LBNL-57628

Why Evolutionary Biology and Genome Sciences Need Each Other

Jeffrey Boore
Evolutionary Genomics Department Head

PARC - May 5, 2005

Overview

- The genome - What is it and what is it good for?
- The transition from sequencing the human to comparing genomes
- Stories at the interface of evolutionary biology and genomics

What is a genome?

Collective term for the complete DNA sequence of an organism

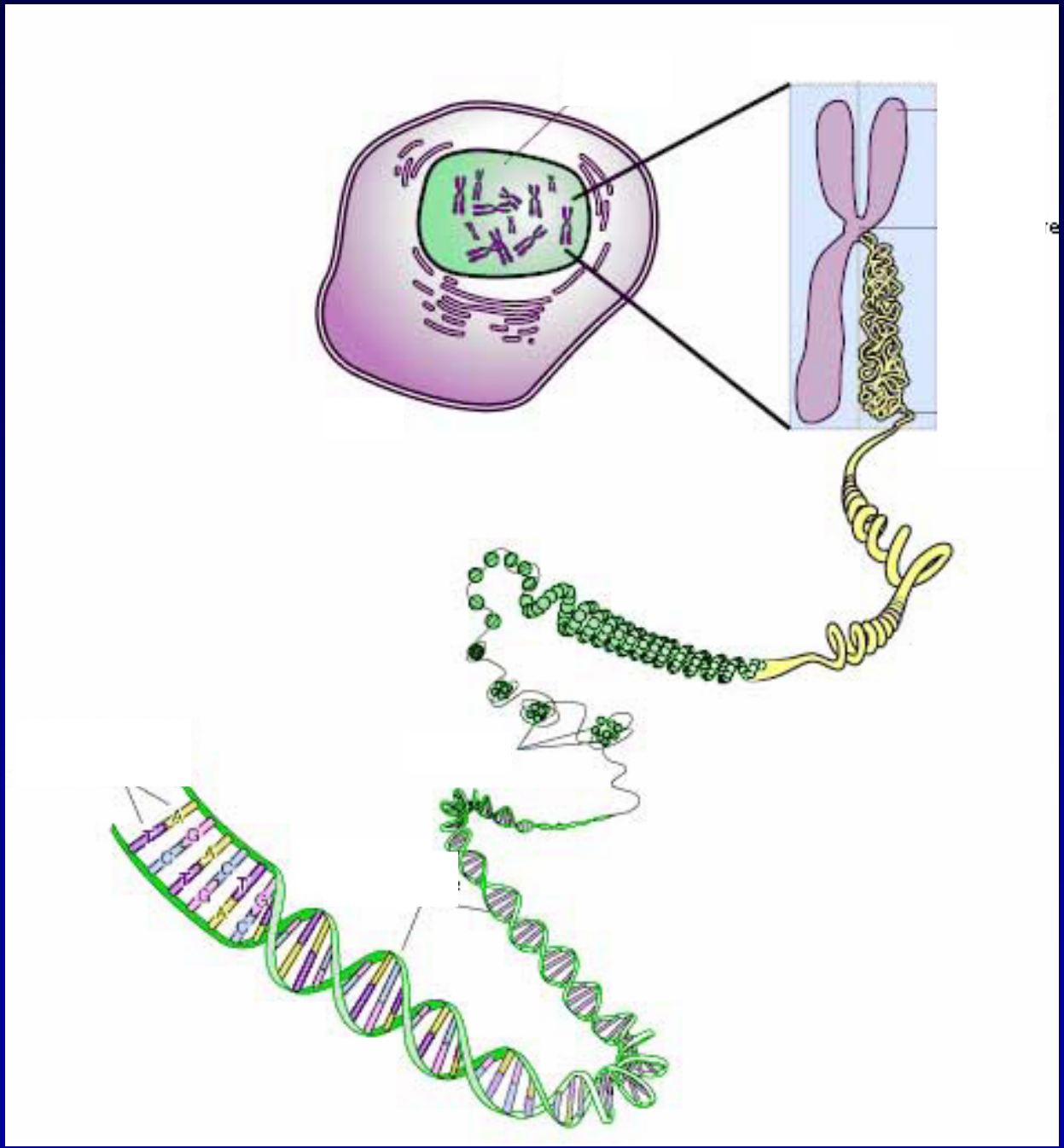
The identical genome sequence is present in each cell of the body

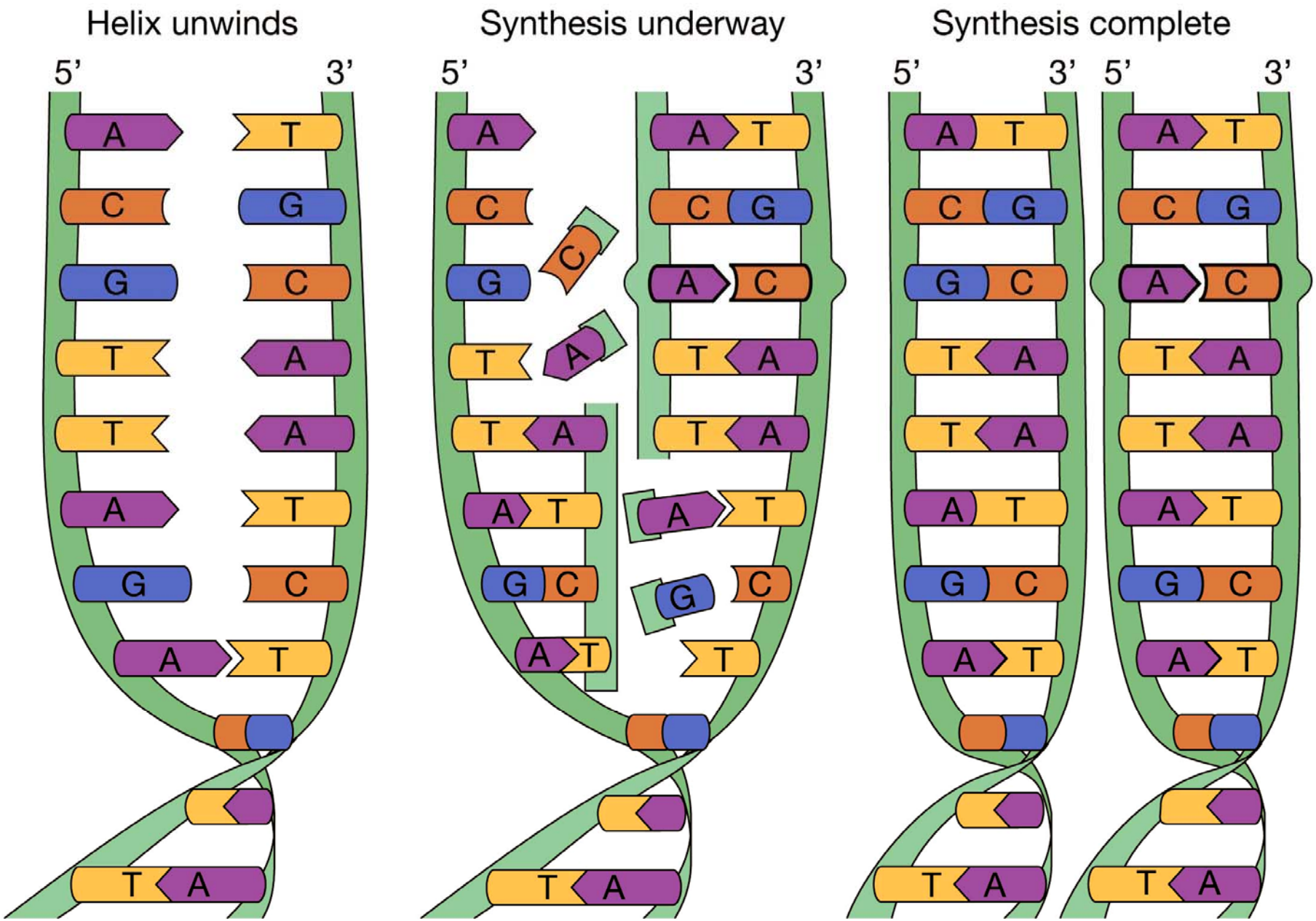
What is DNA sequencing?

DNA is a polymer of four chemicals: A, G, C, T.

There are two strands that are “complementary”, A and T pair and G and C pair. Knowing the sequence of bases in one strand tells us the sequence of the other.

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





Copyright © 2004 Pearson Prentice Hall, Inc.

...T A C T A G T T T G A A G G T T T C A...

A T G A T C A →

DNA sequence

Anneal primer and extend with polymerase+dNTPs

BUT -- include some ddNTPs, which terminate the growing chain, each with a different florescent label

...A T G A T C A

...A T G A T C A **A**

...A T G A T C A **A A**

...A T G A T C A **A A C**

...A T G A T C A **A A C T**

...A T G A T C A **A A C T T C**

...A T G A T C A **A A C T T C**

...A T G A T C A **A A C T T C C**

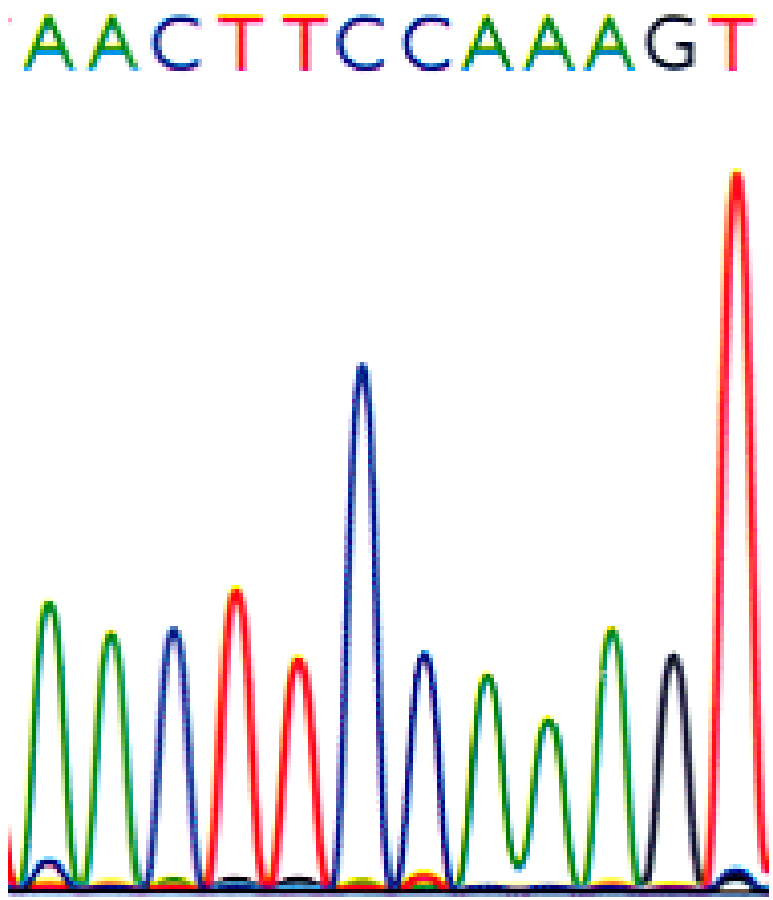
...A T G A T C A **A A C T T C C A**

...A T G A T C A **A A C T T C C A A**

...A T G A T C A **A A C T T C C A A A**

...A T G A T C A **A A C T T C C A A A G**

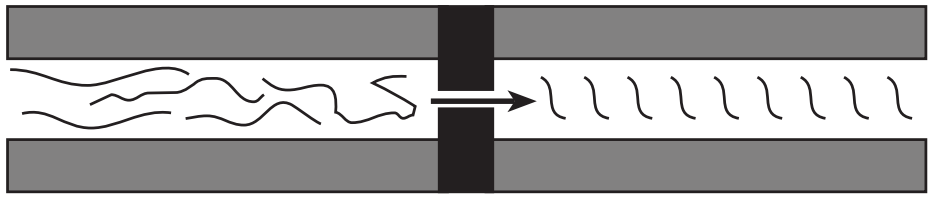
```
...A T G A T C A
...A T G A T C A A
...A T G A T C A A A
...A T G A T C A A A C
...A T G A T C A A A C T
...A T G A T C A A A C T T C
...A T G A T C A A A C T T C
...A T G A T C A A A C T T C C
...A T G A T C A A A C T T C C A
...A T G A T C A A A C T T C C A A
...A T G A T C A A A C T T C C A A A
...A T G A T C A A A C T T C C A A A G
```



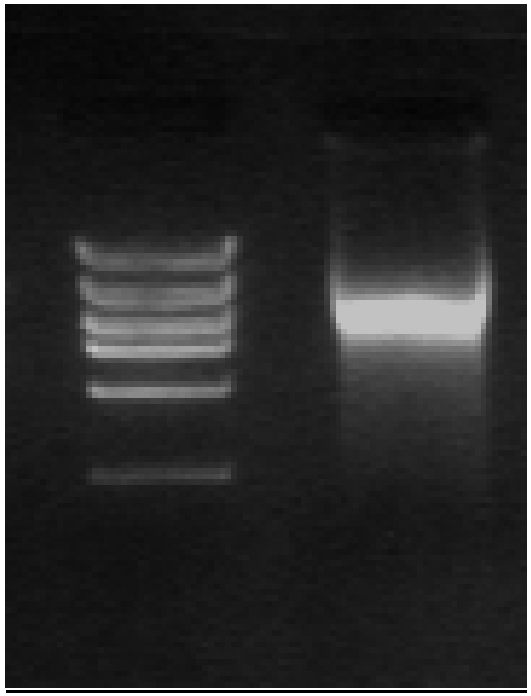
Separate by passing through a matrix driven by an electrical field, detect the florescent color of each fragment

Limitation is the ability of the matrix to separate with one nucleotide resolution as the fragments get longer

A

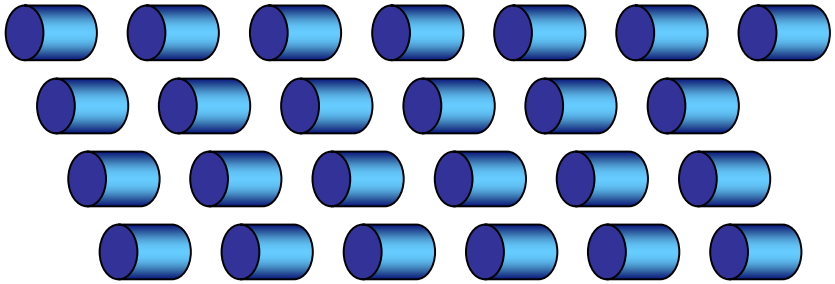


B

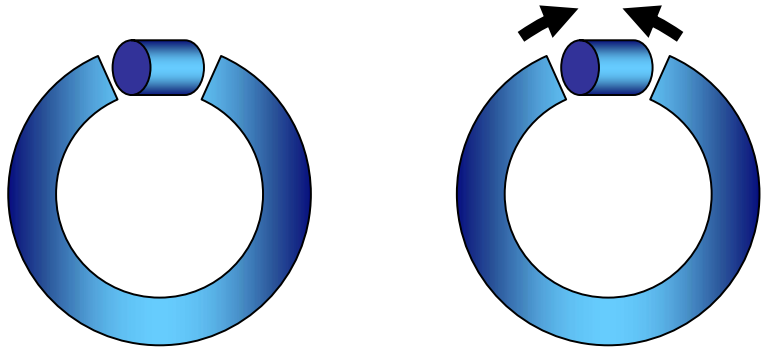


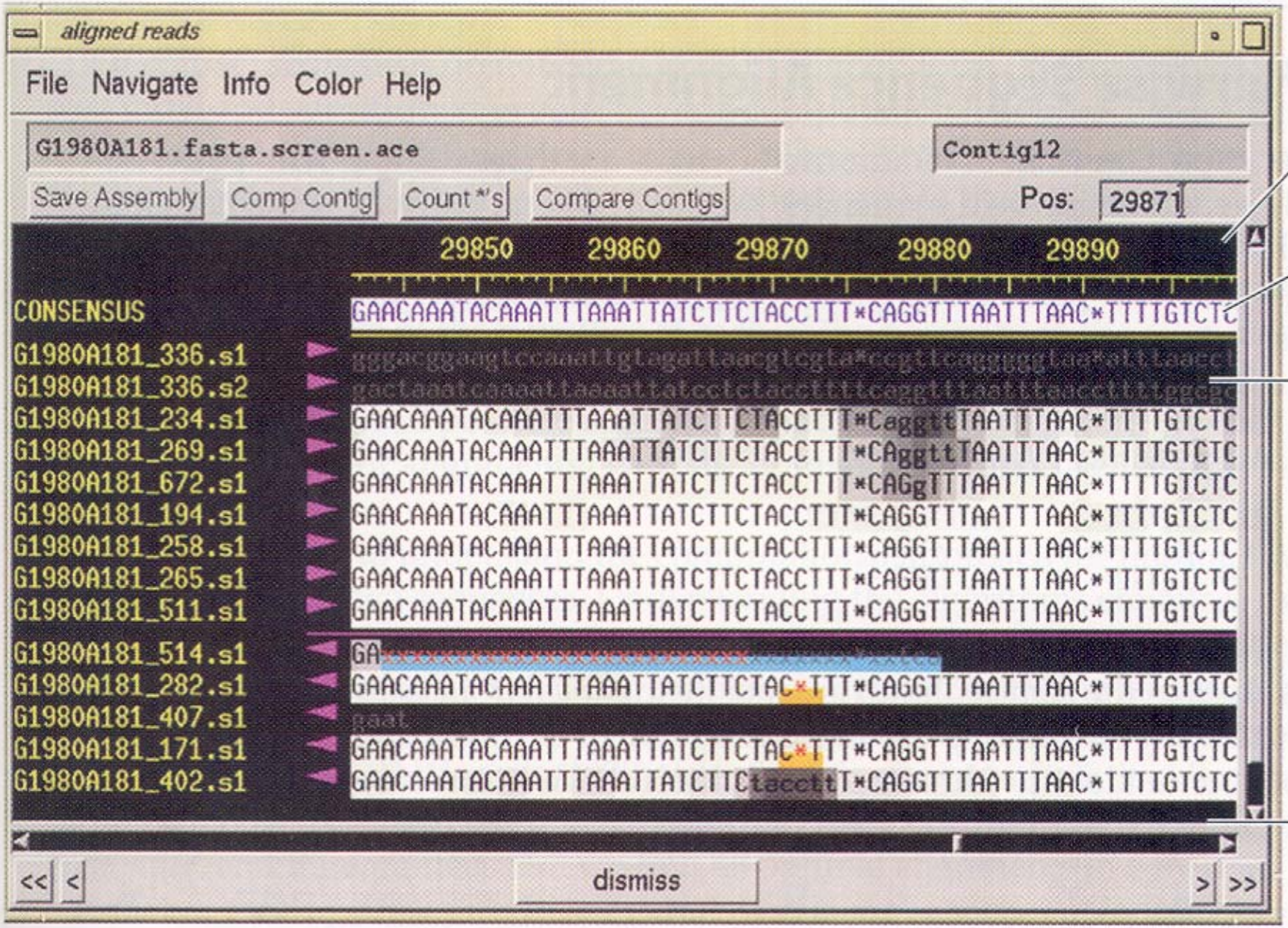
Large DNA fragment

Sheared randomly into small pieces



Clone, sequence from each end




aligned reads
File Navigate Info Color Help
G1980A181.fasta.screen.ace
Contig12
Save Assembly
Comp Contig
Count *'s
Compare Contigs
Pos: 29871
29850 29860 29870 29880 29890
CONSENSUS
G1980A181_336.s1
G1980A181_336.s2
G1980A181_234.s1
G1980A181_269.s1
G1980A181_672.s1
G1980A181_194.s1
G1980A181_258.s1
G1980A181_265.s1
G1980A181_511.s1
G1980A181_514.s1
G1980A181_282.s1
G1980A181_407.s1
G1980A181_171.s1
G1980A181_402.s1
dismiss
Position on contig
Consensus sequence
Independent sequence reads

Incredible increase in rate of DNA sequencing

Major genome centers built for HGP:

1. **Joint Genome Institute**

   Walnut Creek, California

2. **Sanger Centre**

   Hinxton, England

3. **Whitehead Institute**

   Boston, Massachusetts

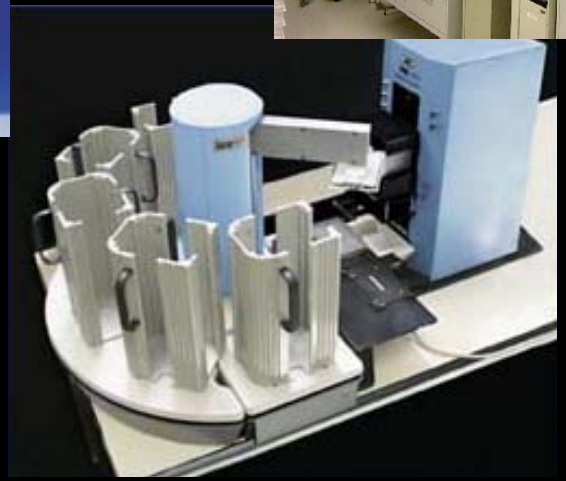
4. **Washington University**

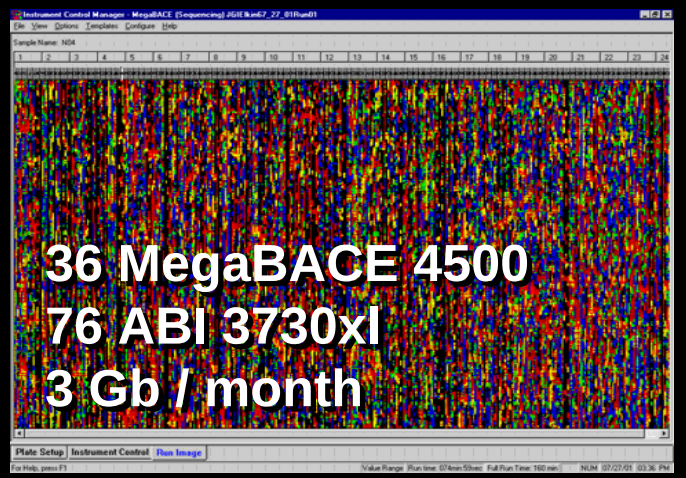
   St. Louis, Missouri

5. **Baylor University**

   Houston, Texas

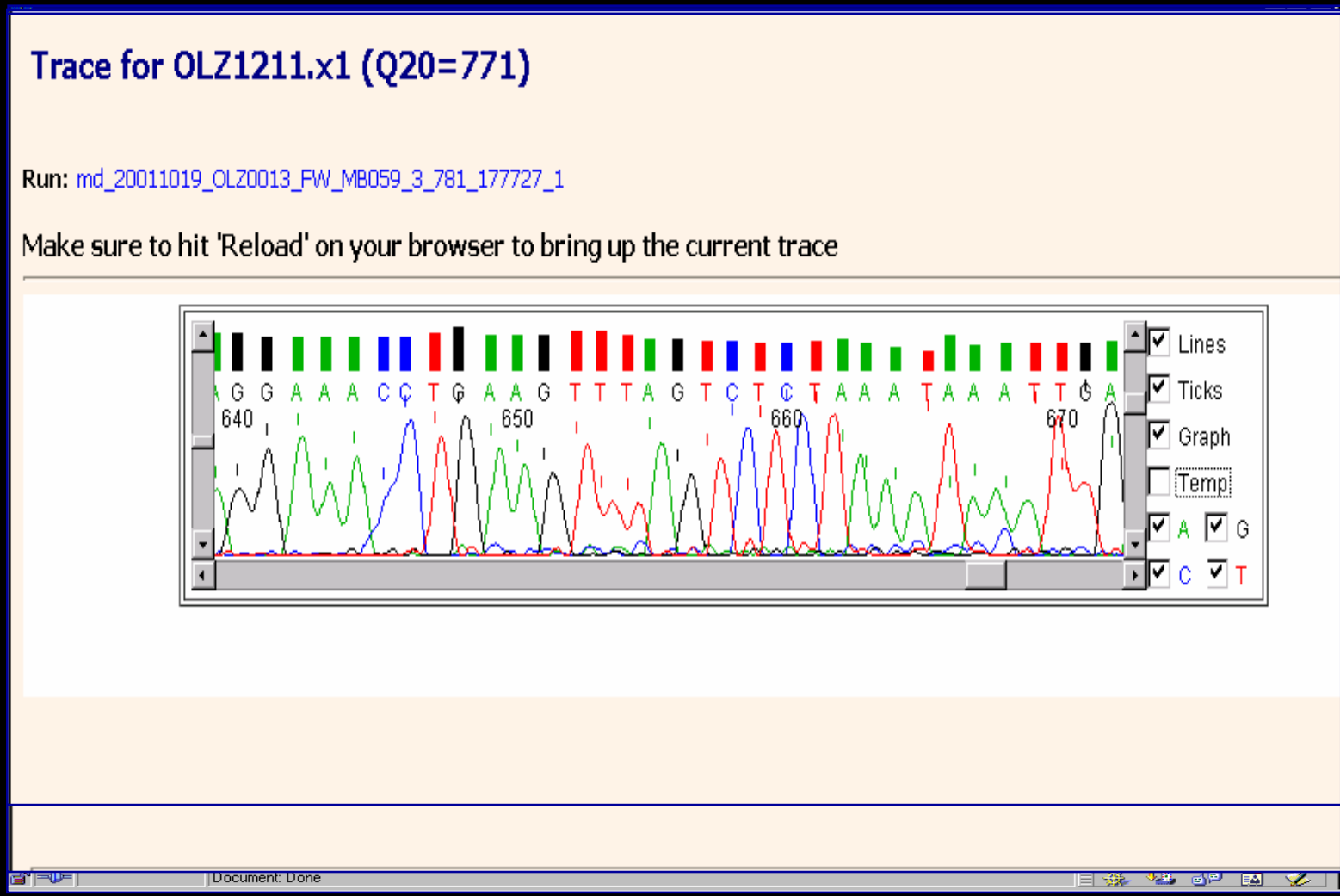
GENEMACHINES
HydroShear



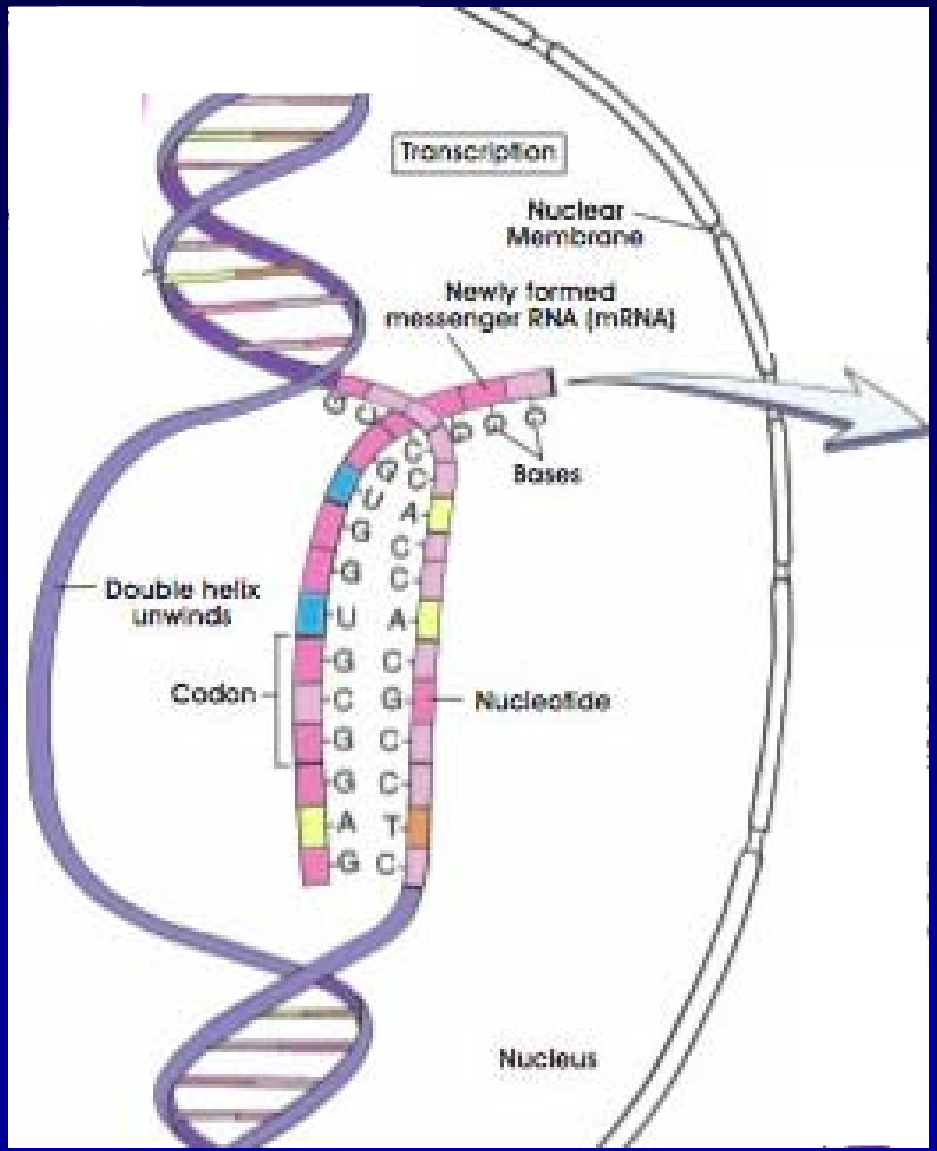

36 MegaBACE 4500
76 ABI 3730xl
3 Gb / month

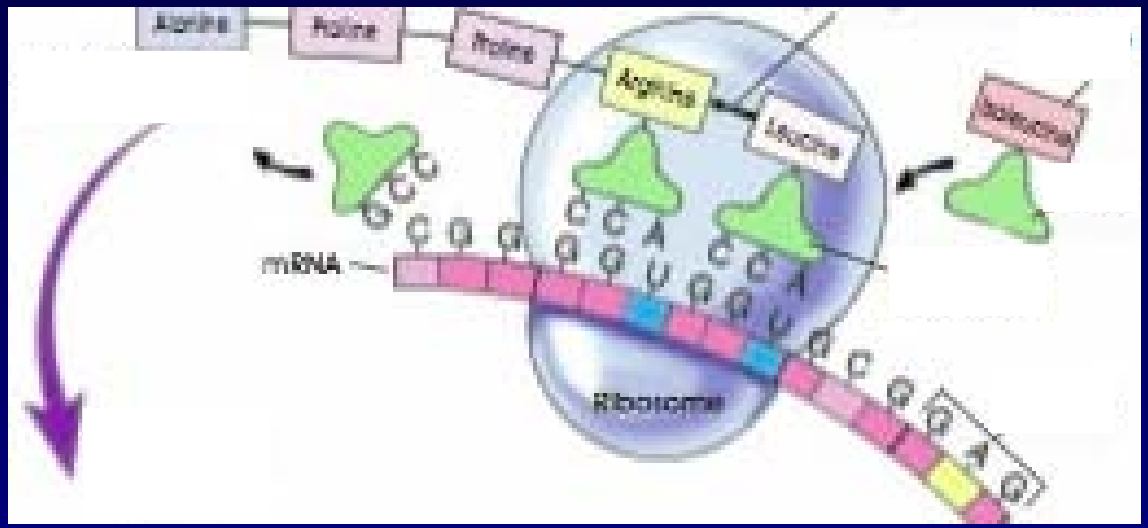
Online tracking of progress

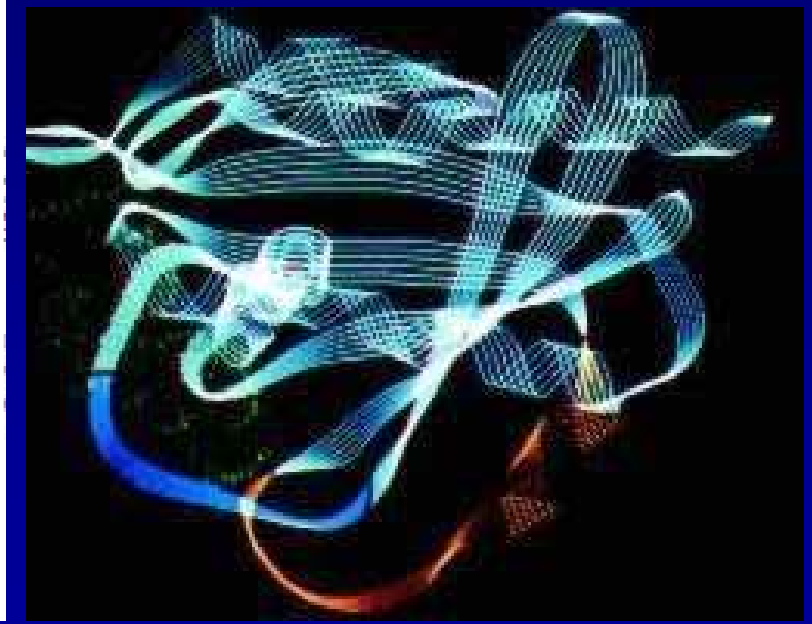
LIMS uses bar code readers at every step and allows real time tracking of all reagents, personnel, and processes



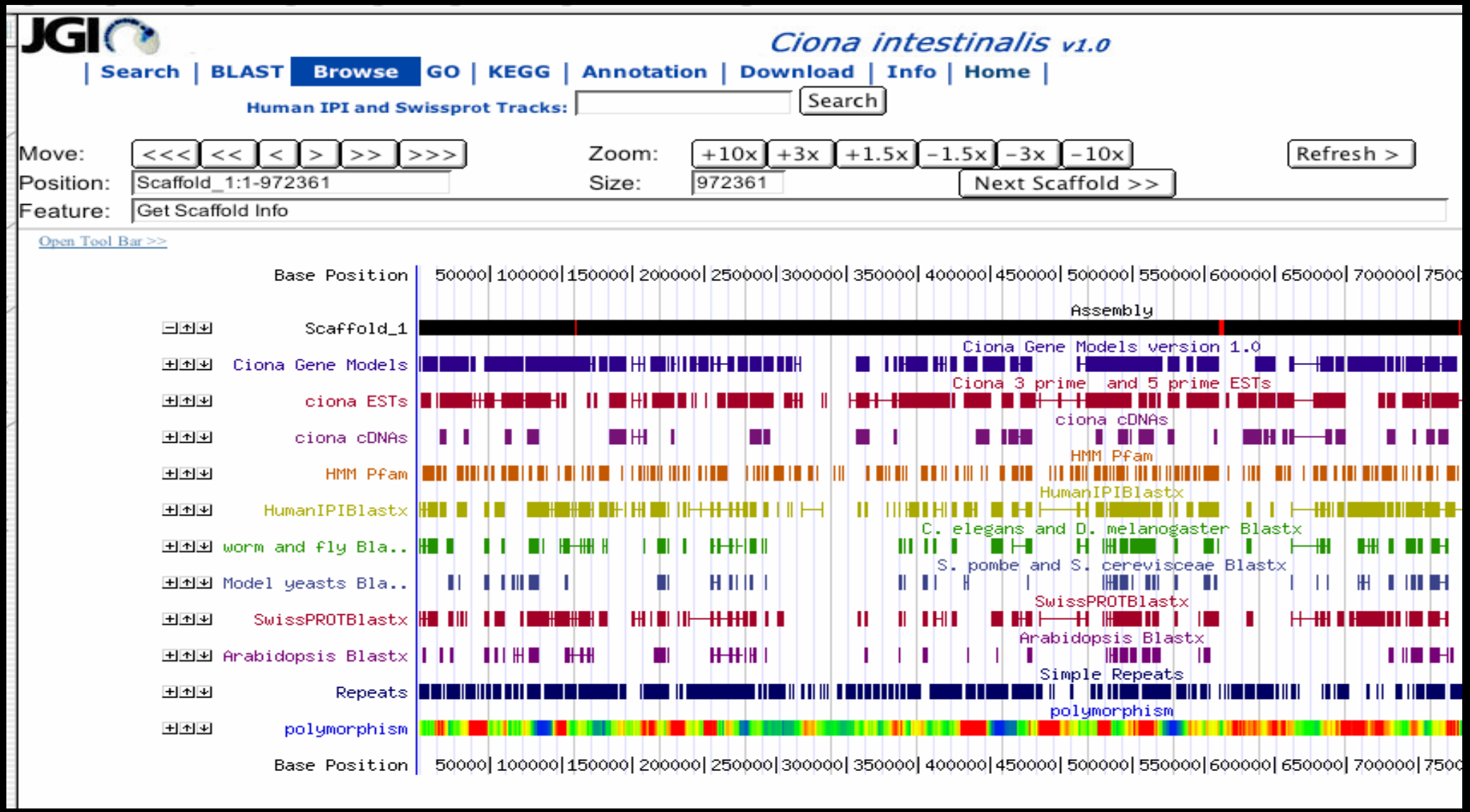
What is a gene?







How do we identify genes?
Ab initio, comparisons to ESTs, comparisons to genes of other organisms



What do we mean by “a gene for . . .”?

The difference between a locus and an allele

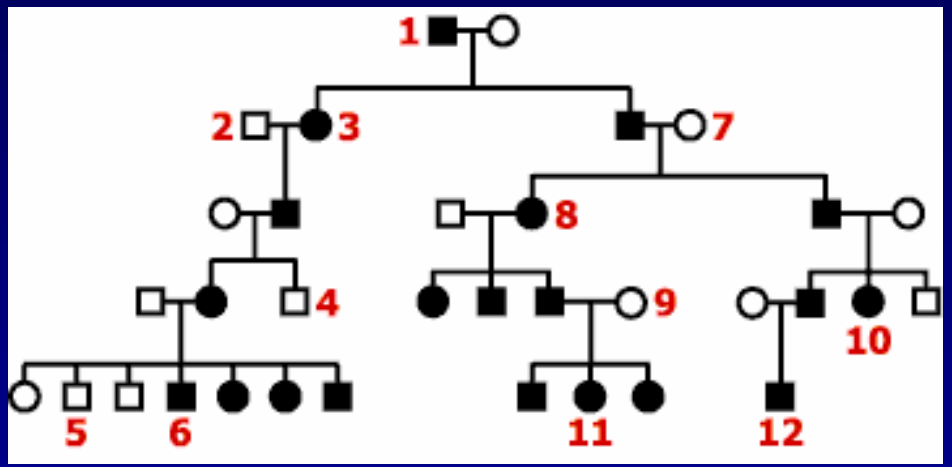
One from mom and one from dad

Recessive mutations are more common than dominant, because there are more ways to break than to adopt a new function

Why is having the human genome sequence cool?

One example - The candidate gene approach

Method prior to HGS



Pedigree including affected individuals
Find physical “markers” in the genome, track co-segregation
Sequence DNA clones corresponding to markers
“Walk” out from this sequence, find additional markers
Repeat MANY times until a candidate gene is found

Method after the HGS

Identify a set of candidate genes
See if they vary between affected and unaffected people (not necessarily related)

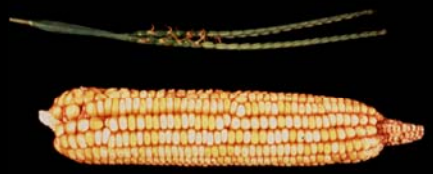
Potentially reduces decades of work to months

The transition of the centers to comparative genomics

Why sequence other genomes?

