Genome of the Asian longhorned beetle (*Anoplophora glabripennis*), a globally significant invasive species, reveals key functional and evolutionary innovations at the beetle-plant interface

Duane D. McKenna[1,2]*, Erin D. Scully[3], Yannick Pauchet[4], Kelli Hoover[5], Roy Kirsch[4], Scott M. Geib[6], Robert F. Mitchell[7,8], Robert M. Waterhouse[9,10], Seung-Joon Ahn[4], Deanna Arsala[11], Joshua B. Benoit[12], Heath Blackmon[13], Tiffany Bledsoe[7], Julia H. Bowsher[14], André Busch[4], Bernarda Calla[6], Hsu Chao[15], Anna K. Childers[16], Christopher Childers[17], Dave J. Clarke[1], Lorna Cohen[11], Jeffery P. Demuth[13], Huyen Dinh[15], HarshaVardhan Doddapaneni[15], Amanda Dolan[18], Jian J. Duan[19], Shannon Dugan[15], Markus Friedrich[20], Karl M. Glastad[21], Michael A. D. Goodisman[21], Stephanie Haddad[1], Yi Han[15], Daniel S. T. Hughes[15], Panagiotis Ioannidis[9], J. Spencer Johnston[22], Jeffery W. Jones[20], Leslie A. Kuhn[23], David R. Lance[24], Chien-Yueh Lee[17,25], Sandra L. Lee[15], Han Lin[17,25], Jeremy A. Lynch[11], Armin P. Moczek[26], Shwetha C. Murali[15], Donna M. Muzny[15], David R. Nelson[27], Subba R. Palli[28], Kristen A. Panfilio[29], Dan Pers[11], Monica F. Poelchau[17], Honghu Quan[11], Jiaxin Qu[15], Ann M. Ray[30], Joseph P. Rinehart[16], Hugh M. Robertson[31], Richard Roehrdanz[16], Andrew J. Rosendale[12], Seunggwan Shin[1], Christian Silva[18], Alex C. Torson[14], Iris M. Vargas Jentzsch[29], John H. Werren[18], Kim C. Worley[15], George Yocum[16], Evgeny M. Zdobnov[9], Richard A. Gibbs[15], and Stephen Richards[15]*

*Corresponding authors

Duane D. McKenna, Department of Biological Sciences, University of Memphis, 3700 Walker Ave., Memphis, TN 38152 USA. Email: dmckenna@memphis.edu

Stephen Richards, Human Genome Sequencing Center, Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030 USA. Email: stephenr@bcm.edu.

Abstract

Background: The beetle family Cerambycidae Latreille (longhorned beetles; >35,000 species) is the most diverse radiation of wood-feeding animals on Earth. However, relatively little is known about the genomic basis of wood-feeding (xylophagy) in beetles. We undertook genome and transcriptome sequencing and annotation, gene expression assays, studies of plant cell wall degrading enzyme substrate specificity, and other functional and comparative genomic studies of the Asian longhorned beetle, *Anoplophora glabripennis*, a globally significant invasive insect species capable of inflicting severe feeding damage on many important orchard, ornamental and forest tree species. Complementary comparative studies of genes encoding key enzymes involved in the digestion of woody plant tissues or the detoxification of plant allelochemicals were undertaken with the genomes of the Asian longhorned beetle and 14 additional insects, including the newly sequenced emerald ash borer beetle (*Agrilus planipennis*) and bull-headed dung beetle (*Onthophagus taurus*) genomes, both of which were studied for the first time.

Results: The Asian longhorned beetle genome encodes a uniquely diverse arsenal of enzymes that possess the ability to degrade the main polysaccharide networks in plant cell walls, detoxify plant allelochemicals, and otherwise facilitate specialized feeding on woody plants. The Asian longhorned beetle has the metabolic plasticity needed to feed on plant species with different chemistries, permitting colonization of a diverse range of host plants, and contributing to its highly invasive nature. Its metabolic capacity is further expanded through affiliations with gut microbes. Large expansions of chemosensory genes involved in the reception of pheromones and plant kairomones are consistent with the complexity of chemical cues used by the Asian longhorned beetle to find host plants and mates.

Conclusions: Our studies reveal that amplification and functional divergence of genes associated with specialized feeding on plants, including genes previously shown to have been originally obtained by beetles via horizontal gene transfer from fungi and bacteria, were fundamental to the addition, expansion and enhancement of the metabolic repertoire of the Asian longhorned beetle, certain other beetles, and to a lesser degree, other phytophagous insects. Our results thus begin to establish a genomic basis for the evolutionary success of insects – especially beetles – on plants.

Keywords: Chemoperception, Detoxification, Glycoside hydrolase, Horizontal gene transfer, Phytophagy, Xylophagy

Background

Beetles (order Coleoptera; >400,000 described extant species) account for more than 20% of metazoans. The causes of this apparent “inordinate fondness” [1] are widely debated, but the evolution of specialized trophic interactions with plants – such as wood-feeding (xylophagy) – is assumed to have played an important role [2, 3]. The beetle family Cerambycidae Latreille (>35,000 species; longhorned beetles) is the most diverse radiation of wood-feeding animals on Earth. Most species complete their entire development while feeding exclusively on the tissues of woody plants. Recent work has established the Asian longhorned beetle (*Anoplophora glabripennis*) as a model for studies of the digestive physiology of wood-feeding beetles (see references herein). *A. glabripennis* is a globally significant invasive species, capable of inflicting severe damage on many economically-important orchard, ornamental and forest trees (>100 species) [4]. Its potential economic impact in the United States alone, if uncontrolled, has been conservatively estimated at $889 billion (adjusted for inflation, May 2016) [5]. Early stage *A. glabripennis* larvae are specialized wood-borers, feeding in galleries under bark in the subcortical tissue and phloem of both healthy and susceptible living trees (Fig. 1). Larger, later stage larvae tunnel deep into the heartwood, where they continue feeding and complete development. Adults are comparatively short-lived external feeders, consuming small amounts of tissue from host tree leaves and twigs [4].

Nitrogen, free amino acids, and protein are typically scarce in wood, and access to sugars, minerals, and other key nutrients is severely impeded by lignified plant cell walls. Furthermore, woody plant tissues contain a diversity of allelochemicals that must be detoxified or sequestered when eaten [6]. Successful feeding on woody plants therefore requires specialized metabolic adaptations. The genomes of *A. glabripennis* and certain other phytophagous beetles are known to contain genes encoding plant cell wall degrading enzymes (PCWDEs) [7-9]. PCWDEs degrade cellulose, hemicellulose or pectin (the main polysaccharide networks in plant cell walls), liberating sugars, minerals and other nutrients from woody plant tissues. Some

cerambycid PCWDEs were originally obtained via HGT from fungi or bacteria, and have subsequently diversified to form multi-gene families [10]. This is in contrast to other wood feeding insects, e.g., termites and some ants and cockroaches, which have broadly similar metabolic capabilities conveyed by symbionts whose genomes contain many of the same families of genes [11]. Additionally, lignin is degraded during passage through the *A. glabripennis* gut [12], suggesting a role for enzymes secreted into the gut by the beetle, its gut microbiota, or both parties. *In vitro*, PCWDEs and lignin-degrading enzymes encoded by the genomes of insects and their symbionts may be important in a wide range of biotechnological processes including the production of biofuels and food [7, 8].

We investigated the genomic basis of specialized phytophagy on woody plants by *A. glabripennis* through genome and transcriptome sequencing and annotation, comparative genomic analyses, gene expression assays, and functional genomic studies. Complementary comparative analyses involving the *A. glabripennis* genome and 14 additional insect genomes, including two additional beetles whose genomes are studied here for the first time – the emerald ash borer (*Agrilus planipennis,* family Buprestidae), and the bull-headed dung beetle (*Onthophagus taurus,* family Scarabaeidae) – were undertaken to reconstruct broader patterns in the evolution of insect (especially beetle) genes encoding enzymes involved in the digestion of woody plant tissues or detoxification of plant allelochemicals.

Results and discussion

General genome features

134X sequence coverage of the *A. glabripennis* genome was generated and assembled from a single female *A. glabripennis* larva, creating a draft genome reference assembly of 710 Mb with contig and scaffold N50s of 16.5 Kb and 659 Kb, respectively (Additional file 1: Table S3). While the *A. glabripennis* genome (female: 981.42 ±3.52 Mb, male: 970.64 ±3.69 Mb) is much larger than the four existing published beetle genomes (ranging from 163-208 Mb) [13-16], it is

average-sized for the order Coleoptera (mean=974 Mb) [17]. As in other draft genome assemblies, repetitive heterochromatin sequences could not be assembled, accounting for the differences between assembled sequence and genome sizes. The proportion of un-assembled genome in *A. glabripennis* is similar to that seen in other insect genome assemblies. 22,035 gene models were annotated using a customized MAKER pipeline [18]. Manual curation involved 1,144 gene models (Additional file 1: Table S4; Additional file 2: Table S6). The automated annotations and manual curations were merged into a non-redundant Official Gene Set (OGS v1.2) with 22,253 protein-coding gene models and 66 pseudogenes (Additional file 2: Table S6), in contrast to the 13,526-19,222 gene models reported for existing published beetle genomes. The completeness of the *A. glabripennis* genome assembly and OGS were assessed using benchmarking sets of universal single-copy orthologs (BUSCOs) [19] and compared with 14 other insect genomes (Fig. 2). The *A. glabripennis* gene set had slightly fewer missing BUSCOs (~3.3%) than most of the other genomes studied. Comparing BUSCO results from the *A. glabripennis* OGS to those obtained from searching the entire genome sequence, the number of missing genes was reduced, indicating that some genes were missed during the automated annotation process. Nonetheless, except for unassembled heterochromatin and other repetitive regions, the *A. glabripennis* genome is well represented and of high quality.

OrthoDB orthology delineation [20] revealed that *A. glabripennis* has a conserved core of 5,029 genes classified in orthologous groups (OGs) with orthologs from the 14 other insect genomes studied (Fig. 3). *A. glabripennis* has a high number of widespread orthologs (6,880 total) in OGs that are not universal but nevertheless have representatives from each of the three sets of species studied (see Methods and Additional file 1: Section I.6). About half (3,346) of these genes are maintained as single-copy orthologs, while the remainder (3,534) appear to have duplicated. Such duplications are more frequent in *A. glabripennis* than in most of the other species, but are not as extreme as in *Acyrthosiphon pisum* (pea aphid, family Aphididae) (8,779). Examining OGs with orthologs from only two of the three species sets showed that the

Coleoptera have maintained more ancient orthologs than the Diptera and Lepidoptera. Of the five Coleoptera genomes studied, *A. glabripennis* has the most Coleoptera-specific genes (5,229), suggestive of a high degree of adaptive novelty. Of these, 1,210 have identifiable orthologs in the other beetles and 2,789 show no clear orthology but do have homologs in other arthropods, i.e., they are likely divergent gene copies, consistent with the large numbers of paralogs in the *A. glabripennis* genome. This leaves a small set of 1,003 unique *A. glabripennis* genes with no homology to the other arthropod genes. A phylogenomic analysis of orthologs (Fig. 2) places *A. glabripennis* sister to *Dendroctonus ponderosae* (mountain pine beetle, family Curculionidae), as expected [21, 22].

In addition to glycoside hydrolase (GH) family genes (discussed below), sixteen HGT candidates were found from bacteria to *A. glabripennis,* and junctions between the insertion and flanking sequences were confirmed in multiple libraries (Additional file 1: Table S7). Four candidates were from bacteria most closely related to *Wolbachia*. Other represented potential sources include *Rickettsia, Calothrix, Clostridium* and *Brachyspira*. None of these HGT candidates showed significant expression in RNA-seq reads for adult males, females or larvae, although this does not rule out expression in other stages or tissue-specific expression of these candidates below detection in whole organism RNA-seq. Following HGT, insertions will either degrade by mutation and deletion, or (occasionally) evolve into functional genes. The sixteen HGTs above are likely recent insertions. Recent insertions have similarly been detected in other arthropod genomes using the DNA based pipeline [23, 24]. In contrast, the GH HGTs are more ancient insertions that have evolved into functional genes [25-29]. No microbial scaffolds were found in the *A. glabripennis* assembly, likely because the tissues used for sequencing (see Additional file 1) are not known to be associated with microbes.

A. glabripennis harbors similar numbers and kinds of genes involved in growth, development and reproduction as *T. castaneum* (and other insects) (Additional file 1: Section VI). Some of these gene clusters (e.g., homeodomain transcription factors) correlate in scale with its genome

size (~5X larger than *T. castaneum*) but also show *A. glabripennis*-specific paralogous expansion and gene dispersal. Key components of the genetic mechanisms underlying diapause in other insects were also found in the *A. glabripennis* genome. In contrast, *A. glabripennis* appears to posses an incomplete methylation machinery, including the maintenance methyltransferase DNMT1, but lacking the *de novo* methyltransferase DNMT3, which was lacking from both the genome assembly and the unassembled raw reads (Additional file 1: Section VI.10). While a similar situation is found in both *T. castaneum* and *Drosophila melanogaster* (common fruit fly, family Drosophilidae), many other insects, including other beetles such as *O. taurus* [30] and *Nicrophorus vespilloides* [13] (burying beetle, family Silphidae), have retained the complete machinery. A full description of the genes studied in the *A. glabripennis* genome can be found in the supplementary materials (Additional file 1).

