BMC Genomics

RESEARCH ARTICLE Open Access

Sequencing and annotation of the *Ophiostoma ulmi* genome

Shima Khoshraftar[1], Stacy Hung[3], Sadia Khan[1], Yunchen Gong[2], Vibha Tyagi[5], John Parkinson[3,4], Mohini Sain[5], Alan M Moses[1*] and Dinesh Christendat[1*]

Abstract

Background: The ascomycete fungus *Ophiostoma ulmi* was responsible for the initial pandemic of the massively destructive Dutch elm disease in Europe and North America in early 1910. Dutch elm disease has ravaged the elm tree population globally and is a major threat to the remaining elm population. *O. ulmi* is also associated with valuable biomaterials applications. It was recently discovered that proteins from *O. ulmi* can be used for efficient transformation of amylose in the production of bioplastics.

Results: We have sequenced the 31.5 Mb genome of *O.ulmi* using Illumina next generation sequencing. Applying both *de novo* and comparative genome annotation methods, we predict a total of 8639 gene models. The quality of the predicted genes was validated using a variety of data sources consisting of EST data, mRNA-seq data and orthologs from related fungal species. Sequence-based computational methods were used to identify candidate virulence-related genes. Metabolic pathways were reconstructed and highlight specific enzymes that may play a role in virulence.

Conclusions: This genome sequence will be a useful resource for further research aimed at understanding the molecular mechanisms of pathogenicity by *O. ulmi*. It will also facilitate the identification of enzymes necessary for industrial biotransformation applications.

Background

Ophiostomoids are the most common Mycelial fungi associated with bark beetles. Within this group is *Ophiostoma ulmi,* the causative agent of the first incident of one of the most destructive plant diseases, Dutch elm disease (DED), starting from the early 1910s in Europe and North America [1]. The far more aggressive species *Ophiostoma novo-ulmi* accounts for a second DED pandemic which was initially recorded in Britain and is believed to have spread to North America from Central Europe in the early 1940s [2,3]. As a consequence of both occurrences, the majority of mature Dutch elm trees were destroyed in North America, Europe and central and southwest Asia. These incidents had tremendous economic impacts on the global forestry and horticultural industries. Unfortunately, bark beetle disease is still a major threat to the remaining North American elm trees, especially in Western Canada yet very few resources are directed towards their control because the molecular basis for *O. ulmi* pathogenicity is still not understood [4-6].

DED is a result of the bark beetle attacking the bark of trees and penetrating into the soft tissue where they feed on nutrients within the phloem [7,8] Concurrently, Ophiostoma fungi are transferred by the beetles to the phloem network where they colonize on the soluble tissues and block the transport of nutrients and water throughout the trees. This colonization of Ophiostoma and competition for nutrients produces an irreversible disease phenotype in mature elm trees, leading to the eventual death of these trees. The two subgroups *O. novo-ulmi* and *O. ulmi* are classified as aggressive and non-aggressive, with *O. novo-ulmi* being the aggressive species [9-12]. Further, they have distinct biological differences, such as growth rate, temperature optimum and colony appearance. As expected, the non-aggressive species is a weak elm pathogen in contrast to the aggressive *O. novo-ulmi* species, but both species produce

* Correspondence: alan.moses@utoronto.ca; dinesh.christendat@utoronto.ca
[1]Department of Cell & Systems Biology University of Toronto, Toronto, Canada
Full list of author information is available at the end of the article

characteristically distinct levels of the toxin protein, cerato-ulmin which is important in protecting infectious propagules from desiccation and is thought to act as a parasitic fitness factor for the organism [13-18].

Interestingly, attempts to cross *O. novo-ulmi* with *O. ulmi* are frequently rejected by the aggressive *O. novo-ulmi* female fungi. The hybrid progenies from successful crosses are usually of low competitive fitness and show low growth rate, decreased pathogenicity, low cerato-ulmin protein production and usually sterile females [19,20]. The differences in biological traits between the aggressive and nonaggressive species point towards incompatibility in their genomic composition. Thus, the genome sequence of these two species would present a unique opportunity for comparative analysis to understand the basis for their pathogenicity, and genomic incompatibility.

Biotransformation strategies have been developed for the use of *O. ulmi* protein extracts in the production of thermoplastic materials. While the protein identities and composition of such mixtures remain uncharacterized due to a lack of an available genome sequence, the quality and consistency of the thermoplastics produced is sufficient for the manufacturing of certain products [21]. Such approaches are highly attractive from an environmental standpoint for the use of renewable resources in manufacturing processes. Therefore, the application of *O. ulmi* has gained tremendous interest in recent years and has resulted in multiple patents. Unlike white-rot and brown-rot fungi (Phanerochaete) whose genomes are sequenced and annotated and are used in a plethora of commercial applications [22], *O. ulmi's* recent emergence in commercial application was based solely on its ability to modify plants polysaccharides. This rather coarse approach in utilizing *O. ulmi* protein extracts in polysaccharides biotransformation is restricted because of a lack of a sequenced genome. Similar to the Phanerochaete fungi, the sequencing of *O. ulmi* would provide tremendous opportunities for its use in industrial applications.

Here, we report a first draft of the genome sequence and annotation of *Ophiostoma ulmi* strain W9. To validate the quality of the gene annotations, we employed EST sequences from *Ophiostoma novo-ulmi* [23], mRNA-seq sequences, and ortholog sequences from three other fungi, *Grosmannia clavigera* [24,25] and two model organisms *Neurospora crassa* [26] and *Saccharomyces serevisiae* [27]. We found that multiple lines of evidence support the quality of the gene annotations. An initial search for genes involved in the pathogenicity of the fungus was performed. The availability of the complete genome sequence should facilitate further studies of *Ophiostoma ulmi* and may be an important step toward development of molecular strategies for controlling DED.

Results

Genome sequence assembly and annotation

Using a whole genome shotgun sequencing strategy, we sequenced the *O. ulmi* genome to an average coverage of 200× by paired-end sequencing (see Methods). A 31.5-Mb genome sequence was obtained by assembling approximately 164 million Illumina reads. Genome statistics are given in Table 1.

To annotate the genome, we combined an *ab initio* method with a comparative gene-finding approach. First we obtained gene models using the *ab initio* method GeneMark ES-v2.0 [28]. The main motivation for this approach was that it takes as input data the genomic sequence alone and requires no other input data such as training sets of known genes from *O. ulmi* or genes from other species. In order to identify conserved genes that were not found by the *ab initio* method, we used Exonerate [29] to align protein sets from *N. crassa* [26] and *G. clavigera* [24,25]. We chose *N. crassa* and *G. clavigera* because they are fully sequenced and annotated and both organisms are closely related to *O. ulmi*. This gene prediction strategy yielded 8639 genes in the *O. ulmi* genome, covering 45% of the total genome. In our final gene set, the vast majority of gene models (8553) were taken from the *ab initio* gene-finding method because most of them have start and stop codons in comparison to genes predicted by Exonerate. Comparative genome analysis revealed that approximately 71% of the annotated genes have orthologs from at least one of the three species, *G. clavigera*, *N. crassa* and *S.cerevisiae*.

