Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

J Comp Neurol. Author manuscript; available in PMC 2017 February 01.

Published in final edited form as:

J Comp Neurol. 2016 February 1; 524(2): 288–308. doi:10.1002/cne.23843.

High-throughput RNA sequencing reveals structural differences of orthologous brain-expressed genes between western lowland gorillas and humans

Leonard Lipovich[1,2], **Zhuo-cheng Hou**[1], **Hui Jia**[1], **Christopher Sinkler**[1], **Michael McGowen**[1], **Kirstin N. Sterner**[3], **Amy Weckle**[1,4,5], **Amara B. Sugalski**[1], **Lenore Pipes**[6], **Domenico Gatti**[7,8], **Christopher E. Mason**[6], **Chet C. Sherwood**[9], **Patrick R. Hof**[10,11], **Christopher W. Kuzawa**[12], **Lawrence I. Grossman**[1], **Morris Goodman**[1,13,*], and **Derek E. Wildman**[1,4,5]

[1]Center for Molecular Medicine and Genetics, Wayne State University, Detroit, MI 48201, U.S.A

[2]Department of Neurology, School of Medicine, Wayne State University, Detroit, MI 48201, U.S.A

[3]Department of Anthropology, University of Oregon, Eugene, OR 97403, U.S.A

[4]Institute for Genomic Biology, University of Illinois, Urbana, IL 61801, U.S.A

[5]Department of Molecular and Integrative Physiology, University of Illinois, Urbana, IL 61801, U.S.A

[6]Department of Physiology and Biophysics, Weill Cornell Medical College, New York, NY 10021, U.S.A

[7]Department of Biochemistry and Molecular Biology, School of Medicine, Wayne State University, Detroit, MI 48201, U.S.A

[8]Cardiovascular Research Institute, School of Medicine, Wayne State University, Detroit, MI 48201, U.S.A

[9]Department of Anthropology, The George Washington University, Washington, DC 20052, U.S.A

[10]Fishberg Department of Neuroscience and Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY 10029, U.S.A

[11]New York Consortium in Evolutionary Primatology, New York, NY 10024, U.S.A

[12]Department of Anthropology, Northwestern University, Evanston, IL 60208, U.S.A

[13]Department of Anatomy and Cell Biology, School of Medicine, Wayne State University, Detroit, MI 48201, U.S.A

[*]deceased Nov. 14, 2010

Conflict of Interest Statement. The authors declare no conflicts of interest.

Roles of Authors. All authors had full access to all of the data in the study and take responsibility for the integrity and accuracy of the data analysis. Study concept and design: LL, MG, CCS, PRH, DEW. Acquisition of data: ZH, AW, DG, LP. Analysis and interpretation of data: LL, DEW, ZH, HJ, DG, CS, MM, ABS, CEM. Drafting of the manuscript: LL, DEW, CCS, PRH, CWK, DG. Critical revision of the manuscript for important intellectual content: KS, PRH, CCS, CWK, LIG, DEW. Statistical analysis: ZH, HJ, MM. Obtained funding: CCS, CEM, PRH, MG, DEW. Administrative, technical, and material support: AW. Study supervision: LL, DEW.

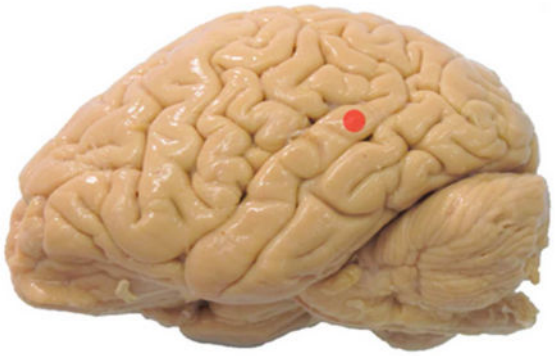
Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Abstract

The human brain, and human cognitive abilities, are strikingly different from those of other great apes despite relatively modest genome sequence divergence. However, little is presently known about the interspecies divergence in gene structure and transcription that might contribute to these phenotypic differences. To date, most comparative studies of gene structure in the brain have examined humans, chimpanzees, and macaque monkeys. To add to this body of knowledge, we analyze here the brain transcriptome of the western lowland gorilla (*Gorilla gorilla gorilla*), an African great ape species that is phylogenetically closely related to humans, but with a brain that is approximately one-third the size. Manual transcriptome curation from a sample of the planum temporale region of the neocortex revealed 12 protein-coding genes and one noncoding-RNA gene with exons in the gorilla unmatched by public transcriptome data from the orthologous human loci. These interspecies gene structure differences accounted for a total of 134 amino acids in proteins found in the gorilla that were absent from protein products of the orthologous human genes. Proteins varying in structure between human and gorilla were involved in immunity and energy metabolism, suggesting their relevance to phenotypic differences. This gorilla neocortical transcriptome comprises an empirical, not homology- or prediction-driven, resource for orthologous gene comparisons between human and gorilla. These findings provide a unique repository of the sequences and structures of thousands of genes transcribed in the gorilla brain, pointing to candidate genes that may contribute to the traits distinguishing humans from other closely related great apes.

Using next-generation RNA sequencing of a *Gorilla* planum temporale (red dot), with transcript-togenome alignments, the authors demonstrate different sequences of orthologous human and gorilla brain messenger RNA and protein molecules. These distinctions relate to the genomic basis of phenotypic differences among hominids, heralding a new era of comparative molecular morphology.

Graphical Abstract



Keywords

alternative splicing; long non-coding RNA (lncRNA); genomics; protein structure; BTBD8; planum temporale

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

INTRODUCTION

Describing the molecular underpinnings of the evolutionarily distinctive features of the human brain is an important goal of comparative neuroscience. There are large differences in brain structure and cognition between humans and other great apes, our closest living relatives (Preuss et al., 2004). The great apes include the chimpanzees, bonobos, and gorillas from Africa and the orangutans from Asia. Humans shared a last common ancestor with chimpanzees and bonobos approximately 6 million years ago, with gorillas 9 million years ago, and with orangutans 16 million years ago (Pozzi et al., 2014; Scally et al., 2012; Meredith et al., 2011; Goodman et al., 1998). Recent studies have revealed differences in gene expression in the brain between humans and chimpanzees, whereas information to compare gene structure and expression with other great apes is more limited (Konopka et al., 2012; Liu et al., 2012; Brawand et al., 2011; Khaitovich et al., 2005; Uddin et al., 2008; Uddin et al., 2004; Enard et al., 2002). The goal of the current study was to analyze the brain transcriptome of the western lowland gorilla (*Gorilla gorilla gorilla*).

Gene expression patterns, as reflected in RNA transcripts measured in harvested tissue samples, provide quantitative insights into gene activity, and can illuminate the development and regulation of neurological structures (Kang et al., 2011; Roberts et al., 2014; Thompson et al., 2014). In addition to such quantitative information about the expression levels of known genes, transcripts can be aligned to the genome, yielding insights about the structures (the precise location of promoters, exons, introns, splice sites, and protein-coding regions) of known and novel genes. Information about the structure of genes is essential for understanding alternative splicing – a process by which one gene gives rise to multiple transcripts and hence, frequently, multiple proteins with potentially distinct biological roles. Past research has compared such patterns of gene expression in human and nonhuman primate samples to clarify empirically and experimentally the molecular mechanisms and evolution of distinctly human neurological traits (Konopka et al., 2012; Liu et al., 2012; Brawand et al., 2011; Khaitovich et al., 2005; Uddin et al., 2008; Uddin et al., 2004; Enard et al., 2002). Recent work has demonstrated that, in addition to gene expression differences, alternative splicing of brain-expressed genes can contribute to regional variation in cortical patterning (Bae et al., 2014), and that some genes involved in learning and memory are alternatively spliced and differentially expressed in brain tissue from humans and other great apes in comparison to other primates (Li et al., 2004).

Past studies in the area of comparative genomics, which focused on multispecies sequence alignments and detection of functional constraint and adaptive evolution in the aligned sequences, have consistently emphasized sequence conservation and the resulting functional similarities rather than differences among multiple species (Birney et al., 2007; Lindblad-Toh et al., 2011; Margulies et al., 2007; Warnefors and Kaessmann, 2013; Uddin et al., 2008). Historically, this work had been dominated by computationally generated, rather than experimental, datasets that were unable to reveal which regions of the genome are transcribed into RNA and, in turn, potentially translated into proteins that influence phenotypic differences between species. To address this limitation, several human transcriptome projects have generated large public RNA sequence collections that have transformed our understanding of human gene structure diversity and alternative splicing

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

(Temple et al., 2009; Takeda et al., 2006; Ota et al., 2004). More recently, these projects have greatly extended our understanding of human gene expression, enabling a catalogue of the brain transcriptome (Hawrylycz et al., 2012). Comparative transcriptome resources in other species, including gorillas, are emerging as well and have resulted in valuable insights into the evolution of gene expression, employing new experimental results to highlight expression modules common among different primates (Brawand et al., 2011), as well as differences, and generating a publicly accessible web-based infrastructure for interspecies primate transcriptome comparisons (Pipes et al., 2013).

Transcript-to-genome alignments at orthologous genes (genes that exist in multiple species and that evolved directly from a single ancestral gene in the common ancestor of those species) among closely related primate species point to a class of interspecies genomic distinctions that cannot be inferred solely by analyzing genome sequences, because they arise from transcriptional phenomena such as alternative splicing. These differences can be inferred from transcript-to-genome alignments as species-specific transcription start sites, splice junctions, and gene termini (3'-ends) at genic regions that are conserved between species. For example, gene structure differences at orthologous loci between humans and chimpanzees are evident as species-specific inclusion or exclusion of exons (Calarco et al., 2007). Lack of nonhuman primate transcriptome sequence data has hampered discovery of such differences, although these distinctions exist between humans and diverse nonhuman primates (Lee and Lipovich, 2008). Human-specific changes in transcription start sites, end sites, and exon usage at conserved genes have been linked to human-specific gene functions relevant to morphogenesis in a recent comparative study of liver tissue from humans, chimpanzees, and macaques (Blekhman et al., 2010).

Numerous studies have extrapolated from human gene structures (which are supported by extensive empirical transcriptome datasets) to infer the predicted structures of the orthologous nonhuman primate genes (in species with little or no available transcriptome data concerning gene structure and expression), on the basis of aligning human transcripts to the nonhuman genomes. But one limitation of this approach is that it cannot infer gene structure differences between closely related species, although such differences can be deduced if empirical collections of transcriptome data are available for all species under analysis (Wood et al., 2013). To infer gene models relevant to brain structure and function in gorillas empirically and more accurately as an alternative to predictive models, we undertook the current study to collect transcriptome data from the planum temporale of an adult male western lowland gorilla (*Gorilla gorilla gorilla*). This method can lead to novel insights about gene structure differences and similarities among primates and may point to biological pathways and molecular functions that could be instrumental in the uniquely human aspects of brain biology.

MATERIALS AND METHODS

Isolation of gorilla RNA

We isolated the RNA from a 41-year old adult male *Gorilla gorilla gorilla* brain from the planum temporale region obtained within 12 hours postmortem. The gorilla had died of natural causes (fibrosing cardiomyopathy, a condition common in great apes). The brain of

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

this gorilla was provided to us by Dr. J. M. Erwin through the Great Ape Aging Project and the Atlanta Zoo. A routine neuropathology evaluation had shown the gross brain specimen and histological samples to be normal. Our analysis was confined to a single individual as the availability of fresh frozen postmortem brain samples from gorillas is extremely limited. Our choice of the planum temporale (Fig. 1) was motivated by the function of this neocortical area in multimodal sensory processing and language (Zheng, 2009) and by the anatomical commonality of planum temporale asymmetry between humans and great apes (Gannon et al., 1998; Hopkins et al., 1998).

