AEM
Journals.ASM.org

Short-Read Sequencing for Genomic Analysis of the Brown Rot Fungus *Fibroporia radiculosa*

Juliet D. Tang,[a] Andy D. Perkins,[b] Tad S. Sonstegard,[c] Steven G. Schroeder,[c] Shane C. Burgess,[d] and Susan V. Diehl[a]

Forest Products, Mississippi State University, Mississippi State, Mississippi, USA[a]; Computer Science and Engineering, Mississippi State University, Mississippi State, Mississippi, USA[b]; USDA ARS Bovine Functional Genomics Laboratory, Beltsville, Maryland, USA[c]; and Institute for Genomics, Biocomputing, and Biotechnology, Mississippi State University, Mississippi State, Mississippi, USA[d]

The feasibility of short-read sequencing for genomic analysis was demonstrated for *Fibroporia radiculosa*, a copper-tolerant fungus that causes brown rot decay of wood. The effect of read quality on genomic assembly was assessed by filtering Illumina GAIIx reads from a single run of a paired-end library (75-nucleotide read length and 300-bp fragment size) at three different stringency levels and then assembling each data set with Velvet. A simple approach was devised to determine which filter stringency was "best." Venn diagrams identified the regions containing reads that were used in an assembly but were of a low-enough quality to be removed by a filter. By plotting base quality histograms of reads in this region, we judged whether a filter was too stringent or not stringent enough. Our best assembly had a genome size of 33.6 Mb, an N50 of 65.8 kb for a *k*-mer of 51, and a maximum contig length of 347 kb. Using GeneMark, 9,262 genes were predicted. TargetP and SignalP analyses showed that among the 1,213 genes with secreted products, 986 had motifs for signal peptides and 227 had motifs for signal anchors. Blast2GO analysis provided functional annotation for 5,407 genes. We identified 29 genes with putative roles in copper tolerance and 73 genes for lignocellulose degradation. A search for homologs of these 102 genes showed that *F. radiculosa* exhibited more similarity to *Postia placenta* than *Serpula lacrymans*. Notable differences were found, however, and their involvements in copper tolerance and wood decay are discussed.

Recent technological achievements in massively parallel sequencing have escalated the rate at which genomes can be sequenced. The most affordable method is the Illumina genome analyzer, which uses reversible dideoxy terminator sequencing chemistry (3). Although the read length is significantly shorter than that with traditional Sanger dideoxy terminator sequencing, Illumina sequencing generates millions of reads, thereby providing the extensive coverage needed to produce an assembly in spite of the higher frequency of base errors and poor read quality. In addition, the read length has increased with technology upgrades, making the technique more attractive for the sequencing of the larger, more complex genomes of eukaryotes.

Now that the race to sequence genomes has begun, we are gaining a much clearer picture of how Illumina sequencing performs in practice. Three eukaryotic genomes have been sequenced to date with the Illumina technology alone or in combination with other platforms. These genomes are those of the giant panda bear (29), the plant-pathogenic ascomycete fungus *Grosmannia clavigera* (10), and another ascomycete, *Sordaria macrospora*, a model organism for fungal morphogenesis (33). The assembly for the giant panda bear was a formidable feat that most ordinary researchers could not afford to duplicate. It used paired-end reads (52-nucleotide [nt] read length) from 37 Illumina libraries (150-kb to 10-kb insert size) to obtain a 56-fold coverage of the genome (29). The draft assembly encompassed 94% of the 2.4-Gb panda genome, predicted 21,000 genes, and led to the discovery of 2.7 million single-nucleotide polymorphisms (SNPs).

On a more modest scale, the *G. clavigera* fungal sequencing project showed that paired-end Illumina reads (42 nt) could be effectively combined with Sanger and Roche platforms to produce a genomic assembly that was improved significantly based on a comparison of N50 with assemblies that lacked the Illumina reads (10). N50 is a measure of assembly quality, since by definition, half the assembly is covered by contigs of a size of N50 or larger. The core assembly was created from paired-end Illumina reads (42-nt read length and 200-bp insert library). Scaffolds were then built by using overlaps with Sanger paired-end reads (average read length of 600 nt from a 40-kb fosmid library) and Roche single-end reads (average read lengths of 100 and 225 nt). Reads were trimmed to remove low-quality sections, and the extent of trimming required was evaluated by counting the number of misassemblies relative to a reference sequence. The draft *G. clavigera* genome was 32.5 Mb and had an N50 of 32 kb, a scaffold N50 of 558 kb, and a total of 162 gaps. The genome was later manually finished with more sequence data, validated with expressed-sequence-tag (EST) data, and then used to predict genes and obtain functional annotations for a transcriptomic analysis (11).

The *S. macrospora* assembly was another notable example, because it used only the Illumina and Roche next-generation sequencing platforms. By varying the number of reads used in the assembly from each platform, the authors of that study showed that the Illumina paired-end reads provided gains in N50 and maximum contig lengths, while the Roche reads dramatically decreased the number and length of gaps (33). The Illumina reads (36 nt) were primarily paired-end reads from two insert libraries

Received 30 August 2011 **Accepted** 9 January 2012
Published ahead of print 13 January 2012
Address correspondence to Juliet D. Tang, jdt57@msstate.edu.
The manuscript has been approved for publication as journal article FP619 of the Forest and Wildlife Research Center, Mississippi State University.
Supplemental material for this article may be found at http://aem.asm.org/.
Copyright © 2012, American Society for Microbiology. All Rights Reserved.
doi:10.1128/AEM.06745-11

(300 bp and 500 bp), providing 85-fold coverage, while the Roche single-end reads (367-nt average read length) were sequenced to 10-fold coverage. The draft genome was estimated to be about 40 Mb. The Illumina assembly had an N50 of 51 kb and 17,956 gaps. By adding the Roche reads, the N50 increased to 117 kb, and the number of gaps dropped to 624. Known syntenic regions between *S. macrospora* and *Neurospora crassa* were then used to produce a scaffold N50 of 498 kb.

Given these successes, we were highly motivated to sequence the genome of the basidiomycete *Fibroporia radiculosa* (*Antrodia radiculosa*). *F. radiculosa* is a brown rot fungus that has been documented to cause a premature failure of wooden stakes treated with copper-based wood preservatives in the field (9). Species of brown rot fungi that show copper tolerance in laboratory tests include *F. radiculosa*, *Postia placenta*, and *Fomitopsis palustris* (19, 20), and all three can secrete high levels of oxalate (7, 20). Oxalate is believed to confer metal tolerance to fungi because it chelates metal to form insoluble metal oxalate crystals (8, 16, 24). A related species, *Serpula lacrymans*, causes a specialized form of brown rot decay called dry rot. It is less copper tolerant and secretes lower levels of oxalate (20, 22). Because the majority of wood preservatives are copper based, an understanding of the mechanisms of copper tolerance has become a priority for research in wood protection. Brown rot fungi are also aggressive decomposers of wood. Their particular mode of attack, however, is selective. They work around the lignin, targeting the rapid deconstruction and utilization of hemicellulose and cellulose (17). Since each cellulose molecule in wood is comprised of an average of 10,000 glucose units (36), the biochemical mechanisms involved in brown rot decay may have potential application for biofuel production.

At the time when this project was begun, very little was known about the genes that brown rot fungi employ to overcome copper-based wood preservatives and degrade wood. Since then, the genomes of *P. placenta* (30) and *S. lacrymans* (12) have been sequenced using the more traditional whole-genome shotgun approach. Our goal was to use the more cost-effective Illumina technology, known as paired-end, short-read sequencing (76-nt read length), to predict as many genes as possible from a single library (300 bp). Because of the higher error rates typically encountered in short-read sequencing, the first impasse that we addressed was read filter stringency. We developed a rational approach for read filtering to find an optimal assembly. Genes and genome size were predicted from the contigs (>3 kb) of this assembly, and comparisons to gene sequences in public databases allowed us to determine gene functions. Putative genes related to survival on copper (i.e., the production of copper oxalate crystals and copper homeostasis) and the oxidative and hydrolytic decay of lignocellulose were identified, and their homologous sequences in *P. placenta* (30) and *S. lacrymans* (12) were determined.

