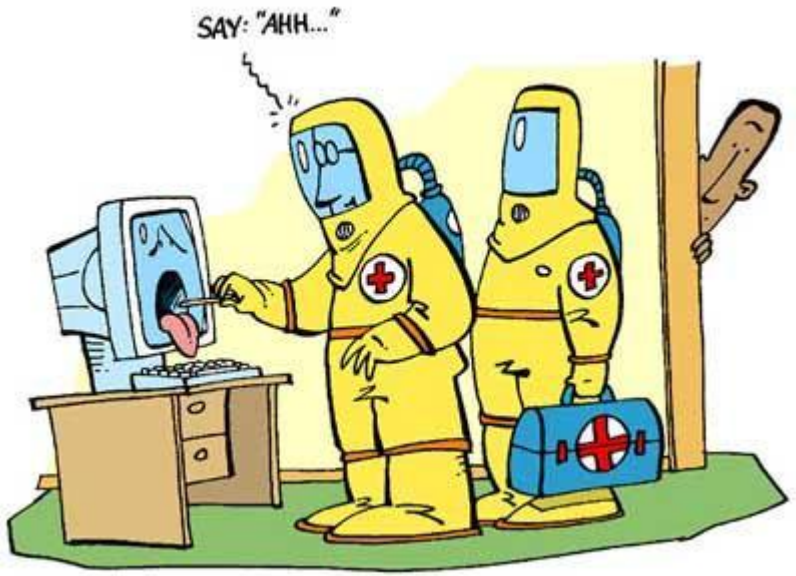
Bioinformatics?



Reads, assembly, annotation, comparative genomics… and a bit of phylogeny

stefano.gaiarsa@unimi.it

PART 4

Case study!

Remember the *Mycoplasma* outbreak?!

The horizontal transfer of a DNA region, gives new functions to our strain

Where does this DNA come from?

Case study!

Remember the *Mycoplasma* outbreak?!

A novel *M. bovis* isolate causes an outbreak

We have sequenced and assembled the genome

We have found a genome insertion with Mauve

We have called genes and we have annotated them

We have compared the gene content with the reference

These novel genes could increase virulence

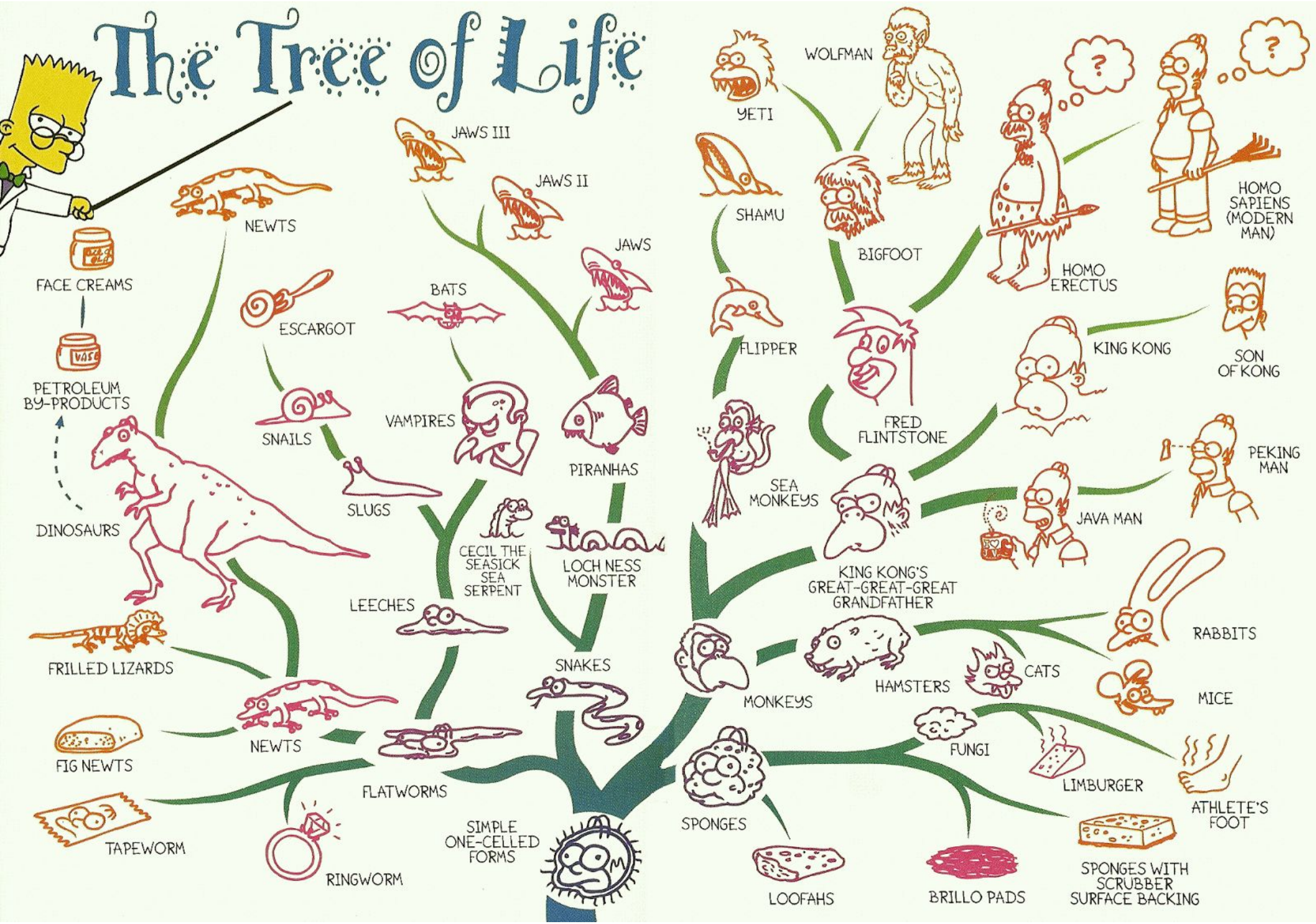
Was there an horizontal transfer of DNA from a donor bacterium?

