# Comparative Genomics and Core Fungal Genes

Tom Hsiang
Dept. Environmental Biology

www.uoguelph.ca/~thsiang/present/2004botany.pdf
Univ. Guelph Botany Seminar, Oct. 26, 2004

# Genomics

- the study of entire genomes, which are the total genetic information of organisms
- Bioinformatics is the use or study of the tools for analysis of biological data such as genomes

# Comparative Genomics

- Comparison of genomes (to give insights into biology, evolution, genetics ….)
- Bioinformaticians have analytical tools, but often not the biological background
- Biologists have data and ideas, but usually not the analytical tools

# Genomic sequencing

- 1995: first non-viral genome (bacterial)
  - *Haemophilus influenzae* (2 Mb)
- 1997: first eukaryote sequenced (fungal)
  - *Saccharomyces cerevisiae* (12 Mb)
- 1998: first animal sequenced (nematode)
  - *Caenorhabditis elegans* (97 Mb)
- 2000: first plant sequenced
  - *Arabidopsis thaliana* (115 Mb)

# Genomic sequencing - animals

- 2000, *Drosophila melanogaster* (fly, 140 Mb)
- 2001, *Homo sapiens* (human, 2.9 Gb)
- 2002, *Fugu rubripes* (fish, 330 Mb)
  - *Ciona intestinalis* (sea squirt, 160 Mb)
  - *Mus musculus* (mouse, 2.5 Gb)
  - *Anopheles gambiae* (mosquito, 280 Mb)
- 2003, *Xenopus tropicalis* (frog, 1.7 Gb)
- 2004, *Rattus norvegicus* (rat, 2.8 Gb)
- 2004, *Bombyx mori* (silkworm, 530 Mb)
- Cow, chimpanzee, bee almost complete

# Genomic sequencing - plants

- 2002, *Oryza sativa* (rice, 400 Mb)
  - Beijing Genomics Institute
    - » var. *indica*
  - International Rice Genome Sequencing Project
    - » var. *japonica*
- 2004, *Populus trichocarpa* (tree, 400 Mb)

# Genomic sequencing - fungi

- Over 20 species to date in publicly accessible databases (more than other higher eukaryotes)
- genome size: 10 to 50 Mb (megabase pairs, or file size in megabytes)
- data in the public domain, funded by government and freely available (some held by private companies)

# Objectives

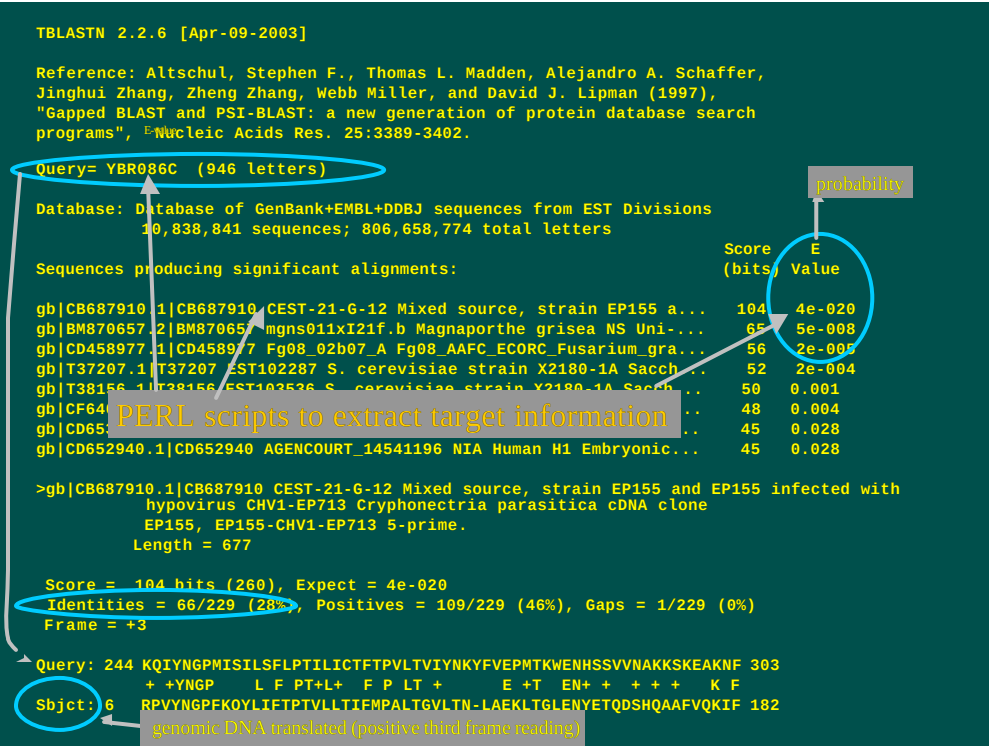
- Find common fungal sequences by comparing each yeast gene to 13 other fungal genomes
- Find core fungal sequences (unique to fungi) by comparing these common fungal sequences to other organisms

