nature biotechnology

Comparative genomic analysis of the thermophilic biomass-degrading fungi *Myceliophthora thermophila* and *Thielavia terrestris*

Randy M Berka[1,15], Igor V Grigoriev[2,15], Robert Otillar[2], Asaf Salamov[2], Jane Grimwood[3], Ian Reid[4], Nadeeza Ishmael[4], Tricia John[4], Corinne Darmond[4], Marie-Claude Moisan[4], Bernard Henrissat[5], Pedro M Coutinho[5], Vincent Lombard[5], Donald O Natvig[6], Erika Lindquist[2], Jeremy Schmutz[3], Susan Lucas[2], Paul Harris[1], Justin Powlowski[4], Annie Bellemare[4], David Taylor[4], Gregory Butler[4], Ronald P de Vries[7,8], Iris E Allijn[7], Joost van den Brink[7], Sophia Ushinsky[4], Reginald Storms[4], Amy J Powell[9], Ian T Paulsen[10], Liam D H Elbourne[10], Scott E Baker[11], Jon Magnuson[11], Sylvie LaBoissiere[12], A John Clutterbuck[13], Diego Martinez[6,14], Mark Wogulis[1], Alfredo Lopez de Leon[1], Michael W Rey[1] & Adrian Tsang[4,15]

Thermostable enzymes and thermophilic cell factories may afford economic advantages in the production of many chemicals and biomass-based fuels. Here we describe and compare the genomes of two thermophilic fungi, *Myceliophthora thermophila* and *Thielavia terrestris*. To our knowledge, these genomes are the first described for thermophilic eukaryotes and the first complete telomere-to-telomere genomes for filamentous fungi. Genome analyses and experimental data suggest that both thermophiles are capable of hydrolyzing all major polysaccharides found in biomass. Examination of transcriptome data and secreted proteins suggests that the two fungi use shared approaches in the hydrolysis of cellulose and xylan but distinct mechanisms in pectin degradation. Characterization of the biomass-hydrolyzing activity of recombinant enzymes suggests that these organisms are highly efficient in biomass decomposition at both moderate and high temperatures. Furthermore, we present evidence suggesting that aside from representing a potential reservoir of thermostable enzymes, thermophilic fungi are amenable to manipulation using classical and molecular genetics.

Rapid, efficient and robust enzymatic degradation of biomass-derived polysaccharides is currently a major challenge for biofuel production. A prerequisite is the availability of enzymes that hydrolyze cellulose, hemicellulose and other polysaccharides into fermentable sugars at conditions suitable for industrial use. The best studied and most widely used cellulases and hemicellulases are produced by *Trichoderma*, *Aspergillus* and *Penicillium* species, and they are most effective over a temperature range from 40 °C to ~50 °C. At these temperatures, complete saccharification of biomass polysaccharides (>90% conversion to fermentable sugars) requires long reaction times, during which hydrolysis reactors are susceptible to contamination. One way to overcome these obstacles is to raise the reaction temperature, thereby increasing hydrolytic rates and reducing contamination risks. However, implementing higher reaction temperatures requires the deployment of enzymes that are more thermostable than the available preparations from mesophilic fungi. Additional advantages of elevated hydrolysis temperatures include enhanced mass transfer, reduced substrate viscosity and the potential for enzyme recycling[1].

Thermophilic fungi represent a potential reservoir of thermostable enzymes for industrial applications. They can also potentially be developed into cell factories to support production of chemicals and materials at elevated temperatures. Enzymes from thermophilic fungi often tolerate higher temperatures than enzymes from mesophilic species, and some show stability at 70–80 °C (refs. 1,2). Notably, it has been reported the cellulolytic activity of some thermophilic species was several times higher than that of the most active cellulolytic mesophiles[3]. Furthermore, biomass-degrading enzymes from thermophilic fungi consistently demonstrate higher hydrolytic capacity[4] despite the fact that extracellular enzyme titers (in grams per liter) are typically lower than those from more conventionally used species

[1]Novozymes, Inc., Davis, California, USA. [2]US Department of Energy Joint Genome Institute, Walnut Creek, California, USA. [3]HudsonAlpha Institute for Biotechnology, Huntsville, Alabama, USA. [4]Centre for Structural and Functional Genomics, Concordia University, Montreal, Quebec, Canada. [5]Architecture et Fonction des Macromolécules Biologiques, CNRS/Universités de Provence/Université de la Mediterranée, Marseille, France. [6]Department of Biology, University of New Mexico, Albuquerque, New Mexico, USA. [7]CBS-KNAW Fungal Biodiversity Centre, Utrecht, The Netherlands. [8]Microbiology and Kluyver Centre for Genomics of Industrial Fermentation, Utrecht University, Utrecht, The Netherlands. [9]Sandia National Laboratory, Albuquerque, New Mexico, USA. [10]Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, Australia. [11]Fungal Biotechnology Team, Pacific Northwest National Laboratory, Richland, Washington, USA. [12]McGill University and Génome Québec Innovation Centre, Montreal, Canada. [13]University of Glasgow, Glasgow, UK. [14]Present address: Broad Institute of MIT & Harvard, Cambridge, Massachusetts USA. [15]These authors contributed equally to this work. Correspondence should be addressed to A.T. (tsang@gene.concordia.ca).

Received 16 May; accepted 18 August; published online 2 October 2011; doi:10.1038/nbt.1976

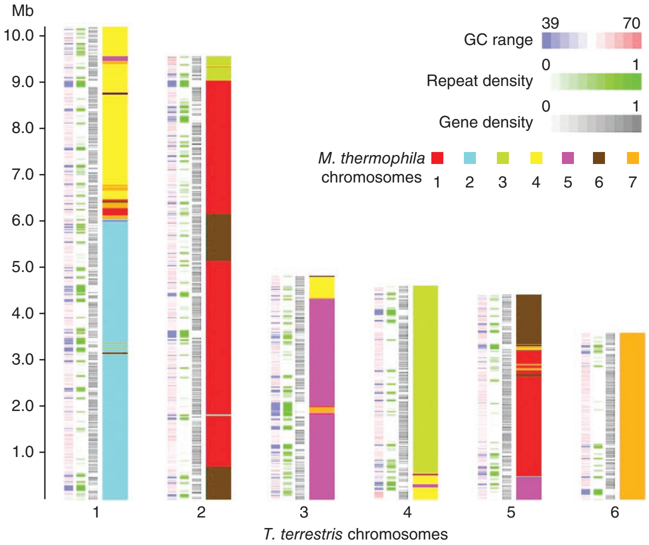

Figure 1 Genome organization of *M. thermophila* and *T. terrestris*. The six chromosomes of *T. terrestris* are mapped to genomic regions from *M. thermophila* (shown as colored blocks in far-right lane). Only major synteny blocks are represented. For each *T. terrestris* chromosome, left-most lane shows G+C content, second lane from left shows repetitive elements and third lane from left shows regions with high gene density.

such as *Trichoderma* or *Aspergillus*. We describe comparative genomic analyses of two thermophilic ascomycete species, *Thielavia terrestris* and *Myceliophthora thermophila,* to our knowledge the first filamentous fungi with finished genomes. These sequences open the way for new industrial applications of the enzymes from these organisms and potential development of thermophilic fungal production hosts.