MATERIALS AND METHODS

Fungus. *F. radiculosa* strain TFFH 294 was kindly provided by Carol Clausen, USDA Forest Service Forest Products Laboratory, Madison, WI. The identity of the strain was verified by the cloning of the internal transcribed spacer (ITS) region after amplification with ITS1 and ITS4 primers (46). The sequenced DNA aligned to two *F. radiculosa* voucher specimens in the NCBI nucleotide database. For DNA isolation, the fungus was grown for 30 days in potato dextrose broth (125 rpm at 25°C), harvested by filtration, rinsed with 100 mM Tris-HCl and 5 mM EDTA (pH 8.0) and then with 70% ethanol, and stored in 70% ethanol at −80°C.

DNA library preparation. Mycelia (1.38 g wet weight) were mechanically disrupted by grinding the hyphae in liquid nitrogen with a mortar and pestle. Nuclear DNA was extracted by using a method developed previously for cotton (34). Proteins were removed by phenol-chloroform extraction, and the DNA was precipitated with ice-cold isopropanol. After the DNA was resuspended in water, it was treated with RNase A (200 ng/μl) for 1 h at room temperature, followed by another phenol-chloroform extraction. Polysaccharides were removed by adding 0.3 volumes of cold ethanol to the mixture, incubating the mixture on ice for 10 min, and then centrifuging the mixture at 7,000 × *g* for 10 min. The purified DNA from the supernatant was precipitated overnight at −20°C with a 1/10 volume of 3 M sodium acetate (pH 6.0) and 2 volumes of 95% ethanol. Following centrifugation (10,000 × *g* for 15 min), the pelleted DNA was washed with 70% ethanol twice to remove residual salt and then resuspended in TE buffer (10 mM Tris-HCl and 1 mM EDTA [pH 8.0]). The concentration was measured with a Nanodrop 1000 spectrophotometer. The yield of genomic DNA was determined to be 69.8 μg/g wet mycelia.

The genomic library was prepared from 10 μg of genomic DNA according to the protocols provided with the kit (Illumina Genomic DNA Sample Prep kit; Illumina, San Diego, CA). Microfluidic chip electrophoresis on an Agilent 2100 Bioanalyzer (DNA 1000 kit; Agilent, Santa Clara, CA) indicated that nebulization (6 min at 34 lb/in^2) produced a broad peak from 300 to 700 bp. After PCR enrichment, the concentration of the library was 13 nM and consisted of a narrow range of fragments centered at about 305 bp.

Short-read sequencing. Short-read sequencing of our library was performed on one paired-end flow cell of Illumina Genome Analyzer IIx (7 lanes of library and 1 lane of a ϕX control). Raw sequence data (76-nt read length) were processed by using Firecrest (image analysis) and Bustard (base calling) as part of the Illumina GA Pipeline, v1.4.0. The sequence data obtained was in SCARF format (Solexa Compact ASCII read format). After the appropriate conversions were performed, FASTQ files were used for quality analysis and filtering, and FASTA files were used for input into the genome assembly tool.

Stringency filters. The data set was filtered to three levels of stringency. The quality scores in the original data set ranged from the worst score of B to the best score of b (in ASCII order). We defined a bad score as any score less than D. Our lowest-stringency filter (filter 1 [F1]) discarded reads with 38 or more bad scores, the moderate stringency filter (F2) discarded reads with one or more bad scores, and the most stringent filter (F3) discarded reads with one or more N or ambiguous base calls. The original data set was denoted DO, and the progressively filtered data sets were designated DF1, DF2, and DF3.

Assembly. The short-read assembly tool that we used was Velvet 0.7.55 (48). For each of the four data sets (DO, DF1, DF2, and DF3), the velveth command was run with 8 or 9 different *k*-mer lengths (*k*). After each velveth command, a velvetg command was executed with the following options: exp_cov auto, min_contig_lgth 100, and ins_length 300. The assemblies with the maximum N50 for each data set were designated VO, VF1, VF2, and VF3, and assembly metrics were obtained from the Velvet output. To obtain a FASTA file of the unused reads, the option unused_reads yes was specified in the velvetg command.

Determination of an optimal assembly. Of VO, VF1, VF2, and VF3, the optimal assembly was selected based on an analysis of the quality of the reads that were used in an assembly but that were of a low-enough quality to be removed by the next level of filtering. These reads were identified by performing a Venn analysis. A Venn analysis was necessary because Velvet does not use all the input reads, nor does it output a file with only the used reads. The three files available to us were the Velvet input read FASTA file (Illumina and Velvet read identifiers), the Velvet unused-read FASTA file (Velvet identifier only), and the FASTQ file of the reads that were removed by the next filter (Illumina identifier only). The first step was to use hash tables to match all reads to their Illumina identifiers. The next step was to use the Illumina identifiers to find the "filtered used reads." The quality of the reads in the filtered-used-read region was assessed by plotting the

percent distribution of the quality scores, the percent distribution of reads with bad scores at each read position, the frequency of N homopolymers (N-mers), and the frequency of N's per read (N is an ambiguous base call when referring to reads). By evaluating these histograms, we were able to judge whether a filter was too stringent or not stringent enough. The optimal assembly determined from this analysis was then used for gene prediction, subcellular localization, and functional annotation.

Gene prediction. GeneMark-ES v2 was used for gene prediction from contigs of ≥3 kb (39). From each gene, we deduced the coding sequence (CDS) and the protein translation. Partial genes, i.e., lacking a start or stop codon, were removed from further analyses. TargetP 1.1 and SignalP 3.0 (13) were used to determine the subcellular localization of the gene products. TargetP determined if the protein was secretory or localized to the mitochondria, and SignalP predicted the presence of a signal peptide or membrane anchor. For genes involved in oxalate metabolism, the detection of a C-terminal peroxisomal target signal was provided by the PTSs Predictor tool (37).

Even though the nuclear condition of the fungus was unknown, the error removal algorithm of Velvet generated one consensus sequence for allele pairs. We verified this by creating a database from each gene plus 500 bp of its upstream sequence and then made pairwise alignments of the database against itself using blastn (1). Our threshold for screening the alignments was a bit score of ≥2,000 and a percent sequence identity of ≥92%.

Functional annotation. The Blast2GO suite (v 2.4.9) automated the process of functional annotation from the deduced protein sequences (18). The analysis was performed in July 2011. The order of analyses was as follows: blastp, map, annotate, InterPro search, mergo GO, and GOSlim. blastp retrieved the top 20 hits in the NCBI nr database and mapped the hits by four methods to obtain their associated gene ontology (GO) terms. The annotate rule found the most specific GO terms and their reliability, and additional terms were then retrieved based on conserved-domain searches of the InterPro database. All GO terms were merged and condensed to their broad functional categories by using the GO-Slim generic database. Query and graphing tools within the Blast2GO suite were used to summarize the results of the annotations.

We mined the Blast2GO output for annotations related to survival on copper (gene products involved in oxalate metabolism and copper homeostasis); the breakdown of pectin, hemicellulose, and cellulose (glycoside hydrolases [GHs] and carbohydrate binding module 1); lignin modification (laccases and ligninases); and the oxidative breakdown of wood by the Fenton reaction (genes involved in H_2O_2 metabolism, iron reduction, and quinone reduction). We also performed a blastp search (1) against the *F. radiculosa* database for genes that were not found in our annotations but that had wood decay functions in other species (E value of $<10^{-50}$). In *Phanerochaete chrysosporium*, these were the genes that encoded the low-molecular-weight glycoproteins (*glp*1 and *glp*2) (38) and cellobiose dehydrogenase, both of which could be involved in iron reduction (28). We also searched for a low-molecular-weight peptide isolated from the non-copper-tolerant brown rot fungus *Gloeophyllum trabeum* called the Gt factor, which has iron-reducing capabilities (42), and an oxalate efflux transporter identified from *F. palustris* (43). Manual curation, which involved the removal of fragments and chimeras (domains that appear in the same gene but that appear to be unrelated) and the validation of conserved domains and top blastp hits, was also performed.

