

OrthoFinder

Uvod v bioinformatiko 24/25

Luka Duniš

luka.dunis@famnit.upr.si

Homologija genov

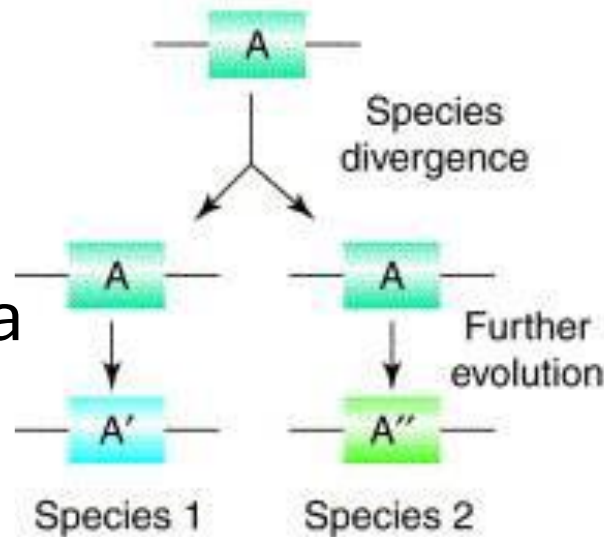
Ortologni geni:

nastanejo iz skupnega prednikovega gena med speciacijo. Gen in njegova glavna funkcija sta večinoma ohranjena.

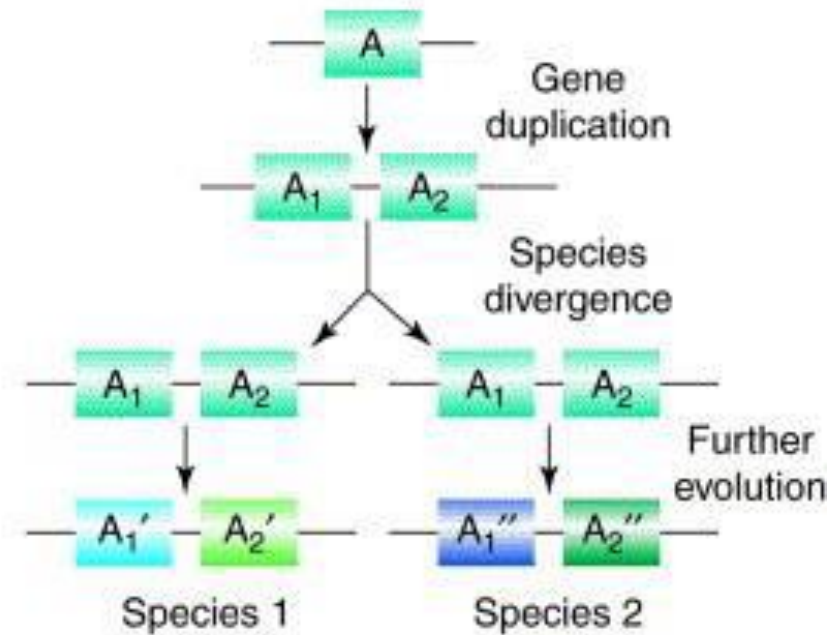
Paralogni geni:

geni znotraj ene vrste ki so nastali s podvojitvijo genov. Paralogi se lahko razlikujejo po zaporedju, sestavi in funkciji.

(a) Orthologous genes



(b) Paralogous genes



Analogni geni

Analogni geni so geni, ki imajo enako ali podobno funkcijo, vendar nimajo skupnega prednika in so zato v nasprotju s homolognimi geni nepovezani.

