The Genome 10K Project:
An Overview of Now and Beyond

Klaus-Peter Koepfli, Ph.D.
Laboratory of Genomic Diversity
Frederick National Laboratory for Cancer Research
Frederick, MD 21702 USA
klauspeter.koepfli527@gmail.com

10,000 Vertebrate Genome Project

Journal of Heredity 2009;100(6):659–674
doi:10.1093/jhered/esp086
Advance Access publication November 5, 2009

Genome 10K: A Proposal to Obtain Whole-Genome Sequence for 10 000 Vertebrate Species

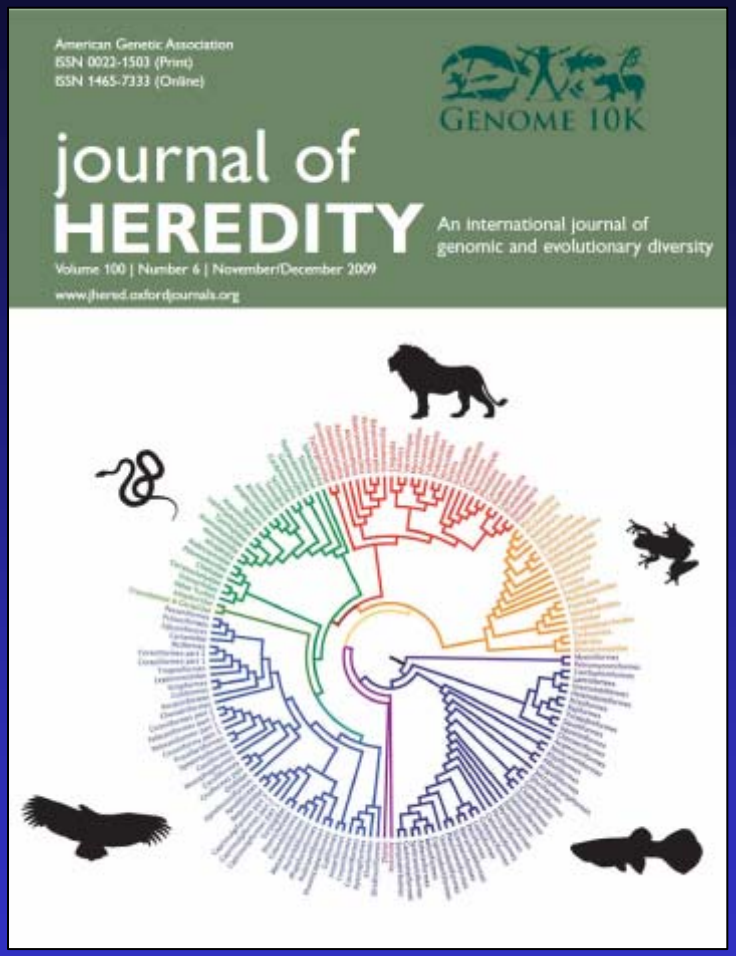
Genome 10K Community of Scientists*

*G10KCOS authors are listed in the Appendix

Abstract

The human genome project has been recently complemented by whole-genome assessment sequence of 32 mammals and 24 nonmammalian vertebrate species suitable for comparative genomic analyses. Here we anticipate a precipitous drop in costs and increase in sequencing efficiency, with concomitant development of improved annotation technology and, therefore, propose to create a collection of tissue and DNA specimens for 10 000 vertebrate species specifically designated for whole-genome sequencing in the very near future. For this purpose, we, the Genome 10K Community of Scientists (G10KCOS), will assemble and allocate a biospecimen collection of some 16 203 representative vertebrate species spanning evolutionary diversity across living mammals, birds, sauropsian reptiles, amphibians, and fishes (i.e. 60 000 living species). In this proposal, we present precise counts for these 16 203 individual species with specimens presently tagged and stipulated for DNA sequencing by the G10KCOS. DNA sequencing has ushered in a new era of investigation in the biological sciences, allowing us to embark for the first time on a truly comprehensive study of vertebrate evolution, the results of which will touch nearly every aspect of vertebrate biological inquiry.

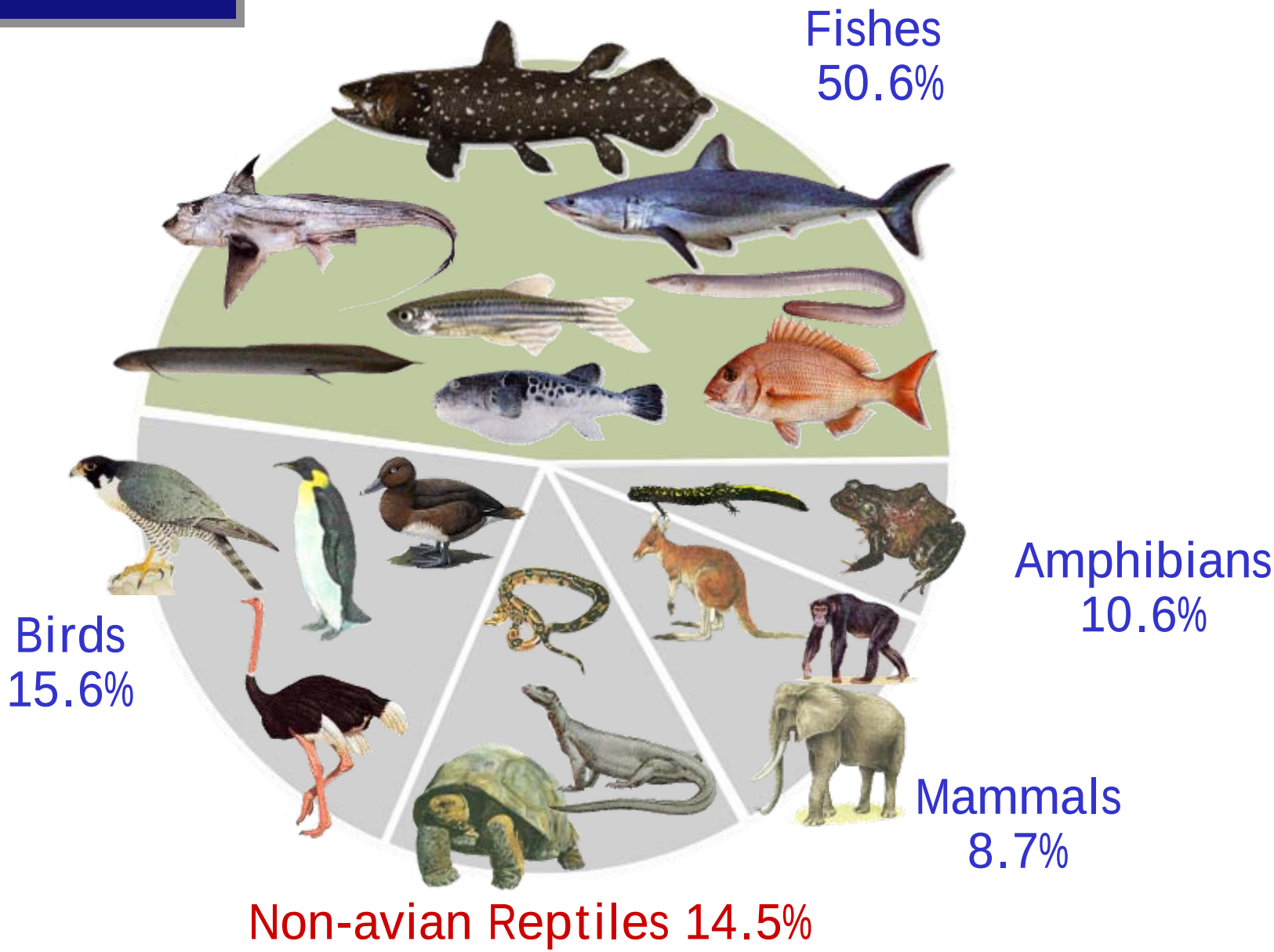
Key words: ancestral state reconstruction, comparative genomics, G10K, molecular evolution, species conservation, vertebrate biology



GENOME 10K


GENOME 10K

Number of Species



Examples of Selected Species

Horned marsupial frog

Monito del Monte

Ocean sunfish

Great white shark

GENOME 10K

Selecting a species: Phylogenetic diversity

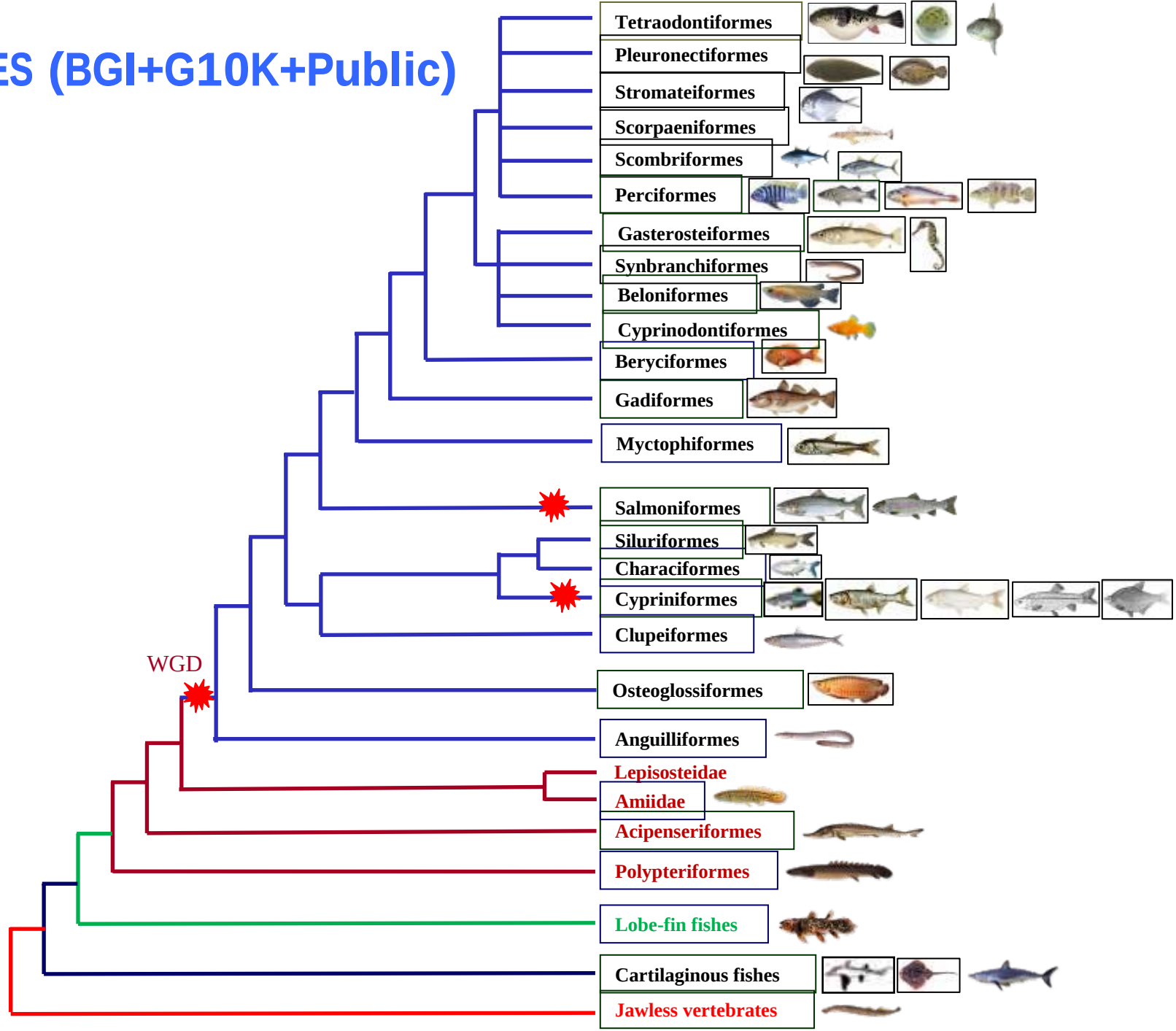
- A selected species captures the phylogenetic depth of a groups' evolutionary radiation maximizing phylogenetic representation across vertebrate evolutionary lineages.

