Package “GePhCort”

Sept . 2013

Title : GePhCort - A genotype-phenotype correlation tool based on phylogenetic analysis.

Version : 1.0

Author : Dr. Farhat Habib, Amol Kolte

Maintainer : Amol Kolte <amolkolte1989@gmail.com>

Description : A purely Unix-based bioinformatics tool to locate phenotype associated genotypic markers (SNPs) by taking into account the phylogenetic relationship among the species. GePhCort has been built using R and python.

Dependencies:

python 2.7 (not 3.0)
ete (2.1 alpha)
python-tk
numpy (1.6.1)
scipy (0.9)
rpy (2.25)

R (>=2.14.0)
ape (<=2.8)
igraph
matrix (1.0)
quadprog (1.5)
phangorn (1.5)

Package Compilation: *GePhCort* do not require package compilation. It mainly uses 2 standalone scripts (resurrect.R and reanimate.py) :

- resurrect.R (Performs ancestral sequence reconstruction)
- reanimate.py (Performs phenotype reconstruction and permutation test)

Keep all the scripts under single location, make sure all the dependancies are satisfied

Running *GePhCort*

Step 1. Perform ancestral sequence reconstruction

```
resurrect.R ⟨sequence_file⟩ ⟨newick_tree_file⟩ ⟨fasta/phylip⟩ ⟨resurrect_output_file⟩
```

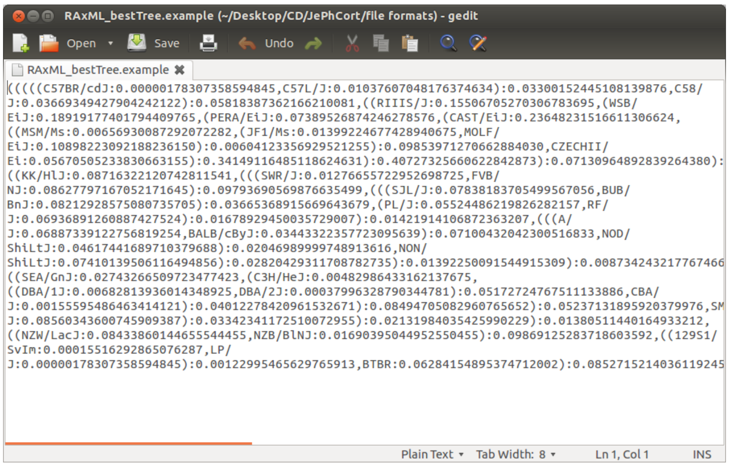
Step 2. Perform ancestral phenotype reconstruction and correlation

```
reanimate.py -s ⟨sequence_file⟩ -t ⟨newick_tree_file⟩ -f ⟨fasta/phylip⟩ -i ⟨Num_iterations_for_permutation_test⟩
-p ⟨phenotype_file⟩ -o ⟨output_file⟩ -r ⟨resurrect_output_file⟩
```

Files and formats : *GePhCort* obligatorily requires three user input text files. The proper formats are discussed below :

i) Sequence file – Nucleotide sequences (SNPs) can be submitted in the standard *fasta* or *phylip* format.

