Lawrence Berkeley National Laboratory
Recent Work

Title
Comparative Analysis of Thermophillic Fungal Genomes

Permalink
https://escholarship.org/uc/item/0p430092

Authors
Otillar, Robert
Salamov, Asaf
Atkins, Alex
et al.

Publication Date
2010-03-24

eScholarship.org Powered by the California Digital Library
University of California

Comparative Analysis of Thermophillic Fungal Genomes

Robert Otillar[1], Asaf Salamov[1], Alex Atkins[1], Frank Korzeniewski[1], Jeremy Schmutz[2], Erika Lindquist[1], Adrian Tsang[3], Randy Berka[4], Igor Grigoriev[1]

[1] DOE Joint Genome Institute, Walnut Creek, CA; [2]Hudson Alpha Institute for Biotechnology, Huntsville, AL; [3]Concordia University, Montreal, Quebec Canada [4]Novozymes, Inc., Davis, CA

Abstract

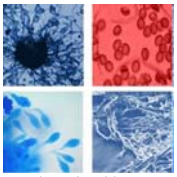
Rapid, efficient, and robust enzymatic degradation of biomass-derived polymers is currently a major obstacle in biofuel production. A key missing component in that process is the availability of enzymes that hydrolyze cellulose, hemicellulose, and other polysaccharides into biofuel substrates at temperatures and chemical conditions suitable for industrial use. Thermophilic fungi are known to excrete enzymes that rapidly degrade polymers including cellulose and hemicellulose at high temperatures, however the genome sequence and full gene complement of thermophillic fungi have not been previously reported. Here we describe the initial sequencing, gene identification, and comparative analysis of two thermophilic fungi, *Thielavia terrestris* and *Sporotrichum thermophile*, with comparison to *Chaetomium globosum,* a closely related non-thermophillic fungal species.



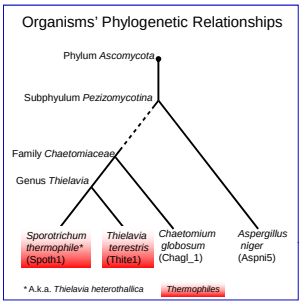
Sporotrichum thermophile
photo Courtesy of Tricia John



Thielavia terrestris
photo courtesy of Randy Berka



Chaetomium globosum

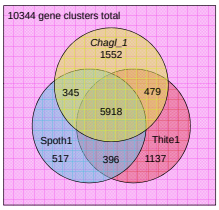


Assembly Comparison	*Sporotrichum thermophile*	*Thielavia terrestris*	*Chaetomium globosum*
Assembly	Spoth1	Thite1	Chagl_1
Genome size(Mbp)	38.7	36.9	34.9
Sequencing read coverage depth	7.66x	10.15x	7x
Reported # of contigs	293	231	2453
# of scaffolds	10	8	37
# of scaffolds >2 Kbp	10	8	32
Scaffold N/L50	3/5.4 Mbp	3/4.6 MB	4/4.7 Mbp
Three largest Scaffolds (Mbp)	7.0,7.0, 5.4	9.5, 6.0, 4.6	6.6, 5.0, 4.7
Sequencing Method	Sanger WGS	Sanger WGS	Sanger WGS
% genome in gaps	0.34%	0.41%	1.58%
% genome in repeats	31%	22%	9%
# genes	8,806	9,815	11,124
Gene density (genes/Mbp)	228	266	319

Note: Organelle DNA was excluded from Spoth1 and Thite1 analysis
Note: Chagl_1 includes 695 gene models that would have been eliminated by JGI length & repeat match filteriNote: Chagl_1 includes 695 gene models that would have been eliminated by JGI length & repeat match filtering

Chaetomiaceae family Gene Clusters

# clusters total	10,344
# Spoth1 clusters	7,176
# Thite1 clusters	7,930
# Chagl_1 clusters	8,294
# Shared across all 3 genomes	57%
# clusters equal across 3 genomes	48%



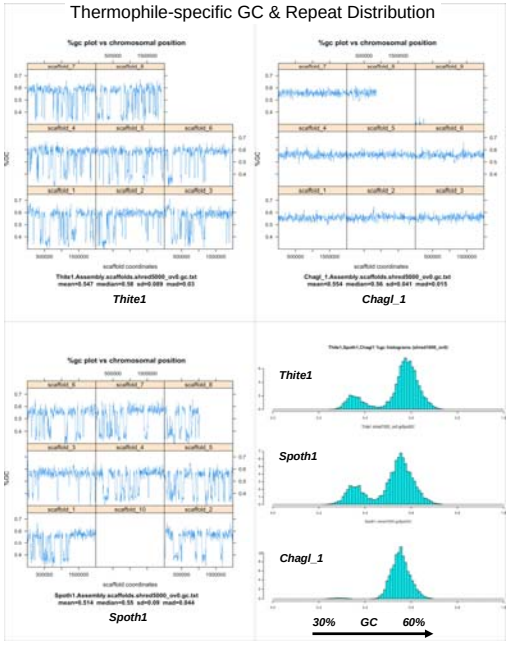
Largest Gene Clusters

	# of genes per cluster			
# clusters	**Spoth1**	**Thite1**	**Chagl_1**	**% total**
4,810	1	1	1	47%
1,436	-	-	1	14%
1,079	-	1	-	10%
482	1	-	-	5%
367	-	1	1	4%
346	1	1	-	3%
269	1	-	1	3%
1,555	*other*			

