Fall 9-25-2013

THE ORIGIN AND MOLECULAR EVOLUTION OF TWO MULTIGENE FAMILIES: G-PROTEIN COUPLED RECEPTORS AND GLYCOSIDE HYDROLASE FAMILIES

Seong-il Eyun
University of Nebraska - Lincoln, seyun2@unl.edu

THE ORIGIN AND MOLECULAR EVOLUTION OF TWO MULTIGENE FAMILIES: G-PROTEIN COUPLED RECEPTORS AND GLYCOSIDE HYDROLASE FAMILIES

by

Seong-il Eyun

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Biological Sciences

Under the Supervision of Professor Etsuko Moriyama

Lincoln, Nebraska

August, 2013

THE ORIGIN AND MOLECULAR EVOLUTION OF TWO MULTIGENE FAMILIES: G-PROTEIN COUPLED RECEPTORS AND GLYCOSIDE HYDROLASE FAMILIES

Seong-il Eyun, Ph.D.

University of Nebraska, 2013

Advisor: Etsuko Moriyama

Multigene family is a group of genes that arose from a common ancestor by gene duplication. Gene duplications are a major driving force of new function acquisition. Multigene family thus has a fundamental role in adaptation. To elucidate their molecular evolutionary mechanisms, I chose two multigene families: chemosensory receptors and glycoside hydrolases. I have identified complete repertoires of trace amine-associated receptors (TAARs), a member of chemosensory receptors, from 38 metazoan genomes. An ancestral-type TAAR emerged before the divergence between gnathostomes (jawed vertebrates) and sea lamprey (jawless fish). Primary amine detecting TAARs (TAAR1-4) are found to be older and have evolved under strong functional constraints. In contrast, tertiary amine detectors (TAAR5-9) emerged later, experienced higher rates of gene duplications, and experienced positive selection that could have affected ligand-binding activities and specificities. Expansions of tertiary amine detectors must have played important roles in terrestrial adaptations of therian mammals. During the primate evolution, TAAR gene losses are found to be a major trend. Relaxed selective constraints found in primate lineages of TAARs support dispensability of these primate genes. Reduced predator exposures owing to the start of arboreal life by ancestoral primates may attribute to this change. For another type of multigene family, glycoside hydrolase (GH) genes were identified in the western corn rootworm, *Diabrotica virgifera virgifera* (Coleoptera: Chrysomelidae). Three GH family genes (GH45, GH48, and GH28) were found only in two coleopteran superfamilies (Chrysomeloidea and Curculionoidea) among insects (except for hemipteran GH28s), indicating their origin from horizontal gene transfer (HGT). Several independent HGTs in fungi and other insects were also detected. Two multigene families in this study are characterized with frequent gene duplications and losses, the birth-and-death process. A high rate of HGTs found in the GH family gene evolution must have accelerated functional evolution. In conclusion, this study showed that birth-and-death process, positive selection, and HGTs, all play a critical role in driving the evolution of multigene families and allow organismal adaptation to novel environmental niches.

ACKNOWLEDGEMENTS

First, I am very grateful to my advisor, Dr. Etsuko Moriyama and the members of Moriyama Lab in which I have interacted. They have provided a continual source of knowledge, guidance, and invaluable assistance in all phases of my doctorial career.

I would like to sincere thank my committee members; Dr. Lawrence G. Harshman, Dr. Blair D. Siegfried, Dr. Jay F. Storz, and Dr. Hideaki Moriyama for their expert knowledge and comments.

I also thank the financial and professional support from the Biological Sciences program at University of Nebraska-Lincoln.

Finally, I want to thank my parents. Thank my wife, Eun Jeong Kim for her continual love, encouragement, and support during the preparation of this dissertation. Thank you my son, Ian Eyun for giving me lovely kisses and hugs.

TABLE OF CONTENTS

LIST OF TABLES

Chapter 1

Chapter 2

Chapter 3

Chapter 4

LIST OF FIGURES

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Abbreviations

cAMP	cyclic adenosine monophosphate
CASR	calcium sensing receptor
CRs	chemosensory receptors
DNA	deoxyribonucleic acid
EC number	Enzyme Commission number
EST	Expressed Sequence Tag
FPR	formyl peptide receptor
GABA	gamma-amino-butyric acid
GDP	guanosine diphosphate
GH	glycoside hydrolase
GPCR	G-protein-coupled receptor
G-protein	GTP-binding protein
GTP	guanosine triphosphate
HGT	horizontal gene transfer
LGT	lateral gene transfer
MOE	main olfactory epithelium
ORs	olfactory receptors
ORF	open reading frame
RGS	regulator of G-protein signaling
RNA	ribonucleic acid
TAAR	trace amine-associated receptor
T1R	Taste receptor type 1
T2R	Taste receptor type 2
TM	transmembrane
PEA	2-phenylethylamine
V1R	vomeronasal receptor type 1
V2R	vomeronasal receptor type 2

Chapter 1

Introduction

1.1 Overview of multigene family and objectives

1.1.1 Gene duplication and multigene family

A multigene family is a group of genes that have descended from a common ancestral gene and therefore have similar functions and similar DNA sequences (Li 1997). A multigene family arises essentially from gene duplication. Gene duplication was probably first been observed in *Drosophila melanogaster*. Bridges (1936) observed a different banding pattern at the region 16A of chromosome X between wild-type and *Bar* mutants. In the mutants, certain banding patterns were “duplicated” indicating a potential role of gene duplication. Later, Ingram (1961) suggested that the myoglobin and hemoglobins α, β, γ, and δ form a family of homologous proteins and they are related to each other by gene duplication events. The term superfamily, a group of mutigene family, was used by Dayhoff (1978) in order to delineate between closely related and distantly related proteins.

Duplicated genes can be lost or fixed. Because they generate functional redundancy, the majority of duplicated genes may become "pseudogenes", nonfunctional sequences of genomic DNA originally derived from functional genes (Jacq et al. 1977; Vanin 1985), and are either unexpressed or functionless (Lynch et al. 2001). The process of pseudogenization can be started through neutral evolution when changes in the genetic background or environment render a formerly useful gene worthless (Li et al. 1981; Balakirev and Ayala 2003). Or it may sometimes occur through positive selection when a previously useful gene becomes harmful to an organism and pseudogenization of such a gene is adaptive (Jeffery et al. 2003; Zhang 2008). Wang et al. (2006) demonstrated the adaptive pseudogenization in humans. They showed that the CASPASE12 gene, a cysteine-aspartic acid protease (caspase) protein participating in inflammatory and innate immune response to endotoxins, is functional in all mammals, but in human this gene became a pseudogene. The functional gene is likely to become deleterious to humans as the null allele is known to be associated with a reduced incidence and mortality of severe sepsis.

Although many duplicated genes are deleted from the genome, some are maintained. The presence of duplicated copies of genes may be beneficial simply because extra amounts of protein or RNA products can be provided (Zhang 2003). Ohno (1970) proposed that gene duplication and subsequent functional divergence of duplicated genes are the most important mechanisms for the evolution of novel gene functions. The following two models can be considered. Neofunctionalization, an adaptive process where one copy mutates into a function that was not present in the pre-duplication gene, is one mechanism that can lead to the retention of both copies. Subfunctionalization, as a neutral process where the two copies partition the ancestral function, has been proposed as an alternative mechanism driving duplicated gene retention in organisms with small effective population sizes (Rastogi and Liberles 2005). Zhang (2003) reviewed comparative genomic studies demonstrating these mechanisms by which duplicate genes diverge in function and contribute to evolution.

1.1.2 Concerted evolution and the birth-and-death model

In early molecular evolutionary studies (before 1970), as shown in hemoglobin genes mentioned above, multigene families were thought to have diverged gradually as the duplicate genes acquired new gene functions (Nei and Kumar 2000). This mode of evolution is called “divergent evolution”. According to this model, if gene duplication preceded the speciation, the sequence difference between duplicated genes within the same species is expected to be as large as those between the different species. However, unexpectedly high sequence similarities within species were reported, and it could not be explained by this model of evolution (Brown and Sugimoto 1973). This suggested that the member genes or nucleotide sequences within a repetitive family do not evolve independently of each other (reviewed in Elder and Turner 1995). The molecular process that leads to homogenization of DNA sequences belonging to a given repetitive family is called “concerted evolution” (Zimmer et al. 1980). Numerous examples of concerted evolution of multigene families have been found, including the 5S DNA family in *Xenupus* (Brown and Sugimoto 1973), the γ-globin genes in primates (Jeffreys 1979), and the chorion multigene family in the silk moth (Hibner et al. 1991).

Later, phylogenetic analyses of the major histocompatibility complex genes and other immune system genes such as immunoglobulin showed a quite different evolutionary pattern and a new model called “birth-and-death evolution” was proposed (Nei and Hughes 1992). With the birth-and-death model, new genes are created by gene duplication and some are retained in the genome for a long time as functional genes, whereas other genes become nonfunctional or eliminated from the genome (Nei and Rooney 2005). Many studies have shown that ribosomal RNA genes, highly conserved histone genes, ubiquitin genes, and chemosensory receptor genes are subject to this type of evolution (Nei and Rooney 2005; Nei et al. 2008).

1.1.3 Horizontal gene transfer

Horizontal gene transfer (HGT, also known as lateral gene transfer, LGT) is the transfer of genetic material between different species. HGTs have been discovered widely in bacteria, protists, fungi, and plants (Syvanen 2012). It is distinct from the normal mode of transmission from parents to offspring, which is commonly known as vertical transfer. Syvanen theorized that HGTs are likely a major evolutionary force because the HGT events have the potential to provide novel functions to animals, allowing adaptation to novel niches, and affect their evolution (Syvanen 1984; Syvanen 1985). A number of such HGT examples involved with eukaryotes are now documented (Keeling 2009; Dunning Hotopp 2011). For example, aphids are the only known animal capable of synthesizing their own carotenoids, and their carotenoid biosynthesis enzymes are derived from fungal genes (Moran and Jarvik 2010). This HGT of carotenoid biosynthesis genes from fungi enabled aphids to avoid predation and thus play a critical role in their survivals. Another example is found in the glycoside hydrolases (GH) gene families that catalyze hydrolysis of the glycoside linkage. Many GH genes found in metazoan genomes are considered to be obtained by HGTs from bacteria or fungi. For many insect herbivores, such acquisition of cellulolytic enzymes is adaptive because it enables them to access the nutritional resources that are most abundant on Earth, cellulose (see more details in 1.3).

1.1.4 Objectives of the research

The scope of this thesis is to elucidate the molecular evolutionary mechanisms of multigene families and their association to functional adaptation. I focused particularly on chemosensory receptors and glycoside hydrolase families.

Molecular evolution of G-protein coupled receptors, especially chemosensory receptors, critically reflects adaptation to the organism’s life. Their evolutionary processes can be often explained by the birth-and-death model. In other words, ecological and behavioral factors can influence the birth and death processes of chemoreceptor families. My working hypothesis is that chemosensory genes show species-specific gene duplications and losses due to their relationships with environmental conditions resulting in larger variation in terms of the gene numbers among the species. Toward this end, I identified complete repertoires of trace amine-associated receptor (TAAR) genes from a wide range of metazoans, and examined the lineage- or species-specific expansions and losses. I also attempted to identify functionally important amino acid sites in these proteins.

As another mechanism of adaptive evolution process, I studied possible HGT events with glycoside hydrolase genes in the western corn rootworm, *Diabrotica virgifera virgifera* (Coleoptera: Chrysomelidae). *D. v. virgifera* is the most serious beetle pest of maize and thus may be specifically adapted to hydrolyze the cellulose. The entire repertoires of glycoside hydrolase genes were identified from the transcriptome of *D. v. virgifera*. I discussed the origin and molecular evolution of glycoside hydrolases among insects, bacteria, and fungi.

1.2 G-protein-coupled receptor superfamily

In the early 1980s, sequencing and subsequent cloning of the bovine retinal photoreceptor, rhodopsin, revealed a novel mammalian protein structure, called G-protein-coupled receptor (GPCR or G protein-linked receptors, GPLRs) (Figure 1.1) (Argos et al. 1982; Ovchinnikov 1982; Hargrave et al. 1983; Nathans and Hogness 1983). The first GPCR whose protein crystal structure was determined was also the bovine rhodopsin (Palczewski et al. 2000). The basic architecture of these receptors is to have seven α-helices. This is why GPCRs are also known as "7-transmembrane receptors". The transmembrane (TM) regions are connected by three intracellular and three extracellular loops with an extracellular N-terminus and an intracellular C-terminus (Figure 1.1). Each of the TM regions are about 25-35 amino acids in length and highly hydrophobic.

GPCRs share a common signaling mechanism in which they interact with heterotrimeric GTP-binding proteins (G-proteins) composed of three subunits (α, β, and γ). Once a ligand activates the GPCRs, G-proteins exchange guanosine diphosphate (GDP) for active guanosine triphosphate (GTP). Through the activation of G-proteins, GPCRs play a central role in eukaryotic signal transduction pathway. The natural ligands for GPCRs are diverse and in a variety of forms such as photons (lights), cations, hormones, and small molecules including biological amines, peptides, lipids, glycoproteins, and sugars. These ligands mediate their messages (*e.g.*, visual, olfactory, and gustatory sensation, intermediary metabolism, neurotransmission, and cell growth) through GPCRs. Many receptor genes have been identified as the result of genome sequencing. However, only a fraction of receptor–ligand interactions have been characterized (Mombaerts 2004). Functions of the most GPCRs are identified on the basis of their sequence similarities and thus are initially unmatched to known natural ligands (Civelli et al. 2013). Many GPCRs are not known to be activated by any known messengers *in vivo* and thus have no known functions. They are called “orphan” GPCRs (Civelli et al. 2006; Chung et al. 2008). For example, more than 70 GPCRs are classified as potential neuromodulator receptors based on the sequence similarities but remain as orphan GPCRs (Civelli 2012).

GPCRs are involved with various mammalian cellular signaling networks as neurotransmission and cellular metabolism. In mutagenesis studies, mutations in GPCRs are found to cause more than thirty human diseases including cancer (Schoeberg et al. 2004). Therefore, these receptors constitute very important novel drug targets for the pharmaceutical industries (Overington et al. 2006). Drugs targeting members of GPCRs command more than 50% of the current market for human therapeutics with annual revenues in excess of $40 billion (Cherezov et al. 2007). Due to their pharmaceutical and biomedical importance, the molecular biology of the GPCRs has been extensively studied in some model organisms.

1.2.1 Classification of GPCRs.

GPCRs represent the largest multigene families in the animal genomes. They comprise 3-10% of the total gene content of animal genomes. In mammalian genomes, their numbers range from 800 to 2,400 (Lagerstrom and Schioth 2008). There are more than 900 GPCRs identified in human (Sällman Almén et al. 2009), more than 1,800 in mouse (Gloriam et al. 2007), roughly 1,500 in the *Caenorhabditis elegans* genome (Bargmann 2006), and about 310 or more in the *Drosophila melanogaster* genome. GPCRs are also present in plant and fungi. However, much fewer numbers of GPCRs have been found in plants and fungi compared to in animals. For example, approximately 20 GPCRs have been identified in the *Arabidopsis thaliana* genome (Moriyama and Opiyo 2010) and 10 GPCRs in the fungal genome of *Neurospora crassa* (Xue et al. 2008).

In addition to being the largest, the GPCR superfamily is the most diverse among membrane-bound receptors (Bockaert and Pin 1999). Sequence similarities among GPCRs can be lower than 25%. For example, the identity between odorant receptor (OR) proteins drop to as low as 40% in human (Glusman et al. 2001) and 16% in *D. melanogaster* (Clyne et al. 1999). Note that there are no absolutely conserved positions among human OR protein sequences (Young et al. 2002). Such high sequence diversity makes it difficult to identify and classify GPCRs. There are several methods to classify GPCRs. Phylogenetic studies showed that *D. melanogaster* GPCRs can be grouped into four families (rhodopsin-like, secretin-like, metabotropic glutamate–like, and atypical 7 TM proteins) (Brody and Cravchik 2000) and human GPCRs by five families (glutamate, rhodopsin, adhesion, frizzled/taste2, and secretin) (Fredriksson et al. 2003). This latter five-family system is known as the GRAFS classification. The GPCRDB database (http://www.gpcr.org/7tm) organizes GPCR sequences using a hierarchical structure based on their binding ligand types, functions, and sequence similarities (Vroling et al. 2011). The current version (ver. 11.3.4) of GPCRDB divides them into three major classes (Class A: Rhodopsin-like family, Class B: Secretin-like family, and Class C: Metabotropic glutamate/pheromone family) and three other divergent groups (cAMP receptors; Vomeronasal receptors, V1R and V3R; and Taste receptors, T2R). The International Union of Pharmacology Committee on Receptor Nomenclature and Drug Classification (NC-IUPHAR) also provides the GPCR classification of human, mouse, and rat proteins for which preliminary evidence for endogenous ligands has been published or there exists a potential link to a disease (Foord et al. 2005; Sharman et al. 2013). NC-IUPHAR classification has the five categories (Class A, Class B, Class C, Frizzled class, and Other 7TM proteins). Table 1.1 shows a classification of GPCR based on the current version of GPCRDB but modified to include four other groups (Insect ORs/GRs, Plant mildew-resistance locus O receptors, Nematode chemoreceptors, and Frizzled/Smoothened family).

1.2.2 Chemosensory receptors

Studies of the chemosensory system at the molecular level began in 1990s. The first chemosensory receptors (CRs) identified were the odorant receptors (ORs) in *Rattus norvegicus* (Buck and Axel 1991). This study showed that the olfactory recognition capacity relies on a set of multigene families and its major player is ORs. ORs play a role in the binding of odorants and the conversion of chemical information into electronic signals in olfactory neurons. For this discovery, Linda Buck and Richard Axel won the 2004 Nobel Prize in Physiology or Medicine. Since then, other types of CRs, *e.g.*, vomeronasal receptor types 1 and 2 (V1R and V2R) (Dulac and Axel 1995; Herrada and Dulac 1997; Matsunami and Buck 1997), taste receptor types 1 and 2 (T1R and T2R) (Adler et al. 2000; Matsunami et al. 2000; Li et al. 2002), trace amine-associated receptors (TAARs) (Borowsky et al. 2001; Bunzow et al. 2001), and formyl peptide receptors (FPRs) (Liberles 2009; Riviere et al. 2009) have been identified in vertebrates. All these vertebrate CR genes are known to be members of the GPCR superfamily. Based on the GPCR classification (shown in Table 1.1), vertebrate chemoreceptors belong to two major classes and two other divergent groups. ORs, TAARs, and FPRs are the members of Class A (Rhodopsin-like family). They are intron-less, encoded in a single exon (except for TAAR2). T1R and V2R genes have complex multiple exon structures (five introns on average) and a long N-terminal. They belong to Class C (Metabotropic glutamate/pheromone family). T2Rs and V1Rs form their own groups, "Taste receptors (T2R)" and "Vomeronasal receptors (V1R)", respectively.

Insects are known to have three different multigene CR families: odorant receptors (ORs) (Clyne et al. 1999; Gao and Chess 1999; Vosshall et al. 1999), gustatory receptors (GRs) (Clyne et al. 2000), and ionotropic glutamate receptors (iGluRs) (Benton et al. 2009). Although vertebrates and insects both have ORs and GRs for detecting odor and taste molecules, they have following significant differences. First, the sequence similarity between vertebrate and insect ORs/GRs is extremely low (*e.g.,* less than 20% between mouse ORs and *D. melanogaster* ORs) and there is no conserved motif between vertebrate and insect OR proteins (Clyne et al. 1999). Second, the transmembrane (TM) topology of insect ORs as well as GRs, although they contain seven TM regions, was found to be inverted compared with that of classic GPCRs (including vertebrate ORs) (see Figure 1.1 for the topology of regular GPCRs) such that the N-terminus is located in the intracellular region in insect receptors (Benton et al. 2006). Third, an insect OR and a ubiquitously expressed co-receptor, Orco (formerly known as OR83b), can act as ligand-gated ion channels (Sato et al. 2008; Wicher et al. 2008). It should be noted that Sato et al. (2008) and Smart et al. (2008) described that G-protein-mediated signaling plays a negligible role in receptor activation and thus the OR complex does not involve G-proteins, whereas Wicher et al. (2008) showed that the OR complex acts as both a GPCR and an ion channel. Boto et al. (2010) furthermore demonstrated that G-proteins (3 Gβ and 2 Gγ subunit) are present in the olfactory sensory neurons bearing ORs. Fourth, in the insect olfactory sensory neurons, odorant-binding proteins (OBPs) mediate chemosensory responses (Laughlin et al. 2008). OBPs have been proposed to serve either as odorant scavengers or carriers that deliver the odorant or pheromone to the receptors (Kaupp 2010). Taken together, these differences imply that insect chemoreceptors may have arisen independently from vertebrate chemoreceptors. Alternatively, the vertebrate types of chemoreceptors may have been lost in insects and the insect types may have been lost in vertebrates.

1.2.3 Chemosensory organs and receptors in vertebrates and insects

Most vertebrates possess three distinct chemosensory organs: the main olfactory epithelium (MOE), the vomeronasal organ (VNO), and the tongue (reviewed in Matsunami and Amrein 2003). MOE is found in almost all vertebrates (except for some marine mammals). The OR genes are predominantly expressed in MOE (Kaupp 2010). TAAR genes are also expressed in the MOE but at a lower level compared to ORs (Borowsky et al. 2001; Liberles and Buck 2006). VNO is absent in birds (Stoddart 1980) while most terrestrial vertebrates possess the paired cigar shaped VNO located just above the roof of the mouth (the rostral end of the nasal cavity). Elephants (*Loxodonta africana*) are known to have a well-developed VNO (Göbbel et al. 2004). However, this organ is absent in some placental (eutherian) groups: catarrhine primates (Maier 1997), cetaceans (Oelschläger 1989), the West Indian manatee (Mackay-Sim et al. 1985), megachiropterans, and some microchiropterans (Cooper and Bhatnagar 1976; Bhatnagar 1980; Wible and Bhatnagar 1996). VNO hosts three CR families: V1Rs, V2Rs, and FPRs. The primary function of vomeronasal receptors (V1Rs and V2Rs) and FPRs is to detect ligands associated with social cues. Traditionally it has been considered that MOE responds to general volatile odor molecules, whereas VNO detects intraspecific pheromonal cues as well as some environmental non-volatile odorants. However, it is now known that the ORs and vomeronasal receptors share some overlapping functions (Sam et al. 2001; Baxi et al. 2006; Zufall and Leinders-Zufall 2007) and their relationships reflect a common history of ecological adaptations (Suárez et al. 2012). Taste recognition is encoded by the T1Rs and T2Rs. They are expressed in the taste buds of the tongue (Adler et al. 2000; Matsunami et al. 2000).

In insects, especially in *Drosophila* species, ORs and olfactory receptor neurons are found in the antenna and the maxillary palp on the head, whereas GRs and taste sensory neurons are scattered on the entire body, including the proboscis, two labial palps, wings, and all legs (Matsunami and Amrein 2003; Vosshall and Stocker 2007). Each neuron expresses a few, possibly just one, ORs or GRs and a few GRs are also expressed in olfactory neurons of the antenna and maxillary palps (Vosshall and Stocker 2007).

1.2.4 Trace amine-associated receptors

Biogenic amines (adrenaline or epinephrine: AD, norepinephrine or noradrenaline: NE or NA, dopamine: DA, serotonin or 5-hydroxytryptamine: 5-HT, and histamine: HA) are enzymatic decarboxylation products of amino acids (Figure 1.2). They are crucial intercellular signaling molecules that function widely as neurotransmitters and neuromodulators (Ringstad et al. 2009; Flames and Hobert 2011). Both of α- and β-adrenergic receptors (adrenoreceptors) are activated by their endogenous agonists AD and NE, which belong to the catecholamine transmitters. (Saavedra 1980; Ho and Chik 2000). Dopamine D_1 and D_2 receptors are stimulated by DA and serotonin receptors such as 5-HT_{1A}, 5-HT_{2A}, and 5-HT_7 receptors by 5HT (Millan et al. 2008; Gogos et al. 2010). All these receptors except for 5-HT_3 receptors, which are ligand-gated cation-permeable ion channels, belong to GPCR Class A (the Rhodopsin-like receptors) (Millan et al. 2008; Ringstad et al. 2009) (Table 1.1).

In addition to these classical amines, there is another class of endogenous amine compounds that are present in mammalian tissues at trace amounts (0.1–10 nM) (Branchek and Blackburn 2003; Zucchi et al. 2006; Broadley 2010). They are called ‘‘trace amines’’ (TAs). They include 2-phenylethylamine (PEA), m-tyramine (m-TYR), ρ-tyramine (ρ-TYR), *meta*-octopamine (m-TA), *para*-octopamine (p-TA), 3-iodothyronamine (T_1AM), tryptamine (TRY), and *N*,*N*-dimethyltryptamine (DMT) (Figure 1.2). TAs are structurally related to classical biogenic amines. They share substantial similarities in their biosynthesis and co-localization in the same neurons (Ledonne et al. 2011). TAs are known to be of importance in invertebrate physiology by interacting with specific plasma membrane GPCRs (Zucchi et al. 2006). Tyramine is found in many common foods and increases blood flow to the brain, which could trigger high blood pressure and headache (Peatfield et al. 1983; Welling 1996).

TAARs were originally identified based on their relatedness to biogenic amine receptors and discovered in search of the receptors activated by the TAs in the brain (Borowsky et al. 2001; Bunzow et al. 2001). In the mouse genome, fifteen functional genes and one pseudogene are known for TAARs. They are classified into nine subfamilies (TAAR1 through TAAR9). In mouse, most of these subfamilies are represented by single copy genes except for TAAR7, which includes five genes and one pseudogene, and TAAR8, which includes three genes (Lindemann et al. 2005). All mouse TAARs except for TAAR1 are expressed in the main olfactory epithelium (MOE) (Liberles and Buck 2006; Fleischer et al. 2007). TAAR1 is expressed in the brain (Borowsky et al. 2001). Liberles and Buck (2006) demonstrated that TAARs also function as chemosensory receptors and are expressed in the main olfactory epithelium (MOE) in mouse. TAAR4, for example, is stimulated by 2-phenylethylamine, which is a carnivore odor that evokes physiological and behavioral responses in two prey species (rat and mouse) (Ferrero et al. 2011). TAARs thus play important roles in sensing predator and prey odors.

Amines have different classes depending on how many of the hydrogen atoms in ammonia are replaced. In primary amines, one of the three hydrogen atoms in the ammonia molecule has been replaced by an alkyl or aromatic. They could be derived from natural amino acids by a single decarboxylation reaction. In tertiary amines, all three hydrogen atoms are replaced by organic substituents (Figure 1.2). Ferrero et al. (2012) showed that TAARs can be classified into two groups based on whether they preferentially detect primary or tertiary amines. TAAR1-4 are stimulated by primary amines (*e.g.*, isoamylamine) while TAAR5-9 detect tertiary amines (*e.g.*, *N*,*N*-dimethylated amines).

Many medical studies have focused on TAs and TAARs. TAs are putative regulatory elements in the brain (Berry 2004) and thus of importance in understanding several human diseases because current studies suggest that a regulatory role of TA system affects some psychiatric disorders such as abuse, insomnia, depression, attention deficit hyperactivity disorder, bipolar, schizophrenia, and other neuropsychiatric diseases (Duan et al. 2004; Wolinsky et al. 2007b; Serretti et al. 2009; Pae et al. 2010). TAAR6 are reported as the candidate genes for schizophrenia (Duan et al. 2004; Vladimirov et al. 2007; Serretti et al. 2009). Interestingly, rat TAAR1 is also activated by classical TAs as well as synthetic analogues such as 3,4-methylenedioxymethamphetamine (MDMA, known as ecstasy), *d*-lysergic acid diethylamide (LSD), and amphetamine (Bunzow et al. 2001).

Only a limited number of molecular evolutionary studies have been done for TAARs. The complete TAAR gene set has been described in nine mammalian species (human, chimpanzee, macaque, mouse, rat, dog, cow, opossum, and platypus) (Lindemann et al. 2005; Grus et al. 2007; Hashiguchi and Nishida 2007), chicken (Mueller et al. 2008), five teleosts (fugu, spotted green pufferfish, stickleback, medaka, and zebrafish), a cartilaginous fish (elephant shark), and a jawless fish (sea lamprey) (Hashiguchi and Nishida 2007; Hussain et al. 2009). These studies showed that the tetrapod genomes have small numbers of

TAAR genes (3–22 genes), while many teleost fishes have higher numbers of TAAR genes compared to tetrapods, ranging from 13 to 109 genes.

1.2.5 Molecular evolution of CRs

CRs are important in mediating behavioral responses to, *e.g.*, food, mates, and predators because CRs are used to detect a wide range of chemical signals. They are thus crucial gateways between environment and perception. Different life history traits such as foraging behavior (herbivore *vs*. carnivore), habitat (aquatic *vs*. terrestrial), and type of foods are expected to play a central role in driving variation in the number of CR genes. For instance, two nocturnal bird species, the brown kiwi (*Apteryx australis*) and the kakapo (*Strigops habroptilus*), have a larger number of OR genes than their closest diurnal relatives (brown kiwi relatives: emu *Dromaius novaehollandiae*, rhea *Rhea americana*, ostrich *Struthio camelus*; kakapo relatives: kaka *Nestor meridionalis*, kea *Nestor notabilis*), suggesting strong ecological niche adaptations such as daily activity patterns (Steiger et al. 2009). Extensive studies of OR genes have been done in teleosts and tetrapods (Alioto and Ngai 2005; Niimura and Nei 2005; Nei et al. 2008). These studies showed that while the tetrapod genomes have a large number of OR genes, ranging from 400 to 2,100, a significant portion of them, in the order of 20–50%, are pseudogenes (Nei et al. 2008) (also see Table 1.2). For example, a total of 802 OR genes were identified in the human genome but at least 52% of them are pseudogenes (Go and Niimura 2008a). In contrast, the mouse genome has 1,391 ORs and has only ~20% pseudogenes.

Nei et al. (2008) suggested that the species-specific gene duplications have important roles in the adaptive evolution to different environments. Bargmann (2006) proposed that CRs, like the immune system, track a moving world of cues generated by other organisms, and must constantly generate, test, and discard receptor genes and coding strategies over the evolutionary time. The expansion and contraction of CRs might be a key to reflect the adaptation to the organism's life at the molecular level. Birth-and-death evolution can be also a random process. As Nei et al. (2008) described, a substantial portion of gene number changes in CR gene families must have been caused by such a random birth-and-death events. Therefore, both the adaptive and non-adaptive evolution can play a role in evolution of CR genes.

The numbers of CRs in insects are significantly fewer than those in vertebrates. However, they are highly divergent and many of them have species-specific gene duplications with no close orthologs (Hansson and Stensmyr 2011). For example, *D. melanogaster* possesses only 62 ORs (encoded by 59 genes) and 73 GRs (encoded by 68 genes) (McBride and Arguello 2007), whereas the red flour beetle (*Tribolium castaneum*) genome has 262 ORs and 62 GRs and the honeybee (*Apis mellifera*) genome has 163 ORs and 10 GRs (Table 1.2). Thus, insect CRs also represent the birth-and-death evolution.

1.2.6 Origin of GPCRs in the basal metazoan

Many GPCR families are shared among a wide range of eukaryotic organisms but several lineage-specific GPCR groups have been also reported: for example, fungal pheromone receptors (STE2 and STE3), mildew-resistance locus O (MLO) receptors in plants, nematode chemoreceptors (serpentine receptors), insect receptors (ORs and GRs), and methuselah (mth, insect Class B). These GPCR families are not present in vertebrate genomes. In addition, several families of GPCRs show no significant sequence similarities to each other. Nordström et al. (2011) suggested that the Rhodopsin family, Adhesion family, and Frizzled family share a common evolutionary origin and are derived from cAMP family, whereas insect ORs and GRs do not share a common origin with vertebrate GPCRs.

On the other hand, a study of coral expressed sequence tags (ESTs) suggested that many genes thought to be invertebrate- or vertebrate-specific may in fact have much older origins, and have been lost during the evolution (Kortschak et al. 2003). Krishnan et al. (2012) provided the evidence of the presence of four of the five main GPCR families in fungi and demonstrated the early evolutionary history of the GPCR superfamily.

Rhodopsins, photosensitive proteins, are found in three domains of life (archaea, eubacteria, and eukaryotes). They can be divided into two types: type I, or microbial, rhodopsins function as light-driven ion transporters and sensory transducers, and are found in γ-proteobacteria, cyanobacteria, archaea, green algae, and fungi, while type II, or metazoan, rhodopsins are found in the photoreceptor cells of animal eyes, and control the activation of hetero-trimeric G-proteins leading to visual reception (Spudich et al. 2000; Beja et al. 2001; Jung 2007). Shen et al. (2013) suggested that type II rhodopsins originated from type I rhodopsins based on the 7-TM structures and a conserved sequence motif (WXXY) in the sixth TM region. These findings suggest that GPCRs could share the common evolutionary origin in basal eukaryotic genomes.

Recently, genome sequences of several basal metazoan have been released (Putnam et al. 2007; Srivastava et al. 2008; Chapman et al. 2010; Srivastava et al. 2010). The genomes of basal metazoan are very attractive for studying the origin of GPCRs because these organisms diverged early in the metazoan evolution after the Kingdom Fungi diverged from the Kingdom Metazoa, more than 700 million years ago (Putnam et al. 2007) (Figure 1.3). For example, a recent study of the *Nematostella vectensis* (sea anemone) genome indicated that the origin of vertebrate ORs can be traced back to the Cnidaria (Churcher and Taylor 2011). Thus, these basal metazoan genomes can be useful in filling the gaps in finding the ancestral characteristics of GPCRs and understanding the divergence and evolution among metazoa, such as between deuterostomes and protostomes and between metazoans and protists.

1.3 Glycoside hydrolase families.

1.3.1 Plant cell walls degradation and cellulase

Plant cell walls are comprised largely of polysaccharides: cellulose, hemicellulose, and pectin, along with ~10% protein and up to 40% lignin (Burton et al. 2010). The plant cell wall degradation process studied in fungi consists of three coordinated steps: depolymerization of lignin or pectin, hemicellulose degradation, and finally cellulose degradation (Gamauf et al. 2012). Hence, plant cell wall digestion requires numerous enzymes including pectinases, ligninases, hemicellulases, and cellulases with diverse substrates (Gilbert 2010). The degradation of pectin chains by polygalacturonases (EC 3.2.1.15) loosens the primary cell wall making the cellulose-hemicellulose network more accessible (Juge 2006). Hemicellulose, which is less rigid than cellulose, is readily degraded by hemicellulases such as xylanases (EC 3.2.1.8).

Cellulose, which is synthesized by terrestrial plants and marine algae, is the most abundant organic compound on Earth. It is a simple carbohydrate polymer, consisting of repeating glucose units linked by β-1,4-glycosidic bonds (Figure 1.4). It is also characterized as insoluble and comprised of nanometer-thick crystalline microfibrils, which are highly resistant to enzymatic hydrolysis (Béguin and Aubert 1994). Cellulase is a general term for cellulolytic enzymes, a family of enzymes that hydrolyze the β-1,4 linkages of cellulose. Three classes are recognized for cellulase on the basis of the mode of enzymatic acting and the substrate specificities: endoglucanases (EC 3.2.1.4), cellobiohydrolases (EC 3.2.1.74 and 3.2.1.91), and β-glucosidases (EC 3.2.1.21) (Watanabe and Tokuda 2010) (Figure 1.4). Cellulases are widespread from microorganisms such as bacteria and fungi to plants. Cellulolytic fungi and bacteria have developed highly complex cellulase systems (Tomme et al. 1995). Plants possess cellulase genes to hydrolyze their cell walls during various developmental stages (Robert et al. 2005). Furthermore, these cellulase systems play a very important role in a wide range of processes ranging from biosphere maintenance (carbon recycling) (Melillo et al. 2002; Brune 2003) to the generation of potentially sustainable energy sources such as glucose, ethanol, hydrogen, and methane (Wyman 2003; Kamm and Kamm 2004; Zhang and Lynd 2005).

1.3.2 Endogenous insect cellulolytic enzymes.

For many insect species, cellulose comprises a major nutritional resource (Breznak and Brune 1994; Watanabe and Tokuda 2010). Until recently, it was wildly accepted that most metazoans do not have endogenous cellulolytic activity or at least is rare and cellulose digestion in insects was mediated by gut-associated microbes such as mixtures of bacteria and protozoa under anaerobic conditions (Martin 1983; Martin 1991; Breznak and Brune 1994). However, this traditional view has been challenged. The number of recent studies have reported the endogenous origin of cellulolytic enzymes in insects (Smant et al. 1998; Watanabe et al. 1998; Calderón-Cortés et al. 2012).

Our current understanding of cellulose digestion in insects has been obtained from study of termite systems. Termites are voracious eaters and an extremely successful group of wood-degrading organisms (Brune and Ohkuma 2011). They are therefore important both for their roles in carbon turnover in the environment and as potential sources of biochemical catalysts for efforts aimed at converting wood into biofuels (Warnecke et al. 2007). Phylogenetically "lower" termites (Mastotermitidae, Termopsidae, Hodotermitidae, Kalotermitidae, Serritermitidae, and Rhinotermitidae) have symbiotic protozoan fauna in the hindgut, which produce cellulases encoded by glycoside hydrolase (GH) family genes, GH5, GH7, and GH45 (Hongoh 2011). The total contribution of symbiotic enzymes in the hindgut of lower termites varies from 12 to 40% for endoglucanases, 62 to 84% for cellobiohydrolases, and 88 to 98% for xylanases (Calderón-Cortés et al. 2012). On the other hand, because “higher” termites (Termitidae) lack cellulolytic protists but still have strong cellulase activity in the midgut, it was believed that they rely solely upon their own endogenous cellulases coded by GH family genes, *e.g.*, GH9 (Brune and Stingl 2005).

However, recent studies showed that higher termites also have diverse bacterial communities including archaea, proteobacteria, bacteroidetes, and spirochaetes (Hongoh 2011). The cellulase activity of hindgut bacteria on highly polymerized cellulose contributes significantly to plant cell wall degradation in higher termites (Tokuda and Watanabe 2007; Warnecke et al. 2007; Zhou et al. 2007). Therefore, there appears to be efficient synergistic enzyme interaction between a complex mixture of bacterial, protozoan, and insect produced enzymes in the termite gut (Zhou et al. 2007). However, an understanding of the exact roles of the host and symbiotic microbiota in the complex process of cellulose degradation is still emerging (Nakashima et al. 2002; Tokuda et al. 2007; Scharf et al. 2011).

1.3.3 Classification of glycoside hydrolases and their distribution in metazoans and insects

Glycoside hydrolases (GH; EC 3.2.1.-) are classified into 132 families and 14 clans according to their amino-acid sequence similarities and their folding patterns by the Carbohydrate-Active enZymes Database (CAZy, http://www.cazy.org) (Cantarel et al. 2009). As shown in Table 1.3, three classes of cellulolytic enzymes are placed into five GH-clans and some are non-classified. The β-glucosidase genes (also known as cellobiases, EC 3.2.1.21) (GH1 and GH3; also the activity is associated with GH5 and GH30) are widely distributed in metazoan species (reviewed in Calderón-Cortés et al. 2012). It has been known that insects lack cellobiohydrolase (also known as exoglucanase) (Scrivener and Slaytor 1994). However, recently Chang et al. (2012) identified a gene in *Anoplophora malasiaca* (spotted longhorn beetle) that exhibited exo-β-glucanase as well as endo-β-glucanase activities. Five endoglucanase genes (GH5, GH7, GH9, GH45, and GH48) and five other GH family genes (GH10, GH11, GH16, GH28, and GH31) have been found in a limited number of metazoan lineages (Markovič and Janeček 2001; Calderón-Cortés et al. 2012). Eight of these GH family genes have been identified in insects. These genes are mapped on the starch and sucrose metabolic pathway in Figure 1.5. The numbers of GH genes identified in beetle species are shown in Figure 1.6.

GH9 are known to have endoglucanase (EC 3.2.1.4), cellobiohydrolase (EC 3.2.1.91), β-glucosidase (EC 3.2.1.21), and exo-β-glucosaminidase (EC 3.2.1.165). Watanabe et al. (1998) identified the first endogenous cellulase gene (GH9) from a termite (*Reticulitermes speratus*). Since then, GH9 genes have been widely identified in arthropods (*e.g.*, pea aphid *Acyrthosiphon pisum*, Cherqui and Tjallingii 2000; Egyptian desert roach *Polyphaga aegyptiaca*, brown-hooded cockroach *Cryptocercus clevelandi*, Lo et al. 2000; garden cricket *Teleogryllus emma*, Kim et al. 2008; western honey bee *Apis mellifera*, Kunieda et al. 2006; red flour beetle *Tribolium castaneum*, Willis et al. 2011; human louse *Pediculus humanus humanus*, XM_002426420), as well as in a mollusk (*Haliotis discus hannai*, abalone) (Suzuki et al. 2003), an urochordate (*Ciona intestinalis*, vase tunicate) (Dehal 2002), a fungus (*Piromyces* sp.) (Steenbakkers et al. 2002), and an amoebozoan (*Dictyostelium discoideum*, slime mold) (Libertini et al. 2004). However, these enzymes are absent in *D. melanogaster*, *Anopheles gambiae*, *Caenorhabditis elegans*, *Saccharomyces pombe*, and *Saccharomyces cerevisiae* genomes (Davison and Blaxter 2005). Because GH9 family genes share several intron positions conserved among four metazoan phyla, these GH genes seem to be derived from an ancient common ancestor (Lo et al. 2003; Davison and Blaxter 2005).

GH5 represents the largest GH family (3,856 sequences as of July 2013 at CAZy) and is assigned into 51 subfamilies based on phylogenetic analysis (Aspeborg et al. 2012). This family can be also assigned to eighteen subgroups according to their substrate specificities; chitosanase (EC 3.2.1.132), β-mannosidase (EC 3.2.1.25), endo-β-1,4-glucanase/cellulase (EC 3.2.1.4), glucan β-1,3-glucosidase (EC 3.2.1.58), licheninase (EC 3.2.1.73), glucan endo-1,6-β-glucosidase (EC 3.2.1.75), mannan endo-β-1,4-mannosidase (EC 3.2.1.78), endo-β-1,4-xylanase (EC 3.2.1.8), cellulose β-1,4-cellobiosidase (EC 3.2.1.91), β-1,3-mannanase (EC 3.2.1.-), xyloglucan-specific endo-β-1,4-glucanase (EC 3.2.1.151), mannan transglycosylase (EC 2.4.1.-), endo-β-1,6-galactanase (EC 3.2.1.164), endoglycoceramidase (EC 3.2.1.123), β-primeverosidase (EC 3.2.1.149), β-glucosylceramidase (EC 3.2.1.45), hesperidin 6-O-α-L-rhamnosyl-β-glucosidase (EC 3.2.1.168), and exo-β-1,4-glucanase/cellodextrinase (EC 3.2.1.74). GH5 genes have been identified in yellow-spotted longicorn beetle (*Psacothea hilaris*), mulberry longicorn beetle (*Apriona germari*), and borer beetle (*Oncideres albomarginata chamela*) (Sugimura et al. 2003; Wei et al. 2006; Calderón-Cortés et al. 2010). While Calderon-Corte et al. (2010) discussed that GH5 likely represents a single ancient origin resulting from a common ancestor rather than HGT based on phylogenetic analysis, later the authors (Calderón-Cortés et al. 2012) described the origin to be unclear. As discussed later (1.3.4), recent studies showed HGT events of GH5 (subfamily 8) in coffee borer beetle *Hypothenemus hampei* (Acuña et al. 2012) and GH5 (subfamily 2) in plant-parasitic nematodes (Rybarczyk-Mydlowska et al. 2012).

GH45 has only the endoglucanase (EC 3.2.1.4) activity. This family has been found among various animals from protists (Li et al. 2003), plant-parasitic nematodes (Smant et al.

1998) to mollusks (Xu et al. 2001; Harada et al. 2004). GH45 genes have also been described from a number of beetle species including *Phaedon cochleariae* (mustard leaf beetle, Chrysomelidae) (Girard and Jouanin 1999), *Ips pini* (pine engraver beetle, Scolytinae) (Eigenheer et al. 2003), *Apriona germari* (mulberry longicorn beetle, Cerambycidae) (Lee et al. 2004), and *Oncideres albomarginata chamela* (borer beetle, Cerambycidae) (Calderón-Cortés et al. 2010). Pauchet et al. (2010) examined GH45 and other genes in four beetle species and described multiple GH45 genes existing within a single species. For example, *Dendroctonus ponderosae* (mountain pine beetle, Curculionoidea) possesses nine GH45 genes (Keeling et al. 2012) and *Leptinotarsa decemlineata* (Colorado potato beetle, Chrysomeloidea) has seven genes (Pauchet et al. 2010). All insect GH45 cellulase are reported only in beetle species, representing two coleopteran superfamilies, Chrysomeloidea and Curculionoidea except for one GH45 gene (ACV50414.1) in *Cryptopygus antarcticus* (Isotomidae, Collembola). Tardigrades are known as the sister group of arthropods and the model species, *Hypsibius dujardini*, has one GH45 (CD449425.1). Calderón-Cortés et al. (2012) reported that *D. melanogaster* has two GH45 genes (EC068056 and CO334668). However, these two sequences are 100% identical to each other in their nucleotide sequences and these corresponding sequences are not present in the genome of *D. melanogaster* (ver. 5.51; http://flybase.org) using BLAST protein sequence similarity search. Moreover, these two sequences are almost identical to the *L. decemlineata* GH45-7 sequence (ADU33351.1) (100% identical in amino acid sequences and only three nucleotide differences). Therefore, these GH45 sequences are very likely to be misidentifications and GH45 is absent in *D. melanogaster*.

GH48 is the most common GH family genes in bacteria. It has endo-β-1,4-glucanase (EC 3.2.1.4), chitinase (EC 3.2.1.4), and cellobiohydrolase (EC 3.2.1.176) activities. Two GH48 genes were isolated from the leaf beetle *Gastrophysa atrocyanea* (leaf beetle, Chrysomelidae) (Fujita et al. 2006). Six GH48 genes in *D. ponderosae*, three genes in *L. decemlineata*, two genes in *Sitophilus oryzae* (Rice weevil, Curculionidae), and *Gastrophysa viridula* (Green dock beetle, Chrysomelidae) are also reported (Pauchet et al. 2010; Keeling et al. 2012).

In addition to these cellulolytic enzyme genes, the gene encoding a pectolytic enzyme polygalactunorase (EC 3.2.1.15), GH28 (GH-N clan, see Table 1.3), is found widespread among bacteria, fungi, and plants, representing the second largest GH family. GH28 enzymes have been found in a phytophagous beetle (*Phaedon cochleariae*, Chrysomelidae) (Girard and Jouanin 1999) and in four beetle species (*G. viridula*, *L. decemlineata*, *S. oryzae*, and *Callosobruchus maculatus*) (Pauchet et al. 2010). In addition to the polygalactunorase activity, the GH28 family enzymes are shown to have activities including exo-polygalacturonase (EC 3.2.1.67), exo-polygalacturonosidase (EC 3.2.1.82), rhamnogalacturonase (EC 3.2.1.171), endo-xylogalacturonan hydrolase (EC 3.2.1.-), rhamnogalacturonan a-L-rhamnopyranohydrolase (EC 3.2.1.40) (Markovič and Janeček 2001). Calderón-Cortés et al. (2012) reported that *D. melanogaster* has one GH28 gene (CO335003). However, again this sequence cannot be found in the present genome of *D. melanogaster* (ver. 5.51). Its nucleotide sequence is 100% identical with the one found in *L. decemlineata* GH28-9 (ADU33363.1). Therefore, again this is likely a misidentification and GH28 gene does not exist in *D. melanogaster*.

GH11 contains only xylanase (EC 3.2.1.8) activity. Xylan is the predominant constituent of the hemicellulose matrix of the plant cell wall and the second most abundant polysaccharide on the earth. Xylanases from the GH11 family are widely distributed in microorganisms but are generally absent in animals (Pauchet and Heckel 2013). Recently, however, two GH11 genes were identified in mustard leaf beetle (*P. cochleariae*), which are likely obtained from γ-proteobacteria through HGT (Kirsch et al. 2012; Pauchet and Heckel 2013). These genes represent the first example of the GH11 family in animals.

GH16 can be assigned to ten subgroups according to their substrate specificities, including xyloglucan:xyloglucosyltransferase (EC 2.4.1.207), keratan-sulfate endo-1,4-β-galactosidase (EC 3.2.1.103), endo-1,3-β-glucanase (EC 3.2.1.39), endo-1,3(4)-β-glucanase (EC 3.2.1.6), licheninase (EC 3.2.1.73), β-agarase (EC 3.2.1.81), κ-carrageenase (EC 3.2.1.83), xyloglucanase (EC 3.2.1.151), endo-β-1,3-galactanase (EC 3.2.1.181), and β-porphyranase (EC 3.2.1.178). Genta et al. (2009) characterized GH16 gene in the midgut of *Tenebrio molitor* (Tenebrionidae, Coleoptera) larvae. Pauchet et al. (2009) found that GH16 was widely distributed in Lepidoptera (*Plutella xylostella*, *Ostrinia nubilalis*, *Spodoptera littoralis*, and *Bombyx mori*). Later, Song et al. (2010) cloned and characterized a GH16 gene (*CaLam*) from the Antarctic springtail, *Cryptopygus antarcticus* (Isotomidae, Collembola).

GH31 are known to have the following activities: α-glucosidase (EC 3.2.1.20), α-1,3-glucosidase (EC 3.2.1.84), sucrase-isomaltase (EC 3.2.1.48 and EC 3.2.1.10), α-xylosidase (EC 3.2.1.177), α-glucan lyase (EC 4.2.2.13), isomaltosyltransferase (EC 2.4.1.-), and α-mannosidase (EC 3.2.1.24). Recently, Wheeler et al. (2013) demonstrated an ancient lepidopteran HGT of a GH31 gene from an *Enterococcus* bacteria. The GH31 genes are also found in the red flour beetle (*Tribolium castaneum*) genome and in *D. v. virgifera* (described in Chapter 4).

1.3.4. Molecular evolution and origin of insect glycoside hydrolase families

Many insect GH genes were found to be subject to species-specific gene duplications. For example, the largest number of GH28 (19 functional) genes was identified in mountain pine beetle (*D. ponderosae*) (Keeling et al. 2012) while only one GH28 gene was found in mustard leaf beetle (*Phaedon cochleariae*) (Pauchet et al. 2010). The GH45 genes in beetle species also have species-specific duplications. Thus GH families show birth-and-death evolutionary patterns as discussed with the chemoreceptor families. This implies that beetle species are specifically adapted to their environments to hydrolyze their food with enzymes.

In addition to species-specific duplications, evolution of GH genes is known to be involved with adaptive HGT events. HGTs of GH genes provide a competitive advantage and can lead to ecological specialization of the recipient. Possible HGTs have been identified in rumen fungal GH5 and GH11 (Garcia-Vallvé et al. 2000), GH16 in *C. antarcticus* (Song et al. 2010), GH5 in *Hypothenemus hampei* (Acuña et al. 2012), GH31in *Bombyx mori* (Wheeler et al. 2013), and GH11 in *P. cochleariae* (Pauchet and Heckel 2013). For example, Acuña et al. (2012) identified a glycoside hydrolase gene (*HhMAN1*, GH5, subfamily 8) from the coffee berry borer beetle (*H. hampei*, Curculionoidea, Coleoptera) and showed the evidence of HGT from bacteria. Interestingly, while this gene was found to be widespread in their broad biogeographic survey, it was not found in two other species: *H. obscurus*, a close relative of *H. hampei* but not a pest of coffee, and *Araecerus fasciculatus* (coffee bean weevil, Anthribidae, Coleoptera), which is a common pest of coffee but polyphagous (a generalist) in contrast to monophagous or a specialist as *H. hampei* is (Gladstone and Hruska 2003; Valentine 2005; Waller et al. 2007). Therefore, acquisition of *HhMAN1* from bacteria appears to provide a rapid adaptation to a specific ecological niche by enabling hydrolysis of galactomannan, which is a potential source of nutrient for *H. hampei* (Acuña et al. 2012).

Therefore, the two multigene families described in this thesis, TAARs and GHs, have a similar evolutionary pattern with high levels of species-specific gene duplications and losses. However, the GH family evolution is unique in that evolution of many insect GHs involves with HGTs. The birth-and-death evolution and HGTs found in GH families are reflected in the fascinating adaptations of insects and other invertebrates toward various environments.

1.4. Organization of the dissertation

This dissertation is divided into following five chapters.

This chapter (Chapter 1) describes the overview of multigene family evolution, the overall objectives of my research, and background on two multigene families.

Chapter 2 describes the functional divergence and molecular evolution of TAARs. Many species-specific TAAR gene duplications and losses contributed to a large variation of TAAR gene numbers among mammals. I found the evidence of positive selection in mammalian specific TAAR groups. This could have contributed to mammalian adaptation to the dynamic land environment.

In Chapter 3, more detailed molecular evolutionary analysis of TAARs in twelve primate genomes is described. Primate genomes have generally smaller numbers of TAARs compared to other mammalian species, and TAAR gene losses seem to be a major trend in the primate evolution. Pseudogenization events are likely to be accelerated in arboreal life and a change of nose shape in Haplorhini species. Particularly in the great apes, the TAAR gene losses by natural selection might have occurred possibly because of a role in susceptibility to psychiatric disorders such as schizophrenia.

Chapter 4 describes the results of molecular evolutionary analysis of another multigene family, glycoside hydrolase (GH), in the western corn rootworm, *D. v. virgifera*, and related coleopteran species. Three types of GH family genes (GH45, GH48, and GH28) were identified. These GH genes were found only in two coleopteran superfamilies, indicating their HGT origin. Several independent HGT events in bacteria, fungi, and other insect are also discussed.

Chapter 5 presents the conclusion of my studies and prospective researches.

1.5. Literature Cited

Acuña R, Padilla BE, Flórez-Ramos CP, Rubio JD, Herrera JC, Benavides P, Lee S-J, Yeats TH, Egan AN, Doyle JJ, Rose JKC. 2012. Adaptive horizontal transfer of a bacterial gene to an invasive insect pest of coffee. *Proc Natl Acad Sci USA*. 109.

Adler E, Hoon MA, Mueller KL, Chandrashekar J, Ryba NJ, Zuker CS. 2000. A novel family of mammalian taste receptors. *Cell*. 100:693-702.

Alioto T, Ngai J. 2005. The odorant receptor repertoire of teleost fish. *BMC Genomics*. 6:173.

Argos P, Pedersen K, Marks MD, Larkins BA. 1982. A structural model for maize zein proteins. *J Biol Chem*. 257:9984-9990.

Aspeborg H, Coutinho PM, Wang Y, Brumer H, Henrissat B. 2012. Evolution, substrate specificity and subfamily classification of glycoside hydrolase family 5 (GH5). *BMC Evol Biol*. 12:186.

Balakirev ES, Ayala FJ. 2003. PSEUDOGENES: Are They “Junk” or Functional DNA? *Annu Rev Genet*. 37:123-151.

Bargmann CI. 2006. Comparative chemosensation from receptors to ecology. *Nature*. 444:295-301.

Baxi KN, Dorries KM, Eisthen HL. 2006. Is the vomeronasal system really specialized for detecting pheromones? *Trends Neurosci*. 29:1-7.

Béguin P, Aubert J-P. 1994. The biological degradation of cellulose. *FEMS Microbiol Rev*. 13:25-58.

Beja O, Spudich EN, Spudich JL, Leclerc M, DeLong EF. 2001. Proteorhodopsin phototrophy in the ocean. *Nature*. 411:786-789.

Benton R, Sachse S, Michnick SW, Vosshall LB. 2006. Atypical membrane topology and heteromeric function of Drosophila odorant receptors in vivo. *PLoS Biol*. 4:240-257.

Benton R, Vannice KS, Gomez-Diaz C, Vosshall LB. 2009. Variant Ionotropic Glutamate Receptors as Chemosensory Receptors in *Drosophila*. *Cell*. 136:149-162.

Berry MD. 2004. Mammalian central nervous system trace amines. Pharmacologic amphetamines, physiologic neuromodulators. *J Neurochem*. 90:257-271.

Bhatnagar KP. 1980. The chiropteran vomeronasal organ: Its relevance to the phylogeny of bats. In: DE Wilson, AL Gardner, editors. Proceedings of the Fifth International Bat Research Conference. Lubbock, Texas: Texas Tech. University Press. p. 289-316.

Bockaert J, Pin JP. 1999. Molecular tinkering of G protein-coupled receptors: an evolutionary success. *EMBO J*. 18:1723-1729.

Borowsky B, Adham N, Jones KA, Raddatz R, Artymyshyn R, Ogozalek KL, Durkin MM, Lakhlani PP, Bonini JA, Pathirana S, Boyle N, Pu X, Kouranova E, Lichtblau H, Ochoa FY, Branchek TA, Gerald C. 2001. Trace amines: Identification of a family of mammalian G protein-coupled receptors. *Proc Natl Acad Sci USA*. 98:8966-8971.

Boto T, Gomez-Diaz C, Alcorta E. 2010. Expression Analysis of the 3 G-Protein Subunits, Gα, Gβ, and Gγ, in the Olfactory Receptor Organs of Adult Drosophila melanogaster. *Chem Senses*. 35:183-193.

Branchek TA, Blackburn TP. 2003. Trace amine receptors as targets for novel therapeutics: legend, myth and fact. *Curr Opin Pharmacol*. 3:90-97.

Breznak JA, Brune A. 1994. Role of Microorganisms in the Digestion of Lignocellulose by Termites. *Annu Rev Entomol*. 39:453-487.

Bridges CB. 1936. The bar "gene" a duplication. *Science*. 83:210-211.

Broadley KJ. 2010. The vascular effects of trace amines and amphetamines. *Pharmacol Ther*. 125:363-375.

Brody T, Cravchik A. 2000. Drosophila melanogaster G Protein-coupled Receptors. *J Cell Biol*. 150:83F-88.

Brown DD, Sugimoto K. 1973. 5S DNAs of *Xenopus laevis* and *Xenopus mulleri*: evolution of a gene family. *J Mol Biol*. 78:397–415.

Brune A. 2003. Symbionts aiding digestion. In: VH Resh, RT Cardé, editors. Encyclopedia of Insects: Academic Press, New York, N.Y. p. 1102–1107.

Brune A, Ohkuma M. 2011. Role of the Termite Gut Microbiota in Symbiotic Digestion. In: DE Bignell, Y Roisin, N Lo, editors. Biology of Termites: A Modern Synthesis: Springer. p. 439–476.

Brune A, Stingl U. 2005. Prokaryotic symbionts of termite gut flagellates: phylogenetic and metabolic implications of a tripartite symbiosis. In: J Overmann, editor. Progress in molecular and subcellular biology. New York, NY: Berlin, Germany: Springer. p. 39–60.

Buck L, Axel R. 1991. A Novel Multigene Family May Encode Odorant Receptors - a Molecular-Basis for Odor Recognition. *Cell*. 65:175-187.

Bunzow JR, Sonders MS, Arttamangkul S, Harrison LM, Zhang G, Quigley DI, Darland T, Suchland KL, Pasumamula S, Kennedy JL, Olson SB, Magenis RE, Amara SG, Grandy DK. 2001. Amphetamine, 3,4-Methylenedioxymethamphetamine, Lysergic Acid Diethylamide, and Metabolites of the Catecholamine Neurotransmitters Are Agonists of a Rat Trace Amine Receptor. *Mol Pharmacol*. 60:1181-1188.

Burton RA, Gidley MJ, Fincher GB. 2010. Heterogeneity in the chemistry, structure and function of plant cell walls. *Nat Chem Biol*. 6:724-732.

Calderón-Cortés N, Quesada M, Watanabe H, Cano-Camacho H, Oyama K. 2012. Endogenous Plant Cell Wall Digestion: A Key Mechanism in Insect Evolution. *Annu Rev Ecol Evol Syst*. 43:45-71.

Calderón-Cortés N, Watanabe H, Cano-Camacho H, Zavala-Páramo G, Quesada M. 2010. cDNA cloning, homology modelling and evolutionary insights into novel endogenous cellulases of the borer beetle *Oncideres albomarginata chamela* (Cerambycidae). *Insect Mol Biol*. 19:323-336.

Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. 2009. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res*. 37:D233-D238.

Chang C-J, Wu CP, Lu S-C, Chao A-L, Ho T-HD, Yu S-M, Chao Y-C. 2012. A novel exo-cellulase from white spotted longhorn beetle (Anoplophora malasiaca). *Insect Biochem Mol Biol*. 42:629-636.

Chapman JA, Kirkness EF, Simakov O, Hampson SE, Mitros T, Weinmaier T, Rattei T, Balasubramanian PG, Borman J, Busam D, Disbennett K, Pfannkoch C, Sumin N, Sutton GG, Viswanathan LD, Walenz B, Goodstein DM, Hellsten U, Kawashima T, Prochnik SE, Putnam NH, Shu S, Blumberg B, Dana CE, Gee L, Kibler DF, Law L, Lindgens D, Martinez DE, Peng J, Wigge PA, Bertulat B, Guder C, Nakamura Y, Ozbek S, Watanabe H, Khalturin K, Hemmrich G, Franke A, Augustin R, Fraune S, Hayakawa E, Hayakawa S, Hirose M, Hwang JS, Ikeo K, Nishimiya-Fujisawa C, Ogura A, Takahashi T, Steinmetz PRH, Zhang X, Aufschnaiter R, Eder M-K, Gorny A-K, Salvenmoser W, Heimberg AM, Wheeler BM, Peterson KJ, Bottger A, Tischler P, Wolf A, Gojobori T, Remington KA, Strausberg RL, Venter JC, Technau U, Hobmayer B, Bosch TCG, Holstein TW, Fujisawa T, Bode HR, David CN, Rokhsar DS, Steele RE. 2010. The dynamic genome of Hydra. *Nature*. 464:592-596.

Cherezov V, Rosenbaum DM, Hanson MA, Rasmussen SGF, Thian FS, Kobilka TS, Choi H-J, Kuhn P, Weis WI, Kobilka BK, Stevens RC. 2007. High-Resolution Crystal Structure of an Engineered Human β_2-Adrenergic G Protein-Coupled Receptor. *Science*. 318:1258-1265.

Cherqui A, Tjallingii WF. 2000. Salivary proteins of aphids, a pilot study on identification, separation and immunolocalisation. *Journal of Insect Physiology*. 46:1177-1186.

Chung S, Funakoshi T, Civelli O. 2008. Orphan GPCR research. *Br J Pharmacol*. 153:S339-S346.

Churcher AM, Taylor JS. 2011. The Antiquity of Chordate Odorant Receptors Is Revealed by the Discovery of Orthologs in the Cnidarian *Nematostella vectensis*. *Genome Biol Evol*. 3:36-43.

Civelli O. 2012. Orphan GPCRs and Neuromodulation. *Neuron*. 76:12-21.

Civelli O, Reinscheid RK, Zhang Y, Wang Z, Fredriksson R, Schiöth HB. 2013. G Protein–Coupled Receptor Deorphanizations. *Annu Rev Pharmacol Toxicol*. 53:127-146.

Civelli O, Saito Y, Wang Z, Nothacker H-P, Reinscheid RK. 2006. Orphan GPCRs and their ligands. *Pharmacol Ther*. 110:525-532.

Clyne PJ, Warr CG, Carlson JR. 2000. Candidate Taste Receptors in *Drosophila*. *Science*. 287:1830-1834.

Clyne PJ, Warr CG, Freeman MR, Lessing D, Kim JH, Carlson JR. 1999. A novel family of divergent seven-transmembrane proteins: Candidate odorant receptors in *Drosophila*. *Neuron*. 22:327-338.

Cooper JG, Bhatnagar KP. 1976. Comparative anatomy of the vomeronasal organ complex in bats. *J Anat*. 122:571–601.

Davison A, Blaxter M. 2005. Ancient origin of glycosyl hydrolase family 9 cellulase genes. *Mol Biol Evol*. 22:1273-1284.

Dayhoff MO, Schwartz RM, Orcutt BC. 1978. A model of evolutionary change in proteins. *Atlas of protein sequece and structure*. 5:345-352.

Dehal P. 2002. The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins. *Science*. 298:2157-2167.

Duan J, Martinez M, Sanders AR, Hou C, Saitou N, Kitano T, Mowry BJ, Crowe RR, Silverman JM, Levinson DF, Gejman PV. 2004. Polymorphisms in the Trace Amine Receptor 4 (TRAR4) Gene on Chromosome 6q23.2 Are Associated with Susceptibility to Schizophrenia. *Am J Hum Genet*. 75:624-638.

Dulac C, Axel R. 1995. A novel family of genes encoding putative pheromone receptors in mammals. *Cell*. 83:195-206.

Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD, Sorensen MV, Haddock SHD, Schmidt-Rhaesa A, Okusu A, Kristensen RM, Wheeler WC, Martindale MQ, Giribet G. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*. 452:745-749.

Dunning Hotopp JC. 2011. Horizontal gene transfer between bacteria and animals. *Trends Genet*. 27:157-163.

Eigenheer AL, Keeling CI, Young S, Tittiger C. 2003. Comparison of gene representation in midguts from two phytophagous insects, Bombyx mori and Ips pini, using expressed sequence tags. *Gene*. 316:127-136.

Elder JF, Jr., Turner BJ. 1995. Concerted Evolution of Repetitive DNA Sequences in Eukaryotes. *The Quarterly Review of Biology*. 70:297-320.

Eyun S, Moriyama H, Hoffmann FG, Moriyama EN. submitted. Molecular Evolution and Functional Divergence of Trace Amine–Associated Receptors. *Genome Biol Evol*.

Ferrero DM, Lemon JK, Fluegge D, Pashkovski SL, Korzan WJ, Datta SR, Spehr M, Fendt M, Liberles SD. 2011. Detection and avoidance of a carnivore odor by prey. *Proc Natl Acad Sci USA*. 108:11235-11240.

Ferrero DM, Wacker D, Roque MA, Baldwin MW, Stevens RC, Liberles SD. 2012. Agonists for 13 Trace Amine-Associated Receptors Provide Insight into the Molecular Basis of Odor Selectivity. *ACS Chem Biol*. 7:1184-1189.

Flames N, Hobert O. 2011. Transcriptional Control of the Terminal Fate of Monoaminergic Neurons. *Annu Rev Neurosci*. 34:153-184.

Fleischer J, Schwarzenbacher K, Breer H. 2007. Expression of Trace Amine–Associated Receptors in the Grueneberg Ganglion. *Chem Senses*. 32:623-631.

Foord SM, Bonner TI, Neubig RR, Rosser EM, Pin J-P, Davenport AP, Spedding M, Harmar AJ. 2005. International Union of Pharmacology. XLVI. G Protein-Coupled Receptor List. *Pharmacol Rev*. 57:279-288.

Fredriksson R, Lagerstrom MC, Lundin L-G, Schiöth HB. 2003. The G-Protein-Coupled Receptors in the Human Genome Form Five Main Families. Phylogenetic Analysis, Paralogon Groups, and Fingerprints. *Mol Pharmacol*. 63:1256-1272.

Fujita K, Shimomura K, Yamamoto K, Yamashita T, Suzuki K. 2006. A chitinase structurally related to the glycoside hydrolase family 48 is indispensable for the hormonally induced diapause termination in a beetle. *Biochem Biophys Res Commun*. 345:502-507.

Gamauf C, Metz B, Seiboth B. 2012. Degradation of Plant Cell Wall Polymers by Fungi. In: K Esser, editor. The Mycota, IX: Fungal Associations. Berlin and New York: Springer. p. 325-340.

Gao Q, Chess A. 1999. Identification of Candidate Drosophila Olfactory Receptors from Genomic DNA Sequence. *Genomics*. 60:31-39.

Garcia-Vallvé S, Romeu A, Palau J. 2000. Horizontal Gene Transfer of Glycosyl Hydrolases of the Rumen Fungi. *Mol Biol Evol*. 17:352-361.

Genta FA, Bragatto I, Terra WR, Ferreira C. 2009. Purification, characterization and sequencing of the major β-1,3-glucanase from the midgut of Tenebrio molitor larvae. *Insect Biochem Mol Biol*. 39:861-874.

Gilbert HJ. 2010. The Biochemistry and Structural Biology of Plant Cell Wall Deconstruction. *Plant Physiology*. 153:444-455.

