Comparative Evolutionary Genomics of Androgen-Binding Protein Genes

Richard D. Emes,1,4 Matthew C. Riley,2 Christina M. Laukaitis,3 Leo Goodstadt,1 Robert C. Karn,2 and Chris P. Ponting1,5

1MRC Functional Genetics Unit, Department of Human Anatomy and Genetics, University of Oxford, Oxford OX1 3QX, United Kingdom; 2Department of Biological Sciences, Butler University, Indianapolis, Indiana 46208, USA; 3Internal Medicine Residency Program, St. Vincent Hospital, Indianapolis, Indiana 46260, USA

Allelic variation within the mouse androgen-binding protein (ABP) α subunit gene (Abpa) has been suggested to promote assortative mating and thus prezygotic isolation. This is consistent with the elevated evolutionary rates observed for the Abpa gene, and the Abpb and Abpg genes whose products (ABPβ and ABPγ) form heterodimers with ABPα. We have investigated the mouse sequence that contains the three Abpa/b/g genes, and orthologous regions in rat, human, and chimpanzee genomes. Our studies reveal extensive “remodeling” of this region: Duplication rates of Abpa-like and Abpg-like genes in mouse are >2 orders of magnitude higher than the average rate for all mouse genes; synonymous nucleotide substitution rates are twofold higher; and the Abpg genomic region has expanded nearly threefold since divergence of the rodents. During this time, one in six amino acid sites in ABPβγ-like proteins appear to have been subject to positive selection; these may constitute a site of interaction with receptors or ligands. Greater adaptive variation among Abpg-like sequences than among Abpa-like sequences suggests that assortative mating preferences are more influenced by variation in Abpg-like genes. We propose a role for ABPβ/β/γ proteins as pheromones, or in modulating odorant detection. This would account for the extraordinary adaptive evolution of these genes, and surrounding genomic regions, in murid rodents.

[Supplemental material is available online at www.genome.org.]
suggesting that they have been subject to adaptive evolution and thus might function in reproduction, immunity, or chemosensation.

The ABPA, Abpβ, and Abpγ subunits are representatives of the secretoglobin protein family. These are small, globular, secreted proteins whose phylectic distribution is currently restricted to the mammals and whose functions are still elusive (Karn 1994; Klug et al. 2000; Reynolds et al. 2002; Laukaitis and Karn 2004). In laboratory experiments, female mice have demonstrated a significant mating preference for males with a comparable Abpa genotype (Laukaitis et al. 1997; Talley et al. 2001). A consequence of their rapid evolution and androgen-binding function (Dlouhy and Karn 1983; Karn and Nachman 1999; Karn et al. 2002), ABP has been proposed to act as a signal in a prezygotic isolation mechanism. Studies of wild mouse populations across a transect of the European hybrid zone in Bohemia partially support this idea. Mice captured in populations in the tails of the domesticus-musculus Abpa allelecline show a weak assortative preference in Y-maze studies, and this pattern of preference across the transect is consistent with incipient reinforcement (Bimova et al. 2004). It is not clear, however, that this effect is strong enough to contribute significantly to the maintenance of the zone. This is because Abpa shows a cline of introgression across a Danish transect of the European hybrid zone that is little different from the clines of markers thought not to be under significant selection (Dod et al. 2004).

The availability of human, mouse, and rat genome sequences provided us with an opportunity to investigate several issues that might clarify the evolution and function of these enigmatic molecules. We sought to understand whether Abpa, Abpb, and Abpg genes and their paralogs arose in the rodent lineage, particularly subsequent to the divergence of rat and mouse lineages, and whether primate genomes contain evidence of Abpa, Abpb, and Abpg genes. In addition, we were interested in whether adaptive evolution has acted preferentially at single sites within Abpa, Abpb, and Abpg gene sequences. Our comparative genomics findings demonstrate how strong adaptive evolution has remodeled a megabase genomic region, its genes, and their codons over a period of less than 12–24 million years.

RESULTS

Our first goal was to predict genes encoding secretoglobin family members, and paralogous pseudogenes, within an orthologous region of mouse, rat, and human genomes. This allowed us to compare coding and noncoding sequences. We were then able to address additional goals, including the reconstruction of evolutionary events, and the prediction of evolutionary rates for genes and codons.

Gene Predictions and Nomenclature

Using previously described mouse Abpa-like and Abpg-like gene sequences as templates (Dlouhy et al. 1987; Karn and Laukaitis 2003), we identified nine and five Abpa-like genes and pseudogenes, respectively, and eight and five Abpg-like genes and pseudogenes, respectively, in the mouse genome (Table 1; Supplemental material). Pseudogenes were predicted on the basis of in-frame stop codons or frame shifts; we recognize that pseudogenes with full-length open reading frames are misassigned as functional genes by this method.

The genes for the ABP subunits α, β, and γ expressed in mouse submaxillary gland were originally designated Abpa, Abpb, and Abpg, respectively (Dlouhy et al. 1985). To describe these sequences, we have extended the original nomenclature by numbering the series of α-like and β/γ-like paralogs as Abpa1, Abpa2... Abpa14, and Abpbg1, Abpbg2... Abpbg13, respectively. Numbering is sequential according to the order in which paired genes (see below) are found in the mouse or rat genome. In this system, the original Abpa is referred to as Abpa11, Abpb as Abpbg11, and Abpg as Abpbg10, but we are not proposing to change the previous designsations of Abpa, Abpb, and Abpg. In the appropriate places in this paper, we use both designations (e.g., Abpa11 [Abpa] to facilitate reference to previous and future papers involving these three genes.