RESULTS

Genomes summary

Among thermophilic fungi, *M. thermophila* and *T. terrestris* are two of the best characterized in terms of thermostable enzymes and cellulolytic activity[1–4]. The fermentation characteristics of these two organisms have been examined and found to be suitable for large-scale production[5,6]. The finished 38,744,216-bp genome of *M. thermophila* and 36,912,256 bp genome of *T. terrestris* contain, respectively, seven and six complete telomere-to-telomere chromosomes (**Fig. 1** and **Table 1**). Their telomeres comprise TTAGGG repeats commonly found in telomeres of filamentous fungi. The two genomes are similar in organization. The major difference occurs in chromosome (Ch)1 of *T. terrestris*, which harbors most of the genes located on Ch2 and Ch4 of *M. thermophila*. In addition, extensive translocation is observed between Ch1/Ch6 of *M. thermophila* and Ch2/Ch5 of *T. terrestris*. The protein coding fractions of the genomes include 9,110 genes in *M. thermophila* and 9,813 genes in *T. terrestris* (**Table 1**); both are smaller than average proteomes of other fungi in the class Sordariomycetes, and substantially smaller than the closely related mesophile *Chaetomium globosum*[7], which has 11,124 predicted genes in a 34.9-Mbp genome. These three species within the family Chaetomiaceae share 6,279 three-way orthologs and extensive synteny with >6,000 genes in syntenic blocks between each pair, including four blocks of >400 genes between *M. thermophila* and *T. terrestris* (**Supplementary Fig. 1** and **Supplementary Table 1**). The breakpoints of the synteny blocks often coincide with AT-rich repetitive regions (**Fig. 1**). The largest gene families in the genomes of *M. thermophila* and *T. terrestris* include transporters (e.g., MFS, ABC, AAA and sugar transporters) and proteins involved in signaling (e.g., protein kinases and WD40) as shown in **Supplementary Table 2**, often with more genes of each type in *T. terrestris*. Several Pfam domains appear to be expanded in the Chaetomiaceae, including glycoside hydrolase families GH61 and GH11, and hypothetical proteins with a DUF1996 domain of unknown function (**Supplementary Table 3**).

Enzymes for biomass degradation

Proteins encoded in the genomes of *T. terrestris* and *M. thermophila* were compared to eight other fungi for genes encoding carbohydrate-active proteins[8] (CAZymes): glycoside hydrolases (GHs), polysaccharide lyases (PLs), carbohydrate esterases, glycosyl transferases (GT) and carbohydrate-binding modules (**Supplementary Table 4**). Like the other fungi examined, the two thermophiles harbor large numbers (>210) of glycoside hydrolases and polysaccharide lyases covering most of the recognized families, albeit with important differences (**Supplementary Figs. 2** and **3** and **Supplementary Tables 5** and **6**). For instance *T. terrestris* is poor in pectin and pectate lyases (no PL1, PL3, PL9 and PL11) and relatively rich in polygalacturonases (seven GH28). In contrast, the reverse is true for *M. thermophila* (five PL1, one PL3 and two GH28). Pectin lyases are most active at neutral to alkaline pH whereas GH28 pectin hydrolases are most active in acidic pH. Consistent with their repertoires of pectinolytic enzymes, *M. thermophila* grows best on pectin at neutral to alkaline pH whereas the growth of *T. terrestris* on pectin is best at acidic pH (**Supplementary Fig. 4**). The two thermophiles can be considered all-purpose decomposers with respect to their CAZymes and their ability to degrade plant polysaccharides (**Supplementary Fig. 5**).

Compared to the paradigmatic cellulase producer, *Trichoderma reesei*, the two thermophiles have similar complements of GH proteins. A major difference is the clear expansion of the GH61 family, and to a lesser extent the GH10 and GH11 xylanases, in members of the family Chaetomiaceae examined in this study (at least 18 GH61 proteins for the Chaetomiaceae and three for *T. reesei*) (**Supplementary Tables 5** and **6**). The GH61 family was originally classified on the basis of very weak endo-1,4-β-D-glucanase activity found in one family member[9]. Recently, it was reported that certain GH61 proteins lack measurable hydrolytic activity by themselves, but in the presence of various divalent metal ions, they can substantially enhance lignocellulosic biomass hydrolysis by cellulases and reduce the amount of cellulase required for hydrolysis of biomass polysaccharides[10]. The expansion of GH61 genes in this group of fungi may have evolved as a modified strategy for deconstruction of biomass polysaccharides compared to that of other species such as *T. reesei* is suggested by the evolutionary relationship of GH61 proteins among selected species in the order Sordariales, whose members can be divided into 25 orthologous clades (designated A–Y) (**Supplementary Fig. 6**). The fact that these organisms have maintained a diverse array of GH61 genes throughout their evolution intimates their importance for degradation of plant cell wall polysaccharides, with differing GH61 types possibly acting on

Table 1 Assembly and gene model statistics

	T. terrestris	*M. thermophila*
G+C content,%	54.7	51.4
Genome size	36.9 Mb	38.7 Mb
No. of chromosomes	6	7
No. of genes	9,813	9,110
Gene length (nt)[a]	1,649	1,733
Exons per gene[a]	2	2

[a] Median values over all genes in an organism.