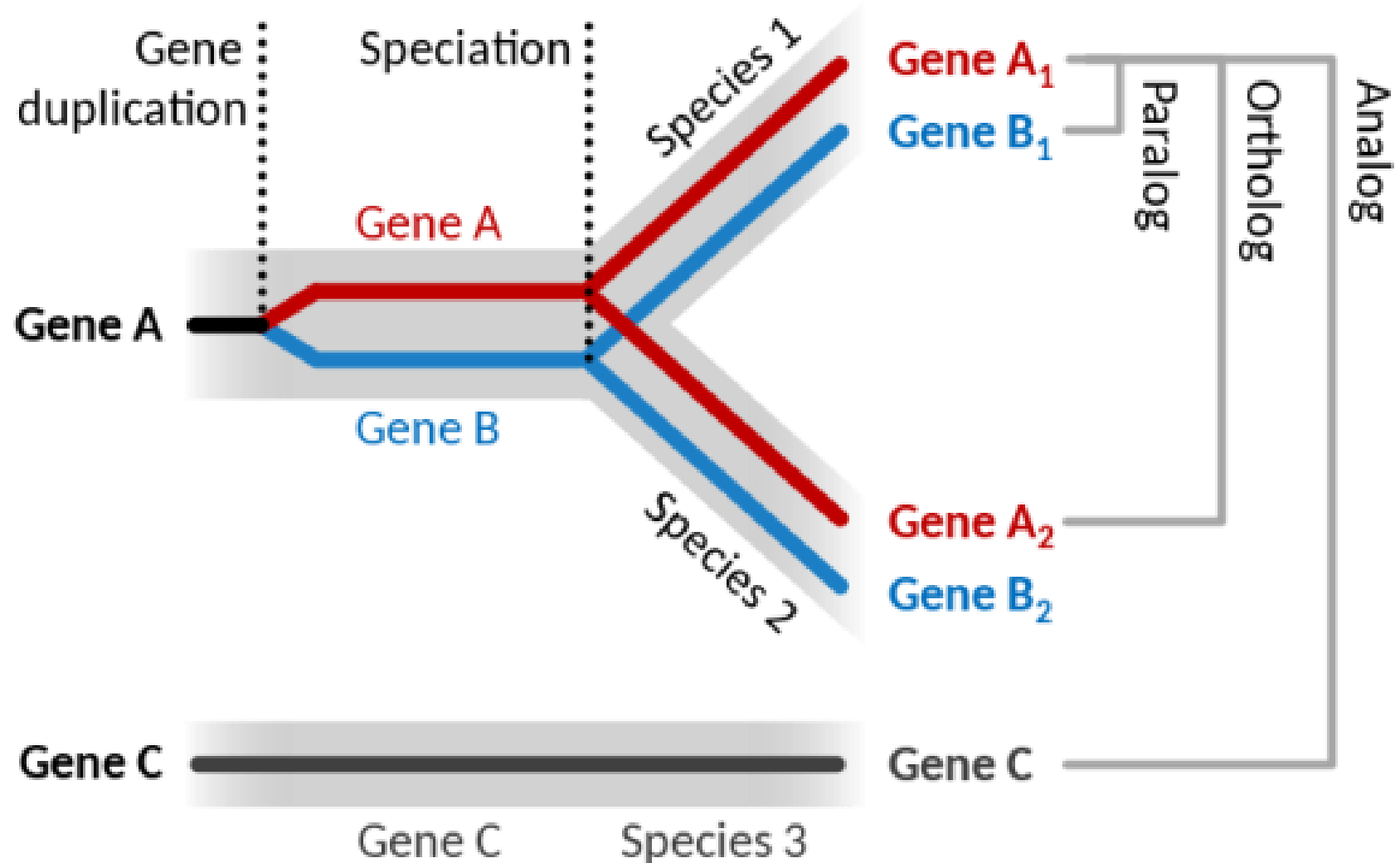


Fig. 1a: Gene Tree

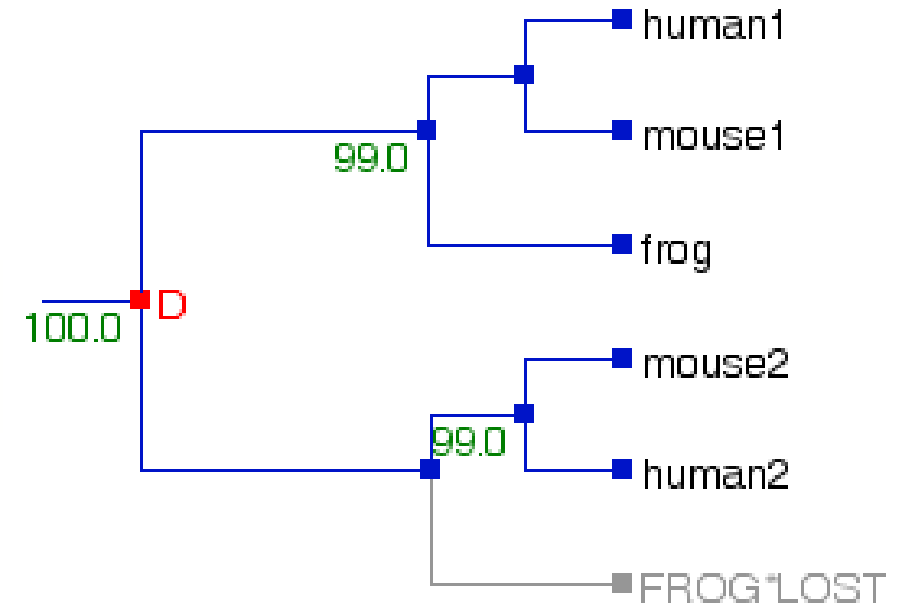
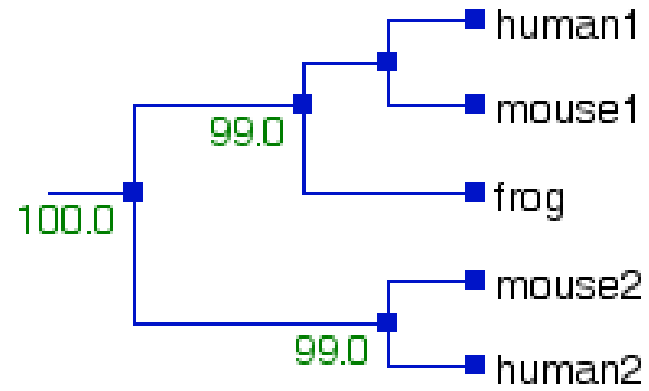
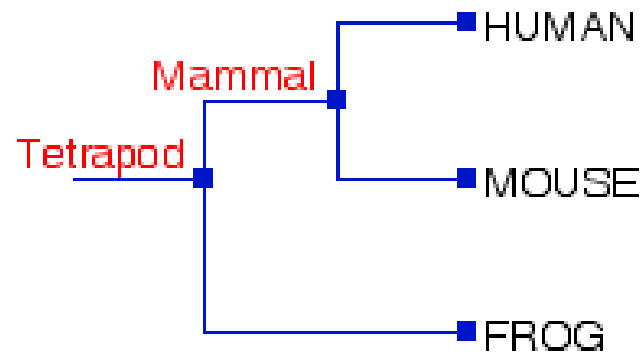


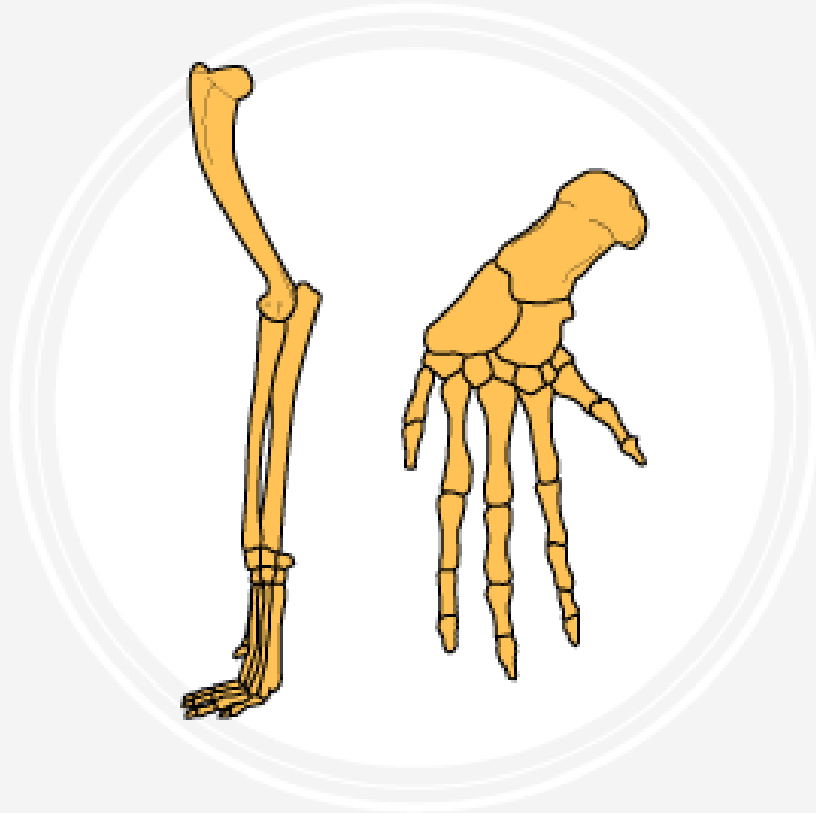
Fig. 1c: Reconciled Gene Tree

Fig. 1b: Species Tree



Primer struktur pri živalih

HOMOLOGOUS VS. ANALOGOUS STRUCTURES



(a) Homologous structures



(b) Analogous structures

OrthoFinder

OrthoFinder je program, ki ga uporabljamo za ugotavljanje filogenije ortologov.