# Fungal predicted genes (ORFs)


- Yeast (*Saccharomyces cerevisiae* ) has ~6000 predicted genes (yeast genome database)
  - ~4000 of these have a predicted function
- *Neurospora crassa* has ~10,000 predicted genes
  - less than 500 of these have a predicted function
- by comparison the human genome has ~30,000 predicted genes (~100,000 gene products) of which ~50% are of unknown function

# Methods of gene comparison

- Yeast genome 12 Mb -> 6,000 predicted genes
- Download genomes & create local databases
- Use Standalone TBLASTN to find homologs


TBLASTN 2.2.6 [Apr-09-2003]
Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs", Nucleic Acids Res. 25:3389-3402.
Query= YBR086C (946 letters)
Database: Database of GenBank+EMBL+DDBJ sequences from EST Divisions
10,838,841 sequences; 806,658,774 total letters
probability
Score E
Sequences producing significant alignments: (bits) Value
gb|CB687910.1|CB687910 CEST-21-G-12 Mixed source, strain EP155 a... 104 4e-020
gb|BM870657.2|BM87065 mgns011xI21f.b Magnaporthe grisea NS Uni-... 65 5e-008
gb|CD458977.1|CD458977 Fg08_02b07_A Fg08_AAFC_ECORC_Fusarium_gra... 56 2e-005
gb|T37207.1|T37207 EST102287 S. cerevisiae strain X2180-1A Sacch .. 52 2e-004
gb|T38156.1|T38156 EST103536 S. cerevisiae strain X2180-1A Sacch .. 50 0.001
gb|CF64 .. 48 0.004
gb|CD65 .. 45 0.028
gb|CD652940.1|CD652940 AGENCOURT_14541196 NIA Human H1 Embryonic... 45 0.028
PERL scripts to extract target information
>gb|CB687910.1|CB687910 CEST-21-G-12 Mixed source, strain EP155 and EP155 infected with
hypovirus CHV1-EP713 Cryphonectria parasitica cDNA clone
EP155, EP155-CHV1-EP713 5-prime.
Length = 677
Score = 104 bits (260), Expect = 4e-020
Identities = 66/229 (28%), Positives = 109/229 (46%), Gaps = 1/229 (0%)
Frame = +3
Query: 244 KQIYNGPMISILSFLPTILICTFTPVLTVIYNKYFVEPMTKWENHSSVVNAKKSKEAKNF 303
+ +YNGP L F PT+L+ F P LT + E +T EN+ + + + + K F
Sbjct: 6 RPVYNGPFKQYLIFTPTVLLTIFMPALTGVLTN-LAEKLTGLENYETQDSHQAAFVQKIF 182
genomic DNA translated (positive third frame reading)

# 6356 Yeast predicted genes

3340 Yeast genes with homologs in 12 of 14 fungi

Ascomycetes

*Aspergillus fumigatus*
*Aspergillus nidulans*
*Candida albicans*
*Gibberella zeae*
*Magnaporthe grisea*
*Neurospora crassa**
*Podospora anserina*
*Trichoderma reesei*
*(Saccharomyces cerevisiae*)*

