Genomics **63,** 227–245 (2000)
doi:10.1006/geno.1999.6030, available online at http://www.idealibrary.com on

Sequence, Structure, and Evolution of a Complete Human Olfactory Receptor Gene Cluster

Gustavo Glusman,* Alona Sosinsky,* Edna Ben-Asher,* Nili Avidan,* Dina Sonkin,* Anita Bahar,* André Rosenthal,† Sandra Clifton,‡ Bruce Roe,‡ Concepción Ferraz,§ Jacques Demaille,§ and Doron Lancet*,[1]

*Department of Molecular Genetics and The Crown Human Genome Center, The Weizmann Institute of Science, Rehovot 76100, Israel; †Institut für Molekulare Biotechnologie, Postfach 100813, D-07708 Jena, Germany; ‡Department of Chemistry and Biochemistry, University of Oklahoma, Norman, Oklahoma 73019; and §Institut de Genetique Humaine, UPR 1142, CNRS, 141, rue de la Cardonille, 34396 Montpellier Cedex 5, France

Received July 23, 1999; accepted October 18, 1999

The olfactory receptor (OR) gene cluster on human chromosome 17p13.3 was subjected to mixed shotgun automated DNA sequencing. The resulting 412 kb of genomic sequence include 17 OR coding regions, 6 of which are pseudogenes. Six of the coding regions were discovered only upon genomic sequencing, while the others were previously reported as partial sequences. A comparison of DNA sequences in the vicinity of the OR coding regions revealed a common gene structure with an intronless coding region and at least one upstream noncoding exon. Potential gene control regions including specific pyrimidine:purine tracts and Olf-1 sites have been identified. One of the pseudogenes apparently has evolved into a CpG island. Four extensive CpG islands can be discerned within the cluster, not coupled to specific OR genes. The cluster is flanked at its telomeric end by an unidentified open reading frame (C17orf2) with no significant similarity to any known protein. A high proportion of the cluster sequence (about 60%) belongs to various families of interspersed repetitive elements, with a clear predominance of LINE repeats. The OR genes in the cluster belong to two families and seven subfamilies, which show a relatively high degree of intermixing along the cluster, in seemingly random orientations. This genomic organization may be best accounted for by a complex series of evolutionary events. © 2000 Academic Press

INTRODUCTION

Environmental stimuli are recognized by sensory neurons, and this information is transmitted to the brain, where it is decoded to provide an internal representation of the external world. The vertebrate olfactory system is exquisitely adapted for recognition and discrimination among a large number of odorants, with high sensitivity and specificity (Laurent, 1997; Pilpel *et al.*, 1998). The initial step in olfactory discrimination involves the interaction of odorant molecules with a large repertoire of specific receptors.

Olfactory receptor (OR) genes encode G-protein-coupled seven-transmembrane proteins (Buck and Axel, 1991). Unlike the somatic gene recombination and mutation mechanisms that account for immunoglobulin diversity, the OR repertoire diversity seems to be germline-inherited. The OR gene superfamily is the largest in the mammalian genome. It is estimated to consist of several hundred genes in mammalian species and about 100 genes in catfish (reviewed in Mombaerts, 1999), suggesting a large expansion of the OR repertoire in higher vertebrates. A given OR was shown to be expressed by about 0.1% of the sensory neurons within the rodent olfactory epithelium (Vassar *et al.*, 1993; Ressler *et al.*, 1993). Estimating 500–1000 OR genes in the rat genome (Buck and Axel, 1991), these findings are consistent with the phenomenon of clonal and allelic exclusion in ORs (Lancet, 1991; Chess *et al.*, 1994; Malnic *et al.*, 1999), in which a neuron expressing a given receptor does not activate expression of other ORs. The multiplicity of receptors reflects the needs of a combinatorial coding system, in which each receptor may bind many odorants and each odorant binds several receptors (Lancet, 1986; Malnic *et al.*, 1999), as analyzed by a probabilistic model (Lancet *et al.*, 1993).

The OR repertoire contains a large percentage of pseudogenes that may be important for the generation and maintenance of diversity. The especially large number of OR pseudogenes in the human genome (up to ~70%) (Rouquier *et al.*, 1998) may reflect a loss

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under Accession Nos. AC007194, AF087915–AF087930, and AF155225.

[1] To whom correspondence should be addressed. Telephone: 972-8-9343683. Fax: 972-8-9344112. E-mail: bmlancet@weizmann.weizmann.ac.il.

of functional genes in the "microsmatic" primates (Sharon *et al.*, 1999).

Many of the human OR genes appear in genomic clusters with 10 or more members (Ben-Arie *et al.*, 1994; Glusman *et al.*, 1996; Vanderhaeghen *et al.*, 1997; Carver *et al.*, 1998; Trask *et al.*, 1998). An estimated total number of 500 human OR genes would indicate 30–50 such clusters, about half of which have been identified by cloning or by fluorescence *in situ* hybridization (FISH) on almost all human chromosomes (Rouquier *et al.*, 1998). In mouse, in which the estimated OR number may reach more than 1000, 12 clusters have so far been identified by genetic linkage on seven different chromosomes (Sullivan *et al.*, 1996). OR genomic clustering also was indicated by Southern hybridization analysis in dog (Issel Tarver and Rine, 1996) and by genomic mapping in zebrafish (Barth *et al.*, 1997). The complete collection of OR-containing genomic regions has been termed the "olfactory subgenome" (Ben-Arie *et al.*, 1993; Glusman *et al.*, 1996), estimated to encompass ~1% of the entire genome of mammalian species.

The availability of genomic sequences surrounding OR genes provides a unique opportunity to study the evolution of this multigene superfamily and to trace the mechanisms or genome dynamics that may have been responsible for its current size and variety. Using sequence comparison, ORs are classified into families (>40% amino acid identity) and subfamilies (>60% amino acid identity) (Ben-Arie *et al.*, 1994). An analysis of clusters in human (Ben-Arie *et al.*, 1994; Trask *et al.*, 1998), mouse (Sullivan *et al.*, 1996), and zebrafish (Barth *et al.*, 1997) indicates that each cluster may contain members of several subfamilies or even families. This suggests that present-day OR clusters have evolved in a complex path, involving ancient precursor gene duplications, as well as more recent within-cluster gene duplications. Conversely, genes of a given subfamily may be found in more than one cluster (Sullivan *et al.*, 1996; Rouquier *et al.*, 1998), suggesting that clusters may be duplicated, in part or in their entirety. In the latter case, this may occur via a duplication process that generates paralogous regions on different chromosomes. Repetitive genomic DNA elements (e.g., *Alu* and LINE) were suggested to have a crucial role in mediating recombination events that lead to OR gene duplications (Glusman *et al.*, 1996). Finally, the olfactory subgenome has been hypothesized to be "exclusive" in the sense that no non-OR genes have been found interspersed with OR genes (Glusman *et al.*, 1996).

The OR coding regions are uninterrupted by introns in the genome (Ben-Arie *et al.*, 1994) like many G-protein-coupled receptors (Gentles and Karlin, 1999), though the possibility of at least one exception has been reported (Walensky *et al.*, 1998). Characterization of human and murine OR genes revealed an intron separating a noncoding leader exon and from the coding exon (Glusman *et al.*, 1996; Asai *et al.*, 1996) and showed that transcription is initiated from a region upstream from the leader exon (Asai *et al.*, 1996; Walensky *et al.*, 1998; Qasba and Reed, 1998). The mechanism of control that generates the complex pattern of odorant receptor expression still remains largely unknown.

Partial genomic sequences of OR gene clusters have been published (Glusman *et al.*, 1996; Brand-Arpon *et al.*, 1999), giving initial genomic insights into the organization of OR genes, their structure and evolution, as well as some hints on potential mechanisms for transcriptional control. We present here the sequencing and analysis of the first complete OR gene cluster. The full analysis and annotation of the sequence, as well as ancillary information, can be viewed at http://bioinfo.weizmann.ac.il/papers/C17olf_cluster.

MATERIALS AND METHODS

Reagents and equipment. Cosmids were from library ICRFc105 isolated from human cell line LCL127 (Nizetic *et al.*, 1991) from the Resource Center, Primary Database of the German Human Genome Project (Nizetic *et al.*, unpublished results). Ten clones of the 80 cosmids covering the cluster were chosen for sequencing as follows: F03103 (cos17), D10132 (cos26), H07155 (cos32), B01193 (cos39), F06137 (cos46), E06173 (cosL53), E06184 (cos58), F1155 (cos65), H0468 (cos68), and D093 (cos73). In addition, two PAC clones C10910Q3 (P8) and E02527Q3 (P123) from the whole-genome library LLNLP704 (Ioannou *et al.*, 1994) were sequenced. Additional cosmids mapped (written in pairs of ICRFc105 number–our number) were H1241–4; D0345–6; E0364–7; C0435–9; A08110–19; G06112–20; F04113–21; G11124–22; A02138–27; B09144–29; H09113–42; F08120–44; D01121–45; A06163–51; G10182–56; F09183–57; B0595–61; D0759–62; B1015–63; F082–66; E101–69; C127–70; C117–71; D0569–74; A107–75, and H0689–76. Additional PACs mapped (written in pairs of LLNLP704 number–our number) were N01235Q19–1; E13239Q19–2; M121198Q4–3; B211178Q4–4; M15660Q3–5; I17730Q3–6; J10811Q3–7; C10910Q3–8; E10912Q3–9; F05891Q3–10; M15947Q3–11; N04302Q19–101; N21613Q3–102; M21613Q3–103; K22613Q3–104; E24597Q3–106; P18817Q3–108; B02928Q3–109; P041058Q3–110; P041064Q3–111; L091077Q3–112; A091041Q3–113; P021089Q4–119; E02527Q3–123; P16680Q3–128; M22845Q3–129; P241019Q3–131; and P231019Q3–132.

Mapping of PACs. The PAC clones used in the current work were obtained from RZPD using DNA probes prepared form the cosmids at the ends of the three cosmid contigs described (Ben-Arie *et al.*, 1994), i.e., cosmids 26, 53, 58, and 68. All the PAC clones thus obtained were subjected to PCR analysis with primers specific for several OR coding regions (ORs 93, 201, 2, 7, 30, 23, 24, 208, 209, 210, 4, and 31), as well as with some cosmid ends (Fig. 1).

Generation of sequencing templates. Except for cosmid 65, all the clones were sequenced using the shotgun strategy (Bodenteich *et al.*, 1993; Rowen and Koop, 1994). Cosmid or PAC DNA was sheared either by sonication or by nebulization, and the ends were repaired by treatment with T4 DNA polymerase followed by Klenow treatment or alternatively by treatment with mung bean nuclease followed by T4 DNA polymerase. The repaired DNA was size-fractionated on a 0.8% agarose gel, and fragments of 0.8–1.5 kb were excised and purified with a Qiagen gel extraction kit (Qiagen Gmbh, Germany) for ligation with M13 RF phage DNA (Novagen). Alternatively, fragments of 2–6 kb were excised and purified from low-melting-point 0.8% agarose gels with gelase (Epicentre) or a Qiagen kit as above, for ligation to pBluescript (Stratagene) or pUC18 (Pharmacia) vectors. Ligation was performed with a Rapid Ligase kit (Boehringer) or with a Fast Link kit (Epicentre) according to the manufacturer's instructions. The ligated DNA was used for transformation of XL1 Blue competent cells (Stratagene). DNA from single clones was subjected to direct sequencing, or PCR products of these clones were subjected to sequencing reactions. When direct sequencing was applied, DNA was prepared by Qiagen kits; either an M13 extraction kit for single-stranded DNA or the turbo miniprep kit for double-stranded DNA was used according to the manufacturer's instructions (Qiagen Gmbh, Germany). These kits were used in their 96-well format with the 96-manifold apparatus, which was connected to a Biomek 2000 robot (Beckman). Some clones were also prepared by a cleared lysate filter-based protocol (Chissoe *et al.*, 1995) and sequenced as described (Bodenteich *et al.*, 1993). When PCR products were to be sequenced, they were cleaned by a 96-well Gel Filtration Block (Edge BioSystems) prior to fluorescence labeling.

Sequencing reactions. DNA was labeled either by fluorescence-labeled primers or by fluorescent dye terminators—Prism cycle sequencing and Big-Dyes kits (Perkin–Elmer/Applied Biosystems)—and analyzed on ABI 373 or ABI 377 sequencers.

Finishing and gap closure. Finishing of cosmid 65 as well as finishing of cosmids 17, 68, and PAC 8 was performed using the differential extension with nucleotide subsets (DENS) method (Raja *et al.*, 1997). Briefly, in this method single-stranded DNA is synthesized by PCR and is then subjected to DNA sequencing by primer walking using a presynthesized primer library. Cosmid 65 was used as template to sequence the 10-kb gap between cosmids 46 and 58. Primers for synthesis of the desired segment on cosmid 65 were designed using the programs Oligo (Rychlik, 1995) and Amplify (Engels, 1993) based on known sequence from the overlapping cosmids 46 and 58. Sequence finishing of the other clones was performed by standard primer walking along the regions where sequence was in doubt.

Sequence assembly. Assembly was performed using Sequencher 3.1 software from GeneCodes Corp. and/or phrap (University of Washington). Based on experience with the sequencing methodology in other projects, we estimate the precision of the consensus sequence to be over 99.9%. Additional quality control was obtained by comparison of overlapping, independently sequenced clones. The cluster sequence has been deposited with GenBank under Accession No. AC007194. The full coding sequences for the 17 OR genes in the cluster have been deposited with GenBank under Accession Nos. AF087915–AF087930 and AF155225.

Sequence analysis. Sequences were analyzed using the GESTALT Workbench (Glusman and Lancet, in preparation). Briefly, GESTALT is a Perl-based workbench for automated large-scale genomic sequence analysis, comparison, and annotation. GESTALT integrates and depicts graphically the output of diverse sequence analysis algorithms, including database searches, gene modeling tools, recognition of interspersed repeats, statistical ORF analysis, and compositional analyses, as well as user annotation.

