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

Highly conserved regions in the 5' region of human olfactory receptor genes

H.F. Tobar[2]*, P.A. Moreno[2,3]* and P.E. Vélez[1,2]*

[1]Department of Biology, Universidad del Cauca, Popayán, Colombia
[2]Group of Molecular, Environmental, and Cancer Biology BIMAC, Universidad del Cauca, Popayán, Colombia
[3]School of Systems and Computer Engineering, Universidad del Valle, Santiago de Cali, Colombia

*All authors contributed equally to this study.
Corresponding author: P.E. Vélez
E-mail: pvelez@unicauca.edu.co

Genet. Mol. Res. 8 (1): 117-128 (2009)
Received October 27, 2008
Accepted November 28, 2008
Published February 10, 2009

ABSTRACT. Regulation of human olfactory receptor (hOR) genes is a complex process of control and signalization with various structures and functions that are not clearly understood. To date, nearly 390 functional hOR genes and 462 pseudogenes have been discovered in the human genome. Enhancer models and trans-acting elements for the regulation of different hOR genes are among the few examples of our knowledge concerning regulation of these genes. We looked for upstream control elements that might help explain these complex control mechanisms. To analyze the human olfactory gene family, we looked for functional genes and pseudogenes common to all hOR genes obtained from public databases. Subsequently, we analyzed sequences upstream of the transcription start sites with data mining and bioinformatics tools. We found two highly conserved regions, which we called HCR I and HCR II, upstream of the transcription start sites in 77 hOR genes and 87 pseudogenes. These regions showed possible enhancer functions common to both genes and pseudogenes, an intriguing feature that may be associated with the expression of pseudogenes. Based on these HCRs, we propose a structural model of gene regulation for the olfactory gene family.

Key words: Olfactory receptor gene; Promoter region;
Highly conserved region; Regulation; Enhancers

BACKGROUND

The sense of smell in humans is an important and complex mechanism, where different molecular structures, functions, and regulatory processes converge. The integration of these characteristics allows the detection of many different odorant molecules with high sensitivity and specificity. The responsibility for the perception of smells is assigned to signal transduction proteins in the sensory neuron called olfactory receptors (OR). These proteins make up a larger gene family in many animal species (Buck and Axel, 1991). In the human genome, this gene family is one of the most representatives especially that on chromosome 11 since 40% of the whole set of human OR (hOR) genes is found in this chromosome (Taylor et al., 2006). Olfactory receptor genes have interesting characteristics that indicate a complex regulation mechanism. Some of them are the presence of a single OR gene in an olfactory neuron (Chess et al., 1994; Malnic et al., 1999), the organization of axons of the olfactory neurons according to the type of receptor that is presented (Wang et al., 1998; Feinstein and Mombaerts, 2004; Barnea et al., 2004), and the presence of functional genes and pseudogenes in different tissues (Zhang et al., 2007).

The regulation of these different structural and functional characteristics needs the integration of specific control mechanisms such as enhancers, silencers, transcription factors, repressors, activator binding sites, and possible unknown mechanisms of control and signalization. The regulation of the OR genes in mice is associated with O/E-like and homeodomain binding sites with possible enhancer function (Michaloski et al., 2006); however, it is not clearly understood. Many studies on olfactory receptor genes in mice show different conserved regions and transcription factors upstream from the transcription start sites (TSS), but it is only for a small number of genes (Qasba and Reed 1998; Sosinsky et al., 2000; Lane et al., 2001; Vassalli et al., 2002; Michaloski et al., 2006; Hoppe et al., 2000, 2003, 2006). All these works taken together show poor understanding of the upstream control elements present in each hOR. Thus, we proposed the examination of the promoter regions of the hOR genes from the HORDE (Human Olfactory Receptor Data Exploratorium) database to establish the existence of cis-acting regulatory elements with a possible role in hOR gene regulation.

MATERIAL AND METHODS

Database for functional hOR genes

The hOR genes with a non-coding 5' sequence of 1000 bp corresponding to the promoter region were downloaded from the HORDE database (Olender et al., 2004), available at the website: http://bioportal.weizmann.ac.il/HORDE/ (HG 18, November 22, 2007). The Tables Browser Tool of the UCSC Genome Browser, http://genome.ucsc.edu (Hinrichs et al., 2006) was used to download the sequences. Subsequently, we defined the extramembrane motifs in the coding region of the gene, according to Zozulya et al. (2001), and defined the functional genes and pseudogenes using the Niimura and Nei (2003) method.

Detection of regularities on the promoter region

Given the lack of similarity in these upstream regions, multiple alignments were made by using several sequences to search for regularities in the promoter region of the hOR genes. These alignments were conducted via the ClustalW (Thompson et al., 1994) and T-Coffee (Notredame et al., 2000) algorithms.

The conserved regions detected in each alignment were sought in the whole set of hOR genes, and re-aligned again. This step was necessary to find all possible genes with conserved regions.

The possible regulatory function for each conserved region was determined through the BLAST algorithm (Altschul et al., 1997) on the Transcription Regulatory Regions Database (TRRD; Kolchanov et al., 2002).

Finally, the cis-acting regulatory elements with a possible gene regulatory role were examined by means of the weeder prediction tools (Pavesi et al., 2004). The Sequence Logo Tool (Crooks et al., 2004) was used to represent the structural characteristics of each conserved region.

