PR10_Evolucijska in komparativna genomika ter filogenomika

The DNA Record is a Living Chronicle of Evolution

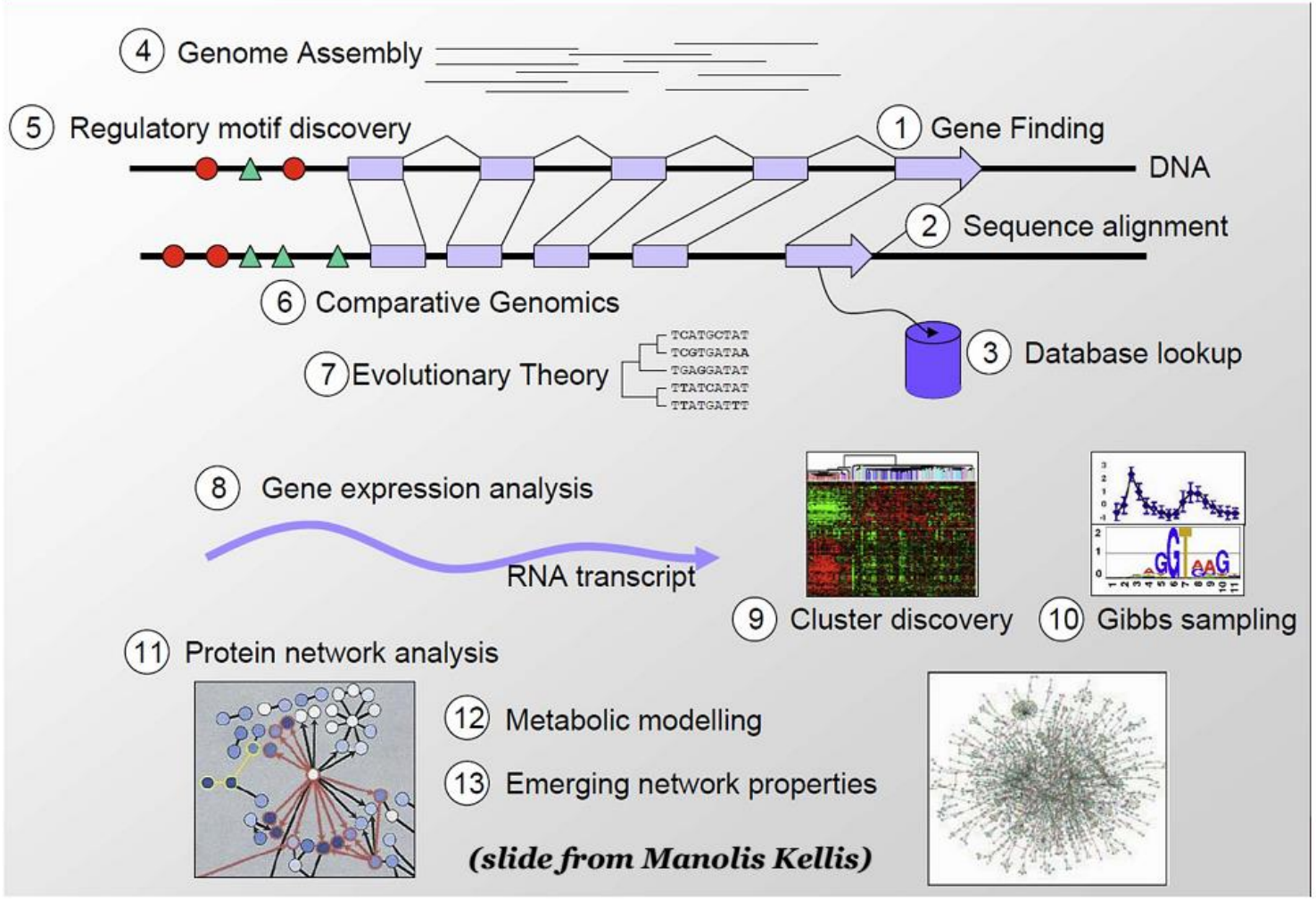
- **Information in genomes is vital to reconstructing the processes and patterns of evolution**

- **Knowledge of evolution is a powerful guide to interpreting genomes**

What is a Genome Like?

ACAACCCCTCCACCTCATGTACCTGCGGACTCTCCTCCAGTCACAGCTCAGGCAGTCCACTTTGCAACCCCTAAACCTCAAAACCGGTTT
GACGTTCTGTTAGACGAACAACTATGATATATCGACCCCGCCTAAGAACGGAGCCTCTGTCAGTGCTCCAGCTGAACGTAGGCCGCGGG
CCAGCCACTCATGAAATCGCCCTTTCATTAGCATACTCTAGCGGCATTGATATCATCCTTATACAGGAGCCATACATATATACTGACCTCA
GCCGGCAAATCACAAAAAGGCACCCATCATACGAGTGCTTCTCCCCAACAGACAGCTGGCTTGTAAGCGGTCGACCCCGGGTCCTCACC
TATGTCCGGAAAAAGATGGGCATTCGGGCCTCTCAGCTCCGCCCTCAGCCAATAGATCAAGATGTTCTCTCAGACCTTCTTCTACTACAG
ATCCTCTCCCGCTCTGGACAATCTGCATTGATAATCAACATCTATAATGCTCCAATCGGCTCAATCAGGTCAGGTGAGGCTGCAAAAGCG
CTTACACTCCTGCCTGACTCCTACTTTTCCCAGCCTACCGTGCTTGCCGGCGACTTCAACCTACTACATAGCAGGTGGCAGCCATCACTG
CATTGCAGCCCTACCACCTTTGCTGAGCCATTTGTTGACTGGCTTGATCGCCTAGGGCTGGTTCTTATCTCCGAGATAGACCAGCCTACAC
ACGATAGAGGCAACGTTCTTGACCTCACTTTCGCCTCCAGCTCCCTAGCACTGGCAGGGTCGAGTACCAGGATAGCAAGTCATTTAGAGT
CAACATCAGATCATCGGCCACTCCTCACCACCATGCCATGGAGCCAGAGATTCACAGAGGCAGCTCAGAAACTGAGATTTGATACATTA
GACCACCCTCGCTTCCTCTCACTACTCAGTTCCCACCTTGCTGTCATTGAATGCTCAGCTACAACAGAAGAGGGCCTGGACAGTCTAGCT
CATGGGTTAACCTTAGCAACTGCTAGTGCGTATAAAGGCTCTGCTAGGAGCTCCTTGGCGCAGGGAATAGGTCAGCCATGGTGGAATATT
GACTGCAGAAAAGCGTTGCAAGACTTCCGCTTAGGTCTCTGTTCAAGAAACGACTTCCGTCGGATAACTAGACGGTCTAAATAGCAGTTC
TGGCGAGATAAACTTACCGCAGTGACACAGATCAAAGATGTCTTTGACATAAGCAAGTGACATAAGTTTACAGGATCTTATCGAAACCCT
CCACTAAACGACCCTTTAAGGCCAAACAGCCCTCCAGCAGGGGCTCTGAATGAGAAACAAGACGTATTAGTCCGTAATCTTCTTCAGAAT
ACTGCTGAAGCGGGTGATATTGTCATAGGCTATGGCCTGGGCTGTGGTTGTCAGCCATGCCCTCAACCATAGAACATTCTAGAAGAACCA
TCGGGAAGAGGTTGGAACCCAGTGGAAGTTTGGGAACATGTATATAAGAAGGAGAGGGAGATGTATCTGCCTATTTCTCTCTCCAAGTCT
GCGATATTCGTTAATACATTATACAGGATTGCCAGTTGAAAACAATACTGCCTACGCCCGTCACAGGTACTGCAGTTTCCAACAAGAATC
AACGCTCGACCCGGCAATTATGGCTCAAGGTTAGACTACGTCCTGTGTAGCCTTGATATGCAAGATTAGTTCTGCGATTTGAATATCTAAG
AGGATCTAATGGTAAGCCCCAAGGCTGCCATGGCTTTATTGTAGATTGATTTTCTAGCTGACAATATGCAATTTGGGACAGGGATCTGATG
ATTGTCCGGTTTATGCTGTCTTCAAAAATGTTATACGCCTCGGCGAAGAAGAGGTCAACATTAAATGAGCCCTCCTGGGATGTTTAAAGAT
GGCGAGCGTCAGCAGGAATACTCTACTAAATATCTTCTGCCTACATCAGGGCGCTTAATACCAGAATTTAACAAGCGGAGGAGGATCAA
GGACATGTTCTTGCGTAAACCATCAGCCAACGTATAGAGACCGACGACGAACATCCTGACATTGAGATATTTTACCTCTAGTCAGGAAAA
GGGAACAGCACCCGCTATTTTGGAGAGTGCTGCCAGCGTCATAGCTACCTGCCAGCCTGTAGTAGCTGCTGACAGCACTCAAATGAAAG
AAGTTATTCGTAAGAGCTCTCAGAAATATGAGACAGGTTCCCCTGTCTCAGTCCAGTATTTGACATCGGGTTCAGCCCAATCATCAACAC
CCCCCACTGCTGGACAGAGGACTCTAAAGGGGTTCTTCAAACTTAAAAGTGGTCTAGCCAGCCAAATGGCCATAGCCCAGGATCCTGCA
ACAGTGTCTACTATGCCAACGAAACAACCAGCCGCATCCCCTACAAAATCTACCCCAGTTACAGAACCTCCTGCACTGGAAGCATTACTG
ACAGCTCCCGCTGGTGAAGCTTCTCCAGGAGAACAGCCAAATTCCGCGACTCCTACAGCTCCCGCTTCACCCCAAAGCAATGATACTATT
ATCGATCCCATTGTCAGCAAGGAAGATTGGTCAAAGCTCTTCACTAAAAAGCCCATTCCCAAGTGCGAGGGCCACCAGGAACCATGTTT
CAGTCTGACAACTAAGAAGCCTGGCATCAACTGCGGAAGATCGTTCTGGATCTGTTTGAGACCCCTTGGGCCCAGCGGAAACAAGGAAA
AGGGGATACAGTGGCGATTTCCTACATTCATATGGGCCAGCGATTGGAACCCTTCCGCTCCGTAGATTTTCTGTCTGGGGCAACTTCTTTT
TGCGATAGTGTAACGATACCCGGTTTTATACTTAGAAGGCTACGAATGGTATGATGTATCATGGTTTCAATGATAAGACATTTCGTCAAGT

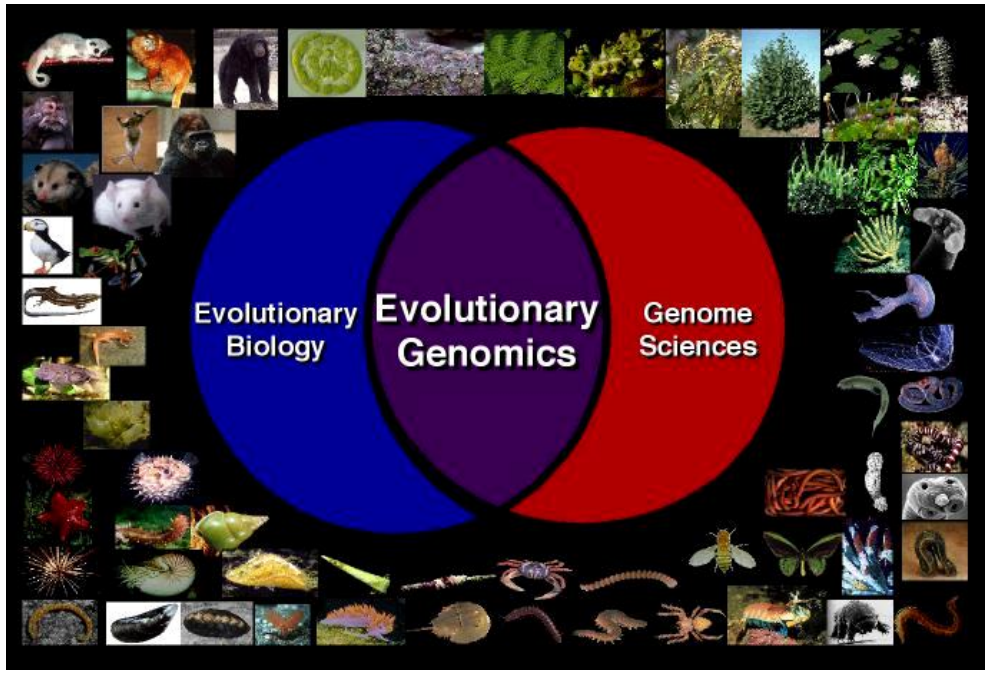
Understanding Sequences Requires Tools and Evolution



What is a Genome Like?

ACAACCCCTCCACCTCATGTACCTGCGGACTCTCCTCCAGTCACAGCTCAGGCAGTCCACTTTGCAACCCCTAAACCTCAAAACCGGTTT
GACGTTCTGTTAGACGAACAACTATGATATATCGACCCCGCCTAAGAACGGAGCCTCTGTCAGTGCTCCAGCTGAACGTAGGCCGCGGG
CCAGCCACTCATGAAATCGCCCTTTCATTAGCATACTCTAGCGGCATTGATATCATCCTTATACAGGAGCCATACATATATACTGACCTCA
GCCGGCAAATCACAAAAAGGCACCCATCATACGAGTGCTTCTCCCCAACAGACAGCTGGCTTGTAAGCGGTCGACCCCGGGTCCTCACC
TATGTCCGGAAAAAGATGGGCATTCGGGCCTCTCAGCTCCGCCCTCAGCCAATAGATCAAGATGTTCTCTCAGACCTTCTTCTACTACAG
ATCCTCTCCCGCTCTGGACAATCTGCATTGATAATCAACATCTATAATGCTCCAATCGGCTCAATCAGGTCAGGTGAGGCTGCAAAAGCG
CTTACACTCCTGCCTGACTCCTACTTTTCCCAGCCTACCGTGCTTGCCGGCGACTTCAACCTACTACATAGCAGGTGGCAGCCATCACTG
CATTGCAGCCCTACCACCTTTGCTGAGCCATTTGTTGACTGGCTTGATCGCCTAGGGCTGGTTCTTATCTCCGAGATAGACCAGCCTACAC
ACGATAGAGGCAACGTTCTTGACCTCACTTTCGCCTCCAGCTCCCTAGCACTGGCAGGGTCGAGTACCAGGATAGCAAGTCATTTAGAGT
CAACATCAGATCATCGGCCACTCCTCACCACCATGCCATGGAGCCAGAGATTCACAGAGGCAGCTCAGAAACTGAGATTTGATACATTA
GACCACCCTCGCTTCCTCTCACTACTCAGTTCCCACCTTGCTGTCATTGAATGCTCAGCTACAACAGAAGAGGGCCTGGACAGTCTAGCT
CATGGGTTAACCTTAGCAACTGCTAGTGCGTATAAAGGCTCTGCTAGGAGCTCCTTGGCGCAGGGAATAGGTCAGCCATGGTGGAATATT
GACTGCAGAAAAGCGTTGCAAGACTTCCGCTTAGGTCTCTGTTCAAGAAACGACTTCCGTCGGATAACTAGACGGTCTAAATAGCAGTTC
TGGCGAGATAAACTTACCGCAGTGACACAGATCAAAGATGTCTTTGACATAAGCAAGTGACATAAGTTTACAGGATCTTATCGAAACCCT
CCACTAAACGACCCTTTAAGGCCAAACAGCCCTCCAGCAGGGGCTCTGAATGAGAAACAAGACGTATTAGTCCGTAATCTTCTTCAGAAT
ACTGCTGAAGCGGGTGATATTGTCATAGGCTATGGCCTGGGCTGTGGTTGTCAGCCATGCCCTCAACCATAGAACATTCTAGAAGAACCA
TCGGGAAGAGGTTGGAACCCAGTGGAAGTTTGGGAACATGTATATAAGAAGGAGAGGGAGATGTATCTGCCTATTTCTCTCTCCAAGTCT
GCGATATTCGTTAATACATTATACAGGATTGCCAGTTGAAAACAATACTGCCTACGCCCGTCACAGGTACTGCAGTTTCCAACAAGAATC
AACGCTCGACCCGGCAATTATGGCTCAAGGTTAGACTACGTCCTGTGTAGCCTTGATATGCAAGATTAGTTCTGCGATTTGAATATCTAAG
AGGATCTAATGGTAAGCCCCAAGGCTGCCATGGCTTTATTGTAGATTGATTTTCTAGCTGACAATATGCAATTTGGGACAGGGATCTGATG
ATTGTCCGGTTTATGCTGTCTTCAAAAATGTTATACGCCTCGGCGAAGAAGAGGTCAACATTAAATGAGCCCTCCTGGGATGTTTAAAGAT
GGCGAGCGTCAGCAGGAATACTCTACTAAATATCTTCTGCCTACATCAGGGCGCTTAATACCAGAATTTAACAAGCGGAGGAGGATCAA
GGACATGTTCTTGCGTAAACCATCAGCCAACGTATAGAGACCGACGACGAACATCCTGACATTGAGATATTTTACCTCTAGTCAGGAAAA
GGGAACAGCACCCGCTATTTTGGAGAGTGCTGCCAGCGTCATAGCTACCTGCCAGCCTGTAGTAGCTGCTGACAGCACTCAAATGAAAG
AAGTTATTCGTAAGAGCTCTCAGAAATATGAGACAGGTTCCCCTGTCTCAGTCCAGTATTTGACATCGGGTTCAGCCCAATCATCAACAC
CCCCCACTGCTGGACAGAGGACTCTAAAGGGGTTCTTCAAACTTAAAAGTGGTCTAGCCAGCCAAATGGCCATAGCCCAGGATCCTGCA
ACAGTGTCTACTATGCCAACGAAACAACCAGCCGCATCCCCTACAAAATCTACCCCAGTTACAGAACCTCCTGCACTGGAAGCATTACTG
ACAGCTCCCGCTGGTGAAGCTTCTCCAGGAGAACAGCCAAATTCCGCGACTCCTACAGCTCCCGCTTCACCCCAAAGCAATGATACTATT
ATCGATCCCATTGTCAGCAAGGAAGATTGGTCAAAGCTCTTCACTAAAAAGCCCATTCCCAAGTGCGAGGGCCACCAGGAACCATGTTT
CAGTCTGACAACTAAGAAGCCTGGCATCAACTGCGGAAGATCGTTCTGGATCTGTTTGAGACCCCTTGGGCCCAGCGGAAACAAGGAAA
AGGGGATACAGTGGCGATTTCCTACATTCATATGGGCCAGCGATTGGAACCCTTCCGCTCCGTAGATTTTCTGTCTGGGGCAACTTCTTTT
TGCGATAGTGTAACGATACCCGGTTTTATACTTAGAAGGCTACGAATGGTATGATGTATCATGGTTTCAATGATAAGACATTTCGTCAAGT

Transposon Protein Binding Site Exon Intron



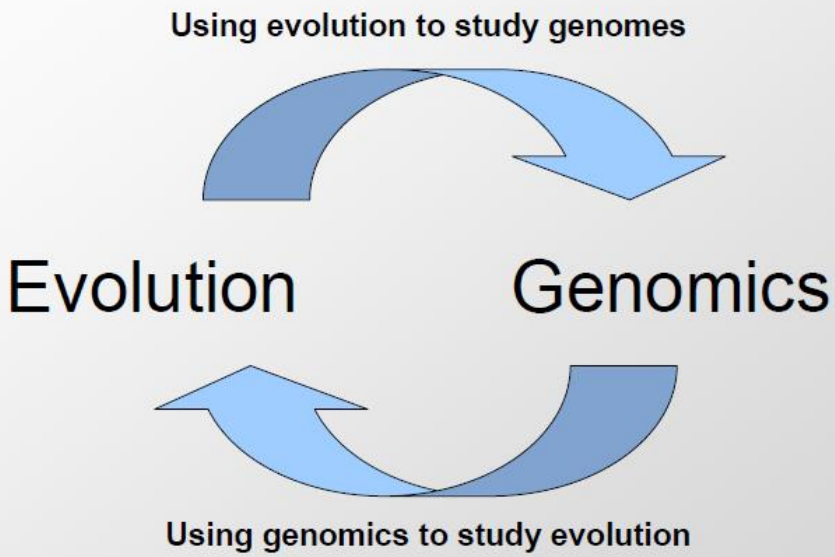
Evolucijska oz. komparativna genomika (Evolutionary and Comparative genomics)

Evolutionary Genomics

- Just as molecular evolution is at the scientific core of molecular biology and genetics, evolutionary genomics is at the core of genomics

- Where did genes and genomes come from? (How and why?)
  - A sequence by itself is nearly useless
  - What restrictions are placed on sequences as they change over time?
    - What potential do they have for non-functional change?
    - What potential do they have for malfunction, and can we predict it?
  - How were new molecular innovations created?
  - How do residues interact to effect functional stasis or change?
  - How did diverse molecules work together to create physiological change?
  - What did ancestral molecules look like and how did they function?

