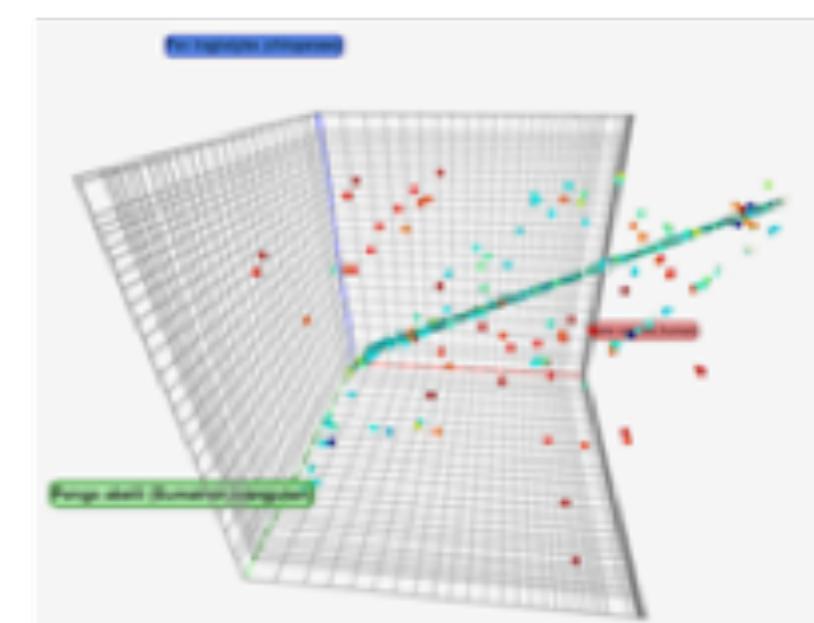
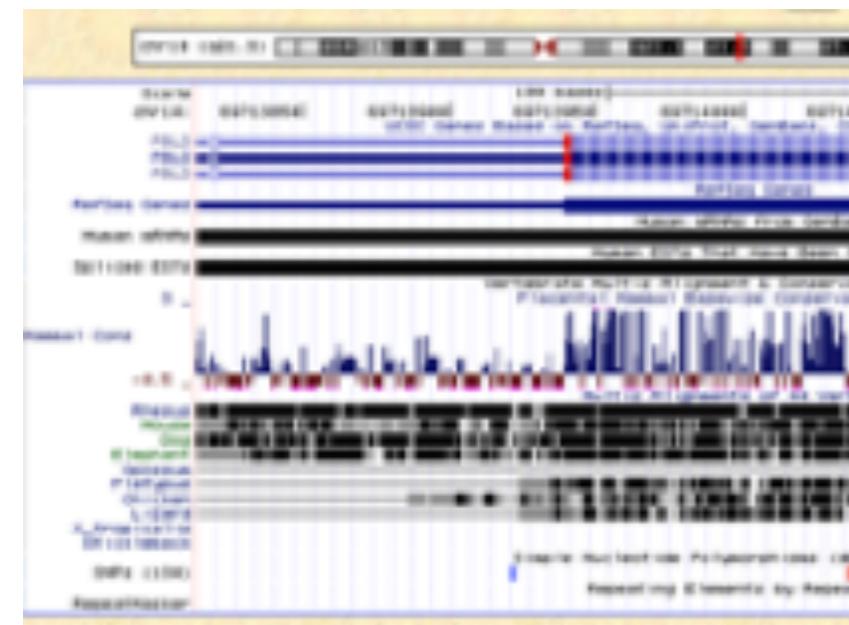


Computational Genomics

Introduction To Comparative Genomics





The Genomic Landscape: *circa 2010*

**Eric Green, M.D., Ph.D.
Director, NHGRI**



A Brief Introduction to the Logic for Comparative Genomics

**Mapping
the Human Genome**

~1990 to ~2000

**The Human
Genome Project**

**Sequencing
the Human Genome**

~1998 to ~2003

**Interpreting
the Human Genome
Sequence**

~2003 to ???

***Beyond*
The Human
Genome Project**

~3,000 bp (0.0001%) of Human Genome Sequence

TGCCCGGAACTTTCGGCTCTAAGGCTGTATTTGATATACGAAAGGCACATTTCCCTTCAAAATGCACCTTGCAAACGTAACAG
GAACCCGACTAGGATCATCGGGAAAAGGAGGAGGAGGAAGGCAGGCTCCGGGAAGCTGGTGGCAGCGGGCCTGGGTCTGGCGGACCCCTGA
CGCGAAGGAGGGTCTAGGAAGCTCTCCGGGAGCCGGTCTCCCGCCGGTGGCTTCTGTCCAGCGTTGCCAACTGGACCTAAAGAGAGG
CCGCGACTGTCGCCACCTGCCGGATGGGCCTGGTCTGGCGGTAGGACACGGACCTGGAAGGAGCGCGCGAGGGAGGGAGGCTGGAGTC
AGAATCGGAAAAGGGAGGTGCCGGCGAGGGAGCGAAGGAGGAGGAGGAAGGAGGAGCGGGAGGGTAGCTGGCGGGGTGCGTAGTGGTGG
GAAAGCCGCTAGAGCAAATTGGGCCGGACCAGGCAGCACTCGCTTTAACCTGGCAGTGAAGGCGGGAAAGAGCAAAGGAAGGGTGG
TGTGCGGAGTAGGGTGGGTGGGGAAATTGGAAGCAAATGACATCACAGCAGGTAGAGAAAAAGGGTAGCGGCAGGCACCCAGAGTAGTAG
GTCTTGGCATTAGGAGCTTGAGCCCAGACGCCCTAGCAGGGACCCCAGCGCCCGAGAGACCAGTCAGAGGGTCGCGCTCTGGAAAAGGCCAGCGT
TGTCTCCAAACTTTTCAGGTGAGAAGGTGGCCAACCGAGCTCGGAAAGACACGTGCCACGAAAGAGGGAGGGCGTGTATGGTTGG
TGGGTAAAGGAATAAGCAGTTTAAAAAGATGCGCTATCATTGTTGAAAGAAAATGTGGTATTGTTAGAATAAAACAGAAAGCATT
AGAAGAGATGGAAGAATGAACTGAAGCTGATTGAATAGAGAGCCACATCTACTTGCAACTGAAAAGTTAGAATCTCAAGACTCAAGTACGCTACT
ATGCACTTGTTTATTCATTCTAAGAAAATTTGTTAATAAGTACCTAAGTATGGTTATTGGTTTCCCCCTCATGCCTTGG
ACACTTGATTGTCTTGGCACATACAGGTGCCATGCCTGCATATAGTAAGTGCTCAGAAAACATTCTTGACTGAATTCAAGGAAACAAAAATT
TTGGGTAGGTAGAAAATATGCTAAAGTATTGTTATGAGACTGGATATATCTAGTATTGTCACAGGTAAATGATTCTCAAAAATTG
AAAGCAAATTGTTGAAATATTGTTAAAAAGTTACTTCACAAGCTATAAATTAAAAGCCATAGGAATAGATACCGAAGTTATATCAA
CTGACATTAAATAAAATTGTTACTCATAGCCTAATGTGATGAGCCACAGAACGCTGCAAACCTTAATGAGATTTTAAAATAGCATCTAAGTCGG
AATCTTAGGCAAAGTGTGTTAGATGTAGCACTTCATATTGAAGTGTTCTTGATATTGCATCTACTTGTTCCCTGTTATTACTGGTGTGA
ATGAATGAATAGGTACTGCTCTCTGGACATTACTGACACATAATTACCAATGAATAAGCATACTGAGGTATCAAAAAGTCAAATATGT
TATAAAATAGCTCATATATGTGTTAGGGGGAGGAATTAGCTTACATCTCTCTTATGTTAGTTCTCTGCATGTGCAGTTAACCTGGAAC
TCCGGTGCTAAGGAGAGACTGTTGGCCCTTGAAGGGAGAGCTCCTCCCTGGATGAGAGAGAAGGACTTTACTCTTGAATTATCTTTGTGT
TGATGTTATCCACCTTGTACTCCACCTATAAAATCGGTTATCTATTGATCTGTTCTAGTCCTTATAAAAGTCAAAATGTTAACCTGGCAT
AAATTATAGACTTTTAGCAGAGAACCTTGAAGGAACCTAAATGCCAACAGTCTAAAATGCAGTTTCAAGAAGAATGAATATTGATGGATA
GTTCTAAATACTAATGAACCTTAAAATAGCTTACTATTGATCTGTCAAAGTGGTTTATATAATTCTTTACAAATCACCTGACACATT
AATATAGGTTAAAAATGCTATCAGGCTGGTTGCAAAGAAAATGTATTACAAAGGCTGCTAAGTGTGTTAGAGCATACTCATTCTGTTCTCC
AAAATATTCTATAAGGTCTTAAGAATAGGTATGTTTAAAAGTTAAGTCCTACTATTATAGGAACGTACAATCACCTAAAATACCAATGA
TTACAAAATTCCTCTGGCCTCTGGACTGCAATTCTAAAAGTGTAAAAACATATTCTGCATTAAGTTAGGCAGTATTGCTTAGTTCAAA
GTGGTAGGCTTGGAGTCAGATTATTGATTCAAGATCCTACATCTACTGTTAGTAGCTCTGTTGCCTGAGGCAGGTCCCTAACATCTGTG
TGTGACTTGACCTTAAAATTGGAGACTGTCATAGGGTTAATCCCTGAGAAAATGAATGTGAAAAGTTAGCCTAATGTTAACTGCTATTATT
ATGGATTACCATATTCACATTCACTACAGTACATGCACCTGTTAATATAAGATGCTCAATTCTTGAGTATAATTGTGACTCTCAAT
CTGGATATGCAATGAGTGGGCCTGTATGAGAATTAAATTGAAAATTGTTACATGGCCTTACCAAGATATAAGGAAACACGTCACATG
TTCTATTGTATGTTAAATGCCTTAGAATTAACTTCTGAATAGGATCCCTCAGTTGAGAGTCATAAAAGAGTAAAATTATTGTT

The Human Genome... by the Numbers

~5% of Human Genome Sequence is Constrained Across Mammals (and Presumed Functional)

