Supporting Information

Iskow et al. 10.1073/pnas.1205199109

SI Materials and Methods

Samples. Six chimpanzee and six rhesus macaque lymphoblastoid cell lines (LCLs) had their transcriptomes sequenced by RNAseq and analyzed as in ref. 1. The chimpanzee LCLs were obtained from Coriell under catalog numbers AG18358, AG18359, S003659, S003610, S003641, and S004973. The "AG" samples are no longer publically available. The rhesus macaque LCLs were obtained from the New England Primate Research Center at Harvard Medical School: Mm 153–99, Mm 150–99, Mm 173–02, Mm 256–95, Mm 265–95, and Mm 303–97. The six human LCLs were already sequenced by RNA sequencing (RNAseq), and the data are publically available (2). The samples chosen for further analysis were NA18502, NA18507, NA18517, NA18510, NA19238, and NA19239. These are all well-characterized HapMap samples from the Yoruba (i.e., YRI) population and are available for purchase through Coriell. Four samples from each species were run on our array comparative genomic hybridization (aCGH) platform (human, NA18517, NA19238, NA19239, NA18502; chimpanzee, S003659, AG18358, AG18359, S003641; rhesus macaque, Mm 265–95, Mm 256–95, Mm 173–02, Mm 303–97). One extra rhesus macaque sample was run on aCGH (Mm 118–99). In addition to the human, chimpanzee, and rhesus macaque samples that were run on the aCGH platform, two gorilla samples (PR00053 and PR001103) and two orangutan samples (PR00648 and PR00724) were also run.

Array Design and Bias. Agilent catalog probes for human (hg18) were downloaded from eArray and aligned against the chimpanzee (panTro2) and rhesus macaque (rheMac2) reference genomes by using BLAT (3). Those probes that had at least one 100% match to both reference genomes were similarity filtered based on the human reference genome (i.e., probes with only one perfect hit to hg18 were considered for the array; Dataset S1).

This array should not be considered as a genome-wide, unbiased copy number difference (CND) discovery array. As the probes were required to have 100% identity across human, chimpanzee, and rhesus macaque, the probes cluster together in regions of high conservation (Fig. S1*A*). We defined targeted clusters as regions with three or more probes with adjacent probes within 2,792 nt of each other (the 75th percentile of interprobe distance). As such, we were able to interrogate 93,514 targeted clusters on our array platform (Fig. S1 *B* and *C*). Such clusters are enriched for conserved genomic regions among primates (Fig. S1*D*). Additionally, we examined how much of the genome could be accessed by 10 or more consecutive probes each within 10 kb of each other, as these were regions which could be reliably called as CND by our algorithm. There are 22,246 accessible regions covering 31.6% of the human genome and with a median size of 33.3 kb. As some regions have hundreds of probes, multiple nonoverlapping CNDs can be found in a single accessible region. All arrays were hybridized by using standard Agilent protocols and were scanned with 3-μm resolution.

Analyzing Gorilla and Orangutan Samples on Array Platform. Given that our probes were chosen based on sequence conservation, the majority of probes represent the sequences of the last common ancestor of human, chimpanzee, and rhesus macaque, which lived 25 Mya. Sequences that have gone unchanged for such a long time are likely under evolutionary constraint. Although lineage-specific mutations can occur, we expected the majority of the array probes to have 100% identity in other great apes and Old World monkeys. To assess whether the probes on the array are also conserved among other great apes, we aligned the probe sequences to the orangutan (ponAbe2) and gorilla (gorGor1) reference genomes by using BLAT.

Despite having lower-quality reference genomes, the majority of probes could be aligned to ponAbe2 and gorGor1 (Fig. S2). For each probe, the highest scoring alignment was analyzed. Only 78% of the aligned probes had 100% identity and 100% coverage to the gorilla genome and only 62% for orangutan. This portion of probes was lower than we had expected, and we suspect that some of the mismatches in the alignments may, in fact, be a result of dips in quality in the reference genomes as opposed to true mismatches with the probe sequences (Fig. S2*C*). Regardless, we chose a very conservative approach when analyzing orangutan and gorilla (OG) samples on the array. First, when performing feature extraction on the array image files, we performed dye bias normalization by using only probes that had 100% identity in all five primate species examined, were autosomal, and were not known to overlap any primate copy number variants (CNVs)/CNDs. From these, we chose the 10,000 highest-scoring probes (Agilent scores their probes based on, e.g., Tm, GC content, hairpin ΔG, sequence complexity). Next, for the OG samples, we removed probes that did not have 100% identity or 100% coverage in their respective reference genomes. The resulting data, therefore, have a different probe spacing and effective resolution than the human, chimpanzee, and rhesus macaque samples (HCR), which is why we chose to analyze these species separately.

Creating CNV/CND Call Sets. As the array has a small medium probe spacing (Fig. S1*A*), it is more similar in design to targeted arrays than it is to unbiased discovery arrays. Unfortunately, commonly used CNV analysis software platforms do not have calling algorithms that can handle the bias and nuance of every possible custom array platform. Thus, we chose multiple strategies to look at CNDs among our samples. First, we used standard rank segmentation. Next, we used the initial calls from the rank segmentation approach to reduce the number of false negatives. Third, we used an independent calling method on targeted regions of interest. In addition, the human reference, NA10851, is male, whereas some samples were male and some were female. Female samples appear to have hundreds of gains along chromosome X relative to NA10851; therefore, we chose to disregard calls made on chromosomes X and Y.

Array data (Agilent Feature Extraction files) were imported into Nexus 5.0t and analyzed by using the Rank Segmentation algorithm. Specifically, the following parameters were used:

Significance threshold	1.0×10^{-9}
Maximum contiguous probe spacing, kb	100
Minimum number probes per segment	10
High gain	1.0
Gain	0.5
Loss	−0.4
Big loss	−1.0
Outliers to remove, %	5.0

By using these stringent criteria, and removing calls on chromosomes X and Y, the initial call set had 964 HCR CNDs and 274 OG CNDs. Subtle differences in the quality of array runs can alter the resolution. As we were interrogating conserved regions of the genome, we thought it unlikely that there would be recurrent, multiallelic CNDs within a species, as these are enriched in divergent regions (4). Thus, we merged overlapping calls within species and used the smallest CND as output, because it presumably has the most accurate breakpoints (Fig. S3). Next, these calls were merged across all samples with 50% reciprocal overlap to determine which CNDs were present across samples and species leading to 407 nonredundant CND regions for HCR and 153 for OG.

As the initial call set relied upon stringent criteria (e.g., the default setting is a three-probe minimum per segment, whereas we used a 10-probe minimum), we tried to reduce the number of false negatives. For each of these 407 HCR and 153 OG CND regions, the average $\log_2$ ratio in each sample was determined. As the initial call set was stringent, the criteria were lowered for samples for which a call was made in the same species. For example, if one of the 407 HCR CNDs was called as a gain in some but not all of the chimpanzee samples, the remaining chimpanzee samples were inspected for whether the average $\log_2$ ratio across this region was greater than 0.3. If so, those additional chimpanzee samples were also labeled as having the gain. For losses, a threshold of −0.3 was used. Using these criteria, 49% of the 407 HCR regions and 85% of 153 OG regions were present at a frequency of one in at least one species and were considered fixed CNDs in those species. The higher percentage of fixed CNDs in OG samples is likely a result of fewer samples from each species being examined.

