Comparative Genomics of the Pine Pathogens and Beetle Symbionts in the Genus Grosmannia

Sepideh Massoumi Alamouti,1 Sajeet Haridas,1,2 Nicolas Feau,3 Gordon Robertson,4 Jörg Bohlmann,3,5 and Colette Breuil*,1

1Department of Wood Science, University of British Columbia, Vancouver, British Columbia, Canada
2DOE Joint Genome Institute, Walnut Creek, California
3Department of Forest and Conservation Sciences, University of British Columbia, Vancouver, British Columbia, Canada
4British Columbia Cancer Agency Genome Sciences Centre, Vancouver, British Columbia, Canada
5Michael Smith Laboratories, University of British Columbia, Vancouver, British Columbia, Canada

*Corresponding author: E-mail: Colette.Breuil@ubc.ca.

Associate editor: John Logsdon

Abstract

Studies on beetle/tree fungal symbionts typically characterize the ecological and geographic distributions of the fungal populations. There is limited understanding of the genome-wide evolutionary processes that act within and between species as such fungi adapt to different environments, leading to physiological differences and reproductive isolation. Here, we assess genomic evidence for such evolutionary processes by extending our recent work on Grosmannia clavigera, which is vectored by the mountain pine beetle and Jeffrey pine beetle. We report the genome sequences of an additional 11 G. clavigera (Gc) sensu lato strains from the two known sibling species, Grosmannia sp. (Gs) and Gc. The 12 fungal genomes are structurally similar, showing large-scale synteny within and between species. We identified 103,430 single-nucleotide variations that separated the Grosmannia strains into divergent Gs and Gc clades, and further divided each of these clades into two subclades, one of which may represent an additional species. Comparing variable genes between these lineages, we identified truncated genes and potential pseudogenes, as well as seven genes that show evidence of positive selection. As these variable genes are involved in secondary metabolism and in detoxifying or utilizing host-tree defense chemicals (e.g., polyketide synthases, oxidoreductases, and mono-oxygenases), their variants may reflect adaptation to the specific chemistries of the host trees Pinus contorta, P. ponderosa, and P. jeffreyi. This work provides a comprehensive resource for developing informative markers for landscape population genomics of these ecologically and economically important fungi, and an approach that could be extended to other beetle–tree-associated fungi.

Key words: fungi, beetle, genomics, pathogen, pine, symbiont.

Introduction

Over tens of millions of years, conifer forests around the world have provided unique ecological niches for native bark beetles and their fungal symbionts. Interactions between conifer hosts, bark beetle vectors, and their fungal associates have influenced the evolution of tree chemical defenses (e.g., terpenoids), beetles, and fungal symbionts (Seybold et al. 2000; Farrell et al. 2001; Jordal 2013). Although beetle–tree-associated fungi have significant effects on forest ecosystems, knowledge has improved only recently about the specificity for host trees or beetle vectors in this group of fungi (Wingfield et al. 1993; Kurz et al. 2008). Currently, little is known about the genetic differences that are associated with speciation and adaptation in this group of fungi. Fungal diversification and specialization for hosts may depend on genetic differences that include genomic rearrangements, gene losses/duplications, and coding and noncoding sequence variants that may be under selective pressure in particular genes (Aguileta et al. 2009; Stukenbrock et al. 2010; Manning et al. 2013). The extent of adaptive processes at the genome level can be quantified by identifying genomic differences within and between fungal lineages that have recently diverged and specialized onto different host trees (Stukenbrock et al. 2010).

In North America, tree-inhabiting beetles and their fungal symbionts are among the most diverse and damaging forest pests (Harrington 2005; Jordal and Cognato 2012). For example, in western Canada alone, the mountain pine beetle (MPB; Dendroctonus ponderosae) and its fungal associates have killed over 18 million hectares of Pinus contorta forests (http://www.nrcan.gc.ca/forests/canada/sustainable-forest-management/criteria-indicators/13241, last accessed March 19, 2014), dramatically altering forest ecosystem dynamics and forest-dependent economic activities (Kurz et al. 2008). Further, the recent spread of the MPB–fungal complex into Alberta and Saskatchewan and into P. banksiana raises the risk that the epidemic will spread eastward into and potentially across Canada’s boreal forests (Cullingham et al. 2011). Of the fungal associates, the ophiostomatoid (Sordariomycetes, Ascomycota) Grosmannia clavigera sensu lato is crucial to the epidemic as an obligate symbiont of MPB and a pathogen of P. contorta that can kill living trees through...
Given that host defense chemicals vary among pine species can tolerate, detoxify, and utilize host defense chemicals. 2010; Wang et al. 2013). This work suggested that aspects of the functional genomics of the fungus, including 30-Mb genome assembly consisted of 18 supercontigs and genome-wide single-nucleotide polymorphism (SNP) variations relative to the Gs reference sequence. We applied comprehensive assessment of intra and interspecies genomic populations of Gc and Gs. We sequenced 11 strains, assem-

citive analysis of evolutionary divergence in distinct lineages and whether it provides further evidence of ecological and/or geographic divergence in these fungi. Focusing on SNPs that are predicted to alter proteins, we assess evidence for fungal adaptation to different species of pine (P. contorta, P. jeffreyi, and P. ponderosa). We identify genes that show evidence of adaptive selection and relate these variations to differences in fungal ecology and biology. 

Results

Genome Assembly, Orthologs Determination, and Single-Nucleotide Variants

For the 11 Grosmannia strains, we obtained genome sequence assemblies ranging from 27.7 to 32.4 Mb (table 1, supplementary tables S1 and S2, Supplementary Material online). We found no significant evidence of genome rearrangements for any of the sequenced strains compared with the slkw1407 reference genome (supplementary fig. S1 and table S2, Supplementary Material online). The 11 strains shared more than 8,000 genes with an average sequence identity of 98 ± 0.4% between Gs and Gc genomes. On average, only 3% of genes were missing or highly divergent (~70% sequence identity) relative to the reference gene models (supplementary table S3, Supplementary Material online, and fig. 1). Sequence assemblies are available in National Center for Biotechnology Information (NCBI) under Genomes BioProject PRJNA: 239888.

Assessing coverage for variant calling, we noted that between 86% and 99% of the filtered reads mapped to the slkw1407 genome sequence, providing an average read depth between 22× and 58× per strain (table 2 and supplementary table S4, Supplementary Material online). On average, 94.1% of the slkw1407 genome (i.e., ~27.4 Mb) was covered by ≥5 mapped reads, with a range of 90.0–97.8% coverage across the 11 genomes.

We compared the variants called by SAMtools and Genome Analysis Toolkit (GATK), which showed a high percentage of overlapping single-nucleotide variations (SNVs) (n = 91,763) between the two methods, and used the SAMTools results because it generated fewer unique calls (12.7% of total 105,104) than GATK (21.9% of total 117,449). Of 198,362 putative variants, 105,104 SNVs and 9,907 indels passed quality control and filtering, yielding 115,011 high-confidence differences across the 12 Grosmannia genomes. After removing ambiguous calls that are likely to represent errors in the reference genome assembly, we obtained 103,430 SNV sites with a mean transition-to-transversion ratio of 3.4 (supplementary table S5, Supplementary Material online). We estimated a false-negative rate of 4.4 × 10^-6 or one per 24,590 nts and a false-positive rate of 0.046% for the sequenced regions.

Functional Classification of Genomic Variants

We classified nucleotide variants for their potential functional and/or adaptive significance by characterizing the level of beetles, trees, and fungi provide unique systems for understanding ecological divergence or speciation (Thompson 1994; DiGuistini et al. 2011; Massoumi Alamouti et al. 2011). Theoretical studies suggest that dispersal of the plant pathogen between hosts, and aspects of the pathogen life cycle can promote ecological divergence; for example, reproduction is frequently asexual, and sexual recombination is constrained because it occurs within a host’s tissues (Giraud et al. 2006, 2008). Concordant with this theoretical framework, protein-coding gene families have identified two cryptic species within G. clavigera (Gc, Massoumi Alamouti et al. 2011). One species (Grosmannia sp. [Gs]) is an exclusive associate of MPB and its primary host tree P. contorta, whereas the other (Gc) is found on localized populations of MPB andJPB where these beetles colonize the closely related P. jeffreyi and P. ponderosa. Although the two Grosmannia lineages can occur in the same geographic region (e.g., California), no evidence of gene flow between Gs and Gc was detected based on sequence analysis of 15 nuclear coding loci, suggesting that host tree species and beetle population dynamics are important factors in the evolution and divergence of these fungi (Massoumi Alamouti et al. 2011).

