Orthology analyses of 394 fungal proteomes

Elisabet Sjökvist

February 22, 2016

Outline

Project backgound

- How do pathogens evolve?
  - how has the secretome evolved, and how does that differ between fungi with different nutritional strategies?
  - how has the effector repetoar evolved?
  - what is the composition of carbohydrate active enzymes and lignocellolytic enzymes?

Genomes

- 564 genomes in Ensembl
- 659 genomes in NCBI
- 567 genomes in JGI, not all published
- 394 publicly avaliable proteomes of “different species”

Orthology analyses

- All vs. all blast ~ 4 weeks
  - 395 fasta files
  - 395 blast databases
  - min. 30aa
  - max one stop codon at end of sequence
- Orthofinder ~ 9 days
  - SequenceIDs.txt
  - SpeciesIDs.txt
  - 156025 blast result files

Distribution of cluster sizes

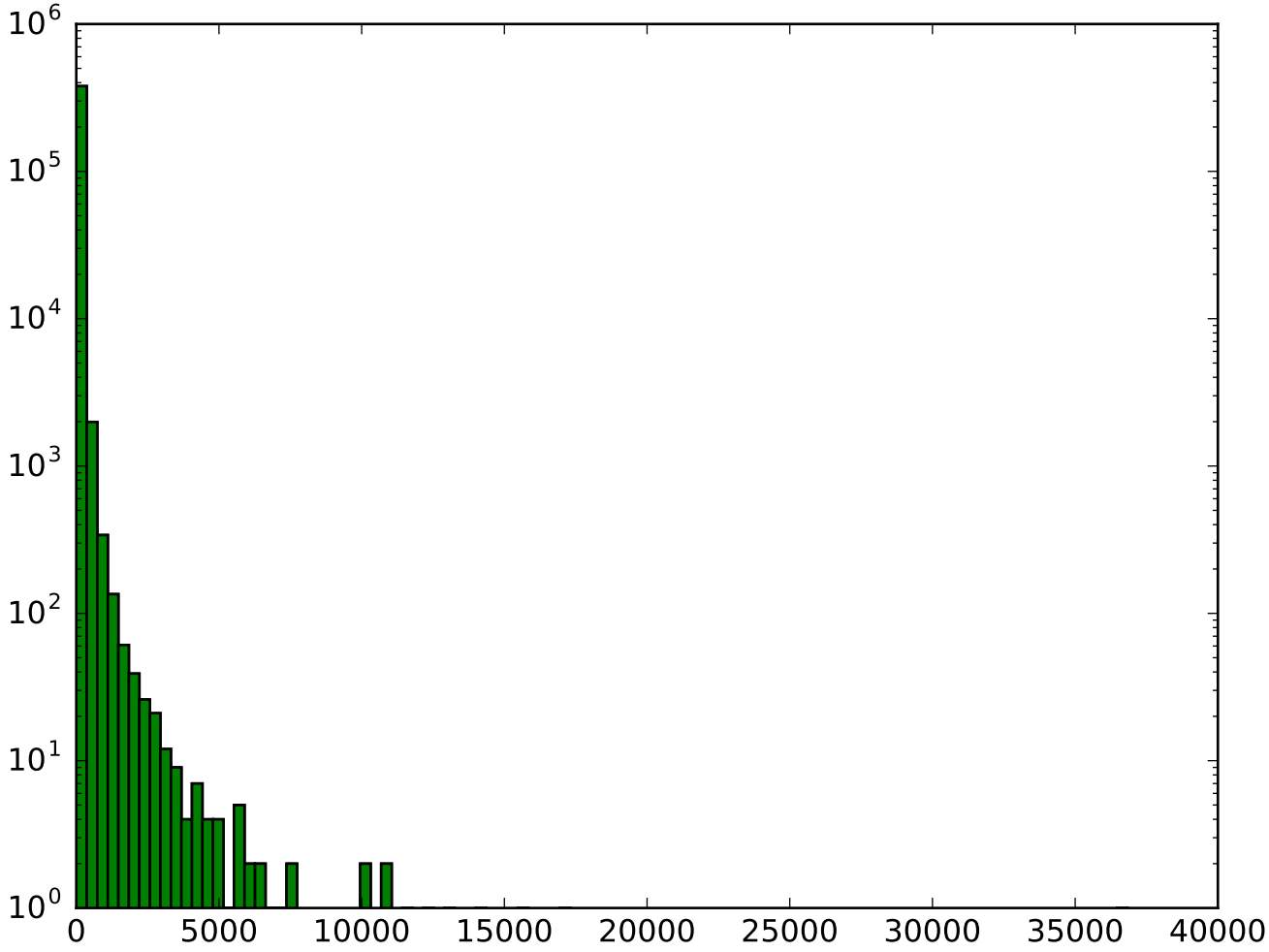

Figure : Distribution of cluster sizes, y-axis = nr of clusters, x-axis = proteins per cluster

