Available online at www.sciencedirect.com

SCIENCE DIRECT®

Genomics 83 (2004) 361–372

GENOMICS

www.elsevier.com/locate/ygeno

The canine olfactory subgenome☆

Tsviya Olender,[a,1] Tania Fuchs,[a,1] Chaim Linhart,[b] Ron Shamir,[b] Mark Adams,[c] Francis Kalush,[c] Miriam Khen,[a,1] and Doron Lancet[a,*]

[a] *Department of Molecular Genetics and the Crown Human Genome Center, The Weizmann Institute of Science, Rehovot 76100, Israel*
[b] *School of Computer Science, Sackler Faculty of Exact Sciences, Tel-Aviv University, Tel Aviv 69978, Israel*
[c] *Celera Genomics, 45 West Gude Drive, Rockville, MD 20850, USA*

Received 30 June 2003; accepted 11 August 2003

Abstract

We identified 971 olfactory receptor (OR) genes in the dog genome, estimated to constitute ~80% of the canine OR repertoire. This was achieved by directed genomic DNA cloning of olfactory sequence tags as well as by mining the Celera canine genome sequences. The dog OR subgenome is estimated to have 12% pseudogenes, suggesting a functional repertoire similar to that of mouse and considerably larger than for humans. No novel OR families were discovered, but as many as 34 gene subfamilies were unique to the dog. "Fish-like" Class I ancient ORs constituted 18% of the repertoire, significantly more than in human and mouse. A set of 122 dog–human–mouse ortholog triplets was identified, with a relatively high fraction of Class I ORs. The elucidation of a large portion of the canine olfactory receptor gene superfamily, with some dog-specific attributes, may help us understand the unique chemosensory capacities of this species.

Keywords: Odorant receptors; Sequence analysis DNA; Sequence homology; Multigene family

Dogs became domesticated 10,000 years ago, and selective breeding has since created hundreds of dog strains that have particular desirable properties [1]. While there is considerable heterogeneity in behavioral traits, all dogs are macrosmatic animals, relying highly on their sense of smell and displaying superb sensitivity and selectivity [2,3]. Yet, the molecular basis of such prominent chemosensory capacities remains largely unknown.

The ability to detect and discriminate millions of odors in vertebrates is mediated by a superfamily of G-protein-coupled olfactory receptor (OR) proteins [4–7]. The complete repertoires of genes coding for these have been recently elucidated in the human [8,9] and mouse [10,11] genomes. Clusters of the ORs, each containing up to 100 intronless coding regions (~1000 bp long), are found on most mammalian chromosomes [12–14]. These have presumably formed by a continuous process of gene and cluster duplication that resulted in a superfamily of >1000 genes. ORs have also been studied in other vertebrate species [15], but their complete olfactory subgenomes remain to be described. Sequence comparisons of human and mouse showed that OR proteins in both mammals are composed of similar families, with clear-cut orthology relations often visible. Certain functional motifs tend to be more conserved in mouse than in human [11].

The most striking difference between the mouse and the human OR repertoires is the number of nonfunctional genes. While only about 20% of the mouse OR arsenal are apparent pseudogenes, the fraction is higher than 60% among human ORs, suggesting a marked reduction of function in humans [8–11]. Otherwise, the mammalian OR repertoires show strong similarity, in particular as regards to the family subdivisions. Both human and mouse ORs are divided into Class I (present in both fish and tetrapods) and Class II (present only in tetrapods). Further subdivisions are 17 families (members of which show >40% amino acid sequence identity) and more than 300 subfamilies (each including proteins with >60% amino acid sequence identity) [8,15]. Others have proposed alternative ways to classify the OR gene repertoire [9,10].

Several authors have studied canine OR genes as well as cDNAs expressed in olfactory epithelium and testis. The

☆ Supplementary material is available online at http://bioinformatics.weizmann.ac.il/HORDE/lancetpub/genomics2003.

* Corresponding author. Fax: +972-8-9344112.
E-mail address: doron.lancet@weizmann.ac.il (D. Lancet).

[1] These authors contributed equally to this article.

0888-7543/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2003.08.009

first relevant study reported the cloning and sequencing of 11 OR genes from cDNA of both tissues [16]. Later, 11 additional dog OR genes were isolated from testis cDNA [17]. Using probes related to 4 OR genes, a genomic Southern blot analysis was carried out in 26 dog breeds [18]. This revealed minor differences among breeds and a pronounced conservation of the apparent OR repertoire size.

Canine genome mapping has been a center of interest [19], and an integrated map of the canine genome, incorporating detailed cytogenetic, radiation hybrid, and meiotic information, has been recently published [20]. A search in the adjoining databases (http://www-recomgen.univ-rennes1.fr/doggy.html) revealed only five ORs among the 320 canine genes mapped. Recently, a survey of canine expressed sequence tags (ESTs) has yielded 8000 ESTs from several tissues, not including olfactory epithelium [21]. This collection does not contain any OR sequences.

Thus, the genomic analysis of the canine olfactory receptor subgenome has remained limited. We therefore launched an effort to obtain a more complete knowledge of OR genes in this species. Celera Genomics has carried out a large-scale canine genome sequencing project and achieved X 1.3 coverage, without assembly. We report here data mining of this canine genome repository, enhanced by experimentally obtained OR sequences resulting from whole-genome PCR amplification and subcloning. This led to the identification of a total of 971 canine OR coding regions and afforded a comparative analysis of the dog, mouse, and human OR subgenomes.

Results

Canine olfactory sequence tags

To obtain efficient coverage of the canine OR repertoire, we employed the olfactory sequence tag (OST) approach [22], whereby pairs of degenerate primers were used to amplify OR clone sublibraries from canine genomic DNA, followed by subcloning and sequencing. Ten primers were designed according to human full-length OR coding sequences (http://bioinformatics.weizmann.ac.il/HORDE). The resulting combined degeneracy of the primer pairs ranged between 1.2×10^6 and 2.25×10^{10} (Table 1). These primers together with a primer pair previously used [23] enabled us to amplify OR gene segments corresponding to the conserved region between TM2 and TM7 in the OR protein. Fig. 1A shows that the highly degenerate primers gave better theoretical and experimental coverage of new genes compared to primers with lower degeneracy. Fig. 1B suggests that these primers were effective in capturing canine ORs of Class II, but less so for Class I ORs.

Table 1
Degenerate primers for deciphering the canine OR repertoire[a]

Side	Name	Primer sequence	Degeneracy
5′	L150a	YTNCAYNNNCCHATGTAYTWYYTBCT	147,456
3′	R150b	KTYCTNARRSTRTADAYNANDGGRTT	147,456
3′	R150a	TTYCKNARRSTRTADAYNANDGGRTT	147,456
5′	L40e	ATGTAYTTYTTCCTNDSHHWBYTSKC	41,472
3′	R40e	TTYYTMAVVSTRTADATNAVRGGRTT	41,472
5′	L40a	YTNCAYDMNCCHATGTAYTWYYTYCT	36,864
3′	R40a	TTYCTBARRSTRTARAYNANDGGRTT	36,864
5′	L8	CTBCAYDMNCCHATGTAYTTYYTYCT	6912
3′	R8	TTYCTCARDSTRTARATNADDGGRTT	6912
5′	L5a	CTBCAYDMNCCYATGTAYTTYYTYCT	4608
5′	L5	CCCATGTAYTTBTTYCTCDSYAAYYTRTC	1152
3′	R5	AGRCWRTANATGAANGGRTTCANCAT	1024

[a] Degenerate primers designed using a training set of 719 full-length human OR genes. L and R stand for 5′ and 3′ side of the primers.