Open reading frame analysis. The significance of each observed open reading frame (ORF) of length L was estimated by calculating an expectation value $E(L)$ as the probability of finding an ORF of length L or longer, times the number of possible such ORFs (approximately the length of the sequence). The probability for length $\geq L$ was calculated assuming an exponential distribution with the extension parameter being the frequency of stop codons in the sequence, using either the observed stop frequency in the entire sequence or the expected value for the local G+C content.

Identification of coding regions. Statistically significant open reading frames were studied by database searches, unless recognized to belong to repetitive elements by RepeatMasker (Smit and Green, 1997). The entire genomic sequence obtained was analyzed using FASTY (Pearson *et al.*, 1997) against a database of translated OR sequences (Glusman *et al.*, in preparation) as well as by dot-plot to representative OR nucleotide sequences. GenScan (Burge and Karlin, 1997) and fgenes 1.6 (Solovyev and Salamov, 1997) were used to build comprehensive gene models within the cluster sequence. For each OR coding region identified in the cluster, the prediction success was calculated as the fraction of its nucleotides predicted to be within a coding exon, in the proper strand.

Identification of CpG islands. The local concentration of CpG dinucleotides was calculated as the contrast value (CV) or ratio between observed and expected frequency, as CV = $[\mathrm{CpG}]/[C][G]$, where $[C]$ indicates frequency of C nucleotides, etc. CpG dinucleotides are underrepresented in the human genome (Karlin *et al.*, 1998). CpG islands are defined as regions over 200 bp with CpG CV > 0.6 and G+C content above 50%.

Phylogenetic analysis. The conceptually translated OR sequences from this cluster were compared to additional human OR sequences, chosen to represent Class II families 1–7 from several chromosomal locations. Fish and human Class I representatives are added for comparison. The human β-3 adrenergic receptor (HSB3A) was used as outgroup. Multiple alignment and neighbor-joining analysis were performed using ClustalX (Higgins *et al.*, 1996) with default parameters. Confidence was estimated using 1000 rounds of bootstrapping. Phylogenetic trees were drawn using TreeView (Page, 1996).

Divergence time estimation. This was performed (Glusman *et al.*, 1996) by comparing nucleotide sequences on which no selection is assumed to take place. The estimated substitution level (ESL) was calculated using the one-parameter model (Jukes and Cantor, 1969) and then translated to million years ago (Mya) as described for the $\psi\eta$-globin gene locus (Bailey *et al.*, 1991) with substitution rates (expressed as 10^{-9} substitutions/site/year) of: 1.1 for the last 19.2 Mya (gibbon/human divergence), 1.7 for the period 25.0–19.2 Mya (cercopithecoid/hominoid divergence), 1.9 for the period 34.2–25.0 Mya (platyrrhine/catarrhine divergence), 3.5 for the period 55.0–34.2 Mya (strepsirhini/haplorhini divergence), and 5.0 before 55 Mya (mammalian-wide). These figures reflect the "hominoid slowdown" in nucleotide sequence mutation frequencies (Bailey *et al.*, 1991) and do not reflect $\psi\eta$-globin-specific evolution rates.

Gene structure prediction. To predict potential upstream noncoding exons, the genomic environment of each OR coding region in the cluster (except for the 5′ truncated OR17-25) was extensively analyzed to generate a gene model. The genomic region of each OR gene was defined to include up to 15 kb upstream from the start codon and 5 kb downstream from the termination codon. The relevant genomic sequence employed was trimmed for seven OR genes, to avoid overlaps: for OR17-228, 12.5 kb were used (downstream from OR17-40); OR17-24 and OR17-40 have a common upstream region in opposite orientations, and therefore half (8.9 kb) was taken for each. Similarly, 11.1 kb were used for OR17-201 and OR17-2. From our analysis, the 5′ genomic region of OR17-4 includes at least 12.8 kb; the genomic region of OR17-210 was correspondingly trimmed to 10.4 kb. Several exon prediction programs based on different algorithms were used, including GenScan (Burge and Karlin, 1997) (suboptimal exon cut-off used: 0.1), GRAIL II (Xu *et al.*, 1994), Genie (Kulp *et al.*, 1996), and the programs fgene, fgenes, hexon, and fex from the GeneFinder package (Solovyev *et al.*, 1995; Solovyev and Salamov, 1997). Potential exons recognized by at least three programs were analyzed further. Dot-plot analysis was used to determine the extents of the duplicated regions within subfamilies, and only exons within such regions were considered as conserved and therefore potentially functional. Dot-plot analysis and sequence alignments were analyzed by GeneAssist 1.1 from ABI, Perkin–Elmer.

Prediction of additional gene structure elements. The acceptor splicing sites for the predicted coding exons were detected by the programs SPL from the GeneFinder package (Solovyev and Salamov, 1997) and SSPNN (Brunak *et al.*, 1991). Donor splicing sites for the potential upstream exons were detected by the SPL program with an LDF value of 0.85 used as cut-off between strong and weak sites. Polyadenylation signals were detected using POLYAH (Solovyev and Salamov, 1997). Potential promoters and corresponding transcription start sites (TSS) were identified using TSSG and TSSW (Solovyev and Salamov, 1997) with minimal score 0.4 and by PPNN (Reese *et al.*, 1996) with minimal score 0.8.

Control region analysis. To detect any significant similarities among potential control regions, the oligonucleotide analysis tool (van Helden *et al.*, 1998) from the Yeast Regulatory Tools (van Helden *et al.*, in preparation) was used. We implemented also a variant that relaxes the requirements on the patterns found, allowing the detection of similar patterns in addition to identical patterns. The sequences were also analyzed using the segment pair overlap method implemented in MACAW (Schuler *et al.*, 1991), as well as the Gibbs sampler as implemented with the Yeast Regulatory Tools. The location of binding sites for members of the two families of transcription factors NF-1 and O/E were examined by MatInspector V2.2 (Quandt *et al.*, 1995) using the TransFac database. Olf-1 and NF-1 sites were also mapped by the Word Mapper tool of the GESTALT Workbench (Glusman and Lancet, in preparation) using the consensus sequences TCCCNNRRGRR and GCTGGCANNNTGCCAG, respectively (R represents purines). Potential recombinatorial signal sequences (Sakano *et al.*, 1981) were mapped using the consensus CACTGTG $(N)_x$GGTTTTTGT (where *x* is 12 or 23).

RESULTS AND DISCUSSION

Complete Sequence of an Olfactory Receptor Gene Cluster

Mapping and sequencing. We have obtained 412 kb of contiguous genomic sequence encompassing the OR gene cluster on human chromosome 17. The sequence is a composite of 12 cosmid and 2 PAC clones (Fig. 1). Additional PAC clones (P110 and P111, see Fig. 1) that overlap with cos58 and extend the cluster map at its telomeric end have been identified, but PCR analysis with OR-specific OR5B/OR3B degenerate primers (Ben-Arie *et al.*, 1994) suggested that they were devoid of additional OR coding regions. The final size of the cluster sequence fits the ∼400 kb estimated in the initial characterization of this cluster (Ben-Arie *et al.*, 1994). STS marker 506 (D17S126) is present within this cluster (in cos73) as originally mapped by PCR (Ben-Arie *et al.*, 1994). In addition, STS marker D17S1548 (WI-5436) is present at the end of this cluster (in cos58). D17S1548 is mapped to 48.8 cR from the 17p telomere or 4.521 Mb according to the UDB map of chromosome 17 (Chalifa-Caspi *et al.*, 1997); see http://bioinformatics.weizmann.ac.il/udb.

Three cosmid contigs have been described (Ben-Arie *et al.*, 1994), and their orientation and distances have been estimated by free-chromatin FISH. Analysis of several PAC clones covering this region enabled us to correct the physical map of the cluster, as shown in Fig. 1. The original improper mapping of several cosmids was found to derive from the existence of a large genomic duplication, described below.

An unclonable region. A 2.6-kb fragment in cosmids 17 and 68 (open boxes, Fig. 1) was particularly refractory to M13 subcloning and significantly underrepresented in sequenced shotgun subclones. Closure was accomplished by DENS primer walking (Raja *et al.*, 1997). Interestingly, this segment posed no shotgun cloning problems upon direct sequencing from a partially deleted cosmid clone (R28, Fig. 1). Analysis of the resulting sequence shows that the fragment is located between very old MIR (SINE) and Charlie (DNA/MER1 type) repeats, which are ∼20% divergent from their respective consensus sequences. No internal repeats or palindromes were detected in this apparently unclonable segment, but it was found to be singularly G+C-poor (30% overall, down to 25% in the middle), culminating with an A+T low-complexity region.

The Detected OR Genes

The genomic sequence in the OR cluster was compared with a database of OR gene sequences at both the nucleotide and the translated amino acid levels. In total, 17 OR coding regions were recognized (Table 1), confirming the presence of 11 of the OR genes described in the initial report on this cluster (Ben-Arie *et al.*, 1994), as well as 6 formerly undetected coding regions. Approved nomenclature symbols (Glusman *et al.*, in preparation) are listed in Table 1. From sequence analysis only, 6 of the 17 OR coding regions are pseudogenes, while the remaining 11 are apparently functional. The expression of OR17-93 and OR17-40 has been previously shown experimentally (Ben-Arie *et al.*, 1994; Crowe *et al.*, 1996). We have experimental evidence that all the remaining apparently functional genes are transcribed (Sosinsky *et al.*, in preparation) except for OR17-6, which may turn out to be a pseudogene.

We have previously reported (Glusman *et al.*, 1996) the sequence analysis of a cosmid (cos39) covering the middle of the cluster, which encodes two genes (OR17-40 and OR17-228) and two fused, truncated pseudogenes (OR17-24 and OR17-25, Fig. 2a). The genomic sequence of cosmids D53 and L53 confirmed the existence of OR17-32, an allelic variant of OR17-2 that differs from it by only 2 bp of 648 bp (Sharon *et al.*, in preparation), indicating that the individual from whom the cosmid library was created was heterozygous at this locus. Similarly, the presence of the OR17-23 pseudogene was confirmed, but its OR17-90 variant was not detected in the sequenced cosmids nor in population studies (Sharon *et al.*, in preparation). In contrast, OR17-30 occurs as two almost identical but disjointed copies in the cluster: the newly detected OR gene (hereafter referred to as OR17-31) appears to be the OR gene closest to the telomeric end of the cluster.

Two additional OR pseudogenes (OR17-208 and OR17-1) and two additional, apparently functional OR genes (OR17-6 and OR17-7) were detected. OR17-208 has an in-frame stop codon (Fig. 2a) but otherwise is apparently intact, suggesting that this is a relatively recent mutation in a gene from family 1. Indeed, its chimpanzee orthologue lacks this stop codon (Sharon *et al.*, 1999). OR17-1 harbors several alterations that render it a pseudogene, including four frameshifting mutations (Fig. 2a). These four novel OR regions represent three new subfamilies within family 1 (Fig. 3).

Of the 17 OR coding regions in this cluster, 6 (3 pseudogenes and 3 apparently functional genes) were detected only by genomic sequencing. The 11 previ-

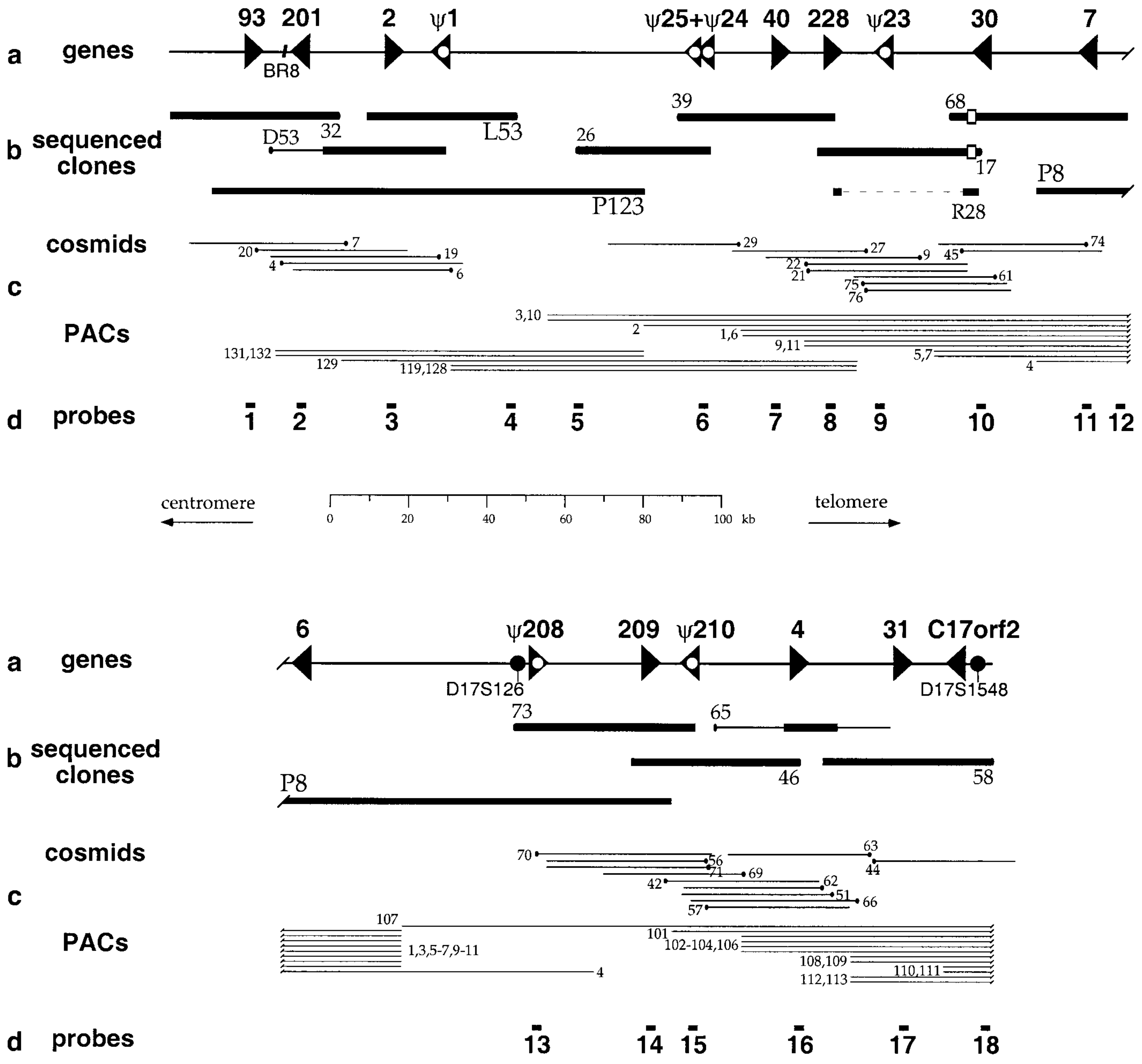


FIG. 1. Physical map of the olfactory receptor gene cluster on human chromosome 17p13.3. (**a**) Location and orientation from 5′ to 3′ of the OR coding regions (arrowheads) and the D17S126 and D17S1548 markers. BR8 represents the approximate DNA breakpoint of Miller-Dieker syndrome patient BR8 (Ben-Arie *et al.*, 1994). Pseudogenes are indicated by ψ and white circles. (**b**) The sequenced cosmid and PAC clones as detailed under Materials and Methods. Thin lines indicate regions not sequenced. The dashed line in cosmid R28 indicates the region deleted in this clone. Open boxes in cosmids 17 and 68 indicate the unclonable region. (**c**) The approximate extents of all additional cosmid and PAC clones mapped. (**d**) The PCR probes used for mapping PACs.

