B1) You have a set of 40 complete, annotated genomes. Briefly define what would be your strategy to reconstruct the relationships among these species.
The supertree approach combines smaller trees from different subsets of species to construct a larger phylogenetic tree. It integrates information from individual trees to infer relationships among species present in multiple subsets.

The supermatrix approach is a method used to construct a phylogenetic tree by combining sequence data from multiple genes or genomic regions into a **single** matrix. This matrix includes information from all the genes or regions and is used to infer the relationships among the species.

To reconstruct the relationships among the 40 complete, annotated genomes using the supertree approach:
1. Use methods to infer orthology (clustering methods for example
2. Perform a MSA
3. Perforn the phylogenetic inference
    1. Run NJ or UPGMA algorithm
4. Run species reconciliation approach or species-overlap algorithm to detect duplication/loss events

Since we are working with a lot of genomes, we should use an heuristic algorithm.

B2) Mention two alternative methods to predict orthology and paralogy relationships, briefly describe their basis and discuss the main advantages and disadvantages for each method in comparion to the other one.
Best Reciprocal Hits (BRH) is a method for predicting orthology relationships. It detects one-to-one orthologies by identifying pairs of proteins that reciprocally match each other as the best hits. However, BRH is highly influenced by paralogy, which can lead to false negatives in certain cases. It is used only for closely related species where there are not many duplications.

InParanoid is an improved version of BRH that addresses some of its limitations: it identifies orthologs and considers in-paralogs. InParanoid works well for pairwise comparisons and can be extended to handle multi-species analyses using the Multi-Paranoid approach. It takes into account more complex evolutionary scenarios and can capture recent duplications. However, it can be computationally intensive and its accuracy may be influenced by the quality of the DBs used.

In summary, BRH is a simpler and faster method with a higher risk of false negatives but a lower rate of false positives. InParanoid improves upon BRH by considering in-paralogs and complex scenarios, but it is more computationally demanding and sensitive to DB quality and parameter choices.

C1) Given this species tree (left) and this gene tree (right), where in the gene tree the species code is found at the end of the gene code, indicate the number of times that each node in the species tree (internal and terminal) has been duplicated in the gene tree using the species overlap algorithm.
Duplication in nodes: A, D, H

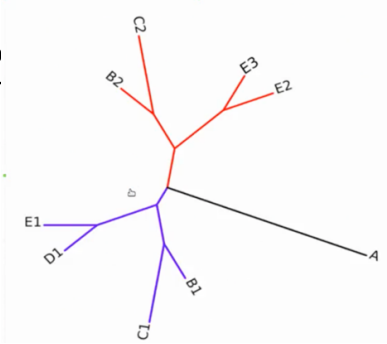
Exam problem: we compare two closely related species in terms of gene order conservation and have obtained the following dotplot. Circle the inversions with an “I” and the chr fusions with an “F”.
There are no whole chr inversions since, the reference sequence in one species is in one way and in the other it's in the other way.
We find chr fusion in G and H (two different chrs) in the y-axis species, and in the x-axis species are fused in chr 5.

Question 1.
Provide the shortest and most inclusive definition of a gene.
Any sequence of DNA or RNA that codes for a molecule that has a function

Question 2.

What is a phylogenetic profile? How can we use them to know the function of an uncharectized gene?
Describes the presence or absence of a protein in a set of genomes. Similarity between profiles is an indicatory of functional coupling between gene products.
If our objective is to characterize an uncharacterized gene, we can search for a gene that has a similar phylogenetic profile, meaning that if there are similarities between both profiles, then both genes are likely to be involved in similar biological processes (this does **not** mean that they have the same function).

Question 3
Given the gene tree below, where letters indicate species and number gen
of species B, and in which A can be used as an out-group). Using the algorith
Using the species-overlap algorithm…
- List of all orthologs of B2: A, C2, E2, E3
- List of all paralogs of B2: B1, C1, D1, E1
- Write the most likely species tree in Newick format: (((D, E), (C, B)), A);



→ Determine orthologs: partiendo del MRCA, mirar cuales son de
de diferent especie a B2

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)
- Explain what type of plot is this one and its utility

This is a dot plot. It's a graphical method for comparing two biological sequences and identifying regions of close similarity. Lines represent regions of similarity between the two bacterial strains.

- What could be the unit of the axes?

Kilobases (Kb)

- Can you reconstruct what rearrangement(s) occured during strain 2 evolution from its strain 1 ancestor.

Translocation in the 10th place, where 10 Kb have been moved to the 80th place. This translocated region has also been inverted.

Question 1
What is the mirror tree approach? What is its purpose and what is it based on? Can you explain how it works?
The mirror tree approach compares gene trees and their distance matrices to investigate coevolution and protein interactions between organisms. It is based on the idea that if two proteins have coevolved, their gene trees and distance matrices will exhibit a strong correlation. By measuring the correlation between the matrices, the approach identifies potential interactions between proteins of different organisms.

Question 2
Provide the shortest and most inclusive definition of a protein domain. What is a promiscuous domain?
Conserved part of a given protein sequence and tertiary structure that can evolve, function and exist independently of the rest of the protein. Promiscuous domains are domains that are present in more than one protein family.

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).
Using the species-overlap algorithm…
a) Indicate which genes are orthologous or paralogous to each of the human genes.

- Paralog to H1: H2, H3, C2
- Ortholog to H1: C1, R1 → FROM DIFFERENT SPECIES
- Paralog to H2: H1, C2, H3
- Ortholog to H2: C1, R1 → FROM DIFFERENT SPECIES
- Paralog to H3: H1, H2, C1
- Ortholog to H3: C2, R1 → FROM DIFFERENT SPECIES

b) Considering the human lineage as a reference, sort the paralogs of H1 as in- and out- paralogs.
- In-paralogs: H2
- Out-paralogs: H3, C2

Question 4

This is a typical volcano plot obtained after comparing GE 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?

- log-scaled P-value (Y axis) and log-scaled fold change (X axis).

The p-value indicates the probability of obtaining the observed differential GE by chance. By taking the negative logarithm, smaller p-values are displayed as larger values on the y-axis, emphasizing **greater statistical significance**.

The fold change indicates the magnitude of difference in GE levels between the two conditions. Positive values indicate **upregulation** (higher expression in condition B), while negative values indicate **downregulation** (higher expression in condition A).

b) What is represented by each of the dots? A gene

What is the difference between green, brown, and black dots?

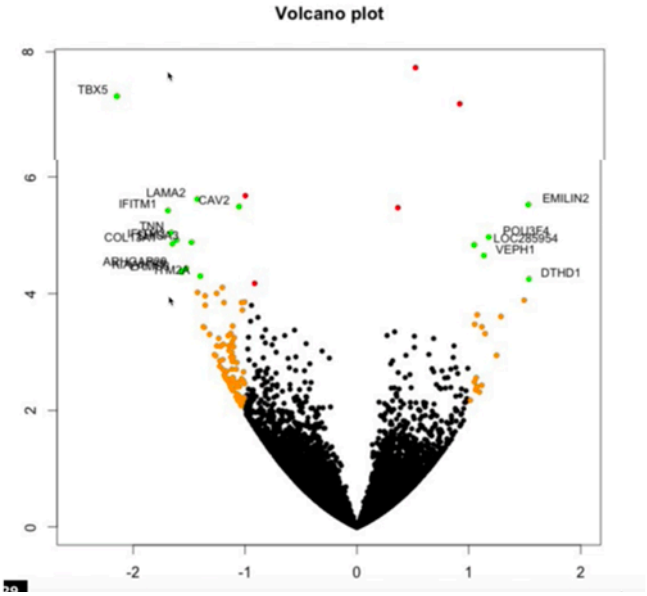
Black dots are not significant and don't have a differential expression (fold change values close to zero or negligible changes, resulting in non-significant p-values).

Brown dots have a differential expression but are not significant (noticeable fold change but do not meet the defined statistical threshold for significance).

Green dots have a significant differential expression (both a considerable fold change and a statistically significant p-value are shown, indicating a robust change in expression).

And those on the left, and right? If they are up or down regulated.

c) Can you identify the gene that changed most of its expression? TBX5



Multiple choice questions

Which of the following entities would necessarily be considered a gene?

- A transcript that is translated into protein that has an enzymatic activity. A transcript is not a gene.
- Region of DNA that binds proteins that open the chromatin. This is an enhancer. To be a gene it needs encode for a molecule that has a function.
- Region of DNA that is transcribed into RNA molecule that serves as a scaffold for enzymes
- All correct

Which statement is correct regarding orthology and paralogy?

- Two homologous genes that are in genomes from different species are necessarily orthologs. Homologous genes in the **same** genome are paralogs, but not all homologous genes in **different** genomes are orthologs.
- Orthologs, as compared to paralogs, share a higher level of homology
- The best bidirectional hit approach is more prone to error when comparing closely related species.
- Clustering approaches define orthologous groups which contain both paralogs and orthologs.

Maximum likelihood phylogenetic reconstruction methods

- Provide the probability (likelihood) that the reconstructed tree is correct. They aim to find the tree and model parameters that maximize the likelihood of the observed data given a specific evolutionary model.
- Generally use a method called joined estimation to compute the likelihood of a tree over a set of parameters.
- The 2 options above are correct
- Uses exhaustive searches. They employ heuristic algorithms that explore the parameter space to find the optimal tree topology and model parameters, and do not exhaustively evaluate all possible trees.

Genes that have co-evolved in terms of presence/absence (they are present or absent in the same species, so their phylogenetic profiles will be really similar).

- The comparison of their phylogenetic profile will show high jaccard indexes.
- The comparison of the phylogenetic profiles will show high hamming distances. Low
- 2 are correct
- Their molecular functions are likely to be equivalent. Having very similar genetic profiles indicates that they are related in the same process (not the same function).

We want to predict orthology and paralogy relationships using a phylogenetic approach coupled with a reconciliation method.

- A node is a duplication node only when there are common species on either side of the node. This is true if we were using species overlap
- A node is a speciation node only when there are common species on either side of the node.
- A node is a duplication node only when there are no common species on either side of the node. A node can be a speciation node when there is no common species on either side of the node.
- A node is a speciation node only when there are no common species on either side of the node.

1. What statement is correct:

- genes can be made of DNA or RNA
- genes are all transcribed regions of a genome. Not all regions of DNA/RNA are transcribed into functional molecules
- genes are all the functional regions of a genome. There are non-coding regions in the genome that do not encode genes.
- all the statements are correct

2. Out-paralogs (D → S), as compared to in-paralogs (S → D), are more:

- Appropriate to build species tree
- similar
- numerous
- divergent.

These duplicates have had more time to accumulate genetic changes and mutations, leading to greater sequence divergence between them. In contrast, in-paralogs are gene duplicates that arise from duplication events within the same genome after the speciation event, and thus they are expected to be more similar to each other.

3. What statement is correct:
- a species tree comprises speciation and duplication events. Gene trees, instead of species trees
- branch lengths indicate level of confidence of a tree partition. Branch lengths represent measures evolutionary change, such as the number of substitutions or time estimates.
- maximum likelihood is the fastest tree reconstruction model. Computational speed can vary depending on the dataset size, model complexity, and available computational resources.
- there are more possible rooted trees than unrooted trees

The number of possible rooted trees is always **one** more than the number of unrooted trees.

4. Regarding the reconstruction of gene trees:
- they are always rooted
- genes that contain duplications cannot be used
- only orthologous genes can be used
- only homologous genes can be used

5. Gibbs sampling is a method to:
- discover conserved motifs in a set of sequences
- compute the similarity of two genomes in terms of gene order. Whole-genome alignment algorithms are more appropriate for this purpose.
- find shared structures between two RNA sequences. RNA folding algorithms or comparative RNA structure prediction approaches are more appropriate for this purpose.
- calculate support of a phylogeny. This task is done by bootstrapping or Bayesian inference.

6. Which signature is strongly suggestive of two genes having the same molecular function:
- the two genes encode the same protein domains
- all the answers are correct
- the two genes are co-expressed in the same tissue. It does not guarantee the same molecular function (maybe they are just being part of the same pathway or regulatory network.)
- the two genes appear fused in the genome of another species. Gene fusion events can occur for various reasons and may have different functional implications.

Protein domains correspond to specific functional units or motifs within proteins. Genes encoding the same protein domains are more likely to have similar or overlapping molecular functions.

7. When can we say that a function (function A) is enriched in a gene set?
- when the function A is the most common among the genes in the set
- all the statements are correct
- when there are more genes with function A in the set than expected by chance in random samplings of the genome
- when the genes with function A are more expressed than the set of the genes. Enrichment analysis focuses on the overrepresentation of genes with a particular function in a gene set, irrespective of their expression levels.

8. Which of the following statements about InParanoid is true?
- provides phylogenetic trees for several gene families
- is a method to predict protein domains
- none of the other answers is correct
- is a method to predict alternative splicing

9. RPKM unit of GE is obtained by:
- dividing reads by gene length and adjusting the GC content
- dividing read counts by total library size and adjusting GC content

- dividing read counts by total library size
- dividing read counts by total library size and gene/transcript length

This normalization method allows for the comparison of GE levels between samples while accounting for differences in library size and gene length.

