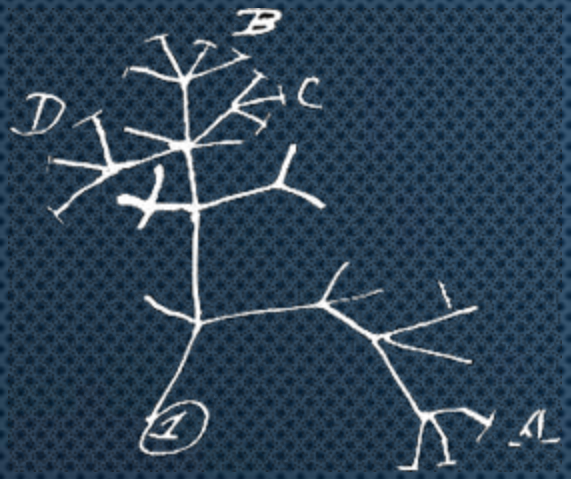


USING GENOMES TO RETRACE THE TREE OF LIFE

Dr. Bob Literman

NSF PostDoctoral Fellow

University of Rhode Island

3/20/2019

ACKNOWLEDGEMENTS

OUR BIG QUESTIONS

1) How are the species on Earth related?

2) Where can we find reliable data to help us reconstruct the tree of life?

WHY SHOULD WE CARE ABOUT SPECIES RELATIONSHIPS?

- Understanding species relationships is fundamental for:
  - Studying the evolution of traits
  - Effective biodiversity protection
  - Pathogen/disease transmission and treatment
  - Understanding the origins of life on Earth

HOW CAN WE FIGURE OUT SPECIES RELATIONSHIPS?

1) We **collect data** that can be compared among species
2) When species share more, they are grouped together
3) Then, those groups are paired with similar groups until a final tree is built

https://evolution.berkeley.edu/evolibrary/article/phylogenetics_07

	Vertebrae?	Bony skeleton?	Four limbs?	Amniotic egg?*	Hair?	Two post-orbital fenestrae?**
Sharks and relatives	YES	no	no	no	no	no
Ray-finned fishes	YES	YES	no	no	no	no
Amphibians	YES	YES	YES	no	no	no
Primates	YES	YES	YES	YES	YES	no
Rodents and rabbits	YES	YES	YES	YES	YES	no
Crocodiles and relatives	YES	YES	YES	YES	no	YES
Dinosaurs and birds	YES	YES	YES	YES	no	YES

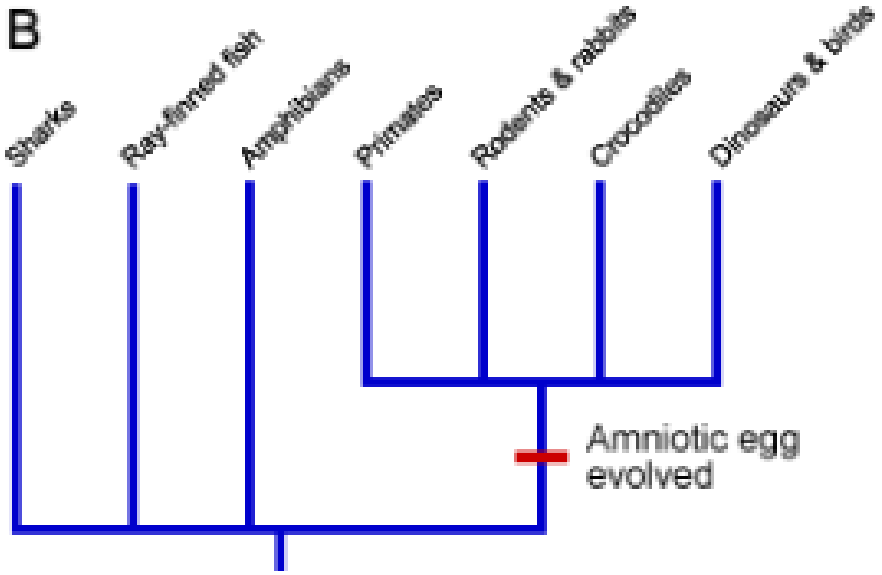
*amniotic egg:
an egg in which the embryo is surrounded by the moisture-retaining amnion membrane

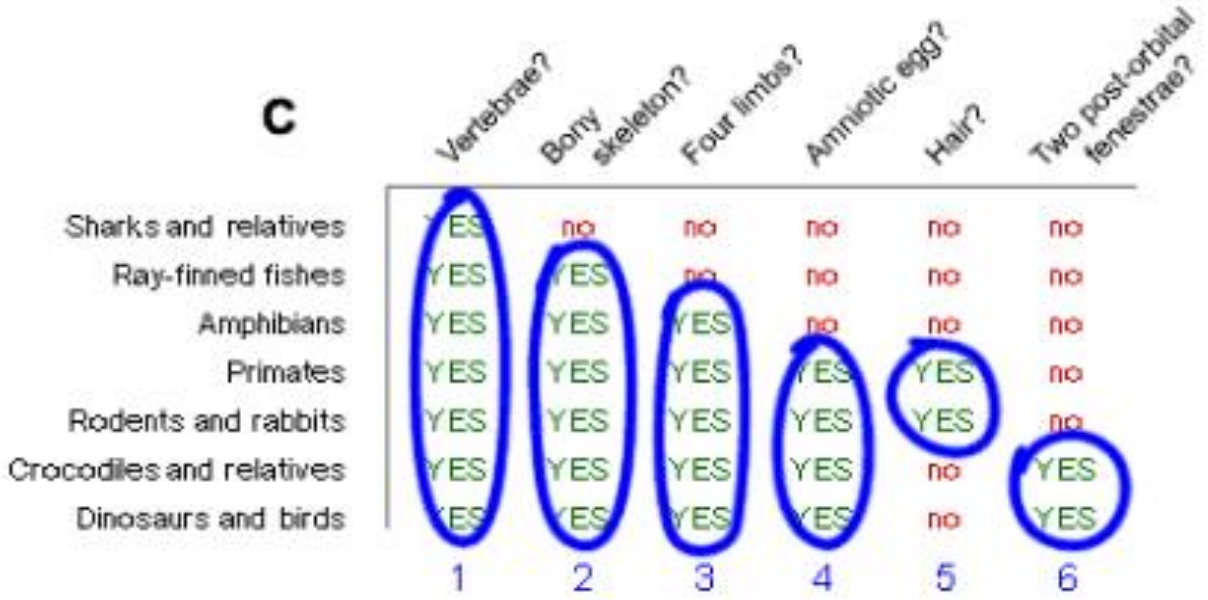
**post-orbital fenestrae:
holes in the skull behind the eye

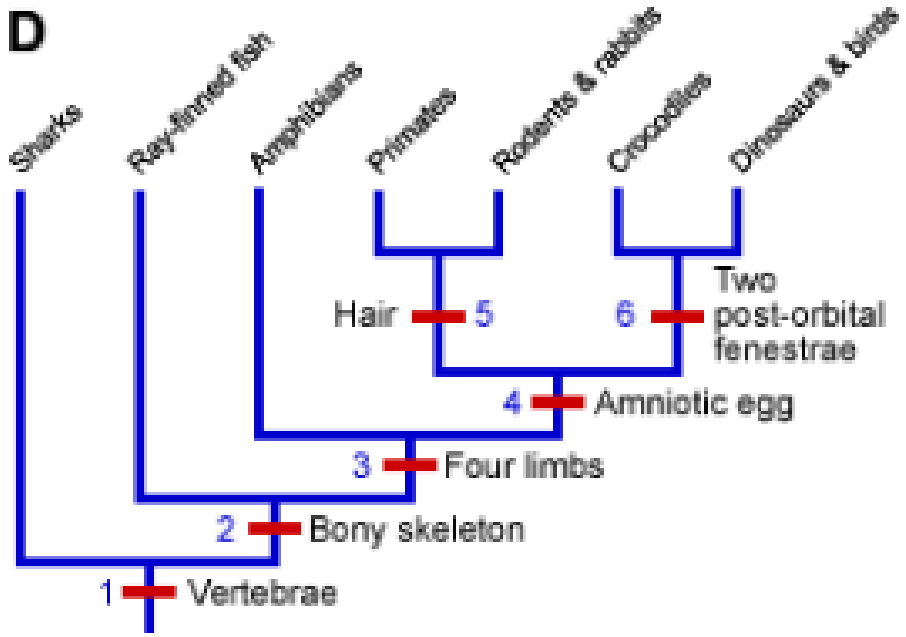
A

	Vertebrae?	Bony skeleton?	Four limbs?	Amniotic egg?	Hair?	Two post-orbital fenestrae?
Sharks and relatives	YES	no	no	no	no	no
Ray-finned fishes	YES	YES	no	no	no	no
Amphibians	YES	YES	YES	no	no	no
Primates	YES	YES	YES	YES	YES	no
Rodents and rabbits	YES	YES	YES	YES	YES	no
Crocodiles and relatives	YES	YES	YES	YES	no	YES
Dinosaurs and birds	YES	YES	YES	YES	no	YES

B




C
Vertebrae?
Bony skeleton?
Four limbs?
Amniotic egg?
Hair?
Two post-orbital fenestrae?
Sharks and relatives
Ray-finned fishes
Amphibians
Primates
Rodents and rabbits
Crocodiles and relatives
Dinosaurs and birds
1
2
3
4
5
6

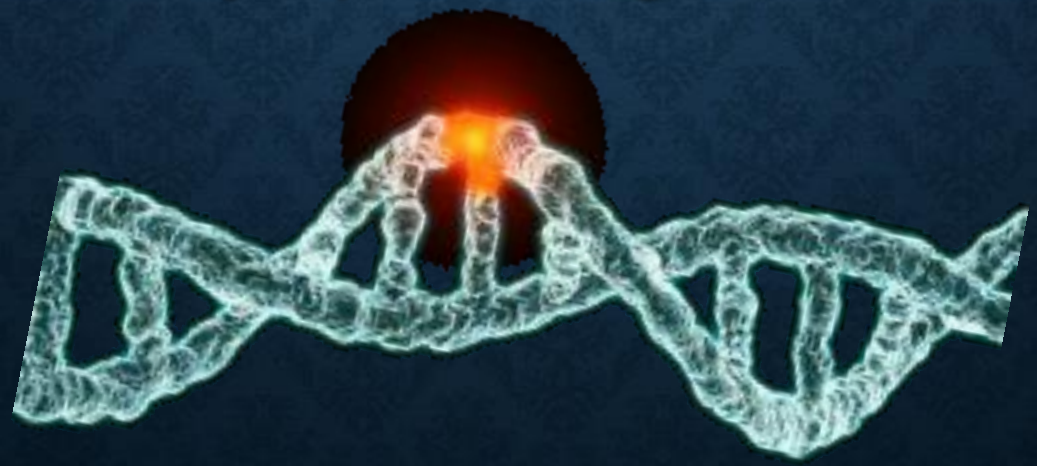

D
Sharks
Ray-finned fish
Amphibians
Primates
Rodents & rabbits
Crocodiles
Dinosaurs & birds
Hair
5
6
Two post-orbital fenestrae
4
Amniotic egg
3
Four limbs
2
Bony skeleton
1
Vertebrae

What types of data can be used to build an evolutionary tree?

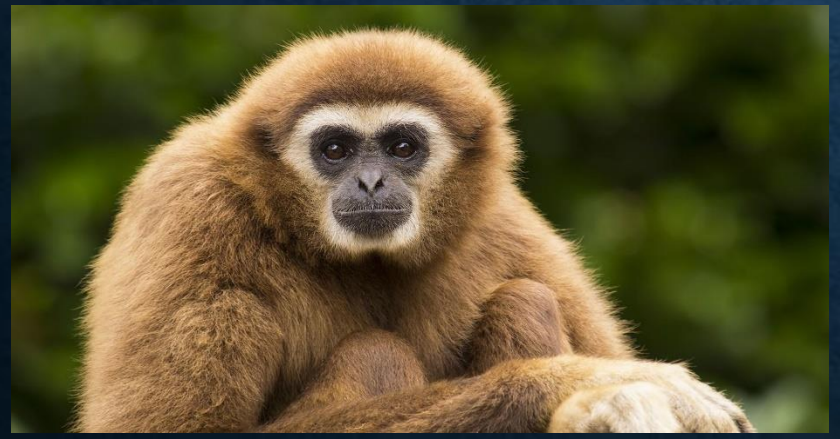
- Morphological data (good for fossils)
- Biochemical data (good for bacteria)
- Genetic data
  - Why genetic data?

DNA IS ALWAYS MUTATING

- DNA bases can be changed, inserted, or removed from a genome for a number of reasons
- We can use this variation to reconstruct the evolutionary history by looking for shared mutations



