ing a short acoustic cue from the submarine's final position. In these trials the seal directly approached the submarine's position in an unusually fast reaction (see Fig. 2B, last frame).

What might be the detection range of a trail-following seal for prey fish? Even after $>$3 min, the wake behind a swimming goldfish contains water velocities that are significantly higher than background noise (*9*) and exceed the sensitivity threshold of the whiskers of harbor seals (*4*). Given that a herring swimming at a sustained speed of $\sim$1 m s$^{-1}$ (*16*) leaves a hydrodynamic trail just as stable as that of a goldfish, it might be detectable for a seal even when the herring is more than 180 m away. However, for a reliable estimate of the maximum detection range, we need to learn more about background noise in the wild as well as the aging of fish-generated trails under natural conditions.

Because a swimming seal itself produces considerable water movements that certainly affect the whiskers, the detection of fish-generated water movements was thought to be hardly possible (*17*). However, preliminary results from our laboratory suggest that seals may have overcome this problem by a simple mechanism. As a function of swim speed and their biomechanical properties, the whiskers of a swimming seal vibrate with characteristic frequencies. A hydrodynamic trail intersected by the seal will cause a modulation of this characteristic vibration that might be sensed by the seal.

Our results describe a system for spatial orientation in the aquatic environment that can explain successful feeding of pinnipeds in dark and turbid waters. The sensory ability of hydrodynamic trail-following may be also important to other species equipped with hydrodynamic receptor systems.

### References and Notes

1. W. W. L. Au, *The Sonar of Dolphins* (Springer, New York, 1993).
2. H. Bleckmann, *Reception of Hydrodynamic Stimuli in Aquatic and Semiaquatic Animals* (Fischer-Verlag, Stuttgart, Jena, New York, 1994).
3. S. Coombs, P. Görner, H. Münz, *The Mechanosensory Lateral Line. Neurobiology and Evolution* (Springer, New York, 1989).
4. G. Dehnhardt, B. Mauck, H. Bleckmann, *Nature* **394**, 235 (1998).
5. M. J. Weissburg, M. H. Doall, J. Yen, *Philos. Trans. R. Soc. London B* **353**, 701 (1998).
6. R. W. Davis *et al.*, *Science* **283**, 993 (1999).
7. H. Bleckmann, T. Breithaupt, R. Blickhahn, J. Tautz, *J. Comp. Physiol. A* **168**, 749 (1991).
8. R. Blickhan, C. Krick, D. Zehren, W. Nachtigall, *Naturwissenschaften* **79**, 220 (1992).
9. W. Hanke, C. Brücker, H. Bleckmann, *J. Exp. Biol.* **203**, 1193 (2000).
10. For the generation of linear hydrodynamic trails, the independently moving submarine was powered by a single propeller only. Two lateral steering propellers allowed the generation of curved trails. The submarine was always started with the steering propellers in the vertical plane, but rotated once around its longitudinal axis while running. Depending on the submarine's list, four inclination contacts switched on the steering propellers when these were in the almost horizontal plane after some meters of straight run. The new course was not predictable, but if the submarine rotated fast, it changed its course sharply (Fig. 2A), if the rotation was slowly the resulting course was a left-hand or right-hand curve (Fig. 2B). The speed of the submarine was $\sim$2 m s$^{-1}$ and $\sim$1.5 m s$^{-1}$, for linear and curved trails, respectively. Experiments were recorded by a camcorder installed $\sim$5 m above the pool. Video recordings were analyzed off-line frame by frame. The straight part of the hydrodynamic trail was described by measuring the direction and velocity of the particle movements (PIV) (*9*). A thin horizontal layer of laser light ($\lambda = 650$ nm) was laid about 30 cm below the water surface. A CCD-camera was installed 1 m from the edge of the pool, and neutrally buoyant seeding particles (Vetosint 1101, Hüls AG, Germany) were put into the water. The submarine was started in the depth of the laser plane at a distance of about 3 m. Several runs were performed with the submarine passing the camera in increasing lateral distances. Before each measurement, the direct current (DC) components of the water velocities were below 10 mm s$^{-1}$; velocities of about 5 mm s$^{-1}$ were typical. Particle movements were analyzed manually (Scion Image; Scion, Frederick, MD) or with a cross-correlation technique (MatLab; MathWorks, Natick, MA) in order to describe direction, form and velocity of the water flow.
11. Web fig. 1 is available on *Science* Online at www.sciencemag.org/cgi/content/full/293/5527/102/DC1.
12. R. J. Adrian, *Annu. Rev. Fluid Mech.* **23**, 261 (1991).
13. Web fig. 2 is available on *Science* Online (*11*).
14. U. K. Müller, B. L. E. Van Den Heuvel, E. J. Stamhuis, J. J. Videler, *J. Exp. Biol.* **200**, 2893 (1997).
15. W. J. Bell, *Searching Behaviour* (Chapman & Hall, London, 1991).
16. J. J. Videler, Ed., *Fish Swimming* (Chapman & Hall, London, 1993), pp. 1–22.
17. D. H. Levenson, R. J. Schusterman, *Mar. Mamm. Sci.* **15**, 1303 (1999).
18. The experimental animals were treated in accord with the official German regulations for research on animals. We thank D. Adelung, G. Nogge, S. Prange, and I. Röbbecke for their support during this study. Financed by grants of the Deutsche Forschungsgemeinschaft (G.D.).

8 March 2001; accepted 24 May 2001

# Human Chromosome 19 and Related Regions in Mouse: Conservative and Lineage-Specific Evolution

Paramvir Dehal,[1,2,4] Paul Predki,[1,5] Anne S. Olsen,[1,2] Art Kobayashi,[1,2] Peg Folta,[1,2] Susan Lucas,[1,2] Miriam Land,[1,8] Astrid Terry,[1,2] Carol L. Ecale Zhou,[1,2] Sam Rash,[1,2] Qing Zhang,[1,5] Laurie Gordon,[1,2] Joomyeong Kim,[1,2] Christopher Elkin,[1,3] Martin J. Pollard,[1,6] Paul Richardson,[1,5] Dan Rokhsar,[1,7] Ed Uberbacher,[1,8] Trevor Hawkins,[1,5] Elbert Branscomb,[1,2] Lisa Stubbs[1,2]*

**To illuminate the function and evolutionary history of both genomes, we sequenced mouse DNA related to human chromosome 19. Comparative sequence alignments yielded confirmatory evidence for hypothetical genes and identified exons, regulatory elements, and candidate genes that were missed by other predictive methods. Chromosome-wide comparisons revealed a difference between single-copy HSA19 genes, which are overwhelmingly conserved in mouse, and genes residing in tandem familial clusters, which differ extensively in number, coding capacity, and organization between the two species. Finally, we sequenced breakpoints of all 15 evolutionary rearrangements, providing a view of the forces that drive chromosome evolution in mammals.**

Spanning 65 to 70 Mb and estimated to contain 1100 genes, human chromosome 19 (HSA19) is one of the smallest and most gene-dense of human chromosomes (*1*, *2*). A clone-based physical map spanning all but the centromeric regions of the chromosome with seven gaps (*3*) has provided the framework for HSA19 sequence, which to date includes 35 Mb of finished sequence and 22 Mb of high-quality draft. The solidly anchored clone framework and high percentage of ordered and oriented contigs generated through application of a plasmid paired-end sequencing strategy (*4*) have rendered unfinished portions of HSA19 draft sequence particularly amenable to annotation and analysis. Comparing human DNA sequence with that of other species has proved to be an especial-

[1]DOE Joint Genome Institute, Walnut Creek, CA 94598, USA. [2]Genomics and [3]Engineering Divisions, Lawrence Livermore National Laboratory, Livermore, CA 94550, USA. [4]Department of Genetics, University of California Davis, Davis, CA 95616, USA. [5]Genomics and [6]Engineering Divisions, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. [7]Physics Department, University of California Berkeley, Berkeley, CA 94720, USA. [8]Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA.

