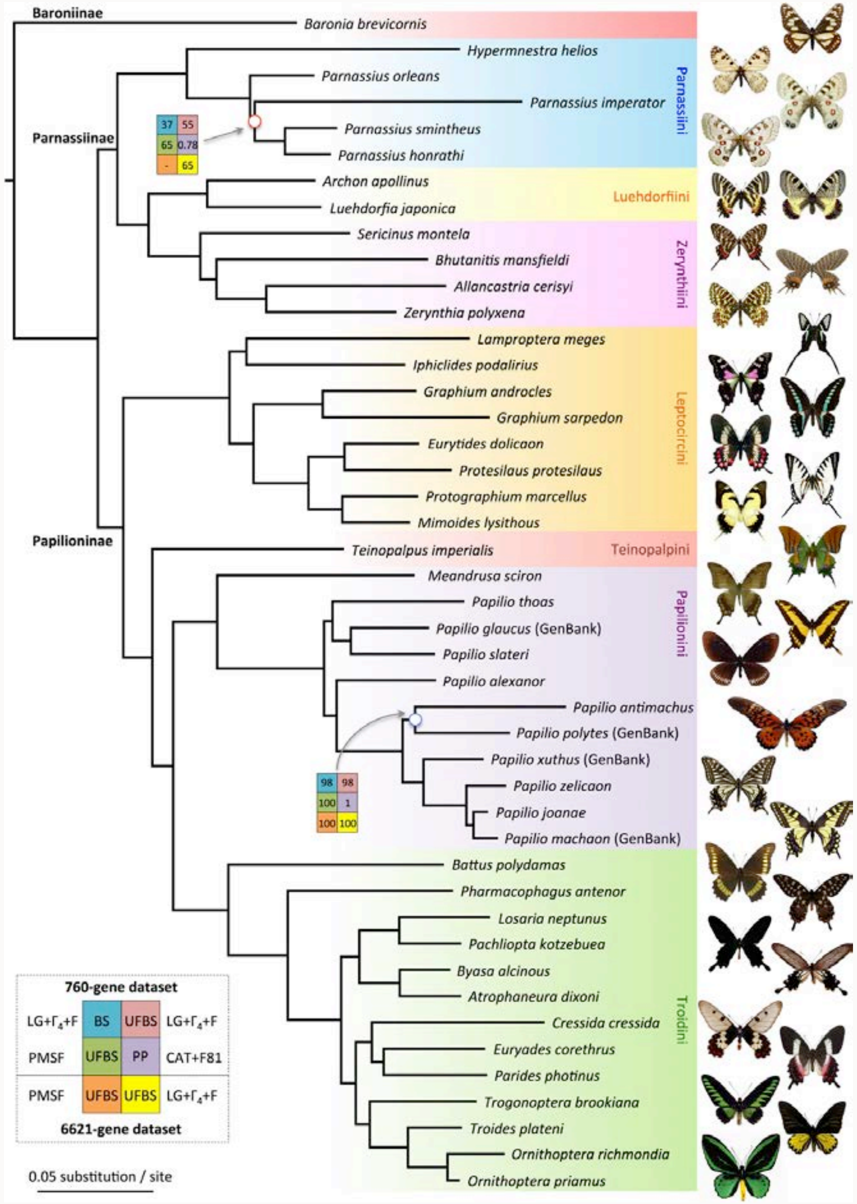
All nodes but one had BV = 100 or PP = 1

All subfamilies, tribes, genera are monophyletic

Low-quality libraries are well placed (but long branches)

Signatures of genomic adaptations?

Taxon sampling = 45 Papilionidae + 16 outgroups



Allio *et al.* 2020 – Syst. Biol.

Are the differences between the sequences due to natural selection?

1. Studying the substitutions in the amino acid sequences with:

- **Synonymous substitutions** = distance between the sequences (ignoring selection)

  ↳ *dS* = number of substitutions in synonymous codons

- **Non-synonymous substitutions** = either positive selection or drift

  ↳ *dN* = number of substitutions in non-synonymous codons

Kimura 1983, Yang 2006, Zhang *et al.* 2005 – Mol. Biol. Evol

Are the differences between the sequences due to natural selection?