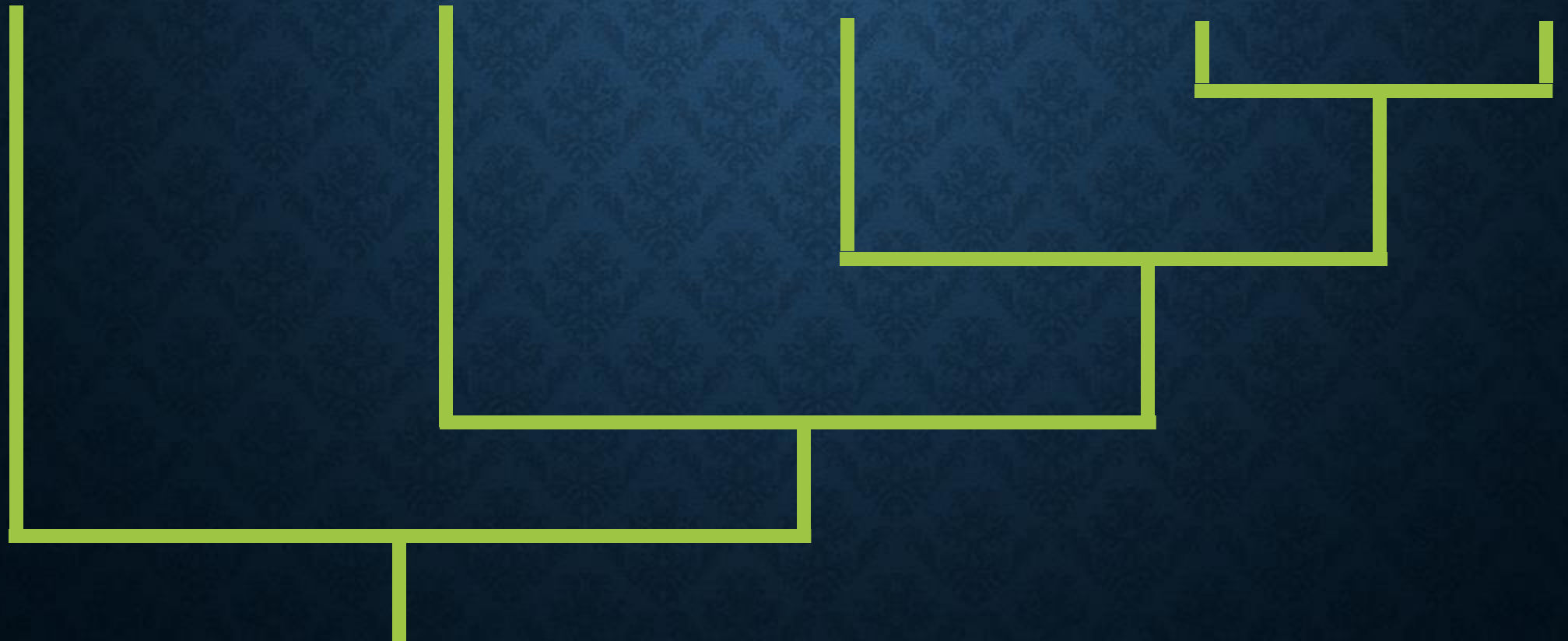
DNA IS ALWAYS MUTATING



DNA IS ALWAYS MUTATING



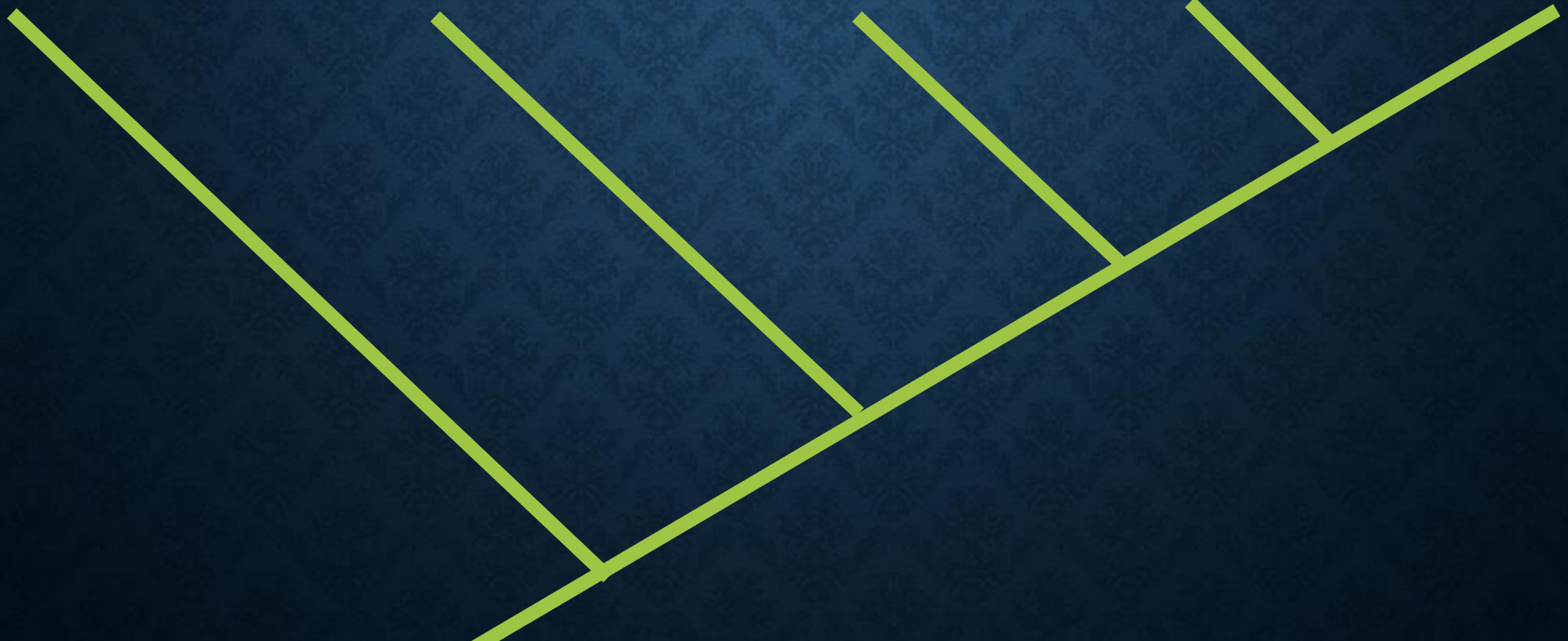




DNA IS ALWAYS MUTATING

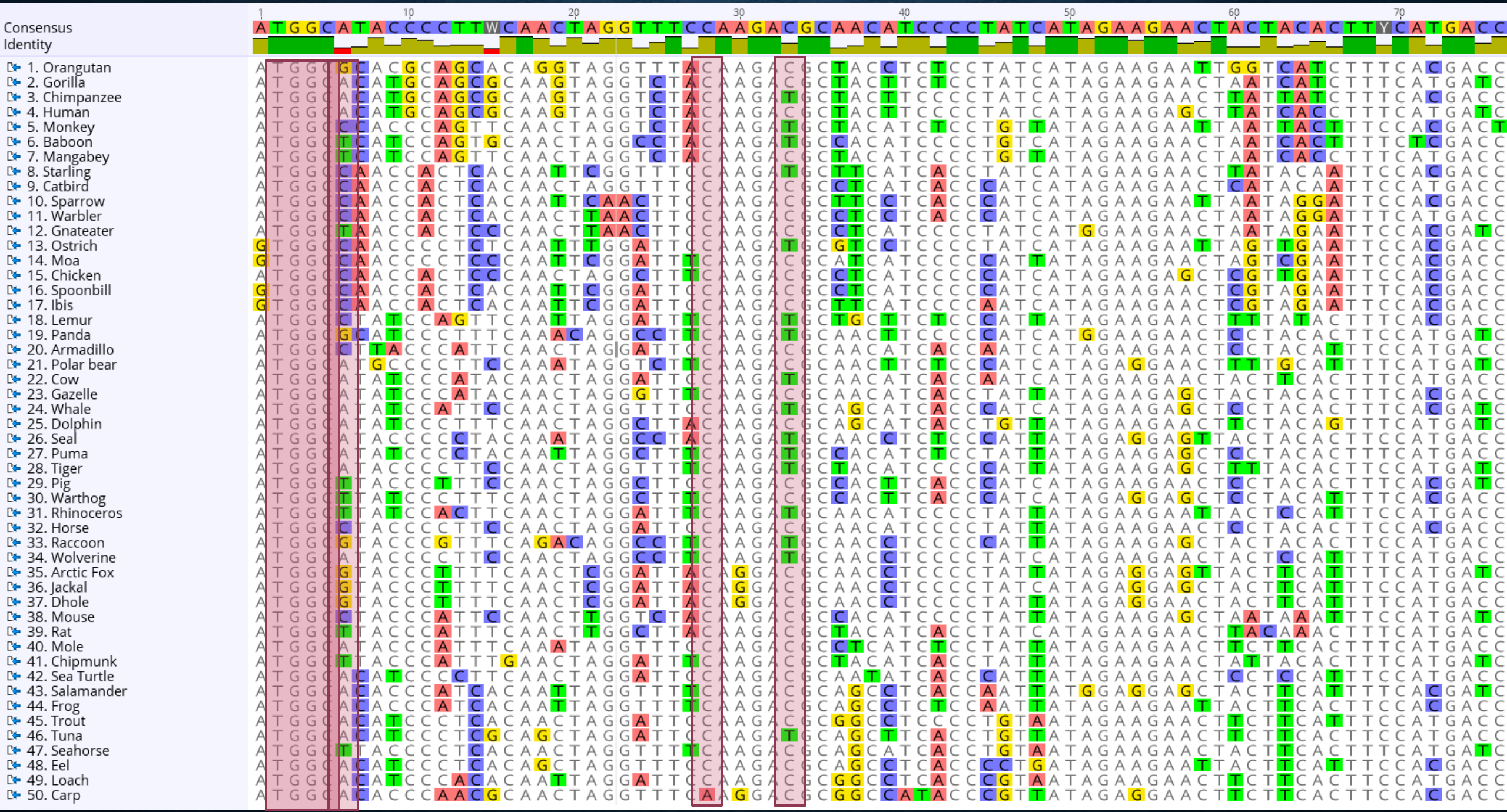


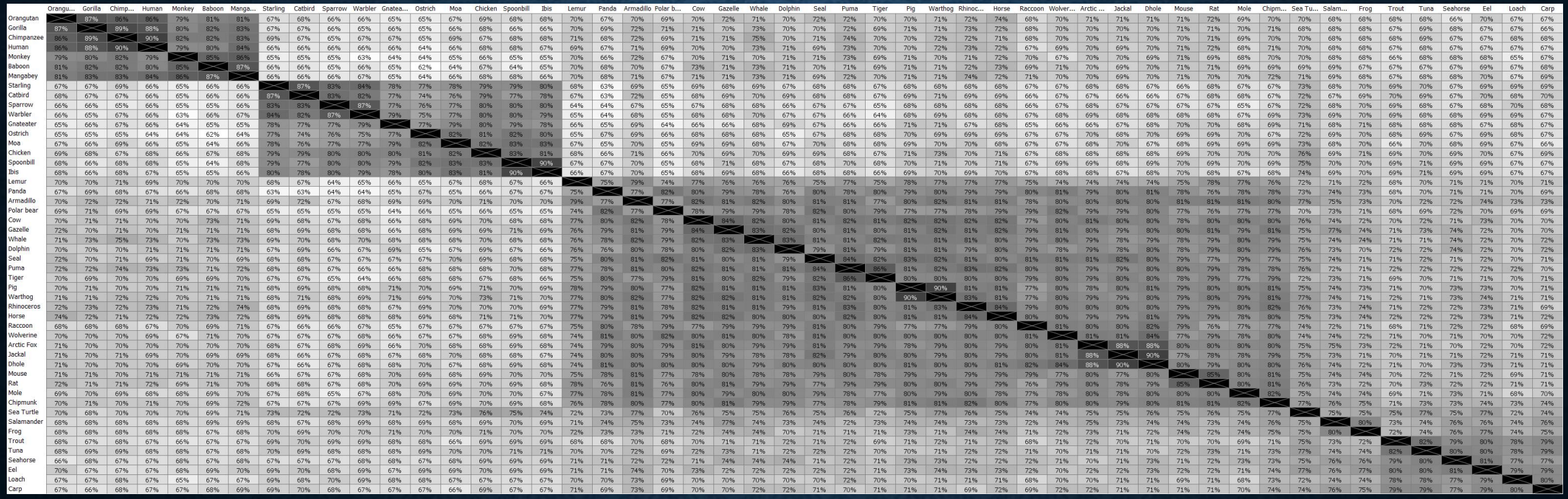


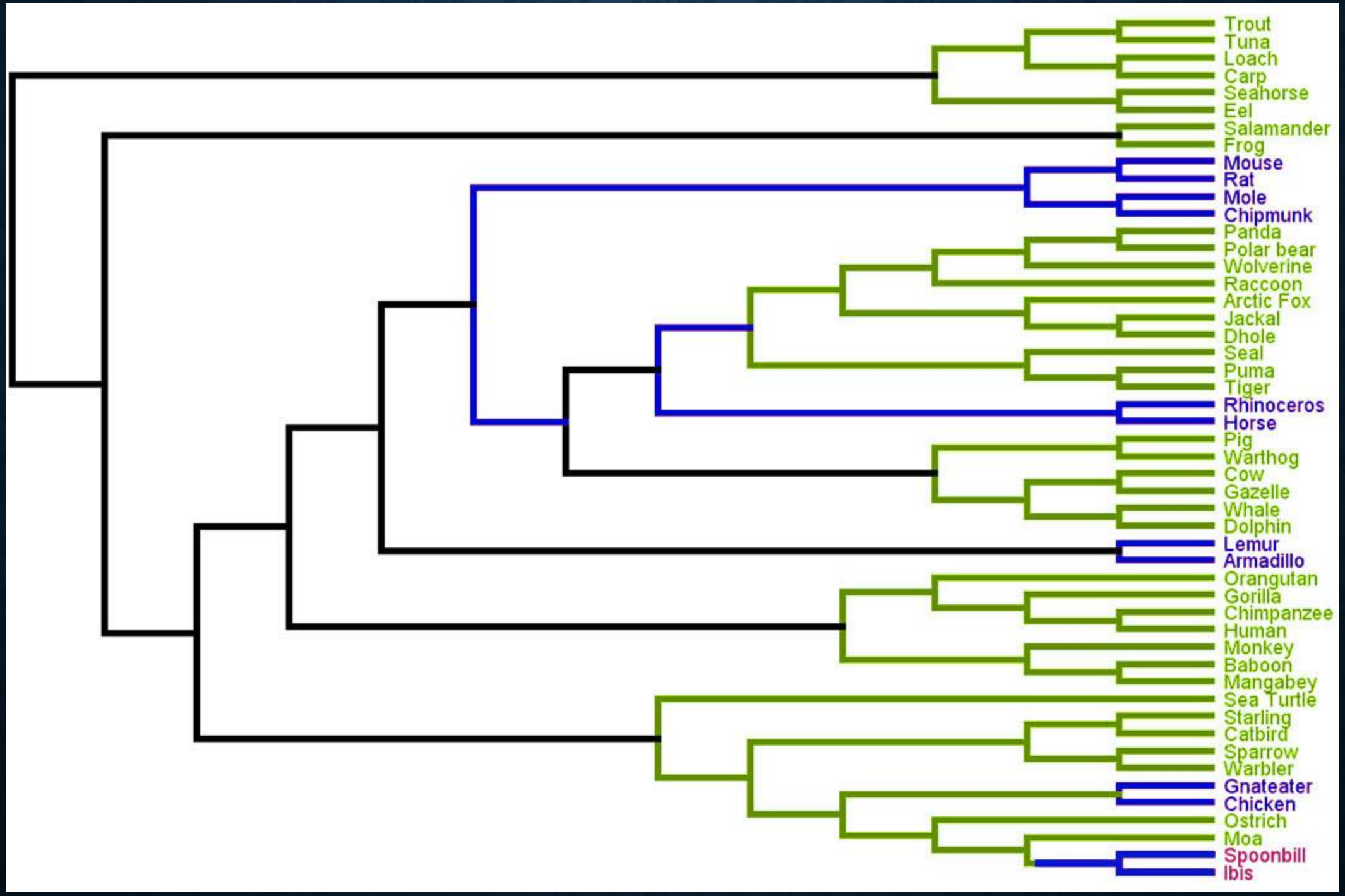


SINGLE GENE PHYLOGENETICS

- 1970's and 1980's
- Sequencing a single gene could take months or years
- Single genes were used to build phylogenetic trees
- COXII: Mitochondrial gene involved in cell respiration


Consensus
Identity
1. Orangutan
2. Gorilla
3. Chimpanzee
4. Human
5. Monkey
6. Baboon
7. Mangabey
8. Starling
9. Catbird
10. Sparrow
11. Warbler
12. Gnateater
13. Ostrich
14. Moa
15. Chicken
16. Spoonbill
17. Ibis
18. Lemur
19. Panda
20. Armadillo
21. Polar bear
22. Cow
23. Gazelle
24. Whale
25. Dolphin
26. Seal
27. Puma
28. Tiger
29. Pig
30. Warthog
31. Rhinoceros
32. Horse
33. Raccoon
34. Wolverine
35. Arctic Fox
36. Jackal
37. Dhole
38. Mouse
39. Rat
40. Mole
41. Chipmunk
42. Sea Turtle
43. Salamander
44. Frog
45. Trout
46. Tuna
47. Seahorse
48. Eel
49. Loach
50. Carp




Trout
Tuna
Loach
Carp
Seahorse
Eel
Salamander
Frog
Mouse
Rat
Mole
Chipmunk
Panda
Polar bear
Wolverine
Raccoon
Arctic Fox
Jackal
Dhole
Seal
Puma
Tiger
Rhinoceros
Horse
Pig
Warthog
Cow
Gazelle
Whale
Dolphin
Lemur
Armadillo
Orangutan
Gorilla
Chimpanzee
Human
Monkey
Baboon
Mangabey
Sea Turtle
Starling
Catbird
Sparrow
Warbler
Gnateater
Chicken
Ostrich
Moa
Spoonbill
Ibis

WHY MIGHT SINGLE-GENE TREES DIFFER FROM THE SPECIES TREE?

- **Homoplasy**: New mutations on top of old mutations
- **Convergent Evolution**: Different species evolving the same sequence independently, not through descent
- **Lack of Data:** V d b

WHY MIGHT SINGLE-GENE TREES DIFFER FROM THE SPECIES TREE?

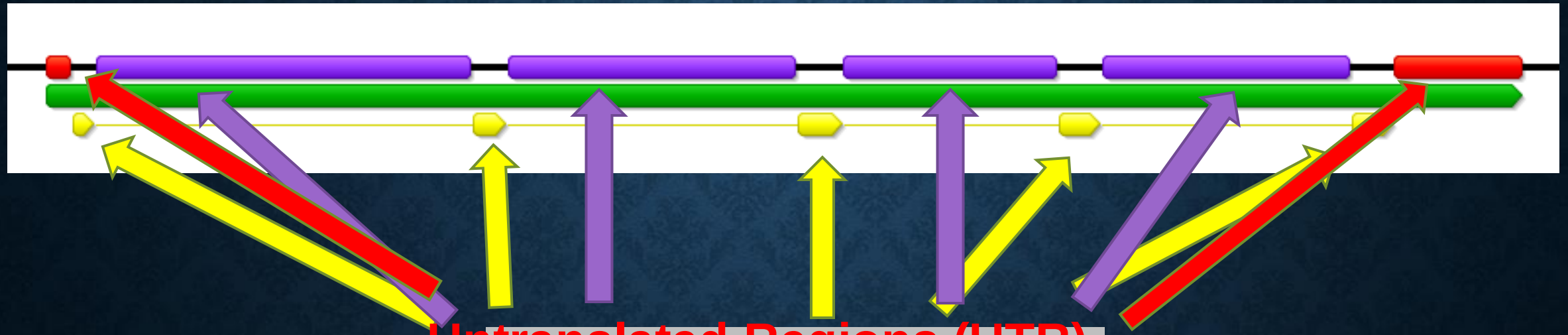
- If we are interested in building species trees, and individuals genes or loci are prone to these issues, what can we do?
- Use more data!
- Sequencing of whole genomes is getting cheaper by the day

WHAT'S IN A GENOME?

- Genomes contain the genetic material required to build an organism
- Instruction guide for:
    - What to build
    - When to build it
- Different parts of the genome perform different functions and are impacted differently by selection

WHAT'S IN A GENOME?

- Protein-coding genes provide instructions to build specific proteins



Untranslated Regions (UTR)
- Important for regulation and mRNA stability/transport
- 5' (near start), 3' (near end)

Introns
- Spacers between exons

WHAT'S IN A GENOME?

- Other genomic regions include:
  - Long-noncoding RNA (lncRNA)
  - Small RNAs (smRNA)
  - Pseudogenes
  - Noncoding genes
  - Unannotated/Intergenic

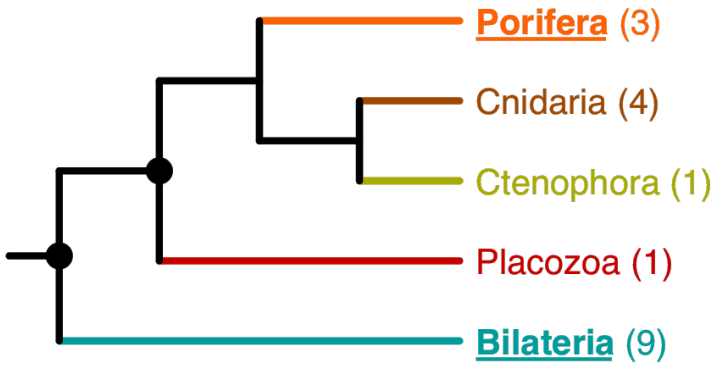
WAS USING BIGGER DATASETS THE MAGIC BULLET?

A

Schierwater et al. (2009)

15 mitochondrial and 34 nuclear genes

GTR and WAG models

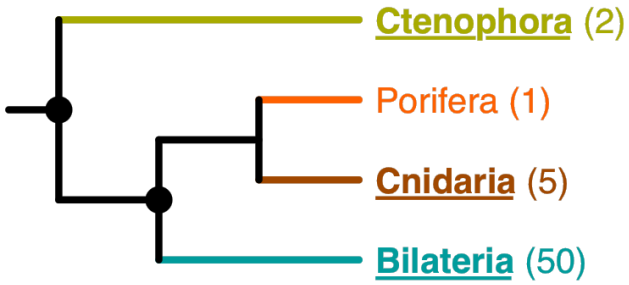


B

Dunn et al. (2008)

6 mitochondrial and 144 nuclear genes

WAG and CAT models

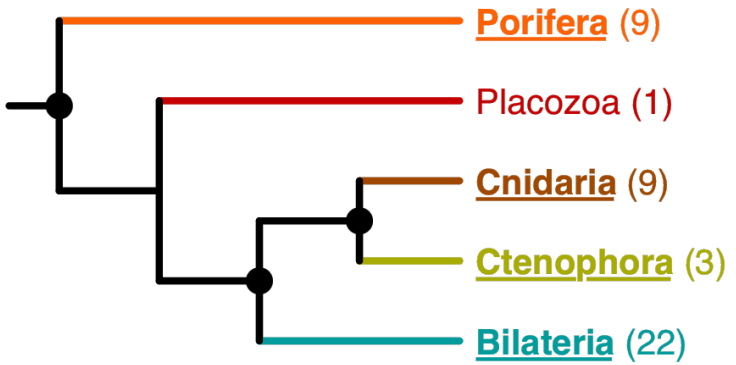


C

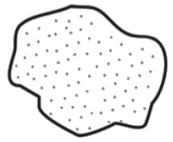
Philippe et al. (2009)

128 nuclear genes

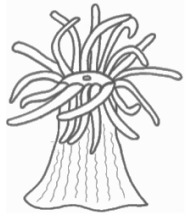
CAT model



Porifera

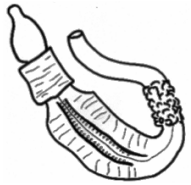


Placozoa



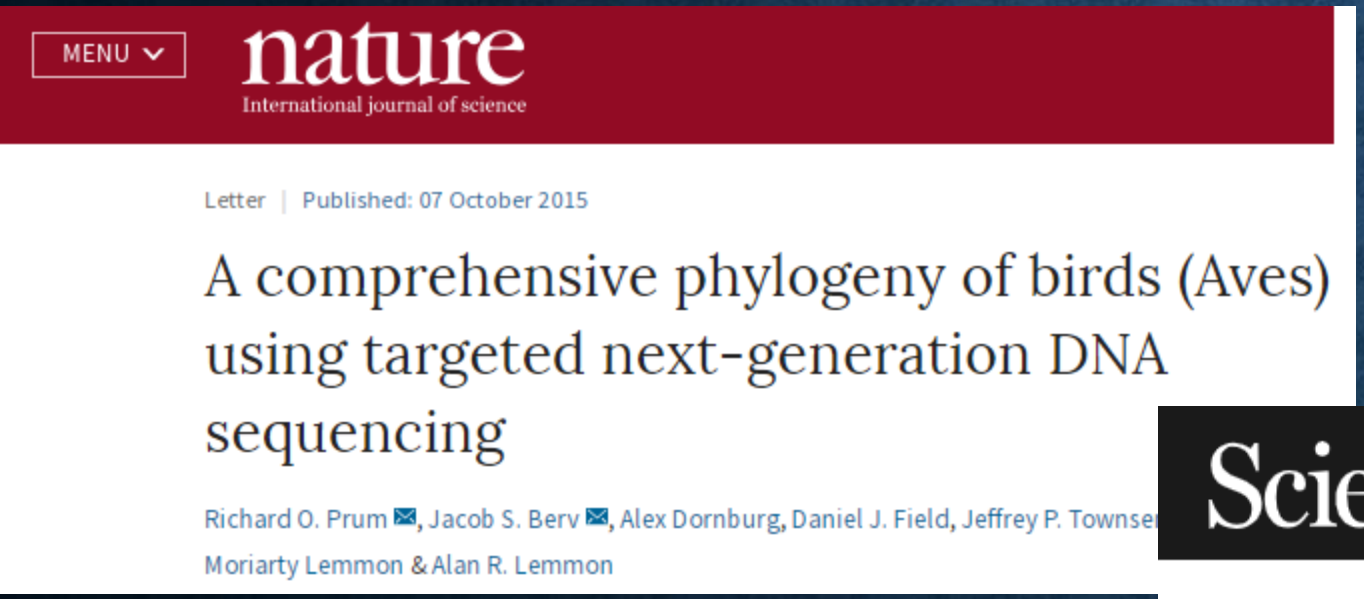
Cnidaria

Ctenophora



Bilateria

WAS USING BIGGER DATASETS THE MAGIC BULLET?



Species 1

Species 2

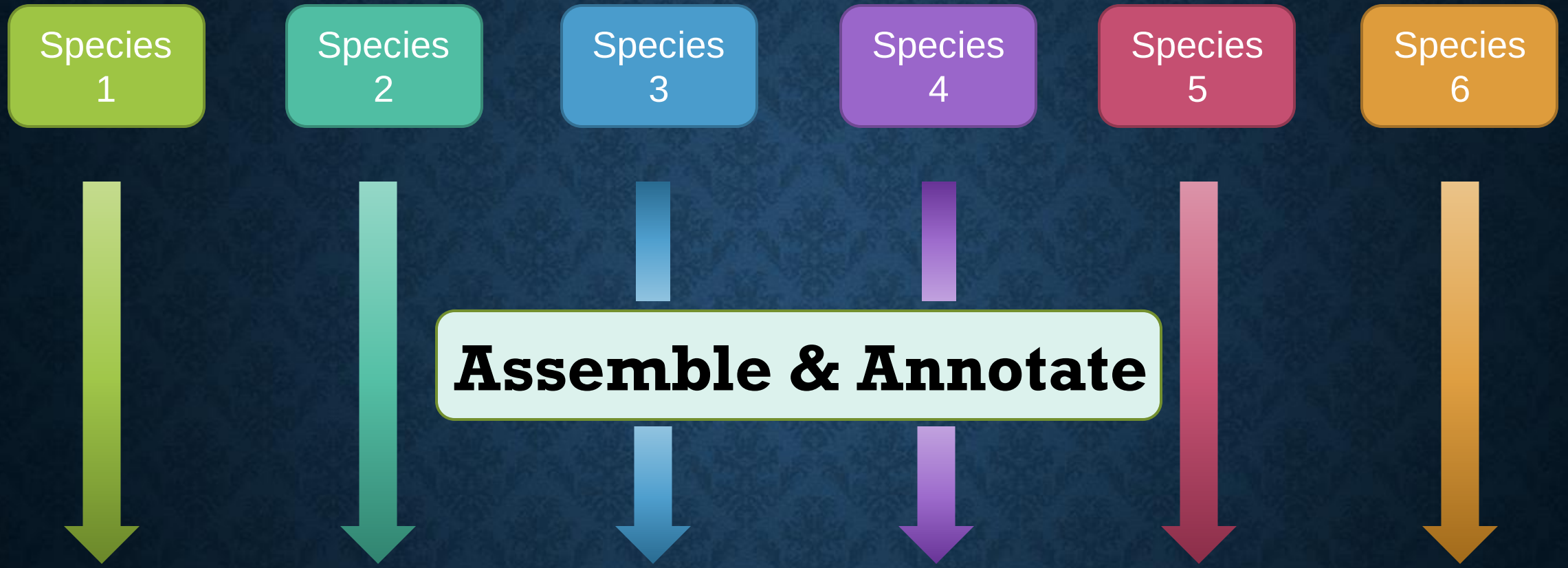
Species 3

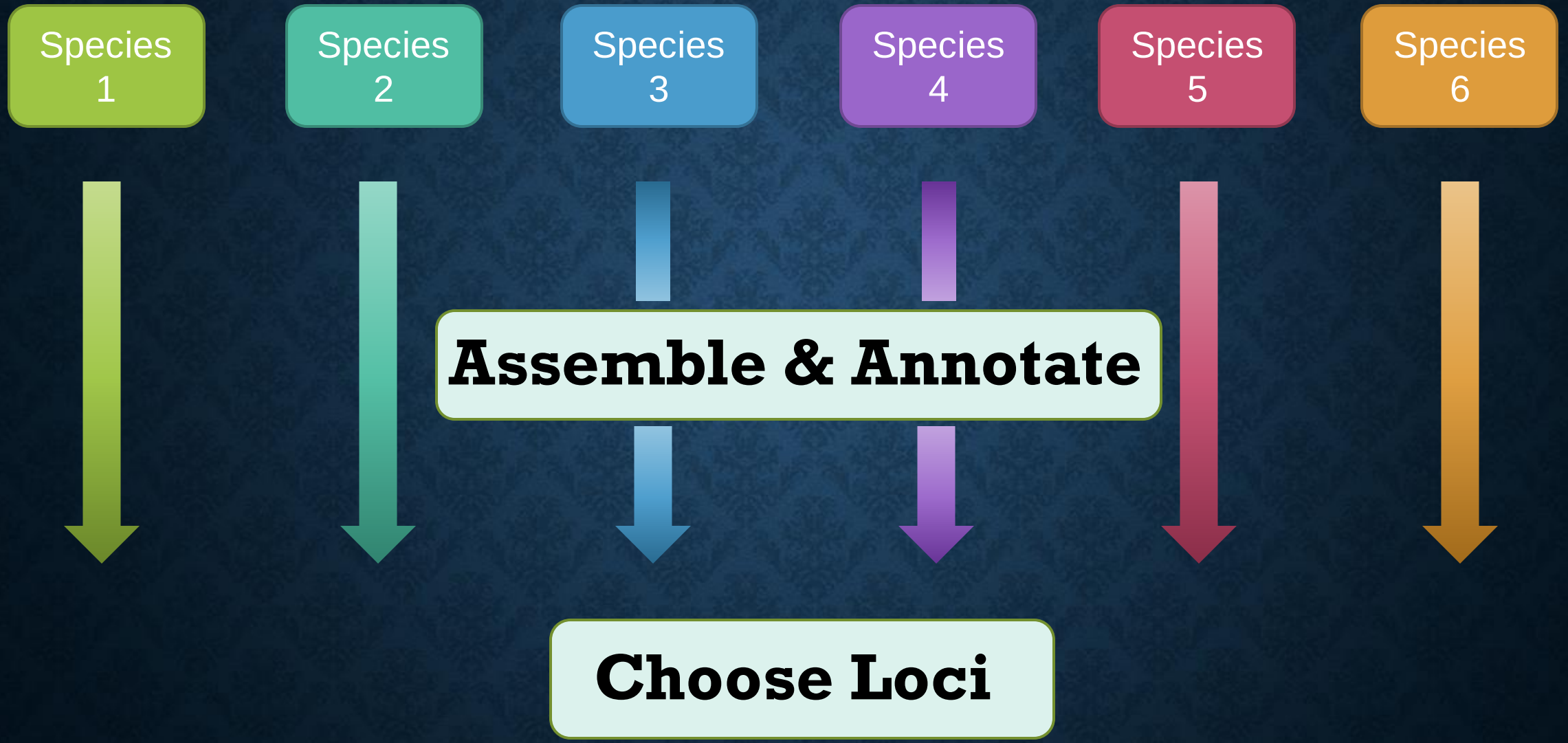
Species 4

Species 5

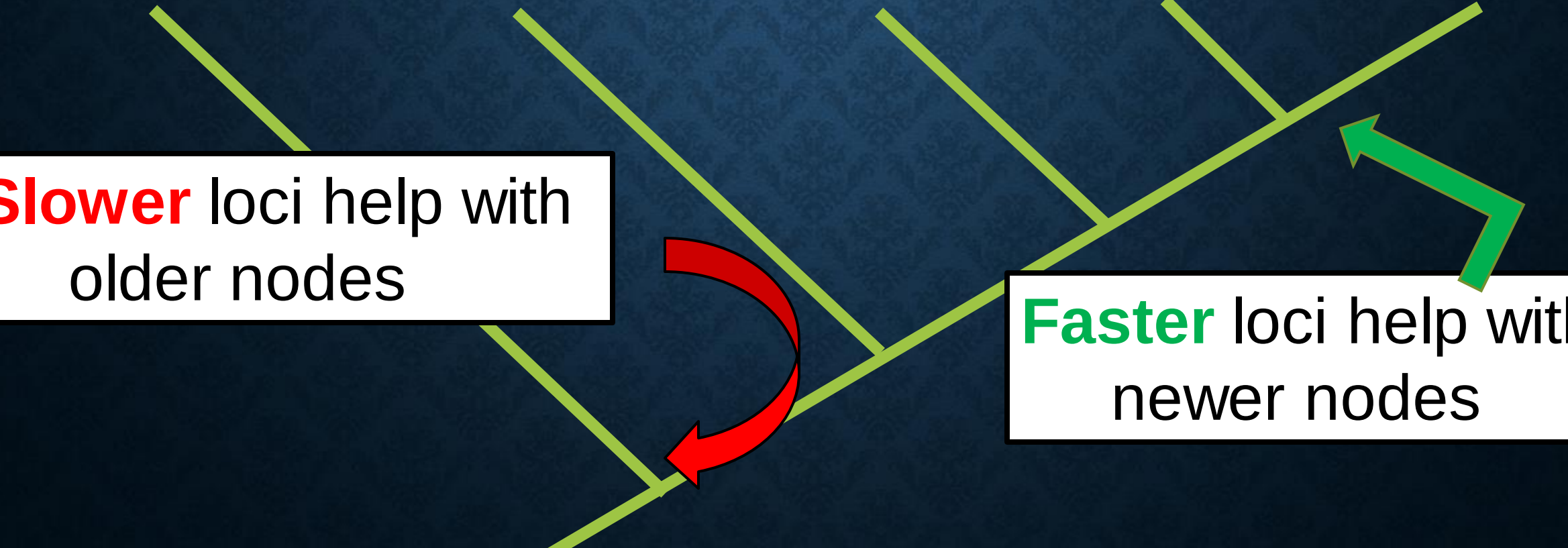
Species 6

Traditional Pipeline


Species 1
Species 2
Species 3
Species 4
Species 5
Species 6
Assemble & Annotate


Species 1
Species 2
Species 3
Species 4
Species 5
Species 6
Assemble & Annotate
Choose Loci

CHOOSING LOCI FOR TREE-BUILDING



Slower loci help with older nodes

Faster loci help with newer nodes

WHAT'S THE BIG PROBLEM?

