LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-JRNL-215844

Evolutionary expansion and divergence in a large family of primate-specific zinc finger transcription factor genes

A. T. Hamilton, S. Huntley, M. Tran-Gyamfi, D. Baggott, L. Gordon, L. Stubbs

October 3, 2005

Genome Research

Evolutionary expansion and divergence in a large family of primate-specific zinc finger transcription factor genes

Aaron T. Hamilton[1], Stuart Huntley[1], Mary Tran-Gyamfi[1], Dan Baggott[1], Laurie Gordon[1], and Lisa Stubbs[1,2]

[1]Genome Biology Division, Lawrence Livermore National Laboratory, P.O. Box 808, L-441, Livermore, CA 94550

[2]To whom correspondence should be addressed

Telephone: (925) 422-8473
FAX: (925) 422-8473
Email: stubbs5@llnl.gov
Running title: Primate-specific zinc finger genes

[ABSTRACT]

Although most genes are conserved as one-to-one orthologs in different mammalian orders, certain gene families have evolved to comprise different numbers and types of protein-coding genes through independent series of gene duplications, divergence and gene loss in each evolutionary lineage. One such family encodes KRAB-zinc finger (KRAB-ZNF) genes, which are likely to function as transcriptional repressors. One KRAB-ZNF subfamily, the ZNF91 clade, has expanded specifically in primates to comprise more than 110 loci in the human genome, yielding large gene clusters in human chromosomes 19 and 7 and smaller clusters or isolated copies at other chromosomal locations. Although phylogenetic analysis indicates that many of these genes arose before the split between old world monkeys and new world monkeys, the ZNF91 subfamily has continued to expand and diversify throughout the evolution of apes and humans. The paralogous loci are distinguished by sequence divergence within their zinc finger arrays indicating a selection for proteins with different DNA binding specificities. RT-PCR and in situ hybridization data show that some of these ZNF genes can have tissue-specific expression patterns, however many KRAB-ZNFs that are near-ubiquitous could also be playing very specific roles in halting target pathways in all tissues except for a few, where the target is released by the absence of its repressor. The number of variant KRAB-ZNF proteins is increased not only because of the large number of loci, but also because many loci can produce multiple splice variants, which because of the modular structure of these genes may have separate and perhaps even conflicting regulatory roles. The lineage-specific duplication and rapid divergence of this family of transcription factor genes suggests a role in determining species-specific biological differences and the evolution of novel primate traits.

INTRODUCTION

Although most vertebrate genes are conserved as one-to-one orthologs in different species, not all genes have followed this conservative evolutionary path. In particular, certain gene families are represented by distinct numbers and types of genes in different vertebrate lineages due to ongoing series of gene duplications, divergence, and gene loss. Two of the best-known examples include genes encoding olfactory receptors and immune receptors, which vary not only between species but also differ dramatically in copy number and protein coding sequence in individual humans. The lineage-specific and individual differences in these families have had important functional consequences, yielding significant inter- and intraspecies variation in acuity and sensitivity in the sense of smell or the response to tumors/infected cells, respectively (Young et al., 2002; Gilad et al., 2005; Trowsdale et al., 2001; Sambrook et al., 2005; Nei et al., 1997; Li and Nei, 2005). Many other gene families have expanded independently in different vertebrate lineages to generate functionally distinct genes, include those encoding bitter taste receptors (Fischer et al., 2005), pheromone receptors (Lane et al., 2004), nuclear receptors (Bertrand et al., 2004), cytochrome P450 proteins (Nelson, 2003), protocadherins (Noonan et al., 2004), and tyrosine kinases (Shiu and Li, 2004) to name just a few. Each gene family has a unique evolutionary history defined by different modes of duplication, selection pressures, and the evolutionary timing, extent, and lineage-specificity of family expansion.

One of the most dramatic examples of lineage-specific expansion and divergence in vertebrates has involved a specific family of transcription factor genes, encoding proteins in which a chromatin-interaction domain called KRAB (kruppel-associated box) is associated with tandem arrays of *Kruppel*-type (C2H2) zinc finger motifs (Shannon et al., 1998; Looman et al., 2002; Shannon et al., 2003). The KRAB domain confers a potent transcriptional repressor function to the proteins by specific interactions with a corepressor, KAP1 (or TRIM28) which serves to recruit chromatin deacetylation machinery (Ayanathan et al., 2003). KRAB-zinc-finger (ZNF) genes are a recent invention, first arising around the time of tetrapod divergence and duplicating aggressively to comprise a family of more than 400 active members in the human

genome (Huntley et al., submitted). Most KRAB-ZNF genes reside in large familial clusters generated by repeated rounds of tandem *in situ* duplication of ancestral genes, although distributed single-gene segmental duplicates are also observed. Because of ongoing rounds of gene duplication and gene loss, most conserved KRAB-ZNF clusters contain a substantial number of lineage-specific genes; in addition, entire clusters that are unique to certain mammalian lineages have been described (Krebs et al., 2005; Eichler et al., 1998; Huntley et al., submitted). Newly duplicated genes appear to diverge quickly through positive selection on single-nucleotide changes and structural changes within the ZNF arrays that are likely to affect DNA binding properties and potentially, target choice for the duplicated transcription factors (Shannon et al., 2003; Hamilton et al., 2003; Schmidt and Durrett 2004; Krebs et al., 2005)

We have speculated that the rapid expansion and lineage-specific diversification of these genes allow the "fine-tuning" of transcriptional regulation and diversification of regulatory networks in evolution (Hamilton et al., 2003; Huntley et al., submitted). Alteration of regulatory networks in a lineage-specific manner could contribute to population differences and play a role in speciation. The functional features and evolutionary histories of lineage-specific KRAB-ZNF genes are particularly interesting in this regard. One large cluster of primate-specific KRAB-ZNF genes near the centromere of human chromosome 19 was identified in earlier studies (Bellefroid et al., 1995; Eichler et al., 1998), although the ages and evolutionary histories of most of the genes remains uncertain. Our recent survey has shown that this single gene cluster contains 39 functional KRAB-ZNF genes and indicated that most of these are in a subfamily that contains additional recent relatives at other chromosomal sites (Huntley et al., submitted). Here we present a detailed analysis of this large primate-specific clade of transcription factor genes with a focus on investigating the most recent duplication events. In particular, we were interested in estimating the patterns and rates of structural and functional divergence of the member genes. Through the larger functional pathways and regulatory networks the predicted proteins influence through their regulatory activities, these more recent changes may help explain some aspects of transcriptome differences that have been documented between humans and our closest relatives among

the apes (Enard et al., 2002; Khaitovich et al., 2004, Preuss et al., 2004), and potentially, individual differences within our own species.

RESULTS

The ZNF91 subfamily

A distinction must be made between the HSA19p12 KRAB-ZNF gene cluster *per se*, which contains two genes not closely related to other loci in this region, and the larger group of related genes to which most KRAB-ZNF genes in the 19p12 region belong. We will refer to this related clade of primate-specific genes as the "ZNF91 subfamily", in reference to one of the best-known members of the group (Bellefroid et al., 1993). In addition to HSA19p12 genes (occupying a region defined by hg17 sequence coordinates chr19: 19,639,970-24,116,402 which extends from 19p12 into 19p13.1) members of the ZNF91 subfamily include members of a large clustered group that spans the centromere of chromosome 7 (chr7: 55,729,371-57,343,922 and chr7: 61,990,432-64,310,197). Together the HSA7 and HSA19 gene clusters contain over 100 loci about half of which are intact genes and are predominated by ZNF91 subfamily members (Table 1). BLAT and BLAST searches using sequences from the KRAB A, spacer, and zinc fingers of ZNF91 and several other genes identified additional related loci in other chromosomal locations, including HSA4p (chr4 : 43,215-482,891; containing known gene ZNF141), and HSA1q (243,434,894-243,821,086; including known gene SBZF3) (Figure 1).

To identify functional genes and pseudogenes, we compared ZNF91-related BLAST matches to the recently completed human KRAB-ZNF gene catalog (Huntley et al., submitted; http://www.znf.llnl.gov). The chromosome 19p12 cluster contains 40 loci capable of encoding functional KRAB-ZNF proteins ("full-ORF" genes) plus multiple pseudogene sequences, most of which correspond to partial duplication events (Huntley et al., submitted). All but two of the 39 full-ORF HSA19p12 KRAB-ZNF genes (ZNF101 and ZNF14) are closely related to ZNF91 based on comparisons of the KRAB and/or spacer sequences. The related centromere-spanning HSA7 gene cluster and HSA4p cluster contain 16 and 5 full-ORF KRAB-ZNF genes, respectively. The HSA1q gene cluster includes 6 intact genes; 2 of these genes are members of the ZNF91 subfamily, 3 are related to ZNF101, and one gene corresponds to an unrelated gene of the

SCAN-KRAB type (Huntley et al., submitted). Other members of the ZNF91 subfamily identified by sequence similarity include 15 pseudogene loci in four groups on chromosome Y, and 24 other isolated loci scattered across 16 chromosomes. Five of these scattered loci are also full-ORF KRAB-ZNF genes and the pseudogenes include 6 processed pseudogenes that appear to be retroposed copies of ZNF91 subfamily members. All told, the ZNF91 subfamily comprises at least 65 full-ORF protein-coding genes, or nearly one-sixth of the 405 human KRAB-ZNF loci identified in our recent study (Huntley et al., submitted) (Table 1). While ZNF91-subfamily genes are by no means the only primate-specific KRAB-ZNF loci in the human gene set, this subfamily does represent the largest clade of genes in the recently expanded group.

Phylogenetic Analysis

To examine the evolutionary history of the ZNF91 subfamily we generated phylogenetic relationship trees based on sequence regions including coding and noncoding sequences. The spacer (tether) region of ZNF91 and its relatives contains a sequence pattern distinguishing KRAB-ZNF loci of this subfamily (Bellefroid et al., 1995) and this region also typically corresponds to the most divergent and potentially informative protein-coding sequences in KRAB-ZNF genes (Shannon and Stubbs, 1998). Sequences encoding the spacer, together with flanking regions from the pre-spacer intron and the two 5'-most zinc fingers, (not counting remnant fingers in the spacer) were therefore chosen as the focus of the phylogenetic analysis. Phylogenetic trees constructed using the neighbor-joining method revealed multiple well-supported clades, some of which are highlighted in Figure 2 and discussed below.

Genes residing within the large gene clusters located in centromeric regions of HSA19p and HSA7 are intermingled throughout the phylogeny. The pattern is suggestive of a history in which one cluster was spawned from another (probably through the duplication or translocation of multiple loci) some time after the start of ZNF91 subfamily expansion, in early primate evolution. The phylogeny further suggests that after cluster separation, the family continued to expand by tandem duplications of genes within each cluster. However, the history of this subfamily also included the spread of genes to multiple locations throughout the human genome.

The clusters on chromosome 19 and chromosome 1 are particularly interesting because they both contain a mix of genes from more than one KRAB-ZNF subfamily. Two genes located at the p-telomeric end of the large HSA19p12cluster, ZNF101 and ZNF14, are members of the "KRAB-C subfamily" and have very divergent spacer sequences that made alignment with the ZNF91 relatives difficult. These genes contain the KRAB C motif (Looman et al., 2004) instead of the KRAB B exon that is typical for most other KRAB zinc finger genes, and share this trait with genes in another recently expanding HSA19p gene cluster (chr19: 11569297-12601676). Because the HSA1 gene cluster also contains ZNF101-related genes intermixed with ZNF91 subfamily loci, it is likely that an ancient progenitor gene cluster including both subfamilies of genes was duplicated before the massive expansion of the ZNF91 subfamily.

Also of note are several clades comprised of genes residing in locations other than HSA1, 7 or 19. The 'Y-chromosome based clade' (Fig. 2) inlcudes members on four different chromosomes including some members that map within the HSA7 cluster. The most significant expansion of this subgroup, in terms of copy number, has occurred on the Y chromosome. All of the Y-linked loci correspond to 'fingers-only' fragmentary pseudogenes, however, their relatives on HSA7 are intact. The chromosome Y loci may therefore have originated from a set of pseudogenes on chromosome 7 that were translocated and then further duplicated. HSAYp11.2 contains a group of six loci, and three of these have apparently been co-duplicated in a series of segmental duplications to create three additional clusters on Yq11.223-11.23. These three recently duplicated chromosomal segments are greater than 99% similar to each other over a distance of nearly 400,000 kb, an apparently very recent segmental duplication (Bailey et al., 2002; Cheng et al., 2005). However, this sequence similarity alone does not mean that these represent human-specific duplication events. The Y chromosome contains large regions that have undergone concerted evolution (Skaletsky et al., 2003; Rozen et al., 2003) and it is likely that the true age of these duplicates (referred to as 'ZNF381P' in the supplemental data of Skaletsky et al., 2003) has been obscured.

A distinct clade highlighted on Figure 2 corresponds to a group of dispersed, singleton loci that appear to have arisen through distributed segmental duplications. No clear cluster-bound progenitor is apparent for this group (it may have been lost from the

human genome after the spread of the dispersed or 'orphon' [Graur and Li, 2000] loci). This dispersed clade contains several recently-originated loci that appear to be novel in great apes and perhaps even in humans (up to >99% identity over variable segmental-duplication lengths of <30 to160kb). These loci are probably pseudogenes, since they contain stop codons in the KRAB B exons (although these could be spliced out), and contain ZNF exons that encode a limited number of intact zinc fingers (Huntley et al., submitted).

