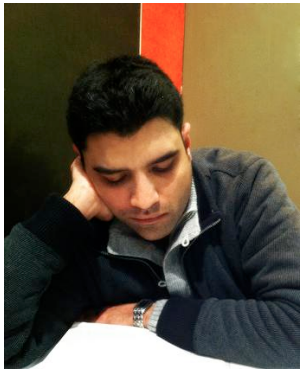
TAXAassign

Tutorial v0.2

Project Team

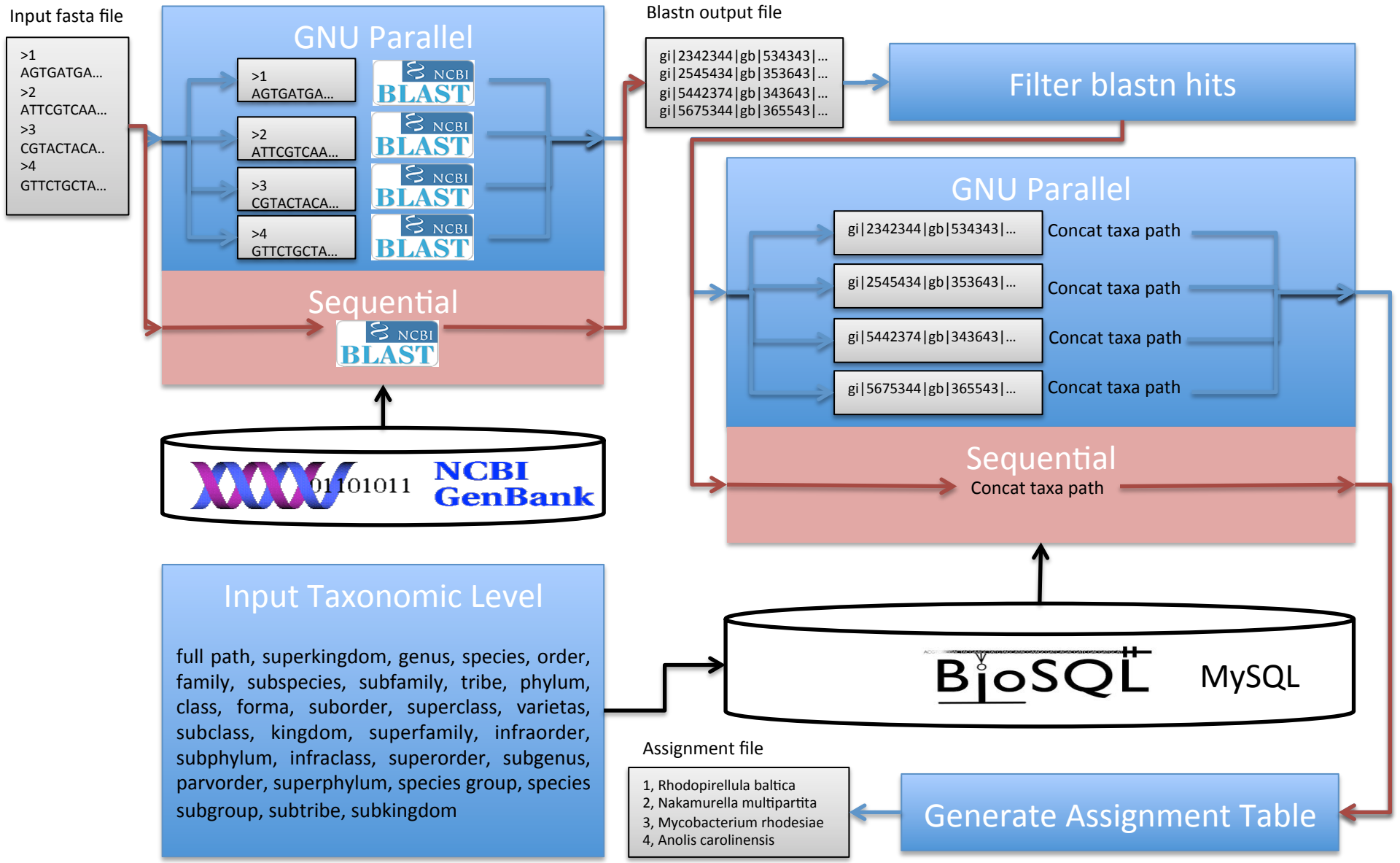
Umer Zeeshan Ijaz
Research Fellow (Infrastructure and Environment)
University of Glasgow, School of Engineering, Glasgow
http://userweb.eng.gla.ac.uk/umer.ijaz/



