Functional annotation and pangenomic analysis of the *Leucoagaricus gongylophorus* LEU184964 genome a mutualistic fungus of the ant *Atta mexicana*

Freddy Castillo-Alfonso[a], Gabriela Cejas-Añón[a], José Utrilla Carreri[b], Cecilio Valadez-Cano[c], Juan Carlos Sigala Alanis[d], Silvie LeBorgne[d], Alfonso Mauricio Sales-Cruz[d], Juan Gabriel Vigueras-Ramírez[d] and Roberto Olivares-Hernández[d]

Posgrado en Ciencias Naturales e Ingeniería, Universidad Autónoma Metropolitana, Unidad Cuajimalpa, Av. Vasco de Quiroga 4871, Col. Santa Fe Cuajimalpa, Delegación Cuajimalpa, Ciudad de México, 05348, México, Mexico[a]

Instituto de Biotecnología, Universidad Nacional Autónoma de Mexico, Avenida Universidad 2001, Chamilpa, 6221011 Cuernavaca, Morelos, Mexico[b]

Department of Biology, University of New Brunswick 10 Bailey Drive, Fredericton, New Brunswick E3B 5A3, Canada[c]

Departamento de Procesos y Tecnología, Universidad Autónoma Metropolitana, Unidad Cuajimalpa, Av. Vasco de15Quiroga 4871, Col. Santa Fe Cuajimalpa, Delegación Cuajimalpa, Ciudad de México, 05348, México, Mexico[d]

ABSTRACT

In this work we report the genome assembly of the fungus *Leucoagaricus gongylophorus*, a fungus that establishes a mutualistic relationship with the *Atta mexicana* ant. The assembly is based on the data generated with Illumina and Roche 454 sequencing technologies. In order to evaluate the quality of the genome assembly we performed a single and hybrid assembling processes. The single assembly with Illumina data compile 48141 contigs, and the hybrid assembly, Roche-Illumina, a total of 48287 contigs; with a N50 value of 8573 and 5214, respectively. Using the gene prediction tool Augustus, it was possible to predict 8,666 genes for the single assembly and 11,690 genes for the hybrid assembly. The database used for the functional annotation was the Basidomycetes division of the RefSeq protein database of NCBI. From the protein sequence alignment, 11252 out of the 11690, matched with an E-value of 10E-50 and 3150 proteins have an EC number assigned. As part of the search for enzymes with the capability to hydrolyze biomass we found 391 CAZymes, 52

FOLymes and 38 possible proteases. To identify the biological relationships with other Leucoagaricus, a comparative genomics analysis was carried out using 4 genomes composing the pangenome for these species with a total of 18052 gene clusters, out of this number, 383 gene clusters belong to the coregenome and 17669 to the accessory genome. The principal relationships among these species are the functional GO terms.

IMPORTANCE

This paper presents the first genomic assembly for the basidiomycete fungus *Leucoagaricus gongylophorus* LEU184964, a symbiont of the ant *Atta mexicana*. Combining two sequencing platforms it was possible to obtain a high-precision functional annotation for this organism containing the highest number of genes found for this genus. A great abundance of enzymes of industrial interest such as CAZymes and FOLymes were found for this assembly, allowing to explore the metabolic capabilities for the degradation of lignocellulosic material and its potential use in different bioprocesses.

KEYWORDS: Genome assembly, Gene prediction, Genome functional annotation, Pangenomic analysis .

INTRODUCTION

Leucoagaricus gongylophorus is a basidiomycete that establishes a mutualistic relationship with leafcutter ants belonging to the genera *Atta* and *Acromycex* (1). These ants collect plant material to cultivate the basidiomycete fungus. In this natural condition, the fungus presents a type of filamentous growth, forming a network similar to a white sponge on the substrate and it is responsible for producing the necessary enzymes to degrade plant polymers and release sugars; a fraction of this sugars is used by the fungus to satisfy its energy requirements, the another fraction is accumulated in the form of glycogen within globular structures called gonglydians, which serve as food for the ants (2).

Various groups have worked on the association of *L. gongylophorus* with various species of ants and have reported various aspects of its metabolic capabilities, as well as the sequences of genes, proteins and genomes (3), (4), (5). The number of sequences reported for this organism is low compared to other widely characterized organisms such as *Aspergillus niger*. In the GenBank Database there are registered 24,007 nucleotide

sequences and 143 protein sequences for *L. gongylophorus*, while for *A. niger* there are 85,933 nucleotide sequences and 169,950 protein sequences. The genetic material in eukaryotic systems differ in terms of composition, structure, organization, and complexity, these variations have a direct impact on the assembly of a post-sequencing genome (6). In general, fungal genomes are compact, with high gene densities, low levels of repetitive content and fewer introns. Size variations can be considerable, we can find differences among genomes ranging from 982 Mb (*Wallemia sebi*) to 130.65Mb (*Dendrothele bispora*) (7).

The technological advance and development of different sequencing platforms has drastically reduced the cost of sequencing a genome, due to this phenomenon the number of sequenced genomes has increased considerably (8,9). State-of-the-art sequencing technologies generate blazingly fast, high-throughput, high-quality sequencing data even though they differ across platforms. The criterion to choose a sequencing technology include read lengths, accuracy, price, and the time required to complete a sequencing run. Whole genome assembly projects have typically used a combination of two or more "short-read" and "long-read" genomic libraries. Second generation technologies such as Illumina and Roche 454, generally start with DNA fragmentation, DNA end-repair, adapter ligation, surface attachment, and in situ amplification. These “short-read” sequencing technologies involve the massively parallel sequencing of short reads, whereby millions of individual sequencing reactions occur in parallel (10). *De novo* genome assembly typically involves pairing and combining large numbers of small DNA fragments based on their overlapping regions into contiguous stretches called contigs (11). One of the main objectives of a correct genomic assembly is to generate contigs of the largest possible size (12). Ultralong contigs provide complete and uninterrupted sequence information across entire genes and, more recently, even allow separation of the different chromosomes for diploid and polyploid organisms.

Gene prediction tools operate algorithms to find defined structures within contigs such as introns and exons. Augustus is a gene and protein prediction tool, often referred to as AB

initio gene predictors because they use mathematical models rather than external evidence to identify genes and determine their intron-exon structures (13).

As the number of sequenced and assembled genomes has increased, various analyzes such as pan-genomics have emerged for their comparison and withdraw biological meaning. Various methods such as EUPAN(14), GET HOMOLOGUES (15), and PanVC (16) can be used to generate pangenes from different sets of species. The pangenome is made up of the core genome and the accessory genome (17). The coregenome contains the set of gene clusters conserved in all the analyzed species, generally essential genes for an organism. The accessory genome is formed by of clusters of specific genes that can form isolated sets within an organism generally related to functions that are not specific to that particular species.

The main objective of the present study is to construct the first genome assembly for the *Leucoagaricus gongylophorus* LEU18496 symbiont of the *Atta mexicana* ant and to functionally annotate it using different genomic and pangenomic tools to characterize the genome and metabolism of this symbiont.

RESULTS

The single (Illumina MiSeq) and hybrid (Illumina MiSeq + Roche 454) genome assemblies of *L. gongylophorus* LEU18496 have different metrics (Table 1). The total length of assembled pair bases was higher in the single assembly.

Table 1: Comparison of the single (Illumina MiSeq) and hybrid (Illumina MiSeq + Roche 454) genome assemblies obtained for *L. gongylophorus* LEU18496.

	Single	**Hybrid**	**Aylward, 2013[a]**
Total length	151 379 115	137 005 739	91 322 395
Contigs number	48 141	48 287	58 433
N50	8573	5214	2096
N90	978	1136	699

Mismatches (N's)	798	0	375
GC Content (%)	36.23	36.09	35.03
Predicted genes	8666	11690	5497[b]
Technology	Illumina	Illumina + Roche 454	Roche 454
Assembler	SPAdes v. 3.9.0	SPAdes v. 3.15.0	Newbler v. 2.3

[a]Source is from: https://www.ncbi.nlm.nih.gov/assembly/GCA_000382605.1

[b]Gene prediction performed in this work using the genome assembly reported by Aylward *et al*., 2013.

In Table 1 it can be seen that the value of N50 was 8573 for the single assembly, this is 1.64 times greater than the hybrid assembly, however, N90 for the hybrid assembly is 1.21 times greater. The similar number of total contigs in the single and hybrid genome assemblies is similar, 48141 and 48287, and the same GC content with 36.23% and 36.09% for the Single and Hybrid respectively. The number of mismatches (N's), or total unassigned bases, was 798 pairs in single assembly and for the hybrid assembly there were no mismatches. The added Ns in the found false gaps increases the size of the genome by 798 bp. This is an amount detected by the N's per 100 kpb metric with a value of 0.53. The elimination of mismatches was carried out during the cleaning process as part of its submission to the NCBI platform; the single assembly was not submitted. This depuration removes the possible contamination, although this process can fractionate some contigs and change metrics presented above.

The plot of points obtained shows a continuous solid line that corresponds to a match between the analyzed sequences, identical sequences will obviously have a diagonal line in the center of the matrix (20), 57.53 % of they have an identity greater than 75% (Supplementary material 2 for more details). The difference in the total number of assembled bases and their correct alignment within the contigs can explain part of the differences in the metrics analyzed in Table 1.

Annotation and comparative genomics of genome assemblies of *L. gongylophorus* LEU18496

The number of predicted genes differs between the single and the hybrid genome assemblies, 8666 and 11690, respectively. These numbers are higher than the 5497 those

previously predicted for the assembly reported by Aylward et al., in 2013 for *L. gongylophorus* Ac12.

Sequence homology analysis using BLASTp performed on both sets of predicted genes allowed the identification of 8582 identical genes between both sets (25-1); of this number, 8364 sequences have alignments greater than 80% and an Evalue=10−75 (Supplementary Material 3).

