Androgen-binding protein (Abp) evolutionary history: Has positive selection caused fixation of different paralogs in different taxa of the genus Mus ?

Robert Karn, Golbahar Yazdanifar, Željka Pezer, Pierre Boursot, Christina Laukaitis

Androgen-binding protein (Abp) evolutionary history: Has positive selection caused fixation of different paralogs in different taxa of the genus *Mus*?

Robert C. Karn[1], Golbahar Yazdanifar[2], Željka Pezer[3], Pierre Boursot[4] and Christina M. Laukaitis[5]

[1]Butler University, Indianapolis, IN, 46208, USA; [2]Department of Medicine, College of Medicine, University of Arizona, Tucson, Arizona 85724 USA; [3]Ruđer Bošković Institute, Bijenička 54, 10000 Zagreb, Croatia; [4]Institut des Sciences de l'Evolution Montpellier, Université de Montpellier, CNRS, IRD, 34095 Montpellier cedex 5, France; [5]Carle Health and Carle Illinois College of Medicine, University of Illinois, Urbana-Champaign, IL 61801 USA

[2]gyazdanifar@arizona.edu; [3]zpezer@irb.hr; [4]pierre.boursot@umontpellier.fr; [5]laukaiti@illinois.edu

Corresponding author: Robert Karn, Department of Biology (emeritus), Butler University, Indianapolis, IN, USA rkarn@butler.edu

Key words: Androgen-binding protein; gene family expansion; alternative paralogs; copy number variant; positive selection; structural variation

Running title: Evolutionary history of the *Abp* expansion in *Mus*

Significance: The *Androgen-binding protein (Abp)* gene family is much larger in the house mouse than in the rat and most other vertebrates. This comparative genomics study identified 206 unique *Abp* sequences in six *Mus* taxa to trace their evolution through the genus. We previously suggested that salivary-expressed alleles of *Abpa27* are under positive selection. We found evidence for duplication of *Abpa27* in two taxa and asked whether previous selection tests were done with the assumption that the *Abpa27* genes were orthologous when they were not. Our new evidence suggests that positive selection acted on paralogs rather than alleles. Deletion of different *Abpa27* gene copies in different taxa, therefore, may better explain the apparent positive selection than fixation of different alleles.

Abstract

Comparison of the *Androgen-binding protein* (*Abp*) gene regions of six *Mus* genomes provides insights into the evolutionary history of this large murid rodent gene family. We identified 206 unique *Abp* sequences and mapped their physical relationships. At least 48 are duplicated and thus present in more than two identical copies. All six taxa have substantially elevated LINE1 densities in *Abp* regions compared with flanking regions, similar to levels in mouse and rat genomes, although non-allelic homologous recombination (NAHR) seems to have only occurred in *Mus musculus domesticus*. Phylogenetic and structural relationships support the hypothesis that the extensive *Abp* expansion began in an ancestor of the genus *Mus*. We also found duplicated *Abpa27*'s in two taxa, suggesting that previously reported selection on *a27* alleles may have actually detected selection on haplotypes wherein different paralogs were lost in each. Other studies reported that *a27* gene and species trees were incongruent, likely because of homoplasy. However, L1MC3 phylogenies, supposed to be homoplasy-free compared to coding regions, support our paralog hypothesis because the L1MC3 phylogeny was congruent with the *a27* topology. This paralog hypothesis provides an alternative explanation for the origin of the *a27* gene that is suggested to be fixed in the three different subspecies of *Mus musculus* and to mediate sexual selection and incipient reinforcement between at least two of them. Finally, we ask why there are so many *Abp* genes, especially given the high frequency of pseudogenes and suggest that relaxed selection operates over a large part of the gene clusters.

Introduction

With the development of dideoxy chain-termination DNA sequencing (Sanger, et al. 1977), the sequences of individual genes and early genomes produced a wealth of single-nucleotide polymorphism (SNP) and small insertion/deletion (indel) data. These have been invaluable for studies of molecular evolution, but since then, it has become apparent that SNPs and small indels explain only a small proportion of trait heritability (Manolio, et al. 2009; Eichler, et al. 2010). Massively parallel sequencing and microarray methods led to the discovery of small microscopic and sub-microscopic variations, structural and quantitative chromosomal rearrangements referred to collectively as structural variation (SV; Alkan, et al. 2011; Huddleston and Eichler 2016). SV includes duplications, deletions, copy-number variants (CNVs), insertions, inversions and translocations ≥50 bp in length (Sudmant, et al. 2015). They contribute substantially to the genetic diversity of genomes and therefore are of much interest for studies of cancer genetics, rare diseases and evolutionary genetics (Spielmann, et al. 2018).

SVs have been important in evolutionary studies, e.g. the role of CNVs in reproductive isolation, a requirement for biological speciation (Loire, et al. 2017). They are also of interest in studies of the evolution of gene families in genomes (Higuchi, et al. 2010; Young, et al. 2016; Pezer, et al. 2017; Zhou, et al. 2017; Wang, et al. 2018; Clifton, et al. 2020). Gene families are important sources of phenotypic diversification and genetic innovation and are therefore central to understanding phenotypic change and the molecular origins of heritable variation (Janousek, et al. 2016; Clifton, et al. 2020). Their early genomic origins are, however, still poorly understood because the evolutionary processes that produce tandemly arrayed genes (TAGs) in these families are obscured by the high degree of sequence identity that have arisen from recent and rapid gene duplications (Pan and Zhang 2008; Karn and Laukaitis 2009; Pavlopoulou, et al. 2010; Uriu, et al. 2021). We have studied and described the extensive *Androgen-binding protein* (*Abp*) gene family in the mouse genome (mm10; hereinafter "reference genome"). Our long-term goals are to understand the origin of this large and recently expanded gene family and to trace the evolutionary history of the expansion, including the role of SV, especially CNV, the mechanisms of duplication, and the contributions of retrotransposons (RTs).

ABPs are members of the secretoglobin (SCGB) superfamily. These small, soluble cytokine-like proteins share significant amino acid sequence with uteroglobin (UG; Karn 1994; Laukaitis, et al. 2005) and share the UG tertiary structure of a four-helix bundle in a boomerang configuration (Callebaut, et al. 2000). The first SCGB superfamily member identified was blastokinin (Krishnan and Daniel 1967), which was renamed UG when it was found to be secreted in large amounts by the rabbit endometrium around the time of embryo implantation (Beier 1968; reviewed in Mukherjee and Chilton 2000). The precise physiological role of UG is not yet known despite the production of two mouse lines with targeted disruptions of different regions of the UG-encoding gene (*SCGB1A1*; reviewed in Mukherjee, et al. 2007). Most of the emphasis of early work was on biochemistry and physiology, although unfortunately the SCGB research field has contracted substantially since the 2000 review. SCGBs are thought to be involved in autocrine, paracrine and endocrine signaling as immunomodulating agents, a role best understood for the SCGB3A2 protein that reduces airway inflammation and injury in mice (Chiba, et al. 2006; Cai and Kimura 2015; Yoneda, et al. 2016). In the lung, the protein encoded by *SCGB1A1* is called CC10 and CC16 (Club Cell Secretory Protein). This protein is upregulated after tissue damage, in premature infants, and in allergy/asthma (Almuntashiri, et al. 2020), suggesting a role in detoxifying harmful substances inhaled into the lungs.

Mouse ABPs are produced primarily by glands of the face and neck (Laukaitis, et al. 2005; Karn, et al. 2014; Karn and Laukaitis 2014) and, as their name suggests, some bind male sex steroid hormones with a clear specificity (Dlouhy and Karn 1983; Karn 1998; Karn and Clements 1999). Mouse salivary ABPs mediate assortative mate selection based on subspecies recognition that potentially limits gene exchange between subspecies where they meet (Laukaitis, et al. 1997; Talley, et al. 2001; Laukaitis and Karn 2012). There is also evidence that ABP constitutes a system of incipient reinforcement across the European hybrid zone where house mouse subspecies make secondary contact (Vošlajerová Bímová, et al. 2011).

ABP proteins are secreted dimers composed of an alpha subunit, encoded by an *Abpa* gene, connected by disulfide-bridging to a betagamma subunit, encoded by an *Abpbg* gene (reviewed in

Laukaitis and Karn 2012). They are a mammalian novelty with variable evolutionary histories in those mammals studied to date. The last common ancestor of the eutherian and metatherian lineages 180 MYA (van Rheede, et al. 2006) had both *Abpa* and *Abpbg* genes, with three each in the metatherian *Monodelphis* (the opossum; Laukaitis, et al. 2008). By contrast, the entire *Abp* gene region in most mammals amounts to only a single *Abpa-Abpbg* gene pair as exemplified by the ground squirrel (*Spermophilus tridecemlineatus*), an outgroup for the purpose of this study. A few species, however, show varying degrees of expansion of the region (e.g. opossum, cattle, horse, rabbit, house mouse and rat) and evidence to date suggests that these expansions are independent of one another (Laukaitis, et al. 2008). The house mouse has the largest expansion with 64 total genes spanning 3 Mb (i.e. ~0.1% of the total genome; Laukaitis, et al. 2008).

The *Abpa* and *Abpbg* subunit genes in the mouse genome are most commonly associated in 5'-5' pairs, (<*a-bg*> or <*bg-a*>, where the arrows point in the 3' directions), with intervening sequence lengths of 5-20 kb. The mouse reference genome (mm10) has 27 of these gene pairs, called "modules", with ten singletons (Pezer, et al. 2017). The mouse reference genome *Abp* cluster is ten times the size of that in the rat genome (rn3) which has only three modules and no singletons (Laukaitis, et al. 2008; Karn and Laukaitis 2009). These two rodent *Abp* gene families appear to have expanded independently (Emes, et al. 2004) and the 64 *Abp* genes in the mouse genome group into five ancestral clades that may have their origins as far back in the evolutionary past as the ancestor of the genus *Mus* (Laukaitis, et al. 2008).

Our previous investigations examined the expansion of the large *Abp* gene family in the genome of C57BL/6, a laboratory research strain derived primarily from *Mus musculus domesticus*, to develop a history of the evolutionary processes involved (Karn and Laukaitis 2009; Janousek, et al. 2013; Janousek, et al. 2016). The comparative genomics study we report here used high-quality genomes representing enough other mouse species to determine how extensive and how widespread the expansion was in the genus *Mus*. Our goals were to resolve old issues concerning *Abp* gene duplication and to reveal emerging properties of the expansion in house mice. To achieve these goals, we interrogated six recently reported genomes of members of the genus *Mus* and compared their *Abp* gene families. These include: *Mus*

musculus domesticus (strain WSB), *M. m. musculus* (strain PWK), *M. m. castaneus* (strain CAS), *M. spretus* (*spr*), *M. caroli* (*car*) and *M. pahari* (*pah*) as well as *Rattus norvegicus* (*Rn*). Hereinafter we will refer to the subspecies of *M. m. musculus* by their strain names. We identified a total of 206 unique *Abp* genes from the six *Mus* genomes that we relate to the *Abp* genes in the reference genome and to each other. We determined their chromosomal positions and explored the roles of retrotransposons (RTs) in their evolutionary histories. Some of the revelations from this work required us to extensively revise what we thought we knew about *Abp* evolutionary history, and to discard some of it altogether.

