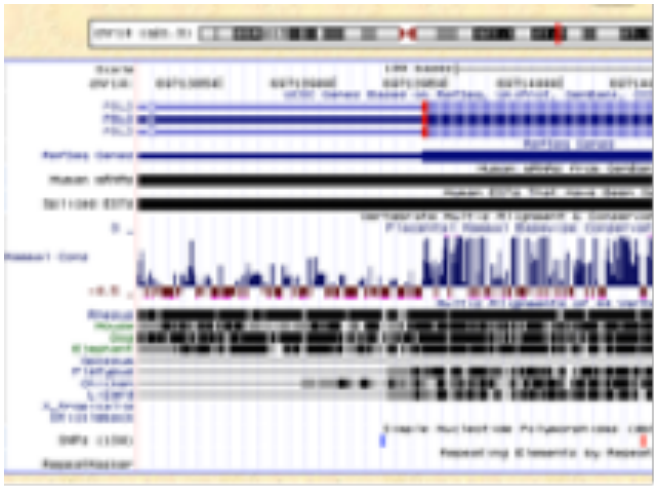
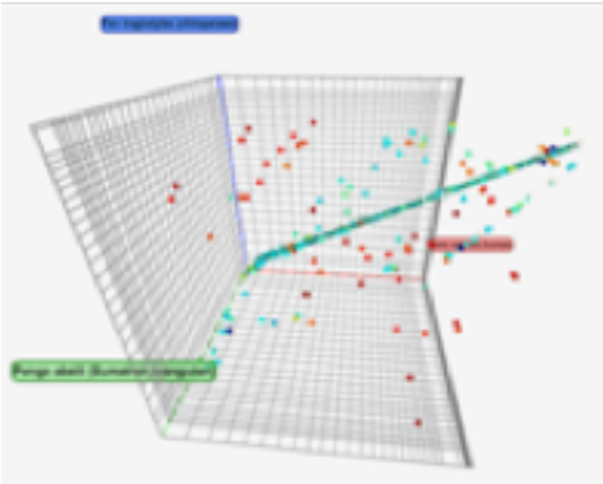
Computational Genomics

Introduction
To
Comparative Genomics





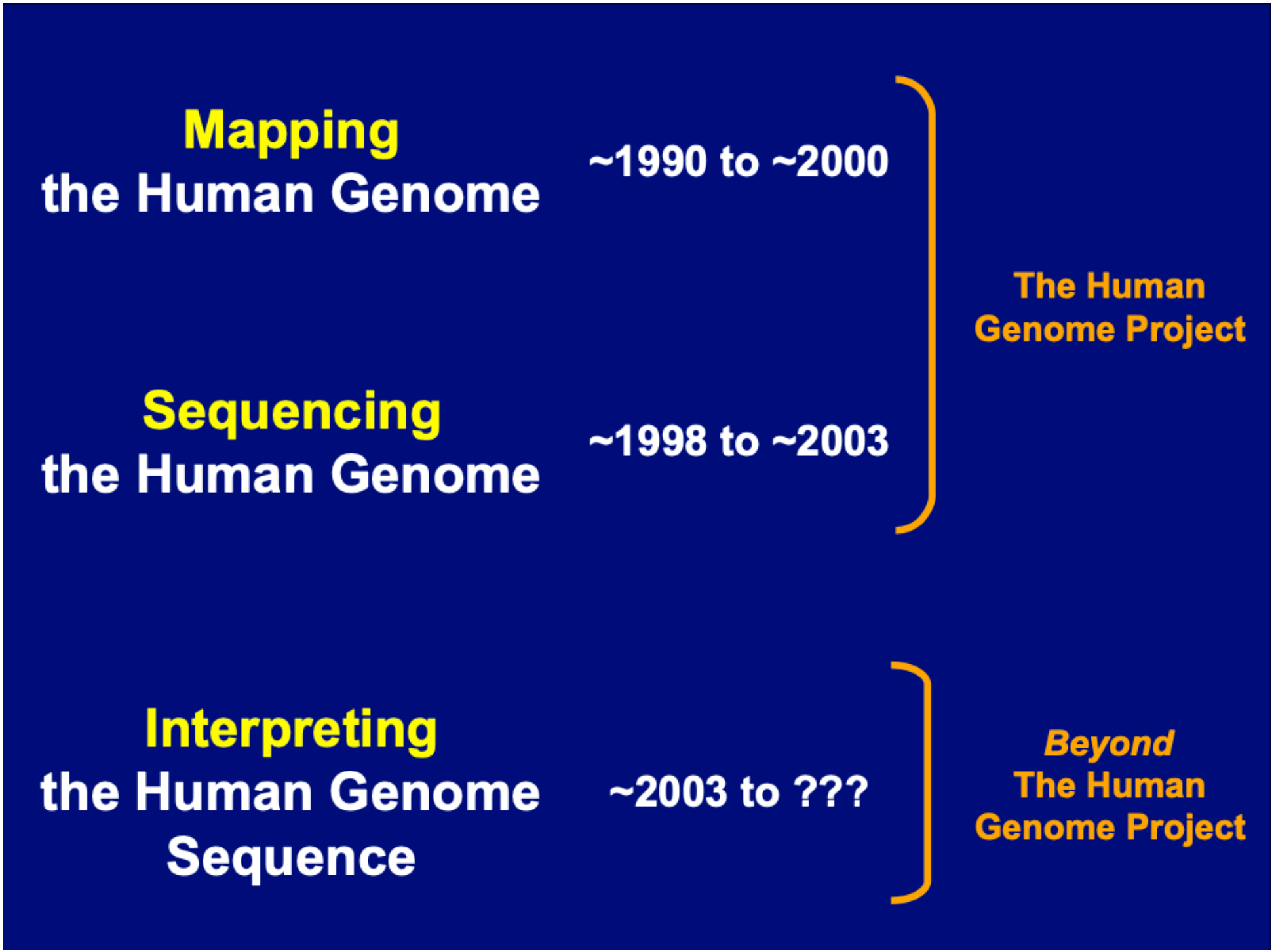
National Human Genome Research Institute

Division of Intramural Research

The Genomic Landscape: *circa* 2010

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

A Brief Introduction to the Logic for Comparative Genomics



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

TGCCGCGGAACTTTTCGGCTCTCTAAGGCTGTATTTTGATATACGAAAGGCACATTTTCCTTCCCTTTTCAAAATGCACCTTGCAAACGTAACAG
GAACCCGACTAGGATCATCGGGAAAAGGAGGAGGAGGAGGAAGGCAGGCTCCGGGGAAGCTGGTGGCAGCGGGTCCTGGGTCTGGCGGACCCTGA
CGCGAAGGAGGGTCTAGGAAGCTCTCCGGGGAGCCGGTTCTCCCGCCGGTGGCTTCTTCTGTCCTCCAGCGTTGCCAACTGGACCTAAAGAGAGG
CCGCGACTGTCGCCCACCTGCGGGATGGGCCTGGTGCTGGGCGGTAAGGACACGGACCTGGAAGGAGCGCGCGCGAGGGAGGGAGGCTGGGAGTC
AGAATCGGGAAAGGGAGGTGCGGGGCGGCGAGGGAGCGAAGGAGGAGAGGAGGAAGGAGCGGGAGGGGTGCTGGCGGGGGTGCGTAGTGGGTGGA
GAAAGCCGCTAGAGCAAATTTGGGGCCGGACCAGGCAGCACTCGGCTTTTAACCTGGGCAGTGAAGGCGGGGGAAAGAGCAAAAGGAAGGGGTGG
TGTGCGGAGTAGGGGTGGGTGGGGGGAATTGGAAGCAAATGACATCACAGCAGGTCAGAGAAAAAGGGTTGAGCGGCAGGCACCCAGAGTAGTAG
GTCTTTGGCATTAGGAGCTTGAGCCCAGACGGCCCTAGCAGGGACCCCAGCGCCCGAGAGACCATGCAGAGGTCGCCTCTGGAAAAGGCCAGCGT
TGTCTCCAAACTTTTTTTTCAGGTGAGAAGGTGGCCAACCGAGCTTCGGAAAGACACGTGCCCACGAAAGAGGAGGGCGTGTGTATGGGTTGGGTT
TGGGGTAAAGGAATAAGCAGTTTTTAAAAAGATGCGCTATCATTCATTGTTTTGAAAGAAAATGTGGGTATTGTAGAATAAAACAGAAAGCATTA
AGAAGAGATGGAAGAATGAACTGAAGCTGATTGAATAGAGAGCCACATCTACTTGCAACTGAAAAGTTAGAATCTCAAGACTCAAGTACGCTACT
ATGCACTTGTTTTATTTCATTTTTCTAAGAAACTAAAAATACTTGTTAATAAGTACCTAAGTATGGTTTATTGGTTTTCCCCCTTCATGCCTTGG
ACACTTGATTGTCTTCTTGGCACATACAGGTGCCATGCCTGCATATAGTAAGTGCTCAGAAAACATTTCTTGACTGAATTCAGCCAACAAAAATT
TTGGGGTAGGTAGAAAATATATGCTTAAAGTATTTATTGTTATGAGACTGGATATATCTAGTATTTGTCACAGGTAAATGATTCTTCAAAAATTG
AAAGCAAATTTGTTGAAATATTTATTTTGAAAAAAGTTACTTCACAAGCTATAAATTTTAAAAGCCATAGGAATAGATACCGAAGTTATATCCAA
CTGACATTTAATAAATTGTATTCATAGCCTAATGTGATGAGCCACAGAAGCTTGCAAACTTTAATGAGATTTTTTAAAATAGCATCTAAGTTCGG
AATCTTAGGCAAAGTGTTGTTAGATGTAGCACTTCATATTTGAAGTGTTCTTTGGATATTGCATCTACTTTGTTCCTGTTATTATACTGGTGTGA
ATGAATGAATAGGTACTGCTCTCTCTTGGGACATTACTTGACACATAATTACCCAATGAATAAGCATACTGAGGTATCAAAAAAGTCAAATATGT
TATAAATAGCTCATATATGTGTGTAGGGGGGAAGGAATTTAGCTTTCACATCTCTCTTATGTTTAGTTCTCTGCATGTGCAGTTAATCCTGGAAC
TCCGGTGCTAAGGAGAGACTGTTGGCCCTTGAAGGAGAGCTCCTCCCTGTGGATGAGAGAGAAGGACTTTACTCTTTGGAATTATCTTTTTGTGT
TGATGTTATCCACCTTTTGTTACTCCACCTATAAAATCGGCTTATCTATTGATCTGTTTTCCTAGTCCTTATAAAGTCAAAATGTTAATTGGCAT
AAATTATAGACTTTTTTTTAGCAGAGAACTTTGAGGAACCTAAATGCCAACCAGTCTAAAAATGCAGTTTTCAGAAGAATGAATATTTCATGGATA
GTTCTAAATACTAATGAACTTTAAAATAGCTTACTATTGATCTGTCAAAGTGGGTTTTTATATAATTTTCTTTTTACAAATCACCTGACACATTT
AATATAGGTTAAAAAATGCTATCAGGCTGGTTTGCAAAGAAAATGTATTACAAAGGCTGCTAAGTGTGTTAAGAGCATACTCATTTCTGTTCTCC
AAAATATTTCATAAGGTGCTTTAAGAATAGGTATGTTTTTAAAAGTTAAGTTCCTACTATTTATAGGAACTGACAATCACCTAAAATACCAATGA
TTACAAACTTCCTTCTGGCCTTCTGGACTGCAATTCTAAAAGTGTAAAAAACATATTTTCTGCATTAAGTTAGGCAGTATTGCTTAGTTTTCAAA
GTGGTAGGCTTTGGAGTCAGATTATTTTGATTCAGATCCTACATCTACTGTTTAGTAGCTCTGTTGCCTGAGGCAGGTCCCTTAACATCTCTGTG
TGTGACTTGACCTTTAAAATTTGGAGACTGTCATAGGGGTTAATCCCTTGAGAAAATGAATGTGAAAAGTTAGCCTAATGTTAACTGCTATTATT
ATGGATTACCATATTTTCACATTCATCACAGTACATGCACCTTGTTAATATAAGATGCTCAATTCATCTTTGAGTATAATTTTGTGACTCTCAAT
CTGGATATGCAATGAGTGGGCCTGTATGAGAATTTAATTTATGAAAAATTGTGTTTCACATGGCCTTACCAGATATACAGGAAACACGTCACATG
TTTCTATTGTATGTTGTTAAATGCCTTAGAATTTAACTTTCTGAATAGGATCCCTTCAGTTTGAGAGTCATAAAAGAGTAAAATTATTATGGTAT