For targeted calls in HCR samples, we generated nonredundant lists of regions of interest (excluding regions on chromosome X and Y). For transcription factors (TFs), this list was based on Gene Ontology classifications and was downloaded from PANTHER (5). For exons, human GenCode gene annotations were used. For genes that overlapped, the longest possible transcript was used to create a nonredundant list of exons. For long intergenic noncoding RNAs (lincRNAs), genomic locations were obtained from Cabili et al. (6) and duplicate lincRNAs were removed. The ultraconserved elements (UCEs) used in this study are a merged set from human/mouse/rat, human/mouse/dog, and human/chicken comparisons (7, 8). By definition, UCEs are conserved over long evolutionary distances and are therefore enriched for probes on our array platform. In fact, >92.5% of autosomal UCEs are tagged by at least one probe (Fig. S4).

By using the hg18 genomic coordinates for these lists of targeted regions, we assessed those overlapped by at least three probes (including probes that only partially overlap the region of interest). We then performed a one-sided paired *t* test for the mean $\log_2$ ratio within the region being consistent with 0. For TFs and exons, we used a *P* value of ≤0.05 and a Bonferroni correction. Next, we required that the mean $\log_2$ within the region is more than 0.5 for duplications and less than −0.9 for deletions. Such requirements are stringent, and these analyses likely have many false negatives; however, those regions identified as CNDs are more likely to have multiple copy gains/losses among our samples. For lincRNAs and UCEs, we required a *P* value of ≤0.02 and absolute $\log_2$ ratio cutoffs of ±0.5.

Examining Reference Effects. Twenty-three CNDs were found in both chimpanzee and rhesus macaque samples. It is possible that CNDs independently occurred in these species; however, they could also result from array reference effects. Deviations from a copy number of 2 in the array reference sample (NA10851) can give the appearance of CNDs in our samples. To test for reference effects, we examined whether CNDs found in both chimpanzee and rhesus macaque also overlapped with gorilla and orangutan CNDs. In addition, the array reference has been sequenced to >40 fold coverage. Thus, read depth *z*-scores can be used to estimate whether NA10851 copy number deviates from two at specific sites. *z*-scores were created for 1-kb bins across the genome. The median *z*-score was then pulled for bins overlapping each HCR CND. Those regions with *z*-scores greater than 2.5 or less than −2.5 were considered potential reference effects and are indicated in Dataset S2.

RNAseq Analysis. RNA was extracted from LCLs by using an RNeasy Mini Kit (Qiagen). Illumina sequencing libraries were created from cDNA synthesized from polyadenylated RNA as described previously (9). RNAseq read mapping, normalization, and gene expression analyses were conducted as previously described (1). Briefly, reads were mapped to the human (hg18), chimpanzee (panTro2), and rhesus macaque (rheMac2) reference genomes by using MAQ version 0.6.8 (10) with default parameters. To compare gene expression levels across species, we only used reads mapping to exons for which clear orthologues exist across human, chimpanzee, and rhesus macaque (as identified and described in ref. 1). To obtain a measure of gene expression levels, we used the sum of the number of reads mapping to all of the orthologous exons within a gene, excluding reads that (*i*) did not map to any of the three-species orthologous exons, (*ii*) had a MAQ quality score lower than 10, and (*iii*) mapped to multiple locations in the genome. To account for different overall numbers of reads in each lane, we divided by the total number of reads mapped to genes in that lane.

To identify differentially expressed (DE) genes, we used the same statistical framework and model outlined by Blekhman et al. (1). Specifically, this framework extends the Poisson mixed-effects model described by Marioni et al. (9) to model the number of reads mapping to each gene. We evaluated differential gene expression in a genewise manner, testing for significant differences between each pair of species separately. For each pairwise analysis, we tested a null model assuming no difference in gene expression between species (while accounting for gene length and the total number of reads in each lane) against the alternative model assuming that the expression of a gene differs between the two species. Evidence for significant differences in expression between the species was determined by using *P* values based on the likelihood ratio (LR) statistic. To correct for multiple testing, we calculated a false discovery rate by using the approach of Storey and Tibshirani (11).

Please note that ambiguous reads (i.e., those mapping equally to multiple sites) are discarded. As such, we are essentially biasing away from detecting increased expression in regions that have been recently duplicated. Such regions are possible CNDs. Therefore, the expression differences observed for *KANK1* (Fig. 3*A*) and *ZNF669* (Fig. 4*C*) are very conservative, and likely underestimate the true expression level differences in LCLs. Likewise, the general effect of CND genes on differential gene expression (Fig. 2) is likely underestimated.

Validations. We attempted quantitative PCR (qPCR) validation on 17 CNDs by using ABI TaqMan copy number assays and SYBR Green, and were able to confirm 94% in at least one sample from the same species. As a result of TaqMan assay locations being proprietary, we cannot ensure that the sequences of the primers are conserved in nonhuman primates. We do know, however, some general information about their conservation. One of the TaqMan assays does not map to the chimpanzee reference genome (Hs02999435_cn). Three TaqMan assays mapped to the chimpanzee reference genome more than once with perfect identity (Hs06432529_cn, Hs04657205_cn, and Hs04315747_cn). Two assays do not map to the orangutan genome (Hs01623634_cn and Hs03351761_cn). Validation information can be found in Dataset S4. Sequence mismatches between the TaqMan probes and nonhuman primate DNA may account for some of the invalidated CNDs. Copy number counts were determined by using CopyCaller software.

In addition, we used NanoString Copy Count technology to verify the gain of *ZNF669* in rhesus macaques following NanoString's recommended protocol. We also performed SYBR Green qPCR for the CND overlapping the *KANK1* gene. Standard qPCR protocols were used. In brief, qPCR reactions were performed in triplicate. A water template negative control was used for each primer set. NA10851 was the reference sample because it was also the reference for aCGH experiments. Serial dilutions of NA10851 were used to create a standard curve for each primer set. Primers were specifically designed to amplify across primate species with minimal bias. Control primer sequences (i.e., prUCE primers) were derived from an UCE that has no evidence of copy number variation among primates. Primer sequences are as follows:

KANK1 forward	TGTGGATGACATACAGAAGGGAAA
KANK1 reverse	GTCTCCATCAGTGTCTTGGTGACAT
prUCE forward	CTCGCTGAGCCCCTTCTCTAA
prUCE reverse	AAAAGGCGATTGTCTGGAGTCTC

After the initial qPCR depicted in Fig. 3*B*, we performed additional qPCRs on an independent set of primate samples (13 chimpanzees, 10 bonobos, 9 black-tufted marmosets, 5 orangutans, 3 hamadryas baboons, 4 olive baboons, 2 ring-tailed lemurs, and 10 gorillas). From these experiments, we confirmed the gain of *KANK1* in all chimpanzees and in one of the bonobo samples. We did not, however, confirm a gain of *KANK1* in the two additional ring-tailed lemur samples.

Determining the Fraction of Genes Affected by CNDs. Based on the array platform and CND calling algorithm parameters described here, we determined the subset of the ~14,000 genes that were analyzed by RNAseq that could be assessed by our array platform. We defined accessible genes as those overlapped by at least three probes. We also examined only those genes for which RNAseq analysis was possible for each pairwise comparison between the three species. By using this set of genes (n = 10,837), we determined the number of genes that were DE between human and chimpanzee (HC), between human and rhesus macaque (HR), or between chimpanzee and rhesus macaque (CR) by using a false discovery rate of ≤1%. Next, we assessed whether these DE and array-accessible genes overlapped CNDs (e.g., when examining HC DE genes, we determined the fraction of these that overlapped either a human CND or a chimpanzee CND). For HC, HR, and CR genes, the percent overlapping relevant CNDs was 2.6%, 3.2%, and 3.1%, respectively, and many of these genes were DE in liver and LCLs (Fig. S5). We also examined those genes found to be DE between pairs of species based on RNAseq in liver (1). For this analysis, we found that 2.2%, 3.1%, and 3.1% of HC, HR, and CR DE genes overlapped with CNDs, respectively. We performed the same analysis looking only at genes with known and predicted TF function based on PANTHER classification (n = 1,258). In doing so, the portions of genes DE in HC, HR, and CR that could possibly be explained by overlapping CNDs were 3.6%, 3.4%, and 4.7%, respectively, in LCLs, and 3.2%, 2.4%, and 4.5%, respectively, in liver tissue.