We first determined whether the expansion that produced this large family began in an ancestor of the genus *Mus*. If that is so, we would expect that different *Mus* taxa with diverse distributions and habitats would still share an evolutionary history of their *Abp* gene region expansions. The counter, or null, expectation is that their *Abp* evolutionary histories would be dissimilar, even to the point of being unique in some taxa as we have found in the independent rat duplication. Testing this hypothesis requires answering three questions: 1) Did duplication occur primarily by <*a*-*bg*> or <*bg*-*a*> modules? 2) Does a phylogeny of *Abp* modules in these six genomes support five ancestral clades (five deeply rooted, monophyletic gene groups shown in Fig. 3 of Laukaitis, et al. 2008)? and 3) Is the temporal history of the appearance and expansion of those clades consistent among the six taxa?

The results of this genome comparison brought us closer to an understanding of a long-standing conundrum in the evolutionary history of mouse *Abp* genes, specifically that key genes expressed in salivary glands and secreted into saliva have phylogenies non-congruent with the species phylogeny. Karn, et al. (2002) studied the complex history of *Abpa* (later *Abpa27* or *a27*), a gene proposed to participate in a sexual isolation mechanism in house mice. They observed an abnormal intron phylogeny for *a27* with an unexpected topology wherein *M. musculus* is not monophyletic and its subspecies stand as outgroups relative to other Palearctic species (*M. spretus* [*spr*], *M. spicilegus*, and *M. macedonicus*). Could assessing the copy numbers (CN) of *a27* in the lineage of the genus *Mus* resolve this issue?

In this process, we revisited the question of how selection has influenced the expansion history of the *Abp* gene family. The evolution of gene families is still poorly understood and there is sparse evidence

that an increased number of specific genes offers a selective advantage (Hastings, et al. 2009), although changes (increase or decrease) in the copy number of dosage-sensitive genes can cause clear selective disadvantage (reviewed in Harel and Lupski 2018). Early evolutionary studies indicated that CNVs might be advantageous because the genes involved are often those that encode secreted proteins and/or are enriched for "environmental" functions, including olfaction, immunity, toxin metabolism and reproduction. Such genes were reported to be under positive selection because they contain higher than average frequencies of nonsynonymous mutations (Johnson, et al. 2001; Nguyen, et al. 2006; Perry, et al. 2007; Emerson, et al. 2008; Nguyen, et al. 2008; Xue, et al. 2008; Sjodin and Jakobsson 2012). Others, however, have suggested instead that a non-adaptive explanation could account for their previous observations (Nguyen, et al. 2006). Finally, is it possible that these six *Abp* clusters are experiencing a form of genome instability in which large blocks of genes are being gained and lost by non-allelic homologous recombination (NAHR), possibly representing runaway gene duplication (Janousek, et al. 2016)?

Results and Discussion

Comparative genomics of the Abp *gene families in the genus* Mus

Interrogation of six wild-derived mouse genomes identified the following number of *Abp* paralogs (defined as unique DNA sequences) in each taxon (**Table 1; Fig. 1**): 11 in *Mus pahari* (*pah*), 33 in *M. caroli* (*car*), 35 in *M. spretus* (*spr*), 43 in *Mus musculus domesticus* (strain WSB), 38 in *M. m. musculus* (strain PWK), and 46 in *M. m. castaneus* (strain CAS) (**supplementary tables S1-S6**). This compares with 64 found in the reference genome (mm10, C57BL/6, B6; Laukaitis, et al. 2008) and we named the 206 *Mus Abp* sequences on those. We did this by: a) comparing their associations in *Abpa* and *Abpbg* phylogenies with the reference genes and ancestral Clades 1-5 (**Fig. 2; supplementary figs. S1, S2**), b) mapping intra-genome linear relationships relative to those of the paralogs (**Fig. 3**); and c) examining the associations of *Abpa* and *Abpbg* genes in putative modules (**supplementary tables S1-S6**). Hereinafter, specific modules will be abbreviated M followed by a number, both in italics, e.g. *M9* and *M10* for the <*bg9-a9*> and <*bg10-a10*> modules, respectively. We evaluated the numbers of *Abp* genes that could be expressed and the numbers of putative pseudogenes (**supplementary table S7;** see also **Table 1** and **supplementary tables S1-S6**) and compared them to those in the reference genome (Karn, et al. 2014). The result showed high percentages of pseudogenes in the six *Mus* gene families (WSB, 58%; PWK, 47%; CAS, 50%; *spr*, 53%; *car*, 48%; and *pah*, 36%; comparable to mm10, 53%).

Copy number analysis

Initially, we attempted to estimate *Abp* gene copy number with CNVnator software (Abyzov, et al. 2011). That approach yielded suspiciously low numbers of small CNVs across the *Abp* gene regions of the six *Mus* genomes, very likely because of numerous gaps in the 1504 assemblies. Instead, we calculated CNs based on differences in read depth between *Abp* genes and putative single-copy regions (**supplementary table S8**). Each "unique" gene sequence may be present in a number of copies in a diploid organism. Those that have two copies likely have one on each chromosome (i.e. alleles), while those with copy number >2 have been duplicated and any with CN<2 have been subject to deletion. The numbers of unique *Abp* genes and pseudogenes and the inferred total gene numbers, including

duplications, are summarized in **Table 1**. The discrete numerical CN estimates from direct read-depth calculations on the 206 paralogs we found in this study (**supplementary tables S1-S6**) are consistent with previous analyses in the three subspecies of *M. musculus* and in *M. spretus* (Pezer, et al. 2017). Altogether, 85% (40/47) of the copy number variable genes appear in the proximal region (defined as *M1-M12;* **supplementary tables S1-S6**) and 95% belong to ancestral Clade 2, which has the largest number of paralogs. The expansion history of these regions hints at a complex sequence of events causing rapid *Abp* gene expansion in the genus *Mus*.

Technical challenges

The genome data we analyzed are of high quality, however, problems remain:1) the earlier 1504 builds were used to mine *Abp* sequences from the six genomes because they contain genes not found in the later builds; 2) there seem to be assembly problems, including unexpected gene orders, in the 1504 builds; 3) it is not possible to determine the locations of the duplicated gene copies found in the CN analysis; and 4) there are many gaps that make it difficult to estimate efficiency of gene-finding.

Our previous studies have shown extensive structural variation in WSB and much less in PWK, CAS and *spr* (Pezer, et al. 2017). We were able to draw these conclusions because we compared the sequences of multiple individuals. Here, we were not able to detect homozygous deletions given that the CN calculations are based on the genome read depth of a single inbred individual from which the assembly was made. In our study, if there was more than average read depth at a certain locus of the assembled genome, we called it duplication (i.e. amplification), if there was less than average, we called it deletion. Thus, it is reasonable to assume that the variation in this region is even higher than we can see by sequencing and assembling the genome from only one individual. In order to find more paralogous genes and to detect possible copy number variation in them, one would need to sequence more individuals of the same taxon or population.

While the 1504 builds of these genomes provided the largest number of *Abp* genes, they were mapped to the reference genome, which may have created or perpetuated assembly problems. The very high levels of *Abp* sequence identity (≥95%) and the use of short reads may have caused additional issues.

For example, the proximal and some central genes in *spr*, PWK and CAS do not share the same order as they do in the reference genome, nor do any two of them share a single, alternative pattern. This is especially evident in ancestral Clade 1 (i.e. *M1* and *M2* in **Fig. 3**). We suggest several possible explanations for the "scrambled" appearance of the *Abp* genes in the 1504 builds: a) some of them are misidentified; b) the genome builds placed them incorrectly; and/or c) small chromosomal rearrangements occurred locally. The absence of a single, alternative order favors choice b): underlying assembly problems caused by high sequence identity and high density of repetitive sequences.

Assembly problems are expected in genome regions containing segmental duplications (SDs) because they are repeated sequences with high pairwise similarity. SDs may collapse during the assembly process causing the region to appear as a single copy in the assembly when it is actually present in two copies in the real genome (Morgan, et al. 2016). Moreover, individual genes and/or groups of genes may appear to be out of order compared with the reference and other genomes. In some studies, genotyping of sites within SDs is difficult because variants between duplicated copies (paralogous variants) are easily confounded with allelic variants (Morgan, et al. 2016). Latent paralogous variation may bias interpretations of sequence diversity and haplotype structure (Hurles 2002), and ancestral duplication followed by differential losses along separate lineages may result in a local phylogeny that is discordant with the species phylogeny (Goodman, et al. 1979). Concerted evolution may also cause difficulties if, for example, local phylogenies for adjacent intervals are discordant due to nonallelic gene conversion between copies (Dover 1982; Nagylaki and Petes 1982).

The annotations of these sequences were complicated because existing programs for identifying orthologs between sequenced taxa (Altenhoff, et al. 2019) were not applicable to our data. The databases these programs interrogate do not include many of these newly-sequenced taxa of *Mus* and also do not include the complete sets of gene predictions we make here. Thus, we had to manually predict both gene sequences and orthology/paralogy relationships. This is a problem facing other groups working with complex gene families in other non-model organisms (Denecke, et al. 2021). Most importantly, we treated the problem of orthology in our own, original way. Our conclusion is that orthology is not applicable to at

least one of the *Abpa27* paralogs, and possibly to other paralogs (*Abpa26, Abpbg26, Abpbg25;* **Fig. 5**), probably due to the apparent frequencies of duplication and deletion and this is precisely the interesting point of our study.

Comparison of the gene orders of the six *Mus Abp* regions with the reference genome suggests perturbed synteny of many *Abp* genes (**Fig. 3**). Overall, the ***proximal*** region *(M1-12* with some singletons) shows significant differences among the six taxa while the ***distal*** region (*M20-27*, singletons *bg34* and *a30*) has gene orders in the six taxa much more like the same regions in the reference genome. The ***central*** region (from singleton *a29* through *M19*, with some singletons) in WSB is unique in that it includes the penultimate and ultimate duplications, shown above the blue triangle in **Fig. 3** (Janousek, et al. 2013). The order of proximal and distal genes in *car* agrees relatively well with that in the reference genome *Abp* region, considering how early *car* diverged from the lineage compared to *spr*, PWK and CAS. In those three taxa, however, the proximal and some central genes do not share the same order as in the reference genome, nor do any two of them share a single, alternative pattern, especially *M1* and *M2*. We note that **Fig. 3** also shows that functional copies of either *Abpa* or *Abpbg* in some species are 1-to-1 orthologs to pseudogenized versions of these genes, which is interesting from the standpoint of patterns of gene family evolution.

Gaps in the assembly might account for the perturbed synteny among different taxa. They might be true deletions in the sequenced genomes when compared with the mouse reference genome used to guide the genome assembly. Counting from the start position of the first *Abp* gene to the end position of the last gene in the *Abp* gene region, there are in total 1933, 1115, 637, 488, 75 and 32 assembly gaps in the 1504 builds of the genome for WSB, CAS, *spr*, PWK, *car* and *pah*, respectively (**supplementary fig. S4**). We looked for reads with perfect matches that span gap positions and found that only 2-6% of gaps in the *Abp* region could be closed in this way and over 50% of them are not supported by more than two to four reads. At best, only a small fraction of gaps can be considered as truly missing sequences in the *de novo* assemblies, and the majority of the gaps actually represent unknown sequences. In short, there are too many gaps in this region to obtain deeper insights into its structure. This illustrates the technical

limitations related to assembly of this particular region. With deeper sequencing, additional gene and intervening sequences may be identified, and the gene orders may change. For example, a significant number of new *Abp* genes were found in the mouse genome from the first report (Emes, et al. 2004) to the second (Laukaitis, et al. 2008).