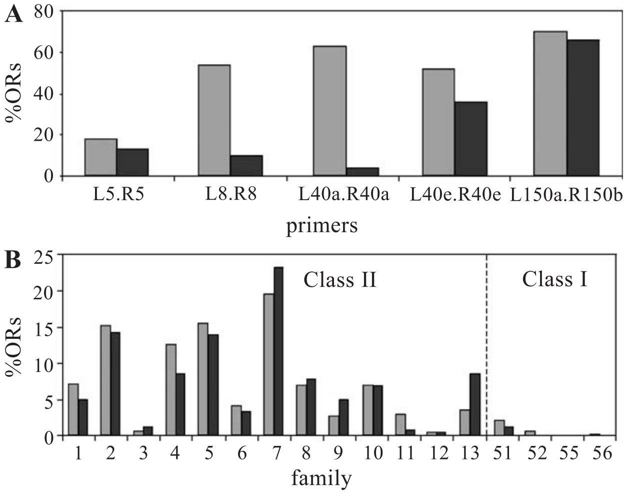

Fig. 1. Performance of OR-specific degenerate primers. (A) The theoretical percentage of genes out of the HORDE database (revision 38, 719 human OR genes) that match each primer pair (gray) and the relative amount of canine ORs experimentally amplified by each primer pair (black). We allowed up to three mismatched nucleotides (both sides combined) between the primer pair and the OR sequences. L and R stand for 5′ and 3′ primers (Table 1). (B) Primer performance for the 17 OR families (numbers on the abscissa), comparing theoretical (gray) versus experimental (black).

A total of 1200 OST clones were sequenced on both the forward and the reverse strands. Assembly of 1809 successful sequence reads, using 97% stringency over a minimum overlap of 40 bp, revealed 456 contigs, including 163 singletons. This amounts to an average sequence depth of 3.6 in the nonsingleton OSTs. OR contigs were conceptually translated and subjected to an all-against-all comparison by BLASTP. Using a cutoff of 98% identity, a set of 246 nonredundant canine OR protein sequences was obtained.

The OSTs were classified into families, as shown in Fig. 1B. They were found to span all Class II OR families and only family 51 of Class I ORs. The low representation of families 3 (three representatives) and 12 (one member), and the absence of three of the four Class I OR families, likely reflects the primer design bias related to the low fraction of certain OR families in the training set. Based on a survey of the nonsingleton OSTs, we estimate that 14.6% of the OSTs contain frame disruptions and are likely pseudogenes.

Canine genome data mining

The X 1.3 canine genome repositories of Celera Genomics (from a male standard poodle) were scrutinized for OR coding regions. Five BLASTN and four TBLASTN rounds were conducted using different OR sets each with 20 dog OSTs and human sequences as probes (Table 1s, supplementary material). These resulted in 2560 OR-positive sequence read hits. Consecutive rounds led to decreasing numbers of output reads (Fig. 1s, supplementary material), suggesting an asymptotic approach to complete search coverage. The sequence reads length ranged from 184 to 781 bp, with the most frequent fragment size of 649 bp (Fig. 2s, supplementary material).

The canine OR sequence compendium

The 2560 Celera OR-positive reads were subjected to sequence assembly. Optimal assembly parameters were sought such that reads of the same OR gene would be in one contig, while pairs of highly similar, recently duplicated genes would be mostly separated. Assembly was performed using a set of four different stringency parameters: 99, 98, 97, and 96% over a minimum overlap of 40 bp. We then utilized the knowledge that the average per-base coverage in the entire Celera canine sequence data set is X 1.3, corresponding to an average depth of 1.79 for the sequenced clones. The actual sequence coverage values for the above stringency parameters were respectively 1.50, 1.71, 1.79, and 1.83. The last three values were deemed consistent, within experimental error, with the computed Poisson-based coverage. We opted for a conservative stringency value of 98% to minimize erroneous clone merging.

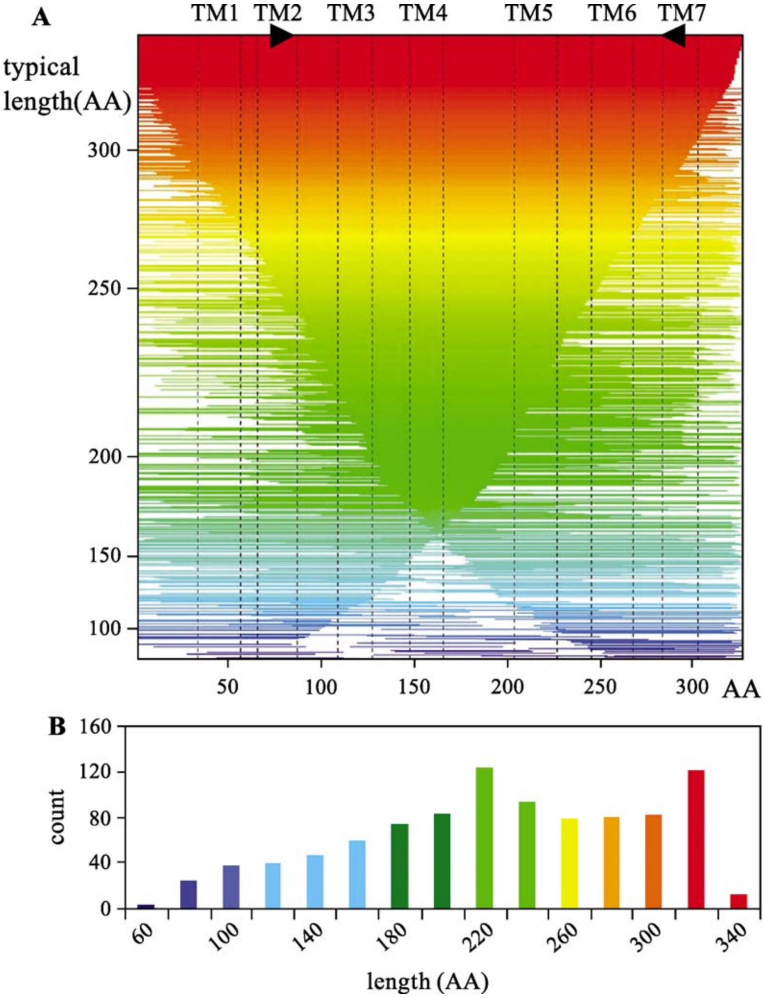

Fig. 2. Positional coverage of assembled OR sequences. (A) Pictorial multiple alignment of the repertoire, with each canine OR gene represented by a horizontal line indicating its extent within the ~300-amino-acid (AA)-long coding region. Colors represent coverage length from <100 residues (blue/lightgray) to full length (red/darkgray). The position of the sequence along the alignment was determined against a highly curated alignment of 191 human and mouse ORs. It is possible that for very short sequences two or more nonoverlapping segments are derived from the same OR gene. The vertical dashed lines indicate the positions of the seven transmembrane (TM) regions and the arrows on top show the rough positions of the degenerate primers. (B) The count distribution of canine ORs with different lengths of determined sequence.