Case study!

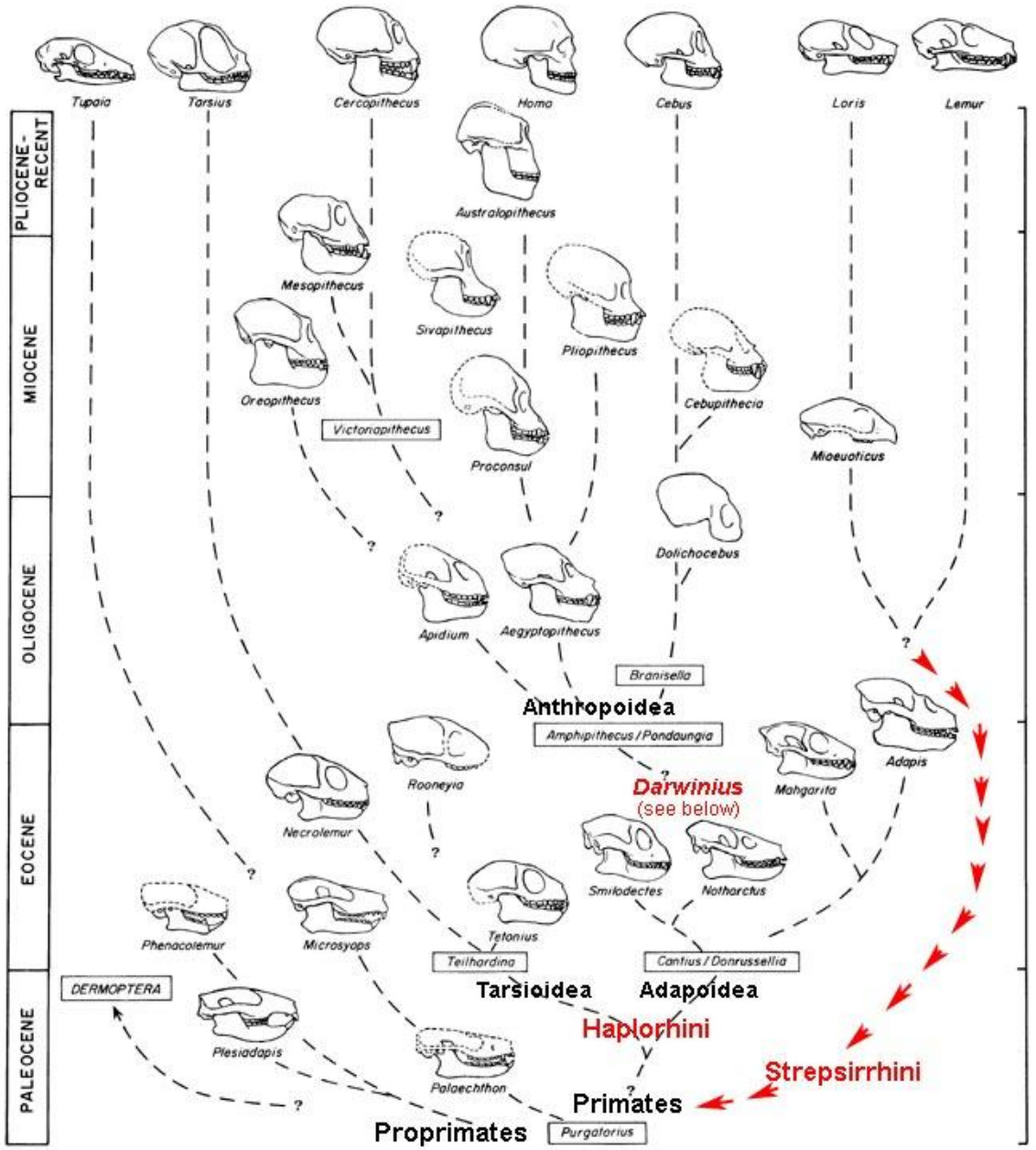
Remember the *Mycoplasma* outbreak?!

