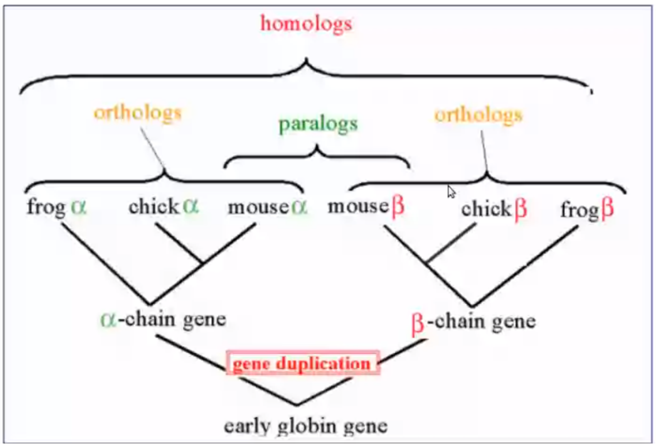
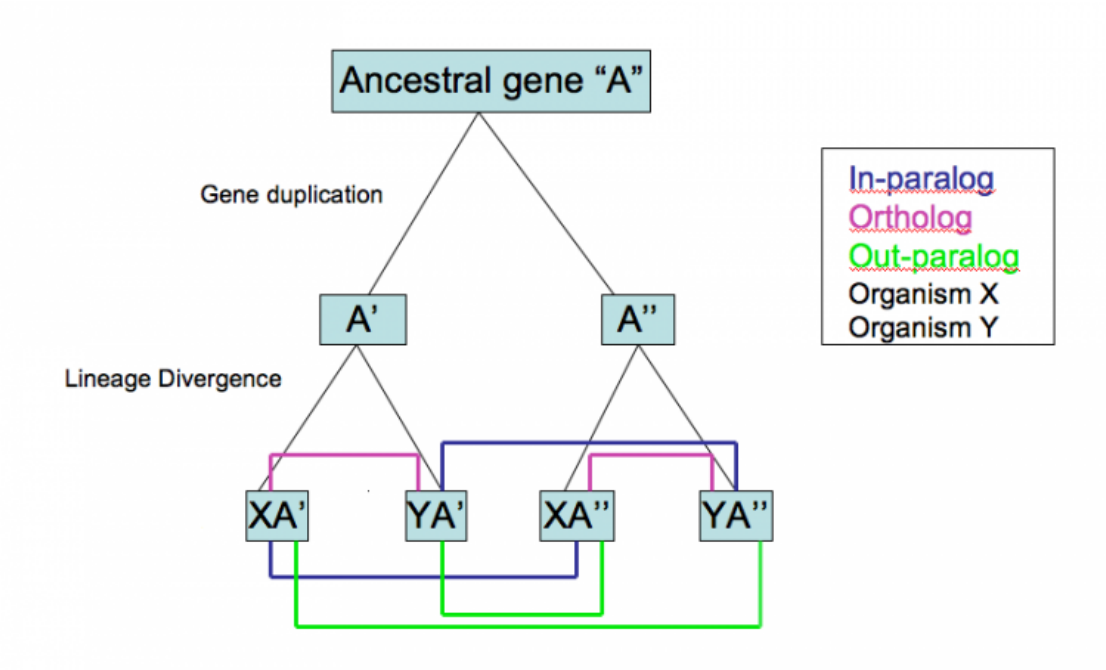
- **De Novo Origin Of Genes:** process by which new genes evolve from DNA sequences that were ancestrally non-genic.
- **Homology Based On Functional Inference:** If a sequence determines structure, which determines function.
- **Protein Domain:** Conserved part of a given protein sequence and tertiary structure that can evolve, function and exist independently of the rest of the protein.
- **Ontology:** The study of 'being'. ONtology oftens deals with questions concerning what entities exist
- **Enrichment Analysis:** method to identify clae of genes or proteins that are over-represented in a large set of genes or proteins, and may have an association with disease phenotypes.
    - **Fisher's Exact Test:** tests for the exact probability of observing a deviation from background population in a sample. Equivalent to chi-squared test but can be used for cases when the number of values is small.
    - **Correction For Multiple Testing:** Test for differences in the same sample of many categories (can perform a Fisher's exact test for each category). P-values should be corrected for multiple testing
- **Go Annotation:** encompasses the practice of capturing data about a gene product, GO annotation uses terms from the GO ontology to do so.
- **Pathway:** series of interactions that leads to a certain product or change.
- **Homology:** statement about common evolutionary ancestry of two sequences. Can only be true or false.
- **Analogous Structures:** Similar function but independent origin.
- **Similarity:** degree of likeness between two sequences, usually expressed as a percentage of similar or identical residues over a given length of the alignment.
- **E-value:** is the expectation value. The number of sequences that would be expected to have that score or higher if the query sentence were compared against a database containing unrelated sequences.
- **Orthology:** homologous sequences from the same ancestors which are separated from each other after a speciation event.
- **Paralogy:** paralogous genes are genes that are related via duplication events in the last common ancestor of species being compared.