Christopher Quince
Reader in Biological Systems Modelling
University of Glasgow, School of Engineering, Glasgow
http://userweb.eng.gla.ac.uk/christopher.quince/

Acknowledgements: Umer Zeeshan Ijaz's work is supported by a Technology Strategy Board (TSB) funded research grant "Development of instrumental and bioinformatic pipelines to accelerate commercial applications of metagenomics approaches" shared between a team of academics (Centre for Genomic Reseach, University of Liverpool and Computational Microbial Genomics Group, University of Glasgow) and commerical experts (Unilever (lead), Skylene and BioControl).

TAXAassign Pipeline



TAXAassign v0.2 Features

- The taxonomic assignment is resolved using NCBI's Taxonomy and running NCBI's Blast against locally-installed NCBI's nt database to minimize execution time.
- Version 0.2 has many orders of magnitude improvement in speed over 0.1.
- It also supports taxonomic assignments at user-specified taxonomic level.
- To minimize the execution time, we use GNU Parallel, a shell tool for executing jobs in parallel on multicore computers. We split the sequence file into fixed size chunks and then run blastn in parallel on these chunks on separate cores. For a 16SrRNA dataset comprising 1000 most abundant OTU sequences, matching at most 100 reference sequences against a local NCBI's NT database took 18.9 minutes on 45 cores. A speedup of 30 times or more is achieved this way.
- To find Genbank ID to Taxa ID mapping, one can use gi_taxid_nucl.dmp.gz from ftp://ftp.ncbi.nih.gov/pub/taxonomy/ which is updated every Monday around 2am EST on NCBI's website. However, searching Taxa IDs through this beastly 4.4GB+ unzipped file is time consuming and slows down the pipeline as noticed in the previous version. To improve the performance, instead of using this file, we use the latest version of NCBI blast 2.28+ which also gives Taxa ID for each hit and so we skip this step altogether in this version.
- To improve the time for finding parent Taxa IDs for a given Taxa ID, instead of using the flat file taxdump.tar.gz from the above ftp site, we use BioSQL and host NCBI's taxonomy data in a local MySQL server. Furthermore, we use GNU Parallel to run multiple SQL queries in parallel on multiple records of blast output file thus reducing the execution time significantly
- Both NCBI's taxonomy database on MySQL server and local nt database can be updated frequently by submitting a cron job on the server scheduled to run when the server is less busy i.e. at night time and thus the information does not get outdated.
- All the time consuming steps in the TAXAassign pipeline are not repeated on reruns. Thus, if pipeline has finished processing the data and you want to run it again with a different taxonomic level, it will skip blasting the fasta file. To see what you have already done, output gets logged in TAXAassign.log file in the current folder. This is useful for debugging, should a problem arise.
- If you have the OTUs abundance table (in csv or tsv format) along with OTUs sequences, the output file generated from TAXAassign can then be used with http://userweb.eng.gla.ac.uk/umer.ijaz/bioinformatics/convIDs.pl to annotate the table.
- For better understanding of BioSQL, refer to my tutorial http://userweb.eng.gla.ac.uk/umer.ijaz/bioinformatics/BIOSQL_tutorial.pdf

