〔東京家政大学研究紀要　第53集 (2), 2013, pp.17～30〕

High quality draft genome sequence analysis of the edible mushroom *Grifola frondosa*

Sato, Masayuki[1,2,3]*, Kurahashi, Atsushi[1]* Ezaki, Masahiro[4] Takeda, Aya[4] Uemura, Yasuo[4] Nishi, Tatsunari Nishibori, Kozo[4] and Fujimori, Fumihiro[2,3,†]

[1] *Yukiguni Maitake Co., Ltd., 89 Yokawa, Minamiuonuma-shi, Niigata 949-6695, Japan*
[2] *HyphaGenesis Inc., 4025-6, Tamagawa-gakuen 6 cho-me, Machida, Tokyo 194-0041 Japan*
[3] *Laboratory of Biological Science and Technology, Tokyo Kasei University, 1-18-1 Kaga, Itabashi-ku, Tokyo 173-8062, Japan*
[4] *Genaris, Inc., Leading Venture Plaza 2-30875-1 Ono-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0046 Japan*

Accepted for publication 20 December 2012

Abstract

The edible mushroom *Grifola frondosa* (Maitake mushroom) is mass-produced by artificial cultivation. However, the mechanisms of fruiting body differentiation are not well understood. To understand this mechanism by molecular biological approaches, we present a high quality draft sequence of the *G. frondosa* genome obtained using Roche 454 GS FLX Titanium and Illumina Genome Analyzer IIx technologies. A total of 280 scaffolds were assembled in the 33.8 Mb genome sequence using optimal assembly approaches combining 454 and Illumina data. The 16,097 gene models were predicted within the genome. Using ESTs from 13 different developmental stages, the number of genes predicted increased to approximately 1.5-fold that of the *ab initio* program that uses only the query genomic sequence as input data. Functional annotation of these models revealed an extensive set of wood- and other polysaccharide-degrading enzymes. Understanding this lignocellulolytic enzyme system will help improve utilization of woody sawdust as a substrate in the cultivation medium. Our study demonstrates that combining 454 and Illumina data substantially improves *de novo* genome assembly, and that EST-based gene prediction represents an effective approach for improving sequence accuracy.

Key words : *Grifola frondosa*; genome sequence; next-generation sequencing; *de novo* sequence assembly; lignocellulolytic enzyme

(1) Introduction

Grifola frondosa (Maitake mushroom) is a well-known edible basidiomycete mushroom that is commercially produced by artificial cultivation. It has also been reported to possess various potential bioactive properties such as anti-tumor, anti-viral, anti-diabetic, anti-oxidant, and immune system-enhancing effects[1-5]. *G. frondosa* has been investigated primarily for its pharmacological components; however, other molecular biological research on this species has not been extensively reported. The cultivation process for *G. frondosa* occurs in three stages: spawn run, primordia development and fruiting body differentiation. During mushroom cultivation, fruiting body differentiation is induced by controlling various environmental factors, such as temperature, light, humidity, and other physical stimuli. The combination of these factors induces genes required for the initiation of fruiting body differentiation, although the exact mechanisms involved are not clear. The morphological changes that occur during cultivation probably involve complicated networks of transcriptional regulation. Therefore, molecular biological information is necessary to enhance our knowledge of the mechanisms of fruiting body differentiation, mating pheromones, and signal transduction.

In previous studies, we identified genes associated

* These authors contributed equally to this work
† To whom correspondence should be addressed. Fax: +81-3-3961-0024; E-mail: fujimori@tokyo-kasei.ac.jp

with each phase of *G. frondosa* cultivation by transcriptome analysis[6]. The genes related to cultivation were assigned expected roles; however, their exact functions remain unclear. To clarify the roles of these genes of unknown function, functional analysis such as gene targeting is necessary in *G. frondosa*. Genome information will be necessary to achieve this purpose, but until now, no genome sequences of *G. frondosa* have been available.

Initial drafts of most eukaryotic genome sequences have been obtained using Sanger sequencing technology. During the last few years, next-generation sequencing techniques have superseded Sanger sequencing for large-scale sequencing projects because of significant improvement in throughput and cost-efficiency[7]. Consequently, the number of fungal genome-sequencing projects has dramatically increased over the last few years. The genomes of several basidiomycetes species have been released by the Fungal Genomics Program at the Joint Genome Institute[8], including those for wood-decaying fungi (*Phanerochaete chrysosporium*[9], *Trametes versicolor*[10], *Dichomitus squalens*[10], *Serpula lacrymans*[11], *and Postia placenta*[12]), inedible mushroom-forming fungi (*Coprinopsis cinerea*[13], *Schizophyllum commune*[14], and *Laccaria bicolor*[15]), and cultivated mushroom-forming fungi (*Agaricus bisporus and Pleurotus ostreatus*). The availability of these genome sequences not only provides baseline genomic information, but can also be utilized to understand the genomic potential of other species through comparative studies of genome structure, gene evolution, and metabolic and regulatory pathways. The expansion of genome sequence information motivated us to sequence the genome of *G. frondosa*.

Here, we present the whole-genome *de novo* sequence of *G. frondosa* sequenced using Roche 454 and Illumina platforms. Roche 454 sequencing has been the most widely used platform for *de novo* genome sequencing mainly because of it can produce read lengths comparable to or exceeding those obtained with Sanger sequencing. Illumina sequencing platforms, on the other hand, have been used in resequencing applications because numerous short reads can be generated at relatively low cost. We determined optimal assembly methods using 454 and Illumina read data to improve accuracy and cost-efficiency. Gene prediction was performed using transcriptome data obtained from a previous study[6] to improve the accuracy of gene prediction. To evaluate the genome sequence data of *G. frondosa*, functional annotation of predicted gene models was performed to target the lignocellulose degradation enzymes involved in the spawn run stage that is an important step in mushroom production. This genomic information helps elucidate the molecular mechanisms of mushroom development, including systems for extracellular digestion of complex polymers in natural substrates and fruiting body differentiation.

(2)Materials and Methods

Strain and culture conditions

G. frondosa Gf-M1, a monokaryotic strain isolated from dikaryotic strain M51, was used as DNA donor for sequencing. The strain was routinely maintained in potato dextrose agar (Difco) in darkness at 25℃. For extraction of genomic DNA, mycelia were cultured in GPY liquid medium containing 2% glucose, 0.2% polypeptone, 0.2% Bacto yeast extract (Difco), 0.05% magnesium sulfate, and 0.05% potassium dihydrogenphosphate at 25℃, stationary for 2 weeks in the dark.

Preparation of genomic DNA for sequencing

Vegetative mycelium was harvested and then ground to a fine powder under liquid nitrogen using a mortar. For genomic DNA isolation, homogenized mycelial powder was extracted with ISOPLANT II (NIPPON GENE, Tokyo, Japan), according to the manufacturer's instructions. Quantity and quality of genomic DNA was analyzed using PicoGreen assay (Molecular Probes, Eugene, OR) and NanoDrop spectrophotometer (Thermo Fisher Scientific, Wilmington, DE).