One HSA11 gene, ZNF195, is an example of an isolated, dispersed intact ZNF91-subfamily gene. This gene is not a recent duplicate such as those mentioned above; its spacer sequence is much more divergent from other ZNF91 subfamily members and does not group strongly with any particular clade, however it encodes a complete KRAB-ZNF protein and is expressed. ZNF195 demonstrates that duplicated ZNF genes separated from the cluster environment can be preserved as a functional gene in a new location by natural selection. We have dubbed such dispersed intact ZNF genes "scouts" because of the possibility that this mode of duplication, which appears to have been utilized much less frequently than tandem *in situ* events, has the potential to seed gene clusters at new chromosomal sites (Ohno, 1970; Eichler et al., 1998).

A possible example of this is another group of genes clustered together at the telomere of HSA4p. The ZNF141 cluster genes form a monophyletic group (Fig.2) and therefore may have arisen through tandem duplications from a single progenitor locus.

In addition to tandem *in situ* and dispersed segmental duplications, the ZNF91 subfamily includes six dispersed, intronless loci; these copies correspond clearly to KRAB-ZNF processed pseudogenes and most can be related to a progenitor locus or at least a clade of related loci.

Gene expansion and recently duplicated genes

Although the primate-specific expansion of the ZNF91 subfamily created gene copies on many chromosomes, most gene duplications were *in situ* tandem duplication events concentrated in two locations: the centromeric clusters of chromosomes 19 and 7. In the phylogeny these genes are intermixed (Fig.2), indicating that the progenitor cluster was already expanding before the event that split off the chromosome 7 cluster from the

chromosome 19 cluster. There are several well-supported clades within each cluster that do not have members in the other location, and therefore represent groups of loci that have been generated by duplication since the HSA19 and HSA7 gene clusters were split. As illustrated in Figures 1 and 2, most of the well-supported related groups of HSA19 genes are cluster neighbors, indicating each clade has expanded with new copies landing close to their progenitors. However, the neighboring genes are often found in opposite orientations and are sometimes separated by more distantly related genes; these arrangements suggest that internal rearrangements within the cluster have also occurred. In striking contrast to this relatively simple pattern, the well-supported clades within the chromosome 7 cluster are widely interspersed. Many clades contain members located on both sides of the centromere (Fig. 1, Fig 2). These data suggest that clade expansion was followed by a series of duplications or inversions around the centromere, as hypothesized for another ZNF cluster on HSA10 (Tunnacliffe et al., 1993).

Alignment of non-coding sequences from within the duplicated regions permitted an approximate age to be determined for specific duplication events that formed this primate-specific subfamily. Pairwise comparisons of intron, 3'UTR, and other noncoding sequences between intact loci of different clusters showed divergence levels consistent with the scenario that much of the expansion of the ZNF91 gene subfamily occurred before the split between the Catarrhini (old world monkeys+hominoids [apes and humans]) and Platyrrhini (new world monkeys) which is estimated to have occurred around 35 million years ago (Li et al., 1987; Li, 1997; Glazko and Nei 2003). For instance, the most similar HSA19p/HSA7 gene comparisons show about 85% non-coding nucleotide sequence identity (not shown). This pattern is in line with the conclusions from previous studies on members of the ZNF91 subfamily (Bellefroid et al., 1995, Eichler et al., 1998). However, noncoding sequence similarity was above 90% in alignments of many ZNF91 subfamily gene pairs within the HSA19 and HSA7 clusters (plus in the aforementioned Y-chromosome and dispersed clades of pseudogene loci). At least 18 genes from the ZNF91 subfamily are included in known recent segmental duplications (Bailey et al., 2002) not counting pseudogenes and examples where only small segments of loci were in recognized 'recent duplications'. For XX comparisons we detected sequence similarity >93-95% in the introns, indicating that these duplications

may be more recent that the ape/old world monkey split (Li et al., 1987; Li, 1997; Liu et al., 2003), although this includes the nine Y-chromosome loci which may have undergone recent gene conversion. These data indicate that certain members of the ZNF91 family have continued to duplicate throughout the evolution of apes and humans.

Focus on two clades of recent gene duplicates

Several HSA19p12 clades show evidence of recent duplication involving intact genes and may reveal the initial stages of diversification in the zinc-finger arrays of new genes. Two that were more intensely studied have been termed the ZNF431 clade and the ZNF492 clade, named after member genes involved in recent duplication events. The ZNF431 clade (highlighted in blue-green on Figs. 1 and 2) includes ZNF430, ZNF431, ZNF100, ZNF493, LLNL618, LLNL744, and the pseudogenes LLNL745 and LLNL1008. The ZNF492 clade (yellow in Figs. 1 and 2) is comprised of ZNF492, LLNL1168, LLNL622, the pseudogene LLNL621, and a processed pseudogene on chromosome 9, LLNL1040. The genes in each clade were identified as closely related based on comparisons of the spacer, pre-spacer intron, and other sequences such as the 3'UTR, with up to 93-98% identity in some cases suggesting that the genes may have duplicated within the hominoid branch leading to humans.

ZNF431 clade

The proposed alignment of zinc finger arrays for several members of the ZNF431 clade is shown in Fig.3. The closest relatives within the group (based on intron, spacer and 3'UTR alignments) are ZNF431 and LLNL618 (93.7% nucleotide sequence similarity in the included intronic segment), and ZNF100 plus LLNL745 (93% similar over the same segment). PCR using unique primers across a panel of ape genomic DNA samples (Fig. 4) showed that several of the genes in the clade date back to the old world monkeys, an observation that was confirmed by BLAT searches of the available chimpanzee and rhesus genome sequences. No ortholog for LLNL618 was found in rhesus by PCR or in sequence searches, and LLNL745-homologous sequences were not found in either rhesus or chimpanzee by PCR or BLAT. However, a homolog of the LLNL745 pseudogene was detected by PCR and sequencing off of gorilla DNA. It is

possible the LLNL745 locus was deleted in chimpanzee+bonobo. Here we note, however, that PCR and sequence searches can only provide a minimum for the age of each locus. For example, failure to detect a gene copy by PCR could result from divergence in primer sites, and specific genes may be missing from the draft chimpanzee and rhesus genome sequences.

For some loci, the alignments of available orthologous zinc finger arrays and the PCR results indicated a conservation of finger number and arrangement across species. However, there are exceptions; the human version of ZNF493 has extra zinc fingers compared to chimpanzee and rhesus (Fig.5). A finger-array deletion has also occurred in chimpanzee LLNL744 relative to its orthologs in human and rhesus, and for ZNF100 the rhesus genomic sequence contains an additional finger motif not found in the human or chimpanzee (Fig. 3). The LLNL745 locus also has a different number of fingers in human and the gorilla sequence (not shown) but is a pseudogene in both.

By contrast, paralog divergence in this family has frequently involved changes in the number and arrangement of zinc fingers. Sister loci, ZNF431 and LLNL618, differ by the insertion of an extra zinc finger motif in the array and have non-synonymous changes in most of the remaining zinc finger motifs (Fig. 3) while similar changes also distinguish the finger-array sequences of their other paralogs. LLNL744 and ZNF493 display much more dramatic differences in their respective zinc finger arrays when compared to the rest of the genes in this clade and even to each other. Most of the individual finger motifs in LLNL744 and ZNF493 are divergent enough from those in their relatives to hinder alignment beyond the 5' end of the spacer+zinc-finger coding exon, a situation usually encountered when comparing older genes. It is possible that this is due to rapid evolution of the amino acid sequences of multiple fingers, which when combined with finger gains and losses could mask homology. ZNF493 has apparently gained fingers through internal duplications (for instance, finger #6 is related in sequence to #15 and #8 to #18). The human-specific change in the ZNF493 fingers exon relative to other primates suggests that structural divergence of this gene is an ongoing process. The expansion in finger number in ZNF493 appears to be more complex than a simple "block" duplication of multiple fingers creating a sudden array-length increase, as was detected in ZNF43

(Lovering and Trowsdale, 1991), ZNF91 (Bellefroid et al., 1993), or ZNF208 (data not shown).

ZNF492 clade

This clade of rapidly-evolving genes includes five loci, of which ZNF492 and LLNL1168 have the same number of zinc fingers, and have pre-spacer intronic sequences that are 98.6% similar over the included segment (Fig. 6). These two genes are included in a recent segmental duplication of the whole loci with over 97% overall sequence similarity (Bailey et al., 2002). This very high level of similarity is notable considering that the available chimpanzee and rhesus genome assemblies contain only a single sequence each that is the reciprocal BLAT match to ZNF492 or LLNL1168 (again with the caveat that the coverage of these genomes is probably incomplete). However, PCR results (Fig. 7) revealed several ZNF492-clade sequences in bonobo, gorilla, and orangutan which may be modified versions of the loci known in humans or represent independent recent duplication events; this clade of genes may be unusually evolutionarily dynamic.

The similarity between the two human genes, ZNF492 and LLNL1168, does not signify redundancy. Despite the overall similarity in the ZNF arrays if these two genes one zinc finger in human ZNF492 has become 'degenerate' due to a mutation disrupting the structurally critical C2H2 pattern of the zinc finger motif. Degenerate fingers cannot bind DNA and potentially affect the overall functional properties of the ZNF array. It should be noted that the chimpanzee and rhesus orthologs of ZNF492 each also contains a degenerate finger motif, although these mutations affect different fingers in each species (Fig. 6). Also, LLNL1168 and ZNF492 may not have the same translation start sites, although it is probable not all splice variants are known for these genes.

Remaining ZNF492 clade members LLNL622, LLNL621, and LLNL1040 are present in the chimpanzee genome but not represented in the available rhesus scaffolds at the time of this submission. While these results are consistent with the high similarity and probable recent origin of LLNL621 and LLNL1040, LLNL622 is divergent enough from the rest to be present in old world monkeys, and this locus may have been deleted in the macaques or is missed by the current sequence assembly. LLNL622 contains a stop

codon that eliminates several 3' end zinc fingers that remain intact in ZNF492 and LLNL1168 (demonstrating another mechanism that can alter the DNA-binding region of these genes), and has other finger-array changes that readily distinguish it from its paralogs. However, a constant number of finger motifs are usually conserved between orthologous loci, except for LLNL621 in human vs. chimp. LLNL1040 lacks introns and is considered a processed pseudogene; although it does appear to retain an open reading frame containing the KRAB and several fingers. LLNL1040 and LLNL621 are most closely related but LLNL621 could not be the progenitor of LLNL1040 unless LLNL621 was formerly a fully functional gene that has since lost its KRAB, or perhaps both had another functional relative in other primates that is missing in humans.

Zinc finger sequence evolution

Most of the results detailed above deal with changes in zinc finger motif number and arrangement. Another way in which the zinc finger array can change is through nucleotide substitutions that alter the sequence of amino acids that are involved in DNA binding. In C2H2 zinc fingers, amino acid positions numbered from -1 to 6 relative to alpha helix of the 'finger-like' loop are the most variable in sequence reflecting their role in determining DNA binding specificity for the proteins. In particular positions -1, 3 and 6 considered to be those most critical for target recognition (Choo and Klug 1994). These amino acid positions have been shown for certain sets of duplicated genes to evolve under positive selection, presumably reflecting a drive to create new proteins with different DNA binding capabilities (Hamilton et al., 2003).

The finger array sequences of aligned paralogous and cross-primate orthologous loci in the ZNF431 and ZNF492 clades were tested for evidence of natural selection by comparing the pairwise values for dN (nonsynonymous changes per nonsynonymous site) and dS (synonymous changes per synonymous site). The Nei-Gojobori (1986) method for calculating dN and dS and tests for selection were implemented in the program MEGA (Kumar et al., 2001). In analyses including complete ZNF array sequences, most comparisons had a dN/dS ratio below 1 (the trend towards purifying selection; see table S2). A Z-test for selection indicated significant evidence for purifying selection in almost all comparisons, with the notable exceptions of many comparisons involving

pseudogenes which would have a relaxation of selection. These results reflect the fact that most amino acids comprising the zinc finger motif are structural, and therefore very highly conserved. However, when only the sets of six (amino acids –1 through 6 relative to the helix, omitting #4) or three (–1, 3,and 6) amino acid positions in the DNA binding regions were included in the alignment, a trend towards higher dN/dS ratios was observed, primarily in comparisons between paralogs where dN/dS was often >1 (Table S2). Alignments involving only the DNA binding sequences involve a small number of positions and there was no significant indication of positive selection; however, the trend seen here is similar to that observed in other comparisons between related zinc finger arrays (Shannon et al., 2003).

Expression of full length genes and alternative splicing

To provide further clues to the functional roles of ZNF91 subfamily genes, we examined tissue-specific expression of selected loci using RTPCR. Many genes of this family are widely expressed, but we also found examples of genes that are transcribed in a very limited number of tissues. For example, LLNL746 is expressed only in testis (Figure 8). Because many known KRAB-ZNF genes are alternatively spliced to include or eliminate specific types of motifs (Lovering and Trowsdale, 1991; Bellefroid et al., 1993), we also examined expression of potential splice variants using primers designed to detect different exons. Results of these experiments confirmed that splicing isoforms are generated by a number of the primate-specific genes. For example, ZNF85 produces a transcript with the typical KRAB+ZNF structure as well as an alternate transcript that includes a portion of the intron between the KRAB B exon and the spacer. This splice variant, which is expressed differently from the full-length isoform, would be predicted to produce a protein with a KRAB domain but no zinc fingers.

In another example, confirmed ZNF43 splice variants differ in 5' UTR exon structure; one isoform initiates with one exon 5' of the KRAB-A exon, in a HERV70/ERV1 LTR repeat, while the second starts further upstream, has three 5' exons, and although it skips the first exon of the aforementioned splice variant its apparent first exon is also within a HERV70/ERV1 LTR repeat (Fig. 8). The predicted translation of the first splice variant has a start codon in the first exon and therefore a complete KRAB-

A, but the second splice variant is predicted to have a truncated KRAB-A. These variants therefore would have both different promoters and different potential to repress their targets. Predicted first exons on many (but not all) of the genes in the cluster (including examples from the ZNF431 and ZNF492 clades and ZNF85 in addition to the two ZNF43 isoforms above) overlap with the same type of LTR repeat sequence. If the promoter and first exon are originating from this region, there may have been an ancient association of this repeat with the standard ZNF91-clade duplicon. The association of LTR repeats may indicate that the repeats are involved in driving gene expression as has been suggested previously for other genes (Di Cristofano et al., 1995). However, without additional experimentation we cannot be certain that these predicted first exons actually correspond to promoters for the ZNF91 genes.