Genomic Arrangements of Abpa, Abpb, and Abpg Gene Homologs

Abpa- and Abpbg-like genes and pseudogenes were found to be ordered nonrandomly on the mouse genome (Fig. 1):

- Of the 13 Abpa-like genes or pseudogenes, 11 were found immediately adjacent to an Abpbg-like gene or pseudogene.
- These 11 Abpa-like/Abpbg-like pairs are oriented on opposing strands in a head-to-head (5′-to-5′; bidirectional) manner.
- The pairs’ coding regions are separated by only short sequences (median 7.6 kb).
- The pairs are arranged in two clusters of Abpa-like/Abpbg-like gene pairs with opposite transcriptional orientations: two Abpa-like/Abpbg-like pairs, followed by nine inverted Abpbg-like/Abpa-like pairs.
- Six, of a possible seven, pairs appear to contain both full-length genes, rather than containing pseudogenes.

These observations immediately suggest that transcription, on complementary strands, of Abpa-like and Abpbg-like gene pairs may be coregulated using common, or else overlapping, regulatory elements (Trinklein et al. 2004), thereby facilitating coexpression, or antagonistic expression, of both Abpa-like and Abpbg-like genes. This is consistent with the head-to-head orientation of Abpa (Abpa11) and Abpb (Abpbg11) genes whose products function cooperatively in a heterodimer. Thus we expect that expression of each of the four remaining Abpa-like/Abpbg-like homolog pairs, including that which includes Abpg (Abpbg10) and Abpa10, are also coregulated. Indeed, Abpa2 and Abpbg2 represent paired genes that, according to GenBank entries AAB67069 and AAQ72534, are both expressed in the same tissue, the lacrimal gland.

Using identical methods, we were able to predict only three Abpa-like genes, two Abpbg-like genes, and one Abpg-like pseudogene in the rat genome. Gaps in the rat genome assembly are too small to account for the greatly diminished number of Abpa and Abpbg genes and pseudogenes relative to mouse. Similar to the mouse, these six sequences are arranged in bidirectional Abpa-like/Abpbg-like pairs (Fig. 1).

Finally, as previously (MGSC 2002), we found no apparently functional Abpa or Abpbg genes in the human genome. However, single Abpa-like and Abpbg-like pseudogenes are present in the orthologous genomic region on human Chromosome 19 and chimpanzee Chromosome 20 (Fig. 1). The human and chimpanzee Abpa-like pseudogene contains a single stop codon (Fig. 2), and its 3′-exon could not be identified, and thus may have been deleted from the primate genome. Only the 5′-exon of the human and chimpanzee Abpbg-like pseudogene could be found (Fig. 2). As expected, this is in close proximity (2.9 kb) to the 5′ of the Abpa-like pseudogene, and retains the head-to-head orientation of the gene pairs seen in the two rodents (Fig. 1).

The larger number of Abpa-like and Abpbg-like homologs in mouse, relative to rat and human, is also reflected in a larger...
size of its orthologous region (Fig. 1): mouse ∼1.3 Mb, rat ∼0.5 Mb, and human ∼0.6 Mb. (The large size of the human region, despite its single *Abpa*-like and *Abpbg*-like homologs, is likely due to the presence of multiple KRAB-box zinc finger genes that appear to have infiltrated this region during primate evolution.)

### Table 1. Positions of Genes in the *Abpa/bg* Orthologous Genomic Region Between *Scn1b* and *Uble1b* in the Three Species *Mus musculus*, *Rattus norvegicus*, and *Homo sapiens*

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Start</th>
<th>Finish</th>
<th>Strand*</th>
<th>Pseudogeneb</th>
<th>Median $K_s/K_s$</th>
<th>Celera comparisonc</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mouse Chromosome 7</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Scn1b</em></td>
<td>86240364</td>
<td>86250239</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td><em>Abpa1</em></td>
<td>86406893</td>
<td>86408040</td>
<td>—</td>
<td>1.075</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td><em>Abpbg1</em></td>
<td>86413577</td>
<td>86415531</td>
<td>+</td>
<td>0.908</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td><em>Abpbg2</em></td>
<td>86471441</td>
<td>86472350</td>
<td>—</td>
<td>0.778</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td><em>Abpa2</em></td>
<td>86479128</td>
<td>86480275</td>
<td>+</td>
<td>1.075</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td><em>Abpbg3</em></td>
<td>86615057</td>
<td>86616828</td>
<td>—</td>
<td>0.709</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td><em>Abpa3</em></td>
<td>86624816</td>
<td>86625956</td>
<td>+</td>
<td>1.018</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td><em>Uble1b</em></td>
<td>86684151</td>
<td>86711255</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td><strong>Rat Chromosome 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Scn1b</em></td>
<td>86240364</td>
<td>86250239</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td><em>Abpa1</em></td>
<td>86406893</td>
<td>86408040</td>
<td>—</td>
<td>1.075</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td><em>Abpbg1</em></td>
<td>86413577</td>
<td>86415531</td>
<td>+</td>
<td>0.908</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td><em>Abpbg2</em></td>
<td>86471441</td>
<td>86472350</td>
<td>—</td>
<td>0.778</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td><em>Abpa2</em></td>
<td>86479128</td>
<td>86480275</td>
<td>+</td>
<td>1.075</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td><em>Abpbg3</em></td>
<td>86615057</td>
<td>86616828</td>
<td>—</td>
<td>0.709</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td><em>Abpa3</em></td>
<td>86624816</td>
<td>86625956</td>
<td>+</td>
<td>1.018</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td><em>Uble1b</em></td>
<td>86684151</td>
<td>86711255</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td><strong>Human Chromosome 19</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>UBA2 (Uble1b)</em></td>
<td>39611152</td>
<td>39652635</td>
<td>+</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td><em>SCG8A1P</em></td>
<td>39759479</td>
<td>39760446</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td><em>SCG8A2</em></td>
<td>39776303</td>
<td>39777308</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td><em>SCG8A3P</em></td>
<td>39828407</td>
<td>39828797</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td><em>SCG8A4</em></td>
<td>39848385</td>
<td>39849843</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td><em>Abpbg1</em></td>
<td>39890294</td>
<td>39890500</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td><em>Abpa7</em></td>
<td>39893447</td>
<td>39893862</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td><em>Scn1b</em></td>
<td>40213542</td>
<td>40223192</td>
<td>+</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

*a* Positions and orientation of transcription were determined using the BLAT server at UCSC (http://genome.cse.ucsc.edu).