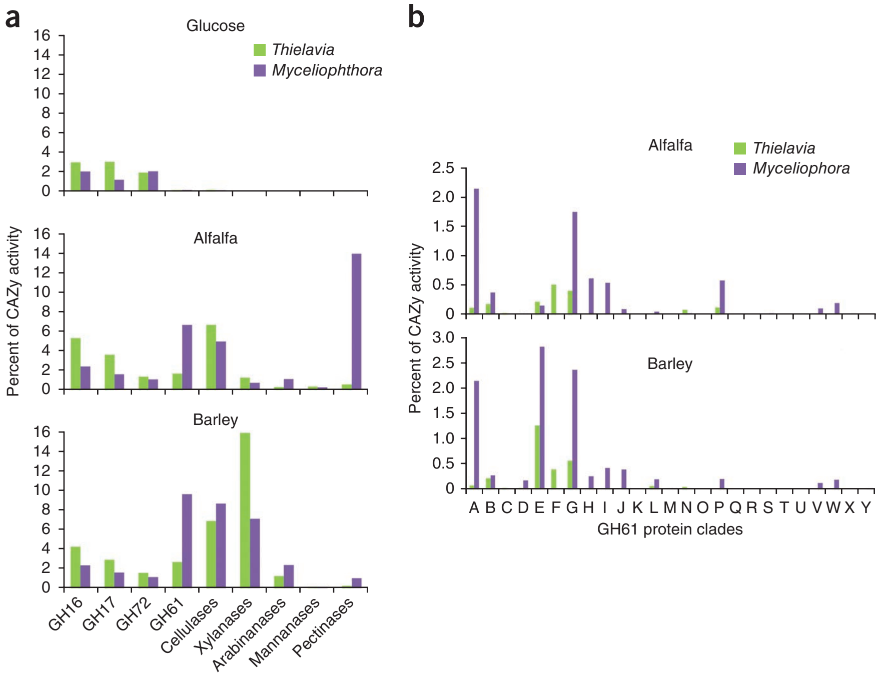

Figure 2 Analysis of transcription profiles. (**a**) Expression of CAZymes genes of the thermophiles cultured on glucose, alfalfa straw and barley straw. Gene activity is presented as percentage of total CAZymes gene activity of the three culture conditions of each organism. For the present analysis, cellulases include endoglucanases and cellobiohydrolases from GH5, GH6, GH7, GH12 and GH45; xylanases refer to GH10 and GH11 endoxylanases; arabinanases are endoarabinanases and arabinosidases from GH43, GH51 and GH62; mannanases are GH5 and GH26 endomannanases; and pectinases include polygalacturonases, rhamnogalacturonases, pectin lyases and pectin esterases from GH28, PL1, PL3, PL4, CE8 and CE12. (**b**) Transcript profiles of GH61 orthologs in *M. thermophila* and *T. terrestris*. The homologs of GH61 of *M. thermophila* and *T. terrestris* are organized in clades as shown in **Supplementary Figure 6**. Gene activity is presented as percentage of total CAZymes gene activity of the three culture conditions of each organism. Genes of several clades (e.g., K, O, R, S and T) are not upregulated in any of the growth conditions. None of the genes encoding GH61 proteins are upregulated during growth on glucose (**Supplementary Tables 5–7**).

assorted substrates and/or possessing varied biochemical properties. Differential expression of discrete GH61 subtypes (noted below) supports this view.

Transcript profiles on biomass substrates

To examine the strategy used by these thermophiles for decomposition of plant cell wall polysaccharides, we used RNA-Seq to compare transcript profiles during growth on barley straw or alfalfa straw to growth on glucose. Alfalfa was chosen to represent dicotyledonous plants, whereas barley was used to represent monocotyledon plants. The major difference between these materials is that the carbohydrates from barley cell wall are mainly cellulose and hemicellulose with a negligible amount of pectin[11], whereas alfalfa cell wall contains pectin and xylan in roughly similar proportions, each consisting of 15–20% of total carbohydrates[12].

We observed notable differences between the transcriptional profiles of genes encoding different classes of carbohydrate-active enzymes (**Fig. 2a** and **Supplementary Tables 5–7**). As expected, the genes encoding enzymes used for modification of fungal cell walls (e.g., GH16, GH17 and GH72 proteins) are expressed at similar levels during growth on glucose and plant straws for the two organisms. In contrast, transcripts encoding enzymes that deconstruct plant cell walls are upregulated only during growth on barley or alfalfa straws. For growth on barley straw, the induced transcripts correspond closely with the substrate composition; genes for cellulolytic and xylanolytic enzymes are highly upregulated, followed by genes for arabinanases, mannanases and to a lesser extent pectinolytic enzymes. However, such simple matching of gene activity and substrate constituents does not extend to growth on alfalfa straw, especially for *T. terrestris*. Though upregulated when compared to growth on glucose, transcripts encoding xylanolytic enzymes during growth on alfalfa straw are kept at a relatively low level in both organisms. Transcripts encoding pectin lyases are highly upregulated for *M. thermophila* but not for *T. terrestris*, accentuating alternative strategies used by these two organisms for degradation of pectin. *T. terrestris* degrades pectin using primarily hydrolase activity (GH28), whereas *M. thermophila* employs predominantly pectin lyases. When grown on complex substrates, several genes encoding GH61 proteins are highly upregulated in the thermophiles, particularly on barley straw for *M. thermophila*.

Fungi that decompose plant biomass typically possess multiple genes for degradation of a given polysaccharide polymer, and the thermophiles are no exception. Orthologs in these two genomes display similar patterns of expression. Only a subset of genes encoding a given enzyme activity is upregulated. Moreover, the same subset of genes is upregulated in different growth conditions. For example, the orthologs in Clades A, B, E, G and P of GH61 are upregulated under growth in complex substrates for both thermophiles (**Fig. 2b**). An even more striking correlation between transcript levels and orthologs is evident for the GH6 and GH7 cellulases (**Supplementary Fig. 7**) where the transcript profiles for the orthologs of the two organisms are essentially identical. With the exception of the pectinolytic enzymes, the correlation between expression profiles and orthologs extends to many of the lignocellulolytic proteins (**Supplementary Table 7**).