DISCUSSION

ZNF91 and related genes were among the first KRAB-ZNF loci to be described, and the HSA19p12 gene cluster has been known for several years to be primate-specific (Eichler et al., 1998; Bellefroid et al., 1995). Indeed, this single subfamily comprises a substantial fraction of the estimated total of 55-62 kruppel-type ZNF genes involved in recent segmental duplications (Bailey et al, 2002; Huntley et al., subm). As elaborated here, the genes have expanded primarily through ongoing rounds of single-gene tandem *in situ* duplications, but have also increased in number through segmental duplications inserted into distant chromosomal sites. In addition to chromosome rearrangement events that have split conserved clusters into unlinked locations (Dehal et al., 2001; Huntley et al., submitted), these distributed duplication events may have been a major source for seeding new lineage-specific clusters over evolutionary time.

In agreement with previous estimates (Bellefroid et al., 1995, Eichler et al., 1998), data presented here for the full set of human genes indicate that the first expansion of the ZNF91 family occurred before the old world monkey/new world monkey split. However, we also define more recent clades containing duplicated genes with noncoding nucleotide similarity >93-95% (the approximate divergence for neutral sequences in ape species vs. old world monkeys) and even up to 98% similarity (approaching the level where a duplication could be unique to great apes or human-specific). In confirmation of these

estimates, many of the duplicons from this subfamily overlap with the recent segmental duplications surveyed by Bailey et al., (2002). Therefore, although the ZNF91 subfamily may have its origins in early primate evolution, specific members of the subfamily have clearly continued to duplicate through the rise of apes and humans.

Evidence of rapid evolution within the zinc-finger array of ZNF91 was first noted in a study of the gene in different primates (Bellefroid et al., 1995). This type of rapid structural divergence, involving positive selection on DNA-binding amino acids and deletions and duplications of tandemly arranged ZNF motifs, has since been shown to be a general property of genes in the KRAB-ZNF family (Shannon et al., 2003; Hamilton et al., 2003; Krebs et al., 2005; and this report). Aside from very recent duplications such as ZNF492 and LLNL1168, we documented very few examples of ZNF91 subfamily paralog sets that have retained the same number and arrangement of zinc finger motifs. These data suggest that such rearrangements are very frequent, and provide a major avenue of rapid structural divergence for newly duplicated ZNF genes.

In addition, the homology of individual fingers in many sets of primate-specific paralogs is masked by the amino acid substations, particularly within the variable amino acid positions located in DNA-binding regions of each ZNF motif (Choo and Klug 1994; Greisman and Pabo 1997). In some proteins, shifts in the linear order of zinc fingers are accompanied by amino acid substitutions, creating more complicated alignments between related finger arrays. In common with other studies on various sets of KRAB-ZNF genes (Looman et al., 2002, Shannon et al., 2003; Schmidt and Durrett 2004), we documented a trend toward positive selection in comparisons of ZNF91-subfamily paralogs, especially in amino acid positions that are critical to determining DNA binding specificities. Together these data strongly suggest that the primate-specific KRAB-ZNF genes diverged rapidly after duplication in ways that can be predicted to alter DNA recognition sites and binding properties of the proteins.

By contrast, most orthologous comparisons showed greater conservation of amino acid sequence, especially within the DNA-binding ZNF arrays. However, we also documented examples of ongoing finger-array structural changes in established genes. In the ZNF91 subfamily, these include the human-specific addition of zinc fingers in ZNF493, the chimpanzee-specific modification of LLNL744, and cases of fingers

becoming degenerate. Hypothetically at least, the alteration in DNA binding specificity of zinc fingers due to amino acid sequence changes might be 'tracking' the concurrent evolution of the target DNA sequences, for example, adapting to substitutions at the target sites, while finger number change or degeneracy of internal fingers could improve binding to targets which acquired short insertions and deletions.

Alternative splicing offers another way to increase the total number and diversity of zinc finger proteins, and transcripts encoding different protein isoforms were documented for ZNF91 subfamily genes. Alternate transcripts that skip the KRAB A, KRAB-B or both effector-encoding exons are typical for this family (e.g. Lovering and Trowsdale, 1991; Bellefroid et al., 1993) The KRAB B domain enhances the activity of the dominant KRAB A domain, and isoforms that do not include both effectors may exhibit reduced levels of repressor activity (Vissing et al., 1995). Since the KRAB A is key to gene repression, isoforms that skip KRAB effector exons altogether should lack repressor activity. Although functions for "fingers-only" isoforms have not been established, they may serve a 'competitive interference' function, displacing full length proteins at DNA binding sites (Chong et al, 1995; Chen et al., 1998). Finally, several genes of the ZNF91 subfamily appear to use alternative 5' ends, which may place specific isoforms under different types of transcriptional regulation. KRAB-ZNF genes were apparently an exception to the suggested inverse relationship between gene copy number and splice variant number across gene families reported by Kopelman et al., (2005); due to the structure of these genes, splice variants and paralogous loci may both be selected for simultaneously as they provide different sources for diversity: isoforms result in functional changes by the splicing of modular effector motifs and expression changes by utilizing alternate 5'UTRs, while target-specificity changes occur by the alteration of the ZNF array in new gene copies.

What kinds of functional roles do these primate-specific transcriptional repressors play? A potential connection between ZNF91 subfamily genes and immune-system function has been widely discussed (Mark et al., 2001; Nishimura et al., 2001), suggesting a role in the 'arms race' between the body's defense mechanisms and the ever-changing suite of threats. This proposed role fits nicely with the rapid evolutionary divergence of the KRAB-ZNF family, and with known expression patterns and functional

data for a small number of genes. For example, ZNF91, ZNF43 and relatives are highly expressed in lymphoid or myeloid cell lines (Lovering and Trowsdale, 1991; Bellefroid et al., 1991; Bellefroid et al., 1993, Mark et al., 1999, Mark et al., 2001). In addition, ZNF91 has been reported as a putative transcriptional repressor for FcgammaRIIB, an immunoglobulin receptor (Nishimura et al., 2001).

However, RTPCR data presented here and information from other sources (Su et al., 2004) indicate a wider range of expression patterns and functions for the ZNF91 subfamily and other primate-specific genes. The testis-specific expression of family member LLNL746 provides an excellent example, and suggests a role in reproduction, another pathway that is potentially operating under intense evolutionary selection (Nielsen et al., 2005 mention positively selected KRAB-ZNF genes involved in spermatogenesis). Available data indicates that KRAB-ZNF genes are expressed in widely divergent patterns, with family members displaying high expression in brain, muscle, glandular tissues, and a wide range of reproductive organs (Su et al., 2004). In addition many ZNF91 subfamily members are expressed in nearly ubiquitous patterns. Here is should be noted, however, that widespread patterns of expression do not necessarily imply non-specific or housekeeping functions for these genes. For example, transcriptional repressor NRSF, which inhibits expression of neural genes in non-neural cells, is widely expressed but exerts a profound role on neurological development (Chong et al., 1995; Chen et al., 1998). It may therefore be the few tissues with lowest expression levels that are most revealing in terms of predicting functions for certain types of KRAB-ZNF genes. Other genes from the ZNF91 subfamily with known or predicted functions include TIZ ("TRAF6-inhibitory zinc finger protein"), which may indirectly regulate osteoclast differentiation from macrophages/ hematopoietic progenitor cells (Shin et al., 2002), ZNF85 which has been suggested to play a role in primate-specific modifications to the spermatogenesis pathway (Poncelet et al., 1998), ZNF43 which has been implicated in maintaining the undifferentiated state of Ewing sarcoma cells (its downregulation allowed neuronal differentiation; Gonzalez-Lamuno et al., 2002), and ZNF253 (ZNF411) was linked to the MAP kinase signaling pathway (Liu et al., 2004).

Because KRAB-ZNF genes are expressed in diverse tissues and probably impact a wide range of biological pathways, and since functions for so few genes of this class are

known, the impact of their evolutionary change remains a mystery. Although data is still incomplete, it is clear that large numbers of lineage-specific KRAB-ZNF genes, including many recent duplicates, also exist in rodent, canine, and other mammalian groups (Shannon et al., 2003; Hamilton et al., 2003; Krebs et al., 2005; Huntley et al., submitted). We conjecture that the prolific creation of lineage-specific KRAB-ZNF genes has provided a major mechanism for fine-scale tuning of mammalian regulatory networks, providing a major driver for evolution and speciation. The duplication and divergence of the KRAB-ZNF family, especially still-expanding subfamilies like the ZNF91-related genes, may have played an important part in the evolution of higher primates, and recent changes in the repertoire of these regulatory genes could underlie gene expression differences in the testes, regions of the brain, or other tissues that have been the focus of research on what makes humans distinct from our closest relatives (Enard et al., 2002; Khaitovich et al., 2004, 2005).

The modularity of KRAB-ZNF genes allows for multiple mechanisms of diversification to expand the number of distinct protein products. Even in the still-expanding ZNF91 subfamily, it is difficult to find paralogs that have the same number and arrangement of zinc fingers, testifying to the extreme diversity of targets these genes could be regulating. In addition, each locus can produce multiple splice variants that are functionally distinct, or have alternate promoters and therefore may themselves be regulated differently. That so many evolutionary lineages which acquired the KRAB-ZNF combination have ended up with large numbers but very different sets of these genes demonstrates the flexibility and adaptability of this type of regulator. KRAB-ZNF genes may not be the foundation of regulatory pathways, but they are the building blocks for the continuing evolution of our complexity.

METHODS

Phylogenetic analysis

Evolutionary analyses were done using the "tether" or spacer region of each locus, a region that has been shown to be diagnostic for ZNF91 subfamily members (Bellefroid et al., 1993). Although alignments of the KRAB exon were also prepared these were not combined with the spacer alignment due to results in a study of a rodent gene family in which the KRAB was homogenized between neighboring clustered genes within the mouse and rat, while the fingers array and the first exon + promoter region retained orthologous relationships between the mouse and rat genes (A. T. Hamilton, unpubl.). The spacer region was extracted from genes in the 19p12 cluster using the batch sequence retrieval capability of the LLNL Biosciences zinc finger gene catalog website (Huntley et al., submitted; http://www.znf.llnl.gov). For each spacer region, additional sequence was added before it (pre-spacer intron) and after it (including the first 2 good zinc fingers). The same was done for the related genes in the centromeric chromosome 7 cluster(s) and the telomeric p-arm chromosome 4 cluster. The spacer element sequence from ZNF91 and was then compared against the genome databases using the BLAST and BLAT search tools to find additional loci that shared this feature and could be added to the alignment. Additional genes were also examined if the KRAB A phylogeny (Huntley et al, submitted) showed their KRAB A exon sequences were similar to those of genes known to be in the ZNF91 subfamily. The KRAB data was also used to select the outgroups from the cluster on HSA19q13.41-42. However, the internal arrangement of the clades in the tree was found to be the same when alternate sets of spacer or fingers segments from various families were used as outgroup sequences. To gain a fully informative view of the evolutionary history of this subfamily we included both genes and fingers-containing pseudogene sequences in the phylogenetic analysis. The spacer+flanking sequence tree contained 116 loci. (Table 1 contains all the loci, plus other finger-containing loci in the major clusters, with gene/pseudogene designations).

After the removal of repetitive elements defined by Repeatmasker (Smit et al., 1996-2004) from the pre-spacer intron sequence, alignments of the spacer+flanking regions were made using Clustal X 1.81 (Thompson et al., 1997). The alignment was manually checked using SeAl (Rambaut, 1996). The total length of each sequence differed due to repeat removal, insertions, and deletions; a typical example, ZNF431, had 300 bp of pre-spacer intronic sequence and 465 bp of exonic sequence comprising of the

spacer (which contains the remnants of what may have been a degenerate zinc finger motif) and the first two good zinc fingers. The processed pseudogene loci sequences were also included, but since these had no intronic sequences, only the spacer and fingers region was included in the alignment. In regions of the sequence alignment where not all loci had sequence, those loci would have 'missing data' for those positions. Parts of the intron alignment where there was uncertainty with the alignment of many loci due to numerous indels were excluded from the phylogenetic analysis. The PAUP 4.0b10 package (Swofford, 2002) was used to generate trees using mean character differences and the neighbor-joining (NJ) method (Saitou and Nei 1987). The NJ trees were evaluated with 1000 rounds of bootstrapping (Felsenstein 1985). PAUP was also used to construct parsimony trees from the same data which did not differ greatly from the neighbor-joining results (trees not shown). For the aligned intronic sequences, divergence times were estimated based on the neutral evolution rates for pseudogenes or introns (Li et al., 1987; Li, 1997; Chen and Li 2001; Liu et al., 2003). The average human/chimpanzee difference in neutral non-repetitive sequences is usually estimated at slightly over 1% sequence divergence, while about 3% is seen for human vs. orangutan, 5-7% for human vs. old world monkeys, and 11-14% for human vs. new world monkeys.