Plant cell wall degradation

86 glycoside hydrolase (GH) family genes (Fig. 4 and Table 1; Additional file 1: Figure S18 and Tables S9, S17) were manually annotated in the *A. glabripennis* genome, more than are known from any other insect. These include a large expansion of 57 GH1 genes, which putatively exhibit (amongst others) β-glucosidase and β-galactosidase activities. Only 15 GH1 genes are known from *T. castaneum* [15], and only 19 from *D. ponderosae* [14]. We manually annotated 11 putative endo- and exoglucanases (cellulases), members of GH9, subfamily 2 of GH5, GH45 and GH48, and 18 GH28 genes encoding putative pectin-degrading polygalacturonases. Previous work has shown that a number of GH family genes have been acquired from microbes by HGT [e.g., references 23-29; Table 1], and Figure 4 shows the distribution of these and endogenous GHs in the 15 arthropod genomes studied herein. The genome of *A. glabripennis* was unique among the 15 species studied in containing matches to GH5 (IPR001547; see Fig. 4), whose members exhibit predominantly endo- and/or exo-glucanase, mannanase and xylanase activities.

Table 1. Plant cell wall degrading enzymes identified in the *A. glabripennis* genome assembly by manual annotation. Genes encoding GH9 cellulases have an ancient origin in animals [25]. The other beetle-derived GH families involved in plant cell wall digestion have a more recent origin and were putatively obtained via HGT from bacteria or fungi. GH5 subfamily 2 genes were likely acquired via HGT from Bacteroidetes [26]. GH45 genes were likely acquired by the last common ancestor (LCA) of the Phytophaga via HGT from a fungus [27, 28]. Amino acid sequences of beetle GH48 cellulases are similar to bacterial cellobiosidases, but their function(s) remain unclear; they may have evolved to scavenge nitrogen by degrading chitin in the gut or diet [31], e.g., from host plant tissues containing fungi, or from fungi resident in the gut (e.g., yeasts, *Fusarium solani*) which are thought to concentrate nitrogen and synthesize essential amino acids [9, 29, 32]. GH48s are constitutively highly expressed in *A. glabripennis* larvae (Fig. 5), and their induction in larvae feeding in a nutrient poor environment (reported herein) is consistent with a putative role in nutrient scavenging. They were most likely acquired by the LCA of the Phytophaga via HGT from a bacterial donor [27, 29]. GH28 genes were likely acquired by the LCA of the Phytophaga via HGT from an ascomycete fungus and subsequently expanded and diversified, but lost in the longhorned beetle subfamily Lamiinae (which includes *A. glabripennis*). After this loss, a GH28 gene was apparently re-acquired by Lamiinae via HGT from a fungal donor [10].

Gene family	Putative function	Genes total	Pseudogenes
	Cellulose/Hemicellulose Degradation		
GH9	Endo-β-1,4-glucanase	1	0
GH45	Endo-β-1,4-glucanase	2	0
GH5 subfamily 2	Endo/exo-β-1,4-glucanase	6	0
GH48	Reducing end-acting cellobiohydrolase	2	0
GH1	β-glucosidase (myrosinase, cyanogenic β-glucosidase)	57	3
	Pectin Degradation		
GH28	Polygalacturonase	18	0

We investigated diet-dependent regulation of GH family genes via an RNA-Seq based differential expression analysis of *A. glabripennis* larvae feeding on artificial diet versus the wood of living sugar maple trees, a preferred host. All GH5 and GH45 cellulases were expressed at least 2-fold higher in larvae feeding in sugar maple (Fig. 5) and have likely roles in converting cellulose into more easily digestible cello-oligosaccharides. Over 30 GH1 genes were most highly expressed in larvae feeding in sugar maple. Many of these genes are putative β-glucosidases and likely convert cellobiose and other oligosaccharides released from the plant cell wall into monosaccharides. GH1 enzymes can have broad catalytic and substrate specificities, so GH1 genes induced in larvae feeding in sugar maple could also function as β-xylosidases, β-glucuronidases, β-galactosidases, β-mannosidases, or exo-β-1,4-glucanases, serving to hydrolyze substrates released from the hemicellulose matrix. Additionally, many β-glucosidases also have known roles in detoxification [33, 34] (see below). Twelve GH28 genes showed elevated expression in larvae feeding in sugar maple, and their homologs are known to function as polygalacturonases in relatives of *A. glabripennis* [7, 10]. Thus, pectinous components of plant primary cell walls may serve as a significant source of sugars for early instar *A. glabripennis* larvae. GH35 genes were also induced in *A. glabripennis* larvae feeding in sugar maple. These had highest scoring BLAST alignments to β-galactosidase and could play roles in processing β-1,4 linked galactose oligomers released from the plant cell wall matrix. GH30 genes were also highly induced in larvae feeding in sugar maple. While some of these were expressed in both larvae and adults, two were expressed exclusively in larvae (AGLA015835 and AGLA015837) and may be important for digesting components of plant secondary cell walls. Consistent with this hypothesis, these two GH30 genes were strongly upregulated in insects feeding in sugar maple compared to artificial diet with log fold change expression values of 6.7 (FDR=1.14e-05) and 6.0 (FDR=1.83e-07). Additionally, three other GH30 genes were more highly expressed in larvae feeding in sugar maple including AGLA015834 (logFC=5.0; FDR=2.96e-11), AGLA015831 (logFC=1.96; FDR=0.029), and

AGLA001694 (logFC=1.80; FDR=0.05). Although the expression patterns of these genes seem consistent with a role in breaking down secondary cell wall polysaccharides in the larval stage, the precise reactions catalyzed by these gene products could not be predicted based on electronic annotations.

To determine substrate specificity and the contribution of enzymes encoded by GH family genes to the metabolism of plant cell wall polysaccharides, 15 of the 18 known *A. glabripennis* GH28 genes (putative polygalacturonases) were functionally characterized *in vitro*. Heterologous expression succeeded for all but GH28-4 (AGLA010098) (Additional file 1: Figure S5). Most GH28 proteins were active against at least one homogalacturonan polymer in plate assays. A group of phylogenetically related proteins, GH28-1 (AGLA010095), -2 (AGLA010096), -3 (AGLA010097) and -5 (AGLA010099), all located in tandem on one genomic scaffold, showed no activity against homogalacturonan polymers (Additional file 1: Figures S5, S6B, S7). However, they did exhibit exopolygalacturonase activity, similar to a previously characterized GH28 from a near relative of *A. glabripennis* [7] (Additional file 1: Figure S6C). GH28-11 (AGLA002350), the only polygalacturonase expressed in both *A. glabripennis* larvae and adults [7], and GH28-17 (AGLA025090), both functioned as endopolygalacturonases; however, accumulation of galacturonic acid monomers was also observed for GH28-11, indicating that it could also function as an exopolygalacturonase (Additional file 1: Figure S6C). Overall, the repertoire of GH28 enzymes encoded by the *A. glabripennis* genome contains both endo- and exo-polygalacturonases and is able to act on substrates with varying degrees of methylation. These enzymes are highly complementary, allowing *A. glabripennis* to efficiently decompose pectinous homogalacturonan polymers present in the primary cell walls of living woody plant tissues.

Six GH5 genes, two GH45 genes, and one GH9 gene were also functionally characterized *in vitro*. GH5-1 (AGLA002353) functioned as an endo-β-1,4-xylanase (EC 3.2.1.8), GH5-2 (AGLA002352), GH5-5 (AGLA006972), GH45-1 (AGLA005419) and GH45-2 (AGLA005420)

functioned as endo-β-1,4-glucanases (EC 3.2.1.4), and GH5-2 showed endo-β-1,4-xyloglucanase activity (EC 3.2.1.151) (Additional file 1: Figures S8B, S9). GH5-2 also hydrolyzed carboxymethylcellulose (CMC), indicating that enzymes encoded by this gene possess the ability to endohydrolyse the 1,4-β-D-glucosidic linkages in both CMC and xyloglucan and may function to degrade both cellulose and components of hemicellulose *in vivo*. GH5-3 (AGLA002354), GH5-4 (AGLA002351), GH5-6 (AGLA016376) and GH9 (AGLA010313) did not harbor any enzymatic activity against the substrates tested, indicating that they are not endo-acting enzymes. To investigate how GH5 enzymes degrade their substrates, the products were subsequently analyzed by thin layer chromatography (TLC) (Additional file 1: Figure S8C and Methods), validating the roles of GH5-1 as a xylanase, GH5-2 as a dual-acting xyloglucanase/endoglucanase, and GH5-5 as an endoglucanase. Furthermore, although no zone of clearing was observed for GH5-6 in an agarose diffusion assay, accumulations of glucose and cellobiose were observed via TLC after incubation with CMC, suggesting that it functions as an exo-β-1,4-glucanase (Additional file 1: Figure S8C). None of these enzymes had the ability to degrade crystalline cellulose substrates. However, Geib et al. [32] observed activity against Avicel in enzyme extracts prepared from larval *A. glabripennis* guts. This suggests that (a) GH5 and GH45 cellulases act synergistically *in vivo* to degrade these substrates, (b) other *A. glabripennis*-encoded enzymes besides those characterized in this study possess the ability to degrade Avicel, or (c) that enzymes produced by the gut microbial community are responsible for the aforementioned previously observed activity. Notably, the cellulases encoded by numerous members of the *A. glabripennis* gut microbial community possess carbohydrate binding domains, which could enhance the efficiency of these enzymes against crystalline substrates by allowing them to bind and degrade their substrates in a processive manner [29, 32]. Thus, the *A. glabripennis* genome encodes at least 3 families of cellulases and hemicellulases (subfamily 2 of GH5, GH9 and GH45) and one family of polygalacturonases (GH28) that provide it with an arsenal of enzymes capable of degrading the

main polysaccharides of the cellulose and hemicellulose networks in both primary and secondary plant cell walls.

GH28, GH45 and subfamily 2 of GH5 were collectively detected only in the three phytophagous beetle genomes studied (*A. glabripennis*, *A. planipennis* and *D. ponderosae*) (Fig. 4; Additional file 1: Figure S18) and were lacking from the 12 other insect genomes. Specifically, GH28 was detected in *A. glabripennis*, *A. planipennis* and *D. ponderosae*, GH45 was detected only in *A. glabripennis* and *D. ponderosae* (sister taxa in our phylogeny, spanning the basal split in the clade Phytophaga [36] (Fig. 2), and subfamily 2 of GH5 was detected exclusively in *A. glabripennis*. Subfamily 2 of GH5 genes have been found in at least one other cerambycid [7] and may be unique to superfamily Chrysomeloidea (leaf beetles, cerambycids and their relatives). *A. glabripennis*, *A. planipennis* and *D. ponderosae* are all specialized phytophages belonging to species-rich taxonomic groups of beetles that feed on the subcortical tissues of woody plants and interact with specialized suites of gut microbes. Interestingly, the genomes of the wood-feeding termites *Macrotermes* and *Zootermopsis* lack all three of the aforementioned gene families. However, these genes are present in the genomes of their gut symbionts. This is in contrast to the phytophagous beetles we studied, whose ancestors obtained these genes (in their genomes) via HGT from bacteria and fungi [8, 14] (Additional file 1: Figures S5, S9). These genes subsequently diversified in beetle genomes to form multi-gene families [10]. Notably, the GH28 family genes we annotated in *A. planipennis* were apparently acquired independently (via HGT from an ascomycete fungus donor) from those in *A. glabripennis* and *D. ponderosae*. Independently-acquired GH28 genes are also known from phytophagous Hemiptera in the species-rich family Miridae [37].

GH1 family genes can encode enzymes having both digestive and non-digestive functions. 23 *A. glabripennis* GH1 sequences had ~44% identity to sequences annotated as myrosinases (MYR) [30] in the *T. castaneum* genome [38]. One sequence closely matches known myrosinase active site motifs. For some insects, including flea beetles, myrosinases are known

to synergize alarm or aggregation pheromones [39, 40]. Non-Brassicacaeous, woody plant sources of glucosinolytes, which are the substrates detoxified by myrosinase, are present in the *A. glabripennis* native range [41]. An additional possibility is that one or more of these *A. glabripennis* sequences is a cyanogenic β-glycosidase [34]. Toxic cyanogenic glycosides are used by some plants (including known hosts of *A. glabripennis*) as a defense against insect-feeding, analogous to the myrosinase system. Interestingly, five *A. glabripennis* GH1 sequences are intermediate in similarity to known myrosinases and a known cyanogenic β-glycosidase (Additional file 1: Figure S16).

Microbes in the gut of *A. glabripennis* are known to have definitive roles in nutrient biosynthesis and nutrient recycling, helping the beetle to thrive under nutrient poor conditions [32, 42, 43]. *A. glabripennis* microbes encode an arsenal of laccases, peroxidases, aldo-keto reductases, dyp-type peroxidases [29], and at least one lignin peroxidase, which is encoded by a fungal symbiont belonging to the *Fusarium solani* species complex [44]. Several of the aforementioned genes are actively expressed in the *A. glabripennis* larval midgut [32]. While these enzymes have not been functionally characterized *in vitro*, they may facilitate lignin degradation in the *A. glabripennis* gut. The *A. glabripennis* genome itself may also encode genes that facilitate lignin degradation. *A. glabripennis* encodes eight genes with hemocyanin domains, three of which are significantly more highly expressed in larvae feeding in sugar maple, including the gene models AGLA002479 (2.1 log-fold upregulation), AGLA002478 (2.5 log-fold upregulation), and AGLA001233 (3.4 log-fold upregulation). All three genes were originally thought to function as storage hexamer proteins. However, the ability of at least one termite-derived hemocyanin highly expressed in salivary glands to oxidize model lignin compounds and other aromatic compounds *in vitro* [45], and the high expression levels of these three genes in multiple organisms that feed in wood [46], could signal that they work synergystically with gut microbes in *A. glabripennis* to facilitate oxidative degradation of prominent linkages in the lignin polymer and/or other biopolymers *in vivo*.