The Human Genome… by the Numbers

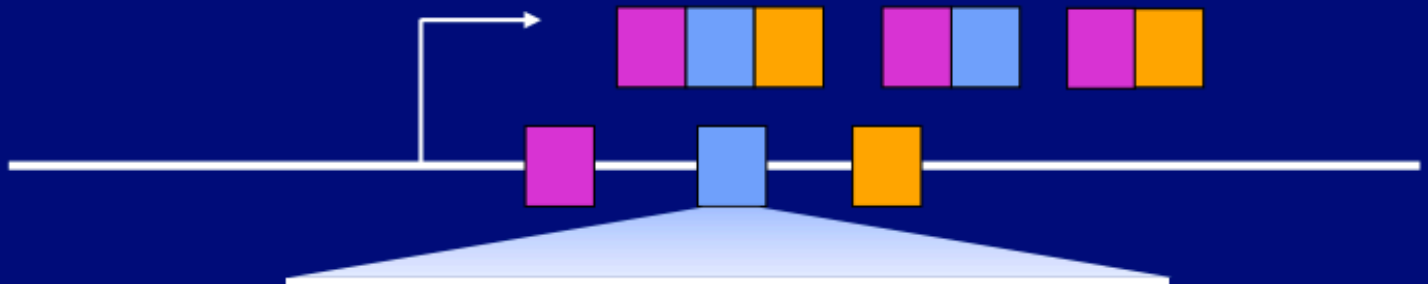
~5% of Human Genome Sequence is Constrained Across Mammals (and Presumed Functional)

- 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

TGCCGCGGAACTTTTCGGCTCTCTAAGGCTGTATTTTGATATACGAAAGGCACATTTTCCTTCCCTTTTCAAAATGCACCTTGCAAACGTAACAG
GAACCCGACTAGGATCATCGGGAAAAGGAGGAGGAGGAGGAAGGCAGGCTCCGGGGAAGCTGGTGGCAGCGGGTCCTGGGTCTGGCGGACCCTGA
CGCGAAGGAGGGTCTAGGAAGCTCTCCGGGGAGCCGGTTCTCCCGCCGGTGGCTTCTTCTGTCCTCCAGCGTTGCCAACTGGACCTAAAGAGAGG
CCGCGACTGTCGCCCACCTGCGGGATGGGCCTGGTGCTGGGCGGTAAGGACACGGACCTGGAAGGAGCGCGCGCGAGGGAGGGAGGCTGGGAGTC
AGAATCGGGAAAGGGAGGTGCGGGGCGGCGAGGGAGCGAAGGAGGAGAGGAGGAAGGAGCGGGAGGGGTGCTGGCGGGGGTGCGTAGTGGGTGGA
GAAAGCCGCTAGAGCAAATTTGGGGCCGGACCAGGCAGCACTCGGCTTTTAACCTGGGCAGTGAAGGCGGGGGAAAGAGCAAAAGGAAGGGGTGG
TGTGCGGAGTAGGGGTGGGTGGGGGGAATTGGAAGCAAATGACATCACAGCAGGTCAGAGAAAAAGGGTTGAGCGGCAGGCACCCAGAGTAGTAG
GTCTTTGGCATTAGGAGCTTGAGCCCAGACGGCCCTAGCAGGGACCCCAGCGCCCGAGAGACCATGCAGAGGTCGCCTCTGGAAAAGGCCAGCGT
TGTCTCCAAACTTTTTTTCAGGTGAGAAGGTGGCCAACCGAGCTTCGGAAAGACACGTGCCCACGAAAGAGGAGGGCGTGTGTATGGGTTGGGTT
TGGGGTAAAGGAATAAGCAGTTTTTAAAAAGATGCGCTATCATTCATTGTTTTGAAAGAAAATGTGGGTATTGTAGAATAAAACAGAAAGCATTA
AGAAGAGATGGAAGAATGAACTGAAGCTGATTGAATAGAGAGCCACATCTACTTGCAACTGAAAAGTTAGAATCTCAAGACTCAAGTACGCTACT
ATGCACTTGTTTTATTTCATTTTTCTAAGAAACTAAAAATACTTGTTAATAAGTACCTAAGTATGGTTTATTGGTTTTCCCCCTTCATGCCTTGG
ACACTTGATTGTCTTCTTGGCACATACAGGTGCCATGCCTGCATATAGTAAGTGCTCAGAAAACATTTCTTGACTGAATTCAGCCAACAAAAATT
TTGGGGTAGGTAGAAAATATATGCTTAAAGTATTTATTGTTATGAGACTGGATATATCTAGTATTTGTCACAGGTAAATGATTCTTCAAAAATTG
AAAGCAAATTTGTTGAAATATTTATTTTGAAAAAAGTTACTTCACAAGCTATAAATTTTAAAAGCCATAGGAATAGATACCGAAGTTATATCCAA
CTGACATTTAATAAATTGTATTCATAGCCTAATGTGATGAGCCACAGAAGCTTGCAAACTTTAATGAGATTTTTTAAAATAGCATCTAAGTTCGG
AATCTTAGGCAAAGTGTTGTTAGATGTAGCACTTCATATTTGAAGTGTTCTTTGGATATTGCATCTACTTTGTTCCTGTTATTATACTGGTGTGA
ATGAATGAATAGGTACTGCTCTCTCTTGGGACATTACTTGACACATAATTACCCAATGAATAAGCATACTGAGGTATCAAAAAAGTCAAATATGT
TATAAATAGCTCATATATGTGTGTAGGGGGGAAGGAATTTAGCTTTCACATCTCTCTTATGTTTAGTTCTCTGCATGTGCAGTTAATCCTGGAAC
TCCGGTGCTAAGGAGAGACTGTTGGCCCTTGAAGGAGAGCTCCTCCCTGTGGATGAGAGAGAAGGACTTTACTCTTTGGAATTATCTTTTTGTGT
TGATGTTATCCACCTTTTGTTACTCCACCTATAAAATCGGCTTATCTATTGATCTGTTTTCCTAGTCCTTATAAAGTCAAAATGTTAATTGGCAT
AAATTATAGACTTTTTTTTAGCAGAGAACTTTGAGGAACCTAAATGCCAACCAGTCTAAAAATGCAGTTTTCAGAAGAATGAATATTTCATGGATA
GTTCTAAATACTAATGAACTTTAAAATAGCTTACTATTGATCTGTCAAAGTGGGTTTTTATATAATTTTCTTTTTACAAATCACCTGACACATTT
AATATAGGTTAAAAAATGCTATCAGGCTGGTTTGCAAAGAAAATGTATTACAAAGGCTGCTAAGTGTGTTAAGAGCATACTCATTTCTGTTCTCC
AAAATATTTCATAAGGTGCTTTAAGAATAGGTATGTTTTTAAAAGTTAAGTTCCTACTATTTATAGGAACTGACAATCACCTAAAATACCAATGA
TTACAAACTTCCTTCTGGCCTTCTGGACTGCAATTCTAAAAGTGTAAAAAACATATTTTCTGCATTAAGTTAGGCAGTATTGCTTAGTTTTCAAA
GTGGTAGGCTTTGGAGTCAGATTATTTTGATTCAGATCCTACATCTACTGTTTAGTAGCTCTGTTGCCTGAGGCAGGTCCCTTAACATCTCTGTG
TGTGACTTGACCTTTAAAATTTGGAGACTGTCATAGGGGTTAATCCCTTGAGAAAATGAATGTGAAAAGTTAGCCTAATGTTAACTGCTATTATT
ATGGATTACCATATTTTCACATTCATCACAGTACATGCACCTTGTTAATATAAGATGCTCAATTCATCTTTGAGTATAATTTTGTGACTCTCAAT
CTGGATATGCAATGAGTGGGCCTGTATGAGAATTTAATTTATGAAAAATTGTGTTTCACATGGCCTTACCAGATATACAGGAAACACGTCACATG
TTTCTATTGTATGTTGTTAAATGCCTTAGAATTTAACTTTCTGAATAGGATCCCTTCAGTTTGAGAGTCATAAAAGAGTAAAATTATTATGGTAT

Coding Sequences (i.e., Genes)



First Letter	Second Letter: T	Second Letter: C	Second Letter: A	Second Letter: G	Third Letter
T	TTT Phe, TTC Phe, TTA Leu, TTG Leu	TCT Ser, TCC Ser, TCA Ser, TCG Ser	TAT Tyr, TAC Tyr, TAA Stop, TAG Stop	TGT Cys, TGC Cys, TGA Stop, TGG Trp	T C A G
C	CTT Leu, CTC Leu, CTA Leu, CTG Leu	CCT Pro, CCC Pro, CCA Pro, CCG Pro	CAT His, CAC His, CAA Gln, CAG Gln	CGT Arg, CGC Arg, CGA Arg, CGG Arg	T C A G
A	ATT Ile, ATC Ile, ATA Ile, ATG Met	ACT Thr, ACC Thr, ACA Thr, ACG Thr	AAT Asn, AAC Asn, AAA Lys, AAG Lys	AGT Ser, AGC Ser, AGA Arg, AGG Arg	T C A G
G	GTT Val, GTC Val, GTA Val, GTG Val	GCT Ala, GCC Ala, GCA Ala, GCG Ala	GAT Asp, GAC Asp, GAA Glu, GAG Glu	GGT Gly, GGC Gly, GGA Gly, GGG Gly	T C A G