Tissue was homogenized with Trizol and total RNA was isolated following manufacturer specifications (Invitrogen, Carlsbad, CA). RNA was purified using the RNeasy Midi Kit and cleaned using the RNase-free DNase kit (Qiagen, Valencia, CA). RNA concentration and quality were analyzed using the Nanodrop 1000 (Thermo Fisher Scientific, Wilmington, DE) and Agilent Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA). Purified RNA (2.5 μg) was provided to the Applied Genomics Technology Center at Wayne State University, where it was sequenced utilizing the Illumina GAIIx protocol (part #1004898, San Diego, CA).

Gorilla brain transcriptome sequence analysis

Raw 76-bp paired-end (PE) sequence tags from internal regions of transcripts were computationally assembled into transcriptome scaffolds using published methods detailed further below (Langmead et al., 2009; Trapnell et al., 2010) and the manufacturer's instructions (Illumina, San Diego, CA). Tags were first computationally clustered into contigs. Subsequently, scaffolds, representing partial or full-length transcript models with multiple independent contig support, were constructed from these contigs. Some contigs could not be joined into any scaffolds and are hereafter referred to as singleton contigs or "singletons."

We obtained 16,696,937 PE short reads totaling 2,537,934,424 bp that met the GAPipeline quality criteria. These reads were used for assembly and annotation. PE short reads were assembled into scaffolds using SOAPdenovo (Li et al., 2010). The assembled transcripts included both scaffolds and singletons. There were 209,588 assembled transcripts, comprising 54,283,005 bp of total transcribed sequence. The longest transcript was 6,210 bp. The N50 of the assembled transcripts was 434 bp.

We assembled the raw data into 45,275 sequences of 300 nucleotides (nt) or longer and processed all sequences with RepeatMasker (http://www.repeatmasker.org). We then applied the UCSC BLAT tool (Kent, 2002a) to these sequences in order to identify the subset of the gorilla transcripts that mapped unambiguously to the hg18 human genome assembly. There were no assembled gorilla transcripts of 300 nt or greater length that either failed to align to the human genome or aligned to multiple human genomic locations after RepeatMasker was used. Therefore, we used the 300-nt minimum threshold upon observing that at 100-nt and 200-nt thresholds both genomic non-mapping and genomic multi-mapping persisted for some transcripts after RepeatMasker (data not shown). The 45,275 gorilla transcripts with unique human genomic mappings and transcript lengths ≥300 nt

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

represented partial-length or full-length gorilla transcript models that mapped to known locations in the human genome (Supplementary Dataset 1).

To identify uniquely mapped gorilla transcripts that represented putative orthologs of specific human genes, we defined the genomic region (chromosome, start, and end) covered by each gorilla transcript, and queried that genomic region for overlap with RefSeq protein-coding (NM), RefSeq noncoding-RNA (NR), and other long noncoding-RNA (Jia et al., 2010) genes in the UCSC GoldenPath transcript-to-genome alignment files ref_all (for RefSeq) and all_mrna (for non-RefSeq transcripts). Only exonic overlaps (corresponding to UCSC blocks) from the complete genomewide set of query results (Supplementary Dataset 2) were scored (Fig. 2).

If any partial (at least one base) or complete overlap of genomic positions (start-end ranges on the same chromosome) was detected, the corresponding gorilla transcript was flagged as an "overlap" with the matched human transcript. At this stage, we did not further examine putative introns manifesting as deletions in gorilla transcript to human genome alignments for gap classification as GT-AG introns versus other classes of gaps. Because the gorilla transcript assemblies were unoriented (i.e. no strand information was included), we did not distinguish sense-strand from antisense-strand coverage of the known human transcripts by the gorilla transcript assemblies. However, we manually confirmed gorilla transcript orientations for a subset of the mappings as described in the next section. The gorilla transcript assemblies for the most part were not long enough to provide complete start-to-end coverage of a human cDNA. Therefore, the true ends of each gorilla transcript detected might reside either at or beyond the boundaries of each assembled RNAseq-inferred gorilla transcript model.

In addition to tabulating gorilla transcripts whose human genomic mappings overlapped with the genomic mappings of the known human transcripts, we searched for gorilla transcripts that mapped near but not within known human transcripts. This additional search was motivated by the finding that untranslated regions (UTRs) of protein-coding genes can often extend more than 10,000 bp (10 kb) beyond the end of a gene, and can be missing or misannotated in publicly available consensus models of that gene (Miura et al., 2013), as well as by the fact that noncoding-RNA transcriptional units, distinct from known protein-coding genes, often occur near those known genes and may *cis*-regulate the known genes (Lipovich et al., 2010). Gorilla transcripts near but not within known human genes were flagged as "flank10k." A distance of under 10 kb between any one of the two ends of a gorilla transcripts and the nearest human RefSeq or lncRNA was sufficient for a "flank10k" designation. All positional relationships between gorilla transcripts and human protein-coding and noncoding-RNA genes are summarized in Supplementary Dataset 3, and are stated in detail in Supplementary Dataset 4. For each human gene, we have also listed the sequences of all gorilla scaffolds matching that gene. We are reporting these gene names and the corresponding gorilla transcript identifiers, with sequences, in Supplementary Datasets 5, 6, and 7, for protein-coding genes, RefSeq noncoding-RNA genes, and non-RefSeq long noncoding-RNA genes, respectively.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

To quantify transcript abundance, and to thereby complement the unquantified SOAPdenovo transcriptome assemblies, we also mapped the gorilla transcriptome to the most recent release of the gorilla genome gorGor Version 3.1 (Scally et al., 2012). Reads were aligned using Bowtie v. 0.12.7 (Langmead et al., 2009) and subsequently, splice junctions were identified using TopHat v. 1.3.3 (Trapnell et al., 2010). Cufflinks v. 1.2.1 was used to quantify expression values using fragments per kb of exon per million mapped reads (FPKM; Trapnell et al., 2010) with and without the annotated gorilla gene models from Ensembl (version 72). RNAseq metrics were computed using Picard Tools (http://picard.sourceforge.net/).

Analysis of human-gorilla gene structure differences affecting brain-expressed genes

To identify human-gorilla gene structure differences affecting the 5'-ends (transcription start sites, promoters) and 3'-ends of conserved genes, we computationally identified all uniquely mapped gorilla transcriptome scaffolds that either started at least 5 kb before the transcription start site (the 5'-end) of a known human gene or ended at least 5 kb after the terminus (the 3'-end) of a known human gene (Fig. 3). All such scaffolds were used as BLAT queries and subjected to manual annotation in the UCSC Genome Browser (Kent et al 2002b; http://genome.ucsc.edu). Gorilla transcripts with zero BLAT mappings, or multiple equally good BLAT mappings, were discarded. For BLAT multiple alignments generated for other gorilla transcripts, the alignment with the highest score and percent identity was selected. BLAT results were visualized in the browser to confirm any computationally identified differences in splicing of the gorilla scaffold transcript relative to the human transcripts (including RefSeq, GenBank mRNAs, and all ESTs) aligning at the same genomic location. Such differences, once manually confirmed, were operationally defined as gorilla-specific (gene structure elements unique to gorilla) if they did not occur in any UCSC-mapped human RefSeq, mRNA, or EST sequences, and if the splice junction of the apparent unique gorilla exon (first GT splice donor for first exons, last AG splice acceptor for last exons) was canonical along the human genome at the splice junction location that was suggested by each alignment of an RNAseq-inferred gorilla transcript to the human genome.

We manually annotated each gorilla transcript containing a gorilla-specific 5'- or 3'-end. In contrast to the automated pipeline, manual annotation enabled us to determine the direction of transcription (biological orientation) of each gorilla transcript. We determined the orientation of the gorilla sequence by comparing the sequence with the matched human RefSeq gene structure at the locus, human cDNA or EST sequences mapping to the locus, and orientation of the splice junctions (GT-AG introns) represented by the human RefSeq, cDNA, and EST sequences at each locus. Direction of the read was given by the orientation that allowed the most canonical junctions – the presence of GT nucleotides at the splice donor site and AG nucleotides at the splice acceptor site, with transcript orientation in the 5'->3' GT->AG direction. Sequences without canonical splice junctions, or without a single orientation yielding a majority of canonical junctions, were discarded. Sequences with human RefSeq, cDNA, and/or EST matches representing human splice isoforms whose genomic structure was identical to that of the gorilla isoform were discarded because they indicated that the 5'- or 3'-end gene extension discovered from the gorilla transcriptome

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

data was not unique to gorilla. Sequences with non-contiguous ESTs or a single, unspliced EST were analyzed further. Poor-quality ESTs were discarded. We classified each gorilla-human transcript-to-genome alignment into one of four types: uninteresting (no gene structure differences between human and gorilla could be discerned from comparing the gorilla assembled transcript with the human gene structure at the locus where the gorilla transcript aligned); 5'-end extension or difference (a putative unique transcription start site and initial exon in the gorilla, not supported by any human RefSeq, cDNA, or EST data); 3'-end extension or difference (a putative unique transcription end site and last exon in the gorilla); and high-complexity (including cases where a single gorilla assembled transcript appeared to match multiple human genes) (Fig. 4). We refer to the 5'- and 3'-ends as putative because there is no guarantee of full-length 5'- and 3'-end capture by our transcriptome sampling method; therefore, the gorilla transcript assemblies may or may not capture the actual mRNA termini, and may be limited to internal regions of the gorilla transcripts.

We also examined the conservation of nucleotide sequences around splice junctions to determine whether any gorilla-specific splicing events affecting gene 5'- or 3'-end structures were dependent on non-conserved or primate-specific splice sites. Conservation was checked using the UCSC Genome Browser visual conservation peak track and the 44 aligned vertebrate species MultiZ tracks (genome.ucsc.edu) by zooming to base-level resolution in the browser to reveal the evolutionary depth of conservation for each splice-site GT and AG dinucleotide.

To determine the impact, if any, of each putative gorilla-specific 5'- or 3'-end gene structure difference on the protein-coding sequence of the conserved gene, we detected open reading frames (ORFs) in the gorilla ortholog transcript sequence and compared them with the human ORFs inferred from the RefSeq and/or full-length cDNA of the human gene. Sequences lacking positive-strand ORFs by ORF Predictor (Jia et al., 2010) were analyzed by the NCBI ORF Finder (http://www.ncbi.nlm.nih.gov/projects/gorf/). If a sequence had multiple ORFs found with either method, the longest predicted protein sequence was used. We pairwise-aligned the translated gorilla and human orthologous protein sequences by BL2SEQ (Tatusova and Madden, 1999). All gorilla-specific N- or C-terminal amino acid sequences, as well as gene-termini differences not affecting the protein sequence of orthologous genes, were annotated.

Homology modeling and molecular dynamics

A complete three-dimensional model of gorilla BTBD8 protein was built with Prime 2.1 (Schrödinger LLC, New York, NY) using the X-ray structure of 5-lipoxygenase (PDB 3O8Y), SPOP (PDB 3HQI), and human klhl11 (Kelch-like protein 11; PDB 3I3N) as templates (Fig. 5). Molecular dynamics (MD) run in a fully solvated environment was performed using Desmond (D.E. Shaw Research).