10. When is appropriate to use InterPro?
- all answers are correct
- when you have a genomic DNA sequence and are interested in gene annotation. Gene annotation involves identifying coding regions, promoters, and regulatory elements within the DNA sequence, which is performed using specialized gene prediction tools and DBs.
- when you want to perform structural alignment of protein sequences. InterPro is not designed for performing structural alignments of protein sequences.
- when you want to know the function of an amino acid sequence or set of sequences.

1. If two genes are orthologous to each other, then
a) They must be syntenic. Not necessarily located on the same chr or chromosomal region
b) All responses are correct
c) They must belong to different species
d) They must have the same function

Orthologs are found always in different species.

2. What statement is correct?
a) Neighbor Joining is more prone to long branch attraction than Maximum likelihood
b) Bayesian analysis is faster than NJ methods. Bayesian analysis requires more computational resources and time compared to NJ methods since it involves complex calculations and Markov chain Monte Carlo (MCMC) simulations, which can be computationally intensive.
c) Maximum parsimony is recommended for distantly related sequences. Maximum parsimony is less suitable for distantly related sequences due to the accumulation of multiple substitutions over time. ML and Bayesian methods are preferred as they can better account for evolutionary complexities.
d) There are more possible unrooted trees than rooted ones.

NJ is known to be susceptible to the long branch attraction phenomenon, where long branches on a phylogenetic tree can appear to be more closely related. ML methods are more robust against this issue.

3. What is the purpose of a gene concatenation approach?
a) To build a species tree
b) To detect genome rearrangements. Comparative genomics or genome assembly techniques are used.
c) To predict the function of a gene
d) To find syntenic genes

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. Jaccard index measures the similarity between two sets in ranges from 0 to 1
c) They have Hamming distance close to 0
d) They have low mutual information. Mutual information measures the statistical dependence between two variables. It does not directly measure the similarity or dissimilarity between phylogenetic profiles.

Hamming distance measures the number of positions at which two sequences differ. When two phylogenetic profiles are very similar, their Hamming distance would be close to 0, indicating minimal differences between the profiles.

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. SAM files do not directly store the quality scores, they contain alignment information, such as read alignments to a reference genome.
b) GFF file is required for fastq data trimming
c) SAM file is a binary version of BAM file. SAM files are plain text files, while BAM files are compressed binary files.
d) Each gene identifier has 4 lines in fastq file

In a FASTQ file, each sequence read is represented by four lines:
1. Identifier line (starts with '@') containing information about the read.
2. Sequence line containing the actual DNA/RNA sequence.
3. Quality score line representing the quality of each base call in the sequence
4. Optional comment line (starts with '+') providing additional information or annotations.

1. What would constitute a gene according to the modern definition
- a piece of DNA that is replicated and transcribed. Not all replicated and transcribed DNA represents a gene; many regions of the genome do not code for functional products and are not considered genes.
- a piece of DNA that regulates the transcription of another DNA. While regulatory elements (promoters, enhancers) are important for GE, they are not genes themselves. They do not encode functional products like proteins or RNAs.
- a piece of DNA that is transcribed as an RNA that inhibits the expression of another RNA
- any DNA that, when deleted confers a phenotype

Refers to ncRNA genes, which produce RNA molecules that play a role in inhibiting or regulating the expression of other genes.

2. The function of a gene…
- it can be described by GO terms only if there is experimental evidence
- it remains constant over the course of evolution
- is best determined by comparing its sequence with a database
- depends on the structural propertie of the molecule that is encoded by the gene

Understanding the structural properties of the gene product can provide insights into the gene's function.

3. Which sentence is wrong?
- GO is the only controlled vocabulary that can be used to functionally annotate genes. There are other controlled vocabularies and resources available for gene functional annotation such as the Enzyme Commission (EC) numbers and the KEGG pathways.
- GO terms can be used for genes with no experimental information. GO terms can be assigned to genes based on computational predictions or annotations derived from other sources, even in the absence of experimental data.
- GO terms are useful to detect what functions might be over-represented in a given dataset. GO terms can be used in enrichment analysis

- GO entries inform on the source of annotation. Each GO term has an associated evidence code that indicates the source of the annotation, such as experimental evidence, computational analysis, or curated information from DBs.

4. Regarding homology:
- paralogous genes are homologs
- all statements are correct
- orthologous genes are homologs
- homologs are genes that share a common ancestor

5. What is a promiscuous domain
- Domains that have many different functions
- Domains that can bind many other protein sequences
- all statements are correct
- Domains that can be present in many different protein families

6. Orthologs, as compared to homologs, are more
- likely to have complementary functions. While orthologs can have similar functions due to their shared evolutionary history, it is not guaranteed that they will have complementary functions.
- all options are correct
- appropriate to reconstruct a species tree
- similar in terms of their sequence. Orthologs can have similar sequences due to their common ancestry, but they can also undergo sequence divergence over time.

Orthologs are genes that diverged through a speciation event and are retained in different species. Their conservation across species makes them ideal for reconstructing species trees.

7. Select the incorrect statement regarding phylogenetic trees:
- the newick format is a graphical representation of phylogenetic trees with edges and nodes. The newick format is a textual representation
- bootstrap values provide information on how much support a give clade has from the analysis. Higher bootstrap values indicate stronger support for the corresponding clade, indicating that the clade is more likely to be accurate.
- in a phylogenetic tree all edges can be rotated without changing the topology
- for a given topology, there are more rooted trees than unrooted trees

8. Out-paralogs are…
- not necessarily encoded in the same genome
- gene that result from duplication
- more distantly divergent as compared to in-paralogs
- all the responses are correct

9. Associate the correct questions and answers:
- Alpha and beta globin are: paralogous
- Wings of bats and bird are: non-homologous, analogous
- Mammals are: monophyletic
- Winged animals are: polyphyletic

10. Phylogenetic reconstruction methods based on Bayesian analysis:
- Need a set of prior probabilities for the parameters of the model
- are generally the fastest among tree reconstruction models
- need to be run on a bootstrapped set of alignments to assess the support of the tree partitions
- 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:
- Chr painting: is an algorithm to detect large chromosomal rearrangements

- Sorting by reversals: is an algorithm to find the minimum number of rearrangements between two series of elements
- Gibbs sampling: is an algorithm to discover enriched motifs that are enriched in a set of sequences

12. Regarding the reconstruction of species trees:
- The more species you compare, the fewer conserved genes you can use in gene concatenation approach
- all existing methods rely on sequence alignments at some point
- the super-tree approach is the fastest and most accurate method
- genes that contain duplications cannot be used

As the number of species increases, the chances of finding genes conserved across all species decrease.

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

16. Adjacent genes are:
- Co-regulated. Adjacency does not always imply co-regulation.
- part of a gene cluster. Adjacency does not always imply clustering.
- neighbors to each other
- all answers are correct

Adjacent genes are indeed neighbors to each other. They are located in close proximity on the same chr, with no intervening genes or significant genomic distance between them

Which of the following entities would necessarily be considered a “gene” by the modern definition.

a) all statements are correct
b) a DNA region which is transcribed into a tRNA
c) a protein that has enzymatic activity
d) a RNA transcript that inhibits the function of other RNAs

Multiple choice (4 points)
Trimming
GO ontology
Mauve → gene inversions
Chr painting → Not specifically for gene inversions
Gene definition → any sequence of DNA or RNA that codes for a molecule that has a function
Hidden Markov Models → can't be provided by a single protein sequence, must be provided from MSA
Orthologs vs paralogs
Length branches

- Phylogram (sequence divergence = amout of change)
- Cladogram (nothing)
- Chronogram (time)

Part B) (3 points)
Tree reconstruction → Monophyletic/polyphyletic

Sorting by reversals (Burned Pancake problem)

Part C) (3 points)
Species-specific duplications

Longest duplication newick format (nodo con más duplicaciones)

Gene duplication and losses

Dot plot/Circos plot/Volcano plot

Session 1. Genes and their functions

Gene (modern gene definition): sequence of DNA or RNA which codes for a molecule that has a function. Both protein-coding and non-coding genes are genes because they **code** a function.

When talking about genes, DNA/RNA itself doesn't have a function. It codes for a protein that has a function.

- Promoters and enhancers have a regulatory function, thus they do not code for specific protein/RNA products (reason why they are not included).
- Pseudogenes are also excluded because their product has no function.
- ncRNA (tRNA, rRNA, etc) have already a function and they are not translated.

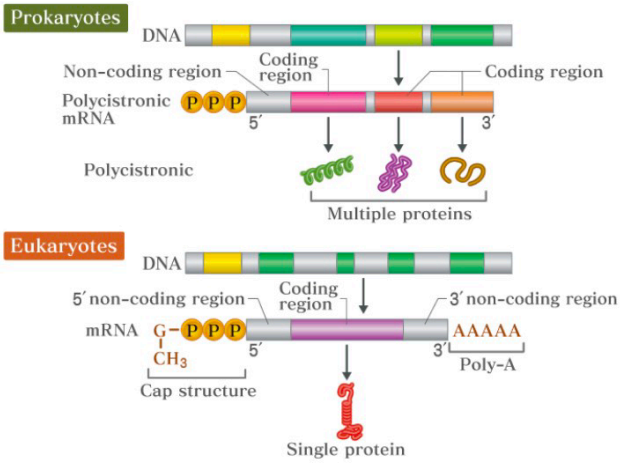
Gene structure and expression

Bacteria and archea → Cells without a nucleus

- Organized into operons: different coding regions next to each other, transcribed into a single mRNA and translated into different proteins.
- No single mRNA for each gene but a polycistronic mRNA that carries the information for more than one gene.
- This ensures that all the genes of a certain pathway are expressed at the same time

Animals, plants, fungi → Cells with a nucleus

- Genes are not in one continuous piece but interrupted by introns (will be removed during splicing).
- Transcripts are modified by adding a Poly-A tail and a Cap structure (distinguishes mRNA from being degraded).
- RNA goes into ribosomes (out of the nucleus) to code for a ncRNA or for a single protein.



If genes encode molecules that encode functions, the cell needs to decide when it needs a function to be or not be expressed

- Proteins: TFs (activators, repressors)
- Sequences: promoters, enhancers, TF binding sites

Genes must be expressed in order to synthesize proteins. Thus, they are activated when they are needed. This gives information about the function of the gene.
The expression of genes is tightly **controlled by** enhancers and silencers, which forces the DNA/RNA polymerase (TFs, etc) to bind or be removed from the DNA.

One gene may code for more than one protein and thus, for more than one function. This depends on the environment (time, tissue…), alternative splicing events, etc.

Why there are different sets of genes in different species?

Genes can be **duplicated** and accumulate mutations that provoque the formation of a new gene.
De novo origin of genes

- Process by which new genes evolve from DNA sequences that previously were not encoding for functional molecules. Genes can originate from nothing. We observe this making an alignment of the genome of closely related species.
- Possible factors: pervasive transcription (transcription occurring across the genome), transposition of promoters/enhancers ↑ expression, mutations originating promoters (↑ expression, extending ORFs, optimizing codon usage),

There can also be a **deletion/addition** of a nt that changes the ORF. Thus, the STOP codons are not readed and we obtain a larger or smaller gene.

Functional role of genes

- **Essential/Non-essential**: if you remove the gene, the cell will die
- **Constitutive**: always expressed

Different levels in which to talk about the function of a gene:

- **Molecular** function: refers to the actual roles of molecules.

- **Cellular** function: refers to the role you have in the cell. You can have the same molecular activity and have different cellular functions. For example: two different TFs can be found in different pathways.
- **Phenotype** function: refers to the effect on the morphology, physiology… of the organisms.

What determines the function of the gene?
- **Function** is determined by the **structure** of the molecule (protein or ncRNA) that is doing the function.
- **Structure** is determined by the **sequence** since each sequence has a propensity to fold in a given way.

The regions that are essential for the function of the protein are more conservated. The sequence may be different but the key structure remains the same.

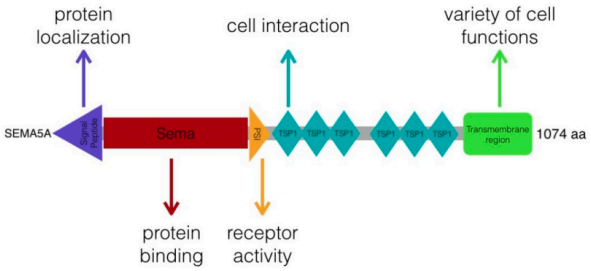
Homology based functional inference
Considers that the sequence determines the structure and the structure determines the function. By just having the sequence we can compare it to the sequence of another protein from which we know the structure (or function) and deduce the structure (or function).
Homology-based prediction is good at predicting molecular function, but not at higher levels (cellular and phenotypic functions).

Protein domains and domain shuffling
Proteins can have a modular structure, in which different parts of the protein do different things. The protein does not act as a whole. The different domains do different functions.
e.g.: a TF may have a domain that is devoted to the binding of DNA and another domain that attracts other proteins. This can lead to confusion when doing a BLAST (homology based predictions) because maybe the similarities of a protein is only restricted to a single domain.