The Genetic Code

The Human Genome… by the Numbers

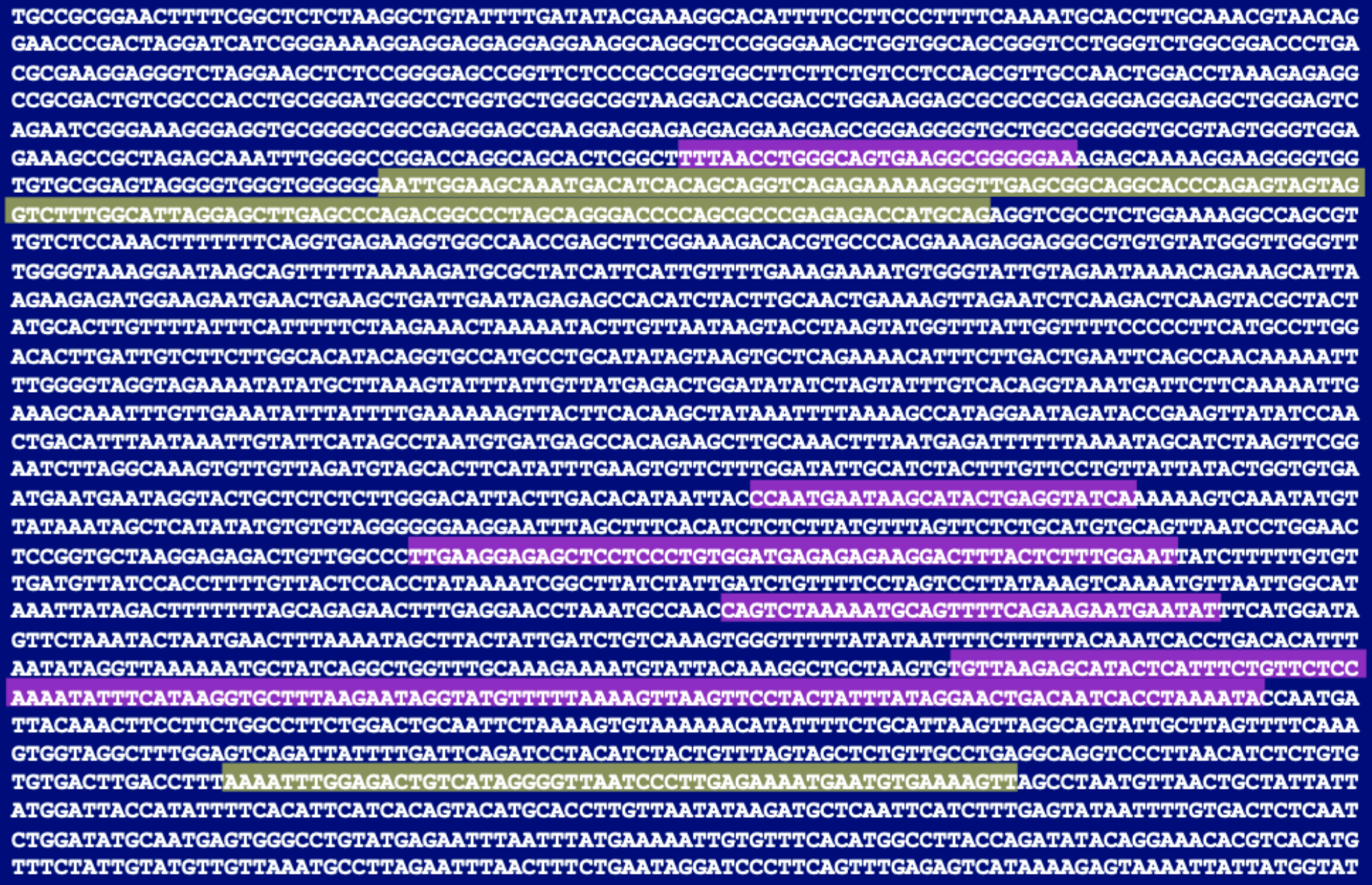
~5% of Human Genome Sequence is Constrained Across Mammals (and Presumed Functional)

- 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

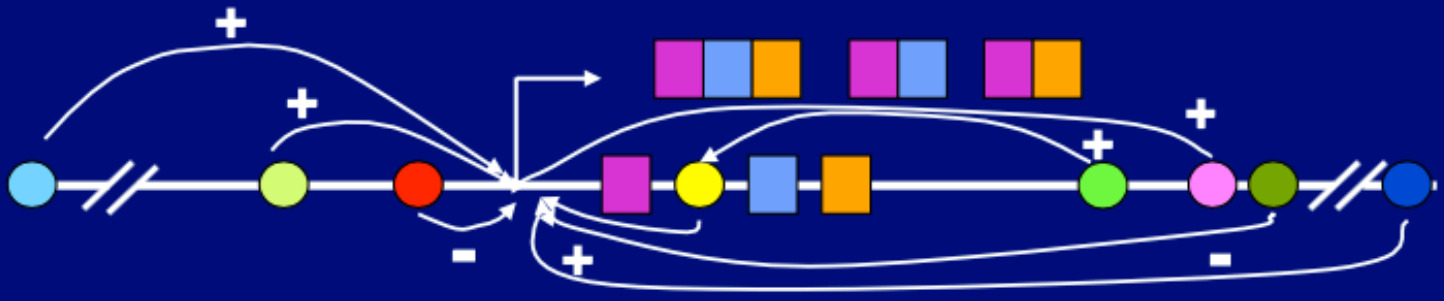
~1.5% Encodes for Protein (Genes)

- Corresponds to ~18-22K Genes
- Many More than ~22K Different Proteins
- Good Inventory at Present

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

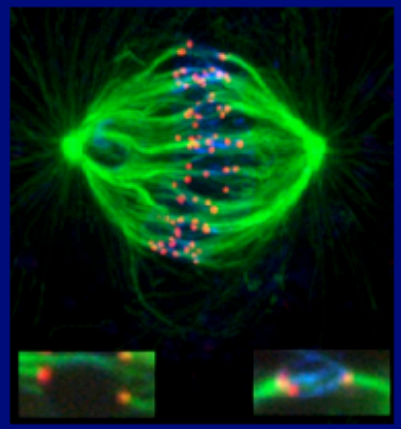


Non-Coding Functional Sequences

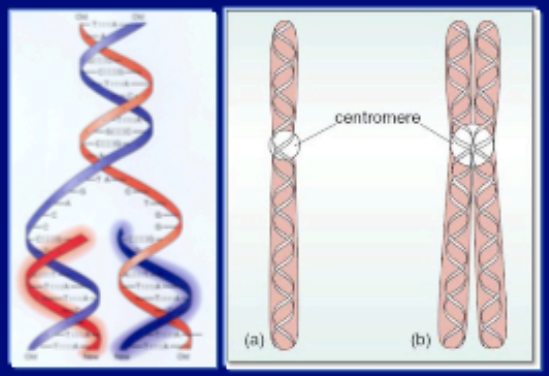


Gene Regulation

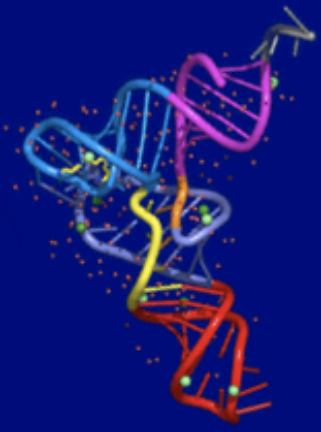
Chromosome Packaging



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)

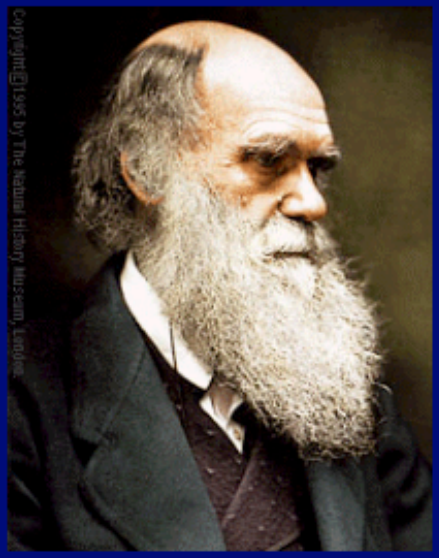
- Corresponds to ~18-22K Genes
- Many More than ~22K Different Proteins
- Good Inventory at Present

~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)



Charles Darwin (1809-1882)

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

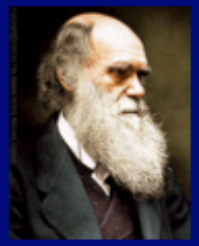
–Eric Lander

Comparative Sequence Analysis

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

Species A

```
GATCGTCTAGAATCTCGAGATC
TCTGAGAGTCGTGGGAAACTGT
GTGATGTGACGATTTAGCCACA
GTTACGTGTGAGAGATGTATGA
TGCACCTGACCCGGGTTTCACT
CTCAACGACTCACTCCACCTCA
GAGGCCCACCGCCGCTGTGCAC
TACCGAGATACACGATACCTAC
ACAGGTGTGACACACCCCTACC
CGTCCACCACACGACTCACTCC
ACCTCAGAGGCCCACCGCCGCT
GTGCACTACCGAGATACACGAT
ACCTACACAGGTGTGACACACG
ATCCTTACCACACTTACACATT
ACCATATATCCACCTACCACAC
ATACCTACCCCATTGCACACCT
ATTATTATTACCGGGACCGAGG
```



Compare

Species B

```
TATCGGCTAGAATCTCGAGATC
TCTGAGAGTCGTGGGAAACTGT
GTGATGTGACTAGCCACAGTTA
CGTGTGAGAGATGTATGATGCA
CCTGACCCGGGTTTCACTCTCA
ACGACTCACTCCACCTCAGAGG
CCCACCGCCGCTGTGCACGTCC
ACCACGATCCTTACCACACTTA
CACATCACTCTCAACGACTCAC
TCCACCTCAGAGGCCCACCGCC
GCTGTGCACGTCCACCACGATC
CTTACCACACTTACACATTACC
ATATATCCACCTACCACACATA
CCTTACCATATATCCACCTACC
ACACATACCTACCCCATTGCAC
ACCTATTATTATTACCGAGGGA
GAGGGGTGACCACACTGTGACA
```

```
GATCGTCTAGAATCTCGAGATC
TCTGAGAGTCGTGGGAAACTGT
GTGATGTGACTAGCCACAGTTA
CGTGTGAGAGATGTATGATGCA
CCTGACCCGGGTTTCACTCTCA
ACGACTCACTCCACCTCAGAGG
CCCACCGCCGCTGTGCACGTCC
ACCACGATCCTTACCACACTTA
CACATCACTCTCAACGACTCAC
TCCACCTCAGAGGCCCACCGCC
GCTGTGCACGTCCACCACGATC
CTTACCACACTTACACATTACC
ATATATCCACCTACCACACATA
CCTTACCATATATCCACCTACC
ACACATACCTACCCCATTGCAC
ACCTATTATTATTACCGAGGGA
GAGGGGTGACCACACTGTGACA
```