- Choosing specific loci can bias results
- Assembling and annotating genomes is data- and time-intensive
- Lots of data that doesn't support your question

Species 1

Species 2

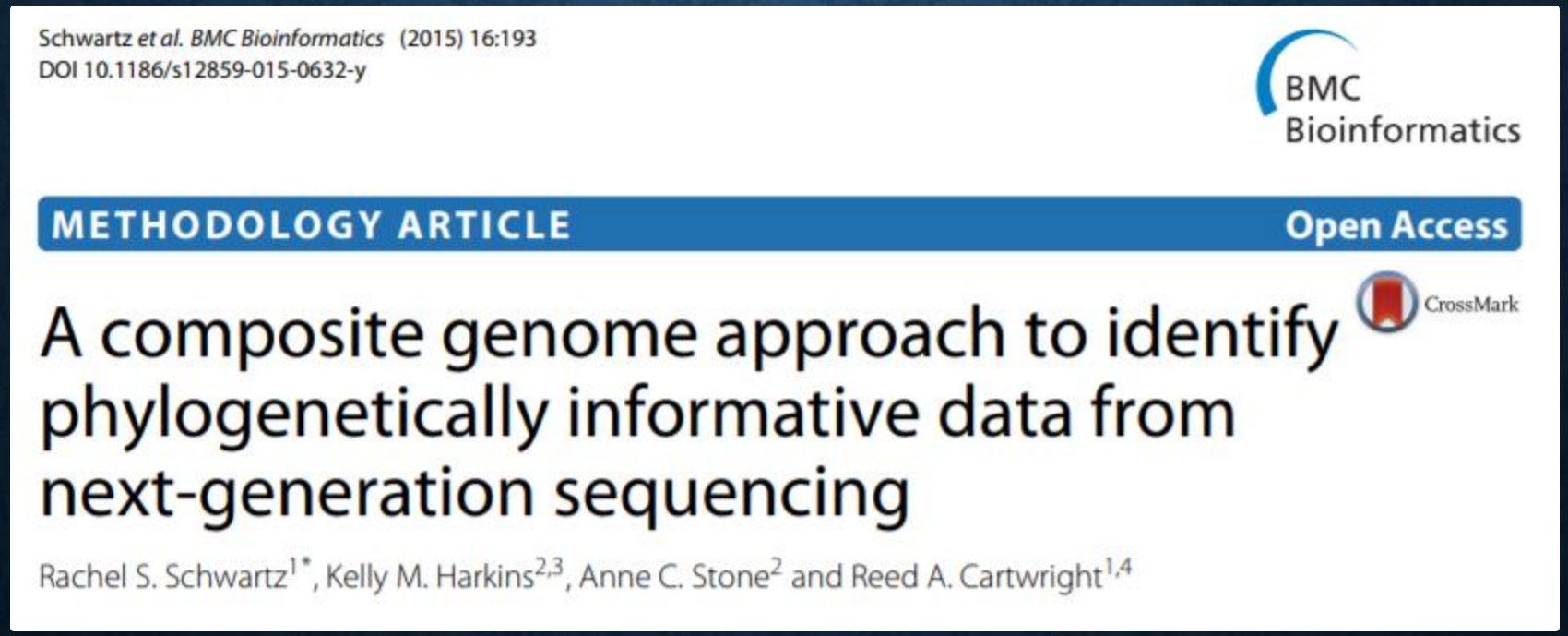
Species 3

Species 4

Species 5

Species 6

SISRS Pipeline



Schwartz *et al. BMC Bioinformatics* (2015) 16:193
DOI 10.1186/s12859-015-0632-y

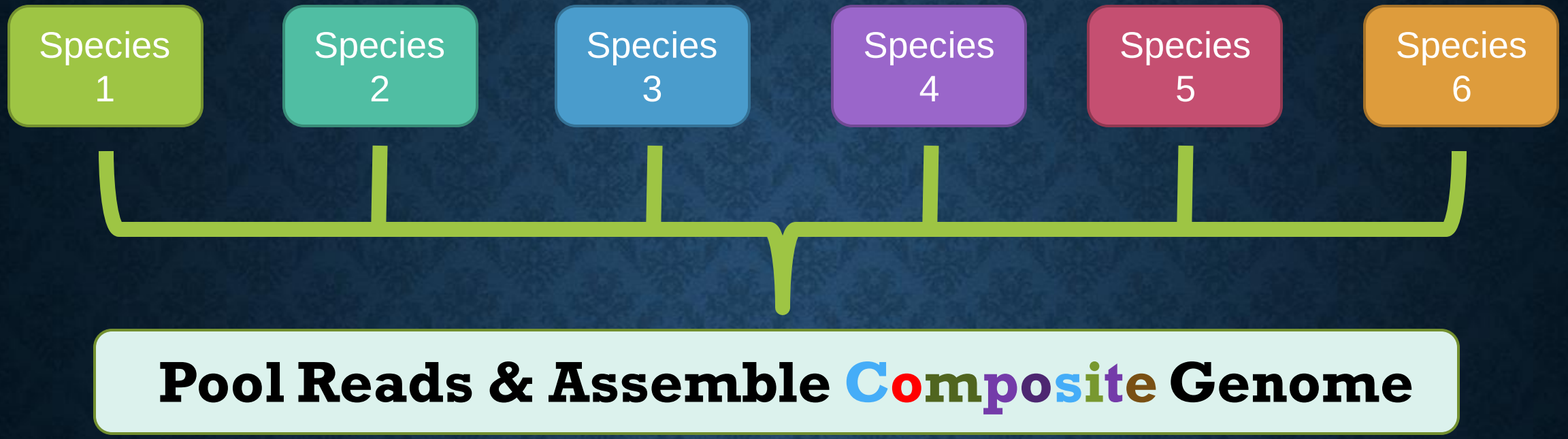
BMC Bioinformatics

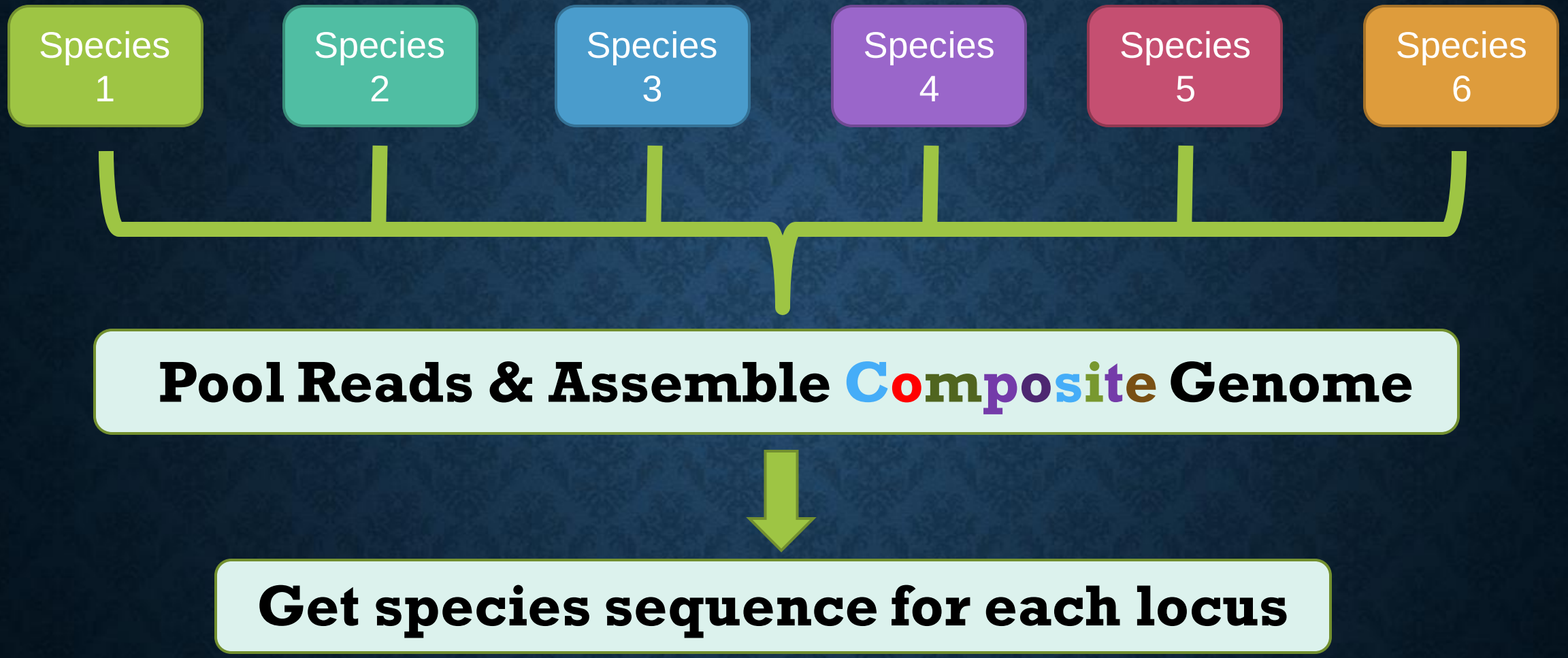
METHODOLOGY ARTICLE Open Access

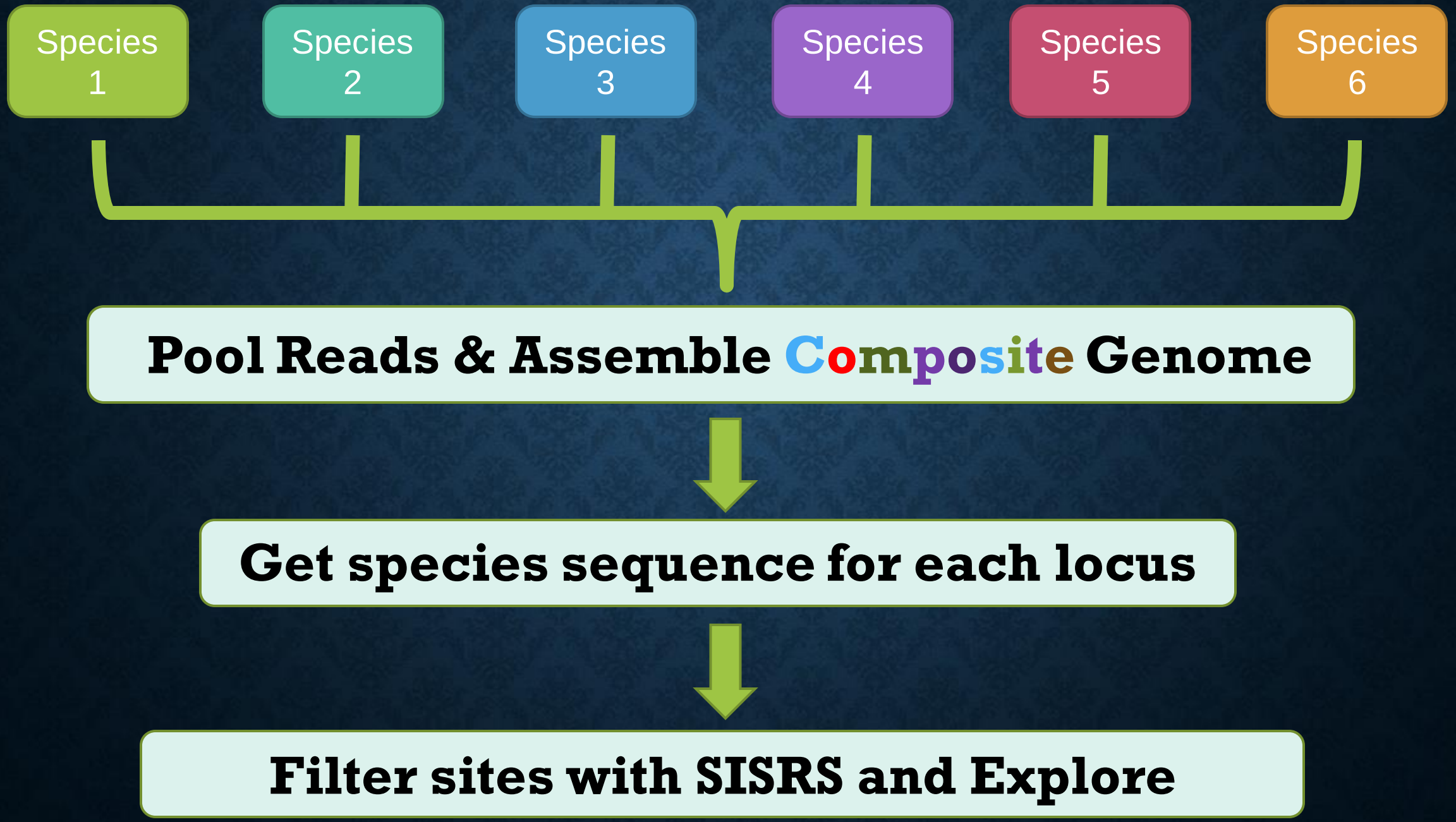
CrossMark

A composite genome approach to identify phylogenetically informative data from next-generation sequencing

Rachel S. Schwartz[1*], Kelly M. Harkins[2,3], Anne C. Stone[2] and Reed A. Cartwright[1,4]


Species 1
Species 2
Species 3
Species 4
Species 5
Species 6
Pool Reads & Assemble Composite Genome


Species 1
Species 2
Species 3
Species 4
Species 5
Species 6
Pool Reads & Assemble Composite Genome
Get species sequence for each locus


Species 1
Species 2
Species 3
Species 4
Species 5
Species 6
Pool Reads & Assemble Composite Genome
Get species sequence for each locus
Filter sites with SISRS and Explore

SISRS FILTERING

- SISRS removes sites that:
  - Are invariant (No help for relationships)
  - Are singletons (One species varies from rest)
  - Are hypervariable (Sites prone to homoplasy)

- SISRS does all filtering WITHOUT any knowledge of annotation types

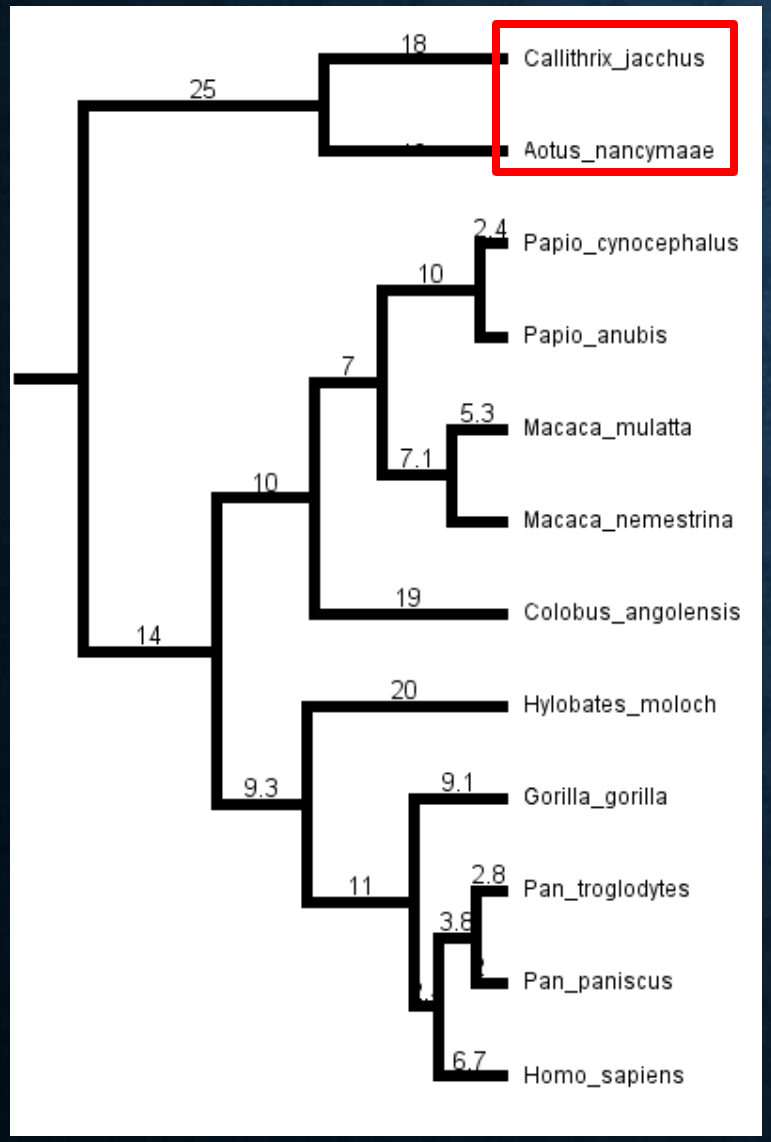
OUR STUDY QUESTIONS

1. Does SISRS filtered data help us to reconstruct accurate species-level phylogenies?
2. What parts of the genome are supporting older and newer nodes?

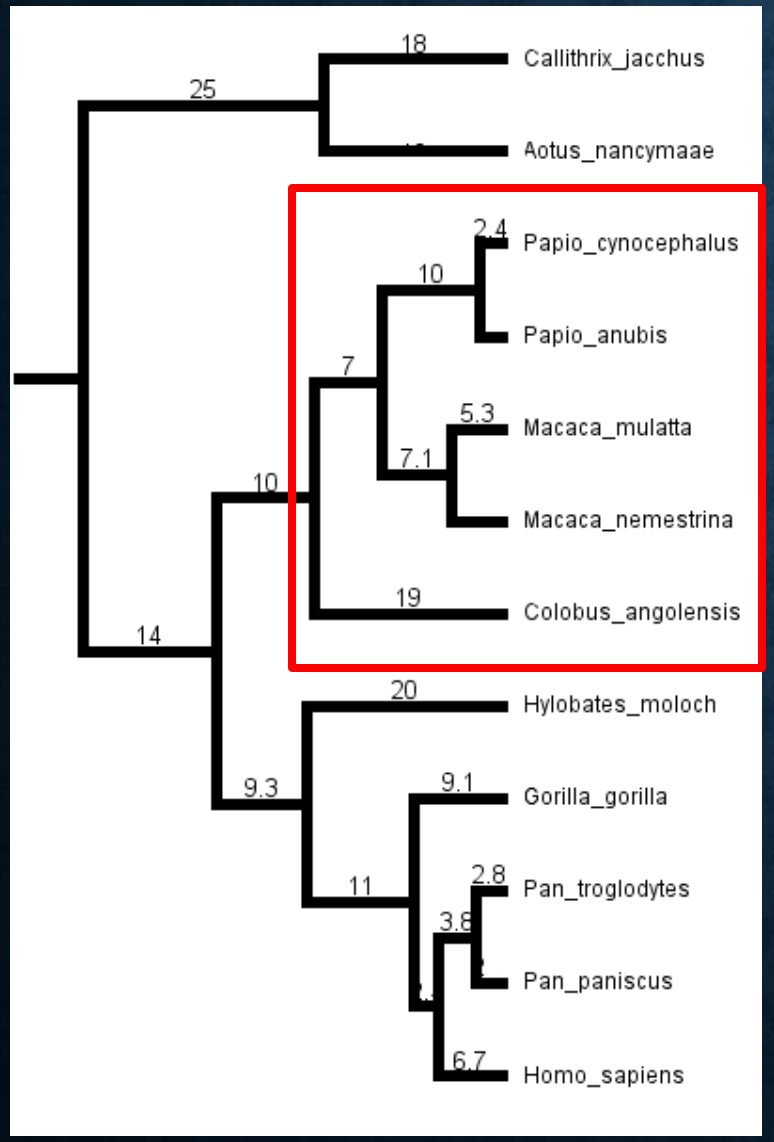
DATA COLLECTION

- Downloaded publicly available DNA-Seq data for:
  - 12 Primates (Reference Genome: Human)
  - 12 Rodentia (Reference Genome: Mouse)
  - 12 Cetartiodactyla (Reference Genome: Cow)
- Taxa from each group have well-supported evolutionary relationships

