SNP discovery in *Leptographium longiclavatum*, a mountain pine beetle-associated symbiotic fungus, using whole-genome resequencing

Dario I. Ojeda[1,], Braham Dhillon[1], Clement K.M. Tsui[1], and Richard C. Hamelin[1,2,]

[1]Department of Forest and Conservation Sciences, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z4.

[2]Natural Resources Canada, Canadian Forest Service, Laurentian Forestry Centre, 1055 du P.E.P.S., Québec, Canada, G1V 4C7

Keywords: bioinformatics, bark beetle, Ophiostomatoid fungi , Sequenom Iplex Gold Assay, next generation sequencing, blue-stain fungi.

Correspondence: Richard Hamelin, Tel: (+1) 604 827 4411; Fax (1+) 604 822 9102; E-mail: richard.hamelin@ubc.ca, dario.alayon@gmail.com.

Running title: SNP discovery in *Leptographium* with resequencing

Abstract

Single nucleotide polymorphisms (SNPs) are rapidly becoming the standard markers in population genomics studies; however, their use in non-model organisms is limited due to the lack of cost-effective approaches to uncover genome-wide variation and the large number of individuals needed in the screening process to reduce ascertainment bias. To discover SNPs for population genomics studies in the fungal symbionts of the mountain pine beetle (MPB), we developed a roadmap to discover SNPs and to produce a genotyping platform. We undertook a whole-genome sequencing approach of *Leptographium longiclavatum* in combination with available genomics resources of another MPB symbiont, *Grosmannia clavigera*. We sequenced 71 individuals pooled into four groups using the Illumina sequencing technology. We generated between 27 and 30 million reads of 75 bp that resulted in a total of 1, 181 contigs longer than 2 kb and an assembled genome size of 28.9 Mb (N_{50}=48 kb, average depth = 125x). A total of 9, 052 proteins were annotated and between 9, 531 and 17, 266 SNPs were identified in the four pools. A subset of 206 genes (containing 574 SNPs, 11% false positives) was used to develop a genotyping platform for this species. Using this roadmap we developed a genotyping assay with a total of 147 SNPs located in 121 genes using the Illumina® Sequenom iPLEX Gold. Our preliminary genotyping (success rate = 85%) of 304 individuals from 36 populations supports the utility of this approach for population genomics studies in other MPB fungal symbionts and other fungal non-model species.

Introduction

The mountain pine beetle (MPB, *Dendroctonus ponderosae*) is the most destructive pest of pine forests in western North America. In Canada alone, 17 million hectares of lodgepole pine (*Pinus contorta*) forests have been destroyed in the last decade (Carroll *et al.* 2004; Kurtz *et al.* 2008). Although MPB is a native bark beetle with a natural distribution range extending from northern Mexico to central British Columbia (BC) and southwestern Alberta (AB), it has expanded into northern BC and northern AB, and it has become a threat to the boreal forest of Canada during the current epidemics (Carroll *et al.* 2004). Recent data showed that MPB is able to attack jack pine, which is the dominant conifer in the Prairies (Cullingham *et al.* 2011). The devastation to lodgepole pine forests has caused severe economic losses to the timber industry, and changed BC from a carbon fixer to carbon emitter (Carroll *et al.* 2004; Kurtz *et al.* 2008).

MPB carries a number of fungal associates in special structures adapted to transport fungal symbionts known as mycangia and on the exoskeleton of the body. This group of MPB associates comprises a large number of sapstain fungi as well as some tree pathogens. The most important fungal species associated with the MPB are *Grosmannia clavigera, Leptographium longiclavatum* and *Ophiostoma montium* (Ophiostomatales, Ascomycetes). These fungi produce numerous sticky asexual spores that are easily dispersed by the beetles (Lee *et al.* 2006a). *Grosmannia clavigera* is associated with MPB and is believed to play a crucial role in MPB attacks to pines (DiGustini 2011). It is capable of killing mature or young lodgepole pines in the absence of MPB when trees are artificially inoculated at a density similar to that of a mass MPB attack (Yamaoka *et al.* 1995; Lee *et al.* 2006a). *Grosmannia clavigera* grows rapidly in the host tree phloem and through the sapwood where it produces melanin, a pigment that stains the wood and blocks the tree's water transport system (Yamaoka *et al.* 1995; Solheim & Krokene 1998;

Rice *et al.* 2007). In contrast, *O. montium* is considered a weak pathogen even though it is frequently isolated from MPB-attacked trees (Solheim & Krokene 1998).

Leptographium longiclavatum shares similar morphological characteristics and evolutionary history with *G. clavigera* (Lim *et al.* 2004; Lee *et al.* 2005, 2006a). It is also pathogenic to lodgepole pines and causes necrosis around the inoculation points in both the phloem and the sapwood (Lee *et al.* 2006a). Inoculation experiments have shown that it is capable of infecting jack pine and lodgepole x jack pine hybrids in northern AB (Rice *et al.* 2007). Since it is a fungal associate of MPB, it is important to better characterize the population structure of *L. longiclavatum*. This information would be important and useful for comparing and contrasting the population structure, adaptation and evolution of this particular species in relation with other MPB fungal associates. The ecological roles that the different fungal species play in the MPB systems are still poorly understood. A better understanding of the selection pressures shaping their evolutionary history would be useful.