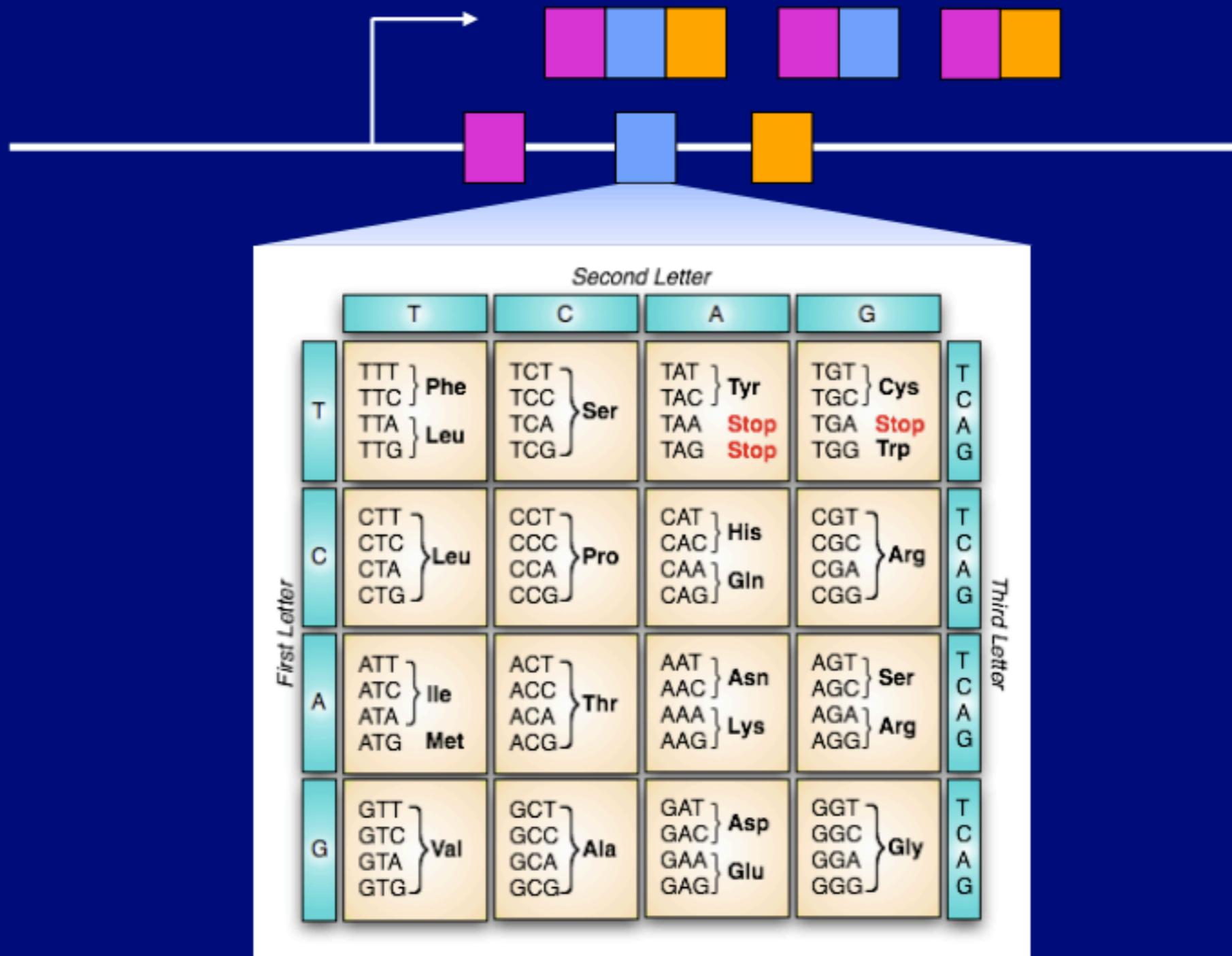
5% of 3B Bases = ~150M Bases

**Do NOT Yet Know the Position of these ~150M Functional Bases
Lower Bound for the Amount that is Functional**

~3,000 bp (0.0001%) of Human Genome Sequence

TGCCCGGGAACTTTGGCTCTCTAAGGCTGTATTTGATATACGAAAGGCACATTTCCCTTCAAAATGCACCTGCAAACGTAACAG
GAACCCGACTAGGATCATGGGAAAAGGAGGAGGAGGAAGGCAGGCTCCGGGAAGCTGGTGGCAGCGGGCCTGGCTGGCGGACCTGA
CGCGAAGGAGGGCTAGGAAGCTCTCCGGGAGCCGGTTCTCCGCCGGTGGCTCTCTGTCCTCCAGCGTTGCCAACTGGACCTAAAGAGAGG
CCGCGACTGTCGCCAACCTGCGGGATGGGCCTGGTGGCTGGCGTAAGGACACGGACCTGGAAGGAGCGCGCGAGGGAGGGCTGGAGTC
AGAATCGGGAAAGGGAGGTGCGGGCGAGGGAGCGAAGGAGGAGGAGGAAGGAGCGGGAGGGTGCTGGCGGGGTGCGTAGTGGTGG
GAAAGCCGCTAGAGCAAATTGGGGCGGACCAGGCAGCACTCGGCTTTAACCTGGCAGTGAAGGCAGGGAAAGAGCAAAGGAAGGGTGG
TGTGGAGTAGGGTGGTGGGGAAATTGGAAGCAAATGACATCACAGCAGGTCAAGAGAAAAAGGGTTGAGCGGCAGGCACCCAGAGTAGTAG
GTCTTGGCATTAGGAGCTTGAGCCCAGACGCCCTAGCAGGGACCCAGCGCCGAGAGACCATGCAGAGGTGCCCTGGAAAAGGCCAGCGT
TGTCTCAAACCTTTTCAGGTGAGAAGGTGCCAACCGAGCTCGAAAGACACGTGCCACGAAAGAGGGAGGGCTGTATGGTTGG
TGGGTAAAGGAATAAGCAGTTTAAAAAGATGCGCTATCATTATTGTTGAAAGAAAATGTGGTATTGAGAATAAAACAGAAAGCATT
AGAAGAGATGGAAGAATGAAGCTGATTGAATAGAGAGCCACATCTACTTGCACACTGAAAAGTTAGAATCTCAAGACTCAAGTACGCTACT
ATGCACTTGTATTTCATTCTAAGAAACTAAAAACTTGTAAATAAGTACCTAAGTATGGTTATTGGTTTCCCCCTCATGCCTTGG
ACACTTGATTGTCTTGGCACATACAGGTGCCATGCCTGCATATAGTAAGTGCTCAGAAAACATTCTTGAUTGAATTGCCAACAAAAATT
TTGGGTAGGTAGAAATATGCTAAAGTATTGTTATGAGACTGGATATATCTAGTATTGTCACAGGTAAATGATTCTTCAAAATTG
AAAGCAAATTGTTGAAATATTATTGAAAAAGTTACTTCACAAGCTATAAATTAAAAGCCATAGGAATAGATAACGAAGTTATATCAA
CTGACATTAAATAATTGTATTCATAGCCTAATGTGATGCCACAGAAGCTGCCAAACTTAAATGAGATTTTAAAATAGCATCTAAGTCGG
AATCTTAGGCAAAGTGTGTTAGATGTGACTCATATTGAAGTGTCTTGGATATTGCATCTACTTGTCTGTATTATACTGGTGTGA
ATGAATGAATAGGTACTGCTCTCTGGACATTACTGACACATAATTACCAATGAATAAGCATACTGAGGTATCAAAAGTCAAATATGT
TATAAATAGCTCATATATGTGTAGGGGGAAAGGAATTAGCTTCACATCTCTTGTATTCTGCATGTGCAGTTAACCTGGAAC
TCCGGTGCTAAGGAGAGACTGTTGCCCTGAGGAGAGCTCCCTGTGGATGAGAGAGAAGGACTTACTCTTGGATTATCTTGT
TGATGTTATCCACCTTGTACTCCACCTATAAAATCGGCTTATCTATTGATCTGTTCTAGTCCTATAAAAGTCAAATGTTAACCTGGCAT
AAATTATAGACTTTTAGCAGAGAACCTTGAGGAACCTAAATGCCAACAGTCTAAAATGCAGTTTCAAGAAGAATGAATATTCACTGGATA
GTTCTAAATACTAATGAACCTTAAAGAAAATGCTACTATTGATCTGTCAAAGTGGTTTATATAATTCTTTACAAATCACCTGACACATT
AATATAGTTAAAAATGCTATCAGGCTGGTTGCAAAGAAAATGTATTACAAAGGCTGTAAGTGTAAAGAGCATACTCTGTTCTCC
AAAATATTCTATAAGGTGCTTAAAGAATAGGTATGTTTAAAAGTTAGTCCTACTATTATAGGAACCTGACAATCACCTAAACATGAA
TTACAAACTCCTCTGGCCTCTGGACTGCAATTCTAAAAGTGTAAAAACATATTCTGCATTAAGTTAGGCAGTATTGCTTAGTTCAAA
GTGGTAGGCTTGGAGTCAGATTATTGATTCACTACATCTACTGTTAGTAGCTCTGTTGCCTGAGGCAGGTCCCTAACATCTGTG
TGTGACTTGACCTTAAAATTGGAGACTGTCATAGGGTTAATCCCTGAGAAAATGAATGTGAAAAGTTAGCCTAATGTTAACTGCTATTATT
ATGGATTACCATATTTCACATTCACTACAGTACATGCACCTGTTAATATAAGATGCTCAATTCTTGTAGTATAATTGACTCTCAAT
CTGGATATGCAATGAGTGGCCTGTATGAGAATTAAATTGAGAAAATTGTGTTCACATGGCCTTACCAAGATACAGGAAACACGTACATG
TTCTATTGTATGTTAAATGCCTTAGAATTAACTTCTGAATAGGATCCCTCAGTTGAGAGTCATAAAAGAGTAAAATTATTGGTAT

Coding Sequences (i.e., Genes)



The Genetic Code

The Human Genome... by the Numbers

~5% of Human Genome Sequence is Constrained Across Mammals (and Presumed Functional)

5% of 3B Bases = ~150M Bases

**Do NOT Yet Know the Position of these ~150M Functional Bases
Lower Bound for the Amount that is Functional**

~1.5% Encodes for Protein (Genes)

Corresponds to ~18-22K Genes

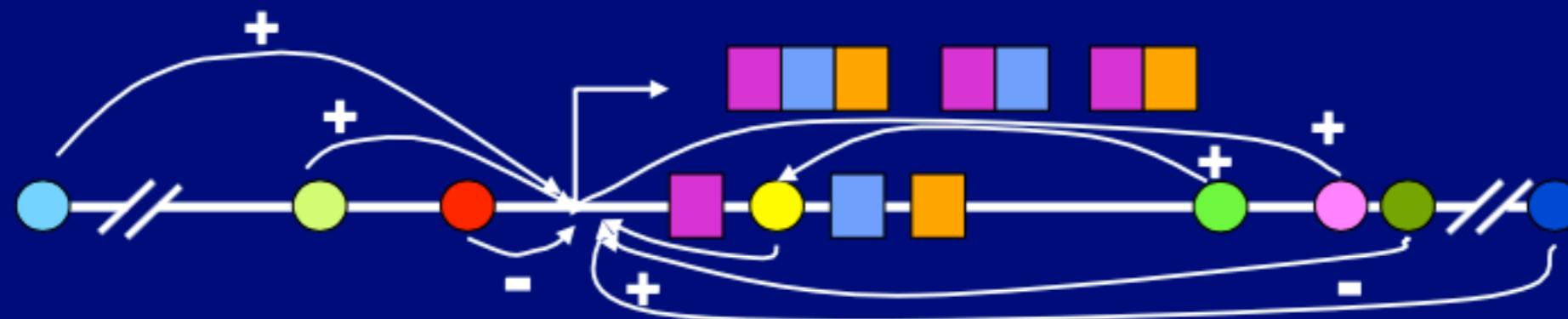
Many More than ~22K Different Proteins

Good Inventory at Present

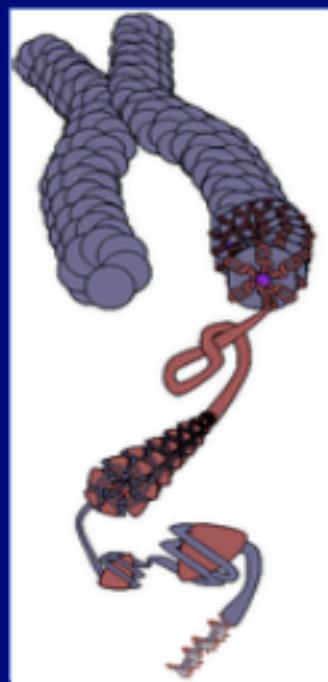
~3,000 bp (0.0001%) of Human Genome Sequence

TGCCCGGAACTTTGGCTCTCAAGGCTGTATTTGATATA
CGAACCCGACTAGGATCATCGGGAAAAGGAGGAGGAGGAAGGCAGGCAGGCTCCGGGGAAAGCTGGTGGCAGCGGGTCCTGGGTCTGGCGGACCCCTGA
CGCGAAGGAGGGTCTAGGAAGCTCTCCGGGGAGCCGGTTCTCCCGCCGGTGGCTTCTGTCCAGCGTTGCCAACTGGACCTAAAGAGAGG
CCGCGACTGTCGCCACCTGCAGGGATGGGCCTGGTGCTGGCGGTAGGACACGGACCTGGAAGGAGCGCGCGAGGGAGGGAGGCTGGGAGTC
AGAATCGGGAAAGGGAGGTGCGGGGGCGAGGGAGCGAAGGAGGAGGAGGAAGGAGGAGGGAGGGTGCTGGCGGGGTGCGTAGTGGGTGGA
GAAAGCCGCTAGAGCAAATTGGGGCGGACCAGGCAGCACTCGGCTTTAACCTGGCAGTGAAGGCGGGAAAGAGCAAAGGAAGGGTGG
TGTGCGGAGTAGGGTGGGTGGGGGAATTGGAAGCAAATGACATCACAGCAGGTAGAGAAAAAGGGTTGAGCGGCAGGCACCCAGAGTAGTAG
GTCTTGCCATTAGGAGCTTGAGCCCAGACGGCCCTAGCAGGGACCCCAGCGCCCGAGAGACCATGCAGAGGTGCCCTTGAAAAGGCCAGCGT
TGTCTCCAAACTTTTCAGGTGAGAAGGTGGCCAACCGAGCTCGGAAAGACACGTGCCACGAAAGAGGGAGGGCGTGTATGGTTGGGTT
TGGGGTAAAGGAATAAGCAGTTTAAAAAGATGCGCTATCATTGTTGAAAGAAAATGTGGTATTGAGATAAAAACAGAAAGCATTAA
AGAAGAGATGGAAGAATGAACTGAAGCTGATTGAATAGAGAGCCACATCTACTGCAACTGAAAAGTTAGAATCTCAAGACTCAAGTACGCTACT
ATGCACTTGTTTATTCATTTCAGGAAACTAAAAACTAAACTTGTAAATAAGTACCTAAGTATGGTTATTGGTTCCCTCATGCCTTGG
ACACTTGATTGTCTTGGCACATACAGGTGCCATGCCTGCATATAGTAAGTGCTCAGAAAACATTCTTGAUTGAATTCAAGCCAAACAAAAATT
TTGGGGTAGGTAGAAAATATGCTTAAAGTATTGTTATGAGACTGGATATCTAGTATTGTCACAGGTAAATGATTCTTCAAAATTG
AAAGCAAATTGTTGAAATATTGTTGAAAAAAAGTTACTTCACAAGCTATAAATTAAAAGCCATAGGAATAGATACCGAAGTTATCCAA
CTGACATTAAATAATTGTTACTCATAGCTTAATGTGATGAGCCACAGAACGCTGCAAACCTTAATGAGATTTTAAAATAGCATCTAAGTCGG
AATCTTAGGCAAAGTGTGTTAGATGTAGCACTTCATATTGAAAGTGTCTTGGATATTGCATCTACTTGTCTGTATTACTGGTGTGA
ATGAATGAATAGGTACTGCTCTCTGGACATTACTTGACACATAATTACCCAAATGAATAAGCATACTGAGGTATCAAAAAGTCAAATATGT
TATAAATAGCTCATATATGTGTAGGGGGAGGAATTAGCTTACATCTCTTGTGTTAGTTCTCTGCATGTGCAGTTAACCTGGAAC
TCCGGTGCTAAGGAGAGACTGTTGGCCCTTGAGGGAGAGCTCCTCCCTGGATGAGAGAGAAGGACTTTACTCTTGAATTATCTTTGTGT
TGATGTTATCCACCTTGTACTCCACCTATAAAATCGGCTTATCTATTGATCTGTTCTAGTCCTTATAAAAGTCAAATGTTAACCTGGCAT
AAATTATAGACTTTTAGCAGAGAACCTAAATGCCAACAGTCTAAAATGCAGTTCAAGAAGAATGAATATTGATGGATA
GTTCTAAATACTAATGAACCTTAAAGCTACTATTGATCTGTCAAAGTGGTTTATATAATTCTTACAAATCACCTGACACATT
AATATAGGTTAAAAATGCTATCAGGCTGGTTGCAAAGAAAATGTATTACAAAGGCTGCTAAGTGTGTTAGAGCATACTCATTCTGTTCTCC
AAAATATTCTATAAGGTGTTAAGAATAGGTATGTTTAAAAGTTAAGTCCTACTATTATAGGAACGTACAATCACCTAAAATACCAATGA
TTACAAACTCCCTCTGGCCTCTGGACTGCAATTCTAAAAGTGTAAAAACATATTCTGCATTAAGTTAGGCAGTATTGCTTAGTTCAAA
GTGGTAGGCTTGGAGTCAGATTATTGATTCAAGATCCTACATCTACTGTTAGTAGCTCTGTTGCCTGAGGCAGGTCCCTAACATCTCTGTG
TGTGACTTGACCTTAAAATTGGAGACTGTCATAGGGTTAATCCCTGAGAAAATGAATGTGAAAAGTTAGCCTAATGTTAACTGCTATT
ATGGATTACCATATTTCACATTCACTACAGTACATGCACCTGTTAATATAAGATGCTCAATTCTTGTGAGTATAATTGTGACTCTCAAT
CTGGATATGCAATGAGTGGGCCTGTATGAGAACCTTATGAAAAATTGTGTTCACATGCCCTTACCAAGATACAGGAAACACGTACATG
TTCTATTGTATGTTAAATGCCTTAGAATTACTGAAATAGGATCCCTCAGTTGAGAGTCATAAAAGAGTAAAATTATTGTT