Expression of candidate genes in the human brain

We tested whether each of the genes found to have unique exons in the gorilla transcriptome is expressed in the human brain using the Allen Brain Atlas, a publicly available database

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

providing gene expression data from various regions of the developing and adult human brain (Hawrylycz et al., 2012). Specifically, we catalogued gene expression measurements from both the human temporal lobe and the human total brain in the form of FPKM and compared this to the gorilla gene expression values from our Cufflinks analysis. In addition, we visually interrogated the human brain microarray resource at these authors' website (http://brain-map.org), viewing the expression levels of all probes, for each of our 13 genes, across all of the distinct individual donor planum temporale samples. Default heat map colors on that website are in the green-red scale where green signifies relatively low expression and red relatively high expression (http://help.brain-map.org/display/humanbrain/Microarray+Data).

RESULTS

An unbiased survey of the gorilla brain transcriptome

Nearly 20% of our raw reads were intergenic and an additional 12% were intronic with respect to public (Ensembl, version 72) gorilla gene models (Supplementary Dataset 8), attesting to a likely substantial presence of previously unrecognized splice variants and entirely novel transcripts in our dataset. This was expected because the public gorilla gene models are largely homology-driven predictions that are based on human genes and the complexity of the gorilla brain transcriptome was not fully sampled prior to our study.

We then proceeded to align the gorilla transcriptome to the human genome and human gene models. Of the 45,275 SOAPdenovo assembled gorilla transcript sequences with unique human genome mappings, 32,634 assembled gorilla transcripts (72%) overlapped exons of protein-coding RefSeq genes, 18,628 were within 10 kb of a protein-coding RefSeq gene (41%), 3,828 (8%) overlapped exons of noncoding-RNA genes from the two noncoding-RNA catalogs that we used (the NCBI RefSeq database and our dataset from Jia et al., 2010), and 7,738 (17%) mapped within 10kb of a noncoding-RNA gene (Supplementary Dataset 9). Note that the total does not equal 45,275 because some transcripts belonged simultaneously to multiple categories, for example overlapping a RefSeq gene while also within 10 kb of a noncoding-RNA gene. The alignment of the transcriptome to the gorilla genome using TopHat and Cufflinks produced evidence for the expression of 45,504 transcripts (i.e., these transcripts have FPKM > 0), roughly equal to the number of transcripts resulting from the SOAPdenovo analysis. However, only 39,432 of these had FPKM > 1 whereas 9,578 had FPKM >10.

Along the human genome, 17,173 human protein-coding RefSeq transcripts overlapped gorilla transcripts, and 12,385 human protein-coding RefSeq transcripts mapped within 10 kb of gorilla transcripts. In the gorilla transcriptome data, 20,640 nonredundant protein-coding human RefSeq transcripts (not the sum of the previous two counts, because some transcripts appear in both of the previous categories) were matched. Therefore, two-thirds of the 30,221 nonredundant protein-coding human RefSeq transcripts are represented in the gorilla brain transcriptome. NCBI DAVID GeneID redundancy assessment of the 20,640 represented transcripts indicated that they corresponded to 10,511 distinct human protein-coding genes because some genes are represented by multiple RefSeq transcripts due to alternative initiation, termination, or splicing.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Of the human noncoding-RNA transcripts (from the RefSeq NR and the Jia et al., 2010 datasets), 3,828 overlapped gorilla transcripts along the human genome, and 7,738 nonredundant human noncoding-RNA transcripts mapped within 10 kb of gorilla transcripts along the human genome, for a total of 4,227 nonredundant human noncoding-RNA transcripts matched by the gorilla transcriptome data and representing 1,870 nonredundant ncRNA genes.

Only 412 (<1%) of the 45,275 gorilla transcripts mapped more than 10 kb away from any protein-coding or noncoding-RNA gene on the human genome, did not match any GenBank mRNAs, and therefore could not be annotated using simple genomic positional gene matching approaches. Of the 412, 102 overlapped segmental duplications (allowing for the possibility that they are related to known genes somewhere else, in another copy of the same duplication family). The 310 remaining transcripts may represent entirely gorilla-specific genes, or very distant alternative splices of known genes that, due to the partial-length coverage of their originating gorilla transcripts, could not be matched to the known genes from which they had originated (Supplementary Dataset 10).

We examined the extent to which protein-coding sequences (CDS) of known-gene human RefSeq transcripts were covered by our assembled gorilla transcriptome sequences. More than 1,000 protein-coding RefSeq transcripts (of the 17,173 that had exonic overlaps with the gorilla sequences) had greater than 90% coverage of their protein-coding ORFs by the gorilla transcript assemblies, including the transcripts of 288 nonredundant RefSeq genes with 100% coverage (Supplementary Dataset 11), and more than 5,000 had more than half of their ORFs covered by the gorilla transcript sequences. Therefore, this gorilla neocortical transcriptome comprises a potentially useful empirical, rather than homology-based or gene prediction-driven, resource that can be used in future comparative analyses of orthologous genes in the two species (Fig. 6).

As an additional quality control measure, we evaluated the coverage of the human genome by the alignable gorilla RNAseq transcriptome on a by-chromosome basis. If the coverage is devoid of genomic-position biases, then the longest human chromosomes are expected to have the greatest numbers of coding and noncoding genes matched by the putatively orthologous gorilla transcripts, whereas the shortest and the most gene-poor human chromosomes should have the smallest numbers of genes matched by gorilla transcripts. This expectation matches the observed human genome distribution of the alignable gorilla transcript mappings, which highlights both the higher gene counts on the largest human chromosomes and the dips in gene numbers on the known gene-poor chromosomes 13, 18, and 21 (Supplementary Dataset 13).

Manual annotation of gene structure difference pipeline results confirms human-gorilla gene structure differences at 13 loci

Automated search for gorilla assembled transcripts starting more than 5 kb before or ending more than 5 kb after known human protein-coding RefSeq genes revealed 374 putative gene structure differences between the two species in our SOAPdenovo data (Supplementary Dataset 14). However, the majority of the computationally detected differences were classified as false positives by our UCSC-based manual annotation of each locus, performed

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

as described in Materials and Methods. Transcript redundancy and locus complexity comprised a major explanation for the pipeline's high false-discovery rate. For some genes with multiple RefSeq transcripts, the pipeline scored a greater than 5 kb discrepancy between the human and gorilla transcript boundaries based on one RefSeq transcript, disregarding the transcriptional unit's additional RefSeq transcripts that may fail to display a discrepancy. More commonly, we observed that the extremely conservative RefSeq gene model construction employed by the NCBI excluded EST and full-length cDNA support continuing far beyond the boundaries of the RefSeq model. Because our pipeline was RefSeq-driven, it would flag any locus where the gorilla transcript's boundary was more than 5 kb away from the corresponding (start or end) boundary of the human RefSeq gene model, regardless of the presence and prevalence of human non-RefSeq evidence whose genomic structure may have much more closely corresponded to that inferred from the gorilla transcript. Only 13 loci remained after our manual UCSC-based alignment of the RNAseq-inferred SOAPdenovo gorilla brain transcript sequences to the annotated human genome (Table 1 and Supplementary Dataset 15); nine of the 13 also appeared in the Cufflinks analysis, while CATSPER2, OSTN, PPARGC1A, and RFPL1S lacked Cufflinks quantifications. The three major types of interspecies gene structure differences are illustrated by actual examples from amongst the 13 loci in Fig. 7.

Although all gorilla-specific splice sites and exons we detected are in genomic sequences conserved in other species, each gorilla-specific splice event, and the corresponding gorilla gene exon structure, is completely absent in public human RefSeq, cDNA, and EST transcriptome data. In view of the extraordinary depth of human transcriptome sequencing, this finding suggests that these splices and alternative exons indeed do not occur in humans. To our knowledge, this work is also the first to show that these genes, including the 13 with human-gorilla gene structure differences identified by us here, are expressed in the gorilla brain at the RNA level. Although the bioGPS human gene expression database (Wu et al., 2009) indicates that most of those 13 genes may be expressed in the human brain in at least very low levels, only one (RFPL1S, a noncoding-RNA gene) is brain-specific in human, and one additional gene (BTBD8) is highly expressed in human fetal brain in addition to expression in other organs. Thus, we have identified a combination of gorilla brain expression and gorilla-human gene structure differences for 11 genes that were not previously known to fit either of those two criteria. None of our 13 genes with human-gorilla gene structure differences correspond to any of the 180 "human-specific exon usage" alternatively spliced genes identified by a recent study of evolutionary lineage-specific alternative splicing in primates that did not include gorillas (Blekhman et al., 2010).

A PantherDB (Mi et al., 2013) search with these 13 genes as the query and the 17,173 gorilla-matched RefSeq transcripts as the background did not identify any significant enrichment or depletion in the 13 genes of any functional categories profiled by any of the five PantherDB search modes (biological process, molecular function, pathway, subcellular localization, PANTHER protein class), presumably because this is too small a gene list for an enrichment calculation attempt. Only two genes (RFPL1S and PPARGC1A) had public database support of the gorilla splice variants by full-length cDNA sequences from species other than human and gorilla.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Gorilla and human orthologous gene expression profiles are similar

We computed a normalized expression value for each of these 13 genes in both gorilla and human (Supplementary Dataset 16). We next used our planum temporale RNAseq dataset (Supplementary Datasets 17 and 18) for gorilla. We then retrieved the Allen Brain Atlas quantitated human RNAseq results from two datasets: temporal lobe and total brain. Of the 13 genes with unique structures in gorilla relative to human, 10 were detectable (i.e., transcribed) in gorilla and both human samples. Furthermore, when ranked in descending order by the gorilla expression level, four of the top five orthologous genes among these 10 (*CWC15*, *GBF1*, *E2F1*, *OSTN*) had the same relative rankings of expression levels in all three human samples, indicating similar expression levels and the same rank order in both gorilla and human. One notable exception was *PPP1R3C*, which was the third most highly expressed gene in both human samples but the seventh in gorilla. We also interrogated the Allen Brain Atlas human microarray datasets, which are available for the planum temporale, screening all probes across all individual donors. We found that ten of our 13 genes (all except CWC15, APOH, and RNF17) were uniformly detectable in the human planum temporale at high to medium levels across all probes and donors. (CWC15 was the sole discrepancy between the RNAseq and the microarray, which could be explicable by the use of the temporal lobe for the former and planum temporale for the latter.) As such, the Allen Brain Atlas microarray results show that the majority of these genes are expressed in the human planum temporale, although the microarray probes were designed based on the known human splice isoforms of these genes and therefore inherently could not have captured the human counterparts of the specific novel splice isoforms from the gorilla.

Two genes with previously known brain expression have nucleotide and protein sequence differences between gorilla and human

Brain expression of BTBD8 has been previously reported in human (Xu et al., 2004), and our bioGPS query indicated that RFPL1S is expressed in human brain as well. The gorilla assembled transcript matching BTBD8, a nuclear protein containing BTB/POZ and Kelch/ankyrin domains, has a unique 3’-end exon generated by splicing into the 5’-UTR of the same-strand downstream gene KIAA1107. The recovery of this transcript from the adult gorilla brain adds to the available insights from human BTBD8 expression datasets, which indicate mainly fetal-brain expression, and suggests that characterization of the gorilla protein and comparison of the gorilla and human spatiotemporal expression patterns of BTBD8 in the brain may elucidate the function of the gorilla-specific protein BTBD8 sequence.

