Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

The olfactory receptor universe - from whole genome analysis to structure and evolution

Tsviya Olender[1]*, Ester Feldmesser[1]*, Tal Atarot[1], Miriam Eisenstein[2] and Doron Lancet[1]

[1]Department of Molecular Genetics and Crown Human Genome Center, and [2]Chemical Research Support Unit, Weizmann Institute of Science, Rehovot 76100, Israel
*These authors contributed equally to this study.
Corresponding author: T. Olender
E-mail: tsviya.olender@weizmann.ac.il

Genet. Mol. Res. 3 (4): 545-553 (2004)
Received October 4, 2004
Accepted December 15, 2004
Published December 30, 2004

ABSTRACT. Olfactory receptors (ORs) constitute the largest gene-family in the vertebrate genome. We have attempted to provide a comprehensive view of the OR universe through diverse tools of bioinformatics and computational biology. Among others, we have constructed the Human Olfactory Receptor Data Exploratorium (HORDE, http://bioportal.weizmann.ac.il/HORDE/) as a free online resource, which integrates information on ORs from different species. We studied the genomic organization of 853 human ORs and divided the repertoire into 135 clusters, accessible through our new cluster viewer feature. An analysis of intact and pseudogenized ORs in different clusters, as well as of OR expression patterns, is provided, relevant to OR transcription control. Coding single nucleotide polymorphisms were integrated; these are to be used for genotype-phenotype correlation studies. HORDE allows a unique opportunity for discerning protein structural and functional information of the individual OR proteins. By applying novel data analysis strategies to the >3000 OR genes of mouse, dog and human within HORDE, we have generated a set of refined rhodopsin-based homology models for ORs. For model improvement, we employed a novel analysis of specific positions along the seven transmembrane helices at which prolines generate helix-breaking kinks. The model shows family-specif-

ic structural features, including idiosyncratic kink patterns, which lead to significant differences in the inferred odorant binding site structure. Such analyses form a basis for a comprehensive sequence-based classification of OR proteins in terms of potential odorant binding specificities.

Key words: Olfactory receptor, HORDE, Computational data mining, Database, Homology modeling, Sequence analysis

INTRODUCTION

Olfaction, the sense of smell, detects and discriminates thousands of odorant molecules. This capacity is made possible by the olfactory receptor (OR) protein superfamily, belonging to the hyperfamily of G-protein coupled (7-transmembrane helix) receptors. ORs are expressed on the membrane of olfactory sensory neurons and enable the transduction of the chemical signal to neuronal action potentials. ORs are distributed in clusters on most human chromosomes, which is evidence for the evolutionary processes that are responsible for their genome expansion by gene duplication and conversion (Glusman et al., 2000; Niimura and Nei, 2003). The completion of the human and several other mammalian genomes enabled the elucidation of the complete human OR repertoire (Glusman et al., 2001; Zozulya et al., 2001; Niimura and Nei, 2003), as well as its comparison with those of the mouse, dog and chimpanzee (Glusman et al., 2000; Young et al., 2002; Zhang and Firestein, 2002; Gilad et al., 2004; Olender et al., 2004). These efforts deciphered ~900 genes in human and >1200 ORs in mouse, dog and chimpanzee. These OR sequences are catalogued in HORDE, the Human Olfactory Receptor Data Exploratorium (http://bip.weizmann.ac.il/HORDE). HORDE was established in 2000 based on the first draft of the human genome (Fuchs et al., 2000; Glusman et al., 2001) and it has been under continuous development and improvement since. It aims at keeping the most updated status of the human OR repertoire, as inferred from the progress of human genome sequencing. In addition, it presents to the user a global overview on OR genes and proteins, including their structure, function and evolution.

There have been novel enhancements in HORDE, aimed at supplying tools that will facilitate future research in this field. These are focused on obtaining additional information, such as genomic organization, transcription, expression profiles, and single nucleotide polymorphisms (SNPs) with known allele frequencies.

