Genomic Analysis of the Olfactory Receptor Region of the Mouse and Human T-Cell Receptor α/δ Loci

Robert P. Lane,1,3,6 Jared C. Roach,1,4 Inyoul Y. Lee,1,4 Cecilia Boysen,2,5 Arian Smit,1,4 Barbara J. Trask,1,3 and Leroy Hood1,4

1Department of Molecular Biotechnology, University of Washington, Seattle, Washington 98195, USA; 2California Institute of Technology, Division of Biology, Pasadena, California 91125, USA

We have conducted a comparative genomic analysis of several olfactory receptor (OR) genes that lie immediately 5′ to the V-α gene segments at the mouse and human T-cell receptor (TCR) α/δ loci. Five OR genes are identified in the human cluster. The murine cluster has at least six OR genes; the first five are orthologous to the human genes. The sixth mouse gene has arisen since mouse-human divergence by a duplication of a ~10-kb block. One pair of OR paralogs found at the mouse and human loci are more similar to each other than to their corresponding orthologs. This paralogous “twinning” appears to be under selection, perhaps to increase sensitivity to particular odorants or to resolve structurally-similar odorants. The promoter regions of the mouse OR genes were identified by RACE-PCR. Orthologs share extensive 5′ UTR homology, but we find no significant similarity among paralogs. These findings extend previous observations that suggest that OR genes do not share local significant regulatory homology despite having a common regulatory agenda. We also identified a diverged TCR-α gene segment that uses a divergent recombination signal sequence (RSS) to initiate recombination in T-cells from within the OR region. We explored the hypothesis that OR genes may use DNA recombination in expressing neurons, e.g., to recombine ORs into a transcriptionally active locus. We searched the mouse sequence for OR-flanking RSS motifs, but did not find evidence to suggest that these OR genes use TCR-like recombination target sequences.

Chemosensory systems are among the oldest forms of communication between organisms and their environment. Throughout evolution, chemosensory receptor repertoires have undergone extensive diversification. Expansion and contraction of olfactory receptor (OR) gene families, recombination, gene conversion, translocation, and positive selection for functional change (Ben-Arie et al. 1993; Ngai et al. 1993; Glusman et al. 1996; Trask et al. 1998a; Sharon et al. 1999) are all hallmarks of a rapidly evolving olfactory subgenome. This propensity for change in OR repertoires may reflect the biological demands for adaptation to narrow, species-specific niches. The OR gene family is the largest gene family in mammalian genomes, with approximately 1000 genes arrayed in clusters at multiple chromosomal locations (Buck and Axel 1991; Trask et al. 1998b; Mombaerts 1999; Glusman et al. 2001).

In mammalian olfactory systems, the internal representation of the complex odorant world is accomplished largely by virtue of one fundamental organizing principle: Each neuron that binds odorants is dedicated to a single allele of a single receptor gene (Chess et al. 1994). Thus, odor quality is encoded by discrete patterns of neuronal activity that result from the specific subset of ORs stimulated by an odorant or odorant mixture (Vassar et al. 1993, 1994).

The transcriptional mechanisms responsible for ensuring that only a single OR gene is expressed per neuron are unknown. Transgenic experiments have shown that ~3 kb surrounding an OR gene is sufficient to achieve normal expression patterns (Queva and Reed 1998), yet comparative analyses of paralogous genes in three human and two mouse OR clusters have failed to reveal significant conservation in putative regulatory regions (Bulger et al. 2000; Sosinsky et al. 2000; Lane et al. 2001).

The striking similarities between the olfactory and immune systems have provoked speculation that the two systems might use a common regulatory strategy. Both systems achieve recognition of a vast array of ligands by dedicating each ligand-binding cell to a single receptor allele, which is selectively expressed from a large genomic repertoire. In the immune system, selective receptor expression is accomplished by DNA recombination at both the TCR and immunoglobulin loci. This strategy generates receptor diversity and permits adaptability and heritability in antigen-recognizing cells. Programmed DNA editing has emerged recently in evolution as a viable developmental strategy for gene control (e.g., Gierl et al. 1989; Klar 1990; Muller et al. 1991; Haselkorn 1992; Prescott 1992) and is an appealing model for regulation in the olfactory system. Recombination could ensure singular gene transcription in olfactory neurons and long-term commitment of basal cells responsible for regenerating the olfactory neuroepithelium. The apparent lack of extensive pro-
moter homology among OR genes might be explained if OR transcription requires recombination into an active locus (or loci). The observation that recombination-activating genes (RAG), key components of the recombination mechanisms of the immune system, are expressed in the olfactory neurons of two different vertebrate species (Jessen et al. 1999, 2001) lends further credence to this model.