The unique 3’-end exon of the BTBD8 gorilla transcript encodes a 34-amino acid C-terminal extension, distinct from the human C-terminal, but with protein database matches in other mammals. This observation suggests that most likely this splice event was lost in humans rather than gained in gorillas. We searched for homologous sequences in the Protein Data Bank (PDB). Based on the search, we built a full-atom homology model of the entire gorilla BTBD8 protein, using Prime 2.1. The model (Fig. 8) comprises two back-to-back BTB domains followed by a C-terminal BACK (BTB and C-terminal Kelch) domain. The critical difference between the human protein and the gorilla protein is the presence in the latter of two extra α-helices encoded by the unique 3’-end exon of the BTBD8 gorilla transcript.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

RFPL1S is an endogenous lncRNA transcriptional unit antisense to the RFPL1 gene, a member of the RING-B30 family (Seroussi et al., 1999), which is the second gene with known brain expression that we identified in our study. The human RFPL1S lncRNA has an almost completely brain-specific expression pattern, with the highest levels in the amygdala, the prefrontal cortex, and the cingulate cortex, whereas the expression of RFPL1 is considerably more diverse and reaches its highest levels outside of the brain (bioGPS; data not shown).

A viral infection response gene has a gorilla-specific first exon contributing an N-terminal protein extension

GBF1, a GDP/GTP exchange factor essential for Golgi complex assembly, is involved in numerous functions in protein trafficking and cell cycle replication; inhibition of GBF1 with siRNA causes cell-cycle arrest in G_0/G_1 (Citterio et al., 2007). GBF1 is a key host factor that aids viral replication upon infection and has been co-opted for replication by diverse viruses that infect humans, including poliovirus, enterovirus, and hepatitis C (Belov et al., 2010; Goueslain et al., 2010; van der Linden et al., 2010), although GBF1 has not been studied in gorillas. The gorilla assembled transcript that maps uniquely to the human GBF1 gene contains an upstream exon and an upstream intron splice donor, neither of which is observed in any of the numerous human cDNAs and ESTs representing this gene in public transcriptome data. The splice donor GT is conserved throughout placental mammals. While the human GBF1 first exon is entirely 5'-UTR, the gorilla first exon in the planum temporale sample encodes both the 5'-UTR and the first 6 amino acids of the protein sequence. The human initiator methionine is the seventh amino acid in the *in silico* protein translation of the gorilla transcript.

Two spermatogenesis genes exhibit interspecies differences

Conserved male-reproduction genes functionally associated with speciation through their contributions to the generation of reproductive barriers (Clark et al., 2009) have undergone positive selection in multiple eukaryotic lineages, and reproductive system expression is enriched among non-conserved primate-specific genes as well (Tay et al., 2009 and references therein). Two of the 13 brain-expressed genes with human-gorilla gene structure differences also function in spermatogenesis. Human RNF17, expressed primarily in the testis, is a protein containing ring-finger and tudor domains (that may interact with histone tails as well as contribute to the Piwi RNA pathway [Ozboyaci et al., 2011; Siomi et al., 2010]), and is essential for spermiogenesis (Pan et al., 2005). Our brain-expressed gorilla-specific splice variant, not supported by any public human transcriptome data, uses the conserved GT splice donor of the penultimate human intron, misses the last two human exons containing the human C-terminus, and splices instead through a canonical AG splice acceptor into a LINE1 retrotransposon located downstream of the 3' boundary of human RNF17. Translation stop is predicted *in silico* to occur after an 11-amino acid C-terminal ORF contributed by the LINE1 element.

CATSPER2 is a voltage-gated calcium channel gene that functions in sperm motility and has been associated with male infertility (Hildebrand et al., 2010). The gorilla-specific splice variant that we observed lacks any human transcriptome support in GenBank and represents

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

intergenic splicing of two same-orientation nearby genes: CATSPER2P1 (an upstream unprocessed pseudogene of CATSPER2, a probable product of an ancestral CATSPER2 tandem duplication) and CATSPER2 itself. The first intron of CATSPER2 is duplicated, and the splice donor of the upstream unprocessed pseudogene is spliced directly into the splice acceptor of CATSPER2 over nearly 100 kb of genomic sequence. As the first exon of both genes is entirely 5'-UTR, the gorilla splice variant uses the CATSPER2 ORF and would not encode any unique gorilla-specific protein sequence.

Four genes relevant to brain energy metabolism have gorilla-specific upstream exons

Human *PPP1R3C* encodes an inhibiting subunit of protein phosphatase 1, expressed primarily outside of the brain. It is regulated by HIF-1 and promotes glycogen accumulation in hypoxic conditions (Shen et al., 2010). The gorilla splice variant detected by our RNAseq analysis of the planum temporale adds two upstream exons and two introns. Although the splice acceptor of the second intron of this splice variant is shared with the splice acceptor of the first intron of the canonical human transcript, the other three splice sites utilized in the gorilla are not supported by any human transcriptome data in GenBank, and two of those splice sites are primate-specific. The predicted translation of the gorilla transcript removes the first five amino acids of the human PPP1R3C protein.

Osteocrin (OSTN) is a peptide expressed in muscle and adipocytes that is regulated by insulin and inhibits glucose uptake (Moffatt and Thomas, 2009), and is therefore potentially related to metabolic disorders. Neither its expression in the brain nor its splicing in the gorilla had been investigated prior to our study. We have identified a unique upstream exon of this gene in the gorilla that lacks any public human transcriptome evidence. This osteocrin splice variant has no effect on the ORF.

Apolipoprotein H (*APOH*), which is elevated in type 2 diabetes (a disorder that is a known risk factor for late-onset Alzheimer's disease [for review see Dickstein et al., 2010]), is a marker of cardiovascular disease risk (Castro et al., 2010) and is mostly expressed in the liver. The gorilla data provide evidence of its expression in the brain and suggest that a novel transcription start site and two cryptic splice sites are utilized in gorilla. These sequences correspond to the 5'-UTR of human APOH and do not affect protein-coding exons.

PPARGC1A (co-activator 1-alpha of the peroxisome proliferator-activated receptor gamma) is the last of four metabolism-related genes with gorilla-specific splice isoforms in our data. The alignment of the gorilla transcript to the annotated human genome reveals a novel upstream exon and an upstream intron splice donor not supported by any human transcriptome data available at the time of our analysis, whereas the first intron's splice acceptor is shared with the canonical human first-intron splice acceptor of this gene. The ORF of the gorilla isoform contains eight amino acids of gorilla-specific N-terminal sequence (Table 1).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Endogenous antisense transcription and intergenic splicing events correspond to gorilla-specific gene structures

Six of the 13 genes with gorilla-specific gene structure features (RNF17, CATSPER2, CWC15, BTBD8, E2F1, and RFPL1S) are associated with one or both of two types of genomic structure complexity that are often not conserved among closely related species (Akiva et al., 2006; Lipovich et al., 2010 and references therein): intergenic splicing (i.e., the combination of transcripts from distinct genes) and endogenous *cis*-antisense (i.e., double-stranded, bidirectional) transcription. Two genes functional in transcriptional control and post-transcriptional regulation are affected by unique genomic structure complexity in the gorilla planum temporale. E2F1, a key transcription factor, is a cell cycle regulator that controls pancreatic beta cell proliferation (Fajas et al., 2010). An assembled transcript in the gorilla dataset represents intergenic splicing between the upstream gene PXMP4 (a peroxisome membrane protein) and E2F1, although the splicing event interrupts the ORFs of both genes, precluding the translation of a chimeric protein product. The CWC15 gene, homologous to a yeast gene encoding a spliceosome-associated protein and therefore putatively related to post-transcriptional gene expression regulation through a function in splicing, has two upstream 5'-UTR exons in the gorilla that are not supported by any human transcriptome data. The first of those two exons is *cis*-antisense to the KDM4DL gene, 55 kb upstream of CWC15 along the human genome.

DISCUSSION

Based on an analysis of genes expressed in the gorilla neocortex, we have shown gene structure differences between human and gorilla. These differences are evident from transcript-to-genome alignments, but not from genome sequences alone. Additionally, we have presented evidence that the gene structure differences have the potential to result in protein structure differences. We did this by describing previously unknown splicing events at 13 loci conserved in gorilla and human and placing these events into the broader context of potential human-gorilla interspecies gene structure differences. This number is a conservative lower bound, because our computational pipeline was built to identify differences at gene boundaries, rather than internal differences at cassette exons. Both terminal and internal differences in gene structure have the potential to affect protein sequence, and both should therefore be considered in future comparative interrogations of primate transcriptomes. Although assigning a biological role to these gorilla transcripts in the brain is challenging at present, their existence illustrates the potential of RNAseq-based studies.

We carried out this analysis because sequencing of transcriptomes from nonhuman primate brains facilitates inference of species-specific splicing and other regulatory events that may underlie functional differences between humans and our close phylogenetic relatives. Lineage-specific differences in these phenomena may contribute to the genomic basis of interspecies phenotypic distinctions because of their impact on the length and sequence of orthologous proteins, untranslated gene regions, and non-coding RNAs. Our analysis documents the actual genomic structures (promoters, exons, introns, and splice sites) of 10,511 protein-coding, and 1,870 noncoding-RNA, neocortex-expressed genes from a

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

western lowland gorilla that are alignable to their orthologous counterparts along the human genome.

Interspecies gene structure differences within primates, including gene boundaries as well as splice site usage, have not been described for any of the 13 genes we identified in the current study. Four of the 13 genes have functions in energy utilization and metabolism. This is notable because the primate brain is known for its high energy demand (Bauernfeind et al., 2014; Kuzawa et al., 2014) and has been accompanied by accelerated evolution of the human cortical metabolome (Bozek et al., 2014). Interestingly, studies have demonstrated positive selection operating on genes of the anthropoid electron transport chain (Grossman et al., 2004; Schmidt et al., 2005; Pierron et al., 2011), by comparative genomics and gene expression studies to human-specific evolutionary changes (Uddin et al., 2008), and by distinct expression patterns of genes that function in glucose metabolism and mitochondrial function (Uddin et al., 2004; Preuss et al., 2004; Cáceres et al., 2003).

One of the genes identified in our analyses with known functions in energy utilization is PPARGC1A, a regulator of mitochondrial biogenesis and oxidative metabolism. It is expressed in all tissues with high oxidative energy turnover as a metabolic regulator (Dominy et al., 2009). In brain, it activates reactive oxygen species detoxification (St-Pierre et al., 2006) and thus is neuroprotective (Róna-Vörös and Weydt, 2010; Helisalmi et al., 2008). The gorilla variant splices into the equivalent of human exon 2. It contains eight amino acids not encoded by known human transcripts. A pig ovary full-length cDNA (GenBank accession number: AK235681) recapitulates the gorilla isoform, suggesting a human loss (not a gorilla gain) in evolution, and adding 14 amino acids to this N-terminal extension. The 22 novel residues provide a potential new acetylation site consistent with the consensus for GCN5 (Basu et al., 2009), the major PGC-1α N-acetyltransferase. Our results indicate brain expression, and a unique promoter and first-exon genomic architecture, in the gorilla for this central regulator of mitochondrial energy metabolism (Spiegelman 2007; Rodgers et al., 2008).