GENOME 10K

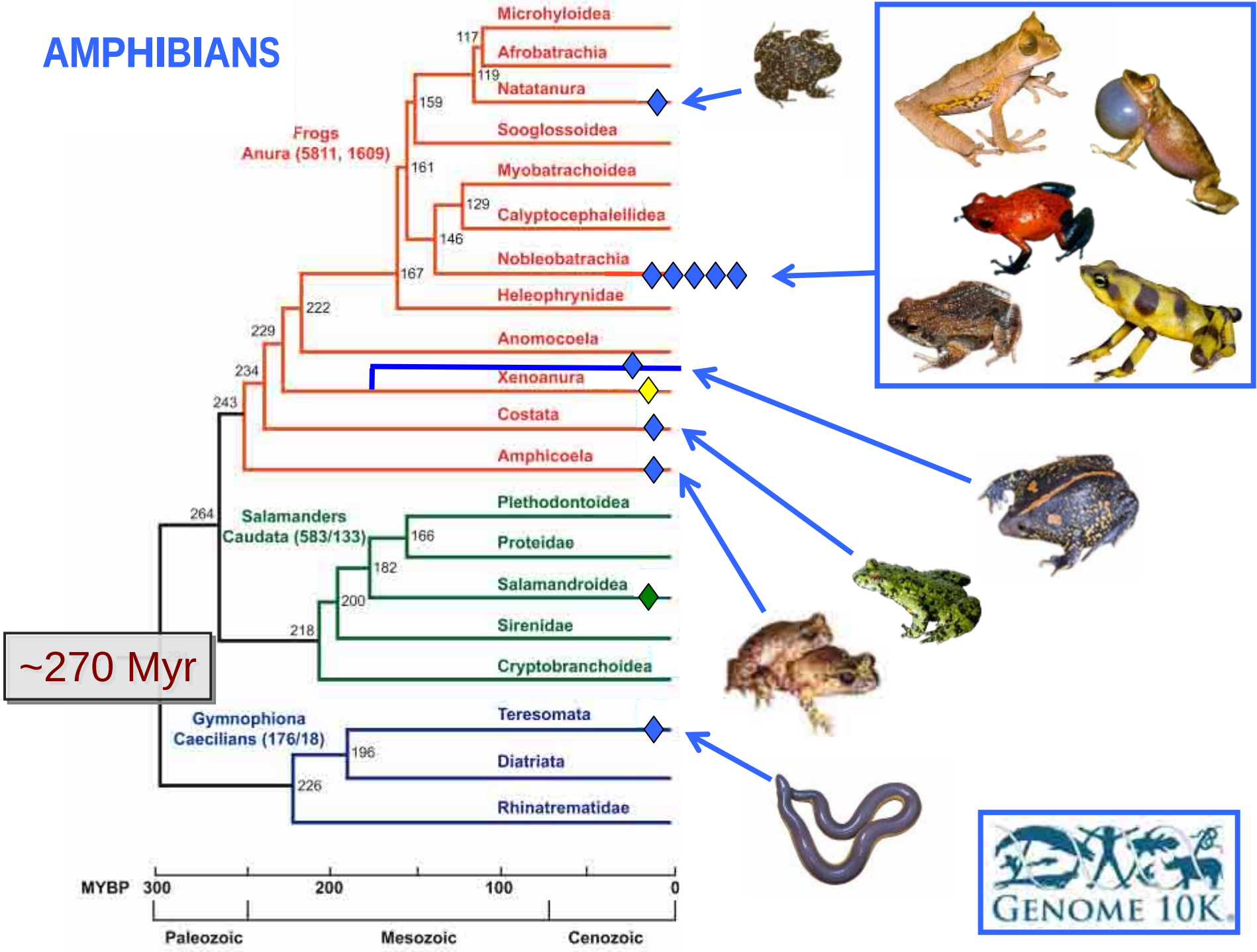
The most diverse group of vertebrates

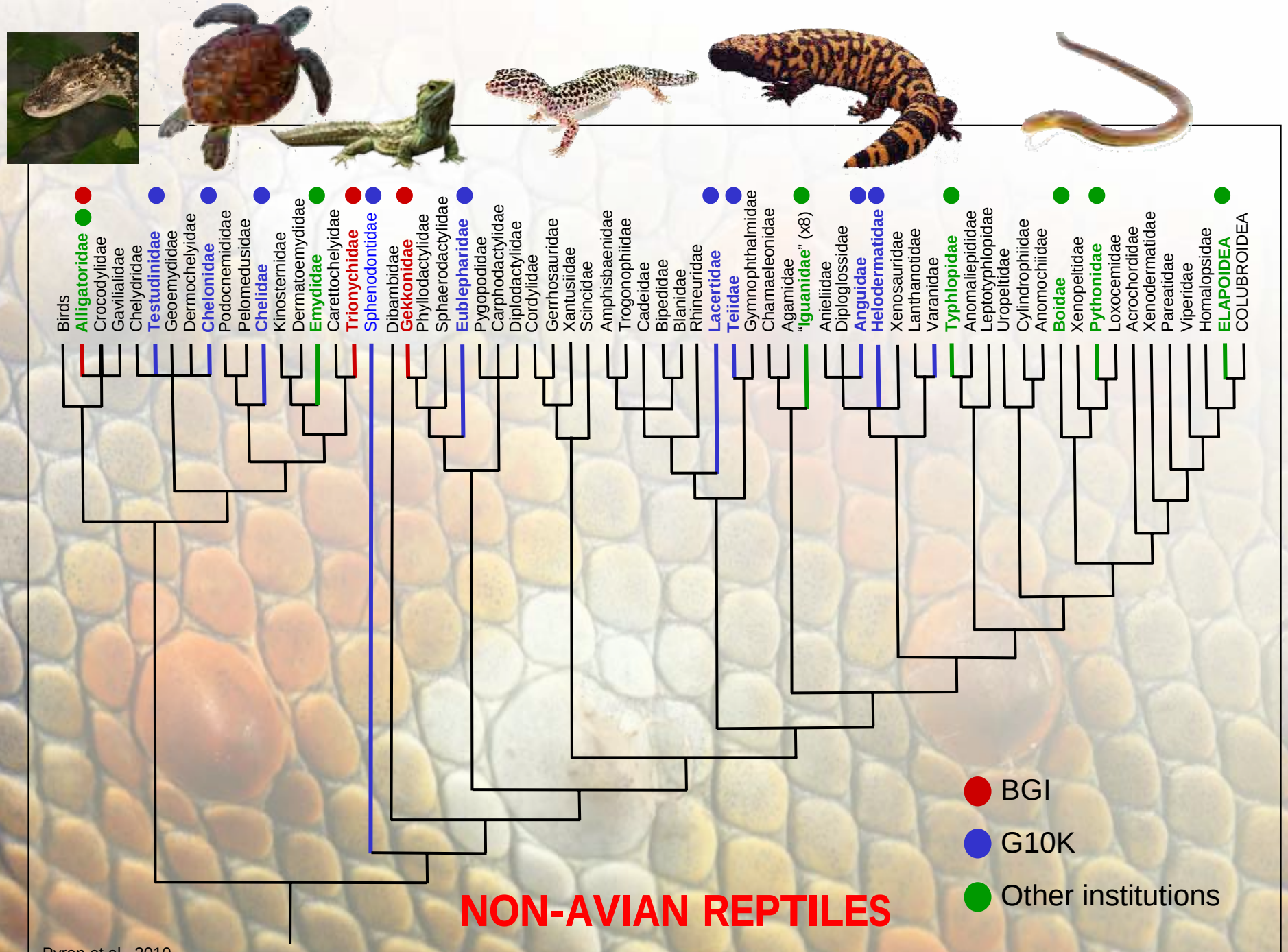
Modified from Meyer, 2005

FISHES (BGI+G10K+Public)



AMPHIBIANS

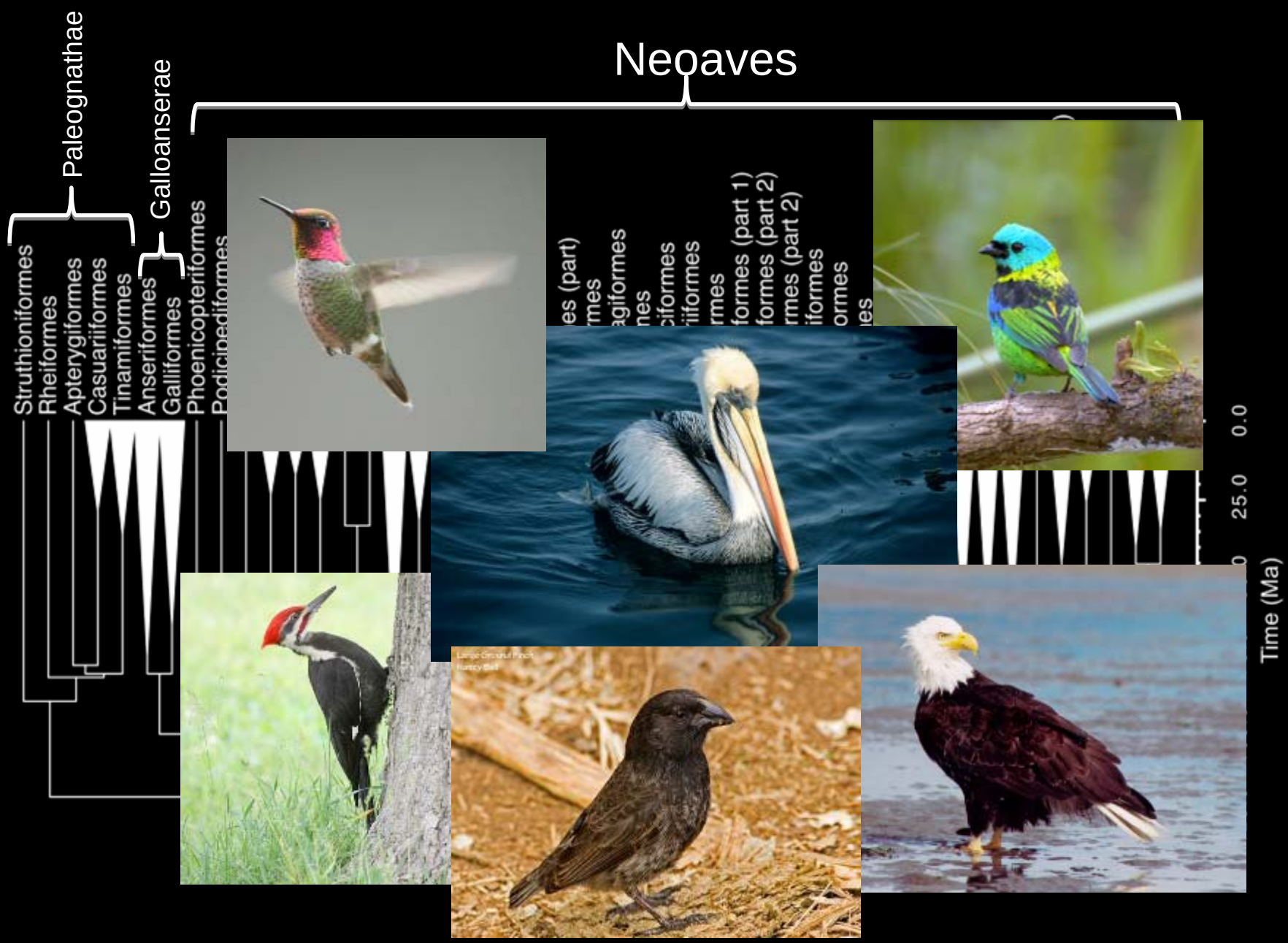

Microhyloidea
Afrobatrachia
Natatanura
Sooglossoidea
Frogs
Anura (5811, 1609)
Myobatrachoidea
Calyptocephalellidea
Nobleobatrachia
Heleophrynidae
Anomocoela
Xenoanura
Costata
Amphicoela
Plethodontoidea
Salamanders
Caudata (583/133)
Proteidae
Salamandroidea
Sirenidae
Cryptobranchoidea
Gymnophiona
Caecilians (176/18)
Teresomata
Diatriata
Rhinatrematidae
117
119
159
161
129
146
167
222
229
234
243
264
166
182
200
218
196
226
~270 Myr
MYBP 300
200
100
0
Paleozoic
Mesozoic
Cenozoic
GENOME 10K