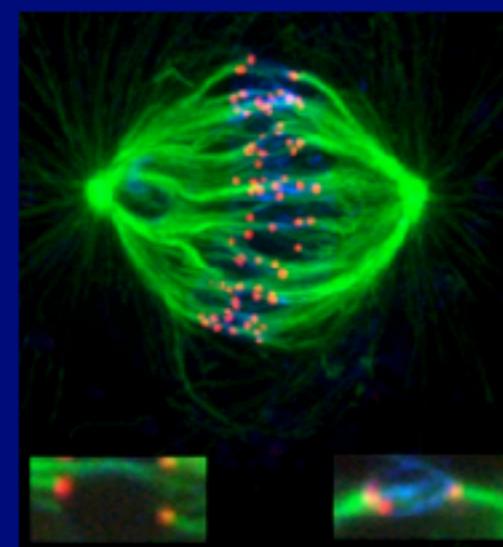
Non-Coding Functional Sequences



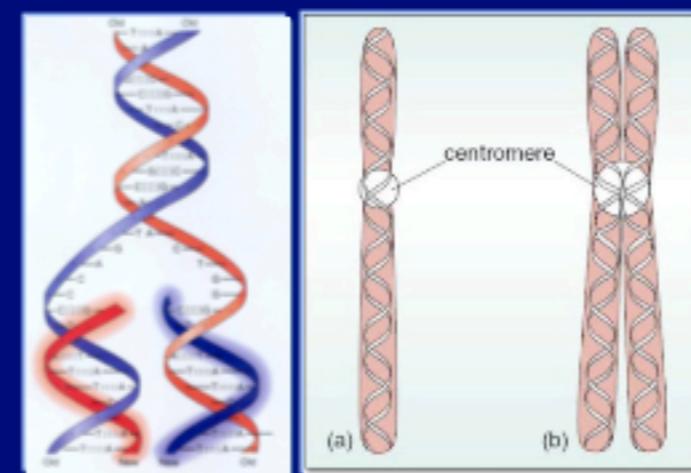
Gene
Regulation



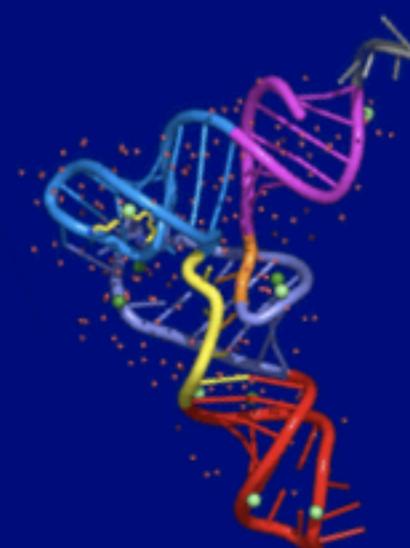
Chromosome
Packaging



Chromosome
Segregation



Chromosome
Replication



Non-Coding
RNAs



The Human Genome... by the Numbers

~5% of Human Genome Sequence is Constrained Across Mammals (and Presumed Functional)

5% of 3B Bases = ~150M Bases

Do NOT Yet Know the Position of these ~150M Functional Bases
Lower Bound for the Amount that is Functional

~1.5% Encodes for Protein (Genes)

Corresponds to ~18-22K Genes

Many More than ~22K Different Proteins

Good Inventory at Present

~3.5% Functional But Non-Coding

Gene Regulatory Elements

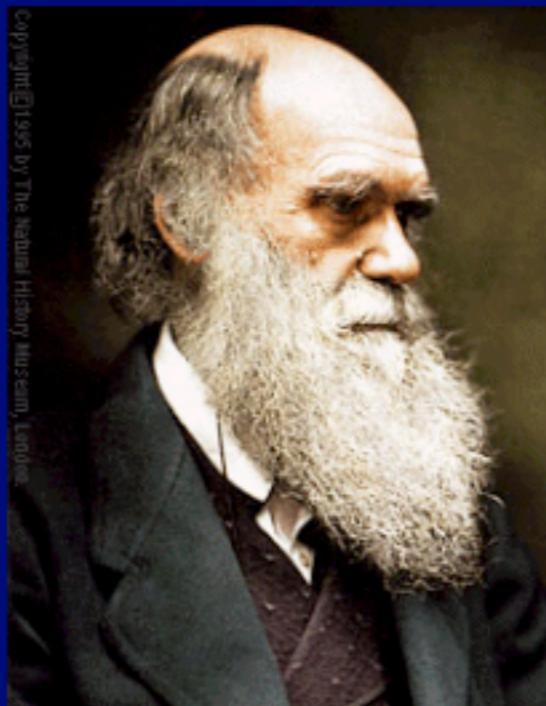
Chromosomal Functional Elements

Undiscovered Functional Elements (NOT Yet in Textbooks!)

Poor Inventory at Present

"It is not the strongest of the species that survives, nor the most intelligent that survives. It is the one that is the most adaptable to change."

(Attributed to Darwin)



Copyright © 1995 by The Natural History Museum, London.

Charles Darwin (1809-1882)

"For the last three and a half billion years, evolution has been taking notes."

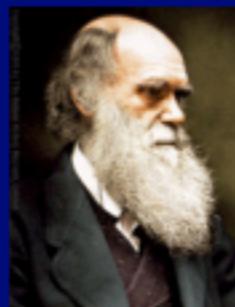
—Eric Lander

Comparative Sequence Analysis

*Using the 'Experiments of Evolution'
to Decode the Human Genome*

Species A

```
GATCGTCTAGAATCTCGAGATC  
TCTGAGAGTCGTGGAAACTGT  
GTGATGTGACGATTAGCCACA  
GTTACGTGTGAGAGATGTATGA  
TGCACCTGACCCGGGTTCACT  
CTCAACGACTCACTCCACCTCA  
GAGGCCACCAGCCGCTGTGCAC  
TACCGAGATAACACGATAACCTAC  
ACAGGTGTGACACACCCCTACC  
CGTCCACCACACGACTCACTCC  
ACCTCAGAGGCCACCGCCGCT  
GTGCACTACCGAGATAACAGAT  
ACCTACACAGGTGTGACACACG  
ATCCTTACCACTTACACATT  
ACCATATATCCACCTACCACAC  
ATACCTACCCATTGCACACCT  
ATTATTATTACCGGGACCGAGG
```



Species B

```
TATCGGCTAGAATCTCGAGATC  
TCTGAGAGTCGTGGAAACTGT  
GTGATGTGACTAGCCACAGTTA  
CGTGTGAGAGATGTATGATGCA  
CCTGACCCGGGTTCACTCTCA  
ACGACTCACTCCACCTCAGAGG  
CCCACCGCCGCTGTGCACGTCC  
ACCACGATCCTTACCACTACTTA  
CACATCACTCTCAACGACTCAC  
TCCACCTCAGAGGCCACCGCC  
GCTGTGCACGTCCACCGATC  
CTTACCACTTACACATTACC  
ATATATCCACCTACCACACATA  
CCTTACCATATATCCACCTACC  
ACACATACCTACCCATTGCAC  
ACCTATTATTATTACCGAGGGA  
GAGGGGTGACCACACTGTGACA
```

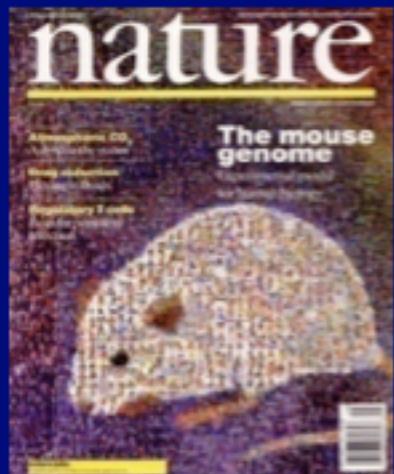
Compare



```
GATCGTCTAGAATCTCGAGATC  
TCTGAGAGTCGTGGAAACTGT  
GTGATGTGACTAGCCACAGTTA  
CGTGTGAGAGATGTATGATGCA  
CCTGACCCGGGTTCACTCTCA  
ACGACTCACTCCACCTCAGAGG  
CCCACCGCCGCTGTGCACGTCC  
ACCACGATCCTTACCACTACTTA  
CACATCACTCTCAACGACTCAC  
TCCACCTCAGAGGCCACCGCC  
GCTGTGCACGTCCACCGATC  
CTTACCACTTACACATTACC  
ATATATCCACCTACCACACATA  
CCTTACCATATATCCACCTACC  
ACACATACCTACCCATTGCAC  
ACCTATTATTATTACCGAGGGA  
GAGGGGTGACCACACTGTGACA
```

Sequences in Common (i.e., 'Conserved' or 'Constrained')

Vertebrate Genome Sequences



Mouse



Rat



Chicken



Chimpanzee



Dog



Macaque



Monodelphis



Platypus



Cow



Pufferfish



Diverse Landscape of Genome Sequencing

Human

Mouse

Rat

Pufferfish

Zebrafish

Chicken

Chimpanzee

Dog

Cow

Xenopus

Monodelphis

Macaque

Platypus

Marmoset

etc....

More Species = More Power

A Model of the Statistical Power
of Comparative Genome Sequence Analysis

Sean R. Eddy

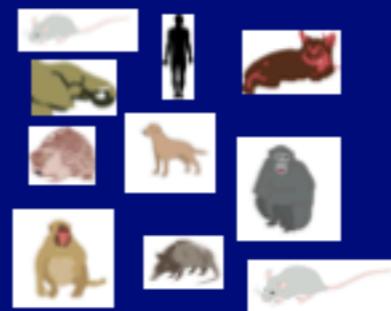
PLoS Biology (2005)



