MYCOLOGY, 2018
VOL. 9, NO. 1, 59–69
https://doi.org/10.1080/21501203.2017.1419995

The genome sequence of the soft-rot fungus *Penicillium purpurogenum* reveals a high gene dosage for lignocellulolytic enzymes

Wladimir Mardones[a], Alex Di Genova[b,c,d,e], María Paz Cortés[b,d,e], Dante Travisany[b,d,e], Alejandro Maass[d,e,f] and Jaime Eyzaguirre[a]

[a]Facultad de Ciencias Biológicas, Universidad Andrés Bello, Santiago, Chile; [b]Facultad de Ingeniería y Ciencias, Universidad Adolfo Ibáñez, Santiago, Chile; [c]Erable Team, INRIA Grenoble, Montbonno, France; [d]Center for Mathematical Modeling, University of Chile, Santiago, Chile; [e]Center for Genome Regulation, University of Chile, Santiago, Chile; [f]Department of Mathematical Engineering, University of Chile, Santiago, Chile

ABSTRACT

The high lignocellulolytic activity displayed by the soft-rot fungus *Penicillium purpurogenum* has made it a target for the study of novel lignocellulolytic enzymes. We have obtained a reference genome of 36.2 Mb of non-redundant sequence (11,057 protein-coding genes). The 49 largest scaffolds cover 90% of the assembly, and Core Eukaryotic Genes Mapping Approach (CEGMA) analysis reveals that our assembly captures almost all protein-coding genes. RNA-seq was performed and 93.1% of the reads aligned to the assembled genome. These data, plus the independent sequencing of a set of genes of lignocellulose-degrading enzymes, validate the quality of the genome sequence. *P. purpurogenum* shows a higher number of proteins with CAZy motifs, transcription factors and transporters as compared to other sequenced *Penicillia*. These results demonstrate the great potential for lignocellulolytic activity of this fungus and the possible use of its enzymes in related industrial applications.

ARTICLE HISTORY
Received 3 October 2017
Accepted 18 December 2017

KEYWORDS
Penicillium purpurogenum; genome sequencing; lignocellulose biodegradation; CAZymes; RNA-seq; Illumina

1. Introduction

Lignocellulose is by far the most abundant renewable resource on earth, and it represents a highly valuable source of raw material for different industrial processes, among them the production of second-generation biofuels such as bioethanol (Ragauskas et al. 2006). Lignocellulose contains several polysaccharides, mainly cellulose, hemicelluloses and pectin. Hydrolysis of these polysaccharides can be achieved chemically or enzymatically, the second method having the advantage of being environmentally friendly and free of potentially poisonous side-products (Blanch 2012). The saccharification process to obtain the monosaccharide components requires a variety of enzymes particularly due to the complex structure of the hemicelluloses and pectin. These enzymes are secreted by a number of bacteria and fungi (Dehority et al. 1962; Kang et al. 2004). A thorough understanding of the lignocellulose-degrading systems is important in order to prepare enzyme cocktails that can efficiently degrade a variety of lignocellulose raw materials of different chemical composition (Rosenberg 1978). Although the study of individual enzymes (purification and characterisation) may give valuable information on their properties, it is a rather time-consuming process. The recent developments in the "omics" sciences and bioinformatics have provided novel techniques for the analysis of genomes, transcriptomes and secretomes of cells and organisms, allowing a rapid identification and characterisation of sets of enzymes (Conn 2003).

Fungi are the preferred sources of lignocellulolytic enzymes. Among them, the genera *Trichoderma* (Merino and Cherry 2007) and *Aspergillus* (Kang et al. 2004) have been the subject of detailed studies and several of their enzymes have been used for industrial applications. Less known are the enzymes from *Penicillia*. However, several members of this genus have been reported as good producers of cellulases and xylanases. A strain of *Penicillium decumbens* (Liu et al. 2013) has been used for industrial-scale cellulase production in China since 1996, and Gusakov and Sinitsyn

CONTACT Jaime Eyzaguirre jeyzaguirre@unab.cl
Wladimir Mardones and Alex Di Genova contributed equally to this work.
Supplemental data for this article can be accessed here.

(2012) have reviewed the production of cellulases by a set of *Pencillium* species, some of whose enzymes show superior performance to those of the better-known *Trichoderma*.

Our laboratory has used as a model for the study of lignocellulolytic enzymes a locally isolated strain of *Penicillium purpurogenum* (Musalem et al. 1984), which was registered in the American Type Culture Collection (ATCC) as MYA-38. This soft-rot fungus grows on a variety of lignocellulolytic natural carbon sources (i.e. sugar beet pulp, corncob, etc.) (Steiner et al. 1994). It secretes to the medium a large number of cellulose and xylan-degrading enzymes, some of which have been characterised and sequenced (Chávez et al. 2006; Ravanal et al. 2010), demonstrating a high and plastic lignocellulolytic enzyme activity, which may potentially have industrial applications in raw material processing.

In this article, we introduce the genome sequence of *P. purpurogenum* (MYA-38) generated using a hybrid approach involving second- and third-generation sequencing technologies (Illumina (Bentley et al. 2008) and PacBio (Eid et al. 2009), respectively), together with gene models and analysis of novel genes focused on lignocellulolytic activity. *P. purpurogenum* (MYA-38) was found, among the *Penicillia* analysed, to have the highest number of CAZymes. An RNA-seq analysis of the fungus grown on sugar beet pulp, and the independent sequencing of a set of genes of lignocellulose-degrading enzymes confirm the quality of the gene models. Due to the high number of CAZymes identified, we propose that this *Penicillium* strain is a powerful source of enzymes for the industrial lignocellulose biodegradation process.

