PHYLOGENETIC RECONSTRUCTION /Roadmap/

BLAST

1. Potential homologs candidates detection will depend on:
   a) Sample size
   b) E-value
   c) identity

Clean fasta headers[1]

2. `cut –d “ ” –f 1 myBlast.fasta > myBlast_short.fasta`

Obtain Taxonomy

3. Extract ACCESSION numbers:
   `grep “>” myBlast_short.fasta | sed –E ‘s/>//’ | tr “\n” “ ”`

4. Run the script:
   `./get_Taxa.sh > myTAXA_info.txt`

Verify the required format for the TAXA table: tabs, spaces, Non-alpha-numeric characters, etc.

ACCESSION to Gis[2]

5. Run the script:
   `perl change_fasta_headers.pl myBlast_short.fasta > myBlast_GIs.fasta`

Redundancy[3]

6. READ instructions. Use the program CD-HIT:
   `cd-hit –i myBlast_GIs.fasta –o myBlast_GIs.nr98.fasta –c 0.98`

Alignment[4]

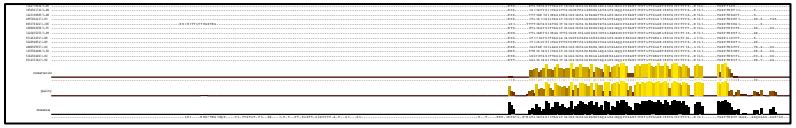
7. Use the program MAFFT:
   `mafft myBlast_GIs.nr98.fasta > myBlast_GIs.nr98.aln.fasta`

8. Rename tree label list (two columns)

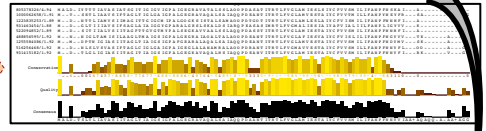
1321786652	Bacteria_-_Proteobacteria_-_Deltaproteobacteria
406895154	Bacteria_-_environmental_samples_-_-_uncultured_bac
700405439	Bacteria_-_Proteobacteria_-_Deltaproteobacteria_-_De
780797372	Bacteria_-_Proteobacteria_-_Deltaproteobacteria_-_De
499677068	Bacteria_-_Chlorobi_-_Chlorobia_-_Chlorobiales_-_Chl
522212918	Bacteria_-_Acidobacteria_-_Acidobacteriales_-_Acidob

Manual edition of the alignment?

“Before of manual edition”



“After of a manual edition”



Fasta to Phylip format

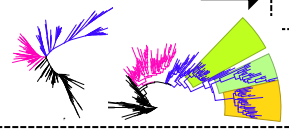
9. Transform FASTA to PHYLIP format (use online resources):
   `myBlast_GIs.nr98.aln.fasta` ➔ `sample.phylip` ➔ `myBlast_GIs.nr98.aln.phy`

Phylogenetic tree[5]

10. Use RAxML to reconstruct a phylogenetic tree:
    `raxmlHPC –m PROTGAMMAWAG –s myBlast_Gis.nr98.aln.phy –n myBlast_Gis.nr98.raxml –p 12345`

11. RENAME the tree:
    `nw_rename RAxML_result.myBlast_Gis.nr98.raxml myLABELS.txt > myTREE_labels.tree`

Bootstrapping[6]