4 Species



50 bp



26 Species



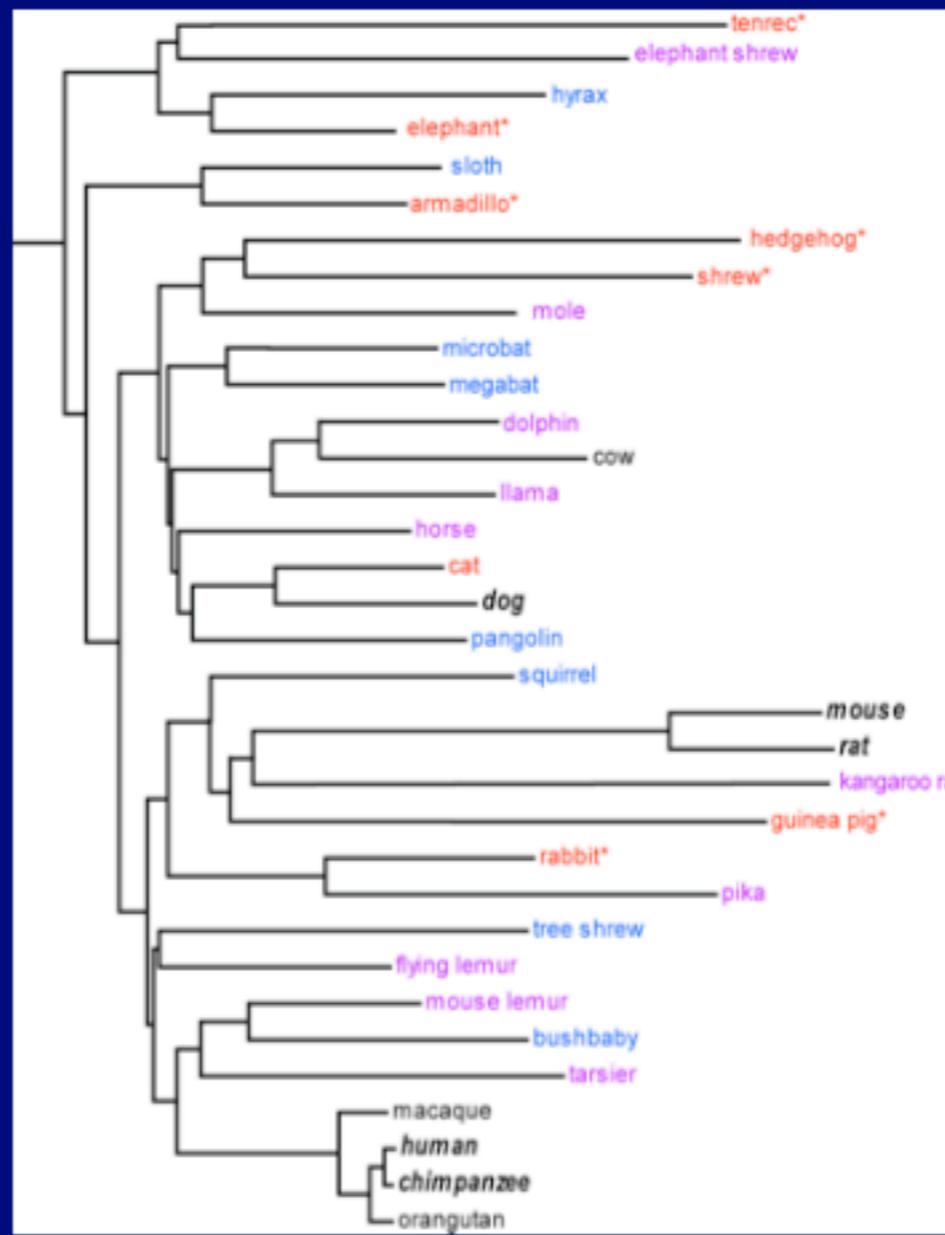
6 bp

'Light Sampling' of Many Mammalian Genomes

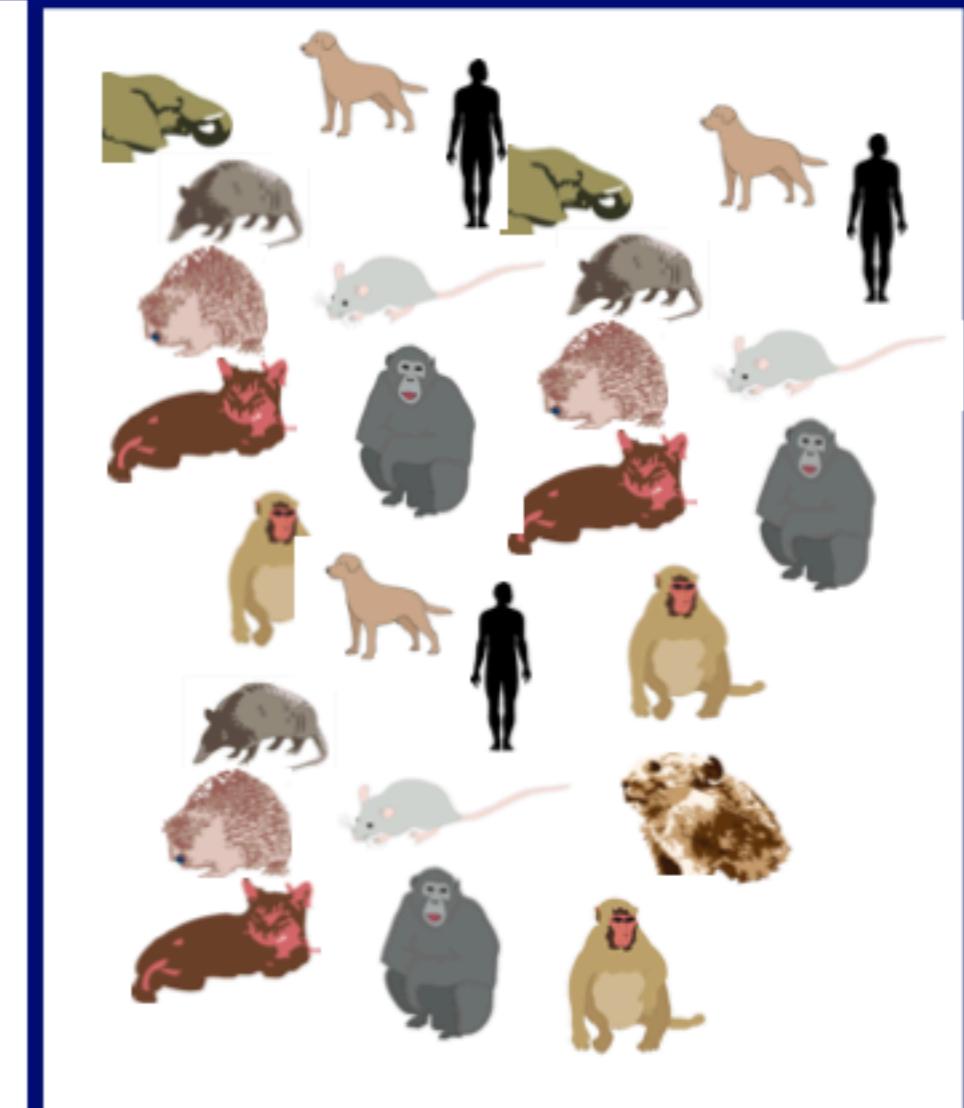
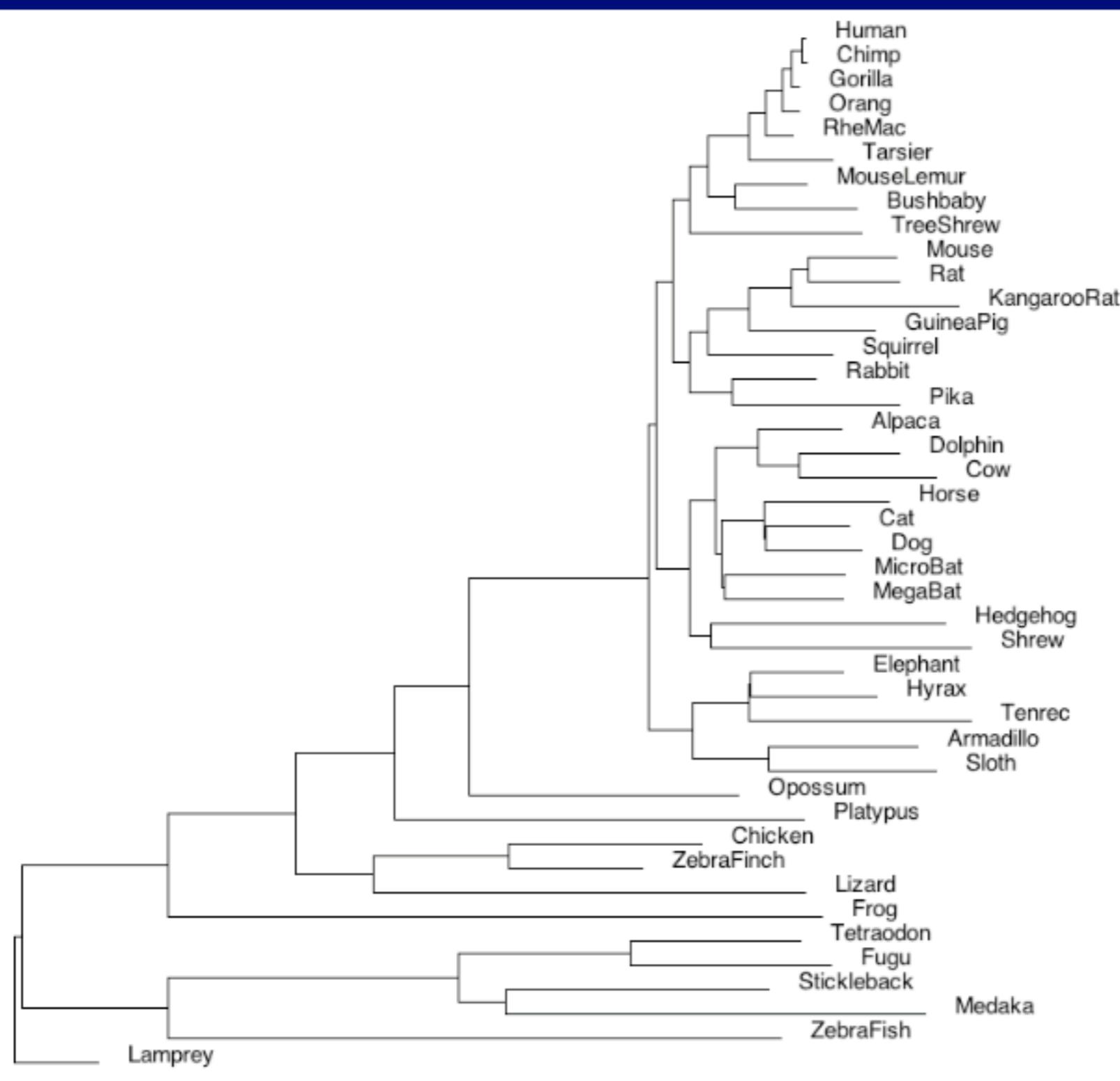
An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing

Elliott H. Margulies*,†, Jade Vinson‡, NISC Comparative Sequencing Program*§¶, Webb Miller§, David B. Jaffe‡,
Kerstin Lindblad-Toh‡, Jean Chang‡, Eric D. Green*§, Eric S. Lander‡, James C. Mullikin*§***, and Michele Clamp**

PNAS (2005)



22 Additional Mammalian Genome Sequences (@ Low Redundancy)



Diverse Landscape of Genome Sequencing

Human

Mouse

Rat

Pufferfish

Zebrafish

Chicken

Chimpanzee

Dog

Cow

Xenopus

Monodelphis

Macaque

Platypus

Marmoset

etc...

Armadillo

Elephant

Tenrec

Rabbit

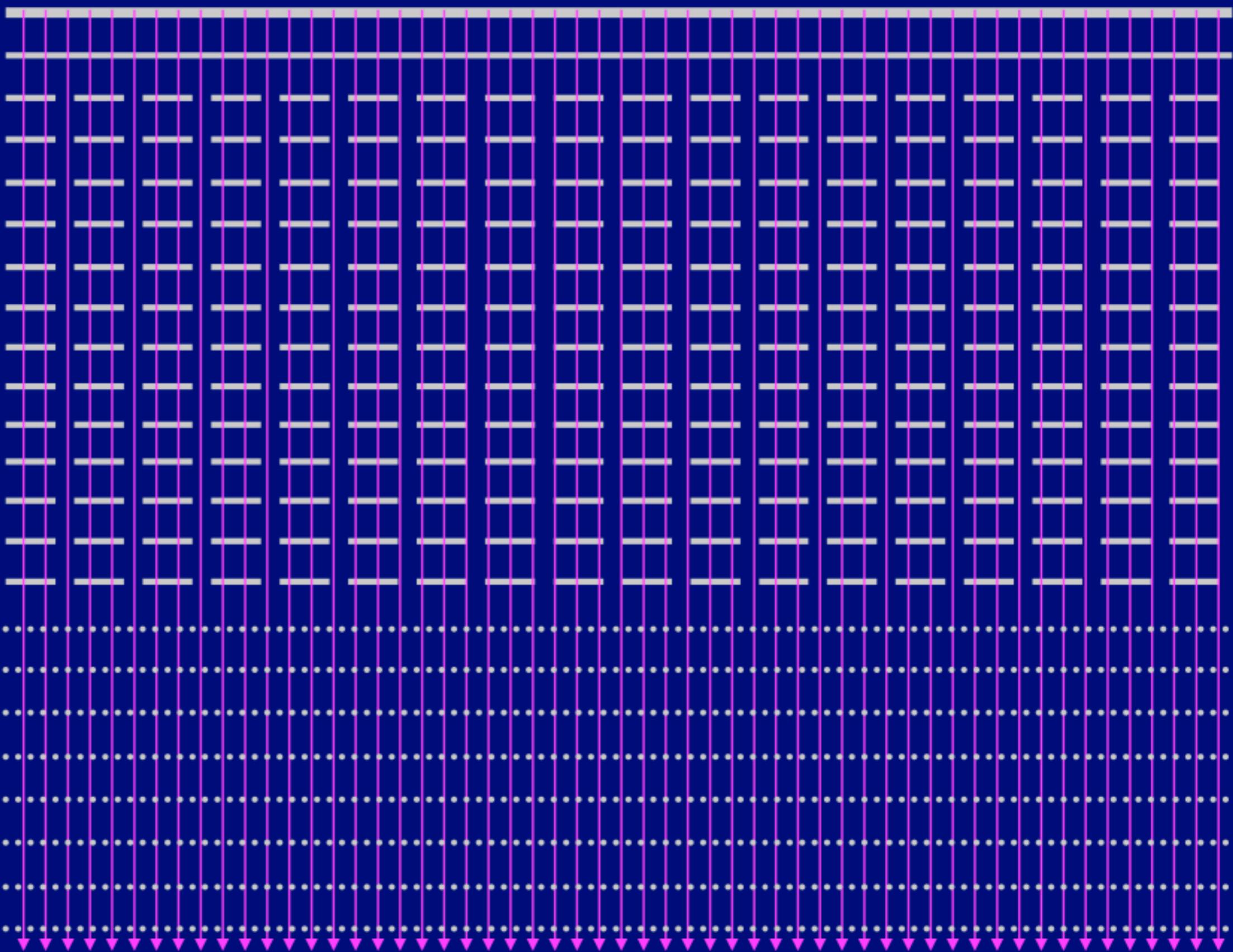
Cat

Shrew

Guinea Pig

Hedgehog

(and others...)



Multi-Species Sequence Comparisons

GATCGTCTAGAATCTCGA
GATCTCTGAGAGTCGTGG
GAAACTGTGTGATGTGAC
TAGCCACAGTTACGTGTG
AGAGATGTATGATGCACC
TGACCCGGGTTCACTCT
CAACGACTCACTCCACCT
CAGAGGCCAACGCCGCT
GTGCACGTCCACCGAT
CCTTACCACTTACACA
TTACCATATATCCACCTA
CCACACATACTACCCCCA
TTGCACACCTATTATTAT
TACC

GATCGTCTAGAATCTCGA
GATCTCTGAGAGTCGTGG
GAAACTGTGTGATGTGAC
TAGCCACAGTTACGTGTG
AGAGATGTATGATGCACC
TGACCCGGGTTCACTCT
CAACGACTCACTCCACCT
CAGAGGCCAACGCCGCT
GTGCACGTCCACCGAT
CCTTACCACTTACACA
TTACCATATATCCACCTA
CCACACATACTACCCCCA
TTGCACACCTATTATTAT
TACC

GATCGTCTAGAATCTCGA
GATCTCTGAGAGTCGTGG
GAAACTGTGTGATGTGAC
TAGCCACAGTTACGTGTG
AGAGATGTATGATGCACC
TGACCCGGGTTCACTCT
CAACGACTCACTCCACCT
CAGAGGCCAACGCCGCT
GTGCACGTCCACCGAT
CCTTACCACTTACACA
TTACCATATATCCACCTA
CCACACATACTACCCCCA
TTGCACACCTATTATTAT
TACC

GATCGTCTAGAATCTCGAG
ATCTCTGAGAGTCGTGGGA
AACTGTGTGATGTGACTAG
CCACAGTTACGTGTGAGAG
ATGTATGATGCACCTGACC
CGGGTTTCACTCTCAACGA
CTCACTCCACCTCAGAGGC
CCACCGCCGCTGTGCACGT
CCACCACTGATCCTTACAC
ACTTACACATTACCATATA
TCCACCTTACACACATACC
TACCCCATGACACCTAT
TATTATTACCACTGAACTAC
TACCCACCTATTATTAT
TACCGTCTAGAGGGAGCTA

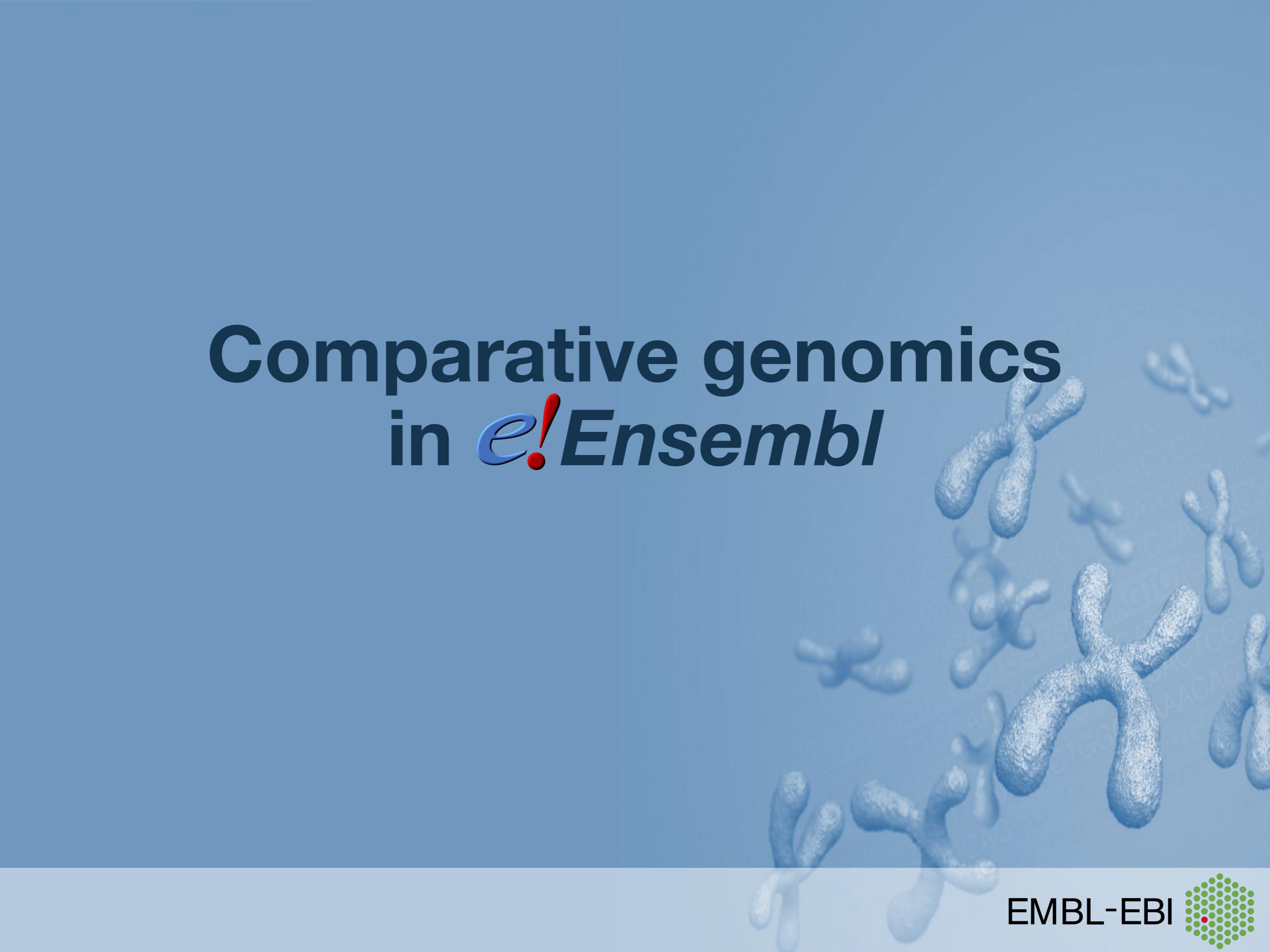
GATCGTCTAGAATCTCGA
GATCTCTGAGAGTCGTGG
GAAACTGTGTGATGTGAC
TAGCCACAGTTACGTGTG
AGAGATGTATGATGCACC
TGACCCGGGTTCACTCT
CAACGACTCACTCCACCT
CAGAGGCCAACGCCGCT
GTGCACGTCCACCGAT
CCTTACCACTTACACA
TTACCATATATCCACCTA
CCACACATACTACCCCCA
TTGCACACCTATTATTAT
TACC