Gene homologs of our curated set of copper tolerance and wood decay genes were identified by a blastp search of the *P. placenta* MAD 698-R v1.0 protein and *S. lacrymans* S7.9 v2.0 filtered gene model databases (http://www.jgi.doe.gov/) at E values of $<10^{-100}$ (or E values of $<10^{-50}$ for proteins with <320 residues). If retrieved sequences showed more than 95% amino acid identity or were truncations of a longer sequence, only one gene of the pair was retained. Retrieved sequences were also manually curated before they were included in the final gene count.

Data repository. The contigs can be downloaded from EMBL/GenBank (http://www.ncbi.nlm.nih.gov/bioproject) under WGS Bio-Project 72357 as the v1.0 assembly. The sequences, top blastp hits, and localization motifs of the manually curated set of copper tolerance and wood decay genes are also summarized in Table S1 in the supplemental material.

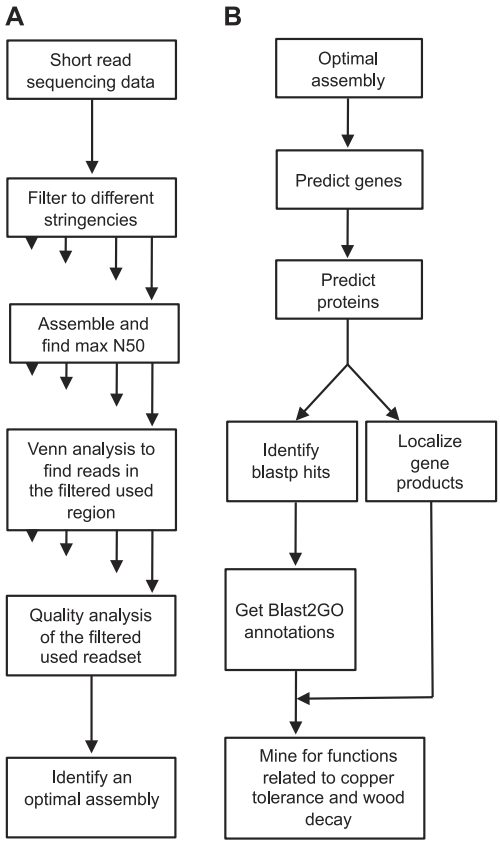

FIG 1 Pipeline used for genome assembly (A) and annotation (B). (A) The short-read sequencing data were filtered at different stringencies to produce the original and three filtered data sets (indicated by the four arrows). Each data set was assembled with multiple *k*-mer values to find the assembly with the maximum N50. Each maximum N50 assembly was subjected to a Venn analysis to identify the reads that were in the "filtered used" region. Quality analysis of the filtered used reads indicated which assembly was optimal. (B) Contigs from this optimal assembly served as the input for gene prediction. Genes were translated to proteins. The protein sequences were used in two workflows. The left workflow produced the top blastp hit and the gene ontology annotation. The right workflow involved signal motif identification. The combined annotations were then mined for functions related to wood decay and copper tolerance.

RESULTS

Pipeline. An overview of the steps that we developed for genome assembly and annotation is shown Fig. 1. The first goal was to identify an optimal assembly from the progressively filtered read sets (Fig. 1A). The second goal was to predict genes from the contigs in the optimal assembly and use the protein translations to determine subcellular localization and function (Fig. 1B). The genomic resource was then mined for functions related to copper tolerance wood decay.

Short-read sequencing and filtering. Error rates of the ϕX control for each paired-end read were 1.20% and 1.09%. The frequency of the base quality scores for the original data set is shown in Fig. S1A in the supplemental material. There was a total of 8.9 Gb in the data set, with the majority (>2.5 Gb) exhibiting high-

TABLE 1 Metrics from the original (VO) and filtered (VF1, VF2, and VF3) Velvet assemblies that had the maximum N50 values[a]

Assembly	*k*	Max N50 (kb)	Mean *k*-mer coverage (×)	Max contig length (kb)	Estimated genome size (Mb)	No. of used reads (M)	No. of unused reads (M)	No. of contigs >100 bp (K)
VO	45	66.2	61.6	341.2	33.1	86.6	31.1	15.4
VF1	51	65.8	57.0	347.0	33.6	85.1	28.7	16.9
VF2	37	23.7	50.1	148.1	31.0	46.1	16.8	11.2
VF3	37	24.0	52.5	148.1	30.9	46.1	16.7	14.3

[a] Half the assembly is covered by contigs with a size of N50 or larger; *k*-mer coverage is the number of times a *k*-mer was seen among the reads; *k*, *k*-mer length; max, maximum; M, million; K, thousand.

quality scores. About 0.5 Gb, however, had the lowest-quality score of B. Having defined a bad score as ≤D, a histogram of the frequency of bad scores by read position showed an exponential increase, indicating that the relationship was cumulative; i.e., once a read went bad, the rest of the read was also likely to be bad (see Fig. S1B in the supplemental material). The number of reads and number of ambiguous base calls in each data set after progressive filtering are listed in Table S2 in the supplemental material. DO had 117.7 million reads. F1 removed 4.0 million reads and 5.3 million N's. F2 removed an additional 50.9 million reads and 7.0 million N's, and F3 removed another 49,000 reads and 49,700 N's.

Assembly. The effects of various *k*'s on the N50 of the Velvet assemblies from the original data set DO and the filtered data sets DF1, DF2, and DF3 are shown in Fig. S2 in the supplemental material. In general, the lowest N50 values were at the extremes of the *k* values tested, with one maximum N50 at some intermediate value of *k*. An unexplained drop in the N50, however, was observed for the DF2 assembly at a *k* of 35. Maximum N50 values for DO, DF1, DF2, and DF3 were at *k* values of 45, 51, 37, and 37, respectively.

Optimal assembly. Based on the assembly metrics alone, it was difficult to assess which of the four maximum N50 assemblies was optimal (Table 1). The VO assembly had the greatest maximum N50 (66.2 kb) and the highest average *k*-mer coverage (61.6×). The VF1 assembly, on the other hand, had the largest *k* (k of 51) and the longest maximum contig length (347 kb). Maximum N50 values of VF2 and VF3 were both about one-third of those of VO and VF1, and the maximum contig lengths were both between one-third and one-half of those of VO and VF1. The estimated genome size ranged from 30.9 to 33.6 Mb, with the largest being from VF1. The number of reads used in the VO and VF1 assemblies was not quite twice the number of used reads from VF2 and VF3, and the number of unused reads was about one-third of the number of used reads, regardless of the assembly. The number of contigs (>100 bp) was greatest for the VF1 assembly (16.9 thousand [16.9K]) and fewest for the VF2 assembly (11.2K). Overall, the majority of the contigs were quite short (e.g., only 861 contigs were ≥3 kb in the VF1 assembly). By converting *k*-mer coverage to nucleotide coverage (47), we obtained the following nucleotide coverage values for each maximum N50 assembly: 146× for VO, 167× for VF1, 95× for VF2, and 100× for VF3.

The Venn analysis gave us a more concrete guide for identifying an optimal assembly. We were able to use the distribution of quality scores and N's of the filtered used reads to directly assess how the filters were affecting the assemblies. For the VO assembly, there were 28.2 million unused reads that were not filtered, 2.9 million unused reads that were filtered (moot reads), and 1 million filtered used reads (see Fig. S3A in the supplemental material). The intersection is moot because the reads were already removed by Velvet, and subsequent removal by the filter was redundant. For the VF1 assembly, there were 14.0 million unused reads that were filtered, 14.7 moot reads, and 36.2 million filtered used reads (see Fig. S3B in the supplemental material).

A closer examination of the 1 million filtered used reads of the VO assembly showed that they were of very low quality and had many ambiguous base calls (Fig. 2). The majority (58%) of the bases had bad scores (B or D) (Fig. 2A). More than 50% of the reads had bad scores starting at read position 35, and 100% of the reads had bad scores starting from read position 44 (Fig. 2B). There was an abundance of long N-mers ≥5 (Fig. 2C), and the number of N's per read was ≥5 (Fig. 2D). There were about 6,000 N-mers greater than a 35-mer in length, and there were 6,400 reads that had more than 40 N's per read. The frequencies of N-mers of lengths of 1 to 4 ranged from 71,564 for an N-mer of 1 to 1,147 for an N-mer of 4. The frequencies of 1 to 4 N's per read ranged from 52,268 to 307, respectively. Since *k* equals 45 for the VO assembly, the *k*-mers from these used reads would certainly contain long stretches of low-quality sequence. Therefore, we reasoned that the accuracy of the VO assembly could be improved by removing these 1 million reads with F1. This was corroborated by the assembly metrics (Table 1), where we saw an increase in the specificity or *k* of the VF1 assembly ($k = 45$ for VO, and $k = 51$ for VF1), with only a minor reduction in the maximum N50 (maximum N50 = 66.2 for VO, and maximum N50 = 65.8 for VF1).

