# Genetic analyses for taxonomy

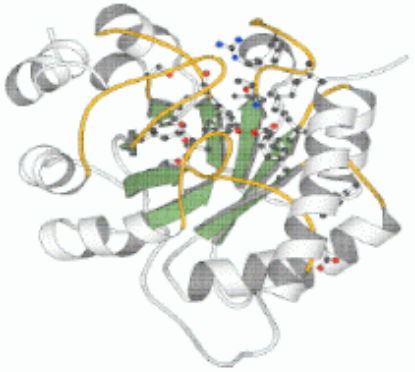
1. Some background details
2. Current and future sequencing
3. Collecting and storing genetic resources

4 Why do phylogenetic trees sometimes disagree with other datasets?

# 1. Some background details

DNA is transcribed into mRNA, which is translated into amino acids

# Central Dogma: DNA -> RNA -> Protein



DNA

transcription

RNA

translation

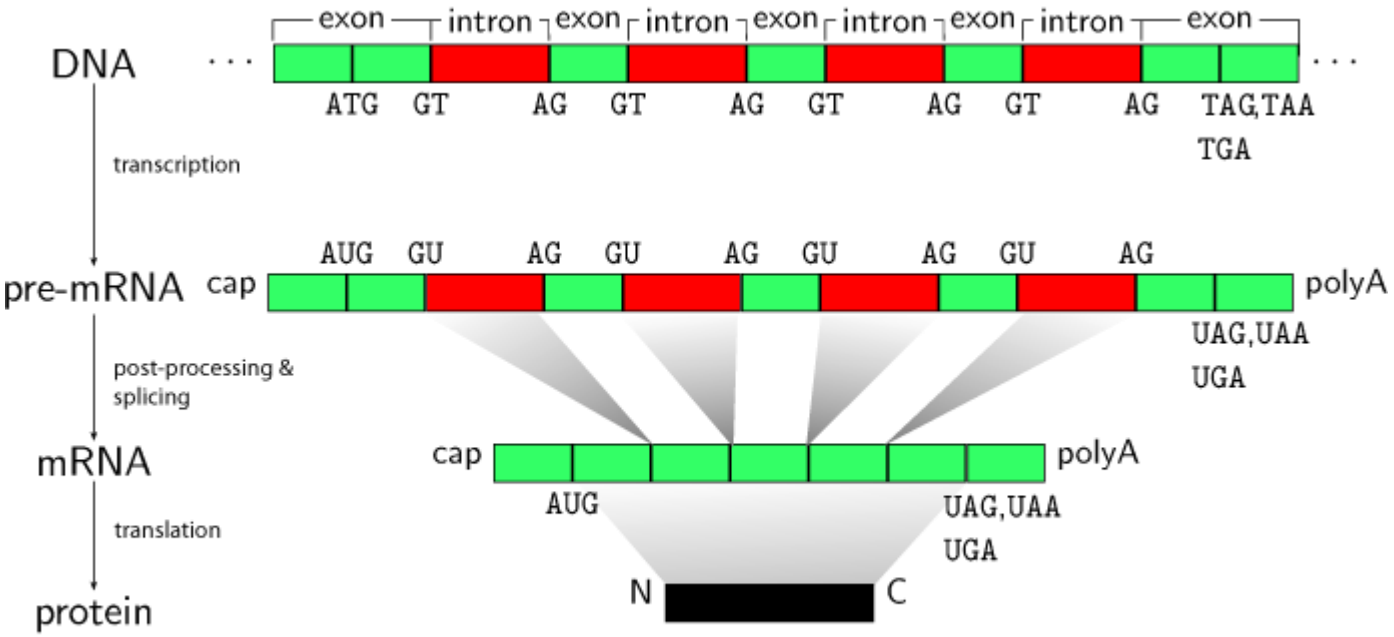
Protein

CCTGAGCCAACTATTGATGAA

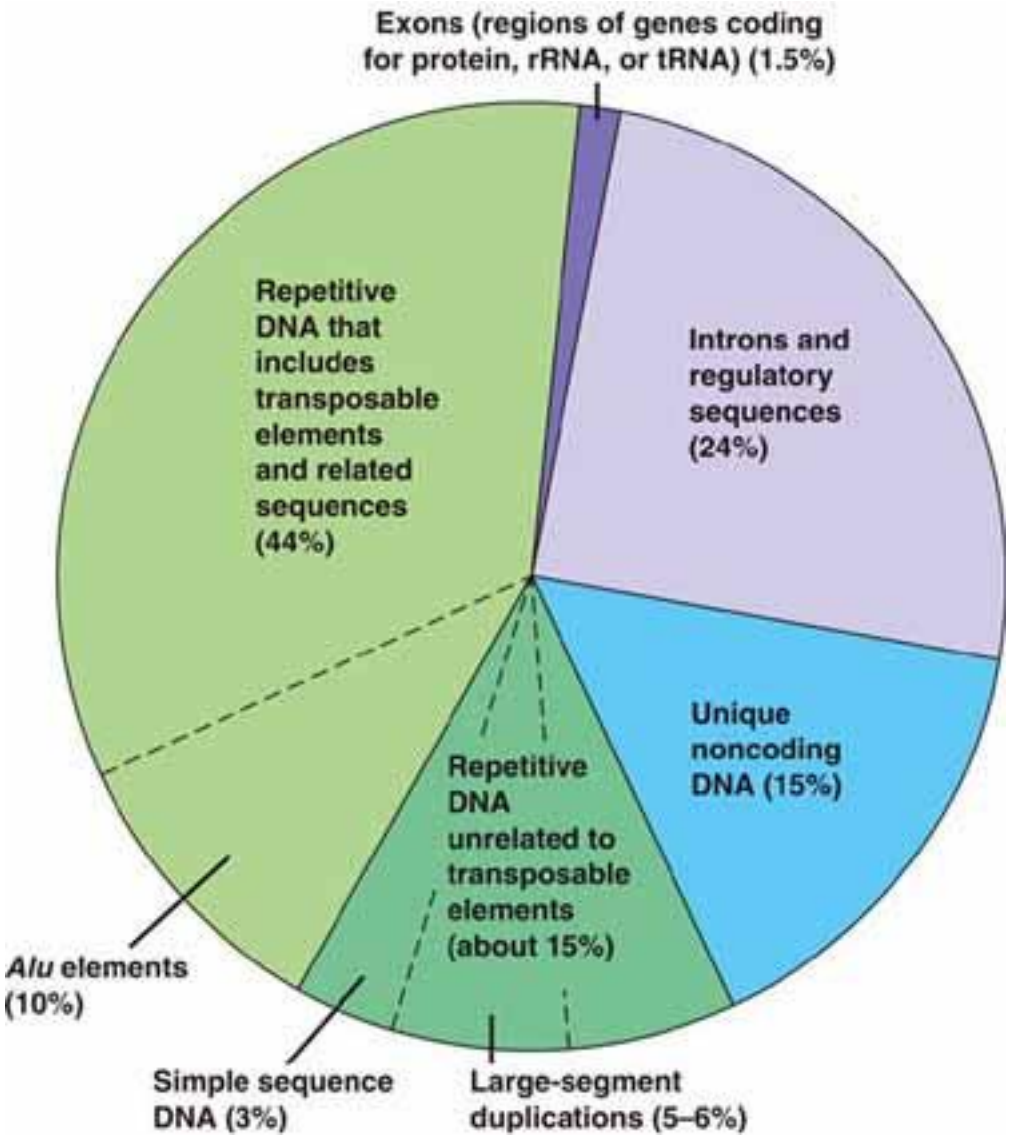
CCUGAGCCAACUAUUGAUGAA

PEPTIDE

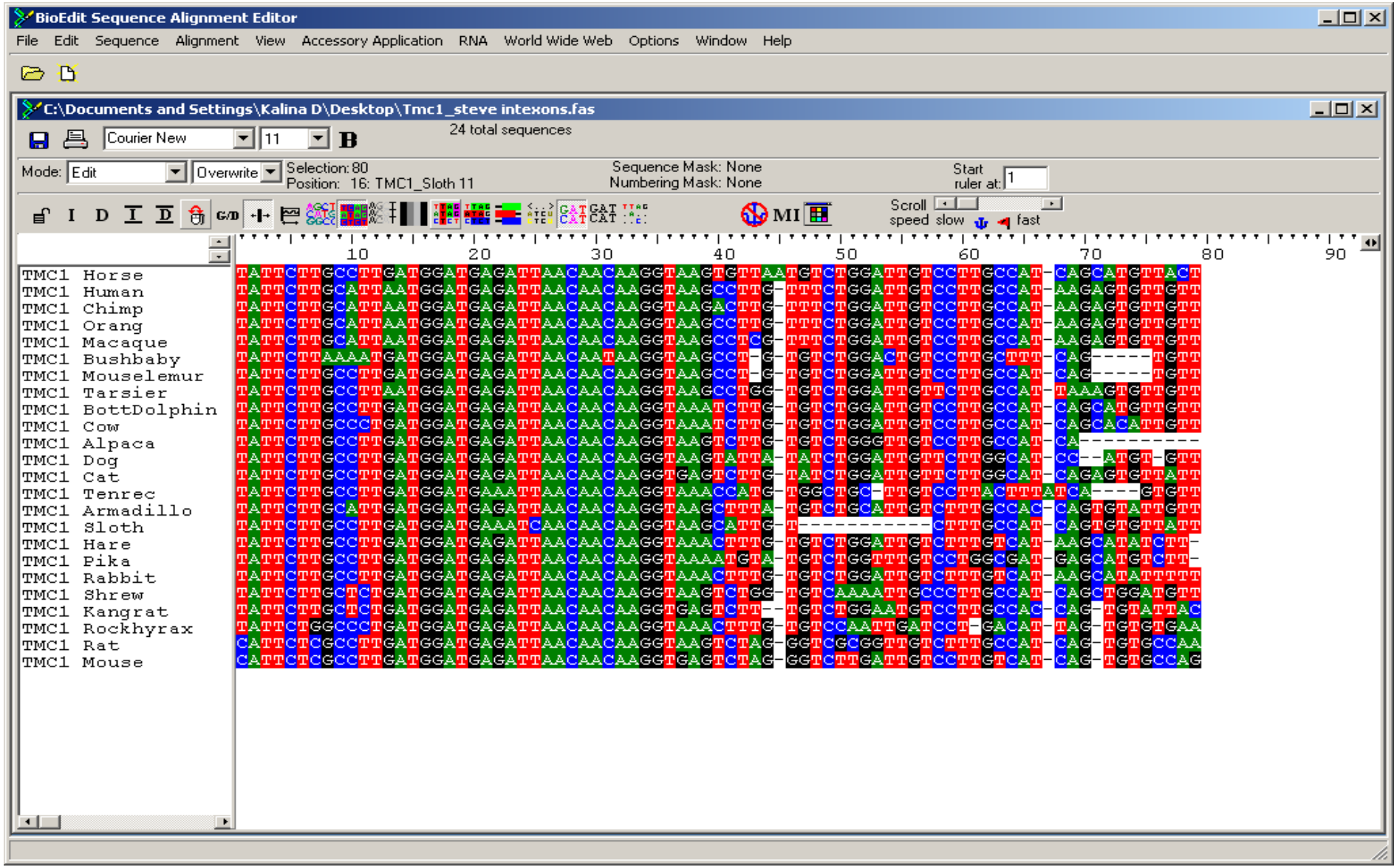
Genes are made up of introns and exons. Introns can be very long and are removed by splicing in gene expression.
This leads to concatenated exons for each gene (equivalent to the coding sequence or “CDS”)

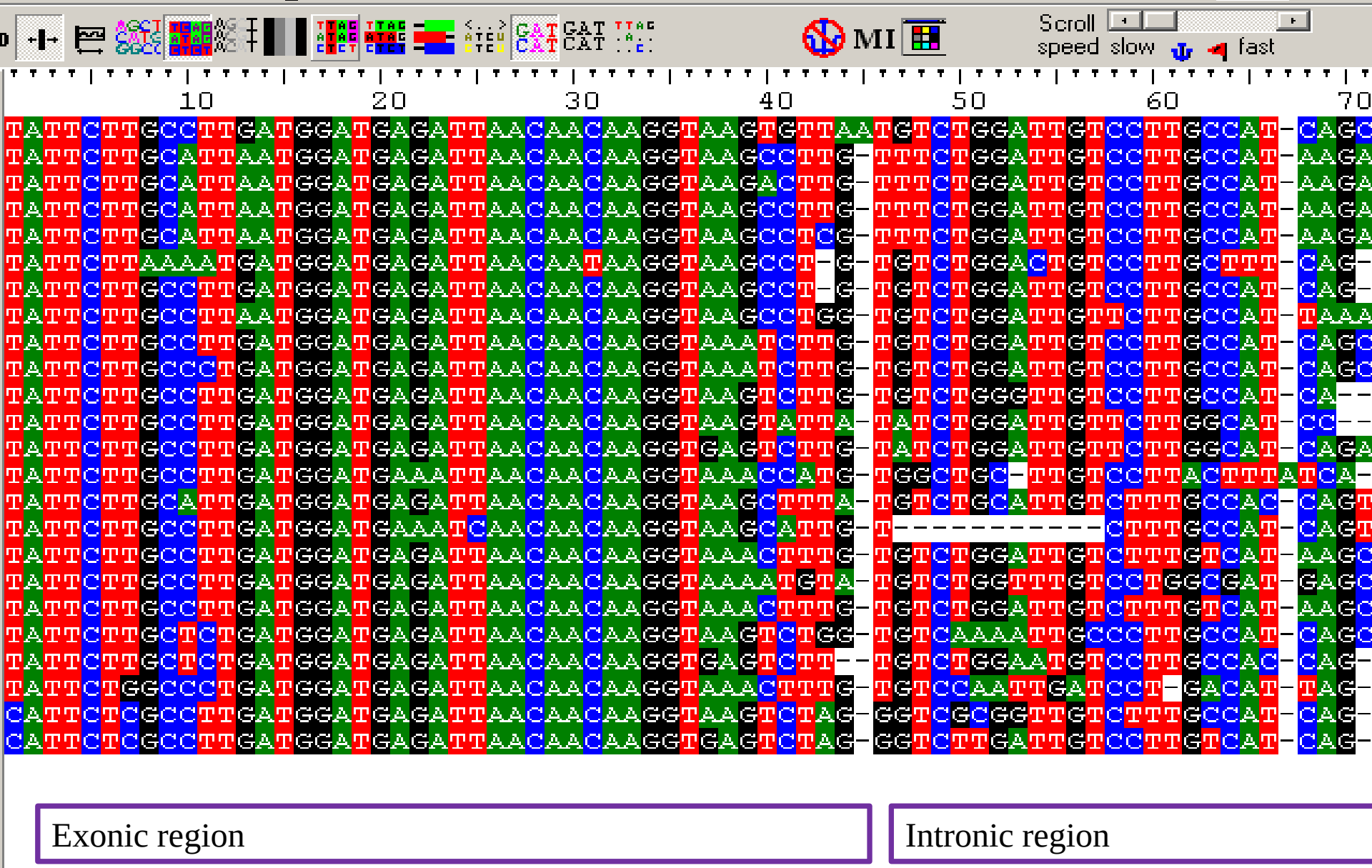


As well as introns within genes, there are large gaps of non-coding sequence between genes. In fact, only around 1% of the mammalian genome is made up of coding DNA)



Exons (coding parts of the gene) tend to be relatively conserved across taxa. Introns are more variable. Below we can see part of an exon and an intron.