Determining Enrichment Using PANTHER. When determining the enrichment for genes of a particular function among the CND genes, it was important to ensure an appropriate background set was used. Several different minimum probe cutoffs were used to determine whether a gene could be considered "covered" by the array. Note that, although a 10-probe minimum was used in rank segmentation CND calling, probes do not have to be within the gene to be part of a CND overlapping a gene. Thus, as the number of probes required to overlap genes increases, the analysis becomes more conservative. Regardless of the background set used for enrichment analysis, TFs, and other gene regulators, are among the most enriched (see display table below).

TFs often have multiple paralogues within a given genome, so we were interested in exploring whether other paralogous gene families were overlapped by CNDs. In general, highly identical paralogues in humans will be depleted for probes on our array platform, as we required probe sequences to be unique in hg18. Likewise, paralogous genes that are highly divergent between species (e.g., APOBEC3 or TRIM genes), will also be depleted for probes because of the lack of conservation between human, chimpanzee, and rhesus macaques. In addition to examining enrichment based on molecular function in PANTHER, we used the eggNOG orthologous gene database to examine genes with known paralogues in primates (12). By using the "Primates nonsupervised orthologous groups and their proteins" data set available on the eggNOG Web site, we filtered for those with orthologues in the species used in this study. Afterwards, 774 genes with known paralogues and RefSeq IDs remained. Of these, 239 (30.9%) were covered by at least three probes on the array platform and 18 were at least partially overlapped by CND calls (coordinates are in hg18; see display table on following page). As previously reported, olfactory receptor genes were overlapped by CNDs (*OR2T33* and *OR8B2*) (13). Olfactory receptors have extreme levels of copy number variation in humans and nonhuman primates, presumably because of an initial rapid gene expansion followed by genomic drift (13–15).

ZNF669. When determining the number of RNAseq reads that mapped to *ZNF669* in human, chimpanzee, and rhesus macaque, we initially used gene annotations to evaluate which reads to count as mapping. The human reference genome has one copy

Probes required to overlap gene	1	3	5	10
Genes in ref set	14,710	13,763	12,940	1,108
Genes in test set (≥1 bp overlap with CND)	462	444	430	384
Top hit	Nucleic acid binding (P = 0.000000039)	Nucleic acid binding (P = 0.0000000168)	Nucleic acid binding (P = 0.00000000737)	Nucleic acid binding (P = 0.0000000104)
Second hit	DNA binding (P = 0.0000304)	DNA binding (P = 0.000022)	Binding (P = 0.0000233)	RNA binding (P = 0.0000151)
Third hit	TF activity (P = 0.0000523)	Binding (P = 0.0000249)	DNA binding (P = 0.0000235)	DNA binding (P = 0.000055)
Fourth hit	Transcription regulator activity (P = 0.0000523)	TF activity (P = 0.000053)	TF activity (P = 0.0000668)	Binding (P = 0.000061)
Fifth hit	RNA binding (P = 0.0000583)	Transcription regulator activity (P = 0.000053)	Transcription regulator activity (P = 0.0000668)	TF activity (P = 0.000221)

Chromosome	Start	Stop	Paralogous gene	Fraction covered by CNDs
chr15	39918302	39927638	*PLA2G4B*	1
chr1	227473501	227507141	*RAB4A*	0.2
chr4	254463	279944	*ZNF732*	1
chr5	138920934	138988202	*UBE2D2*	0.204
chr10	131523485	131652081	*EBF3*	0.176
chr13	52089605	52115920	*HNRNPA1L2*	0.697
chr11	123757507	123758449	*OR8B2*	1
chr1	15675182	15690482	*CELA2B*	0.488
chr11	133651484	133694668	*GLB1L3*	1
chr1	246502776	246503739	*OR2T33*	1
chr2	176680329	176682562	*HOXD11*	1
chr3	114948555	115013595	*ATP6V1A*	0.493
chr1	89602023	89626307	*GBP6*	1
chr22	22363047	22371363	*RGL4*	0.762
chr2	138438277	138490404	*HNMT*	1
chr3	140545550	140558577	*MRPS22*	0.984
chr6	160131481	160139451	*MRPL18*	1
chr16	28298400	28322663	*EIF3CL*	1

of ZNF669 (chr1:245329927–245334251, hg18). The chimpanzee reference has four copies (chr1:228218708–228222643, chr19_random:134710–273219, chr5:182772677–182877441, chr8:143218075–143881301, panTro2). The rhesus macaque reference has nine copies (chr11:52648978–52649807, chr11:8275051–8277070, chr12:66918153–66941468, chr13:136794067–136800374, chr13:137462627–137468939, chr20:3327012–3511474, chr5:4553128–4554808, chr6:93501415–93507381, chr9:37910816–38015622, rheMac2). Several of these genes, however, are divergent from both the human *ZNF669* and from each other within each species. Thus, it is unlikely that these annotated reference genome genes are responsible for the increased copy number we observed by using aCGH.

For a more accurate determination of the copy number of *ZNF669* in the reference genomes of chimpanzee and rhesus macaques, the human *ZNF669* predicted mRNA was aligned to the other genomes by BLAT. The predicted mRNA of the best hit was then used as a query in BLAT for each genome. Those BLAT alignments with at least 90% identity and span were considered as being copies of *ZNF669* present in the reference genome. Thus, based on our analysis, there is one copy of *ZNF669* in the chimpanzee reference genome (chr1:228218640–228222470, panTro2) and 16 copies in the rhesus macaque reference genome (chr1:209876777–209882841, chr11:8275054–8277068, chr11:11380435–11382002, chr11:52648843–52649847, chr13:70622517–70627550, chr13:136794229–136799997, chr13:137462983–137468766, chr16:44866620–44867624, chr18:73536331–73542364, chr2:97054946–97067156, chr3:119351092–119356890, chr4:154309940–154310930, chr5:4553130–4554809, chr5:84097917–84098921, chr6:93501585–93507379, chr8:8623544–8625138, rheMac2). These coordinates were used to determine the read depth of *ZNF669* for Fig. 4*C*.

For examining whether the additional *ZNF669* copies in rhesus macaque are diverging from each other and potentially under positive selection, we used the LR analysis from PAML (16, 17). The LR analysis compares neutral selection (i.e., M1) and positive selection (i.e., M2). M1 allows two ω-site classes to be estimated from the data, $\omega0 < 1$ or $\omega1 = 1$. The ω parameter indicates the underlying nonsynonymous/synonymous rate ratio. M2 allows an additional ω-site class value to be estimated from the data, $\omega2 > 1$. When the LR analysis suggests that positive selection (>1) has occurred at any of the loci of interest, selected sites are further analyzed under the M2 model with the Bayesian approach ($P > 99\%$). A phylogenetic tree for *ZNF669* was constructed using the neighbor-joining (NJ) method implemented in MEGA version 5. The stability of internal nodes was assessed by bootstrap analysis with 1,000 replicates (Fig. S6).

Analyzing Pseudogenes. ***Determining pseudogenes from log$_2$ ratios.*** The average $\log_2$ ratios across exons was determined as described earlier for a nonoverlapping list of Ensembl genes. For those genes that had at least one exon gained, we pulled the average $\log_2$ ratios across all of the introns. As the array is biased toward conserved regions, there were very few probes in individual introns; thus, the average across all introns for a given gene was needed. We then created a "pseudogene score" for each gene in each sample with a gained exon (Fig. S7*A*). The average $\log_2$ ratio for the introns was subtracted from the average $\log_2$ ratio for the gained exon(s). The resulting value was then divided by the average $\log_2$ ratio for the gained exon(s). If the introns have $\log_2$ ratios near zero, indicating that at least 1 exon was gained whereas introns were not, we expect a pseudogene score near 1. After visual inspection by using the Integrative Genomics Viewer and examination of processed pseudogenes present in reference genomes, a cutoff of ≥0.7 was used to define processed pseudogenes.