Defeat disease and parasites

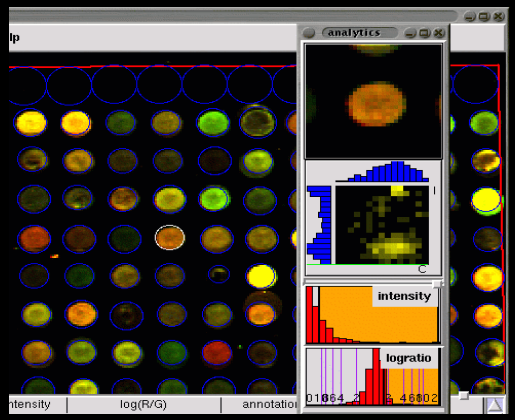
Modify organisms



Create “designer” microbes

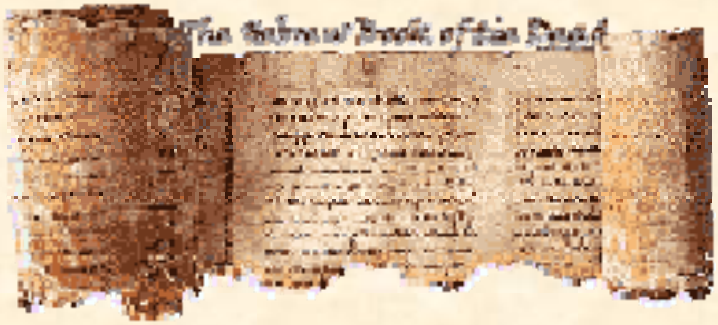
Understand basic biological processes

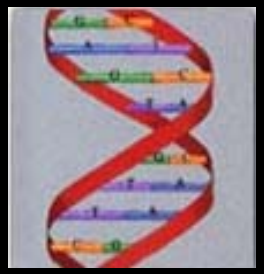
By providing reagents for functional genomics

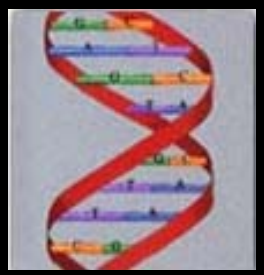


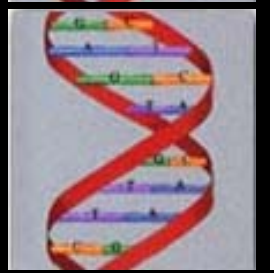
By enabling candidate gene approaches

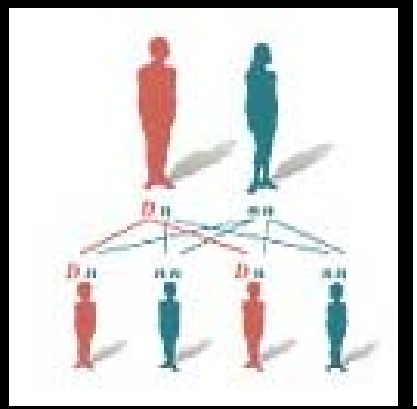
Correlate novel genomic features with novel traits (morphological, physiological, behavioral, etc.) to identify candidate genes

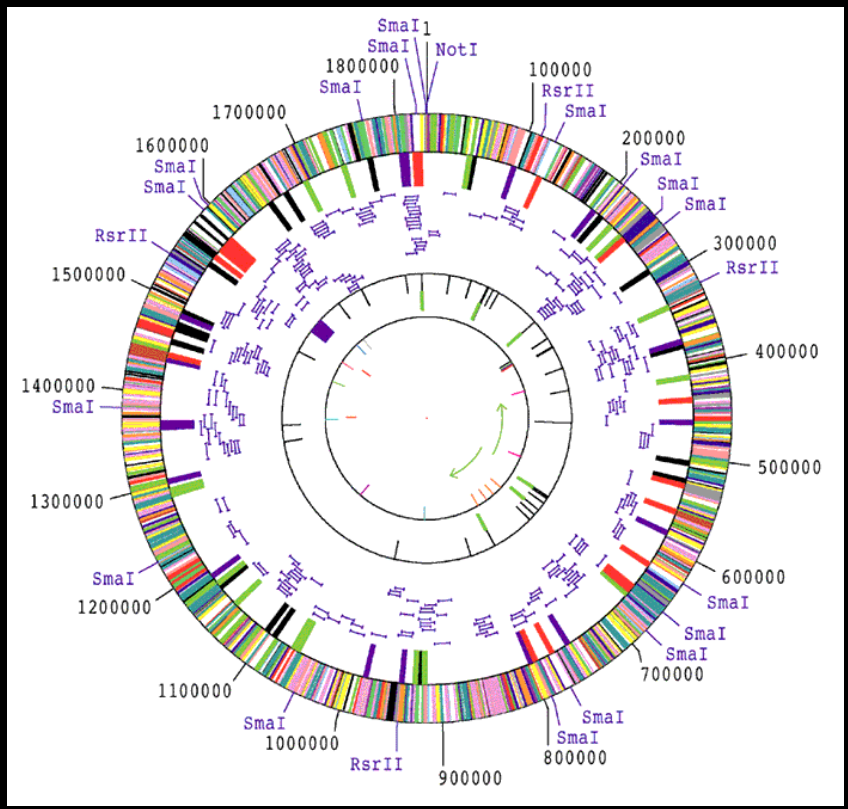




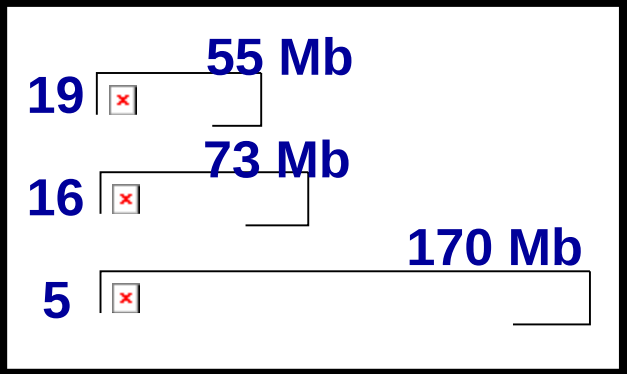




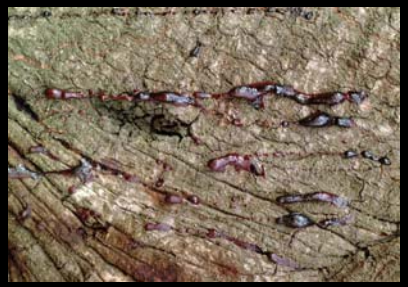



SmaI
SmaI
NotI
1
100000
RsrII
SmaI
200000
SmaI
SmaI
SmaI
300000
RsrII
400000
500000
600000
SmaI
SmaI
SmaI
700000
SmaI
SmaI
800000
900000
RsrII
1000000
SmaI
1100000
SmaI
1200000
1300000
SmaI
1400000
SmaI
1500000
RsrII
1600000
SmaI
SmaI
1700000
1800000
SmaI

JGI sequencing projects to date



Homo sapiens
(3 chromosomes)



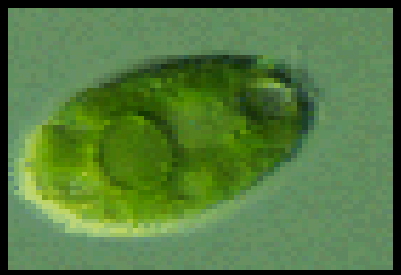
Phytophthora sojae,
P. ramorum

Ciona

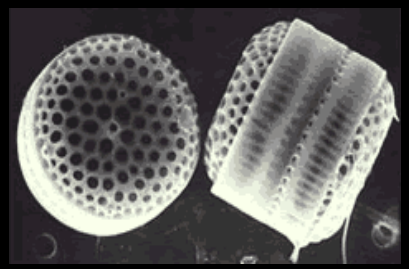
Fugu

Populus

Mus
(1 chromosome)

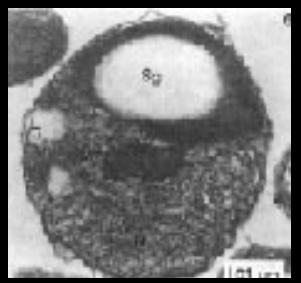


Chlamydomonas



Thalassiosira

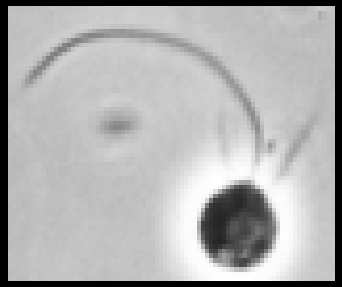
Phanerochaete



Ostreococcus

Xenopus

Some JGI sequencing projects underway



Monosiga

Reniera

Nematostella

Capitella

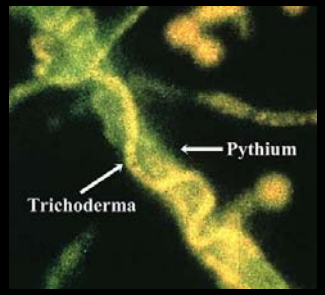
Helobdella

Lottia

Daphnia

Branchiostoma

Emiliania



Trichoderma

Phaeodactylum

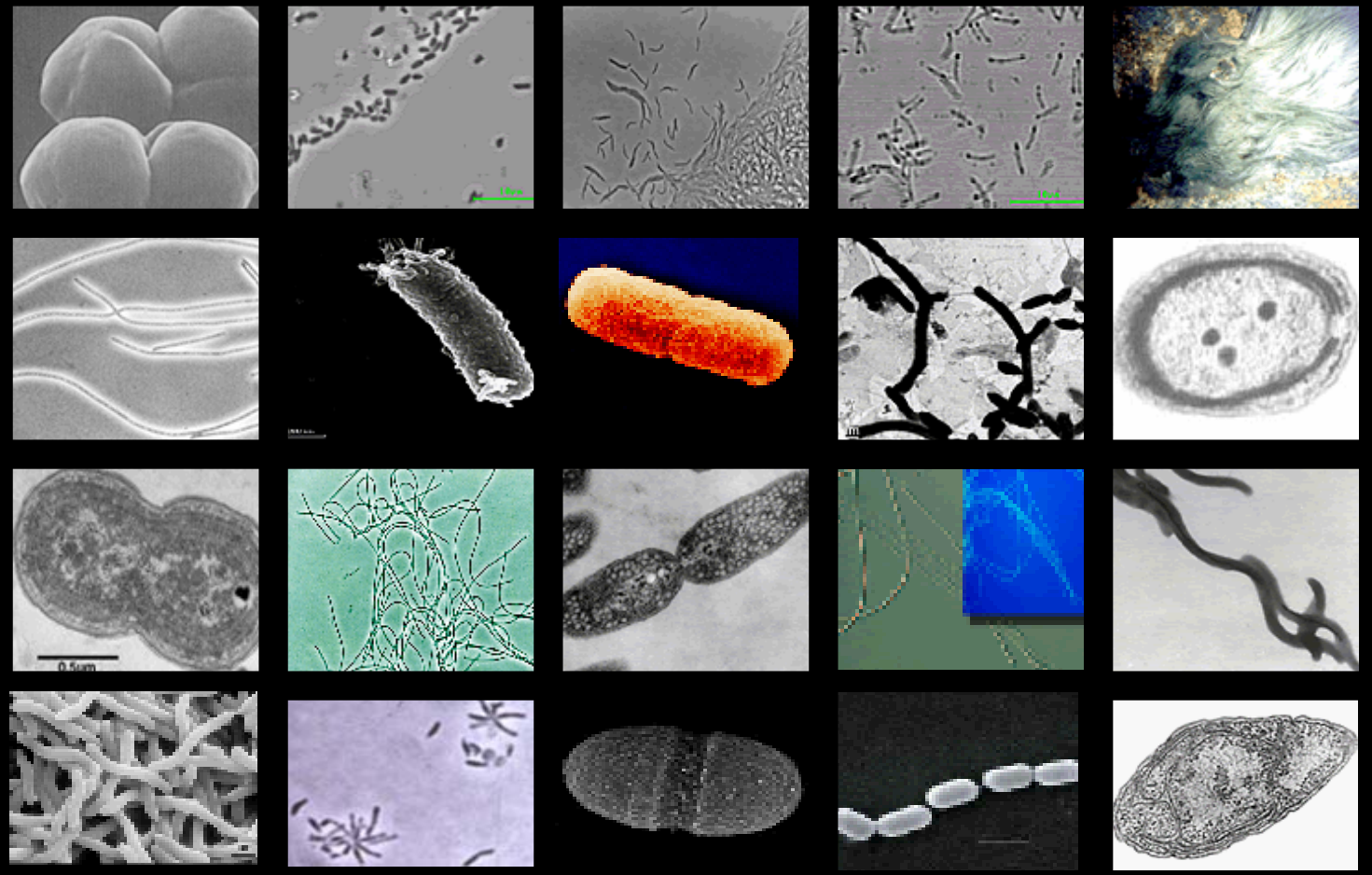
Physcomitrella

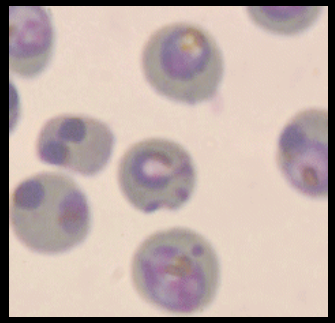
Selaginalla

JGI
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Over 100 prokaryotic genomes

Studied for many reasons, including understanding carbon cycling, life in extreme environments, and toxic waste degradation

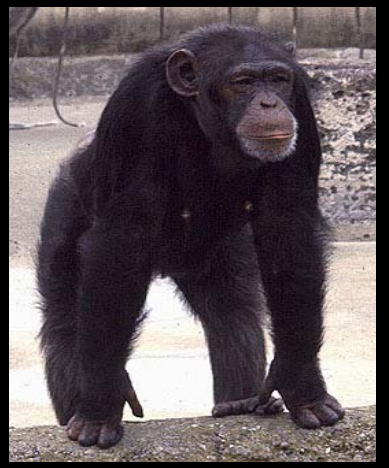




Plasmodium

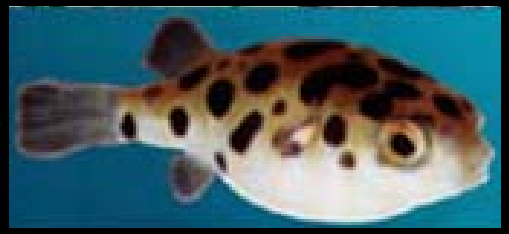
Canis

Bombyx



Pan

Gallus



Tetraodon

Giardia

Arabidopsis

Drosphila (2)

Other sequencing projects complete

Caenorhabditis (2)

Apis

Anopheles

Oryza

Rattus

Other vertebrate projects underway

Macaque

Cow

Opossum

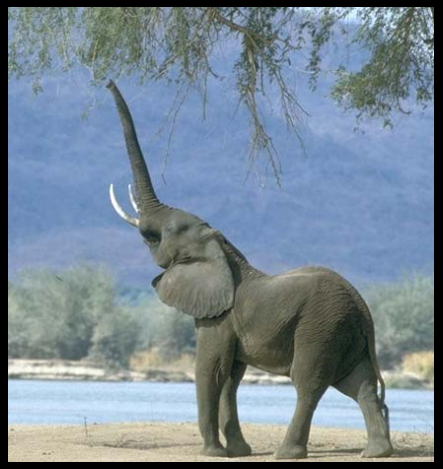
Marmoset

Orangutan

Cat

Rabbit

Guinea pig


Elephant

Hedgehog

Tenrec

Shrew

Lamprey

Zebrafish

Pig

Bat

Armadillo

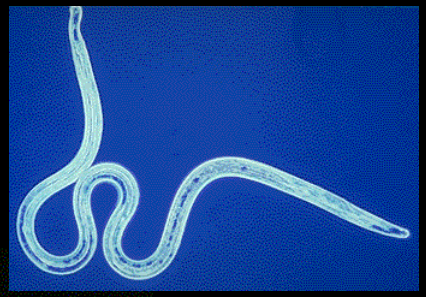
Other non-vertebrate projects underway

Strongylocentrotus

Schmidtea

Tribolium

Saccoglossus



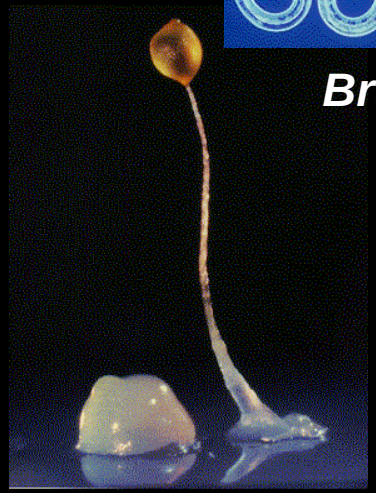
Brugia

Oxytricha

Pristionchus

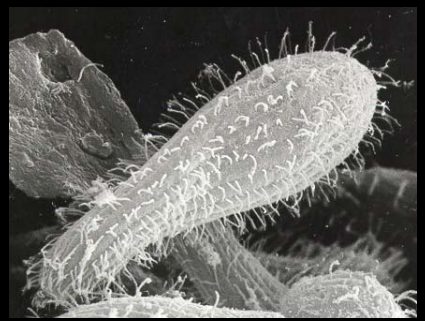
Biomphalaria

Drosophila
many species

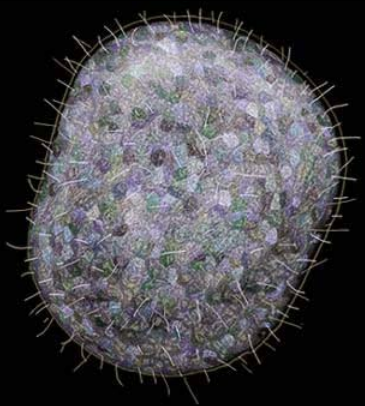


Dictyostelium

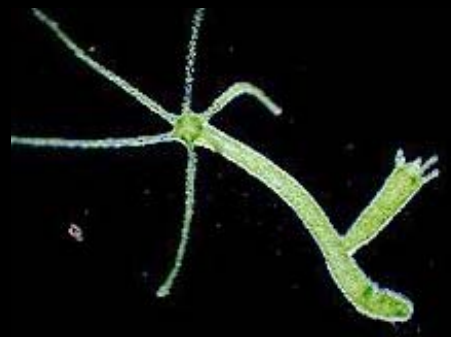
Trichinella



Tetrahymena



Trichoplax



Hydra

Fungal genome initiative

DONE

Aspergillus nidulans
Burkholderia thailandensis
Candida guilliermondii
Candida tropicalis
Candida lusitaniae
Chaetomium globosum
Coccidioides immitis
Coprinus cinereus
Cryptococcus neoformans
Fusarium graminearum
Magnaporthe grisea
Neurospora crassa
Rhizopus oryzae
Saccharomyces cerevisiae
Stagonospora nodorum
Ustilago maydis

IN PROCESS

Botrytis cinerea
Candida albicans
Fusarium verticillioides
Lodderomyces elongisporus
Pneumocystis carinii
Uncinocarpus reesii

Getting into the JGI Pipeline

Work for other agencies (Grants, Agreements)

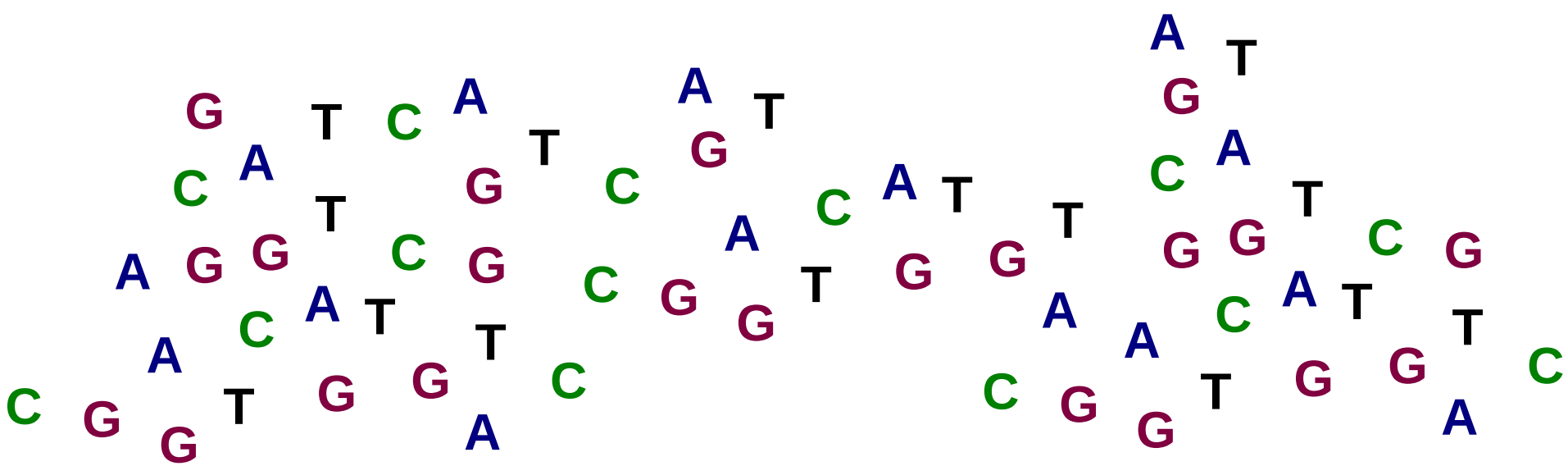
$$$

DOE: GTL and Microbial Genomes

40%

DOE: Community Sequencing Program

60%



JGI CSP
COMMUNITY SEQUENCING PROGRAM

60% of JGI's sequencing capacity
(~1.8 Gb/mo.)
Paid for by DOE, but distributed without regard to DOE's mission

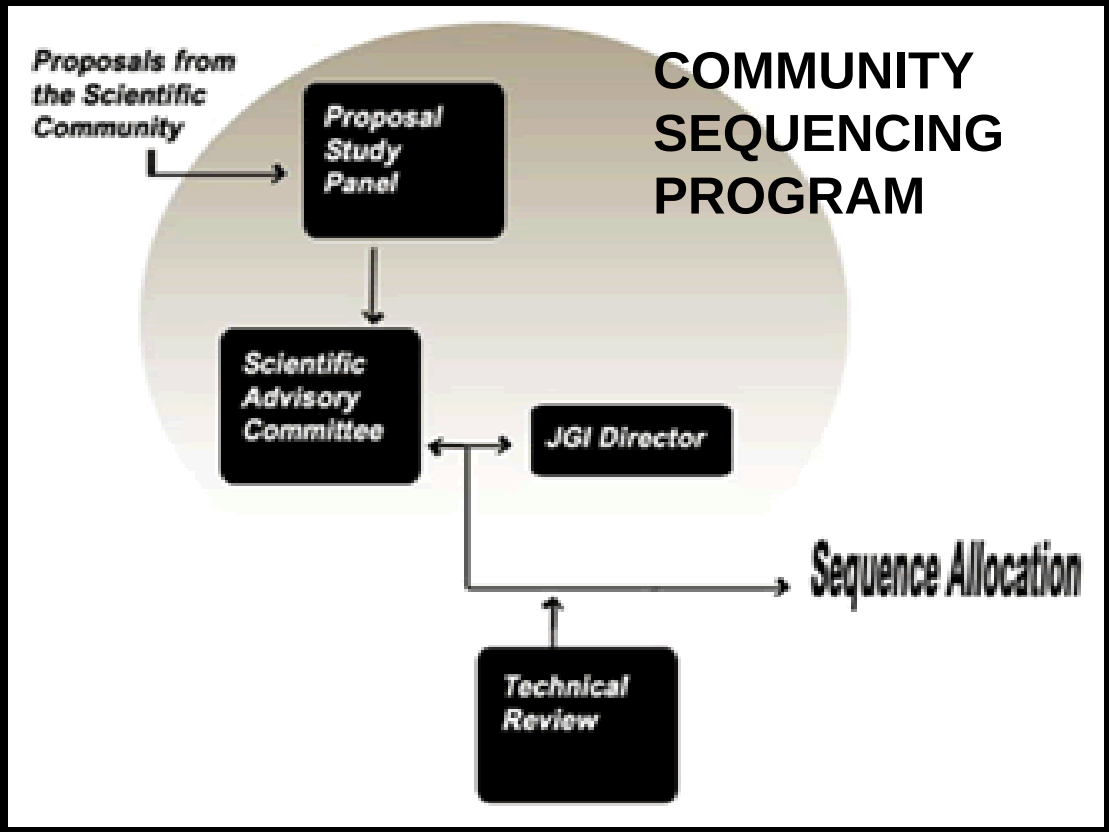
Scoring criteria
- Scientific merit
- Technical feasibility
- Capability of the proposer
- Amount of JGI resources