Figure 3 charts a similar analysis of the 36.2 million filtered used reads of the VF1 assembly. This time, however, the majority of the reads were of acceptable quality. Only 10% of the bases had bad scores (Fig. 3A), and the percentage of reads with a bad score did not exceed 50% until read position 74 (Fig. 3B). Even if we consider that 10% of 36.2 million reads equals 3.62 million reads with bad scores, the frequencies of numbers of N-mers of ≥5 (Fig. 3C) and N's per read of ≥5 were extremely low. The longest N-mer was a 22-mer, and there were only 27 N-mers greater than a 10-mer. The maximum number of N's per read was 26, and there were only 595 reads with more than 10 N's per read (Fig. 3D). The frequencies of numbers of N-mers of <5 were as follows: 3.4 million for an N-mer of 1; 53,083 for an N-mer of 2; 757 for an N-mer of 3; and 287 for an N-mer of 287. The frequencies of <5 N's per read were as follows: 3.3 million for 1 N per read; 116,784 for 2 N's per read; 5,872 for 3 N's per read; and 1,539 for 4 N's per read. Given that the total number of bases in the contigs of the VF1 assembly was 6.5 Gb, it was unlikely that these single N′s would affect the assembly since they were well below the 1% SNP rate that was shown previously not to affect accuracy of the Velvet assembly in tests with simulated data (48). Therefore, we concluded that VF1 was our optimal assembly, obviating the need to

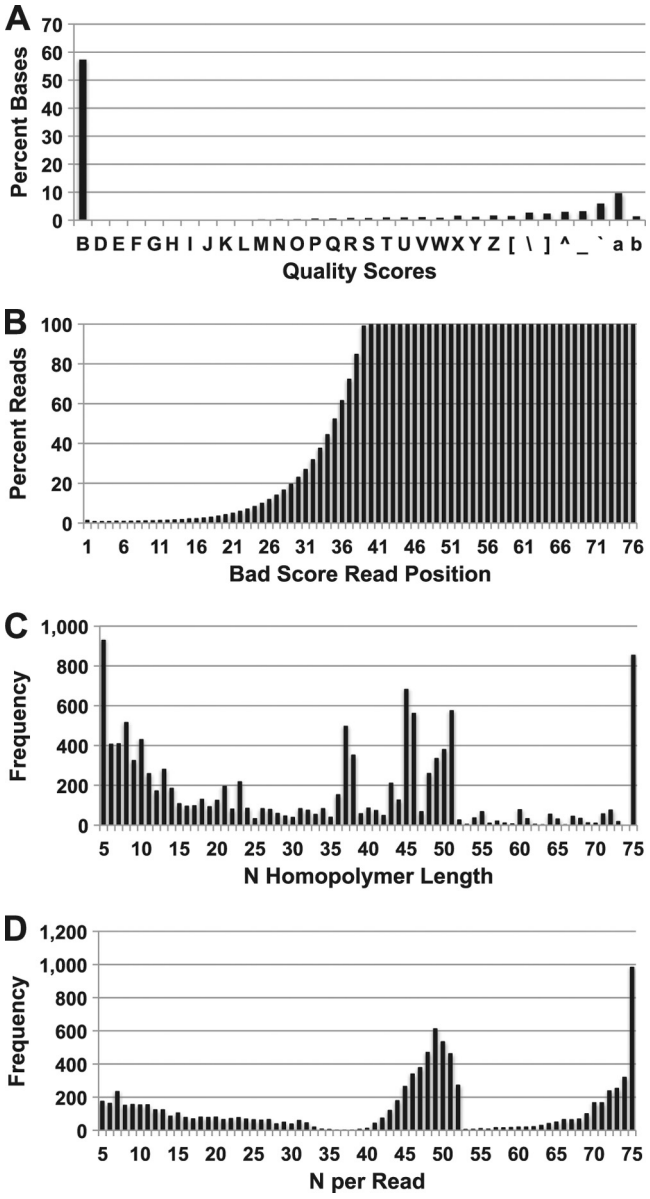

FIG 2 Quality analysis of the reads in the original data set that were in the filtered used region (see the shaded region of Fig. S3A in the supplemental material). The bar charts plot the percent distribution of the quality scores (A), the percentage of reads with a bad base by read position (B), the frequency of N homopolymers (C), and the frequency of N's per read (D). The filter removed reads with 38 or more bad bases.

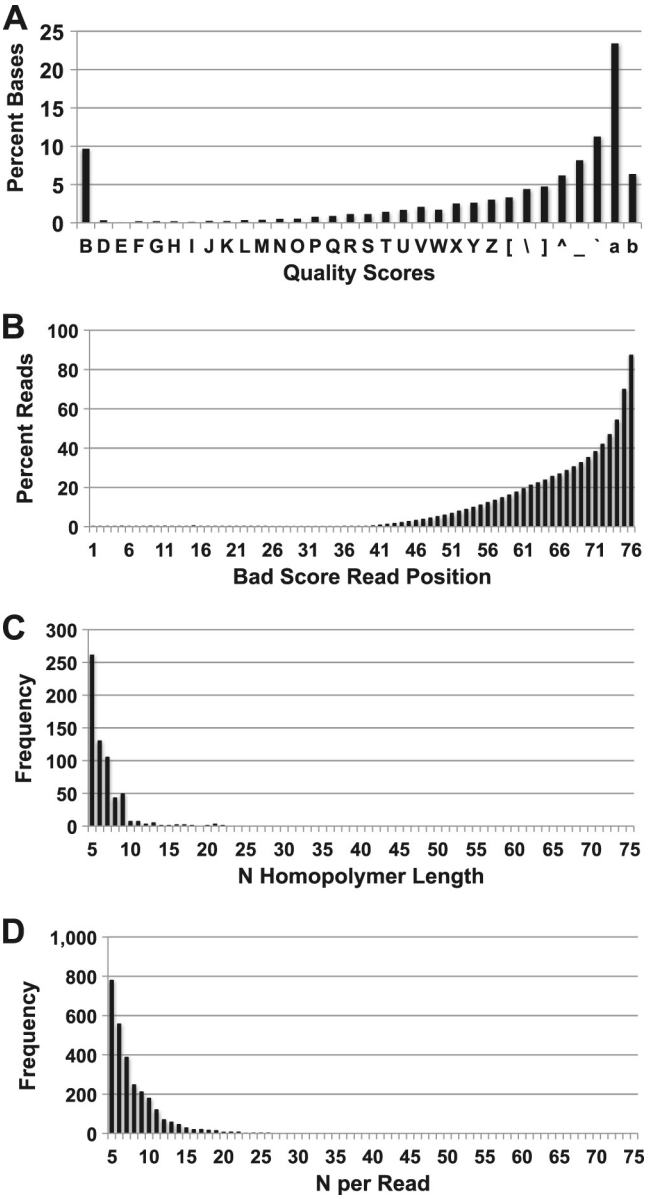

FIG 3 Quality analysis of the reads in the first filtered data set that were in the filtered used region (see the shaded region of Fig. S3B in the supplemental material). The bar charts plot the percent distribution of the quality scores (A), the percentage of reads with a bad base by read position (B), the frequency of N homopolymer lengths (C), and the frequency of N's per read (D). The filter removed reads with one or more bad bases.

carry VF2 and VF3 through a similar Venn analysis. In hindsight, we can now further interpret the assembly metrics of Table 1. Comparing VF1 to VF2, we observed a 1.4× loss of specificity (*k* dropped from 51 to 37, respectively) and a 2.8× decrease in the maximum N50 (65.8 and 23.7, respectively). The decreases in maximum N50 and maximum contig length were larger than the drop in *k*-mer coverage, suggesting that high-stringency filters caused the complete dropout of contigs rather than a progressive decrease in coverage.