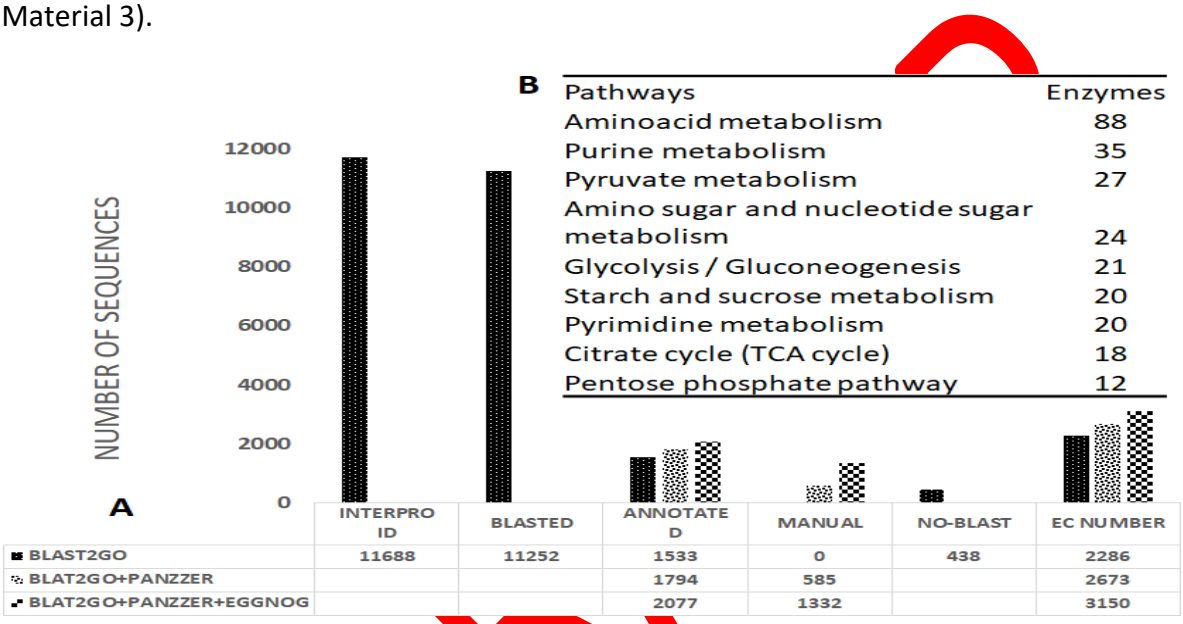


Figure 1: Merged functional annotation for 11690 proteins obtained from hybrid genome assembly by combining different gene annotation tools. B: Best represented metabolic pathways based on relative abundance of enzymes obtained by KEEG pathways.

According to the results shown in Figure 1 (panel A), it was possible to find 11,252 positive alignments with proteins reproduced for the Basidiomycetes family, representing 96.25% of validated genes in the assembly. From this set of proteins, it was possible to assign 2286 EC numbers using Blast2Go. Subsequently, when re-annotating using the functional annotations obtained by PANNZER and EggNog Mapper, it was possible to increase the number of proteins with enzymatic function to 2673 and 3150, respectively. Out of the total of 11,252 proteins found, 3,150 correspond to a metal genome for 26.94%. The best represented KEGG metabolic pathways found during functional annotation include nitrogenous base metabolism, amino acid metabolism. Central carbon metabolism is also widely represented,

fundamentally by pyruvate metabolism and the Glycolysis/Gluconeogenesis pathway (Figure 1, (panel B)).

As part of the description of the metabolism of *L. gongylophorus* LEU18496 made during the functional annotation process, special attention was paid to pathways pertaining to central carbon metabolism. An analysis of metabolic pathways using KEEG Pathways (22) yielded the following results for these pathways according to the number of enzymes found: 27 Pyruvate metabolism, 21 Glycolysis/Gluconeogenesis, 18 Citrate cycle (TCA cycle) and 12 Pentose Phosphate Pathway in scale from largest to smallest represented (Supplementary material 4)

In the case of the functional annotation of enzymes with CAZymes and FOLymes activity, we found 391 CAZymes, 52 FOLymes and 38 possible proteases in the assembled genome of *L. gongylophorus* LEU18496 (Table 2).

TABLE 2: Proteins with CAZymes, FOLymes and Proteases activity found in the hybrid genomic assembly of *L. gongylophorus* LEU18496

Group	Family	Annotation	Enzymes
CAZymes	GH	Hydrolase	54
	GH1	Glucolyase	2
	GH2	Galactosidase	12
	GH3	Glucosidase/Xylosidase	35
	GH5	Cellulases	16
	GH10	Xylanase	2
	GH15	Glucoamilase	6
	GH16	1,3-*β*-glucanase	14
	GH19	Chitinase	13
	GH38	Mannosidase	6
	GH71	α-1,3-glucanase	3
	GH74	Endoglucanase	16
	GT	GlycosylTransferase	31
	GT2	Chitin synthase	22
	GT5	α-1,3-glucan synthase	14
	GT8	Glycogen synthase	4
	GT20	Trehalose-phosphate synthase	5
	GT47	GalactosylTransferase	25
	PL	Polysaccharide Lyases	5
	PL1	Pectine/Pectate lyase	5
	PL4	Rhamnogalacturonan endolyase	3
	CE	Carbohydrate Esterase	3
	CE4	Xylan esterase	6

	CE5	Cutinase	2
	CE16	Acetylesterase	3
	PE	Pectinase	5
	MN	MannosylTransferase	26
	E	Estearase	6
	R	Reductase	2
	AA	Auxiliary Activities	38
	CBM	Carbohydrate-Binding Module	7
	GO	Glyoxal oxidase	22
FOLymes	LO	Laccase	27
	PO	Peroxidase	3
Protease	P	Protease	38

The Cazymes group was the most abundant group with Hydrolases (GH) and Glucosidase/Xylosidase (GH3) superfamilies accounting for 54 and 35 predicted enzymes, respectively. Another significant group found was the Auxiliary Activities (AA) superfamily with 38 enzymes. Due to the natural habitat of *Leucoagaricus* (24), the appearance of enzymes related to the degradation of complex carbohydrates was to be expected, as can be seen, 16 cellulases (GH5), 2 xylanases (GH10), 13 chitinases (GH19), 16 endoglucanases (GH74) and 5 pectinases (PE) were found.

We performed comparative genomics of the predicted protein sequences of the hybrid assembly of *L. gongylophorus* LEU18496 and other genome assemblies of the same genus available in NCBI.

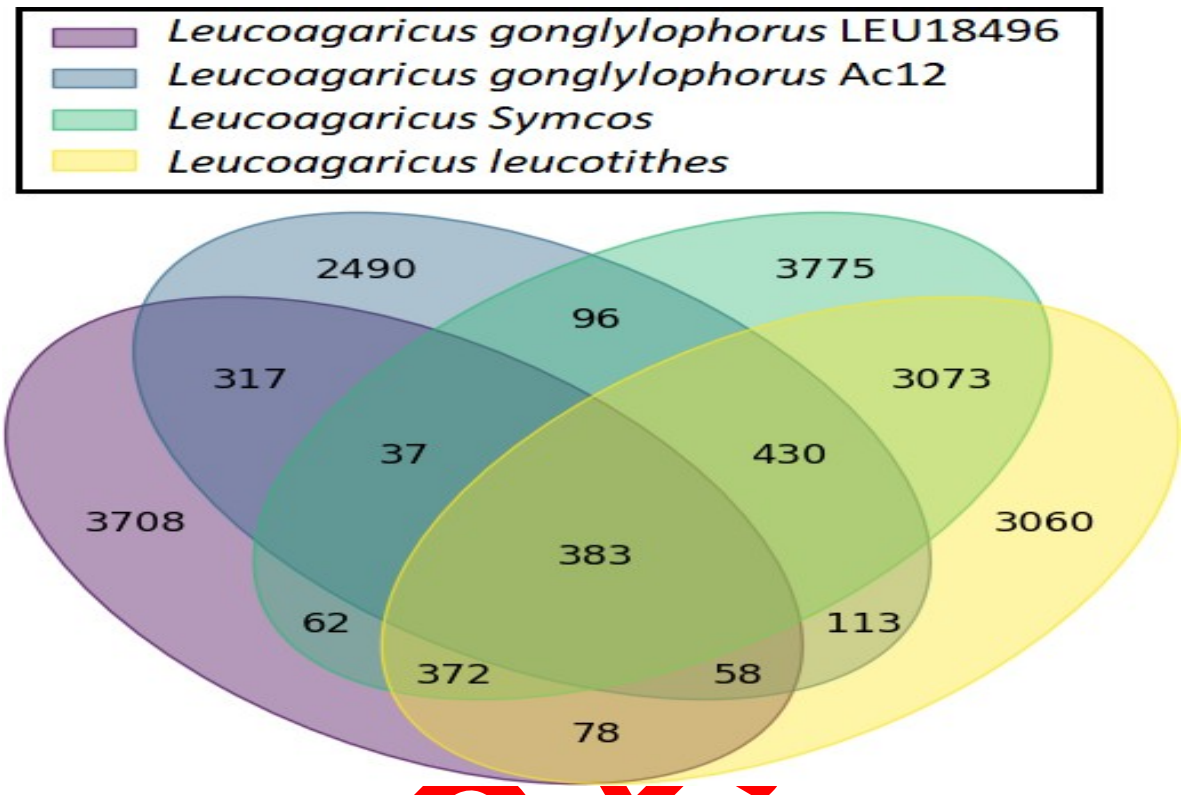


Figure 2 Consensus core and pan-genome obtained from the clusters using the GET HOMOLOGES algorithms COGtriangles (-COG) and OMCL (-M).

The coregenome (core) and pangenome (pan) of these species were obtained through the clustering algorithms -COG and BDH, the combination of several methods allowed to obtain the consensus pangenome: divided in the core and pangenome and thus rule out overestimates. In Figure 2, we list the results corresponding to the pan-genome formed by 18052 gene clusters (Supplementary material 5). Only 2% of the gene clusters (383) are part of the core genome with at least one gene from each genome including protein sequences from the CAZymes, FOLymes and proteases groups and involved in protein and metabolism The pangenome is divided into the coregenome and the accessory genome: 207 with 3775 *L. SymCos* and 3708 *L. gongylophorus* LEU18496, these two species contribute with the highest number of gene clusters to the accessory genome, it should be noted that they are the largest proteomes used in this analysis. *L. gongylophorus* LEU18496 and *L. gongylophorus* AC12 share 317 exclusive gene clusters, a low value given that they belong to the same species.

The accessory genome of all genomes combined encodes 17,669 (98%) protein clusters. This encoded 5,266 (29.8%) protein sets shared by at least two isolates and 12,403 (70.2%) unique clusters found in a single isolate. According to a gene enrichment analysis performed for the coregenome, the most represented functions are transferase activity 10%, oxidoreductase activity 9% and protein binding and metabolism with 7% each (Supplementary material 6)

DISCUSSION

The genome assembly for *Leucoagaricus gongylophorus* LEU18496 was constructed using sequencing data obtained for the Illumina MiSeq and GS FLX+ Roche 454 platforms. Hybrid assemblies using Illumina and Roche 454 technologies have been used for several organisms such as *Hevea brasiliensis* (25) and *Mytilus coruscus* (26). Pootakham *et al*., in 2017 used clean Roche 454 reads and Illumina contigs to assemble the genome of *Hevea brasiliensis* using the Newbler software (Roche Applied Science, Indianápolis, EE. UU.) (27). The total size of 989,097 assembled contigs derived from 454 and Illumina data was 868 Mb, with an N50 contig length of 1,316 bp. Therefore, according to these previous studies, it should be expected that the hybrid assembly Illumina-454 Roche would improve the metrics obtained for individual assemblies (Brown *et al*., 2012a), although this was not our case, the result is in agreement with the work carried out by Utturkar *et al*., in 2014 (28).