Determined by peer review

Much interest - 134 proposals this year

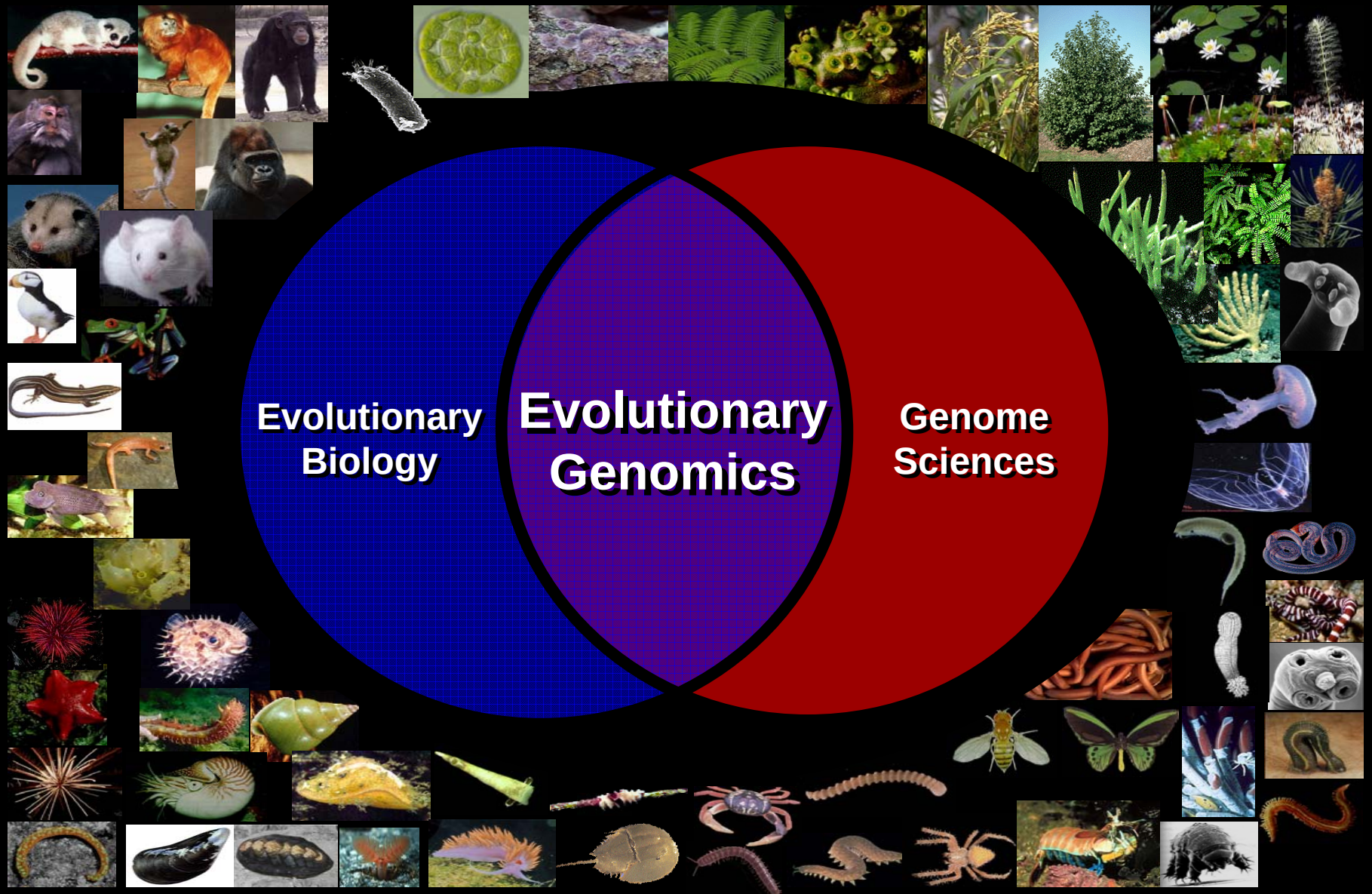
Data release policy - For the community

NIH has similar program



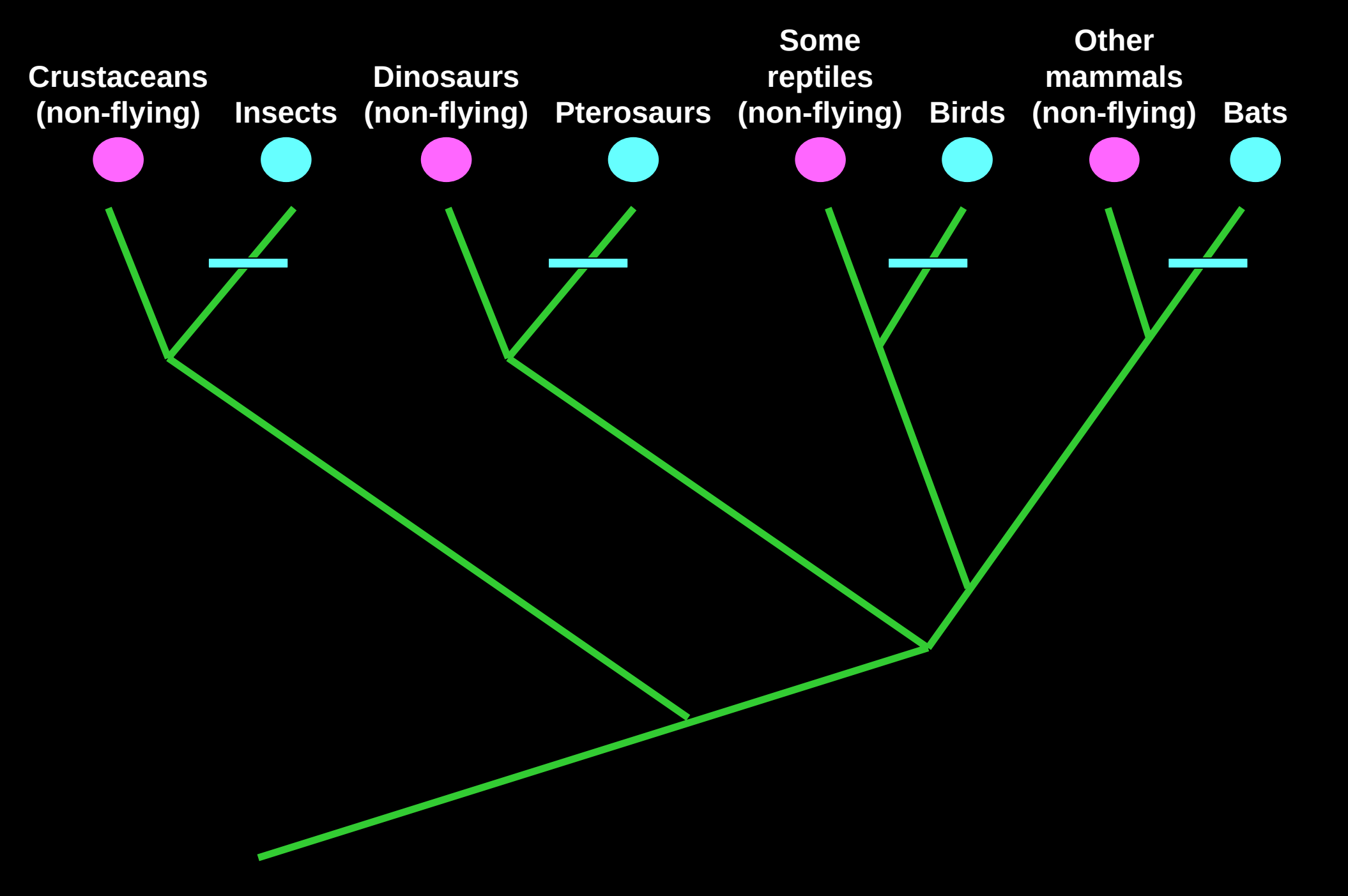
1. Raw shotgun sequence
2. Shotgun sequencing of BACs/fosmids
3. EST/cDNA sequencing
4. Targeted resequencing
5. Finishing
6. Prokaryotes, protists, plants, fungi, animals, organelles

The New "Modern Synthesis" Period



How can the principles of evolutionary biology illuminate genome comparisons?

How many times has flight evolved?


Crustaceans (non-flying)
Insects
Dinosaurs (non-flying)
Pterosaurs
Some reptiles (non-flying)
Birds
Other mammals (non-flying)
Bats

These same principles can be applied to many genome level features, such as:

Gene rearrangements
Gene duplications
Chromosome level changes
Metabolic pathways
Developmental pathways
Gene regulation patterns

Gene Clustering

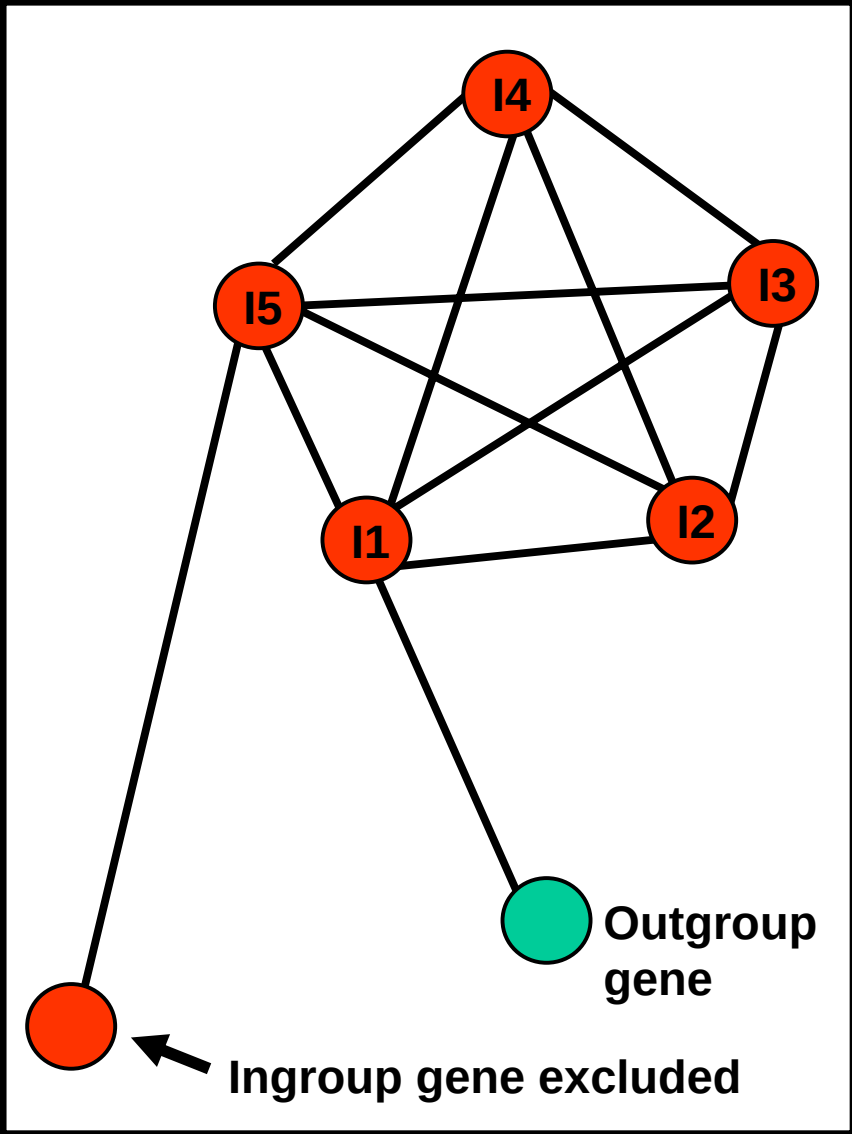
All-against-all Blastp

Global alignment of each gene pair identified and calculation of distance

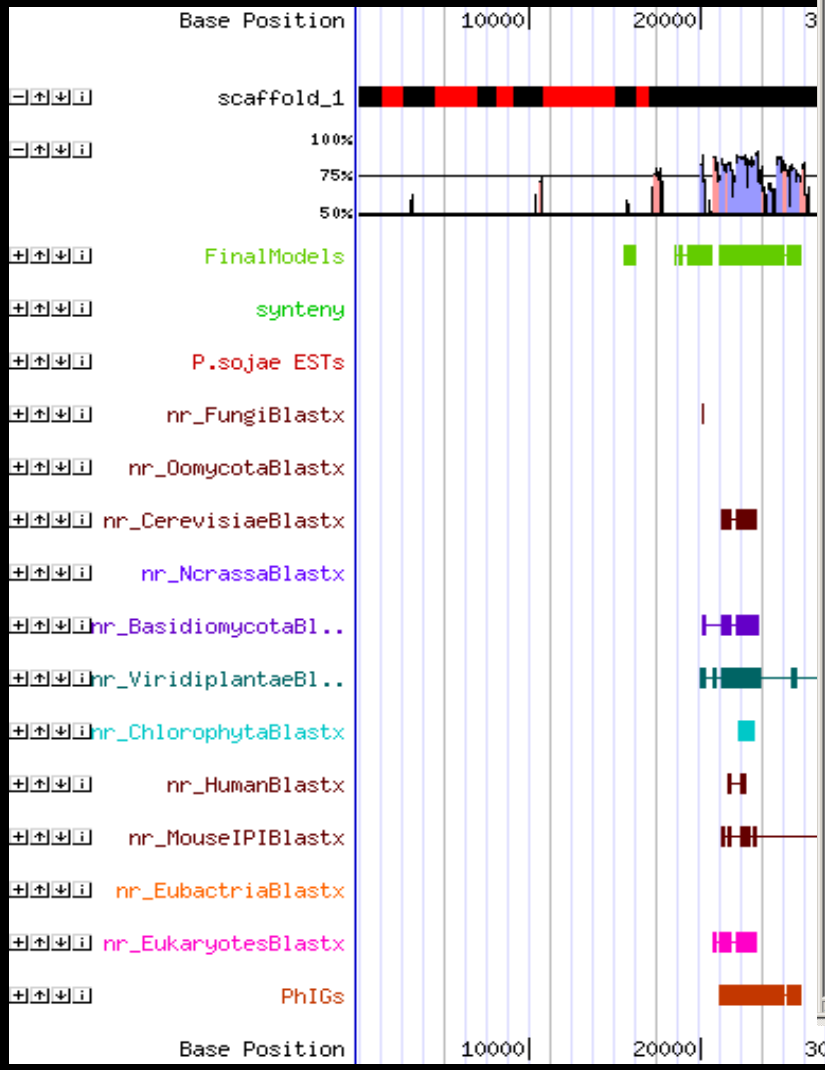
Construct a graph with each protein as a node and raw scores as weight for each edge

Use this graph to seed a search of all proteins to collect those with scores greater than the seed. No gene may be used more than once.

This ensures that each cluster contains the descendents of a single gene in the common ancestor.



PhIGs - Interface



Cluster Viewer - Microsoft Internet Explorer

File Edit View Favorites Tools Help

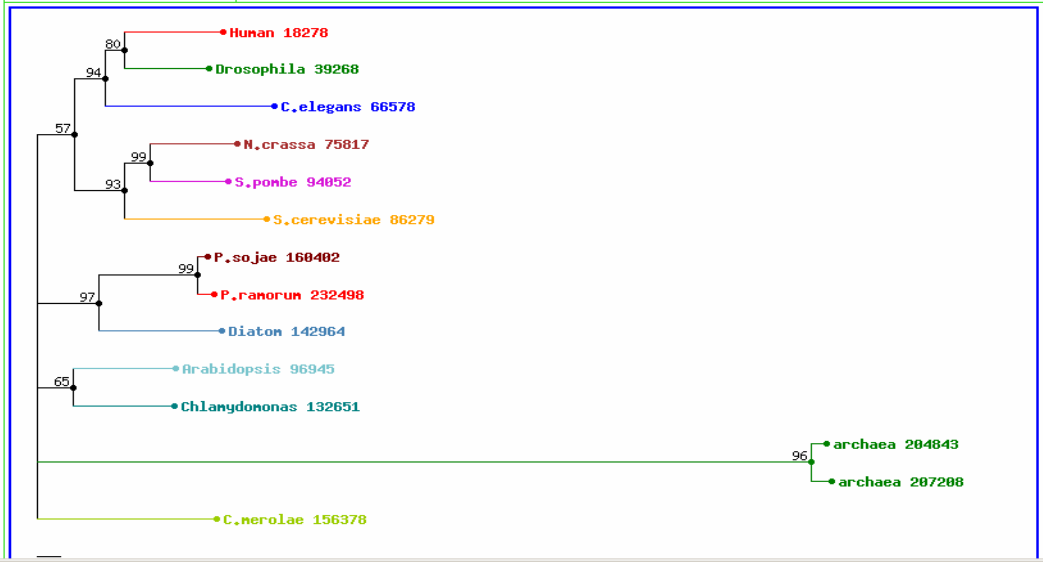
Back Search Favorites

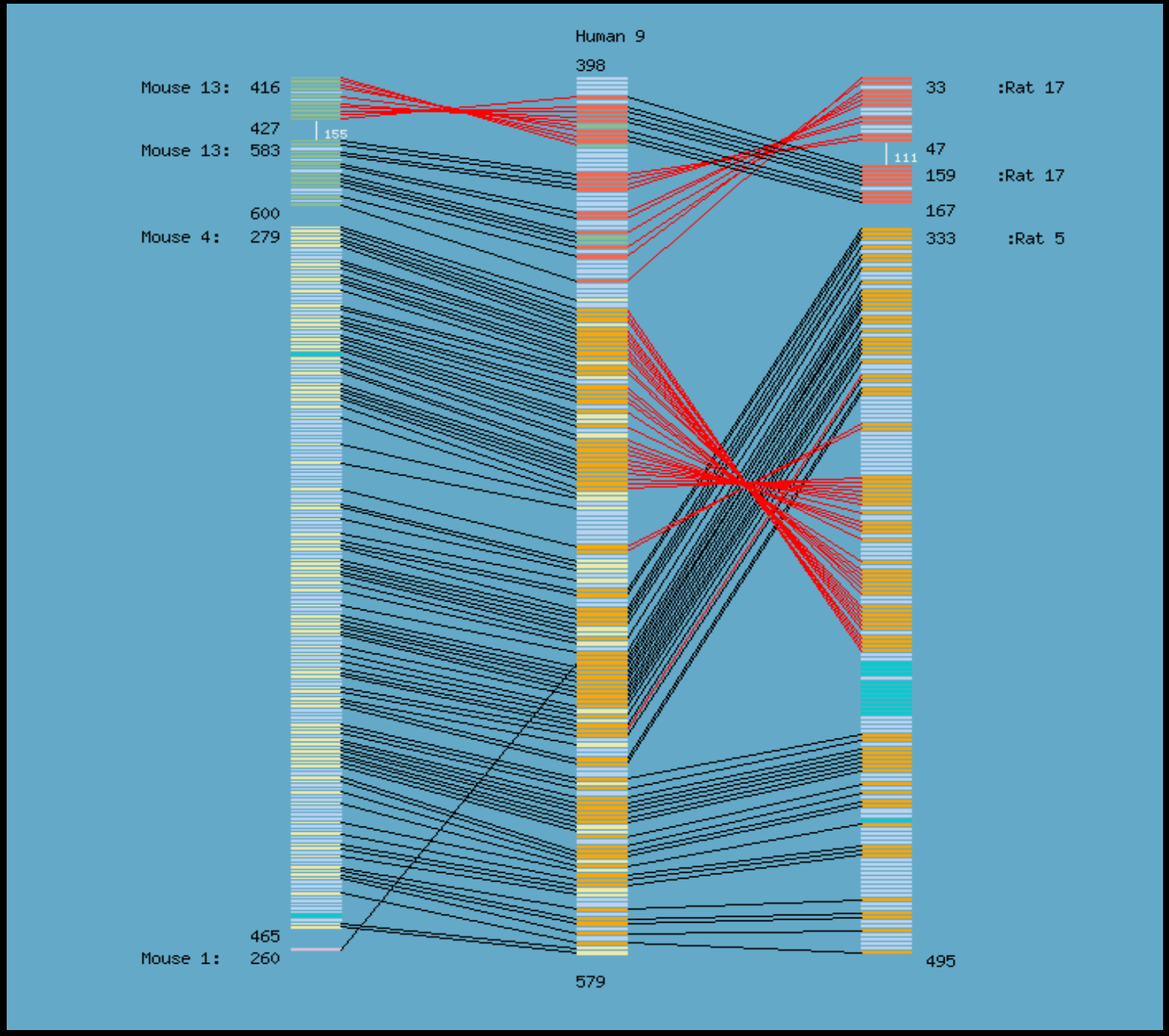
Address http://durant.jgi-psf.org/cgi-bin/psdCluster2.pl?host=durant&database=PhIGs1&clusterId=22926