Recently, we reported the genome sequence of a Gs strain (slkw1407) isolated from P. contorta trees in the epidemic region of Canada (DiGuistini et al. 2011). Approximately 30-Mb genome assembly consisted of 18 supercontigs and 8,312 protein-coding gene models. We characterized some aspects of the functional genomics of the fungus, including its interaction with host-defense chemicals (Hesse-Orce et al. 2010; Wang et al. 2013). This work suggested that Grosmannia can tolerate, detoxify, and utilize host defense chemicals. Given that host defense chemicals vary among pine species (Keeling and Bohlmann 2006; Gerson et al. 2009; Boone et al. 2011; Hall, Yuen, et al. 2013; Hall, Zerbe, et al. 2013), here we hypothesize that genes involved in host-pathogen interactions, secondary metabolite production, and fungal interactions and differentiation, such as cytochrome P450s, mono-oxygenases, membrane proteins such as ATP-binding cassette (ABC) and major facilitator superfamily transporters, polyketide synthases (PKS) genes, and vegetative incompatibility genes, may have diverged to a greater extent than other genes in response to selection in different host environments.

In this work, we use the reference Gs genome to enable comparative analysis of evolutionary divergence in distinct populations of Gc and Gs. We sequenced 11 strains, assembled their draft genome sequences, and reported a comprehensive assessment of intra and interspecies genomic variations relative to the Gs reference sequence. We applied genome-wide single-nucleotide polymorphism (SNP) phylogenies of 12 Grosmannia strains and gene genealogies of additional strains to test whether the genome data set confirm our recent genealogical study that Gs and Gc are distinct lineages and whether it provides further evidence of ecological and/or geographic divergence in these fungi.
intra- and interspecific differences in different genomic regions. From 103,430 SNVs across the 12 Grosmannia genomes, we identified 36,017 variants within the slkw1407 gene models. Of these genic variants, 5,826 were intronic and 30,191 were in coding exons, 14,889 of which were synonymous and 15,302 nonsynonymous (supplementary table S5, Supplementary Material online). Of the nongenic variants, 24,589 were located in our predicted approximately 6,000 kb gene-flanking regions and 42,880 were intergenic. Because gene models in slkw1407 can overlap (DiGuistini et al. 2009), 56 of the genic SNVs were identified in more than one gene region (e.g., a variant in a coding region and in an intronic region).

Among the coding variants, 262 variants in 218 genes were predicted either to cause a premature stop codon \( (n = 226) \) or to eliminate a stop codon \( (n = 36) \). Of these 262 variants, 92 that had truncated proteins and 3 that had lost a stop codon occurred in only one genome, 155 were found in at least two genomes, and 12 were observed in all 11 genomes. The latter 12 variants may indicate that the slkw1407 genome sequence has an error or a low-frequency allele in these positions. For this analysis, we removed the 12 variants that occurred in all 11 genomes, as well as the 95 SNVs that were found in only one genome, which removed 85 genes. Of the remaining 133 genes, 85 were slkw1407 gene models with known functions \( (n = 86 \) for 71 genes with premature stop SNV; \( n = 14 \) for 13...
genes with stop-loss SNVs; and one gene showed both a stop-gain and a stop-loss SNV, supplementary table S6, Supplementary Material online). Blast2Go enrichment analysis of genes with known functions identified enrichment of stop codon variants for members with oxidoreductase activity (31%, \( P < 0.001 \)) within both biological process (BP) and molecular function (MF) classifications, followed by genes involved in transmembrane transporters (16.9%) and nucleotide-binding activities (18%) in BP and MF, respectively (supplementary table S7, Supplementary Material online). Some of the enriched oxidoreductases belonged to gene families with known roles in detoxification (supplementary tables S6

**Fig. 1.** *Grosmannia* SNP phylogenomics, gene content, and amino acid similarity. (a) MP analysis of 103,430 SNPs among 12 *Grosmannia* genomes. The analysis was best described by a single unrooted tree with consistency index of 0.79 and 0.97 when including only Gs (GsRef, GsB3, and GsC1) and Gc (GcC2 and GcB1) from distinct populations. All branches have 100% BS support and posterior probabilities of 1.0. The scale bar indicates the number of SNPs along each branch. A, B, C, and M are the collection sites. GsB1 is the reference genome. Pc, Pj, and Pp are the host tree species (table 1). The gray box highlights Gs strains from epidemic regions. (b) Genome-wide pairwise amino acid identity between 8,312 *Grosmannia* reference gene models and homologous proteins in the 11 other strains.
and S7, Supplementary Material online). For example, a flavoprotein mono-oxygenase (CMQ-6740) in the slkw1407

protein mono-oxygenase (CMQ-6740) in the slkw1407
gene cluster (fig. 2a,b), which was proposed to have a role
detoxification and/or utilization of host-tree defense chem-
icals (DiGuistini et al. 2011), showed a stop codon in both Gc
strain from P. jeffreyi. For this gene, we confirmed the variant
by slkw1407 EST and RNA-seq data (supplementary table S6,
Supplementary Material online), as well as by an independent
polymerase chain reaction (PCR) validation of additional
strains (total Gc n = 16, Gs = 12 and two other species
n = 4), showing that this mutation is unique to the Gc strains
from P. jeffreyi (fig. 2c).

Divergence Classification of Genomic Variants

Across the 12 Grosmannia genomes, approximately 67%
(n = 70,018) of the total number of SNVs were parsimony
informative in that multiple strains contained alternate nu-
cleotide bases. The remaining SNVs (n = 33,412) were unique
differences (i.e., singletons) in that only one strain showed the
alternate nucleotide base. To characterize intra and interspe-
cific variants, we assigned the informative polymorphisms
(SNP) to three classes: fixed, exclusive, and shared (table 3).
Most SNPs were either fixed (n = 37,712) or were exclusive
to the nine Gs (18,871) or the three Gc (n = 9,685) strains; the
rest (n = 3,750) were the shared polymorphisms present in both
species. Within Gs, the eight resequenced genomes dif-
terences as the reference strain and the epidemic strains
from the two strains from the localized California population.

The AWClust resolved Grosmannia genomes into four clusters corresponding to Gs and Gc lineages that
each formed additional subclusters according to the geo-
graphic regions and host tree associates of the fungal taxa
(supplementary fig. S2, Supplementary Material online).

Clustered and Phylogenomic Analysis of SNVs

We assessed genetic distance and phylogenetic relationships
among Grosmannia genomes by the AWClust nonparametric
clustering (Gao and Starmer 2008) and phylogenetic analyses
of SNV data. The AWClust resolved Grosmannia genomes
into four clusters corresponding to Gs and Gc lineages that
each formed additional subclusters according to the geo-
graphic regions and host tree associates of the fungal taxa
(supplementary fig. S2, Supplementary Material online). The
maximum parsimony (MP) and Bayesian phylogenetic trees
supported the results from cluster analysis and showed identi-
tical tree topologies that only differed in the placement of the
slkw1407 reference strain either within the Gs isolates from
Alberta or those from Rocky Mountains (fig. 1, MP tree). The
MP tree provided high statistical support (bootstrap
[BS] = 100% and PP = 1.0) for the positioning of Gs and Gc
strains into two divergent clades and for additional subclades
within each clade. As expected, slkw1407 grouped within the
P. contorta-infesting Gs strains, which formed a distinct clade
from the Gc strains. In the Gc clade, the Gc holotype that had
been isolated from MPB-infested P. ponderosa was in a differ-
ent subclade than the two P. jeffreyi associates. Within the Gs
strains from MPB-epidemic regions in British Columbia,
Alberta, and Rocky Mountains were significantly separated
from the two strains from the localized California population.
This pattern was also consistent with the SNP density for the
latter two genomes, which showed almost twice as many
differences as the reference strain and the epidemic strains
(supplementary table S5, Supplementary Material online).

Table 2. Summary of the Genomic and Gene Coverage Data in the 11 Sequenced Genomes.