Determining enrichment for miRNA binding sites within pseudogenes. A list of predicted miRNA binding sites was downloaded from the TargetScan miRNA Regulatory Sites track in the University of California, Santa Cruz (UCSC),Genome Browser for hg18 (3, 18). The number of miRNA binding sites in processed pseudogenes was compared with the number of "accessible" genes with miRNA binding sites. Accessible genes were defined as having at least one exon overlapped by at least three probes. By using accessible genes as a background set, we performed a χ^2 test with Yates continuity correction to determine whether processed pseudogenes had proportionately more miRNA binding sites (Fig. 5*C*). For enrichment of specific miRNA binding sites among processed pseudogenes, analysis was performed with GREAT (19) by using the same background of accessible genes. The seven miRNAs whose binding sites were enriched among processed pseudogenes had false discovery rates of less than 5% (Dataset S9).

Determining identity of pseudogene with source gene. By using pseudogenes we identified by aCGH, we determined the sequence identity shared with their source genes. We used Ensembl gene annotations to pull just the coding sequence from the UCSC Genome Browser. As this analysis relies on proper annotation in the reference genome, some pseudogenes were omitted from this analysis if they had no annotated gene or if the processed pseudogene was annotated as the only version of the gene. We also omitted pseudogenes that were present in multiple species and those present in gorilla and orangutan. We then used BLAT to align the coding sequence of the source gene to its respective reference genome. A pseudogene was defined as being present in the reference genome if it was at least 100 bp long, had >87% identity to the BLAT query, and was missing introns. The percent identity of the pseudogene was obtained from the BLAT search results and used in Fig. S7*C*.

Determining expression of pseudogenes by RNAseq. We aligned the RNAseq reads to the chimpanzee and rhesus macaque reference genomes in conjunction with a splice-junction library based on the Ensembl gene predictions for each genome. The alignment step was performed by using Bowtie (20) with parameters -v 2 -m 1 –best –strata, and gene expression values were calculated by using RSEQtools (21). For our analysis, we omitted uninformative reads, those that map to gene–pseudogene pairs equally well. The informative reads, those that map uniquely to the pseudogenes, indicate possible expression of that pseudogene. Altogether, six chimpanzee and eight rhesus macaque processed pseudogenes appear to be expressed (chimpanzee, chr9:87174717–87176925, chr12:20287033–20290231, chr1:113222655–113226403, chr7: 123648134–123651355, chr1:229704643–229706370, chr6: 58092602–58094923; rhesus, chr6:75565933–75567781, chr2: 138453155–138456445, chrX:110376088–110378151, chr16: 57445181–57447220, chr6:136334200–136335883, chr7: 95598206–95600045, chr18:55546985–55548855, chr5:117053700–117055696, Fig. S7).

UCEs. UCEs were defined previously (7, 8). CND UCEs were defined as being entirely overlapped by a CND call made by using standard rank segmentation methods or the targeted method described earlier (Dataset S10). Next, the overlap of CND UCEs with DE genes was determined. Also, if a CND UCE did not overlap a DE gene, the nearest DE gene was determined. A UCE in an alternatively spliced 3′ exon of the *SYNCRIP* gene co-occurs with reduced expressed of *SYNCRIP* and the nearby gene, *SNX14* (Fig. S8).

lincRNAs. Based on our initial call set using the Nexus software, 133 lincRNAs were >50% covered by CNDs. As we did with TFs, we pulled the average $\log_2$ ratio across lincRNAs to assess for deviations from zero. Of the 7,185 autosomal lincRNAs, we could assess 2,277 (32%) with our stringent CND calling method. Of these, the vast majority were neutral in copy number (i.e., had the same copy number as the human reference). When combining the targeted calls with the Nexus calls, we had 200 CND lincRNAs in human, chimpanzee, and/or rhesus macaque (Dataset S11).

For studying the expression of lincRNAs in human, chimpanzee, and rhesus macaque, we started with the lincRNA exons as defined in supplemental table 1 of Cabili et al. (6) which were in hg19 coordinates and used the Galaxy LiftOver tool to get the hg18, panTro2, and rheMac2 coordinates. Those exons that could not be lifted over to all three genomes were removed from the analysis. From these, we mapped reads by using MAQ (10). Ambiguous mappings were discarded. The number of mapped reads per lincRNA was normalized by the total number of reads per sample and by the length of the exon in a given species (Fig. S8 *D* and *E*).

SI Discussion

It should be noted here that, since speciation, individual nucleotide changes, indels, mobile element insertions, and other genomic changes have occurred on each primate lineage, and many mechanisms of dosage compensation and regulation can act together to fine-tune expression (22). As such, it is difficult to tease apart the role of single CNDs on gene expression. Furthermore, although a gene (or a regulatory region) may be encompassed by a CND, this does not necessarily mean it will be DE. For example, *HYDIN* is a gene that is duplicated in humans (23). The original gene is expressed in many tissues, however, the paralogue is expressed solely in the brain (24). For this study, DNA and RNA were both derived from LCLs. Thus, even though we can detect the *HYDIN* duplication with our aCGH platform, we were not able to detect differential gene expression between human and other species for this gene.

The number of CNDs we observed in each species increases with the evolutionary distance from the human reference (Fig. 1*A*). This trend indicates that CNDs accumulate relatively consistently over time. As we interrogated only conserved regions across primates, and thus those likely to have function, and because we required the presence of the probe sequences in the human, chimpanzee, and rhesus macaque reference genomes, we expected that the majority of CNDs would be duplications relative to the human reference. Of the 407 nonredundant CNDs from our initial call set, however, there were 241 (59%) gains and 156 (41%) losses (Dataset S2). We found several examples of duplicated genes that were not present or were misannotated in their reference genomes. For example, we predict that *ZNF669* is ∼50 copies in all five rhesus macaques that we examined; however, there are only nine annotated copies of *ZNF669* in the rhesus macaque reference. Three of these nine copies have low sequence identity with the human *ZNF669* (<85%) and likely do not contribute to the increased hybridization on our array platform. Based on sequence identity (using BLAT on the human mRNA and looking for hits with >90% identity), there are 16 copies of *ZNF669* in the rhesus macaque reference. In general, highly homologous genes are difficult to distinguish from one another and, therefore, are difficult to unambiguously assemble into contigs. As such, duplications are often condensed during genome assembly (25, 26). From this study, we were able to identify gene duplications that appear to be fixed or high-frequency based on array results, but one or more copies are missing in its reference genome.

In addition to the orthologue to human *KANK1*, the chimpanzee reference sequence has evidence of at least one paralogous *KANK1* on unincorporated contigs. Based on these contigs, the chimpanzee specific duplication containing *KANK1* has >97% identity with its paralogue, whereas the coding sequence has 99% identity and has an intact ORF. In addition, LR analysis in PAML indicates that both copies of *KANK1* are under purifying selection in chimpanzee. This finding indicates that both the original gene and its duplication have retained protein coding function.

Copy number-different genes that are also DE between primates tend to be highly expressed in gonads. By using microarray expression data across 32 tissues with three replicates for each tissue (27), we assessed the tissue specificity for ∼90% of the genes from Dataset S6. Most of these genes are expressed in many tissues and all are expressed to some degree in bone marrow, which explains why we were able to detect their expression in LCLs. Several of the genes show peaks of expression for ovaries and/or testes (Fig. S9).

