Sequencing and assembly of small eukaryotic genomes & comparative genomics

Dr. Tom Hsiang

Environmental Sciences

email: thsiang@uoguelph.ca

www.uoguelph.ca/~thsiang/present/1403genomics.pdf

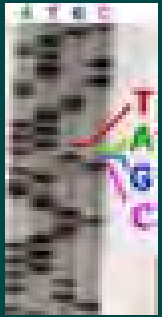
Presentation Outline

- Past & NextGen (next generation) sequencing
  - before & now
- Research Objective
  - sequence fungal genomes
- Methods
  - timeline of research
- Results
  - output of several assembly programs
- Conclusions
  - can it be done? now? later?
- Flowchart for genome sequencing & assembly
- PART II: Comparative Genomics

Sequencing Technologies

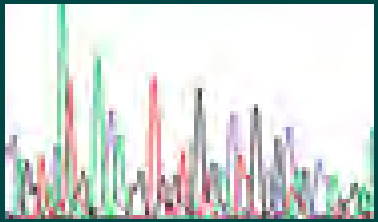
- 1970s: Gilbert (PNAS74:560), Sanger (PNAS74:5463), Sanger chain termination method

autoradiograph

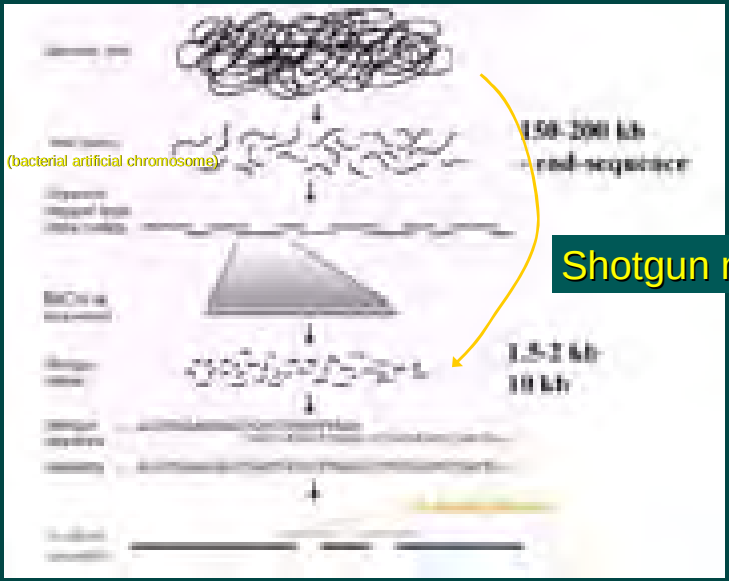


- 1987, ABI first automated sequencer
  - ABI 377 gel based, ABI 3700 capillary based
  - now up to 700 bp reads

chromatogram



Whole genome sequencing by library



or fosmid (40kb)
(bacterial f-plasmid)

Shotgun method

Sequencing Technologies

Next Generation Sequencing (NGS) Technologies

- 2005 - 454 pyrosequencing
  - » 150-200bp reads (now > 400 bp)
- 2006 - Solexa seq by synthesis
  - » 35bp reads, now > 100 bp
- 2008 - SOLiD = seq by oligo ligation & detection
  - » 50 bp reads

Next Gen Sequencing (nat.rev.micro6:419)

- 454
- SOLiD
- SOLEXA

Genome assembly statistics for plant, animals, fungus and bacterium and $

	Maize	Horse	Panda	*Grosmannia clavigera*	*Pseudomonas syringae*
Genome length	2.3 Gb	2.6 Gb	2.5 Gb	32.5 Mb	6.1 Mb
Sequencing technology	Sanger	Sanger	Illumina	Sanger/454/Illumina	Illumina
Number of contigs	125,325	55,316	198,274	3,361	1,346
Contig N50	40 kb	112 kb	40 kb	32 kb	11 kb
Number of scaffolds	61,161	9,687	81,469	2,322	71
Scaffold N50	76 kb	46 Mb	1.3 Mb	782 kb	317 kb
~ Sequencing cost	$30 million	$15 million	$0.6 million	$100,000	$4,000

Jackman & Birol (2010, Genome Biology 11:202) in US dollars $

"Contig or scaffold N50 is a weighted median statistic such that 50% of the entire assembly (bases) is contained in contigs or scaffolds equal to or larger than this value"

Research Objective

- use Illumina-Solexa technology to obtain the complete genomic sequences for several fungi, and assemble with various programs

Hypotheses

- Small genomes can be sequenced and assembled without a reference sequenced genome, using Illumina technology & genome coverages of 100x and paired-end sequencing
- Genes such as mating type genes can be found in such assemblies, to make inferences about life cycles in nature

What is a Reference Genome?