The distribution of contigs in our optimal assembly, VF1, were as follows: 361 contigs ≥ 20 kb, 3 kb ≤ 500 contigs < 20 kb, and 100 bp ≤ 16,001 contigs < 3 kb. The 861 contigs that were ≥3 kb covered 28.4 Mb, or 84.5% of the genome, for an estimated genome size of 33.6 Mb. Of these 861 contigs, 392 had no internal gaps. The gap frequency histogram (not shown) showed 2 peaks: 123 gaps at ≤10 bp and 24 gaps at 195 bp. The largest single gap was 283 bp. It should be noted that Velvet converts ambiguous base calls in the reads to A's and reserves N's to denote gaps of a known size but an unknown sequence.

Gene prediction. A total of 9,262 complete genes were predicted from the VF1 assembly. Long contigs (≥20 kb) and short contigs (≥3 kb and <20 kb) produced 8,137 and 1,125 gene predictions, respectively. The gene statistics are listed in Table 2, along with values reported previously for *S. lacrymans* (12) and *P.*

TABLE 2 Comparison of gene descriptive statistics for *F. radiculosa*, *S. lacrymans*, and *P. chrysosporium*[a]

Descriptive statistic	Value		
	F. radiculosa	*S. lacrymans*	*P. chrysosporium*
No. of complete genes	9,262	11,238	10,048
Avg gene length (bp)	1,817	1,600	1,667
Avg CDS length (bp)	1,432	1,022	1,366
Avg intron length (bp)	70	77	64
Avg exon length (bp)	221	222	234
Avg no. of exons/gene	6.5	5.6	5.9
% GC content of CDS	53.8	48.0[b]	53.2

[a] Gene statistics for *S. lacrymans* monokaryon S7.9 v1.0 and *P. chrysoporium* v2.0 were taken from data reported previously by Eastwood et al. (12) and Vanden Wymelenberg et al. (41), respectively.
[b] D. C. Eastwood, personal communication.

TABLE 3 Numbers of putative gene models with functions related to copper tolerance in *F. radiculosa* and their homologous sequences in *P. placenta* and *S. lacrymans*

Gene annotation	No. of putative gene models		
	F. radiculosa	*P. placenta*	*S. lacrymans*
Oxalate metabolism			
Citrate synthase	3	3	4
Aconitase	2	4	2
Isocitrate lyase	2	2	2
Glyoxylate dehydrogenase-like[a]	6	5	3
Malate synthase	1	1	1
Malate dehydrogenase	3	2	4
Oxaloacetate hydrolase	1	2	1
Oxalate decarboxylase	7	4	3
Oxalate efflux transporter	0	0	1
Subtotal	25	23	21
Copper homeostasis			
Copper-transporting ATPase	3	4	2
Copper homeostasis CutC[b]	1	1	1
Subtotal	4	5	3
Total	29	28	24

[a] The *F. radiculosa* annotations were mitochondrial cytochrome or protein, but the sequences were homologous (E value of $<10^{-100}$) to the deduced protein of gene 1257, which had a top blastp hit to the cytochrome *c*-dependent glyoxylate dehydrogenase of *F. palustris*.
[b] The E value threshold was lowered to 10^{-50} because at least one of the *F. radiculosa* sequences was short (<320 residues).

chrysosporium (41). All gene statistics for *F. radiculosa* and *P. chrysosporium* were within 10% of each other. Comparing *F. radiculosa* with *S. lacrymans*, average intron and exon lengths were within 10%; the gene lengths, percent GC contents of CDSs; and numbers of exons/gene were within 15%; and numbers of complete genes and average CDS lengths were within 30%.

TargetP recognized 287 gene products as being localized to the mitochondria and 1,213 as being secreted. For the gene products localized to the mitochondria, SignalP identified 188 with signal peptides and 99 with membrane signal anchors. For the secreted products, SignalP found 986 with signal peptides and 227 with signal anchors. SignalP found another 250 gene products not localized by TargetP, 103 of which had signal peptides and 148 of which had signal anchors. A total of 1,750 proteins were localized by TargetP and SignalP.

Annotation. Results of the Blast2GO analysis (9,262 total predicted genes) showed that 5,407 genes (58%) had GO annotations, 985 genes (11%) were mapped but had no annotations, 1,833 genes (20%) could not be mapped, and 1,037 genes (11%) had no product matches in the NCBI nr database (E value threshold set to 1E−3). Overall, most of the translated proteins exhibited a high level of similarity with at least one gene product in the nr database (80%, or 7,459 genes, had top blastp hits with E values equal to or less than 1E−20). The greatest number of top blastp hits were against *P. placenta* (2,973 top hits) and *S. lacrymans* (2,910 top hits). The types (and distributions) of GO level 2 annotations were as follows: metabolic process (3,072 genes), cellular metabolic process (1,702 genes), localization (676 genes), biological regulation (495 genes), cellular component organization (295 genes), response to stimulus (246 genes), signaling (193 genes), developmental process (128), and multicellular organismal process (103 genes).

Genes for copper tolerance. The numbers of putative gene models with functions related to oxalate metabolism are shown in Table 3. A notable difference was the absence of oxalate efflux transporters in *F. radiculosa* and *P. placenta*. There was also more variation in the number of gene homologs detected for glyoxylate dehydrogenase-like proteins and oxalate decarboxylase than for citrate synthase, aconitase, isocitrate lyase, malate synthase, malate dehydrogenase, and oxaloacetate hydrolase. More variation means that one of the species had a gene count that was at least ±3 from the *F. radiculosa* count. Of the *F. radiculosa* oxalate metabolism genes, one isocitrate lyase (gene 415) and three glyoxylate dehydrogenase-like gene products (genes 1257, 1757, and 2821) carried peroxisomal target signals, and four oxalate decarboxylase genes (genes 5380, 6399, 7157, and 8686) carried signal peptide motifs for an extracellular function.

Among the genes with roles in regulating copper concentrations (Table 3), there were three copper-transporting ATPases and one copper homeostasis CutC gene in *F. radiculosa*. The only major difference was that one copper-transporting ATPase gene (gene 4719) had no homolog in *S. lacrymans*. Of the four *F. radiculosa* genes, one copper-transporting ATPase (gene 974) encoded a signal peptide.

Genes for wood degradation. Twelve glycoside hydrolase (GH) families that had putative roles in lignocellulose degradation are listed in Table 4. The majority of the GH genes in *F. radiculosa* (74%) possessed signal peptide motifs. The species that were considered dissimilar lacked homologs to one or more *F. radiculosa* genes. One GH5 cellulase (gene 4419), which had a conserved domain for carbohydrate binding module 1, had no homolog in *P. placenta*, while four of the other GH5 cellulases (genes 440, 508, 509, and 7362) had no homologs in *S. lacrymans*. Of the five GH5 cellulases that had homologs in *S. lacrymans*, however, some were more numerous, and the total number of GH5 cellulases for each of the three species was close (9 or 10 genes). Two GH28 proteins (genes 532 and 3577) had no homologs in *P. placenta*, and four GH28 proteins (genes 189, 1512, 1513, and 3577) lacked homologs in *S. lacrymans*. Two GH43 genes (genes 2569 and 2570) also had no similar sequences in *P. placenta* or *S. lacrymans*. An-

TABLE 4 Numbers of putative gene models with functions related to wood degradation in *F. radiculosa* and their homologous sequences in *P. placenta* and *S. lacrymans*