Analysis of chromosomal and phylogenetic clustering

The definition and classification of chromosomal and phylogenetic clustering were carried out via hOR functional genes and pseudogenes with conserved regions upstream from the transcription start sites. Chromosomal clustering was conducted using information about the gene distribution on the chromosomes or genomic distribution downloaded from the database. For each gene, we defined and implemented a localization nomenclature according to three parameters of distribution in each chromosome: regions, clustering, and sub-clustering, i.e, ≥1000 kb for regions (right R, medium M, and left L); ≥100 kb for clustering (I, II, III, IV, etc.), and ≥10 kb for sub-clustering (a, b, c, d, etc.). See Table 1.

The phylogenetic clustering was defined by constructing a phylogenetic tree with the promoter regions and coding regions of the functional genes and pseudogenes. Each phylogenetic clustering was defined according to the clades with more than 5 genes, and numbered as follows: "clustering C1, C2, C3, etc." The reconstruction of the phylogenetic tree was built using a neighbor-joining method with a bootstrapping value of 1000. The MEGA 4.0 software was downloaded from www.megasoftware.net (Tamura et al., 2007) and used for all phylogenetic analyses.

Finally, chromosomal and phylogenetic clustering was related to the conserved regions in the promoter regions.

Table 1. Nomenclature and location of the human olfactory receptor (hOR) genes and pseudogenes.

Name	Name according to ID	Chr	Start	End	Strand	HCR region	Type	Size	Intergenic distance (ID)
OR4G4P	1R	chr1	41316	43258	(+)	HCR II	Pseudogene	1942	first
OR13Z2P	1MI	chr1	145382936	145384490	(+)	HCR II	Pseudogene	1554	145339678
OR10K1	1MII	chr1	156700977	156702914	(+)	HCR I	Functional	1937	11316487
OR10R3P	1MII	chr1	156726634	156728571	(+)	HCR II	Pseudogene	1937	23720
OR10AE1P	1MIII	chr1	157818038	157819634	(-)	HCR II	Pseudogene	1596	1089467
OR9H1P	1La	chr1	246003836	246005758	(+)	HCR II	Pseudogene	1922	88184202
OR5AT1	1La	chr1	246044729	246046654	(-)	HCR I	Functional	1925	38971
OR11L1	1La	chr1	246070857	246072821	(-)	HCR I	Functional	1964	24203
OR2L8	1Lb	chr1	246177784	246179718	(+)	HCR II	Functional	1934	104963
OR2AK2	1Lb	chr1	246194303	246196261	(+)	HCR I	Functional	1958	14585
OR2L1P	1Lb	chr1	246219193	246221116	(+)	HCR I	Pseudogene	1923	22932
OR2AS2P	1Lc	chr1	246727103	246728583	(+)	HCR II	Pseudogene	1480	505987
OR7E46P	2R	chr2	71117360	71119303	(+)	HCR II	Pseudogene	1943	first
OR7E62P	2R	chr2	71134777	71136796	(+)	HCR I	Pseudogene	2019	15474
OR7E102P	2L	chr2	95575054	95577048	(+)	HCR II	Pseudogene	1994	24438258
OR7E122P	3R	chr3	8703921	8705840	(+)	HCR I	Pseudogene	1919	first
OR7E66P	3Ma	chr3	75479689	75481604	(-)	HCR I	Pseudogene	1915	66773849
OR7E22P	3Ma	chr3	75488327	75490351	(-)	HCR I	Pseudogene	2024	6723
OR7E55P	3Ma	chr3	75502255	75504204	(-)	HCR I	Pseudogene	1949	11904
OR7E121P	3Mb	chr3	75730493	75732514	(-)	HCR I	Pseudogene	2021	226289
OR7E100P	3LI	chr3	113725724	113727737	(-)	HCR I	Pseudogene	2013	37993210
OR7E130P	3LII	chr3	126903910	126905871	(+)	HCR I	Pseudogene	1961	13176173
OR7E29P	3LII	chr3	126912660	126914603	(+)	HCR I	Pseudogene	1943	6789
OR7E93P	3LII	chr3	126925014	126927047	(+)	HCR I	Pseudogene	2033	10411
OR7E53P	3LII	chr3	126934770	126936777	(+)	HCR I	Pseudogene	2007	7723
OR7E97P	3LII	chr3	126947609	126949642	(+)	HCR I	Pseudogene	2033	10832
OR7E129P	3LIII	chr3	131223090	131225113	(-)	HCR I	Pseudogene	2023	4273448
OR7E21P	3LIII	chr3	131236023	131238075	(-)	HCR I	Pseudogene	2052	10910
OR7E99P	4I	chr4	4209108	4211131	(-)	HCR I	Pseudogene	2023	first
OR7E43P	4I	chr4	4226950	4228918	(-)	HCR II	Pseudogene	1968	15819
OR7E85P	4II	chr4	9093449	9095475	(+)	HCR I	Pseudogene	2026	4864531
OR2V1	5	chr5	180483967	180485910	(-)	HCR I	Functional	1943	first
OR2V2	5	chr5	180513550	180515493	(+)	HCR II	Functional	1943	27640
OR2B2	6RI	chr6	27987007	27989076	(-)	HCR II	Functional	2069	first
OR2W6P	6RI	chr6	28012161	28014157	(+)	HCR I	Pseudogene	1996	23085
OR12D2	6RII	chr6	29471457	29473376	(+)	HCR I	Functional	1919	1457300
OR12D1P	6RII	chr6	29492037	29493964	(+)	HCR II	Pseudogene	1927	18661
OR11A1	6RII	chr6	29502454	29504397	(-)	HCR I	Functional	1943	8490
OR2A4	6L	chr6	132063306	132065234	(-)	HCR I	Functional	1928	102558909
OR10AH1P	7R	chr7	5122247	5124244	(+)	HCR I	Pseudogene	1997	first
OR9A3P	7LI	chr7	141208130	141210088	(+)	HCR I	Pseudogene	1958	136083886
OR9A4	7LI	chr7	141264146	141266086	(+)	HCR I	Functional	1940	54058
OR6V1	7LIIa	chr7	142458561	142460498	(+)	HCR I	Functional	1937	1192475
OR2A41P	7LIIb	chr7	143404420	143405880	(+)	HCR I	Pseudogene	1460	943922
OR2A7	7LIIc	chr7	143586726	143588654	(-)	HCR I	Functional	1928	180846
OR7E158P	8a	chr8	11814815	11816713	(-)	HCR I	Pseudogene	1898	first
OR7E161P	8a	chr8	11823487	11825481	(-)	HCR I	Pseudogene	1994	6774
OR7E160P	8b	chr8	11928536	11930560	(-)	HCR I	Pseudogene	2024	103055
OR7E8P	8c	chr8	12586021	12588045	(-)	HCR I	Pseudogene	2024	655461
OR7E15P	8c	chr8	12598175	12600144	(-)	HCR I	Pseudogene	1969	10130
OR7E10P	8c	chr8	12604933	12606918	(-)	HCR I	Pseudogene	1985	4789
OR13C6P	9R	chr9	35981337	35983282	(-)	HCR II	Pseudogene	1945	first
OR2AM1P	9R	chr9	36010709	36012112	(+)	HCR I	Pseudogene	1403	27427
OR7E116P	9LI	chr9	92033201	92035225	(+)	HCR I	Pseudogene	2024	56021089
OR13C8	9LII	chr9	106370271	106372229	(+)	HCR II	Functional	1958	14335046
OR2K2	9LIII	chr9	113129588	113131534	(-)	HCR I	Functional	1946	6757359