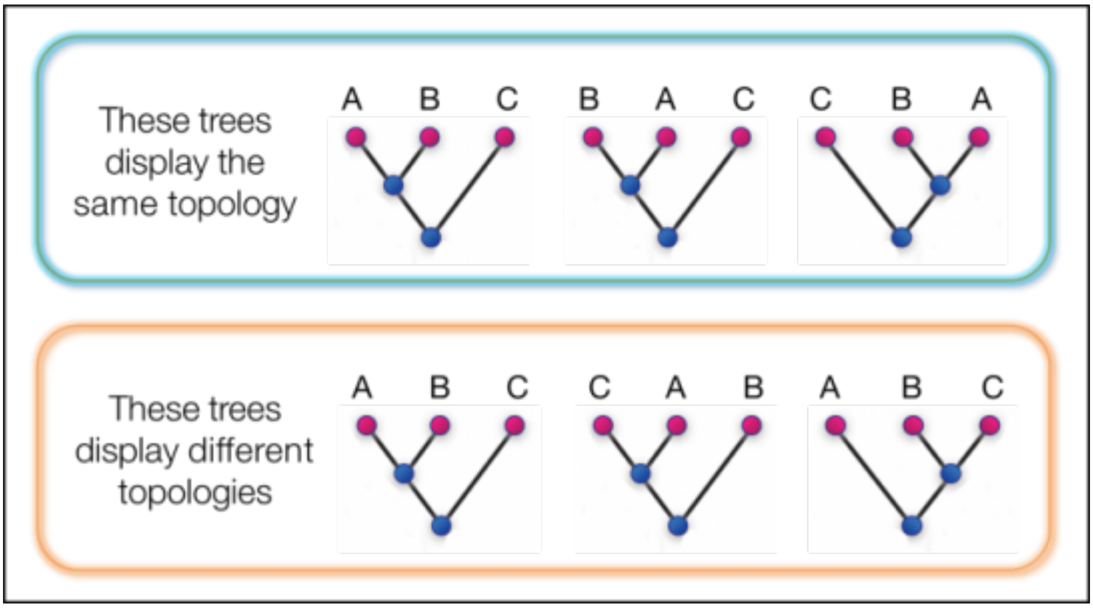


- **Gene Family:** a group of genes that share a common ancestry, gene families have hierarchical evolutionary relationship.
- **In Paralog:**
- **Out Paralog**



- **Orthologous Group:** group of sequences derived from a single gene in a common ancestor. they may include orthologs and in-paralogs
- **Phylogenetic Tree:** is a branching diagram (bipartite graph) showing the inferred evolutionary relationships among various biological species or other entities based on similarities and differences in their physical and/ or genetic characteristics.
- **Paraphyletic Group:** contains not all descendant form a common ancestor
- **Polyphyletic Group:** are from two different ancestors
- **Monophyletic Group:** group with all descendants from one common ancestor.
- **Exhaustive Search:** make all trees first, and the nsee which one best fits the data
- **Heuristic Search:** try to fins a way to find an optimal tree without testing them all. it is also needed an optimal criterion and its not guaranteed to find the best tree.
- **Neighbour-joining:** based on the current distances matrix. is a clustering method to create a tree
- **Maximum Likelihood:** computes the probability of obtaining the data given a defined hypothesis.
- **Bayesian Inference:** a bayesian approach will give the tree or set of trees that is most likely to be explained by the sequences.
- **Phylogenomics:** is the intersection between genomics and evolution. that is, looking at the genome looking from an evolutionary perspective, often using phylogenetics.
- **Phylome:** complete collection of evolutionary histories of all genes encoded in a given genome.
- **Synteny:** describes the physical co-localization of genetic loci on the same chromosome within an individual or species.
- **Motif:** nucleotide or amino acid sequence pattern that is widespread and has a biological significance.
- **Gene Order:** gene orders are the permutation of genome arrangement.
- **Dot Plot:** graphical method for comparing two biological sequences and identifying regions of close similarity after a sequence alignment. Each axis is a sequence.
  - **lines:** The line means regions of similarity.
  - **interpretation:** The continuous lines represent regions of similarity. While the discontinuous lines represent regions of genomic variations.
- **Topology:** topology is the study or analysis of configuration of parts or elements.

- **Convergent Evolution:** is the independent evolution of similar features in species of different periods or epochs in time. Convergent evolution creates analogous structures that have similar form or function but were not present in the last common ancestor of those groups.
- **Compensatory Mutation:** mutations that correct a loss of fitness due to earlier mutations.
- **Co-evolution:** occurs when two or more species reciprocally affect each other's evolution through the process of natural selection.
- **Transcriptome:** can be described as the complete collection of transcripts present in a specific cell, tissue ,organism, etc. at a given time-point.
- **Rna-seq:** used to indicate any RNA sequencing method based on a shotgun approach.
- **Tree topology:** The topology is the branching structure of the tree. It is of particular biological significance because it indicates patterns of relatedness among taxa, meaning that trees with the same topology and root have the same biological interpretation.


These trees display the same topology
A B C
B A C
C B A
These trees display different topologies
A B C
C A B
A B C

- **adjacent genes**
- **phylogenetic profile:** A phylogenetic profile represents a gene or protein family by serving for a taxonomic overview. It stores information about the presence and the absence of that protein in a set of genomes. By clustering identical or similar profiles, proteins with similar functions and potentially interacting are identified.
- **jaccard index:** is a statistic used for gauging the similarity and diversity of sample sets. The Jaccard coefficient measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets
- **hamming distances:** Given any two sequences of equal length, the Hamming distance is the number of positions at which corresponding symbols are different.