Grosmannia clavigera is one of the MPB fungal associate that has been most well characterized and studied (Hesse-Orce *et al.* 2010; Alamouti *et al.* 2011; DiGuistini *et al.* 2011; Tsui *et al.* 2012). A microsatellite study has identified four genetic clusters in this species (Tsui *et al.* 2012). These four genetic clusters have a high level of gene flow among them and individuals among clusters are admixed in origins. Evidence of random mating and genetic equilibrium was also revealed, suggesting the occurrence of sexual reproduction in most populations (Tsui *et al.* 2012). However, the population structure of *L. longiclavatum* is largely undocumented. This species appears to be more prevalent than *G. clavigera* in northern AB (Roe *et al.* 2011b). The populations of *L. longiclavatum* in northern AB were also highly differentiated from those in the Rocky Mountains based on multi-loci sequence data (Roe *et al.* 2011b).

However, the relationships between these populations and those from BC during the MPB epidemic expansion are not fully understood.

The goal of this study is to characterize the populations of *L. longiclavatum* in Canada and the USA in order to establish its population structure and demographic history. In order to do so, a set of SNP markers was developed in *L. longiclavatum* through the construction of genomic DNA libraries and the use of genome and transcriptome reference data from *G. clavigera* (Lim *et al.* 2004; Alamouti *et al.* 2011). The second objective of this study is to develop a SNP genotyping tool using the iPLEX® Gold assay (Sequenom, Inc., San Diego, CA) for this particular species and later apply this strategy to the other MPB fungal species. First, a road map is provided with the strategy used in a species with no previous genomic resources. Second, the results of this experiment, including the number of genes identified and the distribution of SNPs at the genome level, are reported. Finally, a list of SNPs discovered in a selected set of genes involved in the process of host infection and detoxification of terpenoid compounds that are secreted for host defense is reported.

Materials and methods

Fungal materials and culture collection

The fungal cultures were isolated from MPB and infected trees from 16 locations as described in Lee *et al*. (2006a) and Roe *et al.* (2010). They were maintained on malt extract agar (MEA), and the DNA of 71 isolates of *L. longiclavatum* was extracted according to a previous protocol (Roe, *et al.* 2011a). Since four genetic clusters, i. e. northern British Columbia (NBC), southern British Columbia (SBC), epidemics (NPC) and Rocky Mountains (Rocky), have been identified in *G. clavigera* by microsatellite markers (Tsui *et al.* 2012), we hypothesized a similar

geographic structure and pooled the DNA of the *L. longiclavatum* isolates into four clusters that correspond to the spatial distribution pattern in *G. clavigera* with the goal of maximizing the level of genetic variation among and within these clusters (Table S1).

Genome sequencing and assembly

Between nine and twenty-three isolates from each of the four distinct geographical regions were combined into four pools for library construction and sequencing (Table S1). Seventy-five bp paired ends (PE) reads were generated from these four libraries using four lanes on an Illumina Genome Analyzer (GAIIx). Low quality reads were removed before the assembly. Illumina reads from two pools (NBC and SBC) were used to generate a *de novo* genome assembly using ABySS ver. 1.2.3 (Simpson *et al.* 2009), also using a *de* Bruijn graph-based tool in order to assemble an *L. longiclavatum* reference genome.

Gene prediction, annotation and SNP discovery

De novo gene prediction for the *L. longiclavatum* genome was done using GlimmerHMM ver 3.0.1 (Delcher *et al.* 1999; Majoros *et al.* 2004). GlimmerHMM was trained using a protein data-set from a closely related species, *G. clavigera* (DiGuistini *et al.* 2011). For SNP identification, Illumina reads from the four pools were mapped to the *L. leptographium* reference genome (NBC and SBC) using a combination of BWA (Li & Durbin 2009) and SAMtools (Li *et al.* 2009), using the default parameters (the command 'samtools mpileup -C 50 -Q 20 -q 20'). The average genome coverage for the four pools ranged from 51-58X (NBC 58.47, SBC 57.40, NPC 51.86, ROCKY 52.31). Finally, the program snpEff (Cingolani *et al.* 2012) was used to annotate the SNPs discovered in the four pools. In order to determine if the four pools differ in their

number of SNPs, the heterozygosity levels of the common SNPs were calculated, as well as a t-test to determine whether significant differences existed in the number of SNPs among the four pools.

Gene selection and SNP validation using Sanger sequencing

Genes were selected from *G. clavigera* RNA-seq libraries based on their relative expression values (DiGuistini *et al.* 2011). The gene selection strategy aimed to identify genes from three categories: 1) genes overexpressed during treatments; 2) genes with functions involved during host infection and detoxification of terpenoids; and 3) house-keeping genes. This approach allowed the study of demographic as well as adaptive patterns. Therefore, genes with differential expression under specific growing conditions were particularly considered (DiGuistini *et al.* 2011). In total, a subset of 206 genes were selected using this criterion and were classified into 10 groups (Table 1). Orthologs of these *G. clavigera* genes were identified in *L. longiclavatum* and were later used for SNP discovery and validation.