ously detected OR genes in this cluster have between zero and three mismatches in each degenerate PCR primer site or up to four mismatches in total (Fig. 2b). Three of the previously undetected OR coding regions (OR17-1, OR17-25, and OR17-208) have a larger number of mismatches. OR17-31 is almost identical to OR17-30 in the coding region (7 differences of 939 bp, or 99.3% identical). These results underscore the importance of genomic sequencing for reaching a definitive characterization of gene clusters, even when the gene families are well studied.

A recent independent genome-wide sequence survey (Rouquier *et al.*, 1998) of human ORs produced six partial ORs from chromosome 17 that map to this same chromosomal location, although having slight sequence variations. Therefore it is unlikely that additional family 1 and 3 OR genes occur in chromosome 17. However, members of more divergent OR families are present at other loci on chromosome 17, e.g., HTPCR16 in 17q21–q22 (Vanderhaeghen *et al.*, 1997). Indeed, we found this genomic region (GenBank Accession No. AC005962) to include two OR genes belonging

TABLE 1

Characteristics of the OR Coding Regions in the Olfactory Receptor Gene Cluster on Human 17p13.3

Name	HUGO	Start	End	Str	Length	%G+C	CpG	%CpG	ψ	%GS	%FG
OR17-93	*OR1E2*	20,283	21,254	+	972	49.9%	10	1.0%		100	66
OR17-201	*OR3A3*	33,539	32,592	−	948	55.0%	15	1.6%		100	100
OR17-2	*OR1E1*	55,749	56,693	+	945	50.0%	10	1.1%		100	100
OR17-1	*OR1R1P*	68,245	67,250	−	996	65.4%	93	9.3%	Yes	58	(64)
OR17-25	*OR3A5P*	142,943	142,203	−	741	54.3%	9	1.2%	Yes	60	0
OR17-24	*OR3A4P*	143,826	142,944	−	883	54.5%	13	1.5%	Yes	100	0
OR17-40	*OR3A1*	161,571	162,518	+	948	54.1%	18	1.9%		100	47
OR17-228	*OR3A2*	175,237	176,184	+	948	53.9%	18	1.9%		95	100
OR17-23	*OR1D3P*	188,541	187,602	−	940	53.3%	13	1.4%	Yes	85	0
OR17-30	*OR1D4*	213,474	212,536	−	939	52.3%	10	1.1%		100	55
OR17-7	*OR1A1*	238,482	237,553	−	930	47.3%	15	1.6%		100	0
OR17-6	*OR1A2*	256,581	255,652	−	930	45.3%	10	1.1%		100	0
OR17-208	*OR1P1P*	298,430	299,422	+	993	53.3%	20	2.0%	Yes	(55)	0
OR17-209	*OR1G1*	325,764	326,705	+	942	48.9%	12	1.3%		89	0
OR17-210	*OR1E3P*	336,884	335,937	−	948	50.3%	11	1.2%	Yes	(81)	0
OR17-4	*OR1D2*	360,316	361,254	+	939	50.5%	8	0.9%		100	68
OR17-31	*OR1D5*	389,717	390,655	+	939	52.9%	11	1.2%		100	0
Total	17			8/9	3.9%	51.6%	12.7	1.36%	6	99/51	49/11

Note. Trivial name and HUGO nomenclature name; absolute location from the centromeric end of the cluster and strand (Str); length, G+C content, number of CpG dinucleotides and their fraction of the coding region; pseudogene status as predicted from the sequence (ψ); percentage coverage of OR coding regions by predicted exons of GenScan and fgenes (GS and FG, respectively). Figures in parentheses indicate opposite strand predictions (GenScan) or the part in the proper frame (fgenes). Summary line: number of genes in the cluster and their distribution by orientations; fraction of coding sequence in the cluster; average G+C content, average number and percentage of CpG dinucleotides (excluding OR17-1); pseudogene count; average true-positive percentage rate of GenScan and fgenes for genes/pseudogenes.

to family 4: HTPCR16 and a new member of its subfamily (approved symbols (Glusman *et al.*, in preparation) *OR4D1* and *OR4D2*, respectively; see Fig. 3).

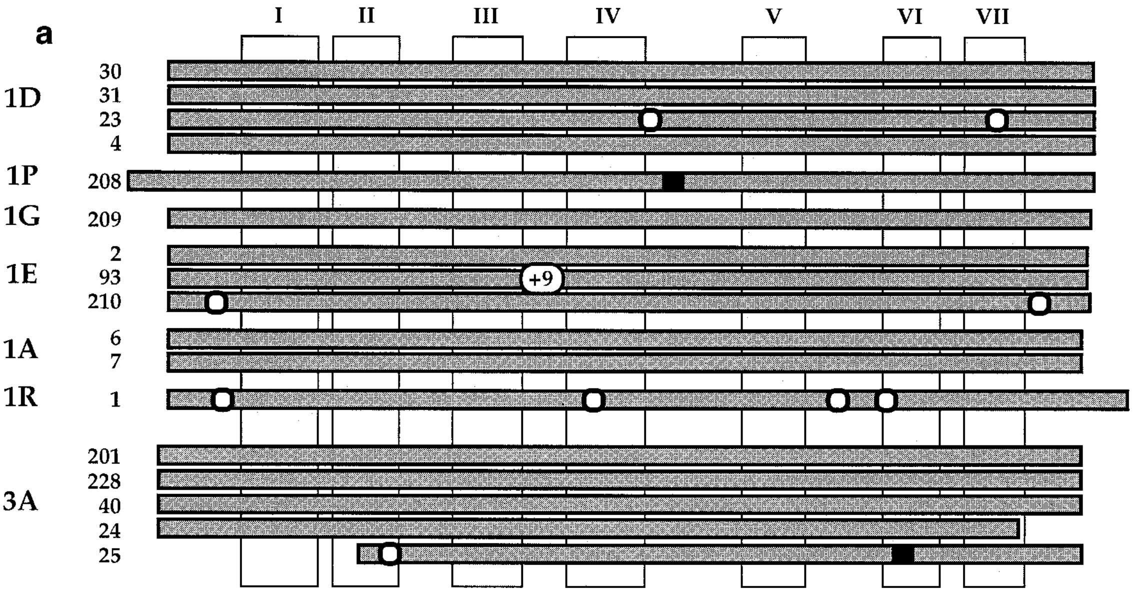
Analytical Gene Prediction

The sequence of the OR cluster was subjected to gene prediction analysis using the GenScan and fgenes programs (Fig. 4), which predicted 33 and 15 genes, respectively. Table 1 summarizes the success rate for each OR coding region (true positives). Overall GenScan yielded better predictions than fgenes, recognizing 99% of the coding sequence for apparently functional genes and 51% of that in pseudogenes, versus 49 and 11% for fgenes, respectively. In two cases antisense-overlapping ORFs (Merino *et al.*, 1994) of pseudogenes were recognized by GenScan: the OR17-208 pseudogene, which is interrupted by an in-frame stop codon, and the OR17-210 pseudogene, which has two frameshifting mutations. The fgenes prediction for the OR17-1 pseudogene (which has several frameshifts; see Fig. 2) partially relies on the wrong frame. Gene modeling programs that do not use sequence comparison generally are unsuitable for modeling pseudogenes, yet a high proportion of OR genes are pseudogenic (Rouquier *et al.*, 1998; Sharon *et al.*, 1999). These programs are trained to identify multi-coding-exon genes, but no introns that interrupt OR coding regions have yet been reported, with one single potential exception (Walensky *et al.*, 1998). Since OR regions can be recognized easily by protein sequence similarity, we conclude that the method of choice for detecting OR genes in new genomic sequence is to compare it to a database of OR sequences using any alignment tool able to incorporate frameshifts, such as FASTX or FASTY (Pearson *et al.*, 1997).

All of the statistically significant ORFs and most of the exons predicted represent OR coding regions (or their complementary strands), high-scoring segments within repetitive sequences (typically fragments of the *pol*-like polypeptide within L1 repeats), or very low-scoring, short exons, none of which finds homologues by blast (not shown).

A Non-OR Candidate Gene

A 297-codon-long ORF (nomenclature symbol C17orf2) was recognized by GenScan as a single-exon gene with a total score of 13.38. A polyadenylation signal is present 134 bp downstream from the stop codon. *C17orf2*, located at the telomeric end of the OR cluster, has a relatively high G+C content (63.2%) and is richer in CpG dinucleotides than expected from its nucleotide composition, especially at its 5′ end. Even though the derived amino acid sequence was analyzed by exhaustive database searching, no homologues were detected. Borderline similarities to EST hits were not improved by clustering, and no particular fold could be assigned to the predicted amino acid sequence (not shown). The possibility exists that this long ORF has no coding content and that its length derives from the expected lower number of stop codons in a G+C-rich region; the composition-corrected expectation value for such an ORF is borderline (0.8). Alternatively, *C17orf2* could represent the first member of a new gene family, in line with the estimate that approx-



```
b OR5B   CCCATGTAYTTBTTYCTCDSYAAYYTRTC      OR3B   AGRCWRTAIATGAAIGGRTTCAICAT

     30  ........c..c..c..Ggcc..cc.C..        30   ..g.tg..a.....a..g....t...
     31  ........c..c..c..Ggcc..cc.C..        31   ..C.tg..g.....a..g....t...
     23  ........c..c..c..Ggcc..cc.C..        23   ..g.tg..g.....a..g..A.t...
      4  ...G....c..c..c..Ggcc..cc.C..         4   ..g.tg..g.....g..a....t...
    208  ..T.....c..c..c..Tgcc.Gct.g..       208   .Tg.tg..a..A..g..g....g..G
    209  ........c..c..t..Agcc..cc.C..       209   .Aa.tg..g..A..g..a....a...
      2  ..T.....t..g..t...agc..ct.g..         2   ..g.tg..g.....g..g....g...
     93  ..TG....t..g..t...agc..ct.g..        93   ..g.tg..g.....g..gG...g...
    210  ..T.....t..g.Gt...agc..ct.g..       210   ..g.tg..g.....g..g....g...
      6  ........t..tC.c..Tgcc..cc.C..         6   ..a.ta..g.....a..a..T.aTGC
      7  ........t..tC.c..Tgcc..cc.C..         7   ..a.tg..g.....a..a..T.a...
      1  ........t.Ac..cT.GggtC.cc.gAG         1   ..g.tg.Tg.....a..g....aGG.
    201  ........c..c..c..GggG..cc.a..       201   ..g.tg..g..A.Gt..g....g...
    228  ........c..c..c..GggG..cc.g..       228   ..g.tg..g..A.Ga..g....g...
     40  ........c..c..c..GggG..cc.a..        40   .Ag.tg..g....Tt..g....g...
     24  ........c..c..c..GggG..ct.g..        24   GTC.ag..g....Gt..g....g...
     25  GATGCTC.cAGgGAGGC.tACTGGACa..        25   ..g.tg..a....Ct..g....g...
```

FIG. 2. The olfactory receptor coding regions. (**a**) The extent of each predicted peptide sequence is indicated on top of the locations of the seven-transmembrane domains (I to VII). Open circles denote frameshifts; black boxes indicate in-frame internal stop codons. The circled +9 indicates the location of a 9-amino-acid duplication in OR17-93 (Ben-Arie *et al.*, 1994). Gene names and subfamily classifications are indicated on the left. (**b**) The sequences recognized by the OR5B and OR3B degenerate primers: for each gene, periods indicate agreement with the primer consensus, lowercase letters indicate the nucleotide found at a primer degenerate position, and shaded uppercase letters indicate mismatches from the primer consensus.

imately 50% of all newly detected genes may represent novel families (Uberbacher *et al.*, 1996).

