Review Article

Human Olfactory Receptor Comparative Sequence Assessments from the Genome

Luis N Marenco[1] and Chiquito J Crasto[2]*

[1]Evolvable Solutions, USA

[2]Center for Biotechnology and Genomics, Texas Tech University, USA

Abstract

We present a review of the results of research to identify olfactory receptor gene sequences from the Human Genome. This effort began soon after the publication of the first draft of the genome. Subsequent publications of the genome have allowed the evolution of OR gene identification with improved accuracy. We performed a comprehensive sequence analysis to identify 100% sequence matches, as well as partial matches by performing the same sequence matching step against random partial regions of sequences. Only sequences supplied to us from the groups that mined the olfactory receptor sequences from the genome or were available online were used in the analysis. The results are publicly available at https://ordb.biotech.ttu.edu/ORDB/info/humanorseqanal. We also performed a similar assessment of mouse olfactory receptor sequences from two groups that assessed the mouse genome, publicly available at https://ordb.biotech.ttu.edu/ORDB/info/mouseorseqanal. Both resources are linked through a list of potential orthologs between the two species' receptor sequences. This review also contains a discussion of nomenclatures of these genes and the resources as information dissemination portals.

*Corresponding author: Chiquito J Crasto, Center for Biotechnology and Genomics, Texas Tech University, Lubbock, Texas, USA

Olfactory Receptor Database

The Olfactory Receptor Database (ORDB) is a resource for genomic and proteomic information related to olfactory and other chemosensory receptors (insect olfactory receptors, fungal pheromone receptors, taste receptors, and vomeronasal receptors). This resource is currently housed at Texas Tech University and maintained by the authors at the Center for Biotechnology and Genomics. The resource is publicly available at https://ordb.biotech.ttu.edu/ORDB.

Its two sister databases, OdorDB (https://ordb.biotech.ttu.edu/odordb) and OdorModelDB (https://ordb.biotech.ttu.edu/oRmodeldb) can also be accessed at the biotech.ttu.edu web link. OdorDB is a repository of odorant organic molecules that have been experimentally shown to activate or inhibit receptor activity. ORModelDB is a resource that houses computational protein models of chemosensory receptors. These include protein models, the results of computational docking, or links to movies following OR-odorant interaction simulation studies.

These databases were part of the SenseLab suite of databases housed at the Yale Center for Medical Informatics, Yale University School of Medicine: that included neuronal databases: CellPropDB, NeuronDB, ModelDB, OdorMapDB, 3DModelDB and MicrocircuitDB. The development of ORDB, OdorDB, and ORModelDB continues under the auspices of the Center for Biotechnology and Genomics at Texas Tech University and is no longer a part of SenseLab.

The Impact of the Discovery of Olfactory Receptors

The identification of olfactory receptor genes followed soon after the publication of the first draft of the human genome [1,2]. The impetus for this effort was bolstered by two primary recognitions: the Nobel Prize in Physiology and Medicine in 2004 [3] to Drs. Linda Buck and

Richard Axel for discovering the first olfactory receptor gene [4]; and the knowledge that olfactory receptors constitute superfamilies of genes numbering in the several hundred. With the subsequent publications of other genomes, the presence of large olfactory gene repertoires was confirmed [5-11] Three publications highlighted the effort to mine the human olfactory receptors in 2001, [12,13] and the fourth, soon after, in 2003 [14]. Olfactory receptors (and other chemosensory receptors) have been deposited in ORDB since its creation 25 years ago. [15,16] ORDB population was automated through the web-based software, AutoPop [17], which facilitated the rapid and remote population of the database.

Summary of the OR sub-genome mining efforts

The efforts of Glusman et al., Zozulya et al., Croning, and Masatoshi and Nei are summarized in the webpage, https://ordb. biotech.ttu.edu/ORDB/info/humanorseqanal. The primary driver for the development of this resource was to unify the olfactory gene identification by mapping these gene names to a primary centralized identifier. The web page consists of six columns titled: Consensus Label, HORDE (Human Olfactory Receptor Data Exploratorium), Senomyx, EBI, ORDB, and Nei Lab.

The genes identified by Glusman et al. [12] (HORDE) have undergone continual evolution through automated and manual curation. In The third column lists the functional OR genes identified at the commercial enterprise Senomyx, Firmenich, a privately owned fragrance and taste company, later acquired Senomyx. The Senoymx effort identified functional genes while in HORDE and the Nei Lab., putative OR pseudogenes were also identified. To mark an OR gene as functional, Senomyx researchers identified sequences that started with a start codon (AUG) and contained seven transmembrane helical regions. A typical complete, functional sequence was 310-330 amino acids long. In the column titled EBI (European Bioinformatics Institute), the unpublished data was supplied by Dr. Michael Croning. The links to the genes direct the user to the Human Genome GPCR discovery resource owned and operated by Dr. Croning. The fifth column is entitled "ORDB." Here, the olfactory gene information is stored in the Olfactory Receptor Database. The last column is titled "Nei Lab." and represents the efforts of Dr. Yoshihito Niimura and Professor Masatoshi Nei at Penn State University. The gene and protein information from Senomyx and the Nei laboratory were supplied by the researchers and do not have a web presence.