Continued on next page

Table 1. Continued.

Name	Name according to ID	Chr	Start	End	Strand	HCR region	Type	Size	Intergenic distance (ID)
OR1L1	9LIV	chr9	124462817	124464745	(+)	HCR II	Functional	1928	11331283
OR1L3	9LIV	chr9	124476231	124478201	(+)	HCR I	Functional	1970	11486
OR1L6	9LIV	chr9	124550949	124552880	(+)	HCR II	Functional	1931	72748
OR5C1	9LIV	chr9	124590034	124591992	(+)	HCR II	Functional	1958	37154
OR1K1	9LIV	chr9	124601224	124603170	(+)	HCR II	Functional	1946	9232
OR7E110P	10R	chr10	15068882	15070476	(-)	HCR I	Pseudogene	1594	first
OR7E26P	10R	chr10	15081109	15083090	(-)	HCR I	Pseudogene	1981	10633
OR7E115P	10R	chr10	15089878	15091884	(-)	HCR I	Pseudogene	2006	6788
OR13A1	10L	chr10	45118894	45120819	(-)	HCR I	Functional	1925	30027010
OR7E12P	11RIa	chr11	3368617	3370555	(-)	HCR I	Pseudogene	1938	first
OR7E117P	11RIb	chr11	3577451	3579475	(-)	HCR I	Pseudogene	2024	206896
OR51D1	11RIIa	chr11	4616598	4618568	(+)	HCR I	Functional	1970	1037123
OR51A9P	11RIIa	chr11	4638639	4640536	(-)	HCR II	Pseudogene	1897	20071
OR51F3P	11RIIa	chr11	4713985	4715884	(-)	HCR I	Pseudogene	1899	73449
OR51F1	11RIIa	chr11	4746789	4748723	(-)	HCR I	Functional	1934	30905
OR52R1	11RIIa	chr11	4781243	4783186	(-)	HCR I	Functional	1943	32520
OR51A7	11RIIb	chr11	4884177	4886111	(+)	HCR II	Functional	1934	100991
OR51P1P	11RIIc	chr11	4991945	4993882	(+)	HCR II	Pseudogene	1937	105834
OR52A5	11RIId	chr11	5109502	5111448	(-)	HCR I	Functional	1946	115620
OR52Z1P	11RIId	chr11	5155527	5157416	(-)	HCR I	Pseudogene	1889	44079
OR51J1	11RIIe	chr11	5379404	5381350	(+)	HCR I	Functional	1946	221988
OR52B6	11RIIf	chr11	5557747	5559687	(+)	HCR I	Functional	1940	176397
OR52N4	11RIIg	chr11	5731548	5733509	(+)	HCR II	Functional	1961	171861
OR56B4	11RIIh	chr11	6084586	6086541	(+)	HCR I	Functional	1955	351077
OR4X2	11MIa	chr11	48222233	48224140	(+)	HCR II	Functional	1907	42135692
OR4C5	11MIb	chr11	48343617	48345593	(-)	HCR II	Functional	1976	119477
OR4C10P	11MIb	chr11	48410349	48412301	(-)	HCR I	Pseudogene	1952	64756
OR4C46	11MII	chr11	51370859	51372784	(+)	HCR I	Functional	1925	2958558
OR7E5P	11MIIIa	chr11	55503167	55505158	(-)	HCR I	Pseudogene	1991	4130383
OR8H2	11MIIIb	chr11	55628096	55630030	(+)	HCR I	Functional	1934	122938
OR8H3	11MIIIb	chr11	55645426	55647360	(+)	HCR I	Functional	1934	15396
OR8J3	11MIIIb	chr11	55660827	55662770	(-)	HCR I	Functional	1943	13467
OR5M8	11MIIIc	chr11	56014491	56016422	(-)	HCR I	Functional	1931	351721
OR5M10	11MIIIc	chr11	56100830	56102773	(-)	HCR I	Functional	1943	84408
OR5AP2	11MIIIc	chr11	56165545	56167494	(-)	HCR I	Functional	1949	62772
OR9G4	11MIIId	chr11	56266884	56268818	(-)	HCR II	Functional	1934	99390
OR5BQ1P	11MIIIe	chr11	56552439	56553773	(+)	HCR II	Pseudogene	1334	283621
OR5AZ1P	11MIIIf	chr11	57441351	57443278	(-)	HCR I	Pseudogene	1927	887578
OR5BD1P	11MIIIf	chr11	57469604	57471519	(-)	HCR II	Pseudogene	1915	26326
OR9Q1	11MIIIg	chr11	57702494	57704422	(+)	HCR I	Functional	1928	230975
OR10Q1	11MIIIg	chr11	57751968	57753923	(-)	HCR II	Functional	1955	47546
OR10Q2P	11MIIIg	chr11	57815878	57817819	(-)	HCR II	Pseudogene	1941	61955
OR5AN1	11MIVa	chr11	58887509	58889440	(+)	HCR I	Functional	1931	1069690
OR4D11	11MIVb	chr11	59026626	59028557	(+)	HCR II	Functional	1931	137186
OR4D9	11MIVb	chr11	59037963	59039903	(+)	HCR II	Functional	1940	9406
OR4D7P	11MIVb	chr11	59054747	59056759	(+)	HCR II	Pseudogene	2012	14844
OR10V1	11MIVc	chr11	59236969	59238894	(-)	HCR II	Functional	1925	180210
OR7E145P	11MVa	chr11	67246507	67248531	(-)	HCR I	Pseudogene	2024	8007613
OR7E11P	11MVa	chr11	67259645	67261594	(-)	HCR I	Pseudogene	1949	11114
OR7E1P	11MVb	chr11	67498283	67500304	(-)	HCR I	Pseudogene	2021	236689
OR7E87P	11MVIa	chr11	70981089	70983107	(+)	HCR I	Pseudogene	2018	3480785
OR7E4P	11MVIa	chr11	71007708	71009726	(+)	HCR I	Pseudogene	2018	24601
OR7E128P	11MVIb	chr11	71281102	71283132	(+)	HCR I	Pseudogene	2030	271376
OR7E126P	11MVIb	chr11	71290880	71292813	(+)	HCR I	Pseudogene	1933	7748
OR6M2P	11La	chr11	123216877	123218821	(-)	HCR I	Pseudogene	1944	51924064
OR4D5	11Lb	chr11	123314535	123316487	(+)	HCR I	Functional	1952	95714
OR10D5P	11Lc	chr11	123429711	123431643	(+)	HCR I	Pseudogene	1932	113224
OR8B3	11Ld	chr11	123771520	123773457	(-)	HCR II	Functional	1937	339877