Kot vhodne podatke obdeluje proteome izbranih vrst, iz njih pa samodejno:

- vzpostavi obsežen niz filogenetskih dreves
- oblikuje ukoreninjeno filogenetsko drevo vrst
- določi vsa razmerja ortologov med geni z uporabo dreves
- ugotovi duplikacije genov in jih poveže z ustreznimi vozlišči na drevesih genov in vrst

Poleg tega, da olajšuje analize, služi kot dragoceno orodje za preverjanje posameznih odnosov genov pred eksperimentalnimi študijami.

Namestitev orodja

```
conda activate [okolje]
```

```
conda install orthofinder
```

Ustvarimo novo mapo v kateri bomo prenesli program, ki vsebuje vhodne podatke. Prenesite program, razširite ga in premaknite se v mapi.

```
wget https://github.com/davidemms/OrthoFinder/releases/latest/download/OrthoFinder.tar.gz
```

```
tar -xzf OrthoFinder.tar.gz
```

```
cd OrthoFinder
```

Primer

Zaženimo program OrthoFinder na primer podatkov (mapa ExampleData), to je zelo majhna zbirka podatkov, zato se bo analiza končala hitro, običajne zbirke podatkov bodo trajale dlje.

`orthofinder -f ExampleData/`

OrthoFinder ustvari mapo v mapi z našimi vhodnimi datotekami in vanj postavi vse rezultate, npr: ExampleData/OrthoFinder/Results_Oct11.

Tako je videti mapa z rezultati:

Name	Size	Modified
Comparative_Genomics_Statistics	10 items	15:24
Gene_Duplication_Events	2 items	15:24
Gene_Trees	325 items	15:24
Orthogroups	5 items	15:24
Orthogroup-Sequences	1,135 items	15:24
Orthologues	4 items	15:24
Resolved_Gene_Trees	325 items	15:24
Single_Copy_Orthologue-Sequences	246 items	15:24
Species_Tree	3 items	15:24
WorkingDirectory	39 items	15:24
Log.txt	756 bytes	15:24

Vaja - Primer izvedbe analize OrthoFinder.

V prejšnjem slidu smo prenesli program OrthoFinder in preverili, ali ga lahko zaženemo na primer podatkov. Zdaj smo pripravljeni za izvedbo lastne analize!

Plan: prenesli bomo proteome za niz vrst, ki jih želimo analizirati, datoteke nekoliko uredili in na teh vrstah zagnali program OrthoFinder.

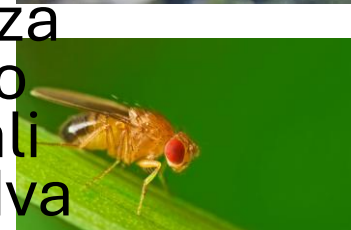
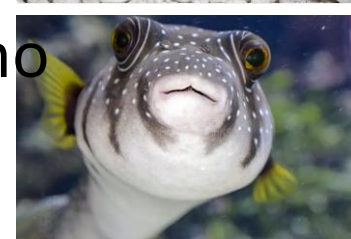
Prenos proteomov za naše vrste

Izvedli bomo filogenomsko analizo na set modelnih vrst: miši, človeka, žabe, cebrice, napihovalka (*Takifugu rubripes*) in vinske mušice (*Drosophila melanogaster*).

Ustvarite mapo "proteoms".

Pojdite na spletno stran <https://www.ensembl.org/>, kjer je na splošno prvo mesto za iskanje proteomov. V razdelku „Favourite genomes“ kliknite na „Human“ (Človek).

OrthoFinder kot vhodne podatke zahteva aminokislinska zaporedja za vse gene, ki kodirajo beljakovine v vaši vrsti. Zaporedja za vsako vrsto morajo biti v ločeni datoteki s končnico „.fa“, „.faa“, „.fasta“, „.fas“ ali „.pep“. Ko je genom neke vrste sekvenciran in na voljo, se izvedeta dva glavna koraka, sestavljanje in anotacija. Sestavljanje je korak v katerem se sestavijo posamezni odčitki v zaporedje genoma. Anotacija je identifikacija zanimivih lastnosti v sestavi genoma, kot so geni, ki kodirajo beljakovine. Zato bodo datoteke, ki jih potrebujemo, pogosto v razdelku z imenom „annotation“. V programu Ensembl na desni strani pod „**Gene annotation**“ kliknite „**Download FASTA**“.





Human (GRCh38.p13) ▾

Search Human (*Homo sapiens*)

Search all categories ▾

Go

e.g. [BRCA2](#) or [17:63992802-64038237](#) or [rs699](#) or [osteoarthritis](#)

Genome assembly: GRCh38.p13 (GCA_000001405.28)



[More information and statistics](#)



[Download DNA sequence \(FASTA\)](#)



[Convert your data to GRCh38 coordinates](#)



[Display your data in Ensembl](#)

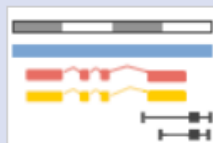
Other assemblies

GRCh37 Full Feb 2014 archive with BLAST, VEP and BioMart ▾

Go



[View karyotype](#)



[Example region](#)

Gene annotation

What can I find? Protein-coding and non-coding genes, splice variants, cDNA and protein sequences, non-coding RNAs.