In order to validate the SNPs predicted in the resequencing pooled data, PCR reactions and Sanger sequencing was performed on a set of 19 isolates of *L. longiclavatum* from a wide geographic distribution within BC and AB (Table 2). DNA was extracted using the same protocol described above. Primers were designed based on the orthologous sequences of *L. longiclavatum* found with the best hit in the reference genome of *G. clavigera* (DiGuistini *et al.* 2011). PCR primers were designed to span the regions that contained predicted SNPs using the Primer Quest^{SM} software (http://www.idtdna.com/Scitools/Applications/Primerquest/. Primers were chosen (using default conditions) to amplify a 400-900 bp product that contained one or

more predicted SNPs. Ten regions were selected that contained 21 predicted SNPs based on our resequencing data set (Table 3).

PCR amplifications of the targeted regions were performed in a 25 µl reaction consisting of 2.5 µl of 10X PCR reaction buffer (Invitrogen, Carlsbad, CA, USA), 0.5 µl of 25mm $MgCl_2$, 1 µl containing 10 mM of each dNTP, 0.2 µl of Platinum® Taq polymerase, 2.5 µl of 10 mM of forward and reverse primers, 2 µl of DNA at a 20 ng/µl concentration, and 15 µl of sterile deionized water. PCR parameters were 94 °C for 3 min, followed by 30 cycles of 94 °C for 30 s, 60 °C for 30 s, 72 °C for 1.5 min, and with a final cycle at 72 °C for 7 min. Amplified products were sequenced on both strands with the BIGDYE version 3.1 ready reaction kit (Applied Biosystems, Carlsbad, CA, USA) on an ABI3730xl Data Analyzer at the CHUL Research Centre (CRCHUL) Sequencing and Genotyping of Université Laval, Quebec, QC, Canada. Chromatograms were trimmed, aligned and edited using Geneious v 6.1 (Drummond *et al.* 2010). The predicted SNPs were compared and validated with the aligned assembly sequence of the 19 isolates using Geneious v 6.1. Polymorphisms were validated by observing the predicted SNPs in the assembly dataset. SNPs were recorded as present when they were present in at least 1% of the population sampled.

Genotyping platform to be used in the Illumina iPLEX Gold assay

In this study a roadmap was developed to find SNPs and to design a SNP genotyping platform in *L. longiclavatum* to be used with the Illumina® Sequenom iPLEX Gold assay (Fig. 1). This platform allows for moderate to high throughput genotyping and multiplexing for up to 40 SNPs in a single panel and makes it possible to process up to 384 samples in parallel (http://www.sequenom.com/iplex). Our study particularly aimed to design four panels (36-plex

each) for a total of 144 SNPs in *L. longiclavatum*. In order to design the SNP genotyping panel, the nucleotide sequence for these *L. longiclavatum* genes plus an additional 200 bp up- and downstream were extracted from the genome assembly and used to design the primers. Our assay was developed at McGill University and Genome Quebec Innovation Centre (Montreal, QC, Canada) using Sequenome iPlex Gold technology (http://gqinnovationcenter.com).

Results

Sequencing, assembly and annotation

A total of 27 to 30 million 75 bp paired end reads (2.02-2.29 Gb) were generated for the four pooled, unindexed *L. longiclavatum* samples on an Illumina GAIIx ABySS (Simpson *et al.* 2009). Initially, a 29.6 Mb *L. longiclavatum* genome was reconstructed, comprising 1,996 contigs longer than 200 bp. Parsing for contig length longer than 2 kb reduced the total contig number to 1,181, with an assembled genome size of 28.9 Mb and an average depth coverage of 125X among the four groups (Fig. S1). The longest contig was 268 kb and the N50 contig length was 49 kb. The *Leptographium longiclavatum* genome with contigs longer than 2 kb was used as reference for both *de novo* gene prediction and SNP discovery.

A total of 9,861 proteins were predicted in the *L. longiclavatum* genome. After parsing for length (> 33 aa) and presence of a start codon, 9,052 resulting proteins were annotated using BLASTP against the NCBI nr database. Only 241 proteins (2.7% of total) did not have a BLAST hit whereas, 97.3% of the *L. longiclavatum* proteins with BLAST results had a top hit to *G. clavigera*.

Reads from the four pooled samples were mapped onto the reference sequence to discover SNPs across the *L. longiclavatum* genome. The number of SNPs per pool ranged

between 9,351 and 17,266, and the number of indels ranged from 1,674 to 2,926. With one change every 1,366 bp, the ROCKY pool showed the highest SNP density, while the lowest was for the NBC pool. The highest number of unique SNPs was also present in the ROCKY pooled samples (Fig. 2). All SNPs common to the four pools were identical at all locations, except for five SNPs in the NBC pool. The lowest proportion of unique SNPs was in the NBC and NPC pools, two populations recently derived. On average, 13% of the putative SNPs were located in exonic regions and about 2% of these SNPs were located in introns. Further details about the distribution of SNPs within each pool can be found in the supplementary information (Table S2).