If this is the case, who is the donor?

We will find this out using PHYLOGENY

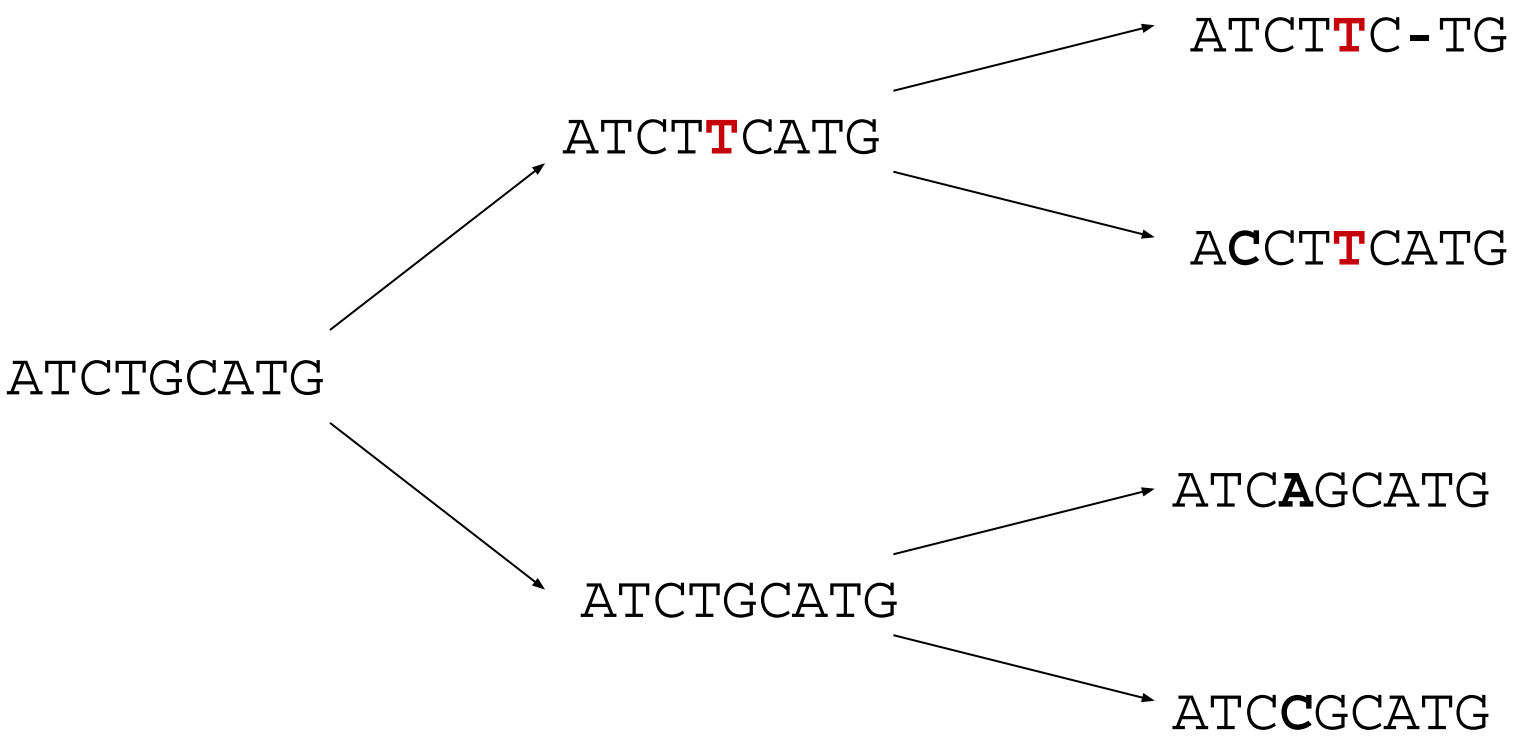

The Tree of Life
JAWS III
JAWS II
JAWS
NEWTS
FACE CREAMS
ESCARGOT
BATS
PETROLEUM BY-PRODUCTS
SNAILS
VAMPIRES
PIRANHAS
DINOSAURS
SLUGS
CECIL THE SEASICK SEA SERPENT
LOCH NESS MONSTER
LEECHES
FRILLED LIZARDS
SNAKES
NEWTS
FIG NEWTS
FLATWORMS
TAPEWORM
RINGWORM
SIMPLE ONE-CELLED FORMS
YETI
WOLFMAN
SHAMU
BIGFOOT
HOMO ERECTUS
HOMO SAPIENS (MODERN MAN)
FLIPPER
KING KONG
SON OF KONG
FRED FLINTSTONE
SEA MONKEYS
JAVA MAN
PEKING MAN
KING KONG'S GREAT-GREAT-GREAT GRANDFATHER
RABBITS
HAMSTERS
CATS
MONKEYS
MICE
FUNGI
LIMBURGER
ATHLETE'S FOOT
SPONGES
LOOFAHS
BRILLO PADS
SPONGES WITH SCRUBBER SURFACE BACKING

Phylogeny



Reconstruction of EVOLUTION, using **differences** and **common** morphological **tracts**