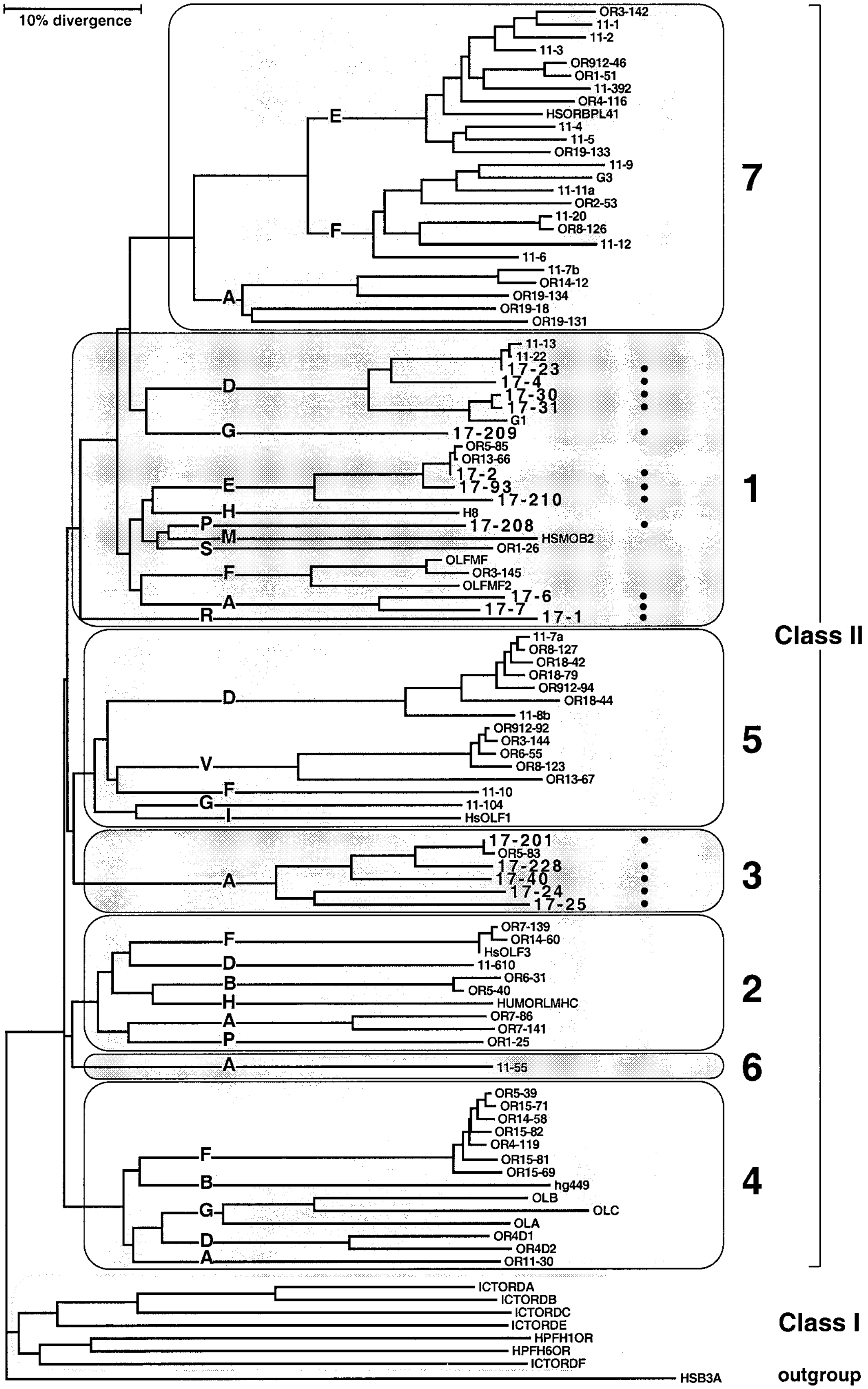
An OR Pseudogene Turned into a CpG Island

We have discovered a striking example of an OR gene whose entire coding region has evolved into an apparent CpG island. The OR17-1 pseudogene has the highest G+C content of all those in the cluster (65.43%, see Table 1) and includes 93 CpG dinucleotides (9.3% of length) while the other OR regions in the cluster have 8–20 CpGs (1 to 2% of length, Table 1). While this pseudogene has high G+C content and many CpG dinucleotides, it lacks Sp1 sites, which prevent remethylation of CpG islands (Brandeis *et al.*, 1994).

While this OR unit lost its protein coding function, it is apparently under new evolutionary constraints. Ancient, Class I (fish-like) OR pseudogenes have been reported to have adopted noncoding functions (i.e., regulatory) as enhancers (Buettner *et al.*, 1998). In addition, the human matrix-attachment region (MAR) reported in GenBank locus HSM0B2 (Nikolaev *et al.*, 1996) is significantly similar to OR genes, apparently being an additional OR pseudogene, this time taking a structural role (Gimelbrant and McClintock, 1997). This MAR is mapped to 19p13.2 and is classified as a Class II, family 1 OR. We therefore hypothesize that OR genes, which are present in the genome in many copies, can also adopt new functions, much as observed for the pseudogenes of retroposons (von Sternberg *et al.*, 1992; Britten, 1994; Hanke *et al.*, 1995).

OR Clusters Contain CpG Islands

Compositional analysis of the complete sequence shows that this cluster belongs in the G+C-poor L isochore (Bernardi, 1993). Four CpG-rich segments (circled 1–4 in Fig. 4) were identified in the cluster.


10% divergence
OR3-142
11-1
11-2
11-3
OR912-46
OR1-51
11-392
OR4-116
HSORBPL41
11-4
11-5
OR19-133
E
11-9
G3
11-11a
OR2-53
11-20
OR8-126
11-12
F
11-6
11-7b
OR14-12
OR19-134
A
OR19-18
OR19-131
7
11-13
11-22
17-23
D
17-4
17-30
17-31
G1
G
17-209
OR5-85
OR13-66
17-2
E
17-93
17-210
H
H8
P
17-208
M
HSMOB2
S
OR1-26
OLFMF
F
OR3-145
OLFMF2
A
17-6
17-7
R
17-1
1
Class II
11-7a
OR8-127
OR18-42
OR18-79
OR912-94
D
OR18-44
11-8b
OR912-92
OR3-144
OR6-55
V
OR8-123
OR13-67
F
11-10
G
11-104
I
HsOLF1
5
17-201
OR5-83
17-228
A
17-40
17-24
17-25
3
OR7-139
F
OR14-60
HsOLF3
D
11-610
B
OR6-31
OR5-40
H
HUMORLMHC
A
OR7-86
OR7-141
P
OR1-25
2
A
11-55
6
OR5-39
OR15-71
OR14-58
OR15-82
OR4-119
F
OR15-81
OR15-69
B
hg449
OLB
G
OLC
OLA
D
OR4D1
OR4D2
A
OR11-30
4
ICTORDA
ICTORDB
ICTORDC
ICTORDE
HPFH1OR
HPFH6OR
ICTORDF
Class I
HSB3A
outgroup

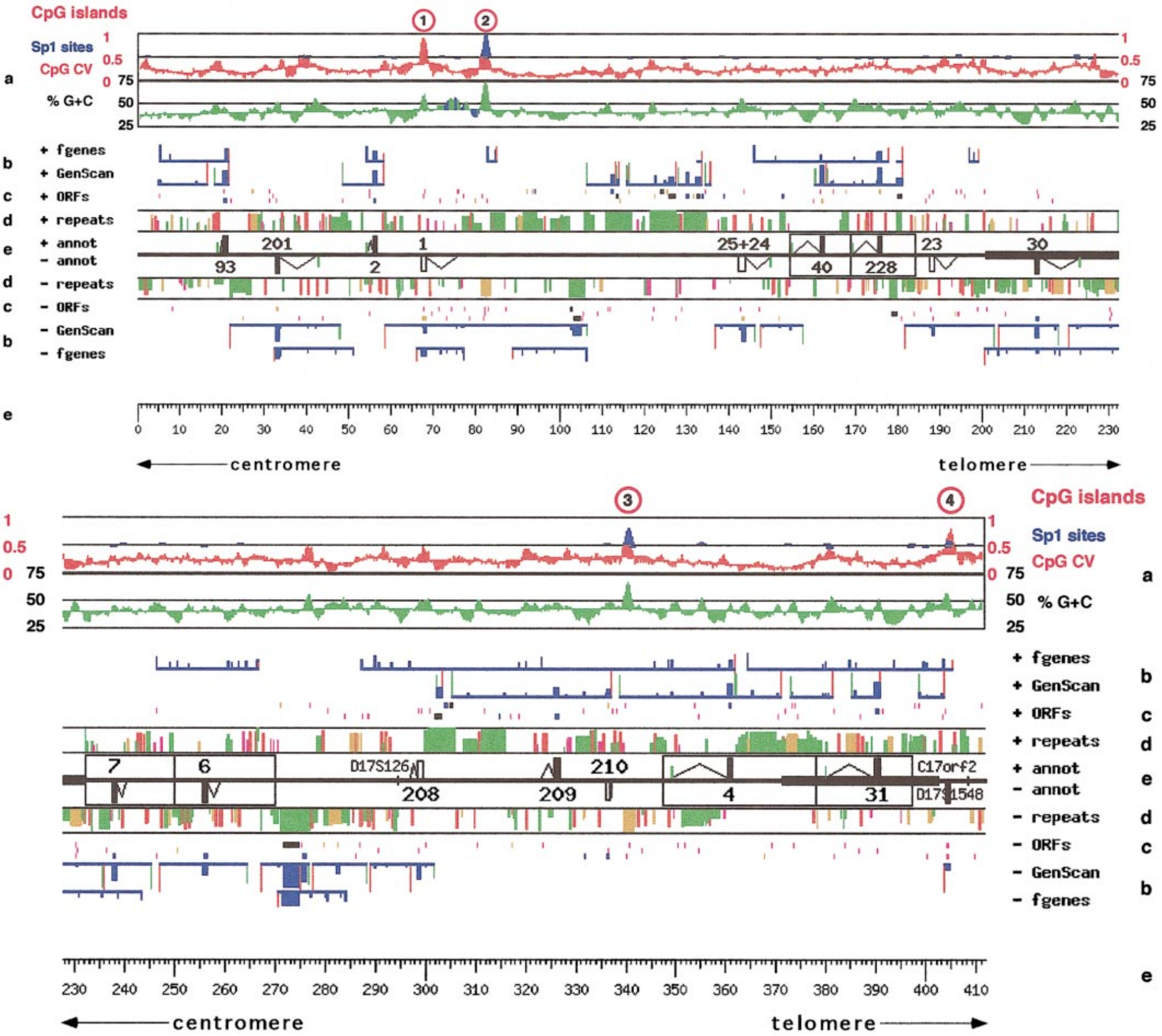


FIG. 4. A sequence map of the cluster generated by the GESTALT Workbench. (**a**) Compositional analyses. CpG contrast values and %G+C are displayed as deviations from the regional average; CpG islands are denoted by circled numbers on top; Sp1 clusters are overlaid in blue; green %G+C stretches belong in the L isochore, and blue stretches belong in the H isochore. (**b**) Gene prediction results (fgenes and GenScan). Predicted exons are displayed in blue, with box height indicating exon quality (the scaling is arbitrary but consistent for each prediction program); complete gene structures are underlined in blue; predicted promoters and poly(A) signals are indicated in green and red, respectively. (**c**) Location of ORFs colored by statistical significance: brown and blue ORFs indicate E( ) value under 1 and 1E-3 for the cluster sequence, black ORFs indicate E( ) value under 1 for the whole genome (3.3E9 bases). (**d**) Repetitive sequences. *Alu*s are denoted in red, MIRs in purple, LINEs in green, other interspersed repeats in brown; box height indicates element youth as percentage identity with the subfamily consensus, from 50% (oldest) to 100% (youngest). (**e**) User annotation. Location, orientation, and intron–exon structure of the OR genes and C17orf2; putative control regions are indicated in green; pseudogenes are indicated by open boxes; also shown are locations of the STS markers. The thicker horizontal bars surrounding OR17-30 and OR17-31 indicate the extent of the duplicated region; the tandem OR17-40/OR17-228, OR17-6/OR17-7, and OR17-4/OR17-31 duplications are enclosed in rectangles. (**e**) Kilobase scale from the centromeric end of the cluster. (**b**) to (**e**) Features on top of the middle line run from 5′ to 3′, and features under the middle line are in the reverse orientation.

FIG. 3. Phylogenetic tree of representative human OR genes of Class II, families 1–7 (shaded). Capital letters on branches denote subfamilies. The OR genes from human chromosome 17 cluster are shown in larger font size and are marked with black dots. The human β-3 adrenergic receptor (HSB3A) is used as an outgroup, and several catfish (*Ictalurus punctatus*) ORs are included as Class I representatives. The bar indicates 10% amino acid divergence along each branch.

The most centromeric CpG island (circled 1) includes the complete coding region of the OR17-1 pseudogene. The most telomeric CpG island (circled 4) includes the long ORF *C17orf2*. The two remaining CpG islands (circles 2 and 3) are derived from recently inserted (<7 million years ago) SVA retroviral elements (Shen *et al.*, 1994). CpG islands 2 and 3 have many Sp1 sites (Brandeis *et al.*, 1994), as indicated by blue peaks over the CpG islands in Fig. 4. CpG island 4 has four Sp1 sites within it, and OR17-1 has none.

Our results indicate that the OR genes in this cluster do not have "private" CpG islands at their 5′ ends: rather, a few CpG islands are present at an average frequency of one island per ~100 kb. These may be regulatory sequences potentially affecting the expression of the entire OR cluster or only part of it. We have similarly observed one CpG island in the 106-kb partial sequence of the human chromosome 3 OR gene cluster (Brand-Arpon *et al.*, 1999), in the range 22–23 kb of GenBank entry AF042089. As in the cluster described here, that kilobase-long CpG island is not associated with any particular OR gene.

High Abundance of Repetitive Sequences

Up to 60% of the cluster sequence is composed of repetitive elements of all known types, including LINEs (40%), SINEs (9%), LTR elements (6%), and DNA transposons (3%). Several instances of repetitive elements retroposing into previous repeats were observed, with up to five levels of repeated insertion into the same locus, a structure we have named the "genomic matrioshka" (Glusman *et al.*, 1998).