A subset of 206 *L. longiclavatum* genes was initially selected for SNP discovery based on the functional annotation and expression pattern of their *G. clavigera* orthologs. Of these, 29 *L. longiclavatum* genes (14%) did not show any SNP across the four pools and were excluded from further analysis. The remaining 177 *L. longiclavatum* genes had a total of 574 candidate SNPs. An additional 25 *L. longiclavatum* genes were omitted based on the SNP location in the gene, i.e. SNPs were present exclusively in the intron or in the UTR region. The remaining 152 *L. longiclavatum* genes had 378 candidate exonic SNPs, with a maximum of 12 exonic SNPs per gene. The maximum proportions of exonic SNPs (32%) were present in one pool, whereas only 16% of the exonic SNPs were shared across the four pools.

SNP validation

A total of 21 putative SNP loci were predicted in the 10 gene regions selected for re-sequencing and only three (15%) were false positives (Table S3). All three false positives were located in two gene regions, namely short chain dehydrogenase reductase (GLEAN_1289) and lignostilbene dioxygenase (GLEAN_684) (Table 3 and S3). Three additional SNPs were found

in addition to those predicted within the 10 gene regions selected for validation. These SNPs were located in three gene regions, namely methyltransferase type 12 (Glean_2132), c2h2 transcription factor (Glean_7264), and an ABC transporter (Glean_8030) (Table S3).

Discussion

Single nucleotide polymorphisms (SNPs) are widely used in humans and model species, and are currently becoming a common marker for a variety of evolutionary and ecological studies in non-models organisms, such as white spruce (*Picea glauca*), Chinook salmon (*Oncorhynchus tshawytscha*), and domestic cattle (reviewed in Garvin *et al.* 2010). The drop of sequencing costs together with the development of a variety of bioinformatics tools has allowed their discovery and application in non-model species (Kumar *et al.* 2012), including some fungal species (e.g., Neafsey *et al.* 2010; Broders *et al.* 2011; Abbott *et al.* 2012; Tollenaere *et al.* 2012).

In this study, we describe the strategy used for SNP discovery in a non-model fungal species, *L. longiclavatum*, with the purpose of developing a genotyping platform for population structure, genetic diversity and the identification of candidate adaptive SNPs involved in host defence detoxification and pathogenicity (Fig. 1). *Leptographium longiclavatum* and *G. clavigera* are blue-stain fungal symbiotic species that are believed to play an important role in the mountain pine beetle outbreak in western North America (Lee *et al.* 2006b; c; DiGuistini *et al.* 2011; Tsui *et al.* 2012). A more comprehensive understanding of the population structure and genetic diversity of these fungal species across their range of distribution will provide additional information that can be incorporated into ecological risk modelling of the current outbreak (James *et al.* 2011; de La Giroday *et al.* 2012; Sambaraju *et al.* 2012).

Our approach successfully demonstrated the combination of short reads sequencing for genome assembly in a non-model species, and the feasibility of gene annotation using available genomic resources of a closely related species. *Grosmannia clavigera* and *L. longiclavatum* are two closely related species (Lee *et al.* 2005; Roe *et al.* 2011b; Tsui *et al.* 2012) and we were able to successfully annotate *L. longiclavatum* proteins based on previous genomic resources from the former species (DiGuistini *et al.* 2011). The close proximity between the two species was also reflected by the high percentage of DNA oligonucleotide primer transferability. For instance, we were able to amplify *G. clavigera* control samples with 90% of the primers we designed for *L. longiclavatum* sequences during our SNP validation step.

We also found a high similarity for the proteins sequences between the two species, as 97.3% of the *L. longiclavatum* proteins annotated had a top hit with annotations in *G. clavigera.* Only 241 proteins had no BLAST hit and might be specific genes coding for unknown functions in *L. longiclavatum*. Further studies, such as transcriptome profiling of these fungi under different growing conditions, are required to better understand the functions of these genes as well as to improve the gene annotation of this species.

In total, we identified from 9,351 and 17,266 SNPs in the four pools studied, with an average of one exonic SNP every 1.3 kb (Table S2). We did not observe significant differences in the total number of SNPs and/or levels of heterozygosity of common SNPs among the four pools, suggesting similar amount of SNPs among the four pools. The SNP frequency in our study is higher than that found in previous reports on other fungal species. For example, an average of one SNP every 599.2 kb was reported in *Ophiognomonia clavigignenti-juglandacearum* (Broders *et al.* 2011), and an average density of 0.3 SNPs per kb was found in *Podosphaera plantaginis* (Tollenaere *et al.* 2012); but 0 to17.5 SNPs per kb were reported in *Fusarium*