<table>
<thead>
<tr>
<th>IDs*</th>
<th>Covered Genomic Bases (%)</th>
<th>Genomic Coverage</th>
<th>Covered Gene Bases (%)</th>
<th>Gene Coverage</th>
<th>Unmapped Reads (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gb1</td>
<td>28,791,583 (98.8)</td>
<td>70×</td>
<td>15,507,591 (99.9)</td>
<td>47×</td>
<td>0.7</td>
</tr>
<tr>
<td>Gb2</td>
<td>27,370,438 (94.0)</td>
<td>27×</td>
<td>15,333,848 (98.8)</td>
<td>30×</td>
<td>1.2</td>
</tr>
<tr>
<td>Gb3</td>
<td>28,464,740 (97.7)</td>
<td>48×</td>
<td>15,316,500 (98.7)</td>
<td>48×</td>
<td>2.6</td>
</tr>
<tr>
<td>Ga1</td>
<td>28,488,061 (97.8)</td>
<td>58×</td>
<td>15,334,490 (98.8)</td>
<td>61×</td>
<td>3.2</td>
</tr>
<tr>
<td>Ga2</td>
<td>26,491,256 (91.0)</td>
<td>35×</td>
<td>15,353,085 (87.2)</td>
<td>78×</td>
<td>5.0</td>
</tr>
<tr>
<td>Ga3</td>
<td>27,146,569 (93.2)</td>
<td>24×</td>
<td>15,203,277 (98.0)</td>
<td>29×</td>
<td>0.8</td>
</tr>
<tr>
<td>Gm1</td>
<td>27,197,413 (93.4)</td>
<td>22×</td>
<td>15,170,992 (97.8)</td>
<td>33×</td>
<td>1.3</td>
</tr>
<tr>
<td>Gc1</td>
<td>28,153,881 (96.6)</td>
<td>50×</td>
<td>15,092,814 (97.3)</td>
<td>57×</td>
<td>4.1</td>
</tr>
<tr>
<td>Gc2</td>
<td>28,084,478 (96.4)</td>
<td>49×</td>
<td>15,031,784 (96.9)</td>
<td>63×</td>
<td>1.0</td>
</tr>
<tr>
<td>Gb1.a</td>
<td>26,267,089 (90.1)</td>
<td>14×</td>
<td>13,636,725 (87.9)</td>
<td>20×</td>
<td>6.8</td>
</tr>
<tr>
<td>Gb1.b</td>
<td>28,360,091 (97.4)</td>
<td>55×</td>
<td>15,399,793 (99.3)</td>
<td>50×</td>
<td>10.1</td>
</tr>
<tr>
<td>Gb1.ab</td>
<td>28,455,969 (97.7)</td>
<td>79×</td>
<td>15,435,487 (99.5)</td>
<td>77×</td>
<td>9.2</td>
</tr>
<tr>
<td>Gc1</td>
<td>27,001,022 (92.7)</td>
<td>25×</td>
<td>15,204,933 (98.0)</td>
<td>29×</td>
<td>2.4</td>
</tr>
<tr>
<td>Gc2.a</td>
<td>26,377,060 (90.5)</td>
<td>36×</td>
<td>13,414,391 (86.5)</td>
<td>54×</td>
<td>13.4</td>
</tr>
<tr>
<td>Gc2.b</td>
<td>27,130,509 (93.2)</td>
<td>42×</td>
<td>14,137,799 (91.1)</td>
<td>56×</td>
<td>12.1</td>
</tr>
<tr>
<td>Gc2.ab</td>
<td>27,940,266 (95.9)</td>
<td>72×</td>
<td>14,868,092 (95.8)</td>
<td>87×</td>
<td>13.6</td>
</tr>
<tr>
<td>Average</td>
<td>27,425,586 (94.1)</td>
<td>37×</td>
<td>14,701,029 (94.7)</td>
<td>47×</td>
<td>5.1</td>
</tr>
</tbody>
</table>

*IDs, “a” and “b” are results from two independent sequence lanes for the same strain and “ab” results from two sequence runs combined for the same strain. The estimated
coverage is based on filtered reads mapped to the slkw1407 reference genome sequence, which is approximately 29.1 Mb after excluding gaps.
Ecological Assessments Using Gene Genealogies of Additional Strains

To support SNP phylogenetic relationships among the 12 *Grosmannia* genomes and to assess the host and distribution ranges of distinct lineages, we sequenced nine gene loci (supplementary table S8, Supplementary Material online) in 16 additional strains from localized populations of MPBs and JPBs in their respective host trees *P. contorta*, *P. ponderosa*, and *P. jeffreyi* (table 1). Genealogies from each of these genes (supplementary fig. S3, Supplementary Material online) and the concatenated phylogeny (fig. 3a) confirmed the genome-wide SNV results noted above by supporting the monophyly...
of the Gc and Gs clades and the following subclades. Within Gc, seven gene trees separated the taxa associated with JPBs (n = 10) in California from the MPB associates infesting *P. ponderosa* trees in British Columbia, California, and South Dakota (n = 6). The Gc–*P. ponderosa* subclade was statistically supported in the concatenated phylogeny (fig. 3a) and in one single-gene tree (CMQ6965–ABC.C, supplementary fig. S3, Supplementary Material online). The phylogeny from concatenated loci was also consistent with geographic isolation within Gs, with five strains from the localized population in California forming a monophyletic clade separated from the epidemic strains, but with low statistical supports. The nine-gene species tree showed identical topology based on ML, MP, and Bayesian analyses with minor differences in the placement of terminal taxa (fig. 3).

**Table 3.** Genome-Wide Characterization of Fixed and Shared Polymorphisms between Gs and Gc Lineages.

<table>
<thead>
<tr>
<th>Genomic Regions</th>
<th>Fixed&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Shared&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Exclusive&lt;sup&gt;c&lt;/sup&gt; to Gs (Parsimony Informative)</th>
<th>Exclusive to Gc (Parsimony Informative)</th>
<th>Total (Parsimony Informative)</th>
<th>Dxy (±SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>37,712</td>
<td>3,750</td>
<td>35,765 (18,871)</td>
<td>26,203 (9,685)</td>
<td>103,430 (70,018)</td>
<td>1.66 (± 0.006)</td>
</tr>
<tr>
<td>Intergenic</td>
<td>10,818</td>
<td>2,458</td>
<td>14,811</td>
<td>14,793</td>
<td>42,880</td>
<td>NC</td>
</tr>
<tr>
<td>Flanking regions</td>
<td>11,148</td>
<td>559</td>
<td>8,308</td>
<td>4,574</td>
<td>24,859</td>
<td>NC</td>
</tr>
<tr>
<td>Intronic</td>
<td>2,755</td>
<td>112</td>
<td>2,001</td>
<td>958</td>
<td>5,826</td>
<td>NC</td>
</tr>
<tr>
<td>Synonymous</td>
<td>6,808</td>
<td>353</td>
<td>4,984</td>
<td>2,744</td>
<td>14,889</td>
<td>NC</td>
</tr>
<tr>
<td>Nonsynonymous</td>
<td>6,116</td>
<td>264</td>
<td>5,601</td>
<td>3,059</td>
<td>15,040</td>
<td>NC</td>
</tr>
<tr>
<td>Stop gain–lost</td>
<td>73–18</td>
<td>4–1</td>
<td>80–14</td>
<td>69–3</td>
<td>226–36</td>
<td>NC</td>
</tr>
</tbody>
</table>

**Note:** NC, not calculated.

<sup>a</sup>Fixed polymorphisms are nucleotide sites, at which all Gs strains differ from all strains of Gc.

<sup>b</sup>Shared polymorphisms are sites for which multiple nucleotides are found in both Gs and Gc strains.

<sup>c</sup>Exclusive polymorphisms are those that are polymorphic in one species and invariant in the other.

Distinct Pattern of Limonene Utilization among *Grosmannia* Lineages

Consistent with results from the genome-wide SNP analyses and nine-gene phylogenies, we showed that although *P. ponderosa* and *P. jeffreyi* associates are genetically very close, they can be characterized with distinct pattern of (+)-limonene utilization. Consistent with *P. jeffreyi* producing a lower level of limonene than *P. contorta* and *P. ponderosa*, we found that no Gc isolates from *P. jeffreyi* grew on (+)-limonene minimum media, in contrast to all Gs isolates from *P. ponderosa*, as well as to all Gs and the closely related species from *P. contorta*, which did grow (fig. 3b).

Signature of Positive and Purifying Selections in *Grosmannia*

To test for positive adaptive selection in Gs–Gc orthologs, we compared the ratio of SNPs within Gs with the sequence divergence between Gs and Gc at nonsynonymous (*P*<sub>n</sub> nonsynonymous polymorphism; *D*<sub>n</sub> nonsynonymous divergence) and synonymous (*P*<sub>s</sub> synonymous polymorphism; *D*<sub>s</sub> synonymous divergence) positions (fig. 4a,c). For the 3,746 Gs–Gc orthologs, we obtained a median –log<sub>10</sub> NI (neutral index) value of less than 0 (–0.05; supplementary figs. S4 and S9, Supplementary Material online), which suggested that the majority of genes (n = 1,755) in our data set are subject to weak purifying selection. We also detected a statistically significant (*P* < 10<sup>–5</sup>) signal of purifying selection in the pooled analysis of all 3,476 genes, using an unbiased estimator of NI (–log<sub>10</sub> NI<sub>Gc</sub> = –0.11, pooled *P* = 3,834, *D*<sub>n</sub> = 5,903, *P*<sub>s</sub> = 3,267, *D*<sub>s</sub> = 6,612). However, only five genes showed significant evidence of purifying selection on a per-gene basis. Although 1,215 genes showed –log<sub>10</sub> NI > 0, indicating fewer amino acid polymorphisms within Gs relative to those between Gs and Gc, we only found 11 genes with statistically significant (*P* ≤ 0.05) excess of protein divergence between the two species (supplementary table S9, Supplementary Material online). Six of the 11 genes were among the 42 *Grosmannia* orthologs showing an excess of nonsynonymous fixed differences between Gs and Gc (i.e., the 1.2% of the 3,476 Gs–Gc polymorphic genes having *D*<sub>n</sub> ≥ 9; supplementary table S9, Supplementary Material online, and fig. 4a–c). Among genes exhibiting the strongest evidence for positive selection (i.e., –log<sub>10</sub> NI > 0), we noted PKS (CMQ_4392, _5323, _5095, and _2677), a nonribosomal peptide synthase (NRPS; CMQ_3566), ABC transporters (CMQ_6634, _6965, _6960), oxidoreductases (CMQ_1999, _5949), and an heterokaryon incompatibility gene (CMQ_742) (table 4 and supplementary table S9, Supplementary Material online). However, no genes were significant for either positive or purifying selection after correction for multiple testing (Benjamini and Hochberg 1995).