Thus, the cluster appears to have been highly permissive to repeated invasion by retroposing elements, especially those of the L1 type, which amount to 38% of the cluster sequence. Indeed, it is apparent that the cluster has been evolutionarily "breaking up" into subclusters separated by long L1-rich stretches (e.g., 75–135 and 270–335 kb, Fig. 4), sometimes engulfing pseudogenes (e.g., OR17-208). *Alu* repeats (amounting to 8% of the sequence) are somewhat clustered in the regions surrounding the OR genes. The high proportion of L1 repeats is consistent with this cluster being part of a low-GC content L isochore (Bernardi, 1993) within a G band (Gardiner, 1995).

Gene Duplications

In previous work (Glusman *et al.*, 1996), we described the analysis of the tandem duplication of an 11-kb-long fragment, mediated by recombination between mammalian-wide interspersed repeats (MIRs) and estimated to have taken place 90–100 million years ago. This recombinatorial event duplicated an entire gene structure, producing two apparently functional copies, which today represent genes from the same subfamily (3A). Analysis of the complete cluster sequence reveals additional instances of large duplications that include entire genes.

A tandem duplication of subfamily 1D genes. The telomeric end of the cluster includes an additional example of an ancient tandem duplication, also apparently mediated by mammalian-wide repeats (of the L1 family). Figure 5a shows a schematic diagram of the genomic region surrounding genes OR17-4 and OR17-31 from subfamily 1D. As in the subfamily 3A duplication, most of the duplicated sequences have diverged significantly, accepting several retroposon insertions and suffering various deletions. The two terminal L1 repeats (of subfamilies L1ME3 and L1M3a) can be discerned, as well as a hybrid L1 repeat between the two duplication arms. Two pairs of segments within the duplication display significantly higher degrees of similarity. These are the intronless coding region and an upstream segment suggested to include a noncoding exon, as well as a putative control region (see below). Excluding these more conserved segments, as well as later retropositions, the estimated substitution level from the original sequence is 15–20%, which corresponds to 56–65 million years ago. This is consistent with the older age (80–90 million years) of the L1M repeats flanking it. It is remarkable that in both tandem duplication events described, the coding region resides in the middle of the duplicated segment, while the putative control region is located in close proximity to its 5′ end. The short distance between the retroposons involved in the duplication mechanisms, and the gene control elements, suggests their location in a structurally more exposed region, potentially yielding an implicit mechanism for duplication of complete gene structures.

Sequence expansion by retroposition. An additional, similarly aged (58–65 million years old) tandem duplication can be discerned, mediated by mammalian-wide L1 repeats and including the complete OR17-6 and OR17-7 (boxed in Fig. 4). The duplicated sequences have also diverged significantly and expanded significantly, from 8–10 to 17–20 kb per duplication arm. This sequence expansion is due to repeated retroposon invasion. An even stronger sequence expansion can be observed following the OR17-4/OR17-31 duplication, with the region surrounding OR17-4 expanding from ~7 to ~30 kb. Most of the added sequence derives from L1 repeats, which enter both the intron and the intergenic sequence.

A recent dispersive duplication. The OR cluster region also contains the results of a very recent event of gene duplication in which 30 kb of sequence containing a full OR gene were copied, within a distance of ~160 kb. Comparison of the genomic sequences surrounding OR17-30 and OR17-31 by GESTALT (Figs. 5a and 5b), dot-plot, and identity plot (Fig. 5c) shows the existence of two distinct regions of similarity: one covers ~24 kb of sequence with >99% nucleotide sequence identity, and the other covers ~8 kb with somewhat lower sequence conservation (95 ± 3%). Both segments include a variety of repetitive sequences. The first duplication

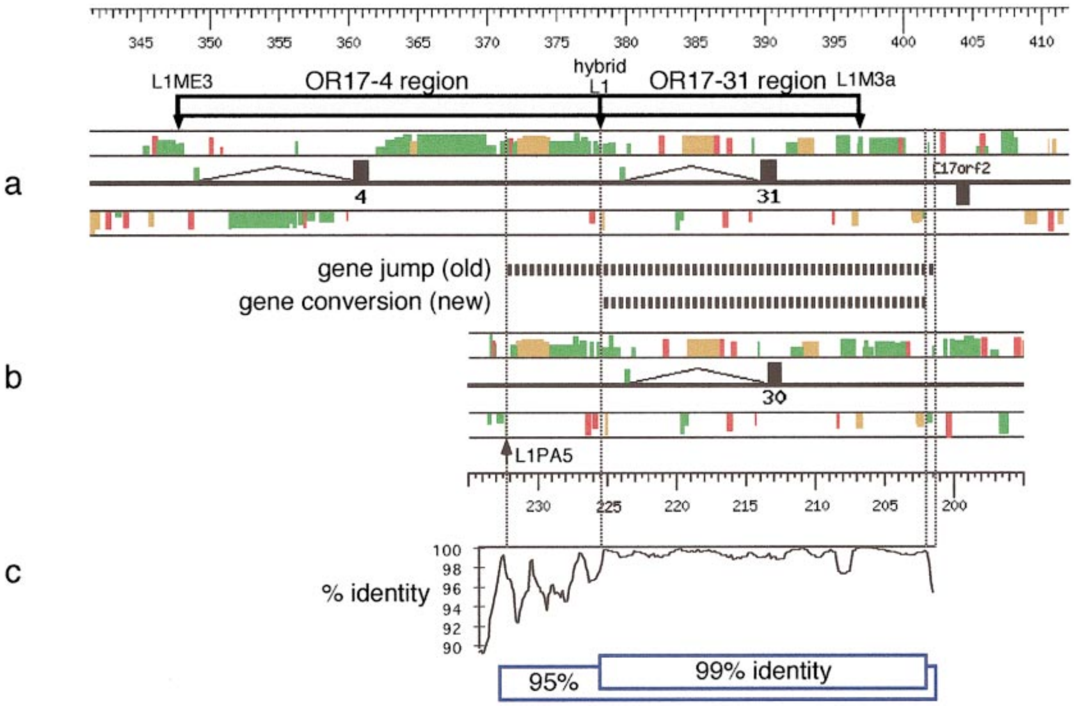


FIG. 5. Gene duplications in subfamily 1D. (**a**) Genomic region of OR17-4 and OR17-31; partial GESTALT map of genes and repeats as in Fig. 4. The old L1 repeats postulated to mediate the duplication event are indicated by arrows. (**b**) Map of the OR30/OR31 gene jumping and conversion events. The sequence segment including OR17-30 is reverse-complemented with respect to the absolute cluster orientation. Thick horizontal dashed lines indicate the extents of the recombinatorial events. Vertical shaded lines are an aid for visualization. (**c**) Identity plot of the duplicated segments including OR17-31 and OR17-30. Depiction of the hypothesis of extensive ectopic copying of 30 kb of sequence (currently 95% identical) followed by the homogenization of 24 kb (currently 99% identical).

segment (99% identity) includes OR17-30 or OR17-31 in its entirety (coding region, upstream intron, noncoding exon, and putative control region). Consistent with the high identity level of this region, no additional repetitive elements have retroposed into it, following the duplication. On the other hand, a young *Alu* repeat (7.1% divergent from *Alu*Y consensus, ∼50 Myr old) has retroposed into the second segment of the OR17-30 copy but not into its OR17-31 counterpart.

To analyze the mechanism leading to this duplication, we examined the sequences at the ends of the duplicated regions. A short (55 bp), very young L1PA5 repeat that flanks the OR17-30 segment at its telomeric end is followed by a 265-bp-long A+T-rich simple-sequence repeat. No short direct repeats flank the L1PA5 element, even though it is a very recent insertion. This L1PA5 retroposon also is not present at the corresponding end of the OR17-31 segment, even though its age is comparable with the divergence between the older duplicated segments. Retroposons enter the genome at specific sites, causing staggered, double-strand breaks (Jurka, 1997). Up to 8 kb of nonhomologous ectopic sequence were shown to be copied in P-element-induced double-strand gap repair in *Drosophila* (Nassif *et al.*, 1994). We hypothesize here that such a mechanism acted in the primate genome using the genomic surroundings of OR17-31 as template. The end result is the duplication of the 30-kb segment, with OR17-31 being the original and OR17-30 being the new copy. A later homogenization event then might have copied the 24-kb sequence including the complete gene, yielding the current structure. Since the sequence homogenized in this later event is entirely contained in the older, larger duplicated segment, the direction of transfer cannot be ascertained.

Cluster Organization and Evolution

Reconstruction of cluster history. The OR gene cluster under study has a very complex organization with genes lying in both orientations: eight genes from centromere to telomere and nine genes from telomere to centromere. A very weak correlation (0.45) can be seen between the orientation of each OR region and whether a gene is apparently functional or is a pseudogene. This suggests the absence of a single, directional "locus control region" for the entire cluster, as it would dictate a preferred orientation for functional genes.

Representatives of seven gene subfamilies are intermixed along the cluster. This is in sharp contrast with the largely unidirectional organization of many known multigene clusters, e.g., homeobox genes (Garcia-Fernandez and Holland, 1994), β-like globins (Fritsch *et al.*, 1980), and also ORs on human chromosome 3 (Brand-Arpon *et al.*, 1999). This uniform cluster organization usually results from repeated tandem duplications and may be functionally important.

The arrangement of the OR genes in the present cluster may be minimally explained by a rather com-

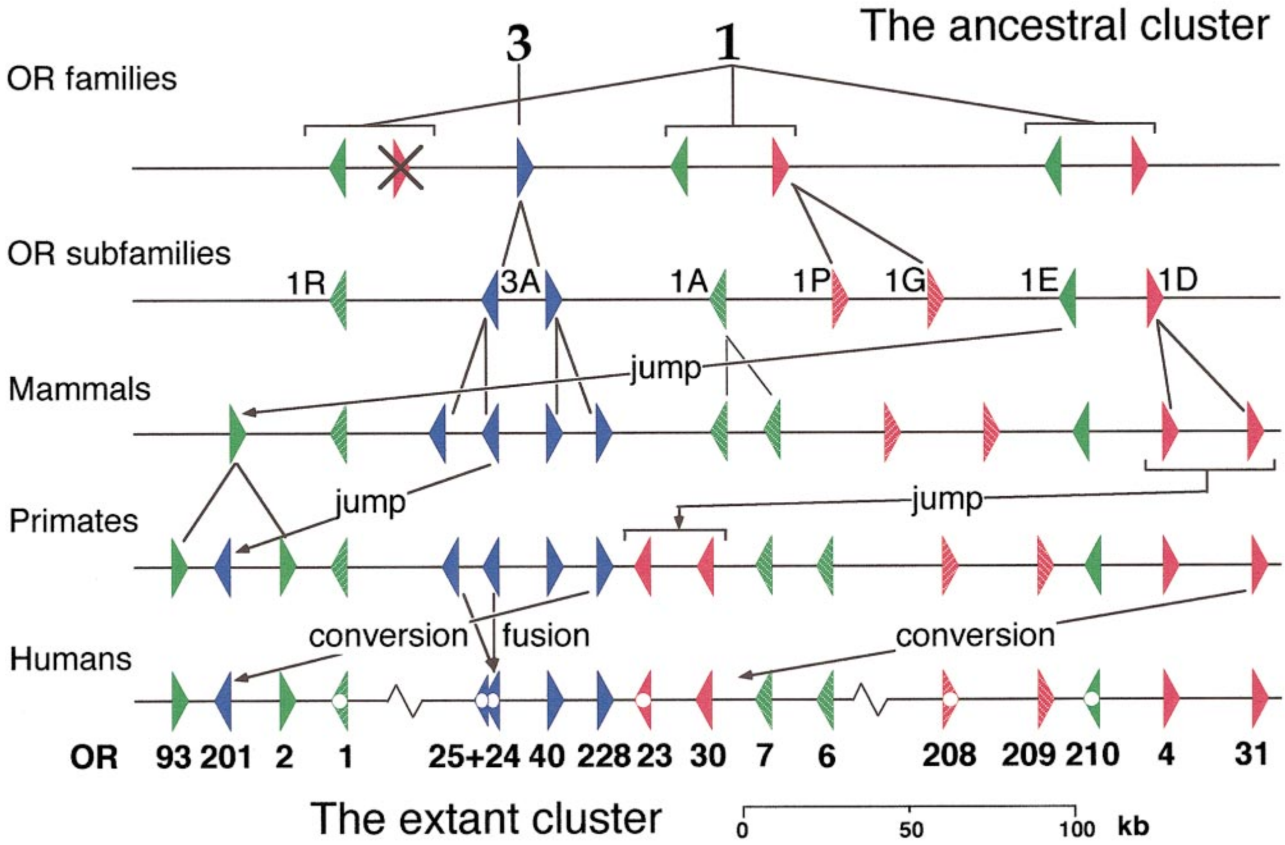


FIG. 6. A hypothetical reconstruction of the evolutionary history of the cluster, indicating duplication, jumping, and conversion events. Gene location, orientation, and pseudogenic status as in Fig. 1. The scale applies only to the map of the extant human cluster. See text for details.

plex series of evolutionary events, including repeated tandem duplications, copying of genes to remote locations within the cluster ("gene jumping"), repeated conversion events, and gene death by point mutation, deletion, and recombination. Based on the family/subfamily classification of the genes in the cluster, and on the estimated times for each duplication event, the possible evolutionary history of the cluster can be reconstructed (Fig. 6). It is likely that an ancestral cluster, composed of only two oppositely oriented genes of family 1, was tandemly duplicated, with each resulting gene becoming a subfamily founder. Later, several local duplication events and gene rearrangements most likely occurred within the cluster, as evidenced by their current high degrees of similarity.

