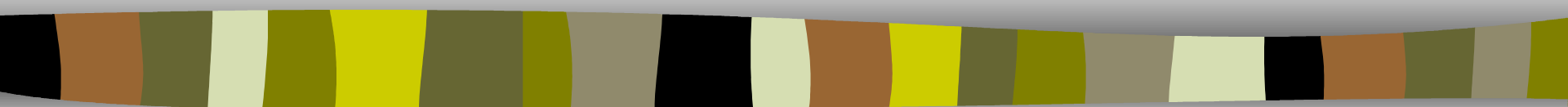


Comparative Genomics: Phylogenetic *footprinting* and *shadowing*



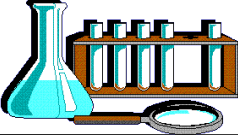

*Sequencing and comparison of yeast species to identify
genes and regulatory elements,*
Manolis Kellis, Nick Patterson, Matthew Endrizzi,
Bruce Birren & Eric S. Lander
Nature **423**, 241 - 254 (15 May 2003)

Understanding genomes

Yeast genome structure

- ~10Mbp
- 70% coding, 15% of intergenic region is regulatory

Elements to identify

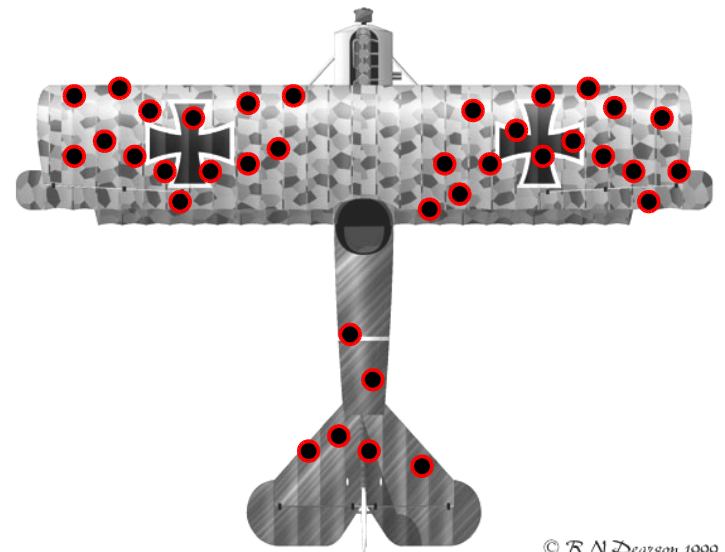
	Genes	Regulatory elements
	cDNA	Systematic mutations upstream
	<i>de novo</i> / other organisms	Clustering / common motif search
Results	4800-6400	~60 motifs

Patterns of bullet holes

Bullet holes in planes after a combat

Aim: protect vulnerable areas

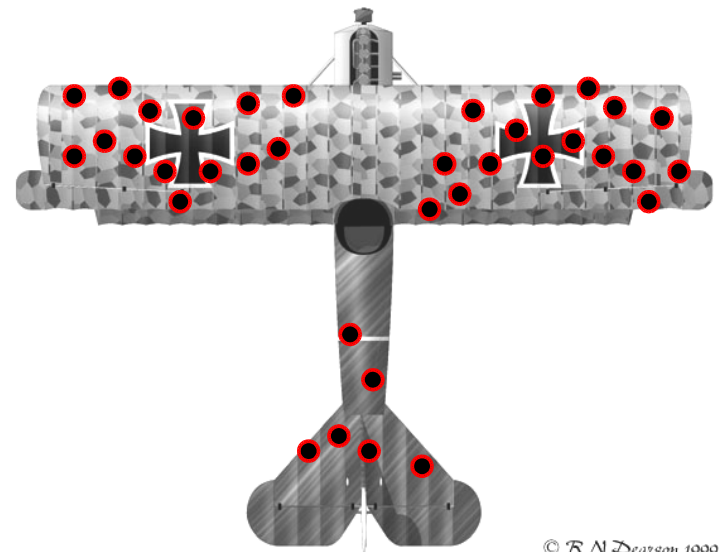
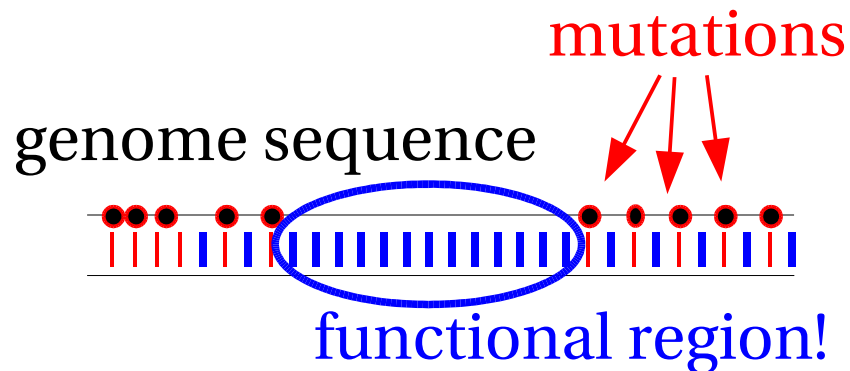
Abraham Wald was asked to analyse patterns of bullet holes