For several well-supported clades containing recent duplications, a more intense scrutiny of the genes' zinc finger sequences and interspecies differences was carried out. The ZNF431 clade and ZNF492 clade were chosen from among the 19p12 cluster's zinc finger loci because both contained intact genes that may have arisen relatively recently in our evolution. Alignments of the zinc finger motifs were done when it was apparent that genes shared enough homologous fingers to permit the tracing of finger-array changes across multiple genes. The zinc finger sequences were also compared to the available chimpanzee and rhesus genomic sequence assemblies or fragments via BLAT searches on the UCSC browser (Kent et al., 2002). When reciprocal best-matching ZNF sequences were found, these were added to the alignment. The draft chimpanzee and rhesus sequences were 'repaired' of frameshift-causing insertions and deletions for the purpose of the alignment after some apparent indels were checked by sequencing of the primate PCRs and not confirmed.

PCR using unique primers for the genes in these clades was done across a panel of primate genomic DNA samples (Coriell phylogenetic panel PRP00001). The samples included chimpanzee (*Pan troglodytes*), bonobo (*Pan paniscus*), gorilla (*Gorilla gorilla*), orangutan (*Pongo pygmaeus*), rhesus monkey (*Macaca mulatta*), pigtailed macaque (*Macaca nemestrina*), common woolly monkey (*Lagothrix lagotricha*), black-handed spider monkey (*Ateles geoffroyi*), red-chested mustached tamarin (*Saguinus labiatus*), and a ring-tailed lemur (*Lemur catta*). The results of course only give a minimum estimate for the age of each gene (failed PCR only means the primer sites are different, not necessarily that the species predates the appearance of the gene).

The zinc finger motif variable amino acid positions were analyzed using alignments of the zinc finger array of genes from two clades, and as mentioned the comparisons included orthologs identified across primate species and between paralogs. Alignments were made of the whole array, and selected amino acid positions (a set of six [-1,1,2,3,5,6] and a set of three [-1,3,6]) implicated as the most vital in determining the potential target-specificity of a zinc finger motif (Choo and Klug 1994; Greisman and Pabo 1997). The MEGA program (Kumar et al., 2001) was used to calculate the number of nonsynonymous changes per nonsynonymous site and synonymous changes per synonymous site, using the modified Nei-Gojobori (1986) method, with pairwise deletion of missing data (caused by fingers aligned with gaps). The dN/dS ratios were calculated for the three subsets of aligned amino acid positions and the Z-test for selection was used to test for a significant indication of either purifying or positive selection. ZNF493 and LLNL744 were not included in the analysis of the ZNF431 clade due to the greater amount of change in the zinc finger sequences which hindered alignment.

Expression analysis

RTPCR was done on an array of human tissue samples (cDNAs made using Superscript (invitrogen) from total RNA and mRNAs purchased from BD Biosciences) using sets of primers designed to be unique to individual genes. Most primer pairs were 5'UTR F and in the spacer region of the spacer+fingers+3'UTR exon, with some additional forward primers for possible alternate 5'UTR ends, or KRAB sequences when the 5'UTR was uncertain. The exon-crossing primer sets were also used to verify co-

expression of KRAB and finger exons in putative gene models from the LLNL zinc finger catalog (Huntley et al, submitted). Also, some initial genes were chosen for RTPCR analysis due to multiple apparent isoforms in public databases; if potential isoforms appeared in the RTPCR analysis the bands were excised from agarose gels and sequenced. The expression patterns of the splice variants of the genes ZNF43 and ZNF85 were compared. RTPCR was done on human cDNA at equal concentration across the panel, confirmed by testing higher-expression and lower-expression housekeeping genes on the panel.

ACKNOWLEDGMENTS

We thank Colleen Elso, Elbert Branscomb and Joomyeong Kim for critical comments on the manuscript. This work was performed under the auspices of the U. S. Department of Energy (DOE) by the University of California, Lawrence Livermore National Laboratory (LLNL) under Contract No. W-7405-Eng-48. The project (04-ERD-084) was funded by the Laboratory Directed Research and Development Program at LLNL.

LIST OF FIGURES/FIGURE CAPTIONS

Figure 1

The map above shows the physical order of the KRAB-zinc finger loci in the four major clusters containing intact KRAB-ZNF genes related to the ZNF91 subfamily. Each finger-containing locus is represented by a blue arrow pointing 5' to 3' to represent the orientation of each locus. Loci that also contain a KRAB have the red bar at the 5' end of the symbol. Isolated red bars are KRAB-only pseudogene fragments, which are not labeled here. The maps are not scaled to show distances between loci or the relative size of the chromosome 7 centromere. Not all labeled genes in the cluster maps are included in the phylogeny in Figure 2 because the spacer region used to create the phylogeny was modified or deleted from the locus, or because they were very divergent and not members of the ZNF91 subfamily, notably the "KRAB C subfamily" loci (blue outline box) on the end of the chromosome 19 cluster and in the chromosome 1 cluster. The chromosome 7 cluster also has a divergent gene on its edge (FLJ39963) but relationships between this gene and other KRAB-ZNF loci are unclear.

Figure 2

Phylogeny of 116 loci related the ZNF91 subfamily and outgroups, using the spacer (tether) region of the spacer+fingers exon, plus the first 2 true fingers and the pre-spacer intron. The tree is a Neighbor-joining tree; 1000-replicate bootstrap values are indicated on branches. Gene names that have a number in front are not from the chromosome 19p12 cluster and the initial number indicates the chromosome the locus is from (i.e., 7_ZNF588). Loci with the ZNF catalog designation LLNLxxx are labeled 'Lxxx'. Colored boxes define certain well-supported clades which are given designations at the right. For the clades including genes from the main clusters, the box colors match the colored bars under the individual member genes on the cluster maps in Figure 1. Note that the "Y chromosome related clade" contains genes from the cluster on chromosome 7

but these are outnumbered by the pseudogene loci that ended up on Y. There is also a processed pseudogene member of this clade (LLNL817, labeled '5_rL817') and two additional pseudogene copies located on HSA 8.

Figure 3

A zinc finger alignment hypothesis for the ZNF431 clade, including the set of human paralogs and their orthologs if found in chimpanzee and rhesus genomic data. Each box represents a zinc finger (gaps are added when one locus has added or deleted a zinc finger so that flanking fingers remain aligned). The amino acid codes inside each box (-1 to 6, relative to the start of the alpha-helix) include the variable positions which are considered to be involved in sequence-specific DNA target recognition and binding. Degenerate fingers are shaded; stops in the sequence are shown for the human sequences, but are faded for LLNL745 because this locus has a stop in the predicted spacer (not shown) and is considered a pseudogene. Frameshifts in the genomic sequences are also indicated, but for the non-human primate sequences these are 'repaired' due to the incomplete nature of the non-human sequence data. (For instance, for ZNF431 our own partial sequencing of the locus did not show the same frameshift-causing mutations.) Most of the fingers of ZNF493 and LLNL744 are not shown due to high sequence divergence and difficulty in alignment with the other members of the clade; these genes are depicted in Fig. 5.

Figure 4

PCR results using locus-specific primers across a panel of primate genomic DNA. The samples are H human (homo sapiens), C chimpanzee (*Pan troglodytes*), B bonobo (*Pan paniscus*), G gorilla (*Gorilla gorilla*), O orangutan (*Pongo pygmaeus*), R rhesus monkey (*Macaca mulatta*), P pigtailed macaque (*Macaca nemestrina*), W common woolly monkey (*Lagothrix lagotricha*), S black-handed spider monkey (*Ateles geoffroyi*), T red-chested mustached tamarin (*Saguinus labiatus*), and L for the ring-tailed lemur (*Lemur catta*) and -C negative control. Vertical lines distinguish the apes and the old world monkeys (the macaques). Weak bands labeled + were confirmed; for unusual bands, the sequencing results are indicated.

Figure 5

The ZNF431 clade contains one additional locus, the pseudogene LLNL1008, which appears to have been a complete or near-complete duplication event but afterwards lost the spacer and perhaps multiple zinc finger motifs. Comparison of regions besides the intron and spacer sequences used for the phylogenetic tree revealed that LLNL1008 clusters with the ZNF431 clade in a KRAB A-based phylogeny (not shown), and its fingers match the dead fingers in the 3'UTRs of ZNF493 and LLNL744. It has lost the diagnostic spacer sequence however, and was therefore not included in the spacer phylogeny for Fig. 2. This diagram of ZNF493, LLNL1008 and LLNL744 shows regions within each duplicated locus which are closely related between these genes and to ZNF430 (gray arrows). As indicated, human ZNF493 has two additional zinc finger motifs compared to chimpanzee and rhesus orthologs.

Figure 6

A zinc finger alignment hypothesis for the ZNF492 clade, including the set of human paralogs and their orthologs if found in chimpanzee and rhesus genomic data. Finger boxes and codes are as in figure 3, Degenerate fingers are shaded; stops in the sequence are shown for the human sequences, but are faded for the loci considered to be pseudogenes (LLNL621 and LLNL1040). Frameshifts in the genomic sequences are also indicated, but for the non-human primate sequences these are 'repaired' due to the incomplete nature of the non-human sequence data. The yellow arrows in rhesus ZNF492 represent a finger in the alignment that is divergent from the same-numbered finger in the gene's human and chimpanzee orthologs, The rhesus ZNF492 fingers 8 and 9 are similar enough to fingers 10 and 11 that in this case, instead of mutational change in the finger sequence, the pattern in this species could be explained by a loss of fingers followed by an internal duplication of the aforementioned finger motifs, restoring the array to the same number of zinc finger motifs but altering homology. For chimpanzee LLNL622, the available genomic sequence is missing the first three finger motifs, but the cross-primate PCR results showed a similar band for chimpanzee and human (Fig.7).

Figure 7

PCR results using locus-specific primers for ZNF492 clade members across a panel of primate genomic DNA. See Figure 4 for the list of species included. Vertical lines distinguish the apes and the old world monkeys (the macaques). Asterisks: PCR using primers designed for ZNF492 also picked up one additional band each in bonobo and gorilla which are very similar in sequence to human members of the ZNF492 clade, but with differences in several finger sequences indicating possible array changes. Primers designed for LLNL1168 picked up this locus in human but not chimpanzee, bonobo or gorilla, but did pick up a ZNF492-like sequence in orangutan. These PCR results cannot distinguish intact loci from fingers-only duplications (such as human LLNL621) and a more intensive sequencing project would be needed to discern the number of gene copies in each species and their location and orientation.

Figure 8

RTPCR gels showing the expression patterns for the KRAB-ZNF gene LLNL746 (a ZNF91 subfamily member on chromosome 19) and alternate splice variants of ZNF85 and ZNF43 targeted by RTPCR. Diagrams at right indicate primer locations (arrows) selected to selectively amplify the depicted isoforms for ZNF43 and ZNF85. At bottom are gels for two 'housekeeping' genes selected as positive controls for the cDNA panels. Guide to the tissues included on the gels: Adi= Adipose tissue; Adr=Adrenal gland; Bpl=Blood, peripheral leukocytes; BoM=Bone marrow; Br=Brain; Brcb=Brain (cerebrum); Bcll=Brain (cerebellum); Bm=Brain(medulla oblongata); Hrt=Heart; Liv=Liver; Lym=Lymph node; MG=Mammary gland; Pan=Pancreas; Ov=Ovary; Pla=Placenta; Pro=Prostate gland; Spl=Spleen; SkM=Skeletal muscle; Tes=Testis; Thr=Thyroid; Thm=Thymus; -C = negative control

Table 1

List of loci included in the phylogenetic analysis (plus additional genes found in clusters with ZNF91 subfamily members that were not included because they were very divergent and members of other subfamilies). Gene/pseudogene calls are from the LLNL ZNF

catalog.

Table S2

List of MEGA results with dN/dS ratios for pairwise comparisons of paralogs and orthologs. On the left are loci from the ZNF431 clade; on the right are comparisons between loci in the ZNF492 clade. Three tables are presented for each clade, using different subsets of the amino acid sequences of the genes. The bottom tables are for the whole zinc finger array (with pairwise deletion of finger sequences that could not be aligned as homologous); the middle and upper tables only include, for each finger motif, the set of six amino acid positions [-1,1,2,3,5,6 relative to the alpha helix] and the subset of three positions within this group [-1,3,6] which have been cited as the most vital in determining the potential target-specificity of each zinc finger motif (Choo and Klug 1994; Greisman and Pabo 1997). Highlighted cells are those in which there was significant evidence for purifying selection according to a Z-test for selection as implemented in MEGA (Kumar et al., 2001).