We next compared the features of the *O. ulmi* genome with those reported for *G. clavigera* and *N. crassa*. The G + C content for all three species is approximately 50%. The number of predicted protein-coding genes for *O. ulmi*, *G. clavigera* and *N. crassa* varies considerably (8639, 8314 and 10,082, respectively) but the average gene density of *O. ulmi* is one gene per 3.6 kb which is very similar to that of *G. clavigera* and *N. crassa*, having one gene per 3.5 kb and one per 3.7 kb respectively. The average gene length of *O. ulmi* is 1.85 kb, slightly higher

Table 1 General characteristics of the *O. ulmi* genome

Size assembled genome (Mb)	31.5
GC content overall (%)	50.02
GC content (coding) (%)	57.8
Protein coding genes	8639
Gene density (genes/kb)	1/3642
Average gene length (bp)	1854
Average number of introns per gene	1.14
Median intron size (bp)	67
Median exon size (bp)	395

than 1.673 kb for *G. clavigera* and 1.67 kb for *N. crassa* (Figure 1a). The mean number of introns per gene is 1.14 for *O. ulmi*, which is somewhat less than 1.86 for *G. clavigera* and 1.7 for *N. crassa* (Figure 1b). The sizes of introns in *O. ulmi* are similar for the three fungal species compared (Figure 1c).

Identification of *O. ulmi* gene orthologs and phylogenetic analysis

To evaluate the evolutionary relationships between *O. ulmi* and other fungi where a complete genome sequence was available, we identified orthologs of *O. ulmi, G. clavigera, N. crassa, and S. cerevisiae*, using Inparanoid [30]. Applying this approach we identified 5784 ortholog with *G. clavigera*, 5517 with *N. crassa* and 2483 with *S. cerevisiae*. Phylogenetic analysis was performed with the amino acid sequences of the 2215 genes for which we have a one-to-one ortholog relationship for each species with *O. ulmi.* We computed the likelihood of four possible trees with *S.cerevisiae* as the outgroup using PAML [31]. The tree, shown in Figure 2, was found with 90% bootstrap support for more than 80% of the orthologous groups. To estimate the distances between species, we took the median value of each branch length (in substitutions per site, Figure 2) from the 2215 genes for which we have a one-to-one ortholog relationship for each species. The total distance between *G. clavigera* and *O. ulmi* is more than 0.2 subs. per site, which is consistent with a previous analysis on two genes [32]. Our analysis indicates that *O. ulmi* is a representative of a divergent clade for which, to our knowledge, no complete genome sequences currently exist.

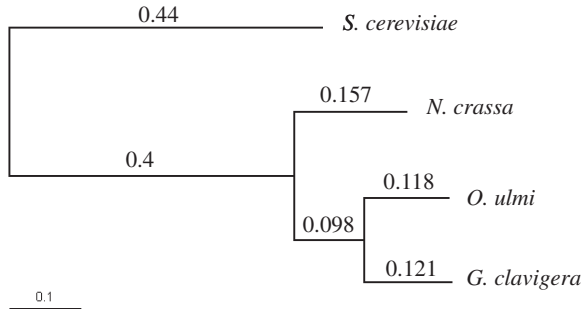

Figure 2 Phylogenic tree. Phylogenetic tree between *O. ulmi* and three other species *G. clavigera*, *N. crassa* and *S. cerevisiae*. Branch length represents the phylogenetic distances in substitution per site.

Validation of genome annotation

In order to provide support for our gene models, we employed three sets of data. First, we used the published expressed sequence tags (EST) library from *O. novo-ulmi* [23]. Using est2genome model with default parameters from Exonerate [29], we mapped the EST data to the *O.ulmi* genome. We then looked for gene models

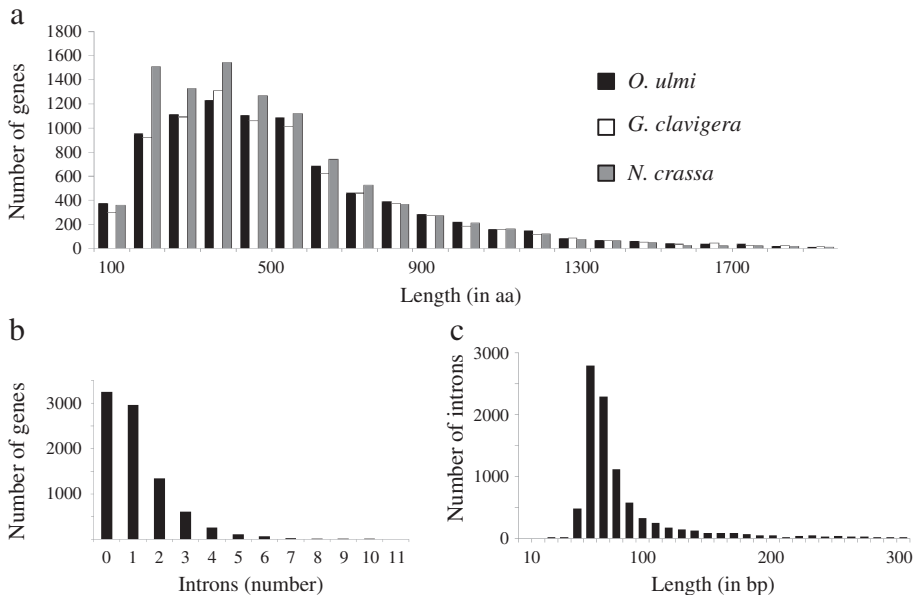

Figure 1 Overview of genome annotation statistics. a) the distribution of the lengths of the genes of the *O. ulmi* with *G. clavigera* and *N. crassa*. **b**) the number of introns in each of the predicted genes in *O. ulmi*. Most of the predicted genes have either zero or one introns and a few of them have more than 5 introns. **c**) distribution of the length of the introns of the gene models. The majority of the introns are 60 bp to 200 bp and 3% of them are between 200 bp and 300 bp. Almost 8% are longer than 300 bp but are not shown in the figure for the purpose of clarity.

that overlapped with positions in the genome mapping back to the EST data. We found that 91% of the gene models had expression evidence from EST data.

We next compared our annotation to *O. ulmi* mRNA-seq read data we generated. Mapping mRNA-seq reads back to the genome using a short read aligner (Bowtie [33]), we found the average coverage to be 30.4 reads per base pair. Computing the coverage for coding segments of the gene models (See Methods), we found the coverage in coding regions to be 48.1 (reads per coding region base pair). Thus, the average coverage is 58% higher within our predicted coding regions than in the genome overall, providing evidence that our gene predictions are enriched for *bona fide* genes in *O. ulmi* (Figure 3). Despite this high average coverage, approximately 20% of the gene models have coverage of less than 1. We speculate that this may result from either lack of expression in the condition that the mRNA was extracted, or errors in gene prediction.

A typical distribution of the mRNA-seq reads for a region of 20 kb in the genome is shown in Figure 4. Gene o753 exemplifies a high-confidence gene prediction because there are approximately 20 reads covering it and the coverage drops to zero at the predicted intron-exon boundaries. On the other hand, gene o758 is an example of imperfect prediction because there is evidence for expression outside of the predicted coordinates.