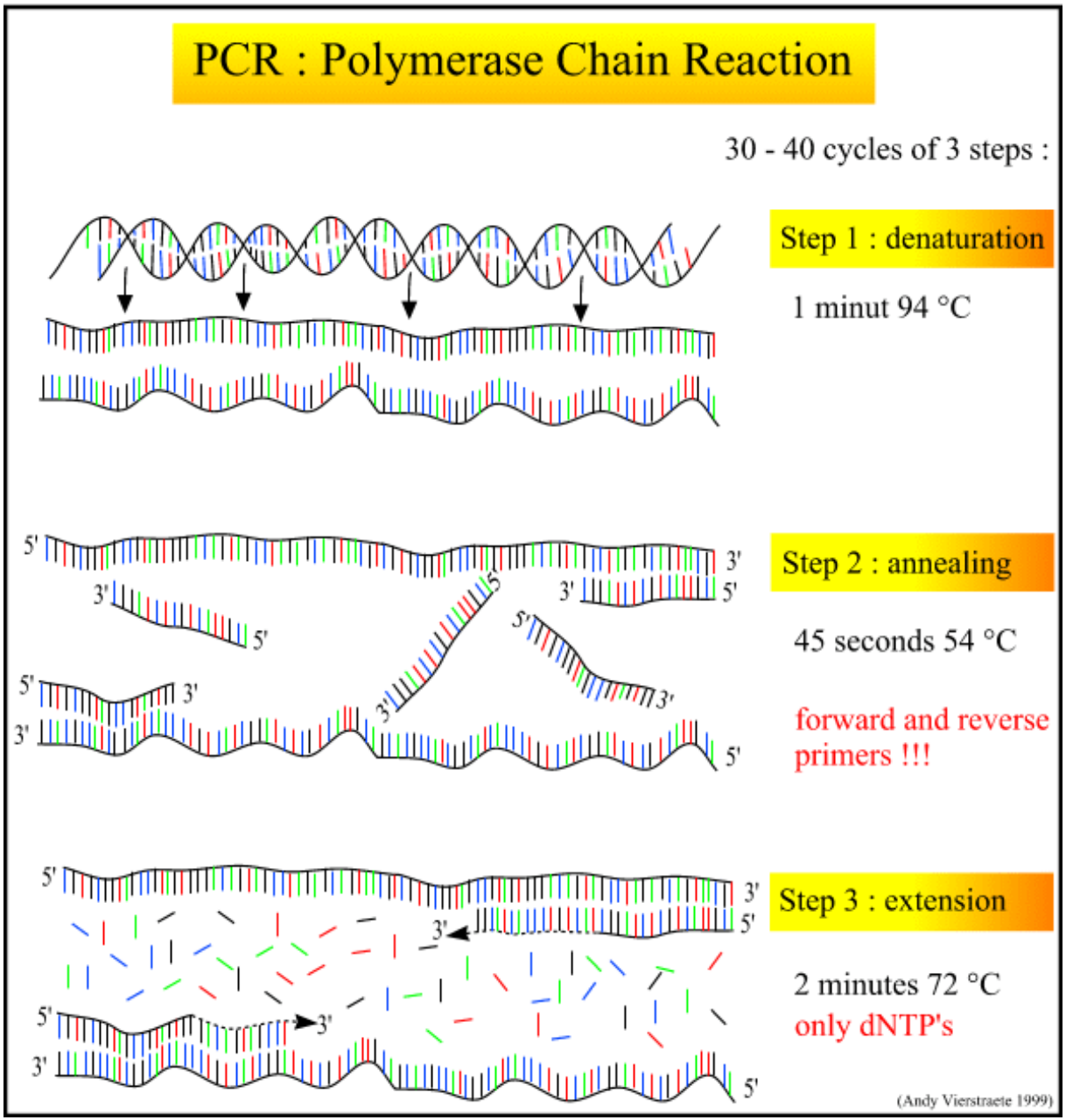





Exonic region

Intronic region

We use PCR to amplify DNA



# 2. Current and future sequencing

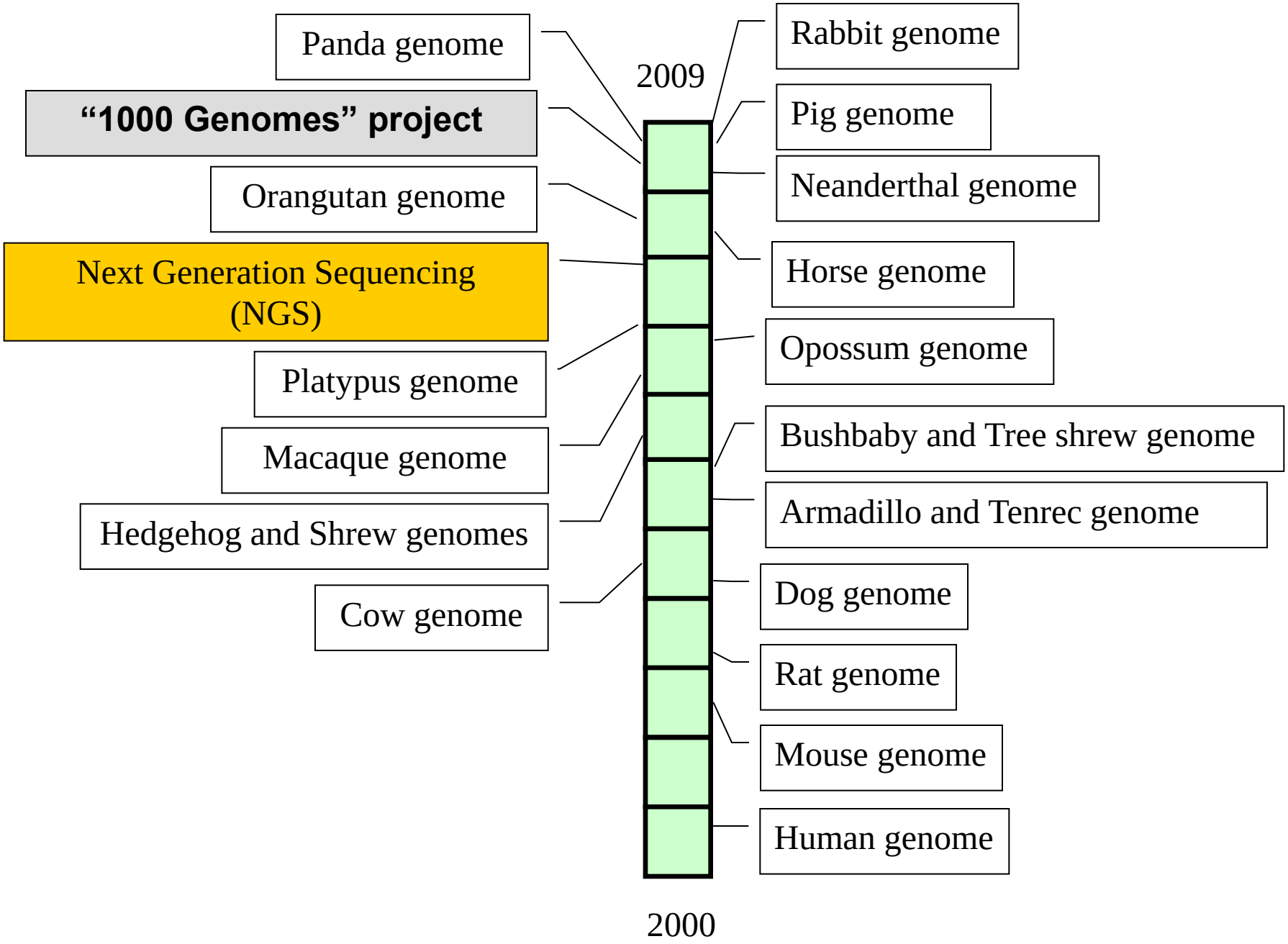
Until 2007 we all used Sanger sequencing.

Several large genome projects were conducted at huge expense

In 2007 several companies released technologies, termed “Next Generation Sequencing”

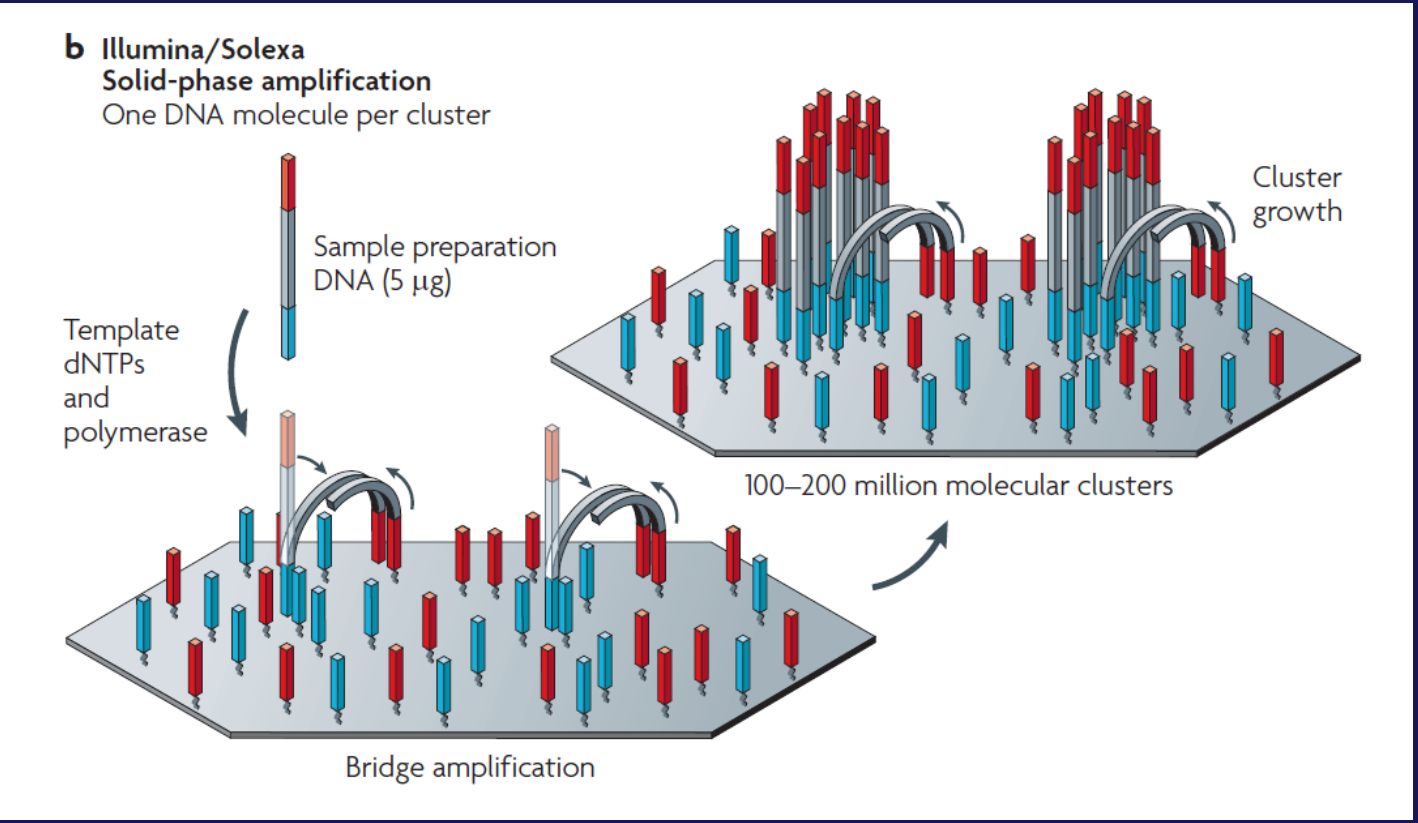
(shotgun sequencing of small reads)

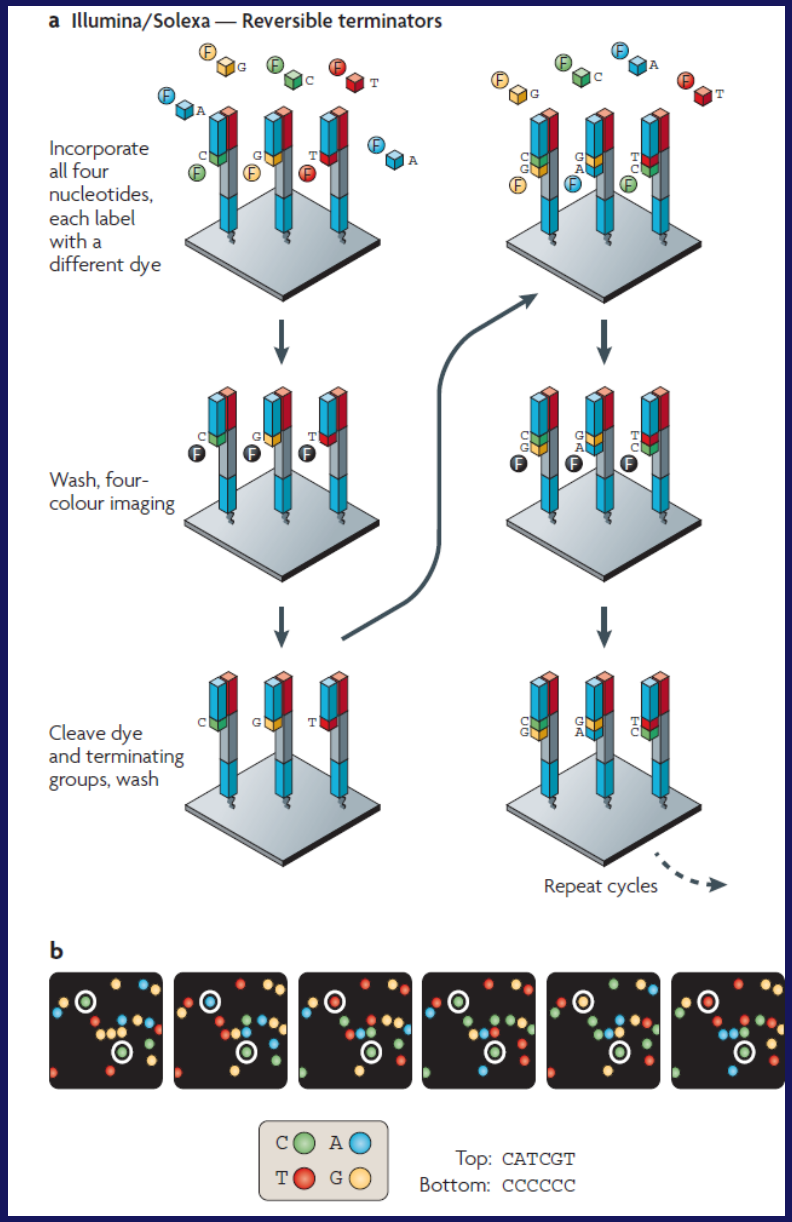
This means more and more genomes are now being produced


Panda genome
“1000 Genomes” project
Orangutan genome
Next Generation Sequencing (NGS)
Platypus genome
Macaque genome
Hedgehog and Shrew genomes
Cow genome
2009
2000
Rabbit genome
Pig genome
Neanderthal genome
Horse genome
Opossum genome
Bushbaby and Tree shrew genome
Armadillo and Tenrec genome
Dog genome
Rat genome
Mouse genome
Human genome

# Next generation sequencing

- Based on shotgun sequencing
- Adaptors containing universal priming sites are ligated to ends of the DNA fragment
- DNA templates amplified clonally to get clusters






