Using Phylogenomics to Predict Novel Fungal Pathogenicity Genes

David DeCaprio, Ying Li, Hung Nguyen

(sequenced Ascomycetes genomes courtesy of the Broad Institute)

Phylogenomics

- Combining whole genome sequences and phylogenetic information to make inferences about gene function

- Phenotypic differences between related organisms can be explained in terms of genomic differences

Plant Pathogens

Magnaporthe grisea	First sequenced pathogen (rice)
Fusarium graminearum	Wheat blight – Worst US pathogen
Stagonospora nodorum	Another wheat blight

Other Ascomycete fungi

Neurospora crassa	Model organism
Aspergillus nidulans	Model organism
Chaetomium globosum	Human pathogen

Dean analysis – Nature 2005

- Identified putative pathogenicity genes in *Magnaporthe grisea*
- Found significant expansion in pth11-related gene family in *M. grisea* vs. *N. crassa*
  - pth11-class is required for pathogenicity (DeZwann 1999)

BUT

- *N. crassa* has RIP: all families are small

Testing the Hypothesis

- Examine pth11-related family in 3 pathogenic species vs. 3 non-pathogenic species

- If this family is pathogenicity related:
  - Expect to see expansion in pathogens
  - Expect to see no expansion in non-pathogens

Purpose of Phylogenetic Analysis

- Ralph Dean's *M. grisea* expansions may not actually be expansions when compared to more data sets and data sets with larger families
- If pth11 required for pathegenicity (appressorium development), other plant-pathogens may also contain expansions yet unidentified

Procedure

- Obtain gene families for six organisms
- Pick out pth11-related family
- Align all genes in family using ClustalW
- Build phylogenetic tree for family
- Bootstrap analysis on the tree
- Examine pathogenic expansions

Find Gene Families

- Obtain match score for each pair of proteins in all genes in all organisms
  - Obtain an all-to-all (but not identity) match of all proteins in all 6 organisms using NCBI-BLAST with default parameters and an E-value cutoff of 1e-10.
  - Combine all hits between a pair of proteins to obtain score between the pair
  - Select highest scoring non overlapping hits per pair. (coverage > 60% of shorter protein AND average identity > 30%)

Find Gene Famlies (cont.)

- COG Single Linkage Clustering
  - Each protein initially in its own cluster
  - Merge clusters with BLAST hits between any of their proteins
  - Each cluster is a gene family

Family Analysis

- Three plant-pathogenic species (*S. nodorum, M. grisea, F. graminearum*) showed most expansions across all families
- 316 genes in PTH11 family:
  - *F. graminearum:* 89
  - *S. nodorum:* 87
  - *M. grisea:* 49
  - *A. nidulans:* 42
  - *C. globosum*: 28
  - *N. crassa:* 21

Procedure: Phylogenetic Analysis

- Aligned all 316 sequences in pth11-related family using ClustalW
- Used Phylip to generate phylogenetic tree (parsimony method)
- Compared portions of tree that match Ralph Dean's *M. grisea* vs *N. crassa* tree
- Choose subset of family to examine further to test pathegenic expansion hypothesis

Phylogenetic Analysis (background)

- Tree-building Algorithms:
  - Distance and Nearest Neighbor Joining
  - Maximum Parsimony
  - Maximum Likelihood
- Sequence order
  - Input sequence order may influence tree found. Solution: randomize (jumble) input

Distance and Neighbor Joining

- Iteratively join closest (distance-wise) nodes
- Distance between two sequences = % sites different between them in an alignment
- Distance between a sequence and a joined node = Average distance between sequence and node

Maximum Parsimony

- Character-based method
- Chooses a tree that minimizes the number of mutational events (substitutions)
- Computationally inexpensive to run

Maximum Likelihood

- Best accounts for variation in sequences
- Likelihood *L* of a tree is the probability of observing the data given the tree
  $L = P(\text{data}|\text{tree})$
- Search all possible trees for one with highest probability P(data | tree)
- Extremely computationally intensive