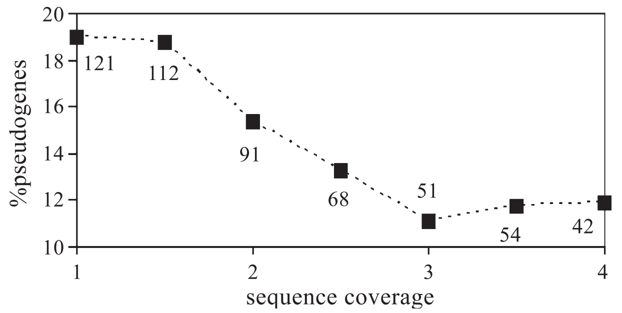

Fig. 3. Canine OR pseudogene fraction as a function of sequence depth. The analysis was carried out on all 121 ORs with full-length sequence. The percentage of ORs with one or more open reading frame disruptions is shown for subsets of this collection with increasing average sequence depth. The numbers next to the data points indicate subset size.

When applying the same stringency value (98%) to the 1809 OST sequences, 81 cases in which read pairs were erroneously separated were observed, and these were corrected manually. Finally, all the OR sequences were conceptually translated and subjected to an all-against-all comparison by BLASTP. Redundancy was removed using a cutoff of 98% identity. This procedure revealed 971 dog OR coding regions (Fig. 2A). This set included 73 ORs covered by OST sequences only, 697 with pure Celera coverage, and 201 with mixed coverage. Note that the 274 ORs covered by OST sequencing include the 246 OST contigs reported above, plus 28 additional sequences that constituted singletons in the OST project but were assembled with Celera sequences.

The combined sequences are derived from two different dog strains, beagle and poodle. We performed an analysis to

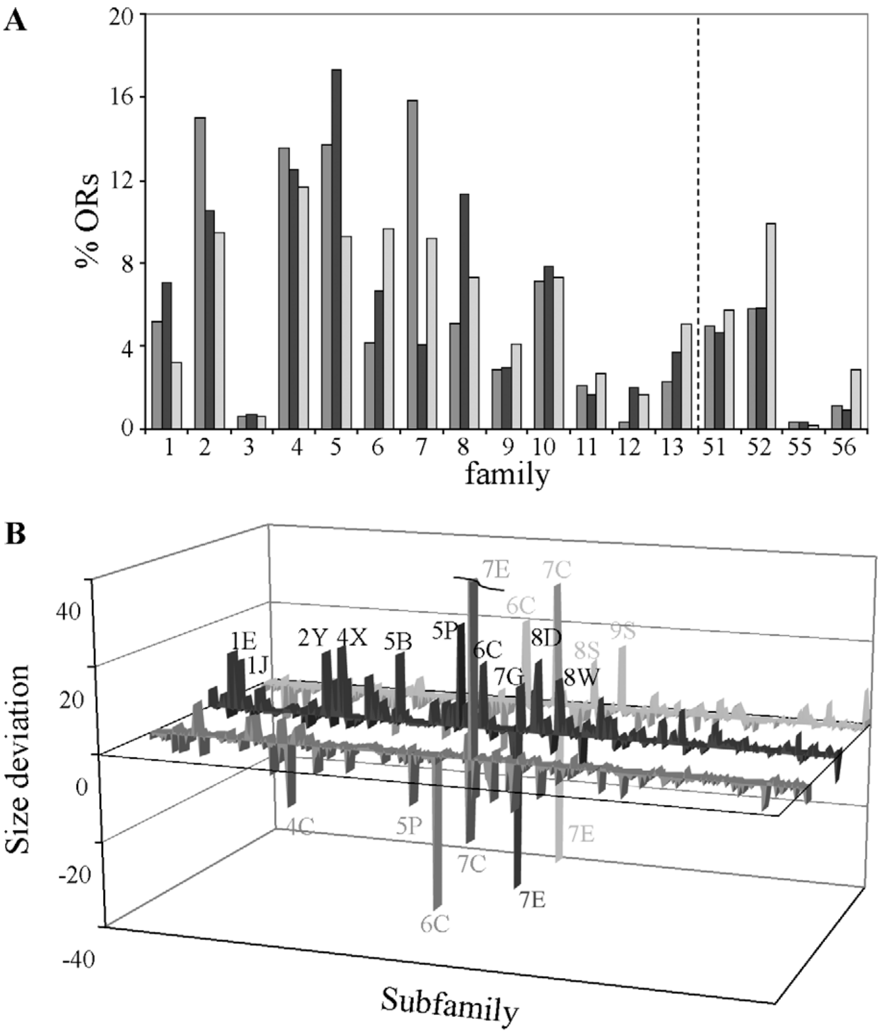

Fig. 4. A comparison of the three mammalian OR repertoires. (A) Distribution of OR family size in each mammalian species, dog (light), mouse (dark), and human (intermediate). Families to the left of the dashed line are Class II (tetrapods only) and to the right are Class I (both fish and tetrapods). (B) Comparison of OR subfamilies. For each subfamily, the deviation from the mammalian-average subfamily size is indicated as vertical displacement. Subfamilies with mostly positive values have undergone species-specific augmentation, while subfamilies with negative values are relatively underrepresented. Labels specify the most expanded or diminished subfamilies with color coding by species as in A. The out-of-scale deviation value for the human subfamily 7E is 81.7.

obtain an estimate of the interstrain sequence variability by looking for positions in which a clear sequence dichotomy occurred at positions with at least X 2 coverage for each of the strains. This was done to eliminate the large majority of variations due to sequencing errors. Of a total of 51,875 positions examined, 88 showed such defined interstrain variation, suggesting a divergence of 0.17%. In an additional 246 positions the sequence version in one of the strains corresponded to a potential polymorphism in the other.

Of the complete set of 971 ORs, 121 sequences had a complete coding region, and an additional 349 covered the OR core region, between TM2 and TM7, the amplification range of the OSTs. The other sequences had partial coverage of the OR coding region, but all were shown to have statistically significant similarity to other OR sequences. The excess of full-length and OST genes is also demonstrated in the length distribution (Fig. 2B), which shows two maxima around the lengths of 220 and 320 amino acids, respectively, corresponding to the extent of an OST and to that of a full-length OR. The complete canine OR sequencing results and analyses may be linked from the Human Olfactory Repertoire Data Exploratorium (HORDE, http://bioinfo.weizmann.ac.il/HORDE).

Pseudogene fraction

Of the 971 canine OR genes, 258 (26.5%) appear to encode pseudogenes, as judged by the existence of in-frame stop codons in the conceptual translation. This fraction likely represents an overestimate due to sequencing errors, particularly due to the relatively low sequence coverage in the Celera sequences. On the other hand, coding region segments that remained unsequenced may harbor additional open reading frame disruptions, generating an underestimate of the pseudogene fraction. To overcome both difficulties, we have focused on the 121 OR sequences that have full length coverage. The pseudogene fraction was then computed for subsets of this collection with increasing sequence coverage (Fig. 3). The pseudogene fraction was found to decrease with increasing sequence coverage, reaching a plateau at about 12%. This asymptotic value represents an approximation of the fraction of disrupted OR genes. It is an underestimate because it does not take into account partial-length OR pseudogenes. Such value is largely in agreement with that of 14% pseudogenes obtained when the OSTs were analyzed separately.