OR modeling can predict structure-function relations within the olfactory system for coupling odorous compounds to their corresponding receptors (Floriano et al., 2000; Vaidehi et al., 2002). However, an experimental 3-dimensional structure of ORs has not yet been developed, hence homology modeling has to be utilized. Despite the availability of a high-resolution crystal structure of bovine rhodopsin (Palczewski et al., 2000) practically no systematic OR modeling based on it has so far been attempted.

Transmembranal signaling upon agonist binding to a GPCR is probably attained by a series of conformational transitions that alter the mutual disposition of the helices and their intracellular interaction with the G protein (Sansom and Weinstein, 2000; Decaillot et al., 2003). The structural basis of such conformational changes lies in the properties of individual transmembranal helices, and in particular, non-helical elements such as proline-induced kinks

that serve as activation switches (Sansom and Weinstein, 2000; Decaillot et al., 2003). The occurrence of non-canonical elements in transmembranal helices must be taken into consideration in the modeling procedure. Here, we present a novel approach, which uses repertoire-wide sequence comparisons, for refinement of rhodopsin-based homology models of ORs. We employed conservation/variability signals and a comprehensive analysis of specific positions along the 7-transmembrane helices at which prolines generate helix-breaking kinks. The procedure produced OR structures that were more suitable for odorant binding prediction.

MATERIAL AND METHODS

The current OR compendium was deciphered using the complete human genome assembly (http://genome.ucsc.edu/goldenPath/releaseLog.html#hg16, NCBI build#34). The data-mining procedure uses a BLAT search (Kent, 2002) with each of previous HORDE's sequences, where all hit locations are collected and extracted from the genome sequence, without applying any cutoff. Another procedure classifies the ORs as pseudo/intact. Pseudogenes are translated via FASTY. For cluster definition, we used 100 kb as a maximal distance between two consecutive ORs within a cluster.

ESTs and mRNAs were extracted from UCSC genome browser tables (http://genome.ucsc.edu/goldenPath/hg16/database/chr#_est.txt.gz and chr#_mrna.txt.gz, # is for the chromosome number) based on genomic location overlap between ORs and the ESTs.

SNP information is currently introduced into HORDE from dbSNP (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=snp). For the analysis of non-canonical element conservation we used all HORDE human, mouse and dog OR sequences that contained up to two frame disruptions. A highly curated multiple alignment described in Man et al., (2004) was used as a template for aligning the sequences. This was achieved with ClustalX (Thompson et al., 1997) and the default parameters. Conservation/variabiliy scores were calculated using the ConSurf algorithm (http://consurf.tau.ac.il/). Homology models of the Rat OR-I7 (ortholog of human OR6A2) were generated using the 'Homology' module of InsightII (Accelrys, Inc., San Diego, CA, USA), and the three-dimensional coordinates of bovine rhodopsin (PDB entry 1F88) as a template.

RESULTS

OR compendium and nomenclature

Our OR digital compendium HORDE, widely used worldwide, contains 853 OR entries, of which 386 have an intact open reading frame; the rest are probable pseudogenes. This collection (HORDE #40) constitutes the most up-to-date repertoire of human ORs, deciphered out of the public and Celera genomic complete human genome assemblies. Although the number of ORs in HORDE is currently less than the 906 in HORDE#38 (Glusman et al., 2001), and the 1024 in HORDE #39 (Safran et al., 2003), it contains 55 novel ORs, and redundant ORs in previous versions were unified. This is a direct result of the improvement in the sequence quality of the human genome. Another outcome of the reduction in sequencing errors is a smaller proportion of pseudogenes in the human repertoire (55% in HORDE #40 relative to 67% in HORDE #39).

To allow ortholog-paralog comparisons, HORDE includes information on ORs from other mammalian species - mouse, dog and chimpanzee. Most of the routines for information extraction are based on automatic data mining, allowing facile update of the database. A key feature in HORDE is a "card" for every OR gene (Figure 1), which contains genomic cluster disposition, SNPs and hyperlinks to other databases, including GeneCards (http://bioinfo.weizmann.ac.il/cards/index.shtml; Safran et al., 2003). Also provided is a widely accepted, HUGO-approved systematic nomenclature (http://www.gene.ucl.ac.uk/nomenclature/genefamily.shtml) that affords an instant guide to the position of a gene in a phylogenetic tree: OR1A2 signifies family 1, subfamily A, member 2. An example of a card is shown in Figure 1.

