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]

[1]Department of Wood Science, University of British Columbia, Vancouver, British Columbia, Canada

[2]DOE Joint Genome Institute, Walnut Creek, California

[3]Department of Forest and Conservation Sciences, University of British Columbia, Vancouver, British Columbia, Canada

[4]British Columbia Cancer Agency Genome Sciences Centre, Vancouver, British Columbia, Canada

[5]Michael 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

doi:10.1093/molbev/msu102 Advance Access publication March 12, 2014

beetle mass colonization (Lee et al. 2006). This fungus forms a symbiotic relationship with MPB and its sister species the jeffrey pine beetle (JPB; *Dendroctonus jeffreyi*). Although the two beetles are morphologically and genetically similar, they are adapted to different host trees (Six and Paine 1997). JPB is highly specialized and colonizes only *P. jeffreyi*, whereas MPB primarily inhabits *P. contorta* but can also successfully colonizes more than 20 pine species but not *P. jeffreyi* (Wood 1982).

Complexes consisting 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 genealogies 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 and JPB 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. 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-positive rate of 4.4×10^{-6} or one per 24,590 nts and a false-negative 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

Table 1. Fungal Strains Used in This Study.

Fungal Species	Beetle Associate	Host Tree	Collection Site (Map No.[a])	Source[b]	Code[c]
Gs	*Dendroctonus ponderosae*	*Pinus contorta* (Pc)	Canada, BC, Kamloops (1)	*(UAMH 11150)*	GsB1
			BC, Houston (2)	*UAMH 11153*	GsB2
				(UAMH 11348)	GsB3
		Pc × *P. banksiana*	Canada, Alberta, Fox Creek (3)	*(UAMH 11353)*	GsA1
				UAMH 11354	GsA2
		Pc	Alberta, Cypress Hills (4)	*UAMH 11347*	GsA3
			USA, Montana, Hidden Valley (5)	*(UAMH 11156)*	GsM1
			USA, California, Sierra Nevada (6)	*(UAMH 11349)*	GsC1
				UAMH 11350	GsC2
				CB 67F21	GsC3
				UAMH 11361	GsC4
				UAMH 11362	GsC5
Gc	*D. ponderosae*	*P. ponderosa* (Pp)	BC, Cache Creek (7)	*(ATCC 18086)*	GcB1
			USA, South Dakota, Black Hills (8)	UAMH 11369	GcS1
				(UAMH 11370)	GcS2
				(UAMH 11371)	GcS3
				(CB 15B29C2)	GcS4
				CB 23B110C5	GcS5
				CB 32B85C10	GcS6
			USA, California, Sierra Nevada (6)	(UAMH 11372)	GcC11
			California, Lassen (9)	UAMH 11373	GcC12
	D. jeffreyi	*P. jeffreyi* (Pj)	California, San Bernardino (10)	*(UAMH 11351)*	GcC1
				(UAMH 11352)	GcC2
				DLS 1560	GcC3
				DLS 1595	GcC4
			California, Lassen (9)	UAMH 11377	GcC6
				(DLS 210)	GcC7
			California, Sierra Nevada (6)	(C 843)	GcC5
				(UAMH 11375)	GcC8
				DLS 681	GcC9
		Pj × Pp	California, Lassen (9)	UAMH 11374	GcC10
Leptographium terebrantis	*D. ponderosae*	Pc	BC	UAMH9722	
		Pc × *P. banksiana*	Canada, Saskatchewan	AU 123-113	
L. longiclavatum	*D. ponderosae*	Pc	BC	CB LKG + T2B	
	D. ponderosae	Pc	BC	UAMH 4876	

Color shades represent the beetle and tree species, from which each fungal strain was isolated.

[a] General map location of collection sites corresponding to supplementary figure S5, Supplementary Material online.

[b] The fungal strain UAMH 11150 used for the reference genome was published by DiGuistini et al. (2011). Strains selected for Illumina sequencing are italicized. Selected for Illumina sequencing, and fungal strains used for the physiological assessment are shown in parentheses. Source of isolates: UAMH, University of Alberta Microfungus Collection and Herbarium, Canada; ATCC, American Type Culture Collection, USA; Isolates beginning with CB, DLS, AU, and C are from culture collections of C. Breuil, University of British Columbia, Canada; D.L. Six, University of Montana, USA; A. Uzunovic, FPInnovations, Canada; and T.C. Harrington, Iowa State University, USA, respectively.

[c] Letters indicate the location and numbers indicate the number of isolates from each location.

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

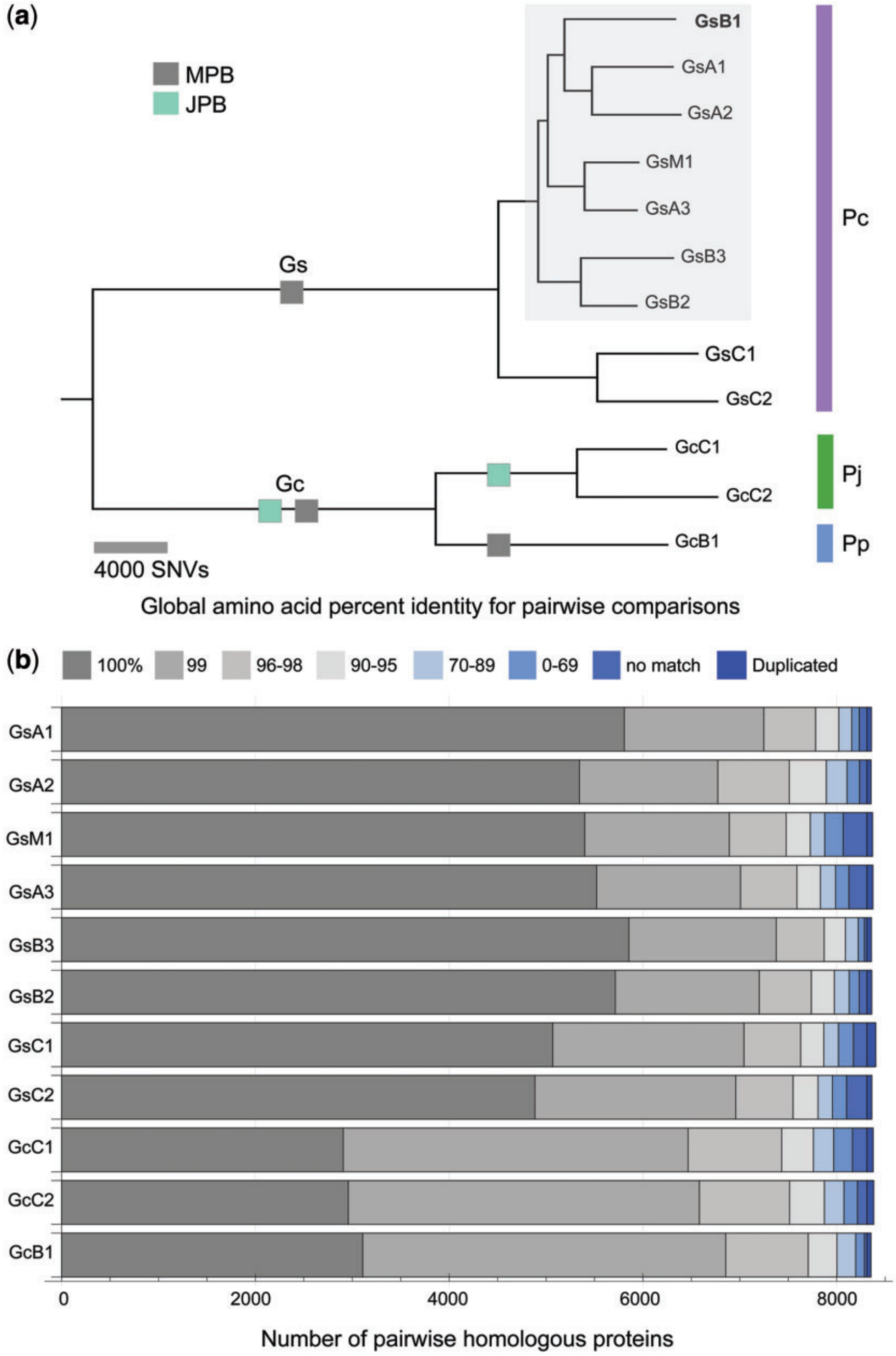


Fig. 1. *Grosmannia* SNP phylogenomics, gene content, and amino acid similarity. (*a*) MP analysis of 103,430 SNPs among 12 *Grosmannia* genomes. The analysis was best described by a single unrooted tree with consistency index of 0.79 and 0.97 when including only Gs (GsRef, GsB3, and GsC1) and Gc (GcC2 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.

