Contrasting genomic diversity in two closely related postharvest pathogens: *Penicillium digitatum* and *Penicillium expansum*

Irene Julca¹²³, Samir Droby⁴, Noa Sela⁵, Marina Marcet-Houben¹²* and Toni Gabaldón¹²⁶*

1) Bioinformatics and Genomics Program. Centre for Genomic Regulation (CRG). Dr. Aiguader, 88. 08003 Barcelona, Spain
2) Universitat Pompeu Fabra (UPF). 08003 Barcelona, Spain.
3) Universitat Autònoma de Barcelona. Barcelona, Spain
4) Department of Postharvest Science, ARO, the Volcani Center, P.O.Box 6, Bet Dagan 50250, Israel
5) Department of Plant Pathology and Weed Research, the Volcani Center, P.O.Box 6, Bet Dagan 50250, Israel
6) Institució Catalana de Recerca i Estudis Avançats (ICREA), Pg. Lluis Companys 23, 08010 Barcelona, Spain.

*Both authors share senior authorship

**Data deposition:** The sequence data produced for this project has been deposited at DDBJ/EMBL/GenBank under accession No. LHSL0000000.
Abstract

*Penicillium digitatum* and *Penicillium expansum* are two closely-related fungal plant pathogens causing green and blue mold in harvested fruit, respectively. The two species differ in their host specificity, being *P. digitatum* restricted to citrus fruits and *P. expansum* able to infect a wide range of fruits after harvest. Although host-specific *Penicillium* species have been found to have a smaller gene content, it is so far unclear whether these different host specificities impact genome variation at the intra-specific level. Here we assessed genome variation across four *P. digitatum* and seven *P. expansum* isolates from geographically distant regions. Our results show very high similarity (average 0.06 SNPs per Kb) between globally distributed isolates of *P. digitatum* pointing to a recent expansion of a single lineage. This low level of genetic variation found in our samples contrasts with the higher genetic variability observed in the similarly distributed *P. expansum* isolates (2.44 SNPs per Kb). Patterns of polymorphism in *P. expansum* indicate that recombination exists between genetically diverged strains. Consistent with the existence of sexual recombination and heterothallism, which was unknown for this species, we identified the two alternative mating types in different *P. expansum* isolates. Patterns of polymorphism in *P. digitatum* indicate a recent clonal population expansion of a single lineage that has reached worldwide distribution. We suggest that the contrasting patterns of genomic variation between the two species reflect underlying differences in population dynamics related with host specificities and related agricultural practices. It should be noted, however, that this results should be confirmed with a larger sampling of strains, as new strains may broaden the diversity so far found in *P. digitatum*. 
Introduction

Worldwide, green mold, caused by *Penicillium digitatum* (Pers: Fr.), and blue mold, caused by *Penicillium expansum* (Link) Thom, are the two most damaging postharvest diseases of fruits (Marcet-Houben et al. 2012; Ballester et al. 2014). *P. digitatum* infects citrus fruits exclusively (Marcet-Houben et al. 2012) while *P. expansum* can infect and develop decay in a large range of fruit crops, among which apples (Buron-Moles et al. 2015), pears (Paster et al. 2009), peaches (Yang et al. 2011), hazel (Yang et al. 2014) and others (Judit-Correia et al. 2010; Cao et al. 2012). Infection of the fruits by both pathogens takes place through surface injuries inflicted in the peel of the fruit during harvest and subsequent handling in the packhouse. Spores present in these injuries become active, germinate, colonize the injured tissue, and produce disease. Besides the economic impact of these fungi for the fruit industry, *P. expansum* is also a potential threat to public health due to its ability to produce patulin, a toxic mycotoxin, which can contaminate infected fruits and their derived products (Morales et al. 2007; Ballester et al. 2014). Other *Penicillium* species are also able to synthesize a wide array of toxins which has increased the interest for these fungi. As a result, the number of sequenced genomes has also increased, with the number of *Penicillium* genome projects reaching more than 40 (NCBI database (Geer et al. 2010). Sequencing at the intra-species level has also received increased attention, due to interest in unveiling the genetic bases of phenotypic variations observed between different strains. For instance, the three *P. expansum* strains sequenced by Ballester et al. (Ballester et al. 2014) presented different levels of Patulin production, with one strain being unable to produce the toxin. Despite its importance, the genomic dynamics of different strains of *Penicillium* species are still poorly understood.

