Using the Taxy tool – a short introduction

Inferring the taxonomic profile of a microbial community from a large collection of anonymous DNA sequencing reads is a challenging task in metagenomics. Because existing methods for taxonomic profiling of metagenomes are all based on the assignment of fragmentary sequences to phylogenetic categories, the accuracy of results largely depends on fragment length. This dependency complicates comparative analysis of data originating from different sequencing platforms or resulting from different preprocessing pipelines. We have developed a read length independent method for taxonomic profiling and we provide the freely available tool Taxy which includes an ultra-fast implementation of that method. Our tests indicate that Taxy results compare well with taxonomic profiles obtained with other methods. However, in contrast to the existing methods, Taxy provides a nearly constant profiling accuracy across all kinds of read lengths and it operates at an unrivaled speed. Currently Taxy is available for Windows operating systems (XP, Vista, 7) and even runs on notebooks with at least 1GB of memory. As input, multiple FASTA DNA sequence files of any size can be used for the estimation of metagenome profiles. The analysis of a large sequence file with a Gbp volume typically requires less than a minute of processing time.

Minimal requirements:

Windows XP, Vista or Windows 7
1 GB RAM
500 Mb of free hard drive space

Recommended:
Windows XP, or Windows 7 (64-bit)
2 GB of RAM

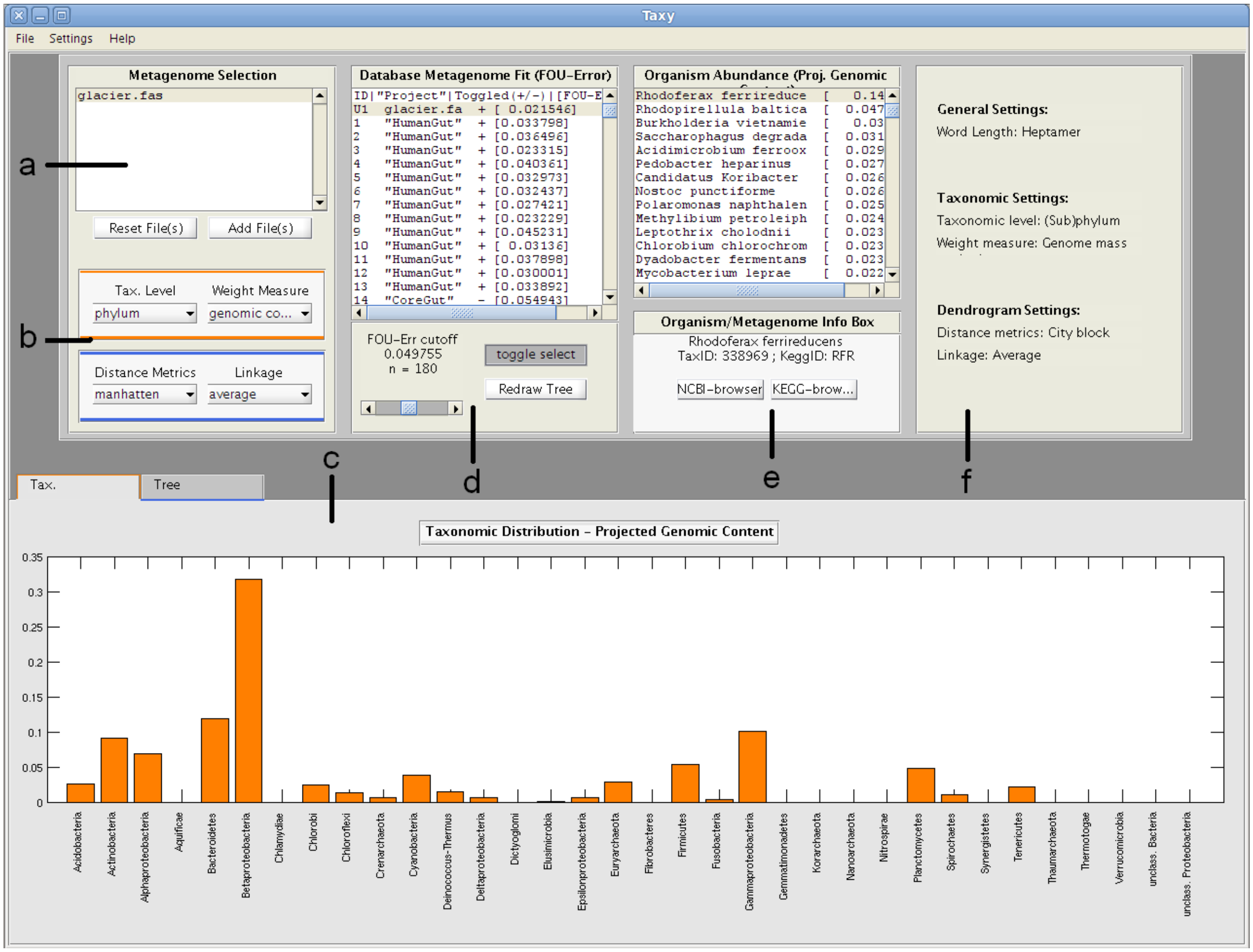
Installation:

To install Taxy, run “Taxy_installer.exe”. Taxy requires the MATLAB Compiler Runtime (MCR). The Taxy installer will automatically launch a packaged MCR version 7.8. installer. In the unlikely case that an MCR version 7.8. is already installed on the system, we recommend to re-install it, using the Taxy packaged MCR installer. In case another version of the MCR is already installed on the system, we recommend to install the Taxy packaged MCR version 7.8. beside it, in order to minimize any conflict with other MCR dependent programs.

Authors:

P. Gumrich and P. Meinicke. Georg August Unversity of Goettingen, Institute of Microbiology and Genetics, Department of Bioinformatics.

User Interface - Quick Overview:



```
glacier.fas
```

```
ID|"Project"|Toggled(+/-)|[FOU-E
U1   glacier.fa   + [ 0.021546]
1    "HumanGut"   + [0.033798]
2    "HumanGut"   + [0.036496]
3    "HumanGut"   + [0.023315]
4    "HumanGut"   + [0.040361]
5    "HumanGut"   + [0.032973]
6    "HumanGut"   + [0.032437]
7    "HumanGut"   + [0.027421]
8    "HumanGut"   + [0.023229]
9    "HumanGut"   + [0.045231]
10   "HumanGut"   + [ 0.03136]
11   "HumanGut"   + [0.037898]
12   "HumanGut"   + [0.030001]
13   "HumanGut"   + [0.033892]
14   "CoreGut"    - [0.054943]
```

FOU-Err cutoff
0.049755
n = 180

```
Rhodoferax ferrireduce   [    0.14
Rhodopirellula baltica   [   0.047
Burkholderia vietnamie   [    0.03
Saccharophagus degrada   [   0.031
Acidimicrobium ferroox   [   0.029
Pedobacter heparinus     [   0.027
Candidatus Koribacter    [   0.026
Nostoc punctiforme       [   0.026
Polaromonas naphthalen   [   0.025
Methylibium petroleiph   [   0.024
Leptothrix cholodnii     [   0.023
Chlorobium chlorochrom   [   0.023
Dyadobacter fermentans   [   0.023
Mycobacterium leprae     [   0.022
```

Rhodoferax ferrireducens
TaxID: 338969 ; KeggID: RFR

General Settings:
Word Length: Heptamer

Taxonomic Settings:
Taxonomic level: (Sub)phylum
Weight measure: Genome mass

Dendrogram Settings:
Distance metrics: City block
Linkage: Average

Keywords in **bold,** see Glossary.

a) Metagenome selection: User supplied metagenomes can be loaded via “**Add File(s)”** or “**Reset File(s)”**. All user supplied metagenomes will be listed in panel a). Single clicking a metagenome will display the taxonomic profile prediction of the selected metagenome and the clustering of taxonomic profiles in panel c). Furthermore, the weights of the database organisms will be displayed in panel e).
b) Analysis settings: In this panel, the user can select the parameters for the taxonomic profile and taxonomic profile clustering analysis. The Parameters of the taxonomic profile analysis are: “**Tax. Level”** and “**Weight measure”.** The parameters of the taxonomic profile clustering are “**Distance Metric**” and “**Linkage Parameter”** .
c) Main analysis panel: The taxonomic distribution of a selected metagenome and clustering of taxonomic profiles of selected metagenomes are displayed in the “**Tree**” and “**Tax”** tabs, respectively. For more information on the selection of metagenomes for taxonomic profile clustering see d).
d) Database metagenome selection: For comparison, 256 database metagenomes are listed in this panel. The FOU measures the fraction of metagenomic DNA that cannot be explained by the mixture of genomic signatures from the 1013 KEGG database organisms. A lower FOU-error corresponds to a

higher explanatory value of the database organisms for the metagenome taxonomic composition. Double clicking a database metagenome will display the taxonomic profile and taxonomic profile clustering analysis in panel c). Single clicking a database metagenome will display the metagenome meta-info in panel e). Metagenomes marked with a "+" are included in the clustering of taxonomic profiles. By default any database metagenome with a FOU-error below 0.05 is included in the clustering. The set of metagenomes included in profile clustering can be altered individually via "**toggle select**" or by adjusting the "**FOU-error cutoff"** . The clustering analysis can be refreshed via "**Redraw Tree"**.

e) Organism abundance and meta-info: The top panel displays a list of all database organisms and their predicted weight within the currently selected metagenome, while the bottom panel lists information about the selected database metagenomes or database organisms

f) Settings overview: All user determined settings are displayed here.

