International Journal of Biochemistry Research & Review
3(4): 328-363, 2013

SCIENCEDOMAIN *international*
www.sciencedomain.org

Evolution of Rodent Pheromones: A Review of the ABPs with Comparison to the ESPs and the MUPs

Robert C. Karn[1*]

[1]*Department of Medicine, College of Medicine, University of Arizona, Tucson, AZ, USA.*

Author's contribution

Author RCK wrote the entire manuscript and approved the final manuscript.

Review Article

Received 2nd July 2013
Accepted 28th July 2013
Published 16th August 2013

ABSTRACT

Three proteinaceous pheromone families, the androgen-binding proteins (ABPs), the exocrine-gland secreting peptides (ESPs) and the major urinary proteins (MUPs) are encoded by large gene families in the *Mus musculus* and *Rattus norvegicus* genomes. The purpose of this article is to review what is known about the evolutionary histories of the the *Abp* gene family expansions in rodents and, where appropriate, to compare them to what is known of the expansions of the *Mup* and *Esp* gene families. The issues important to these histories are the extent of the gene family expansions, the timing of their expansions and the roles played by selection, gene conversion and non-allelic homologous recombination (NAHR). I also compare and contrast the evolutionary histories of all three mouse gene families in light of the proposed functions of their pheromones in mouse communication.

Keywords: Rodent; androgen-binding protein; exocrine gland secreting peptides; major urinary proteins; evolutionary history; gene duplication; gene family expansion.

**Corresponding author: Email: rkarn@butler.edu;*

1. INTRODUCTION

Over the past five years, my colleagues and I have combined a search for the function of androgen-binding protein (ABP) with studies of the mechanism behind the rapid and extensive expansion of the *Abp* region in the mouse genome and the evolution of individual ABP subunits and the genes that encode them. Our hope was that each of these individual approaches would inform the others and that developments made by other research groups studying evolution by gene duplication would contribute new insights to our study.

Early genome studies had already provided some clues to functional relationships. The availability of an increasing number of mammalian genome sequences has enhanced our ability to investigate evolutionary processes and thereby our understanding of gene evolution. Those genes not preserved as single copies in both primate and rodent lineages are subject to frequent duplication, deletion and pseudogene formation [1-3], whereas conserved genes are likely to possess functions that are shared by primates, rodents, and, in all likelihood, by most mammals. By contrast, frequently duplicated genes are more often associated with adaptation and functional innovation [1,4,5]. They often show the footprints of positive selection in elevated ratios of nonsynonymous to synonymous nucleotide substitutions (*dN/dS*; sometimes reported as the rate *Ka/Ks*; [6]) in their coding regions [7-11]. Gene deletion and pseudogene formation events are rare, except among genes that have also been subject to duplication [2,3,12]. When duplication and deletion events are present, the affected gene region may show copy number variation (CNV) and more volatility than other gene regions of similar size [13]. Prevalent among rapidly evolving genes are those involved in immunity, reproduction, chemosensation and toxin metabolism [1]. I will focus this review on the reproduction category of genes especially the evolution of the gene family encoding the rodent ABPs, but I will also make comparisons with the evolution of two other rodent gene families encoding pheromones, the exocrine gland secreting peptides (ESPs) and the major urinary proteins (MUPs).

For the purpose of this review, it is important to emphasize that a high rate of amino acid substitution is a possible signature of adaptive evolution. Proteins involved in reproductive fitness have evolved unusually rapidly across diverse groups of organisms [1,14-16]. In the case of reproductive proteins, coevolutionary cycles involving adaptation and counter adaptation are expected to apply continuous selective pressure, resulting in rapid changes at amino acid sites involved in the function of the protein. These proteins often have roles in sperm competition, host immunity to pathogens, and manipulation of female reproductive physiology and behavior; however, in many other cases, the function of the rapidly evolving protein is unknown.

Recently a great deal of interest has been focused on reproductive proteins encoded by genes, sometimes called speciation genes, that are associated with signatures of positive selection [17-20] and that have functions thought to promote reproductive isolation among closely related species [21,22]. Much emphasis has been given to reproductive genes involved in postzygotic isolation but relatively little to those involved in prezygotic isolation, e.g. proteins with functions such as mediating mate choice [23,24]. And yet there are examples of gene duplication acting as a major source of new gene functions involved in mate selection at the individual and population levels. Among these are the three rodent pheromone protein families introduced above, the ABPs, the ESPs and the MUPs. Each is encoded by genes that have undergone extensive duplication in mice, rats and perhaps other mammals (see for example [25]). Proteins encoded by all three gene families affect

mate selection and thus directly impact gene exchange and thereby evolution and potentially speciation.

2. DISCOVERY AND EARLY STUDIES OF ABP AND ITS GENES

The discovery of the protein now known as androgen-binding protein (ABP) and the development of our understanding of its genetics, biochemistry and function is unusual because it has involved essentially only my laboratory at various institutions, including Indiana University and Butler University during the past three decades, as well as various collaborators at other institutions (reviewed in [26]). Since 2008, our work has continued in the laboratory that I share with Christina Laukaitis at the University of Arizona. Although I began studies of the protein in 1977, its actual description and naming had to await the discovery of a characteristic that distinguished it from other proteins in mouse saliva. Experiments that explored the potential role of testosterone binding in altering the mobility of the male version of the protein at puberty showed instead that the protein in the salivas of both males and females is capable of binding androgen. Based on those observations, we named it Androgen-Binding Protein, and thus the study of ABP was born [26]. In those early studies, we also observed that ABPs are dimers composed of an alpha subunit disulfide-bridged to a beta/gamma subunit [27,28]; (see [25] for nomenclature). Each ABP subunit is a four-helix bundle that takes the boomerang form typical of the secretoglobin superfamily and dimerization of an alpha and a beta/gamma subunit allows binding of ligands with steroid-like structures in the cleft formed by the association of the two subunits [29,30].

The role of ligand binding in the function of ABP is still not well understood. Studies of the specificity of the dimer for steroids show that it binds testosterone and progesterone well and HO-progesterone and dihydrotestosterone (DHT) to a lesser extent but it does not bind either cholesterol or estradiol [31]. The structures of these steroids suggest that binding by ABP is governed by the A ring of the steroid, which is more saturated in testosterone, progesterone, and HO-progesterone than in cholesterol and estradiol. Karn and Clements [32] later showed that the two different dimers in mouse saliva, alpha:beta and alpha:gamma (now A27-BG27 and A27-BG26; see [25]) bind dihydrotestosterone (DHT) and testosterone (T), respectively, with different affinities.

In 1991, we made the important observation that androgen-binding by rodent salivas was taxonomically widespread and suggested that the testosterone-binding characteristic must be common to most or all rodents, since it did not seem to be confined to rodents with only one or another ecological niche (e.g. range, diet, etc.; [33]; see also [34]). This led us to propose two possible explanations: (1) general binding/neutralization of dietary substances (toxins, nutrients, etc.) with structures related to steroids; or (2) binding sex steroids for oral and/or olfactory recognition purposes. It seemed clear to us that our finding of a protein(s) with testosterone binding capability in the salivas of a wide variety of cricetid and murid rodents with very different natural histories augured against the first explanation. We concluded that the ubiquitous appearance of ABP in rodent saliva and its relative specificity for androgen are particularly interesting in light of the second explanation, introducing the hypothesis that ABP has a role in sex and/or subspecies recognition. This was the first suggestion of a reasonably well-defined role for mouse salivary ABP. During a visit to the University of Montpellier II in France the previous year, I typed ABP in the salivas of wild-derived house mouse strains maintained by François Bonhomme and his colleagues. These mildly inbred strains of the three subspecies of *Mus musculus* in Europe and Asia had been produced from samples wild-caught in many, widely separated localities in Western and

Eastern Europe, Africa, and Asia. The result of this survey was the surprising observation that unique *Abpa* alleles appeared to be fixed in each of the three different subspecies of *M. musculus*: *Abpa*^{a} in *M. m. domesticus* (western Europe and the Mediterranean basin), *Abpa*^{b} in *M. m. musculus* (eastern Europe to northern China), and *Abpa*^{c} in *M. m. castaneus* (Southeast Asia and Malaysia). This led us to conclude that *Abpa* underwent significant microevolution in conjunction with *M. musculus* subspeciation [33].

We produced the first alpha subunit cDNA sequence [35] and subsequently sequenced the three *Abpa* alleles fixed in the subspecies of *M. musculus*, and those representing three other full species: *Mus spicilegus*, *M. spretus*, and *M. caroli* [36]. The data supported the notion that *Abpa* microevolution paralleled *M. musculus* subspeciation. We postulated that directional selection is a sufficient explanation both for the large ratio of non-synonymous to synonymous substitutions that we observed in the coding sequences of *Abpa* and for the unusual population distribution of *Abpa* alleles [36]. We also postulated a high level of homoplasy because the phylogeny of the *Abpa* haplotypes was incongruent with the canonical phylogeny of the genus *Mus*. That might reflect cyclical selection of certain amino acid variants that became advantageous at some stage in ABPA evolution. These observations led to the development of the congenic strains used in the ABP-based communication studies described below.

3. PROTEINS INVOLVED IN RODENT COMMUNICATION: AN OVERVIEW

Over the past 30 years, three groups of proteinaceous pheromones have been described in murid rodents: the ABPs described above, the exocrine gland secreting peptides (ESPs) and the major urinary proteins (MUPs). These three proteinaceous pheromones have very different molecular properties and each of them has been associated with a different putative pheromone function. Nonetheless all three of these functions influence some aspect of reproduction and thereby have the potential to influence the evolution of the species.

ABPs mediate assortative mate selection based on subspecies recognition and this function can limit gene exchange between subspecies where they meet [24,37]. The various experimental tests of the idea that mouse salivary ABP influences mate selection behavior have been extensively reviewed in [26] and will not be repeated here, however, they have been central to developing a picture of a possible pheromonal function for ABP. Recognition of one's own species versus a foreign one is an extremely important role in the evolutionary scheme of things. More recently, work has shown that ABP-mediated mate preference across a transect of the house mouse hybrid zone in Europe is a case of reproductive character displacement as predicted by reinforcement [38,39] summarized briefly below.

One of the most important developments in evolutionary thinking was the Biological Species Concept, the idea that the process of speciation that leads to evolution of separate species requires the development of reproductive barriers between gene pools (see [40,41] for reviews of hybridization and speciation). Assortative mating is a potentially efficient prezygotic reproductive barrier and may thus prevent loss of genes into unfit hybrids. While the two *M. musculus* subspecies, *M. m. domesticus* and *M. m. musculus*, that meet to form the hybrid zone in Europe are not absolutely reproductively isolated, the hybrids found in the zone between them have reduced fitness [42-44]. When, as in this case, hybrids are less fit, reinforcement should then amplify consubspecific preference most close to a contact zone resulting in increased prezygotic isolation in sympatry relative to allopatry, a phenomenon called reproductive character displacement [45,46]. If the associated preferences are to

contribute to isolation then we should find assortative mating on both sides of the hybrid zone and we would expect to see reproductive character displacement in the hybrid zone, in terms of enhanced preferences, as predicted by the theory of reinforcement.

Two studies that we did in collaboration with our Czech colleagues provided the first hint that a more complex picture of ABP-based preference might occur in this area of secondary contact of both subspecies. Our first study involved the transition of preference of wild house mice for the different ABP signals in *M. m. domesticus* and *M. m. musculus* across the Czech–Bavarian portion of the European hybrid zone [38]. It appeared to show reproductive character displacement where preferences on the verges of the zone of contact are predicted by the theory of reinforcement to be stronger than preferences farther away from the zone of contact [26]. The second study focused directly on the issue of reinforcement, including both diverged signals (i.e. introgression of *Abpa* alleles across the hybrid zone), and associated preferences [39]. It required the development of an explicit model of the effects of reinforcement on a preference trait cline. Models of ABP behavioral clines including a reinforcement parameter showed significantly better fits than sigmoid cline models for both sexes.The results of that work have been reviewed in [26].

ESPs are small mouse proteinaceous pheromones that were discovered and described in the last decade [47,48]. The secreted peptides are much smaller than either the ABPs or the MUPs and are highly diverged from one another both in terms of their sizes as well as their amino acid sequences. Recently, Yoshinaga et al. [49] have reported the three-dimensional structure of ESP1, which consists of a helical fold composed of three helices, stabilized by an intra-molecular disulfide bridge. Female mice respond to direct facial exposure to an ESP expressed in male exorbital lacrimal glands and released into tear fluid by upregulating c-Fos and egr1 gene expression in vomeronasal sensory neurons [48]. The same response occurs after close contact with the face or bedding of male mice, and a recombinant ESP protein stimulates electrical activity in an isolated female vomeronasal organ. The male response to similar signals is unremarkable [47,48,50].

There is now evidence that mouse ESP1 enhances female sexual receptive behavior, lordosis (the position that some female mammals display when they are ready to mate), upon male mounting and copulation [51]. The structural work of Yoshinaga et al. [49] showed that cysteine mutants unable to form the intra-molecular disulfide bridge were incapable of inducing the c-Fos expression, suggesting that the disulfide bridge is crucial to the biological activity of ESP1. ESP1 is recognized by only one vomeronasal type-2 receptor, V2Rp5 [51-53] and the signal received by that receptor in the female vomeronasal organ (VNO) is transmitted to the accessory olfactory bulb (AOB) in the process of eliciting lordosis behavior. They identified amino acid residues involved in the activation of vomeronasal sensory neurons, showing that a highly charged surface is crucial for the ESP1 activity noted above.

The MUPs are lipocalins with the dominant beta-sheet secondary structure folded into beta-barrels capable of binding small ligands in the internal beta-barrel [54-56] and they are primarily expressed in the liver and the products passed through the kidneys into the urine (for a review, see [54]). Each adult mouse expresses a pattern of 8–14 different MUP isoforms in its urine, which is determined by its genotype and by its sex, because some MUP genes show sex-limited expression [54]. These molecules have been shown to mediate female recognition of potential mates, an individual recognition profile that has been likened to a protein ‘bar code’ [55,57-60]. Although MUP-mediated male recognition may function primarily for avoidance of inbreeding, they may also be important in male–male aggression

[61,62]. Chamero et al. [61] isolated high molecular weight components of male urine that activated dissociated vomeronasal neurons and were sufficient to cause male–male aggressive behavior when painted onto previously castrated males. Other studies have shown that both MUPs [63], and a hypothetical MUP peptide formed from the six N-terminal residues EEARSM [64,65], are androgen-regulated nonvolatile compounds capable of accelerating puberty in female mice.