Girard C, Jouanin L. 1999. Molecular cloning of a gut-specific chitinase cDNA from the beetle *Phaedon cochleariae*. *Insect Biochem Mol Biol*. 29:549-556.

Gladstone S, Hruska A. 2003. Guidelines for Promoting Safer and More Effective Pest Management with Small Holder Farmers: a Contribution to USAID-FFP Environmental Compliance. Georgia, USA: CARE USA.

Gloriam D, Fredriksson R, Schiöth HB. 2007. The G protein-coupled receptor subset of the rat genome. *BMC Genomics*. 8:338.

Glusman G, Yanai I, Rubin I, Lancet D. 2001. The Complete Human Olfactory Subgenome. *Genome Res*. 19:685-702.

Go Y, Niimura Y. 2008. Similar numbers but different repertoires of olfactory receptor genes in humans and chimpanzees. *Mol Biol Evol*. 25:1897-1907.

Göbbel L, Fischer MS, Smith TD, Wible JR, Bhatnagar KP. 2004. The vomeronasal organ and associated structures of the fetal African elephant, *Loxodonta africana* (Proboscidea, Elephantidae). *Acta Zoologica*. 85:41-52.

Gogos A, Kwek P, Chavez C, van den Buuse M. 2010. Estrogen Treatment Blocks 8-Hydroxy-2-dipropylaminotetralin- and Apomorphine-Induced Disruptions of Prepulse Inhibition: Involvement of Dopamine D_1 or D_2 or Serotonin 5-HT_{1A}, 5-HT_{2A}, or 5-HT_7 Receptors. *J Pharmacol Exp Ther*. 333:218-227.

Grus WE, Shi P, Zhang J. 2007. Largest Vertebrate Vomeronasal Type 1 Receptor Gene Repertoire in the Semiaquatic Platypus. *Mol Biol Evol*. 24:2153-2157.

Hansson Bill S, Stensmyr Marcus C. 2011. Evolution of Insect Olfaction. *Neuron*. 72:698-711.

Harada Y, Hosoiri Y, Kuroda R. 2004. Isolation and evaluation of dextral-specific and dextral-enriched cDNA clones as candidates for the handedness-determining gene in a freshwater gastropod, Lymnaea stagnalis. *Dev Genes Evol*. 214:159-169.

Hargrave P, McDowell J, Curtis D, Wang J, Juszczak E, Fong S, Rao J, Argos P. 1983. The structure of bovine rhodopsin. *Biophys Struct Mech*. 9:235-244.

Hashiguchi Y, Nishida M. 2007. Evolution of Trace Amine-Associated Receptor (TAAR) Gene Family in Vertebrates: Lineage-specific Expansions and Degradations of a Second Class of Vertebrate Chemosensory Receptors Expressed in the Olfactory Epithelium. *Mol Biol Evol*. 24:2099–2107.

Herrada G, Dulac C. 1997. A novel family of putative pheromone receptors in mammals with a topographically organized and sexually dimorphic distribution. *Cell*. 90:763-773.

Hibner BL, Burke WD, Eickbush TH. 1991. Sequence identity in an early chorion multigene family is the result of localized gene conversion. *Genetics*. 128:595-606.

Ho AK, Chik CL. 2000. Adrenergic Regulation of Mitogen-Activated Protein Kinase in Rat Pinealocytes: Opposing Effects of Protein Kinase A and Protein Kinase G. *Endocrinology*. 141:4496-4502.

Hongoh Y. 2011. Toward the functional analysis of uncultivable, symbiotic microorganisms in the termite gut. *Cell Mol Life Sci*. 68:1311-1325.

Hunt T, Bergsten J, Levkanicova Z, Papadopoulou A, John OS, Wild R, Hammond PM, Ahrens D, Balke M, Caterino MS, Gomez-Zurita J, Ribera I, Barraclough TG, Bocakova M, Bocak L, Vogler AP. 2007. A comprehensive phylogeny of beetles reveals the evolutionary origins of a superradiation. *Science*. 318:1913-1916.

Hussain A, Saraiva LR, Korsching SI. 2009. Positive Darwinian selection and the birth of an olfactory receptor clade in teleosts. *Proc Natl Acad Sci USA*. 106:4313-4318.

Ingram VM. 1961. Gene evolution and the haemoglobins. *Nature*. 189:704-708.

Jacq C, Miller JR, Brownlee GG. 1977. A pseudogene structure in 5S DNA of Xenopus laevis. *Cell*. 12:109-120.

Jeffery WR, Strickler AG, Yamamoto Y. 2003. To See or Not to See: Evolution of Eye Degeneration in Mexican Blind Cavefish. *Integrative and Comparative Biology*. 43:531-541.

Jeffreys AJ. 1979. DNA sequence variants in the $^{G}\gamma$-, $^{A}\gamma$-, δ- and β-globin genes of man. *Cell*. 18:1-10.

Juge N. 2006. Plant protein inhibitors of cell wall degrading enzymes. *Trends in Plant Science*. 11:359-367.

Jung K-H. 2007. The Distinct Signaling Mechanisms of Microbial Sensory Rhodopsins in Archaea, Eubacteria and Eukarya†. *Photochemistry and Photobiology*. 83:63-69.

Kamm B, Kamm M. 2004. Principles of biorefineries. *Appl Microbiol Biotechnol*. 64:137-145.

Kaupp UB. 2010. Olfactory signalling in vertebrates and insects: differences and commonalities. *Nat Rev Neurosci*. 11:188-200.

Keeling C, Yuen M, Liao N, Docking T, Chan S, Taylor G, Palmquist D, Jackman S, Nguyen A, Li M, Henderson H, Janes J, Zhao Y, Pandoh P, Moore R, Sperling F, Huber D, Birol I, Jones S, Bohlmann J. 2013. Draft genome of the mountain pine beetle, *Dendroctonus ponderosae* Hopkins, a major forest pest. *Genome Biol*. 14:R27.

Keeling CI, Henderson H, Li M, Yuen M, Clark EL, Fraser JD, Huber DP, Liao NY, Docking TR, Birol I, Chan SK, Taylor GA, Palmquist D, Jones SJ, Bohlmann J. 2012. Transcriptome and full-length cDNA resources for the mountain pine beetle, Dendroctonus ponderosae Hopkins, a major insect pest of pine forests. *Insect Biochem Mol Biol*. 42:525-536.

Keeling PJ. 2009. Functional and ecological impacts of horizontal gene transfer in eukaryotes. *Curr Opin Genet Dev*. 19:613-619.

Kim N, Choo YM, Lee KS, Hong SJ, Seol KY, Je YH, Sohn HD, Jin BR. 2008. Molecular cloning and characterization of a glycosyl hydrolase family 9 cellulase distributed throughout the digestive tract of the cricket Teleogryllus emma. *Comp Biochem Physiol B Biochem Mol Biol*. 150:368-376.

Kirsch R, Wielsch N, Vogel H, Svatos A, Heckel D, Pauchet Y. 2012. Combining proteomics and transcriptome sequencing to identify active plant-cell-wall-degrading enzymes in a leaf beetle. *BMC Genomics*. 13:587.

Kortschak RD, Samuel G, Saint R, Miller DJ. 2003. EST Analysis of the Cnidarian *Acropora millepora* Reveals Extensive Gene Loss and Rapid Sequence Divergence in the Model Invertebrates. *Curr Biol*. 24:2190-2195.

Krishnan A, Almén MS, Fredriksson R, Schiöth HB. 2012. The Origin of GPCRs: Identification of Mammalian like *Rhodopsin*, *Adhesion*, *Glutamate* and *Frizzled* GPCRs in Fungi. *PLoS ONE*. 7:e29817.

Kunieda T, Fujiyuki T, Kucharski R, Foret S, Ament SA, Toth AL, Ohashi K, Takeuchi H, Kamikouchi A, Kage E, Morioka M, Beye M, Kubo T, Robinson GE, Maleszka R. 2006. Carbohydrate metabolism genes and pathways in insects: insights from the honey bee genome. *Insect Mol Biol*. 15:563-576.

Lagerstrom MC, Schioth HB. 2008. Structural diversity of G protein-coupled receptors and significance for drug discovery. *Nat Rev Drug Discov*. 7:339-357.

Laughlin JD, Ha TS, Jones DNM, Smith DP. 2008. Activation of Pheromone-Sensitive Neurons Is Mediated by Conformational Activation of Pheromone-Binding Protein. *Cell*. 133:1255-1265.

Ledonne A, Berretta N, Davoli A, Rizzo GR, Bernardi G, Mercuri NB. 2011. Electrophysiological effects of trace amines on mesencephalic dopaminergic neurons. *Front Syst Neurosci*. 5:56.

Lee SJ, Kim SR, Yoon HJ, Kim I, Lee KS, Je YH, Lee SM, Seo SJ, Dae Sohn H, Jin BR. 2004. cDNA cloning, expression, and enzymatic activity of a cellulase from the mulberry longicorn beetle, *Apriona germari*. *Comp Biochem Physiol B*. 139:107-116.

Li L, Frohlich J, Pfeiffer P, Konig H. 2003. Termite gut symbiotic archaezoa are becoming living metabolic fossils. *Eukaryot Cell*. 2:1091-1098.

Li W-H. 1997. Molecular Evolution Sunderland, Massachusetts: Sinauer Associates.

Li W-H, Gojobori T, Nei M. 1981. Pseudogenes as a paradigm of neutral evolution. *Nature*. 292:237–239.

Li X, Staszewski L, Xu H, Durick K, Zoller M, Adler E. 2002. Human receptors for sweet and umami taste. *Proc Natl Acad Sci USA*. 99:4692-4696.

Liberles S, D. 2009. Trace Amine-associated Receptors Are Olfactory Receptors in Vertebrates. *Ann N Y Acad Sci*. 1170:168-172.

Liberles SD, Buck LB. 2006. A second class of chemosensory receptors in the olfactory epithelium. *Nature*. 442:645-650.

Libertini E, Li Y, McQueen-Mason S. 2004. Phylogenetic Analysis of the Plant Endo-β-1,4-Glucanase Gene Family. *J Mol Evol*. 58:506-515.

Lindemann L, Ebeling M, Kratochwil NA, Bunzow JR, Grandy DK, Hoener MC. 2005. Trace amine-associated receptors form structurally and functionally distinct subfamilies of novel G protein-coupled receptors. *Genomics*. 85:372-385.

Lo N, Tokuda G, Watanabe H, Rose H, Slaytor M, Maekawa K, Bandi C, Noda H. 2000. Evidence from multiple gene sequences indicates that termites evolved from wood-feeding cockroaches. *Curr Biol*. 10:801-804.

Lo N, Watanabe H, Sugimura M. 2003. Evidence for the Presence of a Cellulase Gene in the Last Common Ancestor of Bilaterian Animals. *Proceedings: Biological Sciences*. 270:S69-S72.

Lynch M, O'Hely M, Walsh B, Force A. 2001. The probability of preservation of a newly arisen gene duplicate. *Genetics*. 159:1789-1804.

Mackay-Sim A, Duvall D, Graves BM. 1985. The West Indian Manatee *Trichechus manatus* lacks a vomeronasal organ. *Brain, Behavior and Evolution*. 27:186-194.

Maguire JJ, Parker WAE, Foord SM, Bonner TI, Neubig RR, Davenport AP. 2009. International Union of Pharmacology. LXXII. Recommendations for Trace Amine Receptor Nomenclature. *Pharmacol Rev*. 61:1-8.

Maier W. 1997. The nasopalatine duct and the nasal floor cartilages in catarrhine primates. *Zeitschrift für Morphologie und Anthropologie*. 81:289-300.

Markovič O, Janeček Š. 2001. Pectin degrading glycoside hydrolases of family 28: sequence-structural features, specificities and evolution. *Protein Eng*. 14:615-631.

Martin M. 1991. The evolution of cellulose digestion in insects. *Philos Trans R Soc Lond Ser B*. 333:281–288.

Martin MM. 1983. Cellulose digestion in insects. *Comp Biochem Physiol A*. 75:313-324.

Matsunami H, Amrein H. 2003. Taste and pheromone perception in mammals and flies. *Genome Biol*. 4:9.

Matsunami H, Buck LB. 1997. A multigene family encoding a diverse array of putative pheromone receptors in mammals. *Cell*. 90:775-784.

Matsunami H, Montmayeur JP, Buck LB. 2000. A family of candidate taste receptors in human and mouse. *Nature*. 404:601-604.

McBride CS, Arguello JR. 2007. Five Drosophila Genomes Reveal Nonneutral Evolution and the Signature of Host Specialization in the Chemoreceptor Superfamily. *Genetics*. 177:1395-1416.

Melillo JM, Steudler PA, Aber JD, Newkirk K, Lux H, Bowles FP, Catricala C, Magill A, Ahrens T, Morrisseau S. 2002. Soil Warming and Carbon-Cycle Feedbacks to the Climate System. *Science*. 298:2173-2176.

Millan MJ, Marin P, Bockaert J, Mannoury la Cour C. 2008. Signaling at G-protein-coupled serotonin receptors: recent advances and future research directions. *Trends Pharmacol Sci*. 29:454-464.

Mombaerts P. 2004. Genes and ligands for odorant, vomeronasal and taste receptors. *Nature reviews Neuroscience*. 5:263-278.

Moran NA, Jarvik T. 2010. Lateral transfer of genes from fungi underlies carotenoid production in aphids. *Science*. 328:624-627.

Moriyama EN, Opiyo SO. 2010. Bioinformatics of Seven-Transmembrane Receptors in Plant Genomes. In: S Yalovsky, F Baluška, A Jones, editors. Integrated G Proteins Signaling in Plants: Springer Berlin Heidelberg. p. 251-277.

Mueller JC, Steiger S, Fidler AE, Kempenaers B. 2008. Biogenic Trace Amine-Associated Receptors (TAARs) are encoded in avian genomes: evidence and possible implications. *J Hered*. 99:174-176.

Nakashima K, Watanabe H, Saitoh H, Tokuda G, Azuma JI. 2002. Dual cellulose-digesting system of the wood-feeding termite, Coptotermes formosanus Shiraki. *Insect Biochem Mol Biol*. 32:777-784.

Nathans J, Hogness DS. 1983. Isolation, sequence analysis, and intron-exon arrangement of the gene encoding bovine rhodopsin. *Cell*. 34:807-814.

Nei M, Hughes A. 1992. Balanced polymorphism and evolution by the birth-and-death process in the MHC loci. In: K Tsuji, M Aizawa, T Sasazuki, editors. 11th Histocompatibility Workshop and Conference. Oxford, UK: Oxford Univ. Press. p. 27–38.

Nei M, Kumar S. 2000. Molecular Evolution and Phylogenetics. New York, USA: Oxford University Press.

Nei M, Niimura Y, Nozawa M. 2008. The evolution of animal chemosensory receptor gene repertoires: roles of chance and necessity. *Nat Rev Genet*. 9:951-963.

Nei M, Rooney AP. 2005. Concerted and birth-and-death evolution of multigene families. *Annu Rev Genet*. 39:121-152.

Niimura Y, Nei M. 2005. Evolutionary dynamics of olfactory receptor genes in fishes and tetrapods. *Proc Natl Acad Sci USA*. 102:6039-6044.

Nordström KJV, Sällman Almén M, Edstam MM, Fredriksson R, Schiöth HB. 2011. Independent HHsearch, Needleman–Wunsch-Based, and Motif Analyses Reveal the Overall Hierarchy for Most of the G Protein-Coupled Receptor Families. *Mol Biol Evol*. 28:2471-2480.

Oelschläger HA. 1989. Early development of the olfactory and terminalis systems in baleen whales. *Brain, Behavior and Evolution*. 34:171-183.

Ohno S. 1970. Evolution by gene duplication. Berlin, New York,: Springer-Verlag.

Ovchinnikov Y. 1982. Rhodopsin and bacteriorhodopsin: structure-function relationships. *FEBS Lett*. 148:179-191.

Overington JP, Al-Lazikani B, Hopkins AL. 2006. How many drug targets are there? *Nat Rev Drug Discov*. 5:993-996.

Pae C-U, Drago A, Kim J-J, Patkar AA, Jun T-Y, De Ronchi D, Serretti A. 2010. TAAR6 variations possibly associated with antidepressant response and suicidal behavior. *Psychiatry Research*. 180:20-24.

Palczewski K, Kumasaka T, Hori T, Behnke CA, Motoshima H, Fox BA, Trong IL, Teller DC, Okada T, Stenkamp RE, Yamamoto M, Miyano M. 2000. Crystal Structure of Rhodopsin: A G Protein-Coupled Receptor. *Science*. 289:739-745.

Parfrey LW, Grant J, Tekle YI, Lasek-Nesselquist E, Morrison HG, Sogin ML, Patterson DJ, Katz LA. 2010. Broadly Sampled Multigene Analyses Yield a Well-Resolved Eukaryotic Tree of Life. *Syst Biol*. 59:518-533.

Pauchet Y, Freitak D, Heidel-Fischer HM, Heckel DG, Vogel H. 2009. Immunity or Digestion: GLUCANASE ACTIVITY IN A GLUCAN-BINDING PROTEIN FAMILY FROM LEPIDOPTERA. *J Biol Chem*. 284:2214-2224.

Pauchet Y, Heckel DG. 2013. The genome of the mustard leaf beetle encodes two active xylanases originally acquired from bacteria through horizontal gene transfer. *Proceedings of the Royal Society B: Biological Sciences*. 280.

Pauchet Y, Wilkinson P, Chauhan R, ffrench-Constant RH. 2010. Diversity of Beetle Genes Encoding Novel Plant Cell Wall Degrading Enzymes. *PLoS ONE*. 5:e15635.

Peatfield R, Littlewood JT, Glover V, Sandler M, Rose FC. 1983. Pressor sensitivity to tyramine in patients with headache: relationship to platelet monoamine oxidase and to dietary provocation. *J Neurol Neurosurg Psychiatry*. 46:827–831.

Putnam NH, Srivastava M, Hellsten U, Dirks B, Chapman J, Salamov A, Terry A, Shapiro H, Lindquist E, Kapitonov VV, Jurka J, Genikhovich G, Grigoriev IV, Lucas SM, Steele RE, Finnerty JR, Technau U, Martindale MQ, Rokhsar DS. 2007. Sea Anemone Genome Reveals Ancestral Eumetazoan Gene Repertoire and Genomic Organization. *Science*. 317:86-94.

Rastogi S, Liberles DA. 2005. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evol Biol*. 5.

Ringstad N, Abe N, Horvitz HR. 2009. Ligand-Gated Chloride Channels Are Receptors for Biogenic Amines in *C. elegans*. *Science*. 325:96-100.

Riviere S, Challet L, Fluegge D, Spehr M, Rodriguez I. 2009. Formyl peptide receptor-like proteins are a novel family of vomeronasal chemosensors. *Nature*. 459:574-577.

Robert S, Bichet A, Grandjean O, Kierzkowski D, Satiat-Jeunemaître B, Pelletier S, Hauser M-T, Höfte H, Vernhettes S. 2005. An *Arabidopsis* Endo-1,4-β-D-Glucanase Involved in Cellulose Synthesis Undergoes Regulated Intracellular Cycling. *Plant Cell*. 17:3378-3389.

Rybarczyk-Mydlowska K, Maboreke HR, van Megen H, van den Elsen S, Mooyman P, Smant G, Bakker J, Helder J. 2012. Rather than by direct acquisition via lateral gene transfer, GHF5 cellulases were passed on from early Pratylenchidae to root-knot and cyst nematodes. *BMC Evol Biol*. 12:221.

Saavedra J. 1980. Increased adrenaline, beta-adrenoreceptor stimulation and phospholipid methylation in pineal gland of spontaneously hypertensive rats. *Clin Sci*.239s-242s.

Sällman Almén M, Nordström KJV, Fredriksson R, Schiöth HB. 2009. Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin. *BMC Biol*. 7:50.

Sam M, Vora S, Malnic B, Ma W, Novotny MV, Buck LB. 2001. Neuropharmacology: Odorants may arouse instinctive behaviours. *Nature*. 412:142-142.

Sato K, Pellegrino M, Nakagawa T, Nakagawa T, Vosshall LB, Touhara K. 2008. Insect olfactory receptors are heteromeric ligand-gated ion channels. *Nature*. 452:1002-1006.

Scharf ME, Karl ZJ, Sethi A, Boucias DG. 2011. Multiple Levels of Synergistic Collaboration in Termite Lignocellulose Digestion. *PLoS ONE*. 6:e21709.

Schoeberg T, Schulz A, Biebermann H, Hermsdorf T, Rompler H, Sangkuhl K. 2004. Mutant G-protein-coupled receptors as a cause of human diseases. *Pharmacol Ther*. 104:173-206.

Scrivener AM, Slaytor M. 1994. Properties of the endogenous cellulase from *Panesthia cribrata* Saussure and purification of major endo-β-1,4-glucanase components. *Insect Biochem Mol Biol*. 24:223–231.

Serretti A, Pae C-U, Chiesa A, Mandelli L, De Ronchi D. 2009. Influence of TAAR6 polymorphisms on response to aripiprazole. *Prog Neuropsychopharmacol Biol Psychiatry*. 33:822-826.

Sharman JL, Benson HE, Pawson AJ, Lukito V, Mpamhanga CP, Bombail V, Davenport AP, Peters JA, Spedding M, Harmar AJ, NC-IUPHAR. 2013. IUPHAR-DB: updated database content and new features. *Nucleic Acids Res*. 41:D1083-D1088.

Shen L, Chen C, Zheng H, Jin L. 2013. The Evolutionary Relationship between Microbial Rhodopsins and Metazoan Rhodopsins. *The Scientific World Journal*. 2013:10.

Smant G, Stokkermans JP, Yan Y, de Boer JM, Baum TJ, Wang X, Hussey RS, Gommers FJ, Henrissat B, Davis EL, Helder J, Schots A, Bakker J. 1998. Endogenous cellulases in animals: isolation of beta-1, 4-endoglucanase genes from two species of plant-parasitic cyst nematodes. *Proc Natl Acad Sci USA*. 95:4906-4911.

Smart R, Kiely A, Beale M, Vargas E, Carraher C, Kralicek AV, Christie DL, Chen C, Newcomb RD, Warr CG. 2008. Drosophila odorant receptors are novel seven transmembrane domain proteins that can signal independently of heterotrimeric G proteins. *Insect Biochem Mol Biol*. 38:770-780.

Song JM, Nam K, Sun YU, Kang MH, Kim CG, Kwon ST, Lee J, Lee YH. 2010. Molecular and biochemical characterizations of a novel arthropod endo-beta-1,3-glucanase from the Antarctic springtail, Cryptopygus antarcticus, horizontally acquired from bacteria. *Comp Biochem Physiol B Biochem Mol Biol*. 155:403-412.

Spudich JL, Yang C-S, Jung K-H, Spudich EN. 2000. RETINYLIDENE PROTEINS: Structures and Functions from Archaea to Humans. *Annual Review of Cell and Developmental Biology*. 16:365-392.

Srivastava M, Begovic E, Chapman J, Putnam NH, Hellsten U, Kawashima T, Kuo A, Mitros T, Salamov A, Carpenter ML, Signorovitch AY, Moreno MA, Kamm K, Grimwood J, Schmutz J, Shapiro H, Grigoriev IV, Buss LW, Schierwater B, Dellaporta SL, Rokhsar DS. 2008. The *Trichoplax* genome and the nature of placozoans. *Nature*. 454:955-960.

Srivastava M, Simakov O, Chapman J, Fahey B, Gauthier MEA, Mitros T, Richards GS, Conaco C, Dacre M, Hellsten U, Larroux C, Putnam NH, Stanke M, Adamska M, Darling A, Degnan SM, Oakley TH, Plachetzki DC, Zhai Y, Adamski M, Calcino A, Cummins SF, Goodstein DM, Harris C, Jackson DJ, Leys SP, Shu S, Woodcroft BJ, Vervoort M, Kosik KS, Manning G, Degnan BM, Rokhsar DS. 2010. The Amphimedon queenslandica genome and the evolution of animal complexity. *Nature*. 466:720-726.

Steenbakkers PJM, Ubhayasekera W, Goossen HJAM, van Lierop EMHM, van der Drift C, Vogels GD, Mowbray SL, Op den Camp HJM. 2002. An intron-containing glycoside hydrolase family 9 cellulase gene encodes the dominant 90 kDa component of the cellulosome of the anaerobic fungus *Piromyces* sp. strain E2. *Biochem J*. 365:193-204.

Steiger S, Fidler A, Kempenaers B. 2009. Evidence for increased olfactory receptor gene repertoire size in two nocturnal bird species with well-developed olfactory ability. *BMC Evol Biol*. 9:117.

Stoddart DM. 1980. The Ecology of Vertebrate Olfaction: Chapman & Hall, London and New York.

Suárez R, García-González D, De Castro F. 2012. Mutual influences between the main olfactory and vomeronasal systems in development and evolution. *Frontiers in Neuroanatomy*. 6:50.

Sugimura M, Watanabe H, Lo N, Saito H. 2003. Purification, characterization, cDNA cloning and nucleotide sequencing of a cellulase from the yellow-spotted longicorn beetle, Psacothea hilaris. *Eur J Biochem*. 270:3455-3460.

Suzuki K-i, Ojima T, Nishita K. 2003. Purification and cDNA cloning of a cellulase from abalone *Haliotis discus hannai*. *Eur J Biochem*. 270:771-778.

Syvanen M. 1984. Conserved regions in mammalian beta-globins: could they arise by cross-species gene exchange? *J Theor Biol*. 107:685-696.

Syvanen M. 1985. Cross-species gene transfer; implications for a new theory of evolution. *J Theor Biol*. 112:333-343.

Syvanen M. 2012. Evolutionary Implications of Horizontal Gene Transfer. *Annu Rev Genet*. 46:341-358.

Tokuda G, Watanabe H. 2007. Hidden cellulases in termites: revision of an old hypothesis. *Biology Letters*. 3:336-339.

Tokuda G, Watanabe H, Lo N. 2007. Does correlation of cellulase gene expression and cellulolytic activity in the gut of termite suggest synergistic collaboration of cellulases? *Gene*. 401:131-134.

Tomme P, Warren RAJ, Gilkes NR. 1995. Cellulose Hydrolysis by Bacteria and Fungi. In: RK Poole, editor. Advances in Microbial Physiology: Academic Press. p. 1-81.

Tribolium Genome Sequencing Consortium. 2008. The genome of the model beetle and pest Tribolium castaneum. *Nature*. 452:949-955.

Valentine BD. 2005. The scientific name of the coffee bean weevil and some additional bibliography (Coleoptera: Anthribidae: *Araecerus* Schönherr). *Insecta Mundi* 19:247-253.

Van Belleghem SM, Roelofs D, Van Houdt J, Hendrickx F. 2012. *De novo* Transcriptome Assembly and SNP Discovery in the Wing Polymorphic Salt Marsh Beetle *Pogonus chalceus* (Coleoptera, Carabidae). *PLoS ONE*. 7:e42605.

Vanin EF. 1985. Processed Pseudogenes: Characteristics and Evolution. *Annu Rev Genet*. 19:253-272.

Vladimirov V, Thiselton DL, Kuo PH, McClay J, Fanous A, Wormley B, Vittum J, Ribble R, Moher B, van den Oord E, O'Neill FA, Walsh D, Kendler KS, Riley BP. 2007. A region of 35 kb containing the trace amine associate receptor 6 (TAAR6) gene is associated with schizophrenia in the Irish study of high-density schizophrenia families. *Mol Psychiatry*. 12:842-853.

Vosshall LB, Amrein H, Morozov PS, Rzhetsky A, Axel R. 1999. A spatial map of olfactory receptor expression in the Drosophila antenna. *Cell*. 96:725-736.

Vosshall LB, Stocker RE. 2007. Molecular architecture of smell and taste in Drosophila. *Annu Rev Neurosci*. 30:505-533.

Vroling B, Sanders M, Baakman C, Borrmann A, Verhoeven S, Klomp J, Oliveira L, de Vlieg J, Vriend G. 2011. GPCRDB: information system for G protein-coupled receptors. *Nucleic Acids Res*. 39:D309-D319.

Waller JM, Bigger M, Hillocks RJ. 2007. Postharvest and processing pests and microbial problems. In: JM Waller, M Bigger, RJ Hillocks, editors. Coffee Pests, Diseases and Their Management. CABI, Wallingford, UK. p. 325–335.

Wang X, Grus WE, Zhang J. 2006. Gene Losses during Human Origins. *PLoS Biol*. 4:e52.

Warnecke F, Luginbuhl P, Ivanova N, Ghassemian M, Richardson TH, Stege JT, Cayouette M, McHardy AC, Djordjevic G, Aboushadi N, Sorek R, Tringe SG, Podar M, Martin HG, Kunin V, Dalevi D, Madejska J, Kirton E, Platt D, Szeto E, Salamov A, Barry K, Mikhailova N, Kyrpides NC, Matson EG, Ottesen EA, Zhang X, Hernandez M, Murillo C, Acosta LG, Rigoutsos I, Tamayo G, Green BD, Chang C, Rubin EM, Mathur EJ, Robertson DE, Hugenholtz P, Leadbetter JR. 2007. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature*. 450:560-565.

Watanabe H, Noda H, Tokuda G, Lo N. 1998. A cellulase gene of termite origin. *Nature*. 394:330-331.

Watanabe H, Tokuda G. 2010. Cellulolytic Systems in Insects. *Annu Rev Entomol*. 55:609-632.

Wei YD, Lee KS, Gui ZZ, Yoon HJ, Kim I, Zhang GZ, Guo X, Sohn HD, Jin BR. 2006. Molecular cloning, expression, and enzymatic activity of a novel endogenous cellulase from the mulberry longicorn beetle, *Apriona germari*. *Comp Biochem Physiol B Biochem Mol Biol*. 145:220-229.

Welling PG. 1996. Effects of Food on Drug Absorption. *Annual Review of Nutrition*. 16:383-415.

Wheeler D, Redding AJ, Werren JH. 2013. Characterization of an Ancient Lepidopteran Lateral Gene Transfer. *PLoS ONE*. 8:e59262.

Wible JR, Bhatnagar KP. 1996. Chiropteran vomeronasal complex and the interfamilial relationships of bats. *Journal of Mammalian Evolution*. 3:285-314.

Wicher D, Schafer R, Bauernfeind R, Stensmyr MC, Heller R, Heinemann SH, Hansson BS. 2008. Drosophila odorant receptors are both ligand-gated and cyclic-nucleotide-activated cation channels. *Nature*. 452:1007-1011.

Willis JD, Oppert B, Oppert C, Klingeman WE, Jurat-Fuentes JL. 2011. Identification, cloning, and expression of a GHF9 cellulase from Tribolium castaneum (Coleoptera: Tenebrionidae). *J Insect Physiol*. 57:300-306.

Wolinsky TD, Swanson CJ, Smith KE, Zhong H, Borowsky B, Seeman P, Branchek T, Gerald CP. 2007. The Trace Amine 1 receptor knockout mouse: an animal model with relevance to schizophrenia. *Genes, Brain and Behavior*. 6:628-639.

Wyman CE. 2003. Potential Synergies and Challenges in Refining Cellulosic Biomass to Fuels, Chemicals, and Power. *Biotechnology Progress*. 19:254-262.

Xu B, Janson J-C, Sellos D. 2001. Cloning and sequencing of a molluscan endo-β-1,4-glucanase gene from the blue mussel, Mytilus edulis. *Eur J Biochem*. 268:3718-3727.

Xue C, Hsueh Y-P, Heitman J. 2008. Magnificent seven: roles of G protein-coupled receptors in extracellular sensing in fungi. *FEMS Microbiol Rev*. 32:1010-1032.

Young JM, Friedman C, Williams EM, Ross JA, Tonnes-Priddy L, Trask BJ. 2002. Different evolutionary processes shaped the mouse and human olfactory receptor gene families. *Hum Mol Genet*. 11:535-546.

Zhang J. 2008. Positive selection, not negative selection, in the pseudogenization of rcsA in Yersinia pestis. *Proc Natl Acad Sci USA*. 105:E69.

Zhang JZ. 2003. Evolution by gene duplication: an update. *Trends Ecol Evol*. 18:292-298.

Zhang Y-HP, Lynd LR. 2005. Cellulose utilization by Clostridium thermocellum: Bioenergetics and hydrolysis product assimilation. *Proc Natl Acad Sci USA*. 102:7321-7325.

Zhou X, Smith JA, Oi FM, Koehler PG, Bennett GW, Scharf ME. 2007. Correlation of cellulase gene expression and cellulolytic activity throughout the gut of the termite Reticulitermes flavipes. *Gene*. 395:29-39.

Zimmer EA, Martin SL, Beverley SM, Kan YW, Wilson AC. 1980. Rapid duplication and loss of genes coding for the α chains of hemoglobin. *Proc Natl Acad Sci USA*. 77:2158-2162.

Zucchi R, Chiellini G, Scanlan TS, Grandy DK. 2006. Trace amine-associated receptors and their ligands. *Br J Pharmacol*. 149:967-978.

Zufall F, Leinders-Zufall T. 2007. Mammalian pheromone sensing. *Curr Opin Neurobiol*. 17:483-489.

Table 1.1. The classification of the GPCR superfamily.[a]

[Classes]	Examples
Class A: Rhodopsin-like family	Amine receptors, (Rhod)opsin, Olfactory (vertebrates), thyrotropin receptor, Cannabinoid receptors, Melatonin receptor, Leukotriene B4 receptor, Prostanoid receptor
Class B: Secretin-like family	Calcitonin receptor, Glucagon receptor, Parathyroid hormone receptor, Secretin receptor, Diuretic hormone receptor, Methuselah-like proteins (MTH), ERM1, Latrophilin recptor, Cadherin EGF LAG (CELSR), Depsiphilin
Class C: Metabotropic glutamate/pheromone family	Metabotropic glutamate receptor, Calcium sensing receptor, GABA-B receptor, Vomeronasal receptor type 2 (V2R), Taste receptor type 1 (T1R)
[Other groups]	
cAMP receptors	
Vomeronasal receptors (V1R and V3R)	Vomeronasal receptor type 1
Taste receptors T2R	Taste receptor type 2
Putative groups[b]	
Insect chemoreceptors	Odorant receptor, Gustatory receptor
Plant mildew-resistance locus O (MLO)	Plant mildew-resistance locus O
Nematode chemoreceptors	Serpentine receptor, str
Frizzled/Smoothened family	Frizzled, Smoothened

[a] This classification is based on the current version (ver. 11.3.4) of GPCRDB with modifications.

[b] Four putative groups are added from the original GPCRDB (http://www.gpcr.org/7tm_old).

Table 1.2. The numbers of chemosensory receptor genes in (a) vertebrates and (b) insects.[a]

(a)

	OR	**TAAR**[b]	**V1R**	**V2R**	**T1R**	**T2R**
Human	388 (414)	6 (3)	5 (115)	0 (20)	3 (0)	25 (11)
Mouse	1063 (328)	15 (1)	187 (121)	121 (158)	3 (0)	35 (6)
Dog	822 (278)	2 (2)	8 (33)	0 (9)	3 (0)	16 (5)
Cow	1152 (977)	21 (8)	40 (45)	0 (16)	3 (0)	19 (15)
Opossum	1198 (294)	22 (4)	98 (30)	86 (79)	3 (0)	29 (5)
Platypus	348 (370)	4 (1)	270 (579)	15 (112)	NA	NA
Chicken	300 (133)	4 (1)	0 (0)	0 (0)	2 (0)	3 (0)
Xenopus	1024 (614)	7 (0)	21 (2)	249 (448)	0 (0)	52 (12)
Zebrafish	155 (21)	110 (10)	2 (0)	44 (8)	1 (0)	4 (0)

(b)

	OR	**GR**
D. melanogaster	59 (2)[c]	68 (0)[c]
Yellow-fever mosquito	110 (21)	91 (23)
Silkworm	48 (NA)	NA
Red flour beetle	262 (79)	62 (NA)
Honeybee	163 (7)	10 (3)

[a] All numbers are taken from Nei et al. (2008) unless otherwise noted. The numbers of possible pseudogenes are shown in parentheses.

[b,c] The numbers are taken from the following literatures: Eyun et al. (submitted)[b] and McBride and Arguello (2007)[c].

NA: not available.

Table 1.3. Glycoside hydrolase classification by CAZy[a].

GH clans	Families[c]	Shared structural characteristics
GH-A	1, 2, 5, 10, 17, 26, 30, 35, 39, 42, 50, 51, 53, 59, 72, 79, 86, 113, 128	$(\beta/\alpha)_8$
GH-B	7, 16	β-jelly roll
GH-C	11, 12	β-jelly roll
GH-D	27, 31, 36	$(\beta/\alpha)_8$
GH-E	33, 34, 83, 93	6-fold β-propeller
GH-F	43, 62	5-fold β-propeller
GH-G	37, 63	$(\alpha/\alpha)_6$
GH-H	13, 70, 77	$(\beta/\alpha)_8$
GH-I	24, 46, 80	α+β
GH-J	32, 68	5-fold β-propeller
GH-K	18, 20, 85	$(\beta/\alpha)_8$
GH-L	15, 65, 125	$(\alpha/\alpha)_6$
GH-M	8, 48	$(\alpha/\alpha)_6$
GH-N	28, 49	β-helix
Non-Classified[b]	3, 4, 6, 9, 14, 19, 21, 22 23, 25, 29, 38, 40, 41, 44, 45, 47, 52, 54, 55, 56, 57, 58, 60, 61, 64, 66, 67, 69, 71, 73, 74, 75, 76, 78, 81, 82, 84, 87, 88, 89, 90, 91, 92, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 126, 127, 129, 130, 131, 132	

[a] The potential biological functions of various families are described in details at CAZy (http://www.cazy.org).

[b] These GHs cannot be categorized into any existing clans.

[c] The cellulolytic enzymes found in insects are indicated by red fonts.

Three classes of cellulolytic enzymes are indicated by cyan for endoglucanases (EC 3.2.1.4), purple for cellobiohydrolases (EC 3.2.1.74 and 3.2.1.91), and green for β-glucosidases (EC 3.2.1.21). Note that some families have multiple enzymatic activities and they are indicated with multiple colors (*e.g.*, GH1, GH5).

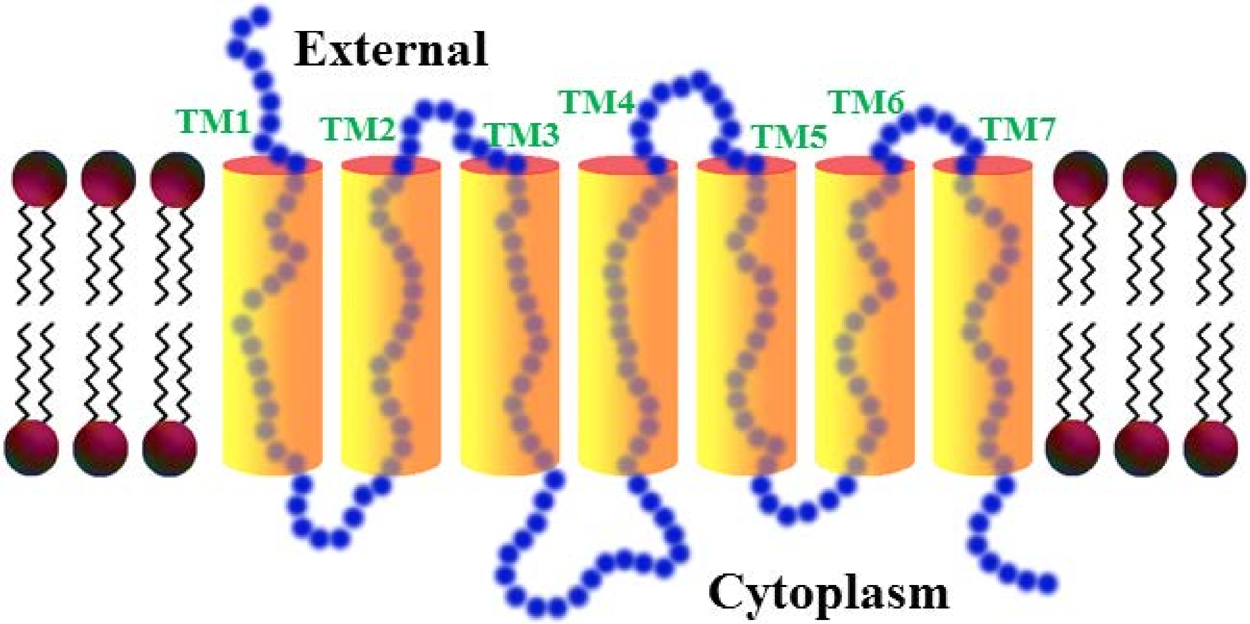
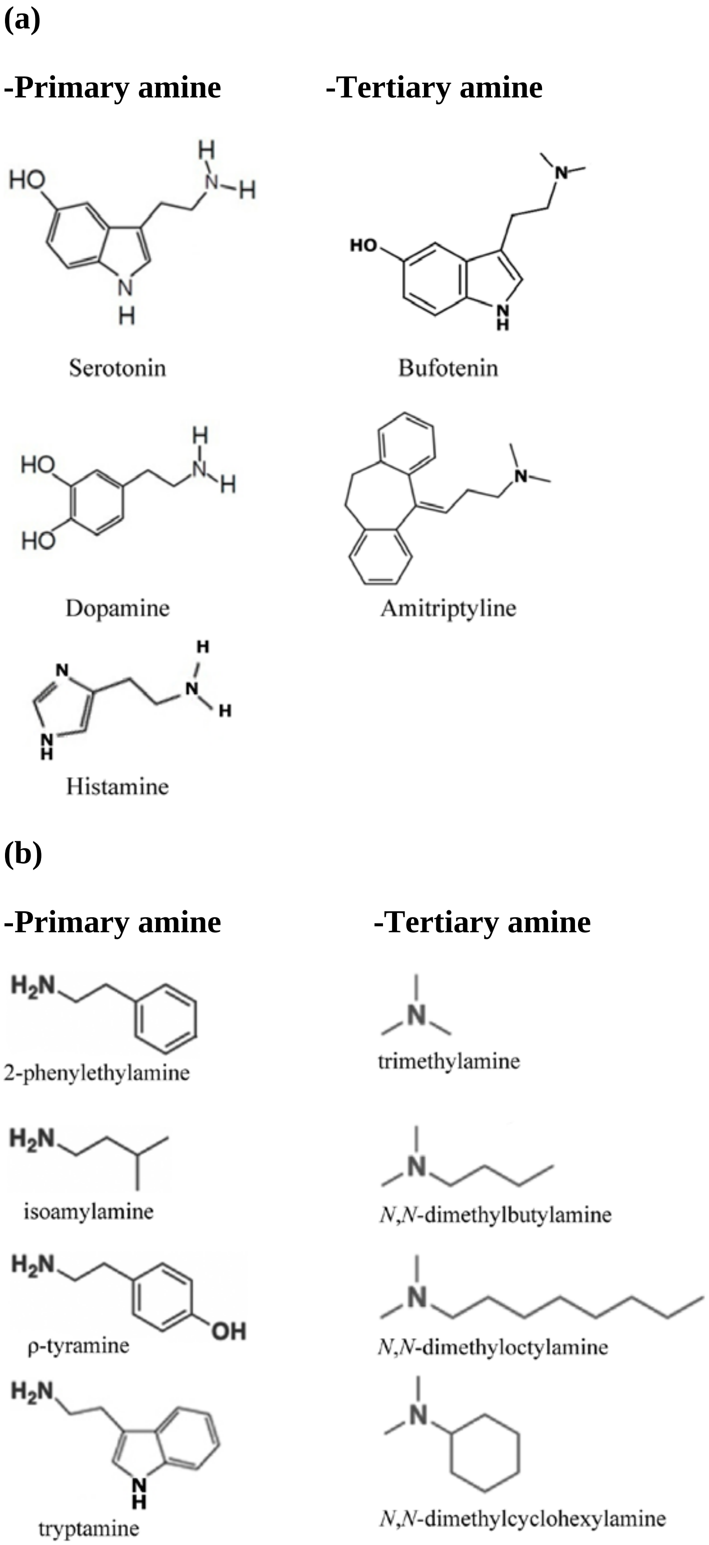


Figure 1.1. A model of a G-protein-coupled receptor protein. Blue-colored dots indicate the amino acids. Seven transmembrane regions (TM1 to TM7) are illustrated with orange cylinders.



(a)

-Primary amine -Tertiary amine

(b)

-Primary amine -Tertiary amine

Figure 1.2. Chemical structures of (a) biogenic amines and (b) trace amines (Branchek and Blackburn 2003; Maguire et al. 2009; Ferrero et al. 2012).

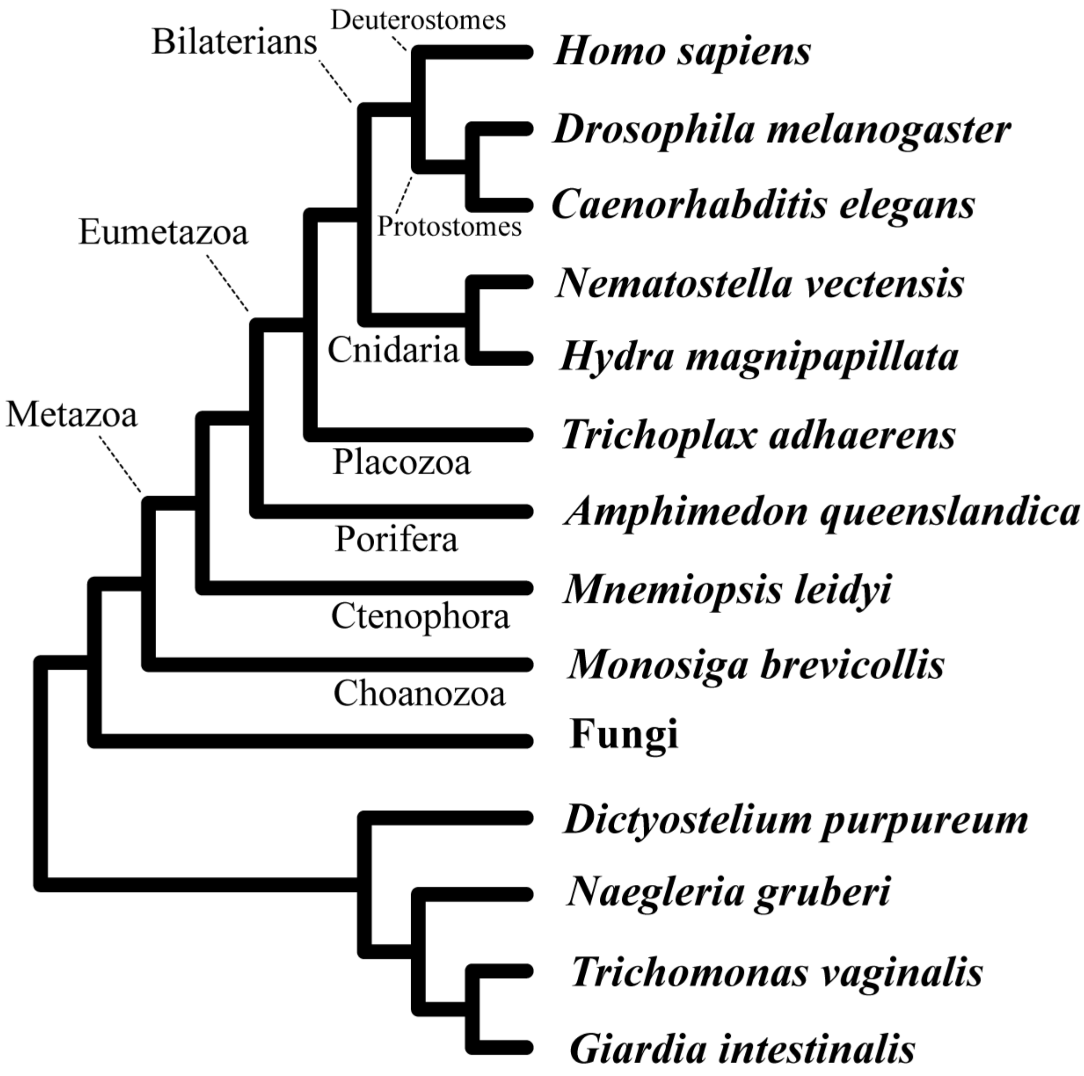


Figure 1.3. Taxonomical relationship among major metazoan animals. The phylogenetic relationship is based on Dunn et al. (2008), Srivastava et al. (2008), Srivastava et al. (2010), and Parfrey et al. (2010).

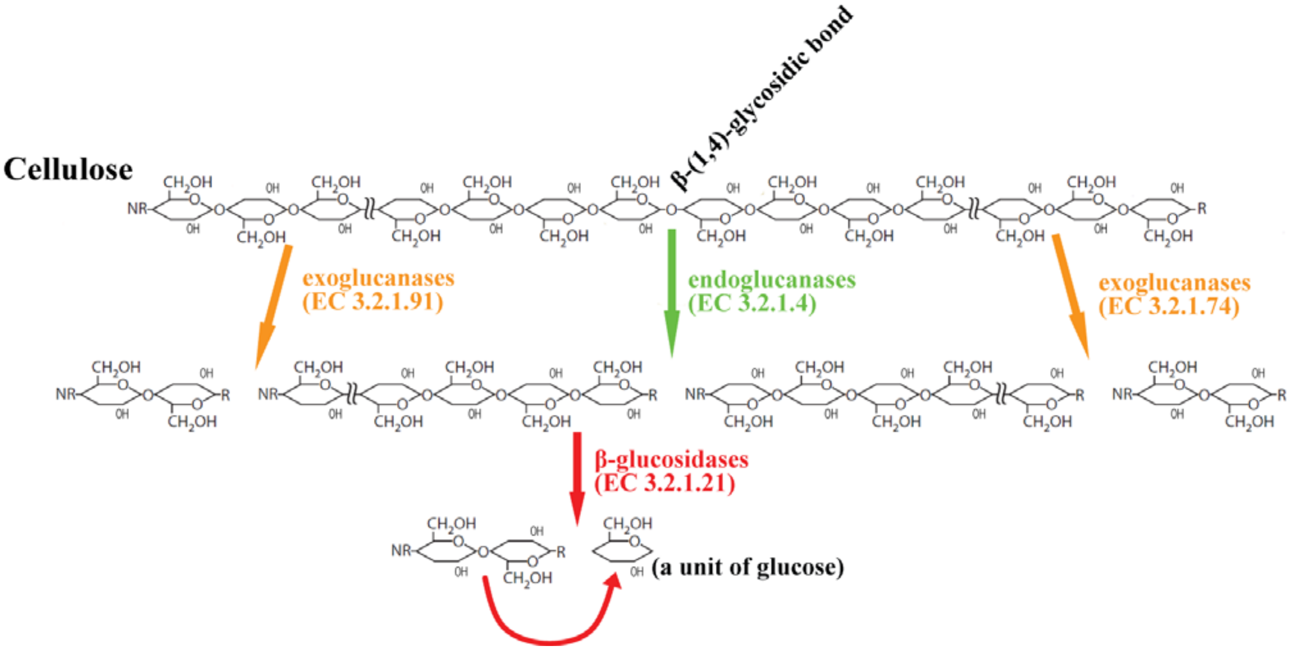


Figure 1.4. Chemical structures of cellulose and the degradation processes (Watanabe and Tokuda 2010). NR and R indicates the non-reducing end and the reducing end, respectively.

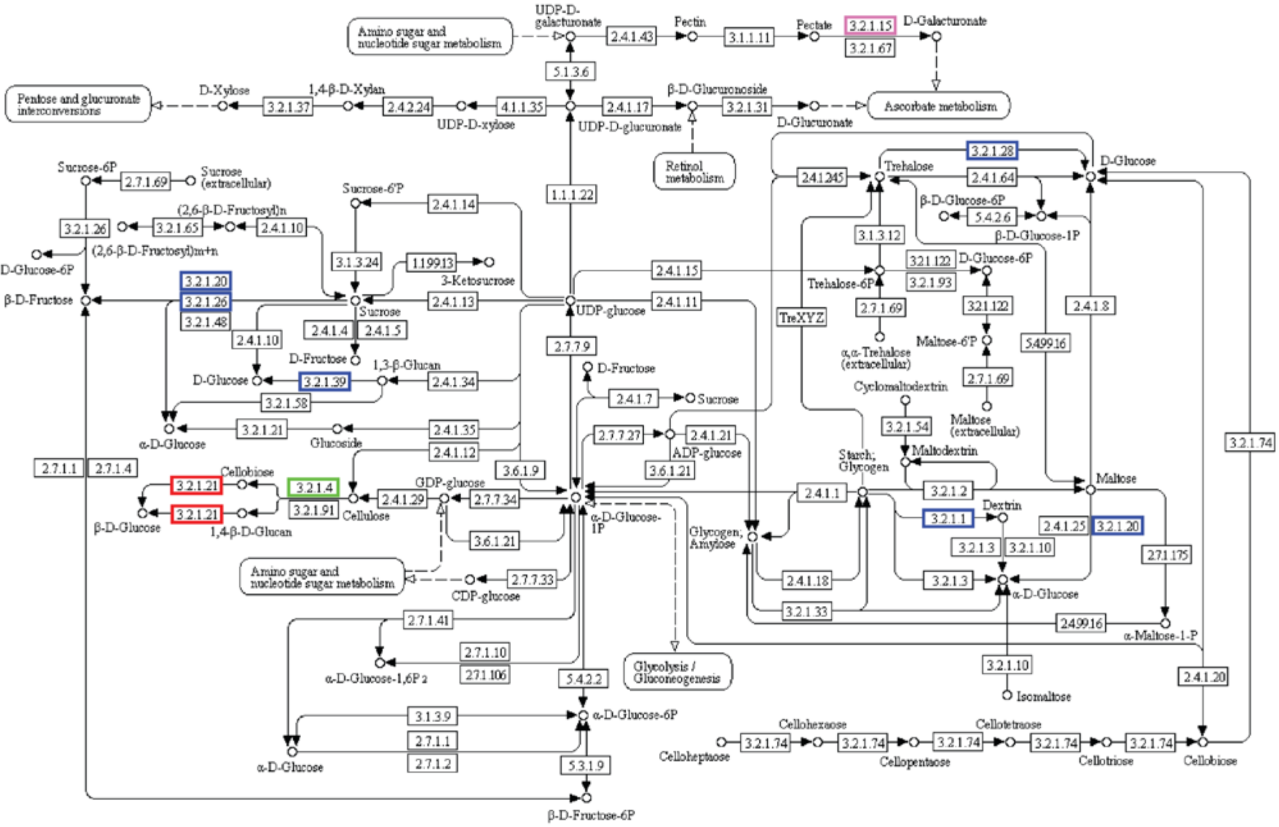
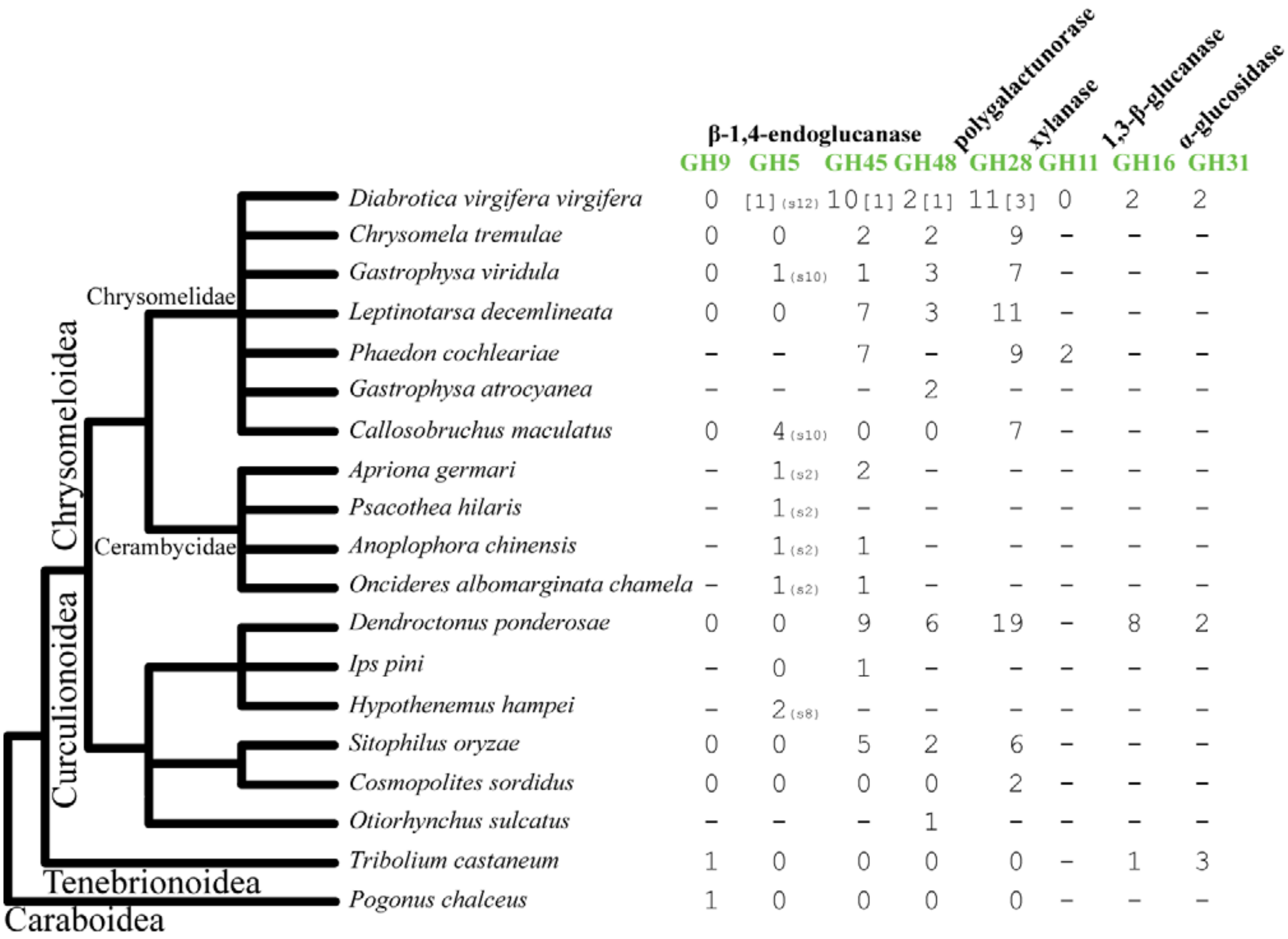


Figure 1.5. The starch and sucrose metabolism pathway generated by the KEGG Automatic Annotation Server (KAAS; http://www.genome.jp/kegg/kaas). The boxes with EC numbers are corresponding to the genes in the pathway. The enzymes found in insects are shown in green for EC 3.2.1.4 (endoglucanases; GH5, GH9, GH45, GH48), red for EC 3.2.1.21 (β-glucosidases; mainly GH1 and GH3), and pink for EC 3.2.1.15 (polygalactunorase; GH28). Other GH families found in insects are also shown in blue boxes and they include: EC 3.2.1.1 (α-amylase, GH13), EC 3.2.1.39 (endo-1,3-β-glucanase, GH16), EC 3.2.1.20 (α-glucosidase, GH13 and GH31), EC 3.2.1.26 (invertase, GH32), and EC 3.2.1.28 (trehalase, GH37).



	β-1,4-endoglucanase				polygalactunorase	xylanase	1,3-β-glucanase	α-glucosidase
	GH9	GH5	GH45	GH48	GH28	GH11	GH16	GH31
Diabrotica virgifera virgifera	0	[1] (s12)	10 [1]	2 [1]	11 [3]	0	2	2
Chrysomela tremulae	0	0	2	2	9	–	–	–
Gastrophysa viridula	0	1 (s10)	1	3	7	–	–	–
Leptinotarsa decemlineata	0	0	7	3	11	–	–	–
Phaedon cochleariae	–	–	7	–	9	2	–	–
Gastrophysa atrocyanea	–	–	–	2	–	–	–	–
Callosobruchus maculatus	0	4 (s10)	0	0	7	–	–	–
Apriona germari	–	1 (s2)	2	–	–	–	–	–
Psacothea hilaris	–	1 (s2)	–	–	–	–	–	–
Anoplophora chinensis	–	1 (s2)	1	–	–	–	–	–
Oncideres albomarginata chamela	–	1 (s2)	1	–	–	–	–	–
Dendroctonus ponderosae	0	0	9	6	19	–	8	2
Ips pini	–	0	1	–	–	–	–	–
Hypothenemus hampei	–	2 (s8)	–	–	–	–	–	–
Sitophilus oryzae	0	0	5	2	6	–	–	–
Cosmopolites sordidus	0	0	0	0	2	–	–	–
Otiorhynchus sulcatus	–	–	–	1	–	–	–	–
Tribolium castaneum	1	0	0	0	0	–	1	3
Pogonus chalceus	1	0	0	0	0	–	–	–

Figure 1.6. Taxonomical relationship of beetle species and the number of glycoside hydrolase genes. All numbers are taken from Pauchet et al. (2010) except for *Diabrotica virgifera virgifera* (this study, Chapter 4), *Cosmopolites sordidus* (this study, Chapter 4), *Tribolium castaneum* (Tribolium Genome Sequencing Consortium 2008; Willis et al. 2011), *Dendroctonus ponderosae* (Keeling et al. 2013), *Phaedon cochleariae* (Kirsch et al. 2012), *Pogonus chalceus* (Van Belleghem et al. 2012), *Gastrophysa atrocyanea* GH28 (Fujita et al. 2006), *Otiorhynchus sulcatus* GH48 (CAH25542), and *Phaedon cochleariae* GH11 (Pauchet and Heckel 2013). GH5 can be classified into 51 subfamilies (Aspeborg et al. 2012) and three subfamilies are found in beetle species; s2 (subfamily 2), s8 (subfamily 8), and s10 (subfamily 10). The taxonomical relationship is obtained from Hunt et al. (2007). NA: not available.

Chapter 2

Molecular Evolution and Functional Divergence of Trace Amine–Associated Receptors

2.0 Abstract for Chapter 2

Trace amine-associated receptors (TAARs) are a member of the G-protein-coupled receptors superfamily and are known to be expressed in olfactory sensory neurons. Only a limited number of molecular evolutionary studies have been done for TAARs. To elucidate how lineage-specific evolution contributed to their functional divergence, 30 metazoan genomes were examined. In total, 493 TAAR gene candidates (including 84 pseudogenes) were identified from 26 vertebrate genomes. TAARs were not identified from non-chordate genomes. An ancestral-type TAAR appeared to have emerged in lamprey. Four therian-specific TAAR subfamilies (one eutherian-specific and three metatherian-specific) were found in addition to previously known nine subfamilies. Many species-specific TAAR gene duplications and losses contributed to a large variation of TAAR gene numbers among mammals. TAARs were classified into two groups based on binding preferences for primary or tertiary amines. Primary amine detecting TAARs (TAAR1-4) are older, generally have single-copy orthologs (no duplication nor loss), and have evolved under strong functional constraints. In contrast, tertiary amine detecting TAARs (TAAR5-9) have emerged more recently and experienced higher rates of gene duplications. Tertiary amine detectors also showed the patterns of positive selection especially in the area surrounding the ligand-binding pocket, which could have affected ligand-binding activities and specificities. Expansions of tertiary amine detecting TAAR genes may have played important roles in terrestrial adaptations of therian mammals. Molecular evolution of the TAAR gene family appears to be governed by a complex, species-specific, interplay between environmental and evolutionary factors.

2.1 Background

While there are other types of biogenic amine receptors such as serotonin-gated cation channel in vertebrates and biogenic amine-gated chloride channels in invertebrates (Ringstad et al. 2009; Flames and Hobert 2011), trace amine-associated receptors (TAARs) and almost all biogenic amine receptors belong to the G-protein-coupled receptor (GPCR) superfamily. They mediate signal transduction in response to a wide variety of stimuli and represent the largest multi-gene family in animal genomes. For example, there are more than 900 GPCRs in human (Sällman Almén et al. 2009) and more than 1,800 in mouse (Gloriam et al. 2007). Within the GPCR superfamily, TAARs as well as biogenic amine receptors belong to the Class A: Rhodopsin-like family (Borowsky et al. 2001). In the mouse genome, for example, fifteen functional genes and one pseudogene are known for TAARs. They are classified into nine subfamilies (TAAR1 through TAAR9). In mouse, most of these subfamilies are represented by single copy genes except for TAAR7, which includes five genes and one pseudogene, and TAAR8, which includes three genes (Lindemann et al. 2005). All mouse TAARs except for TAAR1 are expressed in the main olfactory epithelium (MOE) (Liberles and Buck 2006; Fleischer et al. 2007). TAAR1 is expressed in the brain (Borowsky et al. 2001). The olfactory receptors (ORs) in mammals, another Class A family of GPCRs, also are predominantly expressed in the MOE (Kaupp 2010). The sensory neurons in the mammalian MOE thus have two types of chemosensory receptors, TAARs and ORs.

Only a limited number of molecular evolutionary studies have been done for TAARs. The complete TAAR gene set has been described in nine mammalian species (human, chimpanzee, macaque, mouse, rat, dog, cow, opossum, and platypus) (Lindemann et al. 2005; Grus et al. 2007; Hashiguchi and Nishida 2007), chicken (Mueller et al. 2008), five teleosts (fugu, spotted green pufferfish, stickleback, medaka, and zebrafish), a cartilaginous fish (elephant shark), and a jawless fish (sea lamprey) (Hashiguchi and Nishida 2007; Hussain et al. 2009). These studies showed that the tetrapod genomes have small numbers of TAAR genes (3–22 genes), while many teleost fish have higher numbers of TAAR genes compared to tetrapods, ranging from 13 to 109 genes.

The goal of this study is to understand the molecular evolutionary process of the TAAR gene family. I focused on elucidating how species-specific duplication contributed to their functional divergence among mammals. I identified complete repertoires of TAAR genes and pseudogenes from 30 metazoan genomes, especially from 17 species of mammals. The size of the TAAR family varies significantly among mammals. While the largest number of TAARs, 26 functional genes, was found in the flying fox genome, no functional TAAR genes were found in the dolphin genome. In addition to the previously known nine subfamilies, four subfamilies all therian-specific were found. Among the mammalian-specific TAAR subfamilies, TAAR7 was found to be subject to rapid species-specific gene duplications in many species. TAARs have two different evolutionary patterns. Primary amine detecting TAARs (TAAR1-4) appear to be evolving under strong negative selection, whereas tertiary amine detecting TAARs (TAAR5-9) have significant variations in gene numbers and many of them appear to evolve under the influence of positive selection, reflecting complex species-specific relationships between environmental and evolutionary factors.

2.2 Results and Discussion

2.2.1. Identification of TAAR genes.

Using previously reported TAAR protein sequences as queries, I searched TAAR candidates from 30 metazoan genomes (supplementary table S2.1). A total of 493 TAAR genes (including 84 pseudogenes) were identified from 26 vertebrate genomes (Table 2.1). The analyses failed to identify TAAR candidates in any of the four non-chordate genomes I examined (an amphioxus, two tunicates, and a sea anemone). Gnathostome (jawed vertebrate) paralogs were classified based on sequence similarities and on phylogenetic analyses. Even in distantly related species, a clear orthologous relationship can be distinguished for almost all TAAR genes and thus the nine main subfamilies (TAAR1 to TAAR9) were clearly recognized. I also identified four new mammalian-specific subfamilies (E1 and M1-M3) (described later).

I confirmed the findings of Hashiguchi and Nishida (2007) who identified a novel group of TAARs, TAAR V, found only in teleosts (zebrafish, stickleback, medaka, and spotted green pufferfish) and a frog. In the search using the TAAR V profile hidden Markov model (HMM), I confirmed that TAAR V was found only in the genomes of two teleost fishes (fugu, *Takifugu rubripes*, and spotted green pufferfish, *Tetraodon nigroviridis*) and a frog (*Xenopus tropicalis*) but not in any other tetrapod species I examined.