GATCGTCTAGAATCTCGA
GATCTCTGAGAGTCGTGG
GAAACTGTGTGATGTGAC
TAGCCACAGTTACGTGTG
AGAGATGTATGATGCACC
TGACCCGGGTTCACTCT
CAACGACTCACTCCACCT
CAGAGGCCAACGCCGCT
GTGCACGTCCACCGAT
CCTTACCACTTACACA
TTACCATATATCCACCTA
CCACACATACTACCCCCA
TTGCACACCTATTATTAT
TACC

GATCGTCTAGAATCTCGA
GATCTCTGAGAGTCGTGG
GAAACTGTGTGATGTGAC
TAGCCACAGTTACGTGTG
AGAGATGTATGATGCACC
TGACCCGGGTTCACTCT
CAACGACTCACTCCACCT
CAGAGGCCAACGCCGCT
GTGCACGTCCACCGAT
CCTTACCACTTACACA
TTACCATATATCCACCTA
CCACACATACTACCCCCA
TTGCACACCTATTATTAT
TACC

HUMAN

Comparative genomics in *e!Ensembl*



Training materials

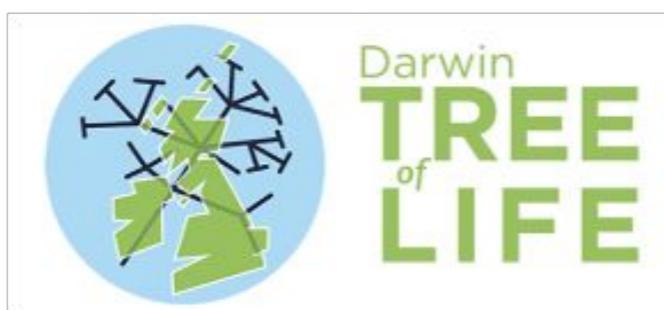


- Ensembl training materials are protected by a CC BY license:
creativecommons.org/licenses/by/4.0/
- If you wish to re-use these materials, please credit Ensembl for their creation
- If you use Ensembl for your work, please cite our papers:
ensembl.org/info/about/publications.html

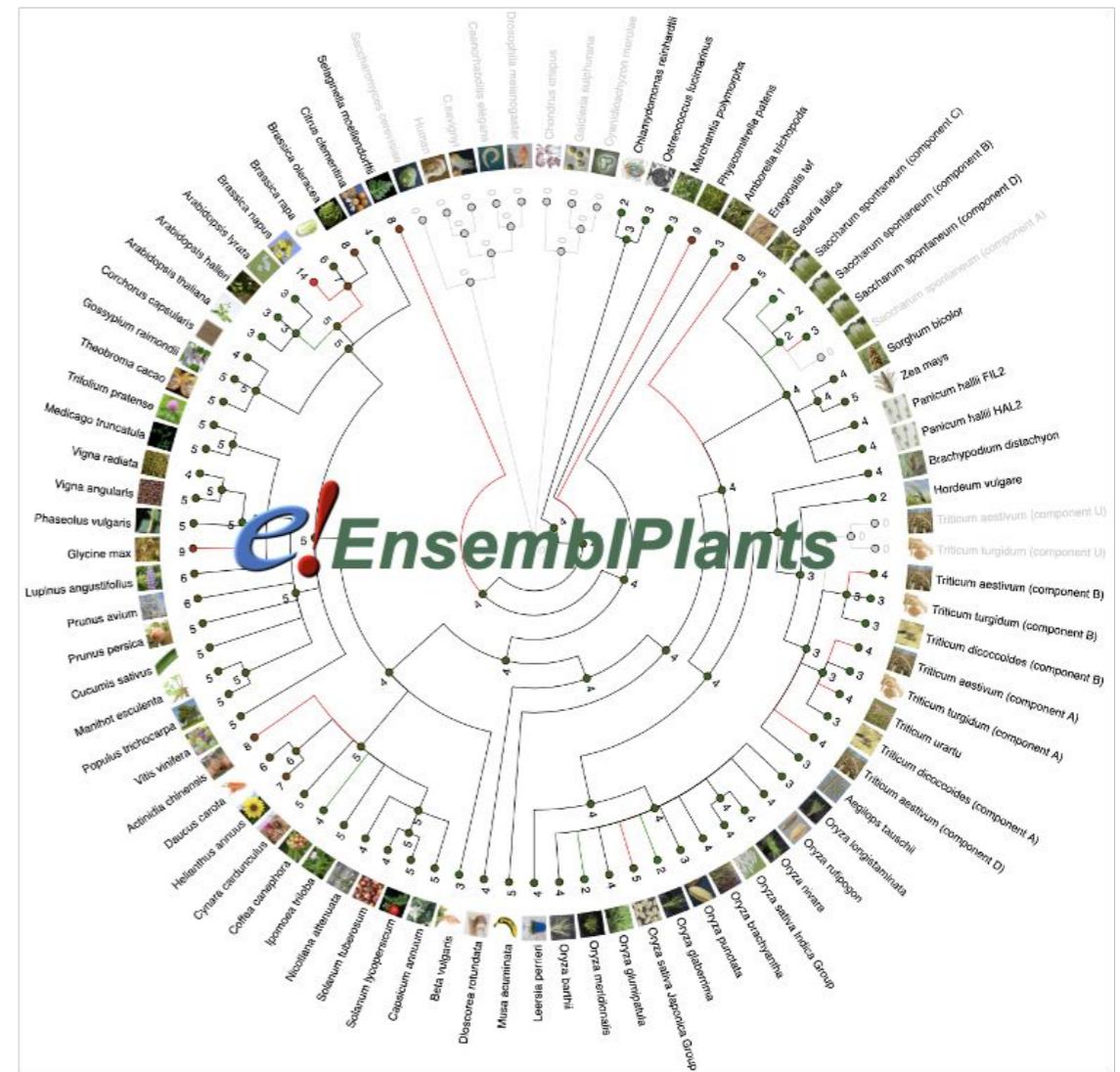
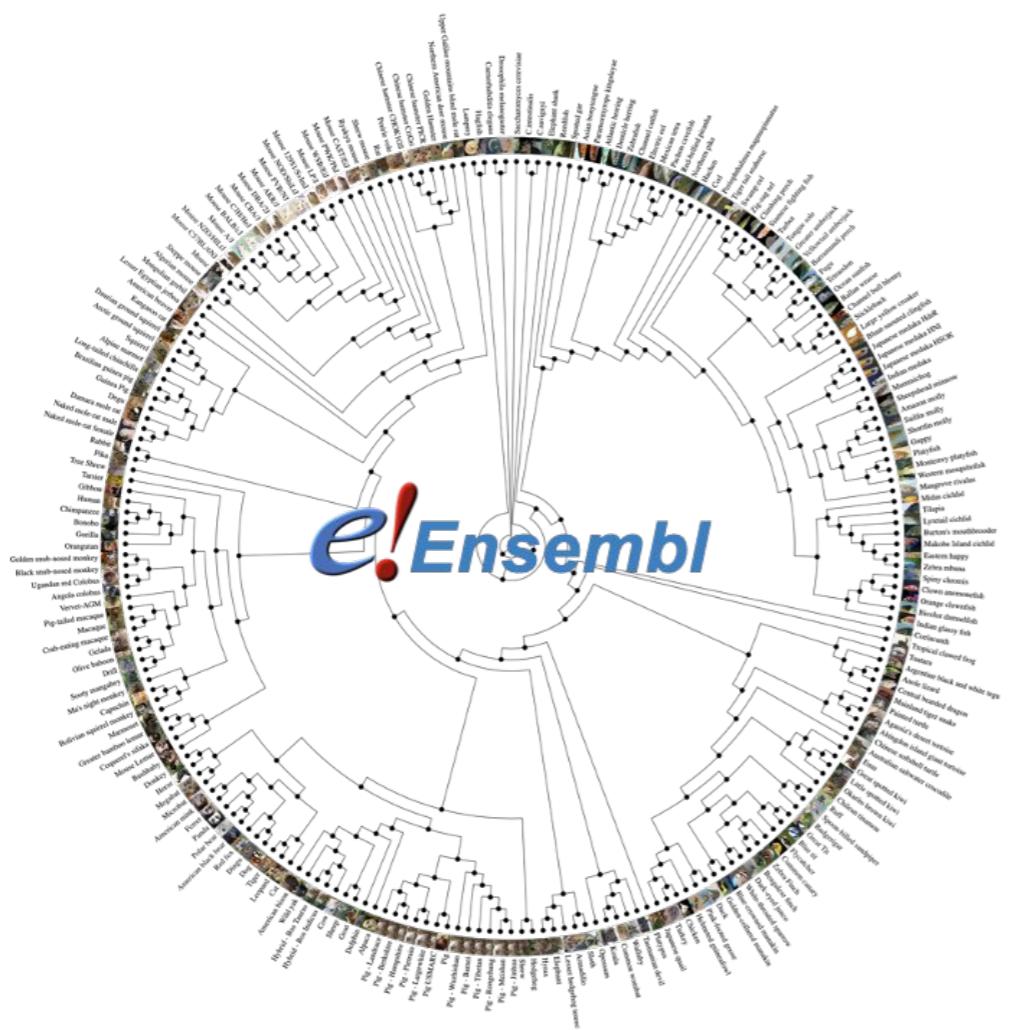
Applications

Comparative genomics allows us to:

- Examine **evolutionary relationship**
- Find **differences** between genomes
- Derive gene function based on **homology**
- Identify highly **conserved regions**



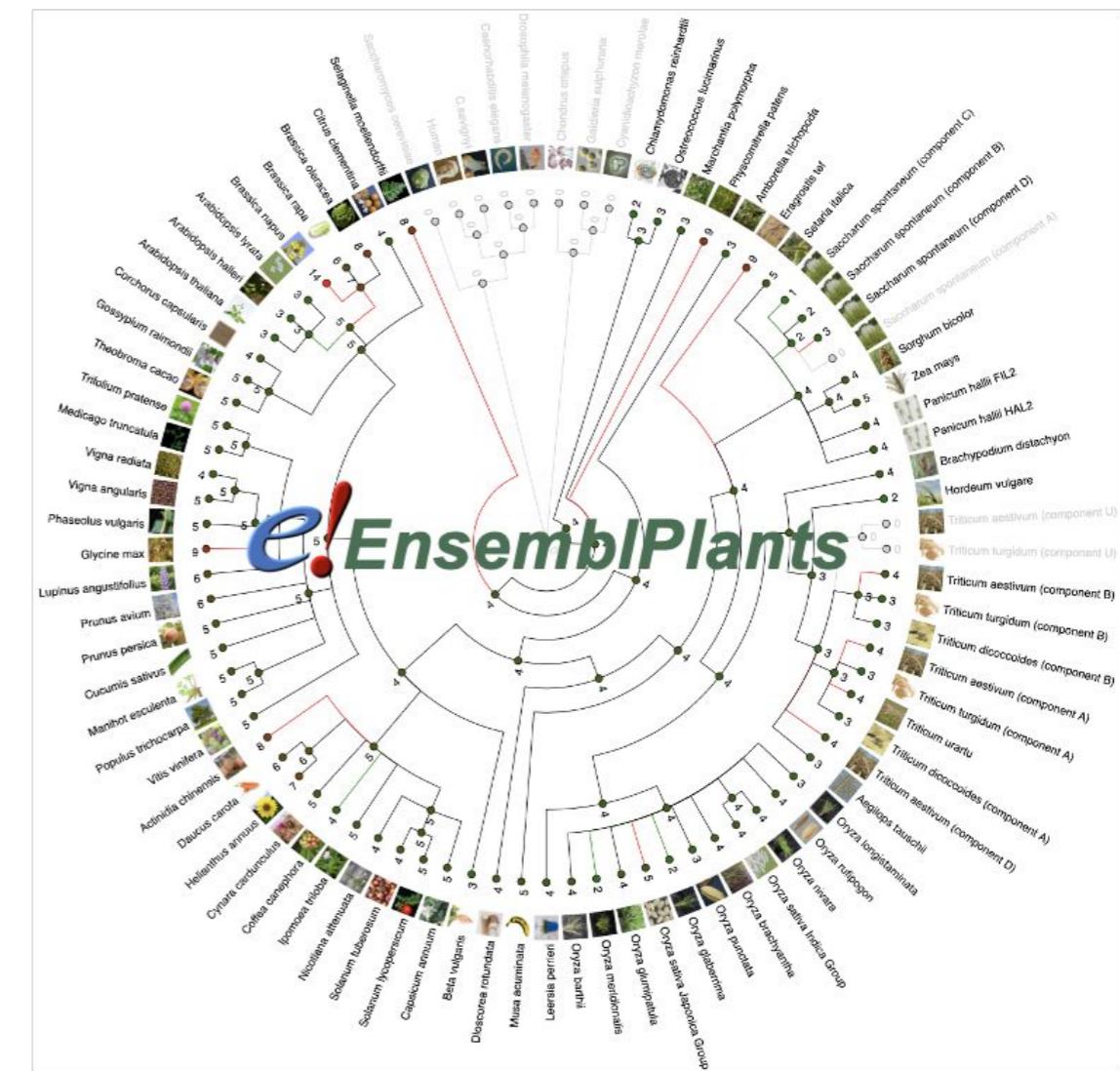
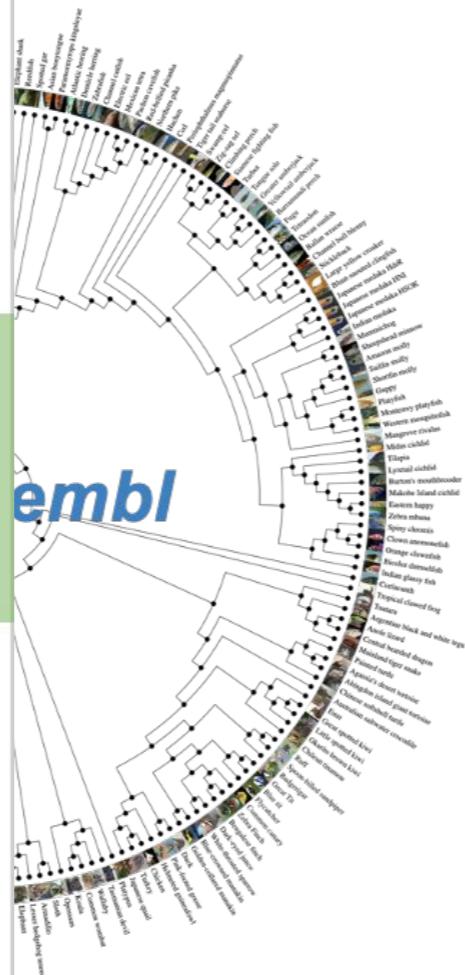
Comparative analysis by taxa