graminearum, a much higher SNP density than that reported here for *L. longiclavatum* (Cuomo *et al.* 2007). There are several reasons explaining the differences of SNPs density reported among these fungal species. First, the technology used to obtain the genome sequences can influence the number of SNPs discovered. Sanger sequencing tends to have fewer false positives than NGS technologies, especially when genome annotation has been improved with EST or transcriptome data. Second, the total number of SNPs discovered can vary according to the source material sequenced. In our case, and that reported for *F. graminearum* (Cuomo *et al.* 2007), SNP discovery was carried out using genomic DNA, while in the case of *P. plantaginis* (Tollenaere *et al.* 2012), SNPs were mined using transcriptome data. It has previously been reported that the distribution of SNPs can be biased towards specific regions of the genome (Cuomo *et al.* 2007); thus a strategy aiming only to a subset of the entire genome will recover less SNPs. Third, differences in SNP calling criteria also affects the number of SNPs recovered. All previous studies mentioned above used different programs and criteria to call SNPs and customized SNP calling criteria can result in fewer SNPs than default parameters in commercial software. Fourth, species that more readily sexually recombine will have a greater frequency of SNPs than those that are clonal and reproduce asexually. *Leptographium longiclavatum* has not been reported to have a sexual stage in the life cycle (Lee *et al.* 2005). However, the fungus has been confirmed to be heterothallic, bearing one of the opposite mating type genes (Tsui *et al.* 2013). All the populations contain isolates of both mating types, suggesting the fungus can reproduce sexually (Tsui *et al.* 2013). Similarly, analyses of the microsatellite data indicated the presence of clones (repeated genotypes) in each location, but the test on the association of genotypes indicated the presence of recombination in some populations (Tsui pers. comm.).

Finally, the genetic relatedness of the samples included during the discovery step can also influence the number of SNPs identified. In our particular case, we based our sample selection for resequencing and SNP discovery on a previous analysis that characterized the population structure of *G. clavigera* (Tsui *et al.* 2012). Although this population structure does not necessary mean that a closely related species will have a similar population structure, our strategy aimed to incorporate the maximum variation possible with the purpose of identifying SNPs in the selected gene regions. Our approach also included historic samples (endemic levels of the outbreak) as well as samples from different years since the beginning of the current outbreak. This approach might explain the higher density of SNPs in comparison with previous studies (Broders *et al.* 2011; Tollenaere *et al.* 2012).

Our *in silico* validation suggests a low percentage of false positives, with an 89% success rate of predicted exonic SNPs in our approach. Similar rates of SNP validation success have been reported in other studies that used NGS technologies for SNP discovery, e.g. 84% success in *Parus major* (Van Bers *et al.* 2010), 90% in sockeye salmon (Everett *et al.* 2011), and between 74.5 and 94% success in several plant species (Novaes *et al.* 2008; Bundock *et al.* 2009; Trick *et al.* 2009; Buggs *et al.* 2010; Fu & Peterson 2011; Geraldes *et al.* 2011).

Because our SNP selection targeted only exonic regions, we were unable to estimate the percentage of recovery in intronic regions, but a previous study in plants (*Populus trichocarpa*) reported higher rates of false positives in intronic compared with exonic regions (Geraldes *et al.* 2011). Nevertheless, our *in silico* validation suggests a low level of false positives within the gene regions selected in *L. longiclavatum*, and we therefore expect that of the total of 17, 266 SNPs predicted, only 15,540 are *bona fide* exonic SNPs.

Our strategy led to the design of four genotyping panels (36-plex each) suited for the iPLEX® Gold assay available from Sequenom. The iPlex Gold assay is a commonly used platform for medium-throughput genotyping, and together with the MassARRAY® System from Sequenom Inc., it has been widely used for fine mapping, validation of genome-wide association studies (GWAS), and routine genetic testing of SNP panels (Ehrich *et al.* 2005; Gabriel *et al.* 2009; Millis 2011). This genotyping platform has been successfully applied to several organisms, including humans, pigs, microorganisms and plants, with reliable results and high proportions of successful SNP validation (Ramos *et al.* 2009; Buggs *et al.* 2010; Bouakaze *et al.* 2011; Ho *et al.* 2011; Shi *et al.* 2011; Syrmis *et al.* 2011).

Previous platforms in other fungal species have used low-throughput genotyping scale and multiplexing. These studies have used SNP-based genotyping platforms that only allow the multiplex of no more than 24 SNPs (Fournier *et al.* 2010; Thomas *et al.* 2012). In contrast to these applications, the sequenom iPLEX platform allows a higher number of SNP multiplexing and sample scaling. This genotyping platform has previously been applied to yeast (Ben-Ari *et al.* 2005) and another fungal species with high call rates (above 98%) (Tollenaere *et al.* 2012). For the particular species in our study, our final assay consisted of four panels with a total of 147 SNPs distributed across 121 gene regions. SNP location and flanking regions for probe design for each of the SNP used in this genotyping platform can be found in the supplementary information (Table S4).

Our panels include a variety of genes involved in host detoxification, stress related-genes, as well as constitutive (nearly neutral) genes (Table 1). This platform will be used to further our understanding of the population structure, genetic diversity, population dynamics and adaptive significance of these genes in a larger dataset comprising almost the entire range of distribution

of this species. Our preliminary genotyping assay with a sample size of 304 samples and 36 populations yielded an overall 85% of calling rate, an 85.9% sample success and 80.2% genotype success rates (Ojeda *et. al*., unpublished data). Additional information regarding SNP heterozygosity can be found in the supplementary information for a subset of four populations (Table S5). These results suggests a high recovery level of suitable SNPs (low levels of false positives) and low ascertain bias (Garvin *et al.* 2010) in a larger data set for this species.