*To whom correspondence should be addressed. E-mail: stubbs5@llnl.gov

ly valuable annotation strategy, identifying sequence elements with important biological functions efficiently from a background of nonconserved DNA (*5*). To provide a tool for HSA19 annotation and for evolutionary studies, we sequenced overlapping bacterial artificial chromosome (BAC) clones spanning all 15 segments of HSA19-mouse homology (*3*, *4*) (Fig. 1). We located reference gene sets (*6*) in the assembled HSA19 sequence (*7*) and syntenically homologous mouse BACs and generated comprehensive sets of GRAIL and Genscan gene models for both species (*8*). We also identified significant sequence matches to nonredundant database entries, expressed sequence tags (ESTs), and 6-frame translations of sequence from the mouse BACs and predicted proteins from the sequenced genomes of *Drosophila*, nematodes, and yeast along the length of HSA19. Finally, we aligned DNA and 6-frame translations of the human sequence and syntenically related mouse BACs to identify 12611 sequence elements that are conserved at significant levels in syntenically related regions of HSA19 and mouse (hereafter termed conserved sequence elements, or CSEs). We grouped sequence matches identified by these different methods into overlapping sets to identify 34733 distinct HSA19 "sequence feature blocks" (SFBs) (*9*), detailed descriptions and displays of which can be found on the project Web site (http://bahama.jgi-psf.org/pub/ch19/).

Some properties of HSA19 CSEs are summarized and compared with exons of HSA19 RefSeq genes in Table 1. Eighty percent of the exons of HSA19 RefSeq (*6*) database entries are represented in the conserved sequence set, indicating that most HSA19 genes should be represented by at least one sequence match in the collection of 12611 CSEs. However, only 38% of HSA19 CSEs correspond to RefSeq exons; an additional 26% contain significant similarities to Locus Link, unigene, EST, or other sequence database entries providing evidence that those sequence features define portions of functional genes. A total of 4546 HSA19 CSEs (36% of total) show no significant similarity to any sequence in public databases, including predicted proteins from the genomes of *Drosophila*, nematodes, and yeast. More than

**Table 1.** Properties of HSA19 conserved sequence elements and RefSeq gene exons.

| | ≤100 bp | | >100 bp | | All | | |
|---|---|---|---|---|---|---|---|
| | Number | Percent* | Number | Percent* | Number | Percent | Average length (bp) |
| CSEs | | | | | | | |
| Total | 6467 | 51.1% | 6144 | 48.7% | 12611 | 100% | 120.7 |
| Associated with Refseq exons | 1097 | 17.0% | 3698 | 60.2% | 4795 | 38% | 197 |
| Associated with EST | 2169 | 33.5% | 3203 | 52.1% | 5372 | 42.6% | 146.5 |
| Associated any database match† | 2758 | 42.6% | 4393 | 71.5% | 7151 | 56.7% | 144.1 |
| No significant database match | 3219 | 49.8% | 1327 | 21.6% | 4546 | 36% | 86.3 |
| RefSeq exons | | | | | | | |
| Total | 1976 | 32.8% | 4046 | 68.2% | 6022 | 100% | 198.3 |
| Associated with CSE | 1340 | 67.8% | 3456 | 85.4% | 4796 | 79.6% | 207.2 |

*In rows labeled "total," percentages reported in columns 2 and 4 report the fraction of total CSEs or exons in each size class; percentages in all other rows report the fraction of CSEs or exons within each size class that are associated with other sequence features. †Numbers correspond to CSEs with a significant similarity (an expected value of $e^{-10}$) to sequences in the nonredundant database or predicted proteins from the genomes of *Drosophila*, nematodes, and yeast.

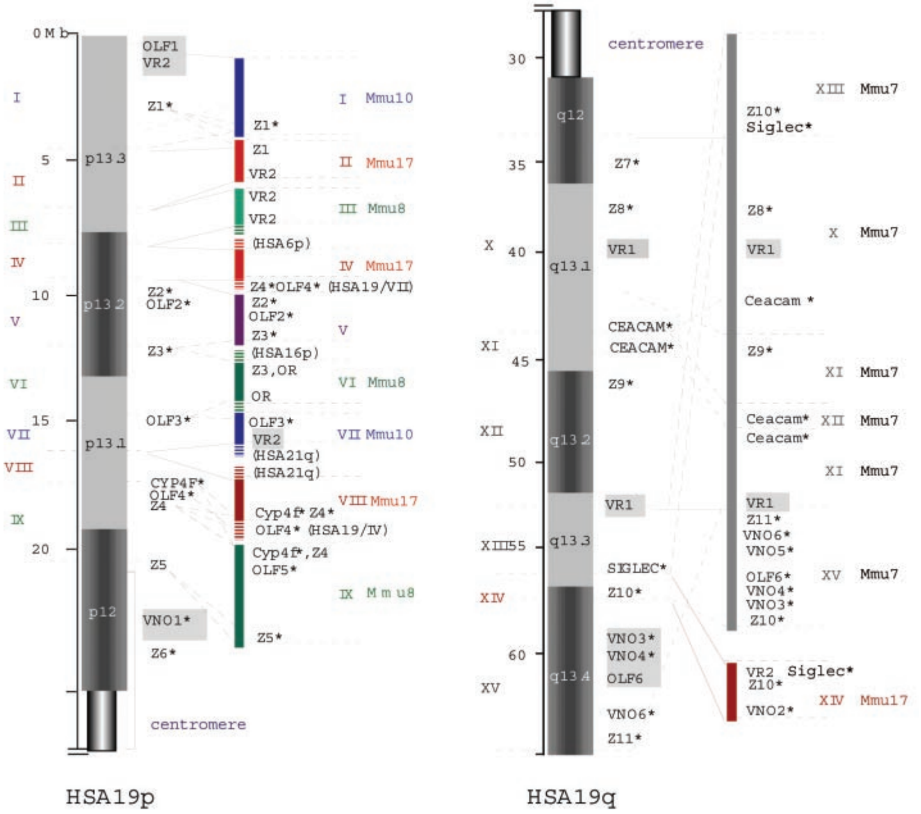

**Fig. 1.** Comparative maps of human chromosome 19. The panel at left illustrates the position of each of the nine segments of HSA19p (labeled I to IX) defining regions of syntenic homology to different intervals of mouse chromosomes 8, 9, 10, and 17; panel at right illustrates relationships between HSA19q and Mmu7 and 17, respectively. In the center of each figure is an ideotype of the respective HSA19 arm with dark and light bands labeled (19p13.3-19q13.4); distance from the telomere of HSA19p, in Mb, represented by numbers on the scale at left. Boundaries of the 15 human-mouse homology segments are shown as dashed horizontal lines; human segments are labeled with roman numerals at far left of each panel. Homology segment labels are colored to correspond to the homologous mouse chromosome, as represented by colored bars to the right of each human chromosome arm (blue, Mmu10; red, Mmu17; green, Mmu8; purple, Mmu9; gray, Mmu7). Mouse homology segments are labeled with roman numerals to the right of each bar, with mouse chromosome number at far right. Hatched ends on the solid bars indicate that mouse BAC sequences crossed the homology segment breakpoint to carry genes with homologs that are not related to HSA19. Symbols of relevant gene families, described in text or in the legend of Table 2, are shown to the right of each ideotype, with an asterisk denoting the presence of a tandemly duplicated cluster of related genes. ZNF, OR, and VR clusters are numbered as summarized in Table 2. Shaded boxes surrounding a gene family symbol indicate the presence of pseudogenes; symbols are not shaded if a single gene in that cluster is functional. Lines between mouse and human maps connect human gene families with corresponding sequences in the related mouse regions; multiple lines drawn between a human locus and the mouse map indicate that homologs of the human gene(s) are split by the evolutionary rearrangement onto two mouse chromosomes. Dashed lines connect human gene family symbols to mouse genes that are duplicated or changed in relative position because of intrachromosomal rearrangements.