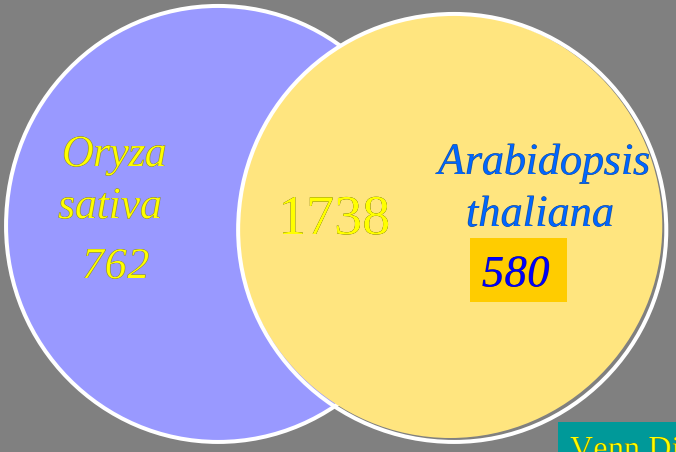
Basidiomycetes

*Cryptococcus neoformans*
*Phanerochaete chrysosporium*
*Ustilago maydis*
*Phakopsora pachyrhizi*
*Coprinus cinerea*

* these genomes are considered completely sequenced; others > 95%?