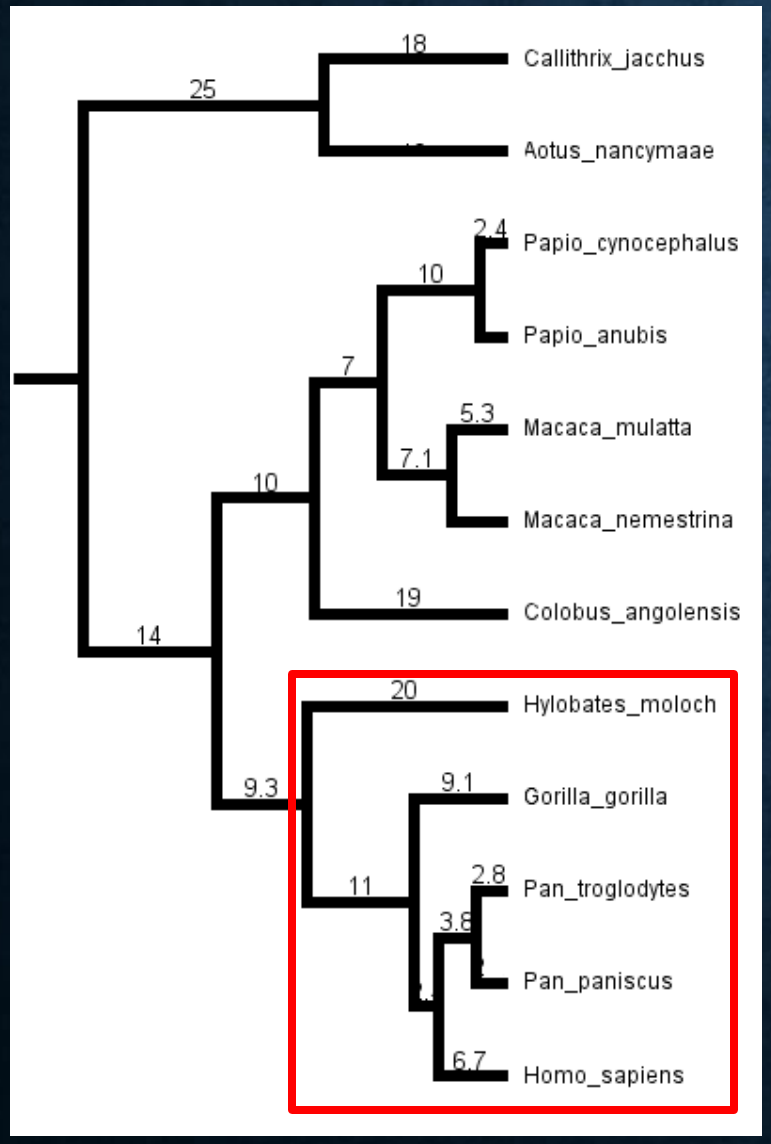
DATA COLLECTION



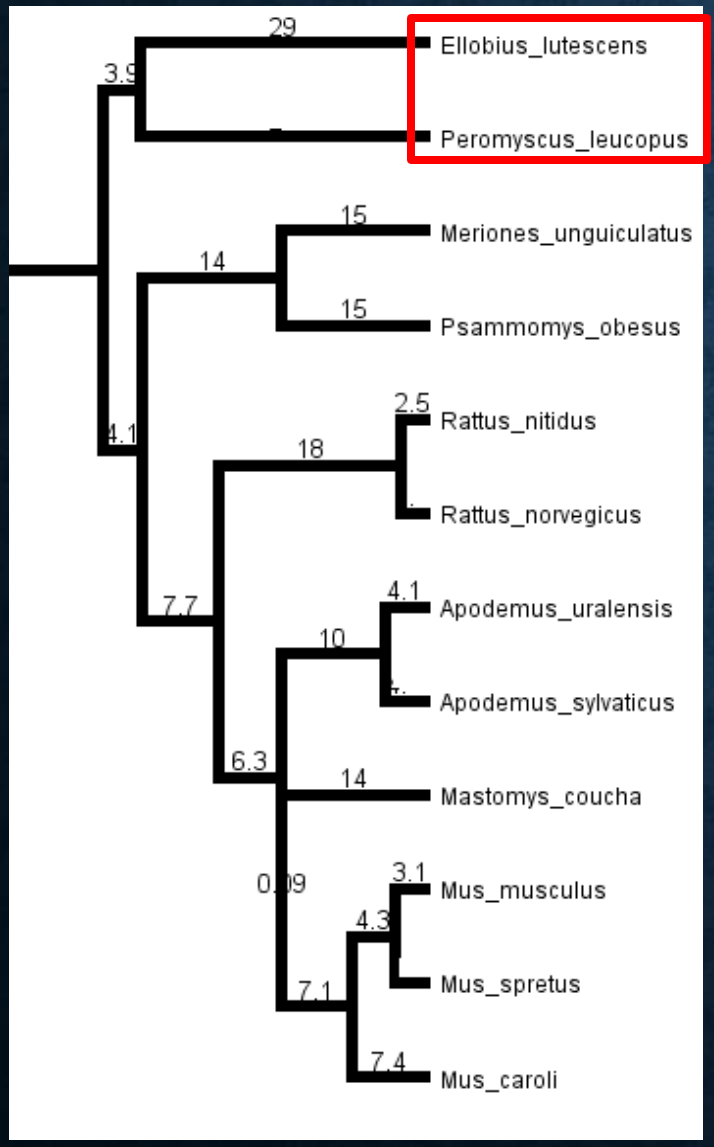
DATA COLLECTION



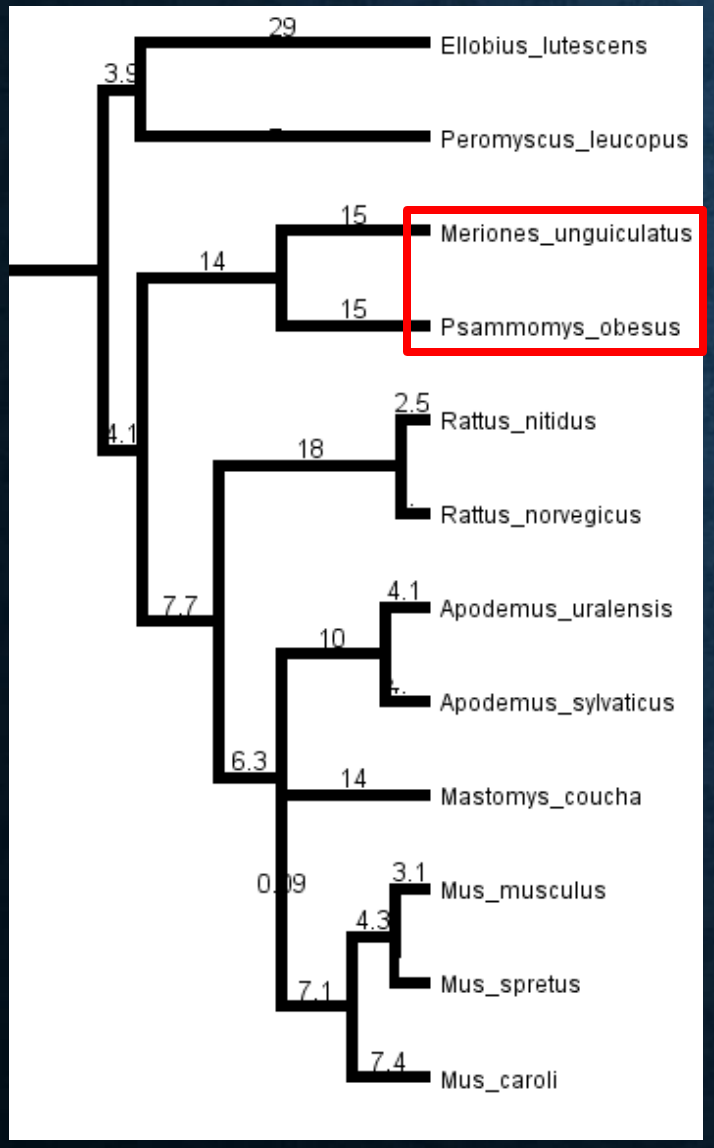
DATA COLLECTION



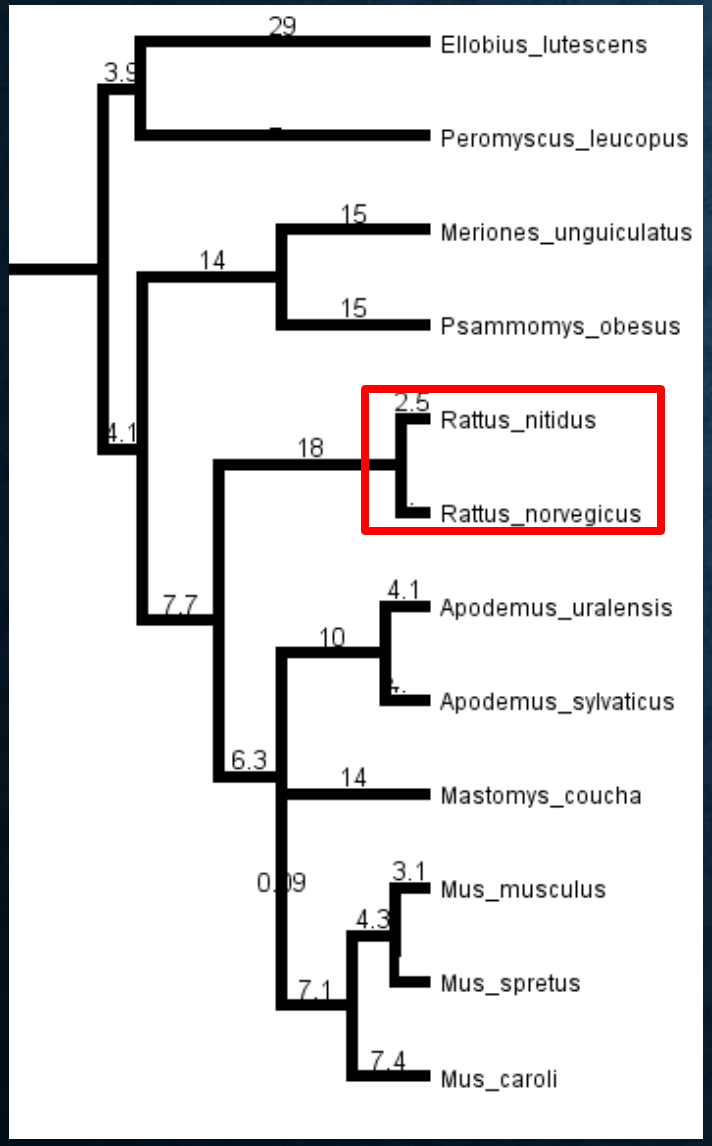
DATA COLLECTION



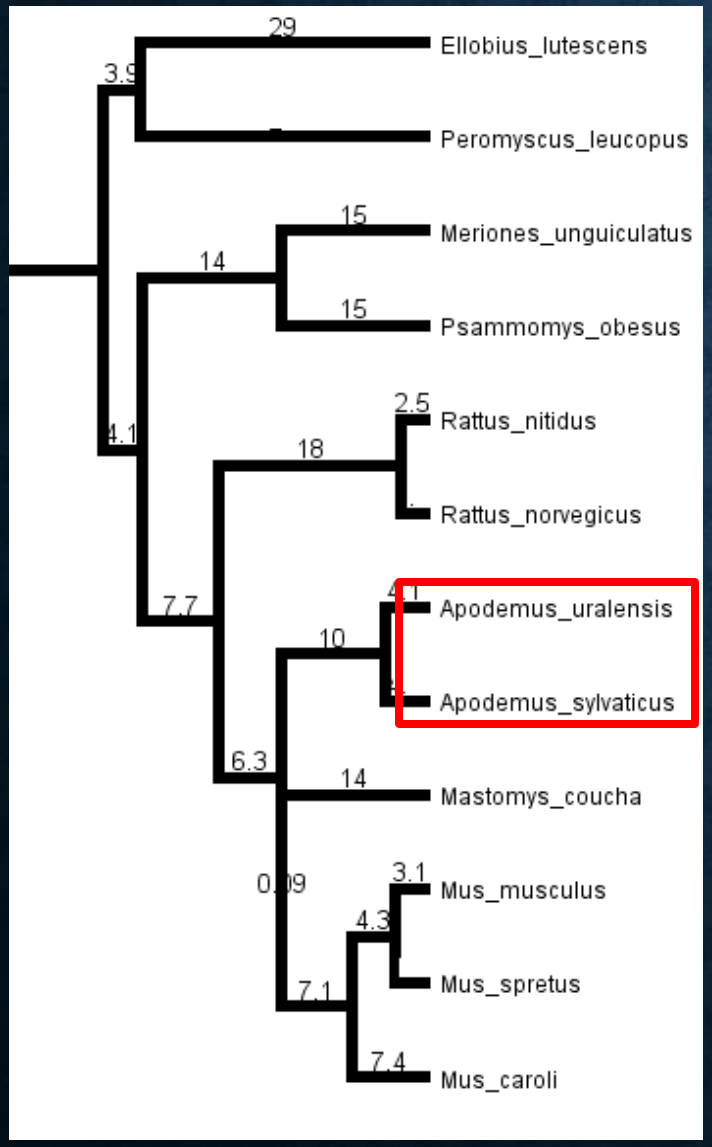
DATA COLLECTION



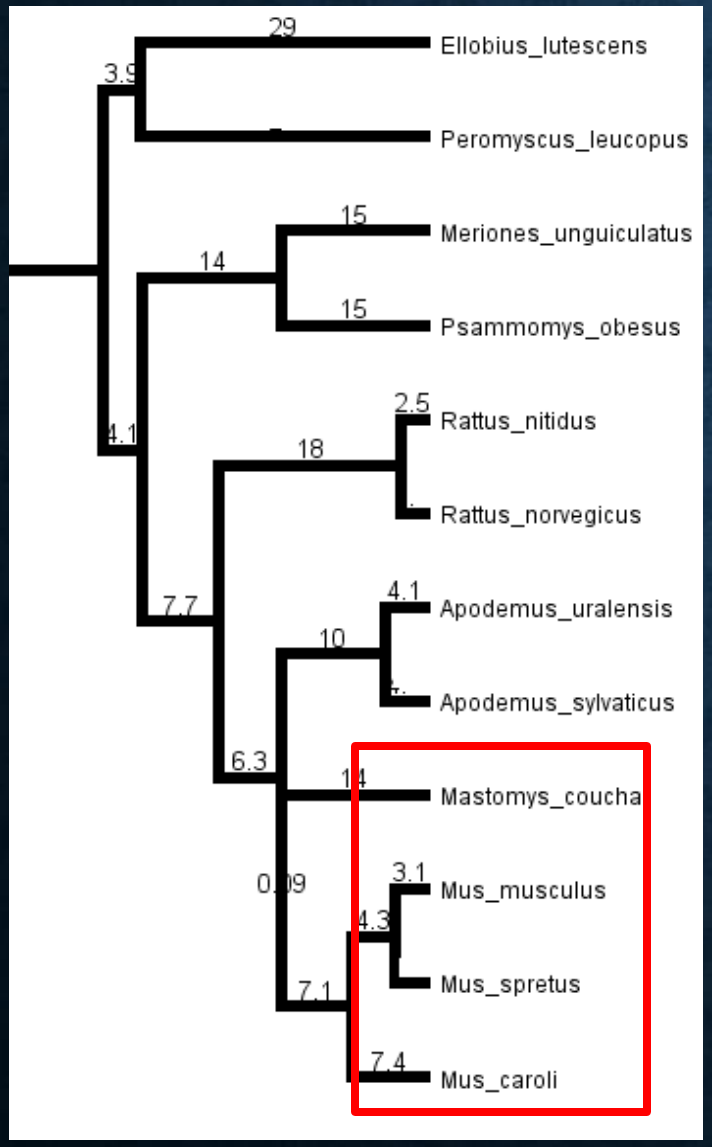
DATA COLLECTION



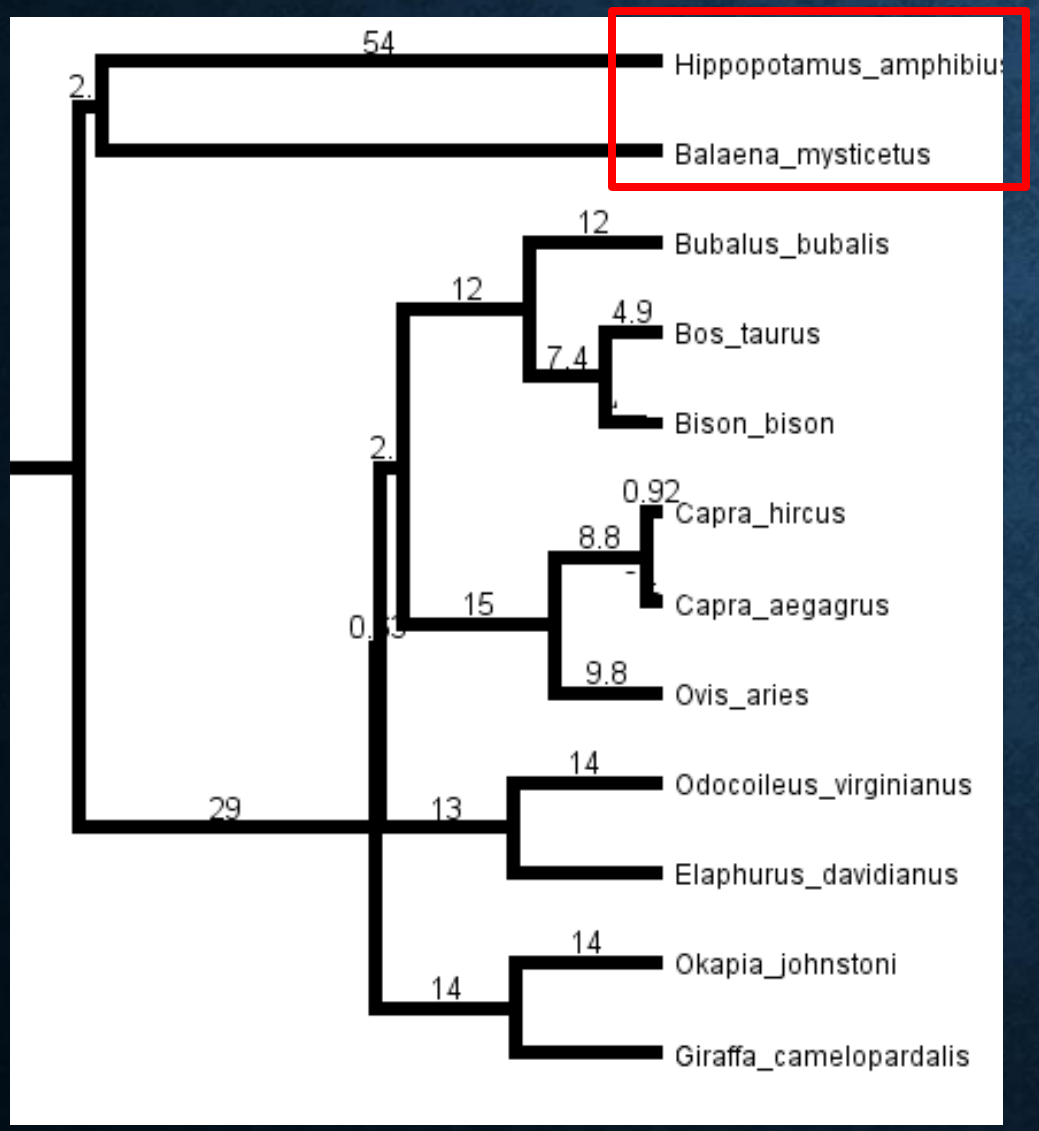
DATA COLLECTION



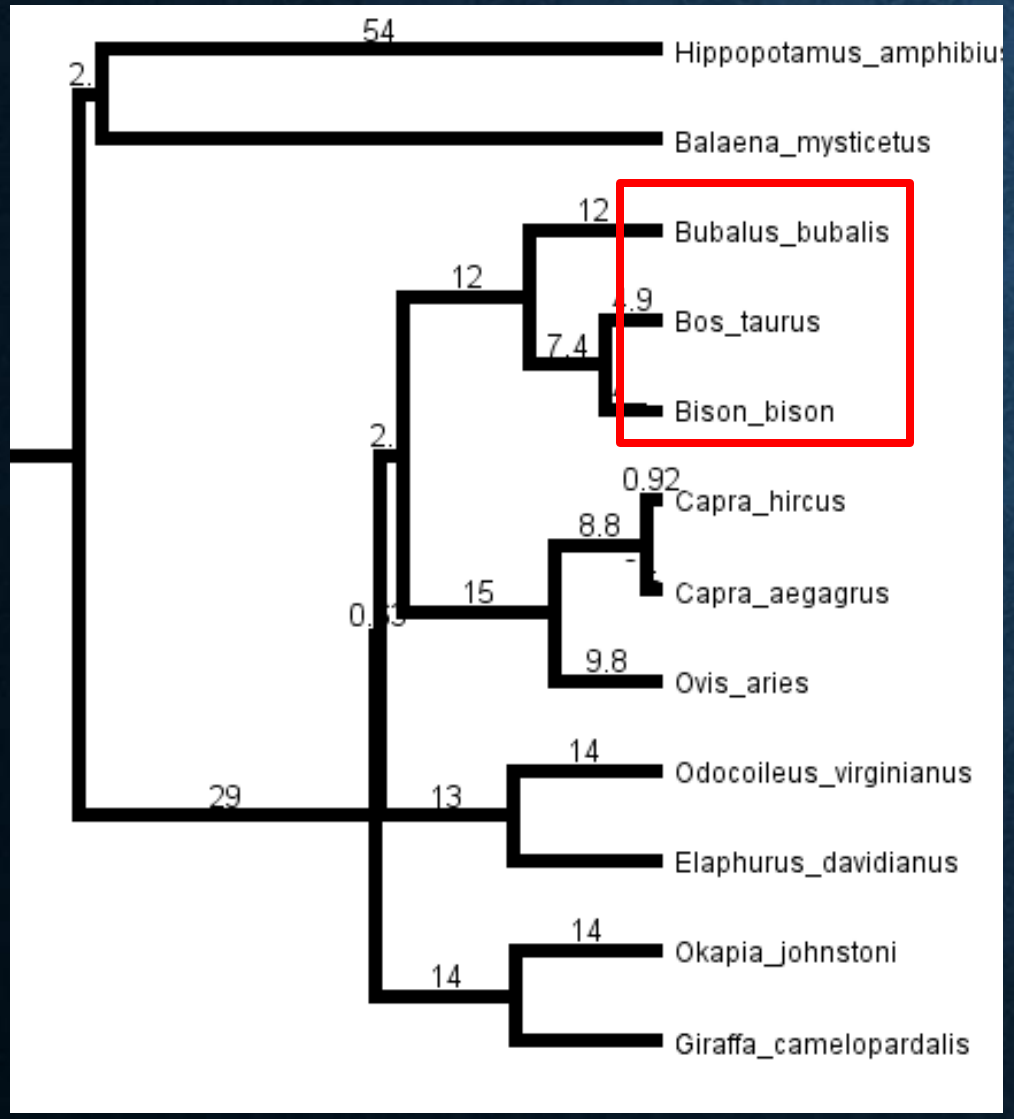
DATA COLLECTION



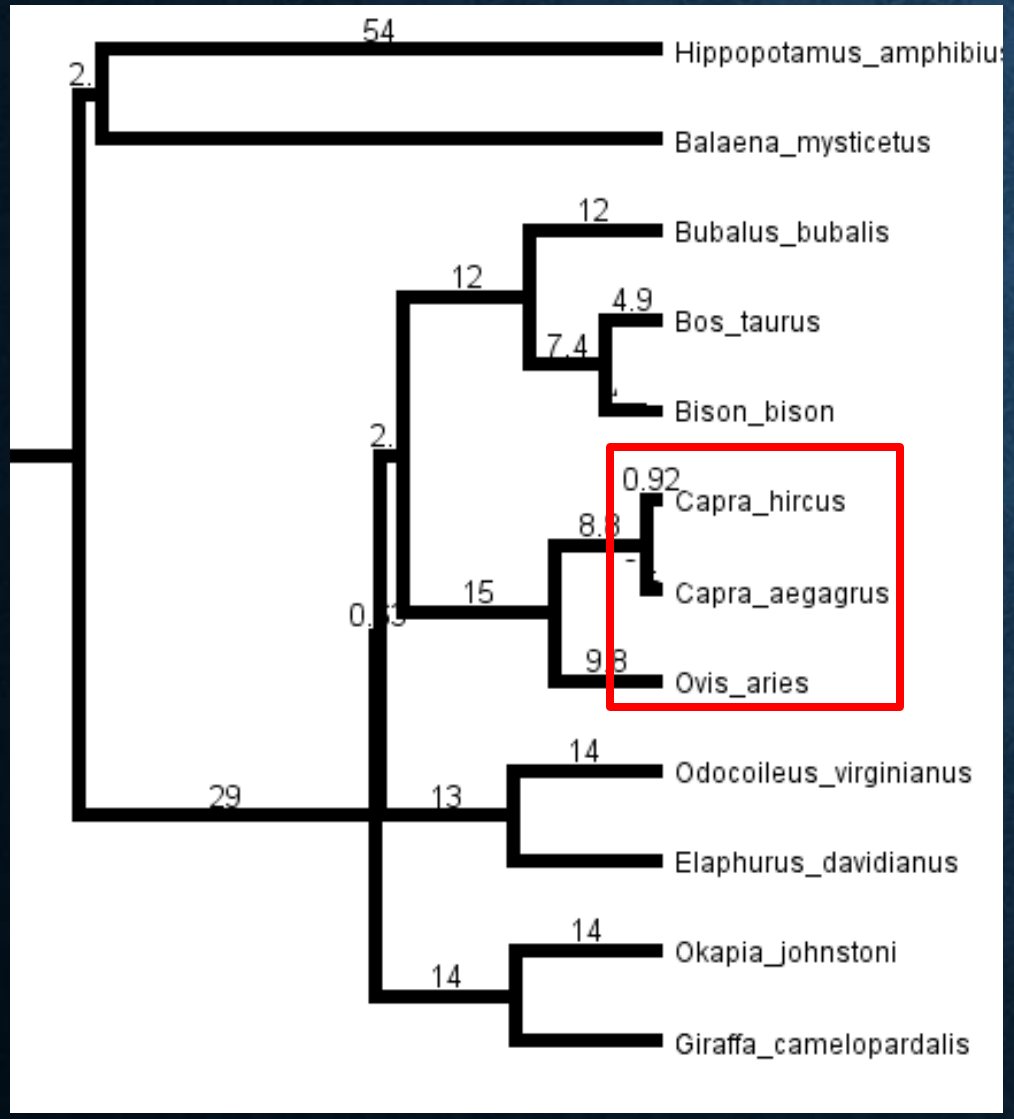
DATA COLLECTION



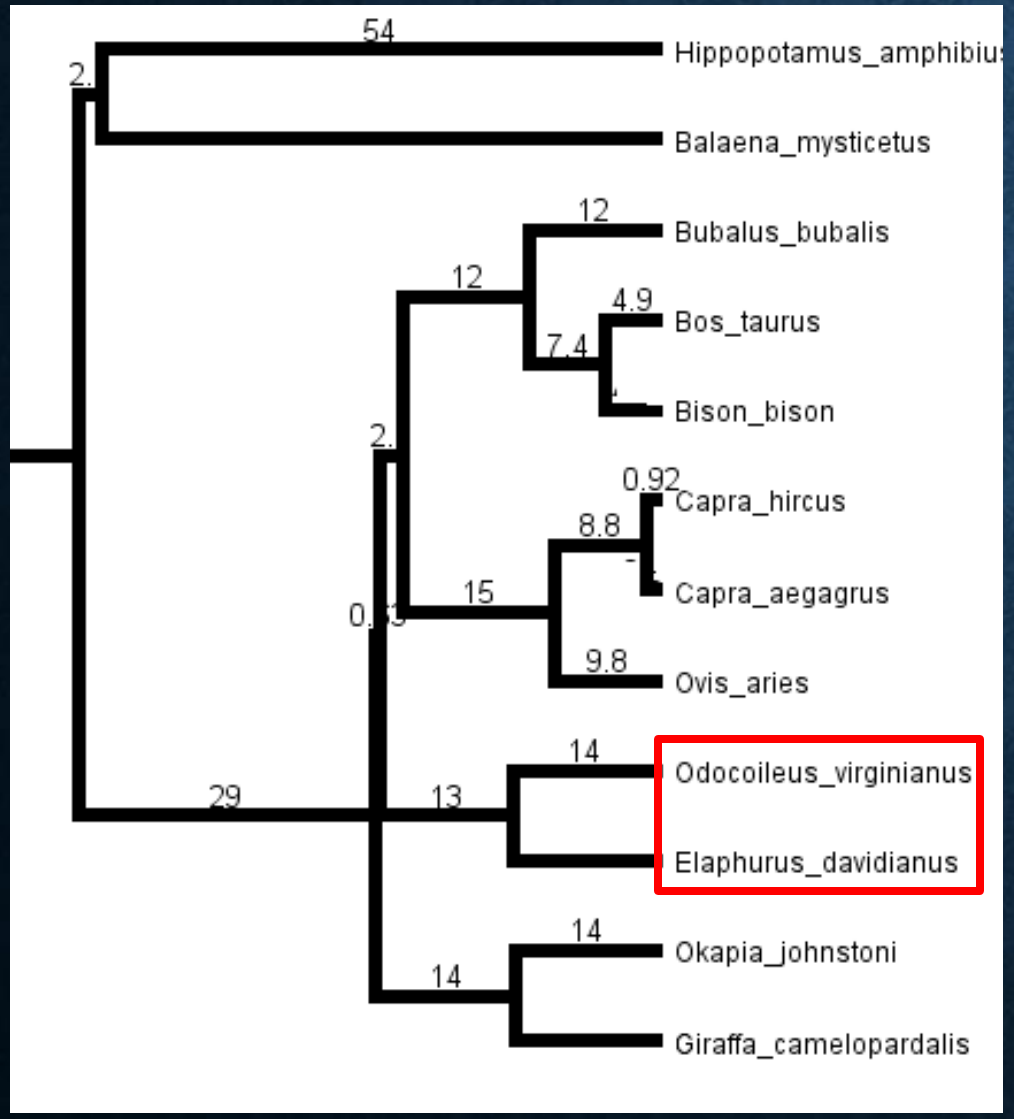
DATA COLLECTION



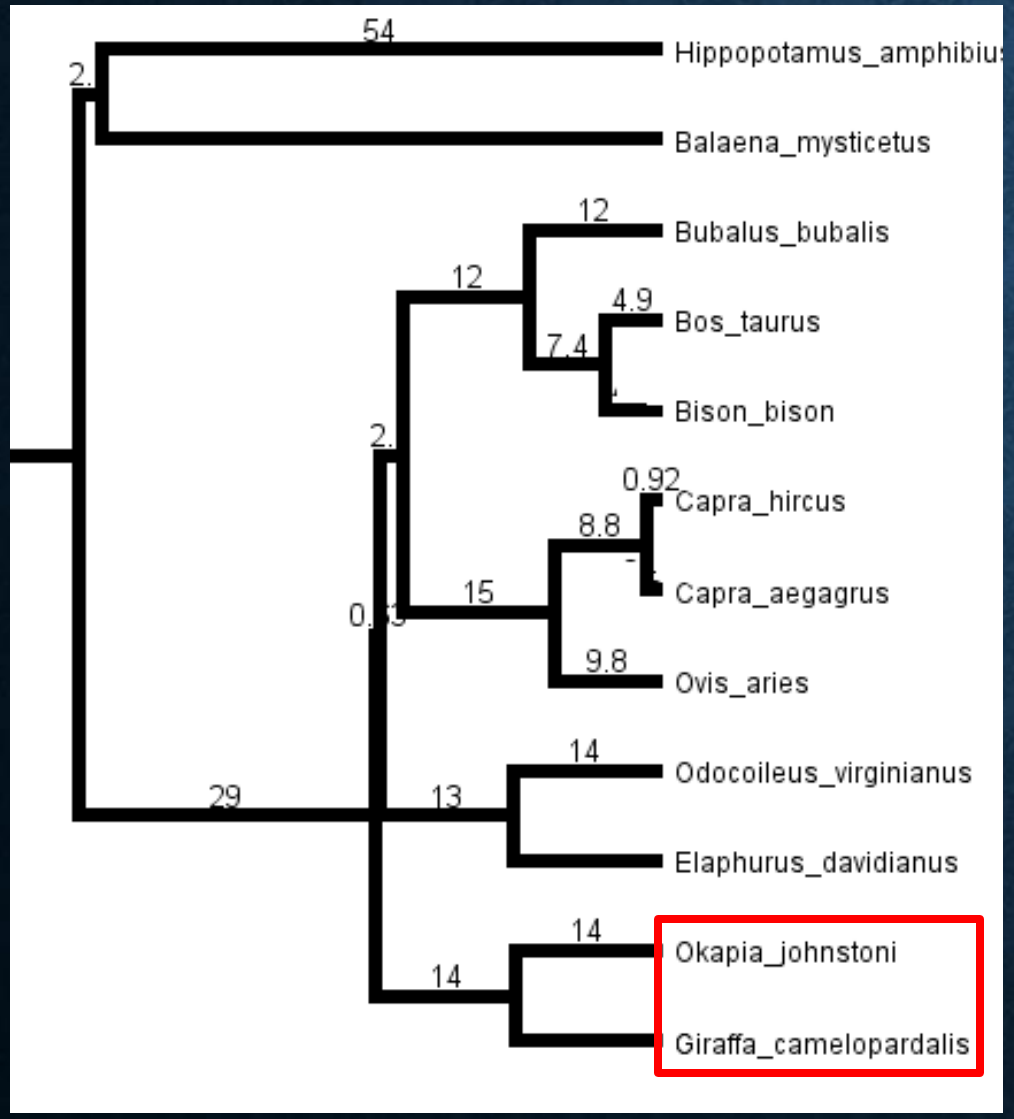
DATA COLLECTION

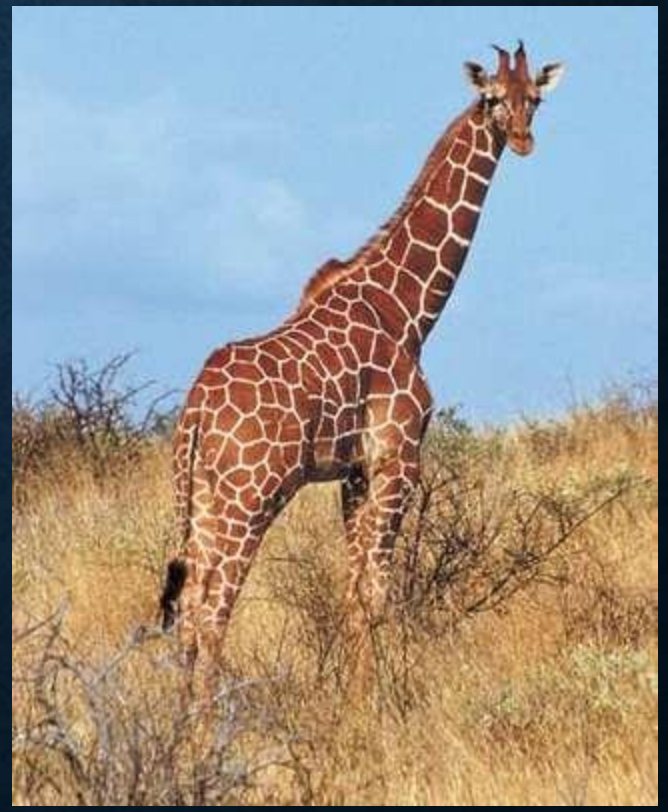


DATA COLLECTION



DATA COLLECTION



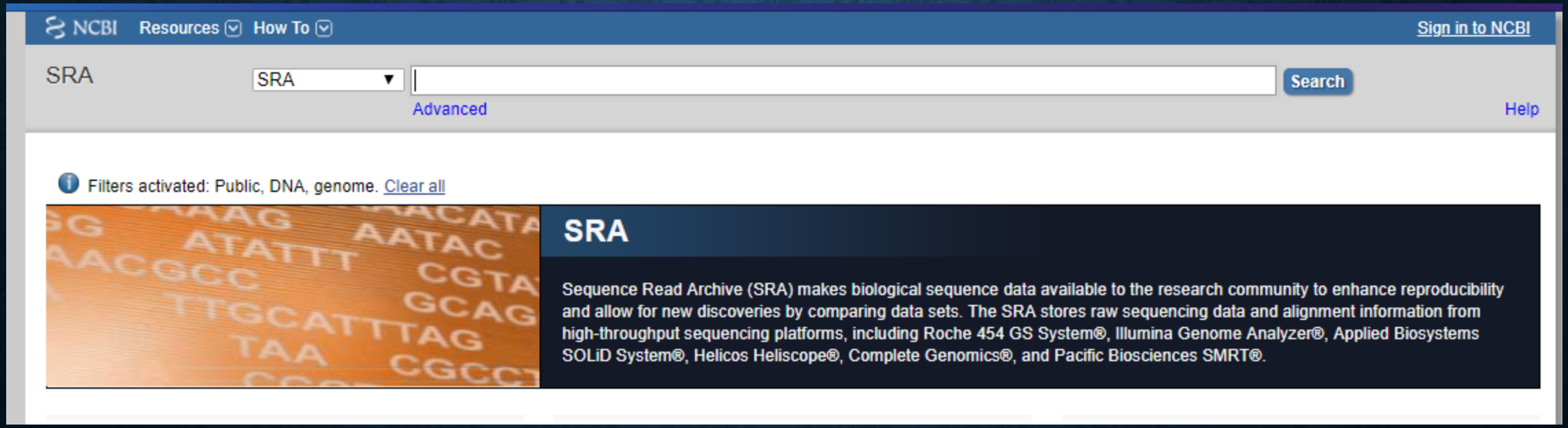


DATA COLLECTION

- Basic Processing Steps

DATA COLLECTION

- Basic Processing Steps
  1. Download DNA-Seq data from NCBI



DATA COLLECTION

- Basic Processing Steps
  1. Download DNA-Seq data from NCBI

```
@SRR5602579.205 205/1
AATGTATGCTGAAAAGGGTGGAAGCTCTTCCAGTTCTAGTCCTGACTTCAAGGCTGTTAAAGAAAAGGCTGAAGATATTGATAGAGGTTTGAGGGGTTTGAAGACTAGAGTGGAGGACACAGAGTTTCAAAGTAAAAGAATTCCTACTCTGGTTCAGGATATAAATGAGGAAAGAAAGAAATGGAGGAT
GTTAAAAGGCGACGTAGG
+
CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGFGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGFCGGGGGGGGGGGGGGGGGGGGFGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGEEEFG6
@SRR5602579.206 206/1
CCACGAATATGGGAAGCCCGGAACAGCGTTCGTCGGCTCGGTGTTTTTACGGAAATGCCCAGGGAGTTTTTTGTACCCTTTTCTTTGCACTATAGTTAATCTATTTTTTTTGTTTCACATTGGGGACAATGTGAACTTTAAGTGTGAGGAGGGGTGTTATTTTGTACACTTTGTTTCA
+
CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGFGGGGGGGGGGGGGGGGGGGGGGGGGGGGFGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGDGGGGGFGFGGGGGGGGGGGGGGCGEGG7CGEGGGFGGGGGFGGGFFGGGGGF
@SRR5602579.207 207/1
AGCCGGGAGTGAATTAATGTCCACTAACTATAGTAAAGTCTGAATCAACTTTGCTGGGGCACCAAAA