2. Materials and methods

2.1. Fungal strain and culture conditions

P. purpurogenum ATCC strain MYA-38 was grown in Mandel's medium as described previously (Hidalgo et al. 1992). Liquid cultures were grown for 4 days at 28°C in an orbital shaker (200 rpm) using 1% glucose or sugar beet pulp as carbon sources. Fungus grown on glucose was used for DNA isolation and genome sequencing; cultures grown on sugar beet pulp were utilised for the transcriptome analysis.

2.2. Genome sequencing and assembly

For genomic DNA preparation, the fungus grown on glucose was filtered and frozen with liquid nitrogen in a mortar, and then powdered with a pestle. The genomic DNA was purified using the Genomic DNA Purification Kit (Thermo Fischer Scientific, USA) according to the manufacturer's instructions.

Genome sequencing was carried out in two stages and with two different technologies. First, Illumina sequencing was performed on a HiSeq2000 instrument utilising two genomic libraries: 180 bp (paired-end) and 2 Kb (mate-pair) insert sizes, with 100 bp read length. Second, PacBio sequencing technology (model RSII and P4C2 chemistry) was used to produce long reads (over 1 Kb).

The genome assembly process was carried out in two steps. First, both Illumina libraries were assembled with ALLPAHTS-LG software (version r43019) (Gnerre et al. 2011) using a 200× coverage in order to build high-quality contigs and scaffolds. Second, the Illumina assembly was scaffolded with the long PacBio reads using SSPACE-Long-Reads version 1.1 (Boetzer and Pirovano 2014).

Finally, the assembly process was validated using CEGMA (Parra et al. 2007), by aligning 248 highly conserved eukaryotic proteins to the resulting scaffolds. Since the CEGMA proteins are highly conserved, alignment methods can identify their exon–intron structures on the assembled genome, thus allowing estimation of the completeness of the assembled genome in terms of gene numbers.

2.3. RNA-seq

The fungus was grown on sugar beet pulp as carbon source for 4 days and was separated from the supernatant by filtration. The mycelium was immediately processed using the TRIzol reagent (Ambion, USA) following the manufacturer's instructions and treated with DNase I (Invitrogen) to eliminate any remnant genomic DNA. The RNA integrity was evaluated by means of the Fragment Analyzer (Advanced Analytical Technologies, Inc., USA). The library was constructed using the TruSeq Stranded mRNA LT kit (Illumina, USA), multiplexed, and sequenced in a HiScanSQ instrument (Illumina) as single reads of 101 base pairs.

Sequenced reads were mapped to the assembled genome using Tophat2 (Trapnell et al. 2009) software. Cufflinks (Trapnell et al. 2012) was used to build and quantify expression levels of RNA-seq transcripts from aligned reads.

2.4. Gene discovery

For gene discovery, a variation of the pipeline described by Haas et al. (2011) was used to annotate protein-coding genes in the assembled genome. This pipeline integrates evidences from different sources to create a consensus gene prediction, including *ab initio* protein and transcript alignments, which are integrated with EVidenceModeler (Haas et al. 2008).

The repetitive and transposable elements of the assembled genome were identified using RepeatModeler (Price et al. 2005) and TransposonPSI (transposonpsi.sourceforge.net). RepeatModeler identifies *de novo* repeat families using k-mer frequency count. These *de novo* repeats plus the latest giriRepbase (Jurka et al. 2005) were then masked using RepeatMasker (Tempel 2012). The resulting hard-masked genome was further processed using TransposonPSI which identifies transposable elements by aligning the DNA sequences to a transposon-protein database using PSI-BLAST, which were then masked using BEDtools (Quinlan and Hall 2010).

For *ab initio* gene prediction, two predictors were used: (1) AUGUSTUS (Stanke and Morgenstern (2005) (using *Aspergillus terreus* and *Aspergillus oryzae* as training sets) and (2) GeneMark-ES (Ter-Hovhannisyan et al. 2008) (self-trained). In addition, protein evidences were included using the following strategy: First, a target protein set was built by mapping the GeneMark-ES and AUGUSTUS predicted proteins against all fungal proteins from the UniProt taxonomic division (Swiss-prot + TrEMBL) using BLASTp. This set was then aligned back to the assembled genome using GenBlastG (She et al. 2011) with a cut-off *e*-value of 10^{-10}. The resulting hits were then used as seed target proteins and mapped to the genome with GeneWise (Birney et al. 2004) in order to establish their most likely position in the genome. A match was considered evidence if the GeneWise alignment represented at least 80% of the original target protein.

The consensus gene model for each locus was produced by source-weighted integration using EVidenceModeler. Weights of 5 and 1 were assigned to GeneWise and *ab initio* predictions, respectively. The final gene models were updated using RNA-seq data with Program to Assemble Spliced Alignments (PASA) (Haas et al. 2003) to include splicing-isoforms, 5′-UTR and 3′-UTR. In addition, the number of genes supported by RNA-seq data was counted.