1. Studying the substitutions in the amino acid sequences with:

- **Synonymous substitutions** = distance between the sequences (ignoring selection)

  ↳ *dS* = number of substitutions in synonymous codons

- **Non-synonymous substitutions** = either positive selection or drift

  ↳ *dN* = number of substitutions in non-synonymous codons

2. Ratio to infer the direction and magnitude of natural selection acting on protein-coding genes:

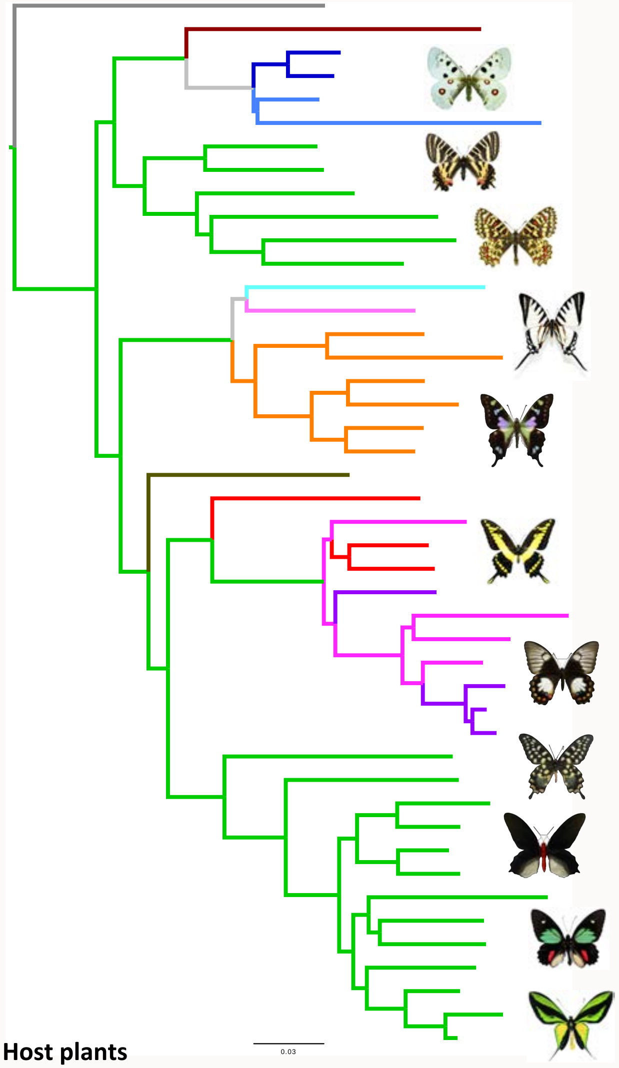
$$\omega = \frac{d_N}{d_S} = \frac{selection + neutral}{neutral}$$

$\omega = 1$: no (neutral) selection

$\omega > 1$: positive selection, it implies positive or Darwinian selection (driving change)

$\omega < 1$: negative selection, it implies purifying or stabilizing selection (acting against change)