Phylogeny

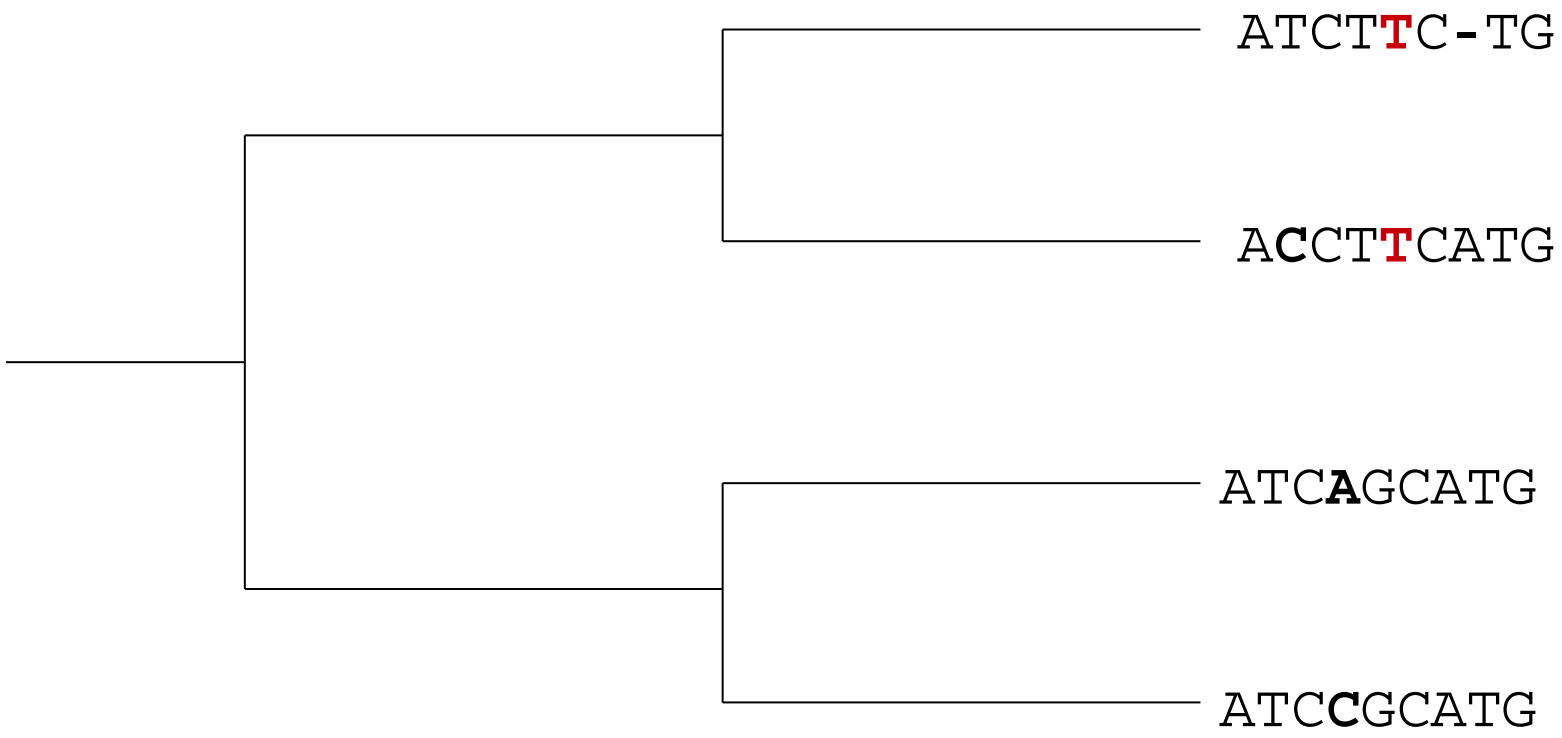


Reconstruction of **EVOLUTION**, using **DNA sequences**

MOLECULAR PHYLOGENY

Phylogeny

We do not have information on the ancestors!!

