The genomic study of an environmental isolate of *Scedosporium apiospermum* shows its metabolic potential to degrade hydrocarbons.

Morales M. Laura Tatiana[1], Gonzalez G. Laura Natalia[2], Restrepo Silvia[3], Vives F. Martha Josefina [4]

[1] Department of Biological Sciences, Universidad de los Andes, Bogotá, Colombia (lt.morales10@uniandes.edu.co)

[2] Universidad de los Andes, Bogotá, Colombia (ln.gonzalez138@uniandes.edu.co)

[3] Department of Biological Sciences, Universidad de los Andes, Bogotá, Colombia (srestrep@uniandes.edu.co)

[4] Department of Biological Sciences, Universidad de los Andes, Bogotá, Colombia (mvives@uniandes.edu.co)

Keywords: bioremediation; metabolic pathway; hydrocarbons; illumina – solexa sequencing

ABSTRACT

Crude oil contamination of soils and waters is a worldwide problem, which has been actively addressed in recent years. Sequencing genomes of microorganisms involved in the degradation of hydrocarbons have allowed the identification of several promoters, genes and degradation pathways of these contaminants, allowing a better understanding of the functional dynamics of microbial degradation. Here, we report a first draft of the 44.2 Mbp genome assembly of an environmental strain of the fungus *Scedosporium apiospermum.* The assembly consisted of 186 high-quality DNA scaffolds with 1.62 % of sequence repeats identified. A total of 11,195 protein-coding genes were predicted including a diverse group of gene families involved in hydrocarbon degradation pathways like dioxygenases and cytochrome P450 monooxygenases. The metabolic pathways identified in the genome can potentially degrade hydrocarbons like chloroalkane/alkene, chorocyclohexane and chlorobenzene, benzoate, aminobenzoate, fluorobenzoate, toluene, caprolactam, geraniol, naphthalene, styrene, atrazine, dioxin, xylene, ethylbenzene and polycyclic aromatic hydrocarbons. The comparison analysis between this strain and the previous sequenced clinical strain showed important differences in the annotated genes involved in the hydrocarbon degradation process.

Keywords: *Scedosporium apiospermum,* genome, degradation of hydrocarbons

INTRODUCTION

Soil and water contamination caused by crude oil or refining processes is a problem of global importance which has generated a great interest in recent years mainly due to accidental spills which have caused serious environmental damages [1]. Since oil is a complex mixture of aromatic and aliphatic hydrocarbons of different molecular weights, its removal is difficult and its permanence in the environment is prolonged [2]. These compounds have gained considerable attention because of their harmful features like presence in all environments (air, soil and water), resistance to degradation, bioaccumulation and carcinogenic activity. Their persistence in the environment increases with their molecular weight and there is a need to develop technologies or processes able to degrade or transform these compounds into less toxic substances [3].

The ability of several organisms, primarily microorganisms (bacteria, fungi and microalgae), to degrade these toxic substances has been extensively studied in recent decades [1, 4-9]. The main goal is to improve the decontamination of the environment via bioremediation, which encompasses technologies that allow the transformation of compounds to less harmful or not harmful forms, with less use of chemicals, energy and time [10]. Microbial bioremediation is very effective, due to the catabolic activity of microorganisms; among these, many species of bacteria, fungi and microalgae have demonstrated the ability of hydrocarbon degradation. This process involves the breakdown of organic molecules through biotransformation in less complex metabolites, or mineralization to water, carbon dioxide or methane [3].

Several strategies have been employed to study these microorganisms and understanding the processes carried out by them. Within these, genomics (based on the analysis of all the genetic information in a cell) has allowed the recognition of promoters, genes and

degradation pathways that influence the construction of more efficient degradative strains relevant in bioremediation process [11, 12]. Genome sequencing of hydrocarbon-degrader organisms has allowed the identification of several genes involved in metabolism and catabolism of aliphatic, aromatic alcohols and compounds, as well as some metals resistance genes [13]. Compared to bacteria, the number of sequenced genomes of fungal species is very low. While there are 218,766 sequenced genomes and sequencing projects from bacteria, there are only 13,699 eukaryotic genomes or sequencing projects to date (https://gold.jgi-psf.org/, September 2016) [14].

Scedosporium apiospermum (teleomorph: *Pseudallescheria apiosperma* [15]) is a fungus belonging to the phylum Ascomycota, which has been isolated from various environments, usually in environments influenced by human activity [16]. This fungus was reported as a hydrocarbon-degrading microorganism since 1998 due to its ability to degrade polluting compounds, such as phenol and p-cresol [17]. One year later, its ability to degrade phenylbenzoate and its derivatives was elucidated [18]. In recent years, studies regarding degradation of complex compounds such as polycyclic aromatic hydrocarbons [19], long-chain aliphatic hydrocarbons and mixtures of these contaminants (unpublished results from our group) [20] have risen. Additionally, the fungus' ability to regenerate granular activated carbon once it has been saturated with phenol was shown in our laboratory (unpublished results).

Therefore, *Scedosporium apiospermum* presents a wide range of opportunities in bioremediation and its genome sequencing can allow the identification of promoters, genes and degradation pathways of hydrocarbons. Indeed, the genomic analysis of this fungus can open possibilities for a better understanding of the functional dynamics of the microbial degradation of contaminants and to improve conditions for effective decontamination

processes in different environments [2]. On the other hand, this fungus has been recognized as a potent etiologic agent of severe infections in immunocompromised and occasionally also in immunocompetent patients [21]. For this reason, in 2014, the genome of an isolate from a cystic fibrosis patient (clinical strain) was sequenced with the aim of gaining knowledge of its pathogenic mechanisms [22]. However, the genome sequence obtained remains as a draft.

Thus, our objective was the complete characterization of the genome sequence of the *S. apiospermum* environmental strain HDO1. For this, we sequenced, assembled, annotated and fully characterized the environmental strain genome, in order to analyze the genes and pathways involved in the degradation process and to assess the unique components of its genome compared to the clinical strain and other sister species.

MATERIALS AND METHODS

Sample

Scedosporium apiospermum environmental strain HDO1 was isolated as a contaminant from assays on bacterial strains able to grow in crude oil as the only carbon source. It was selected for sequencing due its capability to grow in cultures containing aliphatic hydrocarbons of crude oil, naphthalene, phenantrene, phenol and mixtures of these compounds in the laboratory.

DNA extraction and genome sequencing

Fungus growth was carried out in liquid culture (YPG: 1% yeast extract, peptone 1% and 2% glucose) at 30 ° C for 7 days, followed by vacuum filtration, lyophilization and

maceration to have a homogeneous sample. DNA was extracted by the method of CTAB and Phenol/Chloroform/Isoamylic alcohol (Vasco, M. F. et al. 2011) [23]. DNA quality was analyzed by Nanodrop2000 and agarose gel electrophoresis (0.8 %). Quantity was determined by Qubit2.0.

The genome sequencing of the strain was performed by Novo Gene Technology Bioinformatics Co., Ltd. (Beijin, Hong Kong) using high-throughput Illumina-Solexa technology on a Hiseq2500 and employing two libraries: a 250 bp paired-ends library and a 5kpb mate-pairs library. Quality trimming of reads was performed using Trimmomactic 0.23 [24] and quality control was performed using Fastqc 0.11.2 [25]. Coverage and deep of sequencing was analyzed by mapping the reads using Bowtie2-2.2.4 [26], the sam files were converted to bam files for visualization using samtools1.1 [27], and the visualization was made using tablet. 1.15.09.1[28] .

Assembling and *annotation*

The genome was *the novo* assembled using Abyss 1.0.9.20 [29] with a kmer size of 64, scaffolds were generated with SSPACE BASIC 2.0 [30] and gaps were reduced using GapFiller 1.1 [31]. Assembly statistics were obtained by Quast 2.3 [32]. Repetitive elements were identified by RepeatMasker 4.0.5 [33]. Gene prediction and structure annotation was conducted using Augustus 3.0.3 [34]. Functional annotation was performed using Blast2GO 3.1 [35] with BLASTx in the National Center for Biotechnology Information (NCBI) “nr” database [36], allowing the comparison of the sequences aligning and using the Gene Ontology (GO) classification [37]. Protein classification was made using the COG (Cluster of Orthologous Groups) [38], KOG (Eukaryotic Orthologous Groups) [39] and EggNOG [40] databases using Blast2GO v4.0 platform [35]. Annotated

genes were mapped against Kyoto encyclopedia of genes and genomes (KEGG) [41] to its functional analysis and assigned the Enzyme Codes (EC).