- If the same species or a closely related species has been sequenced, then this genome can be used as a scaffold/reference/predictor for the new genome
- If there is no previous genome sequenced for this species, then this is called "denovo sequencing"

What is Genome Coverage?

$$\frac{\text{the total number of bases sequenced}}{\text{number of bases in genome}}$$

e.g. 25 million reads of 100bp length paired-end of a 50 Mb genome gives 100x coverage

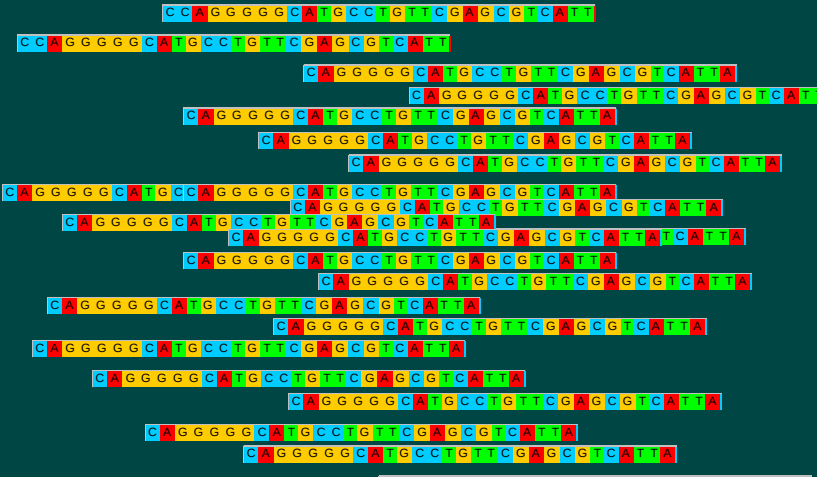
(25,000,000 * 200) / 50,000,000 = **100x**

What is a (paired-end) Read?

Why use Paired-end sequencing?

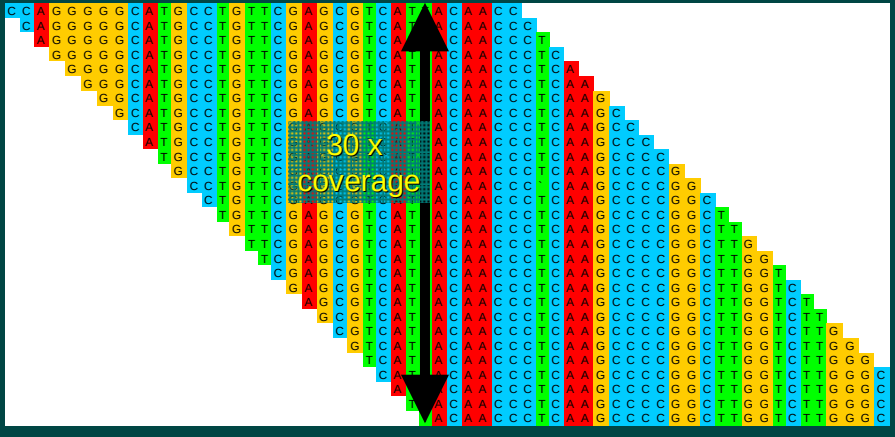
- sequence both ends of DNA fragment (>200 bp)
- gives more information on the position of the small fragments, and better mapping of ambiguous reads than for single ends
- skip over regions difficult to sequence

Assembly of 35 bp reads (millions)



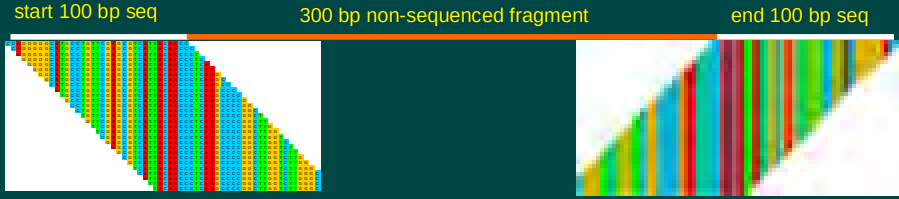
Assembly of 35 bp reads (millions)

Assembly program aligns reads



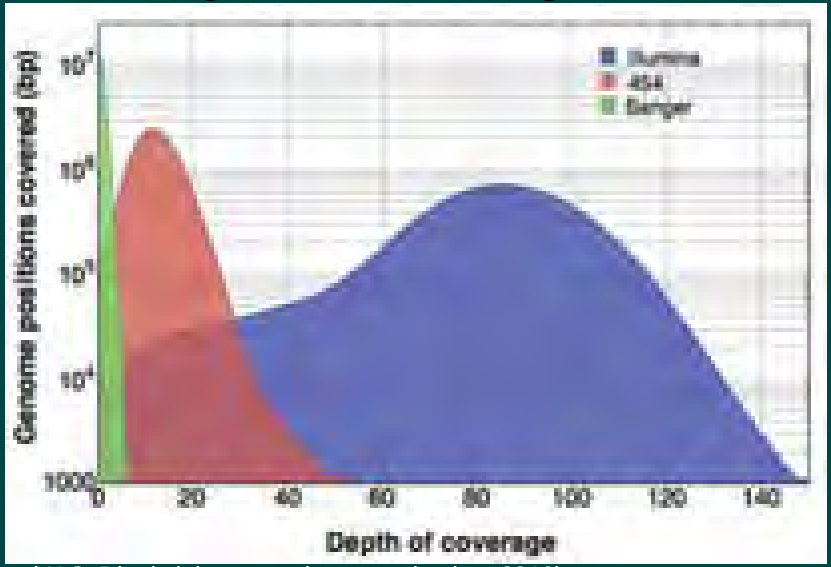
Creates contigs (consensus sequences)

Paired end assembly



Skips over difficult to sequence regions & repeats and creates Scaffolds with “N” between contigs

What is genome coverage?