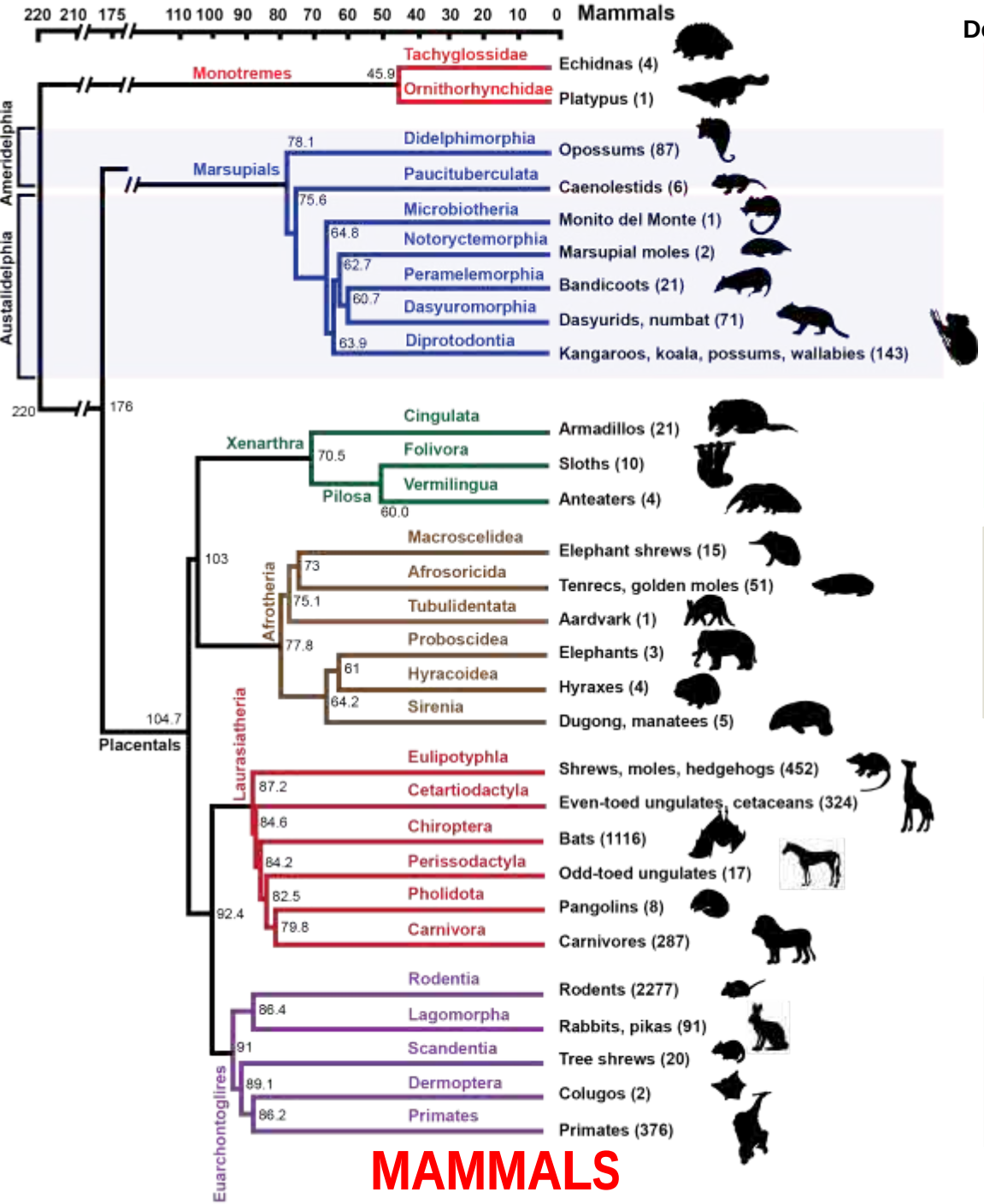
Pyron et al., 2010

After Hackett et al. 2008, *Science* 320: 1763

After Hackett et al. 2008, *Science* 320: 1763



After Hackett et al. 2008, *Science* 320: 1763



Public Done	Public Ongoing	Broad/G10K Proposed	BGI/G10K Done	BGI/G10K Ongoing	TOTAL
–	–	–	–	1	1
1	–	–	–	–	1
1	–	2	–	–	3
–	–	1	–	–	1
–	–	1	–	1	2
–	–	1	–	1	2
–	–	1	–	1	2
1	–	1	–	–	2
1	–	10	–	2	13
1	–	3	–	–	4
1	–	1	–	–	2
–	–	2	–	–	2
–	–	1	–	–	1
1	–	4	–	–	5
–	–	1	–	–	1
2	–	1	–	–	3
1	–	1	–	–	2
–	–	2	–	–	2
2	–	5	–	1	8
2	8	2	3	5	20
1	2	8	–	3	14
1	–	2	–	–	3
–	1	–	–	–	1
2	5	4	5	3	19
4	5	34	–	3	46
2	–	–	–	–	2
1	–	1	–	–	2
–	1	–	–	–	1
8	6	11	1	–	26
37	**28**	**100**	**9**	**21**	**195**

Selecting a species: <u>Scientific community</u>

- The selected species enjoys a biological community with established biological applications for science, comparative medicine, or biology or society.

GENOME 10K

Selecting a species: <u>Scientific context</u>

- For the selected species there exists an explicit scientific value in studying the species historically and for the future

GENOME 10K

Amphibians proposed by G10K

Tissue status:

To be sent directly to BGI, awaiting permit.

Archey's frog: *Leiopelma archeyi*

Scientific motivation:

Represents sister lineage to all other frogs.
Direct development, independently derived.
Critically endangered.

Genome size:	~7.8 Gb
Sex chromos.:	ZW? OW?
IUCN status:	CR
Distribution:	New Zealand
Family:	Leiopelmatidae

Overview

Selecting a species: interesting biological questions

- Extreme or unusual adaptations: biochemical, physiological, morphological

- Phenotypic diversity, convergences, parallelisms

- Sex determining systems, sexual dimorphism



Selecting a species: <u>Popular recognition</u>

- The species enjoys a popular image, recognition and utility, such as domesticated species, conservation icons, national animals, and wildlife icons.

GENOME 10K

Procuring an appropriate sample

- Vouchered specimen or live animal
  - Voucher ID
  - Gender
  - Geographic origin
  - Preservation method
  - Sample type
  - Sample quality
  - DNA quality

The Frozen Zoo®

Mission of the Frozen Zoo

To help preserve the legacy of life on Earth for future generations by establishing and maintaining genetic resources in support of worldwide efforts in research and conservation.

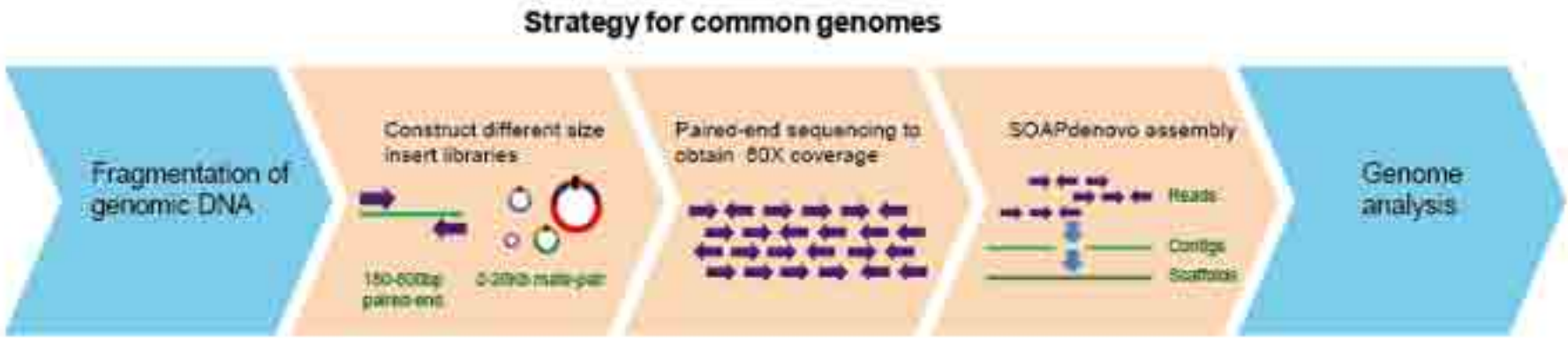
FROZEN ZOO®

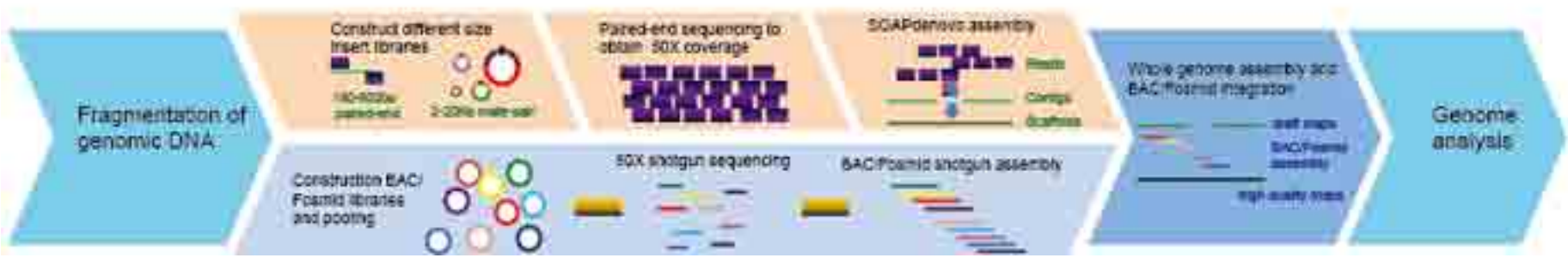
SAN DIEGO ZOO
INSTITUTE FOR
CONSERVATION
RESEARCH

Ideal Sample has:

- DNA barcode (released immediately, prior to genomic sequencing).
- **Vouchered specimen**
- Cell lines for DNA
- Cell lines for RNA (less preferred: field source of mRNA)
- **Access to fresh material**
- Repository of more DNA from same individual
- Repository of DNAs from multiple individuals of same species (for SNP discovery)
- **Large amounts of DNA, 700 µg – 1.0 mg!**
- **<u>Complete permit trail</u>**
- Heterogametic sex when identified (Bank both sexes)
- Possibility of sorting the sex chromosome(s)
- Possibility of physical mapping onto chromosomes? FiSH?
- Possibility of obtaining "trios"

Genome Sequencing Strategies: BGI





FROM: http://www.ldl.genomics.cn/page/pa-ct.jsp

Timetable and Tracking

- Quarterly update from centers - simple format with automated updates of accumulated coverage
- Sample tracking/delivery
- Near real time project management in centralized location (e.g. BGI)
- Updates available to sample donors
- Data must be available in public repository for someone to get credit
- Use Fort Lauderdale rules for submission times
- Assembled data in a reasonable amount of time given achieving particular standards.
- Some kind of group marker paper - with ongoing updates
- Within 6-9 months of assembly of species- have jamboree for white paper etc.
- Barcode release immediately - can also be given a DOI
- DOI assignment of data - can set an embargo: only 1 year sequence traces from WGS projects are to be deposited in a trace archive within one week of production.
- Assemblies are to be deposited in a database as soon as possible after the assembled sequence has met a set of quality evaluation criteria.
- The deposited data should be available for all to use without restriction.