Final exam test:

1. **What statement is correct:**
    a. **genes can be made of DNA or RNA**
    b. **genes are all transcribed regions of a genome**
    c. **genes are all the functional regions of a genome**
    d. **all the statements are correct**

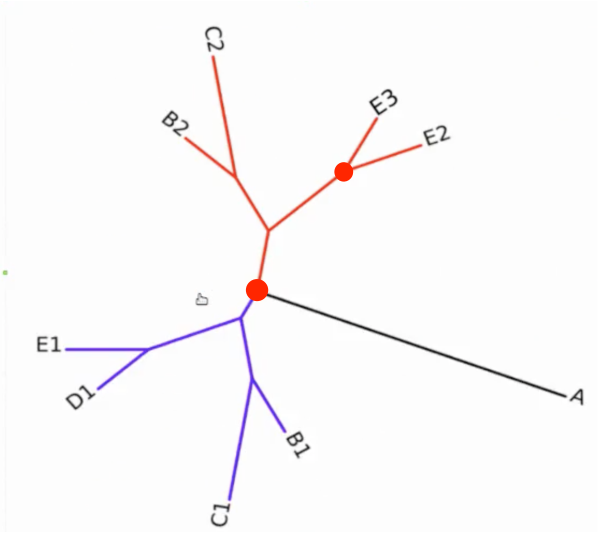
2. Out-paralogs, as compared to in-paralogs, are more:
    a. Appropriate to build species tree
    b. similar
    c. numerous
    d. divergent
3. What statement is correct:
    a. a species tree comprises speciation and duplication events
    b. branch lengths indicate level of confidence of a tree partition
    c. maximum likelihood is the fastest tree reconstruction model
    d. there are more possible rooted trees than unrooted trees
4. Regarding the reconstruction of gene trees:
    a. They are always rooted
    b. genes that contain duplictions cannot be used
    c. only orthologous genes can be used
    d. only homologous genes can be used
5. Gibbs sampling is a method to:
    a. discover conserved motifs in a set of sequences
    b. compute the similarity of two genomes in terms of gene order
    c. find shared structures between two RNA sequences
    d. calculate support of a phylogeny
6. Which signature is strongly suggestive of two genes having the same molecular function:
    a. the two genes encode the same protein domains
    b. all the answers are correct
    c. the two genes are co-expressed in the same tissue
    d. the two genes appear fused in the genome of another species
7. When can we say that a function (function A) is enriched in a gene set?
    a. when the function A is the most common among the genes in the set
    b. all the statements are correct
    c. when there are more genes with function A in the set than expected by chance in random samplings of the genome
    d. when the genes with function A are more expressed than the set of the genes
8. Which of the following statements about InParanoid is true?
    a. provides phylogenetic trees for several gene families
    b. is a method to predict protein domains
    c. none of the other answers is correct
    d. is a method to predict alternative splicing
9. RPKM unit of gene expression is obtained by:
    a. dividing reads by gene length and adjusting the GC content
    b. dividing read counts by total library size and adjusting GC content
    c. dividing read counts by total library size
    d. dividing real counts by total library size and gene/ transcript length
10. When is appropriate to use InterPro?
    a. all answers are correct
    b. when you have a genomic DNA sequence and are interested in gene annotation
    c. when you want to perform structural alignment of protein sequences
    d. when you want to know the function of an amino acid sequence or set of sequences.

Question 1
Provide the shortest and most inclusive definition of a gene

Question 2
What is a phylogenetic profile? How can we use them to know the function of an uncharacterized gene?

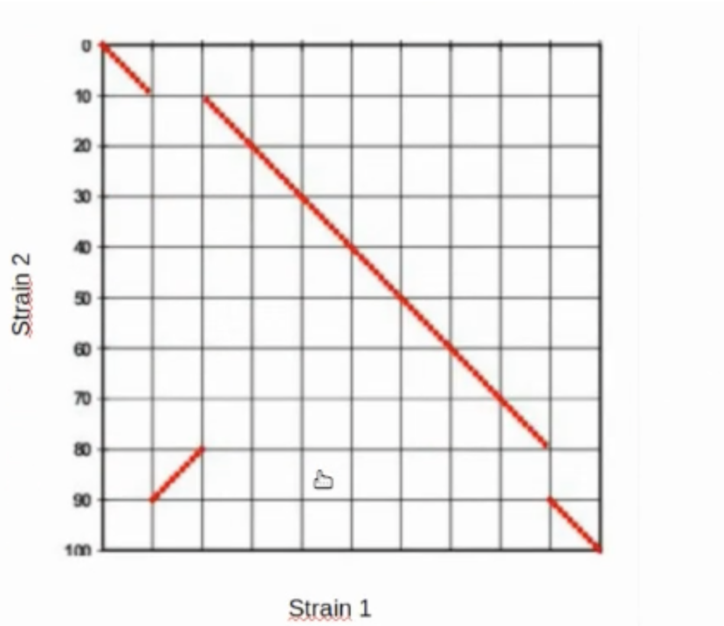
Question 3
Given the gene tree below, where letters indicate species and number genes (i.e. B1 and B2 are two genes of species B, and in which A can be used as an out-group). Using the algorithm of your choice

a) **list all orthologs of B2**
b) **list all paralogs of B2**
c) **write the most likely species tree in newick format**



Question 4
The plot below shows a hypothetical comparison between the genomes of two bacterial strains (Strain 1 and Strain 2, of which strain 2 has been evolved from strain 1)

a) **explain what type of plot is this one and its utility**
b) **what could be the unit of the axes?**
c) **can you reconstruct what rearrangement(s) occured during strain 2 evolution from its strain 1 ancestor.**



Recovery Exam Test:

1. If two genes are orthologous to each other, then
   a) They must be syntenic
   b) All responses are correct
   c) They must belong to different species
   d) They must have the same function
2. What statement is correct?
   a) Neighbor joining is more prone to long branch attraction than maximum likelihood
   b) Bayesian analysis is faster than neighbor joining methods
   c) Maximum parsimony is recommended for distantly related sequences
   d) There are more possible unrooted trees than rooted ones
3. What is the purpose of a gene concatenation approach?
   a) To build a species tree
   b) To detect genome rearrangements
   c) To predict the function of a gene
   d) To find syntenic genes
4. Which signature is strongly suggestive of two genes having the same biological function?
   a) The two genes encode the same protein domains
   b) The two genes are co-expressed across many tissues
   c) The two genes appear fused in the genome of another species
   d) All the answers are correct
5. Sorting by reversals is a method to
   a) Predict gene clusters
   b) Compute the similarity of two genomes in terms of gene order
   c) Discover conserved motifs in a set of sequences
   d) Reconcile gene and species tree