Secretomes and exo-proteomes

In addition to extracellular CAZymes involved in digestion of polysaccharide nutrients, the genomes of *M. thermophila* and *T. terrestris* encode an assortment of hydrolytic and oxidative enzymes that may enhance their ability to forage noncarbohydrate substrates. Collectively, secreted proteins (the secretome) can also provide important information regarding cellular physiology and metabolism in both natural and industrial bioprocessing conditions. The secretomes of *M. thermophila* and *T. terrestris* are predicted to comprise 683 and 789 proteins, respectively (**Supplementary Tables 8** and **9**), of which 569 are homologs. The predicted extracellular proteins (the secretome) include about 180 CAZymes, 40 peptidases, >65 oxidoreductases and >230 proteins of unknown function in each species. Bioinformatic prediction tends to overestimate the number of secreted proteins because the features of some intracellular proteins, in particular those residing in the endoplasmic reticulum, are indistinguishable from secreted proteins. We therefore used mass spectrometry to identify

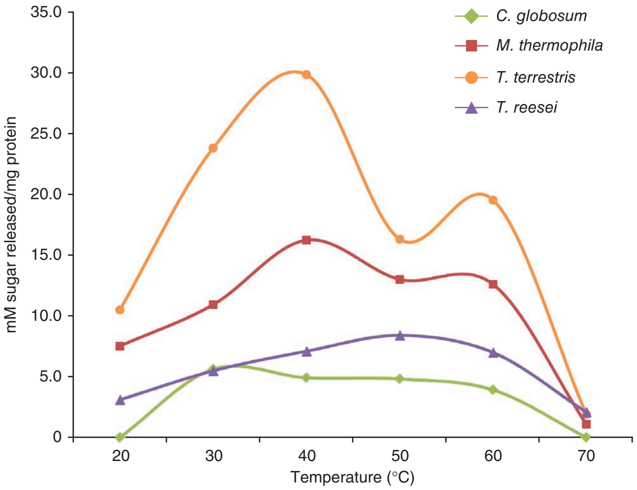

Figure 3 Release of reducing sugars from alfalfa straw by crude extracellular enzymes from thermophilic and nonthermophilic fungi. The crude extracellular enzymes from the fungi cultured on alfalfa straw were used for hydrolysis. The hydrolysis reactions were performed at the temperatures indicated. The final protein concentrations in the reaction mixtures were: *Chaetomium globosum*, 439 µg/ml; *Myceliophthora thermophila*, 422 µg/ml; *Thielavia terrestris*, 362 µg/ml; and *Trichoderma reesei*, 524 µg/ml.

secreted proteins involved in biomass degradation. In general, the genes encoding the identified proteins are expressed at levels higher than those of their paralogs that are not detected extracellularly, especially during growth on agricultural straws (**Supplementary Tables 8** and **9**). Based on transcriptome analysis, the few peptidases detected in the exo-proteomes do not display differential regulation, implying that peptidases are not critical components in biomass degradation. Notably, the genes encoding some hypothetical proteins and oxidoreductases that are detected in the exo-proteomes, and which possess predicted signal peptides, are upregulated when these fungi are cultured on agricultural straws as compared to glucose; for example, Mycth_59005, Mycth_2298860, Mycth_2303335 and Thite_2106069. The role of these secreted proteins in lignocellulose degradation is currently being investigated.

Hydrolysis of polysaccharides in alfalfa straw

Thermophilic fungi are major components of the microflora in self-heating composts. They break down cellulose at a faster rate than prodigious, mesophilic cellulase producers such as *T. reesei* at 40–50 °C (ref. 3). Plant biomass–degrading enzymes characterized from thermophilic fungi have temperature optima between 55 °C and 70 °C (refs. 13–15). Consequently, enzymes from thermophiles are expected to break down plant biomass at a faster rate than enzymes from mesophiles at elevated temperatures. We examined the temperature effects on the hydrolysis of alfalfa straw using enzymes from *M. thermophila*, *T. terrestris*, *C. globosum* and *T. reesei* (**Fig. 3**). The optimum temperature of hydrolysis for enzymes from *T. reesei* occurs at 50 °C. The enzyme mixture from *C. globosum* displays a broad temperature optimum from 30–60 °C. Enzymes from the thermophiles release appreciably higher amounts of reducing sugars than do the enzymes from the mesophiles, with peaks at 40 °C and 60 °C. The biphasic temperature profile of cell wall decomposition suggests that the proteins secreted by the thermophiles contain multiple biomass enzymes, some of which have temperature optima at 60 °C, whereas others have optima around 40 °C.

To determine whether biomass-degrading enzymes from these thermophiles possess distinct temperature optima, we successfully cloned and expressed in *Aspergillus niger* the genes encoding seven xylanases from the thermophiles and tested their biochemical properties. The temperature optima for these xylanases range from 45 °C to 70 °C (**Table 2**). These results suggest that *M. thermophila* and *T. terrestris* not only possess diverse enzyme activities for degradation of plant cell wall polysaccharides, they have also evolved diverse properties for these enzymes that enable efficient hydrolysis over a range of temperatures. The diversity of enzyme activities and properties may help to explain the ubiquity of these organisms in decomposing biomass.

Potential utility of sexual cycle for strain development

The ability to do sexual crossing is rare among fungal cells used for bioproduction. Crossing can facilitate strain improvement stemming from recombinant DNA methods, classical mutagenesis, genome shuffling or natural variation. In this context, we evaluated genes involved in mating and the potential for outcrossing, particularly in *M. thermophila*.

Nearly all thermophilic Chaetomiaceae are either homothallic (self-fertile), as is the case for *C. globosum* and *T. terrestris*, or have been observed only in the asexual state, as is the case for *M. thermophila*. We examined these genomes for homologs of the mating-type genes of *Neurospora crassa*. Each genome possesses a homolog of *N. crassa matA-1* (CHGT_06585, Mycth_2298236 and Thite_2111503), the primary gene responsible for mating-type determination in *mat A* strains. As in *N. crassa*, *matA-2* homologs in the Chaetomiaceae species (CHGT_06585, Mycth_2107654 and Thite_2127265) are immediately adjacent to *matA-1* genes, and presumptive *matA-3* homologs are adjacent to *matA-2* genes (CHGT_06584, 1 Mycth_2115740 and Thite_2111506). We could not identify homologs for *mat a* genes.

The *M. thermophila* genome allowed us to confirm the close relationship between this species and *Myceliophthora heterothallica* (*Thielavia heterothallica*)[16], reported to be heterothallic[17] and for which *mat A* and *mat a* strains have been identified at the molecular level. Using molecular markers based on *M. thermophila* sequences, we confirmed heterothallism and marker segregation in isolates of *M. heterothallica* from diverse locations (to date: New Mexico, Indiana and Germany). Attempts to cross the sequenced strain of *M. thermophila* were not successful. The ability to cross *M. thermophila* would permit optimization of gene combinations with natural and engineered gene variants, as well as development of a complete model organism from this industrially important group. Evidence of repeat induced polymorphism in the sequenced species (**Supplementary Notes**) could represent an obstacle to be overcome to exploit crosses between strains with multicopy transgenes.