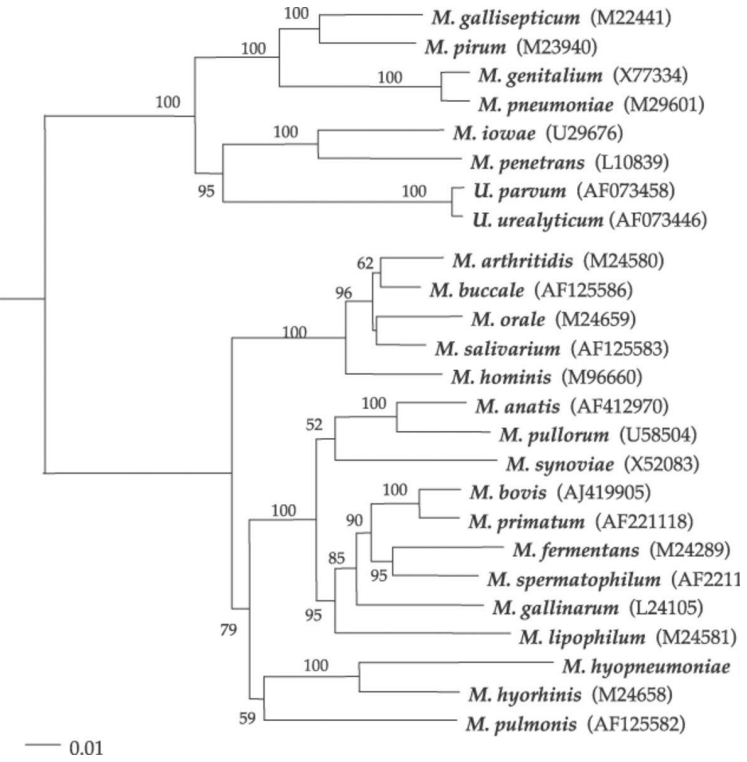


Reconstruction of **EVOLUTION**, using **DNA sequences**

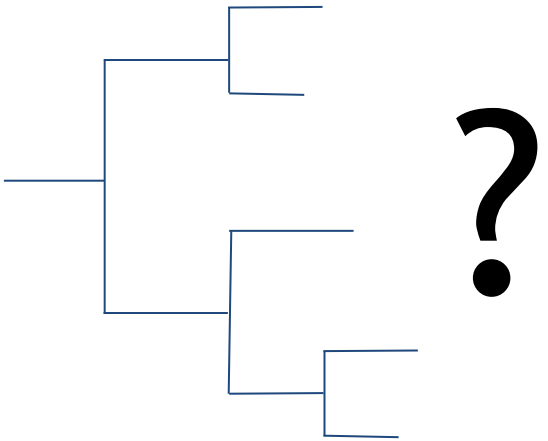
MOLECULAR PHYLOGENY

why Phylogeny?