Example Usage

TAXAassign pipeline uses only one command to find the assignments for list of unknown sequences given in FASTA format. These sequences can originate from both whole shot-gun sequencing or amplicon sequencing experiments (to identify contaminants). The pipeline can operate in both parallel as well as sequential modes by specifying –p switch. You can also specify the quality of blastn hits using –m and –q switches as well as select the total number of reference matches per query sequence using –r switch.

```
[uzi@quince-srv2 ~]$ bash /home/opt/TAXAassign_v0.2/TAXAassign.sh
Script to annotate sequences at different taxonomic levels using  NCBI's taxonomy

Usage:
    bash TAXAassign.sh -f <fasta_file.fasta> [options]
Options:
    -p Turn parallel processing on
    -c Number of cores to use (Default: 10)
    -r Number of reference matches (Default: 10)
    -m Minimum percentage ident in blastn (Default: 97)
    -q Minimum query coverage in blastn (Default: 97)
    -l Taxonomic resolution (Default: species)
       Supported levels:
              full,superkingdom,genus,species,order,family,subspecies,subfamily,tribe,
              phylum,class,forma,suborder,superclass,varietas,subclass,kingdom,superfamily,
              infraorder,subphylum,infraclass,superorder,subgenus,parvorder,superphylum,
              species group,species subgroup,subtribe,subkingdom
```

Example dataset(1)

To test TAXAassign, we will use a small dataset test.fasta comprising 10 unknown sequences. First four sequences are as follows:

```
[uzi@quince-srv2 ~/test]$ head -20 test.fasta
>seq1
TACGAAGGgggCTAGCGTTGCTCGGAATTACTGGGCGTAAAGGGAGCGTAGGCGGACATTTAAGTCAGGGGTGAAATCCC
GGGGCTCAACCTCGGAATTGCCTTTGATACTGGGTGTCTTGAGTATGAgagagGTgtgtgGAACTCCGAGTGTAGAGGTG
AAATTCGTAGATATTCGGAAGAACACCAGTGGCGAAGGCGACacactGGCTCATTACTGACGCTGAGGCTCGAAAGCGTG
GGGAGCAAACAGG
>seq2
TACGTAGGGTGCAAGCGTTGTCCGGAATTACTGGGCGTAAAGAGTTCGTAGGCGGTTTGTCGCGTCGTTTGTGAAAACCA
GCAGCTCAACTGCTGGCTTGCAGGCGATACGGGCAGACTTGAGTACTGCAGGGGAGACTGGAATTCCTGGTGTAGCGGTG
AAATGCGCAGATATCAGGAGGAACACCGGTGGCGAAGGCGGGTCTCTGGGCAGTAACTGACGCTGAGGAACGAAAGCGTG
GGTAGCGAACAGG
>seq3
TACAGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCgcgcgTAGGTGGTTAGTTAAGTTGGATGTGAAATCCC
CGGGCTCAACCTGGGAACTGCATTCAAAACTGACTGACTAGAGTATGGTAGAGGGTGGTGGAATTTCCTGTGTAGCGGTG
AAATGCGTAGATATAGGAAGGAACACCAGTGGCGAAGGCGACCACCTGGACTGATACTGACACTGAGGTGCGAAAGCGTG
GGGAGCAAACAGG
>seq4
TACGGAGGGTGCGAGCGTTAATCGGAATCACTGGGCGTAAAGCGCACGTAGGCTGCTTGGTAAGTCAGGGGTGAAAGCCC
GCGGCTCAACCGCGGAATTGCCTTTGATACTGCCGAGCTAGAGTCCGGGAGAGGGTAGTGGAATTCCAGGTGTAGGAGTG
AAATCCGTAGAGATCTGGAGGAACATCAGTGGCGAAGGCGACTACCTGGACCGGTACTGACGCTGAGGTGCGCAAGCGTG
GGGAGCAAACAGG
```

Example dataset(2)