Bootstrap Analysis

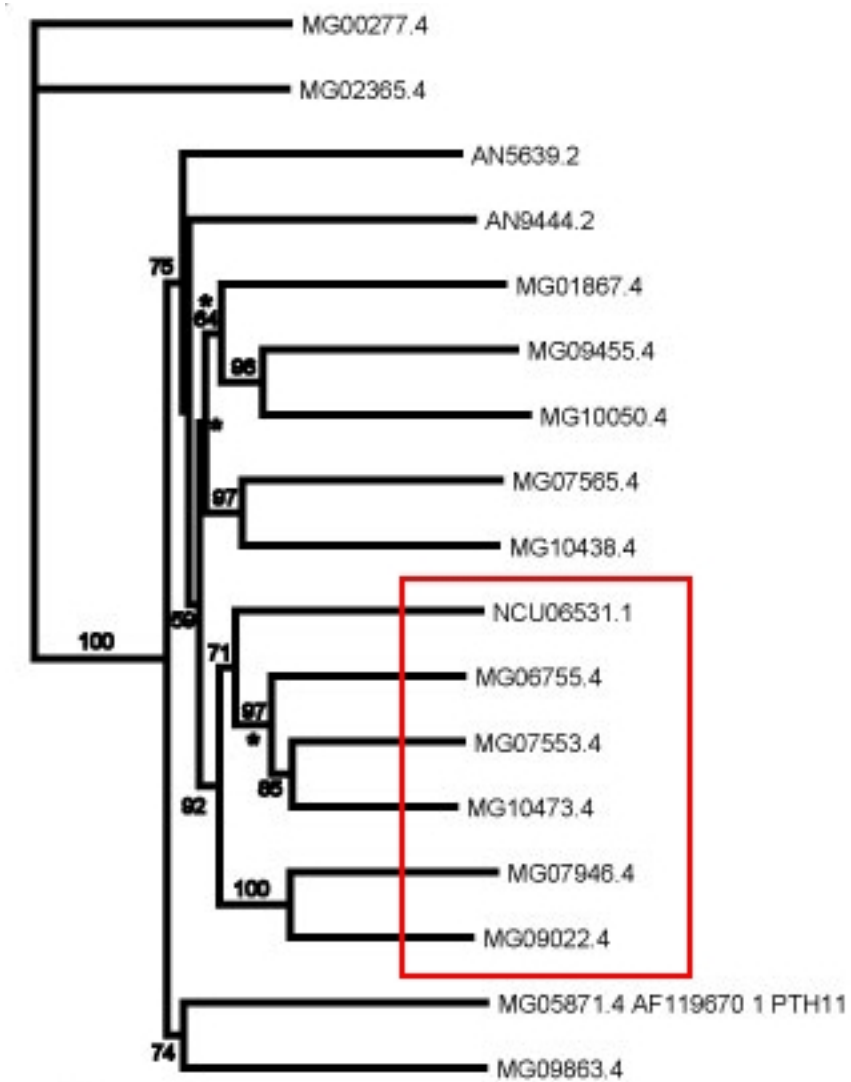
- Statistical technique to measure level of confidence in a previously generated tree
- Resample alignment multiple times and generate a tree for each
- Consensus tree (each branch chosen if it appears in a majority of resamplings) gives confidence values for each branch of tree

Our Initial Tree

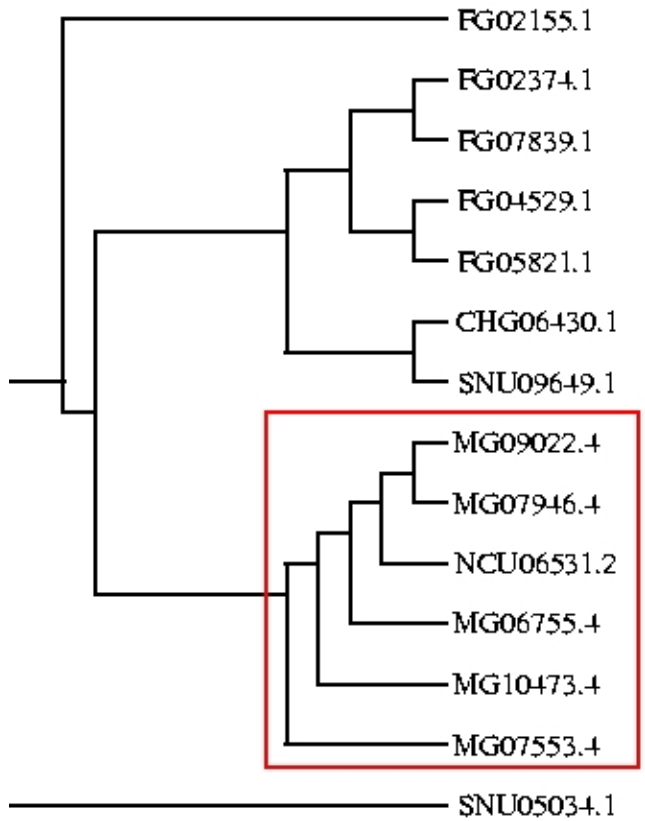
- Generated using maximum parsimony
- Did not have enough time for verification
  - maximum likelihood = too computationally intensive
  - bootstrap analysis = computationally intense and too many iterations
- Difficult to analyze entire tree due to size and complexity