We also applied codon-based models and likelihood estimates of *dN*, *dS*, and *dN*/*dS* (ω) ratios. Divergence estimates were made from Gs–Gc pairwise alignments of the 3,476 orthologs using a codon substitution model that takes into account possible biases such as codon preference and nucleotide composition (Yang and Nielsen 2000). We estimated mean *dS* 0.0032 ± 4.8 × 10<sup>–5</sup>, corresponding to an average of one mutation per 312 synonymous sites between Gs and Gc since the common ancestor. This number was lower than the genome-wide average (one mutation per 446 nts, relative to the slkw1407 reference strain), presumably due to selective constraints in the coding regions. The mean pairwise *dN* (0.0011 ± 8.9 × 10<sup>–5</sup>) was lower than *dS*, reflecting the expected stronger constraints on substitutions that changed amino acids. The overall mean for *dN*/*dS* in the 3,476
orthologous genes (i.e., excluding 289 genes with dS = 0) was 0.3 ± 0.005. This value was similar to the \(-\log_{10} N_{ITG} = -0.11\) value obtained for the McDonald–Kreitman (MK) test, suggesting that a large majority of genes are conserved and evolve with dN/dS less than 1 (supplementary table S9 and fig. S4, Supplementary Material online).

The pairwise dN/dS ratio is a measure of the overall evolutionary constraint averaged across the sequences of the gene and may be too conservative for detecting positively selected sites along a gene. Thus, we applied “site-specific” models to test for further evidence of positive selection within a more divergent subset of the 3,476 orthologous genes,
removing 2,271 genes that had fewer than three fixed \( n = 1,567 \) and/or synonymous \( n = 704 \) differences. For the remaining 1,205 genes, the site-based approach identified 77 genes statistically significant for the positively selected sites \( \omega > 1; P \leq 0.05 \). For the majority of these significant genes \( n = 43 \), the MK test also estimated a summary statistic of positive selection \(-\log_{10} N_{\text{I}} > 0\), indicating an excess of protein divergence by both methods (supplementary table S10, Supplementary Material online). The genes exhibiting the strongest evidence for positively selected sites include PKS (CMQ_5095, _2687, _2677), an ABC transporter (CMQ_6965), CYP450s (CMQ_6107, _3491, _4067), oxidoreductases (CMQ_277, _5685), ankyrin-repeat containing proteins (CMQ_1651, _569), a heat repeat protein (CMQ_7934), and an autophagy protein (CMQ_7167) (table 4). The summary statistics on selection from MK and from PAML generally agreed (supplementary table S10, Supplementary Material online). However, the two methods both identified significant signal for positive selection in only one gene (PKS_5095, table 4). Another two PKS genes (CMQ_4392 and _2677) showed a significant or marginally significant signal for positive selection with both methods before correction for multiple tests. After correction for multiple tests, the signal was no longer significant (PAML \( P = 0.07 \) and MK \( P = 0.09 \)). The number of 43 significant genes (~4%) in our data set is lower than the conventionally accepted significance level of 5% because majority of genes are conserved and evolve with \( \omega \) less than 1. Nonetheless, after correction for multiple testing, we identified at least seven genes that evolved with \( \omega \) greater than one \( (P < 0.0001) \). This indicated that even though the level of divergence between the Gs and Gc was low, there is statistically significant evidence for site-specific positive selection between Grosmannia species. Results for all the genes are available in supplementary

---

**Fig. 4.** Comparison of divergence and polymorphism. (a) Number of synonymous and nonsynonymous Gs–Gc fixed differences (divergences) in 7,340 orthologous gene models. (b) Number of synonymous and nonsynonymous SNPs within Gs strains in 7,340 orthologous gene models. Circular bands with alternating shades of gray represent 36 scaffolds in which the gene models are located. Gray triangles mark genes with the largest numbers of nonsynonymous divergences (e.g., PKS, ANK, and ABC in a) and polymorphisms (e.g., ANK in b). (c) Summary distributions of MK cell entries for the fixed differences and SNPs in 3,476 variable genes (supplementary table S9, Supplementary Material online).
Discussion
In this study, to identify features common across distinct Grosmannia populations and species, we compare the

tables S9 and S10, Supplementary Material online; table 4 shows only the genes with the strongest evidence of positive selection using both PAML “site-model” and the MK test.
genomes of 12 Gc sensu lato strains, representing two known sibling species that have different ecological characteristics (Massoumi Alamouti et al. 2011). We first used genome assemblies to assess changes in genomic structure such as rearrangements and gene gains/losses and then focused on variation at the gene and nucleotide levels. We identified a number of functional variants in genes potentially involved in secondary metabolism and chemical detoxification, reflecting fungal adaptation to the specific chemistries of host trees. The data and results generated are a resource for assessing and characterizing fungal populations in the present MPB epidemic as it continues to spread into new habitats, including the P. banksiana boreal forest as well as in future MPB outbreaks. The approach described here can also be applied to other insect-vectored/tree-colonizing fungi.

**Grosmannia Genomes**

Using Illuma sequencing, we assembled the genome sequences from 11 Grosmannia strains that represent distinct populations of the two sibling species: Gs and Gc (Massoumi Alamouti et al. 2011). We showed that the de novo assemblies in these fungi could be mapped over a large fraction of the Grosmannia (Gs) reference genome (DiGuistini et al. 2011), suggesting that the majority of the assembled contigs, and the genes they contain, lie in regions that are collinear within and between the cryptic species. The extensive similarities in gene content (large-scale synteny) and order (colinearity) (Hane et al. 2011) within a large fraction of aligned contigs suggested that the morphologically cryptic Grosmannia species have diverged recently (Supplementary fig. S1, Supplementary Material online). This is consistent with the previous gene genealogies of Gc sensu lato and a few other close relatives, suggesting that the majority of the assembled contigs, and the genes they contain, lie in regions that are collinear within and between the cryptic species. The extensive similarities in gene content (large-scale synteny) and order (colinearity) (Hane et al. 2011) within a large fraction of aligned contigs suggested that the morphologically cryptic Grosmannia species have diverged recently (Supplementary fig. S1, Supplementary Material online). This is consistent with the previous gene genealogies of Gc sensu lato and a few other close relatives, which suggested that these pine-infesting, beetle-associated taxa have yet to reach a reciprocal monophyly for all the loci (Massoumi Alamouti et al. 2011). Large-scale structural changes can exceed nucleotide evolution in plant pathogens such as Mycosphaerella and Fusarium spp., which retain lineage-specific chromosomal islands or even entire lineage-specific chromosomes (Cuomo et al. 2007; Stukenbrock et al. 2010; Klosterman et al. 2011). In filamentous ascomycetes such structural changes may be attributed to relatively long divergence times or horizontal gene transfer (Hane et al. 2007; Desjardins et al. 2011; Hane et al. 2011; Klosterman et al. 2011). Here, major structural changes that would uniquely distinguish the cryptic Grosmannia species were not evident in our draft assemblies. Instead, the distinct ecological differences and host preferences in these fungi appear to be driven mainly by local nucleotide changes.

**Genome-Wide SNVs in Grosmannia**

Detecting genome-wide nucleotide variants within and between species using high-throughput sequencing depends on two factors: 1) whether the nonreference alleles are present in the strains sequenced and 2) the number of high-quality and accurately mapped reads that overlap the variant sites. The greater than 100,000 novel SNVs that we identified occurred in similar densities in intergenic, regulatory, and coding regions across the 11 strains and provide the first comprehensive assessment of genome-wide intra- and interspecific nucleotide variants for this group of beetle-vectored fungal symbionts. These SNV calls likely somewhat underestimate the total intra- and interspecific nucleotide differences between Grosmannia genomes, given that at least 10% of the reference genome had less than 5× read coverage—a limitation expected for Illumina sequencing of repetitive and GC-rich genomic regions (Li, Ruan, et al. 2008; Wang et al. 2011).

The genome-wide frequencies of nucleotide variants within Grosmannia species were lower than in other filamentous ascomycetes, including the plant pathogens Magnaporthe oryzae, Mycosphaerella graminicola, Sclerotinia sclerotiorum and different Verticillium and Cochliobolus species, as well as human pathogens in the genera Coccioidioides and Paracoccidioides, and the generalist saprophyte Neurospora crassa and species in the genus Aspergillus (Lambrechts et al. 2009; Ma et al. 2010; Neafsey et al. 2010; Amselem et al. 2011; Andersen et al. 2011; Desjardins et al. 2011; Klosterman et al. 2011; McCluskey et al. 2011; Stukenbrock et al. 2011; Xue et al. 2012; Condon et al. 2013). In these fungal species, whole-genome intraspecific SNV densities range from one per 865 nucleotides in the corn pathogen Cochliobolus heterostrophus to one per 132 bases in the human pathogen Paracoccidioides brasiliensis. These numbers are higher than the Grosmannia intraspecific variant frequency of one per 2,133 nucleotides and often higher than nucleotide divergence between Grosmannia sister species (i.e., one per 446 bases). Our intraspecific SNV frequencies were comparable to those of Fusarium graminearum, a global pathogen of cereal crops (Cuomo et al. 2007). This pathogen is a sordariomycete like the ophiostomatoid fungi; it differs from other filamentous ascomycetes, including Gc, because it is homothallic (i.e., self-fertile) and rarely out-crosses (Cuomo et al. 2007; Tsui et al. 2013). Fusarium graminearum’s inbreeding may be associated with lower nucleotide diversity, as is the case in other fungal and Oomycetes genomes (Tyler et al. 2006; Cuomo et al. 2007). The frequency of genome-wide SNVs in the opportunistic human pathogen Aspergillus fumigatus is similar to that for Grosmannia and is surprisingly low compared with its close relatives (Rydhom et al. 2006; Rokas et al. 2007). Aspergillus fumigatus’ low nucleotide variance and its lack of population structure globally have been explained by the worldwide spread of this fungus having occurred too recently for mutations to have accumulated within and between populations (Rydhom et al. 2006).