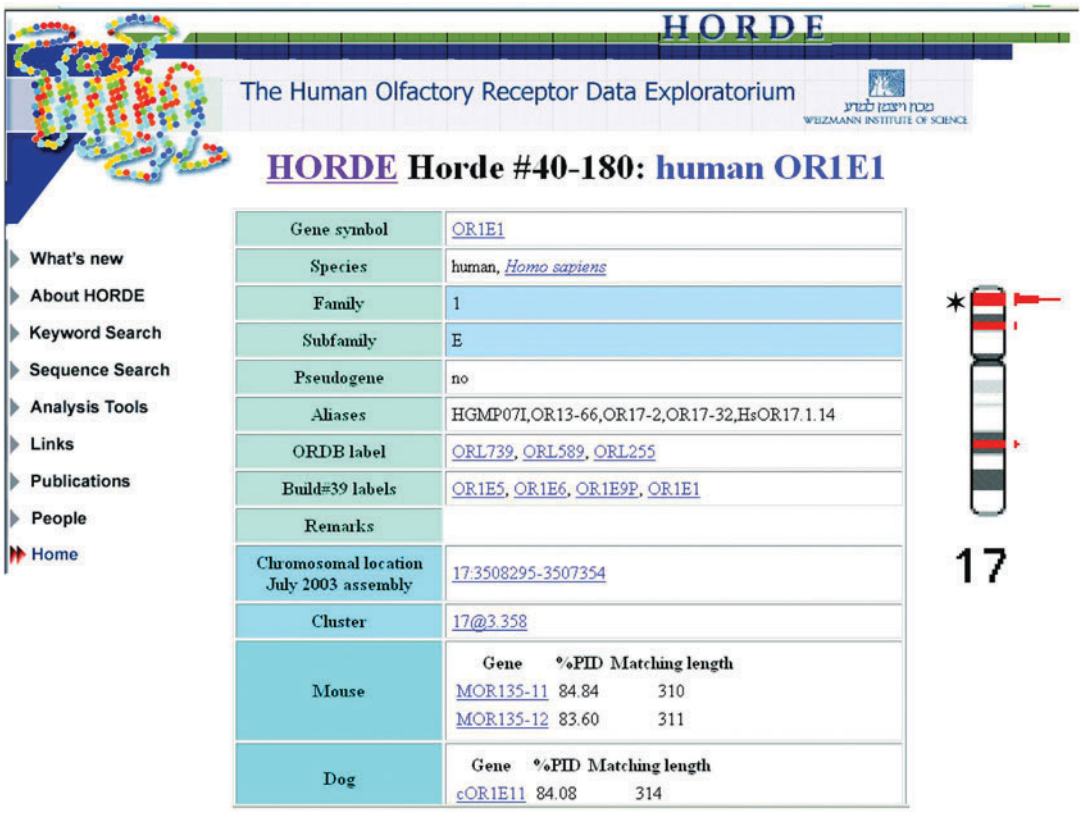

Figure 1. A partial view of the HORDE olfactory receptor (OR) card for OR1E1. On the left is a menu affording easy access to HORDE's data retrieval and analysis tools. The OR1E1 cluster location on chromosome 17 is indicated as a star on the chromosomal image on the right.