Initial results

- Part of our tree matched significant portion of Ralph Dean's tree exactly
- Exact match significant since we aligned genes with genes from 5 other organisms, 2 of which are also pathegenic
- High confidence in *M. grisea* expansion



Dean (2005)



Portion of our initial tree (parsimony)

Courtesy of Ralph A. Dean. Used with permission.
Source: Supplementary Figure S1 in Dean R, et. al. "The genome sequence of the rice blast fungus Magnaporthe grisea." Nature 434, 980-986 (21 April 2005).

Narrowing the scope

- Choose a subtree relevant to our hypothesis
- Confirm structure built by our original tree
- Examine age of expansions
- Chose subtree of 33 members:
  - It contains the exact match to Ralph Dean's tree
  - It contains nearby *S. globosum* and *F. graminearum* expansions

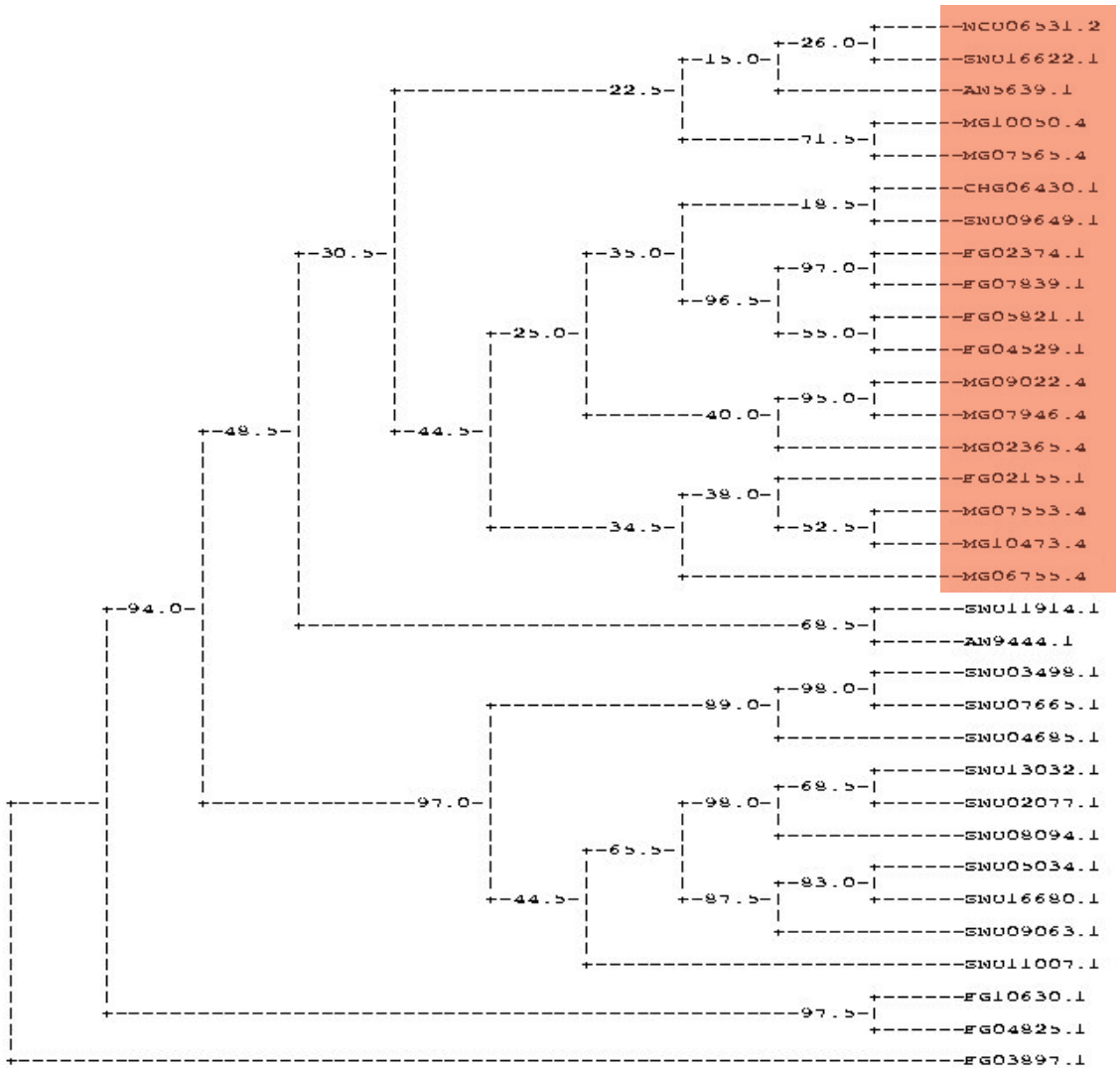
Subtree Analysis

- Realigned sequences
- Built new tree using maximum parsimony (jumbled and ordered), maximum likelihood (ordered), distance and nearest neighbor.
- Ran bootstrap analysis (parsimony - jumbled and ordered, nearest neighbor)
- Removed genes placed as outgroups in the new tree

Subtree Analysis

- Expansion was consistent with pathogenicity
- High level of confidence in expansions
  - Present in trees from different algorithms
  - Supported by bootstrap trials
- 30 genes in resulting subtree:
  - *F. graminearum:* 5
  - *S. nodorum:* 13
  - *M. grisea:* 8
  - *A. nidulans:* 2
  - *C. globosum*: 1
  - *N. crassa:* 1