+
CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGFGG
@SRR5602579.208 208/1
TCTAAAATCATTTGCACACTCACAAGCAAACGAAAATTCCTTTCGTTTGCCACTTCAAACGAATGGGTCCACTTTCATTCGTTTGCATACTCACTAA
+
ACC@AEFGGGGGFCGGGGFEGGGFGCFGGGGECGGGGGCGFGGF7FGGDGGGGGGGGGGFGGGGFFDFEFGGGGGFDFGCCCFFGG,FFFGGGGGGG
@SRR5602579.209 209/1
TATACTAAGTCATTACTCGAAAATCCATTTTTCATTTTAACTTTCCAAAAAGTTACTATAATTAGTAAGTTGCTAGATTTTAACAACAATTCTACATCTTATATTTTCAAACTCTCAATTTCATCATAACATGTTCATTTAAGAATTATTTGTGCCAAAATATCATTTGGAAACTTGTTCTACAAAATC
ATCCCCAAAAAGTTTCCATTTTAGGCAACTATAAGAACATTAACATGAATCTTAGTTTTTCTACCACCATAATACTTCTAGATCATCCTAAAAAGATCGGAAGAGCAAAC
+
CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGFGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGFGGGGGGGGGGGGGGCGCGGGGGGGGGFGFGGGGGGGGGGGGGGGG
GGFGGGGGGGG>DFGGGGGGGFGFCDFFGGGGGGEFFFFFFFFFFFFFF?DFFFFFFFFDBFFBC@EFDBEBE<@EE=CC75>*)0>6:E@:@58>EF29>4-4)5)).5
```

DATA COLLECTION

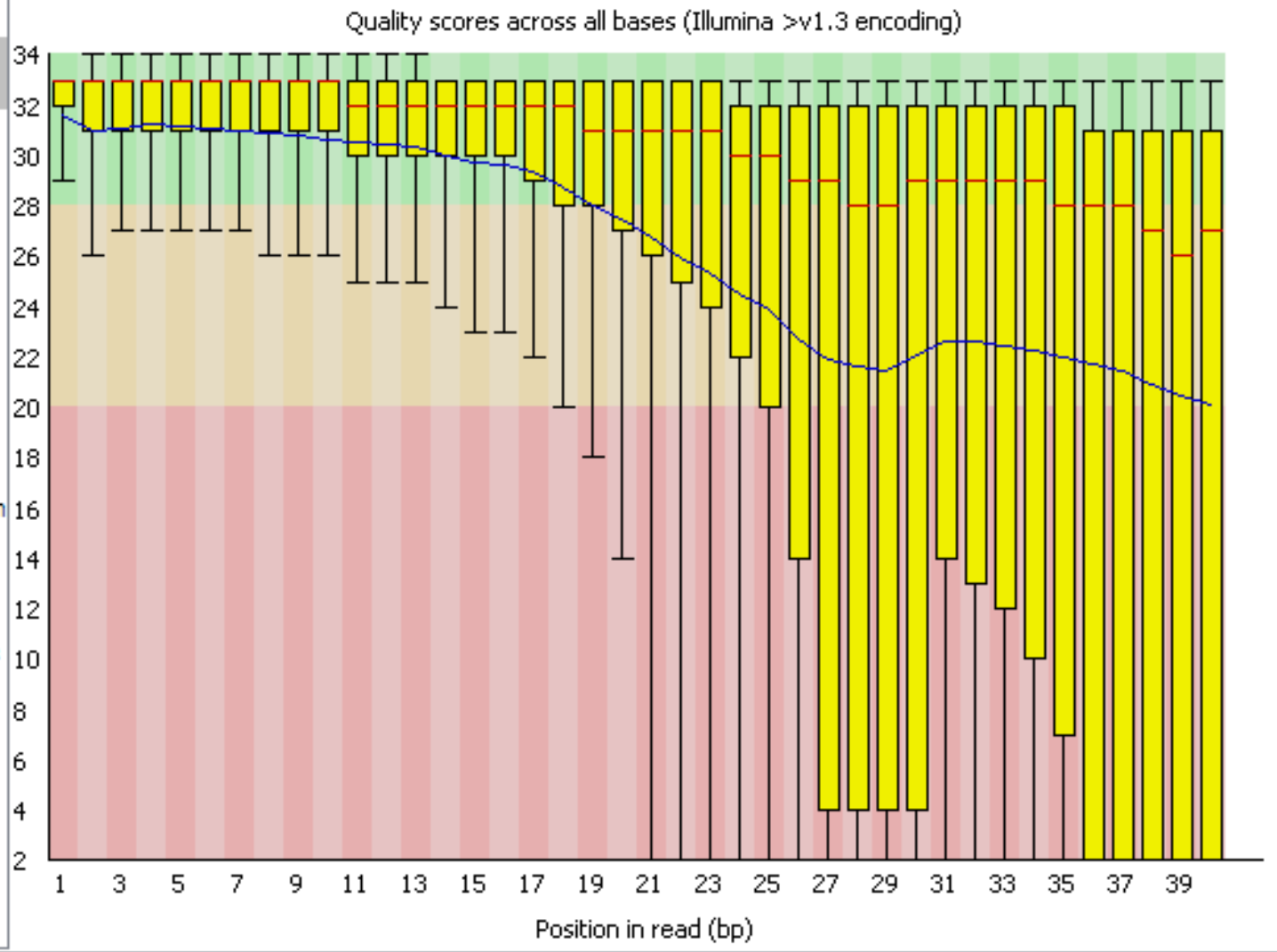
- Basic Processing Steps
  1. Download DNA-Seq data from NCBI
  2. Run FastQC (Look for red flags)

FastQC

File Help

bad_sequence.txt | good_sequence_short.txt

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content



DATA COLLECTION

- Basic Processing Steps
  1. Download DNA-Seq data from NCBI
  2. Run FastQC (Look for red flags)
  3. Trim Data with BBDuk (2.5 TRILLION BASES = 2.5 TB)

DATA COLLECTION

- Basic Processing Steps
  1. Download DNA-Seq data from NCBI
  2. Run FastQC (Look for red flags)
  3. Trim Data with BBDuk (2.5 TRILLION BASES = 2.5 TB)
  4. Assemble the composite genome using Ray

DATA COLLECTION

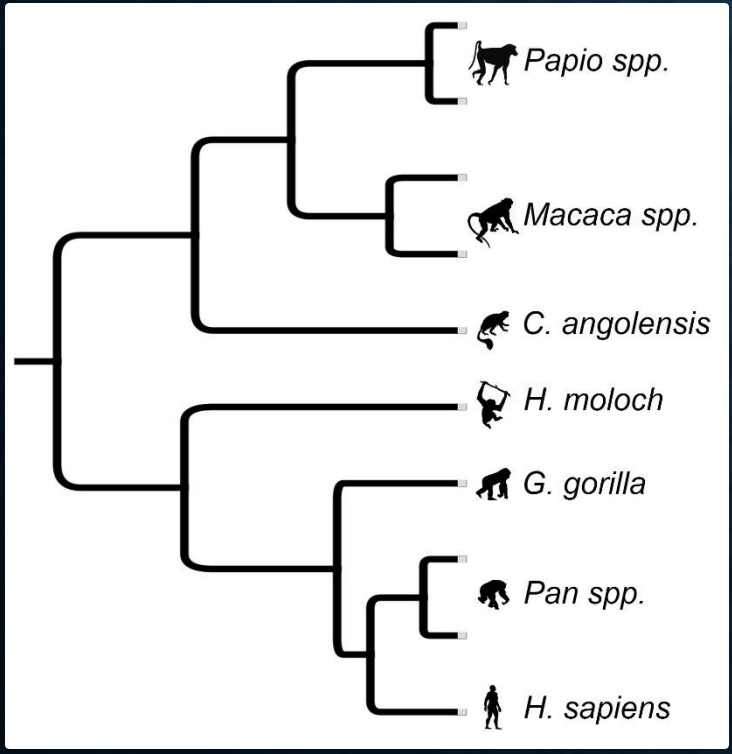
- Basic Processing Steps
  1. Download DNA-Seq data from NCBI
  2. Run FastQC (Look for red flags)
  3. Trim Data with BBDuk (2.5 TRILLION BASES = 2.5 TB)
  4. Assemble the composite genome using Ray
  5. Run data through SISRS Site Filtration