Distinguishing thermophilic from mesophilic fungi

Compared to the closely related mesophile *C. globosum*, the genomes of *M. thermophila* and *T. terrestris* contain larger fractions of repetitive

Table 2 Temperature optima for *M. thermophila* (Mycth) and *T. terrestris* (Thite) xylanases

Gene ID	Family	Optima (°C)
Mycth_100068	GH11	50
Mycth_2121801	GH11	60
Mycth_112050	GH10	60
Mycth_116553	GH10	70
Thite_2117649	GH10	60
Thite_2042100	GH11	50
Thite_2107799	GH11	45

sequences that have low GC content, introducing significant GC variation (**Supplementary Fig. 8**). When comparing the GC content of *M. thermophila* and *T. terrestris* with *C. globosum* and other species within the class Sordariomycetes, it appears that although the genomes of the thermophilic species have a slightly lower genome-level GC content than *C. globosum*, they have a higher GC content in coding regions, which is reflected in the third position of codons (**Supplementary Table 10**). Since G:C pairs are more thermally stable, this may suggest the potential adaptability of protein-coding genes to high temperatures. Approximately 75% of *M. thermophila* codons have a higher GC content at the third nucleotide position (GC3) compared to the corresponding *C. globosum* orthologs. The percentage is even higher when comparing *T. terrestris* with *C. globosum*; 92% of *T. terrestris* codons have a higher GC3. This is in contrast to thermophilic prokaryotes, where analysis of large numbers of sequenced thermophiles and hyperthermophiles did not reveal a correlation of higher GC (and GC3) with thermal adaptability[18]. It remains to be seen whether high GC3 content will hold true for other thermophilic eukaryotes.

Analysis of prokaryotic thermophiles also included multiple attempts to define amino acid 'signatures' of thermophilic adaptations. A seven amino-acid motif IVYWREL has been reported that positively correlates with elevated growth temperatures in 204 complete proteomes of archaea and bacteria[19]. Two groups observed that thermophilic proteins are enriched in glutamine, arginine and lysine amino acid residues and contain lesser amounts of alanine, aspartic acid, asparagine, glutamine, threonine and serine[20,21], and another group found three notable substitutions (lysine to arginine, serine to alanine and serine to threonine) by comparing thermophilic *Corynebacterium efficiens* and mesophilic *C. glutamicum*, and suggested the differences are probably important for thermostability of *C. efficiens* proteins[22]. We searched for these amino acid signatures in filamentous fungi and found that they do not distinguish fungal thermophiles from their mesophilic relatives (**Supplementary Tables 11–14**). On the basis of our comparative analyses of the genomes from two thermophilic fungi, we conclude that their nucleotide and protein features are different from those observed in thermophilic prokaryotes.

We also investigated the possibility that thermophilic fungi possess major differences in processes mediating thermophily including heat shock, oxidative stress, membrane biosynthesis, chromatin structure and modification, and fungal cell wall metabolism. We compared the proteins predicted to be involved in these processes in *C. globosum*, *M. thermophila* and *T. terrestris*, but were unable to find differences that can convincingly be interpreted as the molecular bases that underpin fungal thermophily (**Supplementary Notes** and **Supplementary Tables 15–25**).

Phylogeny and the origins of thermophily among Ascomycota

Thermophilic fungi, defined as fungi that grow better above 45 °C than at 25 °C, have evolved independently in at least two lineages within the phylum Ascomycota, once each within the orders Sordariales and Eurotiales (**Supplementary Fig. 9**). Within the Sordiariales, thermophily is restricted to subgroups of the family Chaetomiaceae. Among fungi more broadly, thermophily also exists in the Zygomycota, but it appears to be rare or absent in the phyla Basidiomycota and Chytridiomycota. The evolutionary trajectory of thermophily is obscured by chaotic taxonomy[23]. Biosystematic efforts have lagged behind research on thermophiles in industry, resulting in a body of literature for these organisms that lacks accurate taxonomic treatments. Aside from the potential for creating disputes and confusion in the commercial realm, the state of taxonomic study hinders efforts to determine the number of sublineages in which thermophily has been gained and lost in the groups where it is common. The fungi that are the subject of this paper provide excellent examples. *M. thermophila* has been placed alternately in the anamorphic genera *Sporotrichum* and *Chrysosporium*, both of which are inappropriate for fungi in the family Chaetomiaceae. Recognition of this organism as a member of the Chaetomiaceae therefore removes the need to assume a separate origin for thermophily in a third order of filamentous ascomycetes. True thermophiles have been assigned to the genera *Chaetomium* and *Thielavia*, along with mesophilic and thermotolerant species. If current taxonomic conclusions are correct, this would imply multiple independent gains or losses of thermophily within the family Chaetomiaceae. Alternatively, it is possible that current classification does not reflect phylogenetic relationships.

DISCUSSION

Thermophilic fungi are ubiquitous organisms commonly found in decomposing organic matter. The biotechnological utility of these fungi has been recognized for many years. The finished genomes for the two thermophiles may serve not only as reference genomes for studies on genome evolution and structure, but might also support targeting and mapping of changes that could facilitate strain construction and improvement. Selection from natural population variability or from mutagenized strains remains a useful tool for strain development. With a finished genome as a scaffold and modern sequencing technologies, resequencing these strains and identifying mutations becomes relatively simple but very helpful for identifying beneficial and deleterious genetic changes.

Thermostability alone does not explain why the enzymes from the thermophilic fungi display higher hydrolytic power than enzymes produced by mesophiles. Crude extracellular enzymes from the thermophiles exhibit higher hydrolytic capacity than their counterparts from mesophiles at temperatures ranging from 30 °C to 60 °C (**Fig. 3**). One explanation is that the enzymes from the thermophiles possess higher specific activity toward lignocellulosic biomass. It is also possible that the thermophiles secrete a broader spectrum of accessory proteins that accelerate biomass degradation. In addition to the GH61 proteins[10], the poorly characterized extracellular proteins that are upregulated when the fungi are cultured on straws (**Supplementary Tables 8** and **9**) may represent new lignocellulose-active proteins.

When expressed in a mesophilic heterologous fungal host, the enzymes from thermophilic fungi usually retain their thermostable character[24,25]. This provides an excellent opportunity to replace specific enzyme components in mesophilic production organisms with better-performing counterparts from thermophiles[10]. Additionally, prospects for further enhancements in the industrial performance of these enzymes through protein engineering appear likely[26].

An intriguing alternative to replacement of individual enzyme components in mesophiles with thermostable orthologs may involve development of thermophilic fungal production hosts. In this regard, such hosts that provide better hyphal morphology in tank fermentations have been shown to result in reduced viscosity and improved productivity, and they can be engineered by using protoplast transformation to introduce recombinant DNA constructs[5,6]. It is also noteworthy that the well-developed industrial production organism previously known as *Chrysosporium lucknowense* C1 was recently reclassified as an isolate of *M. thermophila*[27]. When combined with the prospect of a functional sexual cycle and a finished genome, both of which enable rational design of improved strains, these observations may provide an adequate toolbox for development of efficient thermophilic fungal host strains.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology/.