Domain: conserved part of a protein sequence and structure that can evolve, function, and exist independently of the rest of the protein. Each domain forms a compact 3D structure and often can be independently stable and folded.
Promiscuous domains: domains that appear in diverse families in combination with other domains.
Resources: Pfam, SMART, InterPro

Domains can be described by HMMs, which specify the likelihood of finding a **given residue** in a **given** (relative) **position**. HMMs can derive from MSA and can be used to detect the presence of a given domain in a sequence.

Prediction of protein subcellular localization
There are domains (**motifs**) in the N-terminal that indicate the **subcellular localization**. They are signal peptides. Knowing where a protein is located gives hints on what function it can perform.



How to compare different functions on different organisms/compare genes of different species
- **KEEG**: DB for pathways
- **Enzymatic Commission** (EC) number: each number is a hierarchical description of the function of the enzyme.
- **Gene Ontology** (GO): system that describes functions and relates them with others.
  - It provides a controlled vocabulary of gene and gene product attributes
  - Annotate genes and gene products, and assimilate and disseminate annotation data
  - Provide tools for its easy access.

  It has 3 different ontologies (the study of '*being*'):
  - **Molecular** function: an elemental activity (e.g. transporter, enzyme)
  - **Biological** function: a commonly recognized series of events (e.g. cell division)
  - **Cellular** component: where a gene product is located (e.g. mitochondrial matrix)

Enrichment analysis identifies overrepresented biological terms or gene sets in a given dataset, aiding in functional interpretation. For this:

Fischer's exact test → Correction for multiple testing (FDR, Bonferroni)

Session 2. Comparative sequence analysis

Homology: organs in two species are homologous only if the same structure was present in their MRCA (regardless of their similarity). In other words, the same organ in different animals under every variety of form and function. e.g. forelimbs of bats, birds or dinosaurs (they have a CA).
Analogy: similar function but independent origin (do not derive from a MRCA). e.g. wings of bats, birds or dinosaurs (they don't have a CA).

Extension of homology concept to sequences: two sequences are homologous if they share a CA.
The problem with sequences is that we do not have a fossil record and thus, we can't get the sequence of the CA of birds and reptiles. Then, to check if two sequences are homologous:

- Perform an alignment
- Check if they are really similar and make inference
- Make an hypothesis that 'they come from the same ancestor'

But how similar must they be to consider them homologues?

- **Similarity**: degree of likeness between two sequences, usually expressed as a %, of similar or identical residues over a given length of the alignment. We can not say that *sequences are 50% homologous*, but *have a similarity of 50%*, because you can share a CA or not (there is no middle way ~ pregnant or not).
- Homology is **not** a quantitative measure

Are these two sequences **significantly** similar?
To answer this, we must check how likely it is that such alignment is the result of chance. Thus, we do not have to look at the % of similarity, but the E-value. The score of a BLAST is computed using substitution matrices like BLOSUM62.
If in one position there is a different Aa that has different properties, it will give a ↑ value.
This is more relevant than the identity because we are considering if the Aa have similar properties (not only if they are the same).

- **P-value**: probability of obtaining the same alignment by chance (from 0 to 1).
- **E-value**: expected number of alignments with this score or higher that you would expect by chance when comparing such sequence against a DB containing unrelated sequences (from 0 to a + number).

E-value depends on DB (important when searching in small DBs)

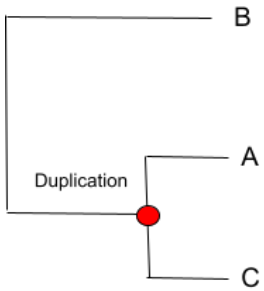
From homology to orthology
All homologous genes come from a CA, but there are two fundamental ways in which two genes can evolve from the CA:

- **Paralogy**: when homology is the result of **gene duplication**, so that both copies have descended side by side during the history of an organism, accumulating different mutations. It can occur between species or within species. e.g. α and β hemoglobin
- **Orthology**: when homology is the result of **speciation**, so that the history of the gene reflects the history of the species. There is a creation of a new species. Occurs always between different species.
  e.g. α hemoglobin in man and mouse

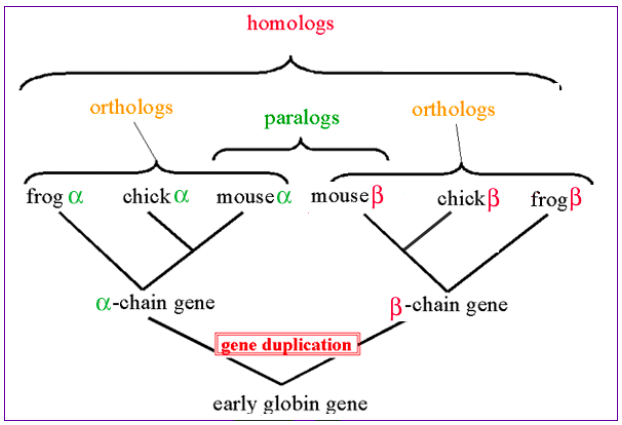
False senteces

1. Orthologs are homologous genes that have the same function
2. Orthologs are homologous genes in different species, while paralogs are homologous genes in the same species
3. The ortholog is the most similar sequence among the homologs in another species
4. If gene A is orthologous to gene B, and gene B is orthologous to gene C, then A and C are orthologous to each other.
5. Orthologs are genes that do not duplicate and, when they exist, they are always present in single copy
6. After a duplication, the orthologous copy is the one that keeps the function of the ancestral gene

Demonstration sentence 4



In this case, A is orthologous to B, B is orthologous to C, but A is not orthologous to C (they are paralogs).



Example

Hemoglobin belongs to a globin family. Initially it was a single gene that duplicated and formed two different chains (α and β) in all vertebrates.

Both chains evolved differently and they specialized in different functions.

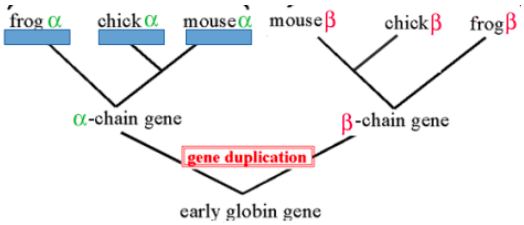
Important concepts

- Orthology definition is purely on evolutionary terms (not functional, not synteny…)
- There is **no limit** on the number of orthologs or paralogs that a given gene can have (when more than one ortholog exist, there is nothing such as the 'true ortholog')
- Many-to-many orthology relationships do exist (**co-orthology**)
- No limit on how ancient/recent is the ancestral relationship of orthologs and paralogs
- Orthology is **non-transitive** (as opposed to homology)

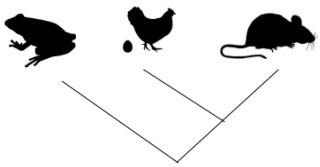
Why predicting orthology is important?

- **Important implications** for **phylogeny**: only sets of orthologous genes are expected to reflect the underlying species evolution (although there are many exceptions) because they show speciation events.

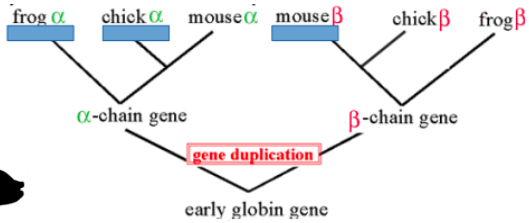
e.g. Imagine that we take one gen from frog, chicken and mouse. All three genes are orthologous to each other, in the α chain for example.



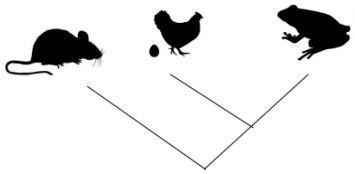
If we make a phylogenetic tree with the 3 sequences, we will retrieve this tree:



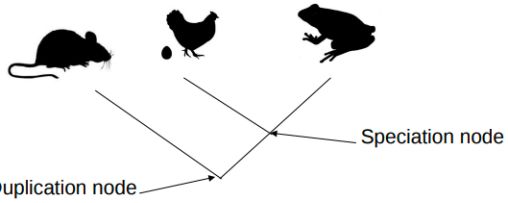
We would not obtain the same result if we use 3 homologous (not orthologous) genes.



We obtain this wrong species tree:



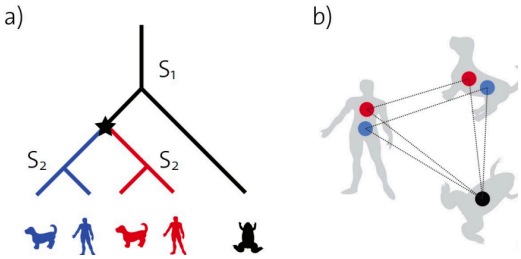
This is because this it not a species tree.



- The most **exact way** of **comparing two or more genomes** in terms of their gene content. Necessary to uncover how genomes evolve.

e.g. Imagine that we have this tree (human, dog, frog).
Regarding a). There is a speciation, a duplication and another speciation. Only by resolving the orthology relationships, we can see that these two human genes are equally related to the frog gene. The frog has not missed a gene.

Regarding b). Working at homology level, we would paint all the genes with the same color and therefore, not know which gene is orthologous to each other. Also, we would not know if there has been a deletion of a gene in frogs or a duplication in dogs and humans.

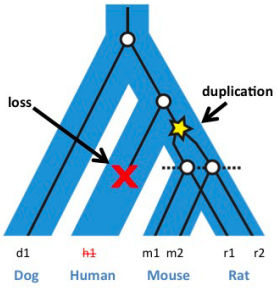


- **Implications** for **functional inference**: orthologs, as compared to paralogs, are more likely to share the same function

e.g. We have a gene that has 4 functions and suffers a duplication. Therefore, there is a redundancy in the genome (not good unless having more quantity of a gene is important). With time, three things can happen:
- Most common thing: one copy is lost (**degeneration/gene loss**). This copy incorporates a deleterious mutation that produces a non-functional gene, the gene is turned into a pseudogene and then, it is lost.
- **Neofunctionalization**: one copy adopts new function(s)
- **Subfunctionalization**: both genes co-exist and one does some of the functions while the other makes the rest ones. Thus, both genes need to be retained.

The process of paralogous retention has something to do with function changes. Then, it is clear that the process of gene mutation and duplication is often associated with processes of changing function. Thus, we must rely more in **orthologs** than in paralogs to **annotate the function**.
Note that orthologs can also change the function. An ortholog protein in a fish and in humans is likely to have a different function.

Gene families
- Group of genes that share a common ancestry (they are **homologs**).
- They have a hierarchical evolutionary relationship (best represented by a tree)
- Members of a gene family can be **orthologs** or **paralogs** between them
- An orthologous group is a (or part of) gene family
- Gene families evolve by **duplication** and **loss** (birth and death)
- Because of loss/duplication dynamics, gene families will vary in size and phylogenetic distribution.



e.g. More than 518 protein kinases are only in humans (they are involved in different pathways, so they do not have the exact same function, but similar).
- We can use homology base function prediction if two proteins come from the same gene family, because the function will be **similar** but **not the same**.

Orthology prediction methods
To predict **homology** we did alignments and check if these were significant by looking at the E-value.
To predict **orthology**, we use other methods.

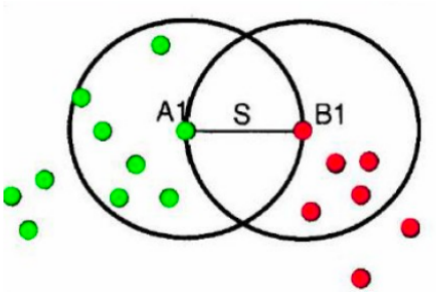
Making **phylogenetic inference** is the classical approach (manually manner)
- Build a gene tree
- Compare to the species tree
- Infer duplications and speciation events
- Assign orthology and paralogy relationships accordingly

Going genome-wide scale
Nowadays, we work with a ↑ number of genes, so we cannot use the classical approach. Everything must be done automatic and "blind" (not looking at the trees).

→ Similarity-based approaches
- **Best Bidirectional** (or **Reciprocal**) **Hits**: method for predicting orthology relationships. It detects 1-to-1 orthologies by identifying pairs of proteins that reciprocally match each other as the best hits. However, BRH is highly influenced by paralogy, which can lead to false negatives in certain cases. It is used only for closely related species where there are not many duplications.

- **InParanoid**: improved version of BRH that addresses some of its limitations: it identifies orthologs and considers in-paralogs. It handles many-to-many.
  - Starts searching for the BRH with a protein of interest.

- Then, it searches within its own genome for its own hit.
- Any hit that is closer to the initial hit than the initial hit to the protein of interest is considered a **paralog**.
- Genes inside the circle will be **in-paralogs*** and genes outside **out-paralogs***.
- It takes into account more complex evolutionary scenarios and can capture recent duplications. However, it can be computationally intensive and its accuracy may be influenced by the quality of the DBs used.