The humanORseqanal.html web page also color codes the gene names for easy understanding of the nature of the genes. A yellow background represents pseudogenes; a green background is for partially matched sequences. This includes sequences that partially matched a complete sequence identified by another research group. A light blue background represents partially matched pseudogenes. For this resource, no additional efforts were made to identify pseudogenes: missense, nonsense, or frame-shift mutations. Information was identified as pseudogenic as identified by the suppliers of the OR gene information. Sequences with indeterminate chromosomal location are identified as:" UNKNOWN." "X" represents sequences matching multiple sequences from a different source. Exactly matching sequences with disputed chromosomal locations (possibly, the result of translocation events) are identified with a red background. A blue background is for OR genes that are disputed functionality-pseudogenic versus functional.

Olfactory Receptor Genes Mined-Functional versus Pseudogenes Per Chromosome

Table 1 represents the number of receptors mined from the drafts of the Human Genome per chromosome. The numbers in parentheses represent the pseudogenes identified. The EBI effort by Dr. Croning did not attempt to identify the ORs by chromosomal location. Indeed, these loci were identified by their clone IDs. In the humanORseqanal.html web page, these genes were tabulated according to chromosomal location following sequence alignment by the author. As mentioned in the previous section, the "Senomyx" column does not have identified pseudogenes, because the researchers only identified full-length, putatively functional OR genes. ORDB is not included in Table 1. It was populated using information from the HORDE, Senomyx, EBI, and Nei Lab. Indeed, before the publication of the OR sub-genomes, ORDB had approximately 300 human OR entries from earlier efforts which, very often, resulted in the identification of a single OR.

Table 1: The total number of OR genes identified by chromosome from HORDE, Senomyx, and Nei Laboratory. The numbers in parenthesis () represent identified pseudogenes. The numbers in [] brackets are those identified putative genes not included in the unified Consensus Name.

Chromosome	HORDE	Senomyx	Nei Lab
1	29 (10) [7]	40 (0) [4]	38 (26) [24]
2	0 (8) [0]	0 (0) [0]	0 (9) [2]
3	4 (18) [0]	4 (0) [0]	3 (20) [8]
4	0 (11) [1]	0 (0) [0]	0 (12) [1]
5	1 (2) [0]	2 (0) [1]	2 (4) [2]
6	23 (24) [6]	32 (0) [10]	17 (22) [7]
7	6 (15) [7]	9 (0) [3]	8 (17) [5]
8	1 (7) [1]	1 (0) [0]	1 (9) [0]
9	20 (13) [6]	26 (0) [1]	25 (14) [4]
10	1 (6) [0]	1 (0) [0]	1 (4) [0]
11	120 (157) [24]	171 (0) [23]	146 (92) [42]

12	3 (5) [0]	0 (0) [0]	6 (22) [12]
13	(6) (0) [0]	0 (0) [0]	0 (8) [0]
14	19 (17) [1]	20 (0) [1]	20 (21) [3]
15	2 (7) [3]	3 (0) [4]	1 (12) [5]
16	4 (3) [1]	4 (0) [3]	3 (1) [1]
17	16 (4) [7]	18 (0) [12]	11 (4) [7]
18	0 (0) [0]	0 (0) [0]	0 (2) [0]
19	16 (18) [6]	20 (0) [7]	15 (21) [9]
20	0 (0) [0]	0 (0) [0]	0 (0) [0]
21	0 (3) [0]	0 (0) [0]	0 (4) [0]
22	1 (0) [0]	1 (0) [0]	1 (0) [0]
X	1 (1) [0]	1 (0) [0]	0 (6) [0]
Y	0 (0) [0]	0 (0) [0]	0 (0) [0]

Classification Schemes and Nomenclature

The disparate nature of the nomenclature represents the challenges of categorizing the olfactory receptor genes. It also represents the complications that arise in our understanding of the role of olfactory receptors in olfaction. This is in no small part because of the substantial number of genes, with significant sequence variation and lack of uniform chromosomal distribution. The HORDE nomenclature relies on a cladistics assessment. Following a multiple sequence analysis after the genes were mined, the groups were identified by families and subfamilies. Families were identified by genes whose sequences were within 40% similarity, and a sub-family was identified for sequences that were within 60% similarity. Families were identified by a number; subfamilies were identified by a letter. The last aspect of this classification system was the number assigned to the member of the family. Thus, the HORDE name "OR10K1" means that the OR belongs to the family "10", subfamily "K" and it is the 1st member of that family and subfamily. The Human Genome Nomenclature Committee adopted the HORDE-driven naming system for The Senomyx nomenclature was primarily different from the HORDE system because it relied on chromosomal location. Also, the sequence alignment process was bootstrapped for only functional genes. The Senomyx researchers identified 347 sequences that aligned in 119 different families distributed across the human chromosomes. The exact Senomyx match for OR10K1 in HORDE is OR01.09.04. According to the Senomyx naming system, this OR is located on Chromosome 1, it belongs to family 9 and is its 4th member.