Kimura 1983, Yang 2006, Zhang *et al.* 2005 – Mol. Biol. Evol

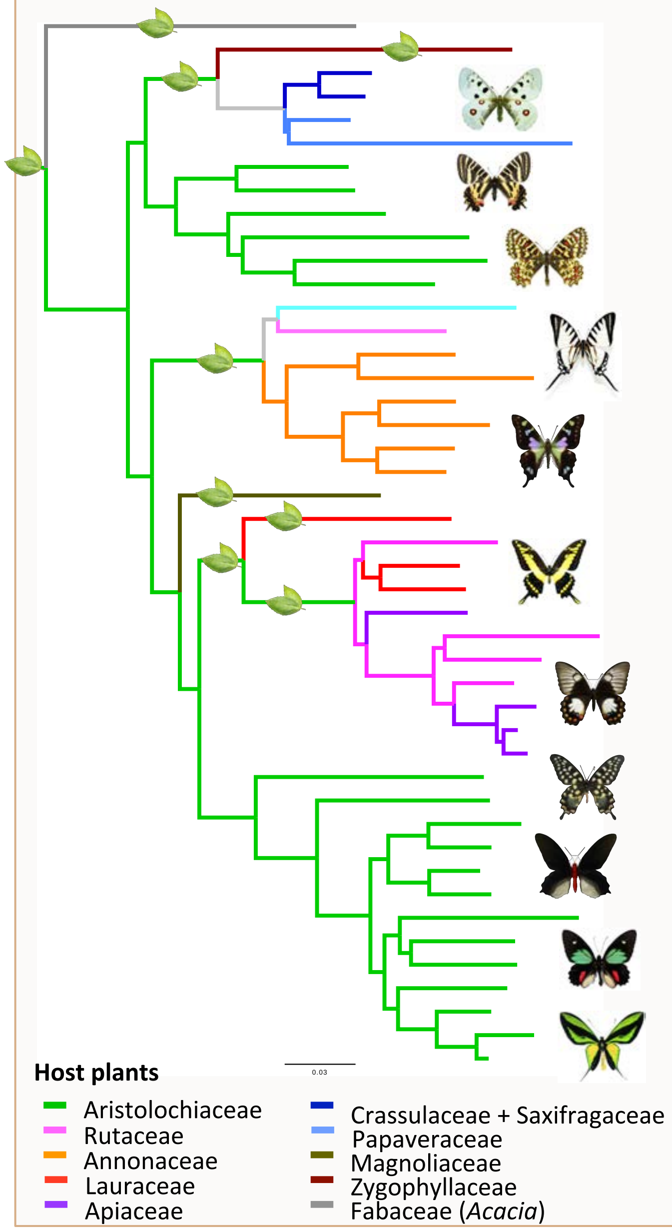


Host plants

- 🟩 Aristolochiaceae
- 🟪 Rutaceae
- 🟧 Annonaceae
- 🟥 Lauraceae
- 🟪 Apiaceae
- 🟦 Crassulaceae + Saxifragaceae
- 🟦 Papaveraceae
- 🟫 Magnoliaceae
- 🟥 Zygophyllaceae
- ⬜ Fabaceae (*Acacia*)

Allio *et al.* 2021 – Nat. Comm.

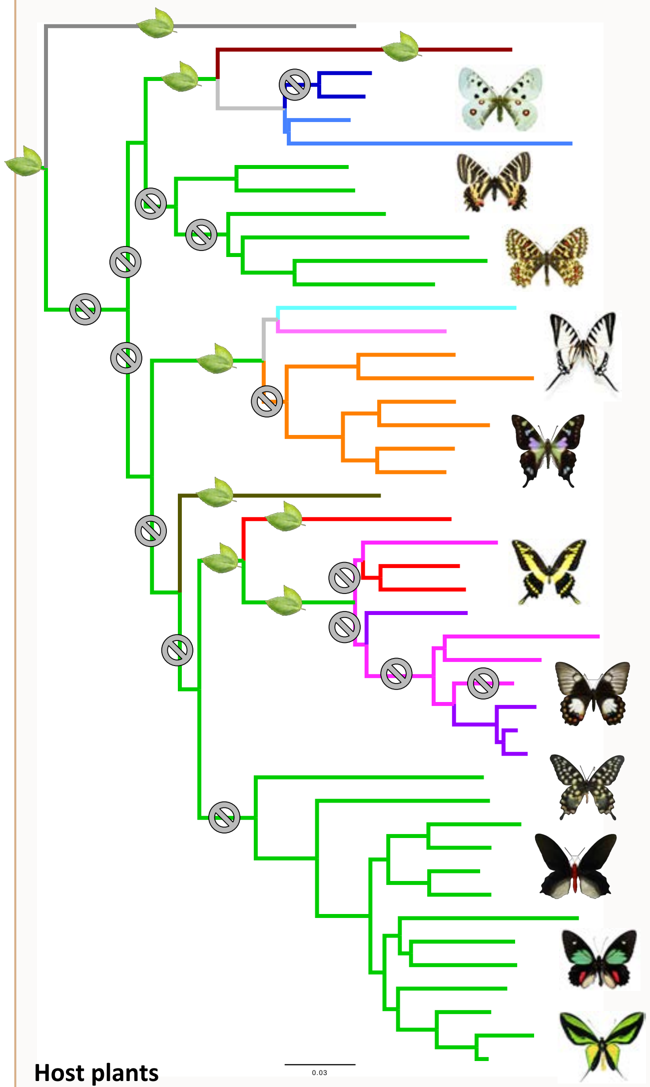
Host-plant shift



Allio *et al.* 2021 – Nat. Comm.

🍃 **Host-plant** shift

🚫 No change



Host plants

🟩 Aristolochiaceae
🟪 Rutaceae
🟧 Annonaceae
🟥 Lauraceae
🟪 Apiaceae

🟦 Crassulaceae + Saxifragaceae
🟦 Papaveraceae
🟫 Magnoliaceae
🟥 Zygophyllaceae
⬜ Fabaceae (*Acacia*)