[More about this genebuild](#)



[Download FASTA](#) files for genes, cDNAs, ncRNA, proteins



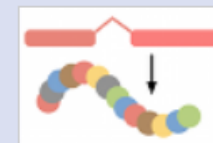
[Download GTF or GFF3](#) files for genes, cDNAs, ncRNA, proteins



[Update your old Ensembl IDs](#)



[Example gene](#)



[Example transcript](#)

Kliknite na mapo „pep“ (ki vsebuje zaporedja peptidov) in prenesite datoteko „Homo_sapiens.GRCh38.pep.all.fa.gz“ v mapo, ki ste jo ustvarili.

Vrnite se na glavno stran Ensembl in ponovite za mouse (*Mus musculus*), zebrafish (*Danio rerio*), tropical clawed frog (*Xenopus tropicalis*, pod 'Amphibians'), fugu (*Takifugu rubripes*, pod 'Fish') in fruit fly (*Drosophila melanogaster*, pod 'Other eukaryotes'. Če je na voljo izbira datotek, izberite datoteko '**.pep.all.fa.gz**'.

V terminal se pomaknemo v mapi ki vsebuje vse datoteke in jih razširimo z

gunzip *.gz

premaknemo se nazaj v mapi od prej

cd ..

Datoteke iz Ensembla bodo vsebovale veliko transkriptov na gen. Če bi OrthoFinder izvajali na teh neobdelanih datotekah, bi to trajalo 10x dlje, kot je potrebno, in bi lahko zmanjšalo natančnost. Uporabili bomo skripto, ki je priložena programu OrthoFinder, da izberemo samo najdaljšo različico zapisa na gen in zaženemo program OrthoFinder na teh datotekah:

```
for f in proteoms/*fa ; do python tools/primary_transcript.py $f ; done
```

Tudi skrajšanje imena datoteke je dobra ideja, saj ohranja rezultate urejene, ker se imena datotek uporabljajo za sklicevanje na vrste, npr. skrajšamo na Homo_sapiens.fa.

Če bi pognali Orthofinder, bi analiza trajala od 20 min do več ur odvisno od računalnika. Če želite pognati, ukaz je:















```
orthofinder -f proteoms/primary_transcripts
```

Za ogled rezultatov, bomo prenesli rezultate analize iz:

https://bioinformatics.plants.ox.ac.uk/davidemms/public_data/Results_model_species.tar.gz

Prvi pregled rezultatov

OrthoFinder privzeto ustvari mapo z rezultati z imenom 'OrthoFinder' znotraj vhodne mape proteomov in vanj postavi rezultate.