Phylogenetic analysis

A phylogenetic analysis was performed using the long subunit (*LSU*) rRNA gene, the internal transcribed spacer (*ITS*) and elongation factor 1-α (*TEF*) sequences from species from *Microascaceae* family [42] [43] and all sequences are shown in the supplementary table 2. The genes sequences were obtained from GenBank (http://www.ncbi.nlm.nih.gov/nuccore). Individual gene regions (*LSU*, *ITS* and *TEF*) were aligned, using MAFFT v. 7.187 [44]. Next, Maximum Likelihood (ML) analyses were performed, using RAxML v.7.6.3 [45] as implemented on the CIPRES portal [46]. The sequence alignment was partitioned into three subsets, each one under a specified model of nucleotide substitution, chosen with the program PartitionFinder [47], but allowing the estimation of different shapes, GTR rates, and base frequencies for each partition. The majority rule criterion implemented in RAxML (-autoMRE) was used to assess clade support. The resulting trees were plotted using FigTree v. 1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/). *Microascus longirostris* and *Scopulariopsis brevicaulis* were used as outgroups.

Comparative genomics

Reads were mapped versus the clinical strain IHEM 14462 using Bowtie2-2.2.4 [26]. The sam files were converted to bam files for the visualization using samtools1.1 [27] and the visualization was made using tablet. 1.15.09.1 [28]. Genomes' comparison between the

environmental strain HDO1 and the clinical strain IHEM 14462 was performed using the alignments made by Quast 4.3 (http://quast.bioinf.spbau.ru/), and MAUVE 20150226 [48]. From ordered output fasta file obtained with MAUVE [48] a new alignment was made with Nucmer at nucleotide level (maximum gap between two adjacent matches in a cluster of 90 bp and a minimum length of a maximal exact match of 20 bp) and Promer at amino acid level (maximum gap between two adjacent matches in a cluster of 30 amino acids and a minimum length of a maximal exact match of 6 amino acids). Nucmer and Promer alignments were plotted using Mumerplot, the last three mentioned tools from mummer 3.0 [49].

RESULTS

Assembling and *annotation*

The draft genome of *S. apiospermum* strain HDO1 was assembled from a –total of 97,208,043 reads using Abyss [29] assembler. The assembly yielded 186 scaffolds (larger than 500 bp) with a genome size of 44.2 Mbp and a G+C content of 49.91% with a mean depth of 541X. The summary of the genome assembly statistics calculated by Quast [32] is shown in Table 1.

Table 1. Genome assembly statistics reported by Quast [32].

Attribute	value
Genome size (pb)	44,195,431
DNA contigs	2,254
DNA Scaffolds	186

Largest contig	5,361,369
GC (%)	49.91
N50	2,801,028
N75	1,627,699
L50	6
L75	12

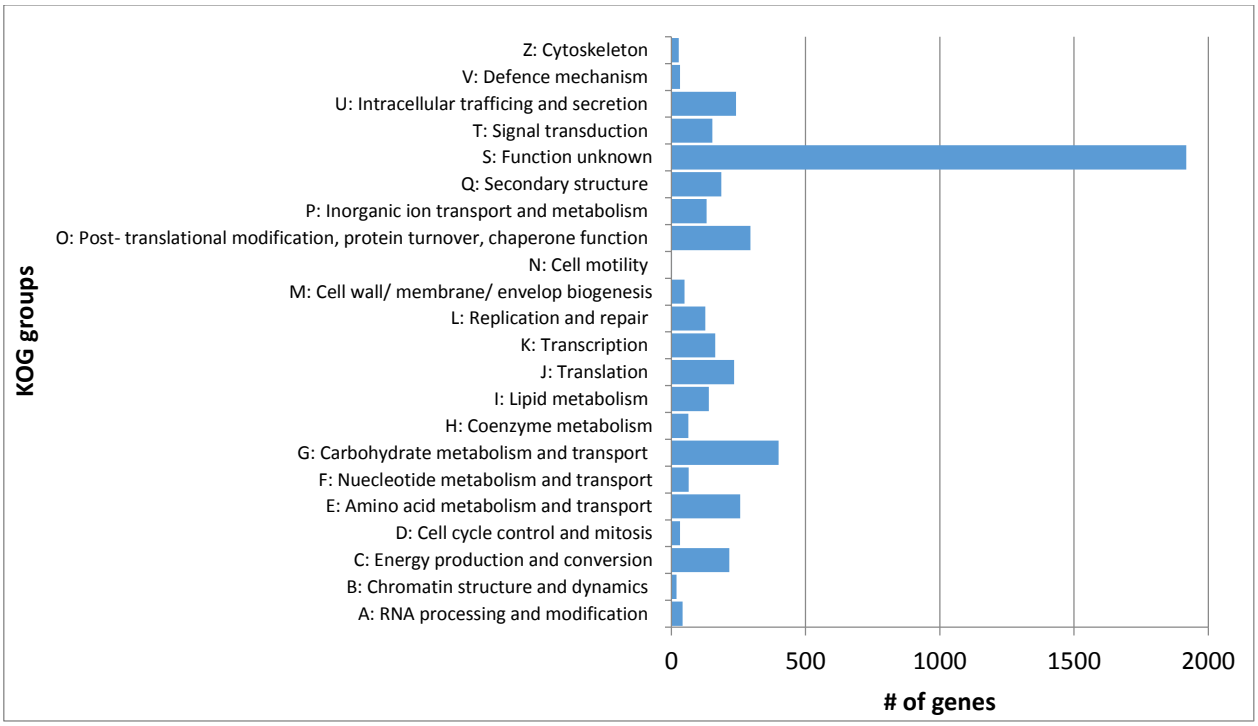
A total of 11,195 protein-encoding genes were predicted using Augustus [34]. Functional annotation showed a total of 8,575 (69.9 % of predicted genes) sequences with predicted function using Blastx [36]. Then, InterProScan [50] and Gene Ontology (GO) [51] permitted the annotation of 7,825 sequences with GO terms, whilst the remaining genes were annotated as hypothetical (25%) and unknown function proteins (12.9%). The statistics of the genome annotation are shown in Table 2. A total of 4,207 genes were assigned to the KOG [52] categories, most of them (60%) were assigned to one or more functional groups and the rest of genes were assigned to the function unknown group (Fig. 1).

Table 2. Gene prediction and annotation.

Attribute	**Value**
Gene predicted	11,195
Genes with function prediction	8,575
Genes with GO terms	7,825
Genes assigned to KOGs	4,789

Gene assigned to KEGG	2,645
Genes annotated as hypothetical	2,798
Genes with unknown function	1,444

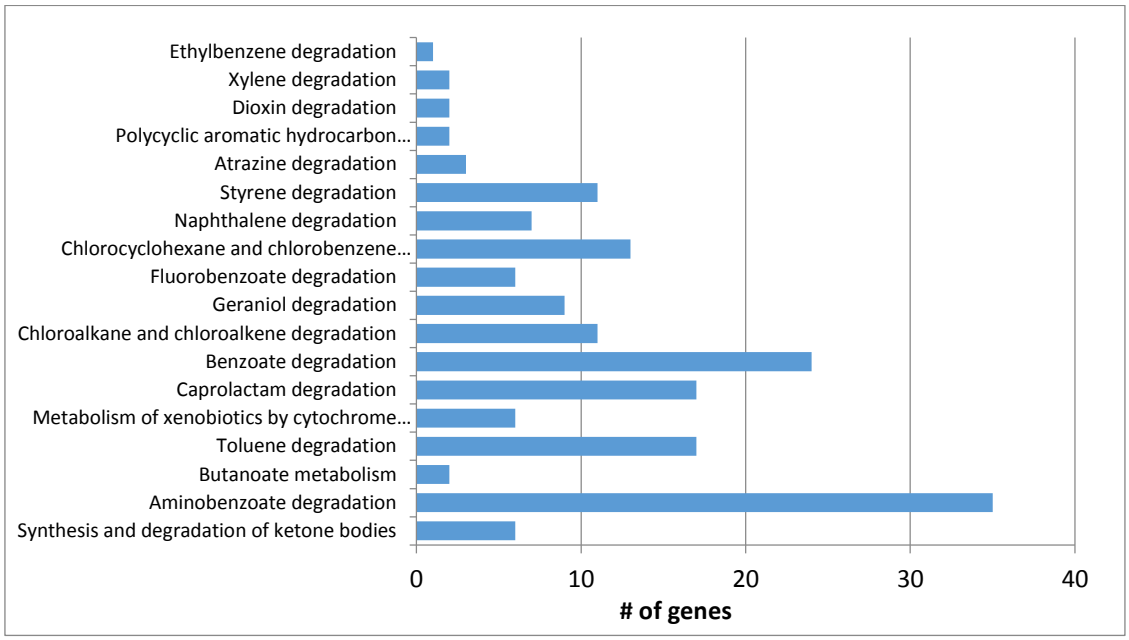
Figure 1. **Number of genes associated with general KOG functional categories.** The bars represent the number of genes related to each KOG category. Most genes were classified in the function unknown category (S), following by genes related to carbohydrate metabolism and transport.