Figure 1

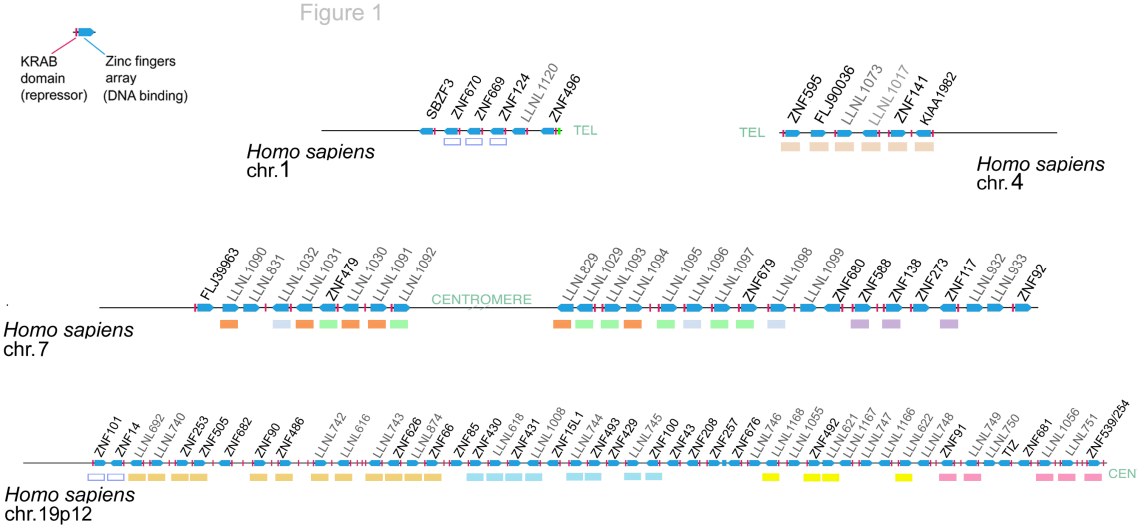

KRAB
domain
(repressor)
Zinc fingers
array
(DNA binding)
SBZF3
ZNF670
ZNF669
ZNF124
LLNL1120
ZNF496
TEL
Homo sapiens
chr.1
ZNF595
FLJ90036
LLNL1073
LLNL1017
ZNF141
KIAA1982
TEL
Homo sapiens
chr.4
FLJ38963
LLNL1090
LLNL831
LLNL1032
LLNL1031
ZNF479
LLNL1030
LLNL1091
LLNL1092
CENTROMERE
LLNL829
LLNL1029
LLNL1093
LLNL1094
LLNL1095
LLNL1096
LLNL1097
ZNF679
LLNL1098
LLNL1099
ZNF680
ZNF588
ZNF138
ZNF273
ZNF117
LLNL932
LLNL933
ZNF92
Homo sapiens
chr.7
ZNF101
ZNF14
LLNL692
LLNL740
ZNF253
ZNF505
ZNF682
ZNF90
ZNF486
LLNL742
LLNL616
LLNL743
ZNF626
LLNL874
ZNF66
ZNF85
ZNF430
LLNL618
ZNF431
LLNL1008
ZNF15L1
LLNL744
ZNF493
ZNF429
LLNL745
ZNF100
ZNF43
ZNF208
ZNF257
ZNF676
LLNL746
LLNL1168
LLNL1055
ZNF492
LLNL621
LLNL1167
LLNL747
LLNL1166
LLNL622
LLNL748
ZNF91
LLNL749
LLNL750
TIZ
ZNF681
LLNL1056
LLNL751
ZNF539/254
CEN
Homo sapiens
chr.19p12

Figure 2

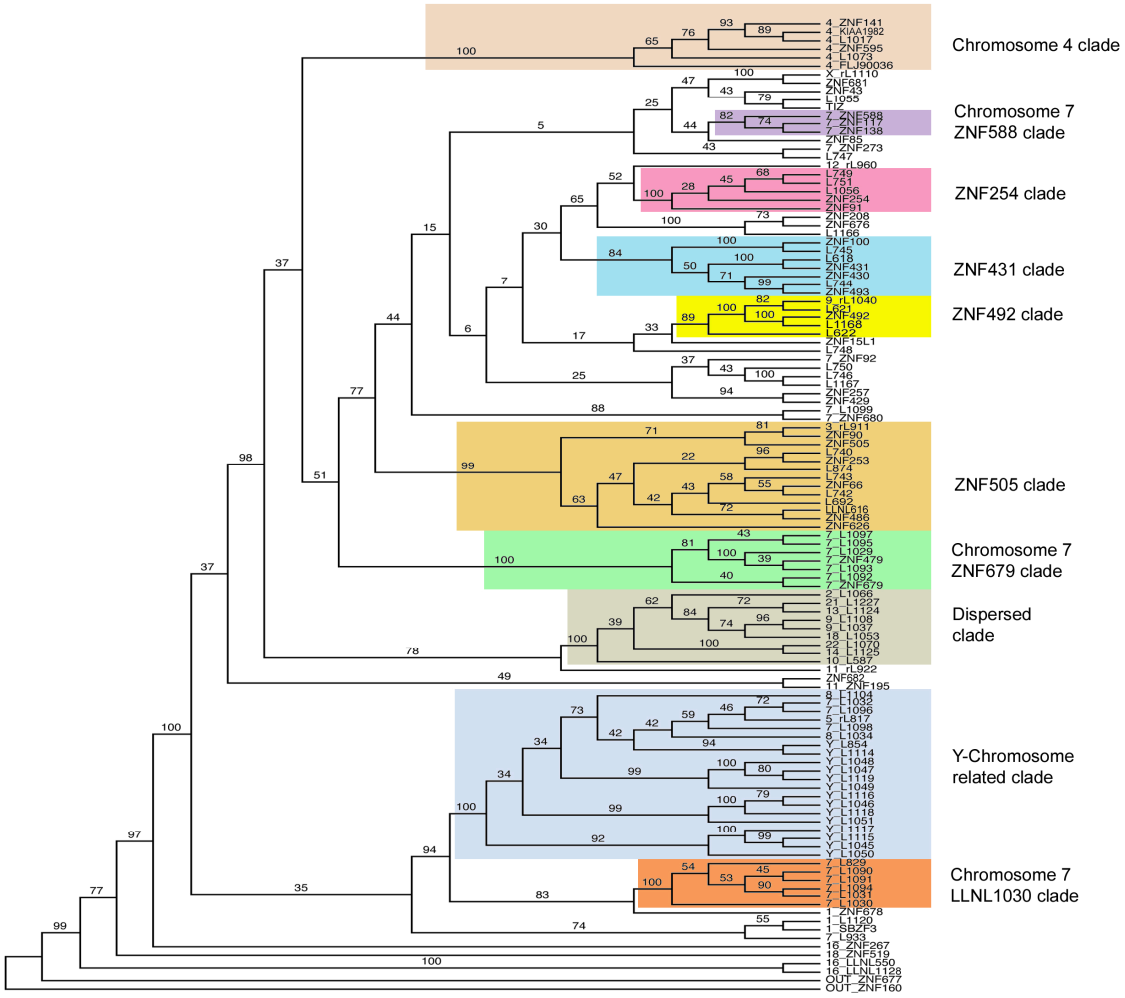

Chromosome 4 clade
Chromosome 7 ZNF588 clade
ZNF254 clade
ZNF431 clade
ZNF492 clade
ZNF505 clade
Chromosome 7 ZNF679 clade
Dispersed clade
Y-Chromosome related clade
Chromosome 7 LLNL1030 clade
4_ZNF141
4_KIAA1982
4_L1017
4_ZNF595
4_L1073
4_FLJ90036
X_rL1110
ZNF681
ZNF43
L1055
TIZ
7_ZNF588
7_ZNF117
7_ZNF138
ZNF85
7_ZNF273
L747
12_rL960
L749
L751
L1056
ZNF254
ZNF91
ZNF208
ZNF676
L1166
ZNF100
L745
L618
ZNF431
ZNF430
L744
ZNF493
9_rL1040
L823
ZNF492
L1168
L822
ZNF15L1
L748
7_ZNF92
L750
L746
L1167
ZNF257
ZNF429
7_L1099
7_ZNF680
3_rL911
ZNF90
ZNF505
L740
ZNF253
L874
L743
ZNF66
L742
L692
LLNL616
ZNF486
ZNF626
7_L1097
7_L1095
7_L1029
7_ZNF479
7_L1093
7_L1092
7_ZNF679
2_L1066
21_L1227
13_L1124
9_L1108
9_L1037
18_L1053
22_L1070
14_L1125
10_L587
11_rL922
ZNF587
11_ZNF195
8_L1104
7_L1033
7_L1096
5_rL817
7_L1098
8_L1034
Y_L854
Y_L1114
Y_L1048
Y_L1047
Y_L1119
Y_L1049
Y_L1116
Y_L1046
Y_L1118
Y_L1051
Y_L1117
Y_L1115
Y_L1045
Y_L1050
7_L829
7_L1090
7_L1091
7_L1094
7_L1031
7_L1030
1_ZNF678
1_L1120
1_SBZF3
7_L933
16_ZNF267
18_ZNF519
16_LLNL550
16_LLNL1128
OUT_ZNF677
OUT_ZNF160

Figure 3

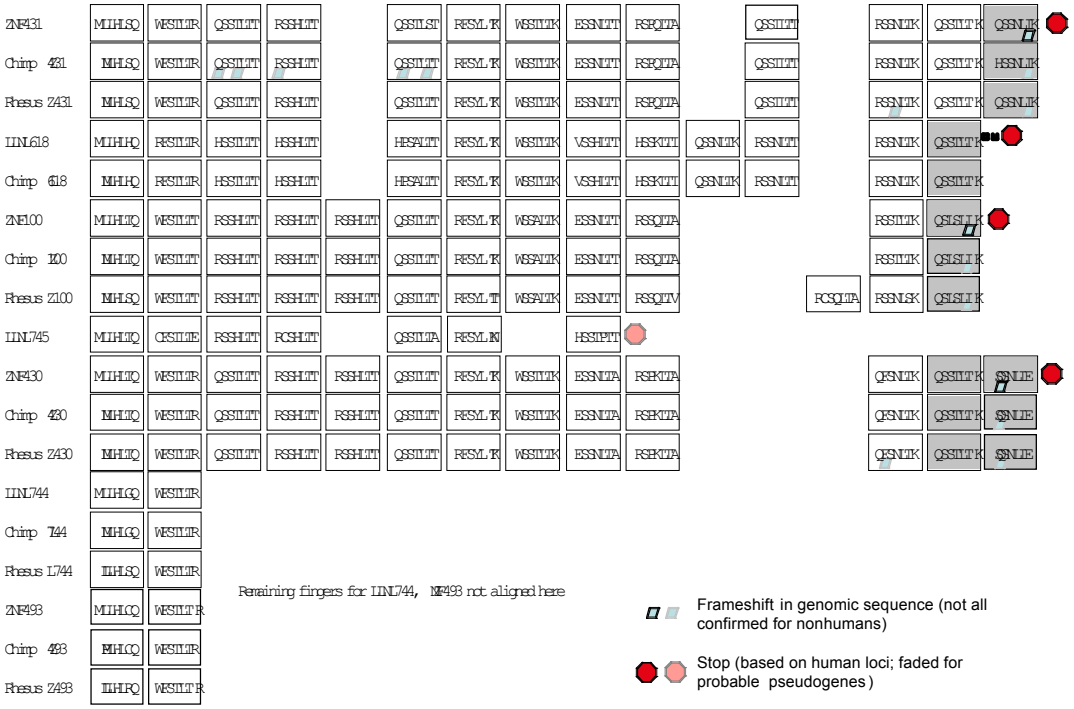

ZNF431
Chimp 431
Rhesus 2431
LINL618
Chimp 618
ZNF100
Chimp 100
Rhesus 2100
LINL745
ZNF430
Chimp 430
Rhesus 2430
LINL744
Chimp 744
Rhesus L744
ZNF493
Chimp 493
Rhesus 2493
Remaining fingers for LINL744, ZNF493 not aligned here
Frameshift in genomic sequence (not all confirmed for nonhumans)
Stop (based on human loci; faded for probable pseudogenes)

Figure 4

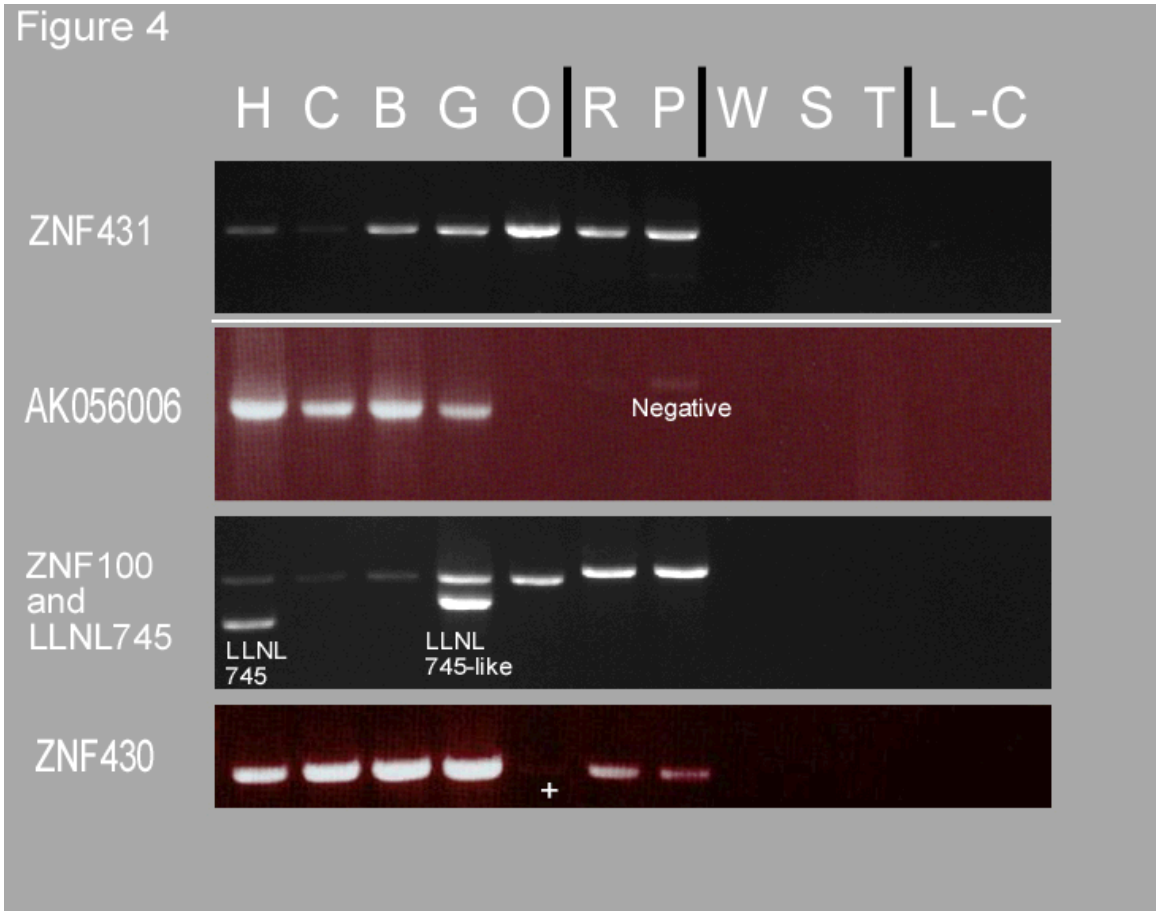

H C B G O R P W S T L-C
ZNF431
AK056006
Negative
ZNF100
and
LLNL745
LLNL
745
LLNL
745-like
ZNF430
+