*b* Pseudogenes were predicted on the basis of in-frame stop codons or frame shifts, and/or missing exons.

*c* Median pairwise $K_s/K_s$ values were calculated using codeml (Yang and Nielsen 2000).

*d* Comparison of genes predicted from C57BL/6J and corresponding genomic DNA from mouse strains 129X1/SvJ, DBA/2J, and A/J from the Celera Genomics subscription database. Gene predictions were compared at the nucleotide level to those of C57BL/6J; *null* sense and sense mutations were predicted by comparing translated cDNA sequences.

*e* Mouse gene coordinates correspond to NCBI build 30 (USCS February 2003 mm3).

*f* Rat gene coordinates correspond to Baylor RGSC v3.1 (USCC June 2003 m3).

*g* Human gene coordinates correspond to NCBI build 34 (UCSC July 2003 hg16).
Comparative Genomics of ABP Genes

Figure 1 A graphical representation of the relative position and transcriptional orientation of the Abpbg-like genes and pseudogenes located on Rattus norvegicus Chromosome 1, Mus musculus Chromosome 7, and Homo sapiens Chromosome 19. Coordinates are taken from genome releases mm3 (Baylor RGSC v3.1), mm3 (NCBI build 30), and hg16 (NCBI build 34), respectively. The 5′-to-3′ orientations of the genes are shown by the direction of the arrowheads. Scn1b and Ube1l1b 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 whereas open arrowheads denote predicted pseudogenes. The sequence of duplication events among rodent genes inferred from phylogenetic trees (see text) is implied from the dendrogram, shown in black. Blue and red lines in this dendrogram represent branches that are not supported by the phylogenetic trees shown in Figure 3 (see text). The scale bar shows approximate genomic distance in megabases. Gaps (≥5 kb) in the genomic assembly of each species are shown as black boxes.

Evolution of Mouse and Rat Abpa-Like and Abpbg-Like Genes

Phylogenetic “5′ trees” were constructed from multiple alignments of nucleotide sequences that lie 5′ upstream to either Abpa-like, or Abpbg-like, sequences (see Methods). These extend far beyond the very short 5′-UTR sequences of these genes and their TATA boxes and transcriptional start sites (Fig. 3A; Laukaitis et al. 2003). Because nongenic regions are less encumbered by selection, as compared with genes (MGSC 2002), these sequences, on average, are more likely to have been subject to neutral, rather than selective, evolution. Phylogenetic trees constructed from them are thus more likely to recapitulate evolutionary events than are trees derived from coding sequences.

The 5′ trees provide detailed insights into the temporal sequence of gene duplications within these mammalian genomic regions. The most parsimonious explanation of the 5′ trees suggests the following sequence of events:

1. The common ancestor of primates and rodents possessed a single gene pair of sequences, one of which was Abpa-like and the other Abpbg-like. This can be inferred from the single (pseudogenic) versions of these genes in human and chimpanzee, and a prediction (see below) that the common ancestor of mouse and rat also only possessed a single gene pair.

2. The mouse Abpa-like and Abpbg-like gene repertoires arose independently from the rat gene repertoires. In the 5′ trees, rat Abpa-like or Abpbg-like sequences form monophyletic groups that exclude mouse sequences (Fig. 3B). Assuming no excision of genes from the rat lineage, this is most parsimoniously explained by single Abpa-like and Abpbg-like genes in the common ancestor of mouse and rat.

3. Each pair of Abpa-like and Abpbg-like genes arose from a single duplication event. Putting aside a single branch in each tree that lacks bootstrap support, the 5′ tree of Abpa-like genes and the 5′ tree of Abpbg-like genes share a common topology (Fig. 3B), when genes paired together on the genomes are considered (see Fig. 1): the two trees share 48,386 out of 49,140 quartets (i.e., 98.5% similarity in quartet distance; Estabrook et al. 1985). Duplication of gene pairs as a single unit appears to have been the dominant mode of duplication in mouse and rat, although five unpaired “singleton” Abpa-like or Abpbg-like sequences are also present in the mouse genome. Some of these may yet be found to be paired when gaps in the genome assembly are filled (Fig. 1).

4. Duplication of a “parent” pair of Abpa- and Abpbg-like genes always resulted in two “daughter” pairs that are juxtaposed on the resulting genome. When the common topology of the phylogenetic trees (Fig. 3B) is superimposed on the gene orders along the two rodent genomes, this never results in a crossing-over of branches (Fig. 1). This implies that “daughter” gene pairs were never interspersed among other gene pairs, but were duplicated side-by-side.

5. Only a single inversion of mouse genes occurred in this region. Abpbg1/Abpbg2 and Abpa1/Abpa2 gene pairs are inverted with respect to all other pairs, yet these form a single monophyletic group. Thus either this single ancestral gene pair was subjected to an inversion event, or else the genomic region encompassing both extant pairs was inverted. (It is also possible that the inversion represents an artifact of the mouse genome assembly.)

Evolution of Other Rodent Abpa-Like and Abpbg-Like Genes

We also constructed trees from amino acid alignments, and from Ks values, including Abpa-like sequences from species of Apodemus (Wickliffe et al. 2002), which are more closely related to the mouse than they are to the rat (Lundrigan et al. 2002). These trees all show that the Apodemus sequences cluster separately from the mouse and rat clusters of Abpa-like sequences (data not shown). Thus, expansions of Abpa-like genes occurred independently for each of the three Mus, Apodemus, and Rattus genera.