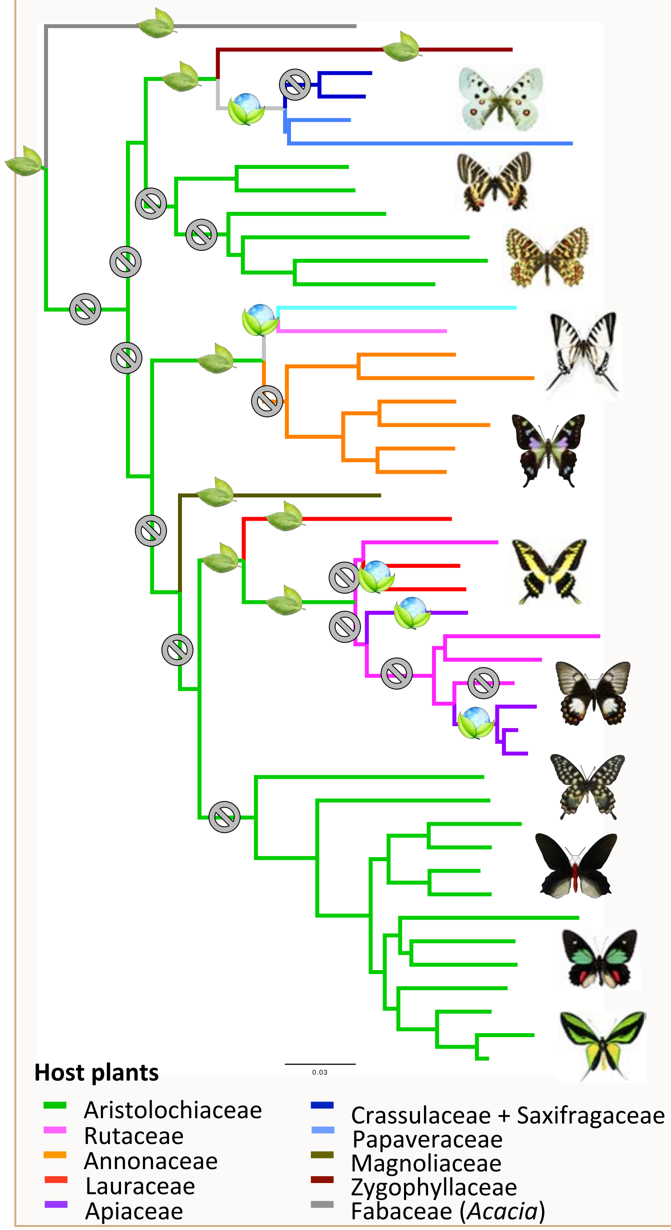
Allio *et al.* 2021 – Nat. Comm.

🍃 **Host-plant** shift

🌍 **Climate** shift

Both **host-plant and climate** shifts

🛇 No change



Host plants

- ■ Aristolochiaceae
- ■ Rutaceae
- ■ Annonaceae
- ■ Lauraceae
- ■ Apiaceae
- ■ Crassulaceae + Saxifragaceae
- ■ Papaveraceae
- ■ Magnoliaceae
- ■ Zygophyllaceae
- ■ Fabaceae (*Acacia*)

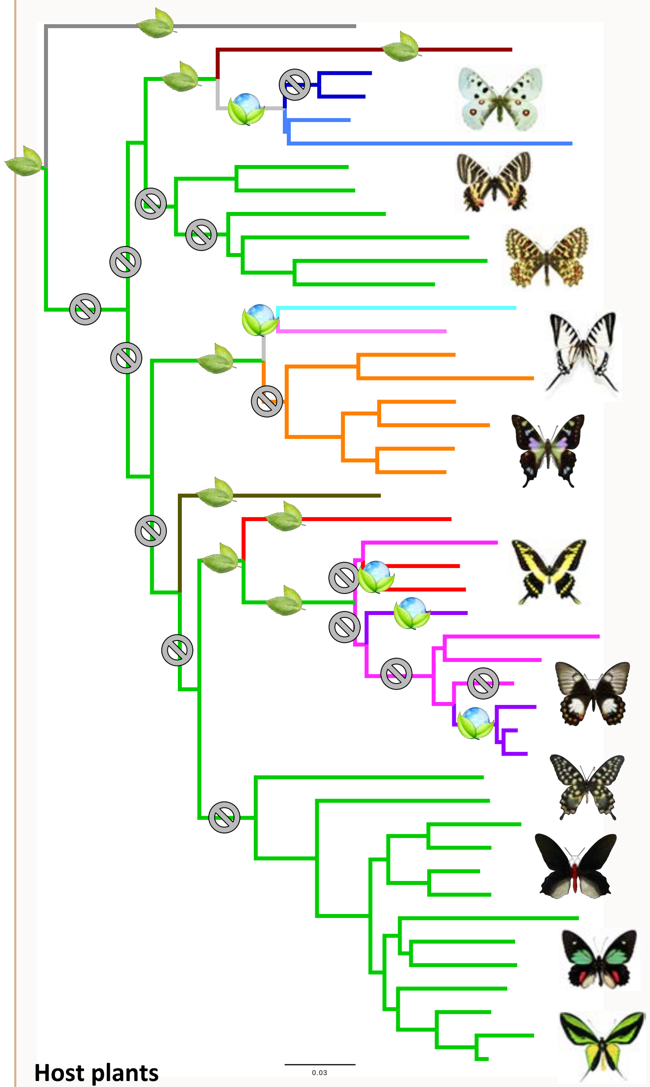
Allio *et al.* 2021 – Nat. Comm.