335 **Detoxification of plant allelochemicals**

336 To gain further insights into the genomic basis of the broad host range of *A. glabripennis* (>100
337 known host tree species) and its concomitant invasiveness, we studied gene families
338 hypothesized to encode key enzymes involved in the detoxification of plant allelochemicals
339 (Additional file 1: Tables S17-S26 and Figures S18-S22). Cytochrome P450s (CYP450;
340 Additional file 1: Figure S21 and Tables S20, S25) encode the most prevalent detoxification
341 enzymes in insects, and participate in many other important physiological processes. A total of
342 106 genes and 19 pseudogenes predicted to encode CYP450s were manually annotated in the
343 *A. glabripennis* genome. 137 genes and 6 pseudogenes were detected by matches to InterPro
344 domains, the third highest number in our comparative genomic study, after the beetles *T.*
345 *castaneum* and *O. taurus*. Examining the CYP450 sub-families showed that *A. glabripennis* had
346 five times as many Group II matches (18 genes; including CYP4 and CYP6) than the average
347 across the other insect species studied. CYP6 enzymes metabolize a wide range of toxic
348 compounds and are known to clear odorants in insect antennae [47]. CYP4 enzymes are
349 involved in cuticular hydrocarbon biosynthesis and have been implicated in insecticide
350 resistance [48]. Supporting their roles in detoxification, 25 CYP450 genes were induced in the
351 guts of *A. glabripennis* larvae feeding in sugar maple, including many genes in *A. glabripennis*-
352 specific clades (Additional file 1: Figure S10). Only two of the genes that were induced
353 (CYP18A1, CYP314A1) occurred in orthologous pairs with *T. castaneum* genes. Therefore,
354 while the many CYP450 ortholog pairs between *T. castaneum* and *A. glabripennis* presumably
355 carry out functions conserved over millions of years of evolution, expansion of several CYP
356 families and the evolution of *A. glabripennis* specific CYP clades relative to *T. castaneum*
357 suggests that these genes have evolved and diversified in *A. glabripennis* as a mechanism to
358 overcome host plant defenses.

359 UDP-glycosyltransferases (UGTs) assist with the detoxification and elimination of
360 xenobiotics (foreign substances such as those produced by parasites) and in the regulation of

361 endobiotics (substances produced, e.g., in response to the presence of parasites). 65 putative
362 UGTs, including 7 pseudogenes, were manually annotated in the *A. glabripennis* genome (Fig.
363 6; Additional file 1: Figures S11, S12, S22 and Tables S21, S26; Additional file 2: Table S16).
364 Only two taxa have so far been reported to harbor a greater number of UGT genes – *Locusta*
365 *migratoria* (the migratory locust, family Acrididae; 68 UGTs) [49], and the aphid *A. pisum* (72
366 UGTs; reported herein via matches to InterPro domains; 58 UGT genes were reported for *A.*
367 *pisum* by Ahn *et al.* [50]). The expansion of UGTs in *A. glabripennis* may be related to its ability
368 to feed on a broad range of healthy host plants, a feature shared with *L. migratoria*.
369 Approximately 92% of *A. glabripennis* UGTs are arranged in a tandem manner and 50 of them
370 were concentrated in just seven clusters. Most UGTs thus appear to have diversified by tandem
371 gene duplication, resulting in increased substrate range of host secondary metabolites by
372 altering the N-terminal substrate binding domain of the enzyme. The largest UGT family
373 observed in *A. glabripennis*, UGT352, is unique to this species and consists of 21 genes. 14
374 UGT352 genes were positioned in the same orientation in a cluster on one scaffold (Fig. 6). An
375 *A. glabripennis*-specific expansion of 7 genes was found in the UGT321 gene family. These
376 expansions may enable *A. glabripennis* to adapt to a wide range of host plant defenses.
377 Consistent with this hypothesis, four UGTs were strongly upregulated in *A. glabripennis* larvae
378 feeding in sugar maple, including two UGT321 genes, and one UGT352. Although only a portion
379 of the potential detoxification genes harbored in the *A. glabripennis* genome were induced while
380 feeding in sugar maple – just one of the many host plants of *A. glabripennis* – the existence of a
381 diverse metabolic repertoire likely helps *A. glabripennis* feed on different host species that
382 produce different defensive compounds.
383 In addition, the *A. glabripennis* genome was found to contain more putative esterases than
384 any of the other insect genomes studied (Additional file 1: Figure S20 and Tables S19, S24).
385 This is due mainly to a large expansion of type-B carboxylesterases (COesterases; IPR002018),
386 most of which are paralogs. COesterases are important for the metabolism of xenobiotics and

for degrading ester bonds linking lignin to hemicellulose in plant secondary cell walls. 107 COesterases were identified in the *A. glabripennis* genome (Additional file 1: Figure S14), more than double the average in the other species studied. Most COesterases occur in large clusters; only 28 (25%) occur as singletons. Two large clades of COesterases, one containing 17 genes and the other 13 genes, were unique to *A. glabripennis*. *A. glabripennis* also had the most genes (8 total) matching the thioesterase domain (IPR001031). COesterases were among the most highly induced genes in *A. glabripennis* larvae feeding in sugar maple and most of the highly induced COesterases belonged to *A. glabripennis*-specific clades and formed tandem repeats in the genome, potentially signifying novel functions related to digestion of woody plant tissues, or detoxification of plant allelochemicals.

Digestive proteinases may play key roles in scavenging nitrogen from plant cell wall proteins or midgut endosymbionts, and may help phytophagous insects cope with proteinase inhibitors produced by plants [51]. *A. glabripennis*-specific expansions of several proteinase OGs were observed in comparison to *T. castaneum* and *D. ponderosae*. The largest were OGs EOG8V724X and EOG8V19NQ, comprising tandem arrays of eight and seven trypsin genes, respectively. Both OGs contain genes predicted to encode secreted serine proteinases. Most proteinase genes were unique to each of the five beetle species studied, suggesting that their evolution occurred largely after speciation and may be correlated with exposure to different digestive enzyme inhibitors and with feeding on different diets. These gene families appear to be highly dynamic and may largely shape the digestive physiology of phytophagous insects.

Sensory biology

A. glabripennis adults use a complex set of chemical and visual cues for host plant and mate finding. We compared the members of four gene families involved in chemoperception (olfaction and gustation) and vision in *A. glabripennis* with those from *T. castaneum* and *D. melanogaster*. We manually annotated 52 odorant binding protein (OBP) genes in the *A. glabripennis* genome (Additional file 1: Figure S23). Most OBPs comprise a large expansion of the minus-C

subfamily, and the remaining genes were placed singly or in small radiations that exhibit the classic 6-cysteine motif. One OBP (AglaOBP51) was identified as a member of the plus-C group, the same as in *T. castaneum* and *D. ponderosae* [52], suggesting that the tendency toward minus-C OBPs originated at least with the beetle infraorder Cucujiformia (~190 Ma) [3]. *A. glabripennis* has 131 odorant receptor (OR) genes in addition to the highly conserved OR co-receptor, Orco (Additional file 1: Figure S24). These include representatives of all seven sub-families of beetle ORs except group 6, and follow the pattern of frequent paralogous radiations typical of insect chemoreceptors. Two new lineages of ORs were identified in *A. glabripennis* and placed as outgroups to OR groups 4, 5, and 6 in *T. castaneum* (Or106-115/126-132 and Or101-103). The function of beetle ORs remains mostly unknown, and receptors have only been characterized from *Megacyllene caryae* (hickory borer, family Cerambycidae) (McarOr3). AglaOr29 is notably sister to McarOr3, which is sensitive to 2-methylbutan-1-ol, a pheromone component of *Megacyllene* [53].

A. glabripennis has an extensive suite of 234 gustatory receptors (GRs), including three conserved candidate CO_2 receptors (Gr1-3), 10 candidate sugar receptors (Gr4-13), and three candidate fructose receptors related to DmGr43a (Gr14-16). The remaining 127 GRs encode 218 receptors through alternative splicing, and presumably belong to the general category of candidate bitter taste receptors, although some likely are also involved in contact pheromone perception [54] – a component of *A. glabripennis* mate finding behavior [55]. *A. glabripennis* has 72 ionotropic receptors (IRs) including orthologs of the conserved co-receptors IR8a and 25a, and of IR21a, 40a, 41a, 68a, 76b, 93a, and 100a. The IR75 lineage consists of 8 genes compared with 6 in *T. castaneum* and 7 in *D. melanogaster*. These are all candidate ORs, while the candidate GRs, represented by the DmIr20a clade of 40 genes [56], consist of 55 genes, compared to 53 in *T. castaneum*, although these two beetles exhibit differential species-specific expansion of gene lineages within this large grouping. Like *T. castaneum* [15, 57], *A. glabripennis* has large OR and GR repertoires compared with *D. melanogaster*, and indeed

most other insects except ants, but their OBP and IR repertoires are more comparable with that of *D. melanogaster* and similar to many other insects (Additional file 1: Table S27). The optical sensitivity of *A. glabripennis* appears to be similar to that of *T. castaneum* [58]. *A. glabripennis* has a single long-wavelength-sensitive opsin and a single UV-sensitive opsin. *A. glabripennis* differs from *T. castaneum*, however, in having the Rh7 opsin, whose function is unknown, and in lacking the c-opsin found in most other insects and other arthropods, which is presumed to have a non-visual function [59].

Conclusions

A. glabripennis possesses a remarkably robust enzymatic repertoire capable of digesting most of the polysaccharides it encounters while feeding on woody host plants (cellulose, xyloglucan, xylan and pectin). Furthermore, diverse suites of detoxification genes, and several classes of digestive proteinases provide *A. glabripennis* with the metabolic plasticity needed to overcome the challenges of feeding on several different host trees, each with a distinct profile of defensive compounds. Many of the paralogs in gene families encoding enzymes typically involved in plant cell wall degradation (PCWDEs) and detoxification occur in large clusters in the *A. glabripennis* genome and appear to have diversified by tandem gene duplication. Large expansions of genes encoding CYP450s, UGTs, COesterases (these three together are sometimes called the defensome; e.g., [60]) and GH1s in the *A. glabripennis* genome are particularly notable, as they are among the largest such repertoires of detoxification genes known in insects. Genes encoding PCWDEs are also uniquely expanded in number in the *A. glabripennis* genome. The *A. glabripennis* genome encodes genes from a remarkable 3 families of putative cellulases (GH5 subfamily 2, GH9 and GH45), and one of these, GH5 subfamily 2, evolved in such a way that it provides the beetle with an arsenal of enzymes possessing the ability to degrade the main polysaccharides of the cellulose and hemicellulose (xylan and xyloglucan) networks in both primary and secondary plant cell walls. *A. glabripennis* also has the ability to degrade lignin,

either through the activities of its gut microbial fauna and/or by way of enzymes encoded in its genome. Our results are notable in including not only an enumeration of genes potentially involved in plant cell wall degradation and detoxification (thus facilitating specialized phytophagy on woody plants and a wide host range), but also results from experimental assessments of gene expression and enzyme activiies.

Acquisition of new genes (here, GH5, GH28 and GH45 family genes) via HGT from bacteria and fungi, followed by gene copy number amplification and functional divergence were fundamental to the addition, expansion and enhancement of the metabolic repertoire of *A. glabripennis*. Our results thus begin to establish both a genomic basis for the invasiveness and broad host plant range of *A. glabripennis*, and more generally, the apparent evolutionary success of beetles on plants.

Methods (More information is available in Additional file 1, and supporting scripts are available at https://github.com/NAL-i5K/AGLA_GB_supp-scripts).

Genome size and DNA and RNA for sequencing. The genome size of 5 male and 5 female adult *A. glabripennis* collected from the former Chicago, IL, USA infestation were estimated via flow cytometry. The *A. glabripennis* specimens sequenced for this project were obtained from a USDA-APHIS colony stocked with the descendants of beetles collected from current and former infestations in IL, NY, and MA, except when noted otherwise in the supplement (Additional file 1: Table S1). The *A. glabripennis* genome was sequenced from DNA that was extracted from a single late instar female larva (G Biosciences, Omniprep kit), whose sex was determined after sequencing (Additional file 1: Figure S3).