KEGG pathway analysis assigned an enzyme code to 2,645 (23.6 %) genes and revealed specific genes involved in the pathways of hydrocarbon degradation. These hydrocarbons are chloroalkane/alkene, chorocyclohexane and chlorobenzene, benzoate, aminobenzoate,

fluorobenzoate, toluene, caprolactam, geraniol, naphthalene, styrene, atrazine, dioxin, xylene, ethylbenzene and polycyclic aromatic hydrocarbons. Also, the analysis revealed the presence of genes involved in metabolism of xenobiotics by cytochrome P450 and in synthesis and degradation of ketone bodies. These results are shown in figure 2.

Figure 2. Distribution of the hydrocarbon degradation genes in KEEG pathways. The bars represent the number of genes mapped in KEEG pathways related to hydrocarbons degradation. Most of the genes were mapped to the benzoate and its derivate compounds as aminobenzoate and fluorobenzoate.



The total number of non-coding repetitions was found using RepeatMasker [33] and was of 1.35 % most repetitions were found to be simple repeats (0.89 %) and low complexity (0.25 %). The complete report of the annotation results for the non-coding repeats sequences can be seen in the supplementary table 1.

Phylogenetic analysis confirms the identity of the strain used

The phylogenetic analysis was made using the genic regions *LSU*, *ITS* and *TEF* and the tree obtained is shown in the Figure 3. It can be observed that the environmental strain HDO1 used in this study clustered with the clinical strain IHEM14464 with good support, and that they are the sister group of *Trichinella spiralis* CBS635.78. The whole group is contained within the *wardamycopsis* lineage described by (Sandoval-Denis, M. *et al.*) [43].

Comparative genomics

The reads obtained from the sequencing were mapped to the draft sequence of the clinical strain IHEM 14464 using Bowtie 2, resulting in an overall alignment of 92.75%. Draft sequences were aligned using the Quast 4.0 web platform (http://quast.bioinf.spbau.ru/) and 88,1 % of the genome sequence of HDO1 strain aligned with the sequence of IHEM 14462 strain (all results are reported in Table S3). The MAUVE [48] alignment showed a high level of similarity between the clinical and the environmental strains (Figure 4). A total of 508 local collinear blocks (LCBs) that correspond to the homologous regions that are shared by the two sequences were found and a few of them are in reverse orientation after eight reordered cycles. Additionally, a total of 634,467 SNPs between the sequences were found. The completed results of the MAUVE [48] analysis are shown in table S4. Genomic features of the comparison between both strains are shown in Table 3.

Figure 3. Phylogenetic Analysis of *S. apiospermum* HDO1. Estimated relationships of *S. apiospermum* HDO1 with *S. apiospermum* IHEM 14462 and other species from the *Microascaceae* family. The tree shows the concatenated analysis of the Internal Transcribed Spacer, the Large Subunit and the Elongation factor gene regions**.** Sequences

from reference strains were used (Table S2). Support values are Bootstrap support values (Maximum Likelihood).

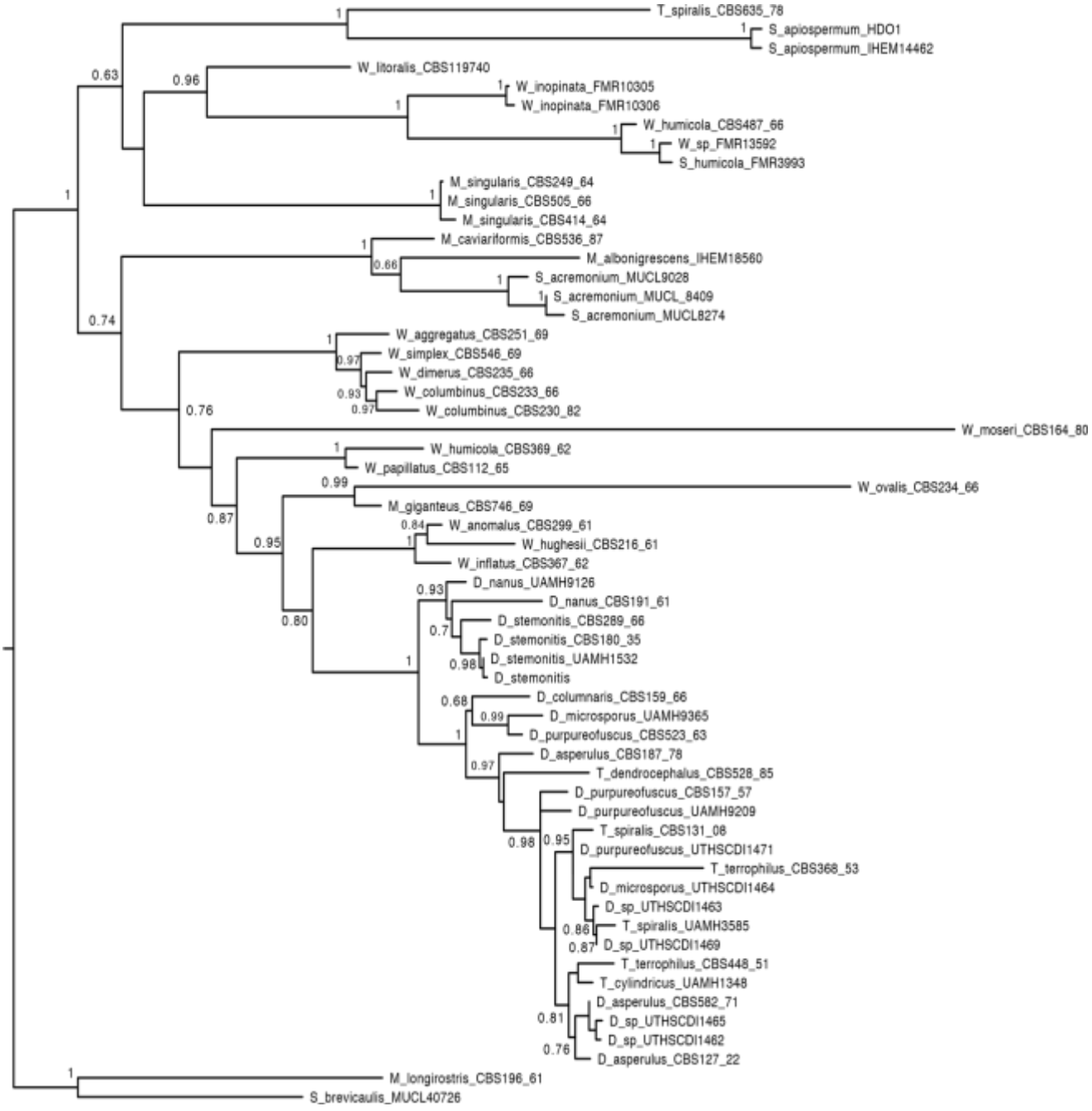
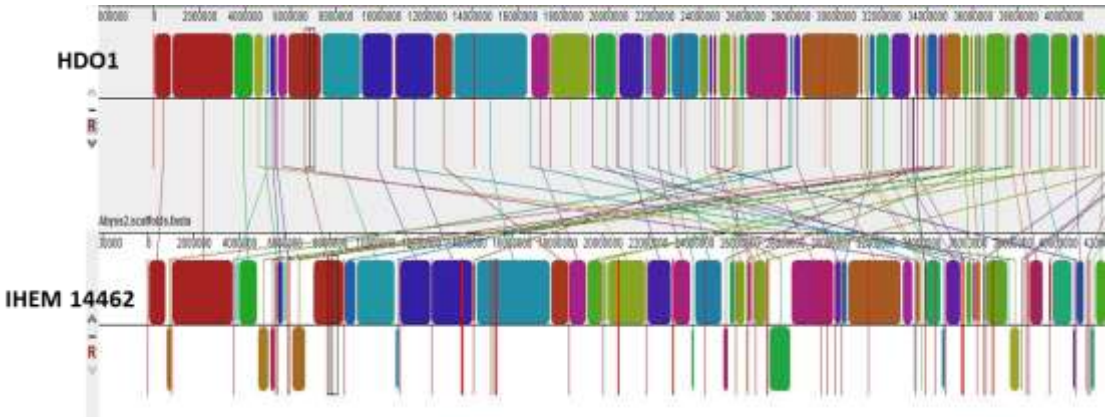


Table 3. Genomic features comparison between HDO1 strain and IHEM 14462 strain[22]. Key features from genomic assembly of *S. apiospermum* HDO1 and *S.*

apiospermum IHEM 14462 shows significant differences in the genome size, number of genes and proteins predicted.