2.2.2. Synteny of TAAR loci among tetrapod species.

TAAR genes in human, mouse, opossum, and chicken are known to be located on a single chromosome, while fish TAARs are scattered over multiple chromosomes (Lindemann et al. 2005; Hashiguchi and Nishida 2007). I analyzed the distribution of the TAAR and other adjacent genes in nine representative tetrapods. The results are summarized in Figure 2.1. The syntenic relationships of TAARs and the adjacent genes are highly conserved as a single gene cluster. At least in amniotic genomes (mammals and chicken), the TAAR genes are all clustered in the specific region of a single chromosome. The average length of intergenic regions between two adjacent TAARs is 12,235 bps for five eutherian species (7,187 bps in frog). The transcriptional orientations are highly consistent among orthologs (Fig. 2.1). I observed many tandem duplications especially in TAAR6, TAAR7, and TAAR8, which are all eutherian specific. All tetrapod TAAR genes I examined are nested between Vanin (VNN) and Syntaxin 7 (STX7) genes. VNN1 is associated with pantetheinase activity (Pitari et al. 2000). STX7 protein forms a SNARE complex and is involved in protein-trafficking (Strömberg et al. 2009). No direct association has been reported for the functions of these adjacent genes and TAARs.

2.2.3. Origin and early evolution of TAARs.

Figure 2.2 shows the phylogeny of the representative TAAR proteins from five tetrapods (mouse, tammar wallaby, platypus, chicken, and frog), three teleosts (fugu, spotted green pufferfish, and zebrafish), a cartilaginous fish (elephant shark), and a jawless fish (sea lamprey). This phylogeny clusters TAAR subfamilies into three strongly supported monophyletic groups: TAAR V, lamprey TAAR-like, and the gnathostome TAAR1 to TAAR9 genes. The TAAR V group is located most basal after the outgroup GPCRs (Fig. 2.2 and supplementary Fig. S2.1). This family seems to have been maintained only in teleost and amphibian lineages but lost from other vertebrates. All 25 TAAR-like proteins found from the sea lamprey (*Petromyzon marinus*) genome form a well-supported monophyletic group (100% bootstrap value, supplementary Fig. S2.1). The phylogenetic analysis indicates that the sea lamprey TAAR-like genes and the gnathostome TAAR1-9 shared the direct common ancestor. Note that, as described in Materials and Methods, the TAAR signature motif (supplementary Fig. S2.2) was only weakly conserved in TAAR V and in the sea lamprey TAAR-like genes, but was present in the majority of the gnathostome members of the TAAR subfamilies. These results suggest that TAAR-like genes of sea lamprey and the TAAR1-9 genes of gnathostomes were derived from the expansion of a single-copy gene present in the common ancestor of jawless fish and jawed vertebrates, and that the well-conserved TAAR motif appeared after jawed vertebrates diverged from jawless fish, about 652 million years ago (MYA) (Blair and Hedges 2005).

Cartilaginous fish represent one of the earliest branches of the gnathostome tree (see the inset of Fig. 2.2). The elephant shark, the representative of this group in this study, possesses two distinct TAAR genes in its genome: TAAR S1a and TAAR S2a. These two elephant shark TAARs maintain the TAAR signature motif (supplementary Fig. S2.3).

Ortholog relationship between the shark TAAR S1a and the tetrapod TAAR1 subfamily was confirmed by their sequence similarities (70% to the mouse TAAR1), reciprocal blastp results, and phylogenetic analysis (Fig. 2.2 and supplementary Fig. S2.1). The shark TAAR S2a was most similar to TAAR4 proteins (63% to the mouse TAAR4). However, this probably reflects the retention of ancestral characteristics by TAAR4 rather than orthology. Phylogenetic placement of TAAR S2a indicates that this shark TAAR gene diverged from the lineage leading to TAAR2-4 (Fig. 2.2) or even all other gnathostome TAARs other than TAAR1 (supplementary Fig. S2.1).

The genomes of teleost fish have generally higher numbers of the TAARs than the tetrapod genomes and the numbers vary significantly among teleost genomes (Hussain et al. 2009). The phylogenetic analysis showed that teleost TAARs are placed in three separate phylogenetic groups (Fig. 2.2 and supplementary Fig S2.1). While one group shows a clear ortholog relationship with the tetrapod TAAR1 subfamily, other two groups have unclear phylogenetic affinities. Hashiguchi and Nishida (2007) also mentioned that the phylogenetic placement of these teleost fish clusters is not fully resolved. With multiple species-specific duplications and frequent loss of the TAAR-signature motif (supplementary Fig. S2.1), however, the evolution of the TAAR genes in teleost fish lineages appears to be unique and largely independent from the evolution of tetrapod TAARs.

2.2.4. Evolution of TAAR subfamilies in tetrapods.

To gain further insights into the evolution of the TAAR subfamilies, I restricted the attention to tetrapods with a focus on mammals, using the TAAR V genes from teleosts and frog as well as the TAAR-like sequences from the sea lamprey as the outgroups (Fig. 2.3). All phylogenetic analyses (Figs. 2.2, 2.3, and supplementary Fig. S2.1) support the TAAR1 subfamily representing the oldest divergence among the gnathostome TAAR lineages. This is consistent with its location at the beginning of the syntenic cluster (Fig. 2.1) and its distribution across all vertebrates including fishes (Table 2.2). They have apparently remained as a single-copy gene in the majority of species analyzed. The remaining TAAR genes in the gnathostome subfamilies are grouped into two separate clades: one that includes the TAAR2-4 genes and the other that includes the TAAR5-9 genes as well as four newly defined mammalian-specific TAAR subfamilies (Fig. 2.3). While there is no significant support for the phylogenetic placement of the shark TAAR S2a, as mentioned before, its position on the phylogenies would suggest that its ortholog gave rise to TAAR2-4 subfamilies (Figs. 2.2 and 2.3) or probably all other TAARs (TAAR2-9, see supplementary Fig. S2.1). TAAR4 is the oldest subfamily among the TAAR2-4 cluster, or likely to be the second oldest among the TAARs because TAAR4 sequences are found among mammals and frog (see Table 2.1 and Fig. 2.1). It must have appeared prior to the split between amphibians and amniotes and had been subsequently lost in the common ancestor of reptiles and birds. The phyletic distribution and phylogenetic arrangement of the TAAR2 and TAAR5 genes would indicate that the origin of these subfamilies predates the origin of amniotes. Since TAAR2 and TAAR3 cluster together with a high bootstrap support (100%) and because of the presence of chicken and lizard TAAR2 genes, their origin must also predate the origin of amniotes. All other TAAR subfamilies in the phylogeny form a monophyletic group and are restricted to mammals, suggesting that they derived from a single-copy TAAR gene. In mammals, descendants from this gene duplicated multiple times to give rise to the TAAR6 to TAAR9 subfamilies as well as to four therian-specific subfamilies described in the next section (M1-M3 and E1).

In summary, I classify TAAR subfamilies into four separate groups based on the timing of their inferred emergence (see Fig. 2.3 inset). TAAR1, the only TAAR that does not function as an olfactory receptor, is the oldest subfamily, as its origin probably predates the deepest split among gnathostomes. All TAARs except for TAAR1 are selectively expressed in olfactory epithetlium. Thus the expression pattern changed after TAAR4 and newer TAARs diverged from TAAR1. TAAR4 is at least as old as tetrapods. Among other younger subfamilies, the origins of TAAR2 and TAAR5 are traced back to the common ancestor of amniotes, whereas all others are apparently derived from mammalian-specific duplications. Many of these timing estimates will have to be re-evaluated once detailed analyses of amphibian and sauropsid TAAR repertoires become possible.

In general, non-therian amniotes such as birds (*Gallus gallus* and *Taeniopygia guttata*), anole lizard (*Anolis carolinensis*), and platypus (*Ornithorhynchus anatinus*) have smaller numbers of TAAR genes than therian mammals (Table 2.1). Although based on the timing of their origins, these lineages would be expected to include members of five TAAR subfamilies, TAAR1-5, these genomes have retained only up to four subfamilies. Note also that the frog (*Xenopus tropicalis*) genome has only copies of the two oldest types of TAARs (TAAR1 and TAAR4). The older types of TAAR subfamilies (TAAR1-5) exist as single-copy genes in each genome except for the expansion of TAAR4 in three genomes (frog, opossum, and elephant). In amniotes, in most instances for these older types of TAAR gene subfamilies, only one of the duplicated copies has remained functional, as in the case with tenrec TAAR1a/1bP and TAAR2a/2bP/2cP, hedgehog TAAR1a/1bP, and common shrew

TAAR4a/4bP ('P' indicating a pseudogene). The two exceptions to this pattern are the chicken TAAR2a/2b and horse TAAR5a/5b where both duplicated genes have intact structures.

2.2.5. Therian TAAR subfamilies.

The more recently diverged TAAR subfamilies (TAAR6-9, M1-M3, and E1) are apparently restricted to therian mammals (eutherians and metatherians; Table 2.1 and Figs. 2.1, 2.3, and supplementary Fig. S2.1) and must have emerged after the divergence between Prototheria and Theria (230 - 166 MYA) (Murphy et al. 2004; Bininda-Emonds et al. 2007) (the cluster is supported by 99% bootstrap value in the maximum likelihood phylogeny). TAAR6-8 are all eutherian specific. In addition, I found eutherian- and three metatherian-unique TAAR subfamilies (TAAR E1 and TAAR M1-M3, respectively) in this cluster. Three metatherian (tammar wallaby and opossum) TAAR groups are highly supported (>99% by at least one method; Fig. 2.3). While TAAR M1 is a single-copy gene, TAAR M2 and TAAR M3 show species-specific expansions. Although the TAAR E1 subfamily is not highly supported (less than 70% bootstrap values in the maximum-likelihood and neighbor-joining phylogenies but 0.85 posterior probability in the Bayesian phylogeny), it forms a distinct cluster consistently in the three different phylogenetic reconstructions. TAAR E1 is found only in a few species of mammals: in two species of Laurasiatheria (common shrew and hedgehog) and in two species of Afrotheria (tenrec and african elephant) (see Table 2.1 for details). Therefore, TAAR E1 must have been present in early eutherians but have been lost in the ancestral lineage of Euarchontoglires (human, mouse, and rat) as well as in many Laurasiatheria species.

2.2.6. Gain and loss of TAAR genes among mammals.

The number of TAAR genes varies widely among the mammals I examined, ranging from 0 in dolphin to 26 in flying fox (Table 2.1). Frequent gene gains have occurred particularly in therian-specific TAAR genes (species-specific duplications are shown with blue branches in Fig. 2.3).

As shown in Table 2.1 and Figure 2.1, the human genome does not have functional copies of TAAR3, TAAR4, and TAAR7. Stäubert et al. (2010) showed that pseudogenization of TAAR3 and TAAR4 happened before the divergence of human and orangutan (for TAAR3) or gorilla (for TAAR4). Interestingly, they also showed that independent pseudogenizations have also occurred in the marmoset/tamarin lineages for both TAAR3 and TAAR4. My preliminary search showed that in parallel to human, common marmoset (*Callithrix jacchus*) also lost TAAR7 (no pseudogene is found). In fact, the marmoset genome has only two functional TAAR genes: TAAR1 and TAAR5. All other five TAAR sequences I found were pseudogenes. Marmoset appears to have the fewest number of functional TAARs following dolphin and dog (Table 2.1). Fewer gene numbers in primates have been reported also for the OR gene family (supplementary Table S2.1) (Go and Niimura 2008b; Dong et al. 2009), which has been associated with poor olfaction senses in primate species (Hayden et al. 2010).

The most extreme reduction in TAAR repertoire is seen in the bottlenosed dolphin (*Tursiops truncatus*) genome, which apparently has no functional TAAR gene, and only possesses three pseudogenes (TAAR1P, TAAR9aP, and TAAR9bP). As an interesting concordance, the dolphin appears to have also lost most but 26 of the functional OR genes (supplementary Table S2.1) (also Hayden et al. 2010). My preliminary study shows that dolphin genome carries only three and four intact vomeronasal type-1 and type-2 receptor genes, respectively, and no functional gene but three pseudogenes of the Taste 1 (sweet taste) receptor. In general, dolphin appears to be a group of mammals that have the smallest number of chemoreceptors, apparently associated with their secondary adaptation for the aquatic environment and with the TAAR genes following the trend.

The dog genome has only two functional TAARs (TAAR4 and TAAR5) and two pseudogenes (TAAR1P and TAAR2P). On the contrary, a large number of OR genes (822 functional genes) with a small proportion of pseudogenes (25.3%) are found in the dog genome compared to other tetrapod species (supplementary Table S2.1) (also Niimura and Nei 2007). The TAAR1 pseudogenization seems to be a recent event. It must have happened after the divergence from feliforms because TAAR1s are all pseudogenes in wild gray wolf and four other caniforms but it is intact in cats (Vallender et al. 2010). The reliance on the higher number of ORs in the dog may have led to the reduction of TAARs due to their possibly overlapping functions.

The flying fox (*Pteropus vampyrus*) genome carries the largest number of TAARs (26 genes and 10 pseudogenes) while another Chiroptera, little brown bat (*Myotis lucifugus*), has a smaller number of TAARs (6 genes and 1 pseudogene). The larger number of TAARs in flying fox is caused, on one hand, by the flying fox-specific duplications of TAAR6 and TAAR7, and on the other hand, by the loss of TAAR6-8 in little brown bat. It is possible that the functions of TAAR6 and TAAR7 subfamilies may be related to dietary difference between fruit-eating flying fox and insectivorous little brown bat. TAAR7 especially is most prone to duplicate among TAAR subfamilies (Table 2.1 and Fig. 2.3), and as described later, positive selection is detected in some TAAR7 genes. I should note, however, that no difference has been observed between these two Chiroptera species in terms of evolutionary patterns (*e.g.*, selection and gene numbers) in other chemoreceptor genes such as sweet taste receptors (Zhao et al. 2010), ORs (Hayden et al. 2010), and vomeronasal sensitivity (Zhao et al. 2011a). The sensory trade-off hypothesis has been considered for enhanced color-vision in primates and their often reduced or inactivated chemosensory genes (Gilad et al. 2004; Zhao et al. 2009; however, Matsui et al. 2010). A similar scenario may be considered for echolocating insectivorous little brown bat, which lost three TAAR genes. However, laryngeal echolocation appears to have evolved earlier than the divergence of the two Chiroptera species I examined (Teeling 2009), and as mentioned above, no such associated difference is known for other chemoreceptors in these or other Chiroptera species. It is thus difficult to apply the trade-off hypothesis in this case.

The numbers of OR and TAAR genes both vary widely among mammalian genomes (supplementary Table S2.1 for the number of OR genes). In general, their numbers appear to be correlated. The dolphin genome has only 26 OR genes and no TAARs. Primates and platypus have relatively small numbers of OR as well as TAAR genes. Rodents (mouse and rat), cow, and opossum all have large numbers of both OR and TAAR genes. Exceptions are, as mentioned before, the dog genome where many TAAR gene functions seem to have been displaced with ORs (more than 800 functional genes are found), and the two Chiroptera genomes where TAAR gene numbers vary significantly (6 vs. 26) while similar numbers of ORs are found between them. The two chemoreceptor families thus seem to have complex relationships in response to both environmental and evolutionary factors.

2.2.7. Functional differentiation among TAAR subfamilies.

TAARs are classified into two groups based on the types of ligands (amines) they detect (Ferrero et al. 2012). TAAR1-4 are stimulated by primary amines (*e.g.*, isoamylamine), which can be derived from natural amino acids by a single decarboxylation reaction. TAAR5-9, on the other hand, detect tertiary amines (*e.g.*, *N*,*N*-dimethylated amines). As phylogenetic analyses clearly showed (Figs. 2.2 and 2.3), the tertiary amine preferring TAARs (TAAR5-9) cluster together (also see Ferrero et al. 2012), and these newer type of TAARs emerged from an ancestral type, primary amine preferring TAAR.

The "differential tuning hypothesis" has been put forth to explain why tetrapods have two olfactory systems: the main olfactory system (MOS) and the vomeronasal system (VNS) (Leinders-Zufall et al. 2000; Grus and Zhang 2008). It is suggested that receptors expressed in MOS are broadly-tuned generalists that can detect an overlapping set of ligands and thus are more likely to be conserved, while receptors expressed in VNS are narrowly-tuned specialists and would evolve in a more lineage-specific manner. Grus and Zhang (2008) tested this hypothesis and showed that VNS-expressed vomeronasal receptors (V1Rs and V2Rs) in tetrapods have abundant lineage-specific gene gains and losses. They found opposite patterns in MOS-expressed ORs and TAARs.

In this study, differences in evolutionary patterns were also found among the TAAR subfamilies. Figure 2.4 compares the number of TAAR genes among TAAR subfamilies for each therian species. While very few species-specific gene duplications were observed in primary amine detecting TAAR subfamilies (TAAR1-4), multiple species-specific duplications were found in tertiary amine detecting TAARs (TAAR5-9). Other newer TAAR subfamilies (TAAR E1 and M1-M3) belong to the same cluster with TAAR5-9. They are potentially tertiary amine detectors and also have multiple duplications. It should be noted that Grus and Zhang (2008) observed such a difference between TAAR1-5 and TAAR6-9 in mouse and opossum. This study analysis clarified and expanded the two evolutionary patterns among TAAR subfamilies.

2.2.8. Different evolutionary patterns in primary and tertiary detecting TAARs.

In order to test possible differences in evolutionary patterns between primary and tertiary detecting TAARs, I estimated the average ω (the ratio of nonsynonymous to synonymous distances, d_N/d_S) for each TAAR subfamily. As shown in Figure 2.4 (see also supplementary Table S2.2), the average ω's were about two times higher in tertiary amine detectors than in primary amine detectors (ω ranging from 0.0774 to 0.1807 for TAAR1-4 and from 0.1388 to 0.3512 for TAAR5-9, E1, and M1-M3; the difference between two groups is significant with *P* = 0.005 by one-tailed *t*-test and *P* = 0.0253 by Mann-Whitney *U* test).

I selected four representative TAAR subfamilies: two primary detectors (TAAR1 and TAAR3) and two tertiary detectors (TAAR7 and TAAR8) and tested which lineage(s) show(s) significantly different ω using the PAML branch models (Yang 2007). Estimated ω's were significantly larger in TAAR7 compared to other lineages (*P* < 0.0001; Tests 1, 4, and 5 in supplementary Fig. S2.4). ω was also significantly larger in TAAR8 when compared against primary amine TAAR lineages (*P* = 0.0031; Test 2 in supplementary Fig. S2.4). Thus, the nonsynonymous substitutions in these tertiary amine detecting TAAR subfamilies were substantially accelerated after the divergence from older primary amine detecting TAARs.

I next tested with the site models for the possibility of positive selection in each TAAR subfamily. The tests showed a highly significant support of positive selection for TAAR7 (*P* < 0.0001) and a weak but significant support (*P* = 0.0327) for TAAR8 (supplementary Table S2.3). To further confirm the occurrence of positive selection in tertiary amine detectors, I tested using the branch-site models that can detect a short episode of positive selection occurring in a small fraction of amino acids (Zhang et al. 2005). Based on the results obtained above, I chose TAAR7 and TAAR8 for this test. As summarized in Table 2.4, significant results were found in two branches in TAAR7 and one branch in

TAAR8. These branches are also shown in red in Figure 2.3. It further supports that the evolution of tertiary amine detecting TAARs has been partly driven by positive selection.

2.2.9. Positive-selection sites are located in the potential ligand-binding sites in the TAAR proteins.

For TAAR7 and TAAR8, the amino acid sites under positive selection were identified using the Bayes Empirical Bayes (BEB) inference (Yang et al. 2005). Eleven sites were identified with the site models (supplementary Table S2.2) and six sites with the branch-site models (supplementary Table S2.3). Four of eleven sites identified in TAAR7 (positions $137^{4.39}$, $155^{4.57}$, 184, and $188^{5.36}$) and one of five sites identified in TAAR8 (position $194^{5.42}$) had their posterior probabilities higher than 0.95, a strong indication of positive selection. The spatial distribution of these sixteen positive-selection sites on the TAAR proteins is illustrated in Figure 2.5 (see supplementary Fig. S2.5 for more details). Thirteen sites are present in the extracellular loop regions, especially in the second extracellular loop (EC2), and in the extracellular-ends of TM regions. They are particularly concentrated in the area surrounding the predicted main ligand-binding pocket (see supplementary Fig. S2.6). The seven positively selected sites in TAAR7 and TAAR8 (positions $103^{3.32}$, $104^{3.33}$, $159^{4.61}$, 184, 186, $190^{5.38}$, and $194^{5.42}$) correspond to residues identified to be directly involved with ligand-binding on β-adrenergic receptors 1 and 2 (Kleinau et al. 2011; Warne et al. 2011; Warne et al. 2012) (see supplementary Fig. S2.6 for the details). Positions $104^{3.33}$ and $155^{4.57}$ were identified to be under positive selection in TAAR7 (supplementary Table S2.3 and Figs. 2.5(c) and 2.6).

A mutational study of the β_2-adrenergic receptor demonstrated that replacement of two amino acids (corresponding to positions $151^{4.53}$ and $155^{4.57}$ in human TAAR1) significantly affected the receptor expression and agonist-stimulated activity (Chelikani et al. 2007). An amino acid mutation corresponding to position $104^{3.33}$ in human TAAR1 also rescued the low expression of the mutant. Therefore, these positively selected positions are potentially important in functions including the folding and ligand-binding.

Ferrero et al. (2012) demonstrated that mutating two amino acids closely located to possible ligand-binding sites in TM3 ($108^{3.37}$ and $109^{3.38}$) between those found in the mouse TAAR7e (SS) and those in TAAR7f (YC) dramatically reversed the ligand responsiveness. In PAML site-model (M8) analysis of TAAR7, these two sites have relatively high ω's (1.022 and 0.902) although their posterior probabilities were lower than 0.3. It should be also noted that there were two other sites whose ω's were larger than 1.0 ($100^{3.29}$ and $196^{5.43}$) although their probabilities were low (0.35 and 0.45, respectively). The position $100^{3.29}$ is one of the ligand-binding sites (supplementary Fig. S2.6). Furthermore, although their posterior probabilities were not high (0.414 and 0.46), two other ligand-binding neighboring sites, $28^{1.37}$ and $152^{4.54}$, have also high ω's (1.231 and 1.104) with PAML site-model (M8) analysis of TAAR8.

The ligand-binding space in the Rhodopsin-like GPCR proteins is consisted of a deeper main ligand-binding crevice and a shallower minor binding pocket (Nygaard et al. 2009; Rosenkilde et al. 2010). The latter area is considered to be important for receptor activation, and the residues surrounding the minor pocket are highly conserved especially among TAARs (supplementary Fig. S2.6). Interestingly, the position $103^{3.32}$ in TM3 was found to be under positive selection in TAAR7, and it is located at the boundary between the two binding pockets. Kleinau et al. (2011) showed that six of the twenty nine residues identified as ligand-binding sites are conserved among biogenic amine receptors including human TAARs and adrenergic receptors, and considered them to be determinants of the ligand-binding regions among these receptors. All but one ($103^{3.32}$) of these positions are in fact highly conserved among the TAARs I examined. Kleinau et al. (2011) further pointed out that six additional ligand-binding residues in human TAAR1 are identical or similar to those of biogenic amine receptors. They speculated that this similarity could explain the ligand promiscuity of TAAR1. I confirmed that these residues are also conserved in all other TAAR1s while residues in the corresponding positions in tertiary amine detecting TAARs are more diverse (see supplementary Fig. S2.6 for the details).

2.2.10. Changes of amino acid properties in positive-selection sites.

Many amino acid changes found in the positively selected sites are those altering physicochemical properties (supplementary Fig. S2.7). I examined these substitutions using TreeSAAP (Woolley et al. 2003; McClellan et al. 2005). Side-chain changes involving volume, torsion angles, hydrophobicity, and charge found in positively selected positions as well as their neighboring sites were shown to be under positive destabilizing selection (*P* < 0.001). Pairwise TreeSAAP analysis also showed that many long branches found in the TAAR7 family (*e.g.*, flying fox 7h and cow 7c in fig. 3) may also be under such positive destabilizing selection. Of particular interests is three changes identified in the tenrec/elephant lineage of TAAR7 using branch-side models. All three changes (positions 161, 177, and $188^{5.36}$) involve acquisition of serine residues. Changes involving serines are also found in two other highly significantly supported positions ($155^{4.57}$ in TAAR7 and $194^{5.42}$ in TAAR8). All these changes are located within or at the border of the EC2 region. Although the positions are not consistent, for β_1AR, serine residues in TM5 (positions $194^{5.42}$, $195^{5.43}$, and $198^{5.46}$) have been reported to be critical for agonist binding and receptor activation (Strader et al. 1989; Sato et al. 1999). Structural analysis of β_1AR by Warne et al. (2011) indicated that the ligand-induced rotamer conformational changes of these serine residues and stabilization of the contracted ligand-binding pocket (through hydrogen-bonding interactions between the ligand and these residues) dictate the efficacy of ligand. Therefore, the changes found in these positive-selection sites may have played an important role in defining ligand-binding activities and specificities among tertiary amine detecting TAAR subfamilies.

2.3 Conclusion

Molecular evolutionary analysis of metazoan TAARs showed that an ancestral-type TAAR emerged in lamprey. The conserved TAAR motif appeared after jawed vertebrates diverged from jawless fish. Among mammalian TAARs, older types of TAAR subfamilies (TAAR1-4) are primary amine detecting receptors. They are more conserved and maintained as single-copy genes in each genome except for TAAR4. Newer types of mammalian TAARs (TAAR5-9, M1-M3, and E1) are tertiary amine preferring receptors. They are found only in therian mammals and have experienced frequent species-specific duplications. My evolutionary analysis found evidence of positive selection distributed around the ligand-binding sites in TAAR7 and TAAR8 proteins. These changes could have affected ligand-binding activities and specificities in these TAARs. It may have contributed to therian mammal's adaptation to the dynamic land environments by allowing finer discrimination among a diverse array of volatile amines. Specific ecological conditions in some species may have led to additional duplications or losses of especially tertiary amine detecting TAARs. Furthermore, birth and death processes of two chemoreceptor families (ORs and TAARs) seem to be under the influence of both environmental and evolutionary factors. Further studies on TAAR evolution and their functions will provide more insights into functional divergence of chemosensory receptors.

2.4 Materials and methods

2.4.1. Query and genome sequences.

Previously reported TAAR genes were used as search queries. The sequences were obtained from Lindemann et al. (2005) and from Hashiguchi and Nishida (2007). Genomic sequences were obtained from multiple sources (supplementary Table S2.1). It includes 17 mammals (14 eutherians, 2 metatherians, and 1 prototherian), two birds, one reptile, one frog, two teleost fishes, elephant shark, as well as four non-chordate species. Note that the zebrafish and sea lamprey TAARs obtained from Hashiguchi and Nishida (2007) are also included in this analysis.

2.4.2. TAAR gene mining.

Similarity search was performed using the Basic Local Alignment Search Tool (BLAST, ver. 2.2.17) programs (Altschul et al. 1990). The default parameters were used for `tblastn` except for setting the effective length of database (option `-z`) to 1.1×10^{10}. This was done to obtain E-values comparable among different sizes of genomes and equivalent to those from the search against the non-redundant (NR) protein database at the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov). The E-value threshold of 1×10^{-30} was used to identify TAAR gene candidates from each genome.

The putative TAAR genes were verified by searches using `blastp` against the NR database. A putative protein was considered to be a TAAR candidate if the top hit from the blastp search was a previously known TAAR. The TAAR candidates newly identified were subsequently used as queries against their genomes again to find any additional candidates. These steps were recursively performed until no other TAAR candidate sequences were detected from each genome.

TAAR V has been found only from a limited number of species (Hashiguchi and Nishida 2007). For more sensitive search, I built a profile HMM with five TAAR V protein sequences from frog (*Xenopus tropicalis*, XP_002935532), zebrafish (*Danio rerio*, XP_001337671), spotted green pufferfish (*Tetraodon nigroviridis*, CAF93600), stickleback (*Gasterosteus aculeatus*, Hashiguchi and Nishida 2007), and medaka (*Oryzias latipes*, Hashiguchi and Nishida 2007). Each genome was searched using the `hmmbuild` and `hmmsearch` programs of the HMMER package (ver. 3.0) (Eddy 2011) with default parameters.

The TAAR genes are intron-less and encoded in a single exon. TAAR2 genes, also known as GPR58, are exceptions and have two exons. To determine exon-intron boundaries for TAAR2, a profile HMM was built from human, mouse, and rat TAAR2 protein sequences using the HMMER package (ver. 2.3.2) (Durbin et al. 1998). Using this profile HMM, the coding sequences were predicted using GeneWise (ver. 2.2) (Birney et al. 2004).

2.4.3. TAAR signature motif.

TAAR proteins have a unique peptide motif that is absent from all other known GPCRs (Lindemann et al. 2005). This motif is located within the seventh transmembrane (TM) region, and defined as $NSX_2NPX_2[Y/H]X_3YXWF$ where X_n represents any n amino acid residue(s) (supplementary Fig. S2.2(a)). The motif is most strongly conserved in the TAAR3 family (supplementary Fig. S2.2(b)). All tetrapod TAAR proteins identified in this study have this motif, while all lamprey TAAR-like and five TAAR V proteins have only weakly conserved motifs. Motifs found in the corresponding regions of the lamprey TAAR-like and TAAR V proteins are $XSX_2NPX_2[Y/F]X_6F$ and $NSX_2NPX_2YX_3[H/N]XS[Y/F]$, respectively. In many teleost fish TAAR proteins, the motif is only weakly conserved or lost completely. In supplementary Figure S2.1, the distribution of teleost fish TAARs among vertebrate TAARs as well as the conservation of the motif is illustrated.

2.4.4. Multiple sequence alignments.

Multiple alignments of TAAR protein sequences were generated using MAFFT with the L-INS-i algorithm (ver. 6.24) (Katoh and Toh 2008), MUSCLE (ver. 3.7) (Edgar 2004), ProbCons (ver. 1.12) (Do et al. 2005), and PRALINE (Heringa 1999), each with the default parameters. Alignments were adjusted manually when necessary. For consistency, all amino acid positions shown in this study are numbered based on the human TAAR1 sequence in the alignment given in supplementary Figure S2.6. Position numbers are also presented using the scheme proposed by Ballesteros and Weinstein (1995). In the Ballesteros-Weinstein system, the most conserved residue in each TM region among all rhodopsin GPCRs is assigned the position index “50” and the rest of the positions within each TM region are numbered accordingly. In this study the Ballesteros and Weinstein numbers are based on the TM regions of the turkey β_1-adrenergic receptor (β_1AR, P07700) sequence obtained from the GPCRDB Web server (http://www.gpcr.org/7tm) (Vroling et al. 2011). These numbers are given as superscripts. All TAAR sequences and alignments are available from: http://bioinfolab.unl.edu/emlab/TAAR

2.4.5. Phylogenetic analysis.

Phylogenetic relationships were reconstructed by the maximum-likelihood method with the PROTGAMMAJTT model (JTT matrix with gamma-distributed rate variation) using RAxML (ver. 7.0.4) (Stamatakis 2006). The neighbor-joining phylogenies (Saitou and Nei 1987) were reconstructed by using `neighbor` of the Phylip package (ver. 3.67) (Felsenstein 2005). The protein distances were estimated using `protdist` of the Phylip package with the JTT model with the gamma-distributed rate variation (α=1.3004 was estimated using the maximum-likelihood method implemented RAxML) (Yang 1994).

Bayesian inference of phylogeny was performed using MrBayes (v3.1.2) (Huelsenbeck and Ronquist 2001) with the JTT substitution model with the gamma-distributed rate variation (α=1.3004). The Markov chain Monte Carlo search was run for 10^6 generations, with a sampling frequency of 10^3, using three heated and one cold chain and with a burn-in of 10^2 trees. In addition to TAAR sequences, eight representative biogenic amine receptors (BARs), four cow opsin sequences, as well as eight representative dog ORs were included in phylogenetic analysis. OR sequences were used as the outgroup. Non-parametric bootstrapping with 1000 pseudo-replicates (Felsenstein 1985) was used to estimate the confidence of branching patterns for the maximum-likelihood and neighbor-joining phylogenies. Presentation of the phylogenies was done with FigTree (http://tree.bio.ed.ac.uk/software/figtree). All phylogenies are available from: http://bioinfolab.unl.edu/emlab/TAAR.

2.4.6. Transmembrane protein topology prediction.

HMMTOP (ver. 2.1) (Tusnady and Simon 2001) and Phobius (ver. 1.01) (Kall et al. 2007) were used to predict the transmembrane protein topology, which includes N-terminal, transmembrane (TM), intercellular loop (IC), extracellular loop (EC), and C-terminal regions.

2.4.7. Tests of selection patterns.

Selection patterns were tested using the maximum-likelihood framework developed by Goldman and Yang (1994). The site-, branch-, and branch-site models implemented in `codeml` of the PAML (Phylogenetic Analysis by Maximum Likelihood) package (version 4.5) were used (Yang 2007). I first used the site-model M0 (one-ratio, ω, for all sites) to estimate the d_N/d_S (ω) for each TAAR subfamily. Two sets of likelihood-ratio tests (LRTs; d.f. = 2) were performed for positive selection: M1a (two site-classes, nearly neutral model: $0 < \omega_0 < 1$ and $\omega_1 = 1$) *vs.* M2a (three site-classes including positive selection: $0 < \omega_0 < 1$, $\omega_1 = 1$, and $\omega_2 > 1$) and M7 (beta distribution and $0 < \omega < 1$) *vs.* M8 (beta distribution and $\omega > 1$).

Using the branch models, I performed LRTs with d.f. = 1 between a one-ratio model (R1; the same ω for all branches) and a two-ratio model (R2; two independent ω's) (Yang 1998; Yang and Nielsen 2002). As illustrated in supplementary Figure 2.4, each test was set up to compare primary amine detecting TAAR lineages (TAAR1 and TAAR3) against tertiary amine detecting receptor lineages (TAAR7 and TAAR8).

I also used the branch-site models in order to detect positively selected sites along specific branches (Yang and Nielsen 2002; Zhang et al. 2005). In these models, positive selection was allowed on a specific, "foreground", branch, and the LRTs (d.f. = 1) were performed against null models that assume no positive selection. The branch-site test of positive selection ("Test 2" in Zhang et al. 2005) has four site classes: 0, 1, 2a, and 2b. For the site classes 0 and 1, all codons are under purifying selection ($0 < \omega_0 < 1$) and under neutral evolution ($\omega_1 = 1$), respectively, on all branches. For the site classes 2a and 2b, positive selection is allowed on the foreground branches ($\omega_2 \geq 1$) but the other, "background", branches are under purifying selection ($0 < \omega_0 < 1$) and under neutral evolution ($\omega_1 = 1$), respectively. For the null model, ω_2 is fixed as 1. For this analysis, TAAR7 and TAAR8 subfamilies were tested. For each subfamily phylogeny, tests were done using each branch (from both internal and terminal branches) as the foreground. The numbers of tests performed were 61 and 26 for TAAR7 and TAAR8, respectively.

All PAML analyses were carried out using the F3X4 model of codon frequency (Goldman and Yang 1994). The level of significance (*P*) for the LRTs was estimated using a χ^2 distribution with given degrees of freedom (d.f.) and the test statistic calculated as twice the difference of log-likelihood between the models ($2\Delta ln\text{L}= 2[ln\text{L}_1 – ln\text{L}_0]$ where L_1 and L_0 are the likelihoods of the alternative and null models, respectively). Positively selected amino acid sites are identified based on Bayes Empirical Bayes posterior probabilities (Yang et al. 2005).

2.4.8. Analysis of selection on amino acid properties.

Possible selection on changes in amino acid properties were examined by TreeSAAP (version 3.2) (Woolley et al. 2003; McClellan et al. 2005). The program reconstructs the ancestral character states at each node based on a given phylogeny. Observed amino acid substitutions are analyzed in the context of 539 physicochemical properties (downloaded from http://dna.cs.byu.edu/treesaap) (Kawashima et al. 2008) and their magnitude of change (in 8 categories, with 1 being the most conservative and 8 the most radical). Based on the methods by Xia and Li (1998), McClellan and McCracken (2001), and McClellan et al. (2005), observed differences are compared against the expected differences under the neutrality. The most radical changes (categories 6 - 8) with significant positive z-scores (> 3.09; $P < 0.001$) are considered to be under positive-destabilizing selection. In order to confirm if the results are not affected by the phylogenetic topologies I used, I also performed pairwise analysis of TreeSAAP. Pairwise comparisons were done for 16 flying fox TAAR7, 29 other mammalian TAAR7, and 16 TAAR8 sequences. TreeSAAP results are available from: http://bioinfolab.unl.edu/emlab/TAAR.

2.4.9. Protein structural homology modeling.

Homology modeling of TAAR protein structures was performed using the SWISS-MODEL Web server (http://swissmodel.expasy.org) (Arnold et al. 2006). The same template, the B-chain of the turkey (*Meleagris gallopavo*) β_1AR (4AMJ), was selected for the human TAAR1, elephant TAAR7a, and mouse TAAR8a proteins. See supplementary Figure S2.5 for the details on TAAR protein structural modeling. The graphical representation of TAAR structures was prepared with PyMOL (version 1.3) (DeLanoScientific, San Carlos, CA).

2.5 Acknowledgements

I thank Wendy Grus (Knome, Inc) and Peng Shi (Kunming Institute of Zoology, Chinese Academy of Science) for providing us TAAR sequences. I also thank David McClellan (Bigelow Laboratory for Ocean Sciences) for his helpful suggesions for TreeSAAP analysis. This work was supported by in part the National Library of Medicine at the National Institute of Health (grant number R01LM009219 to E.N.M.) and the National Science Foundation (EPS-0903787 to F.G.H).

2.6 Literature Cited

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol*. 215:403-410.

Arnold K, Bordoli L, Kopp J, Schwede T. 2006. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics*. 22:195-201.

Ballesteros JA, Weinstein H. 1995. Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. *Methods Neurosci*. 25:366-428.

Bininda-Emonds ORP, Cardillo M, Jones KE, MacPhee RDE, Beck RMD, Grenyer R, Price SA, Vos RA, Gittleman JL, Purvis A. 2007. The delayed rise of present-day mammals. *Nature*. 446:507-512.

Birney E, Clamp M, Durbin R. 2004. GeneWise and Genomewise. *Genome Res*. 14:988-995.

Blair JE, Hedges SB. 2005. Molecular Phylogeny and Divergence Times of Deuterostome Animals. *Mol Biol Evol*. 22:2275-2284.

Borowsky B, Adham N, Jones KA, Raddatz R, Artymyshyn R, Ogozalek KL, Durkin MM, Lakhlani PP, Bonini JA, Pathirana S, Boyle N, Pu X, Kouranova E, Lichtblau H, Ochoa FY, Branchek TA, Gerald C. 2001. Trace amines: Identification of a family of mammalian G protein-coupled receptors. *Proc Natl Acad Sci USA*. 98:8966-8971.

Chelikani P, Hornak V, Eilers M, Reeves PJ, Smith SO, RajBhandary UL, Khorana HG. 2007. Role of group-conserved residues in the helical core of beta2-adrenergic receptor. *Proc Natl Acad Sci USA*. 104:7027-7032.

Churcher AM, Taylor JS. 2011. The Antiquity of Chordate Odorant Receptors Is Revealed by the Discovery of Orthologs in the Cnidarian *Nematostella vectensis*. *Genome Biol Evol*. 3:36-43.

Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res*. 14:1188-1190.

Do CB, Mahabhashyam MSP, Brudno M, Batzoglou S. 2005. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res*. 15:330-340.

Dong D, He G, Zhang S, Zhang Z. 2009. Evolution of Olfactory Receptor Genes in Primates Dominated by Birth-and-Death Process. *Genome Biol Evol*. 1:258-264.

Durbin R, Eddy SR, Krogh A, Mitchison GJ. 1998. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids: Cambridge University Press, Cambridge UK.

Eddy SR. 2011. Accelerated Profile HMM Searches. *PLoS Comput Biol*. 7:e1002195.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 32:1792-1797.

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

Felsenstein J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6.: Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Ferrero DM, Wacker D, Roque MA, Baldwin MW, Stevens RC, Liberles SD. 2012. Agonists for 13 Trace Amine-Associated Receptors Provide Insight into the Molecular Basis of Odor Selectivity. *ACS Chem Biol*. 7:1184-1189.

Flames N, Hobert O. 2011. Transcriptional Control of the Terminal Fate of Monoaminergic Neurons. *Annu Rev Neurosci*. 34:153-184.

Fleischer J, Schwarzenbacher K, Breer H. 2007. Expression of Trace Amine–Associated Receptors in the Grueneberg Ganglion. *Chem Senses*. 32:623-631.

Gilad Y, Wiebe V, Przeworski M, Lancet D, Pääbo S. 2004. Loss of Olfactory Receptor Genes Coincides with the Acquisition of Full Trichromatic Vision in Primates. *PLoS Biol*. 2:e5.

Gloriam D, Fredriksson R, Schiöth HB. 2007. The G protein-coupled receptor subset of the rat genome. *BMC Genomics*. 8:338.

Go Y, Niimura Y. 2008. Similar Numbers but Different Repertoires of Olfactory Receptor Genes in Humans and Chimpanzees. *Mol Biol Evol*. 25:1897-1907.

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

Grus WE, Shi P, Zhang J. 2007. Largest Vertebrate Vomeronasal Type 1 Receptor Gene Repertoire in the Semiaquatic Platypus. *Mol Biol Evol*. 24:2153-2157.

Grus WE, Zhang J. 2008. Distinct Evolutionary Patterns between Chemoreceptors of 2 Vertebrate Olfactory Systems and the Differential Tuning Hypothesis. *Mol Biol Evol*. 25:1593-1601.

Hashiguchi Y, Nishida M. 2007. Evolution of Trace Amine-Associated Receptor (TAAR) Gene Family in Vertebrates: Lineage-specific Expansions and Degradations of a Second Class of Vertebrate Chemosensory Receptors Expressed in the Olfactory Epithelium. *Mol Biol Evol*. 24:2099–2107.

Hayden S, Bekaert M, Crider TA, Mariani S, Murphy WJ, Teeling EC. 2010. Ecological adaptation determines functional mammalian olfactory subgenomes. *Genome Res*. 20:1-9.

Heringa J. 1999. Two strategies for sequence comparison: profile-preprocessed and secondary structure-induced multiple alignment. *Comput Chem*. 23:341-364.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*. 17:754-755.

Hussain A, Saraiva LR, Korsching SI. 2009. Positive Darwinian selection and the birth of an olfactory receptor clade in teleosts. *Proc Natl Acad Sci USA*. 106:4313-4318.

Kall L, Krogh A, Sonnhammer ELL. 2007. Advantages of combined transmembrane topology and signal peptide prediction-the Phobius web server. *Nucleic Acids Res*. 35:W429-432.

Katoh K, Toh H. 2008. Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform*. 9:286-298.

Kaupp UB. 2010. Olfactory signalling in vertebrates and insects: differences and commonalities. *Nat Rev Neurosci*. 11:188-200.

Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. 2008. AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res*. 36:D202-D205.

Kleinau G, Pratzka J, Nurnberg D, Gruters A, Fuhrer-Sakel D, Krude H, Kohrle J, Schoneberg T, Biebermann H. 2011. Differential modulation of Beta-adrenergic receptor signaling by trace amine-associated receptor 1 agonists. *PLoS ONE*. 6:e27073.

Leinders-Zufall T, Lane AP, Puche AC, Ma W, Novotny MV, Shipley MT, Zufall F. 2000. Ultrasensitive pheromone detection by mammalian vomeronasal neurons. *Nature*. 405:792-796.

Liberles SD, Buck LB. 2006. A second class of chemosensory receptors in the olfactory epithelium. *Nature*. 442:645-650.

Lindemann L, Ebeling M, Kratochwil NA, Bunzow JR, Grandy DK, Hoener MC. 2005. Trace amine-associated receptors form structurally and functionally distinct subfamilies of novel G protein-coupled receptors. *Genomics*. 85:372-385.

Matsui A, Go Y, Niimura Y. 2010. Degeneration of Olfactory Receptor Gene Repertories in Primates: No Direct Link to Full Trichromatic Vision. *Mol Biol Evol*. 27:1192-1200.

McClellan DA, McCracken KG. 2001. Estimating the Influence of Selection on the Variable Amino Acid Sites of the Cytochrome b Protein Functional Domains. *Mol Biol Evol*. 18:917-925.

McClellan DA, Palfreyman EJ, Smith MJ, Moss JL, Christensen RG, Sailsbery JK. 2005. Physicochemical Evolution and Molecular Adaptation of the Cetacean and Artiodactyl Cytochrome b Proteins. *Mol Biol Evol*. 22:437-455.

Mueller JC, Steiger S, Fidler AE, Kempenaers B. 2008. Biogenic Trace Amine-Associated Receptors (TAARs) are encoded in avian genomes: evidence and possible implications. *J Hered*. 99:174-176.

Murphy WJ, Pevzner PA, O'Brien SJ. 2004. Mammalian phylogenomics comes of age. *Trends Genet*. 20:631-639.

Nei M, Niimura Y, Nozawa M. 2008. The evolution of animal chemosensory receptor gene repertoires: roles of chance and necessity. *Nat Rev Genet*. 9:951-963.

Niimura Y. 2009. On the Origin and Evolution of Vertebrate Olfactory Receptor Genes: Comparative Genome Analysis Among 23 Chordate Species. *Genome Biol Evol*. 2009:34-44.

Niimura Y, Nei M. 2007. Extensive Gains and Losses of Olfactory Receptor Genes in Mammalian Evolution. *PLoS ONE*. 2:e708.

Nygaard R, Frimurer TM, Holst B, Rosenkilde MM, Schwartz TW. 2009. Ligand binding and micro-switches in 7TM receptor structures. *Trends Pharmacol Sci*. 30:249-259.

Pitari G, Malergue F, Martin F, Philippe JM, Massucci MT, Chabret C, Maras B, Duprè S, Naquet P, Galland F. 2000. Pantetheinase activity of membrane-bound Vanin-1: lack of free cysteamine in tissues of Vanin-1 deficient mice. *FEBS Letters*. 483:149-154.

Ringstad N, Abe N, Horvitz HR. 2009. Ligand-Gated Chloride Channels Are Receptors for Biogenic Amines in *C. elegans*. *Science*. 325:96-100.

Rosenkilde MM, Benned-Jensen T, Frimurer TM, Schwartz TW. 2010. The minor binding pocket: a major player in 7TM receptor activation. *Trends Pharmacol Sci*. 31:567-574.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*. 4:406-425.

Sällman Almén M, Nordström KJV, Fredriksson R, Schiöth HB. 2009. Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin. *BMC Biol*. 7:50.

Sato T, Kobayashi H, Nagao T, Kurose H. 1999. Ser^{203} as well as Ser^{204} and Ser^{207} in fifth transmembrane domain of the human β_2-adrenoceptor contributes to agonist binding and receptor activation. *Br J Pharmacol*. 128:272-274.

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

Stäubert C, Böselt I, Bohnekamp J, Römpler H, Enard W, Schöneberg T. 2010. Structural and Functional Evolution of the Trace Amine-Associated Receptors TAAR3, TAAR4 and TAAR5 in Primates. *PLoS ONE*. 5:e11133.

Strader CD, Candelore MR, Hill WS, Sigal IS, Dixon RA. 1989. Identification of two serine residues involved in agonist activation of the β-adrenergic receptor. *J Biol Chem*. 264:13572-13578.

Strömberg S, Agnarsdóttir Mt, Magnusson K, Rexhepaj E, Bolander Å, Lundberg E, Asplund A, Ryan D, Rafferty M, Gallagher WM, Uhlen M, Bergqvist M, Ponten F. 2009. Selective Expression of Syntaxin-7 Protein in Benign Melanocytes and Malignant Melanoma. *J Proteome Res*. 8:1639-1646.

Taylor WR. 1997. Residual colours: a proposal for aminochromography. *Protein Eng*. 10:743-746.

Teeling EC. 2009. Hear, hear: the convergent evolution of echolocation in bats? *Trends Ecol Evol*. 24:351-354.

Tusnady GE, Simon I. 2001. The HMMTOP transmembrane topology prediction server. *Bioinformatics*. 17:849-850.

Vallender E, Xie Z, Westmoreland S, Miller G. 2010. Functional evolution of the trace amine associated receptors in mammals and the loss of TAAR1 in dogs. *BMC Evol Biol*. 10:51.

Vroling B, Sanders M, Baakman C, Borrmann A, Verhoeven S, Klomp J, Oliveira L, de Vlieg J, Vriend G. 2011. GPCRDB: information system for G protein-coupled receptors. *Nucleic Acids Res*. 39:D309-D319.

Warne T, Edwards PC, Leslie AG, Tate CG. 2012. Crystal Structures of a Stabilized β_1-Adrenoceptor Bound to the Biased Agonists Bucindolol and Carvedilol. *Structure*. 20:841-849.

Warne T, Moukhametzianov R, Baker JG, Nehme R, Edwards PC, Leslie AGW, Schertler GFX, Tate CG. 2011. The structural basis for agonist and partial agonist action on a β_1-adrenergic receptor. *Nature*. 469:241-244.

Woolley S, Johnson J, Smith MJ, Crandall KA, McClellan DA. 2003. TreeSAAP: Selection on Amino Acid Properties using phylogenetic trees. *Bioinformatics*. 19:671-672.

Xia X, Li W-H. 1998. What Amino Acid Properties Affect Protein Evolution? *J Mol Evol*. 47:557-564.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol*. 39:306-314.

Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol*. 15:568-573.

Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol*. 24:1586-1591.

Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol*. 19:908-917.

Yang Z, Wong WSW, Nielsen R. 2005. Bayes Empirical Bayes Inference of Amino Acid Sites Under Positive Selection. *Mol Biol Evol*. 22:1107-1118.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an Improved Branch-Site Likelihood Method for Detecting Positive Selection at the Molecular Level. *Mol Biol Evol*. 22:2472-2479.

Zhao H, Rossiter SJ, Teeling EC, Li C, Cotton JA, Zhang S. 2009. The evolution of color vision in nocturnal mammals. *Proc Natl Acad Sci USA*. 106:8980-8985.

Zhao H, Xu D, Zhang S, Zhang J. 2011. Widespread losses of vomeronasal signal transduction in bats. *Mol Biol Evol*. 28:7-12.

Zhao H, Zhou Y, Pinto CM, Charles-Dominique P, Galindo-Gonzalez J, Zhang S, Zhang J. 2010. Evolution of the sweet taste receptor gene Tas1r2 in bats. *Mol Biol Evol*. 27:2642-2650.

Table 2.1. The number of TAAR genes identified in the 30 animal genomes.

Group/Species name	Common name	Total	Number of TAAR subfamily genes[b]																	
		number[a]	T1	T2	T3	T4	T5	T6	T7	T8	T9	TE1	TM1	TM2	TM3	TV	TFI	TFII	TFIII	TL
[Euarchontoglires]																				
Homo sapiens	human	6 (3)	1	1	0 (1)	0 (1)	1	1	0 (1)	1	1	0	0	0	0	0	0	0	0	0
Mus musculus	house mouse	15 (1)	1	1	1	1	1	1	5 (1)	3	1	0	0	0	0	0	0	0	0	0
Rattus norvegicus	Norway rat	17 (2)	1	1	1	1	1	1	7 (2)	3	1	0	0	0	0	0	0	0	0	0
[Laurasiatheria]																				
Bos taurus	cow	21 (8)	1	1[c]	1	1	1	5 (2)	7 (4)	3 (2)	1	0	0	0	0	0	0	0	0	0
Tursiops truncatus	bottlenosed dolphin	0 (3)	0 (2)	0	0	0	0	0	0	0	0 (1)	0	0	0	0	0	0	0	0	0
Equus caballus	horse	11 (4)	1	1	1	1	2	1	1 (1)	2 (1)	1 (1)	0 (1)	0	0	0	0	0	0	0	0
Canis familiaris	dog	2 (2)	0 (1)	0 (1)	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Pteropus vampyrus	Malayan flying fox	26 (10)	1	1[c]	0 (1)	1	1	4 (6)	16	1 (3)	1	0	0	0	0	0	0	0	0	0
Myotis lucifugus	little brown bat	6 (1)	1	1[c]	1	1	1	0 (1)	0	0	1	0	0	0	0	0	0	0	0	0
Sorex araneus	common shrew	9 [1] (3)	1	[1] (1)	1	1 (1)	1	3	2	0	0 (1)	2	0	0	0	0	0	0	0	0
Erinaceus europaeus	hedgehog	6 [2] (4)	[1] (1)	1[c]	1	0 (2)	0 (1)	[1]	2	0	1	1	0	0	0	0	0	0	0	0
[Afrotheria]																				
Echinops telfairi	lesser hedgehog tenrec	9 [1] (7)	1 (1)	1[c] (2)	1	1	0	0 (1)	2 (1)	1	[1]	2 (2)	0	0	0	0	0	0	0	0
Loxodonta africana	African elephant	9 [3] (3)	[1]	1[c]	[1]	1 [1]	1	1	2	2 (3)	0	1	0	0	0	0	0	0	0	0
[Xenarthra]																				
Dasypus novemcinctus	nine-banded armadillo	5 (4)	1	1[c]	1	0	1	0	1 (1)	0 (2)	0 (1)	0	0	0	0	0	0	0	0	0
[Marsupialia]																				
Macropus eugenii	tammar wallaby	18 [1] (3)	(1)	[1]	1	1	1	0	0	0	1	0	1 (1)	9 (1)	4	0	0	0	0	0
Monodelphis domestica	opossum	22 (4)	1	1[c]	1	3 (1)	1	0	0	0	7 (1)	0	1	2 (1)	5 (1)	0	0	0	0	0
[Prototheria]																				
Ornithorhynchus anatinus	platypus	4 (1)	1	1[c]	1	1 (1)	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[Sauropsida]																				
Gallus gallus	chicken	4 (1)	1	2[c]	0	0	1 (1)	0	0	0	0	0	0	0	0	0	0	0	0	0
Taeniopygia guttata	zebra finch	1 (0)	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Anolis carolinensis	Carolina anole	3 (0)	1	1[c]	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
[Amphibia]																				
Xenopus tropicalis	pipid frog	7 (0)	1	0	0	5	0	0	0	0	0	0	0	0	0	1	0	0	0	0
[Teleostei]																				
Takifugu rubripes	fugu (Japanese pufferfish)	18 (1)	0	0	0	0	0	0	0	0	0	0	0	0	0	1	14 (1)	0	3	0
Tetraodon nigroviridis	spotted green pufferfish	34 (3)	0	0	0	0	0	0	0	0	0	0	0	0	0	1	12 (3)	0	21	0
Danio rerio	zebrafish	110 (10)[d]	1	0	0	0	0	0	0	0	0	0	0	0	0	1	92 (8)	11	6 (2)	0
[Chondrichthyes]																				
Callorhinchus milii	elephant shark	2 (3)	1 (1)	0	0	1 (2)[e]	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[Agnatha]																				
Petromyzon marinus	sea lamprey	25 (3)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	25 (3)
[Cephalochordata]																				
Branchiostoma floridae	amphioxus	0																		
[Urochordata]																				
Ciona intestinalis	vase tunicate	0																		
Ciona savignyi	tunicate	0																		
[Cnidaria]																				
Nematostella vectensis	sea anemone	0																		

[a]TAAR gene candidates are divided into three categories: intact, incomplete, and pseudogenes. The first number shown is that of "intact" genes, which contain full-length open reading frames with seven complete transmembrane regions. The number of "incomplete" genes due to incomplete genome sequences (*e.g.*, long ambiguous sequences such as a run of N's or contig ends) or incompletely identified exons (*e.g.*, TAAR2, see below) is given in square brackets. The number in parentheses is that of possible pseudogenes, which contain premature stop codons or frame-shifting insertions or deletions.

[b]T1-T9, TE1, TM1-TM3, TV, TFI-III, and TL indicate TAAR1-9, TAAR E1, TAAR M1-M3, TAAR V, fish-specific TAAR I-III, and lamprey TAAR-like genes, respectively. The group names of TAAR V and fish-specific TAAR I-III are given by Hashiguchi and Nishida (2007).

[c]Only the exon2 sequences (coding 304 to 331 amino acids) were identified from these TAAR2 genes. The exon1 (coding 8 to 20 amino acids) can be located more than 6000 bp upstream.

[d]The sequences are from Hashiguchi and Nishida (2007). I classified them into five subfamilies.

[e]These three shark sequences (S2a, S2bP, and S2cP) are most similar to TAAR4. However, as I described, these shark TAARs may have diverged from the ancestral TAARs before the divergence of TAAR2-4 (see also phylogenies in Figs. 2.2, 2.3, and supplementary Fig S2.2).

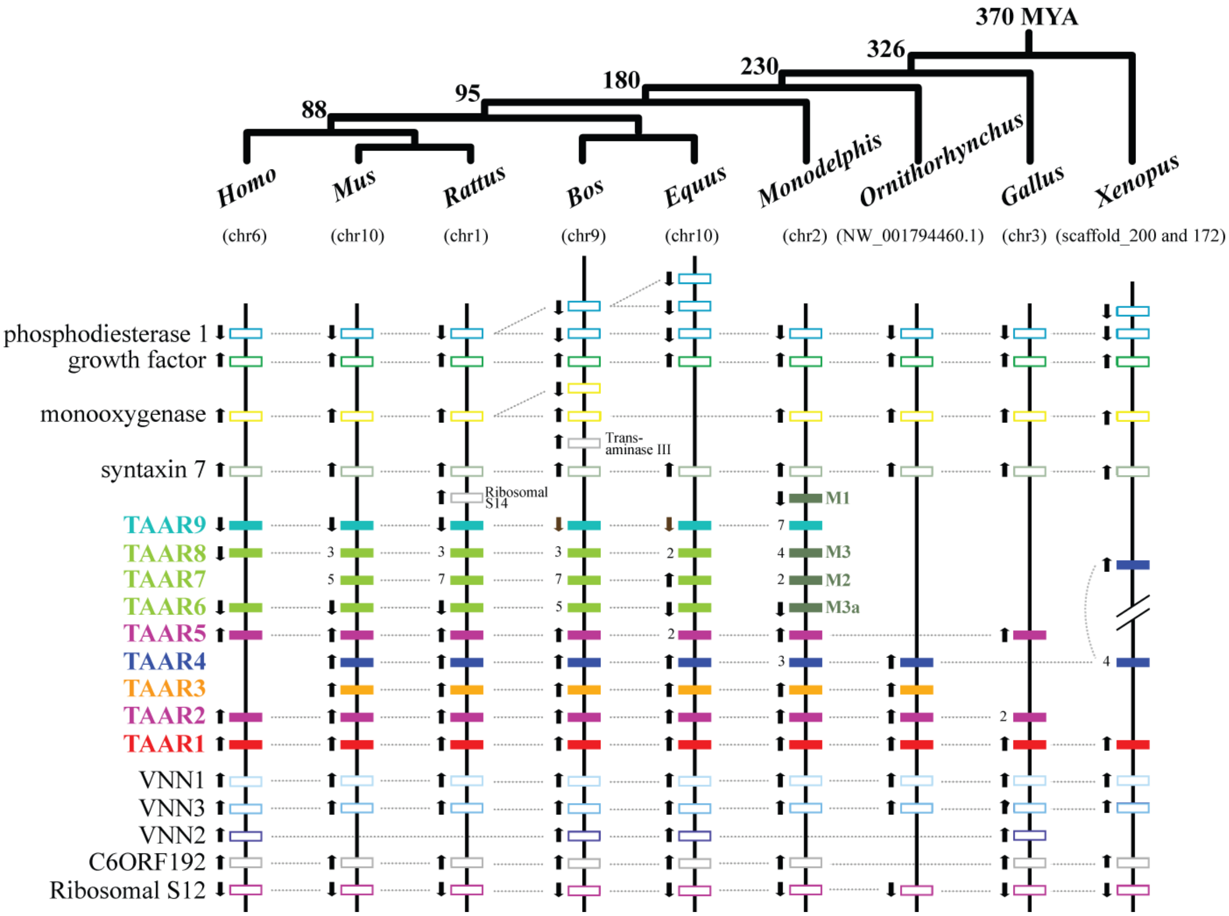


Figure 2.1. Syntenic relationship of the TAAR genes in nine vertebrate genomes. Only genomes in which all TAAR genes are located in one chromosome or no more than two scaffolds were examined. TAAR and adjacent non-TAAR genes are depicted by the closed and open boxes, respectively. Gene locations are not in scale. Black arrows indicate transcriptional directions. TAAR genes with tandemly duplicated functional copies are shown with dark gray boxes along the copy numbers. Metatherian-specific TAARs (M1, M2, and M3) are shown with light gray boxes. A current consensus of the tetrapod phylogeny with their approximate divergence times (million years ago; MYA) is illustrated at the top (Murphy et al. 2004; Blair and Hedges 2005). The chromosome or scaffold numbers are shown below the genus names.

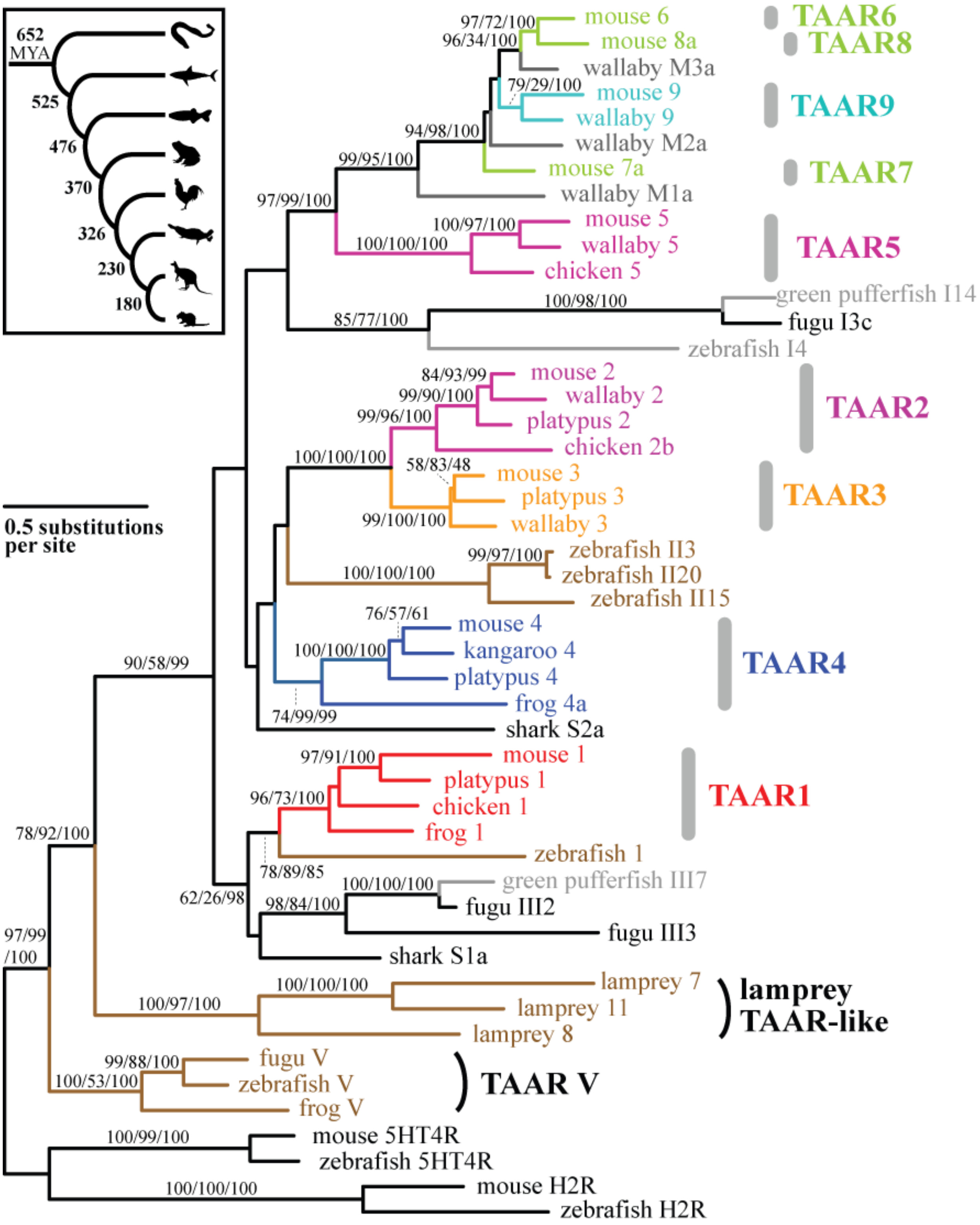

652
MYA
525
476
370
326
230
180
0.5 substitutions per site
97/72/100
96/34/100
mouse 6
mouse 8a
wallaby M3a
79/29/100
mouse 9
wallaby 9
94/98/100
wallaby M2a
99/95/100
mouse 7a
97/99/100
wallaby M1a
100/97/100
mouse 5
100/100/100
wallaby 5
chicken 5
100/98/100
green pufferfish I14
85/77/100
fugu I3c
zebrafish I4
84/93/99
mouse 2
99/90/100
wallaby 2
99/96/100
platypus 2
chicken 2b
100/100/100
58/83/48
mouse 3
platypus 3
99/100/100
wallaby 3
99/97/100
zebrafish II3
100/100/100
zebrafish II20
zebrafish II15
76/57/61
mouse 4
100/100/100
kangaroo 4
platypus 4
frog 4a
74/99/99
shark S2a
90/58/99
97/91/100
mouse 1
platypus 1
96/73/100
chicken 1
frog 1
zebrafish 1
78/92/100
78/89/85
100/100/100
green pufferfish III7
62/26/98
98/84/100
fugu III2
fugu III3
97/99
/100
shark S1a
100/100/100
lamprey 7
100/97/100
lamprey 11
lamprey 8
99/88/100
fugu V
100/53/100
zebrafish V
frog V
100/99/100
mouse 5HT4R
zebrafish 5HT4R
100/100/100
mouse H2R
zebrafish H2R
TAAR6
TAAR8
TAAR9
TAAR7
TAAR5
TAAR2
TAAR3
TAAR4
TAAR1
lamprey
TAAR-like
TAAR V

Figure 2.2. The maximum-likelihood phylogeny of TAAR proteins from ten representative animals. Four representative biogenic amine receptors (5HT4R: serotonin receptors, and H2R: histamine receptors) are used as the outgroup. The numbers at internal branches show the bootstrap support values (%) for the maximum-likelihood and neighbor-joining phylogenies and the posterior probability (%) for the Bayesian phylogeny in this order. Supporting values are shown only for the internal branches that have at least one method supporting higher than 70%. For TAAR V, teleost fish TAARs, and lamprey TAAR-like, I followed the gene names given by Hashiguchi and Nishida (2007). The inset illustrates a current consensus of the vertebrate phylogeny with their approximate divergence times (MYA) (Murphy et al. 2004; Blair and Hedges 2005).