18278	ENSG00000062822.1	2.7.7.7) (DNA polymerase delta subunit p125). [Source:SWISSPROT;Acc:P28340]	POLD1:HUGO	Human	1108	19	1131
39268	CG5949		DNApol-delta:flybase_symbol	Drosophila	1092	3L	1801
66578	F10C2.4		F10C2.4:wormbase_gene	C.elegans	1082	V	2892
75817	NCU01192.1	hypothetical protein		N.crassa	1105	3.45	37
86279	YDL102W	CDC2		S.cerevisiae	1098	IV	156
94052	SPBC336.04	DNA polymerase delta (catalytic subunit) (PMID 1960723)	cdc6	S.pombe	1086		
96945	DPOD_ARATH	DNA polymerase delta catalytic subunit (EC 2.7.7.7)	Q9LVN7	Arabidopsis	1081		
132651	162097	x		Chlamydomonas	1080	scaffold_28	112
142964	113984	x		Diatom	1096	scaffold_10	156
156378	CMN199C	DNA polymerase delta catalytic chain		C.merolae	1084		
160402	144243	x		P.sojae	1094	scaffold_174	7
232498	72220	x		P.ramorum	1158	scaffold_110	11
204843	Q8PVG1	DNA polymerase delta catalytic subunit (EC 2.7.7.7)	Q8PVG1	archaea	933		
207208	Q8TSB3	DNA-directed DNA polymerase	Q8TSB3	archaea	937		

Tree

MsaId	method	
734	Clustalw 1.83 default	view
Tree Method	TREE-PUZZLE v5.2 JTT Gamma 8 Rate categories	

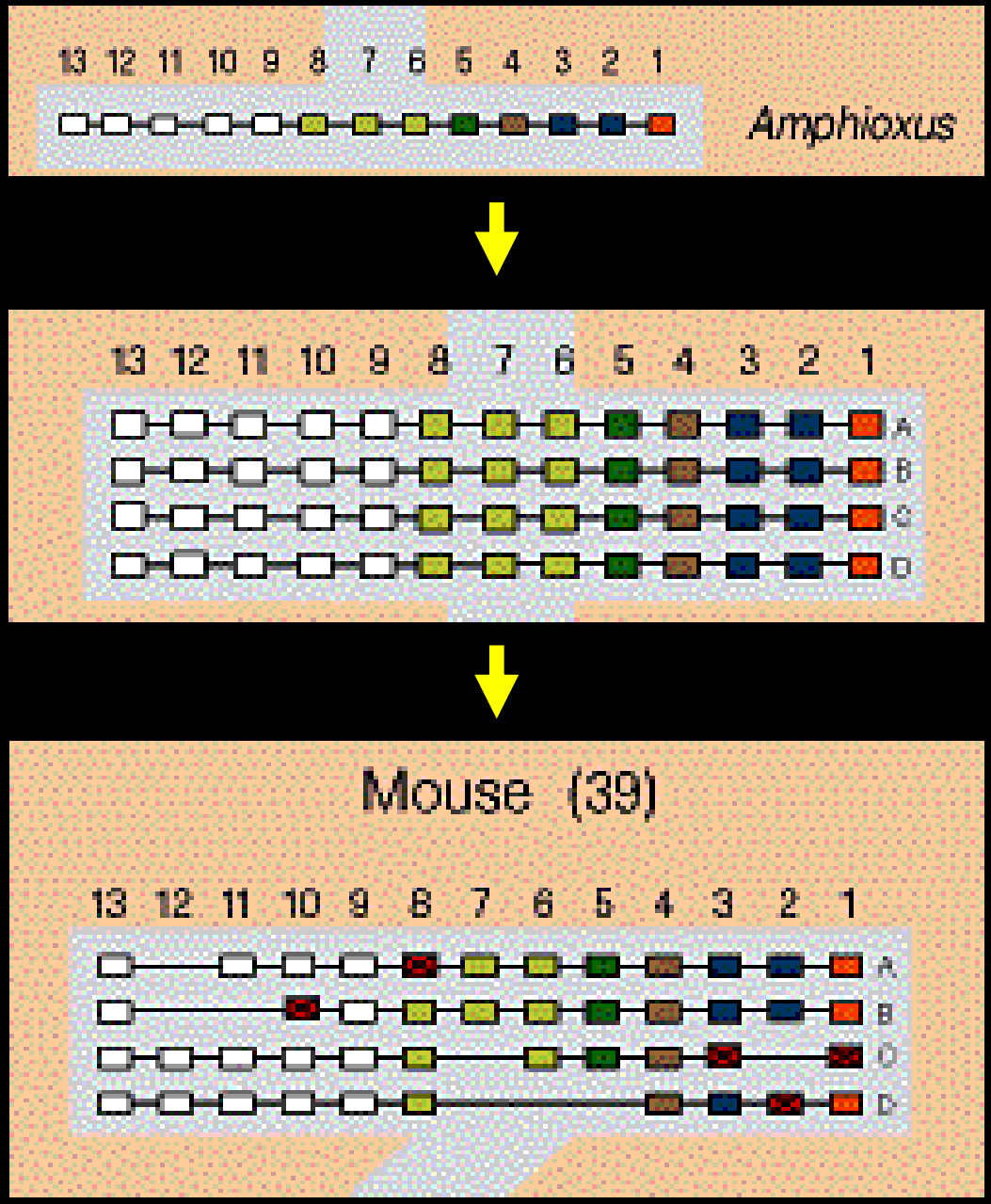



Human 9
398
Mouse 13: 416
427
Mouse 13: 583
600
Mouse 4: 279
465
Mouse 1: 260
579
155
33 :Rat 17
47
111
159 :Rat 17
167
333 :Rat 5
495

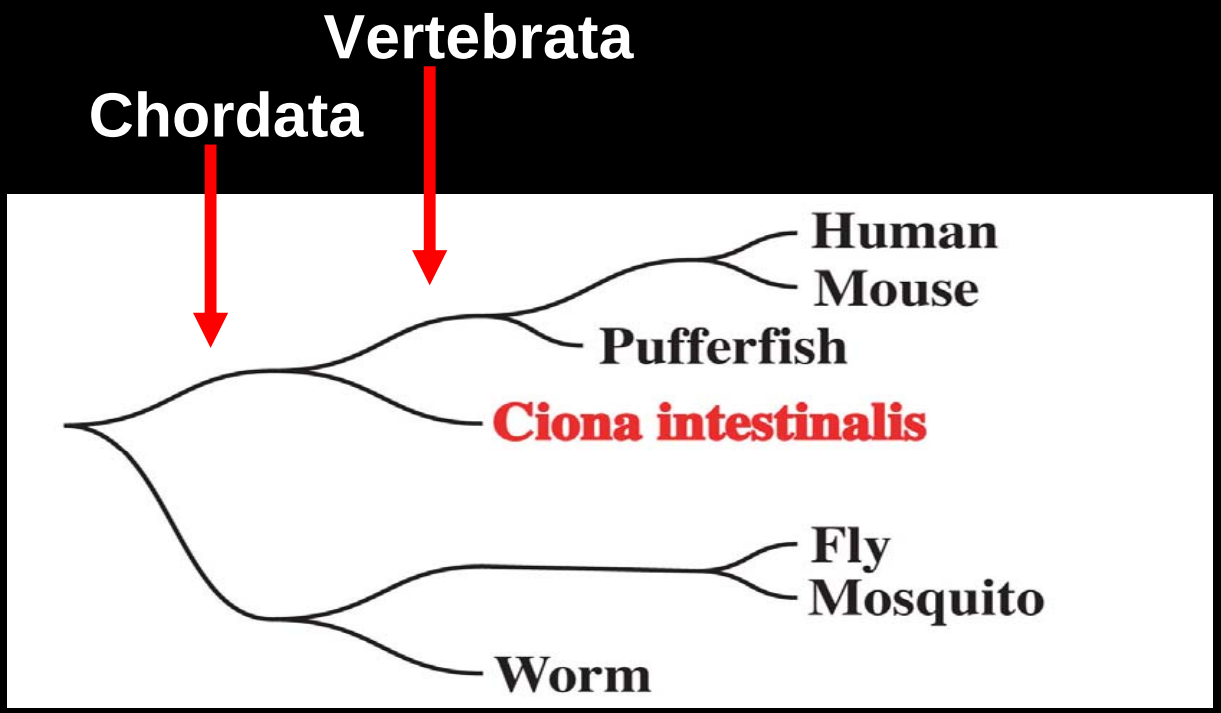
Duplications in Hox clusters first caused some to consider that whole genome duplications may have occurred at the base of Vertebrata.

Alternatives to whole genome duplication:

- **Tandem gene duplications**
- **Large segment duplication**
- **Polysomy**



(figure from Meyer review)



To differentiate:

- Timing of duplications
- Gene family membership
- Arrangement of duplications

http://shake.jgi-psf.org/cgi-bin/psdCluster.pl

Back Forward Stop Refresh Home AutoFill Print Mail

Address: http://shake.jgi-psf.org/cgi-bin/psdCluster.pl?clusterId=26568&database=proteinClusterCi4 go

UCB LBL Google NCBI JGI Production db

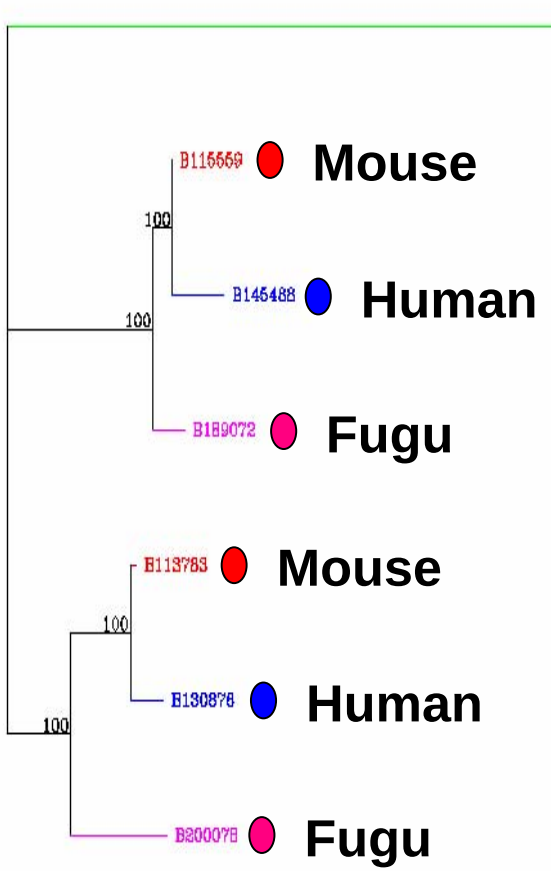
MSA

msaId 2466

method clustalw default, trimmed gaps removed

Tree Method puzzle, 8 gamma rate categories

Tree [lh=-5099.580494](A280:0.35353,((B115559:0.00001,B145488:0.03011)100:0.01099,B189072:0.01853) 100:0.08331,((B113783:0.00395,B130876:0.01945)100:0.03426,B200078:0.05559) 100:0.03623);



Ciona

Mouse

Human

Fugu

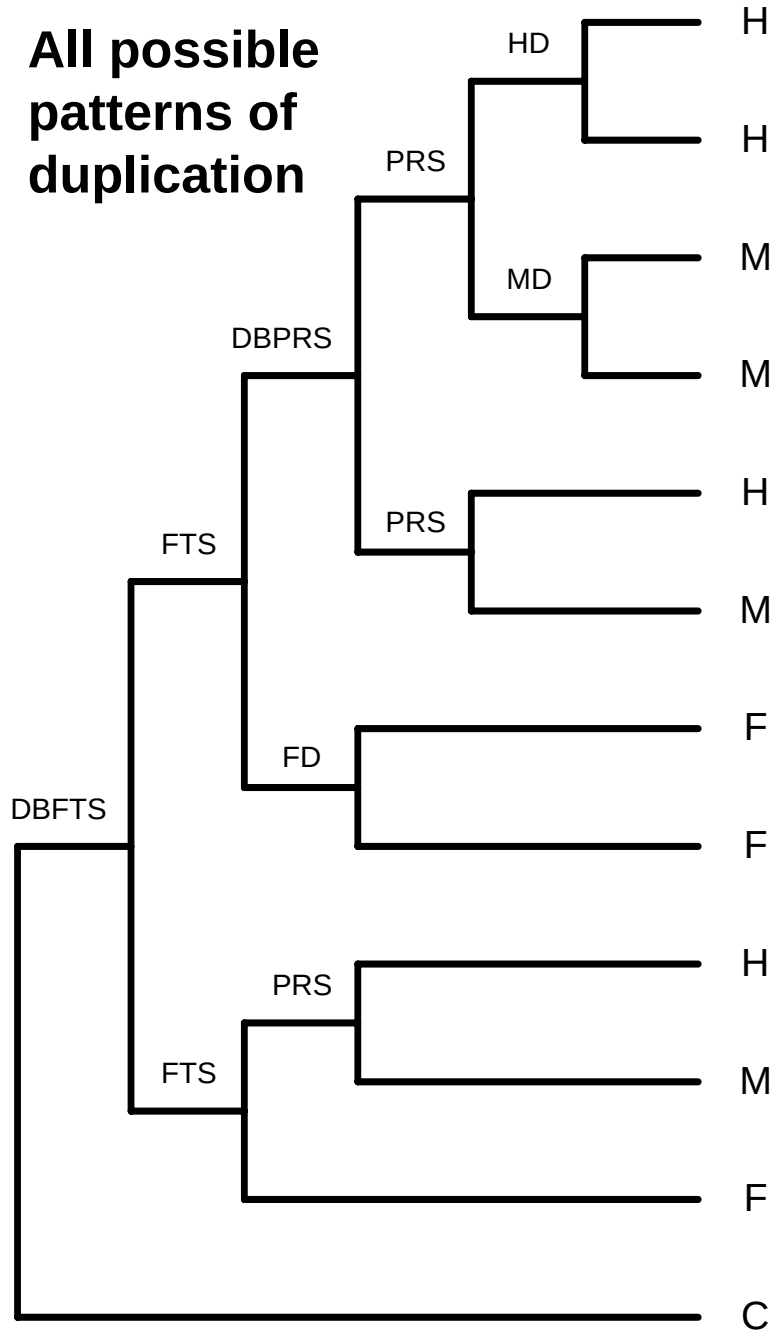
Mouse

Human

Fugu

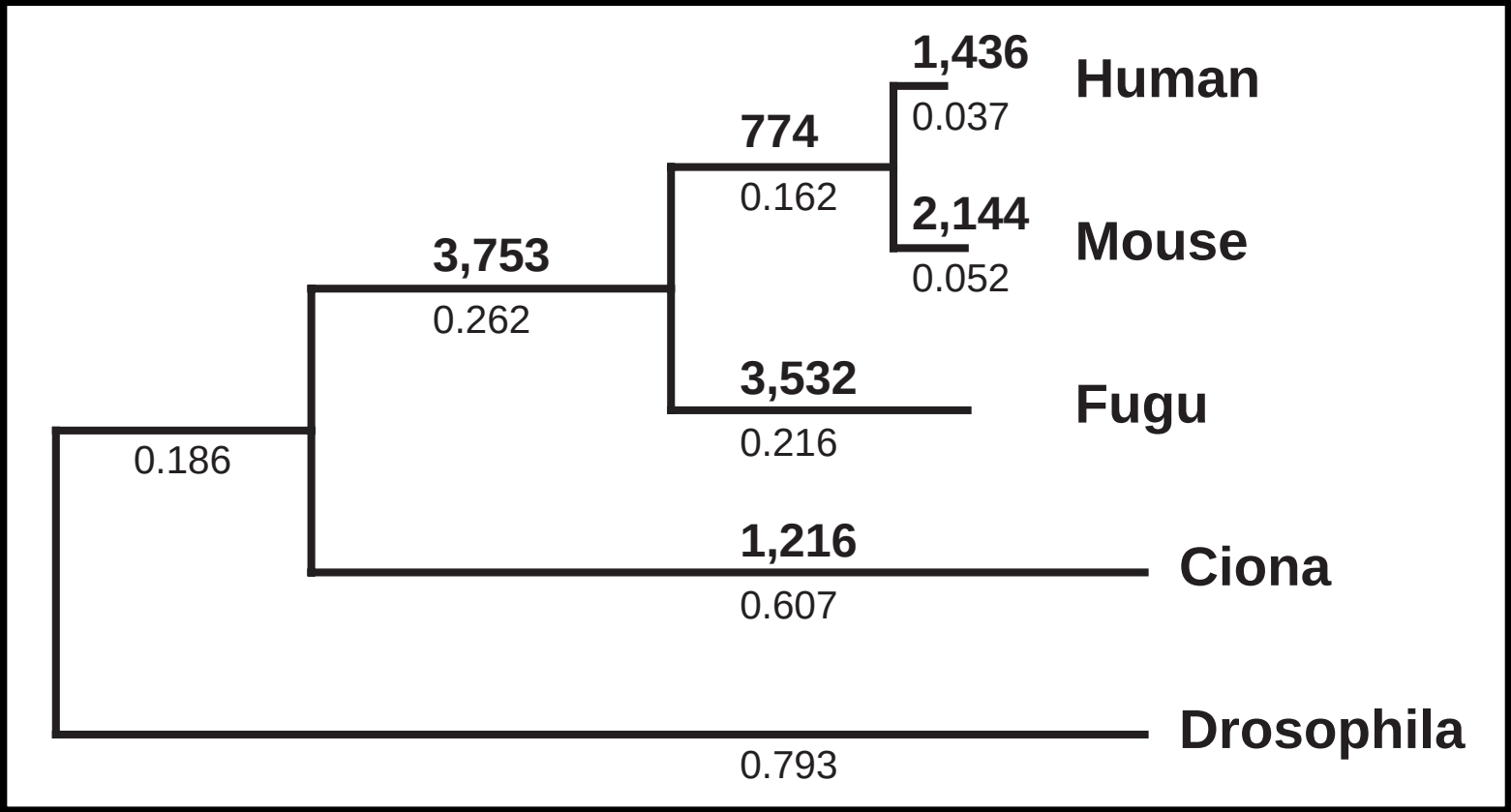
2 member family supporting a duplication prior to the split of these vertebrates

All possible patterns of duplication

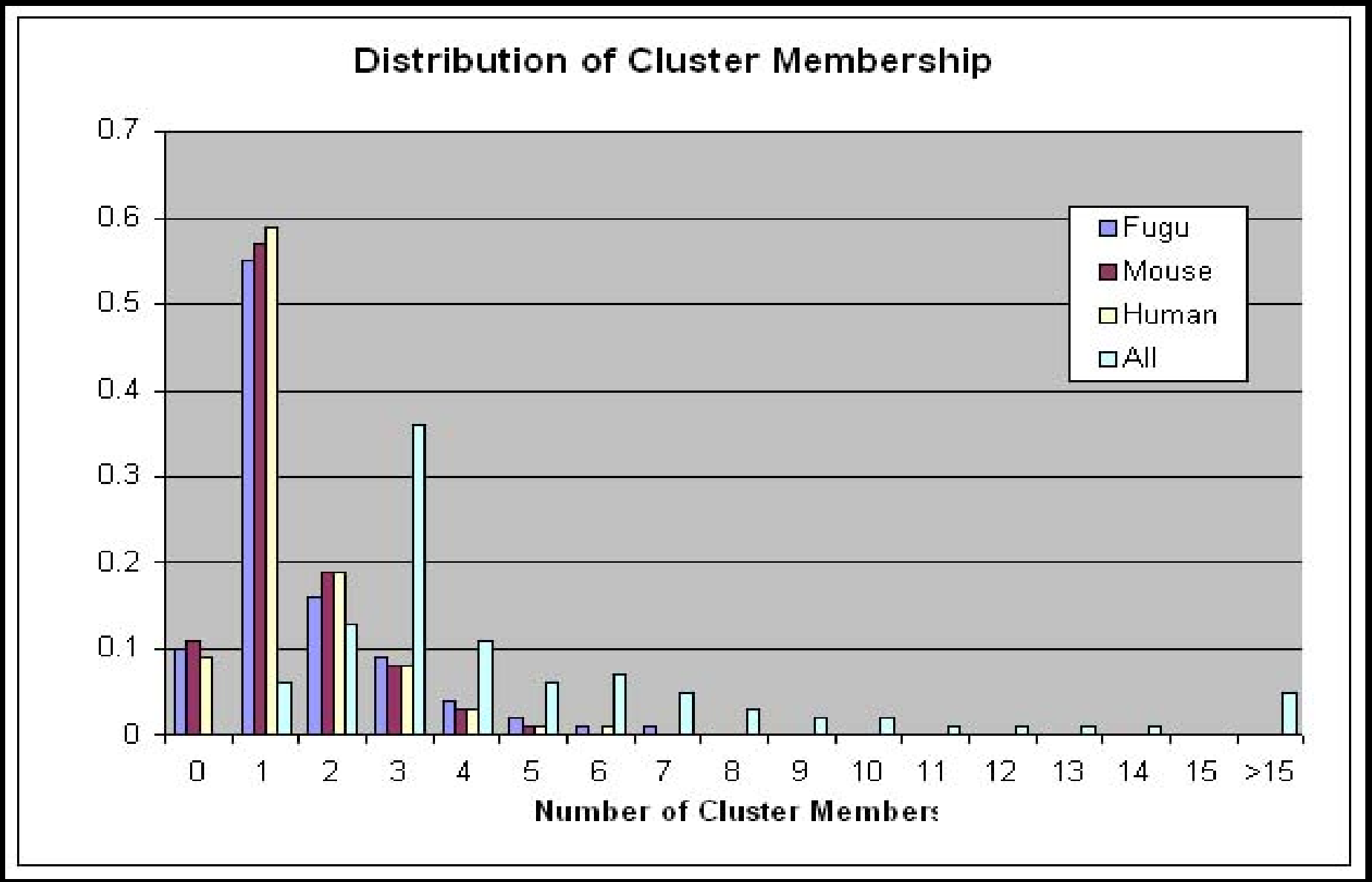


54% of all gene clusters show no evidence of duplication whatsoever, having exactly one copy in each genome.