# 3340 genes common to 12 of 14 fungi compared to plant genomes

# 3340 genes common to 12 of 14 fungi compared to plant genomes




Venn Diagram

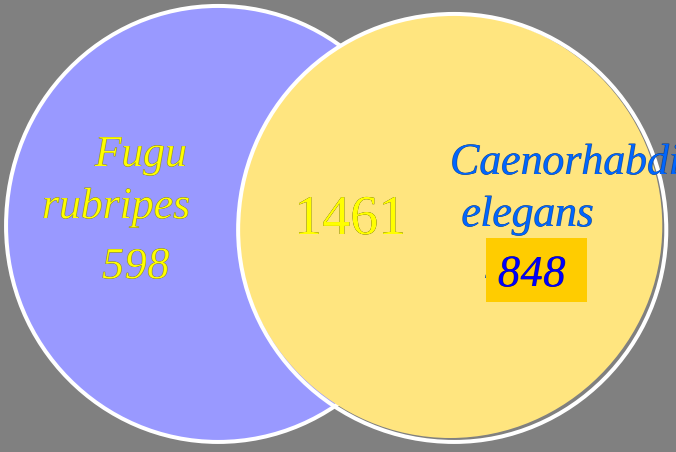
Total: 3080

# 3340 genes common to 12 of 14 fungi compared to animal genomes

*Fugu rubripes*
*2059*

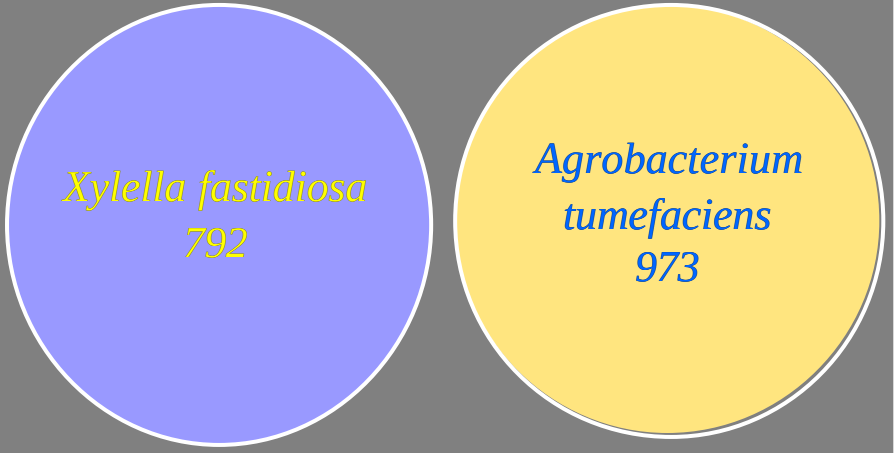
*Caenorhabditis elegans*
*2309*

# 3340 genes common to 12 of 14 fungi compared to animal genomes

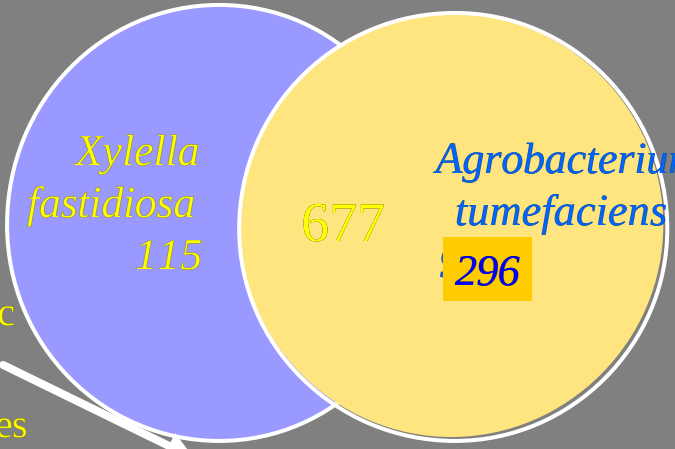



Total: 2907

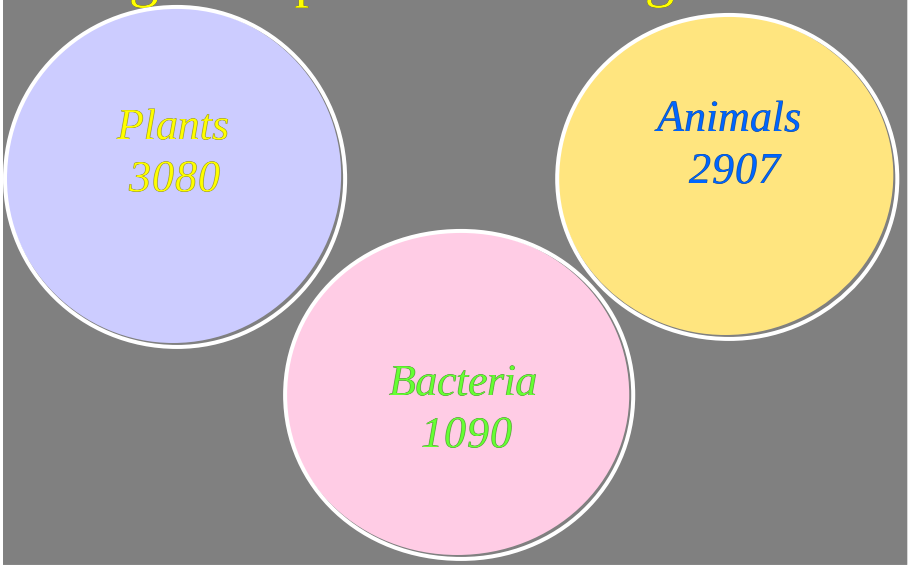
# 3340 genes common to 12 of 14 fungi compared to bacterial genomes



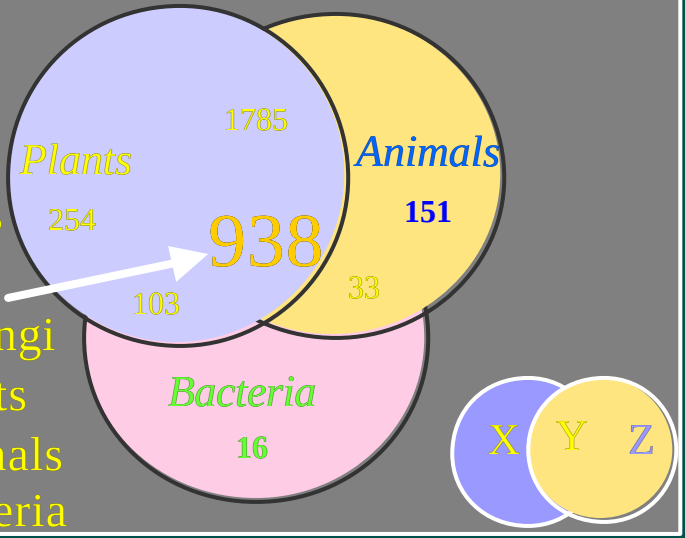
# 3340 genes common to 12 of 14 fungi compared to bacterial genomes