Accession codes. Assembly and annotation data for *M. thermophila* and *T. terrestris* are available through JGI MycoCosm Genome Portal at http://jgi.doe.gov/fungi and at DDBJ/EMBL/GenBank under chromosome accession numbers CP003002-CP003008 and CP003009-CP003014, respectively. The transcriptome data are available under GEO accession number GSE27323.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

The genome sequencing and analysis were conducted by the US Department of Energy Joint Genome Institute and supported by the Office of Science of the US Department of Energy under contract no. DE-AC02-05CH11231. The work on transcriptomes, enzyme characterization and the *Myceliophthora* exo-proteome was supported by the Cellulosic Biofuel Network of the Agriculture Bioproducts Innovation Program of Agriculture and Agri-Food Canada, Genome Canada and Genome Québec.

AUTHOR CONTRIBUTIONS

The final text of the manuscript was written by R.M.B. and A.T., and reviewed by I.V.G.; who together also coordinated the overall analysis. I.V.G. coordinated both genome projects at the Joint Genome Institute. R.M.B. prepared the genomic DNA of *T. terrestris* and T.J. the DNA of *M. thermophila*. A.T. coordinated the transcriptome and exo-proteome work, and analyzed the transcriptomes. S.L. and E.L. led genome and cDNA sequencing. J.G. and J.S. finished and assembled both genomes. R.O. and A.S. annotated and analyzed the genomes, synteny and GC content. I.R. processed the RNA-Seq data and analyzed the cell wall proteins. N.I. coordinated the sample preparation for transcriptome analysis and analyzed the lignocellulolytic proteins. B.H., P.M.C. and V.L. performed the comparative analysis of the carbohydrate-active proteins. C.D. conducted the enzymatic hydrolysis of straws and M.-C.M. prepared the samples for transcriptome and exo-proteome analysis. D.O.N. analyzed the mating types and phylogeny of thermophilic fungi. E.L. coordinated the cDNA synthesis and EST analysis. A.B. coordinated the cloning and expression of xylanase genes. D.T. characterized the biochemical properties of the xylanases. R.P. de V., I.E.A, and J. van den B. examined the growth on different substrates. P.H. analyzed the GH61 proteins and J.P. membrane biogenesis. G.B. analyzed the secretomes. S.U. and R.S. analyzed the chromatin structure and dynamics. A.J.P. examined melanin pigment biogenesis. I.T.P. and L.D.H.E. analyzed transporters. S.E.B. analyzed secondary metabolism. J.M. examined oxidative stress. M.W. reviewed proteases and peptidases. S.L. examined the exo-proteomes. A.J.C. looked for repeat-induced polymorphisms. D.M. contributed computational tools for viewing *T. terrestris* transcriptome data. A.L.de L. and M.W.R. examined oxidoreductases and chitinases, respectively.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/nbt/index.html.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Margaritis, A. & Merchant, R.F.J. Thermostable cellulases from thermophilic microorganisms. *Crit. Rev. Biotechnol.* **4**, 327–367 (1986).
2. Margaritis, A. & Merchant, R. Production and thermal stability characteristics of cellulase and xylanase enzymes from *Thielavia terrestris. Biotechnol. Bioeng. Symp.* **13**, 426–428 (1983).
3. Tansey, M.R. Agar-diffusion assay of cellulolytic ability of thermophilic fungi. *Arch. Mikrobiol.* **77**, 1–11 (1971).
4. Wojtczak, G., Breuil, C., Yamada, J. & Saddler, J.N. A comparison of the thermostability of cellulases from various thermophilic fungi. *Appl. Microbiol. Biotechnol.* **17**, 82–87 (1987).
5. Jensen, E.B. & Boominathan, K.C. Thermophilic fungal expression system. US Patent 5,695,985 (1997).
6. Jensen, E.B. & Karuppan, C.B. Thermophilic fungal expression system. US Patent 5,602,004 (1997).
7. *Chaetomium globosum* Genome Database (Broad Institute, 2005). <http://www.broadinstitute.org/annotation/genome/chaetomium_globosum>.
8. Henrissat, B. & Davies, G. Structural and sequence-based classification of glycoside hydrolases. *Curr. Opin. Struct. Biol.* **7**, 637–644 (1997).
9. Karlsson, J. *et al.* Homologous expression and characterization of Cel61A (EG IV) of Trichoderma reesei. *Eur. J. Biochem.* **268**, 6498–6507 (2001).
10. Harris, P.V. *et al.* Stimulation of lignocellulosic biomass hydrolysis by proteins of glycoside hydrolase family 61: structure and function of a large, enigmatic family. *Biochemistry* **49**, 3305–3316 (2010).
11. Pahkala, K. *et al.* Production of bioethanol from barley straw and reed canary grass: a raw material study. *15th European Biomass Conference and Exhibition*. Berlin, Germany, May 7–11, 2007 (ETA, Florence, Italy and WIP, Munich, 2007).
12. Dien, B.S. *et al.* Chemical composition and response to dilute-acid pretreatment and enzymatic saccharification of alfalfa, reed canarygrass, and switchgrass. *Biomass Bioenergy* **30**, 880–891 (2006).
13. Kaur, G., Kumar, S. & Satyanarayana, T. Production, characterization and application of a thermostable polygalacturonase of a thermophilic mould Sporotrichum thermophile Apinis. *Bioresour. Technol.* **94**, 239–243 (2004).
14. Vafiadi, C., Topakas, E., Biely, P. & Christakopoulos, P. Purification, characterization and mass spectrometric sequencing of a thermophilic glucuronoyl esterase from Sporotrichum thermophile. *FEMS Microbiol. Lett.* **296**, 178–184 (2009).
15. Roy, S.K., Dey, S.K., Raha, S.K. & Chakrabarty, S.L. Purification and properties of an extracellular endoglucanase from Myceliophthora thermophila D-14 (ATCC 48104). *J. Gen. Microbiol.* **136**, 1967–1971 (1990).
16. van den Brink, J., Samson, R.A., Hagen, F., Boekhout, T. & de Vries, R.P. Phylogeny of the industrial relevant, thermophilic genera *Myceliophthora* and *Corynascus. Fungal Divers.* published online, doi:10.1007/s13225-13011-10107-z (28 May 2011).
17. von Klopotek, A. Thielavia heterothallica spec. nov., die perfekte Form von Chrysosporium thermophilum. *Arch. Microbiol.* **107**, 223–224 (1976).
18. Galtier, N. & Lobry, J.R. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *J. Mol. Evol.* **44**, 632–636 (1997).
19. Zeldovich, K.B., Berezovsky, I.N. & Shakhnovich, E.I. Protein and DNA sequence determinants of thermophilic adaptation. *PLoS Comput. Biol.* **3**, e5 (2007).
20. Glyakina, A.V., Garbuzynskiy, S.O., Lobanov, M.Y. & Galzitskaya, O.V. Different packing of external residues can explain differences in the thermostability of proteins from thermophilic and mesophilic organisms. *Bioinformatics* **23**, 2231–2238 (2007).
21. Wang, G.-Z. & Lercher, M.J. Amino acid composition in endothermic vertebrates is biased in the same direction as in thermophilic prokaryotes. *BMC Evol. Biol.* **10**, 263 (2010).
22. Nishio, Y. *et al.* Comparative complete genome sequence analysis of the amino acid replacements responsible for the thermostability of Corynebacterium efficiens. *Genome Res.* **13**, 1572–1579 (2003).
23. Mouchacca, J. Heat-tolerant fungi and applied research work: a synopsis of name changes and synonymies. *World J. Microbiol. Biotechnol.* **16**, 881–888 (2000).
24. Berka, R.M., Rey, M.W., Brown, K.M., Byun, T. & Klotz, A.V. Molecular characterization and expression of a phytase gene from the thermophilic fungus *Thermomyces lanuginosus. Appl. Environ. Microbiol.* **64**, 4423–4427 (1998).
25. Murray, P. *et al.* Expression in and characterisation of a thermostable family 3 β-glucosidase from the moderately thermophilic fungus *Talaromyces emersonii. Protein Expr. Purif.* **38**, 248–257 (2004).
26. Voutilainen, S.P., Murray, P.G., Tuohy, M.G. & Koivula, A. Expression of *Talaromyces emersonii* cellobiohydrolase Cel7A in *Saccharomyces cerevisiae* and rational mutagenesis to improve its thermostability and activity. *Protein Eng. Des. Sel.* **23**, 69–79 (2010).
27. Visser, H. *et al.* Development of a mature fungal technology and production platform for industrial enzymes based on a *Myceliophthora thermophila* isolate, previously known as *Chrysosporium lucknowense* C1. *Ind. Biotechnol.* **7**, 214–223 (2011).