Patterns of bullet holes

Phylogenetic footprinting:

- Airplane – genome
- Bullet hole – mutation
- Combat – natural selection





Comparative genomics: intro

More than one genome
in hand

– no additional information

Look at strongly
conserved regions

```
...CGATGACTATTA...  
...CGATGACTA-TA...  
...C---GAGTATTA...  
...CGATGACTATTA...
```

“Because evolution relentlessly tinkers with genome sequence and tests the results by natural selection, such [functional] elements should stand out by virtue of having a greater degree of conservation”

quantity of data ->
quality of results



We start with...

Input: 4 *Saccharomyces*

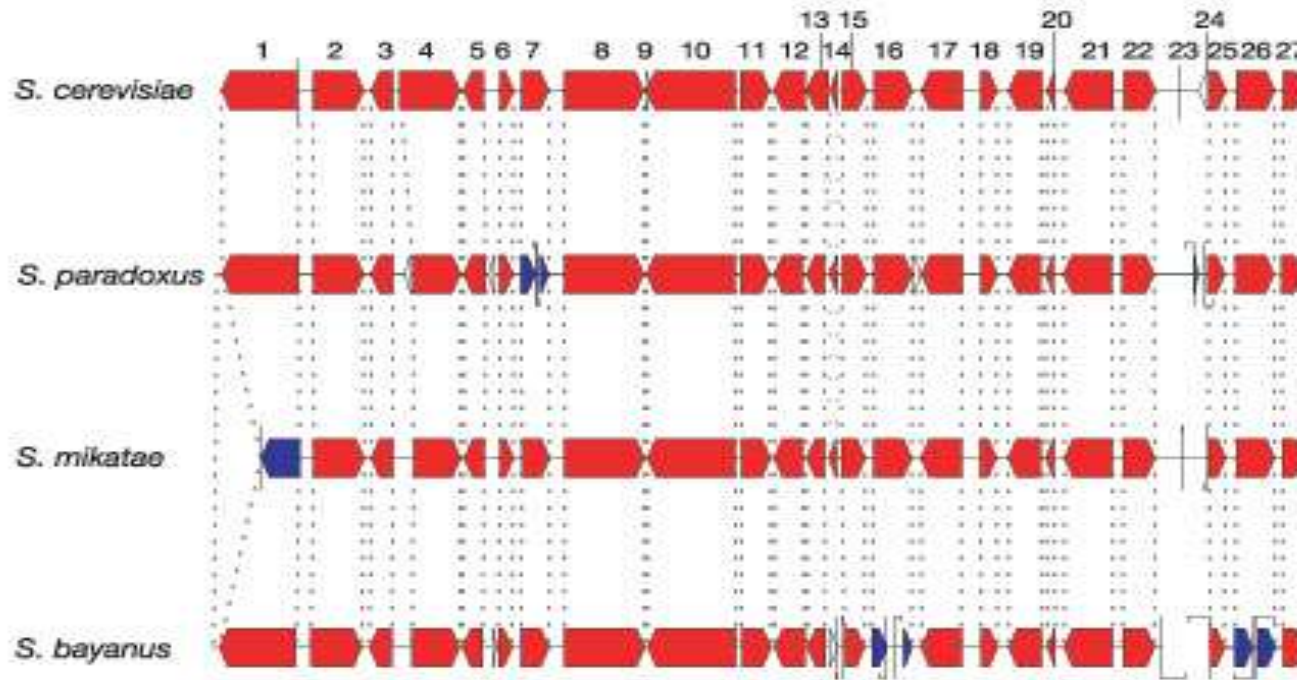
- *S. cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. bayanus*
- 5 to 20 mln years since the divergence of species
- **Divergent enough** to introduce noise where needed
- **Related enough** for orthologues to be easily detectable

Genome alignment

...CGATGACTATTA..
...CGATGACTA-TA..
...C---GAGTATTA..
...CGATGACTATTA..