1. Blekhman R, Marioni JC, Zumbo P, Stephens M, Gilad Y (2010) Sex-specific and lineage-specific alternative splicing in primates. *Genome Res* 20:180–189.
2. Pickrell JK, et al. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464:768–772.
3. Kent WJ, et al. (2002) The human genome browser at UCSC. *Genome Res* 12:996–1006.
4. Gokcumen O, et al. (2011) Refinement of primate copy number variation hotspots identifies candidate genomic regions evolving under positive selection. *Genome Biol* 12:R52.

5. Thomas PD, et al. (2003) PANTHER: A library of protein families and subfamilies indexed by function. *Genome Res* 13:2129–2141.
6. Cabili MN, et al. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* 25:1915–1927.
7. Derti A, Roth FP, Church GM, Wu CT (2006) Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants. *Nat Genet* 38:1216–1220.
8. Bejerano G, et al. (2004) Ultraconserved elements in the human genome. *Science* 304: 1321–1325.
9. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* 18:1509–1517.
10. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 18:1851–1858.
11. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* 100:9440–9445.
12. Powell S, et al. (2012) eggNOG v3.0: Orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res* 40(database issue):D284–D289.
13. Gilad Y, Przeworski M, Lancet D, Lancet D, Pääbo S (2004) Loss of olfactory receptor genes coincides with the acquisition of full trichromatic vision in primates. *PLoS Biol* 2:E5.
14. Hasin Y, et al. (2008) High-resolution copy-number variation map reflects human olfactory receptor diversity and evolution. *PLoS Genet* 4:e1000249.
15. Nozawa M, Kawahara Y, Nei M (2007) Genomic drift and copy number variation of sensory receptor genes in humans. *Proc Natl Acad Sci USA* 104:20421–20426.
16. Yang Z, Nielsen R, Goldman N, Pedersen AM (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155:431–449.
17. Yang Z (1997) PAML: A program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 13:555–556.
18. Lewis BP, Burge CB, Bartel DP (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 120:15–20.
19. McLean CY, et al. (2010) GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* 28:1630–1639.
20. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10:R25.
21. Habegger L, et al. (2011) RSEQtools: A modular framework to analyze RNA-Seq data using compact, anonymized data summaries. *Bioinformatics* 27:281–283.
22. Ying-Fei Chang A, Liao B-Y (2011) DNA methylation rebalances gene dosage after mammalian gene duplications. *Mol Biol Evol* 29:133–144.
23. Doggett NA, et al. (2006) A 360-kb interchromosomal duplication of the human HYDIN locus. *Genomics* 88:762–771.
24. Brunetti-Pierri N, et al. (2008) Recurrent reciprocal 1q21.1 deletions and duplications associated with microcephaly or macrocephaly and developmental and behavioral abnormalities. *Nat Genet* 40:1466–1471.
25. Quinlan AR, et al. (2010) Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res* 20:623–635.
26. Sudmant PH, et al.; 1000 Genomes Project (2010) Diversity of human copy number variation and multicopy genes. *Science* 330:641–646.
27. Dezso Z, et al. (2008) A comprehensive functional analysis of tissue specificity of human gene expression. *BMC Biol* 6:49.

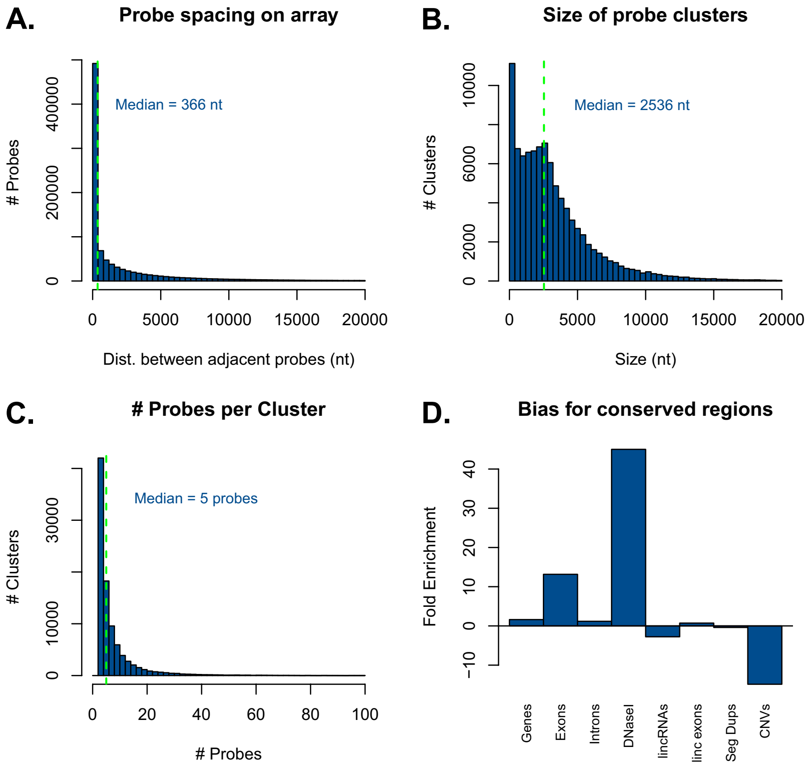

Fig. S1. The array comprises probes with 100% identity in human, chimpanzee, and rhesus macaque. (*A*) Histogram of probe spacing. Probes with 100% sequence identity between human, chimpanzee, and rhesus macaque reference genomes were used to populate the array. As a result, the probes cluster together in regions with high conservation. Shown here is a histogram of the distance between adjacent probes. There is a clear skew toward small distances with a median probe spacing of 366 nt, indicating how probes tend to be near other probes. (*B*) Histogram of cluster size. Clusters were defined as having a three-probe minimum. By using a cutoff of 2,792 nt between adjacent probes (the 75th percentile for interprobe distance), probes were placed into clusters (n = 93,514). Clusters have a median size of 2,536 nt. (*C*) Histogram of the number of probes per cluster. Clusters contained as many as 1,333 probes and a median of five probes. A subset of the data are shown for better visualization (the tail on the right side of the curve continues until 1,333 probes). (*D*) Fold enrichment was calculated by taking the fraction of probes covered by, for example, exons, divided by the fraction of the genome covered. Depletion values were obtained by dividing −1 by the enrichment calculation. Regions known to be functional and conserved at the sequence level tend to be enriched for probes on the array. Conversely, regions known to be divergent are depleted for probes.