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

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

IDs[a]	Covered Genomic Bases (%)	Genomic Coverage	Covered Gene Bases (%)	Gene Coverage	Unmapped Reads (%)
GsB1 (control)	28,791,583 (98.8)	70×	15,507,591 (99.9)	47×	0.7
GsB2	27,370,438 (94.0)	27×	15,333,848 (98.8)	30×	1.2
GsB3	28,464,740 (97.7)	48×	15,316,500 (98.7)	48×	2.6
GsA1	28,488,061 (97.8)	58×	15,334,490 (98.8)	61×	3.2
GsA2	26,491,256 (91.0)	35×	13,533,085 (87.2)	78×	5.0
GsA3	27,146,569 (93.2)	24×	15,203,277 (98.0)	29×	0.8
GsM1	27,197,413 (93.4)	22×	15,170,992 (97.8)	33×	1.3
GsC1	28,153,881 (96.6)	50×	15,092,814 (97.3)	57×	4.1
GsC2	28,084,478 (96.4)	49×	15,031,784 (96.9)	63×	4.0
GcB1.a	26,267,089 (90.1)	14×	13,636,725 (87.9)	20×	6.8
GcB1.b	28,360,091 (97.4)	55×	15,399,793 (99.3)	50×	10.1
GcB1.ab	28,455,969 (97.7)	79×	15,435,487 (99.5)	77×	9.2
GcC1	27,001,022 (92.7)	25×	15,204,933 (98.0)	29×	2.4
GcC2.a	26,377,060 (90.5)	36×	13,414,391 (86.5)	54×	13.4
GcC2.b	27,130,509 (93.2)	42×	14,137,799 (91.1)	56×	12.1
GcC2.ab	27,940,266 (95.9)	72×	14,868,092 (95.8)	87×	13.6
Average	27,425,586 (94.1)	37×	14,701,029 (94.7)	47×	5.1

[a] 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.

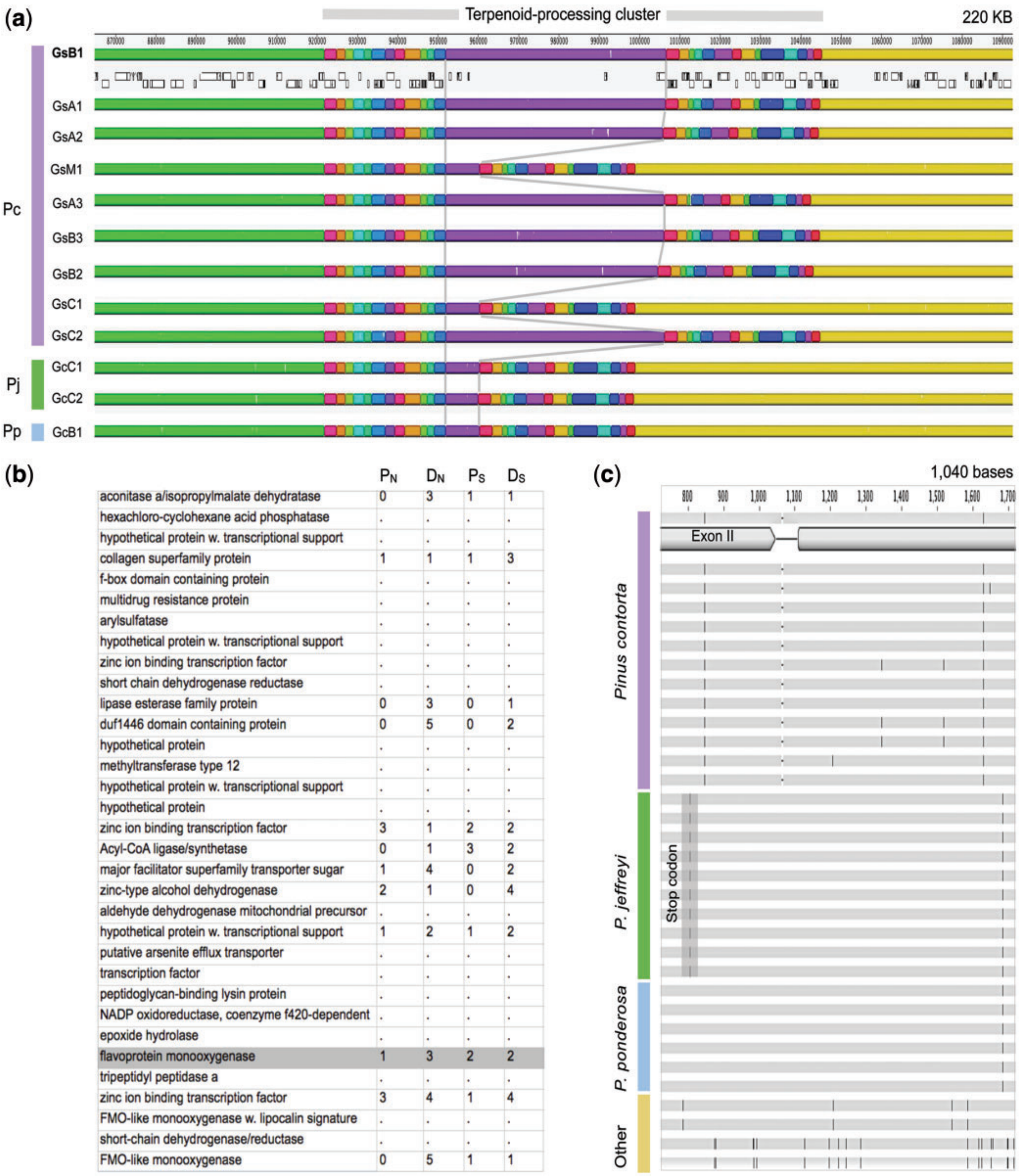
and S7, Supplementary Material online). For example, a flavoprotein mono-oxygenase (CMQ-6740) in the slkw1407-gene cluster (fig. 2*a*,*b*), which was proposed to have a role in detoxification and/or utilization of host-tree defense chemicals (DiGuistini et al. 2011), showed a stop codon in both Gc strains 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. 2*c*).

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 nucleotide 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 interspecific 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 differed from the slkw1407 reference genome by an average of 12,859 SNPs and 3,315 short indels, corresponding to one SNP per 2,133 nucleotides in the approximately 27.4 Mb covered regions (table 2 and supplementary table S5, Supplementary Material online). In contrast, the three Gc strains showed an average of 61,512 SNPs and 6,878 short indels, corresponding to one SNP per 446 nucleotides. The mean single-nucleotide divergence between the Gs and Gc genomes was estimated at 1.66 (±0.006), which was, respectively, approximately 7 and 11 times higher than mean intraspecific divergence for Gc (0.24 ± 0.002) and Gs (0.15 ± 0.0006).

Clustering 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 geographic 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 identical 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 different 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).