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

Vertebrate Genome Sequences

Mouse

Rat

Chicken

Chimpanzee

Dog

Macaque

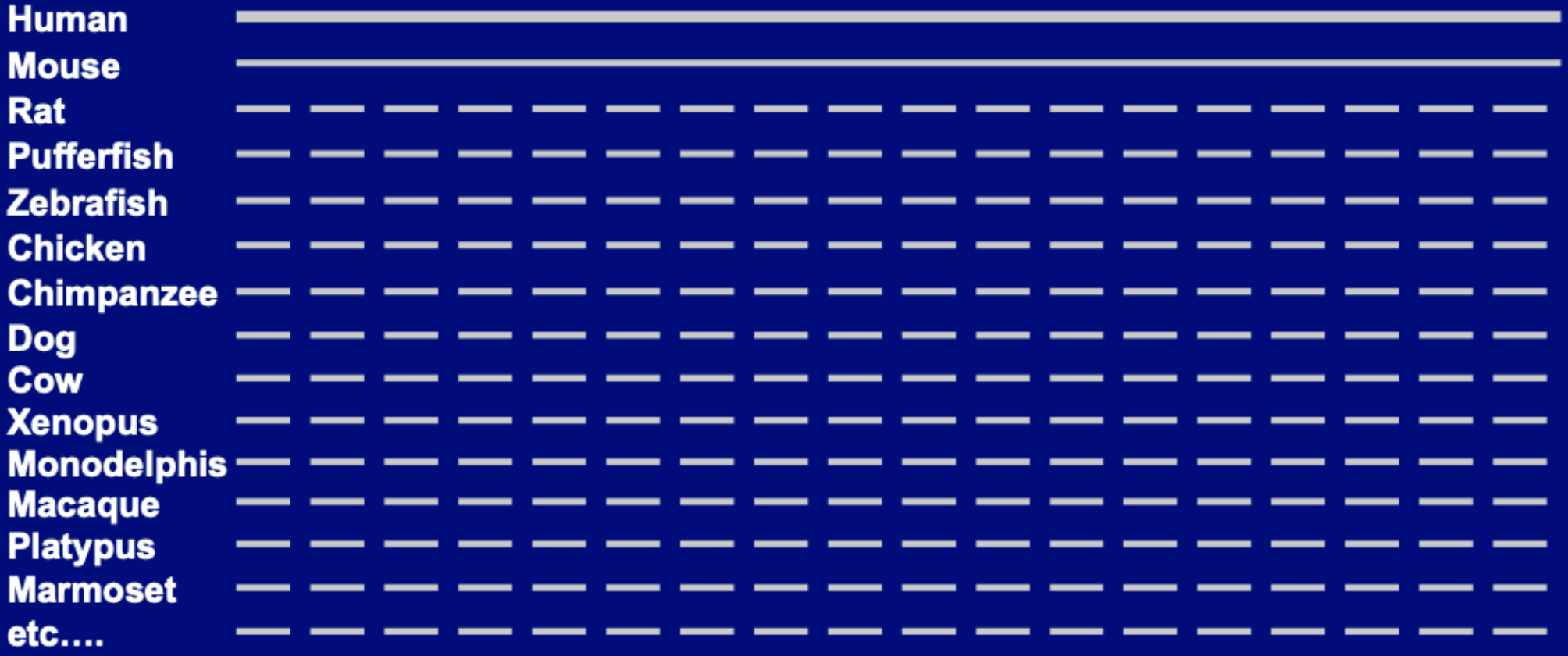
Monodelphis

Platypus

Cow

Pufferfish

Diverse Landscape of Genome Sequencing

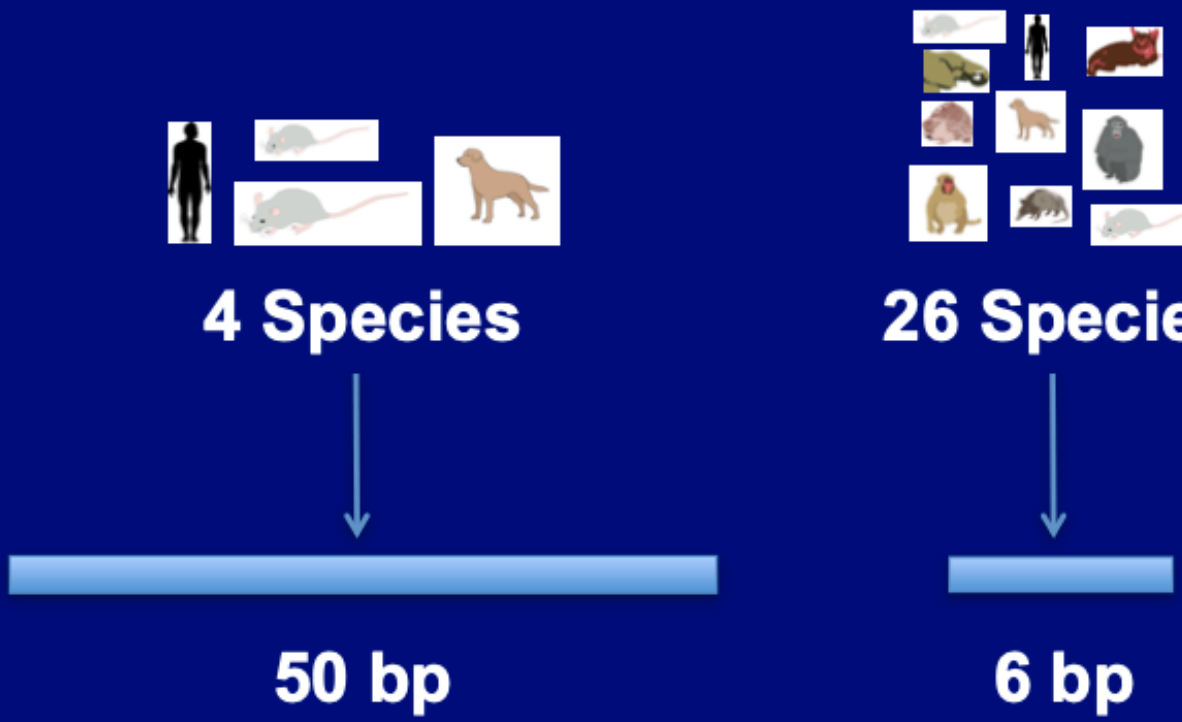


More Species = More Power

A Model of the Statistical Power of Comparative Genome Sequence Analysis

Sean R. Eddy

PLoS Biology (2005)

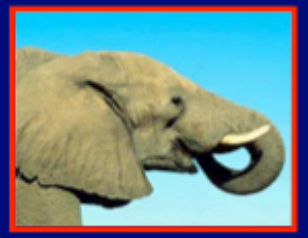


‘Light Sampling’ of Many Mammalian Genomes

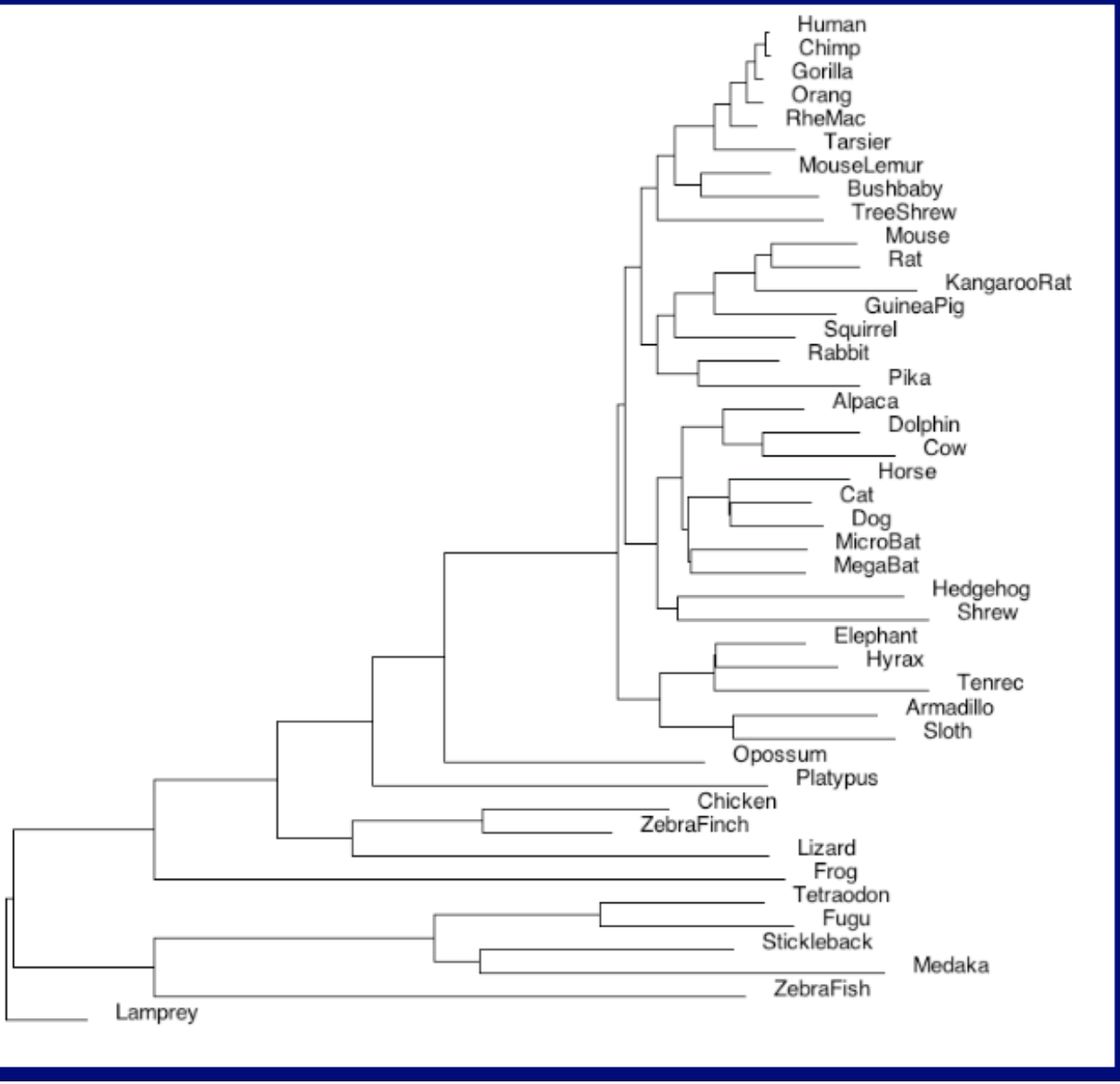
An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing

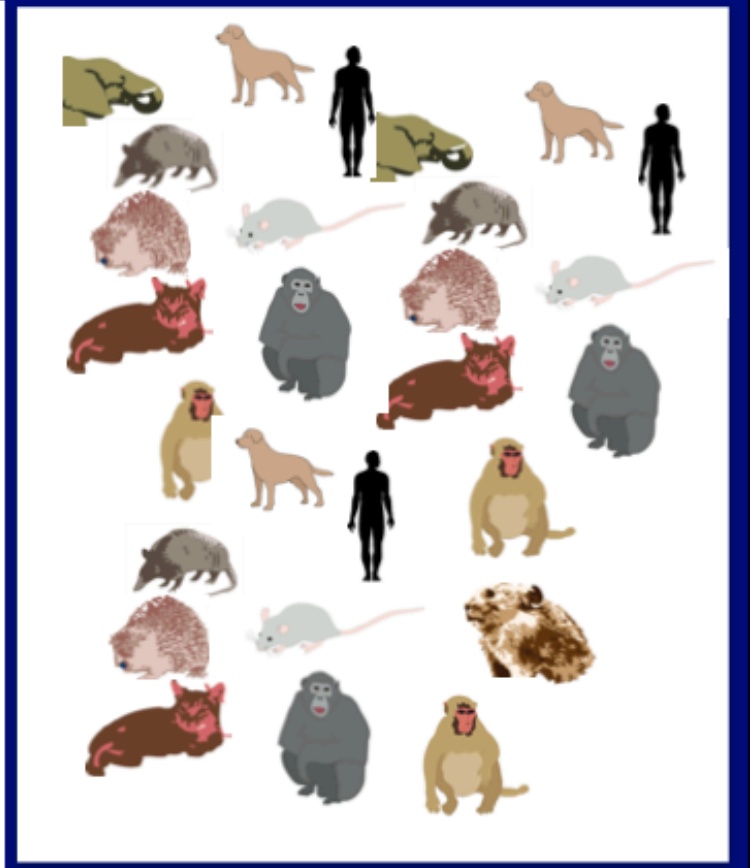
Elliott H. Margulies*†, Jade Vinson†‡, NISC Comparative Sequencing Program*§¶, Webb Miller‖, David B. Jaffe‡, Kerstin Lindblad-Toh‡, Jean Chang‡, Eric D. Green*§, Eric S. Lander‡, James C. Mullikin*§**, and Michele Clamp‡**

PNAS (2005)

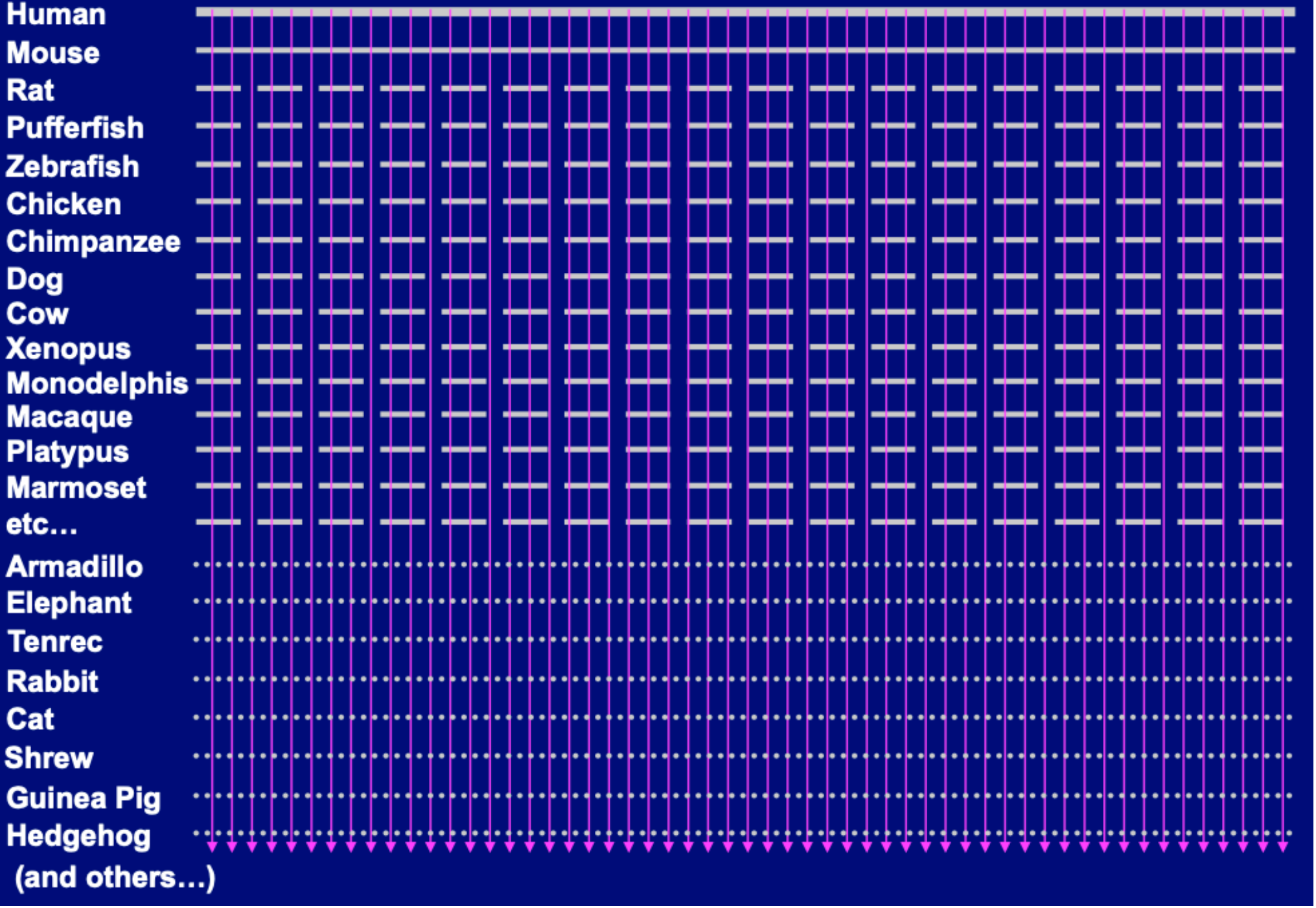


22 Additional Mammalian Genome Sequences (@ Low Redundancy)





Diverse Landscape of Genome Sequencing



Multi-Species Sequence Comparisons

Comparative genomics in e!Ensembl

EMBL-EBI

Training materials

- Ensembl training materials are protected by a CC BY license:

  creativecommons.org/licenses/by/4.0/

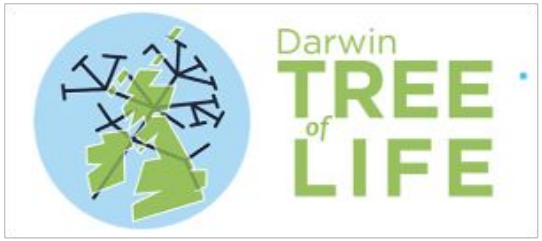
- If you wish to re-use these materials, please credit Ensembl for their creation
- If you use Ensembl for your work, please cite our papers:

  ensembl.org/info/about/publications.html

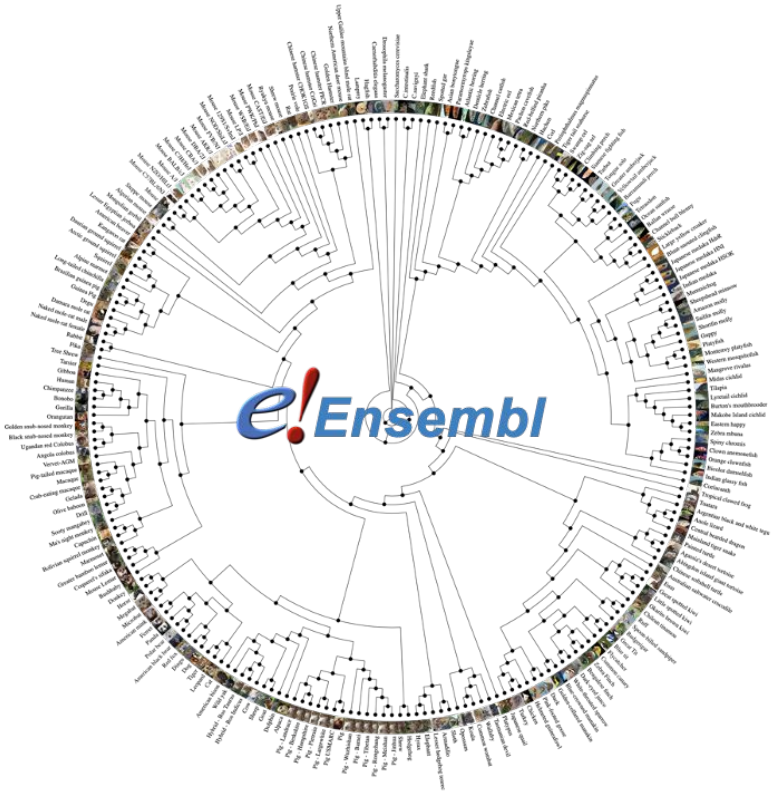
Applications

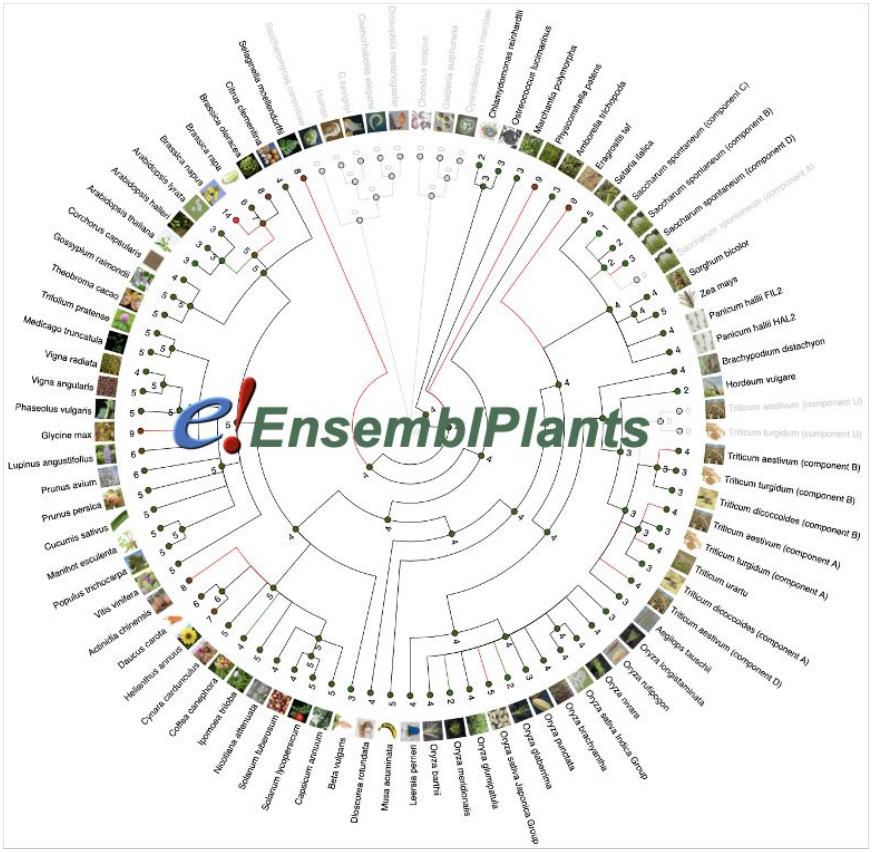
Comparative genomics allows us to:

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



Comparative analysis by taxa

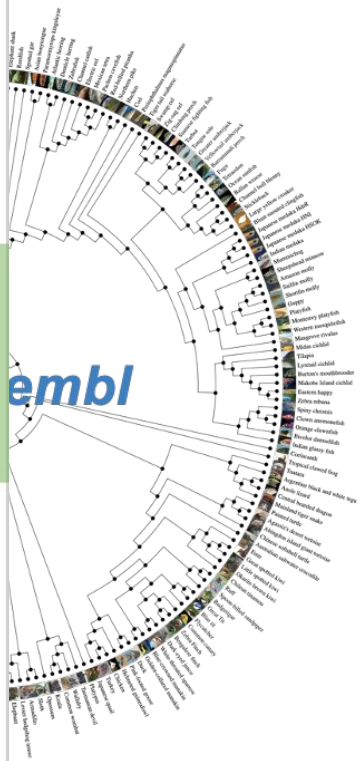


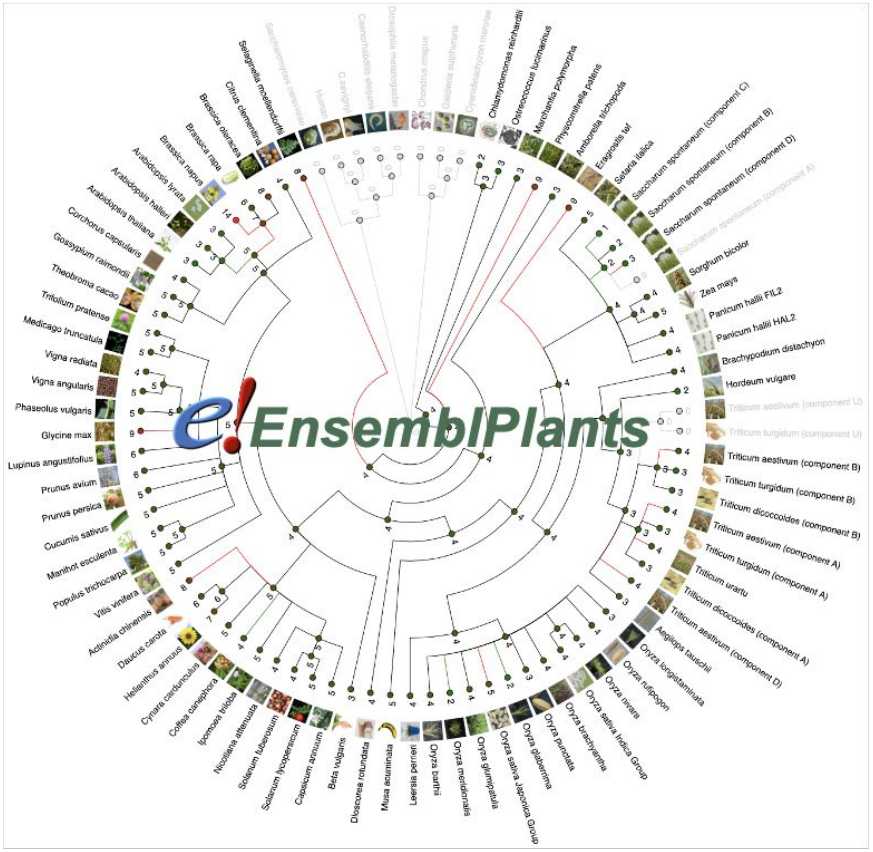


Comparative analysis by taxa

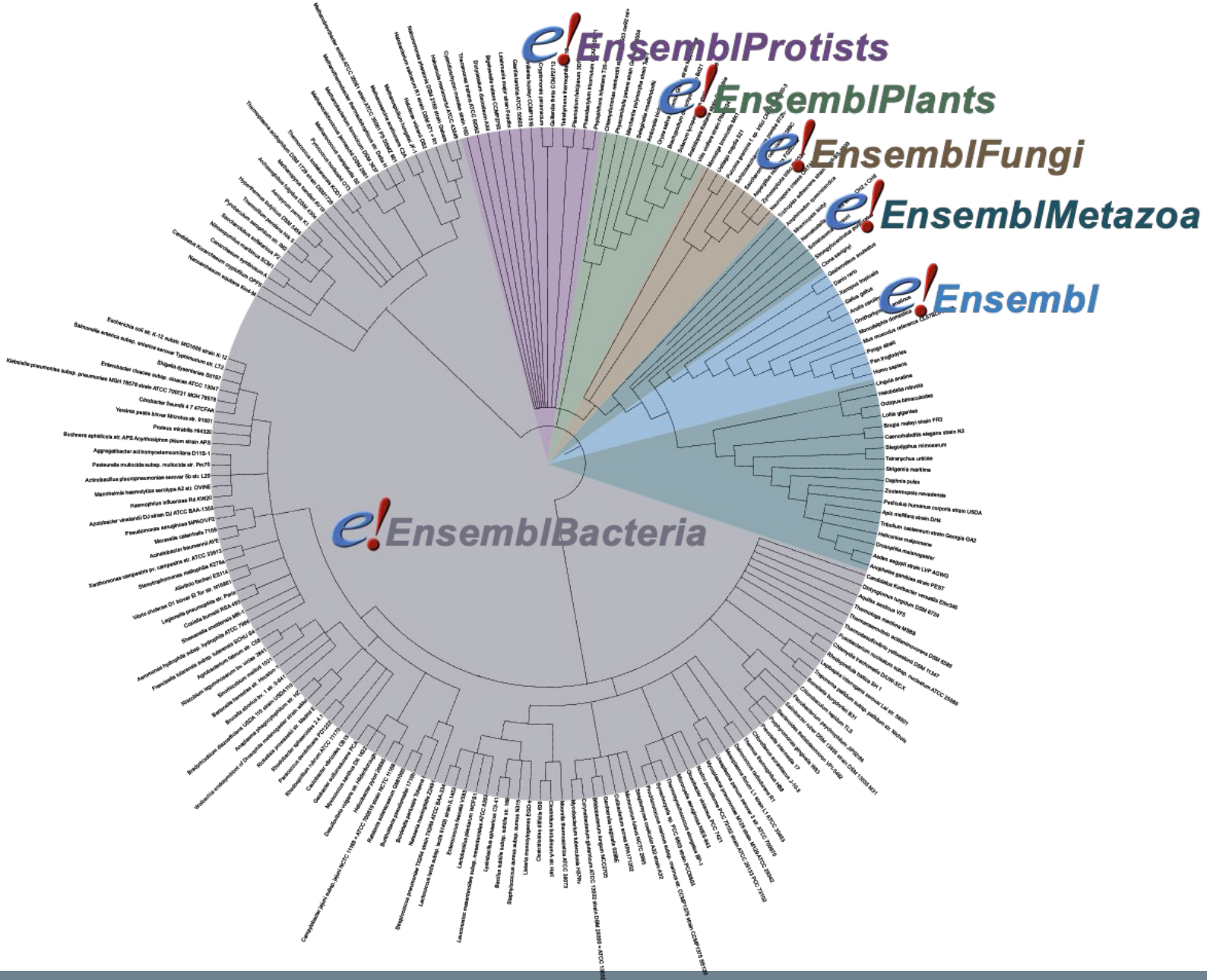
Gene-based displays

- **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





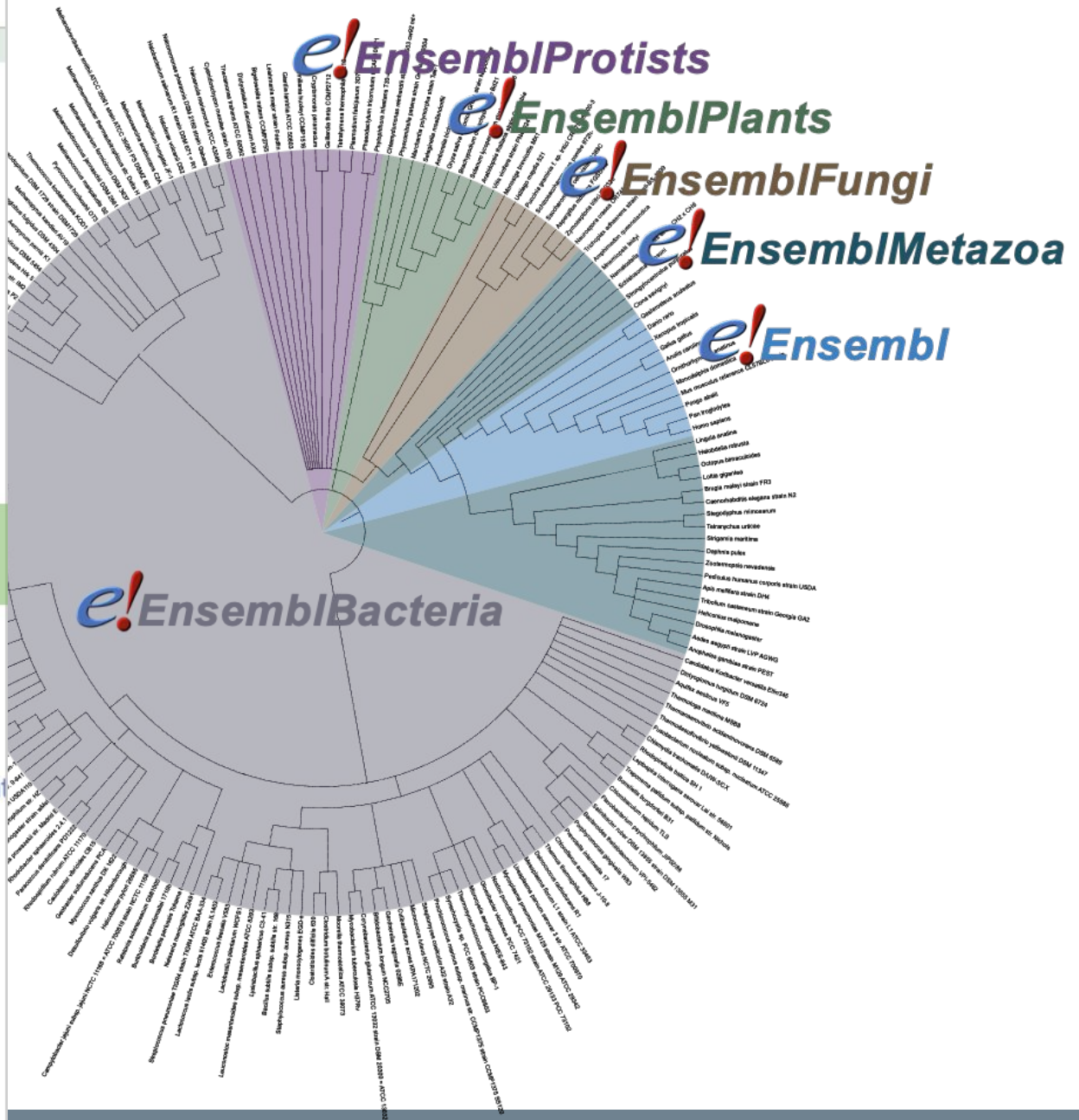
Pan-taxonomic compara



Pan-taxonomic compara