Figure 5

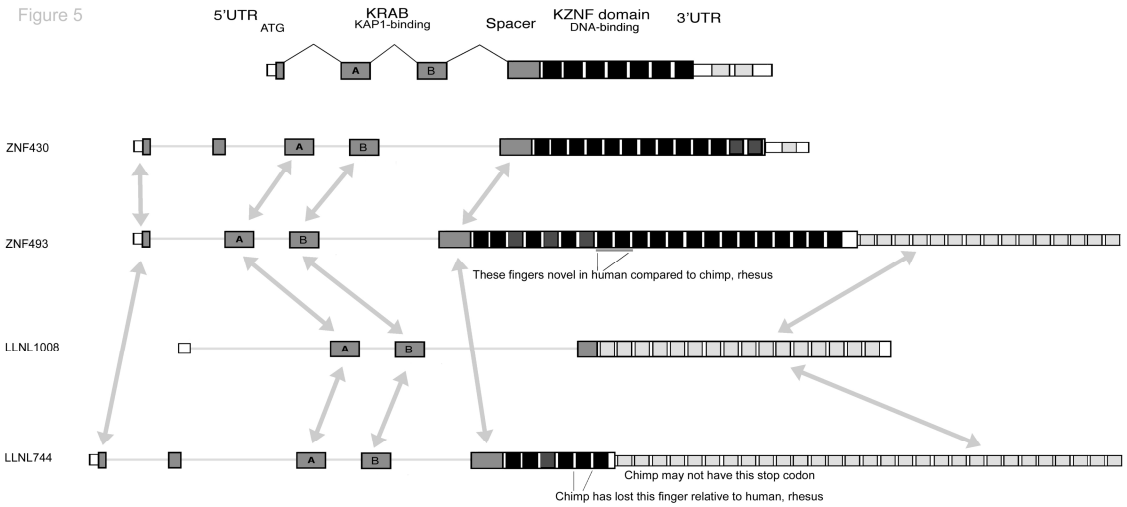
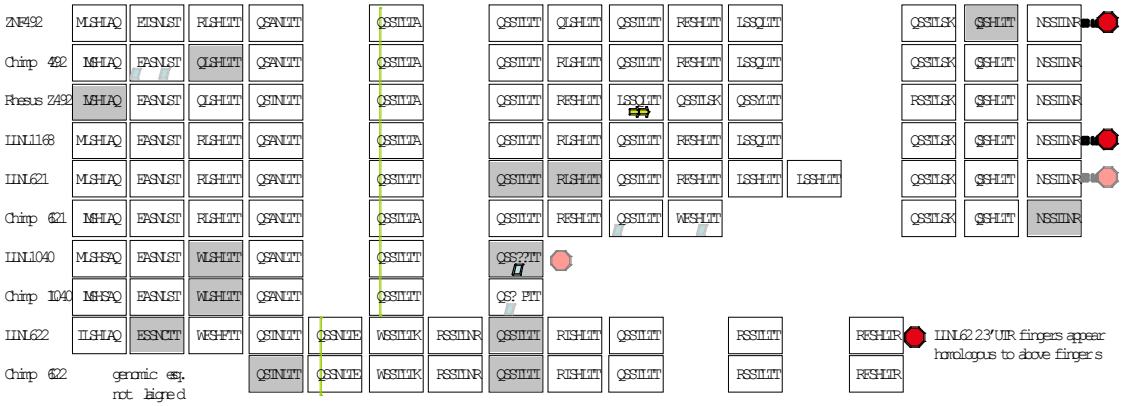

5'UTR
ATG
KRAB
KAP1-binding
Spacer
KZNF domain
DNA-binding
3'UTR
A
B
ZNF430
ZNF493
These fingers novel in human compared to chimp, rhesus
LLNL1008
LLNL744
Chimp may not have this stop codon
Chimp has lost this finger relative to human, rhesus

Figure 6



Frameshift in genomic sequence (not all confirmed for nonhumans)

Stop (based on human loci; faded for probable pseudogenes)

Indicates possible alternate alignment of Rhesus ZNF492 fingers 8,9 with human ZNF492 fingers 10,11 (leaving gaps)

Figure 7

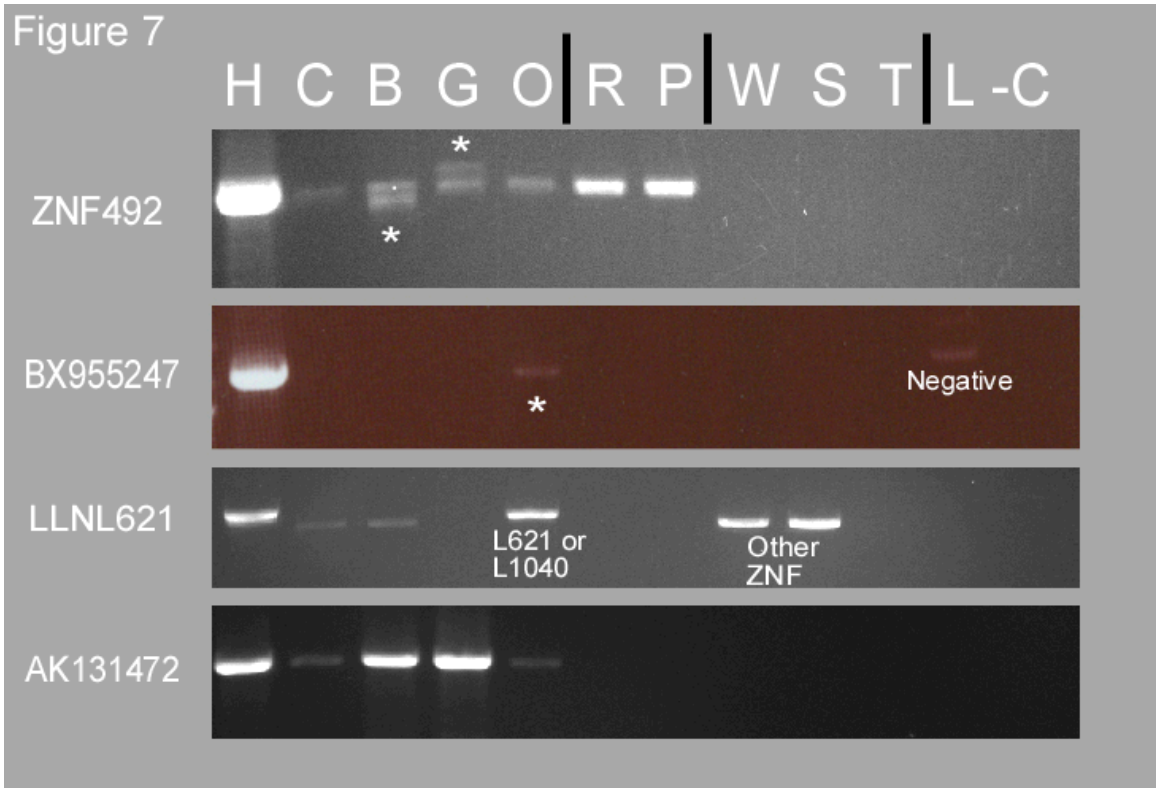

H C B G O R P W S T L -C
ZNF492
BX955247
LLNL621
AK131472
Negative
L621 or
L1040
Other
ZNF

Figure 8

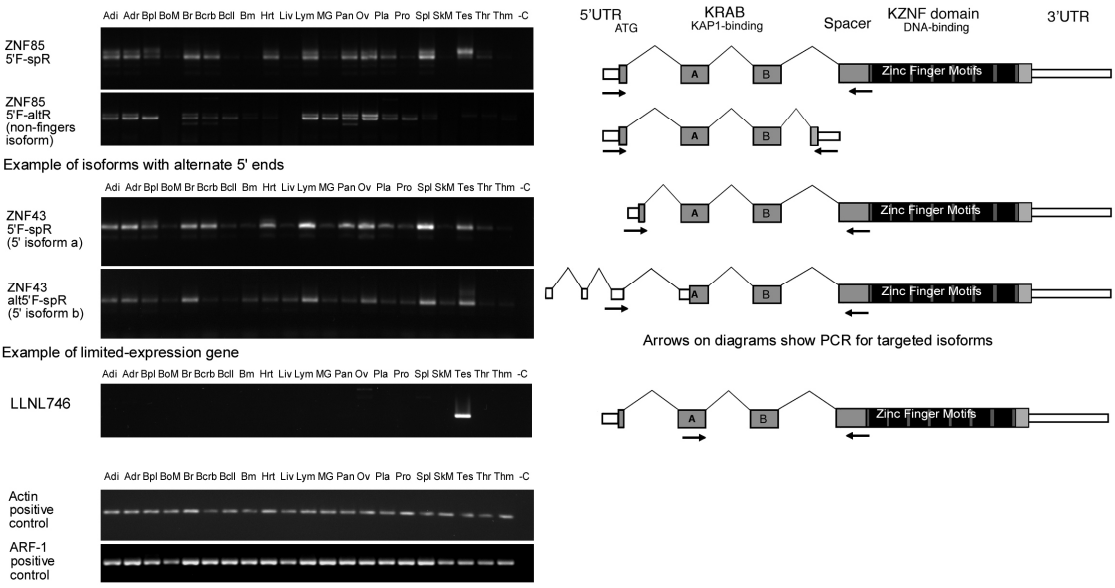

Example of isoform that loses the fingers exon
Adi Adr Bpl BoM Br Bcrb Bcll Bm Hrt Liv Lym MG Pan Ov Pla Pro Spl SkM Tes Thr Thm -C
ZNF85 5'F-spR
ZNF85 5'F-altR (non-fingers isoform)
5'UTR
ATG
KRAB
KAP1-binding
Spacer
KZNF domain
DNA-binding
3'UTR
A
B
Zinc Finger Motifs
Example of isoforms with alternate 5' ends
ZNF43 5'F-spR (5' isoform a)
ZNF43 alt5'F-spR (5' isoform b)
Arrows on diagrams show PCR for targeted isoforms
Example of limited-expression gene
LLNL746
Actin positive control
ARF-1 positive control

Table 1. List of zinc finger containing loci related to ZNF91 and in clusters containing ZNF91-subfamily genes

Gene name(s)	Chromosome	In a cluster?	Catalog status as of submission	
LLNL1120	1	yes	gene (putative)	
SBZF3 (ZNF695)	1	yes	gene (known)	
ZNF124	1	yes	gene (known)	NOT IN TREE (different subfamily)
ZNF496	1	yes	gene (known)	NOT IN TREE (different subfamily)
ZNF669	1	yes	gene (known)	NOT IN TREE (different subfamily)
ZNF670	1	yes	gene (known)	NOT IN TREE (different subfamily)
ZNF678	1		gene (known)	
LLNL1066	2		pseudogene (fragment)	
LLNL911	3		processed pseudogene	
FLJ90036	4	yes	gene (known, no KRAB)	
KIAA1982	4	yes	gene (known)	
LLNL1017	4	yes	gene (putative)	
LLNL1073	4	yes	pseudogene	
ZNF141	4	yes	gene (known)	
ZNF595	4	yes	gene (known)	
LLNL817	5		processed pseudogene	
FLJ39963	7	yes	gene (known)	NOT IN TREE
LLNL829	7	yes	pseudogene (fragment)	
LLNL831	7	yes	pseudogene (fragment)	NOT IN TREE
LLNL932	7	yes	pseudogene (fragment)	NOT IN TREE
LLNL933	7	yes	pseudogene (fragment)	
LLNL1029	7	yes	gene (putative)	
LLNL1030	7	yes	pseudogene (fragment)	
LLNL1031	7	yes	pseudogene (fragment)	
LLNL1032	7	yes	pseudogene	
LLNL1090	7	yes	pseudogene (fragment)	
LLNL1091	7	yes	pseudogene (fragment)	
LLNL1092	7	yes	gene (putative)	

LLNL1093	7	yes	gene (putative)	
LLNL1094	7	yes	pseudogene (fragment)	
LLNL1095	7	yes	gene (putative)	
LLNL1096	7	yes	gene (putative)	
LLNL1097	7	yes	gene (putative)	
LLNL1098	7	yes	gene (putative)	
LLNL1099	7	yes	pseudogene (fragment)	
ZNF92	7	yes	gene (known)	
ZNF117	7	yes	gene (known, no KRAB)	
ZNF138	7	yes	gene (known)	
ZNF273	7	yes	gene (known)	
ZNF479	7	yes	gene (known)	
ZNF588	7	yes	gene (known)	
ZNF679	7	yes	gene (known)	
ZNF680	7	yes	gene (known)	
LLNL1034	8		pseudogene (fragment)	
LLNL1104	8		pseudogene (fragment)	
LLNL1037	9		pseudogene	
LLNL1040	9		processed pseudogene	
LLNL1107	9		pseudogene (fragment)	NOT IN TREE
LLNL1108	9		pseudogene	
LLNL587	10		pseudogene	
ZNF195	11		gene (known)	
LLNL922	11		processed pseudogene	
LLNL960	12		processed pseudogene	
LLNL1124	13		pseudogene	
LLNL1125	14		pseudogene	
ZNF267	16		gene (known)	
LLNL550	16		gene (known)	
LLNL1128	16		pseudogene (fragment)	
ZNF519	18		gene (known)	
LLNL1053	18		pseudogene	
LLNL616	19	yes	pseudogene	
LLNL618	19	yes	gene (novel)	
LLNL621	19	yes	pseudogene (fragment)	
LLNL622	19	yes	gene (known)	
LLNL692	19	yes	gene (known)	
LLNL740	19	yes	gene (putative)	
LLNL742	19	yes	pseudogene	
LLNL743	19	yes	gene (putative)	
LLNL744	19	yes	gene (novel)	
LLNL745	19	yes	pseudogene	
LLNL746	19	yes	gene (putative)	
LLNL747	19	yes	gene (putative)	
LLNL748	19	yes	gene (putative)	
LLNL749	19	yes	gene (putative)	
LLNL750	19	yes	pseudogene (fragment)	
LLNL751	19	yes	gene (putative)	
LLNL874	19	yes	pseudogene	