The strategy reported here can be further applied to SNP discovery and to the design of genotyping assays for other fungal species involved in the mountain pine beetle outbreak, as well as other species that lack available genomic resources.

Acknowledgments

Funding for this research was provided by Genome Canada, Genome BC, Genome Alberta and the Government of Alberta (AAET/ AFRI-859-G07) in support of the Tria I and Tria II Projects (http://www.thetriaproject.ca). The authors thank B. Lai and S. Beauseigle (UBC) for technical assistance. We are grateful to Colette Breuil for continued support and assistance. Fungal isolates used in the study were graciously provided by A. Roe, A. Rice, F. Sperling, C. Breuil and S. Lee.

References

Abbott CL, Gilmore SR, Lewis CT *et al.* (2012) Development of a SNP genetic marker system based on variation in microsatellite flanking regions of Phytophthora infestans. *Canadian Journal of Plant Pathology*, **32**, 440–457.

Alamouti SM, Wang V, Diguistini S *et al.* (2011) Gene genealogies reveal cryptic species and host preferences for the pine fungal pathogen Grosmannia clavigera . *Molecular Ecology*, **20**, 2581–2602.

Ben-Ari G, Zenvirth D, Sherman A *et al.* (2005) Application of SNPs for assessing biodiversity and phylogeny among yeast strains. *Heredity*, **95**, 493–501.

Van Bers NEM, Van Oers K, Kerstens HHD *et al.* (2010) Genome-wide SNP detection in the great tit Parus major using high throughput sequencing. *Molecular Ecology*, **19 Suppl 1**, 89–99.

Bouakaze C, Keyser C, Gonzalez A *et al.* (2011) MALDI-TOF MS-based SNP genotyping assay using the iPLEX Gold technology for identification of Mycobacterium tuberculosis complex species and lineages. *Journal of Clinical Microbiology*, JCM.

Broders KD, Woeste KE, San Miguel PJ, Westerman RP, Boland GJ (2011) Discovery of single-nucleotide polymorphisms (SNPs) in the uncharacterized genome of the ascomycete Ophiognomonia clavigignenti-juglandacearum from 454 sequence data. *Molecular Ecology Resources*, **11**, 693–702.

Buggs R., Chamala S, Wu W *et al.* (2010) Characterization of duplicate gene evolution in the recent natural allopolyploid Tragopogon miscellus by next-generation sequencing and Sequenom iPLEX MassARRAY genotyping. *Molecular Ecology*, **19 Suppl 1**, 132–146.

Bundock PC, Eliott FG, Ablett G *et al.* (2009) Targeted single nucleotide polymorphism (SNP) discovery in a highly polyploid plant species using 454 sequencing. *Plant Biotechnology Journal*, **7**, 347–354.

Carroll AL, Taylor SW, Régnière J, Safranyik L (2004) Effects of climate change on range expansion by the mountain pine beetle in British Columbia. In: *Mountain Pine Beetle Symposium: Challenges and Solutions* (eds Shore TL, Brooks JE, Stone JE), pp. 223–232. Natural Resources Canada, Canadian Forest Service, Pacific Forestry Centre, Information Report BC-X-399.

Cingolani P, Platts A, Wang L *et al.* (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*, **6**, 80–92.

Cullingham CI, Cooke JEK, Dang S *et al.* (2011) Mountain pine beetle host-range expansion threatens the boreal forest. *Molecular Ecology*, **20**, 2157–2171.

Cuomo CA, Güldener U, Xu J-R *et al.* (2007) The Fusarium graminearum genome reveals a link between localized polymorphism and pathogen specialization. *Science*, **317**, 1400–1402.

Delcher AL, Harmon D, Kasif S, White O, Salzberg SL (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Research*, **27**, 4636–4641.

DiGuistini S, Wang Y, Liao NY *et al.* (2011) Genome and transcriptome analyses of the mountain pine beetle-fungal symbiont Grosmannia clavigera, a lodgepole pine pathogen.

Proceedings of the National Academy of Sciences of the United States of America, **108**, 2504–2509.

Drummond AJ, Ashton B, Buxton S *et al.* (2010) Geneious pro v4.7. *Available from www.geneious.com*.

Ehrich M, Böcker S, Van Den Boom D (2005) Multiplexed discovery of sequence polymorphisms using base-specific cleavage and MALDI-TOF MS. *Nucleic Acids Research*, **33**, e38.

Everett MV, Grau ED, Seeb JE (2011) Short reads and nonmodel species: exploring the complexities of next-generation sequence assembly and SNP discovery in the absence of a reference genome. *Molecular Ecology Resources*, **11**, 93–108.

Fournier A, Widmer F, Enkerli J (2010) Development of a single-nucleotide polymorphism (SNP) assay for genotyping of Pandora neoaphidis. *Fungal Biology*, **114**, 498–506.

Fu Y, Peterson GW (2011) Developing genomic resources in two Linum species via 454 pyrosequencing and genomic reduction. *Molecular Ecology Resources*, **12**, 492–500.