	P_N	D_N	P_S	D_S
aconitase a/isopropylmalate dehydratase	0	3	1	1
hexachloro-cyclohexane acid phosphatase	.	.	.	.
hypothetical protein w. transcriptional support	.	.	.	.
collagen superfamily protein	1	1	1	3
f-box domain containing protein	.	.	.	.
multidrug resistance protein	.	.	.	.
arylsulfatase	.	.	.	.
hypothetical protein w. transcriptional support	.	.	.	.
zinc ion binding transcription factor	.	.	.	.
short chain dehydrogenase reductase	.	.	.	.
lipase esterase family protein	0	3	0	1
duf1446 domain containing protein	0	5	0	2
hypothetical protein	.	.	.	.
methyltransferase type 12	.	.	.	.
hypothetical protein w. transcriptional support	.	.	.	.
hypothetical protein	.	.	.	.
zinc ion binding transcription factor	3	1	2	2
Acyl-CoA ligase/synthetase	0	1	3	2
major facilitator superfamily transporter sugar	1	4	0	2
zinc-type alcohol dehydrogenase	2	1	0	4
aldehyde dehydrogenase mitochondrial precursor	.	.	.	.
hypothetical protein w. transcriptional support	1	2	1	2
putative arsenite efflux transporter	.	.	.	.
transcription factor	.	.	.	.
peptidoglycan-binding lysin protein	.	.	.	.
NADP oxidoreductase, coenzyme f420-dependent	.	.	.	.
epoxide hydrolase	.	.	.	.
flavoprotein monooxygenase	1	3	2	2
tripeptidyl peptidase a	.	.	.	.
zinc ion binding transcription factor	3	4	1	4
FMO-like monooxygenase w. lipocalin signature	.	.	.	.
short-chain dehydrogenase/reductase	.	.	.	.
FMO-like monooxygenase	0	5	1	1

Fig. 2. Intra-/interspecific variants in the terpenoid-processing gene cluster. (*a*) Alignment of 12 homologous supercontigs shows complete synteny and colinearity for all *Grosmannia* strains. GsB1: *Grosmannia* reference genome. Locally collinear blocks (LCB, shown in the same color) have similar sizes among strains, and the 33 gene models that are potentially involved in limonene utilization or detoxification are in complete synteny. The apparent indel (purple LCB) is located in an intergenic region that may have been subjected to assembly error for a few Gs and Gc strains. To rule out the possibility of assembly error, a contig longer than the indel (puple LCB) is required, but these data were missing for all the strains showing the deletion. (*b*) The 33 orthologous genes in the terpenoid-processing cluster showed relatively low numbers of polymorphisms (nonsynonymous, P_N; synonymous: P_S) and divergence (nonsynonymous, D_N; synonymous, D_S). Dots represent genes with less than two coding nucleotide differences. (*c*) The flavoprotein mono-oxygenase (CMQ-6740) has a stop codon in its second exon that is unique to the Gc isolates from *Pinus jeffreyi*.

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. 3*a*) confirmed the genome-wide SNV results noted above by supporting the monophyly

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

Genomic Regions	Fixed[a]	Shared[b]	Exclusive[c] to Gs (Parsimony Informative)	Exclusive to Gc (Parsimony Informative)	Total (Parsimony Informative)	Dxy (±SD)
Total	37,712	3,750	35,765 (18,871)	26,203 (9,685)	103,430 (70,018)	1.66 (± 0.006)
Intergenic	10,818	2,458	14,811	14,793	42,880	NC
Flanking regions	11,148	559	8,308	4,574	24,589	NC
Intronic	2,755	112	2,001	958	5,826	NC
Synonymous	6,808	353	4,984	2,744	14,889	NC
Nonsynonymous	6,116	264	5,601	3,059	15,040	NC
Stop gain–lost	73–18	4–1	80–14	69–3	226–36	NC

Note.—NC, not calculated.

[a] Fixed polymorphisms are nucleotide sites, at which all Gs strains differ from all strains of Gc.

[b] Shared polymorphisms are sites for which multiple nucleotides are found in both Gs and Gc strains.

[c] Exclusive polymorphisms are those that are polymorphic in one species and invariant in the other.

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. 3*a*) 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*a*, ML tree).

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 Gc isolates from *P. ponderosa*, as well as to all Gs and the closely related species from *P. contorta*, which did grow (fig. 3*b*).

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_N, nonsynonymous polymorphism; D_N, nonsynonymous divergence) and synonymous (P_S, synonymous polymorphism; D_S, synonymous divergence) positions (fig. 4*a*–*c*). For the 3,746 Gs–Gc orthologs, we obtained a median $-\log_{10}$ NI (neutrality 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 ($P < 10^{-05}$) signal of purifying selection in the pooled analysis of all 3,476 genes, using an unbiased estimator of NI ($-\log_{10}$ $\mathrm{NI}_{TG} = -0.11$, pooled $P_N = 3{,}834$, $D_N = 5{,}903$, $P_S = 3{,}267$, $D_S = 6{,}612$). However, only five genes showed significant evidence of purifying selection on a per-gene basis. Although 1,215 genes showed $-\log_{10}$ NI > 0, indicating fewer amino acid polymorphisms within Gs relative to those between Gs and Gc, we only found 11 genes with statistically significant ($P \leq 0.05$) 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 $D_N \geq 9$; supplementary table S9, Supplementary Material online, and fig. 4*a*–*c*). Among genes exhibiting the strongest evidence for positive selection (i.e., $-\log_{10}$ 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 d*N*, d*S*, and d*N*/d*S* (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 d*S* $0.0032 \pm 4.8 \times 10^{-5}$, 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 d*N* ($0.0011 \pm 8.9 \times 10^{-19}$) was lower than d*S*, reflecting the expected stronger constraints on substitutions that changed amino acids. The overall mean for d*N*/d*S* in the 3,476