The presence of members of family 3 within the cluster most likely occurred because, as depicted in Fig. 6, following the initial duplication of the ancestral two-gene cluster, one of the genes (crossed out in Fig. 6) was replaced (e.g., by conversion) by an "invading" family 3 founder gene. Alternatively, the ancestral cluster may have included the founder family 3 gene in addition to the family 1 members. The subfamily 1R founder may then represent an additional, ancient gene rearrangement event. To clarify this, it will be necessary to characterize the orthologous cluster in more remote vertebrate species (Lapidot *et al.*, in preparation). While comparison to a paralogous cluster that contains family 3 genes would also be informative, this is currently impossible, since only one additional family 3 member has been described in the human genome to date (OR5-83, see below).

Intriguingly, the family 3 "domain" in the middle of the cluster is on average more G+C-rich (42.9%) than the family 1 domains (40.2 and 40.7% centromeric and telomeric to the 3A region, respectively). Moreover, family 3 coding regions are in general more G+C-rich than family 1 members (Table 1). These observations suggest that the founder family 3 gene derived from a different genomic environment characterized by a higher G+C content, i.e., an H-type isochore (Bernardi, 1993). After integration into the L isochore of this cluster, the composition of the originally G+C-rich sequence apparently has changed to reflect that of the new environment, but the original G+C richness is still apparent, especially in the coding regions.

Several events can be discerned in which a gene is duplicated to a remote location along the cluster (i.e., gene jumping), which account for most of the intermixing of subfamilies. Such events could involve gene retroposition (Brosius, 1999). Some of the duplication events involving genes of the same subfamily are mediated by mammalian-wide repetitive elements (MIR and L1M). Since these likely occurred over 100 million

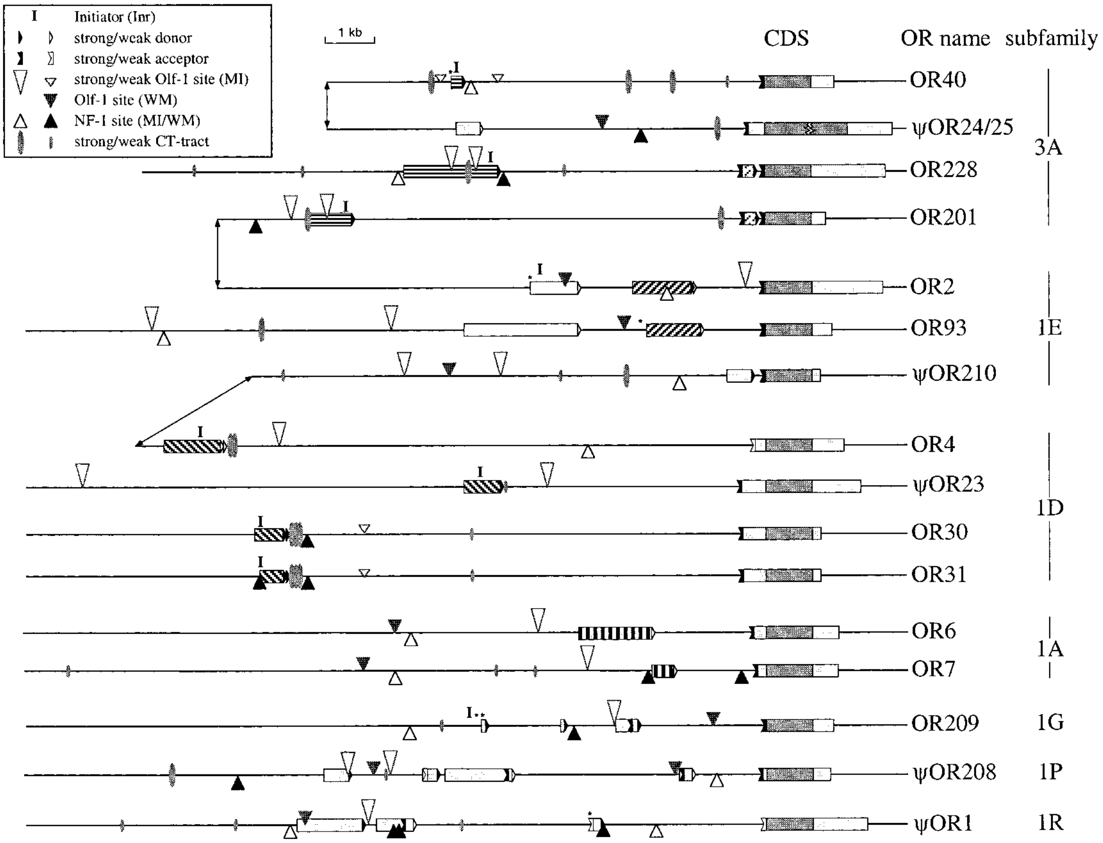


FIG. 7. Predicted OR gene structures. Dark gray boxes represent OR open reading frames for coding sequences (CDS). Light gray boxes represent the nontranslated parts of coding exon, from the predicted splicing acceptors to the predicted sites for polyadenylation. Predicted groups of upstream noncoding exons with similar sequences are denoted by boxes filled with identical patterns. Light gray upstream exons do not show similarity to any known OR genomic sequences. Upstream exons begin at predicted TSSs and end at predicted splicing acceptors. TSSs predicted by two programs at the same location are denoted by asterisks. MI, MatInspector. WM, WordMapper.

years ago, this gene cluster was established significantly before the mammalian radiation, potentially at the amphibian stage. The ancestral mammalian cluster is therefore predicted to have included 8–13 genes (Fig. 6). Since then, the genes in this cluster apparently have undergone little further amplification by repeated tandem duplication, in the primate lineage.

On the other hand, there is evidence for recent genetic exchange with other genomic loci, with several instances of genes from this cluster being copied into other chromosomes. The genomic clone G1 (Selbie *et al.*, 1992) is almost identical to OR17-4, but its genomic location is unknown. OR11-13 (*OR1D7P*, GenBank Accession No. AF065866) and OR11-22 (*OR1D6P*, GenBank Accession No. AF065868), on chromosome 11 (Buettner *et al.*, 1998), are >99% identical in nucleotide sequence to OR17-23. OR13-66 (U86222) on chromosome 13 (Rouquier *et al.*, 1998) is identical to OR17-2. Strikingly, OR5-83 (U86272) and OR5-85 (U86274) on chromosome 5 (Rouquier *et al.*, 1998) also are almost identical to OR17-201 and OR17-2, respectively, suggesting that at least 30 kb of chromosome 17 sequence, including genes from two different families (3 and 1), were duplicated into chromosome 5. It is therefore apparent that duplications of single OR genes between different chromosomes are not uncommon, without necessitating concomitant duplication of extensive genomic regions (Trask *et al.*, 1998).

The Conserved Structure of OR Genes

To predict the intron–exon structure of the genes in the cluster, including potential upstream noncoding exons, we analyzed the genomic environment of each OR coding region, concentrating on the features conserved between ORs that belong to the same subfamily. One to four upstream, noncoding exons were predicted for each OR gene in the cluster (Fig. 7). Exons contained within repetitive sequences were eliminated from the analyzed set, since we aimed to recognize similarity due to exon conservation, rather than due to similarity between repeats. The sequences of the predicted exons for each OR gene were aligned with the upstream genomic regions of all other ORs from the same subfamily, to recognize potential upstream exons conserved between ORs that belong to the same subfamily. In general, the noncoding exons predicted for genes of a given subfamily displayed sequence similarity, but the upstream exons of genes from different subfamilies or different families were much more divergent.

Subfamily 3A. The previously predicted upstream exon for OR17-40 and OR17-228 (Glusman *et al.*, 1996) also was present in OR17-201. An additional short exon was predicted for OR17-201 and OR17-228, in close proximity to the coding exon.

Subfamily 1D. An upstream exon was predicted for the four ORs from subfamily 1D (OR17-4, OR17-23, OR17-30, and OR17-31). The potential upstream exon for pseudogene OR17-23 was identified only by fex and GenScan.

Subfamily 1E. A single upstream exon was predicted for OR17-93. Its counterpart upstream of OR17-2 was predicted only by fex. It is worth noting that this predicted exon is located within a very old L2 repeat, which apparently was present before the gene duplication leading to OR17-2 and OR17-93. Part of this ancient repeat may have adopted a structural function as an upstream, noncoding exon. Regions of similarity between these two genomic sequences that contain no predicted exons are shown as open boxes in Fig. 7. The genomic sequence surrounding the OR17-210 pseudogene does not show any significant similarity with the upstream sequences of OR17-2 or OR17-93.

Subfamily 1A. One of the three upstream exons predicted for OR17-7 shows sequence similarity with the upstream exon predicted for OR17-6 (recognized only by Grail).

Subfamilies 1R, 1P, and 1G. OR17-1 and OR17-209 have three predicted upstream exons, while one upstream exon is predicted for OR17-208 (Fig. 7). Since full genomic sequences of additional ORs from subfamilies 1R, 1P, and 1G are unknown, further subfamily comparative analysis for OR17-1, OR17-208, and OR17-209 could not be performed.

Prediction of splicing sites. The two splice site prediction programs (SPL and SSPNN) complemented the exon prediction programs as they can detect potential cryptic or suboptimal sites. Splice acceptor sites were found to be localized 6–471 bp upstream to the start codon of all analyzed ORs (Fig. 7). The coding exons of two genes (OR17-4 and OR17-1) had weak acceptor sites with scores less than 0.70 according to SPL and less than 0.80 according to SSPNN. Acceptor sites also were predicted for the putative internal upstream exons of OR17-201, OR17-228, OR17-1, OR17-208, and OR17-209.

Donor splice sites for upstream exons were detected by the SPL program (score for weak sites less than 0.85). Interestingly, "donor doublets" were predicted for the upstream exons of all subfamily 1D ORs (OR17-4, OR17-23, OR17-30, and OR17-31). The observed donor doublet consensus sequence is GCAG-mACrGAgCAsTGG**GT**AGG**GT**syGkmyrbCTCAGsCy, where the boldface, underlined GTs five nucleotides apart represent the alternative splicing donors, and capitalized bases are conserved in the four sequences studied. This suggests that alternative splicing occurs in this subfamily.

Prediction of polyadenylation signals. A POLYAH predicted (Solovyev and Salamov, 1997) polyadenylation site occurs 3′ to the coding region for each OR gene in the cluster, indicating 3′-UTRs of 200 to 1500 bp (Fig. 7).

Prediction of transcription start sites. Potential promoters and corresponding transcription start sites (TSSs) were predicted by TSSG and TSSW (Solovyev and Salamov, 1997) and by PPNN (Reese *et al.*, 1996). The TSSs predicted by both PPNN and by either TSSG or TSSW are marked as asterisks in Fig. 7. In addition, the very highly scoring TSS predicted for one of the upstream exons of OR17-209 is indicated by a double asterisk in Fig. 7. The predicted promoters are all TATA-less, like the promoters of other olfactory-specific genes (Wang *et al.*, 1993) and as suggested by the preliminary analysis (Glusman *et al.*, 1996). Initiator (Inr) sequences (Javahery *et al.*, 1994) are present in the upstream regions of ORs from subfamilies 3A (excluding the OR17-24 pseudogene), and 1D, as well as OR17-2 and OR17-209 (Fig. 7). The predicted Inr sites do not coincide with the promoters predicted by TSSG and TSSW but are located within 800 bp of suitable splice donor sites.

Potential Transcriptional Control Signals

The availability of the complete sequence of the cluster provided us with the first opportunity to compare the upstream genomic regions for the OR genes that are clustered and that might be expected to share common control features. A dot-plot and ClustalX alignment comparison of the 15 kb upstream from each of 16 OR ORFs in the cluster (no upstream sequences are available for OR17-25) showed significantly conserved segments within subfamilies, but no extensive sequence conservation of upstream regions either between subfamilies or between families. It is apparent, therefore, that genes belonging to different subfamilies have diverged significantly in their upstream regions.

No recombinatorial signal sequences. It can be hypothesized that the clonal exclusion of ORs is at least partially based on somatic recombination, which would then join an OR gene to a putative locus control region. Somatic recombination joins gene segments in immunoglobulin heavy-chain genes via recombinatorial signal sequences, or RSSs (Sakano *et al.*, 1981). The WordMapper tool was used to detect such signals in this cluster. No suitable RSSs were found in the genomic environments upstream of the OR coding regions.