Luo *et al*., in 2012 evaluated the advantages and limitations of the Roche 454 and Illumina platforms for assembling metagenomic samples, demonstrating that using Illumina technology assemble longer and more accurate contigs, despite substantially shorter read length relative to Roche 454, although this technology may be advantageous for resolving sequences with repetitive structures or palindromes, given the substantially longer read length (29). DiGuistini *et al*., in 2009, worked on a genome assembly of the filamentous fungus *Grosmannia clavigera* by combining sequencing technologies and assembly methods, the assembly generated from Illumina data alone produced contigs with N50 with a value of 24500 bases. In contrast, an assembly from Roche 454 reads conformed contigs with N50 with a value of 7800 bases. On the other hand, the Roche 454 array contained approximately 2.5 Mb of sequences not found in the Illumina array. These investigations

show that Illumina produces larger contigs but Roche 454 improves the resolution of areas rich in repetitive sequences by increasing sequencing depth (30). In Table 1 is presented a similar number of contigs for both assemblies, nevertheless this difference renders a significant difference in the prediction of genes. In particular, when using the hybrid assembly with the lowest N50, a greater number of genes can be predicted. This indicates that a more fragmented genomic assembly can favor the appearance of genes when using *ab initio* tools such as Augustus. Denton *et al*., in 2014 (31) demonstrated using *Drosophila melanogaster* genomes that the fragmentation of the genome increases the number of genes, a similar result was obtained by Zhang *et al*., in 2014 when they were working with the assembly draft and the annotation of the *Macaco rhesus* (32). A high fragmentation of a genome frequently leads to overestimation of genes (citation), at this point is where a functional annotation is necessary to validate the predictions (33) (34).

By performing a hybrid genomic assembly, it was possible to identify 11,690 possible genes. With the functional annotation process it was possible to assign functions to 11,552 (98.81 %), among thes genes, CAZymes group with 391 proteins in total, while only 52 are found in the FOLymes group. It is known that *L. gongylophorus*, as part of the symbiosis with the ant is exposed to lignocellulosic material, promoting the expression of enzyme with diverse activity for the hydrolysis of the material. During this colonization, the fungus secretes enzymatic cocktails known as CAZymes, which is why a greater relative abundance of these proteins in the genomic assembly is expected. These results are related to those obtained by Aylward *et al*., 2013, who reported the presence of enzymes CAZymes, FOLymes and proteases from a metagenomic assembly of *L. gongylophorus* symbiont of the ant *A. cephalotes*, where the highest relative abundance of proteins found was for the CAZymes group (18,35). White rot fungi such as Leucoagaricus efficiently degrade plant biomass, especially aromatic compounds due to the presence of CAZymes and FOLymes enzymes. The sequencing of the genome of several basidiomycetes reveals the importance and conservation of these enzyme groups (24).

The analysis of CAZymes families showed that the GH superfamily (Hydrolases) is widely represented, the large number found corresponds to the wide variety of metabolic processes to which these enzymes are associated (36) and to the low annotation functional that can be found in databases for enzymes belonging to the genus *Leucoagaricus*. Other families of enzymes that participate in the degradation of lignocellulosic substrates were found: the GH74family that participates in the formation of cellobiose as part of the degradation of cellulose and the PE family involved in the degradation of pectin and the GH3 family involved in xylan degradation. Several investigations have shown the capability of *L. gongylophorus* to grow on different lignocellulosic substrates (37), which explains the fact that these enzymes are present in the assembly of this fungus. Ike et al., in 2015 demonstrated the presence of FOLymes enzymes in *L. gongylophorus*, as part of the study, they characterized two laccases (38), in this study 19 laccases were identified in the annotation. The activity of these enzymes is specific for the elimination of toxic components present in the medium and thus, favor the growth of the organism that produces them on substrates of this nature (39), hence their synthesis by *L. gongylophorus* is important.

Genetic variation increases among the species, and it has been reported that different species can considerably vary their genome even when they are phylogenetically close (40). The coregenome obtained in the genomic comparison contains a low number of gene clusters relative to the size of the analyzed genomes and the number of genomes, although it should be noted that these species of fungi analyzed belong to the same genus, but all have a mutualistic relationship with different species of ants, which distances them evolutionarily. Although, pangenomic analyzes base their search on sequence homology and functional annotation of protein sequences (17), in our case the genus Leucoagaricus has a low level of annotation in different databases such as Uniprot and NCBI and this might drive the genetic variation observed.

As various research groups have reported, the size of the coregenome found should be substantially larger (41,42), although various elements may support the result obtained in this investigation. The dot plot analysis performed for genomic assemblies of the same species shows considerable variations in the alignment of genomic sequences, thus the

predictions of gene products from these sequences vary considerably. Thus, when comparing proteomes belonging to species of poorly annotated species, it is difficult to find phylogenetically conserved functions. Highlighting the conservation in the coregenome of enzymes involved in the synthesis of glycogen and the degradation of lignocellulosic material, distinctive capacity of the genus *Leucoagaricus* (43)(44)

MATERIALS AND METHODS

Fungal strain *L. gongylophorus* LEU18496 (GenBank accession number: KJ419350.1) was isolated from leaf cutter ants *Atta mexicana*, Coatepec, Veracruz, México. The strain was propagated on malt extract agar (MEA-LP), contained (g/L): malt extract 20 g, bacteriological peptone 5 g, yeast extract 2 g and agar 20 g, the pH was adjusted at 5.0. Cultures were incubated at 27°C ± 0.5°C in the dark. Fungal suspension was obtained by adding few milliliters of sterile water and scrapping off the agar surface with glass beads. The fungal strain was maintained at 4°C and in filter paper strips soaked with glycerol 20% (v/v) at -80°C.

Genomic DNA extraction *L. gongylophorus* LEU18496 was grown in 125 mL Erlenmeyer flask with 30 mL of yeast nitrogen base medium (Difco 233520) without amino acids and 4.6 g*L−1 of (N H4)2SO4 and 20 g*L^{-1} of glucose, as nitrogen source carbon source, respectively. pH was adjusted at 5.2. Cultures were inoculated with 1 mL of fungal suspension and maintained at 27°C ± 0.5°C and 150 rpm. To extract the genomic DNA, we used the ZR Fungal/Bacterial DNA MiniPrep kit (Zymo Research) following the manufacturer's instructions. Integrity and purity of the DNA samples were verified in gel and nanodrop (Thermo), the 260/230 nm and 260/280 nm absorbance ratios were 1.8 and 1.96, respectively. The final concentration measurement by fluorescence was 16 ng/µL. Genomic DNA was sequenced at Langebio, Cinvestav (Irapuato, Mexico) using sequencing by synthesis (SBS) with Illumina MiSeq platform 2x250 format with coverage of ± 54x and

sequences of ± 250 bases in length and by massive pyrosequencing with GS FLX+ Roche 454 platform with coverage of ± 20x and sequences of ± 650 bases in length .

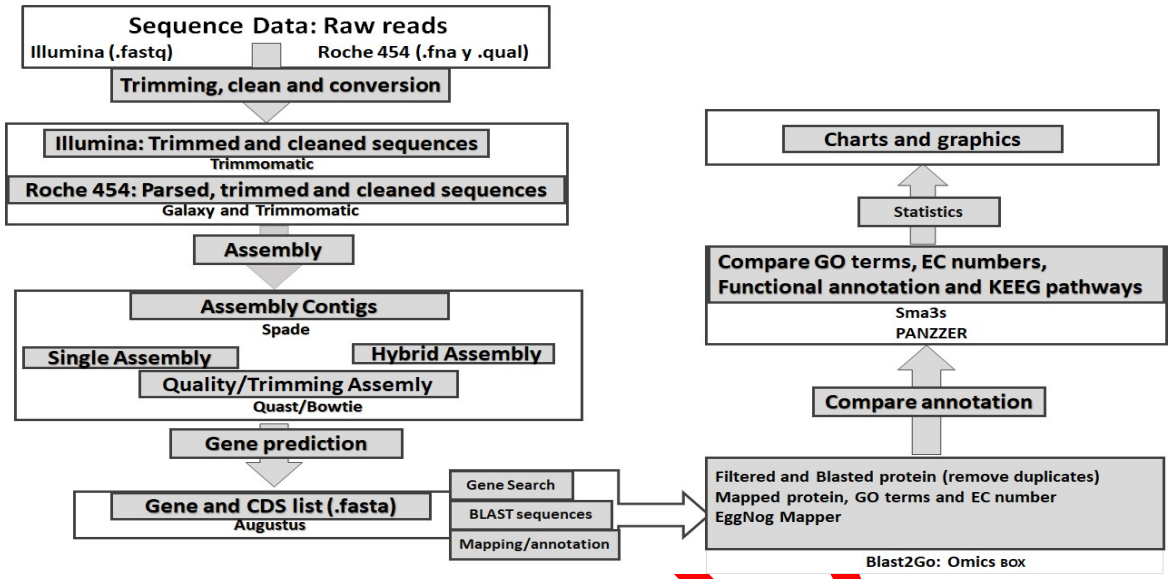


Figure 3 Flow diagram of the assembly process and functional annotation of the genome of *L. gongylophorus* LEU18496.

Raw reads obtained with both sequencing platforms were processed for quality control and to remove adapters using Trimmomatic Version 0.39 (46) . The data obtained for the Roche 454 platform (.fna and .qual files) were converted to the .fastq format using the galaxyproject.org platform (47) using the FASTQ manipulation tool (48) to be used in the genomic assembly process. We used SPADES Version 3.15.0 (49) to perform single (with only Illumina MiSeq data) and Hybrid (Illumina + Roche data) genome assemblies of *L. gongylophorus* using the k-mer options -k33, 55, 77, 99, 111, 127 and the –careful parameter to reduce errors due to base correction. The hybrid genome assembly was submitted as novel genome and assembly for L. gongylophorus to the NCBI database https: //www.ncbi.nlm.nih.gov/(50) (Bioproject SUB10717702, BioSample SAMN23428376).

Gene prediction and functional annotation of the Hybrid genome assembly of *L. gongylophorus* LEU18496

Gene prediction of both genome assemblies of *L. gongylophorus* was performed with the Augustus pipeline Version 3.5.0 (13). For the training of the Augustus tool, 5090 gene sequences reported for the genus *Leucoagaricus* in the NCBI database (51) were used. From this set of genes, random sets of genes were extracted using RandomSplit.pl to carry out several trainings to the tool. Genes for the *L. gongylopohurus* Ac12 genome assembly (BioSample: SAMN02981481, BioProject: PRJNA179280) were also predicted using the same training parameters described above.