As part of our evaluation of the gene model predictions, we have also included gene models with orthologs to at least one of the three other comparative species *G. clavigera*, *N.crassa* and *S. cerevisiae*. Significantly, 70.4% of our predictions were found to have 1-to-1 orthologs in other species and given that orthologs are often similar in function [34], this suggests that most of our predicted genes can be assigned a function based on comparison with genes of other well-studied organisms. This further supports the assertion that the gene models described here are accurate.

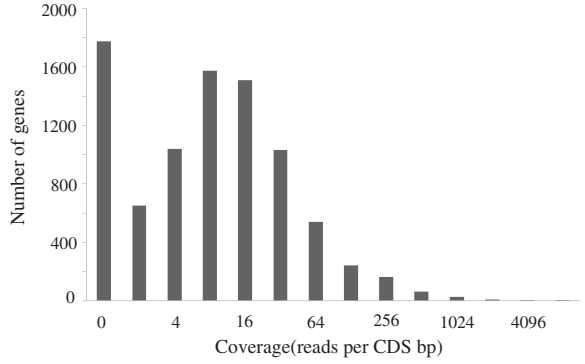

Figure 3 Distribution of coverage for the coding regions of the gene models. The coverage for the coding regions of the predicted genes is shown. Almost 80% have the coverage greater than 1.

Overall, ~99% of the gene models have at least one type of evidence with ~62% of them having all three types of evidence (Figure 5). This analysis indicates that our draft genome assembly and gene annotation is of high quality.

Protein domain analysis

Previous studies suggested that specific genes of a pathogen are important for its pathogenicity [35,36]. For the fungi *O. ulmi* and the closely related species *O. novo-ulmi*, it was suggested that a hydrophobic protein Cerato-ulmin and a colony type gene col1 were directly correlated to the fungi causing DED in elm trees [37-39]. To check for the occurrence of other related genes (homologues), we searched our predicted gene set using these genes as queries. For both the hydrophobic protein Cerato-ulmin and col1, high confident single matches are found using Blast-2.2.25 [40] (e-value 6e-51 and 5e-86 respectively). While consistent with the accuracy of our gene predictions, these searches did not identify new pathogenicity related genes.

Using the sequenced *O. ulmi* genome, we performed a global domain analysis by searching the entire predicted gene set for protein domains from the Pfam database [41]. In total, 5069 protein domains were found in our *O. ulmi* gene set. Comparison of the protein domains amongst the three fungal species showed that *O. ulmi* has 3993 families in common with *G. clavigera* and 4155 families in common with *N. crassa,* while 605 of the identified domains are found only in *O. ulmi*. However, of those, 205 are domains of unknown function with little information available for them. The remaining 400 unique protein domains are not among those known to play a crucial role in pathogenicity and host-plant cell-wall degradation, such as glycosyl hydrolase, glycosyl transferase and oxidases. Further, we did not observe significant expansion of protein families important in virulence. For instance, the glycosyl hydrolase family is represented by 145 genes in *O. ulmi* compared with 130 genes in *G. clavigera* and 167 genes in *N. crassa*. Overall, our comparison of domain content with *G. clavigera* and *N. crassa* suggests that these three species are very similar. Furthermore, the gene families appear to be highly conserved since no outstanding expansions of protein domain families could be detected in *O. ulmi*.

As another approach to identify virulence factors, we searched for gene models previously known to be associated with pathogenicity in other organisms. To do this, *O. ulmi* gene models were compared against PHI-base [36], a database of 924 fungal-verified virulence and pathogenicity related genes (Table 2)[42,43]. Comparing our results to *G. clavigera* and *N. crassa,* we find that the number of genes from *O. ulmi*, *G. clavigera*, and *N. crassa* with matches to pathogenicity related genes

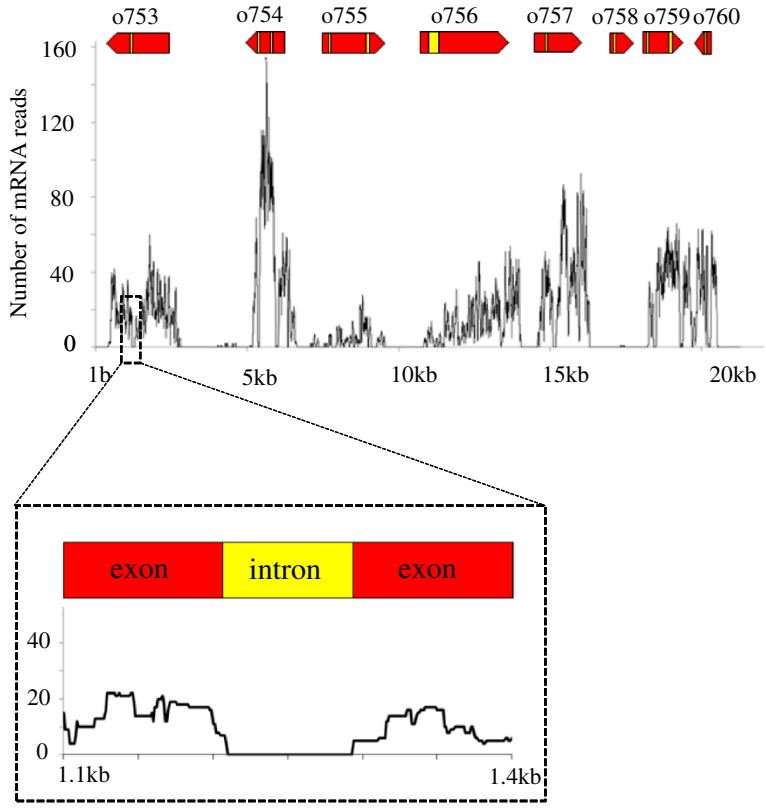

Figure 4 Examples of mRNA-seq coverage. The number of mRNA-seq reads mapped is plotted as a function of genomic coordinate. Eight predicted genes are displayed by red arrows, (o753 to o760). Yellow sections inside the red arrows refer to the intron parts of the gene models. A zoomed version of the area in which the intron appears is given in the box below the figure.