Differences in genome-wide frequency of SNVs among filamentous fungi may be in part due to differences in their life histories and dispersal processes. Ascomycetes comparative genomics have largely focused on saprotrophs that have broad host ranges and on pathogens that have the ability to survive for extended periods as free-living saprophytes without a specific host. Such fungi tend to have more stable population sizes and higher genetic variation in natural populations (Thompson 1994; Barrett et al. 2008). In contrast, fungal symbionts such as Grosmannia have limited and specific ecological niches (beetle vectors and host trees) and are
more likely to experience local population outbreaks, crashes, and recolonization than generalist and saprophytic fungi (Thompson 1994; Six and Paine 1999; Carroll et al. 2006; Smith et al. 2010; Roe et al. 2011; Tsui et al. 2012). After such crashes, long periods of low endemic population sizes are expected for both the beetle and its associated fungi. Such cycles promote loss of genetic variance within populations and generate between-population genetic differences, through genetic drift and adaptive selection. Consistent with the above, our results show that although Grosmannia fungi have lower overall genome-wide frequencies of nucleotide variants than other filamentous fungi, their SNVs support distinguishing two cryptic species and also suggest phylogenetically and biogeographically structured lineages that may include at least one additional species.

**Grosmannia SNV Phylogenomics**

Using genome-wide SNVs, we generated a high-resolution phylogeny that separated the 12 Grosmannia strains into two divergent monophyletic clades, confirming our previous genealogy discrimination of the Gs and Gc sister cryptic species (Massoumi Alamouti et al. 2011). If the two species share extensive polymorphism through introgression or incomplete lineage sorting due to a recent split from a common ancestor, we would expect that inter- and intraspecific nucleotide differences would be correlated (Avise 2004; Kulathinal et al. 2009). Here, no such correlation was evident; interspecific nucleotide divergence was significantly ($P < 0.01$) higher than the mean intraspecific variation within Gc and Gs, suggesting that gene flow between Grosmannia cryptic species was weak or absent. This was consistent with the low level of homoplasy in our SNV phylogeny (consistency index = 0.97) and with our previous genealogies using population-level samples (Massoumi Alamouti et al. 2011). The statistical support for each Grosmannia SNV phylogenetic group indicates that we can detect lineage-specific variants and so may be able to identify functional variants that are likely important to Gs and Gc adaptation to distinct ecological niches or to divergence of other phylogenetic groups resolved here.

Our SNV phylogeny divided the epidemic Gs strains into well-supported phylogenetic groups that were also identified previously using AFLP and microsatellite markers (Lee et al. 2007; Tsui et al. 2012). In addition, within Gs, we found a more divergent subclade, separating the strains from localized populations in California from the epidemic British Columbia subpopulations. The average pairwise nucleotide divergence between California and epidemic phylogenetic groups were more than twice as large compared with divergences within and among epidemic groups, likely due to California location being distant, in the southernmost part of the species’ range, along the Great Basin Desert (Wood 1982). Although genetic structures within localized Gs populations have not been documented before, they have been reported for the MPB populations using AFLP markers (Mock et al. 2007). MPB populations in California were more divergent compared with those from other epidemic and most of the localized populations, consistent with our results on the fungal associate. This consistency reflects the coevolutionary association between the beetle and the fungus, as suggested for other similar insect–fungal associations (Marin et al. 2009). MPB divergence based on AFLP makers was not significantly higher than expected for the isolation by distance, and it was suggested to correlate with the phylogenetic pattern of *P. contorta* trees experiencing a northward expansion into British Columbia and the Northwest Territories since the last glaciation period (Marshall et al. 2002; Mock et al. 2007). For the fungal associate, whether or not the Gs–California lineage warrants recognition as a species would require sampling additional isolates from the localized populations infesting *P. contorta* trees in the southern and eastern portion of the species’ range, preferably using SNV makers optimized for this application (Morin et al. 2004). Our previous network analysis on a 15-gene concatenated data set of the population-level samples from California and epidemic regions had shown incongruence among gene genealogies, inferring the evidence of either incomplete lineage sorting or recombination (Massoumi Alamouti et al. 2011). Either of these processes could be occurring in Gs populations. They may well have resulted from a recent species divergence maintaining high population size during the ongoing epidemics, a typical scenario in incomplete lineage sorting (Maddison and Knowles 2006). Recombination is also likely and indicative of the potential lack of species structures within Gs when phylogenomic analyses are applied to population-level samples.

Within the Gc clade, our whole-genome SNV phylogeny indicates host-specific differentiation in Grosmannia by separating the JPB associate from the holotype isolated from MPB-infested *P. ponderosa* (Pp) Consistency index in British Columbia (Robinson-Jeffrey and Davidson 1968). Consistent with these results, the protein-coding combined phylogeny of additional Gc strains suggested that one lineage (Gc–Pp) is exclusively associated with the JPB infesting the host tree *P. jeffreyi* in California, whereas the other (Gc–Pp) was only found on MPBs infesting *P. ponderosa* trees. The Gc from *P. ponderosa* host species in different geographic areas (i.e., BC, South Dakota, and California) was genetically closer than those collected from different host species (*P. jeffreyi*) in the same geographic region in California. Although our data from *P. ponderosa* trees were limited, preventing us from assessing the extent of host-specificity across the MPB-localized US populations, or the role of geographical isolation in speciation, overall, our results suggest that speciation process in these fungi can be attributed to the host-tree species and the geographic isolation of the host species from the current epidemics.

The genome-wide SNV divergence between the Gc–Pj and Gc–Pp was only twice as large as the intraspecific differences, reflecting the recent divergence of these lineages. A recently diverged population may represent an early stage in speciation, which begins when populations become genetically separated through geographical isolation or through ecological selection, and when adaptation acts as barrier to gene flow, and leads to genetically cohesive populations that are called
species because they are “segments of separately evolving lineages” (de Queiroz 2007). The genealogical nondiscordance criterion (Dettman et al. 2003) and the phylogeny of nine informative (i.e., genes randomly selected because of their potential fixed differences between the P. ponderosa and P. jeffreyi associates) protein-coding loci suggest that Gc–Pj and Gc–Pp are independent evolutionary lineages. The SNV phylogeny and gene genealogies were further supported by our current ecological data showing that each lineage was associated with distinct beetle and tree host species. Further characterization of lineage-specific SNVs at a population level would strengthen evidence for the work reported here, which suggests that lineages within Gc likely warrant recognition as genealogical and ecological species.

**Grosmannia Genes Involved in Host Adaptation and Ecological Divergence**

A combination of life history traits and selection imposed by host trees may have promoted speciation and ecological divergence in *Grosmannia* lineages, as shown for many plant pathogenic fungi (Giraud et al. 2006; Stukkenbrock and McDonald 2008; Giraud et al. 2010). In pine trees, phenolics and terpenoids from oleoresin are key constitutive and inducible chemical defenses (Keeling and Bohlmann 2006; Boone et al. 2011). Although monoterpenes (e.g., β-phellandrene and limonene) and heptane (a straight-chain alkane found in the oleoresin of P. jeffreyi) are toxic to many pathogens and insects, beetle–fungal complexes have evolved efficient mechanisms to survive and become established in such environments (DiGuistini et al. 2011; Wang et al. 2013). For Gs, functional genomics and transcriptomic data suggest that ABC transporters, genes associated with oxidative stress responses and fatty acid β-oxidation pathways, and gene clusters that contain cytochrome P450s, dehydrogenases, and mono-oxygenases are involved in overcoming tree defenses (Hesse-Orce et al. 2010; DiGuistini et al. 2011; Wang et al. 2013). However, chemical defense systems differ quantitatively and qualitatively between species of pine and between different populations within a pine species (Keeling and Bohlmann 2006; Gerson et al. 2009; Boone et al. 2011; Hall, Yuen, et al. 2013; Hall, Zerbe, et al. 2013). For example, P. jeffreyi has lower level of limonene and higher level of heptane than P. contorta (Mirov and Hasbrouck 1976; Paine and Hanlon 1994; Smith 2000). Limonene is one of the most toxic defense chemicals for bark beetle–fungal complexes (Raffa et al. 2001; Raffa et al. 2005); it influences MPB-attack density in epidemic regions, and it is found at high concentrations in *P. ponderosa* populations that have been subject to beetle–fungal outbreaks (Sturgeon 1979; Clark et al. 2010). Given this, host preferences among *Grosmannia* lineages may reflect different abilities to survive and adapt to host chemicals or to other biotic and abiotic stresses inside the host.