86% of the CSEs without significant database matches are clustered within 5 kb of known gene components, and depending on their location and protein coding capacity, these CSEs represent candidates for undiscovered exons (e.g., 5′-ends of partially sequenced genes) or regulatory elements. The probability that a CSE corresponds to a known expressed sequence increases markedly with element length: The average size of a CSE that corresponds to portion of a RefSeq exon is 197 base pairs (bp), whereas the average CSE with no sequence similarity is 86 bp in length (Table 1; Fig. 2). Combining evidence from all types of sequence features, we have predicted about 1200 HSA19 genes, including a hand-curated set of 892 loci defined by matches to RefSeq, locus link, and unigene sequences (*6*) and 128 genes based solely on clustered CSEs that are anchored by nonunigene EST sequence matches. Except for 28 candidate genes defined only by isolated clusters of open reading frame (ORF)–containing CSEs, all predicted HSA19 genes are associated with high probability human or mouse cDNA sequence matches. Locations and properties of these known and predicted HSA19 genes are summarized in Web table A (*10*), and a more detailed version of the gene catalog linked directly to regional displays, DNA sequence, and sequence features that define each locus is available on the project Web site. Our analyses indicate that the number of genes that will be found uniquely through cross-species comparisons will be small; the true and very important benefit of human-mouse sequence alignments will be in the further definition of known genes and conservation-based confirmation of hypothetical genes predicted by other methods.

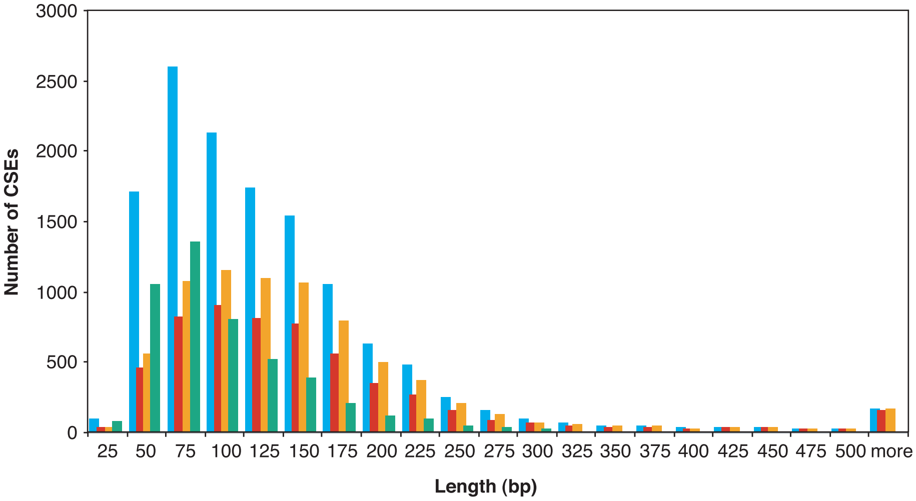

**Fig. 2.** Characteristics of HSA19 CSEs as a function of element length. A histogram showing the distribution of lengths of the 12611 HSA19 CSEs is shown, with numbers along the *x* axis representing the maximum length within each 25-bp bin. The total number of CSEs in each bin is represented by blue bars. The number of CSEs in each bin that are associated with ESTs (red bars), other significant sequence database matches (gold bars, including known genes, nonredundant database entries, or predicted proteins from genomes of *Drosophila*, nematode, and yeast), or not associated with any signficant sequence match (green bars) are also shown.

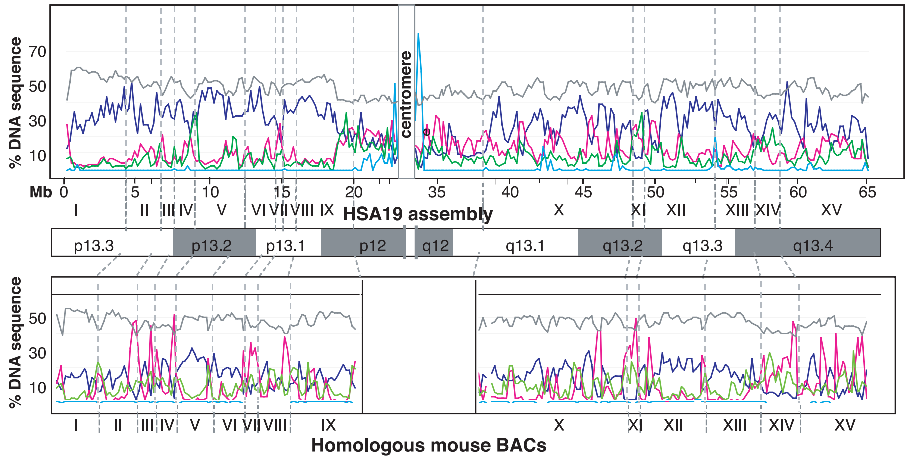

**Fig. 3.** Distribution of repeat sequences in HSA19 and related mouse clones. The repeat content of HSA19 is plotted along the length of the chromosome in the top panel, with SINES (dark blue) LINEs (pink), LTR (green), and satellite repeat sequences (light blue) plotted as a percentage of the total DNA sequence in 200-kb segments along the length of the assembled HSA19 sequence. GC content of the 200-kb segments is also plotted, in gray. The centromere of HSA19 has not been sequenced, and data from that region are not included. Below the human map is an ideogram of HSA19 linking positions along the assembled sequence and the chromosome's cytogenetic banding pattern. Scale below the human repeat plot shows location measured from the HSA19p telomere, in Mb. The bottom panel shows a similar plot of repeat and GC content of mouse BAC clones, arranged so that syntenically related mouse and human sequences are aligned. Dashed vertical lines show positions of homology segment breakpoints; homology segments are labeled below each panel with roman numerals.

For an overview of gene conservation, we focused first on analysis of the 892 established HSA19 genes. Forty-one of these genes (4.6%) are located in HSA19 positions corresponding to gaps in the mouse BAC coverage; clear homologs of all but 3 of the remaining 892 loci were found in syntenically homologous mouse BACs. The three loci not found in mouse BAC sequence—*GOV*, *PPP2R1A*, and DKFZp434d1335—identified highly similar mouse sequences in expressed sequence databases but are also closely related to genes located elsewhere in the human genome; the mouse genes may therefore represent paralogous rather than orthologous loci. In general, we found orthologous gene pairs arranged in syntenically conserved positions in all aligned HSA19 and mouse homology segments. Curiously, however, despite identical gene content, several HSA19 regions are substantially larger than related intervals in mouse (*3*). For example, homology segment I spans 4.1 Mb in HSA19p13.3 but only 2.5 Mb in Mmu10, and genes making up homology segment II occupy 2.7 Mb in HSA19 but only 1.8 Mb in Mmu17. In these intervals and others, which together make up ~50% of HSA19-related mouse DNA, mouse genes are shorter and packed more tightly together because of differences in the number of repetitive elements, particularly short interspersed nuclear elements (SINEs), located both between and within genes. The difference is apparently chromosome-wide: The HSA19 sequence contains a substantially larger number of SINEs (which comprise 27.6% of sequenced HSA19 DNA) than related mouse regions (12.5%) (Web table B; Fig. 3) (*11*). These differences in SINE repeat content and associated changes in interval length represent the major difference between single-copy gene regions of HSA19 and syntenically homologous mouse DNA. Relative to other human chromosomes, HSA19 DNA is particularly rich in SINE repeats (*2*), and it is therefore unclear whether the repeat-driven genomic expansions we have observed for this human chromosome will be seen genome-wide.

Thirty-one percent of HSA19 known and predicted genes (340 genes) are not single copies but are members of large, tandemly clustered gene families, and most of these familial clusters are represented by syntenically homologous clusters in mouse (*3*). We found lineage-specific differences in coding capacity of many different tandem HSA19 gene families, including some that have been described in previous reports (*12*, *13*). Because of their large numbers and potential to influence species-specific aspects of biology, we focused on analysis of genes encoding *Krüppel*-type (C2H2) zinc-finger (ZNF) proteins, which encode putative transcription factors, olfactory receptors (OR), and putative pheromone receptors (vomeronasal receptors, or VR) (*14*). *Krüppel*-type ZNF genes make up one of the largest human families, with at least 700 members genome wide (*1*, *2*). HSA19 carries a disproportionate share of the human ZNF gene repertoire: 262 distinct C2H2 finger-containing segments were identified in the HSA19 sequence (Web table C) (*15*). The genes are clustered in 11 different HSA19 locations, and most clusters contain sets of highly similar sets of genes that appear to have arisen by tandem in situ duplications of ancestral copies (Table 2; Fig. 4). Despite their clear evolutionary relationships, many homologous HSA19 and related mouse clusters contain strikingly different complements of ZNF genes. Evolutionary analysis of a subset of 160 HSA19 and 101 mouse ZNF genes containing a *Krüppel*-associated box (KRAB) motif (*16*) suggests that different founder genes have been duplicated, lost, and selected independently in each conserved cluster since the divergence of primate and rodent lineages (*17*) (Fig. 4). In one extreme example, a single human ZNF gene [ZNF-9906, a gene associated with SFBs 9906 (KRAB) and 9907 (ZNF domain)] is represented by a cluster of 12 closely related ZNF genes in homologous mouse regions (cluster Z4; Fig. 5). Two other mouse ZNF clusters are represented by only one or a few genes in related HSA19 locations (clusters Z1 and Z5), and human-specific cluster expansions were also observed (e.g., cluster Z8; Table 2; Web tables C through F).