First, we check the assignments at “genus” level.

```
[uzi@quince-srv2 ~/test]$ bash /home/opt/TAXAassign_v0.2/TAXAassign.sh -m 99 -q 99 -f test.fasta -l genus
TAXAassign v0.2. Copyright (c) 2013 Computational Microbial Genomics Group, University of Glasgow, UK
[2013-05-16 02:56:19] Using  /home/opt/ncbi-blast-2.2.28+/bin/blastn
[2013-05-16 02:56:19] Using  /home/opt/TAXAassign_v0.2/scripts/blast_concat_taxon.py
[2013-05-16 02:56:19] STEP 1: Blast against NCBI's nt database with minimum percent ident of 99, maximum of 10
reference sequences, and evalue of 0.0001 in blastn.
[2013-05-16 02:56:30] blastn took 11 seconds for test.fasta.
[2013-05-16 02:56:30] test_B.out generated successfully!
[2013-05-16 02:56:30] STEP 2: Filter blastn hits with minimum query coverage of 99.
[2013-05-16 02:56:30] test_BF.out generated successfully!
[2013-05-16 02:56:30] STEP 3: Annotate blastn hits with NCBI's taxonomy data at genus level
[2013-05-16 02:56:30] test_BFT.out generated successfully!
[2013-05-16 02:56:30] STEP 4: Generate taxonomic assignments table from blastn hits.
[2013-05-16 02:56:30] test_ASSIGNMENTS.csv generated successfully!
[2013-05-16 02:56:30] SUMMARY: Reads assigned at genus level are 10/10.
[uzi@quince-srv2 ~/test]$ cat test_ASSIGNMENTS.csv
seq1,Brevundimonas
seq2,Rhodococcus
seq3,Pseudomonas
seq10,Microbacterium
seq4,Desulfovibrio
seq5,Pedobacter
seq6,Pseudomonas
seq7,Pseudomonas
seq8,Brevundimonas
seq9,Ralstonia
```

Example dataset(3)

Next, we will count how many reads were assigned to unique taxa.

```
[uzi@quince-srv2 ~/test]$ cat test_ASSIGNMENTS.csv | awk -F"," '{print $2}' | sort | uniq -c
      2 Brevundimonas
      1 Desulfovibrio
      1 Microbacterium
      1 Pedobacter
      3 Pseudomonas
      1 Ralstonia
      1 Rhodococcus
```

Next, we will check the taxonomic assignments at “species” level only to find two hits, which are less than we expected.

```
[uzi@quince-srv2 ~/test]$ bash /home/opt/TAXAassign_v0.2/TAXAassign.sh -m 99 -q 99 -f test.fasta -l species
TAXAassign v0.2. Copyright (c) 2013 Computational Microbial Genomics Group, University of Glasgow, UK
[2013-05-16 03:02:42] Using  /home/opt/ncbi-blast-2.2.28+/bin/blastn
[2013-05-16 03:02:42] Using  /home/opt/TAXAassign_v0.2/scripts/blast_concat_taxon.py
[2013-05-16 03:02:42] STEP 1: Blast against NCBI's nt database with minimum percent ident of 99, maximum of 10
reference sequences, and evalue of 0.0001 in blastn.
[2013-05-16 03:02:42] test_B.out already exists. Skipping this step.
[2013-05-16 03:02:42] STEP 2: Filter blastn hits with minimum query coverage of 99.
[2013-05-16 03:02:42] test_BF.out already exists. Skipping this step.
[2013-05-16 03:02:42] STEP 3: Annotate blastn hits with NCBI's taxonomy data at species level
[2013-05-16 03:02:42] test_BFT.out generated successfully!
[2013-05-16 03:02:42] STEP 4: Generate taxonomic assignments table from blastn hits.
[2013-05-16 03:02:42] test_ASSIGNMENTS.csv generated successfully!
[2013-05-16 03:02:42] SUMMARY: Reads assigned at species level are 2/10.
[uzi@quince-srv2 ~/test]$ cat test_ASSIGNMENTS.csv
seq4,Desulfovibrio vulgaris
seq9,Ralstonia solanacearum
```

To understand better, we resolve the full path using –l full. The resulting file is shown on next page

Example dataset(4)