There is somewhat higher number of duplications inferred to be at the base of vertebrates, but duplications are common everywhere on the tree, and this could otherwise indicate a greater commonality of individual gene or segmental duplications or a lower rate of loss of duplicated genes.

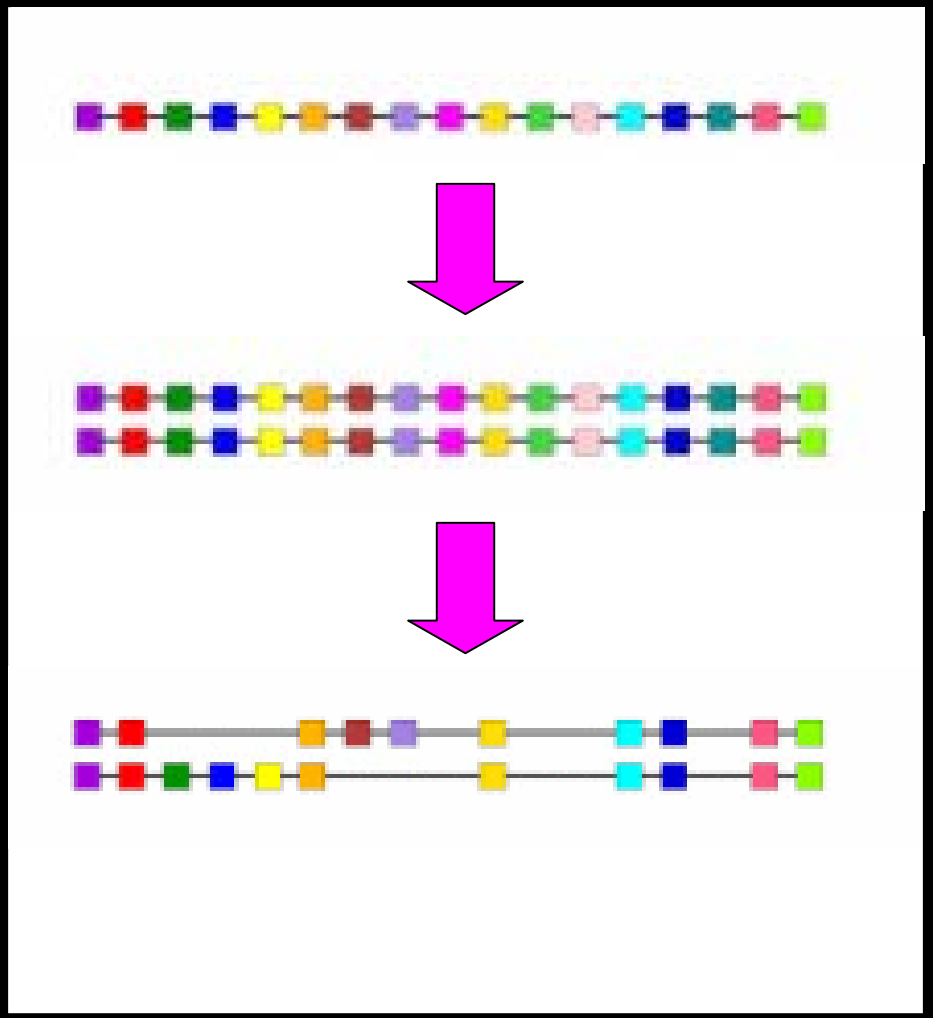


There is no peak at 4 for gene family membership

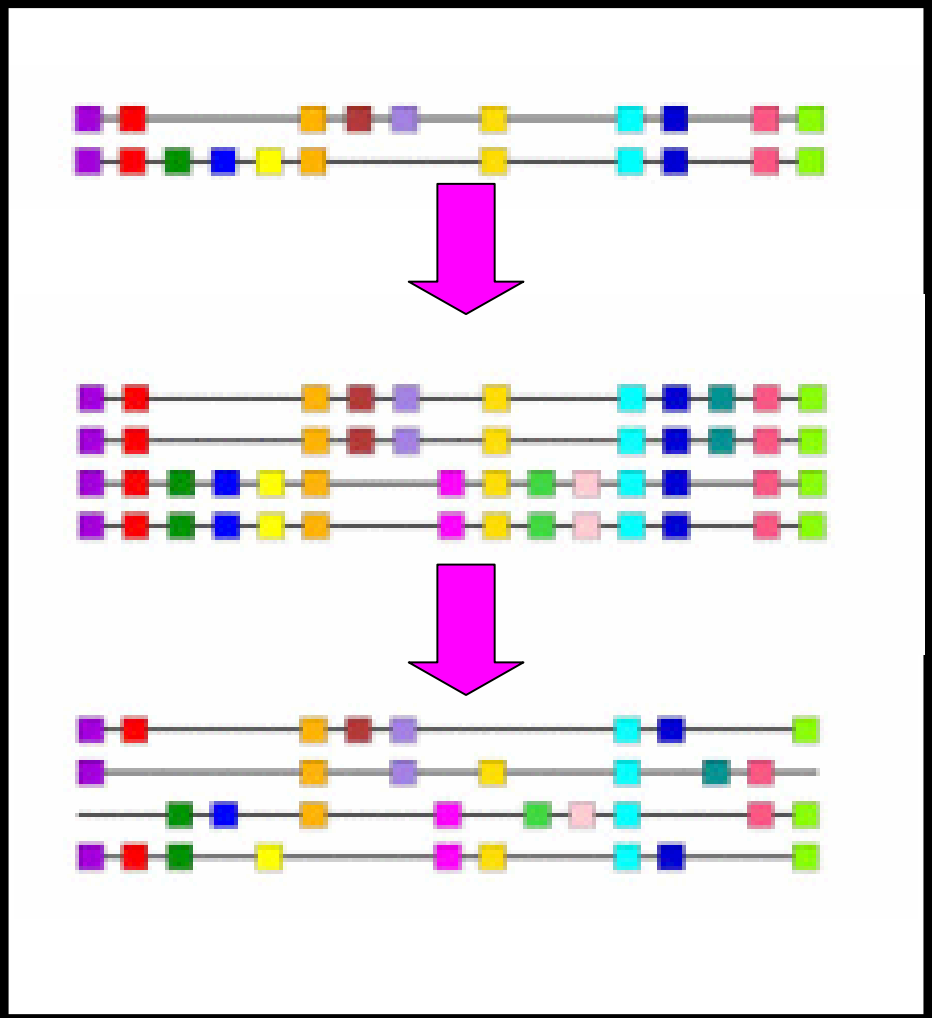


“Tetraparalogons” result from 2 whole genome duplications

Genome duplication followed by gene losses

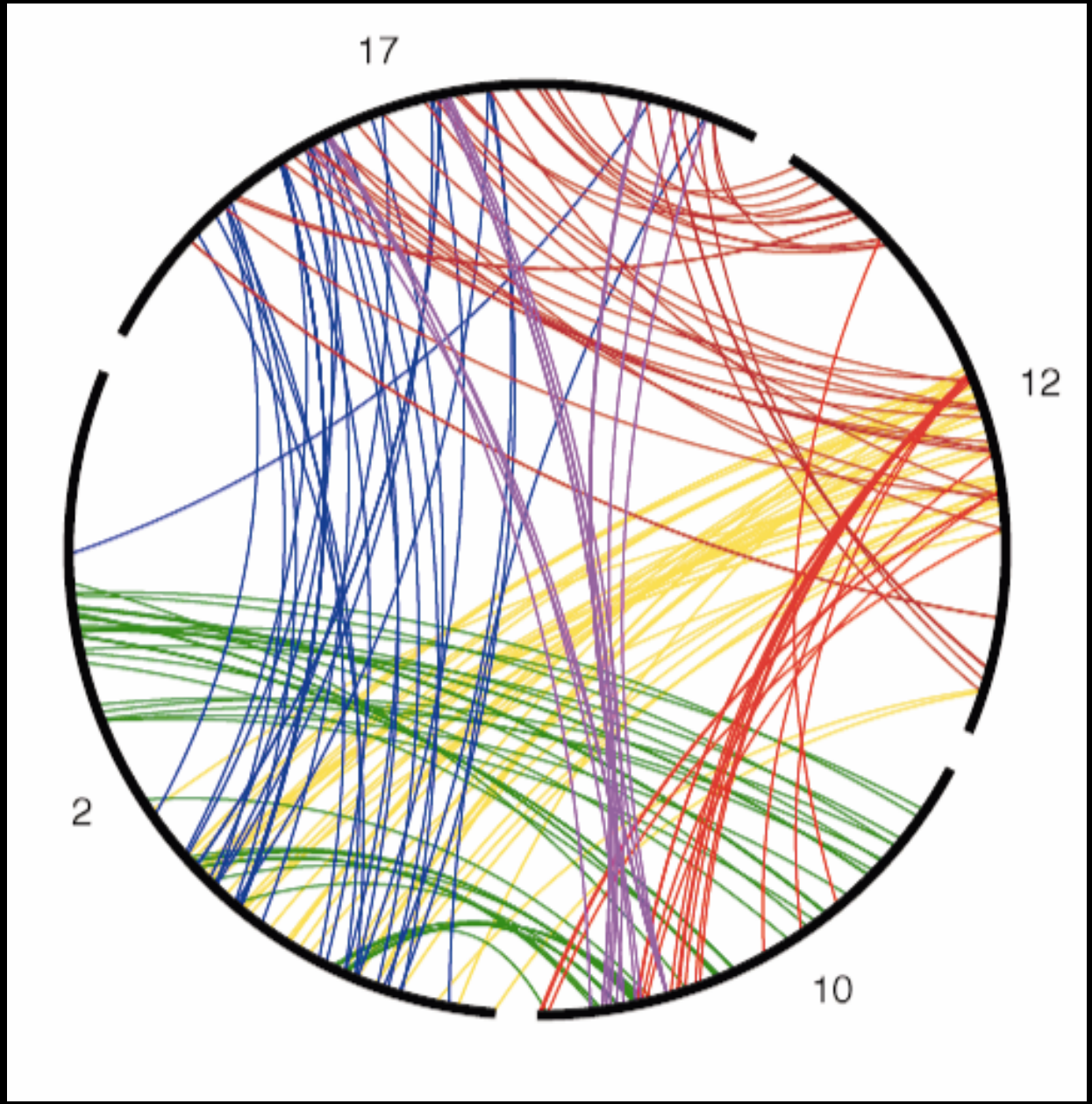


Second genome duplication followed by gene losses



Results for the human genome (best annotated)

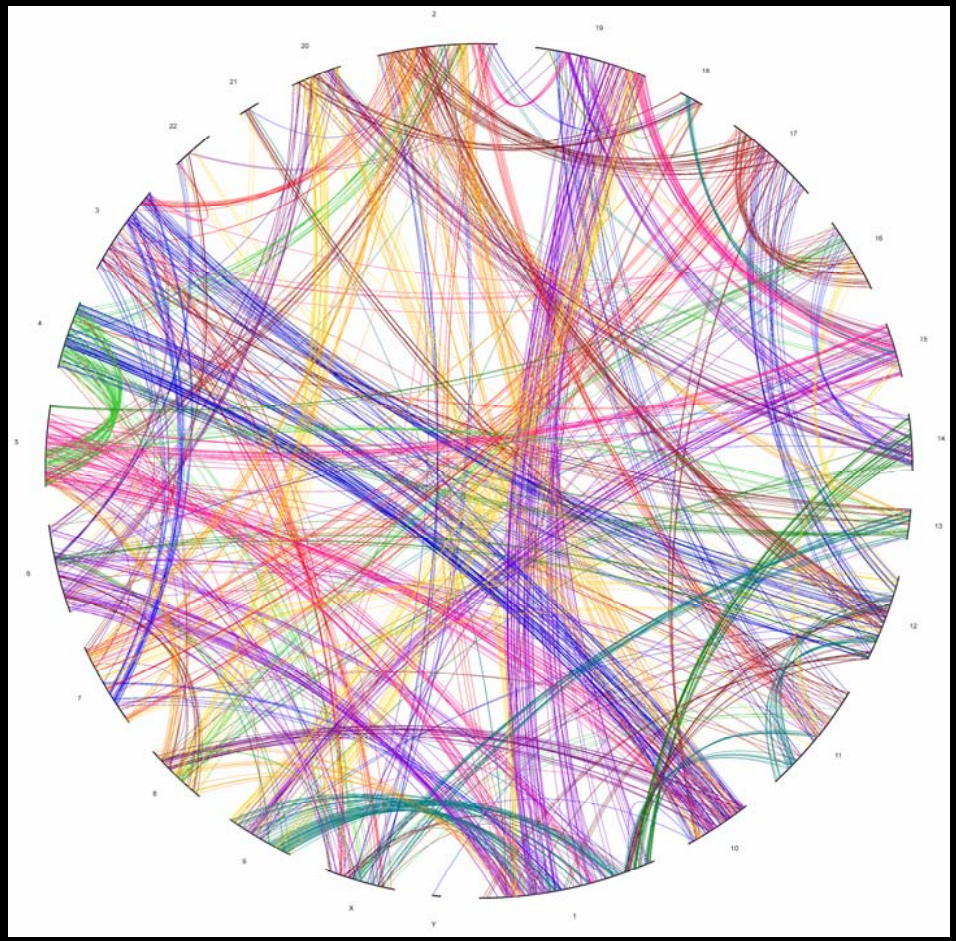
- This is ONLY the subset of early duplicating genes
- Human chromosomes show tetra-paralogy
- Signal is detectable despite
  - Re-arrangements
  - Gene loss
  - Gene conversion
  - Subsequent duplication



This involves nearly every chromosome and covers 72% of the human genome
An estimated 92% of genes from these duplications have been lost

Tetra-paralogy by chromosome

Chromosome	Total Genes	Coverage	% Coverage
1	2,165	1,624	75.0
2	1,455	1,063	73.1
3	1,138	810	71.2
4	849	812	95.6
5	1,008	875	86.8
6	1,113	998	89.7
7	1,063	536	50.4
8	788	759	96.3
9	844	788	93.4
10	839	786	93.7
11	1,415	280	19.8
12	1,088	597	54.9
13	377	338	89.7
14	709	658	92.8
15	679	618	91.0
16	946	842	89.0
17	1,222	884	72.3
18	306	25	8.2
19	1,377	789	57.3
20	636	582	91.5
21	261	0	0.0
22	528	321	60.8
X	869	700	80.6
Y	110	0	0.0
Genome	21,785	15,685	72.0



How can genomic data uncover the evolutionary history of organisms?

The way that life is NOT arranged

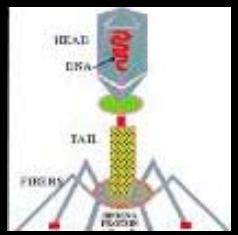
Human

Mus

Xenopus

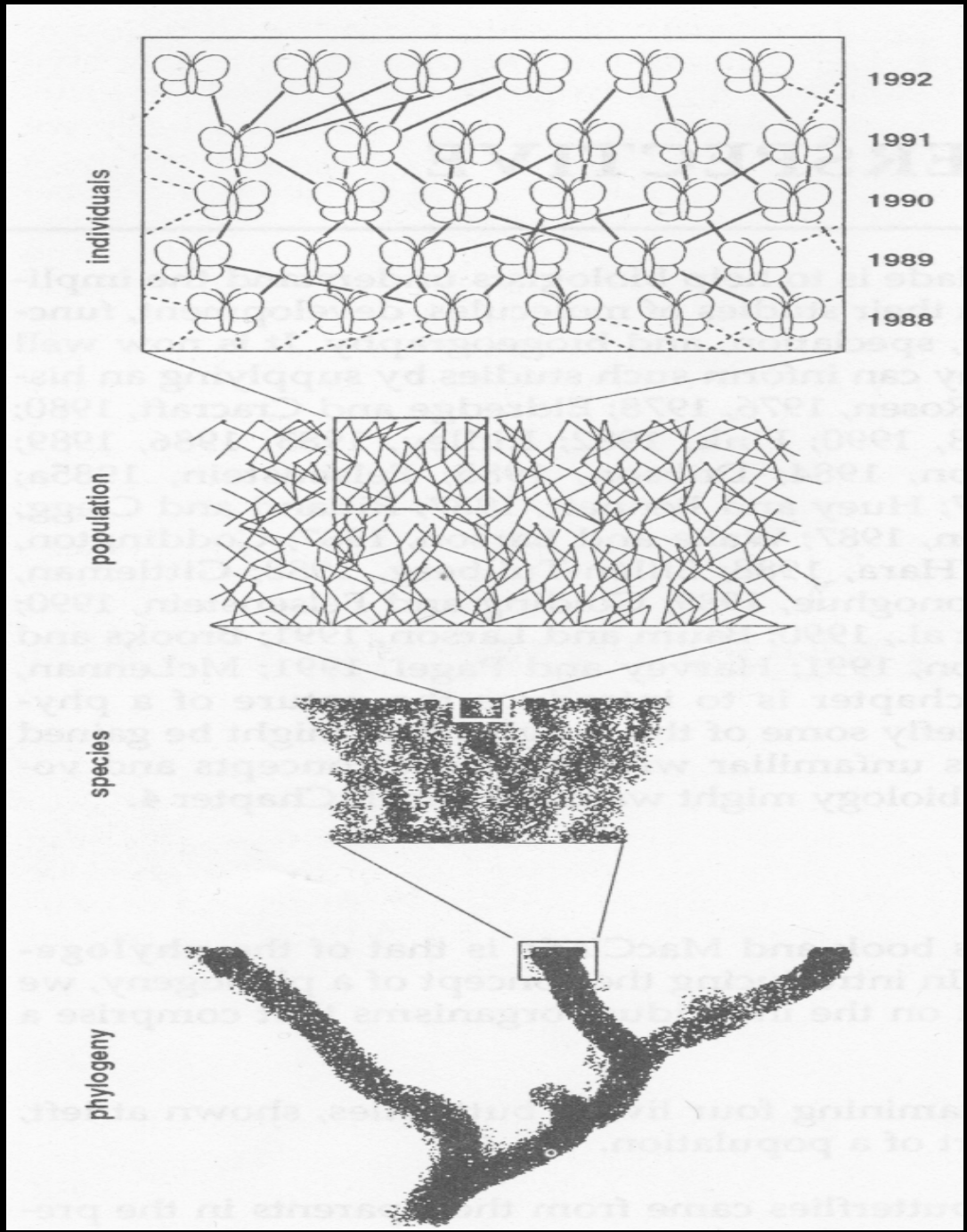
Drosophila

E. coli



Virus

The way that life IS arranged

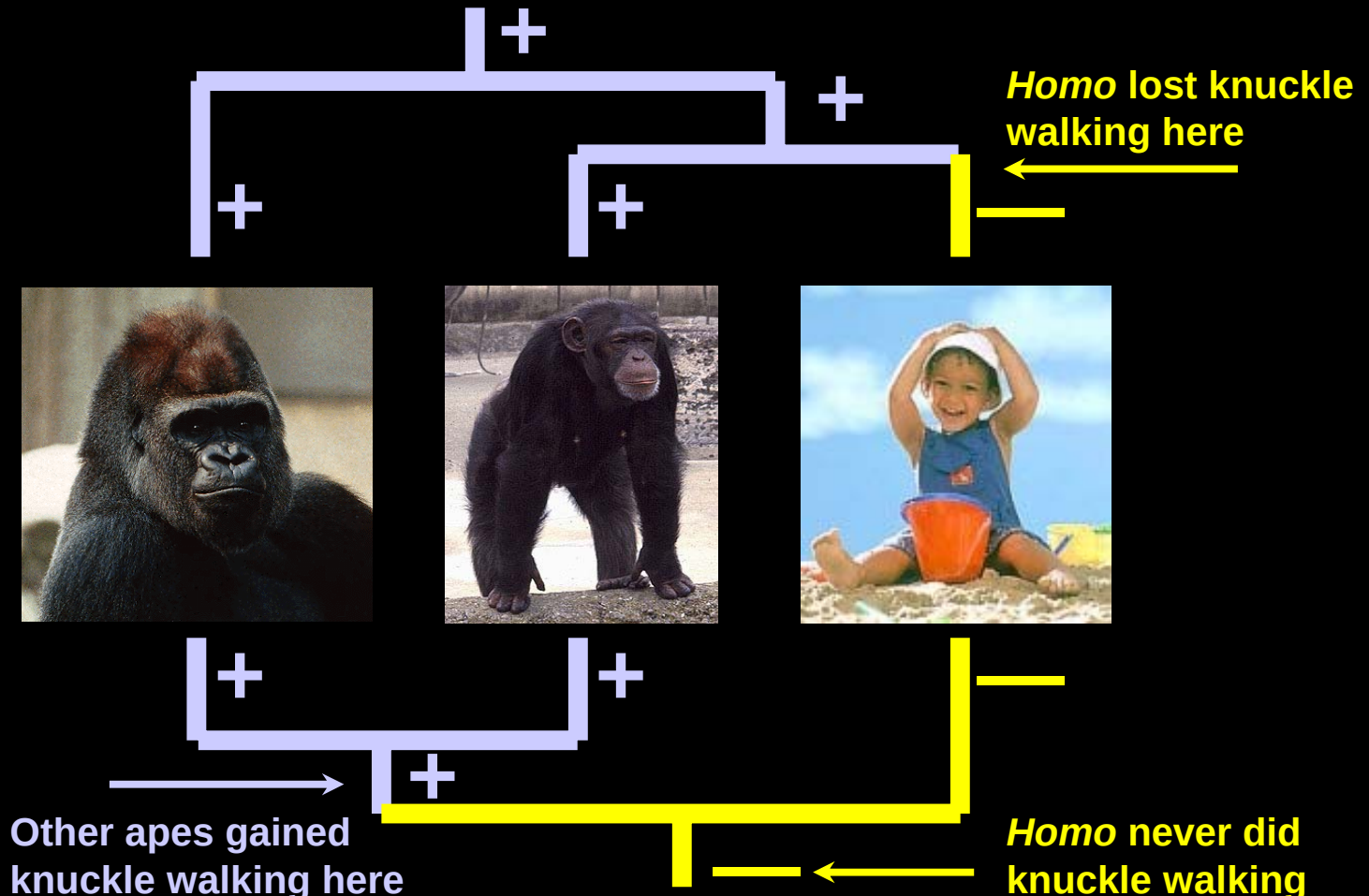


Why does this matter?