All of the 262 HSA19 ZNF loci and all but 10 related mouse sequences contain ORFs capable of encoding proteins with at least five contiguous zinc fingers (Web tables C and D). One hundred ninety-one HSA19 ZNF genes are represented by high probability sequence matches in the EST databases (>97% nucleotide identity in >100 bp), indicating that the genes are actively expressed. ZNF proteins containing KRAB domains are thought to function as transcriptional repressors (*18*), and the idea that different mammalian lineages are actively generating and selecting distinct regulatory protein repertoires from a constantly changing pool is an intriguing one. The observation that most ZNF copies have retained coding capacity suggests that the duplicated genes are not entirely redundant to parental copies in function. Sequence variation within the internally duplicated, microsatellite-like finger regions may provide one path to rapid functional diversity. In

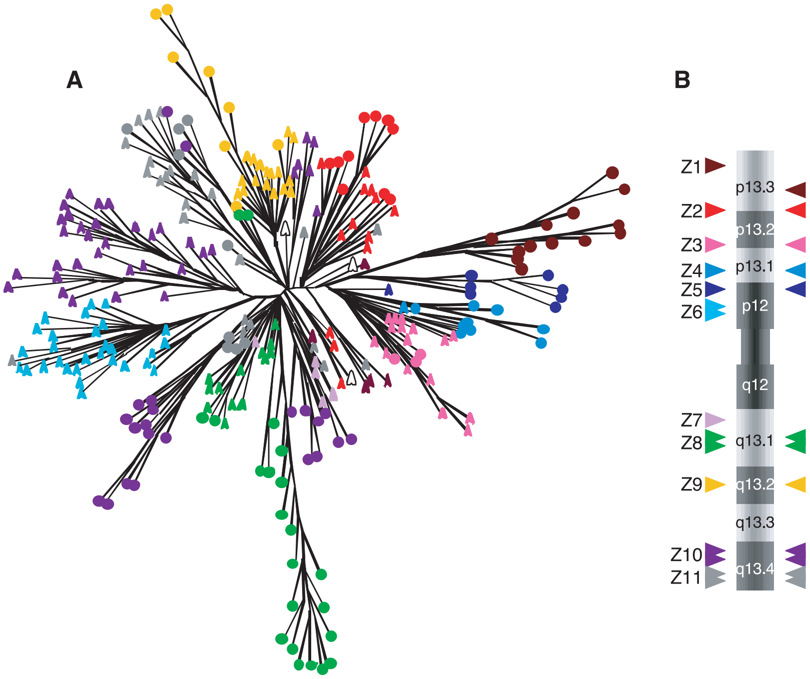

**Fig. 4.** Evolutionary relationships between KRABA-containing ZNF genes in HSA19 and related mouse regions. (**A**) An evolutionary tree showing relationships between KRAB-encoding motif nucleotide sequences of a representative set of 101 mouse and 160 human ZNF genes was estimated with the PHYLIP tree generating program. Only mouse KRAB sequences that were well anchored to specific positions in the comparative map were analyzed. Information regarding the sequences used to generate the tree is summarized in Web tables D and E. Mouse genes are represented by colored circles and human genes are denoted by arrowheads; symbol colors representing each gene's cluster location in HSA19 and related mouse DNA, as indicated by colored arrows on either side of the HSA19 ideotype, in (B). Unfilled black symbols represent singleton ZNF genes. (**B**) Approximate positions of ZNF clusters Z1-Z11 relative to the HSA19 map, as summarized in Table 2. Arrows at left of the chromosome figure show positions of human clusters, color-coded to correspond to symbols representing resident genes on the evolutionary tree, in (A). Double arrows point to large clusters comprising genes distributed over distances of >1 Mb. Arrows at right of the figure show the relative position of mouse clusters in related homology segments; mouse clusters Z6 and Z7 were not sequenced.

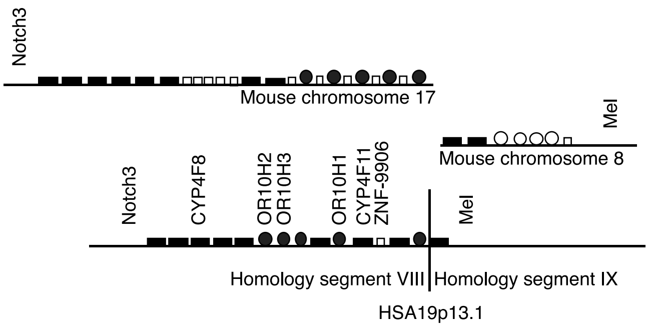

**Fig. 5.** Arrangement of CYP4F, OLFR, and ZNF genes at the border of homology segments VIII and IX. Map at bottom illustrates the arrangement of CYP4F, OR, and ZNF genes and unique genes, *NOTCH3* and *MEL*, that flank the breakpoint of homology segments VIII and IX in HSA19p13.1. Gene family members are represented by symbols as follows: CYP4F genes, filled rectangles; ZNF9906-related ZNF genes, open boxes; OLF4 cluster OR genes, filled circles; OLF5 cluster OR genes, open circles. The names of family members that correspond to known genes are listed above the corresponding symbols. Above the human map are maps representing the arrangement of related genes in homologous regions of Mmu8 and Mmu17, respectively. Multiple relatives of the single human ZNF gene, ZNF-9906, are distributed on either side of the rearrangement breakpoint in mouse, as are multiple members of the CYP4F family. Relatives of genes in HSA19p13.1 cluster, OLF4 (all members of OR sequence subfamily 10) are all found on Mmu17. The closest human relatives of Mmu8 cluster OLF5 are found in human clusters OLF2 and OLF3, located in human homology segments V and VII, respectively.

the case of one HSA19 ZNF cluster, even highly similar duplicates display distinct patterns of tissue-specific expression (*19*, *20*), suggesting another mode by which the newly minted repressor genes might acquire new function. The large number of actively expressed ZNF genes and their diversity in humans and mice suggest that the different mammalian lineages have invested a substantial amount of evolutionary capital in constructing and fine tuning networks designed to regulate the expression of genes.

Human and mouse genomes each contain roughly 900 OR genes (*1*, *2*), encoding proteins that recognize distinct types of olfactants and that function in different regions of the olfactory epithelium (*21*). Acuity of the olfactory sense is reduced in humans when compared with other mammals, and recent loss of function at conserved OR loci has been proposed as a major contributing factor (*22*). Sequence analysis of one pair of homologous human and mouse OR clusters revealed a complex pattern of lineage-specific gene duplication and loss (*23*), but human and mouse OR clusters have not been widely compared. HSA19 contains 49 OR loci distributed in four major clusters (clusters OLF1 to OLF4) (Fig. 1, Table 2, Web table G). Cluster OLF1 is located near the HSA19p telomere and contains only degenerated OR loci, and we did not attempt to isolate related mouse sequences. Mouse and human clusters OLF2, OLF3, and OLF4 contain similar gene sets and are clearly homologous, but the arrangement of mouse OR genes also suggests that the genes have been dispersed in the course of evolution. For example, mouse cluster OLF5 has no obvious counterpart but contains the closest sequenced relatives of human loci found in clusters OLF2 and OLF3 (Web table G). We also found three HSA19 and five mouse OR singletons, related in sequence to clustered HSA19 OR sequences but isolated from those clusters and surrounded by unrelated genes. Two human and four mouse singleton loci have complete ORFs indicating that they encode functional receptors (Web table G). The possibility that these OR singletons are functional is of interest because it provides a counterpoint to the idea that clustered organization is required for OR gene function. Mouse cluster OLF6 [corresponding to the Olfr5 family (*24*)] is represented by a single degenerated OR locus in the homologous human region, suggesting either gene loss in primates or recent cluster expansion in the rodent lineage (*25*). Twenty-seven of the 49 HSA19 OR sequences are short degenerated fragments disrupted by repeat insertions or multiple mutations; only 22 HSA19 human loci contain complete ORF. Overall, a substantially larger fraction of mouse genes has retained capacity to encode functional proteins. Fifteen of the 20 completely sequenced mouse OR genes are capable of encoding a functional protein, and none of the 6 partially sequenced genes contained mutations that would disrupt the ORF (Table 3; Web table G). These observations support that notion that conserved human and mouse loci differ substantially in coding capacity (*22*) but also indicate that the duplication and loss of genes, and indeed whole clusters, have played important roles in determining the different olfactory capabilities of the two species.