We encountered several instances of complex loci (Engström et al., 2006) - places in the genome where multiple genes overlap each other in the same locus. The start of the gorilla GBF1 transcript sequence mapped to the protein-coding region of the last exon of the human PITX3 gene (paired-like homeodomain 3, a regulator of eye development), with GBF1 transcription taking place in the antisense orientation relative to PITX3. In view of the fact that pathogens including Ebola, SIV, and *Plasmodium falciparum* malaria can cross the species barrier, infecting both gorillas and humans (Liu et al., 2010; Duval et al., 2010; Genton et al., 2012; Neel et al., 2010), the human-gorilla distinctions of the viral replication accessory GBF1 are likely to contribute to the broader human-gorilla differences in infection susceptibility and to the respective histories of shared infections and pathogen transfers.

We have previously shown that endogenous antisense lncRNAs are key regulators of the protein-coding loci that they overlap, including specifically in the brain (Lipovich et al., 2012; Lipovich et al., 2010) and therefore the possibility that RFPL1S is a brain-specific regulator of RFPL1 warrants further investigation. *Cis*-antisense regulation of another brain-expressed gene, NEFH (neurofilament H) by RFPL1S is suggested by the unexpected

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

overlap of the gorilla-specific RFPL1S first exon and the NEFH N-terminal exon, both of which reside 30 kb away from the previously known human RFPL1S locus boundary. NEFH is a neuronal biomarker of brain injury (Siman et al., 2009) and of neurodegeneration in Alzheimer's disease (Hof et al., 1990; Bussière et al., 2003a,b; Hof and Morrison, 2004). The novel gorilla exon is absent from human RFPL1S but present in another nonhuman primate, the long-tailed macaque, *Macaca fascicularis* (GenBank accession: AB172344). RFPL1S does not encode a protein.

Human-gorilla gene structure differences have potentially diverse functional and evolutionary implications. The gorilla splice variant of HRH2 presented a unique C-terminus of 63 amino acids. HRH2 functions in the cell cycle (Xu et al., 2008). The discovery of its gorilla-specific splice variant may have implications for differences between primate species in cell cycle regulation, a factor that might be relevant to the up to fourfold variation in the incidence of spontaneous malignant tumors between closely related primate species (Takayama et al., 2008). TTLL13, a tubulin tyrosine ligase-like gene, has a unique upstream exon in its matching gorilla assembled transcript.

Another example of the implications of human-gorilla gene structure differences is provided by our structural model of the gorilla BTBD8 protein, which consists of four domains (Fig. 8). Its two central domains contain a BTB subdomain (Zollman et al., 1994; Adams et al., 2000). The C-terminal domain of BTBD8 has a BACK fold; the BACK domain is found in the majority of proteins that contain both a BTB domain and Kelch repeats (Stogios and Privé 2004), and is involved in ubiquitination. Hence, our results point to the possible existence of differences in protein ubiquitination between humans and gorillas. Interspecies gene structure differences that do not affect protein-coding sequence may also contribute to functional and phenotypic distinctions because they can lead to differential regulation of orthologous genes by RNA-binding proteins acting at 5'-UTRs and by microRNAs acting at 3'-UTRs.

Our approach of sequencing the gorilla planum temporale transcriptome enabled us to construct more accurate gene models in gorilla. Our results suggest that the gorilla planum temporale transcribes more than 40,000 RNA molecules. Deeper-coverage comparative sequencing of transcribed regions of the genomes of multiple primate species should help to characterize the expressed genes in these different regions accurately.

An unpublished directional pooled gorilla RNAseq dataset (Christopher E. Mason, 2014, personal communication) from the Non-Human Primate Transcriptome Resource (Pipes et al., 2013), yielded support for 8 of our 13 novel gorilla-specific exons, indicating that these splice events – despite their modest FPKM counts – are recurrent and detectable in other gorilla datasets and tissues. Furthermore, we mounted the unassembled, unstranded raw RNAseq data from the 11 gorilla samples of Brawand et al. (2011; NCBI SRA Bioproject 143627) in a custom instance of the UCSC Genome Browser. We witnessed clear support of our gorilla-specific exons in seven of our 13 cases (APOH, CATSPER2, HRH2, PPARGC1A, PPP1R3C, RNF17, and TTLL13) and complete absence of support of these exons at three loci. This further supports the conclusion that these splice events are recurrently detectable in independently derived datasets from other gorillas and diverse

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

tissues. As the remaining three loci involved intergenic splicing, we could not interpret the Brawand et al unstranded peaks for those genes without transcriptomic contigs. Also, five of the 13 genes matched interspecies conserved gene expression modules, according to the companion web resource of Brawand et al. (2011: http://www2.unil.ch/cgi-bin/cgiwrap/wwwcbg/devel/isaweb.py). Four genes resided in modules with brain and neuronal functions. BTBD8 and OSTN shared two modules (IDs: p169, p187) in human and chimpanzee, but not in the gorilla. This is intriguing with regard to the potential for human-gorilla splicing differences that we suggest at these loci. Even though OSTN lacks known function in the brain, it participates in six other modules pertaining to neuronal functions. CWC15 is found in three modules linked to neuron and oligodendrocyte functions in human but not in gorilla, whereas E2F1 is found in four modules, of which three are shared by human and gorilla.

The availability of appropriate resources for nonhuman primate transcriptome sequencing, including snap-frozen, high-quality postmortem brain specimens from rare primate species, is extremely limited. Therefore, our effort represents an important contribution to the literature in this field, even though we only analyzed the brain of a single gorilla.

It is tempting to speculate that the unique upstream exons we have detected at several genes in this gorilla transcriptome survey correspond to unique transcription start sites, and therefore potentially to promoters that are utilized specifically or selectively in the gorilla. However, several limitations preclude a definitive test of this hypothesis. First, our transcriptome sampling protocol does not distinguish between transcripts that are, or are not, full-length. Second, lack of comparable data for other primates or, indeed, sequences at comparable depth specifically for the human planum temporale, leaves open questions such as whether the unique upstream exons absence in humans is due to a human-specific loss taking place on the human terminal branch in evolution, or to a gorilla-specific gain. Third, we derived these results from a single gorilla.

We make the assumption that no changes other than substitutions and very small indels distinguish the human and gorilla genome regions containing the orthologous genes in our study. Although large-scale changes, generated by segmental duplications and genomic rearrangements, are relatively rare, this assumption may impact the interpretation of gorilla transcript to human genome alignments at regions that have undergone recent large-scale genomic remodeling.

Although orthologous genes may retain similar functions in multiple closely related species, minor changes in their post-transcriptional regulation or encoded protein, such as the changes that can be inferred from interspecies orthologous gene structure differences, may in fact reflect protein function or gene regulation differences that separate one species from another. Even gene structure differences that do not affect protein sequence may carry a functional load because they may result in different patterns of post-transcriptional regulation in humans and gorillas. These differences, which we have detected from the alignment of a nonhuman primate transcriptome to the human genome, may represent molecular fossils of adaptive changes accompanying past speciation events that ultimately gave rise to humans and other great apes.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Acknowledgments

Other Acknowledgments. We thank Michael J. Milligan and Jason Caravas for technical assistance with analysis of RNAseq datasets, and Dr. Joseph M. Erwin for his help in making the study of the gorilla brain specimen possible through the Great Ape Aging Project. We are indebted to Drs. Daniel C. Anderson (Yerkes National Primate Research Center) and Rita McManamon (University of Georgia College of Veterinary Medicine) for providing access to the postmortem brain specimen and pathological documentation of the case.

Grant Sponsor: The James S. McDonnell Foundation; Grant numbers: 22002078 and 220020293. Grant Sponsor: National Science Foundation; Grant numbers: BCS 0827546, BCS 0827531, BCS 0550209, and the XSEDE supercomputing cluster (MCB 120116). Grant Sponsor: National Institutes of Health; Grant numbers: AG014308, GM069840, NS076465, and Office of Research Infrastructure Programs (ORIP) R24 RR032341 and ORIP/OD P51 OD011132.

LITERATURE CITED

Adams J, Kelso R, Cooley L. The kelch repeat superfamily of proteins: propellers of cell function. Trends Cell Biol. 2000; 10:17–24. [PubMed: 10603472]

Akiva P, Toporik A, Edelheit S, Peretz Y, Diber A, Shemesh R, Novik A, Sorek R. Transcription-mediated gene fusion in the human genome. Genome Res. 2006; 16:30–36. [PubMed: 16344562]

Bae BI, Tietjen I, Atabay KD, Evrony GD, Johnson MB, Asare E, Wang PP, Murayama AY, Im K, Lisgo SN, Overman L, Šestan N, Chang BS, Barkovich AJ, Grant PE, Topçu M, Politsky J, Okano H, Piao X, Walsh CA. Evolutionarily dynamic alternative splicing of GPR56 regulates regional cerebral cortical patterning. Science. 2014; 343:764–768. [PubMed: 24531968]

Basu A, Rose KL, Zhang J, Beavis RC, Ueberheide B, Garcia BA, Chait B, Zhao Y, Hunt DF, Segal E, Allis CD, Hake SB. Proteome-wide prediction of acetylation substrates. Proc Natl Acad Sci U S A. 2009; 106:13785–13790. [PubMed: 19666589]

Bauernfeind AL, Barks SK, Duka T, Grossman LI, Hof PR, Sherwood CC. Aerobic glycolysis in the primate brain: reconsidering the implications for growth and maintenance. Brain Struct Funct. 2014; 219:1149–1167. [PubMed: 24185460]

Belov GA, Kovtunovych G, Jackson CL, Ehrenfeld E. Poliovirus replication requires the N-terminus but not the catalytic Sec7 domain of ArfGEF GBF1. Cell Microbiol. 2010; 12:1463–1479. [PubMed: 20497182]

Birney E. ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007; 447:799–816. [PubMed: 17571346]

Blekhman R, Marioni JC, Zumbo P, Stephens M, Gilad Y. Sex-specific and lineage- specific alternative splicing in primates. Genome Res. 2010; 20:180–189. [PubMed: 20009012]

Bozek K, Wei Y, Yan Z, Liu X, Xiong J, Sugimoto M, Tomita M, Pääbo S, Pieszek R, Sherwood CC, Hof PR, Ely JJ, Steinhauser D, Willmitzer L, Bangsbo J, Hansson O, Call J, Giavalisco P, Khaitovich P. Exceptional evolutionary divergence of human muscle and brain metabolomes parallels human cognitive and physical uniqueness. PLoS Biol. 2014; 12(5):e1001871. [PubMed: 24866127]

Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, Albert FW, Zeller U, Khaitovich P, Grützner F, Bergmann S, Nielsen R, Pääbo S, Kaessmann H. The evolution of gene expression levels in mammalian organs. Nature. 2011; 478:343–348. [PubMed: 22012392]

Bussière T, Gold G, Kövari E, Giannakopoulos P, Bouras C, Perl DP, Morrison JH, Hof PR. Stereologic analysis of neurofibrillary tangle formation in prefrontal cortex area 9 in aging and Alzheimer's disease. Neuroscience. 2003a; 117:577–592. [PubMed: 12617964]