Due to lack of space here, details are given on the next slide.

```
[uzi@quince-srv2 ~/test]$ bash /home/opt/TAXAassign_v0.2/TAXAassign.sh -m 99 -q 99 -f test.fasta -l full
[uzi@quince-srv2 ~/test]$ cat test_ASSIGNMENTS.csv
seq1,cellular organisms;Bacteria;environmental samples;uncultured bacterium;;cellular organisms; Bacteria; Proteobacteria; Alphaproteobacteria; Caulobacterales; Caulobacteraceae; Brevundimonas; Brevundimonas nasdae;; cellular organisms; Bacteria; Proteobacteria; Alphaproteobacteria; Caulobacterales; Caulobacteraceae; Brevundimonas; Brevundimonas sp. 7-8;; cellular organisms; Bacteria; Proteobacteria; Alphaproteobacteria; Caulobacterales; Caulobacteraceae; Brevundimonas;Brevundimonas sp. JNU-L071;; cellular organisms; Bacteria; Proteobacteria; Alphaproteobacteria; Caulobacterales; Caulobacteraceae; Brevundimonas; Brevundimonas vesicularis;; cellular organisms; Bacteria; Proteobacteria; Alphaproteobacteria; Caulobacterales; Caulobacteraceae; Brevundimonas; environmental samples; uncultured Brevundimonas sp.;; cellular organisms; Bacteria; Proteobacteria; environmental samples; uncultured proteobacterium
seq2,cellular organisms; Bacteria; Actinobacteria; Actinobacteria; Actinobacteridae; Actinomycetales; Corynebacterineae; Nocardiaceae; Rhodococcus; Rhodococcus erythropolis;; cellular organisms; Bacteria; Actinobacteria; Actinobacteria; Actinobacteridae; Actinomycetales; Corynebacterineae; Nocardiaceae; Rhodococcus; Rhodococcus sp. NSi17;; cellular organisms; Bacteria; environmental samples; uncultured bacterium;; unclassified sequences; environmental samples; prokaryotic environmental samples; uncultured prokaryote
seq3,cellular organisms; Bacteria; environmental samples; uncultured bacterium;; cellular organisms; Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas; Pseudomonas fluorescens group; Pseudomonas brenneri;; cellular organisms; Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas; Pseudomonas fluorescens group; Pseudomonas fluorescens;; cellular organisms; Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas; Pseudomonas sp. 33(2013);; cellular organisms; Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas; Pseudomonas sp. JR37;; cellular organisms; Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas; Pseudomonas sp. P080
seq10,cellular organisms; Bacteria; Actinobacteria; Actinobacteria; Actinobacteridae; Actinomycetales; Micrococcineae; Microbacteriaceae; Microbacterium; Microbacterium laevaniformans
seq4,cellular organisms; Bacteria; environmental samples; uncultured bacterium;; cellular organisms; Bacteria; Proteobacteria; delta/epsilon subdivisions; Deltaproteobacteria; Desulfovibrionales; Desulfovibrionaceae; Desulfovibrio; Desulfovibrio sp. BDN100T;; cellular organisms; Bacteria; Proteobacteria; delta/epsilon subdivisions; Deltaproteobacteria; Desulfovibrionales; Desulfovibrionaceae; Desulfovibrio; Desulfovibrio sp. BH3;; cellular organisms;Bacteria; Proteobacteria;delta/epsilon subdivisions; Deltaproteobacteria; Desulfovibrionales; Desulfovibrionaceae; Desulfovibrio; Desulfovibrio sp. BH8;; cellular organisms; Bacteria; Proteobacteria;delta/epsilon subdivisions; Deltaproteobacteria; Desulfovibrionales; Desulfovibrionaceae; Desulfovibrio; Desulfovibrio sp. M21;; cellular organisms; Bacteria; Proteobacteria; delta/epsilon subdivisions; Deltaproteobacteria; Desulfovibrionales; Desulfovibrionaceae; Desulfovibrio; Desulfovibrio vulgaris;Desulfovibrio vulgaris DP4;; cellular organisms; Bacteria;Proteobacteria; delta/epsilon subdivisions; Deltaproteobacteria; Desulfovibrionales; Desulfovibrionaceae; Desulfovibrio; Desulfovibrio vulgaris; Desulfovibrio vulgaris RCH1;; cellular organisms; Bacteria; Proteobacteria; delta/epsilon subdivisions; Deltaproteobacteria; Desulfovibrionales; Desulfovibrionaceae; Desulfovibrio; Desulfovibrio vulgaris; Desulfovibrio vulgaris str. Hildenborough;; cellular organisms; Bacteria; Proteobacteria; delta/epsilon subdivisions; Deltaproteobacteria; Desulfovibrionales; Desulfovibrionaceae; Desulfovibrio; environmental samples; Desulfovibrio sp. enrichment culture clone VN_TX2-2;; cellular organisms; Bacteria; Proteobacteria; delta/epsilon subdivisions; Deltaproteobacteria; Desulfovibrionales; Desulfovibrionaceae; Desulfovibrio; environmental samples; uncultured Desulfovibrio sp.