2.5. Functional annotation

Alignments were carried out using BLASTp version 2.2.29+ and were filtered for the first 10 hits with a cut-off *e*-value of 10^{-5} as suggested in the literature (Clarke et al. 2013; Liu et al. 2013). BLASTp alignments against the non-redundant RefSeq protein database (22 January 2014) (Ye et al. 2006; Pruitt et al. 2012), UniProt (22 January 2014) (Consortium 2014) (including Swiss-Prot, fungal taxonomic division and uniref90) and KEGG (Release 69.0, 1 January 2014) (Kanehisa et al. 2002) were performed to assign general protein function profiles. Protein domains were assigned using InterProScan 5.2-45.0 (Quevillon et al. 2005) (including Pfam 27.0 (Punta et al. 2012), SUPERFAMILY 1.75 (Wilson et al. 2009), SMART 7 (Letunic et al. 2012), TIGRFAMs 13.0 (Haft et al. 2013), TMHMM 2.0c (Krogh et al. 2001), PROSITE 20.99 (Sigrist et al. 2013) and PANTHER 8.1 (Mi et al. 2013) databases). Gene ontology (GO) annotations were determined using InterProScan and TransporterTP (Li et al. 2009). Enzyme Commission codes (EC numbers) and KEGG Orthology (KO) numbers were assigned using the KEGG database and BLASTp. PRI enzyme functions were assigned using PRIAM (Enzyme rel. of 6 March 2013) (Claudel-Renard et al. 2003). Targeting and transmembrane signals were obtained from SignalP4.1 (Petersen et al. 2011), NetNES1.1 (la Cour et al. 2004), TargetP1.1 (Emanuelsson et al. 2000) and SecretomeP2.0 (Bendtsen et al. 2004). Transporter signals were determined using TransporterTP. Carbohydrate-activating enzymes (CAZymes) were predicted using dbCAN (Yin et al. 2012) (cutoff *e*-value of 10^{-5}).

2.6. Comparative genomics and phylogenetic analysis

The annotated genes from *P. purpurogenum* were compared with 21 previously annotated genomes from

members of the phylum Ascomycota (listed in Table S1). Twenty are from the family Trichocomaceae (12 *Penicillium*, 5 *Aspergillus* and 3 *Talaromyces*). A cellulolytic-enzyme secreting fungus from a different family, *Trichoderma reesei*, was used as outgroup.

The assembled and annotated genomes from the Ascomycota reference genomes were downloaded from the JGI database (http://genome.jgi.doe.gov/). All species were assessed and compared for the number of CAZy proteins and InterPro families.

Orthologous genes between the Ascomycota members and *P. purpurogenum* were predicted using the OrthoMCL program (Li et al. 2003). A total of 252,277 proteins were clustered with OrthoMCL using an *e*-value cut-off of 10^{-5} and a moderate inflation value of 2.5 for the MCL (Enright et al. 2002) cluster method. The analysis produced 19,051 clusters, which were used to build the orthologous dendrogram. In order to compute the distance between the Ascomycota genomes, a membership matrix was built using orthologous clusters as rows and the genomes as columns, assigning a value of 1 if the genome belonged to the cluster and 0 if not. The final dendrogram was computed using Manhattan distances between all-pairs of genomes.

In parallel, a phylogenomics tree was constructed considering *P. purpurogenum*, the 20 aforementioned members of the phylum Ascomycota and using *T. ressei* as outgroup. In order to build the phylogenomics tree, the most conserved of the orthologous clusters shared among the 22 species were selected. For each of these clusters, a multiple alignment of their corresponding amino acid sequences was performed using MUSCLE (Edgar 2004) v.3.8.3152. Groups with gaps greater than 10% of the alignment length were discarded and from this subset, only those where over 60% of their alignment length were fully or strongly conserved residues were kept, resulting in 147 clusters. Misaligned sections from these clusters sequences were removed using GBlocks 0.91b53 (Castresana 2000) and the resulting sequences were concatenated.

This concatenated sequence was used as an input for a Bayesian phylogenetic analysis performed using MrBayes v 3.2 (Huelsenbeck and Ronquist 2001). The analysis was carried out partitioning by gene and applying the Jones model to each partition, as it was the one with the highest posterior probability (1.0) with parameters unlinked across partitions. Two independent runs of 1,000,000 generations were conducted with tree sampling every 1000 generations. A consensus tree was built from the two runs using a burn-in value of 25%. Finally, FastTree (Price et al. 2009) was used to produce another phylogenomics tree using the aforementioned concatenated sequence to confirm MrBayes results by a different approach. FastTree was run with default parameters for amino acid sequence.

In addition, for a set of members of the *Penicillium, Aspergillus* and *Talaromyces* genera, sequences were downloaded for marker genes BenA, CaMand RPB2 (β-tubulin, calmodulin and RNA polymerase II second largest subunit, respectively) (Table S2). This set of sequences was used to perform a Blast search against *P. purpurogenum* genes.

2.7. Manual annotation of CAZymes

dBCAN searches for genes coding for glycosyl hydrolases (GH), glycosyl transferases (GT), polysaccharide lyases (PL) and carbohydrate esterases (CE) from the CAZy database were performed. From the results obtained, all CAZy genes with an *e*-value of less than 10^{-40} were curated with the help of Apollo (Lewis et al. 2002). In addition, each entry was subjected to BLASTp and CD-search (Marchler-Bauer and Bryant 2004) to manually confirm or improve the automatic annotation.

2.8. Accession numbers

Genome and RNASeq sequences were submitted to the sequence read archive (SRA) and can be downloaded from BioProject SRP055745. SRA numbers for Illumina-overlapping, Illumina-jumping, PacBio and RNA-seq reads are SRR1823665, SRR1823673, SRR1823957 and SRR1824012, respectively. Assembled genome and related annotation files can be downloaded from http://ppurdb.cmm.uchile.cl/material.

3. Results and discussion

3.1. Genome sequencing and assembly

The Illumina sequencing process produced 221.7 million reads, representing a raw coverage of 568×. The PacBio sequencing produced 975,179 reads

larger than 1 kb, representing a raw coverage of 44.25×, with an N50 of 1771 base pairs.