***In-paralogs**: derived from duplications that occurred after/subsequent to the speciation event (S → D)
***Out-paralogs**: derive from a duplication event that occurred before the speciation event (D → S)

- **COG**, **MCL-clustering approach**: we can perform multiple comparisons between different genomes and then build a network of BLAST hits.
  - Make a BLAST of all against all and then, build a network of these relationships where each node is a protein that belongs to different species and the edges are BLAST relationships.
  - The most modern methods have weighted relationships, depending on the E-value of the BLAST.
  - In the network, we can find orthologous groups (all genes derive from a single gene of a CA).
  - The term orthologous group* is confusing, because we cannot only find orthologs.
  - Also, we need to define how we go from networks to families (we can change the threshold).

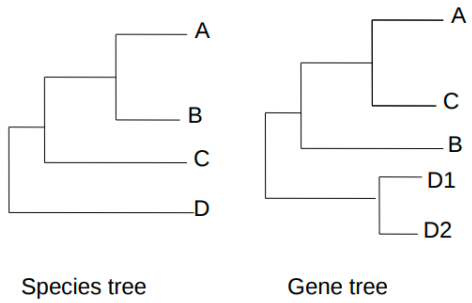
***Orthologous groups**: group of sequences derived from a single gene in a CA. they may include orthologs and in-paralogs. Each orthologous group has implicit the specification of an ancestral species of reference (speciation node).

This 3 methods use BLAST and therefore, they are really fast.

→ Phylogeny-based methods
Methods based on phylogeny were not used at a large scale due to limitations in computational power (phylogenetics is costly). However these has changed due to new algorithms.

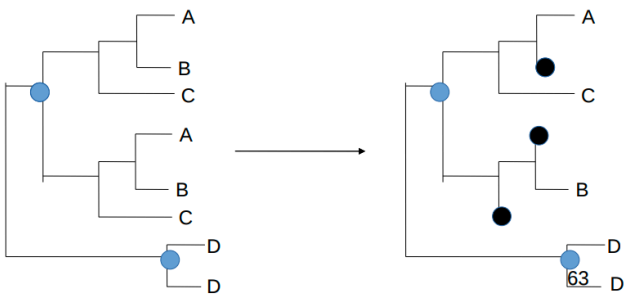
They reconstruct the evolution of a gene family (phylogenetics), detect duplication and speciation nodes and predict orthology and paralogy accordingly.
The two main methods for predicting duplication and speciation nodes from a tree are:

- **Species tree reconciliation** (RIO, Ensemble).
  - **Hard reconciliation**: we want to reconcile every node. It resolves any incongruence between gene tree and species tree by introducing the minimal number of gene duplications and losses.
  - **Soft reconciliation**: allows incongruences below a given support value. Only a gene tree is **congruent** when it has the same topology as the species tree.



We have a species tree and a gene tree. Are they congruent? No, because in gene tree, A is closer to C than to B.

So, this method will reconcile the species tree and gene tree by adding the minimal number of duplications or losses.



Reconciliation with the species tree provides information on speciation and duplication nodes in a tree. It only works when:

- We know the **true species tree**.
- The gene tree is **correct** and **reflects** the species evolution, meaning that genes are only inherited vertically.

How often do we have the true species tree and the correct gene tree reflects the species evolution? There is a degree of uncertainty and we cannot reliably use reconciliation because of topological variability (phylogenetic artifacts, insuficient phylogenetic signal, lineage sorting, gene conversion, etc).

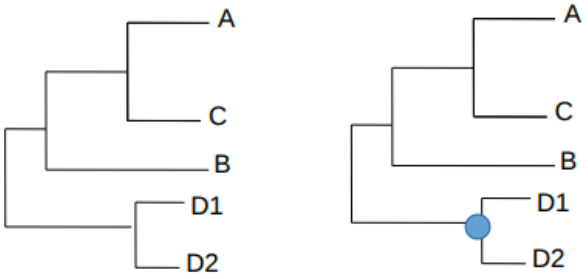
To deal with this topological variability we implemented a species-overlap algorithm.

- **Species-overlap algorithm**

  We simply explore the tree and for every partition in the tree, we ask the question: *Is there a species overlap?* If there is no species overlap, we put a speciation. Otherwise, we put a duplication.

  It does not require a species tree, but needs to know the species to which the genes belong. In essence, can be seen as a reconciliation with an unresolved species tree.

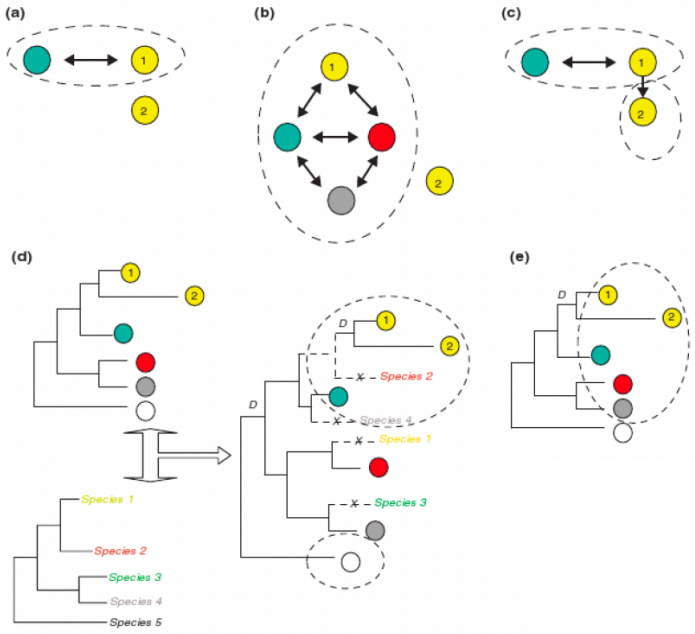
In the last example, we would obtain (note that we do not use a species tree):



When comparing these two methods, we obtain similar prediction values, meaning that they both say that it's orthologous or not. But the species overlap algorithm has a higher sensitivity.

Summary

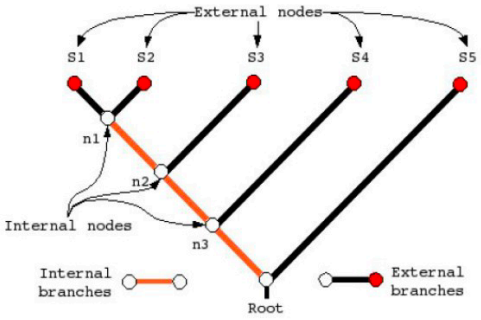
a) Best bidirectional hits
b) COG, MCL-clustering approach
c) InParanoid
d) Tree reconciliation
e) Species-overlap (PhylomeDB)



Session 3. Phylogenetic analysis

Phylogenetic tree: branching diagram (bipartite graph) showing the inferred evolutionary relationships among various biological species or other entities (e.g. sequences) based on similarities and differences in their physical and/or genetic characteristics.

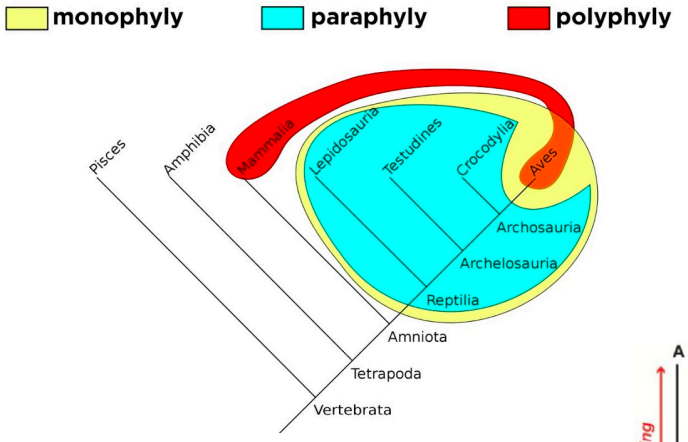
- **Species tree**: shows the evolutionary relationships between the different species.
- **Gene tree**: represents evolutionary relationships between genes. In this case, nodes can represent duplications or speciations.

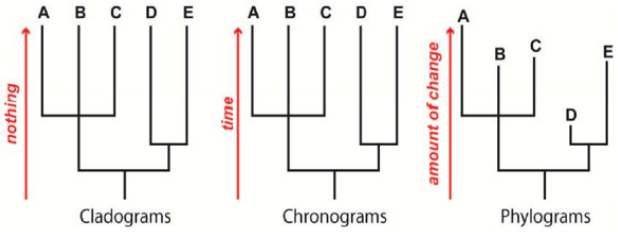


Trees contain internal and external nodes, and branches.

- **External nodes** are sequences representing genes, populations or species.
- **Internal nodes** can contain the ancestral information of the clustered species.
- **Branch** defines the relationship between sequences in terms of descent and ancestry.

- **Monophyletic group**: group of organisms that includes an ancestral species and all of its descendants, but **no other organisms**. e.g. Birds, reptiles
- **Paraphyletic group**: group of organisms that includes an ancestral species and some, but **not all**, of its descendants. e.g. Modern reptiles (excluding birds)
- **Polyphyletic group**: group of organisms that includes multiple evolutionary lineages that do **not** share a CA. e.g. Warm-blood animals (mammals and birds) and cold-blooded animals (reptiles and fish)





Trees

The **topology** is the branching structure of the tree.
The **length** of the **horizontal** axis is not relevant.
The **length** of the **vertical** axis can mean different things:

- In **cladograms**, it does not mean anything because there is a constant mutation rate.
- In **chronograms** it represents time
- In **phylograms** it represents the amount of change (sequence divergence)

We can represent trees in vertical or horizontal mode, or in different shapes. If we rotate branches, the tree remains the same but with another configuration.

To represent a tree, we can use the **Newick format** (textual representation of phylogenetic trees with edges and nodes that explains the hierarchical relationships).
Trees can be rooted or unrooted. There are several strategies that can be used to root a tree (e.g. midpoint, outgroup). For every topology there are always **more rooted** trees **than unrooted** trees.

Number of rooted trees with n OTUs: $(2n-3)!! = \frac{(2n-3)!}{2^{n-2}(n-2)!}$ Number of unrooted trees with n OTUs: $(2n-5)!! = \frac{(2n-5)!}{2^{n-3}(n-3)!}$

How do we reconstruct a phylogenetic tree?

Phylogenetic trees can be based on anything that can tell us similarities and differences.
Nowadays, molecular phylogenetics is the most widely used method because we can obtain sequences of different species very easily. Then, these sequences are compared using alignments.

To find the best tree, we can use two approaches:

- **Exhaustive search**: make all trees first and then see which one best fits the data. This is not possible for a large number of sequences.
- **Heuristic search**: try to find a way to find an optimal tree (hopefully the best) without testing them all. We also need an optimality criterion and we are not guaranteed to find the best tree, but we save time.

Phylogenetic approaches

- **Distance-based methods** are the fastest but they are not the best methods when obtaining the correct tree. If there are no errors, the correct tree can be obtained in polynomial time. Otherwise, optimization problems are NP-hard.
- **Maximum Parsimony** and **Probabilistic methods** are NP-hard.

Bootstrapping

The numbers that appear in a branch of a tree are support values that are computed by a bootstrap. It does a sampling with replacement, so that it mixts the order of the sequence to check the **robustness** of the tree.

We will obtain a tree for each alignment and then we can make a consensus tree. Each value of a branch represents the % of times that branch appeared in other trees. Normally, ↓ supported node correspond to short branches (there has been a short number of changes).

Maximum parsimony: given a number of sequences, it builds a tree minimizing the number of mutations for each position of the sequence.
→ *Problem*: when comparing sequences that are really far away in evolution, the number of changes is not representative to the number of mutations. This is due to the fact that there is a saturation. Also, because we have to evaluate many trees to evaluate which is the best one, it is not used often.

Neighbour Joining
We start with some sequences from which we do not know the relationship between them. Then, we compare all sequences to each other to obtain the pairwise distances. We find which is the closest pair and make a cluster. We repeat for the other remaining sequences.
This method is very fast because we only need pairwise comparisons. The problem is that it is also affected by the saturation of mutations when comparing sequences that are really far from each other. But, because of convergent evolution, this method is used a lot.

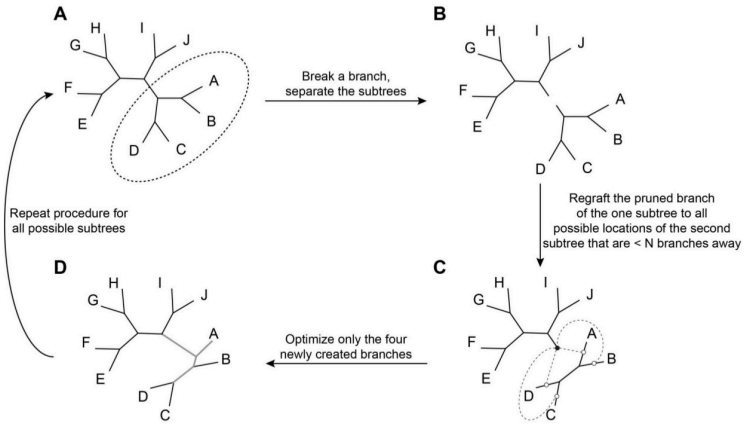
Statistical methods: try to work with probabilities of observing one tree.
→ **Maximum Likelihood**
Computes the probability of observing the data given an hypothesis. In our case, data is the alignment and hypothesis is the tree. So, we are computing the probability that a tree computes a certain alignment.
- **D**: alignment
- **H**: model of evolution, tree topology, branch lengths, parameters of the model
- Each H will have a certain probability of producing the data **P(D|H)**. But best H will be that of greatest P

Likelihood function is **not** the probability of an hypothesis being correct. Thus, the probability of observing the data has nothing to do with the probability that the underlying model is correct.