Bussière T, Giannakopoulos P, Bouras C, Perl DP, Morrison JH, Hof PR. Progressive degeneration of nonphosphorylated neurofilament protein-enriched pyramidal neurons predicts cognitive impairment in Alzheimer's disease: a stereologic analysis of prefrontal cortex area 9. J Comp Neurol. 2003b; 463:281–302. [PubMed: 12820162]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Cáceres M, Lachuer J, Zapala MA, Redmond JC, Kudo L, Geschwind DH, Lockhart DJ, Preuss TM, Barlow C. Elevated gene expression levels distinguish human from nonhuman primate brains. Proc Natl Acad Sci USA. 2003; 100:13030–13035. [PubMed: 14557539]

Calarco JA, Xing Y, Cáceres M, Calarco JP, Xiao X, Pan Q, Lee C, Preuss TM, Blencowe BJ. Global analysis of alternative splicing differences between humans and chimpanzees. Genes Dev. 2007; 21:2963–2975. [PubMed: 17978102]

Canning P, Bullock AN. New strategies to inhibit KEAP1 and the Cul3-based E3 ubiquitin ligases. Biochem Soc Trans. 2014; 42:103–107. [PubMed: 24450635]

Castro A, Lázaro I, Selva DM, Céspedes E, Girona J, NúriaPlana, Guardiola M, Cabré A, Simó R, Masana L. APOH is increased in the plasma and liver of type 2 diabetic patients with metabolic syndrome. Atherosclerosis. 2010; 209:201–205. [PubMed: 19878946]

Citterio C, Vichi A, Pacheco-Rodriguez G, Aponte AM, Moss J, Vaughan M. Unfolded protein response and cell death after depletion of brefeldin A-inhibited guanine nucleotide-exchange protein GBF1. Proc Natl Acad Sci U S A. 2007; 105:2877–2882. [PubMed: 18287014]

Clark NL, Gasper J, Sekino M, Springer SA, Aquadro CF, Swanson WJ. Coevolution of interacting fertilization proteins. PLoS Genet. 2009; 5:e1000570. [PubMed: 19629160]

Dickstein DL, Walsh J, Brautigam H, Stockton SD Jr, Gandy S, Hof PR. Role of vascular risk factors and vascular dysfunction in Alzheimer's disease. Mt Sinai J Med. 2010; 77:82–102. [PubMed: 20101718]

Dominy JE Jr, Lee Y, Gerhart-Hines Z, Puigserver P. Nutrient-dependent regulation of PGC-1alpha's acetylation state and metabolic function through the enzymatic activities of Sirt1/GCN5. Biochim Biophys Acta. 2010; 1804:1676–1683. [PubMed: 20005308]

Duval L, Fourment M, Nerrienet E, Rousset D, Sadeuh SA, Goodman SM, Andriaholinirina NV, Randrianarivelojosia M, Paul RE, Robert V, Ayala FJ, Ariey F. African apes as reservoirs of *Plasmodium falciparum* and the origin and diversification of the Laverania subgenus. Proc Natl Acad Sci USA. 2010; 107:10561–10566. [PubMed: 20498054]

Enard W, Khaitovich P, Klose J, Zöllner S, Heissig F, Giavalisco P, Nieselt-Struwe K, Muchmore E, Varki A, Ravid R, Doxiadis GM, Bontrop RE, Pääbo S. Intra- and interspecific variation in primate gene expression patterns. Science. 2002; 296:340–343. [PubMed: 11951044]

Engström PG, Suzuki H, Ninomiya N, Akalin A, Sessa L, Lavorgna G, Brozzi A, Luzi L, Tan SL, Yang L, Kunarso G, Ng EL, Batalov S, Wahlestedt C, Kai C, Kawai J, Carninci P, Hayashizaki Y, Wells C, Bajic VB, Orlando V, Reid JF, Lenhard B, Lipovich L. Complex loci in human and mouse genomes. PLoS Genet. 2006; 2:e47. [PubMed: 16683030]

Fajas L, Blanchet E, Annicotte JS. The CDK4-pRB-E2F1 pathway: a new modulator of insulin secretion. Islets. 2010; 2:51–53. [PubMed: 21099295]

Farimani AB, Min K, Aluru NR. DNA base detection using a single-layer MoS2. ACS Nano. 2014; 8:7914–7922. [PubMed: 25007098]

Gannon PJ, Holloway RL, Broadfield DC, Braun AR. Asymmetry of chimpanzee planum temporale: humanlike pattern of Wernicke's brain language area homolog. Science. 1998; 279:220–222. [PubMed: 9422693]

Genton C, Cristescu R, Gatti S, Levréro F, Bigot E, Caillaud D, Pierre JS, Ménard N. Recovery potential of a western lowland gorilla population following a major Ebola outbreak: results from a ten year study. PLoS One. 2012; 7:e37106. [PubMed: 22649511]

Goodman M, Porter CA, Czelusniak J, Page SL, Schneider H, Shoshani J, Gunnell G, Groves CP. Toward a phylogenetic classification of Primates based on DNA evidence complemented by fossil evidence. Mol Phylogenet Evol. 1998; 9:585–598. [PubMed: 9668008]

Goueslain L, Alsaleh K, Horellou P, Roingeard P, Descamps V, Duverlie G, Ciczora Y, Wychowski C, Dubuisson J, Rouillé Y. Identification of GBF1 as a cellular factor required for hepatitis C virus RNA replication. J Virol. 2010; 84:773–787. [PubMed: 19906930]

Schmidt TR, Wildman DE, Uddin M, Opazo JC, Goodman M, Grossman LI. Accelerated evolution of the electron transport chain in anthropoid primates. Trends Genet. 2004; 20:578–585. [PubMed: 15475118]

Gumucio DL, Lockwood WK, Weber JL, Saulino AM, Delgrosso K, Surrey S, Schwartz E, Goodman M, Collins FS. The -175T----C mutation increases promoter strength in erythroid cells: correlation

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

with evolutionary conservation of binding sites for two trans-acting factors. Blood. 1990; 75:756–761. [PubMed: 1688723]

Gumucio DL, Shelton DA, Blanchard-McQuate K, Gray T, Tarle S, Heilstedt-Williamson H, Slightom JL, Collins F, Goodman M. Differential phylogenetic footprinting as a means to identify base changes responsible for recruitment of the anthropoid gamma gene to a fetal expression pattern. J Biol Chem. 1994; 269:15371–15380. [PubMed: 7515056]

Hawrylycz MJ, Lein ES, Guillozet-Bongaarts AL, Shen EH, Ng L, Miller JA, van de Lagemaat LN, Smith KA, Ebbert A, Riley ZL, Abajian C, Beckmann CF, Bernard A, Bertagnolli D, Boe AF, Cartagena PM, Chakravarty MM, Chapin M, Chong J, Dalley RA, Daly BD, Dang C, Datta S, Dee N, Dolbeare TA, Faber V, Feng D, Fowler DR, Goldy J, Gregor BW, Haradon Z, Haynor DR, Hohmann JG, Horvath S, Howard RE, Jeromin A, Jochim JM, Kinnunen M, Lau C, Lazarz ET, Lee C, Lemon TA, Li L, Li Y, Morris JA, Overly CC, Parker PD, Parry SE, Reding M, Royall JJ, Schulkin J, Sequeira PA, Slaughterbeck CR, Smith SC, Sodt AJ, Sunkin SM, Swanson BE, Vawter MP, Williams D, Wohnoutka P, Zielke HR, Geschwind DH, Hof PR, Smith SM, Koch C, Grant SG, Jones AR. An anatomically comprehensive atlas of the adult human brain transcriptome. Nature. 2012; 489:391–399. [PubMed: 22996553]

Helisalmi S, Vepsäläinen S, Hiltunen M, Koivisto AM, Salminen A, Laakso M, Soininen H. Genetic study between SIRT1, PPARD, PGC-1alpha genes and Alzheimer's disease. J Neurol. 2008; 255:668–673. [PubMed: 18438697]

Hildebrand MS, Avenarius MR, Fellous M, Zhang Y, Meyer NC, Auer J, Serres C, Kahrizi K, Najmabadi H, Beckmann JS, Smith RJ. Genetic male infertility and mutation of CATSPER ion channels. Eur J Hum Genet. 2010; 18:1178–1184. [PubMed: 20648059]

Hof PR, Cox K, Morrison JH. Quantitative analysis of a vulnerable subset of pyramidal neurons in Alzheimer's disease: I. Superior frontal and inferior temporal cortex. J Comp Neurol. 1990; 301:44–54. [PubMed: 2127598]

Hof PR, Morrison JH. The aging brain: morphomolecular senescence of cortical circuits. Trends Neurosci. 2004; 27:607–613. [PubMed: 15374672]

Hopkins WD, Marino L, Rilling JK, MacGregor LA. Planum temporale asymmetries in great apes as revealed by magnetic resonance imaging (MRI). Neuroreport. 1998; 9:2913–2918. [PubMed: 9760145]

Jia H, Osak M, Bogu GK, Stanton L, Johnson R, Lipovich L. Genome-wide computational identification and manual annotation of human long noncoding RNA genes. RNA. 2010; 16:1478–1487. [PubMed: 20587619]

Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL. NCBI BLAST: a better web interface. Nucleic Acids Res. 2008; 36:W5–W9. [PubMed: 18440982]

Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, Li M, Sousa AM, Pletikos M, Meyer KA, Sedmak G, Guennel T, Shin Y, Johnson MB, Krsnik Z, Mayer S, Fertuzinhos S, Umlauf S, Lisgo SN, Vortmeyer A, Weinberger DR, Mane S, Hyde TM, Huttner A, Reimers M, Kleinman JE, Sestan N. Spatio-temporal transcriptome of the human brain. Nature. 2011; 478:483–489. [PubMed: 22031440]

Kent WJ. BLAT -- The BLAST-Like Alignment Tool. Genome Res. 2002a; 4:656–664. [PubMed: 11932250]

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res. 2002b; 12:996–1006. [PubMed: 12045153]

Khaitovich P, Hellmann I, Enard W, Nowick K, Leinweber M, Franz H, Weiss G, Lachmann M, Pääbo S. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. Science. 2005; 309:1850–1854. [PubMed: 16141373]

King MC, Wilson AC. Evolution at two levels in humans and chimpanzees. Science. 1975; 188:107–116. [PubMed: 1090005]

Konopka G, Friedrich T, Davis-Turak J, Winden K, Oldham MC, Gao F, Chen L, Wang GZ, Luo R, Preuss TM, Geschwind DH. Human-specific transcriptional networks in the brain. Neuron. 2012; 75:601–617. [PubMed: 22920253]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Kuzawa CW, Chugani HT, Grossman LI, Lipovich L, Muzik O, Hof PR, Wildman DE, Sherwood CC, Leonard WR, Lange N. Metabolic costs and evolutionary implications of human brain development. Proc Natl Acad Sci USA. 2014; 111:13010–13015. [PubMed: 25157149]

Lander ES. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature. 2001; 409:860–921. [PubMed: 11237011]

Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009; 10:R25. [PubMed: 19261174]

Lee TM, Lipovich L. Structural differences of orthologous genes: Insights from human- primate comparisons. Genomics. 2008; 92:134–143. [PubMed: 18606524]

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, Yang H, Wang J. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. 2010; 20:265–272. [PubMed: 20019144]

Li Y, Qian YP, Yu XJ, Wang YQ, Dong DG, Sun W, Ma RM, Su B. Recent origin of a hominoid-specific splice form of neuropsin, a gene involved in learning and memory. Mol Biol Evol. 2004; 21:2111–2115. [PubMed: 15282331]