The assembled genome has 158 scaffolds with a length of 36.2 Mb (Table 1); 90% of the genome length is covered by only 49 scaffolds (L90 = 49), indicating that the assembly is highly continuous. The assembly length covers 92% of the estimated genome size obtained from the k-mer spectrum analysis of 39.2 Mb at a $K = 25$ scale (Figure S1). This estimated genome size, however, is not consistent with a previously reported genome length (Chávez et al. 2001b), which was obtained with contour-clamped homogeneous electric field gel electrophoresis, resulting in an estimated genome size of 21.2 Mb. We attribute this discrepancy to a lack of resolution of the larger chromosomes by pulse field electrophoresis. The estimated size of the genome (36.2 Mb) is consistent with results obtained for other related fungi, such as *Penicillium chrysogenum* (van den Berg et al. 2008) and *Aspergillus flavus* (37 Mb) (Nierman et al. 2015).

The scaffolding with PacBio reads incremented the scaffold N50 fourfold in relation to the high-coverage using only Illumina reads (Table 1). The genome was assembled with the highest sequence coverage and it is, to our knowledge, the first *Penicillium* assembled with PacBio technology.

CEGMA analyses (Table S1) indicate that 99.6% of the 248 ultra-conserved core eukaryotic genes are present in the assembled genome, and 98.4% of them were considered complete; in addition, our CEGMA numbers are in agreement with respect to the reference fungal genomes (Table S1). Further, 93.1% of the RNA-seq reads aligned to the assembled genome. Collectively, these results indicate that our assembled genome is highly continuous and captures the majority, if not all, the protein-coding genes present in the genome.

Table 1. Assembly summary.

	Illumina assembly	PacBio + Illumina
Number of contigs	836	687
Total contigs size (bp)	35,743,205	35,951,513
Min contig length (bp)	1000	947
Max contig length (bp)	462,380	665,523
N50 contigs (kb)	131	230
Number of scaffolds	582	158
Total scaffolds size (bp)	35,858,128	36,207,399
Min scaffolds length (bp)	1000	977
Max scaffolds length (bp)	1,235,734	2,800,880
N50 scaffolds (kb)	213.4	838.183
N90 scaffolds (kb)	12.2 (249)	188.1 (49)

The N90 and N50 scaffolds are the length of the scaffolds covering 90 and 50% of the genome, respectively. The N90 and N50 were calculated using 39 Mb as genome size. In parenthesis are the numbers of contigs/scaffolds required to cover 90% of the genome length.

3.2. Repeat masking and gene discovery

A total of 65 families of repetitive elements were *ab initio* found and classified with RepeatModeller. 22%, 14%, 9% and 55% were classified as long terminal repeats (LTR), DNA transposons, long interspersed elements and unknown elements, respectively. RepeatModeller families in conjunction with Repbase (3 February 2014) masked 9.83% (3.8 Mb) of the assembled genome. The major family corresponds to LTR elements, specifically to the Gypsy family, which represents 51% of the masked sequences. Similar results have been reported previously for fungal genomes (Muszewska et al. 2011).

A total of 111 transposable elements were identified which masked another 0.2% (76 Kb) of the genome. A total of 3.8 Mb was thus masked, representing a 10% of the assembled genome, which is consistent with the estimated repeat percentage from the k-mer analysis.

The *ab initio* gene predictors rendered 33,544 gene models. A total of 1,642,820 non-redundant consolidated proteins from UniProt were mapped against our gene models, identifying a total of 10,811 unique homologous proteins. This set gave 16,764 hits against the whole genome with GeneBlast. 12,559 gene models were obtained with GeneWise using proteins as evidence. Using EVidenceModeler and PASA, 11,057 consolidated genes were found (Table 2). These predicted gene sequences account for 59% of the *P. purpurogenum* genome, with an average gene length of 2105 bp. On average, each gene contains 3.3 exons and 2.2 introns (Table 2). Comparison of gene models between the chosen fungal genomes (Table S1) shows that our genome has on average larger genes and transcripts. This is explained by the annotation of Untranslated region (UTR)-tails with PASA and the absence of UTR-tails annotations on the reference fungal genomes. No significant difference on the average protein length, number of exons and average intron length is observed when comparing genomes, supporting the aforementioned observation.

A total of 9585 genes (87%) were confirmed by at least 10 reads from RNA-seq data, indicating that our

Table 2. Relevant statistics of gene discovery.

Attribute	Stats
Number of genes	11,057
Number of mRNAs	11,555
Number of exons	37,145
Number of coding DNA sequences	35,980
Number of introns	24,653
Exons per gene	3.3
Coding DNA sequencesper gene	3.2
Introns per gene	2.2
Genes alternative spliced	454
Percentage of genes alternative spliced (%)	4.1
Average gene length (bp)	2105
Average exon length (bp)	612
Average intron length (bp)	92
Average distance between genes (bp)	1349
Genome coding (%)	59
Number of tRNAs	170

A consensus gene set was built with EVM and PASA using evidence from *ab initio*, protein and transcript alignments.

gene models are well supported and in agreement with transcriptome data.

Out of the 11,057 genes predicted, 94% (10,422) were successfully annotated (Table S3) using standard functional protein databases. We were able to assign InterPro Number, GO Numbers and CAZy IDs to 80%, 63% and 8% of the gene models (Table S3), respectively.

3.3. Further validation of the genome sequence

In addition to RNA-seq data, the quality of the genome sequence can be validated by the result of independent sequencing of individual genes. Table 3 lists a set of lignocellulose-degrading enzymes which have been studied and sequenced either prior to the availability of the draft genome or as a result of mining the genome. In all cases, the gene sequences obtained utilising the Sanger method agree with the sequence predicted by the genome. These studies, which show properties of novel enzymes, are also proof that *P. purpurogenum* is a powerful source for the discovery and analysis of new lignocellulolytic enzymes with potential biotechnological applications.