Changes in gene contents and in gene products are central mechanisms in fungal genome evolution. Genes lost or in the process of being lost through pseudogenization have been shown in plant pathogens (Stukkenbrock et al. 2010; Marce-Houben et al. 2012; Raffaele and Kamoun 2012; de Wit et al. 2012) and in the closely related human-pathogenic yeasts *Candida albicans* and *C. dubliniensis* (Moran et al. 2011). Similarly, 1.3% of the protein-coding genes in the Gs and Gc genomes contain premature stop codons, indicating that the genes may have been pseudogenized. Twenty-two of these genes have oxidoreductase activity, including those with known roles in stress response and detoxification like short-chain dehydrogenases, cytochrome P450s, and mono-oxygenases (Hesse-Orce et al. 2010; DiGuistini et al. 2011; Lah et al. 2013). Among those, 20 appear to have been lost in both Gc–Pp and Gc–Pj or in only one of these lineages. For example, a flavoprotein mono-oxygenase identified in the Gs gene cluster potentially involved in terpenoid detoxification and/or utilization (DiGuistini et al. 2011) has been pseudogenized in all the Gc–Pj strains tested *(n = 10). Sequencing of additional Gc–Pp and Gs strains and two related species confirmed that the stop codon is unique to the P. jeffreyi associates. Our physiological assessment using limonene as sole carbon source also showed that all *Grosmannia* fungi including three species from P. contorta (Gs, *Leptographium longiclavatum*, and L. terrebranis) and Gc strains from P. ponderosa were able to grow on limonene as a sole carbon source, but none of the Gc strains from P. jeffreyi grew or survived in this condition. These results agree with our ongoing work that shows that Gs requires mono-oxygenases to use limonene as a carbon source (Wang Y, Lim L, Lah L, Bohlmann J, Breuil C. unpublished data). Although we have natural and laboratory-made mutants for some of the enzymes, additional functional characterization needs to be carried out on a large scale to confirm how these enzymes modify or degrade monoterpenes, including limonene. Despite these limitations, our initial combined results suggest that a number of genes with potential roles in Gs host adaptation are inactivated or are evolving to become pseudogenes in Gc lineages. Because P. jeffreyi produces lower levels of monoterpenes (including limonene) than pine species in epidemic and localized populations (Mirov and Hasbrouck 1976; Paine and Hanlon 1994; Smith 2000), the Gc–Pj lineage may no longer require certain genes for processing some defense chemicals. The Gc lineages likely have more pseudogenes than we report here, because we have characterized only those caused by stop codons and not those due to indels and/or frameshift mutations. Host specificity seems to contribute to functional loss of genes and pseudogenes formation. Because lineage-specific pseudogenes may be unnecessary genes for colonizing particular *Pinus* species, we anticipate additional gene losses in Gs and Gc lineages in the future.

We assessed both purifying and positive selection in *Grosmannia* protein-coding genes. Under the assumption that synonymous changes are neutral, purifying selection is inferred when the ratio of nonsynonymous to synonymous substitutions is less than 1.0, whereas positive selection pressure is usually inferred when the ratio is greater than 1.0 (Wright and Andolfatto 2008). Similar to other filamentous ascomycetes, our genome-wide characterization of Gs–Gc protein-coding evolution showed that most genes evolve under purifying selection *(dN/dS = 0.3 ± 0.005)*, reflecting overall evolutionary constraints on protein-coding genes.
(Gu et al. 2005; Nielsen et al. 2005; Rokas 2009; Sharpton et al.
2009; Stukenbrock et al. 2010, 2011). In contrast, among all
the variable Grosmannia genes, only 43 showed significant
evidence for positive selection (i.e., before correction for mul-
tiple testing, P < 0.05), which is not surprising given the close
similarity between Gs and Gc orthologs (dS = 0.0032). We
note, however, that current divergence-based selection meth-
ods have limited statistical power for closely related species
(Li, Costello, et al. 2008; Oleksyk et al. 2010), and consequently,
we may have missed some genes with weaker signs of selec-
tion. For instance, sequence diversity and divergence in our
data suggested that 1,215 candidate genes were showing some
signs of adaptive selection (i.e., neutrality index [NI] –
log10 NI > 0), but the test was only significant for 11 genes
(P < 0.05).

Genes showing evidence of positive selection are likely
functionally important in divergence and/or ecological adap-
tation of Grosmannia fungi. The most significant examples of
evidence for positive selection were the four PKSs, one NRPS,
and three ABC transporters. The PKSs and NRPS families are
key enzymes for producing secondary metabolites, which are
involved in fungal host colonization and pathogenicity
(Kroken et al. 2003; Collemare et al. 2008; de Wit et al.
2012). On the basis of our ABC domain and phylogenetic
analyses (data not shown), the three membrane transporters
are classified in the ABC-C subfamily and so have potential
roles in either host-chemical defenses or secondary metabo-
lite export (Kovalchuk and Driessen 2010). Other genes with
putative functions in chemical detoxification or utilization
included an oxidoreductase, an isoflavone reductase, and
two cytochrome P450s (DiGuistini et al. 2011; Lah et al.
2013). We also found that some of these genes had putative
role in nutrient uptake (a ferric reductase and a monocarbox-
ylate permease). Other genes were potentially involved in cell
signaling (e.g., histidine kinase and phospholipase), fungal de-
development, and growth (e.g., membrane copper amine oxi-
dase, cell morphogenesis, autophagy protein, heat-repeat
protein, and hit finger domain protein) and a few with puta-
tive roles in protein–protein interactions or self-/nonself-
recognition (e.g., two ankyrin repeat proteins and a hetero-
karyon incompatibility protein) (Luhtala 2004; Fedorova et al.
2005; Bahn et al. 2006; Kohler et al. 2006; Liu and Gelli 2008;
Soanes et al. 2008; Pollack et al. 2009). In summary, for
Grosmannia lineages that are adapted to different pine
trees, our results suggest that many of the genes that are
evolving under positive selection are involved in secondary
metabolite synthesis and secretion, host-chemical detoxifica-
tion and stress responses, nutrient uptake from the host
plants, and hyphal growth and differentiation. Adding other
closely related species such as L. terebrantis and L. longiclav-
tum would increase the phylogenetic depth of our genome
data sets and the statistical power of the selection analyses.

In conclusion, we have used the Grosmannia genomes to
show relationships between ecology and biological functions
that are maintained or that diverge during colonization of a
range of pine host trees, which are themselves adapting to
changing environmental conditions. Although the fungal
population has expanded and contracted repeatedly over at
least several hundred years, large-scale synteny, with con-
served gene content and order, suggests that these closely
related strains adapt to different pine hosts largely through
local nucleotide changes. This genome-wide SNV data set is a
phylogenetic resource that can be extended into a more
comprehensive characterization of Grosmannia species ecol-
y and population structure.

Materials and Methods
Fungal Samples
We sequenced eight Gs genomes from two distinct popula-
tions of MPB-infested P. contorta trees: 1) epidemic regions in
Canada and the United States and 2) localized populations in
small geographically isolated outbreaks in California. We also
sequenced three genomes of the sibling species Gc. The sib-
ling group included two strains from JPB-infested P. jeffreyi
trees in California, as well as the Gc holotype described by
Robinson-Jeffrey and Davidson (1968) from MPB-infested P.
ponderosa trees in British Columbia. We deposited cultures
of these fungi at the University of Alberta Microfungus
Collection and Herbarium, along with additional Gs and Gc
strains used for SNP validations, phylogenies, and physiolog-
ical studies (table 1, supplementary fig. S5, Supplementary
Material online).

Illumina Paired-End Library Construction, Sequencing,
and Assembly
Fungal mycelia were grown on 2% malt extract (MEA; 33 g
Oxoid malt extract agar, 10 g Technical agar No.3, and 1
distilled water) overlaid with cellophane. DNA from the my-
celia was extracted using the method of Möller et al. (1992).
DNA samples were processed at the Genome Science Center
(GSC, Vancouver, BC, Canada) for paired-end sequencing
following Illumina protocols (Illumina, Hayward, CA). The li-
brary for each strain was amplified in a single flow cell and
sequenced to either 50 or 76 nucleotide base (nt) reads on the
Illumina Genome Analyzer (GA) II or IIX following the man-
ufacturer specifications.