- Fluoro-labelled dNTPS washed over the strand
- Each time a photo is taken

# NGS by Illumina

2 years

Up to 6.5 Gb per day

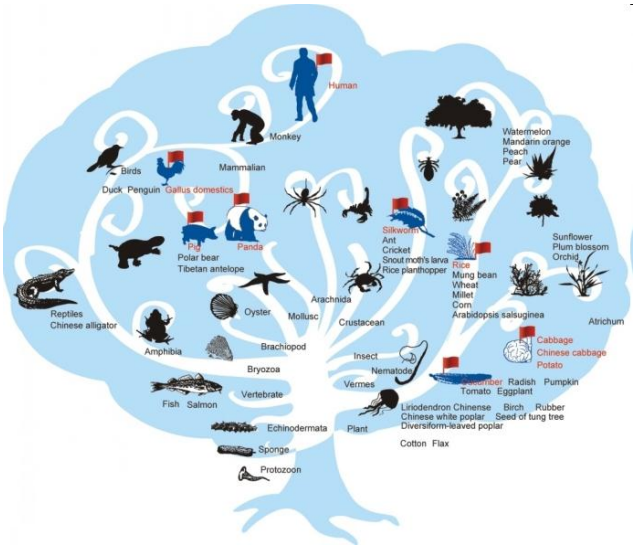
640 million paired-end reads

Up to 25 Gb per day

2 billion paired-end reads

~60x coverage of a human genome in a single run for under $10,000

2019




**BGI launches "1000 Genome" initiative**

Tasmanian devil genome

Gorilla and Gibbon genome

Myotis genome

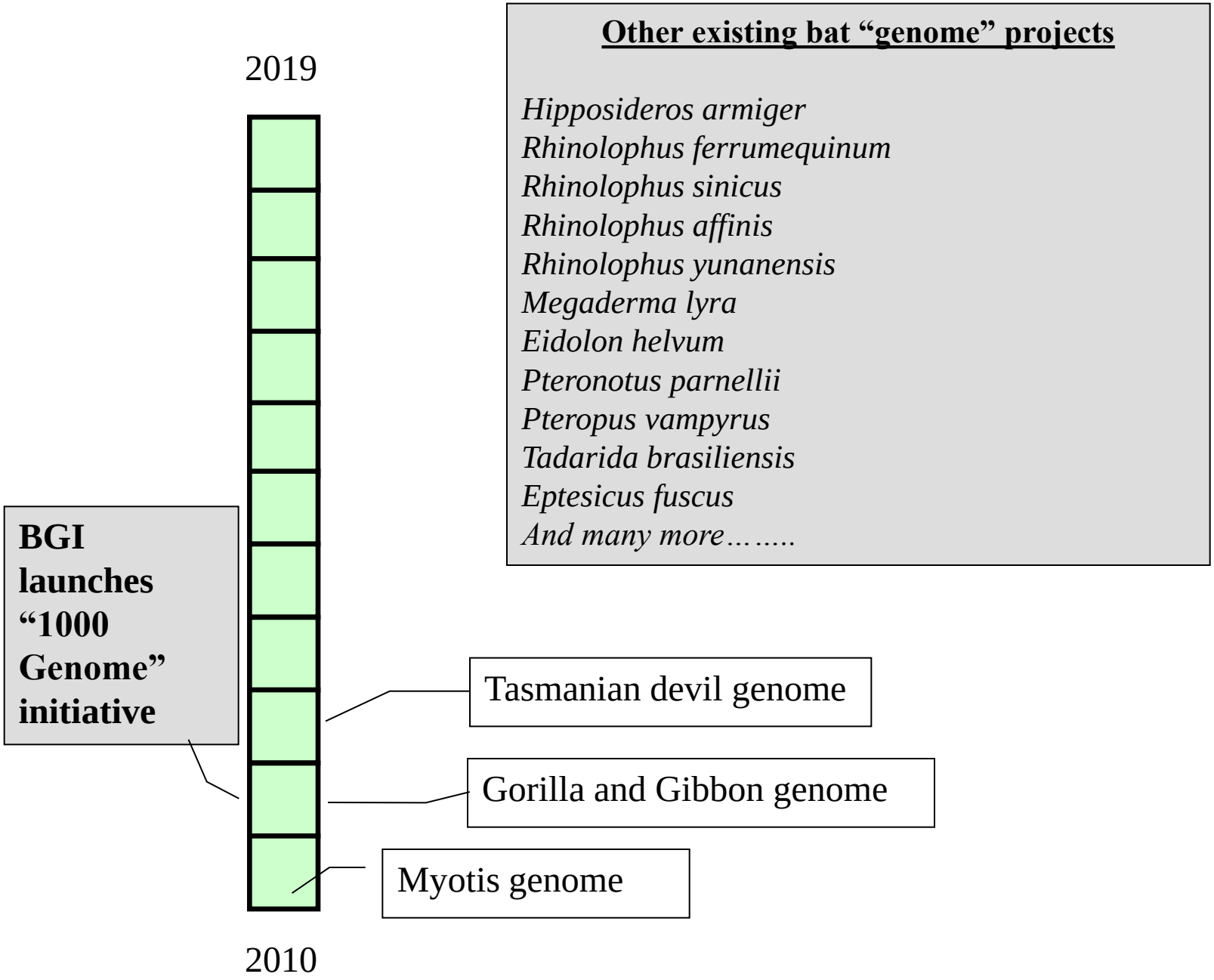
2010

**In Process**

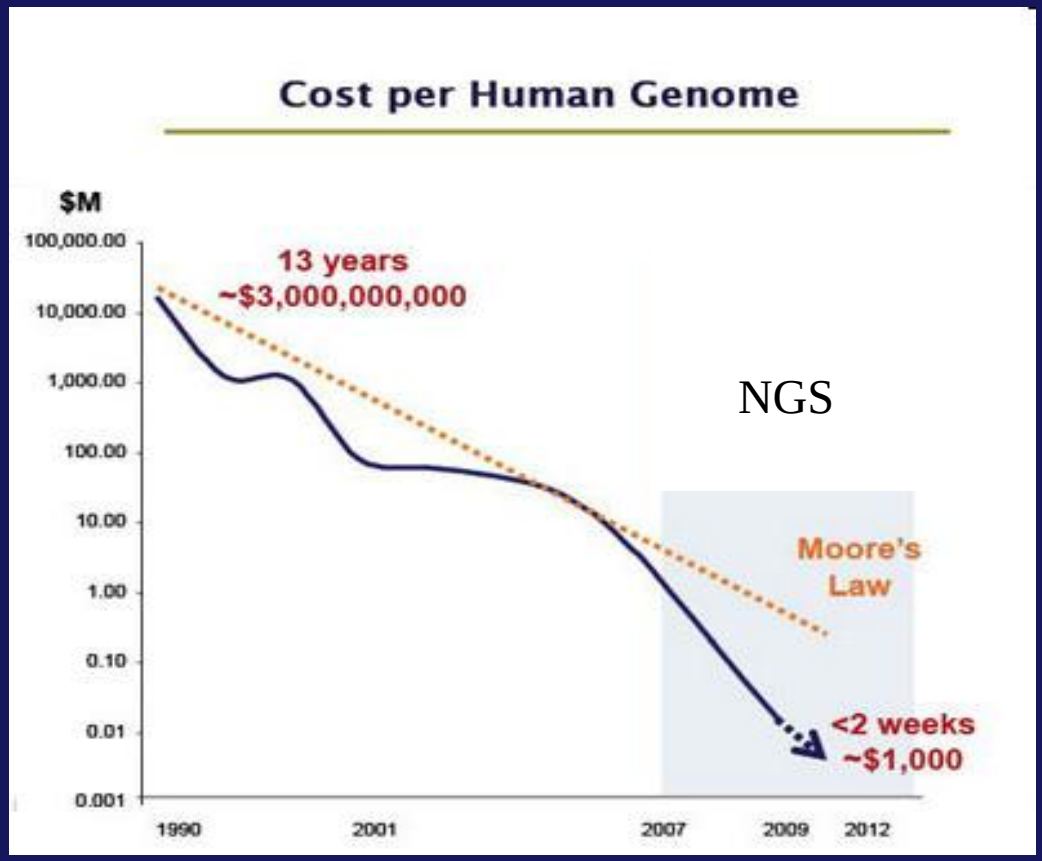
Tibetan antelope
Polar bear
Camel
Puma

**Next**

Hyrax
Potto
Wombat
Chinese dolphin
Donkey
Porpoise
Asian lion
Beluga whale
Giraffe
Aardwolf
Whale
Mole-rat
Hamster


2019
Other existing bat "genome" projects
Hipposideros armiger
Rhinolophus ferrumequinum
Rhinolophus sinicus
Rhinolophus affinis
Rhinolophus yunanensis
Megaderma lyra
Eidolon helvum
Pteronotus parnellii
Pteropus vampyrus
Tadarida brasiliensis
Eptesicus fuscus
And many more……
BGI launches "1000 Genome" initiative
Tasmanian devil genome
Gorilla and Gibbon genome
Myotis genome
2010

# Affordability?




- Gene sequencers outpace microchips
- Numerous companies promise genomes for 1K USD within just 2-5 years
- This means these methods will be within our financial reach.
- Some predict a genome will be less than 100 dollars in a few years.

# Within one or two years from now

- Single molecule approaches (this could mean our museum samples are useful for genome sequencing)

- Ultra portable sequencers (could be useful for field work)

Oxford Nanopore's "minION"

# So how will we benefit from these methods?

Development of new DNA markers

Examples of using genome comparative data

Future of phylogenomics and population genomics

# Development of new DNA markers I

- Microsatellite discovery by mining published genome data

Shikano *et al. (*2010) *BMC Genomics* **11**: 334
Sequenced genomes for microsatellite marker development in nine-spined sticklebacks

- Microsatellite development by low coverage genome sequencing

Abdelkrim (2009) *BioTechniques* 46: 185-192
blue duck DNA → 454 sequencing → 17215 reads → >200 loci → 24 primer sets

**Table 2. Summary Results of the Development of Microsatellite Markers Following the Genomic Approach on Three Other Species**

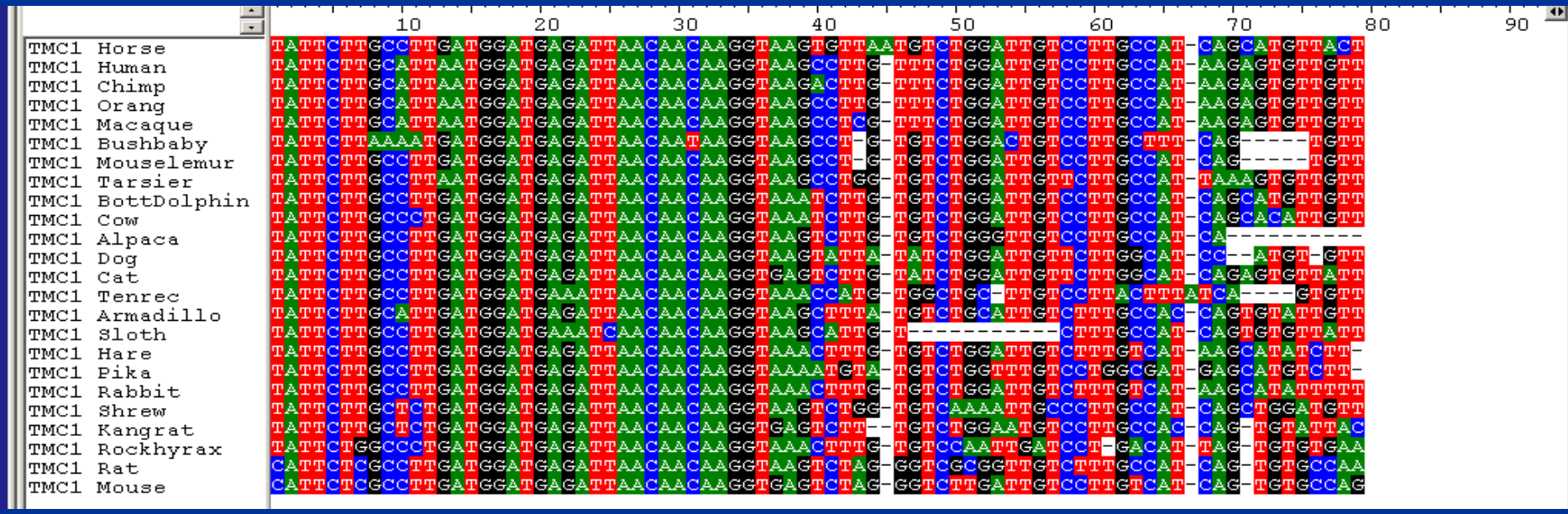
| Species | Class | Number of reads | Minimum number of repeats* | Number of microsatellites detected | Number of potential primer pairs |
|---|---|---|---|---|---|
| *Maritrema novaezealandensis*[1] | Trematoda | 31120 | 5,4,4 | 676 | 46 |
| *Motuweta isolata*[2] | Insecta | 52059 | 4,4,4 | 472 | 134 |
| *Powelliphanta augusta*[2] | Gastropoda | 35196 | 6,6,6 | 2541 | 170 |

*For di-, tri-, and tetranucleotide, respectively
[1]Unpublished data provided by Yuri Springer (University of Otago)
[2]Unpublished data provided by Thomas Buckley (Landcare Research)

# Development of new DNA markers II

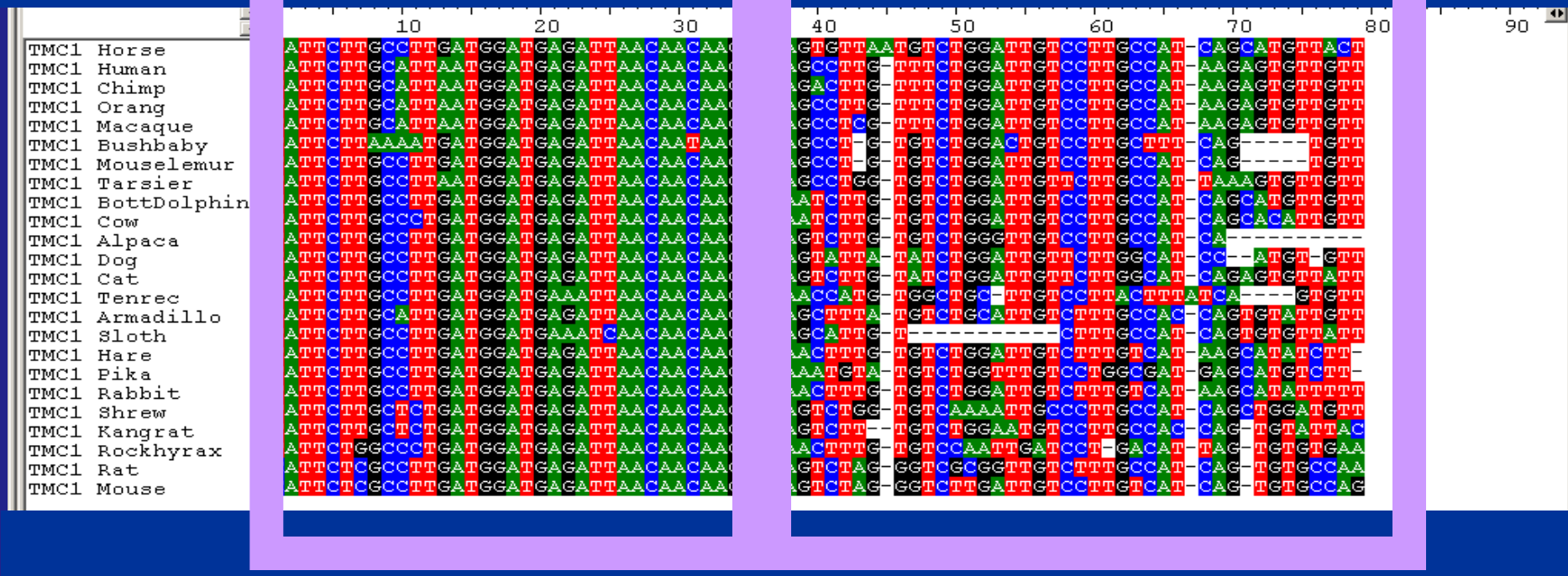
- Non-coding and coding DNA from mining genome data

"Transmembrane channel-like protein 1" gene



- Non-coding and coding DNA from mining genome data

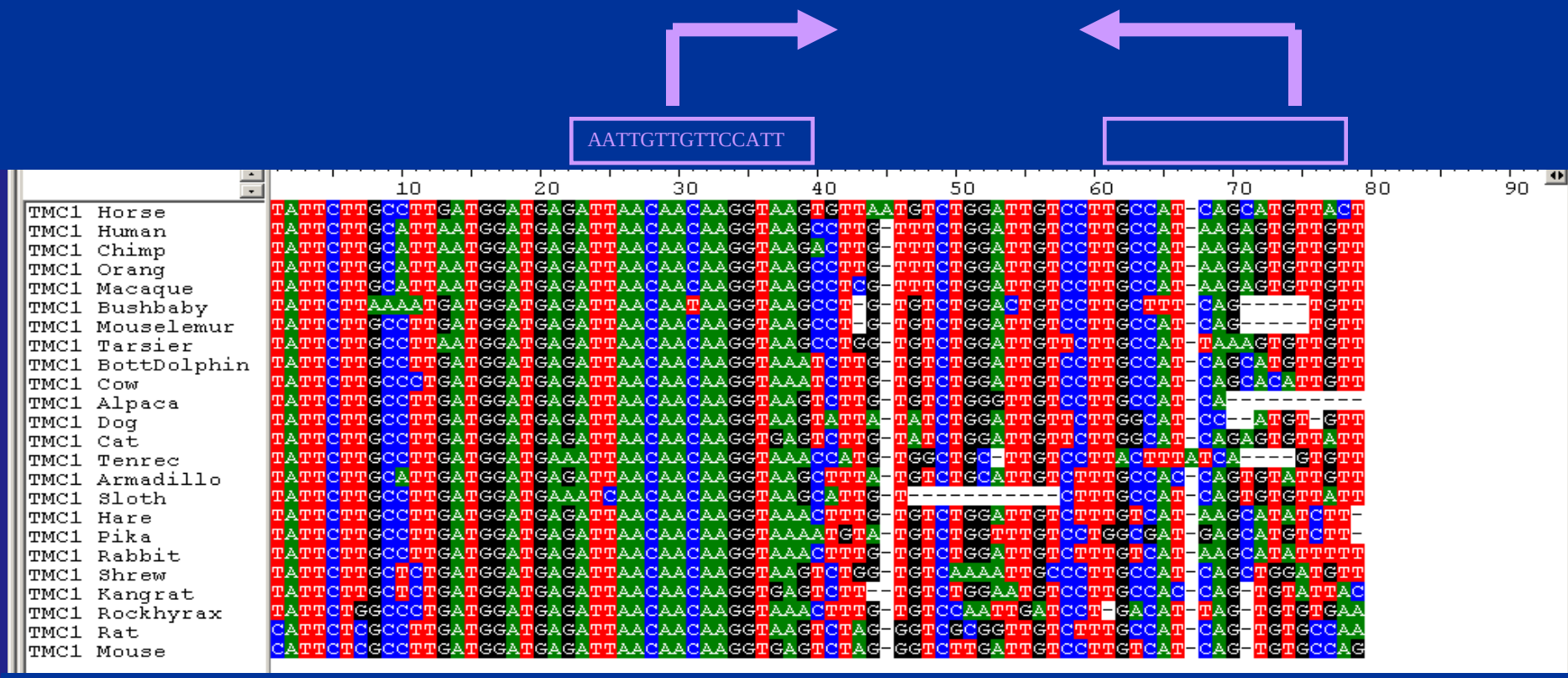
“Transmembrane channel-like protein 1” gene



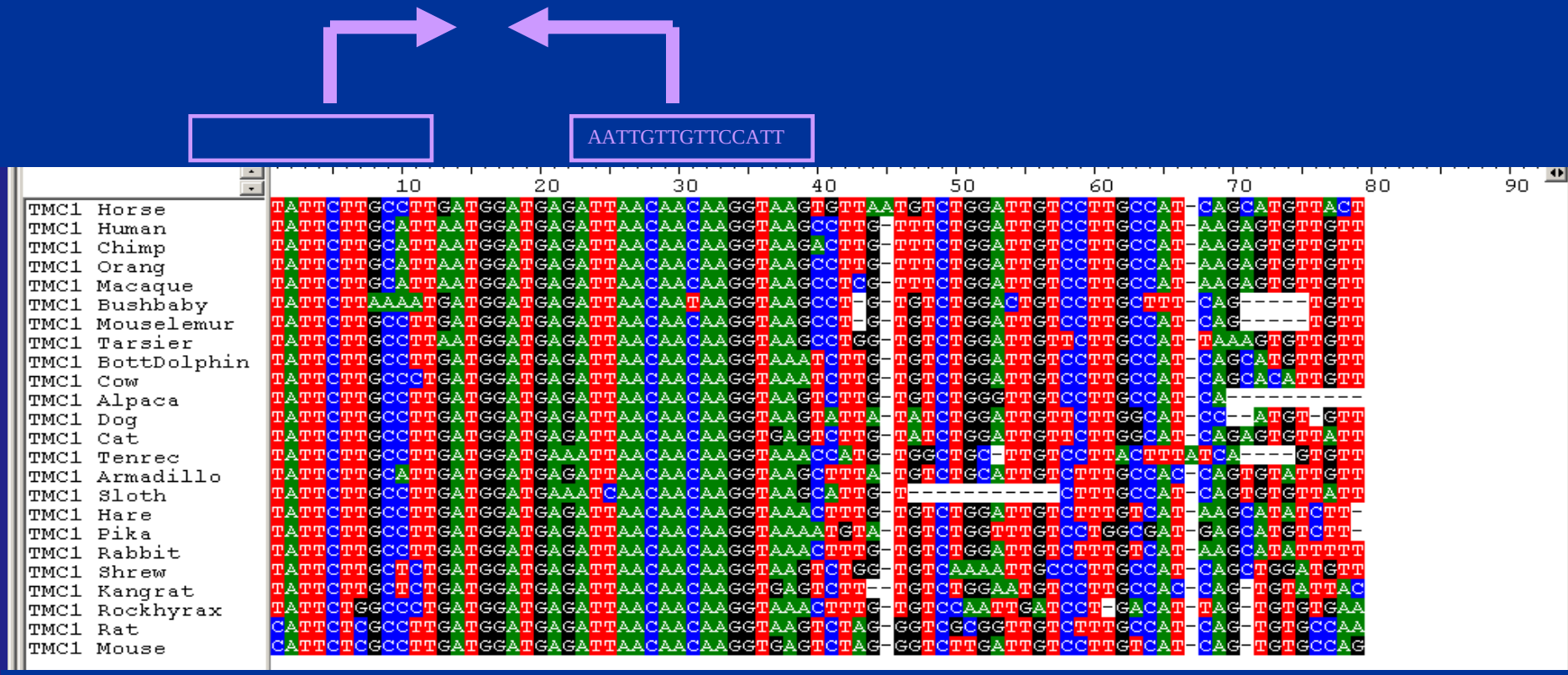

Intron

Exon

“Exon-primed intron-crossing sequences” (EPICs)



# Nuclear protein coding loci (NPCL)



We need to be prepared for these new methods and technologies.