Finally, it is important to consider a possible role for gene conversion because if it can be ruled out, then estimates of gene age are reliable. Otherwise, in addition to obscuring orthology, it can also cause problems for estimating the role of natural selection (Casola and Hahn 2009) and the potential role of pseudogenes as donors of new mutations for functional genes in the vicinity (Casola, et al. 2012). Several previous studies suggested that gene conversion has played little if any role in the evolution of the *Abp* gene region. Laukaitis, et al. (2008) discounted the explanation that nonallelic gene conversion caused the low divergence of the *Abp* sequences they studied because the phylogenetic tree of ribosomal protein *L23a* pseudogenes suggests that they frequently co-duplicated with *Abpa-Abpbg* gene modules.

Laukaitis and Karn (2012) pursued this at a smaller scale by analyzing the 64 *Abp* genes in the reference genome with GENECONV to look for evidence of short gene conversion tracks. In the case of the *Abpa* paralogs, GENECONV identified no inner (conversion between genes within alignment) and no outer (conversion with genes outside alignment) fragments that were globally significant, suggesting that there is no compelling evidence of gene conversion in *Abpa* paralogs. In the case of the *Abpbg* paralogs, the analysis identified only one inner (*Abpbg26* and *Abpbg34*) and two outer fragments (*Abpbg5p* and *Abpbg19*) that were globally significant. They also calculated the GC content of the *Abp* gene region because sequences undergoing frequent gene conversion, either ectopic or allelic, are expected to become GC rich (Galtier, et al. 2001; Galtier, et al. 2009). Their results showed that the average GC content in the *Abp* gene region is low (~41–42%) compared with genes undergoing gene conversion, such as ribosomal operons and transfer RNAs which have much higher GC contents (Galtier, et al. 2001). They concluded that gene conversion has made a minimal, but not nonexistent, contribution to the evolutionary history of the *Abp* gene family. Moreover, it certainly was not significant enough to have confounded the

phylogenetic inference presented by Laukaitis, et al. (2008), and it should not have adversely affected their analysis of recently duplicated products.

Overall, despite problems in estimating the efficiency of gene finding because of non-uniformity of read coverage, we propose that we missed relatively few unique gene sequences because we: a) used three different methods of gene-finding to interrogate the six high-quality genomes; b) found genes with disparate sequences, even in taxa diverged many MYR apart; and c) identified novel *Abp* genes without counterparts in the reference genome (e.g. *bgI, bgJ, bgKp, aL, aMp* and *aN* in *car*). Similarity in the numbers of protein-coding and noncoding genes found in the *car* and *pah* genomes compared with the mouse and rat reference genomes (Thybert, et al. 2018) further supports this conclusion.

Contributions of structural variants

There appears to be a pattern of fewer genes in the earlier-diverging taxa (**Fig. 1** and **Fig. 3**) with a large jump from *pah* (11) to *car* (33). To the extent that deletions occurred, they seem clearly to have been outnumbered by duplications, but that may be an ascertainment bias because "presence" is easier to identify than "absence". Insertions and translocations are also not immediately recognizable, however either could have happened locally and that might explain the "scrambled" gene orders in the Palearctics, especially on the left flank of **Fig. 3**.

The *Abp* regions in the Palearctic taxa that we analyzed in this study have much higher LINE1 (L1) densities than their flanking regions, consistent with those in the mouse and rat genomes (Janousek, et al. 2013) and the *car* and *pah* genomes (Thybert, et al. 2018), regardless of how populous the *Abp* genes are in any individual taxon (**supplementary fig. S3**). By contrast, we found intragenic RTs only in *car_a1bp* (a lineage-specific invasion of L1Md_A into Exon 3) and an insertion of IAP1-MM_I-int, a member of the ERVK family of LTRs, into the *a30* paralogs of *spr*, WSB, PWK and CAS.

Pezer, et al. (2017) used CNVnator to show that the penultimate duplication segment (*M7-M12* and *bg33-M19*) of the large-block NAHR duplication pattern in the mouse genome is seen only in WSB and other inbred strains (their Fig. 6), as well as in six wild *Mus* genomes (their Fig. 2). That is consistent with the derivation of the mouse genome (C57BL/6) from an *M. m. domesticus* mouse. Taken together,

these genes, along with those of the ultimate duplication (*M14-M17*) shown above the blue triangle at the top of **Fig. 3** represent these NAHR duplications (Janousek, et al. 2013). By contrast, Pezer, et al. (2017) found no such correspondence in the CNVnator patterns of the other three taxa (PWK, CAS and *spr*) or in those of inbred strains derived from them. Here we have actual paralog data to test the idea that genes that vary in copy number in the other five *Mus* taxa are not in these large, contiguous blocks (**supplementary tables S1-S6**), despite also having enrichment in L1 sequences (**Fig. S3**). Instead, some show unique duplications (e.g. *car_a1ap, a1bp* and *a1cp*; CAS_*a12a* and *a12b*), which are not paired in modules indicating that they might have duplicated after these taxa diverged from the lineage.

The Abp *family expansion: Modules, ancestral clades and the growth of three regions*

We estimate that 154/206 (~75%) of unique *Abp* genes in these six taxa are pair-members of modules (**supplementary tables S1-S6**). The 11 *Abp* genes of *pah* are organized into five <*bg-a*> modules and only one singleton with the *Abpbg* genes on one strand and the *Abpa* genes on the other, as expected of tandemly arrayed genes (TAGs). *car* has a total of 33 genes with 22 in 11 modules; *spr* has a total of 35 genes with 24 in 12 modules; WSB has a total of 43 with 38 in 19 modules; PWK a total of 38 with 30 in 15 modules; and CAS has a total of 46 with 30 in 15 modules.

Because some of the new *Abp* genes we found in the six *Mus* taxa are incomplete, we could not build a phylogeny using only the large intron as done previously for *Abp*'s in the mouse reference genome (Laukaitis, et al. 2008). This raised the concern that positive selection on coding regions would bias the phylogenetic trees. To test for this, we needed an unselected sequence common to both genes within a module. Since most of the *Abp* genes are organized into modules, we searched the intra-modular sequences between the 5' ends of the *Abpa* and *Abpbg* gene partners for common sequences.

Supplementary table S9 shows that a LINE1, L1MC3, is nearly ubiquitous within modules of the reference genome and within those of the six *Mus* taxa (93/103, 90%), suggesting that it is ancestral to the genus *Mus*. It does not occur in any of the three rat *Abp* modules, however, there is an L1MC3 in the single intramodular sequence of the ground squirrel, an outgroup to these taxa. That suggests that this RT was lost in the rat ancestor following divergence from the rodent lineage. **Figure 4** shows a phylogeny

constructed from L1MC3 sequences in the modules of the new genes and the reference genome rooted on the L1MC3 in *pah_M24* because it is the only module common to all six *Mus* taxa (**Fig. 2; supplementary table S1**). The *pah_M24* is also the only *M24* that has an L1MC3, suggesting that this RT was lost from the lineage following divergence of *pah*. In general, the module phylogeny has a topology congruent with the gene (*Abpa* and *Abpbg*) phylogenies in **Fig. 2**, suggesting that the **Fig. 2** phylogenies built on the genes themselves were not biased by the combination of coding regions and introns that were available to use.

The module-based phylogeny we made using L1MC3 was valuable for the insights it provided into the ancestral clades in the reference genome (those most deeply rooted in the *Mus* phylogeny; Laukaitis, et al. 2008). **Figure 4** defines the relationships of *pah* modules to several of these ancestral clades: 1) *pah_M3* and *car_M3* group with the Palearctic *M3*s on the left flank of the large ancestral Clade 2 in the reference genome, while *pah_MU* is basal to the rest of that ancestral clade; 2) *M24* is the sole occupant of ancestral Clade 3 and is found in all six of the *Mus* genomes (**supplementary tables S1-S6** and **Fig. 2**), however, it appears alone here because only the *M24* in *pah* has L1MC3. Comparison of the two *Abp* subunit gene phylogenies in **Fig. 2** with the module phylogeny in **Fig. 4** suggests that Ancestral Clade 1 is more closely related to *M3* than it is to any of the other modules in Clade 2. In fact, the *bg3* clade in the *Abpbg* phylogeny groups with Clade 1, not with Clade 2 as is the case with the *a3* clade. As well, the L1MC3 of *M3* has the shortest branch with Clade 1 in **Fig. 4** and M3 lies physically next to *M2* as might be expected for tandem duplication products at least when it occurred.

Figure 2 shows that the duplication that gave birth to the ancestor of *M25* and the ancestor of *M26-27-MX* occurred in an ancestor of the *Mus* lineage, prior to the divergence of *pah*, since it is older than the divergence between *pah_MX* and *M26-27*. Thus, the duplication that gave rise to *M25* is older than that which gave rise to *M26-27.* The duplication that gave rise to *M1-M2* (Clade 1) must also have occurred previously to the divergence of *pah,* confirming the status of Clade 1 as ancestral.

In summary, Clades 1-5 are confirmed as ancestral, although clearly Clades 4 and 5 are closely related. Clade 2 began expanding in the ancestor of *car* and the Palearctic taxa, and some copies survived

and were duplicated, while other paralogs died after the divergence of the Palearctics (**Fig. 2; supplementary table S2;** see also **Fig. 3**). This clade is larger and more complex in the three subspecies of *M. musculus* and seems to have been the source of most of the volatility identified when comparing the *Abp* gene regions of 15 inbred strains to the mouse genome using the Mouse Paralogy Browser (Karn and Laukaitis 2009).

Modules *M24, MX* and *MY* in *pah* (**supplementary table S2**) may represent the ancestors of the entire right-flank in *car* (the segment in the mouse genome stretching from *M24* to *a30)*. We did not find a "classical" ancestral Clade 1 (*M1-M2*) in *pah*, because *aU*, *bgUp* and *aVp* are not in the reverse order (i.e. switched strands) in relation to the other *pah* genes/modules, as Clade 1 is in the other four taxa (**Fig. 3**). One possibility, however, is that they do represent *pah* Clade 1 but the strands on the other four taxa represent the outcome of an event that occurred between the divergence of *pah* and the other four, perhaps during the massive genome rearrangement that followed divergence of *M. pahari* from the ancestral lineage and before divergence of *M. caroli* 3-6 MYA (Thybert, et al. 2018).

The central gene region (ancestral Clade 2), is smaller and less complex in *pah*, probably only represented by *M3*. However, in *car*, it is comprised of nearly 20 genes: *M3*, three *a28*-like paralogs, eight genes variously related to *M21-23* and six more deeply rooted paralogs (*aL, aMp, aNp, bgI, bgJ*, and *bgKp)*, which likely explains the jump from 11 genes in *pah* to 33 in *car* (see above). The gene numbers making up the populous and volatile central region in the *M. musculus* subspecies are consistently larger than in the other three taxa. Ancestral Clade 4 (*M25*) is seen only in the Palearctic taxa, however, it had to have a progenitor in the ancestor of *Mus* because it is basal to *M26* and *M27* (**Fig. 2** and **4**). So, *M25* was either deleted or we failed to find it in both *pah* and CAS.

Taken together, our observations on the *Abp* gene family expansion, the modules, the Clades and the growth of the three regions, provide strong support for the idea that expansion of the large reference genome *Abp* family began in an ancestor of the genus *Mus.* They also suggest that most or all of the *Abp* genes in these six *Mus* genomes are related as branches within one or another of the five ancestral Clades. The alternative would have been independent expansions, similar to the rat *Abp* region where individual

paralogs are not orthologous with those in the genus *Mus*. Another way of thinking about this is that most of the *Abp*s in *Mus* have orthologs in some or all of the six taxa we studied. That suggests that they evolved from a shared lineage while none of them have orthologs in the rat, which apparently had an independent expansion.