LLNL792	19	yes	pseudogene (fragment)	NOT IN TREE
LLNL1008	19	yes	pseudogene	NOT IN TREE
LLNL1055	19	yes	pseudogene (fragment)	
LLNL1056	19	yes	pseudogene	
LLNL1166	19	yes	gene (putative)	
LLNL1167	19	yes	gene (putative)	
LLNL1168	19	yes	gene (putative)	
TIZ (ZNF675)	19	yes	gene (known)	
ZNF14	19	yes	gene (known)	NOT IN TREE (different subfamily)
ZNF15L1	19	yes	gene (known)	
ZNF43	19	yes	gene (known)	
ZNF66 (LLNL581)	19	yes	gene (novel)	
ZNF85	19	yes	gene (known)	
ZNF90	19	yes	gene (novel)	
ZNF91	19	yes	gene (known)	
ZNF100	19	yes	gene (known)	
ZNF101	19	yes	gene (known)	NOT IN TREE (different subfamily)
ZNF208	19	yes	gene (known)	
ZNF253	19	yes	gene (known)	
ZNF254	19	yes	gene (known)	
ZNF257	19	yes	gene (known)	
ZNF429	19	yes	gene (known)	
ZNF430	19	yes	gene (known)	
ZNF431	19	yes	gene (known)	
ZNF486 (LLNL614)	19	yes	gene (known)	
ZNF492 (LLNL620)	19	yes	gene (known)	
ZNF493	19	yes	gene (known)	
ZNF505	19	yes	gene (known)	
ZNF626	19	yes	gene (known)	
ZNF676 (LLNL348)	19	yes	gene (novel)	
ZNF681	19	yes	gene (known)	
ZNF682	19	yes	gene (known)	
LLNL1227	21		pseudogene	
LLNL1070	22		pseudogene	
LLNL1110	X		processed pseudogene	
LLNL854	Y	4 groups	pseudogene (fragment)	
LLNL855	Y	4 groups	pseudogene (fragment)	NOT IN TREE
LLNL1045	Y	4 groups	pseudogene (fragment)	
LLNL1046	Y	4 groups	pseudogene (fragment)	
LLNL1047	Y	4 groups	pseudogene (fragment)	
LLNL1048	Y	4 groups	pseudogene (fragment)	
LLNL1049	Y	4 groups	pseudogene (fragment)	
LLNL1050	Y	4 groups	pseudogene (fragment)	

LLNL1051	Y	4 groups	pseudogene (fragment)	
LLNL1114	Y	4 groups	pseudogene (fragment)	
LLNL1115	Y	4 groups	pseudogene (fragment)	
LLNL1116	Y	4 groups	pseudogene (fragment)	
LLNL1117	Y	4 groups	pseudogene (fragment)	
LLNL1118	Y	4 groups	pseudogene (fragment)	
LLNL1119	Y	4 groups	pseudogene (fragment)	

Table 2 (supplemental): Values for dN/dS (non-synonymous substitutions per non-synonymous site over synonymous substitutions pe
and their orthologs (if found) in chimpanzee and rhesus. '3 sites' refers to amino acid positions -1, 3, and 6 only; '6 sites' refers to posi

3 sites

	Chimp ZNF431	Human ZNF431	Rhesus ZNF431	Human LLNL618	Chimp LLNL618	Chimp ZNF100	Human ZNF100	Rhesus ZNF100	Human LLNL745	Rhesus ZNF430	Human ZNF430
ChimpZ431											
ZNF431	0										
RhesusZ431	**	0									
LLNL618	2.8194	1.84545	2.8194								
ChimpLLNL618	2.8194	1.84545	2.8194	0							
ChimpZ100	0.7185	0.71852	0.7185	2.037267	2.037267						
ZNF100	1.1395	1.13953	1.1395	3	3	0					
RhesusZ100	1.3061	1.30612	1.3061	2.753968	2.753968	0.23134	0.3605				
LLNL745	0.8606	0.86059	0.8606	1.229258	1.229258	0.53333	0.4841	0.5908			
RhesusZ430	0.336	0.336	0.336	2.506173	2.506173	1.11475	1.7564	3.61905	1.258865		
ZNF430	0.336	0.336	0.336	2.506173	2.506173	1.11475	1.7564	3.61905	1.258865	**	
ChimpZ430	0.336	0.336	0.336	2.506173	2.506173	1.11475	1.7564	3.61905	1.258865	**	**

6sites

	Chimp ZNF431	Human ZNF431	Rhesus ZNF431	Human LLNL618	Chimp LLNL618	Chimp ZNF100	Human ZNF100	Rhesus ZNF100	Human LLNL745	Rhesus ZNF430	Human ZNF430
ChimpZ431											
ZNF431	0.4375										
RhesusZ431	**	0.4375									
LLNL618	1.1026	1.00735	1.1026								
ChimpLLNL618	1.303	1.17094	1.303	0							
ChimpZ100	0.6667	0.75	0.6667	1.180645	1.386364						
ZNF100	0.8462	0.96154	0.8462	1.396947	1.663636	0					
RhesusZ100	0.6607	0.73451	0.6607	1.108571	1.293333	0.30769	0.4138				
LLNL745	0.8504	0.90254	0.8504	1.227612	1.227612	0.78788	0.7259	0.76636			
RhesusZ430	0.2846	0.33588	0.2846	0.933775	1.076336	1.1831	1.6154	1.60345	1.519084		
ZNF430	0.3364	0.3964	0.3364	0.815029	0.933775	0.93333	1.1831	1.19231	1.213415	0	
ChimpZ430	0.3333	0.39669	0.3333	0.797814	0.901235	0.77064	0.9438	1.10227	1.005051	0	0

whole fingers array

	Chimp ZNF431	Human ZNF431	Rhesus ZNF431	Human LLNL618	Chimp LLNL618	Chimp ZNF100	Human ZNF100	Rhesus ZNF100	Human LLNL745	Rhesus ZNF430	Human ZNF430
ChimpZ431											
ZNF431	0.1429										
RhesusZ431	0.1053	0.13043									
LLNL618	0.6038	0.61682	0.7333								
ChimpLLNL618	0.6176	0.62745	0.7442	0.333333							
ChimpZ100	0.3155	0.31609	0.3503	0.624204	0.631579						
ZNF100	0.3272	0.32738	0.3642	0.644737	0.657534	0					
RhesusZ100	0.3503	0.34356	0.3631	0.643312	0.655629	0.12048	0.1282				
LLNL745	0.6935	0.68235	0.8756	1.171569	1.171569	1.11268	1.1679	1.30833			
RhesusZ430	0.4173	0.41667	0.5	0.739726	0.751773	0.45968	0.479	0.36184	0.864583		
ZNF430	0.4677	0.46512	0.5701	0.765957	0.779412	0.50877	0.5321	0.40426	0.982659	0.13636	
ChimpZ430	0.4488	0.44697	0.5455	0.748252	0.76087	0.45968	0.479	0.37584	0.959538	0.03846	0.09091

r synonymous site) for pairwise comparisons of the zinc finger motif arrays of sets of paralogous loci
itions -1,1,2,3,5,6, while 'whole fingers array' includes all positions in each finger motif.

3 sites

	Human ZNF492	Chimp ZNF492	Rhesus ZNF492	Human LLNL1168	Human LLNL621	Chimp LLNL621	Human LLNL1040	Chimp LLNL1040	Human LLNL622
ZNF492									
Chimp Z492	**								
Rhesus Z492	1.70435	1.44348							
LLNL1168	**	**	1.58772						
LLNL621	1.16129	1.16129	1.20667	0.774194					
ChimpL621	0.72222	0.72222	1.26357	0.371429	**				
LLNL1040	0.54412	0.92647	1.08	0.746269	0.75758	1.39474			
ChimpL1040	0.71014	1.08696	1.12195	0.898551	1.08824	1	**		
LLNL622	0.67723	0.58182	0.75485	0.532982	0.62271	0.56693	1.350365	1.350365	
ChimpLLNL622	0.80108	0.71781	1.04658	0.717808	0.94144	0.93443	1.6753247	1.675325	0

6 sites

	Human ZNF492	Chimp ZNF492	Rhesus ZNF492	Human LLNL1168	Human LLNL621	Chimp LLNL621	Human LLNL1040	Chimp LLNL1040	Human LLNL622
ZNF492									
Chimp Z492	**								
Rhesus Z492	1.15929	0.9646							
LLNL1168	0.85714	0.42857	1.02655						
LLNL621	0.86207	0.65517	1.02655	0.428571					
ChimpL621	0.5625	0.41667	0.9537	0.270833	0.625				
LLNL1040	0.42105	0.5	0.82119	0.421053	0.31667	1.6875			
ChimpL1040	0.48718	0.57692	0.85806	0.487179	0.40984	1.17647	**		
LLNL622	0.60702	0.55667	0.63545	0.526846	0.55118	0.81013	1.3786408	1.358491	
ChimpLLNL622	0.54192	0.49849	0.70522	0.498489	0.6278	0.73171	0.8628571	0.861878	0

whole fingers array

	Human ZNF492	Chimp ZNF492	Rhesus ZNF492	Human LLNL1168	Human LLNL621	Chimp LLNL621	Human LLNL1040	Chimp LLNL1040	Human LLNL622
ZNF492									
Chimp Z492	1.28571								
Rhesus Z492	0.56701	0.51429							
LLNL1168	0.57143	0.35714	0.51						
LLNL621	0.93548	0.73684	0.84615	0.742857					
ChimpL621	0.5283	0.49057	0.63441	0.403509	0.63333				
LLNL1040	0.49275	0.54286	0.48	0.449275	0.56667	0.32292			
ChimpL1040	0.47436	0.51899	0.47015	0.435897	0.53623	0.32075	0.375		
LLNL622	0.48387	0.44255	0.46862	0.459091	0.55051	0.5567	0.6489362	0.631313	
ChimpLLNL622	0.35417	0.31538	0.37339	0.343096	0.46392	0.44681	0.3813953	0.381395	0.733333

Yellow highlight indicates comparisons that showed significant purifying selection

REFERENCES

Ayyanathan, K., M.S. Lechner, P. Bell, G.G. Maul, D.C. Schultz, Y. Yamada, K. Tanaka, K. Torigoe, and F.J. Rauscher, 3rd. 2003. Regulated recruitment of HP1 to a euchromatic gene induces mitotically heritable, epigenetic gene silencing: a mammalian cell culture model of gene variegation. *Genes Dev* **17:** 1855-1869.

Bailey, J.A., Z. Gu, R.A. Clark, K. Reinert, R.V. Samonte, S. Schwartz, M.D. Adams, E.W. Myers, P.W. Li, and E.E. Eichler. 2002. Recent segmental duplications in the human genome. *Science* **297:** 1003-1007.

Bellefroid, E.J., D.A. Poncelet, P.J. Lecocq, O. Revelant, and J.A. Martial. 1991. The evolutionarily conserved Kruppel-associated box domain defines a subfamily of eukaryotic multifingered proteins. *Proc Natl Acad Sci U S A* **88:** 3608-3612.

Bellefroid, E.J., J.C. Marine, T. Ried, P.J. Lecocq, M. Riviere, C. Amemiya, D.A. Poncelet, P.G. Coulie, P. de Jong, C. Szpirer, and et al. 1993. Clustered organization of homologous KRAB zinc-finger genes with enhanced expression in human T lymphoid cells. *Embo J* **12:** 1363-1374.

Bellefroid, E.J., J.C. Marine, A.G. Matera, C. Bourguignon, T. Desai, K.C. Healy, P. Bray-Ward, J.A. Martial, J.N. Ihle, and D.C. Ward. 1995. Emergence of the ZNF91 Kruppel-associated box-containing zinc finger gene family in the last common ancestor of anthropoidea. *Proc Natl Acad Sci U S A* **92:** 10757-10761.

Bertrand, S., F.G. Brunet, H. Escriva, G. Parmentier, V. Laudet, and M. Robinson-Rechavi. 2004. Evolutionary genomics of nuclear receptors: from twenty-five ancestral genes to derived endocrine systems. *Mol Biol Evol* **21:** 1923-1937.

Chen, Z.F., A.J. Paquette, and D.J. Anderson. 1998. NRSF/REST is required in vivo for repression of multiple neuronal target genes during embryogenesis. *Nat Genet* **20:** 136-142.

Chen, F.C. and W.H. Li. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am J Hum Genet* **68:** 444-456.

Cheng, Z., M. Ventura, X. She, P. Khaitovich, T. Graves, K. Osoegawa, D. Church, P. DeJong, R.K. Wilson, S. Paabo, M. Rocchi, and E.E. Eichler. 2005. A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* **437:** 88-93.

Chong, J.A., J. Tapia-Ramirez, S. Kim, J.J. Toledo-Aral, Y. Zheng, M.C. Boutros, Y.M. Altshuller, M.A. Frohman, S.D. Kraner, and G. Mandel. 1995. REST: a mammalian silencer protein that restricts sodium channel gene expression to neurons. *Cell* **80:** 949-957.

Choo, Y. and A. Klug. 1994. Toward a code for the interactions of zinc fingers with DNA: Selection of randomized fingers displayed on phage. *Proc Natl Acad Sci U S A* **91:** 11163-11167.