Continued on next page

Table 1. Continued.

Name	Name according to ID	Chr	Start	End	Strand	HCR region	Type	Size	Intergenic distance (ID)
OR8A2P	11Ld	chr11	123834671	123836632	(+)	HCR II	Pseudogene	1961	61214
OR7E148P	12R	chr12	8471526	8473477	(-)	HCR I	Pseudogene	1951	first
OR7E149P	12R	chr12	8481267	8483246	(-)	HCR I	Pseudogene	1979	7790
OR5BT1P	12LIa	chr12	47065402	47067355	(-)	HCR I	Pseudogene	1953	38582156
OR8S1	12LIb	chr12	47204683	47206617	(+)	HCR I	Functional	1934	137328
OR5BS1P	12LIb	chr12	47238933	47240864	(+)	HCR II	Pseudogene	1931	32316
OR7E47P	12LII	chr12	50786370	50788314	(+)	HCR I	Pseudogene	1944	3545506
OR6C74	12LIII	chr12	53926340	53928274	(+)	HCR I	Functional	1934	3138026
OR6C1	12LIII	chr12	53999652	54001586	(+)	HCR II	Functional	1934	71378
OR6C75	12LIII	chr12	54044163	54046097	(+)	HCR I	Functional	1934	42577
OR6C65	12LIII	chr12	54079581	54081515	(+)	HCR I	Functional	1934	33484
OR7E36P	13R	chr13	40903401	40905394	(-)	HCR I	Pseudogene	1993	first
OR7E155P	13R	chr13	40911974	40913998	(-)	HCR I	Pseudogene	2024	6580
OR7E111P	13L	chr13	67373378	67375357	(+)	HCR I	Pseudogene	1979	26459380
OR7E33P	13L	chr13	67382135	67384081	(+)	HCR I	Pseudogene	1946	6778
OR11H2	14Ra	chr14	19250939	19252882	(-)	HCR I	Functional	1943	first
OR4K3P	14Rb	chr14	19406200	19408142	(-)	HCR I	Pseudogene	1942	153318
OR4K1	14Rb	chr14	19472667	19474598	(+)	HCR I	Functional	1931	64525
OR4K17	14Rc	chr14	19654500	19656434	(+)	HCR I	Functional	1934	179902
OR11H5P	14Rc	chr14	19746297	19748194	(+)	HCR II	Pseudogene	1897	89863
OR7E105P	14L	chr14	51292459	51294453	(+)	HCR I	Pseudogene	1994	31544265
OR4G6P	15	chr15	100295277	100297219	(-)	HCR II	Pseudogene	1942	unique
OR2C1	16	chr16	3344943	3346877	(+)	HCR II	Functional	1934	unique
OR1P1P	17a	chr17	3003938	3005896	(-)	HCR I	Pseudogene	1958	first
OR1R1P	17b	chr17	3234975	3236915	(+)	HCR II	Pseudogene	1940	229079
OR1E1	17b	chr17	3247514	3249454	(-)	HCR I	Functional	1940	10599
OR3A3	17b	chr17	3269631	3271574	(+)	HCR I	Functional	1943	20177
OR1E2	17b	chr17	3282918	3284885	(-)	HCR II	Functional	1967	11344
OR4G3P	19I	chr19	44063	46005	(+)	HCR II	Pseudogene	1942	first
OR4F17	19I	chr19	60680	62593	(+)	HCR I	Functional	1913	14675
OR2Z1	19IIa	chr19	8701392	8703332	(+)	HCR I	Functional	1940	8638799
OR1M4P	19IIb	chr19	9054613	9056074	(-)	HCR I	Pseudogene	1461	351281
OR1M1	19IIb	chr19	9063922	9065859	(+)	HCR II	Functional	1937	7848
OR7G2	19IIb	chr19	9073949	9075919	(-)	HCR II	Functional	1970	8090
OR7G1	19IIb	chr19	9086508	9088439	(-)	HCR II	Functional	1931	10589
OR7G15P	19IIb	chr19	9093761	9095163	(-)	HCR II	Pseudogene	1402	5322
OR7D2	19IIb	chr19	9156459	9158393	(+)	HCR I	Functional	1934	61296
OR7E25P	19IIb	chr19	9174929	9176908	(+)	HCR I	Pseudogene	1979	16536
OR7D4	19IIb	chr19	9185579	9187513	(-)	HCR I	Functional	1934	8671
OR7E24	19IIb	chr19	9221721	9223736	(+)	HCR I	Functional	2015	34208
OR7C1	19IIIa	chr19	14770990	14772948	(-)	HCR I	Functional	1958	5547254
OR7A5	19IIIa	chr19	14799098	14801053	(-)	HCR I	Functional	1955	26150
OR7A10	19IIIa	chr19	14812764	14814689	(-)	HCR II	Functional	1925	11711
OR7A11P	19IIIa	chr19	14887196	14889213	(+)	HCR I	Pseudogene	2017	72507
OR7C2	19IIIa	chr19	14912302	14914257	(+)	HCR II	Functional	1955	23089
OR1I1	19IIIb	chr19	15057878	15059821	(+)	HCR I	Functional	1943	143621
OR10H3	19IIIc	chr19	15712204	15714150	(+)	HCR II	Functional	1946	652383
OR10H5	19IIIc	chr19	15764860	15766803	(+)	HCR II	Functional	1943	50710
OR10H1	19IIIc	chr19	15778895	15780847	(-)	HCR I	Functional	1952	12092
OR1AB1P	19IIId	chr19	16022788	16024525	(+)	HCR II	Pseudogene	1737	241941