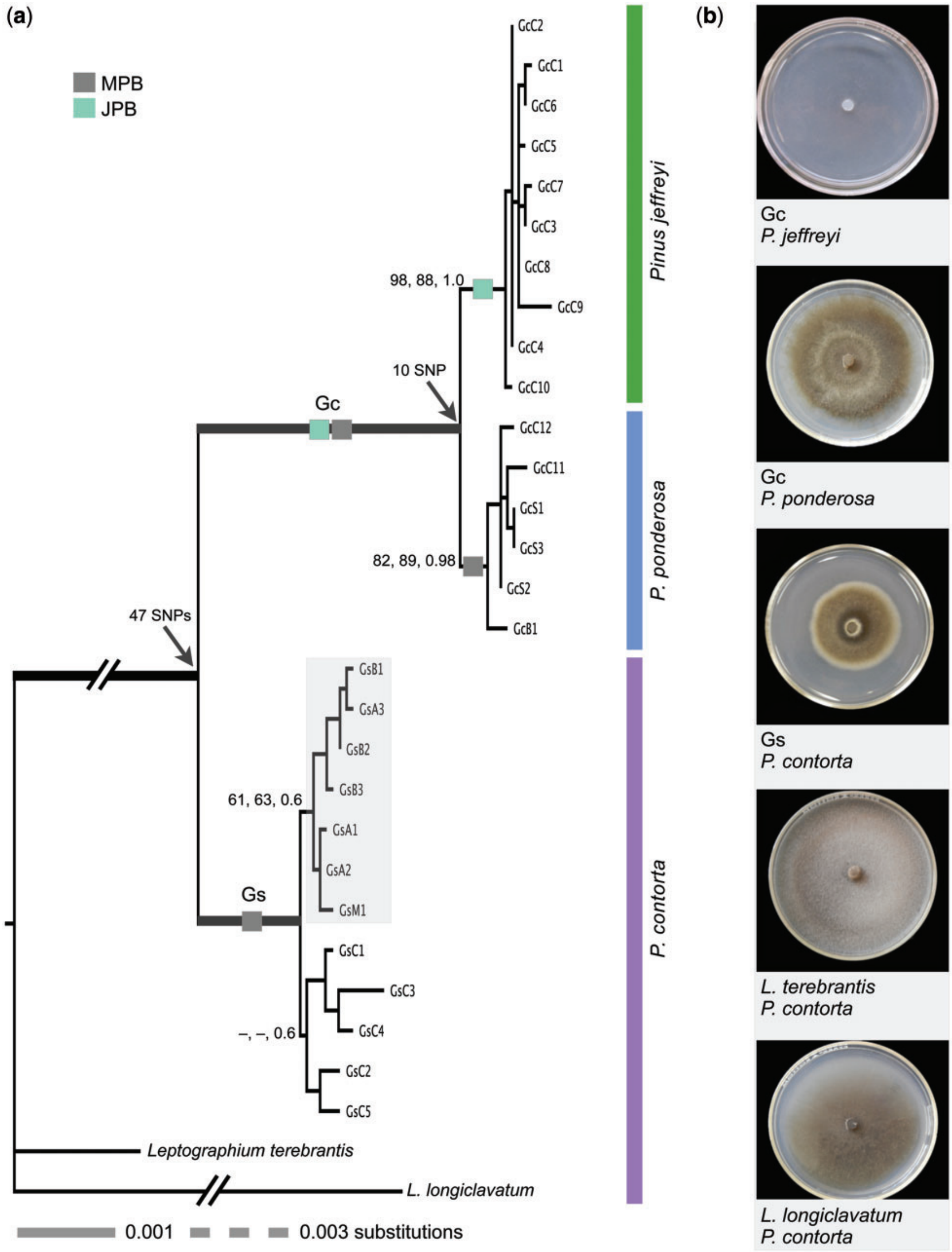


Fig. 3. *Grosmannia* species phylogeny correlated with the host-tree species of different phylogenetic lineages and fungal lineage tolerance to a related host-defence chemical. (*a*) ML analysis of a concatenated nine-gene data set subdivides Gs and Gc strains into three well-supported clades according to host tree species. Thick branches indicate nodes with 100% support from ML, MP, and Bayesian analyses. The gray box highlights Gs strains from epidemic regions. Arrows indicate total numbers of fixed differences between Gs–Gc, Gc–*Pinus ponderosa,* and Gc–*P. jeffreyi* lineages. The tree is rooted with the outgroup taxa *Leptographium longiclavatum* and *L. terebrantis.* Dashed line indicates an adjustment of scale. (*b*) Physiological assessment comparing (+)-limonene utilization as a carbon source in Gs and Gc lineages and close relatives. The growth of the fungal lineages from *P. contorta* and *P. ponderosa* indicates their ability to tolerate and utilize this toxic chemical.

orthologous genes (i.e., excluding 289 genes with dS = 0) was 0.3 ± 0.005. This value was similar to the $-\log_{10}$ $\mathrm{NI}_{\mathrm{TG}} = -0.11$ 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,

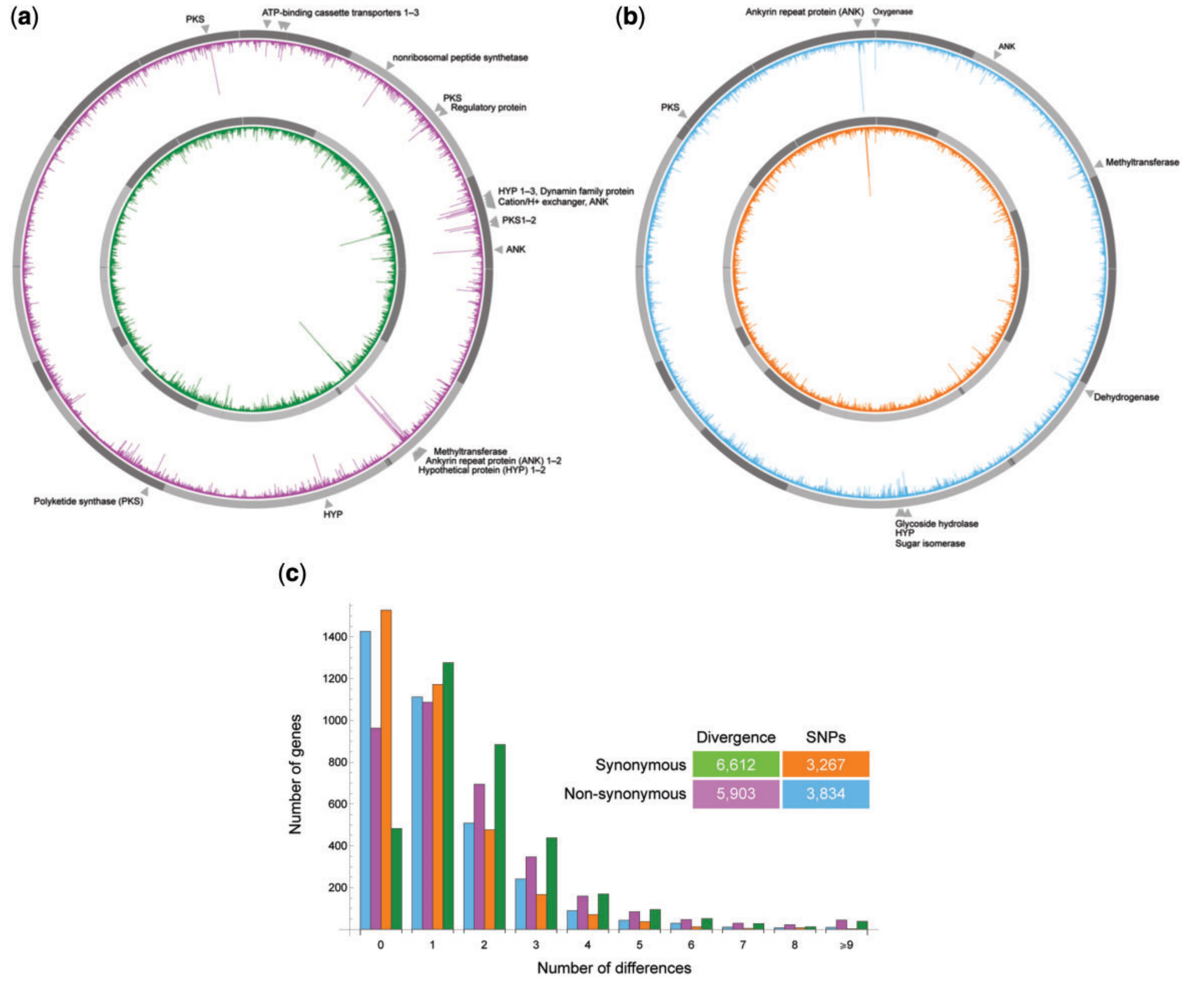


Fig. 4. Comparison of divergence and polymorphism. (*a*) Number of synonymous and nonsynonymous Gs–Gc fixed differences (divergences) in 7,340 orthologous gene models. (*b*) Number of synonymous and nonsynonymous SNPs within Gs strains in 7,340 orthologous gene models. Circular bands with alternating 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).

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 $-\log_{10}$ NI > 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 authophagy 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 ω less than 1. Nonetheless, after correction for multiple testing, we identified at least seven genes that evolved with ω 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

Table 4. The Top 39 Genes Showing Evidence of Positive Selection.[a]