New Features

Olfactory receptor cluster information

ORs are disposed in clusters on all human chromosomes, except 20 and Y. It has been suggested that this genomic disposition is related to the regulation of OR expression, where every sensory neuron expresses only one allele of a single OR locus (Serizawa et al., 2003). Such an expression control mechanism may be essential for olfactory neuronal networking and

information processing. The *cluster viewer* is a novel feature in HORDE that facilitates browsing of OR clusters along human chromosomes. HORDE includes information on 135 clusters, ranging in size from nearly 100 ORs to singletons. The latter category constitutes 7% of all ORs, and nearly a half of these belong to subfamily 7E, an unusual group largely composed of pseudogenes (Newman and Trask, 2003). Large clusters were found to contain a higher fraction of intact ORs. Clusters with up to five ORs contain only 20% intact genes (64 such clusters containing a total of 111 OR genes), clusters with six to nine ORs contain 35% intact genes (11 such clusters containing a total of 85 OR genes), while clusters with more than 10 ORs contain an average amount of 55% intact genes (20 such clusters containing a total of 563 OR genes).

Olfactory receptor expression profiles in human tissues

ORs are nominally expressed in the olfactory epithelium, the primary sensory organ of smell. However, several publications report the expression of ORs in other tissues, such as the testis (Vanderhaeghen et al., 1997). Our newly performed integration of gene expression information into HORDE provides an experimental tool to promote future research relevant to this issue (Figure 2). HORDE provides two types of expression information - global expression patterns of 302 ORs extracted from publicly available whole-genome DNA arrays, as well as 700 ESTs and 147 mRNAs supporting the transcription potential of 215 OR genes. Both data types show a curious pattern of global OR expression in a large number of tissues. This will be described in full elsewhere (Feldmesser, E., unpublished results). The EST and mRNA genomic placements were further explored to reveal new information on the 5'- and 3'-untranslated regions (5'- and 3'-UTRs) of 154 ORs. This is a fundamental step for analyzing OR gene structure promoter regions.

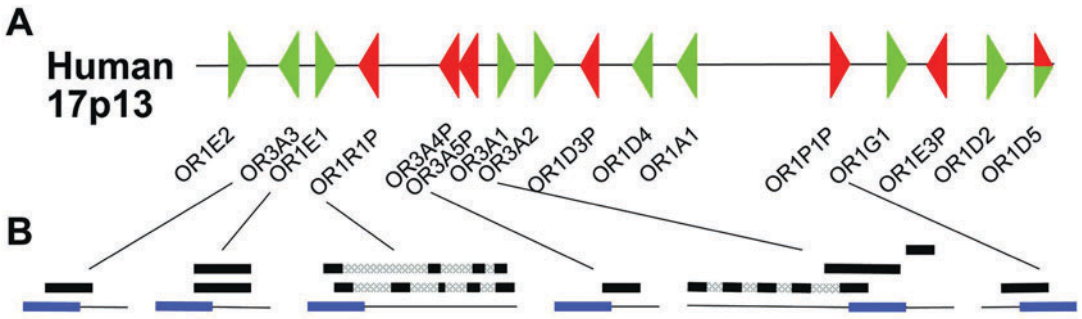

Figure 2. Transcription map of cluster 17@3.36 shown as an example. **A**, The complete cluster structure. Green triangles are intact genes and red ones are pseudogenes. The direction of the triangle indicates the transcription strand. **B**, Database-identified expressed sequence tags (ESTs) in the cluster, aligned to six of the olfactory receptors (ORs). Coding regions are shown in blue, while EST-based exons are shown in black above each coding region. Inferred introns are shown in gray (not to scale).

Single nucleotide polymorphism summaries

SNPs represent human genetic variation. In the OR context, they might have a role in the genetic variation of human olfactory sensitivity, including specific anosmia or odor blindness

(Amoore et al., 1968; Wysocki and Beauchamp, 1984). SNP information is extracted from databases in reference to the OR genomic location, and it is classified as synonymous or non-synonymous. Currently, HORDE contains a total of 1858 SNPs in 568 ORs. Four SNPs that generate in-frame stop codons are classified as segregating pseudogenes (Menashe et al., 2003).