Genome sequencing and assembly. An enhanced Illumina-ALLPATHS-LG [61] sequencing and assembly strategy was employed. We sequenced four libraries of nominal insert sizes 180 bp, 500 bp, 3 kb and 8 kb at genome coverages of 59.7X, 45.8X, 58.7X and 20.5X respectively. Sequencing was performed on Illumina HiSeq2000s generating 100bp paired end reads. Reads

were assembled using ALLPATHS-LG (v35218) and further scaffolded and gap-filled using in-house tools Atlas-Link (v.1.0) and Atlas gap-fill (v.2.2) (https://www.hgsc.bcm.edu/software/). Data for the *A. glabripennis* genome has been deposited in the GenBank/EMBL/DDBJ Bioproject database under the accession code PRJNA163973 (Additional file 1: Table S3). Raw genomic sequence data is deposited in the GenBank/EMBL/DDBJ sequence read archive under the accession codes of SRX326764, SRX326768, SRX326767, SRX326766, and SRX326765. The genome assembly has been deposited to GenBank under the accession GCA_000390285.1. RNA-seq datasets used in gene prediction are deposited to the GenBank/EMBL/DDBJ sequence read archive under the accession codes SRX873913 and SRX873912.

Automated annotation. The *A. glabripennis* genome assembly was subjected to automatic gene annotation using a MAKER 2.0 [62] annotation pipeline tuned for arthropods. Both protein and RNA-seq evidence from extant arthropod gene sets were used to guide gene models. The genome assembly was first subjected to *de novo* repeat prediction and Core Eukaryotic Genes Mapping Approach (CEGMA) analysis [63] to generate gene models for initial training of the *ab initio* gene predictors. Three rounds of training of the Augustus [64] and SNAP [65] gene predictors within MAKER were used to bootstrap to a high quality training set. RNA-seq data from *A. glabripennis* adult males and females was used to identify exon-intron boundaries. Finally, the pipeline used a nine-way homology prediction with human, *D. melanogaster* and *C. elegans*, and InterPro Scan5 to allocate gene names. The automated gene set is available from the BCM-HGSC website (https://www.hgsc.bcm.edu/asian-long-horned-beetle-genome-project) and at the National Agricultural Library (https://i5k.nal.usda.gov).

Community Curation. The *A. glabripennis* genome was curated to improve the structural and functional annotations of genes and gene families of interest using the Web Apollo manual curation tool [66] (Additional file 1: Table S4; Additional file 2: Tables S5, S6). Web Apollo is an interactive, web-based manual curation tool that visualizes user-generated annotation changes

515 in real time, allowing remote collaboration on annotations. The *A. glabripennis* genome
516 coordinator (D. McKenna, University of Memphis) organized a group of experts to manually
517 curate genes or gene families of interest in Web Apollo. Web Apollo
518 (https://apollo.nal.usda.gov/anogla/jbrowse/) tracked all evidence used for the MAKER gene
519 predictions, as well as an additional RNA-Seq dataset that was not used in the generation of the
520 MAKER gene predictions. The manually curated models were inspected for quality, including
521 overlapping models, internal stop codons within the CDS, gff3 formatting errors, and mixed
522 transcript types within gene models. The quality-corrected models were then merged with the
523 MAKER-predicted gene set to generate an official gene set (OGS), followed by post-processing
524 to ensure curation information was transferred adequately. A full list of conditions for mRNA,
525 gene, exon and CDS are listed in Additional file 1: Table S5. All functional information was
526 included in the OGS. Information on the *A. glabripennis* genome project is collated at the i5k
527 Workspace [67] (https://i5k.nal.usda.gov/Anoplophora_glabripennis), and the genome, transcript
528 and protein sets can be searched via BLAST and browsed via the JBrowse genome browser
529 [68] (https://apollo.nal.usda.gov/anogla/jbrowse). All manually curated genes and transcripts
530 and their curation actions are provided in a supplemental table (Additional file 2: Table S6).
531 Additional details on annotation methods are provided in the Supplementary materials.
532 **Assessing orthology and the quality of genome assembly and annotation.** Orthology data
533 from OrthoDB v8 [20] with a total of 87 arthropod species were analyzed to identify orthology
534 and homology assignments of *A. glabripennis* genes with those of other beetles and
535 representative species from six other insect orders. The gene sets of *A. planipennis* and *O.*
536 *taurus* (unpublished data, manuscript in preparation; Fig. 2) were mapped to OrthoDB v8
537 orthologous groups (OGs) to include them in the analysis. The selected species include several
538 that feed on plants and were partitioned into three species sets - 5 Coleoptera, 5
539 Lepidoptera/Diptera, and 5 outgroup insects. Arthropod OGs were queried with custom Perl
540 scripts to identify OGs with genes from all three species sets (across 15 species), just two sets

(across 10 species), or restricted to a single set (across 5 species). To be considered shared, orthologous groups were required to contain genes from at least two species in each set. For those shared among all three sets (a total of 7,376 OGs), the numbers of single-copy and multi-copy orthologs were summed across all OGs for each species. Lineage-restricted genes without orthologs were assessed for significant homology (e-value <1e-05) to other arthropod genes from OrthoDB or for significant homology (e-value <1e-05) to genes from their own genomes (self-only homology). The completeness of the *A. glabripennis* genome assembly and annotated Offical Gene Set (OGS) were assessed using BUSCOs [19]. We compared the results from *A. glabripennis* to those from 14 other insect genomes (Figure 2B; Additional file 1: Figure S1). We used the Arthropoda gene set, which consists of 2,675 single-copy genes that are present in at least 90% of Arthropoda.

Identification of bacterial to eukaryote horizontal gene transfers. HGTs were identified as described in Wheeler et al. [69]**.** Briefly, we used BLASTN to compare genomic scaffolds against a bacterial database containing 1,097 complete bacterial genome sequences downloaded from the National Center for Biotechnology Information (NCBI). Regions with significant bacterial identity (E value<1e−5) were then compared to a second database containing representative animal genomes (see Wheeler et al. [69] for list of animal species) obtaining a corresponding "animal" BLASTN E value score. If the animal E value score was less than the bacterial E value score the sequence was excluded as a slowly evolving highly conserved gene. Candidates were then further annotated manually for flanking eukaryotic genes and junctions between eukaryotic and bacterial sequences in the libraries. For glycoside hydrolases, the same methods were used, but additionally, we simply BLASTed the genome using sequences of known, characterized PCWDEs found in phytophagous beetles [8-10] including *Apriona japonica* [7], a close relative of *A. glabripennis*.

Differential expression analysis of *A. glabripennis* larvae feeding on sugar maple versus artificial diet. Five pairs of adult male and female *A. glabripennis* were allowed to maturation

feed on fresh twigs collected from Norway maples (*Acer platanoides*, family Aceraceae) for two weeks. After this period, the beetles were allowed to mate and oviposit into potted sugar maple trees (*Acer saccharum*) maintained in a USDA-approved quarantine greenhouse for two weeks. The trees were harvested approximately 60 days after the eggs hatched and four third-instar larvae were collected. Four third-instar larvae feeding on artificial diet [70] were also harvested. Larvae were surface sterilized, dissected, and their midguts were removed and frozen in liquid nitrogen. RNA was isolated, and ribosomal RNA was depleted from the sample using Ribominus Eukaryotic Kit for RNA-Seq (Life Technologies). The enriched mRNA was further polyA purified and multiplexed Illumina libraries were constructed using the TruSeq RNA Sample Prep kit (Illumina, San Diego, CA). Samples were pooled and sequenced on a single Illumina HiSeq lane at the University of Delaware Biotechnology Institute (Newark, DE) to generate approximately 13 million 101 nt paired end reads per sample. Forward reads were trimmed and quality filtered using ea-utils (https://code.google.com/p/ea-utils/) and high quality reads of at least 75 nt in length were mapped to the *A. glabripennis* reference genome assembly using Tophat [71]. Read counts that mapped to each locus (version v0.5.3 annotations) were summed using HTSeq [72]; reads that spanned multiple features were summed using the union mode and reads that did not map uniquely to a single region in the genome were discarded. Differential expression analysis was performed using edgeR [73]. Features with less than 10 mapped reads were removed from the analysis, read counts were normalized by quantile normalization, and variances were estimated using tagwise dispersions. Statistical analysis was performed using Fisher's exact tests; features were flagged as differentially expressed if they had a log fold change greater than 1.0 and an adjusted p-value of < 0.05. Experiment-wise false discovery rate (FDR) was estimated at 0.05. The raw Illumina reads used for the differential expression analysis have been deposited into NCBI's Sequence Read Archive (SRA) and are associated with Bioproject PRJNA279780. The read counts used

592 to compute differential expression have been deposited in Gene Expression Omnibus (GEO)
593 under the accession GSE68149.
594 **In vitro functional characterization of plant cell wall degrading enzymes.** *A. glabripennis*
595 larval samples were obtained from D. Lance (USDA-APHIS-PPQ). Larvae were chilled on ice
596 and cut open; midguts from 1.5 month old, 4 month old and 8 month old larvae were collected
597 and stored in an excess of RNA Later solution (Ambion) prior to shipping. RNA was
598 subsequently isolated using the innuPREP RNA Mini Kit (Analytik Jena) according to the
599 manufacturer's protocol. Genomic DNA contamination was removed by DNAse treatment
600 (TURBO DNAse, Ambion) for 30 min at 37 °C. Midgut RNA was further purified using the
601 RNeasy MinElute Clean up Kit (Qiagen) following the manufacturer's protocol and eluted in 20
602 µl of RNA storage solution (Ambion). Integrity and quality of the RNA samples were determined
603 using the RNA 6000 Nano LabChip kit (Agilent Technologies) on an Agilent 2100 Bioanalyzer
604 (Agilent Technologies) according to the manufacturer's instructions.
605 Open reading frames encoding putative PCWDEs were amplified by PCR using gene-
606 specific primers. The forward primer was designed to introduce a 5' Kozak sequence, and the
607 reverse primer was designed to omit the stop codon. Equal amounts of total RNA prepared from
608 midguts either of 1.5 month old or 4 month old or 8 month old larvae were pooled, and 1 µg total
609 RNA from this pool was used to generate first strand cDNAs using the SMARTer RACE cDNA
610 Amplification Kit (BD Clontech), following the manufacturer's instructions. These cDNAs were
611 subsequently used as templates for PCR amplifications. PCR products were cloned into the
612 pIB/V5-His TOPO/TA (Invitrogen) vector, in frame with a V5-$(His)_6$ epitope at the carboxyl-
613 terminus. Constructs were transfected into insect *Sf*9 cells, cells were grown to confluence, and
614 expression of the recombinant proteins was validated as described previously [7]. Diffusion
615 assays were performed using 1% agarose Petri dishes in McIlvaine buffer (pH 5.0) containing
616 one of the following substrates: 0.1% carboxymethylcellulose (CMC, Sigma-Aldrich); 0.1%
617 beechwood xylan (Sigma-Aldrich); 0.1% xyloglucan from tamarind seeds (Megazyme); 0.1%

618 pectin from citrus peels (Sigma-Aldrich); 0.1% demethylated polygalacturonic acid (Megazyme)
619 Enzyme activity was detected using a 0.1% Congo Red solution as described previously [7].
620 TLC analysis of hydrolysis reaction products was also performed. The culture medium of
621 transiently transfected cells was first dialyzed against distilled water at 4 °C for 48 h, using
622 Slide-A-Lyzer Dialysis Cassettes with a 10 kDa cut off, before being desalted with Zeba Desalt
623 Spin Columns 7 kDa cut off (both Thermo Scientific), according to the manufacturer´s
624 instructions. Enzyme assays (20 ul) were set up using 14 µl of dialyzed and desalted crude
625 enzyme extracts mixed with 4 µl of a 1% substrate in solution in a 20 mM McIlvaine buffer (pH
626 5.0). For GH5-1 to -6, the following substrates were tested: carboxymethyl cellulose (CMC),
627 avicel (suspension), beechwood xylan and xyloglucan. For GH28s, the following substrates
628 were tested: demethylated polygalacturonic acid and pectin from citrus peels. The activity of
629 GH28s on 10 µg/µl aqueous solution of tri- and di-galacturonic acid was also tested. Enzyme
630 assays were incubated and plates developed as described previously [7].
631 Amino acid alignments were carried out using MUSCLE version 3.7 on the Phylogeny.fr web
632 platform (http://www.phylogeny.fr) [74] and were inspected and corrected manually when
633 needed. Bayesian analyses were carried out in MrBayes 3.1.2 [75]. Two runs were conducted
634 for the dataset showing agreement in topology and likelihood scores. To obtain support from a
635 second independent method, maximum likelihood analyses were also performed using MEGA5
636 [76]. The robustness of each analysis was tested using 1,000 bootstrap replicates.
637 **Comparative genomics of phytophagy and detoxification across Insecta.** Gene families
638 and sub-families associated with phytophagy (particularly xylophagy) and polyphagy or
639 detoxification were identified by searching for matches to relevant InterPro domains in the
640 complete gene sets from the genomes of 15 exemplar insect species. These included 5 beetles:
641 *A. glabripennis*, *D. ponderosae*, *T. castaneum*, *A. planipennis* (unpublished), and *O. taurus*
642 (unpublished); 5 basal insects: *Zootermopsis nevadensis* (dampwood termite, family
643 Termopsidae), *Pediculus humanus* (human louse, family Pediculidae), *A. pisum*, *Apis mellifera*

644 (honey bee, family Apidae), and *Nasonia vitripennis* (jewel wasp, family Pteromalidae); 2
645 lepidopterans: *Plutella xylostella* (diamondback moth, family Plutellidae), and *Danaus plexippus*
646 (Monarch butterfly, family Nymphalidae); and 3 dipterans: *Mayetiola destructor* (Hessian fly,
647 family Cecidomyiidae), *D. melanogaster*, and *Anopheles gambiae* (African malaria mosquito,
648 family Culicidae). Protein domains were annotated with InterProScan5 [77] using the following
649 domain libraries: PfamA-27.0, PrositeProfiles-20.97, SMART-6.2, SuperFamily-1.75, and
650 PRINTS-42.0. The gene families examined included glycoside hydrolases, peptidases,
651 esterases, cytochrome P450s, and UDP-glucosyltransferases.
652 The classifications based on InterPro domain counts were used only for those cases where
653 the maximum gene count in a given species was greater than 5 (i.e., at least one species had a
654 potential expansion of more than 5 genes). The orthology status of each of these identified
655 genes was assessed using OrthoDB v8 [20] to determine if the gene was found as a single-copy
656 ortholog, or with co-orthologs, or whether it showed homology to the domain but was not
657 classified in any orthologous group. The results of the counts of each relevant domain type and
658 the orthology status for the identified genes are given in Additional file 1: Tables S17-S26.
659 Domains were selected for plotting from the complete list to avoid redundant domains (e.g. sub-
660 families rather than families, and just one of N/C-terminal domains). For each gene family, the
661 bar charts were plotted with largest sub-family at the bottom and smallest at the top, showing
662 the counts for each sub-family per species (Additional file 1: Figures S18-S22). The orthology
663 status of genes in the sub-family bar charts (i.e., those plotted and where at least one species
664 has >5 genes) show the totals in each species partitioned into single-copy and multi-copy
665 orthologs, and homologs (Additional file 1: Tables S19-S23).