Assembly and Annotation: Challenges and Solutions

THE ASSEMBLATHON

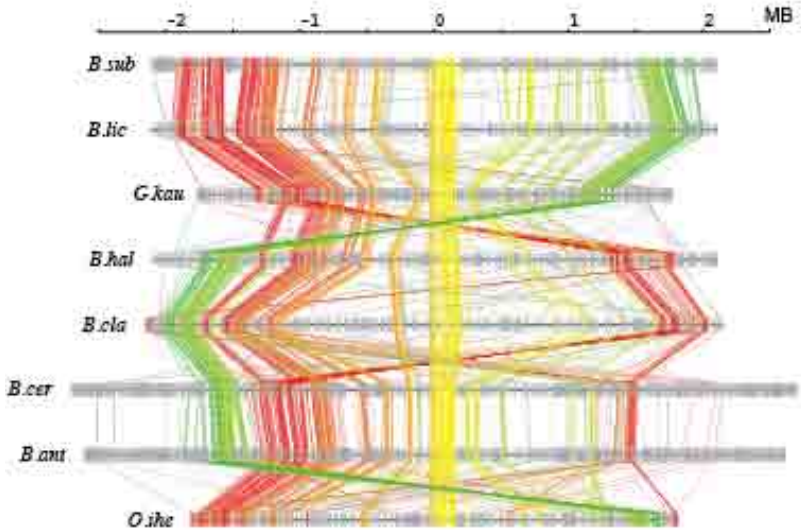
http://assemblathon.org/
Earl et al. 2011 *Genome Research*

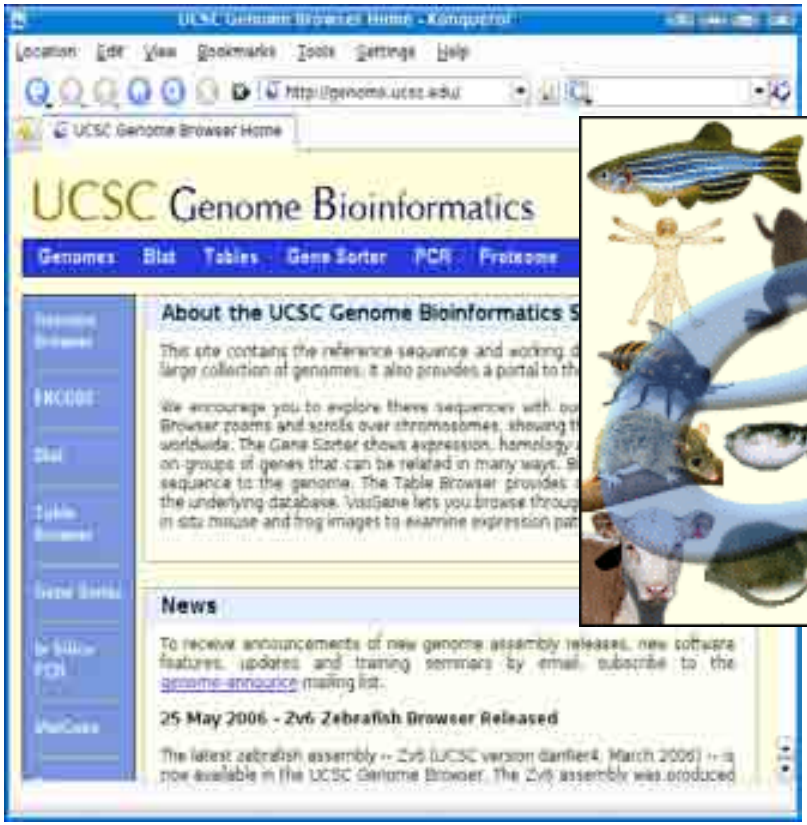
http://gage.cbcb.umd.edu/
Salzberg et al. 2012 *Genome Research*

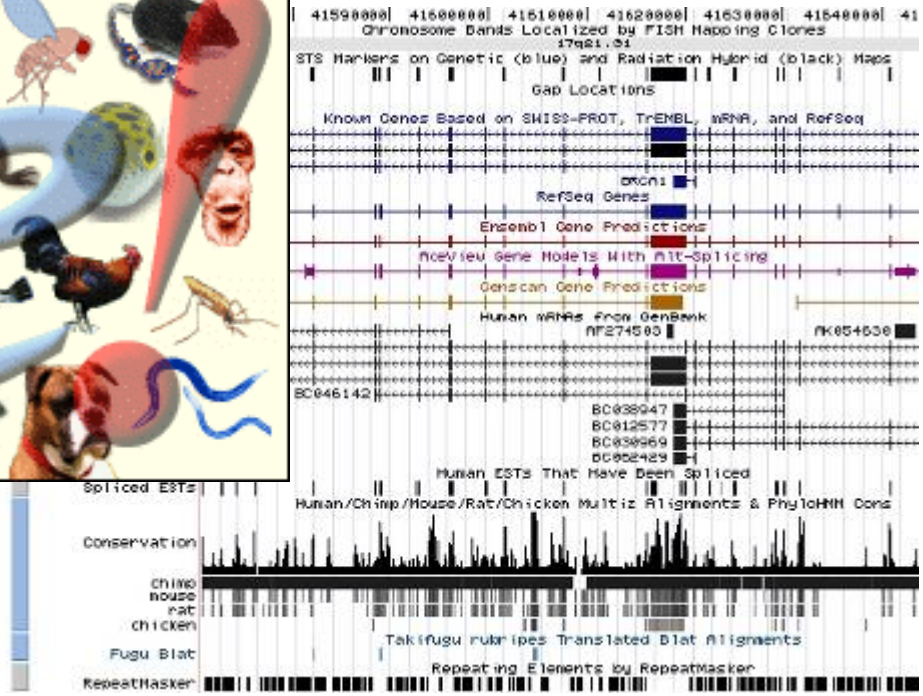
Data Access and Analysis

Alignathon

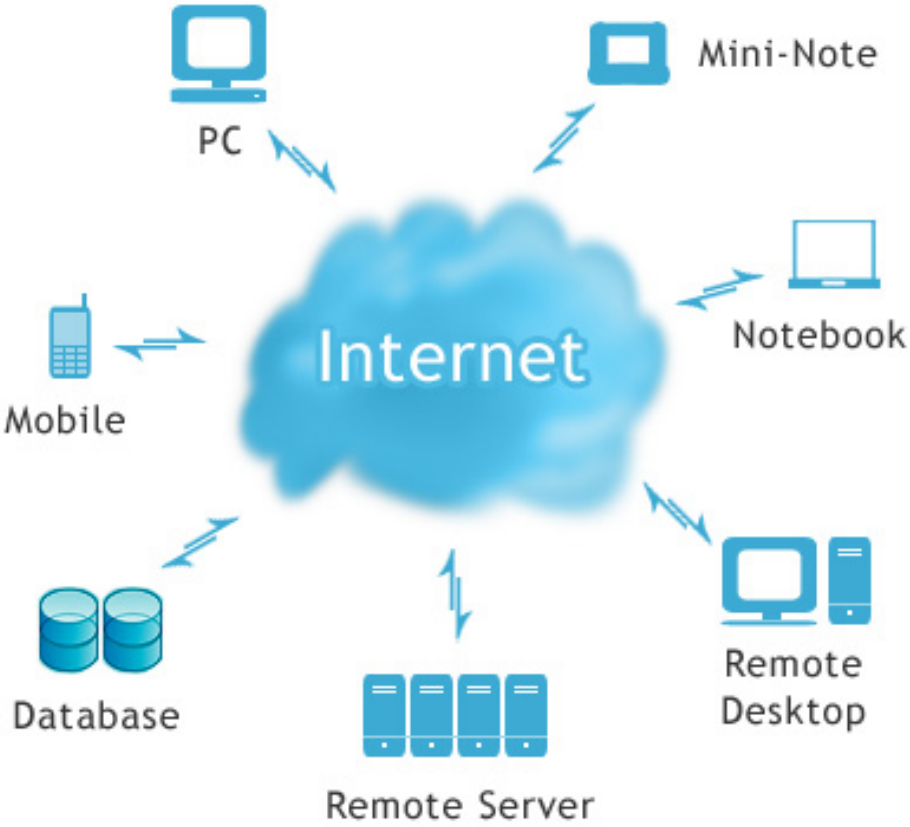
http://compbio.soe.ucsc.edu/alignathon/







Data Access and Release



http://www.gigasciencejournal.com/

‘BGI-BOX’
Cloud Computing Solution

BGI-G10K Carnivoran Genomes: Genomic Ambitions

- Phylogenomics
- Comparative Genomics
- Population Genomics
- Physiological/Adaptation Genomics