ONLINE METHODS

Additional details on the methods are provided in the **Supplementary Methods**.

Genome sequencing and assembly. Genomic DNA extracted from *Myceliophthora thermophila* (*Sporotrichum thermophile*) strain ATCC 42464 and *Thielavia terrestris* strain NRRL 8126 were used for whole genome shotgun sequencing. All sequencing reads were collected with standard Sanger sequencing protocols on ABI 3730XL capillary sequencing machines and assembled with ARACHNE[28]. Low-quality regions and gaps were computationally selected and sequenced with custom primers. After the completion of the automated rounds, we selected further reactions manually to finish the genomes. We resolved smaller repeats by transposon hopping with 8 kb plasmid clones. We shotgun sequenced and finished fosmid clones to fill large gaps, resolve larger repeats or to resolve chromosome duplications and extend into telomere regions. The finished genomes of *M. thermophila* and *T. terrestris* consist of seven and six chromosomes, respectively, comprising 38,744,216 bp and 36,912,256 bp of finished sequence with an estimated error rate of less than one error in 100,000 base pairs (**Table 1**).

Construction of cDNA libraries and analysis of expressed sequence tags (ESTs). Poly A^+ RNA was isolated from total RNA (pooled RNA from cells grown in rich medium for *T. terrestris* and 1% cellulose and 1% pectin pooled culture from *M. thermophila*) using the Absolutely mRNA Purification Kit and manufacturer's instructions (Stratagene). Synthesis and cloning of cDNA was a modified procedure based on the SuperScript plasmid system with Gateway technology for cDNA synthesis and cloning (Invitrogen). Plasmid DNA for sequencing was produced by rolling circle amplification[29] (TempliPhi, GE Healthcare). Subcloned inserts were sequenced from both ends. A total of 33,559 *T. terrestris* ESTs including 5,548 from external sources and 44,939 *M. thermophila* ESTs including 11,392 from external sources were processed through the JGI EST pipeline separately for each genome.

Genome annotation. Chromosome sequences were masked using RepeatMasker[30] and the RepBase library of 234 fungal transposable elements[31]. Gene modeling on the masked assembly was performed using *ab initio* Fgenesh[32] and Genemark-ES[33]; homology-based Fgenesh+[32] and Genewise[34] seeded by BLASTx alignments of NCBI's nr (nonredundant) protein database against the assembly; cDNA-based EST_map (http://www.softberry.com/) seeded by EST contigs. EST BLAT alignments[35] were used to add or extend exons for gene models. Because multiple gene models were generated for each locus, a single representative model was algorithmically chosen based on model quality. Genes for tRNAs were predicted using tRNAscan-s.e.m[36]. The term gene model refers to protein-coding genes unless otherwise noted. The *C. globosum* genome assembly and gene models, used for comparison to the two thermophiles, were downloaded from the Broad Institute *Chaetomium globosum* Database at http://www.broadinstitute.org/annotation/genome/chaetomium_globosum. All predicted gene models were functionally annotated using SignalP[37], TMHMM[38], InterProScan[39], Blastp[40] against the nr database, and hardware-accelerated double-affine Smith-Waterman alignments (deCypherSW; http://www.timelogic.com/decypher_sw.html) against SwissProt, the Kyoto Encyclopedia of Genes and Genomes (KEGG)[41], and the Eukaryotic Orthologous Groups of proteins database (KOG)[42]. The Enzyme Commission (EC) numbers (http://www.expasy.org/enzyme/) of KEGG hits were assigned to gene models and mapped to corresponding KEGG pathways. InterPro and SwissProt hits were used to assign Genome Ontology (GO) terms[43] to gene models. Multigene families were predicted with the Markov clustering algorithm (MCL)[44], using Blastp alignment scores between proteins as a similarity metric. Manual curation of automated annotations was performed using the JGI MycoCosm Genome Portal (http://jgi.doe.gov/fungi).