Studies of genetic polymorphism in other species from the *Trichocomaceae* family are scarce, but available data point to the existence of several globally distributed lineages, so that isolates from distant locations can show high similarity while, at the same time, isolates from nearby locations...
may show larger differences. For instance, a comparison between two strains of *Penicillium chrysogenum* from China and USA revealed a very low level of sequence divergence between the two strains measured as the density of single nucleotide polymorphisms (SNP) between the two strains (~0.08 SNPs/Kb) (Wang et al. 2014). In contrast, a comparison of the *P. chrysogenum* type strain to *P. chrysogenum KF-25* showed a high level of protein sequence divergence where proteins shared only 75.1 % similarity on average, suggesting they are actually different species (Peng et al. 2014). In *Aspergillus fumigatus*, the comparison of two strains, A1163, a derivative strain of CBS144-89/CEA10 isolated from a patient in Paris, and Af293 isolated from a patient in UK, showed an identity of 99.8% at nucleotide level between the core genes (Fedorova et al. 2008). A genome-wide nucleotide diversity analysis showed that the genetic diversity of the 8 *A. oryzae* strains from Japan (1.05 SNPs/Kb) is ~25% of that found in 8 *A. flavus* strains from different parts in the world (2.06 SNPs/Kb) (Gibbons et al. 2012). In *A. niger*, the type strain CBS513.88, from The Netherlands, was compared to *A. niger SH2* from China, which showed a high level of similarity to the type strain with a SNP density of 0.051 SNPs/Kb (Yin et al. 2014) while the second strain (ATCC1015 from USA) had a more than hundred higher average SNP density of 8 SNPs/Kb (Andersen et al. 2011). Comparison between this latter strain and a fourth *A. niger* strain (ATCC9029, also from USA) showed an intermediate level of divergence (2 SNPs/Kb) (Andersen et al. 2011). Remarkably, in all cases, SNPs were shown to cluster in hypervariable regions in the genome rather than being uniformly distributed. In the strain *A. niger SH2* these hypervariable regions were located in regions where gene order was not conserved with respect to other *Aspergillus* species. This effect was first observed in *A. oryzae* where the comparison of two strains showed that proteins with mutations were more often found in genes whose synteny was not conserved among other *Aspergillus* species (Umemura et al. 2012). These regions were often subtelomeric and were enriched in genes encoding secondary metabolism clusters, hydrolases and transporters.
In addition, the increase of genetic variability can be linked to sexual recombination between diverged strains (Horn et al. 2009; Böhm et al. 2015; Teixeira et al. 2015). Until some decades ago the fungal kingdom was considered to comprise a high proportion of asexual species. Now, a growing number of studies in diverse species previously thought to be asexual, show that they are indeed able to reproduce sexually at some stage of their life cycle, or under some specific conditions (Taylor et al. 2015; Ropars et al. 2012; Böhm et al. 2015). Sex in fungi is controlled by the MAT loci, which encode key regulators of mating, production of pheromones and pheromone receptors (Fowler et al. 1999). The fungal kingdom includes outbreeding (heterothallism) systems with at least two and up to thousands of different mating types and inbreeding (homothallism) systems (Ni et al. 2011). Specifically, in Ascomycetes, MAT regions can vary in size from 1-6 Kb and contain between one and three MAT genes, which encode homeodomain, α-domain and/or high mobility group (HMG)-domain transcription factors (Dyer 2008). Sexual reproduction is driven by mating compatibility and is mainly regulated by two MAT idiomorphs (dissimilar sequences occupying the same locus on the chromosome). Heterothallic individuals require the interaction of two partners of opposite mating types with compatible MAT idiomorphs in order to reproduce. Homothallic organisms are self-fertile and can present primary or secondary homothallism (pseudohomothallism), in the first one both idiomorphs are present in the same individual genome, either fused together or in different genomic regions; in the second one, an individual that has several mating types in the same genome is able to switch between them in order to reproduce with an individual of an opposite mating type (Ni et al. 2011; Billiard et al. 2012; Almeida et al. 2015).

Of note, the above mentioned studies focus on generalist species that inhabit a variety of niches. Indeed A. niger can be found in mesophilic environments such as decaying vegetation or soil and plants (Schuster et al. 2002) and P. chrysogenum is found in moist soil in habitats ranging from degraded forests (Jha et al. 1992) to arctic subglacial ice (Sonjak et al. 2006). Both species can colonize decaying fruits of a wide variety of species, likely resulting from their opportunistic
saprophytic lifestyle. The initial sequencing of the *P. digitatum* genome revealed that strains isolated from Spain and China had nearly identical mitochondrial sequences, which was suggestive of a recent global expansion of this species (Marcet-Houben et al. 2012). On the other hand, *P. expansum* presented a higher degree of variability in the three sequenced strains (Ballester et al. 2014). The authors suggested that these differences in genome variability were correlated with smaller gene content and higher host-specificity (Ballester et al. 2014). Currently, there are four sequenced strains for *P. digitatum* (Marcet-Houben et al. 2012; Sun et al. 2013) and seven for *P. expansum* (Ballester et al. 2014; Yang et al. 2014; Yu et al. 2014; Li et al. 2015; Nordberg et al. 2014) which doubles the number of genomes used in previous analyses. Here we assessed patterns of intraspecific variation in the two species, which extended previous observation of contrasting levels of variability. In addition, we reconstructed the evolutionary history of the sequenced lineages and assessed the contribution of recombination among divergent strains to patterns of genome variability.