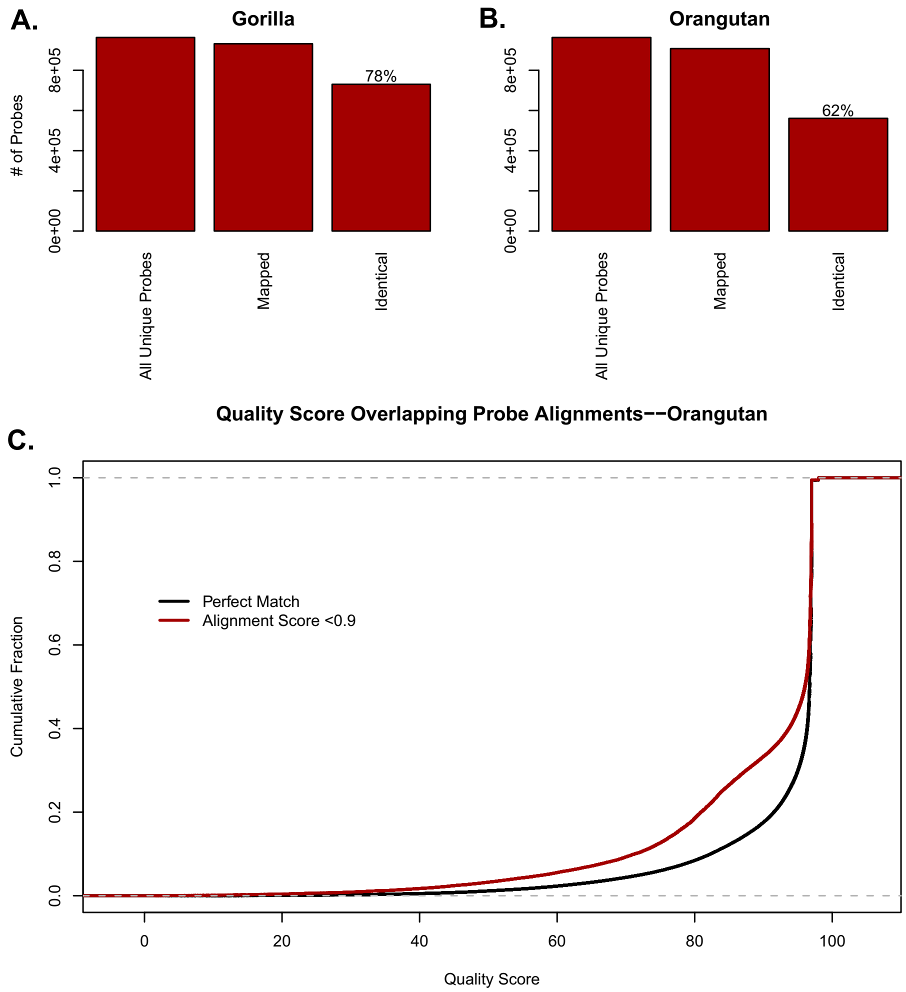

Fig. S2. The vast majority of probes align to the orangutan and gorilla reference genomes. Probe sequences were aligned to the gorGor1 (*A*) and ponAbe2 (*B*) reference genomes using BLAT. The best scoring alignment from each probe was assessed for identity, coverage, and gaps. "Identical" probes are those with 100% identity, 100% coverage, and no gaps in its alignment. (*C*) Probes with lower alignment scores (more mismatches to the orangutan reference genome) were more likely to overlap lower quality sequence in the orangutan reference. Quality scores were obtained from the UCSC Genome Browser for the orangutan reference genome (ponAbe2). Shown here is a cumulative fraction plot for alignments of probe sequences from the array that have 100% identity, 100% coverage, and no gaps, and compared with probe sequence alignments that had a BLAT alignment score <0.9.

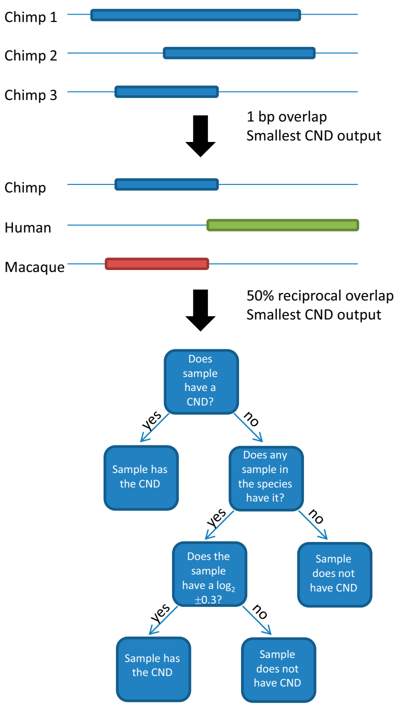

Fig. S3. CND calling schematic. Calls were initially made by using the parameters (listed in *SI Materials and Methods*) in the Nexus Copy Number Analysis software package. To reduce bias caused by variation in resolution between arrays, the original calls were merged into a nonredundant call set. To reduce the rate of false negatives, this set was then used to discover additional CNDs in samples of the same species as those with merged Nexus CND calls.

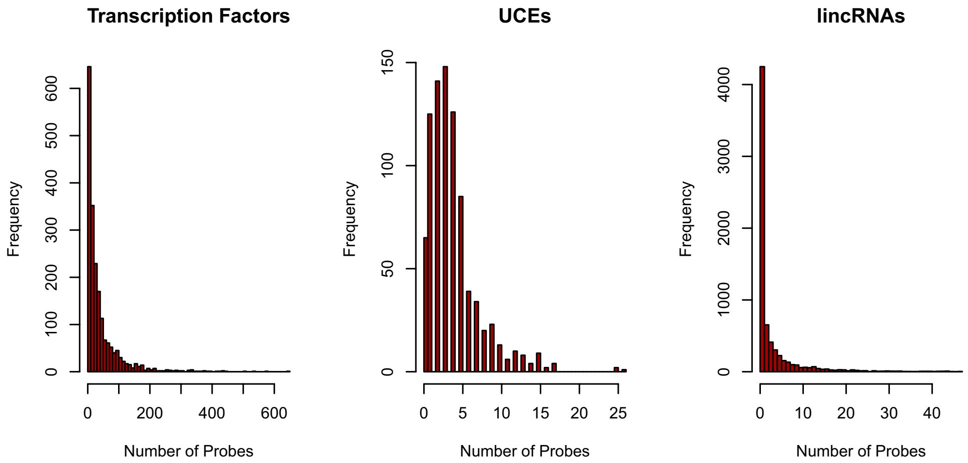

Fig. S4. Coverage of targeted regions by array platform. The number of probes in regions targeted for CND analysis. Regions overlapped by a minimum of three probes were included in downstream analyses. Regions on chromosomes X and Y were omitted from this analysis. For lincRNAs, only overlap with up to 47 probes are shown (the 97.5th percentile). As many as 684 probes overlapped single lincRNAs, but, for visualization purposes, a subset of data are plotted here.

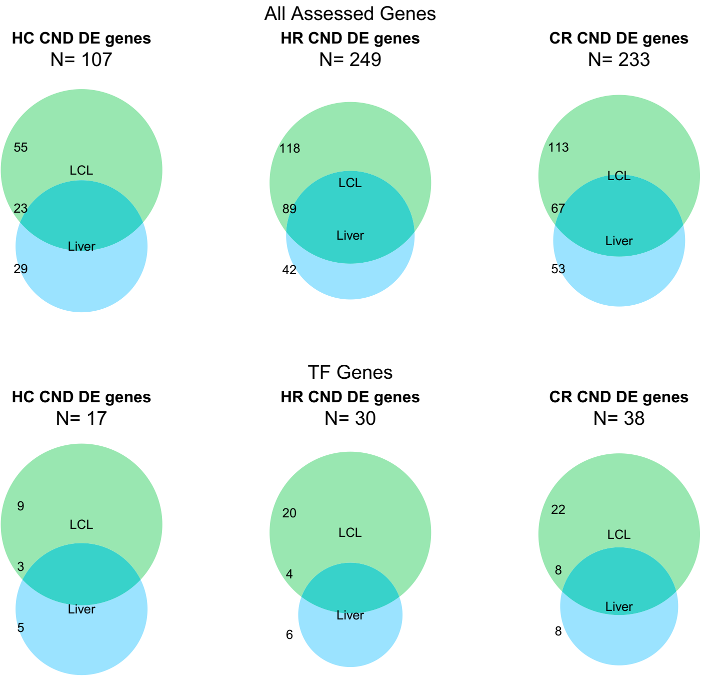

Fig. S5. DE and CND genes in LCLs and liver tissue. Using Dataset S6, which contains genes DE between pairs of species for which there is also a CND between those species, we determined the overlap between LCL and liver tissue. Note the substantial overlap for DE and CND genes between tissues, indicating that the copy number difference may be a common source of the differential expression levels. The number of genes in each part of the Venn diagram is plotted.