Genome sequencing and assembly

De novo genome sequencing of *G. frondosa* was performed using the Roche 454 GS FLX Titanium platform and the Illumina Genome Analyzer IIx (GAIIx) platform. All library preparations and sequencing followed the manufacturer's instructions. For 454 sequencing, a fragment library and a mate-pair library with an insert size of ～3 kb were

prepared, from which a total of 2.6 million reads were obtained. For Illumina sequencing, a paired-end library with a fragment size of ～500 bp was prepared. The sequences obtained were 76 bases long and 117 million reads passed the filtering process. Illumina reads were then subjected to a custom algorithm to further refine the data quality. Briefly, trimming and filtering was carried out in such a way that low quality stretches on the 3' ends were trimmed off, and those read pairs with an average Phred score lower than 15 or length shorter than 35 bp were fully discarded.

Combining 454 and Illumina reads, assembly was carried out in the following multiple ways: a) assembly of 454 reads alone using Newbler v2.5.3 and Velvet[16]; b) hybrid assembly of 454 and Illumina reads using Celera assembler[17], Velvet, and Newbler v.2.6; c) assembly of 454 reads by Newbler v2.5.3 and Illumina reads by Velvet independently, followed by assembling the resulting contigs using AMOS[18] and Velvet; and d) assembling 454 reads, followed by refinement of the resulting scaffolds using Illumina reads (custom approach). The custom assembly approach is depicted in Figure 1. Briefly, Illumina reads were mapped against the scaffolds using BWA[19] to detect and correct consensus base-call errors, then subsequently to close intra-scaffold gaps.

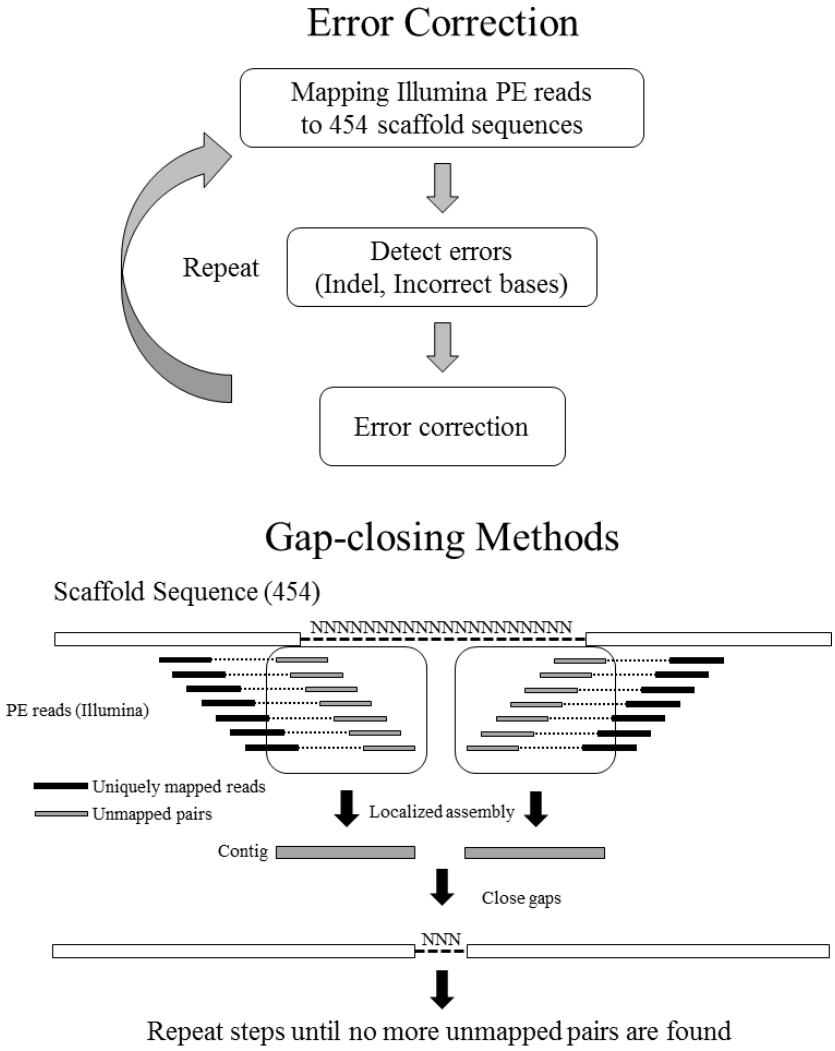

Fig. 1 Custom assembly approach (d) overview.
Illumina paired-end (PE) reads were mapped against the 454 scaffold sequences using BWA to detect and correct consensus base-call errors and to subsequently close intra-scaffold gaps.

Gene prediction

Gene models in the genome of *G. frondosa* were predicted using AUGUSTUS[20]. For gene prediction using AUGUSTUS, pre-trained parameter sets for *L. bicolor* bundled in the software distribution were applied to the draft genome sequence of *G. frondosa* without providing any hints for gene prediction. To accomplish more accurate gene prediction, 534,290 EST sequences obtained from 13 different developmental stages[6] of *G. frondosa* were then utilized as hints for gene prediction by AUGUSTUS.

Functional annotation

All predicted gene models were annotated using the annotation pipeline shown in Figure 2. First, automatic basic annotation was carried out using BLAST searches (NCBI, NIH, Bethesda, MD), GO mapping (Institution, Location), Enzyme Commission (EC) (Institution, Location) number mapping, and InterProScan (EMBL, Heidelberg, Germany) using Blast2GO V.2.5.1[21]. BLAST search parameters included an e-value $\leq 1 \times 10^{-3}$ and a bit score > 40 against the NCBI non-redundant protein sequence database. Then, all annotated genes were assigned with CAZy[22] and FOLy[23] clan number using automatically annotation results as hints.

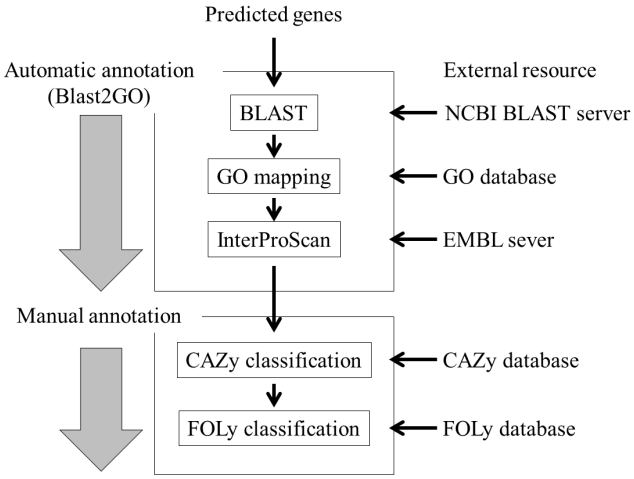

Fig. 2 The annotation pipeline.