It is important to think about how we maximise the potential benefit of our samples, particularly for our ongoing collecting and research

# Collecting and storing genetic resources

The best approaches to collecting and storing bat material for genetic analysis will depend on the needs of the samples

Very often we cannot predict the future technologies so it is important to take care now to safeguard the value of our material in the future

**Most genetic analyses are based on DNA**
**Uses of DNA work include:**

Sequencing for phylogenetic and phylogeographic analyses

Species identification via bar coding (COI) and other loci

Species ID etc

Microsatellite genotyping

Functional genes

**Advantages of DNA**

Relatively stable

Evenly distributed across all cells
(same result from muscle versus wing versus liver)

Advantages of using introns as a source of variation

**Disadvantages of DNA**

Introns and inter-genic areas can also make primer design difficult

Exonic within genes areas might be far apart from each other

Lots of Intergenic sequence

**Work on RNA is becoming more important**

RNA can be studied to determine expression in different tissue types

No introns or Intergenic regions, so get more gene sequence per dollar

**Disadvantages of RNA**

Degrades rapidly

Need more material to get enough

Need multiple tissue to obtain all genes

For amplification, need to convert to DNA first

**DNA and RNA need to collected and stored differently**

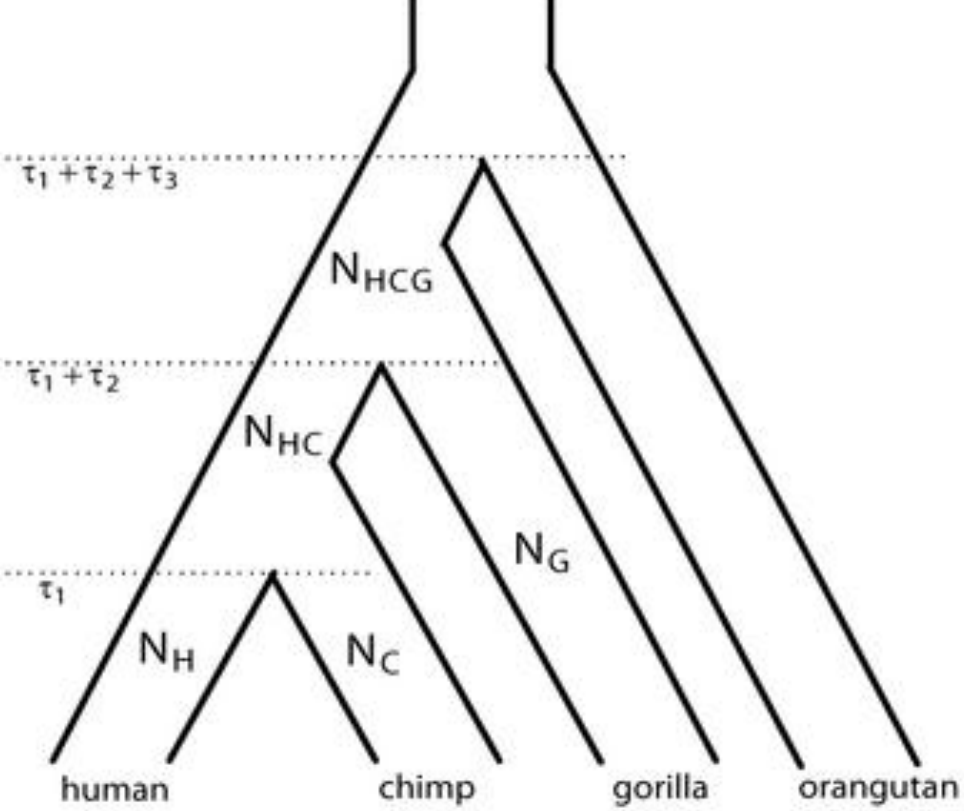
# Tissue preservation methods

| | |
|---|---|
| RNAlater | Liquid nitrogen |
| 100% Ethanol | Dry ice |
| 70% Ethanol | Tissue lysis buffer |
| Formulin | AllProtect |
| IMS | Silica gel |
| VTM | Freezing (-20) |
| DMSO | Freezing (-80) |

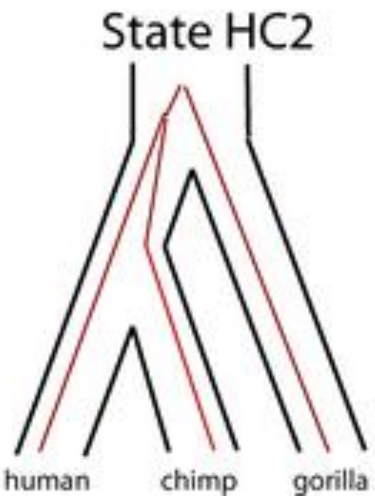
# Why do phylogenetic trees sometimes disagree with other datasets?

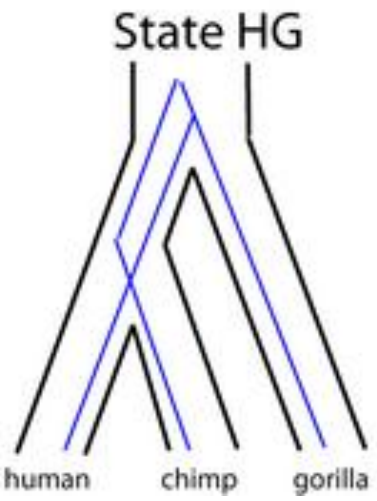
# Why do phylogenetic trees sometimes disagree with other datasets?

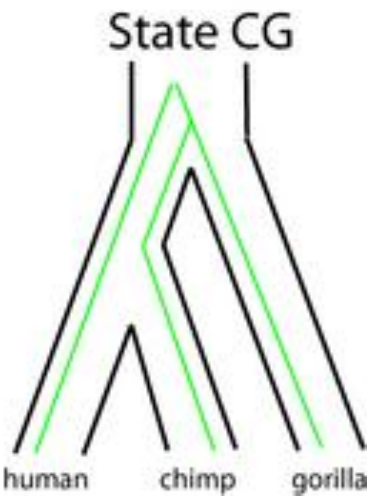
1. Incomplete sorting


$\tau_1+\tau_2+\tau_3$
$N_{HCG}$
$\tau_1+\tau_2$
$N_{HC}$
$N_G$
$\tau_1$
$N_H$
$N_C$
human
chimp
gorilla
orangutan

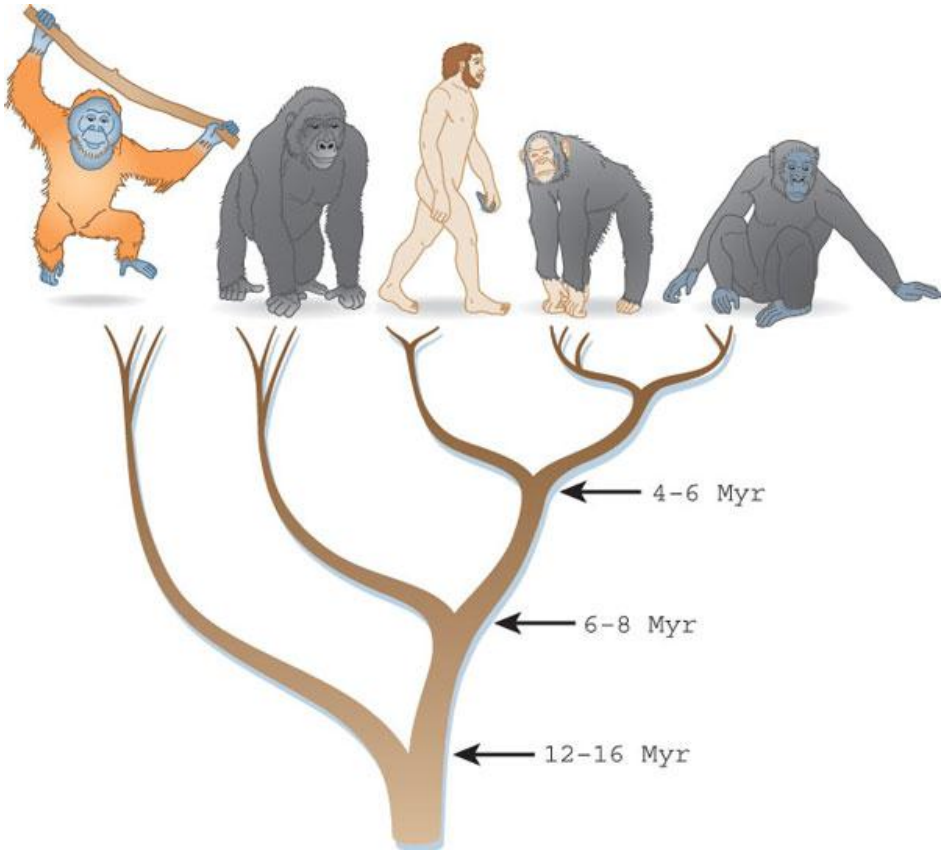

State HC1
human
chimp
gorilla


State HC2
human
chimp
gorilla


State HG
human
chimp
gorilla

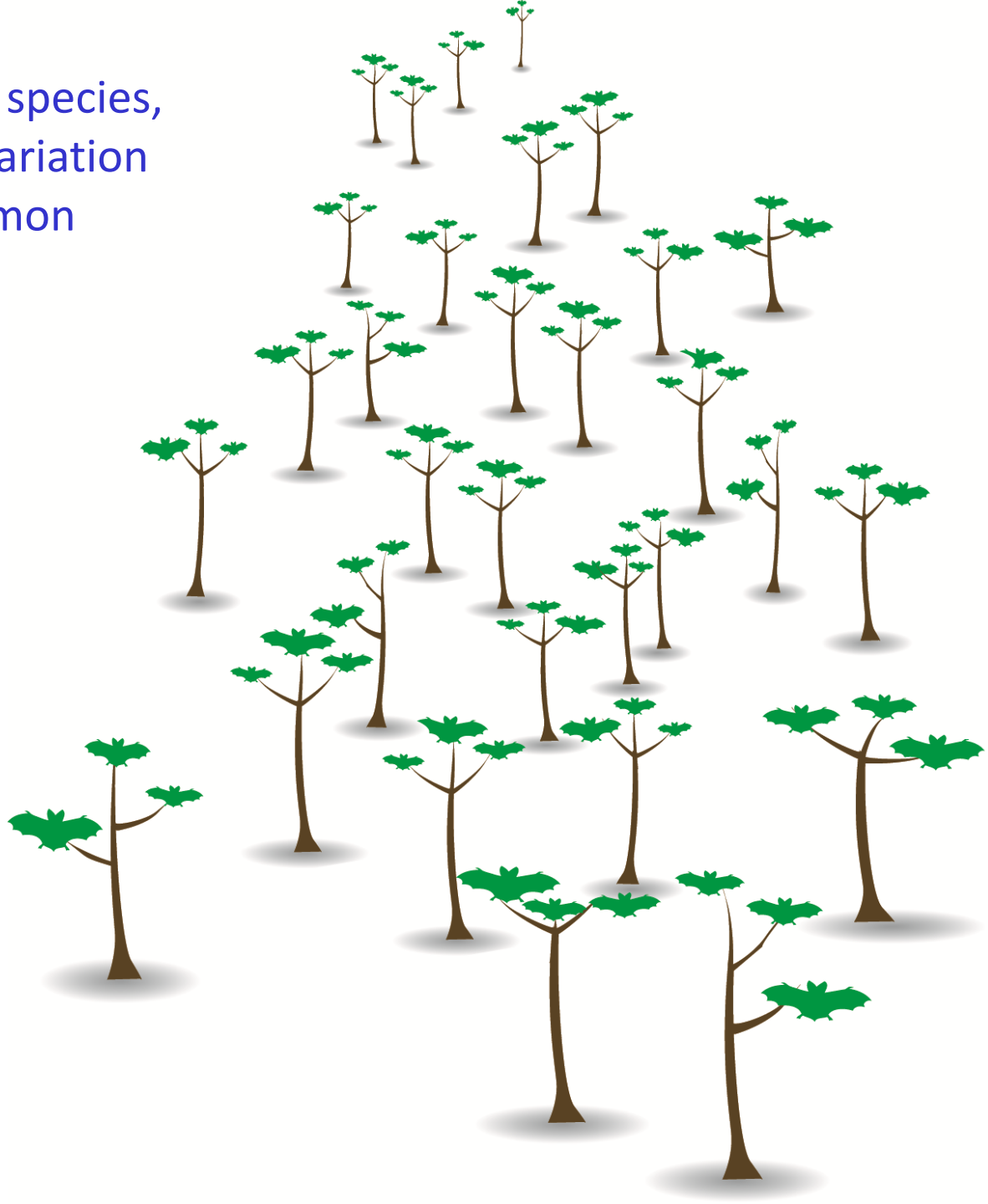

State CG
human
chimp
gorilla

For 1% of genome, humans more closely related to orang utans than to chimps

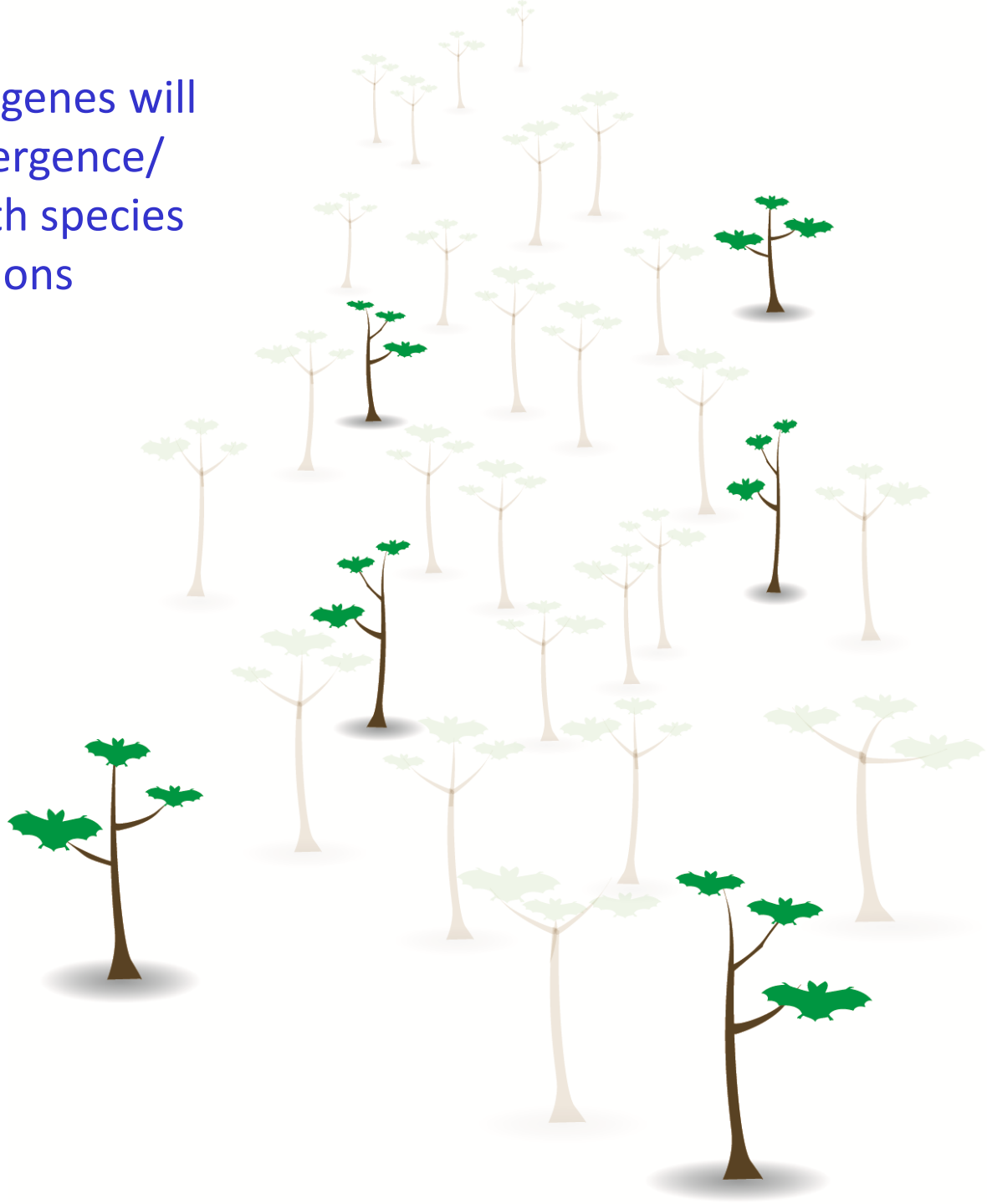



Genome Res. 2011 March; 21(3): 349–356.