Genes	Gene Description	MK[b]					PAML[c]				
		P_N	D_N	P_S	D_S	*P* Values	M1a.M2a	M7.M8	*P* Values	d*N*/d*S* (%)	dS
CMQ_6634	ABC transporter	1	10	3	1	*	NS	NS	NS	NA	0.002
CMQ_5949	Putative oxidoreductase	0	5	3	1	*	NS	NS	NS	NA	0.005
CMQ_742	Heterokaryon incompatibility	1	9	3	1	*	NS	NS	NS	NA	0.004
CMQ_3665	Thermotolerance protein	1	4	4	0	*	NS	NS	NS	NA	0.001
CMQ_3566	NRPS	0	21	4	13	*	NS	NS	NS	NA	0.003
CMQ_5323	PKS	1	17	4	5	*	NS	NS	NS	NA	0.003
CMQ_5095	PKS	1	20	4	9	*	21	26	**	55 (1.3)	0.005
CMQ_4392	PKS	0	29	6	8	**	5	5	*P* = 0.07	5.3 (NA)	0.006
CMQ_2677	PKS	3	8	2	5	0.09	26	26	**	157 (0.18)	0.007
CMQ_2687	PKS	3	7	2	2	NS	30	31	**	577 (0.24)	0.001
CMQ_1651	Ankyrin repeat protein	1	27	0	9	NS	27	23	**	35 (4.70)	0.010
CMQ_569	Ankyrin repeat protein	1	49	3	31	NS	12	13	**	6 (12.5)	0.061
CMQ_5699	Peroxin	1	5	1	2	NS	35	37	**	999 (0.08)	0.005
CMQ_7934	Heat repeat protein	2	4	3	2	NS	18	19	**	306 (0.20)	0.001
CMQ_8021	Heat shock transcription factor	1	3	4	0	NS	8	8	*	158 (0.60)	0.001
CMQ_6965	ABC transporter	1	10	3	3	NS	8	8	**	164 (0.60)	0.002
CMQ_1224	Membrane copper amine oxidase	0	4	2	2	NS	7	7	*	75 (0.70)	0.004
CMQ_277	Isoflavone reductase	1	3	4	2	NS	32	32	**	58 (0.70)	0.005
CMQ_5685	2-dehydropantoate 2-reductase	1	1	3	2	NS	10	10	**	453 (0.06)	0.003
CMQ_6107	Cytochrome p450 monooxygenase	0	5	1	2	NS	10	16	**	67 (1.15)	0.007
CMQ_3491	Cytochrome p450 monooxygenase	1	4	1	2	NS	7	7	*	61 (1.13)	0.006
CMQ_4067	Cytochrome p450 monooxygenase	0	5	1	4	NS	7	7	*	230 (0.06)	0.015
CMQ_1868	C2h2 finger domain containing protein	1	3	2	2	NS	7	7	*	81 (0.48)	0.010
CMQ_938	GTP cyclohydrolase	1	2	2	3	NS	7	7	*	80 (0.40)	0.011
CMQ_5377	Glycoside hydrolase	3	4	2	1	NS	7	7	*	113 (0.60)	0.003
CMQ_1846	C-terminal hydrolase	1	2	2	3	NS	7	7	*	252 (0.18)	0.003
CMQ_4864	Cell morphogenesis protein	1	7	3	4	NS	7	7	*	73 (0.70)	0.003
CMQ_7167	Autophagy protein	1	6	1	3	NS	8	8	*	152 (0.60)	0.002
CMQ_555	Death-box RNA helicase	2	4	2	1	NS	8	8	*	147 (0.97)	0.002
CMQ_3835	Dash complex subunit	0	5	1	2	NS	7	7	*	74 (0.80)	0.005
CMQ_6415	tRNA-guanine transglycosylase	1	2	2	2	NS	12	12	**	935 (0.03)	0.005
CMQ_906	Monocarboxylate permease	1	5	1	2	NS	7	6	*	72 (2.20)	0.005
CMQ_4470	Sensor histidine kinase response	0	8	1	9	ns	7	7	*	821 (0.01)	0.005
CMQ_2502	Phospholipase	0	2	1	2	NS	11	11	**	884 (0.03)	0.003
CMQ_3485	mtDNA inheritance protein	1	4	1	3	NS	8	8	*	156 (0.33)	0.007
CMQ_1204	Inositol polyphosphate phosphatase	2	2	3	2	NS	8	8	*	148 (0.56)	0.002
CMQ_4773	ATP-dependent DNA helicase	2	2	3	2	NS	7	7	*	170 (0.20)	0.004
CMQ_599	ATP-binding endoribonuclease	3	2	4	2	NS	14	14	**	103 (0.50)	0.006
CMQ_3121	60S ribosomal protein	1	6	1	3	NS	7	7	*	82 (0.80)	0.003

[a] List of genes that showed the strongest evidence of positive selection using MK test and PAML "site-model." Genes that were found significant for positive selection by both methods or those that after correction for multiple testing remained significant are highlighted. Results are only shown for genes with a putative function. NS, not significant; NA, not applicable.

[b] P_N, number of nonsynonymous polymorphisms within Gs; D_N, number of nonsynonymous differences between Gs and Gc; P_S, number of synonymous polymorphisms within Gs; D_S, number of synonymous differences between Gs and Gc; *P* value, significant excess of nonsynonymous divergence in MK test using the Fisher's exact test (*$P < 0.05$ or **$P < 0.01$).

[c] M1a–M2a and M7–M8, twice the difference in the natural logs of the likelihoods of the two models being compared. This value is used in a likelihood ratio test along with 2 degrees of freedom; *P* value is an uncorrected value from χ^2 distribution to indicate the confidence with which the null model can be rejected; d*N*/d*S*, nonsynonymous/synonymous substitution ratio ($\omega = dN/dS$) under M8 model of variable ω ratios among sites and the percent of codons placed in that class. Amino acid positions identified in the class of codons evolving under positive selection in M8 (posterior probability > 0.90) are listed in supplementary table S10, Supplementary Material online.

*$P < 0.05$.

**$P < 0.01$.

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.

Discussion

In this study, to identify features common across distinct *Grosmannia* populations and species, we compare the 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 Illumina 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, 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 *Coccidioides* and *Paracoccidioides*, and the generalist saprophyte *Neurospora crassa* and species in the genus *Aspergillus* (Lambreghts 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 (Rydholm 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 (Rydholm 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 gene 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 gene 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–Pj) 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; Stukenbrock 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 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 (Stukenbrock et al. 2010; Marcet-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. terebranis*) 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 imitations, 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 ($dN/dS = 0.3 \pm 0.005$), 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 multiple 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 methods 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 selection. 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] – $\log_{10}$ 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 adaptation 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 metabolite export (Kovalchuk and Driessen 2010). Other genes with putative functions in chemical detoxification or utilization included an oxidoreductase, an isoflavone reductase, and three 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 monocarboxylate permease). Other genes were potentially involved in cell signaling (e.g., histidine kinase and phospholipase), fungal development, and growth (e.g., membrane copper amine oxidase, cell morphogenesis, autophagy protein, heat-repeat protein, and hit finger domain protein) and a few with putative roles in protein–protein interactions or self-/nonself-recognition (e.g., two ankyrin repeat proteins and a heterokaryon 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 detoxification 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. longiclavatum* 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 conserved 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 ecology and population structure.

Materials and Methods

Fungal Samples

We sequenced eight Gs genomes from two distinct populations 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 sibling 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 physiological 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 l distilled water) overlaid with cellophane. DNA from the mycelia 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 library 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 manufacturer 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 (slkw1407; NCBI Genome PID: 39837; DiGuistini et al. 2009). Contigs that did not align with slkw1407 were ignored for our analysis. The assembly statistics after ordering and orientation are shown in supplementary table S2, Supplementary Material online.

Gene Predictions and Ortholog Determination

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 either 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

We performed variant calling among the *Grosmannia* genomes by mapping reads from each strain to the slkw1407 reference genome sequence. Before read mapping, we filtered raw reads to remove low-quality and duplicate sequences using PRINSEQ lite v0.17.1 (Schmieder and Edwards 2011). We discarded reads that failed the Illumina chastity filter, contained uncalled bases, and had an average Phred-scaled quality of less than ten in the last 20 base calls. For the retained reads, the initial (5′) five nucleotides showed GC-content bias (data not shown) and were trimmed, leaving 45 and 71 nt reads for mapping. We also filtered potential duplicate reads resulting from amplification of the identical DNA fragments during library preparation and sequencing. The numbers of reads used for SNV calling and processing steps are shown in supplementary table S4, Supplementary Material online.

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 BEDtools 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 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 matrixes 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. Monophylies supported by both BS $\geq$ 70% and PP $\geq$ 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 monoterpene.

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 ($D_N + D_S = 0$; $n = 487$). 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 (P_N:P_S) or divergent (D_N:D_S), 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 (NITG; 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 (d*N*) and synonymous (d*S*) sites for each gene, by setting parameters as follow: 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 d*S* values were determined by plotting d*N* as a function of d*S*, 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 ($D_N + D_S \geq 3$) and/or synonymous ($D_S + P_S \geq 3$) 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 d*N*/d*S* for a site to be less than, equal to, or greater than 1. If d*N*/d*S* 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 $0 \leq dN/dS \leq 1$ 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 χ^2 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

Aguileta G, Lengelle J, Marthey S, Chiapello H, Rodolphe F, Gendrault A, Yockteng R, Vercken E, Devier B, Fontaine MC, et al. 2009. Finding candidate genes under positive selection in nonmodel species: examples of genes involved in host specialization in pathogens. *Mol Ecol.* 19:292–306.

Amselem J, Cuomo CA, van Kan JAL, Viaud M, Benito EP, Couloux A, Coutinho PM, de Vries RP, Dyer PS, Fillinger S, et al. 2011. Genomic analysis of the necrotrophic fungal pathogens *Sclerotinia sclerotiorum* and *Botrytis cinerea*. *PLoS Genet.* 7:e1002230.

Andersen MR, Salazar MP, Schaap PJ, van de Vondervoort PJI, Culley D, Thykaer J, Frisvad JC, Nielsen KF, Albang R, Albermann K, et al. 2011. Comparative genomics of citric-acid-producing *Aspergillus niger* ATCC 1015 versus enzyme-producing CBS 513.88. *Genome Res.* 21:885–897.