Phylogenetic analysis

The 15 putative peroxidase sequences from *G. frondosa* and 66 complete peroxidase sequences that were present in the NCBI database or the Joint Genome Institute comprised the dataset. Phylogenetic analyses were conducted in MEGA5[24], and the evolutionary history of each gene was inferred using the UPGMA method[25]. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. Evolutionary distances were computed using the Poisson correction method[26] and are expressed as the number of amino acid substitutions per site. All positions containing gaps and missing data were eliminated.

(3)Results

Genome sequencing and assembly

We sequenced the genome of *G. frondosa* monokaryotic strain Gf-M1 using a combination of Roche 454 GS FLX Titanium and Illumina GAIIx. The 454 platform generated fragment reads of 458 mega-base-pair (Mbp) of sequence data (14× coverage) and mate-pair reads of 336 Mbp of sequence data (10× coverage). The Illumina paired-end reads of 7 Gbp (78.3%) of sequence data (205× coverage) survived the trimming/filtering process (data not shown). In total 7.8 billion bases were used for downstream analysis, giving a total coverage of about 230 genome equivalents. Combining 454 and Illumina data in *de novo* assembly is a promising approach because both technologies can complement platform-dependent systematic errors or coverage bias. However, there has been no bioinformatics pipeline established to date for assembling combined 454 and Illumina *de novo* sequence data. In this study, various strategies for whole-genome assembly were tested using 454 and Illumina data. Initially, we assembled genome sequence using only 454 or only Illumina data, respectively (Table 1 a). The assembly generated from 454 data alone yielded a total contig number of 1,884. In contrast, the assembly generated from Illumina data alone yielded a total contig number of 5,533. The results suggest that the longer read lengths from the 454 approach can facilitate the assembly of genomes in the absence of a reference genome, when compared with the Illumina short reads. Subsequently, we examined assembly methods including a hybrid reads approach using both 454 and Illumina reads, and a hybrid contigs approach using the contigs assembled from 454 and Illumina reads separately (Table 1 b, c).

Table 1 Basic statistics of each assembly strategy and assembler combination

Assembly strategy	(a)			(b)		(c)		(d)
Sequence data	I 454	II illumina	III 454+Illumina	IV 454+Illumina	V 454+Illumina	VI I + II	VII I + II	VIII 454+Illumina
Assembler	Newbler v2.5	Velvet	Celera	Velvet	Newbler v2.6	AMOS	Velvet	custom
Contig* total	1,884	5,533	N/A	5,228	2,215	6,665	11,104	1,186
Scaffold total	280	N/A	482	N/A	358	N/A	N/A	280
N50 scaffold length (bp)	311,974	94,007	220,935	157,841	230,810	173,887	32,605	312,243
Sum. gap length (bp)	647,199	171,790	8,373	1,659,751	743,542	310,229	0	393,664

* Contigs in this table refer to all scaffolds and non-scaffolded contigs combined.
Assembly (a) was generated from only 454 or only Illumina data, respectively (Basic Approach). Assembly (b) was generated from combining 454 and Illumina reads (Hybrid Reads Approach). Assembly (c) was generated from the contigs assembled 454 and Illumina reads separately (Hybrid Contigs Approach). Assembly (d) was generated from assembly of 454 data processed using Illumina data for both error-correction and gap-closure (Custom Approach).

However, no remarkable advantages were observed with these methods, when compared with assembly of 454 data alone. Instead, we then examined a custom approach that used Illumina data for both error correction and gap closure against assembled 454 data (Table 1 d, Fig. 1). The resulting custom approach improved the assembly, decreasing the number of contigs and the sum of gap length. In conclusion, we were able to obtain the best result using the custom approach, whereby 280 scaffolds (N50 = ~312 kb) were assembled in the total genome length of approximately 33.8 Mb. Genome statistics are presented in Table 2.

Gene prediction

Gene model prediction for *G. frondosa* genome was generated by AUGUSTUS. AUGUSTUS can be used as an *ab initio* program, that is, as a program that uses only one single genomic sequence as input information. Furthermore, it is able to combine information from the genomic sequence under study with external hints from various sources of information. Initially, gene prediction was performed by an *ab initio* method using a program based on full gene models from *L. bicolor,* resulting in a total of 10,505 genes predicted in the *G. frondosa* genome. Subsequently, we carried out gene predictions using a collection of *G. frondosa* ESTs from 13 different developmental stages[6] as hints. This approach gener-

Table 2 Genome characteristics of *G. frondosa* and other basidiomycetes

	G. frondosa	*S. commune*[14]	*A. bisporus*[8]	*P. ostreatus*[8]	*L. bicolor*[15]
Genome statistics					
Sequence technology	454+Illumina	Sanger	Sanger	454+Sanger	Sanger
Genome size (Mb)	33.8	38.5	30	35.6	64.9
Scaffold total	280	36	29	572	665
Contig total	1,186	352	254	3,272	4,398
Estimated depth	×229	×8.29	×8.5	×21.14	×25
Gene statistics					
Total gene models	16,097	13,210	10,438	12,206	23,132
Average gene length (bp)	1,453	1,795	1,764	1,692	1,549
Average protein length (aa)	377	447.8	425.89	413	356
Exons per gene	5.07	5.7	6.05	6.1	5.28
Exon length (bp)	224.1	249.3	234.22	215	220
Intron length (bp)	77.1	79	70.96	77	92

Table 3 Comparision of Gene prediction accuracy

Prediction program	Hints	Predicted gene number	Avg. mRNA length (bp)	Gene hit numbers in public data
AUGUSTUS 2.5.5	No Hints	10,505	2207.6	5,813
AUGUSTUS 2.5.5	EST data	16,097	1195.9	6,310

The first line shows results using the *ab initio* program of AUGUSTUS. The second line shows results from a collection of *G. frondosa* ESTs from 13 developmental stages as hints.

ated a set of 16,097 predicted genes, approximately 1.5-fold that of the *ab initio* method (Table 3). These results suggest that the use of EST data as an additional source of information dramatically improved the accuracy of gene model prediction.

Annotation

The result of the Blast2GO analysis of 16,097 total predicted gene models showed that 10,942 genes (68%) had BLAST hits. The greatest number of top BLAST hits were against *T. versicolor* (4,757 top hits) and *D. squalens* (3,084 top hits) (Fig. 3).

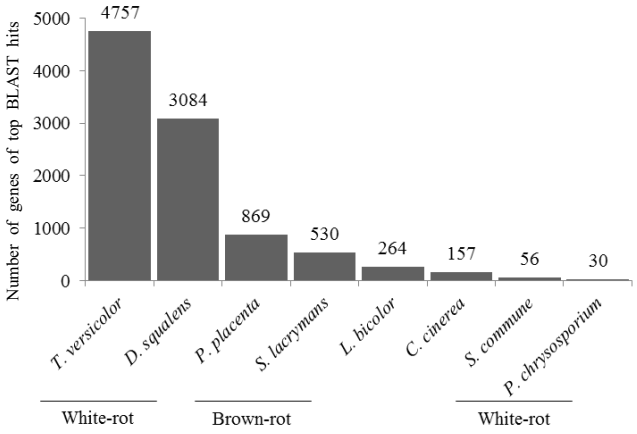

Fig. 3 Top-Hits species distribution.