Each of the groups of proteinaceous pheromones, the ABPs, ESPs, and MUPs, constitutes a large family of proteins and each is encoded by a large gene cluster on different chromosomes in the mouse genome. The functions of the ABPs and MUPs have been reviewed recently [26, 54], and the ESPs were described so recently that only one specific function in reproduction, lordosis, has been described to date. This review is focused on recent developments in our knowledge of the evolution of rodent androgen-binding protein (*Abp*) genes, especially the role of retrotransposons in non-allelic homologous recombination (NAHR) implicated in the very recent expansion of this region of the mouse genome. I also compare and contrast the evolutionary history of the *Abp* gene family with the histories of the other two, those encoding the *Mup* and *Esp* genes, including the roles played by gene conversion and selection in their evolution.

4. GENETIC CONTROL OF ABP

The discovery of mouse salivary ABP and it's gene, originally *Abp*, has been reviewed in [26], which also contains a discussion of the distribution of the protein in the salivas of many different rodents and the development of congenic strains for behavioral testing.

Biochemical genetic studies of *Abp* in my laboratory in conjunction with recombinant inbred (RI) strain studies at Jackson Laboratory resulted in a paper that suggested the existence of two dimeric forms of ABP in mouse saliva [28]. We hypothesized that these share a common subunit that we named alpha (encoded by *Abpa*, later revised to *Abpa27*; [25]) but differ in their other subunit, such that one is an alpha–beta dimer and the other an alpha–gamma dimer, where beta is encoded by gene *Abpb* and gamma by gene *Abpg*. The RI study in that paper showed that at least the *Abpa* gene resided on chromosome 7.

A purification protocol [35,66] that we used to obtain the amino acid composition of ABP also allowed us to obtain the alpha subunit of the protein and, with the assistance of Mark Hermodson at Purdue University, we produced its partial amino acid sequence. We used that to design mixed DNA oligomers for probing colony lifts of submaxillary gland cDNAs kindly provided by Kenneth Gross at Roswell Park Memorial Cancer Institute. That work resulted in the first cDNA sequence and thereby the putative amino acid sequence of the alpha subunit [35]. Ten years later, my laboratory published the complete cDNA sequences and putative protein sequences of the alpha, beta and gamma subunits [30] and in that same year we published their gene sequences and their chromosome 7 genetic map, confirming the original "three gene hypothesis" [67].

In 2002, a collaboration was launched between Chris Ponting's laboratory in Oxford, England and my laboratory in the USA that resulted in expanding our appreciation of the size of the *Abp* gene region in the mouse. Rather than finding only the three *Abp* genes we had proposed earlier, we found 27 paralogs (Fig. 1 A), including 14 *Abpa*-like and 13 *Abpb*/*Abpg*-like genes (Fig. 1A; [68]). Clearly, a modified nomenclature was required and we used *Abpa* and *Abpbg* prefixes with numeric suffixes for that purpose. Over the entire map of the *Abp* region on mouse chromosome 7, eleven *Abpa* and *Abpbg* pairs occured in a 5'–5'

orientation and we numbered these 1–11 with higher numbers used for unpaired genes. Karn and Laukaitis [13] later dubbed this structure with an *Abpa-Abpbg* pair in a 5'–5' orientation an <*Abpa-Abpbg*> module (usually abbreviated <*a-bg*>), where the arrows point in the 3' directions, because it appears to be the unit of duplication in those mammals with multiple pairs. The Emes et al. [68] study also showed that the rat genome contained at least three each of the *Abpa* and *Abpbg* paralogs, also in 5'-5' modules, but that the rat paralog complement had expanded independently of that in the mouse genome. By contrast only two *Abp* pseudogenes were found in the human and chimpanzee genomes and we speculated that these genes may have been silenced in the primates.

In the Emes et al. [68] report, we noted that there were several large gaps in the incomplete *Abp* gene region in the build of the mouse genome we studied (Fig. 1A), and we suggested that there might have been more *Abp* genes hidden in those gaps. We later on revisited the mouse *Abp* gene family in a new genome build with a complete *Abp* gene region that allowed us to complete the *Abp* gene complement, including 30 *Abpa* paralogs and 34 *Abpbg* paralogs (Fig. 1B; [25]). Twenty seven pairs of each of these two genes appear as <*Abpa-Abpbg*> (2 modules) or <*Abpbg-Abpa*> (25 modules), leaving only three *Abpa* and seven *Abpbg* paralogs unpaired. In that study, we also presented evidence that the extensive *Abp* gene family seen in the mouse genome began its expansion in the ancestor of the subgenus *Mus* (Fig. 2). We noted that other independent *Abp* expansions had occurred in the European wood rat (*Apodemus*), rat, rabbit, cattle and a marsupial (*Monodelphis*). By contrast, the genomes of other mammals, many of which had low coverage, showed only one *Abpa* and *Abpbg* either as potentially expressed genes or as pseudogenes.

Our laboratory at the University of Arizona continued the work with the *Abp* gene regions in the mouse and rat genomes [13]. An evaluation of the pattern of *Abp* paralogs and interspersed repeated elements (Fig. 3A) suggested that the mechanism for this was duplication of blocks containing combinations of <*Abpa-Abpbg*> modules and single *Abpbg* paralogues. Thus we proposed that parts of the *Abp* region of the mouse genome have duplicated as low-copy repeats (LCRs) consistent with the general mechanism described by [69,70] as non-allelic homologous recombination (NAHR) [70,71]. The pattern in Fig. 3 suggested that the most recent of these block duplications produced the genes <*Abpbg14p-Abpa14p*> *Abpbg31p* <*Abpbg15p-Abpa15*> (hereinafter abbreviated *14-31-15*) and <*Abpbg16p-Abpa16p*> *Abpbg32p* <*Abpbg17p-Abpa17*> (abbreviated *16-32-17*) (Fig. 3B). Apparently this was so recent as not yet to be fixed in the mouse genome as indicated by copy number variation (CNV) that we found for the *14-31-15* and *16-32-17* segment of the *Abp* gene region of the house mouse genome (Fig. 4).

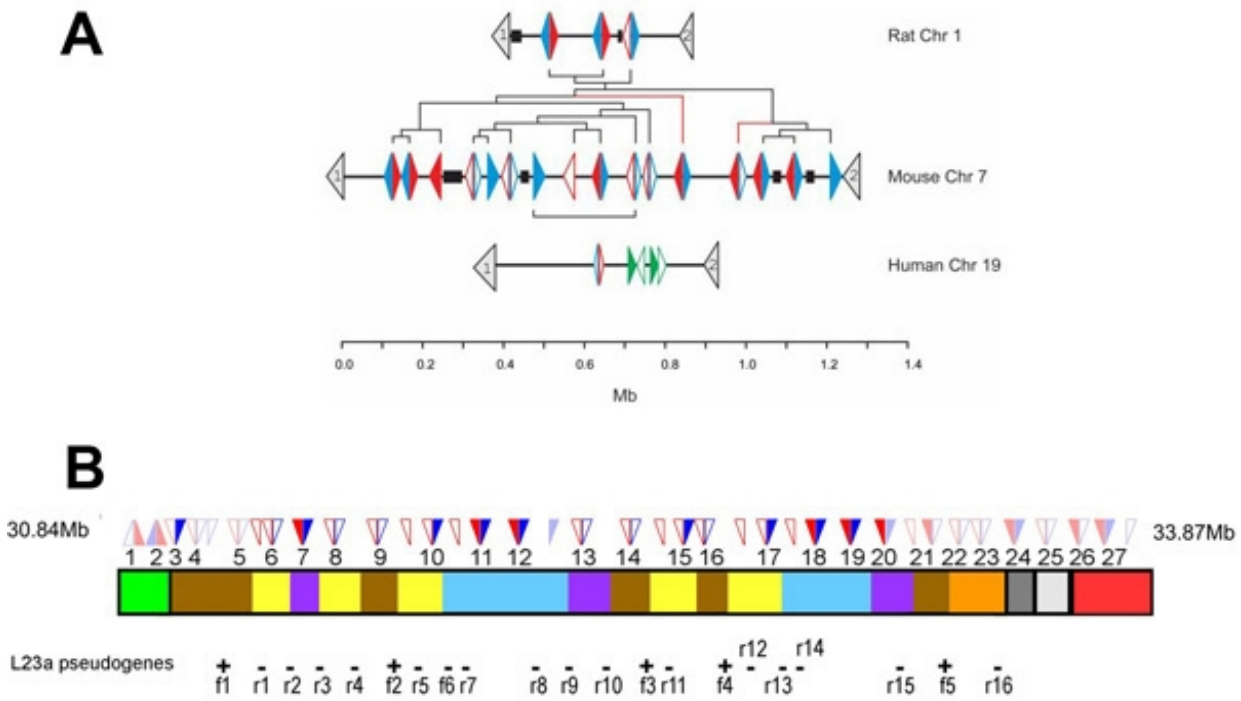

Fig. 1. The Abp gene family linkage groups

Adapted from [68] (Panel A) and from [25] (panel B; original publisher: BioMed Central) with permission

(A) A graphical representation of the relative position and transcriptional orientation of the Abpa- and Abpbg-like genes and pseudogenes located on Rattus norvegicus chromosome 1, Mus musculus chromosome 7, and Homo sapiens chromosome 19 [68]. The 5'–3' orientations of the genes are shown by the direction of the arrowheads. Scn1b and Uble1b genes, which lie in orthologous genomic regions in all three species, are numbered 1 and 2, respectively. Abpa-like genes are shown in blue, Abpbg-like genes in red, and primate SCGB4A1–4(P) genes in green. Filled arrowheads represent predicted functional genes; open arrowheads denote predicted pseudogenes. Gaps (>5 kb) in the genomic assembly of each species are shown as black boxes.

(B) A graphical representation of the relative position and transcriptional orientation of the Abpa and Abpbg genes and pseudogenes located on Mus musculus chromosome 7 [25]. Predicted genes and pseudogenes are indicated as triangles. Abpa paralogues and Abpbg paralogues are colored and filled as in A above. The original paralogues of [68] are faded, while the new paralogues described by [25] are in bright colors. Ribosomal protein L23a pseudogenes on the forward and reverse strands are shown as '+' and '–' symbols, respectively, and are numbered at the bottom. See [25] for other details.

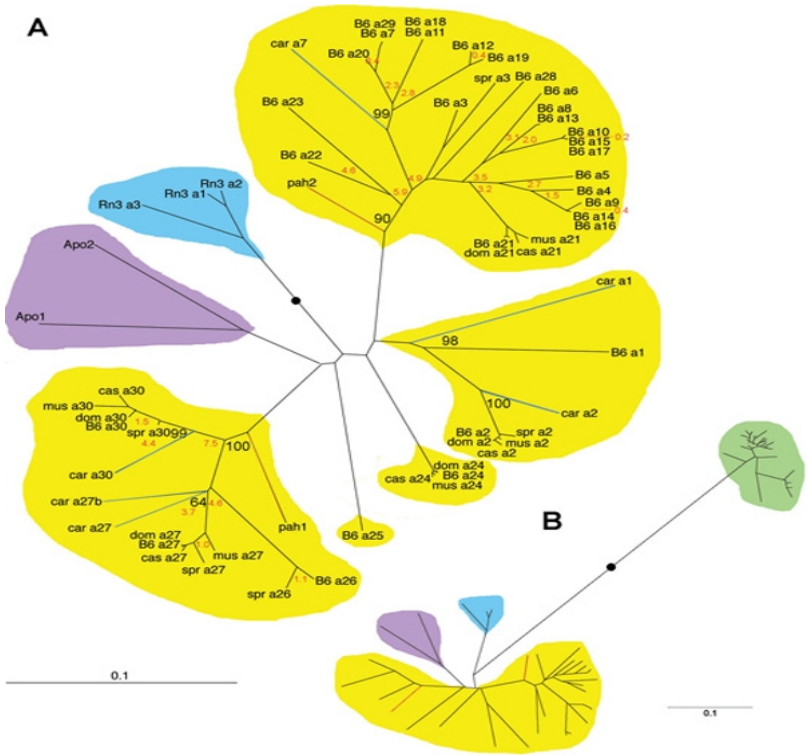

Fig. 2. Evolutionary relationship between Abpa paralogues in rodents and the rabbit

From [25]; (original publisher: BioMed Central) with permission

(A) The 30 Abpa genes found in mm8 are shown as B6_a1–a30. Corresponding genes found in subgenus Mus taxa are abbreviated: dom (M. m. domesticus), mus (M. m. musculus), cas (M.m. castaneus), spr (M. spretus) and car (M. caroli). Abpa genes found in other murid taxa are abbreviated: pah (M. pahari), Apo (Apodemus) and Rn3 (2004 Rattus genome release). Subgenus Mus Abpa clades are shaded in yellow. In the case of two of these clades, an M. pahari paralogue appears as an outgroup (branch shaded in red). M. caroli paralogues serve as outgroups to M. musculus clades (branches shaded in blue). The Apodemus clade is shaded purple and the rat clade blue. The bootstrap values for all internal nodes except two exceed 60%; key bootstrap values are shown in black typeface. Divergence times in millions of years (MY) are shown in red typeface

(B) NJ phylogeny of intron 2 from rabbit and rodent Abpa genes. The rabbit clade is shown with a green background; the rodent clades are shaded as in (A). In both panels, the black dots represent the probable roots of these phylogenies, the locations of which are supported by the locations of rat (A) and rabbit (B) genes which serve as outgroups.