🍃 **Host-plant** shift

🌍 **Climate** shift

🌍🍃 Both **host-plant and climate** shifts

⊘ No change



Host plants

🟩 Aristolochiaceae
🟪 Rutaceae
🟧 Annonaceae
🟥 Lauraceae
🟪 Apiaceae
🟦 Crassulaceae + Saxifragaceae
🟦 Papaveraceae
🟫 Magnoliaceae
🟥 Zygophyllaceae
⬜ Fabaceae (*Acacia*)

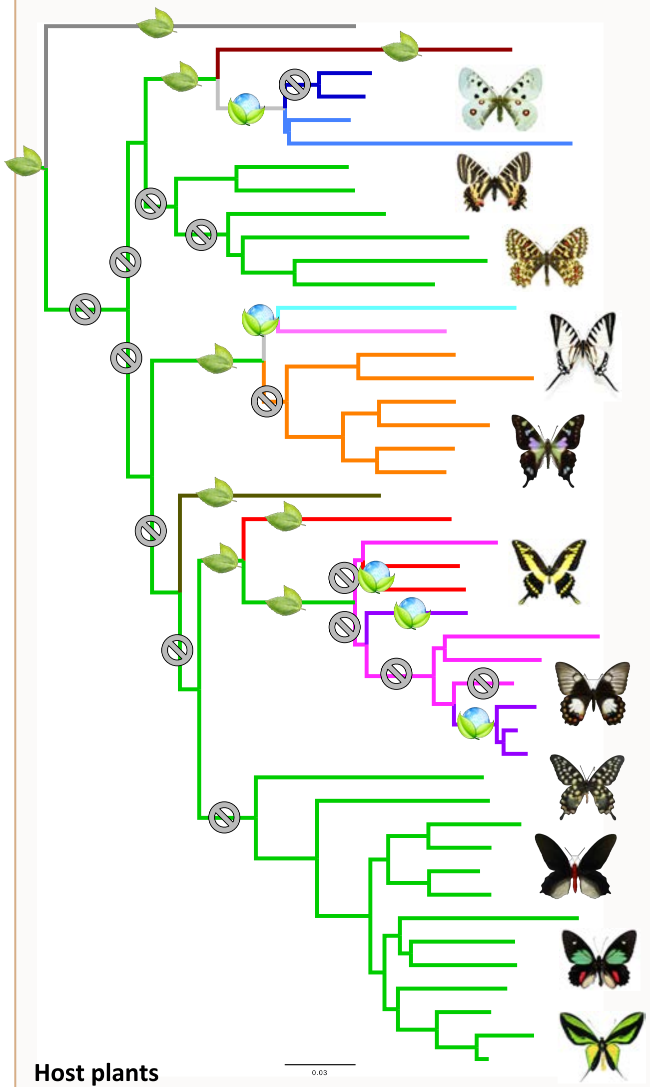
Allio *et al.* 2021 – Nat. Comm.

Host-plant shift

Climate shift

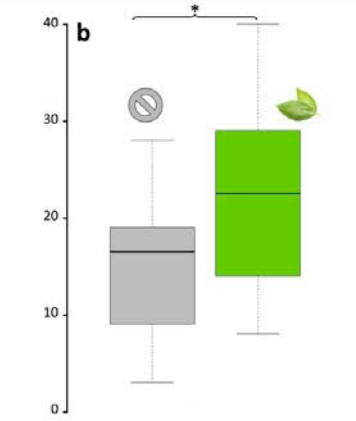
Both **host-plant and climate** shifts

No change



Host plants

- 🟩 Aristolochiaceae
- 🟪 Rutaceae
- 🟧 Annonaceae
- 🟥 Lauraceae
- 🟪 Apiaceae
- 🟦 Crassulaceae + Saxifragaceae
- 🟦 Papaveraceae
- 🟫 Magnoliaceae
- 🟥 Zygophyllaceae
- ⬜ Fabaceae (*Acacia*)