666 **Declarations**

667 **Acknowledgements**

We thank S. Raschka (MI State Univ.), and K. Morris, V. Lopez, S. Gula, and B. Walters (Xavier Univ.), M. Munoz-Torres (Lawrence-Berkeley National Lab), B. Wurlitzer and D. Schnabelrauch (Max Plank Inst. for Chemical Ecology), K. Kuhn and E. Aparicio (USDA ARS BIIR), and J. Johnson (ND State Univ.) for their assistance with this research, and D. Hibbett (Clarke Univ.) for comments on an early draft of the manuscript.

Funding

Funding for genome sequencing, assembly and automated annotation was provided by NHGRI grant U54 HG003273 to RAG. The i5k pilot project (https://www.hgsc.bcm.edu/arthropods) assisted in sequencing of the *A. glabripennis*, *A. planipennis* and *O. taurus* genomes. We acknowledge funding from the University of Memphis FedEx Institue, U.S. NSF grant DEB1355169 and USDA-APHIS cooperative agreement 15-8130-0547-CA to DDM, USDA NIFA Grant 2015-67013-23287 to KH and ES, NSF grant DEB1257053 to JHW, NSF grants IOS1120209 and IOS1256689 to APM, University of Cincinnati Faculty Development Research Grant to JBB, Marie Curie International Outgoing Fellowship PIOF-GA-2011-303312 to RMW., Swiss NSF awards 31003A-125350 and 31003A-143936 to EMZ, Max-Planck-Gesellschaft to YP, RK and AB, German Research Foundation (DFG) Grants PA 2044/1-1 and SFB 680 to KAP, and NIH postdoctoral training grant 5K12 GM000708-15 to RFM.

Additional files

Additional file 1: Supplementary figures, tables, methods, and other text. (DOCX 38.2 kb)

Additional file 2: Large supporting tables. (XLSX 352 kb)

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data

693 All sequence data is publically available at the NCBI, bioproject number PRJNA167479.
694 Additional file 1: Table S3 gives specific details of accession numbers for specific raw and
695 assembled DNA and RNA sequences. In addition, gene models and a browser are available at
696 the National Agricultural Library https://i5k.nal.usda.gov/Anoplophora_glabripennis.

697 **Competing interests**

698 The authors have no competing interests.

699 **Authors' contributions**

700 DDM and SR conceived, managed and coordinated the project. DL provided specimens for
701 sequencing. DDM performed DNA and RNA extractions. SR, HuD, YH, HaD, DMM and RAG
702 managed library preparations and sequencing. SR, SLL and HC constructed libraries and
703 performed sequencing. SCM, JQ, DSTH, SR and KCW performed the genome assembly and
704 automated gene prediction. DDM, MFP, CC, CL and HL developed and implemented
705 WebApollo manual curation. RMW, EMZ and PI performed orthology and phylogenomic
706 analyses. DDM, SA, DA, AB, JBB, TB, JB, BC, LC, MAD, MF, KMG, MADG, SH, IMVJ, JSJ,
707 JWJ, LK, RK, JAL, RM, DRN, SRP, KAP, DP, SP, YP, HQ, AR, JR, HR, RLR, AJR, EDS, AT
708 and GY participated in manual curation and contributed to subprojects and/or the Supporting
709 Information (SI). JSJ analyzed genome size. HB and JPD studied genome organization and
710 chromosome synteny. KMG and MADG conducted analyses of DNA methylation. AD, CS and
711 JW studied bacterial horizontal gene transfers. RK, AB and YP performed in vitro functional
712 characterization of PCWDEs. EDS, KH and SMG studied gene expression. LK, AMR and EDS
713 studied myrosinase and cyanogenic β-glycosidase-like sequences. RMW, JJD, SMG, DDM, AM
714 and EDS contributed to comparative genomic analyses. DDM, EDS, KH, LAK, JBB, SMG, YP
715 and SR wrote the manuscript. DDM, SR, DJC and SS organized the SI. All authors approved
716 the final manuscript.

717 **Author details**

[1]Department of Biological Sciences, University of Memphis, Memphis, TN USA 38152. [2]Feinstone Center for Genomic Research, University of Memphis, Memphis, TN USA 38152. [3]USDA, Agricultural Research Service, Center for Grain and Animal Health, Stored Product Insect and Engineering Research Unit, Manhattan, KS 66502. [4]Department of Entomology, Max Planck Institute for Chemical Ecology, 07745 Jena, Germany. [5]Department of Entomology and Center for Chemical Ecology, The Pennsylvania State University, University Park, PA, USA 16802. [6]USDA, Agricultural Research Service, Daniel K Inouye US Pacific Basin Agricultural Research Center, Tropical Crop and Commodity Protection Research Unit, Hilo, HI, USA 96720. [7]Center for Insect Science and Department of Neuroscience, University of Arizona, Tucson, AZ, USA 85721. [8]Department of Biology, University of Wisconsin Oshkosh, Oshkosh, WI, USA 54901. [9]Department of Genetic Medicine and Development and Swiss Institute of Bioinformatics, University of Geneva, Geneva 1211, Switzerland. [10]The Massachusetts Institute of Technology and The Broad Institute of MIT and Harvard, Cambridge, MA USA 02142. [11]Department of Biological Sciences, University of Illinois at Chicago, Chicago, IL USA 60607. [12]Department of Biological Sciences, University of Cincinnati, Cincinnati, OH USA 45221. [13]Department of Biology, University of Texas at Arlington, Department of Biology, Arlington, TX, USA 76019. [14]Department of Biological Sciences, North Dakota State University, Fargo, ND USA 58108. [15]Human Genome Sequencing Center, Department of Human and Molecular Genetics, Baylor College of Medicine, Houston, TX USA 77030. [16]USDA, Agricultural Research Service, Red River Valley Agricultural Research Center, Biosciences Research Laboratory, Fargo, ND USA 58102. [17]USDA, Agricultural Research Service, National Agricultural Library, Beltsville, MD USA 20705. [18]Department of Biology, University of Rochester, Rochester, NY USA 14627. [19]USDA, Agricultural Research Service, Beneficial Insects Introduction Research, Newark, DE, USA 19713. [20]Department of Biological Sciences, Wayne State University, Detroit, MI USA 48202. [21]School of Biology, Georgia Institute of Technology, Atlanta, GA, USA 30332. [22]Department of Entomology, Texas A&M University, College Station, TX USA 77843.

[23]Department of Biochemistry and Molecular Biology, Department of Computers Science and Engineering, and Department of Fisheries and Wildlife, Michigan State University, East Lansing, MI USA 48824. [24]USDA, Animal and Plant Health Inspection Service, Plant Pest and Quarantine, Center for Plant Health Science and Technology, Otis Laboratory, Buzzards Bay, MA, USA 02542. [25]Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei 10617, Taiwan. [26]Department of Biology, Indiana University, Blomington, IN, USA 47405. [27]Department of Microbiology, Immunology, and Biochemistry, University of Tennessee Health Science Center, Memphis, TN USA 38163. [28]Department of Entomology, University of Kentucky, Lexington, KY USA 40546. [29]Institute for Developmental Biology, University of Cologne, 50674 Cologne, Germany. [30]Department of Biology, Xavier University, Cincinnati, OH USA 45207. [31]Department of Entomology, University of Illinois at Urbana-Champaign, Urbana, IL USA 61801.

Author e-mail addresses

Alex C. Torson: alex.s.torson@my.ndsu.edu
Amanda Dolan: amanda.r.avery@rochester.edu
André Busch: abusch@ice.mpg.de
Andrew J. Rosendale: andrew.rosendale@uc.edu
Ann M. Ray: raya6@xavier.edu
Anna K. Childers: Anna.Bennett@ARS.USDA.GOV
Armin P. Moczek: armin@indiana.edu
Bernarda Calla: calla2@illinois.edu
Chien-Yueh Lee: chien-yueh.lee@ars.usda.gov
Christian Silva: christian.silva@rochester.edu
Christopher Childers: christopher.childers@ars.usda.gov
Dan Pers: dpers3@uic.edu
Daniel S.T. Hughes: Daniel.Hughes@bcm.edu
Dave J. Clarke: djclarke@memphis.edu
David R. Lance: David.R.Lance@aphis.usda.gov
David R. Nelson: drnelson1@gmail.com
Deanna Arsala: darsal2@uic.edu
Donna M. Muzny: donnam@bcm.edu
Duane D. McKenna: dmckenna@memphis.edu
Erin D. Scully: Erin.Scully@ARS.USDA.GOV
Evgeny M. Zdobnov: Evgeny.Zdobnov@unige.ch
George Yocum: george.yocum@ars.usda.gov
Han Lin: hlin@ars.usda.gov
HarshaVardhan Doddapaneli: doddapan@bcm.edu
Heath Blackmon: coleoguy@gmail.com

Honghu Quan: hquan2@uic.edu
Hsu Chao: hchao@bcm.edu
Hugh M. Robertson: hughrobe@uiuc.edu
Huyen Dinh: hdinh@bcm.edu
Iris M. Vargas Jentzsch: ivargasj@uni-koeln.de
J. Spencer Johnston: spencerj@TAMU.EDU
Jeffery P. Demuth: jpdemuth@uta.edu
Jeffery W. Jones: jonesjefferyw@gmail.com
Jeremy A. Lynch: jlynch42@uic.edu
Jian J. Duan: Jian.Duan@ARS.USDA.GOV
Jiaxin Qu: jiaxinqu@yahoo.com
John H. Werren: jack.werren@rochester.edu
Joseph P. Rinehart: joseph.rinehart@ars.usda.gov
Joshua B. Benoit: benoitja@ucmail.uc.edu
Julia H. Bowsher: julia.bowsher@ndsu.edu
Karl M. Glastad: karlglastad@gatech.edu
Kelli Hoover: kxh25@psu.edu
Kim C. Worley: kworley@bcm.edu
Kristen A. Panfilio: kpanfili@uni-koeln.de
Leslie A. Kuhn: kuhnl@msu.edu
Lorna Cohen: cohen36@uic.edu
Markus Friedrich: friedrichwsu@gmail.com
Michael A.D. Goodisman: michael.goodisman@biology.gatech.edu
Monica F. Poelchau: monica.poelchau@ars.usda.gov
Panagiotis Ioannidis: panos.ioannidis@gmail.com
Richard A. Gibbs: agibbs@bcm.edu
Richard Roehrdanz: richard.Roehrdanz@ars.usda.gov
Robert F. Mitchell: mitchellr@uwosh.edu
Robert M. Waterhouse: robert.waterhouse@gmail.com
Roy Kirsch: rkirsch@ice.mpg.de
Sandra L. Lee: sllee@bcm.edu
Scott M. Geib: Scott.Geib@ars.usda.gov
Seung-Joon Ahn: sahn@ice.mpg.de
Seunggwan Shin: sshin4@memphis.edu
Shannon Dugan: sdugan@bcm.edu
Shwetha C. Murali: ShwethaCanchi.Murali@bcm.edu
Stephanie Haddad: stephanyhaddad@gmail.com
Stephen Richards: stephenr@BCM.EDU
Subba R. Palli: RPALLI@UKY.EDU
Tiffany Bledsoe: nikitaur1@gmail.com
Yannick Pauchet: ypauchet@ice.mpg.de
Yi Han: yhan@bcm.edu

References

1. Hutchinson GE. Homage to Santa Rosalia, or why are there so many kinds of animals? Am Nat. 1959;93:145–59.

2. McKenna DD, Sequeira AS, Marvaldi AE, Farrell BD. Temporal lags and overlap in the diversification of weevils and flowering plants. Proc Natl Acad Sci U S A. 2009;106:7083–8.

3. McKenna DD, Wild AL, Kanda K, Bellamy CL, Beutel RG, Caterino MS, Farnum CW, Hawks DC, Ivie MA, Jameson ML, et al. The beetle tree of life reveals that Coleoptera survived end-Permian mass extinction to diversify during the Cretaceous terrestrial revolution. Syst Entomol. 2015;40:835–80.