Genome sequence data were also available from several mouse strains other than C57BL/6J (129X1/SvJ, DBA/2J, and A/J mouse strains; Mural et al. 2002). These data were investigated to determine whether Abpa-like and Abpbg-like genes are heterogeneous in number and sequence among different laboratory strains. Using identical methods, we identified each of the C57BL/6J 14 Abpa-like and 13 Abpbg-like (pseudo)genes in these other strains. In no case did we find a C57BL/6J gene to be a pseudogene in the other strains (and vice versa). These genes only differ by single-nucleotide substitutions in seven genes (six missense and one sense substitutions). The protein-coding sequences of the remaining 20 genes were identical at the nucleotide level in these strains. The six missense mutations are at sites distinct from predicted ω′ sites (see below).
One Abpa-like and two Abpbg-like genes (and two Abpa-like and two Abpbg-like pseudogenes) were additionally identified that are absent from the current C57BL/6J genome assembly. However, these may yet be found to lie within gaps in the current C57BL/6J assembly (Fig. 1). Indeed, identical sequences to six of these seven genes or pseudogenes are present in unassembled C57BL/6J sequence (data not shown).

Primate Abpa-Like and Abpbg-Like Pseudogenes and Secretoglobin Evolution

S' to the human Abpa- and Abpbg-like pseudogenes on human Chromosome 19, we identified three additional genes and one pseudogene; these lie within the orthologous Scn1b-Uble1p region (Fig. 1). These are secretoglobin homologs: for example, a search of the Pfam database (Bateman et al. 2004) with the conceptual translation of the 5'-most of these genes reveals significant sequence similarity to uteroglobins (E = 2.5 × 10^{-5}). We name these genes SCGB4A1–4(P), following the proposals of the Secretoglobin Nomenclature Committee (Klug et al. 2000). SCGB4A1P and SCGB4A3P appear to be pseudogenes: the former’s exon 1 has been translocated to lie 3’ to exons 2 and 3, whereas the latter contains an in-frame stop codon (Fig. 2). These four sequences appear to have arisen from two recent duplications because SCGB4A1P and SCGB4A4 are highly sequence-similar, as are SCGB4A2 and SCGB4A3P. Thus, it is possible that the sequence-dissimilar gene pair SCGB4A2 and SCGB4A4 encodes protein subunits of a heterodimer, similar to that seen for mouse ABPβ and β or γ, and for cat fed EL. But all SCGB4A1P are present in the current chimpanzee genome assembly (Fig. 1).

From EST data, it appears that SCGB4A2 is expressed in the eye and ovarian cancerous tissue (GenBank accessions BU738523, BM716941, A1073890, A1472332, A1821523, A1821558, BX105421, AA290868, AA481857, AA290985, and AA291047). Interestingly, SCGB4A1P, a likely pseudogene, has corresponding ESTs isolated from lung, multiple sclerosis lesions, infant brain, and in head and neck tissues (GenBank accessions BM984683, CD173361, NS7568, T16687, B1570711, respectively). In addition to these data, we assayed cDNAs prepared from a variety of human tissues. Only SCGB4A2 was successfully amplified and was isolated from pancreas and spleen cDNAs (data not shown).

Abpa-Like and Abpbg-Like Pseudogenes

The set of mouse Abpa, Abpb, and Abpg homologs contains a surprisingly high proportion (37%) of predicted pseudogenes. Pseudogene formation appears to have occurred following acquisition of frame shifts, exon deletions, or stop codons, or the substitution of critical residues such as an initiating methionine (in mouse Abpbg9), or structurally important cysteine residues. In all 10 of these cases, disruptions to the open reading frames of mouse Abpa-like or Abpbg-like proteins appear to have occurred independently, because no two disruptions coincide in the alignment (Fig. 2). This implies that when duplication within the mouse genome occurred, it always involved functional Abpa-like and Abpbg-like genes, rather than pseudogenes. This would not have been expected if an unusually high duplication rate within this region was responsible for the elevated number of observed gene duplicates. Rather, it is likely that fixation of functional paired gene duplicates is the outcome of recurrent episodes of positive selection where selective advantage has accrued to individuals with relatively rare gene duplications.

Evolutionary Rates

Using codeml, we calculated the pairwise Ks/Ka values among mouse and rat Abpa-like or among Abpbg-like genes (Table 1; Supplemental Table 1). The medians of these values are unusually high: 0.99 for Abpa-like homologs, and 0.86 for Abpbg-like homologs. These values are similar to those calculated previously (Karn and Nachman 1999; Karn et al. 2002; MGSC 2002; Karn and Laukaitis 2003) and indicate that these genes have experienced relaxed selective constraints and/or adaptive evolution.

Median values of Ks calculated for mouse Abpa-like and Abpbg-like pairwise comparisons were 0.27 and 0.53, respectively. These values are significantly higher than the median Ks value (0.197) between mouse and rat ortholog pairs used in the initial analysis of the rat genome (RGSC 2004). Only 16.4% and 0.3% of mouse-rat ortholog pairs are characterized by Ks values >0.27 and >0.53, respectively.

Because we believe that these mouse genes arose by duplication after the mouse-rat divergence, the elevated neutral substitution rates imply that mouse Abpa-like and Abpbg-like genes have suffered an unusually high mutation rate relative to other genes. Similar conclusions were reached for rat Abpa-like and Abpbg-like genes (data not shown).

We were able to exclude the possibility that these Ks increases are simply due to abnormal G+C content or values of ω, the ratio of transitions to transversions (data not shown). Moreover, these are not due to alignment inaccuracies because the low frequencies of insertion/deletion positions (Fig. 2) provide little margin of error in the alignment process. Instead, a combination of two evolutionary processes may have led to the elevation of these Ks values. First, the genomic region in which the Abpa-like and Abpbg-like genes lie may have been unusually hypermutable, with inflated mutation rates relative to the rest of the genome. Second, substitutions at synonymous sites may have been elevated due to mutational correlation of adjacent bases (Bains 1992) such as methylated CG dinucleotides.