Gabriel S, Ziaugra L, Tabbaa D (2009) SNP genotyping using the Sequenom MassARRAY iPLEX platform. In: *Current Protocols in Human Genetics* (eds Haines JL, Korf BR, Morton CC, *et al.*) John Wiley & Sons, Inc.

Garvin MR, Saitoh K, Gharrett AJ (2010) Application of single nucleotide polymorphisms to non-model species: a technical review. *Molecular Ecology Resources*, **10**, 915–934.

Geraldes A, Pang J, Thiessen N *et al.* (2011) SNP discovery in black cottonwood (Populus trichocarpa) by population transcriptome resequencing. *Molecular Ecology Resources*, **11 Suppl 1**, 81–92.

Hesse-Orce U, DiGuistini S, Keeling CI *et al.* (2010) Gene discovery for the bark beetle-vectored fungal tree pathogen Grosmannia clavigera. *BMC Genomics*, **11**, 536.

Ho DW, Yiu WC, Yap MK *et al.* (2011) Genotyping performance assessment of whole genome amplified DNA with respect to multiplexing level of assay and its period of storage. *PLoS ONE*, **6**, e26119.

James PMA, Coltman DW, Murray BW, Hamelin RC, Sperling FAH (2011) Spatial genetic structure of a symbiotic beetle-fungal system: toward multi-taxa integrated landscape genetics (S Joly, Ed,). *PLoS ONE*, **6**, e25359.

Kumar S, Banks TW, Cloutier S (2012) SNP discovery through next-generation sequencing and its applications. *International Journal of Plant Genomics*, **2012**, 1–15.

Kurtz W, Dymond C, Stinson G *et al.* (2008) Mountain pine beetle and forest carbon feedback to climate change. *Nature*, **452**, 987–990.

De La Giroday H-MC, Carroll AL, Aukema BH (2012) Breach of the northern Rocky Mountain geoclimatic barrier: initiation of range expansion by the mountain pine beetle. *Journal of Biogeography*, **39**, 1112–1123.

Lee S, Kim J, Breuil C (2005) Leptographium longiclavatum sp. nov., a new species associated with mountain pine beetle, Dendroctonus ponderosae. *Mycological Research*, **109**, 1162–1170.

Lee S, Kim J, Breuil C (2006a) Pathogenicity of Leptographium longiclavatum associated with Dendroctonus ponderosae to Pinus contorta. *Canadian Journal of Forest Research*, **36**, 2864–2872.

Lee S, Kim J, Breuil C (2006b) Diversity of fungi associated with the mountain pine beetle, Dendroctonus ponderosae and infested lodgepole pines in British Columbia. *Fungal Diversity*, **22**, 91–105.

Lee S, Kim JJ, Breuil C (2006c) Diversity of fungi associated with the mountain pine beetle, Dendroctonus ponderosae and infested lodgepole pines in British Columbia. *Fungal Diversity*, **22**, 91–105.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*, **25**, 1754–1760.

Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Lim YW, Alamouti SM, Kim JJ, Lee S, Breuil C (2004) Multigene phylogenies of Ophiostoma clavigerum and closely related species from bark beetle-attacked Pinus in North America. *FEMS Microbiology Letters*, **237**, 89–96.

Majoros WH, Pertea M, Salzberg SL (2004) TigrScan and GlimmerHMM: two open-source ab initio eukaryotic gene-finders. *Bioinformatics*, **20**, 2878–2879.

Millis MP (2011) Medium-throughput SNP genotyping using mass spectrometry: multiplex SNP genotyping using the iPLEX® Gold assay. (JK DiStefano, Ed,). *Methods In Molecular Biology Clifton Nj*, **700**, 61–76.

Neafsey DE, Barker BM, Sharpton TJ *et al.* (2010) Population genomic sequencing of Coccidioides fungi reveals recent hybridization and transposon control. *Genome Research*, **20**, 938–946.

Novaes E, Drost DR, Farmerie WG *et al.* (2008) High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome. *BMC Genomics*, **9**, 312.

Ramos AM, Crooijmans RPMA, Affara NA *et al.* (2009) Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology (L Orban, Ed,). *PLoS ONE*, **4**, 13.

Rice AV, Thormann MN, Langor DW (2007) Virulence of, and interactions among, mountain pine beetle associated blue-stain fungi on two pine species and their hybrids in Alberta. *Canadian Journal of Botany*, **85**, 316–323.

Roe AD, James P, Rice AV, Cooke JEK, Sperling FAH (2011) Spatial community structure of mountain pine beetle fungal symbionts across a latitudinal gradient. *Microbial Ecology*, **62**, 347–360.

Roe AD, Rice AV, Bromilow SE, Cooke JEK, Sperling FAH (2010) Multilocus species identification and fungal DNA barcoding: insights from blue stain fungal symbionts of the mountain pine beetle. *Molecular Ecology Resources*, **10**, 946–959.

Roe AD, Rice AV, Coltman DW, Cooke JEK, Sperling FAH (2011) Comparative phylogeography, genetic differentiation and contrasting reproductive models in three fungal symbionts of a multipartite bark beetle symbiosis. *Molecular Ecology*, **20**, 584–600.