Functional annotation of predicted protein sequences was performed using the Blast2Go platform Version 6.0.3 (52), the Protein ANNotation with Z-scoRE (PANNZER) Version 2 tool (53) and the EggNog Mapper web system (54). For the annotation of the previously predicted genes, the strategy described in Figure 2 was used. First, the genes were annotated by sequence homology using local databases built on the Blast2Go platform, then they were annotated using the tool itself and the Mapping, Annotation and INTERPRO options offered by it. Joint data annotation was manually checked and compared with functional annotations using PANZZER and EggNog Mapper. Finally, using Blast2Go's manual annotation option, the sequences were re-annotated considering all functional annotations obtained.

Comparative genomics using the annotated sequence of *L. gongylophorus* LEU18496.

For the pangenomic analysis, 4 proteomes were used: the proteome obtained for *L. gongylophorus* LEU18496 and *L. gongylophorus* Ac12 using Augustus and the proteomes reported in NCBI for *L. SymCos* (PRJNA295288) and *L. leucothites* (PRJNA496460). Predicted genes of *Leucoagaricus* genome assemblies were grouped into putative families (clusters of orthologous genes) with GET_HOMOLOGUES (15) using the OrthoMCL v1.4 (55) and COGtriangles (56) algorithms: min %coverage in BLAST pairwise alignments (range [1-100],default=75) and max E-value (default=1e-05,max=0.01) as control parameters. The compare_cluster.pl and parse_pangenome_matrix.pl scripts from GET_HOMOLOGUES pipeline were used to compile the corresponding pangenome matrix and calculate the core (genes present in

95% or more of the MAGs) and accessory (genes present in less than 95% of the MAGs) genomes. Visualization of the pangenome data was performed using the Upset program (57).

Charts and Data Analysis .

All data presented were filtered and analyzed using the Microsoft Excel statistical package (58). The bar chart presented in Figure 1 was built using Matlab (Mathworks, Inc., Natick, MA, USA) and the Venn diagram in Figure 2 was made using a python script in PyCharm 2021.3.3 (59).

SUPPLEMENTAL MATERIAL

Supplementary material 1 Metrics corresponding to the comparison of the size distribution of the contigs between different genomic assemblies of Leucoagaricus gongylophorus analyzed.

Supplementary material 2 Linearity analysis of genes belonging to the hybrid assembly performed for *Leucoagaricus gonglyophorus* LEU18496 and the genomic assembly reported for *Leucoagaricus gonglyophorus* Ac12.

Supplementary material 3 Results of the comparison between the hypothetical proteins predicted for the constructed genomic assemblies using Blast2Go.

Supplementary material 4 Results of a KEEG PATHWAYS analysis performed in Blast2Go to determine the enzymes contained in the assembly and involved in central carbon metabolism.

Supplementary material 5. This material contains graphs related to the GET HOMOLOGUES analysis performed for 4 genomes belonging to species of the genus Leucoagaricus

Supplementary material 6 Gene enrichment analysis using Blast2go for gene clusters belonging to the coregenome of the 4 species analyzed

REFERENCES

1. Zani RdOA. 2020. Bactérias do abdômen de formigas cortadeiras (Subtribo Attina): cultivo, sequenciamento e fisiologia. .
2. Rønhede S, Boomsma JJ, Rosendahl S. 2004. Fungal enzymes trans¬ferred by leaf-cutting ants in their fungus gardens. Mycol research 108 (1):101–106.
3. Kooij PW, Rogowska-Wrzesinska A, Hoffmann D, Roepstorff P, Boomsma JJ, Schiøtt M. 2014. Leucoagaricus gongylophorus uses leaf-cutting ants to vector proteolytic enzymes towards new plant sub¬strate. The ISME journal 8 (5):1032–1040.
4. Bich GA, Castrillo ML, Villalba L, Zapata PD. 2017. Isolation of the symbiotic fungus of Acromyrmex pubescens and phylogeny of Leuco¬agaricus gongylophorus from leaf-cutting ants. Saudi journal biologi¬cal sciences 24 (4):851–856.
5. Melo CR, Oliveira BMS, Santos ACC, Silva JE, Ribeiro GT, Blank AF, Araújo APA, Bacci L. 2020. Synergistic effect of aromatic plant essential oils on the ant Acromyrmex balzani (Hymenoptera: Formi-cidae) and antifungal activity on its symbiotic fungus Leucoagaricus gongylophorus (Agaricales: Agaricaceae). Environ Sci Pollut Res 27 (14):17303–17313.
6. Chaisson MJ, Wilson RK, Eichler EE. 2015. Genetic variation and the de novo assembly of human genomes. Nat Rev Genet 16 (11):627–640.
7. Mohanta TK, Bae H. 2015. The diversity of fungal genome. Biol pro¬cedures online 17 (1):1–9.
8. Jackman SD, Birol İ. 2010. Assembling genomes using short-read sequencing technology. Genome Biol 11 (1):1–4.
9. Goenka SD, Gorzynski JE, Shafin K, Fisk DG, Pesout T, Jensen TD, Monlong J, Chang PC, Baid G, Bernstein JA, et al.. 2022. Accelerated identification of disease-causing variants with ultra-rapid nanopore genome sequencing. Nat Biotechnol p 1–7.
10. Kchouk M, Gibrat JF, Elloumi M. 2017. Generations of sequencing technologies: from first to next generation. Biol Med 9 (3).

11. Luo J, Wei Y, Lyu M, Wu Z, Liu X, Luo H, Yan C. 02 2021. A compre¬hensive review of scaffolding methods in genome assembly. Briefings Bioinform 22 (5). doi:10.1093/bib/bbab033. Bbab033.
12. Jauhal AA, Newcomb RD. 2021. Assessing genome assembly quality prior to downstream analysis: N50 versus BUSCO. Mol Ecol Resour 21 (5):1416–1421.
13. Hoff KJ, Stanke M. 2019. Predicting genes in single genomes with AUGUSTUS. Curr protocols bioinformatics 65 (1):e57.
14. Hu Z, Sun C, Lu Kc, Chu X, Zhao Y, Lu J, Shi J, Wei C. 2017. EUPAN en¬ables pan-genome studies of a large number of eukaryotic genomes. Bioinformatics 33 (15):2408–2409.
15. Vinuesa P, Contreras-Moreira B. 2015. Robust identification of orthologues and paralogues for microbial pan-genomics using GET_HOMOLOGUES: a case study of pIncA/C plasmids, p 203–232. In Bacterial Pangenomics. Springer.
16. Khan AW, Garg V, Roorkiwal M, Golicz AA, Edwards D, Varshney RK. 2020. Super-pangenome by integrating the wild side of a species for accelerated crop improvement. Trends plant science 25 (2):148–158.
17. McCarthy CG, Fitzpatrick DA. 2019. Pan-genome analyses of model fungal species. Microb genomics 5 (2).
18. Aylward FO, Burnum-Johnson KE, Tringe SG, Teiling C, Tremmel DM, Moeller JA, Scott JJ, Barry KW, Piehowski PD, Nicora CD, et al.. 2013. Leucoagaricus gongylophorus produces diverse enzymes for the degradation of recalcitrant plant polymers in leaf-cutter ant fun¬gus gardens. Appl Environ Microbiol 79 (12):3770–3778.
19. Cabanettes F, Klopp C. 2018. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. PeerJ 6:e4958.
20. Seibt KM, Schmidt T, Heitkam T. 2018. FlexiDot: highly customiz¬able, ambiguity-aware dotplots for visual sequence analyses. Bioin¬formatics 34 (20):3575–3577.

21. Stanke M, Morgenstern B. 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic acids research 33 (suppl_2):W465–W467.

22. Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research 28 (1):27–30.

23. Drula E, Garron ML, Dogan S, Lombard V, Henrissat B, Terrapon N. 2022. The carbohydrate-active enzyme database: functions and literature. Nucleic acids research 50 (D1):D571–D577.

24. Qin W, et al.. 2016. Recent developments in using advanced sequenc¬ing technologies for the genomic studies of lignin and cellulose de¬grading microorganisms. Int journal biological sciences 12 (2):156.

25. Pootakham W, Sonthirod C, Naktang C, Ruang-Areerate P, Yoocha T, Sangsrakru D, Theerawattanasuk K, Rattanawong R, Lekawipat N, Tangphatsornruang S. 2017. De novo hybrid assem¬bly of the rubber tree genome reveals evidence of paleotetraploidy in Hevea species. Sci Reports 7 (1):1–15.

26. Li R, Zhang W, Lu J, Zhang Z, Mu C, Song W, Migaud H, Wang C, Bekaert M. 2020. The whole-genome sequencing and hybrid assem¬bly of Mytilus coruscus. Front Genet 11:440.

27. Nederbragt AJ. 2014. On the middle ground between open source and commercial software-the case of the Newbler program. Genome biology 15 (4):1–2.

28. Utturkar SM, Klingeman DM, Land ML, Schadt CW, Doktycz MJ, Pelletier DA, Brown SD. 2014. Evaluation and validation of de novo and hybrid assembly techniques to derive high-quality genome se¬quences. Bioinformatics 30 (19):2709–2716.

29. Luo C, Tsementzi D, Kyrpides N, Read T, Konstantinidis KT. 2012 Direct comparisons of Illumina vs. Roche 454 sequencing technolo¬gies on the same microbial community DNA sample. PloS one 7 (2):e30087.

30. DiGuistini S, Liao NY, Platt D, Robertson G, Seidel M, Chan SK, Docking TR, Birol I, Holt RA, Hirst M, et al.. 2009. De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. Genome biology 10 (9):1–12.

31. Denton JF, Lugo-Martinez J, Tucker AE, Schrider DR, Warren WC, Hahn MW. 2014. Extensive error in the number of genes inferred from draft genome assemblies. PLoS computational biology 10 (12):e1003998.

32. Zhang X, Goodsell J, Norgren RB. 2012. Limitations of the rhesus macaque draft genome assembly and annotation. BMC genomics 13 (1):1–9.

33. Schrider DR, Costello JC, Hahn MW. 2009. All human-specific gene losses are present in the genome as pseudogenes. J Comput Biol 16 (10):1419–1427.

34. Vinson JP, Jaffe DB, O'Neill K, Karlsson EK, Stange-Thomann N, An¬derson S, Mesirov JP, Satoh N, Satou Y, Nusbaum C, et al.. 2005. As¬sembly of polymorphic genomes: algorithms and application to Ciona savignyi. Genome research 15 (8):1127–1135.

35. Aragona M, Minio A, Ferrarini A, Valente MT, Bagnaresi P, Orrù L, Tononi P, Zamperin G, Infantino A, Valè G, et al.. 2014. De novo genome assembly of the soil-borne fungus and tomato pathogen Pyrenochaeta lycopersici. BMC genomics 15 (1):1–12.