For sister species, shared variation common



Only a few genes will show divergence/ sorting with species divisions



# Why do phylogenetic trees sometimes disagree with other datasets?

1. Incomplete sorting

2. Long branch attraction

# Why do phylogenetic trees sometimes disagree with other datasets?

1. Incomplete sorting

2. Long branch attraction

3. Introgression

# Introgression

Movement of genes from one taxon to another following mating

Previous thought to be uncommon in wild mammals

Now known to be widespread in mammals, incl. bats!

Hybridization between black (*Pteropus alecto*) and grey-headed (*P. poliocephalus*).
Webb & Tidemann (1995) Australian Mammalogy, 18, 19-26.

Hybridization in Peters′ tent-making bat (*Uroderma bilobatum*: Phyllostomidae).
Hoffmann et al (2003) Molecular Ecology, 12, 2981-2993.

Berthier et al (2006) Hybridization between *Myotis myotis* and *Myotis blythii*.
Proceedings of the Royal Society B: Biological Science, 273, 3101-3109.

Hulva et al (2010) Hybridisation in the genus *Pipistrellus*.
Molecular Ecology, 19, 5417-5431.

Mao et al (2010) Historical hybridisation in *Rhinolophus pearsoni* and *R yunanensis*. Molecular Ecology, 19, 1352-1366.

Mao et al (2010) Hybridisation in *Rhinolophus affinis* subspecies.
Molecular Ecology, 19, 2754-2769.

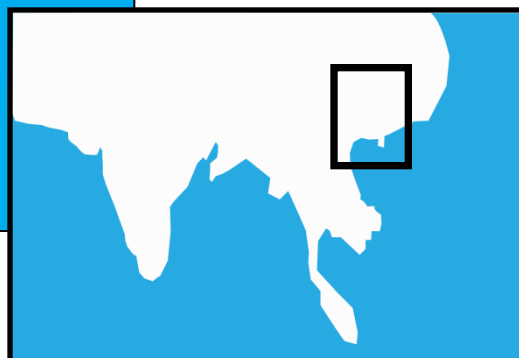
Nesi et al (2011) Possible introgression between *Epomophorus gambianus* and *Micropteropus pusillus*
Comptes Rendus Biologies, 334, 544-554.

**Example 1: *Rhinolophus pearsoni* and *Rhinolophus yunanensis***

***R. p. pearsoni***

***R. p. chinensis***

***R. yunanensis***



**Example 1: *Rhinolophus pearsoni* and *Rhinolophus yunanensis***

***R. p. pearsoni***

***R. p. chinensis***

***R. yunanensis***

SPECIES TREE

Mt DNA tree (2 genes)