OR structural modeling

We aimed to build an infrastructure for OR modeling and virtual odorant screening that would enable coupling of ORs to their cognate ligands. This requires the generation of accurate OR models and an examination of their capacity to predict odorant recognition. We have generated models of well-studied ORs based on the 2.8-Å structure of bovine rhodopsin (Palczewski et al., 2000). We developed a comprehensive approach for OR modeling, which incorporates the use of sequence conservation/variability signal, molecular dynamics simulations and helix kink analysis. This approach was applied to produce an improved structural model for ORs. In the improved model for one OR (the rat I7 receptor), a key lysine residue K164, which may be involved in odorant binding (Vaidehi et al., 2002), along with other putative binding residues (Man et al., 2004), all can generate a binding pocket suitable for odorant accommodation (Figure 3). In the original structural model, K164 pointed outwards from the binding site.

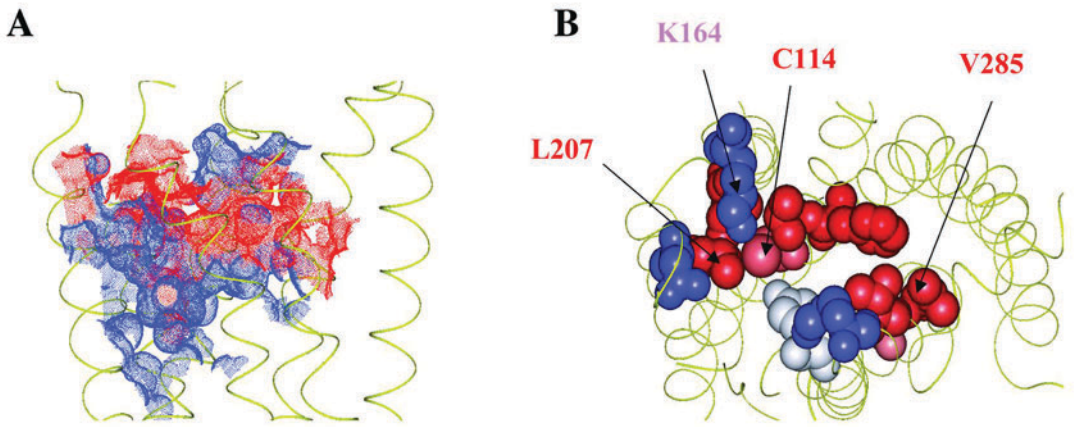

Figure 3. A three-dimensional homology model of I7 after model improvement. **A**, Side view. The red Connolly surface depicts the binding site. **B**, Top view, from the extracellular side. Residues that are known to participate in odorant binding are colored according to their hydrophobicity (blue for hydrophilic and red for hydrophobic).

We analyzed the conservation of the rhodopsin non-canonical elements, particularly proline kinks, within the OR family. In rhodopsin, these were shown to play a major role in structural and functional attributes. Specifically, the presence or absence of any kink-forming residues in transmembrane regions of the OR family may indicate a significantly different transmembrane conformation, when compared to rhodopsin. Based on a manually curated multiple alignment, we assigned a "kink pattern signature" for each OR (Figure 4).

The results of the analysis suggest that transmembrane helices H1 and H6 are not kinked in the same way as in rhodopsin, as they have conserved prolines in different positions.