36. Casa Villegas MF. 2018. Caracterización de glicosidasas y permeasas fúngicas implicadas en el transporte y metabolismo de azúcares. PhD thesis. Universitat Politècnica de València.

37. Maya-Yescas ME, Revah S, Le Borgne S, Valenzuela J, Palacios-González E, Terrés-Rojas E, Vigueras-Ramírez G. 2021. Growth of Leucoagaricus gongylophorus Möller (Singer) and production of key enzymes in submerged and solid-state cultures with lignocellulosic substrates. Biotechnol Lett 43 (4):845–854.

38. Ike PTL, Moreira AC, de Almeida FG, Ferreira D, Birolli WG, Porto ALM, Souza DHF. 2015. Functional characterization of a yellow lac¬case from Leucoagaricus gongylophorus. Springerplus 4 (1):1–11.

39. Reinhammar B. 2018. Laccase, p 1–36. In Copper proteins and cop¬per enzymes. CRC press.

40. Plissonneau C, Hartmann FE, Croll D. 2018. Pangenome analyses of the wheat pathogen Zymoseptoria tritici reveal the structural basis of a highly plastic eukaryotic genome. BMC biology 16 (1):1–16.

41. Kim Y, Gu C, Kim HU, Lee SY. 2020. Current status of pan-genome analysis for pathogenic bacteria. Curr opinion biotechnology 63:54–62.

42. Tao Y, Jordan DR, Mace ES. 2020. A graph-based pan-genome guides biological discovery. Mol Plant 13 (9):1247–1249.

43. Kermarrec A, Decharme M, Febvay G. 2019. Leaf-cutting ant symbi¬otic fungi: a synthesis of recent research. Fire Ants Leaf-Cutting Ants p 231–246.

44. CEJAS AÑON G. 2021. Evaluación del metabolismo de producción del glucógeno y enzimas CAZymes y FOLymes de Leucoagaricus gongylophorus. .

45. Vigueras G, Paredes-Hernández D, Revah S, Valenzuela J, Olivares-Hernández R, Le Borgne S. 2017. Growth and enzy¬matic activity of Leucoagaricus gongylophorus, a mutualistic fungus isolated from the leaf-cutting ant Atta mexicana, on cellulose and lignocellulosic biomass. Lett applied microbiology 65 (2):173–181.

46. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trim¬mer for Illumina sequence data. Bioinformatics 30 (15):2114–2120.

47. Jalili V, Afgan E, Gu Q, Clements D, Blankenberg D, Goecks J, Tay¬lor J, Nekrutenko A. 2020. The Galaxy platform for accessible, repro¬ducible and collaborative biomedical analyses: 2020 update. Nucleic acids research 48 (W1):W395–W402.

48. Blankenberg D, Gordon A, Von Kuster G, Coraor N, Taylor J, Nekrutenko A, Team G. 2010. Manipulation of FASTQ data with Galaxy. Bioinformatics 26 (14):1783–1785.

49. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, et al.. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J computational biology 19 (5):455–477.

50. Coordinators NR. 2018. Database resources of the national center for biotechnology information. Nucleic acids research 46 (Database issue):D8.

51. Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovan-skaya R, Leipe D, Mcveigh R, O'Neill K, Robbertse B, et al.. 2020. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database 2020.

52. Conesa A, Götz S. 2008. Blast2GO: a comprehensive suite for func¬tional analysis in plant genomics. Int journal plant genomics 2008.

53. Törönen P, Medlar A, Holm L. 2018. PANNZER2: a rapid functional annotation web server. Nucleic acids research 46 (W1):W84–W88.

54. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. 2021. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol biology evolution 38 (12):5825–5829.

55. Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H. 2014. Up-Set: visualization of intersecting sets. IEEE transactions on visualiza¬tion computer graphics 20 (12):1983–1992.

56. Li, L., Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome research, 13(9), 2178-2189.

57. Kristensen, D. M., Kannan, L., Coleman, M. K., Wolf, Y. I., Sorokin, A., Koonin, E. V., & Mushegian, A. (2010). A low-polynomial algorithm for assembling

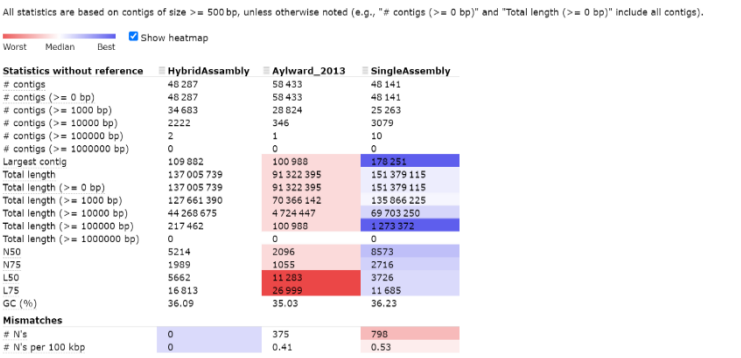
clusters of orthologous groups from intergenomic symmetric best matches. Bioinformatics, 26(12), 1481-1487.

58. Anderson DR, Sweeney DJ, Williams TA, Camm JD, Cochran JJ. 2020. Modern business statistics with Microsoft Excel. Cengage Learn¬ing.

59. Nguyen Q. 2019. Hands-on application development with PyCharm: Accelerate your python applications using practical coding techniques in PyCharm. Packt Publishing Ltd.

Suplemmentary Material 1

Metrics corresponding to the comparison of the size distribution of the contigs among different genomic assemblies of *Leucoagaricus gongylophorus* analyzed.



All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Worst Median Best ☑ Show heatmap

Statistics without reference	HybridAssembly	Aylward_2013	SingleAssembly
# contigs	48 287	58 433	48 141
# contigs (>= 0 bp)	48 287	58 433	48 141
# contigs (>= 1000 bp)	34 683	28 824	25 263
# contigs (>= 10000 bp)	2222	346	3079
# contigs (>= 100000 bp)	2	1	10
# contigs (>= 1000000 bp)	0	0	0
Largest contig	109 882	100 988	178 251
Total length	137 005 739	91 322 395	151 379 115
Total length (>= 0 bp)	137 005 739	91 322 395	151 379 115
Total length (>= 1000 bp)	127 661 390	70 366 142	135 866 225
Total length (>= 10000 bp)	44 268 675	4 724 447	69 703 250
Total length (>= 100000 bp)	217 462	100 988	1 273 372
Total length (>= 1000000 bp)	0	0	0
N50	5214	2096	8573
N75	1989	1055	2716
L50	5662	11 283	3726
L75	16 813	26 999	11 685
GC (%)	36.09	35.03	36.23
Mismatches			
# N's	0	375	790
# N's per 100 kbp	0	0.41	0.53

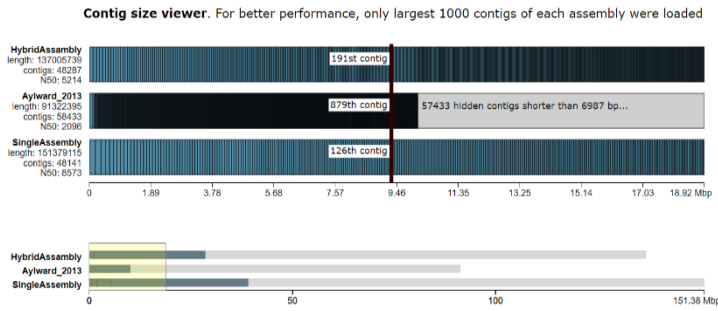
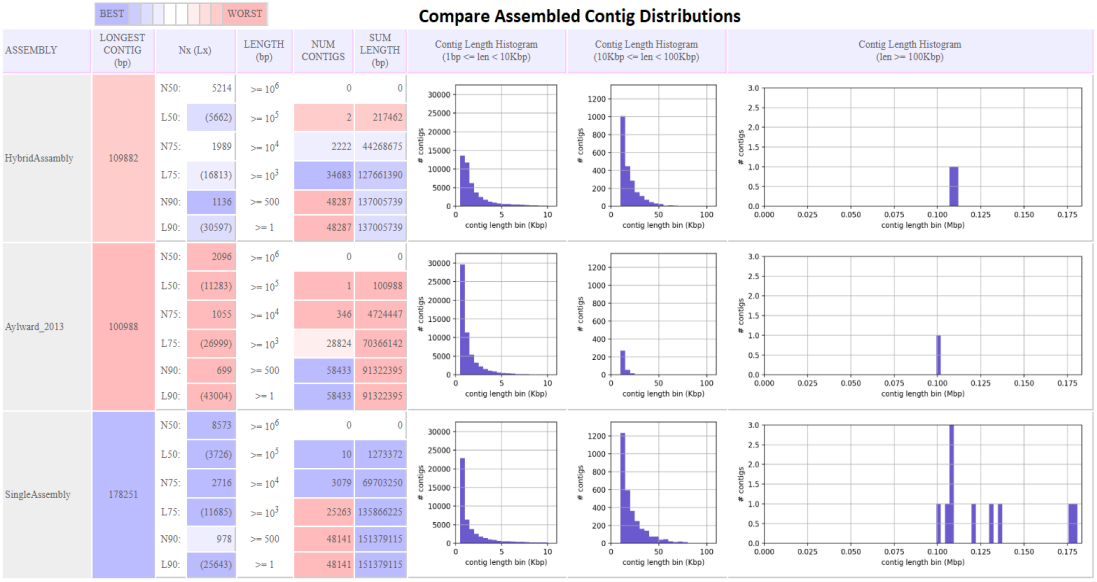


FIG1: Report obtained using the QUAST tool



Compare Assembled Contig Distributions

BEST WORST

ASSEMBLY	LONGEST CONTIG (bp)	Nx (Lx)	LENGTH (bp)	NUM CONTIGS	SUM LENGTH (bp)	Contig Length Histogram (1bp <= len < 10Kbp)	Contig Length Histogram (10Kbp <= len < 100Kbp)	Contig Length Histogram (len >= 100Kbp)
HybridAssembly	109882	N50: 5214	>= 10^6	0	0	# contigs / contig length bin (Kbp)	# contigs / contig length bin (Kbp)	# contigs / contig length bin (Mbp)
		L50: (5662)	>= 10^5	2	217462			
		N75: 1989	>= 10^4	2222	44268675			
		L75: (16813)	>= 10^3	34683	127661390			
		N90: 1136	>= 500	48287	137005739			
		L90: (30597)	>= 1	48287	137005739			
Aylward_2013	100988	N50: 2096	>= 10^6	0	0	# contigs / contig length bin (Kbp)	# contigs / contig length bin (Kbp)	# contigs / contig length bin (Mbp)
		L50: (11283)	>= 10^5	1	100988			
		N75: 1055	>= 10^4	346	4724447			
		L75: (26999)	>= 10^3	28824	70366142			
		N90: 699	>= 500	58433	91322395			
		L90: (43004)	>= 1	58433	91322395			
SingleAssembly	178251	N50: 8573	>= 10^6	0	0	# contigs / contig length bin (Kbp)	# contigs / contig length bin (Kbp)	# contigs / contig length bin (Mbp)
		L50: (3726)	>= 10^5	10	1273372			
		N75: 2716	>= 10^4	3079	69703250			
		L75: (11685)	>= 10^3	25263	135866225			
		N90: 978	>= 500	48141	151379115			
		L90: (25643)	>= 1	48141	151379115			

FIG 2: Report obtained using the Compare Assembled tool in Kbase.