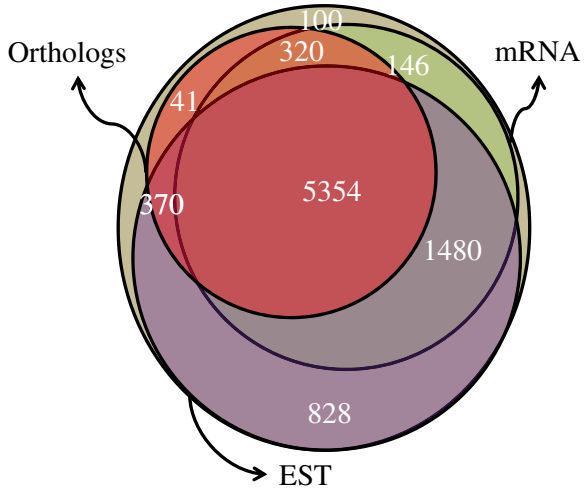

Figure 5 Summary of genome annotation validation. The biggest circle shows the total number of gene models predicted and every other circle represents a subset of gene models that are supported by any of the three types of evidences. The gene models that have evidence from EST data from *O. novo-ulmi* are referred to by EST data in the figure, those having ortholog genes from at least one of the mentioned three species referred to by Orthologs and those with evidence from mRNA-seq data from *O. ulmi* referred to by mRNA-seq.

from PHI-base were similar across all three species (610, 598, and 611, respectively); this indicates that there is no obvious expansion in the number of pathogen related genes in *O. ulmi*. While total numbers were similar, a small number of genes from PHI-base represented unique matches for each species (Table 2). These genes are important because they could be responsible for specific characteristics in each species. Three PHI-base genes were found to match only to *O. ulmi* (Table 3). One of these genes, CTB7 from *Cercospora nicotianae*, is annotated as a putative FAD/FMN- or NADPH-dependent oxidoreductase in the cercosporin toxin biosynthetic pathway of *C. Nicotianae*. Cercosporin is a toxin which plays an important role in the pathogenicity

Table 2 Comparison of the number of PHI-base pathogen genes found in the three species

Organism	Number of pathogen genes (from PHI-base database) found in the organism	Number of unique pathogens genes found in the organism
O. ulmi	610	3
G. clavigera	598	2
N. crassa	611	7

Table 3 PHI-base pathogen genes found in *O. ulmi* not in *N. crassa* and *G. clavigera*

Pathogen name	Description
PHI:48|CnLAC1|BAD91825|TX:5207|Cryptococcus neoformans|Reduced virulence	A laccase enzyme which catalyzes the synthesis of melanin in the presence of phenolic compounds [43]
PHI:876|MGG_11671|EDK03349|TX:148305| Magnaporthe grisea|Reduced virulence	hypothetical protein similar to reverse transcriptase
PHI:1048|CTB7|ABK64184|TX:29003|Cercospora nicotianae|Reduced virulence	Encodes putative FAD/FMN- or NADPH-dependent oxidoreductases in the cercosporin toxin biosynthetic pathway of C. nicotianae [44]

of many phytopathogenic *Cercospora* species [44] and may therefore represent an important virulence factor in *O. ulmi*.

Metabolic network reconstruction

With the Pfam and PHI-base analyses indicating overall gene content of *O. ulmi* to be similar to other organisms, we attempted a more detailed analysis of the pathogen's metabolism. Reconstruction of the metabolic network for *O. ulmi* was achieved by integrating several automated datasets together with ortholog mappings to *S. cerevisae*. In total 1,378 genes (representing 16% of the genome) map to enzymatic activity based on EC number annotation. This number aligns well with the yeast consensus metabolic reconstruction, consisting of 832 genes (representing 13% of its genome) [45]. *O. ulmi* shares 79% (615/783) of its enzymes with *S. cerevisae*.

Mapping *O. ulmi* enzymes to KEGG metabolic pathways provides a unique perspective to highlight pathways that are either conserved amongst fungi (represented by yeast) and have been lost in *O. ulmi*, or that are unique to *O. ulmi* and may play a role in causing disease in its host, the elm tree (represented by the model plant genome *A. thaliana*). For instance, pathways for fatty acid biosynthesis and the metabolism of D-arginine and D-ornithine have noticeably fewer enzymes in *O. ulmi* compared to *S. cerevisae* (Figure 6). Other pathways such as anthocyanin biosynthesis and photosynthesis related are plant-specific and as expected appear to be missing from the fungal species. Of interest are a number of pathways that have noticeably greater proportions of enzymes in *O. ulmi* compared to *S. cerevisae* including glycosphingolipid biosynthesis (ganglio series), glycan degradation (N-glycans, gangliosides), lipoic acid metabolism, and drug metabolism by enzymes other than cytochrome P450 (see Figure 6). These activities, particularly those related to glycan degradation, are likely to play an important role in the success and survival of *O. ulmi* as a phytopathogen.

Endopolygalacturonase (ePG) has been identified in *O. ulmi* (DETECT prediction: o7823), and is involved in cell wall degradation. ePG belongs to the polygalacturonase (PG) family of enzymes that catalyze the hydrolysis of pectin compounds which comprise 30% of the primary cell wall in plants [46]. Previous studies have implicated PGs as virulence factors in other phytopathogens including *Botrytis cinerea* [47] and *Alternaria citri* [48] where the enzyme could assist host invasion, tissue destruction and similar processes associated with plant disease. A recent study assessing the role of ePG in *O. ulmi*, however, suggests that the enzyme functions as a parasitic fitness factor as opposed to a virulence factor, given that targeted disruption of the gene led to a reduction in pectin-degrading activity and not a lethal phenotype [49]. Other pectinase enzymes such as pectin methylesterase (BLAST and PRIAM prediction: o5231) and pectinase (BLAST and PRIAM prediction: o3878) present in *O. ulmi* likely act in concert to contribute to successful invasion of the host. The production of PG enzymes is particularly important for the success and survival of *O. ulmi* as it is a pathogen that enters the host directly through a pre-existing wound and therefore lacks specialized penetration structures [50]. Moreover, a role as a minor virulence factor is possible and when combined with other virulence factors, ePG represents a potential target for the control of DED.

In general, core essential pathways such as those related to amino acid, carbohydrate, energy, and nucleotide metabolism are highly conserved across both fungi and plants (see Figure 6). Interestingly, a recent study by Oliveira *et al.* [51] showed that elm trees inoculated with *O. novo-ulmi* had significantly reduced contents of glucose, fructose, starch and sucrose, suggesting that carbohydrate metabolism pathways are important to the pathogenicity of the fungus. Consistent with this hypothesis, the genes encoding carbohydrate metabolism enzymes seem to be highly expressed in *O. ulmi* based on our mRNA-seq data (Figure 6). Moreover, functional categorization of an EST library for *O. novo-ulmi* revealed that the majority of EST sequences associated with metabolism had the greatest representation in carbohydrate metabolism [23]. These results suggest that, while metabolic reconstruction predicts *O. ulmi* has similar enzyme complements for a number of pathways such as those involved in carbohydrate metabolism to *S. cerevisae* and *A. thaliana*, expression profiles are an essential component to assessing the functionality of specific pathways. In addition, the close phylogenetic