RESULTS

hOR functional genes and pseudogenes

A total of 851 hOR gene sequences were downloaded from the HORDE database, from which 389 are functional genes and 462 are pseudogenes. These values correspond to those found in known databases and previously reported studies (Zozulya et al., 2001; Glusman et al., 2001; Malnic et al., 2004; Niimura and Nei, 2003).

Regularities in the promoter regions of the hOR genes

Our approach resulted in the discovery of two highly conserved regions (HCR) upstream from the TSS. We designated these regions HCR I and HCR II for 111 and 53 genes, respectively. They represent a structural model of gene regulation for the hOR gene (Figure 1). This number of genes represents 19% of the hOR gene family, and each HCR is present in several functional genes and pseudogenes. Each HCR has different structural characteristics as shown in Table 2.

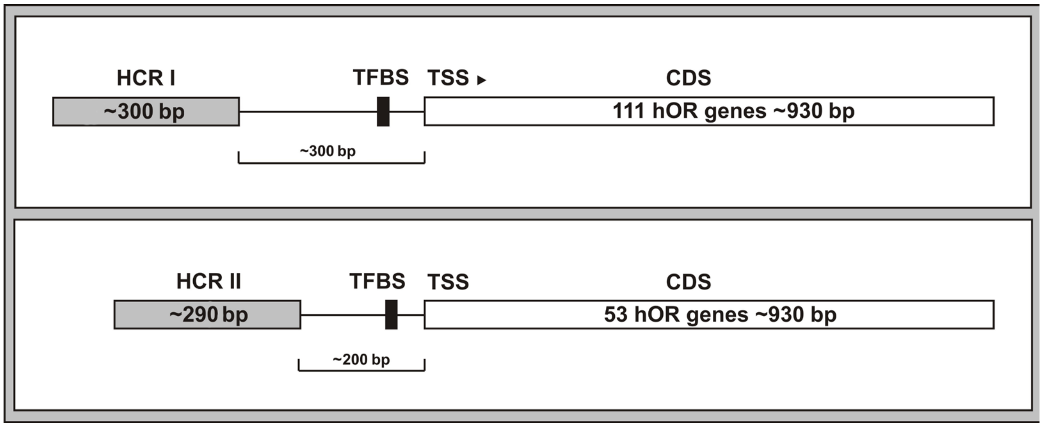

Figure 1. Structural model of gene regulation for the human olfactory gene family. Structural model of human olfactory receptor (hOR) genes with the highly conserved regions (HCR I and HCR II). The blocks represent the structure of the gene with its average length. Each gene has one HCR and a possible transcription factor binding site (TFBS) upstream from the transcription start sites (TSS).

Table 2. Structural characteristics of the highly conserved regions, HCR I and HCR II.

Region	Number of functional genes	Number of pseudogenes	Percentage of similarity	HCR length	Distance HCR-TSS
HCR I	49	62	≈75%	≈300 bp	≈300 bp
HCR II	28	25	≈70%	≈290 bp	≈200 bp

TSS = transcription start sites.