Avise JC. 2004. Molecular markers, natural history, and evolution, 2nd revised ed. Sunderland (MA): Sinauer Associates.

Bahn Y-S, Kojima K, Cox GM, Heitman J. 2006. A unique fungal two-component system regulates stress responses, drug sensitivity, sexual development, and virulence of *Cryptococcus neoformans*. *Mol Biol Cell.* 17:3122–3135.

Barrett LG, Thrall PH, Burdon JJ, Linde CC. 2008. Life history determines genetic structure and evolutionary potential of host-parasite interactions. *Trends Ecol Evol.* 23:678–685.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Stat Methodol.* 57:289–300.

Boone CK, Aukema BH, Bohlmann J, Carroll AL, Raffa KF. 2011. Efficacy of tree defense physiology varies with bark beetle population density: a basis for positive feedback in eruptive species. *Can J For Res.* 41: 1174–1188.

Carroll A, Aukema B, Raffa K, Linton DA, Smith G, Lindgren B. 2006. Mountain pine beetle outbreak development: the endemic—incipient epidemic transition. Victoria (BC): Canadian Forest Service.

Clark EL, Carroll AL, Huber DPW. 2010. Differences in the constitutive terpene profile of lodgepole pine across a geographical range in British Columbia, and correlation with historical attack by mountain pine beetle. *Can Entomol.* 142:557–573.

Clement M, Posada D, Crandall KA. 2000. TCS: a computer program to estimate gene genealogies. *Mol Ecol.* 9:1657–1659.

Collemare J, Billard A, Böhnert HU, Lebrun M-H. 2008. Biosynthesis of secondary metabolites in the rice blast fungus *Magnaporthe grisea*: the role of hybrid PKS-NRPS in pathogenicity. *Mycol Res.* 112: 207–215.

Condon BJ, Leng Y, Wu D, Bushley KE, Ohm RA, Otillar R, Martin J, Schackwitz W, Grimwood J, MohdZainudin N, et al. 2013. Comparative genome structure, secondary metabolite, and effector coding capacity across *Cochliobolus* pathogens. *PLoS Genet.* 9: e1003233.

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21: 3674–3676.

Cullingham CI, Cooke JEK, Dang S, Davis CS, Cooke BJ, Coltman DW. 2011. Mountain pine beetle host-range expansion threatens the boreal forest. *Mol Ecol.* 20:2157–2171.

Cuomo CA, Güldener U, Xu J-R, Trail F, Turgeon BG, Pietro AD, Walton JD, Ma L-J, Baker SE, Rep M, et al. 2007. The *Fusarium graminearum* genome reveals a link between localized polymorphism and pathogen specialization. *Science* 317:1400–1402.

Darling AE, Mau B, Perna NT. 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* 5: e11147.

de Queiroz K. 2007. Species concepts and species delimitation. *Syst Biol.* 56:879–886.

de Wit PJGM, van der Burgt A, Ökmen B, Stergiopoulos I, Abd-Elsalam KA, Aerts AL, Bahkali AH, Beenen HG, Chettri P, Cox MP, et al. 2012. The genomes of the fungal plant pathogens *Cladosporium fulvum* and *Dothistroma septosporum* reveal adaptation to different hosts and lifestyles but also signatures of common ancestry. *PLoS Genet.* 8: e1003088.

Desjardins CA, Champion MD, Holder JW, Muszewska A, Goldberg J, Bailão AM, Brigido MM, Ferreira ME, Garcia AM, Grynberg M, et al. 2011. Comparative genomic analysis of human fungal pathogens causing Paracoccidioidomycosis. *PLoS Genet.* 7:e1002345.

Dettman JR, Jacobson DJ, Taylor JW. 2003. A multilocus genealogical approach to phylogenetic species recognition in the model eukaryote *Neurospora*. *Evolution* 57:2703–2720.

DiGuistini S, Liao NY, Platt D, Robertson G, Seidel M, Chan SK, Docking TR, Birol I, Holt RA, Hirst M, et al. 2009. De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. *Genome Biol.* 10:R94.

DiGuistini S, Wang Y, Liao NY, Taylor G, Tanguay P, Feau N, Henrissat B, Chan SK, Hesse-Orce U, Alamouti SM, et al. 2011. Genome and transcriptome analyses of the mountain pine beetle–fungal symbiont *Grosmannia clavigera*, a lodgepole pine pathogen. *Proc Natl Acad Sci U S A.* 108:2504–2509.

Farrell B, Sequeira A, O'Meara B, Normark B, Chung J, Jordal B. 2001. The evolution of agriculture in beetles (Curculionidae: Scolytinae and Platypodinae). *Evolution* 55:2011–2027.

Fay JC, Wyckoff GJ, Wu CI. 2001. Positive and negative selection on the human genome. *Genetics* 158:1227–1234.

Fedorova ND, Badger JH, Robson GD, Wortman JR, Nierman WC. 2005. Comparative analysis of programmed cell death pathways in filamentous fungi. *BMC Genomics* 6:177.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39:783–791.

Gao X, Starmer JD. 2008. AWclust: point-and-click software for non-parametric population structure analysis. *BMC Bioinformatics* 9:77.

Gerson EA, Kelsey RG, St Clair JB. 2009. Genetic variation of piperidine alkaloids in *Pinus ponderosa*: a common garden study. *Ann Bot.* 103: 447–457.

Giraud T, Gladieux P, Gavrilets S. 2010. Linking the emergence of fungal plant diseases with ecological speciation. *Trends Ecol Evol.* 25: 387–395.

Giraud T, Refrégier G, Le Gac M, de Vienne DM, Hood ME. 2008. Speciation in fungi. *Fungal Genet Biol.* 45:791–802.

Giraud T, Villaréal LM, Austerlitz F, Le Gac M, Lavigne C. 2006. Importance of the life cycle in sympatric host race formation and speciation of pathogens. *Phytopathology* 96:280–287.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol.* 11: 725–736.

Gu Z, David L, Petrov D, Jones T, Davis RW, Steinmetz LM. 2005. Elevated evolutionary rates in the laboratory strain of *Saccharomyces cerevisiae*. *Proc Natl Acad Sci U S A.* 102: 1092–1097.

Hall DE, Yuen MMS, Jancsik S, Quesada AL, Dullat HK, Li M, Henderson H, Arango-Velez A, Liao NY, Docking RT, et al. 2013. Transcriptome resources and functional characterization of monoterpene synthases for two host species of the mountain pine beetle, lodgepole pine (*Pinus contorta*) and jack pine (*Pinus banksiana*). *BMC Plant Biol.* 13:80.

Hall DE, Zerbe P, Jancsik S, Quesada AL, Dullat H, Madilao LL, Yuen M, Bohlmann J. 2013. Evolution of conifer diterpene synthases: diterpene resin acid biosynthesis in lodgepole pine and jack pine involves monofunctional and bifunctional diterpene synthases. *Plant Physiol.* 161:600–616.

Hane JK, Lowe RGT, Solomon PS, Tan K-C, Schoch CL, Spatafora JW, Crous PW, Kodira C, Birren BW, Galagan JE, et al. 2007. Dothideomycete–plant interactions illuminated by genome sequencing and EST analysis of the wheat pathogen *Stagonospora nodorum*. *Plant Cell* 19:3347–3368.

Hane JK, Rouxel T, Howlett BJ, Kema GH, Goodwin SB, Oliver RP. 2011. A novel mode of chromosomal evolution peculiar to filamentous Ascomycete fungi. *Genome Biol.* 12:R45.

Haridas S, Breuill C, Bohlmann J, Hsiang T. 2011. A biologist's guide to de novo genome assembly using next-generation sequence data: a test with fungal genomes. *J Microbiol Methods.* 86:368–375.

Harrington T. 2005. Ecology and evolution of mycophagous bark beetles and their fungal partners. In: Vega F, Blackwell M, editors. Ecological and evolutionary advances in insect–fungal associations. New York: Oxford University Press. p. 257–291.

Hesse-Orce U, DiGuistini S, Keeling CI, Wang Y, Li M, Henderson H, Docking TR, Liao NY, Robertson G, Holt RA, et al. 2010. Gene discovery for the bark beetle-vectored fungal tree pathogen *Grosmannia clavigera*. *BMC Genomics* 11:536.