A more dramatic example of primate-specific gene loss was observed in two gene families encoding putative pheromone receptors (vomeronasal receptor genes, VR1 and VR2) (*26*, *27*). A number of active rodent VR1 and VR2 genes have been identified, but only a single functional human VR gene, *V1RL1*, has been described (*28*). Loss of function at conserved VR loci has been suggested to parallel a substantial reduction in pheromone-detection activity in humans (*29*), but the evolutionary history of the VR gene families and their prevalence, coding status, and organization in different mammalian genomes are not known. *V1RL1* is located in HSA19q13.4, and we uncovered a cluster of 10 related genes, including at least 6 mouse VR1 genes with complete ORFs, in the syntenically homologous region of Mmu7 (cluster VNO7, Table 2; Web table H). A comprehensive search revealed 26 VR loci in HSA19 DNA; only *V1RL1* is potentially

**Table 2.** ZNF, OLFR, and VR gene families in HSA19 and related mouse regions. NC indicates that mouse BACs related to these human regions were not isolated. Parentheses surrround HSA19 positions that would be predicted given the location of mouse genes that do not have identified HSA19 counterparts.

| Family | Cluster* | Human location (Mb)† | Human genes | | Mouse location | Mouse genes | |
|---|---|---|---|---|---|---|---|
| | | | Loci | ORF§ | | Loci | ORF§ |
| OR | OLF1 | 0.1 | 6 | 0 | NC | NC | NC |
| VR2 | S | 0.3 | 1 | 0 | Mmu10 | 0 | 0 |
| ZNF | Z1 | 2.8‖ (4.1) | 5 | 5 | Mmu10 | 16 | 16 |
| | | | | | Mmu17 | | |
| VR2 | S | (6.7) | 0 | 0 | Mmu17 | 1 | 1 |
| | | | | 0 | Mmu8 | 1 | 1 |
| ZNF | S | 6.7 | 1 | 1 | Mmu8 | 0 | 0 |
| VR2 | S | (7.9) | 0 | 0 | Mmu8 | 1 | 1 |
| OR | OLF2 | 8.5‖ (15.65) | 21 | 8 | Mmu9 | 7‡ | 2 |
| ZNF | Z2 | 9.3 | 13 | 13 | Mmu9 | 10 | 10 |
| | | | | | Mmu9 | 3‡ | 3 |
| ZNF | Z3 | 11 | 23 | 23 | Mmu8 | 1 | 1 |
| OR | S | (11.8) | 0 | 0 | Mmu8 | 1 | 1 |
| OR | OLF3 | 14‖ (15.65) | 14 | 7 | Mmu10 | 2‡ | 0 |
| | | | | | Mmu17 | 1 | 1 |
| OR | S | 14.75 | 1 | 1 | Mmu8 | 1 | 1 |
| OR | S | 14.8 | 1 | 1 | Mmu10 | 1 | 1 |
| VR2 | S | (14.8) | 0 | 0 | Mmu10 | 1 | 0 |
| OR | OLF4 | 15.35 | 5 | 5 | Mmu17 | 4‡ | 4 |
| | | | | | Mmu17 | 11 | 11 |
| ZNF | Z4 | 15.45 | 1 | 1 | Mmu8 | 1 | 1 |
| OR | OLF5 | (15.65) | 0 | 0 | Mmu8 | 5 | 2 |
| ZNF | Z5 | 19.5 | 1 | 1 | Mmu8 | 9 | 9 |
| ZNF | Z6 | 19.5 | 30 | 30 | NC | NC | NC |
| VR1 | VNO1 | 21 | 6 | 0 | NC | NC | NC |
| ZNF | Z7 | 39 | 7 | 7 | NC | NC | NC |
| ZNF | Z8 | 42 | 25 | 25 | Mmu7 | 30 | 27 |
| VR1 | S | 45.5 | 1 | 0 | Mmu7 | 1 | 0 |
| ZNF | Z9 | 49.1 | 22 | 22 | Mmu7 | 10 | 10 |
| VR1 | S | 53.5 | 1 | 0 | Mmu7 | 1 | 0 |
| ZNF | Z10 | 56.5 | 13 | 13 | Mmu7 | 6 | 6 |
| | | | 26 | 26 | Mmu17 | 46 | 42 |
| VR2 | S | (57) | 0 | 0 | Mmu17 | 8 | 3 |
| VR2 | VN02 | (58) | 0 | 0 | Mmu17 | 8 | 3 |
| VR1 | VNO3 | 58 | 3 | 0 | Mmu7 | 3 | 3 |
| VR2 | VNO4 | 60.8 | 3 | 0 | Mmu7 | 3 | 3 |
| OR | OLF6 | 61 | 1 | 0 | Mnu7 | 3 | 3 |
| VR2 | VNO5 | 62 | 0 | 0 | Mmu7 | 7‡ | 6 |
| VR1 | VNO6 | 62.5 | 2 | 1 | Mmu7 | 10 | 6 |
| ZNF | Z11 | 60 | 60 | 60 | Mmu7 | 20‡ | 17 |

*Cluster numbers or S = singleton loci. ZNF, OR, and VR gene sequences are described in detail in Web tables C through H. †Human positions reported indicate the start of the cluster, measured in Mb from the HSA19p telomere in assembled sequence. ‡These mouse clusters were not completely cloned and sequenced, and numbers may not reflect the full gene content. §Criteria for defining ORF lengths for the different gene types are as defined in the text. ‖ The closest relatives of some or all of the genes in these human clusters are located in nonhomologous positions because of evolutionary rearrangements; parentheses surround the HSA19 position, in Mb, that corresponds to the location of rearranged genes in mouse.

functional. We also identified 36 VR loci in HSA19-related mouse DNA, including at least 17 mouse genes with complete ORFs (Web table H). The obvious loss of function in all but 1 of 26 human VR genes is consistent with a precipitous loss of pheromone receptor capacity in primates. By contrast, the similarity between different mouse VR gene copies (*30*) indicates that rodents have generated new VR receptor functions through active rounds of gene duplication.

HSA19 exists as a single, conserved linkage group in most primates (*31*), and linkage within each respective HSA19 chromosome arm is well conserved in dogs, cats, and cattle (*32–34*). Significant conservation of HSA19 gene synteny is also seen in zebrafish (*35*). Therefore, the major differences that distinguish HSA19 from related mouse DNA are chromosome fission events that appear to have occurred specifically in rodents. To gain a chromosome-wide picture of forces that may have driven these rearrangements, we examined mouse and human sequence surrounding the borders of all 15 homology segments (Fig. 1; Table 3). Sequences related to the HSA19p telomere, HSA19q telomere, HSA19q centromere, and several other homology segment ends are fused to other HSA19-related sequences in mouse (Fig. 6) indicating that intrachromosomal rearrangements accompanied or preceded many of the fission events that split the conserved HSA19p linkage group into fragments in rodents.