(W.S. Diguistini, personal communication, 2010)

Methods – Ascomycota tested

Species	Class	Order
Diplodia pinea	Dothideomycetes	Botryosphaeriales
Discula destructiva	Sordariomycetes	Diapothales
Gaeumannomyces cylindrosporus	Sordariomycetes	Magnaporthales
Neonectria faginata	Sordariomycetes	Hypocreales
Sclerotinia homoeocarpa	Leotiomycetes	Helotiales
Volutella buxi	Sordariomycetes	Hypocreales

Methods

- Spring 2011 (my lab)
  - grew and harvested mycelium (500 mg) of fungal hyphae grown on cellophane over PDA
  - extracted DNA with Qiagen DNeasy kit
- June 2011 (my lab)
  - sent to Sequencing Center 1 µg of genomic DNA of each of six fungal species
- July 2011 (Sequencing Center)
  - created genomic sequencing libraries
  - 3 lanes with 2 libraries per lane of 100bp paired-end runs on an Illumina-Solexa GAIIx machine

Methods

- Sept 2011 (Sequencing Center)
  - gave us data on external hard drive (2 Tb)
- Oct & Nov 2011 (my lab)
  - assembled reads with 3 assembly programs
    - » Abyss
    - » SOAPdenovo
    - » Velvet
- Dec 2011 (my lab)
  - predicted genes with Augustus (prediction pgm)
  - searched for mating type genes in assemblies

Results

- millions of 100bp records (fastq)

```
@SOLEXA5:2:120:19806:18845#0/1
TACTAGAA.CTCTAATGATATTACTATACCCTCAGTAATTAACTGGGCCACTTAAAAAGTTTCCTTTAAAAGTTTA
+SOLEXA5:2:120:19806:18845#0/1
cccca^_UBa``b``cacccccccccccccccacc^cccccb`^caacccTcccbaaac`^cbccYcbYY^`abb^
```

Results – sequence reads

Species	Reads
Diplodia pinea	27,291,386
Discula destructiva	28,036,272
Gaeumannomyces cylindrosporus	29,479,478
Neonectria faginata	30,697,374
Sclerotinia homoeocarpa	28,436,528
Volutella buxi	35,064,118

Assembly stats: *Diplodia pinea*

Program	Contigs/Scaffolds	N50
Abyss 1.2.7	12,818 s	44 kb
SOAP 1.04	1,705 s	74 kb
Velvet 1.1.05	17,811 c	5.9 kb

Contigs = number of assembled pieces from raw reads

Scaffolds = joining of contigs with N's in between

Highest N50 (weighted average size) calculated from various K-values, cutoffs, etc.

Results – N50 assembly stats

Species	Abyss	SOAP	Velvet
Diplodia pinea	44 kb	74 kb	5.9 kb
Discula destructiva	21 kb	35 kb	9.0 kb
Gaeumannomyces cylindrosporus	58 kb	82 kb	23 kb
Neonectria faginata	13 kb	18 kb	1.4 kb
Sclerotinia homoeocarpa	48 kb	57 kb	10 kb
Volutella buxi	44 kb	56 kb	7.1 kb

Results – assembly size (Mb)

Species	Abyss	SOAP	Velvet
Diplodia pinea	38.4	37.0	38.3
Discula destructiva	47.9	47.2	49.0
Gaeumannomyces cylindrosporus	44.5	42.5	43.4
Neonectria faginata	48.5	44.0	44.8
Sclerotinia homoeocarpa	45.9	44.7	45.6
Volutella buxi	29.0	29.3	29.5

Results – assembly stats

Species	Reads	~Size	Coverage
Diplodia pinea	27,291,386	38 Mb	144 x
Discula destructiva	28,036,272	48 Mb	117 x
Gaeumannomyces cylindrosporus	29,479,478	43 Mb	137 x
Neonectria faginata	30,697,374	46 Mb	133 x
Sclerotinia homoeocarpa	28,436,528	45 Mb	126 x
Volutella buxi	35,064,118	29 Mb	263 x

Results – gene predictions with Augustus

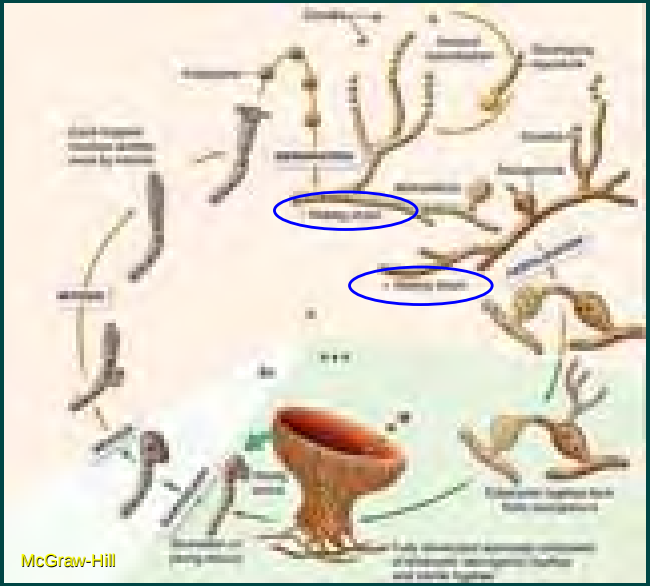
Species	Genes
Diplodia pinea	~13,000
Discula destructiva	~12,000
Gaeumannomyces cylindrosporus	~13,000
Neonectria faginata	~14,000
Sclerotinia homoeocarpa	~ 9,000
Volutella buxi	~ 9,000

Results

Species	Core
D. pinea	246
D. destructiva	245
G. cylindrosporus	247
N. faginata	247
S. homoeocarpa	247
V. buxi	245

Parra et al. (2009, NAR37:289) created a core eukaryotic set of 248 genes to evaluate genome assemblies & predictions. Core means genes out of the 248 core set that have an e-value of ≤6e-4, implying that these assemblies contain most of the core genes, and hence should have the other genome-specific genes.