Gene annotation	No. of putative gene models		
	F. radiculosa	*P. placenta*	*S. lacrymans*
Glycoside hydrolases			
GH2 possible β-mannosidase	3	6	2
GH3 possible β-glucosidase or β-xylosidase	6	6	9
GH5 cellulase	9	10	9
GH5 BglC endoglucanase	3	7	3
GH6 cellobiohydrolase	0	0	1
GH10 related to endo-β-xylanase[a]	3	3	1
GH12 related to endo-β-glucanase[a]	2	3	2
GH28 polygalacturonase and rhamnogalacturonase[c]	8	7	4
GH43	3	1	1
GH51 α-L-arabinofuranosidase	1	2	1
GH53 arabinogalactan endo-1,4-β-galactosidase	1	2	1
GH61[a]	2	3	3
GH115 α-glucuronidase	2	1	1
Subtotal	43	51	38
Cellulose binding			
Carbohydrate binding module 1[b]	2	0	8
Subtotal	2	0	8
H_2O_2 metabolism			
Alcohol oxidase and alcohol oxidase-like	6	6	4
Aryl-alcohol oxidase	4	3	0
Copper radical oxidase	3	3	3
Catalase	3	6	3
Subtotal	16	18	10
Lignin modification			
Laccase	2	3	4
Multicopper oxidase	2	3	2
Low-redox peroxidase	1	2	0
Subtotal	5	8	6
Iron redox cycling			
Iron reductase	3	5	2
Quinone reductase	4	7	3
Iron binding glycoprotein[a]	1	1	0
Cellobiose dehydrogenase	0	0	2
Subtotal	8	13	7
Total	73	90	69

[a] The E value threshold was lowered to 10^{-50} because at least one of the *F. radiculosa* sequences was short (<320 residues).
[b] The number of genes with a carbohydrate module 1 binding domain was detected by a keyword search of each species database.
[c] Gene 4419 is also a GH5 cellulase.

other prominent difference was the absence of GH6 cellobiohydrolase in *F. radiculosa* and *P. placenta*. The remaining four families (GH2, GH3, GH5 BglC endoglucanase, and GH10) were distinguished by differences in the numbers of homologs detected, but no consistent pattern was observed. The GH12, GH51, GH61, and GH115 families all had similar numbers of homologs (within ±1 of the *F. radiculosa* gene count). Another related finding concerned the number of genes with cellulose binding motifs (Table 4). There were 2 genes with carbohydrate binding module 1 in *F. radiculosa*, none in *P. placenta*, and 8 in *S. lacrymans*.

Among the genes involved in H_2O_2 metabolism (Table 4), a notable discovery was that *S. lacrymans* had no homolog to the four *F. radiculosa* aryl alcohol oxidases (genes 5573, 8290, 8291, and 8299). *S. lacrymans* also had two fewer homologs than either species when alcohol oxidase-like genes were compared, and both *F. radiculosa* and *S. lacrymans* had three catalase homologs, compared to six for *P. placenta*. The numbers of copper radical oxidases, on the other hand, were more consistent across species. The numbers of signal peptides found for the *F. radiculosa* alcohol oxidase-like protein, aryl-alcohol oxidase, copper radical oxidase, and catalase genes were 0, 3, 3, and 1, respectively.

A difference observed among the lignin modification genes (Table 4) was the presence of low-redox peroxidase homologs in *F. radiculosa* and *P. placenta* but not in *S. lacrymans.* In addition, *S. lacrymans* had two more laccase homologs than *F. radiculosa.* The numbers of multicopper oxidases, however, were similar among all three species. Motifs for extracellular localization were found on all the *F. radiculosa* lignin modification genes except for one multicopper oxidase.

A comparison of the iron redox genes displayed some interesting differences (Table 4). Although an iron binding glycoprotein was found in *F. radiculosa* and *P. placenta*, none was detected in *S. lacrymans.* In contrast, neither *F. radiculosa* nor *P. placenta* had cellobiose dehydrogenase genes, but two homologs were found in *S. lacrymans.* The numbers of homologs to the *F. radiculosa* iron reductases and quinone reductases varied, with *P. placenta* and *S. lacrymans* exhibiting the highest and lowest gene counts, respectively. Another feature of the iron redox genes of *F. radiculosa* was the relatively low number of secretion motifs. One iron reductase had a signal peptide, and one iron binding glycoprotein had a signal anchor. Genes showing similarity to the Gt factor (15 amino acids) were not detected.

DISCUSSION

Our results showed that it was entirely feasible to produce a comprehensive set of structurally and functionally annotated genes for a basidiomycete fungus using only short-read sequencing. Our approach utilized a paired-end strategy because it was known to increase the N50 by 3-fold compared to single-end reads in *Escherichia coli* (5). The rationale for selecting a 76-nt read length was because it approximated the 60-nt read length barrier of *Saccharomyces cerevisiae* (5). The read length barrier is the threshold above which assemblies fail to improve and below which assemblies deteriorate (5). Furthermore, we selected Velvet because it was known to produce highly accurate assemblies (6, 10, 33), even in the presence of 1% sequencing errors or single-nucleotide polymorphisms (48). This was critical, since short-read sequencing has a higher error rate than Sanger methods, and in our case, we did not know whether the genome that we sequenced was haploid or diploid. Our search for allelic pairs produced few candidates,

suggesting either that DNA isolations came from haploid hyphae or that the error removal algorithms in Velvet caused sequence differences of allelic pairs to be simplified to a single consensus sequence.

Once we knew that we could produce an assembly with our data, we proceeded to develop a systematic method for refining the assembly. Our approach used stepwise filters to create smaller, higher-quality data sets. By varying the k for each data set, we were able to find the value of *k* that produced the assembly with the maximum N50 value (47). Having identified the maximum N50 assemblies for the unfiltered and filtered data sets, our next goal was to find the threshold for filtering. We developed a Venn analysis to identify the filtered used reads and then assessed their quality. By performing this analysis, we showed that the best assembly from the four data sets (original data set plus 3 filtered data sets) was not the one with the largest maximum N50 but, rather, was the assembly with the maximum *k* or specificity. Analysis of the quality of the filtered used reads showed that a high-stringency filter caused reductions in N50 with only minor gains in accuracy, while a low-stringency filter produced modest gains in N50 at the expense of accuracy.

The gene prediction tool GeneMark-ES v2 was chosen because it was shown previously to be accurate and sensitive when tested on the genomes of nine different fungal species (39). For basidiomycetes like *F. radiculosa*, the branch point sequence, which guides lariat formation during splicing, is conserved (27). GeneMark allows for the presence and absence of branch point sequences in the intron model, which gives the algorithm more flexibility to correctly locate intron boundaries. This version of GeneMark also predicts genes *ab initio*, meaning that the hidden Markov model trains directly on the assembly, and a separate EST data set to train the algorithm is not needed.

However, genomic analysis based on short-read sequencing alone has its limitations. The first is that without EST or other RNA sequence data, there was no way to validate CDS predictions or predict genes from splice sites that use donor and acceptor sequences other than the canonical GT-AG introns. The frequencies of GC-AG introns are 0.6% in *Caenorhabditis elegans* (14), 0.7% in mammals (4), and 1.0 to 1.2% in the Ascomycota (35). For the Basidiomycota, genome surveys of noncanonical splice site frequencies have not yet been reported, but initial data suggest that the rate may be as high as 3%. In a survey of four genes, 1/38 introns had a GC-AG splice site in *Armillaria mellea* (31). Another limitation is that without longer reads from a another platform, we were unable to join the contigs into the longer scaffolds that characterized the *G. clavigera* (10) and *S. macrospora* (33) assemblies. In spite of these limitations, we were still able to predict the sequences of over 9,000 genes, assign functional annotations to 58% of the genes, and then identify 102 genes with potential roles in copper tolerance and wood decay.

A comparison of the *F. radiculosa* gene descriptive statistics with those obtained previously for *P. chrysosporium* (41) and *S. lacrymans* (12) showed that the average exon length was within 6%, the average intron length was within 10%, and the average number of exons/gene was within 15%. Since these wood decay saprobes are from different taxonomic orders, we would expect the greatest evolutionary pressure to be exerted on keeping the average exon length the most similar, which was what we observed. A comparison of the deduced protein sequence similarity showed that *F. radiculosa* was most closely related to *P. placenta*, with *S. lacrymans* coming in right behind. This was expected, however, since *F. radiculosa* and *P. placenta* exhibit similar biologies and taxonomical relationships. They are both highly copper tolerant (19, 20). They cannot decay wood unless it is in direct contact with moisture, and they belong to the order Polyporales. *S. lacrymans*, on the other hand, is less copper tolerant (20, 22), has the ability to translocate water and decay wood that is not in direct contact with moisture, and belongs to the order Boletales.