The role of selection in Mus Abp *gene evolution:* R*econciling topologies of the gene and species trees*

Studies of selection on *Abp* genes have focused on *a27, bg27*, and *bg26*, the three saliva-expressed paralogs because of evidence that *Abp* has a role in sexual selection between house mouse subspecies (Laukaitis, et al. 1997; Talley, et al. 2001; Bímová, et al. 2005). Hwang, et al. (1997) observed a high nonsynonymous/synonymous substitution ratio (*dN/dS*) in their *Abpa* (now *a27*) sequence data from six *Mus* taxa and proposed that directional selection was a sufficient explanation of their data. They envisioned the possibility of cyclical selection of certain amino acid variants that became advantageous at some stage and they posited that homoplasy occurred in the phylogeny of the *Abpa* haplotypes that was incongruent with the canonical phylogeny of the genus. Karn and Nachman (1999) used the HKA test (Hudson, et al. 1987) to investigate patterns of DNA sequence variation at *a27* within and between species of mice. Their results provided evidence that selection has shaped the evolution of *Abpa* in house mice and was consistent with a recent adaptive fixation (a selective sweep) at or near *Abpa.* They also calculated the ratio of nonsynonymous substitutions to synonymous substitutions on a per-site basis (*K*a/*K*s) for the *Mus* sequences of Hwang, et al. (1997). Based on the combined observations of no variation at *a27* within *M. m. domesticus* and uniformly high *K*a/*K*s values between species, they suggested that positive directional selection has acted recently at this locus. Laukaitis, et al. (2012) assessed site-specific positive selection on the coding sequences of three genes, *a27*, *bg26* and *bg27*, in five *Mus* taxa using the program CODEML in the PAML package (Yang 2007). They concluded that at least two (*a27, bg26*) of the three genes encoding the subunits of ABP dimers evolved under positive selection and suggested that the third one may have also.

These selection tests were based on the assumption that the *a27* genes in the subspecies of *M. musculus* are orthologs and thus that the studied variants were alleles. However, some genes have a

phylogeny at variance with the species phylogeny and Karn, et al. (2002) suggested that the *M. musculus* taxa are not monophyletic and its subspecies are outgroups relative to other Palearctic species. Here, we provide evidence that *pah* and *car* both appear to have duplications of modules related to *M27*, specifically *MX* and *MY* in *pah*; as well as *M27a* (<*bg27a-a27a*>) and *M26*/*27b* (<*bg26-a27bp*>) in *car* (**Figs. 2, 3,** and **5**). These extra *M27* modules are not found in the Palearctic taxa that have their *a27* topologies incongruent with that of the species phylogeny (Karn, et al. 2002). Such duplications and deletions may also have occurred in the ancestor of the Palearctics, so that the copies we observe now are not necessarily all orthologous. That could provide a parsimonious explanation for why the gene phylogeny is incongruent with the species phylogeny. Interestingly, **Fig. 2** shows that clades *a26*, *bg25* and *bg26* are also noncongruent with the species phylogeny.

Karn, et al. (2002) discussed and discarded an explanation for the incongruent gene and species trees that was based on a hypothetical duplication that produced two copies of *a27* in an early ancestor(s). In this view, different *a27* paralogs were fixed or lost producing very different "*a27*" sequences in *M. m. domesticus* and *M. m. musculus* that were not orthologous. The critical point is that, if duplication of *M27* and related modules led to fixation of different paralogs in *M. m. musculus* and the other Palearctic species/subspecies, then the many selection tests reported for *a27*, *bg26* and other ancestral Clade 5 genes (Karn and Nachman 1999; Laukaitis, et al. 2003; Laukaitis, et al. 2012) were done with the assumption that they were orthologous in all the Palearctic taxa when they were not.

In this study we discovered that *pah* and *car* both have two modules that appear to be duplicated ancestral versions of *M27* (**Fig. 5**). Later in *Mus* evolutionary history, random mutation could have created a situation with two haplotypes segregating in a population, one haplotype having paralog *a27a* with paralog *a27b* deleted and another haplotype that retained paralog *a27b* with paralog *a27a* deleted. These would fit the description of "pseudoalleles" if tandem duplication had produced the two paralogs. Assuming that the population gave rise to two separate migrations (as in the case of the progenitors of *M. m. domesticus* and *M. m. musculus*), selection and/or drift could have increased the frequency of paralog *a27a* in one population and conversely paralog *a27b* in the other, perhaps even to fixation in both. If

individual animals in the two subpopulations could sense the different salivary proteins expressed by the pseudoalleles, it might have led to olfactory recognition resulting in homosubspecific selection and eventually incipient reinforcement. The ancestors of the *M. m. domesticus* and *M. m. musculus* subspecies made secondary contact 5000-10,000 years ago, forming what is now the European mouse hybrid zone (Boursot, et al. 1993; Sage, et al. 1993). It seems clear from the literature that "*a27*", whether orthologous or paralogous in these two subspecies, mediates sexual selection and constitutes a system of incipient reinforcement at the mouse hybrid zone (Vošlajerová Bímová, et al. 2011).

This emergent property highlights perhaps the most important contribution of the module trees because previous explanations of the topology of these genes tended to cite homoplasy as the result of strong positive selection (Hwang, et al. 1997; Karn, et al. 2010; Laukaitis, et al. 2012). One of the reasons we used L1MC3 to build *Abp* module phylogenies is that L1 RTs are thought to be homoplasy-free regions compared to gene regions (Verneau, et al. 1998; Semple and Steel 2002; Alexeev and Alekseyev 2018). Since the abnormal topology in the *a27* phylogeny is not eliminated by building a phylogeny with the intramodular L1MC3 **(Fig. 4**), we conclude that it is the result of SV rather than homoplasy. Moreover, it also shows that other genes in *M25* and *M26*, which are related by descent to *a27* as the result of duplications, also have abnormal topologies. Coupled with the observation that *a27* and *bg27* have duplicates in *pah* and *car*, this better supports the notion that duplication produced two copies of *M27*, which then were differentially eliminated, causing one to be fixed in an ancestor of PWK and the other in an ancestor of the rest of the Palearctic taxa. We feel that this explanation, rather than explanations such as the occurrence of secondary genetic exchanges along the lineages leading to the Palearctic taxa (Karn, et al. 2002), is more parsimonious and better fits the data we report here.

Why are there so many Abp *genes in the genus* Mus?

Only genes found in ancestral Clade 5 of the reference genome are expressed in salivary glands and secreted into saliva, while many more genes from three of the other clades are expressed in lacrimal glands and secreted into tears (Karn, et al. 2014). This led to the proposal that, early in the expansion of the mouse *Abp* gene family, neo- or subfunctionalization occurred to create this clear-cut partitioning of

Abp expression between these glands of the face and neck. Their proposal raises the possibility that the function(s) of ABP proteins in tears differs from those of ABP proteins in saliva. We have already reviewed the functions of the paralogs expressed in ancestral Clade 5 above (mate recognition, sexual selection, incipient reinforcement) and no other functions have yet been found for the *Abp* paralogs in ancestral Clades 1-4.

So why are there so many other *Abp* genes in the members of the genus *Mus*? We have searched repeatedly for other potential functions for *Abp* genes but as yet have found none (Karn and Dlouhy 1991; Chung, et al. 2017). An equally important question is: Why does there appear to be so much genome instability and polymorphism of CNV in naturally occurring WSB populations (Karn and Laukaitis 2009; Pezer, et al. 2017), suggesting runaway gene duplication (Janousek, et al. 2016), and why is this not found in the other Palearctic taxa (Pezer, et al. 2017) and this report?

Nguyen, et al. (2008) questioned their previous proposal (Nguyen, et al. 2006) that CNVs are often retained in the human population because of their adaptive benefit. Rather, they showed that genic biases of CNVs are best explained, not by positive selection, but by reduced efficiency of selection in eliminating deleterious changes from the human population. We propose here that this might also apply to *Abp* genes in mouse populations. They are environmental genes (*sensu* Nguyen, et al. 2006; Nguyen, et al. 2008) associated with SDs and so are subject to frequent duplication, deletion and pseudogene formation (Lander, et al. 2001; Waterston, et al. 2002; Gibbs, et al. 2004; Perry, et al. 2008; Karn and Laukaitis 2009; Sjodin and Jakobsson 2012). This may be why so much volatility has been observed, especially in the central region of the *Abp* cluster in the reference genome (Karn and Laukaitis 2009) and in the six *Mus* genomes we studied here. It could also explain why >50% of *Abp* paralogs in the reference genome are pseudogenes (Karn, et al. 2014) and why we found similarly high percentages of pseudogenes in the six *Mus* taxa. These observations suggest that the environment of the *Abp* gene region may be more permissive for duplication in the sense of retention of pseudogenized duplicates. On the one hand, this might facilitate duplication followed by neo- and/or subfunctionalization (e.g. Karn, et al. 2014) while, on

the other hand, volatility may have resulted from an elevated rate of gene birth and death (Karn and Laukaitis 2009; Janousek, et al. 2013; Pezer, et al. 2017).

Broader evolutionary implications

Given that most of the emphasis of past secretoglobin work was on biochemistry and physiology, there is little information on the evolution of this superfamily, a regrettable situation that may account for the lack of a holistic picture of secretoglobin function(s). We have pursued questions about the evolution of the extensive *Abp* families in house mice but it seems that we and others in the field have been remiss in not considering the cytokine nature of these molecules in an evolutionary context. For example, gene families, such as those involved in chemosensation, reproduction, host defense and immunity, and toxin metabolism that are expanded, usually as tandem duplications, in one lineage are often expanded in another (Janousek, et al. 2013). Cytokines are part of the body's immune response to infection, and considering that in conjunction with their detoxifying capability and with ABP's reproductive function (mate recognition), we propose that there may be a connection among these seemingly diverse capabilities. Karn, et al. (2014) observed that a partitioning of *Abp* expressions occurred early in the evolution of the ancestral clades in the genome of the ancestor of the genus *Mus*. They suggested that neo- and/or subfunctionalization was responsible for partitioning the expression of eleven *Abp* paralogs from ancestral Clades 1, 2 and 3 into lacrimal gland/tears, and three *Abp* paralogs from ancestral Clade 5 into submandibular gland/saliva. Again considering possible multiple functions, this time in the context of the large *Abp* gene family in the mouse, could it be that a small group of these secretoglobins initially served in an immune and/or detoxification capacity but that one or more mutated to take on a communication function in the reproductive sense of sexual selection either exclusively or in addition to the earlier function? This might explain the extensive expansion of the *Abp* families not only in rodents but also in a few other more distant mammals.

Conclusions

We identified 206 unique *Abp* gene sequences in the genomes of six taxa of the genus *Mus* and mapped their relative positions in those *Abp* clusters. Our CN estimates suggest that the total number of paralogs is closer to 300. We present evidence that the roots of the mouse reference genome expansion in the ancestor of the *Mus Abp* lineage had substantially elevated L1 densities over their *Abp* regions. Further, we suggest that previous analyses of selection on *a27* alleles fixed in different species and subspecies of *Mus* actually detected selection on different paralogs that resulted from tandem duplication (i.e., pseudoalleles) of *a27* in an ancestor of *Mus*. This alternative explanation is based on our finding that *pah* and *car* both have two modules that appear to be duplicated ancestral versions of *M27*. These alternative *a27* paralogs could have arisen as haplotypes having lost different copies through random mutation. If these became recognizable by olfaction, it could have led to sexual selection and eventually incipient reinforcement. We offer evidence for this paralog hypothesis from our module phylogenies built with L1MC3 because L1 RTs are thought to be homoplasy-free regions compared to gene regions. Fixation of different *a27* paralogs in the subspecies by selection is consistent with; 1) different *a27* sequences fixed in each of the three subspecies of *M. musculus*; 2) finding that ABP mediated sexual selection and incipient reinforcement in previous behavioral analyses; and 3) the incongruence of the gene and species phylogenies which was explained previously by homoplasy.