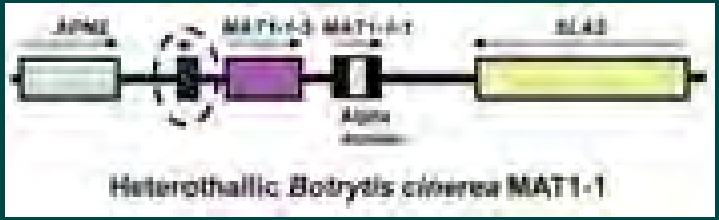
Ascomycete mating types

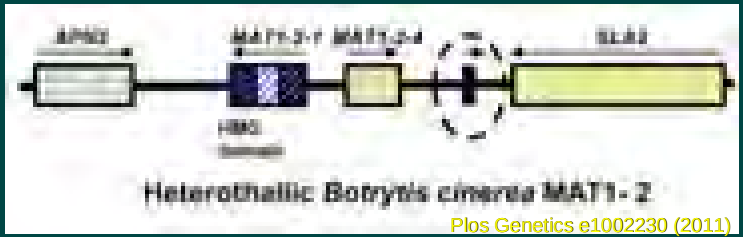


McGraw-Hill

Heterothallism

- two genotypes of different mating type needed for sexual reproduction (MAT1 and MAT2)





Plos Genetics e1002230 (2011)

Homothallism

- both MAT1 and MAT2 present in a genotype, which can produce sexual spores by itself

Plos Genetics
e1002230 (2011)

Results: mating type genes

Contigs from the various assemblies were compared against many ascomycete MAT1 & MAT2 genes

Isolate	MAT1	MAT2
Diplodia pinea SP16a	Yes	Yes
Discula destructiva 10115	No	Yes
Gaeumannomyces cylindrosporus 08145	No	Yes
Neonectria faginata 11171	Yes	Yes
Sclerotinia homoeocarpa SH84	No	Yes
Volutella buxi 09052	No	Yes

Dogwood anthracnose

- caused by *Discula destructiva*

Hudler

Results: mating type genes (M. Stanescu)

- MAT1-2 primers were designed from genome
- all isolates tested to date (~50) have MAT1-2
- other mating type was not introduced when the fungus arrived in North America in the 1970's?
- if introduced, perhaps another epidemic?

Volutella Blight (caused by *Volutella buxi*)

Results: mating type genes (Amy Shi)

- MAT1-2 primers designed from genome
- In Ontario isolates, 16 MAT1-2, 12 no band
- What is in these 12 isolates?

Results: mating type genes (Amy Shi)

- Conserved MAT1-1 primers designed from other species gave multiple bands
- Used SLA2 and APN2 in semi-nested PCR (to fish out the MAT1-1 gene
- sexually reproducing with 1:1 ratio?
  - (12 MAT1-1 to 16 MAT1-2)

Conclusions

- many secrets in a genome
- if a small genome is already sequenced
  - assemble another genotype for < $1000
- If not sequenced yet (*de novo*)
  - need >100x coverage and different sized libraries (paired-end 200bp, mate-pair 3kb, >$2000?)
- Is a $1000 *de novo* genome possible?
  - Maybe, but only if your team learns how to put the billions of nucleotides and millions of pieces together!

Fungal genomes sequenced

Species	Date	Scaf-N50	Cost
Microdochium majus	2011/06	116 kb	
Diplodia pinea	2011/09	74 kb	
Bipolaris hawaiiensis	2012/01	740 kb	
Venturia inaequalis	2012/08	43 kb	
Cylindrocladium buxicola	2012/12	420 kb	
Pestalotiopsis funerea	2013/02	450 kb	
Pestalotia sp.	2013/03	1383 kb	
Fusarium culmorum	2013/07	910 kb	
Fusarium proliferatum	2013/12	1591 kb	

Flowchart of a genome sequencing and assembly process

- Figure 8 (with text explanations) from
  - Haridas S, Breuill C, Bohlmann J, Hsiang T. 2011. A biologist's guide to de novo genome assembly using next-generation sequence data: a test with fungal genomes. Journal of Microbiological Methods 86:368-375.
  - available at www.uoguelph.ca/~thsiang/pubs
  - glossary of sequencing-related words provided

Flowchart of genome sequencing & assembly

4.1) Prepare DNA for submission *to Sequencing Centre (Illumina technology)*
- regular DNA extraction methods with kits are sufficient

⇓

4.2) Select a depth of coverage and specify library fragment size
- a single lane of 100 bp reads may be sufficient *(GAIIx = 10 Gb/lane)*
- *HiSeq-2000 (30 Gb/lane) multiplexed (up to 12 samples in one lane)*
- typical insert sizes are 200 bp to 10 kb
- give these instructions to sequencing center to make DNA library *(typically $250 per library, and $2500-3000/lane)*