Comparative Genomics

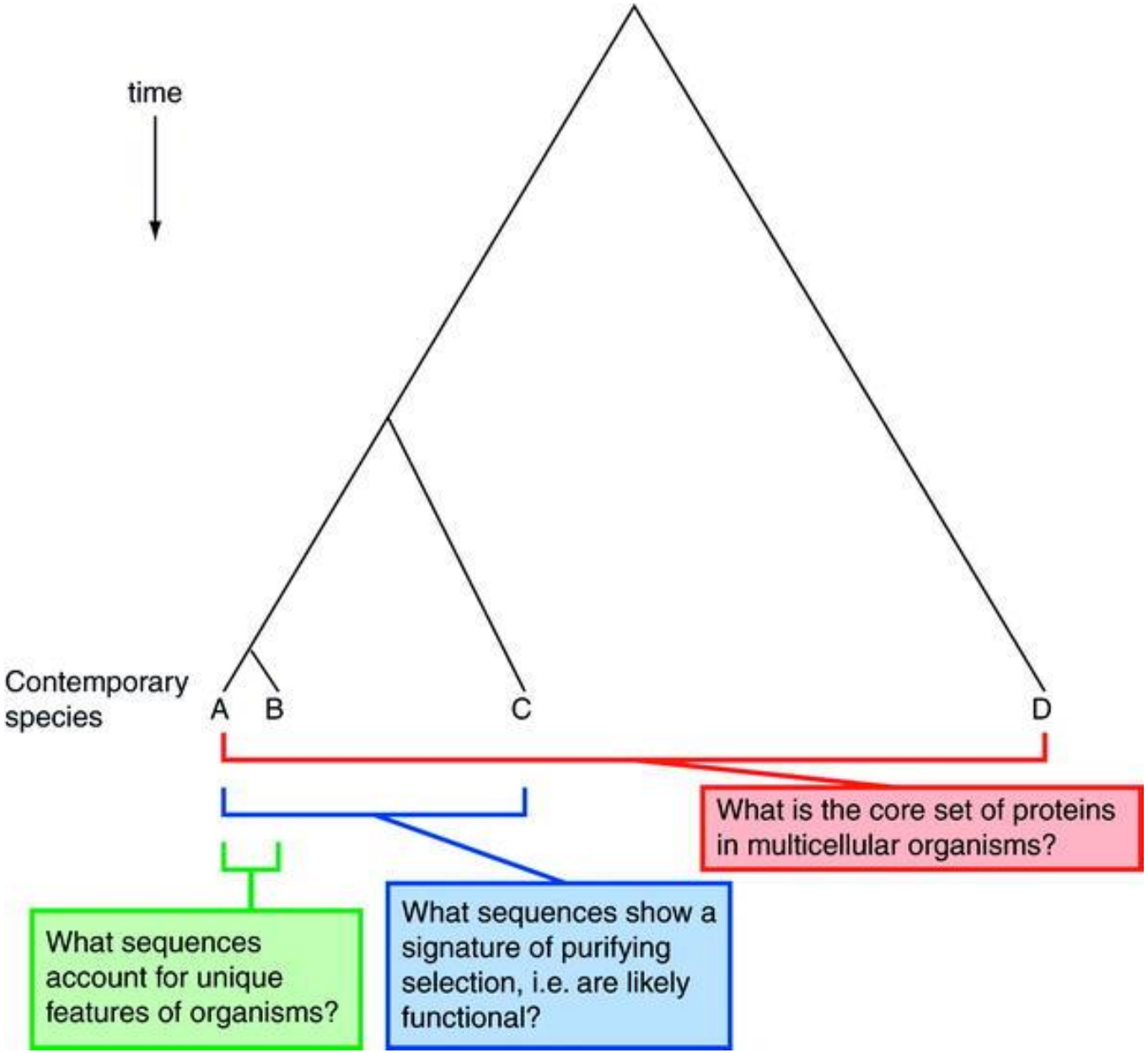


Evolution → Genomes

- **Gene identification**
- **Regulatory motif discovery**
- **Post-transcriptional regulation**

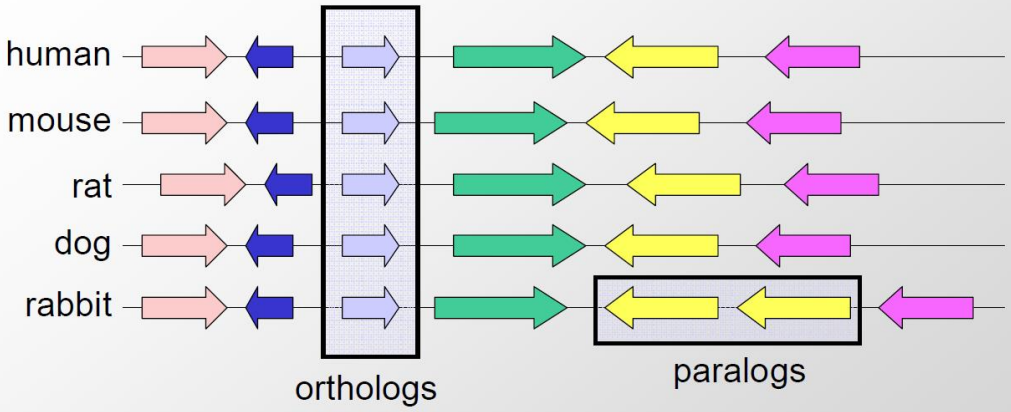
Genomes → Evolution

- **Genome duplication**
- **Network evolution**
- **Motif evolution**



Comparisons of Genomes at Different Phylogenetic Distances Are Appropriate to Address Different Questions

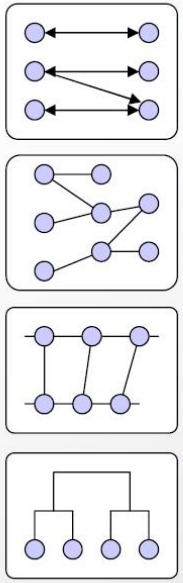
Orthologs and paralogs



- **Orthologs** arise by *speciation*
  - typically keep same function
- **Paralogs** arise by *duplication*
  - typically take on new functions

Ortholog identification a prerequisite to genomic studies

Current methods for ortholog finding



- Pair-wise sequence comparison
  - Best bi-directional BLAST hits
  - Focuses on one-to-one orthologs (no duplications)
- Hit clustering methods
  - Detect clusters in graph of pair-wise hits
  - Difficulty to separate large connected components
- Synteny methods
  - Detect conserved regions, stretches of nearby hits
  - Genome alignment methods focus on best hits
- Phylogenetic methods
  - Phylogeny of family clusters orthologs near each other
  - Traditionally applied to specific families (not genome-wide)

What is comparative genomics

There are many ways that genomes can be compared

- Whole genome
  - Genome size
  - Genome alignments
  - Synteny (gene order conservation)
  - Gene number
  - Anomalous regions

- Gene-centric
  - Gene families and unique genes
  - Gene clustering by function

- Gene sequence variations
  - Codon usage, SNPs, inDels, pseudogenes

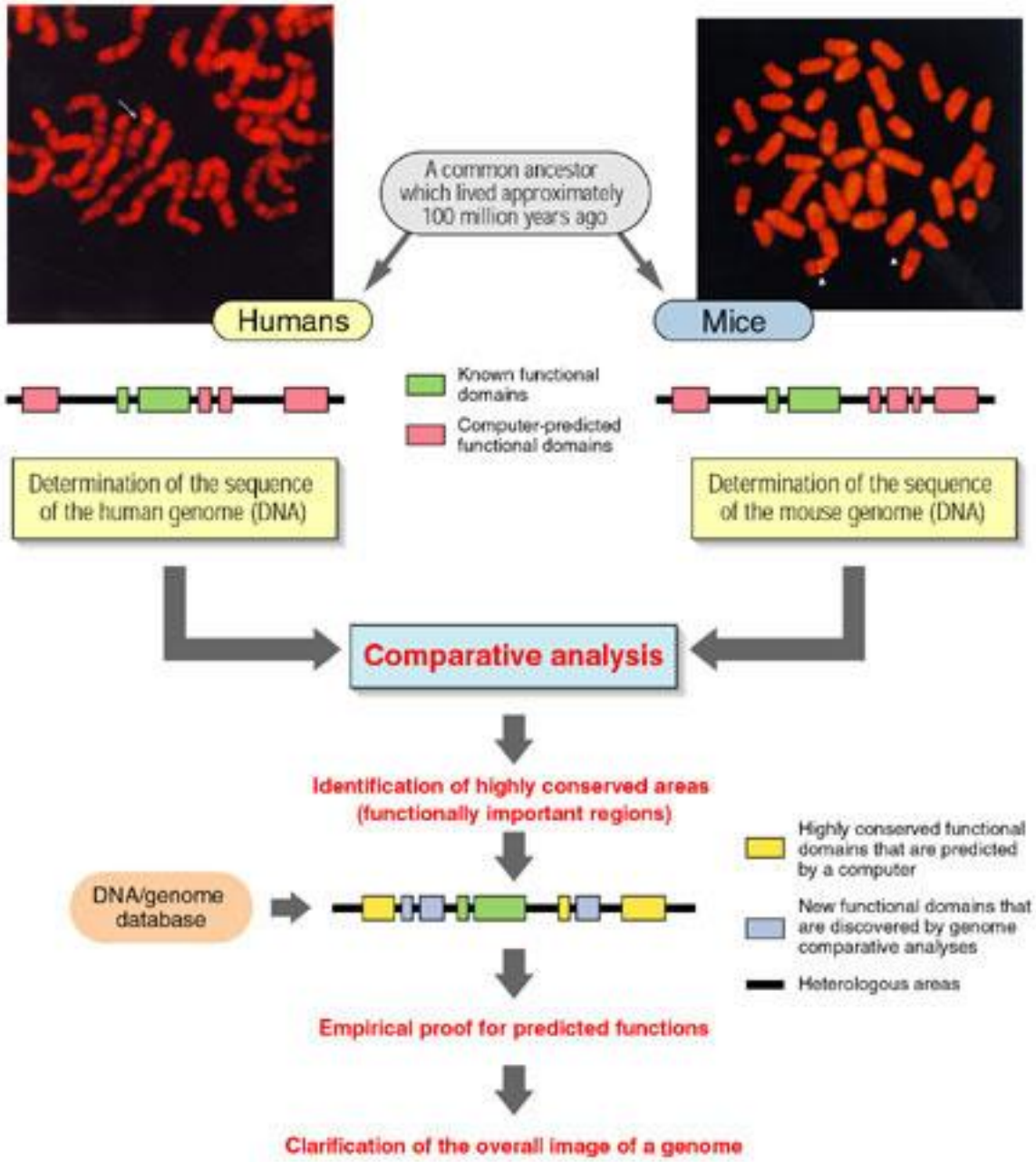
Why Comparative Genomics?

1. Conservation over long evolutionary distances suggests functional constraints
2. Lack of conservation over short distances may be indicative of adaptive evolution
3. Helps us identify both coding and non-coding genes and regulatory elements
4. Characterizing the differences between organisms reveals mechanisms of change
5. Allows us to achieve a greater understanding of vertebrate evolution
6. Leveraging knowledge between species for annotation and inference of function
7. Tells us what is common and what is unique between different species at the genome level
8. The function of human genes and other regions may be revealed by studying their counterparts in simpler model organisms

Comparative genomics

- Discover what lies hidden in genomic sequence by comparing sequence information.
- Main areas
  - Whole genome alignment
  - Gene prediction
  - Regulatory element prediction
  - Phylogenomics
  - Pharmacogenetics

- Affected by evolutionary aspects
  - Mutational forces (introduce random mutations)
  - Selection pressures
  - ⇒ Ratio of non-synonymous to synonymous substitutions
  - ⇒ Mutation rates lower or higher than neutral


A common ancestor which lived approximately 100 million years ago
Humans
Mice
Known functional domains
Computer-predicted functional domains
Determination of the sequence of the human genome (DNA)
Determination of the sequence of the mouse genome (DNA)
Comparative analysis
Identification of highly conserved areas (functionally important regions)
DNA/genome database
Highly conserved functional domains that are predicted by a computer
New functional domains that are discovered by genome comparative analyses
Heterologous areas
Empirical proof for predicted functions
Clarification of the overall image of a genome

Comparing sequences, methods.

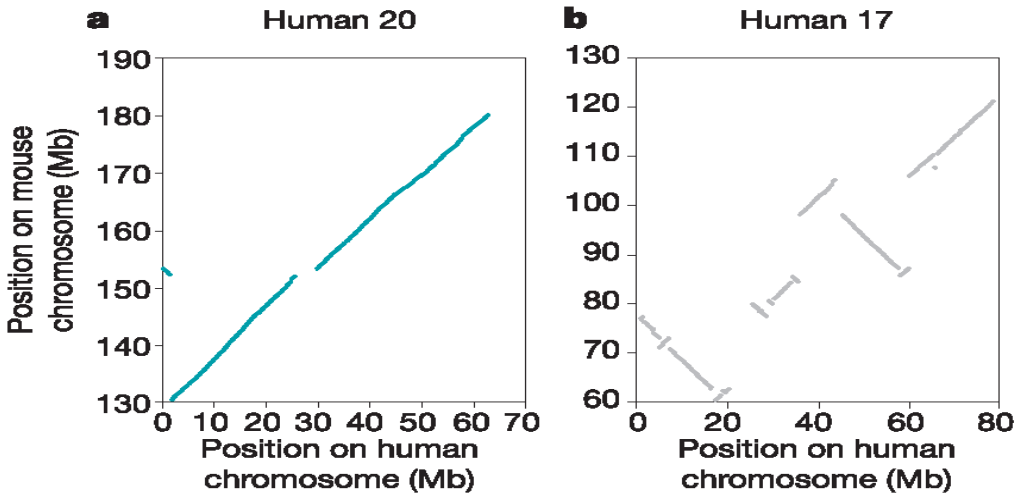
- Pairwise comparison of sequences (alignments)
  - proteins or genes
  - variety of local alignment tools like BLAST, Smith-Waterman etc.
  - multiple sequence comparisons (ClustalW, Muscle etc.)
  - results may be dependent on alignment settings

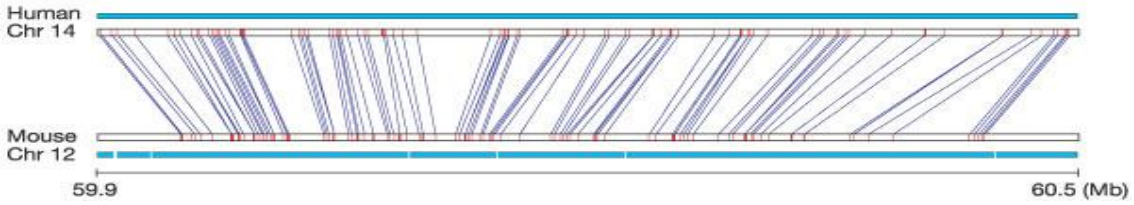
```
>HS.239752 #DB UNIGENE  #DE GNL|UG|HS_S1731711 HOMO SAPIENS NUCLEAR RECEPTOR
           SUBFAMILY 2, GROUP F, MEMBER 6 (NR2F6), MRNA
           /CDS=(132,1346) /GB=NM_005234 /GI=20070198 /UG=HS.239752
           /LEN=1830  #OS HOMO SAPIENS (HUMAN)  #IT UNIGENE
         Length = 1830

 Score =  206 bits (104), Expect = 1e-51
 Identities = 334/408 (81%), Gaps = 2/408 (0%)
 Strand = Plus / Plus

Query: 626  tgggcatcgagaacatctgcgagctggccgcgcgcctgctcttcagcgcccgtcgagtggg 685
            |||||||||| ||| | ||||||||||| ||||| ||||||||||| |||| |||||||
Sbjct: 721  tgggcatcgacaacgtgtgcgagctggcggcgcggctgctcttcagcaccgtggagtggg 780

Query: 686  cccgcaacatccccttcttcccggatctgcagatcaccgaccaggtgtccctgctacgcc 745
            | ||| ||   |||||||||| || |||| | |  |||||||||| | ||||| ||||
Sbjct: 781  cgcgccacgcgcccttcttccccgagctgccggtggccgaccaggtggcgctgctgcgcc 840

Query: 746  tcacctggagcgagctgttcgtgctcaacgcggcccagtgctctatgccgctgcacgtgg 805
            | | ||||||||||| |||||||| ||||||| |||    |  |||| ||||||  ||
Sbjct: 841  tgagctggagcgagctcttcgtgctgaacgcggcgcaggcggcgctgcccctgcacacgg 900
```

```
Sequence: REFSEQ:NM_012387
Sequence: REFSEQ:NM_011059
Sequence: REFSEQ:NM_017230
CLUSTAL W (1.82) multiple sequence alignment

NM_012387       ------------------------------------------------------------
NM_011059       AGAAGGCACGCTCCTACCCGGCAGCCGGGTGTCCCCAAGGAGATCTGGGGAGCCGGGGAG
NM_017230       ------------------------------------------------------------GAG

NM_012387       --------------AGCCA--GAGGGACGAGC-TAGCCCGACGATG---------GCCC-
NM_011059       CAGCTGGGGCTTGTGGCTGCTGGGCACCCAAGCTGACCCGACCAGGAGACTGGAAGCTCG
NM_017230       T-GCTGGGTTTGGTGGCCA--AAGCCCCCATCCTAGGCCGGC-ATG---------TCTCT
                                **       *   * *    *   *** * * *          * *

NM_012387       --AGGGGACATT--------GATCCGTG----TGACCCCAGAGCAGCCCACCCATGCCGT
NM_011059       ACACAGGATGGCCTCCCCGAGAGCTGTGCAGCTGTCTCTGAGAAAGCCGACTCACGCTGT
NM_017230       GCAGAGGAC--T--------G-TACGTG----TATCCCTGGAACACCCAACCAGCGCTGT
                  *  ***            *    ***    *  * *      * ** **    ** **

NM_012387       GTGTGTGCTGGGCACCTTGACTCAGCTTG-ACATCTGCAGCTCTGCCCCTGAGGAC-TGC
NM_011059       GTGCGTGGTAGGTGTGGAGACACTG-GTGAACGTCTATAGTGATGTGCCCAAAGGTGCCA
NM_017230       GTGTGTGGCAGGGGTTGAGAC-CATCGTGGACATTTATGGGTCGGTTCCTGAGGGCACAG
                *** ***   **      *** *    ** ** * *   *    *  **  * *
```

Comparing sequences, methods.