In this paper, we provide a comparative genomic analysis of the mouse and human OR clusters that are found 5' to the TCR genes in both species. We identify orthologous relationships, characterize recent gene block duplication events, and describe paralogous ORs that appear to be subject to strong selection to be maintained as highly similar pairs. We have used 5' RACE-PCR to identify transcription start sites and find that orthologs share extensive noncoding homology largely contained within the transcriptional unit. Our analysis reveals no strong sequence conservation, TATA-boxes, or conserved transcription initiator sites in OR promoter regions. We identify a functional TCR V-gene segment (Vα1), which is significantly diverged from the other Vα segments and uses a recombination signal sequence (RSS) that differs markedly from consensus RSSs. Therefore, we were curious if the divergent RSS of Vα1 might be a relic of an ancestral recombination system, perhaps used by surrounding OR genes. However, no Vα-like RSSs are apparent near the OR genes. Thus, we find no evidence to support the hypothesis that RAG-mediated recombination plays a direct role in OR regulation.

RESULTS AND DISCUSSION

We have identified six OR genes in ∼200 kb of genomic sequence immediately 5' to the mouse TCR-α/β locus and five OR genes in the corresponding human region (Fig. 1). No further OR homology is found in the ∼65 kb of available sequence beyond hOR1. This ∼65-kb region contains three non-OR genes: the Hsa12 zinc finger gene, a gene encoding a methyl transferase, and the 3' end of a gene (KIAA0737) whose function has not been characterized. CpG islands are associated with the upstream regions of Hsa12 and the methyltransferase gene. Available sequence in mouse extends 15.5 kb beyond mOR1, and no non-OR genes are detected in this region. Thus, it is possible that the mouse OR cluster extends further in this direction.

The first five OR genes in the mouse locus are orthologs to the five OR genes at the human locus. A molecular tree (Fig. 2) illustrates that the relative position and orientation of orthologs have been maintained within their respective clusters. Pairwise identities between orthologs range between 83% and 88%, consistent with levels of conservation observed between orthologs at the mouse and human P2- and β-globin-associated OR loci (Bulger et al. 2000; Lane et al. 2001). Overall, pairwise paralogous nucleotide identities range from 55% to 98%, indicative of both ancient paralogous relationships and very recent duplications within the clusters. The mOR6 gene, for example, is the result of a mouse-specific duplication of a ∼10-kb block containing mOR5. The mOR5

![Figure 1](image-url) Gene map and RSS profiling of the olfactory receptor regions 5' to the mouse and human TCR-α/β loci. (A) The ∼200-kb regions upstream of the TCR-α/β loci are shown. Olfactory receptor (OR) genes are shown as black flags, and putative orthologs are indicated by thick gray lines that connect the mouse and human maps. The five human OR genes were named OR1G3, OR1G1P, OR1G2, OR4E2, and OR4E1P by Glusman et al. (2001). The mOR4, mOR5, and mOR6 genes are identical to previously identified mOR83, mOR10, and mOR28 OR genes, respectively (Tsuboi et al. 1999). V-gene segments are shown as striped flags (Vα1), which position of a conserved 2-kb region of putative regulatory orthology is shown as open rectangles within the ∼80-kb region identified previously by Serizawa et al. (2000). Processed pseudogenes are shown as gray flags (R: retinoblastoma-binding protein-like; A: Arp3/actin2-like; U: ubiquitin-like; T: TBX2-like; P: proteosome component C8-like; S: signal transferase; ZNF: Hsa12 zinc finger genes) are indicated by black arrows, and individual exons are indicated by vertical lines (width according to exon size) beneath these arrows. In all cases, flags point in predicted transcriptional directions, and “X”s on the stems indicate pseudogenes. Block duplications are shaded and boxed. (B) RSS profiling for both strands (FOR, REV) of the mouse OR region. For each direction, the outer rectangles (F1, R1): A Hamming distance was computed between every known functional Vα and Vβ segment RSS for every position in the region. Positions (along with scores and the name of the most similar Vα-gene segment RSS outside the rectangle) below cutoff threshold 1.1 are indicated in the figure (vertical lines within the rectangles for both strands). Stronger RSS signals below cutoff threshold 1.0 are bolded. The inner rectangles (F2, R2): the positions of conserved heptamer motifs (CACAGTG). Open boxes between forward and reverse plots indicate positions of OR coding sequences. Closed rectangles upstream of the mOR2, mOR4, mOR5, and mOR6 open boxes indicate the positions of 5' UTRs (introns and exons).
logs is anomalously high given the age of the duplication that occurred before mouse and human diverged, we postulate that branch length (50 nt changes) is shown.