- Branches experiencing host shifts have **MORE genes under positive selection**

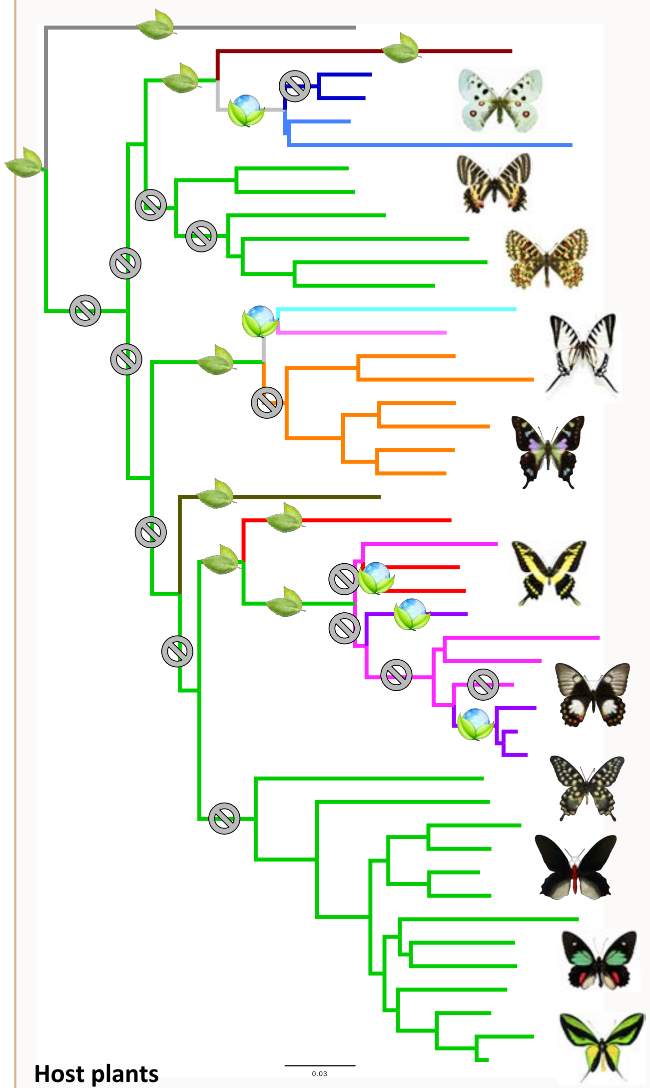
Allio *et al.* 2021 – Nat. Comm.

🍃 **Host-plant** shift

🌍 **Climate** shift

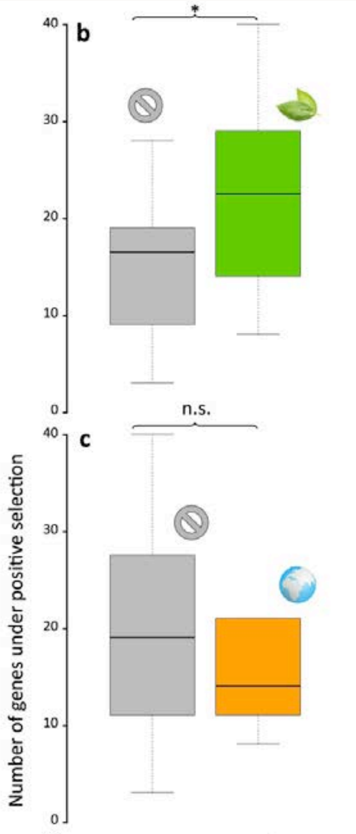
🌍🍃 Both **host-plant and climate** shifts

🚫 No change



Host plants

▬ Aristolochiaceae
▬ Rutaceae
▬ Annonaceae
▬ Lauraceae
▬ Apiaceae
▬ Crassulaceae + Saxifragaceae
▬ Papaveraceae
▬ Magnoliaceae
▬ Zygophyllaceae
▬ Fabaceae (*Acacia*)



- Branches experiencing host shifts have **MORE genes under positive selection**
- Branches experiencing environmental shifts have **NOT more genes under positive selection**