- Whole genome comparisons
  - Large stretches of sequence
  - Divergence up to 450Mya (*fugu*-human) with sufficient similarity remaining.
  - BLAT, BLASTZ, Phusion/BlastN
    - Seeding strategy → alignment extension → gapped alignments

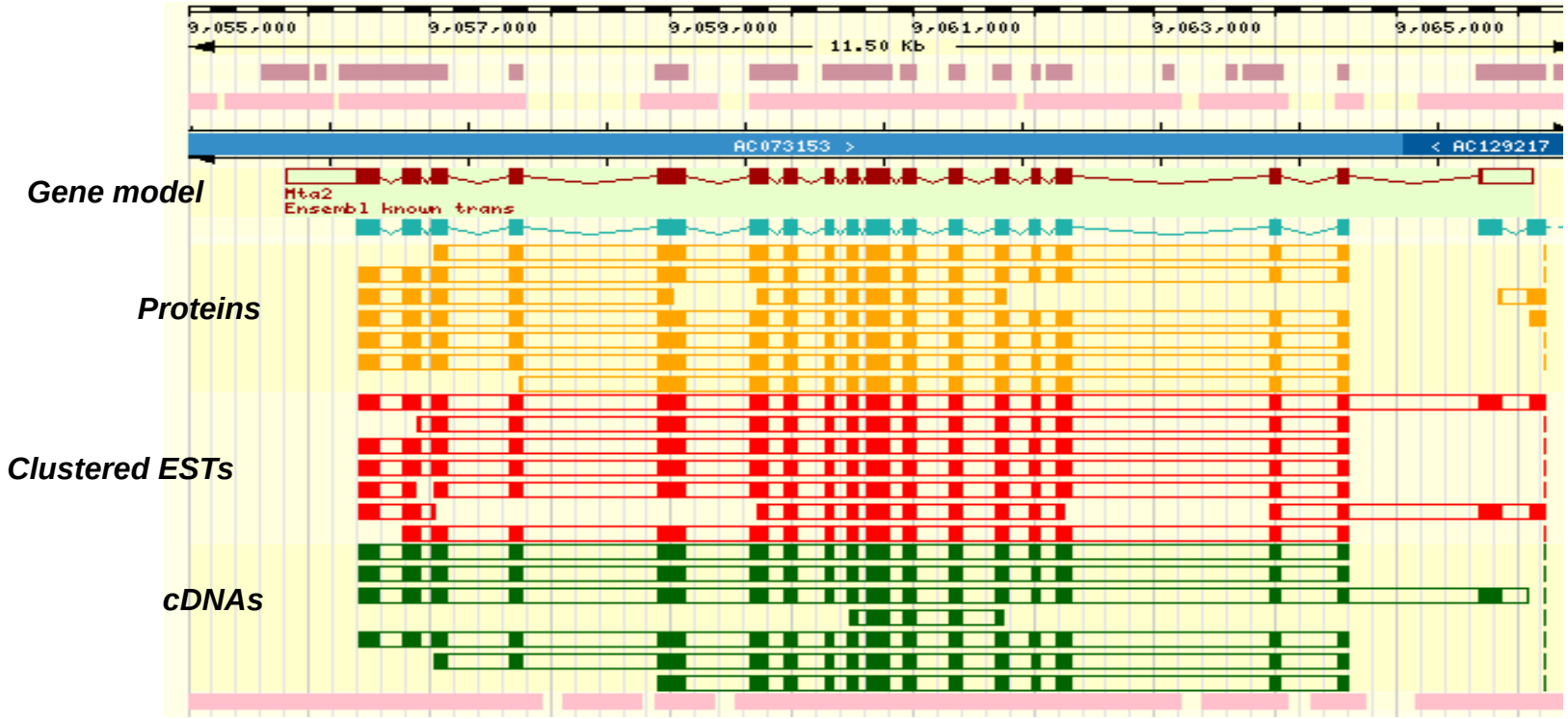




Gene prediction

- Comparing sequences has contributed enormously to the accuracy of gene prediction.

- Evidence based method.
  - Use cDNAs, ESTs and proteins from various organisms.
  - Apply gene feature rules.

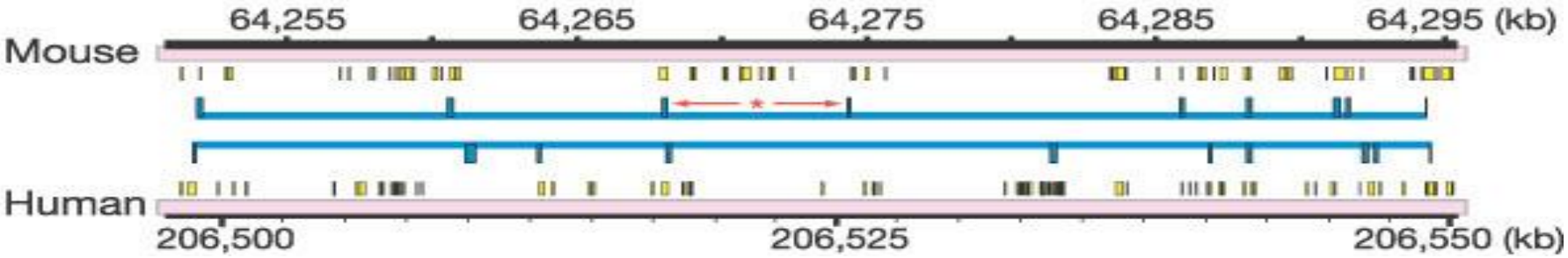


Gene prediction

- *De novo* methods.
  - Alignment of genomic sequences
  - Splicing rules and other gene features

  - *De novo* gene prediction by comparing sequences attempts to model a negative selection of mutations. Areas with less mutations are conserved because the mutations where detrimental for the organism.
  - Prediction of similar proteins in both genomes.

Newly predicted protein in mouse and human, similar to the disease related gene dystrophin.



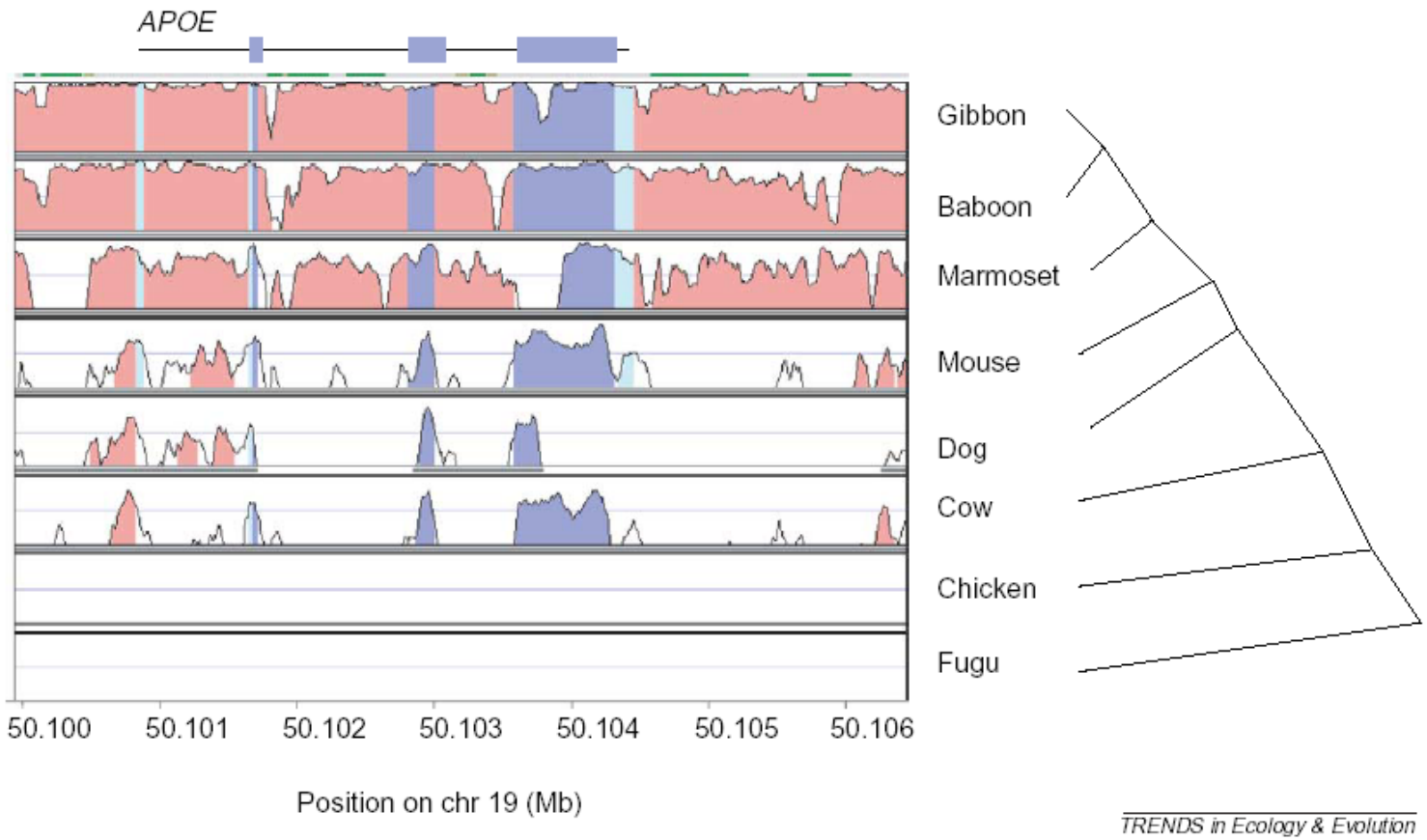
Regulatory element prediction

- The complexity of higher eukaryotes and their relatively low number of genes can be explained partially through the importance of transcriptional regulation.

- Identification of RE’s will have an extensive impact in understanding gene expression patterns (expression intensity, tissue specificity), relations within expression patterns and inferring biological systems or networks.

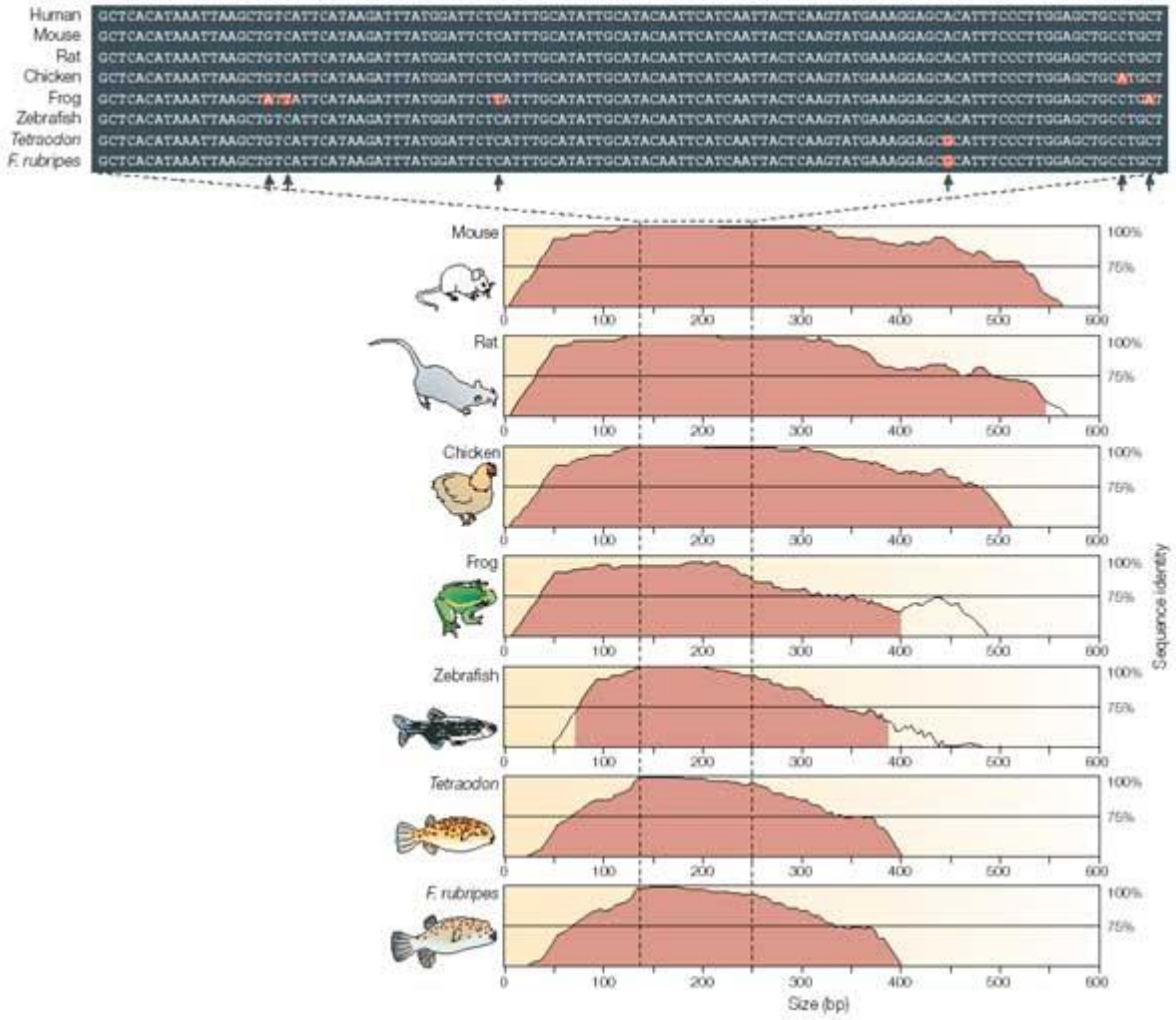
Regulatory element prediction

- No formal models for regulatory motifs
- Attempt to find conserved regions or motifs based on the global alignment of similar sequences of different organisms (*phylogenetic footprinting*).
  - Which species to compare? Evolutionary distance?
  - What regions around gene models to investigate? 5’ and 3’ flanking regions, introns?
  - Take expression patterns into account?
  - How does evolution affect RE’s?



Identifying conserved DNA sequence elements. Alignment of genome sequences and tools for visualizing those alignments have proven remarkably useful in identifying how evolution has resulted in the conservation of genetic elements and the erasure of nonessential sequence. The figure shows a multi-species alignment of a homologous region to chr 19 in humans spanning the gene encoding apolipoprotein E.

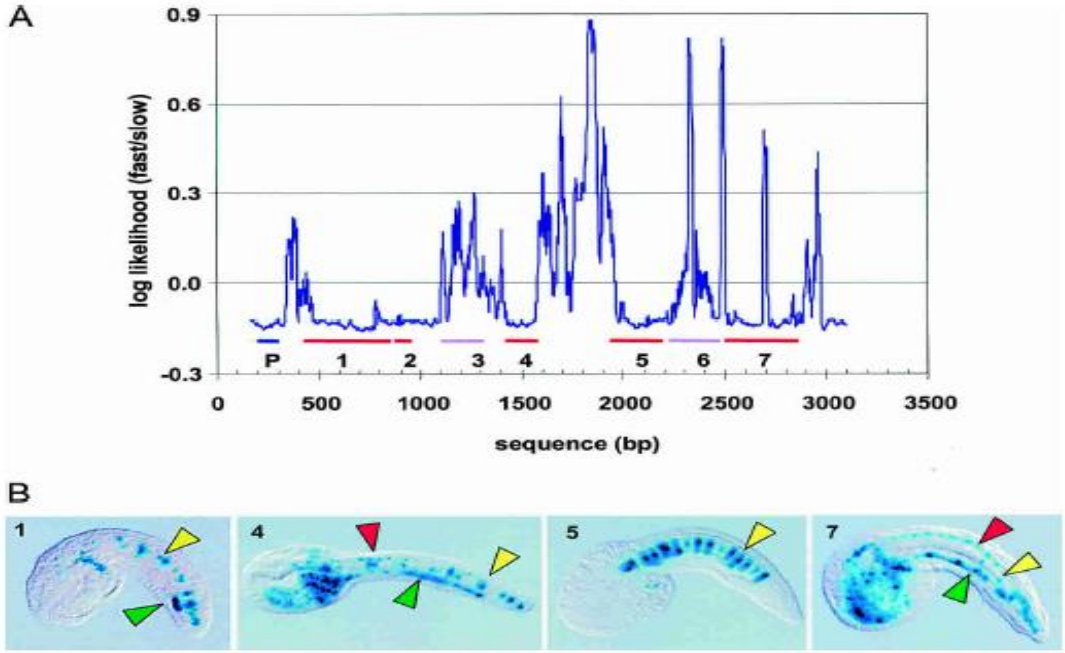
Extreme conservation in enhancers that are shared by human and fish



Comparative genomics at the vertebrate extremes

- Intraspecies sequence comparisons allow identification of species specific sequences
  - Phylogenetic shadowing
  - Requires high rate of polymorphism

- Comparison among primates show human specific sequences
  - Analysis of regulatory sequence of ApoA (involved in human heart disease)



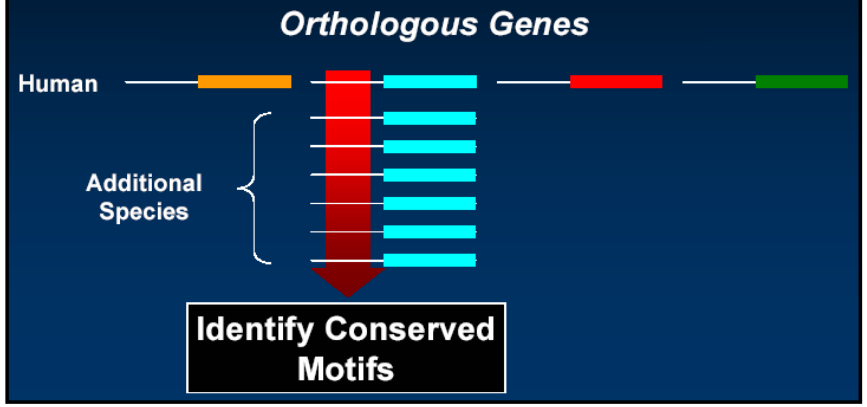
A. Mutation rate analysis of *Ciona intestinalis* 5` region of the *forkhead* gene. B. Validation of identified potential regulatory elements in *Ciona* larvae.