Suplemmentary Material 2

Linearity analysis of genes belonging to the hybrid assembly performed for Leucoagaricus gonglyophorus LEU18496 and the genomic assembly reported for Leucoagaricus gonglyophorus Ac12

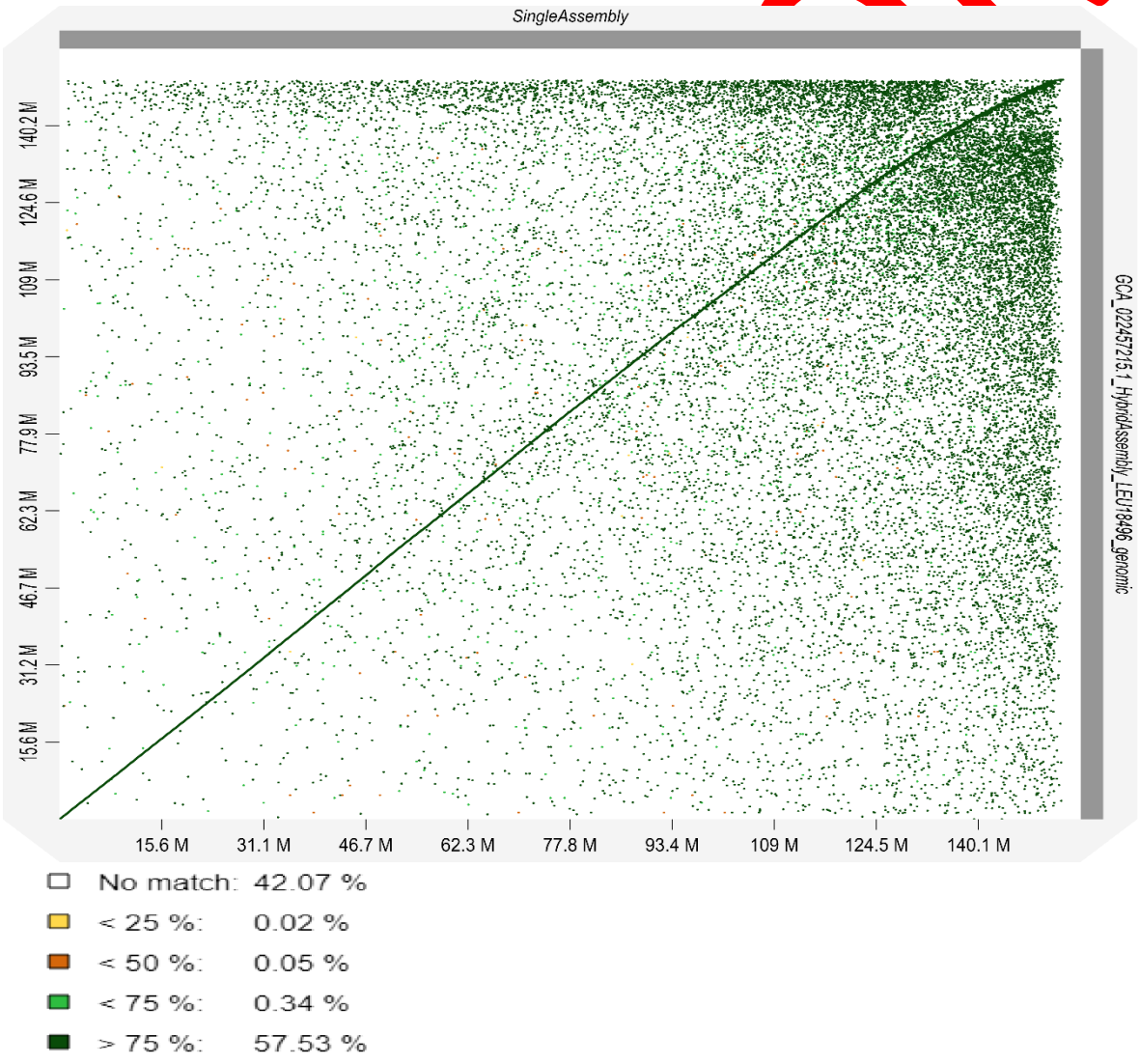


FIG3: Dot plot obtained using the Dgenies tools for the alignment análisis of the hybrid and single assembly.

Suplemmentary Material 3

Results of the comparison between the hypothetical proteins predicted for the constructed genomic assemblies using Blast2Go

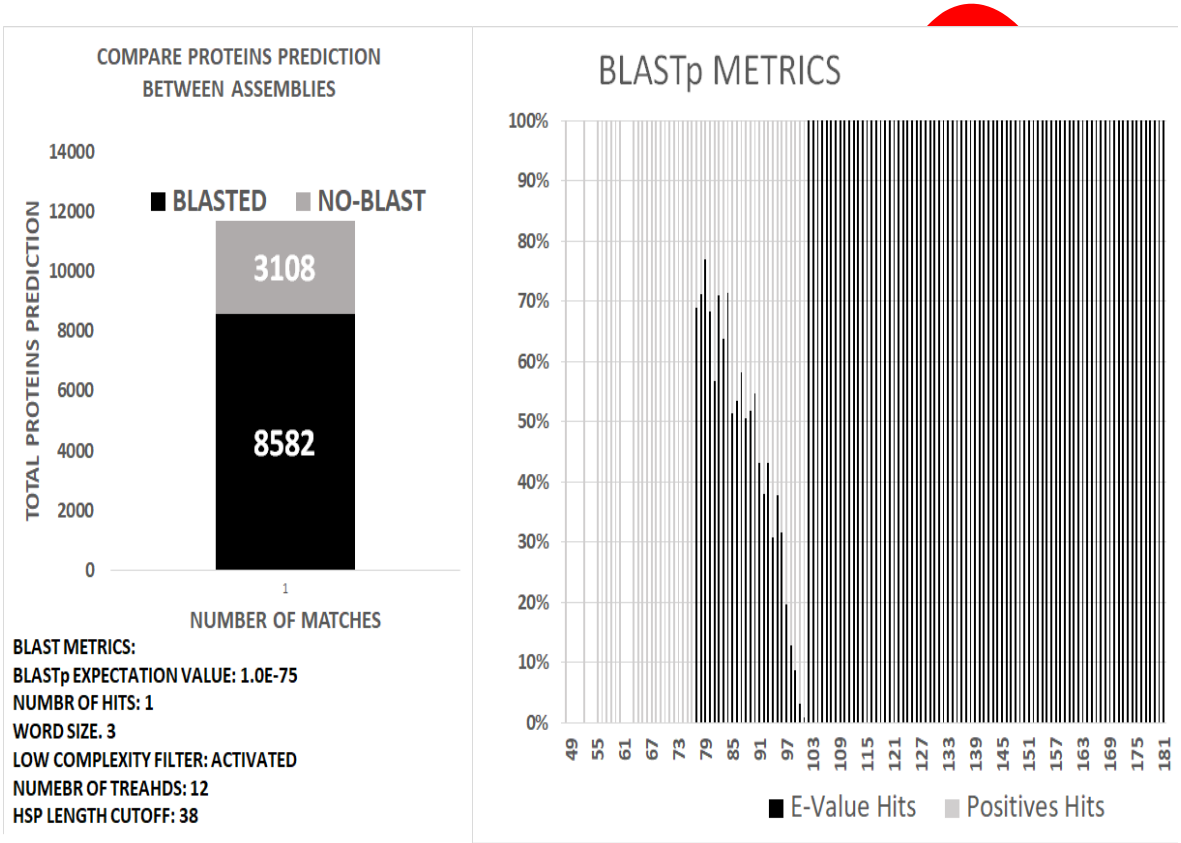


FIG4: Dot plot obtained using the Dgenies tools for 2 genomes

Suplemmentary Material 4

Results of a KEEG PATHWAYS analysis performed in Blast2Go to determine the enzymes contained in the assembly and involved in central carbon metabolism