Biological interpretations hinge on knowing phylogenetic relationships

+ / —
for knuckle walking

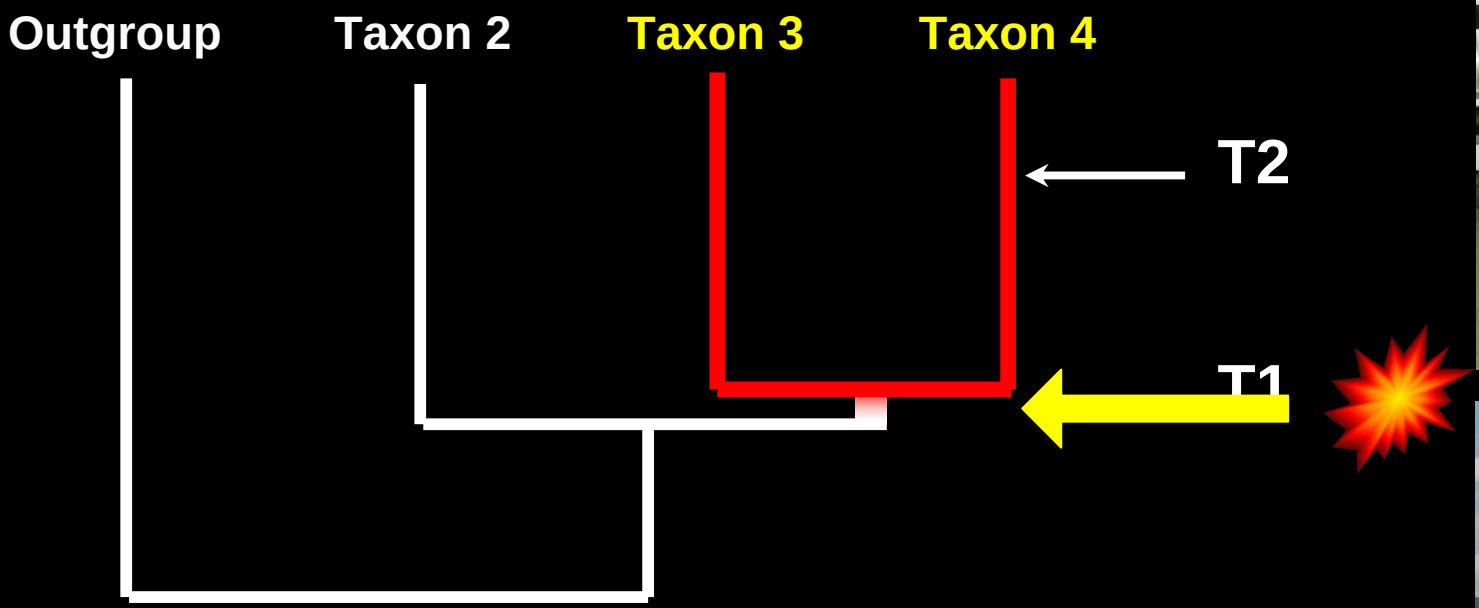
This principle applies to all traits, whether morphological, behavioral, physiological, molecular, etc.

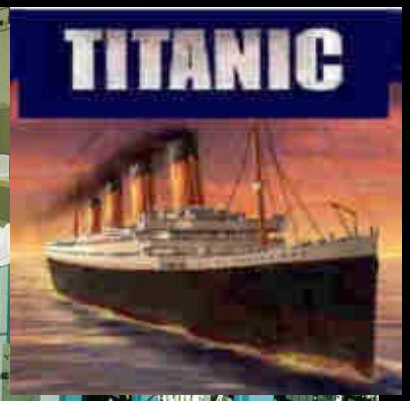


Although comparing DNA sequences has revolutionized our understanding of evolutionary relationships, some branches have remained recalcitrant.

Why?

The T1/T2 ratio problem





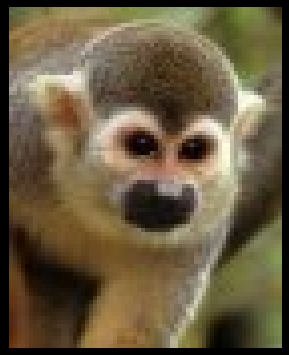
Clock-like characters are guaranteed to fail when the T1:T2 ratio is extreme

Packet watch from the body of postal clerk John March, National Postal Museum in Washington

Lemur

Loris

Tarsius

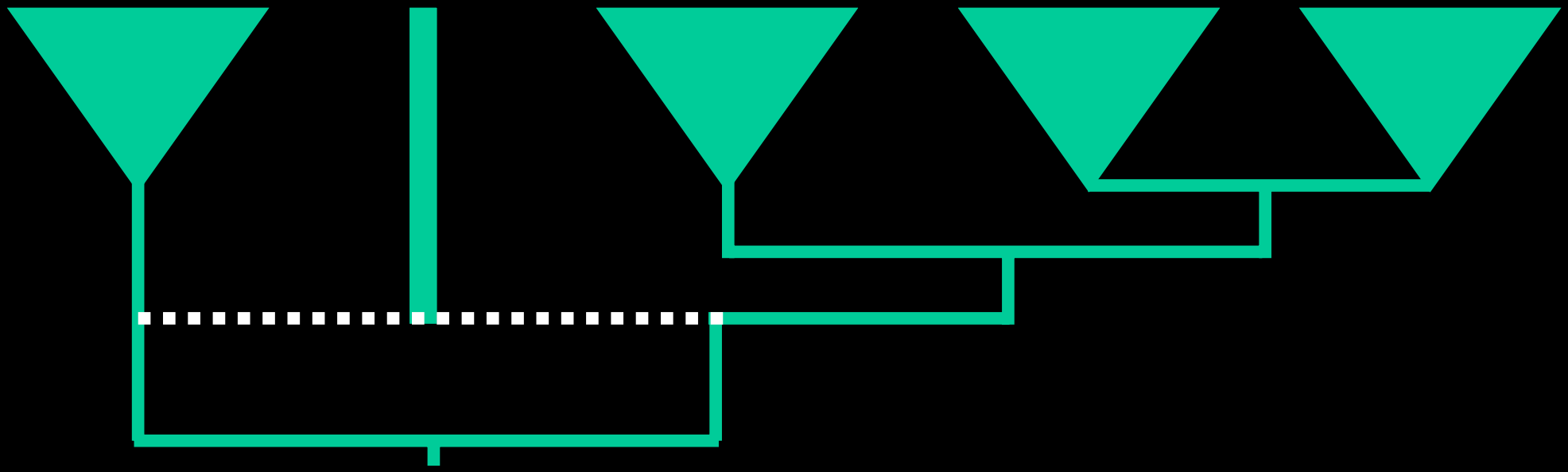


Squirrel monkey

Baboon

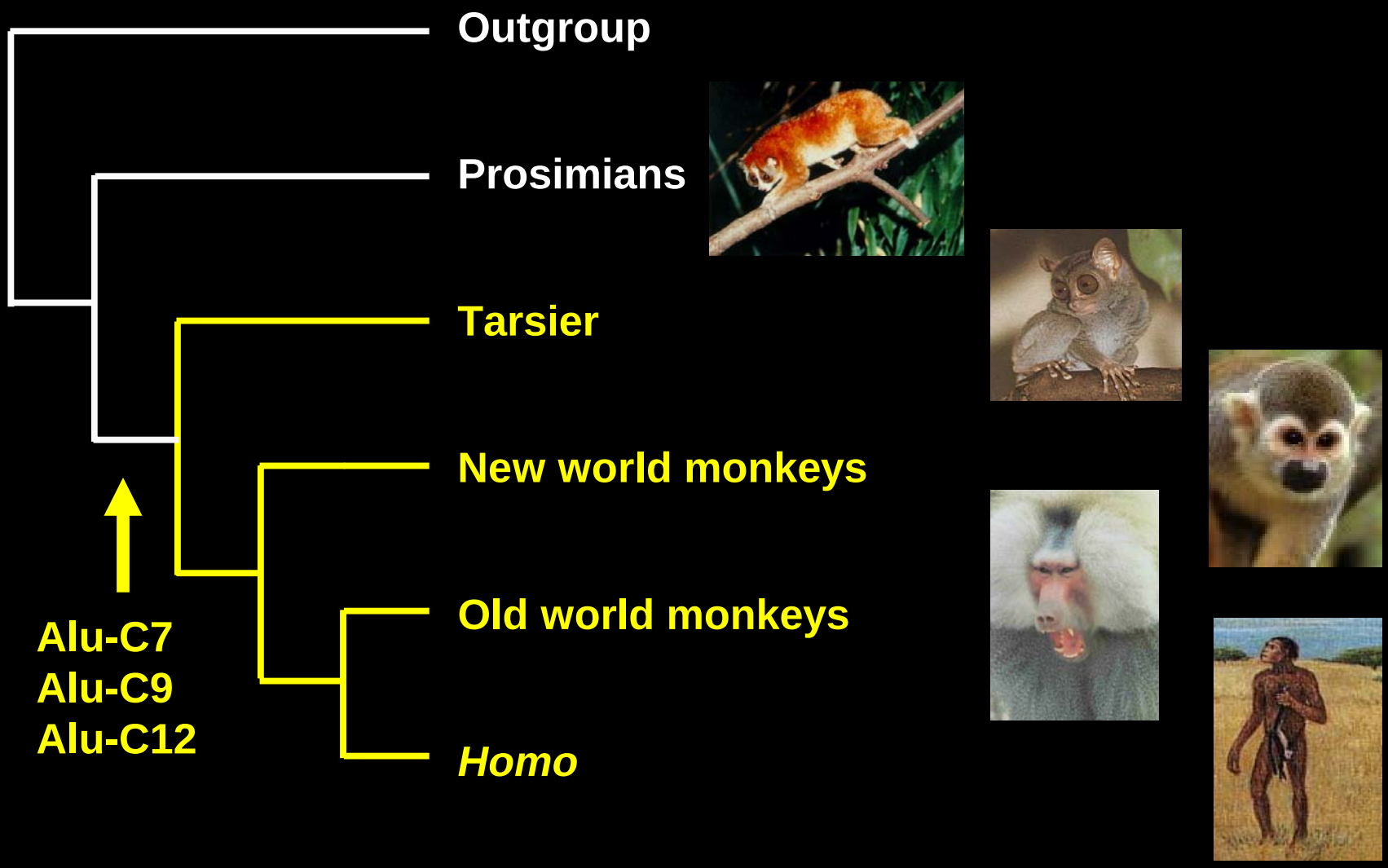
Chimpanzee

Prosimians Tarsiers NWM OWM Apes



Beyond Linear Sequence Comparisons

Schmitz, Ohme, and Zischler, 2001 Genetics 157:777-784.
PCR survey of 118 intronic SINE element positions, sequence phylogenetically informative ones



Crustaceans

Insects

Arthropods . . .

Cheliceriforms

Myriapods

Polyphyletic

? ?
? ?
? ?
?

Myriapoda
Insecta
Onychophora
Chelicerata
Crustacea

} Uniramia

Mandibulate

Myriapoda
Insecta
Crustacea
Chelicerata
Onychophora

} Mandibulata

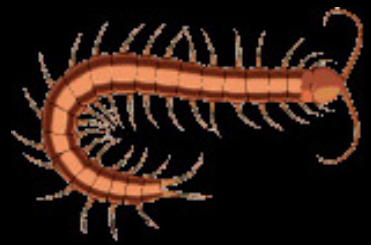
"TCC"

Myriapoda
Insecta

} Terrestrial

Crustacea
Chelicerata
(Trilobita)
Onychophora

} "TCC"