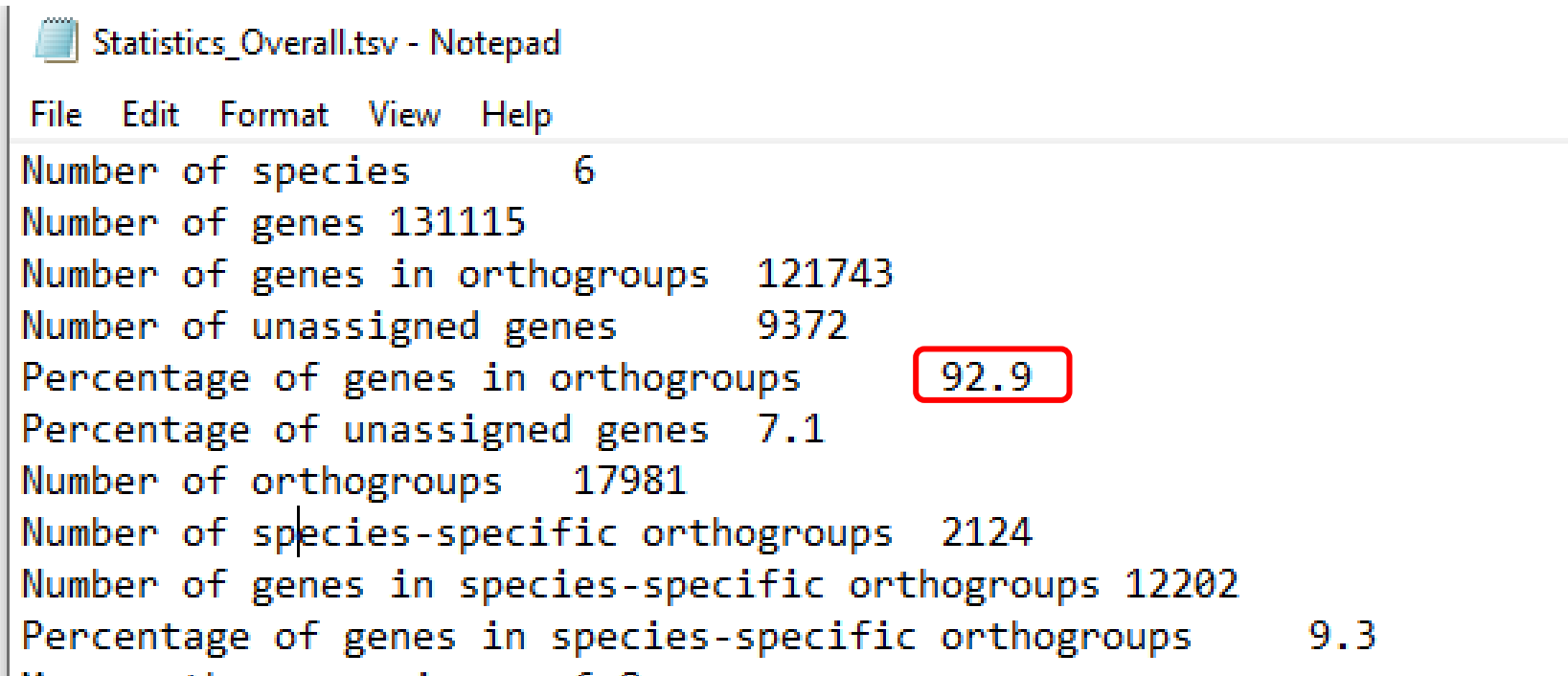
Name	▲	Size	Modified
 Comparative_Genomics_Statistics		10 items	14:15
 Gene_Duplication_Events		2 items	14:15
 Gene_Trees		13,340 items	14:15
 Orthogroups		5 items	14:10
 Orthogroup_Sequences		27,353 items	14:10
 Orthologues		6 items	14:14
 Phylogenetically_Misplaced_Genes		10 items	14:14
 Putative_Xenologs		6 items	14:14
 Resolved_Gene_Trees		13,340 items	14:15
 Single_Copy_Orthologue_Sequences		1,755 items	14:10
 Species_Tree		2 items	14:14
 WorkingDirectory		61 items	14:15
 Citation.txt		2.5 kB	14:17
 Log.txt		848 bytes	14:15

Quality control: odstotek genov v ortoskupinah

Ortoskupina je niz genov, ki izhajajo iz enega gena v zadnjem skupnem predniku vseh obravnavanih vrst.

Najprej želimo preveriti, koliko genov je bilo uvrščenih v ortoskupine. Ta informacija je v

Comparative_Genomics_Statistics/Statistics_Overall.tsv



```
Statistics_Overall.tsv - Notepad
File Edit Format View Help
Number of species      6
Number of genes 131115
Number of genes in orthogroups 121743
Number of unassigned genes 9372
Percentage of genes in orthogroups 92.9
Percentage of unassigned genes 7.1
Number of orthogroups 17981
Number of species-specific orthogroups 2124
Number of genes in species-specific orthogroups 12202
Percentage of genes in species-specific orthogroups 9.3
```

To je precej dobro, na splošno je dobro, če je vsaj 80 % genov uvrščenih v ortoskupine. Manj kot to pomeni, da verjetno manjkajo ortološka razmerja, ki dejansko obstajajo za nekatere od preostalih genov; najverjetnejši vzrok za to je slaba izbira vrst. Preverimo tudi odstotke na podlagi posamezne vrste. Najdeno v `Comparative_Genomics_Statistics/Statistics_PerSpecies.tsv`.

Tako kot druge datoteke „.tsv“ iz programa OrthoFinder je tudi to najbolje pregledati v programu za preglednice, kot sta Excel ali LibreOffice Calc. Te datoteke bodo morda v vašem računalniku pravilno obdelane samodejno ali pa mu boste morali izrecno povedati, da so tabelarno omejene. Tukaj je prikazano, kako to storite v programu LibreOffice.