4. Meng PS, Hoover K, Keena MA. Asian longhorned beetle (Coleoptera: Cerambycidae), an introduced pest of maple and other hardwood trees in North America and Europe. J Integr Pest Manag. 2015;6:1–13.

5. Nowak DJ, Pasek JE, Sequeira RA, Crane DE, Mastro VC. Potential effect of *Anoplophora glabripennis* (Coleoptera : Cerambycidae) on urban trees in the United States. J Econ Entomol. 2001;94:116–22.

6. Berenbaum MR. The chemistry of defense: theory and practice. Proc Natl Acad Sci U S A. 1995;92:2–8.

7. Pauchet Y, Kirsch R, Giraud S, Vogel H, Heckel DG. Identification and characterization of plant cell wall degrading enzymes from three glycoside hydrolase families in the cerambycid beetle *Apriona japonica*. Insect Biochem Mol Biol. 2014;49:1–13.

8. Pauchet Y, Wilkinson P, Chauhan R, Ffrench-Constant RH. Diversity of beetle genes encoding novel plant cell wall degrading enzymes. PLoS One. 2010;5:1–8.

9. Scully ED, Hoover K, Carlson JE, Tien M, Geib SM. Midgut transcriptome profiling of *Anoplophora glabripennis*, a lignocellulose degrading cerambycid beetle. BMC Genomics. 2013;14:850.

10. Kirsch R, Gramzow L, Theissen G, Siegfried BD, Ffrench-Constant RH, Heckel DG, Pauchet Y. Horizontal gene transfer and functional diversification of plant cell wall degrading polygalacturonases: Key events in the evolution of herbivory in beetles. Insect Biochem Mol Biol. 2014;52:33–50.

11. Ohtoko K, Ohkuma M, Moriya S, Inoue T, Usami R, Kudo T. Diverse genes of cellulase homologues of glycosyl hydrolase family 45 from the symbiotic protists in the hindgut of the termite *Reticulitermes speratus*. Extremophiles. 2000;4:343–9.

12. Geib SM, Filley TR, Hatcher PG, Hoover K, Carlson JE, Jimenez-Gasco Mdel M, Nakagawa-Izumi A, Sleighter RL, Tien M. Lignin degradation in wood-feeding insects. Proc Natl Acad Sci U S A. 2008;105:12932–7.

13. Cunningham CB, Ji L, Wiberg RA, Shelton J, McKinney EC, Parker DJ, Meagher RB, Benowitz KM, Roy-Zokan EM, Ritchie MG, et al. The genome and methylome of a beetle with complex social behavior, *Nicrophorus vespilloides* (Coleoptera: Silphidae). Genome Biol Evol. 2015;7:3383–9.

14. Keeling CI, Yuen MM, Liao NY, Docking TR, Chan SK, Taylor GA, Palmquist DL, Jackman SD, Nguyen A, Li M, et al. Draft genome of the mountain pine beetle, *Dendroctonus ponderosae* Hopkins, a major forest pest. Genome Biol. 2013;14**:**R27.

15. Richards S, Gibbs RA, Weinstock GM, Brown SJ, Denell R, Beeman RW, Gibbs R, Beeman RW, Brown SJ, Bucher G, et al. The genome of the model beetle and pest *Tribolium castaneum*. Nature. 2008;452**:**949–55.

16. Vega FE, Brown SM, Chen H, Shen E, Nair MB, Ceja-Navarro JA, Brodie EL, Infante F, Dowd PF, Pain A. Draft genome of the most devastating insect pest of coffee worldwide: the coffee berry borer, *Hypothenemus hampei*. Sci Rep. 2015;5**:**12525.

17. Hanrahan SJ, Johnston JS. New genome size estimates of 134 species of arthropods. Chromosome Res. 2011;19**:**809–23.

18. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinf. 2011;12:491.

19. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31**:**3210–2.

20. Kriventseva EV, Tegenfeldt F, Petty TJ, Waterhouse RM, Simao FA, Pozdnyakov IA, Ioannidis P, Zdobnov EM. OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. Nucleic Acids Res. 2015;43**:**D250–6.

21. Haddad S, McKenna DD. Phylogeny and evolution of superfamily Chrysomeloidea (Coleoptera: Cucujiformia). Syst Entomol. In Press.

22. McKenna DD. Molecular Phylogenetics and Evolution of Coleoptera**.** In: Beutel RG, Leschen RAB (editors). *Handbook of Zoology Volume IV Arthropoda: Insecta Part 38 Coleoptera, Beetles Volume 3: Morphology and Systematics (Phytophaga).* Berlin: Walter de Gruyter; 2014. 1–10.

23. Benoit, JB, Adelman ZN, Reinhardt K, Dolan A, Poelchau M, Jennings EC, Szuter EM, Hagan RW, Gujar H, Shukla JN, Zhu F, Mohan M, Nelson DR, Rosendale AJ, Derst C, Resnik V, Wernig S, Menegazzi P, Wegener C, Peschel N, Hendershot JM, Blenau W, Predel R, Johnston PR, Ioannidis P, Waterhouse RM, Nauen R., Schorn C, Ott MC, Maiwald F, Johnston JS, Gondhalekar AD, Scharf ME, Peterson BF, Raje KR, Hottel BA, Armisén D, Crumière AJ, Refki PN, Santos ME, Sghaier E, Viala S, Khila A, Ahn SJ, Childers C, Lee CY, Lin H, Hughes DS, Duncan EJ, Murali SC, Qu J, Dugan S, Lee SL, Chao H, Dinh H, Han Y, Doddapaneni H, Worley KC, Muzny DM, Wheeler D, Panfilio KA, Vargas Jentzsch IM, Vargo EL, Booth W, Friedrich M, Weirauch MT, Anderson MA, Jones JW, Mittapalli O, Zhao C, Zhou JJ, Evans JD, Attardo GM, Robertson HM, Zdobnov EM, Ribeiro JM, Gibbs RA, Werren JH, Palli SR, Schal C, Richards S. Unique features of a global human ectoparasite identified through sequencing of the bed bug genome. Nat Commun. 2016;7:10165. doi: 10.1038/ncomms10165.

24. Zhao C, Escalante LN, Benatti TR, Qu J, Chellapilla S, Waterhouse RM, Wheeler D, Anderson MN, Bao R, Batterton M, Behura S, Blankenburg KP, Caragea D, Carolan JC, Chen H, Coyle M, El-Bohssini M, Francisco L, Friedman M, Gibbs RA, Gil N, Grace T,

909 Grimmelikhuijzen CJP, Han Y, Hauser F, Herndon N, Holder M, Jackson L, Javaid M,
910 Jhangiani SN, Johnson AJ, Kalra D, Korchina V, Kovar C, Lara F, Lee SL, Liu X, Löfstedt
911 C, Mata R, Mathew T, Muzny DM, Nagar S, Nazareth L, Ngo D, Okwuonu G, Ongeri F,
912 Perales L, Pu L-L, Robertson HM, Scherer SE, Shreve JT, Simmons D, Subramanyam S,
913 Thornton RL, Viswami V, Weissenberger GM, Williams CE, Worley KC, Zhu D, Zhu Y,
914 Harris MO, Shukle RH, Werren JH, Zdobnov EM, Chen M-S, Brown SJ, Stuart JJ,
915 Richards S. A Massive Expansion of Effector Genes Underlies Gall-Formation in the
916 Wheat Pest Mayetiola destructor. Curr Biol. 2015;25(5):613-620.
917 doi:10.1016/j.cub.2014.12.057.

918 25. Davison A, Blaxter M. Ancient origin of glycosyl hydrolase family 9 cellulase genes.
919 Mol Biol Evol. 2005;22**:**1273–84.

920 26. Chang CJ, Wu CP, Lu SC, Chao AL, Ho THD, Yu SM, Chao YC. A novel exo-cellulase
921 from white spotted longhorn beetle (*Anoplophora malasiaca*). Insect Biochem Mol
922 Biol. 2012;42**:**629–36.

923 27. Eyun SI, Wang HC, Pauchet Y, Ffrench-Constant RH, Benson AK, Valencia-Jimenez A,
924 Moriyama EN, Siegfried BD. Molecular evolution of glycoside hydrolase genes in the
925 western corn rootworm (*Diabrotica virgifera virgifera*). PLoS One. 2014;9**:**e102818.

926 28. Palomares-Rius JE, Hirooka Y, Tsai IJ, Masuya H, Hino A, Kanzaki N, Jones JT, Kikuchi
927 T. Distribution and evolution of glycoside hydrolase family 45 cellulases in
928 nematodes and fungi. BMC Evol Biol. 2014;14**:**69.

929 29. Scully ED, Geib SM, Hoover K, Tien M, Tringe SG, Barry KW, Glavina del Rio T,
930 Chovatia M, Herr JR, Carlson JE. Metagenomic profiling reveals lignocellulose
931 degrading system in a microbial community associated with a wood-feeding beetle.
932 PLoS One. 2013;8**:**e73827.

933 30. Choi JH, Kijimoto T, Snell-Rood E, Tae H, Yang Y, Moczek AP, Andrews J. Gene
934 discovery in the horned beetle *Onthophagus taurus*. BMC Genomics. 2010;11**:**703.

935 31. Fujita K, Shimomura K, Yamamoto K, Yamashita T, Suzuki K. A chitinase structurally
936 related to the glycoside hydrolase family 48 is indispensable for the hormonally
937 induced diapause termination in a beetle. Biochem Bioph Res Co. 2006;345**:**502–7.

938 32. Scully ED, Geib SM, Carlson JE, Tien M, McKenna D, Hoover K. Functional genomics
939 and microbiome profiling of the Asian longhorned beetle (*Anoplophora*
940 *glabripennis*) reveal insights into the digestive physiology and nutritional ecology of
941 wood feeding beetles. BMC Genomics. 2014;15**:**1096.

942 33. Hopkins RJ, van Dam NM, van Loon JJ. Role of glucosinolates in insect-plant
943 relationships and multitrophic interactions. Annu Rev Entomol. 2009;54**:**57–83.

944 34. Møller BL. Functional diversifications of cyanogenic glucosides. Curr Opin Plant Biol.
945 2010;13**:**338–47.

946 35. Geib SM, Jimenez-Gasco Mdel M, Carlson JE, Tien M, Hoover K. Effect of host tree
947 species on cellulase activity and bacterial community composition in the gut of
948 larval Asian longhorned beetle. Environ Entomol. 2009;38**:**686–99.

36. Robertson JA, Slipinski A, Moulton M, Shockley FW, Giorgi A, Lord NP, Mckenna DD, Tomaszewska W, Forrester J, Miller KB, et al. Phylogeny and classification of Cucujoidea and the recognition of a new superfamily Coccinelloidea (Coleoptera: Cucujiformia). Syst Entomol. 2015;40:745–78.

37. Celorio-Mancera Mde L, Carl Greve L, Teuber LR, Labavitch JM. Identification of endo- and exo-polygalacturonase activity in *Lygus hesperus* (Knight) salivary glands. Arch Insect Biochem Physiol. 2009;70:122–35.

38. Magrane M, UniProt Consortium. UniProt Knowledgebase: a hub of integrated protein data. Database. 2011;2011:bar009.

39. Beran F, Pauchet Y, Kunert G, Reichelt M, Wielsch N, Vogel H, Reinecke A, Svatos A, Mewis I, Schmid D, et al. *Phyllotreta striolata* flea beetles use host plant defense compounds to create their own glucosinolate-myrosinase system. Proc Natl Acad Sci U S A. 2014;111:7349–54.

40. Tóth M, Csonka E, Bartelt RJ, Cosse AA, Zilkowski BW. Similarities in pheromonal communication of flea beetles *Phyllotreta cruciferae* Goeze and *Ph*. *vittula* Redtenbacher (Coleoptera, Chrysomelidae). J Appl Entomol. 2012;136:688–97.

41. Fahey JW, Zalcmann AT, Talalay P. The chemical diversity and distribution of glucosinolates and isothiocyanates among plants. Phytochem. 2001;56(1):5-51.

42. Ayayee P, Rosa C, Ferry JG, Felton G, Saunders M, Hoover K. Gut microbes contribute to nitrogen provisioning in a wood-feeding cerambycid. Environ Entomol. 2014;43:903–12.

43. Ayayee PA, Jones SC, Sabree ZL. Can C-13 stable isotope analysis uncover essential amino acid provisioning by termite-associated gut microbes? Peerj. 2015;3:e1218.

44. Scully ED, Hoover K, Carlson J, Tien M, Geib SM. Proteomic analysis of *Fusarium solani* isolated from the Asian longhorned beetle, *Anoplophora glabripennis*. PLoS One. 2012;7:e32990.

45. Qiu HF, Geng A, Zhu DC, Le YL, Wu J, Chow NW, Wu JHD, Sun JZ. Purification and characterization of a hemocyanin (Hemo1) with potential lignin-modification activities from the wood-feeding termite, *Coptotermes formosanus* Shiraki. Appl Biochem Biotech. 2015;175:687–97.

46. King AJ, Cragg SM, Li Y, Dymond J, Guille MJ, Bowles DJ, Bruce NC, Graham IA, McQueen-Mason SJ. Molecular insight into lignocellulose digestion by a marine isopod in the absence of gut microbes. Proc Natl Acad Sci U S A. 2010;107:5345–5350.