Examination of mouse and human breakpoint sequences revealed a number of common features. Ten breakpoints lie in the midst of clustered gene families with the break splitting closely related family members onto separate mouse chromosomes or regions (Fig. 1; Table 3). The five ZNF gene families, three OR gene clusters, and the single cluster of CYP4F genes located in HSA19p are, without exception, po-

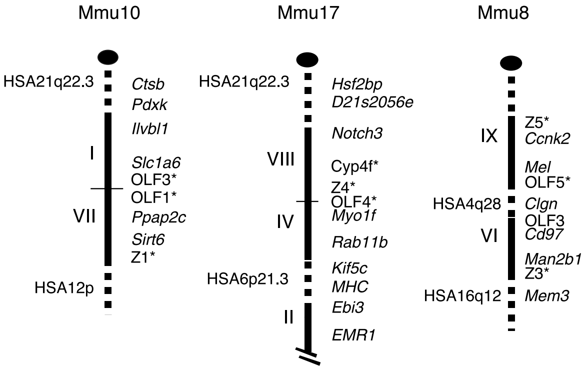

**Fig. 6.** Organization of HSA19p-related segments in mouse chromosomes 8, 10, and 17. The arrangement of different HSA19 homology segments in mouse chromosomes 8, 10, and 17 are shown. Segments I/VII and VIII/IV are contiguous in mouse chromosomes 10 and 17, respectively. Solid bars denote regions of the mouse chromosomes that are related to HSA19, with homology segments identified by Roman numerals at the left of each figure. Mouse DNA related to other human chromosomes is represented by dashed bars; the human chromosome regions to which these adjacent mouse segments are related are identified by numbers at left, where known. Gene names or gene family designations, as explained in the text, are shown at the right of each figure.

**Table 3.** Summary of DNA sequence content at homology segment borders. Abbreviations are as follows: CEACAM, carcinoembryonic antigen/pregnancy-specific glycoprotein family; SIGLEC, sialic acid glycoprotein family; FPR1, formyl peptide receptor 1 family; CYP4F, cytochrome P450 subfamily 4F; RV, endogenous retrovirus sequence; L1, LINE1 repetitive element. -rs marks sequences that are related but not identical to established genes.

| Breakpoint* | Sequence content: human | Flanking genes | Mouse Chr† | Sequence content: mouse | Adjacent homology‡ |
|---|---|---|---|---|---|
| ptel/I | OLF1 L1# | *PPAP2C* | 10 | L1# | H19p (VII) |
| I/II | Simple sequence repeats§ | SIRT6‖ | 10 | Z1‖, RPL37a-rs, L1#, RV# | HSA12q22-q24 |
| | | EB13 | 17 | Z1; RPL23a-rs; L1#, RV# | HSA6q |
| II/III | L1# | EMR1 | 17 | VR2, L1#, RV# | HSA8q13 |
| | | INSR | 8 | VR2, L1#, RV# | HSA13q34 |
| III/IV | RPL23rs, RV# | SCYA25 | 8 | VR2; ZNF; RPL23a-rs, L1#, RV# | HSA13q34 |
| | | RAB11B | 17 | L1#, RV#, Kife2¶ | HSA6p21.3 |
| IV/V | Z2, OLF2, RV# | MYO1E | 17 | Z2, RV, OLF4¶ | HSA19p (VIII) |
| | | FBL12 | 9 | Z2, OLF2, L1¶, RV¶ | HSA11q21 |
| V/VI | | ACP5 | 9 | Z3, RV¶ | HSA11q22 |
| | Z3, RV# | MAN2B1 | 8 | Z3, RV#, L1¶, Mem3¶ | HSA16q |
| VI/VII | | NDUFB7 | 8 | OLF3, L1#, RV#, Clgn¶ | HSA4q |
| | OLF3 | SLC1A6 | 10 | OLF3, L1#, RV# | HSA19p13.3(I) |
| VII/VIII | | ILVBL1 | 10 | Notch3-rs, Pdxk¶, Cstb¶ | HSA21q22.3 |
| | | NOTCH3 | 17 | Notch3, Pdxk-rs¶, Hsf2bp¶ | HSA21q22.3 |
| VIII/IX | CYP4F¶, OLF4, Z4, L1#, RV# | CYP4F2 | 17 | Cyp4f#, Z4, OLF4 | HSA19p13.2(IV) |
| | | MEL | 8 | Cyp4f, OLF5, L1#, RV# | 4q28 |
| IX/pcen | Z5, Z6 L1# | CCNK2 | 8 | Z5, LI#, RV# | 8p21 |
| qcen/X | Alpha satellite repeats | UQCRFS1 | 7 | RV#, RPL21-rs | HSA19q (XIII) |
| X/XI | | LIPE‖ | 7 | Ceacam#, L1# | HSA19q (XII) |
| | CEACAM# | *CEACAM* | 7 | Ceacam# | Hsa19q (XII) |
| XI/XII | PSG#, L1H, RV# | CEACAM | 7 | Ceacam#, L1#, RV#, Ppp5c¶ | HSA19q (XII) |
| | | XRCC1 | 7 | Ceacam, L1#, RV#, Argefh1¶ | HSA19q (X) |
| XII/XIII | ZNF, VR1 | LIG1 | 7 | VR1, L1#, RV# | HSA19q (XV) |
| | | EmP3 | 7 | ZNF, RV# | HSA11p14 |
| XIII/XIV | Z10, SIGLEG#, FPR1#, L1# | ETFB | 7 | Z10, Siglec# | HSA19q (X) |
| | | HAS1 | 17 | Z10, Siglec#, Fpr1¶, L1#, RV# | HSA16p13.3 |
| XIV/XV | Z10, VN03#, L1#, RV# | ZNF160 | 17 | Z10 | HSA16p13.3 |
| | | PKCC | 7 | Z10, VNO3# | Mmu7-cen |
| XV/qtel | Z11, VNO6# | ZNF42 | 7 | Z11, VNO6# | HSA19q (XII) |

*Roman numerals refer to homology segments defined by a given breakpoint; pcen and qcen, HSA19 p and q centromeres, respectively; ptel and qtel, HSA19 p and q telomeres, respectively. †Mouse chromosome. ‡Human chromosomal region to which adjacent mouse sequences are related. For regions that abut other HSA19 segments in mouse, the neighboring mouse homology segment number is shown in parentheses. §Taken from (*44*). ‖Genes located in different positions within human and mouse homology segments because of internal inversions. ¶Known genes identified on a breakpoint-spanning mouse BAC that are derived from adjacent regions of human homology, as indicated in column 6. #Indicates the presence of a cluster of at least two related genes; ZNF (Z), OR (OLF), and VR (VNO) gene cluster numbers shown in column two are defined in Table 2.

sitioned across or immediately adjacent to sites of evolutionary rearrangements. In six cases, the evolutionary break is associated with a significant expansion of the mouse gene family. One example involves the division of mouse CYP4F genes and the 12 mouse relatives of the singleton human ZNF gene, ZNF-9906, described above, onto the separated ends of homology segments VIII and IX in mouse (Mmu17 and Mmu8, respectively) (Fig. 5; Tables 2 and 3). Clear evidence of gene duplication at sites of rearrangement was also observed outside of clustered gene families. For example, in HSA19p13.1, *ILVBL1* and *NOTCH3* are neighboring genes that lie ~17 kb apart, whereas mouse *Ilvbl1* and *Notch3* are separated onto different chromosomes (Mmu10 and Mmu17, respectively). Except for a complex duplication of sequences related to *Notch3* and *Pdxk* (the ortholog of an HSA21 gene), the arrangement of HSA19p13.1 and HSA21q22.3 genes appears to be the product of a simple reciprocal exchange. The similarity between Mmu10 and Mmu17 duplicates indicates that the rearrangement occurred 20 to 25 million years ago, long after the separation of primate and rodent lineages (*36*). The arrangement of nearly perfect repeat structures at the junctions of the two mouse regions suggests a direct association between the duplication and translocation events (Web fig. 1).

Finally, we noted especially high concentrations of tandemly organized L1 repeats and retrovirus-associated LTR sequences at sites of evolutionary breaks (Fig. 3, Table 3). On average, mouse breakpoint clones contained 21.7% L1 sequences, more than a twofold increase over the average L1 repeat content HSA19-related BAC set as a whole (9.2%). Recombination between L1s and other repeated sequences has been documented to drive rearrangements associated with human disease (*37*), and concentration of repeats and duplicated genes at virtually all HSA19 breakpoints indicates that similar mechanisms are also driving chromosome reorganization on the evolutionary scale. The large arrays of ZNF and OR genes, each spanning several hundred kilobases and scattered liberally around the genome, may also present exceptional targets for nonhomologous pairing. The tandemly clustered architecture is therefore likely to be both a reflection of, and a predisposing force for, the rapid evolution of these families.