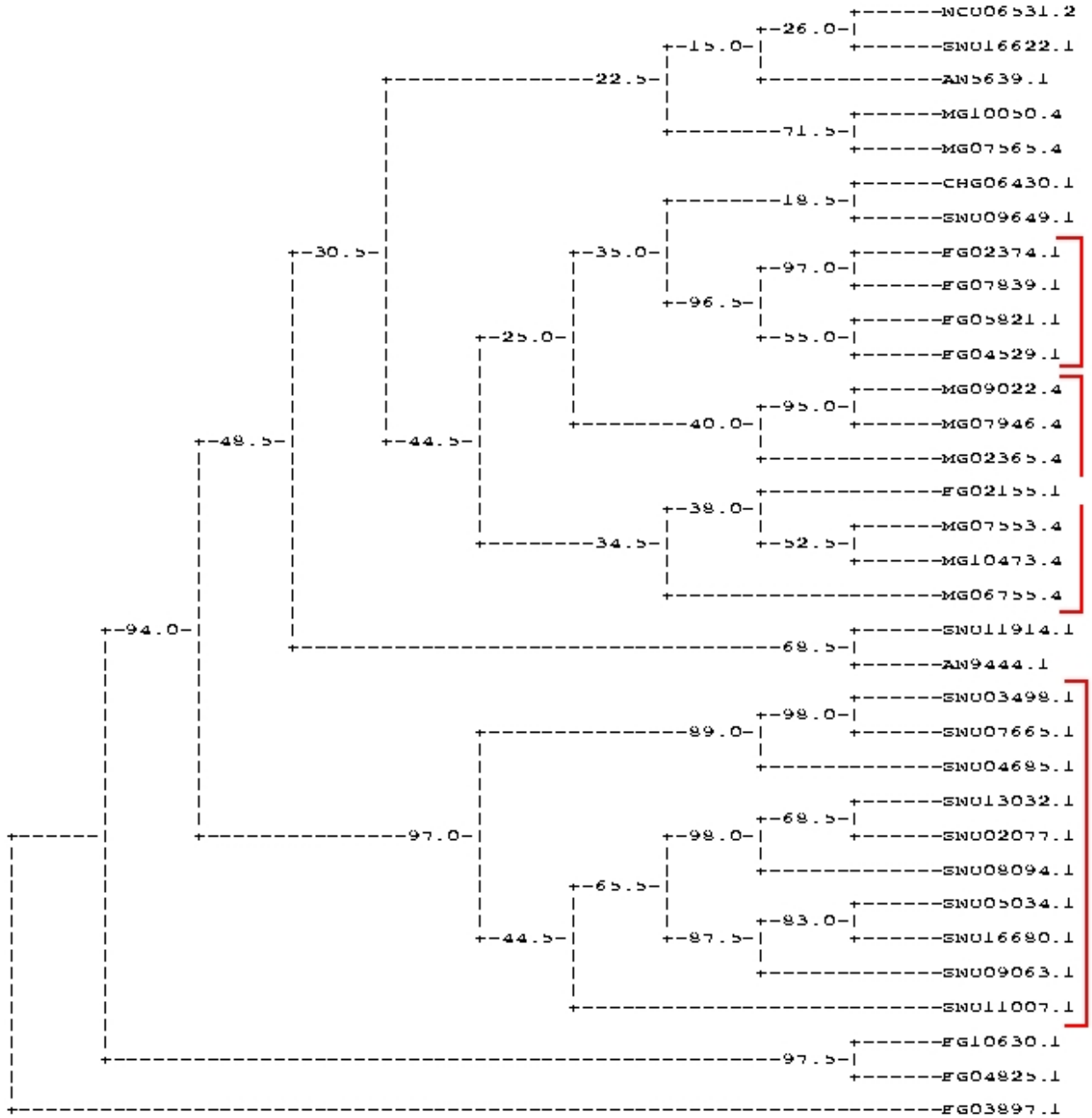
CFEM Domain


NCU06531.2
SNU16622.1
AN5639.1
MG10050.4
MG07565.4
CHG06430.1
SNU09649.1
FG02374.1
FG07839.1
FG05821.1
FG04529.1
MG09022.4
MG07946.4
MG02365.4
FG02155.1
MG07553.4
MG10473.4
MG06755.4
SNU11914.1
AN9444.1
SNU03498.1
SNU07665.1
SNU04685.1
SNU13032.1
SNU02077.1
SNU08094.1
SNU05034.1
SNU16680.1
SNU09063.1
SNU11007.1
FG10630.1
FG04825.1
FG03897.1
26.0
15.0
22.5
71.5
18.5
35.0
97.0
96.5
25.0
55.0
95.0
30.5
40.0
48.5
44.5
38.0
34.5
52.5
94.0
68.5
98.0
89.0
68.5
97.0
98.0
65.5
93.0
44.5
87.5
97.5

Domains

- CFEM domain
  - Present in 23 of 316 sequences in larger family
  - Present 18 contiguous sequences in our tree
- PFamB_10167
  - Present in 29 genes of our subtree
  - Unable to test larger family due to time constraints

Species Specific Expansions



F. graminaerum

M. grisea

S. Nodorum

Results

- Confirmed expansion in *M. grisea* as reported by Ralph Dean.
- This gene family expansion was consistent with pathogenicity across 6 organisms
- Expansions species-specific: each plant pathogen developed expansions independently
- Predicted 13 putative pathogenicity genes from *S. nodorum* and 4 in *F. graminearum*

Weaknesses in our Approach

- Initial tree was unfiltered: 316 genes were generated using BLAST only.
  - Improvement: filter out those that do not contain the CFEM protein domain or the PFamB_10167 domain
- We chose our subtree somewhat arbitrarily because we located putative *F. graminearum* and *S. globosum* expansions
  - Improvement: given time and a filtered family sequence, analyze all expansions across all genes in family to estimate significance