6. When two phylogenetic profiles are very similar
   a) All the statements are correct

  b) They have Jaccard index higher than 1
  c) They have hamming distance close to 0
  d) They have low mutual information

7. Which of the following statements about UniProt is wrong?
   a) Contains information of manually annotated proteins
   b) Allows you to convert other database identifiers to UniProt identifiers but not vice versa
   c) Contains information of computationally annotated proteins
   d) Is a freely accessible database of protein sequence and functional information
8. Which of the following statements about KEGG is true?
   a) It provides a genome browser that acts as a single point of access to annotated genomes for mainly vertebrate species
   b) Is a database for intensively studies model organisms
   c) Is a collection of databases dealing with genomes, biological pathways, diseases, drugs, and chemical substances
   d) Provides complete proteome sets for organisms whose genomes have been completely sequenced
9. Which of the following statement is true?
   a) Phred quality score of reads is stored in SAM and fastq files
   b) GFF file is required for fastq data trimming
   c) SAM file is a binary version of BAM file
   d) Each gene identifier has 4 lines in fastq file

Question 1

What is the mirror tree approach? What is its purpose and what is it based on? Can you explain how it works?

It is used to compare interaction between proteins. Look for distance matrices between proteins from a tree and then obtain the correlation.

Question 2

Provide the shortest and most inclusive definition of a protein domain. What is a promiscuous domain?

Question 3

Given the following gene tree in newick format:

(R1,(((H1,H2),C1),(H3,C2)))

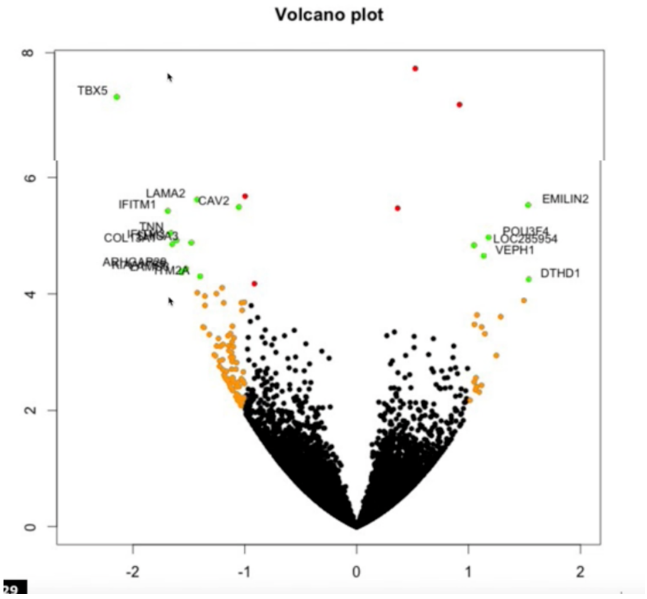
Where R represents Rat genes, H human genes, and C chimpanzee genes. Using the algorithm of your choice (indicate it).

   a) **Indicate which genes are orthologous or paralogous to each of the human genes.**
   b) **Considering the human lineage as a reference, sort the paralogs of H1 as in- and out- paralogs.**

Question 4

This is a typical volcano plot obtained after comparing gene expression of genes in two conditions (A versus B).

a) **Can you explain what is represented in each of the axes, and what units are typically used?**
b) **What is represented by each of the dots? What is the difference between green, brown, and black dots? And those on the left, and right?**
c) **Can you identify the gene that changed most of its expression?**



QUIZ ANSWERS:

1. **What would constitute a gene according to the modern definition**
   a. **a piece of DNA that is replicated and transcribed**
   b. **a piece of DNA that regulates the transcription of another DNA**
   c. **a piece of DNA that is transcribed as an RNA that inhibits the expression of another RNA**
   d. **any DNA that, when deleted confers a phenotype**
2. **The function of a gene…**
   a. **it can be described by GO terms only if there is experimental evidence**
   b. **it remains constant over the course of evolution**
   c. **is best determined by comparing its sequence with a database**
   d. **depends on the structural propertie of the molecule that is encoded by the gene**
3. **Which sentence is wrong?**
   a. **GO is the only controlled vocabulary that can be used to functionally annotate genes**
   b. **gene ontology terms can be used for genes with no experimental information**
   c. **GO terms are useful to detect what functions might be over-represented in a given dataset**
   d. **Gene ontology entries inform on the source of annotation**
4. **Regarding homology:**

   a. paralogous genes are homologs
   b. all statements are correct
   c. orthologous genes are homologs
   d. homologs are genes that share a common ancestor

5. What is a promiscuous domain
   a. Domains that have many different functions
   b. Domains that can bind many other protein sequences
   c. all statements are correct
   d. domains that can be present in many different protein families
6. Orthologs, as compared to homologs, are more
   a. likely to have complementary functions
   b. all options are correct
   c. appropriate to reconstruct a species tree
   d. similar in terms of their sequence
7. Select the incorrect statement regarding phylogenetic trees:
   a. the newick format is a graphical representation of phylogenetic trees with edges and nodes
   b. bootstrap values provide information on how much support a give clade has from the analysis
   c. in a phylogenetic tree all edges can be rotated without changing the topology
   d. for a given topology, there are more rooted trees than unrooted trees
8. Out paralogs are…
   a. not necessarily encoded in the same genome
   b. gene that result from duplication
   c. more distantly divergent as compared to in-paralogs
   d. all the responses are correct
9. Associate the correct questions and answers:
   a. Alpha and beta globin are: paralogous
   b. Wings of bats and bird are: non-homologous
   c. Mammals are: monophyletic
   d. winged animals are: polyphyletic
10. Phylogenetic reconstruction methods based on Bayesian analysis:
    a. Need a set of prior probabilities for the parameters of the model
    b. are generally the fastest among tree reconstruction models
    c. need to be run on a bootstrapped set of alignments to assess the support of the tree partitions
    d. generally use a method called joint estimation to compute the likelihood of a tree over a set of parameters
11. Match the following concepts and algorithms with the correct context:
    a. Chromosome painting: is an algorithm to detect large chromosomal rearrangements
    b. Sorting by reversals: is an algorithm to find the minimum number of rearrangements between two series of elements
    c. Gibbs sampling: is an algorithm to discover enriched motifs that are enriched in a set of sequences
12. REgarding the reconstruction of species trees:
    a. The more species you compare, the fewer conserved genes you can use in gene concatenation approach
    b. all existing methods rely on sequence alignments at some point