Estimating the canine OR repertoire size

An estimate of the OR repertoire size of the dog may be obtained by analyzing the degree of overlap between the two independent sequencing efforts, namely the OSTs and the Celera whole-genome set. While the former set is biased by primer choice, the latter is presumably unbiased. Thus, by asking how much coverage of the OSTs set was attained by Celera sequences one may estimate f, Celera's overall coverage fraction for the OR gene repertoire. Thus, since 201 of the 274 OSTs also had Celera coverage, $f = 201/274 = 0.733$ is obtained. Therefore, since the Celera-based mining yielded a total of 898 ORs, we estimate that the complete canine genome contains $898/0.733 = 1225$ ORs. Thus, our entire effort appears to have led to the discovery of 971 of 1225 ORs, hence of 79.3% of the canine olfactory subgenome.

An independent method for estimating the dog OR repertoire size relates to the number of OR-positive reads obtained from the canine, 1.3 X coverage genome (see Materials and methods). Using the number of reads $h = 2560$, an average length of Celera's reads $l = 649 \pm 71$ bp, and an effective size of a single OR gene $c = 1144 \pm 111$ bp, we obtain an estimation of $\sim 1100 \pm 100$ in the complete canine OR repertoire. This is an underestimate because it is known that our Celera genome data mining has likely not reached completeness.

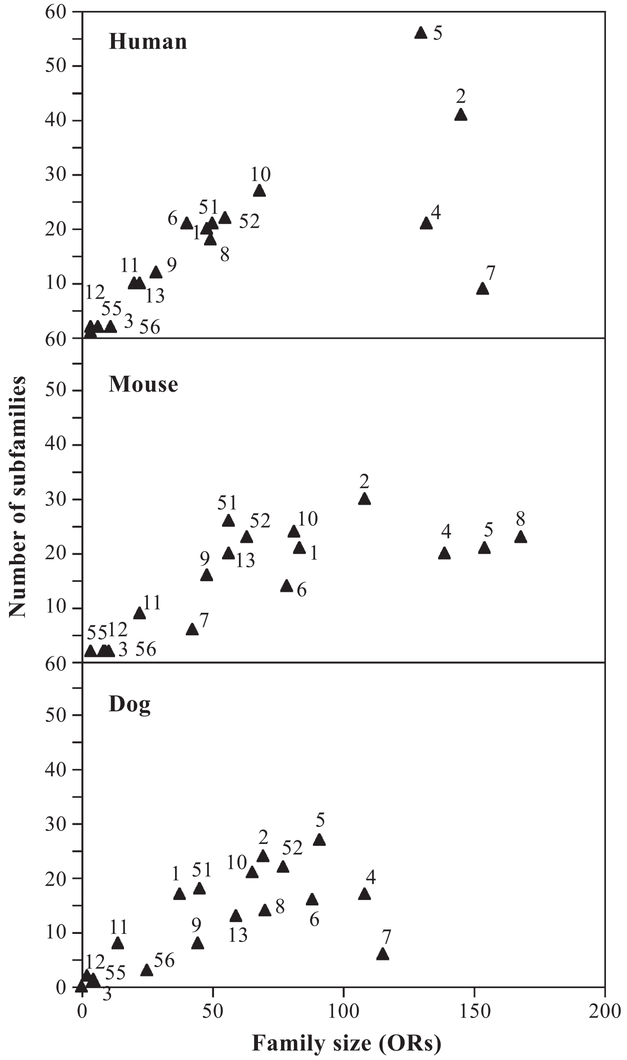

Fig. 5. Correlation between subfamily count and OR count for different OR families. The inverse of the slope reflects the "gene duplication rate" (number of genes per subfamily), which is higher for dog and mouse compared to the human repertoire. The numbers next to the data points indicate the family number.

Using a power law extrapolation of the increments of read counts after each data-mining round, we estimate that we have failed to spot at most 10% of the total number of reads. Thus, a more realistic estimation is ~1200 ORs in the complete canine repertoire, in good agreement with the first estimate.

OR families and subfamilies

We have subdivided the dog OR repertoire into families and subfamilies on the basis of evolutionary divergence, using the same algorithm as that used for the human OR classification [8]. To perform a phylogenetic comparison of dog, mouse, and human, we classified also in the same manner the mouse repertoire [10,11]. Fig. 4A shows the distribution of OR family sizes in the three species. One clear observation is that the three mammalian repertoires are composed of the same 17 families and in similar proportions. For example, in all three species, families 4 and 5 are among the most expanded and families 3, 12, and 55 are the smallest. This indicates that the families represent mainly ancient divergence events, predating the common ancestor of rodents, primates, and carnivores.

An interesting observation is the relatively high percentage of Class I ORs in the dog: 18.8% relative to 11.7 and 12.2% in mouse and human, respectively. Taking into consideration the OSTs primer bias against Class I OR families, the difference may even be higher in the complete canine OR repertoire.

The more recent evolutionary dynamics of the OR repertoire can be demonstrated in the change of subfamily count and size. As previously shown for human ORs [8], for most of the families in dog and mouse, the average number of ORs per subfamily is similar. This is manifested in a linear relationship between the number of genes and the number of subfamilies per family (Fig. 5). There are, however, significant deviations, such as families 4 (for all three species), families 2 and 7 in human, families 5 and 8 in mouse, and family 7 in dog, all of which have a higher average gene count per subfamily. Part of the explanation could be that certain subfamilies have undergone a recent set of gene duplication events. A comple-