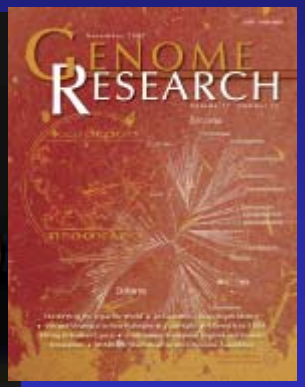

华大基因 BGI Premier Scientific Partner

GENOME 10K

The First Carnivoran Genomes

- **Dog (*Canis familiaris*) - 2003, 2005**

- **Cat (*Felis catus*) - 2007**



- **Giant Panda (*Ailuropuda melanoleuca*) - 2010**

The Next Carnivoran Genomes: The BGI-G10K species

-**Polar bear (*Ursus maritimus*)**

-**Cheetah (*Acinonyx jubatus*)**

-**Lion (*Panthera leo*)**

-**Red fox (*Vulpes vulpes*)**

-**Red panda (*Ailurus fulgens*)**

-**Spotted hyena (*Crocuta crocuta*)**

The Next Carnivoran Genomes: Other research groups

-**Domestic ferret (*Mustela putorius furo*)**

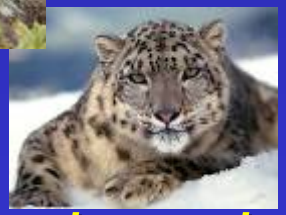
-**Wolverine (*Gulo gulo*)**

-**Weddell seal (*Leptonychotes wedellii*)**

-**Tiger (*Panthera tigris*)**

-**Iberian lynx (*Lynx pardinus*)**

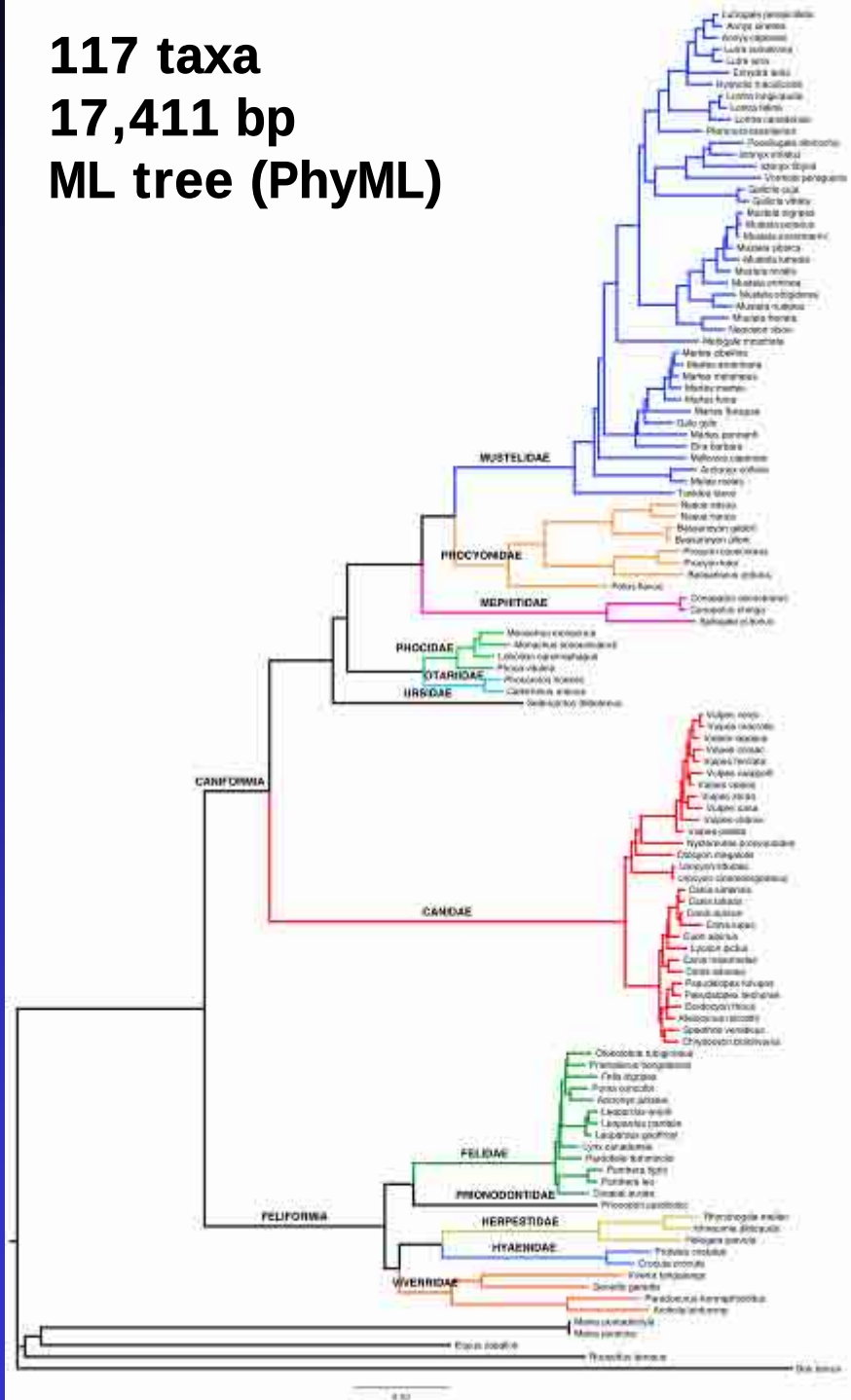
-**Snow leopard (*Panthera uncia*)**



-**Small Indian mongoose (*Herpestes javanicus*)**

The Carnivora Tree of Life

117 taxa
17,411 bp
ML tree (PhyML)



~250 species expected,
24 loci (~22,000 bp)
Koepfli et al.
unpublished data

The power of genomes: Identifying NEW markers effeciently

1147 primer pairs: Housley et al. 2006, *BMC Genomics*

↓

Genome assemblies of cat, dog, and giant panda

↓

Electronic PCR (*e*PCR)

http://www.ncbi.nlm.nih.gov/projects/e-pcr/

+ primer selection criteria

↓

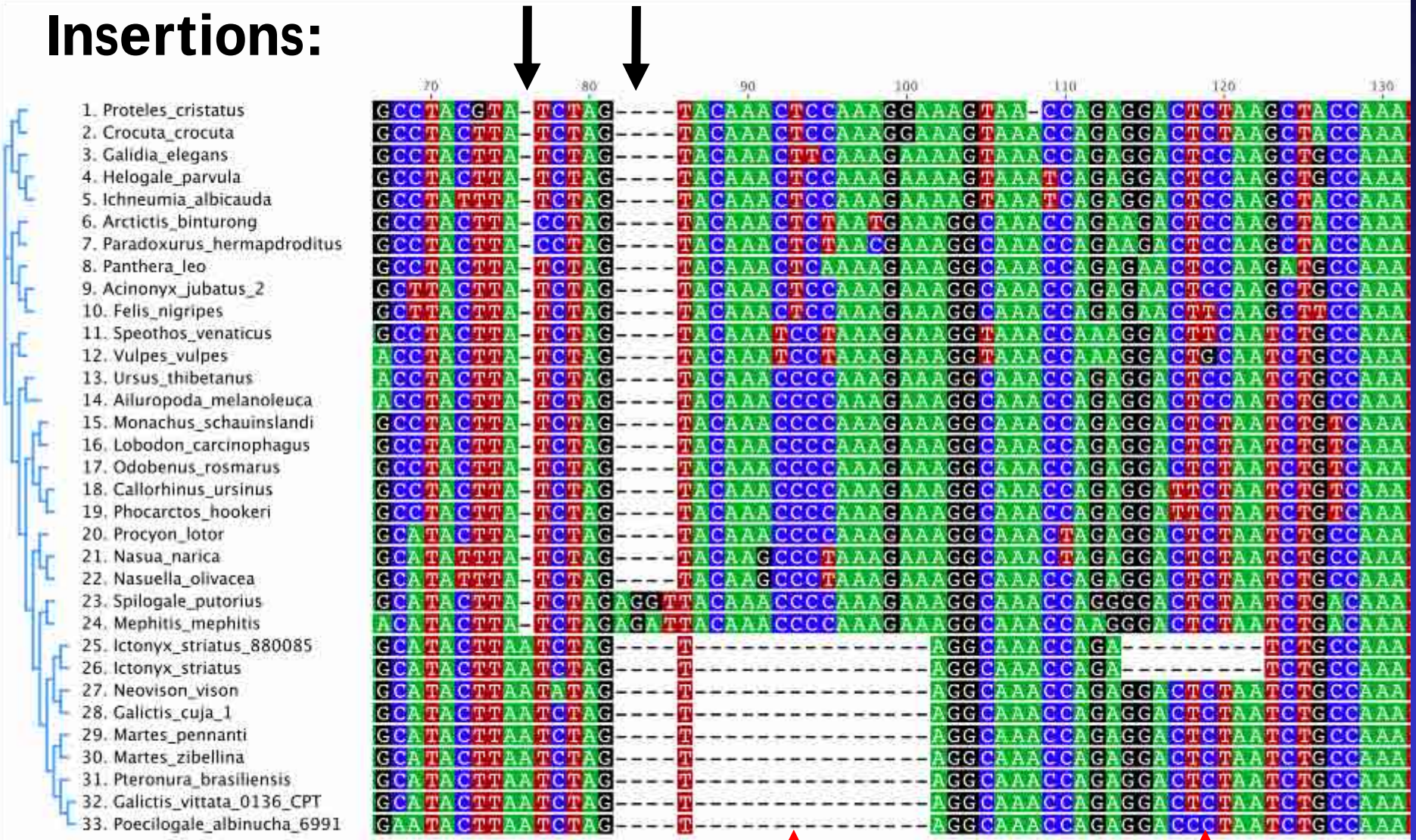
<u>140 primer pairs selected</u>

↓

Order 50 primer pairs for testing

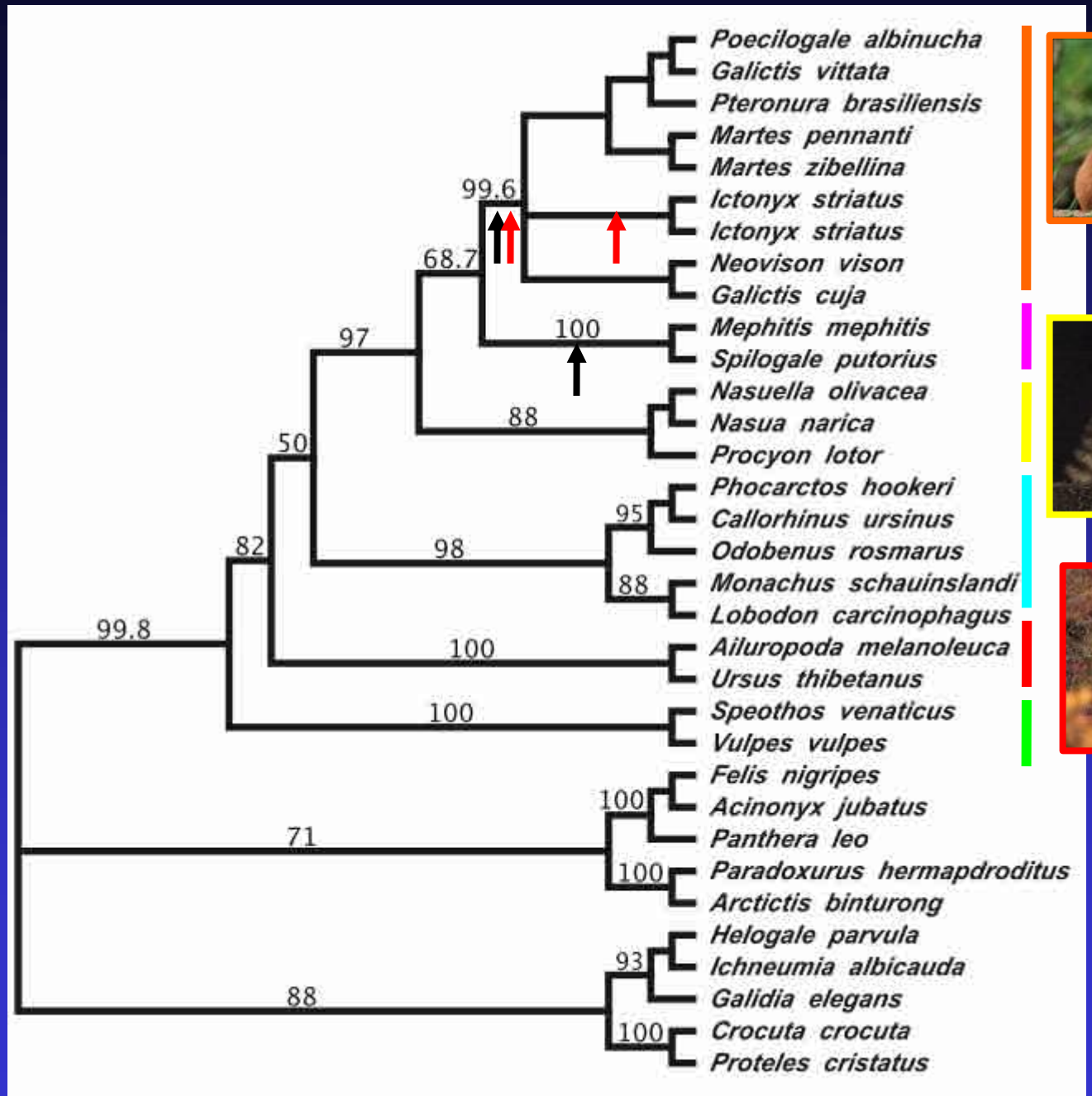
INDELs in *PFKFb1* intron as markers of carnivore phylogeny