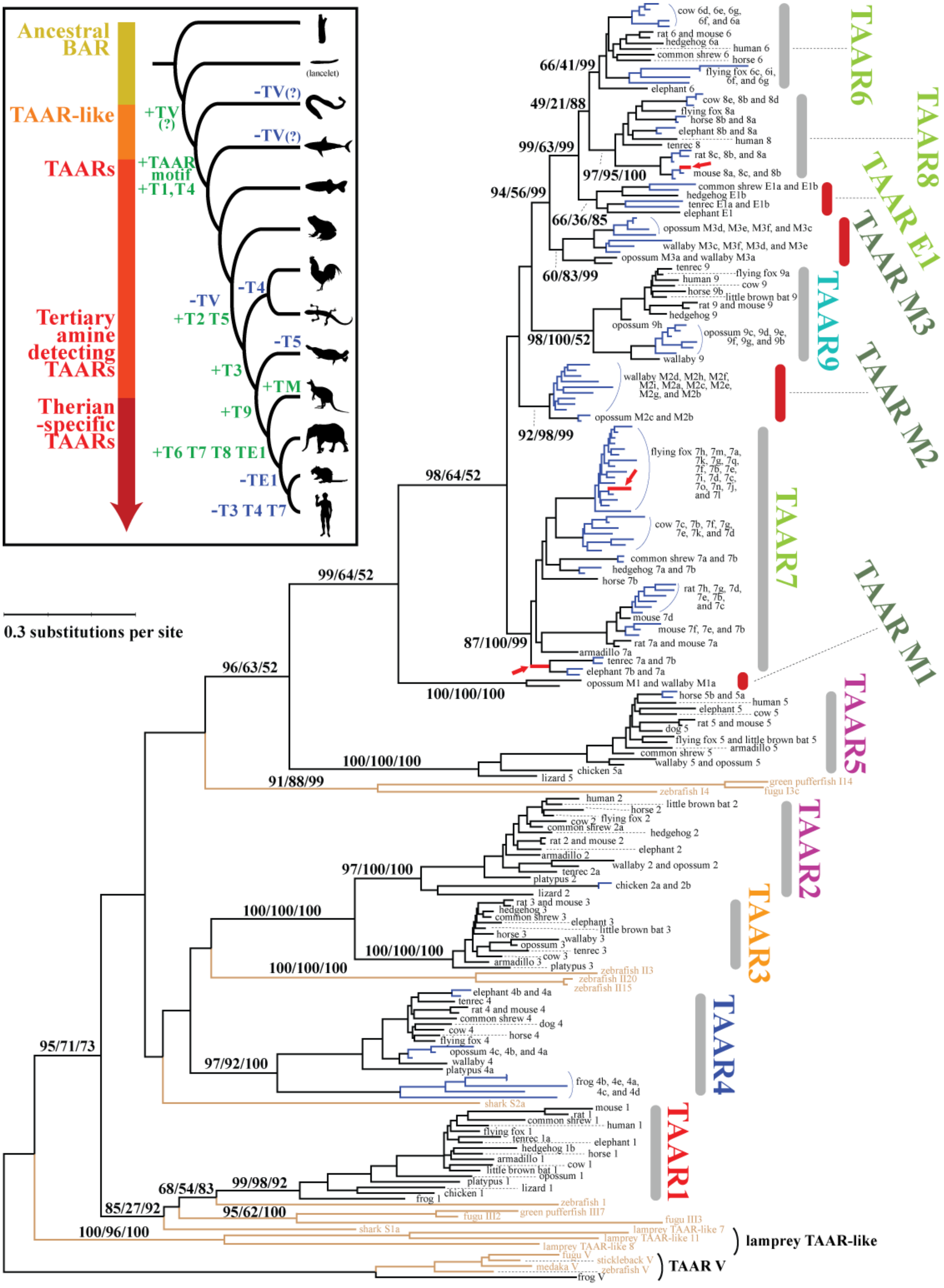

Ancestral BAR
TAAR-like
TAARs
Tertiary amine detecting TAARs
Therian -specific TAARs
(lancelet)
+TV (?)
-TV(?)
-TV(?)
+TAAR motif
+T1,T4
-TV
+T2 T5
-T4
-T5
+T3
+TM
+T9
+T6 T7 T8 TE1
-TE1
-T3 T4 T7
0.3 substitutions per site
66/41/99
49/21/88
99/63/99
97/95/100
94/56/99
66/36/85
60/83/99
98/100/52
92/98/99
98/64/52
99/64/52
87/100/99
100/100/100
96/63/52
100/100/100
91/88/99
97/100/100
100/100/100
100/100/100
100/100/100
95/71/73
97/92/100
68/54/83
99/98/92
85/27/92
95/62/100
100/96/100
cow 6d, 6e, 6g, 6f, and 6a
rat 6 and mouse 6
hedgehog 6a
human 6
common shrew 6
horse 6
flying fox 6c, 6i, 6f, and 6g
elephant 6
cow 8e, 8b and 8d
flying fox 8a
horse 8b and 8a
elephant 8b and 8a
human 8
tenrec 8
rat 8c, 8b, and 8a
mouse 8a, 8c, and 8b
common shrew E1a and E1b
hedgehog E1b
tenrec E1a and E1b
elephant E1
opossum M3d, M3e, M3f, and M3c
wallaby M3c, M3f, M3d, and M3e
opossum M3a and wallaby M3a
tenrec 9
flying fox 9a
human 9
cow 9
horse 9b
little brown bat 9
rat 9 and mouse 9
hedgehog 9
opossum 9h
opossum 9c, 9d, 9e, 9f, 9g, and 9b
wallaby 9
wallaby M2d, M2h, M2f, M2i, M2a, M2c, M2e, M2g, and M2b
opossum M2c and M2b
flying fox 7h, 7m, 7a, 7k, 7g, 7q, 7f, 7b, 7e, 7i, 7d, 7c, 7o, 7n, 7j, and 7l
cow 7c, 7b, 7f, 7g, 7e, 7k, and 7d
common shrew 7a and 7b
hedgehog 7a and 7b
horse 7b
rat 7h, 7g, 7d, 7e, 7b, and 7c
mouse 7d
mouse 7f, 7e, and 7b
rat 7a and mouse 7a
armadillo 7a
tenrec 7a and 7b
elephant 7b and 7a
opossum M1 and wallaby M1a
horse 5b and 5a
human 5
elephant 5
cow 5
rat 5 and mouse 5
dog 5
flying fox 5 and little brown bat 5
armadillo 5
common shrew 5
wallaby 5 and opossum 5
chicken 5a
lizard 5
green pufferfish I14
fugu I3c
zebrafish I4
human 2
little brown bat 2
horse 2
cow 2
flying fox 2
common shrew 2a
hedgehog 2
rat 2 and mouse 2
elephant 2
armadillo 2
wallaby 2 and opossum 2
tenrec 2a
platypus 2
chicken 2a and 2b
lizard 2
rat 3 and mouse 3
hedgehog 3
common shrew 3
elephant 3
little brown bat 3
horse 3
wallaby 3
opossum 3
tenrec 3
cow 3
armadillo 3
platypus 3
zebrafish II3
zebrafish II20
zebrafish II15
elephant 4b and 4a
tenrec 4
rat 4 and mouse 4
common shrew 4
dog 4
cow 4
horse 4
flying fox 4
opossum 4c, 4b, and 4a
wallaby 4
platypus 4a
frog 4b, 4e, 4a, 4c, and 4d
shark S2a
mouse 1
rat 1
common shrew 1
human 1
flying fox 1
tenrec 1a
elephant 1
hedgehog 1b
horse 1
armadillo 1
cow 1
little brown bat 1
opossum 1
platypus 1
lizard 1
chicken 1
frog 1
zebrafish 1
fugu III2
green pufferfish III7
fugu III3
shark S1a
lamprey TAAR-like 7
lamprey TAAR-like 11
lamprey TAAR-like 8
lamprey TAAR-like
fugu V
stickleback V
medaka V
zebrafish V
frog V
TAAR V
TAAR6
TAAR8
TAAR E1
TAAR M3
TAAR9
TAAR M2
TAAR7
TAAR M1
TAAR5
TAAR2
TAAR3
TAAR4
TAAR1

Figure 2.3. The maximum-likelihood phylogeny of TAAR proteins from 24 gnathostome genomes. All functional proteins in tetrapods, nine representative teleost proteins, and two elephant shark TAARs are included in the analysis. TAAR V as well as the TAAR-like sequences from the sea lamprey are used as the outgroup. The numbers at internal branches show the bootstrap support values (%) for the maximum-likelihood and neighbor-joining phylogenies and the posterior probability (%) for the Bayesian phylogeny in this order. Supporting values are shown only for the major internal branches that have at least one method supporting higher than 70%. Blue-colored branches indicate the species-specific gene duplications within a cluster supported by higher than 80% of bootstrap values or posterior probability for all methods. Red-colored branches and arrows indicate those identified to be under positive selection by the branch-site models of PAML analysis (see supplementary Table S2.3). Brown-colored branches indicate nine representative teleost TAARs, elephant shark TAARs, and TAAR-like proteins from sea lamprey. The inset illustrates the evolution of vertebrate TAARs with approximate timing of various gain (green) and loss (blue) events. The vertebrate phylogeny is based on Blair and Hedges (2005).

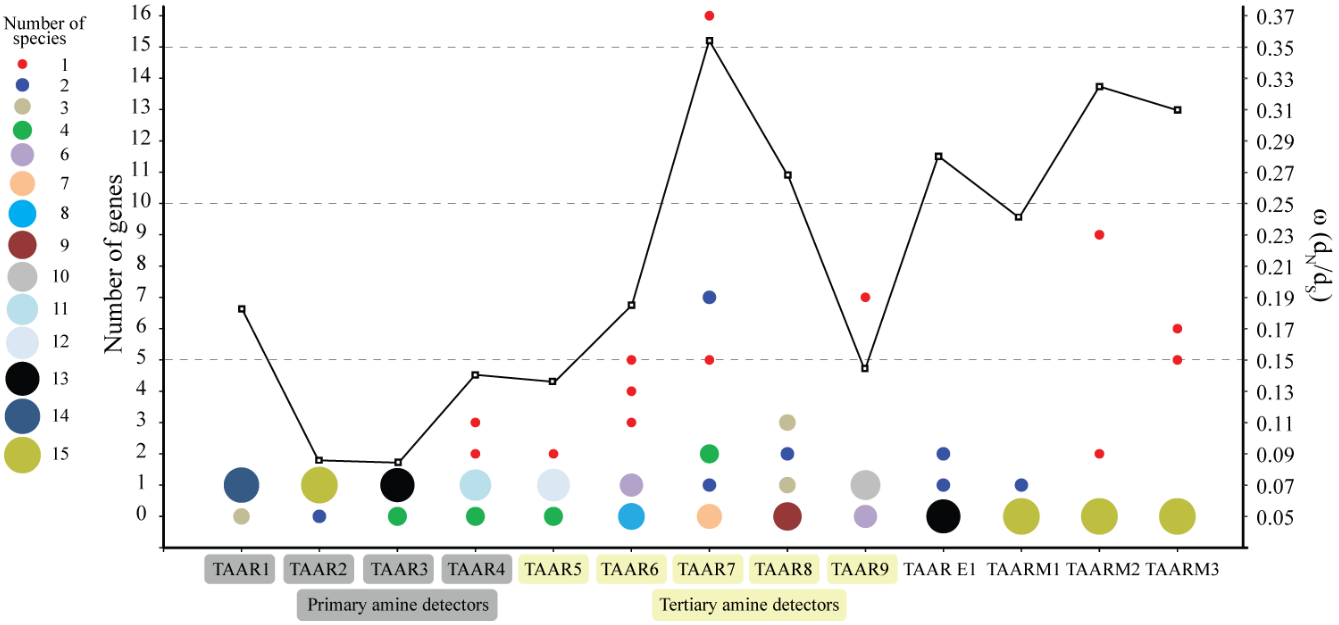


Figure 2.4. The number of TAAR genes within each TAAR subfamily for each therian species. The size of bubbles denotes the number of species. The average ω (d_N/d_S) calculated by the PAML M0 model for each TAAR subfamily is also plotted (open squares).

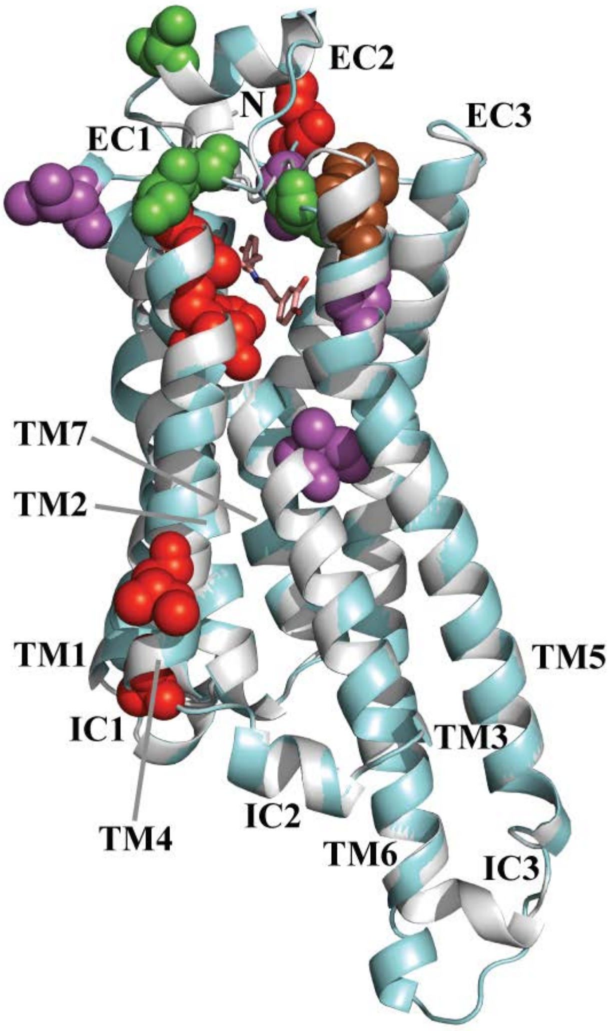


Figure 2.5. The 3D-structural model of the elephant TAAR7a protein (cyan) superimposed with the turkey β$_1$-adrenergic receptor (β$_1$AR, gray). The ligand of the β$_1$AR, dobutamine, is shown with the stick model. Positively selected sites are indicated by red (detected by the site model in TAAR7), green (detected by the branch-site model in flying fox TAAR7c and elephant TAAR7a), purple (detected by the site model in TAAR8), and brown (detected by the branch-site model in mouse TAAR8a). The transmembranes (TM) and internal/external loop (IC1-3 and EC1-3) regions as well as N-terminal (N) are labeled. The C-terminal is invisible locating behind TM1. See supplementary figure S2.5 for more details.

Chapter 3

Pseudogenization of Trace Amine–Associated Receptor Genes in Primates

3.0 Abstract for Chapter 3

Trace amine-associated receptors (TAARs) are considered as a second class of vertebrate olfactory receptors. TAARs consist of thirteen subfamilies in placental mammals and can be classified into two groups based on binding preferences for primary or tertiary amines. All therian-specific TAARs are tertiary amine detecting receptors. In Chapter 2, we showed that they underwent lineage-specific gene duplications in most mammals. However, in some primate lineages, no TAAR duplication was found and their genomes have smaller numbers of TAAR genes than other mammals. In order to elucidate what evolutionary force drives such lower number of TAARs, I conducted exhaustive mining of TAAR genes from twelve primate and northern treeshrew genomes. I found a total of 99 TAAR genes (including 48 pseudogenes) from the 12 primate genomes. They have in general smaller numbers of TAARs (ranging from 1 in white-cheeked gibbon and 8 in bushbaby) and have had only gene losses but no gene gains. Primates have lost all TAAR7 and most of TAAR8, although in other mammals these genes showed the patterns of positive selection. Pseudogenization events are likely to be accelerated in arboreal life and the change in the nose shape of Haplorhini species after the divergence from Strepsirrhini.

3.1 Background

Trace amines (TAs) are endogenous amine compounds that include 2-phenylethylamine (PEA), m-tyramine (m-TYR), ρ-tyramine (ρ-TYR), *meta*-octopamine (m-TA), *para*-octopamine (p-TA), 3-iodothyronamine (T_1AM), tryptamine (TRY), and *N*,*N*-dimethyltryptamine (DMT). TAs are putative regulatory elements in the brain (Berry 2004) and thus of importance in understanding several human brain diseases. Current studies suggest that regulatory roles of the TA system affect brain diseases such as substance abuse, insomnia, depression, attention deficit hyperactivity disorder, bipolar, schizophrenia, and other neuropsychiatric diseases (Premont et al. 2001; Duan et al. 2004; Wolinsky et al. 2007a; Serretti et al. 2009; Pae et al. 2010). The trace amine-associated receptors (TAARs) were discovered in search of the receptors activated by TAs in the brain (Borowsky et al. 2001; Bunzow et al. 2001). The TAAR6 gene (also known as TRAR4 or TA4) is reported to be associated with schizophrenia (Duan et al. 2004; Vladimirov et al. 2007; Serretti et al. 2009). Interestingly, rat TAAR1 is not only activated by classical TAs but also by synthetic analogues such as 3,4-methylenedioxymethamphetamine (MDMA, known as ecstasy), *d*-lysergic acid diethylamide (LSD), and amphetamine (Bunzow et al. 2001). Furthermore, TAAR4 is stimulated by 2-phenylethylamine, which is a carnivore odor that evokes physiological and behavioral responses in two prey species (mouse and rat) (Ferrero et al. 2011). TAs and TAARs are therefore important in understanding many psychiatric human disorders as well as critical roles in sensing predator and prey odors.

As shown in Chapter 2, placental mammals are known to have more TAAR subfamilies, in total thirteen (TAAR1-9, E1, and M1-3), than archosaurs and amphibians. However, as also shown in Chapter 2, the human and marmoset genomes have a smaller numbers of TAAR genes than other mammalian species. Stäubert et al. (2010) demonstrated that pseudogenization of TAAR3 and TAAR4 happened before the divergence of human and gorilla. They also showed that independent pseudogenizations have occurred in the marmoset lineage for both TAAR3 and TAAR4. While mammalian TAAR7 genes are found to be under positive selection, this gene has been lost in the human and marmoset genomes (Chapter 2). These findings prompted me to further characterize the evolutionary patterns of the TAAR genes in primates. Although nearly complete genomes have been released from twelve primates, chemosensory evolution among these primates has not been investigated thoroughly and is still poorly understood. Furthermore, the study of primate genomes is of importance and it will provide insight into human adaptation. In this chapter, I identified complete repertoires of functional TAAR genes and pseudogenes from twelve primate genomes. I examined their gene structures and their evolutionary patterns. I found that the primate genomes have generally smaller numbers of TAARs compared to other mammals. All primate TAARs remained as a single copy gene. Most of pseudogenization events except for TAAR1 have occurred independently in different Haplorhini species after the divergence from Strepsirrhini. Selective constraint for primate TAARs is weakher than that for other mammalian orthologues. I speculate that the TAAR pseudogenizations are resulted from natural selection possibly because of a role in susceptibility to some brain diseases.

3.2 Results and Discussion

3.2.1. Identification of TAAR genes in primates.

TAAR gene candidates were searched from twelve primate genomes as well as the northern treeshrew (*Tupaia belangeri*) genome (supplementary Table S3.1). A total of 99 TAAR genes (including four incomplete genes and 48 pseudogenes) were identified from the twelve primate genomes (Table 3.1). No intact TAAR7 sequences were identified from the twelve primate and northern treeshrew genomes. The numbers of TAAR genes varied significantly among the primate species, ranging from one in white-cheeked gibbon (*Nomascus leucogenys*) to seven in bushbaby (*Otolemur garnettii*). Compared to other mammalian species (Table 2.2), these primate genomes have in general smaller numbers of functional TAAR genes.

3.2.2. Phylogenetic analysis of primate TAARs.

To clarify the evolutionary relationships among the primate TAAR genes, phylogenetic analyses were performed (Figure 3.1). Three sea lamprey TAAR-like proteins were used as outgroups because they are the most ancestral among all TAAR genes (Chapter 2). In Figure 3.1, each cluster of eight main subfamilies (TAAR1 to TAAR9) is strongly supported by high bootstrap values (> 81% in the maximum likelihood phylogeny, > 99% in the neighbor-joining, and all 100% posterior probability in the Bayesian phylogeny). As described in Chapter 2, TAARs are divided into two groups: the primary-amine detectors (from TAAR1 to TAR4) and the tertiary-amine detectors (from TAAR5 to TAAR9) (see also Ferrero et al. 2012). Our phylogenetic analysis supports this two-group classification (>96% by at least one method; Figure 3.1).

Phylogenetic analysis was also done based on the concatenated TAAR protein supermatrix (2,809 amino acids) including twelve primates, treeshrew, mouse, and rat (all belong to Euarchontoglires). The resultant phylogeny shown in supplementary Figure S3.1 is consistent with the recent primate phylogenetic studies and known taxonomical relationship among these species (Fabre et al. 2009; Perelman et al. 2011).

3.2.3. Pseudogenization of TAARs in primates.

All TAAR subfamilies except for TAAR1 have experienced pseudogenization in different primate lineages. In Figure 3.2, all pseudogenization and gene loss (as well as gain) events throughout the primate evolution are summarized.

In other mammals, the TAAR7 subfamily has the highest level of gene number variation. It has also been shown to evolve under the influence of positive selection (Chapter 2). On the contrary, all 12 primate genomes examined in this study do not have an intact TAAR7. Bushbaby (*Otolemur garnettii*) belongs to the suborder Strepsirrhini and, with lemurs, diverged off from other primates earlier. Bushbaby possesses all eight but TAAR7 subfamily. No identifiable TAAR7 gene nor pseudogene exists in the bushbaby genome. The order Scandentia, which includes northern treeshrew (*T. belangeri*), is closest to the primates (Murphy et al. 2007; Prasad et al. 2008). The northern treeshrew genome also possesses all TAAR subfamilies except TAAR7. However, the mouse and rat genomes have multiple TAAR7 genes (5-7 genes). Therefore, TAAR7 genes have been maintained until the euarchontoglirean lineages, which include mouse, cow, horse, and elephant, but subsequently lost in the common ancestor of primate and scandentia (Figure 3.2).

Pseudogenization of TAAR2 gene may have happened very recently after the divergence of human and chimpanzee and even after before the divergence of chimpanzee and bonobo (4.5 - 1 MYA) (Prufer et al. 2012) (see Materials and Methods) (supplementary Figure S3.2(a)). Indels associated with pseudogenization in TAAR3 (supplementary Figure S3.2(b)) and in TAAR4 (supplementary Figure S3.2(c)) seem to have occurred in the lineage leading to the African apes (subfamily Homininae). In TAAR8 sequences, two nucleotide deletions at positions 748 and 749 are shared in the lineage leading to Anthropoidea (infraorder Simiiformes) except for human TAAR8 and orangutan TAAR8P (supplementary Figure S3.2(d)). It is more likely that pseudogenization of TAAR8 happened independently at the same positions in different primate lineages. An alternative explanation, more parsimonious but less plausible, is that these two deletions occurred in the common ancestor of Anthropoidea and subsequently TAAR8 sequences in human and orangutan have been resurrected by gaining the two missing nucleotides. Such an implausible resurrection event was attributed to a member of IRG (immunity-related GTPases) in human evolution (Bekpen et al. 2009).

The white-cheeked Gibbon (*Nomascus leucogenys*) possesses only one intact TAAR (supplementary Table S3.1). Hylobatidae is known to have extremely rapid chromosome evolution (Roberto et al. 2007; Misceo et al. 2008). All human TAAR genes are located on chromosome 6, which corresponds to six chromosomes (NLE1, NLE3, NLE8, NLE17, NLE18, and NLE22) in the white-cheeked gibbon genome (Roberto et al. 2007). Although all gibbon TAAR genes are located in one single scaffold (GL397266.1), I speculated that a higher rate of segmental rearrangements may render the relaxation of the negative selection and acts as a driving force in TAAR gene loss.

As described in Chapter 2, TAARs are classified into two groups based on binding preferences for primary or tertiary amines. Tertiary-amine detecting TAARs (TAAR5-9) are therian-specific receptors, which have recently emerged after the divergence between prototherian and therian mammals (230 – 166 million years ago; MYA). They are subjected to rapid species-specific tandem gene duplication in most mammalian species (Chapter 2). For example, the cow (*Bos taurus*) genome possesses sixteen tertiary-amine detecting TAARs (5 TAAR6s, 7 TAAR7s, 3 TAAR8s, and one TAAR9). In euarchontoglirean species, two rodentia genomes (mouse and rat) also have a high number of TAAR7 (5 - 7 genes) and TAAR8 (3 genes). This trend is consistently observed in the northern treeshrew genome, which has theeshrew-specific duplicated copies of TAAR6 (two) and TAAR8 (five) (Figure 3.1). In the primate lineage, however, no extra copy (gene gain or gene duplication) of TAARs was found. Possible gene duplication events were identified only in orangutan TAAR3 and gorilla TAAR8 but all of them appeared to be pseudogenes. Furthermore, based on their shared stop codons (for TAAR3P) and frame-shifting insertions (TAAR8P), their duplications must have happened after these changes happened making them pseudogenes. All functional primate TAAR genes have apparently remained as a single copy gene. This is different from what we observed typically in mammalian TAAR evolution.

3.2.4. Dispensability of primate TAARs.

Mammalian TAARs are known to have low fractions of pseudogenes (Hashiguchi and Nishida 2007). For instance, mouse and rat genomes have 15 and 17 intact TAARs but only 1 and 2 pseudogenes, respectively (Gloriam et al. 2007). Stäubert et al. (2010) pointed that, however, pseudogenization events in TAAR3-5 are more frequent in primates than in other mammals. Disruptions of all TAARs except for TAAR1 are found more often in primate TAARs than other mammalians (Table 3.1).

Pseudogenization events are more frequent in Haplorhini (including Simiiformes and Tarsiiformes) after the divergence (87 million years ago, Perelman et al. 2011) from Strepsirrhini (including Lemuriformes and Lorisiformes). Although primates are generally divided into two suborders, Simiiformes and Prosimii (Tarsiiformes, Lemuriformes, and Lorisiformes), the other classification of Haplorhini and Strepsirrhini was divided on the basis of the features of the nose shape. While the name "Strepsirrhini" is derived from a “curly" nostril on the rhinarium (moist area of the nasal tip in mammals or wet nose, an ancestral condition), Haplorhini means “simple nose” that lacks a rhinarium. Mammals with rhinarium are known to have very sensitive and more acute olfaction capacity. In addition to the loss of a rhinarium, the size of the main olfactory epithelium (MOE, the back of the nose into which air flows) is reduced in Haplorhini primates compared to Strepsirrhini (Barton 2006). The sensory neurons in the mammalian MOE have two types of chemosensory receptors, TAARs and ORs (Kaupp 2010). All mouse TAARs except for TAAR1 are expressed in MOE (Liberles and Buck 2006; Fleischer et al. 2007). Thus, the loss of a rhinarium and smaller size of MOE in haplorhines is very likely associated to their decreased reliance on olfaction sensitivities (Smith and Rossie 2006). Furthermore, the degeneration of OR genes in primates have been observed (Gilad et al. 2006; Matsui et al. 2010). Therefore, frequent pseudonization of TAARs found especially in the Haplorhini lineage can be also considered to be the results of relaxed selection due to their decreased reliance on olfaction.

Gradual degeneration seems to be the major trend in the evolution of primate TAARs. More than half of TAARs had been lost due to multiple independent pseudogenization events particularly in the Hominoidea genomes except for human (Table 3.1). Ferrero et al. (2011) demonstrated that TAARs play important roles in sensing predator and prey odors. For example, rat and mouse TAAR4 is stimulated by 2-phenylethylamine, which is a carnivore odor from mountain lion, tiger, and jaguar. However, all African apes have lost TAAR4. The primate ancestor probably arose as arboreal animals and their characteristics still remain as adaptations to this life style (shortened rostrum with stereoscopic vision, opposable hallux and pollux, and highly mobile radius and ulna in the forelimb) (Cartmill and Smith 2009). Living in trees significantly reduces the predator exposures and makes escaping from many ground-living predators easier (Hart 2007). Thus, it is conceivable that arboreal life adapted by primate species may have decreased the reliance on chemosensing of predators, leading to non-functionalization of primate TAARs. Interestingly, primate OR gene repertoires are also in a phase of deterioration (Dong et al. 2009). On the contrary, many cercopithecid species including macaques and baboons still have a wide array of predators such as leopard, tiger, or cheetah and thus they display a variety of behaviors in response to the threat of predation (Enstam 2007). They have functional TAAR4 and a higher number (~ 6 genes) of TAARs among haplorhines. It implies that the pseudogenization of TAAR4 was not the shared ancestral event, but rather lineage-specific multiple independent events.

3.2.5. Selection patterns among TAAR subfamilies.

If the TAAR genes in primates have less critical functions, they should have evolved under relaxed selective constraints. In order to examine the level of selective constraints, the ratio of nonsynonymous to synonymous distances (d_N/d_S or ω) was estimated for each TAAR subfamily among primates. Generally, the average ω for each TAAR subfamily was higher in primate than that in non-primate mammalian orthologs. Overall average ω for TAAR subfamilies was higher in primate than that in non-primate mammalian orthologs (the overall average ω's are 0.2232 from primates and 0.1523 from non-primate mammals) (supplementary Table S3.2). This indicates that primate TAARs are subject to relaxed purifying selection. Alternatively, a limited number of sites of these proteins may have been under positive selection. Furthermore, the overall average ω (0.3813) is significantly higher when estimated only using Haplorhini primates than when using non-primate mammals (supplementary Table S3.2).

3.2.6. Different selective forces operating on TAAR subfamilies.

In order to confirm the hypothesis that there are different selective pressures within primate TAAR subfamilies, PAML tests based on the branch models (Yang 2007) were applied. It tests an alternative hypothesis where two ω's are allowed in specific branches against the null hypothesis with a single ω in all branches. Using this test, the levels of selective pressures were compared between Haplorhini TAARs and Strepsirrhini TAARs. The alternative hypothesis with two ω's was found to be significantly better than the null hypothesis with a single ω for most of the TAAR subfamilies ($P < 0.01$ for TAAR2, TAAR3, TAAR4, TAAR6, and TAAR9; supplementary Table S3.3). Estimated ω's were about two or three times higher in Haplorhini (ω_1 in R2) compared to the Strepsirrhini lineages (ω_0 in R2) (supplementary Table S3.3) indicating further relaxed selective constrains in Haplorhini TAAR subfamilies after the divergence from Strepsirrhini TAARs. Similar tests were done within Haplorhini TAARs comparing each different group against others. Significant difference was observed only when tarsier TAARs were compared against others. Tarsier TAAR2 and TAAR4 genes showed significantly lower ω's compared to those estimated from other Haplorhini primates (Catarrhini) ($P<0.01$) (supplementary Table S3.4).

Furthermore, PAML tests with branch-site models, which can detect a short episode of positive selection occurring in a small fraction of amino acids (Zhang et al. 2005), were performed. The tests were conducted both including and excluding pseudogenes. The models that allowed $\omega > 1$ had a significant fit to the data on chimpanzee TAAR6 ($P < 0.0001$) and marginally significant on human TAAR2 ($P < 0.05$) (supplementary Table S3.5).

3.2.7. Positive-selection sites located in the potential ligand-binding sites.

The amino acid sites under positive selection signatures were identified with the PAML branch-site models using the Bayes Empirical Bayes inference (Yang et al. 2005). Three sites were identified in human TAAR2 (positions 2, $99^{3.28}$, and $130^{3.59}$) and six sites in chimpanzee TAAR6 (positions 7, $96^{3.25}$, $97^{3.26}$, $114^{3.43}$, $115^{3.44}$, and $195^{5.43}$) (supplementary Table S3.5). Four of six sites identified in chimpanzee TAAR6 (positions $96^{3.25}$, $97^{3.26}$, $114^{3.43}$, and $115^{3.44}$) had their posterior probabilities higher than 95%, indicating strong positive selection. For their spatial distribution of these nine positive-selection sites, homology modeling of TAAR protein structures was performed (Figure 3.3). Six of the nine amino acids identified as being under positive selection are located in or near the extracellular regions of the receptors (including two in the N-terminal region) (Figure 3.3). Three positively selected sites in human TAAR2 (position $99^{3.28}$) and chimpanzee TAAR6 (positions $96^{3.25}$ and $195^{5.43}$) correspond to residues identified to be directly involved with ligand-binding on β-adrenergic receptors 1 and 2 (Kleinau et al. 2011; Warne et al. 2011; Warne et al. 2012). Thus, these substitutions may have affected ligand-binding activities and specificities of these TAARs.

The results in this study indicated that the TAAR family genes in primates, particularly those in haplorhines, have undergone relaxed selection. Because Strepsirrhini species still have almost all TAARs, the major morphological transition from Strepsirrhini to Haplorhini is most likely associated to this change in selection constraints. Relaxed selection on TAAR subfamilies has caused accumulation of multiple independent mutations, resulting in non-functionalization of multiple TAAR genes. This gradual degeneration process in the primate TAARs has been accompanied also with possible positive selection in recent human and chimpanzee evolution.

Wang et al. (2006) demonstrated a possible case of adaptive pseudogenization in human. They showed that while the CASPASE12 gene, a cysteine-aspartic acid protease (caspase) protein participating in inflammatory and innate immune response to endotoxins, is functional in all mammals, the null allele of this gene has been nearly fixed in human population. The functional gene is likely deleterious to humans as the null allele is known to be associated with a reduced incidence and mortality of severe sepsis. Similarly, if TAAR pseudogenizations confer lowered susceptibility to psychiatric disorders (Boulton 1980; Premont et al. 2001; Branchek and Blackburn 2003), it would be beneficial for primate evolution. So far, however, any evidence was observed that non-human primates are affected by the same psychiatric disorders as humans. Further studies on primate TAAR functions will provide more insights into how primate TAAR function would be different for those with mammalians. This can be related to clinical implications and can provide further insight into therapeutic potentiality.

3.3 Conclusion

We have identified TAAR genes in twelve primate genomes and demonstrated that they have in general a smaller number of TAARs compared to other mammalian species. Primate TAARs have experienced only gene losses but no gene gains. The TAAR genes in primates appear to be under relaxed selection, shown as higher ω. Pseudogenization of TAAR genes are likely to be accelerated after the change of the nose shape in Haplorhini species. Relaxed selection in primate TAARs has resulted in multiple independent mutations and smaller numbers compared to other mammalians

3.4 Materials and Methods

3.4.1. Genome sequences and TAAR gene mining.

Thirteen genomic sequences were obtained from multiple sources (supplementary Table S3.1). Previously reported TAAR sequences (Chapter 2) were used as queries. Similarity search was performed using the Basic Local Alignment Search Tool (BLAST, ver. 2.2.26) programs (Altschul et al. 1990). The method for mining TAAR genes is essentially the same as Chapter 2. TAAR candidates were subsequently used as queries against their genomes again to find any additional candidates. These steps were recursively performed until no other TAAR candidate sequences were detected from each genome. For the TAAR naming, we followed the nomenclature of Maguire et al. (2009).

Dong et al. (2012) shows the number of TAARs from five primate genomes (*Homo sapiens*, *Pan troglodytes*, *Pongo pygmaeus abelii*, *Macaca mulatta*, and *Callithrix jacchus*. However, the number of TAARs is not same with that of ours (*e.g.*, 2 in ours and 1 in from *Callithrix jacchus*). Also, they showed that human has 5 TAARs (TAAR1, TAAR2, TAAR3, TAAR4, and TAAR5) but 6 TAARs (TAAR1, TAAR2, TAAR5, TAAR6, TAAR8, and TAAR9) in ours. They used automatic data-mining methods which are probably not completed to mine all TAARs.

The TAAR genes are intron-less and encoded in a single exon. TAAR2 genes, also known as GPR58, are exceptions and have two exons. To determine exon and intron boundaries for TAAR2, the coding sequences were predicted using GeneWise (ver. 2.2) (Birney et al. 2004). Highly conserved first exons were found in six primates (human, chimpanzee, bonobo, gorilla, orangutan, and rhesus macaque). The average length of six TAAR2 introns is 6,070 bps (6,042 bps in orangutan to 6,097 in chimpanzee).

The chimpanzee TAAR2P (NG_004780.2) is likely a pseudogene due to a nucleotide deletion (nucleotide position 861 in human TAAR2), which is shared by the bonobo TAAR2P. Note that the bonobo TAAR2P (XP_003827712) from the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov) is annotated as an intact gene and has “N” in that position with “low quality position”. However, this bonobo gene has a nucleotide deletion in the same position with the chimpanzee TAAR2P as well as a unique stop codon at N-terminal. Therefore, we consider this gene as a pseudogene and call it TAAR2P.

3.4.2. Multiple sequence alignments and phylogenetic analysis.

Multiple alignments of TAAR protein sequences were generated using MAFFT with the L-INS-i algorithm (ver. 7.050b) (Katoh and Standley 2013). All TAAR sequences and alignments are available from: http://bioinfolab.unl.edu/emlab/primate_TAAR. All amino acid positions shown in this study are numbered based on the human TAAR1 sequence in the alignment. The Ballesteros and Weinstein system numbering is shown as a superscript according to the turkey β1-adrenergic receptor (β1AR, P07700) sequence. All pseudogenes identified in this study are included in the multiple alignments and the phylogenetic analysis after removing the codon to have frame-shifting insertions/deletions or in-frame stop codons. To generate the TAAR protein supermatrix, eight TAAR subfamily alignments (2,809 amino acid sequences) from twelve primates, treeshrew, mouse, rat, and cow were concatenated.

Phylogenetic relationships were reconstructed by the maximum-likelihood method with the PROTGAMMAJTT substitution model (JTT matrix with gamma-distributed rate variation) using RAxML (ver. 7.0.4) (Stamatakis 2006). The neighbor-joining phylogenetic method (Saitou and Nei 1987) was performed using the Phylip package (ver. 3.67) (Felsenstein 2005). The protein distances were estimated with the JTT substitution model with gamma-distributed rate variation with α=1.3 (Yang 1994) estimated from the maximum-likelihood method implemented RAxML. Bayesian phylogenetic inference was performed using the MrBayes v3.1.2 package (Huelsenbeck and Ronquist 2001) with the JTT substitution model with gamma-distributed rate variation. The Markov chain Monte

Carlo search was run for 10^6 generations, with a sampling frequency of 10^3, using three heated and one cold chain and with a burn-in of 10^2 trees. Non-parametric bootstrapping with 1000 pseudo-replicates (Felsenstein 1985) was used to estimate the confidence of branching patterns for the maximum-likelihood and neighbor-joining methods. Presentation of the phylogenies was done with FigTree (http://tree.bio.ed.ac.uk/software/figtree). All phylogenies are available from: http://bioinfolab.unl.edu/emlab/primate_TAAR.

3.4.3. Transmembrane protein topology prediction.

To predict the transmembrane protein topology, which includes N-terminal, transmembrane (TM), intercellular loop (IC), extracellular loop (EC), and C-terminal regions, we used HMMTOP (ver. 2.1) (Tusnady and Simon 2001) and Phobius (ver. 1.01) (Kall et al. 2007).

3.4.4. Tests of selection patterns.

The branch-specific and branch-site models implemented in `codeml` of the PAML (Phylogenetic Analysis by Maximum Likelihood) package (version 4.5) were used (Yang 2007). The one-ratio model (M0) for estimating an equal ω ratio for all branches in the phylogeny was compared against the free-ratio model, which assumes an independent ω for each branch. For the branch models, I performed LRTs with d.f. = 1 between a one-ratio model (R1; the same ω for all branches) and a two-ratio model (R2; two independent ω's) (Yang 1998; Yang and Nielsen 2002). The branch-site models were applied to detect positively selected sites along specific branches (Yang and Nielsen 2002; Zhang et al. 2005). Positively selected amino acid sites are identified based on Bayes Empirical Bayes posterior probabilities (Yang et al. 2005). In these models, positive selection was allowed on a specific, "foreground", branch, and the likelihood-ratio tests (LRTs) (d.f. = 1) were performed against null models that assume no positive selection. The branch-site test of positive selection ("Test 2" in Zhang et al. 2005) has four site classes: 0, 1, 2a, and 2b. For the site classes 0 and 1, all codons are under purifying selection ($0 < \omega_0 < 1$) and under neutral evolution ($\omega_1 = 1$), respectively, on all branches. For the site classes 2a and 2b, positive selection is allowed on the foreground branches ($\omega_2 \geq 1$) but the other, "background", branches are under purifying selection ($0 < \omega_0 < 1$) and under neutral evolution ($\omega_1 = 1$), respectively. For the null model, ω_2 is fixed as 1. For each subfamily phylogeny, tests were done using each branch (from both internal and terminal branches) as the foreground. All PAML analyses were carried out using the F3X4 model of codon frequency (Goldman and Yang 1994). The level of significance (*P*) for the LRTs was estimated using a χ^2 distribution with given degrees of freedom (d.f.) and the test statistic calculated as twice the difference of log-likelihood between the models ($2\Delta ln\mathrm{L}= 2[ln\mathrm{L}_1 – ln\mathrm{L}_0]$ where L_1 and L_0 are the likelihoods of the alternative and null models, respectively). We performed LRTs with d.f. = 4 between a one-ratio model and a free-ratio model.

3.4.5. Protein structural homology modeling. Homology-based structural modeling of TAAR proteins was performed using the SWISS-MODEL Web server (http://swissmodel.expasy.org) (Arnold et al. 2006). The same template, the B-chain of the turkey (*Meleagris gallopavo*) β_1-adrenergic receptor (β_1AR; 4AMJ), was selected for the human TAAR2 and chimpanzee TAAR6 proteins. The root mean squared deviation (RMSD) and the QMEAN score for the human TAAR2 and chimpanzee TAAR6 are 2.30 Å and 2.20 Å and 0.251 and 0.241. The graphical representation of TAAR structures was prepared with PyMOL (version 1.3) (DeLanoScientific, San Carlos, CA).

3.5 Literature Cited

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol*. 215:403-410.

Arnold K, Bordoli L, Kopp J, Schwede T. 2006. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics*. 22:195-201.

Bailer U, Leisch F, Meszaros K, Lenzinger E, Willinger U, Strobl R, Gebhardt C, Gerhard E, Fuchs K, Sieghart W, Kasper S, Hornik K, Aschauer H. 2000. Genome scan for susceptibility loci for schizophrenia. *Neuropsychobiology*. 42:175-182.

Balakirev ES, Ayala FJ. 2003. PSEUDOGENES: Are They "Junk" or Functional DNA? *Annu Rev Genet*. 37:123-151.

Bekpen C, Marques-Bonet T, Alkan C, Antonacci F, Leogrande MB, Ventura M, Kidd JM, Siswara P, Howard JC, Eichler EE. 2009. Death and resurrection of the human IRGM gene. *PLoS Genet*. 5:e1000403.

Berry MD. 2004. Mammalian central nervous system trace amines. Pharmacologic amphetamines, physiologic neuromodulators. *J Neurochem*. 90:257-271.

Birney E, Clamp M, Durbin R. 2004. GeneWise and Genomewise. *Genome Res*. 14:988-995.

Borowsky B, Adham N, Jones KA, Raddatz R, Artymyshyn R, Ogozalek KL, Durkin MM, Lakhlani PP, Bonini JA, Pathirana S, Boyle N, Pu X, Kouranova E, Lichtblau H, Ochoa FY, Branchek TA, Gerald C. 2001. Trace amines: Identification of a family of mammalian G protein-coupled receptors. *Proc Natl Acad Sci USA*. 98:8966-8971.

Boulton AA. 1980. The properties and potential function of some brain trace amines. *Prog Clin Biol Res*. 39:291-303.

Branchek TA, Blackburn TP. 2003. Trace amine receptors as targets for novel therapeutics: legend, myth and fact. *Curr Opin Pharmacol*. 3:90-97.

Bunzow JR, Sonders MS, Arttamangkul S, Harrison LM, Zhang G, Quigley DI, Darland T, Suchland KL, Pasumamula S, Kennedy JL, Olson SB, Magenis RE, Amara SG, Grandy DK. 2001. Amphetamine, 3,4-Methylenedioxymethamphetamine, Lysergic Acid Diethylamide, and Metabolites of the Catecholamine Neurotransmitters Are Agonists of a Rat Trace Amine Receptor. *Mol Pharmacol*. 60:1181-1188.

Cao Q, Martinez M, Zhang J, Sanders AR, Badner JA, Cravchik A, Markey CJ, Beshah E, Guroff JJ, Maxwell ME, Kazuba DM, Whiten R, Goldin LR, Gershon ES, Gejman PV. 1997. Suggestive Evidence for a Schizophrenia Susceptibility Locus on Chromosome 6q and a Confirmation in an Independent Series of Pedigrees. *Genomics*. 43:1-8.

Cartmill M, Smith FH. 2009. The Human Lineage: Wiley-Blackwell, Hoboken, New Jersey.

Dong D, Jin K, Wu X, Zhong Y. 2012. CRDB: Database of Chemosensory Receptor Gene Families in Vertebrate. *PLoS ONE*. 7:e31540.

Duan J, Martinez M, Sanders AR, Hou C, Saitou N, Kitano T, Mowry BJ, Crowe RR, Silverman JM, Levinson DF, Gejman PV. 2004. Polymorphisms in the Trace Amine Receptor 4 (TRAR4) Gene on Chromosome 6q23.2 Are Associated with Susceptibility to Schizophrenia. *Am J Hum Genet*. 75:624-638.

Enstam KL. 2007. Effects of Habitat Structure on Perceived Risk of Predation and Anti-Predator Behavior of Vervet (*Cercopithecus aethiops*) and Patas (*Erythrocebus patas*)

Monkeys. In: SL Gursky, KAI Nekaris, editors. Primate Anti-Predator Strategies Springer US. p. 308-338.

Fabre PH, Rodrigues A, Douzery EJ. 2009. Patterns of macroevolution among Primates inferred from a supermatrix of mitochondrial and nuclear DNA. *Mol Phylogenet Evol*. 53:808-825.

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

Felsenstein J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Ferrero DM, Lemon JK, Fluegge D, Pashkovski SL, Korzan WJ, Datta SR, Spehr M, Fendt M, Liberles SD. 2011. Detection and avoidance of a carnivore odor by prey. *Proc Natl Acad Sci USA*. 108:11235-11240.

Ferrero DM, Wacker D, Roque MA, Baldwin MW, Stevens RC, Liberles SD. 2012. Agonists for 13 Trace Amine-Associated Receptors Provide Insight into the Molecular Basis of Odor Selectivity. *ACS Chem Biol*. 7:1184-1189.

Gloriam D, Fredriksson R, Schiöth HB. 2007. The G protein-coupled receptor subset of the rat genome. *BMC Genomics*. 8:338.

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

Hart D. 2007. Predation on Primates: A Biogeographical Analysis. In: SL Gursky, KAI Nekaris, editors. Primate Anti-Predator Strategies: Springer. p. 27-59.

Hashiguchi Y, Nishida M. 2007. Evolution of Trace Amine-Associated Receptor (TAAR) Gene Family in Vertebrates: Lineage-specific Expansions and Degradations of a Second Class of Vertebrate Chemosensory Receptors Expressed in the Olfactory Epithelium. *Mol Biol Evol*. 24:2099–2107.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*. 17:754-755.

Jeffery WR, Strickler AG, Yamamoto Y. 2003. To See or Not to See: Evolution of Eye Degeneration in Mexican Blind Cavefish. *Integrative and Comparative Biology*. 43:531-541.

Kall L, Krogh A, Sonnhammer ELL. 2007. Advantages of combined transmembrane topology and signal peptide prediction-the Phobius web server. *Nucleic Acids Res*. 35:W429-432.

Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol*. 30:772-780.

Kleinau G, Pratzka J, Nurnberg D, Gruters A, Fuhrer-Sakel D, Krude H, Kohrle J, Schoneberg T, Biebermann H. 2011. Differential modulation of Beta-adrenergic receptor signaling by trace amine-associated receptor 1 agonists. *PLoS ONE*. 6:e27073.

Lerer B, Segman RH, Hamdan A, Kanyas K, Karni O, Kohn Y, Korner M, Lanktree M, Kaadan M, Turetsky N, Yakir A, Kerem B, Macciardi F. 2003. Genome scan of Arab Israeli families maps a schizophrenia susceptibility gene to chromosome 6q23 and supports a locus at chromosome 10q24. *Mol Psychiatry*. 8:488-498.

Levinson DF, Holmans P, Straub RE, Owen MJ, Wildenauer DB, Gejman PV, Pulver AE, Laurent C, Kendler KS, Walsh D, Norton N, Williams NM, Schwab SG, Lerer B, Mowry BJ, Sanders AR, Antonarakis SE, Blouin J-L, DeLeuze J-F, Mallet J. 2000.

Multicenter Linkage Study of Schizophrenia Candidate Regions on Chromosomes 5q, 6q, 10p, and 13q: Schizophrenia Linkage Collaborative Group III. *Am J Hum Genet*. 67:652-663.

Lewis CM, Levinson DF, Wise LH, DeLisi LE, Straub RE, Hovatta I, Williams NM, Schwab SG, Pulver AE, Faraone SV, Brzustowicz LM, Kaufmann CA, Garver DL, Gurling HMD, Lindholm E, Coon H, Moises HW, Byerley W, Shaw SH, Mesen A, Sherrington R, O'Neill FA, Walsh D, Kendler KS, Ekelund J, Paunio T, Lönnqvist J, Peltonen L, O'Donovan MC, Owen MJ, Wildenauer DB, Maier W, Nestadt G, Blouin J-L, Antonarakis SE, Mowry BJ, Silverman JM, Crowe RR, Cloninger CR, Tsuang MT, Malaspina D, Harkavy-Friedman JM, Svrakic DM, Bassett AS, Holcomb J, Kalsi G, McQuillin A, Brynjolfson J, Sigmundsson T, Petursson H, Jazin E, Zoëga T, Helgason T. 2003. Genome Scan Meta-Analysis of Schizophrenia and Bipolar Disorder, Part II: Schizophrenia. *Am J Hum Genet*. 73:34-48.

Li W-H, Gojobori T, Nei M. 1981. Pseudogenes as a paradigm of neutral evolution. *Nature*. 292:237–239.

Lindholm E, Ekholm B, Shaw S, Jalonen P, Johansson G, Pettersson U, Sherrington R, Adolfsson R, Jazin E. 2001. A Schizophrenia-Susceptibility Locus at 6q25, in One of the World's Largest Reported Pedigrees. *Am J Hum Genet*. 69:96-105.

Maguire JJ, Parker WAE, Foord SM, Bonner TI, Neubig RR, Davenport AP. 2009. International Union of Pharmacology. LXXII. Recommendations for Trace Amine Receptor Nomenclature. *Pharmacol Rev*. 61:1-8.

Misceo D, Capozzi O, Roberto R, Dell'Oglio MP, Rocchi M, Stanyon R, Archidiacono N. 2008. Tracking the complex flow of chromosome rearrangements from the Hominoidea Ancestor to extant Hylobates and Nomascus Gibbons by high-resolution synteny mapping. *Genome Res*. 18:1530-1537.

Murphy WJ, Pringle TH, Crider TA, Springer MS, Miller W. 2007. Using genomic data to unravel the root of the placental mammal phylogeny. *Genome Res*. 17:413-421.

Olson MV. 1999. When Less Is More: Gene Loss as an Engine of Evolutionary Change. *Am J Hum Genet*. 64:18-23.

Olson MV, Varki A. 2003. Sequencing the chimpanzee genome: insights into human evolution and disease. *Nat Rev Genet*. 4:20-28.

Pae C-U, Drago A, Kim J-J, Patkar AA, Jun T-Y, De Ronchi D, Serretti A. 2010. TAAR6 variations possibly associated with antidepressant response and suicidal behavior. *Psychiatry Research*. 180:20-24.

Perelman P, Johnson WE, Roos C, Seuánez HN, Horvath JE, Moreira MAM, Kessing B, Pontius J, Roelke M, Rumpler Y, Schneider MPC, Silva A, O'Brien SJ, Pecon-Slattery J. 2011. A Molecular Phylogeny of Living Primates. *PLoS Genet*. 7:e1001342.

Prasad AB, Allard MW, Green ED, Program NCS. 2008. Confirming the phylogeny of mammals by use of large comparative sequence data sets. *Mol Biol Evol*. 25:1795-1808.

Premont RT, Gainetdinov RR, Caron MG. 2001. Following the trace of elusive amines. *Proc Natl Acad Sci USA*. 98:9474-9475.

Prufer K, Munch K, Hellmann I, Akagi K, Miller JR, Walenz B, Koren S, Sutton G, Kodira C, Winer R, Knight JR, Mullikin JC, Meader SJ, Ponting CP, Lunter G, Higashino S, Hobolth A, Dutheil J, Karakoc E, Alkan C, Sajjadian S, Catacchio CR, Ventura M, Marques-Bonet T, Eichler EE, Andre C, Atencia R, Mugisha L, Junhold J, Patterson N, Siebauer M, Good JM, Fischer A, Ptak SE, Lachmann M, Symer DE, Mailund T, Schierup MH, Andres AM, Kelso J, Paabo S. 2012. The bonobo genome compared with the chimpanzee and human genomes. *Nature*. 486:527-531.

Roberto R, Capozzi O, Wilson RK, Mardis ER, Lomiento M, Tuzun E, Cheng Z, Mootnick AR, Archidiacono N, Rocchi M, Eichler EE. 2007. Molecular refinement of gibbon genome rearrangements. *Genome Res*. 17:249-257.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*. 4:406-425.

Serretti A, Pae C-U, Chiesa A, Mandelli L, De Ronchi D. 2009. Influence of TAAR6 polymorphisms on response to aripiprazole. *Prog Neuropsychopharmacol Biol Psychiatry*. 33:822-826.

Smith T, Rossie J. 2006. Primate Olfaction: Anatomy and Evolution. In: WJ Brewer, D Castle, C Pantelis, editors. Olfaction and the Brain: Cambridge University Press. p. 135-166.

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

Stäubert C, Böselt I, Bohnekamp J, Römpler H, Enard W, Schöneberg T. 2010. Structural and Functional Evolution of the Trace Amine-Associated Receptors TAAR3, TAAR4 and TAAR5 in Primates. *PLoS ONE*. 5:e11133.

Tusnady GE, Simon I. 2001. The HMMTOP transmembrane topology prediction server. *Bioinformatics*. 17:849-850.

Vladimirov V, Thiselton DL, Kuo PH, McClay J, Fanous A, Wormley B, Vittum J, Ribble R, Moher B, van den Oord E, O'Neill FA, Walsh D, Kendler KS, Riley BP. 2007. A region of 35 kb containing the trace amine associate receptor 6 (TAAR6) gene is associated with schizophrenia in the Irish study of high-density schizophrenia families. *Mol Psychiatry*. 12:842-853.

Wang X, Grus WE, Zhang J. 2006. Gene Losses during Human Origins. *PLoS Biol*. 4:e52.

Warne T, Edwards PC, Leslie AG, Tate CG. 2012. Crystal Structures of a Stabilized β_1-Adrenoceptor Bound to the Biased Agonists Bucindolol and Carvedilol. *Structure*. 20:841-849.

Warne T, Moukhametzianov R, Baker JG, Nehme R, Edwards PC, Leslie AGW, Schertler GFX, Tate CG. 2011. The structural basis for agonist and partial agonist action on a β_1-adrenergic receptor. *Nature*. 469:241-244.

Wolinsky T, Swanson C, Smith K, Zhong H, Borowsky B, Seeman P, Branchek T, Gerald C. 2007. The Trace Amine 1 receptor knockout mouse: an animal model with relevance to schizophrenia. *Genes Brain Behav*. 6:628 - 639.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol*. 39:306-314.

Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol*. 15:568-573.

Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol*. 24:1586-1591.

Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol*. 19:908-917.

Yang Z, Wong WSW, Nielsen R. 2005. Bayes Empirical Bayes Inference of Amino Acid Sites Under Positive Selection. *Mol Biol Evol*. 22:1107-1118.

Zhang J. 2008. Positive selection, not negative selection, in the pseudogenization of rcsA in Yersinia pestis. *Proc Natl Acad Sci USA*. 105:E69.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an Improved Branch-Site Likelihood Method for Detecting Positive Selection at the Molecular Level. *Mol Biol Evol*. 22:2472-2479.

Table 3.1. The number of TAAR genes identified in the 13 animal genomes.

Group/	[a]Total	Number of TAAR subfamily genes[b]								
Species name	number	T1	T2	T3	T4	T5	T6	T7	T8	T9
[Simiiformes]										
Homo sapiens	6 (3)	1	1	0 (1)	0 (1)	1	1	0 (1)	1	1
Pan troglodytes	3 (6)	1	0 (1)	0 (1)	0 (1)	1	1	0 (1)	0 (1)	0 (1)
Pan paniscus	2 (7)	1	0 (1)	0 (1)	0 (1)	1	0 (1)	0 (1)	0 (1)	0 (1)
Gorilla gorilla	3 (7)	1	1	0 (1)	0 (1)	1	0 (1)	0 (1)	0 (2)	0 (1)
Pongo pygmaeus abelii	4 (6)	1	0 (1)	0 (2)	1	1	0 (1)	0 (1)	0 (1)	1
Nomascus leucogenys	1 (3)	1	0	0	0 (1)	0 (1)	0	0	0	0 (1)
Macaca mulatta	6 (3)	1	1	1	1	1	1	0 (1)	0 (1)	0 (1)
Papio hamadryas	5 [1] (2)	1	[1]	1	1	1	1	0	0 (1)	0 (1)
[Simiiformes]										
Callithrix jacchus	2 (6)	1	0 (1)	0 (1)	0 (1)	1	0 (1)	0	0 (1)	0 (1)
[Tarsiiformes]										
Tarsius syrichta	2 [1] (4)	0	[1]	1	1	0 (1)	0 (1)	0	0 (1)	0 (1)
[Lemuriformes]										
Microcebus murinus	6 [1] (1)	1	[1]	1	1	0 (1)	1	0	1	1
[Lorisiformes]										
Otolemur garnettii	7 [1] (0)	1	[1]	1	1	1	1	0	1	1
[Scandentia]										
Tupaia belangeri	8 [4] (5)	0 (1)	[1]	[1]	1	[1]	2 (2)	0	5 (2)	[1]
[Rodentia]										
Mus musculus	15 (1)	1	1	1	1	1	1	5 (1)	3	1
Rattus norvegicus	17 (2)	1	1	1	1	1	1	7 (2)	3	1
[Artiodactyla]										
Bos taurus	21 (8)	1	1	1	1	1	5 (2)	7 (4)	3 (2)	1

[a]TAAR gene candidates are divided into three categories: intact, incomplete, and pseudogenes. The first number shown is that of "intact" genes, which contain full-length open reading frames with seven complete transmembrane regions. The numbers of the “incomplete” genes due to contig ends and the pseudogenes due to premature stop codons or frame-shifting insertions or deletions are given in square brackets and in parentheses, respectively.

[b]T1-T9 indicate TAAR1 to TAAR9.

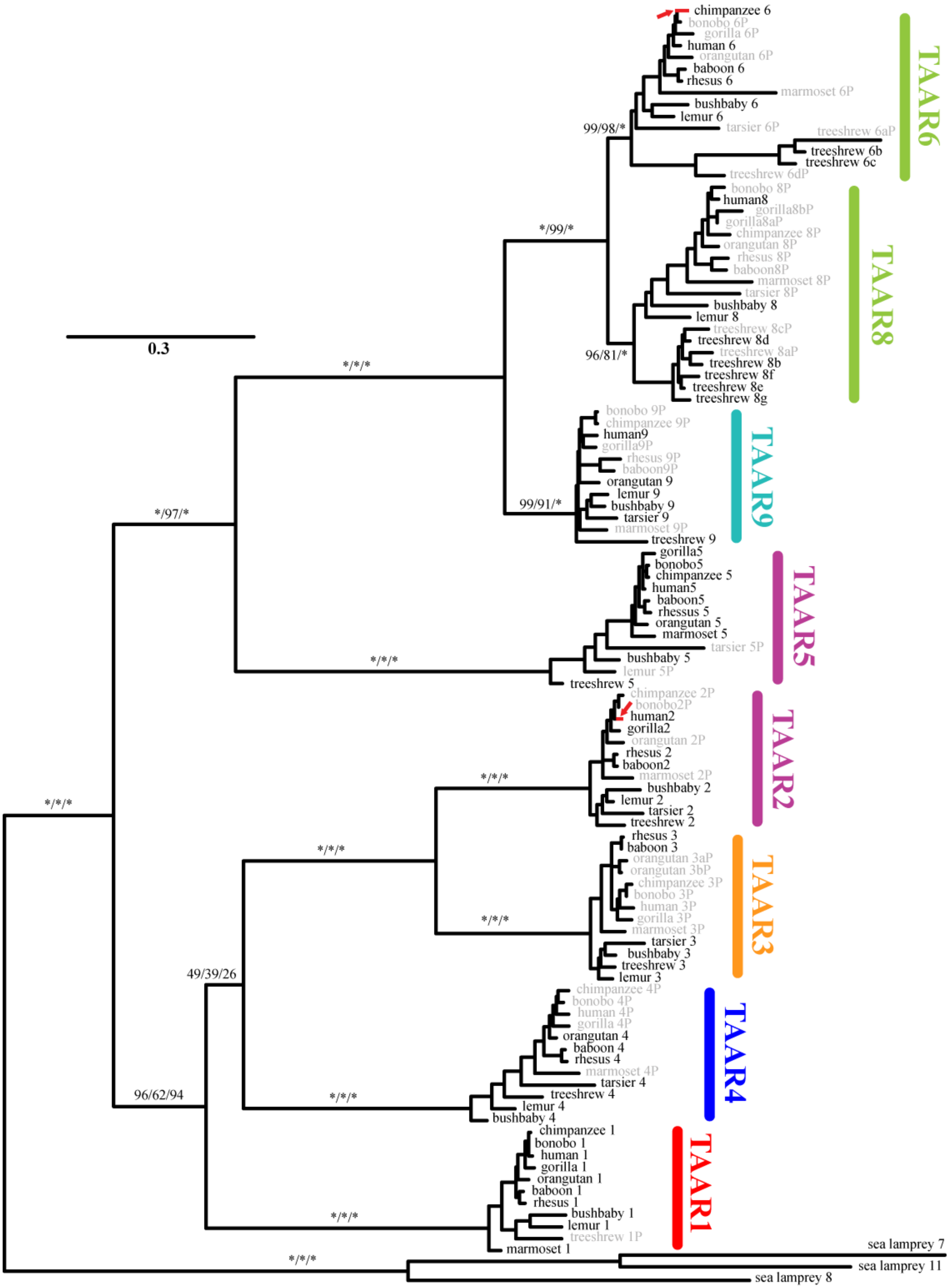
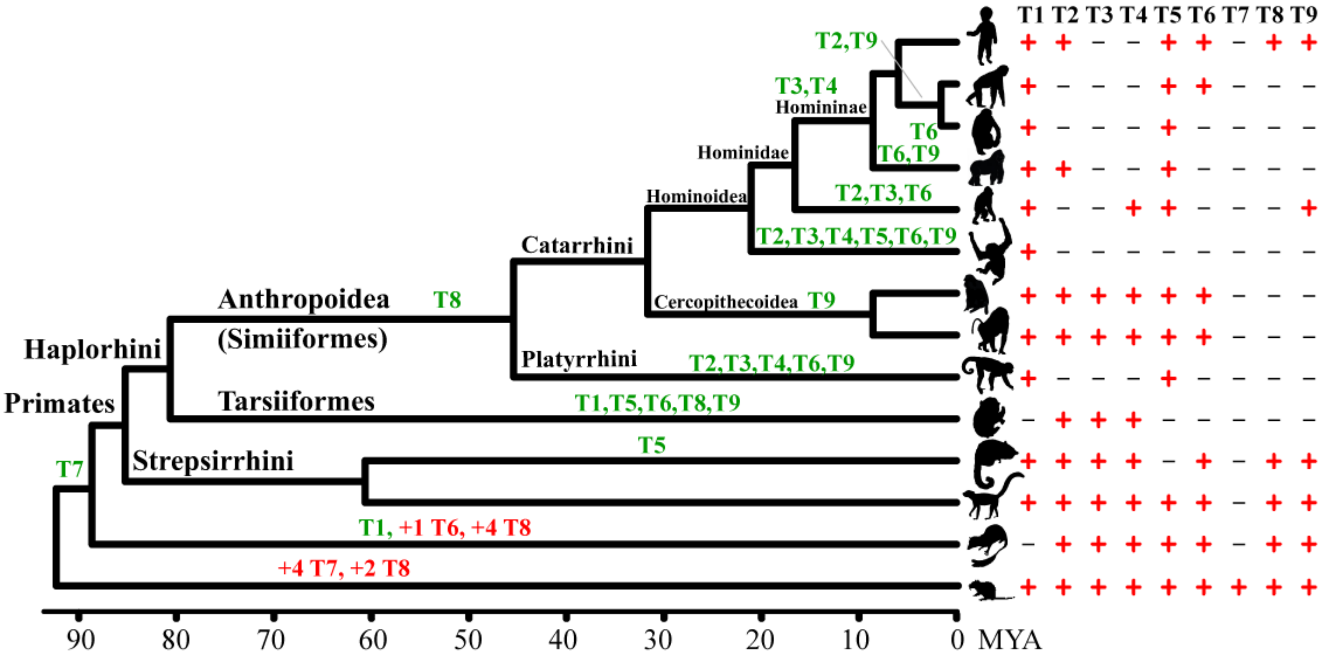
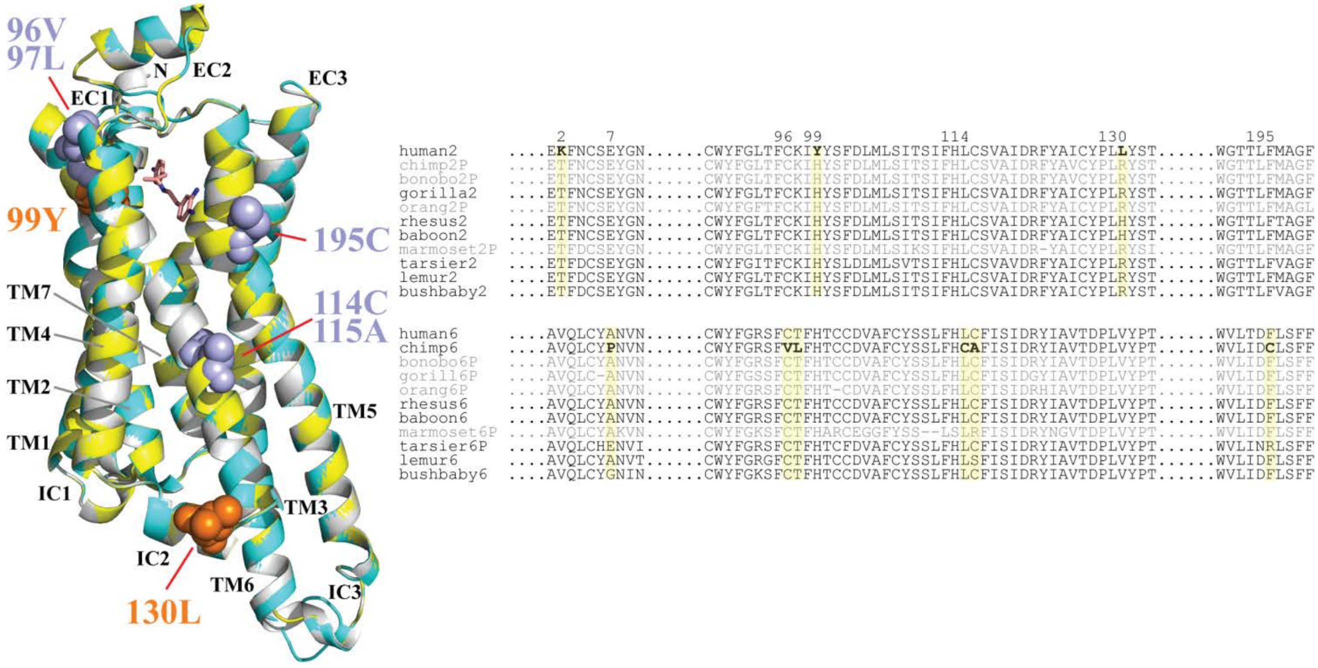

chimpanzee 6
bonobo 6P
gorilla 6P
human 6
orangutan 6P
baboon 6
rhesus 6
marmoset 6P
bushbaby 6
lemur 6
tarsier 6P
treeshrew 6aP
treeshrew 6b
treeshrew 6c
treeshrew 6dP
99/98/*
TAAR6
bonobo 8P
human8
gorilla8bP
gorilla8aP
chimpanzee 8P
orangutan 8P
rhesus 8P
baboon8P
marmoset 8P
tarsier 8P
bushbaby 8
lemur 8
treeshrew 8cP
treeshrew 8d
treeshrew 8aP
treeshrew 8b
treeshrew 8f
treeshrew 8e
treeshrew 8g
/99/
96/81/*
TAAR8
0.3
//*
bonobo 9P
chimpanzee 9P
human9
gorilla9P
rhesus 9P
baboon9P
orangutan 9
lemur 9
bushbaby 9
tarsier 9
marmoset 9P
treeshrew 9
99/91/*
TAAR9
/97/
gorilla5
bonobo5
chimpanzee 5
human5
baboon5
rhessus 5
orangutan 5
marmoset 5
tarsier 5P
bushbaby 5
lemur 5P
treeshrew 5
//*
TAAR5
chimpanzee 2P
bonobo2P
human2
gorilla2
orangutan 2P
rhesus 2
baboon2
marmoset 2P
bushbaby 2
lemur 2
tarsier 2
treeshrew 2
//*
TAAR2
//*
rhesus 3
baboon 3
orangutan 3aP
orangutan 3bP
chimpanzee 3P
bonobo 3P
human 3P
gorilla 3P
marmoset 3P
tarsier 3
bushbaby 3
treeshrew 3
lemur 3
//*
TAAR3
//*
49/39/26
chimpanzee 4P
bonobo 4P
human 4P
gorilla 4P
orangutan 4
baboon 4
rhesus 4
marmoset 4P
tarsier 4
treeshrew 4
lemur 4
bushbaby 4
96/62/94
//*
TAAR4
chimpanzee 1
bonobo 1
human 1
gorilla 1
orangutan 1
baboon 1
rhesus 1
bushbaby 1
lemur 1
treeshrew 1P
marmoset 1
//*
TAAR1
sea lamprey 7
sea lamprey 11
sea lamprey 8
//*

Figure 3.1. Evolutionary relationships of 116 TAARs from twelve primates and northern treeshrew. The phylogenetic tree is reconstructed by the maximum-likelihood method. Three sea lamprey TAAR-like proteins were used as outgroups. The numbers at internal branches show the bootstrap support values (%) for the neighbor-joining and maximum-likelihood phylogenies and the posterior probability (%) for the Bayesian phylogeny in this order, with asterisks indicating scores of 100%. Supporting values are shown only for the major internal branches. Gray-colored names indicate pseudogenes. Two red-colored branches and arrows indicate those identified to be under positive selection by the PAML branch-site models (see supplementary Table S3.5).