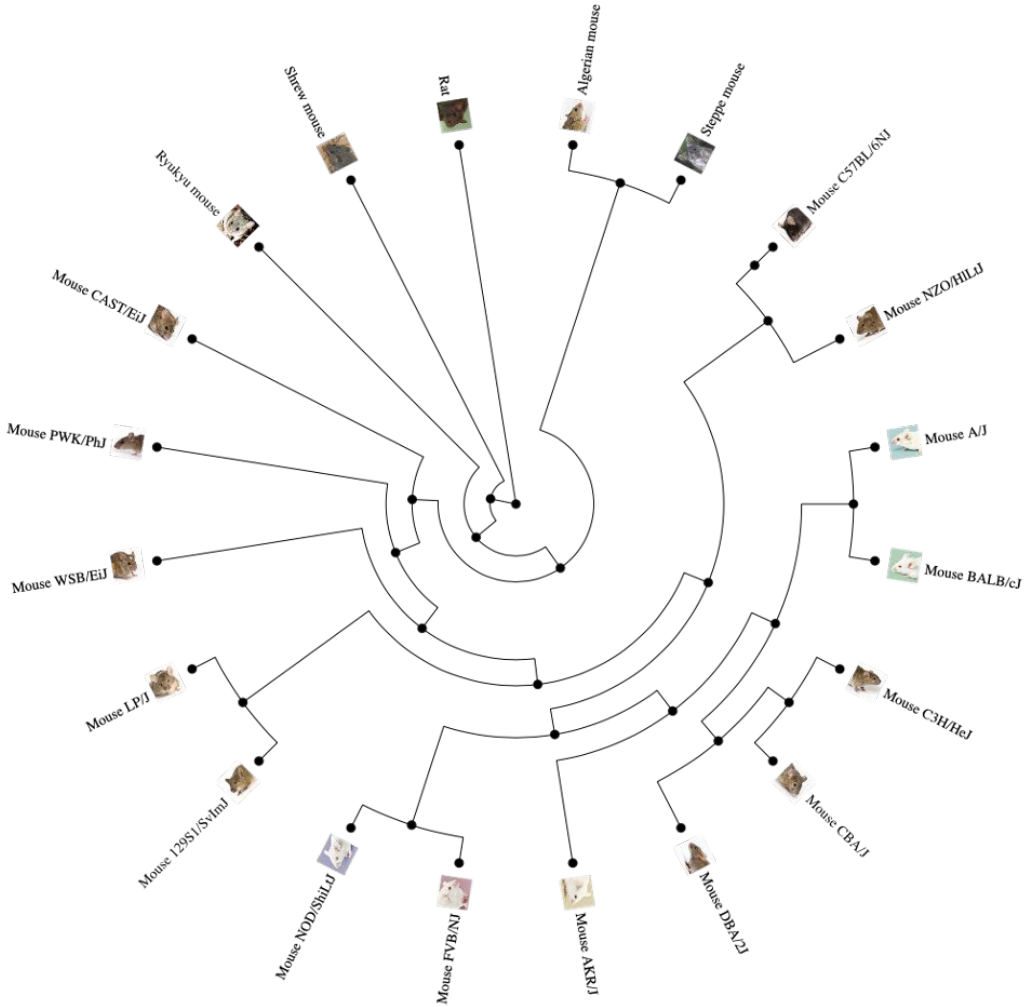
Gene-based displays

- 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



EMBL-EBI

Mouse compara



Types of data

Gene-based resources (found in the gene tab)

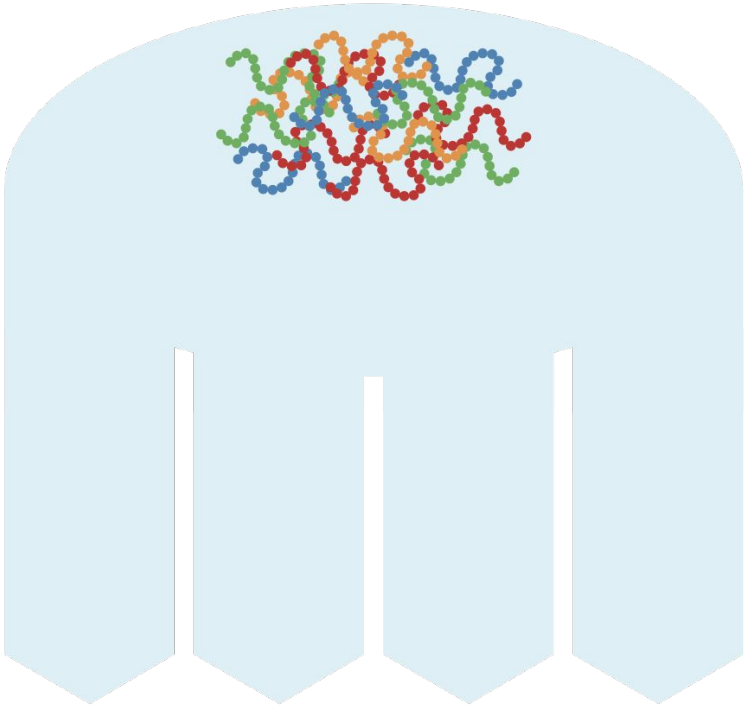
- Phylogenetic trees and tree-inferred homology
- Protein trees
- ncRNA trees
- Stable ID mapping
- Protein families

Sequence-based resources (found in the location tab)