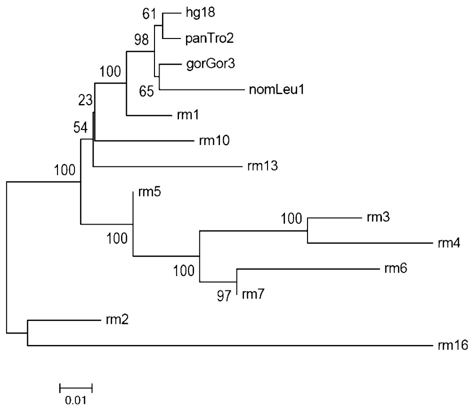

Fig. S6. NJ tree based on the coding sequences of *ZNF669*. NJ tree of *ZNF669* among five primate species [human (hg18), chimpanzee (panTro2), gorilla (gorGor3), gibbon (nomLeu1), and rhesus macaque (rm1-rm13)]. Note rm1 to rm13 represent different copies of this TF gene in the rhesus macaque genome. The numbers at each branch indicate percent bootstrap probabilities (1,000 bootstraps) and the length of the scale below corresponds to 1% expected substitutions per site.

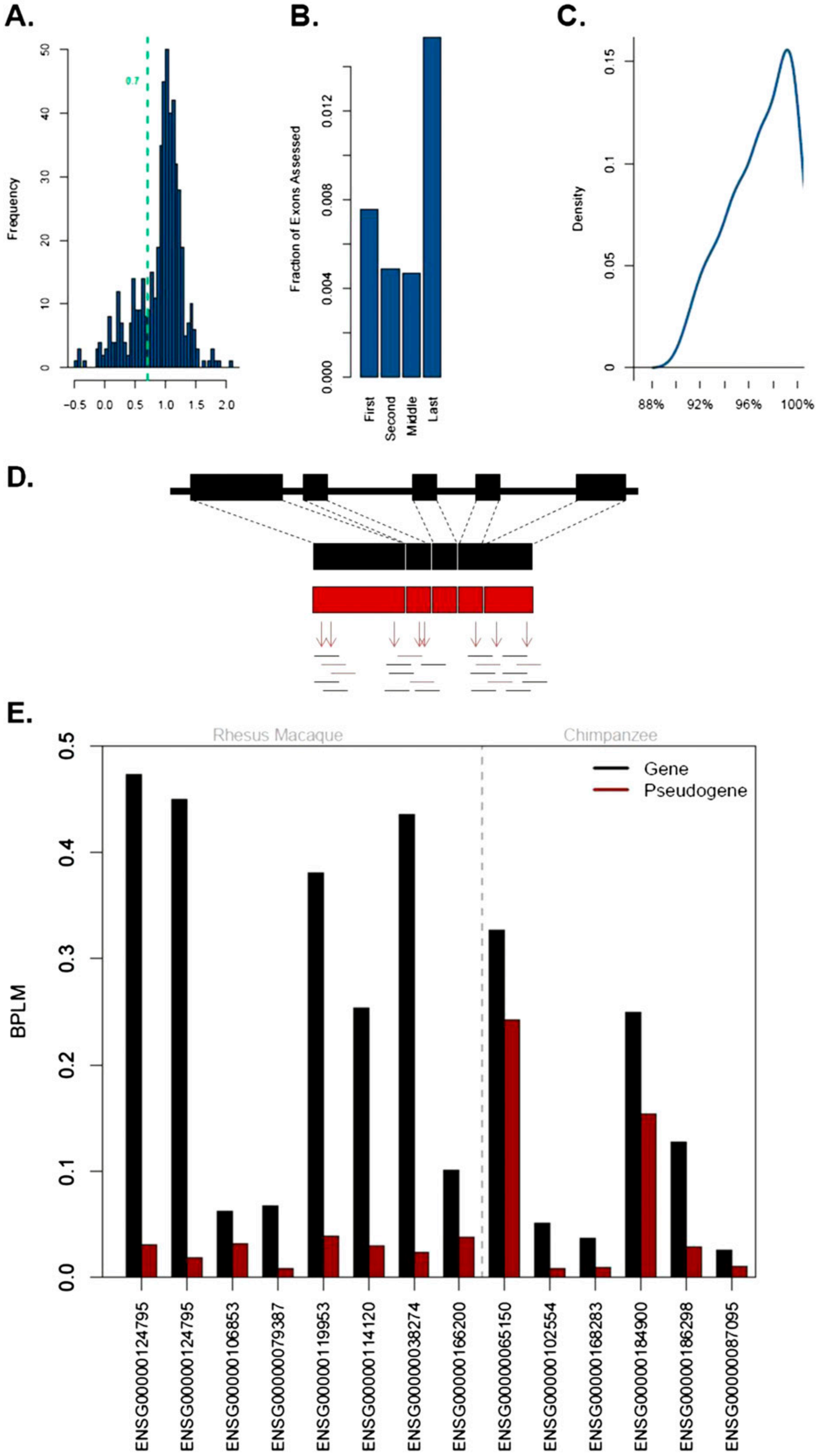

Fig. S7. Analysis of processed pseudogenes. (*A*) Histogram of pseudogene scores across samples with genes having gained exons. The green line indicates a cutoff of 0.7, which was used to define which gained exons were likely part of processed pseudogenes. This cutoff was chosen after visual evaluation in Integrative Genomics Viewer and comparing vs. those pseudogenes found in reference genomes. (*B*) Gained exons found in this study were enriched for 3′ UTRs, further supporting the idea that processed pseudogenes resulting from retrotransposition are common among primates. (*C*) Density plot of percent identity of pseudogenes with their source genes. (*D*) Processed pseudogenes can be expressed in LCLs. Schematic for determining whether processed pseudogenes are expressed using RNAseq. For those pseudogenes that were present in the reference genome (red), we aligned them to their source gene exons

Legend continued on following page

(black). Differences between the gene and pseudogene (red arrows) were considered informative sites for aligning RNAseq reads. Reads that map better to the pseudogene than the reconstructed exon-only source gene were indicative of pseudogene expression. (*E*) RNAseq reads from LCLs were mapped to their respective reference genomes. The reads were subsequently used to calculate expression values by counting the number of mapped bases per gene/pseudogene alignment length per million mapped reads (BPLM). The *x*-axis labels are Ensembl gene IDs based on the hg18 orthologous genes. ENSG00000124795 is listed twice because this gene had two processed pseudogenes in the rhesus macaque reference genome, both of which show evidence of expression. The first eight IDs are for pseudogenes in rhesus macaque and the last six values are for pseudogenes in chimpanzee. Only informative sites (i.e., differences between the parent gene and pseudogene) were used for analysis, so calculating BPLMs allowed us to compare expression between the pseudogene and its parent gene. These BPLMs may strongly underestimate real expression levels, as most of the sites between pseudogene and parent gene are identical and, thus, are not informative (i.e., do not allow for unambiguous read mapping). Please note that pseudogenes typically have lower expression compared with their parent genes. In three cases, the expression of the pseudogene is >50% of the parent gene's expression level. Pearson correlation between the pseudogene and parent gene expression is 0.26 and is insignificant ($P = 0.38$). This lack of correlation suggests that the observed pseudogene expression is independent of the parental gene expression, thus substantiating the fact that the observed pseudogene expression is real and not the result of mismapping of reads from the parent gene (in such a case, one would expect there to be correlation of the pseudogene and parent gene expression levels).