3.4. Manually annotated CAZyme genes present in the *P. purpurogenum* genome

In order to obtain a more precise identification of CAZy genes and their enzymes, a manual annotation of a subset of the CAZymes (*e*-value of 10^{-40}) from those originally assigned by dbCAN with an *e*-value of 10^{-5} was performed. The result is presented in Table S4. A total of 302 genes have been manually annotated, coding for 216 GH, 49 GT, 6 PL and 31 CE, thus confirming the high potential for lignocelluloses degrading enzymes present in the fungus.

3.5. Comparative functional analysis

In order to identify potentially unique novel functions for *P. purpurogenum*, we constructed a tree map of GO-terms using 688 genes without orthologs among the selected Ascomycota species arranged according to their uniqueness (Supek et al. 2011). Figure S2 shows a uniqueness for genes related to glycolipid transport, Krebs cycle, proteolysis, metabolism of L-arabinose and stress response.

In Figure S3 we compare the functional profile of Ascomycota genomes using InterPro domains. We observe that *P. purpurogenum* has, on average, lower number of genes related to specific central metabolic processes such as protein kinase signals, methyltransferases, oxidoreductases and alpha/beta hydrolases. In addition, it potentially has a lower capacity to incorporate amino acids into the cell and has a lower number of genes associated to Zinc and homeodomain transcription factors. However, we observe in *P. purpurogenum* a higher

Table 3. Characterised lignocellulose-degrading enzymes from *P. purpurogenum* whose gene sequence was determined independently.

Enzyme	CAZy family	Gene ID	GenBank	Reference
Endoxylanase A	GH 10	PPSCF00028.25	AAF71268	Chávez et al. (2001a)
Acetyl xylan esterase II	CE 5	PPSCF00020.380	AAC39371	Gutiérrez et al. (1998)
Arabinofuranosidase 1	GH 54	PPSCF00061.89	AAK51551	Carvallo et al. (2003)
Arabinofuranosidase 2	GH 51	PPSCF00010.19	EF490448	Fritz et al. (2008)
Arabinofuranosidase 3	GH 43	PPSCF0002.743	FJ906695	Ravanal et al. (2010)
Arabinofuranosidase 4[a]	GH 54	PPSCF00024.186	AGR66205	Ravanal and Eyzaguirre (2015)
Pectin lyase	PL 1	PPSCF00015.779	KC751539	Pérez-Fuentes et al. (2014)
Exo-arabinanase[a]	GH 93	PPSCF00035.45	KP313779	Mardones et al. (2015)

[a]The gene was mined from the genome and later sequenced.

capacity in terms of gene dosage for functions related to pectin lyase, major facilitator superfamily and glycoside hydrolases (involved in plant cell-wall degradation), general substrate transport, gene regulation and degradation of biomass. The overrepresented terms suggest that *P. purpurogenum* is more specialised in the transport of elements across the cellular membrane (IPR016196 and IPR005828). This is a signal that the fungus has a stronger performance in the secretion of extracellular proteins and import of extracellular sugars. In addition, it is interesting that *P. purpurogenum* has an elevated number of proteins with IPR011050 domain, a domain associated to pectin lyases; this is consistent with the good growth of the fungus when cultivated on rich pectin carbon sources such as sugar beet pulp.

The differences in the number of genes related to transcription factors, pectin lyases and glycoside hydrolases, with respect to the average found in Ascomycota genomes, suggest that *P. purpurogenum* is an interesting model to explore novel mechanics for biodegradation of lignocellulose and its regulation.

To provide further insight into the aforementioned capacity of *P. purpurogenum* for biodegradation of lignocellulosic compounds, we compared the Ascomycota genomes using the Carbohydrate-ActiveEnZymes classification system. The CAZy family diversity per genome is shown in Figure S4 (the raw data of putative predicted CAZYmes utilised to construct Figure S4 have not been included in the article due to their excessive length. They are available from the authors upon request). There is no great difference in terms of CAZy family numbers among the Ascomycota genomes as shown in Figure S5 (a large core of 129 CAZy families is shared by all Ascomycota genomes). However, Figure S6 shows that the number of genes related to each CAZy family varies among the genomes. Thus, we may conclude that the performance for biodegradation of lignocellulose may be more related to gene dosage than to differences in family diversity, and in this respect *P. purpurogenum* presents a higher potential than the other fungi analysed.

3.6. Comparative phylogenetic analysis

CEGMA analysis shows that all Ascomycota genomes in Table S1 are complete in terms of gene models, thus allowing comparative genomic analysis at the gene level without the bias produced by incomplete genomes.

Orthologous analysis whit OrthoMCL produced 19,052 ortholog groups. To represent pairs distances among Ascomycota genomes in terms of ortholog clusters, we built an orthologous dendrogram (Figure S7) based on Manhattan distances (seeSection 2). The orthologous dendrogram locates *P. purpurogenum* close to other *Penicillia* and far from *Talaromyces* in terms of shared orthologous clusters (on average 5502 clusters of distance to the *Talaromyces* species).

To confirm the previous observation, we devised a phylogenetic analysis using a core of 147 conserved clusters among Ascomycota genomes (Figure 1), using *T. reesei* as outgroup and the super-matrix phylogenetic approach (see Section 2).

Our phylogenetics analysis (Figure 1) locates *P. purpurogenum* in the *Penicillium* genus and recovers the known phylogenetic relations among *Trichoderma, Aspergillus, Talaromyces* and *Penicillium* (Samson et al. 2011; Houbraken et al. 2014) where *Penicillium* is located closer to *Aspergillus* than to *Talaromyces*. Also, within *Penicillium*, we recover the current two subgenus classification (Houbraken et al. 2014), which corresponds to *Aspergilloides* (*Penicillium* A) and *Penicillium* (*Penicillium* B) and we locate *P. purpurogenum* as member of the subgenus *Aspergilloides*.