Carbohydrate-active enzyme

Several recent studies have demonstrated a strong relationship between the repertoire of carbohydrate-active enzymes (CAZymes) in fungal genomes and their saprophytic lifestyle. A total of 324 *G. frondosa* genes were functionally annotated as belonging to CAZymes families defined by the CAZy database[22]. These putative CAZy genes contain 211 candidate glycoside hydrolases (GH), 69 candidate glycosyltransferases (GT), 25 candidate carbohydrate esterases (CE), and 8 candidate polysaccharide lyases (PL). Comparison of the number of GH families related to the degradation of plant cell wall polysaccharides (cellulose, hemicellulose, xyloglucan, and pectin) are listed in Table 4. The number of GH families in *G. frondosa* is close to that in other white-rot species, especially in *T. versicolor* and *D. squalens*. The families GH5 with CBM1, GH6, and GH7, which include cellulose and cellobiohydrolases involved in cellulose degradation, had similar patterns to the number of homologs in the white-rot species, but *S. lacrymans, P. placenta, A. bisporus*, and *L. bicolor* lacked some or all homologs. *G. frondosa* had the lack of family GH11 and numerous members of family GH28 in common with *T. versicolor, D. squalens*, and the brown-rot species. In addition, *G. frondosa* had similar pattern for the number of members of GH61, which are of unknown function, but were recently found to influence the activity of cellulases on lignocellulose[27] in white-rot species.

Fungal oxidative lignin enzymes

White-rot fungi, like *G. frondosa*, are known to degrade lignin, which is a complex aromatic polymer that comprises wood and the lignified elements of plants. White-rot fungi catalyze the initial depolymerization of lignin by secreting an array of oxidases and peroxidases that generate highly reactive and nonspecific free radicals, which in turn undergo a complex series of spontaneous cleavage reactions. The fungal enzymes involved in lignin degradation are classified as FOLymes[23]. FOLymes are divided into two categories: lignin oxidases (LO families) and lignin-degrading auxiliary enzymes that generate H_2O_2 for peroxidases (LDA families). Annotation of the gene model in *G. frondosa* genome showed 48 candidate FOLyme genes, including nine laccase genes (LO1), 17 peroxidase genes (LO2), two cellobiose dehydrogenase genes (LO3), five aryl-alcohol oxidase genes (LDA1), nine glyoxal oxidase genes (LDA3), a glucose oxidase gene (LDA6), a

Table 4 Comparision of candidate CAZyems of G. frondosa with other basidiomycetes

CAZymes classification	Target carbohydrate	White rot					Brown rot		Leaf litter		Saprophytic	Ectomycorrhizal
		G. frondosa (Polyporales)	*T. versicolor*[10] (Polyporales)	*D. squalens*[10] (Polyporales)	*P. chrysosporium*[10] (Polyporales)	*S. commune*[14] (Agaricales)	*S. lacrymans*[11] (Boletales)	*P. placenta*[12] (Polyporales)	*C. cinerea*[15] (Agaricales)	*A. bisporus*[8] (Agaricales)	*P. ostreatus*[8] (Agaricales)	*L. bicolor*[15] (Agaricales)
GH 5 (CBM 1)	Cellulose	4	4	4	4	2	3	0	1	3	3	1
GH 6	Cellulose	1	1	1	1	1	1	0	5	0	3	0
GH 7	Cellulose	3	4	4	9	2	0	0	7	0	16	0
GH 10	Hemicellulose	5	6	5	6	5	1	3	5	2	3	0
GH 11	Hemicellulose	0	0	0	1	1	0	0	6	2	2	6
GH 28	Pectin	7	11	7	4	3	7	11	3	4	4	0
GH 43	Pectin, hemicellulose	3	3	7	4	19	2	1	4	2	3	0
GH 51	Pectin, hemicellulose	2	2	2	2	2	1	1	1	0	1	0
GH 61	Plant cell wall (uncertain target)	13	18	15	14	22	5	2	33	5	7	8
GH 74	Xyloglucan	1	1	1	4	1	1	0	1	1	3	0

GH, total number glycoside hydrolase modules; GH#, modules within individual glycoside hydrolase families; GH 5 (CBM 1), glycoside hydrolase family 5 modules associated with family 1 carbohydrate-binding module.

Table 5 Comparision of candidate FOLymes of G. frondosa with other basidiomycetes

	FOLyme classification	White rot					Brown rot		Leaf litter		Saprophytic
		G. frondosa (Polyporales)	*T. versicolor*[10] (Polyporales)	*D. squalens*[10] (Polyporales)	*P. chrysosporium*[9] (Polyporales)	*S. commune*[14] (Agaricales)	*S. lacrymans*[11] (Boletales)	*P. placenta*[11] (Polyporales)	*C. cinerea*[15] (Agaricales)	*A. bisporus*[8] (Agaricales)	*P. ostreatus*[8] (Agaricales)
LO 1	Laccase and other related multicopper	9	10	13	0	6	6	8	17	12	6
LO 2	peroxidase	17	26	12	16	0	0	1	1	2	7
LO 3	Cellobiose dehydrogenase	2	1	1	1	1	2	0	1	1	1
	Subtotal	28	37	26	17	7	8	9	19	15	14
LDA 1	Aryl-alcohol oxidase	5	1	1	3	2	6	8	18	1	12
LDA 3	Glyoxal oxidase	9	5	5	1	2	3	5	0	0	0
LDA 4	Pyranose oxidase	1	2	0	1	1	0	0	0	0	0
LDA 6	Glucose oxidase	1	0	0	1	7	0	0	1	0	0
LDA 7	Benzoquinone reductase	1	1	1	4	4	2	2	2	0	0
LDA 8	Alcohol oxidase	3	4	4	0	4	1	2	0	2	1
	Subtotal	20	13	11	10	20	12	17	21	3	13
	Total	48	50	37	27	27	20	26	40	18	27