**Material and Methods**

**Fungal strains and genome data**

Four strains of *P. digitatum* (Table 1) and seven strains of *P. expansum* (Table 2) were used in this study. These strains were isolated in different parts of the world and from different substrates (Figure 1A). The two Spanish strains of *P. digitatum* were isolated from orange and grapefruit (Marcet-Houben et al. 2012), the Chinese strain from citrus (Sun et al. 2013) and the Israeli one was isolated from decayed grapefruit and was sequenced within this study (see below). The Chinese, Israeli, Spanish and three American strains of *P. expansum* were isolated from decayed apples (Li et al. 2015; Ballester et al. 2014; Yang et al. 2014; Yu et al. 2014), while the additional American strain was isolated from hazel “*Corylus avellana* L.” and was initially misidentified as *Penicillium aurantiogriseum* (Yang et al. 2014), and corrected later (Ballester et al. 2014). Based on their higher
quality, the Spanish *P. digitatum* (PHI26) and the Israeli *P. expansum* (PEXP) strains were used as reference for their respective species. Due to the lack of publicly stored data for two strains of *P. expansum* (R19 and T01) we simulated sequencing paired-end reads (70 bases read length, and 100 bases insert size) based on their genome assembly using wgsim 0.3.1-r13 tool (Li 2012) setting the error parameter to zero (rRXe = 0).

**Genome sequencing of *P. digitatum* strain PDC 102**

Here we report the genome sequence of *P. digitatum* strain PDC 102, originally isolated from decayed grapefruit. The genome of *P. digitatum* was sequenced using Illumina HiSeq 2000 at the high throughput sequencing unit, Technion – Israel Institute of Technology, Haifa, Israel. A total of 83,360,197 paired-end reads with a length of 100 nucleotides were generated with genome coverage of approximately 500-fold. The paired-end reads were assembled into 883 contigs with an N50 contig length of 70,904 bp by the ABysS (Simpson et al. 2009) software followed by SSPACE (Boetzer et al. 2011) and GapFiller (Boetzer and Pirovano 2012) software to create longer scaffolds with paired-end data. This whole genome shotgun project has been deposited at DDBJ/EMBL/GenBank under accession No. LHSL00000000. The version described in this paper is the first version (accession No. LHSL00000000). Gene prediction and annotation were done using the maker website pipeline (Cantarel et al. 2008) and the Augustus software pipeline (Stanke et al. 2004).

**Detection of Single Nucleotide Polymorphisms**

To improve the reference genomes, we mapped their reads onto their respective reference genome using BWA 0.7.6a-r433 (Li and Durbin 2009) and we identified SNPs with GATK (McKenna et al. 2010), using thresholds for mapping quality (>40) and read depth (>20) as filters. Then we substituted the reference nucleotides with the highly reliable SNPs, as they are most likely the results of sequence errors in the assembly. Then, sequence reads from each strain were mapped onto
the respective improved reference genome for the SNP calling, using the same tools described
before (BWA and GATK), but this time, lowering the read depth to 10. To study patterns of SNP
distribution along the genomes, we plotted Venn diagrams using InteractiVenn tool (Heberle et al.
2015) and for each strain, we plotted the SNP density per contig for non-overlapping windows of 1
Kb.

**Population genetics parameters**

We estimated the Nei's gene diversity index (Nei 1973) in each species. To investigate the variation
of non-synonymous and synonymous SNPs in the coding regions per strain, we compared non-
synonymous changes per non-synonymous site (πN) to synonymous changes per synonymous site
(πS) by assuming that ¾ of all sites are non-synonymous.