We attempted to quantify the contributions of these two evolutionary processes by examining the neutral substitution rate within the nucleotide sequences lying 5′ upstream of either Abpa-like or Abpbg-like sequences; these may be assumed to be free from natural selection. The median number of substitutions per site between pairs of Abpa-like sequences was 0.251, and that
for Abpbg-like sequences was 0.216; these calculations only compared a mouse with a rat sequence to provide nucleotide distances for orthologous sequences. These values are only slightly elevated when compared with an average of 0.174 substitutions per neutral site found in a rat–mouse genome-wide comparison (Fig. 5B, below; 0.083 + 0.091; RGSPC 2004). This appears to rule out hyper-mutability of this genomic region as an explanation of the significant elevations of $K_S$ values.

**Concerted Evolution**

We considered whether the apparent monophyly of mouse Abpa-like or Abpbg-like genes (Fig. 3B) might instead be caused by the effects of concerted evolution (Li 1997). On one hand, the unusually high $K_S$ values observed between mouse and rat orthologs argue against concerted evolution. This is because concerted evolution acts to homogenize sequences, and low, rather than high, $K_S$ values would thus be expected. Furthermore, concerted evo-
Figure 4  Multiple nucleotide sequence alignment of mouse and rat Abpbg-like exons 3 and surrounding genomic DNA. Genomic DNA corresponding to exon 3 (98 positions) and 100 nucleotide positions of both flanking intronic and 3’-UTR sequence was aligned with HMMER, and manually adjusted. We found that 81.3%, 50.5%, and 92.6% of the sites in the intron, exon, and 3’-UTR, respectively, exhibited ≥70% consensus. In these calculations, positions with fewer than 50% gaps were considered. The 14 codons of exon 3 corresponding to predicted ω* sites are shown by horizontal bars.
lution usually is observed only in coding regions (e.g., Shibata and Yamazaki 1995), and the monophyly of mouse Abpa-like or Abpbg-like genes was inferred from 5' flanking, rather than coding, sequences (Fig. 3). On the other hand, the equivalence of inactivating frameshift mutations in mouse Abpbg6 and rat Abpbg3 pseudogenes (Fig. 2) suggests that concerted evolution might be obscuring the orthology relationship between these two sequences. The true evolutionary history of these genes thus appears to have involved widespread gene duplication and coding sequence diversification, with perhaps more limited episodes of concerted evolution within coding sequence.

**Secretoglobin Protein Structure**

The detection of multiple Abpa-like and Abpbg-like genes in rodents yielded an opportunity to compare their sequences with the homologous chains of the cat allergen Fel dl (chains 1 and 2, respectively), whose protein tertiary structure has been determined (Kaiser et al. 2003). ABP-like sequences are closely related to cat Fel dl chain 1, whereas ABPβ-like sequences are closely related to Fel dl chain 2. Thus, the heterodimeric structure of ABPαβ and ABPβα is recapitulated by the sequence similar Fel dl chains 1 and 2. Conservation of primary and quaternary structure indicates that the genome of the eutherian common ancestor of cats, rodents, and primates contained a similar gene pair.

A multiple alignment of ABPαβ-like conceptual amino acid sequences from mouse, rat, human, and chimpanzee genes (Fig. 2) demonstrates that these share the same four α-helix secondary structure that has been observed in structures of rabbit urotoglobin and cat Fel dl (Mornon et al. 1980; Kaiser et al. 2003). They also share the three cysteine residues that form single intrachain and interchain disulfide bonds in the Fel dl heterodimer. In addition, the Abpbg-like genes encode C-terminal extensions that are predicted to form a fifth α-helical structure, which is present in the ABPβγ-like Fel dl chain 2 but absent from other secretoglobins, such as Fel dl chain 1 and urotoglobin (Karn and Laukaitis 2003). These structural observations assisted the interpretation of evolutionary rate findings described below.

**Positive Selection of Abpa-Like and Abpbg-Like Genes**

By comparing the gene sequences of mouse and rat Abpa-like or Abpbg-like homologs, it is evident that their exons are changing more rapidly than are their introns and 3’ untranslated regions (UTRs; Fig. 4). This is a hallmark of positive selection, rather than neutral or purifying evolution. We compared the conservation of nucleotide sequence within exon 3 with that within 100 alignment positions 5’ upstream in intron 2, and 100 positions downstream in the 3’ UTR for Abpbg-like homologs (Fig. 4); a similar comparison has been recently reported for Abpb and Abpbg (Laukaitis et al. 2003). The intron and 3’ UTR sequences were highly conserved between mouse and rat homologs: 81.3% and 92.6% of sites, respectively, contain a nucleotide that is conserved in at least 70% of sequences. These are significantly higher levels than conservation levels in exon 3, where the corresponding figure is only 50.5% (Fig. 4).

Next we used codeml to predict codons ("ω" sites) that have been subjected to positive selection. Five and 28 such ω+ sites were identified for mouse Abpa-like or Abpbg-like homologs, respectively (Fig. 2). These results demonstrate that adaptation has been the dominant evolutionary force during recent divergence of these rodent homologs. We mapped these ω+ sites to the crystal structure of the cat Fel dl heterodimer (Fig. 5). Although predicted ω+ sites appear to be relatively uniformly distributed along the protein sequence, a clear pattern is revealed when the sites are examined in relation to the predicted three-dimensional structure: they are overrepresented among exposed amino acids (Table 2), and they form a pronounced cluster on one face of the heterodimer (Fig. 5). Thus, the majority of these ω+ sites are available to participate in binding interactions. Moreover, they are entirely absent from the dimer’s internal cavity (Fig. 5) and are distinct from sites in urotoglobin thought to bind internal ligands (Callebaut et al. 2000; Kaiser et al. 2003; Karn and Laukaitis 2003).

Several secretoglobins are known to bind hydrophobic ligands within an interior cavity formed at the interface of two homologous chains (Beato and Baier 1975; Parker et al. 1982; Dluhy and Karn 1983; Callebaut et al. 2000), and the cat Fel dl crystal structure contains strong electron density within this cavity, indicating the presence of a bound ligand of unknown type (Kaiser et al. 2003). Our results suggest that positive selection has not acted strongly on sites that confer specificity on binding internal ligands. Rather, it has constantly remodeled a single face of the heterodimer that consists of all but one of the five α helices of ABPβγ-like homologs.