Detection of a specific CT tract. A global comparison of all 16 sequences found, as expected, significantly shared patterns among members of each subfamily. Seven of the 16 genes were therefore selected as representatives of the different subfamilies for further analysis (OR17-7, OR17-31, OR17-2, OR17-209, OR17-

```
228   -6027     aaccggctttgaagaaagCTTtTCCCTcTTaTCTCccttctggggcctcctcct
 40   -6730     agccaaggttggagaaagCTTcTCCCTtTTgTCTCcattcttgtgcctccttct
 40   -2766     attagtatattcctgtttCTTtTCCCTtTTcTCTCtctcctataccatgtaatt
208  -12051 c   tacggagatggtcccttcCTTtTCCCTccTcTCTCtctctctttcccttccttc
 30   -9773     cgcctcaggctcatgctcCTTcTCCCTcTTccCTCttatcttctcctccatttt
 31   -9752     cgcctcaggctcatgctcCTTcTCCCTcTTccCTCttatcttctcctccatttt
  4  -11077     catctcagcctcatgctctTTcTCCCTcTTcTCTCttatcttctctcttatctt
201    -880 c   aaaaaaaaaaaagatagaaTTtTCCCatTTtTCTCtgagggatcaacccaaact
201   -9262     aaccagctttgaagaacaCTTtcCCCTcTTaTaTCccttctggggcctcctcct
210   -2824 c   cctctccctcctctccctCcTcTCCCTcTcgTCTCcccttccacggtctccct
 24    -951 c   aactccaccatcactgagCTTtTCaCTcTTcTgTCtcctggttctaattacgca
 30   -9766     ggctcatgctccttctccCTctTCCCTcTTaTCTtctcctccatttctctcat
 31   -9745     ggctcatgctccttctccCTctTCCCTcTTaTCTtctcctccatttctctcat
  4  -11037     ttctctcttatcttctctCTTcTCCaTtTTcTCTgtcacacaaacatactcaca
  1   -6147 c   tgttagtgaccctggcctCTTcTCCCTggTtTtTCccaccctgagctacaacca
 30   -9749     cctcttccctcttatcttCTccTCCaTtTTcTCTCatatttgtgggataaaaaa
 31   -9728     cctcttccctcttatcttCTccTCCaTtTTcTCTCatatttgtgggataaaaaa
 40   -1848     tcttttctcatatctgtcaTTtTCtCTcTTtTCTCattgcatcttgtcatcttt
 93  -10203 c   tttgtaccctttgacctaCaTcTCCCTaTTccCTCctccttgagtaaattgttt
208   -7691 c   agccaccatcacaatgacCTgtTgCCTcTTcTgTCctctacatcatcggaaata
209   -6589     aggtacacatggaaatcgtTTtTCtCTccTgTCTCtcatccaattccatacctc
 23   -5301     ttccttcctccctccctcCTTcTCttTcTTtTCTttttctcttttctttctctt
 23   -5286     ctccttctctttcttttctTTtTCtCTtTTcTtTCtctttccttccttcattcc
 23   -5158     ctctttccttccttctttCTTcTttCctTTtTCTCtttcccttccttccttcct
 23   -5086     tttctttctttcctttctCTTtTtCtTtTTcTtTCttccttccttcctttttct
  7   -5464     gtgcccgtcttcttttttCTTtTCCaTtTTtTtTttttagatggggtctcacac
  7   -4654     tcttgatagcatacacaaaTTtTCCCTtTatTtTCcactaaatttttgacgtt
  7  -14138 c   agctttcttcactcatttCTatTCttTtTTtTCTCcctttactaaattatttcc

CT-tract consensus  yy   m yy yw   yyycTtYTCccTYTTyTcTcyywyyy      yyymyyyw
```

FIG. 8. Multiple alignment of the CT-tract sequences, listing gene name, pattern position relative to the ATG codon, orientation ("c" for complementary strand), and actual sequence. In the consensus line, uppercase indicates consensus by plurality of 85 or 90% for unambiguous and ambiguous bases, respectively. Y denotes pyrimidines; M denotes A or C; W denotes A or T. Shaded bases indicate matches to the consensus in positions where the consensus is unambiguous.

208, OR17-1, and OR17-40, for subfamilies 1A, 1D, 1E, 1G, 1P, 1R, and 3A, respectively). The genomic environments of the 7 selected genes were examined with the oligonucleotide analysis tool (van Helden *et al.*, 1998) from the Yeast Regulatory Tools (van Helden *et al.*, in preparation), using pattern length of 8. As expected, the highest scores (representing patterns present in most of the sequences) corresponded to patterns that are part of *Alu* repeats. Ignoring these, the highest scoring patterns were seen for pyrimidine:purine (Y:R) tracts, the CA repeat, and CpG-containing patterns. A similar analysis on sequences in which interspersed repeats were premasked again gave highest scores to Y:R tracts. Similar results were obtained by analyzing both strands simultaneously (not shown).

When the segment pair overlap method of MACAW (Schuler *et al.*, 1991) was used to detect longer conserved sequences, a pyrimidine-rich segment (hereafter named CT tract) with consensus CTTYTCCCTYTT-NTCTCY was found. Using the Word Mapper tool of the GESTALT Workbench, the positions with significant similarity to this consensus were detected and mapped (Figs. 7 and 8) and also could be correlated with the predicted splice donor sites in a subfamily-specific fashion. Specifically, the CT tract is contained within the putative control region conserved in genes of subfamily 3A (Glusman *et al.*, 1996), as well as in the noncoding conserved sequences of subfamily 1D.

To study the generality of these findings, the Gibbs sampler method was used through the Web interface of the Yeast Regulatory Tools (van Helden *et al.*, in preparation). Using patterns of various lengths, but especially ≧30, only Y:R tracts that comap with the CT-tract motif detected using MACAW were detected. The sequences surrounding the CT tracts are enriched in C+T beyond the specific consensus sequence described (Fig. 8).

Therefore, the most significant pattern common to most potential control regions, beyond trivial similarities deriving from either historical conservation (between genes in one subfamily) or sequence repetition (of retroposons), was the presence of pyrimidine:purine tracts, which are located near splice donors. The CT tracts could in principle be an olfactory-specific recombinatorial signal. On the other hand, their location 3′ to the putative upstream exons weakens this possibility, as such exons and splicing signals would be missing from the selected gene.

Pyrimidine:purine tracts have been shown to promote unwinding of the double-helix (Bucher and Yagil, 1991) and to be implicated in regulation of transcription and in posttranslational regulation (Valcarcel and Gebauer, 1997). Within the observed tracts, a specific motif (CT tract) could be defined as consensus, suggesting the conservation of specific patterns for transcription factor binding.

Mapping of transcription factor binding sites. Two families of transcription factors are expressed in the neurons of the olfactory epithelium: the O/E family, including Olf-1, Olf-2, and Olf-3 (Wang *et al.*, 1997); and the NF-1 family (Baumeister *et al.*, 1999). Olf-1 and NF-1 binding has been demonstrated for promoters of the olfactory-specific genes: OMP, type III adenylyl cyclase, and olfactory cyclic nucleotide gated channel. For $G_{olf\alpha}$, only Olf-1 binding has been shown (Wang *et al.*, 1993; Baumeister *et al.*, 1999). Several potential binding sites for O/E and NF-1 transcription factors now have been identified in the genomic surroundings (up to 15 kb upstream) of each OR gene in the cluster (Fig. 7).

Using MatInspector, one or two Olf-1 sites were found with scores above 0.850 for most of the analyzed

ORs (Fig. 7). Predicted Olf-1 sites for OR17-40, OR17-30, and OR17-31 have scores from 0.820 to 0.835. Additional Olf-1 sites were observed upstream of OR17-24, OR17-2, OR17-93, OR17-210, OR17-6, OR17-7, OR17-209, OR17-208, and OR17-1 when mapping the Olf-1 consensus using the Word Mapper tool. Generally, the Olf-1 sites for these ORs were predicted with a lower score than Olf-1 sites for other olfactory neuron-specific genes. This most likely is because the previously described Olf-1 sites were located in the rat genomic sequences, with the sole exception of the human OMP Olf-1 site (Buiakova *et al.*, 1994). Therefore, a low score for human predicted Olf-1 sites might reflect interspecies differences. In addition, Olf-1, Olf-2, and Olf-3 bind to Olf-1 sites of olfactory neuron-specific genes with different affinity (Wang *et al.*, 1997). Thus, it is likely that different members of the O/E transcription factor family bind *in vivo* to OR Olf-1 and to other Olf-1 sites. Except for the OR17-23 pseudogene, all analyzed gene upstream regions were found to contain strong potential NF-1 binding sites, but no conserved patterns for NF-1 localization could be distinguished within the analyzed OR subfamilies.

CONCLUSIONS

A large human genomic region including a cluster of 17 genes of the OR superfamily has now been fully sequenced and characterized. The only potential non-OR gene identified was at the telomeric margin, suggesting that this uninterrupted cluster evolved by repeated expansion. The inferred primordial cluster, suggested to have been established in an early amphibian ancestor, presumably included only a few OR genes, which gave rise to the two different gene families observed in the extant sequence. The cluster has not evolved by simple tandem multiplication of its initial components, but has apparently grown in complexity by several recombinatorial events, some of which are relatively recent. For some of the recombinations, a mechanism may be discerned, involving interspersed repeats (retroposons). The interspersed repeats represent 60% of the sequence in the cluster and belong mainly to the LINE family of retroposons, though SINEs and DNA transposons are also present. The intergenic distances vary significantly (5–67 kb) and are related to the amount of inserted repetitive sequences. At this stage, it is unclear whether repetitive sequences within the OR genes affect their expression to any extent.

Significantly, the observed recombinatorial events involve complete genes, suggesting an evolutionary mechanism for preserving intact gene structures upon duplication. The common gene structure has been delineated by computational analysis of the OR genes. This was found to include an intronless terminal coding exon, terminated by a signal for polyadenylation (0.15–1.5 kb downstream from the stop codon) and preceded by introns (0.5–11 kb long) and by one or two short, noncoding upstream exons. The resulting common gene structure is consistent with that which we previously described for OR genes belonging to family 3, with the addition of the possible existence of more than one upstream noncoding exon for each gene. The functional role of this stereotyped structure is still unknown. The upstream noncoding exons might play a role in the control of mRNA fate or subcellular localization.

When the complete genes in this OR cluster are compared, several levels of conservation may be discerned. Within each subfamily, the coding sequences are most conserved, the putative control regions and noncoding upstream exons show an intermediate level of conservation, while the introns and the intergenic sequences are the least conserved. Between subfamilies, the overall intron–exon structure of the genes is more conserved than the specific location and the quality of the relevant splice signals, while the putative control sequences are the most divergent, with only their pyrimidine:purine tracts and Olf-1 transcription factor binding sites conserved. The Olf-1 transcription factor binding sites may play an important role in olfactory-specific transcription of the OR genes, while the pyrimidine:purine tracts, previously shown to promote melting of the double-helix for transcription initiation, may serve an auxilliary control function.

A sizable fraction (6 of 17) of the coding regions in the cluster are pseudogenes. One of these (OR17-1) apparently has shifted function to become a CpG island. Other examples are known where OR genes adopted new, noncoding functions, e.g., promoters and matrix attachment regions. This appears to indicate that OR coding regions have a special plasticity, allowing them to evolve new functionalities. A potential explanation of this versatility, as well as the prevalence of pseudogenes, may reside in the variability within the OR superfamily and the partial functional redundancy of OR genes.

Clustering of the OR genes may play an important role for initiation of their transcription by common enhancers. Each of the identified OR genes appears to have its own, independent TATA-less promoter region. This finding and the apparent lack of recombinatorial signal sequences suggest the importance of *trans*-acting factors for regulating the excluded cellular expression of single OR genes in the cluster, rather than a somatic DNA rearrangement mechanism. The cluster includes CpG islands, potentially affecting OR gene expression. Two of the observed CpG islands derive from recently inserted SVA retroviral elements, presumably absent from the genomes of New World monkeys and other mammals. Further work will be required to ascertain the functional role of these potential regulatory signals.

ACKNOWLEDGMENTS

We thank Dr. Dror Sharon, Dr. Irit Rubin, and Yitzhak Pilpel for fruitful discussions and valuable suggestions, and Dr. Jaime Prilusky for unwavering support. Doron Lancet holds the Ralph and Lois Chair in Neurogenomics and is supported by the Crown Human Genome Center, a Ministry of Science grant to the National Laboratory for Genome Infrastructure, the National Institutes of Health (DC00305), the Krupp Foundation, the German–Israeli Foundation for scientific research and development, and the Glasberg, Levy, Nathan, Brunschweig, and Levine funds.

REFERENCES

Asai, H., Kasai, H., Matsuda, Y., Yamazaki, N., Nagawa, F., Sakano, H., and Tsuboi, A. (1996). Genomic structure and transcription of a murine odorant receptor gene: Differential initiation of transcription in the olfactory and testicular cells. *Biochem. Biophys. Res. Commun.* **221:** 240–247.

Bailey, W. J., Fitch, D. H., Tagle, D. A., Czelusniak, J., Slightom, J. L., and Goodman, M. (1991). Molecular evolution of the psi eta-globin gene locus: Gibbon phylogeny and the hominoid slowdown. *Mol. Biol. Evol.* **8:** 155–184.

Barth, A. L., Dugas, J. C., and Ngai, J. (1997). Noncoordinate expression of odorant receptor genes tightly linked in the zebrafish genome. *Neuron* **19:** 359–369.

Baumeister, H., Gronostajski, R. M., Lyons, G. E., and Margolis, F. L. (1999). Identification of NFI-binding sites and cloning of NFI-cDNAs suggest a regulatory role for NFI transcription factors in olfactory neuron gene expression. *Brain Res. Mol. Brain Res.* **72:** 65–79.

Ben-Arie, N., Lancet, D., Taylor, C., Khen, M., Walker, N., Ledbetter, D. H., Carrozzo, R., Patel, K., Sheer, D., Lehrach, H., and North, M. A. (1994). Olfactory receptor gene cluster on human chromosome 17: Possible duplication of an ancestral receptor repertoire. *Hum. Mol. Genet.* **3:** 229–235.

Ben-Arie, N., North, M., Khen, M., Gross-Isserof, R., Walker, N., Horn-Saban, S., Gat, U., Natochin, M., Lehrach, H., and Lancet, D. (1993). "Olfactory Reception: from Signal Modulation to Human Genome Mapping," Vol. 1, pp. 122–126, Springer-Verlag, Berlin/New York, Koseinenkin Kaikan, Sapporo, Japan.

Bernardi, G. (1993). The isochore organization of the human genome and its evolutionary history—A review. *Gene* **135:** 57–66.

Bodenteich, A., Chissoe, S., Wang, Y. F., and Roe, B. A. (1993). Shotgun cloning as the strategy of choice to generate templates for high-throughput dideoxynucleotide sequencing. *In* "Automated DNA Sequencing and Analysis Techniques" (J. C. Venter, Eds.), pp. 42–50, Academic Press, London.

Brand-Arpon, V., Rouquier, S., Massa, H., de Jong, P. J., Ferraz, C., Ioannou, P. A., Demaille, J. G., Trask, B. J., and Giorgi, D. (1999). A genomic region encompassing a cluster of olfactory receptor genes and a myosin light chain kinase (MYLK) gene is duplicated on human chromosome regions 3q13–q21 and 3p13. *Genomics* **56:** 98–110.

Brandeis, M., Frank, D., Keshet, I., Siegfried, Z., Mendelsohn, M., Nemes, A., Temper, V., Razin, A., and Cedar, H. (1994). Sp1 elements protect a CpG island from de novo methylation. *Nature* **371:** 435–438.