Finding functional regions through genome sequence comparisons

- "Phylogenetic footprinting"
  - Comparing sequences of divergent genomes to find functional regions, eg. Human-mouse
  - Will not reveal elements if background divergence is low
  - Thus cannot reveal novel functional elements in recent lineages, such as great apes
- "Phylogenetic shadowing" (D. Boffelli et al. Science 2003)
  - Using comparisons of close relatives to find regions of conservation
  - Allows detection of elements with varying levels of constraint
  - Eg. Comparing humans to other apes and monkeys

Phylogenetic Footprinting

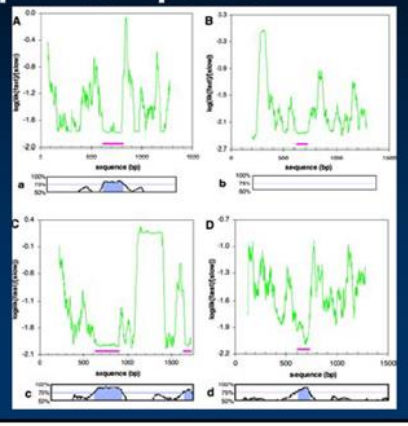
- FootPrinter – http://bio.cs.washington.edu/software.html
- Takes the phylogeny into account



Phylogenetic Shadowing

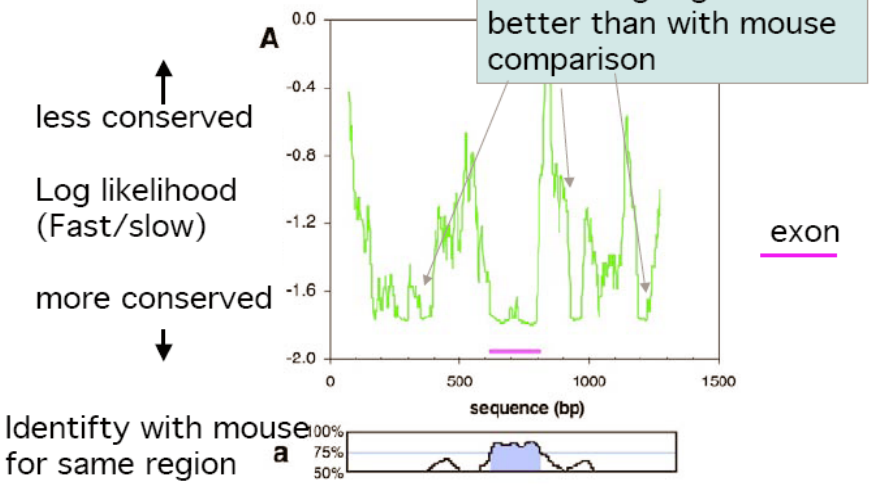
Boffelli et al. (2003) *Science* **299**:1391-1394.

- Identifying sequence *differences* between multiple primate species



Phylogenetic shadowing to find conserved regions

Resolution of conserved non-coding regions much better than with mouse comparison



Boffelli et al. Science 2003

Whole Genome Alignments

- Functional sequences often evolve more slowly than non-functional sequences, therefore sequences that remain conserved *may* perform a biological function.

- Comparing genomic sequences from species at different evolutionary distances allows us to identify:
  - Coding genes
  - Non-coding genes
  - Non-coding regulatory sequences

Genome Browsers

UCSC Genome Bioinformatics

http://genome.ucsc.edu

e! project Ensembl

http://www.ensembl.org

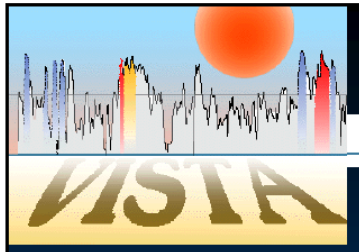
NCBI Map Viewer

http://www.ncbi.nlm.nih.gov/mapview/

Summary of Alignment Tools

- PipMaker (`blastz`)
- VISTA (`avid`)
- Lagan and mLagan (glocal alignments)
  - http://lagan.stanford.edu/
- Box 1 from:

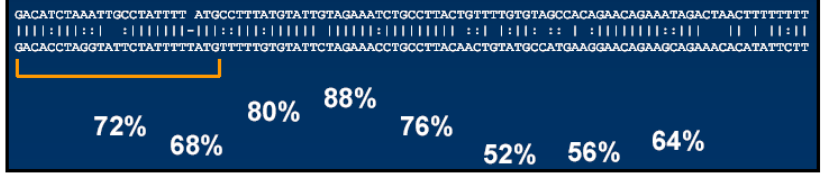
Ureta-Vidal, Ettwiller, and Birney (2003) Comparative Genomics: Genome-Wide Analysis in Metazoan Eukaryotes *Nature Reviews Genetics* **4:** 251-262



http://www-gsd.lbl.gov/vista/

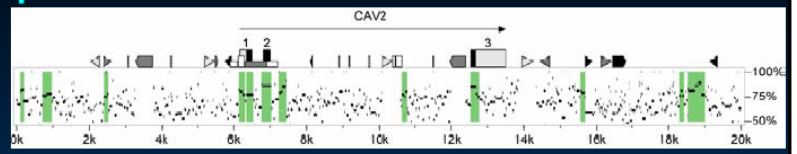
VISUALIZATION TOOLS FOR ALIGNMENTS

- Global Alignment (`avid`)
  - Bray et al. (2003) *Genome Res* **13**:97-102
- Sliding Window Approach to Visualization
  - Plot Percent Identity within a Fixed Window-Size, at Regular Intervals

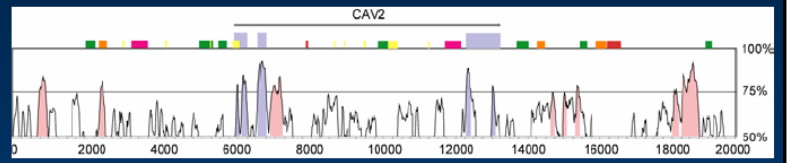


What's Your Preference?

PipMaker



VISTA



Analysis of Comparative Sequence Data

Sequence Conservation → Identification of Functional Elements

- Coding Sequences (i.e. Genes)
  - Relatively EASY to identify
  - Basic understanding of the 'language'
  - Complementary datasets available (ESTs, cDNAs)
- Non-Coding Functional Sequences
  - HARD to identify
  - Very little idea of what to look for
  - Virtually no complementary datasets

Comparative Genomics

- Find sequences that have diverged less than we expect
  These sequences are likely to have a functional role
- Our expectation is related to the time since the last common ancestor

Comparative Genomics

Similarity in Identity

↓

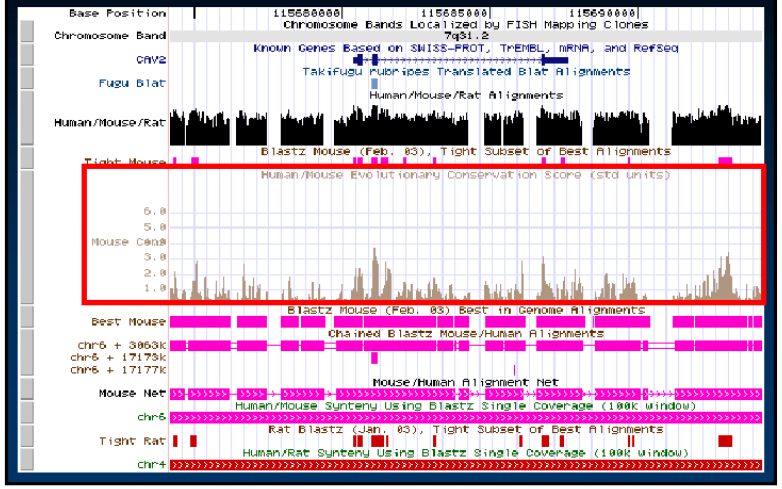
Similarity in Function

My name is Elliott. I am presenting a seminar about comparative genomics.

Mon nom est Elliott. Je vous présente un séminaire à propos de génomique comparative.

Meine name ist Elliott. Ich präsentiere ein Seminar über Comparativ Genomics.

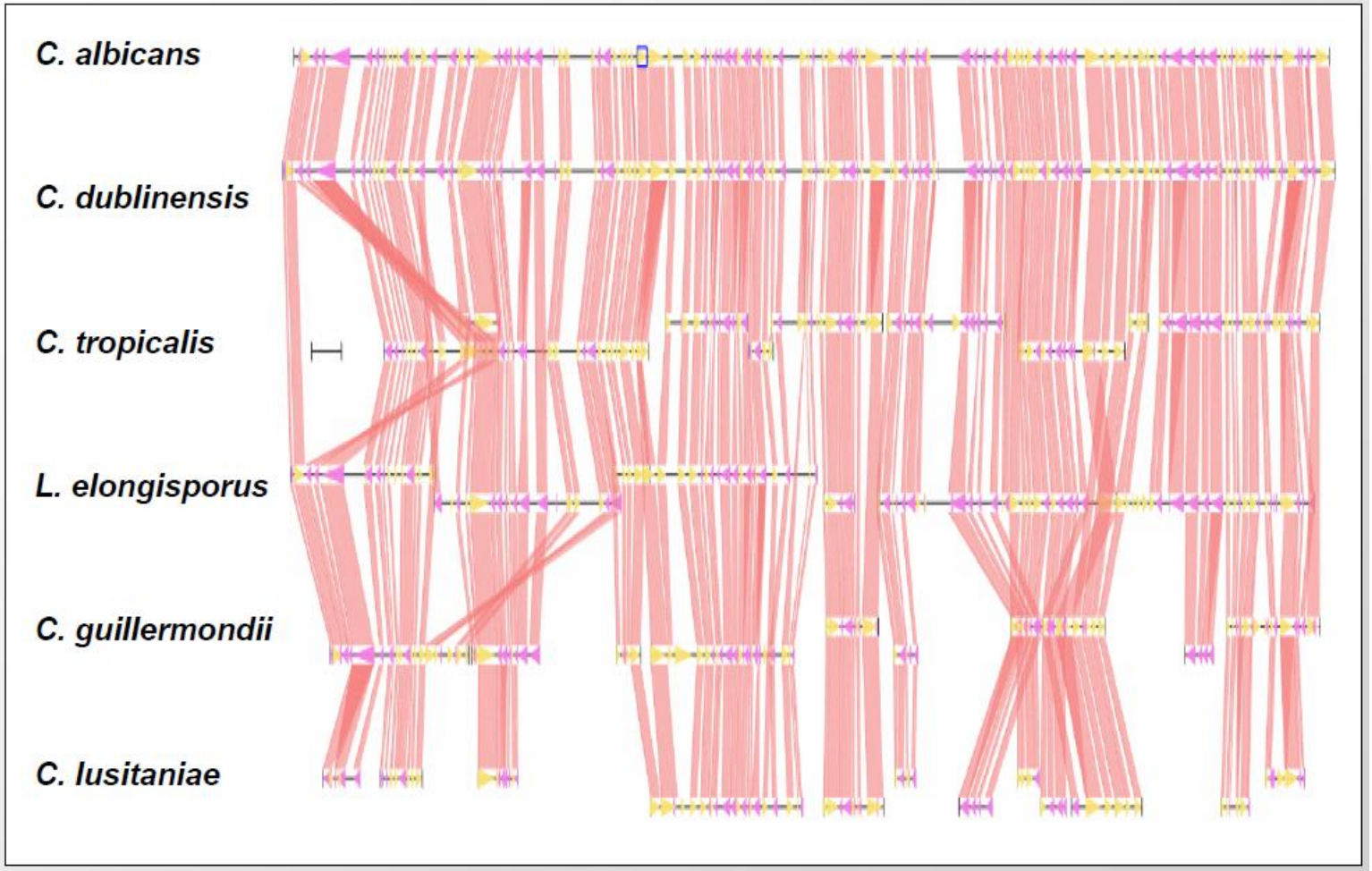
Comparative Sequence Tracks



Whole genome comparison

- Conservation of synteny!
  - Cross-reference of any genetic traits (diseases!) from one organism (eg mouse) to genes in the syntenic regions in the other organism (eg human).
- Genome expansion and contraction
  - Genome duplications, segmental duplications: important mechanism for generating new genes.
- (G+C) content, CpG islands
  - Reflect different mutational or DNA repair processes?
- Repeats
  - Transposable elements are a main force in reshaping genomes. TE's (or remainders thereof) can be used to measure evolutionary forces acting on the genome.
  - Neutral mutation rate.

Extensive conservation of gene order in multiple species



Comparing Genome Size
The ‘C-value paradox’

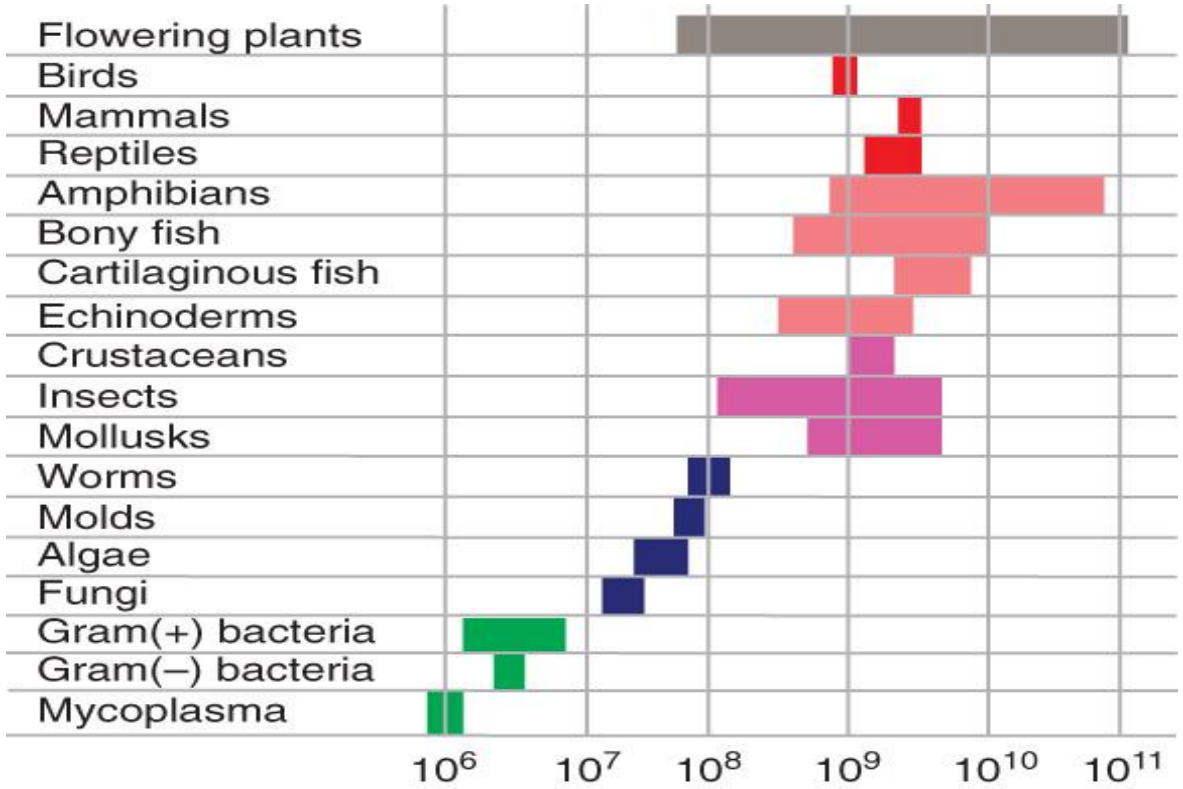
Genome size does *NOT* correlate with organismal complexity

Why Are Some Genomes So Large?

- There is no clear correlation between genome size and genetic complexity.