# 3340 genes common to 12 of 14 fungi compared to other genomes



# 3340 genes common to 12 of 14 fungi compared to other genomes



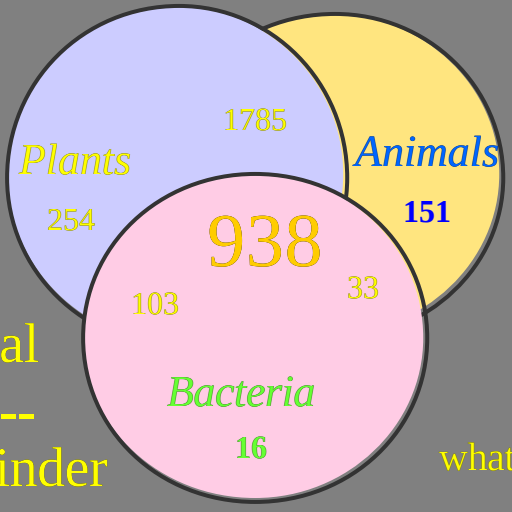

Core genes found in
12 to 14 fungi
1 or 2 plants
1 or 2 animals
1 or 2 bacteria

# What are these 938 core genes?

- ~150 function unknown
- ~ 400 involved in metabolism
- ~ 200 involved in DNA/RNA processing
- ~ 150 involved in signalling

Note: functional annotation from Yeast Genome Database

# 3340 genes common to 12 of 14 fungi compared to other genomes




3280 Total

---

60 Remainder

what are they?

# What are these 60 fungal genes?

- are they really found only in fungi?
  - sent to BLAST against GenBank NR , EST, GSS, and HTGS databases
- **43** have a non-fungal match at E-value $\leq 10^{-5}$ so these are not exclusively fungal (many have homologs in the human or mouse genomes). Contaminants?
- this leaves **17** genes found only in fungi

# What are these 17 fungal genes?

- 5 have unknown function
- 2 involved in protein biosynthesis
- 2 involved in transport
- 7 have miscellaneous functions
- 1 involved in sporulation (clearly fungal)

Note: functional annotation from Yeast Genome Database

# Characteristics of the 17 genes?

- Scattered across yeast genome (16 chr.)
- GC content not different from other yeast genes
- Codon usage patterns similar
- Deletion mutants (Yeast Genome Database)
  - 1 non-viable phenotype; 2 severe growth defects
- Patents
  - 7 in U.S. Patents on antifungal drug discovery

# Characteristics of the 5 unknowns?

- The nt sequences of 5 yeast genes of unknown function were aligned with homologs in other fungi, and the most conserved regions matched against GenBank NR
  - 4 of the 5 found to have conserved portions matching membrane or membrane-related proteins

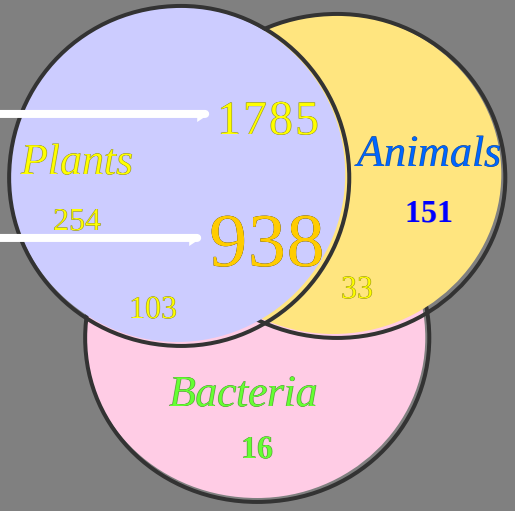
# Age of yeast genes (6355 total)