**Example 1: *Rhinolophus pearsoni* and *Rhinolophus yunanensis***

***R. p. pearsoni***

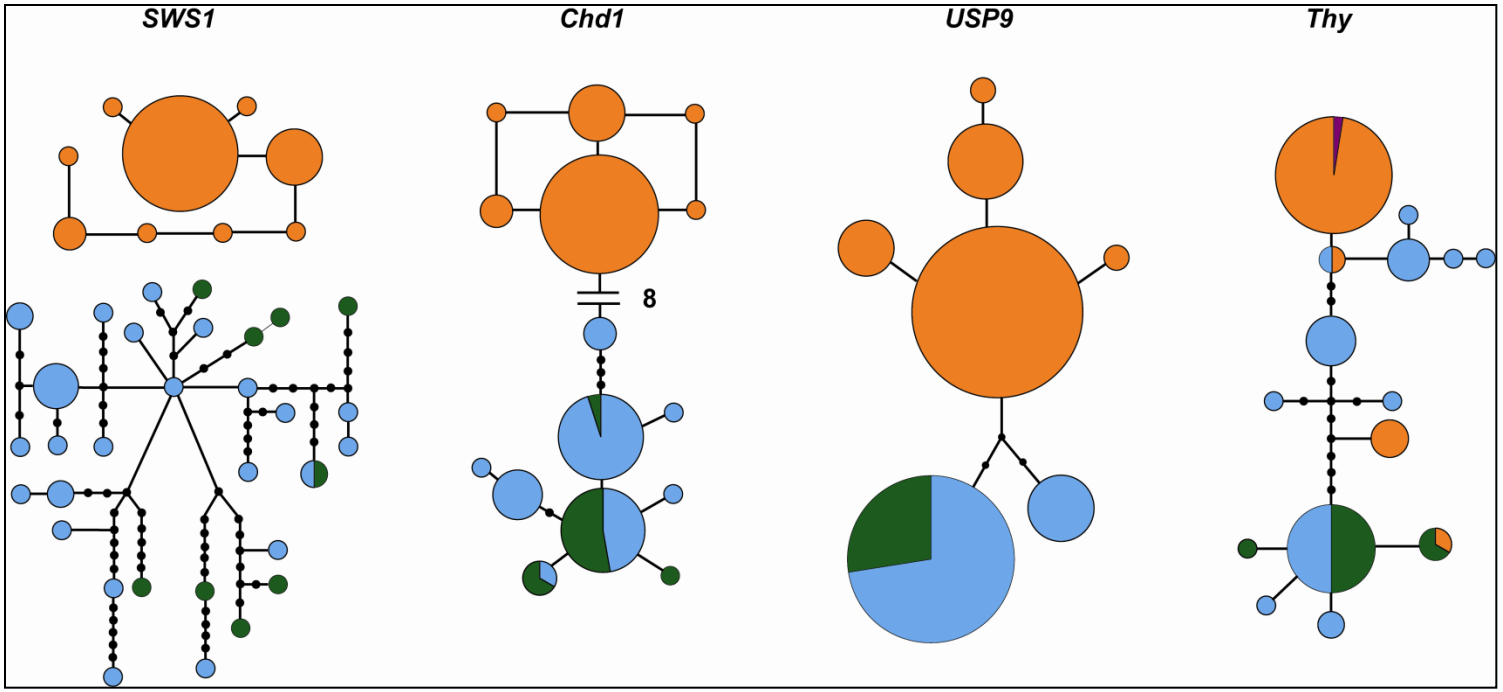
***R. p. chinensis***

***R. yunanensis***

Used the SWS1 pseudogene

Nuclear intron networks

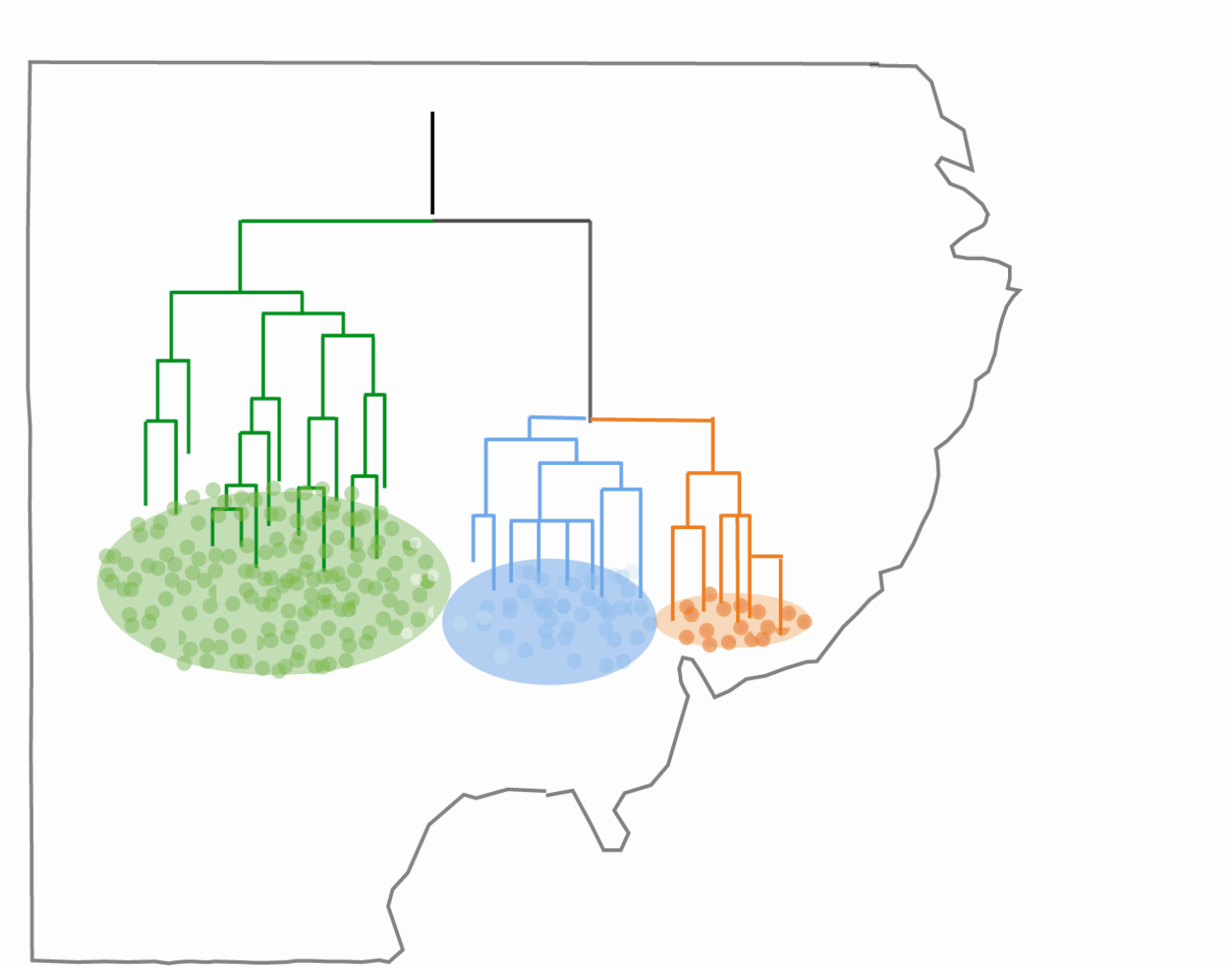
*R. p. chinensis*



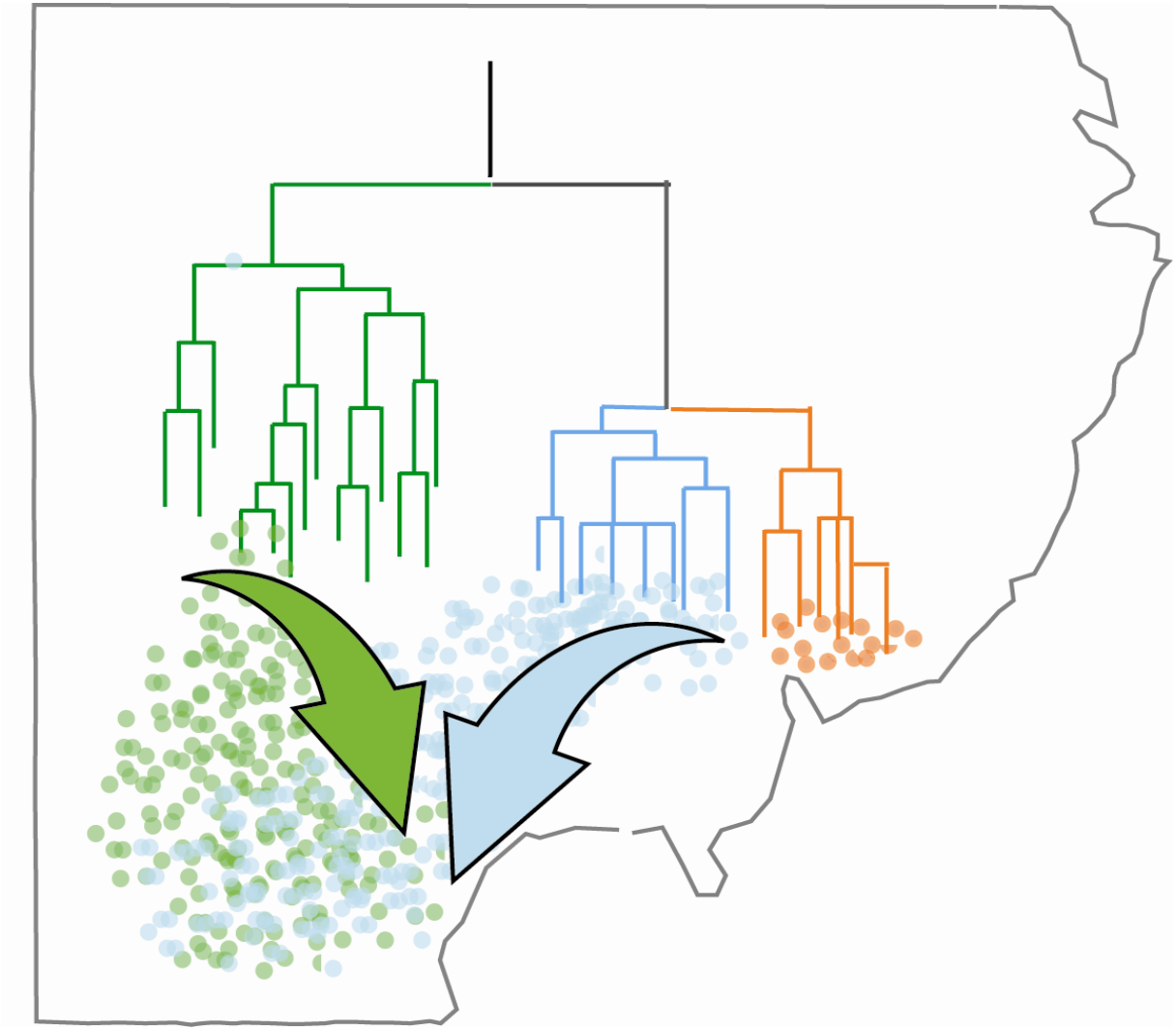

*R. yunanensis*

*R. p. pearsoni*

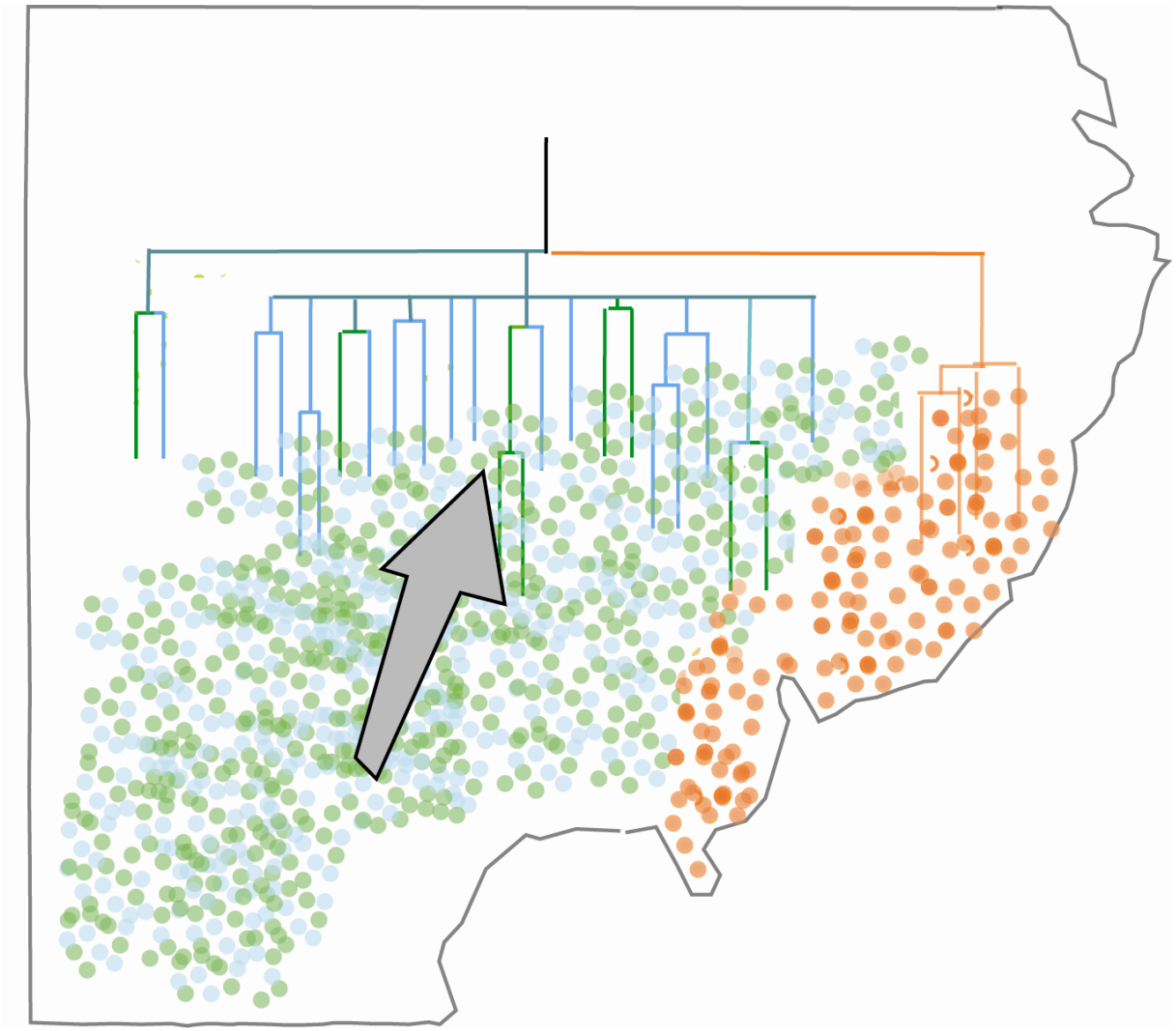
History of the nuclear genes studied



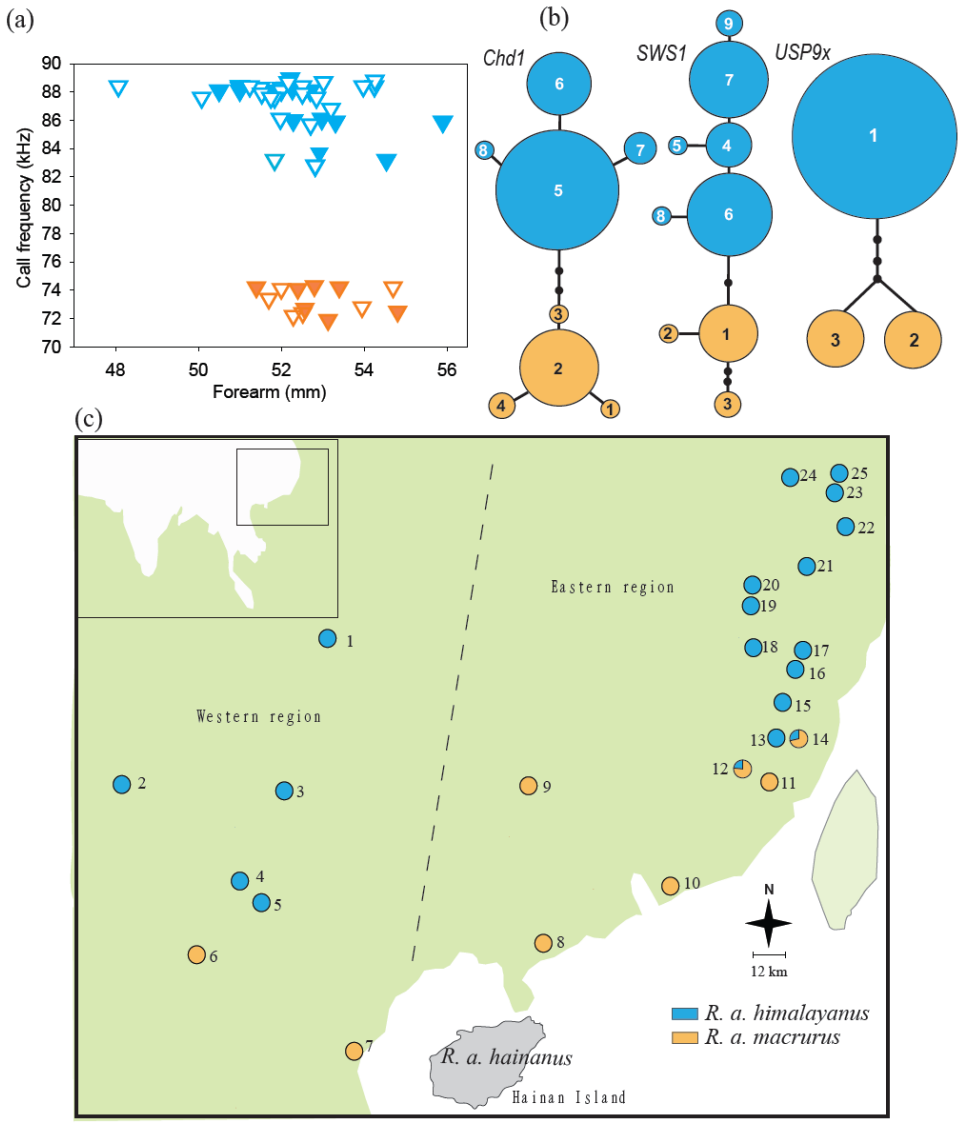
History of the nuclear genes studied



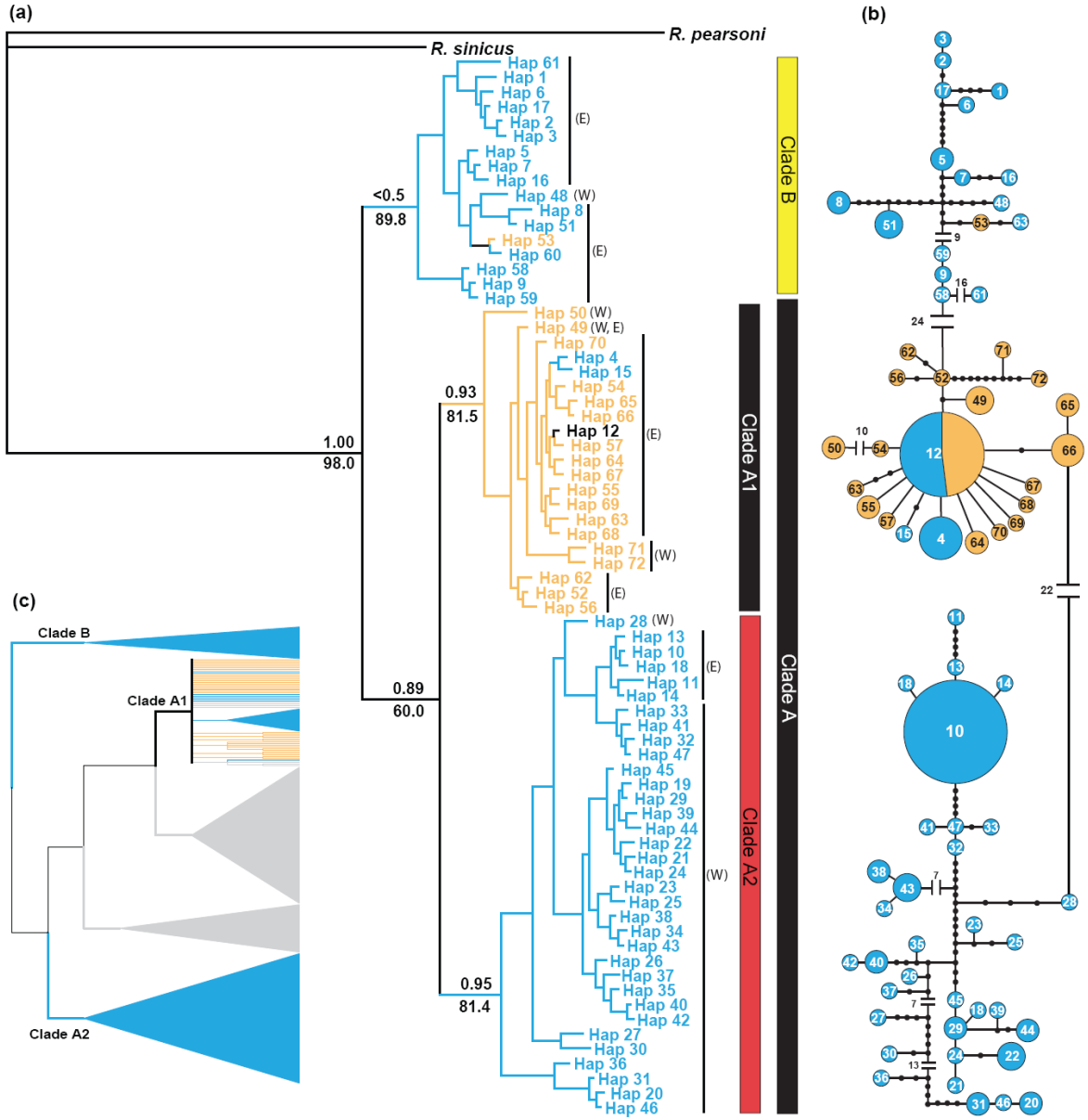
History of the nuclear genes studied



# Example 2: *Rhinolophus affinis himalayanus* and *R. a. macrurus*



# Example 2: *Rhinolophus affinis himalayanus* and *R. a. macrurus*



## Detecting Introgression

Taxa must have a contact zone or have been in contact in the past

Often a geographical pattern

More commonly detected in mtDNA (barcoding caveat)

More common where one taxon has undergone population expansion

Neutral genes typically flow from resident taxon to the invading taxon

# Why do phylogenetic trees sometimes disagree with other datasets?

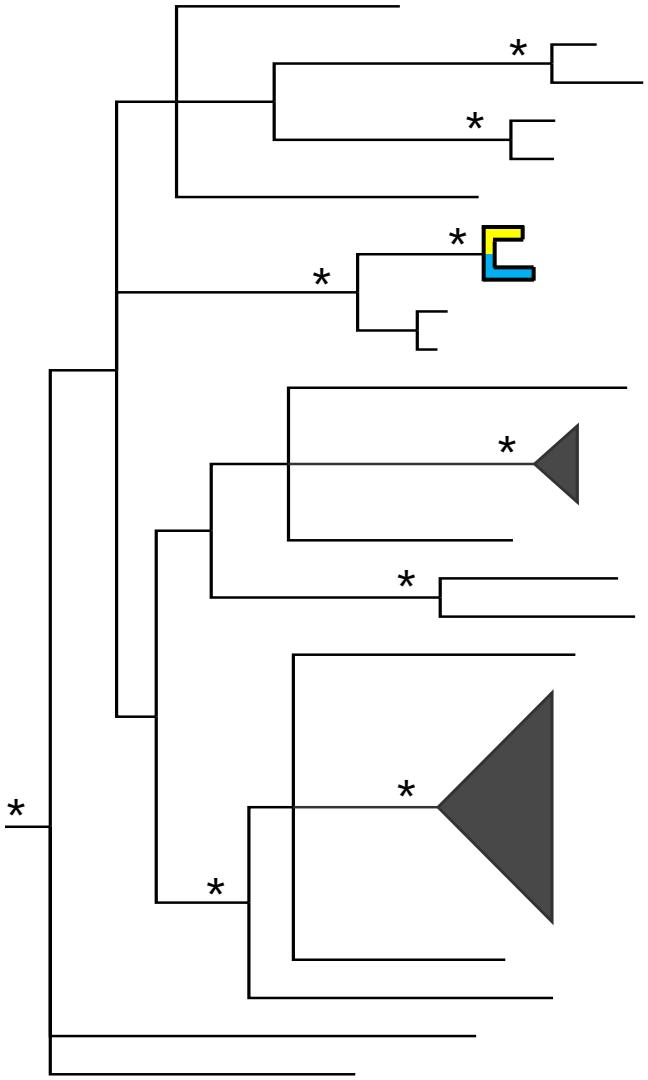
1. Incomplete sorting

2. Long branch attraction

3. Introgression

4. Homoplasy

# Bayesian tree of *Murina* based on mtDNA COI (637 bps)



*M. gracilis*

*M. recondita*

* posterior probability > 0.95

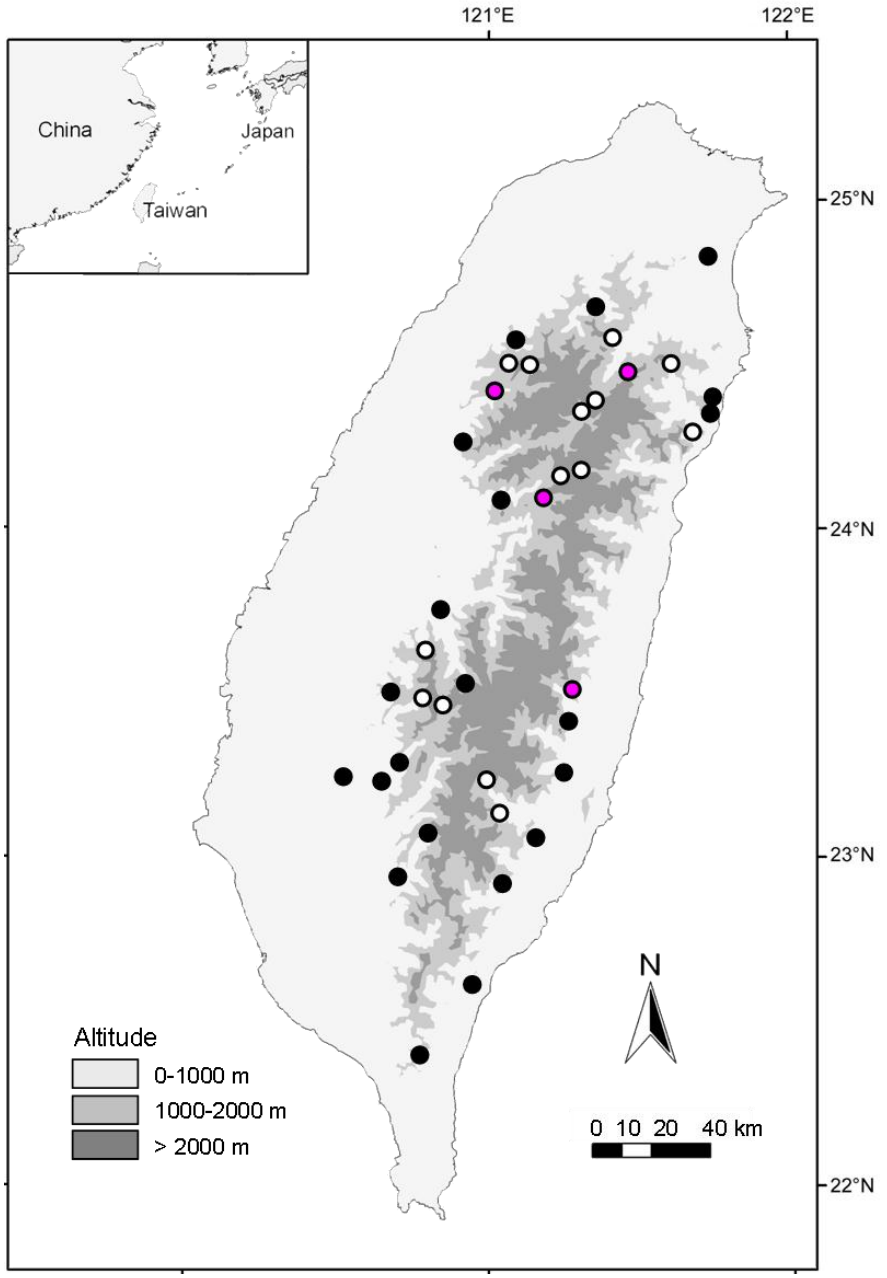
*Murina gracilis* (○)

≥ 1500 m ASL

*M. recondita* (●)

≤ 1500 m ASL

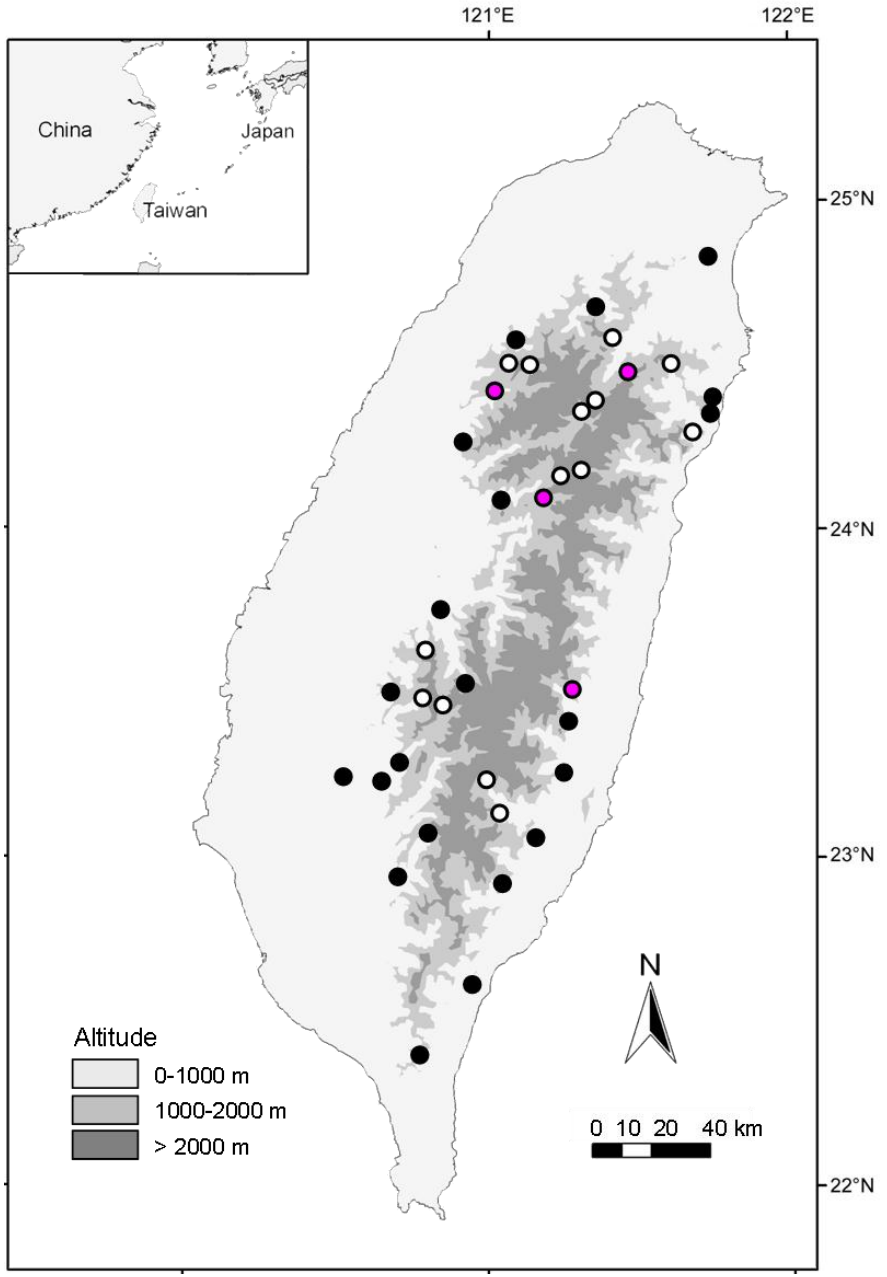
Both species (●)