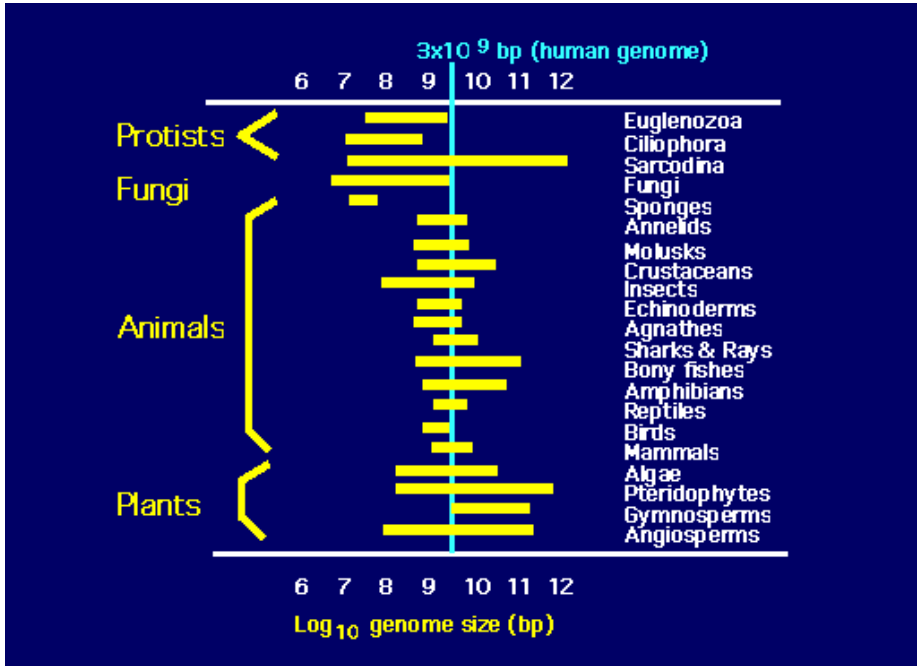
- **C-value –** The total amount of DNA in the genome (per haploid set of chromosomes)

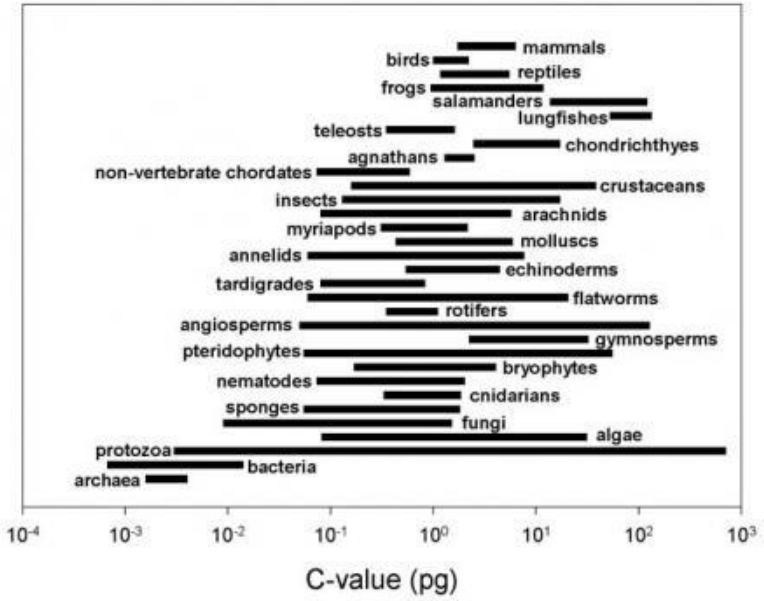
- **C-value paradox –** The lack of relationship between the DNA content (C-value) of an organism and its coding potential.

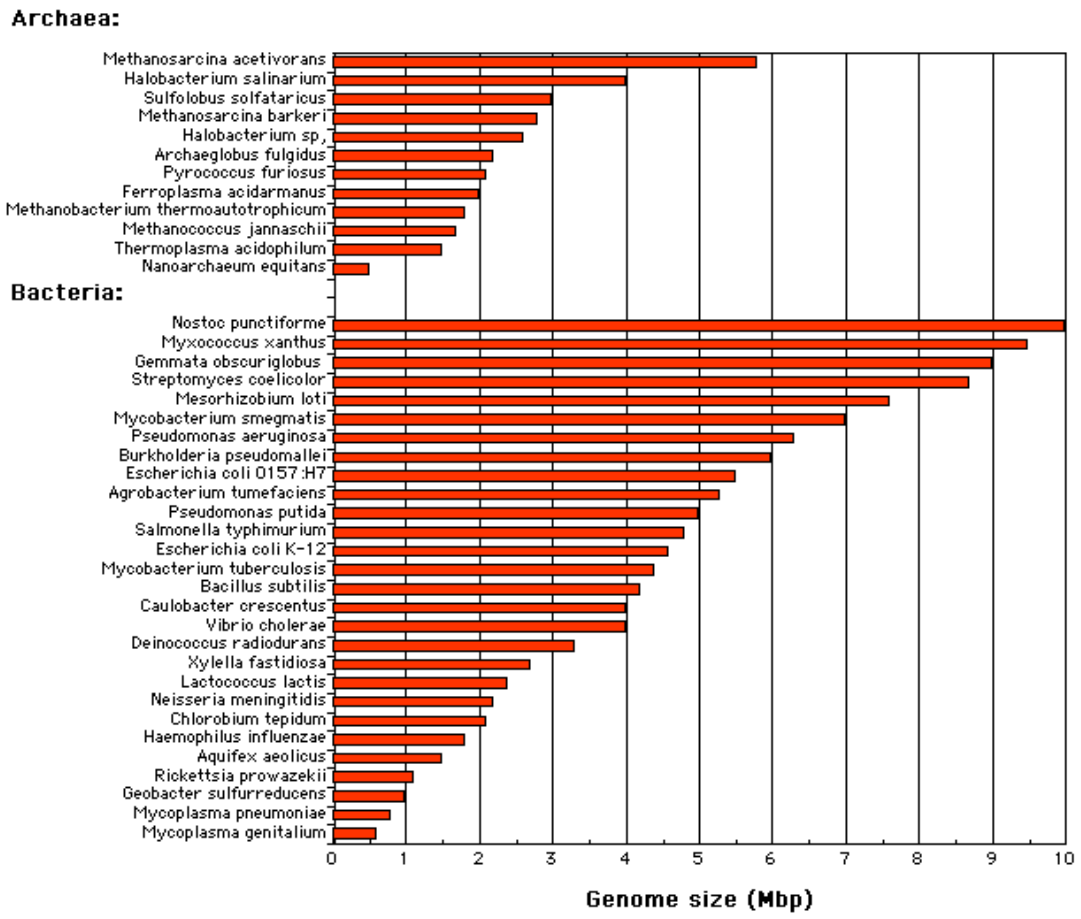


Haploid Genome Size (log scale)

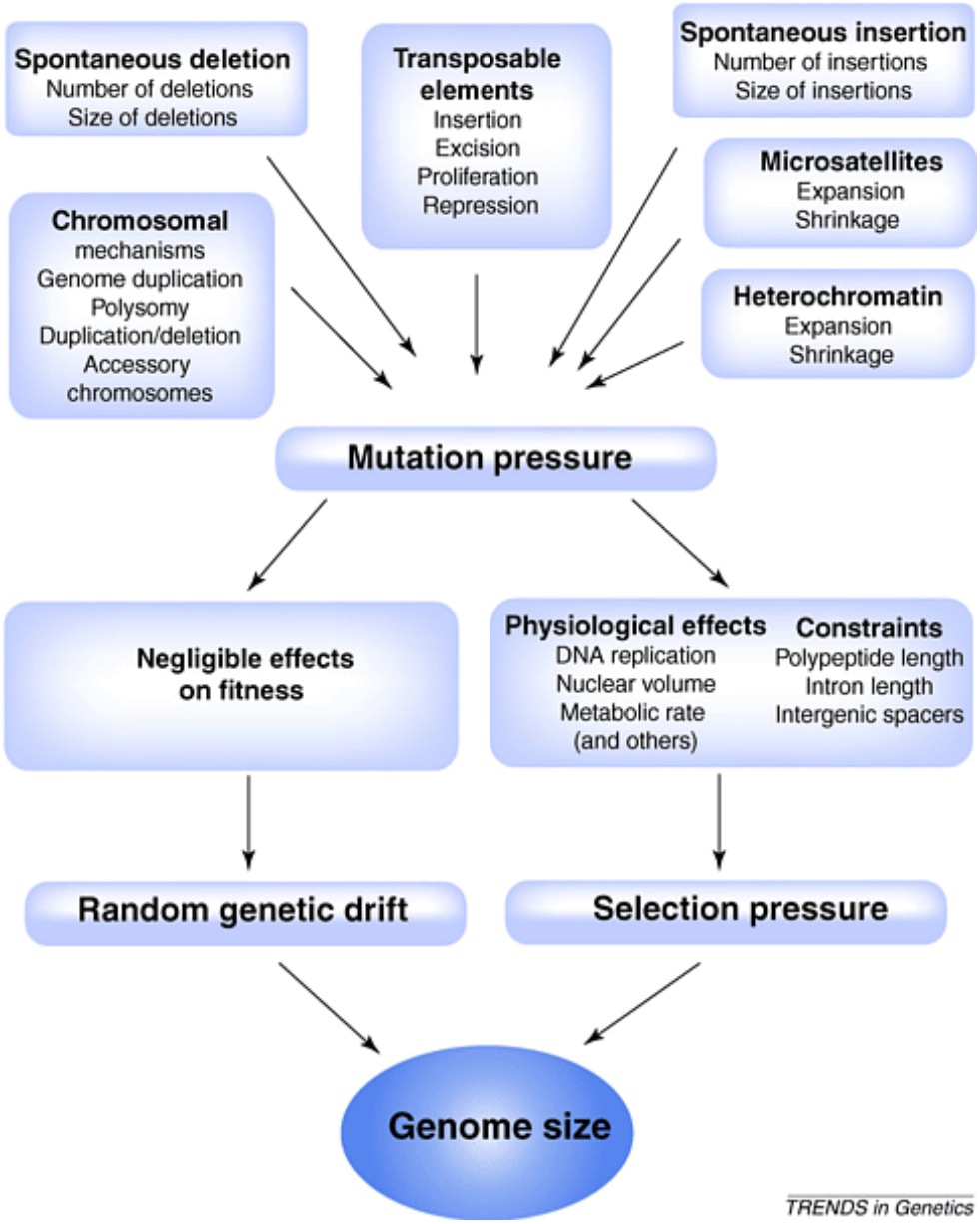
Genome sizes and C-values

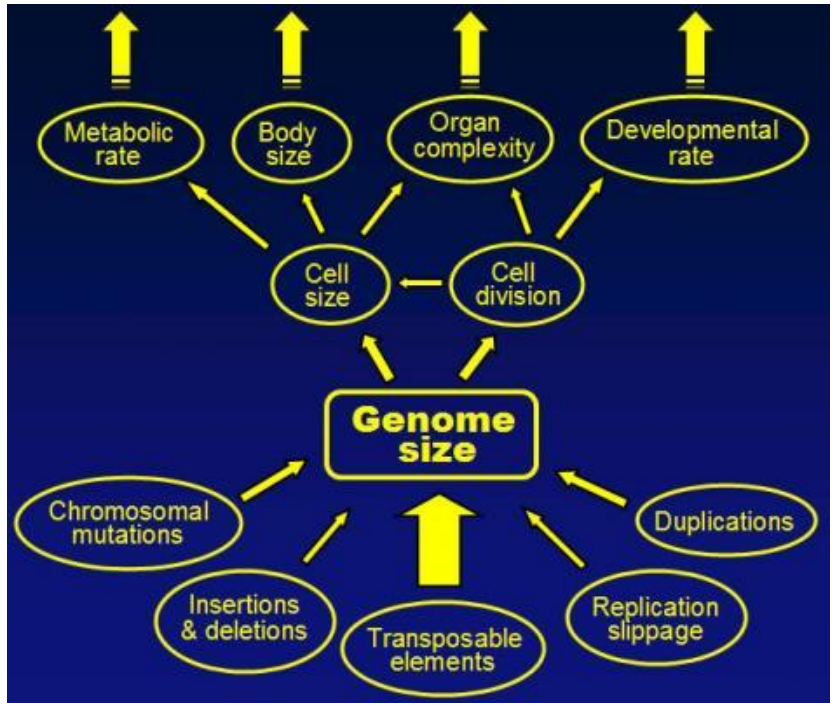


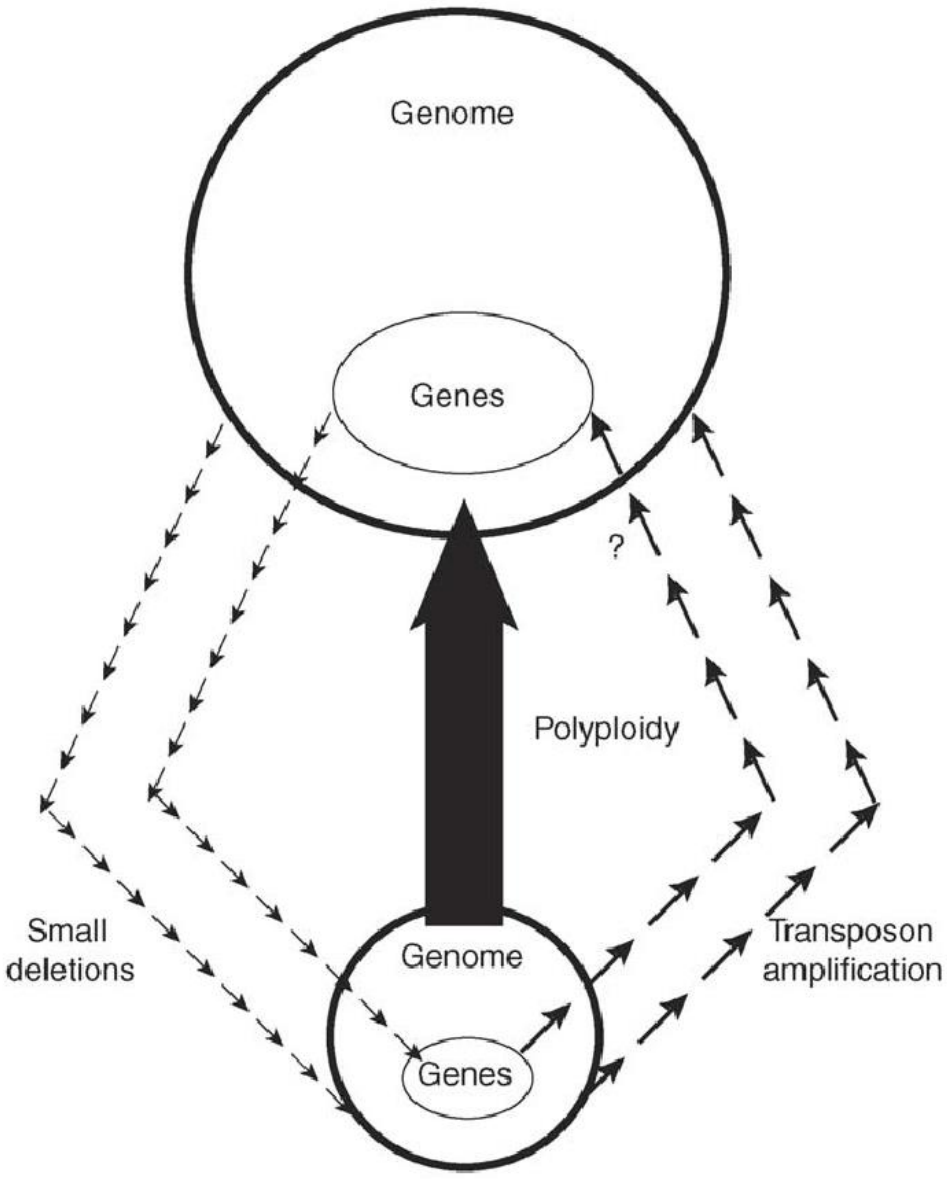




The forces affecting genome size evolution (Petrov, TIG 2001)

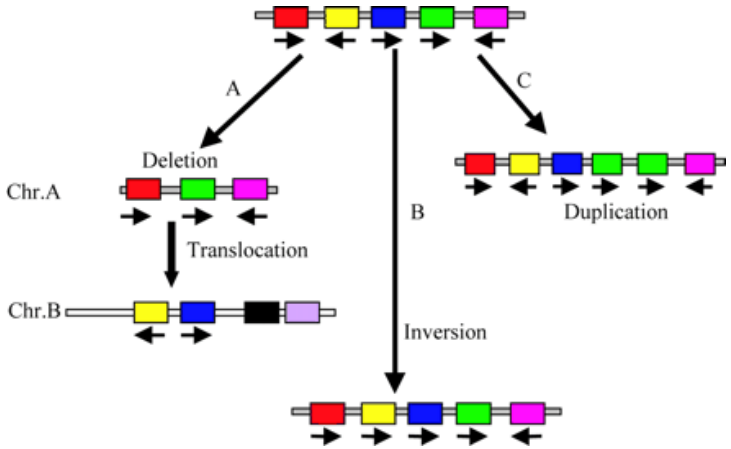




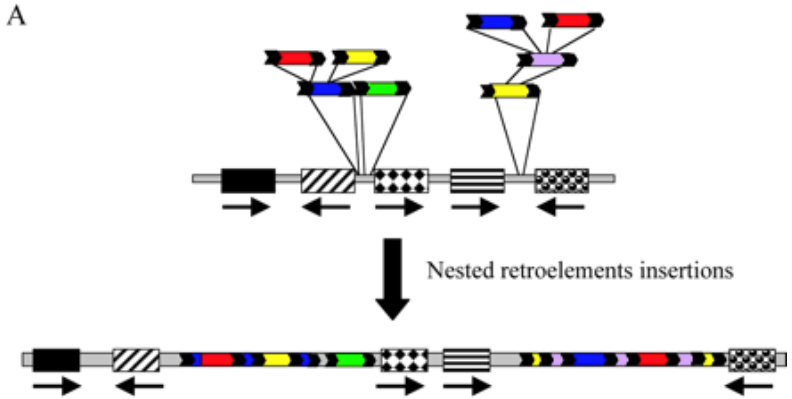


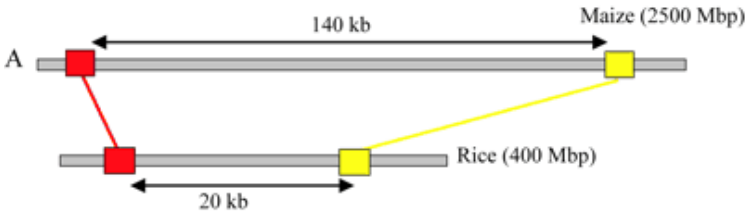
Current Opinion in Genetics & Development

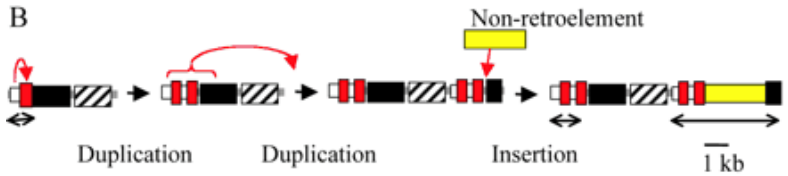
Processes that generate qualitative and quantitative variation in gene number and overall DNA content in plant nuclear genomes. Small arrows indicate minor events, and large arrows indicate large events.