DATA COLLECTION

- Basic Processing Steps
  1. Download DNA-Seq data from NCBI
  2. Run FastQC (Look for red flags)
  3. Trim Data with BBDuk (2.5 TRILLION BASES = 2.5 TB)
  4. Assemble the composite genome using Ray
  5. Run data through SISRS Site Filtration
  6. Extract SISRS sites for analysis

DATA PROCESSING

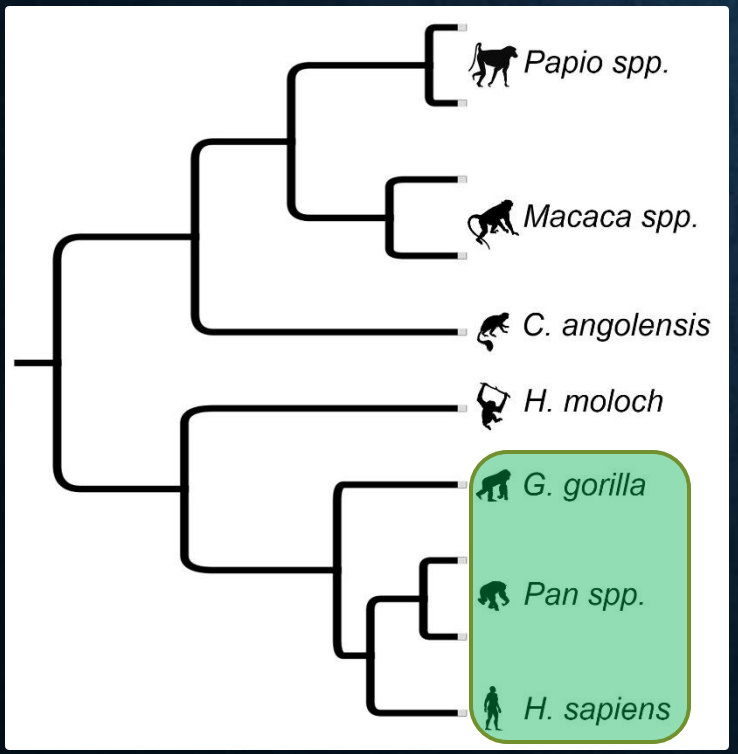
- Each SISRS site was scored as either in agreement or disagreement with the reference tree



1. AotNan	A	C	C	C	T	G	C	A	C	A
2. CalJac	A	C	C	C	T	G	C	G	C	A
3. ColAng	A	C	C	T	C	T	T	G	T	A
4. PapAnu	A	C	C	T	C	T	T	A	T	A
5. PapCyn	A	C	C	T	C	T	T	A	T	A
6. MacMul	A	T	C	T	C	T	T	G	T	A
7. MacNem	A	T	C	T	C	T	T	G	T	A
8. HylMol	A	C	C	T	C	T	T	G	T	A
9. GorGor	G	T	C	T	C	T	T	G	T	A
10. PanPan	G	C	T	T	C	T	T	G	T	G
11. PanTro	G	C	T	T	C	T	T	G	T	G
12. HomSap	G	C	T	T	C	T	T	G	T	G

DATA PROCESSING

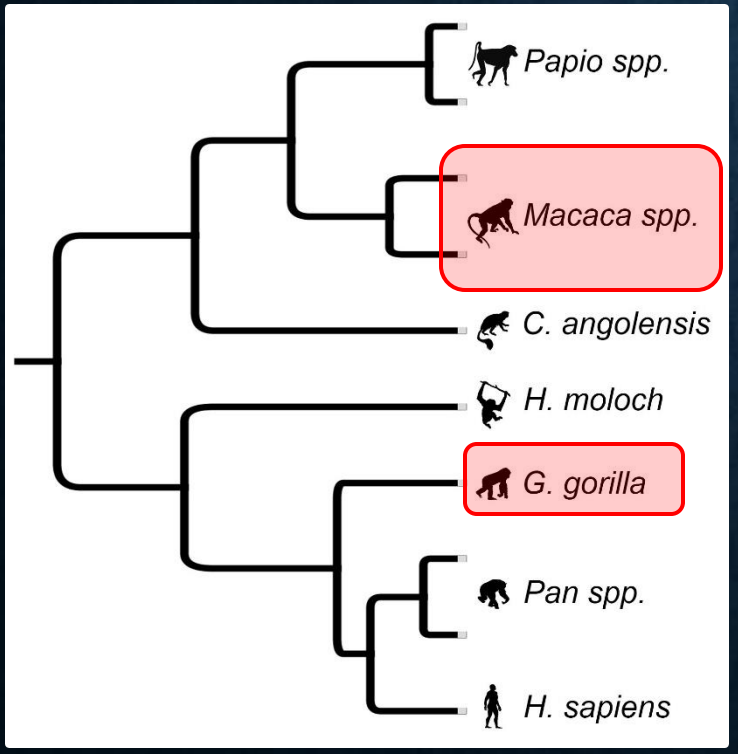
- Each SISRS site was scored as either in agreement or disagreement with the reference tree



1. AotNan	A	C	C	C	T	G	C	A	C	A
2. CalJac	A	C	C	C	T	G	C	G	C	A
3. ColAng	A	C	C	T	C	T	T	G	T	A
4. PapAnu	A	C	C	T	C	T	T	A	T	A
5. PapCyn	A	C	C	T	C	T	T	A	T	A
6. MacMul	A	T	C	T	C	T	T	G	T	A
7. MacNem	A	T	C	T	C	T	T	G	T	A
8. HylMol	A	C	C	T	C	T	T	G	T	A
9. GorGor	G	T	C	T	C	T	T	G	T	A
10. PanPan	G	C	T	T	C	T	T	G	T	G
11. PanTro	G	C	T	T	C	T	T	G	T	G
12. HomSap	G	C	T	T	C	T	T	G	T	G

DATA PROCESSING

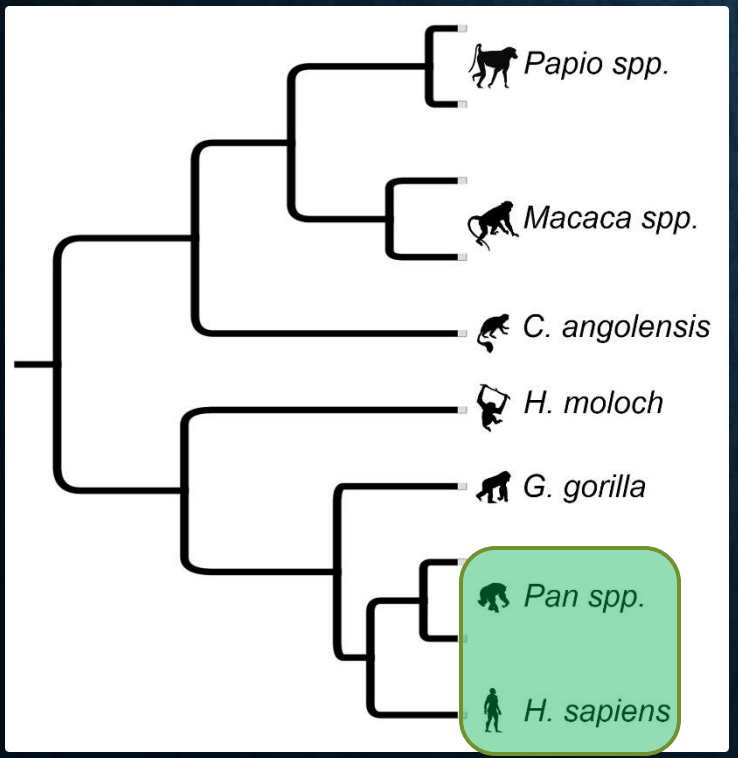
- Each SISRS site was scored as either in agreement or disagreement with the reference tree



1. AotNan	A	C	C	C	T	G	C	A	C	A
2. CalJac	A	C	C	C	T	G	C	G	C	A
3. ColAng	A	C	C	T	C	T	T	G	T	A
4. PapAnu	A	C	C	T	C	T	T	A	T	A
5. PapCyn	A	C	C	T	C	T	T	A	T	A
6. MacMul	A	T	C	T	C	T	T	G	T	A
7. MacNem	A	T	C	T	C	T	T	G	T	A
8. HylMol	A	C	C	T	C	T	T	G	T	A
9. GorGor	G	T	C	T	C	T	T	G	T	A
10. PanPan	G	C	T	T	C	T	T	G	T	G
11. PanTro	G	C	T	T	C	T	T	G	T	G
12. HomSap	G	C	T	T	C	T	T	G	T	G

DATA PROCESSING

- Each SISRS site was scored as either in agreement or disagreement with the reference tree



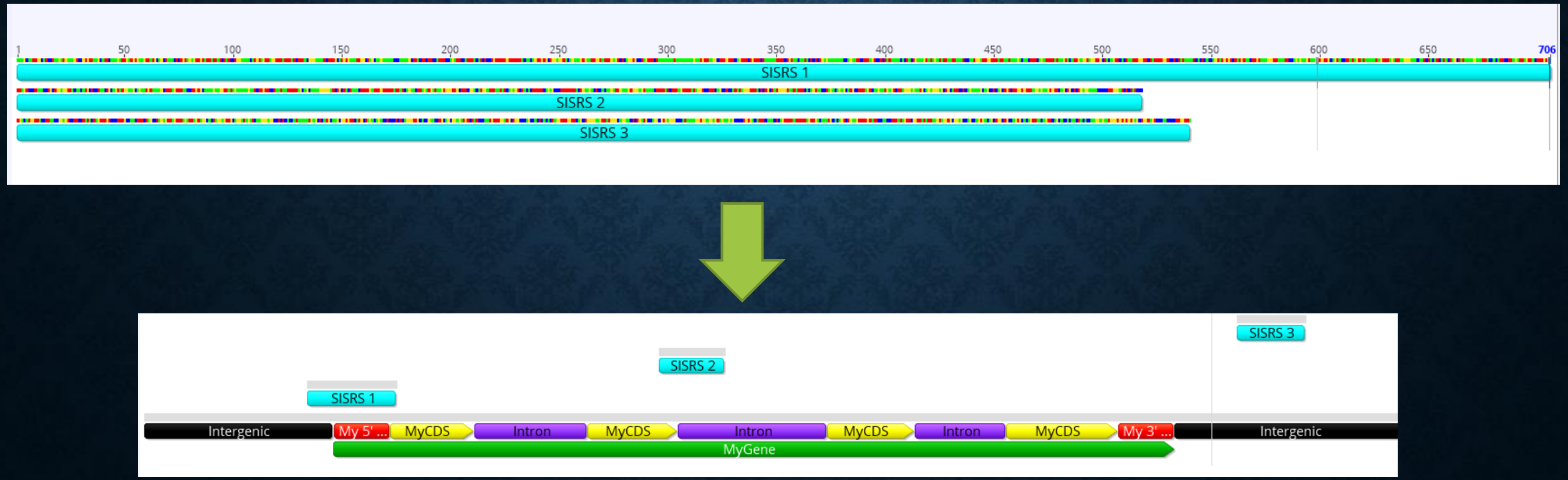
1. AotNan	A	C	C	C	T	G	C	A	C	A
2. CalJac	A	C	C	C	T	G	C	G	C	A
3. ColAng	A	C	C	T	C	T	T	G	T	A
4. PapAnu	A	C	C	T	C	T	T	A	T	A
5. PapCyn	A	C	C	T	C	T	T	A	T	A
6. MacMul	A	T	C	T	C	T	T	G	T	A
7. MacNem	A	T	C	T	C	T	T	G	T	A
8. HylMol	A	C	C	T	C	T	T	G	T	A
9. GorGor	G	T	C	T	C	T	T	G	T	A
10. PanPan	G	C	T	T	C	T	T	G	T	G
11. PanTro	G	C	T	T	C	T	T	G	T	G
12. HomSap	G	C	T	T	C	T	T	G	T	G

DATA PROCESSING

- Every site in the composite genome was mapped to the reference genome

DATA PROCESSING

- Every site in the composite genome was mapped to the reference genome

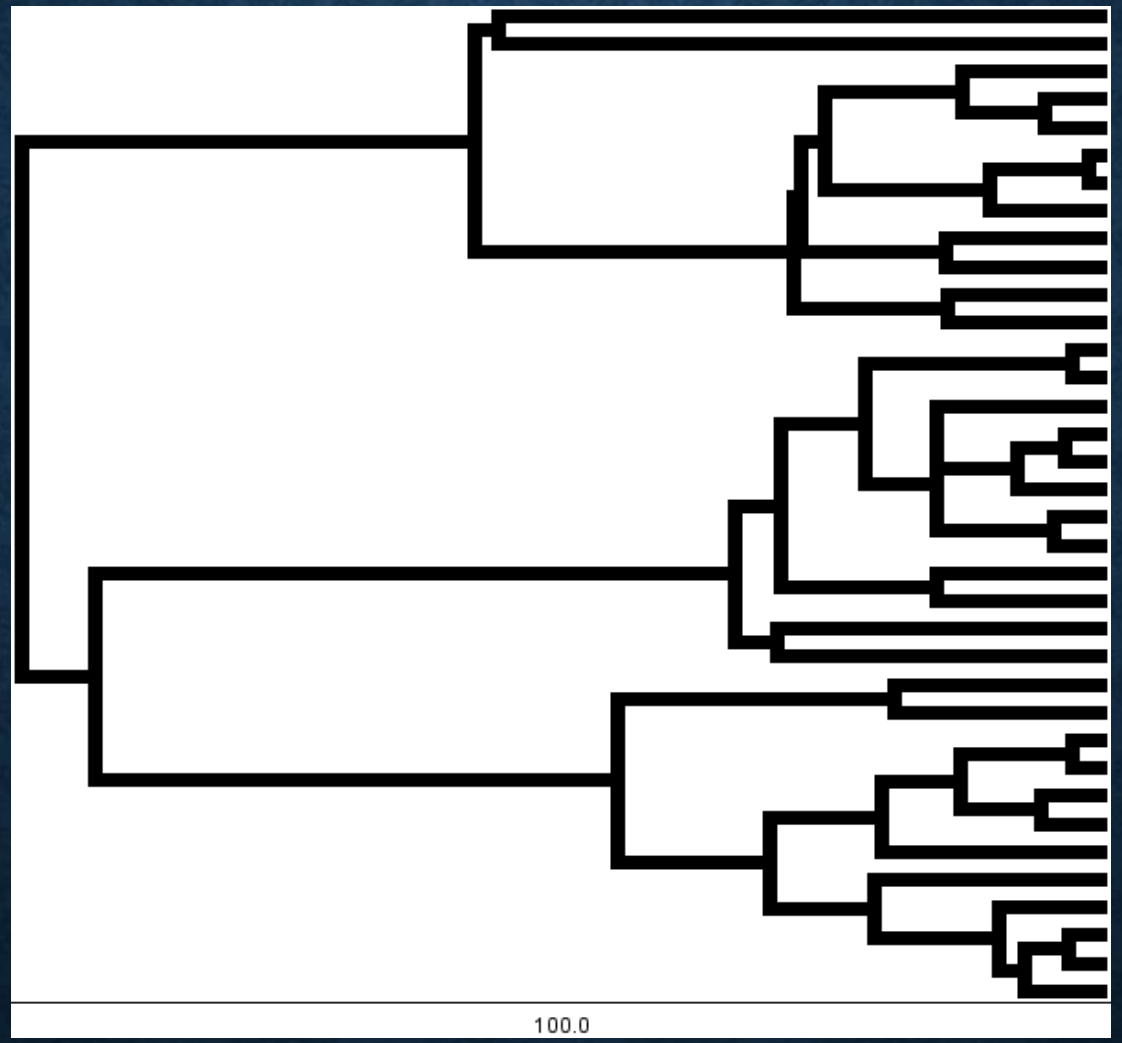


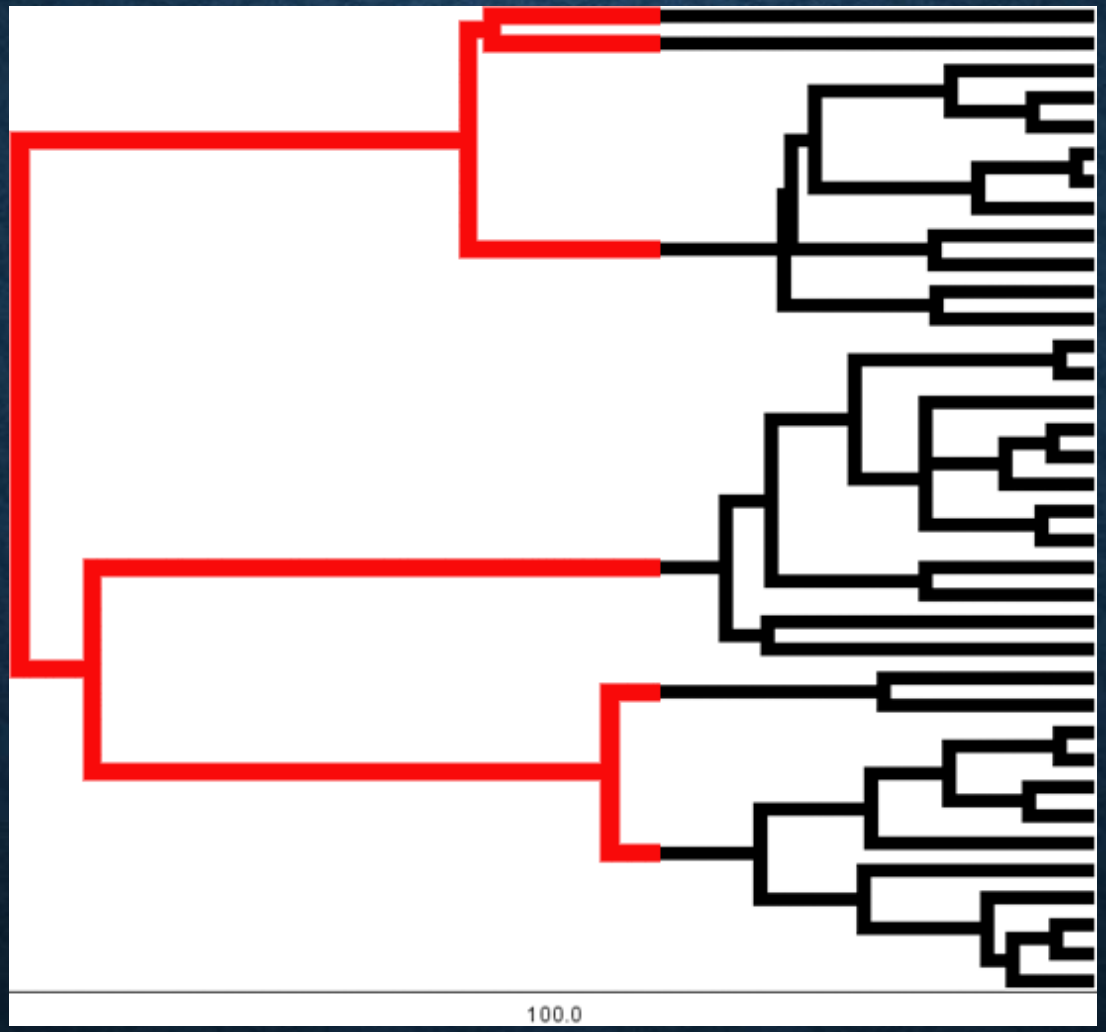
DATA PROCESSING

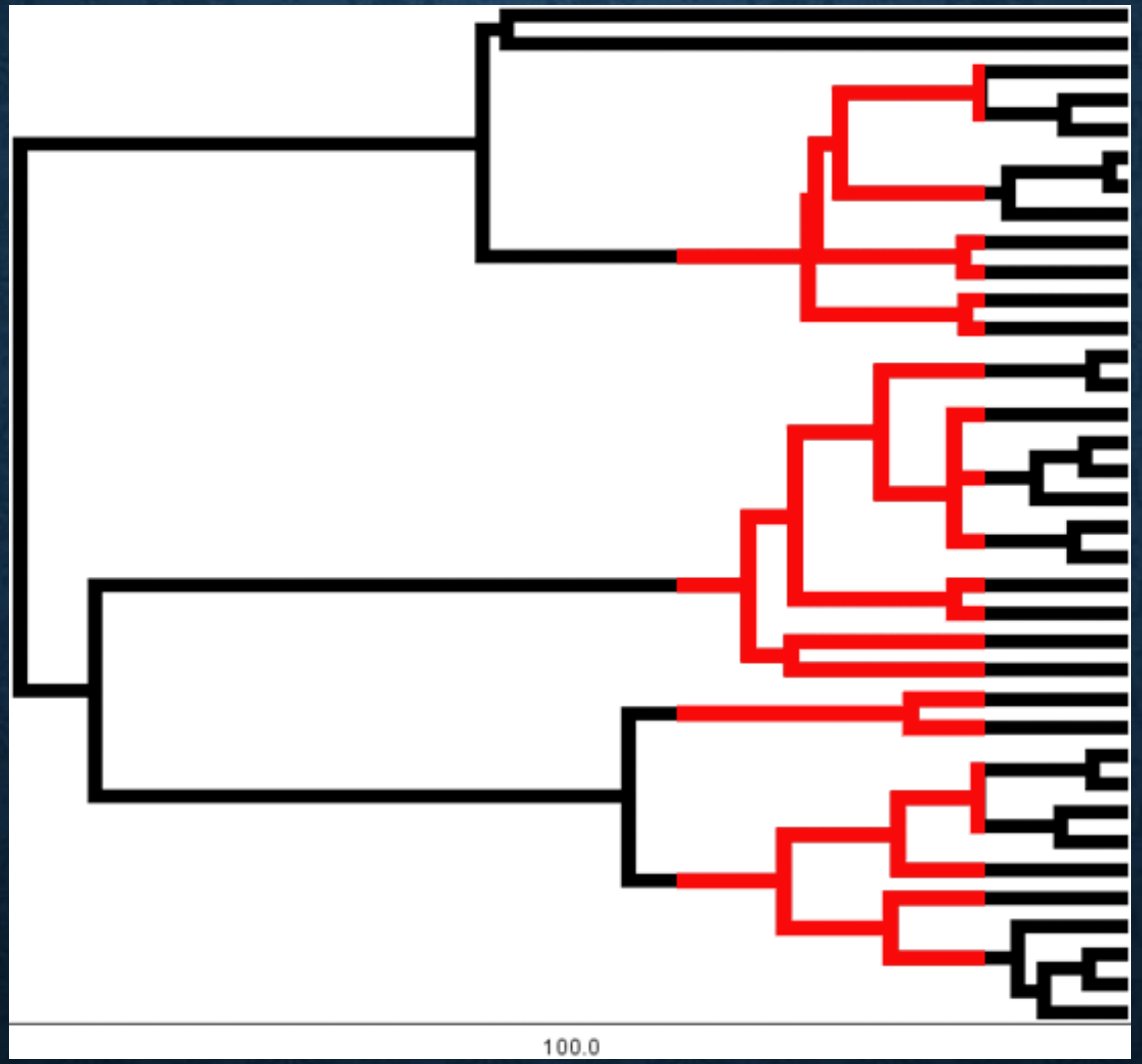
- Every site in the composite genome was mapped to the reference genome