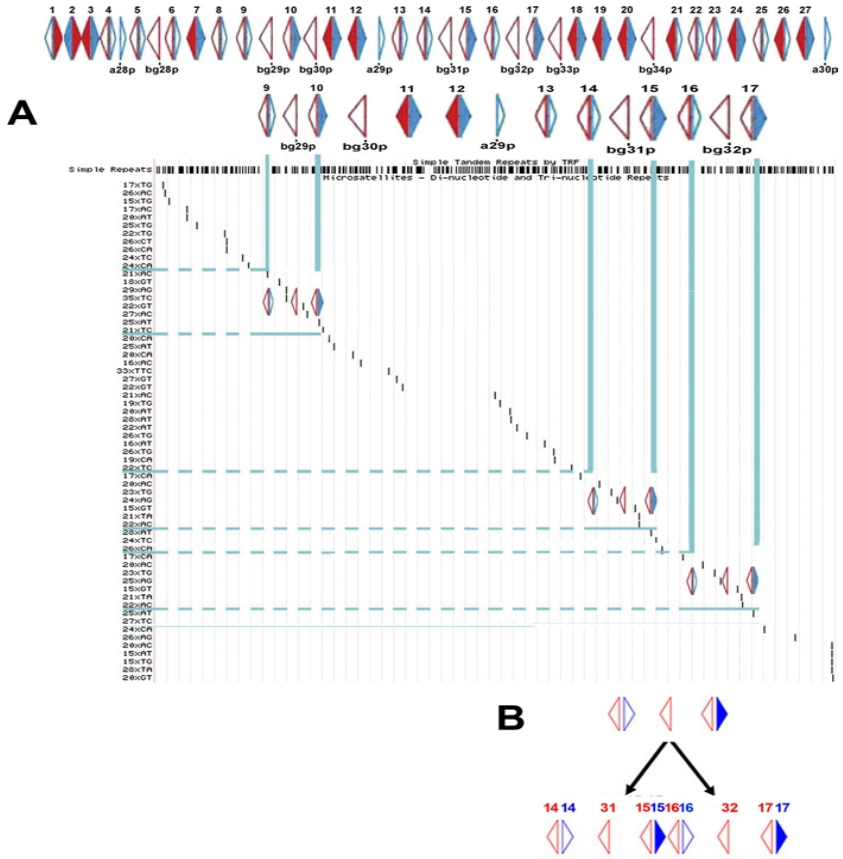

Fig. 3. Use of microsatellite patterns to search for duplicated blocks of Abp genes

Adapted from [13] with permission

(A): the complete map of *Abp* genes modified from [25] with selected duplications shown. Arrow colors, and fill and numbering depict *Abp* genes, psuedogenes and gene modules as in Fig.1. Horizontal blue bars delineate the areas corresponding to the repeated microsatellites, and these were matched to the *Abp* linkage map (vertical blue bars) to look for corresponding repeated patterns of *Abp* genes and pseudogenes. Blocks identified in this fashion include ,*bg9p-a9p.-,bg29p-,bg10p-a10.*, ,*bg14p-a14p.-,bg31p-,bg15p-a15*., and ,*bg16pa16p.-,bg32p-,bg17p-a17*.

(B): Diagram of the most recent duplication in the *Abp* gene region with colors, fill and numbering as above.

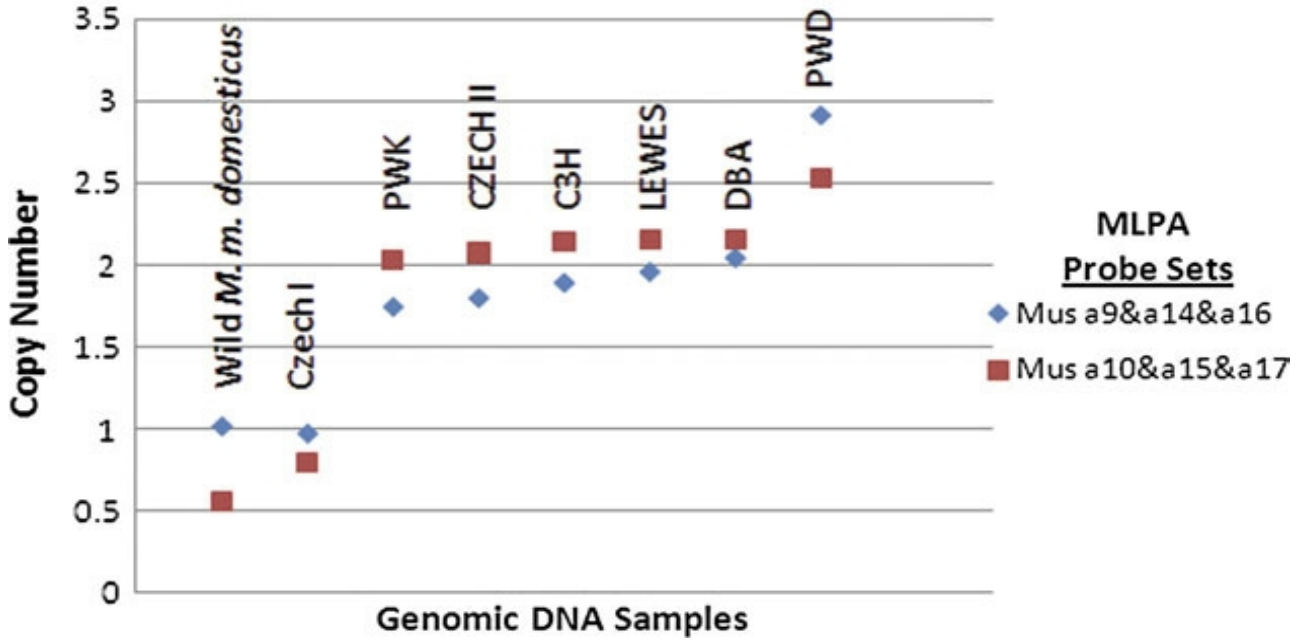

Fig. 4. CNV of a9p/a14p/a16p, and a10/a15/a17

From [13] with permission.

Multiplex ligation probe assay (MLPA) was used to quantitate copy number. The graph shows the copy number for each probe set indexed to values for known single-copy genes

The major question arising from our 2009 study was: How did this mechanism occur? The evolution of gene families is still poorly understood, despite the appearance of an ever increasing number of sequenced genomes. Many of them have expanded much faster than expected based on random gene gain and loss. Transposable elements (i.e. retrotransposons) and selection have been cited as main causes of gene family expansion and contraction [72]. They are associated with local recombination [73-77], perpetuate ectopic recombination [78-83] and are enriched at the breakpoints of segmental duplications in various organisms [84-87]. Because these repeat elements also represent highly homologous sequences, an increase in their local densities may have caused instability in that region of the genome and consequently caused an increase in the rate of NAHR, as proposed for Alu elements in the human genome [85]. While repeat enrichment at the junctions of segmental duplications is associated with only ~12% of all duplications [87], this subgroup may represent the tandem duplications responsible for active expansion of gene families, such as the Alu repeats on human chromosome 22 [88].

We collaborated with Václav Janoušek at Charles University in Prague, to examine the role of repeat element sequences in the expansions of the mouse and rat *Abp* gene families, focusing on those genes because their expansions in the mouse and rat genomes occurred after the divergence of the two species [89]. This involved searching the *Abp* gene region for evidence that retrotransposons contributed to the gene family expansions, and possibly served as the substrates for NAHR. It also resulted in further characterization of the pattern of accumulation of repeats in the *Abp* region, thereby forging a putative link between accumulation of retrotransposons and *Abp* gene family expansion.

Using dot plot analysis, we studied the most recent duplication in the *Abp* region of the mouse genome (*14-31-15* and *16-32-17*; Fig. 3B) and identfied *L1Md_T* repeats on the left flank of *bg14p*, between *a15* and *bg16p*, and on the right flank of *a17* (Fig. 5). Further analysis yielded a candidate breakpoint in the *L1Md_T* elements that flank the *14-31-15* and *16-32-17* duplication, shown in detail in Fig. 5B. A 50 bp sequence is repeated on the left flank of *bg14p*, but appears in only one copy in the *L1Md_T* sequences on the right of *a15* and *a17*. We proposed that the misalignment that created the break probably occurred during replication at the point where synthesis of a new strand had proceeded just to the end of the GGTT (in the first green bracket). Had a hairpin loop formed in strand 1 at that moment, the newly synthesized GGTT end might have slipped ahead to line up with the AGCA further downstream. The resulting mismatch could have been stabilized by the TTT ahead of both tetra-nucleotides and the continuing synthesis of the duplicated 51 bp of sequence beyond them. Such slippage would be expected in the gap sequence because it is rich in a core unit of G followed by three or more Ts. However this happened, destabilization at the mis-paired AGCA/GGTT (second green bracket) would then set the stage for the impending break by which NAHR produced the duplication of the *Abp* genes in this region. Finding this breakpoint in the *L1Md_T* elements that flank the most recently duplicated *Abp* gene blocks strongly supports the hypothesis that this duplication occurred by NAHR [13]. Most reports of such breakpoints in the literature involved those that produce duplications of genes or parts of genes in cancer and other diseases in somatic cells. These instances of NAHR are examples of mitotic recombinations producing clinically significant aberrations but not stable increases in gene copy number in the genome. By contrast, we described breakpoints in repeat elements that produced duplicated *Abp* paralogs by meiotic NAHR in the germ line, a unique observation in studies of gene duplication [89].

Our study also showed that L1 and ERVII retrotransposons are considerably denser in the *Abp* regions than in 1 Mb flanking regions (Fig. 6), while other repeat types are depleted in

the *Abp* regions compared to flanking regions. L1 retrotransposons preferentially accumulated in the *Abp* gene regions after lineage separation and roughly followed the pattern of *Abp* gene expansion. By contrast, the proportion of shared vs. lineage-specific ERVII subfamilies in the *Abp* region resembles the rest of the genome. These high densities of L1 and ERVII repeats in the *Abp* region and their abrupt transitions at the *Abp* gene region boundaries suggest that their higher densities are tightly associated with the *Abp* genes. We also reported that the major contribution to the total L1 density occurred after the split of the two lineages in both genomes, with clear overlap between the accumulation pattern of L1 elements and the *Abp* gene family expansion, at least in the mouse genome. Regardless of whether the higher densities of L1 repeats are a cause or a consequence of the gene family expansion, our study demonstrated the putative link between the accumulation of these elements and the gene family expansion. By contrast, the accumulation pattern of ERVII repeats is complex with a considerable portion of the total ERVII density predating the mouse-rat lineage split, similar to genome-wide patterns.

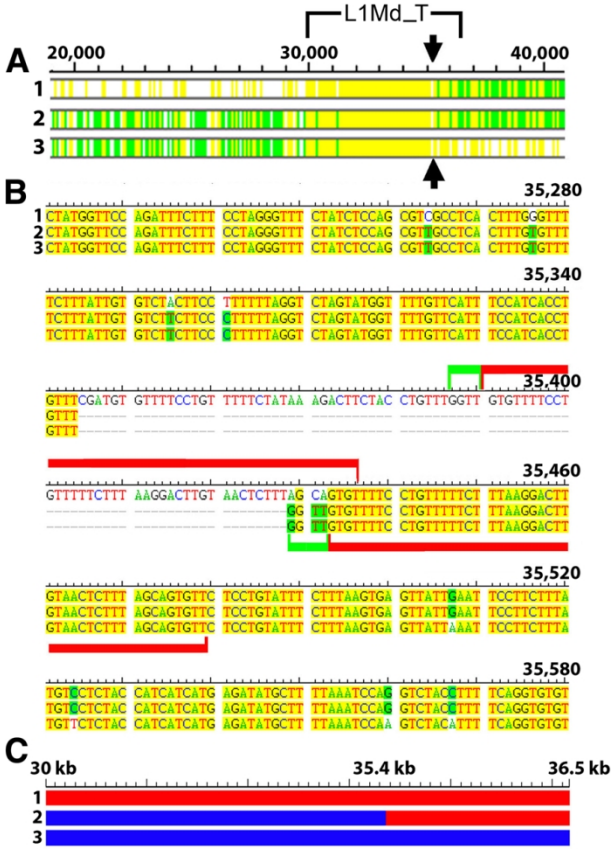

Fig. 5. Breakpoints for the most recent Abp gene duplication

From [89]; (original publisher: BioMed Central) with permission

A) The region of alignment in all three sequences between ~30 kb and ~36 kb includes the nearly full-length L1Md_T sequence. The aligned sequences are shown as bars with the L1Md_T element set off by a bracket above the bars. Bar 1: the 33 kb region immediately to the left of Abpbg14p; bar 2: the L1Md_T repeat on the right flank of Abpa15; and bar 3: the region containing the L1Md_T repeat to the right of Abpa17. Bars in the upper part of the figure show the alignment of the three sequences over slightly more than 20 kb. Regions that align in all three sequences are tinted yellow; regions that align in two of three are tinted green and regions that do not align in any of the three are untinted.

B) An alignment of 360 bp of the three sequences that surround the gap shown by black arrows in panel A. Two duplications that occur within the gap are depicted with red bars. We bracketed in green a GGTT preceding the rest of the duplication, which is marked with a red bar/bracket. We note that there are also shorter indels, e.g., TGTGTTTTCCTGTTTTTC, within the gap. Proximal to and at four nucleotides in the gap (GGTT), sequences 2 and 3 are identical (i.e. seven of eleven divergent sites shown in the figure). However, distal to the gap, 2 is identical to 1, while 3 differs at four divergent sites C) Bars representing the entire L1Md_T sequences 1, 2 and 3 show that, for 384 divergent sites proximal to the breakpoint, 2 is identical to 3; for 127 divergent sites following the breakpoint 2 is identical to 1.

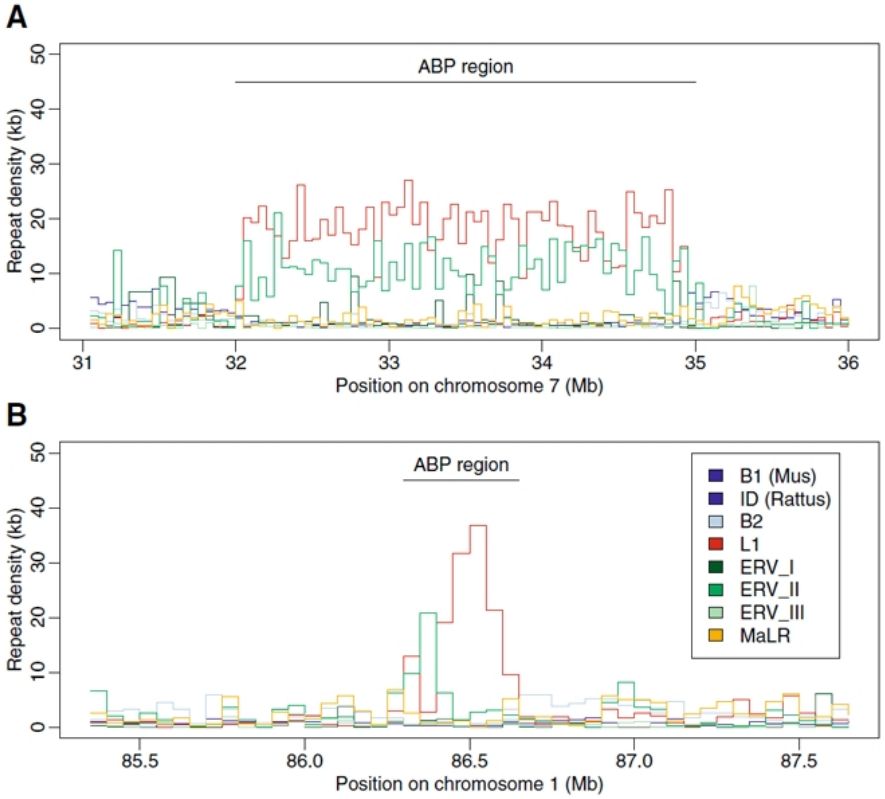

Fig. 6. Densities of repeat families

on chromosome 7 in mouse (A) and chromosome 1 in rat (B) genomes within 50 kb windows plotted along the physical position of the chromosome in the Abp gene family regions and their one Mb flanking regions. Densities are based on the RepeatMasker output provided at the UCSC Genome Browser website for mouse (NCBI M37/mm9) and rat (Baylor 3.4/rn4) genomes. From [89]; (original publisher: BioMed Central) with permission.