Finally, we propose that the roles of cytokines in immune response to infection and detoxification be considered in further work on the evolutionary histories of the secretoglobins generally and the ABPs specifically. A unique aspect of ABPs in the secretoglobin superfamily evolution is that some diverse mammal species have *Abp* families that expanded extensively. Genes involved in adaptation and functional innovation are subject to frequent duplication, deletion and pseudogene formation and prevalent among rapidly evolving genes are those involved in immunity, reproduction, chemosensation and toxin metabolism. Considering possible multiple functions in the context of the large *Abp* gene family in the mouse, we ask if a small group of them initially served in an immune and/or detoxification capacity but that one or more mutated to take on a communication function in the reproductive sense of sexual

selection. That might explain the extensive expansion of *Abp* families not only in rodents but in a few other more distant mammals.

Materials and Methods

Genomic sequences

Paired-end Illumina HiSeq2000 sequencing data were obtained from the European Nucleotide Archive (ENA) as fastq files for WSB (*Mus musculus domesticus*, ENA Accession number ERS076380); PWK (*M. m. musculus*, ERS076378); CAST (*M. m. castaneus*, ERS076381); and *M. spretus* (ERS076388, ERS138732). The 1504 builds of the genomes of WSB, PWK, CAST and *M. spretus* were obtained from Sanger Mouse Genomes project (ftp://ftp-mouse.sanger.ac.uk/ (Keane, et al. 2011; Lilue, et al. 2018)) as were *M. caroli* and *M. pahari* (Thybert, et al. 2018). Although the later *de novo* assemblies uncoupled from the reference genome had better overall statistics than the 1504 builds, more unique *Abp* sequences were identified on chromosome 7 (chromosome 1 in *pah*) in the 1504 builds (Thybert, et al. 2018). Because they yielded the largest number of *Abp* genes from each genome and the most parsimonious set of gene assignments to chromosomes, we used the gene predictions and coordinates of the 1504 builds.

Data mining genomes for Abp sequences

We employed the BLASTn function in DNA Workbench v3.1.0 (https://www.ncbi.nlm.nih.gov/tools/gbench/), hmmersearch in HMMER/ 3.1 (https://github.com/EddyRivasLab/hmmer), and Exonerate/2.2.0 (https://www.ebi.ac.uk/about/vertebrate-genomics/software/exonerate) to search for sequences similar to *Abp* genes identified previously in rodents. We then searched again with the newly identified sequences but did not obtain additional genes. The sequences we identified by these methods were searched manually for start and stop codons and for donor and acceptor intron splice sites. We also verified the flanking genes *Scn1b* and *Uba2* (formerly *Uble1b*). Once a mouse taxon *Abp* gene sequence was identified bioinformatically, it was verified by designing a set(s) of primers, amplifying it in genomic DNA, and sequencing it (UAGC core facility, University of Arizona) as previously reported (Laukaitis, et al. 2005). The DNA samples used as PCR templates for these six genus *Mus* taxa were obtained from Jackson Laboratory (Bar Harbor, ME). We aligned the sequences we obtained in the laboratory with the corresponding data-mined sequences using DNAsis max/3.0 (MiraiBio, San Francisco, CA). This allowed

us to verify and/or correct the data-mined sequences and complete some that were not full-length. **supplementary tables S1-S6** contain the gene descriptions by taxon with their GenBank number; **supplementary table S7** shows their gene/pseudogene status.

Sequence coverage and calculation of gene copy numbers (CN)

We first attempted to estimate copy number (CN) of *Abp* genes using CNVnator software (Abyzov, et al. 2011), but due to numerous gaps in the *Abp* regions of the 1504 build assemblies (see below), this yielded suspiciously low numbers of small CNVs across the *Abp* gene regions of the six mouse genomes. Hence, we calculated the CNs based on differences in read depth between *Abp* genes and supposedly single-copy regions. With samtools (Li, et al. 2009) we extracted the number of reads mapped to each *Abp* gene and calculated the coverage as (read count/gene length) x (average read length). In the same way we also calculated the coverage for the 1000 randomly chosen regions of 2 kb in length of each *Abp*-containing chromosome where CNVnator did not reveal CNVs. Regions with less than 10% of reads of low mapping quality (defined as MAPQ < 20) were chosen for calculating the average coverage for single-copy sequences (**supplementary table S8)**. We derived a diploid copy number for each *Abp* gene by dividing the coverage of the *Abp* gene with the average coverage for single-copy sequences. Very few reads in any of the *Abp* genes were of low quality (MAPQ<20). In the case of two sequences of *car_a28a* and *b* (**supplementary table S2**), multiple mapping locations could have inflated their apparent coverage. This made only a minor contribution (<1% of the 206 genes we found) to CN determination. We have shown that the average GC content in the *Abp* gene region is in line with the genome average (Karn and Laukaitis 2009). Hence, we assumed that GC bias is not an issue for read-depth based inference of copy number in the *Abp* region.

Analysis of gaps in the assemblies

To examine the possibility that some of the gaps represent truly missing sequences rather than undefined difficult to assemble regions, we looked for read evidence to dismiss gap placement in the assembly. For that reason, we deleted all gaps (N's) in the reference chromosome where the *Abp* cluster resides in order to create a new reference chromosome without gaps. We next performed mapping and

subsequent processing steps with the gap-free reference as described above. We searched for reads that span gap positions and are either ≥ 80 nucleotides (nt) long or have mapping quality > 0. The reads were also required to exactly match the gap-free reference over the entire read length and to have at least 8 nt of sequence around the predicted gap position (i.e. that the gap position is not at the very end of the read).

Retrotransposon content

The data table "rmsk" was obtained from the UCSC FTP server for C57Bl/6 and the Sanger Mouse Genomes project for six mouse taxa (see above). Data for LINEs was extracted (Wicker, et al. 2007; Kapitonov and Jurka 2008). Sliding windows of 100 kb with 10 kb steps were created across each genome assembly using the "bedtools makewindows" command of bedtools. The number of bases within each window that is covered by LINEs was calculated using bedtools coverage (https://github.com/arq5x/bedtools). When gaps were present within the assembly, two coordinate systems were created: one before the gap removal and one after the gap removal, and positions of LINEs and *Abp* genes were converted between these (https:// github.com/ucscGenomeBrowser/kent). The density plots with gaps removed were made using the ggplot2 package in R.

The rmsk RT data table was obtained from the Sanger Mouse Genomes project for six mouse taxa (ftp://ftp-mouse.sanger.ac.uk/ Keane, et al. 2011; Lilue, et al. 2018). Positions of L1MC3 Elements were extracted from the rmsk table and were filtered using the *Abp* intra-module coordinates. L1MC3 sequences were then obtained using "bedtools getfasta" command of bedtools (https://github.com/arq5x/bedtools). Those intra-module sequences negative for L1MC3 when searched in that manner were searched again by aligning L1MC3 sequences from other modules and, in some cases this revealed the retrotransposon in the intra-module sequences.

Data analysis

We assigned exons and introns to the verified and/or corrected DNA sequences of the six taxa of *Mus musculus* by aligning them with the known exon and intron sequences of four *Abpa* and four *Abpbg* genes from the mouse genomes (*a2, a27, a24, a27, bg2, bg7, bg24 and bg27*). The donor and acceptor splice sites were identified and the exons were assembled into putative mRNAs and translated *in silico*. From the translations, we identified each gene as either a potentially expressed gene or as a pseudogene if it had either a disruption in the coding region and/or a noncanonical splice site (Emes, et al. 2004). **Tables S1-S6** show the disruptions for the putative pseudogenes. MAFFT was used to align the *Abp* gene sequences from the genus *Mus* and the mouse and rat reference genomes, IQtree (http://www.iqtree.org; Trifinopoulos, et al. 2016) was used to build maximum likelihood phylogenetic trees that were visualized with FigTree v1.4.3 (http://tree.bio.ed.ac.uk/software/figtree). Initially, we built trees with the larger intron b, that lies between Exons 2 and 3, in order to avoid bias caused by selection (Laukaitis, et al. 2008). Comparisons with trees constructed with the full genes (ATG to the stop codon) showed essentially the same topologies and allowed us to include partial sequences lacking most or all of intron b. Bootstrap values (1000 repetitions) were obtained with the MAFFT ultrafast bootstrap approximation. L1MC3 RTs from the intramodular regions were aligned and used for producing MAFFT and IQTree.

Acknowledgements

This work was supported by the National Cancer Institute at the National Institutes of Health [grant number P30 CA23074] for laboratory infrastructure. The authors gratefully acknowledge Willie Swanson for helpful discussions and Miloš Macholán for additional recommendations.

Data Availability Statement

All sequence data are released into GenBank and their Accession numbers are listed in **supplementary tables S1-6, supplementary material** online.

Author Contributions

RCK, GY and CML conceived of the project, mined the *Abp* and *L1MC3* sequence data, designed primers and sequenced the genes and built phylogenies. ZP did the *Abp* module alignments, the copy number analyses and the gap analyses. PB and RCK assessed the evolutionary forces acting on *Abp* orthologs vs. paralogs. All the authors participated in writing the manuscript.

References

Abyzov A, Urban AE, Snyder M, Gerstein M. 2011. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res 21:974-984.

Alexeev N, Alekseyev MA. 2018. Combinatorial Scoring of Phylogenetic Trees and Networks Based on Homoplasy-Free Characters. J Comput Biol 25:1203-1219.

Alkan C, Coe BP, Eichler EE. 2011. Genome structural variation discovery and genotyping. Nat Rev Genet 12:363-376.

Almuntashiri S, Zhu Y, Han Y, Wang X, Somanath PR, Zhang D. 2020. Club Cell Secreted Protein CC16: Potential Applications in Prognosis and Therapy for Pulmonary Diseases. J Clin Med 9.

Altenhoff AM, Glower NM, Dessimoz C. 2019. Inferring Orthology and Paralogy. In: Anisimova M, editor. Evolutionary Genomics: Statistical and Computational Methods. New York, NY: Humana Press. p. 149-176.

Beier HM. 1968. Uteroglobin: a hormone-sensitive endometrial protein involved in blastocyst development. Biochim Biophys Acta 160:289-291.

Bímová B, Karn RC, Pialek J. 2005. The role of salivary androgen-binding protein in reproductive isolation between two subspecies of house mouse: Mus musculus musculus and Mus musculus domesticus. Biological Journal of the Linnean Society 84:349-361.

Boursot P, Auffray J-C, Britton-Davidian J, Bonhomme F. 1993. The evolution of house mice. Annu Rev Ecol Syst 24:119-152.

Cai Y, Kimura S. 2015. Secretoglobin 3A2 Exhibits Anti-Fibrotic Activity in Bleomycin-Induced Pulmonary Fibrosis Model Mice. PLoS One 10:e0142497.

Callebaut I, Poupon A, Bally R, Demaret JP, Housset D, Delettre J, Hossenlopp P, Mornon JP. 2000. The uteroglobin fold. Ann N Y Acad Sci 923:90-112.

Casola C, Conant GC, Hahn MW. 2012. Very low rate of gene conversion in the yeast genome. Mol Biol Evol 29:3817-3826.

Casola C, Hahn MW. 2009. Gene conversion among paralogs results in moderate false detection of positive selection using likelihood methods. J Mol Evol 68:679-687.

Chevret P, Veyrunes F, Britton-Davidian J. 2005. Molecular phylogeny of the genus *Mus* (Rodentia:Murinae) based on mitochondrial and nuclear data. Biol J Linn Soc 84:417-427.

Chiba Y, Kurotani R, Kusakabe T, Miura T, Link BW, Misawa M, Kimura S. 2006. Uteroglobin-related protein 1 expression suppresses allergic airway inflammation in mice. Am J Respir Crit Care Med 173:958-964.

Chung AG, Belone PM, Bimova BV, Karn RC, Laukaitis CM. 2017. Studies of an Androgen-Binding Protein Knockout Corroborate a Role for Salivary ABP in Mouse Communication. Genetics 205:1517-1527.

Clifton BD, Jimenez J, Kimura A, Chahine Z, Librado P, Sanchez-Gracia A, Abbassi M, Carranza F, Chan C, Marchetti M, et al. 2020. Understanding the Early Evolutionary Stages of a Tandem

Drosophilamelanogaster-Specific Gene Family: A Structural and Functional Population Study. Mol Biol Evol 37:2584-2600.

Denecke S, Rankic I, Driva O, Kalsi M, Luong NBH, Buer B, Nauen R, Geibel S, Vontas J. 2021. Comparative and functional genomics of the ABC transporter superfamily across arthropods. BMC Genomics 22:553.

Dlouhy SR, Karn RC. 1983. The tissue source and cellular control of the apparent size of androgen binding protein (Abp), a mouse salivary protein whose electrophoretic mobility is under the control of sex-limited saliva pattern (Ssp). Biochem Genet 21:1057-1070.

Dover G. 1982. Molecular drive: a cohesive mode of species evolution. Nature 299:111-117.

Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH. 2010. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11:446-450.

Emerson JJ, Cardoso-Moreira M, Borevitz JO, Long M. 2008. Natural selection shapes genome-wide patterns of copy-number polymorphism in Drosophila melanogaster. Science 320:1629-1631.

Emes RD, Riley MC, Laukaitis CM, Goodstadt L, Karn RC, Ponting CP. 2004. Comparative evolutionary genomics of androgen-binding protein genes. Genome Res 14:1516-1529.

Galtier N, Duret L, Glemin S, Ranwez V. 2009. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. Trends Genet 25:1-5.

Galtier N, Piganeau G, Mouchiroud D, Duret L. 2001. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics 159:907-911.

Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE, et al. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428:493-521.

Goodman M, Czelusniak J, William Moore G, Romero-Herrera AE, Matsuda G. 1979. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. Systematic Zoology 28:132-163.

Guy L, Kultima JR, Andersson SG. 2010. genoPlotR: comparative gene and genome visualization in R. Bioinformatics 26:2334-2335.

Harel T, Lupski JR. 2018. Genomic disorders 20 years on-mechanisms for clinical manifestations. Clin Genet 93:439-449.

Hastings PJ, Lupski JR, Rosenberg SM, Ira G. 2009. Mechanisms of change in gene copy number. Nat Rev Genet 10:551-564.

Higuchi DA, Cahan P, Gao J, Ferris ST, Poursine-Laurent J, Graubert TA, Yokoyama WM. 2010. Structural variation of the mouse natural killer gene complex. Genes Immun 11:637-648.

Huddleston J, Eichler EE. 2016. An Incomplete Understanding of Human Genetic Variation. Genetics 202:1251-1254.

Hudson RR, Kreitman M, Aguade M. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics 116:153-159.

Hurles M. 2002. Are 100,000 "SNPs" useless? Science 298:1509; author reply 1509.

Hwang JM, Hofstetter JR, Bonhomme F, Karn RC. 1997. The microevolution of mouse salivary androgen-binding protein (ABP) paralleled subspeciation of Mus musculus. J Hered 88:93-97.

Janousek V, Karn RC, Laukaitis CM. 2013. The role of retrotransposons in gene family expansions: insights from the mouse Abp gene family. BMC Evol Biol 13:107.

Janousek V, Laukaitis CM, Yanchukov A, Karn RC. 2016. The Role of Retrotransposons in Gene Family Expansions in the Human and Mouse Genomes. Genome Biol Evol 8:2632-2650.

Johnson ME, Viggiano L, Bailey JA, Abdul-Rauf M, Goodwin G, Rocchi M, Eichler EE. 2001. Positive selection of a gene family during the emergence of humans and African apes. Nature 413:514-519.

Kapitonov VV, Jurka J. 2008. A universal classification of eukaryotic transposable elements implemented in Repbase. Nat Rev Genet 9:411-412; author reply 414.

Karn RC. 1994. The mouse salivary androgen-binding protein (ABP) alpha subunit closely resembles chain 1 of the cat allergen Fel dI. Biochem Genet 32:271-277.

Karn RC. 1998. Steroid binding by mouse salivary proteins. Biochem Genet 36:105-117.

Karn RC, Chung AG, Laukaitis CM. 2014. Did androgen-binding protein paralogs undergo neo- and/or Subfunctionalization as the Abp gene region expanded in the mouse genome? PLoS One 9:e115454.

Karn RC, Clements MA. 1999. A comparison of the structures of the alpha:beta and alpha:gamma dimers of mouse salivary androgen-binding protein (ABP) and their differential steroid binding. Biochem Genet 37:187-199.

Karn RC, Dlouhy SR. 1991. Salivary androgen-binding protein variation in Mus and other rodents. J Hered 82:453-458.

Karn RC, Laukaitis CM. 2009. The mechanism of expansion and the volatility it created in three pheromone gene clusters in the mouse (Mus musculus) genome. Genome Biol Evol 1:494-503.

Karn RC, Laukaitis CM. 2014. Selection shaped the evolution of mouse androgen-binding protein (ABP) function and promoted the duplication of Abp genes. Biochemical Society Transactions 42:in press.

Karn RC, Nachman MW. 1999. Reduced nucleotide variability at an androgen-binding protein locus (Abpa) in house mice: evidence for positive natural selection. Mol Biol Evol 16:1192-1197.

Karn RC, Orth A, Bonhomme F, Boursot P. 2002. The complex history of a gene proposed to participate in a sexual isolation mechanism in house mice. Mol Biol Evol 19:462-471.

Karn RC, Young JM, Laukaitis CM. 2010. A candidate subspecies discrimination system involving a vomeronasal receptor gene with different alleles fixed in M. m. domesticus and M. m. musculus. PLoS One 5.

Keane TM, Goodstadt L, Danecek P, White MA, Wong K, Yalcin B, Heger A, Agam A, Slater G, Goodson M, et al. 2011. Mouse genomic variation and its effect on phenotypes and gene regulation. Nature 477:289-294.

Krishnan RS, Daniel JC, Jr. 1967. "Blastokinin": inducer and regulator of blastocyst development in the rabbit uterus. Science 158:490-492.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860-921.

Laukaitis C, Karn RC. 2012. Recognition of subspecies status mediated by androgen-binding protein (ABP) in the evolution of incipient reinforcement on the European house mouse hybrid zone. In: Macholan M, Munclinger P, Baird SJ, Pialek J, editors. Evolution of the House Mouse. Cambridge, UK: Cambridge University Press.

Laukaitis CM, Critser ES, Karn RC. 1997. Salivary Androgen-Binding Protein (Abp) Mediates Sexual Isolation in Mus Musculus. Evolution 51:2000-2005.

Laukaitis CM, Dlouhy SR, Emes RD, Ponting CP, Karn RC. 2005. Diverse spatial, temporal, and sexual expression of recently duplicated androgen-binding protein genes in Mus musculus. BMC Evol Biol 5:40.

Laukaitis CM, Dlouhy SR, Karn RC. 2003. The mouse salivary androgen-binding protein (ABP) gene cluster on chromosomes 7: characterization and evolutionary relationships. Mamm Genome 14:679-691.

Laukaitis CM, Heger A, Blakley TD, Munclinger P, Ponting CP, Karn RC. 2008. Rapid bursts of androgen-binding protein (Abp) gene duplication occurred independently in diverse mammals. BMC Evol Biol 8:46.

Laukaitis CM, Mauss C, Karn RC. 2012. Congenic strain analysis reveals genes that are rapidly evolving components of a prezygotic isolation mechanism mediating incipient reinforcement. PLoS One 7:e35898.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078-2079.

Lilue J, Doran AG, Fiddes IT, Abrudan M, Armstrong J, Bennett R, Chow W, Collins J, Collins S, Czechanski A, et al. 2018. Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci. Nat Genet 50:1574-1583.

Loire E, Tusso S, Caminade P, Severac D, Boursot P, Ganem G, Smadja CM. 2017. Do changes in gene expression contribute to sexual isolation and reinforcement in the house mouse? Mol Ecol 26:5189-5202.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. 2009. Finding the missing heritability of complex diseases. Nature 461:747-753.

Morgan AP, Holt JM, McMullan RC, Bell TA, Clayshulte AM, Didion JP, Yadgary L, Thybert D, Odom DT, Flicek P, et al. 2016. The Evolutionary Fates of a Large Segmental Duplication in Mouse. Genetics 204:267-285.

Mukherjee A, Chilton BE. 2000. The uteroglobin/clara cell protein family. Annals of the New York Academy of Sciences 932:1-356.

Mukherjee AB, Zhang Z, Chilton BS. 2007. Uteroglobin: a steroid-inducible immunomodulatory protein that founded the Secretoglobin superfamily. Endocr Rev 28:707-725.

Nagylaki T, Petes TD. 1982. Intrachromosomal gene conversion and the maintenance of sequence homogeneity among repeated genes. Genetics 100:315-337.

Nguyen DQ, Webber C, Hehir-Kwa J, Pfundt R, Veltman J, Ponting CP. 2008. Reduced purifying selection prevails over positive selection in human copy number variant evolution. Genome Res 18:1711-1723.

Nguyen DQ, Webber C, Ponting CP. 2006. Bias of selection on human copy-number variants. PLoS Genet 2:e20.

Pan D, Zhang L. 2008. Tandemly arrayed genes in vertebrate genomes. Comp Funct Genomics:545269.

Pavlopoulou A, Pampalakis G, Michalopoulos I, Sotiropoulou G. 2010. Evolutionary history of tissue kallikreins. PLoS One 5:e13781.

Perry GH, Ben-Dor A, Tsalenko A, Sampas N, Rodriguez-Revenga L, Tran CW, Scheffer A, Steinfeld I, Tsang P, Yamada NA, et al. 2008. The fine-scale and complex architecture of human copy-number variation. Am J Hum Genet 82:685-695.

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, et al. 2007. Diet and the evolution of human amylase gene copy number variation. Nat Genet 39:1256-1260.

Pezer Z, Chung AG, Karn RC, Laukaitis CM. 2017. Analysis of Copy Number Variation in the Abp Gene Regions of Two House Mouse Subspecies Suggests Divergence during the Gene Family Expansions. Genome Biol Evol 9.

Sage RD, Atchley WR, Capanna E. 1993. House mice as models in systematic biology. Syst. Biol. 42:523-561.

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74:5463-5467.

Semple C, Steel M. 2002. Tree Reconstruction from Multi-State Characters. Advances in Applied Mathematics 28:169-184.

Sjodin P, Jakobsson M. 2012. Population genetic nature of copy number variation. Methods Mol Biol 838:209-223.

Spielmann M, Lupianez DG, Mundlos S. 2018. Structural variation in the 3D genome. Nat Rev Genet 19:453-467.

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH, et al. 2015. An integrated map of structural variation in 2,504 human genomes. Nature 526:75-81.

Talley HM, Laukaitis CM, Karn RC. 2001. Female preference for male saliva: implications for sexual isolation of Mus musculus subspecies. Evolution 55:631-634.

Thybert D, Roller M, Navarro FCP, Fiddes I, Streeter I, Feig C, Martin-Galvez D, Kolmogorov M, Janousek V, Akanni W, et al. 2018. Repeat associated mechanisms of genome evolution and function revealed by the Mus caroli and Mus pahari genomes. Genome Res 28:448-459.

Trifinopoulos J, Nguyen LT, von Haeseler A, Minh BQ. 2016. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. Nucleic Acids Res 44:W232-235.

Uriu K, Kosugi Y, Ito J, Sato K. 2021. The Battle between Retroviruses and APOBEC3 Genes: Its Past and Present. Viruses 13.

van Rheede T, Bastiaans T, Boone DN, Hedges SB, de Jong WW, Madsen O. 2006. The platypus is in its place: nuclear genes and indels confirm the sister group relation of monotremes and Therians. Mol Biol Evol 23:587-597.

Verneau O, Catzeflis F, Furano AV. 1998. Determining and dating recent rodent speciation events by using L1 (LINE-1) retrotransposons. Proc Natl Acad Sci U S A 95:11284-11289.

Vošlajerová Bímová B, Macholán M, Baird SEB, Munclinger P, Laukaitis CM, Karn RC, Luzynski K, Tucker P, Piálek J. 2011. Reinforcement selection acting on the European house mouse hybrid zone. Molecular Ecology 20:2403-2424.

Wang Y, Li X, Lin Y, Wang Y, Wang K, Sun C, Lu T, Zhang M. 2018. Structural Variation, Functional Differentiation, and Activity Correlation of the Cytochrome P450 Gene Superfamily Revealed in Ginseng. Plant Genome 11.

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420:520-562.

White MA, Ane C, Dewey CN, Larget BR, Payseur BA. 2009. Fine-scale phylogenetic discordance across the house mouse genome. PLoS Genet 5:e1000729.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, et al. 2007. A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8:973-982.

Xue Y, Sun D, Daly A, Yang F, Zhou X, Zhao M, Huang N, Zerjal T, Lee C, Carter NP, et al. 2008. Adaptive evolution of UGT2B17 copy-number variation. Am J Hum Genet 83:337-346.

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

Yoneda M, Xu L, Kajiyama H, Kawabe S, Paiz J, Ward JM, Kimura S. 2016. Secretoglobin Superfamily Protein SCGB3A2 Alleviates House Dust Mite-Induced Allergic Airway Inflammation in Mice. Int Arch Allergy Immunol 171:36-44.

Young ND, Zhou P, Silverstein KA. 2016. Exploring structural variants in environmentally sensitive gene families. Curr Opin Plant Biol 30:19-24.

Zhou P, Silverstein KA, Ramaraj T, Guhlin J, Denny R, Liu J, Farmer AD, Steele KP, Stupar RM, Miller JR, et al. 2017. Exploring structural variation and gene family architecture with De Novo assemblies of 15 Medicago genomes. BMC Genomics 18:261.

Table 1. *Abpa* and *Abpbg* genes in each wild-derived mouse genome (B6 refers to mouse reference genome (mm10)).

Taxon	Number of genes (unique)	Number of *Abpa* genes				Number of *Abpbg* genes				Represented Clades
		Total (unique)	Number of genes (unique)	Number of pseudogenes (unique)	Lineage-specific	Total (unique)	Number of genes (unique)	Number of pseudogenes (unique)	Lineage-specific	
B6	64 (58)	30 (27)	14 (13)	16 (14)	7	34 (31)	12 (11)	22 (20)	7	1-5
WSB	79 (43)	39 (21)	21 (8)	18 (13)	NA	40 (22)	18 (10)	22 (12)	NA	1-5
PWK	41 (38)	20 (18)	13 (11)	7 (7)	1	21 (20)	10 (10)	11 (10)	1	1-5
CAS	72 (46)	36 (22)	24 (11)	12 (11)	2	36 (24)	26 (18)	10 (6)	3	1-5
spr	65 (35)	30 (17)	12 (6)	18 (11)	3	35 (18)	23 (11)	12 (7)	4	1-5
car	40 (33)	21 (17)	9 (6)	12 (11)	9	19 (16)	13 (10)	6 (6)	8	1-3,5
pah	11 (11)	6 (6)	3 (3)	3 (3)	4	5 (5)	4 (4)	1 (1)	3	3, 5

Figure Legends

Figure 1. A canonical phylogeny of murid rodents adapted from Chevret, et al. (2005) to show the divergence of *M. m. domesticus* (strain WSB) from the ancestor of *M. m. musculus* (strain PWK) and *M.m. castaneus* (strain CAS) (White, et al. 2009; Keane, et al. 2011). The seven taxa are differentiated by color. The black numbers under each taxon are the different gene sequences we found and the grey numbers indicate total additional copies (CN) above the diploid number (the CN for each gene is given in **supplementary tables S1-S6**). CN was not determined for the rat genome *Abp* region.

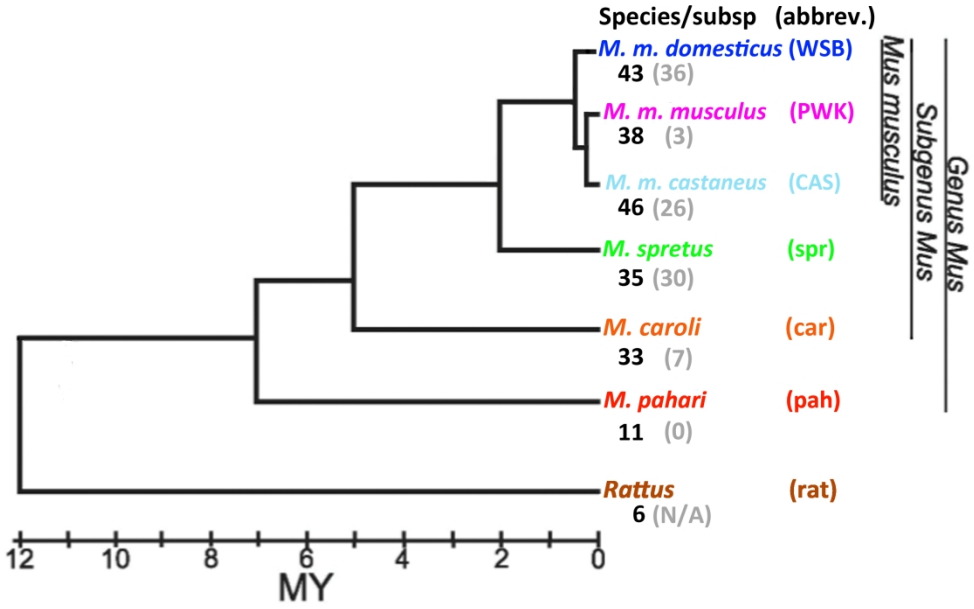
Figure 2. Gene phylogenies of murid rodent *Abpa* (Panel A) and *Abpbg* (Panel B) genes rooted to the independent Rat clade (brown) and basal *Mus* root (gray). Paralogs from five ancestral clades (Laukaitis, et al. 2008) are indicated by color-coding of branches, red (1), green (2), purple (3), yellow (4), and blue (5), represented by colored bars around the periphery of the phylogeny. The taxon-specific colors of Fig. 1 are used for the gene names (not italicized) and genes that root more deeply than individual B6 clades are named with capital letters (e.g. *pah_aW*). Bootstrap values are shown in black. See **supplementary figures S1-S2** for parts of these trees broken out to make them easier to read.

Figure 3. Relationships between *Abp* paralogs identified in six rodent taxa. Genes that could be expressed are solid-filled blue (*Abpa*) or red (*Abpbg*) arrows while putative pseudogenes are unfilled arrows. Taxon abbreviations with their chromosome in parentheses are shown in a phylogeny on the left. A block of genes unique to the C57BL/6 (B6) reference genome is shown above a blue triangle at the top of the figure and represents genes of the ultimate and penultimate duplications in the WSB lineage (Pezer, et al. 2017). Orthologs between taxa are connected by bands with the colors of the Clades shown in **Fig. 2** and the large Clade numbers at the top of the figure are colored the same. Starting from *pah*, genes with no ortholog in the next taxon are highlighted in blue. Genes are shown in a proportional scale, however regions between genes are reduced tenfold to enable clearer representation of gene relationships. Synteny was plotted with genoPlotR package (Guy, et al. 2010).

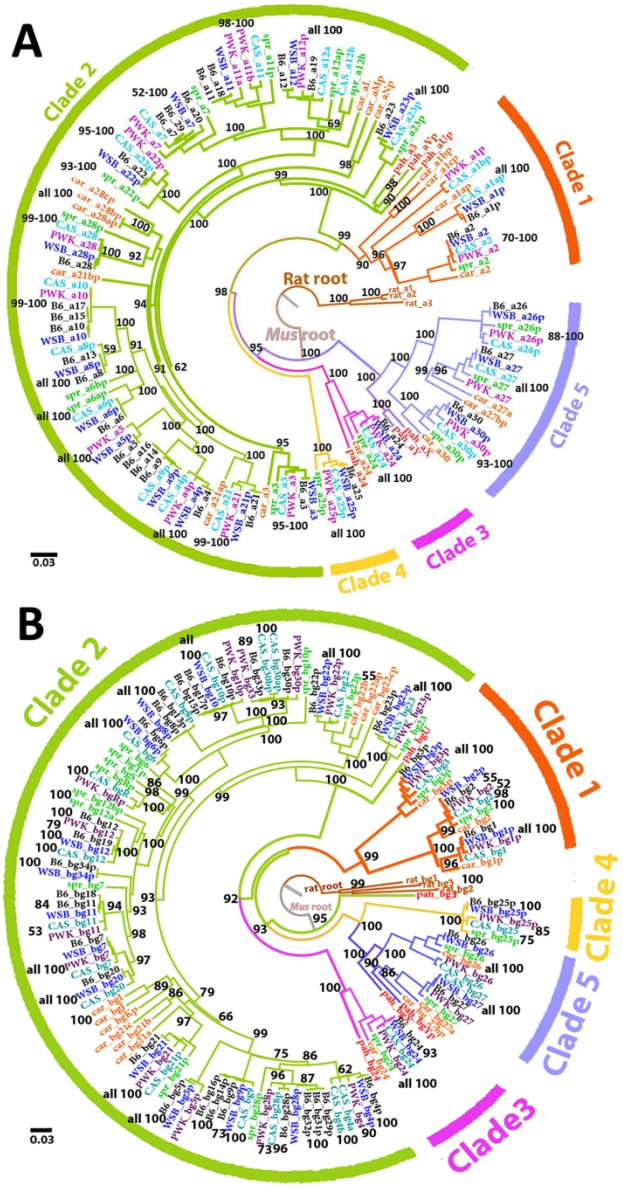
Figure 4. Module phylogeny constructed with a LINE1, L1MC3, that is nearly ubiquitous in the intra-modular sequences of gene modules in the genome mouse (mm10) and in the six *Mus* taxa. Five ancestral

clades are labeled in red (1), green (2), purple (3, represented only by pah-M24), yellow (4), and blue (5) and the gene-specific colors are the same as in **Fig. 2**. Bootstrap values >50 are shown in black.