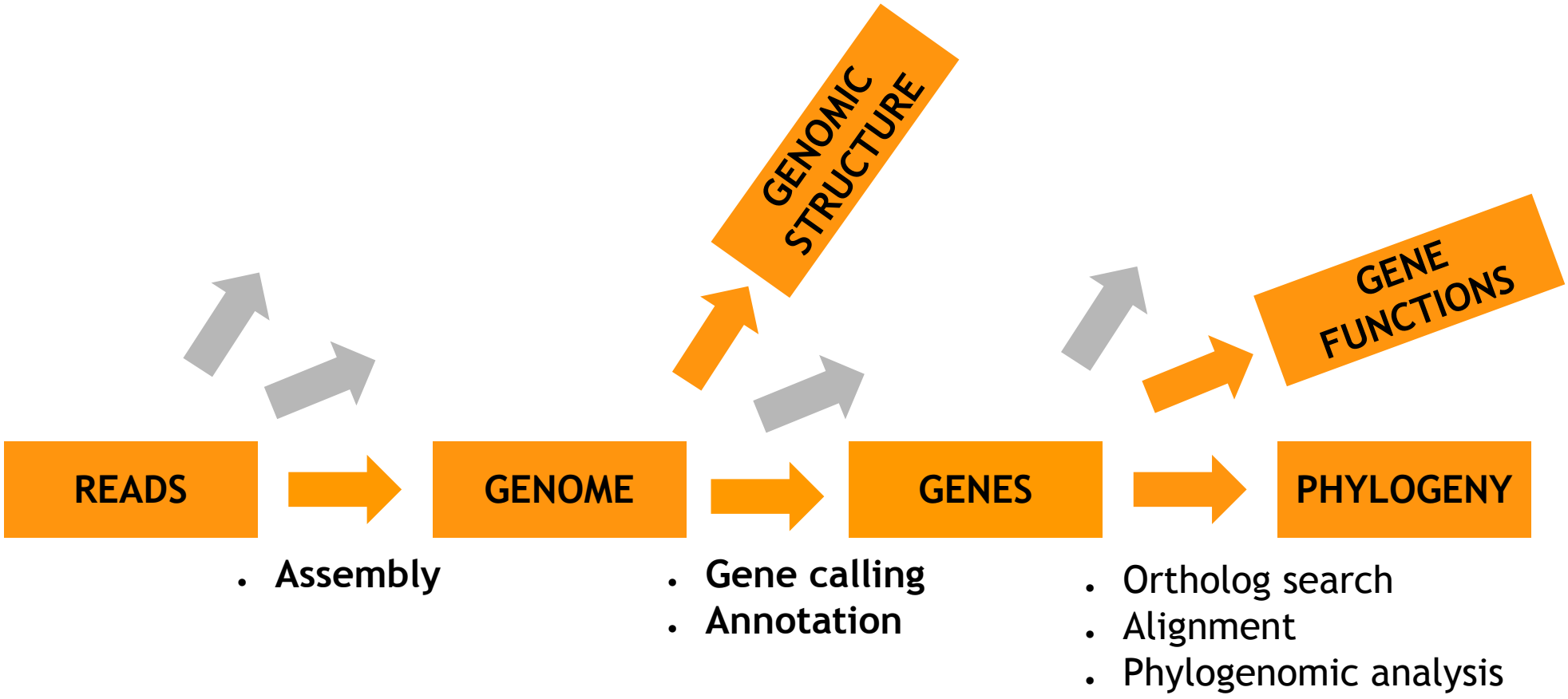
The molecular phylogeny of the genome will show the evolution of the species (in our case *M. bovis*)



The molecular phylogeny of the inserted region will resemble the evolution of the **DONOR organism**!!