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

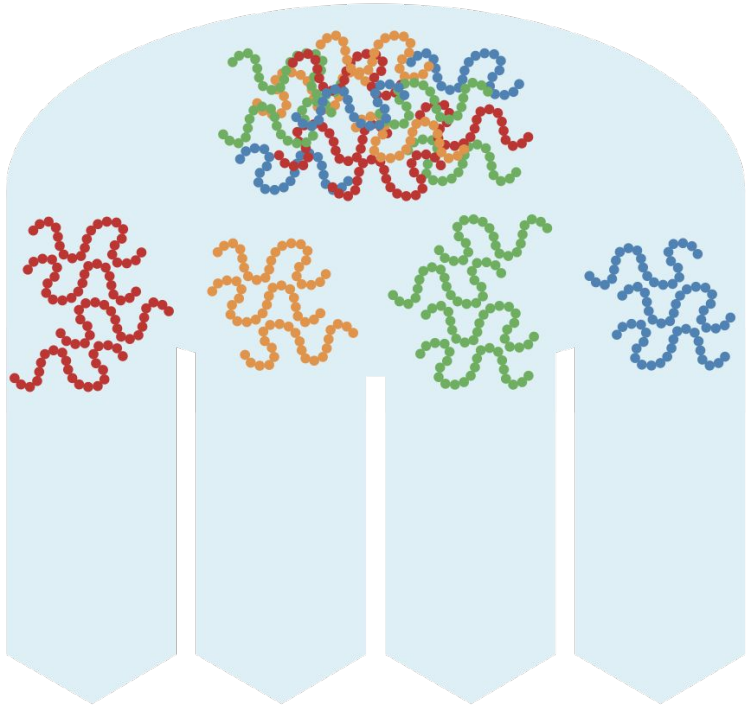
Gene/protein trees

1. Representative translation of each gene from all species



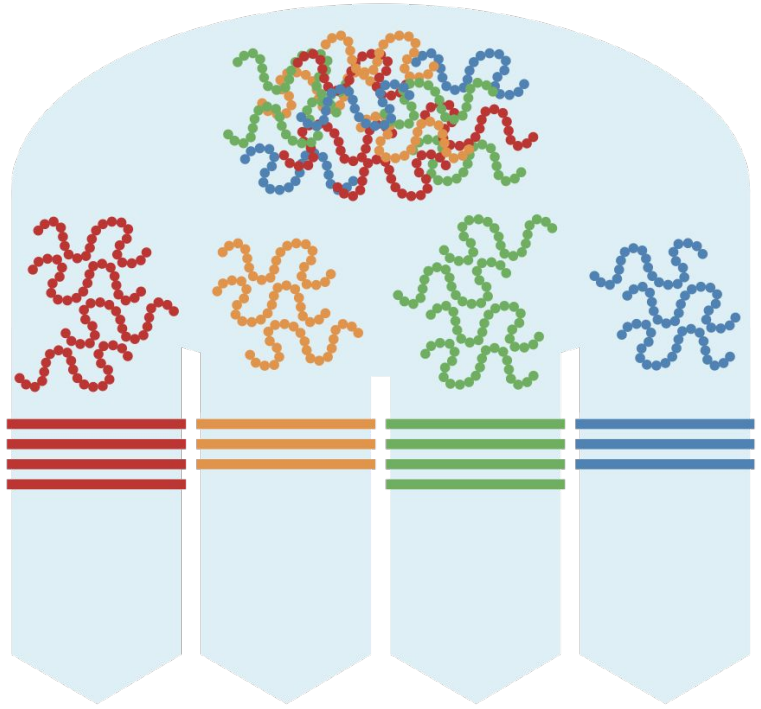
Gene/protein trees

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



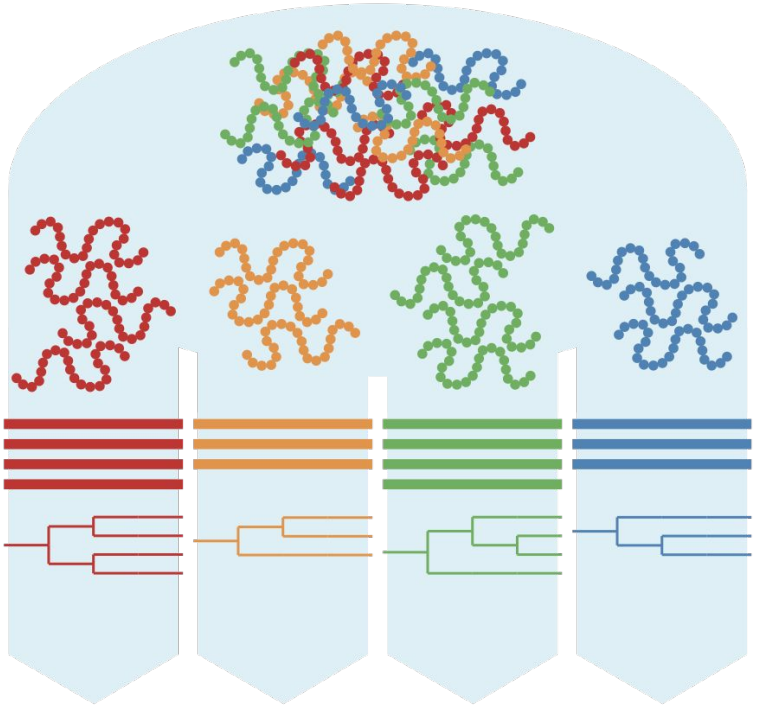
Gene/protein trees

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



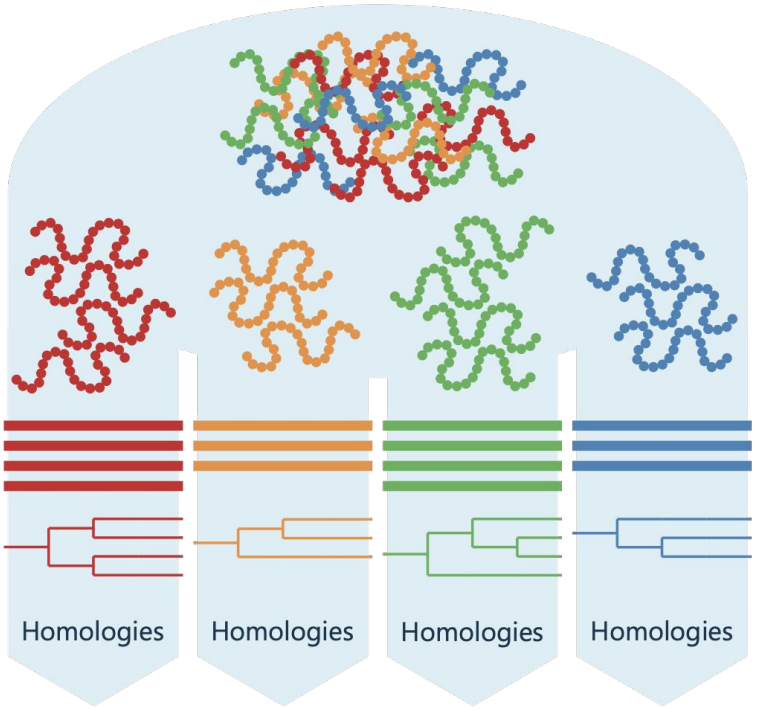
Gene/protein trees

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



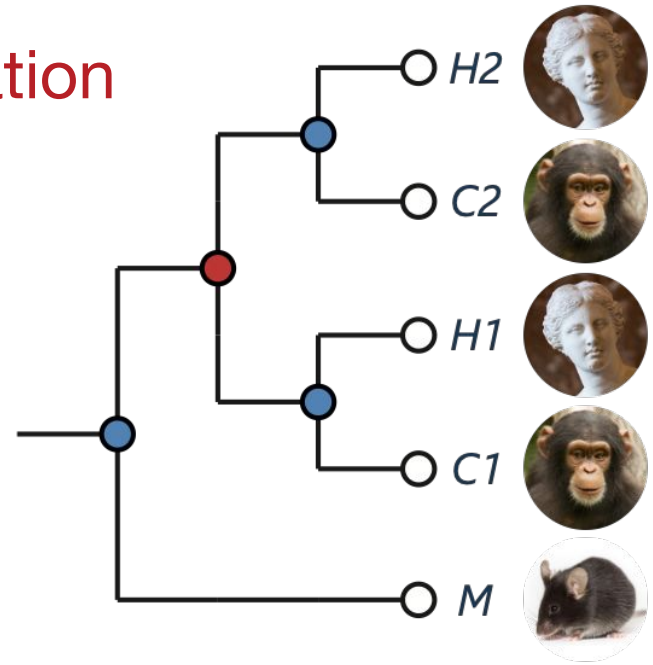
Gene/protein trees

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-/paralogue inference

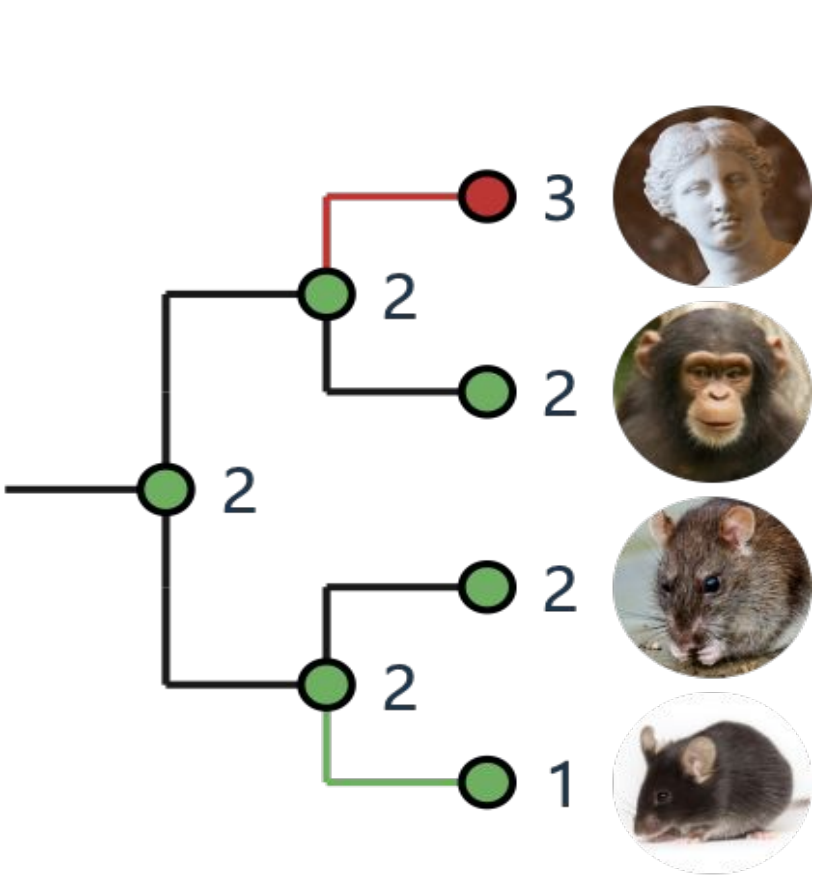


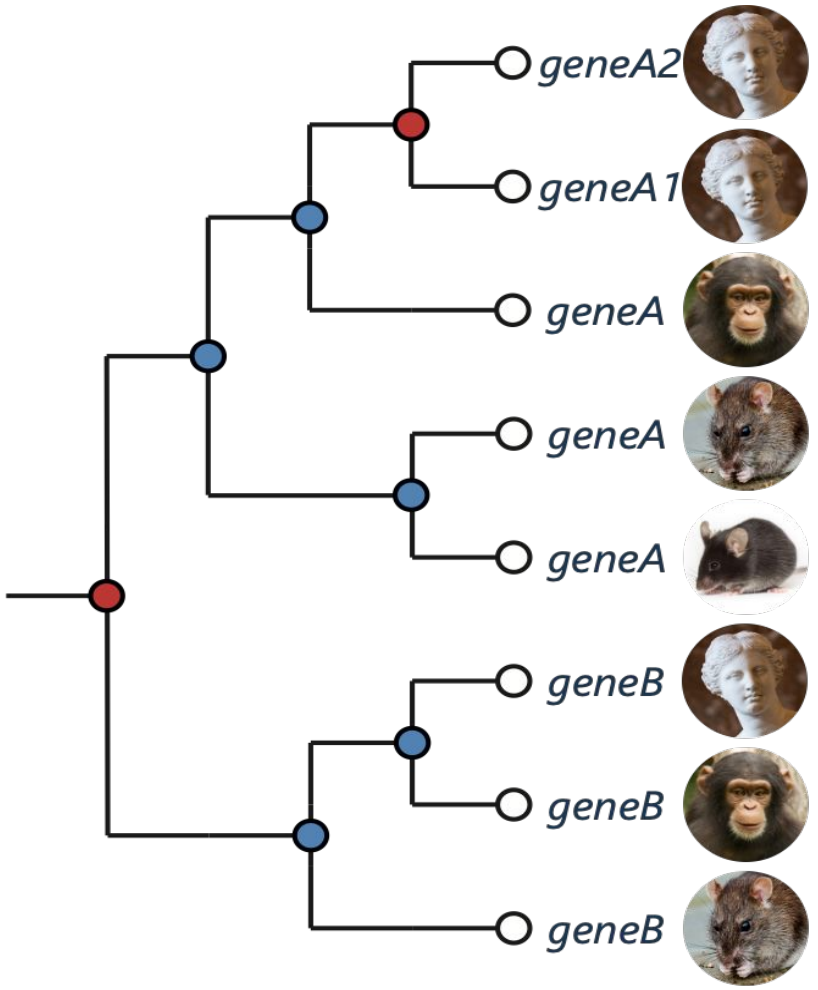
Homology relationships

- 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

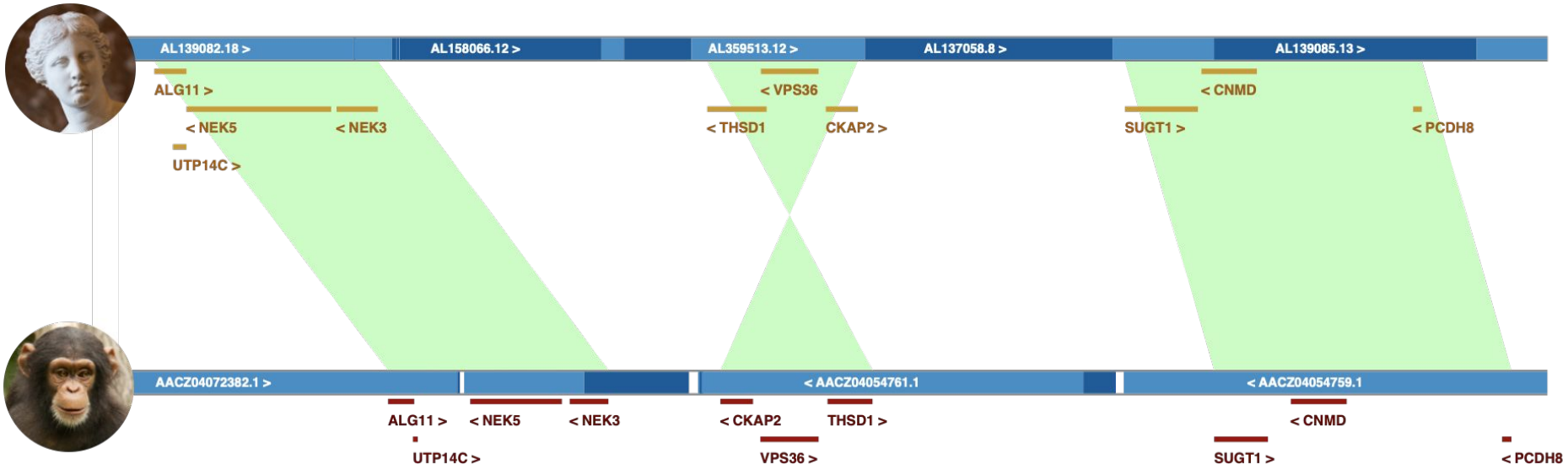
Pairwise alignments with BLASTZ (older) or LASTZ (newer):

- 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