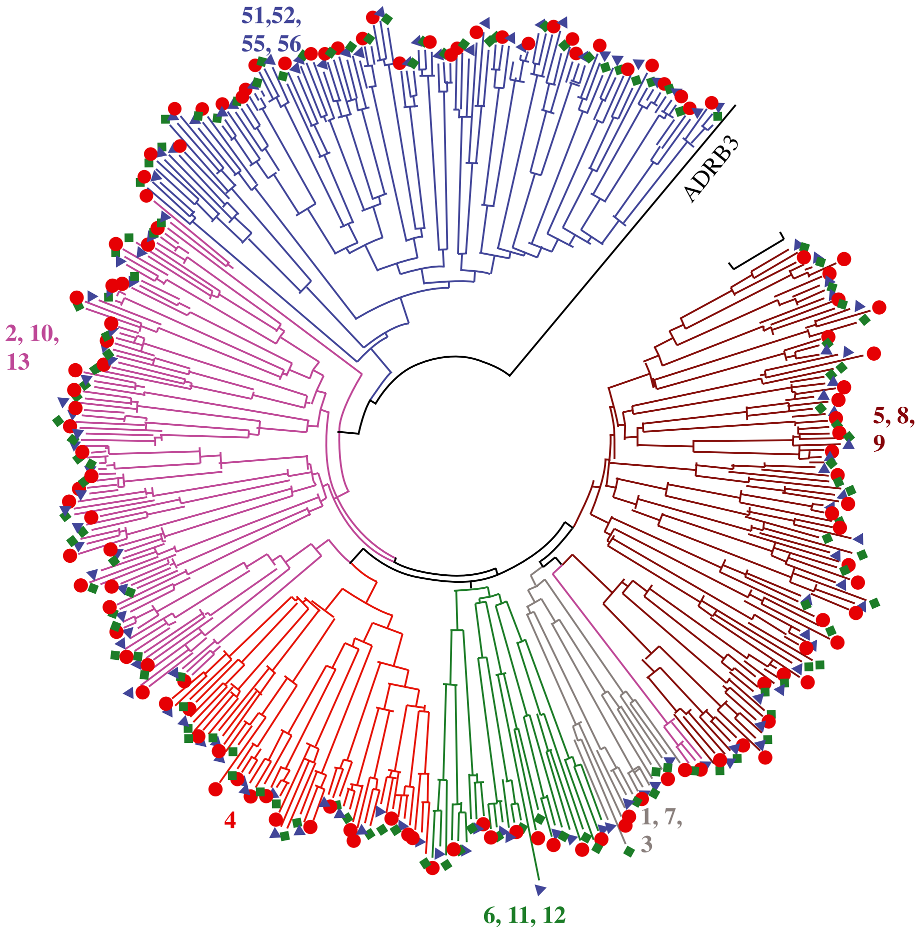

Fig. 6. A dendrogram analysis of 122 three-way OR mammalian orthologs. The branch colors signify groups of related OR families, indicated by numbers. Nodes are annotated with red circles (human), green diamonds (mouse), and blue triangles (dog). The multiple alignment was obtained with the CLUSTALX program [49] using the default parameters. The phylogenetic tree was derived by the MEGA2 program [51], with the neighbor-joining algorithm [27], and distances were calculated using the number of amino acid differences. Human β-adrenergic receptor type 3 (ADRB3) served as outgroup.

mentary explanation is the appearance of new subfamilies due to rapid deviation from the consensus of genes that turned into pseudogenes. It is particularly interesting that the slope of the graph is apparently lower for both mouse and dog compared to human, probably related to both explanations, i.e., fewer subfamily-specific duplications and many more pseudogenes.

Among a total of 412 subfamilies defined in all three mammals, there are 34 dog-specific subfamilies, compared to 60 human-specific subfamilies and 69 mouse-unique ones. However, the low number in dog could be augmented when the entire canine OR subgenome is deciphered. The existence of species-specific OR subfamilies is consistent with repertoire diversification that postdated the species divergence (Fig. 4B).

Detection of potential orthologs

Orthologs are defined as genes from different species that derive from a single gene in the last common ancestor, often also sharing functional attributes [24]. In the OR gene superfamily, assignment of orthologous pairs is more difficult, since there is almost no information about their functional specialization. The postspeciation gene duplication events generate additional difficulties. Thus, in an attempt to identify the most rigorously defined orthologs we have employed mutual best hit searches, based on sequence identity scores [25]. In addition, we have requested that the candidates share at least 78% identity over the entire protein, as previously described [26]. The analysis resulted in the identification of 234 mutual OR pairs in the mouse–dog comparisons, 229 in the dog–human comparison, and 344 in the human–mouse relationships. The latter higher figure probably relates to the fact that it involves the only comparison in which both repertoires are rather complete.

When a three-way analysis was performed, 122 ortholog triplets were found. These orthologs share a high sequence similarity of 83.8% on average and are proposed to constitute the most ancient and conserved core of the OR repertoire. Neighbor-joining analysis of a multiple sequence alignment [27] for the orthologous triplets supports the above identification (Fig. 6). The set of three-way orthologs contained representatives from all OR families, except for the smallest families, 3 and 12 (Fig. 7A). Interestingly, the Class I “fish-like” families had a larger proportion of orthologous triplets relative to Class II (tetrapod-specific) families (Fig. 7B). Also notable is that only 39.4% of the human genes involved in triples are pseudogenes, compared to the 63% in the total human repertoire [8]. Because for

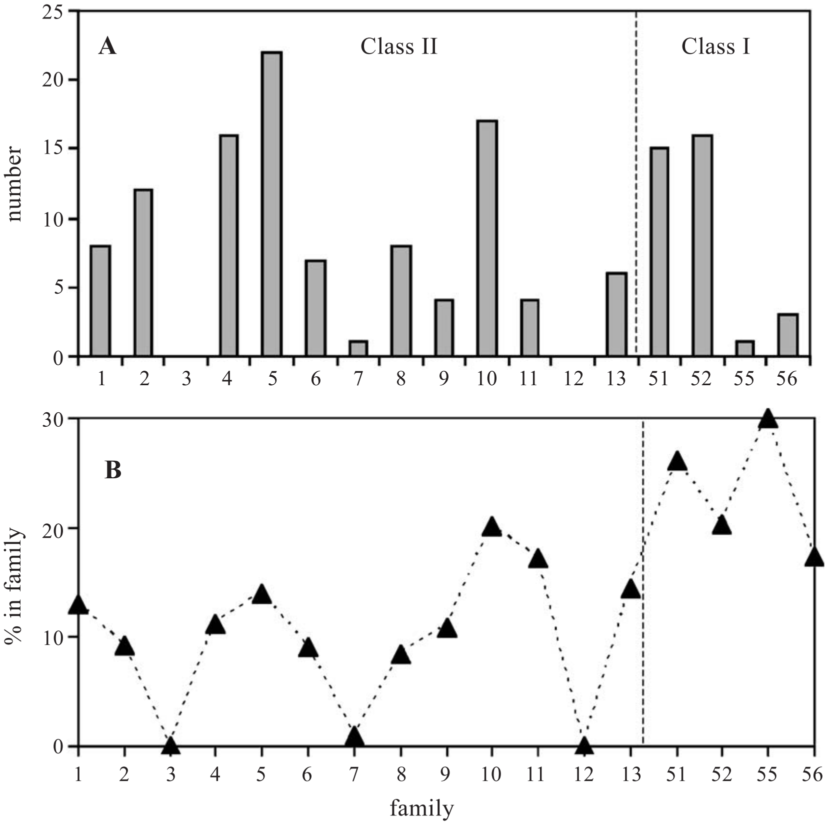

Fig. 7. Analyses of the three-way mammalian OR orthologs. (A) Representation of 121 three-way orthologs in the 17 OR gene families. (B) The percentage of three-way orthologs in each family. The normalization was performed relative to the three-mammal average. The vertical dashed line separates Class II (left) and Class I (right).

mouse and dog the pseudogene fraction within the three-way orthologs is about the same as in their entire OR repertoire (17.3 and 24.6%, respectively), the dichotomy in human underscores the functional importance of the three-way-ortholog subset.

Based on the classification into families and subfamilies, and on ortholog identification, we assigned a symbol for each of the dog OR genes. We have used the HORDE nomenclature system [15], with "*cOR*" as a root symbol. Triple orthologs were assigned the exact human symbol. Otherwise, the number within the subfamily was assigned in a continuation of human symbols. This information will be used to enhance the HORDE database, originally established for human ORs, with the newly defined canine ORs.