Lindblad-Toh K. Broad Institute Sequencing Platform and Whole Genome Assembly Team, Baylor College of Medicine Human Genome Sequencing Center Sequencing Team, Genome Institute at Washington University. A high-resolution map of human evolutionary constraint using 29 mammals. Nature. 2011; 478:476–482. [PubMed: 21993624]

Lipovich L, Johnson R, Lin CY. MacroRNA underdogs in a microRNA world: evolutionary, regulatory, and biomedical significance of mammalian long non-protein-coding RNA. Biochim Biophys Acta. 2010; 1799:597–615. [PubMed: 20951849]

Lipovich L, Dachet F, Cai J, Bagla S, Balan K, Jia H, Loeb JA. Activity-dependent human brain coding/noncoding gene regulatory networks. Genetics. 2012; 192:1133–1148. [PubMed: 22960213]

Liu W, Li Y, Learn GH, Rudicell RS, Robertson JD, Keele BF, Ndjango JB, Sanz CM, Morgan DB, Locatelli S, Gonder MK, Kranzusch PJ, Walsh PD, Delaporte E, Mpoudi-Ngole E, Georgiev AV, Muller MN, Shaw GM, Peeters M, Sharp PM, Rayner JC, Hahn BH. Origin of the human malaria parasite *Plasmodium falciparum* in gorillas. Nature. 2010; 467:420–425. [PubMed: 20864995]

Liu X, Somel M, Tang L, Yan Z, Jiang X, Guo S, Yuan Y, He L, Oleksiak A, Zhang Y, Li N, Hu Y, Chen W, Qiu Z, Pääbo S, Khaitovich P. Extension of cortical synaptic development distinguishes humans from chimpanzees and macaques. Genome Res. 2012; 22:611–622. [PubMed: 22300767]

Maitra RD, Kim J, Dunbar WB. Recent advances in nanopore sequencing. Electrophoresis. 2012; 33:3418–3428. [PubMed: 23138639]

Margulies EH, Cooper GM, Asimenos G, Thomas DJ, Dewey CN, Siepel A, Birney E, Keefe D, Schwartz AS, Hou M, Taylor J, Nikolaev S, Montoya-Burgos JI, Löytynoja A, Whelan S, Pardi F, Massingham T, Brown JB, Bickel P, Holmes I, Mullikin JC, Ureta-Vidal A, Paten B, Stone EA, Rosenbloom KR, Kent WJ, Bouffard GG, Guan X, Hansen NF, Idol JR, Maduro VV, Maskeri B, McDowell JC, Park M, Thomas PJ, Young AC, Blakesley RW, Muzny DM, Sodergren E, Wheeler DA, Worley KC, Jiang H, Weinstock GM, Gibbs RA, Graves T, Fulton R, Mardis ER, Wilson RK, Clamp M, Cuff J, Gnerre S, Jaffe DB, Chang JL, Lindblad-Toh K, Lander ES, Hinrichs A, Trumbower H, Clawson H, Zweig A, Kuhn RM, Barber G, Harte R, Karolchik D, Field MA, Moore RA, Matthewson CA, Schein JE, Marra MA, Antonarakis SE, Batzoglou S, Goldman N, Hardison R, Haussler D, Miller W, Pachter L, Green ED, Sidow A. Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. Genome Res. 2007; 17:760–774. [PubMed: 17567995]

Marques-Bonet T, Girirajan S, Eichler EE. The origins and impact of primate segmental duplications. Trends Genet. 2009; 25:443–454. [PubMed: 19796838]

McFarlin SC, Barks SK, Tocheri MW, Massey JS, Eriksen AB, Fawcett KA, Stoinski TS, Hof PR, Bromage TG, Mudakikwa A, Cranfield MR, Sherwood CC. Early Brain Growth Cessation in Wild Virunga Mountain Gorillas (*Gorilla beringei beringei*). Am J Primatol. 2013; 75:450–463. [PubMed: 23208801]

Meredith RW, Janečka JE, Gatesy J, Ryder OA, Fisher CA, Teeling EC, Goodbla A, Eizirik E, Simão TL, Stadler T, Rabosky DL, Honeycutt RL, Flynn JJ, Ingram CM, Steiner C, Williams TL, Robinson TJ, Burk-Herrick A, Westerman M, Ayoub NA, Springer MS, Murphy WJ. Impacts of

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. Science. 2011; 334:521–524. [PubMed: 21940861]

Mi H, Muruganujan A, Casagrande JT, Thomas PD. Large-scale gene function analysis with the PANTHER classification system. Nat Protoc. 2013; 8:1551–1566. [PubMed: 23868073]

Miller DJ, Duka T, Stimpson CD, Schapiro SJ, Baze WB, McArthur MJ, Fobbs AJ, Sousa AM, Sestan N, Wildman DE, Lipovich L, Kuzawa CW, Hof PR, Sherwood CC. Prolonged myelination in human neocortical evolution. Proc Natl Acad Sci USA. 2012; 109:16480–16485. [PubMed: 23012402]

Miura S, Kai Y, Kamei Y, Ezaki O. Isoform-specific increases in murine skeletal muscle peroxisome proliferator-activated receptor-gamma coactivator-1alpha (PGC-1α) mRNA in response to β2-adrenergic receptor activation and exercise. Endocrinology. 2008; 149:4527–4533. [PubMed: 18511502]

Miura P, Shenker S, Andreu-Agullo C, Westholm JO, Lai EC. Widespread and extensive lengthening of 3' UTRs in the mammalian brain. Genome Res. 2013; 23:812–825. [PubMed: 23520388]

Moffatt P, Thomas GP. Osteocrin--beyond just another bone protein? Cell Mol Life Sci. 2009; 66:1135–1139. [PubMed: 19234809]

Neel C, Etienne L, Li Y, Takehisa J, Rudicell RS, Bass IN, Moudindo J, Mebenga A, Esteban A, Van Heuverswyn F, Liegeois F, Kranzusch PJ, Walsh PD, Sanz CM, Morgan DB, Ndjango JB, Plantier JC, Locatelli S, Gonder MK, Leendertz FH, Boesch C, Todd A, Delaporte E, Mpoudi-Ngole E, Hahn BH, Peeters M. Molecular epidemiology of simian immunodeficiency virus infection in wild-living gorillas. J Virol. 2010; 84:1464–1476. [PubMed: 19906908]

Ota T, Suzuki Y, Nishikawa T, Otsuki T, Sugiyama T, Irie R, Wakamatsu A, Hayashi K, Sato H, Nagai K, Kimura K, Makita H, Sekine M, Obayashi M, Nishi T, Shibahara T, Tanaka T, Ishii S, Yamamoto J, Saito K, Kawai Y, Isono Y, Nakamura Y, Nagahari K, Murakami K, Yasuda T, Iwayanagi T, Wagatsuma M, Shiratori A, Sudo H, Hosoiri T, Kaku Y, Kodaira H, Kondo H, Sugawara M, Takahashi M, Kanda K, Yokoi T, Furuya T, Kikkawa E, Omura Y, Abe K, Kamihara K, Katsuta N, Sato K, Tanikawa M, Yamazaki M, Ninomiya K, Ishibashi T, Yamashita H, Murakawa K, Fujimori K, Tanai H, Kimata M, Watanabe M, Hiraoka S, Chiba Y, Ishida S, Ono Y, Takiguchi S, Watanabe S, Yosida M, Hotuta T, Kusano J, Kanehori K, Takahashi-Fujii A, Hara H, Tanase TO, Nomura Y, Togiya S, Komai F, Hara R, Takeuchi K, Arita M, Imose N, Musashino K, Yuuki H, Oshima A, Sasaki N, Aotsuka S, Yoshikawa Y, Matsunawa H, Ichihara T, Shiohata N, Sano S, Moriya S, Momiyama H, Satoh N, Takami S, Terashima Y, Suzuki O, Nakagawa S, Senoh A, Mizoguchi H, Goto Y, Shimizu F, Wakebe H, Hishigaki H, Watanabe T, Sugiyama A, Takemoto M, Kawakami B, Yamazaki M, Watanabe K, Kumagai A, Itakura S, Fukuzumi Y, Fujimori Y, Komiyama M, Tashiro H, Tanigami A, Fujiwara T, Ono T, Yamada K, Fujii Y, Ozaki K, Hirao M, Ohmori Y, Kawabata A, Hikiji T, Kobatake N, Inagaki H, Ikema Y, Okamoto S, Okitani R, Kawakami T, Noguchi S, Itoh T, Shigeta K, Senba T, Matsumura K, Nakajima Y, Mizuno T, Morinaga M, Sasaki M, Togashi T, Oyama M, Hata H, Watanabe M, Komatsu T, Mizushima-Sugano J, Satoh T, Shirai Y, Takahashi Y, Nakagawa K, Okumura K, Nagase T, Nomura N, Kikuchi H, Masuho Y, Yamashita R, Nakai K, Yada T, Nakamura Y, Ohara O, Isogai T, Sugano S. Complete sequencing and characterization of 21,243 full-length human cDNAs. Nat Genet. 2004; 3:40–45. [PubMed: 14702039]

Ozboyaci M, Gursoy A, Erman B, Keskin O. Molecular recognition of H3/H4 histone tails by the tudor domains of JMJD2A: a comparative molecular dynamics simulations study. PLoS One. 2011; 6:e14765. [PubMed: 21464980]

Pan J, Goodheart M, Chuma S, Nakatsuji N, Page DC, Wang PJ. RNF17, a component of the mammalian germ cell nuage, is essential for spermiogenesis. Development. 2005; 132:4029–4039. [PubMed: 16093322]

Perez SE, Raghanti MA, Hof PR, Kramer L, Ikonomovic MD, Lacor PN, Erwin JM, Sherwood CC, Mufson EJ. Alzheimer¹s disease pathology in the neocortex and hippocampus of the Western lowland gorilla (*Gorilla gorilla gorilla*). J Comp Neurol. 2013; 521:4318–4338. [PubMed: 23881733]

Pierron D, Opazo JC, Heiske M, Papper Z, Uddin M, Chand G, Wildman DE, Romero R, Goodman M, Grossman LI. Silencing, positive selection and parallel evolution: busy history of primate cytochromes c. PLoS One. 2011; 6:e26269. [PubMed: 22028846]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Pipes L, Li S, Bozinoski M, Palermo R, Peng X, Blood P, Kelly S, Weiss JM, Thierry-Mieg J, Thierry-Mieg D, Zumbo P, Chen R, Schroth GP, Mason CE, Katze MG. The non-human primate reference transcriptome resource (NHPRTR) for comparative functional genomics. Nucleic Acids Res. 2013; 41:D906–D914. [PubMed: 23203872]

Pozzi L, Hodgson JA, Burrell AS, Sterner KN, Raaum RL, Disotell TR. Primate phylogenetic relationships and divergence dates inferred from complete mitochondrial genomes. Mol Phylogenet Evol. 2014; 75:165–183. [PubMed: 24583291]

Preuss TM, Cáceres M, Oldham MC, Geschwind DH. Human brain evolution: insights from microarrays. Nat Rev Genet. 2004; 5:850–860. [PubMed: 15520794]

Rivest S. Regulation of innate immune responses in the brain. Nat Rev Immunol. 2009; 9:429– 439. [PubMed: 19461673]

Roberts TC, Morris KV, Wood MJ. The role of long non-coding RNAs in neurodevelopment, brain function and neurological disease. Philos Trans R Soc Lond B Biol Sci. 2014; 369:1652.