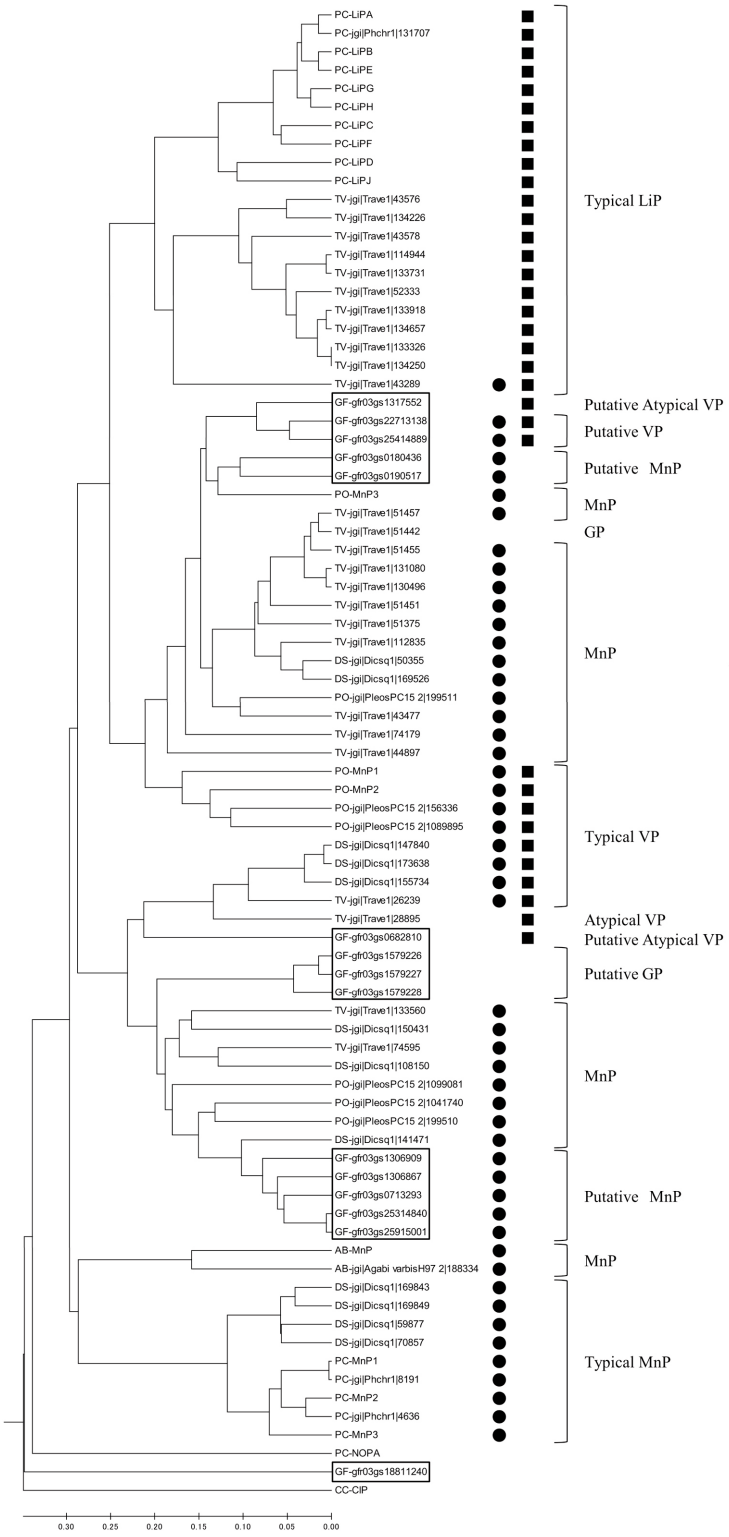

Fig.4 Evolutionary relationships and structural-functional classification of basidiomy-cete peroxidases.
The evolutionary analysis was performed with MEGA 5 using Poisson-corrected evolutionary distances and the UPGMA clustering method. The structure/function classification was based on the presence in the protein sequence or molecular structure of the sequence or structural motifs characteristic of the above peroxidases. The Mn(II)-binding site is composed of three acidic residues (circles) and exposed catalytic tryptophan residue (squares). As a result of the structural analysis, seven putative MnP, two putative VP, two putative atypical VP, and three putative GP were identified in G. frondosa. Species abbreviations: AB, A. bisporus; CC, C. cinerea; DS, D. squalens; GF, G. frondosa; PC, P. chrysosporium; PO, P. ostreatus; and TV, T. versicolor. Protein sequence entries: AB-MnP; CAG 27835, CC-CIP; CAA 50060, PC-LiPA; AAA 53109, PC-LiPB; AAA 33741, PC-LiPC; AAA 33739, PC-LiPD; CAA 33621, PC-LiPE; AAA 33738, PC-LiPF; AAA 33736, PC-LiPG; ABT 17199, PC-LiPH; AAA 56852, PC-LiPJ; AAD 46494, PC-MnP 1; AAA 33744, PC-MnP 2; AAA 33745, PC-MnP 3; AAB 39652, PC-NOPA; AAU 82081, PO-MnP 1; AAA 84396, PO-MnP 2; CAB 51617, PO-MnP 3; BAA 33449.

benzoquinone reductase gene (LDA 7), and three alcohol oxidase genes (LDA 8) (Table 5). The FOLyme component of *G. frondosa* is apparently different from that of the brown-rot species, which lack peroxidase genes belonging to the LO 2 family. These components of *G. frondosa* also differed from those of white-rot fungi such as *P. chrysosporium* and *S. commune*, which lack LO 1 and LO 2, respectively. In particular, *G. frondosa* has a large number of peroxidase genes and glyoxal oxidase genes belonging to the LO 2 and LDA 3 families, respectively. The FOLyme repertoires of *G. frondosa* are similar to those of *T. versicolor* and *D. squalens*. Because peroxide and free radicals generated by peroxidases are considered key components of ligninolysis, we have focused on LO 2 family. These peroxidases are classified into four groups based on structure/function considerations. Lignin peroxidase proteins (LiP) are defined as possessing an exposed redox-active tryptophan residue homologous to the Trp^{171} residue in *P. chrysosporium* LiP-H 8 (encoded by lipA)[28]. Manganese peroxidase proteins (MnP) are defined as possessing a Mn(II)-binding site composed of three residues homologous to the Glu^{35}, Glu^{39}, and Asp^{179} residues of *P. chrysosporium* MnP 1[29]. Versatile peroxidase proteins (VP) are defined as possessing both the Trp^{171} homolog and Mn(II)-binding residues[30]. There are also low redox-potential peroxidases that lack both of the two catalytic sites mentioned above that are defined as generic peroxidases (GP)[31].

To establish the evolutionary relationships among *G. frondosa* peroxidases, their deduced protein sequences were compared with those of the other peroxidases from basidiomycetes. This evolutionary analysis was performed by MEGA 5[24] using Poisson-corrected evolutionary distances and the UPGMA clustering method. The structure/function-based classification was based on the presence of sequence or structural motifs characteristic of the peroxidase groups above in the protein sequence or molecular structure. The dendrogram showed a well-defined cluster corresponding to a typical MnP (from *P. chrysosporium* and *D. squalens*) and a typical LiP (from *P. chrysosporium* and *T. versicolor*). Most candidate peroxidases from *G. frondosa* were located in a heterogenous cluster containing MnP and typical VP from *P. ostreatus, T. versicolor,* and *D. squalens* (Fig. 4). The seven *G. frondosa* peroxidases possessing only the Mn(II)-binding site were classified as putative MnP. The two peroxidases possessing both types of catalytic sites were classified as putative VP, whereas two gene models were classified as atypical VP that lacked 1 or 2 acidic residues at the Mn(II)-binding site. Three models lacking both catalytic sites were classified as putative GP. The model of the remaining (gfr 03 gs 18811240) was not grouped as well as CIP[32] and NOPA[33], respectively, from *C. cinerea* and *P. chrysosporium*. On the other hand, LiP genes identified in both *P. chrysosporium* and *T. versicolor* were not found in the *G. frondosa* genome.