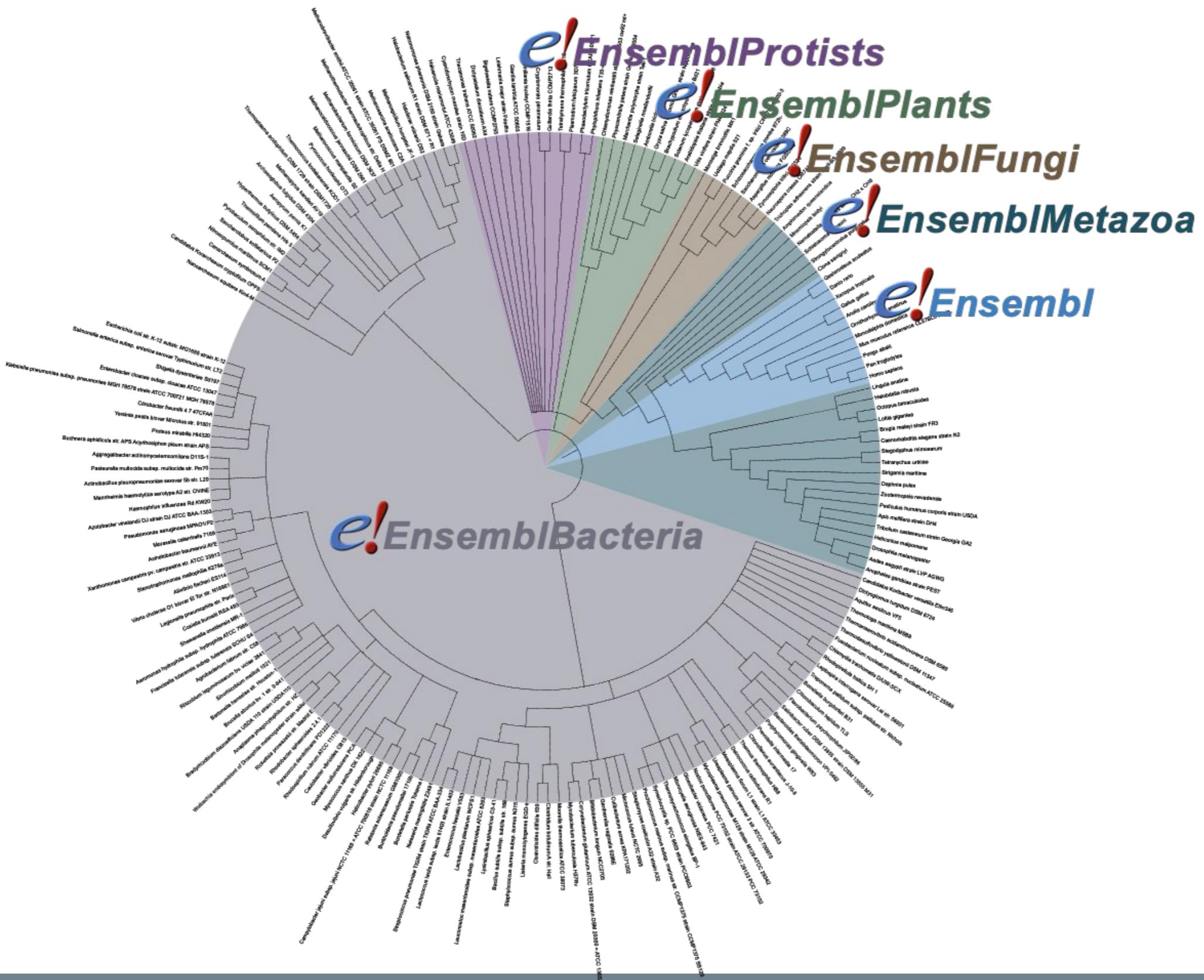
Comparative analysis by taxa

Gene-based displays

- **Summary**
 - Splice variants
 - Transcript comparison
 - Gene alleles
- **Sequence**
 - Secondary Structure
 - Gene families
 - Literature
- **Plant Compara**
 - Genomic alignments
 - Gene tree
 - Gene gain/loss tree
 - Orthologues
 - Paralogues
- **Pan-taxonomic Compara**
 - Gene Tree
 - Orthologues
- **Ontologies**
 - GO: Biological process
 - GO: Cellular component
 - GO: Molecular function
 - PO: Plant anatomical entity
 - PO: Plant structure development
- **Phenotypes**
- **Genetic Variation**
 - Variant table
 - Variant image
 - Structural variants
- **Gene expression**
- **Pathway**
- **Regulation**
- **External references**
- **Supporting evidence**
- **ID History**
 - Gene history



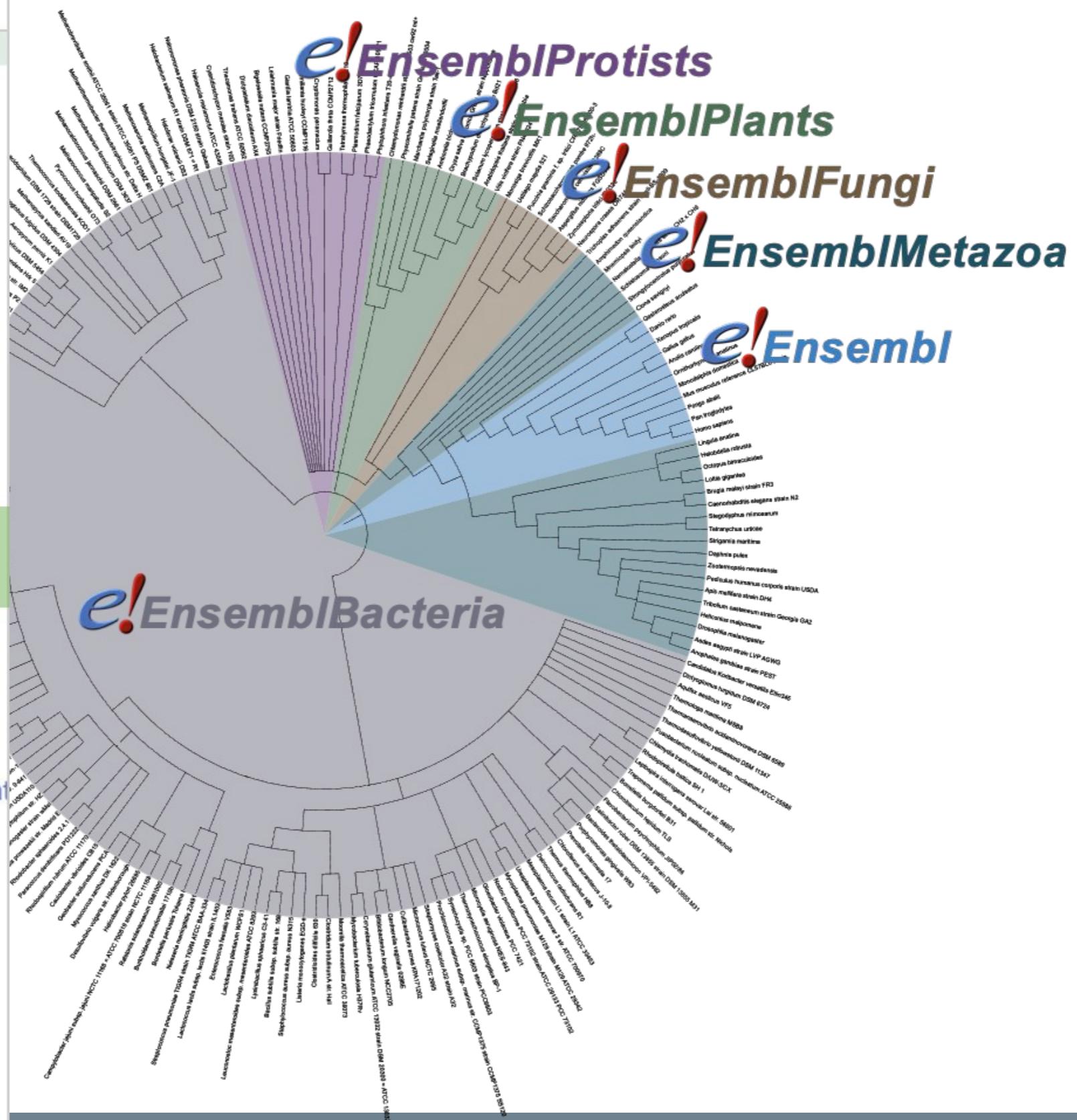
Pan-taxonomic compara



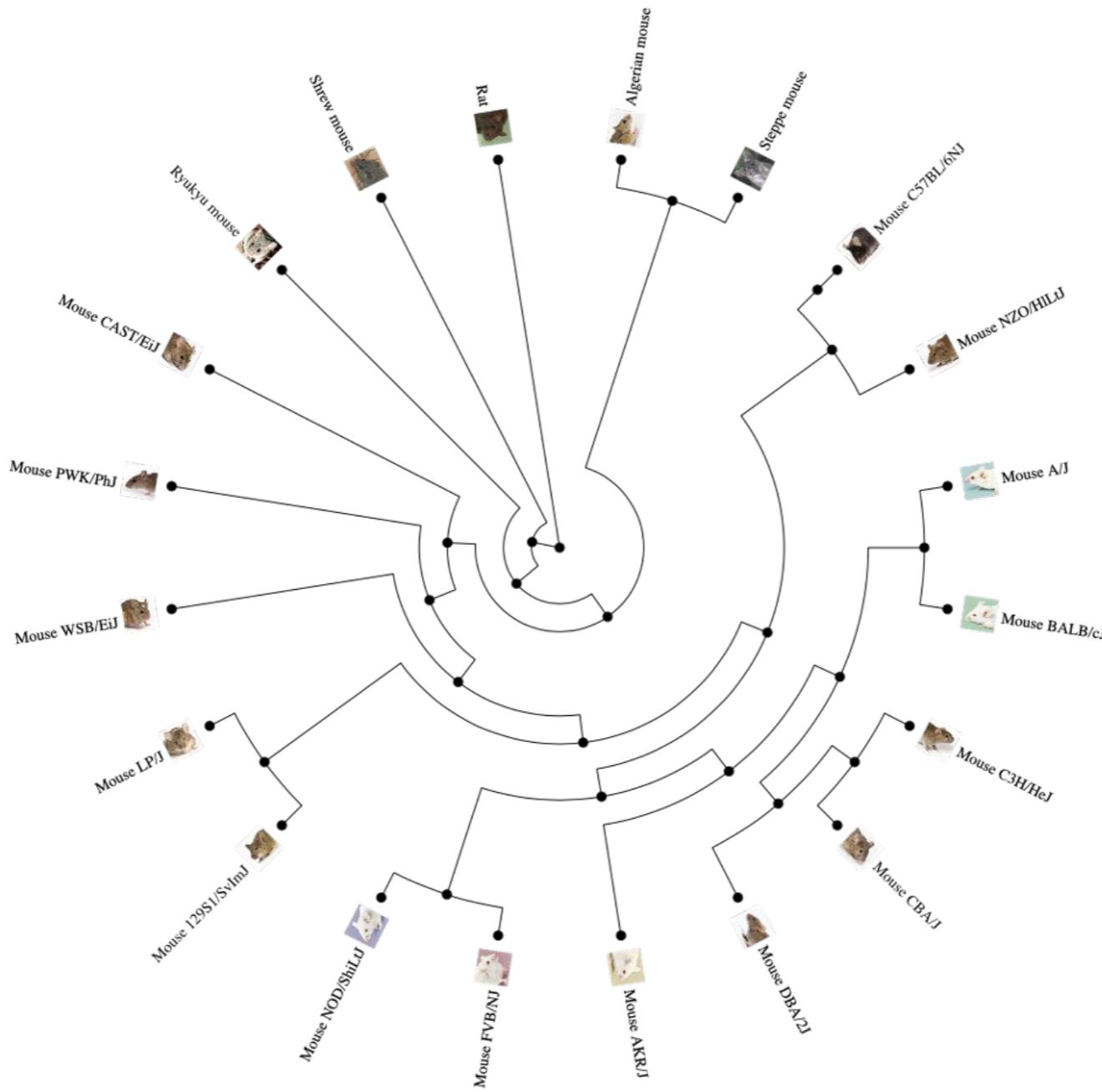
Pan-taxonomic compara

Gene-based displays

- Summary
- Splice variants
- Transcript comparison
- Gene alleles
- Sequence
 - Secondary Structure
- Gene families
- Literature
- Plant Compara
 - Genomic alignments
 - Gene tree
 - Gene gain/loss tree
 - Orthologues
 - Paralogues
- Pan-taxonomic Compara
 - Gene Tree
 - Orthologues
- Ontologies
 - GO: Biological process
 - GO: Cellular component
 - GO: Molecular function
 - PO: Plant anatomical entity
 - PO: Plant structure development
- Phenotypes
- Genetic Variation
 - Variant table
 - Variant image
 - Structural variants
- Gene expression
- Pathway
- Regulation
- External references
- Supporting evidence
- ID History
 - Gene history



Mouse compara



Types of data

Gene-based resources (found in the gene tab)