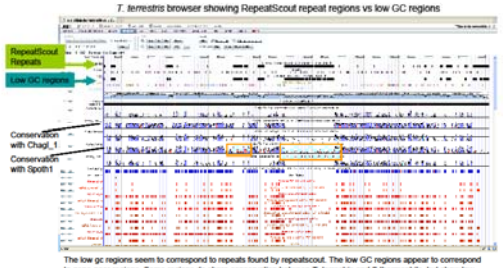
~1/2 of clusters are 1:1:1 ortholog candidates

85% of clusters have =< 1 gene per species

Thermophile-specific GC & Repeat Distribution



T. terrestris browser showing RepeatScout repeat regions vs low GC regions



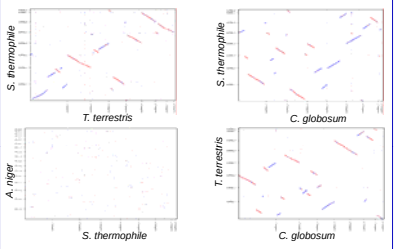
The low gc regions seem to correspond to repeats found by repeatscout. The low GC regions appear to correspond to gene-poor regions. Some regions do share conservation between T. terrestris and S.thermophile, but show low conservation with other fungi, e.g. C. globosum, as illustrated in the orange-boxed regions.

- Thermophiles have undergone substantial, spatially-distributed repeat expansion
- The repeat expansion is characterized by low GC regions (~30-40% GC)
- Chagl_1 scaffold_9, 10, 12, 13, 14, 16, 17, 18, 19, 21, 23, 24, 26, 27, 28, 30, 33, 34 are nearly entirely covered by repeats & lo GC regions. Other than these ~all-lo GC scaffolds, there are only 5 lo GC 1kb segments in the rest of the genome. (data not shown)

Genome conservation

VISTA DNA Alignment Coverage

genomes		Total	UTR	CDS
S. thermophile	*T. terrestris*	68%	82%	87%
S. thermophile	*C. globosum*	66%	88%	93%
T. terrestris	*C. globosum*	67%	83%	89%
S. thermophile	*A. niger*	12%	6%	46%



- Thermophiles' exon sequence conservation *lower* than with C. globosum
- Long syntenic regions across all 3 *Chaetomiaceae* family members
- Comparatively low synteny visible between 3 family members and *A. niger*

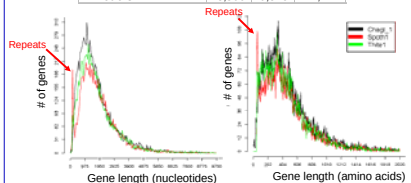
Thermophile-Specific Functional Domains (PFAM)

Thermophile-only domains:

# domains total	# Spoth1	# Thite1	# Chagl_1	# Aspni5	descriptions
3	1	2	0	0	Cupin superfamily (DUF985)
3	1	2	0	0	O-Glycosyl hydrolase family 30
2	1	1	0	0	Alanine racemase, N-terminal domain
2	1	1	0	0	Chromo shadow domain
2	1	1	0	0	Conserved hypothetical protein CHP02453
2	1	1	0	0	DASH complex subunit Dad4
2	1	1	0	0	Double-stranded DNA-binding domain
2	1	1	0	0	Dpy-30 motif
2	1	1	0	0	Eukaryotic protein of unknown function (DUF866)
2	1	1	0	0	Glycosyl hydrolases family 25
2	1	1	0	0	GRIM-19 protein
2	1	1	0	0	HIUase/Transthyretin family
2	1	1	0	0	Molydopterin dinucleotide binding domain
2	1	1	0	0	NHL repeat
....	(80 total)				

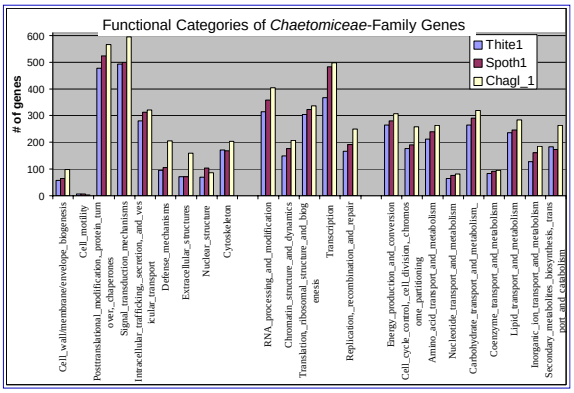
Gene-Model Structure Comparison

Model Statistics	Spoth1	Thite1	Chagl_1
gene length (bp)	1,879	1,823	1,785
transcript length (bp)	1,630	1,549	1,481
protein length (aa)	478	466	493
exons per gene	3.02	3.05	3.12
exon length (bp)	540	508	475
intron length (bp)	125	136	145
# models	8,806	9,815	11,124



- Thermophiles have fewer genes, but no other striking patterns vs. *C. globosum*

Functional Categories of *Chaetomiceae*-Family Genes



This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.