T1	T2	T3	T4	T5	T6	T7	T8	T9
+	+	–	–	+	+	–	+	+
+	–	–	–	+	+	–	–	–
+	–	–	–	+	–	–	–	–
+	+	–	–	+	–	–	–	–
+	–	–	+	+	–	–	–	+
+	–	–	–	–	–	–	–	–
+	+	+	+	+	+	–	–	–
+	+	+	+	+	+	–	–	–
+	–	–	–	+	–	–	–	–
–	+	+	+	–	–	–	–	–
+	+	+	+	–	+	–	+	+
+	+	+	+	+	+	–	+	+
–	+	+	+	+	+	–	+	+
+	+	+	+	+	+	+	+	+

Figure 3.2. TAAR gene gains and losses in primate genomes. TAAR gene gain (red color) and gene loss (including pseudogenization) (green color) events are shown along the branches. T1-T9 indicate TAAR1 to TAAR9. A current consensus of primate phylogenies with their approximate divergence times (million years ago; MYA) was obtained from Perelman et al. (2011). The outgroup used is a mouse. The species are listed in the same order as shown in Table 3.1 (from the top, except rat and cow).



```
                  2    7                96 99             114             130            195
human2       ....EKFNCSEYGN......CWYFGLTFCKIYYSFDLMLSITSIFHLCSVAIDRFYAICYPLLYST......WGTTLFMAGF
chimp2P      ....ETFNCSEYGN......CWYFGLTFCKIHYSFDLMLSITSIFHLCSVAIDRFYAVCYPLRYST......WGTTLFMAGF
bonobo2P     ....ETFNCSEYGN......CWYFGLTFCKIHYSFDLMLSITSIFHLCSVAIDRFYAVCYPLRYST......WGTTLFMAGF
gorilla2     ....ETFNCSEYGN......CWYFGLTFCKIHYSFDLMLSITSIFHLCSVAIDRFYAICYPLRYST......WGTTLFMAGF
orang2P      ....ETFNCSEYGN......CWYFGFTFCKIHYSFDLMLSITSIFHLCSVAIDRFYAICYPLRYST......WGTTLFMAGL
rhesus2      ....ETFNCSEYGN......CWYFGLTFCKIHYSFDLMLSITSIFHLCSVAIDRFYAICYPLHYST......WGTTLFTAGF
baboon2      ....ETFNCSEYGN......CWYFGLTFCKIHYSFDLMLSITSIFHLCSVAIDRFYAICYPLHYST......WGTTLFMAGF
marmoset2P   ....ETFDCSEYGN......CWYFGLTFCKIHYSFDLMLSIKSIFHLCSVAIDR-YAICYPLRYSI......WGTTLFMAGF
tarsier2     ....ETFDCSEYGN......CWYFGITFCKIHYSLDLMLSVTSIFHLCSVAVDRFYAICYPLRYST......WGTTLFVAGF
lemur2       ....ETFDCSEYGN......CWYFGLTFCKIHYSFDLMLSITSIFHLCSVAIDRFYAICYPLRYST......WGTTLFMAGF
bushbaby2    ....ETFDCSEYGN......CWYFGLTFCKIHYSFDLMLSITSIFHLCSVAIDRFYAICYPLRYST......WGTTLFVAGF

human6       ....AVQLCYANVN......CWYFGRSFCTFHTCCDVAFCYSSLFHLCFISIDRYIAVTDPLVYPT......WVLTDFLSFF
chimp6       ....AVQLCYPNVN......CWYFGRSFVLFHTCCDVAFCYSSLFHCAFISIDRYIAVTDPLVYPT......WVLIDCLSFF
bonobo6P     ....AVQLCYANVN......CWYFGRSFCTFHTCCDVAFCYSSLFHLCFISIDRYIAVTDPLVYPT......WVLIDFLSFF
gorill6P     ....AVQLC-ANVN......CWYFGSSFCTFHTCCDVAFCYSSLFHLCFISIDGYIAVTDPLVYPT......WVLIDFLSFF
orang6P      ....AVQLCYANVN......CWYFGRSFCTFHT-CDVAFCYSSLFHLCFISIDRHIAVTDPLVYPT......WVLIDFLSFF
rhesus6      ....AVQLCYANVN......CWYFGRSFCTFHTCCDVAFCYSSLFHLCFISIDRYIAVTDPLVYPT......WVLIDFLSFF
baboon6      ....AVQLCYANVN......CWYFGRSFCTFHTCCDVAFCYSSLFHLCFISIDRYIAVTDPLVYPT......WVLIDFLSFF
marmoset6P   ....AVQLCYAKVN......CWYFGKSFCTFHARCEGGFYSS--LSLRFISIDRYNGVTDPLVYPT......WVLIDFLSFF
tarsier6P    ....AVQLCHENVI......CWYFGRSFCTFHTCFDVAFCYSSLFHLCFISIDRYIAVTDPLVYPT......WVLINRLSFF
lemur6       ....AVQLCYANVT......CWYFGRGFCTFHTCCDVAFCYSSLFHLSFISIDRYIAVTDPLVYPT......WVLIDFLSFF
bushbaby6    ....AVQLCYGNIN......CWYFGKSFCTFHTCCDVAFCYSSLFHLCFISIDRYIAVTDPLVYPT......WVLIDFLSFF
```

Figure 3.3. The 3D-structural model and partial sequence alignments of TAAR2 and TAAR6 proteins. (a) The 3D-structural model of the human TAAR2 (yellow) and chimpanzee TAAR6 protein (cyan) superimposed with the turkey β1-adrenergic receptor (β1AR, gray). The ligand of the β_1AR, dobutamine, is shown with the stick model. Positively selected sites are indicated by orange (human TAAR2) and dark cyan (chimpanzee TAAR6). Two positive selected sites (positions 2 and 7) are not shown due to the lack of 3D protein model. (b) The partial sequence alignment of primate TAAR2 and TAAR6. The nine residues predicted to be under positive selection are shown in boldfaces (indicated by yellow boxes). The pseudogenes are in grey-colored.

Chapter 4

Molecular Evolution of the Glycoside Hydrolase Gene Families in the Western Corn Rootworm (*Diabrotica virgifera virgifera*)

4.0 Abstract for Chapter 4

Cellulose is an important nutritional resource for a number of insect herbivores. Digestion of cellulose and other polysaccharides in plant-based diets requires several types of enzymes including a number of glycoside hydrolase (GH) families. In a previous study, we showed that a single GH45 gene is present in the midgut tissue of the western corn rootworm, *Diabrotica virgifera virgifera* (Coleoptera: Chrysomelidae). However, the presence of multiple enzymes was also suggested by the lack of a significant biological response when the expression of the gene was silenced by RNA interference. In order to clarify the entire repertoire of cellulose-degrading enzymes in *D. v. virgifera*, we performed next-generation sequencing and assembled transcriptomes from the tissue of three different developmental stages (eggs, neonates, and third instar larvae). Results of this study revealed the presence of thirty three genes that potentially encode GH enzymes belonging to six families (GH45, GH48, GH28, GH16, GH31, and GH5). *D. v. virgifera* possesses the largest and second largest numbers of GH45 and GH28 genes, respectively, among insects where these genes have been identified. Three GH family genes (GH45, GH48, and GH28) are found almost exclusively in two coleopteran superfamilies (Chrysomeloidea and Curculionoidea) among insects, indicating the possibility of their acquisitions by horizontal gene transfer rather than vertical transmission from the ancestral insect species. Acquisition of GH genes by horizontal gene transfers and subsequent lineage-specific GH gene expansion appear to have played important roles for phytophagous beetles in specializing on particular groups of host plants and in the case of *D. v. virgifera*, its close association with maize.

4.1 Background

The western corn rootworm, *Diabrotica virgifera virgifera* (Chrysomelidae, Coleoptera), is the most serious and economically important beetle pest of maize (*Zea mays* L.) in the U.S. Corn Belt in terms of direct crop losses and the cost of control measures including synthetic insecticides (Levine and Oloumi-Sadeghi 1991; Sappington et al. 2006). An economic analysis has indicated that costs of control and yield loss associated with *D. v. virgifera* damage exceed $1 billion annually and a recent estimate would likely be larger (Metcalf 1986; Dun et al. 2010). *D. v. virgifera* larvae are primarily responsible for the damage to corn, which obtain nourishment and cause the majority of economic damage via root feeding. Larval feeding also weakens the structural support of the root, thus reducing plant stability and grain yield (Sutter et al. 1990; Spike and Tollefson 1991; Gray and Steffey 1998; Urias-Lopez and Meinke 2001). Damage to corn roots as a result of larval feeding can further cause physiological stress to the plants leading to reduced yield (Riedell 1990; Godfrey et al. 1993b; Godfrey et al. 1993a; Hou et al. 1997).

The glycoside hydrolases (GH; EC 3.2.1.-) gene (also known as glycosidases or glycosyl hydrolases) families are a widespread group of enzymes to catalyze hydrolysis of the glycoside linkages. GH genes are classified into 132 families and 14 clans according to their amino-acid sequence similarities and their folding patterns based on the Carbohydrate-Active enZymes Database (CAZy, http://www.cazy.org) (Cantarel et al. 2009). Three classes of cellulases (endoglucanases: EC 3.2.1.4, cellobiohydrolases: EC 3.2.1.74 and

3.2.1.91, and β-glucosidases: EC 3.2.1.21) are placed into five GH-clans (GH-A, GH-B, GH-C, GH-K, and GH-M) although some have not been classified (Cantarel et al. 2009). Genes encoding cellulases and other GH families have been identified from a number of phytophagous coleopterans belonging to the superfamilies Chrysomeloidea, which includes long-horned beetles and leaf beetles, and Curculionoidea (weevils) (Calderón-Cortés et al. 2010; Pauchet et al. 2010; Watanabe and Tokuda 2010; Pauchet and Heckel 2013). A β-1,4-endoglucanase (EC. 3.2.1.4) gene belonging to the GH family 9 was also isolated and characterized from the red flour beetle *Tribolium castaneum* (Coleoptera: Tenebrionidae) (Willis et al. 2011).

Valencia et al. (2013) cloned and characterized a novel β-1,4-endoglucanase gene (*DvvENGaseI*, JQ755253) belonging to the GH family 45 from the western corn rootworm *Diabrotica virgifera virgifera* (Coleoptera: Chrysomelidae), an important insect pest of maize (*Zea mays* L.) in the United States (Siegfried et al. 2005; Valencia et al. 2013). They showed that suppression of *DvvENGaseI* expression by RNA interference resulted in only slight developmental delays suggesting that this gene might be a part of the larger system of cellulose degrading enzymes (Valencia et al. 2013). The goal of this study is focused on the exploration of genetic diversity among GH family genes in *D. v. virgifera*, especially focusing on its larval stages. In order to identify the diversity of GH family genes encoding cellulase and other plant cell wall degrading enzymes expressed in *D. v. virgifera* larvae, we sequenced the transcriptomes covering three different developmental stages (eggs, neonates, and midgut from third instar larvae) using next-generation technologies. I identified six types of GH family genes that encode three types of cellulases (GH45, GH48, and GH5) as well as a pectinase (GH28), an endo-1,3-β-glucanase (GH16), and an α-glucosidase (GH31). I found large numbers of GH45 and GH28 genes from the *D. v. virgifera* transcriptomes, one of the largest and second largest so far known among coleopteran species studied. The analyses also suggested multiple horizontal transfer events of GH45, GH48, and GH28 genes from bacteria and fungi to the common ancestor of chrysomelid and curculionid beetles, as well as to other herbivorous insects. Acquisition and subsequent expansion of GH gene copies in phytophagous beetles may have been adaptive and have played important roles for them to specialize in feeding on particular host plants.

4.2 Results and Discussion

4.2.1 Sequencing and *de novo* assembly of *D. v. virgifera* transcriptome.

Using Illumina paired-end as well as 454 Titanium sequencing technologies, in total ~700 gigabases were sequenced from cDNA prepared from eggs (15,162,017 Illumina paired-end reads after filtering), neonates (721,697,288 Illumina paired-end reads after filtering), and midguts of third instar larvae (44,852,488 Illumina paired-end reads and 415,742 Roche 454 reads, both after filtering) (see Supplementary Table S4.1 for details). *De novo* transcriptome assembly was performed using Trinity (Grabherr et al. 2011) for each of three samples as well as for the pooled dataset (see Materials and Methods and Supplementary Tables S4.1, S4.2, and S4.3 for the comparative analysis of assembly programs and other details). The *D. v. virgifera* transcriptome assembled from the pooled dataset included 163,871 contigs (the average length: 914 bp) (Table 4.1).

4.2.2 Identification of GH family genes from *D. v. virgifera* transcriptomes.

A total of thirty three potential genes belonging to six different GH families (GH45, GH48, GH28, GH16, GH31, and GH5) were identified from our *D. v. virgifera* transcriptome. In Figure 4.1, numbers of these GH family genes in *D. v. virgifera* are compared with those found in other coleopteran species. While the enzymes encoded by GH45, GH48, and GH5 are known to have β-1,4-endoglucanase (EC. 3.2.1.4) activity, GH28 gene encodes a pectolytic enzyme, α-1-4-polygalactunorase (EC 3.2.1.15) (Cantarel et al. 2009), GH16 encodes an endo-1,3-β-glucanase (EC 3.2.1.39) (Genta et al. 2009; Bragatto et al. 2010), and GH31 an α-glucosidase (EC 3.2.1.20) (Wheeler et al. 2013).

4.2.3 GH45

Eleven GH45 family genes were identified from the *D. v. virgifera* transcriptome (Supplementary Figure S4.1). Ten of these sequences covered the entire coding region. The partial GH48-2 sequence was also confirmed in the draft *D. v. virgifera* genome. The average length of the complete GH45 coding sequences is 717 bp (ranging from 615 to 741 bp, coding from 205 to 247 amino acids). Four of them (GH45-1, GH45-4, GH45-7, and GH45-10) were highly expressed (> 100 reads per kilobase of per million mapped reads or RPKM) especially in the third-instar larval midgut and neonate samples but not expressed in the egg samples (Supplementary Table S4.4). We have previously identified GH45-7 as *DvvENGaseI* (JQ755253) (Valencia et al. 2013). This gene exhibits the highest expression among the eleven GH45 genes and also the highest among all GH genes identified in the present study (Supplementary Table S4.4).

GH45 genes have been described from a number of coleopteran species belonging to the suborder Polyphaga (*e.g.*, (Girard and Jouanin 1999) (Eigenheer et al. 2003) (Lee et al. 2004), and (Calderón-Cortés et al. 2010)). Similarity searches against the NCBI non-redundant protein database as well as ten insect genomes confirmed that within insects, GH45 genes are found only in two polyphagan coleopteran superfamilies, Chrysomeloidea and Curculionoidea. As shown in Figure 4.1, multiple GH45 genes have been identified in some species, and based on available sequences, *D. v. virgifera* appears to have the largest number of GH45 genes (11 genes) among coleopteran species, and probably among any known invertebrates where this gene has been identified.

In addition to these coleopteran GH45 sequences, a sequence similar to GH45 has been identified from the springtail *Cryptopygus antarcticus* (ACV50414.1, described also in (Calderón-Cortés et al. 2010)), which belongs to one of the basal hexapodan orders, Collembola (Gao et al. 2008). Another sequence similar to GH45 was also reported from the water bear *Hypsibius dujardini* (phylum Tardigrada, a sister group of arthropods) (CD449425.1, mentioned also in Davison and Blaxter 2005). GH45 genes have also been reported among various metazoans from protists (Li et al. 2003) to plant-parasitic nematodes (Smant et al. 1998) and mollusks (Xu et al. 2001; Harada et al. 2004). In order to understand the evolutionary process that has led to the diversity of coleopteran GH45 genes, a maximum-likelihood phylogeny was reconstructed including GH45 proteins from eleven coleopteran species as well as other metazoans mentioned above, fungi, and bacteria (Figure 4.2 and Supplementary Figure S4.2). Our phylogenetic analysis suggests that all coleopteran GH45 genes are monophyletic although the support was weak (≤ 66% bootstrap supports). Several species-specific gene duplications were found in coleopteran species (shown with blue branches in Figure 4.2). While all bacterial GH45 proteins, except for *Myxococcus stipitatus* sequence, formed a monophyletic group, relationships among fungal and metazoan sequences were unresolved. Although the exact origins are not clear, we conclude that multiple horizontal gene transfer (HGT) events of GH45 genes likely happened particularly to the insect, springtail, and water bear lineages.

Pauchet et al. (2010) showed that a clade of Curculionoidea GH45 proteins (Group 1 in Figure 4.2) is the only one that utilizes Glu rather than Asp as a putative proton donor position. In all but one *D. v. virgifera* GH45 proteins, this position is also conserved with Asp (GH45-9 has Val; Supplementary Figure S4.2). I also found some varied residues at the proton donor sites including Asn in *H. dujardini* (Tardigrada), Thr in *Leptosphaeria maculans* (fungus), Ser in *Alternaria alternate* (fungus), and Glu in *Myxococcus stipitatus* (bacterium).

4.2.4 GH48

I identified three GH48 genes from *D. v. virgifera*: two complete (1,926 bp, 642 amino acids) and one partial (374 bp, 124 amino acids) (Supplementary Figure S4.3). This partial GH48 gene sequence (GH48-2) was also confirmed in the draft *D. v. virgifera* genome. Similar to GH45 genes, GH48 genes have been identified from many polyphagan coleopterans especially from the two superfamilies (Chrysomeloidea and Curculionoidea) (Fujita et al. 2006; Pauchet et al. 2010; Keeling et al. 2012) (Figure 4.1). Consistent with the results obtained by Pauchet et al. (2010), the number of GH48 genes found in coleopterans was smaller than those of GH45 and GH28 genes.

Fujita et al. (2006) isolated two GH48 genes (active phase-associated proteins, APAP I and II; shown as Gatr GH48-1 and -2 in Figure 4.3) from a leaf beetle *Gastrophysa atrocyanea*. While neither glucanase nor cellobiohydrolase activity was detected with *G. atrocyanea* GH48-1, it exhibited chitinase activity. *G. atrocyanea* GH48-1 was shown to be necessary for diapause termination in adults (Fujita et al. 2006). Based on our phylogenetic analysis, *G. atrocyanea* GH48-1 was found to be closer to *D. v. virgifera* GH48-2 (Figure 4.3). However, the expression level of the *D. v. virgifera* GH48-2 was not confirmed from our egg and larval samples (Supplementary Table S4.4). While *D. v. virgifera* GH48-1 also had very low expression, GH48-3 was found to be expressed more in larvae than in eggs.

GH48 is one of the most common GH family genes in bacteria (Berger et al. 2007). Apart from their presence in bacteria and in coleopterans, this family has been reported from three fungal species (*Neocallimastix patriciarum*: AEX92722.1, *Piromyces equi*: AAN76735.1, and *Piromyces* sp.: AAN76734.1). None of the ten insect genomes had GH48 family genes. Figure 4.3 (and Supplementary Figure S4.4) shows the maximum-likelihood phylogeny of GH48 proteins from coleopterans as well as from fungi and bacteria. This disparate and limited distribution of GH48 genes in two related coleopteran superfamilies and in three fungal species but not in any other eukaryotes, clearly indicates at least two independent HGT events: one from bacteria to the ancestral coleopteran lineage before the divergence of the two coleopteran superfamilies and the other from bacteria to the ancestral lineage before the divergence of the three fungal species.

The three fungal GH48 sequences belong to the family Neocallimastigaceae (phylum Neocallimastigomycota). These fungi are isolated in the digestive tracts of ruminant and non-ruminant mammals and herbivorous reptiles (Ljungdahl 2008). Rumen fungi have been reported to obtain catalytic enzymes from bacterial sources by HGT events. For example, GH5 (endoglucanase, EC 3.2.1.4) and GH11 (xylanase, EC 3.2.1.8) genes found in *Orpinomyces joyonii* and *Orpinomyces* sp. (phylum Neocallimastigomycota) are considered to be bacterial origin (Garcia-Vallvé et al. 2000). GH5 genes in the rumen fungus, *Neocallimastix patriciarum*, have also been suggested to have originated from bacteria (*Streptococcus equinus* and *Ruminococcus albus*) (Hung et al. 2012). Therefore, although our similarity search and phylogenetic analysis did not show a clear relationship with any known bacterial species, the three rumen fungal GH48 genes are very likely to be another examples of HGT from bacteria.

4.2.5 GH28

GH28 genes encode polygalactunorase (pectinase, EC 3.2.1.15). Eleven intact and three partial GH28 sequences were identified in the *D. v. virgifera* transcriptome. The average length of the complete GH28 coding sequences was 1,027 bp (343 amino acids, ranging from 1,062 to 1,116 bp) (Supplementary Figure S4.5). Gene expression, especially in larvae, was confirmed from the majority of these eleven intact GH28 gene candidates (Supplementary Figure S4.4). For the three partial sequences (GH28-8, 10, and 14), although their expression was either very low or confirmed neither in eggs nor in larvae, these partial sequences were found in the draft genome. Multiple copies of GH28 genes have been found in a number of coleopteran species belonging to the two superfamilies (Chrysomeloidea and Curculionoidea) (Girard and Jouanin 1999; Pauchet et al. 2010). While the largest number of GH28 (19 functional genes) was found in mountain pine beetle (*D. ponderosae*) (Keeling et al. 2012; Keeling et al. 2013) and *D. v. virgifera* has the second largest number, 11 (and 3 possible pseudogenes), only two GH28 genes were found in banana root borer (*Cosmopolites sordidus*) (Figure 4.1). In addition to a large variation in the gene number, our phylogenetic analysis confirmed many species-specific GH28 gene duplications in coleopterans (Figure 4.4).

Pauchet et al. (2010) showed that GH28 genes can be divided into two clades. GH28 enzymes from *Callosobruchus maculatus* (bean beetle) are more closely related to bacterial GH28 enzymes and they form the subgroup B, while all other beetle GH28 enzymes are more closely related to fungal and plant bug (Hemiptera) enzymes forming the subgroup A (Pauchet et al. 2010). Although two plant bug species (*Lygus hesperus* and *Lygus lineolaris*, Hemiptera) were reported to have several GH28 genes (Allen and Mertens 2008; Celorio-Mancera et al. 2008), we failed to identify GH28 in the ten insect genomes including two from hemipterans *Rhodnius prolixus* (a blood-sucking bug) and *Acyrthosiphon pisum* (pea aphid). Among insects, in addition to the two plant bug species, GH28 genes were found only in two coleopteran superfamilies (Chrysomeloidea and Curculionoidea). Our phylogenetic analysis showed that these plant bugs as well as all coleopteran GH28 genes except for those of *C. maculatus* are nested within fungal GH28 cluster (Figure 4.4 and Supplementary Figure S4.6). Consistent with what Pauchet et al. (2010) indicated, seven

GH28 genes identified from *C. maculatus* clustered with GH28 genes from bacteria (all Gram-negative bacteria) (≤76% bootstrap supports). Therefore, GH28 genes currently found in coleopterans and plant bugs are most likely acquired by three independent HGT events: from a Gram-negative bacteria to *C. maculatus*, from a fungus to a hemiptera, and from a fungus to an ancestral coleopteran before the divergence of the two superfamilies.

4.2.6 GH16 family genes

I identified two GH16 genes in the *D. v. virgifera* transcriptome which exhibit full length of sequences (450 and 499 amino acids in GH16-1 and GH16-2) (Supplementary Figure S4.7). They were not highly expressed among any of the three libraries that were sequenced (Supplementary Figure S4.4). GH16 genes are widely found in insects (*e.g.*, Genta et al. 2009 and Pauchet et al. 2009) and have been reported in a springtail *C. antarcticus*, which is believed to have originated by HGT from bacteria (Song et al. 2010). Similarity searches further confirmed the wide distribution of GH16 within arthropods, fungi, and bacteria, but not in plants, nematodes, or protists. It has also been reported from a scallop, *Chlamys albidus* (AAZ04385.1) (Kovalchuk et al. 2009) as well as in a sea urchin, *Strongylocentrotus purpuratus* (XP_003725438.1) and in a tunicate *Ciona intestinalis* (XP_002126690.1). GH16 genes appear to be one of the most common GH family genes among invertebrates.

Figure 4.5 shows the phylogeny of GH16 protein sequences from four coleopteran species (*Tribolium castaneum*, *T. molitor*, *D. ponderosae*, and *D. v. virgifera*) and other insects as well as some other metazoans, fungi, and bacteria. Metazoan GH16 proteins are clearly clustered into two major groups. GH16 proteins in Group 2 have highly conserved catalytic nucleophile and proton donor sites (Glu for both, except for Tyr and Ser in *Daphnia pulex* GH16-3), while those in Group 1 have Gln and Phe (or Tyr) residues for those sites (Supplementary Figure S4.7). The two *D. v. virgifera* GH16 genes belong to Group 1. In this group, the enzymatic activity of the *T. molitor* GH16 (Q76D12.1) is known as endo-1,3-β-glucanase (EC 3.2.1.39) (Genta et al. 2009). Therefore, the same enzyme activity can be considered for the two *D. v. virgifera* GH16 gene products. Note that GH16 genes from *D. ponderosae* are divided into the two groups, and their catalytic site residues are also different (Supplementary Figure S4.7). In Group 2, the GH16 gene from *Spodoptera frugiperda* (Armyworm, Lepidoptera) (SLam, ABR28478.2) has been characterized and shown also as β-1,3-glucanase (EC 3.2.1.39), which hydrolyzes only β-1,3-glucan (Bragatto et al. 2010). Therefore, the amino acid changes found between these two GH16 groups in the catalytic sites do not seem to have affected the endoglucanase activity.

4.2.7 GH5 gene

A short sequence similar to part of the GH5 gene candidate was identified in the *D. v. virgifera* transcriptome (317 bp corresponding to 105 amino acids) (Supplementary Figure S4.8a). Among the 51 GH5 subfamilies (Aspeborg et al. 2012), coleopteran GH5 genes known so far belong to three subfamilies (2, 8, and 10) (Supplementary Figure 4.8b). Phylogenetically, the short *D. v. virgifera* GH5 sequence is closer to fungal GH5 sequences belonging to the subfamily 12 (Supplementary Figure S4.8b). We should, however, note that we failed to confirm the corresponding sequence in the draft *D. v. virgifera* genome.

Furthermore, the expression of this sequence was not confirmed with confidence (Supplementary Table S4.4). Therefore, we consider the existence of a GH5 gene in *D. v. virgifera* to be inconclusive.

4.2.8 Absence of GH9 gene

GH9 candidate sequence was not identified in the *D. v. virgifera* transcriptome. Among beetle species, GH9 is present in *T. castaneum* (Tenebrionoidea) (Willis et al. 2011). We also found a GH9 gene sequence from the transcriptome of the carabid beetle, *Pogonus chalceus* (salt marsh beetle, Caraboidea, Adephaga). However, GH9 gene appears to be absent among chrysomelids and curculionids. Because *P. chalceus* is placed as the most basal species in Coleoptera (Hunt et al. 2007) (Figure 4.1), GH9 was likely maintained in the common ancestor of Coleoptera and the lineage leading to Tenebrionoidea. GH9 must have been subsequently lost in the common ancestor of Chrysomeloidea and Curculionoidea.

We confirmed that three GH families (GH45, GH48, and GH28) were absent from the transcriptomes of *P. chalceus* (Van Belleghem et al. 2012) and the genome of *T. castaneum* (Tribolium Genome Sequencing Consortium 2008). The loss of GH9 and gain of GH45, GH48, and GH28 can be traced back at least to the common ancestor of chrysomelids and curculionids. Although GH9 and three enzymes (GH48, GH45, and GH28) do not share sequence similarities and have different 3D structural features, Watanabe and Tokuda (2010) suggested, for example, a possible convergent evolution in terms of enzymatic function based on the same substrate specificities (*e.g.*, β-1,4 linkages) with GH9 and GH45. GH9 and GH28 utilize the inverting glycosidase mechanism, which only allows polysaccharide hydrolysis (Sinnott 1990). Thus, their functional similarities may have allowed the laterally acquired genes to replace the role of the lost GH9.

4.2.9 GH31 family genes

Two GH31 genes, which encode an α-glucosidase (EC 3.2.1.20), were identified from the *D. v. virgifera* transcriptome (Supplementary Figure S4.9). The full length of coding sequences (1,236 bp encoding 411 amino acids, and 1,338 bp encoding 445 amino acids) were observed for both genes (Supplementary Figure S4.9). One of them (GH31-1) was highly expressed (> 200 RPKM) especially in the third-instar larval midgut (Supplementary Table S4.4). GH31 genes are found in a wide range of organisms, from bacteria, protists, fungi, vertebrates, to plants (Supplementary Figure S4.10) indicating these genes sharing an ancient common ancestor.

4.2.10 Gene expression.

When expression levels of GH gene candidates we identified from the *D. v. virgifera* transcriptomes were compared between egg and larval (neonate and third instar) samples, all but two (GH28-10 and GH16-2) were expressed significantly more in larval stages. We found that the majority of GH45, GH28, and GH31 genes are expressed more in the third-instar larval midgut samples compared to egg and neonate samples with some genes (GH45-7, GH28-6, and GH31-1) showing much higher expression (Supplementary Table S4.4). Polygalactunorase gene expression and enzyme activity has previously been reported from the gut of another corn rootworm species, *Diabrotica undecimpunctata howardi,* spotted cucumber beetle by Shen et al. (2003). Kirsch et al. (2012) examined the expression levels of several GH family genes including GH28 and GH45 genes in *P. cochleariae* larvae and adults. They are expressed more in the guts both in larvae and adults (Kirsch et al. 2012). *D. v. virgifera* GH28 and GH45 are also expressed more in gut samples than the egg sample. Polygalactunorases are known to loosen the primary cell wall and make cellulose-hemicellulose network more accessible to enzymatic digestion (Juge 2006). With its high number of GH45 and GH28 genes and their high expression in larval midgut tissue, *D. v. virgifera* may utilize β-1,4-endoglucanase as well as polygalactunorase activities in larval midgut to assist in the digestion of corn root cell walls as an initial degradation step.

4.2.11 Horizontal gene transfer of GH genes.

Our current study indicated that the three GH gene families (GH45, GH48, and GH28) are unique to the two coleopteran superfamilies (Chrysomeloidea and Curculionoidea) and generally absent from other insects except in plant bugs (GH28) and in a springtail (GH45). These results imply that these genes are likely not vertically inherited from the ancestral species but acquired by HGT events from bacteria and fungi to the common ancestor of Chrysomeloidea and Curculionoidea.

Recently, Acuña et al. (2012) identified a GH5 gene (*HhMAN1*) from the coffee berry borer (*Hypothenemus hampei,* Curculionoidea) and showed evidence of HGT from bacteria. Interestingly, this gene was not found in two other related species; *H. obscurus,* which is not a pest of coffee, and *Araecerus fasciculatus* (coffee bean weevil, Anthribidae, Coleoptera), which is a common pests of coffee but polyphagous on a number of different plant families (a generalist) in contrast to the monophagous or specialist *H. hampei* (Gladstone and Hruska 2003; Valentine 2005; Waller et al. 2007). Therefore, acquisition of *HhMAN1* from bacteria appears to have provided a rapidly acquired adaptation that enables hydrolysis of galactomannan, a nutrient source for *H. hampei* (Acuña et al. 2012). Other examples of possible HGTs include: rumen fungi GH5 and GH11 from rumen bacteria *Fibrobacter succinogenes* (Garcia-Vallvé et al. 2000), GH16 in *C. antarcticus* from bacteria (Song et al. 2010), GH31 in *Bombyx mori* from an *Enterococcus* bacteria (Wheeler et al. 2013), and GH11 in *P. cochleariae* from γ-proteobacteria (Pauchet and Heckel 2013). We also found evidence of several independent HGT events such as fungal GH48, bacterial GH45, and plant bug GH28. Although HGT events are often detected in prokaryotes (Dunning Hotopp 2011), GH families seem to be characterized by a high rate of HGT events in various animals. Such an acquisition may be important to these organisms' ability to adapt to novel niches.

4.3 Conclusion

I have identified six GH family genes from the transcriptomes of *D. v. virgifera*. It is likely that three GH families (GH45, GH48, and GH28) were obtained by HGT events in the common ancestor of Chrysomeloidea and Curculionoidea. Rapid birth-and-death processes have been also observed among these GH genes. A large number of GH enzymes owing to their species-specific duplications in *D. v. virgifera* could have contributed to the successful adaptation to its niche by providing more efficient hydrolyzation of corn cell walls.

4.4 Materials and Methods

4.4.1 Next generation sequencing.

Sample collection, preparation, and total RNA extraction were conducted in Blair Lab. The 454 pyrosequencing experiments of larval midgut samples were completed using Roche GS-FLX titanium sequencer at the Core for Applied Genomics and Ecology, University of Nebraska-Lincoln. The transcriptome sequencing for the egg and larval midgut samples with an insert size of 300 bp was done on Illumina Genome Analyzer II platform at the Center for Biotechnology, University of Nebraska-Lincoln. The neonate samples were sequenced with an insert size of 500 bp on Illumina HiSeq2000 system at the Durham Research Center, University of Nebraska Medical Center. In total, 16.6 gigabases (Gb) (read length 75 bp) of egg RNA, 33 Gb (read length 75 bp) of larval midgut RNA, and 662 Gb (read length 101 bp) of neonate RNA were sequenced.

4.4.2 *de novo* assembly of *D. v. virgifera* transcriptomes.

Because sequencing errors can cause difficulties for the assembly algorithm, we applied a stringent quality filter process. For 454 reads, the adapter and poly(A/T) sequences were trimmed using PRINSEQ (Schmieder and Edwards 2011). 454 reads that have abnormal read length (<50 bp or >1000 bp) or where the average quality was less than 20 were removed. The Illumina paired-end reads that did not have the minimum quality score (20 per base for egg and midgut samples or 30 per base for neonate samples) across the whole read were removed. Note that the quality scores of 20 (Q20) and 30 (Q30) correspond to 1% and 0.1% expected error rates, respectively. We also removed all Illumina reads that have any unknown nucleotide 'N'.

After the filtering processes, we performed *de novo* transcriptome assembly for each of three samples. We used four different short read assemblers: Newbler (ver. 2.5) (Roche, 454 Life Sciences; used only for 454 read assembly), Mira (ver. 3.4.0) (Chevreux et al. 2004), Velvet/Oasis (ver. 1.2.03) (Zerbino and Birney 2008), and Trinity (release 2013-02-25) (Grabherr et al. 2011). The k-mer size of 25 was used for all programs. Mira could be used only for 454 read assembly from the third instar larval samples and for the Illumina read assembly from the egg samples because of the large memory requirement. The results of these assemblies are summarized in Supplementary Tables S4.1. The number of assembled transcripts varied among the different assemblers, ranging from 37,181 by Trinity to 165,361 for Velvet/Oasis for 454 reads (larval midgut sample) and from 56,135 by Velvet/Oasis to 72,638 by Trinity for Illumina reads (egg sample). The average length and N50 of contigs were generally longer with the Trinity assembly (Supplementary Tables S1). Results of BLAST similarity search (ver. 2.2.26) (Altschul et al. 1990) against the UniProt protein database (http://www.uniprot.org) (The UniProt Consortium 2013) showed that fractions of contigs that had highly significant hits (E-value $\leq 10^{-100}$) were larger with the Trinity (18.9%) and Velvet/Oasis assemblies (~19.4%) than the Mira assembly (11%) although the difference was not significant ($P = 0.69$ for E-value $\leq 10^{-100}$ and $P = 0.61$ for all E-values by *t*-test between Trinity and Mira (Supplementary Figure S4.11). Note that Zhao et al. (2011b) showed the highest accuracy with Trinity among methods specialized in *de novo* transcriptome assemblies such as SOAPdenovo (Li et al. 2009), ABySS (Birol et al.

2009), Velvet/Oasis, and Trinity (they did not include Mira in their comparison). We also attempted the hybrid assemblies using two different sequencing platforms (454 and Illumina Genome Analyzer II) for the third instar larval midgut sample as well as for the pooled egg and third instar larval midgut samples (Supplementary Table S4.2). Furthermore, we performed assembly using the dataset pooled from egg (produced by Illumina Genome Analyzer II), third instar larval midgut (produced by Illumina Genome Analyzer II), and neonate samples (produced by Illumina HiSeq2000) (Table 4.1). Among all of these assemblies, the Trinity assembly using the pooled Illumina dataset had the longest average length of contigs and N50, even longer than the hybrid assemblies including 454 reads. With this assembly, more GH gene candidates were also identified. Therefore, we used this Trinity assembly using the pooled dataset as the most inclusive "combined *D. v. virgifera* transcriptome" for this study (Table 4.1).

4.4.3 Gene expression analysis.

To compare the gene expression levels, the paired-end reads were mapped onto our combined *D. v. virgifera* transcriptome using bowtie (ver. 1.0.0) (Langmead et al. 2009) with 0 mismatch. The numerical values of gene expression were measured by RPKM (reads per kilobase per million mapped reads) to normalize for the number of sequencing reads and total read length (Mortazavi et al. 2008). RPKM values above 0.3 (Ramsköld et al. 2009) as well as having more than 10 reads was used as the threshold for gene expression.

4.4.4 Identification of GH family genes.

Previously reported insect GH sequences were obtained for GH45 (Lee et al. 2004; Calderón-Cortés et al. 2010; Pauchet et al. 2010), for GH48 (Fujita et al. 2006; Pauchet et al. 2010), for GH28 (Pauchet et al. 2010), for GH9 (Willis et al. 2011), for GH5 (Acuña et al. 2012; Pauchet and Heckel 2013), for GH16 (Kim et al. 2000), and for GH31 (Willis et al. 2011) (see Supplementary Table 5). Using these sequences as queries, we searched GH gene candidates against our combined *D. v. virgifera* transcriptome using BLAST (ver. 2.2.24) similarity search (Altschul et al. 1990). All *D. v. virgifera* GH gene candidate sequences were also confirmed by BLASTN similarity search against the draft *D. v. virgifera* genome sequence (Hugh M. Robertson, personal communication). All *D. v. virgifera* GH gene candidate sequences were also confirmed by BLASTN similarity search against the draft *D. v. virgifera* genome sequence (Hugh M. Robertson, personal communication). The criterion we used to identify alternative spliced isoforms was to have a more than 60 bp of 100% identical region among the candidate sequences. I do not find any possible alternative spliced isoforms for GH sequences.

4.4.5 Similarity search

I performed the BLAST similarity searches (version 2.2.26) using *D. v. virgifera* GH sequences as queries against the non-redundant (NR) protein database at the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov) as well as ten insect genomes (*D. melanogaster*, *Anopheles gambiae*, *Aedes aegypti*, *Bombyx mori*, *Apis mellifera*, *Nasonia vitripennis*, *Solenopsis invicta*, *Ixodes scapularis*, *Rhodnius prolixus*, and *Acyrthosiphon pisum*).

4.4.6 Multiple sequence alignments and phylogenetic analysis.

Multiple alignments of GH protein sequences were generated using MAFFT (ver. 7.050b) with the L-INS-i algorithm, which uses a consistency-based objective function and local pairwise alignment with affine gap costs (Katoh and Standley 2013). Phylogenetic relationships were reconstructed by the maximum-likelihood method using RAxML (ver. 7.0.4) (Stamatakis 2006) with the PROTGAMMAJTT substitution model (JTT matrix with gamma-distributed rate variation). The neighbor-joining phylogenies (Saitou and Nei 1987) were reconstructed by using `neighbor` of the Phylip package (ver. 3.67) (Felsenstein 2005). The protein distances were estimated using `protdist` of the Phylip package with the JTT model. Non-parametric bootstrapping with 1000 pseudoreplicates (Felsenstein 1985) was used to estimate the confidence of branching patterns. FigTree (http://tree.bio.ed.ac.uk/software/figtree) was used to display the phylogenetic trees.

4.5 Acknowledgements

I would like to thank Drs. Kim Walden and Hugh M. Robertson (University of Illinois at Urbana-Champaign, USA) for letting us use the draft genome sequence of *D. v. virgifera* and Steven M. Van Belleghem for the transcriptome of *Pogonus chalceus* (salt marsh beetle).

4.6 References

Acuña R, Padilla BE, Flórez-Ramos CP, Rubio JD, Herrera JC, Benavides P, Lee S-J, Yeats TH, Egan AN, Doyle JJ, Rose JKC. 2012. Adaptive horizontal transfer of a bacterial gene to an invasive insect pest of coffee. *Proc Natl Acad Sci USA*. 109.

Allen ML, Mertens JA. 2008. Molecular Cloning and Expression of Three Polygalacturonase cDNAs from the Tarnished Plant Bug, *Lygus lineolaris*. *J Insect Sci*. 8:27.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol*. 215:403-410.

Aspeborg H, Coutinho PM, Wang Y, Brumer H, Henrissat B. 2012. Evolution, substrate specificity and subfamily classification of glycoside hydrolase family 5 (GH5). *BMC Evol Biol*. 12:186.

Berger E, Zhang D, Zverlov VV, Schwarz WH. 2007. Two noncellulosomal cellulases of Clostridium thermocellum, Cel9I and Cel48Y, hydrolyse crystalline cellulose synergistically. *FEMS Microbiology Letters*. 268:194-201.

Birol I, Jackman SD, Nielsen CB, Qian JQ, Varhol R, Stazyk G, Morin RD, Zhao Y, Hirst M, Schein JE, Horsman DE, Connors JM, Gascoyne RD, Marra MA, Jones SJ. 2009. De novo transcriptome assembly with ABySS. *Bioinformatics*. 25:2872-2877.

Bragatto I, Genta FA, Ribeiro AF, Terra WR, Ferreira C. 2010. Characterization of a β-1,3-glucanase active in the alkaline midgut of *Spodoptera frugiperda* larvae and its relation to β-glucan-binding proteins. *Insect Biochem Mol Biol*. 40:861-872.

Calderón-Cortés N, Watanabe H, Cano-Camacho H, Zavala-Páramo G, Quesada M. 2010. cDNA cloning, homology modelling and evolutionary insights into novel endogenous cellulases of the borer beetle *Oncideres albomarginata chamela* (Cerambycidae). *Insect Mol Biol*. 19:323-336.

Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. 2009. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res*. 37:D233-D238.

Celorio-Mancera MP, Allen ML, Powell AL, Ahmadi H, Salemi MR, Phinney BS, Shackel KA, Greve LC, Teuber LR, Labavitch JM. 2008. Polygalacturonase causes lygus-like damage on plants: cloning and identification of western tarnished plant bug (*Lygus hesperus*) polygalacturonases secreted during feeding. *Arthropod-Plant Interactions*. 2:215-225.

Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Müller WEG, Wetter T, Suhai S. 2004. Using the miraEST Assembler for Reliable and Automated mRNA Transcript Assembly and SNP Detection in Sequenced ESTs. *Genome Res*. 14:1147-1159.

Davison A, Blaxter M. 2005. Ancient origin of glycosyl hydrolase family 9 cellulase genes. *Mol Biol Evol*. 22:1273-1284.

Dun Z, Mitchell PD, Agosti M. 2010. Estimating Diabrotica virgifera virgifera damage functions with field trial data: applying an unbalanced nested error component model. *Journal of Applied Entomology*. 134:409-419.

Dunning Hotopp JC. 2011. Horizontal gene transfer between bacteria and animals. *Trends Genet*. 27:157-163.

Eigenheer AL, Keeling CI, Young S, Tittiger C. 2003. Comparison of gene representation in midguts from two phytophagous insects, Bombyx mori and Ips pini, using expressed sequence tags. *Gene*. 316:127-136.

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

Felsenstein J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Fujita K, Shimomura K, Yamamoto K, Yamashita T, Suzuki K. 2006. A chitinase structurally related to the glycoside hydrolase family 48 is indispensable for the hormonally induced diapause termination in a beetle. *Biochem Biophys Res Commun*. 345:502-507.

Gao Y, Bu Y, Luan Y-X. 2008. Phylogenetic Relationships of Basal Hexapods Reconstructed from Nearly Complete 18S and 28S rRNA Gene Sequences. *Zoological Science*. 25:1139-1145.

Garcia-Vallvé S, Romeu A, Palau J. 2000. Horizontal Gene Transfer of Glycosyl Hydrolases of the Rumen Fungi. *Mol Biol Evol*. 17:352-361.

Genta FA, Bragatto I, Terra WR, Ferreira C. 2009. Purification, characterization and sequencing of the major β-1,3-glucanase from the midgut of *Tenebrio molitor* larvae. *Insect Biochem Mol Biol*. 39:861-874.

Girard C, Jouanin L. 1999. Molecular cloning of a gut-specific chitinase cDNA from the beetle *Phaedon cochleariae*. *Insect Biochem Mol Biol*. 29:549-556.

Gladstone S, Hruska A. 2003. Guidelines for Promoting Safer and More Effective Pest Management with Small Holder Farmers: a Contribution to USAID-FFP Environmental Compliance. Georgia, USA: CARE USA.

Godfrey LD, Meinke LJ, Wright RJ. 1993a. Affects of Larval Injury by Western Com Rootworm (Coleoptera: Chrysomelidae) on Gas Exchange Parameters of Field Corn. *J Econ Entomol*. 86:1546-1556.

Godfrey LD, Meinke LJ, Wright RJ. 1993b. Vegetative and Reproductive Biomass Accumulation in Field Com: Response to Root Injury by Western Com Rootworm (Coleoptera: Chrysomelidae). *J Econ Entomol*. 86:1557-1573.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotech*. 29:644-652.

Gray ME, Steffey KL. 1998. Corn rootworm (Coleoptera: Chrysomelidae) larval injury and root compensation of 12 maize hybrids: an assessment of the economic injury index. *J Econ Entomol*. 91:723–740.

Harada Y, Hosoiri Y, Kuroda R. 2004. Isolation and evaluation of dextral-specific and dextral-enriched cDNA clones as candidates for the handedness-determining gene in a freshwater gastropod, Lymnaea stagnalis. *Dev Genes Evol*. 214:159-169.

Hou X, Meinke L, Arkebauer T. 1997. Soil moisture and larval western corn rootworm injury: influence on gas exchange parameters in corn. *Agronomy Journal*. 89:709–717.

Hung YL, Chen HJ, Liu JC, Chen YC. 2012. Catalytic efficiency diversification of duplicate beta-1,3-1,4-glucanases from Neocallimastix patriciarum J11. *Appl Environ Microbiol*. 78:4294-4300.

Hunt T, Bergsten J, Levkanicova Z, Papadopoulou A, John OS, Wild R, Hammond PM, Ahrens D, Balke M, Caterino MS, Gomez-Zurita J, Ribera I, Barraclough TG, Bocakova M, Bocak L, Vogler AP. 2007. A comprehensive phylogeny of beetles reveals the evolutionary origins of a superradiation. *Science*. 318:1913-1916.

Juge N. 2006. Plant protein inhibitors of cell wall degrading enzymes. *Trends in Plant Science*. 11:359-367.

Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol*. 30:772-780.

Keeling C, Yuen M, Liao N, Docking T, Chan S, Taylor G, Palmquist D, Jackman S, Nguyen A, Li M, Henderson H, Janes J, Zhao Y, Pandoh P, Moore R, Sperling F, Huber D, Birol I, Jones S, Bohlmann J. 2013. Draft genome of the mountain pine beetle, *Dendroctonus ponderosae* Hopkins, a major forest pest. *Genome Biol*. 14:R27.

Keeling CI, Henderson H, Li M, Yuen M, Clark EL, Fraser JD, Huber DP, Liao NY, Docking TR, Birol I, Chan SK, Taylor GA, Palmquist D, Jones SJ, Bohlmann J. 2012. Transcriptome and full-length cDNA resources for the mountain pine beetle, *Dendroctonus ponderosae* Hopkins, a major insect pest of pine forests. *Insect Biochem Mol Biol*. 42:525-536.

Kim Y-S, Ryu J-H, Han S-J, Choi K-H, Nam K-B, Jang I-H, Lemaitre B, Brey PT, Lee W-J. 2000. Gram-negative Bacteria-binding Protein, a Pattern Recognition Receptor for Lipopolysaccharide and β-1,3-Glucan That Mediates the Signaling for the Induction of Innate Immune Genes in Drosophila melanogaster Cells. *J Biol Chem*. 275:32721-32727.

Kirsch R, Wielsch N, Vogel H, Svatos A, Heckel D, Pauchet Y. 2012. Combining proteomics and transcriptome sequencing to identify active plant-cell-wall-degrading enzymes in a leaf beetle. *BMC Genomics*. 13:587.

Kovalchuk SN, Bakunina IY, Burtseva YV, Emelyanenko VI, Kim NY, Guzev KV, Kozhemyako VB, Rasskazov VA, Zvyagintseva TN. 2009. An *endo*-(1→3)-β-D-glucanase from the scallop *Chlamys albidus*: catalytic properties, cDNA cloning and secondary-structure characterization. *Carbohydrate Research*. 344:191-197.

Langmead B, Trapnell C, Pop M, Salzberg S. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 10:R25.

Lee SJ, Kim SR, Yoon HJ, Kim I, Lee KS, Je YH, Lee SM, Seo SJ, Dae Sohn H, Jin BR. 2004. cDNA cloning, expression, and enzymatic activity of a cellulase from the mulberry longicorn beetle, *Apriona germari*. *Comp Biochem Physiol B*. 139:107-116.

Levine E, Oloumi-Sadeghi H. 1991. Management of Diabroticite Rootworms in Corn. *Annu Rev Entomol*. 36:229-255.

Li L, Frohlich J, Pfeiffer P, Konig H. 2003. Termite gut symbiotic archaezoa are becoming living metabolic fossils. *Eukaryot Cell*. 2:1091-1098.

Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J. 2009. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*. 25:1966-1967.

Ljungdahl LG. 2008. The Cellulase/Hemicellulase System of the Anaerobic Fungus Orpinomyces PC-2 and Aspects of Its Applied Use. *Ann N Y Acad Sci*. 1125:308-321.

Mertz B, Gu X, Reilly PJ. 2009. Analysis of functional divergence within two structurally related glycoside hydrolase families. *Biopolymers*. 91:478-495.

Metcalf R. 1986. Forward. In: JL Krysan, TA Miller, editors. Methods for the study of pest diabrotica. New York: Springer-Verlag. p. vii–xv.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat Methods*. 5:621-628.

Pauchet Y, Freitak D, Heidel-Fischer HM, Heckel DG, Vogel H. 2009. Immunity or digestion: glucanase activity in a glucan-binding protein family from Lepidoptera. *J Biol Chem*. 284:2214-2224.

Pauchet Y, Heckel DG. 2013. The genome of the mustard leaf beetle encodes two active xylanases originally acquired from bacteria through horizontal gene transfer. *Proceedings of the Royal Society B: Biological Sciences*. 280:20131021.

Pauchet Y, Wilkinson P, Chauhan R, ffrench-Constant RH. 2010. Diversity of Beetle Genes Encoding Novel Plant Cell Wall Degrading Enzymes. *PLoS ONE*. 5:e15635.

Ramsköld D, Wang ET, Burge CB, Sandberg R. 2009. An Abundance of Ubiquitously Expressed Genes Revealed by Tissue Transcriptome Sequence Data. *PLoS Comput Biol*. 5:e1000598.

Riedell WE. 1990. Rootworm and mechanical damage effects on root morphology and water relations in maize. *Crop Science*. 30:628-631.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*. 4:406-425.

Sappington TW, Siegfried BD, Guillemaud T. 2006. Coordinated Diabrotica genetics research: accelerating progress on an urgent insect pest problem. *American Entomologist*. 52:90-97.

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

Shen Z, Denton M, Mutti N, Pappan K, Kanost MR, Reese JC, Reeck GR. 2003. Polygalacturonase from *Sitophilus oryzae*: Possible horizontal transfer of a pectinase gene from fungi to weevils. *J Insect Sci*. 3:24.

Siegfried BD, Waterfield N, Ffrench-Constant RH. 2005. Expressed sequence tags from Diabrotica virgifera virgifera midgut identify a coleopteran cadherin and a diversity of cathepsins. *Insect Mol Biol*. 14:137-143.

Sinnott M. 1990. Catalytic mechanisms of enzymatic glycosyl transfer. *Chem Rev*. 90:1171–1202.

Smant G, Stokkermans JP, Yan Y, de Boer JM, Baum TJ, Wang X, Hussey RS, Gommers FJ, Henrissat B, Davis EL, Helder J, Schots A, Bakker J. 1998. Endogenous cellulases in animals: isolation of beta-1, 4-endoglucanase genes from two species of plant-parasitic cyst nematodes. *Proc Natl Acad Sci USA*. 95:4906-4911.

Song JM, Nam K, Sun YU, Kang MH, Kim CG, Kwon ST, Lee J, Lee YH. 2010. Molecular and biochemical characterizations of a novel arthropod endo-beta-1,3-glucanase from the Antarctic springtail, Cryptopygus antarcticus, horizontally acquired from bacteria. *Comp Biochem Physiol B Biochem Mol Biol*. 155:403-412.

Spike BP, Tollefson JJ. 1991. Yield Response of Corn Subjected to Western Corn Root worm (Coleoptera: Chrysomelidae) Infestation and Lodging. *J Econ Entomol*. 84:1585-1590.

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

Sutter GR, Fisher JR, Elliott NC, Branson TF. 1990. Effect of Insecticide Treatments on Root Lodging and Yields of Maize in Controlled Infestations of Western Corn Rootworms (Coleoptera: Chrysomelidae). *J Econ Entomol*. 83:2414-2420.

The UniProt Consortium. 2013. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res*. 41:D43-D47.

Tribolium Genome Sequencing Consortium. 2008. The genome of the model beetle and pest Tribolium castaneum. *Nature*. 452:949-955.

Urias-Lopez MA, Meinke LJ. 2001. Influence of western corn rootworm (Coleoptera: Chrysomelidae) larval injury on yield of different types of maize. *J Econ Entomol*. 94:106-111.

Valencia A, Alves AP, Siegfried BD. 2013. Molecular cloning and functional characterization of an endogenous endoglucanase belonging to GHF45 from the western corn rootworm, Diabrotica virgifera virgifera. *Gene*. 513:260-267.

Valentine BD. 2005. The scientific name of the coffee bean weevil and some additional bibliography (Coleoptera: Anthribidae: *Araecerus* Schönherr). *Insecta Mundi* 19:247-253.

Van Belleghem SM, Roelofs D, Van Houdt J, Hendrickx F. 2012. *De novo* Transcriptome Assembly and SNP Discovery in the Wing Polymorphic Salt Marsh Beetle *Pogonus chalceus* (Coleoptera, Carabidae). *PLoS ONE*. 7:e42605.

Waller JM, Bigger M, Hillocks RJ. 2007. Postharvest and processing pests and microbial problems. In: JM Waller, M Bigger, RJ Hillocks, editors. Coffee Pests, Diseases and Their Management. CABI, Wallingford, UK. p. 325–335.

Watanabe H, Tokuda G. 2010. Cellulolytic Systems in Insects. *Annu Rev Entomol*. 55:609-632.

Wheeler D, Redding AJ, Werren JH. 2013. Characterization of an Ancient Lepidopteran Lateral Gene Transfer. *PLoS ONE*. 8:e59262.

Willis JD, Oppert B, Oppert C, Klingeman WE, Jurat-Fuentes JL. 2011. Identification, cloning, and expression of a GHF9 cellulase from *Tribolium castaneum* (Coleoptera: Tenebrionidae). *J Insect Physiol*. 57:300-306.

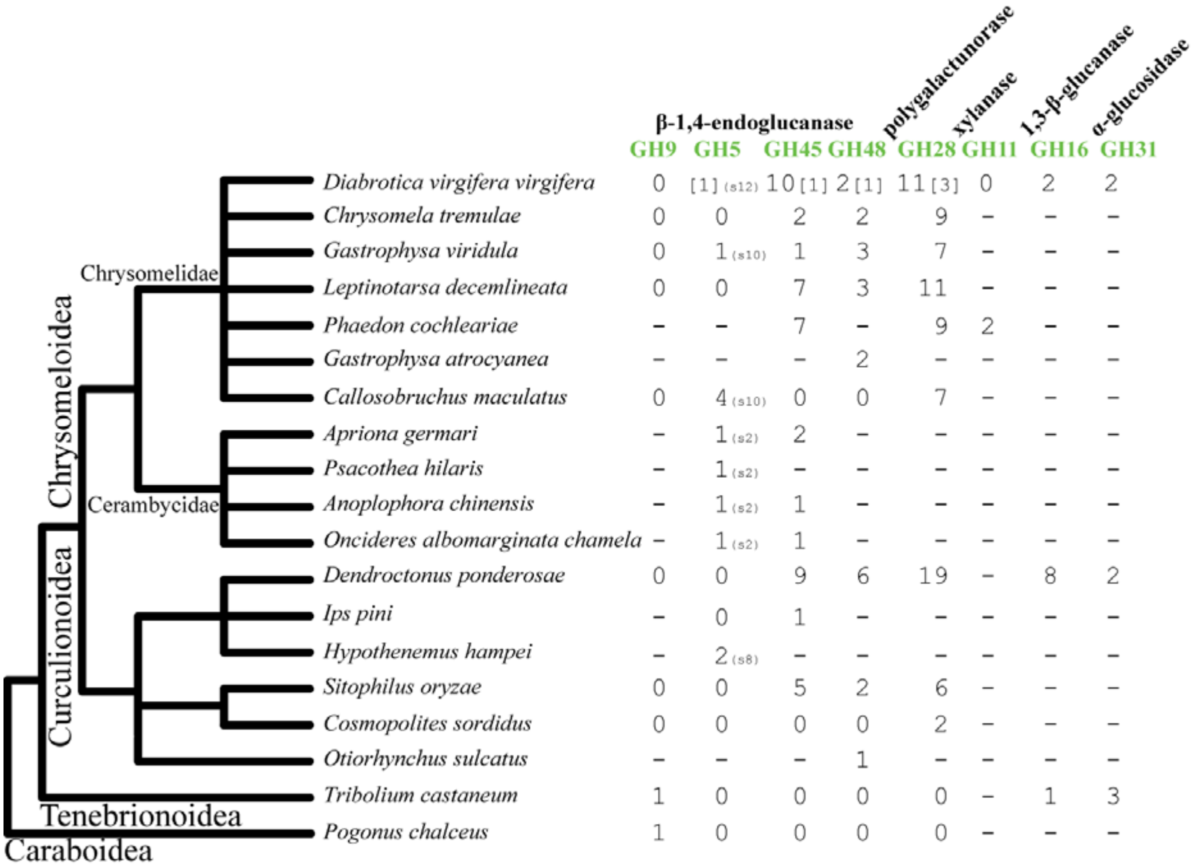
Xu B, Janson J-C, Sellos D. 2001. Cloning and sequencing of a molluscan endo-β-1,4-glucanase gene from the blue mussel, Mytilus edulis. *Eur J Biochem*. 268:3718-3727.

Zerbino DR, Birney E. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*. 18:821-829.

Zhao QY, Wang Y, Kong YM, Luo D, Li X, Hao P. 2011. Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. *BMC bioinfo*. 12 Suppl 14:S2.

Table 4.1. Summary of the *D. v. virgifera* transcriptome assembly using the pooled dataset.

Samples	Egg, neonates, and third larval midgut
Number of paired-end reads (base pairs) before filtering	1,462.2 x 10^6 (144,690 x 10^6 bp)
Number of paired-end reads (base pairs) after filtering	781.7 x 10^6 (77,393 x 10^6 bp)
Assembly program used	Trinity (2013-02-25)
Total number of contigs	163,871
Average contig length (range)	914 bp (201 – 31,064 bp)
N50 length	1,396 bp



	β-1,4-endoglucanase				polygalactunorase	xylanase	1,3-β-glucanase	α-glucosidase
	GH9	GH5	GH45	GH48	GH28	GH11	GH16	GH31
Diabrotica virgifera virgifera	0	[1] (s12)	10 [1]	2 [1]	11 [3]	0	2	2
Chrysomela tremulae	0	0	2	2	9	-	-	-
Gastrophysa viridula	0	1 (s10)	1	3	7	-	-	-
Leptinotarsa decemlineata	0	0	7	3	11	-	-	-
Phaedon cochleariae	-	-	7	-	9	2	-	-
Gastrophysa atrocyanea	-	-	-	2	-	-	-	-
Callosobruchus maculatus	0	4 (s10)	0	0	7	-	-	-
Apriona germari	-	1 (s2)	2	-	-	-	-	-
Psacothea hilaris	-	1 (s2)	-	-	-	-	-	-
Anoplophora chinensis	-	1 (s2)	1	-	-	-	-	-
Oncideres albomarginata chamela	-	1 (s2)	1	-	-	-	-	-
Dendroctonus ponderosae	0	0	9	6	19	-	8	2
Ips pini	-	0	1	-	-	-	-	-
Hypothenemus hampei	-	2 (s8)	-	-	-	-	-	-
Sitophilus oryzae	0	0	5	2	6	-	-	-
Cosmopolites sordidus	0	0	0	0	2	-	-	-
Otiorhynchus sulcatus	-	-	-	1	-	-	-	-
Tribolium castaneum	1	0	0	0	0	-	1	3
Pogonus chalceus	1	0	0	0	0	-	-	-

Figure 4.1. Distribution of glycoside hydrolase family genes among polyphagan coleopterans. All numbers are taken from Pauchet et al. (2010) except for *D. v. virgifera* (this study, partial sequences in square brackets), *Cosmopolites sordidus* (this study), *Pogonus chalceus* (this study) from the transcriptome (Van Belleghem et al. 2012), *Tribolium castaneum* (Tribolium Genome Sequencing Consortium 2008; Willis et al. 2011), *Dendroctonus ponderosae* (Keeling et al. 2013), *Phaedon cochleariae* GH45 and GH28 (Kirsch et al. 2012), *Phaedon cochleariae* GH11 (Pauchet and Heckel 2013), *Gastrophysa atrocyanea* GH28 (Fujita et al. 2006), and *Otiorhynchus sulcatus* GH48. GH5 genes are classified into 51 subfamilies (Aspeborg et al. 2012) and four subfamilies are found in coleopteran species as shown in the above table: s2 (subfamily 2), s8 (subfamily 8), s10 (subfamily 10), and s12 (subfamily 12). Accession numbers for all coleopteran GH genes included in this study are found in Supplementary Table S6. The taxonomical relationship is based on Hunt et al. (2007). *Pogonus chalceus* (Suborder Adephaga) is shown as the outgroup. '-': not determined.

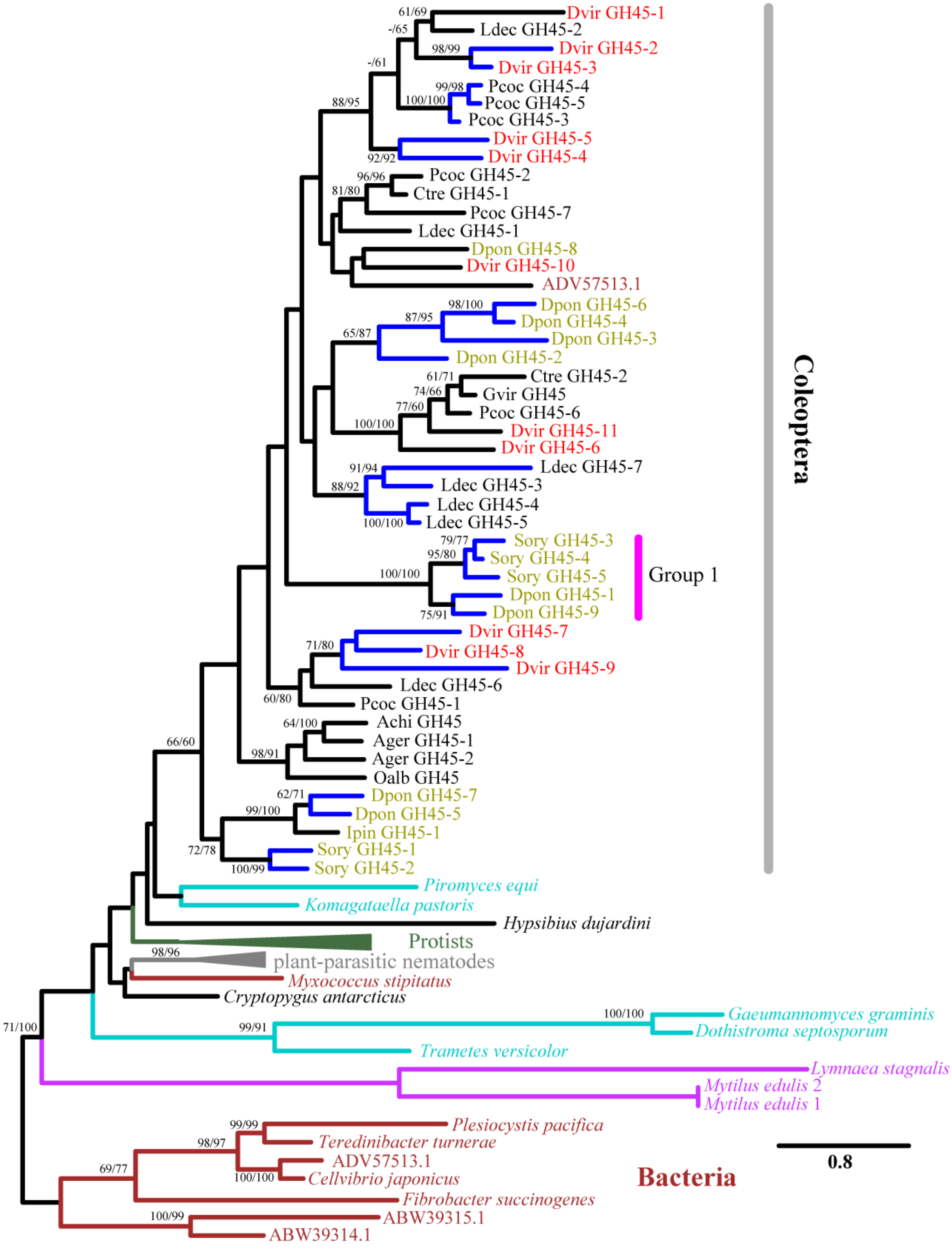

61/69
Dvir GH45-1
-/65
Ldec GH45-2
98/99
Dvir GH45-2
-/61
Dvir GH45-3
99/98
Pcoc GH45-4
88/95
100/100
Pcoc GH45-5
Pcoc GH45-3
Dvir GH45-5
92/92
Dvir GH45-4
96/96
Pcoc GH45-2
81/80
Ctre GH45-1
Pcoc GH45-7
Ldec GH45-1
Dpon GH45-8
Dvir GH45-10
ADV57513.1
98/100
Dpon GH45-6
87/95
Dpon GH45-4
65/87
Dpon GH45-3
Dpon GH45-2
61/71
Ctre GH45-2
74/66
Gvir GH45
77/60
Pcoc GH45-6
100/100
Dvir GH45-11
Dvir GH45-6
91/94
Ldec GH45-7
88/92
Ldec GH45-3
Ldec GH45-4
100/100
Ldec GH45-5
79/77
Sory GH45-3
95/80
Sory GH45-4
100/100
Sory GH45-5
Dpon GH45-1
75/91
Dpon GH45-9
Group 1
Coleoptera
71/80
Dvir GH45-7
Dvir GH45-8
Dvir GH45-9
Ldec GH45-6
60/80
Pcoc GH45-1
64/100
Achi GH45
66/60
Ager GH45-1
98/91
Ager GH45-2
Oalb GH45
62/71
Dpon GH45-7
99/100
Dpon GH45-5
Ipin GH45-1
72/78
Sory GH45-1
100/99
Sory GH45-2
Piromyces equi
Komagataella pastoris
Hypsibius dujardini
Protists
98/96
plant-parasitic nematodes
Myxococcus stipitatus
Cryptopygus antarcticus
100/100
Gaeumannomyces graminis
71/100
99/91
Dothistroma septosporum
Trametes versicolor
Lymnaea stagnalis
Mytilus edulis 2
Mytilus edulis 1
99/99
Plesiocystis pacifica
98/97
Teredinibacter turnerae
0.8
69/77
ADV57513.1
100/100
Cellvibrio japonicus
Bacteria
Fibrobacter succinogenes
100/99
ABW39315.1
ABW39314.1

Figure 4.2. The maximum-likelihood phylogeny of GH45 proteins. Forty seven GH45 protein sequences from eleven coleopteran species are included. Their species abbreviations are found in Supplementary Table S4.5. Olive-colored names indicate the coleopteran species belonging to the superfamily Curculionoidea and all other coleopteran sequences colored in black belong to the superfamily Chrysomeloidea. In addition to coleopterans, sequences are included from two mollusks (*Mytilus edulis* and *Lymnaea stagnalis*, shown in purple), *Cryptopygus antarcticus* (Collembola), *Hypsibius dujardini* (Tardigrada), 24 protists (shown in dark green), a plant-parasitic nematode (*Bursaphelenchus xylophilus*, 10 sequences, shown in grey), 5 representative fungi (shown in cyan, chosen from 138 sequences), and 7 representative bacteria (shown in brown, chosen from 18 sequences). Bacterial sequences were used as outgroups. The numbers at internal branches show the bootstrap support values (%) for the maximum-likelihood and neighbor-joining phylogenies in this order. Supporting values are shown for the internal branches that have at least one support higher than 60%. Blue-colored branches indicate the species-specific gene duplications within a cluster supported by higher than 70% of bootstrap values. The scale bar represents the number of amino acid substitutions per site. Supplementary Figure S2 shows the identical phylogeny with all details.