Evolutionary inferences

The present data, together with the availability of the complete mouse and human OR repertoires, provide an opportunity for a three-way phylogenetic analysis. Recent studies of placental mammals suggested several possible scenarios for the evolutionary relationship of dog, mouse, and human. One study, based on the comparison of 18 gene segments, suggests that the dog divergence occurred before the mouse–human separation [28]. Other studies, utilizing mitochondrial genes, suggested that the mouse separation from the other two species occurred first [29] or that all three separated from one another concomitantly [30]. To obtain an appreciation of which scenario shows better consistency with the data of ORs, we analyzed the pairwise identity scores of all the DNA sequences of the 122 three-way orthologs. Our results show that the average human–dog identity score (81.2%) is significantly higher than that of human–mouse (75.4%) and dog–mouse (75.7%) ($p = 2.3 \times 10^{-13}$, using Kolmogorov–Smirnov test). Thus, these results tentatively support the scenario in which mouse diverged earlier. However, at the protein level, the respective values are 84.5, 82.6, and 84.2, suggesting equal evolutionary distance. Altogether, these results do not justify a strong claim that one of the mammalian OR repertoires has undergone a considerable degree of independent, species-specific evolution.

Discussion

We have employed a comprehensive experimental and in silico approach to decipher a large fraction of the dog OR repertoire. A combination of de novo DNA sequencing and genome-wide data mining enabled us to define 971 OR genes.

A previous implementation of the OST approach, as part of the DEFOG scheme, deciphered successfully a third of the human repertoire [23]. In this study, we have used a set of degenerate primers designed according to human sequences to obtain 246 canine OR sequences. This further demonstrates the power of the method and its capacity to cross boundaries between two mammalian species.

Previously, only 20 dog OR sequences had been published [16–18], and thus the present report represents significant progress in the elucidation of the OR repertoire of this mammalian species. Of the previously reported 20 canine ORs, 16 were found by our procedure, again indicating roughly 80% coverage, in agreement with the value of 79.3% coverage computed here by another method.

Macrosmatic and microsmatic species

Dogs are macrosmatic animals, which rely highly on their sense of smell, although some evidence points to the contrary [31,32]. Many experiments attest that the canine olfactory capacities much exceed those of humans [33]. A recently published article in support of this notion describes a behavioral test that indicates that dogs can distinguish different species of animals by their odors and even perform olfactory differentiation of individuals within a species [34]. Also identification of individual humans by dogs can be presented as evidence in many courts of law, similar to fingerprints [35].

Several neuroanatomical features can underlie the superb smelling ability of macrosmatic animals, such as dog or mouse. In addition to a larger area of sensory epithelium, the fraction of their brain structures devoted to processing olfactory information is considerably larger than for microsmatic animals such as primates [36].

The elucidation of most of the canine OR repertoire may provide essential tools for assessing the contribution of the OR gene repertoire to such dichotomy. We estimate that the full canine repertoire contains about 1200 genes, similar in size to the mouse OR repertoire, 1296 [10] and 1500 [11]. The fraction of canine OR pseudogenes, which is estimated to be as low as 12%, is also not significantly different compared to the value in mouse (20% [10,11]). In humans, the OR repertoire is composed of 1116 genes, of which 67% are pseudogenes [37]. The actual size of the dog and mouse functional OR repertoire is therefore three times higher than in human, well correlated with the macrosmatic/microsmatic dichotomy. The partial data available for several primate species [38,39] suggest that, indeed, the olfactory ability is correlated with the proportion of functional genes. Taken together, the data reported here suggest that the dog's functional OR repertoire is not unusually large and that the most unusual of the dog-specific chemosensory traits may be ascribable to central nervous system characteristics rather than to sensory epithelial characteristics.

Class I Receptors in the dog

ORs belonging to Class I compose the most ancient group in the olfactory repertoire. They were originally identified in fish [40] and subsequently found to be intermixed with Class II (mammalian-type) ORs in amphibian species [41]. In human, 120 Class I ORs have been identified, organized in a single cluster on chromosome 11 [8]. Their number in the mouse repertoire is even somewhat larger, i.e., 147 genes [10], and they are concentrated on the syntenic chromosome 7 at 85–88 Mb, in two clusters ([10] and A. Sadot et al., unpublished). It was suggested that Class II ORs are specialized for recognizing airborne odorants, whereas Class I ORs in fish and amphibia bind nonvolatile odorants such as amino acids [42]. Therefore their relatively large number in both the human and the mouse repertoires was surprising. Expression of Class I ORs has already been reported in rat [43] and in human [44]. Our observation that Class I ORs are more conserved in the three-way mammalian comparison than Class II confirms the notion that these receptors may be centrally involved in mammalian olfaction. The fact that the canine OR repertoire contains more Class I receptors, relative to human and mouse, might relate to special olfactory capacities in the dog.

Species-specific subfamily expansions

The observation that several subfamilies are larger in each of the three mammalian genomes compared to the two others could be related to distinct chemosensory requirements of the individual species. However, it should be noted that no direct evidence exists at present for functional significance related to either family or subfamily, and a species-specific subfamily does not necessarily mean species-specific odorant detection. We detected 34 dog-specific subfamilies and, in addition, 4 subfamilies that are significantly expanded in the canine olfactory subgenome. The most notable human expansion is that of subfamily 7E, which contains 127 genes compared to only 7 and 2 in mouse and dog repertoires, respectively [22,45]. However, all but one of the human subfamily 7E ORs are pseudogenes, and therefore this expansion may have limited functional consequences. In contrast, the discovery of dog-expanded subfamilies such as 6C and 7C, and even more so of canine-specific OR subfamilies (e.g., 8S and 9S), could be indicative of newly evolved functions.

Materials and methods

Degenerate primers design

Degenerate primers were designed using the HYDEN software (Highly Degenerate Primers: http://www.cs.tau.ac.il/~rshamir/hyden/HYDEN.htm). Given a set of DNA family-related sequences, HYDEN constructs degenerate primer pairs that match many of the given sequences with up to three mismatched bases. The full algorithmic details are described in [46]. For this work we have designed 10 primers, based on 719 human full-length ORs listed in the HORDE database version 38 as described in Table 1. The primers L5 and R5 were designed previously [23].

Amplification and subcloning

Canine genomic DNA from an individual beagle (Novagen,Madison, WI, USA) was used. PCR were performed with different combinations of the degenerate primer pairs described in Table 1. Reactions were carried out in a total volume of 25 μl, containing 0.2 mM concentration of each deoxynucleotide (Promega, Madison, WI, USA); 50 pmol of each primer; PCR buffer containing 1.5 mM $MgCl_2$, 50 mM KCl, 10 mM Tris, pH 8.3; 1 unit of *Taq* DNA polymerase (Boehringer Mannheim, Mannheim, Germany); and 50 ng of genomic DNA. PCR conditions were as follows: 35 cycles of 1 min denaturation at 94°C, 1 min annealing at 55°C, and 1 min extension at 72°C. The first step of denaturation and the last step of extension were both 3 min.