**DISCUSSION**

Our comparative genomics findings demonstrate that the mouse genomic sequence containing Abpa-like and Abpbg-like homologs has been subject to an exceptional degree of change over the relatively short interval of time (12–24 million years; Adkins et al. 2001; Springer et al. 2003) since the divergence of mouse and rat lineages. Rates of gene duplication, pseudogene formation, and synonymous and nonsynonymous nucleotide substitution are all significantly higher than for most other rodent genes. Our findings show that this genomic region has been subject to extensive and sustained incidents of positive selection.

Phylogenetic tree (Fig. 3B) and gene order (Fig. 1) data strongly indicate that the last common ancestor of rats and mice possessed only a single Abpa- and Abpbg-like gene pair. Thus the past 12–24 million years has experienced 14-fold and 13-fold expansions of the Abpa- and Abpbg-like mouse gene (and pseudogene) repertoires, respectively. Concomitantly, the mouse Scn1b–Ubl1b orthology region has been enlarged several fold. Duplications are likely to have been of genes rather than pseudogenes, often with a subsequent loss of function and acquisition of reading frame disruptions and truncations. In rat, expansions have been more modest, resulting in only three genes and pseudogenes of each type.

We note that the rapidity of gene duplication and pseudogene formation inferred for these families is exceedingly rare. On average, only 3% of extant mouse genes arose from duplication since the last common ancestor with rat, and the rat and mouse genome sizes differ by only ~6% (RGSPC 2004). Although the Scn1b–Ubl1b regions in mouse and rat share a common evolutionary origin, they have been subject to a rare and extensive remodeling in relatively recent times.

Rapid evolutionary change was also observed within coding sequences. Per-gene median $K_s/K_o$ values for Abpa-like and Abpbg-like genes are close to 1, which might indicate either neutral or adaptive evolution. In contrast, site-specific evolutionary rate estimates indicated the past action of positive selection among both Abpa-like and Abpbg-like genes: five and 28 codons, respectively, were identified as having experienced positive selection. The low number of such ω+ sites in Abpa-like, relative to Abpbg-like sequences may indicate that the former may include several nonfunctional pseudogenes, albeit with full-length open reading frames, that are evolving neutrally.

Loss of selective constraints on a gene eventually leads to
pseudogene formation. Relatively few rodent genes have single pseudogenic counterparts in the human (C.P. Ponting and C. Webber, unpubl.). Those that do, such as genes for olfactory receptors and odorant-binding proteins (Emes et al. 2004), predominantly have functions in chemosensation, perhaps reflecting a diminution of the sense of smell in humans (Glusman et al. 2001). The rapid expansion of gene families independently in mouse and rat has also been observed for rodent pheromone genes (Emes et al. 2004). These analogies suggest that the Abpa-like and Abpbg-like genes of the primate/rodent common ancestor may have possessed chemosensation-related functions and that ABP paralogs may act as pheromones.

An intriguing picture of the functional regions of ABP is developing from this and other studies. ABP subunits are probably four- or five-helix bundles with the unusual boomerang shape of uteroglobin/clara proteins (Callebaut et al. 2000; Karn and Laukaitis 2003). As in the case of uteroglobin/clara, the subunits form a hydrophobic cavity at the interface of the two dimers (Callebaut et al. 2000) wherein a ligand having the A-ring structure of testosterone and progesterone (Karn 1998) can be

---

**Figure 5** Site-specific $K_e/K_s$ analysis of Abpa-like and Abpbg-like genes mapped to the tertiary structure of cat Fel dl. $\omega^+$ codons that are predicted to be under positive selection for mouse paralogs are mapped to a ribbon representation of the tertiary structure of feline major allergen Fel dl (PDB 1POU; Kaiser et al. 2003). (A) Ribbon diagram of the Fel dl dimer. Side chains of mapped predicted $\omega^+$ sites are highlighted in red. A single ligand molecule (2-methyl-2,4-pentanediol) present in the crystal structure is colored orange. (B) A representation of the molecular surface of the Fel dl chain 1 and 2 heterodimer. Mapped ABP$\alpha$ $\omega^+$ sites are shown in blue, and ABP$\beta/\gamma$ $\omega^+$ sites are shown in red. (C) Mapped $\omega^+$ sites predicted for the Abpa-like mouse paralogs. (D) Mapped $\omega^+$ sites predicted for the Abpbg-like mouse paralogs. (E) Prediction of the ABP$\beta/\gamma$ terminal extension and the mapped location of identified $\omega^+$ sites. The ABP$\beta/\gamma$ terminal extension is predicted to form a helical structure of ~24 residues and is shown to scale. Swiss-PDBviewer (http://www.expasy.org/spdbv/; Guex et al. 1999) was used for all structural manipulations, and POV-Ray (http://www.povray.org) was used to generate images.
bound. Secretoglobins have six conserved residues (residues F6, L13, V21, F28, M41, and I63) in uteroglobin/clara oriented toward the interior of the cavity. Of these, one (F28) functions in maintenance of the dimer interface, and the other five correspond to the ligand. The alternative combinations of β and γ subunits provide for sequence variation at three positions (corresponding to uteroglobin/clara L13, M42, and I63): this limited variation has been offered as the explanation of different binding affinities to testosterone and DHT (Karn and Nachman 1999; Karn and Laukaitis 2003). Overall, however, these internal residues are highly conserved, in contrast with the exterior residues (Karn and Laukaitis 2003). Overall, however, these internal residues are highly conserved, in contrast with the exterior residues that may help answer the broader question of the functional attributes to assortative mating in mouse populations. If so, the amino acid sequences of known Abpa, Abpb, and Abpg genes (Karn and Laukaitis 2003; Laukaitis et al. 2003) and predicted Abpa-like genes were used to search the genomes of M. musculus strain C57BL/6J (NCBI build 30, mm3), Rattus norvegicus (m3), and Homo sapiens (NCBI build 34, hg16) using the BLAT server at UCSC (Kent 2002). These searches identified orthologous genomic segments in mouse, rat, and human that contain Abpa and Abpg homologs. These segments are flanked by Scn1b and Uble1b genes in all three genomes. To obtain gene predictions, the genomic sequences that intervene between these flanking genes were used as templates for gene prediction (on both strands) using both GeneWise (Birney and Durbin 2000) and hmmssearch (Eddy 1998), each of which uses hidden Markov models (HMMs; Eddy 1995, 1996). GeneWise searches used HMMs constructed from amino acid multiple sequence alignments; hmmssearch used HMMs constructed from nucleotide multiple sequence alignments of known Abpa-like or Abpb-like gene sequences extracted from the genome.