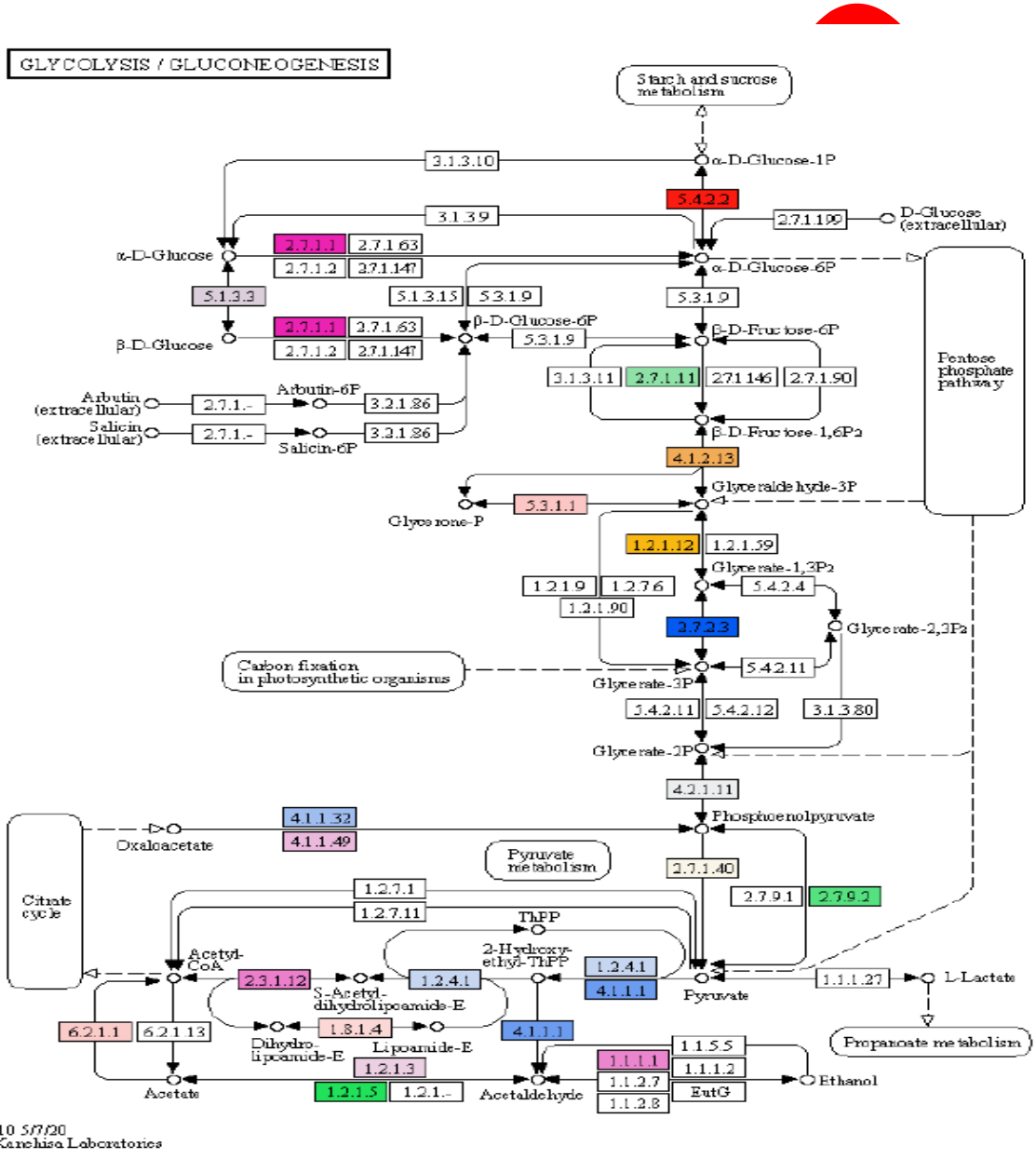


FIG5: Enzymes found for the Glycolysis/Guconeogenesis pathway

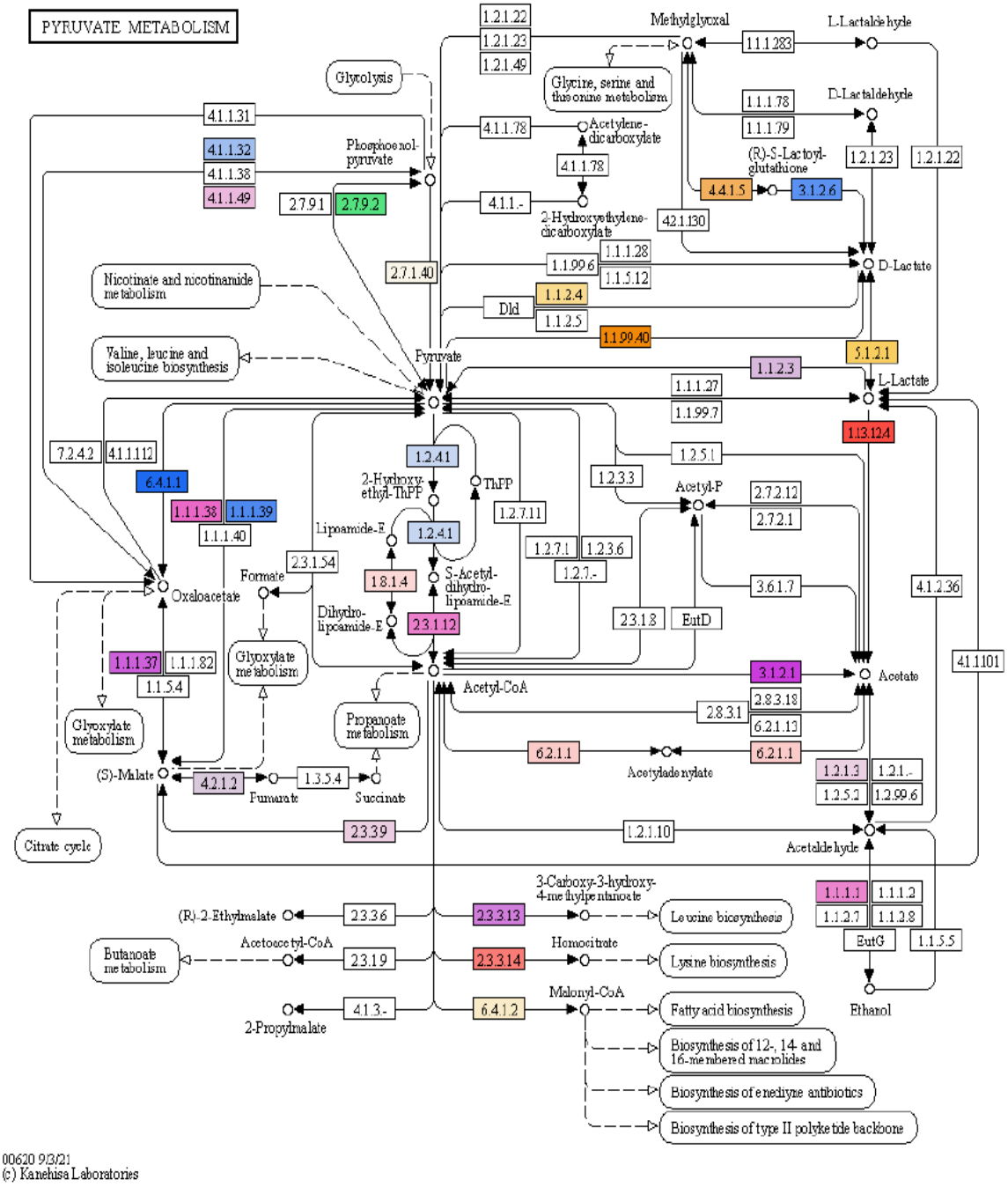


FIG6: Enzymes found for the Pyruvate matabolism pathway

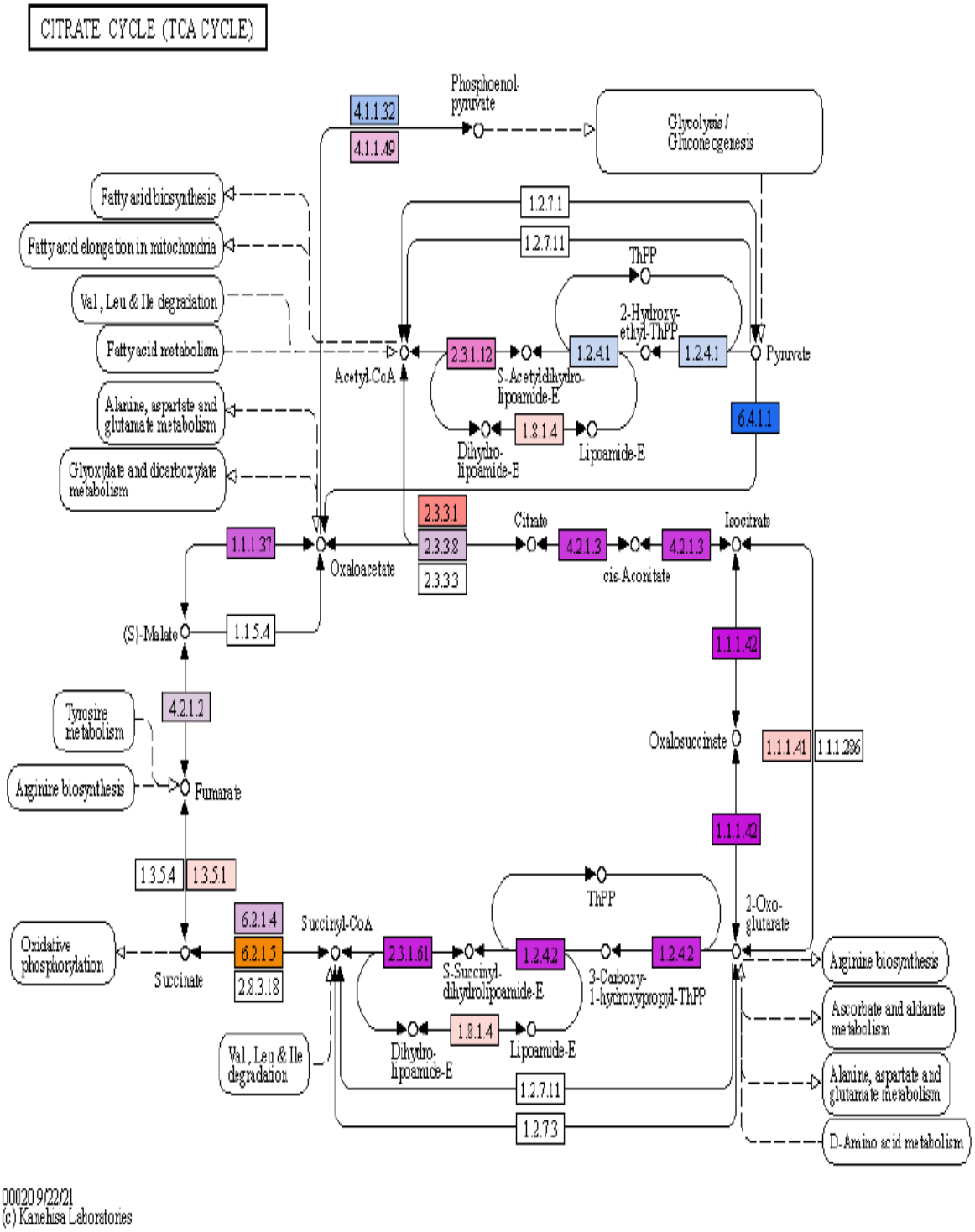


FIG7: Enzymes found for the Tricabolxilic acid cicle

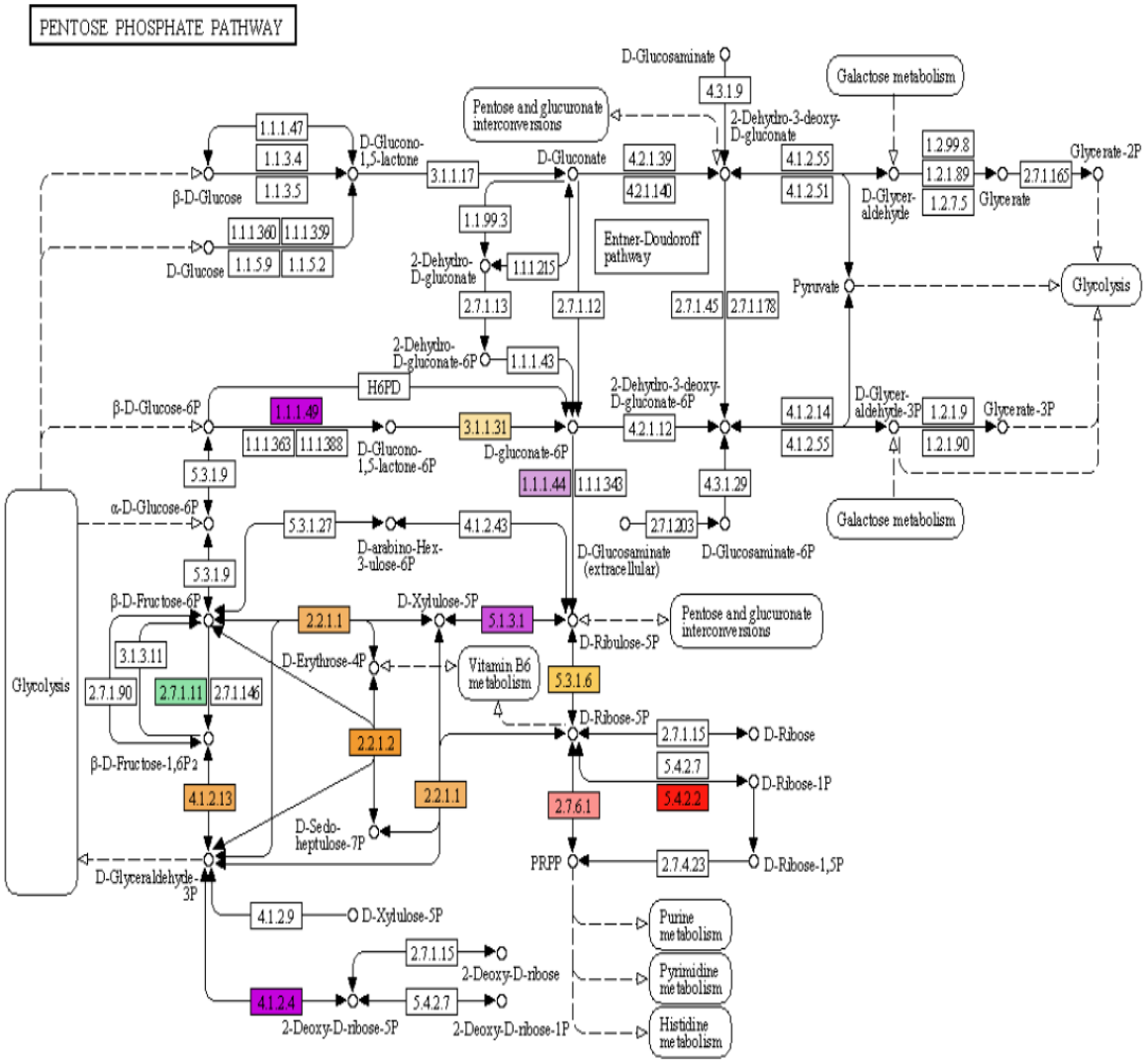


FIG8: Enzymes found for the Pentose Phosphate pathway

Suplemmentary Material 5

This material contains graphs related to the GET HOMOLOGUES analysis performed for 4 genomes belonging to species of the genus Leucoagaricus