(4) Discussion

G. frondosa is a well-known edible basidiomycete mushroom that can be commercially produced by artificial cultivation systems and that possesses various potential biological activities including anti-tumor and other therapeutic properties. However, our understanding of the molecular biological characteristics of *G. frondosa* is still very limited. Here, we present a high-quality draft genome sequence for *G. frondosa* generated solely by next-generation sequencing combining Roche 454 GS FLX Titanium and Illumina GAIIx platforms. Further, we showed that combining 454 and Illumina data in de novo assemblies is useful for improving accuracy and cost-efficiency.

To determine the best way to assemble a fungal genome *de novo* using only next generation sequencing data, we tested various assembly strategies and different assembly software combinations using 454 and Illumina data. We were able to obtain the best results with the custom approach using Illumina data for both error-correction and gap-closure against assembled 454 data. The custom approach resulted in a drastic reduction in the number of contigs and the sum of gap length compared to assemblies developed from a single type of data set. The 454 and Illumina sequencing methods both have shortcomings which make it difficult to perform *de novo* assembling alone. The 454 sequencing system offers superior read lengths and mate-pair information that is useful for generating assemblies with longer contigs and larger scaffolds. However, 454 reads are prone to yielding indel error accumulation in homopolymers, which is

considered a major technical problem of 454 pyrosequencing technology[34),35)]. On the other hand, with Illumina sequencing, these errors due to homopolymeric regions is rare[36)] and provide notably accurate consensus sequence generation. Illumina sequencing also generates a huge number of short reads that are sufficient for covering the entire region of the genome, whereas assemblies of Illumina reads are more fragmented than those of 454 reads. Therefore, this study suggests that our custom approach has led to good results by taking advantage of the strengths of other sequencing technologies and that it outperforms other typically used methods.

Gene model prediction for the *G. frondosa* genome was performed using AUGUSTUS, because it can be used not only as an *ab initio* program using only one single genomic sequence as input information, but can also combine information from the genomic sequence under study with external hints from various sources of information. To determine the best parameters for gene prediction in the *G. frondosa* genome, we tested both an *ab initio* program and also the addition of extrinsic gene expression data for *G. frondosa*[6)]. As a result, the number of genes predicted increased to approximately 1.5-fold that of the *ab initio* method, when ESTs of *G. frondosa* were included as external hints (Table 3). Gene prediction in eukaryotic organisms is more difficult than in prokaryotic genomes, because of low gene density due to the presence of introns in coding regions. In particular, genes in fungi exhibit significant variation in exon-intron structure[37)]. Furthermore, programs designed for recognizing intron/exon boundaries for a particular organism or group of organisms may not recognize all intron/exon boundaries. Thus, EST-based gene prediction is an effective approach for *de novo* whole-genome sequencing, because the performance of *ab initio* gene-finding algorithms depends greatly on which species were used to generate gene modeling parameters.

White-rot fungi are known to degrade all woody cell wall components. In contrast, brown-rot fungi efficiently degrade cellulose but only modify lignin, leaving a polymeric residue. *G. frondosa* is white-rot fungus, and analysis of its genome sequence revealed a large assortment of genes potentially involved in degradation of wood and other plant cell-wall polysaccharides. In general, these fungi can degrade wood and forest litter, and fully mineralize all cell-wall polysaccharides composed of cellulose, hemicellulose, and pectin. For efficient degradation of these polysaccharides, these fungi produce an extensive set of carbohydrate-active enzymes (CAZymes). Initially, we focused in particular on those CAZymes involved in plant cell-wall polysaccharide degradation. The genome of *G. frondosa* encodes at least one gene from each gene family involved in the degradation of cellulose, hemicellulose, xyloglucan, and pectin. *G. frondosa* also showed many similarities to *T. versicolor* and *D. squalens* in the number and type of putative genes. These results suggest a correlation between the habitat of each species and the number of CAZymes encoded by their genomes. In recent studies, *G. frondosa* utilized oak sawdust plus corn bran as a substrate for growth and fruiting body development, producing enzymes associated with degradation of these materials, including endoglucanase (probably including members of the GH 5 and GH 7 families), exoglucanase (probably including members of the GH 6 and GH 7 families), β-glucosidase (probably including members of the GH 1 and GH 3 families), and endoxylanase (probably including members of the GH 10 and GH 11 families)[38)]. In particular, these enzyme activities increased during the colonization phase, reaching a peak when the substrate was fully colonized at early spawn run (20-30 d after inoculation). Thus, information about candidate CAZymes annotated from the *G. frondosa* genome could improve our understanding of the parameters affecting growth and fruiting and help to improve culture conditions for commercial cultivation. In addition, the *G. frondosa* genome is rich in members of the glycosyl hydrolase family GH 28. The GH 28 family enzymes include endo- and exo-polygalacturonase that catalyze the hydrolysis of pectin, a complex carbohydrate present in the cell walls and middle lamella of plants[39)]. *G. frondosa* possesses a large number of genes associated GH 28 family, suggesting the possibility that *G. frondosa* can utilize a variety of plant tissues as substrates for vegetative mycelium growth. In general, mushroom production has focused on developing more efficient substrate formulas to improve yield and quality, and shorten cultivation period. The information discovered here will provide

a useful starting point for exploration of more suitable and cost-effective mushroom culture substrates. Fungal polygalacturonase enzymes are also utilized in industrial food processes, such as fruit juice clarification and baby food preparation[40]. Thus, candidate enzymes from the GH28 family in *G. frondosa* may have potential for biotechnological applications, although much more work needs to be done to understand the characteristics of these enzymes.

Unlike the digestion of polysaccharides, lignin degradation results from the concerted action of several oxidoreductases such as laccases, ligninolytic peroxidases, and peroxide-generating oxidases[41]. *G. frondosa* possesses at least one gene in each FOL family involved in the degradation of lignin. Compared with model white-rot fungi, such as *P. chrysosporium* and *S. commune, G. frondosa* possesses a complete set of ligninolytic peroxidases including those from the LO1, LO2, and LO3 families. In particular, *G. frondosa* has a large number of peroxidase genes and glyoxal oxidase genes belonging, respectively, to the LO2 and LDA3 families. Enzymatic studies of *P. chrysosporium* have demonstrated physiological connections between peroxidases in the LO2 family and glyoxal oxidase from the LDA3 family[42]. Thus, the presence of genes coding for these enzymes in its genome suggests that *G. frondosa* probably degrades lignin using highly reactive and nonspecific free radicals generated by peroxidases and extracellular H_2O_2 generated by glyoxal oxidase. Because members of the LO2 family are considered key components of ligninolysis in *G. frondosa*, we performed structure/function classification based upon the presence of structural motifs characteristic of peroxidases. As a result, 15 candidate peroxidase genes were classified as seven putative MnP, two putative VP, two putative atypical VP, three putative GP, and several that remained unclassified. In our previous study, microarray experiments performed for all eight stages of the *G. frondosa* cultivation process showed that putative MnP were specifically expressed during spawn run, whereas atypical VP and GP exhibited no significant changes in expression level throughout the cultivation process[6]. These results suggest that MnP may play a key role in lignin digestion in *G. frondosa*. On the other hand, LiP genes identified in both *P. chrysosporium* and *T. versicolor* were not found in the *G. frondosa* genome. Montoya et al. reported that qualitative assays for LiP revealed no activity during solid state fermentation in *G. frondosa*[37], a result consistent with the absence of LiP genes in the present study.