Sambaraju KR, Carroll AL, Zhu J *et al.* (2012) Climate change could alter the distribution of mountain pine beetle outbreaks in western Canada. *Ecography*, **35**, 211–223.

Shi A, Chen P, Vierling R *et al.* (2011) Multiplex single nucleotide polymorphism (SNP) assay for detection of soybean mosaic virus resistance genes in soybean. *Theoretical and Applied Genetics*, **122**, 445–457.

Simpson JT, Wong K, Jackman SD *et al.* (2009) ABySS: A parallel assembler for short read sequence data. *Genome Research*, **19**, 1117–1123.

Solheim H, Krokene P (1998) Growth and virulence of mountain pine beetle associated blue-stain fungi, Ophiostoma clavigerum and Ophiostoma montium. *Canadian Journal of Botany*, **76**, 561–566.

Syrmis MW, Moser RJ, Whiley DM *et al.* (2011) Comparison of a multiplexed MassARRAY system with real-time allele-specific PCR technology for genotyping of methicillin-resistant Staphylococcus aureus. *Clinical Microbiology and Infection*, **17**, 1804–10.

Thomas E, Pakala S, Fedorova ND, Nierman WC, Cubeta MA (2012) Triallelic SNP-mediated genotyping of regenerated protoplasts of the heterokaryotic fungus Rhizoctonia solani. *Journal of Biotechnology*, **158**, 144–150.

Tollenaere C, Susi H, Nokso-Koivisto J *et al.* (2012) SNP design from 454 sequencing of Podosphaera plantaginis transcriptome reveals a genetically diverse pathogen metapopulation with high levels of mixed- genotype infection. *PLoS ONE*, **7**, e52492.

Trick M, Long Y, Meng J, Bancroft I (2009) Single nucleotide polymorphism (SNP) discovery in the polyploid Brassica napus using Solexa transcriptome sequencing. *Plant Biotechnology Journal*, **7**, 334–346.

Tsui CK-M, Diguistini S, Wang Y *et al.* (2013) Unequal recombination and evolution of the mating-Type (MAT) loci in the pathogenic fungus Grosmannia clavigera and relatives. *G3*, **3**, 465–80.

Tsui CKM, Roe AD, El-kassaby YA *et al.* (2012) Population structure and migration pattern of a conifer pathogen, Grosmannia clavigera, as influenced by its symbiont, the mountain pine beetle. *Molecular Ecology*, **21**, 71–86.

Yamaoka Y, Hiratsuka Y, Maruyama PJ (1995) The ability of Ophiostoma clavigerum to kill mature lodgepole-pine trees. *European Journal of Forest Pathology*, **25**, 401–404.

Data accessibility

The contigs resulting from the SNP validation were deposited in GenBank under accession nos. XXXXXXXX-XXXXXXXX.

Author Contributions

D.I.O., B.D., C.K.M.T. and R.C.H. conceived the study. B.D. performed the genome assembly and SNP discovery, D.I.O ran the SNP validation and the analyses. D.I.O. and C.K.M.T. wrote the manuscript. All authors have read and approved the manuscript.

Figure Legends

Fig. 1 Strategy used for the design and development of the SNP genotyping platform in *Leptographium longiclavatum* using whole genome resequencing.

Fig. 2 Distribution of the SNPs discovered at the genome level in each of the four pools of *Leptographium longiclavatum*. NPC = Epidemic, NBC = northern British Columbia, SBC = southern British Columbia, and ROCKY = Rocky Mountains.

Tables

Table 1 Gene categories of the 206 genes selected for SNP validation and further design of the genotyping panel in *Leptographium longiclavatum*. The numbers for each category represent the final number of SNPs included in the final Illumina® Sequenom iPLEX Gold assay genotyping platform. LPPE = lodgepole pine phloem extract, CFEM = cystein-rich fungal extracellular membrane.

Table 2 Samples used in the validation of the SNPs predicted in the whole genome resequencing of *Leptographium longiclavatum*.

Table 3 Distribution of the predicted SNPs in the 10 gene regions used during the validation step using Sanger sequencing of *Leptographium longiclavatum*. Primer sequences, the size of each fragment amplified and GenBank accession numbers are provided for each gene region.

Supplementary Information

Table S1 Samples pooled in the whole genome resequencing of *Leptographium longiclavatum*.

Table S2 Summary of the statistics during the SNP discovery in the four pools of *Leptographium longiclavatum*. NPC = Epidemic, NBC = northern British Columbia, SBC = southern British Columbia, and ROCKY = Rocky Mountains.

Table S3 Predicted and validated SNPs for each of the 10 gene regions selected for the validation process. Validated SNPs are indicated in grey, while novel SNPs discovered during the validation process are indicated as unshaded. NA= not applicable, SNP not observed during the validation.

Table S4 SNP position and flanking regions used in the design of the sequenom iPLEX® Gold assay at McGill University and Genome Quebec Innovation Centre (Montreal, QC, Canada). Gleans in bold were used in the SNP validation.

Table S5 Levels of heterozygosity in four populations of *Leptographium longiclavatum* using the four panels of SNPs obtained with the Sequenom Iplex assay.