**Prediction of Primate Secretoglobin Genes**

In the human genome, the Scn1b–Uble1b intervening region contains a region aligned, by BLAT at the UCSC genome browser, to cat major allergen Fel d1 mRNA (accession no. M77341), a known secretoglobin. This indicated that secretoglobin genes or pseudogenes are present within this orthologous genomic region. tBLASTn searches of human ESTs using the conceptual translations of cat Fel d1 as a query were used to identify several homologous human ESTs that are mapped to the Scn1b–Uble1b orthologous segment. Predicted genes were mapped to the H. sapiens genome using BLAT. Introns and exons were predicted by requiring consistency between a multiple alignment of genomic sequences and transcript sequences mapped by BLAT to the genome; splice-site consensus sequences were required for genes. BLASTn searches of the November 2003 Pan troglodytes assembly obtained from Ensembl (http://wwwensembl.org) were used to identify potential chimpanzee orthologs of these candidate human genes.

**Nucleotide Sequence Multiple Alignments**

Genomic DNAs corresponding to each of the Abpa-like and Abpb-like genes were obtained from the genome browser at UCSC using positions obtained using BLAT. Among the Abpa-like genes, Abpa11 (Abpa) and Abpa10 exhibited greatest sequence similarity; their genomic DNA sequences were aligned using CLUSTAL W (Thompson et al. 1994). This alignment was used to generate an HMM using HMMer (Eddy 1998), which was then used to align all Abpa genomic DNA sequences. The same procedure was used to align the Abpb-like genes using Abpbg10 (Abpg) and Abpbg11 (Abpb) genes for the starting alignment. The corresponding genomic sequences of 129X1/SvJ, DBA/2, and A/J

---

### Table 2. The Relative Solvent Accessibility of ABPβγω* Sites That Map to the Recombinant Cat Allergen Fel D1 (PDB 1PUO)

<table>
<thead>
<tr>
<th></th>
<th>ABPω*</th>
<th>ABPβγω*</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All sites</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Mouse and rat ω</strong>*</td>
<td>17%</td>
<td>17%</td>
</tr>
<tr>
<td><strong>Mouse only ω</strong>*</td>
<td>29%</td>
<td>29%</td>
</tr>
<tr>
<td><strong>No. of residues</strong></td>
<td>68</td>
<td>67</td>
</tr>
<tr>
<td><strong>No. buried</strong></td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td><strong>No. intermediate</strong></td>
<td>25</td>
<td>33</td>
</tr>
<tr>
<td><strong>No. exposed</strong></td>
<td>42</td>
<td>43</td>
</tr>
<tr>
<td><strong>Buried (%)</strong></td>
<td>7.4</td>
<td>7.9</td>
</tr>
<tr>
<td><strong>Intermediate (%)</strong></td>
<td>32.4</td>
<td>28.4</td>
</tr>
<tr>
<td><strong>Exposed (%)</strong></td>
<td>60.3</td>
<td>62.7</td>
</tr>
</tbody>
</table>

The solvent accessibility values of cat Fel d1 residues equivalent to ABPωβγω*ω* residues were calculated from the structure coordinates using the DSSP algorithm (Kabsch and Sander 1983). Relative accessibility scores were partitioned into three states: buried (<9% relative accessibility), intermediate (9%–35% relative accessibility), and exposed (>36% relative accessibility), as described previously (Rost and Sander 1994).

---

**METHODS**

**Gene Prediction and Annotation**

A large-scale study of the mouse genome previously identified nine candidate Abpa-like genes lying in close proximity on Chromosome 7 (MGSC 2002; Emes et al. 2003). We used the amino acid conceptual translations of these candidate genes, and gene identification and gene building techniques, to identify orthologous genomic regions and homologous mouse, rat, and primate genes.
mouse strains from the Celera Genomics subscription database were also obtained and analyzed using these HMMs. The positions of introns and exons were predicted using evidence from BLAT searches using the predicted protein sequences and also by maintaining conserved intron donor and acceptor splice sites (Zhang 1998).

Amino Acid Sequence Multiple Alignments
Conceptual amino acid sequences were translated from these gene predictions and aligned using CLUSTAL W (Thompson et al. 1994). Conceptual amino acid sequences of pseudogenes were prepared by applying frame shifts, where required, and ignoring any stop codons. All pairs of these gene sequences were extracted from the multiple alignments and their pairwise values of $K_s$ and $K_s$ calculated using codeml, an application from the PAML package (v1.3.1) of phylogenetic software (Yang 1997) with ambiguity positions ignored.

Predictions of Evolutionary Relationships
To overcome problems associated with building trees from rapidly evolving gene sequences, genomic DNA sequences 5’ to the translational start site of each gene were aligned. We assume that these DNA regions have evolved neutrally and thus trees (so-called 5’ trees) constructed from these alignments, may be used to infer evolutionary relationships among Abp paralogous groups. Genomic DNA, the repeat elements present in the intervening genomic DNA of the Abpa and Abpbg gene pairs were first masked using the RepeatMasker Web server (http://ftp.genome.washington.edu/cgi-bin/RepeatMasker). The intervening sequences between Abpa11 and Abpbg11, or Abpa10 and Abpbg10, were then aligned by CLUSTAL W and manually edited to minimize gaps; HMMs were generated from these alignments. All repeat-masked genomic DNA sequences were then aligned using these HMMs and the program hmmalign (Eddy 1998), and the aligned DNA then edited to reduce gap positions.