47. Keeling CI, Henderson H, Li M, Dullat HK, Ohnishi T, Bohlmann J. CYP345E2, an antenna-specific cytochrome P450 from the mountain pine beetle, *Dendroctonus ponderosae* Hopkins, catalyses the oxidation of pine host monoterpene volatiles. Insect Biochem Mol Biol. 2013;43:1142–51.

48. Qiu XH, Pan J, Li M, Li Y. PCR-RFLP methods for detection of insecticide resistance-associated mutations in the house fly (*Musca domestica*). Pestic Biochem Phys. 2012;104:201–5.

49. Wang XH, Fang XD, Yang PC, Jiang XT, Jiang F, Zhao DJ, Li BL, Cui F, Wei JN, Ma CA, et al. The locust genome provides insight into swarm formation and long-distance flight. Nat Commun. 2014;5:1–9.

50. Ahn SJ, Vogel H, Heckel DG. Comparative analysis of the UDP-glycosyltransferase multigene family in insects. Insect Biochem Mol Biol. 2012;42:133–47.

51. Jongsma MA, Bakker PL, Peters J, Bosch D, Stiekema WJ. Adaptation of *Spodoptera exigua* larvae to plant proteinase-inhibitors by induction of gut proteinase activity insensitive to inhibition. Proc Natl Acad Sci U S A. 1995;92:8041–5.

52. Andersson MN, Grosse-Wilde E, Keeling CI, Bengtsson JM, Yuen MMS, Li M, Hillbur Y, Bohlmann J, Hansson BS, Schlyter F. Antennal transcriptome analysis of the chemosensory gene families in the tree killing bark beetles, *Ips typographus* and *Dendroctonus ponderosae* (Coleoptera: Curculionidae: Scolytinae). BMC Genomics. 2013;14:198.

53. Mitchell RF, Hughes DT, Luetje CW, Millar JG, Soriano-Agaton F, Hanks LM, Robertson HM. Sequencing and characterizing odorant receptors of the cerambycid beetle *Megacyllene caryae*. Insect Biochem Mol Biol. 2012;42:499–505.

54. Zhang A, Oliver JE, Chauhan K, Zhao B, Xia L, Xu Z. Evidence for contact sex recognition pheromone of the Asian longhorned beetle, *Anoplophora glabripennis* (Coleoptera: Cerambycidae). Naturwissenschaften. 2003;90:410–3.

55. Hoover K, Keena M, Nehme M, Wang SF, Meng P, Zhang AJ. Sex-specific trail pheromone mediates complex mate finding behavior in *Anoplophora glabripennis*. J Chem Ecol. 2014;40:169–80.

56. Koh TW, He Z, Gorur-Shandilya S, Menuz K, Larter NK, Stewart S, Carlson JR. The *Drosophila* IR20a clade of ionotropic receptors are candidate taste and pheromone receptors. Neuron. 2014;83:850–65.

57. Engsontia P, Sanderson AP, Cobb M, Walden KKO, Robertson HM, Brown S. The red flour beetle's large nose: An expanded odorant receptor gene family in *Tribolium castaneum*. Insect Biochem Mol Biol. 2008;38:387–97.

58. Jackowska M, Bao R, Liu Z, McDonald EC, Cook TA, Friedrich M. Genomic and gene regulatory signatures of cryptozoic adaptation: Loss of blue sensitive photoreceptors through expansion of long wavelength-opsin expression in the red flour beetle *Tribolium castaneum*. Front Zool. 2007;4:24.

59. Friedrich M, Wood EJ, Wu M. Developmental evolution of the insect retina: insights from standardized numbering of homologous photoreceptors. J Exp Zool B Mol Dev Evol. 2011;316:484–99.

60. Marrone V, Piscopo M, Romano G, Ianora A, Palumbo A, Costantini M. Defensome against toxic diatom aldehydes in the sea urchin *Paracentrotus lividus*. PLoS One. 2012;7:e31750.

61. Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A. 2011;108:1513–8.

62. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinf. 2011;12:491.

63. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 2007;23:1061–7.

64. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 2008;24:637–44.

65. Korf I. Gene finding in novel genomes. BMC Bioinf. 2004;14:59.

66. Lee E, Helt GA, Reese JT, Munoz-Torres MC, Childers CP, Buels RM, Stein L, Holmes IH, Elsik CG, Lewis SE. Web Apollo: a web-based genomic annotation editing platform. Genome Biol. 2013;14:R93.

67. Poelchau M, Childers C, Moore G, Tsavatapalli V, Evans J, Lee CY, Lin H, Lin JW, Hackett K. The i5k Workspace@NAL--enabling genomic data access, visualization and curation of arthropod genomes. Nucleic Acids Res. 2015;43:D714–9.

68. Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH. JBrowse: a next-generation genome browser. Genome Res. 2009;19:1630–8.

69. Wheeler D, Redding AJ, Werren JH. Characterization of an Ancient Lepidopteran Lateral Gene Transfer. PLoS ONE 2013; 8(3): e59262. doi:10.1371/journal.pone.0059262

70. Keena MA. Pourable artificial diet for rearing *Anoplophora glabripennis* (Coleoptera : Cerambycidae) and methods to optimize larval survival and synchronize development. Ann Entomol Soc Am. 2005;98:536–47.

71. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25:1105–11.

72. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31:166–9.

73. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.

74. Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, Dufayard JF, Guindon S, Lefort V, Lescot M, et al. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res. 2008;36:W465–9.

75. Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 2003;19:1572–4.

76. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. 2011;28:2731–9.

77. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30:1236–40.

78. Haack RA, Law KR, Mastro VC, Ossenbruggen HS, Raimo BJ. New York's battle with the Asian long-horned beetle. J Forest. 1997;95:11–5.

79. Haack RA, Herard F, Sun JH, Turgeon JJ. Managing invasive populations of Asian longhorned beetle and citrus longhorned beetle: a worldwide perspective. Annu Rev Entomol. 2010;55:521–46.

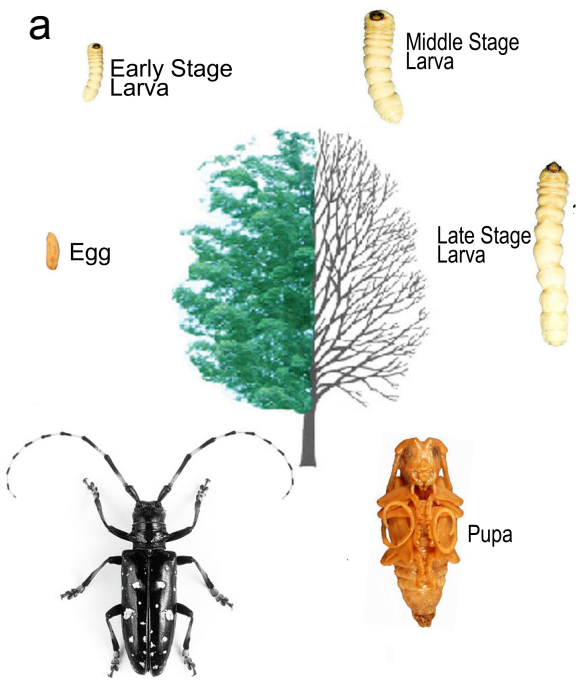
80. Hu JF, Angeli S, Schuetz S, Luo YQ, Hajek AE. Ecology and management of exotic and endemic Asian longhorned beetle *Anoplophora glabripennis*. Agr Forest Entomol. 2009;11:359–75.

81. Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri T, Beutel RG, et al. Phylogenomics resolves the timing and pattern of insect evolution. Science. 2014;346:763–7.

1096 Fig. 1

1097 *A. glabripennis*, the Asian longhorned beetle, is a high profile invasive pest species capable of
1098 inflicting severe damage on its hosts, which include many important orchard, ornamental and
1099 forest tree species. a Life cycle (adapted from Michael Bohne, used with permission; image of
1100 adult female courtesy of Barbara Strnadova, used with permission). b Wood dissected to
1101 expose feeding *A. glabripennis* larva (image courtesy of Kelli Hoover, used with permission). c,
1102 d Adult *A. glabripennis* (images courtesy of Damon Crook, used with permission). Early stage
1103 larvae are specialized wood-borers, feeding in galleries under the bark of host trees (in the
1104 subcortical tissue and phloem). Larger, later stage larvae tunnel deep into the heartwood
1105 (mature xylem) of their hosts, where they continue feeding and complete development [4].
1106 Adults are comparatively short-lived external feeders, consuming small amounts of tissue from
1107 host leaves and twigs. *A. glabripennis* is broadly polyphagous on woody angiosperms. It is
1108 native to eastern Asia, but has recently become established in several countries in North
1109 America, Europe, and beyond, via solid wood packing material. *A. glabripennis* is a globally
1110 significant pest, whose economic impact in the U.S. alone, if uncontrolled, has been
1111 conservatively estimated at $889[1] billion [5]. It is capable of attacking both healthy and
1112 susceptible trees [78] and is broadly polyphagous, feeding on at least 100 species of woody
1113 angiosperms worldwide [4, 79, 80].

[1] Adjusted for inflation May 2016.



1115

1116

1117

1118

1119

Fig. 2

Phylogenetic relationships and estimates of completeness among the 15 insect genomes studied. a Maximum likelihood (ML) phylogenetic tree based on amino acid sequences from 523 orthologs. All nodes have 100% ML bootstrap support. The tree was rooted with *Zootermopsis nevadensis*. Asterisks indicate genomes that were sequenced via i5k and are analyzed herein for the first time. Estimated divergence times are shown along branches subtending the crown group nodes they refer to, and were obtained from [3] for Coleoptera, and [81] for all others. b The completeness of both genome assemblies and official gene sets (OGSs) of each of the insects were assessed using 2,675 arthropod benchmarking universal single-copy orthologs (BUSCOs). For each species, the bottom bar in the histogram shows the OGS-based results, whereas the top bar shows the genome-based results. Images courtesy of: Nicolas Gompel (DMELA), Scott Bauer/USDA-ARS (MDEST), Chris Lewis (PXYLO), Didier Decouens (DPLEX), Barbara Strnadova (AGLAB), Klaus Bolte (DPOND), Kohichiro Yoshida (TCAST), Rafal Celadyn (OTAUR), PA Dept. of CNR (APLAN), Elizabeth Cash (NVITR), Gary McClellan (AMELL), John & Kendra Abbott/Abbott Nature Photography (PHUMA), Sandy Rae (APISU), Don Loarie (ZNEVA).

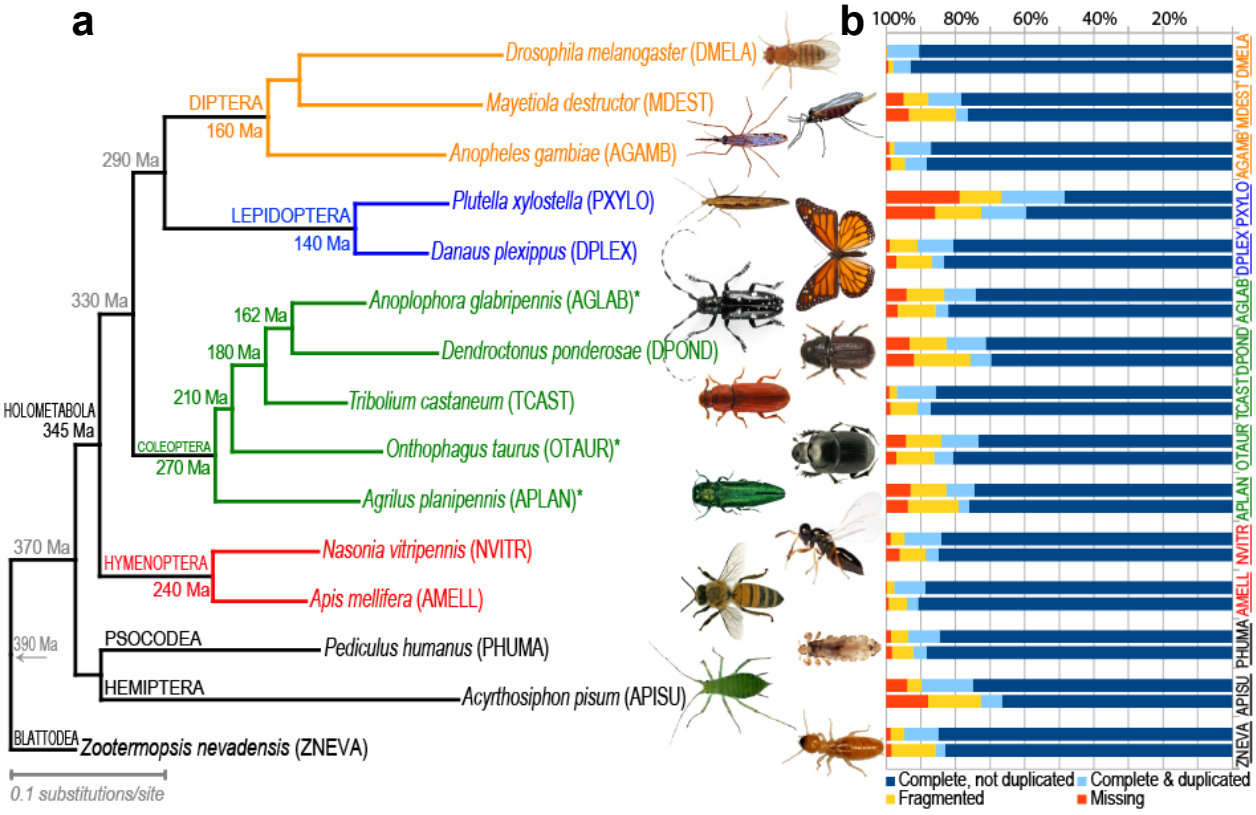


Fig. 3

Orthology and homology assignments of *A. glabripennis* genes with those of 14 other insect species. A conserved core of about 5,000 orthologs per species (5,029 *A. glabripennis* genes) is maintained in orthologous groups with gene members from all 15 species, about half with a single gene (dark purple) and half with multiple copies (light purple). A variable fraction of genes is less well maintained but still widespread (green) with orthologs in at least two species from each of the three sets of insect species. Lineage-restricted genes include those with orthologs only within each set (pink), with recognizable homology to other arthropod genes (white) or their own genes (cyan), or without any significant homology (gray). The numbers of orthologous groups (OGs) are shown with area-proportional boxes for the set intersections and the lineage-restricted orthologs. See Methods for orthology classification details.