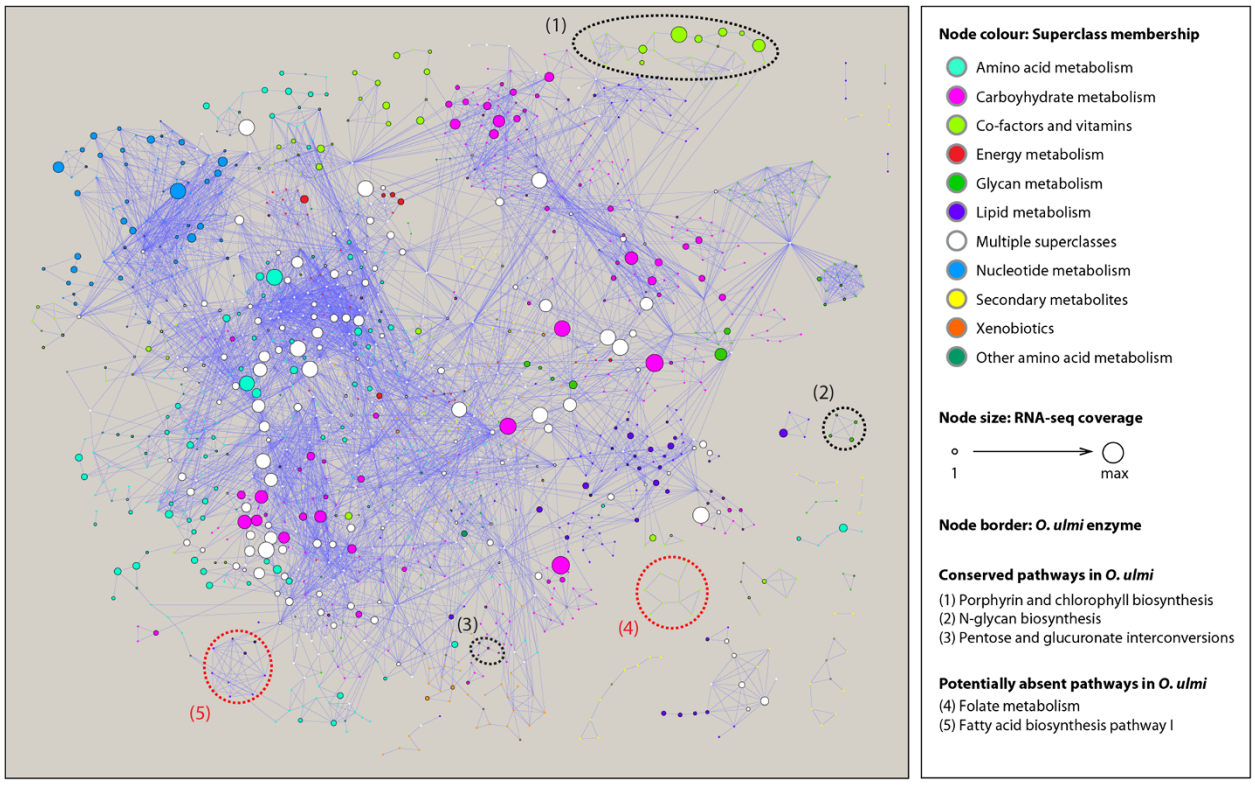

Figure 6 Metabolic network of *O. ulmi*. Node size corresponds to RNA-seq expression coverage, with background nodes (borderless) included for context of KEGG pathways. Groups of enzymes dotted-circled in black appear to be conserved in the fungus while enzymes dotted-circled in red appear to be absent in the fungus.

relationship to *O. novo-ulmi* might also hint that the pathogenic role of *O. ulmi* is at least partially a result of decreasing the plant's carbohydrates, consequently reducing the efficiency of photosynthesis, and eventually leading to plant senescence.

Discussion

Ophiostoma ulmi caused the first emergence of DED, one of the most destructive plant diseases in the last 100 years [1]. In addition, because of its starch modification characteristics, *O. ulmi* has been used in industry for bioplastic production [21]. However, compared to the more aggressive species *O. novo-ulmi*, little is known about the basic biology of *O. ulmi*.

In this paper, we sequenced the genome of *O. ulmi* using next generation sequencing. Our genome sequence annotation contains 8639 gene models. EST data from the closely related species *O. novo-ulmi*, mRNA-seq data and orthologous genes in other species provide strong evidence for the quality of our annotation. Using genome-scale analysis, we estimated the phylogenetic relationship and distance of *O. ulmi* to *N. crassa* and *G. clavigera*. Finally, we compared the protein domains and matches to PHI-base in our gene models with two other fungal species, *G. clavigera* and *N. crassa* to search for genetic features that may yield important clues about the different lifestyles of the species.

Through metabolic reconstruction, we identified certain families of enzymes that may play a role in the virulence of the fungus. Significantly, we identify a cell wall degrading enzyme, ePG, which may be involved in host-pathogen interactions. The contribution of the gene to virulence has been examined in other fungi, with evidence demonstrating that ePG is required for full virulence and infection of the host tissue [47]. Nevertheless, the full role of the enzyme in pathogenicity for *O. ulmi* has yet to be elucidated.

Conclusion

Our contribution here was to generate a high-quality genome sequence and annotation for *O. ulmi*. With this in hand, future research will achieve a deeper understanding of the processes by which the fungi colonize and, break down cell wall components and damage elm trees. Furthermore, because of the starch modification and plastic improvement features of the fungus, availability of the fungus genome may help develop new industrial processes for bioplastic production.

Methods

Genomic DNA extraction

DNA was extracted from *Ophiostoma ulmi* using Qiagen kit. The fungus was cultured for three days and then after centrifuging the pellet (spores) was washed 3–4 times with distilled water and centrifuged again. The spores were then freeze dried. 500 mg of these spores were crushed using liquid nitrogen and then the powder was suspended in 30 ml of Cell Suspension Solution. 150 μl of Cell lytic solution was then added to it. This was mixed by inverting the tubes 25 times and then incubated at 37 C for 30 minutes. The mixture was then centrifuged and the pellet was suspended in 30 ml of cell suspension solution and 10 ml of Protein precipitation solution. This was mixed by vortex and centrifuged for 3 minutes. The supernatant from this was added to 30 ml of Isopropanol and mixed well. After centrifuging for 1 minute the pellet was washed carefully with 70% Ethanol. This was again centrifuged for 1 minute and then the pellet was air dried. The pellet was then suspended in 5 ml of DNA Hydration Solution and 150 μl of RNase A solution was added and incubated at 37°C for 60 minutes and at 65°C for 60 minutes to dissolve the DNA. This was incubated at room temperature overnight with gentle shaking. The Purity of DNA was evaluated by determining its spectroscopic ratio at A260/A280 nm.

Genome sequencing and assembly

Genomic DNA was sequenced with Illumina GAIIx as paired-end (PE) reads and mate-paired (MP) reads. Insert size for PE library is ~220 base pairs, and insert size for MP library is ~3000 base pairs. In total we obtained 64,563,784 pairs of PE reads of length 38 base pairs, and 39,690,603 pairs of MP reads of length 40 base pairs. Quality reads are extracted based on the criteria of Illumina pipeline for genome assembly, they are of 86% of the PE reads and 89% of the MP reads. We carried out two rounds of *de novo* assembly of the genome. In the first round, PE and MP reads are trimmed to different lengths and assembled with Velvet [52] separately. For each length several k-mer sizes were tested and the length that gave the maximum N50 was identified. In the second round, the PE and MP reads of the best length from the first round were assembled together with Velvet using several k-mer sizes, and the assembly with the maximum N50 was picked up as the final assembly. There are 3,415 contigs in the final assembly, the largest contig contains 3,256,915 base pairs, and total base pairs of all contigs are 31,466,092. N50 of the final assembly is 1,009,735 base pairs. 164,295,551 reads out of 180,796,760 total reads were used in the final assembly.