Insertions:



Deletions:

Carnivore phylogeny based on *PFKFB1* intron

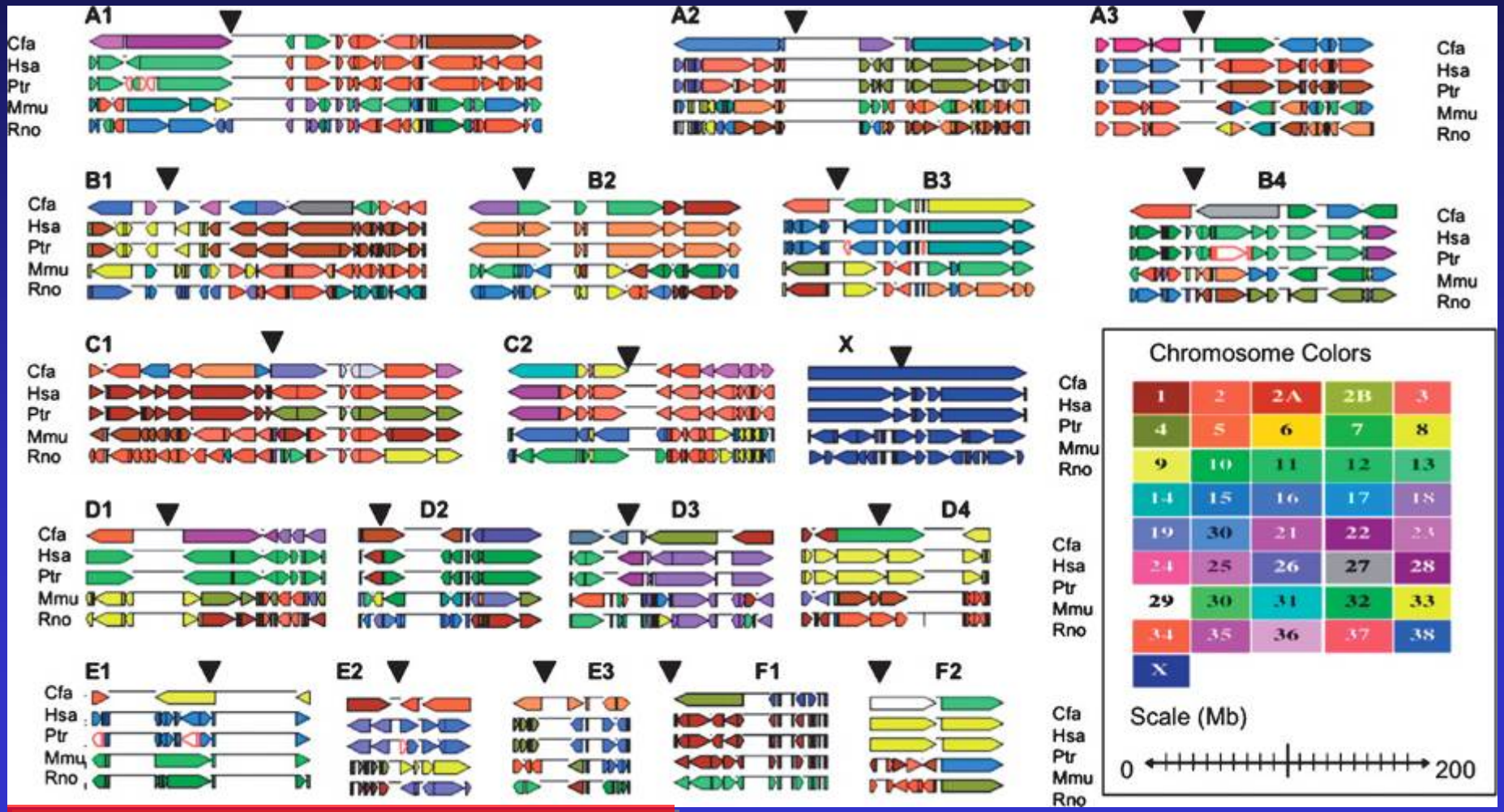


Other Phylogenomic Markers

- Insertions and deletions
- Intron gain and loss
- Retroposon integrations
- Signature sequences
- Gene duplications
- Genomic rearrangements
- Microinversions
- Ultra-conserved elements

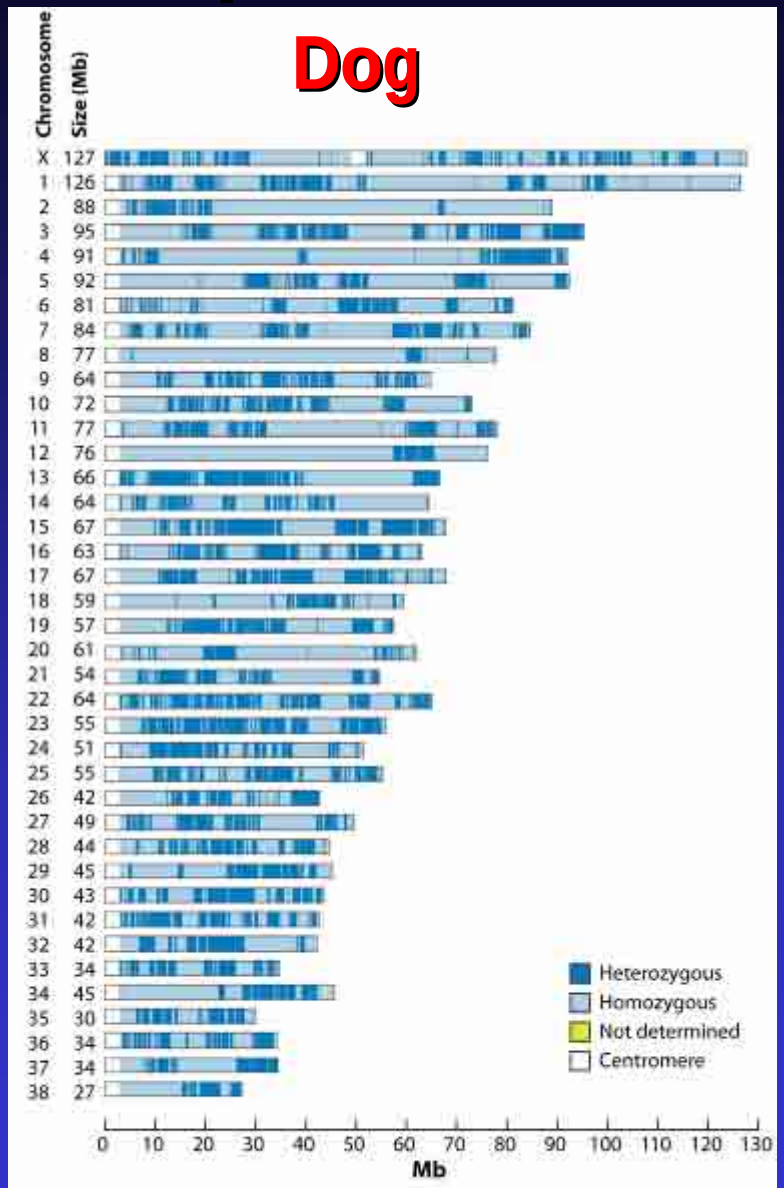
Comparative Genomics: Genome Architecture

Homologous synteny blocks (HSBs) of the cat genome as compared to corresponding syntenic blocks in five mammalian species: (Cfa) Canis familiaris, (Hsa) Homo sapiens, (Ptr) Pan troglodytes, (Mmu) Mus musculus, and (Rno) Rattus norvegicus.