Anchoring with ORFs

Aligning region in between



50kb segment; arrow – direction of ORF, red – 1-1 match; blue – multiple match

Genome evolution - nucleotides

S. cerevisiae . . . CGTTGACTATTA . . .
Other yeast . . . CGATGGCTA-TA . . .

Nucleotide identity:

	<i>S. paradoxus</i>	<i>S. mikatae</i>	<i>S. bayanus</i>
coding	90%	85%	80%
intergenic	80%	70%	60%

2x faster!



Genome evolution - nucleotides

Measures of variation for **all species** (*multiple alignment*)

S. cerevisiae . . . **CGATGCCTATTC** . . .
S. paradoxus . . . **CGATGGCTA**-TA . . .
S. mikatae . . . **C**---GAG**TATTA** . . .
S. bayanus . . . **CGATTACTATGA** . . .

	identity	gap	frame shift
coding	60%	1.3%	0.14%
intergenic	30%	14%	10.2% +stop codons
difference	2x	10x	75x



Genes identification

Solution:

- Dummy method: long region without stop codon becomes putative ORF
- We reject putative ORF based on the gene region characteristics

Expected quality:

- High sensitivity of dummy method
- High specificity thanks to divergence



Genes identification

Results: gene catalogue revision

- From ~4000 named genes only 15 rejected. One mistake.
- 5538 genes (before: 6062)
- Different start/stop codons for 5% of genes
- ~60 new introns



Regulatory elements identification

Regulatory motifs

- Short 6-15bp
- Hard to identify

A small investigation of Gal4 motif shows conservation rates:

Motif:	Random	Gal4	Difference
Intergenic	3%	12.5%	4x
Coding	7%	3%	
Difference	$\frac{1}{2}x$	4x	

Detection of regulatory elements

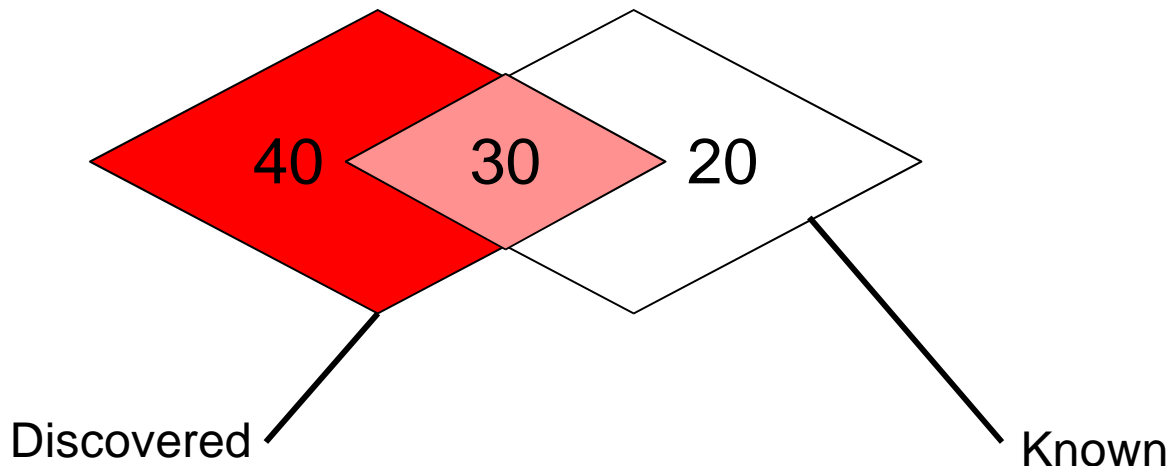
Make up random motif of form:

CGAxxxxxTGG

any nucleotides

Check validity with the test

Results





Summary

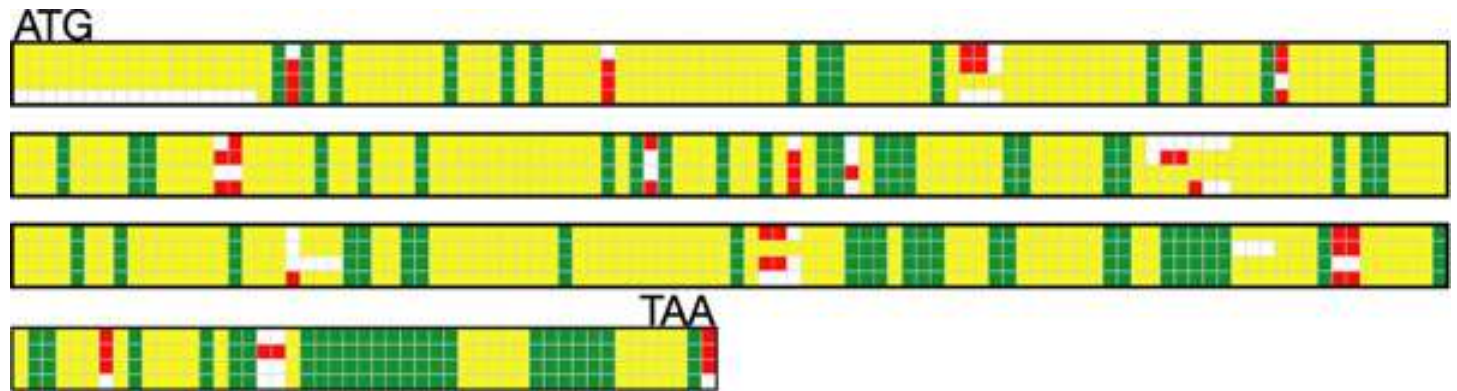
Results

- Exact count of 5538 genes
- Better gene boundaries
- New introns identified
- New motifs found

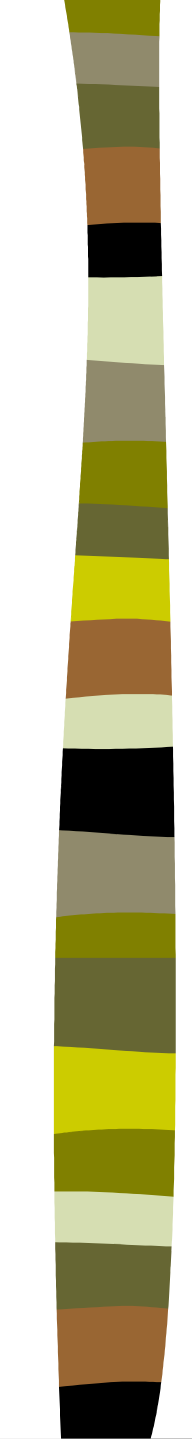
Human



- Coding part of genome 2% (70% in yeast), regulatory motifs 3% of intergenic sequences (15% in yeast)
- repetitions