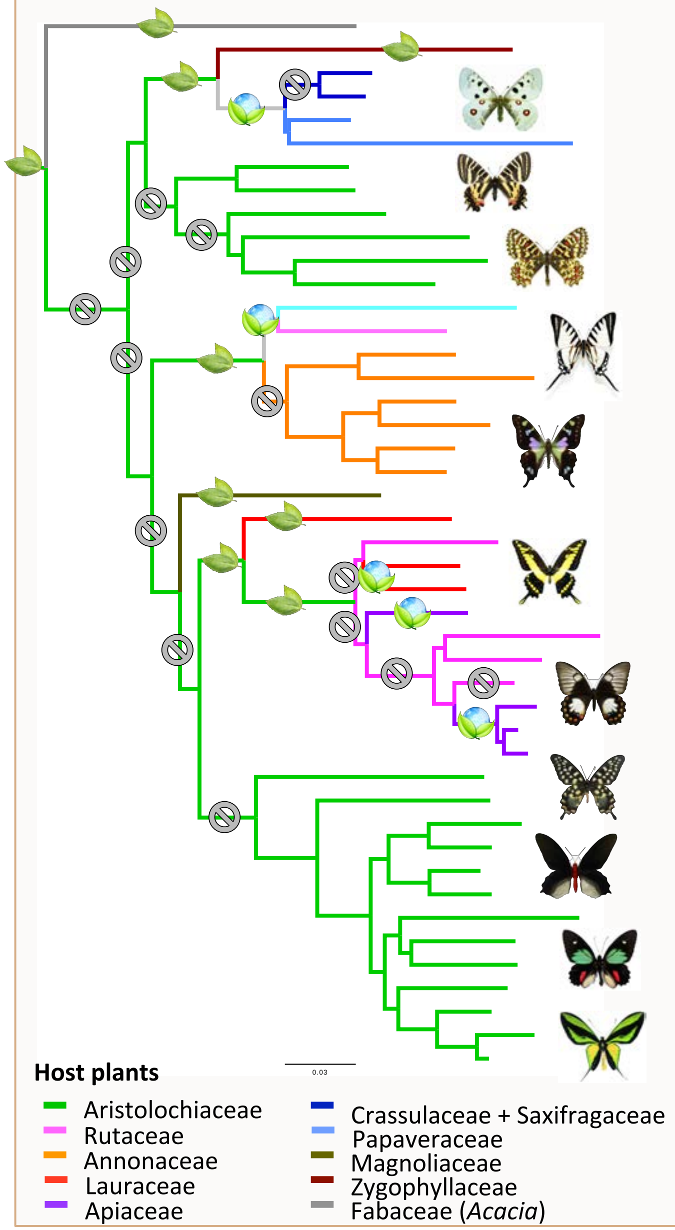
Allio *et al.* 2021 – Nat. Comm.

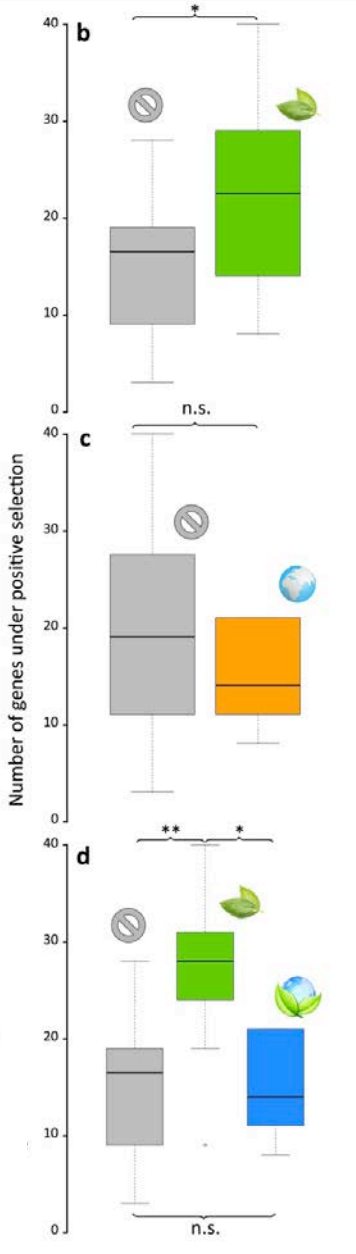
Host-plant shift

Climate shift

Both **host-plant and climate** shifts

No change





- Branches experiencing host shifts have **MORE genes under positive selection**
- Branches experiencing environmental shifts have **NOT more genes under positive selection**
- Branches experiencing both host and environmental shifts have **NOT more genes under positive selection**

Allio *et al.* 2021 – Nat. Comm.