Import

Character set: Unicode (UTF-8) ▼

Language: Default - English (UK) ▼

From row: 1 − +

Separator Options

☐ Fixed width ☒ Separated by

☒ Tab ☐ Comma ☐ Semicolon ☐ Space ☐ Other

☐ Merge delimiters String delimiter: " ▼

Other Options

☐ Format quoted field as text ☐ Detect special numbers

Fields

Column type: ▼

	Standard	Standard
1		Danio_rer
2	Number of genes	30313
3	Number of genes in orthogroups	28236
4	Number of unassigned genes	2077
5	Percentage of genes in orthogroups	93.1
6	Percentage of unassigned genes	6.9
7	Number of orthogroups containing species	13472
8	Percentage of orthogroups containing species	74.9

Help

OK

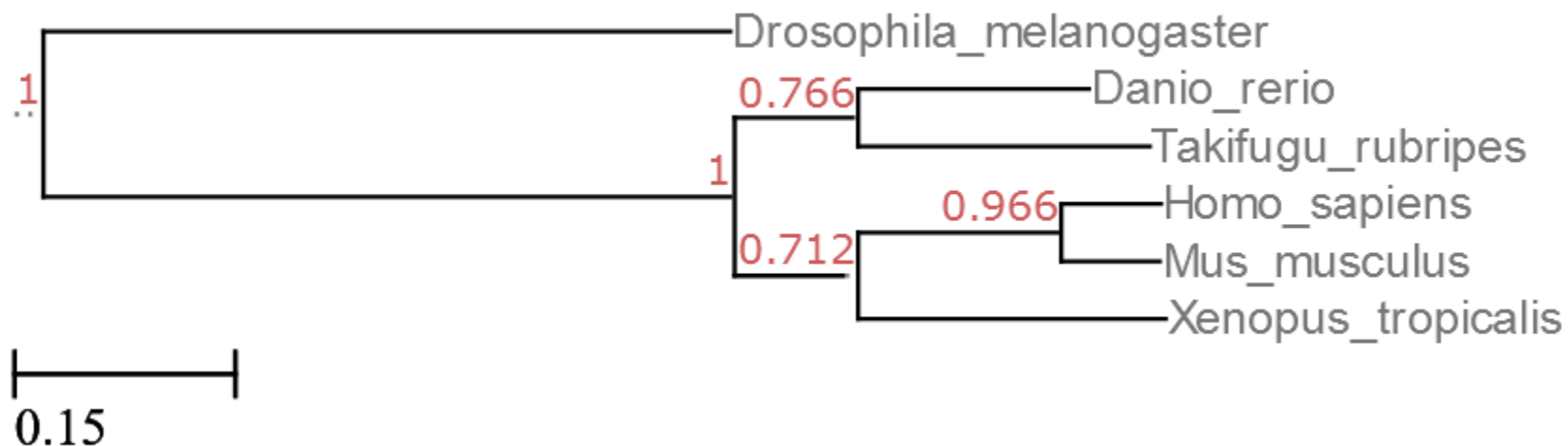
Cancel

Ko odpremo to datoteko, vidimo, da imajo vsi vretenčarji več kot 90% svojih genov dodeljenih v ortoskupine, medtem ko ima mušica dodeljenih približno 76% genov. To je verjetno posledica izbira vrst. Vse pet vrste vretenčarjev so si relativno sorodne, medtem ko je bila izbira vrst pri mušici slaba. Da bi to izboljšali, bi morali vključiti nekaj vrst, ki bi razbile dolge veje v drevesu vrst, ki ločujejo te vrste od vseh drugih.

Filogenetsko drevo vrst

Za to bomo uporabili spletno orodje **Phylogenetic tree (newick) viewer**, <http://etetoolkit.org/treeview/>.

Odprite datoteko Species_Tree/SpeciesTree_rooted.txt v orodju, ali pa odprite datoteko v Text editor in prilepite vrednosti. Kliknite na View tree!



Vidimo, da je mušica na daljši veji kot druge vrste. Če veste, kako naj bi bilo videti drevo vrst, preverite, ali se drevo ujema s tem, kar ste pričakovali. Drevo, ki ga je tukaj izpeljal program OrthoFinder, je pravilno.