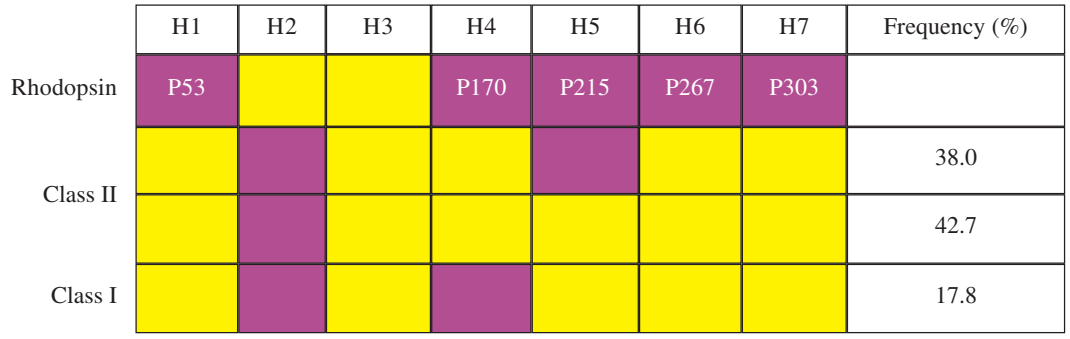

Figure 4. Proline patterns in the olfactory receptor (OR) repertoire. The figure shows the conservation/non-conservation of rhodopsin proline residues within the OR family. The top row specifies the kinked positions in rhodopsin transmembranal helices (H). Each row represents a different pattern, where purple squares are proline residues and yellow squares are others. Three patterns are predominant in the OR repertoire, two in class II, and one in class I. The abundance of each pattern is indicated on the right column.

A proline residue in H4 is conserved only among the evolutionary ancient "fish-like" (Freitag et al., 1995) class I ORs, while this position in class II ORs is not occupied by prolines or other kink-forming residues, such as S, T, C or G. These findings may indicate putative structural differences between class I and the newer tetrapod-specific class II ORs. In general, class I ORs are structurally more related to rhodopsin compared to class II ORs. This implies that straightforward modeling of ORs based on rhodopsin structure will probably generate more accurate models for class I ORs compared to class II ORs.

DISCUSSION

The availability of a centric database with the complete human OR universe, including concisely integrated information on every OR, is a fundamental asset for future studies in the field. This is demonstrated here, where the entire OR kink-residue space was explored, based on sequence information available in the HORDE. The different "kink pattern signature" yielded by class I ORs relative to class II are in-line with previous reports that "fish-like" ORs might have special functional significance in olfaction (Sun et al., 1999; Glusman et al., 2001; Olender et al., 2004). It points to a general approach for improving OR structural models according to their family signature.

Through HORDE, we discovered partial UTRs for more than 150 ORs. This is a significant progress, relative to the <30 UTRs reported so far. Future development in HORDE will focus on providing additional information on features related to genetic variation, full OR structure, OR promoters, and modes of expression.

ACKNOWLEDGMENTS

D. Lancet holds the Ralph and Lois Chair in Human Genetics. Research supported by the Crown Human Genome Center, by an Israel Ministry of Science and Technology grant to the National Knowledge Center in Genomics, and by the Abrahan and Judith Goldwasser Foundation.

REFERENCES

Amoore, J.E., Venstrom, D. and **Davis, A.R.** (1968). Measurement of specific anosmia. *Percept. Mot. Skills 26*: 143-164.

Decaillot, F.M., Befort, K., Filliol, D., Yue, S., Walker, P. and **Kieffer, B.L.** (2003). Opioid receptor random mutagenesis reveals a mechanism for G protein-coupled receptor activation. *Nat. Struct. Biol. 10*: 629-636.

Floriano, W.B., Vaidehi, N., Goddard 3rd, W.A., Singer, M.S. and **Shepherd, G.M.** (2000). Molecular mechanisms underlying differential odor responses of a mouse olfactory receptor. *Proc. Natl. Acad. Sci. USA 97*: 10712-10716.

Freitag, J., Krieger, J., Strotmann, J. and **Breer, H.** (1995). Two classes of olfactory receptors in *Xenopus laevis*. *Neuron 15*: 1383-1392.

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

Gilad, Y., Man, O. and **Glusman, G.** (2004). A comparison of the human and chimpanzee olfactory receptor gene repertoires. *Genome Res.* (in press).

Glusman, G., Sosinsky, A., Ben-Asher, E., Avidan, N., Sonkin, D., Bahar, A., Rosenthal, A., Clifton, S., Roe, B., Ferraz, C., Demaielle, J. and **Lancet, D.** (2000). Sequence, structure, and evolution of a complete human olfactory receptor gene cluster. *Genomics 63*: 227-245.

Glusman, G., Yanai, I., Rubin, I. and **Lancet, D.** (2001). The complete human olfactory subgenome. *Genome Res. 11*: 685-702.