12. PLOT the tree, identify the HGT, and make the tree pretty in FigTree.

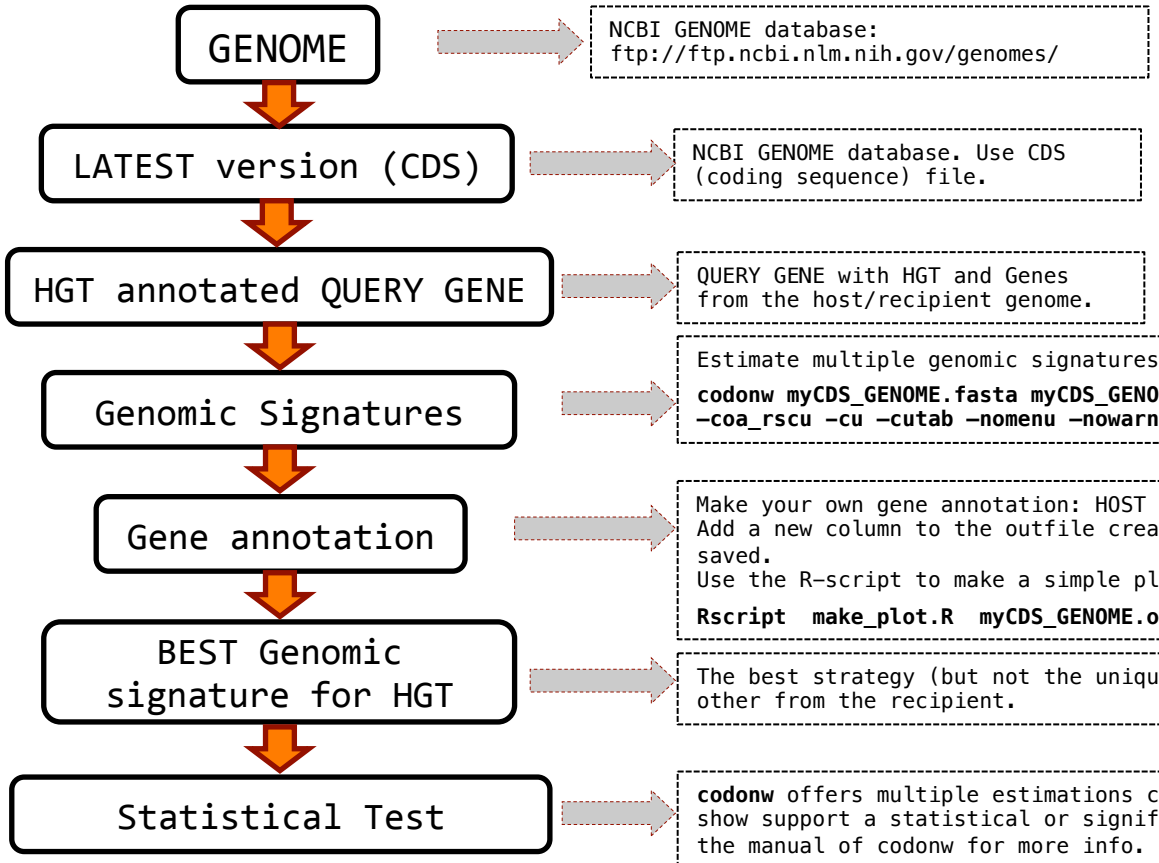
13. Final and crucial step to support the tree topology. **FIRST**, use RAxML to make a FAST-BOOTSTRAPPING and ML search:

`raxmlHPC-PTHREADS –f a –m PROTGAMMAWAG –p 12345 –x 12345 –# 100 –s myBlast_Gis.nr98.aln.phy –n myBlast_Gis.nr98.100_BOOTSTRAP_TREE.raxml`

SECOND, make a CONSENSUS tree using PHYLIP: `consense` ➔ `myBlast_Gis.nr98.100_BOOTSTRAP_TREE.raxml` ➔ `Y`

THIRD, rename the tree branch labels.

PARAMETRIC APPROACH /Roadmap/



NOTES on "Phylogenetic Reconstruction"

Each step is critical, here are some advices:

1. Clear the fasta header helps to manipulate the data, making it less confuse.

2. GI number contains max 10 characters and they are unique for each gene.

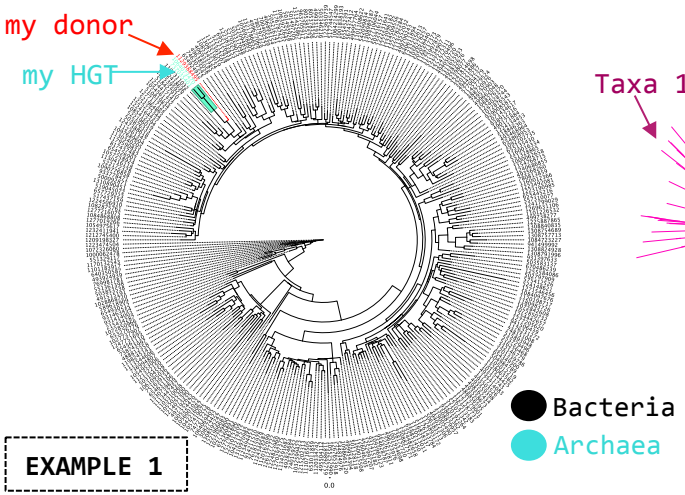
3. Why make a redundancy test? Gene copies or highly similar genes (e.g. from strains) might be remove.

4. Delete regions manually in the alignment has to be done in many cases. Why? For instance, it is used as criterion to improve the alignment or the homologs searching.