5. OVERVIEW OF *ABP* GENE FAMILY EXPANSION IN THE MOUSE GENOME

We suggested that the massive expansion of the *Abp* gene region in the genus *Mus* must have begun by a mechanism different from that which occurred later on, one that better explains the origin of the single <*Abpa-Abpbg*> modules seen in the genomes of most mammals [13]. In a study by Katju and Lynch [90], two-thirds of tandem gene duplicates are in inverse orientation with respect to one another. The inverse (i.e. the 5'–5') orientation of the original pair of *Abpa* and *Abpbg* genes (an <*Abpa-Abpbg*> module) fits this description and we proposed that the original duplication of a single ancestral *Abp* gene in an early

mammalian ancestor produced two paralogues in inverse adjacent order and these evolved into the original <*Abpa-Abpbg*> pair, consistent with the most widespread *Abp* gene configuration in mammals [25]. We also suggested that the first duplication of an <*Abpa-Abpbg*> module occurred in this way in the ancestor of the genus *Mus* because the <*Abpa1-Abpbg1*> . . . <*Abpa2-Abpbg2*> set, constituting one of the oldest clades, is in the inverse orientation to all the other *Abp* paralogues [13]. Following that original duplication, one of the modules was likely the ancestor of <*a1-bg1*> . . . <*a2-bg2*>, and the other was the ancestor of <*bg26-a26p*> . . . <*bg27-a27*> . . . <*a30* (unpaired) and possibly <*bg24-a24*> and <*bg25p-a25p*>. This is because these sets of modules are the oldest in the *Abp* gene group, comprising four of the five ancestral clades of genes described by Laukaitis et al. [25]. It is likely that their large fifth clade arose later, from one or several modules that duplicated by NAHR as described below.

The fact that the most recent *Abp* duplicates in the center of this large group of genes appear in direct, not inverse, order with respect to all members of their clade, and with respect to three of the four other clades suggests that the mechanism of duplication changed following this scenario. This new mechanism was NAHR, by which blocks of multiple modules, sometimes including unpaired *Abpa* and *Abpbg* paralogues, were duplicated, rather than by the relatively slower mechanism of duplication of single <*Abpa-Abpbg*> modules by primer slipping during DNA replication [91].

Our dot plot analyses [89] showed that the *14-31-15* and *16-32-17* duplication was preceded by a duplication of a much larger gene block in the large and most volatile clade in the center of the *Abp* gene region (Fig. 7). This duplication block contained the progenitors of what are now < *Abpbg-Abpa* > modules *7*, *8*, *9*, *10*, *11* and *12*, as well as the single *Abpbg* pseudogenes *29* and *31* and it duplicated to create all the genes identified in the region shown in Final Product (bottom of Figure 1A) except the <*Abpbg*X-*Abpa*X > *Abpbg*Y < *AbpbgZ-AbpaZ* > progenitor of *14-31-15* and *16-32-17*. This accounted for all the other genes from the < *Abpbg*7-*Abpa*7 > to the < *Abpbg*19-*Abpa*19 > modules and, in the process, the *Abpbg*29 gene in the original < *Abpbg*29-*Abpa*29 > module in the center of the duplication product was eliminated.

This and the subsequent *14-31-15* and *16-32-17* duplication pushed the ancestral gene sets apart, leaving the more diverged sequences on the flanks, reminiscent of the mechanism proposed by [92,93], and it helps explain the very high rate of duplication that that was necessary to create the dramatic expansion of the *Abp* region in the mouse genome. This kind of gene region expansion has been called the 'snowball effect' by Kondrashov and Kondrashov [94] and is supported by our observations using the Mouse Paralogy Browser, which suggest that gene birth and death accelerated by the ability to duplicate and delete numerous paralogues in large blocks is common in the *Abp* gene region and created substantial volatility there [13]. The direct repeat nature of the most recent *Abp* duplicates is also consistent with duplication by NAHR, which relies on LCRs flanking the duplicating region and commonly produces direct orientation. Are the *L1Md_T* repeat elements identified by Janoušek et al. [89] responsible for this process?

Since L1 and LTR (including ERVII) repeat families are enriched at junctions of segmental duplications in the mouse and rat genome [84,86,87,95], we speculated that selection for increased gene copy number resulting from densely packed repeat elements was the cause of the association [89]. However, there are alternative explanations as we discussed in our paper. Additional research will be necessary to determine whether selection was indeed

involved and we suggest that the *Abp* gene region is an excellent candidate system for such studies.

6. ABP, ESP AND MUP REGIONS IN MAMMALIAN GENOMES

As described above for rodent genomes, the *Abp* gene arrangement is most often found as an alpha-beta/gamma. pair, the <*Abpa-Abpbg*> module with arrows pointing in the 5'-5' directions, in all mammals [25, 68]. The basal situation in the mammal genome appears to be a single <*a-bg*> module, sometimes with one or more pseudogenes, such as in the little brown bat, horse, cat, dog, squirrel and tree shrew. Interestingly, the primate lineage, including human, chimpanzee, and possibly macaque, apparently has only a pseudogenized pair, suggesting that these genes have been silenced, at least in the Great Ape lineage [25]. However, there have been independent expansions in opossum, cattle, mouse, rat and rabbit genomes involving multiple alpha and/or beta/gamma paralogs and the most extensive of these has already been described in detail above [25].

Kimoto et al [48] described finding 38, 10 and 0 *Esp* paralogs, respectively, in the mouse, rat and human genomes (Fig. 8). Since other genomes remain to be interrogated for their *Esp* complements, it is not possible to determine the basal condition in mammals more widely. The rodent *Esp* gene clusters appear to be older than their *Abp* and *Mup* gene clusters because clades sharing mouse and rat *Esp* genes are common in the rodent *Esp* phylogeny (Fig. 8). This suggests that these *Esp* gene expansions, at least for many, if not most paralogs, began in an ancestor predating the *Mus*/*Rattus* divergence Kimoto et al [48].

Two papers describing the expansions of the *Mup* region in the mouse and rat genomes appeared from different laboratory groups in 2008 [55, 56]. I have only summarized their conclusions here, using the nomenclature of [55]. A single *Mup* gene without evidence of a pseudogene(s) appears to be the basal situation in mammals such as the dog, pig, baboon, chimpanzee, bush-baby and orangutan but not in humans where only a pseudogene with an altered donor splice site has been observed. However, at least two lineage-specific expansions have been found, one in the horse (three *Mup* paralogs) and the other in the grey mouse lemur (*Microcebus murinus*; at least two *Mup* gene paralogs and one presumptive pseudogene). The numbers of *Mup* genes in rodent genomes contrast strongly with what is found in other mammals in that they have large gene clusters that expanded independently in the *M. musculus* and *R. norvegicus* genomes (Fig. 9).

The thread that runs through all three rodent gene families, *Abp*s, *Esp*s and *Mup*s, is that they have greatly expanded in the mouse genome and to a lesser extent in the rat genome. It is tempting to envision that these two lineages, or a common ancestor in the case of *Esp* genes, encountered circumstances in their evolutionary histories where massive increases in the numbers of these three kinds of genes were adaptive. The next section considers two forces, gene conversion and selection, which might have acted during the expansions of these three gene regions in the mouse and rat genomes.

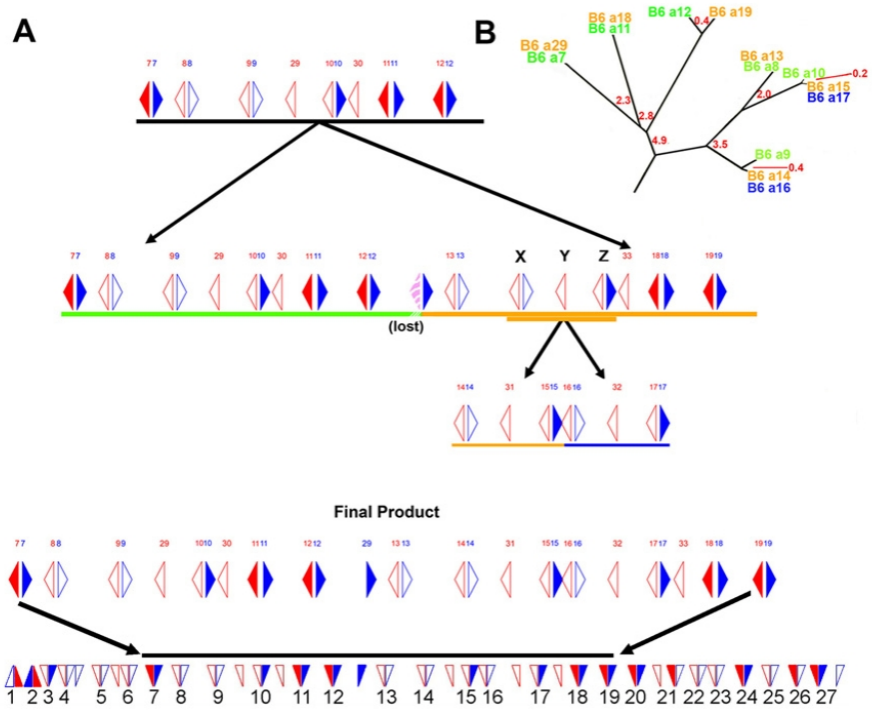

Fig. 7. A model for recent Abp gene duplication events

Modified from [89] (original publisher: BioMed Central) with permission

A) Partial map of Abp genes with arrows colored, filled and numbered as described for Fig. 1 (modified from [89]; original publisher: BioMed Central). Two duplications are depicted: 1) a large block of genes that duplicated to create the products underlined in green and orange; 2) those paralogs/modules designated X, Y and Z in the block on the right created products that are underlined in orange (14-31-15) and blue (16-32-17). The center portion of the Abp gene region created by these two duplications is shown as "Final Product" at the bottom of the figure. The complete linkage map of the Abp region showing where the Final Product is located appears below it.

B) A phylogeny of the genes in the two duplications shown in A. The partial Abp phylogeny (panel B) was modified from Laukaitis et al. [25], who produced an NJ phylogeny of intron 2 from rodent Abpa genes (their Figure 3, yellow clade at the top of the figure). The Abp branches derived from the mouse genome (labeled B6) were retained and the other branches removed. Paralog products are in a typeface color matching the bars that designate their places in the duplications and age estimates in Panel A. The age estimates of the duplications calculated by Laukaitis et al. [25] were also retained and are shown in red typeface.

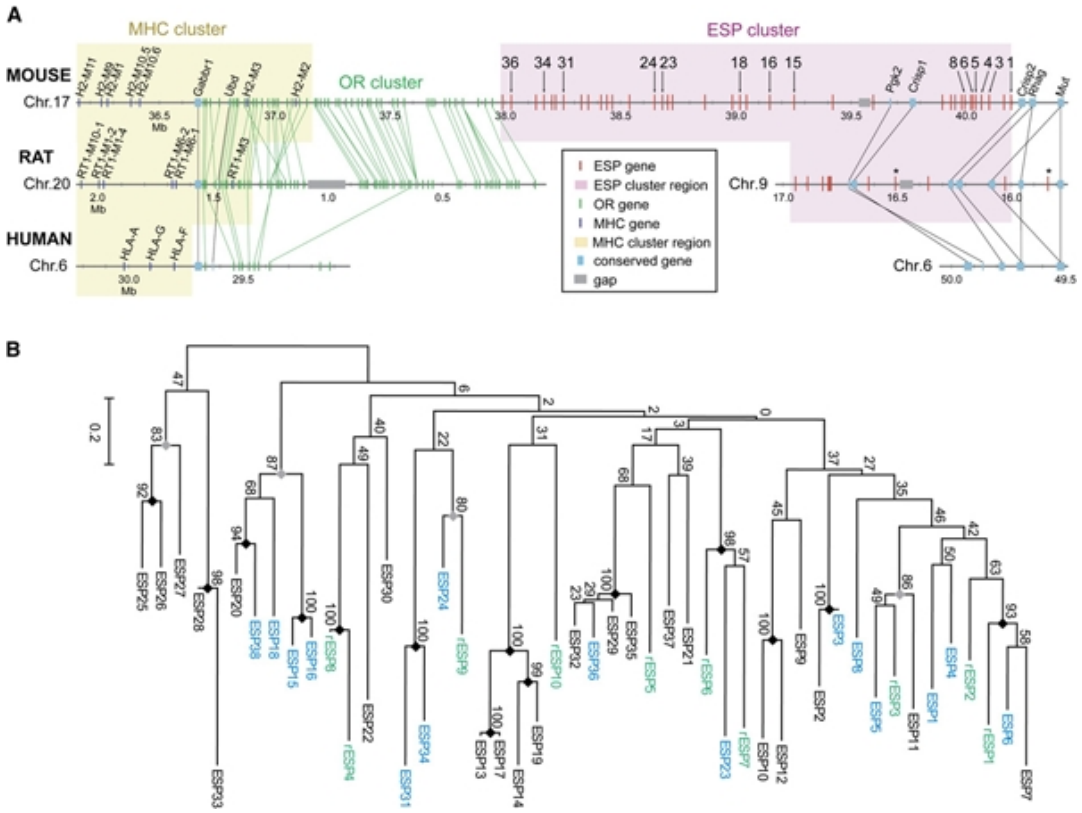

Fig. 8. The ESP Gene Family

From [48] with permission

(A) Genomic location of the ESP gene family in various species. The locations of the last exons of ESP genes are shown as red lines and numbered from ESP1 up to ESP37, counting toward the centromere. ESP38 has not been mapped to a chromosome. The numbers of the expressing ESP genes are indicated. The genes conserved among mouse, rat, and human are shown as blue lines. Colored boxes represent the ESP region (pink) and the MHC cluster (yellow). OR genes are shown in green. Gaps are shown in gray boxes. The two rat ESP genes indicated with an asterisk are identical, possibly because of an error in the database. (B) Phylogeny of ESP genes in mouse and rat. Squares indicate nodes whose bootstrap values are at least 90% (black) and 70% (gray). The expressing ESP genes in mouse are indicated in blue. The rat ESP genes are shown in green. The scale bar represents 0.2 amino acid substitutions per site.