To confirm that these regions are not part of a possible exon in an interrupted gene, we conducted an exploratorium search of the hOR receptor interrupted genes using the gbk file from GenBank (Human, May 2004 - hg 17, NCBI Build 35). We found that only 5 genes (OR3A2, OR52E5, OR56A3, OR7G1, and OR8S1) are not single-exon.

The BLAST search on the TRRD (Kolchanov et al., 2002) allowed for the discovery of many different and conserved motifs in each HCR, as shown in Figure 2. However, the most important motifs were predicted via the weeder prediction tool (Pavesi et al., 2004). These motifs represent O/E-like and homeodomain binding sites in about 90% of genes with an HCR. It is noteworthy that the HCR II motifs were also defined in the olfactory receptor genes in mice and represent possible negative control function of the gene (Michaloski et al., 2006; Hoppe et al., 2006).

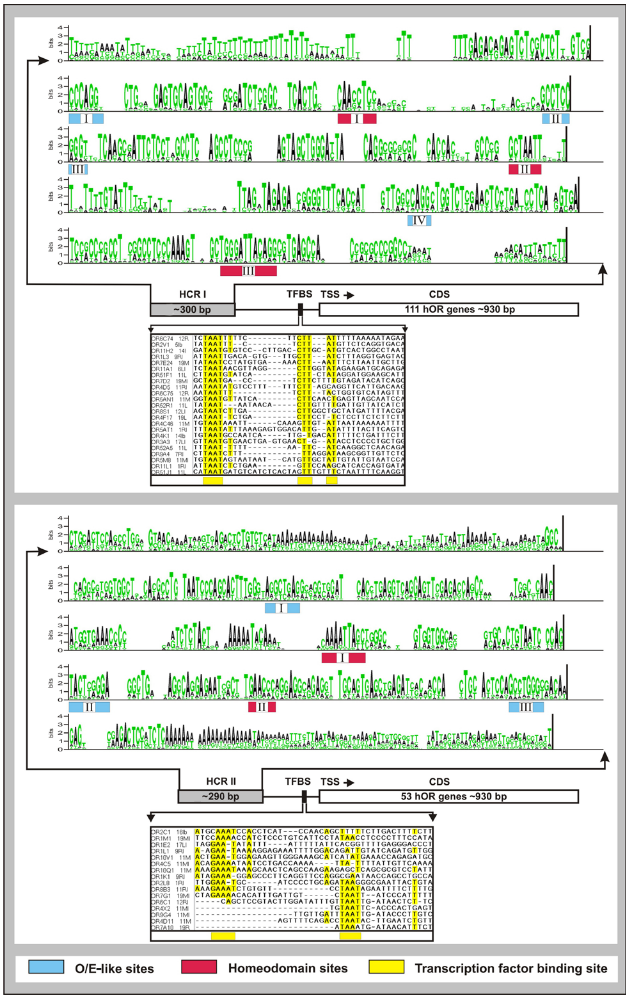

Figure 2. Highly conserved regions (HCR). Highly conserved regions with the most representative motifs and the most conserved transcription factors in some genes. TFBS = transcription factor binding sites; TSS = transcription start sites.

Michaloski et al. (2006) associated the O/E-like and homeodomain motifs with a possible enhancer function, as witnessed herein by comparing the HCR sequences with TRRD. This possible function was validated by looking for particular transcription factor sites for each HCR, -100 bp from the TSS of each gene. We used the Hctata Tool (http://zeus2.itb.cnr.it/~webgene/wwwHC_tata.html) to predict any common TATA-box in the sequence. A multiple alignment using the T-Coffee tool, and visual inspection then permitted the definition of the transcription factor binding sites in the sequences. As a result of this process, different TATA-box and transcription factor binding sites were found in different groups of genes with HCR I and HCR II, as shown in Figure 2. The number of genes with these transcriptional factor binding sites is about 46% for HCR I and 35% for HCR II. These results indicate that the HCR may act on different transcription factors and in different positions, as previously reported in mice by Michaloski et al. (2006).

Chromosomal and phylogenetic clustering analyses

The cluster distribution we established was similar to that previously reported for the complete set of hOR genes (Zozulya et al., 2001; Glusman et al., 2001; Niimura and Nei, 2003). We defined 33 chromosomal clusterings, as shown in Figure 3A. Gene distribution in the regions is not a particular characteristic for chromosomes 4, 5, 8, 17, and 19; however, the distribution in sub-clusters is found on chromosomes 1, 3, 7, 8, 11, 12, 14, 17, and 19. Chromosome 11 has the largest number of genes with an HCR, and these genes are distributed in many clusters and sub-clusters mainly in the R and M regions of the chromosome.