**Phylogenetic analysis**

A species tree was built using a gene concatenation approach. We used *Penicillium camemberti*
as an out-group. We selected the orthologous proteins predicted during the phylome reconstruction of
*P. expansum* (Ballester et al. 2014). Genes that had one-to-one orthologs in the three species of
interest (*P. camemberti, P. digitatum* and *P. expansum*) (Gabaldón 2008) were selected. After
discarding genes with no SNP in any of the three species we kept 5,892 protein-coding genes for
further analyses. We first did a protein alignment using the pipeline described by Huerta-Cepas
(Huerta-Cepas et al. 2011). Briefly, this alignment strategy consists of doing six alignments using
three different programs (MUSCLE v3.8.31 (Edgar 2004), MAFFT v6.861b (Katoh et al. 2002),
Kalign v2.04 (Lassmann and Sonnhammer 2005) and applying them in forward and reverse
sequence orientation. The six resulting alignments are then combined using M-COFFEE as
implemented in T-COFEE v8.80 (Notredame et al. 2000). The alignment is then back-translated to
coding sequence (CDS). After concatenating all the alignments, the resulting nucleotide alignment
was trimmed using trimAl v1.4 (Capella-Gutiérrez et al. 2009) to remove all positions with gaps in
the alignment. The final alignment had 9,321,774 positions. A phylogenetic tree was then
reconstructed from this alignment using PhyML (Guindon and Gascuel 2003) using
the GTR model and other options as default. 100 bootstrap repetitions were calculated.

**Genomic Recombination**

To compute the number of recombination events we selected the four strains of *P. digitatum* and
four strains of *P. expansum* (PEXP, PEX1, NRRL 62431 and ATCC 24692). LDhat v. 2.2 (Auton
and McVean 2007) was used to infer the minimum number of recombination events (Rmin) per
species. In addition, for *P. expansum*, we used RDP4 (Martin and Rybicki 2000) on the largest
contigs (>500 Kb) to analyse the recombination breakpoints and recalculate the SNP density
without recombination tracks (Martin and Rybicki 2000).

**MAT locus search**

In order to search the presence of MAT idiomorph regions in *P. digitatum* and *P. expansum*, we
performed a blast search of the two idiomorphs available in NCBI database (Geer et al. 2010) in the
species closely related to *P. digitatum* and *P. expansum* (Supplementary Table S1) against each
genome.

**Results and Discussion**

**Genome sequence of *P. digitatum* PDC 102**

To increase the genomic information of *P. digitatum* we undertook the sequencing of an additional
strain of this species: PDC 102, isolated from decayed grapefruit in Israel. This genome was
sequenced using Illumina HiSeq 2000 pair-end approach (see Materials and Methods) and the final
assembly resulted in a genome size (24.9 Mb) similar to other sequenced strains of *P. digitatum*
(Marcet-Houben et al. 2012). The assembly includes 883 contigs with an N50 contig length of
70,904 bp and a coverage of 500x (Table 1). The genome has an average G+C content of 48.9%. A
total of 8,070 predicted proteins (minimum length of 44 amino acids, maximum length of 6,923 amino acids) were identified in this strain.

**Patterns of polymorphism in *P. expansum* and *P. digitatum***

Including the newly sequenced *P. digitatum* PDC 102 genome we have a total of four *P. digitatum* genomes and seven *P. expansum* genomes isolated from different geographical locations (Figure 1A). This allowed us to derive maps of genome variation in both species (see Materials and Methods). For each species we mapped the read data to the chosen reference assembly of the same species to detect SNPs (see Materials and Methods). These analyses showed that in *P. digitatum* there were few polymorphic sites, with an average of 0.06 SNPs per Kb (Table 3, Figure 1B). In contrast, the same analysis in *P. expansum* showed a 40-fold larger amount of genetic variation, with an average of 2.44 SNPs per Kb (Table 3, Figure 1C). These differences were also apparent when considering Nei’s index of genetic diversity (Hs) which was 47 times higher in *P. expansum* as compared to *P. digitatum* (Table 3). These results are consistent with previous comparisons done with the two Spanish strains of *P. digitatum*, and supported by the comparison of the mitochondrial sequence of the Chinese strain (Marcet-Houben et al. 2012) and three completely sequenced strains of *P. expansum* (Ballester et al. 2014). Interestingly, for *P. expansum* we found no correlation between the number of SNPs shared between the strains and their geographical distribution. For instance, the strains that shared more SNPs are from USA (ATCC 24692) and China (T01), while the strains that shared less SNPs are both from USA (PEX2 and R19). In *P. digitatum*, the strains from China (Pd01-ZJU) and Israel (PDC 102) shared more SNPs (Figure 1B).