Parameter	IHEM 14462	HDO1
Size (Mb)	43.44	44.26
Content G-C (%)	50.4	49.91
Predicted genes	10,919	11,195
Predicted proteins	8.375	8,575

Figure 4. **MAUVE [48] alignment of draft genome sequence of HDO1 strain and draft genome sequence of IHEM 14462 strain**. The figure represents the locally contiguous blocks (LCBs) that both sequences share and these blocks are connected by lines to show its positions in the genomes. Up is visualized the sequence of HDO1 strain and down the sequence re-ordered from IHEM 14462 strain [22]. Blocks that are shown below indicate regions that have the reverse sequence related to the HDO1 sequence.



Nucmer [49] comparison at the nucleotide level is shown in figure 5. This analysis allowed revealing that a high number of forward matches are in the greatest scaffolds of HDO1 genome sequence and reverse matches are more common in the smallest scaffolds. Promer [49] analysis showed similar pattern between sequences, this means that the differences and similarities of the nucleotides remain when these are translate to amino acid. The Promer comparison at the protein level is shown in figure 6. Nucmer and Promer plots also permitted to reveal rearrangements, insertions and deletions between both sequences.

Figure 5. Dot plot analysis at the nucleotide level comparing the HDO1 and IHEM14462 strains. It shows the alignment of the genome sequence of IHEM 14462 strains (y axis) against HDO1 genome sequence (x axis). The red color lines and dots represent the forward matches between the both genome sequences while the blue color ones represent reverse matches.

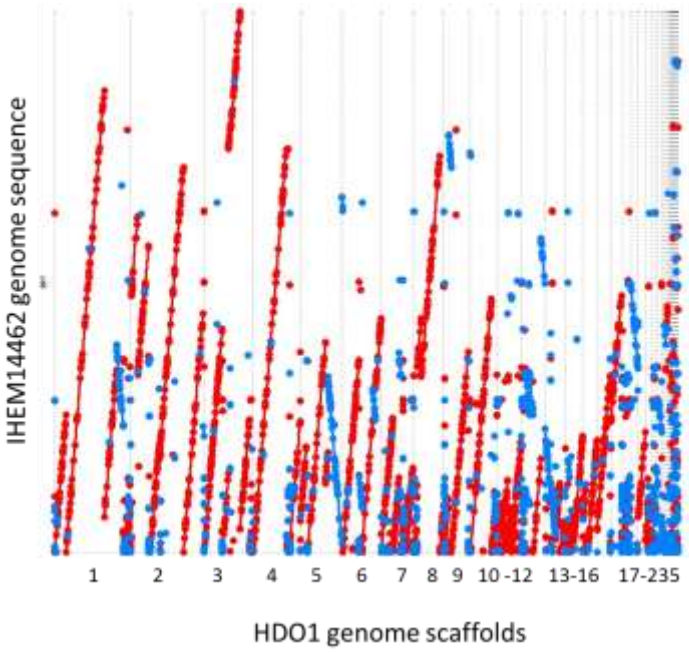
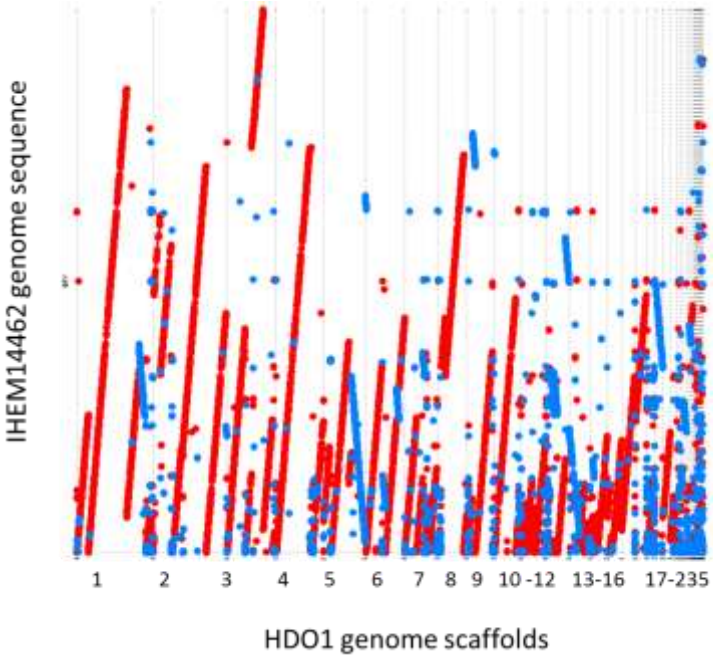


Figure 6. **Dot plot analysis at the protein level comparing the HDO1 and IHEM14462 strains.** It shows the alignment of the genome sequence translated of IHEM 14462 strains (y axis) against HDO1 genome sequence (x axis). The red color lines and dots represent the forward matches between the both genome sequences while the blue color ones represent reverse matches.



Genes involved in hydrocarbon biodegradation pathways

Several genes involved in hydrocarbon biodegradation pathways were annotated in the genome. In Table 4 the genes already reported in the clinical strain [22] are shown. Data revealed that some genes are involved in several degradation pathways, principally corresponding to aromatic hydrocarbon metabolism (polycyclic aromatic hydrocarbons and phenolic compounds) and cytochrome P450 system. The number of these genes annotated for each strain can also be seen in the table and these values show a higher number of genes

in the environmental strain HDO1. The genes solely found in the draft sequence of the strain HDO1 are reported in table 5. These genes comprise genes belonging to the aromatic hydrocarbons degradation pathways and they complete some of the pathways in which genes found in both strains are also involved. Genes involved in the degradation of others organic compounds like toluene, lignin and xylenol were found (Table 5).

Table 4. Annotated genes involved in hydrocarbons degradation pathways.

Gene	Pathway	# of genes in HDO1	# of genes in IHEM 14462 [22]
Cytochrome P450 monooxygenase (EC:1.14.13.12)	PAHs degradation, alcane biodegradation [53, 54]	79	44
Phenol hydroxylase	Phenol degradation [17]	4	4
Epoxide hydrolase (EC:3.3.2.9)	PAHs degradation [55]	3	2
Oxidoreductase	Organic compund metabolism [56]	13	8
Salicylate hydroxylase (EC:1.14.13.24)	Naphthalene degradation [57]	13	4
Laccase (EC:1.10.3.2)	PAHs degradation [58]	2	2
Catechol 1,2-dioxygenase (EC:1.13.11.1)	Phenol degradation [17]	4	4
2,4-dichlorophenol 6-monooxygenase (EC:1.14.13.7; EC:1.14.13.20)	Chlorinated phenols degradation [59]	5	5
2,3-dihydroxybenzoate decarboxylase (EC:4.1.1.46)	2,3-dihydroxybenzoate degradation [60]	1	1
Carboxy-cis,cis-muconate cyclase	Phenol degradation [17]	4	2
Phenylacetate 2-hydroxylase	Homogentisate degradation [61]	1	2
2-nitropropane dioxygenase (EC:1.13.12.16)	Nitroalquene oxidation [62]	4	4

Biphenyl-2,3-diol 1,2-dioxygenase	Byphenyl degradation [63]	1	2
Dienelactone hydrolase	Chloroaromatic degradation [64]	2	7
Vanillyl-alcohol oxidase (EC:1.1.3.38)	Aromatic degradation [65]	4	6
Cyclopentanone 1,2-monooxygenase (EC:1.14.13.8; EC:1.14.13.16)	Cyclopentanol degradation [66]	2	2
Tyrosinase	Phenolic compounds degradation [67]	1	3
Total number of genes		143	102

Table 5. Annotated genes found only in the HDO1 strain

Genes in HDO1	Pathway	# of genes
3-oxoadipate enol-lactonase	Phenol degradation [68]	5
5-carboxymethyl-2-hydroxymuconate isomerase	Homoprotocatechuate degradation pathway [69]	1
Trihydroxytoluene oxygenase	2,4-dinitrotoluene degradation [70]	1
Benzoate 4-monooxygenase (EC:3.6.1.3)	Benzoate degradation [18]	5
Diphenol oxidase	Phenolic compounds degradation [71]	1
Cyclohexanone monooxygenase (EC:1.14.13.8)	Cyclohexane degradation [72]	4
Lignostilbene dioxygenase (EC:1.13.11.43)	Lignin degradation [73]	2
Gentisate 1,2-dioxygenase	PAHs degradation [74]	2
2-keto-4-pentenoate hydratase (EC:3.7.1.5)	Benzoate degradation [75]	1
carboxymuconolactone decarboxylase	Protocatechuate degradation [76]	1
3-(3-hydroxy-phenyl)propionate 3-hydroxycinnamic acid hydroxylase	Phenyl propionate degradation [77]	2
3-hydroxybenzoate 6-	Xylenol [78] and 3-	1

hydroxylase	hydroxybenzoate degradation [79]	
3-hydroxyisobutyrate dehydrogenase (EC:1.1.1.44; EC:2.1.1.43)	Aromatic compounds metabolism [80]	5
Total number of genes		30