An Example: ORF rejected by the test

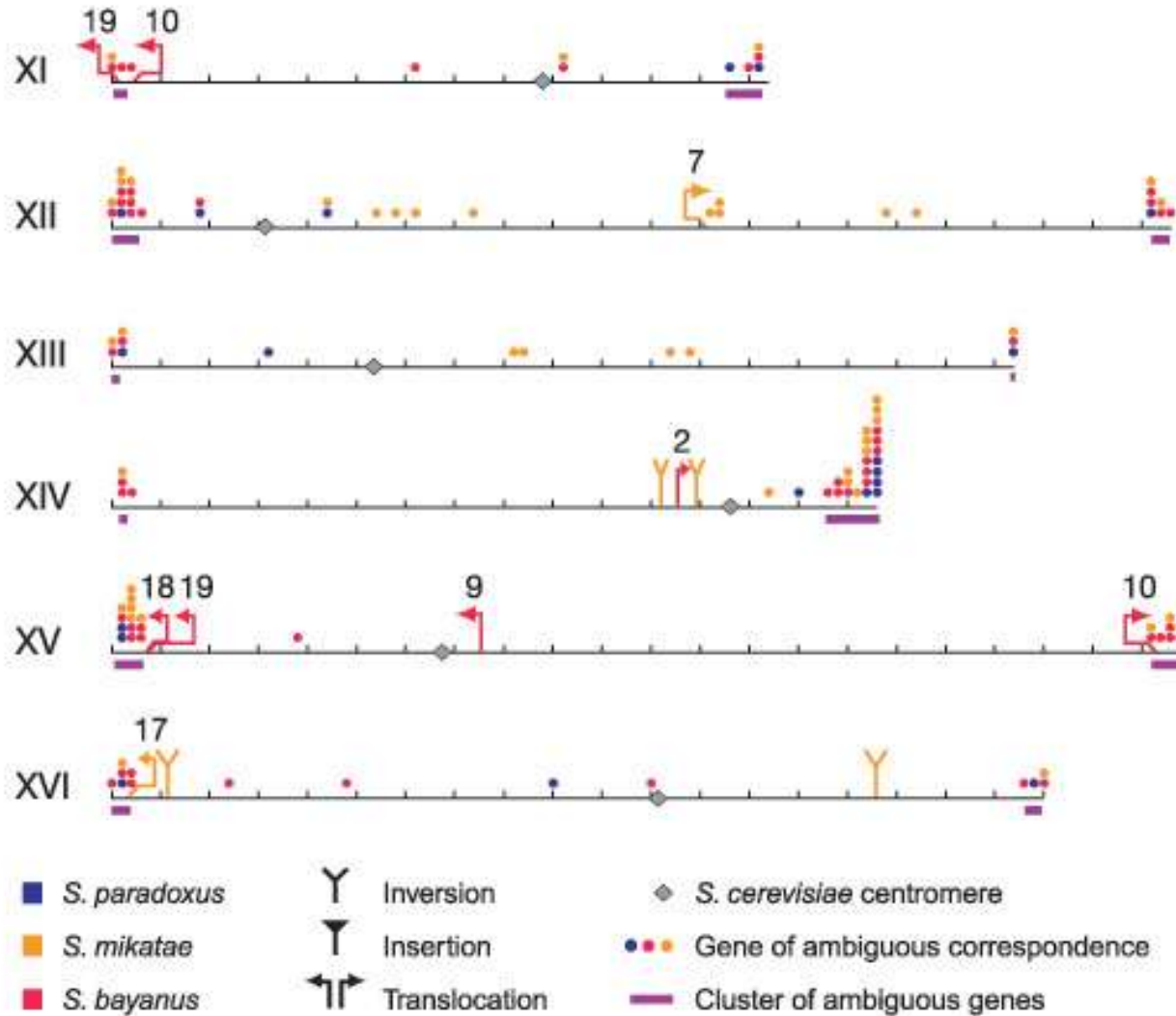


Motif recognition



	Gal4
<p>a motif</p> 	?
 <p>genes</p>	?
Difference	?