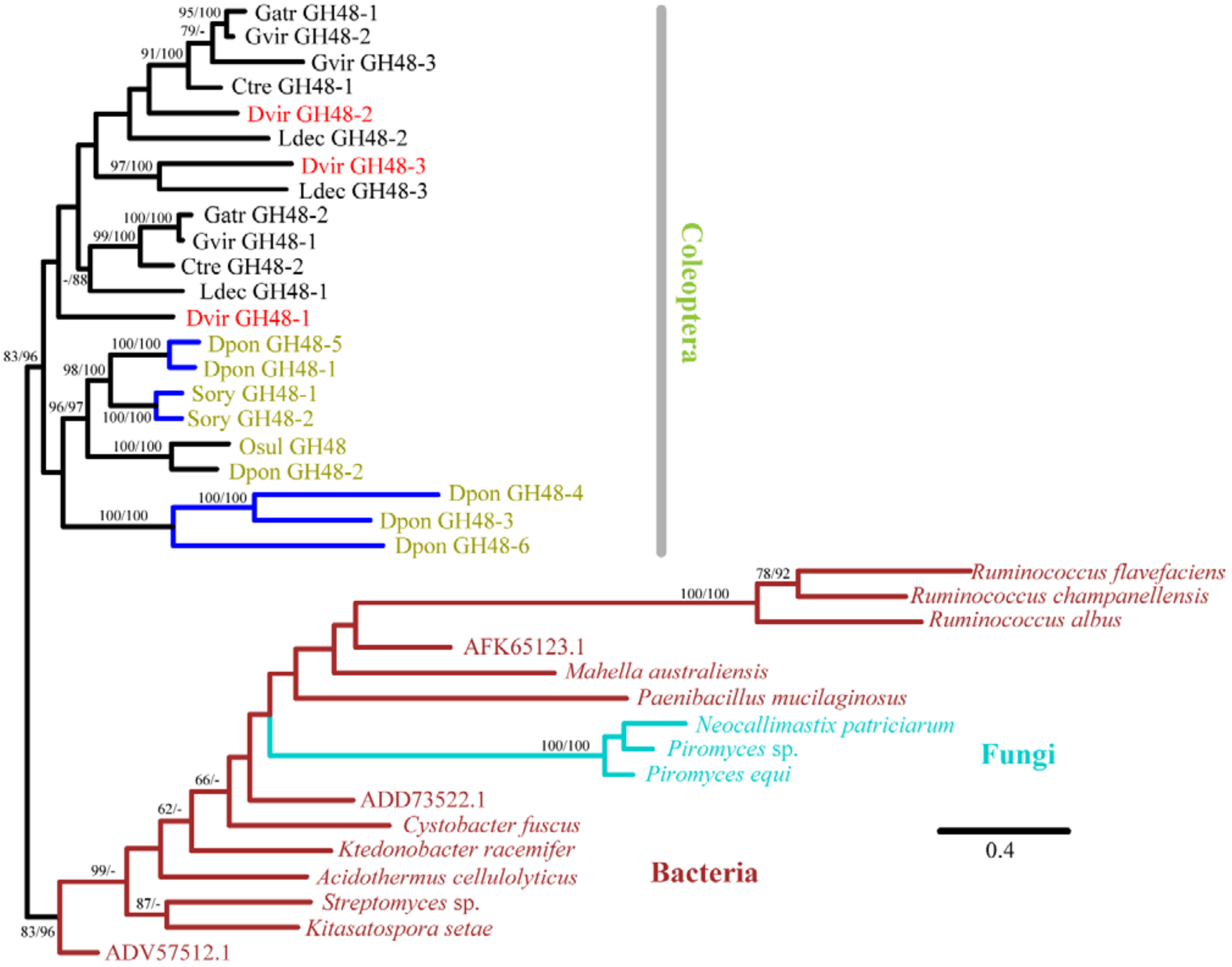

95/100
Gatr GH48-1
79/-
Gvir GH48-2
91/100
Gvir GH48-3
Ctre GH48-1
Dvir GH48-2
Ldec GH48-2
97/100
Dvir GH48-3
Ldec GH48-3
100/100
Gatr GH48-2
99/100
Gvir GH48-1
Ctre GH48-2
-/88
Ldec GH48-1
Dvir GH48-1
83/96
100/100
Dpon GH48-5
98/100
Dpon GH48-1
Sory GH48-1
96/97
100/100
Sory GH48-2
100/100
Osul GH48
Dpon GH48-2
100/100
Dpon GH48-4
100/100
Dpon GH48-3
Dpon GH48-6
Coleoptera
78/92
Ruminococcus flavefaciens
100/100
Ruminococcus champanellensis
Ruminococcus albus
AFK65123.1
Mahella australiensis
Paenibacillus mucilaginosus
Neocallimastix patriciarum
100/100
Piromyces sp.
Fungi
Piromyces equi
66/-
ADD73522.1
Cystobacter fuscus
62/-
0.4
Ktedonobacter racemifer
99/-
Acidothermus cellulolyticus
Bacteria
87/-
Streptomyces sp.
Kitasatospora setae
83/96
ADV57512.1

Figure 4.3. The maximum-likelihood phylogeny of GH48 proteins. Twenty two GH48 protein sequences from seven coleopteran species are included. Their species abbreviations are found in Supplementary Table S5. Olive-colored names indicate the coleopteran species belonging to the superfamily Curculionoidea, and all other coleopteran sequences colored in black belong to the superfamily Chrysomeloidea. In addition to coleopterans, sequences are included from 13 representative bacteria (shown in brown, chosen from 653 sequences) and 3 fungi (shown in cyan). When species names are not known for bacterial sequences, those sequences are labeled with the accession numbers. Numbers of sequences from the same specie or groups are shown in parentheses next to their names. Bacterial sequences were used as outgroups. The numbers at internal branches show the bootstrap support values (%) for the maximum-likelihood phylogenies and neighbor-joining in this order. Supporting values are shown for the internal branches that have at least one support higher than 60%. Blue-colored branches indicate the species-specific gene duplications within a cluster supported by higher than 100% of bootstrap values. The scale bar represents the number of amino acid substitutions per site. Supplementary Figure S4.5 shows the identical phylogeny with all details.

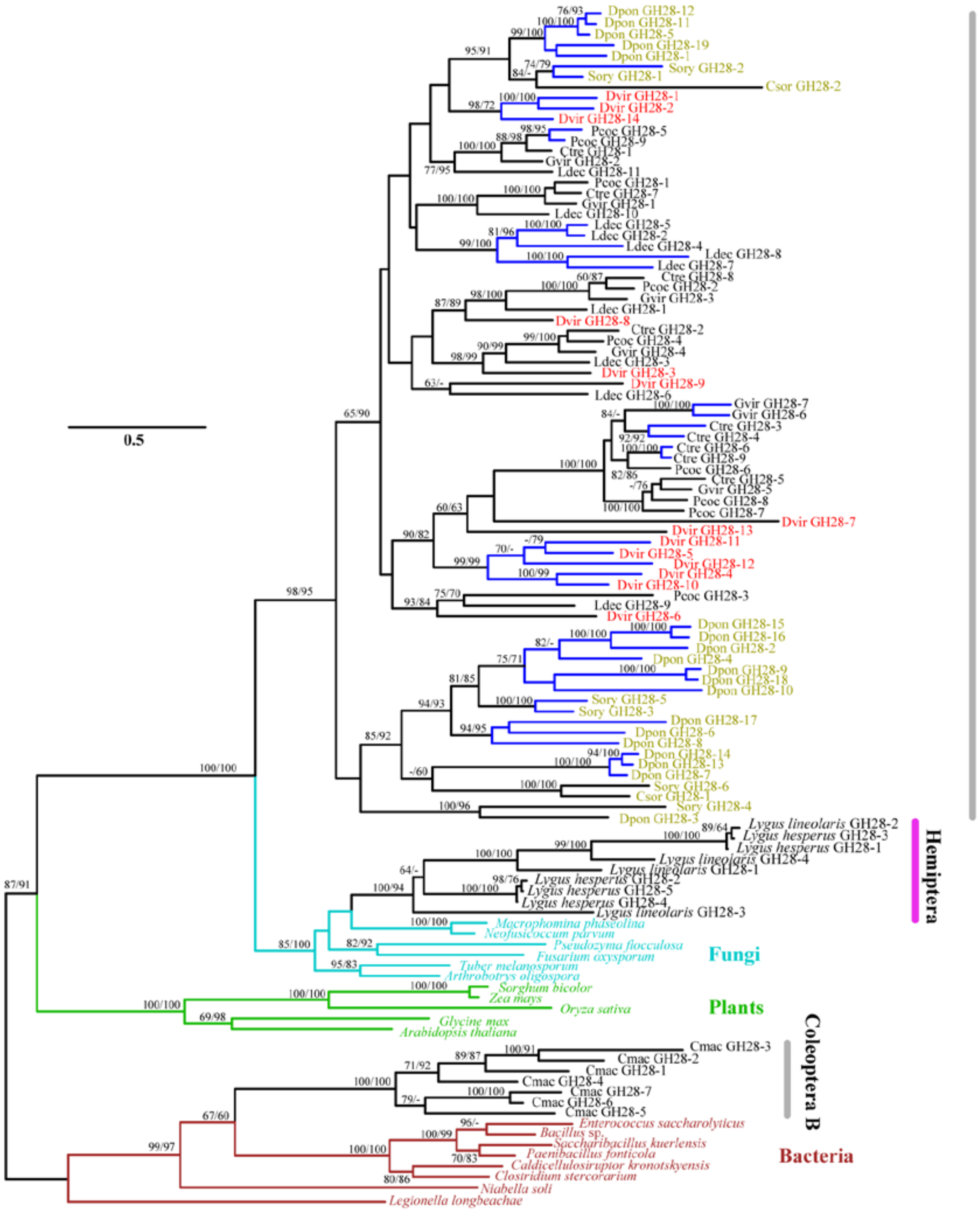

76/93
Dpon GH28-12
100/100
Dpon GH28-11
99/100
Dpon GH28-5
95/91
Dpon GH28-19
Dpon GH28-1
74/79
Sory GH28-2
84/-
Sory GH28-1
Csor GH28-2
100/100
Dvir GH28-1
98/72
Dvir GH28-2
Dvir GH28-14
98/95
Pcoc GH28-5
88/98
Pcoc GH28-9
100/100
Ctre GH28-1
Gvir GH28-2
77/95
Ldec GH28-11
100/100
Pcoc GH28-1
100/100
Ctre GH28-7
Gvir GH28-1
Ldec GH28-10
100/100
Ldec GH28-5
81/96
Ldec GH28-2
99/100
Ldec GH28-4
Ldec GH28-8
100/100
Ldec GH28-7
60/87
Ctre GH28-8
100/100
Pcoc GH28-2
98/100
Gvir GH28-3
87/89
Ldec GH28-1
Dvir GH28-8
Ctre GH28-2
99/100
Pcoc GH28-4
90/99
Gvir GH28-4
98/99
Ldec GH28-3
Dvir GH28-3
63/-
Dvir GH28-9
Ldec GH28-6
0.5
65/90
100/100
Gvir GH28-7
84/-
Gvir GH28-6
Ctre GH28-3
92/92
Ctre GH28-4
100/100
Ctre GH28-6
100/100
Ctre GH28-9
Pcoc GH28-6
82/86
Ctre GH28-5
-/76
Gvir GH28-5
Pcoc GH28-8
60/63
100/100
Pcoc GH28-7
Dvir GH28-7
90/82
Dvir GH28-13
-/79
Dvir GH28-11
70/-
Dvir GH28-5
99/99
Dvir GH28-12
100/99
Dvir GH28-4
Dvir GH28-10
98/95
75/70
Pcoc GH28-3
93/84
Ldec GH28-9
Dvir GH28-6
100/100
Dpon GH28-15
100/100
Dpon GH28-16
82/-
Dpon GH28-2
75/71
Dpon GH28-4
100/100
Dpon GH28-9
81/85
Dpon GH28-18
Dpon GH28-10
94/93
100/100
Sory GH28-5
Sory GH28-3
Dpon GH28-17
94/95
Dpon GH28-6
85/92
Dpon GH28-8
94/100
Dpon GH28-14
100/100
Dpon GH28-13
-/60
Dpon GH28-7
100/100
Sory GH28-6
Csor GH28-1
100/96
Sory GH28-4
Dpon GH28-3
Coleoptera A
100/100
89/64
Lygus lineolaris GH28-2
100/100
Lygus hesperus GH28-3
Lygus hesperus GH28-1
99/100
100/100
Lygus lineolaris GH28-4
64/-
Lygus lineolaris GH28-1
98/76
Lygus hesperus GH28-2
100/94
100/100
Lygus hesperus GH28-5
Lygus hesperus GH28-4
87/91
Lygus lineolaris GH28-3
Hemiptera
100/100
Macrophomina phaseolina
Neofusicoccum parvum
85/100
82/92
Pseudozyma flocculosa
Fusarium oxysporum
Fungi
95/83
Tuber melanosporum
Arthrobotrys oligospora
100/100
Sorghum bicolor
100/100
100/100
Zea mays
Oryza sativa
Plants
69/98
Glycine max
Arabidopsis thaliana
100/91
Cmac GH28-3
89/87
Cmac GH28-2
71/92
Cmac GH28-1
100/100
Cmac GH28-4
100/100
Cmac GH28-7
79/-
Cmac GH28-6
Cmac GH28-5
Coleoptera B
67/60
96/-
Enterococcus saccharolyticus
100/99
Bacillus sp.
99/97
Saccharibacillus kuerlensis
100/100
Paenibacillus fonticola
70/83
Caldicellulosiruptor kronotskyensis
80/86
Clostridium stercorarium
Bacteria
Niabella soli
Legionella longbeachae

Figure 4.4. The maximum-likelihood phylogeny of GH28 proteins. Eight four GH28 protein sequences from eight coleopteran species are included. Their species abbreviations are found in Supplementary Table S4.5. Olive-colored names indicate the coleopteran species belonging to the superfamily Curculionoidea, and all other coleopteran sequences colored in black belong to the superfamily Chrysomeloidea. In addition to coleopterans, sequences are included from plant bugs (*Lygus hesperus* and *Lygus lineolaris*, Hemiptera; 9 sequences), 6 representative fungi (shown in cyan, chosen from 651 sequences), 8 representative bacteria (shown in brown, chosen from 42 sequences), and 5 representative plants (shown in green, chosen from 491 sequences). Bacterial sequences were used as outgroups. The numbers at internal branches show the bootstrap support values (%) for the maximum-likelihood and neighbor-joining phylogenies in this order. Supporting values are shown for the internal branches that have at least one support higher than 60%. Blue-colored branches indicate the species-specific gene duplications within a cluster supported by higher than 70% of bootstrap values. The scale bar represents the number of amino acid substitutions per site. Supplementary Figure S6 shows the identical phylogeny with all details.

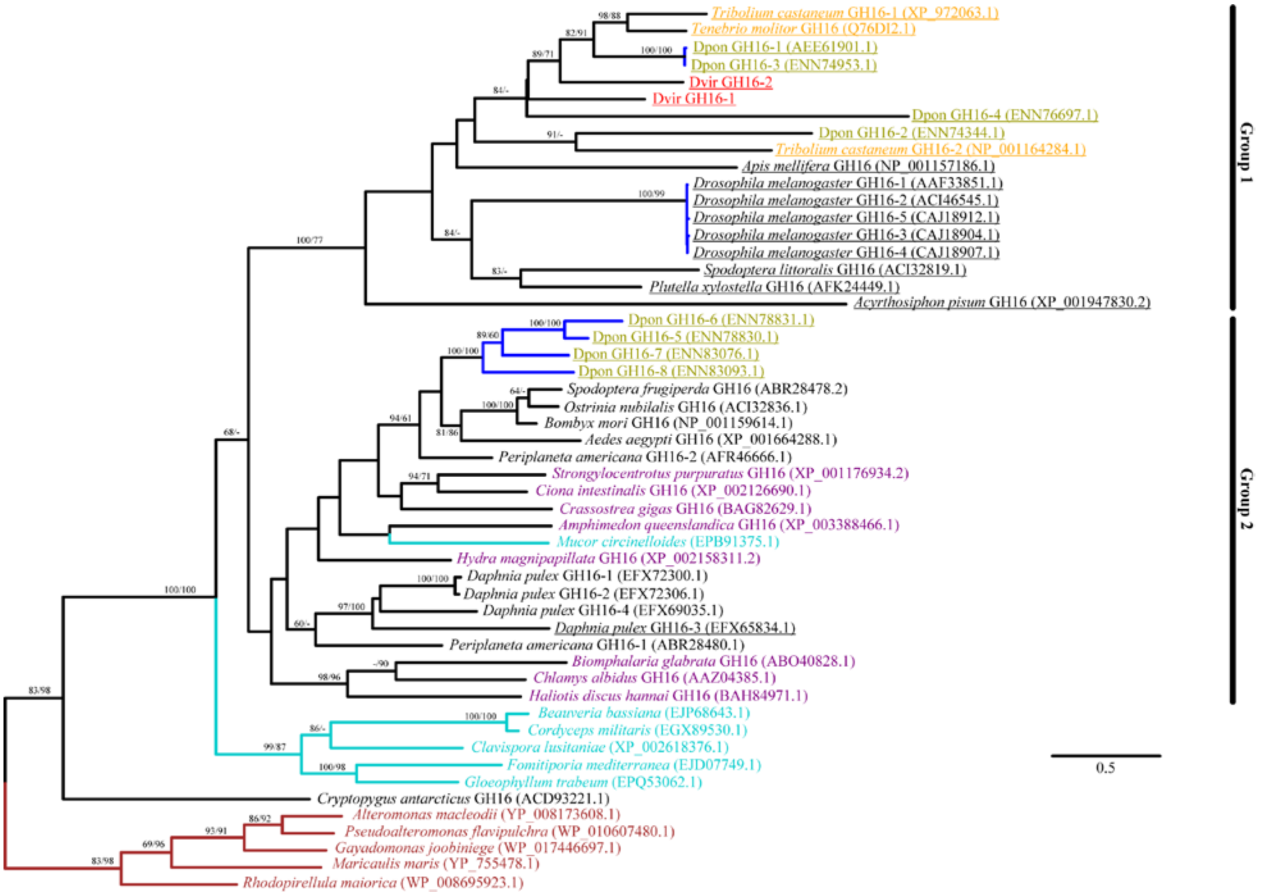

98/88
Tribolium castaneum GH16-1 (XP_972063.1)
82/91
Tenebrio molitor GH16 (Q76DI2.1)
89/71
100/100
Dpon GH16-1 (AEE61901.1)
Dpon GH16-3 (ENN74953.1)
Dvir GH16-2
84/-
Dvir GH16-1
Dpon GH16-4 (ENN76697.1)
91/-
Dpon GH16-2 (ENN74344.1)
Tribolium castaneum GH16-2 (NP_001164284.1)
Apis mellifera GH16 (NP_001157186.1)
100/99
Drosophila melanogaster GH16-1 (AAF33851.1)
Drosophila melanogaster GH16-2 (ACI46545.1)
Drosophila melanogaster GH16-5 (CAJ18912.1)
Drosophila melanogaster GH16-3 (CAJ18904.1)
Drosophila melanogaster GH16-4 (CAJ18907.1)
100/77
84/-
83/-
Spodoptera littoralis GH16 (ACI32819.1)
Plutella xylostella GH16 (AFK24449.1)
Acyrthosiphon pisum GH16 (XP_001947830.2)
100/100
Dpon GH16-6 (ENN78831.1)
89/60
Dpon GH16-5 (ENN78830.1)
100/100
Dpon GH16-7 (ENN83076.1)
Dpon GH16-8 (ENN83093.1)
64/-
Spodoptera frugiperda GH16 (ABR28478.2)
100/100
Ostrinia nubilalis GH16 (ACI32836.1)
94/61
Bombyx mori GH16 (NP_001159614.1)
68/-
81/86
Aedes aegypti GH16 (XP_001664288.1)
Periplaneta americana GH16-2 (AFR46666.1)
94/71
Strongylocentrotus purpuratus GH16 (XP_001176934.2)
Ciona intestinalis GH16 (XP_002126690.1)
Crassostrea gigas GH16 (BAG82629.1)
Amphimedon queenslandica GH16 (XP_003388466.1)
Mucor circinelloides (EPB91375.1)
Hydra magnipapillata GH16 (XP_002158311.2)
100/100
100/100
Daphnia pulex GH16-1 (EFX72300.1)
Daphnia pulex GH16-2 (EFX72306.1)
97/100
Daphnia pulex GH16-4 (EFX69035.1)
60/-
Daphnia pulex GH16-3 (EFX65834.1)
Periplaneta americana GH16-1 (ABR28480.1)
-/90
Biomphalaria glabrata GH16 (ABO40828.1)
98/96
Chlamys albidus GH16 (AAZ04385.1)
Haliotis discus hannai GH16 (BAH84971.1)
83/98
100/100
Beauveria bassiana (EJP68643.1)
86/-
Cordyceps militaris (EGX89530.1)
99/87
Clavispora lusitaniae (XP_002618376.1)
100/98
Fomitiporia mediterranea (EJD07749.1)
Gloeophyllum trabeum (EPQ53062.1)
Cryptopygus antarcticus GH16 (ACD93221.1)
86/92
Alteromonas macleodii (YP_008173608.1)
93/91
Pseudoalteromonas flavipulchra (WP_010607480.1)
69/96
Gayadomonas joobiniege (WP_017446697.1)
83/98
Maricaulis maris (YP_755478.1)
Rhodopirellula maiorica (WP_008695923.1)
Group 1
Group 2
0.5

Figure 4.5. The maximum-likelihood phylogeny of representative GH16 family proteins. Thirteen GH16 protein sequences from four coleopteran species (shown in blue and red) are included. *D. v. virgifera* sequences are shown in red. The underbars indicate the sequences to have different the catalytic nucleophile and proton donor sites. Metazoan except for arthropods, fungal (6 chosen from 204 sequences), and bacterial sequences (5 chosen from 209 sequences) are indicated by purple, cyan, and brown, respectively. Bacterial sequences were used as outgroups. The numbers at internal branches show the bootstrap support values (%) from 1000 pseudo-replications for maximum-likelihood and the neighbor-joining phylogenies in this order. Only bootstrap values higher than 60% are shown. The scale bar represents the number of amino acid substitutions per site.

Chapter 5

Conclusion and Future Works

In this dissertation, two multigene families, chemosensory receptors and glycoside hydrolase enzymes, were studied for their origin and evolutionary mechanisms as well as their functional adaption. For chemosensory receptors, I focused on trace amine-associated receptors (TAARs), which are a member of the G-protein-coupled receptors (GPCR) superfamily. For another multigene family, glycoside hydrolase (GH) genes, a member of carbohydrate active enzymes, were examined and I focused on three GH families belonging to two cellulolytic (GH45 and GH48) and one pectolytic (GH28) enzymes.

In Chapter 2, I have studied the origin and molecular evolutionary mechanisms on TAARs. An ancestral-type TAAR has emerged before the divergence of gnathostomes from jawless fish (sea lamprey). Older types of TAAR subfamilies (TAAR1-5) except for TAAR4 are more conserved and maintained as single-copy genes (no duplication nor loss) in each genome. Newer types of mammalian TAARs (TAAR6-9, E1, and M1-3) are found only in therian mammals and they have experienced frequent species-specific duplications. Generally, older type of TAARs is primary amine detecting receptors while newer types are in general tertiary amine detecting receptors. Positive selection was observed around the ligand-binding sites in TAAR7 and TAAR8 proteins among tertiary amine detecting receptors. These changes could have affected ligand-binding activities and specificities in these TAARs. This may have contributed to mammalian adaptation to the dynamic land environment by allowing finer discrimination among a diverse array of volatile amines. Different ecological factors may have led to additional duplications or losses of some TAARs in response to specific ecological conditions in some species, and thus the birth and death processes of TAARs seem to be under the influence of both environmental and evolutionary factors.

In Chapter 3, I have identified TAAR genes in twelve primate genomes. Primates have in general a smaller number of TAARs compared to other mammalian species. The ancestral species of primates arose as arboreal animals and arboreal life must have made easy escape from many ground-living predators possible and significantly reduced the predator exposures. The dispensability of primate TAAR genes must have significantly affected the TAAR evolutionary patterns. No gene duplications were observed and the average ω (d_n/d_s) for primate TAARs was higher than that for other mammalian orthologs, implying relaxed selective constraints. Pseudogenization events were likely to be accelerated by the change of nose shape in Haplorhini species. In the great apes, the TAAR gene losses by natural selection might have occurred possibly due to a role in susceptibility to psychiatric disorders.

In Chapter 4, I have performed detailed mining of the GH genes in the transcriptome of *Diabrotica virgifera virgifera*. The results showed that three types of GH family genes (GH45, GH48, and GH28) have been obtained by HGT events in the common ancestor of two coleopteran superfamilies, Chrysomeloidea and Curculionoidea. Large numbers of cellulase genes (11 GH45 and 3 GH48) and their species-specific duplications in *D. v. virgifera* could have contributed to the successful adaptation to its niche, specifically for hydrolyzing the corn starch.

In this dissertation, I showed that two multigene families are characterized with high levels of gene duplications and losses. Many multigene families are considered to be subject to concerted evolution. However, chemoreceptor especially TAAR and three GH families represent fascinating birth-and-death evolution. HGT events are the major evolutionary mechanism particularly in GH genes. Therefore, the dynamic birth-and-death process and horizontal gene transfer have played a critical role in driving the evolution of multigene families and allowed adaptation of organisms to novel environmental niches.

For my prospective studies, I will examine the origin of GPCRs. As described in Chapter 1.2.6, the evolutionary relationships of GPCRs between deuterostomes and protostomes and between metazoans and protists are unclear. It has been debated which of them share a common origin because several families of GPCRs show no significant sequence similarities to each other (Nordström et al. 2011). Furthermore, insect ORs exhibit non-canonical features such as inverted 7-TM topology (N-terminus is found in the intracellular region), acting as ligand-gated ion channels, and mediated by OBPs. Nordström et al. (2011) demonstrated that insect ORs and GRs do not share a common origin with vertebrate GPCRs. These differences between insect and vertebrate chemosensory receptors imply that insect chemoreceptors may have arisen independently from vertebrate chemoreceptors or that the losses of vertebrate type of chemoreceptors in insect and invertebrate types in vertebrates may have lost. I plan to search for the entire sets of GPCR candidates from five basal metazoans (Cnidaria, Placozoa, Porifera, Ctenophora, and Choanozoa) and three protists (Mycetozoa, Percolozoa, and Metamonada), and elucidate the possible common origin of CRs between deuterostomes and protostomes.

References

Nordström KJV, Sällman Almén M, Edstam MM, Fredriksson R, Schiöth HB. 2011. Independent HHsearch, Needleman–Wunsch-Based, and Motif Analyses Reveal the Overall Hierarchy for Most of the G Protein-Coupled Receptor Families. *Mol Biol Evol*. 28:2471-2480.

Supplementary Materials

Chapter 2

Table S2.1. The animal genomes used in Chapter 2.

Group/species	Order	Sources[a]	Coverage or version	Number of OR genes[b]	Number of TAAR genes[b]
[Euarchontoglires]					
Homo sapiens	Primate	NCBI (BUILD.37.2)	-	388 (414)[c]	6 (3)
Mus musculus	Rodentia	NCBI (BUILD.38.1)	-	1063 (328)[c]	15 (1)
Rattus norvegicus	Rodentia	NCBI (BUILD.4.1)	-	1259 (508)[c]	17 (2)
[Laurasiatheria]					
Bos taurus	Cetartiodactyla	BC	7.1×	970 (1159)[c]	21 (8)
Tursiops truncatus	Cetacea	BI	2.59×	26[c]	0 (3)
Equus caballus	Perissodactyla	BI	6.79×	NA	11 (4)
Canis familiaris	Carnivora	BI	7.6×	822 (278)[c]	2 (2)
Pteropus vampyrus	Chiroptera	BI	2.63×	672[d]	26 (10)
Myotis lucifugus	Chiroptera	BI	1.84×	659[d]	6 (1)
Sorex araneus	Insectivora	BI	1.92×	NA	9 [1] (3)
Erinaceus europaeus	Insectivora	BI	1.86×	NA	6 [2] (4)
[Afrotheria]					
Echinops telfairi	Afrosoricida	BI	1.90×	NA	9 [1] (7)
Loxodonta africana	Proboscidea	BI	1.94×	NA	9 [3] (3)
[Xenarthra]					
Dasypus novemcinctus	Cingulata	WU	2.11×	NA	5 (4)
[Marsupialia]					
Macropus eugenii	Diprotodontia	Ens	2.0×	NA	18 [1] (3)
Monodelphis domestica	Didelphimorphia	BI	6.8×	1198 (294)[c]	22 (4)
[Prototheria]					
Ornithorhynchus anatinus	Monotremata	WU	6.0×	348 (370)[c]	4 (1)
[Sauropsida]					
Gallus gallus	Galliformes	WU	6.6×	211 [89] (133)[e]	4 (1)
Taeniopygia guttata	Passeriformes	WU	6.3×	NA	1 (0)
Anolis carolinensis	Squamata	BI	6.3×	112 [4] (30)[e]	3 (0)
[Amphibia]					
Xenopus tropicalis	Anura	JGI	7.65×	824 [200] (614)[e]	7 (0)
[Teleostei]					
Takifugu rubripes	Tetraodontiformes	IMC	8.7×	47 [39] (39)[e]	18 (1)
Tetraodon nigroviridis	Tetraodontiformes	Gen	8.2×	11 [4] (19)[e]	34 (3)
Danio rerio	Cypriniformes	-	-	154 [1] (21)[e]	110 (10)[g]
[Chondrichthyes]					
Callorhinchus milii	Chimaeriformes	IMC	1.4×	1 [1] (0)[e]	2 (3)
[Agnatha]					
Petromyzon marinus	Petromyzontiformes	UCSC	Ver.2	32 [8] (27)[e]	25 (3)
[Cephalochordata]					
Branchiostoma floridae	Amphioxiformes	JGI	8.1×	31 [3] (9)[e]	0
[Urochordata]					
Ciona intestinalis	Enterogona	JGI	11×	0 (0)[e]	0
Ciona savignyi	Enterogona	ASL (v2.1)	-	0 (0)[e]	0
[Cnidaria]					
Nematostella vectensis	Actiniaria	JGI	7.8×	45[f]	0

[a]Data source abbreviations. ASL: the Arend Sidow Lab at Stanford University (http://mendel.stanford.edu/sidowlab/ciona.html), BC: Baylor College of Medicine Human Genome Sequencing Center (http://www.hgsc.bcm.tmc.edu), BI: Broad Institute at MIT (http://www.broad.mit.edu), Ens: Ensembl Genome Browser (http://www.ensembl.org), Gen: Genoscope (http://www.genoscope.cns.fr), IMC: the Institute of Molecular and Cellular Biology (http://www.imcb.a-star.edu.sg), JGI: the Joint Genome Institute (http://www.jgi.doe.gov), NCBI: National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov), WU: the Genome Sequencing Center at Washington University School of Medicine (http://genome.wustl.edu), and UCSC the University of California-San Cruz (UCSC) Genome Browser (http://genome.ucsc.edu/).

[b]Gene candidates are divided into three categories: intact, incomplete, and pseudogenes. See table 1 for the details.

[c-g]The numbers were taken from the following literatures: Nei et al. (2008)[c], Hayden et al. (2010)[d], Niimura (2009)[e], Churcher and Taylor (2011)[f], and Hashiguchi and Nishida (2007)[g].

NA: not available.

Table S2.2. The results of PAML site-model analysis for TAAR subfamilies.

TAAR subfamily[a]	**ω (M0)**	**2Δ*ln*L**[b]		**Positively selected sites**[c]
		M2a–M1a	**M8–M7**	
TAAR1 (14)	0.1807	0 (1)	0.00038 (0.9998)	
TAAR2 (15)	0.0783	0 (1)	0.00408 (0.9980)	
TAAR3 (13)	0.0774	0 (1)	0.00532 (0.9973)	
TAAR4 (15)	0.1406	0 (1)	0.12241 (0.9406)	
TAAR5 (14)	0.1388	0 (1)	3.33783 (0.1885)	
TAAR6 (14)	0.1891	0.2721 (0.8728)	2.1408 (0.3429)	
TAAR7 (45)	0.3512	**28.3281 (<0.0001)**	**36.6892 (<0.0001)**	103$^{3.32}$ (0.69), 104$^{3.33}$ (0.74), **137$^{4.39}$ (0.97)**, 142$^{4.44}$ (0.89), **155$^{4.57}$ (1.00)**, 159$^{4.61}$ (0.85), **184 (0.99)**
TAAR8 (16)	0.2698	0 (1)	**6.84249 (0.03267)**	94 (0.59), 111$^{3.40}$ (0.78), 186 (0.62), **194$^{5.42}$ (0.95)**
TAAR9 (17)	0.1479	0 (1)	0.00024 (0.9999)	
TAAR E1 (6)	0.2835	0 (1)	0.00001 (1)	
TAAR M1 (2)	0.2444	0.0171 (0.9915)	0.06897 (0.9661)	
TAAR M2 (11)	0.3277	1.3045 (0.5209)	5.59743 (0.06089)	
TAAR M3 (9)	0.3102	0 (1)	0.32545 (0.8498)	

[a] The number of the TAAR subfamily genes tested is given in parentheses.

[b] Likelihood-ratio test statistics. *P*-values (shown in parentheses) are obtained based on a χ^2 distribution with d.f. = 2. Significant *P*-values (< 0.05) are shown in boldfaces.

[c] Positively selected amino acid sites using the Bayes Empirical Bayes inference with the model M8. The same sites were identified with the model M2a except for two sites (94 and 186). Posterior probabilities are given in parentheses, shown in boldfaces when $P > 0.95$. The position numbers are based on the alignment shown in Figure 2.10. The numbering of the Ballesteros-Weinstein scheme is shown in superscripts.

Table S2.3. The results of PAML branch-site model analysis.[a]

TAAR subfamily[b]	**Foreground branch**	**2Δ*ln*L**[c]	**Proportion of site class**	**ω**	**Positively selected sites**[d]
TAAR7 (45)	flying fox TAAR7c	3.9934 (0.0457)	0: 0.68747, 1: 0.29542, 2a: 0.01197, 2b: 0.00514	$\boldsymbol{\omega_0}$=0.11593, $\boldsymbol{\omega_1}$=1, $\boldsymbol{\omega_2}$=140.19823	A162 (0.657), I184 (0.599)
TAAR7 (45)	tenrec-elephant TAAR7	**7.2427 (0.0071)**	0: 0.69211, 1: 0.29130, 2a: 0.01167, 2b: 0.00491	$\boldsymbol{\omega_0}$=0.11524, $\boldsymbol{\omega_1}$=1, $\boldsymbol{\omega_2}$=169.33093	S161 (0.581), S177 (0.522), **S188$^{5.36}$ (0.973)**
TAAR8 (16)	mouse TAAR8a	6.0053 (0.0142)	0: 0.82235, 1: 0.17302, 2a: 0.00383, 2b: 0.00081	$\boldsymbol{\omega_0}$=0.14625, $\boldsymbol{\omega_1}$=1, $\boldsymbol{\omega_2}$=777.9954	F190$^{5.38}$ (0.935)

[a] Only the results where the given foreground branch having positive selection is supported significantly are listed. These branches are indicated with red color and arrows in Figure 2.6.

[b] The number of the TAAR subfamily genes tested is given in parentheses.

[c] Likelihood-ratio test statistics. *P*-values (shown in parentheses) are obtained based on a χ^2 distribution with d.f. = 1. *P*-values smaller than 0.01 are shown in boldfaces.

[d] Positively selected amino acid sites using the Bayes Empirical Bayes inference. Posterior probabilities are shown in parentheses, in boldfaces when $P > 0.95$. The position numbers are based on the alignment in Figure 2.10. The numbering of the Ballesteros-Weinstein scheme is shown in superscripts.

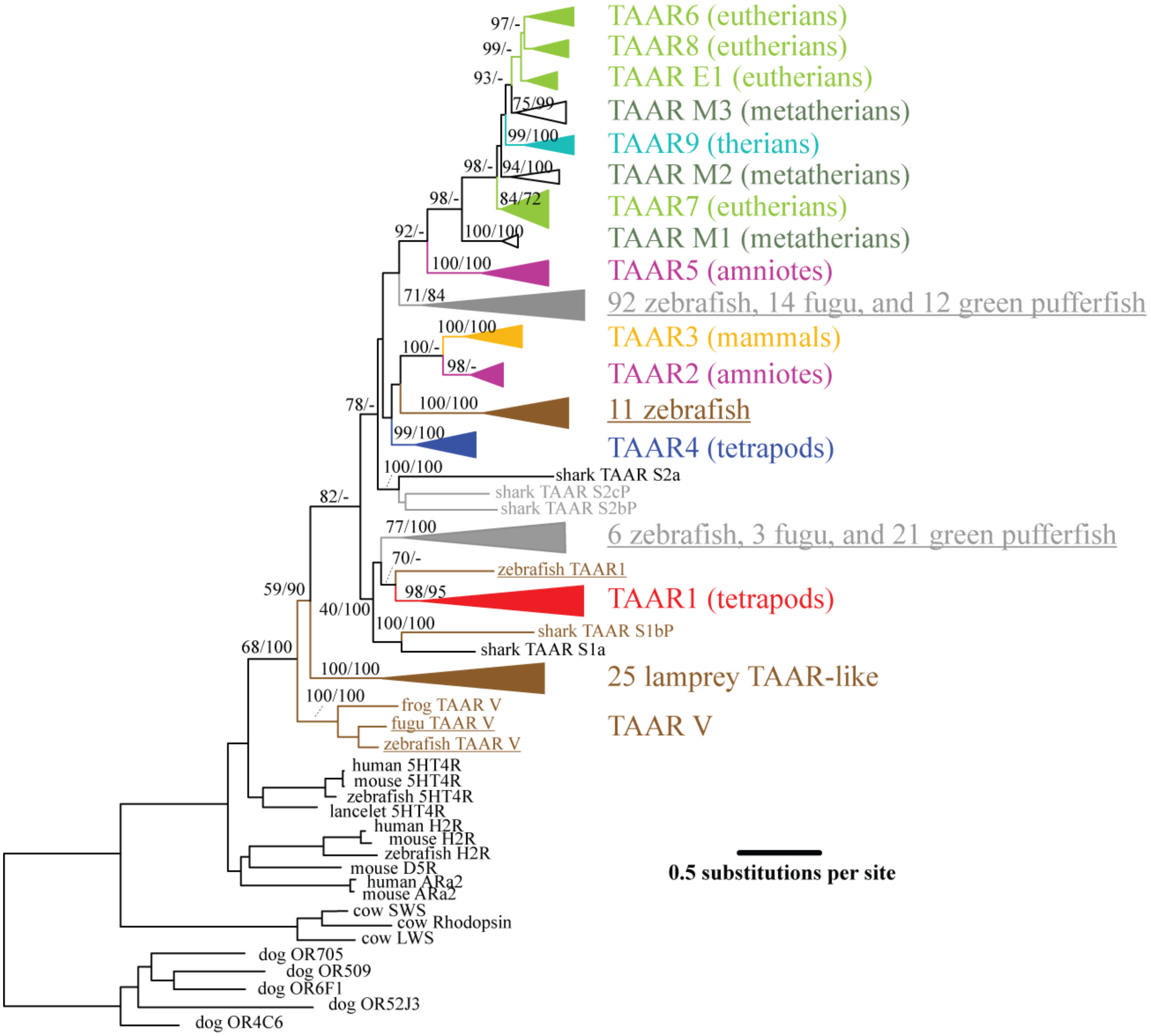

97/-
TAAR6 (eutherians)
99/-
TAAR8 (eutherians)
93/-
TAAR E1 (eutherians)
75/99
TAAR M3 (metatherians)
99/100
TAAR9 (therians)
98/-
94/100
TAAR M2 (metatherians)
98/-
84/72
TAAR7 (eutherians)
92/-
100/100
TAAR M1 (metatherians)
100/100
TAAR5 (amniotes)
71/84
92 zebrafish, 14 fugu, and 12 green pufferfish
100/100
TAAR3 (mammals)
100/-
98/-
TAAR2 (amniotes)
78/-
100/100
11 zebrafish
99/100
TAAR4 (tetrapods)
100/100
shark TAAR S2a
shark TAAR S2cP
82/-
shark TAAR S2bP
77/100
6 zebrafish, 3 fugu, and 21 green pufferfish
70/-
zebrafish TAAR1
59/90
98/95
TAAR1 (tetrapods)
40/100
100/100
shark TAAR S1bP
68/100
shark TAAR S1a
100/100
25 lamprey TAAR-like
100/100
frog TAAR V
fugu TAAR V
TAAR V
zebrafish TAAR V
human 5HT4R
mouse 5HT4R
zebrafish 5HT4R
lancelet 5HT4R
human H2R
mouse H2R
zebrafish H2R
0.5 substitutions per site
mouse D5R
human ARa2
mouse ARa2
cow SWS
cow Rhodopsin
cow LWS
dog OR705
dog OR509
dog OR6F1
dog OR52J3
dog OR4C6

Figure S2.1. The maximum-likelihood phylogeny of TAAR proteins from 25 vertebrates. Ten representative biogenic amine receptors (5HT4R: serotonin receptors, H2R: histamine receptors, D5R: dopamine receptors, and ARa2: adrenergic receptors), three cow opsins, and five representative dog olfactory receptors (ORs) are included as the outgroup. The numbers at internal branches show the bootstrap support values (%) for the maximum-likelihood phylogeny and the posterior probability (%) for the Bayesian inference phylogeny. Support values are shown only for the major internal nodes. Three metatherian-specific and one eutherian-specific TAAR groups are indicated as TAAR M1-M3 and TAAR E1, respectively. Teleost fish proteins are indicated with underline. Brown-colored branches indicate the protein lineages where all proteins have weakly conserved motifs (see Materials and Methods). Two teleost fish clusters colored in gray have TAARs with mixed types of motifs: conserved, weakly conserved, or lost. Note also that the phylogenetic placement of these teleost fish clusters is not resolved.

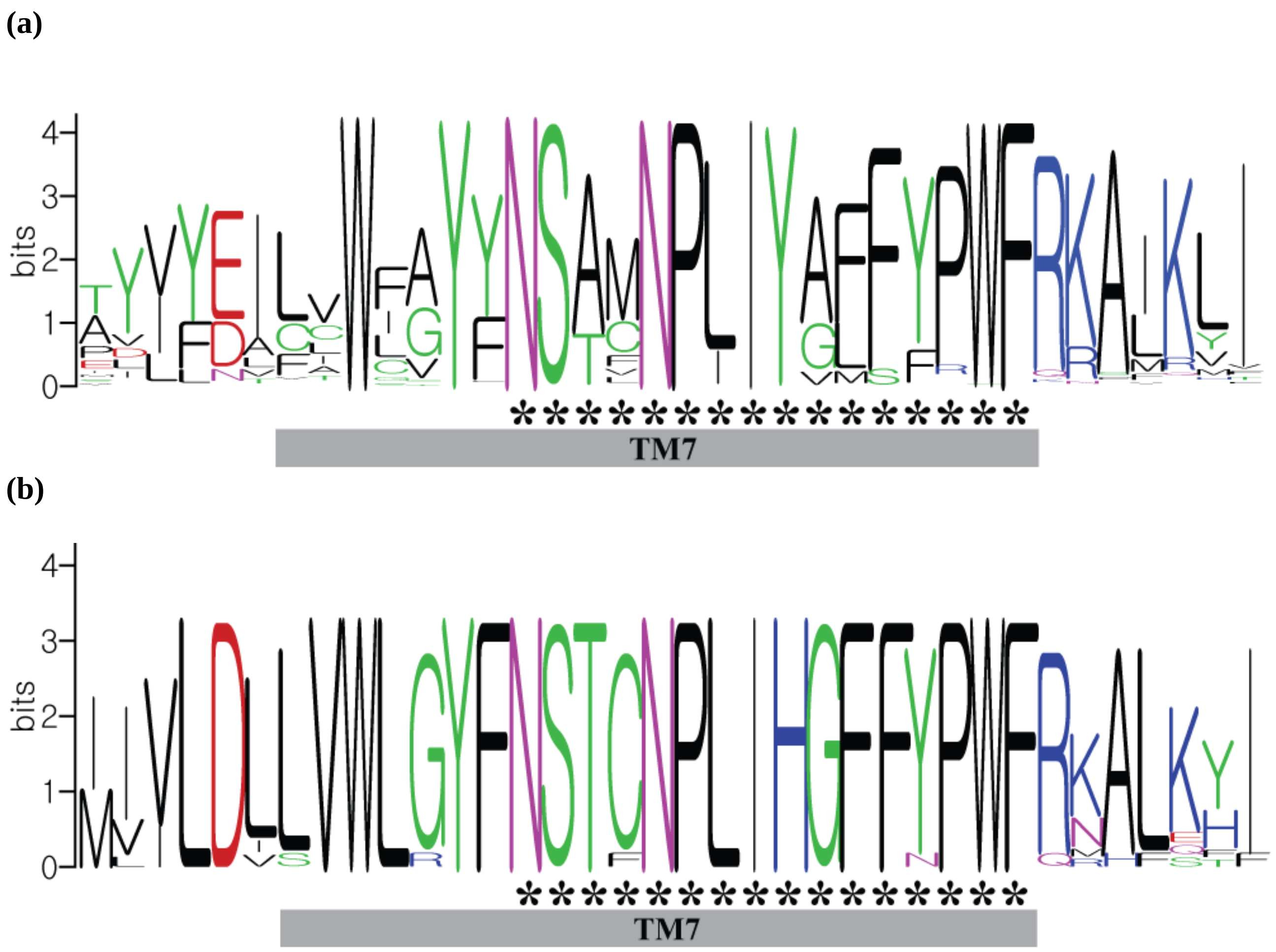


Figure S2.2. Conserved TAAR signature motifs found from TAAR subfamilies (a) and from the TAAR3 subfamily (b). Conserved amino acid patterns based on the multiple sequence alignments from positions 291 – 326 (numbering according to the mouse TAAR3: NP_001008429) are shown using the sequence logo (http://weblogo.berkeley.edu) (Crooks et al. 2004). 209 sequences from TAAR1-9, M1-M3, and E1 (a) and 13 sequences from TAAR3 (b) were included in each multiple alignment. The height of each amino-acid letter is proportional to its frequency of occurrence in a given position. The known TAAR signature motif (NSX$_2$NPX$_2$[Y/H]X$_3$YXWF) corresponds to the positions marked with *. The location of the seventh transmembrane region (indicated as TM7) was predicted using Phobius (Kall et al. 2007).

>shark TAAR S1a

LCYESVNGSCPRAIRSTGVRITLYLLAVLAILVTLFGNMLVIISIAHFKQLHTPTNYLVF
SLAIADFLLGCIVMPYSLIRSIESCWYFGILFCKLHTSFDLVLCAASIIHLCCISVDRYY
AVCDPLKYKTTITVSTVLIMICLSWALSFLVGFVIIFLELHLIEIKDFYYHEIACFGGCT
LMMGKVCALVYSTISFYFPAFIMVCIYTKIYLVAKKQARTINNLSRKVQPINEGNSIASQ
RSERKAAKTLGIVMGVFILCWSPYFVCDSIEPFIKYSTPPVLFDAFFWVGYL**NSTFNPMI**
YGFFYSWFRKALKIILTCKIFAPDSSRINLF

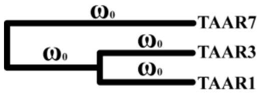
>shark TAAR S2a

MNSINLENSEDLQYCFEFNMSCPKSIRSTTTTVTMYIFITISIVITILGNSVVMISILHF
KQLQTPTNYLVLSLAFVDFLMGFFVLPFSMVRSVETCWYFGDTFCDIHSTLDVVLTTVSI
YNLCFIAIDRYYAVCEPLLYSIKMTLPMTALIITLNWLFAIIYGSCVFLSEFTKKASGHY
RTTISCKGSCIEYRFGGHMDALIVLFIPTFIILGIYLKIYFVQRKHARKIGNMPNNINSK
EEINVRVLQTKEKTAAKNQGVVMGIFVLSWLPFYLSSIINPYLNFATPPILFEAFTWFGF
F**NSAFNPVLYAFFYPWF**RTALKSILTCQILRPESSIMNLFPE

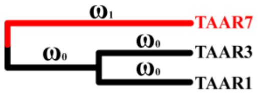
Figure S2.3. TAAR signature motifs found in the two elephant shark (*Callorhinchus milii*) TAAR protein sequences. The TAAR motif regions are highlighted with yellow. The seven transmembrane regions predicted by Phobius (Kall et al. 2007) are indicated with underline.

R1 **R2** **2Δ*ln*L**

[Test 1: TAAR1 (14) and TAAR3 (13) *vs.* TAAR7 (45)]



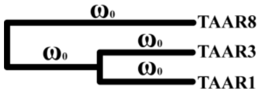
ω_0 = 0.2012



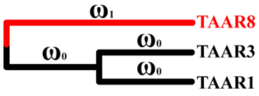
ω_0 = 0.1102, ω_1 = 0.3638

46.2655 (< **0.0001**)

[Test 2: TAAR1 (14) and TAAR3 (13) *vs.* TAAR8 (16)]



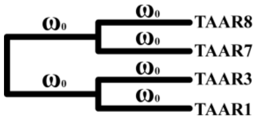
ω_0 = 0.1601



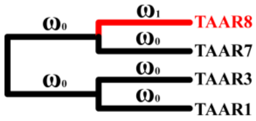
ω_0 = 0.1360, ω_1 = 0.2589

8.7584 (**0.0031**)

[Test 3: TAAR1 (14), TAAR3 (13), and TAAR7 (45) *vs.* TAAR8 (16)]



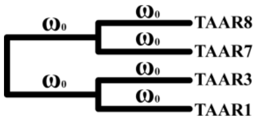
ω_0 = 0.2077



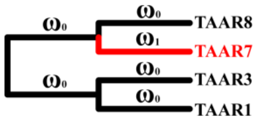
ω_0 =0.1991, ω_1 = 0.2615

1.3960 (0.2374)

[Test 4: TAAR1 (14), TAAR3 (13), and TAAR8 (16) *vs.* TAAR7 (45)]



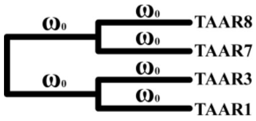
ω_0 = 0.2077



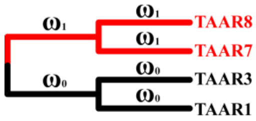
ω_0 =0.1363, ω_1 = 0.3658

36.4144 (< **0.0001**)

[Test 5: TAAR1 (14) and TAAR3 (13) *vs.* TAAR7 (45) and TAAR8 (16)]



ω_0 = 0.2077



ω_0 =0.1069, ω_1 = 0.3314

47.7429 (< **0.0001**)

Figure S2.4. PAML branch-model tests between primary amine detecting TAARs (TAAR1 and TAAR3) and tertiary amine detecting TAARs (TAAR7 and TAAR8). All tests were performed comparing the two hypotheses: R1 (a single ω for all branches) and R2 (two independent ω's: ω_1 for the red lineage and ω_0 for the black lineages). The number of the genes included in each TAAR subfamily is given in parentheses after the subfamily name. For the likelihood ratio test statistics, 2 $\boxed{ln}$L, *P*-values (shown in parentheses) are obtained based on a χ^2 distribution with d.f. = 1. Significant *P*-values (< 0.05) are shown in boldfaces.

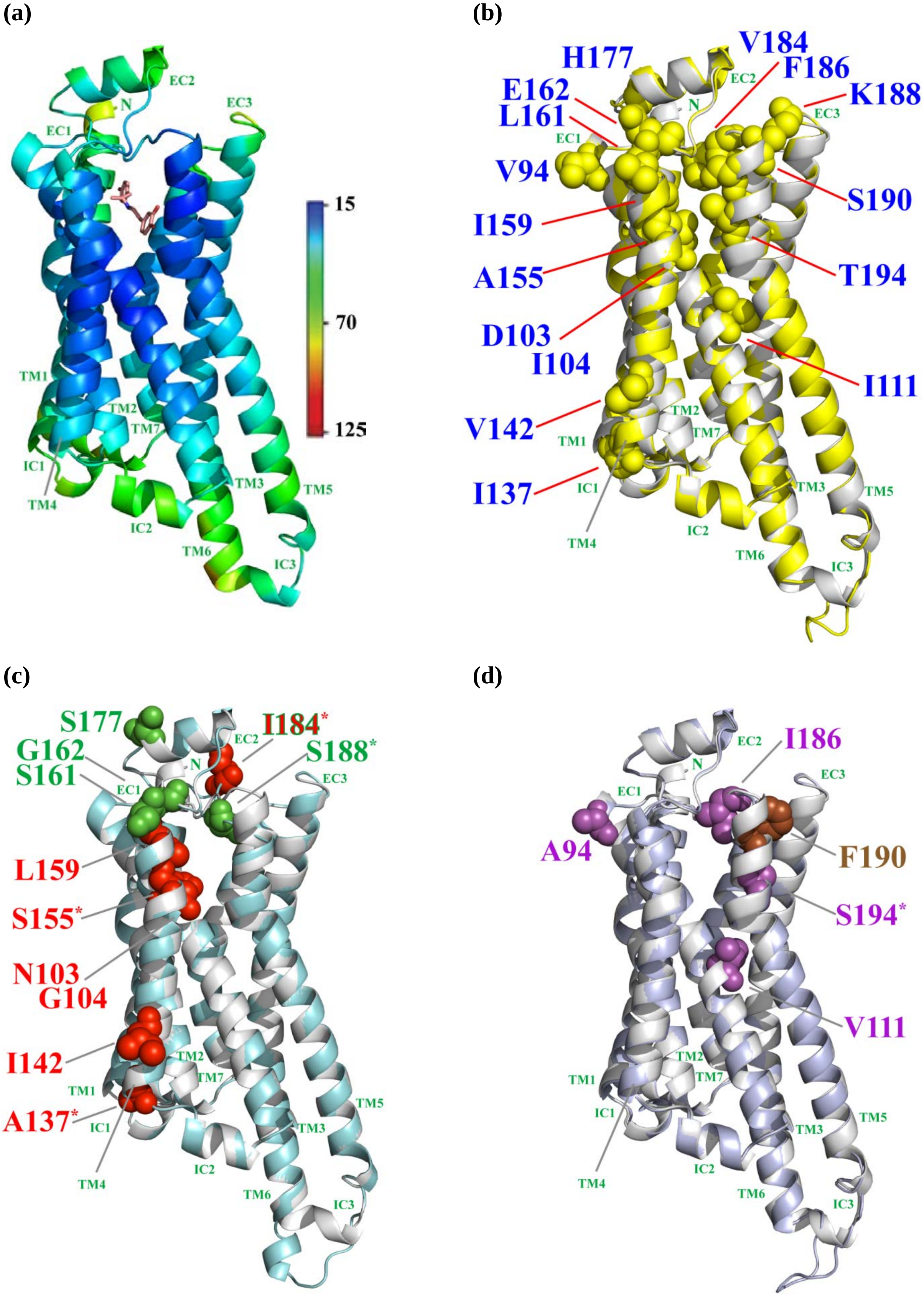

(a)
EC2
N
EC3
EC1
15
70
125
TM1
TM2
TM7
IC1
TM4
TM3
TM5
IC2
TM6
IC3
(b)
H177
V184
F186
E162
K188
L161
EC1
EC2
EC3
V94
S190
I159
A155
T194
D103
I104
I111
V142
TM1
TM2
TM7
I137
IC1
TM3
TM5
TM4
IC2
TM6
IC3
(c)
S177
I184*
G162
S188*
S161
EC1
EC2
EC3
N
L159
S155*
N103
G104
I142
TM2
TM7
TM1
A137*
IC1
TM5
TM3
TM4
IC2
TM6
IC3
(d)
EC2
I186
N
EC3
EC1
A94
F190
S194*
V111
TM1
TM2
TM7
IC1
TM5
TM3
TM4
IC2
TM6
IC3

Figure S2.5. Modeling of the 3D-structure of TAAR proteins. The same template, the B-chain of the turkey β_1-adrenergic receptor (a: β_1AR, PDB: 4AMJ), was selected by SWISS-MODEL (http://swissmodel.expasy.org; Arnold et al. 2006) for modeling protein structures of the human TAAR1 (b: NP_612200), elephant TAAR7a (c: XP_003404143), and mouse TAAR8a (d: NP_001010830) (all E-values < 0.001; their sequence similarities against β_1AR, P07700, are 49.3%, 46.2%, and 43.9%, respectively). The 3D-structure of the 4AMJ (a) is color-coded based on the temperature factors (B-factors), ranging from 15.74 (blue) to 124.95 (red) (see color scale in the figure). The average B-factor is 45.52. The ligand for the β_1AR, dobutamine, is shown with the stick model. Note that the template protein contains truncations at N-terminus, third intracellular loop, and C-terminus as well as some thermostabilizing point mutations to improve expression and to obtain crystals (Warne et al. 2012). None of these positions were, however, overlapped with those identified to be under positive selection (see Fig. 10 for more details). Predicted protein structures of the human TAAR1 (b: yellow), elephant TAAR7a (c: cyan), and mouse TAAR8a (d: light blue) are superimposed with the template structure (gray) using PyMOL. The QMEAN4 Z-scores given by SWISS-MODEL were -8.27, -8.02, and -8.37 (raw scores: 0.234, 0.250, and 0.228), respectively. The overall root-mean-square deviations (RMSDs) given by PyMOL were 0.054 Å, 0.055Å, and 0.054 Å, respectively. The N-terminal 15, 25, 23 amino acids (aa) and the C-terminal 19, 16, and 16 aa, respectively, were excluded from the modeling due to insufficient sequence similarity. Positive-selection sites identified by the PAML analysis in elephant TAAR7a (c) and mouse TAAR8a (d) are indicated by red and purple (site models) and by green and brown (branch-site models). Position 184 in elephant TAAR7a was identified by both site and branch-site models. Sites identified with higher than 0.95 posterior probabilities are indicated with asterisks. See Tables 2.3 and 2.4 for details on PAML analysis. All amino acid sites corresponding to these positive-selection sites are also mapped on human TAAR1 by yellow spheres for comparison (b). All amino acid position numbers are according to the human TAAR1 sequence. The transmembrane (TM) and internal/external loop (IC1-3 and EC1-3) regions as well as the N- terminal (N) are labeled in each structure. The C-terminal is invisible locating behind TM1. See Figure 2.10 for the alignment and more detailed information on these sequences.

```
                                  1         .       21         .        41         .
                                                     1.30                          1.61 2.34
4AMJ           MG[                           ]AELLSQQWEAGMSLLMALVVLLIVAGNVLVIAAIGsTQR
β1AR           MGDGWLPPDCGPHNRSGGGATAAPTGSRQVSAELLSQQWEAGMSLLMALVVLLIVAGNVLVIAAIGRTQR
humanTAAR1     -----------------MMPFCHNIINISCVKNNWSNDVRASLYSLMVLIILTTLVGNLIVIVSISHFKQ
mouseTAAR3     ---------MDLIYIPEDLSSCPKFGNKSCPPTNRSFRVRMIMYLFMTGAMVITIFGNLVIIISISHFKQ
elephantTAAR7a --------MSTELSPPAPVQLCYENLNGSCVKTSYSPGPRVMLYLVFGSGAVLAVFGNLLVMISMLHFKQ
mouseTAAR8a    ---------MTSNFSQPALQLCYENTNGSCIKTPYSPGPRVILYMVYGFGAVLAVCGNLLVVISVLHFKQ
                                                          <-----------TM1--------><---

             54     61          .        81        .        101        .        121
                                               2.67  3.21
4AMJ           LQTLTNLFITSLACADLVvGLLVVPFGATLVVRGTWLWGSFLCElWTSLDVLCVTASIETLCVIAIDRYLA
β1AR           LQTLTNLFITSLACADLVMGLLVVPFGATLVVRGTWLWGSFLCECWTSLDVLCVTASIETLCVIAIDRYLA
humanTAAR1     LHTPTNWLIHSMATVDFLLGCLVMPYSMVRSAEHCWYFGEVFCKIHTSTDIMLSSASIFHLSFISIDRYYA
mouseTAAR3     LHSPTNFLILSMATTDFLLGFVIMPYSMVRSVESCWYFGDSFCKFHASFDMMLSLTSIFHLCSIAIDRFYA
elephantTAAR7a LHSPANFLIASLACADFLVGVTVMPFSTVRSVESCWYFGEIYCTFHSCFNGSFCYASIFHLCFISVDKFIA
mouseTAAR8a    LHSPANFLIASLASADFLVGISVMPFSMVRSIESCWYFGDAFCSLHSCCDVAFCYSSVLHLCFISVDRYIA
                                                      △         △△      △
               IC1--><---------TM2---------><------EC1-------><---------TM3---------><

             125    .         .        151                  .        181        .
                        3.65 4.39                  4.61                       5.36
4AMJ           ITSPFRYQSLMTRARAKVIICTVWAISALVSFLPIMMHWWRDEDPQ-ALKCYQDPGCCDFVTNRAYAIASS
β1AR           ITSPFRYQSLMTRARAKVIICTVWAISALVSFLPIMMHWWRDEDPQ-ALKCYQDPGCCDFVTNRAYAIASS
humanTAAR1     VCDPLRYKAKMNILVICVMIFISWSVPAVFAFGMIFLELNFKGAEEIYYKHVHCRGGCSVFFSKISGVLTF
mouseTAAR3     VCDPLHYTTTMTVSMIKRLLAFCWAAPALFSFGLVLSEANVSGMQS-YEILVACFNFCALTFNKFWGTILF
elephantTAAR7a VTDPLIYPTRFTASVSGICIAFSWLLSILYSFSLLCSGANETGLEE-LVSGLSCVGGCQIAVNSNWVFVNF
mouseTAAR8a    VTDPLVYPTKFTVSVSGICISISWILPLVYSSAVFYTGISAKGIES-LVSALNCVGGCQIVINQDFVLISF
                           ▲   △            ▲   △ △△              △      ▲ △ ▲ △   ▲
               -------IC2-------><-------TM4-------><------------EC2------------><----

             196  201         .                  231                               241
                                    5.65                                             6.26
4AMJ           IISFYIPLLIMIFVaLRVYREAKEQIRKIDR[                          ]ASKRKTSRVMlM
β1AR           IISFYIPLLIMIFVYLRVYREAKEQIRKIDRCEGRFYGSQEQPQPPPLPQHQPILGNGRASKRKTSRVMAM
humanTAAR1     MTSFYIPGSIMLCVYYRIYLIAKEQARLISDANQKLQIGL----------------------EMKNGISQS
mouseTAAR3     TTCFFTPGSIMVGIYGKIFIVSRRHARALSDMPANTKG------------------------AVGKNLSKK
elephantTAAR7a V-LFFIPTLVMIIVYSKIFLVAKQQARKIESLSNKTETSS----------------------DSYKDRVAK
mouseTAAR8a    L-LFFIPTLVMIILYSKIFLVAKQQAVKIETSVSGNRGESSS--------------------ESHKARVAK
               -----TM5---------><----------------------------IC3----------------------

             245    .        261          .                 .        301        .
                                                6.62 7.28                 7.54
4AMJ           REHKALKTLGIIMGVFTLCWLPFFLVNIVNVFNRDLVPDWLFVaFNWLGYANSAmNPIIYCR-SPDFRKAF
β1AR           REHKALKTLGIIMGVFTLCWLPFFLVNIVNVFNRDLVPDWLFVFFNWLGYANSAFNPIIYCR-SPDFRKAF
humanTAAR1     KERKAVKTLGIVMGVFLICWCPFFICTVMDPFLHYIIPPTLNDVLIWFGYLNSTFNPMVYAFFYPWFRKAL
mouseTAAR3     KDRKAAKTLGIVMGVFLACWLPCFLAVLIDPYLDYSTPIIVLDLLVWLGYFNSTCNPLIHGFFYPWFRKAL
elephantTAAR7a RERKAAKTLGIAVIAFLISWLPYFIDIVIDAFLGFITPTYIYEILVWFAYYNSAMNPLIYAFFYPWFRKAI
mouseTAAR8a    RERKAAKTLGVTVVAFMVSWLPYTIDALVDAFMGFITPAYVYEICCWGTYYNSAMNPLIYAFFFPWFKKAI
               ------><------TM6-------><-------EC3-------><---------TM7--------->

             316  321         .
4AMJ           KRLLaFPRKADRRLhhhhhh[
β1AR           KRLLCFPRKADRRLHAGGQPAPLPGGFISTLGSPEHSPGGTWSDCNGGTRGGSESSLEERHSKTSRSESKM
humanTAAR1     KMMLFGKIFQKDSSRCKLFLELSS-----------------------------------------------
mouseTAAR3     QFIVSGKIFRSNSDTANLFPEAH------------------------------------------------
elephantTAAR7a KLIITGKVLRENSSTTNLFSD--------------------------------------------------
mouseTAAR8a    KLILSGEILKGHSSTANLFSE--------------------------------------------------

4AMJ                                                                    ]
β1AR           EREKNILATTRFYCTFLGNGDKAVFCTVLRIVKLFEDATCTCPHTHKLKMKWRFKQHQA
humanTAAR1     -----------------------------------------------------------
mouseTAAR3     -----------------------------------------------------------
elephantTAAR7a -----------------------------------------------------------
mouseTAAR8a    -----------------------------------------------------------
```

Figure S2.6. Multiple alignment of the four TAAR and the turkey β$_1$-adrenergic receptor proteins. Protein sequences of two primary amine detecting TAARs (human TAAR1: NP_612200 and mouse TAAR3: NP_001008429) and two tertiary amine detecting TAARs (elephant TAAR7a: XP_003404143 and TAAR8a: NP_001010830) are aligned with the sequence of the turkey β$_1$-adrenergic receptor (β$_1$AR: P07700). The position number at the top of the alignment starts at the beginning of the human TAAR1 sequence. Position numbers based on the scheme proposed by Ballesteros and Weinstein (1995) are also shown diagonally for the start and end of each transmembrane region of β$_1$AR. Approximate regions for transmembranes (TM1-TM7), intracellular loops (IC1-IC3), and extracellular loops (EC1-EC3) are indicated below each alignment block. The first lines of the alignment show the sequence the protein structure (4AMJ) is based on. In order to improve expression and to obtain crystals, eight thermostabilizing point mutations, a His-tag at the C-terminus, and truncations (at N-terminus, third intracellular loop, and C-terminus) were introduced (Warne et al. 2012). These changes are indicated by lower cases and square brackets in the 4AMJ sequence. Residues assigned for alpha helices in 4AMJ are shown with white letters on black background. 26 residues suggested to involve with agonist binding to the β$_1$AR are shown with blue background (Warne et al. 2011; Warne et al. 2012). For the β$_1$AR and TAAR protein sequences, residues predicted to be in transmembrane regions by Phobius (Kall et al. 2007) are shown with gray background. The residues surrounding the main and minor ligand-binding pockets in the β$_1$AR are shown with cyan and magenta background (Nygaard et al. 2009; Rosenkilde et al. 2010). 29 ligand-binding sites identified by Kleinau et al. (2011) are shown with green background in the human TAAR1. Among them, the residues conserved among human TAARs (including both primary amine detectors and tertiary amine detectors), adrenergic receptors, as well as other biogenic amine receptors are shown with red fonts. Those in the human TAAR1 identical or similar to the residues in the corresponding position of biogenic amine receptors are shown with yellow fonts. Positively selected sites identified by PAML analysis are shown with triangles below the alignment: red and green are sites identified by the site and branch-site models, respectively, in TAAR7, and purple and brown are sites identified by the site and branch-site models, respectively, in TAAR8. Closed triangles indicate sites identified with posterior probabilities higher than 0.95. See Tables 2.3 and 2.4 for details.