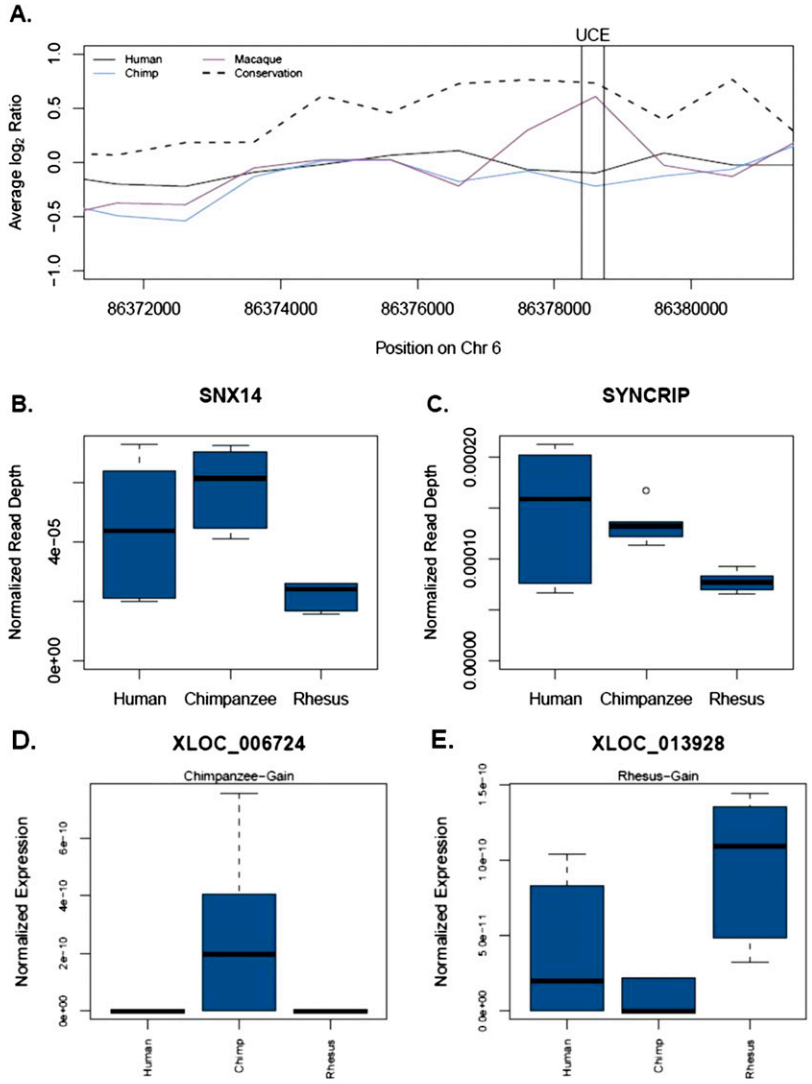

Fig. S8. UCEs and lincRNAs are copy number different and may affect expression. (*A*) Average $\log_2$ ratios across 1-kb bins were averaged by species. The location of a CND UCE is indicated by vertical black lines. Note the increase in $\log_2$ ratio for rhesus macaques across the UCE. Log_{10} *P* values from the conservation track of the UCSC Genome Browser were also binned and plotted as a dashed line. The higher the dashed line (i.e., the lower the *P* value), the more likely the region is evolving under nonneutral conditions. (*B* and *C*) The number of RNAseq reads mapping to *SNX14* and *SYNCRIP* in each sample was normalized by the total number of reads per sample. The lower expression in rhesus macaques is significant compared with human or chimpanzee with an FDR of less than 1%. (*D* and *E*) CND lincRNAs that are also DE between species. lincRNAs are as defined by Cabili et al. (6). lincRNA exon hg19 coordinates were lifted over to hg18, panTro2, and rheMac2 to identify orthologous exons. The number of RNAseq reads mapping to lincRNA exons was normalized by the length of the exon based on the reference genomes and normalized by the total number of RNAseq reads per sample.

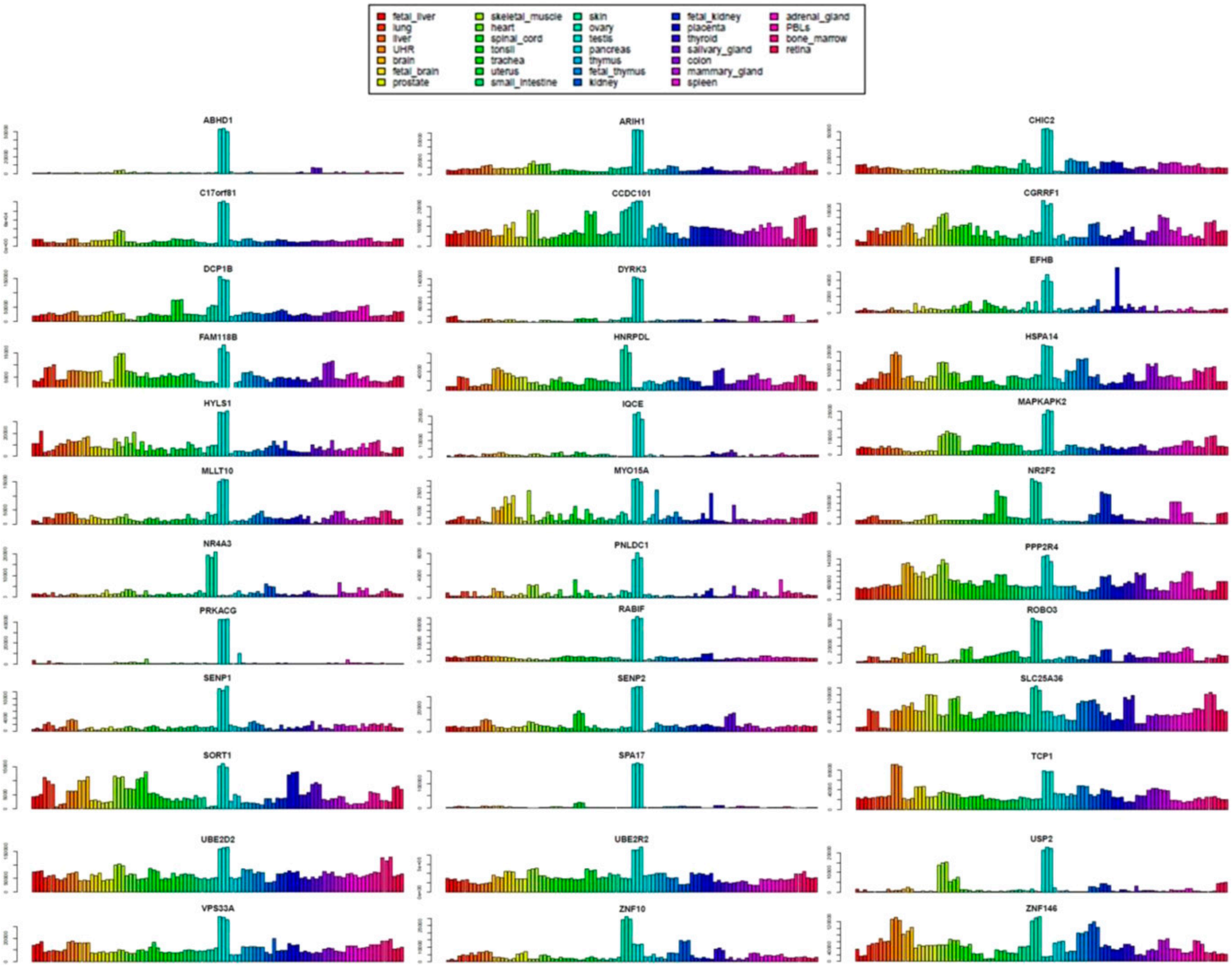

Fig. S9. Expression of CND and DE genes in various human tissues. Expression data were acquired from Dezso et al. (27) under Gene Expression Omnibus accession no. GDS3113. The CND and DE genes in Dataset S6 were examined for expression in this data set. Those in this figure are genes appearing to have higher expression in gonads than most other tissues.

Other Supporting Information Files

Datasets S1–S11 (XLSX)