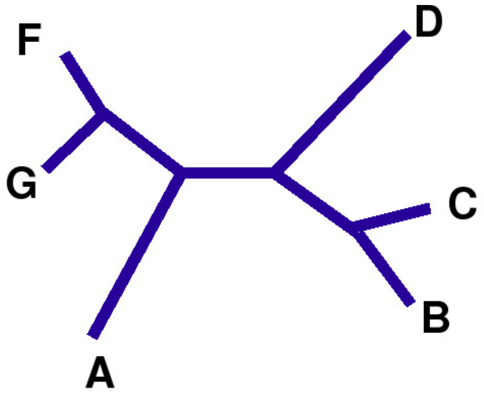
    c. the super-tree approach is the fastest and most accurate method
    d. genes that contain duplications cannot be used

13. Which statement is correct regarding protein domains
    a. Promiscuous domains mediate binding to other domains
    b. they can exert a function independently of the rest of the protein
    c. they are longer than 200 amino acids
    d. they are separated by introns in eukaryotic genes.

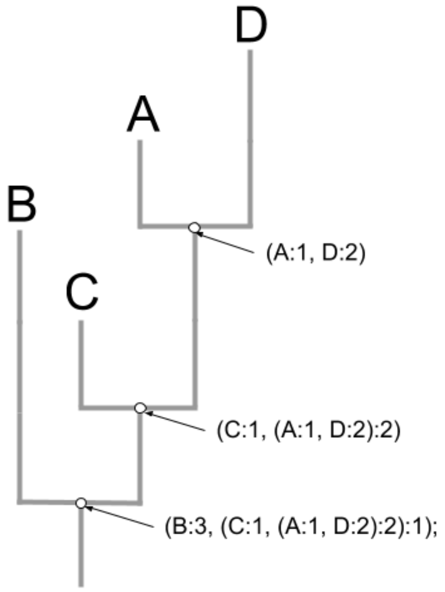
14. Genes that co-evolved, being similarly present or absent form genomes:
    a. the comparison of their phylogenetic profiles will show high Jaccard Indexes
    b. all the answers are correct
    c. the comparison of their phylogenetic profiles will show high Hamming distances
    d. their molecular functions are likely to be equivalent.

15. Maximum likelihood phylogenetic reconstruction methods:
    a. all the answers are correct
    b. generally use a method called joint estimation to compute the likelihood of a tree over a set of parameters
    c. provide the probability (likelihood) that the reconstructed tree is correct
    d. uses exhaustive search

16. Adjacent genes are:
    a. co-regulated
    b. part of a gene cluster
    c. neighbors to each other
    d. all answers are correct



(((F , G) , A) , (B , C) , D)



Gene annotation methods:

1. **phylogenetic profiling:** Phylogenetic profiling is a technique in which the joint presence or absence of two traits across a large number of species is used to infer meaningful biological connections. Similarity between profiles is an indicator of functional coupling between gene products: the greater the similarity, the greater the likelihood of proteins sharing membership in the same pathway or cellular system. Because of this property, uncharacterized proteins can be assigned putative functions, based on the similarity of their profiles with those of known proteins.
   - Genes with similar phylogenetic profiles tend to be involved in the same biological process.
   - Genes with complementary phylogenetic profiles tend to have a similar biochemical function.
2. **chromosomal colocalization:** Intrachromosomal colocalization can strengthen co-expression and co-modification of neighboring gene pairs and their conservation across species
3. **homology based methods**

4. **congruence**
    - A gene tree is congruent when It has the same topology as the species tree.
    - Congruence between gene trees and species trees can give us some evolutionary information.

Paralogy/orthology + methods

Original definitions of orthology and paralogy by Walter Fitch:

- Where the homology is the result of gene duplication so that both copies descended side by side during the history of an organism, the genes should be called paralogous.
- Where the homology is the result of speciation so that the history of the gene reflects the history of the species the genes should be called orthologous.
- In-paralogs and out-paralogs are definitions that depend on a given speciation event.
- In-paralogs are paralogs that follow a given speciation event.

Methods to infer paralogy and orthology relationships: Methods based in phylogeny were not used at a large scale due to limitations in computational power. However, now fast pipelines and algorithms are available: Ensembl trees, PhylomeDB, TreeFarm.

