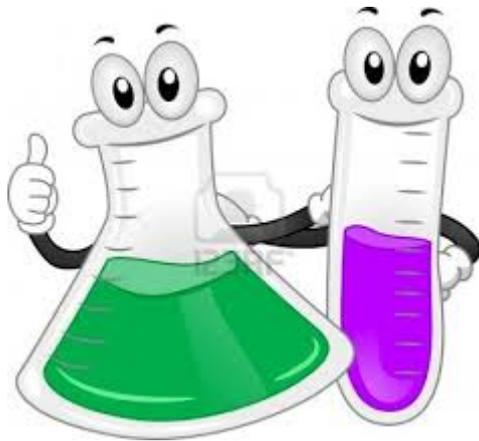


# Session 2: Orthology assignment

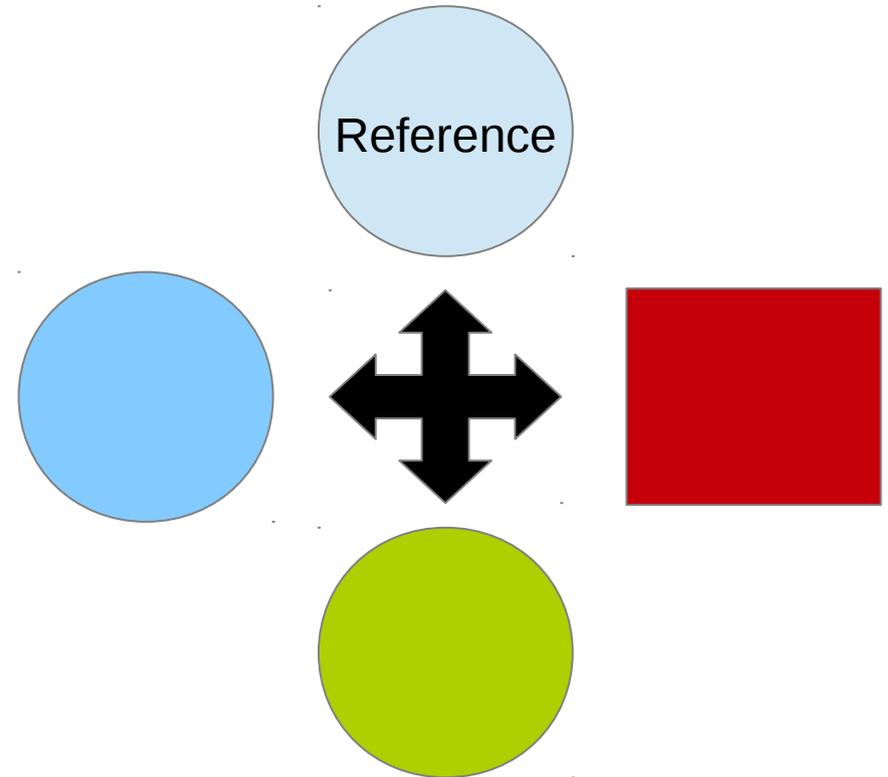
# What do genes do and how did they evolve?



Doing experiments for each element we have identified to discover their function is impossible right now.



Predictions need to be verified.



Comparative genomics. Tries to exploit genome data to produce biological knowledge. Similarities and differences between genomes can provide information.

Time (My)