Back to the main questions

1. ***What are the ancestral host plants of Papilionidae?***

   ➡ Fed on Aristolochiaceae, very likely on the genus *Aristolochia*

2. ***Temporal congruence?***

   ➡ Both Papilionidae and *Aristolochia* originated ≈55 Ma

3. ***Biogeographic congruence?***

   ➡ Both Papilionidae and *Aristolochia* originated ≈55 Ma in Beringia

4. ***Signatures of genomic adaptations and on speciation rates?***

   ➡ Genome-wide adaptations

Back to the main questions

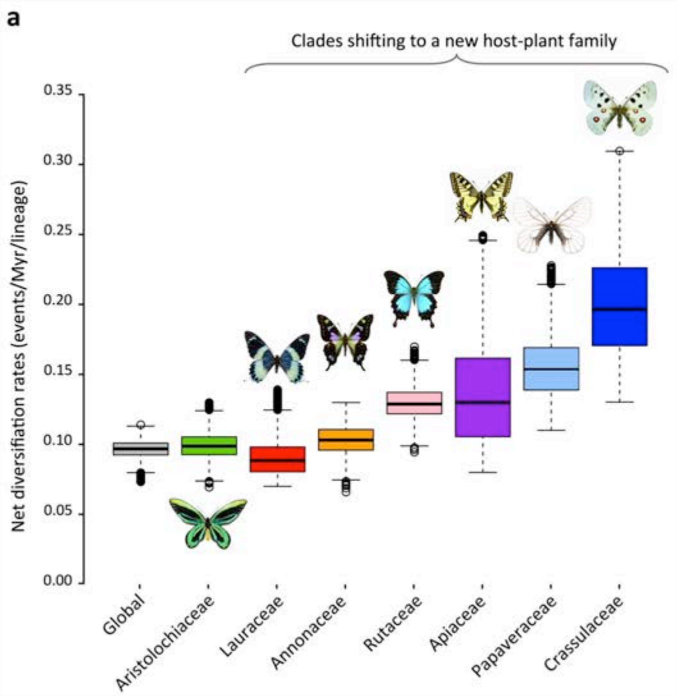
1. W

2. Te

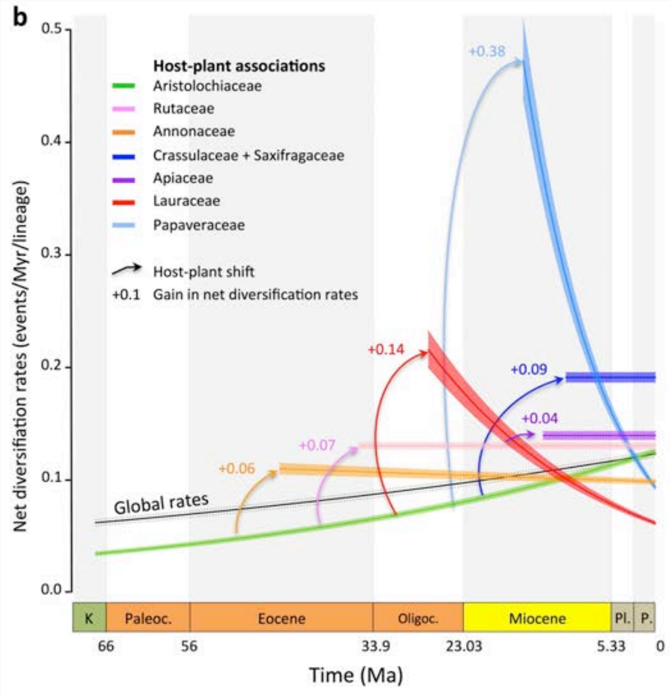
3. Bi

4. Sig … es?

Trait-dependent constant-rate models



Clade-specific time-dependent models



Back to the main questions

1. ***What are the ancestral host plants of Papilionidae?***

   ➡ Fed on Aristolochiaceae, very likely on the genus *Aristolochia*

2. ***Temporal congruence?***

   ➡ Both Papilionidae and *Aristolochia* originated ≈55 Ma

3. ***Biogeographic congruence?***

   ➡ Both Papilionidae and *Aristolochia* originated ≈55 Ma in Beringia

4. ***Signatures of genomic adaptations and on speciation rates?***

   ➡ Genome-wide adaptations and speciation rates increase with host-plant shifts

To summarize …

- Papilionidae = excellent biological model to study insect-plant interactions:

1. Ancestral host plant = Aristolochiaceae, highly toxic
2. Both groups originated at the same time and place
3. Strong niche conservatism of host-plant preferences
4. When butterflies shifted to new plants: increases of speciation rates and genome-wide signatures of positive selection

However… this is an ongoing work with many possibilities

1. Genomic data is huge: ontology of genes?
2. Increase the taxon sampling for whole genomes = more shifts
3. Transcriptomes

Acknowledgments

- <u>People:</u>

PI: Fabien Condamine

Entomologists: Felix Sperling, Adam Cotton, Gaël Kergoat

Botanists: Stephan Wanke, Oscar Pérez-Escobar, Guillaume Chomicki

Genomics: Benoit Nabholz

Molecular biology: Anne-Laure Clamens

- <u>Funding:</u>

Marie Curie IOF grant (BIOMME project)
ERC Starting Grant (GAIA project)

Thank you for your attention