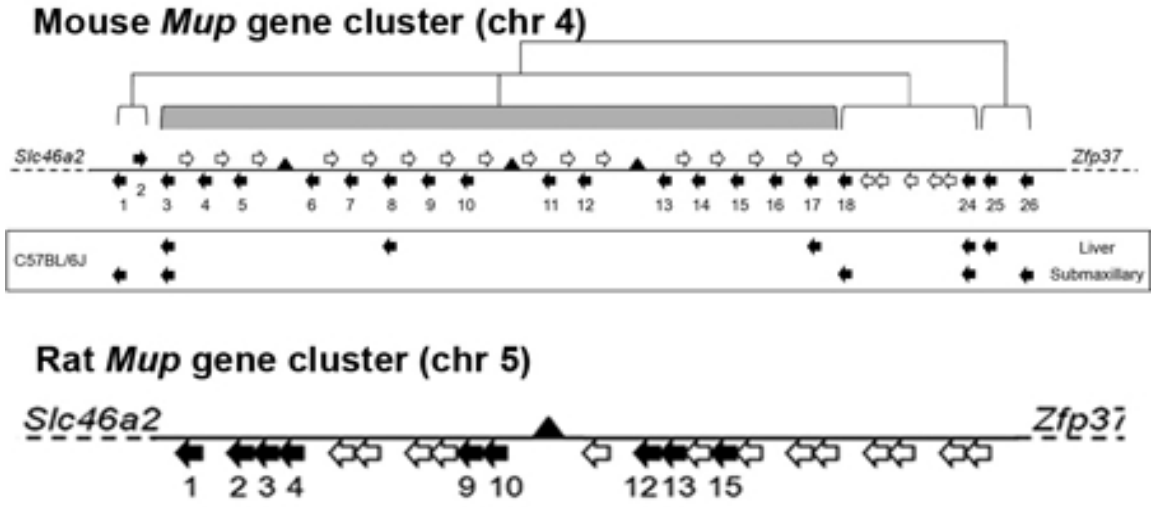

Fig. 9. The mouse *Mup* gene cluster

From [55] with permission

Black arrows indicate direction of coding genes, numbered beneath, in the mouse genome. White arrows indicate direction of pseudogenes. Gaps in the genome are indicated by black triangles. The genes are arranged in two classes based on phylogeny, Class A in open brackets and Class B shaded grey. Genes expressed in male C57BL/6J liver and submaxillary glands indicated by black arrows, by RNA expression analysis.

7. THE CONTRIBUTIONS OF GENE CONVERSION TO THE EVOLUTIONARY HISTORY OF THE THREE GENE REGIONS

The processes responsible for the three gene family expansions in the mouse and rat genomes have undoubtedly resulted in recent gene birth (duplication) and death (deletion), with or without concomitant selection. It is possible that at least some of these events in these evolutionary histories have been obscured by gene conversion and that possibility must be evaluated before considering the role of selection in the evolution of members of the gene family.

In the case of the *Abp* expansion, it was important to consider the possibility that the homogeneity of *Abpa* genes we observed might also have arisen because of non-allelic gene conversion, as previously observed for interferon-alpha [96] and globin [97] genes, for example, rather than by gene duplication [25]. Our studies of *Abp* genes in the mouse genome also revealed that ribosomal protein L23a pseudogenes appear to have frequently co-duplicated with *Abpa-Abpbg* gene pairs (Fig. 1B). These L23a pseudogene duplications must be of recent origin since similar sequences are absent from the syntenic region in the rat. Because the phylogenies of the pseudogenes and their associated *Abp* genes are topologically equivalent, *Abpa-Abpbg* gene pairs appear to have arisen primarily by duplication, presumably via nonallelic homologous recombination, rather than by sequence homogenization after non-allelic gene conversion events.

In spite of the conclusion based on the L23a pseudogenes we tested the question of gene conversion on a local scale by analyzing *Abp* genes with the program GENECONV to look for evidence of short gene conversion tracks. In the case of the *Mup* gene region, Clark et al [98] compared the exonic sequences of four mouse *Mup* genes and cDNA sequences and concluded that an ancestral gene conversion event occurred in some exons. More recently, there has been some speculation that gene conversion played a role in the evolution of the *Mus musculus* Class B *Mup* genes because of the similarity of the gene coding regions and the proteins they encode [55, 56]. Therefore, we extended the GENECONV analysis to include the *Mup* genes using the *Mup* nomenclature of [55]. In addition, we reported the results of the first investigation of the contribution of gene conversion on *Esp* paralogs. GENECONV seeks aligned DNA or protein segments for which a pair of sequences is sufficiently similar to suggest that gene conversion occurred. These are classified as inner or outer fragments. Inner fragments are evidence of a possible gene conversion event between ancestors of two sequences in the alignment. Outer fragments are runs of unique sites that may be evidence of past gene conversion events that originated from outside of the alignment or else from within the alignment but such that evidence of the source has been destroyed by later mutation or gene conversion. (see http:// www.math.wustl.edu/~sawyer/geneconv/gconvdoc.html).

Our GENECONV analysis of *Abpa* paralogs 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 [13]. Our analysis of the *Abpbg* paralogs identified only one inner (*Abpbg26* and *Abpbg34*) and two outer fragments (*Abpbg5p* and *Abpbg19*) that were globally significant. Sequences undergoing frequent gene conversion, either ectopic or allelic, are expected to become GC rich [99,100]. Therefore, we also calculated the GC content of the *Abp* gene region and found that the average GC content in the *Abp* gene region is low, about 41–42%, compared with genes undergoing gene conversion, such as

ribosomal operons and transfer RNAs which have much higher GC contents [100]. We concluded that gene conversion has made a minimal, but not nonexistent, contribution to the evolutionary history of the *Abp* gene family. It certainly has not been significant enough to have confounded the phylogenetic inference presented by Laukaitis et al. [25], and it should not adversely affect further analysis of recently duplicated *Abp* genes.

Our study of gene conversion in the *Mup* genes makes an interesting comparison to the lack of an appreciable contribution of gene conversion to the mouse *Abp* gene expansion [13, 25] because the GENECONV results we reported suggest that gene conversion has also played little if any role in the expansion of the *Mup* gene family [101]. Specifically, we found no evidence for appreciable gene conversion in the *M. musculus* Class A and Class B *Mup* genes and pseudogenes, nor did we find such evidence in the *R. norvegicus Mup* genes. We recognize that GENECONV has low power for detecting conversion events when divergence between duplicates is very low [102], as in the case of the Class B *Mup* genes, and it has also been shown to have high false negative rates [103]. These limitations would have been of greater concern, had we only analyzed the very similar exonic sequences of the mouse Class B *Mup* genes, however, our GENECONV analyses included both the exons and introns of all four *Mup* gene groups we analyzed. This is important because the nucleotide divergences that we subsequently calculated for the Class B *Mup* introns (see below) exceeded by three-fold those of the exons. Moreover, the collective intron sequence between exons encoding the secreted Class B *Mup*s was 3.6 times as large as the total coding exon size. We concluded that we should have detected more evidence of gene conversion in the *Mup* genes, if it exists, than we did given that gene conversion is not expected to act on exons alone. Thus, although the GENECONV program has recognized limitations, we should have detected a significant level of gene conversion in our analysis of whole *Mup* genes, in spite of the conservation of the coding regions of the Class B *Mup*s. Instead we argued that the substantially lower nucleotide divergences in the relatively smaller exons most likely reflect the action of purifying selection on the Class B MUPs. We also calculated the GC content of the mouse and rat *Mup* gene regions and found the following average GC contents in the four sets of *Mup* paralogs: Class A *Mup* genes, 39.89%; Class B genes, 41.31%; Class B pseudogenes, 39.76%; and rat genes, 45.46%. These GC contents in the various rodent *Mup* gene regions are relatively low compared with genes undergoing gene conversion [99,104,105]. We note, however, that there is conflicting data on whether increased GC content is consistent with gene conversion [102]. Nonetheless, we feel that the low GC contents support the conclusion from the GENECONV analyses of the whole *Mup* genes that conversion has contributed minimally to the expansions of these gene families.

Given this apparently consistent picture of *Mup* and *Abp* gene evolution, it was a striking contrast to find evidence of extensive gene conversion in many *Esp* genes, although we did not find it in all of them [101]. The *Esp* paralogs involved were all found in inner fragments and none in outer fragments. Those in the inner fragments were identified with the same short DNA sequence that ranged from 20–30 bp, depending on whether a mismatch was allowed (Fig. 10). Perhaps one of our most important observations was that a number of the *Esp* inner fragments revealed by GENECONV involved both a mouse paralog and a rat paralog. That finding is consistent with the conclusion of Kimoto et al [48] that the *Esp* gene expansion, at least for many/most paralogs, began in an ancestor predating the *Mus*/*Rattus* divergence. Thus it seems that the *Abp*s and the *Mup*s expanded without much contribution from gene conversion, while the expansion of the older *Esp* family shows significant evidence that gene conversion was involved in a region that affected the proximal part of the coding region of the secreted peptides. Interestingly, the *Esp* paralogs involved in inner

fragments in our analysis had GC content values that are not particularly high compared to other genes that have undergone gene conversion, at odds with expectations.

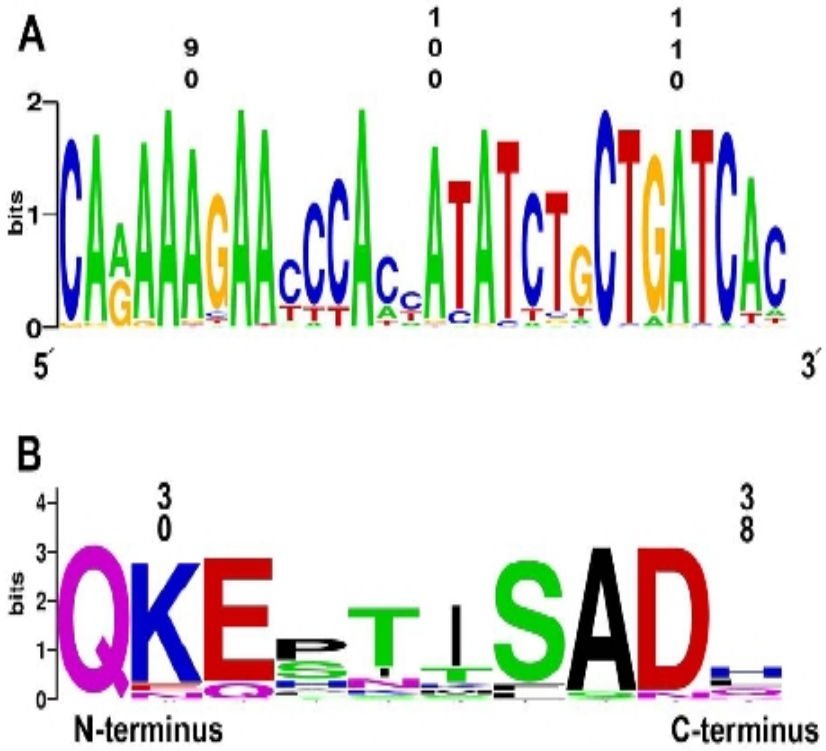

Fig. 10. WEBLOGO of the inner fragment shared by 21/38 mouse and 9/10 rat Esp genes

From [101] with permission

(A): The nucleotide sequence in the gene converted region for the expressed mouse Esp genes and the rat Esp genes involved in gene conversion.

(B): The translation of the inner fragment sequence. The y-axis values are bits, the maximum entropy for the given sequence type (log_2 4 = 2 bits for DNA/RNA, log_2 20 = 4.3 bits for protein; weblogo.berkeley.edu/info.html).

8. THE CONTRIBUTIONS OF SELECTION TO THE EVOLUTIONARY HISTORY OF THE THREE GENE EXPANSIONS

Studies of the contribution of selection to *Abpa* evolutionary history include: 1) high values of *Ka*/ *K*s of *Abpa* (now *Abpa27*), *Abpb* (*Abpbg27*) and *Abpg* (*Abpbg26*) coding regions [36,106,107]; 2) high ω (i.e. *dN/dS*) values from CODEML analysis of the numerous *Abpa* and *Abpbg* paralogs arising from rapid gene duplication [25, 68]; 3) elevated nucleotide divergences in the coding region for the secreted *Abpbg26* and *Abpbg27* subunits compared to their introns or the coding region for their signal peptides [67]; and 4) identification of sites under selection in subunits secreted into mouse saliva [108]. This last report will be reviewed here.

It appears, in light of mouse saliva proteome data [108,109], that *Abpa27*, *Abpbg26* and *Abpbg27* are the only *Abp* genes expressed in the glands that contribute proteins to mouse

salivas. We obtained the sequences of the *a27*, *bg27* and *bg26* genes in five taxa of *Mus*, including the three subspecies of *Mus musculus*, as well as *Mus spicilegus* and *Mus spretus* and constructed gene phylogenies and found that each of the three genes has a substantial number of nonsynonymous substitutions in its coding region. We investigated the evolutionary history of all three of these *Abp* genes using the CODEML program to identify selected sites, i.e. to look for the footprints of adaptive evolution [108].

Our CODEML analysis supports the notion that both subunits, A27 and BG26 (the products of the *Abpa27* and *Abpbg26* genes, respectively), have a history of adaptive evolution driven by positive selection on a few sites at the surface of one face of the dimer they form, consistent with analysis on *Abp* paralogs [68]. We suggested that these two ABP subunits evolved rapidly because they form one of the dimers secreted in relatively large quantity into the salivas of mice and we also proposed that and that their evolutionary histories may not be independent. This is not unexpected given that at least A27 appears to be involved in incipient reinforcement on the hybrid zone where *M. m. domesticus* and *M. m. musculus* made secondary contact [39] but it must do so in conjunction with at least one BG monomer because all ABP subunits described to date are paired in an alpha-beta/gamma dimer.

Four amino acid sites of A27 apparently evolved under positive selection as have at least two of BG26, however, it is less clear whether the rapid evolution of A27 and BG26 has been shared by BG27. The CODEML result for bg27 was non-significant, however, there were three sites in the sequence of BG27 identified as positively selected at a Bayes empirical Bayse (BEB) [110] posterior probability threshold of .95. Whether to accept such sites as selected, given the non-significant CODEML result, is controversial but it is nonetheless interesting that two of the three sites are at the surface of one face of the dimer formed with the A27 subunit resulting in a model that looks very similar to that of the A27-BG26 dimer (Fig. 11). Our selection results suggest that at least A27 and BG26 may have coevolved because they are the subunits of the same functional entity, the A27-BG26 dimer found in mouse saliva. BG27 may have also coevolved with A27 because these two subunits form a dimer found in mouse saliva at a level nearly equal-molar with the A27-BG26 dimer [32]. The coevolution hypothesis is supported by the fact that both of the residues under selection in BG26, and two of the three residues under selection in BG27, share the same exterior face of the dimer with the four residues under selection in A27. This location of all but one of these selected residues suggests that these ABP dimers interact with another molecule(s).