DISCUSSION

A high-quality draft genome sequence of the first environmental isolate of the fungus *Scedoporium apiospermum* was obtained. The assembly features obtained for the draft sequence were similar to other fungal genome sequence projects [22, 81, 82]. A robust gene annotation was made with 69.9 % of the predicted sequences with a possible function assigned with the nr database of NCBI [36]. Support for our protein dataset was obtained by Gene Ontology terms [51] (69.9 %), KOG classifications [39] (42 %) and KEGG analysis [41] (23 %). Although most of the proteins were assigned some function, it is important to notice that the greatest group of these was classified in the unknown function group in KOG classification. Phylogenetic analysis confirmed the identity of our environmental strain as *Scedosporium apiospermum*.

A thorough comparative analysis showed some important differences between the genome draft sequences of the clinical and the environmental strain sequenced here. These differences were evident in the genome size of the assemblies and the number of predicted gene. Indeed, our assembly had 783.135 bp (1.77 % of genome size) and 276 coding sequences more than the clinical strain. The remarkable difference in the number of annotated genes involved in hydrocarbons degradation pathways could be attributed to the pipeline followed to annotate genes. For the clinical strain the CDSs found were annotated against TrEmbl database [83] that only comprises UniProtKB/Swiss-Prot

(http://www.uniprot.org/), while in this study, we used nr (non-redundant protein sequences) database of NCBI (https://www.ncbi.nlm.nih.gov/protein) which has a wider coverage because it comprises sequences obtained from another databases like GenPept, TPA (Third-Party Annotation Sequences), PIR (The Protein Information Resource), PRF (Protein Research Foundation), PDB (Protein Data Bank Protein), NCBI RefSeq and UniProtKB/Swiss-Prot (http://www.uniprot.org/). Since the repetitive elements of the genome were estimated as only 1.35 %, it is highly probable that the difference in size can be attributed to some of the elements involved in functional categories.

Although the assembly was robust with only 186 scaffolds containing the 44.2 Mb of the genome, and we could detect 8,575 genes, we still do not have a complete finished genome. This was evidenced when we compared the clinical and the environmental strains and we observed a high number of inversions, rearrangements and translocations, at the nucleotide level as well as at the amino acid level. Among other purposes, a finished genome is required to obtain an a complete reference genome with an ordered sequence to make a comparative or global genome study [84]. To finish our sequence genome we have to consider that this process is divided in three steps: gap closure, assembly validation and refinement [85]. To achieve the gap closure of the whole sequence some strategies like directed-PCR and primer-walking approaches can be used [86] or we can sequence long reads using Pacific Biosciences [87] to obtain longer fragments of the genome that can help to assemble the draft scaffolds.

Despite the previous mentioned limitation, our main objective was achieved: the complete annotation of the genome and, particularly, of the genes belonging to a major class of protein families involved in fungal catabolism of organic pollutants. We could identify genes coding for proteins that have the ability to oxidize aromatic compounds like

dioxygenases or monooxygenases. Among these, we could predict dioxygenases such as 2-nitropropane dioxygenase, extracellular dioxygenase (EC:1.13.11), gentisate 1,2-dioxygenase, intradiol ring-cleavage dioxygenase (EC:1.13.11), lignostilbene dioxygenase (EC:1.13.11.43), catechol 1,2-dioxygenase (EC:1.13.11.1), Biphenyl-2,3-diol 1,2-dioxygenase, aromatic ring-opening dioxygenase, and 4-hydroxyphenylpyruvate dioxygenase (EC:1.13.11.27). These enzymes have great importance because, along with NADH-dependent flavin reductase and [2Fe-2S] redox centers, they catalyze the transformation of several aromatic compounds to dihydrodiols [88], allowing the complete mineralization of these compounds to CO_2 and H_2O (with the participation of other specific enzymes). Another enzyme family identified among the annotated genes was cytochrome P450. These enzymes have an interesting catabolic potential because they do not have substrate specificity and can catalyze epoxidation and hydroxylation of several organic pollutants like dioxins, nonylphenol and PAHs [89]. Genes for extracellular proteins like laccases and tyrosinase (known as phenoloxidase enzymes), which have the ability to degrade several groups of organic compounds due to their non-specificity action, were annotated in the genome. These enzymes produce organic radicals beyond one electron abstraction; those free radicals can be transformed by several reactions that include the ether cleavage in dioxins, quinone formations from PAHs and chlorophenols [90]. These extracellular enzymes are extremely important because of their potential in biotechnological applications [91, 92]. Moreover, several oxidoreductases, hydrolases, dehydroxylases, isomerases and transferases were also predicted in the studied strain. However, extracellular enzymes such as lignin and manganese peroxidases could not be identified yet.

Catabolic proteins of *Scedosporium apiospermum* involved in phenol, p-cresol and phenylbenzoate degradation pathway previously reported by (Clauβen and Schmidt) [17, 18] like phenol 2-monooxygenase and cathecol 1,2 dioxygenase were identified. However, hydroquinone hydroxylase, 4-hydroxybenzoate 3-hydroxylase, hydroxiquinone 1,2 dioxygenase, protocatechuate 3,4 dioxygenase and maleylacetate reductase could not be found, suggesting that these proteins could be classified among the proteins annotated as hypothetical or with an unknown function or that they can be in the gap regions of the genome assembly.

The structural and functional information of the genome sequence of *S. apiospermum* has allowed advancing in the understanding of the ability of this fungus to degrade several kinds of xenobiotic compounds and offers an opportunity to propose its use or its enzymes in bioremediation or bioaugmentation processes.

Acknowledgements

Acknowledgements to the School of Sciences, Universidad de los Andes, for the funding provided to the project ‘Sequencing of the whole genome of fungi strain capable of degrading hydrocarbons; to Departamento Administrativo de Ciencia, Tecnología e Innovación Colciencias, project: “Hydrocarbons biodegradation using a microbial consortia (fungi-bacteria) immobilized on activated carbon’, and for “beca-pasantía joven investigador” from “young researchers and innovators program 2014”. Announcement 645-2014. Code 0017-2013.

References

1. Atlas, R.M., *Microbial degradation of petroleum hydrocarbons: an environmental perspective.* Microbiological reviews, 1981. **45**(1): p. 180.
2. Baker, S.E., et al., *Fungal genome sequencing and bioenergy.* Fungal Biology Reviews, 2008. **22**(1): p. 1-5.
3. Haritash, A. and C. Kaushik, *Biodegradation aspects of polycyclic aromatic hydrocarbons (PAHs): a review.* Journal of hazardous materials, 2009. **169**(1): p. 1-15.
4. Leahy, J.G. and R.R. Colwell, *Microbial degradation of hydrocarbons in the environment.* Microbiological reviews, 1990. **54**(3): p. 305-315.
5. April, T.M., J. Foght, and R. Currah, *Hydrocarbon-degrading filamentous fungi isolated from flare pit soils in northern and western Canada.* Canadian Journal of Microbiology, 1999. **46**(1): p. 38-49.
6. Semple, K.T., R.B. Cain, and S. Schmidt, *Biodegradation of aromatic compounds by microalgae.* FEMS Microbiology letters, 1999. **170**(2): p. 291-300.
7. Smith, M.R., *The biodegradation of aromatic hydrocarbons by bacteria*, in *Physiology of Biodegradative Microorganisms*. 1991, Springer. p. 191-206.
8. Grady Jr, C.L., *Biodegradation of toxic organics: status and potential.* Journal of Environmental Engineering, 1990. **116**(5): p. 805-828.
9. Samanta, S.K., O.V. Singh, and R.K. Jain, *Polycyclic aromatic hydrocarbons: environmental pollution and bioremediation.* TRENDS in Biotechnology, 2002. **20**(6): p. 243-248.
10. Providenti, M.A., H. Lee, and J.T. Trevors, *Selected factors limiting the microbial degradation of recalcitrant compounds.* Journal of industrial Microbiology, 1993. **12**(6): p. 379-395.
11. Hickey, W.J., S. Chen, and J. Zhao, *The phn island: a new genomic island encoding catabolism of polynuclear aromatic hydrocarbons.* Frontiers in microbiology, 2012. **3**: p. 125.
12. Hickey, W.J., *Development of tools for genetic analysis of phenanthrene degradation and nanopod production by Delftia sp. Cs1-4.* Frontiers in microbiology, 2011. **2**: p. 187.
13. Desai, C., H. Pathak, and D. Madamwar, *Advances in molecular and "-omics" technologies to gauge microbial communities and bioremediation at xenobiotic/anthropogen contaminated sites.* Bioresource Technology, 2010. **101**(6): p. 1558-1569.
14. Reddy, T., et al., *The Genomes OnLine Database (GOLD) v. 5: a metadata management system based on a four level (meta) genome project classification.* Nucleic acids research, 2014: p. gku950.
15. Gilgado, F., et al., *Heterothallism in Scedosporium apiospermum and description of its teleomorph Pseudallescheria apiosperma sp. nov.* Medical mycology, 2010. **48**(1): p. 122-128.
16. Kaltseis, J., et al., *Ecology of Pseudallescheria and Scedosporium species in human-dominated and natural environments and their distribution in clinical samples.* Medical Mycology, 2009. **47**(4): p. 398-405.
17. Claußen, M. and S. Schmidt, *Biodegradation of phenol and p-cresol by the hyphomycete Scedosporium apiospermum.* Research in microbiology, 1998. **149**(6): p. 399-406.