orthologs are indicated within orthologous clades. Scale bar for pseudogenes are shown in light gray. Percent nucleotide identities for mOR5, and mOR6 are shown in parentheses (25). The two human genes at the TCR-γ segment (V1.1) does not have an ortholog in mouse. The functional form of this gene could play a role in OR targeting in neurons and be a potentially important cofactor in the effort to express OR genes in heterologous cell types.

All six mouse OR genes have complete open-reading frames and are, therefore, presumably functional. So far, we have identified cDNAs for four of the mouse genes. The 5’ RACE-PCR products for these four mouse OR transcripts indicate that each has at least one upstream intron (Fig. 4). Transcription start sites range from 4–7 kb upstream of the coding sequence. In no case do we find introns that span exons of other genes.

We have examined noncoding sequences in the OR clusters for conserved motifs that might be involved in the regulation of these genes. PipMaker analyses (Schwartz et al. 2000) show that, with the exception of recent duplications, noncoding sequence has been conserved only between orthologs, and this homology typically extends only a few hundred base pairs upstream of transcription start sites (TSSs). We find strong non-TATA promoter signals upstream of some but not all OR TSSs (Fig. 4). Regions upstream of the TSSs lack homology with other OR clusters and other gene families represented in GenBank. Because OR transgenes with as little as 3 kb of upstream genomic sequence transcribe in the appropriate cell types and within the native zones of the olfactory epithelium (Qasba and Reed 1998), it is likely that cis motifs play a role in OR transcriptional regulation. Our results suggest that putative cis regulatory sequences may be small and/or scattered, thus requiring more refined techniques to identify.

The expression of the mOR6 transgene is dependent on...
sequences that reside well within the TCR locus, 45–125 kb upstream of the mOR6 coding sequence (Serizawa et al. 2000, in which the mOR6 gene was named mOR28). The ∼80-kb region required for mOR6 expression contains three Vα gene segments of the TCR cluster, a small region of similarity to vacuolar proton ATPase, and a 250-bp region of homology.

Figure 3 Dotplot analysis of the mOR2-mOR3 and hOR2-hOR3 region. A concatenated sequence file containing 71.3-kb mouse genomic sequence surrounding the mOR2-mOR3 region and 94.4-kb human genomic sequence surrounding the hOR2-hOR3 region was plotted against itself. Above the diagonal is the plot of unmasked sequence; below the diagonal is a plot of RepeatMasker sequence. Breaks in the diagonal line indicate positions of masked repeats. Along the vertical axis, repeat content is summarized by color-coded bars: low complexity (light gray), simple repeats (dark gray), LTRs (brown), SINEs (yellow), LINEs (green), and the L1MA4 noted in the text (red). Along the top axis, black arrows indicate the positions of the mOR2, mOR3, hOR2, and hOR3 coding regions; gray arrows indicate the positions of Vα1 gene segments; red rectangles indicate the position of the L1MA4 repeat inserted within the hOR2 block. Within the plot, red lines indicate the regions of the L1MA4 repeat. The duplication unit extends beyond the L1MA4 repeat, although this is not obvious in the dotplot because the sequences preceding the insertion are short and disrupted by numerous Alu repeat insertions. Black boxes in the plot surround homologous portions of the OR gene blocks. The Vα1 gene homology is noted with light gray shaded boxes in the plot. Note that the mOR2 block has more extensive homology with the hOR3 block than the hOR2 block (upper right/lower left quadrants). Also note the extensive hOR2 and hOR3 paralogous block homology (lower right quadrant) as compared to the mOR2 and mOR3 paralogous blocks, which are more significantly diverged (upper left quadrant).
with mouse type IIB intracisternal A-particle (IAP). Within this 80-kb putative regulatory region, we find a 2-kb region 68 kb upstream of the mOR6 gene that is homologous to a region 33 kb upstream of the hOR5 gene at the human locus (Fig. 1). Within this 2-kb noncoding region are four patches of especially high-sequence homology between mouse and human: an 84-bp sequence with 82% identity, a 38-bp sequence with 89% identity, a 20-bp sequence with 100% identity, and a 28-bp sequence with 93% identity. This cross-species homology may be the consequence of selective pressure. Therefore, these specific sequences are candidate regulatory motifs that could account for the mOR6 transgene result. If this region is also required for the transcription of the other OR genes at this locus, it could be a locus-control region (LCR) or an insulator to partition the TCR and olfactory regulatory domains. This orthology resides at the boundary between the olfactory and TCR clusters, an appropriate position for a genomic insulator.