Holloway AK, Lawniczak MKN, Mezey JG, Begun DJ, Jones CD. 2007. Adaptive gene expression divergence inferred from population genomics. *PLoS Genet.* 3:e187.

Jordal BH. 2013. Deep phylogenetic divergence between Scolytoplatypus and Remansus, a new genus of Scolytoplatypodini from Madagascar (Coleoptera, Curculionidae, Scolytinae). *Zookeys* 352:9–33.

Jordal BH, Cognato AI. 2012. Molecular phylogeny of bark and ambrosia beetles reveals multiple origins of fungus farming during periods of global warming. *BMC Evol Biol.* 12:133.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30:3059–66.

Keeling CI, Bohlmann J. 2006. Genes, enzymes and chemicals of terpenoid diversity in the constitutive and induced defence of conifers against insects and pathogens. *New Phytol.* 170:657–675.

Klosterman SJ, Subbarao KV, Kang S, Veronese P, Gold SE, Thomma BPHJ, Chen Z, Henrissat B, Lee Y-H, Park J, et al. 2011. Comparative genomics yields insights into niche adaptation of plant vascular wilt pathogens. *PLoS Pathog.* 7:e1002137.

Kohler GA, Brenot A, Haas-Stapleton E, Agabian N, Deva R, Nigam S. 2006. Phospholipase A2 and phospholipase B activities in Fungi. *Biochim Biophys Acta.* 1761:1391–1399.

Kovalchuk A, Driessen A. 2010. Phylogenetic analysis of fungal ABC transporters. *BMC Genomics* 11:177.

Kroken S, Glass NL, Taylor JW, Yoder OC, Turgeon BG. 2003. Phylogenomic analysis of type I polyketide synthase genes in pathogenic and saprobic ascomycetes. *Proc Natl Acad Sci U S A.* 100: 15670–15675.

Kulathinal RJ, Stevison LS, Noor MAF. 2009. The Genomics of speciation in *Drosophila*: diversity, divergence, and introgression estimated using low-coverage genome sequencing. *PLoS Genet.* 5:e1000550.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. *Genome Biol.* 5:R12.

Kurz WA, Dymond CC, Stinson G, Rampley GJ, Neilson ET, Carroll AL, Ebata T, Safranyik L. 2008. Mountain pine beetle and forest carbon feedback to climate change. *Nature* 452:987–990.

Lah L, Haridas S, Bohlmann J, Breuil C. 2013. The cytochromes P450 of *Grosmannia clavigera*: genome organization, phylogeny, and expression in response to pine host chemicals. *Fungal Genet Biol.* 50: 72–81.

Lambreghts R, Shi M, Belden WJ, Decaprio D, Park D, Henn MR, Galagan JE, Bastürkmen M, Birren BW, Sachs MS, et al. 2009. A high-density single nucleotide polymorphism map for *Neurospora crassa*. *Genetics* 181:767–781.

Lee S, Hamelin RC, Six DL, Breuil C. 2007. Genetic diversity and the presence of two distinct groups in *Ophiostoma clavigerum* associated with *Dendroctonus ponderosae* in British Columbia and the northern Rocky Mountains. *Phytopathology* 97: 1177–1185.

Lee S, Kim J-J, Breuil C. 2006. Diversity of fungi associated with mountain pine beetle, *Dendroctonus ponderosae*, and infested lodgepole pines in British Columbia. *Fungal Divers.* 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, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* 25:2078–2079.

Li H, Ruan J, Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* 18: 1851–1858.

Li YF, Costello JC, Holloway AK, Hahn MW. 2008. "Reverse ecology" and the power of population genomics. *Evolution* 62:2984–2994.

Liu M, Gelli A. 2008. Elongation factor 3, EF3, associates with the calcium channel Cch1 and targets Cch1 to the plasma membrane in *Cryptococcus neoformans*. *Eukaryot Cell.* 7:1118–1126.

Luhtala N. 2004. Bro1 coordinates deubiquitination in the multivesicular body pathway by recruiting Doa4 to endosomes. *J Cell Biol.* 166: 717–729.

Ma LJ, van der Does HC, Borkovich KA, Coleman JJ, Daboussi M-J, Di Pietro A, Dufresne M, Freitag M, Grabherr M, Henrissat B, et al. 2010. Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. *Nature* 464:367–373.

Maddison WP, Knowles LL. 2006. Inferring phylogeny despite incomplete lineage sorting. *Syst Biol.* 55:21–30.

Manning VA, Pandelova I, Dhillon B, Wilhelm LJ, Goodwin SB, Berlin AM, Figueroa M, Freitag M, Hane JK, Henrissat B, et al. 2013. Comparative genomics of a plant-pathogenic fungus, *Pyrenophora tritici-repentis*, reveals transduplication and the impact of repeat elements on pathogenicity and population divergence. *G3* 3:41–63.

Marcet-Houben M, Ballester A-R, de la Fuente B, Harries E, Marcos JF, González-Candelas L, Gabaldón T. 2012. Genome sequence of the necrotrophic fungus *Penicillium digitatum*, the main postharvest pathogen of citrus. *BMC Genomics* 13:646.

Marin M, Preisig O, Wingfield BD, Kirisits T, Wingfield MJ. 2009. Single sequence repeat markers reflect diversity and geographic barriers in Eurasian populations of the conifer pathogen *Ceratocystis polonica*. *Forest Pathol.* 39:249–265.

Marshall DH, Newton C, Ritland K. 2002. Chloroplast phylogeography and evolution of highly polymorphic microsatellites in lodgepole pine (*Pinus contorta*). *Theor Appl Genet.* 104:367–378.

Massoumi Alamouti S, Wang V, Diguistini S, Six DL, Bohlmann J, Hamelin RC, Feau N, Breuil C. 2011. Gene genealogies reveal cryptic species and host preferences for the pine fungal pathogen *Grosmannia clavigera*. *Mol Ecol.* 20:2581–2602.

McCluskey K, Wiest AE, Grigoriev IV, Lipzen A, Martin J, Schackwitz W, Baker SE. 2011. Rediscovery by whole genome sequencing: classical mutations and genome polymorphisms in *Neurospora crassa*. *G3* 1: 303–316.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* 351:652–654.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: a map reduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20: 1297–1303.

Mirov NT, Hasbrouck J. 1976. The story of pines, 1st ed. Bloomington (IN): Indiana University Press.

Mock KE, Bentz BJ, O'neill EM, Chong JP, Orwin J, Pfrender ME. 2007. Landscape-scale genetic variation in a forest outbreak species, the mountain pine beetle (*Dendroctonus ponderosae*). *Mol Ecol.* 16: 553–68.

Möller EM, Bahnweg G, Sandermann H, Geiger HH. 1992. A simple and efficient protocol for isolation of high molecular weight DNA from filamentous fungi, fruit bodies, and infected plant tissues. *Nucleic Acids Res.* 20:6115–6116.

Moran GP, Coleman DC, Sullivan DJ. 2011. Comparative genomics and the evolution of pathogenicity in human pathogenic fungi. *Eukaryot Cell.* 10:34–42.

Morin PA, Luikart G, Wayne RK, the SNP workshop group. 2004. SNPs in ecology, evolution and conservation. *Trends Ecol Evol.* 19: 208–216.

Neafsey DE, Barker BM, Sharpton TJ, Stajich JE, Park DJ, Whiston E, Hung C-Y, McMahan C, White J, Sykes S, et al. 2010. Population genomic sequencing of Coccidioides fungi reveals recent hybridization and transposon control. *Genome Res.* 20:938–946.

Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ, et al. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* 3:e170.

Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148:929–936.

Nylander JAA, Wilgenbusch JC, Warren DL, Swofford DL. 2008. AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. *Bioinformatics* 24: 581–583.

Oleksyk TK, Smith MW, O'Brien SJ. 2010. Genome-wide scans for footprints of natural selection. *Philos Trans R Soc Lond B Biol Sci.* 365: 185–205.

Paine TD, Hanlon CC. 1994. Influence of oleoresin constituents from *Pinus ponderosa* and *Pinus jeffreyi* on growth of mycangial fungi from *Dendroctonus ponderosae* and *Dendroctonus jeffreyi*. *J Chem Ecol.* 20:2551–2563.