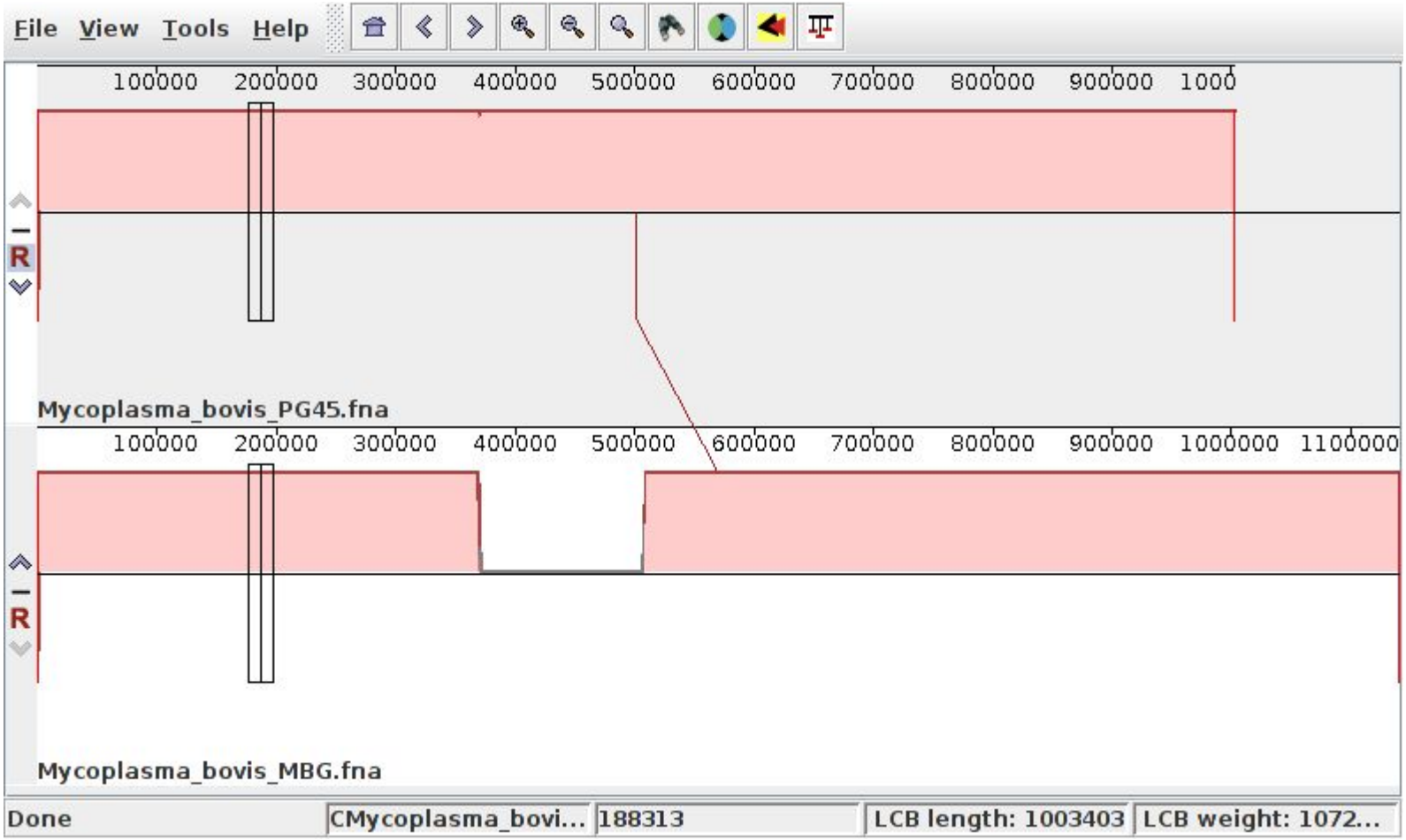
What now?