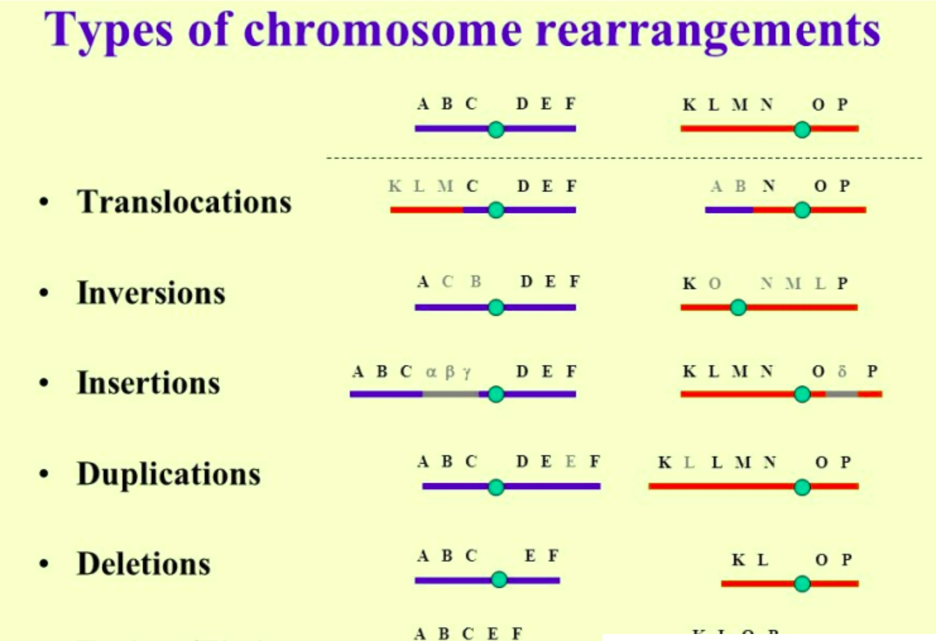
1. **Phylogenetic inference:** Reconciliation
    a. Build a gene tree
    b. Compare to the species tree
    c. Infer duplications and speciation events
    d. Assign orthology and paralogy relationships accordingly
2. **Similarity based approaches:**
    a. Best reciprocal hits: Detects all orthologies as one-to-one. Highly affected by paralogy. Low rate of false positives but high rates of false negatives. This is the simplest and fastest method.
    b. COG-like approach: Exploits multi-species information. Predicts clusters of orthologous (in-paralogs). Can be used at different stringent levels.
3. **Clustering methods:**
    a. Orthologous groups: Each orthologous group has implicit the specification of an ancestral species of reference (a speciation node).

Databases:

- **InParanoid:** Orthology
- **Pfam:** Protein families and hidden markov models
- **Uniprot:** protein sequence and functional information, many entries being derived from genome sequencing projects
- **PhylomeDB:** Trees
- **GenBank**
- **EMBL**
- **DDBJ**
- **Swiss-Prot**
- **InterPro:** Provides functional analysis of proteins by classifying them into families and predicting domains and important sites. **Interproscan:** Tool that allows to scan a sequence for matches against the InterPro signature databases.
- **Ensembl:** Gene sequence, splice variants and annotation at genome, gene and protein level
- **QuickGo:** Go annotations

- **KEGG:** Pathways
- **Ensembl**: genome browser for vertebrate and model organisms genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation.
- **Prosite:**

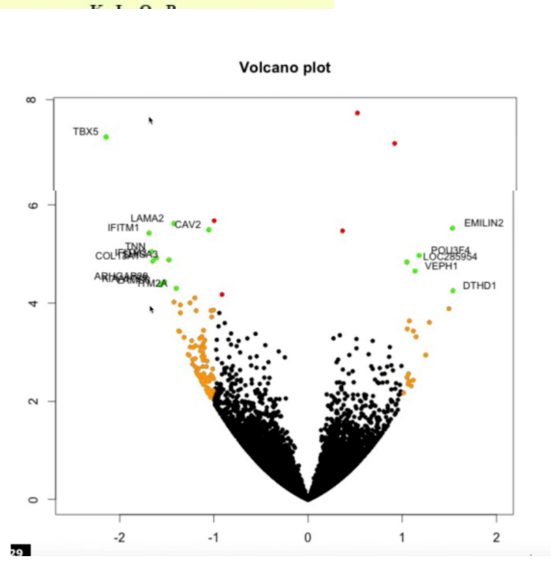
Homology, Orthology and Paralogy, are relationships between GENES or GENETIC LOCI not SPECIES.



Types of chromosome rearrangements

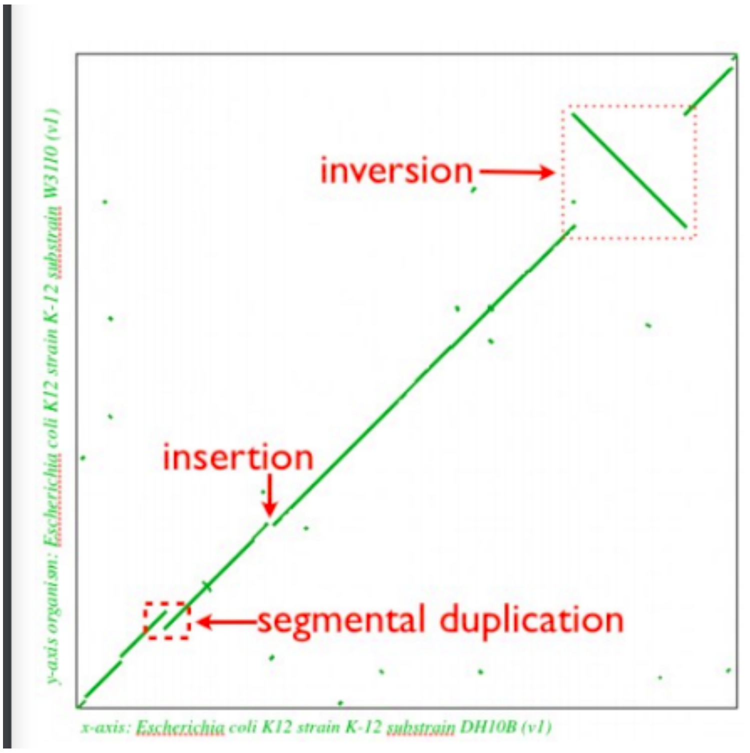
- Translocations
- Inversions
- Insertions
- Duplications
- Deletions
- Fusions/Fissions