18. Clauβen, M. and S. Schmidt, *Biodegradation of phenylbenzoate and some of its derivatives by Scedosporium apiospermum.* Research in microbiology, 1999. **150**(6): p. 413-420.
19. Reyes-César, A., et al., *Biodegradation of a mixture of PAHs by non-ligninolytic fungal strains isolated from crude oil-contaminated soil.* World Journal of Microbiology and Biotechnology, 2014. **30**(3): p. 999-1009.
20. Sandoval, C., J.C. Pijarán, and M. Vives-Florez, *Remoción de hidrocarburos por hongos filamentosos.* Universidad de los Andes (Master's Thesis), 2012.
21. Guarro, J., et al., *Scedosporium apiospermum: changing clinical spectrum of a therapy-refractory opportunist.* Medical Mycology, 2006. **44**(4): p. 295-327.
22. Vandeputte, P., et al., *Draft genome sequence of the pathogenic fungus Scedosporium apiospermum.* Genome announcements, 2014. **2**(5): p. e00988-14.
23. Vasco, M.F., et al., *Recovery of mitosporic fungi actively growing in soils after bacterial bioremediation of oily sludge and their potential for removing recalcitrant hydrocarbons.* International Biodeterioration & Biodegradation, 2011. **65**(4): p. 649-655.
24. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data.* Bioinformatics, 2014: p. btu170.
25. Andrews, S.F. and Q. Fast, *A quality control tool for high throughput sequence data. 2010*, 2015.
26. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nature methods, 2012. **9**(4): p. 357-359.
27. Li, H., et al., *The sequence alignment/map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-2079.
28. Milne, I., et al., *Tablet—next generation sequence assembly visualization.* Bioinformatics, 2010. **26**(3): p. 401-402.
29. Simpson, J.T., et al., *ABySS: a parallel assembler for short read sequence data.* Genome research, 2009. **19**(6): p. 1117-1123.
30. Boetzer, M., et al., *Scaffolding pre-assembled contigs using SSPACE.* Bioinformatics, 2011. **27**(4): p. 578-579.
31. Boetzer, M. and W. Pirovano, *Toward almost closed genomes with GapFiller.* Genome biology, 2012. **13**(6): p. 1.
32. Gurevich, A., et al., *QUAST: quality assessment tool for genome assemblies.* Bioinformatics, 2013. **29**(8): p. 1072-1075.
33. Smit, A.F., R. Hubley, and P. Green, *RepeatMasker*. Published on the web at http://www. repeatmasker. org, 1996.
34. Stanke, M., et al., *AUGUSTUS: a web server for gene finding in eukaryotes.* Nucleic acids research, 2004. **32**(suppl 2): p. W309-W312.
35. Conesa, A., et al., *Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.* Bioinformatics, 2005. **21**(18): p. 3674-3676.
36. Altschul, S.F., et al., *Basic local alignment search tool.* Journal of molecular biology, 1990. **215**(3): p. 403-410.
37. Consortium, G.O., *The Gene Ontology (GO) database and informatics resource.* Nucleic acids research, 2004. **32**(suppl 1): p. D258-D261.
38. Tatusov, R.L., et al., *The COG database: a tool for genome-scale analysis of protein functions and evolution.* Nucleic acids research, 2000. **28**(1): p. 33-36.

39. Koonin, E.V., et al., *A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes.* Genome biology, 2004. **5**(2): p. 1.
40. Powell, S., et al., *eggNOG v3. 0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges.* Nucleic acids research, 2012. **40**(D1): p. D284-D289.
41. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic acids research, 2000. **28**(1): p. 27-30.
42. Sandoval-Denis, M., et al., *Scopulariopsis, a poorly known opportunistic fungus: spectrum of species in clinical samples and in vitro responses to antifungal drugs.* Journal of clinical microbiology, 2013. **51**(12): p. 3937-3943.
43. Sandoval-Denis, M., et al., *Phylogeny and taxonomic revision of Microascaceae with emphasis on synnematous fungi.* Studies in Mycology, 2016. **83**: p. 193-233.
44. Katoh, K. and H. Toh, *Parallelization of the MAFFT multiple sequence alignment program.* Bioinformatics, 2010. **26**(15): p. 1899-1900.
45. Stamatakis, A., *RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models.* Bioinformatics, 2006. **22**(21): p. 2688-2690.
46. Miller, M.A., W. Pfeiffer, and T. Schwartz. *Creating the CIPRES Science Gateway for inference of large phylogenetic trees*. in *Gateway Computing Environments Workshop (GCE), 2010*. 2010. IEEE.
47. Lanfear, R., et al., *PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses.* Molecular biology and evolution, 2012. **29**(6): p. 1695-1701.
48. Darling, A.E., B. Mau, and N.T. Perna, *progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement.* PloS one, 2010. **5**(6): p. e11147.
49. Kurtz, S., et al., *Versatile and open software for comparing large genomes.* Genome biology, 2004. **5**(2): p. 1.
50. Zdobnov, E.M. and R. Apweiler, *InterProScan–an integration platform for the signature-recognition methods in InterPro.* Bioinformatics, 2001. **17**(9): p. 847-848.
51. Ashburner, M., et al., *Gene Ontology: tool for the unification of biology.* Nature genetics, 2000. **25**(1): p. 25-29.
52. Yong Jung, W.S.L., Sang; Wook Kim, Chul; Kim, Hyun-Soon; Ran Min, Sung; Moon, Jae Sun; Kwon, Suk-Yoon; Jeon, Jae-Heung; Sun Cho, Hye figshare., *Eukaryotic Orthologous Groups (KOG) classification of the assembled loci.* figshare., 2014.
53. Martinez, D., et al., *Genome sequence of the lignocellulose degrading fungus Phanerochaete chrysosporium strain RP78.* Nature biotechnology, 2004. **22**(6): p. 695-700.
54. Vatsyayan, P., et al., *Broad substrate cytochrome P450 monooxygenase activity in the cells of Aspergillus terreus MTCC 6324.* Bioresource technology, 2008. **99**(1): p. 68-75.
55. Cerniglia, C.E., et al., *Fungal transformation of naphthalene.* Archives of microbiology, 1978. **117**(2): p. 135-143.
56. Asgher, M., et al., *Recent developments in biodegradation of industrial pollutants by white rot fungi and their enzyme system.* Biodegradation, 2008. **19**(6): p. 771-783.