Britten, R. J. (1994). Evolutionary selection against change in many Alu repeat sequences interspersed through primate genomes. *Proc. Natl. Acad. Sci. USA* **91:** 5992–5996.

Brosius, J. (1999). Many G-protein-coupled receptors are encoded by retrogenes. *Trends Genet.* **15:** 304–305.

Brunak, S., Engelbrecht, J., and Knudsen, S. (1991). Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J. Mol. Biol.* **220:** 49–65.

Bucher, P., and Yagil, G. (1991). Occurrence of oligopurine · oligopyrimidine tracts in eukaryotic and prokaryotic genes. *DNA Seq.* **1:** 157–172.

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

Buettner, J. A., Glusman, G., Ben-Arie, N., Ramos, P., Lancet, D., and Evans, G. A. (1998). Organization and evolution of olfactory receptor genes on human chromosome 11. *Genomics* **53:** 56–68.

Buiakova, O. I., Krishna, N. S., Getchell, T. V., and Margolis, F. L. (1994). Human and rodent OMP genes: Conservation of structural and regulatory motifs and cellular localization. *Genomics* **20:** 452–462.

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268:** 78–94.

Carver, E. A., Issel-Tarver, L., Rine, J., Olsen, A. S., and Stubbs, L. (1998). Location of mouse and human genes corresponding to conserved canine olfactory receptor gene subfamilies. *Mamm. Genome* **9:** 349–354.

Chalifa-Caspi, V., Rebhan, M., Prilusky, J., and Lancet, D. (1997). The Unified Database (UDB): A novel genome integration concept. *Genome Digest* **4:** 15.

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

Chissoe, S. L., Bodenteich, A., Wang, Y.-F., Wang, Y.-P., Burian, D., Clifton, S. W., Crabtree, J., Freeman, A., Iyer, K., Li, J. A., Ma, Y., McLaury, H.-J., Pan, H.-Q., Sarhan, O., Toth, S., Wang, Z., Zhang, G., Heisterkamp, N., Groffen, J., and Roe, B. A. (1995). Sequence and analysis of the human ABL gene BCR gene and regions involved in the Philadelphia chromosomal translocation. *Genomics* **27:** 67–82.

Crowe, M. L., Perry, B. N., and Connerton, I. F. (1996). Olfactory receptor-encoding genes and pseudogenes are expressed in humans. *Gene* **169:** 247–249.

Engels, W. R. (1993). Contributing software to the internet: The Amplify program. *Trends Biochem. Sci.* **18:** 448–450.

Fritsch, E. F., Lawn, R. M., and Maniatis, T. (1980). Molecular cloning and characterization of the human beta-like globin gene cluster. *Cell* **19:** 959–972.

Garcia-Fernandez, J., and Holland, P. W. (1994). Archetypal organization of the amphioxus Hox gene cluster. *Nature* **370:** 563–566.

Gardiner, K. (1995). Human genome organization. *Curr. Opin. Genet. Dev.* **5:** 315–322.

Gentles, A. J., and Karlin, S. (1999). Why are human G-protein-coupled receptors predominantly intronless? *Trends Genet.* **15:** 47–49.

Gimelbrant, A. A., and McClintock, T. S. (1997). A nuclear matrix attachment region is highly homologous to a conserved domain of olfactory receptors. *J. Mol. Neurosci.* **9:** 61–63.

Glusman, G., Clifton, S., Roe, R., and Lancet, D. (1996). Sequence analysis in the olfactory receptor gene cluster on human chromosome 17: Recombinatorial events affecting receptor diversity. *Genomics* **37:** 147–160.

Glusman, G., Sharon, D., Kalush, F., Clifton, S., Roe, B., Ferraz, C., Demaille, J., Ben-Asher, E., and Lancet, D. (1998). Genome dynamics, polymorphisms and mutations in a 350 kb human olfactory receptor gene cluster. *Math. Model. Sci. Comput.* **9:** 30–44.

Hanke, J. H., Hambor, J. E., and Kavathas, P. (1995). Repetitive *Alu* elements form a cruciform structure that regulates the function of the human CD8 alpha T cell-specific enhancer. *J. Mol. Biol.* **246:** 63–73.

Higgins, D. G., Thompson, J. D., and Gibson, T. J. (1996). Using CLUSTAL for multiple sequence alignments. *Methods Enzymol.* **266:** 383–402.

Ioannou, P. A., Amemiya, C. T., Garnes, J., Kroisel, P. M., Shizuya, H., Chen, C., Batzer, M. A., and de Jong, P. J. (1994). A new bacteriophage P1-derived vector for the propagation of large human DNA fragments. *Nat. Genet.* **6:** 84–89.

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

Javahery, R., Khachi, A., Lo, K., Zenzie-Gregory, B., and Smale, S. T. (1994). DNA sequence requirements for transcriptional initiator activity in mammalian cells. *Mol. Cell. Biol.* **14:** 116–127.

Jukes, T. H., and Cantor, C. R. (1969). Evolution of protein molecules. *In* "Mammalian Protein Metabolism" (H. N. Munro, Eds.), pp. 21–123, Academic Press, New York.

Jurka, J. (1997). Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc. Natl. Acad. Sci. USA* **94:** 1872–1877.

Karlin, S., Campbell, A. M., and Mrazek, J. (1998). Comparative DNA analysis across diverse genomes. *Annu. Rev. Genet.* **32:** 185–225.

Kulp, D., Haussler, D., Reese, M. G., and Eeckman, F. H. (1996). A generalized hidden Markov model for the recognition of human genes in DNA. *Ismb* **4:** 134–142.

Lancet, D. (1986). Vertebrate olfactory reception. *Annu. Rev. Neurosci.* **9:** 329–355.

Lancet, D. (1991). Olfaction. The strong scent of success. *Nature* **351:** 275–276. [News]

Lancet, D., Sadovsky, E., and Seidemann, E. (1993). Probability model for molecular recognition in biological receptor repertoires: Significance to the olfactory system. *Proc. Natl. Acad. Sci. USA* **90:** 3715–3719.

Laurent, G. (1997). Olfactory processing: Maps, time and codes. *Curr. Opin. Neurobiol.* **7:** 547–553.

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

Merino, E., Balbas, P., Puente, J. L., and Bolivar, F. (1994). Antisense overlapping open reading frames in genes from bacteria to humans. *Nucleic Acids Res.* **22:** 1903–1908.

Mombaerts, P. (1999). Molecular biology of odorant receptors in vertebrates. *Annu. Rev. Neurosci.* **22:** 487–509.

Nassif, N., Penney, J., Pal, S., Engels, W. R., and Gloor, G. B. (1994). Efficient copying of nonhomologous sequences from ectopic sites via P-element-induced gap repair. *Mol. Cell. Biol.* **14:** 1613–1625.

Nikolaev, L. G., Tsevegiyn, T., Akopov, S. B., Ashworth, L. K., and Sverdlov, E. D. (1996). Construction of a chromosome specific library of human MARs and mapping of matrix attachment regions on human chromosome 19. *Nucleic Acids Res.* **24:** 1330–1336.

Nizetic, D., Zehetner, G., Monaco, A. P., Gellen, L., Young, B. D., and Lehrach, H. (1991). Construction, arraying, and high-density screening of large insert libraries of human chromosomes X and 21: Their potential use as reference libraries. *Proc. Natl. Acad. Sci. USA* **88:** 3233–3237.

Page, R. D. (1996). TreeView: An application to display phylogenetic trees on personal computers. *Comput. Appl. Biosci.* **12:** 357–358.

Pearson, W. R., Wood, T., Zhang, Z., and Miller, W. (1997). Comparison of DNA sequences with protein sequences. *Genomics* **46:** 24–36.

Pilpel, Y., Sosinsky, A., and Lancet, D. (1998). Molecular biology of olfactory receptors. *Essays Biochem.* **33:** 93–104.

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

Quandt, K., Frech, K., Karas, H., Wingender, E., and Werner, T. (1995). MatInd and MatInspector: New fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.* **23:** 4878–4884.

Raja, M., Zevin-Sonkin, D., Shwartzburd, J., Rozovskaya, T. A., Sobolev, I. A., Chertkov, O., Ramanathan, V., Lvovsky, L., and Ulanovsky, L. E. (1997). DNA sequencing using differential extension with nucleotide subsets (DENS). *Nucleic Acids Res.* **25:** 800–805.

Reese, M. G., Harris, N. L., and Eeckman, F. H. (1996). "Large Scale Sequencing Specific Neural Networks for Promoter and Splice Site Recognition," World Scientific, Singapore.

Ressler, K. J., Sullivan, S. L., and Buck, L. B. (1993). A zonal organization of odorant receptor gene expression in the olfactory epithelium. *Cell* **73:** 597–609.

Rouquier, S., Taviaux, S., Trask, B. J., Brand-Arpon, V., van den Engh, G., Demaille, J., and Giorgi, D. (1998). Distribution of olfactory receptor genes in the human genome. *Nat. Genet.* **18:** 243–250.

Rowen, L., and Koop, B. F. (1994). Zen and the art of large-scale genomic sequencing. *In* "Automated DNA Sequencing and Analysis" (M. D. Adams, C., Fields, and J. C. Venter, Eds.), pp. 167–174, Academic Press, New York.

Rychlik, W. (1995). Selection of primers for polymerase chain reaction. *Mol. Biotechnol.* **3:** 129–134.

Sakano, H., Kurosawa, Y., Weigert, M., and Tonegawa, S. (1981). Identification and nucleotide sequence of a diversity DNA segment (D) of immunoglobulin heavy-chain genes. *Nature* **290:** 562–565.

Schuler, G. D., Altschul, S. F., and Lipman, D. J. (1991). A workbench for multiple alignment construction and analysis. *Proteins* **9:** 180–190.

Selbie, L. A., Townsend Nicholson, A., Iismaa, T. P., and Shine, J. (1992). Novel G protein-coupled receptors: A gene family of putative human olfactory receptor sequences. *Brain Res. Mol. Brain Res.* **13:** 159–163.

Sharon, D., Glusman, G., Pilpel, Y., Khen, M., Gruetzner, F., Haaf, T., and Lancet, D. (1999). Primate evolution of an olfactory receptor cluster: Diversification by gene conversion and recent emergence of pseudogenes. *Genomics* **61:** 24–36.

Shen, L., Wu, L. C., Sanlioglu, S., Chen, R., Mendoza, A. R., Dangel, A. W., Carroll, M. C., Zipf, W. B., and Yu, C. Y. (1994). Structure and genetics of the partially duplicated gene RP located immediately upstream of the complement C4A and the C4B genes in the HLA class III region. Molecular cloning, exon–intron structure, composite retroposon, and breakpoint of gene duplication. *J. Biol. Chem.* **269:** 8466–8476.

Smit, A. F. A., and Green, P. (1997). RepeatMasker at http://repeatmasker.genome.washington.edu/cgi-bin/RM2_req.pl.

Solovyev, V., and Salamov, A. (1997). The Gene-Finder computer tools for analysis of human and model organisms genome sequences. *Ismb* **5:** 294–302.

Solovyev, V. V., Salamov, A. A., and Lawrence, C. B. (1995). Identification of human gene structure using linear discriminant functions and dynamic programming. *Ismb* **3:** 367–375.

Sullivan, S. L., Adamson, M. C., Ressler, K. J., Kozak, C. A., and Buck, L. B. (1996). The chromosomal distribution of mouse odorant receptor genes. *Proc. Natl. Acad. Sci. USA* **93:** 884–888.

Trask, B. J., Friedman, C., Martin-Gallardo, A., Rowen, L., Akinbami, C., Blankenship, J., Collins, C., Giorgi, D., Iadonato, S., Johnson, F., Kuo, W. L., Massa, H., Morrish, T., Naylor, S., Nguyen, O. T., Rouquier, S., Smith, T., Wong, D. J., Youngblom, J., and van den Engh, G. (1998). Members of the olfactory receptor gene family are contained in large blocks of DNA duplicated polymorphically near the ends of human chromosomes. *Hum. Mol. Genet.* **7:** 13–26.

Uberbacher, E. C., Xu, Y., and Mural, R. J. (1996). Discovering and understanding genes in human DNA sequence using GRAIL. *Methods Enzymol.* **266:** 259–281.

Valcarcel, J., and Gebauer, F. (1997). Post-transcriptional regulation: The dawn of PTB. *Curr. Biol.* **7:** R705–R708.

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.

van Helden, J., André, B., and Collado-Vides, J. (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.* **281:** 827–842.

van Helden, J., André, B., and Collado-Vides, J. A Web site for the analysis of regulatory sequences in the yeast *Saccharomyces cerevisiae.* In preparation.

Vassar, R., Ngai, J., and Axel, R. (1993). Spatial segregation of odorant receptor expression in the mammalian olfactory epithelium. *Cell* **74:** 309–318.

von Sternberg, R. M., Novick, G. E., Gao, G. P., and Herrera, R. J. (1992). Genome canalization: The coevolution of transposable and interspersed repetitive elements with single copy DNA. *Genetica* **86:** 215–246.

Walensky, L. D., Ruat, M., Bakin, R. E., Blackshaw, S., Ronnett, G. V., and Snyder, S. H. (1998). Two novel odorant receptor families expressed in spermatids undergo 5′-splicing. *J. Biol. Chem.* **273:** 9378–9387.

Wang, M. M., Tsai, R. Y., Schrader, K. A., and Reed, R. R. (1993). Genes encoding components of the olfactory signal transduction cascade contain a DNA binding site that may direct neuronal expression. *Mol. Cell. Biol.* **13:** 5805–5813.

Wang, S. S., Tsai, R. Y. L., and Reed, R. R. (1997). The characterization of the Olf-1/EBF-like HLH transcription factor family: Implications in olfactory gene regulation and neuronal development. *J. Neurosci.* **17:** 4149–4158.

Xu, Y., Mural, R., Shah, M., and Uberbacher, E. (1994). Recognizing exons in genomic sequence using GRAIL II. *Genet. Eng.* **16:** 241–253.