Transcriptome analysis. The thermophiles were cultured at 45 °C with shaking at 150 r.p.m. in 10× TDM[45] containing 2% glucose or 2% agricultural straws (alfalfa or barley straws ground to 0.5 mm lengths). Mycelia were harvested at early growth phase; 21 h for *M. thermophila* and 24–28 h for *T. terrestris*. Total RNA samples were isolated from mycelia as described[46]. Sequencing was performed using the RNA-Seq method of Illumina's Solexa IG at either the DNA Core Facility of the University of Missouri or at the McGill University-Génome Québec Innovation Centre. The RNA-Seq reads, 38–42 nt in length, from each mRNA sample were mapped against a combination of the genomic sequence and the spliced sequences with Bowtie[47] using the 'best' strata option. The depth of mapped read coverage at each genome position was calculated using the WIGGLES program bundled with TopHat[48]. To analyze differential expression, we took the transcript and exon definitions from the filtered models of version 2.0 of genome annotation. The number of reads mapping to each transcript was estimated by integrating the coverage depth over the annotated exons of the transcript, dividing by the read length, and rounding to an integer. The Bayesian posterior probability of differential expression was estimated from the read counts for each transcript using the R package baySeq v 1.4 (ref. 49). To aid interpretation, we also calculated FPKM (fragments per kilobase of transcript per million mapped reads) values from the counts using the transcript lengths and the total number of mapped reads from each sample.

Secretome and exo-proteome analysis. Sequence-based prediction of extracellular proteins (secretome) was processed as described[50]. For exo-proteome analysis, proteins secreted by *M. thermophila* after 30 h of growth in barley and alfalfa straws were resolved on one-dimensional SDS-PAGE, fractionated into 12 bands, and in-gel digested with trypsin. Proteins secreted by *T. terrestris* after 96 h of growth in cellulose and xylose were resolved by two-dimensional gel electrophoresis. After staining with Coomassie Blue, over 100 spots from each gel were excised and in-gel digested with trypsin. Peptides eluted from the gel fragments were analyzed by tandem mass spectrometry.

Hydrolysis of polysaccharide biomass. To obtain extracellular proteins, we grew the fungi in 10× TDM containing 2% alfalfa straw (0.5 mm length), at 45 °C for *M. thermophila* and *T. terrestris*, 34 °C for *C. globosum* and 24 °C for *T. reesei*. The cultures were harvested when the peak cellulase activity was reached: 30 h for *M. thermophila*, 40 h for *T. terrestris* and *C. globosum*, and 70 h for *T. reesei*. The culture filtrates containing extracellular proteins were obtained by removing the mycelia and residual substrates by filtering through Miracloth (Calbiochem), and followed by centrifugation at 10,000*g* for 30 min. Cleared supernatant fluids were concentrated using the Vivaflow200 (Sartorius Stedim) with a polyethersulfone membrane and 10 kDa cutoff. One (1) ml of supernatant was aliquoted into individual tubes containing 2% alfalfa (wt/vol). Reaction mixtures were incubated for 4 h at 20 °C, 30 °C, 40 °C, 50 °C, 60 °C and 70 °C. Reducing sugar was estimated by the 2,2′-bicinchoninate (BCA) method. Protein concentration was determined by the Bradford method. The activity was calculated as mM of reducing sugar released per mg of protein.

Cloning and expression of xylanase genes. Genes of *M. thermophila* and *T. terrestris* predicted to encode endo-1,4-β-xylanase were cloned and expressed in *Aspergillus niger*. Briefly, oligonucleotides were designed to be complementary to the start and stop regions of the target genes, then used to PCR amplify cDNA prepared from the two thermophiles. The amplified cDNAs were cloned into the *A. niger* expression vector using the Gateway recombination method (Invitrogen) and used to transform *A. niger*.

Enzyme activity assays. Cellulase and xylanase activities were determined by measuring the reducing sugar released from the substrates using the BCA reagent in 96-well microplate format. Carboxymethyl cellulose (CMC-4M) and birchwood xylan, both obtained from Sigma-Aldrich, were used to determine the activity of cellulase and xylanase, respectively. The assays were performed in 50 mM Britton-Robinson buffer (50 mM boric acid, 50 mM acetic acid and 50 mM phosphoric acid) at the indicated temperatures for 30 min. Following incubation, 10 µl of the reaction mixture was added to 190 µl of BCA reagent, incubated at 80 °C for 40 min for color development, and the absorbance of the resultant mixtures was read at 562 nm. Glucose and xylose were used to prepare standard curves.

28. Jaffe, D.B. Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13**, 91–96 (2003).
29. Detter, J.C. *et al.* Isothermal strand-displacement amplification applications for high-throughput genomics. *Genomics* **80**, 691–698 (2002).
30. Smit, A.F.A., Hubley, R. & Green, P. RepeatMasker Open–3.0. 1996–2010. <http://www.repeatmasker.org/> (2010).

31. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
32. Salamov, A.A. Ab initio gene finding in *Drosophila* Genomic DNA. *Genome Res.* **10**, 516–522 (2000).
33. Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y.O. & Borodovsky, M. Gene prediction in novel fungal genomes using an *ab initio* algorithm with unsupervised training. *Genome Res.* **18**, 1979–1990 (2008).
34. Birney, E. Using GeneWise in the *Drosophila* annotation experiment. *Genome Res.* **10**, 547–548 (2000).
35. Kent, W.J. BLAT—The BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
36. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
37. Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Int. J. Neural Syst.* **8**, 581–599 (1997).
38. Melén, K., Krogh, A. & von Heijne, G. Reliability measures for membrane protein topology prediction algorithms. *J. Mol. Biol.* **327**, 735–744 (2003).
39. Zdobnov, E.M. & Apweiler, R. InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847–848 (2001).
40. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
41. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32**, D277–D280 (2004).
42. Koonin, E.V. *et al.* A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol.* **5**, R7 (2004).
43. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**, D258–D261 (2004).
44. Enright, A.J., Van Dongen, S. & Ouzounis, C.A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).
45. Roy, B.P. & Archibald, F. Effects of kraft pulp and lignin on Trametes versicolor carbon metabolism. *Appl. Environ. Microbiol.* **59**, 1855–1863 (1993).
46. Semova, N. *et al.* Generation, annotation, and analysis of an extensive *Aspergillus niger* EST collection. *BMC Microbiol.* **6**, 7 (2006).
47. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
48. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
49. Hardcastle, T.J. & Kelly, K.A. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* **11**, 422 (2010).
50. Tsang, A., Butler, G., Powlowski, J., Panisko, E. & Baker, S. Analytical and computational approaches to define the *Aspergillus niger* secretome. *Fungal Genet. Biol.* **46**, S153–S160 (2009).

doi:10.1038/nbt.1976