Database structure

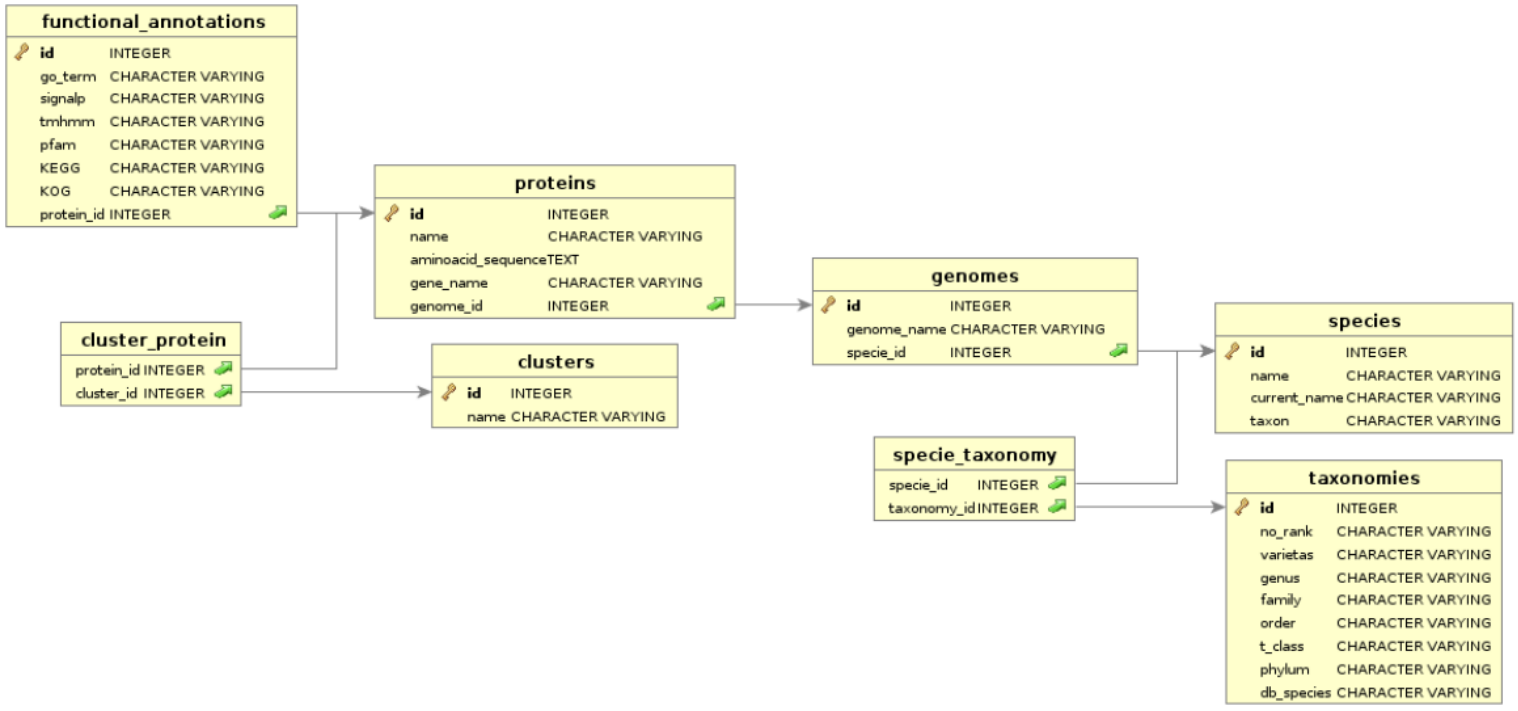

Figure : Organization

fun395=# select count(taxonomies.id), taxonomies.phylum from taxonomies group by taxonomies.phylum; count | phylum ——-+————————— 11 | undef 3 | Chytridiomycota 234 | Ascomycota 1 | Glomeromycota 88 | Basidiomycota 19 | Microsporidia 1 | Blastocladiomycota 1 | Cryptomycota 1 | Entomophthoromycota (9 rows)

Nr. of protein families as a function of species sampled

- random sampling of 20-380 taxa with 20sp interval x 20
- clusters where taxa have at least 1 protein present