Plots

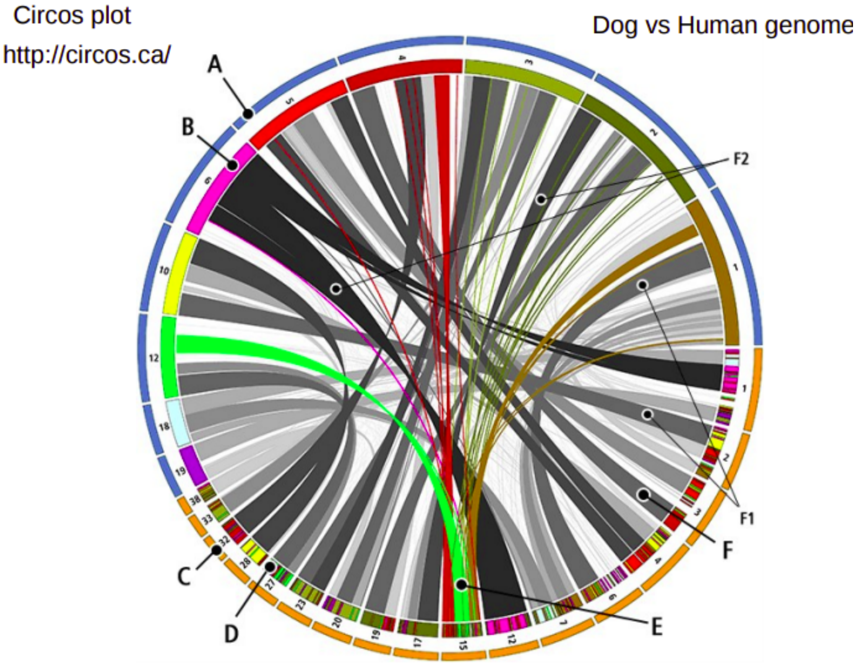


Volcano plot: Represents the p-value and the log2fold change (how much the expression changes in a logarithmic scale). The more to the extremes, the more change in expression values. Since we also account for the p-value, the most differentially expressed genes (the most significant ones) are the ones at the top. Each dot represents a gene. Black dots represent genes that do not pass the threshold of p-value and log2fold change. Orange dots only pass the log2fold change value. Red dots only pass the p-value threshold. Green dots are the best results and pass both thresholds.

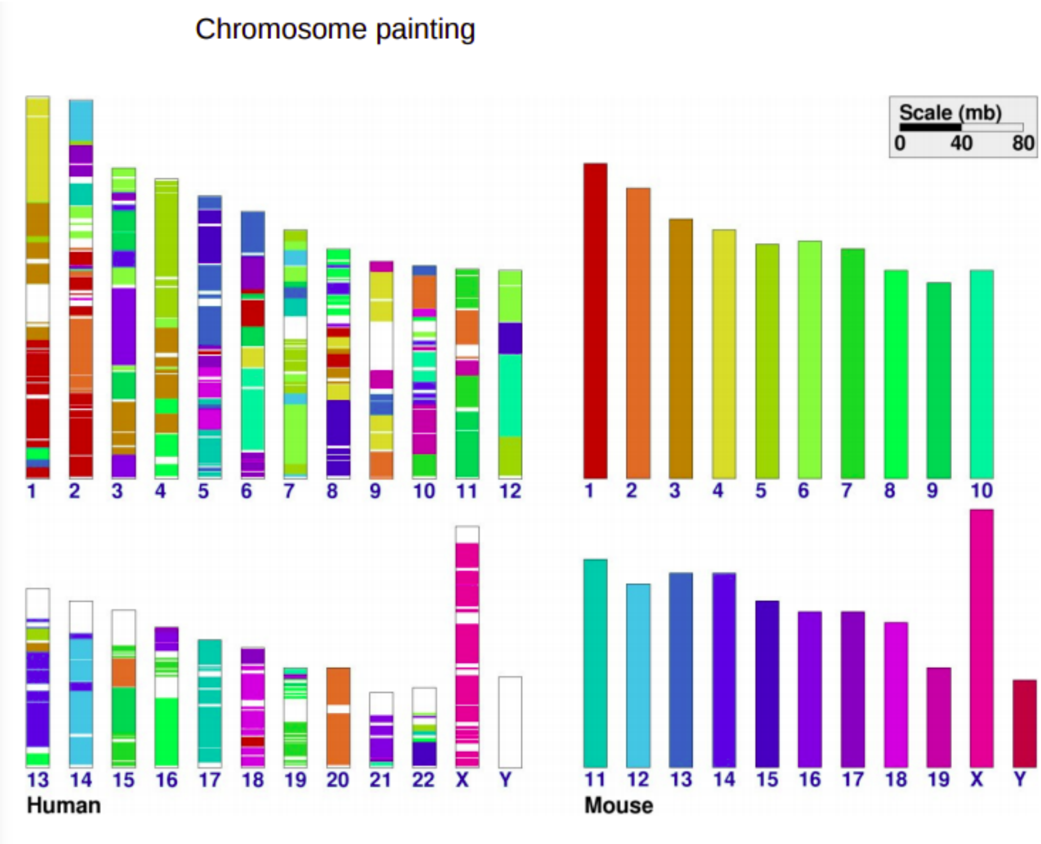
Dot plot: To compare genomes. One genome is represented in one axis and another genome in the other. Each dot represents a region/ fragment of hoology. If we compared a genome with itself we would observe a diagonal line with no gaps.