⇓

4.3) Set up or obtain access to computing facilities *(& learn to use)*
- set up Linux on a PC with >16 Gb RAM or
- access a High Performance Computing Cluster (*e.g. SHARCNET*)

⇓

Generation of sequence data at sequencing center (2-4 mo)

NGS genomic DNA library preparation

- get fragments of target size (e.g. 500 bp, with the two 100bp ends sequenced)
- current rate-limiting step for NGS
- Illumina library prep ~$250, or you can do it yourself with kits for < $100

Flowchart of genome sequencing & assembly

Generation of sequence data at sequencing center

⇓

4.4) Download or copy and convert the raw sequence data

- use software to download the gigabytes of data
- convert the data format for processing

⇓

4.5) Choose assembly program (e.g. Abyss, SOAP, Velvet)

- choose between several (freely) available software programs (Linux)

⇓

4.6) Choose k-mer for assembly

- lower values have higher sensitivity, while higher values have higher specificity
- scripts can be used to produce a wide range of k-mer assemblies

⇓

4.7) Accept that your Illumina assembly is incomplete

- current technological limitations cannot overcome issues of repetitive regions

Future Research

- Having a genome sequence does not mean it is accurate or complete
- Other data is needed to verify genes, such as EST or RNAseq
- Also, comparison of same genotype with completed sequencing projects done with other technologies is needed to confirm the assemblies and gene predictions
- This type of research is underway….

Acknowledgements

- NSERC for providing base funding for this work
- A. Darbyson, B. Nash and A. Shi helped prepare DNA for submission to Sequencing Centre
- SHARCNET for access to computer clusters

Comparative Genomics

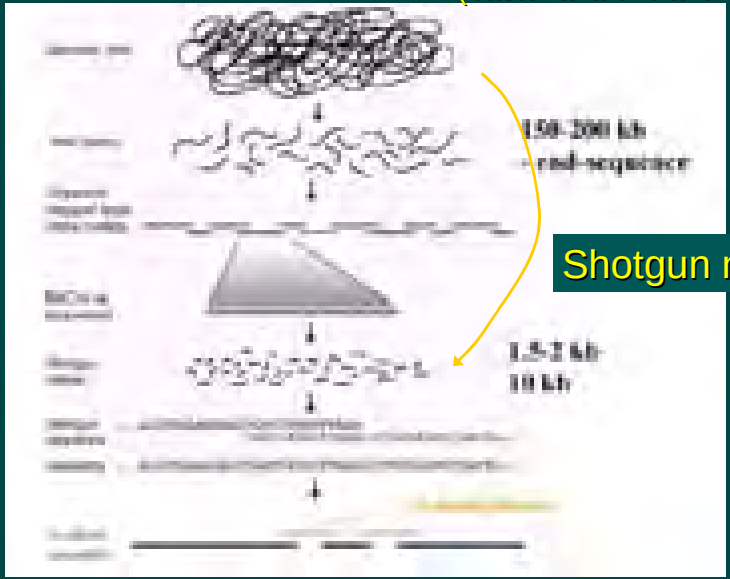
- definitions
- common fungal genes
- core set of genes among various Kingdoms
- horizontal Gene transfer

Genomic sequencing - history

- 1995: first complete 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)
- 2001: *Homo sapiens* (human, 2.9 Gb)

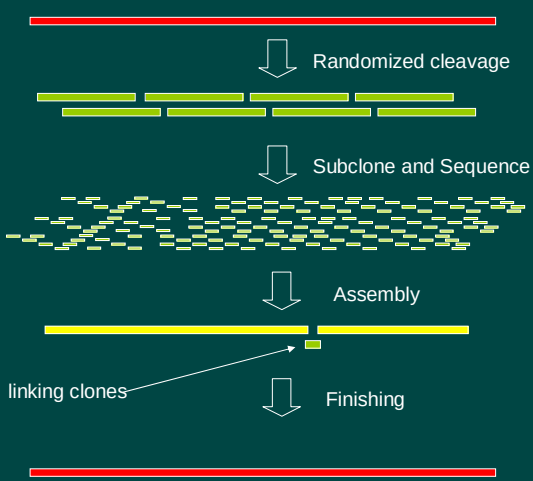
Whole genome sequencing by BAC

(bacterial artificial chromosome)



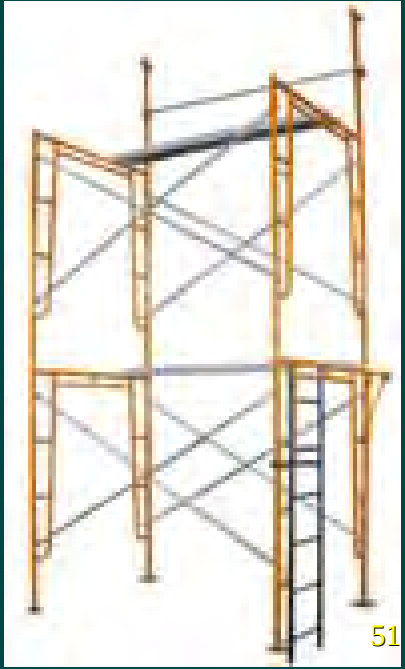
Shotgun method

Whole genome sequencing by shotgun



What is a scaffold?