Genomes were assembled from Illumina paired-end reads
of 200 base DNA fragments using the ABYSS assembler v1.2.7
(Simpson et al. 2009). Reads that passed the chastity filter
(Haridas et al. 2011) were assembled with a relative short
kmer (25–31 nt) for higher sensitivity. The resulting contigs
were used as single end reads along with the original paired
end data and reassembled with a higher kmer (35–61 nt),
which has a higher specificity (supplementary table S1,
Supplementary Material online). The assembly was cleaned
and gaps closed using Anchor (www.bcgsc.ca/platform/
bioinfo/software, last accessed March 20, 2014). Ambiguous
base calls were resolved by mapping the reads back to the assembly using BWA v0.5.9 (Li and Durbin 2009)
and calling consensus bases using SAMtools “mpileup” v0.1.18
(Li et al. 2009). Assembled contigs and scaffolds larger than
200 nucleotides were ordered and oriented using MUMmer
(Kurtz et al. 2004) based on the Grosmannia published
genome (sklw1407; NCBI Genome PID: 39837; DiGuistini
et al. 2009). Contigs that did not align with slkw1407 were

Downloaded from https://academic.oup.com/mbe/article-abstract/31/6/1454/2925720 by guest on 16 January 2019
We performed variant calling among the draft genome using the homology-based gene predictor genBlastG (She et al. 2011). We used protein sequences from the slkw1407 reference genome as the query ($N = 8,312$) and the genome of another strain as the target database. Gene annotations and pairwise homology between the slkw1407's gene models and those from each genome were assigned based on genBlastG hits with an $E$-value cutoff of $\leq 1e-10$ and a query coverage of $>50\%$. The genBlastG output can result in redundant gene predictions when the query gene belongs to a multigene family, paralogous genes, or tandem gene duplications. Given this, for downstream analyses, we applied a filtering procedure, so that each genomic region would contain only one gene prediction with the highest global sequence percent identity (PID) to the query. The filtering procedure was carried out as follows: 1) all gene predictions were sorted by PID, 2) for each two overlapping gene model, if the overlapping region was $> 5\%$ of the length for each gene, then only the prediction with higher PID was kept. 3) all gene models were required to have PID $\geq 70\%$ to the query, and 4) to avoid assigning paralogs to query-target pairs, the best match had to have a PID $10\%$ higher than the next best match. Nonoverlapping gene models with high similarity to the same query were reported as putative paralogs and were removed from analysis. For genes with alternative splicing variants, the longest transcript was selected to represent the gene. After filtering incomplete genes and discarding genes with frame shifts, which could have been caused by the draft quality of the genomes, only high-quality 1:1 orthologous genes were retained for analysis. Gene models for all Grosmannia genomes are available upon request in annotation files.gff. Supplementary table S11, Supplementary Material online, summarizes genBlastG output used to find the pairwise homology between reference gene models and those of each draft genome.

Mapping and Variant Calling

To detect Grosmannia orthologs, we first generated gene annotations for each draft genome using the homology-based gene predictor genBlastG (She et al. 2011). We used protein sequences from the slkw1407 reference genome as the query ($N = 8,312$) and the genome of another strain as the target database. Gene annotations and pairwise homology between the slkw1407's gene models and those from each genome were assigned based on genBlastG hits with an $E$-value cutoff of $\leq 1e-10$ and a query coverage of $>50\%$. The genBlastG output can result in redundant gene predictions when the query gene belongs to a multigene family, paralogous genes, or tandem gene duplications. Given this, for downstream analyses, we applied a filtering procedure, so that each genomic region would contain only one gene prediction with the highest global sequence percent identity (PID) to the query. The filtering procedure was carried out as follows: 1) all gene predictions were sorted by PID, 2) for each two overlapping gene model, if the overlapping region was $> 5\%$ of the length for each gene, then only the prediction with higher PID was kept. 3) all gene models were required to have PID $\geq 70\%$ to the query, and 4) to avoid assigning paralogs to query-target pairs, the best match had to have a PID $10\%$ higher than the next best match. Nonoverlapping gene models with high similarity to the same query were reported as putative paralogs and were removed from analysis. For genes with alternative splicing variants, the longest transcript was selected to represent the gene. After filtering incomplete genes and discarding genes with frame shifts, which could have been caused by the draft quality of the genomes, only high-quality 1:1 orthologous genes were retained for analysis. Gene models for all Grosmannia genomes are available upon request in annotation files.gff. Supplementary table S11, Supplementary Material online, summarizes genBlastG output used to find the pairwise homology between reference gene models and those of each draft genome.

For each strain, filtered reads were mapped to the slkw1407 reference genome sequence using BWA v0.5.9, with the default parameters (Li and Durbin 2009). Initial mapping results were converted into the indexed and sorted Binary Alignment/Map (BAM) format using SAMtools v0.1.18 (Li et al. 2009) and Picard v1.54 (http://picard.sourceforge.net, last accessed March 20, 2014). To enhance the quality of the alignments for more accurate variant detection, we used GATK (McKenna et al. 2010) to locally realign the BAM files in complex regions, for example, containing insertions/deletions (indels). For each alignment, BWA assigned a mapping quality score (MAPQ). We used reads with MAPQ greater than 0 to estimate the coverage and average read depth of final BWA alignments, using BEGtools v2.13.4 (Quinlan and Hall 2010). The individual BAM data sets are available upon request. Once reads from individual strains were mapped to the slkw1407 genome, we used SAMtools “mpileup” to assess variant sites, applying Base Alignment Quality computation and a $-c50$ argument to minimize alignment artifacts and base-calling errors. Single nucleotide variants (SNVs) were identified using the Bayesian variant calling models implemented in “bcftool” (Li, Ruan, et al. 2008). After consensus base calling, we filtered the initial variants for strand and distance biases ($P$ value $< 0.0001$) using SAMtools “vcfutils.pl.” The final set of high-quality calls also required a candidate site to be biallelic and to meet the following criteria: minimum Phred-scaled base calling score of 20, MAPQs of at least 30, read depths of more than four and less than 250, and a minimum $10\text{nt}$ distance from indels. Variant calls that failed to meet these criteria were likely to be false positives. Because SAMtools “mpileup” assumes a diploid model and our samples represent haploid genomes, we also removed heterozygote calls.

Verification of Variant Calls

To estimate the robustness of SAMtools results, genomic variants were also assessed using the SNV calling algorithm implemented in GATKv1.40 (McKenna et al. 2010). This method also uses a Bayesian model to estimate the likelihood of a site harboring an alternative allele for each sample. GATK was run on the same BAM files as SAMtools, using default parameters. GATK raw-variant calls were filtered in the same manner as the SAMtools calls (see earlier). To estimate SNV false positives in our data set, we generated Illumina paired-end reads for the slkw1407 reference strain (supplementary table S4, Supplementary Material online) and assessed variant calls for these reads mapped against their own published genome and identified 1,796 high-quality SNVs. Because the alternate base was present in all 11 genomes and also in Illumina read alignments from the slkw1407, a large percentage (93.2%) of these changes likely represent errors in the reference genome assembly. We removed these ambiguous calls from the final SNV data set. For 16 additional Gc and Gs isolates (table 1), we also used PCR and Sanger sequencing to validate SNPs in the nine candidate genes listed in
supplementary table S8, Supplementary Material online. For the 11 *Grosmannia* strains, we aligned homologous contigs of the candidate genes to those of slkw1407 genome and gene models using progressive Mauve 2.3.1 (Darling et al. 2010). Primers were designed based on the alignment using Geneious 5.1 (Biomatters Ltd, New Zealand). Amplicons were purified and sequenced at the Nucleic Acid Protein Service Unit at the University of British Columbia (Vancouver, Canada). All sequences and alignment matrices are available at TreeBASE (S15463, ID M21081–89).

Functional Annotations for SNVs
SNVs were annotated as coding (synonymous and nonsynonymous), intronic, flanking, and intergenic, with SNPeffect v.2.0.5 (Reumers et al. 2006), using the slkw1407 genome’s sequence and predicted gene models. We assigned flanking regions (i.e., UTRs and putative regulatory regions) of 1,000 nt upstream and downstream of the initiation/termination codons of the annotated slkw1407 gene models, unless neighboring gene sequences were within this range; for such cases, we truncated the assigned regions. We also characterized, as a set of variants, SNVs that result in the loss or gain of a stop codon, which likely affect the integrity of the protein products. Orthologous genes containing a premature stop codon were labeled as pseudogenes. We assessed the accuracy of stop codon variant calls using expressed sequence tag libraries and RNA-seq data from slkw1407 (DiGuistini et al. 2009) as well as Illumina reads from more than one strain within each Gs and Gc group. Finally, we applied Gene ontology (GO) functional enrichment analysis (MF or BP) on pseudogene candidates. The GO term associations were determined for each slkw1407 reference gene models using Blast2Go v2.5.0 with the default parameters (Conesa et al. 2005). Blast2GO was also used for a GO functional enrichment analysis; for that we performed the Fisher’s exact test with a false discovery rate correction to obtain an adjusted P value between the candidate genes and the whole genome annotation.