GENOMIC STRUCTURE
GENE FUNCTIONS
READS
GENOME
GENES
PHYLOGENY
Assembly
Gene calling
Annotation
Ortholog search
Alignment
Phylogenomic analysis

Phylogeny

For sake of simplicity we will reconstruct the story of a single gene

We must extract our gene of interest from the region that was horizontally transferred



Phylogeny

For sake of simplicity we will reconstruct the story of a single gene

the transferred region is located from **~370000** to **~512000** in the MBG genome

Select the first protein present in this region, and insert it in a **novel empty file**

Phylogeny

We will reconstruct the story of a single gene

Based on our knowledge of the literature, we know that in our area, multiple bacterial pathogens could be present in the farm
and could have donated the gene

We have their genome in our computers

We will look for a **single protein** from the inserted region in **a dataset of potential donor organisms**

YOUR TURN! What tool would you use?

Phylogeny

We will reconstruct the story of a single gene.

1) We will look for the **same protein** from the inserted region in **all the organisms** we want to analyze.

YOUR TURN! What tool would you use?

yes, the Swiss army knife of bioinformatics:

… BLAST!

Phylogeny

INSTRUCTIONS FOR THE EXERCISE:

You will find the gene “MBG_protein.faa” in the “protein_search” folder

You will also find a set of organisms you can use as databases

Pick an organism in which you wish to find which protein is homologous to ours

Run the **makeblastdb** and **blastp** commands
DO NOT use “-max_target_seqs” and “-outfmt”

Phylogeny

OK! Since now you are all expert Blast users:

YOU CAN REPEAT THE BLAST ON THE OTHER 5 GENOMES!

this means doing the BLAST 6 times, or...

Phylogeny

OK! Since now you are all expert Blast users:

YOU CAN REPEAT THE BLAST ON THE OTHER 5 GENOMES!

or you can use the cat command to join the databases:

```
you@yourPC[yourdirectory] cat [first_sequence.fasta] [second_sequence.fasta]
> [joined_sequence.fasta]
```

Phylogeny

We have now detected the homologous proteins from the 6 analyzed genomes

We can **create a multifasta file** containing these 6 genes, plus the input gene from our *M. bovis* genome

No need to use complex (bio)informatics for this task, just copy and paste

Phylogeny

2) We need to **ALIGN** all the proteins, in order to spot the differences

3) then **eliminate some noisy amino acids**

4) Do the phylogeny

Phylogeny is a complex task
There are a lot of possible ways to perform analyses
Some people have dedicated their entire scientific career to molecular phylogeny, so it's not a really 5 minute thing…

..but we are going to take it the easiest way and use **3 simple tools**

Phylogeny

muscle is used to **align** sequences in order to spot **which** base/aminoacid mutated and **what it became**

```
MTSGFFVPGLMSKYNTKEIRESPDKAKIPSS
MTSGFFPKLESKYNTKEIRERMLKPDKAKIDSS
MTSGFFPGLMSKYNTKEIRESMLKPDKAKIDSS
MTSGFFGLESKYNTKEIRESMLKPDKAKIDSS
```


```
MTSGFFVPGLMSKYNTKEIRES---PDKAKIPSS
MTSGFF-PKLESKYNTKEIRERMLKPDKAKIDSS
MTSGFF-PGLMSKYNDKEIRESMLKPDKAKIDSS
MTSGFF--GLESKYNTKEIRESMLKPDKAKIDSS
```

```
you@yourPC[yourdirectory] muscle -in [input.fasta] -out [alignment.fasta]
```

Phylogeny

Gblocks is used to **eliminate** the positions in the alignment, which are not reliable for the phylogeny. i.e the gaps and some adjacent positions that are not reliably aligned


```
you@yourPC[yourdirectory] Gblocks [alignment.fasta] -t p
```

i.e. proteins

Phylogeny

FastTree is used to perform **phylogeny**

The input is an alignment, the program **reads** the **positions with substitutions** and predicts the **evolution** of the sequences

```
you@yourPC[yourdirectory] fasttree < [trimmed_alignment.fasta] > [tree_file]
```

We can look at the output using the software **seaview**
can be launched from the command line
can be searched in the menu
can open files through right-click

Case study!

Remember the *Mycoplasma* outbreak?!

Most probable conclusion:

A DNA insertion from *Staphilococcus aureus*, makes our *Mycoplasma* strain really BAD ASS!

Phylogeny

You can find the same tools inside SeaView:

- **muscle: to align all the sequences**
- **Gblocks: to cut out all the bases that could not be used for phylogeny (i.e. gaps and bases/amino acids with low quality alignment)**
- **phyml: to run the phylogeny**