The Nei Lab naming system relied on both cladistics where members in each clade were then mapped to a chromosome and specifically, a cluster within a chromosome. This mapping is illustrated in Figure 4 of reference 14. The clades were identified by letters and the clusters were numbered by distance from the centromeres-as would be seen in an ideogram representation of a chromosome. Thus, the NEI lab homolog to OR10K1 and OR01.09.04 is HsOR1.4.4. This means that the gene is the fourth member (4) in the fourth cluster on Chromosome 1. The Hs is the two-letter abbreviation of Homo sapiens.

The Consensus label sought to unify these efforts and provide a nomenclature that was a controlled vocabulary that could map identical genes between each group's disparate naming system. Thus, the naming system was HSA (Homo sapiens), Chromosome Number, and member number. Therefore, in the humanORseqanal. html webpage, HSAOR01001 map OR10K1 (HORDE), OR01.09.04 (Senomyx), HGPCR1104 (EBI), ORL3001 (ORDB), Hs1.4.4 (NEI Lab.).

Though this was an ORDB-driven web resource, it was not biased towards the ORDB naming system. The ORDB naming system prefixed its gene names with ORL and not OR because some of the ORs with sequence attributes typical for ORs were found in tissues and organs far removed from the central olfactory system [18].

Not all OR genes mined by the groups were given the HSA, etc., designation. For the designation to be assigned, at least two groups had to have identified a complete and putatively functional gene with a high degree of similarity.

ORDB as the acknowledged, centralized repository was used to store all the OR gene information supplied by users. Where there were sequential agreements as identified on the humanORSeqanal web page, they were deposited as a single entry. Sequences that did not meet the agreement criteria were still deposited in ORDB to ensure that this resource remained comprehensive and as a service to the chemosensory research community.

The curious aspect of these sequence analysis exercises is the matches, or lack thereof, of pseudogenes between the HORDE and the Nei Lab. efforts. All pseudogene sequences were subjected to the same sequence match program–seeking full-length 100% matches, followed by systematic and random partial matches. One would expect that while full-sequence matches were not likely because of nonsense interruptions in the open reading frames, partial matches could have been found. Only seven pseudogene matches between HORDE and NEI Lab. were found across the human olfactory sub-genome. The remaining pseudogenes did not sequentially match and were deposited into ORDB as separate entries.

HumanORSeqanal.html-a knowledge dissemination portal

The humanORseqanal.html resource unifies the classification, nomenclature, and organization of the human olfactory receptor

gene family. It also serves as an information dissemination portal. This is primarily through links from the HORDE resource, HGPCR (contributed by Dr. Croning), and the ORDB database. The Senomyx and NEI Lab OR gene information was obtained through personal communication and is not accessible through the World Wide Web.

The HORDE link for a human OR includes: the family, subfamily, and gene number, identification of whether the gene is functional (or pseudogenic), the aliases to other resources, links to the gene in the HGNC (HUGO Gene Nomenclature Committee) and Genbank, the gene's chromosomal location, cluster, synteny, its closest paralogs (when sequentially compared to other human OR sequences), it's putative orthologs in dog, mouse, rat, cow, orangutan, horse and chimpanzee, the nucleotide and protein sequences, single nucleotide variations from DbSNP or the 1000 Genome Project, and protein haplotypes from the Human Genome Project, and gene models constructed from cDNA sequences.

ORDB link for an OR provides information related to the type of chemosensory receptor, the organism (in this case, human), the source tissue, chromosome number, the original Pubmed ID where this OR was identified, the common names (aliases), the source of the information, the providers of the gene and protein sequences, the length of the sequence, nucleotide and protein sequences, type of sequence (mRNA, cDNA or Genomic DNA), a SWISSPROT link, molecular models, the laboratory where functional analysis to deorphanize this gene was performed and a literature citation, and a microarray experiment link, if any. For most of the genome-determined ORs, the last few instances of information in ORDB will remain incomplete, because of the sheer bulk of available information determined through informatics-driven methods.