Gene prediction methods

Ab initio gene prediction has been done using GeneMark-ES-v2.0 [28] with default parameters. We used this method for two reasons. First, it is designed specifically for fungus genome sequences. In addition, unlike other gene prediction methods, it does not have the bottleneck of a large training set to train their underlying model. It computes the model parameters from the genome sequence. We aligned the protein sequences from the other species *G.clavigera* and *N.crassa* to *O. ulmi* sequences using Exonerate [29]. For protein comparison, the protein2genome model was used and the bestn parameter was set to 1 to find the best matches to the protein sequences. However, Because of the possibility of gene duplication and gene expansion we also included the genes predicted using bestn 10 parameter which were not overlapping with bestn 1 gene models. Our final gene model set was the combination of the genes predicted by *ab initio* and comparative gene predictors. Initially the set comprised of the genes predicted by *ab initio* gene predictor and then the genes that are obtained from comparison with *G. clavigera* and *N. crassa* protein sequences were added to the set if they do not overlap the initial gene set.

mRNA-seq

O. ulmi was grown and harvested under similar condition described above for its genomic DNA extraction. Total RNA isolation was carried out using a Qiagen RNA preparation kit (Qiagen Inc., Mississauga, ON, Canada) by following the supplier instructions for filamentous fungi. cDNA was synthesized at CAGEF using mRNA-seq sample preparation kit following the supplier instructions (Illumina Inc., San Diego, CA). *O. ulmi* mRNA was sequenced with Illumina GAIIx as paired-end (PE) reads as described above. We had approximately 93 million paired-end reads of length 38. In order to evaluate the quality of predicted genes, we mapped the reads back to the genome sequence using Bowtie.0.12.7 [11]. The bowtie-build command was used to build an index from the genome sequence and then we ran Bowtie using bowtie command with default parameters. Approximately 30% of the mRNA-seq read data was mapped to the genomic DNA sequence. The rest of the read data were of low quality and could not be aligned to the sequence. Using the coordinates of mapped reads, the overall average coverage and the coverage for the coding regions of each gene was calculated by dividing the total length of the reads by the total number of base pairs for every desired region.

Phylogenetic analysis

We found orthologous groups among four species *O. ulmi*, *G. clavigera*, *N. crassa* and *S. cervisiae* using Inparanoid_4.1 [30] with the default settings. First the

protein sequences for all the four fungi were searched against each other using BLAST with the default parameters and then the orthologous groups were identified between every two species. We employed these pairwise ortholog groups in building files which contains four gene models each from one of the species and the gene models were pairwise orthologs. This resulted in 2215 files. Then we aligned the gene models for each file using the multiple sequence aligner MAFFT [53] and the phylogenetic analysis was performed with PAML [31]. The parameters for running PAML were as follows: Empirical amino acid substitution model and removing gaps columns. For each alignment we computed the likelihood of four trees: tree 1 (((*O. ulmi*, *G. clavigera*), *N. crassa*), *S. cerevisiae*), tree 2 ((*O. ulmi*, (*G. clavigera*, *N. crassa*)), *S. cerevisiae*), tree 3 ((*O. ulmi*, *G. clavigera*, *N. crassa*), *S. cerevisiae*) and tree 4 (((*O. ulmi*, *N. crassa*), *G. clavigera*), *S. cerevisiae*). 1926 of the alignments (86%) support the tree 1 with 90% bootstrap support cutoff.

Metabolic reconstruction

The gene model for *O. ulmi*, containing 8, 639 genes, was searched against the SwissProt-Uniprot protein database (v 58.0) using the following homology-based enzyme prediction tools: (i) DETECT [54] (cutoff ILS > 0.2, at least 5 positive hits), (ii) BLAST (E-value > 1e-10), (iii) PRIAM [55] (E-value > 1e-10), and (iv) ortholog mappings to Yeast based on OrthoMCL [56,57]. No pathway data for *Ophiostoma* was available from KEGG. The BRENDA resource (Barthelmes, *et al.*, 2007) provided biochemical evidence for four enzymes. The final set of 783 enzymes from *O. ulmi* was obtained by integrating the datasets from BRENDA, DETECT, Yeast orthologs, and enzymes identified by both BLAST and PRIAM. See (data on our website) for gene-EC mappings and for corresponding evidences. Yeast and *Arabidopsis thaliana* EC numbers were obtained by combining species-specific datasets from BRENDA and BioCyc (YeastCyc and AraCyc, respectively).

Pathway heatmap

Ratios of enzyme complements from *O. ulmi*, Yeast, and *A. thaliana* were calculated based on KEGG pathways and grouped according to superclass. Note that KEGG incorporates information for all species available, so many pathways may include enzymes that are not relevant leading to misleading interpretations of pathways that might appear absent or present in the species.

Data availability

The sequence and annotation data are available at (www.moseslab.csb.utoronto.ca/o.ulmi). These include genome sequence, datasets for genes and proteins, a summary of the results from Pfam analyses and a Blast server.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SK (Shima) drafted the manuscript, carried out genome annotation and bioinformatics analyses. SH reconstructed the metabolic network, participated in drafting the manuscript. SK(Sadia), prepared RNA from *O. ulmi*. YG performed genome assembly, participated in drafting the manuscript. VT prepared genomic DNA from *O. ulmi*, participated in drafting the manuscript. JP supervised SH. MS co-supervised SK(Sadia) and VT. AMM supervised SK(Shima) and edited the manuscript. DC co-supervised SK(Sadia) and VT, interpreted the sequencing data and prepared the manuscript. All authors read and approved the final manuscript.

Acknowledgments

This project was funded by the Natural Science and Engineering Council (NSERC) of Canada and the Ontario BioCar initiative. The sequencing of *O. ulmi* genomic and RNA samples was conducted at the Center for the Analysis of Genome Evolution and Function (CAGEF).

Author details

[1]Department of Cell & Systems Biology University of Toronto, Toronto, Canada. [2]Center for Analysis of Genome Evolution and Function, Toronto, Canada. [3]Departments of Biochemistry and Molecular Genetics University of Toronto, Toronto, Canada. [4]Molecular Structure and Function Hospital for Sick Children, Toronto, Canada. [5]Centre for Biocomposites and Biomaterials Processing, Faculty of Forestry, Toronto, Canada.

