AMF Taxonomy Assignment Pipeline

Joyalea Carson-Austin

2025-08-05

Assign taxonomy to 97% OTUs via BLAST top hit

Note: This code was adapted and partially generated with support from OpenAI's ChatGPT (GPT-4, May 2025). Each step was verified and edited by author. To do this, taxonomic assignment was compared to the 'consenus vsearch' plugin set to 97% identity and 95% query coverage performed in QIIME2. As the only difference between these two methods is the inclusion of quality control with the blastn approach, taxonomic assignments should be roughly comparable.

```
blast <- read_tsv("C:/Users/Joyalea/Documents/Bioinformatics_97_BLASTN/blast_results_97.tsv",
                  col_names = c("qseqid", "sseqid", "pident", "length", "mismatch", "gapopen",
                                "qstart", "qend", "sstart", "send", "evalue", "bitscore"))
```

```
## Rows: 166 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr  (2): qseqid, sseqid
## dbl (10): pident, length, mismatch, gapopen, qstart, qend, sstart, send, eva...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
blast_top <- blast %>%
  group_by(qseqid) %>%
  slice_max(order_by = bitscore, n = 1, with_ties = FALSE) %>%
  ungroup()

ref_tax <- read_tsv("C:/Users/Joyalea/Documents/Bioinformatics_97_BLASTN/maarjam_ref_tax_export/taxonomy
                    col_names = c("FeatureID", "Taxon"))
```

```
## Rows: 384 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (2): FeatureID, Taxon
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
blast_tax <- blast_top %>%
  left_join(ref_tax, by = c("sseqid" = "FeatureID")) %>%
  select(FeatureID = qseqid, Taxon) %>%
  distinct(FeatureID, .keep_all = TRUE)

writeLines("#q2:types\tcategorical", "taxonomy_otu97.tsv")
write_tsv(blast_tax, "taxonomy_otu97.tsv", col_names = FALSE)
```

Identify unassigned OTUs from exported FASTA file

```
fasta <- readLines("C:/Users/Joyalea/Documents/Bioinformatics_97_BLASTN/rep_seqs_export/dna-sequences.fa
all_otus <- gsub(">", "", fasta[grepl("^>", fasta)])

matched_otus <- unique(blast$qseqid)
unmatched_otus <- setdiff(all_otus, matched_otus)

write.table(data.frame(FeatureID = unmatched_otus),
            "C:/Users/Joyalea/Documents/Bioinformatics_97_BLASTN/unassigned_ids_97.tsv",
            col.names = TRUE, row.names = FALSE, quote = FALSE, sep = "\t")
```

Reassign 90% identity OTUs at phylum level (manually in excel)

```
filtered_ref_tax <- ref_tax %>%
  filter(str_detect(Taxon, "p__Glomeromycota|p__Mucoromycota"))

blast90_top <- read_tsv("C:/Users/Joyalea/Documents/Bioinformatics_97_BLASTN/blast_results_90.tsv",
                        col_names = c("qseqid", "sseqid", "pident", "length", "mismatch", "gapopen",
                                      "qstart", "qend", "sstart", "send", "evalue", "bitscore")) %>%
  group_by(qseqid) %>%
  slice_max(order_by = bitscore, n = 1, with_ties = FALSE) %>%
  ungroup()
```

```
## Rows: 9510 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr  (2): qseqid, sseqid
## dbl (10): pident, length, mismatch, gapopen, qstart, qend, sstart, send, eva...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
blast90_tax_filtered <- blast90_top %>%
  inner_join(filtered_ref_tax, by = c("sseqid" = "FeatureID")) %>%
  select(FeatureID = qseqid, Taxon)

writeLines("#q2:types\tcategorical", "taxonomy_otu90_filtered.tsv")
write_tsv(blast90_tax_filtered, "taxonomy_otu90_filtered.tsv", col_names = FALSE)
```

Merge 97% and 90% taxonomy (phylum level only for 90% matches)

```
tax97 <- read_tsv("C:/Users/Joyalea/Documents/Bioinformatics_97_BLASTN/taxonomy_otu97_export/taxonomy.ts
                  col_names = c("FeatureID", "Taxon"))
```

```
## Rows: 43 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (2): FeatureID, Taxon
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
tax90 <- read_tsv("C:/Users/Joyalea/Documents/Bioinformatics_97_BLASTN/taxonomy_otu90_filtered_export/ta
                  col_names = c("FeatureID", "Taxon"))
```

```
## Rows: 151 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (2): FeatureID, Taxon
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
tax90_phylum <- tax90 %>% mutate(Taxon = str_extract(Taxon, "^k__Fungi; p__[^;]+"))
tax90_only <- tax90_phylum %>% filter(!FeatureID %in% tax97$FeatureID)
tax_merged <- bind_rows(tax97, tax90_only)

writeLines("#q2:types\tcategorical", "C:/Users/Joyalea/Documents/Bioinformatics_97_BLASTN/taxonomy_merg
write_tsv(tax_merged, "C:/Users/Joyalea/Documents/Bioinformatics_97_BLASTN/taxonomy_merged.tsv", col_nam
```

Export to ReBLAST 90% Assigned Features against NCBI nt database and the AMFungal database

```
taxonomy <- read_tsv("C:/Users/Joyalea/Documents/Bioinformatics_97_BLASTN/taxonomy_otu90_filtered_expor
```

```
## Rows: 150 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (2): Feature ID, Taxon
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Rename column for consistency
taxonomy <- taxonomy %>%
  rename(FeatureID = `Feature ID`)
```

```
# Export unique FeatureIDs
taxonomy %>%
  distinct(FeatureID) %>%
  write_lines("C:/Users/Joyalea/Documents/Bioinformatics_97_BLASTN/likely_glom_90_ids.txt")
```