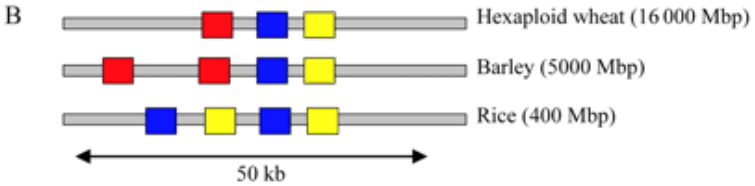


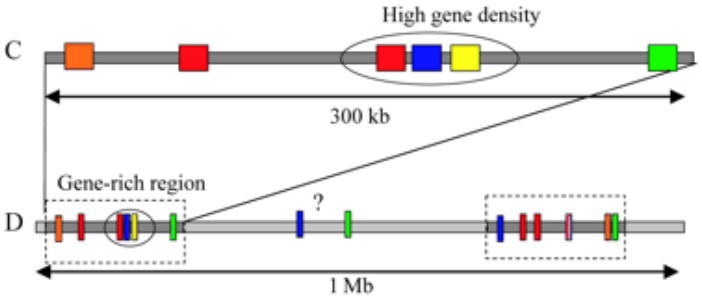
Possible mechanisms of genome expansion in the grass genomes



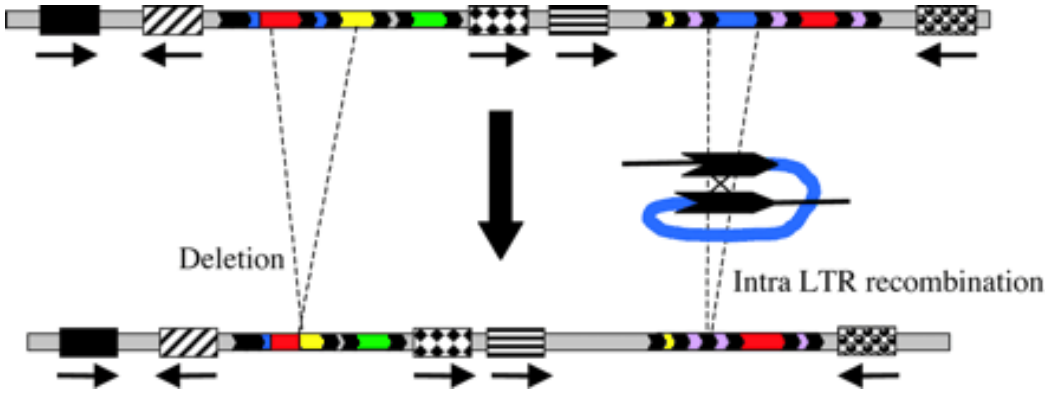








Possible mechanisms leading to genome contraction



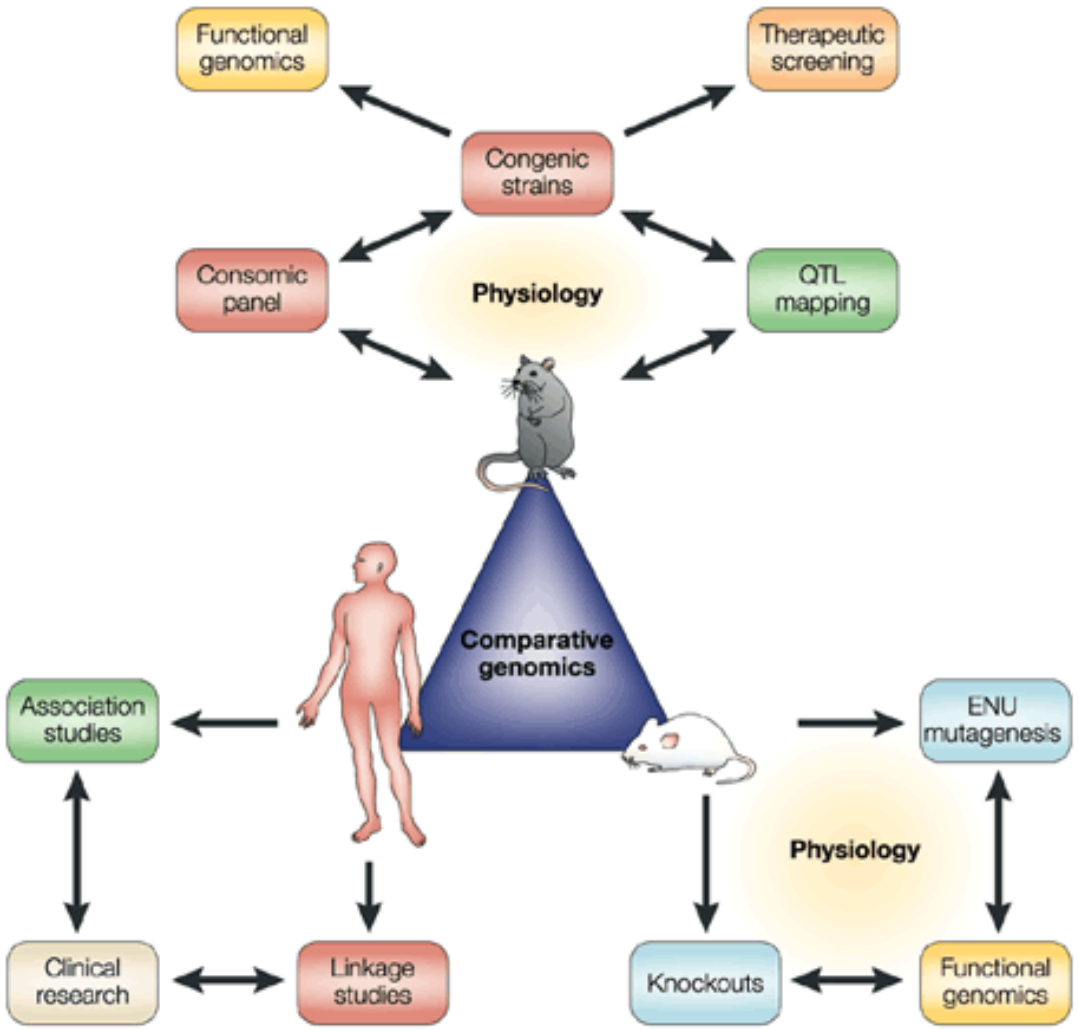
Genome-wide sequence retrieval

- Finding information from whole-genome sequencing projects
  - DNA sequence reads
  - Assembled genomic DNA sequences
  - Annotated genes (RNA genes + protein-encoding genes)
  - Repeats, transposable elements
  - Integrated platform providing both sequence data and functional genomics data

What can we learn from cross-species comparisons?

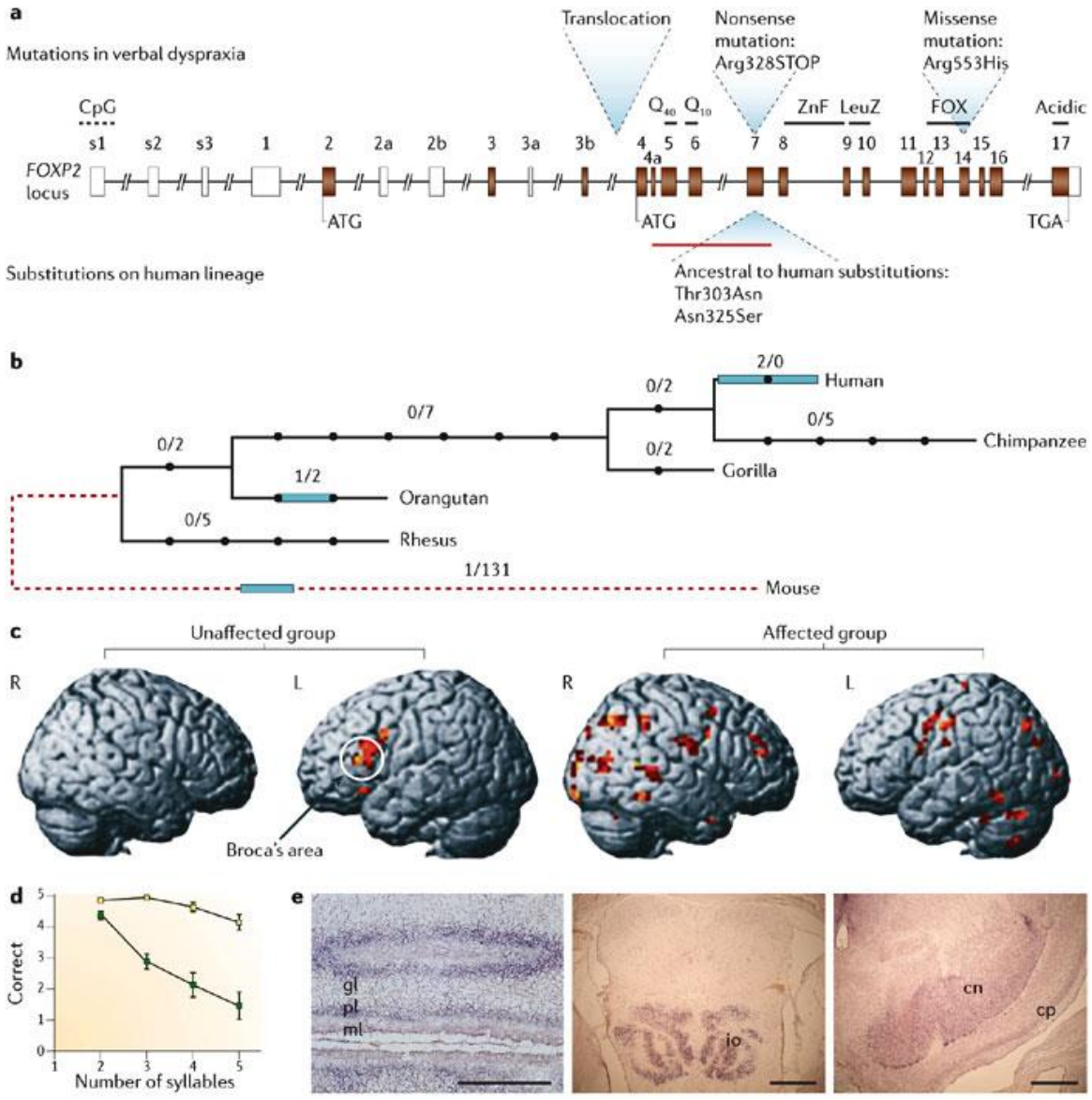
- **Genome conservation**
  - transfer knowledge gained from model organisms to non-model organisms
- **Genome variation**
  - understand how genomes change over time in order to identify evolutionary processes and constraints
- **Detection of functional elements**
  - Coding elements (e.g. exons)
  - Conserved non-coding sequences / elements

Človeške bolezni in komparativna genomika



Nature Reviews | Genetics

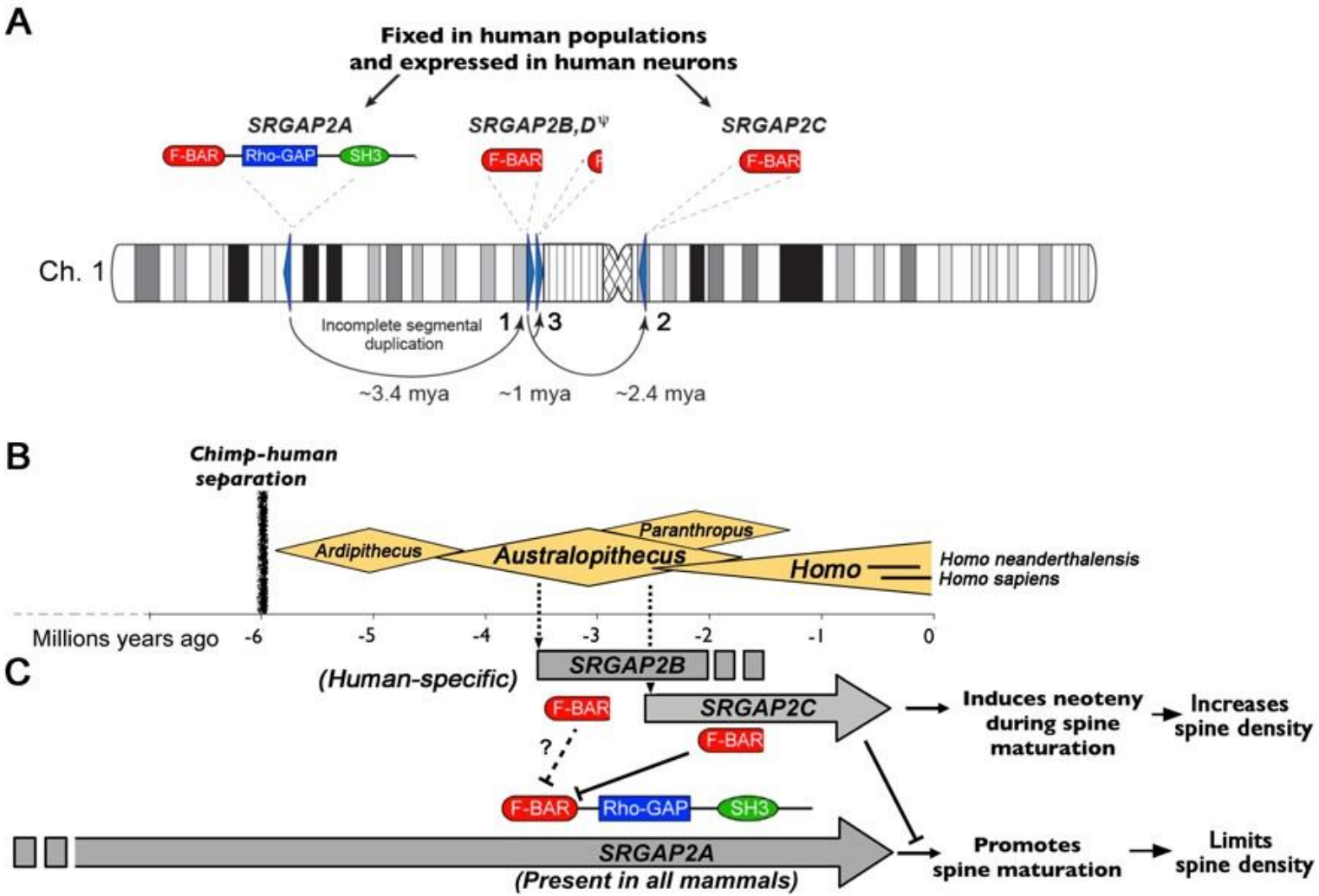
Uporaba komparativne oz. evolucijske genomike



Nature Reviews | Genetics

A multidisciplinary perspective on language evolution

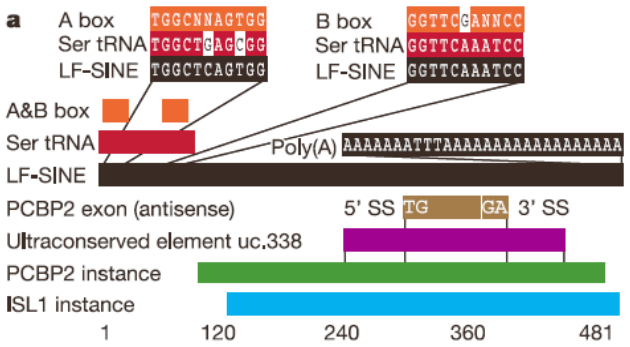
The genomic structure of human **forkhead box P2 (FOXP2)**, showing the location of mutations that cause verbal dyspraxia, which are distinct from sites of evolutionary substitution in the human lineage (filled rectangles, coding exons; white rectangles, non-coding exons). The red bar indicates genomic regions that show evidence of a selective sweep.

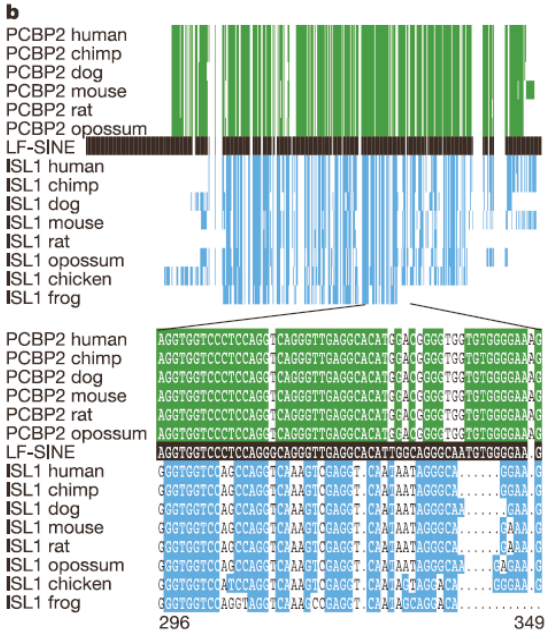


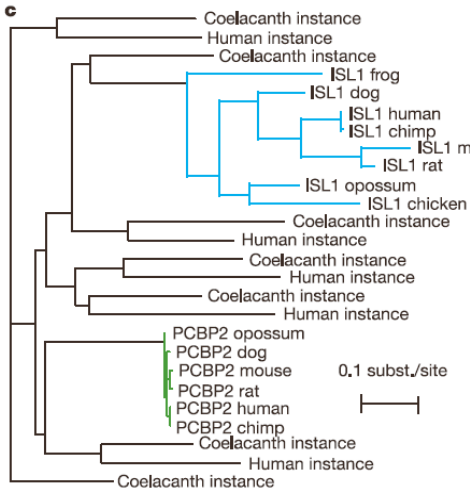
Evolutionary genomics of SRGAP2 and its human-specific paralogs.

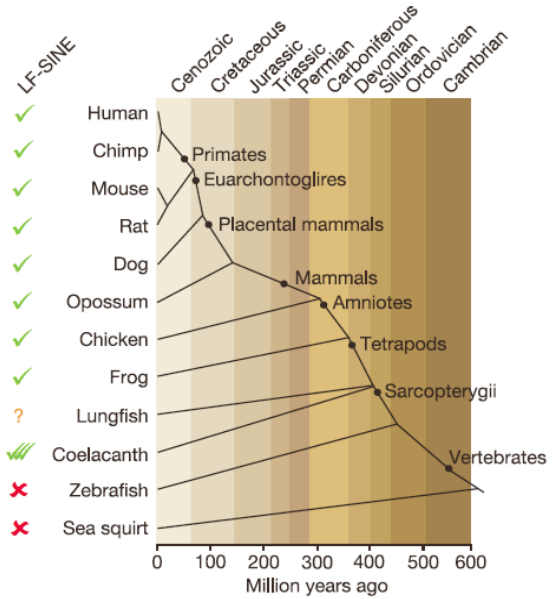
(A) Evolutionary genomic analysis performed by the Eichler's group (Dennis et al. 2012) revealed that SRGAP2 has undergone three human-specific duplications. The first one (SRGAP2B) occurred approx. 3.4 million years ago (mya) and is a partial duplication including only the first nine exons (total of 22 coding exons in the ancestral copy SRGAP2A). The large segmental duplication containing SRGAP2B duplicated approx. 2.4 mya to generate SRGAP2C which also contains the first nine exons. A more recent duplication (SRGAP2D) containing only the first two exons emerged approx. 1 mya. Based on expression analysis as well as copy number variation in the general human population reveals that SRGAP2B and SRGAP2D are most likely pseudogenes. On the other hand, SRGAP2C encodes for a truncated protein (partial F-BAR domain) highly expressed in neurons during human brain development and acts as a strong antagonist of ancestral SRGAP2A, leading to neoteny during spine maturation and increased spine density in vivo.