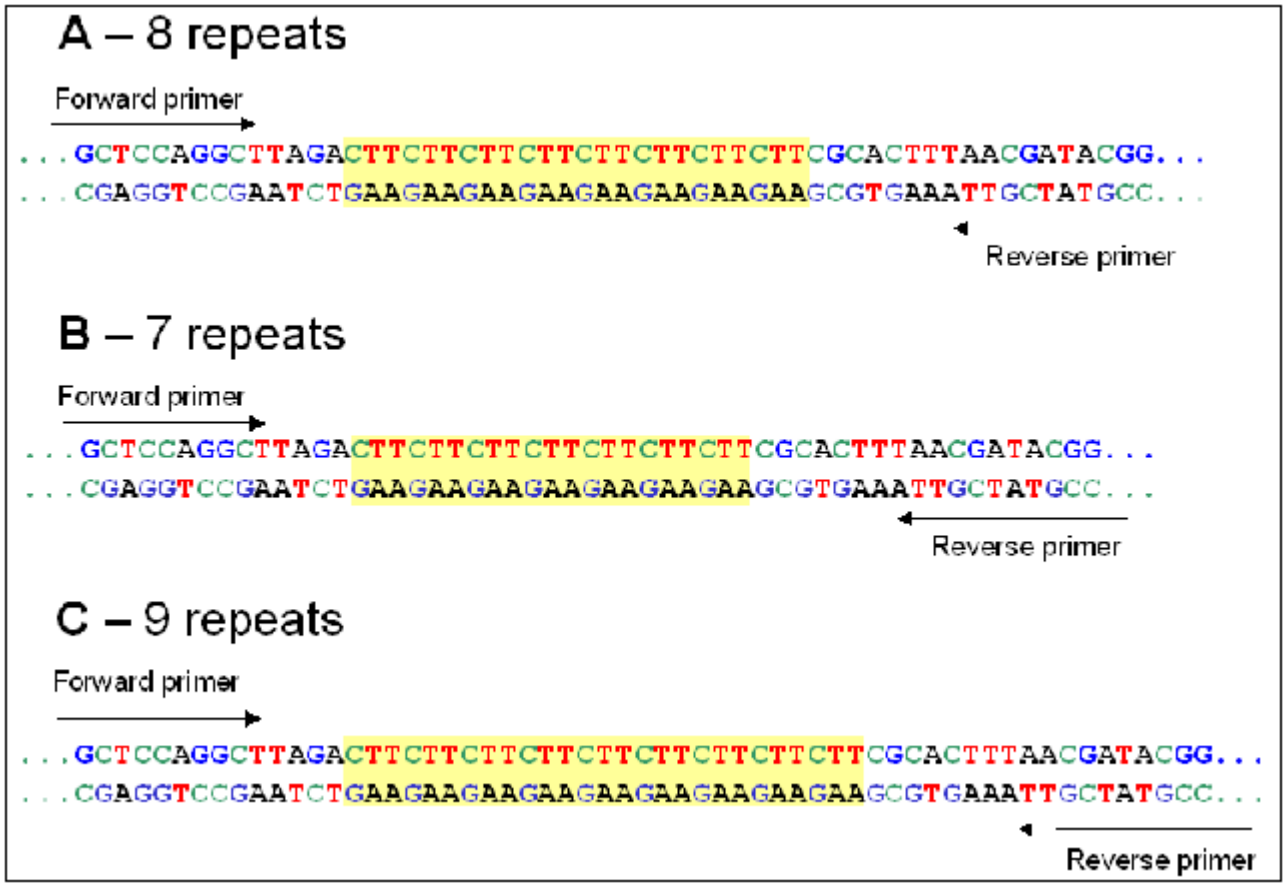


Input: 106 *M. gracilis*

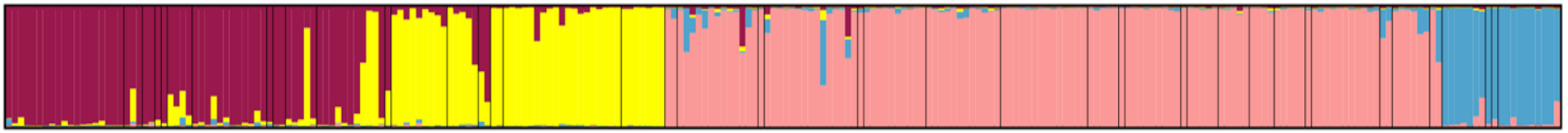
144 *M. recondita*

14 microsatellite loci

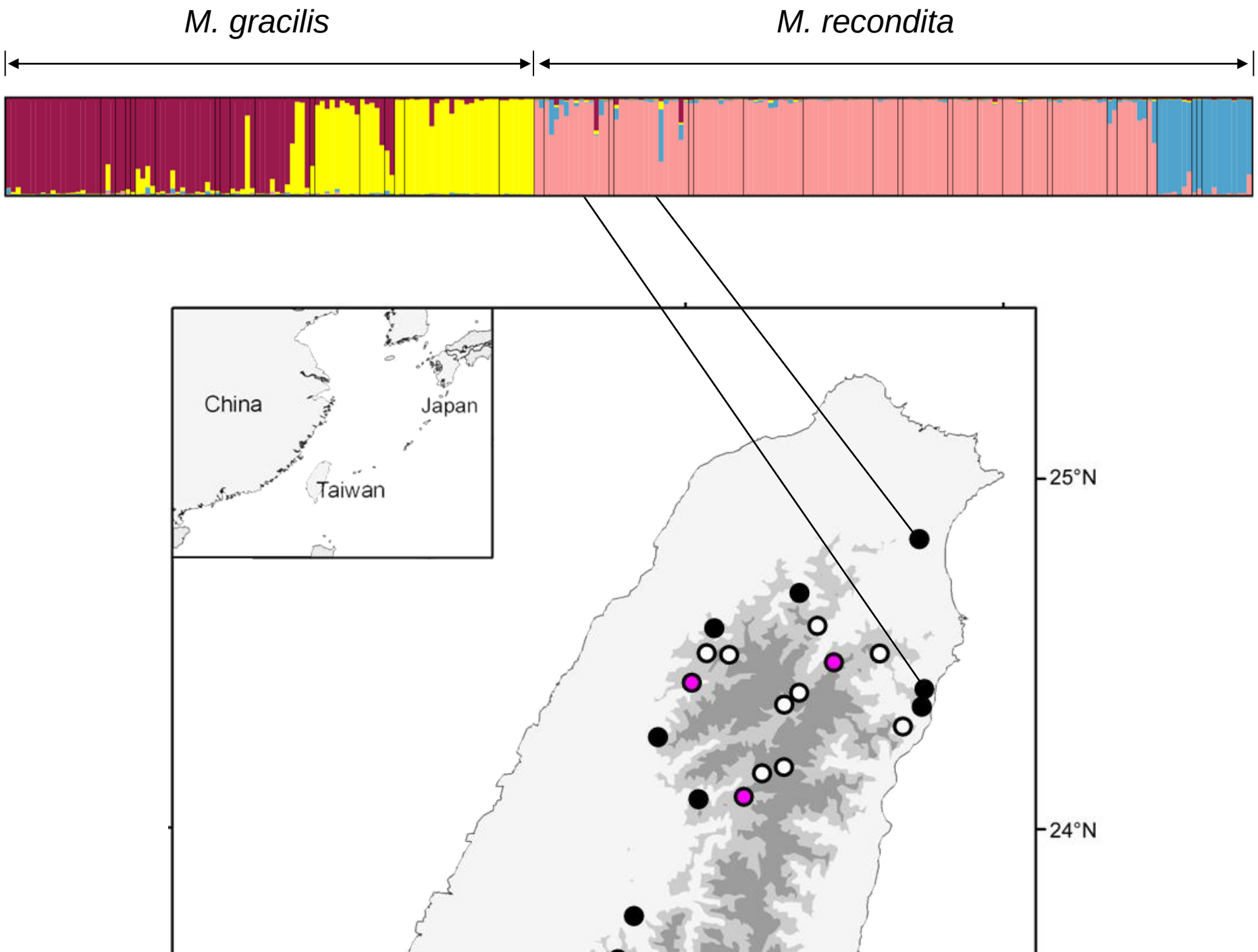


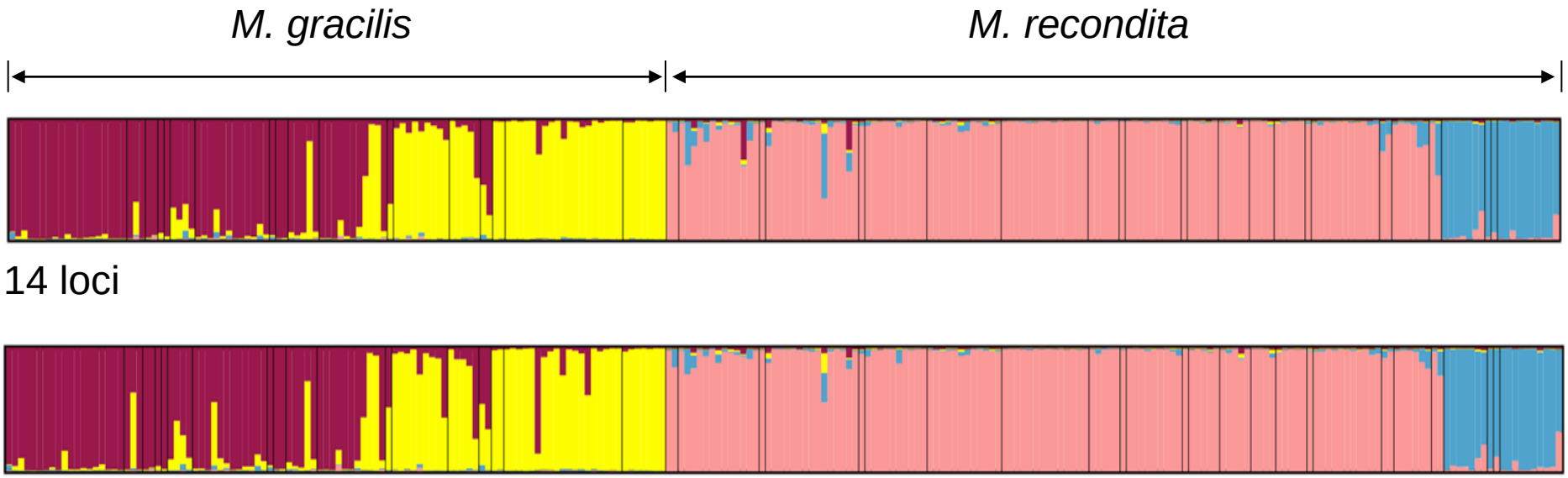

A – 8 repeats
Forward primer
...GCTCCAGGCTTAGACTTCTTCTTCTTCTTCTTCTTCTTCGCACTTTAACGATACGG...
...CGAGGTCCGAATCTGAAGAAGAAGAAGAAGAAGAAGAAGCGTGAAATTGCTATGCC...
Reverse primer
B – 7 repeats
Forward primer
...GCTCCAGGCTTAGACTTCTTCTTCTTCTTCTTCTTCGCACTTTAACGATACGG...
...CGAGGTCCGAATCTGAAGAAGAAGAAGAAGAAGAAGCGTGAAATTGCTATGCC...
Reverse primer
C – 9 repeats
Forward primer
...GCTCCAGGCTTAGACTTCTTCTTCTTCTTCTTCTTCTTCTTCGCACTTTAACGATACGG...
...CGAGGTCCGAATCTGAAGAAGAAGAAGAAGAAGAAGAAGAAGCGTGAAATTGCTATGCC...
Reverse primer

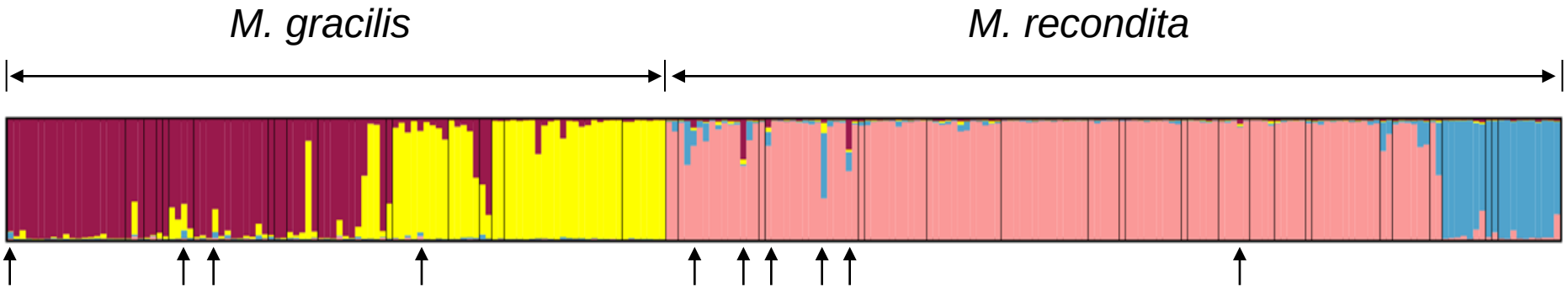
*M. gracilis* *M. recondita*



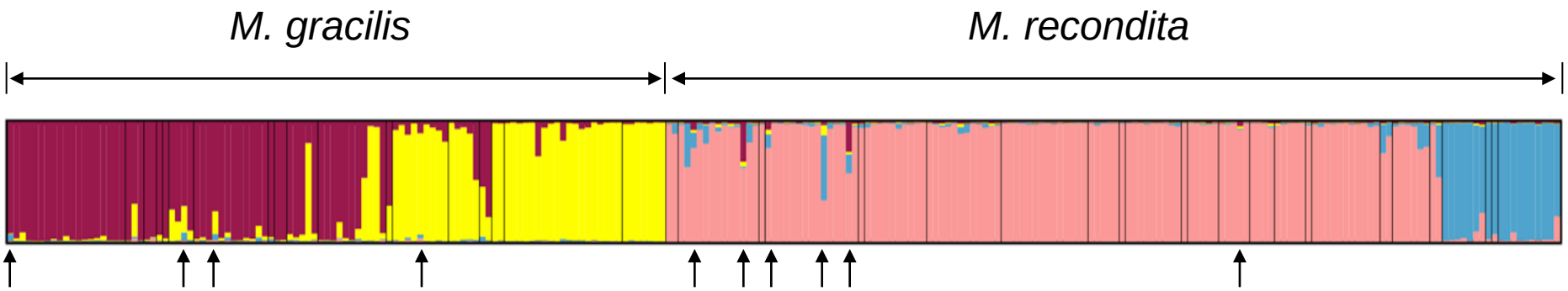
K = 4


M. gracilis
M. recondita
China
Japan
Taiwan
25°N
24°N


M. gracilis
M. recondita
14 loci
13 loci: without locus “A9”




Recent hybridization? Or homoplasy?



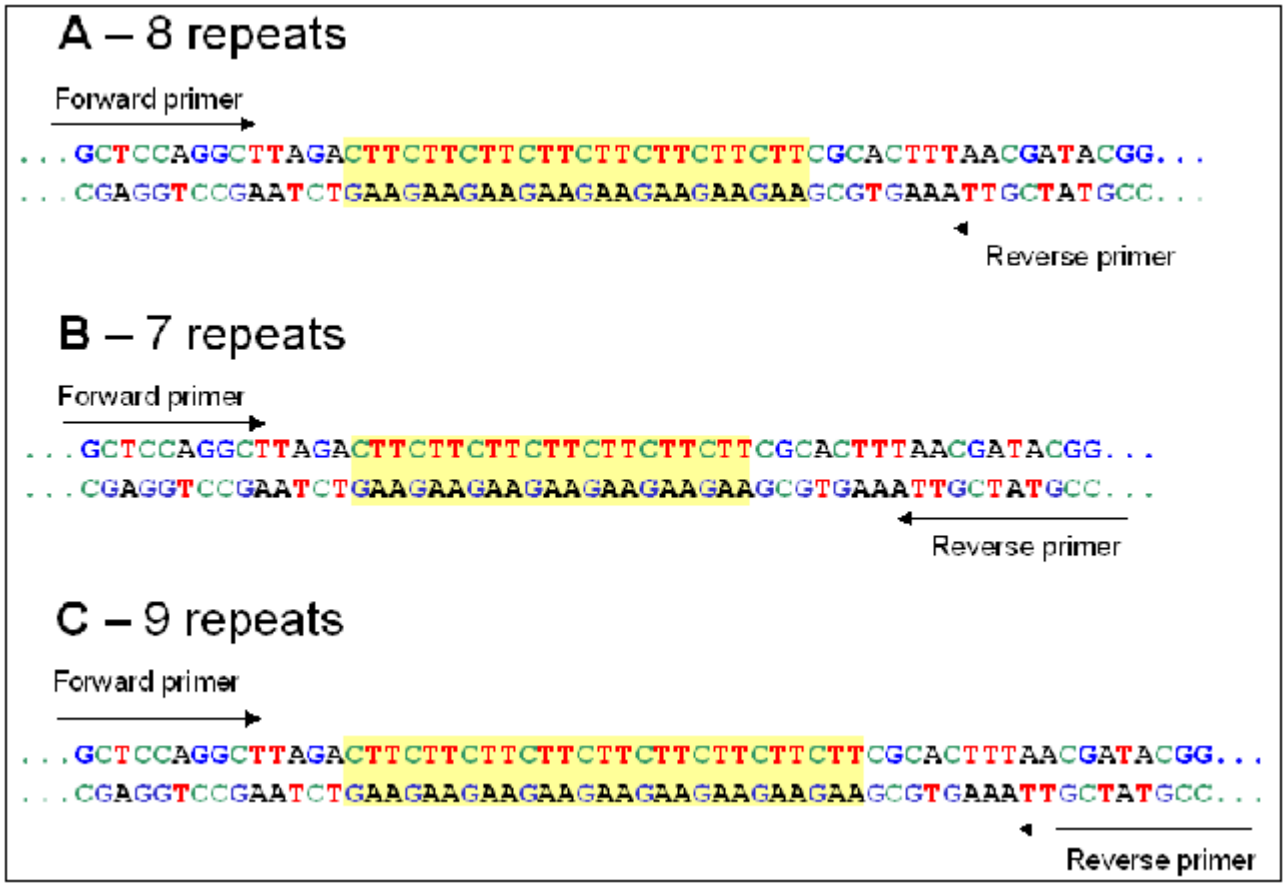

# Recent hybridization? Or homoplasy?

Sequencing the flanking regions of loci showing mixed ancestry for some individuals (↑):

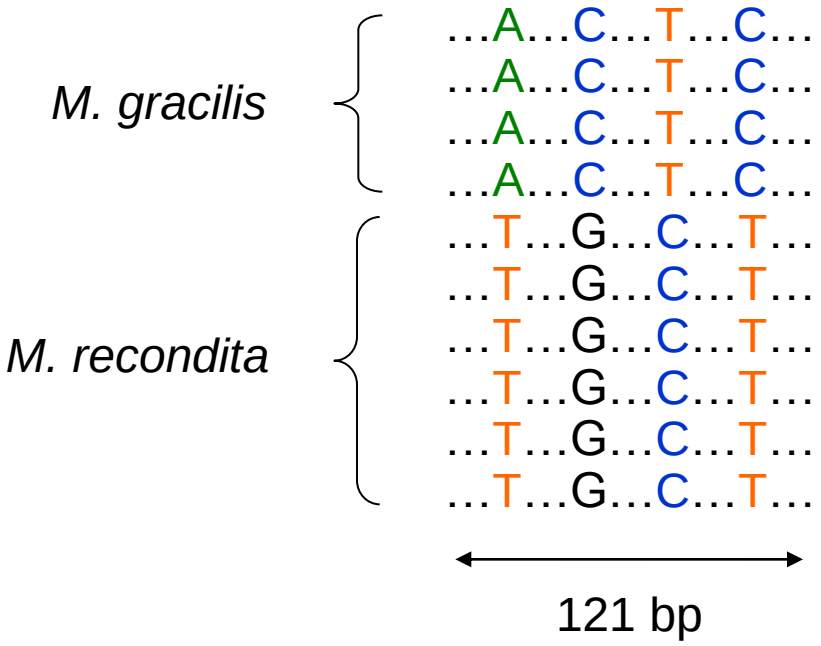
Predictions:

If hybridization: some *M. gracilis* will phylogenetically group with *M. recondita*, and vice versa.