It is interesting to compare these selection results with our earlier work [32] in which we showed that the two different dimers, A27-BG26 and A27-BG27 bind dihydrotestosterone (DHT) and testosterone (T), respectively, with different affinities. How amino acid variation at sites on an exterior face of the molecule might make this possible remains to be determined, however, it is likely that the different binding affinities result from a synergistic affect caused by the interaction of A27 with the BG monomer in each of the two dimers. One possibility is that conformation of the binding pocket depends on which BG subunit is involved, since this pocket is created by the formation of the dimer [29, 30]. Indeed BG26 and BG27 differ at more amino acid sites than the ones shown here to be evolving under positive selection [30]. Thus, while coevolution of A27 with one or both BG26 and BG27 can affect a surface interaction with another molecule, such as a receptor, changes at other amino acid sites independently, or in conjunction with the positively selected site, can affect ligand binding.

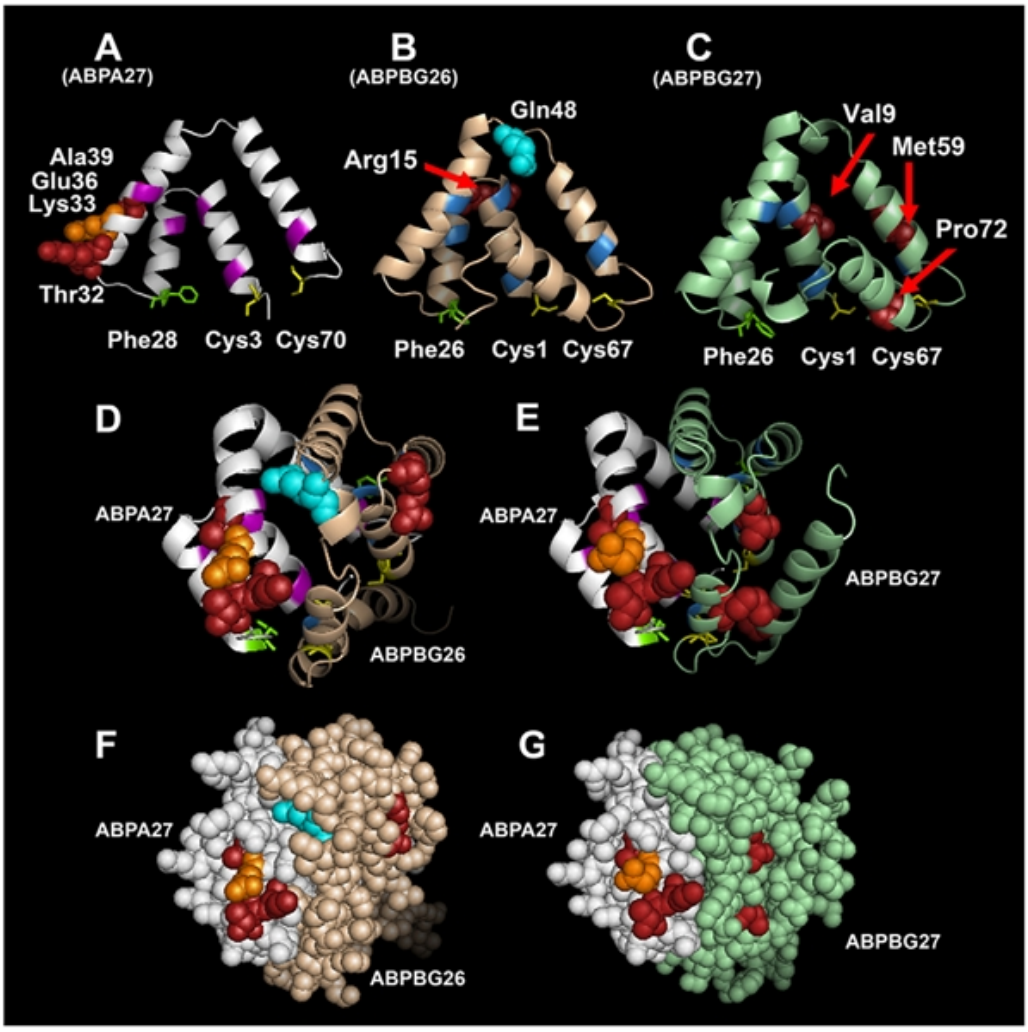

Fig. 11. Positive selection at sites on molecular models of mouse salivary ABPA27 and ABPBG26 and the dimer they form

From [108] with permission

Panels A–E: cartoon-format models with spiral ribbons representing alpha helices and thinner connecting lines representing loops. Panels F and G: illustration of the dimers composed of the two subunits with amino acid residues as filled spheres. Panels A–C: inside views of the three ABP subunits: ABPA27 (white), ABPBG26 (tan) and ABPBG27 (light green), respectively. The conserved residues that line the ligand-binding pocket are represented in purple for the ABPA27 model and blue for the ABPBG26 and ABPBG27 models, with the conserved Phe that coordinates the orientation of the monomers shown in green with stick side chain. The Cys residues that form disulfide bridges uniting the monomers in antiparallel orientation are shown in yellow with stick side chains. The residue in bright blue is selected with a BEB posterior probability of .99; those in red with BEB posterior probabilities of .95; and the one in orange with a BEB posterior probability of .90. Panel D: the ABPA27-ABPBG26 dimer in cartoon format. Panel E: the ABPA27-ABPBG27 dimer in cartoon format. Panel F: a solid model of the ABPA27-ABPBG26 dimer showing all six selected residues on one face. Panel G: a solid model of the ABPA27-ABPBG27 dimer showing six of the seven selected residues on one face (one of the ABPBG27 selected residues is out of sight on the left side).

We produced congenic strains from two strains with different a27 alleles, *Abpa27*^{a} and *Abpa27*^{b} on the same genetic background for use in behavioral testing (reviewed in [26]). At the time that the two congenic strains were used for testing, it was clear that whatever affect A27 has on behavior, it does so as a subunit of a dimer with either BG26 or BG27, since no free monomers had been observed in mouse salivas [28]. Thus it is not surprising that these genes, *bg26* and *bg27*, which are closely linked to *a27* and share its evolutionary history [25], should both also have a role in mediating mate preference behavior. It seems evident that we can conclude this about the A27-BG26 dimer at the very least and perhaps about the A27-BG27 dimer as well.

Given the striking differences in divergence of the members within the two subfamilies of *M. musculus Mup* genes, we considered the possibility that they evolved under different selection regimens. To explore this, we compared nucleotide divergences of the exons to those of the introns. Both intron divergence and the synonymous nucleotide sites in the coding region (represented by *dS*) are for the most part thought to be free of selective constraints and thus their values should be similar. This is because comparisons of homologous DNA sequences for many different genes reported by Hayashida and Miyata [111] showed that silent positions of protein-encoding regions (estimated by *Ks* or, alternatively, *dS*) and introns (which we estimated with nucleotide divergences) evolve at high and remarkably similar rates for different genes. Those authors concluded that the evolutionary clocks at the DNA level in such divergent blocks as silent positions and introns run at essentially the same rates for many different genes over a long period of evolutionary time. By contrast, the coding region is predicted to show higher nucleotide variability than the introns in the case of positive selection. The prediction is the opposite in the case of purifying (negative) selection, wherein the coding region should show reduced nucleotide variability compared to the introns. To make this comparison, we removed the signal peptide coding region from consideration because it is expected to be under different selective constraints than the region encoding the secreted protein.

As expected, the overall nucleotide divergence values that we calculated for the *M. musculus* Class A and B *Mup* concatenated introns agreed well with previously published *dS* values [55]. Nucleotide divergences of the *M. musculus* Class A *Mup* exons and introns were not significantly different from each other (onetailed t test modified from [112]; see [101]) and the nucleotide divergence value that we calculated for the Class A concatenated introns agreed well with the previously published *dS* values [55]. In the case of the *M. musculus* Class B *Mup* genes, the exons showed significantly less nucleotide divergence than the introns. As in the case of the *M. musculus* Class A *Mup*s, the nucleotide divergence values that we calculated for the *M. musculus* Class B concatenated introns agreed well with the previously published *dS* values. In an analysis of *R. norvegicus Mup* exons and introns, the nucleotide divergence in exons significantly exceeded that in the introns. In this case, however, the divergence of the concatenated introns was less (0.059) than the previously published *dS* value (0.098). In light of the lack of evidence for gene conversion, our data suggested that the exons of the mouse Class A and rat *Mup* genes have experienced significant nucleotide substitution in their evolutionary histories while, by comparison, the mouse Class B *Mup* genes seemed to have been under purifying selection.

We applied the CODEML sites analysis to the *Mup* codons as we have done previously for the *Abp* codons. At least two MUP amino acid sites in beta-sheets of each of the mouse Class A and Class B MUPs, as well as in the rat MUPs were identified as having evolved under positive selection (Fig. 12). These sites are in a beta-barrel in the interior of the molecule where they might influence the nature of the ligand preferentially bound. This

stands in strong contrast to the ABP sites under selection in both the alpha and beta/gamma subunits (above), which fall on the surface of one face of the dimer where they could be involved in interaction with other molecules (e.g. receptors; [68,108]). Nonetheless we cannot rule out that one or more of the MUP surface residues might interact with a receptor(s).

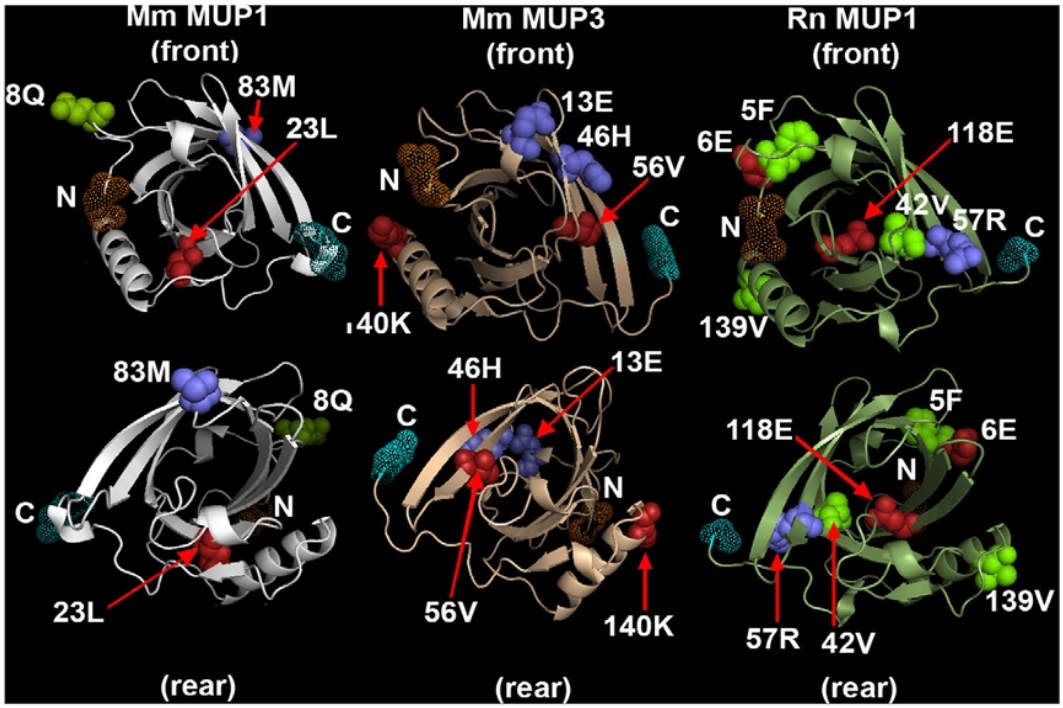

Fig. 12. Positive selection on rodent MUPs

From [101] with permission

Selected sites are plotted on molecular models of mouse MUP1 (left), MUP3 (center) and rat MUP1 (right), representing the mouse Class A, Class B and rat MUPs. Both mouse MUP1 and MUP3 were mapped on the d1znda1 model and rat MUP1was mapped on the d2a2ua model with PyMol. Residues with a BEB posterior probability .99 are in red; a BEB posterior probability .95 are in green; and a BEB posterior probability .90 are in blue. In all the models, alpha-helices are shown as spiral tapes and beta-sheets are shown as flat arrows. The eight-sheet beta-barrel can be seen in the center of each model. At least two of the selected sites map to different beta-sheets in the beta-barrel of all three structures.

Mouse and rat *Esp* genes differ in many ways from the *Abp* and *Mup* genes of the two species. The *Esp* genes are much smaller than those genes and vary widely from each other in the lengths of the secreted ESP peptides they encode. Although their signal peptide coding regions and the proximal ends of the coding regions of their secreted sequences align reasonably well, sequence similarity deteriorates rapidly proceeding toward their 3' ends. We documented substantial gene conversion affecting ~30 bp near the 5' end of the secreted protein coding region in more than half of the 38 mouse *Esp* genes and nearly all ten of the rat *Esp* genes (above). This represents a significant portion of the relatively small coding regions of many of these genes. There is also evidence that the *Esp* gene expansion appears to be older than that in either the *Mup* or *Abp* genes, possibly predating the

divergence of *M. musculus* and *R. norvegicus* [48]. That raises the concern that an *Esp* phylogeny might be biased by the phenomenon of long branch attraction wherein homoplasy will increase the probability that two lineages will evolve the same nucleotide at the same site [113]. The resulting bias in the gene phylogeny will confound tree-based analyses such as CODEML. Given these concerns, we took a different approach to evaluate the possibility that selection has acted on the ESPs, opting to determine *Ka/Ks* on the Exon 3 sequences with and without the converted sequence segment identified with GENECONV (Fig. 13). Our data provide preliminary evidence that at least some *Esp* paralogs experienced positive selection during the expansion of the mouse gene family. Unfortunately, this data does not provide site-specific selection results as was the case with both the ABPs and MUPs, however, it is very likely that CODEML would have given spurious results, particularly as *Esp* alignments deteriorate rapidly proceeding toward their 3' ends.