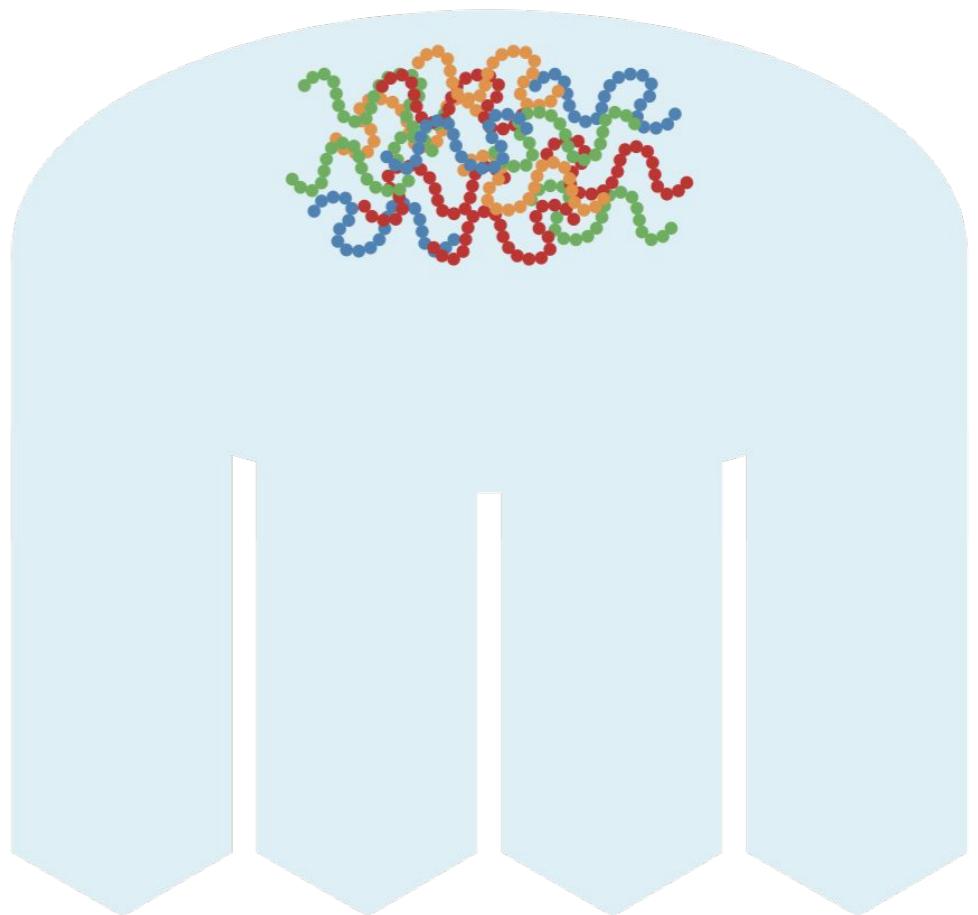
- Phylogenetic trees and tree-inferred homology
- Protein trees
- ncRNA trees
- Stable ID mapping
- Protein families

Sequence-based resources (found in the location tab)

- Whole genome alignments
- Ancestral sequences
- Age of base
- Conservation scores and constrained elements
- Syntenies

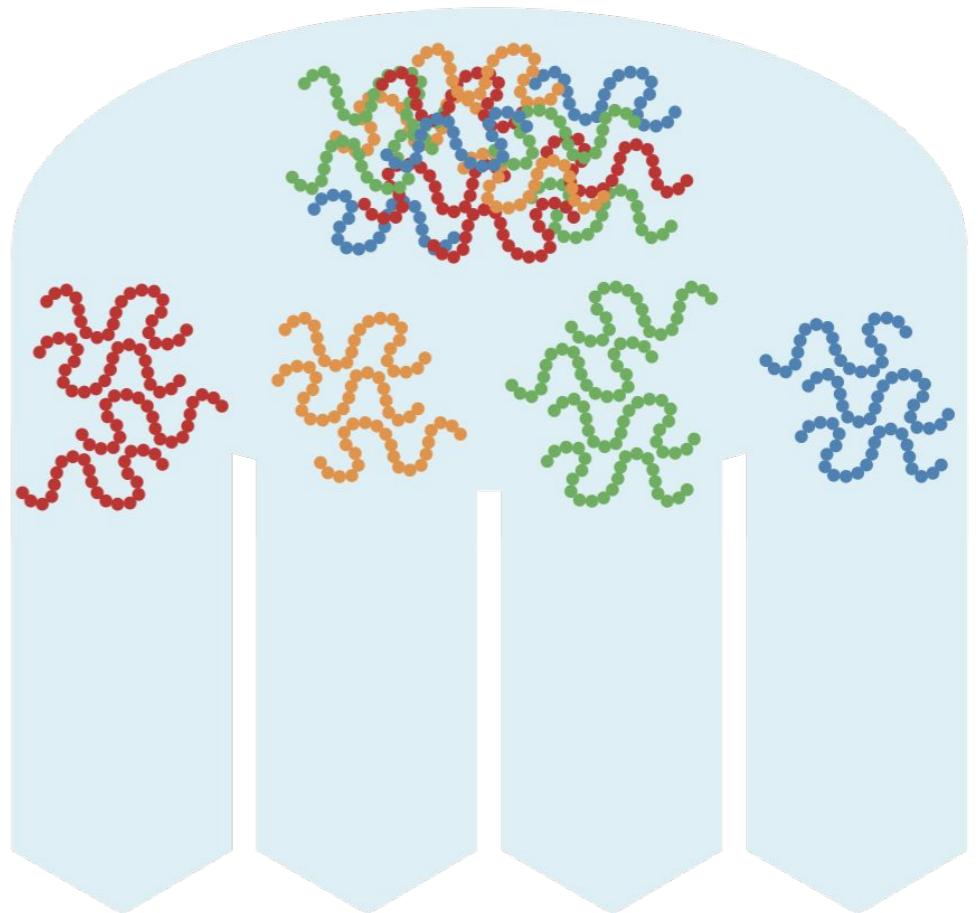
Gene/protein trees

1. Representative translation
of each gene from all
species



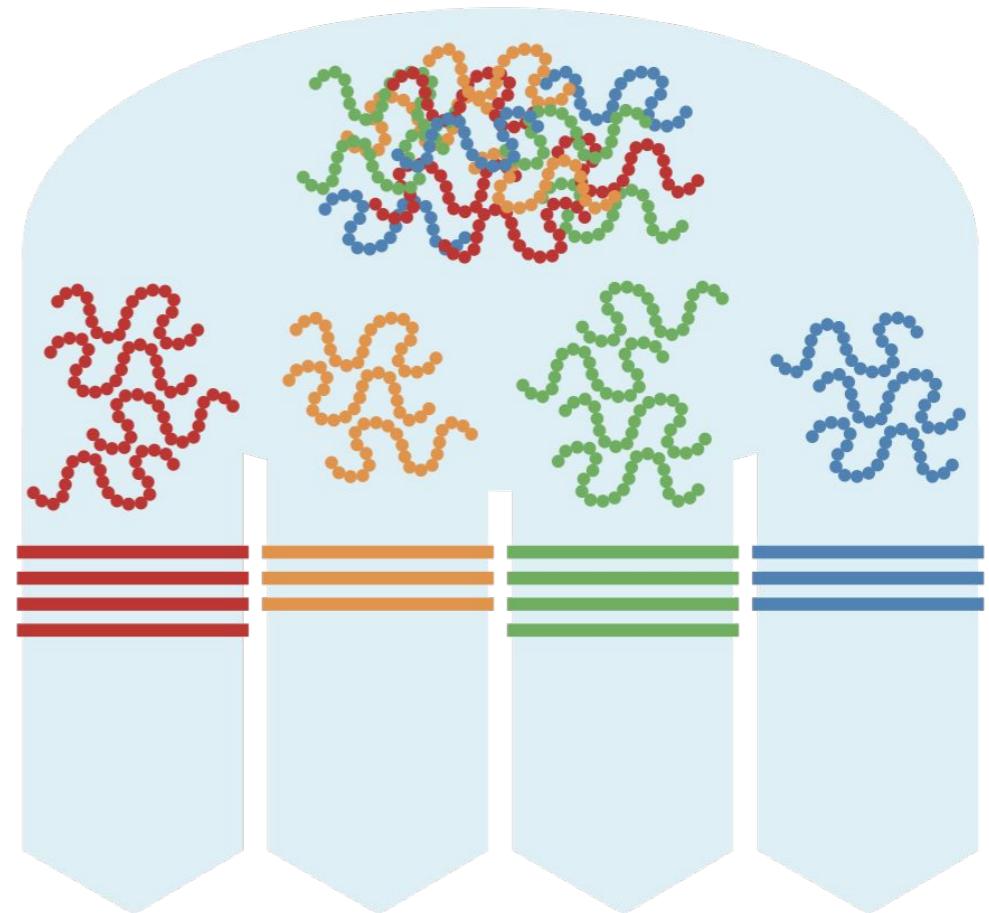
Gene/protein trees

1. Representative translation of each gene from all species
2. All-vs-all HMM search to classify into families or clustering



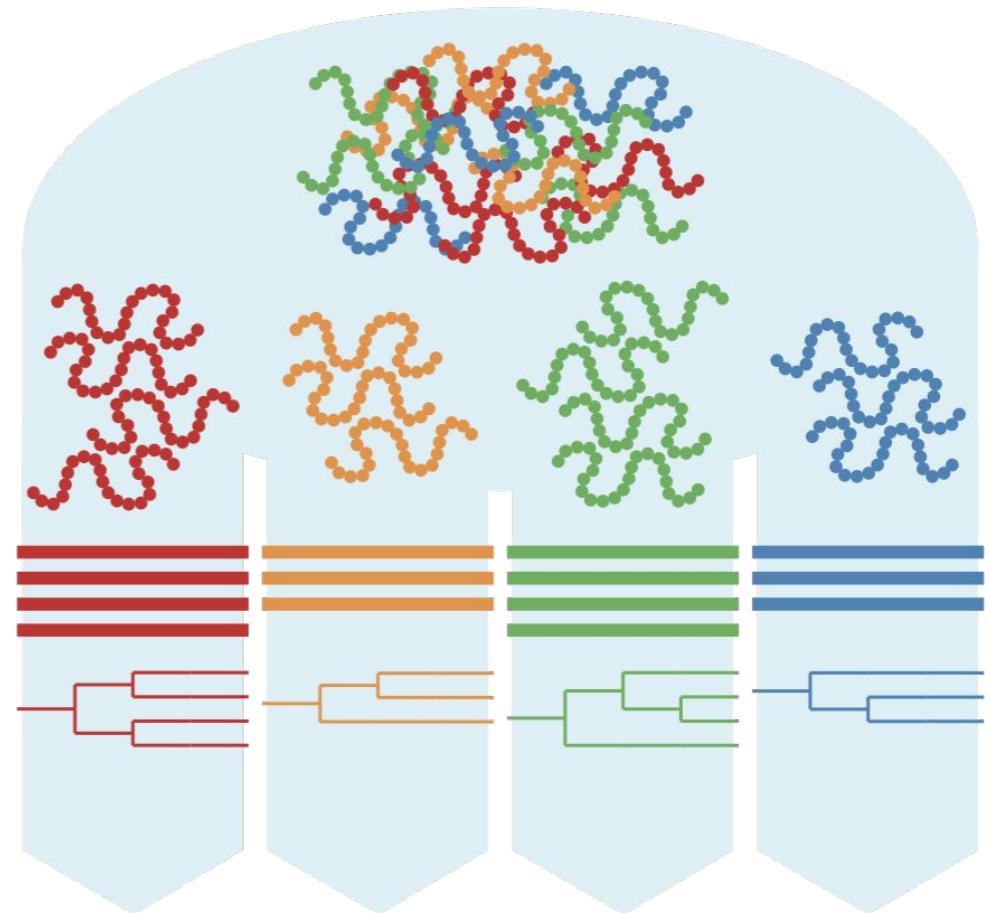
Gene/protein trees

1. Representative translation of each gene from all species
2. All-vs-all HMM search to classify into families or clustering
3. Multiple protein alignment



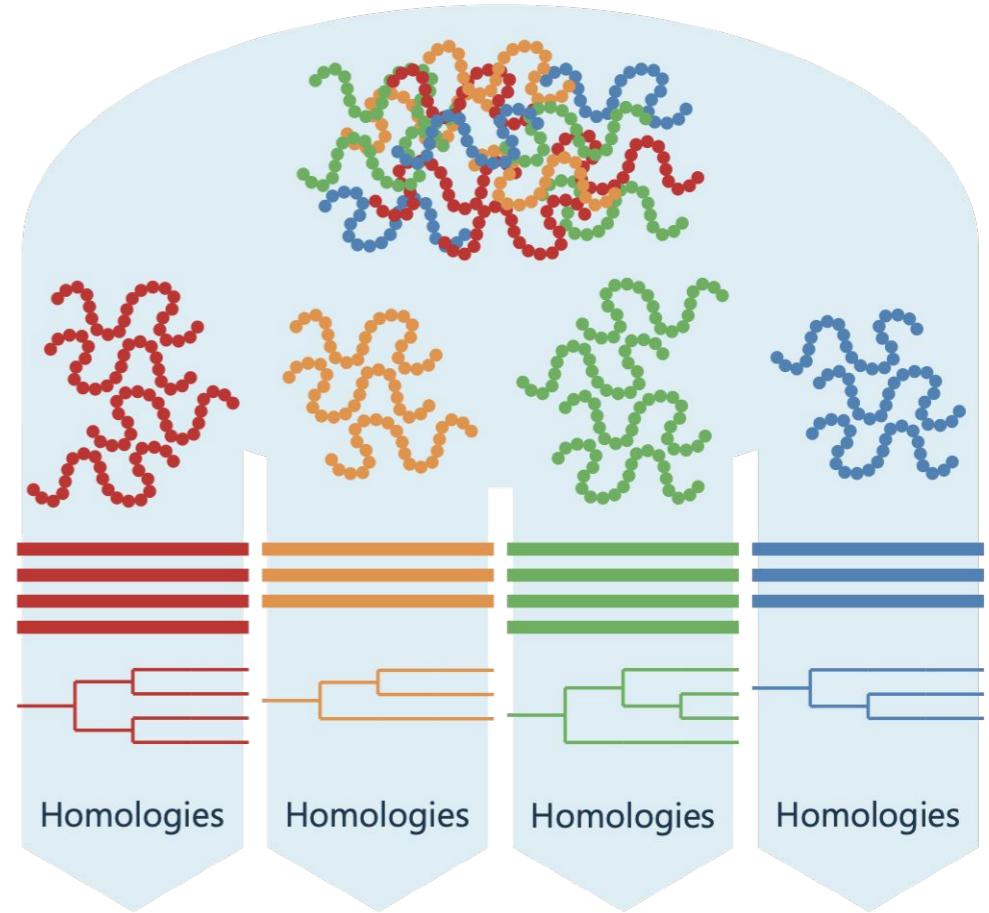
Gene/protein trees

1. Representative translation of each gene from all species
2. All-vs-all HMM search to classify into families or clustering
3. Multiple protein alignment
4. Phylogenetic tree for each aligned cluster and reconciliation against NCBI taxonomy