Weaknesses Cont'd:

- Didn't capture all of Dean's original pth11 family in our family.

References

Dean R, et. al. “The genome sequence of the rice blast fungus *Magnaporthe grisea*.” *Nature*. 2005 Apr 21;434(7036):980-6.

Kulkarni RD, et. al. “Novel G-protein-coupled receptor-like proteins in the plant pathogenic fungus *Magnaporthe grisea*.” *Genome Biol*. 2005;6(3):R24.

Hu G, et. al. “A phylogenomic approach to reconstructing the diversification of serine proteases in fungi.” *J Evol Biol*. 2004 Nov;17(6):1204-14.

DeZwann TM, et. al. “Magnaporthe grisea pth11p is a novel plasma membrane protein that mediates appressorium differentiation in response to inductive substrate cues.” *Plant Cell*. 1999 Oct;11(10):2013-30.

References

- Dean R, et. al. "The genome sequence of the rice blast fungus *Magnaporthe grisea*." *Nature*. 2005 Apr 21;434(7036):980-6.
- DeZwann TM, et. al. "Magnaporthe grisea pth11p is a novel plasma membrane protein that mediates appressorium differentiation in response to inductive substrate cues." *Plant Cell*. 1999 Oct;11(10):2013-30.
- Kulkarni RD, et. al. "Novel G-protein-coupled receptor-like proteins in the plant pathogenic fungus *Magnaporthe grisea*." *Genome Biol*. 2005;6(3):R24.
- Hu G, et. al. "A phylogenomic approach to reconstructing the diversification of serine proteases in fungi." *J Evol Biol*. 2004 Nov;17(6):1204-14.
- PHYLIP (U. Washington) http://evolution.genetics.washington.edu/phylip.html
- CLUSTALW (EMBL-EBI) http://www.ebi.ac.uk/clustalw/
- BLAST (NCBI)
- Broad Institute (MIT) http://www.broad.mit.edu
- PFAM (Sanger) http://www.sanger.ac.uk/Software/Pfam