**Patterns of synonymous and non-synonymous SNPs**

We next analyzed the distribution of SNPs present within protein-coding exons and observed that the number of nonsynonymous SNPs was higher than the number of synonymous SNPs in all the strains of *P. digitatum*, while in *P. expansum* all strains displayed an excess of synonymous SNPs.
We calculated per site non-synonymous and synonymous nucleotide diversity by assuming that $\frac{3}{4}$ of all sites are non-synonymous and computed the ratio between these values (Table 3). Our results indicate an excess in the rate of non-synonymous to synonymous substitutions in *P. digitatum* (0.57-0.73) as compared to *P. expansum* (0.19-0.25). These results indicate a relaxation of purifying selection in *P. digitatum* relative to *P. expansum*. Considering that this excess is distributed along the entire genome and is present in all sampled isolates, an underlying common demographic history of the population should provide an explanation to our observation. A possible explanation is that the excess of non-synonymous SNPs in *P. digitatum* results, at least in part, from a population bottleneck followed by a recent population expansion. Earlier studies based on simulations (Lohmueller et al. 2008; Travis et al. 2007) showed that during a population bottleneck the proportion of non-synonymous SNPs increases. When the population expanded, the fraction of non-synonymous SNPs increases because the increase in population size results in many more mutations, of which most are non-synonymous because of the genetic code. This scenario would also imply that the population growth was recent enough so that purifying selection has not had enough time to reduce the fraction of slightly deleterious non-synonymous SNPs to the equilibrium value expected for a larger population.

**Evolutionary relationships based on SNP data**

We used SNP data to reconstruct the phylogenetic relationships among the strains considered, using *P. camemberti* as an outgroup. We limited the analysis to coding regions that had one to one orthologs in all three species and that had at least one SNP in any *P. expansum* or *P. digitatum* strains. 5,892 single copy genes fulfilled these requirements. These genes were concatenated and a species tree was reconstructed using a maximum likelihood approach (see Materials and Methods). As seen in Figure 2, there was no correlation between the phylogenetic position and the geographical origins of *P. expansum* and *P. digitatum*. Neither did we find a correlation with the host specificity in *P. expansum*. Branch lengths from the common ancestor to *P. digitatum* were
longer than the lengths to *P. expansum*. At the strain level, however, branches in *P. digitatum* were shorter, in consonance with the smaller amount of genomic variation observed (Table 3). Collectively these data indicate that the time to coalescence is longer in *P. expansum* strains as compared to *P. digitatum* strains (i.e. *P. expansum* strains started to diverge earlier).

**Signatures of recombination and heterothallism in the *P. expansum* genome**

Interestingly, we found that the SNPs were not equally distributed along the *P. expansum* genome, but rather they formed clusters (see Figure 3 for an example). Such pattern is reminiscent of possible recombination events, in which a given region of the genome has been introgressed from a genetically different strain. To test this possibility we searched for recombination events in both species (see Materials and Methods). Our results show a much larger number of predicted recombination events in *P. expansum* (0.036 Rmin per Kb, where Rmin is the minimum number of recombination events) versus the 0.004 Rmin per Kb found in *P. digitatum* (Table 3). To assess to what extent recombination may have contributed to the observed larger diversity in *P. expansum* we compared the SNPs density in a subset of the longest 13 contigs, comprising 10.4 Mb, before and after excluding recombination tracks (see Materials and Methods). In the evaluated regions, SNP density decreased from 7.68 to 7.49 SNPs/Kb, indicating that roughly 2.5% of the variation may be attributed to recombination tracks between the strains compared. This result shows some contribution to genetic variability in *P. expansum* of recombination among diverged lineages, although this process can only explain a small part of the larger genetic variability when compared to that found in *P. digitatum* (0.06 SNPs/Kb). The low level of recombination events detected in *P. digitatum* (0.004 Rmin per Kb) (Table 3) is not necessarily indicative of a low prevalence of recombination, as recombination events between very closely related strains are very difficult to detect.

Consistent with sexual crossing as the source of recombination between lineages we found
reproduction idiomorphs of alternative mating types (MAT genes) among the sequenced P. expansum strains. Indeed, we found that the strains NRRL 62431 and R19 of P. expansum carry the MAT1-1 idiomorph as was previously reported for the PEXP, PEX1, PEX2 strains (Ballester et al. 2014). However, the ATCC 24692 and T01 strains had the MAT1-2 idiomorph (Supplementary Figure S1), suggesting the possibility of sexual reproduction in P. expansum as an heterothallic species. The upstream and downstream flanking regions of the MAT genes included, respectively, the genes APN2 and SLA2, which is a structural organization similar to the mating-type loci in other ascomycetes that have sexual reproduction, like P. roqueforti and P. chrysogenum (Böhm et al. 2015; Ropars et al. 2012). These results indicate that P. expansum is an heterothallic species and that mating among diverged strains may have contributed to the observed patterns of genomic variation. To the best of our knowledge this is the first indication of heterothallism in P. expansum.