57. You, I.S., D. Ghosal, and I.C. Gunsalus, *Nucleotide sequence analysis of the Pseudomonas putida PpG7 salicylate hydroxylase gene (nahG) and its 3'-flanking region.* Biochemistry, 1991. **30**(6): p. 1635-1641.
58. Novotný, Č., et al., *Ligninolytic fungi in bioremediation: extracellular enzyme production and degradation rate.* Soil Biology and Biochemistry, 2004. **36**(10): p. 1545-1551.
59. Bollag, J., C. Helling, and M. Alexander, *2, 4-D Metabolism. Enzymic hydroxylation of chlorinated phenols.* Journal of Agricultural and Food Chemistry, 1968. **16**(5): p. 826-828.
60. Anderson, J.J. and S. Dagley, *Catabolism of tryptophan, anthranilate, and 2, 3-dihydroxybenzoate in Trichosporon cutaneum.* Journal of bacteriology, 1981. **146**(1): p. 291-297.
61. Mingot, J.M., M.A. Peñalva, and J.M. Fernández-Cañón, *Disruption of phacA, an Aspergillus nidulans gene encoding a novel cytochrome P450 monooxygenase catalyzing phenylacetate 2-hydroxylation, results in penicillin overproduction.* Journal of Biological Chemistry, 1999. **274**(21): p. 14545-14550.
62. Kido, T. and K. Soda, *Oxidation of anionic nitroalkanes by flavoenzymes, and participation of superoxide anion in the catalysis.* Archives of biochemistry and biophysics, 1984. **234**(2): p. 468-475.
63. Yang, X., et al., *Purification, characterization, and substrate specificity of two 2, 3-dihydroxybiphenyl 1, 2-dioxygenase from Rhodococcus sp. R04, showing their distinct stability at various temperature.* Biochimie, 2008. **90**(10): p. 1530-1538.
64. Schlömann, M., E. Schmidt, and H. Knackmuss, *Different types of dienelactone hydrolase in 4-fluorobenzoate-utilizing bacteria.* Journal of bacteriology, 1990. **172**(9): p. 5112-5118.
65. Fraaije, M.W., M. Pikkemaat, and W. Van Berkel, *Enigmatic Gratuitous Induction of the Covalent Flavoprotein Vanillyl-Alcohol Oxidase in Penicillium simplicissimum.* Applied and environmental microbiology, 1997. **63**(2): p. 435-439.
66. Iwaki, H., et al., *Cloning and characterization of a gene cluster involved in cyclopentanol metabolism in Comamonas sp. strain NCIMB 9872 and biotransformations effected by Escherichia coli-expressed cyclopentanone 1, 2-monooxygenase.* Applied and environmental microbiology, 2002. **68**(11): p. 5671-5684.
67. Ikehata, K. and J.A. Nicell, *Characterization of tyrosinase for the treatment of aqueous phenols.* Bioresource Technology, 2000. **74**(3): p. 191-199.
68. Harwood, C.S. and R.E. Parales, *The β-ketoadipate pathway and the biology of self-identity.* Annual Reviews in Microbiology, 1996. **50**(1): p. 553-590.
69. Roper, D.I. and R.A. Cooper, *Purification, some properties and nucleotide sequence of 5-carboxymethyl-2-hydroxymuconate isomerase of Escherichia coli C.* FEBS letters, 1990. **266**(1-2): p. 63-66.
70. Johnson, G.R., R.K. Jain, and J.C. Spain, *Properties of the trihydroxytoluene oxygenase from Burkholderia cepacia R34: an extradiol dioxygenase from the 2, 4-dinitrotoluene pathway.* Archives of microbiology, 2000. **173**(2): p. 86-90.
71. Williamson, P.R., *Biochemical and molecular characterization of the diphenol oxidase of Cryptococcus neoformans: identification as a laccase.* Journal of Bacteriology, 1994. **176**(3): p. 656-664.

72. Sheng, D., D.P. Ballou, and V. Massey, *Mechanistic studies of cyclohexanone monooxygenase: chemical properties of intermediates involved in catalysis.* Biochemistry, 2001. **40**(37): p. 11156-11167.
73. Kamoda, S., et al., *Purification and Some Properties of Lignostilbene-a,/i-dioxygenase Responsible for the Ca—Cp Cleavage of a Diarylpropane Type Lignin Model Compound from Pseudomonas sp. TMY1009.* Agricultural and biological chemistry, 1989. **53**(10): p. 2757-2761.
74. Luo, S., et al., *Site-directed mutagenesis of gentisate 1, 2-dioxygenases from Klebsiella pneumoniae M5a1 and Ralstonia sp. strain U2.* Microbiological research, 2006. **161**(2): p. 138-144.
75. Perez-Pantoja, D., et al., *Metabolic reconstruction of aromatic compounds degradation from the genome of the amazing pollutant-degrading bacterium Cupriavidus necator JMP134.* FEMS microbiology reviews, 2008. **32**(5): p. 736-794.
76. Eulberg, D., et al., *Characterization of a protocatechuate catabolic gene cluster from Rhodococcus opacus 1CP: evidence for a merged enzyme with 4-carboxymuconolactone-decarboxylating and 3-oxoadipate enol-lactone-hydrolyzing activity.* Journal of bacteriology, 1998. **180**(5): p. 1072-1081.
77. Burlingame, R. and P.J. Chapman, *Catabolism of phenylpropionic acid and its 3-hydroxy derivative by Escherichia coli.* Journal of bacteriology, 1983. **155**(1): p. 113-121.
78. Poh, C.L. and R.C. Bayly, *Evidence for isofunctional enzymes used in m-cresol and 2, 5-xylenol degradation via the gentisate pathway in Pseudomonas alcaligenes.* Journal of Bacteriology, 1980. **143**(1): p. 59-69.
79. Park, M., et al., *Molecular and biochemical characterization of 3-hydroxybenzoate 6-hydroxylase from Polaromonas naphthalenivorans CJ2.* Applied and environmental microbiology, 2007. **73**(16): p. 5146-5152.
80. Masai, E., Y. Katayama, and M. Fukuda, *Genetic and biochemical investigations on bacterial catabolic pathways for lignin-derived aromatic compounds.* Bioscience, biotechnology, and biochemistry, 2007. **71**(1): p. 1-15.
81. Yew, S.M., et al., *The genome of newly classified Ochroconis mirabilis: Insights into fungal adaptation to different living conditions.* BMC genomics, 2016. **17**(1): p. 1.
82. Galagan, J.E., et al., *The genome sequence of the filamentous fungus Neurospora crassa.* Nature, 2003. **422**(6934): p. 859-868.
83. Magrane, M. and U. Consortium, *UniProt Knowledgebase: a hub of integrated protein data.* Database, 2011. **2011**: p. bar009.
84. Mardis, E., et al., *What is finished, and why does it matter.* Genome research, 2002. **12**(5): p. 669-671.
85. Nagarajan, N., et al., *Finishing genomes with limited resources: lessons from an ensemble of microbial genomes.* BMC genomics, 2010. **11**(1): p. 1.
86. Tettelin, H., et al., *Optimized multiplex PCR: efficiently closing a whole-genome shotgun sequencing project.* Genomics, 1999. **62**(3): p. 500-507.
87. English, A.C., et al., *Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology.* PloS one, 2012. **7**(11): p. e47768.
88. Mason, J.R. and R. Cammack, *The electron-transport proteins of hydroxylating bacterial dioxygenases.* Annual Reviews in Microbiology, 1992. **46**(1): p. 277-305.

89. Kasai, N., et al., *Metabolism of mono-and dichloro-dibenzo-p-dioxins by Phanerochaete chrysosporium cytochromes P450.* Applied microbiology and biotechnology, 2010. **86**(2): p. 773-780.
90. Harms, H., D. Schlosser, and L.Y. Wick, *Untapped potential: exploiting fungi in bioremediation of hazardous chemicals.* Nature Reviews Microbiology, 2011. **9**(3): p. 177-192.
91. Duran, N. and E. Esposito, *Potential applications of oxidative enzymes and phenoloxidase-like compounds in wastewater and soil treatment: a review.* Applied catalysis B: environmental, 2000. **28**(2): p. 83-99.
92. Gianfreda, L. and M.A. Rao, *Potential of extra cellular enzymes in remediation of polluted soils: a review.* Enzyme and Microbial Technology, 2004. **35**(4): p. 339-354.

Supplementary material

Figure 1S. Scaffolds size distribution

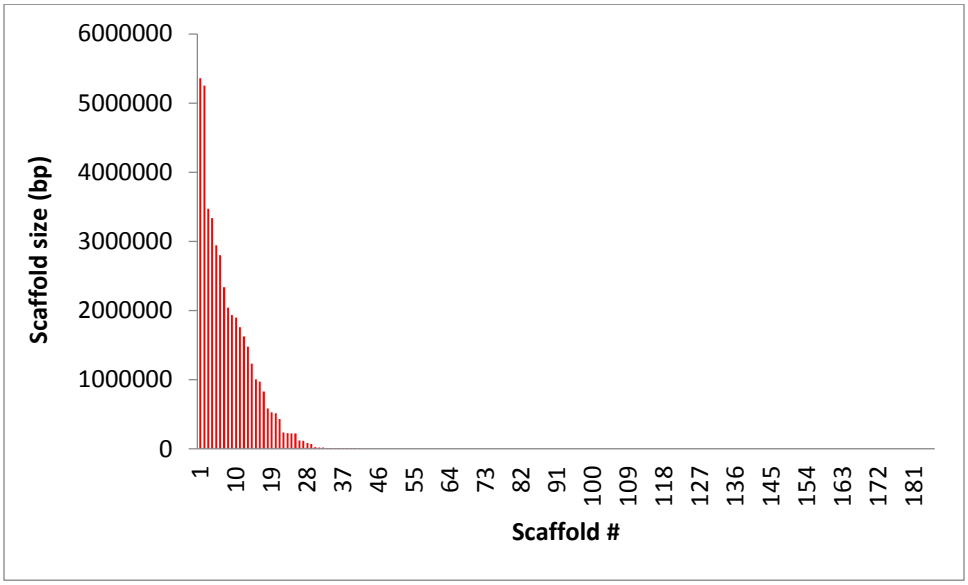


Table S1. Non-coding repeats sequences summary.