Pollack JK, Harris SD, Marten MR. 2009. Autophagy in filamentous fungi. *Fungal Genet Biol.* 46:1–8.

Posada D. 2008. jModelTest: phylogenetic model averaging. *Mol Biol Evol.* 25:1253–1256.

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

Raffa K. 2001. Mixed messages across multiple trophic levels: the ecology of bark beetle chemical communication systems. *Chemoecology* 11: 49–65.

Raffa KF, Aukema B, Erbilgin N, Klepzig KD, Wallin KF. 2005. Recent advances in phytochemistry. Toronto (Canada): Elsevier.

Raffaele S, Kamoun S. 2012. Genome evolution in filamentous plant pathogens: why bigger can be better. *Nat Rev Microbiol.* 10: 417–430.

Rambaut A, Drummond AJ. 2009. Tracer v1.5 [Internet]. [cited 2014 Mar 20]. Available from: http://tree.bio.ed.ac.uk/software/tracer/.

Rand DM, Kann LM. 1996. Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from *Drosophila*, mice, and humans. *Mol Biol Evol.* 13:735–748.

Reumers J, Maurer-Stroh S, Schymkowitz J, Rousseau F. 2006. SNPeffect v2.0: a new step in investigating the molecular phenotypic effects of human non-synonymous SNPs. *Bioinformatics* 22: 2183–2185.

Robinson-Jeffrey RC, Davidson RW. 1968. Three new *Europhium* species with *Verticicladiella* imperfect states on blue-stained pine. *Can J Bot.* 46:1523–1527.

Roe AD, Rice AV, Coltman DW, Cooke JEK, Sperling FAH. 2011. Comparative phylogeography, genetic differentiation and contrasting reproductive modes in three fungal symbionts of a multipartite bark beetle symbiosis. *Mol Ecol.* 20:584–600.

Rokas A. 2009. The effect of domestication on the fungal proteome. *Trends Genet.* 25:60–63.

Rokas A, Payne G, Fedorova ND, Baker SE, Machida M, Yu J, Georgianna DR, Dean RA, Bhatnagar D, Cleveland TE, et al. 2007. What can comparative genomics tell us about species concepts in the genus *Aspergillus*? *Stud Mycol.* 59:11–17.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572–1574.

Rydholm C, Szakacs G, Lutzoni F. 2006. Low genetic variation and no detectable population structure in *Aspergillus fumigatus* compared to closely related *Neosartorya* species. *Eukaryot Cell.* 5: 650–657.

Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27:863–864.

Seybold SJ, Bohlmann J, Raffa KF. 2000. Biosynthesis of coniferophagous bark beetle pheromones and conifer isoprenoids: evolutionary perspective and synthesis. *Can Entomol.* 132:697–753.

Sharpton TJ, Stajich JE, Rounsley SD, Gardner MJ, Wortman JR, Jordar VS, Maiti R, Kodira CD, Neafsey DE, Zeng Q, et al. 2009. Comparative genomic analyses of the human fungal pathogens *Coccidioides* and their relatives. *Genome Res.* 19:1722–1731.

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

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I. 2009. ABySS: a parallel assembler for short read sequence data. *Genome Res.* 19:1117–1123.

Six DL, Paine TD. 1997. *Ophiostoma clavigerum* is the mycangial fungus of the jeffrey pine beetle, *Dendroctonus jeffreyi*. *Mycologia* 89: 858–866.

Six DL, Paine TD. 1999. Allozyme diversity and gene flow in *Ophiostoma clavigerum* (Ophiostomatales: Ophiostomataceae), the mycangial fungus of the Jeffrey pine beetle, *Dendroctonus jeffreyi*. *Can J For Res.* 29:324.

Smith GD, Carroll AL, Lindgren BS. 2010. Facilitation in bark beetles: endemic mountain pine beetle gets a helping hand. *Agric For Entomol.* 13:37–43.

Smith RH. 2000. Xylem monoterpenes of pines: distribution, variation, genetics, function. Berkeley (CA): USDA.

Soanes DM, Alam I, Cornell M, Wong HM, Hedeler C, Paton NW, Rattray M, Hubbard SJ, Oliver SG, Talbot NJ. 2008. Comparative genome analysis of filamentous fungi reveals gene family expansions associated with fungal pathogenesis. *PLoS One* 3:e2300.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.

Stoletzki N, Eyre-Walker A. 2010. The positive correlation between dN/dS and dS in mammals is due to runs of adjacent substitutions. *Mol Biol Evol.* 28:1371–1380.

Stoletzki N, Eyre-Walker A. 2011. Estimation of the neutrality index. *Mol Biol Evol.* 28:63–70.

Stukenbrock EH, Bataillon T, Dutheil JY, Hansen TT, Li R, Zala M, McDonald BA, Wang J, Schierup MH. 2011. The making of a new pathogen: insights from comparative population genomics of the domesticated wheat pathogen *Mycosphaerella graminicola* and its wild sister species. *Genome Res.* 21:2157–2166.

Stukenbrock EH, Jørgensen FG, Zala M, Hansen TT, McDonald BA, Schierup MH. 2010. Whole-genome and chromosome evolution associated with host adaptation and speciation of the wheat pathogen *Mycosphaerella graminicola*. *PLoS Genet.* 6: e1001189.

Stukenbrock EH, McDonald BA. 2008. The origins of plant pathogens in agro-ecosystems. *Annu Rev Phytopathol.* 46:75–100.

Sturgeon KB. 1979. Monoterpene variation in ponderosa pine xylem resin related to western pine beetle predation. *Evolution* 33: 803.

Swofford DL. 2003. PAUP*, phylogenetic analysis using parsimony (*and other methods). Sunderland (MA): Sinauer Associates.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol.* 28:2731–2739.

Thompson JN. 1994. The coevolutionary process. Chicago (IL): University of Chicago Press.

Tsui CK-M, DiGuistini S, Wang Y, Feau N, Dhillon B, Bohlmann J, Hamelin RC. 2013. Unequal recombination and evolution of the mating-type (MAT) loci in the pathogenic fungus *Grosmannia clavigera* and relatives. *G3* 3:465–480.

Tsui CKM, Roe AD, El-Kassaby YA, Rice AV, Alamouti SM, Sperling F A. H, Cooke JEK, Bohlmann J, Hamelin RC. 2012. Population structure and migration pattern of a conifer pathogen, *Grosmannia clavigera*, as influenced by its symbiont, the mountain pine beetle. *Mol Ecol.* 21:71–86.

Tyler BM, Tripathy S, Zhang X, Dehal P, Jiang RHY, Aerts A, Arredondo FD, Baxter L, Bensasson D, Beynon JL, et al. 2006. *Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science* 313:1261–1266.

Wang W, Wei Z, Lam T-W, Wang J. 2011. Next generation sequencing has lower sequence coverage and poorer SNP-detection capability in the regulatory regions. *Sci Rep.* 1:55.

Wang Y, Lim L, DiGuistini S, Robertson G, Bohlmann J, Breuil C. 2013. A specialized ABC efflux transporter *GcABC-G1* confers monoterpene resistance to *Grosmannia clavigera*, a bark beetle-associated fungal pathogen of pine trees. *New Phytol.* 197:886–898.

Wingfield MJ, Seifert KA, Webber JF. editors. 1993. *Ceratocystis* and *Ophiostoma*: taxonomy, ecology, and pathogenicity. Minnesota: American Phytopathological Society Press. p. 1–293.

Wood S. 1982. The bark and ambrosia beetles of North and Central America (Coleoptera, Scolytidae): a taxonomic monograph. *Great Basin Nat Memoirs.* 6:1–1359.

Wright SI, Andolfatto P. 2008. The impact of natural selection on the genome: emerging patterns in *Drosophila* and *Arabidopsis*. *Annu Rev Ecol Evol Syst.* 39:193–213.

Xue M, Yang J, Li Z, Hu S, Yao N, Dean RA, Zhao W, Shen M, Zhang H, Li C, et al. 2012. Comparative analysis of the genomes of two field isolates of the rice blast fungus *Magnaporthe oryzae*. *PLoS Genet.* 8: e1002869.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.

Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol.* 17:32–43.

Yang Z, Swanson WJ. 2002. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Mol Biol Evol.* 19: 49–57.