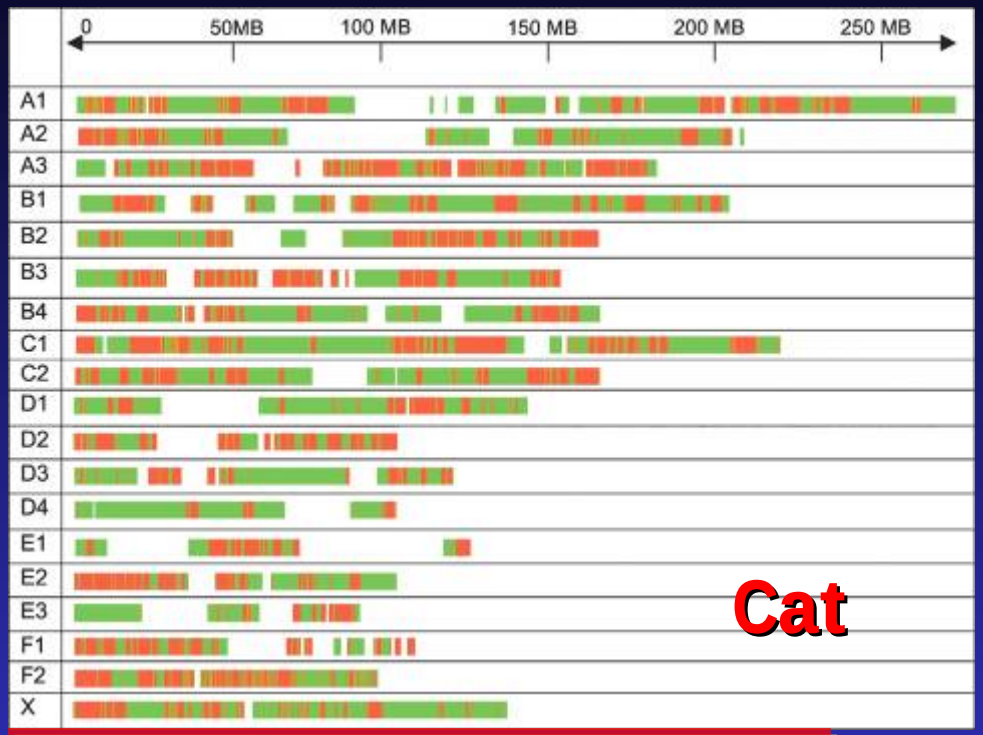


Pontius et al. *Genome Res*. 2007

Comparative Genomics: Genome Architecture

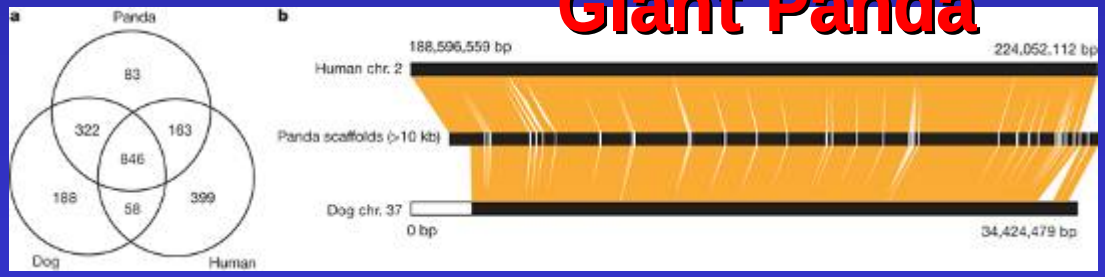


Parker HG, et al. 2010.
Annu. Rev. Genet. 44:309–36



Pontius et al. *Genome Res*. **2007**

Giant Panda



Li *et al*. *Nature* 2009

Population Genomics

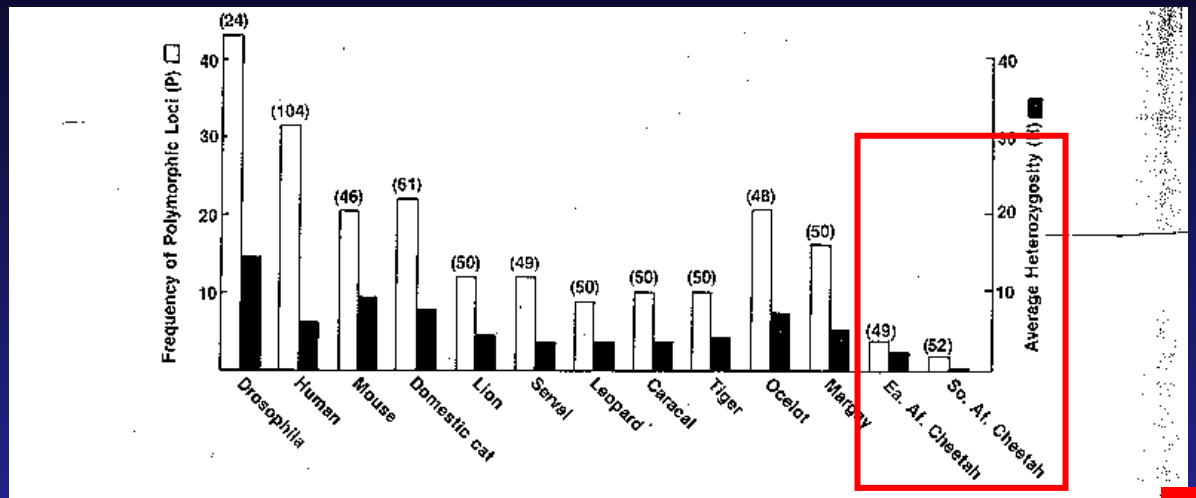
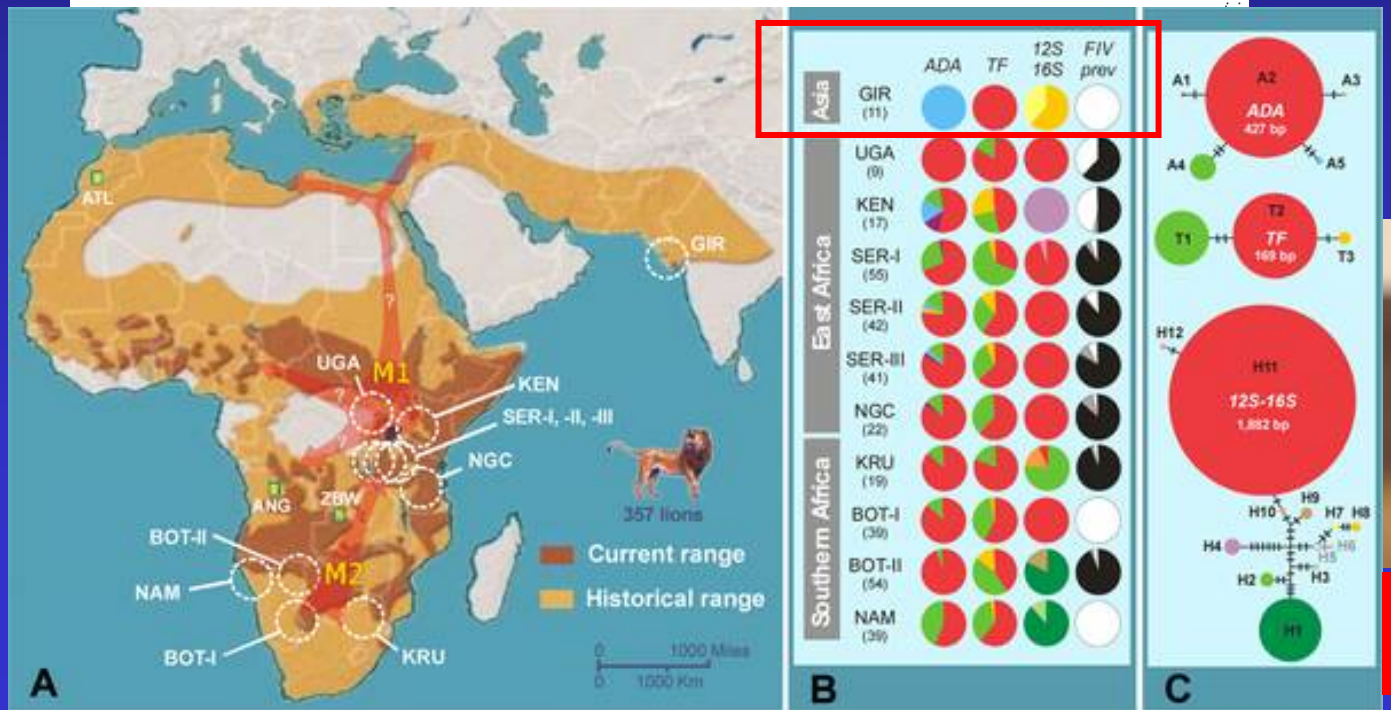


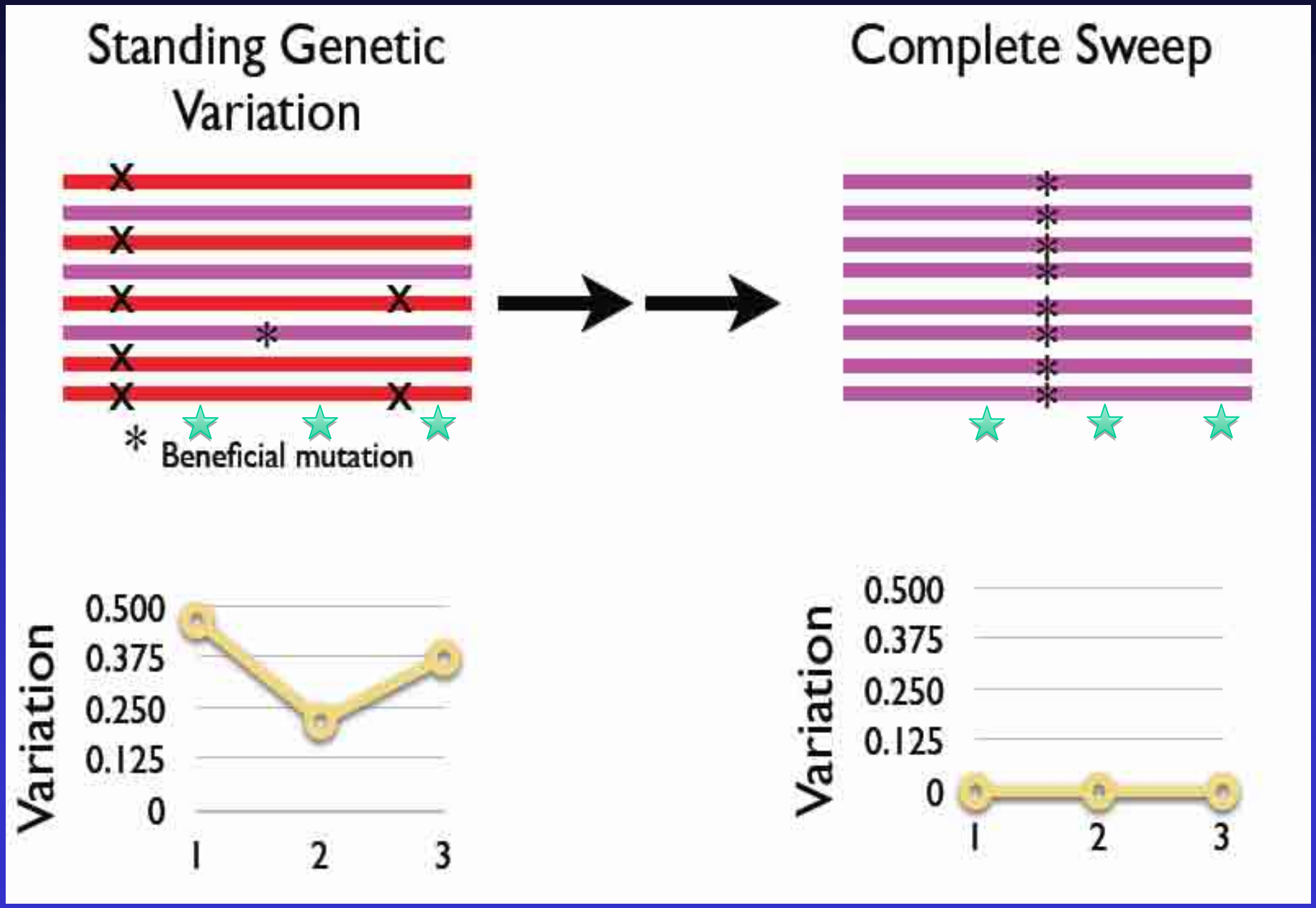
FIG. 1. Estimates of the extent of genetic variation based on allozyme electrophoretic surveys in the east African cheetah (*A. jubatus raineyi*), the south African cheetah (*A. jubatus jubatus*), eight additional feline species (8, 15), and three nonfelid species. The number in parentheses indicates the number of genetic loci that were considered in the estimate. For a review of over 250 such electrophoretic surveys see ref. 24. Enzymes typed and procedures used are as previously described (7, 8, 15).