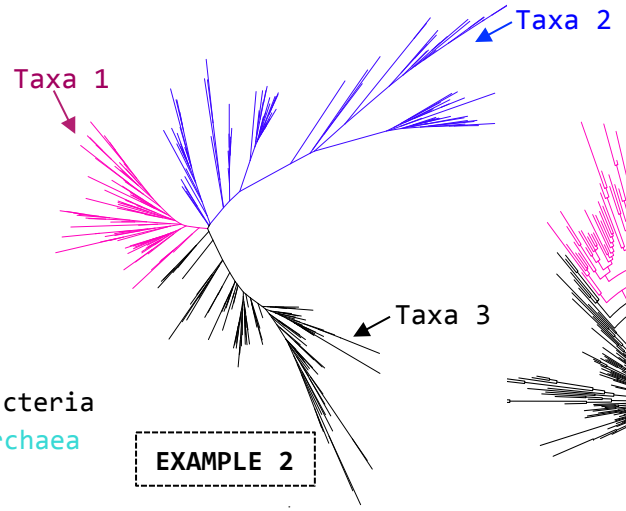
5. Previous to this step, is mandatory—if you don't know how your gene(s) 'evolve'—to identify their best substitution model (aa or nts). Use jModelTest or Prottest to achieve this purpose. Read our manual (step 11). Also, phylogenetic trees based on Maximum Likelihood (ML) or distances are good options, you should try a third approach: the Bayes methods. One good strategy/exercise show both tree topologies, ML and Bayes.

PHYLOGENETIC RECONSTRUCTION

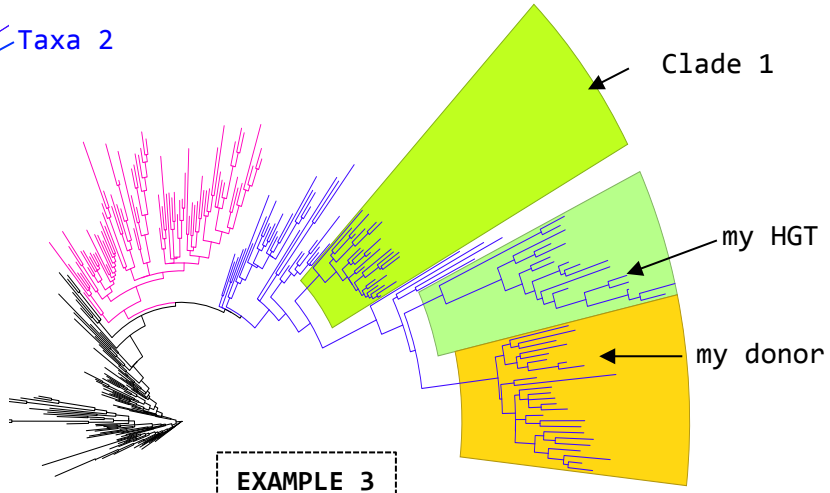
Make your tree pretty using FigTree



EXAMPLE 1

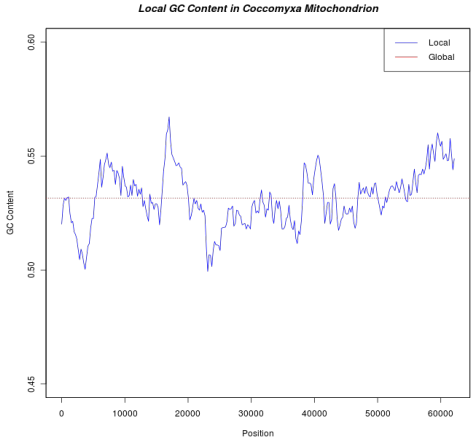


EXAMPLE 2

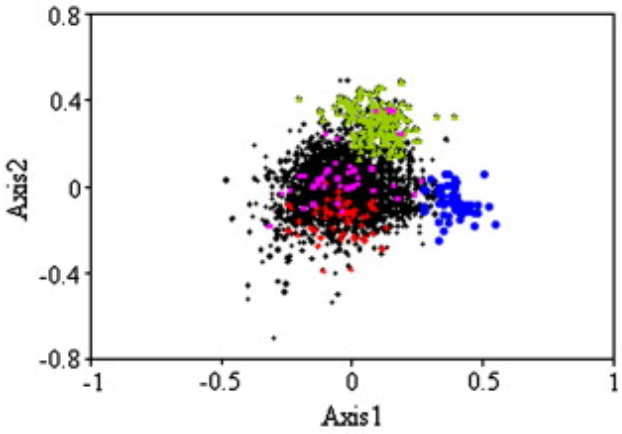


EXAMPLE 3

PARAMETRIC METHODS



EXAMPLE 1



EXAMPLE 2

EXAMPLE 3