- In genomic mapping, a scaffold is a series of contigs that are in the right order but not necessarily connected in one continuous sequence
- a contig (~contiguous) is a set of overlapping joined DNA segments



Comparative genomics

- Comparison of genome sequence data
- Completely sequenced genomes (>95%)
  - > 1000 prokaryotic genomes (Archaea & bacteria)
  - > 100 fungal genomes (especially yeasts)
  - > 20 animal genomes (insects, fishes, mammals)
  - > 10 plant genomes (arabidopsis, rice, poplar, corn, green alga, coffee, grape, soybean)

data from Genomes Online (http://www.genomesonline.org/) and other sources

Research objective: Comparison of the yeast proteome to other fungal genomes to find core fungal genes

Objective 1

- Find common fungal sequences
  - compare each yeast gene to 13 fungal genomes

Objective 2

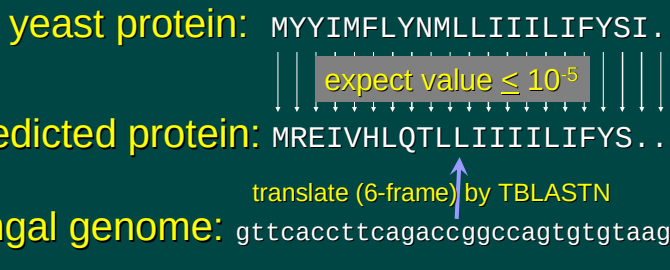
- Assess how many are shared with other taxa
  - compare common fungal sequences to animals, plants & bacteria

Fungal predicted genes (ORFs)

- Yeast (*Saccharomyces cerevisiae*) has ~6000 predicted genes (yeast genome database)
  - ~4000 of these have a predicted function

Methods: sequence comparisons

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



The Expect Value refers to “the number of hits one can expect to see just by chance when searching a database of a particular size” (probability of match by chance) 1e-5

```
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
                                                                 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|BM870657 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|CF640919                                                     a...    48   0.004
gb|CD653105                                                     c...    45   0.028
gb|CD652940.1|CD652940 AGENCOURT_14541196 NIA Human H1 Embryonic...    45   0.028

>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
```

PERL scripts to extract target information

6356 Yeast predicted genes

3340 Yeast genes with homologs in 12 of 14 fungi

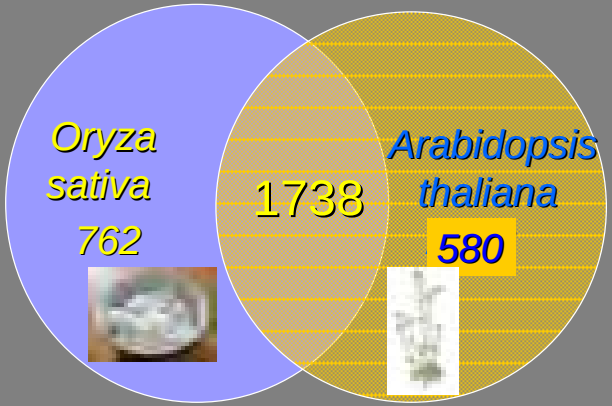
Ascomycota

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

Basidiomycota

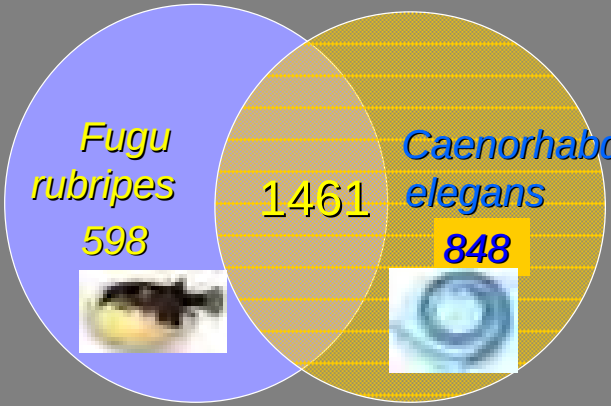
Cryptococcus neoformans
Coprinus cinerea
Phanerochaete chrysosporium
Phakopsora pachyrhizi
Ustilago maydis

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



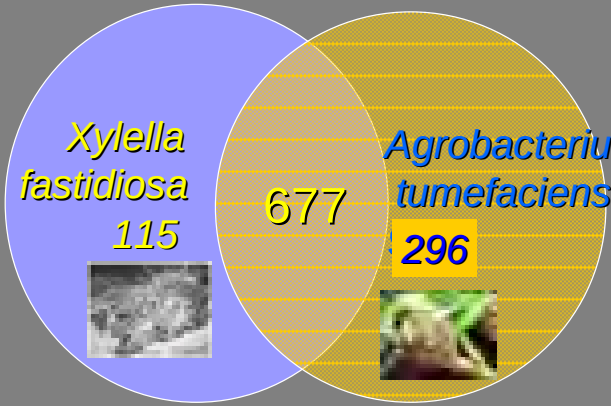
Total: 3080

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



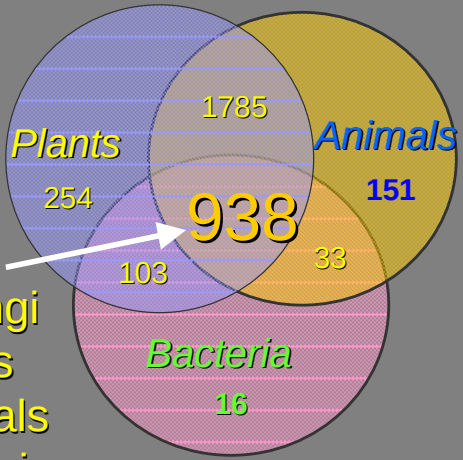
Total: 2907

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



Total: 1090

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



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

What are these 938 core genes?