Concluding remarks
The results reported here from the analysis of the genomes of P. expansum and P. digitatum strains sampled worldwide unveil contrasting patterns of genomic variation. On the one hand P. expansum has a large genetic diversity. Consistently, pairwise comparisons of strains show a relatively high sequence divergence, which does not correlate with the geographical distance between the compared isolates. This indicates that sampled isolates diverged a long time ago but that the geographical structure of the populations may have been broken by migrations. In addition we find compelling evidence for recombination between diverged P. expansum lineages. In this regard the demonstration of the presence of two alternative mating type loci in P. expansum support heterothallism and sexual recombination in this species, thus providing a mechanistic basis for the observed introgressions between diverged lineages. On the other hand P. digitatum shows a remarkably low genetic diversity and only one mating type locus was present among the analyzed strains. In addition we detected a relative excess of non-synonymous SNPs as compared to P. expansum. Taken together these results suggest a very recent clonal expansion for P. digitatum.
How can these two contrasting patterns be explained in the light of their association with species of agricultural interests?. One of the main differences between these two species is their contrasting host specificity. As explained above, \textit{P. expansum} is able to infect a wide range of hosts, whereas \textit{P. digitatum} is specific to citrus fruits. Although any consideration must remain highly speculative at this point, we hypothesize that differences in host specificity coupled with present and past agricultural practices in the respective hosts have shaped the two species differently. Of note, although most fruits were first domesticated several millennia ago, the industrial expansion of key varieties and the start of significant global commerce can be traced back in most cases to the last century. Hence, it is reasonable to consider that expansion and industrialization of agriculture have favored the movement of strains of both species around the world in recent times, which is compatible with the observed absence of a clear geographical structure of the genetic diversity. However, we argue that, being \textit{P. digitatum} a highly host-specific species, its original distribution was narrow, perhaps restricted to certain geographical areas, and thus the expansion affected a single population that thereby colonized new areas. The fact that only one mating type is found in sequenced \textit{P. digitatum} strains may be related to this clonal expansion, although the number of analyzed strains is still too low to reach any conclusion. In contrast, the more versatile \textit{P. expansum} may have had a wider, perhaps global, geographical distribution preceding the start of global commerce. Thus when commercial expansion of several of its host species favored migration, this resulted in favoring genetic admixture between the newly encountered populations, particularly considering that the species is able to reproduce sexually. Such scenario could explain how a similar process (global migration) can result in totally different population genomics patterns depending on the preceding diversity and geographical distribution of the species at hand. Thus whereas global commerce may favor fast clonal expansion of niche specialists it will trigger population admixture in generalists. Naturally, further research is needed to better understand the present and past population dynamics in these and other important plant pathogens.
Acknowledgements


Figure Legends

**Fig. 1.** Approximate geographic locations and host species where the four strains of *P. digitatum* and the seven strains of *P. expansum* analyzed in this study were isolated (A). Venn diagrams indicating shared SNPs in (B) *P. digitatum* using the the PHI26 strain as a reference for the other three strains (Pd1, Pd01-ZJU, PDC 102), (C) *P. expansum* using the PEXP strain as a reference for the other six strains (PEX1, PEX2, NRRL 62431, ATCC 24692, R19, T01).

**Fig. 2.** Maximum likelihood species tree derived from the SNPs found in 5,892 protein-coding genes present in all species considered. All bootstrap values that are not maximal (100%) are indicated in the figure. Scales below indicate substitutions per site. Insets show detail of the subtrees containing *P. expansum* and *P. digitatum* lineages at a larger scale, to better appreciate the differences in branch lengths.

**Fig. 3.** (A) Visualization with the integrative genome viewer (IGV) of the mapping results of a selected 32Kb region (coordinates with respect to the reference: contig129:71,272-103,598). The
region was predicted to contain a recombination track. The three lower frames show depth of coverage graphs for the mapping of the sequenced reads against the reference for PEXP, PEX1 and ATCC 24692 strains, in this order. SNPs are indicated with colors. Note that the reference strain, as expected, has no SNPs, and that PEX1 shows a region of possible recombination (highlighted with a red box) in which the SNP pattern is similar to that in ATCC 24692 whereas the surrounding regions have much fewer and distinct SNPs. (B) Schematic representation of the selected region, in which the two SNP patterns in (A) are indicated as different colors. The red square shows a region in which the strain PEX1 has a SNP pattern (green) which is different from the overall pattern of the rest of the region and that in PEXP (yellow), and similar to the one found in ATCC 24692.