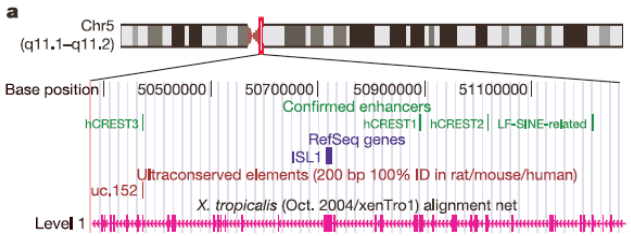
A distal enhancer and an ultraconserved exon are derived from a novel retroposon

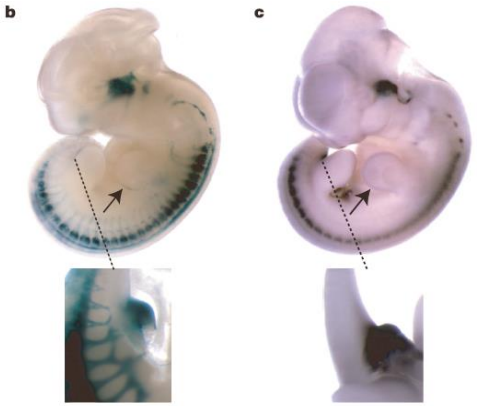


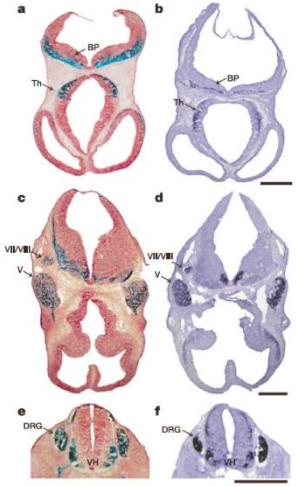












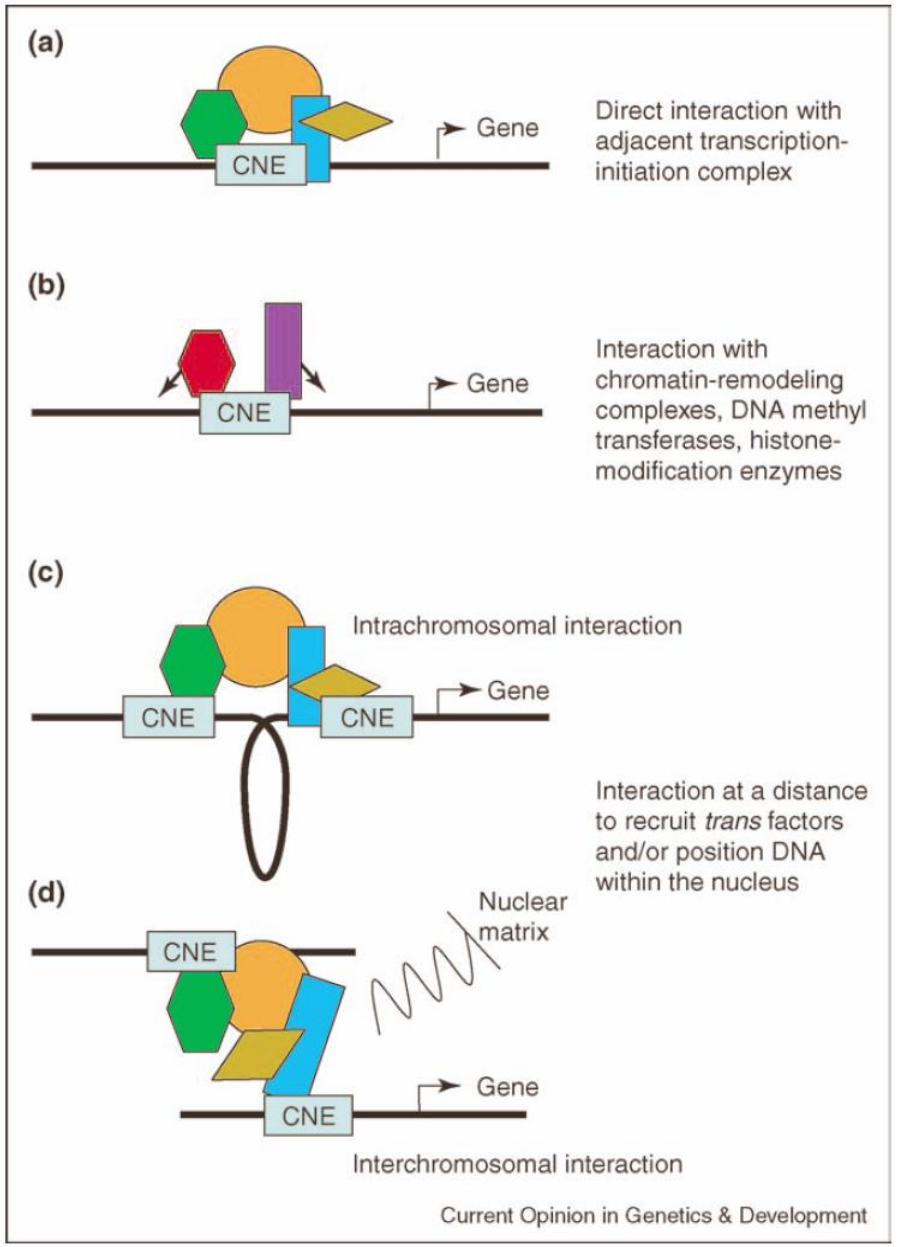
Conserved sequences and the evolution of gene regulatory signals

Potential roles of CNEs in gene regulation.

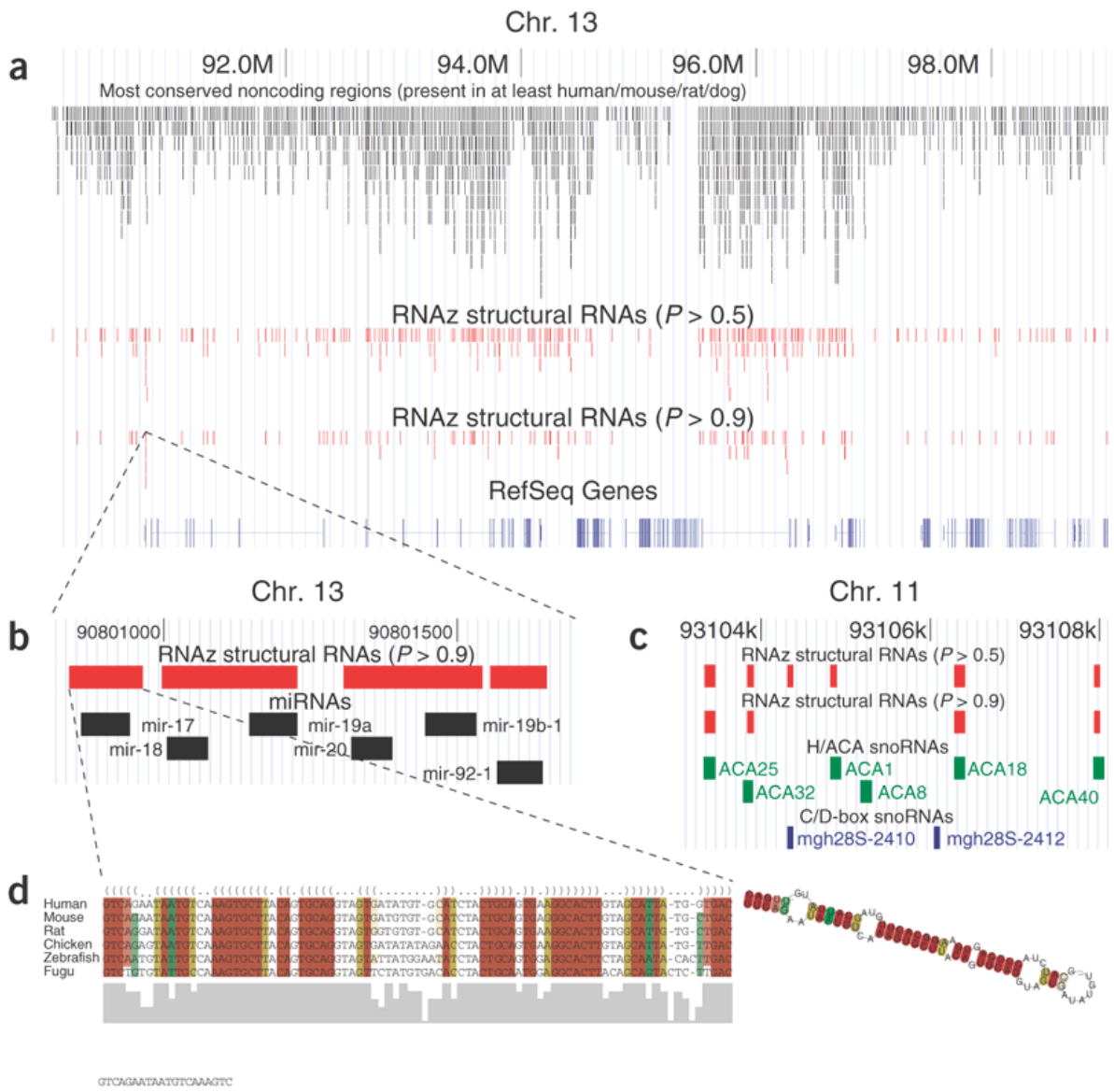
Conserved non-coding sequence elements (CNEs) might have a diverse array of functions related to gene regulation and maintenance of chromosome structure. Some roles for CNEs include (a) acting as cis-regulatory elements; for example, as an enhancer, silencer or insulator to modulate transcription of an adjacent gene; (b) interacting with components of chromatin remodelling complexes, which create 'open' structures that promote active transcription or 'closed' structures that inhibit it; and acting in concert to bridge regulatory elements located at long distances, either (c) on the same chromosome or (d) on another chromosome.

Identifying functional sequences by comparative genomics approaches

The impact of evolution is most readily seen in the pattern of sequence conservation and variation among species. Sequencing of the mouse and rat genomes has revealed remarkable features of evolutionary conservation. About 5% of each genome is under selective constraint, of which only about 1.5% is protein-coding. This conservation extends throughout mammalian species. The importance of gene regulation is highlighted by the fact that two-thirds of the sequence conserved among mammals is not protein-coding. These 'conserved non-coding sequence elements' (CNEs) are largely unique in each genome (i.e. non-repetitive), and, to date, no readily recognizable clusters, classes or subdivisions have been defined that might be useful in further characterizing them.



Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome

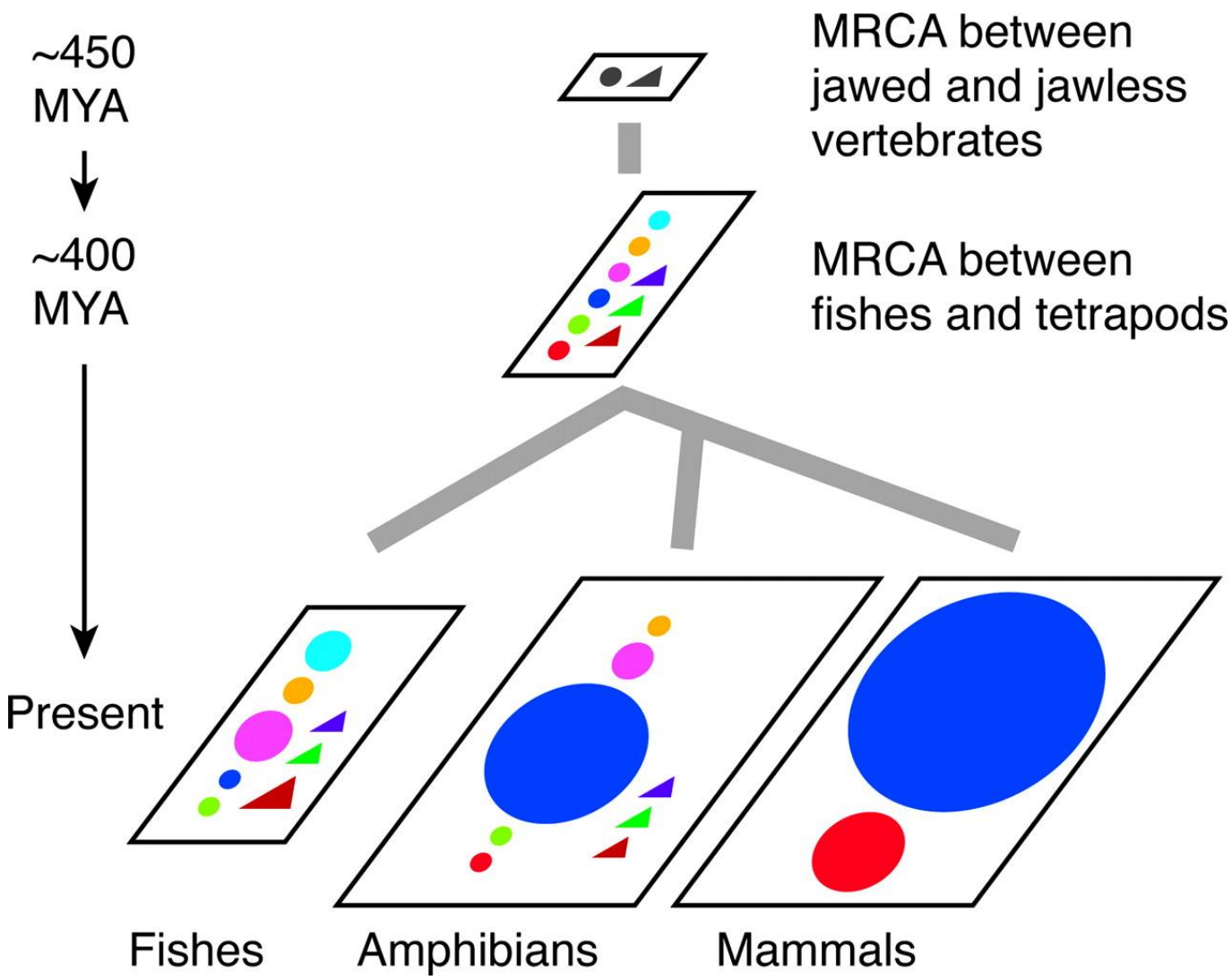


Lineage-specific expansions: Transcription Factor Families Have Much Higher Expansion Rates in Plants than in Animals

After examining the lineage-specific expansion of TF families in two plants, eight animals, and two fungi, we found that TF families shared among these organisms have undergone much more dramatic expansion in plants than in other eukaryotes. The high rate of expansion among plant TF genes and their propensity for parallel expansion suggest frequent adaptive responses to selection pressure common among higher plants.

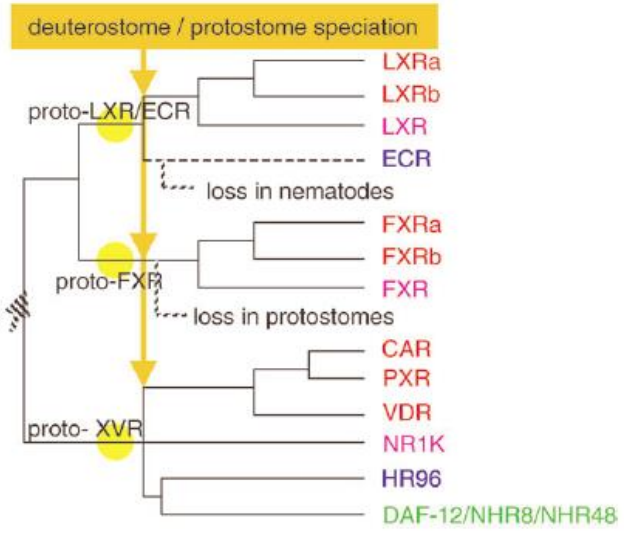
Species	Myb_DNA-binding	zf-C2H2	zf-C3H4	CBFD_NFYB_HMF	AT_hook	PHD	Homeobox	HSF_DNA-bind	bZIP	ARID	SRF-TF	GATA	HLH	CBF	zf-B_box	E2F_TDP	Tub	CXC	AP2	WRKY	SBP	RWP-RK	CG-1	B3	zf-Dof	YABBY	EIN3	DUF573	GRAS	FLO_LFY	NAM	TCP
Saccharomyces cerevisiae	14	46	23	4	6	14	7	5	8	2	4	10	7	2																		
Candida albicans	16	52	27	7	2	15	3	7	7	2	2	9	11	3																		
Schizosaccharomyces	13	22	35	6	5	17	2	2	5	2	3	5	4	2																		
Neurospora crassa	16	75	35	5	16	14	6	2	9	3	2	6	14	2																		
Fusarium graminearum	12	95	30	5	20	15	10	3	10	2	2	6	16	2																		
Magnaporthe grisea	12	71	29	5	18	11	4	3	10	3	1	7	9	2																		
Cryptococcus neoformans	7	37	28	6	6	16	5	4	12	1	2	7	9	2																		
Ustilago maydis	11	33	28	7	12	13	6	4	5	2	2	8	12	2																		
Encephalitozoon cuniculi	9	11	15	5	1	3	10	2	1		1	2			1	2	1															
Homo sapiens	31	641	195	8	17	59	204	5	29	13	5	12	90	2	65	8	4	1					2									
Gallus gallus	29	279	133	6	20	47	126	4	21	10	2	7	70	1	14	8	3	2					1									
Takifugu rubripes	27	362	181	7	17	74	214	2	38	15	8	13	102	2	59	10	5	1					2									
Ciona intestinalis	18	166	69	6	7	36	78		11	6	2	2	39	1	9	5	1	1					1									
Anopheles gambiae	11	322	56	15	6	55	80	1	7	5	2	4	44	1	6	2	1	1					1									
Drosophila melanogaster	17	289	70	6	20	35	93	1	9	6	2	6	49	2	6	3	2	2														
Caenorhabditis briggsae	13	117	46	14	9	17	64	1	7	3	2	5	21	1	6	3	2	1														
Caenorhabditis elegans	15	168	90	6	13	24	83	1	9	4	2	12	42	2	15	4	2	1														
Dictyostelium discoideum	28	27	48	4	6	8	9		15	1	3	11			27	1				1												
Entamoeba hydrolytica	36	21	38	4		1	2	8	2		2	1		1	2																	
Arabidopsis thaliana	280	137	373	27	29	63	87	24	73	9	107	29	147	11	29	8	11	8	145	73	15	14	6	83	36	4	6	19	32	1	101	23
Oryza sativa	286	154	340	31	28	58	96	32	95	7	74	32	174	15	34	9	15	12	167	104	16	9	6	91	34	7	11	7	56	1	133	27
Chlamydomonas	22	10	37	4	13	32	3	2	8	2	1	6	8		1	2	3	1	15	1	11	6										
Thalassiosira pseudonana	35	2	53	6	3	22	5	24	5							3	3	7	9													
Plasmodium falciparum	7	8	22	2	1	5				1					3		1															
Trypanosoma brucei	1	28	18	1											2																	
Giardia lamblia	11	19	33	2		3				1					1	1				1												