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

functional annotation from Yeast Genome Database

Objective 3

- Assess the age of fungal genes based on shared sequences and evolutionary timelines

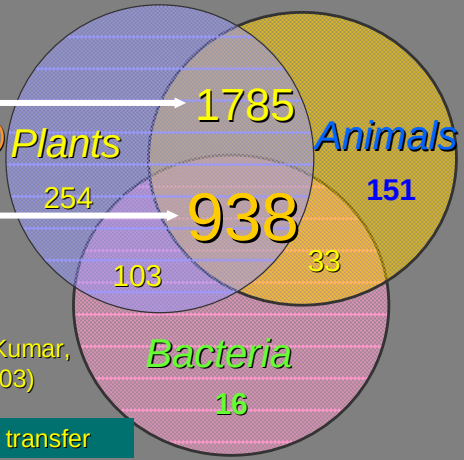
Age of yeast genes (6356 total)

3340 common fungal genes

> 1.2 billion years old
(asco/basidio split)

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

> 2.7 billion years old
(euk/prok split)



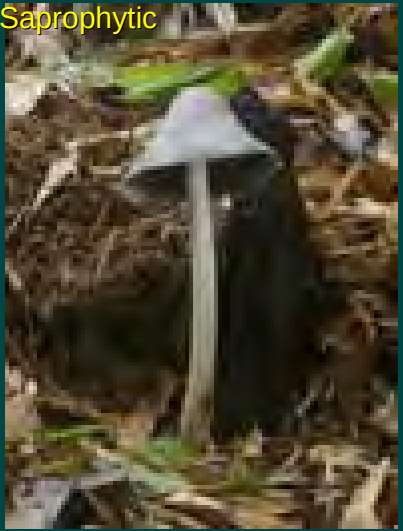
Timeline based on Hedges & Kumar, Trends in Genetics 19:200 (2003)

* assuming no horizontal gene transfer

Objective 4

- Search for evidence of horizontal gene transfer
- Compare the genome of the mycorrhizal mushroom, *Laccaria bicolor,* with available plant genomes and fungal genomes
  - search for homologs shared between *Laccaria* and plants but not other fungi tested
- Compare the genome of a saprophytic, non-mycorrhizal mushroom, *Coprinus cinereus,* with available plant genomes
  - contrast these results with the *Laccaria* results

Coprinus cinereus (inky cap)
Agaricales, Psathyrellaceae
Saprophytic



Laccaria bicolor
Agaricales, Tricholomataceae
Mycorrhizal

http://blogimg.goo.ne.jp/user_image/2a/d0/c1c08294d30cafdde1348763f7b704a4.jpg http://inocybes.free.fr/photos_JJW/Laccaria-bicolor-oignies-2.jpg

Mycorrhizae

- “myco” = fungus and “rhiza” = root
- Symbiotic association between roots & fungi
- Endomycorrhizae
  - Zygomycota fungi
  - 80% of all plants
  - vesicles and arbuscules in plant cells
- Ectomycorrhizae
  - mostly Basidiomycota fungi, some Ascomycota
  - temperate trees and shrubs
  - intercellular, with mantle & Hartig net

Endomycorrhizae

- arose maybe 400 million years ago
- important role in land colonization by plants?
- intimate associations with plants

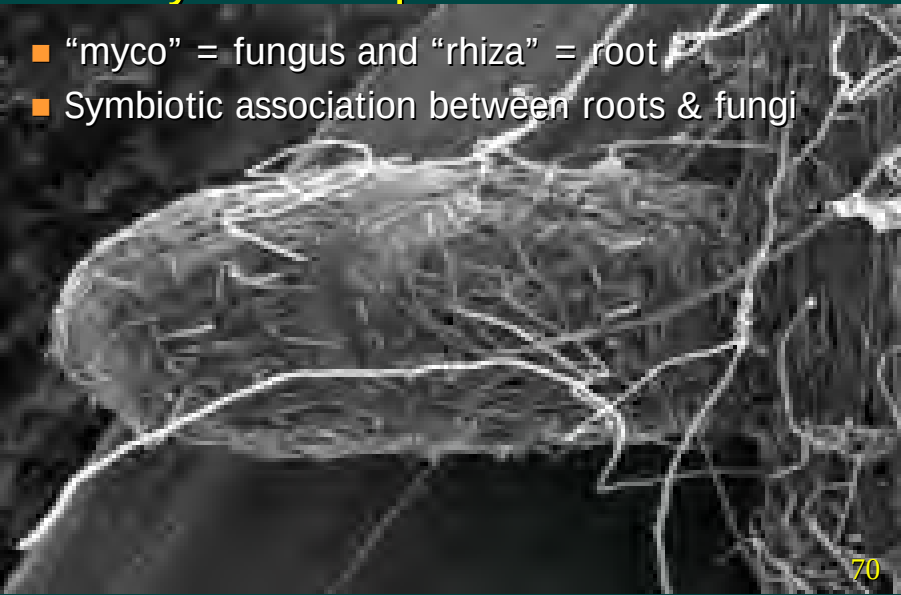
arbuscules for nutrient exchange

vesicles for storage

images from Edmonds, University of Washington

Ectomycorrhizal pine root

- "myco" = fungus and "rhiza" = root
- Symbiotic association between roots & fungi



Ectomycorrhizae

Mycorrhizal associations

- Ancient?
  - establishment of land plants?