We found some notable differences among the putative gene homologs with roles in copper tolerance. The first difference was that only *S. lacrymans* had a homolog of the characterized oxalate efflux transporter from *F. palustris* (43). This finding was surprising, since both *F. radiculosa* and *P placenta* are known to secrete some of the highest levels of oxalate in response to high copper concentrations (20). The absence of genes for oxalate transport may indicate that *F. radiculosa* and *P placenta* have evolved a novel structure for this gene product. The second discovery was that *S. lacrymans* had no homolog of one of the copper-transporting ATPases. Since copper-transporting ATPases prevent intracellular concentrations of copper from becoming toxic in non-wood-decaying fungi (45), it is possible that this homolog may contribute to the higher levels of copper tolerance that distinguish *F. radiculosa* and *P. placenta* from *S. lacrymans.*

A more subtle difference was the variation in the number of genes involved in oxalate metabolism. The greater number of glyoxylate dehydrogenase genes than oxaloacetate hydrolase genes in *F. radiculosa* and *P. placenta* suggests that the shortcut pathway, which produces oxalate directly from glyoxylate, may function in these two highly copper-tolerant brown rot species. This contrasts with the longer pathway that has been described for *F. palustris*, where oxalate production proceeds through the intermediates malate and oxaloacetate (32). Another distinction was that *S. lacrymans* had the fewest homologs of glyoxylate dehydrogenase and oxalate decarboxylase. It remains to be determined, however, if this difference contributes to the smaller amount of oxalate secreted by this species.

Although all three species showed many similarities in the numbers and types of putative genes involved in the oxidative and hydrolytic decay of wood, when differences were found, the general pattern observed was that *F. radiculosa* and *P. placenta* exhibited more similarities than did *F. radiculosa* and *S. lacrymans*. For example, GH6 cellobiohydrolase and cellobiose dehydrogenase were found in *S. lacrymans* but not in *F. radiculosa* and *P. placenta*. Or, we observed the opposite pattern. Aryl alcohol oxidases, a low-redox peroxidase, and an iron binding glycoprotein were found in *F. radiculosa* and *P. placenta* but not in *S. lacrymans.* For the GH5 cellulases and the GH28 pectinases, *F. radiculosa* and *P. placenta* shared more homologs than did *F. radiculosa* and *S. lacrymans.* There were only two exceptions where *F. radiculosa* and *P. placenta* differed. *F. radiculosa* had two GH43 homologs that were absent in both *P. placenta* and *S. lacrymans*, and *P. placenta* lacked any gene products with carbohydrate domain module 1.

Results of gene expression studies have shown an active role of many glycoside hydrolases during the growth of *P. placenta* (30, 40) and *S. lacrymans* (12) on wood (or microcrystalline cellulose). Several of these genes were homologous to the *F. radiculosa* genes. For example, in *S. lacrymans*, five of the GH5 cellulase homologs exhibited increased expression levels on wood (genes 355683, 361086, 362272, 433208, and 433209) (12). Two of these genes (genes 355683 and 433209) had carbohydrate binding module 1

domains and were homologous to *F. radiculosa* gene 4419. Of the GH5 cellulase homologs found in *P. placenta*, four showed increased expression levels (genes 115648, 121713, 121831, and 95568) and one exhibited decreased expression levels (gene 117690) on wood (30, 40). Other homologous genes that showed increased expression levels on wood were GH2 (genes 114395 and 57564), GH3 (gene 46915), GH10 (genes 113670 and 105534), and GH28 (gene 111730) in *P. placenta* (30, 40) and GH10 (gene 349170), GH28 (gene 453971), and GH61 (genes 335267 and 465649) in *S. lacrymans* (12). Thus, it appears that many of these homologous glycoside hydrolases are expressed and are likely responsible for the enzymatic breakdown of pectin, cellulose, and hemicellulose.

Oxidative decay in brown rot fungi involves the highly reactive hydroxyl free radical produced chemically by the Fenton reaction ($Fe^{2+} + H_2O_2 + H^+ \Rightarrow Fe^{3+} + \cdot OH + H_2O$) (2, 21, 25, 26). These highly reactive but short-lived free radicals are capable of randomly fragmenting long molecules of cellulose by scission (25). Unlike the role that the glycoside hydrolase genes play in polysaccharide degradation, there are still many unresolved questions concerning the origins of the Fenton reactants. For example, in oxalate-producing brown rot fungi, laccase has been proposed to be a necessary participant. It causes a one-electron abstraction from hydroquinones to produce semiquinone radicals that, in turn, interact with other chemicals in a series of redox reactions to produce the Fenton reactants (44). Gene expression data linking increased oxalate production with laccase, however, have not yet been reported. Only an increased expression level of laccase has been observed (gene 111314 in *P. placenta* [40] and gene 362730 in *S. lacrymans* [12]), and the increases in expression levels observed on wood were relatively low. It is possible, however, that the lack of evidence could be attributed to the timing of the analysis. Hydroxyl free radical attack has been associated with incipient decay, during which the hydroxyl free radicals cause an increase in wood pore size (15, 23). On the other hand, most studies have been conducted when the gene expression levels of the later-acting hemicellulases and cellulases were high (12, 30, 40).

Another issue is that theoretical calculations predict that the action of laccase alone is sufficient to generate enough Fenton reactants for incipient decay (44). If so, then why do we see increased expression levels of genes for H_2O_2 production during brown and dry rot decay? An extracellular-acting alcohol oxidase and a copper radical oxidase exhibited increased expression levels in *S. lacrymans* (gene 439506) (12) and *P. placenta* (gene 56703) (40), respectively. However, with regard to iron and quinone reduction, there were no differentially expressed quinone reductase genes in *S. lacrymans* (12) but increased expression levels of three iron reductases (30) and one 1,4-benzoquinone reductase (30, 40) in *P. placenta*.

It is clear that we still do not understand many of the complex relationships that tie the observed copper tolerance of brown rot fungi to oxalate production and the generation of Fenton reactants. Furthermore, we cannot ignore that almost half the predicted gene sequences still had no annotations, and for those that did, they are only as good as the evidence codes upon which they are based. Once a genome is sequenced, however, we possess a powerful tool that can systematically interrogate and track the network of biological processes that make nutrients quickly available to the fungus while ensuring its survival during the harsh and changing conditions of extracellular digestion.

ACKNOWLEDGMENTS

This work was supported by grants from the Lucas Biodeterioration Laboratory (Department of Forest Products, Mississippi State University), the Life Sciences and Biotechnology Institute (Mississippi State University), Wood Utilization Research (USDA), and the National Science Foundation (under grant NSF EPS-0903787).

We are grateful to Chuan-Yu Hsu, Cetin Yuceer, Cathy Gresham, and Brandon Malone, who provided advice and/or equipment for the project.