So far we have assumed one evolutionary rate for all the sites. But we know that different parts of a protein evolve at different rates. So, to make our model more realistic we can use different probabilities of change in different regions of the alignment. e.g. Approximation of rate-heterogenicity by a γ-distribution

Heuristic search
We know that there are millions of possible trees and each of them has a likelihood.
Thus, we can imagine a space of possible trees where some of them are more likely than others.

We can explore this space of solutions with an heuristic method, starting at a random place, or the first operation can be a NJ and then use an heuristic method.



→ **Bayesian analysis**
Tries to compute the probability of the tree given the alignment. For this, first the **posterior probability** (likelihood*prior probability) is computed. To do this, we need to know the prior information (probability of knowing the tree before looking at the alignment).
- *Problem*: we do not know the probability of knowing the tree before looking at the alignment.
- *Solution*: prior that are flat distributions (uninformative), meaning that all trees have equal probability (thus, the result is not influenced).

How it works
We need a method to compute the **posterior distribution** and then, use **MCMC** to look at the space of possibilities. In the case of ML, we were navigating until we found the ML point.

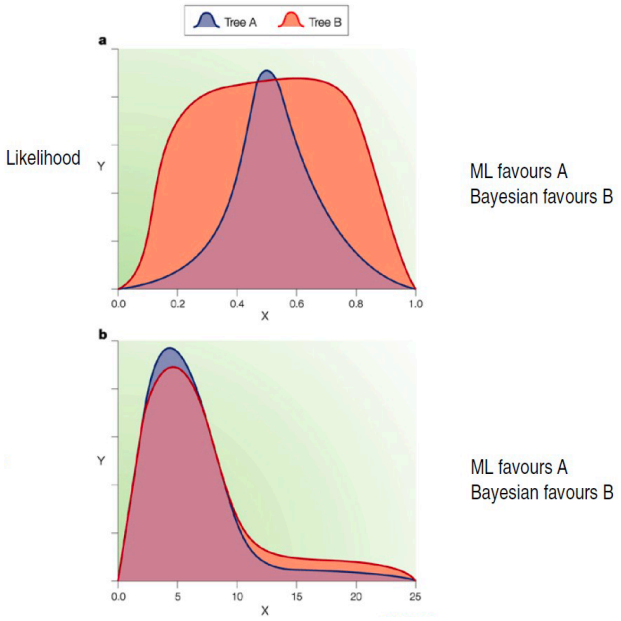
In Bayesian approach, the robot **computes** and **stores** tree **likelihood** in its backpack at each step. At the end, Bayesian approach will have a sampling and therefore, it will have more trees in the picks of the mountain because the algorithm stays longer in the ↑ parts.

Trees will be sampled in proportion to the posterior probability (trees with ↑ probability will be sampled more times).

In Bayesian approach, no bootstrap is needed as 100 000 trees provide support for each tree partition. But, determining when to stop is challenging. Multiple robots explore solutions randomly. When backpacks become similar, signifying full exploration, we can stop the process. Initial trees are also discarded (burn-in).

ML picks the blue tree because it has the highest peak. But, the Bayesian approach chooses the orange one because it has been sampled more times than the blue one. Orange tree is better (BI).

In the second case, blue tree is much better (ML).



ML uses **Joint estimation**. It finds the highest point in the parameter landscape

Bayesian analyses measure the volume under the posterior probability surface, the parameters are integrated (marginalized) to obtain the marginal posterior probability of a tree (**Marginal estimation**)

When there are **few parameters** and **plenty** of **data**, then ML and Bayesian inference will **agree**.
When we have **few data** and a **lot** of **parameters**, then the **BI** is more reliable.

The good thing about this probabilistic-based methods is that we can use **statistical tests** to determine how **more likely** a topology/branch length is with respect to another.

We can compute the Akaike Information Criterion (AIC) to assess which topology is more likely. **AIC** is an estimator of the relative quality of statistical models for a given set of data. Given a set of a model, we can use AIC to choose the one with the minimum value. As we can see, if we have a large number of parameters we will have a ↑ AIC (it penalizes a large number of parameters)

Number of model parameters

Maximum likelihood value

$$\mathrm{AIC} = 2k - 2\ln(\hat{L})$$

This all methods mentioned before are to make a tree. Now, algorithms such as reconciliation algorithm and the species-overlap approach, would be use to infer evolutionary events on the created tree.

Also, we may want to <u>infer relative timing of speciations and duplications</u> (put time on branches of the tree) because normally the branch length represents the number of changes and, not the time.

- We can assume a **molecular clock**. It is said to exist when substitutions accumulate linearly with time. This assumption is violated most of the time, particularly at long evolutionary distances.
- Different species have different μs. For a **constant μ**, neutral substitutions are expected to behave more clock-like than non-neutral substitutions, that is why they are generally used as a proxy for time.
- Other kinds of evidence such as **fossil records** imply some uncertainties.

S substitutions are still used as a proxy. However, variations in GC content, local variations in μs, or repair can introduce **noise**.

Non-synonymous/missense substitutions: result in a change of the encoded Aa in a protein.
Synonymous/silent substitutions: do not change the Aa sequence due to the redundancy of the genetic code.

Radical substitutions: Aa changes in a protein sequence that result in a significant alteration in the chemical properties of the substituted Aa. Indicative of functional shifts, new functions, clades.
Conservative substitutions: involve Aa changes that retain similar chemical properties as the original Aa.

We can use NS mutations to know about functions of proteins. dN/dS is the ratio of NS and S substitutions. It is useful to measure the strength of natural selection acting on protein-coding genes.

- If gene is evolving **neutrally**, we expect a value of **1** (same chance of having a S or NS mutation) e.g. pseudogene, where it does not care about the sequence.
- If gene is **important**, we expect a value **< 1** because it does not accept NS mutations (purifying selection).
- If there is **positive selection**, we expect a lot of NS mutations and a value **> 1**. This is due to the fact that it needed to change of function.

Session 4. Phylogenomics

Intersection between genomics and evolution. In other words, looking at genomes from an evolutionary perspective, using **phylogenetics**. Distinction between phylogenetics/phylogenomics lies in the scale:

- When talking about a single gene or protein in **one or few** species, we are in the **phylogenetics** realm
- When talking about a single gene or protein in **all** species, we are in the **phylogenomics** realm

Phylogenomics is necessary to provide an evolutionary framework to the amount of data generated. It is useful to obtain biological knowledge from sequence data. The **more data**, the **more powerful**.
But, it is computationally demanding. It needs to be automated and proper scalable (e.g. alignments, form a certain point, the **more data**, the **more noisy**).

Phylogenomics to reconstruct species trees

- **Species tree**: represents relationships between **different species**.
- **Gene tree**: represents evolutionary relationships **between genes**.

Most of the species trees are based on molecular data, so there is a relationship between both trees.

How can we use molecular data to reconstruct a species tree?

At the beginning, methods only used information from sequences, without making any alignment. We don't use them any more because they suffer from the effects of convergence (similar traits between distantly related species).

The alternative methods are **sequence-based methods** (use information of the sequence of genes).
There are two main methods (both start by **aligning homologous** genes present in **different species**):

- **Supermatrix**: concatenating alignments of different genes to make a **single tree**. This is expected to better represent the evolution of species than a tree made by a single gene. By concatenating **more residues** into a longer alignment, **precision** in resolving difficult positions of the tree can **improve** (more genes, more signal, noise cancels out). The **more species** we use, the **fewer genes** we can find **shared** between all species.
- **Supertree**: from each alignment of a gene, we make a tree. Then, all the trees are converged to a **consensus tree** (multiple methods to converge trees).

Results obtained in both methods are similar, but the branch lengths change a bit.

Genome-wide phylogenetic analysis (phylome)

We are now interested in studying gene trees.
For each of the n genes in a genome, we examine its evolutionary history using a phylogenetic tree that shows its evolution and homologous relationships across species.

If we do this for all genes, we will obtain a phylome.

- **Phylome**: complete collection of evolutionary histories of all genes encoded in a given genome.

There are two main approaches:

- **Family-based approach**: first build families (orthologous groups), then reconstruct one tree per family
  - Build families (**Ensembl**)
  - Make an alignment per family
  - Phylogenetic reconstruction per alignment
- **Gene-based approach**: sequentially use each gene of interest as a seed to build a gene tree (**PhylomeDB**):
  - Search for homologs
  - Make an alignment per gene + homologs
  - Phylogenetic reconstruction per alignment.

  In this case we will have to make many more trees.

Pipeline to obtain a phylome

Homologs search → MSAs → trimAI → NJ tree → ML trees → MrBayes Tree → Phylome

MSAs to TrimmAI procedure

We can compare three aligners in both forward and reverse modes. However, due to alignment heuristics, we may observe differences in the results. These differences can accumulate due to variations in gap placement (either to the right or left) during the alignment process.

Using the results from the three aligners, we create a consensus alignment. By comparing paired residues from each alignment, we identify stable columns and trim unreliable/noisy parts of the alignment.

These phylomes provide the following information:

- Families that show a particular topology
- Let detect and date duplication events
- Genes that have accelerated evolutionary rates at a particular lineage (because of an adaptation)
- Families expanded at particular lineages
- Footprints of HGT, lineage sorting, gene conversion and other evolutionary processes
- Search for co-evolving genes
- Predict functional properties
- Across-species prediction of orthology and paralogy

There is uncertainty in species trees and topological variability in gene trees.
We can use gene trees from a human phylome to support a topology. But this is not accurate at all.

Possible sources of the incongruence between gene tree and species tree:

- **Analytical factors**: lead to failure in accurately inferring a gene tree (we obtain a wrong gene tree).
  - **Stochastic** error: insufficient sequence length or taxon samples, noise. They are random
  - **Systematic** error: observed data is far depart from model assumptions. These are the most dangerous ones since the model explains poorly the data. Therefore, we will always have that error.

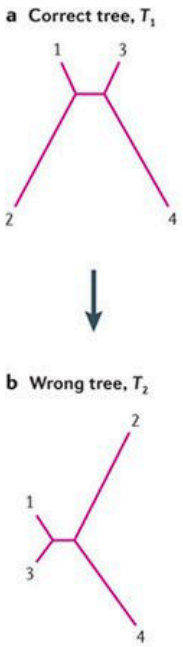
  Problems regarding the methodology or data used.
- **Biological factors**: lead to gene trees that are topologically distinct from each other and from the species tree. Known factors: stochastic lineage sorting, hidden paralogy, HGT, recombination, natural selection.
  Actual biological processes that may result in true incongruences between the gene and species trees.

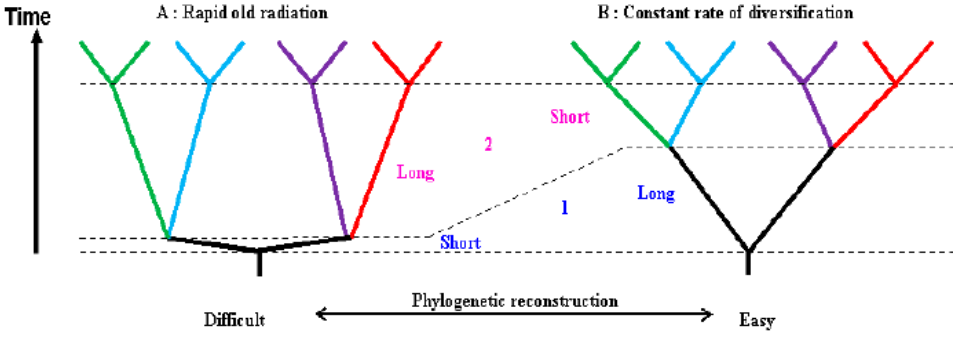
Analytical factors

Most common case is when there is "***fast radiation***", meaning that there is a short internode.
Species that result from that branch will be really similar because there has not been enough time to accumulate enough mutations. Therefore, it will be really **difficult** to obtain the right topology because all species (green, blue, purple and red) can descend from any of the branches.
This also happen when analyzing short sequences because we have even less information.

e.g. Systematic error could be "***long branch attraction artifact***".
When we have a dataset where sequences are close to each other, and other sequences are far from all the rest.

- We will obtain two types of trees and we won't be able to distinguish which one is the correct one.
- Two species that should be far away are put together to minimize the total length of the gene tree.





Biological factors

- **Incomplete Lineage Sorting** (**ILS**)

The speciation process happens at the population level. Each population is genetically diverse, meaning that there are different alleles in the population. Because of that, if there are two consecutive speciation events in a short amount of time, it may happen that some of the alleles distribute in the resulting species in a way that is different to the species tree.