- Intimate?
  - exchange nutrients
  - exchange genes?

20614 *Laccaria* predicted genes

12665 with homologs in at least one of 22 fungi

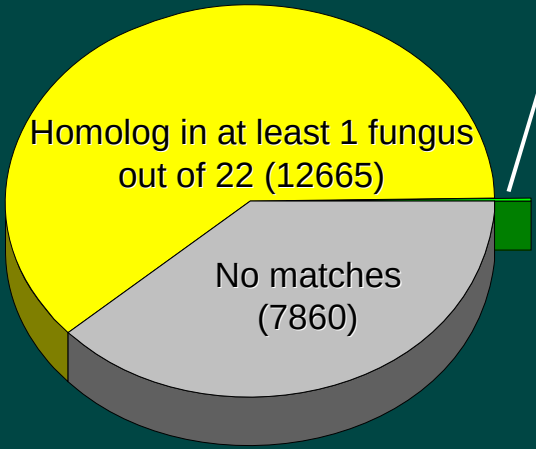
Ascomycota
Ashbya gossypii
Aspergillus fumigatus
Aspergillus nidulans
Botrytis cinerea
Candida albicans
Chaetomium globosum
Gibberella zeae
Magnaporthe grisea
Nectria haematococca
Neurospora crassa
Podospora anserina
Saccharomyces cerevisiae
Schizosaccharomyces pombe

Ascomycota (con't)
Sclerotinia sclerotiorum
Stagonospora nodorum
Trichoderma reesei
Uncinocarpus reesii

Basidiomycota
Phanerochaete chrysosporium
aphyllophorales
Coprinus cinereus
Cryptococcus neoformans
Ustilago maydis

Zygomycota *Rhizopus oryzae*

20614 *Laccaria bicolor* predicted genes compared to fungi & plants



Homolog in at least 1 of these plants, but not fungi
- *Arabidopsis thaliana*
- *Oryza sativa*
- *Populus trichocarpa*

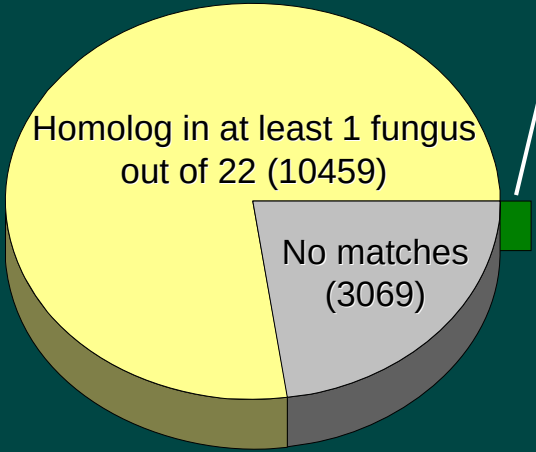
- Total: 89

89 *Laccaria* predicted genes with homologs in plants but not fungi compared to GenBank NR: top hits

- no matches: 17
- other fungi: 16
- animals: 16
- bacteria: 8
- protists: 4
- other: 2

- plants et al. : 26
- plants only: 19
  - e-value < 10^{-10}: 12
  - e-value < 10^{-20}: 7
  - e-value < 10^{-30}: 4
    - » peptidase-like proteins with 47-75% identity (not found in 22 other fungi)

13544 *Coprinus cinereus* predicted genes compared to fungi & plants



Homolog in at least 1 of these plants, but not fungi
- *Arabidopsis thaliana*
- *Oryza sativa*
- *Populus trichocarpa*

- Total: 16 with evalues ranging from 10^{-4} to 10^{-6}

Summary of mycorrhizal HGT

- 16 predicted genes (eval > 10^{-6}) of *Coprinus cinereus* had only a plant match in a test against 22 fungi and 3 plants genomes, vs. 89 predicted genes of *Laccaria bicolor*
- perhaps some of these resulted from HGT
- ongoing lab work to verify these findings

Conclusions

- Just over half (3340) of the genes in yeast (6356) have a homolog in 12 other fungi
- Almost 1000 common fungal genes have homologs in plants, animals and bacteria
- These 1000 genes may have originated before the origin of eukaryotes 2.7 billion years ago (ignoring horizontal gene transfer)
- Preliminary evidence of horizontal gene transfer to fungi from plants
  - *Laccaria bicolor* (from mycorrhizal host?)

Significance of Results

- Explore commonality between living organisms by focusing on highly conserved sequences, which may be useful for phylogeny
- Insight into evolutionary biology of fungi and fungal-plant interactions
  - how do pathogens differ from saprophytes or mutualists?
  - do fungal pathogens or mycorrhizae acquire plant genes to hide themselves from detection?
- J. Molec. Evol. 60:475-483 (obj. 1,2,3) available at: http://www.uoguelph.ca/~thsiang/pubs