HGPCR provides information for the OR gene more from the proteomics perspective: Classification (primarily as rhodopsin-type GPCR), clone-location of the gene, the peptide and cDNA in the Ensembl resource, the predicted membrane protein topology (from Hidden Markov Models estimations and secondary structure prediction methods), prediction of signal peptide, links to the fingerprint regions in the proteins and protein family to which the protein belongs, the result of a multiple sequence alignment and the results as Uniprot Accession IDs. Links to patents of the sequences in EMBL and the predicted protein and cDNA sequences.

Exploration of Human Orthologs from the Mouse Olfactory Subgenome

After the mouse genome was published, two groups independently identified mouse OR genes [10,11]. An analysis similar to that provided in the humanORseqanal.html was created and can be accessed at https://ordb.biotech.ttu.edu/ORDB/info/mouseorseqanal. A comparable review of this resource like that for human ORs is beyond the scope of this review. It is worth noting, however, that the color markup in this web page is the same as that following a sequence analysis of human ORs. This includes a red background when the chromosomal locations of identical sequences are in dispute, especially between chromosomes 2 and 11. All mouse OR information regardless of disputed identities,

chromosomal location, and cd- size were deposited in ORDB. We also carried out a comprehensive multiple sequence alignment between the human and mouse OR gene sequences. High-similarity sequences were identified as orthologs. These links are provided in the mouseORseqanal.html file in the last column. Links for these human orthologs are to the humanORseqanal.html file. Additionally, because of user requests, the mouse alignment web page also has a column of common names that researchers use for each of these sequences.

Conclusion

The human (and mouse) ORSeqanal.html resources represent the efforts of several groups who contributed to showing how OR genes serve as a model for gene families: how they are categorized, distributed across chromosomes, and their evolutionarily determined functionality versus pseudogene city. We anticipate that the contributions of these genes towards our understanding of olfaction will only be strengthened as more efforts towards their deorphanization will be realized.

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409(6822): 860-921.

2. Venter JC, Adams MD, Myers EW, Li PW, Mural R J, et al. (2001) The sequence of the human genome. Science 291(5507): 1304-1351.

3. Firestein S (2005) A nobel nose: the 2004 Nobel Prize in Physiology and Medicine. Neuron 45(3): 333-338.

4. Buck L, Axel R (1991) A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. Cell 65(1): 175-187.

5. Liu A, Funan He, Shen L, Liu R, Wang Z et al. (2019) Convergent degeneration of olfactory receptor gene repertoires in marine mammals. BMC Genomics 20(1): 977.

6. Niimura Y, Matsui A, Touhara K (2014) Extreme expansion of the olfactory receptor gene repertoire in African elephants and evolutionary dynamics of orthologous gene groups in 13 placental mammals. Genome Res 24(9): 1485-1496.

7. Quignon P, Giraud M, Rimbault M, Lavigne P, Tacher S, et al. (2005) The dog and rat olfactory receptor repertoires. Genome Biol 6(10): R83.

8. Rouquier S, Blancher A, Giorgi D (2000) The olfactory receptor gene repertoire in primates and mouse: evidence for reduction of the functional fraction in primates. Proc Natl Acad Sci USA 97(6): 2870-2874.

9. Rouquier S, Giorgi D (2007) Olfactory receptor gene repertoires in mammals. Mutat Res 616(1-2): 95-102.

10. Young, JM, Friedman C, Williams EM, Ross JA, Tonnes-Priddy L, et al. (2002) Different evolutionary processes shaped the mouse and human olfactory receptor gene families. Hum Mol Genet 11(5): 535-546.

11. Zhang X, Firestein S (2002) The olfactory receptor gene superfamily of the mouse. Nat Neurosci 5(2): 124-133.

12. Glusman G, Yanai I, Rubin I, Lancet D (2001) The complete human olfactory subgenome. Genome Res 11(5): 685-702.

13. Zozulya S, Echeverri F, Nguyen T (2001) The human olfactory receptor repertoire. Genome Biol 2(6): RESEARCH0018.

14. Niimura Y, Nei M (2003) Evolution of olfactory receptor genes in the human genome. Proc Natl Acad Sci USA 100(21): 12235-12240.

15. Skoufos E, Healy MD, Singer MS, Nadkarni PM, Miller PL et al. (1999) Olfactory receptor database: a database of the largest eukaryotic gene family. Nucleic Acids Res 27(1): 343-345.

16. Skoufos E, Marenco L, Nadkarni PM, Miller PL, Shepherd GM (2000) Olfactory receptor database: a sensory chemoreceptor resource. Nucleic Acids Res 28(1): 341-343.

17. Crasto C, Marenco L, Miller P, Shepherd G (2002) Olfactory receptor database: a metadata-driven automated population from sources of gene and protein sequences. Nucleic Acids Res 30(1): 354-360.

18. Leslie M (2001) Molecular Biology of Smell. Science 293(5531): 767.