Received: 1 November 2012 Accepted: 28 February 2013
Published: 12 March 2013

References

1. Brasier CM: **Recent genetic changes in the Ophiostoma ulmi population: the threat to the future of the elm**. In *Populations of plant pathogens: Their dynamics and Genetics*. Edited by Wolfe MS, Caten CE. Oxford: Blackwell Publ; 1987:213–226.
2. Brasier CM: **Ophiostoma novo-ulmi sp. nov., causative agent of current Dutch elm disease pandemics.** *Mycopathologia* 1991, **115:**151–161.
3. Lieutier F, Day KR, Battisti A, Grégoire JC, Evans HF: *Bark and Wood Boring Insects in Living Trees in Europe, A Synthesis.* 1st edition. Springer; 2007. 2nd printing.
4. Gagné P, Yang DQ, Hamelin RC, Bernier L: **Genetic variability of Canadian populations of the sapstain fungus ophiostoma piceae.** *Phytopathology* 2001, **91:**369–376.
5. Massoumi Alamouti S, Kim J-J, Humble LM, Uzunovic A, Breuil C: **Ophiostomatoid fungi associated with the northern spruce engraver, Ips perturbatus, in western Canada.** *Antonie Van Leeuwenhoek* 2007, **91:**19–34.
6. Temple B, Pines PA, Hintz WE: **A nine-year genetic survey of the causal agent of Dutch elm disease, Ophiostoma novo-ulmi in Winnipeg, Canada.** *Mycol Res* 2006, **110:**594–600.
7. Martín JA, Solla A, Coimbra MA, Gil L: **Metabolic distinction of Ulmus minor xylem tissues after inoculation with Ophiostoma novo-ulmi.** *Phytochemistry* 2005, **66:**2458–2467.
8. Bleiker KP, Six DL: **Effects of water potential and solute on the growth and interactions of two fungal symbionts of the mountain pine beetle.** *Mycol Res* 2009, **113:**3–15.
9. Hubbes M: **The American elm and Dutch elm disease.** *Forestry Chronicle* 1999, **75:**265–273.
10. Burges HD, Grove JF, Pople M: **The internal microbial flora of the elm bark beetle, Scolytus scolytus, at all stages of its development.** *J Invertebr Pathol* 1979, **34:**21–25.
11. Brasier CM, Kirk SA: **Designation of the EAN and NAN races of Ophiostoma novo-ulmi as subspecies.** *Mycol Res* 2001, **105:**547–554.
12. Jeng R, Hintz WE, Bowden CG, Horgen PA, Hubbes M: **A comparison of the nucleotide sequence of the cerato-ulmin gene and the rDNA ITS between aggressive and non-aggressive isolates of Ophiostoma ulmi sensu lato, the causal agent of Dutch elm disease.** *Curr Genet* 1996, **29:**168–173.