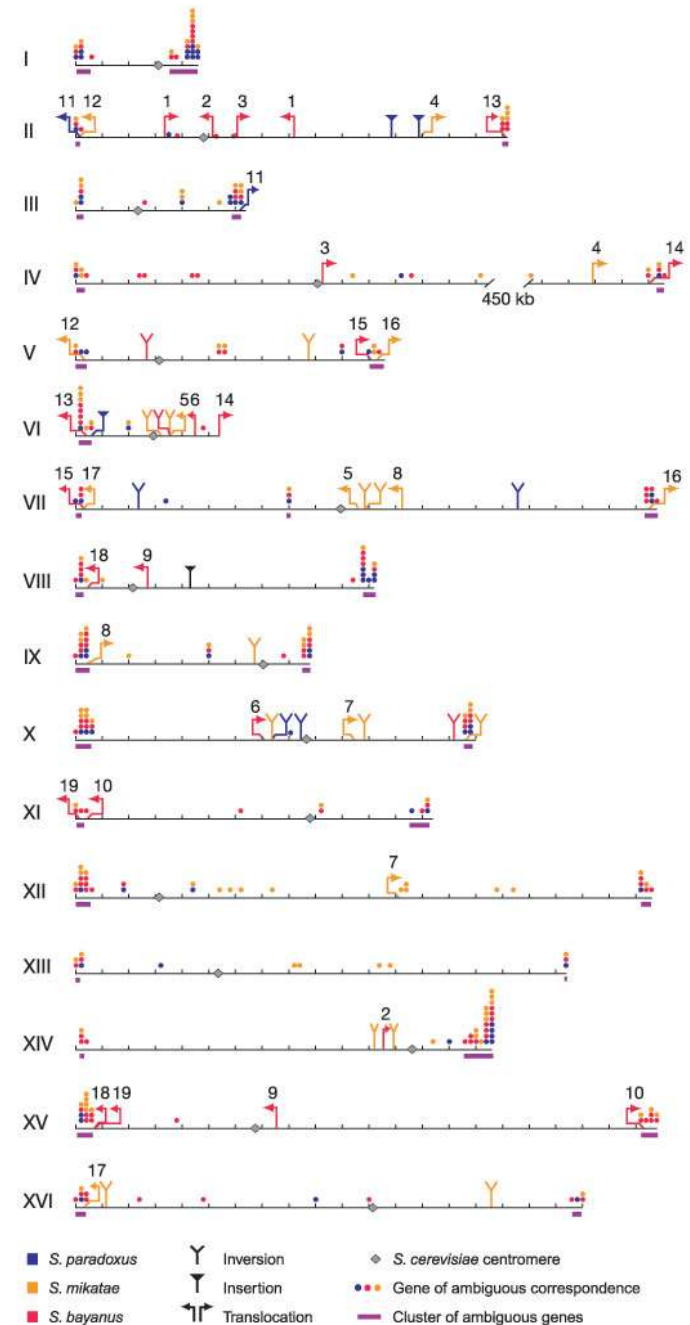
Genome evolution at large scale



Genome evolution at large scale

Telomeres: evolution's workshop

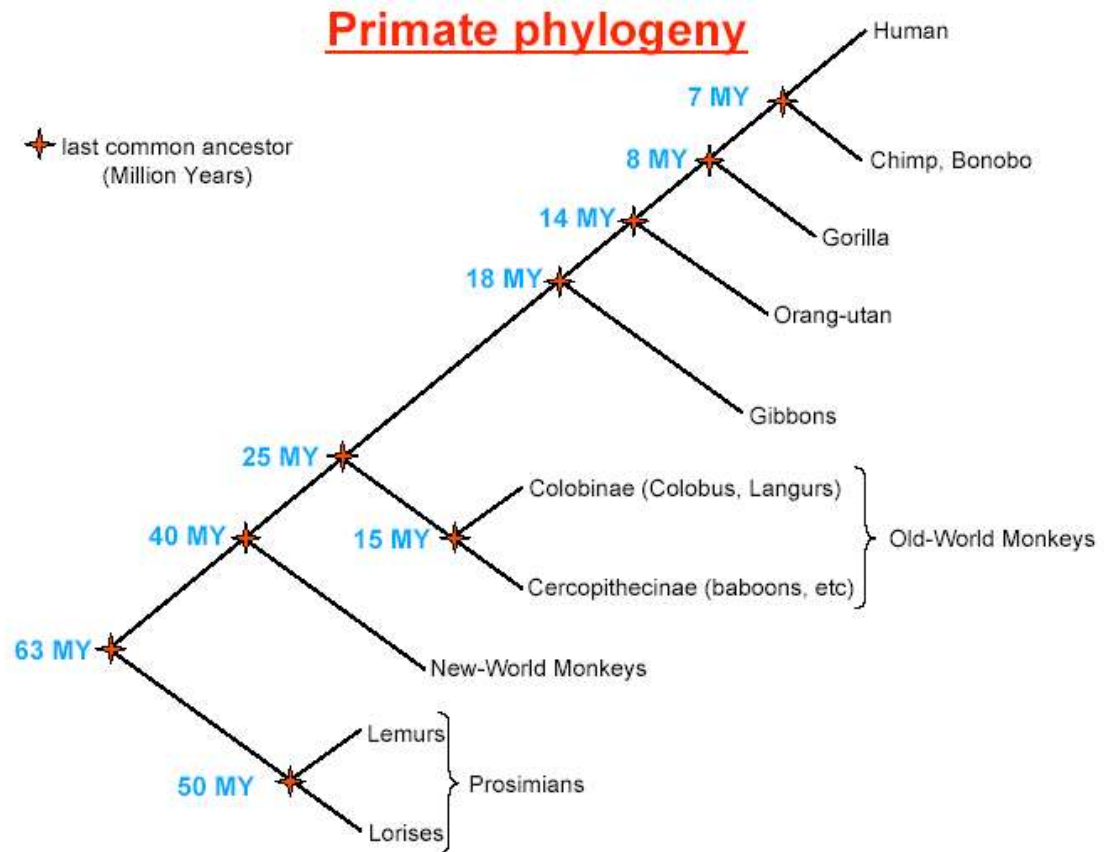
- Clusters of ambiguity
- 7-52kb from each end
- Translocations between telomeres
- Rapid evolution
- Observed in *Plasmodium falciparum* (antigenic variation)



Phylogenetic shadowing

A few closely related species

- Chimp
- Baboon
- monkeys



Phylogenetic shadowing

Additive divergence