Dehal, P., P. Predki, A.S. Olsen, A. Kobayashi, P. Folta, S. Lucas, M. Land, A. Terry, C.L. Ecale Zhou, S. Rash, Q. Zhang, L. Gordon, J. Kim, C. Elkin, M.J. Pollard, P. Richardson, D. Rokhsar, E. Uberbacher, T. Hawkins, E. Branscomb, and L. Stubbs. 2001. Human chromosome 19 and related regions in mouse: conservative and lineage-specific evolution. *Science* **293:** 104-111.

Di Cristofano, A., M. Strazullo, L. Longo, and G. La Mantia. 1995. Characterization and genomic mapping of the ZNF80 locus: expression of this zinc-finger gene is

driven by a solitary LTR of ERV9 endogenous retroviral family. *Nucleic Acids Res* **23:** 2823-2830.

Eichler, E.E., S.M. Hoffman, A.A. Adamson, L.A. Gordon, P. McCready, J.E. Lamerdin, and H.W. Mohrenweiser. 1998. Complex beta-satellite repeat structures and the expansion of the zinc finger gene cluster in 19p12. *Genome Res* **8:** 791-808.

Enard, W., P. Khaitovich, J. Klose, S. Zollner, F. Heissig, P. Giavalisco, K. Nieselt-Struwe, E. Muchmore, A. Varki, R. Ravid, G.M. Doxiadis, R.E. Bontrop, and S. Paabo. 2002. Intra- and interspecific variation in primate gene expression patterns. *Science* **296:** 340-343.

Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39:** 783-791.

Fischer, A., Y. Gilad, O. Man, and S. Paabo. 2005. Evolution of bitter taste receptors in humans and apes. *Mol Biol Evol* **22:** 432-436.

Gilad, Y., O. Man, and G. Glusman. 2005. A comparison of the human and chimpanzee olfactory receptor gene repertoires. *Genome Res* **15:** 224-230.

Glazko, G.V. and M. Nei. 2003. Estimation of divergence times for major lineages of primate species. *Mol Biol Evol* **20:** 424-434.

Gonzalez-Lamuno, D., N. Loukili, M. Garcia-Fuentes, and T.M. Thomson. 2002. Expression and regulation of the transcriptional repressor ZNF43 in Ewing sarcoma cells. *Pediatr Pathol Mol Med* **21:** 531-540.

Graur, D.a.W.-H.L. 2000. *Fundamentals of Molecular Evolution*. Sinauer Associates, Sunderland, Massachusetts.

Greisman, H.A. and C.O. Pabo. 1997. A general strategy for selecting high-affinity zinc finger proteins for diverse DNA target sites. *Science* **275:** 657-661.

Hamilton, A.T., S. Huntley, J. Kim, E. Branscomb, and L. Stubbs. 2003. Lineage-specific expansion of KRAB zinc-finger transcription factor genes: implications for the evolution of vertebrate regulatory networks. *Cold Spring Harb Symp Quant Biol* **68:** 131-140.

Hao, L. and M. Nei. 2005. Rapid expansion of killer cell immunoglobulin-like receptor genes in primates and their coevolution with MHC Class I genes. *Gene* **347:** 149-159.

Huntley, S., D. Baggott, A.T. Hamilton, M. Tran-Gyamfi, S. Yang, J. Kim, L. Gordon, and L. Stubbs. submitted. A comprehensive catalogue of human KRAB-associated zinc finger genes: insights to the evolutionary history of a large family of transcriptional repressors.

Kent, W.J., Sugnet, C. W., Furey, T. S., Roskin, K.M., Pringle, T. H., Zahler, A. M., and Haussler, D. 2002. The Human Genome Browser at UCSC. *Genome Res* **12:** 996-1006.

Khaitovich, P., B. Muetzel, X. She, M. Lachmann, I. Hellmann, J. Dietzsch, S. Steigele, H.H. Do, G. Weiss, W. Enard, F. Heissig, T. Arendt, K. Nieselt-Struwe, E.E. Eichler, and S. Paabo. 2004. Regional patterns of gene expression in human and chimpanzee brains. *Genome Res* **14:** 1462-1473.

Khaitovich, P., I. Hellmann, W. Enard, K. Nowick, M. Leinweber, H. Franz, G. Weiss, M. Lachmann, and S. Paabo. 2005. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science* **309:** 1850-1854.

Kopelman, N.M., D. Lancet, and I. Yanai. 2005. Alternative splicing and gene

duplication are inversely correlated evolutionary mechanisms. *Nat Genet* **37:** 588-589.
Krebs, C.J., L.K. Larkins, S.M. Khan, and D.M. Robins. 2005. Expansion and diversification of KRAB zinc-finger genes within a cluster including Regulator of sex-limitation 1 and 2. *Genomics* **85:** 752-761.
Kumar, S., K. Tamura, I.B. Jakobsen, and M. Nei. 2001. MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics* **17:** 1244-1245.
Lane, R.P., J. Young, T. Newman, and B.J. Trask. 2004. Species specificity in rodent pheromone receptor repertoires. *Genome Res* **14:** 603-608.
Li, W.H., M. Tanimura, and P.M. Sharp. 1987. An evaluation of the molecular clock hypothesis using mammalian DNA sequences. *J Mol Evol* **25:** 330-342.
Li, W.-H. 1997. *Molecular Evolution*. Sinauer Associates,, Sunderland, Massachusetts.
Liu, G., S. Zhao, J.A. Bailey, S.C. Sahinalp, C. Alkan, E. Tuzun, E.D. Green, and E.E. Eichler. 2003. Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. *Genome Res* **13:** 358-368.
Liu, H., C. Zhu, J. Luo, Y. Wang, D. Li, Y. Li, J. Zhou, W. Yuan, Y. Ou, M. Liu, and X. Wu. 2004. ZNF411, a novel KRAB-containing zinc-finger protein, suppresses MAP kinase signaling pathway. *Biochem Biophys Res Commun* **320:** 45-53.
Looman, C., M. Abrink, C. Mark, and L. Hellman. 2002. KRAB zinc finger proteins: an analysis of the molecular mechanisms governing their increase in numbers and complexity during evolution. *Mol Biol Evol* **19:** 2118-2130.
Looman, C., L. Hellman, and M. Abrink. 2004. A novel Kruppel-Associated Box identified in a panel of mammalian zinc finger proteins. *Mamm Genome* **15:** 35-40.
Lovering, R. and J. Trowsdale. 1991. A gene encoding 22 highly related zinc fingers is expressed in lymphoid cell lines. *Nucleic Acids Res* **19:** 2921-2928.
Mark, C., M. Abrink, and L. Hellman. 1999. Comparative analysis of KRAB zinc finger proteins in rodents and man: evidence for several evolutionarily distinct subfamilies of KRAB zinc finger genes. *DNA Cell Biol* **18:** 381-396.
Mark, C., C. Looman, M. Abrink, and L. Hellman. 2001. Molecular cloning and preliminary functional analysis of two novel human KRAB zinc finger proteins, HKr18 and HKr19. *DNA Cell Biol* **20:** 275-286.
Nei, M. and T. Gojobori. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* **3:** 418-426.
Nei, M., X. Gu, and T. Sitnikova. 1997. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proc Natl Acad Sci U S A* **94:** 7799-7806.
Nelson, D.R. 2003. Comparison of P450s from human and fugu: 420 million years of vertebrate P450 evolution. *Arch Biochem Biophys* **409:** 18-24.
Nielsen, R., C. Bustamante, A.G. Clark, S. Glanowski, T.B. Sackton, M.J. Hubisz, A. Fledel-Alon, D.M. Tanenbaum, D. Civello, T.J. White, J.S. J, M.D. Adams, and M. Cargill. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol* **3:** e170.
Nishimura, T., T. Narita, E. Miyazaki, T. Ito, N. Nishimoto, K. Yoshizaki, J.A. Martial, E.J. Bellfroid, H. Vissing, and T. Taniyama. 2001. Characterization of the human

Fc gamma RIIB gene promoter: human zinc-finger proteins (ZNF140 and ZNF91) that bind to different regions function as transcription repressors. *Int Immunol* **13:** 1075-1084.

Noonan, J.P., J. Grimwood, J. Danke, J. Schmutz, M. Dickson, C.T. Amemiya, and R.M. Myers. 2004. Coelacanth genome sequence reveals the evolutionary history of vertebrate genes. *Genome Res* **14:** 2397-2405.

Ohno, S. 1970. *Evolution by Gene Duplication*. Springer-Verlag, Berlin, New York.

Poncelet, D.A., E.J. Bellefroid, P.V. Bastiaens, M.A. Demoitie, J.C. Marine, H. Pendeville, Y. Alami, N. Devos, P. Lecocq, T. Ogawa, M. Muller, and J.A. Martial. 1998. Functional analysis of ZNF85 KRAB zinc finger protein, a member of the highly homologous ZNF91 family. *DNA Cell Biol* **17:** 931-943.

Preuss, T.M., M. Caceres, M.C. Oldham, and D.H. Geschwind. 2004. Human brain evolution: insights from microarrays. *Nat Rev Genet* **5:** 850-860.

Rambaut, A. 1996. Se-Al: Sequence Alignment Editor.

Rozen, S., H. Skaletsky, J.D. Marszalek, P.J. Minx, H.S. Cordum, R.H. Waterston, R.K. Wilson, and D.C. Page. 2003. Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* **423:** 873-876.

Saitou, N. and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4:** 406-425.

Sambrook, J.G., A. Bashirova, S. Palmer, S. Sims, J. Trowsdale, L. Abi-Rached, P. Parham, M. Carrington, and S. Beck. 2005. Single haplotype analysis demonstrates rapid evolution of the killer immunoglobulin-like receptor (KIR) loci in primates. *Genome Res* **15:** 25-35.

Schmidt, D. and R. Durrett. 2004. Adaptive evolution drives the diversification of zinc-finger binding domains. *Mol Biol Evol* **21:** 2326-2339.

Shannon, M., J. Kim, L. Ashworth, E. Branscomb, and L. Stubbs. 1998. Tandem zinc-finger gene families in mammals: insights and unanswered questions. *DNA Seq* **8:** 303-315.

Shannon, M. and L. Stubbs. 1998. Analysis of homologous XRCC1-linked zinc-finger gene families in human and mouse: evidence for orthologous genes. *Genomics* **49:** 112-121.

Shannon, M., A.T. Hamilton, L. Gordon, E. Branscomb, and L. Stubbs. 2003. Differential expansion of zinc-finger transcription factor loci in homologous human and mouse gene clusters. *Genome Res* **13:** 1097-1110.

Shin, J.N., I. Kim, J.S. Lee, G.Y. Koh, Z.H. Lee, and H.H. Kim. 2002. A novel zinc finger protein that inhibits osteoclastogenesis and the function of tumor necrosis factor receptor-associated factor 6. *J Biol Chem* **277:** 8346-8353.

Shiu, S.H. and W.H. Li. 2004. Origins, lineage-specific expansions, and multiple losses of tyrosine kinases in eukaryotes. *Mol Biol Evol* **21:** 828-840.

Skaletsky, H., T. Kuroda-Kawaguchi, P.J. Minx, H.S. Cordum, L. Hillier, L.G. Brown, S. Repping, T. Pyntikova, J. Ali, T. Bieri, A. Chinwalla, A. Delehaunty, K. Delehaunty, H. Du, G. Fewell, L. Fulton, R. Fulton, T. Graves, S.F. Hou, P. Latrielle, S. Leonard, E. Mardis, R. Maupin, J. McPherson, T. Miner, W. Nash, C. Nguyen, P. Ozersky, K. Pepin, S. Rock, T. Rohlfing, K. Scott, B. Schultz, C. Strong, A. Tin-Wollam, S.P. Yang, R.H. Waterston, R.K. Wilson, S. Rozen, and D.C. Page. 2003. The male-specific region of the human Y chromosome is a

mosaic of discrete sequence classes. *Nature* **423:** 825-837.
Smit, A., Hubley, R & Green, P. 1996-2004. RepeatMasker Open-3.0.
Su, A.I., T. Wiltshire, S. Batalov, H. Lapp, K.A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M.P. Cooke, J.R. Walker, and J.B. Hogenesch. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* **101:** 6062-6067.
Swofford, D. 2002. PAUP* Phylogenetic analysis using parsimony (*and other methods). Sinauer Associates, Sunderland, MA.
Thompson, J.D., T.J. Gibson, F. Plewniak, F. Jeanmougin, and D.G. Higgins. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25:** 4876-4882.
Trowsdale, J., R. Barten, A. Haude, C.A. Stewart, S. Beck, and M.J. Wilson. 2001. The genomic context of natural killer receptor extended gene families. *Immunol Rev* **181:** 20-38.
Tunnacliffe, A., L. Liu, J.K. Moore, M.A. Leversha, M.S. Jackson, L. Papi, M.A. Ferguson-Smith, H.J. Thiesen, and B.A. Ponder. 1993. Duplicated KOX zinc finger gene clusters flank the centromere of human chromosome 10: evidence for a pericentric inversion during primate evolution. *Nucleic Acids Res* **21:** 1409-1417.
Vissing, H., W.K. Meyer, L. Aagaard, N. Tommerup, and H.J. Thiesen. 1995. Repression of transcriptional activity by heterologous KRAB domains present in zinc finger proteins. *FEBS Lett* **369:** 153-157.
Young, J.M., C. Friedman, E.M. Williams, J.A. Ross, L. Tonnes-Priddy, and B.J. Trask. 2002. Different evolutionary processes shaped the mouse and human olfactory receptor gene families. *Hum Mol Genet* **11:** 535-546.

WEBSITE REFERENCE

http://znf.llnl.gov:8080, the ZNF database home page