In addition, we compiled a set of sequences for genes BenA, CamA and RPB2 from several species of the genera *Penicillium* (212 sequences), *Aspergillus* (227 sequences) and *Talaromyces* (68 sequences) (Table S2). It has been previously suggested that these genes are well suited as markers for the identification of *Penicillium* species (Visagie et al. 2014). We found that for these three markers, *P. purpurogenum* MYA-38 had the highest identity with sequences from *Penicillium* species, although no exact match was found. The closest match was with *P. ochrochloron* with an average identity of 98.6%. These findings support the classification of our fungus as a *Penicillium*.

4. Conclusions

Combining the PacBio and Illumina technologies we have succeeded in obtaining a high-quality genome sequence for *P. purpurogenum*.

The genome annotation of our strain reveals a higher secretion activity as compared to other fungi

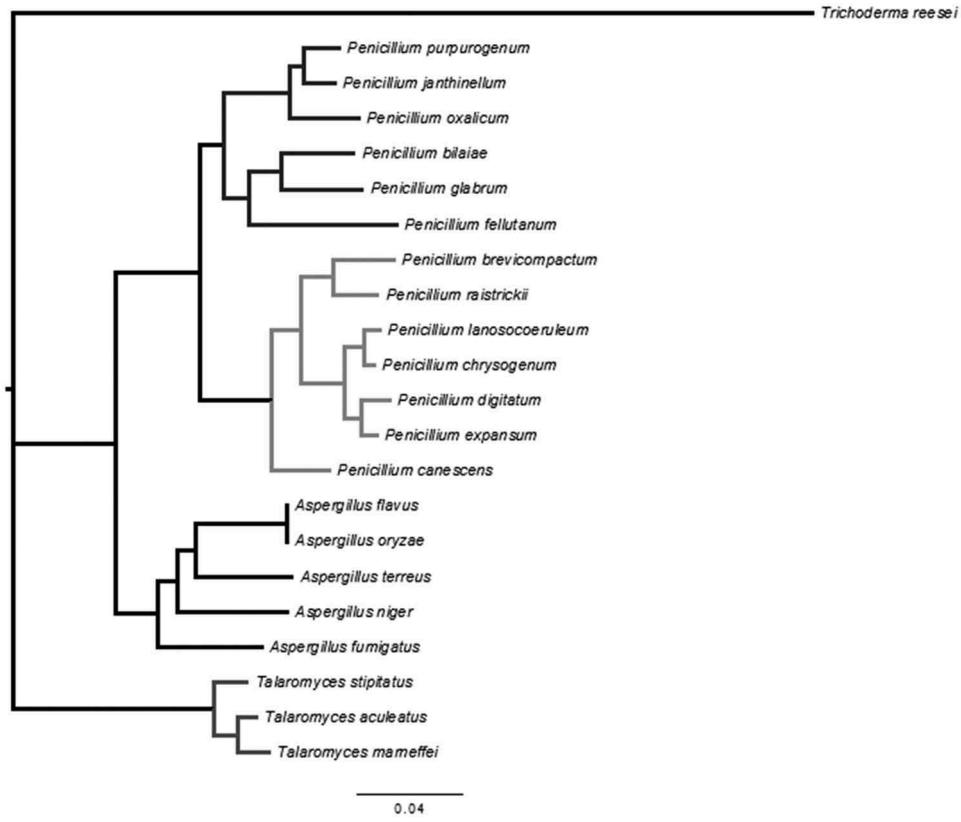

Figure 1. Phylogenetic tree of the phylum Ascomycota inferred using mrBayes and FasTree analysis, considering the deduced amino acid sequences of 147 conserved genes. *T. ressei* was used as outgroup.

analysed, as well as an important number of CAZy and transporter proteins, indicating a high potential for secreting enzymes involved in lignocellulose biodegradation.

The comparative genomic analysis shows that all Ascomycota genomes have a similar repertory and share a large core of CAZyme families. However, they differ in gene dosage, suggesting that this may relate to a different performance in lignocellulose biodegradation.

The availability of a high-quality genome sequence presents a platform for the searching of genes coding for enzymes of potentially novel functions. In addition, this genome is a valuable tool for the analysis and understanding of other fungal genomes.

A preprint of this article has been uploaded in the biorxiv database.

Acknowledgements

The authors thank Dr Shahina Maqbool (Albert Einstein College of Medicine of Yeshiva University, New York, USA) for performing the Illumina DNA sequencing; Dr Rodrigo Gutiérrez and Dr Tatiana Kraiser (Pontificia Universidad Católica de Chile, Santiago, Chile) for the RNASeq sequence and Dr Olivier Fedrigo (Duke University, USA) for the PacBio sequencing.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was funded by grants from the "Fondo Nacional de Ciencia y Tecnología" (FONDECYT) [grant number 1130180] to Jaime Eyzaguirre, Universidad Andrés Bello [grant numbers DI-478-14/R and DI-31-12/R] to Jaime Eyzaguirre, "Fondo de Financiamiento en Areas Prioritarias" (CRG-Fondap) [grant number 1509007] to Alejandro Maass and a "Mejoramiento de la calidad de la Educación Superior" (MECESUP) Fellowship [grant number UAB 0802] to Wladimir Mardones. We acknowledge the contribution of the National Laboratory for High Performance Computing at the Center for Mathematical Modeling, Project PIA ECM-02 – "Comisión Nacional de Investigación Científica y Tecnológica" (CONICYT).

References

Bendtsen JD, Jensen LJ, Blom N, Von Heijne G, Brunak S. 2004. Feature-based prediction of non-classical and leaderless protein secretion. Protein Eng. 17:349–356.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 456:53–59.