SNVs Clustering and Phylogenomics
We used the genome-wide SNV data to determine phylogenomic relationships and the nucleotide divergence among *Grosmannia* genomes. To assess genome-wide-SNV clusters among a relatively small number of *Grosmannia* strains, we used the nonparametric AWclust (Gao and Starmer 2008) R package, because it requires no model assumptions (e.g., Hardy–Weinberg equilibrium) and is based on hierarchical clustering of a distance matrix rather than on allele frequency variation. We compared the clustering results with inferences from MP and Bayesian phylogenetic analyses, for which we concatenated the genomic SNV data set into one continuous sequence for each strain (total character = 103,430). MP trees were identified using PAUP* 4.0b10 (Swofford 2003) by heuristic searches with TBR branch swapping and the MULPARS option, and 100 random sequence additions. Bayesian analyses used MrBayes 3.2 (Ronquist and Huelsenbeck 2003), under the best-fit substitution model selected by the Akaike information criterion implemented in JModelTest 0.1.1 (Posada 2008). Each run consisted of four incrementally heated Markov chains, using default-heating values. The chains were initiated from a random tree and were run for 2 million generations with sampling every 1,000 generations. To assess the confidence of phylogenomic analyses, MP BS values were calculated with 1,000 replicates and the heuristic option (Felsenstein 1985) using PAUP*, and Bayesian posterior probabilities (PP) were inferred with a 50% majority-rule consensus tree that was sampled after the likelihood scores had converged, using MrBayes. The stationarity of likelihood scores for sampled trees was assessed in Tracer v1.5 (Rambaut and Drummond 2009), and the convergence was assessed using cumulative posterior probability plots in AWTY (Nylander et al. 2008) to assess split frequency within and between Markov chain Monte Carlo runs. The roots of the resulting trees were inferred by midpoint rooting. Mean nucleotide divergence (Dxy) was calculated using the maximum composite likelihood method implemented in Mega 5.0 (Tamura et al. 2011) and was averaged across 1,000 BS replicates. The 103,430-SNV-character matrix used in the cluster and phylogenetic analyses is deposited in TreeBASE (S15463, ID M21079).

Gene Genealogies and Concatenated Data Phylogeny
To assess biogeographic traits resolved using the genome-wide SNV data set, we randomly selected nine gene loci (supplementary table S8, Supplementary Material online) that showed putative fixed differences between distinct Gs and Gc populations and sequenced them in 16 additional strains (table 1). For each of the nine gene data sets, we generated MP and statistical-parsimony genealogies using PAUP and TCS v. 1.13 (Clement et al. 2000). Gaps were treated as missing data, and no weighting was introduced in the single-gene analyses. The nine gene loci were concatenated to conduct maximum likelihood (ML) analysis (with 1,000 nonparametric replicates BS) using RAxML-VI-HPC 7.0.4 (Stamatakis 2006), as well as weighted MP, with the weighting inversely proportional to the number of parsimony informative characters at each locus. We also performed Bayesian analyses for each gene and for the combined data set. For Bayesian and MP analyses and for assessing their confidence and best-fit model of sequence evolution, we used the same criteria as those applied to construct SNV phylogenies. Monophilies supported by both BS ≥ 70% and PP ≥ 95% were considered significant. The multigene data sets and related phylogenetic trees are deposited in TreeBASE (S15463, ID M21080–89).

Physiological Assessments
We characterized the monoterpene utilization of (−)-limonene as a carbon source by *Grosmannia* strains from three different pine trees *P. contorta, P. jeffreyi*, and *P. ponderosa*. For this experiment, we selected five Gs and Gc strains from independent samples of each tree species (total *n* = 15, table 1). The 3-day fungal cultures actively growing on MEA were transferred to glass plates containing yeast nitrogen base minimal medium (0.17% YNB, 1.5% granulated agar), where 200 μl of (−)-limonene (Sigma, Oakville, ON) were added.
onto two (2 × 4 cm) strip filter papers that were placed inside the lid of each glass plate. The plates were sealed with DuraSeal film (Laboratory Sealing Film; VWR, Mississauga, Ontario, Canada) and incubated at 22 °C in a sealed glass container. Limonene was resupplied biweekly with the same volume as the initial one; after 6 weeks, the mycelial plugs treated with limonene were transferred to normal MEA plates to check whether the fungus was killed or survived the chemical treatment. The control was YNB minimal medium without monoterpenone.

Detecting Signature of Selection and Rate of Protein Evolution

For selection analyses with Gs–Gc multiple alignments, we first searched for genes that were orthologous to the 8,312 gene models of the reference strain slkw1407. We found an average of 8,064 orthologs for the 11 assemblies, ranging from 7,876 to 8,222 in Gs and 7,973 to 8,198 in Gc (supplementary table S3, Supplementary Material online). We retained orthologs to 7,340 slkw1407 genes that matched at least four of the eight Gs and/or at least two of the three Gc genomes (n = 972) and removed 3,864 of these because they had either fewer than two coding differences (n = 3,377) or zero divergence (Dn + Ds = 0; n = 487). The selection analyses included the remaining 3,476 orthologs, which contained 19,616 nucleotide differences in coding sequences with a median size of 1,749 aligned bases, after excluding the gaps (supplementary table S9, Supplementary Material online). The average number of Gs–Gc genomes in the aligned data sets was n = 8.6.

We then applied different methods to Gs and Gc gene predictions that were orthologous to the slkw1407 gene models for detecting positive selection. First, we compared polymorphisms within Gs (n = 4–8 strains) with fixed substitutions (i.e., divergence) between Gs and Gc sequences. We considered synonymous and nonsynonymous differences and used two Gc strains from P. jeffreyi and P. ponderosa as the outgroup taxa. We used Gs–Gc multiple alignments of all genes with at least two coding differences that were aligned by MAFFTv7.023 (Katoh et al. 2002) for their entire coding regions and applied the MK tests (McDonald and Kreitman 1991) implemented in the MK.pl script (Holloway et al. 2007). We assessed whether the ratio of nonsynonymous and synonymous was statistically independent of differences being polymorphic (Pn/Ps) or divergent (Dn/Ds), using Fisher’s exact test. For each gene, MK results for the direction and degree of departure from neutrality were summarized using the NI (Rand and Kann 1996), after adding one pseudocount to each mutation class to eliminate zero counts. We also reported an unbiased NI estimate for differences across all the genes (NIITG; Stoletzki and Eyre-Walker 2011).

Next, we applied ML methods implemented in the Codeml from PAMLv4.0 (Yang 2007). We estimated Gs–Gc pairwise distances at nonsynonymous (dN) and synonymous (dS) sites for each gene, by setting parameters as follows: seqtype = 1, CodonFreq = 2, Runmode = −2, and the transition–transversion ratio K estimated from the data (Goldman and Yang 1994). For this test, we used pairwise alignments of single coding sequences from each species, generated for all Gs and Gc strains; the number of pairwise comparisons ranged from 8 to 27 per gene. Threshold dS values were determined by plotting dN as a function of dS, excluding outliers from the main distribution. To test for further evidence of positive selection, we applied the “site-specific” models M1a/M2a and M7/M8 (Nielsen and Yang 1998). Only gene alignments displaying more than three fixed (Dn + Ds ≥ 3) and/or synonymous (Dn + Ds ≥ 3) differences were considered for this additional test (Stoletzki and Eyre-Walker 2010). M1a assumes that codons contain only 0 < dN/dS < 1 or dN/dS = 1. We compared this with the alternative model M2a, which allows dN/dS for a site to be less than, equal to, or greater than 1. If dN/dS is significantly greater than 1, then adaptive substitutions are assumed to have occurred to fix nonsynonymous differences between species. If dN/dS < 1, adaptive evolution may still have occurred on some fraction of all differences but cannot be inferred with certainty. We also compared the null model M7, which assumes a beta distribution of 0 ≤ dN/dS ≤ 1 across sites with the alternative model M8, which allows for positive selection (Yang and Nielsen 2000; Yang and Swanson 2002). The log likelihoods for the null and alternative models were used to calculate a likelihood ratio test statistic, which was then compared against the χ² df = 2 distribution (Yang 2007). The positive selection hypothesis was accepted if both alternative models M2a and M8 provided a statistically significant better fit to the data. For all the analyses, we removed low-frequency polymorphisms (singletons) to avoid biases caused by slightly deleterious mutations regarding the prevalence of adaptive divergence (Fay et al. 2001; Li, Costello, et al. 2008).

Supplementary Material

Supplementary figures S1–S5 and tables S1–S11 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank Canada’s Michael Smith Genome Science Center (Vancouver, BC, Canada) for sequencing the 11 Grosmannia genomes. The authors would like to recognize the excellent technical work of Lynette Lim in the CB laboratory. They greatly benefited from discussions with Mary Berbee, Kermit Ritland, Sally Otto, and Ye Wang, whose suggestions improved the manuscript. They thank Dr Richard Hamelin for allowing us to use the selection analysis pipeline developed for the TAIGA project (Tree Agressors Identification using Genomics Approaches) founded by Genome BC and Genome Canada. They thank Ms Karen Reid for excellent project management of the Tria project. This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC) to J.B. and C.B. and by Genome Canada, Genome BC, and Genome Alberta to J.B. and C.B., which supported the Tria project (www.thetriaproject.ca).
References


Cochliobolus sativus ATCC 1015 versus enzyme-producing CBS 513.88. Genome Res. 16:2174–2183.