One model able to account for singular expression of OR genes and consistent with apparent lack of paralogous homology and strong promoters invokes recombination of OR sequences into an active OR locus in the genome. This model predicts that OR genes share signal sequences near the transcriptional unit that would direct recombination into an active locus. Because OR transgenes can be expressed from constructs that lack 3’ noncoding sequences (Qasba and Reed 1998), RSSs in regions upstream of the 5’ UTR would, therefore, be sufficient to direct these putative recombination events. We explored this hypothesis by screening OR regions of the mouse TCR locus for RSS-like motifs using a profile derived from multiple alignments of the known functional V-gene segment RSSs. We identified orphan RSSs (RSSs not associated with V-α segments) in the region, but no pattern of RSSs common to multiple OR genes emerges (Fig. 1). For example, we do not identify RSS motifs immediately 5’ to transcription start sites, which would be expected if these regions were recombined adjacent to an active promoter.

Interestingly, there are few RSS-like sequences other than the functional downstream RSS in the ∼40-kb region surrounding the Vα1 gene, a functional recombination target (cDNA GenBank accession codes: AF012171, X55824, D12895, Z49903, U51446), and the only known functional non-OR gene so far identified within an OR cluster. This apparent RSS void around Vα1 suggests that orphan RSSs are tolerated only if they are not a distraction to functional RSSs.

During these analyses, we discovered that the Vα1 gene segment has a lower-scoring RSS than orphan RSSs in the region. The Vα1 RSS is significantly diverged from the RSS consensus identified for the other functional Vα gene segments (Fig. 5). In addition, the Vα1 gene-coding sequence is significantly diverged from other V-α gene segments (Fig. 6). With the thought that the Vα1 RSS may be more representative of sequence motifs that might be involved with recombination within the olfactory region, we performed two additional searches aimed at identifying more divergent RSSs surrounding OR genes.

First, we searched using the CACAGTG heptamer motif conserved in every known functional
cause of overlapping regulatory features. Indeed, the diverged

Figure 6 A molecular tree of the V-gene segments. The phylo-
genic isolation of the mouse and human Vα1 gene segments (bold) is
illustrated by a molecular tree of vertebrate V-gene segments.

RSS, including the Vα1 RSS. Second, we computed Hamming
distances (for a definition, see http://www.ncbi.nlm.nih.gov/
projects/t1glossary2000/_hamming_distance.html) between
every known RSS and every subsequence in the cluster. We
recorded significant similarity to the Vα1-like RSS or any
other functional RSS variant regardless of similarity to an
average RSS. Although we identified several candidate RSS mo-
tifs by this analysis (Fig. 1), we found none with a highest ho-
mology with Vα1. This result argues against the hypothesis that
the divergent Vα1-like RSS resembles putative olfactory-specific
signals. We also find no RSS motifs at a relative position com-
mon to more than one OR gene. These results argue against the
hypothesis that RAG-mediated recombination involving TCR-
like RSSs occurs to achieve selective expression of OR genes.

However, a DNA recombination model in the olfactory
system cannot be excluded. There are at least 45 human trans-
posase-like genes that, like RAG1 and RAG2, are derived from
transposons (Smit 1999; Lander et al. 2001). Each presumably
has its own target sequence and potential function. In addition,
our computational analyses were confined to relatively
simple comparisons of primary sequences. Subtle recombi-
nation signals, perhaps related to three-dimensional structure or
accompanying cofactor binding sites, might be missed by our
analyses. A definitive test of this model awaits examination of
the genomic context of an expressed OR gene in a homoge-
neous population of neurons.