Kent, W.J. (2002). BLAT - the BLAST-like alignment tool. *Genome Res. 12*: 656-664.

Man, O., Gilad, Y. and **Lancet, D.** (2004). Prediction of the odorant binding site of olfactory receptor proteins by human-mouse comparisons. *Protein Sci. 13*: 240-254.

Menashe, I., Man, O., Lancet, D. and **Gilad, Y.** (2003). Different noses for different people. *Nat. Genet. 34*: 143-144.

Newman, T. and **Trask, B.J.** (2003). Complex evolution of 7E olfactory receptor genes in segmental duplications. *Genome Res. 13*: 781-793.

Niimura, Y. and **Nei, M.** (2003). Evolution of olfactory receptor genes in the human genome. *Proc. Natl. Acad. Sci. USA 100*: 12235-12240.

Olender, T., Fuchs, T., Linhart, C., Shamir, R., Adams, M., Kalush, F., Khen, M. and **Lancet, D.** (2004). The canine olfactory subgenome. *Genomics* (in press).

Palczewski, K., Kumasaka, T., Hori, T., Behnke, C.A., Motoshima, H., Fox, B.A., Le Trong, I., Teller, D.C., Okada, T., Stenkamp, R.E., Yamamoto, M. and **Miyano, M.** (2000). Crystal structure of rhodopsin: A G protein-coupled receptor. *Science 289*: 739-745.

Safran, M., Chalifa-Caspi, V., Shmueli, O., Olender, T., Lapidot, M., Rosen, N., Shmoish, M., Peter, Y., Glusman, G., Feldmesser, E., Adato, A., Peter, I., Khen, M., Atarot, T., Groner, Y. and **Lancet, D.** (2003). Human gene-centric databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE. *Nucleic Acids Res. 31*: 142-146.

Sansom, M.S. and **Weinstein, H.** (2000). Hinges, swivels and switches: the role of prolines in signalling via transmembrane alpha-helices. *Trends Pharmacol. Sci. 21*: 445-451.

Serizawa, S., Miyamichi, K., Nakatani, H., Suzuki, M., Saito, M., Yoshihara, Y. and **Sakano, H.** (2003). Negative feedback regulation ensures the one receptor-one olfactory neuron rule in mouse. *Science 302*: 2088-2094.

Sun, H., Kondo, R., Shima, A., Naruse, K., Hori, H. and **Chigusa, S.I.** (1999). Evolutionary analysis of putative olfactory receptor genes of medaka fish, *Oryzias latipes*. *Gene 231*: 137-145.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and **Higgins, D.G.** (1997). The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res. 24*: 4876-4882.

Vaidehi, N., Floriano, W.B., Trabanino, R., Hall, S.E., Freddolino, P., Choi, E.J., Zamanakos, G. and **Goddard 3rd, W.A.** (2002). Prediction of structure and function of G protein-coupled receptors. *Proc. Natl. Acad. Sci. USA 99*: 12622-12627.

Vanderhaeghen, P., Schurmans, S., Vassart, G. and **Parmentier, M.** (1997). Molecular cloning and chromosomal mapping of olfactory receptor genes expressed in the male germ line: evidence for their wide distribution in the human genome. *Biochem. Biophys. Res. Commun. 237*: 283-287.

Wysocki, C.J. and **Beauchamp, G.K.** (1984). Ability to smell androstenone is genetically determined. *Proc. Natl. Acad. Sci. USA 81*: 4899-4902.

Young, J.M., Friedman, C., Williams, E.M., Ross, J.A., Tonnes-Priddy, L. and **Trask, B.J.** (2002). Differ-

ent evolutionary processes shaped the mouse and human olfactory receptor gene families. *Hum. Mol. Genet. 11*: 1683.

Zhang, X. and **Firestein, S.** (2002). The olfactory receptor gene superfamily of the mouse. *Nat. Neurosci. 5*: 124-133.

Zozulya, S., Echeverri, F. and **Nguyen, T.** (2001). The human olfactory receptor repertoire. *Genome Biol. 2*: research0018.