- **Hidden paralogy** caused by **differential gene loss**, following **duplication**

In this case, a gene duplicates. So, all the species have two paralogs for a long time and suddenly, they all lose one of the paralogs (maybe one or the other). Therefore, the phylogeny we will reconstruct from these genes will be different.

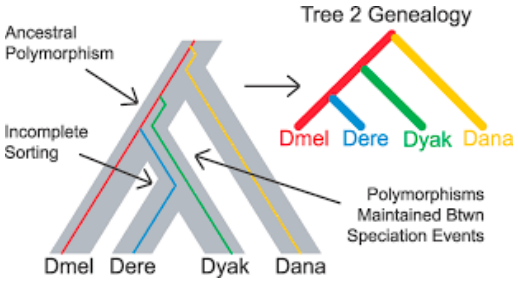
- **Positive selection** can cause tree patterns different from the gene tree by **convergent mutations**.

This is very rare. Two genes appear together in the gene tree because they are really similar to each other because of convergent evolution. e.g. The gene responsible for echolocation is present in bats and dolphins. So, we will find they are close and in the species tree they are really far.

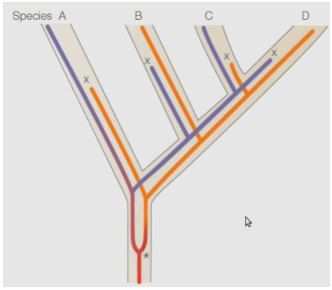
- **Non-vertical evolution**

We represent trees vertically, meaning that there are parents that create childs. But what happens when there are genes that go from one species to another, or lineages that mix? Evolution makes small jumps (*quantum leaps*), meaning there are gene duplications, symbiosis, hybridization and HGT.

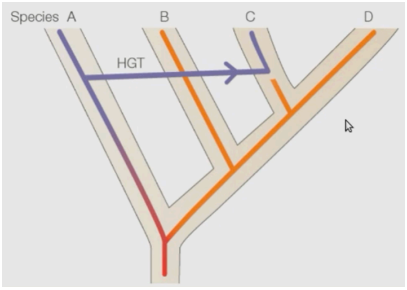
→ **HGT**. Process by which a gene is transferred from an organism of a species to another one from a different species. Maybe this transfer substitutes another gene, so if we make a tree, we will obtain a topology that is different to the species tree.



ILS



Hidden paralogy



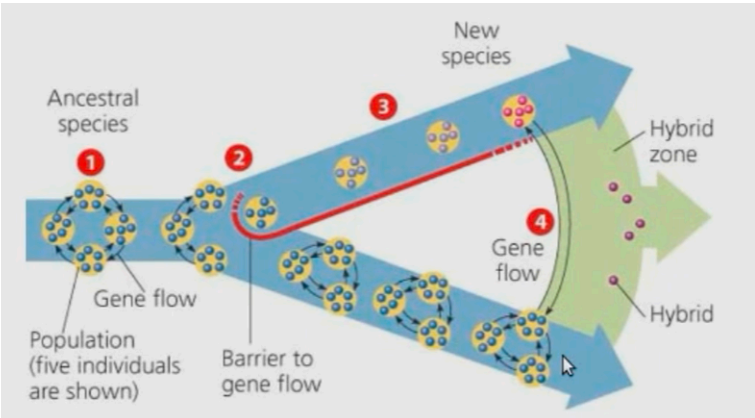
Non-vertical evolution

In this case, A will be next to C, and B next to D (which is wrong).

→ **Hybridization**: it can be seen as a massive HGT of the whole genome. In a populations perspective, we can see the hybridization as:

- A population with gene flow experiences a barrier, leading to population separation and divergence. During this period, populations evolve independently. When the barrier disappears, gene flow resumes, resulting in hybridization between the previously separated populations, forming the hybrids.

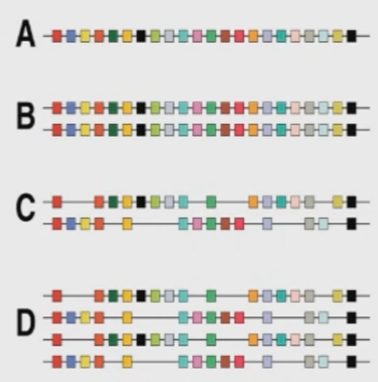
Hybridization can lead to **networks** rather than trees, because in the hybridization we are joining species. When two species mix, the two genes that were orthologs in different species become paralogs (we see two homologs in the same genome). But they are not paralogous genes because they speciated. They will be **homeologs**: set of genes that were originated by speciation and then joined back together in the same genome due to hybridization.



- **Whole-genome duplications**

Hox cluster: group of genes that regulates the development process in vertebrates. These genes are situated together in different chrs. These blocks sometimes had similar arrangements between genes that are more similar to each other.

For example, red genes are always in the first position… We proposed that this comes from two rounds of whole genome duplications. So, we have an original Hox cluster (A) that suffers a duplication. Then some of the paralogs are lost, and there is another duplication.



Session 5. Genome comparisons and evolution of gene order

In pre-sequencing times, a way to compare two genomes (e.g. DNA-DNA hybridization) consists of:

- Obtain both DNA
- Heat to denaturalize (separate strands)
- Combine ssDNA at ↓ temperature
- Cool to allow renaturation of dsDNA
- Determine the degree of hybridization

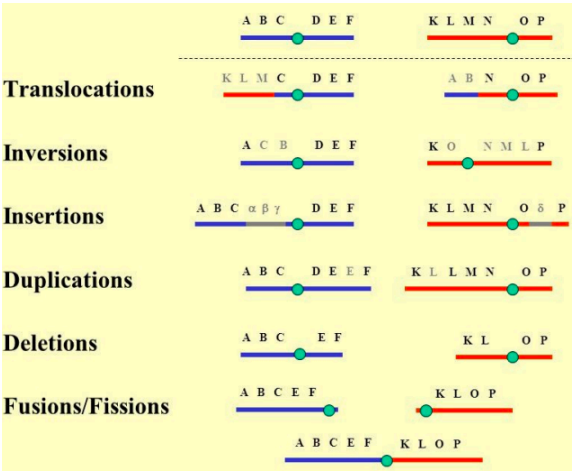
If they are similar, they will remain together longer. A ↑ temperature implies greater genomic similarity.

It was also used electrophoresis (separates entire chrs) for visualizing genome re-arrengements. The smaller chrs run faster in the gel.

Using this methodologies, it was discovered that species do not only diverge by accumulating changes like point mutations in the DNA, but also by the **genomic re-arrangements**.

Types of chrs re-arrangements

Mutational changes in the genome that range from few 100 bp to several Mb, that cause structural variations in the genome.



Comparing genomes in the sequencing era by comparing nt frequencies of two different species:

- Also dint, trint, k-nt frequencies (k-mers).

  k-mer: string of k nts, of length k. The longer the k-mer, the more specific we are.

  We can see how many k-mers we have in both genomes… If the k-mer content is similar, we will suppose that they belong to close related species. There are methods that do this automatically, such as Self Organizing Maps (SOM), used in metagenomic analysis.

  They will try to put in a space closer to each other sequences that have similar tetrant frequencies, and far away sequences that have different tetrant frequencies.

- Compare GC content (also gives information about AT content)

The good thing about this method is that we do not need an alignment, we are just counting and therefore the algorithm is really fast.

To compare genomes, we can also do alignments:

- **Partial alignments**: we break the two genomes into pieces, we make a BLAST of each of these pieces to the other genome, and then make a reciprocal hit. If our piece aligns with an identity >30% and >70% length of the query ORF, then we count it as a hit.

We will store the % of identity and then we will compute the average nucleotide identity (ANI) over the whole genome (we will see that on average 60% of the genome is identical).

- **Whole genome alignments**

All previous methods based on k-mer frequencies do **not** capture types of re-arrangements.

Also, standard sequence alignment algorithms (e.g. Needleman-Wunsch, Smith-Waterman) do not work, as they do **not** handle re-arrangements (because an alignment expects the things to go in a **certain order** and if we have a translocation, it does not know how to proceed) and such large sequences (computationally).

→ **Solution**: align bits and pieces while being aware of the relative **coordinates** of the pieces. They try to find anchors (small regions that align well) and then they try to extend them. Finally, they look at how these different anchors are related to each other.

Whole-genome alignment algorithms: LASTZ, BLAT, Mugsy, megaBLAST, MUMmer, LAGAN, M-LAGAN
→ LAGAN procedure: starts by finding matches between two genomes (diagonal lines indicate similar regions between them). Then, it tries to compute an optimal path that takes the least distance across the box. Finally, it realigns the similar areas using more sophisticated algorithms.
→ Mauve procedure: detects blocks that are homologous across the different genomes. Then tries to find consecutive homologous blocks to form single blocks. Finally, infers re-arrangements.

<u>Which genomes should we align?</u>
For reasonable analysis, genomes should:
- Derive from a sufficient recent CA, so that homologous regions can be identified
- Derive from a sufficiently distant CA, so that sufficiently "interesting" changes are likely to have occurred

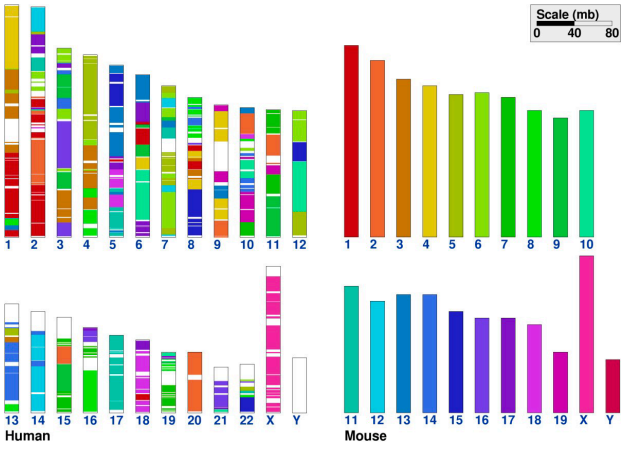
There are many genome browsers (e.g. Ensembl/UCSC) that provide pre-computed alignments with closely-related species.

Motifs: short conserved patterns in DNA/RNA sequences that indicate functional elements.
Gibbs sampling: algorithm to discover enriched/conserved motifs in a set of sequences. Given a window of sequences from different species, it looks for motifs that are over-represented in a given region, not being necesarilly aligned (not in the same position).

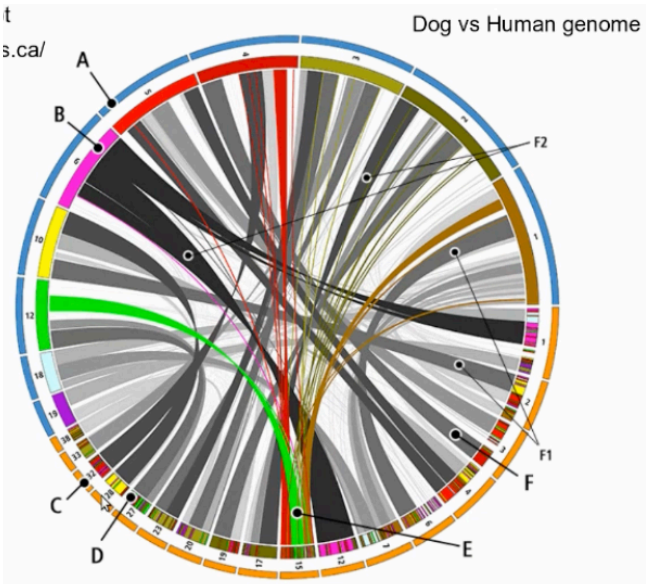
<u>How to represent comparisons of genomes?</u>
- **Chr painting**: algorithm to detect large chromosomal re-arrangements, to check if the location of genes is conserved.

e.g. We pick a color for each mouse chr and paint it in the human chr, and compare the re-arrangements.
We can see that the chr Y does not have homologous regions at the threshold that they use. This is because it evolves much faster than the other chrs since it cannot be repaired by recombination. The X chr is highly conserved because it never recombines with other chrs.

- **Circos plot**: similar approach as chr painting. Cannot only display chromosomal re-arrangements.



Chromosome painting



(top, blue [A]) and dog (bottom, orange [C]) chromosomes. One dimensional similarity mapping between human [B] and dog [D] chromosomes. This mapping provides the chromosome color coding associated with grey ribbons [F]. These grey ribbons are composed of binned synteny regions that fall in the same bundle (see above). The level of grey is proportional to the size of the syntenic regions. Synteny on chromosome 15 is highlighted with colored ribbons [E]. Ribbons that twist such as [F2] indicate inversions, whereas those that don't [F1] indicate regions of synteny on the same strand. (zoom).

- **Mauve**: to visualize re-arrengements between multiple genomes. Mostly used for bacteria since they have samll genomes. Mainly for inversions.

Synteny

In **classical genetics**, synteny describes the physical co-localization of genetic loci on the same chr within an individual or species.