13. Abraham LD, Breuil C: **Isolation and characterization of a subtilisin-like serine proteinase secreted by the sap-staining fungus Ophiostoma piceae.** *Enzyme Microb Technol* 1996, **18:**133–140.
14. Del Sorbo G, Scala F, Parrella G, Lorito M, Comparini C, Ruocco M, Scala A: **Functional expression of the gene cu, encoding the phytotoxic hydrophobin cerato-ulmin, enables Ophiostoma quercus, a nonpathogen on elm, to cause symptoms of Dutch elm disease.** *Mol Plant Microbe Interact* 2000, **13:**43–53.
15. Pazzagli L, Cappugi G, Manao G, Camici G, Santini A, Scala A: **Purification, characterization, and amino acid sequence of cerato-platanin, a new phytotoxic protein from *Ceratocystis fimbriata f. sp. platani*.** *J Biol Chem* 1999, **274:**24959–24964.
16. Hong Y, Cole TE, Brasier CM, Buck KW: **Evolutionary relationships among putative RNA-dependent RNA polymerases encoded by a mitochondrial virus-like RNA in the Dutch elm disease fungus, Ophiostoma novo-ulmi, by other viruses and virus-like RNAs and by the Arabidopsis mitochondrial genome.** *Virology* 1998, **246:**158–169.
17. Tadesse Y, Bernier L, Hintz WE, Horgen PA: **Real time RT-PCR quantification and Northern analysis of cerato-ulmin (CU) gene transcription in different strains of the phytopathogens Ophiostoma ulmi and O. novo-ulmi.** *Mol Genet Genomics* 2003, **269:**789–796.
18. Pipe ND, Brasier CM, Buck KW: **Two natural cerato-ulmin (CU)-deficient mutants of Ophiostoma novo-ulmi: one has an introgressed O. ulmi cu gene, the other has an O. novo-ulmi cu gene with a mutation in an intron splice consensus sequence.** *Mol Plant Pathol* 2000, **1:**379–382.
19. Paoletti M, Buck KW, Brasier CM: **Cloning and sequence analysis of the MAT-B (MAT-2) genes from the three Dutch elm disease pathogens, Ophiostoma ulmi, O. novo-ulmi, and O. himal-ulmi.** *Mycol Res* 2005, **109:**983–991.
20. Paoletti M, Buck KW, Brasier CM: **Selective acquisition of novel mating type and vegetative incompatibility genes via interspecies gene transfer in the globally invading eukaryote Ophiostoma novo-ulmi.** *Mol Ecol* 2006, **15:**249–262.
21. Huang CB, Jeng R, Sain M, Saville B, Hubbes M: **Production, characterization, and mechanical properties of starch modified by ophiostoma SPP.** *BioResources* 2007, **1:**257–269.
22. Suzuki H, MacDonald J, Syed K, Salamov A, Hori C, Aerts A, Henrissat B, Wiebenga A, VanKuyk PA, Barry K, Lindquist E, LaButti K, Lapidus A, Lucas S, Coutinho P, Gong Y, Samejima M, Mahadevan R, Abou-Zaid M, de Vries RP, Igarashi K, Yadav JS, Grigoriev IV, Master ER: **Comparative genomics of the white-rot fungi, Phanerochaete carnosa and P. chrysosporium, to elucidate the genetic basis of the distinct wood types they colonize.** *BMC Genomics* 2012, **13:**444.
23. Hintz W, Pinchback M, Bastide P de la, Burgess S, Jacobi V, Hamelin R, Breuil C, Bernier L: **Functional categorization of unique expressed sequence tags obtained from the yeast-like growth phase of the elm pathogen Ophiostoma novo-ulmi.** *BMC Genomics* 2011, **12:**431.
24. Hesse-Orce U, DiGuistini S, Keeling CI, Wang Y, Li M, Henderson H, Docking TR, Liao NY, Robertson G, Holt RA, Jones SJM, Bohlmann J, Breuil C: **Gene discovery for the bark beetle-vectored fungal tree pathogen Grosmannia clavigera.** *BMC Genomics* 2010, **11:**536.
25. DiGuistini S, Wang Y, Liao NY, Taylor G, Tanguay P, Feau N, Henrissat B, Chan SK, Hesse-Orce U, Alamouti SM, Tsui CKM, Docking RT, Levasseur A, Haridas S, Robertson G, Birol I, Holt RA, Marra MA, Hamelin RC, Hirst M, Jones SJM, Bohlmann J, Breuil C: **Genome and transcriptome analyses of the mountain pine beetle-fungal symbiont Grosmannia clavigera, a lodgepole pine pathogen.** *PNAS* 2011, **108**(9):2504–2509.
26. Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, FitzHugh W, Ma L-J, Smirnov S, Purcell S, Rehman B, Elkins T, Engels R, Wang S, Nielsen CB, Butler J, Endrizzi M, Qui D, Ianakiev P, Bell-Pedersen D, Nelson MA, Werner-Washburne M, Selitrennikoff CP, Kinsey JA, Braun EL, Zelter A, Schulte U, Kothe GO, Jedd G, Mewes W, *et al*: **The genome sequence of the filamentous fungus Neurospora crassa.** *Nature* 2003, **422:**859–868.
27. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG: **Life with 6000 genes.** *Science* 1996, **274:**563–567. 546.
28. Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M: **Gene prediction in novel fungal genomes using an *ab initio* algorithm with unsupervised training.** *Genome Res* 2008, **18:**1979–1990.
29. Slater GSC, Birney E: **Automated generation of heuristics for biological sequence comparison.** *BMC Bioinformatics* 2005, **6:**31.
30. Östlund G, Schmitt T, Forslund K, Köstler T, Messina DN, Roopra S, Frings O, Sonnhammer ELL: **InParanoid 7: new algorithms and tools for eukaryotic orthology analysis.** *Nucleic Acids Res* 2010, **38:**D196–D203.
31. Yang Z: **PAML 4: phylogenetic analysis by maximum likelihood.** *Mol Biol Evol* 2007, **24:**1586–1591.
32. Zipfel RD, de Beer ZW, Jacobs K, Wingfield BD, Wingfield MJ: **Multi-gene phylogenies define Ceratocystiopsis and Grosmannia distinct from Ophiostoma.** *Stud Mycol* 2006, **55:**75–97.
33. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biol* 2009, **10:**R25.
34. Altenhoff AM, Dessimoz C: **Phylogenetic and functional assessment of orthologs inference projects and methods.** *PLoS Comput Biol* 2009, **5:**e1000262.
35. Idnurm A, Howlett BJ: **Pathogenicity genes of phytopathogenic fungi.** *Mol Plant Pathol* 2001, **2:**241–255.
36. Winnenburg R, Baldwin TK, Urban M, Rawlings C, Köhler J, Hammond-Kosack KE: **PHI-base: a new database for pathogen host interactions.** *Nucleic Acids Res* 2006, **34:**D459–D464.
37. Bowden CG, Hintz WE, Jeng R, Hubbes M, Horgen PA: **Isolation and characterization of the cerato-ulmin toxin gene of the Dutch elm disease pathogen, Ophiostoma ulmi.** *Curr Genet* 1994, **25:**323–329.
38. Konrad H, Kirisits T, Riegler M, Halmschlager E, Stauffer C: **Genetic evidence for natural hybridization between the Dutch elm disease pathogens Ophiostoma novo-ulmi ssp. novo-ulmi and O. novo-ulmi ssp. americana.** *Plant Pathology* 2002, **51:**78–84.
39. Dvorak M, Tomsovsky M, Jankovsky L, Novotny D: **Contribution to identify the causal agents of Dutch elm disease in the Czech Republic.** *Plant Protection Science - UZPI* 2007, **43**(4):142–145.
40. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403–410.
41. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer ELL, Eddy SR, Bateman A, Finn RD: **The Pfam protein families database.** *Nucleic Acids Res* 2011, **40:**D290–D301.
42. DiGuistini S, Ralph SG, Lim YW, Holt R, Jones S, Bohlmann J, Breuil C: **Generation and annotation of lodgepole pine and oleoresin-induced expressed sequences from the blue-stain fungus Ophiostoma clavigerum, a Mountain Pine Beetle-associated pathogen.** *FEMS Microbiol Lett* 2007, **267:**151–158.
43. Noverr MC, Williamson PR, Fajardo RS, Huffnagle GB: **CNLAC1 Is Required for Extrapulmonary Dissemination of Cryptococcus neoformans but Not Pulmonary Persistence.** *Infect Immun* 2004, **72:**1693–1699.
44. Chen H-Q, Lee M-H, Chung K-R: **Functional characterization of three genes encoding putative oxidoreductases required for cercosporin toxin biosynthesis in the fungus Cercospora nicotianae.** *Microbiology (Reading, Engl.)* 2007, **153:**2781–2790.
45. Herrgård MJ, Swainston N, Dobson P, Dunn WB, Arga KY, Arvas M, Blüthgen N, Borger S, Costenoble R, Heinemann M, Hucka M, Le Novère N, Li P, Liebermeister W, Mo ML, Oliveira AP, Petranovic D, Pettifer S, Simeonidis E, Smallbone K, Spasić I, Weichart D, Brent R, Broomhead DS, Westerhoff HV, Kirdar B, Penttilä M, Klipp E, Palsson BØ, Sauer U, *et al*: **A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology.** *Nat Biotechnol* 2008, **26:**1155–1160.
46. Juge N: **Plant protein inhibitors of cell wall degrading enzymes.** *Trends Plant Sci* 2006, **11:**359–367.
47. ten Have A, Mulder W, Visser J, van Kan JA: **The endopolygalacturonase gene Bcpg1 is required for full virulence of Botrytis cinerea.** *Mol Plant Microbe Interact* 1998, **11:**1009–1016.
48. Isshiki A, Akimitsu K, Yamamoto M, Yamamoto H: **Endopolygalacturonase is essential for citrus black rot caused by Alternaria citri but not brown spot caused by Alternaria alternata.** *Mol Plant Microbe Interact* 2001, **14:**749–757.
49. Temple B, Bernier L, Hintz WE: **Characterisation of the polygalacturonase gene of the Dutch elm disease pathogen Ophiostoma novo-ulmi.** *New Zealand Journal of Forestry Science* 2009, **39:**29–37.
50. De Lorenzo G, Ferrari S: **Polygalacturonase-inhibiting proteins in defense against phytopathogenic fungi.** *Curr Opin Plant Biol* 2002, **5:**295–299.
51. Oliveira H, Sousa A, Alves A, Nogueira AJA, Santos C: **Inoculation with Ophiostoma novo-ulmi subsp. americana affects photosynthesis, nutrition and oxidative stress in in vitro Ulmus minor plants.** *Environmental and Experimental Botany* 2012, **77:**146–155.

52. Zerbino DR, Birney E: **Velvet: algorithms for de novo short read assembly using de Bruijn graphs.** *Genome Res* 2008, **18:**821–829.
53. Katoh K, Misawa K, Kuma K, Miyata T: **MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform.** *Nucl Acids Res* 2002, **30:**3059–3066.
54. Hung SS, Wasmuth J, Sanford C, Parkinson J: **DETECT–a density estimation tool for enzyme classification and its application to Plasmodium falciparum.** *Bioinformatics* 2010, **26:**1690–1698.
55. Claudel-Renard C, Chevalet C, Faraut T, Kahn D: **Enzyme-specific profiles for genome annotation: PRIAM.** *Nucleic Acids Res* 2003, **31:**6633–6639.
56. Li L, Stoeckert CJ, Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes.** *Genome Res* 2003, **13:**2178–2189.
57. Barthelmes J, Ebeling C, Chang A, Schomburg I, Schomburg D: **BRENDA, AMENDA and FRENDA: the enzyme information system in 2007.** *Nucleic Acids Res* 2007, **35**(Database issue):D511–514.

doi:10.1186/1471-2164-14-162
Cite this article as: Khoshraftar *et al.*: Sequencing and annotation of the *Ophiostoma ulmi* genome. *BMC Genomics* 2013 **14**:162.