Ortologi

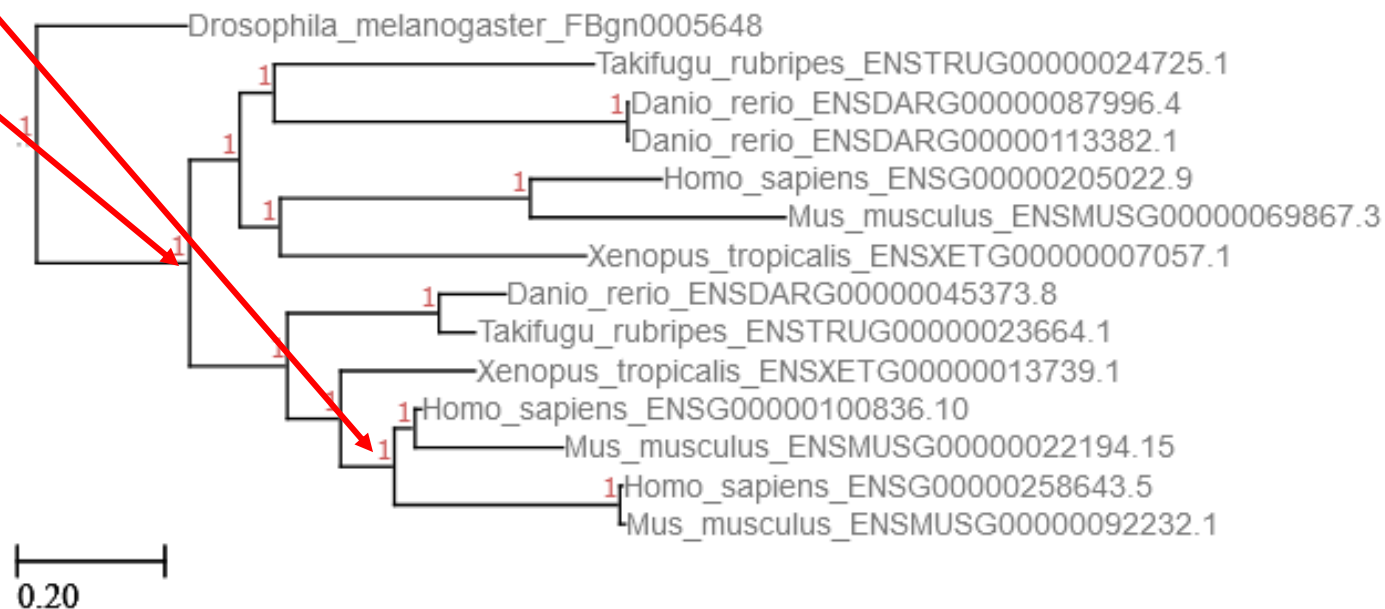
Eden najpogostejših razlogov za uporabo programa OrthoFinder je iskanje ortologa gena, ki nas zanima. Oglejmo si ortologe gena FBgn0005648 iz rodu *Drosophila*. To je zanimiv gen, saj je v liniji, ki vodi do človeka, prišlo do dveh podvojitev genov, kot bomo videli. Poglejmo, kateri so njegovi ortologi pri človeku.

V mapi Orthologues je podmapa za vsako vrsto. Odprite datoteko Orthologues/Orthologues_*Drosophila_melanogaster*/*Drosophila_melanogaster__v__Homo_sapiens*.tsv. Datoteka ima tri stolpce: „Orthogroup“, „*Drosophila_melanogaster*“ in „*Homo_sapiens*“. V preglednici poiščite „FBgn0005648“ in videli boste, da je gen v ortoskupini OG0001189 in da ima pri človeku tri ortologe: ENSG00000205022, ENSG00000100836, ENSG00000258643.

Drevesa genov

Ugotovili smo, da ima FBgn0005648 pri človeku tri ortologe. Oglejmo si gensko drevo in kako so ti trije ortologi nastali. Odprite Gene_Trees/OG0001189_tree.txt s programom s spletnim orodjem Tree viewer.

Če pogledamo gensko drevo, vidimo, da sta se zgodili dve podvojitvi genov, ena skupna vretenčarjem, druga pa človeku in miši. Posledica tega je ortološko razmerje ena proti trem, kar pomeni, da so vsi trije človeški geni enako tesno povezani z enim genom mušice.



Domača naloga 2

Vaša naloga je da ponovite analizo Orthofinder (stran/slide 9 naprej) in dodate še 1 vrsto sorodno z žabo (*Xenopus tropicalis*) in 3 vrste sesalcev.

Rezultat naloge naj bo v obliki poročila in naj vsebuje naslednje:

1. vpišite ukaze, ki ste jih uporabili.
2. odstotek genov v ortoskupinah
3. odstotek genov v ortoskupinah po vrsti
4. filogenetsko drevo vrst. Ali mislite, da je drevo pravilno ali ne? Zakaj?
5. drevo gena FBgn0005648 (mušica).
6. drevo poljubnega gena pri mušici, ki ima vsaj 3 ortologe pri človeku.
7. na obeh drevesih označite, kdaj/kje so se zgodile podvojitve genov.

Domačo nalogo 1 in Domačo nalogo 2 združite in oddate kot en dokument na e-učilnici: **Oddaja domačih nalog.**