The 5’ tree for Abpa-like genes used 300-base regions containing no repeat elements. Similarly, the 1-kb regions 5’ to the translational start site of Abpph-like genes was used. Trees were generated using the neighbor joining (NJ) method (Saitou and Nei 1987), where distances correspond to percentage sequence divergence. Bootstrap values were calculated from 1000 trials with gap positions excluded; trees were visualized using Treeview (Page 1996). NJ trees were also generated from protein sequence alignments. Phylogenetic trees with contemporaneous tips based on $K_s$ values from codeml were generated using the Fitch-Margoliash algorithm (Fitch and Margoliash 1967) as modified from the “kitch” program in Phylip (Felsenstein 1989) and 20 iterations. Topological comparisons of the phylogenetic 5’ trees for Abpa-like and Abpbg-like gene pairs were performed by counting gene quartets common to both trees as a proportion of the total number of possible quartets (Estabrook et al. 1985); for this calculation, all paired genes’ branches were considered. The Qdist program (Mailund and Pedersen 2004) was used for this purpose.

Nucleotide Substitution Levels in Noncoding Genomic DNA
The calculation of nucleotide distances between aligned 5’ genonic DNA sequences for mouse and rat Abpa-like and Abpbg-like genes was performed using baseml (Yang 1994). We used the parametric TN93 substitution model (Tamura and Nei 1993), which takes into account different rates of transversions and transitions. For both Abpa-like and Abpbg-like genes, the medians of all distances between mouse sequences and rat sequences were calculated as an indication of the substitution rates since the divergence of the mouse and rat lineages.

Site-Specific $K_s/\omega$ Analysis
Site-specific $K_s/\omega$ analyses were conducted using codeml (Yang et al. 2000). For each analysis, a cDNA sequence alignment was prepared that corresponded to its amino acid alignment. To ensure that the site-specific $K_s/\omega$ analysis was not influenced by sites with a large number of gaps, all columns with gaps in >25% of the sequences were removed from the cDNA alignment. Although the 5’ trees were used as input files for codeml to generate results described in this report, other tree topologies, such as those generated from $K_s$ estimates, gave essentially identical results.

Codeml uses maximum likelihood to predict sites in a group of cDNA sequences that have been subject to positive selection (Zanotto et al. 1999; Bishop et al. 2000; Yang and Bielawski 2000; Yang et al. 2000; Peek et al. 2001; Schaner et al. 2001; Jansa et al. 2003). Log likelihood values (one-ratio; Goldman and Yang 1994) versus M3 (discrete; Yang et al. 2000); M1 (neutral) versus M2 (selection; Nielsen and Yang 1998); and M7 (β versus M8 (β + ω; Yang et al. 2000). As described previously (Emes et al. 2004), only nonconserved alignment positions predicted to be under positive selection with a posterior probability >0.90 by one codeml model, and >0.50 by at least one other model, were highlighted. We have termed these alignment positions “ω” sites. ω sites were mapped to the crystal structure of the major cat allergen Fel d1 (PDB 1PUO; Kaiser et al. 2003), an engineered heterodimer of two secretoglobin molecules. Fel d1 amino acid sequences were aligned to the family alignment using HMMer, and manually adjusted to minimize insertions and deletions within secondary structure elements. Swiss-PDBviewer (http://www.expasy.org/spdbv/; Guex et al. 1999) was used for structural manipulations, and POVRAY (http://www.povray.org) was used to render images.

Solvent accessibility values of Fel d1 amino acids that align to Abpa, β, and γ ω* residues were calculated using structural information from PDB (http://www.rcsb.org/pdb/) and the DSSP algorithm (Kabsch and Sander 1983); ligands were removed prior to accessibility score estimation. Relative accessibility scores were produced by normalizing these scores by amino-acid-specific maximal accessibility values. These relative scores were partitioned into three states: buried (<9% relative accessibility), intermediate (9%–35% relative accessibility), and exposed (>36% relative accessibility), as previously (Rost and Sander 1994).

Primate-Specific Secretoglobin Gene Expression
The expression profiles of human secretoglobin homologs were obtained using the EST table from the genome browser at UCSC (http://genome.cse.ucsc.edu/). To obtain further expression information, gene-specific primers were designed and used to assay a human multiple tissue cDNA panel (Clontech multiple tissue cDNA panels 1 and 2, product numbers K1420-1 and K1421-1) by PCR. Gene-specific primers for each of the genes were used to amplify 1 µl of a 2× diluted stock of each cDNA under the following cycle conditions: 30 sec at 95°C, five cycles of 3 min at 72°C; 3 min at 95°C, five cycles of 3 min at 70°C; 3 min at 95°C; and 1 min at 94°C. Amplified products were separated by electrophoresis in 2% agarose gels and purified (Minelute, gel extraction; QIAGEN). Then 1 µl of eluted cDNA was ligated to PST-Iblue vector and sequenced on both strands using vector primers.

Acknowledgments
We thank Peter Oliver and Ponting group members for helpful discussions. This work was funded by the UK Medical Research Council, and by the Holcomb Research Institute, Butler Univer-
sity. M.C.R. was supported for his work at Oxford University by a Seitz Scholarship, Butler University. That support is gratefully acknowledged.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES


a Seitz Scholarship, Butler University. That support is gratefully

sity. M.C.R. was supported for his work at Oxford University by

Emes et al.


Nucleic Acids Res. 32: D138-D141.


WEB SITE REFERENCES
http://www.ensembl.org/; Ensembl genome browser.
http://www.rcsb.org/pdb/; the protein data bank.
http://www.ensembl.org/; Ensembl genome browser.
http://www.rcsb.org/pdb/; the protein data bank.