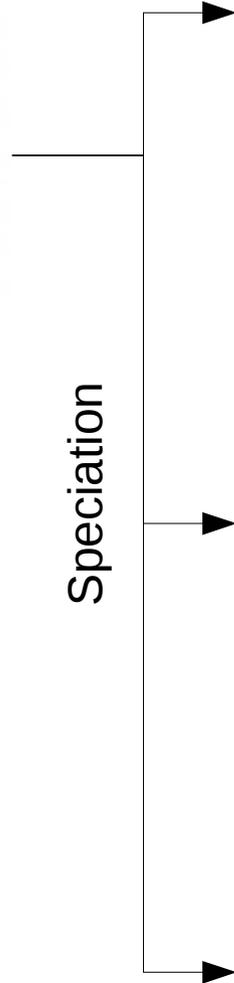
G1 - F1



Duplication



G1-F1  
G1' - F2



Speciation



S1 - G1 - F1  
S1 - G1' - F2



Duplication



S1 - G1 - F1  
S1 - G1' - F2  
S1 - G1'' - F3



S2 - G1 - F1  
S2 - G1' - F2



S2 - G1 - F1  
S2 - G1' - F2



Loss



S3 - G1 - F1  
S3 - G1' - F2



S3 - G1 - F1

G → Gene  
F → Function  
S → Species

NOW: we sequence the three species



—————

S1 – G1

—————

S1 – G2

—————

S1 – G3



—————

S2 – G4

—————

S2 – G5



—————

S3 – G6

After sequencing we will obtain a list of genes for each species but **we will not know the relation between them nor the function they have.**

First approximation to functional annotation: BLAST

**Homologs:** the genes come from the same common ancestor



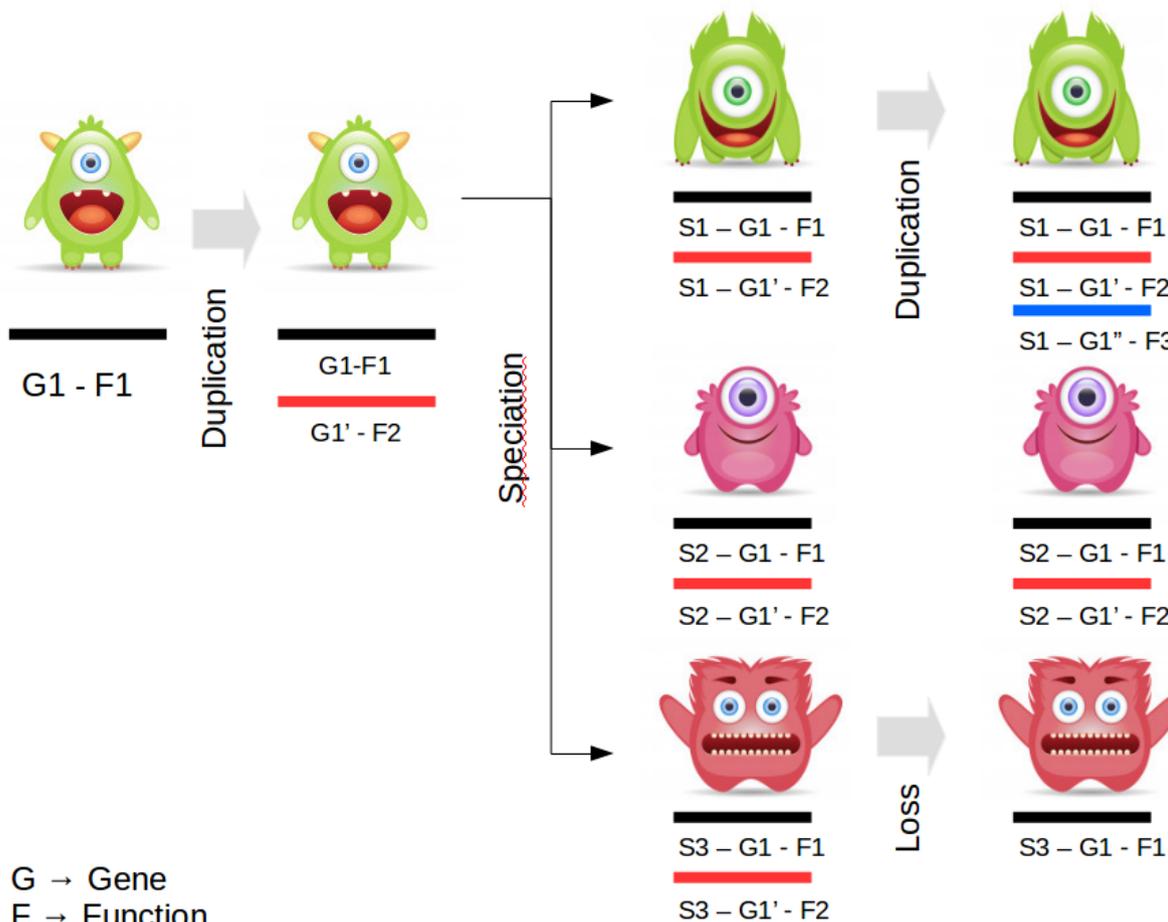
—————

Using blast to search for the function of a gene is a generalization often done in bioinformatics. But as seen in this example, it will not be completely correct.