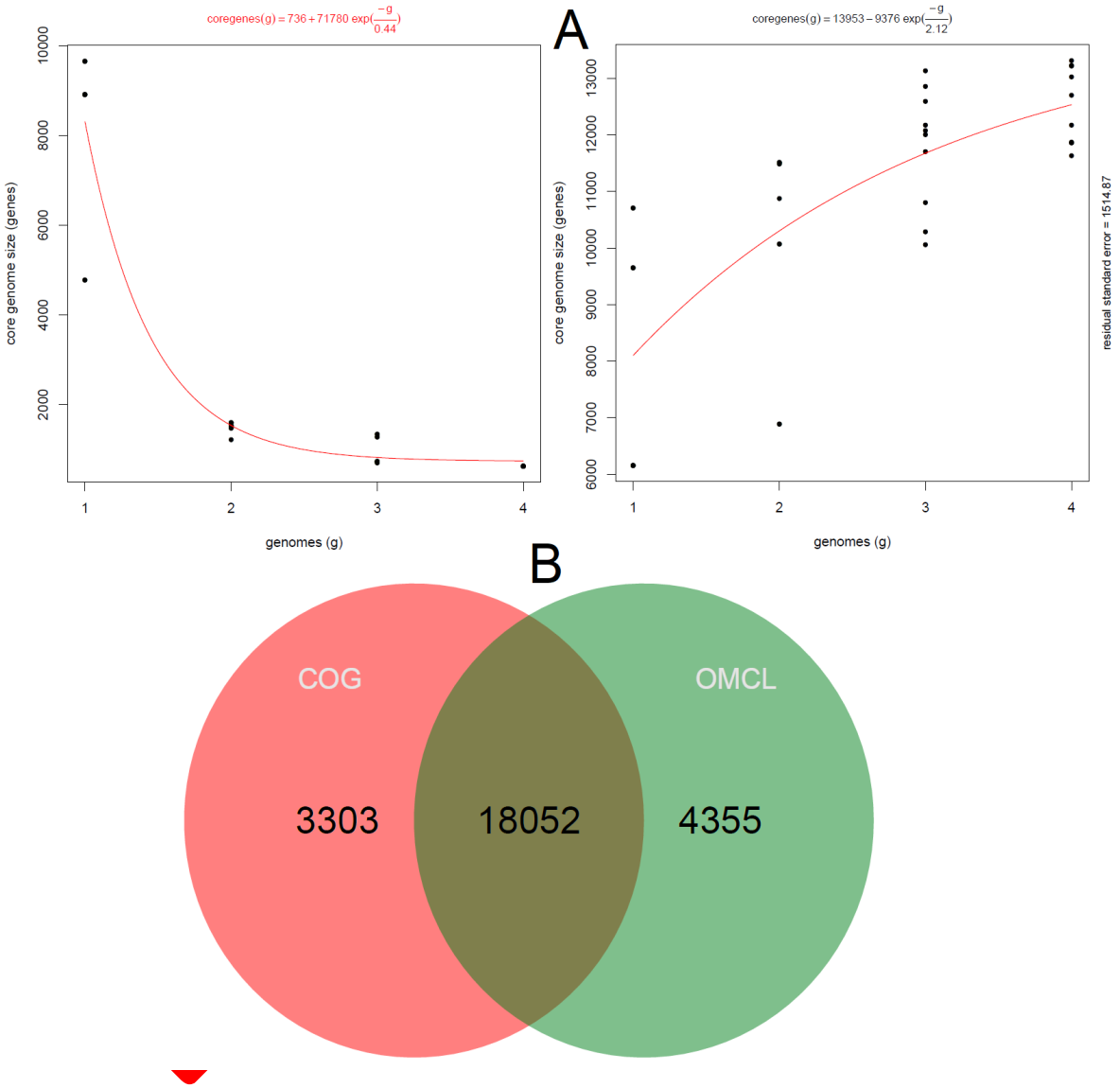


FIG9: GET HOMOLOGUES RESULTS A: mathematical adjustment made for the data obtained by the OMCL algorithm, B: Venn diagram of the418consensus COREGENOMA and B2: Venn diagram of the consensus PANGENOMA obtained by compare cluster python script

Suplemmentary Material 6

Gene enrichment analysis using Blast2go for gene clusters belonging to the coregenome of the 4 species analyzed

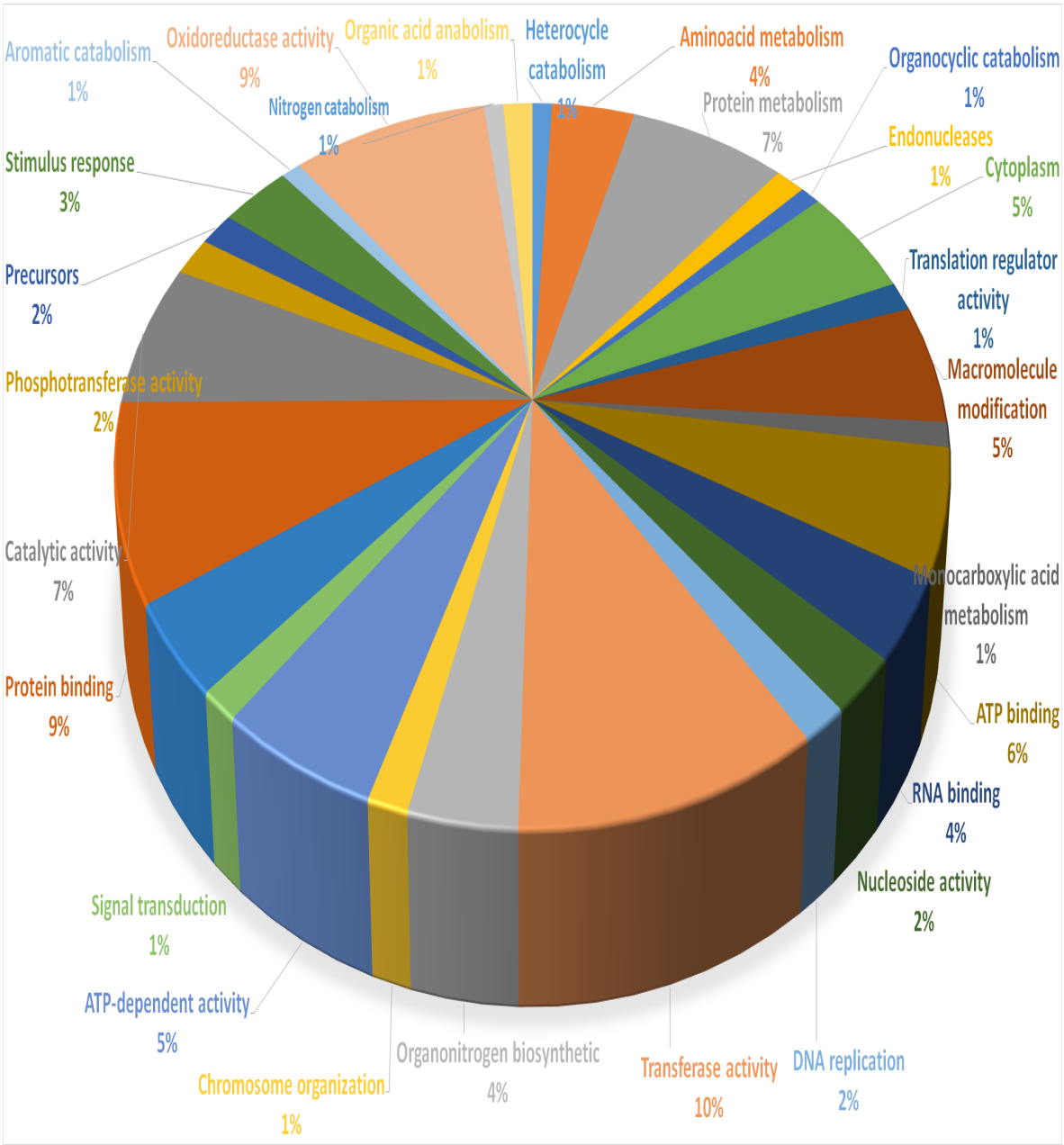


FIG10: Pie plot of Gene Ontology functions found during gene enrichment