The analyses presented here add to the paradox of OR
gene regulation. Although functional studies suggest the ex-
istence of many common levels of transcriptional control,
which together achieve the expression of a single allele of a
single gene in each neuron and zone-specific expression
within the confines of the olfactory epithelium, available ge-
nomic sequences have provided few clues to this regulatory
puzzle. The fact that the TCR and OR gene families have
a similar transcriptional agenda (e.g., allelic exclusion and
restricted expression of only one of a number of similar clus-
tered genes) and are colocized in the genome could be
because of overlapping regulatory features. Indeed, the diverged
Vα1 TCR gene segment is expressed from within an OR ge-
nomic region, and mOR6 transcription is dependent on se-
quences within the TCR genomic region. However, we find no
additional evidence to support the hypothesis that these two
gene families are interdependent or use common regulatory
mechanisms (e.g., recombination) that might account for
their overlapping genomic relationships.

METHODS

Sequence Data
The sequences considered in this paper were generated previ-
ously in our laboratory (Boysen et al. 1997; Glusman et al.
2001) and are available in the GenBank database (accession
codes: mouse TCR α/β locus NT_002581; human TCR α/β
U85199, U85198, U85197, U85196, and U85195). Before the
availability of the genome sequence of the mouse TCR α locus
and the subsequent revision of the nomenclature, mouse Vα1
was known as Vα19. Olfactory receptor gene families have been
named in accordance with genomic position (5’ OR1, OR2 . . .3’)
for convenience, using the prefix “m” for mouse and “h” for human.

The five human OR genes were named OR10G3, OR10G1P, OR10G2, OR4E2, and OR4E1P by

Glusman et al. (2001). The mOR4, mOR5, and mOR6 mouse
OR genes were named mOR83, mOR10, and mOR28 by

Tsuboi et al. (1999).

Isolation of 5′ OR Exons by RACE
The olfactory epithelium from seven B6CBAF1/J adult mice
was dissected, and 1.3 µg of poly(A)+ mRNA was isolated us-
ing oligo(dT) cellulose (Stratagene). Preparation of cDNA and
RACE protocols were essentially as described in the Marathon
cDNA Amplification and Advantage cDNA PCR kits (Clon-
tech), using antisense PCR primers within the coding region
of the mouse OR genes.

Genomic Analysis Tools
Repeat content was determined by RepeatMasker (Smit and
Green, version of June 6, 2000; A.F. Smit and P. Green, un-
publ.) with RepBase 5.03 as a reference repeat library. Map-
ing of noncoding sequence homology was aided by

PipMaker (Schwartz et al. 2000). The following genomic
analysis tools available at the Baylor College of Medicine
Search Launcher (http://www.hgsc.bcm.tmc.edu) were used:

Genie (Kulp et al. 1996; Reese et al. 1997), TSSG
(Solyvoy et al., unpublished), TSSM (Solyvoy et al., in prep.),
NNPF (Reese and Eckman 1995; Reese et al. 1996), and

MatInspector/ TRANSFAC (Quandt et al. 1995).

RSS Analysis
An RSS profile was generated from a multiple alignment of the
RSSs of all known functional V-α genes, in which predictions
of functionality were based on the presence of the V-gene
segment in expressed mRNAs. The inclusion or exclusion of
RSSs of V-gene segments not definitely known to have func-
tion did not significantly impact our results. A profile is a
tabulation of the frequency of each residue at each position in
an alignment. The V-α profile was used to screen the OR re-

loci (Reese and

Eckman 1995; Reese et al. 1996), and

MatInspector/ TRANSFAC (Quandt et al. 1995).

RSS Analysis
An RSS profile was generated from a multiple alignment of the
RSSs of all known functional V-α genes, in which predictions
of functionality were based on the presence of the V-gene
segment in expressed mRNAs. The inclusion or exclusion of
RSSs of V-gene segments not definitely known to have func-
tion did not significantly impact our results. A profile is a
tabulation of the frequency of each residue at each position in
an alignment. The V-α profile was used to screen the OR re-

loci (Reese and

Eckman 1995; Reese et al. 1996), and

MatInspector/ TRANSFAC (Quandt et al. 1995).
with score <1.1 and none <1.0. If a position in the OR region produces a low Hamming score, this indicates that this position is the start of a sequence that is very similar to one of the known functional RSSs, which may or may not be highly similar to the consensus RSS sequence. Sequences were also screened for the conserved sequence 5'-CACGTG heptamer motif found in many RSSs. All analyses were performed in both the forward and reverse directions.

ACKNOWLEDGMENTS

We thank the Richard Axel laboratory at Columbia University, in particular Tyler Cuthforth, for providing mouse OR 5’ RACE data. The authors also thank numerous individuals in the Department of Molecular Biotechnology Core Sequencing facility for their ongoing efforts in this project. This work was supported by the NIH grants R01 DCO4209 and R01 HG01475.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES