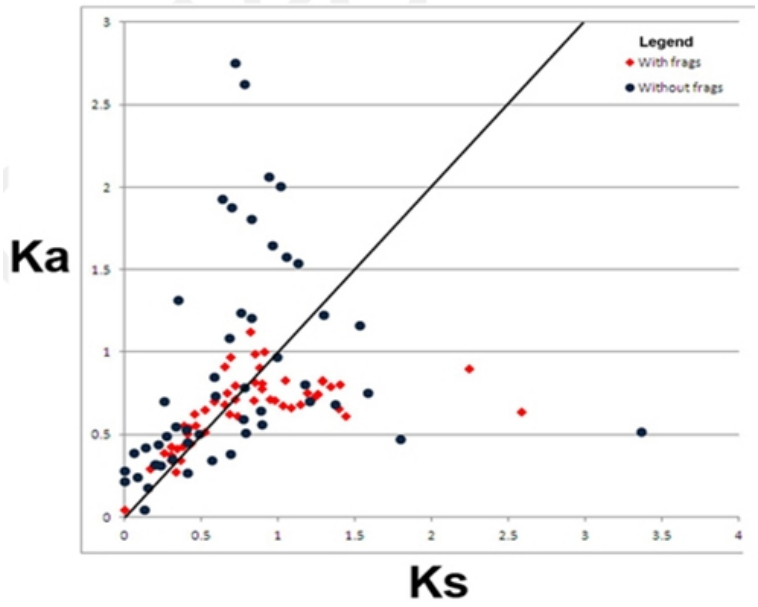

Fig. 13. Ka plotted vs. Ks for selected mouse Esp sequences

From [101] with permission

The line demarcates a slope of 1.0. Each sequence is plotted twice. The red diamonds mark the Esp sequences including inner fragment sequences and the blue dots show the same Esp sequences with the inner fragment sequences removed

9. HOW DID EVOLUTION INFLUENCE PROTEIN FUNCTION?

The proteins encoded by each of the three gene families appear to have evolved a unique type of function that influences a different aspect of reproduction. ABPs have been shown to mediate assortative mate selection, based on subspecies recognition that potentially limits gene exchange between subspecies where they meet [24,37]. Moreover, there is evidence that ABP-mediated mate preference at the edges of the European mouse hybrid zone exhibit reproductive character displacement as predicted by reinforcement [38]. Models with reinforcement components suggest that ABP-mediated mate preference constitutes a system of incipient reinforcement where *M. m. domesticus* and *M. m. musculus* make secondary contact, the house mouse hybrid zone in Europe [39].

MUPs have been shown to mediate female recognition of potential mates to avoid inbreeding and they have also been implicated in male–male aggression and have been reported to accelerate puberty in female mice (reviewed in [54]). Several attempts have been made to connect MUP function to subspecies recognition, as has been done with ABP, however, such a connection seems unlikely for several reasons. One reason is that any heritable signal mediating subspecies recognition and discrimination must involve a gene encoding a protein, or a combination of proteins consistently similar among members of each subspecies but significantly different between the two subspecies to be recognizable [39, 101]. In the case of the ABP system, different *Abpa27*, *Abpbg26* and *Abpbg27* alleles are fixed in *M. m. domesticus* and *M. m. musculus* [33 36] but there is no evidence that different *Mup* genes are fixed in the two subspecies [39, 101]. In fact, the signal used in most of the tests suggested to involve MUPs was urine or bedding in which other constituents capable of firing VNO receptors have been identified, in particular sulfated steroids [114,115] and (methylthio) methanethiol [116]. In short, the specific odorant compounds involved in recognition based on urine have not as yet been characterized [117]. The most serious concern, however, stems from the results of actual mate preference tests that show: 1) wild house mice use self-reference matching of MUP patterns to avoid inbreeding [118] and 2) female house mice show a consistent preference for associating with *Mup* heterozygous males over *Mup* homozygous males when heterozygosity across the rest of the mouse genome was controlled [119]. Thus the preponderance of behavioral evidence supports MUP-based disassortative mating, consistent with the lack of evidence for any *Mup* alleles fixed in different subspecies.

By contrast to the ABPs and MUPs, less is known about the function(s) of the ESPs. At least one of them, ESP1, appears to enhance lordosis and copulation [51], however, the function(s) of the other ESPs are unknown in spite of the fact that fourteen of the remaining 37 are expressed [48]. Lordosis is an intrinsic component of copulation and might be expected to have evolved before the recognition functions of the younger two pheromone gene families described above. This is consistent with the observation that the *Esp* gene family expansion appears to predate the mouse-rat divergence [48]. By contrast, *Abp* and *Mup* gene families appear to have expanded relatively recently and rapidly, duplicating numerous paralogs that already had become pseudogenes in the process.

It should not be surprising that the majority of ABP sites evolving under positive selection are on one face of the surface of the protein [68, 108] and that these are fixed differences between the two subspecies [33,106,107]. As described above, we expect these characteristics for a molecule or a combination of molecules consistently similar among members of either subspecies but sufficiently different between the two to be a recognizable signal for subspecies recognition. These subspecies recognition sites probably evolved under the kind of cyclical selection of certain amino acid variants we have previously observed [36]. Those would have become advantageous at one stage or another in repeated selective sweeps [106, 107]. A recent report suggests that alpha and beta/gamma subunits may have coevolved such sites for harmonious function in the dimeric form that mediates recognition [108].

Our MUP selection analysis suggests that the role of the bound ligand may have equal or even more importance in recognition than specific sites on the surface of the protein and thus selection might rather be directed at sites on the interior of the beta-barrel where ligand binding specificity is determined. This would explain why both classes of mouse MUPs as well as rat MUPs have at least two selected sites on beta-sheet secondary structure in the beta-barrel. It is particularly interesting that we found positively selected sites in the mouse

Class B MUPs given the conserved sequences in this group [55,56]. The nucleotide divergence data we report here suggest that purifying selection has had an especially strong role in the evolution of this group compared to the mouse Class A and rat *Mup* genes. Nonetheless we were able to identify a few specific sites in each group that were subject to positive selection and over half of those were in beta-sheet secondary structure in the beta-barrel where ligands are bound.

It is relatively easy to envision the need of the ABP and MUP communication systems for evolution of multiple paralogs that play different roles individually, or in combination, to satisfy the need for the kinds of functions described above. This will take on added importance if ligands bound by the encoded proteins diversify their functions even more. In both cases, duplication of a progenitor paralog during a rapid and specific gene family expansion, with nucleotide substitutions at nonsynonymous sites driven by positive selection would provide new paralogs with potentially adaptive functions. By contrast, the need for the number of paralogs in the ESP family is not nearly so clear since the only ESP function known at this time is lordosis mediated by ESP1. While it is tempting to speculate that there are undiscovered functions beyond lordosis that require the number of ESPs that are apparently expressed [48], there is not enough additional information about ESPs to explain the gene conversion among so many mouse and rat paralogs, a biased gene conversion that seems to be under some sort of selection. Moreover, there is no obvious explanation for the positive selection suggested by the elevated *Ka/Ks* we reported. More work will have to be done on these interesting peptides to shed light on a potential role for diversity in their functions that might match the diversity in their structures.

10. CONCLUSION

The three rodent gene families, *Abp*s, *Esp*s and *Mup*s, expanded at different times with the *Esp* expansion being the oldest. Selection seems to have been important in all three families but there is convincing evidence for gene conversion only in the *Esp* family. The functions of the genes of each family appear to be quite different as would be expected from their very different evolutionary histories. It is interesting to contemplate what sort of forces caused the primordial genes for the three different pheromones to all undergo dramatic expansions. Early in the expansion process, duplication may have been driven by a need for additional copies of the same gene, in the sense of increasing the volume of the signal. As previous copies underwent functional divergence or died (became pseudogenized), more copies were necessary to maintain the enhanced volume of the original signal, requiring more duplication. On the other hand, rapid divergence of pheromone proteins to produce more unique group identities, as in the MUP bar-code analogy, could also have been the selective factor behind increased duplication. We must learn much more about the functions of the individual proteins to determine the driving forces, which could differ among the three pheromone systems.

ACKNOWLEDGEMENTS

The author wishes to thank Václav Janoušek and Christina Laukaitis for helpful comments.

COMPETING INTERESTS

The author has declared that no competing interests exist.

REFERENCES

1. Emes RD, Goodstadt L, Winter EE, Ponting CP. Comparison of the genomes of human and mouse lays the foundation of genome zoology. Hum Mol Genet. 2003;12(7):701-709.
2. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860-921.
3. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P et al. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002;420(6915):520-562.
4. Hughes AL: Adaptive evolution of genes and genomes. New York: Oxford University Press; 1999.
5. Ohno S: Evolution by gene duplication. New York: Springer Verlag; 1970.
6. Hurst LD, Smith NG. Do essential genes evolve slowly? Curr Biol. 1999;9(14):747-750.
7. Hughes AL, Nei M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature. 1988;335(6186):167-170.
8. Jensen JD, Wong A, Aquadro CF. Approaches for identifying targets of positive selection. Trends Genet. 2007;23(11):568-577.
9. Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 1986;3(5):418-426.
10. Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ et al. A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol. 2005;3(6):e170.
11. Yang ZH, Bielawski JP. Statistical methods for detecting molecular adaptation. Trends in Ecology & Evolution. 2000;15(12):496-503.
12. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE et al. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature. 2004;428(6982):493-521.
13. Karn RC, Laukaitis CM. The mechanism of expansion and the volatility it created in three pheromone gene clusters in the mouse (Mus musculus) genome. Genome Biol Evol. 2009;1:494-503.
14. Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD et al. Natural selection on protein-coding genes in the human genome. Nature. 2005;437(7062):1153-1157.
15. Clark NL, Aagaard JE, Swanson WJ. Evolution of reproductive proteins from animals and plants. Reproduction. 2006;131(1):11-22.
16. Swanson WJ, Vacquier VD. The rapid evolution of reproductive proteins. Nat Rev Genet. 2002;3(2):137-144.
17. Castillo-Davis CI, Kondrashov FA, Hartl DL, Kulathinal RJ. The functional genomic distribution of protein divergence in two animal phyla: coevolution, genomic conflict, and constraint. Genome Res. 2004;14(5):802-811.
18. Dean MD, Clark NL, Findlay GD, Karn RC, Yi X, Swanson WJ, MacCoss MJ, Nachman MW. Proteomics and comparative genomic investigations reveal heterogeneity in evolutionary rate of male reproductive proteins in mice (Mus domesticus). Mol Biol Evol. 2009;26(8):1733-1743.
19. Ellegren H, Parsch J. The evolution of sex-biased genes and sex-biased gene expression. Nat Rev Genet. 2007;8(9):689-698.
20. Karn RC, Clark NL, Nguyen ED, Swanson WJ. Adaptive evolution in rodent seminal vesicle secretion proteins. Mol Biol Evol. 2008;25(11):2301-2310.

21. Lyon JD, Vacquier VD. Interspecies chimeric sperm lysins identify regions mediating species-specific recognition of the abalone egg vitelline envelope. Dev Biol. 1999;214(1):151-159.
22. Orr HA. The probability of parallel evolution. Evolution. 2005;59(1):216-220.
23. Coyne JA, Charlesworth B. Genetics of a pheromonal difference affecting sexual isolation between Drosophila mauritiana and D. sechellia. Genetics. 1997;145(4):1015-1030.
24. Laukaitis CM, Critser ES, Karn RC. Salivary androgen-binding protein (ABP) mediates sexual isolation in Mus musculus. Evolution. 1997;51(6):2000-2005.
25. Laukaitis CM, Heger A, Blakley TD, Munclinger P, Ponting CP, Karn RC. Rapid bursts of androgen-binding protein (*Abp*) gene duplication occurred independently in diverse mammals. BMC Evol Biol. 2008;8:46.
26. Laukaitis C, Karn RC: Recognition of subspecies status mediated by androgen-binding protein (ABP) in the evolution of incipient reinforcement on the European house mouse hybrid zone. In: Evolution of the House Mouse. Edited by Macholan M, Munclinger P, Baird SJ, Pialek J; 2012.
27. Dlouhy SR, Karn RC. 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. 1983;21(11-12):1057-1070.
28. Dlouhy SR, Taylor BA, Karn RC. The genes for mouse salivary androgen-binding protein (ABP) subunits alpha and gamma are located on chromosome 7. Genetics. 1987;115(3):535-543.
29. Callebaut I, Poupon A, Bally R, Demaret JP, Housset D, Delettre J, Hossenlopp P, Mornon JP. The uteroglobin fold. Ann N Y Acad Sci. 2000;923:90-112.
30. Karn RC, Laukaitis CM. Characterization of two forms of mouse salivary androgen-binding protein (ABP): implications for evolutionary relationships and ligand-binding function. Biochemistry. 2003;42(23):7162-7170.
31. Karn RC. Steroid binding by mouse salivary proteins. Biochem Genet. 1998;36(3-4):105-117.
32. Karn RC, Clements MA. 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. 1999;37(5-6):187-199.
33. Karn RC, Dlouhy SR. Salivary androgen-binding protein variation in Mus and other rodents. J Hered. 1991;82(6):453-458.
34. Vandewege MW, Phillips CJ, Wickliffe JK, Hoffmann FG. Evolution of the ABPA Subunit of Androgen-Binding Protein Expressed in the Submaxillary Glands in New and Old World Rodent Taxa. J Mol Evol. 2013;76(5):324-331.
35. Karn RC, Russell R. The amino acid sequence of the alpha subunit of mouse salivary androgen-binding protein (ABP), with a comparison to the partial sequence of the beta subunit and to other ligand-binding proteins. Biochem Genet. 1993;31(7-8):307-319.
36. Hwang JM, Hofstetter JR, Bonhomme F, Karn RC. The microevolution of mouse salivary androgen-binding protein (ABP) paralleled subspeciation of Mus musculus. J Hered. 1997;88(2):93-97.
37. Talley HM, Laukaitis CM, Karn RC. Female preference for male saliva: implications for sexual isolation of Mus musculus subspecies. Evolution. 2001;55(3):631-634.
38. Bímová B, Karn RC, Pialek J. 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. 2005;84(3):349-361.