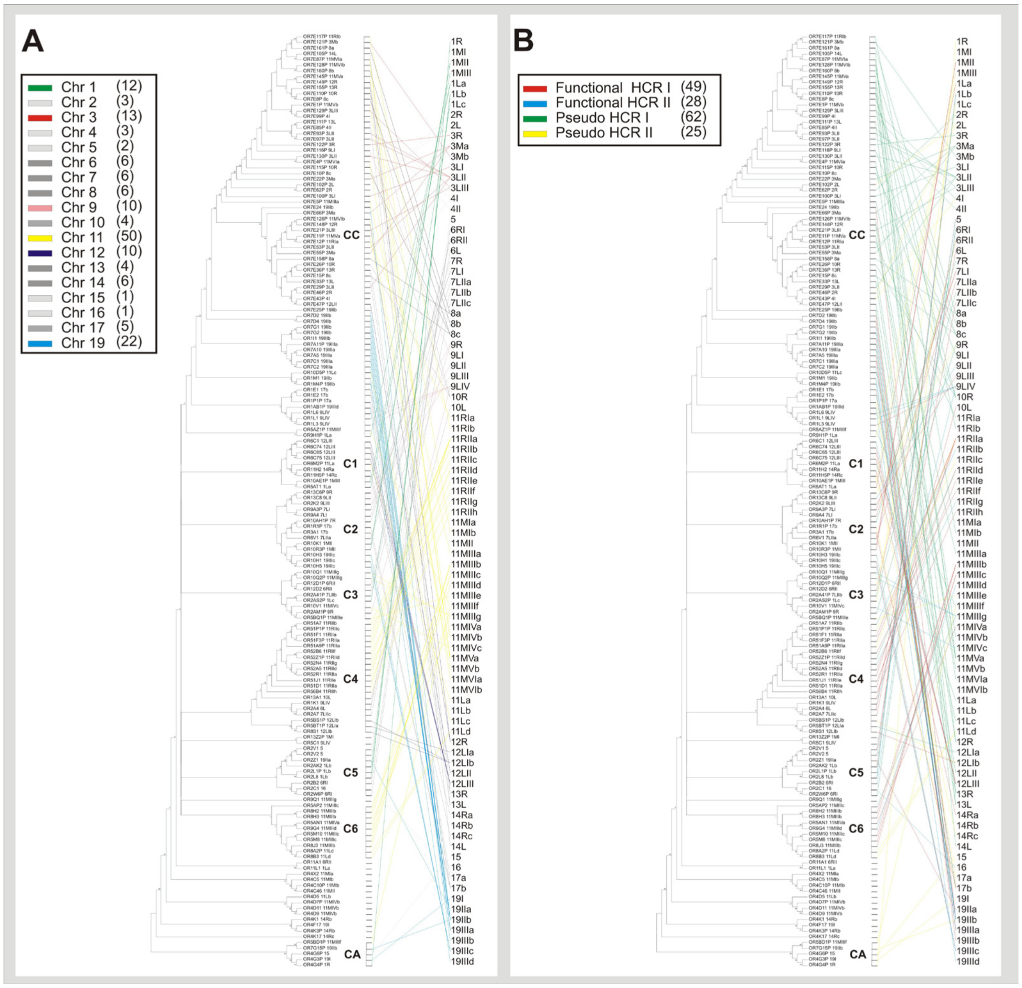

Figure 3. Relationship between phylogenetic and chromosomal clustering and the higly conserved regions (HCR). Phylogenetic tree reconstructed using a neighbor-joining method with a bootstrapping value of 1000. **A.** Phylogenetic clustering and chromosomal clustering of the genes are joined by lines with different colors by each chromosome. **B.** Gene distribution of each HCR on the phylogenetic tree and chromosome location.

We found 8 phylogenetic clusterings for all genes with HCR, which were named and enumerated. Two of these clusters were re-named as CA and CC clusterings. CA corresponds to the ancestral cluster with more than 5 genes, and CC corresponds to the conserved cluster with a large number of genes (49 genes).

All phylogenetic clustering was associated with chromosomal clustering, as shown in Figure 3A, to find a relationship between them. We used this approach to detect relationships between HCR I and HCR II and their possible roles in the function of the gene and their chromosomal distribution (Figure 3B). We found no relationship between the phylogenetic tree topology and most genes with HCR I and HCR II, even with similarity >70% in most conserved motifs. This finding suggests that an HCR is not a common feature for a specific gene with the same function and chromosome location. However, the presence of these regions in 25% of the functional genes is a descriptor parameter to distinguish some members of the OR gene family. It is important to note that of the 62 pseudogenes with an HCR I, 45 are in the CC phylogenetic cluster. However, this clustering is due to the high homology of the HCR rather than the coding sequence of the pseudogenes. Therefore, this cluster was analyzed based on the relationships between the HCR and their chromosome location, not by its phylogenetics arrangement.

DISCUSSION

Two important points are revealed by this study: i) the presence of an HCR in functional genes and pseudogenes, and ii) the relationship between the function of the genes and their location on the chromosomes and the HCR.

Highly conversed regions

To analyze these regions, it is important to consider the structural characteristics of each HCR. Position and length of each HCR are common characteristics both in functional genes and pseudogenes. They display high similarity among themselves, suggesting a similar role for these regions and probably a similar set of specific transcriptional binding factors. These regions (HCRs) in conjunction with other molecular mechanisms could determine the gene expression of the hOR family. For example, a previous microarray study revealed that the expression level of the human olfactory genes and of pseudogenes, even in different tissues (Zhang et al., 2007), can be associated with common mechanisms of regulation or common expression patterns. According to this study, we suggest that the presence of HCR in pseudogenes accounts for their expression in the cell, perhaps at the same levels as the functional genes. If one pseudogene is normally expressed, then the HCR and transcription factors would not be associated with the features by which a pseudogene is defined; for instance, having a coding region less than 250 residues long. This hypothesis has been validated with the chromosomal and phylogenetic analyses explained in the next paragraph.