In conclusion, the genome sequence of *G. frondosa* constitutes a notable contribution to our knowledge base of the molecular mechanisms of mushroom development, including systems for extracellular digestion of natural substrates and fruiting body differentiation. The discovery of numerous lignocellulose degradation enzymes will not only help to improve culture conditions but also contribute to the development of technologies for biomass conversion to reduce carbon emissions. In the future, functional analyses using RNA interference (RNAi), sequence-specific knockdown methods, and heterologously expressed proteins will further characterize candidate genes identified in this study to elucidate mushroom developmental mechanisms.

(5) Acknowledgment

This work was performed mainly by a funding from Yukiguni Maitake Co. Ltd. and was supported partly a grant from the Graduate School of Tokyo Kasei University.

(6) References

1) Masuda Y, Matsumoto A, Toida T, Oikawa T, Ito K, and Nnaba H. *J. Agric. Food Chem.*, 57, 10143-10149 (2009)

2) Obi N, Hayashi K, Miyahara T, Shimada Y, Terasawa K, Watanabe M, Takeyama M, Obi R, and Ochiai H. *Am. J. Chin. Med.*, 36, 1171-1183 (2008)

3) Hong L, Xun M, and Wutong W. *J. Pharm. Pharmacol.*, 59, 575-582 (2007)

4) Yeh JY, Hsieh LH, Wu KT, and Tsai CF. *Molecules.*, 16, 3197-3211 (2011)

5) Deng G, Lin H, Seidman A, Fornier M, D'Andrea G, Wesa K, Yeung S, Cunningham-Rundles S, Vickers AJ, and Cassileth B. *J. Cancer Res. Clin. Oncol.*, 135, 1215-1221 (2009)

6) Kurahashi A, Fujimori F, and Nishibori K. *The Bulletin of Tokyo Kasei University.*, 52, 17-32 (2012)

7） Shendure J and Li H. *Nat. Biotechnol.*, 26, 1135-1145 (2008)

8） Grigoriev IV, Nordberg H, Shabalov I, Aerts A, Cantor M, Goodstein D, Kuo A, Minovitsky S, Nikitin R, Ohm RA, Otillar R, Poliakov A, Ratnere I, Riley R, Smirnova T, Rokhsar D, and Dubchak I. *Nucleic Acids Res.*, 40, D26-32 (2012)

9） Martinez D, Larrondo LF, Putnam N, Gelpke MD, Huang K, Chapman J, Helfenbein KG, Ramaiya P, Detter JC, Larimer F, Coutinho PM, Henrissat B, Berka R, Cullen D, and Rokhsar D. *Nat. Biotechnol.*, 22, 695-700 (2004)

10） Floudas D, Binder M, Riley R, Barry K, Blanchette RA, Henrissat B, Martinez AT, Otillar R, Spatafora JW, Yadav JS, Aerts A, Benoit I, Boyd A, Carlson A, Copeland A, Coutinho PM, de Vries RP, Ferreira P, Findley K, Foster B, Gaskell J, Glotzer D, Górecki P, Heitman J, Hesse C, Hori C, Igarashi K, Jurgens JA, Kallen N, Kersten P, Kohler A, Kues U, Kumar TK, Kuo A, LaButti K, Larrondo LF, Lindquist E, Ling A, Lombard V, Lucas S, Lundell T, Martin R, McLaughlin DJ, Morgenstern I, Morin E, Murat C, Nagy LG, Nolan M, Ohm RA, Patyshakuliyeva A, Rokas A, Ruiz-Duenas FJ, Sabat G, Salamov A, Samejima M, Schmutz J, Slot JC, St John F, Stenlid J, Sun H, Sun S, Syed K, Tsang A, Wiebenga A, Young D, Pisabarro A, Eastwood DC, Martin F, Cullen D, Grigoriev IV, and Hibbett DS. *Science.*, 336, 1715-1719 (2012)

11） Eastwood DC, Floudas D, Binder M, Majcherczyk A, Schneider P, Aerts A, Asiegbu FO, Baker SE, Barry K, Bendiksby M, Blumentritt M, Coutinho PM, Cullen D, de Vries RP, Gathman A, Goodell B, Henrissat B, Ihrmark K, Kauserud H, Kohler A, LaButti K, Lapidus A, Lavin JL, Lee YH, Lindquist E, Lilly W, Lucas S, Morin E, Murat C, Oguiza JA, Park J, Pisabarro AG, Riley R, Rosling A, Salamov A, Schmidt O, Schmutz J, Skrede I, Stenlid J, Wiebenga A, Xie X, Kues U, Hibbett DS, Hoffmeister D, Hogberg N, Martin F, Grigoriev IV, and Watkinson SC. *Science.*, 333, 762-765 (2011)

12） Martinez D, Challacombe J, Morgenstern I, Hibbett D, Schmoll M, Kubicek CP, Ferreira P, Ruiz-Duenas FJ, Martinez AT, Kersten P, Hammel KE, Vanden Wymelenberg A, Gaskell J, Lindquist E, Sabat G, Bondurant SS, Larrondo LF, Canessa P, Vicuna R, Yadav J, Doddapaneni H, Subramanian V, Pisabarro AG, Lavin JL, Oguiza JA, Master E, Henrissat B, Coutinho PM, Harris P, Magnuson JK, Baker SE, Bruno K, Kenealy W, Hoegger PJ, Kües U, Ramaiya P, Lucas S, Salamov A, Shapiro H, Tu H, Chee CL, Misra M, Xie G, Teter S, Yaver D, James T, Mokrejs M, Pospisek M, Grigoriev IV, Brettin T, Rokhsar D, Berka R, and Cullen D. *Proc. Natl. Acad. Sci. U.S.A.*, 106, 1954-1959 (2009)

13） Stajich JE, Wilke SK, Ahrén D, Au CH, Birren BW, Borodovsky M, Burns C, Canbäck B, Casselton LA, Cheng CK, Deng J, Dietrich FS, Fargo DC, Farman ML, Gathman AC, Goldberg J, Guigo R, Hoegger PJ, Hooker JB, Huggins A, James TY, Kamada T, Kilaru S, Kodira C, Kües U, Kupfer D, Kwan HS, Lomsadze A, Li W, Lilly WW, Ma LJ, Mackey AJ, Manning G, Martin F, Muraguchi H, Natvig DO, Palmerini H, Ramesh MA, Rehmeyer CJ, Roe BA, Shenoy N, Stanke M, Ter-Hovhannisyan V, Tunlid A, Velagapudi R, Vision TJ, Zeng Q, Zolan ME, and Pukkila PJ. *Proc. Natl. Acad. Sci. U S A.*, 29, 11889-11894 (2010)