REFERENCES

1. **Altschul SF, et al.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**:3389–3402.
2. **Arantes V, Milagres AM, Filley TR, Goodell B.** 2011. Lignocellulosic polysaccharides and lignin degradation by wood decay fungi: the relevance of nonenzymatic Fenton-based reactions. J. Ind. Microbiol. Biotechnol. **38**:541–555.
3. **Bentley DR, et al.** 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature **456**:53–59.
4. **Burset M, Seledtsov IA, Solovyev VV.** 2000. Analysis of canonical and non-canonical splice sites in mammalian genomes. Nucleic Acids Res. **28**:4364–4375.
5. **Chaisson MJ, Brinza D, Pevzner PA.** 2009. *De novo* fragment assembly with short mate-paired reads: does the read length matter? Genome Res. **19**:336–346.
6. **Chaisson MJ, Pevzner PA.** 2008. Short read fragment assembly of bacterial genomes. Genome Res. **18**:324–330.
7. **Clausen CA, Green F.** 2003. Oxalic acid overproduction by copper-tolerant brown-rot basidiomycetes on southern yellow pine treated with copper-based preservatives. Int. Biodeterior. Biodegradation **51**:139–144.
8. **Clausen CA, Green F, Woodward BM, Evans JW, DeGroot RC.** 2000. Correlation between oxalic acid production and copper tolerance in *Wolfiporia cocos.* Int. Biodeterior. Biodegradation **46**:69–76.
9. **Clausen CA, Jenkins KM.** 2011. Chronicles of *Fibroporia radiculosa* (=*Antrodia radiculosa*) TFFH 294. General technical report 240. U.S. Department of Agriculture Forest Service Forest Products Laboratory, Madison, WI.
10. **Diguistini S, et al.** 2009. *De novo* genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. Genome Biol. **10**:R94.
11. **DiGuistini S, et al.** 2011. Genome and transcriptome analyses of the mountain pine beetle-fungal symbiont *Grosmannia clavigera*, a lodgepole pine pathogen. Proc. Natl. Acad. Sci. U. S. A. **108**:2504–2509.
12. **Eastwood DC, et al.** 2011. The plant cell wall-decomposing machinery underlies the functional diversity of forest fungi. Science **333**:762–765.
13. **Emanuelsson O, Brunak S, von Heijne G, Nielsen H.** 2007. Locating proteins in the cell using TargetP, SignalP and related tools. Nat. Protoc. **2**:953–971.
14. **Farrer T, Roller AB, Kent WJ, Zahler AM.** 2002. Analysis of the role of *Caenorhabditis elegans* GC-AG introns in regulated splicing. Nucleic Acids Res. **30**:3360–3367.
15. **Flournoy DS, Kent Kirk T, Highley TL.** 1991. Wood decay by brown-rot fungi: changes in pore structure and cell wall volume. Holzforschung **45**: 383–388.
16. **Fomina M, et al.** 2005. Role of oxalic acid overexcretion in transformations of toxic metal minerals by *Beauveria caledonica.* Appl. Environ. Microbiol. **71**:371–381.
17. **Goodell B.** 2003. Brown-rot fungal degradation of wood: our evolving view, p 97–117. *In* Goodell B, Nicholas DD, Schultz TP (ed), Wood deterioration and preservation: advances in our changing world. American Chemical Society, Washington, DC.
18. **Gotz S, et al.** 2008. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. **36**:3420–3435.
19. **Green F, Clausen CA.** 2005. Copper tolerance of brown-rot fungi: oxalic acid production in southern pine treated with arsenic-free preservatives. Int. Biodeterior. Biodegradation **56**:75–79.
20. **Green F, Clausen CA.** 2003. Copper tolerance of brown-rot fungi: time course of oxalic acid production. Int. Biodeterior. Biodegradation **51**:145–149.
21. **Halliwell G.** 1965. Catalytic decomposition of cellulose substrates. Biochem. J. **95**:35–40.

22. **Hastrup ACS, Green F, Clausen CA, Jensen B.** 2005. Tolerance of *Serpula lacrymans* to copper-based wood preservatives. Int. Biodeterior. Biodegradation **56**:173–177.
23. **Irbe I, et al.** 2006. On the changes of pinewood (*Pinus sylvestris* L.) chemical composition and ultrastructure during the attack by brown-rot fungi *Postia placenta* and *Coniophora puteana.* Int. Biodeterior. Biodegradation **57**:99–106.
24. **Jarosz-Wilkolazka A, Gadd GM.** 2003. Oxalate production by wood-rotting fungi growing in toxic metal-amended medium. Chemosphere **52**:541–547.
25. **Kirk TK, Ibach R, Mozuch MD, Conner AH, Highley TL.** 1991. Characterization of cotton cellulose depolymerized by a brown-rot fungus, by acid, or by chemical oxidants. Holzforschung **45**:239–244.
26. **Koenigs JW.** 1974. Hydrogen peroxide and iron: a proposed system for decomposition of wood by brown-rot basidiomycetes. Wood Fiber Sci. **6**:66–80.
27. **Kupfer DM, et al.** 2004. Introns and splicing elements of five diverse fungi. Eukaryot. Cell **3**:1088–1100.
28. **Li B, Nagalla SR, Renganathan V.** 1997. Cellobiose dehydrogenase from *Phanerochaete chrysosporium* is encoded by two allelic variants. Appl. Environ. Microbiol. **63**:796–799.
29. **Li R, et al.** 2010. The sequence and *de novo* assembly of the giant panda genome. Nature **463**:311–317.
30. **Martinez D, et al.** 2009. Genome, transcriptome, and secretome analysis of wood decay fungus *Postia placenta* supports unique mechanisms of lignocellulose conversion. Proc. Natl. Acad. Sci. U. S. A. **106**:1954–1959.
31. **Misiek M, Hoffmeister D.** 2008. Processing sites involved in intron splicing of *Armillaria* natural product genes. Mycol. Res. **112**:216–224.
32. **Munir E, Yoon JJ, Tokimatsu T, Hattori T, Shimada M.** 2001. A physiological role for oxalic acid biosynthesis in the wood-rotting basidiomycete *Fomitopsis palustris.* Proc. Natl. Acad. Sci. U. S. A. **98**:11126–11130.
33. **Nowrousian M, et al.** 2010. *De novo* assembly of a 40 Mb eukaryotic genome from short sequence reads: *Sordaria macrospora*, a model organism for fungal morphogenesis. PLoS Genet. **6**:e1000891.
34. **Paterson AH, Brubaker CL, Wendel JF.** 1993. A rapid method for extraction of cotton (*Gossypium* spp.) genomic DNA suitable for RFLP or PCR analysis. Plant Mol. Biol. Rep. **11**:122–127.
35. **Rep M, et al.** 2006. The presence of GC-AG introns in *Neurospora crassa* and other euascomycetes determined from analyses of complete genomes: implications for automated gene prediction. Genomics **87**:338–347.
36. **Rowell RM.** 2005. Handbook of wood chemistry and wood composites. CRC Press, Boca Raton, FL.
37. **Schluter A, Real-Chicharro A, Gabaldon T, Sanchez-Jimenez F, Pujol A.** 2010. PeroxisomeDB 2.0: an integrative view of the global peroxisomal metabolome. Nucleic Acids Res. **38**:D800–D805.
38. **Tanaka H, et al.** 2007. Characterization of a hydroxyl-radical-producing glycoprotein and its presumptive genes from the white-rot basidiomycete *Phanerochaete chrysosporium.* J. Biotechnol. **128**:500–511.
39. **Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M.** 2008. Gene prediction in novel fungal genomes using an *ab initio* algorithm with unsupervised training. Genome Res. **18**:1979–1990.
40. **Vanden Wymelenberg A, et al.** 2010. Comparative transcriptome and secretome analysis of wood decay fungi *Postia placenta* and *Phanerochaete chrysosporium.* Appl. Environ. Microbiol. **76**:3599–3610.
41. **Vanden Wymelenberg A, et al.** 2006. Computational analysis of the *Phanerochaete chrysosporium* v2.0 genome database and mass spectrometry identification of peptides in ligninolytic cultures reveal complex mixtures of secreted proteins. Fungal Genet. Biol. **43**:343–356.
42. **Wang W, Huang F, Mei Lu X, Ji Gao P.** 2006. Lignin degradation by a novel peptide, Gt factor, from brown rot fungus *Gloeophyllum trabeum.* Biotechnol. J. **1**:447–453.
43. **Watanabe T, et al.** 2010. Oxalate efflux transporter from the brown rot fungus *Fomitopsis palustris.* Appl. Environ. Microbiol. **76**:7683–7690.
44. **Wei DS, et al.** 2010. Laccase and its role in production of extracellular reactive oxygen species during wood decay by the brown rot basidiomycete *Postia placenta.* Appl. Environ. Microbiol. **76**:2091–2097.
45. **Weissman Z, Berdicevsky I, Cavari BZ, Kornitzer D.** 2000. The high copper tolerance of *Candida albicans* is mediated by a P-type ATPase. Proc. Natl. Acad. Sci. U. S. A. **97**:3520–3525.
46. **White TJ, Bruns SL, Taylor J.** 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics, p 315–322. *In* Innis MA, Gelfand DH, Sninsky JJ, White TJ (ed), PCR protocols: a guide to methods and applications. Academic Press, San Diego, CA.
47. **Zerbino D.** 2008. Velvet manual version 0.7. http://www.ebi.ac.uk/~zerbino/velvet/.
48. **Zerbino DR, Birney E.** 2008. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Res. **18**:821–829.