Table Legends

Table 1.- Summary of the main assembly and annotation features of the genomes of the four *P. digitatum* strains used for this study.

Table 2.- Summary of the main assembly and annotation features of the genomes of the seven *P. expansum* strains used for this study.

Table 3.- Number of SNPs (total, synonymous and non-synonymous SNPs), density of SNPs (SNPs/Kb), \( \pi_N/\pi_S \) ratio, Nei index (Hs) and minimum number of recombination events (Rmin) per Kb of the *P. digitatum* and *P. expansum* genome strains using as a reference PHI26 and PEXP strain, respectively.
Supplementary Tables and Figures

Supplementary Table S1. NCBI database accession numbers of the idiomorphs of six species closely related to *P. digitatum* and *P. expansum*.

Supplementary Figure S1. IGV read alignment view at 10-kb resolution of the Mat locus region in *P. expansum*. Regions without reads are shown as white spaces and the red boxes mark the position of the mat locus A) Region of the mat1-1 using PEXP strain as a reference, B) Region of the mat1-2, using ATCC 24692 strain as a reference.

References


Fig. 1. Approximate geographic locations and host species where the four strains of *P. digitatum* and the seven strains of *P. expansum* analyzed in this study were isolated (A). Venn diagrams indicating shared SNPs in (B) *P. digitatum* using the PHI26 strain as a reference for the other three strains (Pd1, Pd01-ZJU, PDC 102), (C) *P. expansum* using the PEXP strain as a reference for the other six strains (PEX1, PEX2, NRRL 62431, ATCC 24692, R19, T01).
Fig. 2. Maximum likelihood species tree derived from the SNPs found in 5,892 protein-coding genes present in all species considered. All bootstrap values that are not maximal (100%) are indicated in the figure. Scales below indicate substitutions per site. Insets show detail of the subtrees containing *P. expansum* and *P. digitatum* lineages at a larger scale, to better appreciate the differences in branch lengths.
Fig. 3. (A) Visualization with the integrative genome viewer (IGV) of the mapping results of a selected 32Kb region (coordinates with respect to the reference: contig129:71,272-103,598). The region was predicted to contain a recombination track. The three lower frames show depth of coverage graphs for the mapping of the sequenced reads against the reference for PEXP, PEX1 and ATCC 24692 strains, in this order. SNPs are indicated with colors. Note that the reference strain, as expected, has no SNPs, and that PEX1 shows a region of possible recombination (highlighted with a red box) in which the SNP pattern is similar to that in ATCC 24692 whereas the surrounding regions have much fewer and distinct SNPs. (B) Schematic representation of the selected region, in which the two SNP patterns in (A) are indicated as different colors. The red square shows a region in which the strain PEX1 has a SNP pattern (green) which is different from the overall pattern of the rest of the region and that in PEXP (yellow), and similar to the one found in ATCC 24692.
Table 1.- Summary of the main assembly and annotation features of the genomes of the four *P. digitatum* strains used for this study.

<table>
<thead>
<tr>
<th><em>P. digitatum</em> strains</th>
<th>PHI26</th>
<th>Pd1</th>
<th>Pd01-ZJU</th>
<th>PDC 102</th>
</tr>
</thead>
<tbody>
<tr>
<td>Origin</td>
<td>Spain</td>
<td>Spain</td>
<td>China</td>
<td>Israel</td>
</tr>
<tr>
<td>Sample</td>
<td>orange</td>
<td>grapefruit</td>
<td>citrus</td>
<td>grapefruit</td>
</tr>
<tr>
<td>Sequencing technology</td>
<td>Illumina (HiSeq 2000)</td>
<td>Roche 454</td>
<td>Roche 454, Illumina (GAI)</td>
<td>Illumina (HiSeq 2000)</td>
</tr>
<tr>
<td>Genome size (Mb)</td>
<td>25.6</td>
<td>24.9</td>
<td>26</td>
<td>24.9</td>
</tr>
<tr>
<td>Sequencing coverage</td>
<td>83x</td>
<td>24x</td>
<td>300x</td>
<td>500x</td>
</tr>
<tr>
<td>Number of contigs</td>
<td>287</td>
<td>544</td>
<td>1817</td>
<td>826</td>
</tr>
<tr>
<td>Number of scaffolds</td>
<td>102</td>
<td>54</td>
<td>76</td>
<td>ND</td>
</tr>
<tr>
<td>Number of Large Scaffolds (&gt;100 Kb)</td>
<td>36</td>
<td>26</td>
<td>35</td>
<td>ND</td>
</tr>
<tr>
<td>N50 (base pairs)</td>
<td>878909</td>
<td>1533507</td>
<td>1047382</td>
<td>70904</td>
</tr>
<tr>
<td>GC content (%)</td>
<td>48.9</td>
<td>48.9</td>
<td>48.4</td>
<td>48.9%</td>
</tr>
<tr>
<td>Number of protein-coding genes</td>
<td>9153</td>
<td>8969</td>
<td>9006</td>
<td>8070</td>
</tr>
</tbody>
</table>
Table 2.- Summary of the main assembly and annotation features of the genomes of the seven *P. expansum* strains used for this study.