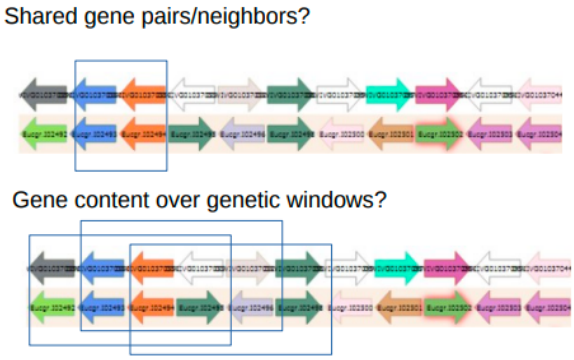
In **comparative genomics**, synteny (or shared synteny) refers to the conservation of block order within two sets of chrs that are being compared with each other.

In other words, conservation of DNA blocks which are close in the chr and in the same relative order. The gene has stayed in the same relative position in the chr, so the neighbors around these gene are the same.

How can we measure gene order conservation? How can we compare the gene order in species A and B? How can we determine how distant they are in terms of synteny?

With nt sequences it was easy, because we only needed to count substitutions, % of identity… But in this case, what should we do?

- We can **count shared gene pairs**
- We can create windows of a certain # of genes and see the number of **genes** that are **shared between** the **two windows**.



Other algorithms:

Sorting by reversals or Pancake flipping problem

- Imagine that you have a pile of pancakes and each pancake has a different size. We want to make a pile like a pyramid (gene **order**).
- To do this, we can only insert a spatula in any place and flip the pile.
- This algorithm obtains the pyramid with the smallest number of movements.
- We also have the **Burned pancake flipping problem**, where the **orientation** matters. This is really similar to genome re-arrangements.
- We can measure how many steps away is one genome from the other. In other words, compute the distance between both genomes by counting the number of steps that are needed.

Dot plot

Method to compare genomes, one against another (represented in the X and Y axis).

Each dot can be a fraction of the genome or a gene. It is a graphical method for comparing two biological sequences and identifying regions of close similarity after a sequence alignment. Each axis is a sequence.

→ Lines mean regions of similarity.

Genes that are physically close to each other in the genome tend to be regulated in a coordinated manner. This can be observed in operons in prokaryotes and the spatial organization of genes in euchromatin regions of eukaryotic chromosomes. Deviations from this trend often indicate **selective pressures** and highlight the importance of gene regulation in maintaining functional coherence.

How can we use gene order to predict function?

If we find genes that are kept together through evolution, we can make an hypothesis saying that '***they are involved in the same pathway or biological process***'. All this is very different from what we were discussing when talking about homology prediction since, in this case genes are not doing the same function, but they are **cooperating** in the same process.

Take into consideration that it is of great help knowing the function of **at least one** gene, because then we can make inferences and discover the function of the other genes. Otherwise, we only know that they are **related** but we don't know what they do.

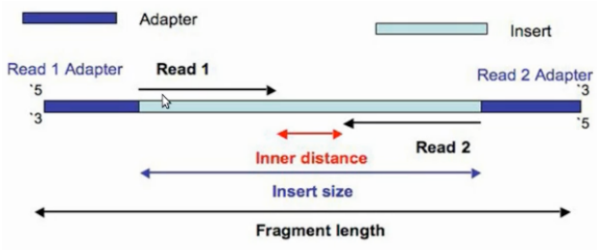
There is one extreme case in which the genes fusionate.

e.g. Tryptophan synthase has two subunits but in yeasts there has been a fusion and thus, there is a single unit. This is an even stronger indication that both genes are related somehow and moreover, they are likely to be physically interacting.

Gene fusion: joining of two separate genes to form a single functional gene.
Gene fission: splitting of a single gene into two or more separate genes, each with its own functionality.

Searching for genes with conserved synteny across many genomes for functional prediction:
STRING: DB of known and predicted protein interactions. These include direct (physical) and indirect (functional) associations; and they are derived from 4 sources

- Genomic context
- ↑-throughput experiments
- (Conserved) co-expression
- Previous knowledge



Detecting genome rearrangements with NGS

We want to detect genomic re-arrangements not only across species, but also within species.
e.g. Cancer genomes have a lot of re-arrangements. So, it is interesting to be able to detect them. How can we do this? We can use sequencing approaches. We just expect a certain distance between reads and if there is a smaller distance, then there is a deletion…

Session 6. Phylogenetic profiling and co-evolution

Co-evolution: occurs when two or more species reciprocally affect each other's evolution through the process of natural selection.
Phylogenetic profile: describes the presence or absence of a protein in a set of genomes. Similarity between profiles is an indicator of functional coupling between gene products.

Why is it useful to compare the gene content of 2 genomes (what genes has genome A and what genes has genome B)? If we compare two genomes we expect them to share the genes that are responsible for making them share the same phenotype/characteristics. The genes that are unique for each of the two genomes may indicate functions, characteristics, traits… that are specific to each species.

Can we extend this to more than two genomes? Yes. But we will need to detect homology or, even better, orthology between the genomes.

Detecting homology involves identifying similarities in DNA or protein sequences between different organisms or genes. It indicates a shared evolutionary origin or common ancestry.
Detecting orthology refers to identifying genes that originated from a common ancestral gene through a speciation event, indicating their shared function across different species.
Orthologous genes retain similar functions, while **homologous** genes may have diverged in function.

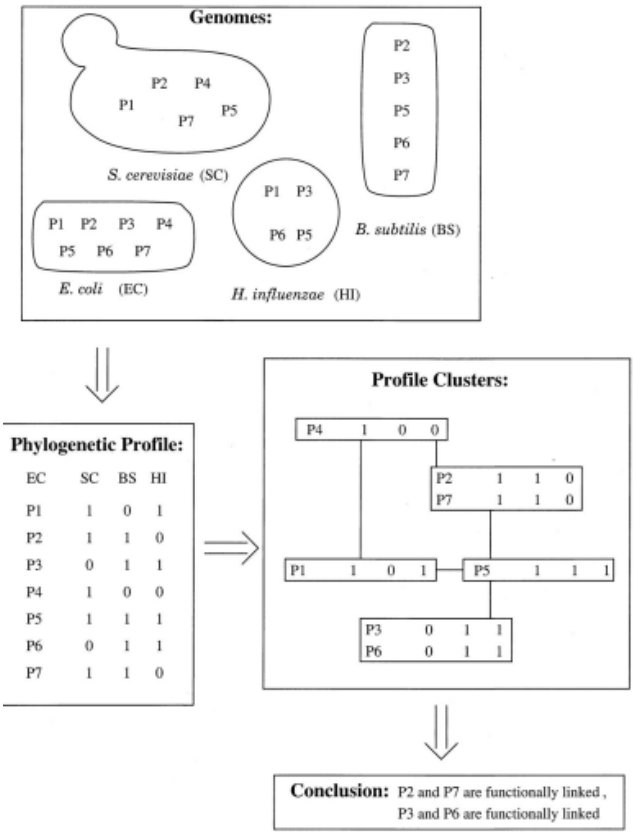
We have 4 different genomes and we are investigating 7 different proteins that can be shared between the species or not.

We can build a matrix that stores a 1 if the protein is codified in the genome or a 0 if not.

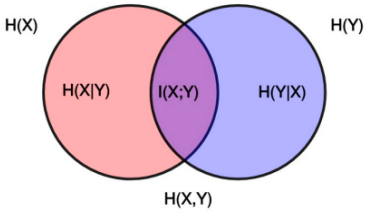
Then we make a clustering: group proteins or genes that have identical or similar profiles.

The genes that share a **similar profile** will tend to be working in the **same biological process.**

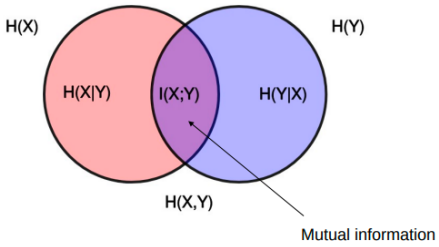
Because if there is a biological process that needs 2 genes, these 2 genes will be present in the genome every time this biological process is required.



Distance measurements between profiles



→ **Hamming distance**
Counting how many different instances we have between the 2 profiles. In other words, how many times there is a species that has a gene and the other **does not**. ↑ distance, ↓ similarity



→ **Mutual information (intersection)**
Counting how many instances are **shared** between the 2 profiles. ↑ mutual information, ↑ similarity

→ **Jaccard index** (intersection/over union)
Counting how many instances are shared between the 2 profiles and dividing it by the total "surface" (all the instances). ↑ Jaccard, ↑

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

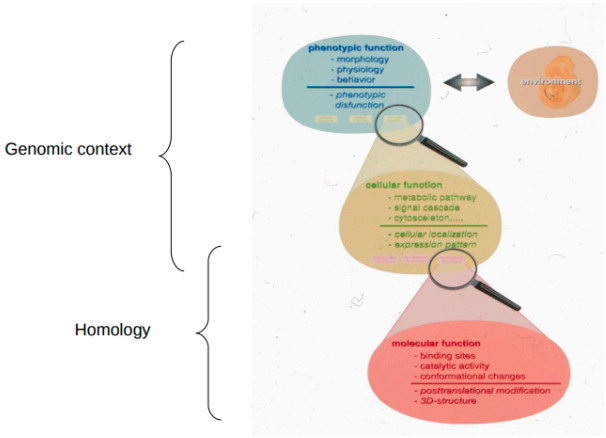
$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

Genes with **complementary phylogenetic profiles** tend to have a **similar biochemical function**. It makes no sense to have different genes that do the same (**redundant**).

Types of genomic context

Methodologies that are based on genomic context (gene fusion/fission, gene order, phylogenetic profiling) give us information that is totally different and non-overlapping with the information obtained with homology based function prediction (a **similar sequence** means having a **similar function**).

Genomic context: if you are always close to a gene that synthesizes tryptophan, maybe you are involved in the same pathway…

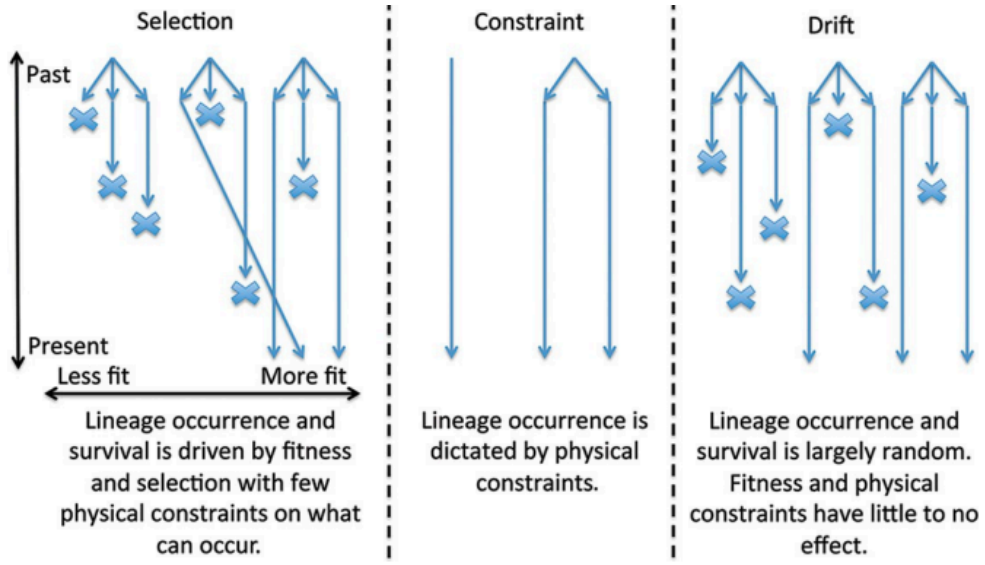


How many genomes must we include in a phylogenetic profiling (predicting the function)? If we add more genomes, we will have a better result. But at a certain number, we will reach a *plato*. Also, there is **no** difference between using a subset with maximally diverse organisms (2 species for each phylum) or a subset with a randomly selected species. Few genomes → choose maximally diverse organisms; Many genomes → doesn't matter random selected or diverse organisms

Reaching a "plateau" refers to the point at which adding more genomes to the analysis does not significantly improve the accuracy or reliability of the predictions. It means that beyond a certain number of genomes, the additional information gained becomes negligible or redundant.

Convergent evolution
They adapt to a similar environment and therefore they acquire very similar traits. It's the independent evolution of similar features. Convergent evolution creates analogous structures that have similar form or function but were not present in the last common ancestor of those groups. E.g. Cactus and Euphorbia.