Highly conserved regions, function and genomic distribution of the genes

Gene distribution in chromosome-specific locations is a very important concept in understanding the molecular process necessary in the development of a function. If a particular set of genes has the same position and the same function, it is possible that all characteristics associated with its expression and regulation will be accomplished. Interestingly, the olfactory gene family is characterized by the regular organization of their structures, such as with axons nestled in the epithelium according to the type of receptor that is expressed (Wang et al., 1998; Feinstein and Mombaerts, 2004; Barnea et al., 2004) and to similar genes in the same location on the chromosomes (Zozulya et al., 2001; Glusman et al., 2001; Niimura and Nei, 2003). However, our study shows that HCR I and HCR II are present in different hOR genes and pseudogenes, and they do not have the same functions or locations on the chromosomes. This fact was shown in mice (Hoppe et al., 2006), where common regulatory regions located upstream from the TSS are associated with the topology of the genes in the olfactory epithelium. Therefore, our phylogenetic tree shows that the gene olfactory regulation mechanisms do not depend on the functionality of the gene; perhaps they are associated with the olfactory epithelium, as reported by Hoppe et al. (2006).

CONCLUSIONS

We found two highly conserved regions upstream from the transcription start sites for 19% of the human olfactory receptor genes. These regions have a possible enhancer function, they are not associated directly with the function of the gene, and are present in functional genes and pseudogenes. These qualities are important contributions for clarifying the regulation of the hOR genes and can explain the expression of the pseudogenes found recently.

ACKNOWLEDGMENTS

We gratefully acknowledge the Instituto Colombiano para el Desarrollo de la Ciencia y la Tecnología, Francisco José de Caldas - COLCIENCIAS, from the Republic of Colombia for supporting this project. We also thank Luis Garreta and Ember Martinez for helping us in the data mining, and all members in the BIMAC research group. Finally, we thank Dr. Albert Leyva for his assistance in the preparation of the manuscript.

REFERENCES

Altschul SF, Madden TL, Schaffer AA, Zhang J, et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389-3402.

Barnea G, O'Donnell S, Mancia F, Sun X, et al. (2004). Odorant receptors on axon termini in the brain. *Science* 304: 1468.

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

Chess A, Simon I, Cedar H and Axel R (1994). Allelic inactivation regulates olfactory receptor gene expression. *Cell* 78: 823-834.

Crooks GE, Hon G, Chandonia JM and Brenner SE (2004). WebLogo: a sequence logo generator. *Genome Res.* 14: 1188-1190.

Feinstein P and Mombaerts P (2004). A contextual model for axonal sorting into glomeruli in the mouse olfactory system. *Cell* 117: 817-831.

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

Hinrichs AS, Karolchik D, Baertsch R, Barber GP, et al. (2006). The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* 34: D590-D598.

Hoppe R, Weimer M, Beck A, Breer H, et al. (2000). Sequence analyses of the olfactory receptor gene cluster mOR37 on mouse chromosome 4. *Genomics* 66: 284-295.

Hoppe R, Frank H, Breer H and Strotmann J (2003). The clustered olfactory receptor gene family 262: genomic organization, promoter elements, and interacting transcription factors. *Genome Res.* 13: 2674-2685.

Hoppe R, Breer H and Strotmann J (2006). Promoter motifs of olfactory receptor genes expressed in distinct topographic patterns. *Genomics* 87: 711-723.

Kolchanov NA, Ignatieva EV, Ananko EA, Podkolodnaya OA, et al. (2002). Transcription Regulatory Regions Database (TRRD): its status in 2002. *Nucleic Acids Res.* 30: 312-317.

Lane RP, Cutforth T, Young J, Athanasiou M, et al. (2001). Genomic analysis of orthologous mouse and human olfactory receptor loci. *Proc. Natl. Acad. Sci. U. S. A.* 98: 7390-7395.

Malnic B, Hirono J, Sato T and Buck LB (1999). Combinatorial receptor codes for odors. *Cell* 96: 713-723.

Malnic B, Godfrey PA and Buck LB (2004). The human olfactory receptor gene family. *Proc. Natl. Acad. Sci. U. S. A.* 101: 2584-2589.

Michaloski JS, Galante PA and Malnic B (2006). Identification of potential regulatory motifs in odorant receptor genes by analysis of promoter sequences. *Genome Res.* 16: 1091-1098.

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

Notredame C, Higgins DG and Heringa J (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* 302: 205-217.

Olender T, Feldmesser E, Atarot T, Eisenstein M, et al. (2004). The olfactory receptor universe - from whole genome analysis to structure and evolution. *Genet. Mol. Res.* 3: 545-553.

Pavesi G, Mereghetti P, Mauri G and Pesole G (2004). Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res.* 32: W199-W203.

Qasba P and Reed RR (1998). Tissue and zonal-specific expression of an olfactory receptor transgene. *J. Neurosci.* 18: 227-236.

Sosinsky A, Glusman G and Lancet D (2000). The genomic structure of human olfactory receptor genes. *Genomics* 70: 49-61.

Tamura K, Dudley J, Nei M and Kumar S (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* 24: 1596-1599.

Taylor TD, Noguchi H, Totoki Y, Toyoda A, et al. (2006). Human chromosome 11 DNA sequence and analysis including novel gene identification. *Nature* 440: 497-500.

Thompson JD, Higgins DG and Gibson TJ (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22: 4673-4680.

Vassalli A, Rothman A, Feinstein P, Zapotocky M, et al. (2002). Minigenes impart odorant receptor-specific axon guidance in the olfactory bulb. *Neuron* 35: 681-696.

Wang F, Nemes A, Mendelsohn M and Axel R (1998). Odorant receptors govern the formation of a precise topographic map. *Cell* 93: 47-60.

Zhang X, De la Cruz O, Pinto JM, Nicolae D, et al. (2007). Characterizing the expression of the human olfactory receptor gene family using a novel DNA microarray. *Genome Biol.* 8: R86.

Zozulya S, Echeverri F and Nguyen T (2001). The human olfactory receptor repertoire. *Genome Biol.* 2: research0018.1-0018.12.