File Menu: "**Add file(s)**" and "**Reset file(s)**" provide an analogous function as in a). "**Export organism weight**s" and "**Export tax distribution**" allow the user to export the taxonomic distribution on the organism level and chosen taxonomic level, respectively.

Settings Menu: "**Set word length**" allows the user to change the oligonucleotide length used for the taxonomic profile prediction.

Additional Information

Database metagenomes can be viewed and a cluster analysis performed prior to loading an user supplied metagenome. However, when selecting reference genomes manually, at least two reference metagenomes have to be included in the clustering set.

A typical use case:

1. User Anne starts Taxy. She decides to change the oligonucleotide length from the default 7-mer to 6-mer under via "**Set word length**" in the **Settings Menu.**

2. Anne then reduces the **Database metagenome selection** by moving the "**FOU-error cutoff"** slider. After the threshold augmentation, she add a couple of individual database metagenomes by activating "**toggle select**" and single clicking the metagenomes she wishes to include. After adding the database metagenomes to the clustering set, she deactivates "**toggle select**".

3. To inspect the clustering before the addition of her own metagenome, she can either double clicks on a database metagenome within the **Database metagenome selection** panel, or alternatively she could have activated the "**Redraw Tree"** button. Because she is not interested in any specific database metagenome, she chooses to press "**Redraw Tree".**

4. After inspecting the taxonomic profile clustering under the "**Tree"** tab in the **Main analysis panel** Anne decides to switch the clustering linkage parameter from average to complete in the "**Linkage Parameter"** drop down menu within the **Analysis settings.**

5. Anne then loads her metagenome via **"Add File(s)"** in the **Metagenome selection.**

6. Anne inspects the taxonomic distribution under the "**Tax**." and the taxonomic profile clustering under the "**Tree**" tab in the **Main analysis panel.**

7. To inspect the meta-data of the database metagenomes clustered close to her own metagenome she single clicks the appropriate database metagenomes within the **Metagenome selection,** with the information displayed in information box of the **Organism abundance and meta-info** panel.

8. Anne double clicks the closest database metagenomes successively to view the taxonomic distribution and exports the species weights via "**Export organism weight**s" in the **File Menu.**

9. Anne then re-selects her own metagenome by single clicking within the **Metagenome selection**, or by double clicking her own metagenome which has been added to the reference metagenome list in the **Database metagenome selection** panel.

10. Anne compares the organism weight distribution of her own metagenome to the database metagenomes and exports it for further analysis. Within the **Organism abundance panel**, she single clicks an organism of interest and clicks on the KEGG link in the info box for further information about the organism.

Glossary:

Add File(s): Opens a single or group of files for metagenomic analysis. Files must be in the multiple FASTA format and contain DNA sequence data only. *[panel a]*

Distance Metrics: Selects the distances metrics used in the clustering of taxonomic profiles. Possible settings: Manhattan distance and Euclidean distance. *[panel b]*

Export organism weights: Exports a list in .txt format of predicted organism weights *[File Menu]*

Export tax. Distribution: Exports a list in .txt format of predicted taxon frequencies at the selected taxonomic level. *[File Menu]*

Linkage Parameter: Selects the linkage parameter for the hierarchical clustering of taxonomic profiles. Possible settings: average linkage and complete linkage *[panel b]*

FOU-error: A measure of confidence, which characterizes how well the 1013 KEGG database organisms explain the metagenome composition. *[panel d]*

FOU-error cutoff: Determines a FOU-error threshold below which database metagenomes are included in the clustering analysis. All user loaded metagenomes are included in the analysis, regardless of FOU-error. *[panel d]*

Redraw Tree: After an augmentation of the database metagenome set for the taxonomic profile clustering analysis, this button re-draws the taxonomic profile clustering tree. *[panel d]*

Reset File(s): Clears the user metagenome file list and loads new metagenomes. *[panel a]*

Set word length: Sets the oligonucleotide length used for the taxonomic profile prediction. Possible settings are 6, 7 and 8-mer. The 8-mer analysis requires at least 1 Gb of memory and a 64 bit operating system. The default setting is 7-mer.*[Setting Menu]*

Tax.: Displays the taxonomic profile prediction at the selected taxonomic level. *[panel c]*

Tax. Level: Sets the taxonomic level at which the taxonomic profile prediction is displayed. Possible settings: phylum, class, order. *[panel b]*

Toggle select: When activated allows the user to (de)select individual database genomes for inclusion in the taxonomic profile clustering analysis. *[panel d]*

Tree: Displays the clustering of predicted taxonomic profiles. *[panel c]*

Weight measure: Selects the weight measure of the taxonomic profile prediction. The primary weight measure of the Taxy taxonomic prediction is genomic content. The secondary weight measure is derived from the genomic content and measures genome/organism count. *[panel b]*