Birney E, Clamp M, Durbin R. 2004. GeneWise and genomewise. Genome Res. 14:988–995.

Blanch HW. 2012. Bioprocessing for biofuels. Curr Opin Biotech. 23:390–395.

Boetzer M, Pirovano W. 2014. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. BMC Bioinformatics. 15:211.

Carvallo M, de Ioannes P, Navarro C, Chávez R, Peirano A, Bull P, Eyzaguirre J. 2003. Characterization of an alpha-L-arabinofuranosidase gene (abf1) from *Penicillium purpurogenum* and its expression. Mycol Res. 107:388–394.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 17:540–552.

Chávez R, Almarza C, Schachter K, Peirano A, Bull P, Eyzaguirre J. 2001a. Structure analysis of the endoxylanase A gene from *Penicillium purpurogenum*. Biol Res. 34:217–226.

Chávez R, Bull P, Eyzaguirre J. 2006. The xylanolytic enzyme system from the genus *Penicillium*. J Biotechnol. 123:413–433.

Chávez R, Fierro F, Gordillo F, Martin F, Eyzaguirre J. 2001b. Electrophoretic karyotype of the filamentous fungus *Penicillium purpurogenum* and chromosomal location of several xylanolytic genes. FEMS Microbiol Lett. 205:379–383.

Clarke M, Lohan AJ, Liu B, Lagkouvardos I, Roy S, Zafar N, Bertelli C, Schilde C, Kianianmomeni A, Burglin TR, et al. 2013. Genome of *Acanthamoeba castellanii* highlights extensive lateral gene transfer and early evolution of tyrosine kinase signaling. Genome Biol. 14:R11.

Claudel-Renard C, Chevalet C, Faraut T, Kahn D. 2003. Enzyme-specific profiles for genome annotation: PRIAM. Nucleic Acids Res. 31:6633–6639.

Conn PM. 2003. Handbook of proteomic methods. Totowa (NJ): Humana Press.

Consortium TU. 2014. Activities at the Universal Protein Resource (UniProt). Nucleic Acids Res. 42:D191–D198.

Dehority BA, Johnson RR, Conrad HR. 1962. Digestibility of forage hemicellulose and pectin by rumen bacteria in vitro and the effect of lignification thereon. J Dairy Sci. 45:508–512.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, et al. 2009. Real-time DNA sequencing from single polymerase molecules. Science. 323:133–138.

Emanuelsson O, Nielsen H, Brunak S, von Heijne G. 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol. 300:1005–1016.

Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30:1575–1584.

Fritz M, Ravanal MC, Braet C, Eyzaguirre J. 2008. A family 51 α-L-arabinofuranosidase from *Penicillium purpurogenum*: purification, properties and amino acid sequence. Mycol Res. 112:933–942.

Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, et al. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. P Natl Acad Sci USA. 108:1513–1518.

Gusakov AV, Sinitsyn AP. 2012. Cellulases from *Penicillium* species for producing fuels from biomass. Biofuels. 3:463–477.

Gutiérrez R, Cederlund E, Hjelmqvist L, Peirano A, Herrera F, Ghosh D, Duax W, Jörnvall H, Eyzaguirre J. 1998. Acetyl xylan esterase II from *Penicillium purpurogenum* is similar to an esterase from *Trichoderma reesei* but without a cellulose binding domain. FEBS Lett. 423:35–38.

Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, et al. 2003. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 31:5654–5666.

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. Genome Biol. 9:R7.

Haas BJ, Zeng Q, Pearson MD, Cuomo CA, Wortman JR. 2011. Approaches to fungal genome annotation. Mycology. 2:118–141.

Haft DH, Selengut JD, Richter RA, Harkins D, Basu MK, Beck E. 2013. TIGRFAMs and genome properties in 2013. Nucleic Acids Res. 41:D387–D395.

Hidalgo M, Steiner J, Eyzaguirre J. 1992. Beta-glucosidase from *Penicillium purpurogenum*: purification and properties. Biotechnol Appl Bioc. 15:185–191.

Houbraken J, Visagie CM, Meijer M, Frisvad JC, Busby PE, Pitt JI, Seifert KA, Louis-Seize G, Demirel R, Yilmaz N, et al. 2014. A taxonomic and phylogenetic revision of *Penicillium* section *Aspergilloides*. Stud Mycol. 78:373–451.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 17:754–755.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 110:462–467.

Kanehisa M, Goto S, Kawashima S, Nakaya A. 2002. The KEGG databases at GenomeNet. Nucleic Acids Res. 30:42–46.

Kang SW, Park YS, Lee JS, Hong SI, Kim SW. 2004. Production of cellulases and hemicellulases by *Aspergillus niger* KK2

from lignocellulosic biomass. Bioresource Technol. 91:153–156.

Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 305:567–580.

la Cour T, Kiemer L, Molgaard A, Gupta R, Skriver K, Brunak S. 2004. Analysis and prediction of leucine-rich nuclear export signals. Protein Eng. 17:527–536.

Letunic I, Doerks T, Bork P. 2012. SMART 7: recent updates to the protein domain annotation resource. Nucleic Acids Res. 40:D302–D305.

Lewis SE, Searle SM, Harris N, Gibson M, Lyer V, Richter J, Wiel C, Bayraktaroglir L, Birney E, Crosby MA, et al. 2002. Apollo: a sequence annotation editor. Genome Biol. 3:Research0082.1.

Li H, Benedito VA, Udvardi MK, Zhao PX. 2009. TransportTP: a two-phase classification approach for membrane transporter prediction and characterization. BMC Bioinformatics. 10:418.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13:2178–2189.