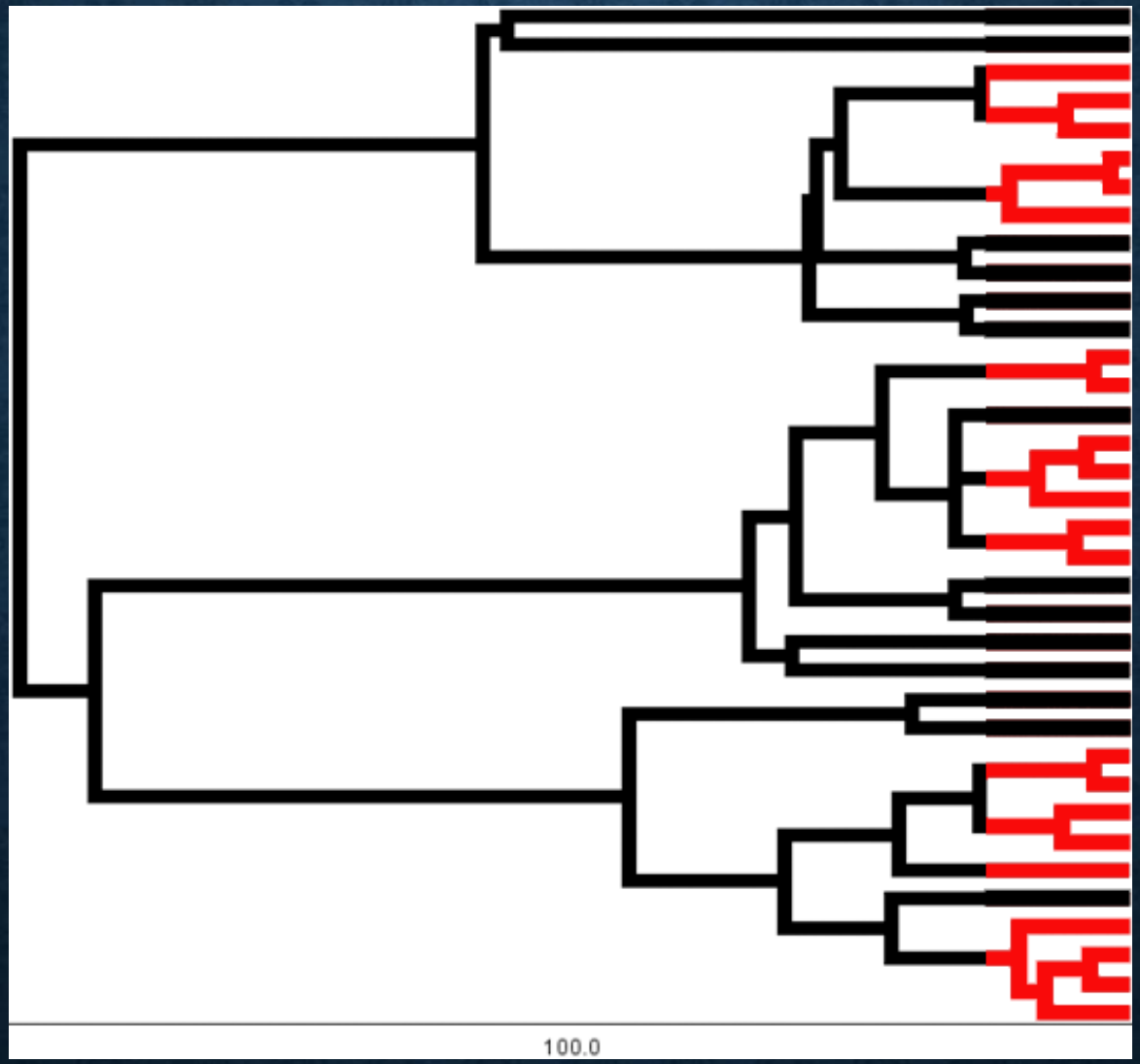
DATA PROCESSING

- Each SISRS site was assigned to one of nine annotation categories
- Reference trees were dated using ‘ape’ in R
- Timed trees (chronograms) were used to analyze differences in annotation utility throughout time


100.0


100.0


100.0


100.0

RESULTS

How many sites did we assemble?

RESULTS

Group	Composite Scaffolds	Mapped Bases	Final SISRS Bases
Primates	3.4M	424M (88%)	11.7M (2.7%)
Rodents	6.5M	372M (39%)	3.12M (0.84%)
Cetartiodactyla	3.4M	348M (73%)	10.2M (2.93%)
Combined	2.2M	110M (34%)	315K (0.29%)

RESULTS

Were certain types of annotations easier or harder for Ray to assemble?

RESULTS

	Composite Genome Annotations: Percent of Reference								
Group	**CDS**	**5'UTR**	**Intergenic**	**Intronic**	**lncRNA**	**ncGenes**	**pGenes**	**smRNA**	**3'UTR**
Primates	28.9%	16.6%	13.6%	18.2%	18.1%	19.0%	6.9%	12.9%	24.0%
Rodents	23.3%	19.8%	13.0%	20.0%	20.0%	21.4%	5.9%	13.6%	24.6%
Cetartiodactyla	24.5%	20.9%	14.2%	16.8%			6.1%	14.4%	23.8%
Combined	10.8%	4.4%	2.9%	4.3%	4.4%	4.4%	1.6%	2.4%	7.3%

Ray had a relatively hard time assembling Intergenic regions, Pseudogenes, and small RNAs

RESULTS

Were certain types of annotations more likely to be filtered in or out by SISRS?

RESULTS

	SISRS Annotations: Percent of Composite								
Group	**CDS**	**5'UTR**	**Intergenic**	**Intronic**	**lncRNA**	**ncGenes**	**pGenes**	**smRNA**	**3'UTR**
Primates	1.18%	1.57%	2.57%	2.96%	2.86%	2.80%	1.75%	1.04%	2.09%
Cetartiodactyla	2.43%	1.98%	2.83%	3.06%			0.91%	1.06%	2.79%
Rodents	4.10%	1.56%	0.73%	0.81%	0.94%	0.87%	0.50%	0.85%	1.81%
Combined	2.46%	1.12%	0.20%	0.20%	0.34%	0.47%	0.14%	0.51%	0.95%

- Groups varied with respect to which annotations were filtered in by SISRS
- Intergenic regions were enriched in all datasets

RESULTS

Were certain types of annotations more likely to give us reliable split information?

RESULTS

	SISRS Annotations: Percent that Support True Splits in the Tree								
Group	**CDS**	**5'UTR**	**Intergenic**	**Intronic**	**lncRNA**	**ncGenes**	**pGenes**	**smRNA**	**3'UTR**
Primates	88.4%	90.5%	90.0%	90.1%	90.2%	90.1%	90.3%	88.7%	90.3%
Rodents	69.7%	78.9%	79.4%	79.2%	78.3%	79.7%	78.1%	78.1%	78.2%
Cetartiodactyla	79.6%	85.3%	84.5%	84.1%			85.5%	84.2%	84.9%
Combined	56.2%	66.1%	71.7%	71.5%	67.7%	69.0%	66.4%	74.8%	64.4%

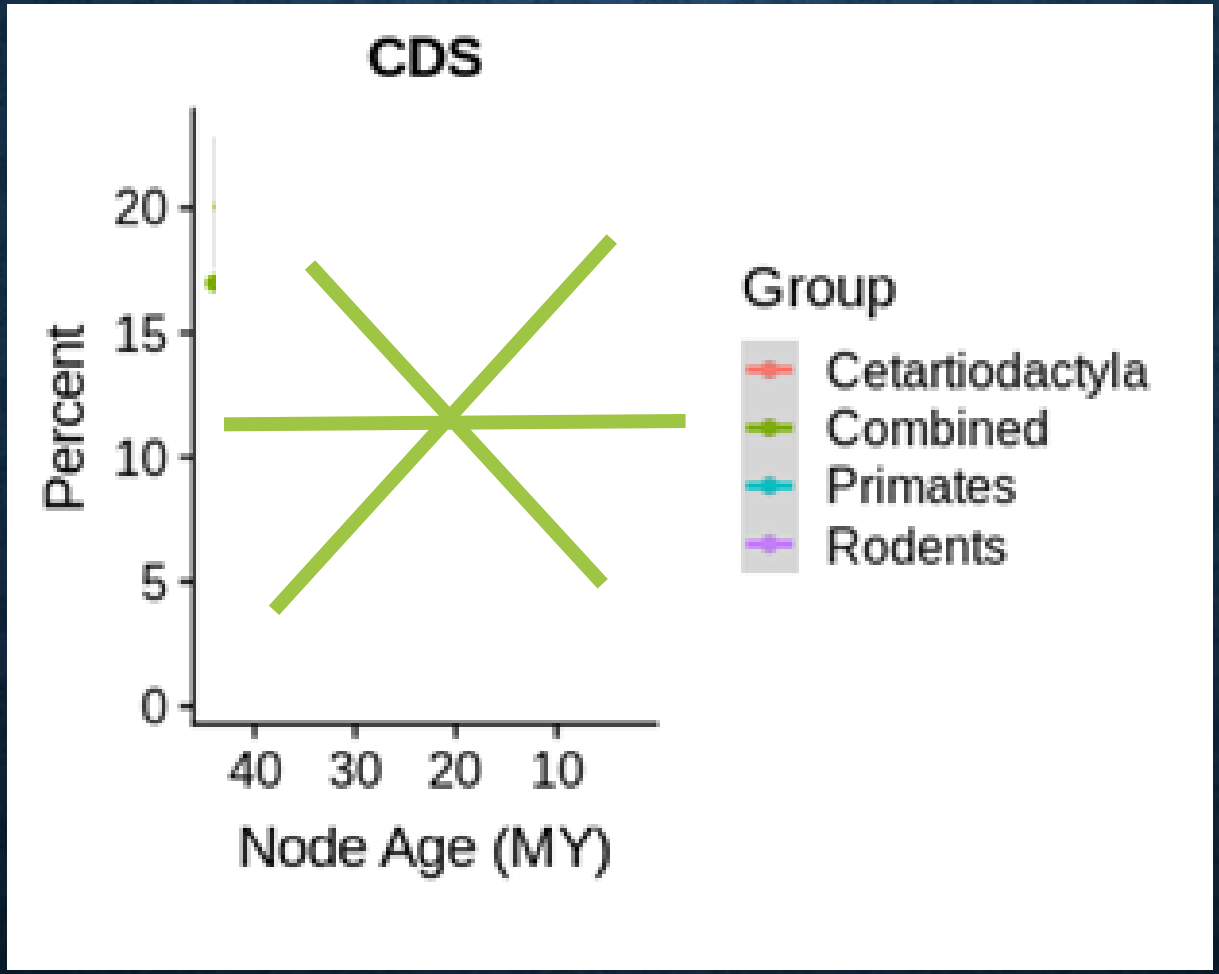
- CDS sites were disproportionately unhelpful
- Intergenic sites were disproportionately helpful in 3 / 4
- Introns vary in their utility

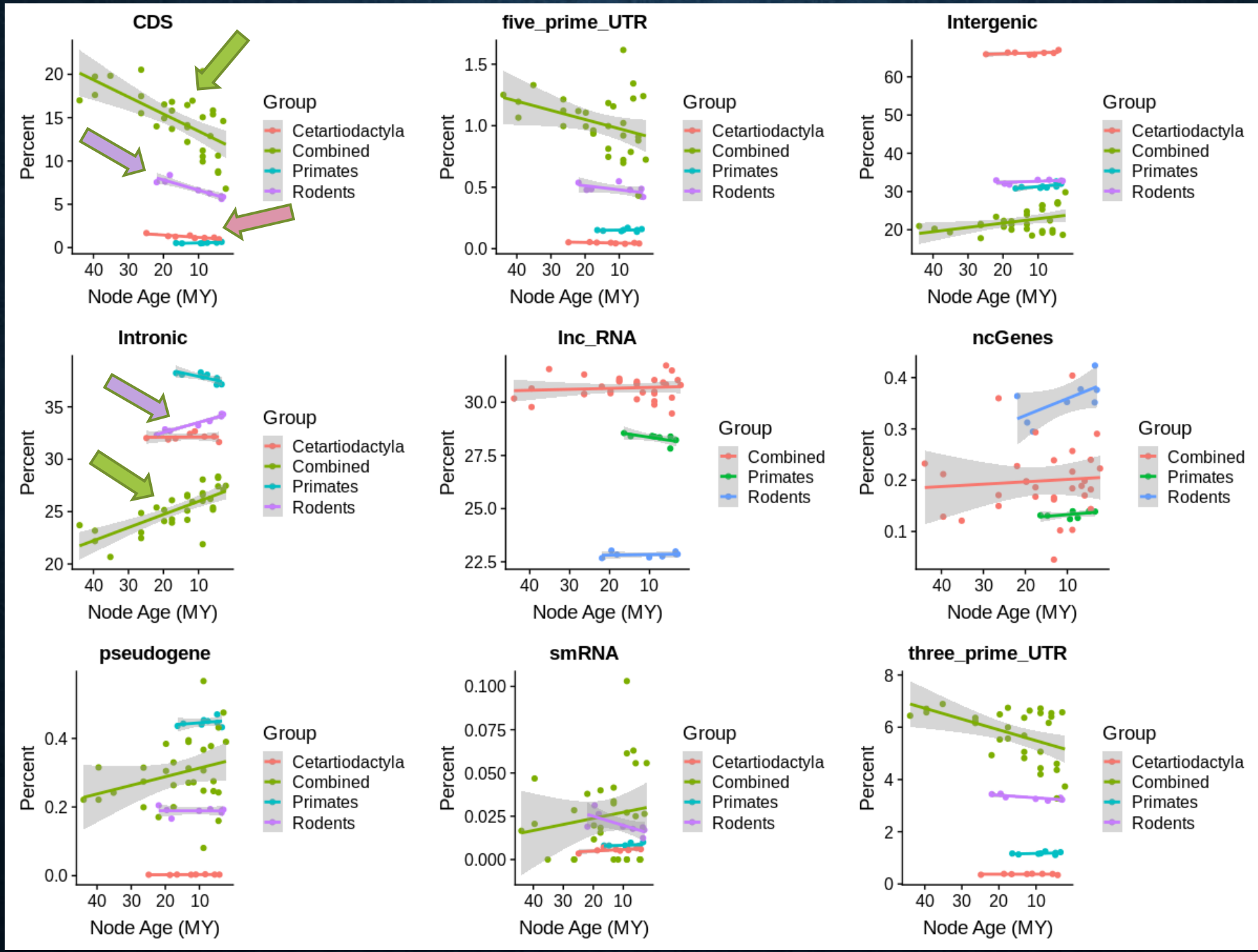
RESULTS

- 33/34 annotation subsets were sufficient to reconstruct the correct species tree
    - Combined smRNA did not
    - There were only 89 total sites
- SISRS filtering reduces genome-scale datasets down to workable and accurate subsets

RESULTS

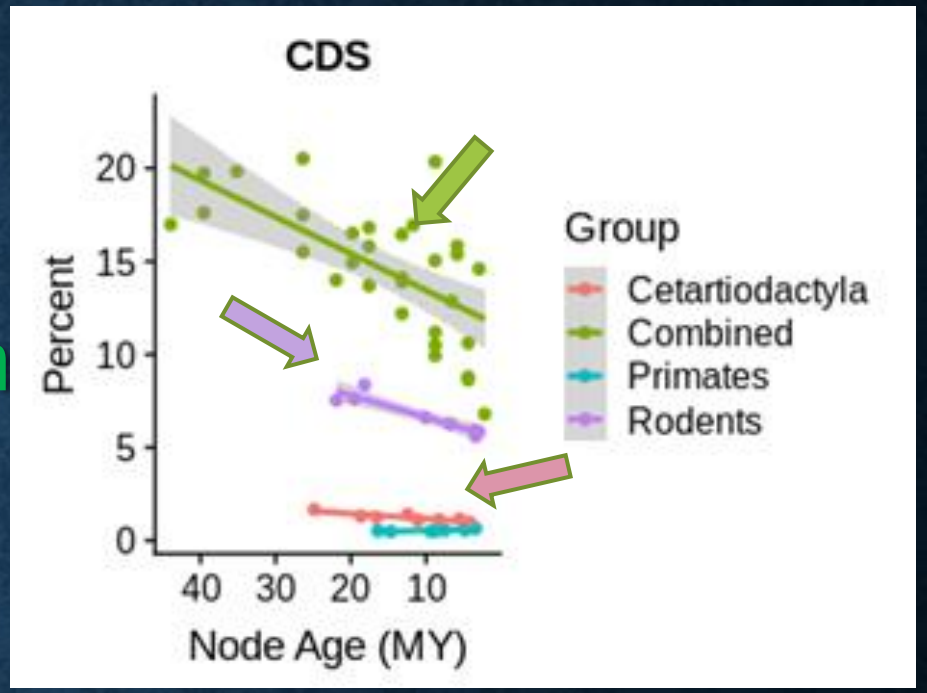
- Are certain annotation types disproportionately shifted towards supporting older or younger nodes?


CDS
Percent
20
15
10
5
0
40
30
20
10
Node Age (MY)
Group
Cetartiodactyla
Combined
Primates
Rodents


CDS
five_prime_UTR
Intergenic
Intronic
lnc_RNA
ncGenes
pseudogene
smRNA
three_prime_UTR
Percent
Node Age (MY)
Group
Cetartiodactyla
Combined
Primates
Rodents

RESULTS

- Most annotation types show no differen
- **Introns**
  - Rodents & Combined
  - As you move from old to young, ~4.8% rise over 43MY
- **CDS**
  - Rodent & Combined + Cetartiodactyla
  - As you move from young to old, ~6.7% rise over 43MY (R/Co.)
  - As you move from young to old, 0.68% rise over 25MY (Ce.)



CONCLUSIONS

- SISRS filtered sites were largely accurate
  - Primates: 90%
  - Rodents: 78%
  - Cetartiodactyla: 84%
  - Combined: 67%

CONCLUSIONS

- What was useful for one group was not always as useful for another
  - Can’t use one rule for choosing the best loci
  - Data-driven site selection takes the guesswork out

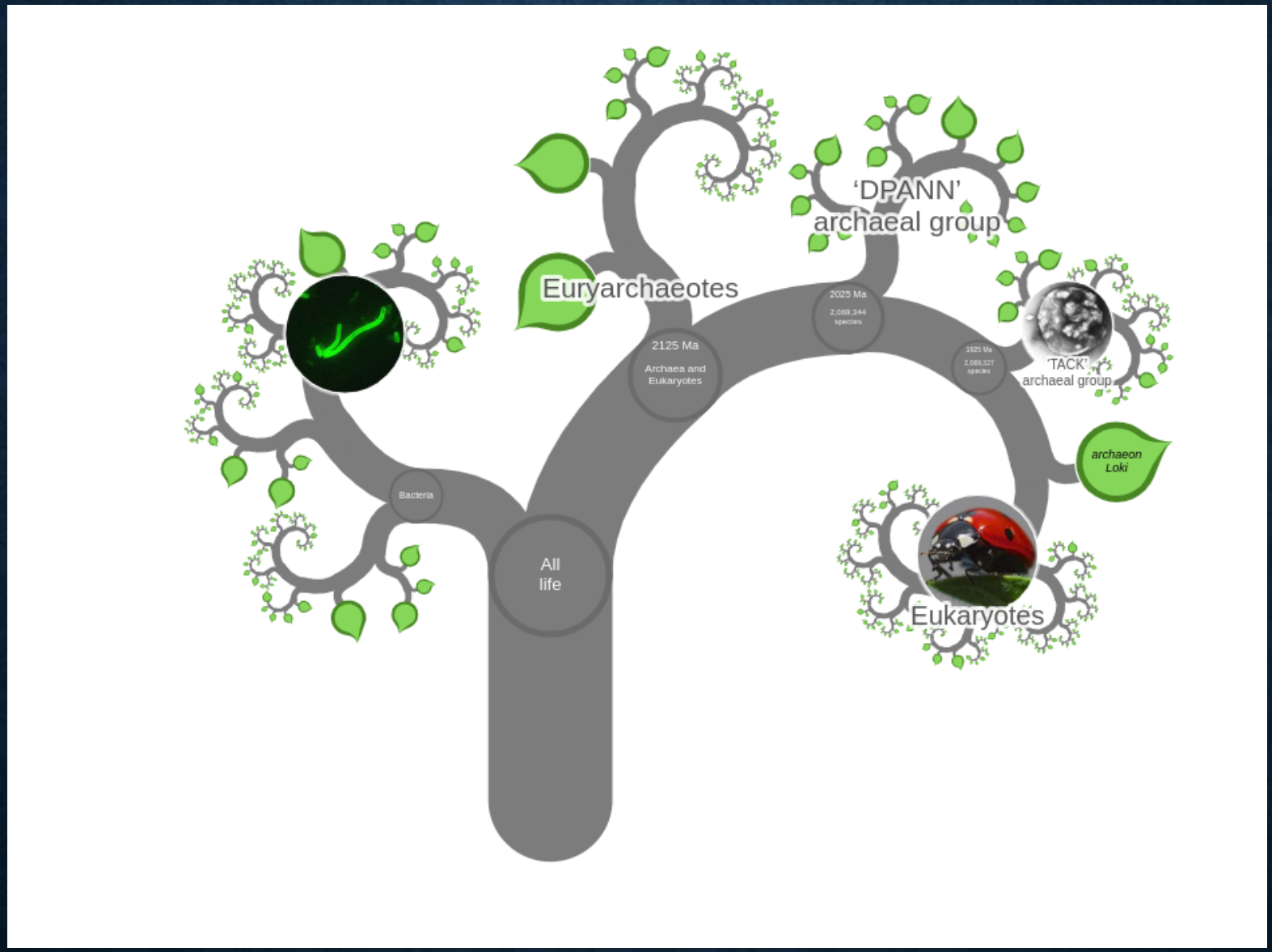
CONCLUSIONS

- Slower evolving CDS were useful for older nodes in 3 out of 4 datasets
- Faster evolving introns were useful for younger nodes in 2 out of 4 datasets

FUTURE WORK

- In the future, I will:
  - Investigate the nature of loci that provide phylogenetic signal versus noise
  - Implement additional SISRS filtration steps to further reduce noise
  - Explore why rodents behaved so much differently
  - Extend the work beyond mammals

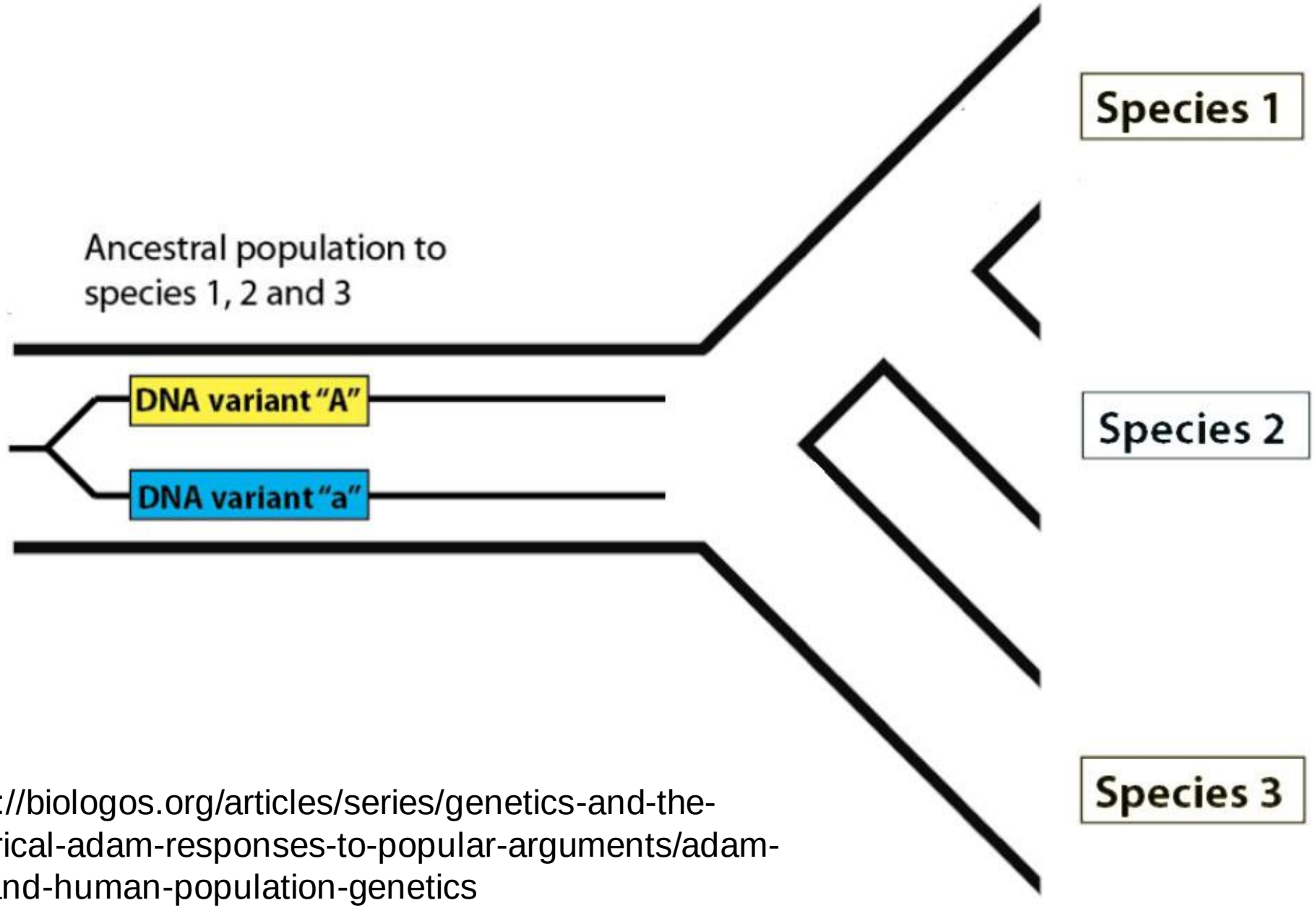

QUESTIONS?