O'Brien *et al. PNAS* 1987

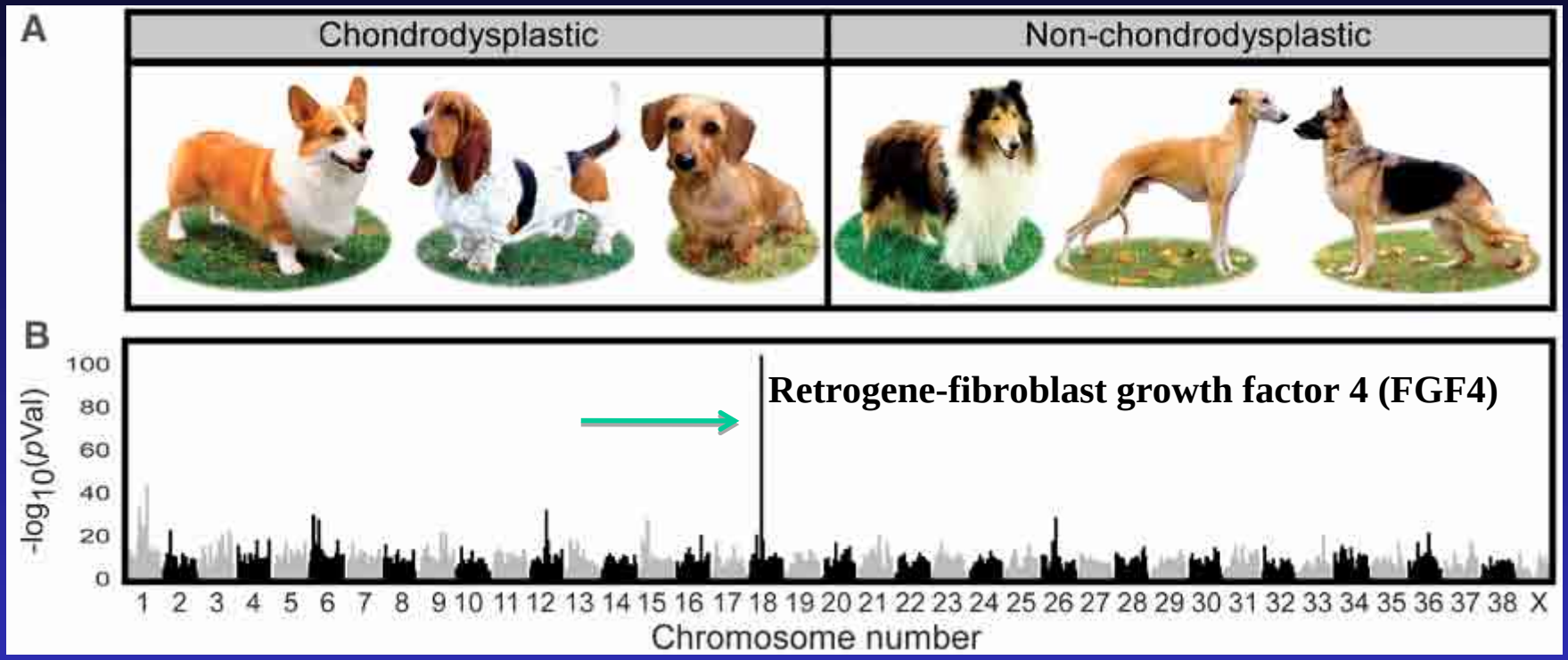


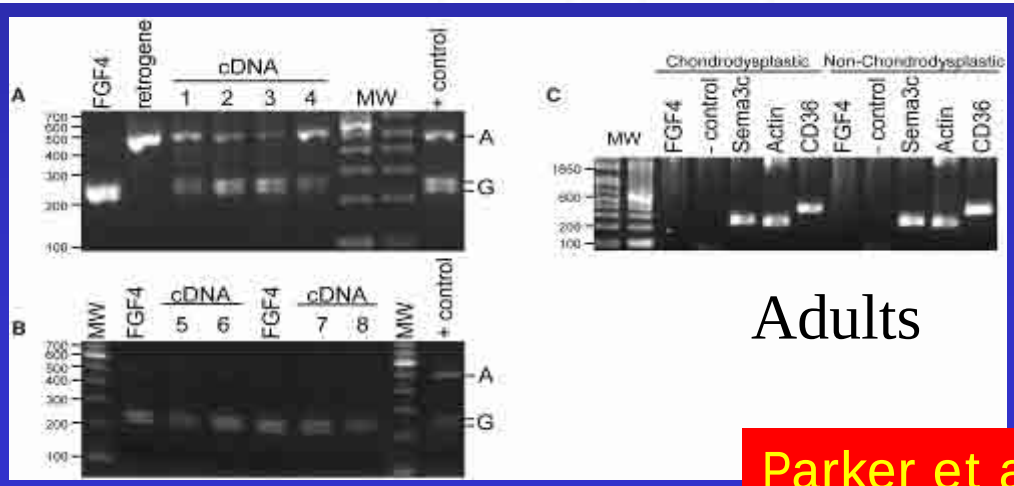
Antunes *et al.* *PLoS Genet.* 2007

Selective Sweep Mapping



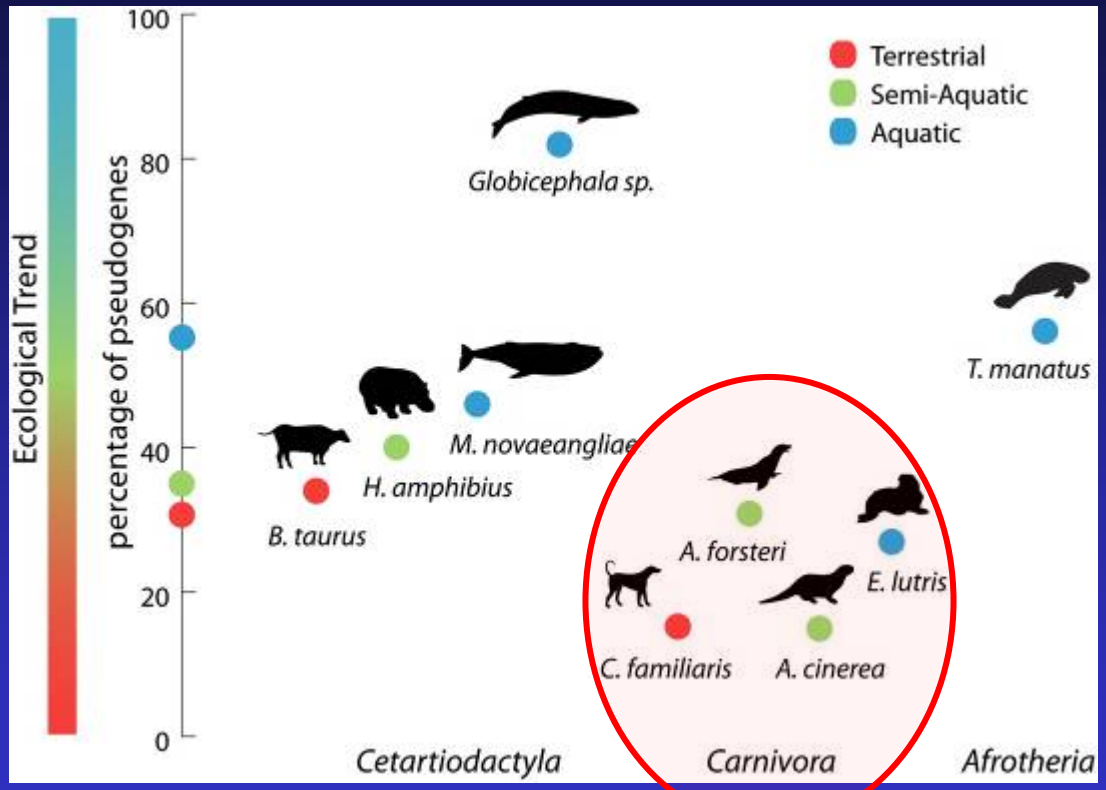
The short legs gene: association mapping

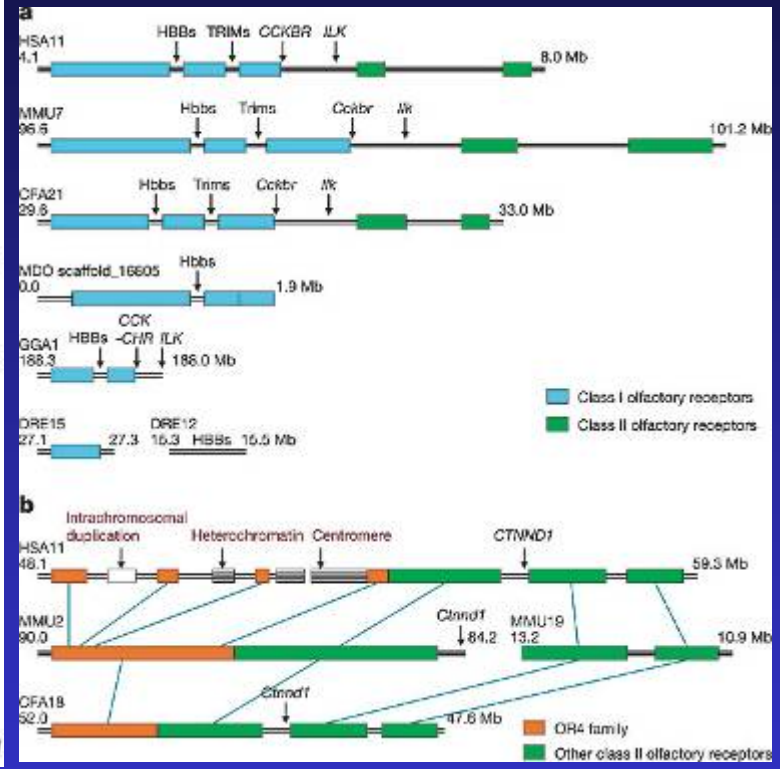




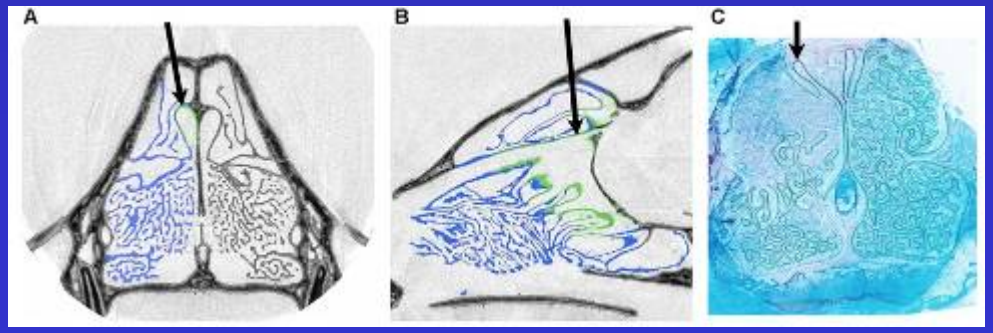
Parker et al. *Science*, 2009

Physiological/Adaptation Genomics: Olfactory Receptor (OR) Genes!





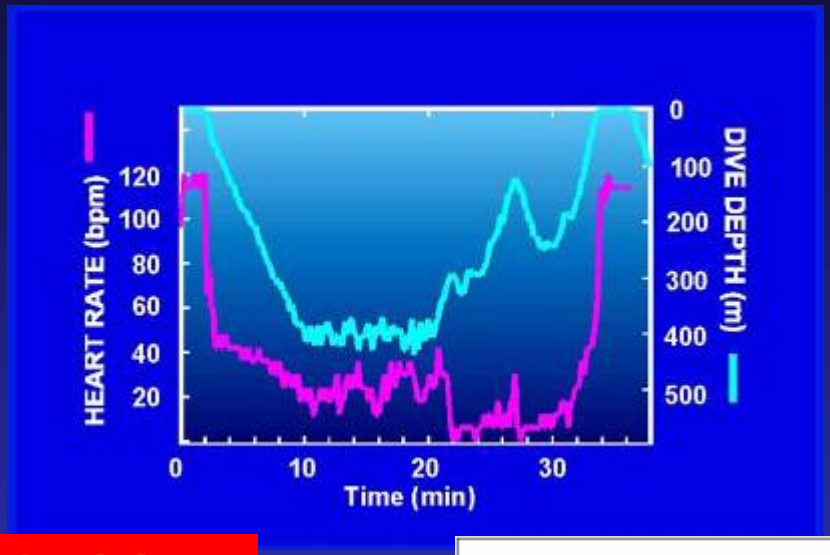
Hayden *et al. Genome Res.* 2010



Physiological/Adaptation Genomics: Hibernation (denning) in bears

- Decrease in metabolism 75%
- Body temperature is maintained
- No loss in bone mass or density
- No loss of muscle mass
  - Inhibition of the ubiquitin-proteasome system

The Genomic Future of Carnivora: – many possible avenues for study…



Northern elephant seal: diving physiology

Sea otter: osmoregulatory physiology

Skunks: chemical ecology

Thank You!

Genome 10K

Steve O'Brien

Warren Johnson

Oliver Ryder

David Haussler

G10KCOS

BGI

Guojie Zhang

Yingrui Li

Jun Wang

BGI team

GENOME 10K

华大基因 BGI Premier Scientific Partner