All primers were modified for subsequent subcloning into the pAMP1 vector (Gibco BRL). The PCR products were subcloned without prior purification, using the Clone Amp System (Gibco BRL) and DH5 competent bacterial cells. Subclones underwent a second PCR amplification with vector-specific primers and were subjected to sequencing. Sequencing reactions were performed in both directions with a dye-terminator cycle sequencing kit (Perkin–Elmer) on an ABI 3700 automated DNA sequencer. After base calling with the ABI Analysis Software (version 3.0), forward and reverse sequences were assembled after removal of the primer sequences using Sequencher (GeneCodes Corp., version 4.1).

Data mining procedures

For the data mining procedure we prepared a set of 119 OR queries designed to cover the maximal OR sequence space (Table 1s, supplementary material). The set was composed of seven groups, each group containing representatives of all 17 mammalian OR families. The first group consisted of family consensus sequences of human ORs (O. Man et al., unpublished). The remaining six groups contained representatives of the 17 mammalian OR subfamilies, but each of a different subfamily. We used sequence collections of human ORs and dog OSTs, with canine OST sequences entered into the list with a higher priority.

Four rounds of TBLASTN were conducted against Celera's X 1.3 coverage of the dog genome, each round using 20 query sequences from the list, beginning with the first sequence. In parallel, five rounds of BLASTN were done, each using 20 query sequences from the list, beginning with sequence 18. For all BLAST searches a target sequence read was considered OR-positive if it contained a contiguous stretch of 40 bp showing ≥ 80% identity.

Estimating the repertoire size

Let h be the number of OR-positive reads obtained from the canine X 1.3 coverage genome. The fraction that h constitutes relative to the total number of reads in Celera's database, s, is an estimate of the probability of a single read overlapping with an OR sequence:

$$p_{OR} = h/s. \tag{1}$$

We can obtain this probability alternatively using

$$p_{OR} = \frac{cr}{g}, \tag{2}$$

where r is the OR repertoire size, c is the effective size of a single OR gene, and g is the canine genome length. The effective size of a single OR gene is given by

$$c = l + l_{OR} - 2\theta, \tag{3}$$

where l and l_{OR} are the average lengths of Celera's reads and OST query, respectively, and θ is the required overlap between the two. Thus, from Eqs. (1) and (2) one obtains

$$r = hg/sc. \tag{4}$$

To solve the equation we used the following relationship, which stems from the definition of sequence coverage,

$$s = \frac{gX}{l}, \tag{5}$$

where X is the sequence coverage. Eqs. (2) and (4) will yield the following formula for the repertoire size:

$$r = \frac{hl}{cX}. \tag{6}$$

Sequence assembly

OSTs and Celera's OR-positive sequences were assembled using the Sequencher software package for Macintosh (GeneCodes Corp, Michigan, USA, version 4.1). Celera's sequences were obtained and entered into the assembly as text files, whereas for the OST sequences the full chromatograms were available.

The comparison between the expected sequence coverage distribution in Celera's whole-genome data and that found in the assembly of OR sequences was done as follows. A Poisson distribution was assumed (Eq. (7)) [47],

$$p_X = \frac{e^{-\lambda}\lambda^X}{X!}, \tag{7}$$

and used to calculate the expected average base coverage for the known canine Celera genome sequence depth parameter ($\lambda = 1.3$). This was achieved using all X values except $X = 0$. The result was compared to coverage values obtained from assemblies with different stringency parameters.

Sequence analysis

An OR was considered to be full length if the following conditions were fulfilled: it had a length of at least 290 amino acids, an ATG codon at the beginning, and the contig that defined the gene covered more that the gene's length.

OR sequences were translated to proteins when they had a single ORF. In other cases they were conceptually translated using FASTY [48] against a core of properly translated OR sequences. The protein sequences showed an excess of frame disruptions at the edges of the proteins, because a quality clipping was not applied on Celera's reads and FASTY translation artifacts. Therefore, frame disruptions at the edges of the protein sequences (up to 1% of the protein length) were trimmed. A special routine was applied whenever possible on sequences longer than 280 amino acids to choose a suitable ATG codon for initiation and to end at a stop codon. A sequence was considered to be an OR if it showed at least 40% identity to another OR (human, mouse, or dog) over at least 100 amino acids or over 80% of the protein sequence, for sequences shorter than 110 amino acids.

Family and subfamily classification of dog and mouse repertoires was performed by amino acid similarity to previously classified OR genes, as described in [15]. For each subfamily we calculated an average size over dog, human, and mouse repertoires. A subfamily was considered to be species-specific enlarged if it was larger than twice the average size and it contained more than five members.

For the alignment of the dog repertoire (Fig. 2A), each dog OR was aligned to a well-curated multiple alignment of mouse and human ORs [52] generated with ClustalX [49]. The alignment was analyzed to define the position of each sequence with regard to the standard OR alignment.

For the dog–human–mouse OR sequence comparisons, we used the HORDE (http://bioinformatics.weizmann.ac.il/HORDE) database and a set of 1570 mouse OR sequences composed from two published mouse repertoires (S. Firestein, Columbia University, personal communication, July 2001, and http://www.fhcrc.org/labs/trask/OR/showAllBACs.html; GenBank Accession Nos. AY072961–AY074256). Each OR from the three repertoires was utilized as a query for BLASTP [50] comparison to all three repertoires. The best hits for each OR were stored in the CORDE database in the format of mySQL tables. BLASTN [50] was used to compare the DNA of the three-way orthologs.

Acknowledgments

This work was supported by German–Israel Foundation Grant G-0506-183.0396 (D. Lancet and R. Shamir) and the Israel Science Foundation (Grant 309/09) (R.Shamir) and the Crown Human Genome Center of the Weizmann Institute (D. Lancet).

References

[1] E.A. Ostrander, F. Galibert, D.F. Patterson, Canine genetics comes of age, Trends Genet. 16 (2000) 117–124.

[2] J.P. Scott, J.L. Fuller, Genetics and the Social Behavior of the Dog, Univ. of Chicago Press, Chicago, 1965.

[3] D.G. Moulton, in: D. Muller-Shwarze, M.M. Mozell (Eds.), Chemical Signals in Vertebrates, Plenum, New York, 1977, pp. 455–464.

[4] J.M. Young, B.J. Trask, The sense of smell: genomics of vertebrate odorant receptors, Hum. Mol. Genet. 11 (2002) 1153–1160.

[5] S. Firestein, How the olfactory system makes sense of scents, Nature 413 (2001) 211–218.

[6] P. Mombaerts, The human repertoire of odorant receptor genes and pseudogenes, Annu. Rev. Genom. Hum. Genet. 2 (2001) 493–510.

[7] D. Lancet, Vertebrate olfactory reception, Annu. Rev. Neurosci. 9 (1986) 329–355.

[8] G. Glusman, I. Yanai, I. Rubin, D. Lancet, The complete human olfactory subgenome, Genome Res. 11 (2001) 685–702.

[9] S. Zozulya, F. Echeverri, T. Nguyen, The human olfactory receptor repertoire, Genome Biol. 2 (2001) (research0018).

[10] X. Zhang, S. Firestein, The olfactory receptor gene superfamily of the mouse, Nat. Neurosci. 5 (2002) 124–133.

[11] J.M. Young, et al., Different evolutionary processes shaped the mouse and human olfactory receptor gene families, Hum. Mole. Genet. 11 (2002) 535–546.