The differences we observed in homologous mouse and human gene families provide a stark contrast to the solid, virtually one-on-one conservation of unique HSA19 genes. These data show that the mammalian single-copy gene repertoire is relatively static, changing slowly over evolutionary time. However, against this stably conserved background, lineage-specific changes continue at a rapid pace within gene families, especially within families for which tandem in situ duplication provides the major mode of expansion. Such lineage-specific differences are not likely to be limited to HSA19; in fact, differential patterns of gene gain and loss have been documented in several well-studied clustered gene families, with the major histocompatibility complex serving as the classic but not the exclusive example [(*13*, *22*), reviewed in (*38*)]. Extrapolating from HSA19 observations, we can expect to find hundreds of new and lost lineage-specific genes as human and mouse genomes are compared. Unique genes also generate duplicates through retrotransposition and other mechanisms at significant rates, and dispersed gene copies may also be selected to carry out lineage-specific functions [e.g., *Apo(a)* (*38*)]. However, most gene copies do not retain function (*39*), and HSA19 and mouse DNA are both littered with fossils of such failed lineage-specific gene duplication events. The high rate at which duplicated ZNF, OR, and other clustered genes have been maintained as intact copies in one or both species may reflect the operation of different evolutionary forces than those acting on single-copy genes. Our studies show that certain types of genes are driving dramatic lineage specific changes in gene repertoire through in situ duplication events; this mechanism, acting in concert with incremental changes in the protein coding and regulatory elements of the stably conserved genes, is likely an important source of evolutionary novelty.

### References and Notes

1. J. C. Venter *et al.*, *Science* **291**, 1304 (2001).
2. The International Human Genome Sequencing Consortium, *Nature* **409**, 860 (2001).
3. J. Kim *et al.*, *Genomics* **74**, 129 (2001). Human and mouse mapping data including sequencing tiling paths and homology links and GenBank assession numbers for all sequenced mouse and human clones are accessible at http://greengenes.llnl.gov.
4. Mouse BAC DNA was isolated from overnight cultures with a commercially available kit (Qiagen, catalog no. 12165). Purified DNA was fragmented with a GeneMachines Hydroshear, end-repaired, and size fractionated on agarose gels. Fragments in the 3- to 4-kb size range were excised, eluted, and blunt-end ligated to *Sma*I-linearized pUC-18. DNA from overnight cultures (180 μl) of these libraries was purified with an automated 96-well SPRI protocol (*40*). These templates were cycle-sequenced with both M13-fw and M13-rv primers with dye-terminator chemistry (AP Biotech, catalog no. US81095). The resulting product was electrophoresed on MegaBACE 1000 capillary sequencers. Raw traces were preprocessed with Cimarron software version 2.1905 (AP Biotech) and base-called with Phred (*41*). Readlengths in libraries typically average 500 to 550 Phred>=20 bases. Individual sequence reads were assembled with Phrap (http://bozeman.mbt.washington.edu/phrap.docs/phrap.html), and contig order/orientation was established with the graphical user interface Finisher (available through William FitzHugh at will@genome.wi.mit.edu). All human and mouse clones were sequenced to a depth of at least 6× coverage. About 40% of the sequenced mouse BACs were assembled as fully ordered and oriented contigs; 50% were "partially ordered and oriented" or assembled into large contigs, not all of which are fully ordered and oriented. The remaining 10% are presently available only as unordered draft sequence. Of the 22 Mb of HSA19 sequence that is presently in draft form, 47% is completed as fully ordered and oriented contigs, 27% is partially ordered, and 26% remains in unordered draft contigs. Quality data, statistics, and sequence updates for each mouse and human clone are available at www.jgi.doe.gov.
5. E. M. Rubin, A. Tall, *Nature* **407**, 267 (2000).
6. HSA19 sequence was used as a query sequence in a BLAST search against various sequence databases. BLAST reports were then parsed to create two undirected graphs for subject hits and their corresponding location on HSA19. Thresholds such as subject type (e.g., RefSeq, locus link, nonredundant protein) percent identity, e score, and/or length are used to filter the subject graph. The subject graph was used to calculate the "best hit path" for each subject; nodes not along a path, or on a path less than the minimum percent coverage, were eliminated creating a directed subject graph. The connected components of the HSA19 graph were found on the basis of the remaining subject nodes such that a common subject defines an edge. This allows HSA19 locations to be grouped together by multiple subjects. Constraints were placed on the HSA19 graph such that locations from only nearby segments of the chromosome could group together. This approach was used in an iterative fashion such that after a location on HSA19 is grouped by a subject at a high threshold, that location is excluded from further analysis and the subject may now hit another HSA19 location at a lower threshold. This strategy eliminates groups of similar genes from collapsing into a single set of HSA19 locations. To create the comparative gene catalog, we parsed RefSeq, Locus Link, and unigene matches from the collection. A gene entry was considered to be conserved if the corresponding syntenically homologous mouse BACs contained BLAST or tblastx sequence matches below an expected value of e 10, with length corresponding to a minimum of 10% of the human RefSeq alignment. Validity of all human and mouse matches was verified by manual examination of sequence matches. Descriptions of RefSeq and Locus Link collections can be found at www.ncbi.nlm.nih.gov/LocusLink/. The collection used was from 9 February 2001. To add a more speculative gene set to the catalog, we identified significant matches to nonunigene human ESTs (≥99% identity over lengths of >100 bp) or mouse cDNA sequences (≥97% identity to homologous mouse BAC sequence) supported by clusters of mouse sequence matches. We also identied 28 sets of clustered ORF-containing conserved elements without significant cDNA sequence matches. The resulting comparative gene catalog can be found, with displays of sequence features, access to sequence and other information at http://bahama.jgi-psf.org/pub/ch19.
7. We generated a framework for comparative sequence alignments by assembling the combined finished and draft HSA19 DNA sequence using clone mapping information and established sequence order and orientation data. HSA19 assembly was performed in two stages with the Paracel CAP4 assembly program (*42*). The contigs of finished, ordered, and oriented, and draft clones were assembled (without quality values) in a pairwise manner following the tiling path resulting in sets of overlapping clones. Assemblies that violated the tiling path, the preestablished clones order and orientation, or had intraclone assemblies were eliminated. Each set of overlapping contigs was then reassembled. A final check broke any assembly that violated tiling path or order and orientation information. Contigs were then placed into groups of overlapping clones. Assemblies within the 195 resulting contigs were then curated by hand to ensure that clone order, overlap, sequence order and orientation information, and gene sequence contiguity were maintained as much as possible. The resulting assembly resembles recent updates of published assembly data (*2*) but, largely because of our curation efforts, preserves gene sequence contiguity more accurately. Contigs containing large regions of partially ordered draft sequence still contain obvious assembly errors; the assembly and annotation will be updated regularly as additional sequence is completed (additional information provided at http://bahama.jgi-psf.org/pub/ch19).
8. The GrailEXP gene finding program (http://compbio.

ornl.gov/grailexp/) was used to build gene models from a combination of EST/mRNA alignments and computationally predicted exon candidates. The databases used in the gene modeling process include RefSeq, TIGR EGAD, the Baylor Human Transcript database, the Univ. of Penn. Database of Transcribed Sequences (DOTS), and Dbest. Genscan (http://genes.mit.edu/GENSCAN.html) was also run on the sequences (after repeat-masking). Because Genscan does not use integrated ESTs or reference gene message information, the gene models generated with this program were not allowed to cross a sequence gap.