	103(3.32)	104(3.33)	137(4.39)	142(4.44)	155(4.57)	159(4.61)	161	162	177	184	188(5.36)
human 1	D	I	I	V	A	I	L	E	H	V	K
mouse 3	D	M	V	R	S	V	S	E	A	L	K
mouse 7a	E	G	A	K	S	L	T	G	T	I	Q
mouse 7b	D	V	A	K	G	L	T	G	T	L	Q
mouse 7d	E	G	A	K	S	L	T	G	T	I	Q
mouse 7e	D	V	A	K	G	I	T	G	T	L	Q
mouse 7f	D	V	A	K	G	I	T	G	T	I	Q
rat 7a	E	G	A	K	S	L	T	G	T	I	Q
rat 7b	D	G	A	K	S	F	T	G	T	I	Q
rat 7c	E	G	A	K	S	L	T	G	T	V	Q
rat 7d	D	V	A	K	G	I	T	G	T	I	Q
rat 7e	D	V	A	K	S	L	T	G	T	L	Q
rat 7g	D	I	A	K	G	I	T	G	T	V	Q
rat 7h	D	M	A	K	G	I	T	G	S	I	Q
cow 7g	D	V	V	M	S	L	T	G	T	I	Q
cow 7f	E	G	V	M	S	F	T	G	T	I	Q
cow 7e	D	G	M	M	S	L	S	G	S	F	Q
cow 7d	D	G	E	M	S	L	T	G	T	I	Q
cow 7c	E	G	V	M	T	L	T	G	T	I	Q
cow 7b	E	G	V	M	T	L	T	G	T	I	Q
cow 7k	D	G	M	M	S	L	T	G	T	V	Q
horse 7b	E	G	A	M	S	L	T	G	T	I	Q
flying fox 7o	D	G	I	I	S	L	T	A	T	I	Q
flying fox 7n	D	G	I	T	Y	F	T	G	T	V	Q
flying fox 7m	D	G	I	M	N	F	T	G	T	G	Q
flying fox 7l	D	G	T	I	S	Y	T	G	T	A	Q
flying fox 7k	D	G	I	M	N	L	T	G	T	G	Q
flying fox 7j	D	G	I	I	S	F	T	G	T	I	Q
flying fox 7i	D	G	T	L	S	F	T	G	T	G	Q
flying fox 7h	D	G	I	M	Y	I	T	G	T	A	Q
flying fox 7q	D	G	I	M	N	L	T	G	T	V	Q
flying fox 7g	D	G	I	M	N	L	T	G	T	G	Q
flying fox 7f	D	G	I	M	S	F	T	G	T	G	Q
flying fox 7e	D	G	T	I	S	L	T	G	T	G	Q
flying fox 7d	D	G	I	I	Y	F	T	G	T	G	Q
flying fox 7c	D	G	I	I	S	L	T	**A**	T	**I**	Q
flying fox 7b	D	G	I	M	S	F	T	G	T	G	Q
flying fox 7a	D	G	I	M	N	L	T	G	T	G	Q
common shrew 7a	E	G	T	L	S	V	T	G	T	V	Q
common shrew 7b	E	G	T	L	S	V	T	G	T	I	Q
hedgehog 7b	E	G	M	V	S	L	T	G	T	I	Q
hedgehog 7a	D	G	M	M	S	F	T	G	T	I	Q
tenrec 7a	E	G	P	S	S	L	**S**	G	**S**	M	**S**
tenrec 7b	E	G	P	S	S	L	**S**	G	**S**	I	**S**
elephant 7a	N	G	A	I	S	L	**S**	G	**S**	I	**S**
elephant 7b	E	G	A	V	S	L	**S**	G	**S**	V	**S**
armadillo 7a	E	G	T	L	S	L	T	G	T	V	Q

	94	111(3.40)	186	190(5.38)	194(5.42)
human 1	V	I	F	S	T
mouse 3	S	I	F	W	L
human 8	K	V	V	W	D
mouse 8a	A	V	I	**F**	S
mouse 8b	A	A	V	W	D
mouse 8c	A	A	V	W	S
rat 8a	T	L	V	W	D
rat 8b	T	L	I	W	S
rat 8c	T	A	V	W	S
cow 8b	R	L	I	W	S
cow 8d	R	L	V	W	S
cow 8e	R	L	I	W	S
horse 8b	R	L	V	W	D
horse 8a	R	L	V	W	D
flying fox 8a	R	L	V	W	D
tenrec 8	Q	L	V	W	D
elephant 8a	Q	V	V	W	D
elephant 8b	Q	V	V	W	D

Figure S2.7. Alignments of the positively selected sites identified in TAAR7 (a) and TAAR8 (b). The position numbers correspond to those given in Figure 2.10. The residues identified by the branch-site models are shown in boldface. The amino acids are color-coded based on their physico-chemical properties using the Taylor color scheme (Taylor 1997). Color-coding is roughly as follows: red for negatively charged (D and E), blue/blueish for positively charged (R, K, and H), green/yellow green for hydrophobic (I, F, V, L, M, and A), blueish green for aromatic (W and Y), purple for large polar (N and Q), and reddish/orange for small (G, T, and S).

Chapter 3

Table S3.1. Taxonomic classification and the genomes used in Chapter 3.

Group/species	Common names	Family	Sources[a]	Quality (Version)
Haplorhini				
[Simiiformes (Catarrhini)]				
Homo sapiens	Human	Hominidae	NCBI	BUILD.37.2
Pan troglodytes	Chimpanzee	Hominidae	WU, Ens	6X
Pan paniscus	Bonobo	Hominidae	NCBI	26X
Gorilla gorilla gorilla	Gorilla	Hominidae	Ens	gorGor3 (Release 63)
Pongo pygmaeus abelii	Sumatran Orangutan	Hominidae	WU, Ens	6X (ver2.0.2)
Nomascus leucogenys	White-cheeked gibbon	Hylobatidae	Ens	Release 68
Macaca mulatta	Rhesus monkey	Cercopithecidae	BC, Ens	6X
Papio hamadryas	Hamadryas baboon	Cercopithecidae	BC	5.3X (Pham_1.0)
[Simiiformes (Platyrrhini)]				
Callithrix jacchus	Common marmoset	Cebidae	WU, Ens	6X (ver.3.2)
[Tarsiiformes]				
Tarsius syrichta	Tarsier	Tarsiidae	BI, Ens	1.82X
Strepsirrhini				
[Lemuriformes]				
Microcebus murinus	Gray mouse lemur	Cheirogaleidae	BI, Ens	1.93X
[Lorisiformes]				
Otolemur garnettii	small-eared bushbaby	Galagidae	BI, Ens	1.5X
Scandentia				
Tupaia belangeri	northern treeshrew	Tupaiidae	BI, Ens	2X

[a] Data source abbreviations. BC: Baylor College of Medicine Human Genome Sequencing Center (http://www.hgsc.bcm.tmc.edu), BI: Broad Institute at MIT (http://www.broad.mit.edu), Ens: Ensembl Genome Browser (http://www.ensembl.org), NCBI: National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov), and WU: the Genome Sequencing Center at Washington University School of Medicine (http://genome.wustl.edu).

Table S3.2. The results of PAML site –model analysis and likelihood ratio statistics for heterogeneity within primate TAAR subfamily.

TAAR subfamily[a]	**M0**	**ω**	**M3**	**2Δ*ln*L**[b]	***P*-values**[c]	**ω in non-primate mammals**	**ω in only haplorhines**
TAAR1 (11)	-2651.6	0.2879	-2644.6	**13.91662**	**0.0076**	0.1802	0.3805
TAAR2 (7)	-2517	0.149	-2511.6	**10.96983**	**0.0269**	0.0759	0.2113
TAAR3 (5)	-2382.1	0.1181	-2376.5	**11.15437**	**0.0249**	0.0774	0.2019
TAAR4 (6)	-2684.7	0.2844	-2665.9	**37.71547**	**<0.0001**	0.1406	0.3325
TAAR5 (9)	-2223.6	0.2546	-2222.9	7.57626	0.1084	0.1384	0.2861
TAAR6 (6)	-2399.2	0.2591	-2391.5	**15.26471**	**0.0042**	0.1951	0.4747
TAAR8 (3)	-2161.2	0.314	-2156.4	**9.65311**	**0.0467**	0.2615	NA
TAAR9 (5)	-1944.4	0.1186	-1944.4	0.03297	0.9999	0.1490	0.7818
Overall Average		0.2232				0.1523	0.3813

[a]The number of the TAAR genes used in the test is given in parentheses.

[b]Likelihood-ratio test statistics.

[c]*P*-values are obtained based on a χ^2 distribution with d.f. = 4. *P*-values smaller than 0.05 are shown in boldfaces.

NA: not available.

Table S3.3. PAML branch-model tests between Haplorhini TAARs and Strepsirrhini TAARs.[a]

	R1	**R2**	**2Δ*ln*L (*P*-values)**
TAAR1 (11)	ω_0 = 0.2879	ω_0 = 0.2371, ω_1 = 0.3509	2.1373 (*P*=0.1438)
TAAR2 (7)	ω_0 = 0.1490	ω_0 = 0.0906, ω_1 = 0.2081	7.7244 (***P*=0.0054**)
TAAR3 (5)	ω_0 = 0.1181	ω_0 = 0.0542, ω_1 = 0.1941	14.3442 (***P*=0.0002**)
TAAR4 (6)	ω_0 = 0.2788	ω_0 = 0.1702, ω_1 = 0.3586	6.9279 (***P*=0.0085**)
TAAR5 (9)	ω_0 = 0.2546	ω_0 = 0.0001, ω_1 = 0.2975	1.4747 (*P*=0.2246)
TAAR6 (6)	ω_0 = 0.2591	ω_0 = 0.1948, ω_1 = 0.4498	7.1277 (***P*=0.0076**)
TAAR8 (3)	ω_0 = 0.314	ω_0 = 0.3927, ω_1 = 0.0001	3.0658 (*P*=0.0799)
TAAR9 (5)	ω_0 = 0.1335	ω_0 = 0.0722, ω_1 = 0.1972	9.5753 (***P*=0.0019**)

[a] All tests were performed comparing two hypotheses: R1 (a single ω for all branches) and R2 (two independent ω's: ω_0 for the Strepsirrhini lineages and ω_1 for the Haplorhini lineages). The number of the genes included in each TAAR subfamily is given in parentheses after the subfamily name. For the likelihood ratio test statistics, 2*ln*L (*P*-values shown in parentheses) are obtained based on a χ^2 distribution with d.f. = 1. Significant *P*-values (< 0.01) are shown in boldfaces.

Table S3.4. PAML branch-model tests within Haplorhini TAARs.[a]

(a) tarsier *vs*. others

	R1	**R2**	**2Δ*ln*L (*P*-values)**
TAAR2 (5)	ω_0 = 0.2113	ω_0 = 0.1465 for tarsier, ω_1 = 0.4547 for Catarrhini	8.2062 (***P*=0.0042**)
TAAR3 (3)	ω_0 = 0.2019	ω_0 = 0.9894 for tarsier, ω_1 = 0.2019 for Cercopithecoidea	0 (*P*=0.9975)
TAAR4 (4)	ω_0 = 0.3325	ω_0 = 0.0001 for tarsier, ω_1 = 0.6845 for Catarrhini	7.464 (***P*=0.0063**)

(b) marmoset *vs*. others.

	R1	**R2**	**2Δ*ln*L (*P*-values)**
TAAR1 (9)	ω_0 = 0.3805	ω_0 = 0.4569 for marmoset, ω_1 = 0.3805 for Catarrhini	0 (*P*=0.9984)
TAAR5 (8)	ω_0 = 0.2962	ω_0 = 0.2311 for marmoset, ω_1 = 0.3409 for Catarrhini	0.7867 (*P*=0.3751)

(c) All other comparisons

	R1	**R2**	**2Δ*ln*L (*P*-values)**
TAAR1 (8)	ω_0 = 0.3975	ω_0 = 0.5231 for Cercopithecoidea, ω_1 = 0.3828 for Hominoidea	0.1902 (*P*=0.6628)
TAAR1 (6)	ω_0 = 0.4964	ω_0 = 0.0001 for gibbon, ω_1 = 0.5245 for Hominidae	0.0864 (*P*=0.7688)
TAAR1 (5)	ω_0 = 0.5724	ω_0 = 0.0001 for orangutan, ω_1 = 0.8696 for Homininae	1.0595 (*P*=0.3033)
TAAR1 (4)	ω_0 = 0.8172	ω_0 = 2.8961 for gorilla, ω_1 = 0.3048 for Hominini	3.5329 (*P*=0.0602)
TAAR1 (3)	ω_0 = 0.2782	ω_0 = 999 for Pan, ω_1 = 0.1845 for human	1.9438 (*P*=0.1633)
TAAR2 (4)	ω_0 = 0.55	ω_0 = 0.4082 for Cercopithecoidea, ω_1 = 0.8115 for Homininae	1.0483 (*P*=0.3059)
TAAR2 (2)	ω_0 = 0.7804	ω_0 = 0.0001 for gorilla,	0.0061 (*P*=0.9378)

		ω_1 = 240.1312 for human	
TAAR3 (2)	ω_0 = 0.5199	ω_0 = 749.49 for rhesus, ω_1 = 0.5195 for baboon	0.0034 (*P*=0.9537)
TAAR4 (3)	ω_0 = 0.6193	ω_0 = 0.5422 for Cercopithecoidea, ω_1 = 859.85 for orangutan	0.0559 (*P*=0.8131)
TAAR5 (7)	ω_0 = 0.344	ω_0 = 1.885 for Cercopithecoidea, ω_1 = 0.293 for Hominidae	3.6163 (*P*=0.0572)
TAAR5 (5)	ω_0 = 0.4654	ω_0 = 0.3732 for orangutan, ω_1 = 0.5964 for Homininae	0.5048 (*P*=0.4774)
TAAR5 (4)	ω_0 = 0.5231	ω_0 = 999 for gorilla, ω_1 = 0.277 for Hominini	3.1409 (*P*=0.0764)
TAAR5 (3)	ω_0 = 0.2291	ω_0 = 100.0409 for Pan, ω_1 = 0.1369 for human	3.7107 (*P*=0.0541)
TAAR9 (2)	ω_0 = 0.7818	ω_0 = 999 for orangutan, ω_1 = 0.0001 for human	0.0247 (*P*=0.875)

[a] All tests were performed comparing two hypotheses: R1 (a single ω for all branches) and R2 (two independent ω's). The number of the genes included in each TAAR subfamily is given in parentheses after the subfamily name. For the likelihood ratio test statistics, 2*ln*L (*P*-values shown in parentheses) are obtained based on a χ^2 distribution with d.f. = 1. Significant *P*-values (< 0.01) are shown in boldfaces.

Table S3.5. The results of PAML branch-site model analysis.[a]

(a) Analysis including pseudogenes.

TAAR subfamily[b]	**Foreground branch**	**2Δ*ln*L[c]**	**Proportion of site class**	**ω**	**Positively selected sites[d]**
TAAR2 (11)	human TAAR2	3.881624 (0.04882)	0: 0.62047, 1: 0.04397, 2a: 0.31335, 2b: 0.02221	$\boldsymbol{\omega_0}$=0.11208, $\boldsymbol{\omega_1}$=1, $\boldsymbol{\omega_2}$=3.04205	K2 (0.848), Y99$^{3.28}$ (0.846), L130$^{3.59}$ (0.782)
TAAR6 (11)	chimpanzee TAAR6	**49.323874 (<0.0001)**	0: 0.71059, 1: 0.27347, 2a: 0.01151, 2b: 0.00443	$\boldsymbol{\omega_0}$=0.20664, $\boldsymbol{\omega_1}$=1, $\boldsymbol{\omega_2}$=999.0000	P7 (0.739), V96$^{3.25}$ (**0.997**), L97$^{3.26}$ (**0.993**), C114$^{3.43}$ (**0.982**), A115$^{3.44}$ (**0.954**), C195$^{5.43}$ (0.726)

(b) Analysis excluding pseudogenes.

TAAR subfamily[b]	**Foreground branch**	**2Δ*ln*L[c]**	**Proportion of site class**	**ω**	**Positively selected sites[d]**
TAAR2 (7)	human TAAR2	5.031986 (0.025)	0: 0.90688, 1: 0.05283, 2a: 0.03807, 2b: 0.00222	$\boldsymbol{\omega_0}$=0.09728, $\boldsymbol{\omega_1}$=1, $\boldsymbol{\omega_2}$=53.19005	K2 (0.944), Y99$^{3.28}$ (0.944), L130$^{3.59}$ (0.833), I257 (0.948), C327 (**0.997**), I332 (0.947)
TAAR6 (6)	chimpanzee TAAR6	**38.049413 (<0.0001)**	0: 0.84791, 1: 0.13563, 2a: 0.0142, 2b: 0.00227	$\boldsymbol{\omega_0}$=0.12816, $\boldsymbol{\omega_1}$=1, $\boldsymbol{\omega_2}$=848.75853	P7 (0.73), E15 (0.847), T16 (0.642), L17 (0.845), V96$^{3.25}$ (**0.997**), L97$^{3.26}$ (**0.992**), C114$^{3.43}$ (**0.99**), A115$^{3.44}$ (**0.964**), C195$^{5.43}$ (0.858)

[a] Only the results where the given foreground branch having positive selection is supported significantly are listed. These branches are indicated with red color and arrows in Figure 3.2.

[b] The number of the TAAR subfamily genes tested is given in parentheses.

[c] Likelihood-ratio test statistics. *P*-values (shown in parentheses) are obtained based on a χ^2 distribution with d.f. = 1. *P*-values smaller than 0.01 are shown in boldfaces.

[d] Positively selected amino acid sites using the Bayes Empirical Bayes (BEB) inference. Posterior probabilities are shown in parentheses, in boldfaces when *P* > 0.95. The position numbers are based on the human TAAR1 sequence.

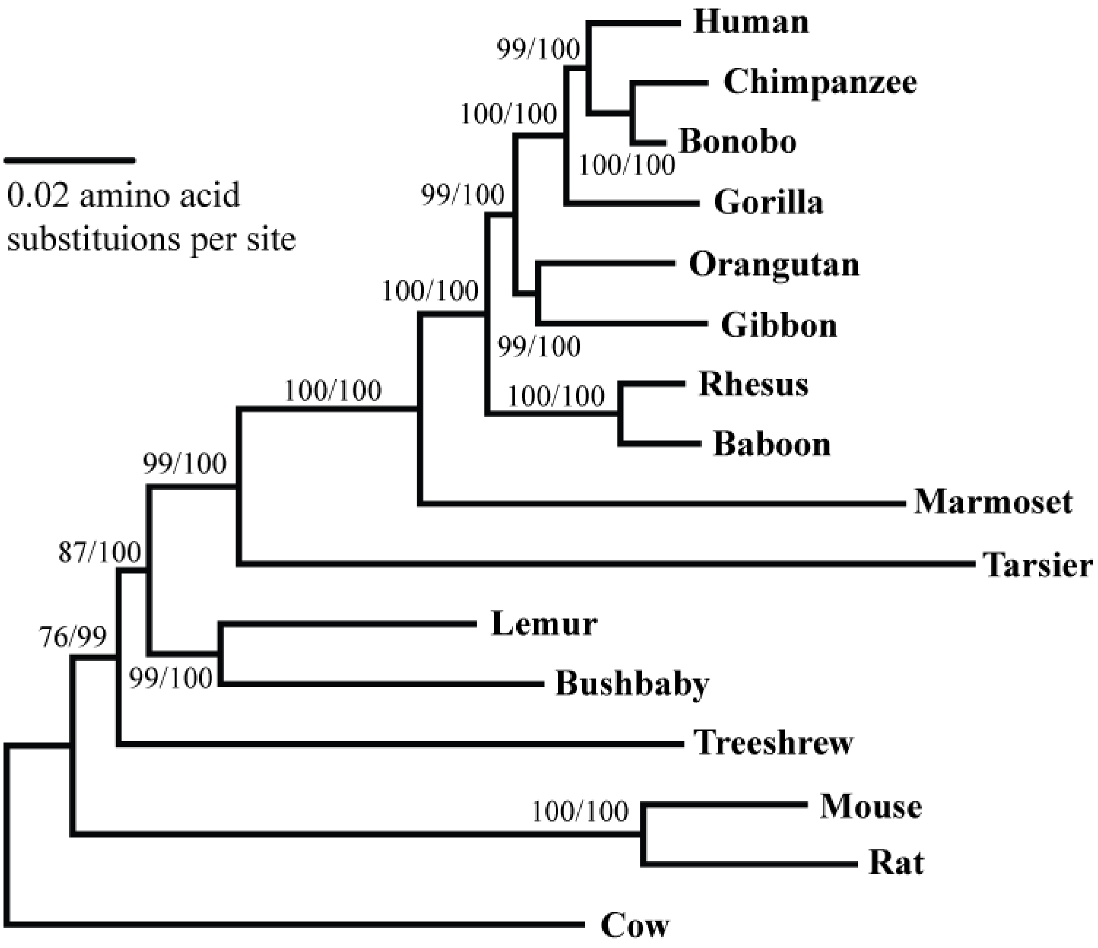


Figure S3.1. The maximum-likelihood phylogeny inferred from the supermatrix dataset. The sequences are based on concatenated 8 orthologous alignments (2809 amino acids). Note that the codons to have frame-shifts and in-frame stop codons were removed in the alignments. Cow is used as the outgroup. The numbers at internal branches show the bootstrap support value (%) for maximum-likelihood method and the posterior probability (%) for the Bayesian inference method.

(a)

```
           808 .           820            .          840             .          860
Human2       GTT TTC TTA TTA TGT TGG TTT CCT TGT TTC TTC ACA ATT TTA TTG GAT CCC TTT TTG AAC
Chimpanzee2P GTT TTC TTA TTA TGT TGG TTT CCT TCT TTC TTC ACA ATT TTA TTG GAT CCC TT- TTG AAC
Bonobo2P     GTT TTC TTA TTA TGT TGG TTT CCT TCT TTC TTC ACA ATT TTA TTG GAT CCC TT- TTG AAC
Gorilla2     GTT TTC TTA TTA TGT TGG TTT CCT TGT TTC TTC ACA ATT TTA TTG GAT CCC TTT TTG AAC
Orangutan2P  GTT TTC TTA TTA TGT TGG TTT CCT TAT TTC TTC ACA ATT TTA TTG GAT CCC TTT TTG AAC
Rhesus2      GTT TTC TTA TTA TGT TGG TTT CCT TGT TTC TTC ACA ATT TTA TTG GAT CCC TTT TTG AAC
Baboon2      GTT TTC TTA TTA TGT TGG TTT CCT TGT TTC TTC ACA ATT TTA TTG GAT CCC TTT TTG AAC
Marmoset2P   GTT TTC TTA TTA TGT TGG TTT CCT TGT TTC TTC ACA ATT TTA TTG GAT CCC TTT TTG AAC
Tarsier2     GTT TTC TTA TTA TGT TGG TTT CCC TGT TTC TTC ACG ATT TTA TTG GAT CCC TTT TTG AAC
Lemur2       GTT TTC TTA TTA TGT TGG TTT CCC TGT TTC TTC ACA ATT TTA TTG GAT CCC TTT CTG AAT
Bushbaby2    GTT TTC TTA TTG TGC TGG TTT CCT TGT TTC TTC ACC ATT TTG TTG GAT CCC TTT CTG AAC
Treeshrew2   GTT TTC TTA CTA TGT TGG TTT CCC TGT TTT TTT ACA ATT TTG TTA GAT CCC TTT TTG AAT
```

(b)

```
           125     .           138    .          150             .          170
Human3P      ATG ATT ATC C-- ACT A-- TTC GGA AAC TTG GTT ATA ATG GTT TCC ATA TCG CAT
Chimpanzee3P ATG ATT ATC --- CCT A-- TTC GGA AAC TTG GTT ATA ATG GTT TCC ATA TCG CAT
Bonobo3P     ATG ATT ATC --- CCT A-- TTC GGA AAC TTG GTT ATA ATG GTT TCC ATA TCG CAT
Gorilla3P    ATG ATT ATC --- ACT A-- TTT GGA AAC TTG GTT ATA ATG GTT TCC ATA TCG CAT
Orangutan3aP ATG ATT ATC --- ACT ACT TTT GGA AAC TTG GTT ATA ATG GTT TCC ATA TCG CAT
Orangutan3bP ATG ATT ATC --- ACT ACT TTT GGA AAC TTG GTT ATA ATG GTT TCC ATA TCA CAT
Rhesus3      ATG ATT ATC --- ACT ATT TTT GGA AAC TTG GTT ATA ATA GTT TCT ATA TCG CAT
Baboon3      ATG ATT ATC --- ACT ATT TTT GGA AAC TTG GTT ATA ATA GTT TCT ATA TCG CAT
Marmoset3P   ATG GTT ATA --- ACT ATT TTT GGC AAC TTG GTT ATA ATG GTT TCC ATG TCT CAT
Tarsier3     ATG GTT ATC --- ACT GTT TTG GGA AAC TTG GTT ATC ATG ACT TCC ATA TCA CAC
Lemur3       ATG ATT ATC --- ACT ATT TTT GGG AAT CTG GTT ATA ATG ATT TCC ATA TCA CAT
Bushbaby3    ATG GTT ATC --- ACC ATT TTT GGA AAT CTG GTT ATA ATG ATT TCC ATA TCC CAT
Treeshrew3   ATG GTT ATC --- ACT ATC TTT GGA AAC TTG GTT ATA ATG ATT TCC ATA TCA CAT
```

(c)

```
           734       .       748     .          760            .           780            .
Human4P      TCA GAA AGC AAA AAA A-- AAG GCA TCC TCT AAA ACA GAA AGC AAG GCC ACC AGG ACC TTA
Chimpanzee4P TCA AAA AGC AAA AAA A-- AAG TCA TCC TCT AAA ACA GAA AGC AAG GCC ACC AGG ACC TTA
Bonobo4P     TCA AAA AGC AAA ATA A-- AAG GCA TCC TCT AAA ACA GAA AGC AAG GCC ACC AGG ACC TTA
Gorilla4P    TCA GAA AGC AAA AAA AA- AAG GCA TCC TCT AAA ACA GAA AGC AAG GCC ACC AGG ACT TCA
Orangutan4   TCA GAA AGC AAA AAA --- AAG GCA TCC TCT AAA ACA GAA AGC AAG GCC ACC AGG ACC TTA
Gibbon4P     TCA GAA AGC AAA AAA --- AAG GCA TCC TCT AAA ACA GAA AGC AAG GCC ACC AGG ACC TTA
Rhesus4      TCA GAA AGC AAA AAA --- AAG ACA TCC TCT AAA ACA GAA AGC AAG GCC ACC AGG ACC TTA
Baboon4      TCA GAA AGC AAA AAA --- AAG ACA TCC TCT AAA ACA GAA AGC AAG GCC ACC AGG ACC TTA
Marmoset4P   TCA GAA AGC AAA AA- --- AAG GCA TCC TCT AAA AGA GAA AGC AAG GCC ACC AGG ACC TTA
Tarsier4     TCA GAA ACC CAA AGG --- AAG GGA TCA TCC AGA AGG GAA AAC AAG GCC ACC AGG ACC TTA
Lemur4       ACA GAA AGC AAA ATG --- AAG GCA TCA TCC AAA AGA GAA AGC AAG GCC ACC AAG ACC CTC
Bushbaby4    CAG AAC --- AGA ATG --- AAG GCA TCG TCC AAG AAA GAA AGC AAG GCC ACC AAG ACC TTA
Treeshrew4   TCA GAA AGC AAA GCA --- AAG --- TCC TCC AAA AAG GAA AGC AAG GCC ACC AAG ACC CTG
```

(d)

```
           700            .          720             .          740            .
Human8       AGT AGC AAA GTA GAA TCA TCC TCA GAG AGT TAT AAA ATC AGA GTG GCC AAG AGA GAG AGG
Chimpanzee8P AGT AGC AAA GTA GAA TCA TCC TCA GAG AGT TAT AAA ATC AGA GTG GCC --G AGA GAG AGG
Bonobo8P     AGT AGC AAA GTA GAA TCA TCC TCA GAG AGT TAT AAA ATC AGA GTG GCC --G AGA GAG AGG
Gorilla8aP   AGT AGC AAA GTA GAA TCA TCC TCA GAG AGT TAT AAA ATC AGA GTG GCC --G AGA GAG AGG
Gorilla8bP   AGT AGC AAA GTA GAA TCA TCC TCA GAG AGT TAT AAA ATC AGA GTG GCC --G AGA GAG AGG
Orangutan8P  AGT AGC AAA GTA GAA TCA TCT TCA GAG AGT TAC AAA ATC AGA GTG GCC AAG AGA AAG AGG
Rhesus8P     AGT AGC AAA GTA GAA TCA TCC TCA G-- AGT TAC AAA ATC AGA GTG GCC --G AGA GAG AGG
Baboon8P     AGT AGC AAA GTA GAA TCA TCC TCA G-- AGT TAC AAA ATC AGA GTG GCC --A AGA GAG AGG
marmoset8P   AGT AGC AAA GTA AAA TCA TCT TCA GAG AGT TAC AAA ATC AGA GTG GCC --G AGA GAG AGG
Tarsier8P    AGT AAC AAA ACA GAA TCA TCC TCA GAG AGT TGC AAA GCC AGA GTG GCC AAG AGA GAG AGA
Lemur8       AGT AGC AAA ACA GAA TCA TCT TCA GAG AGT TAC AAA GCC AGA GTG GCC AAG AGG GAG AGA
Bushbaby8    AGT TGC AAA GCA GAA TCT TCC TCA GGG AGT TAC AAA GCC AGA GTG GCC AGA AGG GAG AGA
Treeshrew8aP GGA AGC AAA ACA AAA TCA ACT TCA GAG AGT TAC AAA GCT AGA GTG GCC AAG AGA GAG AGA
Treeshrew8b  GGT AGC AAA ACA GAA TCA TCT TCA GAG AGT TAC AAA GCT AGA GTG GCC AAG AGA GAG AGA
Treeshrew8cP GGT AGC AAA ACA GAA TCA TCC TCA GAG AGT TAC AAA GCT AGA GTG GCC AAG AGA GAG AGA
Treeshrew8d  GGT AGC AAA ACA GAA TCA TCT TCA GAG AGT TCC AAA GCT AGA GTG GCC AAG AGA GAG AGA
Treeshrew8e  GGT AGC AAA ACA GAA TCA TCT TCA GAG AGT TAC AAA GCT AGA GTG GCC AAG AGA GAG AGA
Treeshrew8f  GGT AGC AAA ACA AAA TCA TCC TCA GAG AGT TAC AAA GCT AGA GTG GCC AAG AGA GAG AGA
Treeshrew8g  GGT AGC AAA ACA GAA TCA TCC TCA GAG AGT TCC AAA GCT AGA GTG GCC AAG AGA GAG AGA


           763        .          780             .          800                .
Human8       GCA GCT AAA ACC CTG GGG GTC ACG GTA CTA GCA TTT GTT ATT TCA --- TGG TTA CCG TAT
Chimpanzee8P GCA GCT AAA ACC CTG GGG GTC ACG GTA CTA GCA TTT GTT ATT TCA --- TGG TTA CCG TAT
Bonobo8P     GCA GCT AAA ACC CTG GGG GTC ACG GTA CTA GCA TTT GTT ATT TCA --- TGG TTA CCG TAT
Gorilla8aP   GCA GCT AAA ACC CTG GGG GTC ACG GTA CTA GCA TTT GTT ATT TCA --- TGG TTA CCG TAT
Gorilla8bP   GCA GCT AAA ACC CTG GGG GTC ACG GTA CTA GCA TTT GTT ATT TCA --- TGG TTA CCG TAT
Orangutan8P  GCA GCT AAA ACC CTC GGG GTC ACG GTA ATA GCA TTT GTT ATT TCA --- TGG TTA CCG TAT
Rhesus8P     GCA GCT AAA ACC CTG GGG GTC ATG GTA ATA GCA TTT ATT ATT TCA --- TGG TTA CCG TAT
Baboon8P     GCA GCT AAA ACC CTG GGG GTC ATG GTA ATA GCA TTT ATT ATT TCA --- TGG TTA CCA TAT
marmoset8P   GCA GCT AGA ATC CT- GGG GTC ATG GTA ATA GCA TTT ATT ATT TCA --- TGG TTA CCG TGT
Tarsier8P    GCA GCT AAA ACC CTG GGA GTC ATG GTG GTA GCA TTT ATG AAT TTT G-- TGG TTA CCA TAT
Lemur8       GCA GCT AAA ACT CTG GGG GTC ACA GTG GTA GCA TTT ATG ATT TCA --- TGG TTA CCA TAT
Bushbaby8    GCA GCT AAA ACT CTG GGG GTC ACA GTA GTA GCA TTT ATG ATT TCA --- TGG TTA CCA TAC
Treeshrew8aP GCA GCG AAA ACC CTG GGG GTT ACA GTG ATA GCT TTC ATG ATT TCA --- TGG TTA CCA TAC
Treeshrew8b  GCA GCG AAA ACC CTG GGG GTC ACA GTG CTA GCC TTC ATG ATT TCA --- TGG TTA CCG TAC
Treeshrew8cP GCA GCG AAA ACC CTG GGG GTC ACA GTG ATA GCC TTC ATG ATT TCA --- TGG TTA CCG TAC
Treeshrew8d  GCA GCA AAA ACC CTG GGG GTT ACA GTG ATA GCC TTC ATG ATT TCA --- TGG TTA CCG TAC
Treeshrew8e  GCA GCG AAA ACC CTG GGG GTC ACA GTG CTA GCC TTC ATG ATT TCA --- TGG TTA CCG TAC
Treeshrew8f  GCA GCG AAA ACC CTG GGG GTC ACA GTG CTA GCC TTC ATG ATT TCA --- TGG TTA CCG TAC
Treeshrew8g  GCA GCA AAA ACT CTG GGG GTC ACA GTG ATA GCC TTC ATG ATT TCA --- TGG TTA CCA TAC
```

Figure S3.2. Partial nucleotide sequence alignment of TAAR2 (a), TAAR3 (b), TAAR4 (c), and TAAR8 (d) from 12 primate and northern treeshrew genomes. The position number at the top of the alignment is based on the human TAAR nucleotide sequence. The indel events are highlighted with yellow. Dashes indicate alignment gaps.

Chapter 4

Table S4.1. Summary statistics for *D. v. virgifera* transcriptome sequencing and assembly.

egg 0-10 day total RNA (Illumina, paired-end)	
Total number of paired-end reads before filtering (length)	38,657,737 (2,899,330,275 bp)
Number of paired-end reads that entered assembly after > Q20 filtering (length)	15,162,017 (1,137,151,275 bp)
Assembly program used	Trinity (2013-02-25)
Total number of contigs	72,638
Average contig length (range)	825 bp (201 – 13,911 bp)
N50 length	1,357 bp
Assembly program used	Velvet/Oasis (ver. 1.2.03)
Total number of contigs	56,135
Average contig length (range)	583 bp (100 – 10,434 bp)
N50 length	850 bp
Assembly program used	Mira (ver. 3.4.0)
Total number of contigs	69,815
Average contig length (range)	520 bp (100 – 13,526 bp)
N50 length	850 bp
Third larval midgut RNA (Illumina, paired-end)	
Total number of paired-end reads before filtering (length)	76,202,715 (5,715,203,625 bp)
Number of paired-end reads that entered assembly after > Q20 filtering (length)	44,852,488 (3,363,936,600 bp)
Assembly program used	Trinity (2013-02-25)
Total number of contigs	72,325
Average contig length (range)	859 bp (201 – 17,831 bp)
N50 length	1,435 bp
Assembly program used	Velvet/Oasis (ver. 1.2.03)
Total number of contigs	96,215
Average contig length (range)	635 bp (100 – 17,673 bp)
N50 length	1,180 bp
Third larval midgut RNA (Roche 454)	
Total number of reads before filtering (length)	664,431 (361,187,777 bp)
Number of reads that entered assembly after filtering (removing the adapters and > Q20)	415,742 (210,423,467 bp)
Assembly program used	Trinity (2013-02-25)
Total number of contigs	37,181
Average contig length (range)	614 bp (201 – 5,044 bp)
N50 length	743 bp

Assembly program used	Newbler (ver. 2.5)
Total number of contigs	45,994
Average contig length (range)	535 bp (51 – 4,098)
N50 length	595 bp
Assembly program used	Velvet/Oasis (ver. 1.2.03)
Total number of contigs	165,361
Average contig length (range)	322 bp (100 – 5,807 bp)
N50 length	481 bp
Assembly program used	Mira (ver. 3.4.0)
Total number of contigs	57,923
Average contig length (range)	762 bp (100 – 3,032 bp)
N50 length	853 bp
Neonates RNA (Illumina Hi-seq, paired-end)	
Total number of paired-end reads before filtering (length)	1,347,291,731 (136,076,464,831 bp)
Number of paired-end reads that entered assembly after > Q30 filtering (length)	721,697,288 (72,891,426,088 bp)
Assembly program used	Trinity
Total number of contigs	155,787
Average contig length (range)	937 bp (201 – 25,737 bp)
N50 length	1,817 bp

Table S4.2. Summary of *D. v. virgifera* transcriptome sequencing and assemblies.

		Egg	**Larval midgut**	**Larval midgut**	**Neonates**
Sequencing platform		Illumina Genome Analyzer II	Illumina Genome Analyzer II	454 Titanium	Illumina HiSeq2000
Read length		75 bp	75 bp	NA	101 bp
Total reads[a]		15.1 M	44.8 M	415,742	721 M
Total number of contigs (average length)	Trinity	72,638 (825 bp)	72,325 (859 bp)	37,181 (614 bp)	155,787 (914 bp)
	Newbler	NA	NA	45,994 (535 bp)	NA
	Velvet/Oasis	56,135 (520 bp)	96,215 (635 bp)	165,361 (322 bp)	NA
	Mira	69,815 (520 bp)	NA	57,923 (762 bp)	NA

[a]Numbers of reads used for assembly after filtering. M: million paired-end.

Table S4.3. Summary statistics for hybrid and pooled-data assembly of *D. v. virgifera* transcriptome.

Hybrid (454 + Illumina) assembly of third larval midgut	
Assembly program used	Trinity (2013-02-25)
Total number of contigs	81,858
Average contig length (range)	862 bp (201 – 17,831 bp)
N50 length	1,396 bp
Assembly program used	Velvet/Oasis (ver. 1.2.03)
Total number of contigs	133,276
Average contig length (range)	425 bp (100 - 16,733 bp)
N50 length	675 bp
Hybrid (454 + Illumina) assembly of egg and third larval midgut	
Assembly program used	Trinity (2012-03-17)
Total number of contigs	101,915
Average contig length (range)	662 bp (201 – 13,611 bp)
N50 length	1,006 bp
The pooled read dataset (egg + neonates + third larval midgut)	
Assembly program used	Trinity (2013-02-25)
Total number of contigs	163,871
Average contig length (range)	914 bp (201 – 31,064 bp)
N50 length	1,396 bp

Table S4.4. Expression analysis of *D. v. virgifera* GH genes identified in Chapter 4.[a]

Genes	Sequence length (bp)	Egg (38,657,737)[b]		Larval midgut (40,096,158)[b]		Neonate (21,864,095)[b]	
		read[c]	RPKM	read[c]	RPKM	read[c]	RPKM
GH45-1	717	1	0.04	888	30.89	4053	258.59
GH45-2	738	2	0.07	485	16.39	343	21.26
GH45-3	726	3	0.11	142	4.88	9	0.57
GH45-4	738	5	0.18	26621	899.63	2135	132.30
GH45-5	726	0	0	79	2.71	12	0.76
GH45-6[d]	612	1	0.04	538	21.92	15	1.12
GH45-7	717	2	0.07	228852	7960.36	61252	3907.24
GH45-8	714	6	0.19	86	3.00	14	0.89
GH45-9	717	0	0	2473	86.02	796	50.78
GH45-10	714	4	0.14	67084	2343.23	11641	745.75
GH45-11	732	2	0.07	656	22.35	31	1.94
GH48-1	1923	1	0.01	26	0.33	3	0.07
GH48-2[d]	372	1	0.06	2	0.13	0	0
GH48-3	1923	2	0.02	965	12.51	302	7.18
GH28-1	1095	0	0	319	7.27	171	7.14
GH28-2	1095	0	0	1098	25.01	384	16.04
GH28-3	1113	0	0	175	3.92	36	1.48
GH28-4	1095	21	0.50	221	5.03	3	0.13
GH28-5	1059	0	0	294	6.92	55	2.38
GH28-6	1089	2	0.05	11750	269.09	707	29.69
GH28-7	1068	1	0.02	192	4.48	18	0.77
GH28-8[d]	849	5	0.15	3	0.08	1	0.05
GH28-9	1098	6	0.14	1105	25.10	184	7.66
GH28-10[d]	810	22	0.70	4	0.12	1	0.06
GH28-11	1059	4	0.09	8	0.18	2	0.09
GH28-12	1065	1	0.02	18	0.42	3	0.13
GH28-13	1101	1	0.02	15	0.33	28	1.16
GH28-14[d]	366	0	0	0	0.14	0	0.07
GH16-1	1353	40	0.76	34	0.63	173	5.85
GH16-2	1500	301	5.19	48	0.80	53	1.62
GH31-1	1237	1870	39.11	12256	247.10	1396	51.62
GH31-2	1338	94	1.82	217	4.04	70	2.39
GH5[d]	317	5	0.41	0	0	0	0

[a]Genes whose RPKM is less than 0.3 or the number of reads is less than 10 are considered as not expressed and marked with grey shade.

[b] The total numbers of paired-end reads before filtering are shown in parentheses. Note that although the reads were not filtered, we used 0 mismatch to map the reads to the assembled transcriptome. Thus reads that included any ambiguity including unknown nucleotide ‘N’ were not counted.

[c] The number of paired-end reads mapped.

[d] The partial ORFs, not including from start to stop codons.

Table S4.5. Beetle species names used in Chapter 4 and accession numbers

	GH9	GH5	GH45	GH48	GH28	GH11	GH16	GH31
[Chrysomeloidea]								
Chrysomela tremulae	None	None	ADU33285.1 ADU33286.1	ADU33283.1 ADU33284.1	ACP18831.1 ADU33275.1 ADU33276.1 ADU33277.1 ADU33278.1 ADU33279.1 ADU33280.1 ADU33281.1 ADU33282.1	-	-	-
Gastrophysa viridula	None	ADU33333.1	ADU33334.1	ADU33335.1 ADU33336.1 ADU33337.1	ADU33338.1 ADU33339.1 ADU33340.1 ADU33341.1 ADU33342.1 ADU33343.1 ADU33344.1	-	-	-
Leptinotarsa decemlineata	None	None	ADU33345.1 ADU33346.1 ADU33347.1 ADU33348.1 ADU33349.1 ADU33350.1 ADU33351.1	ADU33352.1 ADU33353.1 ADU33354.1	ADU33355.1 ADU33356.1 ADU33357.1 ADU33358.1 ADU33359.1 ADU33360.1 ADU33361.1 ADU33362.1 ADU33363.1 ADU33364.1 AEX93414.1	-	-	-
Phaedon cochleariae	-	-	CCJ09450.1 CCJ09451.1 CCJ09452.1 CCJ09453.1 CCJ09454.1 CCJ09455.1 CCJ09456.1	-	CCJ09441.1 CCJ09442.1 CCJ09443.1 CCJ09444.1 CCJ09445.1 CCJ09446.1 CCJ09447.1 CCJ09448.1 CCJ09449.1	CAA76932.1 YP_001984213.1	-	-
Gastrophysa atrocyanea	-	-	-	BAE94320.1 BAE94321.1	-	-	-	-
Callosobruchus maculatus	None	ADU33271.1 ADU33272.1 ADU33273.1 ADU33274.1	None	None	ADU33264.1 ADU33265.1 ADU33266.1 ADU33267.1 ADU33268.1 ADU33269.1 ADU33270.1	-	-	-
Apriona germari	-	AAX18655.1	AAU44973.1 AAR22385.1	-	-	-	-	-
Psacothea hilaris	-	BAB86867.1	-	-	-	-	-	-
Anoplophora chinensis	-	AFN89566.1	AFN89565.1	-	-	-	-	-
Oncideres albomarginata chamela	-	ADI24131.1	ADI24132.1	-	-	-	-	-
[Curculionoidea]								

Dendroctonus ponderosae	None	None	ADU33287.1 ADU33288.1 ADU33289.1 ADU33290.1 ADU33291.1 ADU33292.1 ADU33293.1 ADU33294.1 ADU33295.1	ADU33296.1 ADU33297.1 ADU33298.1 ADU33299.1 ADU33300.1 ADU33301.1	ADU33302.1 ADU33303.1 ADU33304.1 ADU33305.1 ADU33306.1 ADU33307.1 ADU33308.1 ADU33309.1 ADU33310.1 ADU33311.1 ADU33312.1 ADU33313.1 ADU33314.1 ADU33315.1 ADU33316.1 ADU33317.1 ADU33318.1 ADU33319.1 ADU33320.1	-	AEE61901.1 ENN74344.1 ENN74953.1 ENN76697.1 ENN78830.1 ENN78831.1 ENN83076.1 ENN83093.1	ENN70227.1 ENN70228.1
Ips pini	-	None	CB408544	-	-	-	-	-
Hypothenemus hampei	-	ACU52526.1 ACU52527.1	-	-	-	-	-	-
Sitophilus oryzae	None	None	ADU33246.1 ADU33247.1 ADU33248.1 ADU33249.1 ADU33250.1	ADU33251.1 ADU33252.1	ADU33253.1 ADU33254.1 ADU33255.1 ADU33256.1 ADU33257.1 ADU33258.1	-	-	-
Otiorhynchus sulcatus	-	-	-	CAH25542.1	-	-	-	-
[Tenebrionoidea]								
Tribolium castaneum	XP_001810693.1	None	None	None	None.	-	XP_972063	XP_973339.2 XP_973373.1 XP_973404.1
[Diptera]								
Drosophila melanogaster	0	0	0	0	0	-	-	-

'-': not determined.

```
Dvir_GH45-1   MLSL-----KIAVAILSLAG--VTIAQDLTPIPGGKSGDGVTTRYWDCCAPSCAWYPRIHTQNGVPIQTCKADGVTPSDK
Dvir_GH45-2   MKLLVA---IAFLGYVAAGSFGRCPGPDIVPIPGGLSGDGITTTYWDCCAQTCAHRQNVKTDNGIPVQTCAIDGTTNITI
Dvir_GH45-3   MKYLVV---ITFLGYVAAAS--SDRSPEIVPIPGGISGDGITTRYWDCCAPSCAYYGFIKTKNGIPDQTCQIDGVTNSTK
Dvir_GH45-4   MYTGIVNIFLVSIAIVTASS--KESSPDIVAIPGGLRGDAITTRYWDCCVVSCSWDANVHTKNRQPVKSCQKNGATYSTR
Dvir_GH45-5   MKTFTV---FASLIVFGASL--KEPSPEIIPVPGGLSGDAVTTRYWDCCGVSCSWDGIVHTKNGIPVRSCEKDGKTYSTK
Dvir_GH45-6   ------------------------------------------RYWDCCKPTCSWPGNVNYKT--PVKSCQHDGVTAI--
Dvir_GH45-7   MKIAILV--SALVALAVATP--LEQSPEIKFIEKGISGEGTTTRYWDCCKPSCSWRGNVHTPSGVPVASCDRSGVNRV--
Dvir_GH45-8   M---IFN--CFIFSVVLAVT--LAYSPEIKKIVGGKSGYGTTTRYWDCCKPSCAWKENIKTPDMEPIATCATDGVTVV--
Dvir_GH45-9   MIFII----FSLLAFVGLAP--SIDALELTPVEGGLSGNGSTSRYWDCCKPACAWPSNV-PHSPRPVTSCKADGITPI--
Dvir_GH45-10  MIPLPI---LLVLAVATSIK--AEVSPDIIAVPNGLSGKGITTRYWDCCKPSCAWADNVNTPDKQPLKSCRVDGEAVA--
Dvir_GH45-11  MKYTITS--LLLLAAYVAATSLNNQNIVIKKIPGGLSGVGTTTRYWDCCKATCSWPGNVEYKK--PVKACQADGENAN--


Dvir_GH45-1   DLNA-QSGC--EVGGVAYTCTNQSPKIINETLAYTFVAASFAGGLDY-ADCCICLVMDFKG-KLAGKRLLAQVTNTGEA-
Dvir_GH45-2   DQNGIVSGC--RVGGQAFACSNQQPYVVSDTLALGWSAASFTGGIDN-SKCCSCFLLSFKD-QLAGKQMLVQLVNSGTD-
Dvir_GH45-3   DNNA-QSGC--EQGGVAYTCSNQQPSVINDTLAFGWAAASFQGGIDT-SKCCHCILLSFKD-QLAGKQMLVQIVNTGSD-
Dvir_GH45-4   ENNG-NSVCYPDHPGNAYVCNNNSPFVVNSTLAYGFAGVSFQGGADV-EHCCHCYLLSFKG-KLQGKQMLVQTINTGAD-
Dvir_GH45-5   ENNA-QSTCW-NENGPAFTCSNQVPFVINSTLSYGFAAVSFVGSTDT-GHCCQCYLLKFQG-QLKDRELLVQAINTGSD-
Dvir_GH45-6   DPET-QSGC---VGGGAYVCTNQAQRSVNDSIALGFVAAKFIHS-NR-NMCCSCIVFRFKPAELAGKQMVLQVTNTGDDD
Dvir_GH45-7   DANA-KSGC--EGGGSAYMCNSQQPWAVNSTLAYGFGAASFSNGVDV-SLCCACFLLSFKD-QISNKKMIVQVTNTGSD-
Dvir_GH45-8   NASV-QSGC---IGGTSYMCNNQQPFVVNETLGYGFAAVSFSGGVDN-DLCCSCYLLTFQN-QINNKKLVLQFTNTGGD-
Dvir_GH45-9   NPDA-MSGC---ENGTAYTCTNQQPFIVNQTYGYGFAAAYLIGGPSTNNFCCACFLLNFTD-QIKYKHMVVQVTNSGTN-
Dvir_GH45-10  PPND-PSGC--DINGSSFVCNNNQPYVVNSTLSYGFASASFSGGIDT-SMCCSCMLLNFEG-QLKGKQFLVQLTNSGEE-
Dvir_GH45-11  DPEN-ESGC---IGGQSYICTKQSGFAINSTLAYGYVAARFHGT-TR-NMCCSCVLFSFQPQELANKKMLVQVTNTGNA-


Dvir_GH45-1   --LGQNHFDIQMPGGGVGIYNLGCKTQWNAPDDGWGERYGGVTDIKGC-KQLPEQLQEGCRFRFTWMKGVPNPPVSFYQI
Dvir_GH45-2   --LASNHFDLQIPGGGVGIWNHGCDAQWGAGENGWGRRYDGVSSLEEC-CLLPEVLQPGCRFRFQFMEGVYRPNVTFQEV
Dvir_GH45-3   --LNENQFDLQIPGGGVGIFNLGCMTQWGTGEDGWGRRYGGVSSIEEC-SILPEVLQPGCRFRFQFMEGVDNPKVSFQEV
Dvir_GH45-4   --AVAHHFDLQIPGGGVGYNTQGCRIQWNAPENGWGDRYGGVHSEQEC-NQLPWQLQAGCKFRFQFMQGVSNPDVSFQEV
Dvir_GH45-5   --LTTNQFDLQIPGGGVGLYN-GCVKQWNAPVDGWGERYGRVTSVEGC-DQLPVQLQDGCKWRFEYLEGVSNPSATFYEV
Dvir_GH45-6   PHATHNEFDIAMPGSGVGYYTQGCSSQWNADVSKWGDQYGGVHSIEEC-HNLPAHLQPGCEFRFTWMKGYSNPDIEFDEV
Dvir_GH45-7   --LSHNHFDIALPGGGVGIFTQGCHDQWNAPWNGWGDQYGGVHNRGEC-ATLPQALQSGCYFRFDFYQNANNPRMHFDQV
Dvir_GH45-8   --LGSNQFDIALPGGGVGAFNQGCHDQWNAPWTGWGQQYGGISSREECLSLLPKELQSGCLFRFDFMQNANNPQMYFEQV
Dvir_GH45-9   --FDKNEFVIALPGSGVGDHPEGCHDQWNAPWTGWGDQYGGVHMRSECVTLLPEELQEGCKFRFDFMETAANPLVSFQQV
Dvir_GH45-10  --YQTNQFDLGIPGGGVGLFPKGCTAQWNAPSTGWGDLYGGVHTEEEC-NELPEVLQPGCKWRFTFMEGVSNPEVTFYQV
Dvir_GH45-11  PETNTNLFDIAMPGSGVGYYTQGCTSQWHTDVSSWGDQYGGVNSLQEC-YNLPQPLWEGCAFRFNWMLGYSNPDVSFEEV


Dvir_GH45-1   KCPEYFVGVSKCGDL---
Dvir_GH45-2   QCPAELIAVTACGNLNY-
Dvir_GH45-3   KCPAELVAVSACGDLD--
Dvir_GH45-4   KCPSQLVSITGCGDL---
Dvir_GH45-5   KCPSELIAITNCGDRD--
Dvir_GH45-6   VCPKRLTDISGCYPASHP
Dvir_GH45-7   QCPAEIVARSGCSL----
Dvir_GH45-8   ECPAELVKISGCSLPL--
Dvir_GH45-9   VCPDELVKISGCRIPE--
Dvir_GH45-10  QCPRELVERSGCVL----
Dvir_GH45-11  ECPQELLSISGCDPISHP
```

Figure S4.1. GH45 protein sequences identified from the *D*. *v*. *virgifera* transcriptome. GH45-6 includes only a partial GH45 coding sequence. Two potential amino acid residues for the catalytic nucleophile (Asp40) and the proton donor (Asp154) are highlighted with red (Sakamoto and Toyohara 2009). Dashes indicate alignment gaps.

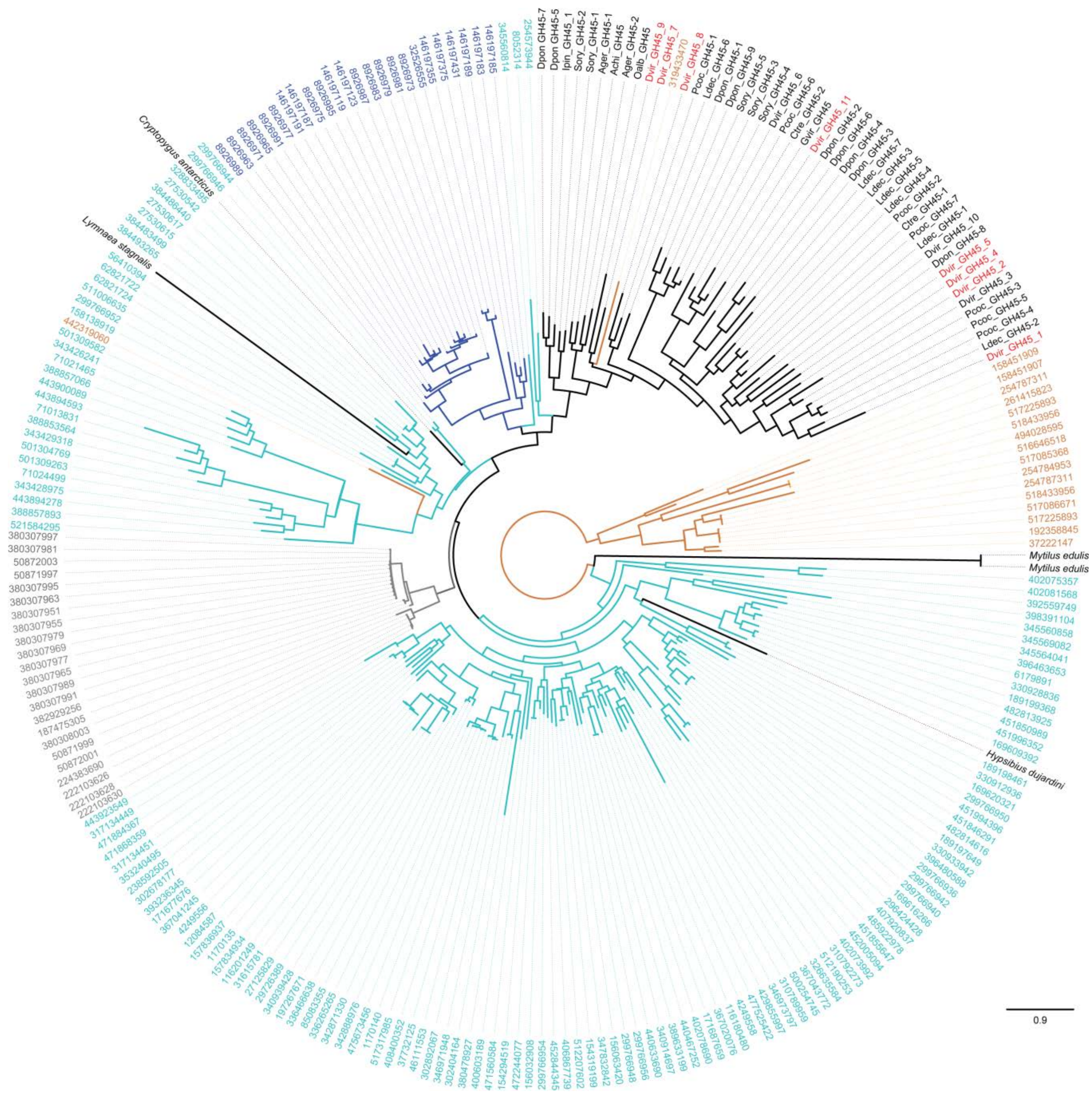


Figure S4.2. The maximum-likelihood phylogeny of GH45 family proteins. Coleopteran proteins included are found in Supplementary Table S5. *D. v. virgifera* sequences are shown in red. Bacterial, fungal, nematode, and protist sequences are indicated by brown, cyan, grey, and blue, respectively. The NCBI gi numbers are shown except for panarthropod species. The scale bar represents the number of amino acid substitutions per site.

```
Dvir_GH48-1  ---MRLGLFVLFCVTSTALAGTYTDRFLTQYRKIHDSNNGYFSKEGIPYHSVETLIVEAPDHGHETTSEAYSYYVWLEAV
Dvir_GH48-2  --------------------------------------------------------------------------------
Dvir_GH48-3  MTPLHLLVLAVIIMNHASCESVYKQRFLEQYNKMHDPNNGYFSSKGIPYHAVETLVVESSDYGHETTSEAHSYYIWLEAM


Dvir_GH48-1  YGKVTGDFSSFNNAWNNLETYIIPVYSSQPTNSFYTPGHPATFIPEQDDPSQYP-SQIDSSVPVGQDPLHQELVNAYGSH
Dvir_GH48-2  --------------------------------------------------------------------------------
Dvir_GH48-3  YGGITNNFSRFNEAWEIMEKYIIPVHESQPNTNLYNPSHPAGYGPEQEYPEDYPVGPVDPPAPVGIDPLYQELVDTYGTS


Dvir_GH48-1  EVYGMHWLLDVDNIYGFGNTPGNCNLGPSAGGPSYINSYQRGSMESVWRTIPQPTCDNFRFGGNHGFLDLFTKDNSYAQQ
Dvir_GH48-2  --------------------------------------------------------------------------------
Dvir_GH48-3  DIYAMHWLTDVDNVYGFGNSPGNCELGPNEPGPSFINTYQRGPRENAWKTIPQPTCDSHKYGGPEGFGPLFSTGD-HAPN


Dvir_GH48-1  WKFTNAPDADARAIQAAYWAGQWAQQSGQLGTIQGTLAKAAKMGDYLRYALFDKYFKQVGNCDNRWSCPGGYGKSSAHYL
Dvir_GH48-2  --------------------------------------------------------------------------------
Dvir_GH48-3  WKYSVAPDADARAIAAAFWASRWATKSGHLSEITDTLQKAGKLGDYLRYCFFDQNFKRIGNCIDPYKCPGGTGKDSAHYL


Dvir_GH48-1  LGWYYAWGGSVDTNGGWAWRIGDSAAHFGYQNPLAAYALANDPNLRPKGATAVSDWQTSLERQLEFYEWLQSAEGAFAGG
Dvir_GH48-2  --------------------------------------------------------------------------GAFGGG
Dvir_GH48-3  LGWYFGWGGSISSEYGYSWRIGDGVAHFGYQNPMAAYALINEPNMTPKGATAVEDWQISLDRQLELYDYLQSVEGAFAGG


Dvir_GH48-1  ATNSINGHYDSPSSDLTANTFHGMYYDWEPVYHNPPSNRWYGMQSWSVDRLAQYYYVTGDSKAKSVLDKWVNWILKETTI
Dvir_GH48-2  ATNTWNGRYDTPPQELTTNTFHGMFYDWEPVYHDPPSNRWYGMQSWSTDRLAQYYYVTGDATAKTLLDKWVKWVISEIKF
Dvir_GH48-3  VSNSWNGRYEQPPEELMDNTFHGMFYNWEPVAYDPPSNQWFGMQPWSTDRLAQYYYITGDDKAKKILDKWVSWIIANTYF


Dvir_GH48-1  EAGKSFKLPSQLSWSGNPPNVHCTINAYTTDVGSASGTARTLAYYAAKANHAQAKEVAKEILDIMWNNFQTSKGVSSPEV
Dvir_GH48-2  E-GTGYTHPDHLEWSGQPPNVHVQVTSYSDDVGTASSTA-----------------------------------------
Dvir_GH48-3  E-GDDYRIPSTLDWVGVPPNVHCKVVYYGNGVGPAAATARTLSYYAARANHAEAKNLAKKILDSLWNLHRTPLGIAVEEQ


Dvir_GH48-1  ADTYTQFNEPVFVPNGWYGTYPKGDVIQSGATFLSLRSWYKSDPDWNKVQTYLNGGSAPTFTYHRFWAQADIAISNGVYG
Dvir_GH48-2  --------------------------------------------------------------------------------
Dvir_GH48-3  PEIH--FNQSVYVPKDFHGVYPNGDVIDSDSTFISMRSFYKNDPQWNKIESYMNGGPAPKFTYHRFWDQTDVALGFGVYG


Dvir_GH48-1  ILFNE
Dvir_GH48-2  -----
Dvir_GH48-3  LLFDE
```

Figure S4.3. GH48 protein sequences identified from the *D*. *v*. *virgifera* transcriptome. GH48-2 includes only a partial GH48 coding sequence. The potential residues for the catalytic nucleophile and the catalytic proton donor are highlighted with magenta and green, respectively (Parsiegla et al. 2008).