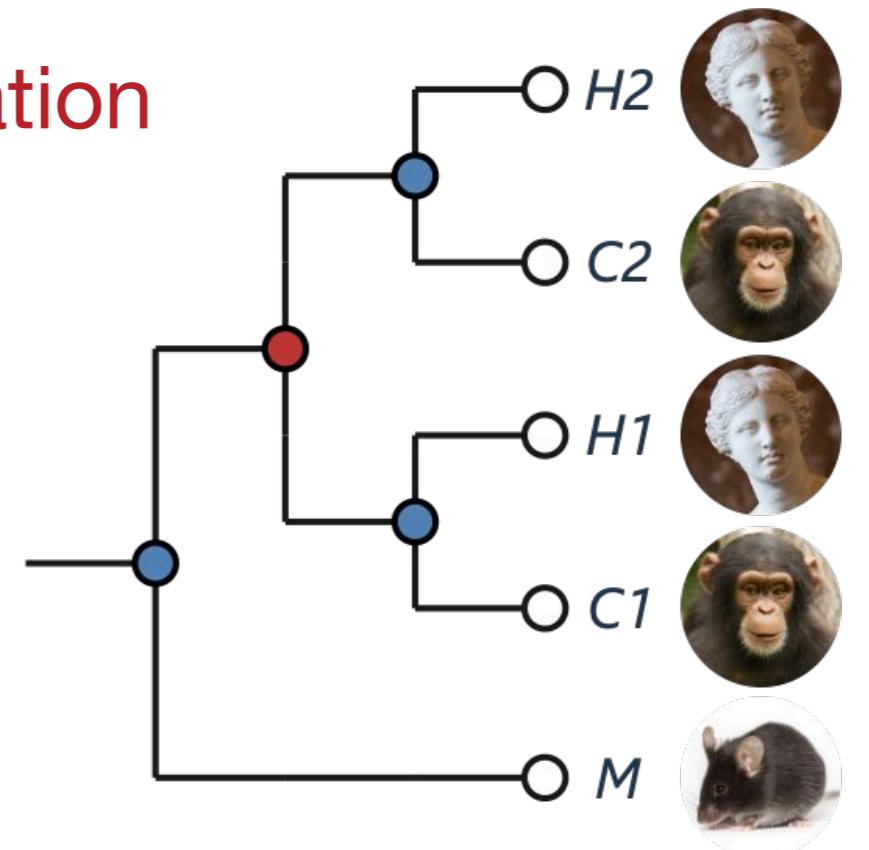
Gene/protein trees

1. Representative translation of each gene from all species
2. All-vs-all HMM search to classify into families or clustering
3. Multiple protein alignment
4. Phylogenetic tree for each aligned cluster and reconciliation against NCBI taxonomy
5. Ortho-/parologue inference

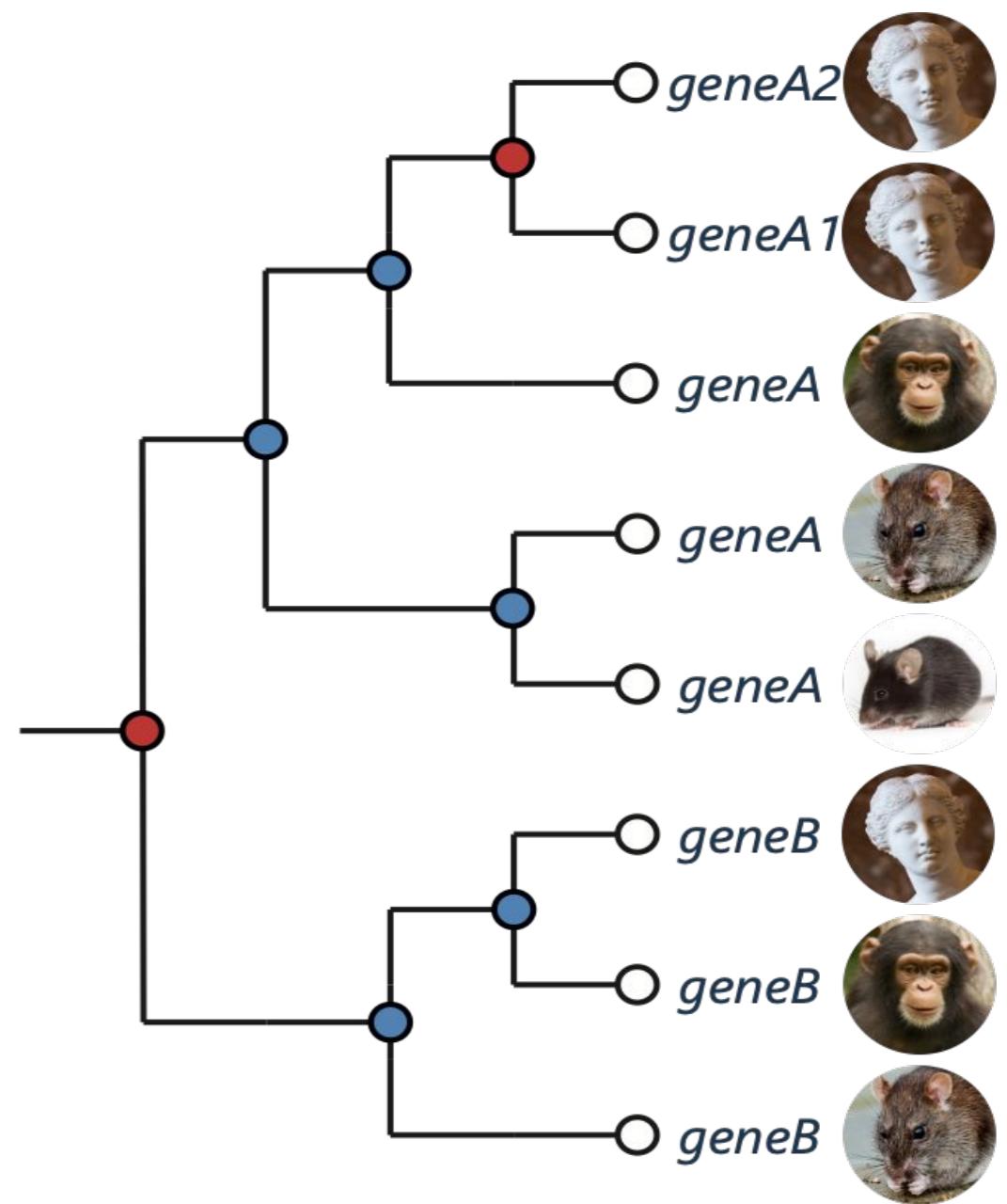
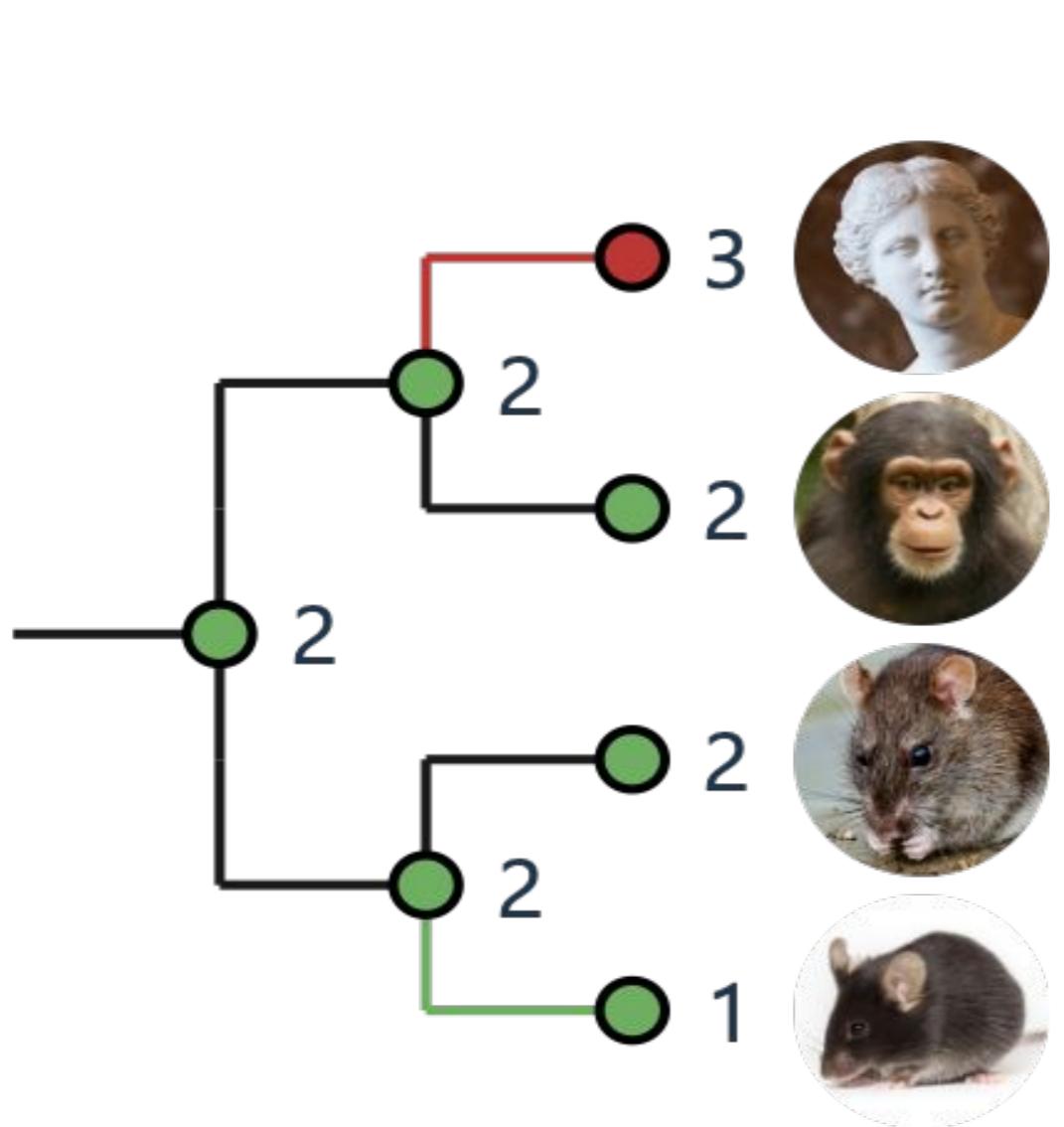


Homology relationships

- Orthologues
 - Genes emerged through a **speciation** event, e.g. $C1$ and $H1$; $C2$ and M ; $H2$ and M
 - 1-to-1: $C1$ and $H1$
 - 1-to-many: M and $H1$, $H2$
- Paralogues
 - Genes emerged through a **duplication** event, e.g. $C1$ and $C2$, $H1$ and $H2$



Gene tree vs gene gain/loss tree



Whole genome alignments: pairwise vs multiple

- To identify highly conserved regions
 - Sequences that evolve slowly
 - Regions likely to be functional
 - Both coding and non-coding
- To support problematic gene predictions
- To define syntenic regions

Pairwise alignments

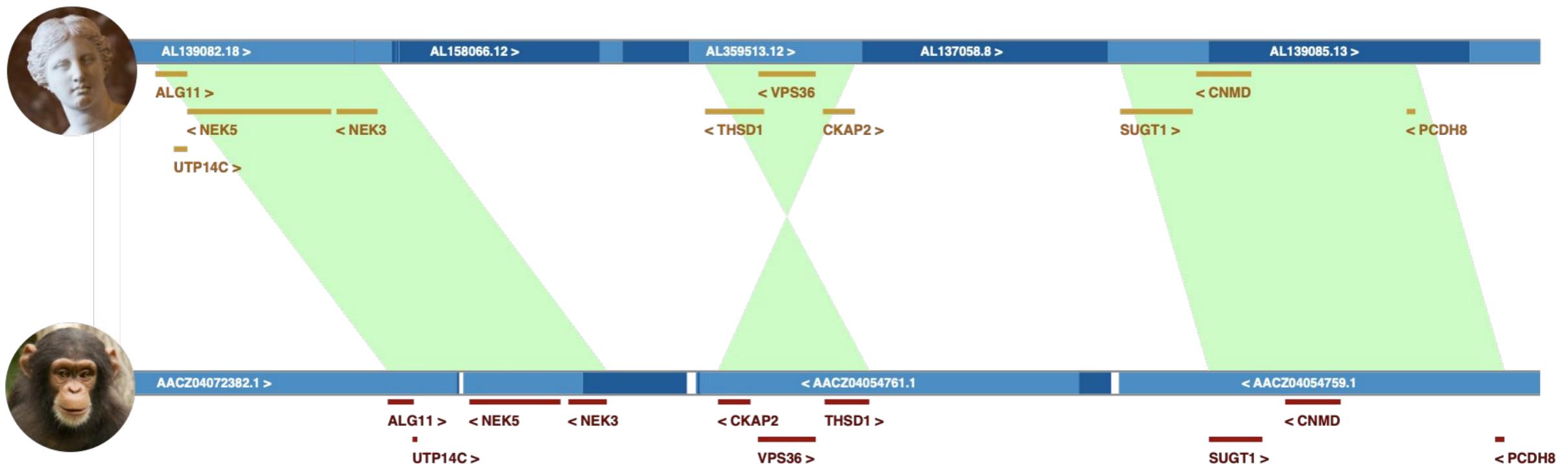
Pairwise alignments with BLASTZ (older) or LASTZ (newer):

- Human vs everything
- Model organisms vs related species
- Agricultural mammals vs each other



Shared synteny

Conserved order of aligned homologous genomic blocks between species (irrespective of orientation):



Multiple alignments



- EPO (Enredo-Pecan-Ortheus)

38 fish; 17 sauropsids; 46 eutherian mammals; 12 primates, 21 murinae

- EPO Extended (formerly “low-coverage”)

Allows fragmented assemblies

65 fish; 27 sauropsids; 99 eutherian mammals; 24 primates; 16 pig breeds and other agricultural mammals

- Mercator-Pecan

65 amniota vertebrates (mammals, birds, reptiles)

More information

Herrero et al.

Ensembl comparative genomics resources

Database: the Journal of Biological Databases and Curation (2016)

epmc.org/abstract/MED/26896847