## 3340 common fungal genes

> 1.1 billion years old
(asco/basidio split)

> 1.6 billion years old
(fungi/animal/plant split)

> 2.7 billion years old
(euk/prok split)




Timeline based on
Hedges & Kumar 2003,
Trends in Genetics 19:200

* fine print: assuming no horizontal transfer

# Methods for comparing genomes

- set up Linux Operating System (free)
  - also works in Windows, but slower
- download genomic data from internet (free)
  - see my web copy for details
- set up Standalone Blast program (free)
  - from www.ncbi.nlm.nih.gov/BLAST/
- learn PERL script programming (free but takes much time for biologists)

# Uses for Comparative Genomics

- evolutionary biology
- phylogenetics (whole genome comparisons)
- targetted drugs (pathogen gene absent in host)
- finding genes (conserved sequences)
- primer design
- whatever you can dream up!

# Current & Future Research

- Discovery of new highly conserved sequences that would be useful in phylogenetics

  rDNA, actin, tubulin, EF, Hsp, RNA polymerase...

- Using entire gene sets (proteomes) to resolve the relationship between plants/fungi/animals
- Evolutionary relationships of slime molds (myxomycota) to other eukaryotes

# Publications

- in press. Comparison of the yeast proteome to other fungal genomes to find core fungal genes. J. Molec. Evol.
- 2004. Comparative analysis of expressed sequence tags from *Colletotrichum*-infected plants. Plant Science 167:481-489.
- 2004. Comparative fungal genomics. Can. J. Plant Pathol. 26:19-30
- 2003. Distinguishing plant and fungal ESTs. J. Microbiological Methods 54:339-351
- http://www.uoguelph.ca/~thsiang/pubs

# Supplemental Information

- NR = nonredundant, 0.5 x $10^9$ letters
- EST = expressed sequence tags, 9.6x$10^9$
- GSS = genome survey sequences, 4.5x$10^9$
- HTGS = unfinished high-throughput genome sequences, 1.2x$10^{10}$

NOTE: Numbers current October 2003

# Methods: fungal genomes

- *Aspergillus fumigatus*
  ftp.sanger.ac.uk/pub/pathogens/A_fumigatus
- *Aspergillus nidulans*
  www-genome.wi.mit.edu/cgi-bin/annotation/aspergillus
- *Cryptococcus neoformans*
  valefor.stanford.edu/group/Cneoformans
- *Fusarium graminearum* (*G. zeae*)
  www-genome.wi.mit.edu/ftp/pub/annotation/fusarium
- *Magnaporthe grisea*
  www-genome.wi.mit.edu/annotation/fungi/magnaporthe

# Methods: fungal genomes

- *Neurospora crassa*
  www-genome.wi.mit.edu/cgi-bin/annotation/neurospora
- *Phanerochaete chrysosporium*
  ftp.jgi-psf.org/pub/JGI_data/WhiteRot
- *Saccharomyces cerevisiae*
  GenBank (www.ncbi.nlm.nih.gov)
- *Schizosaccharomyces pombe*
  ftp.genome.ad.jp/pub/db/ebi/embl/genomes/Eukaryota/spombe

# Methods: other genomes

- *Agrobacterium tumefasciens*
  GenBank (www.ncbi.nlm.nih.gov)
- *Arabidopsis thaliana*
  GenBank (www.ncbi.nlm.nih.gov)
- *Caenorhabidits elegans*
  ftp.wormbase.org/pub/wormbase/DNA_DUMPS
- *Fugu rubripes*
  genome.jgi-psf.org/fugu6
- *Oryzae sativa*
  210.83.138.53/rice/download.php
- *Xylella fastidiosa*
  aeg.lbi.ic.unicamp.br/xf