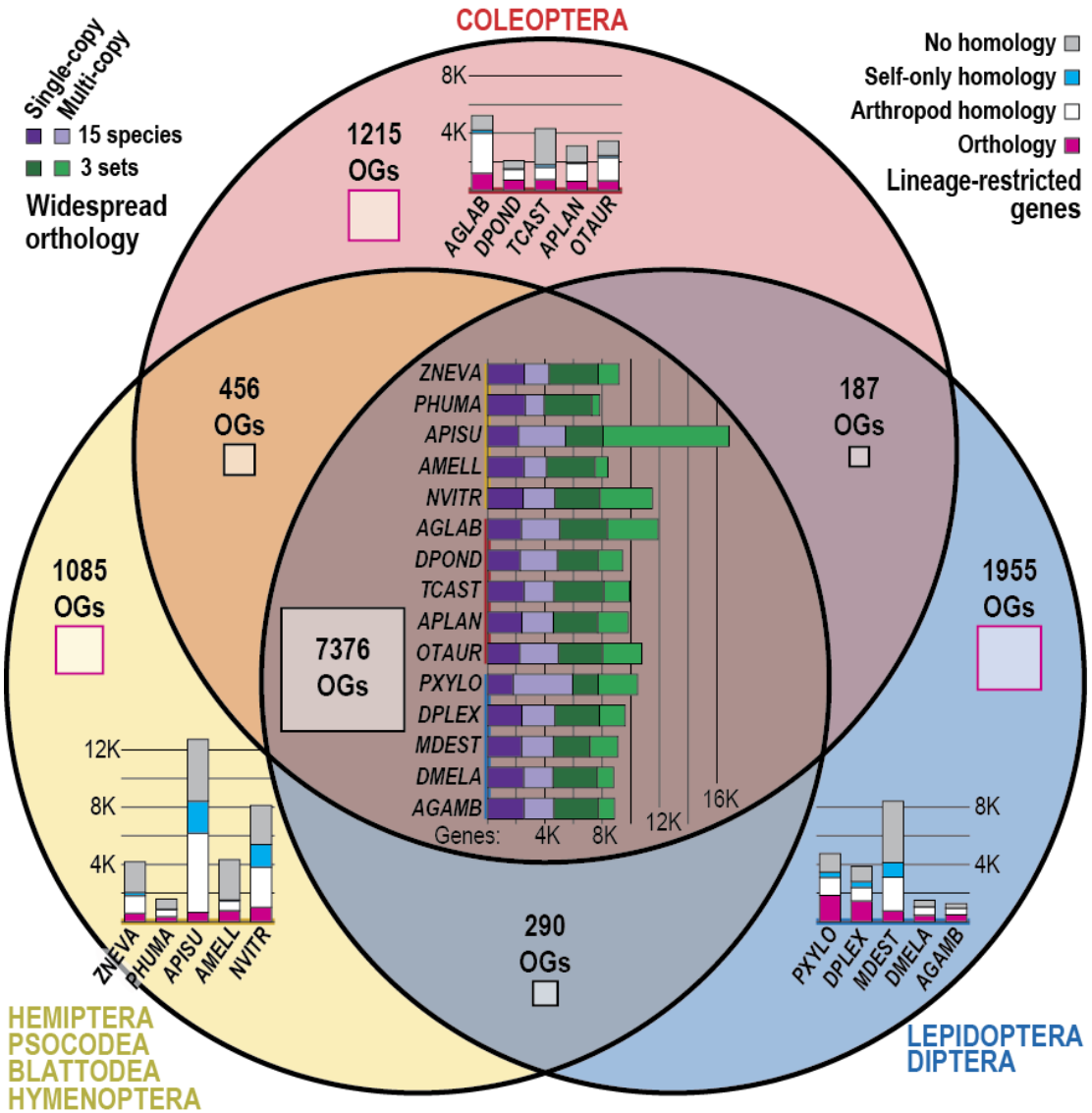


Fig. 4

Sub-family sizes for gycoside hydrolases found in the genome sequences of 15 insect species, including *A. glabripennis*. Species with the maximum gene count for each are indicated with a white asterisk. Among the examined species, *A. glabripennis* showed the most genes with matches to GH domains, the majority of which were found as multi-copy orthologs. This elevated gene count was mainly due to GH Family 1 (IPR001360), members of which exhibit beta-glucosidase, beta-galactosidase, 6-phospho-beta-galactosidase, 6-phospho-beta-glucosidase, lactase-phlorizin hydrolase, beta-mannosidase, and myrosinase activities. Uniquely among the examined species, 6 *A. glabripennis* genes matched GH Family 5 (IPR001547), also known as cellulase family A, whose members exhibit endoglucanase, beta-mannanase, exo-1,3-glucanase, endo-1,6-glucanase, xylanase, and endoglycoceramidase activities. *A. glabripennis* also had 2 matches to the GH Family 45 (IPR000334, endoglucanase activity) also known as cellulase family K, which was also found in *D. ponderosae* (9 copies). Members of GH Family 28 (IPR000743) are pectinases that exhibit polygalacturonase and rhamnogalacturonase activities, and had matches to 16 genes in *A. glabripennis* (18 were identified by manual annotation; 19 were reported in [8]), 16 in *D. ponderosae* and 7 in *A. planipennis* (50 were manually annotated).

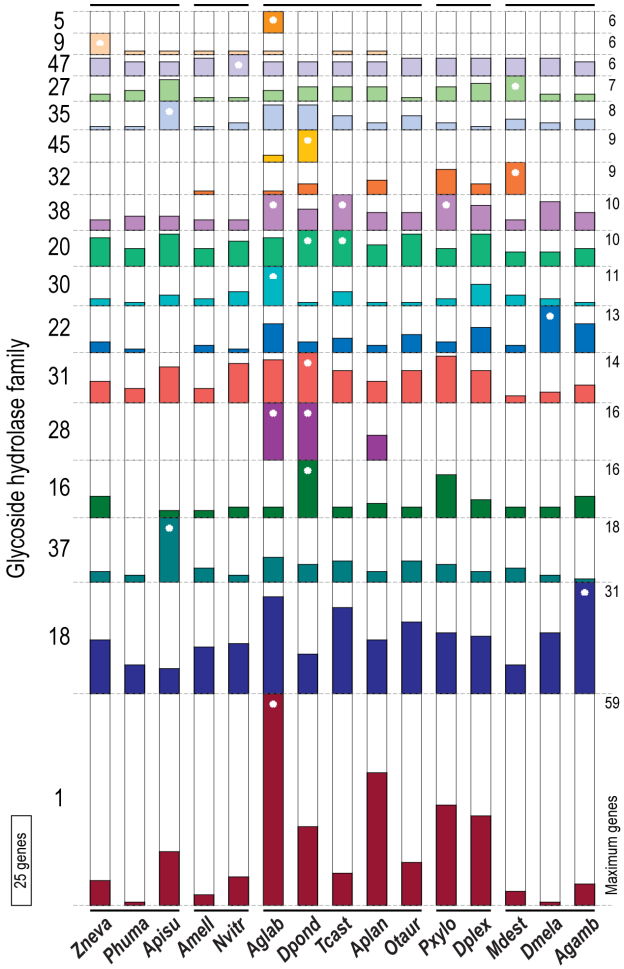


Fig. 5

Heatmap showing expression levels from *A. glabripennis* gycoside hydrolase genes with putative involvement in plant cell wall degradation. Logfold changes in expression levels in genes collected from *A. glabripennis* larvae feeding in the wood of living sugar maple trees are shown versus those from larvae feeding on a nutrient rich artificial diet. While the expression levels of GH genes were variable, several were significantly upregulated in larvae feeding in the wood of living sugar maple.

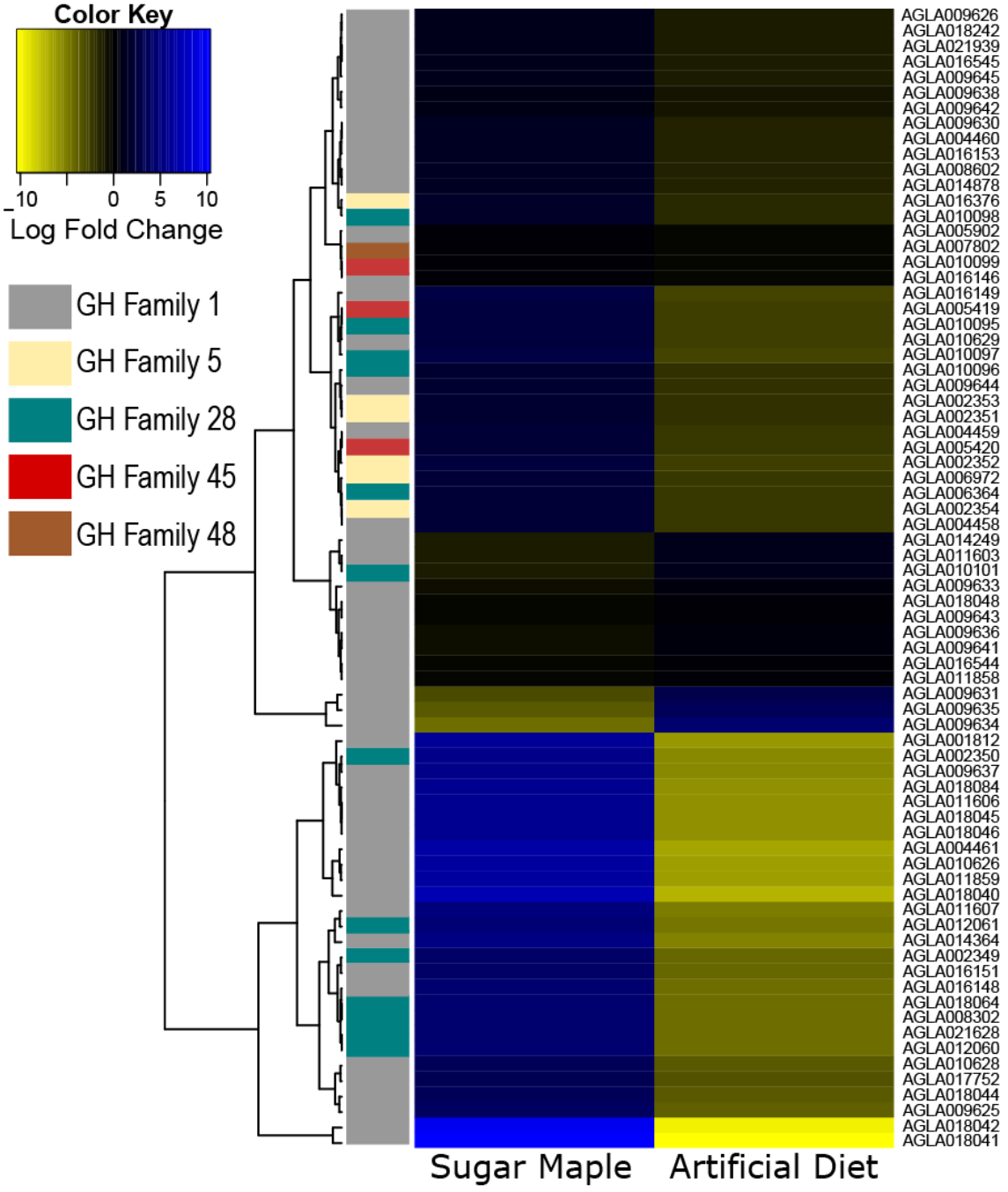


Fig. 6

Phylogenetic tree showing *A. glabripennis* (color) and *T. castaneum* (black) UDP-glycosyltransferases (UGTs), reconstructed from amino acid (aa) sequences using ML inference (MLBS values <70 not shown). Each gene belonging to UGT352, UGT321, and UGT328 consists of 4 exons, with the long first exon (ca. 810 aa) followed by three short exons. Each member of UGT323, UGT324, and UGT325 is composed of 4 exons with the short first exon (ca. 200 aa) and the long second exon (ca. 800 aa) followed by two short exons. UGT312 and UGT353 (AglaUGT_63 and _64) consistently contain genes with 5 exons. Scaffold 72 is shown to illustrate the tandem arrangement typical of *A. glabripennis* UGTs. Photo of *A. glabripennis* courtesy of Barbara Strnadova, used with permission.