14） Ohm RA, de Jong JF, Lugones LG, Aerts A, Kothe E, Stajich JE, de Vries RP, Record E, Levasseur A, Baker SE, Bartholomew KA, Coutinho PM, Erdmann S, Fowler TJ, Gathman AC, Lombard V, Henrissat B, Knabe N, Kües U, Lilly WW, Lindquist E, Lucas S, Magnuson JK, Piumi F, Raudaskoski M, Salamov A, Schmutz J, Schwarze FW, vanKuyk PA, Horton JS, Grigoriev IV, and Wösten HA. *Nat. Biotechnol.*, 28, 957-963 (2010)

15） Martin F, Aerts A, Ahren D, Brun A, Danchin EG, Duchaussoy F, Gibon J, Kohler A, Lindquist E, Pereda V, Salamov A, Shapiro HJ, Wuyts J, Blaudez D, Buée M, Brokstein P, Canbäck B, Cohen D, Courty PE, Coutinho PM, Delaruelle C, Detter JC, Deveau A, DiFazio S, Duplessis S, Fraissinet-Tachet L, Lucic E, Frey-Klett P, Fourrey C, Feussner I, Gay G, Grimwood J, Hoegger PJ, Jain P, Kilaru S, Labbé J, Lin YC, Legué V, Le Tacon F, Marmeisse R, Melayah D, Montanini B, Muratet M, Nehls U, Niculita-

Hirzel H, Oudot-Le Secq MP, Peter M, Quesneville H, Rajashekar B, Reich M, Rouhier N, Schmutz J, Yin T, Chalot M, Henrissat B, Kues U, Lucas S, Van de Peer Y, Podila GK, Polle A, Pukkila PJ, Richardson PM, Rouzé P, Sanders IR, Stajich JE, Tunlid A, Tuskan G, and Grigoriev IV. *Nature*, 452, 88-92 (2008)

16) Zerbino DR and Birney E. *Genome Res.*, 18, 821-829 (2008)

17) Miller JR, Delcher AL, Koren S, Venter E, Walenz BP, Brownley A, Johnson J, Li K, Mobarry C, and Sutton G. *Bioinformatics.*, 24, 2818-2824 (2008)

18) Treangen TJ, Sommer DD, Angly FE, Koren S, and Pop M. *Curr Protoc Bioinformatics.*, Chapter 11:Unit 11.8. (2011)

19) Li H and Durbin R. *Bioinformatics.*, 25, 1754-1760 (2009)

20) Stanke M, and Waack S. *Bioinformatics*, 19, 215-225 (2003)

21) Conesa A, Götz S, Garcia-Gómez JM, Terol J, Talón M, and Robles M. *Bioinformatics*, 21, 3674-3676 (2005)

22) Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, and Henrissat B. *Nucleic Acids Res.*, 37, D233-8 (2009)

23) Levasseur A, Piumi F, Coutinho PM, Rancurel C, Asther M, Delattre M, Henrissat B, Pontarotti P, Asther M, and Record E. *Fungal Genet. Biol.*, 45, 638-645 (2008)

24) Tamura K, Peterson D, Peterson N, Stecher G, Nei M, and Kumar S. *Mol. Biol. Evol.*, 28, 2731-2739 (2011)

25) Sneath, P.H.A. and Sokal, R.R. *Numerical Taxonomy.*, 230-234 (1973)

26) Zuckerkandl E. and Pauling L. *Edited in Evolving Genes and Proteins by V. Bryson and H.J. Vogel.*, 97-166 (1965)

27) Harris PV, Welner D, McFarland KC, Re E, Navarro Poulsen JC, Brown K, Salbo R, Ding H, Vlasenko E, Merino S, Xu F, Cherry J, Larsen S, and Lo Leggio L. *Biochemistry.*, 49, 33005-3316 (2010)

28) Blodig W, Smith AT, Doyle WA, and Piontek K. *J. Mol. Biol.*, 305, 851-861 (2001)

29) Sundaramoorthy M, Kishi K, Gold MH, and Poulos TL. *J. Biol. Chem.*, 272, 17574-17580 (1997)

30) Camarero S, Sarkar S, Ruiz-Dueñas FJ, Martinez MJ, and Martinez AT. *J. Biol. Chem.*, 274, 10324-10330 (1999)

31) Fernändez-Fueyo E, Ruiz-Duenas FJ, Miki Y, Martinez MJ, Hammel KE, and Martinez AT. *J. Biol. Chem.*, 287, 16903-16916 (2012)

32) Baunsgaard L, Dalboge H, Houen G, Rasmussen EM, and Welinder KG. Eur. *J. Biochem.*, 213, 605-611 (1993)

33) Larrondo L, Gonzalez A, Perez Acle T, Cullen D, and Vicuña R. *Biophys. Chem.*, 116, 167-173 (2005)

34) Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, and Rothberg JM. Nature., 437, 376-380 (2005)

35) Huse SM, Huber JA, Morrison HG, Sogin ML, and Welch DM. Genome Biol., 8, R143 (2007)

36) Erlich Y, Mitra PP, delaBastide M, McCombie WR, and Hannon GJ. Nat. Methods., 5, 679-682 (2008)

37) Kupfer DM, Drabenstot SD, Buchanan KL, Lai H, Zhu H, Dyer DW, Roe BA, and Murphy JW. Eukaryotic Cell., 3, 1088-1100 (2004)

38) Montoya S, Orrego CE, and Levin L. *World J. Microbiol. Biotechnol.*, 28, 1533-1541 (2012)

39) Niture SK. Biologia., 63, 1-19 (2008)

40) Lang C and Dornenburg H. *Appl. Microbiol. Biotechnol.*, 53, 366-375 (2000)

41) Ruiz-Dueñas FJ, and Martinez AT. *Microb Biotechnol.*, 2, 164-177 (2009)

42) Kersten PJ., *Proc. Natl. Acad. Sci. U.S.A.*, 87, 2936-2940 (1990)

食用キノコ Grifola frondosa（マイタケ）の高品質ドラフトゲノム解析

論文要旨

マイタケは，人工栽培により大規模に生産されている．しかしながら，その子実体生育のメカニズムについては理解されていない部分が多い．このメカニズムを分子生物学的アプローチにより理解するため，我々は次世代シーケンサーであるロシュ社 454 GS FLX Titanium 及びイルミナ社 Genome Analyzer IIx を用いて，マイタケの高品質なドラフトゲノム配列を取得した．総塩基長 33.8 Mb のゲノム配列から両シーケンスデータを最適に組み合わせたアセンブリ手法により，280 個のスキャホールドが得られた．また，マイタケ栽培工程 13 か所の EST 配列データを予測プログラムの参照配列とすることによって，ゲノム配列から 16,097 個の遺伝子が予測された．予測された遺伝子を機能注釈した結果，木材や多糖類の分解に関わる広範囲の酵素を持つことが明らかになった。これらの情報は，マイタケのリグノセルロース分解に関わる酵素システムの解明に貢献することが期待される．我々の研究は，新規ゲノム配列決定において，454 とイルミナの両データを組み合わせることでアセンブル精度が改善されること，EST 情報を基にした遺伝子予測が予測精度向上のために効果的なアプローチであることを示した．