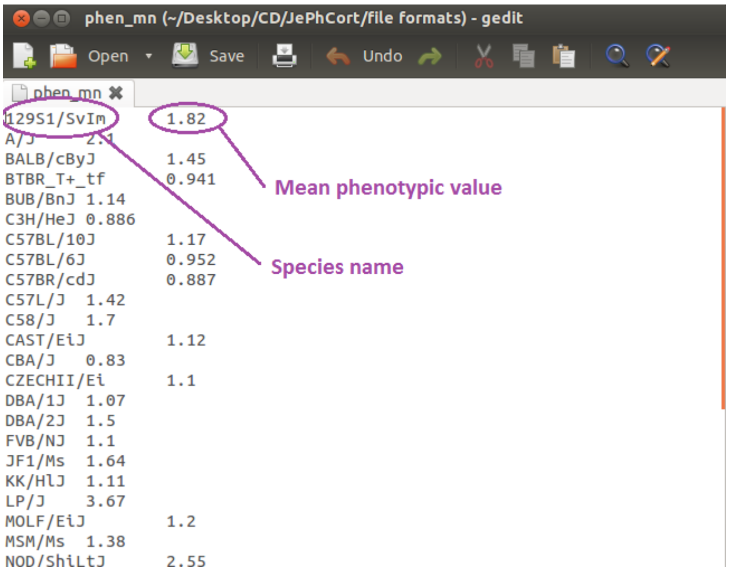
ii) Phylogeny/Tree file – In the standard *newick* format (with branch lengths) as shown in the adjacent figure.



```
(((((C57BR/cdJ:0.00000178307358594845,C57L/J:0.01037607048176374634):0.03300152445108139876,C58/
J:0.03669349427904242122):0.05818387362166210081,((RIIIS/J:0.15506705270306783695,(WSB/
EiJ:0.18919177401794409765,(PERA/EiJ:0.07389526874246278576,(CAST/EiJ:0.23648231516611306624,
((MSM/Ms:0.00656930087292072282,(JF1/Ms:0.01399224677428940675,MOLF/
EiJ:0.10898223092188236150):0.00604123356929521255):0.09853971270662884030,CZECHII/
Ei:0.05670505233830663155):0.34149116485118624631):0.40727325660622842873):0.07130964892839264380):
((KK/HlJ:0.08716322120742811541,(((SWR/J:0.01276655722952698725,FVB/
NJ:0.08627797167052171645):0.09793690569876635499,(((SJL/J:0.07838183705499567056,BUB/
BnJ:0.08212928575080735705):0.03665368915669643679,(PL/J:0.05524486219826282157,RF/
J:0.06936891260887427524):0.01678929450035729007):0.01421914106872363207,(((A/
J:0.06887339122756819254,BALB/cByJ:0.03443322357723095639):0.07100432042300516833,NOD/
ShiLtJ:0.04617441689710379688):0.02046989999748913616,NON/
ShiLtJ:0.07410139506116494856):0.02820429311708782735):0.01392250091544915309):0.00873424321776746
((SEA/GnJ:0.02743266509723477423,(C3H/HeJ:0.00482986433162137675,
((DBA/1J:0.00682813936014348925,DBA/2J:0.00037996328790344781):0.05172724767511133886,CBA/
J:0.00155595486463414121):0.04012278420961532671):0.08494705082960765652):0.05237131895920379976,SM
J:0.08560343600745909387):0.03342341172510072955):0.02131984035425990229):0.01380511440164933212,
((NZW/LacJ:0.08433860144655544455,NZB/BlNJ:0.01690395044952550455):0.09869125283718603592,((129S1/
SvIm:0.00015516292865076287,LP/
J:0.00000178307358594845):0.00122995465629765913,BTBR:0.06284154895374712002):0.08527152140361192
```

Figure 1: *Newick* file format

iii) Phenotype file – This is a simple tab-separated text file in a customized format. The first column in two-column file format represents 'name of the species' and other represents the 'continuous phenotypic value' (as shown in Figure 3).



```
129S1/SvIm      1.82
A/J     2.1
BALB/cByJ       1.45
BTBR_T+_tf      0.941
BUB/BnJ 1.14
C3H/HeJ 0.886
C57BL/10J       1.17
C57BL/6J        0.952
C57BR/cdJ       0.887
C57L/J  1.42
C58/J   1.7
CAST/EiJ        1.12
CBA/J   0.83
CZECHII/Ei      1.1
DBA/1J  1.07
DBA/2J  1.5
FVB/NJ  1.1
JF1/Ms  1.64
KK/HlJ  1.11
LP/J    3.67
MOLF/EiJ        1.2
MSM/Ms  1.38
NOD/ShiLtJ      2.55
```

Figure 2: Two-column phenotype file format

Result Interpretation : *GePhCort* generates a tab-separated text file once the operation is successfully completed. The file consists of five entities, which are explained below.

1. **SNP serial ID** – Every row in the result file represents a genotypic marker *i.e.* an SNP. They are alloted serial numbers (starting from *Zero*) as per their position in the original sequence file. Thus, the n^{th} SNP serial ID in the result represents the $(n+1)^{th}$ SNP in the original file.

2. **p-value** – Lower the p-value, higher is the significance of association between a given SNP and the phenotype. This is the raw p-value.

3. **p.adjusted_FDR** – Corrected p-value after FDR correction.

4. **p.adjusted_Bonferroni** – Corrected p-value after Bonferroni correction.