seq5,cellular organisms; Bacteria; Bacteroidetes/Chlorobi group; Bacteroidetes; Sphingobacteriia; Sphingobacteriales; Sphingobacteriaceae; Pedobacter;environmental samples; uncultured Pedobacter sp.
seq6,cellular organisms; Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas; Pseudomonas sp. BA-75-09;; cellular organisms; Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas; Pseudomonas sp. P13-1;; cellular organisms; Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas; Pseudomonas sp. P28-2;; cellular organisms; Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas; Pseudomonas sp. PB2H;; cellular organisms; Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas; Pseudomonas sp. PPR-1;; cellular organisms; Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas; Pseudomonas sp. PTAS6;; cellular organisms; Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas; Pseudomonas sp. PY2;; cellular organisms; Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas; Pseudomonas sp. PY3;; cellular organisms; Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas; Pseudomonas sp. UA-JF3003;; cellular organisms; Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas; Pseudomonas sp. UA-JF3203
seq7,cellular organisms; Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas; Pseudomonas sp. BA-75-09;; cellular organisms; Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas; Pseudomonas sp. P13-1;; cellular organisms; Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas; Pseudomonas sp. P28-2;; cellular organisms; Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas; Pseudomonas sp. PB2H;; cellular organisms; Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas; Pseudomonas sp. PPR-1;; cellular organisms; Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; roteobacteria; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas; Pseudomonas sp. PY3;; cellular organisms; Bacteria; Proteobacteria; GamPseudomonas; Pseudomonas sp. PTAS6;; cellular organisms; Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas; Pseudomonas sp. PY2;; cellular organisms; Bacteria; Pmaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas; Pseudomonas sp. UA-JF3003;; cellular organisms; Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas; Pseudomonas sp. UA-JF3203
seq8,cellular organisms; Bacteria; environmental samples; uncultured bacterium;; cellular organisms; Bacteria; Proteobacteria; Alphaproteobacteria; Caulobacterales; Caulobacteraceae; Brevundimonas; Brevundimonas nasdae;; cellular organisms; Bacteria; Proteobacteria; Alphaproteobacteria; Caulobacterales; Caulobacteraceae;Brevundimonas; Brevundimonas sp. 7-8;; cellular organisms; Bacteria; Proteobacteria; Alphaproteobacteria; Caulobacterales; Caulobacteraceae; Brevundimonas; Brevundimonas sp. JNU-L071;; cellular organisms; Bacteria; Proteobacteria; Alphaproteobacteria; Caulobacterales; Caulobacteraceae; Brevundimonas; Brevundimonas vesicularis;; cellular organisms; Bacteria; Proteobacteria; Alphaproteobacteria; Caulobacterales; Caulobacteraceae; Brevundimonas; environmental samples; uncultured Brevundimonas sp.;; cellular organisms; Bacteria; Proteobacteria; environmental samples; uncultured proteobacterium
seq9,cellular organisms;Bacteria;environmental samples; uncultured bacterium;; cellular organisms; Bacteria; Proteobacteria; Betaproteobacteria; Burkholderiales; Burkholderiaceae; Ralstonia; Ralstonia solanacearum;; cellular organisms; Bacteria; Proteobacteria; Betaproteobacteria; Burkholderiales; Burkholderiaceae; Ralstonia; Ralstonia solanacearum; Ralstonia solanacearum FQY_4;; cellular organisms; Bacteria; Proteobacteria; Betaproteobacteria; Burkholderiales; Burkholderiaceae; Ralstonia; Ralstonia solanacearum; Ralstonia solanacearum GMI1000;; cellular organisms; Bacteria; Proteobacteria; Betaproteobacteria; Burkholderiales; Burkholderiaceae; Ralstonia; Ralstonia sp. BF07G02
```

Example dataset(5)

- Results in the previous slide are annotated with different colours to make them easy to read.
- Using –l full, we get the full taxonomic path from root node to the leaf node.
- For more than one hits, the taxonomic paths for unique hits are separated by the delimiter ;;
- On close scrutiny, we notice that except seq4 and seq9, the other sequences either end up as partial strains, uncultured taxa, or have missing taxa names in NCBI's taxonomy database.
- Using –l species will thus give us assignments for those query sequences that match against complete genomes in NCBI's nt database.