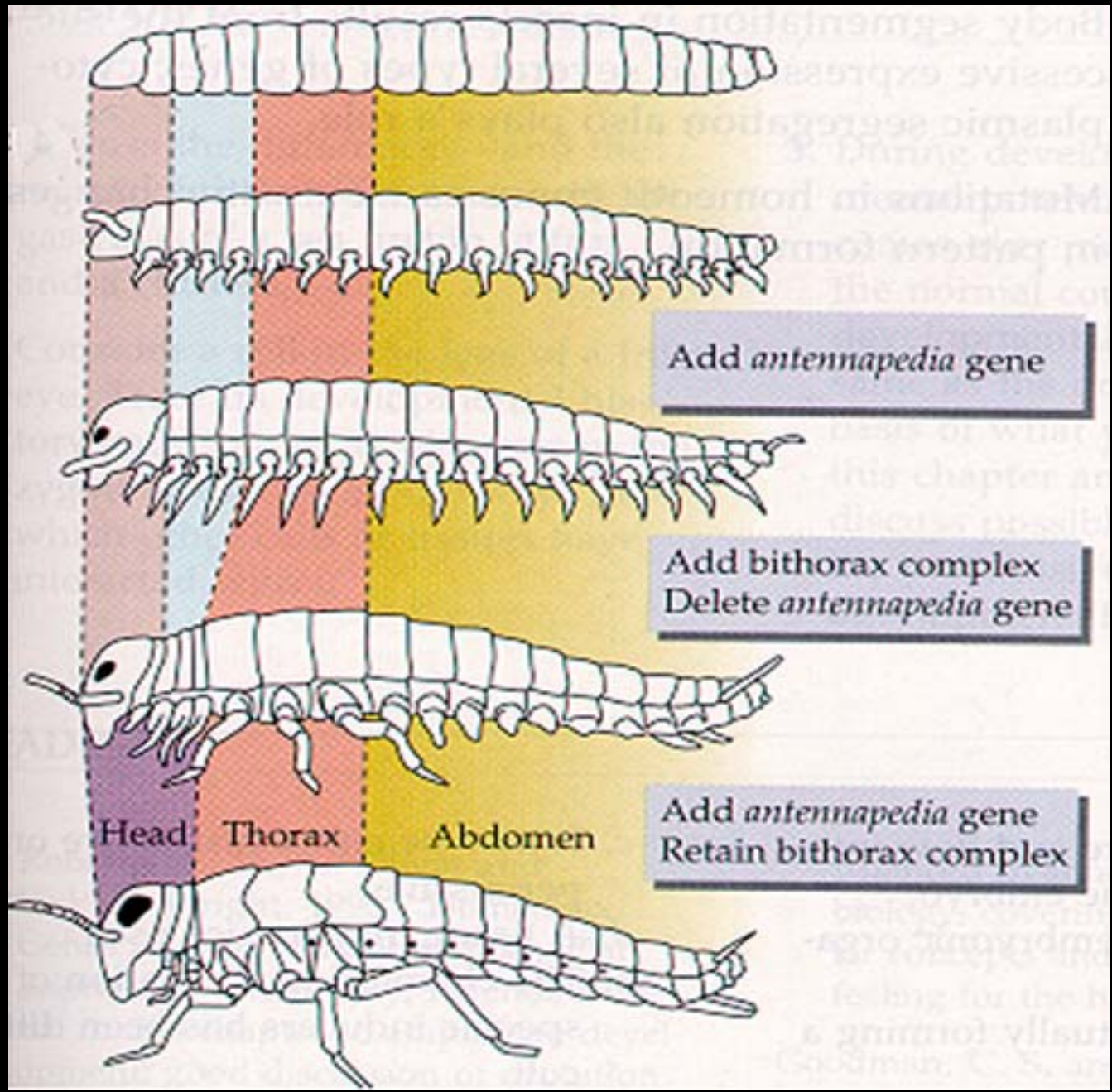
Annelid

Onychophoran

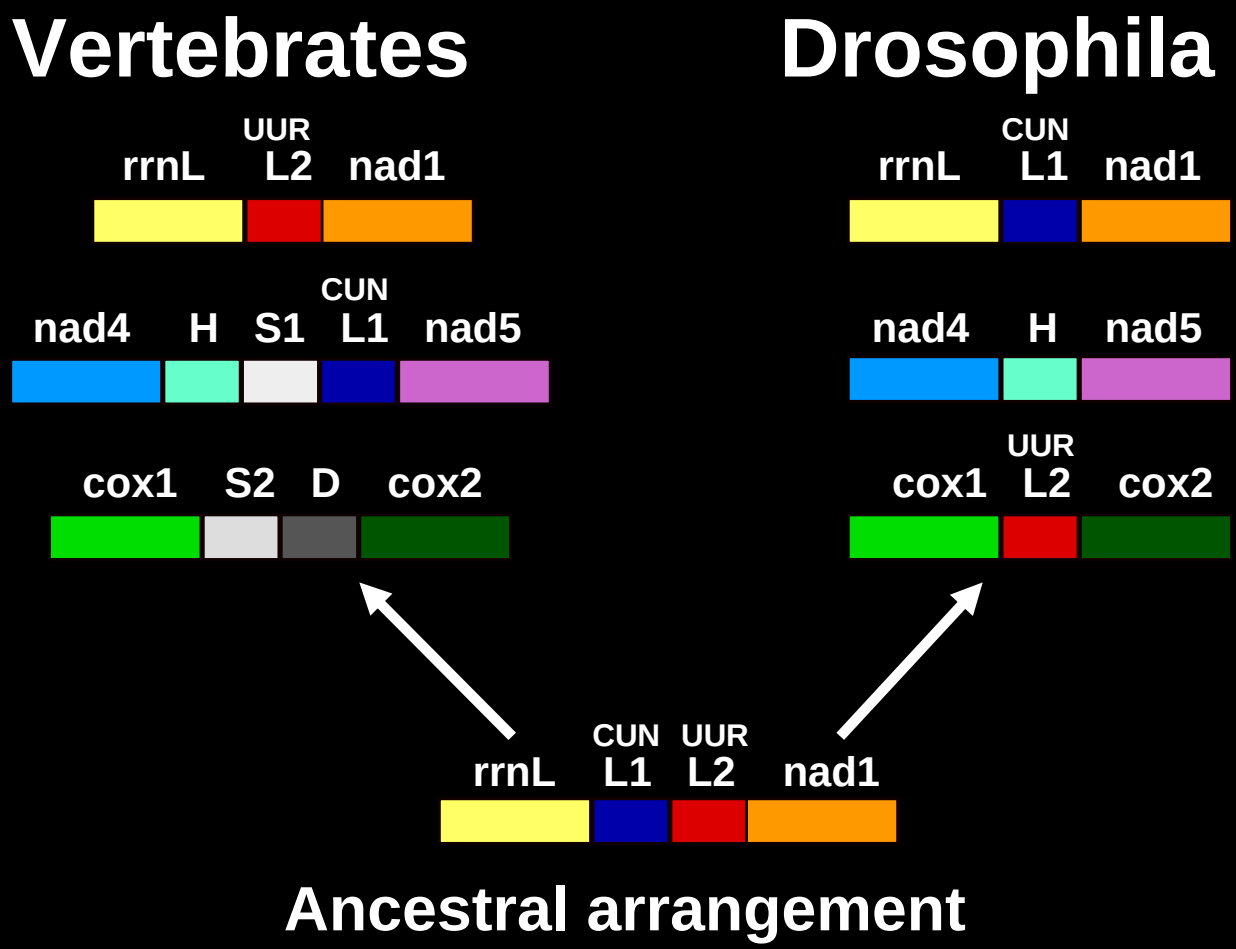
Myriapod

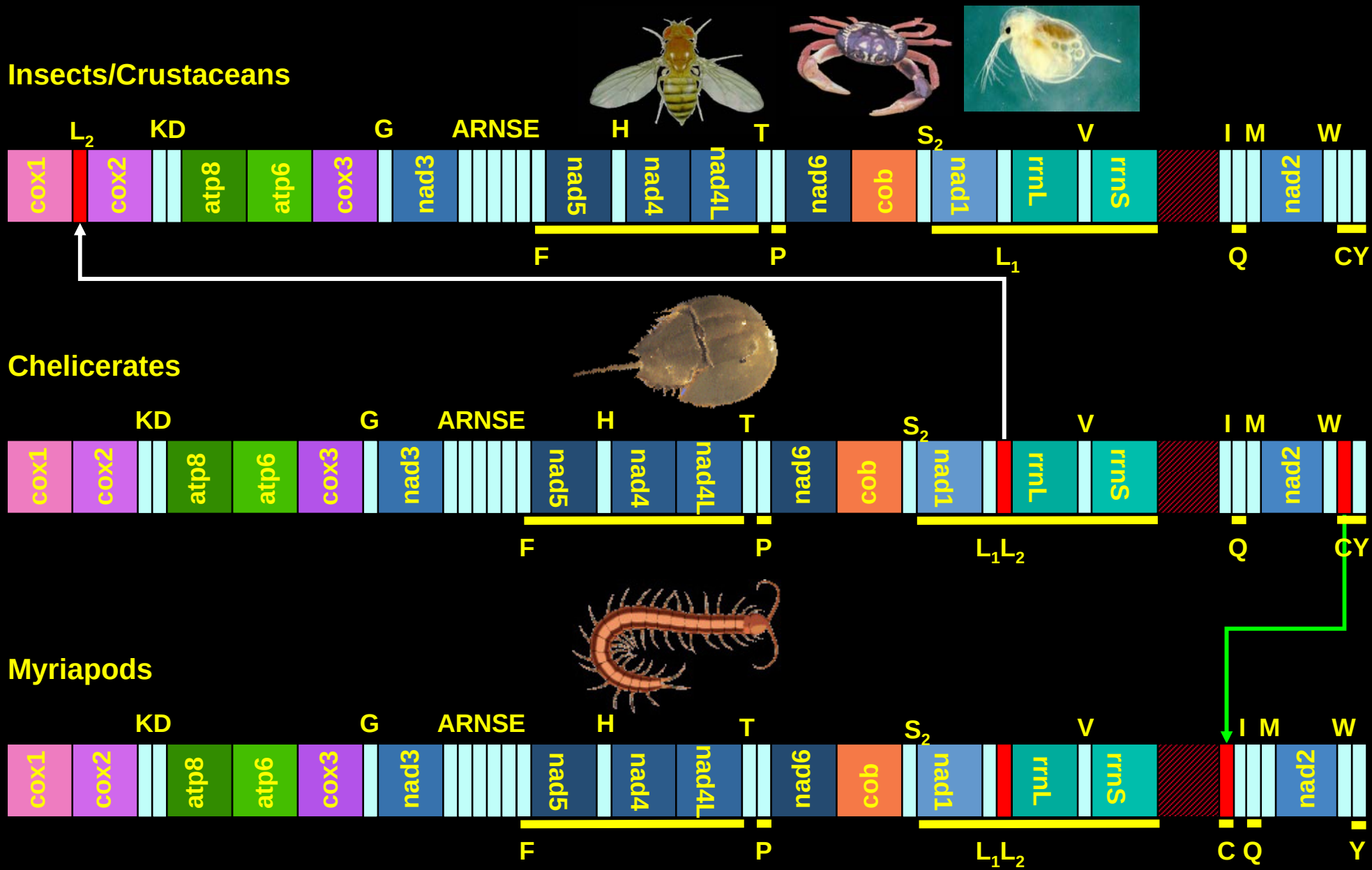
Wingless insect

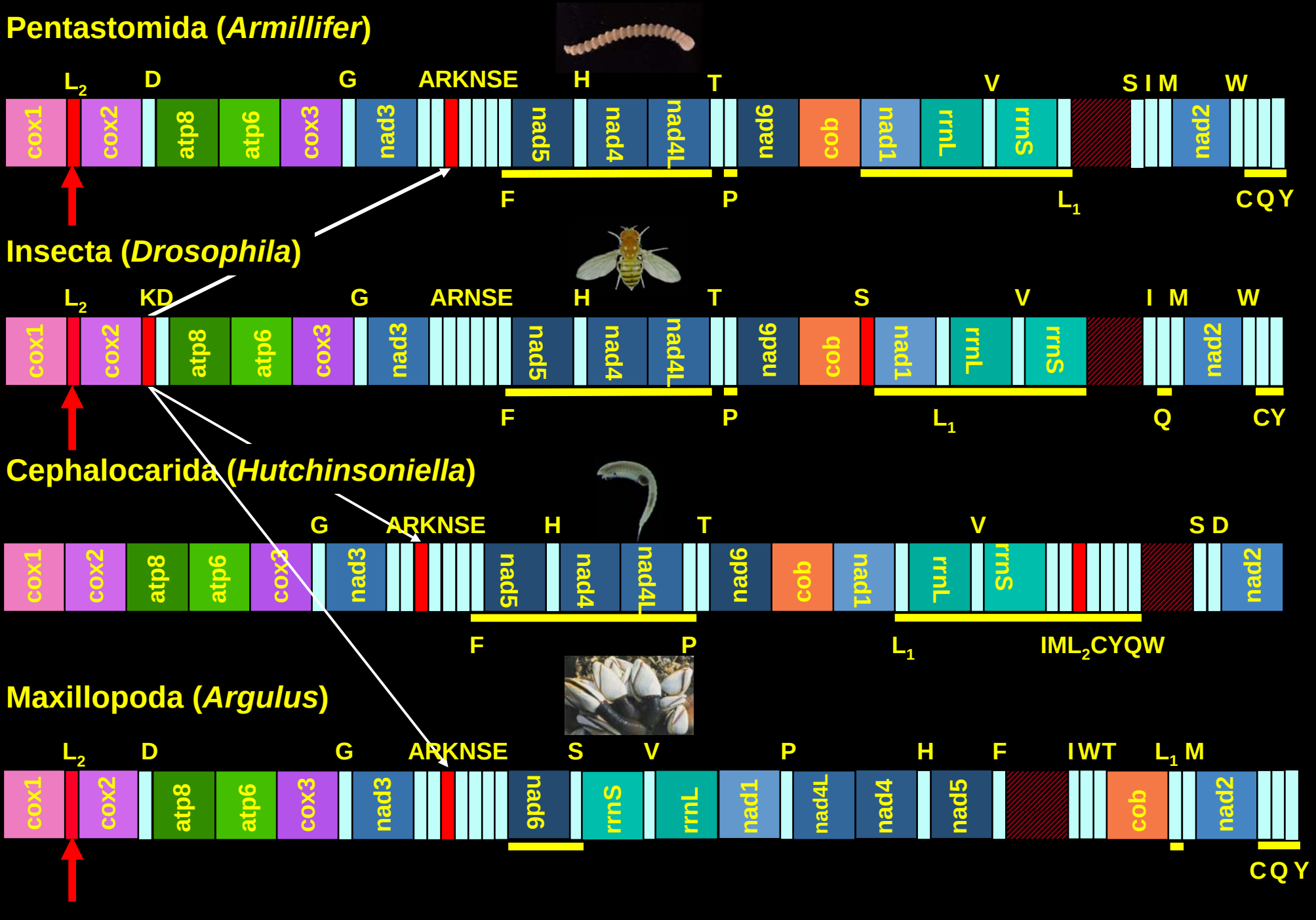
Winged insect



A little background on mtDNAs . . .




Insects/Crustaceans
L2
KD
G
ARNSE
H
T
S2
V
I M
W
cox1
cox2
atp8
atp6
cox3
nad3
nad5
nad4
nad4L
nad6
cob
nad1
rrnL
rrnS
nad2
F
P
L1
Q
CY
Chelicerates
KD
G
ARNSE
H
T
S2
V
I M
W
F
P
L1L2
Q
CY
Myriapods
KD
G
ARNSE
H
T
S2
V
I M
W
F
P
L1L2
C Q
Y


Pentastomida (*Armillifer*)
L2 D G ARKNSE H T V S I M W
cox1 cox2 atp8 atp6 cox3 nad3 nad5 nad4 nad4L nad6 cob nad1 rrnL rrnS nad2
F P L1 C Q Y
Insecta (*Drosophila*)
L2 KD G ARNSE H T S V I M W
cox1 cox2 atp8 atp6 cox3 nad3 nad5 nad4 nad4L nad6 cob nad1 rrnL rrnS nad2
F P L1 Q CY
Cephalocarida (*Hutchinsoniella*)
G ARKNSE H T V S D
cox1 cox2 atp8 atp6 cox3 nad3 nad5 nad4 nad4L nad6 cob nad1 rrnL rrnS nad2
F P L1 IML2CYQW
Maxillopoda (*Argulus*)
L2 D G ARKNSE S V P H F IWT L1 M
cox1 cox2 atp8 atp6 cox3 nad3 nad6 rrnS rrnL nad1 nad4L nad4 nad5 cob nad2
C Q Y

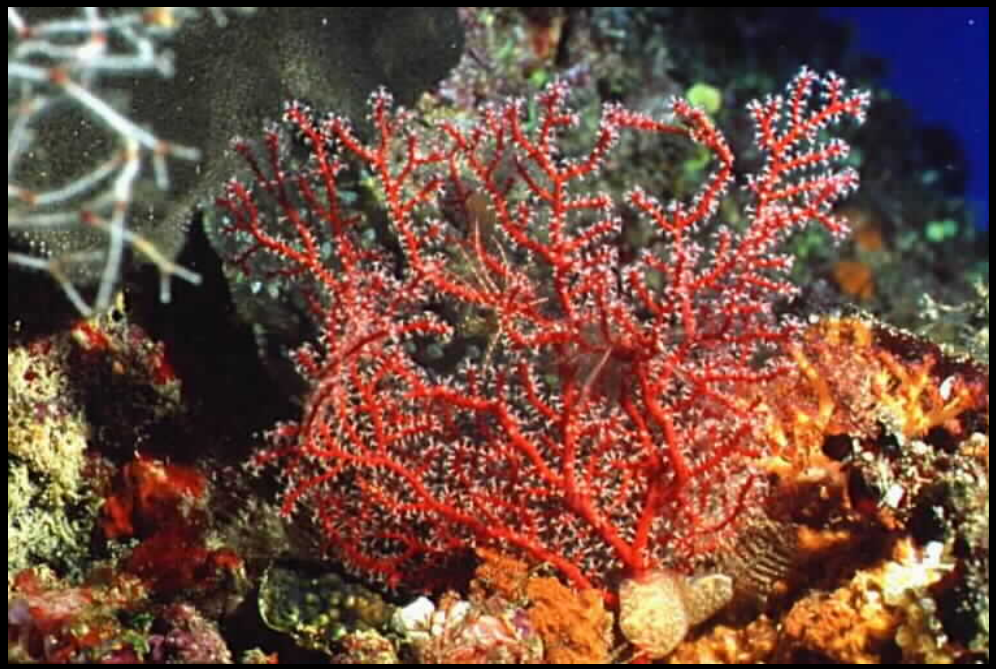
A brief sample of some other projects our Evolutionary Genomics Department has underway

Coral Reef Genomics

Coral reefs harbor a great deal of marine biodiversity and corals sequester CO_2

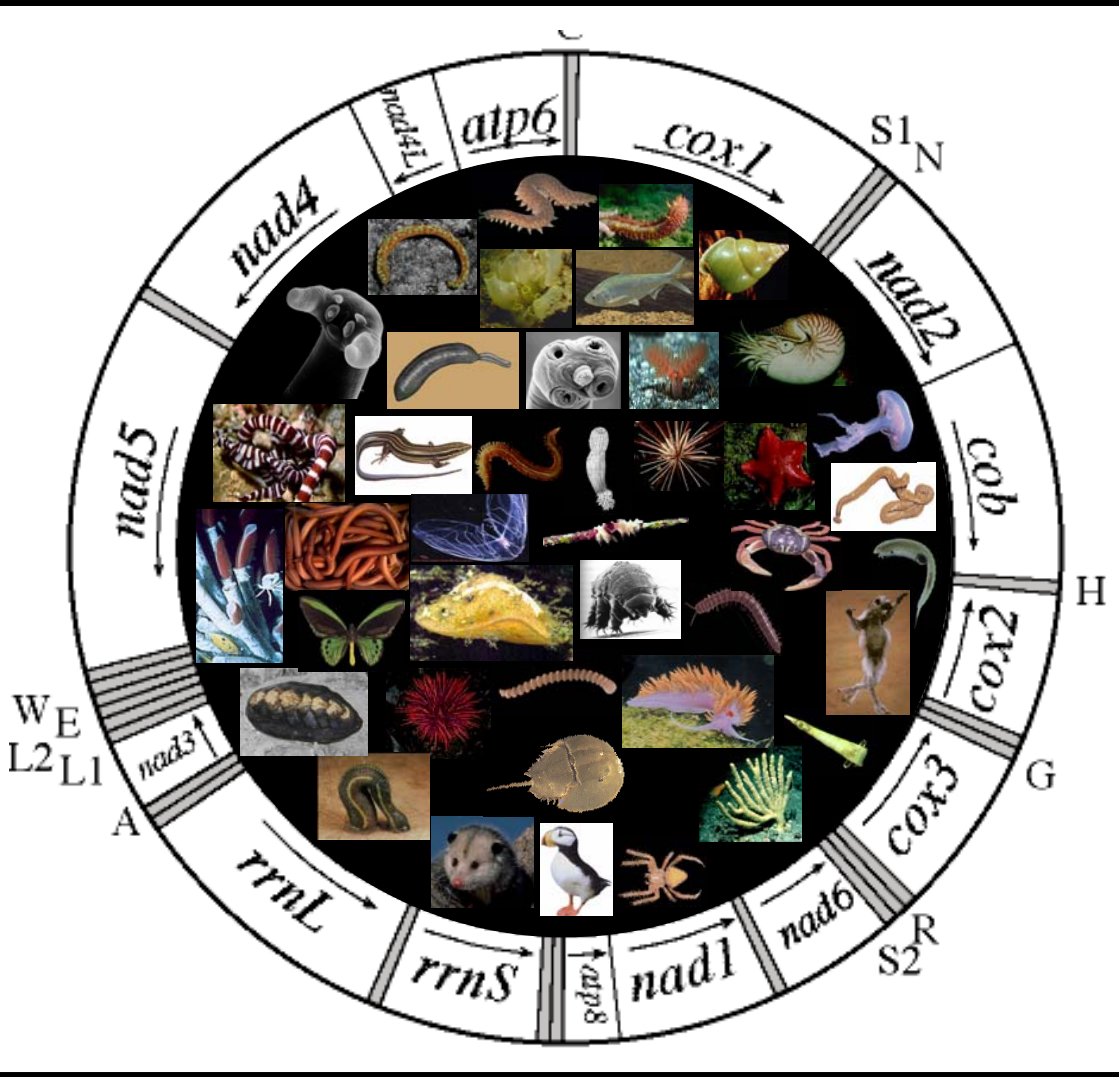
Global warming seems to be causing coral bleaching due to the loss of their symbiotic zooxanthellae

The goal is to understand the roles of various genes in mediating the coral-zooxanthellae relationship and its loss



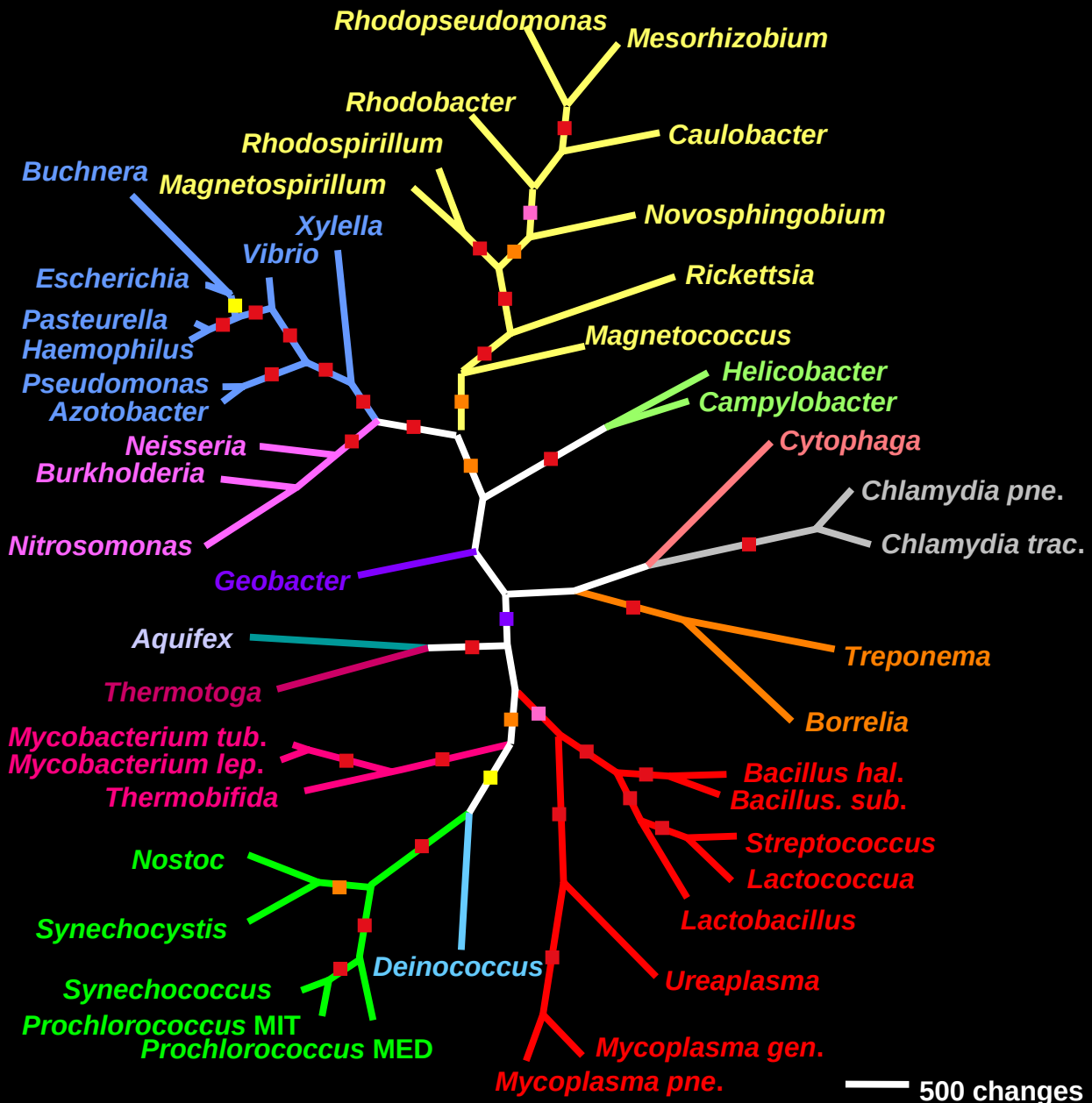
Evolution of stalk-eyed flies

Organelle genomes



Chloroplast genomics

Mitochondrial genomics



Reconstruct phylogeny using large scale genome sequence comparisons

39 orthologous ribosomal proteins

18 newly sequenced bacterial genomes representing many previously unsampled lineages

- Alpha proteobacteria
- Beta proteobacteria
- Delta proteobacteria
- Gamma proteobacteria
- Epsilon proteobacteria
- Bacteroidetes
- Chlamydiae
- Spirochaetes
- Firmicutes
- Deinococcus-Thermus
- Cyanobacteria
- Actinobacteria
- Thermotogae
- Aquificae

- 100% bootstrap support by Neighbor Joining (NJ), Minimum Evolution (ME) and Maximum Parsimony (MP)
- 85-100% bootstrap support by NJ, ME and MP
- 85-100% bootstrap support by NJ and ME
- 85-100% bootstrap support by NJ
- 85-100% bootstrap support by MP

Comparing interesting portions of many genomes

Acknowledgements:

Production and Computational Genomics Departments of the JGI

The members and collaborators of the Evolutionary Genomics Department and the staff and management of the JGI

- Acknowledgment:

  This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC03-76SF00098 and Los Alamos National Laboratory under contract No. W-7405-ENG-36.