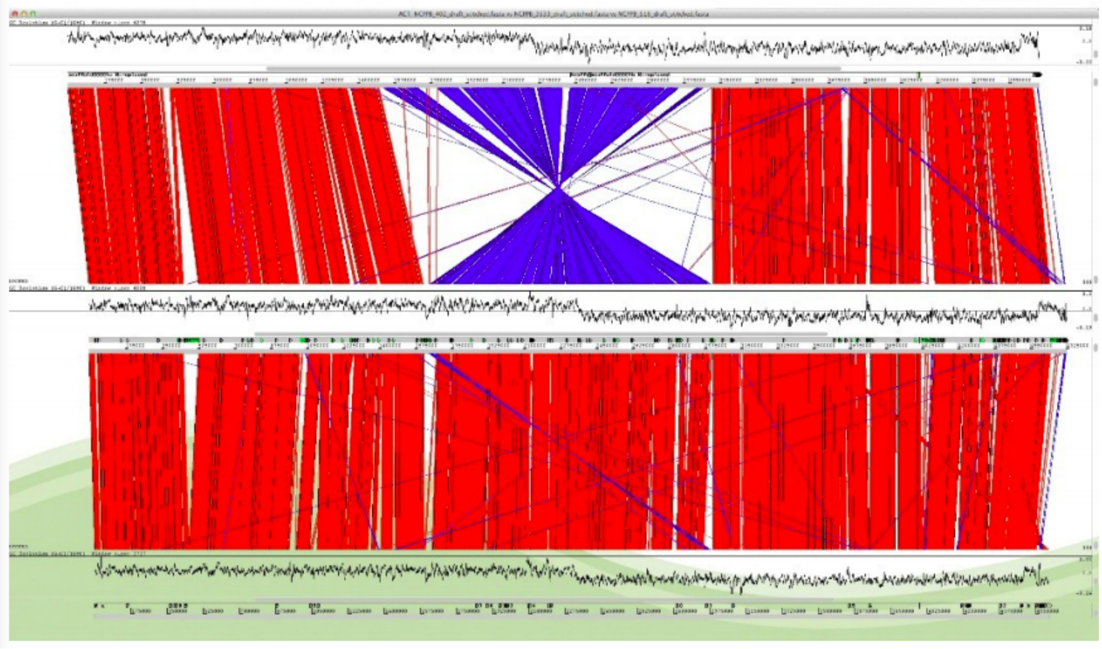
Circos plot: Each segment represents one chromosome, we can compare chromosomes from two different species. There are lines uniting windows of these chromosomes that are similar to each other. As more lines crossing each other and regions that connect to multiple other parts of the other genomes, means that rearrangements have been produced. White regions that have not an homologous regions can be deletions or insertions.



Chromosome painting: To compare genomes we can also use chromosome painting, where you pick a color for each mouse chromosome and paint it in the human chromosome, and compare the rearrangements.
X has the least rearrangements, since it only has chunks from one chromosome (no rearrangements with other chromosomes).



Whole genome alignment: the windows approach to be able to sequence very large genomes. But in this case, we do take into account where a window starts and ends, no como en ANI where we just compute the average. In conclusion, to solve the problem of handling genome rearrangements and also be able to sequence very large genomes, we align bits and pieces while being aware of the relative coordinates of the pieces. Blue lines indicate that a window aligns to a different part of the other genome. Here we can see an inversion.



Index Review:

1. **Genes and their functions:**
    a. **Protein-coding and non-coding genes.**
        Modern gene definition: A piece of DNA that is transcribed int orNA and that inhibits the expression of another RNA.
    b. **Gene structure and expression**
        Splicing.
    c. **Functional roles of genes**
    d. **Relationships between sequence, structure and function and their evolution**
    e. **Homology based functional inference**
    f. **Protein domains and domain shuffling**
    g. **Prediction of protein subcellular localization**
        Levels
    h. **De novo origin of genes**
    i. **Pathways**
    j. **The Gene ontology**
    k. **Enrichment analysis**
2. **Comparative sequence analysis**
    a. **Homology, Paralogy and Orthology**
    b. **Methods for predicting orthology and paralogy (clustering-based and phylogeny based)**
    c. **Gene families**
    d. **Gene duplication**
    e. **Neo- and sub-functionalization**
    f. **Gene family expansions and contractions**
    g. **Adaptation and genome evolution**
3. **Phylogenetic analysis**
    a. **Gene family tree reconstruction**
    b. **Inference of gene duplication and other evolutionary events**
    c. **Detection of functional divergence**
    d. **Dn/Ds analysis**

4. **Phylogenomics:**
    a. **Species tree reconstruction**
    b. **Genome-wide phylogenetic analysis (phylome)**
    c. **Gene tree vs species tree**
    d. **Non-vertical process of evolution**
    e. **Horizontal gene transfer**
    f. **Hybridization**
    g. **Whole genome duplication**
5. **Genome comparison and gene order**
    a. **Genome alignments**
    b. **Detection of conserved regions**
    c. **Conserved motif discovery**
    d. **Genome rearrangements**
    e. **Synteny analysis**
    f. **Prediction of function from conserved gene order.**
6. **Phylogenetic profiling and co-evolution**
    a. **PResence/ absence pattern**
    b. **Convergent evolution**
    c. **Gene tree comparison**
    d. **Co-evolution between genes**
7. **Gene expression analysis**
    a. **Genomic-based methods to assess gene expression**
    b. **Transcriptomics analysis**
    c. **Differential gene expression**
    d. **Functional inference form co-expression networks**
8. **Comparative and functional analysis of genomes**
    a. **The Encode project**
    b. **Long-non-coding RNAs**
    c. **Efforts in other model and non-model species**
    d. **Diversity of life and the tree of life**
    e. **Variation of genome size and organization**
    f. **The C-paradox**
    g. **Extreme genome expansions and reductions**