To be able to assign function properly we will need to group the genes into orthologs and paralogs.

Orthologs: genes that come from a speciation event. Orthologs tend to have the same function.

Paralogs: genes that come from a duplication event. After the duplication, one of the copies will likely adopt a different function.



G → Gene  
 F → Function  
 S → Species

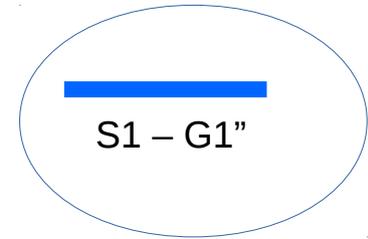
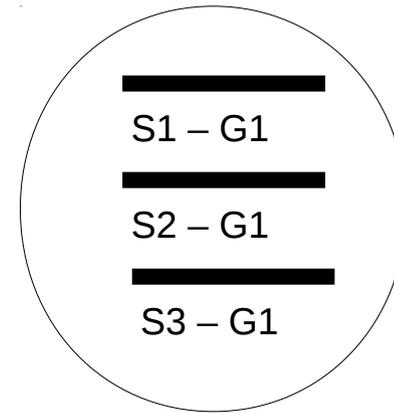
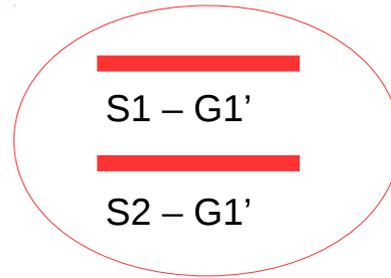
Red and black genes are paralogs.

All black genes are orthologs between them.

All red genes are orthologs between them.

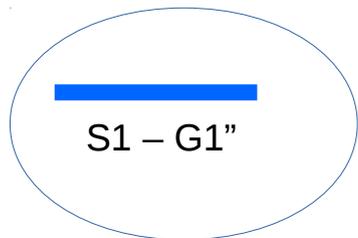
The blue gene is paralogous to the red and black genes.

## Orthologous groups



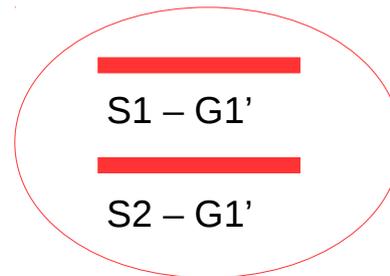
All members of the same orthologous group are assumed to have the same function. If we know the function of one of the members of the group we will assume all the others have the same one.

## Paralogs – inparalogs and outparalogs



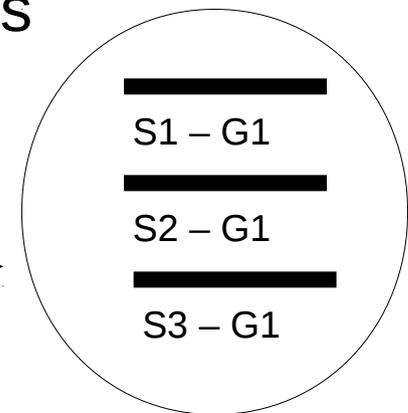
inparalogs

The duplication happened **after** the speciation event



outparalogs

The duplication happened **before** the speciation event



# How can we distinguish between orthologs and paralogs?

█  
S1 – G1

█  
S1 – G2

█  
S1 – G3

█  
S2 – G4

█  
S2 – G5

█  
S3 – G6

Based on similarity (blast based): These methods rely on the similarity between sequences in order to make groups of orthologous genes. They tend to be fast and work well in simple orthology / paralogy scenarios but they can get confused in more complicated settings.

**We will see: BBH, inparanoid and orthoMCL**

Tree based methods: The methods use a phylogenetic tree reconstruction to infer orthology and paralogy events. They tend to take more time, but are more accurate and can deal better with complex scenarios.

**We will see: species overlap and reconciliation**