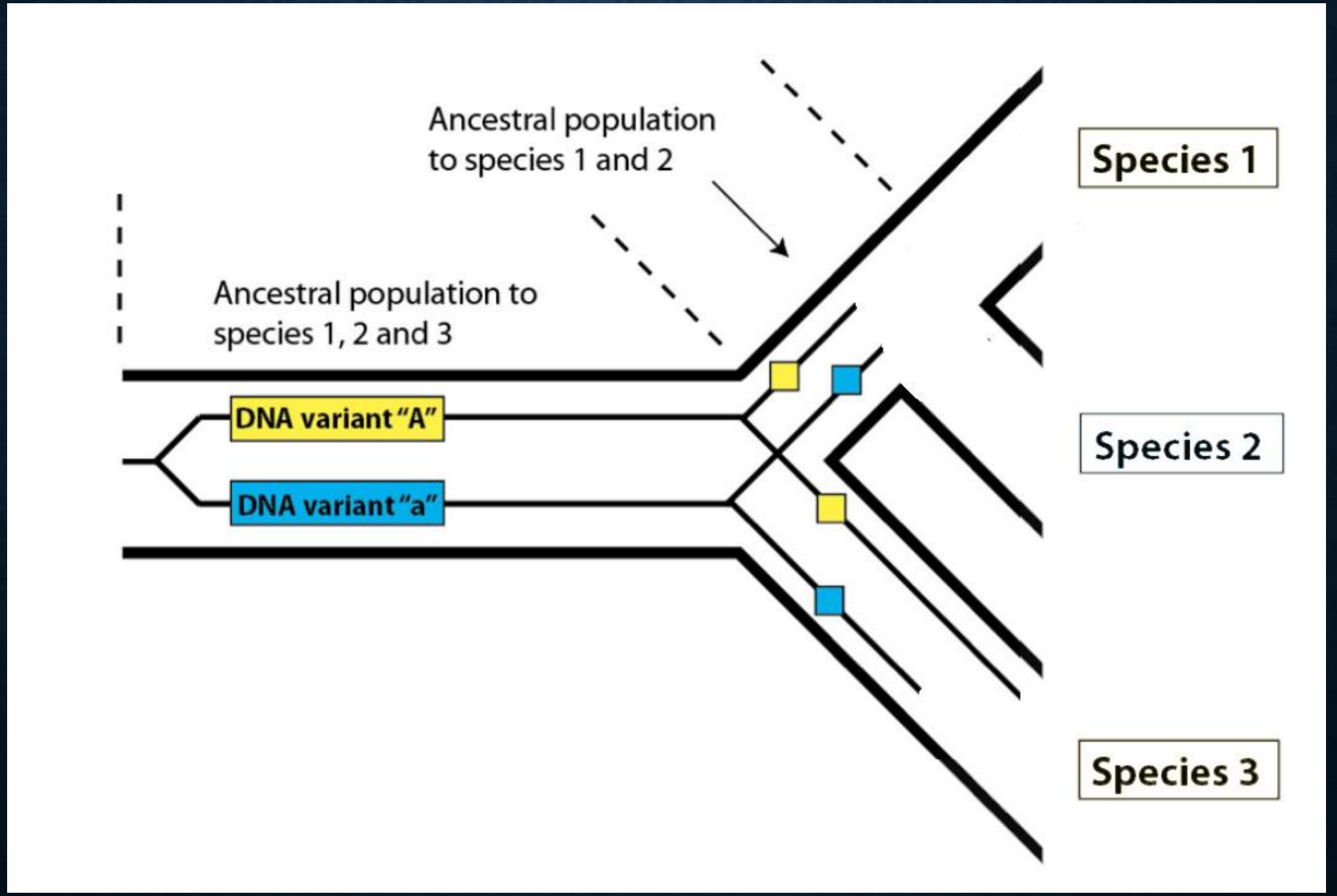
http://www.onezoom.org/

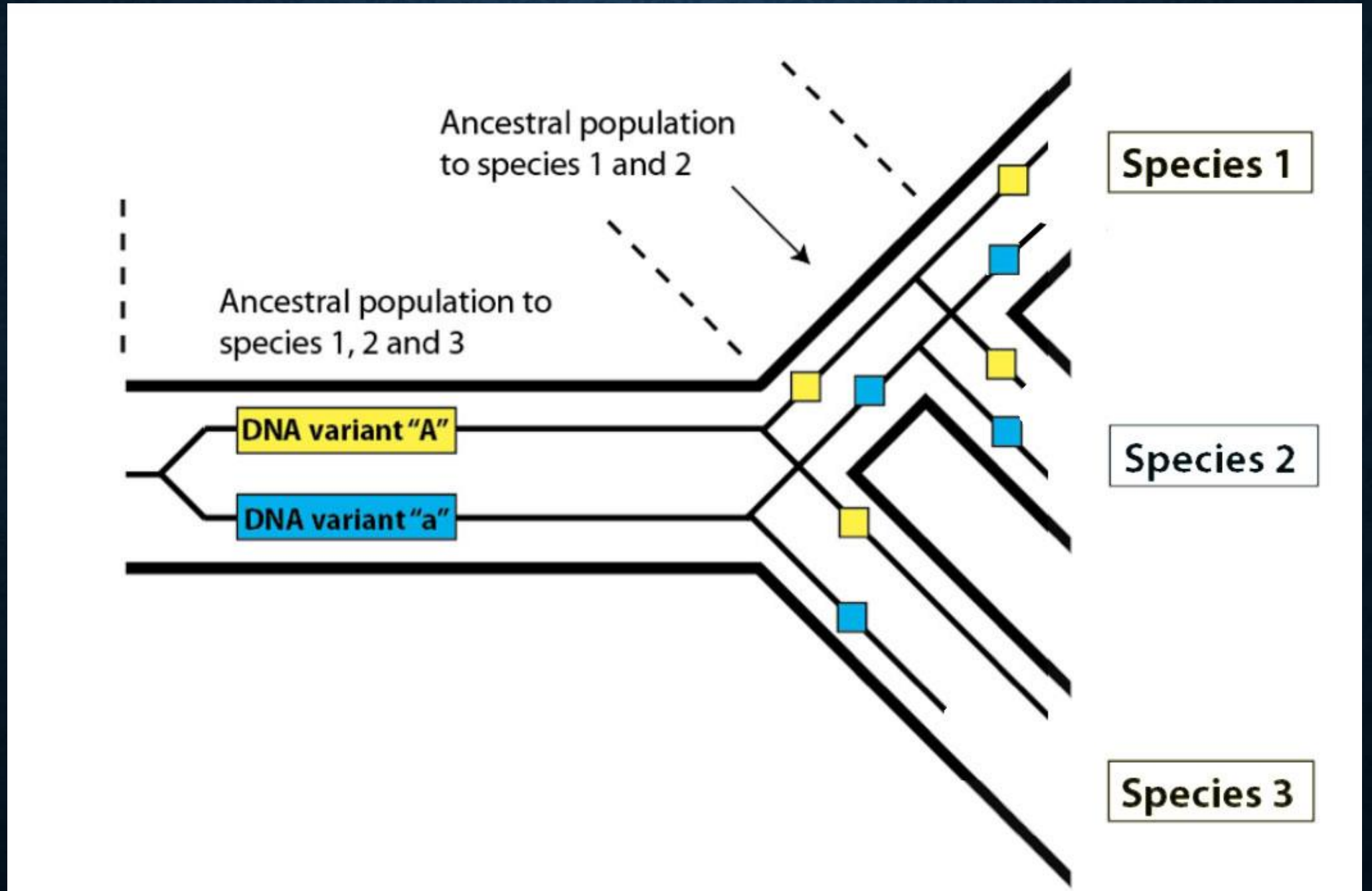
WHY MIGHT SINGLE-GENE TREES DIFFER FROM THE SPECIES TREE?

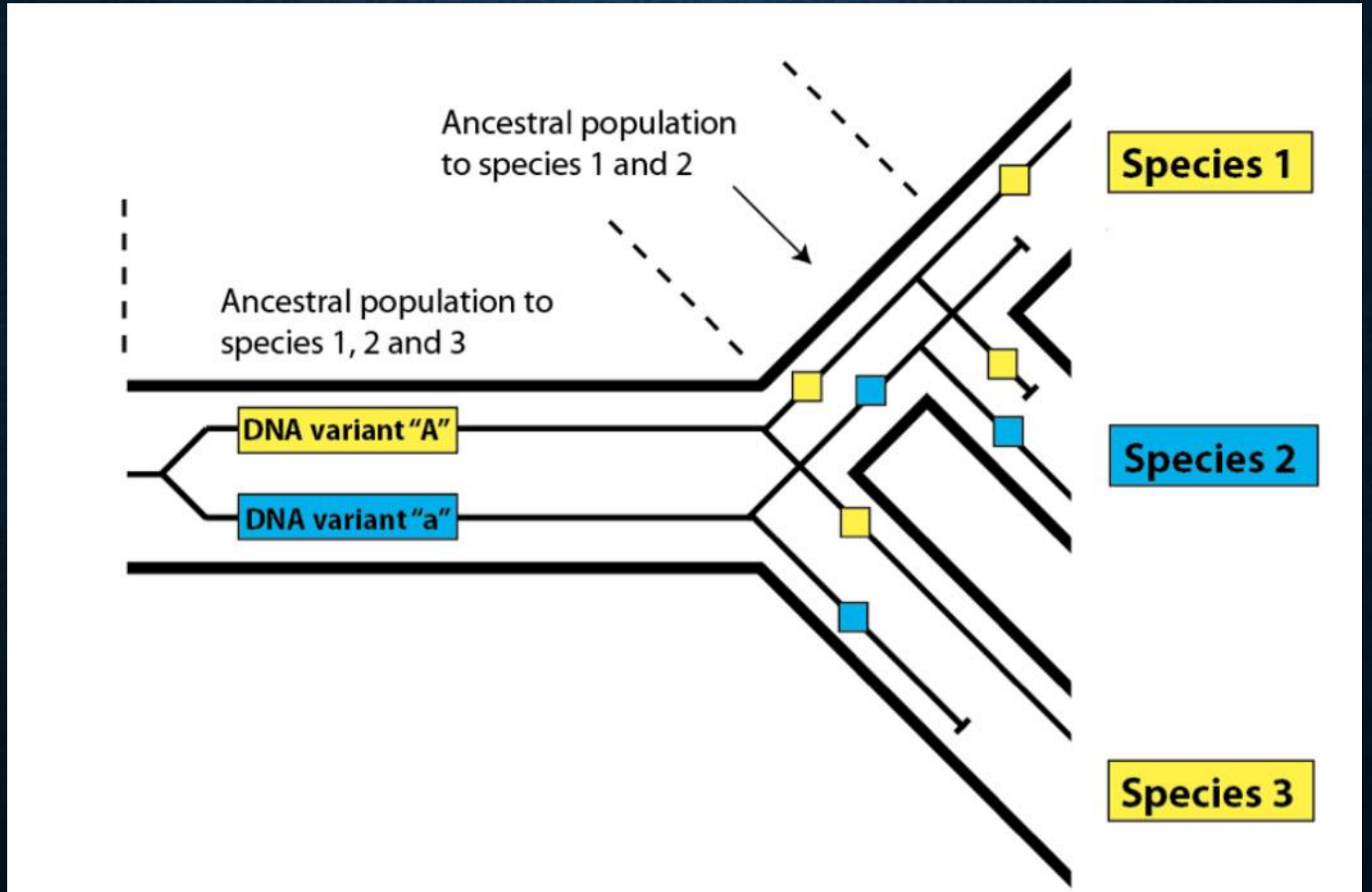
Incomplete Lineage Sorting

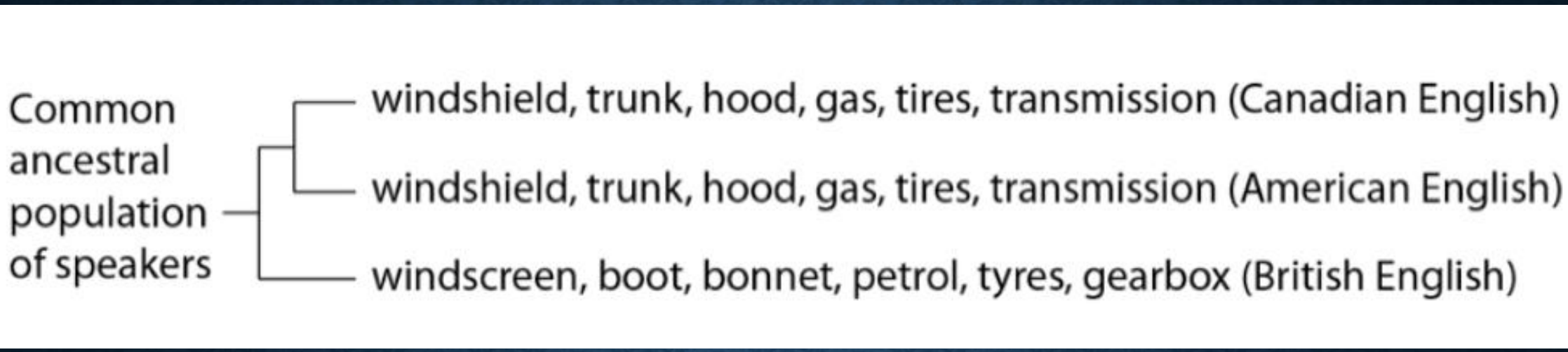


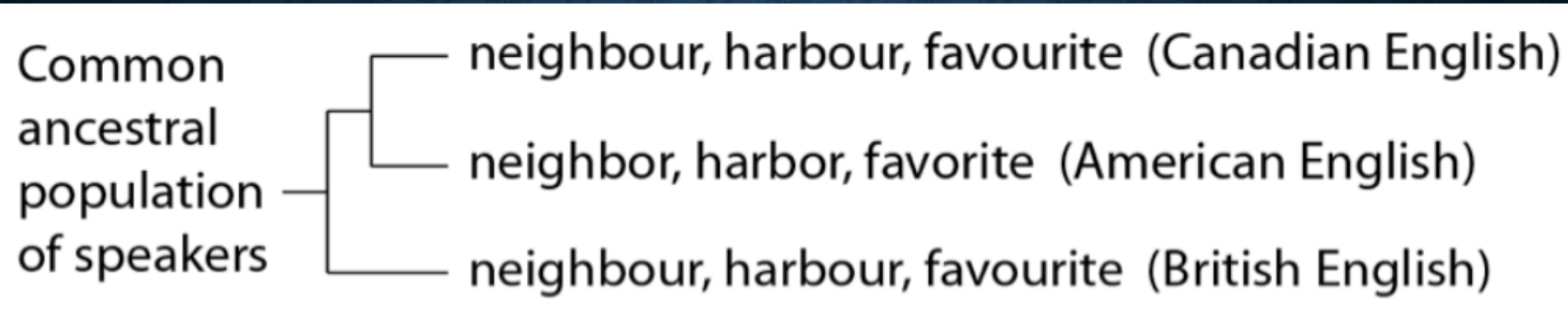
https://biologos.org/articles/series/genetics-and-the-historical-adam-responses-to-popular-arguments/adam-eve-and-human-population-genetics

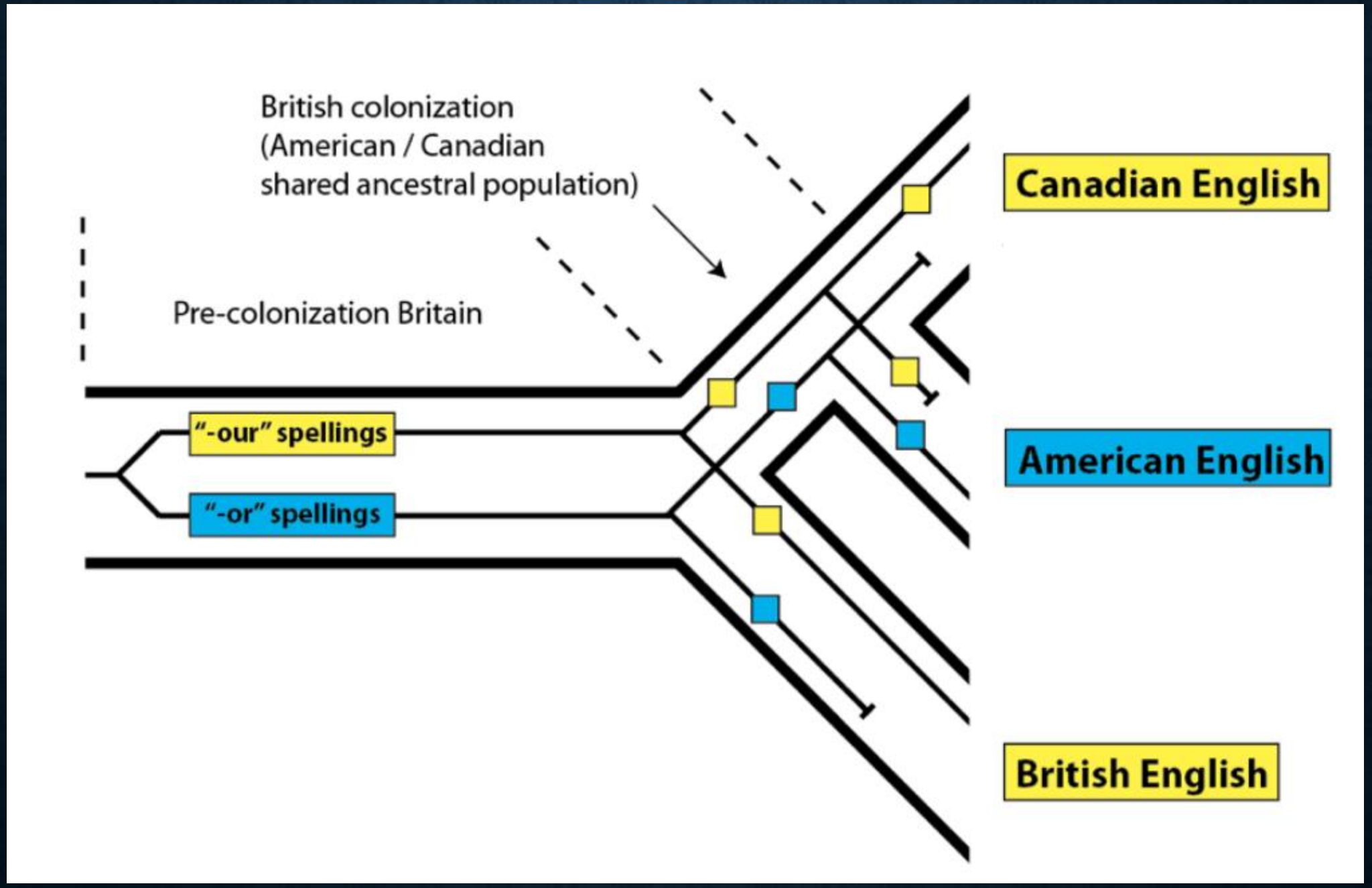

Ancestral population
to species 1 and 2
Ancestral population to
species 1, 2 and 3
DNA variant "A"
DNA variant "a"
Species 1
Species 2
Species 3


Ancestral population
to species 1 and 2
Ancestral population to
species 1, 2 and 3
DNA variant "A"
DNA variant "a"
Species 1
Species 2
Species 3


Ancestral population
to species 1 and 2
Ancestral population to
species 1, 2 and 3
DNA variant "A"
DNA variant "a"
Species 1
Species 2
Species 3


Common ancestral population of speakers
windshield, trunk, hood, gas, tires, transmission (Canadian English)
windshield, trunk, hood, gas, tires, transmission (American English)
windscreen, boot, bonnet, petrol, tyres, gearbox (British English)


Common ancestral population of speakers
neighbour, harbour, favourite (Canadian English)
neighbor, harbor, favorite (American English)
neighbour, harbour, favourite (British English)


British colonization
(American / Canadian
shared ancestral population)
Pre-colonization Britain
"-our" spellings
"-or" spellings
Canadian English
American English
British English