<table>
<thead>
<tr>
<th><em>P. expansum</em> strains</th>
<th>PEXP</th>
<th>PEX1</th>
<th>PEX2</th>
<th>NRRL 62431</th>
<th>ATCC 24692</th>
<th>R19</th>
<th>T01</th>
</tr>
</thead>
<tbody>
<tr>
<td>Origin</td>
<td>Israel</td>
<td>Spain</td>
<td>USA</td>
<td>USA</td>
<td>USA</td>
<td>USA</td>
<td>China</td>
</tr>
<tr>
<td>Sample</td>
<td>apple</td>
<td>apple</td>
<td>apple</td>
<td>nuts</td>
<td>apple</td>
<td>apple</td>
<td>apple</td>
</tr>
<tr>
<td>Genome size (Mb)</td>
<td>32.1</td>
<td>31.1</td>
<td>32.4</td>
<td>32.7</td>
<td>32.5</td>
<td>31.4</td>
<td>33.5</td>
</tr>
<tr>
<td>Sequencing coverage</td>
<td>605x</td>
<td>73x</td>
<td>473x</td>
<td>50x</td>
<td>119x</td>
<td>27x</td>
<td>527x</td>
</tr>
<tr>
<td>Number of contigs</td>
<td>269</td>
<td>1723</td>
<td>377</td>
<td>4775</td>
<td>136</td>
<td>1231</td>
<td>527</td>
</tr>
<tr>
<td>Number of scaffolds</td>
<td>249</td>
<td>1723</td>
<td>345</td>
<td>ND</td>
<td>12</td>
<td>ND</td>
<td>108</td>
</tr>
<tr>
<td>Number of Large Scaffolds (&gt;100 Kb)</td>
<td>80</td>
<td>9</td>
<td>89</td>
<td>ND</td>
<td>8</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>N50 (base pairs)</td>
<td>422053</td>
<td>31189</td>
<td>376691</td>
<td>17305</td>
<td>5330000</td>
<td>48518</td>
<td>5031090</td>
</tr>
<tr>
<td>GC content (%)</td>
<td>47.6</td>
<td>48.2</td>
<td>47.6</td>
<td>48.5</td>
<td>ND</td>
<td>48.24</td>
<td>47.1</td>
</tr>
<tr>
<td>Number of protein-coding genes</td>
<td>11048</td>
<td>10683</td>
<td>11070</td>
<td>11476</td>
<td>1830</td>
<td>10554</td>
<td>11770</td>
</tr>
</tbody>
</table>
Table 3.- Number of SNPs (total, synonymous and non-synonymous SNPs), density of SNPs (SNPs/Kb), πN/πS ratio, Nei index (Hs) and minimum number of recombination events (Rmin) per Kb of the *P. digitatum* and *P. expansum* genome strains using as a reference PHI26 and PEXP strain, respectively.

<table>
<thead>
<tr>
<th></th>
<th>P. expansum strains</th>
<th>P. digitatum strains</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PEX1</td>
<td>PEX2</td>
</tr>
<tr>
<td>SNPs</td>
<td>42559</td>
<td>56969</td>
</tr>
<tr>
<td>SNPs syn</td>
<td>8665</td>
<td>10020</td>
</tr>
<tr>
<td>SNPs nonsyn</td>
<td>6392</td>
<td>7124</td>
</tr>
<tr>
<td>πN/πS</td>
<td>0.24</td>
<td>0.24</td>
</tr>
<tr>
<td>SNPs/Kb</td>
<td>1.368</td>
<td>1.758</td>
</tr>
<tr>
<td>SNPs/Kb average</td>
<td>2.44</td>
<td></td>
</tr>
<tr>
<td>Hs</td>
<td>27.89 x 10^-4</td>
<td></td>
</tr>
<tr>
<td>Rmin/Kb</td>
<td>0.036</td>
<td></td>
</tr>
</tbody>
</table>

* In *P. expansum*, the Rmin was obtained using the strains: PEXP, PEX1, NRRL 62431 and ATCC 24692.