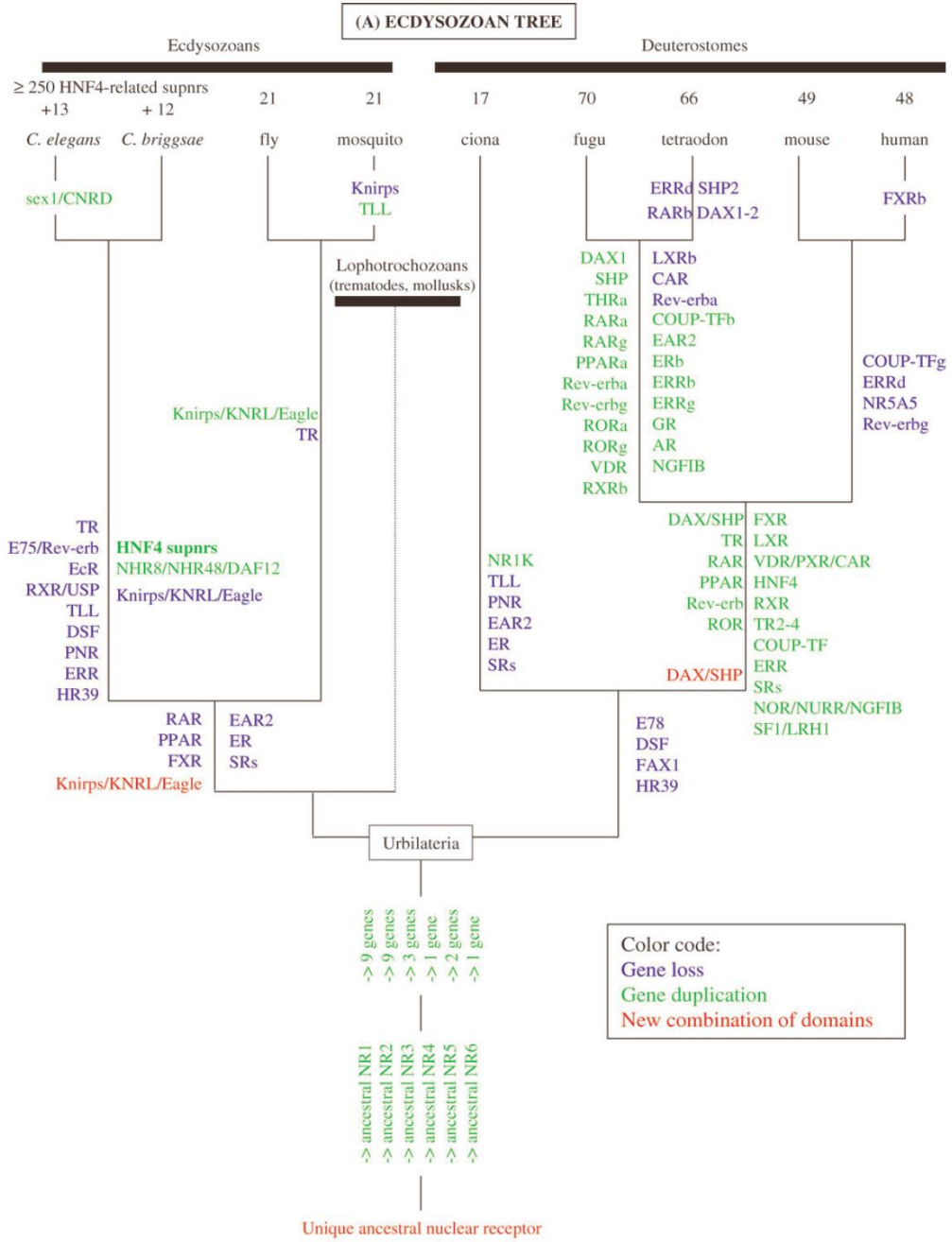
Evolutionary dynamics of vertebrate olfactory receptor (OR) genes



Estimating the tempo and mode of gene family evolution from comparative genomic data

Evolutionary Genomics of Nuclear Receptors: From Twenty-Five Ancestral Genes to Derived Endocrine Systems





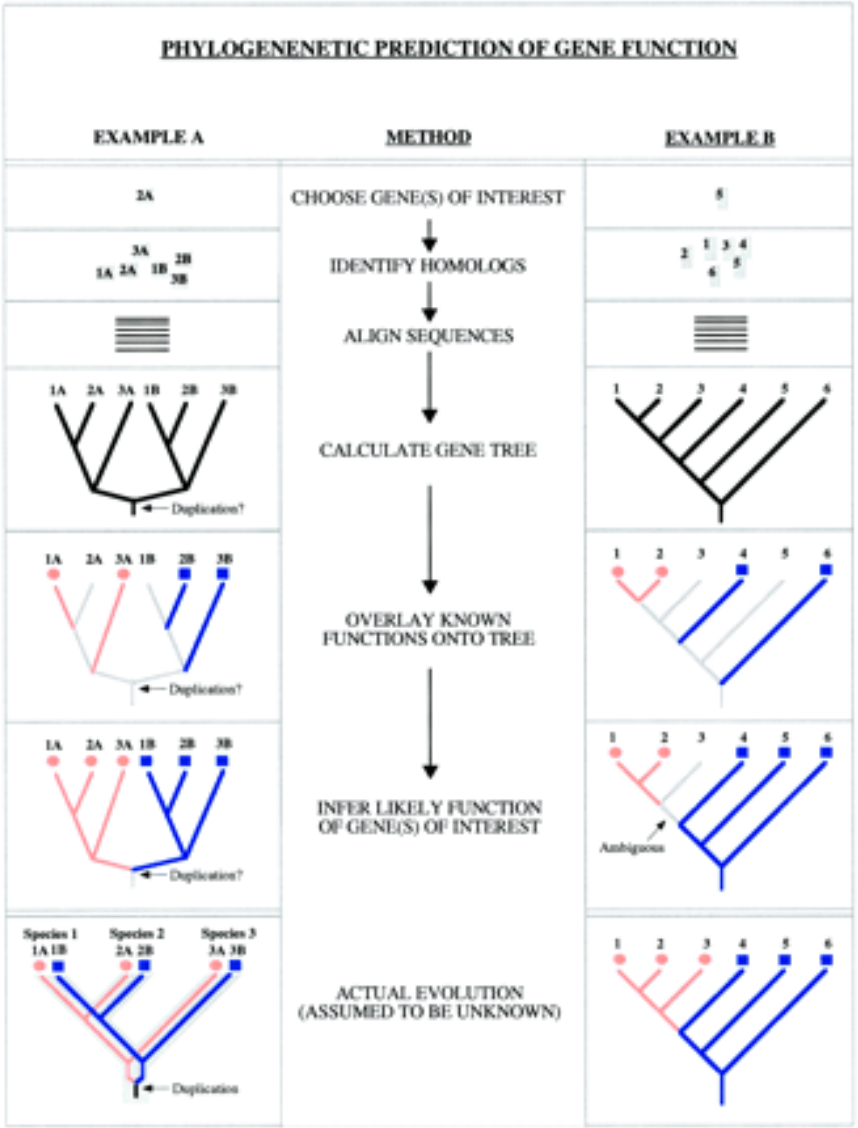
Filogenomika/Phylogenomics

Phylogenomic analysis

- Major step in PA is to infer four main events in the evolutionary history of gene family:
  - 1. Gene origin
  - 2. Gene duplication
  - 3. Horizontal gene transfer (HGT)
  - 4. Gene loss
- Evolutionary distribution pattern (EDP)
  - determined by overlaying gene presence/absence information onto an evolutionary tree of species.
  - EDP reveal a great deal about the evolutionary history of particular genes
- Uneven distribution patterns
  - are difficult (scattered presence and absence throughout the tree)
  - **HGT or gene loss.**
- Gene is present in only one subsection of the species tree:
  - originated in that subsection.

Phylogenomics

(Eisen, Genome Res. 1998)

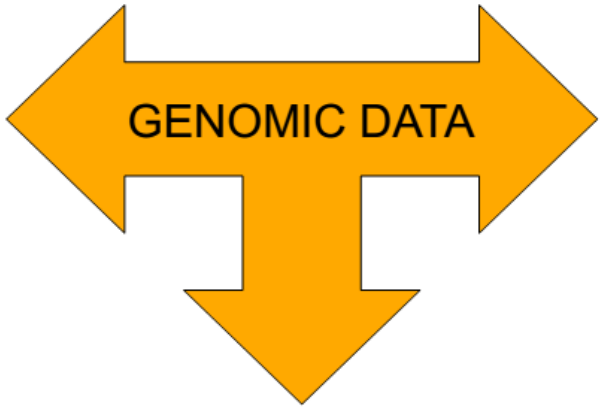


Phylogenomics

- A combination of :
  - genomics (study of function and structure of genes and genomes)
  - molecular phylogenetics (study of evolutionary relationships among organisms)
- Two different aspects :
  - using phylogenetic data to infer functions for DNA and protein sequences
    (**Eisen**. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res. 1998)
  - using genomic data to infer phylogenetic relationships (species trees) and to gain insights into the mechanisms of molecular evolution
    (**O'Brien and Stanyon**. Phylogenomics. Ancestral primate viewed. Nature 1999)

Phylogenomics: three main axes

Reconstruct the evolutionary relationships among organisms



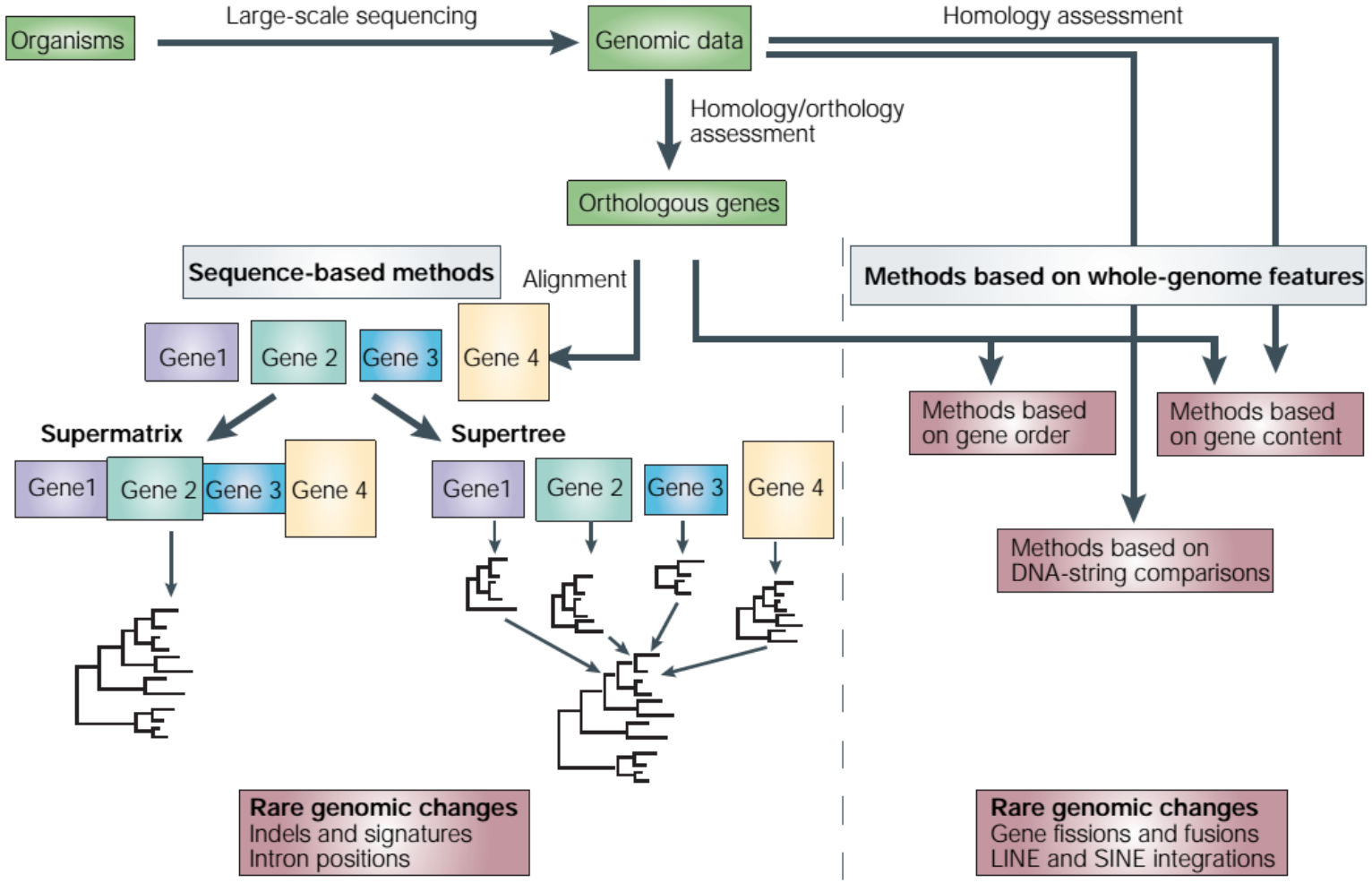
Understand the evolutionary history of genes

- Gene origins
- Gene duplications
- Gene losses
- Horizontal gene transfer
- Functional assignment

Evolutionary history of genomes and biological systems

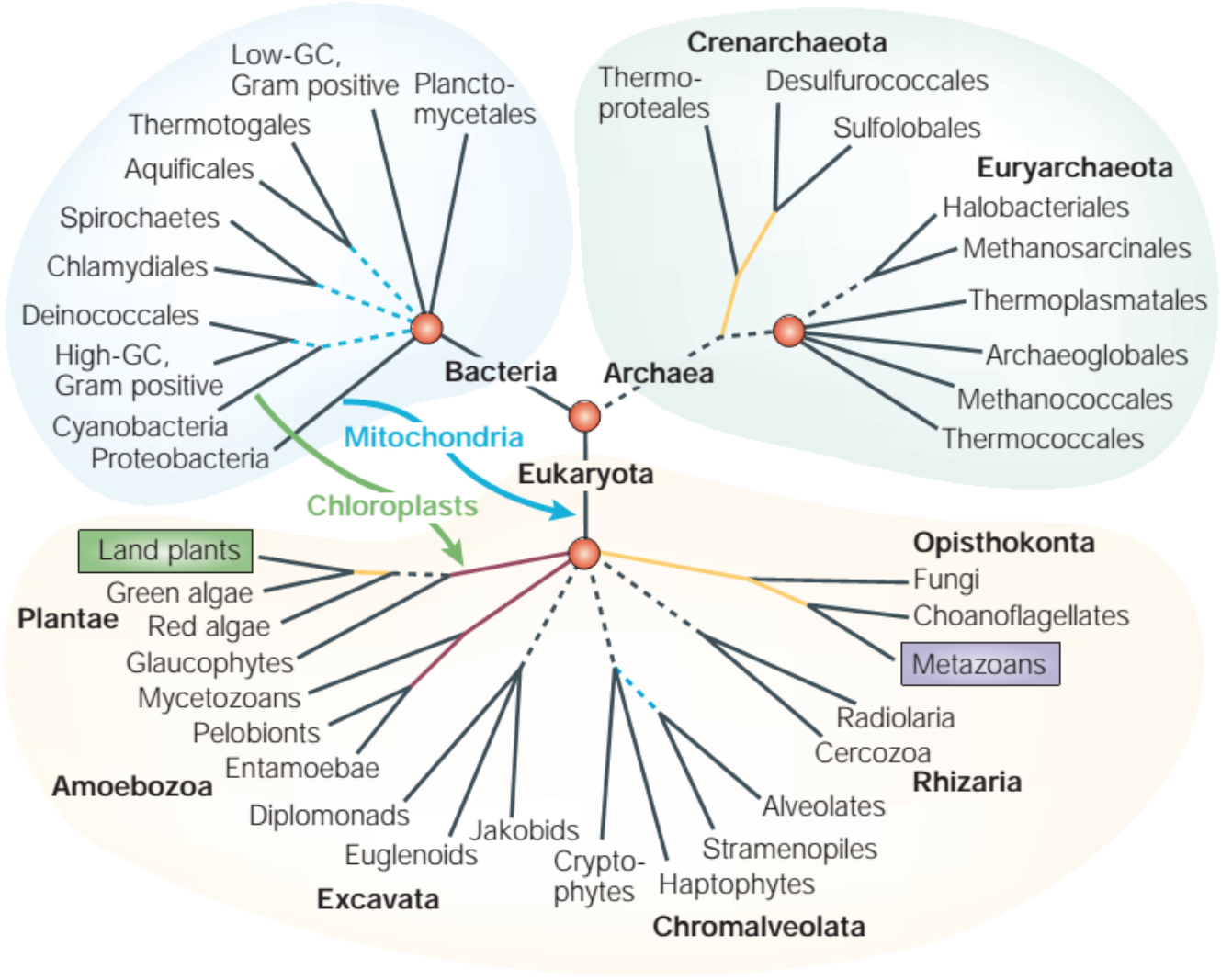
Phylogenomics

- The use of phylogenetic principles to make use of genomic data



Methods of phylogenomic inference

The flowchart shows steps in the inference of evolutionary trees from genomic data. Genomic information is obtained by large-scale DNA sequencing. In general, sets of orthologous genes are then assembled from specific sets of species for phylogenetic analysis. This homology or orthology assessment is a crucial step that is almost always based on simple similarity comparisons (for example, BLAST searches). Most methods used for the subsequent reconstruction of phylogenetic trees are either sequence-based or are based on whole-genome features.



Phylogenomics and the tree of life

A schematic representation showing recent advances and future challenges of the phylogenomic approach for resolving the main branches of the tree of life. This tree aims to represent a consensus view on evolutionary relationships within the three domains — Bacteria, Archaea, and Eukaryota — with hypothetical relationships indicated as dashed lines. The main branches that have been identified (purple) or confirmed (yellow) by phylogenomics are indicated. Blue dashed lines underline putative phylogenetic hypotheses that have been indicated by phylogenomic studies and need further investigation. The main uncertainties for which the phylogenomic approach might provide future answers are pinpointed by red circles. Note that most of the progress brought about by the phylogenomic approach has been realized at a smaller taxonomic scale for land plants, and for placental mammals within the metazoans (see main text). The two well-recognized endosymbiotic events involving bacteria that gave rise to eukaryotic organelles (mitochondria and chloroplasts) are indicated by arrows (blue and green, respectively). Note however, that other horizontal gene transfers and gene duplication events are not represented in this organismal tree, although they do constitute important aspects of genome evolution.

Phylogenomics and the evolutionary history of biological systems