[12] R.P. Lane, et al., Genomic analysis of the olfactory receptor region of the mouse and human T-cell receptor alpha/delta loci, Genome Res. 12 (2002) 81–87.

[13] G. Glusman, et al., Sequence, structure, and evolution of a complete human olfactory receptor gene cluster, Genomics 63 (2000) 227–245.

[14] M. Bulger, et al., Comparative structural and functional analysis of the olfactory receptor genes flanking the human and mouse beta-globin gene clusters, Proc. Natl. Acad. Sci. USA 97 (2000) 14560–14565.

[15] G. Glusman, et al., The olfactory receptor genesuperfamily: data mining classification and nomenclature, Mamm. Genome 11 (2000) 1016–1023.

[16] M. Parmentier, et al., Expression of members of the putative olfactory receptor gene family in mammalian germ cells, Nature 355 (1992) 453–455.

[17] P. Vanderhaeghen, S. Schurmans, G. Vassart, M. Parmentier, Specific repertoire of olfactory receptor genes in the male germ cells of several mammalian species, Genomics 39 (1997) 239–246.

[18] L. Issel Tarver, J. Rine, Organization and expression of canine olfactory receptor genes, Proc. Natl. Acad. Sci. USA 93 (1996) 10897–10902.

[19] P. Werner, et al., Comparative mapping of the diGeorge region in the dog and exclusion of linkage to inherited canine conotruncal heart defects, J. Hered. 90 (1999) 494–498.

[20] M. Breen, et al., Chromosome-specific single-locus FISH probes allow anchorage of an 1800-marker integrated radiation-hybrid/linkage map of the domestic dog genome to all chromosomes, Genome Res. 11 (2001) 1784–1795.

[21] L.E. Palmer, et al., A survey of canine expressed sequence tags and a display of their annotations through a flexible web-based interface, J. Hered. 94 (2003) 15–22.

[22] T. Fuchs, G. Glusman, S. Horn-Saban, D. Lancet, Y. Pilpel, The human olfactory subgenome: from sequence to structure and evolution, Hum. Genet. 108 (2000) 1–13.

[23] T. Fuchs, et al., Defog—a practical scheme for deciphering families of genes, Genomics 80 (2002) 295–302.

[24] W.M. Fitch, Distinguishing homologous from analogous proteins, Syst. Zool. 19 (1970) 99–113.

[25] G.M. Rubin, et al., Comparative genomics of the eukaryotes, Science 287 (2000) 2204–2215.

[26] M. Lapidot, et al., Mouse–human orthology relationships in an olfactory receptor gene cluster, Genomics 71 (2001) 296–306.

[27] N. Saitou, M. Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees, Mol. Biol. Evol. 4 (1987) 406–425.

[28] E. Eizirik, W.J. Murphy, S.J. O'Brien, Molecular dating and biogeography of the early placental mammal radiation, Jo. Hered. 92 (2001) 212–219.

[29] P.S. Corneli, R.H. Ward, Mitochondrial genes and mammalian phylogenies: increasing the reliability of branch length estimation, Mol. Biol. Evol. 17 (2000) 224–234.

[30] Y. Cao, M. Fujiwara, M. Nikaido, N. Okada, M. Hasegawa, Interordinal relationships and timescale of eutherian evolution as inferred from mitochondrial genome data, Gene 259 (2000) 149–158.

[31] P. Nicollini, Lo stimulo olfactoorio e la sua recezione, Arch. Ital. Sci. Farmacol. 4 (1954) 109–172.

[32] M. Laska, A. Seibt, Olfactory sensitivity for aliphatic esters in squirrel monkeys and pigtail macaques, Behav. Brain Res. 134 (2002) 165–174.

[33] D.A. Marshal, D.G. Moulton, Olfactory sensitivity to alpha-ionone in humans and dogs, Am. J. Vet. Res. 46 (1981) 2409–2412.

[34] D.A. Smith, K. Ralls, B. Davenport, B. Adams, J.E. Maldonado, Canine assistants for conservationists, Science 291 (2001) 435.

[35] I.L. Brisbin, S. Austad, S.K. Jacobson, Canine detectives: the nose knows—or does it? Science 290 (2000) 1093.

[36] H. Stephan, G. Baron, H.D. Frahn, Comparative size of brains and brains structures, in: H. Steklis, J. Erwin (Eds.), Comparative Primate Biology Neurosciences, vol. 4, A.R. Liss, New York, 1988, pp. 1–38.

[37] M. Safran, et al., Human gene-centric databases at the Weizmann Institute of Science: Genecards, Udb, Crow 21 and Horde, Nucleic Acids Res. 31 (2003) 142–146.

[38] S. Rouquier, A. Blancher, D. Giorgi, The olfactory receptor gene repertoire in primates and mouse: evidence for reduction of the functional fraction in primates, Proc. Natl. Acad. Sci. USA 97 (2000) 2870–2874.

[39] Y. Gilad, O. Man, S. Paabo, D. Lancet, Human specific loss of olfactory receptor genes, Proc. Natl. Acad. Sci. USA 100 (2003) 3324–3327.

[40] J. Freitag, J. Krieger, J. Strotmann, H. Breer, Two classes of olfactory receptors in Xenopus laevis, Neuron 15 (1995) 1383–1392.

[41] M. Mezler, J. Fleischer, H. Breer, Characteristic features and ligand specificity of the two olfactory receptor classes from Xenopus laevis, J. Exp. Biol. 204 (2001) 2987–2997.

[42] H. Sun, et al., Evolutionary analysis of putative olfactory receptor genes of medaka fish, oryzias latipes, Gene 231 (1999) 137–145.

[43] K. Raming, S. Konzelmann, H. Breer, Identification of a novel G-protein coupled receptor expressed in distinct brain regions and a defined olfactory zone, Recept. Channels 6 (1998) 141–151.

[44] E.A. Feingold, L.A. Penny, A.W. Nienhuis, B.G. Forget, An olfactory receptor gene is located in the extended human beta-globin gene cluster and is expressed in erythroid cells, Genomics 61 (1999) 15–23.

[45] T. Newman, B.J. Trask, Complex evolution of 7e olfactory receptor genes in segmental duplications, Genome Res. 13 (2003) 781–793.

[46] C. Linhart, R. Shamir, The degenerate primer design problem, Bioinformatics 18 (2002) S172–S180.

[47] M. Wendl, et al., Theories and applications for sequencing randomly selected clones, Genome Res. 11 (2001) 274–280.

[48] W.R. Pearson, D.J. Lipman, Improved tools for biological sequence comparison, Proc. Natl. Acad. Sci. USA 85 (1988) 2444–2448.

[49] J.D. Thompson, T.J. Gibson, F. Plewniak, F. Jeanmougin, D.G. Higgins, The clustalX Windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools, Nucleic Acids Res. 24 (1997) 4876–4882.

[50] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.

[51] S. Kumar, K. Tamura, M. Nei, Mega: Molecular Evolutionary Genetics Analysis, edn version 1.01, Edited by Pennsylvania State Univ., University Park, PA, 1993.

[52] O. Man, Y. Gilad, D. Lancet, Prediction of the oderant binding site of the olfactory proteins by human-mouse comparisons, Protein Sci. (2003), in press.