Liu G, Zhang L, Wei X, Zou G, Qin Y, Ma L, Li J, Zheng H, Wang S, Wang C, et al. 2013. Genomic and secretomic analyses reveal unique features of the lignocellulolytic enzyme system of *Penicillium decumbens*. PLoS One. 8:e55185.

Marchler-Bauer A, Bryant SH. 2004. CD-search: protein domain annotations on the fly. Nucleic Acids Res. 32:W327–W331.

Mardones W, Callegari E, Eyzaguirre J. 2015. Heterologous expression of a *Penicillium purpurogenum* exo-arabinanase in *Pichia pastoris* and its biochemical characterization. Fungal Biol. 119:1267–1278.

Merino S, Cherry J. 2007. Progress and challenges in enzyme development for biomass utilization. Adv Biochem Eng Biotechnol. 108:95–120.

Mi H, Muruganujan A, Thomas PD. 2013. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res. 41:D377–D386.

Musalem S, Steiner J, Contreras O. 1984. Producción de celulasas por hongos aislados de madera y suelos del sur de Chile. Boletín Micológico. 2:17–25.

Muszewska A, Hoffman-Sommer M, Grynberg M. 2011. LTR retrotransposons in fungi. PLoS One. 6:e29425.

Nierman WC, Yu J, Fedorova-Abrams ND, Losada L, Cleveland TE, Bhatnagar D, Bennett JW, Dean R, Payne GA. 2015. Genome sequence of *Aspergillus flavus* NRRL 3357, a strain that causes aflatoxin contamination of food and feed. Genome Announc. 3:e00168–15.

Parra G, Bradnam K, Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 23:1061–1067.

Pérez-Fuentes C, Ravanal MC, Eyzaguirre J. 2014. Heterologous expression of a *Penicillium purpurogenum* pectin lyase in *Pichia pastoris* and its characterization. Fungal Biol. 118:507–515.

Petersen TN, Brunak S, von Heijne G, Nielsen H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 8:785–786.

Price AL, Jones NC, Pevzner PA. 2005. De novo identification of repeat families in large genomes. Bioinformatics. 21 (Suppl 1):i351–i358.

Price MN, Dehal PS, Arkin AP. 2009. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol. 26:1641–1650.

Pruitt KD, Tatusova T, Brown GR, Maglott DR. 2012. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res. 40: D130–D135.

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, et al. 2012. The Pfam protein families database. Nucleic Acids Res. 40: D290–D301.

Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R. 2005. InterProScan: protein domains identifier. Nucleic Acids Res. 33:W116–W120.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26:841–842.

Ragauskas A, Williams CK, Davison BH, Britovsek G, Cairney J, Eckert CA, Frederick WJ Jr, Hallett JP, Leak DJ, Liotta CL, et al. 2006. The path forward for biofuels and biomaterials. Science. 311:484–489.

Ravanal MC, Callegari E, Eyzaguirre J. 2010. Novel bifunctional alpha-L-arabinofuranosidase/xylobiohydrolase (ABF3) from *Penicillium purpurogenum*. Appl Environ Microb. 76:5247–5253.

Ravanal MC, Eyzaguirre J. 2015. Heterologous expression and characterization of alpha-L-arabinofuranosidase 4 from *Penicillium purpurogenum* and comparison with the other isoenzymes produced by the fungus. Fungal Biol. 119:641–647.

Rosenberg SL. 1978. Cellulose and lignocellulose degradation by thermophilic and thermotolerant fungi. Mycologia. 70:1–13.

Samson RA, Yilmaz N, Houbraken J, Spierenburg H, Seifert KA, Peterson SW, Varga J, Frisvad JC. 2011. Phylogeny and nomenclature of the genus *Talaromyces* and taxa accommodated in *Penicillium* subgenus *Biverticillium*. Stud Mycol. 70:159–183.

She R, Chu JS, Uyar B, Wang J, Wang K, Chen N. 2011. genBlastG: using BLAST searches to build homologous gene models. Bioinformatics. 27:2141–2143.

Sigrist CJA, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, Bougueleret L, Xenarios I. 2013. New and continuing developments at PROSITE. Nucleic Acids Res. 41:D344–D347.

Stanke M, Morgenstern B. 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. 33:W465–W467.

Steiner J, Socha C, Eyzaguirre J. 1994. Culture conditions for enhanced cellulase production by a native strain of *Penicillium purpurogenum*. World J Microb Biot. 10:280–284.

Supek F, Bošnjak M, Škunca N, Šmuc T, Gibas C. 2011. REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One. 6:e21800.

Tempel S. 2012. Using and understanding RepeatMasker. Method Mol Biol. 859:29–51.

Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M. 2008. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome Res. 18:1979–1990.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 25:1105–1111.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 7:562–578.

van den Berg M, Albang R, Albermann K, Badger JH, Daran J-M, Driessen AJM, Garcia-Estrada C, Fedorova ND, Harris DM, Heijne WHM, et al. 2008. Genome sequencing and analysis of the filamentous fungus *Penicillium chrysogenum*. Nat Biotechnol. 26:1161–1168.

Visagie CM, Houbraken J, Frisvad JC, Hong S-B, Klaassen CH, Perrone G, Seifert KA, Varga J, Yaguchi T, Samson RA. 2014. Identification and nomenclature of the genus *Penicillium*. Stud Mycol. 78:343–371.

Wilson D, Pethica R, Zhou Y, Talbot C, Vogel C, Madera M, Chothia C, Gough J. 2009. SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny. Nucleic Acids Res. 37:D380–D386.

Ye J, McGinnis S, Madden TL. 2006. BLAST: improvements for better sequence analysis. Nucleic Acids Res. 34:W6–W9.

Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. 2012. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 40:W445–W451.