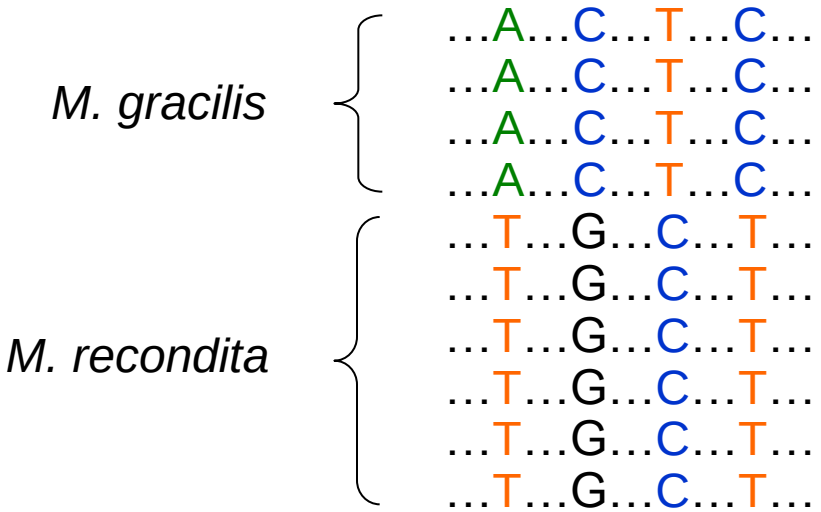
If homoplasy: samples of different species will be phylogenetically separate


A – 8 repeats
Forward primer
...GCTCCAGGCTTAGACTTCTTCTTCTTCTTCTTCTTCTTCGCACTTTAACGATACGG...
...CGAGGTCCGAATCTGAAGAAGAAGAAGAAGAAGAAGAAGCGTGAAATTGCTATGCC...
Reverse primer
B – 7 repeats
Forward primer
...GCTCCAGGCTTAGACTTCTTCTTCTTCTTCTTCTTCGCACTTTAACGATACGG...
...CGAGGTCCGAATCTGAAGAAGAAGAAGAAGAAGAAGCGTGAAATTGCTATGCC...
Reverse primer
C – 9 repeats
Forward primer
...GCTCCAGGCTTAGACTTCTTCTTCTTCTTCTTCTTCTTCTTCGCACTTTAACGATACGG...
...CGAGGTCCGAATCTGAAGAAGAAGAAGAAGAAGAAGAAGAAGCGTGAAATTGCTATGCC...
Reverse primer

# Flanker sequencing result for locus “A9”



# Flanker sequencing result for locus “A9”




The sequencing result for another locus “A122” (209 bp) is also consistent with the prediction of the allele size homoplasy

# Why do phylogenetic trees sometimes disagree with other datasets?

1. Incomplete sorting

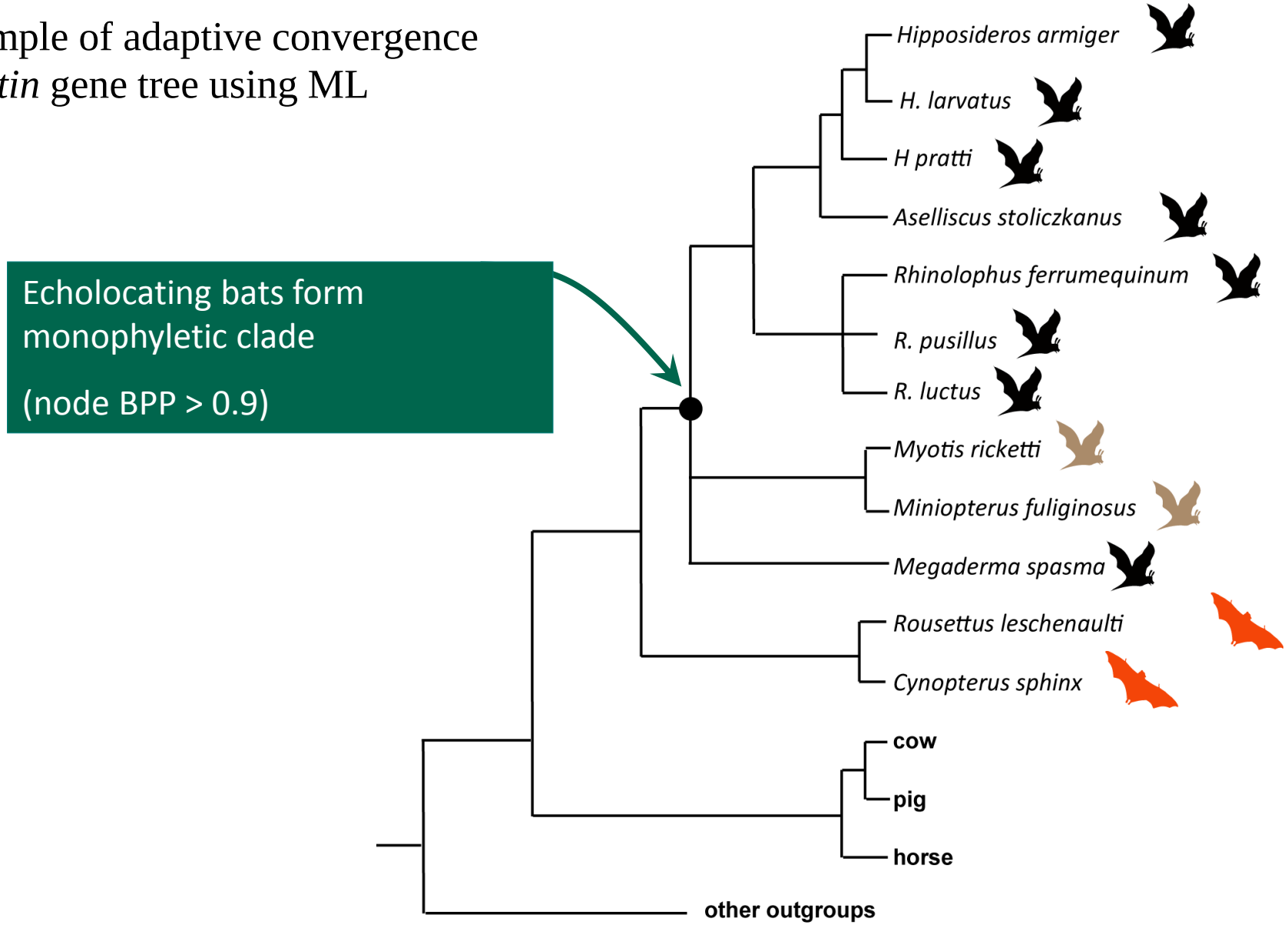
2. Long branch attraction

3. Introgression

4. Homoplasy

5. Adaptive convergence

Example of adaptive convergence
*Prestin* gene tree using ML



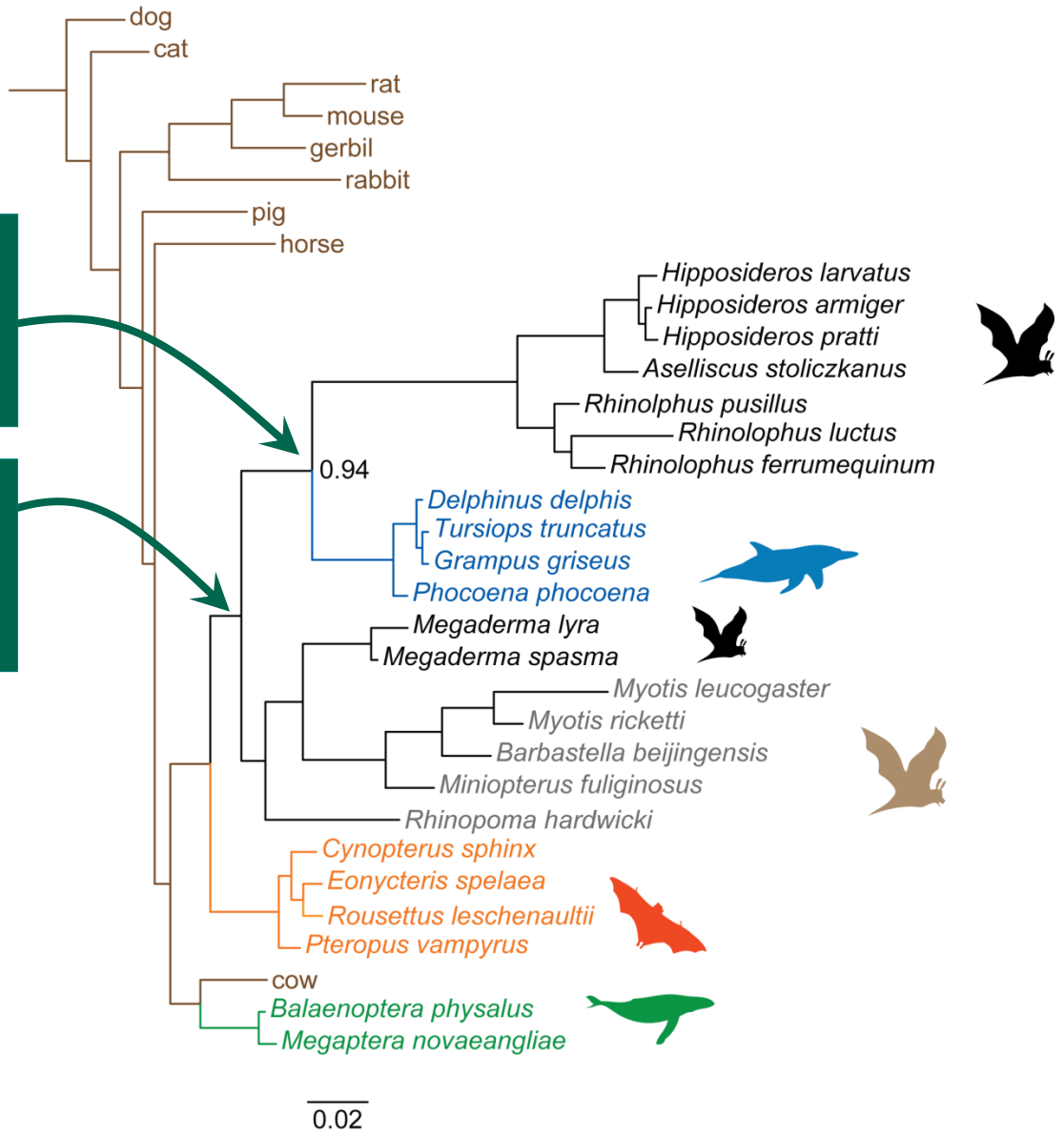
# *Prestin* gene tree using ML

Dolphins & horseshoe bats form monophyletic clade

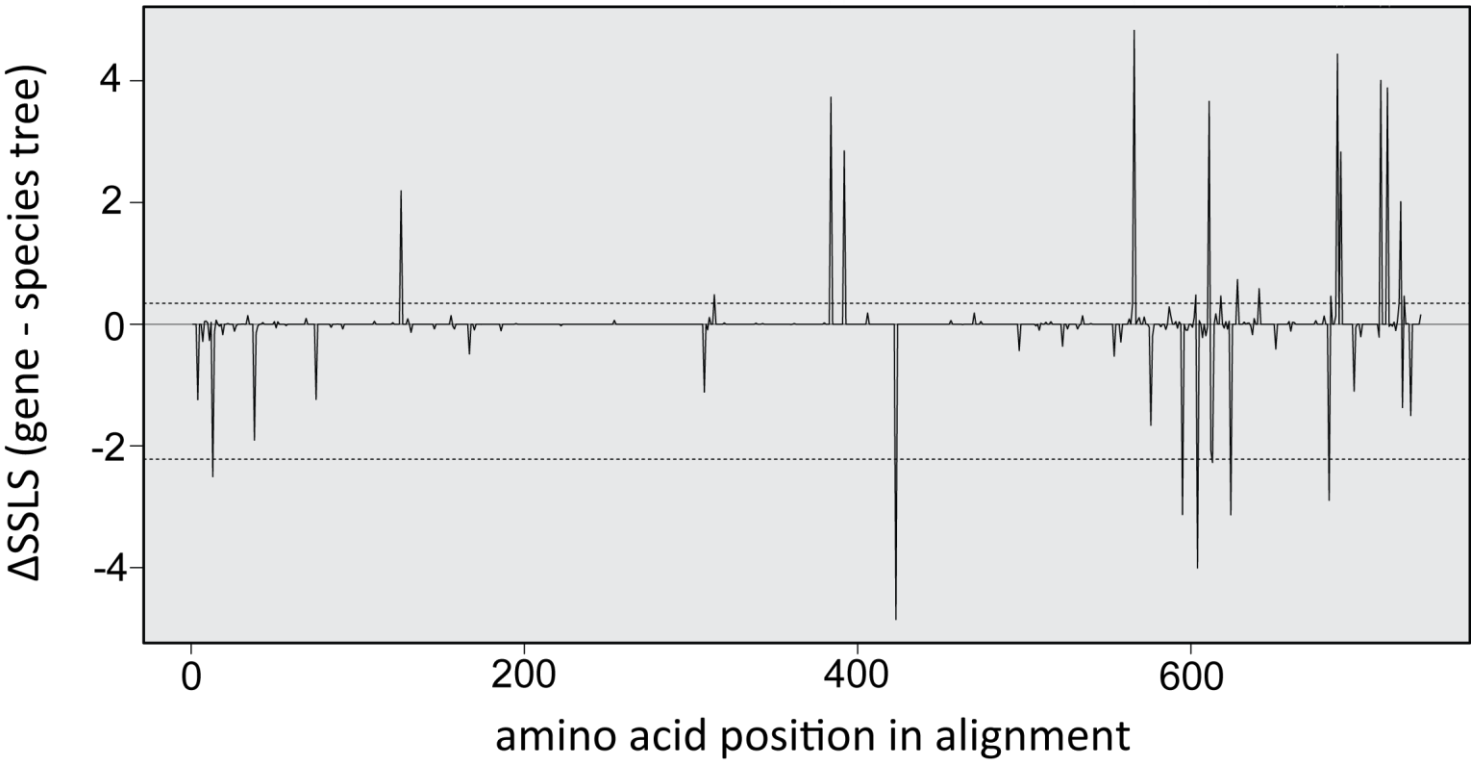
BPP > 0.9

Echolocating bats & cetaceans form monophyletic clade

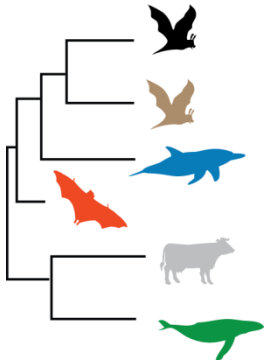
BPP > 0.65

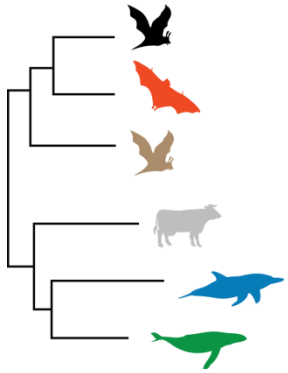


# Revealing site-wise convergence




Support for "wrong tree")





Support for correct tree

# Results from 1200 genes