Ways of explaining convergence:



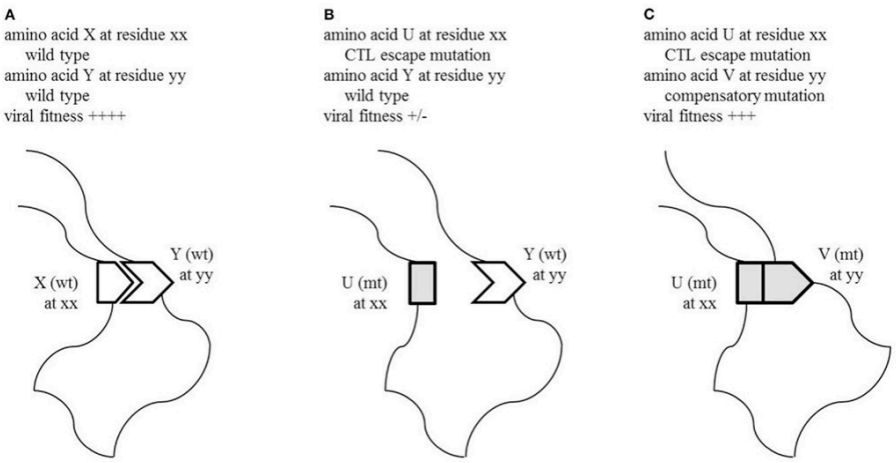
Constraints: in the case of genome rearrangements, maybe two genes can't be separated because they are at both sites of a centromere (you can not break that part because then, the chromosome would be unstable). If there is a constraint, this implies that this has always been like that.

Positive selection can cause tree patterns **different** from the gene tree by convergent mutations. Two species that are very distant can have some genes converged (gene tree will be really different to the species tree).

> **Converged genes**: genes that have evolved independently in different species to perform similar functions despite not sharing a common ancestry.

Co-evolution at the sequence level

The structure of a protein is maintained because of weak interactions (electrostatic interactions) between the Aa of the sequence. If there is a mutation that changes an Aa that is involved in the structure, then it will tend to have a **compensatory mutation** (mutations that correct a loss of fitness due to earlier mutations) to return to the initial (or similar) structure.



So, if in an alignment we see that there are some Aa that tend to **mutate together** (in opposite directions: one goes from + to - and other from - to +), we can make an hypothesis saying that both residues are **interacting**.

We can do this within a protein or also between different proteins. So, we can predict proteins that interact. Therefore, we will need to check for **correlations** between Aa of different proteins.

To do this, sometimes we don't even need to look at the alignments → MIRROR TREE APPROACH:

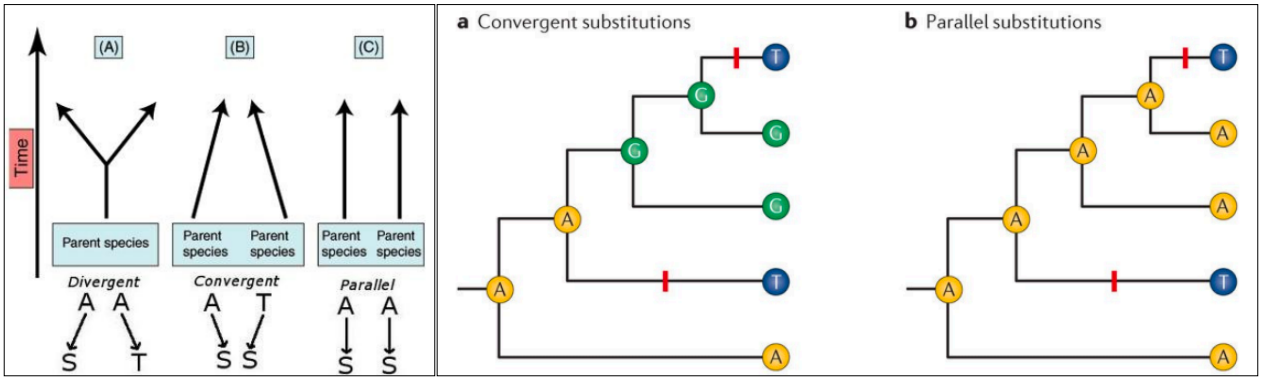
- **Reconstruct** the tree from **each** gene family
- We obtain a **distance matrix** from **each** tree

- **Measure** the **correlation** between the two matrices. If the two trees are similar (similar branch lengths) there will be a ↑ correlation. Then we expect the proteins to interact. If one protein evolves fast, the other will also evolve fast.

We can also look for co-evolution between different organisms that interact (e.g. host parasites, viruses).

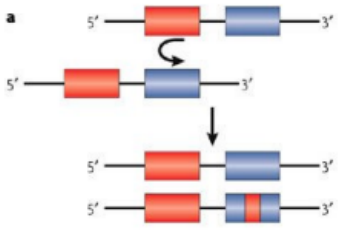
Difference between convergent and parallel evolution

- **Convergent** come from **different nts**
- **Parallel** come from the **same nt** → process in which independent species acquire similar traits while evolving together in the same space and time.



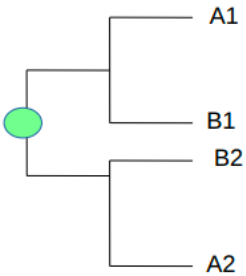
Gene conversion and parallel, concerted evolution

Gene conversion: in the same genome you have two similar genes (e.g paralogs) that can recombine and exchange some material. This happens because of DNA repair.



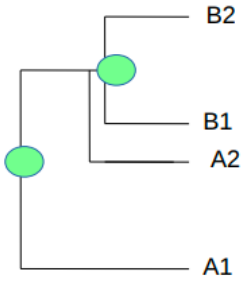
→ This may lead to confusion when doing analysis.
Imagine that we have a gene duplication and both genes diverge (they are different but they are still similar because they are paralogs).

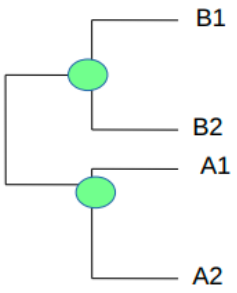
Then, there is a speciation. So, we have species A and species B that both have the two genes. As expected, the same protein of different species is more similar.



If there is gene conversion, B1 converts to B2. In other words, B2 is copied over B1. Then, the gene tree changes, making you think that there has been a duplication in species B.

So, the orthologs are closer than ancient paralogs





If it happens in species B, it can also happen in species A. Then, it looks that there was an speciation at the beginning and then, two parallel duplications.

Paralogs are closer than orthologs, apparent parralel duplication.

Session 7. GE Analysis

One way of knowing the function of a gene is knowing **when** and **where** this gene is expressed. So, we can interrogate the transcriptome under different conditions, tissues… to understand the function of genes.

Transcriptome: complete collection of transcripts present in a specific cell, tissue, organism… at a given time-point. It is very dynamic (reason why every cell has the same DNA information but has a different phenotype). The transcriptional process is highly regulated. The cell controls very well which genes are going to be expressed in each condition (temporally, spatially):

- **Enhancer**: region in the DNA that attracts some activator proteins that will open the chromatin. That will allow other TFs to come and bind to promoter regions, where the transcription will start.
- There are repressors…
- **Methylations** also regulate by promoting/inhibiting depending on the case.

The transcripts will be spliced and then they will be modified in the 3' (addition of poly-A tail) and 5' (cap addition) ends, that will ensure the protection and stability of the transcript.

How do we know when and where a gene is expressed?

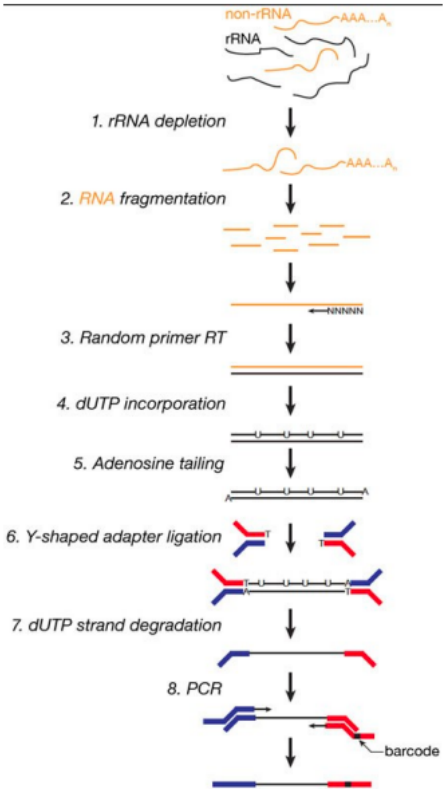
Previous technologies that they used to do this analyses:

- RT-PCR: precise, good but very ↓ throughput
- EST, SAGE: mid-throughput but bad coverage
- DNA microarrays: ↑-throughput, good information on expression levels, but no direct information on splicing, etc… not suitable to discover new transcripts.

- RNA-seq

In most cases we convert the RNA into cDNA and then we do the analysis. Others, like Oxford Nanopore (measures changes in conductivity that goes through the pore. The conductivity depends on the nt that is passing and the modifications) use RNA directly.

Generally, the term RNA-seq is used to indicate any RNA sequencing method based on a **shotgun approach** (not a specific fragment). The advantage of a shotgun, sequence-it-all method, over a tag-based method, is the ability to quantify the expression level of each exon within a transcript, estimate their percent inclusion level and detect (differential) alternative splicing events.



However, it is difficult to identify the exact 3' and 5' ends of transcripts due to various technical biases (random hexamer priming, oligo dT priming) leading to under-representation of sequences near 5' and 3' ends. Exist plenty of different protocols but they have many steps in common:

- rRNA depletion and fragmentation of RNA
- Conversion of RNA into cDNA (performed by oligo dT or random primers)
- Second strand synthesis
- Ligation of adapter sequences at the 3' and 5' ends
- Final amplification

rRNA depletion: rRNA (90%) constitutes the majority of the RNA. Thus, we need to get rid of it.

How can we selectively remove the rRNA?

Use probes/oligos that bind specifically the rRNA. Then, we add an enzyme that degrades the heteroduplex. The problem is that we need to know the sequence of the rRNA.

An alternative that works in eukaryotes is to capture the mRNA using an oligo of T. It will bind to the poly-A tail that is only found in mRNA.

Considerations:

- Bacteria have no poly-A tail.
- Ribosomal depletion kits are based on hybridization to specific sequences so they are optimized for specific species.

Target enrichment

Used to enrich some of the transcripts, but it is not based on the poly-A tail. We have to design probes for the transcripts we want to capture.

Transcript orientation

All tag-based methods are strand specific, meaning that they preserve information about the transcript's orientation, while shotgun methods may be strand-specific or not strand specific.

Strand specificity is important to determine the exact GE levels in the presence of antisense transcription, or for accurate prediction of certain classes of transcripts (e.g. lncRNAs).

Strand-specific methods can be classified into two categories:

- RNA-seq methods based on **ligation of two different adaptors** in a known orientation relative to the 5' and 3' ends
- RNA-seq methods based on **chemical modification** of the RNA, either by bisulfite treatment or by the incorporation of dUTPs during the second-strand cDNA synthesis.

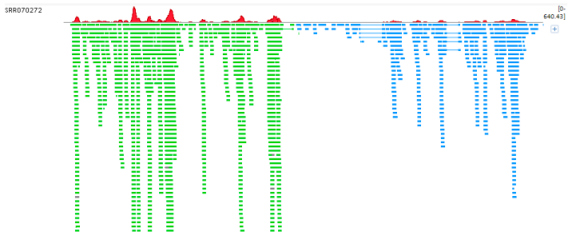
In both cases, the non-modified strand is degraded enzymatically

BI analyses

Imagine that we have done this enrichment. Now we have all the sequences of RNA and we want to analyze them. We have to translate this into the expression of different genes.

The easiest way is to map these reads into a reference genome. We can also do transcriptome assembly (try to reconstruct the transcripts) or de-novo transcriptome assembly (no need of reference genome).

e.g. Results when mapping the reads into a reference genome.

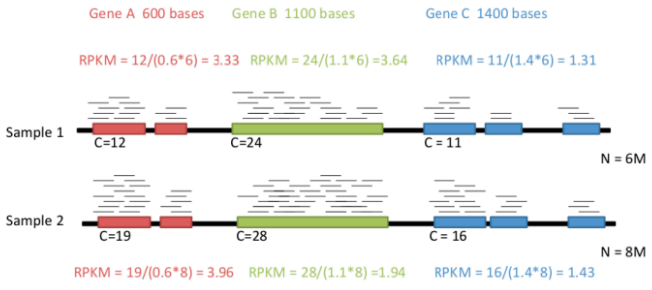


We can see that green gene is more expressed, because it has more reads. But can we quantify it?

The expression of genes is assessed by counting how many reads map to the gene, taking into account read length and total # of reads (RPKM or RFKM).

Note that we are normalizing using the length of the gene (because otherwise, we could have more reads because the gene is longer).

We also take into consideration the number of million reads that we obtained.



Counting rules

- Count reads, not nucleotides
- Count each read at most once.
- Discard a read if
  - it cannot be uniquely mapped
  - its alignment overlaps with several genes
  - the alignment quality score is bad
  - (for paired-end reads) the mates do not map to the same gene

Usually log2FoldChange and p-value thresholds are put to select DGE. Each dot is a gene.

- We want a small p-value and a big FoldChange, meaning that there is a DE that is significant.
- We can use information from RNA-seq to reconstruct intron-exon structures of the genes and also to recognize alternative splicing. This is because when we map against the reference genome, we can see reads that map to a given exon and other reads map to half of one exon and half of another (meaning that there is an intron between).