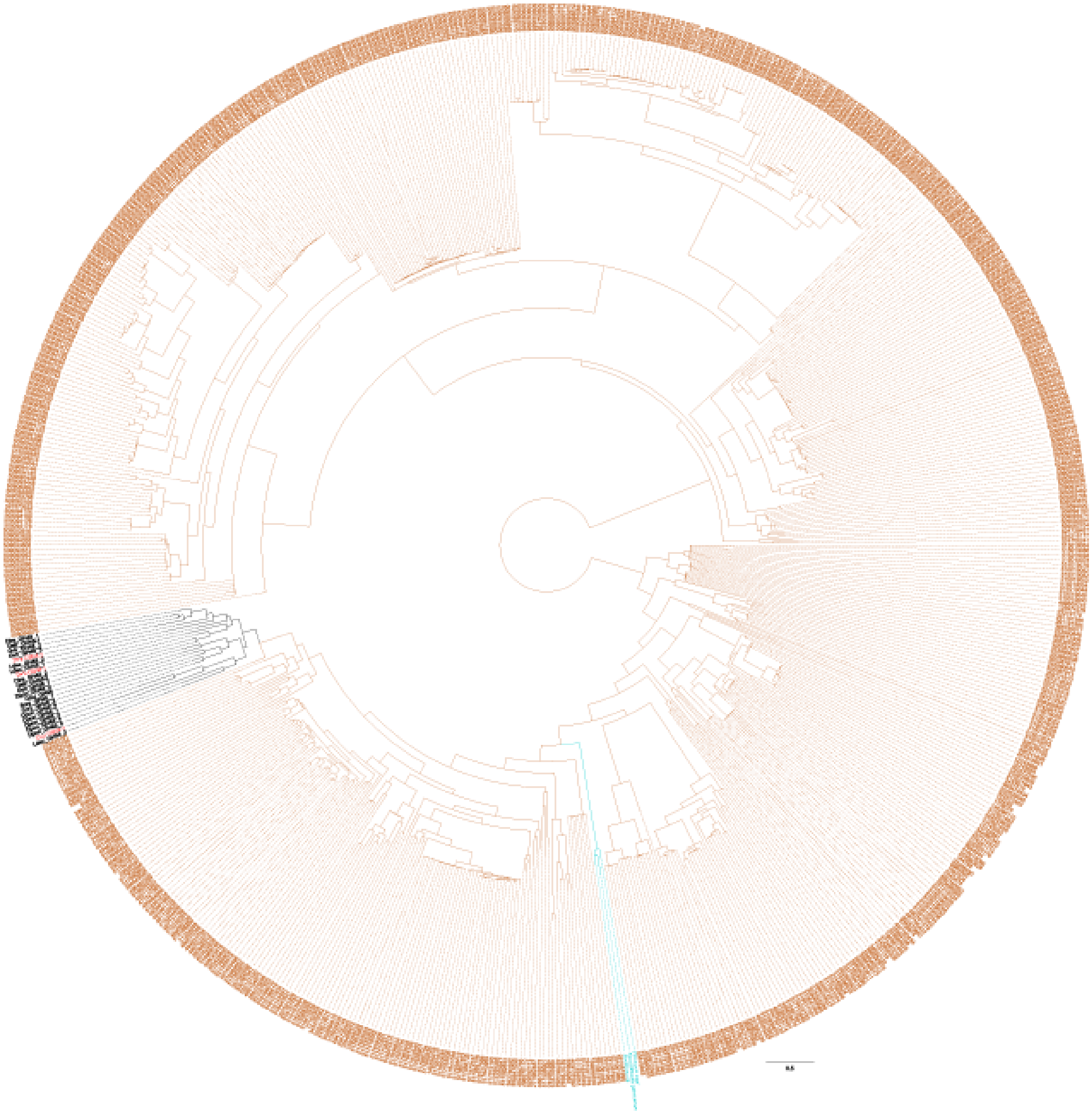


Figure S4.4. The maximum-likelihood phylogeny of GH48 family proteins. Coleopteran proteins included are found in Supplementary Table S5. *D. v. virgifera* sequences are shown in red. Bacterial and fungal sequences are indicated by brown and cyan, respectively. The NCBI gi numbers are shown for bacteria and fungal sequences. The scale bar represents the number of amino acid substitutions per site.

```
Dvir_GH28_1   M--TNLTLLIVFSVIVATIAIPFNST-KNIGDGCTISNIWEVENVVKNCKNIVVNNLYVPGGQKLELKLHSGTVLKFQGT
Dvir_GH28_2   M--ATLTLFLVLCAAVATSAISLNST--NVGAGCTISKIGEVDNVVKNCKNIVINNLSVPGGKTLKLDLHPGTTLKFQGT
Dvir_GH28_3   MYYTIMCYLFLFLLFNAALVICKCSP-----TNCEITNFDQVSDTVHRCSDIIIRNLDVPAGQTLELDLKQGASLTFEGI
Dvir_GH28_4   M-NLFIIENFIVVVLLNSLLFISCVD-----QPCTITNFSQVSEVLQSCKNITISNLNVPAGQQLYLELLNDSSVTFEGV
Dvir_GH28_5   M--------SYTKFLIVAFISTVSAN-----NNCTITEFAQVAEIVKECSNIVINDLVVPAYSTLLLNLKNGSRVTFTGN
Dvir_GH28_6   M--RTIQLFEYFFLCSIAYASNLT-------ASCTISRFDHVDTVVSQCKSITVESFAVPAGQTLKLHLQYGTTLTFNGN
Dvir_GH28_7   MIKTGMSLVLFFLGVVLAQE-----------YDCEINSIDQVLPVIEKCSVITVKNLWVPSGQTLELSLKDNSHLIFDGN
Dvir_GH28_8   --------------------------------------------------------------------------------
Dvir_GH28_9   MSSNKLIYSLLFVVISAAAKSLNE-------DCCTITEYSQVPDVVETCKNIVISNLRVPANKTLNLNLQDGSELTFEGR
Dvir_GH28_10  ---------------------------------------------LNDC-------------------------------
Dvir_GH28_11  M-------LFIYKILVLLIVVSIAAS-----DICTISNYDLVDEALSSCIDIVISNLTVPSGKTLNLNLKERSTVTFDGV
Dvir_GH28_12  M-----CYFNKFSLLLLLYSPLLSKS-----DPCTVTQFSQVAQAVNDCTNLIISNLVVPGGQTLELHLKYGATVTFEGT
Dvir_GH28_13  MGFSVLLFLSLLALISGTSVLQATNNTEAVGDSCTITQYSQVDGVLKSCTNIILSNVEVPSGKSLNLYLRDGSTLTVRGT
Dvir_GH28_14  --------------------------------------------------------------------------------


Dvir_GH28_1   TTFQHSNW-EGPLVEITGSNLHVSGA-GAILDGLGAQYWDGY-GDKGAVKPKFLKIRTT-GSTFDNIHLLNCPRQCVSIL
Dvir_GH28_2   TTFQHTNW-EGPLISISGSNLHVSGS-GAVLDGLGSKYWDGK-GDKGAKKPKFFKIRETTGSTFDSIHLLNCPHQCVSIQ
Dvir_GH28_3   TTFDYTNW-SGPLIRINGSGFTIKGAPGSLLNGQGDLYWDHL-GDKGPKKPQFIKIEAFDGSIIENINLLNCPHHCVYVG
Dvir_GH28_4   ITFGVAQW-KGHLIVVKGHNVIIQGAPGSILNGQGQKYWDGQGGGGGTTKPKFFYIETTGGSIFKNIYLYQCANWCVGIG
Dvir_GH28_5   VLFEVGYW-EGPLLEISGDGVEVQGNAGHIINAQGEKYWDGQGGSGGVTKPRFVVISTTGGSVLRNIYLLNCVYFCVGIH
Dvir_GH28_6   IAFGYSEW-DGPLMWIKGDGITIQGTESHLLNGRGELWWDGHGDHSNKKKPQFMLIQATGNSLLKDIKVKNCPHTCIGIS
Dvir_GH28_7   VTVGVKYQDEVPLIRISGANLFIEGRKDAVINGQGEKYWDGKGIEGKNRKPVLLEISAQ-ESLLKNINIRNCPQKCVNIL
Dvir_GH28_8   ----------GPLVRFRGSQIVVQGAKGSFLDGQGALYWDGMGGNGGVTKPYFFQIETTGGSIFRNIHLLNCPHHCVIIS
Dvir_GH28_9   TYFDYFEW-KGPLVNITGDDLIVRGAPGHVLDGQGELYWDHL-GGKGIKKPKFIRLQGN-NSRYENIYLKNCPVHCASVA
Dvir_GH28_10  ----------------------------SILDAQGEKYWDGQGGAGGVTKPKFFYVQTTGGSILKNIYLLNCAHFCVGVG
Dvir_GH28_11  ITFEVSFR-TGFLVSVAGKNVLVQGAPGSILNGQGEKYWDGF-GDNGVVKPKFFRVATSGGSIFRNIYLLNCPHFCVGVY
Dvir_GH28_12  TVFEVAHW-EGPRIEKKEENVEVQGASRSILNAQGEKYWDGHGGSGGVTKPRFVQISTTGGSVFKNIHLKNCALFCVGIR
Dvir_GH28_13  ISFDVGYN-NIWLVTISGNNIKVIGEKGSLFHGHGEKYWDGHGGSGGVTKPKLLQILNVNNAHFSNINLKNCPMFCTGIT
Dvir_GH28_14  --------------------------------------------DKGNKKPKFFKIQATGGSVFKNINLLNCPHQCVSIQ


Dvir_GH28_1   SSKQTTLTNFNIDVSAGDITHL-ATNTDGFDLSD-SDGITIENSVVRNQDDCVAVNSGKNYHFNKLNCNGGHGLSLSVGM
Dvir_GH28_2   NSKKTTLNNWNIDVAAGDINSL-GHNTDGFDLCE-NEEITIQNSIVHNQDDCVAVNSGKHYHFNKLTCVGGHGLSLSVGT
Dvir_GH28_3   KSDGLTIRGWVIDNSYGDQNNFTGHNTDGFDVSA-ASNLIIEDSTVINQDDCIAIRHGYNILVRNMYCAGGHGLSLSAGF
Dvir_GH28_4   -SKDVIITGWTIDNTAGDKDMI-ALNTDGFSLID-SENVLIENSTIMNQDDCIVVRRGNNMTFRNIKCFGSHGLSFATGF
Dvir_GH28_5   -ASDLTLSGWTIDAVAGNTRG--GLNTDGFGIGN-GQNILIENSVIMNQDDCVVVNSGSDMVFRNLECYGSHGLSFSIGD
Dvir_GH28_6   DSHDITLQHWTIDCQDGDTKG--GANTDGFDIAK-SYKVTIKDTTVRNQDDCICVNQGQHLVFQNMHCIGGHGLSLASGL
Dvir_GH28_7   KSANSSFTGWNIDITDGFKDNV-GVDTHGFAVAN-SSDIIIKESNIINQGDCIVVNQGSDLHFEQIVCRGSQGITVRPEW
Dvir_GH28_8   -STDLTITGWNIDVSAGDKGNL-GHNTDGFDVIY-GENIVIENSIVQNQDDCVAINRGKNMLISNLRCYGGHGISLSVGF
Dvir_GH28_9   VS-NSIIDGWLIDVSEGDKNNFTGHNTDGFDLS--STNLILQNSIVKNQDDCVVVNVGANILVRNMACYGGHGLSISAGF
Dvir_GH28_10  -AKDTTITGWTIDSVAGNKDLI-ALNTDGFGVSSHSDNILIENSVIMNQDDCVVVNQGTNMVFRNLHCYGSHGLSFAVGF
Dvir_GH28_11  -ATDVTLTGWTIDVLAGNTRG--GLNTDGFGIHS-GRNIVVQDSVVMNQDDCVVVNSGTDMIFRNLQCYGSHGLSFSVGS
Dvir_GH28_12  -ASDLTISGWNIDSHEGRKK---GKNTDGFGIAA-GNNIHIENSSVDNQDDGIVVNGGTNMVFNGIKCTGSHGLSFSAGS
Dvir_GH28_13  KAKDLTIDGWNADCAEGDKL---GRNTDGIGISW-SQHVYINNAYIHNQDQCLYVNQGSDMVFTGIHCVGSNGICATAGF
Dvir_GH28_14  NSKQLTISNWNIDVSAGDKNKL-GHNTDGFDISG-SDGVNFEYCTVQNQDDCVAVNSGKNLHFNHMTCSGGHGLSLSIGM


Dvir_GH28_1   SKNDSP--------RNHVEDVTFSNCIVSNSLNGIHIKT-HSDAGKGYINGVEYRNIILKDITNYGINVQQDYQGGHSTG
Dvir_GH28_2   STTDPS--------KNYAEDINFSDCSVSNSRNGIHIKT-HTDGANGYIRGVTYKNIKLSGITHYGINVQQDYNGGGSSG
Dvir_GH28_3   SYTTFQ--------ENTITNVVIKDSVIARSANGIHVKT-HADAYNGRIQNVTYENIFMSGLINYGINVQQDYVNGSATG
Dvir_GH28_4   HETDGPFGHKEDDAEDIATDITFEDCLVANGLYGIHIKT-APNGKRGRIENVLFKNIKLSGIQEDGIYIQQDYGD---IG
Dvir_GH28_5   SHNDDA-------AANTIKNITFSDCLVANGLYGIHVKT-KK--GTGVLTDVTYENIRLSGITEDGIYINQDYGD---IG
Dvir_GH28_6   -WDTYE--------LNTIYNVTFQNSIVENSRNAIHIKTIPVNNKKGEITSITYDNIKLIGISYYAINVQEDYTNDGPTG
Dvir_GH28_7   EY------------ENYIRDVIFDDCTVIEGQTGIRVVT-SPHQPEGYISNVIYRKIHLTGILFRGIDIRQDLDD---EG
Dvir_GH28_8   SHRSYK--------HNTVHNVTFIDCVVARSENGIHVKT-HNDGYLGEIKNVTYKNIEFVDILNYGVNVQQDYANGTSTG
Dvir_GH28_9   SKDDFA--------KNSVYNVIFEDSLVHRSPNGIHVKT-HADSGPGIIQNIIYRNIRFEDINNFALNIQQDYVNGEATG
Dvir_GH28_10  GGRDKP------EDDSVASNITFENCWVANGLYGIHVKT-GAVGNKGRIENVVFRNIKLSGIQEDGIYIQQDYGN---IG
Dvir_GH28_11  KTEENA-------EAGIVQNITFLDSLVANGLYGIHIKT-KK--GSGTIRDVIYENIQLSGITEDGIYINQDYED---IG
Dvir_GH28_12  NTNDHA-------KYATINNITFSNCELKDGAIGIHVKT-KR--GTGLITNVTYDHITMTGMQKDGIYINQDYGD---VG
Dvir_GH28_13  SKTSYE--------ENTTKNITFHNCVLEGGLTGVQVIA-MADGGPGEITDIHFQSIILKGVRQQGVYVQMDYGN---DG
Dvir_GH28_14  SKTDSS--------KN----------------------------------------------------------------
```

```
Dvir_GH28_1   YPTSNIPINGLKLEGVTGSLRS-----GQPVYIFCGN-NACFNFNWSGVSITGGNQQSSCNYHPNGYYC-
Dvir_GH28_2   YATSNIQINGLHLQSVTGSLKS-----GKAVYILCGN-KACSNFNWSGISIYGGNEKNGCNYHPNGFSC-
Dvir_GH28_3   VANNNIPIYNLNLINIRGTVRDSDSEKSMPVYINCGK-SACHEWSWSNINIAGGSNSSICNYTPDGYQC-
Dvir_GH28_4   KQDSNVTIKNLTLKNVYGSLQGIL---TRPIHIFCGNQGTCSEWIFSNINILGGN-RSYCNYQ-------
Dvir_GH28_5   NYSREIEITNLKMSNIYGSVQGVL---TRPVHIVCSN-DKCQNWTWSNINILGGG-KNYCNFQPTEFIC-
Dvir_GH28_6   HPLGNIPVKDLKIHNVYGTMTGSN---SVKAYILCGS-GGCTNWNWSEINVSGAAKPNSCNFTPNGFSC-
Dvir_GH28_7   RPSGNVKITELDISDVKGNMTDKY---VRSVYIWCGP-DGCANWNWSDIDIENAEVENACNFLPNNWSCW
Dvir_GH28_8   NPTNNIPITNLSLINVHGTVKGSH---ATGVYILCGS-AGCIDWNWSEISITGAKRENSCNYVPSGYHC-
Dvir_GH28_9   IPGTNIPIVGLSLDNISGWMKSFNESPTLEALILCGD-GACDKWEFHNIDITGAQNNSICTFQPEGYSC-
Dvir_GH28_10  NLESEVKIHNLTVENVTGSVQGVL---TRPIHIFCGNRSTCDDWKFSGINILGGG-QSYCNYVPDEFHC-
Dvir_GH28_11  NYSREFEIHNLKISNVYGSIQGLL---TRPVHVVCNE-NKCSNWTWSNINILGTG-KSYCNYIPDGFRC-
Dvir_GH28_12  NTTRDFQITNLKVSNVEGSIHGKG---ARAVHIVCND-KKCANWQWSNIDISGGA-KDYCNFHPTGFDC-
Dvir_GH28_13  HPNNNIAVTGLKLSHVTGTVSGNS---ARPYYIKCG--AKCSNWIFNDVQVTGGGVKSSCNYKPSGFNC-
Dvir_GH28_14  ----------------------------------------------------------------------
```

Figure S4.5. GH28 protein sequences identified from the *D*. *v*. *virgifera* transcriptome.

GH28-8, 10, and 14 include only partial coding sequences.

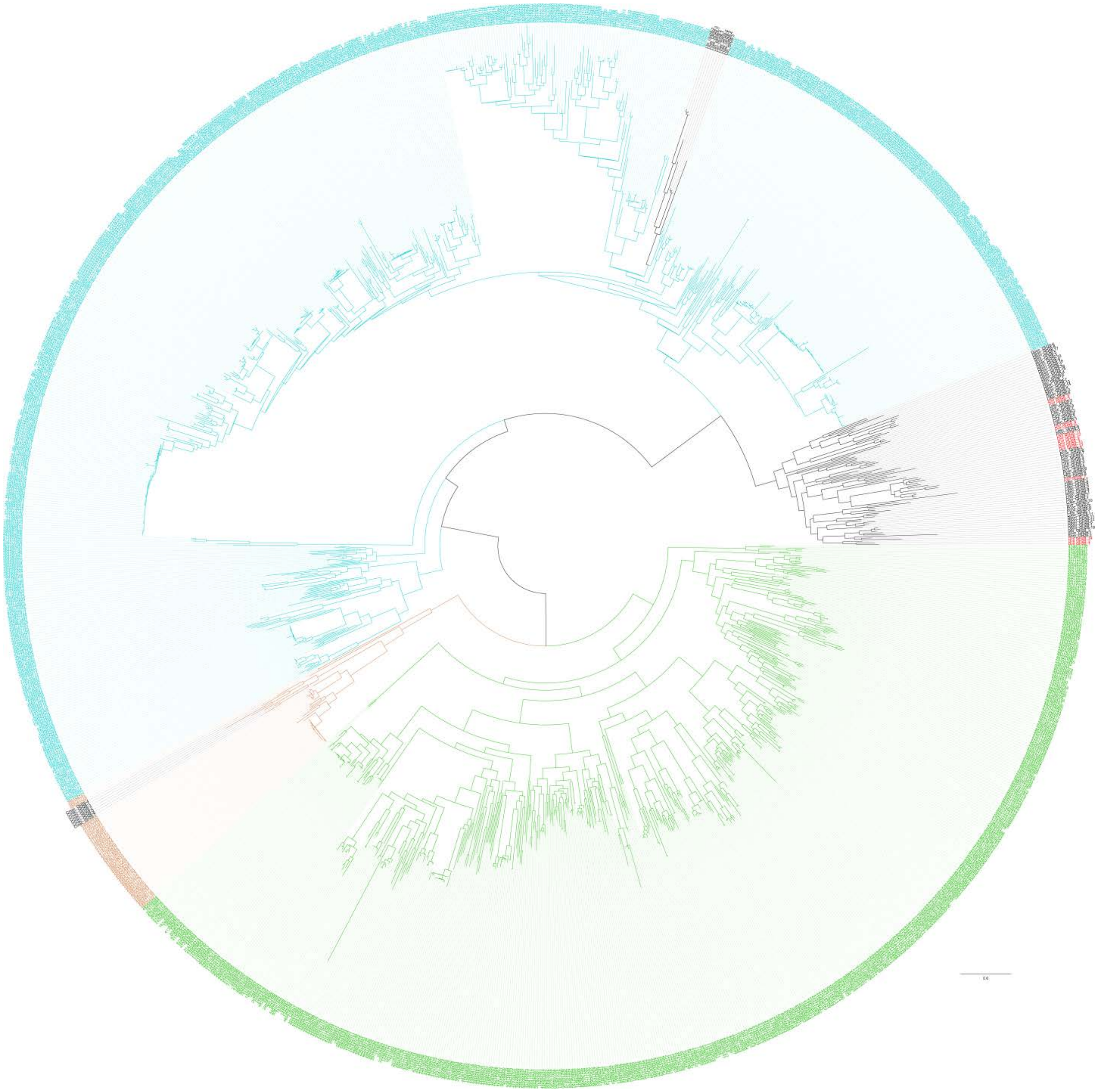


Figure S4.6. The maximum-likelihood phylogeny of GH28 family proteins. Coleopteran proteins included are found in Supplementary Table S5. *D. v. virgifera* sequences are shown in red. Bacterial, fungal, and plant sequences are indicated by brown, cyan, and green, respectively. The NCBI gi numbers are shown except for insect species. The scale bar represents the number of amino acid substitutions per site.

```
Dvir GH16-1 ----MK-----ILFVVI--WFAFVKAQ--YEVPDAVVEVYEPKGFTVSIPDEEGIKLFAFHGKINEDFDGLEAGTFARDILKPRNGSWTF
Dvir GH16-2 ----MYY----FILILC--HVAFVFAQ--YEVPAATVEVYHPKGFSVSIPDEDGIKLFSFHGNINQEMDGREAGTFSRDILRPVNGMWTF
Tmol GH16   ----MKVL---VVFIFC--LVRSTFGQ--FEVPDALVEVFRPRGLRVSIPDQEGIKLFAFHGKINEEMNGREGGTFSRDILKAKNGRWTF
Tcas GH16-1 ----MKEFS--FVVVFY--FITLSLAE--FEVPDALVEVFQPQGLRVSIPDQEGIKLFAFHAKVNEEMNGREGGTFSRDITKAKHGRWTF
Dpon GH16-1 ----MYMWTVCLVLAFC--ACFSH-GQDGFEVPDATVEAFTPRGLRVSIPDQDGIKLFALHAKINEEMNGREAGTFSRDITKAKDGRWTF
Dpon GH16-5 ----MKNLVCYLALIVV--GLV--------------------------------------------------------------------
Dmel GH16-1 ----MADALRFVAWSCCLQLLFLLLGVQGYEVPKAKIDVFYPKGFEVSIPDEEGITLFAFHGKLNEEMEGLEAGTWARDIVKAKNGRWTF
Bmor GH16   ----MWLLTLGVVALIS-------------------------------------------------------------------------
Sfru GH16   ----MWSVLAGVLAIAS-------------------------------------------------------------------------
Pame GH16-1 TRGPLAFVATGLLVLLL-------------------------------------------------------------------------
Pame GH16-2 ----I-------------------------------------------------------------------------------------
Cant GH16   ----MNAFTFPLLLAFC--AFA--------------------------------------------------------------HGAW--
Calb GH16   ----MEPLLCLVLFPLV--A----------------------------------------------------------------------
Hhan GH16   ----MEYSVKYL------------------------------------------------------------------------------
Cgig GH16   ----MSALGVLALVLVV--------TRTVLSIQPAIVEQYNGHGVKFTIPDDGNYDFVAVHYSINQPIAGVGAGQWAFDVHTKQGSSFVH
Cint GH16   ----MFASLVFVLSLAA--------CSHAYSVQQPIISLMEPTGLMFMYPDDGNINLVSYHYSINTPLPDVQAGTYNQDVSESTNGYFTL
Aque GH16   ----MSCYYFLFLLLVT-------------------------------------------------------------------------

Dvir GH16-1 RD--RETRLKKGDIIYYWLYVDYNNGRNTLGYRKTDQQYEVREFSNNP-----NAPVNRVKPTT--------------------------
Dvir GH16-2 SD--RSTKLRVGDKIYYWTYVEYGDEYEKRGYPKDDQVFTVTNLIKKPPGRDKNKDTGESRPSVITEPTTTT-------------TTTTT
Tmol GH16   YD--ANARLKEGDILYYWTYVDYFDGKNKLGYPNDDQKFVVKQLLDKD-----GA------APSVTPPTVTK-------------APPQE
Tcas GH16-1 YD--PYAKLKIGDTIYYWTYVDYFDGKNKLGYTKDDQEFVVRQLLDKE-----KS------PAN---PPVKK-------------PEPEI
Dpon GH16-1 YD--SQAKLAVGDTLYFWTFVDYFDGERKLGFVRDDQFFTITELLPKP-----GAKPPPASVPTVTPVT----------------KSPSI
Dpon GH16-5 ------------------------------------------------------------------------------------------
Dmel GH16-1 RD--RITALKPGDTLYYWTYVIY----NGLGYREDDGSFVVNGYSGNN-----ASPHPPVVPVSTTPWT----------------PPADP
Bmor GH16   ------------------------------------------------------------------------------------------
Sfru GH16   ------------------------------------------------------------------------------------------
Pame GH16-1 ------------------------------------------------------------------------------------------
Pame GH16-2 ------------------------------------------------------------------------------------------
Cant GH16   ------------------------------------------------------------------------------------------
Calb GH16   ------------------------------------------------------------------------------------------
Hhan GH16   ------------------------------------------------------------------------------------------
Cgig GH16   TNDLPAIQVHKGDTVYYWLHAQK----GGTPSELLGQSAVIGDLTTTPKPTTTTTTTTVRPVVTSKTTSAPSGGSHGSGHSSGGGILTQT
Cint GH16   QN--FNVAVVPGDEVNYWVNVIT----STGGYLLTDQTWVAQGPTTTTPAKTNPPTQPPTQPPTNSPTNPPI------------------
Aque GH16   ------------------------------------------------------------------------------------------

Dvir GH16-1 -----------------------------------------EPT--STA---------------KLILFDDFSTK-----RDDFWTVEQR
Dvir GH16-2 TTTTTPSPPTVPNVNICDPS--------------VTVINNGQNT--CKG---------------KLIFSESFNTK-I---KSTSWTFENK
Tmol GH16   HTTLESG---------CKAS---------------VTTKVNERV--CAG---------------EQIFHEDFTT--F---ETNIWRPEVK
Tcas GH16-1 EERLPSG---------CKAS---------------ATISKQKKM--CKG---------------EEIFNENFEN--L---KPELWNREIK
Dpon GH16-1 DTGSNGG---------CKVS---------------TTTVRGKNS--CRG---------------QLIFDSSFDQ--FAKNKANLWTIQRK
Dpon GH16-5 --------------------QSTPAACDVASVTTASGPAYNPPSMLCPG---------------DLIFEDHFDT--L---DLDTWKHEVT
Dmel GH16-1 DIDIRLG---------CTTP--------------KTEVNGAPTR--CAG---------------QLVFVDEFNAAKL---DPNKWKAERK
Bmor GH16   -------------------------ASKACTPSVTTVSGTHAPVTVCSG---------------QLIFADDFVD--F---DLEKWQHENT
Sfru GH16   -------------------------LGAACTPSLTTVSGTHAPVTVCSG---------------ALIFADGFDT--F---DLEKWQHENT
Pame GH16-1 ----------------------------------------GQPG--SPE---------------TLVWQDEFDT--L---NLNEWSHLVT
Pame GH16-2 ------------------------------------------------------------------------------------------
Cant GH16   ----------------------------------------------------------------VLDWEDEFNGG-N---LADRWNFELG
Calb GH16   ----------------------------------------------------------------GAGFRDDFTT--W---NPNNYQIGVS
Hhan GH16   -------------------SASGDVVYHVDGVFTIPAASSLSPRLYRRG---------------NTVFEDSFNSHQL---NPKHWHHEIT
Cgig GH16   MIETQSAGTSGGQSQGGSSVGSSGGTQQVYSGTGYVQHGTSQQSCTSYP---------------CLIFEDNFDF--L---NFETWTHDLT
Cint GH16   ------VSTTEPPATAPPATGPPGVTTTTASSGGSGGTGGGGPAYVCSSYPCDSQCDMSVAPCNGLIFEETWDQ--F---DLNRWQHEIT
Aque GH16   ----------------------------------------TSNG--REL---------------TLALEDEFDT--F---NLSLWKHEIT

Dvir GH16-1 YADAPDYEFVLYV-NKPEVFQIKDSNLHIRPVPSED--IFGNGFLT--SEYDL-----GNSCTAPIGTTDCKRKY---DAGFI-LPP---
Dvir GH16-2 FAGLPDYEFVLYT-NRPEVAFIQDKALVIKPALMDN--VYGPNFVE--QPLDL-----GTSCTGALGTLDCHIRP---DAGFI-LPP---
Tmol GH16   FADKPDYEFVFYR-AGPPNLQVKHHRLTIRPVPSDA--VFGEGFVSRREKVNL-----APACTGVHGSIECVQTP---GAFLI-LPP---
Tcas GH16-1 YAGKPDFEFVLYT-DRKEILSVNNNELTIRPIFTEQ--LFGKEFVSYEHELDL-----GEKCTGIHGTTDCVQKA---DAFLI-LPP---
Dpon GH16-1 FATGPDYEFVIYE-DNPIVLSVENSRLAITPILTDS--LYGEGFVVRPEGFDL-----GEKCTGVRASAECYQTA---LGWRI-IPP---
Dpon GH16-5 MAGGGNGEFEYYR-NSRTNSFTQGGNLHIKPTFLAD--EYGEDFLYSG-TLDI-----TDECTNS-NHNGCLRTG---TSTNI-LNP---
Dmel GH16-1 FSGQPDYEFNVYVDDAPETLCLANGHVVLSTNTMKK--QFKKG---SGESLDL-----GEKCTGQANTHDCVRNG---RTLNDGLPP---
Bmor GH16   LAGGGNWEFQYYN-NNRTNSFTNNGLLYIRPSLTSD--QFGSAFLHSG-RLNIEGGAPADRCTNP-QWYGCERVG---TPTNI-LNP---
Sfru GH16   LAGGGNWEFQYYG-NNRTNSFVRSGSLFIRPSLTSD--EFGEAFLSSG-HWNVEGGAPADRCTNP-QWWGCERTG---TPTNI-LNP---
Pame GH16-1 AWGGGNSEFQYYR-NDRRNSYVRDGILYLRPTWTSA--EYGDDFLYSG-SLSY-PDCNMEPCSST-------------AGQDI-VQP---
Pame GH16-2 -----------YI-NNRSNSFVKDSKLFIKPTLTAD--VYGEGFLSTG-TLKLYGGAPADECTNP-SDWGCERQG---SAANL-LNP---
Cant GH16   CNGWGNNELQCYTDNRGANARQEDGKLVISAV--RE--WWGDG-----------------------------------------VNPDKE
Calb GH16   AWGGGNHEFQVYT-PEPSNLFVRDGSLYIKPTFTRDSRHFTDGNLYYG-TMDLY--HLWNKCTQH-DNNGCQKHSYG-GNSEI-LPP---
Hhan GH16   CWGGGNGEFQMYT-PEAANTYIKNGVLYLKPTFTAD--KFGDDFFQHG-VLDV--KQQWGSCTAA-QDNGCRRQG---AQ----IPP---
Cgig GH16   ASGGGNWEFEYYT-NNRTNSYTKDGKLFIKPTLTAD--NYGEHFLSSG-TLDLWGGEPNSLCTSN-QFWGCSRQG---SPEHI-VNP---
Cint GH16   MSGGGNWEFEYYT-NNRTNSYVRDNTLFIKPTLTAD--HYGEEFLSTG-TLDLWGGSPADLCTMN-AFWGCQRTG---SGSNY-INP---
Aque GH16   LTGGGNWEFEAYL-NNRSNSFVRDGVLYIKPTLLED--QIGLANVENGFTMDIWGGAPADLCTQN-AFFGCLRKSEKYTGGSI-LNP---
```

```
Dvir GH16-1 IASAQLTTKNKVAFKYGKIEVRAKLPKGDWIYPEIYLTPA--NEKYG-LKSQSGQIRIAFTPGNSDLNH-------------VLHGGLTI
Dvir GH16-2 IISAKITTKGKFSFKYGKIEIRAKLPKGDWLYPILTINPV--KDEYG-PGYDSGQITIAFCPGNAVLSH-------------NVYGGIVI
Tmol GH16   VTSAQISTKGKWSFKYGKVEIRAKLPKGDWIYPELYLNPV--NEEYG-PGYASGQIRIAFSGGNEDLCR-------------DLRGGCIL
Tcas GH16-1 VASGRVNTKDKWSFKFGKIEIKAKLPKGDWIYPQLFLNPV--SEEYG-SDYASGQIRVAFLPGNQAMAQ-------------QLYGGCVL
Dpon GH16-1 VISSQLKTKGKFSFKYGKIEVRAKLPKGDWLYPELYLNSE--SEEYG-SGYESGQIRIAFAAGNEGESR-------------KLEGGVIL
Dpon GH16-5 IESARIRTYETFAFKYGTVVARAKVPAGDWLWPAIWLLPS--DYRYG-GWPVAGEVDLVESRGNRNLTDSTGLNIGTQLAFSTLEWGPSL
Dmel GH16-1 MVTAQFSSK-DFSFKYGRVEVRAKMPRAQWVTPQIWLQPR--RPIYGVDDYRSGQLRIAYTRPNGGNLD--------------LYGAAVL
Bmor GH16   IKSARIRTVNSFSFQYGKVEVRAKMPSGDWLWPAIWLMPA--YNKYG-TWPASGEIDLVESRGNKNMFL-NGLHIGTQEAGSTLHYGPFP
Sfru GH16   IKSARVRTVNSFSFRYGRLEVRAKMPAGDWIWPAIWLMPA--YNTYG-TWPASGEIDLVESRGNRNMFH-NGVHIGTQEAGSTLHYGPYP
Pame GH16-1 LQSARISS--SFSFKYGRVEVRAKLPRGDWIWPAIWMLPK--NWVYG-DWPRSGEIDIMESKGNDNYYDSNGVSQGDDRMGSTLHWGPDA
Pame GH16-2 VTSARIRTVDSFSFVYGKVEVKAKLPAGDWLWPAIWLLPR--YNQYG-GWPASGEIDLSEGRGNLNYISPSGQNIGSELSSSTLHFGPFW
Cant GH16   FTSARMTT--KANWLHGKFEMRARLPKGKHLWPAFWMMPQ--NSEYG-GWPRSGEIDITEYRGQR-----------PQQILGTLHFGAAP
Calb GH16   VMSGKITT--NFAMTYGRVNVRAKIPKGDWLWPAIWMLSR--DRSYG-GWPRSGEIDIMESRGNTKAVL-WGQNSGVNYVASTLHWGPDF
Hhan GH16   IMSSKVFS--VASITHGRVEVVAKIPKGDWIWPAIWLLPPGWPWKYG-AWPASGEIDIMESRGNVHLSEANGATQGVDRVLSTIHYGASP
Cgig GH16   IQSARLRSDKAFNFKYGKMEVRAKMPKGDWIWPAIWLLPH--RNAYG-GWPASGEIDVVESRGNTDYHDENGRSQGVDSFGSTLHFGPVY
Cint GH16   IQSARLRTVNSFSFKYGRVEIEAKMPTGDWIWPAMWLLPK--TNSYG-SWPASGEIDICESRGNTDLKDDQGVSHGNDAMGSTLHWGPYW
Aque GH16   IKSARLRTAESFNFKYGKIEVKAKLPIGDWLWPAIWMLPR--HNQYG-VWPSSGEIDIMESRGNAIGYSEG----GYDSFGSTLHWGIDY

Dvir GH16-1 GRSV---AATNYFDKTIESKMSSWSDDFHTFRVDWKPNEISFSVDGTVYGNIY-PPARGLASLGPT--LNL-NTDKWK-EGTLMAPFDQE
Dvir GH16-2 SGSP---VGRKYGLKSISRSS-PWYSSYYRYAVTWNEDGISLSVNDRIYGTIS-PPSGGFSTLAKA--LNIKNADRWK-TGSLFAPFDKE
Tmol GH16   GSRP---AARNYAVKNIVKNSGSWSDDFHKFIVIWKPDQITMMVDDQVYGNIY-PPEGGFVSEAYN--LDLVNVERWR-GGTSFAPFDKE
Tcas GH16-1 GPTT---AARNYALKVIRKTDGLWSDDYHKFTAVWKPDQITLSVDDQVYGYIE-PPRGGFVSDFQNLGLDFEIVERWR-NGTSFAPFDKE
Dpon GH16-1 GSIP---AARKYAMKTIEKTQ-SWTDDFHNFSALWKPDSITLSVDNMVYGTIF-PPEGGFASLATN--LHLKNSDKWK-SGTKIAPFDKE
Dpon GH16-5 EQNQ---YTKTHWVKSN-PD--GYNNDFHLYKVVWSPEGFWFYYDDELIGSVN-PPDGGFWELAGLQDS--DEYNPWS-SGTKMAPFDVE
Dmel GH16-1 FADEPLRSVKNCLKPGTGNNSEDWSDSFHNYTLEWTPRELRWLVDGKEWCVQG-SAKGSFSETTAAGKS-LPQAQKLE-EGTGLAPFDQE
Bmor GH16   GLSG---WERAHWVRRN-SA--GYDTNFHRYQLEWTPDFISFRIDDSEIGRVA-PGNGGFWEYGGFNNR-PGIHNPWR-YGSKMAPFDQK
Sfru GH16   AMNG---WERAHWVRRN-PA--GYNSNFHRYQLEWTPTYLRFSIDDMELGRVT-PGNGGFWEYGGFNSN-PNIENPWR-FGSRMAPFDEK
Pame GH16-1 NHNN---YWRTHWEKSIQDTGTDFADDFHLYGMQWTDNHITFTVDNAEIGTVW-APQDGFWYFGNFEND-PGGTNIWQ-NGNWMAPFDQE
Pame GH16-2 PYNG---YTHAHFEKNT-PAGQGFDKDFHRFQLEWTEDHMQFSIDDEVIGTVA-PGDGGFWELGEFGQQVGTVDNPWQ-YGNKMAPFDQP
Cant GH16   DNKG---DVG----TGERDFPIDFSADFHTFGLDWSPDSMQWLLDDQ-VYHTE-SLQRNFWDGV---------------YNQNGSPFDKN
Calb GH16   NNNR---FQKTHGSKRK-SGGADW-HGWHTYSLDWTAGHIVTYVDNVEIMRIT-TPSQSFWGWGAFSGN-----NIWA-SGGKNAPFDKP
Hhan GH16   SQHR------QQGDSKTSKTGTTWADSFHTYSVDWTAGHIRMDIDNQPVMAWT-TPSQGYWSYSHQSGT-----NVWS-QGGNDAPFDGK
Cgig GH16   GYDP---YEKAHGEMTI-PSG-TLNDDFHIWTLEWDEEHIKVSFEGQEVMNVS-PPPEGFWKLGELDKT--NINNPYKYTNNKMAPFDQE
Cint GH16   PVNA---YEKT-----TKETHGTFASEFHSYVMDWDENKIKFTIDGEELMTVD-PGASGFWEFGEFDTVAPGSDNPWKDTKNKMTPFDQE
Aque GH16   MYNF---FPQTHKSVTI---GTTLANDFHVYGLIWNETYIGTYFDDESNVVLSVPINQSFWSRTGLSTT--YWDNPWV-GAGNNAPFDQE

Dvir GH16-1 MYLTLGVGVGGF--VFKE-----SPSKPWR-NGERNSFQVFNSARQQWQRTWS------DDSKLEVEYVQITSL*---------------
Dvir GH16-2 MYISVGVGAGGL--NFEDKT---DGSKPWR-NYERLSFKKFYQAQQNWSSTWD------EDSRLSVTSIKVWAL*---------------
Tmol GH16   MYLVLGVGVGGH--CFEDRS---DATKPWT-NNDPKSQKKFYQAAAQWGATWS------NASRLEVDYVKVSAL----------------
Tcas GH16-1 MYLSIGVGVGGH--CFEDRS---DGSKPWK-NSDPKGQKNFYKASAQWLPTWD------NSSVLKVDYVKIWAL----------------
Dpon GH16-1 MHIVVGVGAGGH--NFDDRS---DGTKPWF-NNQPISQKEFYKARNQWQSSWK------TEAKLQVEYVKVWALD---------------
Dpon GH16-5 FHLLINLAIGATTGYFPDEANN-PGGKPWR-IGSPTAMTDFWQGKSQWEPTWN---RNTDDSHFIIDYIQVFAI----------------
Dmel GH16-1 FYLTFGLSVGG----FNEYQ---HEIKPWN-ERAPQAQKAFWKEVKKIRDHWL------DEGHMKIDYVKVYSL----------------
Bmor GH16   FYLIINLAVGGTNGFFPDGVKN-PIPKPWW-NNSPTAATDFWNGQGGWLPTWNLNVNDGQDASLQVDYVRVWAL----------------
Sfru GH16   FYLIMNVAVGGTNGFFPDGVSN-PSPKPWW-NGSPTAPRDFWNARSAWLNTWNLNVNDGQDASMQVDYVRIWAL----------------
Pame GH16-1 FNFILNVAVGGT--FFPDNLGN----RPWSWDGHP--MRDFWERRSEWLPTWH-----EEDAAMKIDYIRVYQ-----------------
Pame GH16-2 FYFVLNLACGGVNYYFPDDAQN-PGGKPWL-NTSPAASTDFWNGKNQWLPTWNLDVNNGESAAMQVDYIKVWAL----------------
Cant GH16   FFIILNLAVGGN--FFGGEPFDPSESDGWA------------------------------KNTFEVEYVKKWTWN---------------
Calb GH16   FHLILNVAVGGD--FFADGDY--DVPKPWG-GHNP--MRSFWEARHSWENTWK-----GDEVAMVVDYIEMIPH----------------
Hhan GH16   MSLILNVAVGATNGYFQDSWHNTPHAKPWK-NNSPTAMMDFWKSKQQWQSTWH-----GEDVAMKVKSVKMIQY----------------
Cgig GH16   FFIILNVAVGGV-GFFPDKFRNSPYPKPWN-DKSEFTARDFWNHKSQWYPTWNPDQNDGEQAAMQVDYIRVWKMKP--------------
Cint GH16   FYLILNVAVGGTNGFFPDTWTNGKGAKPWN-NNSPTAFKDFWMGKNSWYPTWQPDVNNGENAAMQVKTIRVWAK----------------
Aque GH16   YYLIMNVAVGGTTGFFPDGPH-----KPWN-NTSPTSVNQFYDAKSSWYPTWD-----GDKSALKIDSVRVWTYSDGATNSPGSGGAKET
```

Figure S4.7. Multiple alignment of two GH16 protein sequences identified from *D. v. virgifera* with orthologs from other species. Species included are: *D. v. virgifera* (Dvir), *Tenebrio molitor* (Tmol, Q76DI2.1), *Tribolium castaneum* (Tcas, XP_972063.1), *Dendroctonus ponderosae* (Dpon, AEE61901.1), *Drosophila melanogaster* (Dmel, AAF33851.1), *Bombyx mori* (Bmor, NP_001159614.1), *Spodoptera frugiperda* (Sfru, ABR28478.2), *Periplaneta americana* (Pame, ABR28480.1 and AFR46666.1), *Cryptopygus antarcticus* (Cant, ACD93221.1), *Chlamys albidus* (Calb, AAZ04385.1), *Haliotis discus hannai* (Hhan, BAH84971.1), *Crassostrea gigas* (Cgig, BAG82629.1), *Ciona intestinalis* (Cint, XP_002126690.1), and *Amphimedon queenslandica* (Aque, XP_003388466.1). Potential residues for the catalytic nucleophile and proton donor are highlighted with magenta and green (Mertz et al. 2009; Bragatto et al. 2010).

(a)

```
Dvir  ------------------------------------------------------------------------------------------------------------------------
Ylip  MSLQLLIDETGNFTDPSGKAVILRGINVAADAKLPAKPFTPSQQKA-GDDFYD--TTVSFVGSPFPLEEADEHFARIKAWGFNTIRYIYTWEALEHEGPGVYDEEFIDYTIAVLRKIGE-
Vdah  -----------------------------------------------TD-----------------------------------------------------------------------
Agos  MLGKIYISQQGEFTDYEGNVVQLRGVNLDPSVKFPQQPRIPTNMPV-DDEFWDGATNVSFVNERLDPKEIEEHMIRLKALGYNCIRYLFTWEALEHGGPGIYDEEYMKYTVMVLKKIKEA
Scer  MPAKIHISADGQFCDKDGNEIQLRGVNLDPSVKIPAKPFLSTHAPIENDTFFEDADKVSFINHPLVLDDIEQHIIRLKSLGYNTIRLPFTWESLEHAGPGQYDFDYMDYIVEVLTRINSV

Dvir  ------------------------------------------------------------------------------------------------------------------------
Ylip  -HGMFAFMDPHQDVWSRFTGGSGAPLWTLYAAGLDPRHCMTTHSALVQNLWDNP----------SKFPKMIWSTNYQKLACQVMFTLFFAGNHFAPKCIINGVNVQDYLQGSFLAAKRHL
Vdah  ------------------------------------------------------------------------------------------------------------------------
Agos  -GGMYVYLDPHQDVWSRFSGGSGAPLWTLHCAGFQPKRFLATEAAILHNYYIDSETQAE----KAQYPEMIWSTNYYRLACQTMFTLFFSGKLFAPKCVINGRNIQDYLQGHFLKAVMTF
Scer  QQGMYIYLDPHQDVWSRFSGGSGAPLWTLYCAGFQPANFLATDAAILHNYYIDPKTGREVGKDEESYPKMVWPTNYFKLACQTMFTLFFGGKQYAPKCTINGENIQDYLQGRFNDAIMTL

Dvir  ------------------------------------------------------------------------------------------------------------------------
Ylip  AERIA--VDQHLVENVVIGWESVNEPNHGLIGYENIHAIPDSQKLRLGPTPTAFECMRMGMGETVEVDNYEFGPFGATKNGTVVIEPKGTLAWL--KDFSECDKIYGWTRGPEWLPGMCI
Vdah  ------------------------------------------------------------------------------------------------------------------------
Agos  YKYIQDNAPELFEENCIIGLETMNEPNCGYLDHPNLRELPRDRQLMKGTTPTAYQSFILGEGFACNIDSYDISLIGARKIGKSFVDPKGKSAWLDATERLELDRAYGWTRPDDWAPG-CI
Scer  CARIKEKAPELFESNCIIGLESMNEPNCGYIGETNLDVIPKERNLKLGKTPTAFQSFMLGEGIECTIDQYKRTFFGFSKGKPCTINPKGKKAWLSAEERDAIDAKYNWERNPEWKPDTCI

Dvir  ----------------------------------------------------------------------------DDDDVA-QFDITSRVVYAPHWYDGLTLLNKRWN-FFNIDYLGVK
Ylip  WAQHGVWEPKT----GKLLKPTYFNDGHSFHGIGSKIDEEVWVNKYFLGYWLAFLATIRQVNKDWLVLMQAPVMQVPPDLVNHPEFNDKRIVYSPHYYDGLTLMNKKWNRLYNVDVVGIL
Vdah  ----------------TLLKKDYFKKNPK---TGLEYTFPNWTQTHFMDGYRRYRDAIRAIHTDCIMIMQYPTLELPPKIKG-TEDDDPKMAFAPHWYDGITLMTKKWNKLWNVDVVGVL
Agos  WRLHGVWDIESKSSKPVLLLPGYFSKCPS---TGEETSMSYFTNKLFLDFYVRYRNQYRELDPDSLLFLEPPVLQEPPYLIG-SDIIDKRTVYACHFYDGMSLMFKSWNRRYNVDTFGFM
Scer  WKLHGVWEIQN-GKRPVLLKPNYFSQPDA---T-------VFINNHFVDYYTGIYNKFREFDQELFIIIQPPVMKPPPNLQN-SKILDNRTICACHFYDGMTLMYKTWNKRIGIDTYGLV

Dvir  RGRYPNYAMAVKIGDKAIRECFRNQLAWIKEEGQGAIGQH-PTVIGEIGIPYDMNGGKSYRSNG--------------------------------------------------------
Ylip  RGKYPSIVLGLRVGESAIRNCLRDQLRFLRKEGLAKIGNF-PCLISEIGIPYDMDDKYAYRTGDYSQQIRALDANQYALEGSKLH-YTLWVYTASNNHKWGDNWNGEDLSLYSKDDA---
Vdah  RGRYWTPALAVKVGETAIRNCFRDQHNYLYKEGKEHLGNH-PCIMTEFGIPYDMDDHYAYKTGDYTSQSAAMDANYFGVEGSGMEGHCLWLYT--NTHEYGDQWNGEDLSIFSHDDKLLP
Agos  RGKYLSPIFGLVFGEANIKRCFRRQLRAMKLEGRRFLGDSVPIFFTEIGMPYDMEGKKAYRDHDYSSQIGANDALGFALEGSNMS-FSLWCYTYINNTTWGDNWNREDFSIWNKEYA-MK
Scer  NKKYSNPAFAVVLGENNIRKCIRKQLSEMQKDAKSMLGKKVPVFFTEIGIPFDMDDKKAYITNDYSSQTAALDALGFALEGSNLS-YTLWCYCSINSHIWGDNWNNEDFSIWSPDDKPLY

Dvir  ------------------------------------------------------------------------------------------------------------------------
Ylip  ----AKQLQKYGGATQTLTNGSADGSQSSEETPPPTYTSYASYYLDSSYLGKTSIGKSIKGRVSSIKGAIRRRNKTAAVPLSSHGDAFKPPPEYVLGARAGEAFIRPCPQVISGKLDSYG
Vdah  TSPAAAAPGPQGE-----------------------------------------------------------------------------------------------------------
Agos  VPRDVVVKTGDAMPNSSINTIVGAESH----------------LTCESRLSDDA----LVLDYS--------------------------------GFRALDAILRPYPVKIHGSFSTAE
Scer  HDTRARTPTPEPSPASTVASVSTSTSKSGSSQPP-------SFIKPDNQLDLDSPSCTLKSDLS--------------------------------GFRALDAIMRPFPIKIHGRFEFAE

Dvir  --------------------------------------------------------------------------------------------------------
Ylip  FDLQKSVFTLKIKGAACGENDKCEGKLLPTTIYLPHYHFLQWATGVSTSSGKWEYDEN-TQILTWWHYEGPQQLQV---KGNIRFITDYIDTANNLSSSQCRSQ
Vdah  -------------------------------------------------------------------------------------------------------F
Agos  FDLERKRYFLEIIARTETEGTT--------SIFLPYYHFPPESTVVSSSSGYYVREQDNNQLLKWCHGGGRQYISIEVTGMGSSYSVQSADSS-------CVIM
Scer  FNLCNKSYLLKLVGKTTPEQIT-----VPTYIFIPRHHFTPSRLSIRSSSGHYTYNTD-YQVLEWFHEPGHQFIEI-CAKSKSRPNTPGSDTSNDLPA-ECVIS
```

(b)

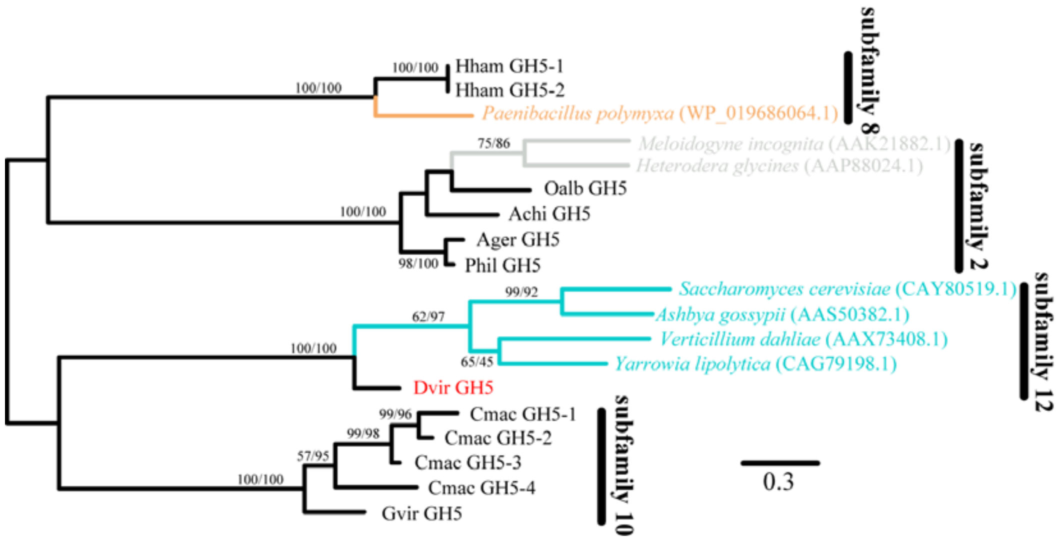

100/100
Hham GH5-1
Hham GH5-2
100/100
Paenibacillus polymyxa (WP_019686064.1)
subfamily 8
75/86
Meloidogyne incognita (AAK21882.1)
Heterodera glycines (AAP88024.1)
Oalb GH5
Achi GH5
100/100
Ager GH5
Phil GH5
98/100
subfamily 2
99/92
Saccharomyces cerevisiae (CAY80519.1)
62/97
Ashbya gossypii (AAS50382.1)
Verticillium dahliae (AAX73408.1)
100/100
65/45
Yarrowia lipolytica (CAG79198.1)
Dvir GH5
subfamily 12
99/96
Cmac GH5-1
99/98
Cmac GH5-2
57/95
Cmac GH5-3
100/100
Cmac GH5-4
Gvir GH5
subfamily 10
0.3

Figure S4.8. Multiple alignment of the potential GH5 protein sequences identified from *D. v. virgifera* with four fungal GH5 sequences (a) and their phylogenetic relationships with other known GH5 sequences (b). The potential amino acid residues for the catalytic nucleophile and catalytic proton donor are highlighted with magenta and green in the

alignment (Larsson et al. 2006). Coleopteran proteins included in the phylogeny are found in Supplementary Table S5. The *D. v. virgifera* sequence is shown in red. Four GH5 subfamilies (2, 8, 10, and 12) are classified according to Aspeborg et al. (2012). The bacterial, nematode, and fungal sequences are indicated by brown, grey, and cyan. The numbers at internal branches show the bootstrap support values (%) from 1000 pseudo-replications for maximum-likelihood and the neighbor-joining phylogenies in this order. Only bootstrap values higher than 70% are shown.

```
Dvir_GH31-1  MVS---RLVLALVGLISTVNGLDNGLARTPPMGWMDWQRFRCLTNCTLYPDECISEKLFR
Dvir_GH31-2  MYKIWFVLAVVVFYLGIDVTPLENGLARTPPMGWLAWERFRCNTDCKNDPENCISENLFR

Dvir_GH31-1  DMADRMAADGYLAAGYEYIMIDDCWSSKERDSKGRLVPDPDRFPSGIKNLSDYIHSKGLK
Dvir_GH31-2  TMADILVNEGYASVGYEYINVDDCWLEKDRSVYGELVPDRVRFPRGMKSLADYVHSKGLK

Dvir_GH31-1  FGIYADYGTLTCAGYPGSKEYLKIDADRFAEWEVDYLKFDGCNSDWIFIDKGYIEMGKHL
Dvir_GH31-2  FGIYEDYGNYTCAGYPGVLGSLQRDAETFASWDVDYVKLDGCYAHPRDMDRGYPEFGFHL

Dvir_GH31-1  NATGRPIVYSCSWPAYQEPNKMQSNYTALAETCNLWRNWDDIDDSWESVTSIIEWFSDNQ
Dvir_GH31-2  NRTGRAMIYSCSWPVYQIYAGMSPNFSAIIEHCNMWRNFDDIQDSWTSVESIIDYYGNNQ

Dvir_GH31-1  DRIGPFSAPGHWNDPDMLVIGNFGLSFEQSKGQMSVWSVMAAPLIMSVDLRTIEPKFRAI
Dvir_GH31-2  DVLIANAGPGHWNDPDMLIIGNFGLSYEQSKTQMAIWAILAAPLLMSVDLRTIRPEYKAI

Dvir_GH31-1  LLNKDAIAVNQDPLGEMGRLVLKKNNIYIWTKKLTAKADGRQPHAIAVLSQRTDGYKYRM
Dvir_GH31-2  LQNRKIIAVDQDPLGIQGRRIYKHKGIEIWSRPITPLYQSYFSYAIAFVNRRTDGTPSDV

Dvir_GH31-1  EFTLKDLNITGPNGFLIKDIFDEDKSVASIADDEPFVLRMAPTGGTLLVATPK-------
Dvir_GH31-2  AVTLKELGLTSPTGYRVEDLY-EDVDYGVLSPQTKIKVKVNPSGVVILRADVQADFNRRI

Dvir_GH31-1  -------------------------K*
Dvir_GH31-2  PFFTTQRPFSSSPLNQVFRVRENGFK*
```

Figure S4.9. GH31 protein sequences identified from the *D*. *v*. *virgifera* transcriptome.

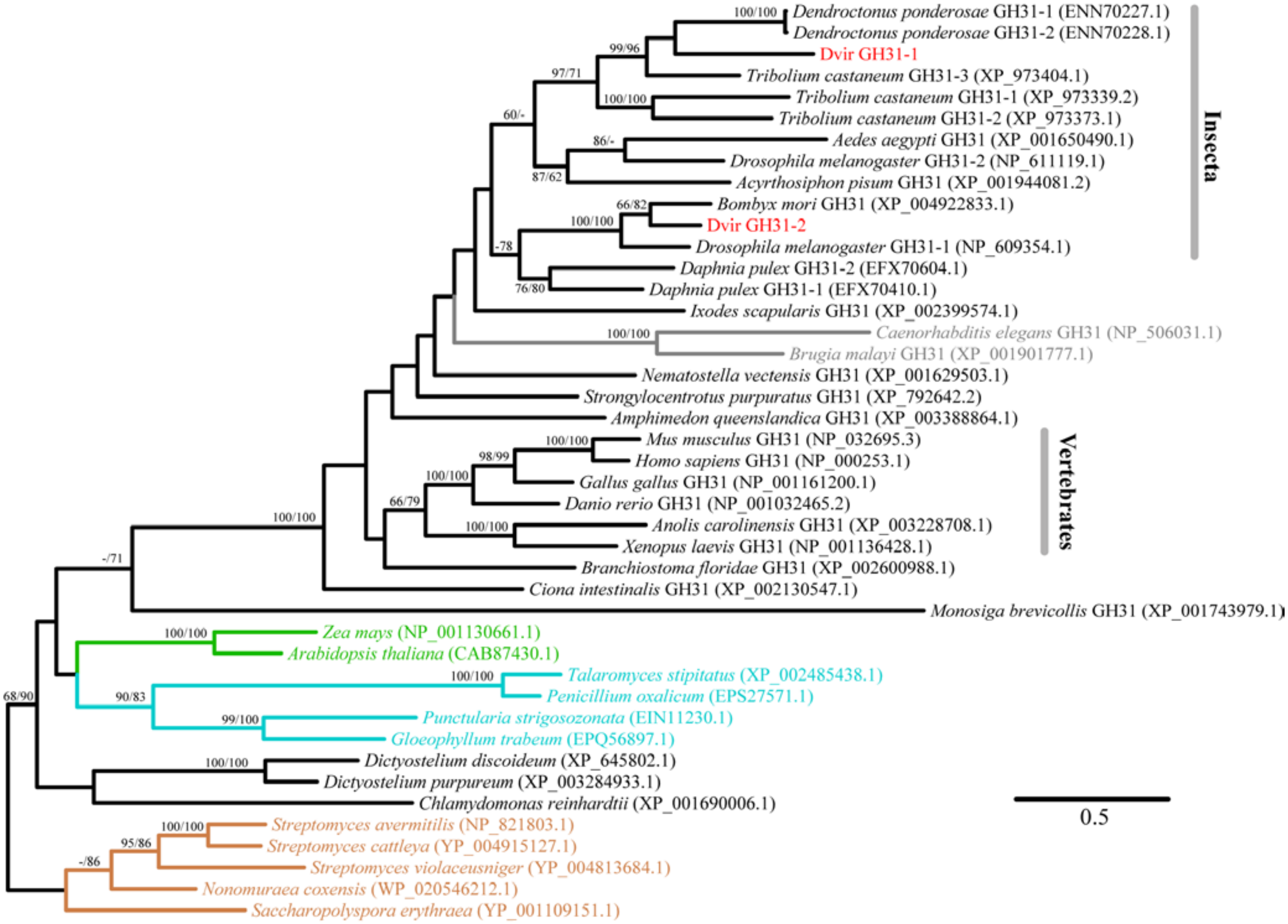

100/100
Dendroctonus ponderosae GH31-1 (ENN70227.1)
Dendroctonus ponderosae GH31-2 (ENN70228.1)
99/96
Dvir GH31-1
97/71
Tribolium castaneum GH31-3 (XP_973404.1)
100/100
Tribolium castaneum GH31-1 (XP_973339.2)
60/-
Tribolium castaneum GH31-2 (XP_973373.1)
86/-
Aedes aegypti GH31 (XP_001650490.1)
Drosophila melanogaster GH31-2 (NP_611119.1)
87/62
Acyrthosiphon pisum GH31 (XP_001944081.2)
66/82
Bombyx mori GH31 (XP_004922833.1)
100/100
Dvir GH31-2
-/78
Drosophila melanogaster GH31-1 (NP_609354.1)
Daphnia pulex GH31-2 (EFX70604.1)
76/80
Daphnia pulex GH31-1 (EFX70410.1)
Ixodes scapularis GH31 (XP_002399574.1)
100/100
Caenorhabditis elegans GH31 (NP_506031.1)
Brugia malayi GH31 (XP_001901777.1)
Nematostella vectensis GH31 (XP_001629503.1)
Strongylocentrotus purpuratus GH31 (XP_792642.2)
Amphimedon queenslandica GH31 (XP_003388864.1)
100/100
Mus musculus GH31 (NP_032695.3)
98/99
Homo sapiens GH31 (NP_000253.1)
100/100
Gallus gallus GH31 (NP_001161200.1)
66/79
Danio rerio GH31 (NP_001032465.2)
100/100
Anolis carolinensis GH31 (XP_003228708.1)
100/100
Xenopus laevis GH31 (NP_001136428.1)
Branchiostoma floridae GH31 (XP_002600988.1)
-/71
Ciona intestinalis GH31 (XP_002130547.1)
Monosiga brevicollis GH31 (XP_001743979.1)
100/100
Zea mays (NP_001130661.1)
Arabidopsis thaliana (CAB87430.1)
100/100
Talaromyces stipitatus (XP_002485438.1)
68/90
90/83
Penicillium oxalicum (EPS27571.1)
99/100
Punctularia strigosozonata (EIN11230.1)
Gloeophyllum trabeum (EPQ56897.1)
100/100
Dictyostelium discoideum (XP_645802.1)
Dictyostelium purpureum (XP_003284933.1)
Chlamydomonas reinhardtii (XP_001690006.1)
0.5
100/100
Streptomyces avermitilis (NP_821803.1)
95/86
Streptomyces cattleya (YP_004915127.1)
-/86
Streptomyces violaceusniger (YP_004813684.1)
Nonomuraea coxensis (WP_020546212.1)
Saccharopolyspora erythraea (YP_001109151.1)
Insecta
Vertebrates

Figure S4.10. The maximum-likelihood phylogeny of representative GH31 family proteins. *D. v. virgifera* sequences are shown in red. Bacterial, fungal, nematode, and plant sequences are indicated by brown, cyan, grey, and green, respectively. Bacterial sequences were used as outgroups. The numbers at internal branches show the bootstrap support values (%) from 1000 pseudo-replications for maximum-likelihood and the neighbor-joining phylogenies. Only bootstrap values higher than 60% are shown.

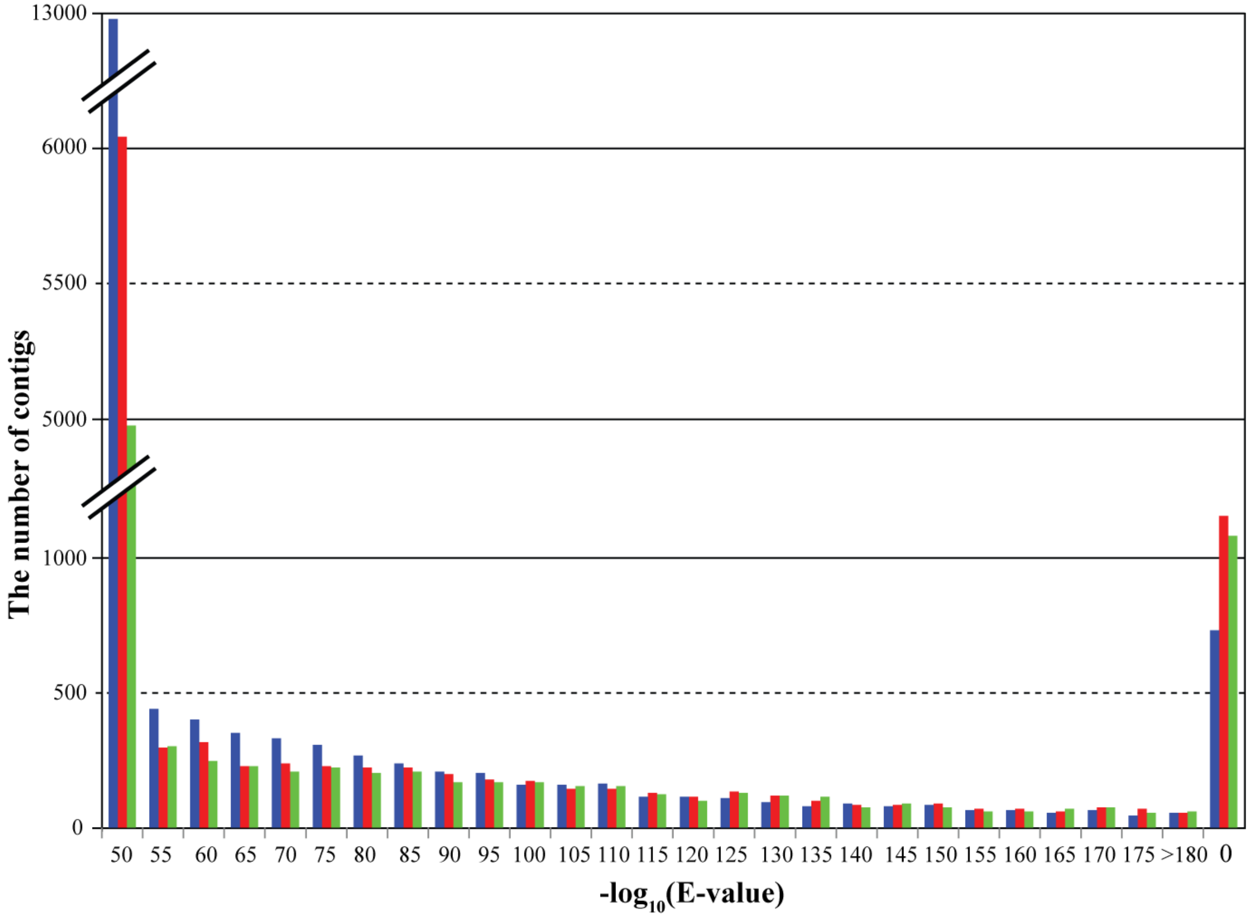


Figure S4.11. The distribution of blastp e-values against the UniProt database using three transcriptome assemblies. The numbers of contigs are 18,173 in Mira (blue), 11,035 in Trinity (red), and 9843 in Velvet/Oasis (green). E-values are transformed to the negative logarithm except for E-value=0. Note that there is no significant difference between Trinity and Mira (*t*-test $P > 0.5$ for both E-value $\leq 10^{-100}$ and using all E-values).

Literature Cited

Arnold K, Bordoli L, Kopp J, Schwede T. 2006. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics*. 22:195-201.

Aspeborg H, Coutinho PM, Wang Y, Brumer H, Henrissat B. 2012. Evolution, substrate specificity and subfamily classification of glycoside hydrolase family 5 (GH5). *BMC Evol Biol*. 12:186.

Ballesteros JA, Weinstein H. 1995. Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. *Methods Neurosci*. 25:366-428.

Bragatto I, Genta FA, Ribeiro AF, Terra WR, Ferreira C. 2010. Characterization of a β-1,3-glucanase active in the alkaline midgut of *Spodoptera frugiperda* larvae and its relation to β-glucan-binding proteins. *Insect Biochem Mol Biol*. 40:861-872.

Calderón-Cortés N, Quesada M, Watanabe H, Cano-Camacho H, Oyama K. 2012. Endogenous Plant Cell Wall Digestion: A Key Mechanism in Insect Evolution. *Annu Rev Ecol Evol Syst*. 43:45-71.

Churcher AM, Taylor JS. 2011. The Antiquity of Chordate Odorant Receptors Is Revealed by the Discovery of Orthologs in the Cnidarian *Nematostella vectensis*. *Genome Biol Evol*. 3:36-43.

Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res*. 14:1188-1190.

Hashiguchi Y, Nishida M. 2007. Evolution of Trace Amine-Associated Receptor (TAAR) Gene Family in Vertebrates: Lineage-specific Expansions and Degradations of a Second Class of Vertebrate Chemosensory Receptors Expressed in the Olfactory Epithelium. *Mol Biol Evol*. 24:2099–2107.

Hayden S, Bekaert M, Crider TA, Mariani S, Murphy WJ, Teeling EC. 2010. Ecological adaptation determines functional mammalian olfactory subgenomes. *Genome Res*. 20:1-9.

Kall L, Krogh A, Sonnhammer ELL. 2007. Advantages of combined transmembrane topology and signal peptide prediction-the Phobius web server. *Nucleic Acids Res*. 35:W429-432.

Kleinau G, Pratzka J, Nurnberg D, Gruters A, Fuhrer-Sakel D, Krude H, Kohrle J, Schoneberg T, Biebermann H. 2011. Differential modulation of Beta-adrenergic receptor signaling by trace amine-associated receptor 1 agonists. *PLoS ONE*. 6:e27073.

Larsson AM, Anderson L, Xu B, Munoz IG, Uson I, Janson JC, Stalbrand H, Stahlberg J. 2006. Three-dimensional crystal structure and enzymic characterization of beta-mannanase Man5A from blue mussel Mytilus edulis. *J Mol Biol*. 357:1500-1510.

Mertz B, Gu X, Reilly PJ. 2009. Analysis of functional divergence within two structurally related glycoside hydrolase families. *Biopolymers*. 91:478-495.

Nei M, Niimura Y, Nozawa M. 2008. The evolution of animal chemosensory receptor gene repertoires: roles of chance and necessity. *Nat Rev Genet*. 9:951-963.

Niimura Y. 2009. On the Origin and Evolution of Vertebrate Olfactory Receptor Genes: Comparative Genome Analysis Among 23 Chordate Species. *Genome Biol Evol*. 2009:34-44.

Nygaard R, Frimurer TM, Holst B, Rosenkilde MM, Schwartz TW. 2009. Ligand binding and micro-switches in 7TM receptor structures. *Trends Pharmacol Sci*. 30:249-259.

Parsiegla G, Reverbel C, Tardif C, Driguez H, Haser R. 2008. Structures of mutants of cellulase Cel48F of Clostridium cellulolyticum in complex with long hemithiocellooligosaccharides give rise to a new view of the substrate pathway during processive action. *J Mol Biol*. 375:499-510.

Rosenkilde MM, Benned-Jensen T, Frimurer TM, Schwartz TW. 2010. The minor binding pocket: a major player in 7TM receptor activation. *Trends Pharmacol Sci*. 31:567-574.

Sakamoto K, Toyohara H. 2009. Molecular cloning of glycoside hydrolase family 45 cellulase genes from brackish water clam *Corbicula japonica*. *Comp Biochem Physiol B*. 152:390-396.

Taylor WR. 1997. Residual colours: a proposal for aminochromography. *Protein Eng*. 10:743-746.

Warne T, Edwards PC, Leslie AG, Tate CG. 2012. Crystal Structures of a Stabilized β_1-Adrenoceptor Bound to the Biased Agonists Bucindolol and Carvedilol. *Structure*. 20:841-849.

Warne T, Moukhametzianov R, Baker JG, Nehme R, Edwards PC, Leslie AGW, Schertler GFX, Tate CG. 2011. The structural basis for agonist and partial agonist action on a β_1-adrenergic receptor. *Nature*. 469:241-244.