Element	Feature	Size (bp)	%
Retroelements	625	116890	0.16
SINEs:	0	0	0.00
Penelope	0	0	0.00
LINEs:	17	1563	0.00
CRE/SLACS	0	0	0.00
L2/CR1/Rex	0	0	0.00
R1/LOA/Jockey	0	0	0.00
R2/R4/NeSL	0	0	0.00
RTE/Bov-B	0	0	0.00
L1/CIN4	0	0	0.00
LTR elements:	608	115327	0.15
BEL/Pao	0	0	0.00
Ty1/Copia	147	23725	0.05
Gypsy/DIRS1	459	91432	0.19
Retroviral	0	0	0.00
DNA transposons	54	6910	0.01
hobo-Activator	5	413	0.00
Tc1-IS630-Pogo	14	2674	0.01
En-Spm	0	0	0.00
MuDR-IS905	0	0	0.00
PiggyBac	0	0	0.00
Tourist/Harbinger	13	931	0.00
Other (Mirage P-element Transib)	2	109	0.00
Rolling-circles	0	0	0.00
Unclassified:	6	536	0.00
Total interspersed repeats:		124336	0.18
Small RNA:	4644	590528	0.02
Satellites:	77	11338	0.02
Simple repeats:	13319	570227	0.89
Low complexity:	2759	140347	0.25

Table S2. Species and genes (accession numbers) used in the phylogenetic analysis.

Modified from [43].

Specie	LSU	ITS	TEF
Sedosporium apiospermum HDO1	---------	--------	------
Sedosporium apiospermum IHEM14462	--------	--------	XM_016787526
Wardomyces moseri CBS164.80	LN851049	LN850995	LN851100
Microascus albonigrescens IHEM18560	LN851004	LM652389	LN851058
Trichurus spiralis CBS635.78	LN851024	LN850977	LN851077
Microascus caviariformis CBS536.87	LN851005	LM652392	LN851059
Scopulariopsis acremonium MUCL9028	LN851001	LM652456	HG380362
Scopulariopsis acremonium MUCL8274	LN851002	LM652457	LN851056
Scopulariopsis acremonium MUCL8409	LN851003	LM652458	LN851057
Wardomyces litoralis CBS119740	LN851055	LN851000	LN851107
Wardomyces inopinata FMR10305	LN851054	LM652498	LN851106
Wardomyces inopinata FMR10306	LN850956	LN850955	LN850957
Wardomyces humicola CBS487.66	LM652554	LM652497	LN851103
Scopulariopsis humicola FMR3993	LN851052	LN850998	LN851104
Wardomyces sp. FMR13592	LN851053	LN850999	LN851105
Microascus giganteus CBS746.69	LN851045	LM652411	LN851096
Wardomyces humicola CBS369.62	LN851046	LN850993	LN851097
Wardomyces papillatus CBS112.65	LN851051	LN850997	LN851102
Wardomyces columbinus CBS233.66	LN851039	LN850990	LN851092
Wardomyces dimerus CBS235.66	LN851040	LN850991	LN851093
Wardomyces simplex CBS546.69	LN851041	LM652379	LN851094
Wardomyces columbinus CBS230.82	LN851038	LN850989	LN851091
Wardomyces aggregatus CBS251.69	LN851037	LM652378	LN851090
Wardomyces ovalis CBS234.66	LN851050	LN850996	LN851101
Wardomyces inflatus CBS367.62	LN851048	LN850994	LN851099
Wardomyces anomalus CBS299.61	LN851044	LN850992	LN851095
Wardomyces hughesii CBS216.61	LN851047	LM652496	LN851098
Trichurus dendrocephalus CBS528.85	LN851013	LN850966	LN851067
Doratomyces asperulus CBS187.78	LN851033	LN850986	LN851086
Doratomyces purpureofuscus UAMH9209	LN851018	LN850971	LN851072
Doratomyces purpureofuscus CBS 157 57	LN851031	LN850984	LN851084

Trichurus spiralis CBS 131 08	LN851021	LN850974	--------
Doratomyces microsporus UTHSCDI1464	LN851027	LN850980	LN851080
Doratomyces sp. UTHSCDI1469	LN851028	LN850981	LN851081
Trichurus terrophilus CBS368.53	LN851023	LN850976	LN851076
Doratomyces purpureofuscus UTHSCDI1471	LN851029	LN850982	LN851082
Doratomyces sp. UTHSC1463	LN851026	LN850979	LN851079
Trichurus spiralis UAMH3585	LN851025	LN850978	LN851078
Trichurus cylindricus UAMH1348	LN851012	LN850965	LN851066
Doratomyces asperulus CBS127.22	LN851006	LN850959	LN851060
Doratomyces asperulus CBS582.71	LN851007	LN850960	LN851061
Doratomyces sp. UTHSCDI1462	LN851008	LN850961	LN851062
Doratomyces sp. UTHSCDI1465	LN851009	LN850962	LN851063
Doratomyces purpureofuscus CBS523.63	LN851014	LN850967	LN851068
Doratomyces microsporus UAMH9365	LN851015	LN850968	LN851069
Doratomyces nanus UAMH9126	LN851017	LN850970	LN851071
Doratomyces stemonitis CBS289.66	LN851032	LN850985	LN851085
Doratomyces stemonitis	LN850952	LN850951	LN850953
Doratomyces stemonitis CBS180.35	LN851019	LN850972	LN851073
Doratomyces stemonitis UAMH1532	LN851020	LN850973	LN851074
Doratomyces nanus CBS191.61	LN851016	LN850969	LN851070
Doratomyces columnaris CBS159.66	LN851010	LN850963	LN851064
Microascus singularis CBS249.64	LN851034	LN850987	LN851087
Microascus singularis CBS505.66	LN851036	LN850988	LN851089
Microascus singularis CBS414.64	LN851035	LM652442	LN851088
Scopulariopsis brevicaulis MUCL40726	LN851042	LM652465	HG380363
Microascus longirostris CBS196.61	LN851043	LM652421	LM652566

1

2

Table S3. Complete genome assembly statistics reported by Quast [32].

Assembly	Scaffolds HDO1
Total length	44220274
Reference length	43437139
GC (%)	49.91
Reference GC (%)	50.36
N50	2801028
NG50	2801028
N75	1627699
NG75	1760834
L50	6
LG50	6
L75	12
LG75	11
# misassemblies	437
# misassembled contigs	31
Misassembled contigs length	42592731
# local misassemblies	669
# unaligned contigs	47 + 36 part
Unaligned length	853113
Genome fraction (%)	88.091
Duplication ratio	1.133
# N's per 100 kbp	1633.76
# mismatches per 100 kbp	1194.41
# indels per 100 kbp	165.94
Largest alignment	1121384
Total aligned length	38328450
NA50	177431
NGA50	181568
NA75	55139
NGA75	59935
LA50	65
LGA50	62
LA75	174
LGA75	164

Table 4S. Mauve alignment report of HDO1 strain as reference sequence against IHEM 14462 strain.

Number of Contigs:	176
Number HDO1 replicons:	235
Number of IHEM14462 bases:	43437139
Number of HDO1 bases:	44236179
Number of LCBs:	508
Number of Blocks:	172
Breakpoint Distance:	142
DCJ Distance:	93
SCJ Distance:	284
Number of SNPs:	634487
Number of Gaps in HDO1:	35477
Number of Gaps in IHEM14462:	35426
Total bases missed in HDO1:	4659051
Percent bases missed:	10,53%
Total bases extra in IHEM14462:	3438714
Percent bases extra:	7,92%
Number of missing chromosomes:	80
Number of extra contigs:	109
Number of Shared Boundaries:	0
Number of Inter-LCB Boundaries:	130
Contig N50:	1380919
Contig N90:	507
Min contig length:	507
Max contig length:	3171699