9. The global comparative analysis of the assembled human contigs was carried out with NCBI's blastall version 2.0.14 program (*43*). Contigs were masked for simple sequence and repeats with RepeatMasker (http://ftp.genome.washington.edu/cgi-bin/RepeatMasker/) and then used to search syntenically homologous mouse BAC clones (using blastn and tblastx) and a series of other data sets including NCBI's EST and nonredundant databases (7 October 2000 release, using blastn), and *Drosophila*, *C. elegans*, and yeast amino acid sequence databases (7 October 2000 release; using tblastx). All BLAST matches with e scores less than $10^{-2}$ were collected, parsed, merged and then used to define "sequence feature blocks" on the basis of overlapping hits. Locus Link, unigene, and IMAGEne IDs were retrieved for all relevant matching human accession numbers. All information was entered into a database for statistical queries and the generation of graphical views. A display of these combined data sets is available on http://bahama.jgi-psf.org/pub/ch19.
10. Web figure and tables are available on *Science* Online at www.sciencemag.org/cgi/content/full/293/5527/104/DC1.
11. Repeat and GC content of 200 kb assembled human fragments and mouse BACs was determined with RepeatMasker (http://ftp.genome.washington.edu/cgi-bin/RepeatMasker/).
12. J. L. Gao *et al.*, *Genomics* **51**, 270 (1998).
13. D. R. Nelson, *Arch. Biochem. Biophys*. **369**, 1 (1999).
14. HSA19 and mouse regions were used as query sequence in a BLAST search against themselves. Results were parsed, self-hits were eliminated, and a graph was created such that each HSA19 or mouse region represents a node and any hit between two regions constitutes an edge whose weight is equal to the e score. For each gene family, an appropriate maximum e score was chosen and all connected nodes found. E scores were chosen by incrementally raising the threshold until groupings for that family were obviously incorrect. For KRABA, ZNF, OLFR, and VNO, the e scores chosen were $e^{-60}$, $e^{-75}$, $e^{-20}$, and $e^{-15}$, respectively. Different e scores reflect relative conservation between members of the family. In KRAB-ZNF genes, the KRABA and zinc-finger domains are each encoded by single exons (*13*), so that there is a one-to-one correspondence between sequence feature blocks containing these elements and individual genes. Because each sequence feature block generally contains only one exon or exon fragment, only single exons of multiexon genes are found; therefore, for other ZNF gene types that contain more than one finger domain, the gene count will be an overestimate. Likewise, OR (*18*) and VR1 (*23*) coding sequences are contained within a single exon, simplifying the analysis of these genes. VR2 genes have a more complex structure; to identify these genes and assess coding capacity, we focused on the analysis of sequences in a single exon encoding the conserved membrane spanning domains (*24*). To find exon fragments, such as the numerous pseudogene fragments found for VR1 and VR2 genes in HSA19, we set the e scores to be relatively high; this caused an increase in the number of false positives. Resulting sets of matches corresponding to all families were therefore examined by hand to determine validity of membership and coding capacity. Details of each hit and tables summarizing the locations and properties of each gene set are summarized in Web tables B through G.
15. E. J. Bellefroid *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 3608 (1991).
16. A multiple alignment of the HSA19 and syntenic mouse KRABA nucleotide sequence was created with ClustalX (*44*). Using PHYLIP (*45*), we generated a consensus distance tree using dnadist, neighbor, and consense programs with 1000 rounds of bootstrapping. The results of this procedure are consensus trees and branch lengths therefore do not represent distance; however, the topology of the tree correctly shows relationships between the groups. Details of the KRAB sequences used in the tree are summarized in Web tables D and E.
17. M. Abrink *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 1422 (2001).
18. M. Shannon, L. Stubbs, *Genomics* **49**, 112 (1998).
19. P. Mombaerts, *Science* **289**, 707 (1999).
20. M. Shannon, L. Stubbs, unpublished data.
21. S. Roquier, A. Blancher, D. Giorgi, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 2870 (2000).
22. M. Lapidot *et al.*, *Genomics* **71**, 296 (2001).
23. S. Sullivan *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 884 (1996).
24. C. Dulac, R. Axel, *Cell* **83**, 195 (1995).
25. J. Kim, manuscript in preparation.
26. H. Matsunami, L. Buck, *Cell* **90**, 775 (1997).
27. I. Rodriguez *et al.*, *Nature Genet.* **26**,18 (2000).
28. D. Giorgi *et al.*, *Genome Res.* **10**, 1979 (2000).
29. F. Richard *et al.*, *Genome Res.* **10**, 644 (2000).
30. P. Dehal *et al.*, data not shown.
31. E. Ostrander, F. Galibert, D. Patterson, *Trends Genet.* **16**, 117 (2000).
32. W. J. Murphy *et al.*, *Genome Res.* **10**, 691 (2000).
33. M. Band *et al.*, *Genome Res.* **10**, 1359 (2000).
34. J. H. Postalthwait *et al.*, *Genome Res.* **10**, 1890 (2000).
35. We examined all alignments in the duplicated region that were more than 500 bp in length and were not part of the *Pdxk* or *Notch3* genes or repetitive elements. The duplication date was estimated by multiplying the total percent identity (89.5%) over all aligned segments (10.7 kb) and dividing the percent difference (10.5%) number by the estimated neutral mutation rate for mouse, as determined by Li and colleagues (*46*).
36. Y. Ji *et al.*, *Genome Res.* **10**, 597 (2000).
37. E. Carver, L. Stubbs, *Genome Res.* **7**, 1123 (1997).
38. K. A. Frazer *et al.*, *Nature Genet.* **9**, 424 (1995).
39. M. Lynch, J. S. Conery, *Science* **290**, 1151 (2000).
40. T. L. Hawkins *et al.*, *Nucleic Acids Res.* **22**, 4543 (1994).
41. B. Ewing *et al.*, *Genome Res.* **8**, 175 (1998).
42. X. Huang, *Genomics* **33**, 21 (1996).
43. S. F. Altschul *et al.*, *Nucleic Acids Res.* **25**, 3389 (1997).
44. J. D. Thompson *et al.*, *Nucleic Acids Res.* **24**, 4876 (1997).
45. J. Felsenstein, Phylogeny Inference Package, version 3.5 (University of Washington, Seattle, WA, 1993).
46. R. Puttagunta *et al.*, *Genome Res.* **10**, 1369 (2000).
47. We thank the staff of the JGI sequencing team for an outstanding effort; a complete list of contributors can be found at photo http://bahama.jgi-psf.org/pub/ch19/people.html. We were provided useful technical discussions and critical comments on the manuscript by J. Boore, B. Wold, and M. Frazier. This research was performed under the auspices of the U.S. Department of Energy at Lawrence Livermore National Laboratory under contract W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract DE-AC03-76SF00098, (managed by the University of California), and Oak Ridge National Laboratory under contract DE-AC05-00OR22725 (managed by UT-Battelle, LLC).

28 February 2001; accepted 1 June 2001

# Ventroptin: A BMP-4 Antagonist Expressed in a Double-Gradient Pattern in the Retina

Hiraki Sakuta,[1] Ryoko Suzuki,[1,3] Hiroo Takahashi,[1,3] Akira Kato,[1,3] Takafumi Shintani,[1,3] Shun-ichiro Iemura,[2] Takamasa S. Yamamoto,[2] Naoto Ueno,[2,3] Masaharu Noda[1,3]*

**In the visual system, the establishment of the anteroposterior and dorsoventral axes in the retina and tectum during development is important for topographic retinotectal projection. We identified chick Ventroptin, an antagonist of bone morphogenetic protein 4 (BMP-4), which is mainly expressed in the ventral retina, not only with a ventral high–dorsal low gradient but also with a nasal high–temporal low gradient at later stages. Misexpression of *Ventroptin* altered expression patterns of several topographic genes in the retina and projection of the retinal axons to the tectum along both axes. Thus, the topographic retinotectal projection appears to be specified by the double-gradient molecule Ventroptin along the two axes.**

Axonal connection patterns in the nervous system often form topographic maps, with nearest neighbor relationships of the projection neurons maintained in their connections within the target. The projection from the retina to the tectum is a good model system for understanding the development of topographic maps. Graded distributions of topographic molecules along the anteroposterior (A-P) (nasotemporal) and dorsoventral (D-V) axes in the retina and tectum, which are derived from the regional specialization along the two axes during retinal and tectal development, control the topographical projection of retinal axons (*1*). Although gradients of diffusible factors and transcription factors are known to control the regional specificities in the retina during development (*2*–*6*), the molecular mechanisms involved are mostly unknown. We performed large-scale screening of region-specific molecules (*7*) [also see supple-

[1] Division of Molecular Neurobiology, National Institute for Basic Biology, [2] Division of Morphogenesis, National Institute for Basic Biology, [3] Department of Molecular Biomechanics, The Graduate University for Advanced Studies, 38 Nishigonaka, Myodaiji-cho, Okazaki 444-8585, Japan.

*To whom correspondence should be addressed. E-mail: madon@nibb.ac.jp