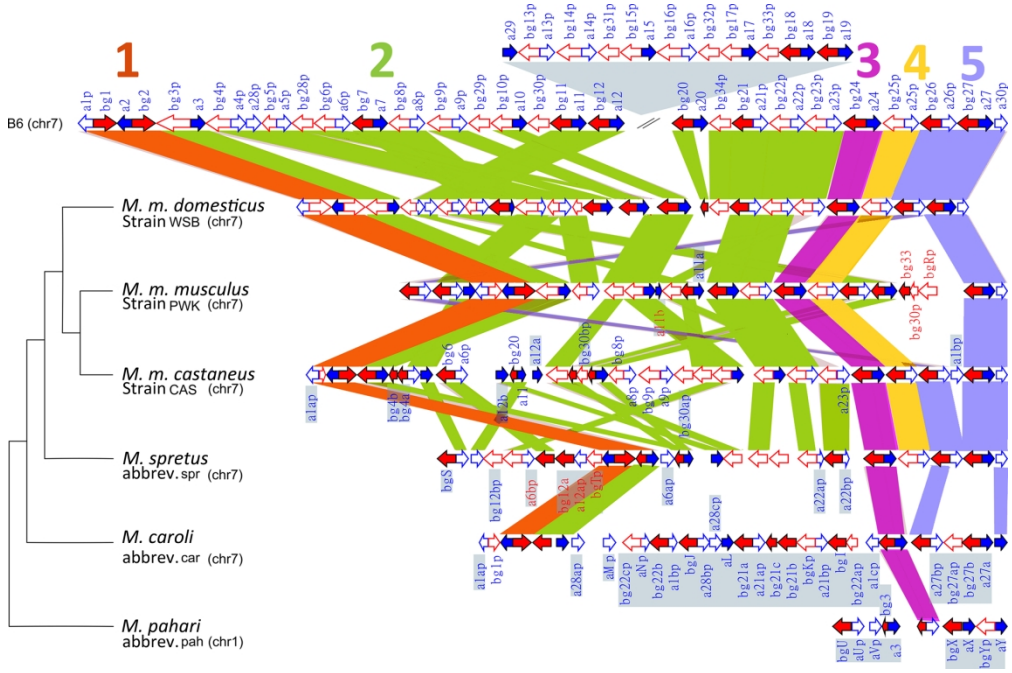
Figure 5. Clades 4 and 5 gene and module phylogenies. Genes and modules with unusual topologies are shown with red asterisks. *Abpa27* (Panel A, center) has the unexpected topology reported by Karn, et al. (2002) where the PWK allele is an outgroup to the *spr* allele. The *a26* genes (Panel A, top) also have an unexpected topology as do the *M27, M26* (Panel C) and *M25* (Panel D) modules, and *bg26* (Panel E) and *bg25* (Panel F) genes. Only *a25* (Panel B) shows an expected topology.



A canonical phylogeny of murid rodents adapted from Chevret, et al. (2005) to show the divergence of M. m. domesticus (strain WSB) from the ancestor of M. m. musculus (strain PWK) and M.m. castaneus (strain CAS) (White, et al. 2009; Keane, et al. 2011). The seven taxa are differentiated by color. The black numbers under each taxon are the different gene sequences we found and the grey numbers indicate total additional copies (CN) above the diploid number (the CN for each gene is given in supplementary tables S1-S6). CN was not determined for the rat genome Abp region.

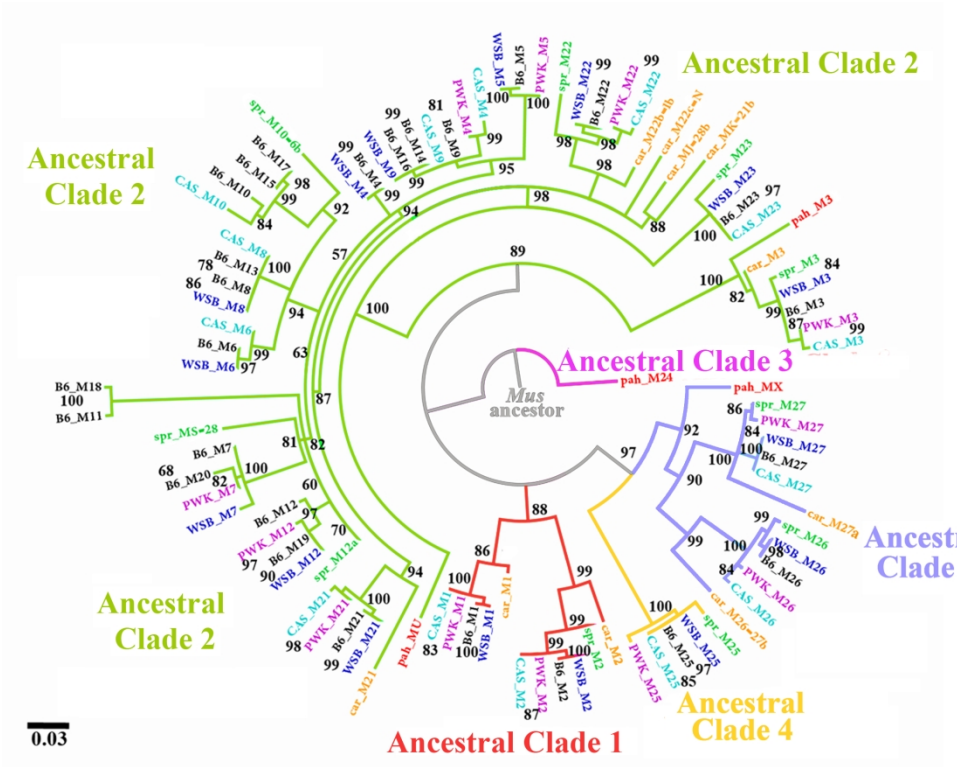


Gene phylogenies of murid rodent Abpa (Panel A) and Abpbg (Panel B) genes rooted to the independent Rat clade (brown) and basal Mus root (gray). Paralogs from five ancestral clades (Laukaitis, et al. 2008) are indicated by color-coding of branches, red (1), green (2), purple (3), yellow (4), and blue (5), represented by colored bars around the periphery of the phylogeny. The taxon-specific colors of Fig. 1 are used for the gene names (not italicized) and genes that root more deeply than individual B6 clades are named with capital letters (e.g. pah_aW). Bootstrap values are shown in black. See supplementary figures S1-S2 for parts of these trees broken out to make them easier to read.

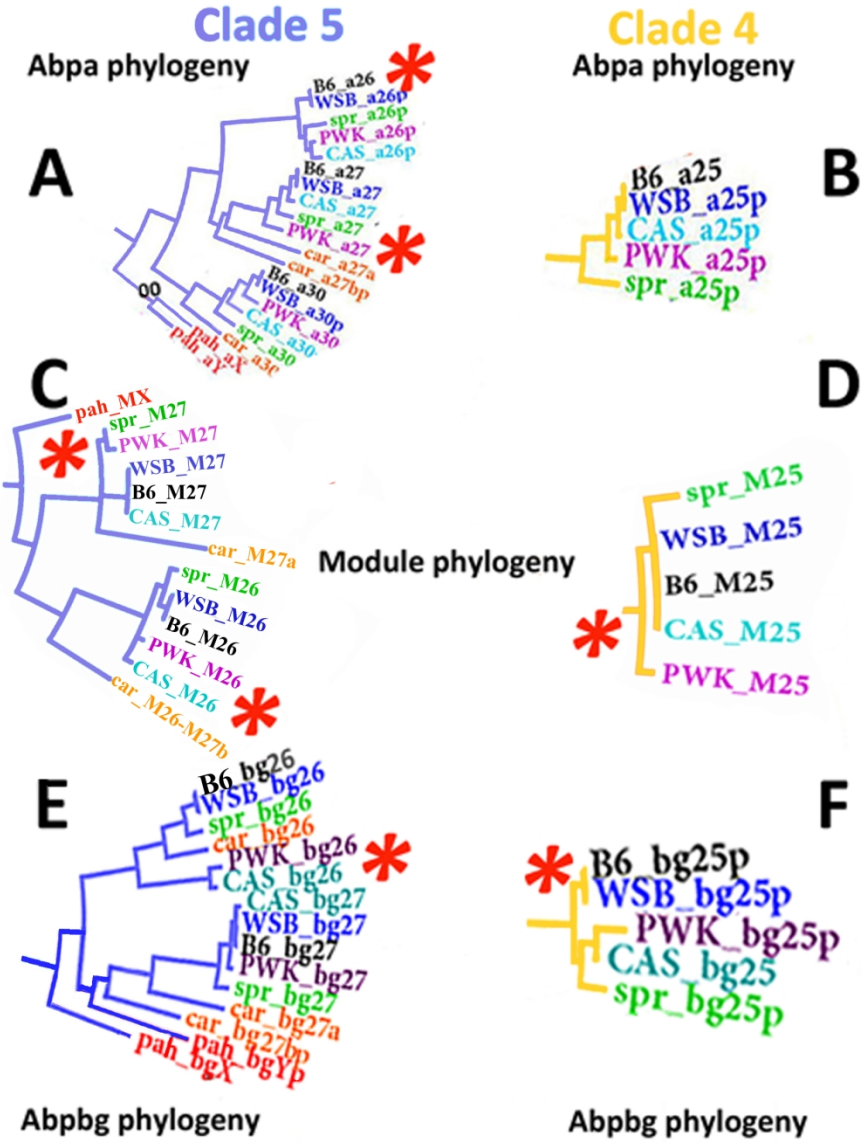


Relationships between Abp paralogs identified in six rodent taxa. Genes that could be expressed are solid-filled blue (Abpa) or red (Abpbg) arrows while putative pseudogenes are unfilled arrows. Taxon abbreviations with their chromosome in parentheses are shown in a phylogeny on the left. A block of genes unique to the C57BL/6 (B6) reference genome is shown above a blue triangle at the top of the figure and represents genes of the ultimate and penultimate duplications in the WSB lineage (Pezer, et al. 2017). Orthologs between taxa are connected by bands with the colors of the Clades shown in Fig. 2 and the large Clade numbers at the top of the figure are colored the same. Starting from pah, genes with no ortholog in the next taxon are highlighted in blue. Genes are shown in a proportional scale, however regions between genes are reduced tenfold to enable clearer representation of gene relationships. Synteny was plotted with genoPlotR package (Guy, et al. 2010).

229x152mm (300 x 300 DPI)



Module phylogeny constructed with a LINE1, L1MC3, that is nearly ubiquitous in the intra-modular sequences of gene modules in the genome mouse (mm10) and in the six Mus taxa. Five ancestral clades are labeled in red (1), green (2), purple (3, represented only by pah-M24), yellow (4), and blue (5) and the gene-specific colors are the same as in Fig. 2. Bootstrap values >50 are shown in black.



Clades 4 and 5 gene and module phylogenies. Genes and modules with unusual topologies are shown with red asterisks. Abpa27 (Panel A, center) has the unexpected topology reported by Karn, et al. (2002) where the PWK allele is an outgroup to the spr allele. The a26 genes (Panel A, top) also have an unexpected topology as do the M27, M26 (Panel C) and M25 (Panel D) modules, and bg26 (Panel E) and bg25 (Panel F) genes. Only a25 (Panel B) shows an expected topology.