39. Vošlajerová Bímová B, Macholán M, Baird SEB, Munclinger P, Laukaitis CM, Karn RC, Luzynski K, Tucker P, Piálek J. Reinforcement selection acting on the European house mouse hybrid zone. Molecular Ecology. 2011;20:2403-2424.
40. Abbott R, Albach D, Ansell S, Arntzen JW, Baird SJ, Bierne N, Boughman J, Brelsford A, Buerkle CA, Buggs R et al. Hybridization and speciation. J Evol Biol. 2013;26(2):229-246.
41. Howard D: Reinforcement: the origin, dynamics, and fate of an evolutionary hypothesis. In: *Hybrid zones and the evolutionary process.* Edited by Harrison R. New York: Oxford University Press; 1993: 46-69.
42. Boursot P, Auffray J-C, Britton-Davidian J, Bonhomme F. The evolution of house mice. Annu Rev Ecol Syst. 1993;24:119-152.
43. Macholán M, Kryštufek B, Vohralík V. The location of the Mus musculus/M. domesticus hybrid zone in the Balkans: clues from morphology. Acta Theriol. 2003;48:177-188.
44. Sage RD, Atchley WR, Capanna E. House mice as models in systematic biology. Syst Biol. 1993;42:523-561.
45. Butlin RK. Reinforcement: an idea evolving. Trends Ecol Evol. 1995;10:432-434.
46. Lemmon AR, Smadja C, Kirkpatrick M. Reproductive character displacement is not the only possible outcome of reinforcement. J Evol Biol. 2004;17(1):177-183.
47. Kimoto H, Haga S, Sato K, Touhara K. Sex-specific peptides from exocrine glands stimulate mouse vomeronasal sensory neurons. Nature. 2005;437(7060):898-901.
48. Kimoto H, Sato K, Nodari F, Haga S, Holy TE, Touhara K. Sex- and strain-specific expression and vomeronasal activity of mouse ESP family peptides. Curr Biol. 2007;17(21):1879-1884.
49. Yoshinaga S, Sato T, Hirakane M, Esaki K, Hamaguchi T, Haga-Yamanaka S, Tsunoda M, Kimoto H, Shimada I, Touhara K et al. Structure of the Mouse Sex Peptide Pheromone ESP1 Reveals a Molecular Basis for Specific Binding to the Class C G-protein-coupled Vomeronasal Receptor. J Biol Chem. 2013;288(22):16064-16072.
50. Kimoto H, Touhara K. Induction of c-Fos expression in mouse vomeronasal neurons by sex-specific non-volatile pheromone(s). Chem Senses. 2005;30 Suppl 1:i146-147.
51. Haga S, Hattori T, Sato T, Sato K, Matsuda S, Kobayakawa R, Sakano H, Yoshihara Y, Kikusui T, Touhara K. The male mouse pheromone ESP1 enhances female sexual receptive behaviour through a specific vomeronasal receptor. Nature. 2010;466(7302):118-122.
52. Haga S, Kimoto H, Touhara K. Molecular characterization of vomeronasal sensory neurons responding to a male-specific peptide in tear fluid: sexual communication in mice. Pure Appl Chem. 2007;79:775–784.
53. Touhara K, Vosshall LB. Sensing odorants and pheromones with chemosensory receptors. Annual review of physiology. 2009;71:307-332.
54. Hurst JL. Female recognition and assessment of males through scent. Behav Brain Res. 2009;200(2):295-303.
55. Logan DW, Marton TF, Stowers L. Species specificity in major urinary proteins by parallel evolution. PLoS One. 2008;3(9):e3280.
56. Mudge JM, Armstrong SD, McLaren K, Beynon RJ, Hurst JL, Nicholson C, Robertson DH, Wilming LG, Harrow JL. Dynamic instability of the major urinary protein gene family revealed by genomic and phenotypic comparisons between C57 and 129 strain mice. Genome Biol. 2008;9(5):R91.

57. Armstrong SD, Robertson DH, Cheetham SA, Hurst JL, Beynon RJ. Structural and functional differences in isoforms of mouse major urinary proteins: a male-specific protein that preferentially binds a male pheromone. Biochem J. 2005;391(Pt 2):343-350.
58. Beynon RJ, Hurst JL. Multiple roles of major urinary proteins in the house mouse, Mus domesticus. Biochem Soc Trans. 2003;31(Pt 1):142-146.
59. Cheetham SA, Thom MD, Jury F, Ollier WE, Beynon RJ, Hurst JL. The genetic basis of individual-recognition signals in the mouse. Curr Biol. 2007;17(20):1771-1777.
60. Robertson DH, Cox KA, Gaskell SJ, Evershed RP, Beynon RJ. Molecular heterogeneity in the Major Urinary Proteins of the house mouse Mus musculus. Biochem J. 1996;316 (Pt 1):265-272.
61. Chamero P, Marton TF, Logan DW, Flanagan K, Cruz JR, Saghatelian A, Cravatt BF, Stowers L. Identification of protein pheromones that promote aggressive behaviour. Nature. 2007;450(7171):899-902.
62. Stowers L, Holy TE, Meister M, Dulac C, Koentges G. Loss of sex discrimination and male-male aggression in mice deficient for TRP2. Science. 2002;295(5559):1493-1500.
63. Clissold PM, Hainey S, Bishop JO. Messenger RNAs coding for mouse major urinary proteins are differentially induced by testosterone. Biochem Genet. 1984;22(3-4):379-387.
64. Clark AJ, Ghazal P, Bingham RW, Barrett D, Bishop JO. Sequence structures of a mouse major urinary protein gene and pseudogene compared. EMBO J. 1985;4(12):3159-3165.
65. Mucignat-Caretta C, Caretta A, Cavaggioni A. Acceleration of puberty onset in female mice by male urinary proteins. J Physiol. 1995;486 (Pt 2):517-522.
66. Dlouhy SR, Nichols WC, Karn RC. Production of an antibody to mouse salivary androgen binding protein (ABP) and its use in identifying a prostate protein produced by a gene distinct from *Abp*. Biochem Genet. 1986;24(9-10):743-763.
67. Laukaitis CM, Dlouhy SR, Karn RC. The mouse salivary androgen-binding protein (ABP) gene cluster on Chromosomes 7: characterization and evolutionary relationships. Mamm Genome. 2003;14(10):679-691.
68. Emes RD, Riley MC, Laukaitis CM, Goodstadt L, Karn RC, Ponting CP. Comparative evolutionary genomics of androgen-binding protein genes. Genome Res. 2004;14(8):1516-1529.
69. Lupski JR. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. Trends Genet. 1998;14(10):417-422.
70. Stankiewicz P, Lupski JR. Genome architecture, rearrangements and genomic disorders. Trends Genet. 2002;18(2):74-82.
71. Shaffer LG, Lupski JR. Molecular mechanisms for constitutional chromosomal rearrangements in humans. Annu Rev Genet. 2000;34:297-329.
72. Hahn MW, Demuth JP, Han SG. Accelerated rate of gene gain and loss in primates. Genetics. 2007;177(3):1941-1949.
73. Arnheim N: Concerted evolution of multigene families. In: Evolution of Genes and Proteins. Edited by Nei M, RK K. Sunderland: Sinauer. 1983;38–61.
74. Arnheim N, Calabrese P, Tiemann-Boege I. Mammalian meiotic recombination hot spots. Annu Rev Genet. 2007;41:369-399.
75. Hurst GD, Werren JH. The role of selfish genetic elements in eukaryotic evolution. Nat Rev Genet. 2001;2(8):597-606.
76. Witherspoon DJ, Watkins WS, Zhang Y, Xing J, Tolpinrud WL, Hedges DJ, Batzer MA, Jorde LB. Alu repeats increase local recombination rates. BMC Genomics. 2009;10:530.

77. Yang S, Arguello JR, Li X, Ding Y, Zhou Q, Chen Y, Zhang Y, Zhao R, Brunet F, Peng L et al. Repetitive element-mediated recombination as a mechanism for new gene origination in Drosophila. PLoS Genet. 2008;4(1):e3.
78. Jurka J, Kohany O, Pavlicek A, Kapitonov VV, Jurka MV. Clustering, duplication and chromosomal distribution of mouse SINE retrotransposons. Cytogenet Genome Res. 2005;110(1-4):117-123.
79. Paigen K, Petkov P. Mammalian recombination hot spots: properties, control and evolution. Nat Rev Genet. 2010;11(3):221-233.
80. Paigen K, Szatkiewicz JP, Sawyer K, Leahy N, Parvanov ED, Ng SH, Graber JH, Broman KW, Petkov PM. The recombinational anatomy of a mouse chromosome. PLoS Genet. 2008;4(7):e1000119.
81. Shifman S, Bell JT, Copley RR, Taylor MS, Williams RW, Mott R, Flint J. A high-resolution single nucleotide polymorphism genetic map of the mouse genome. PLoS Biol. 2006;4(12):e395.
82. Smagulova F, Gregoretti IV, Brick K, Khil P, Camerini-Otero RD, Petukhova GV. Genome-wide analysis reveals novel molecular features of mouse recombination hotspots. Nature. 2011;472(7343):375-378.
83. Song M, Boissinot S. Selection against LINE-1 retrotransposons results principally from their ability to mediate ectopic recombination. Gene. 2007;390(1-2):206-213.
84. Bailey JA, Church DM, Ventura M, Rocchi M, Eichler EE. Analysis of segmental duplications and genome assembly in the mouse. Genome Res. 2004;14(5):789-801.
85. Bailey JA, Liu G, Eichler EE. An Alu transposition model for the origin and expansion of human segmental duplications. Am J Hum Genet. 2003;73(4):823-834.
86. She X, Cheng Z, Zollner S, Church DM, Eichler EE. Mouse segmental duplication and copy number variation. Nat Genet. 2008;40(7):909-914.
87. Zhou Y, Mishra B. Quantifying the mechanisms for segmental duplications in mammalian genomes by statistical analysis and modeling. Proc Natl Acad Sci U S A. 2005;102(11):4051-4056.
88. Guo X, Freyer L, Morrow B, Zheng D. Characterization of the past and current duplication activities in the human 22q11.2 region. BMC Genomics. 2011;12:71.
89. Janoušek V, Karn RC, Laukaitis CM. The evolutionary histories of three rodent pheromone gene families II: The role of LINE1 repeat elements in their gene region expansions. PLoS Genet. 2012:Under review.
90. Katju V, Lynch M. The structure and early evolution of recently arisen gene duplicates in the Caenorhabditis elegans genome. Genetics. 2003;165(4):1793-1803.
91. Chen JM, Chuzhanova N, Stenson PD, Ferec C, Cooper DN. Meta-analysis of gross insertions causing human genetic disease: novel mutational mechanisms and the role of replication slippage. Hum Mutat. 2005;25(2):207-221.
92. Achaz G, Coissac E, Viari A, Netter P. Analysis of intrachromosomal duplications in yeast Saccharomyces cerevisiae: a possible model for their origin. Mol Biol Evol. 2000;17(8):1268-1275.
93. Achaz G, Netter P, Coissac E. Study of intrachromosomal duplications among the eukaryote genomes. Mol Biol Evol. 2001;18(12):2280-2288.
94. Kondrashov FA, Kondrashov AS. Role of selection in fixation of gene duplications. J Theor Biol. 2006;239(2):141-151.
95. Tuzun E, Bailey JA, Eichler EE. Recent segmental duplications in the working draft assembly of the brown Norway rat. Genome Res. 2004;14(4):493-506.
96. Todokoro K, Kioussis D, Weissmann C. Two non-allelic human interferon alpha genes with identical coding regions. EMBO J. 1984;3(8):1809-1812.

97. Slightom JL, Chang LY, Koop BF, Goodman M. Chimpanzee fetal G gamma and A gamma globin gene nucleotide sequences provide further evidence of gene conversions in hominine evolution. Mol Biol Evol. 1985;2(5):370-389.
98. Clark AJ, Chave-Cox A, Ma X, Bishop JO. Analysis of mouse major urinary protein genes: variation between the exonic sequences of group 1 genes and a comparison with an active gene out with group 1 both suggest that gene conversion has occurred between MUP genes. EMBO J. 1985;4(12):3167-3171.
99. Galtier N, Duret L, Glemin S, Ranwez V. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. Trends Genet. 2009;25(1):1-5.
100. Galtier N, Piganeau G, Mouchiroud D, Duret L. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics. 2001;159(2):907-911.
101. Karn RC, Laukaitis CM. The roles of gene duplication, gene conversion and positive selection in rodent *Esp* and *Mup* pheromone gene families with comparison to the *Abp* family. PLoS One. 2012;7(10):e47697.
102. McGrath CL, Casola C, Hahn MW. Minimal effect of ectopic gene conversion among recent duplicates in four mammalian genomes. Genetics. 2009;182(2):615-622.
103. Lawson MJ, Zhang L. Sexy gene conversions: locating gene conversions on the X-chromosome. Nucleic Acids Res. 2009;37(14):4570-4579.
104. Galtier N. Gene conversion drives GC content evolution in mammalian histones. Trends Genet. 2003;19(2):65-68.
105. Kudla G, Helwak A, Lipinski L. Gene conversion and GC-content evolution in mammalian Hsp70. Mol Biol Evol. 2004;21(7):1438-1444.
106. Karn RC, Nachman MW. Reduced nucleotide variability at an androgen-binding protein locus (*Abpa*) in house mice: evidence for positive natural selection. Mol Biol Evol. 1999;16(9):1192-1197.
107. Karn RC, Orth A, Bonhomme F, Boursot P. The complex history of a gene proposed to participate in a sexual isolation mechanism in house mice. Mol Biol Evol. 2002;19(4):462-471.
108. Laukaitis CM, Mauss C, Karn RC. Congenic strain analysis reveals genes that are rapidly evolving components of a prezygotic isolation mechanism mediating incipient reinforcement. PLoS ONE. 2012;7(4):e35898.
109. Karn RC, Laukaitis CM. Positive selection shaped the convergent evolution of independently expanded kallikrein subfamilies expressed in mouse and rat saliva proteomes. PLoS ONE. 2011;6(6):e20979.
110. Yang Z, Wong WS, Nielsen R. Bayes empirical bayes inference of amino acid sites under positive selection. Mol Biol Evol. 2005;22(4):1107-1118.
111. Hayashida H, Miyata T. Unusual evolutionary conservation and frequent DNA segment exchange in class I genes of the major histocompatibility complex. Proc Natl Acad Sci U S A. 1983;80(9):2671-2675.
112. Graur D, Li W-H: Fundamentals of molecular evolution, 2nd edn. Sunderland, MA: Sinauer; 2000.
113. Bergsten J. A review of long-branch attraction. Cladistics. 2005;21(2):163-193.
114. Hsu FF, Nodari F, Kao LF, Fu X, Holekamp TF, Turk J, Holy TE. Structural characterization of sulfated steroids that activate mouse pheromone-sensing neurons. Biochemistry. 2008;47(52):14009-14019.
115. Nodari F, Hsu FF, Fu X, Holekamp TF, Kao LF, Turk J, Holy TE. Sulfated steroids as natural ligands of mouse pheromone-sensing neurons. J Neurosci. 2008;28(25):6407-6418.
116. Dulac C, Wagner S. Genetic analysis of brain circuits underlying pheromone signaling. Annu Rev Genet. 2006;40:449-467.

117. Smadja C, Butlin RK. On the scent of speciation: the chemosensory system and its role in premating isolation. Heredity. 2009;102(1):77-97.
118. Sherborne AL, Thom MD, Paterson S, Jury F, Ollier WE, Stockley P, Beynon RJ, Hurst JL. The genetic basis of inbreeding avoidance in house mice. Curr Biol. 2007;17(23):2061-2066.
119. Thom MD, Stockley P, Jury F, Ollier WE, Beynon RJ, Hurst JL. The direct assessment of genetic heterozygosity through scent in the mouse. Curr Biol. 2008;18(8):619-623

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sciencedomain.org/review-history.php?iid=252&id=3&aid=1878