Rodgers JT, Lerin C, Haas W, Gygi SP, Spiegelman BM, Puigserver P. Nutrient control of glucose homeostasis through a complex of PGC-1alpha and SIRT1. Nature. 2005; 434:113–118. [PubMed: 15744310]

Rodgers JT, Lerin C, Gerhart-Hines Z, Puigserver P. Metabolic adaptations through the PGC-1 alpha and SIRT1 pathways. FEBS Lett. 2008; 582:46–53. [PubMed: 18036349]

Róna-Vörös K, Weydt P. The role of PGC-1α in the pathogenesis of neurodegenerative disorders. Curr Drug Targets. 2010; 11:1262–1269. [PubMed: 20840068]

Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, Herrero J, Hobolth A, Lappalainen T, Mailund T, Marques-Bonet T, McCarthy S, Montgomery SH, Schwalie PC, Tang YA, Ward MC, Xue Y, Yngvadottir B, Alkan C, Andersen LN, Ayub Q, Ball EV, Beal K, Bradley BJ, Chen Y, Clee CM, Fitzgerald S, Graves TA, Gu Y, Heath P, Heger A, Karakoc E, Kolb-Kokocinski A, Laird GK, Lunter G, Meader S, Mort M, Mullikin JC, Munch K, O'Connor TD, Phillips AD, Prado-Martinez J, Rogers AS, Sajjadian S, Schmidt D, Shaw K, Simpson JT, Stenson PD, Turner DJ, Vigilant L, Vilella AJ, Whitener W, Zhu B, Cooper DN, de Jong P, Dermitzakis ET, Eichler EE, Flicek P, Goldman N, Mundy NI, Ning Z, Odom DT, Ponting CP, Quail MA, Ryder OA, Searle SM, Warren WC, Wilson RK, Schierup MH, Rogers J, Tyler-Smith C, Durbin R. Insights into hominid evolution from the gorilla genome sequence. Nature. 2012; 483:169–175. [PubMed: 22398555]

Schmidt TR, Wildman DE, Uddin M, Opazo JC, Goodman M, Grossman LI. Rapid electrostatic evolution at the binding site for cytochrome *c* on cytochrome *c* oxidase in anthropoid primates. Proc Natl Acad Sci USA. 2005; 102:6379–6384. [PubMed: 15851671]

Seroussi E, Kedra D, Pan HQ, Peyrard M, Schwartz C, Scambler P, Donnai D, Roe BA, Dumanski JP. Duplications on human chromosome 22 reveal a novel Ret Finger Protein-like gene family with sense and endogenous antisense transcripts. Genome Res. 1999; 9:803–814. [PubMed: 10508838]

Shen GM, Zhang FL, Liu XL, Zhang JW. Hypoxia-inducible factor 1-mediated regulation of PPP1R3C promotes glycogen accumulation in human MCF-7 cells under hypoxia. FEBS Lett. 2010; 584:4366–4372. [PubMed: 20888814]

Sherwood CC, Cranfield MR, Mehlman PT, Lilly AA, Garbe JA, Whittier CA, Nutter FB, Rein TR, Bruner HJ, Holloway RL, Tang CY, Naidich TP, Delman BN, Steklis HD, Erwin JM, Hof PR. Brain structure variation in great apes, with attention to the mountain gorilla (*Gorilla beringei beringei*). Am J Primatol. 2004; 63:149–164. [PubMed: 15258959]

Siman R, Toraskar N, Dang A, McNeil E, McGarvey M, Plaum J, Maloney E, Grady MS. A panel of neuron-enriched proteins as markers for traumatic brain injury in humans. J Neurotrauma. 2009; 26:1867–1877. [PubMed: 19811094]

Siomi MC, Mannen T, Siomi H. How does the royal family of Tudor rule the PIWI- interacting RNA pathway? Genes Dev. 2010; 24:636–646. [PubMed: 20360382]

Spiegelman BM. Transcriptional control of mitochondrial energy metabolism through the PGC1 coactivators. Novartis Found Symp. 2007; 287:60–63. discussion 63–69. [PubMed: 18074631]

St-Pierre J, Drori S, Uldry M, Silvaggi JM, Rhee J, Jäger S, Handschin C, Zheng K, Lin J, Yang W, Simon DK, Bachoo R, Spiegelman BM. Suppression of reactive oxygen species and

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

neurodegeneration by the PGC-1 transcriptional coactivators. Cell. 2006; 127:397–408. [PubMed: 17055439]

Stogios PJ, Privé GG. The BACK domain in BTB-kelch proteins. Trends Biochem Sci. 2004; 29:634–637. [PubMed: 15544948]

Tadaishi M, Miura S, Kai Y, Kawasaki E, Koshinaka K, Kawanaka K, Nagata J, Oishi Y, Ezaki O. Effect of exercise intensity and AICAR on isoform-specific expressions of murine skeletal muscle PGC-1α mRNA: a role of β_2-adrenergic receptor activation. Am J Physiol Endocrinol Metab. 2011; 300:E341–E349. [PubMed: 21098736]

Takayama S, Thorgeirsson UP, Adamson RH. Chemical carcinogenesis studies in nonhuman primates. Proc Jpn Acad Ser B Phys Biol Sci. 2008; 84:176–178.

Takeda J, Suzuki Y, Nakao M, Barrero RA, Koyanagi KO, Jin L, Motono C, Hata H, Isogai T, Nagai K, Otsuki T, Kuryshev V, Shionyu M, Yura K, Go M, Thierry-Mieg J, Thierry-Mieg D, Wiemann S, Nomura N, Sugano S, Gojobori T, Imanishi T. Large-scale identification and characterization of alternative splicing variants of human gene transcripts using 56,419 completely sequenced and manually annotated full-length cDNAs. Nucleic Acids Res. 2006; 34:3917–3928. [PubMed: 16914452]

Tatusova TA, Madden TL. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbiol Lett. 1999; 174:247–250. [PubMed: 10339815]

Tay SK, Blythe J, Lipovich L. Global discovery of primate-specific genes in the human genome. Proc Natl Acad Sci USA. 2009; 106:12019–12024. [PubMed: 19581580]

Temple G. MGC Project Team. The completion of the Mammalian Gene Collection (MGC). Genome Res. 2009; 19:2324–2333. [PubMed: 19767417]

Thompson CL, Ng L, Menon V, Martinez S, Lee CK, Glattfelder K, Sunkin SM, Henry A, Lau C, Dang C, Garcia-Lopez R, Martinez-Ferre A, Pombero A, Rubenstein JL, Wakeman WB, Hohmann J, Dee N, Sodt AJ, Young R, Smith K, Nguyen TN, Kidney J, Kuan L, Jeromin A, Kaykas A, Miller J, Page D, Orta G, Bernard A, Riley Z, Smith S, Wohnoutka P, Hawrylycz MJ, Puelles L, Jones AR. A high-resolution spatiotemporal atlas of gene expression of the developing mouse brain. Neuron. 2014; 83:309–323. [PubMed: 24952961]

Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009; 25:1105–1111. [PubMed: 19289445]

Uddin M, Wildman DE, Liu G, Xu W, Johnson RM, Hof PR, Kapatos G, Grossman LI, Goodman M. Sister grouping of chimpanzees and humans as revealed by genome-wide phylogenetic analysis of brain gene expression profiles. Proc Natl Acad Sci USA. 2004; 101:2957–2962. [PubMed: 14976249]

Uddin M, Opazo JC, Wildman DE, Sherwood CC, Hof PR, Goodman M, Grossman LI. Molecular evolution of the cytochrome *c* oxidase subunit 5A gene in primates. BMC Evol Biol. 2008; 8:8. [PubMed: 18197981]

van der Linden L, van der Schaar HM, Lanke KH, Neyts J, van Kuppeveld FJ. Differential effects of the putative GBF1 inhibitors Golgicide A and AG1478 on enterovirus replication. J Virol. 2010; 84:7535–7542. [PubMed: 20504936]

Warnefors M, Kaessmann H. Evolution of the correlation between expression divergence and protein divergence in mammals. Genome Biol Evol. 2013; 5:1324–1335. [PubMed: 23781097]

Watanabe H, Fujiyama A, Hattori M, Taylor TD, Toyoda A, Kuroki Y, Noguchi H, BenKahla A, Lehrach H, Sudbrak R, Kube M, Taenzer S, Galgoczy P, Platzer M, Scharfe M, Nordsiek G, Blöcker H, Hellmann I, Khaitovich P, Pääbo S, Reinhardt R, Zheng HJ, Zhang XL, Zhu GF, Wang BF, Fu G, Ren SX, Zhao GP, Chen Z, Lee YS, Cheong JE, Choi SH, Wu KM, Liu TT, Hsiao KJ, Tsai SF, Kim CG, OOta S, Kitano T, Kohara Y, Saitou N, Park HS, Wang SY, Yaspo ML, Sakaki Y. DNA sequence and comparative analysis of chimpanzee chromosome 22. Nature. 2004; 429:382–388. [PubMed: 15164055]

Wood EJ, Chin-Inmanu K, Jia H, Lipovich L. Sense-antisense gene pairs: sequence, transcription, and structure are not conserved between human and mouse. Frontiers Genet. 2013; 4:183.

Wu C, Orozco C, Boyer J, Leglise M, Goodale J, Batalov S, Hodge CL, Haase J, Janes J, Huss JW 3rd, Su AI. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. Genome Biol. 2009; 10:R130. [PubMed: 19919682]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Xu A-J, Kuramasu A, Maeda K, Kinoshita K, Takayanagi S, Fukushima Y, Watanabe T, Yanagisawa T, Sukegawa J, Yanai K. Agonist-induced internalization of histamine H2 receptor and activation of extracellular signal-regulated kinases are dynamin-dependent. J Neurochem. 2008; 107:208–217. [PubMed: 18691388]

Xu J, He T, Wang L, Wu Q, Zhao E, Wu M, Dou T, Ji C, Gu S, Yin K, Xie Y, Mao Y. Molecular cloning and characterization of a novel human BTBD8 gene containing double BTB/POZ domains. Int J Mol Med. 2004; 13:193–197. [PubMed: 14654994]

Yoshioka T, Inagaki K, Noguchi T, Sakai M, Ogawa W, Hosooka T, Iguchi H, Watanabe E, Matsuki Y, Hiramatsu R, Kasuga M. Identification and characterization of an alternative promoter of the human PGC-1alpha gene. Biochem Biophys Res Commun. 2009; 381:537–543. [PubMed: 19233136]

Zhang Y, Huypens P, Adamson AW, Chang JS, Henagan TM, Boudreau A, Lenard NR, Burk D, Klein J, Perwitz N, Shin J, Fasshauer M, Kralli A, Gettys TW. Alternative mRNA splicing produces a novel biologically active short isoform of PGC-1alpha. J Biol Chem. 2009; 284:32813–32826. [PubMed: 19773550]

Zheng ZZ. The functional specialization of the planum temporale. J Neurophysiol. 2009; 102:3079–3081. [PubMed: 19776362]

Zollman S, Godt D, Privé GG, Couderc JL, Laski FA. The BTB domain, found primarily in zinc finger proteins, defines an evolutionarily conserved family that includes several developmentally regulated genes in *Drosophila*. Proc Natl Acad Sci USA. 1994; 91:10717–10721. [PubMed: 7938017]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

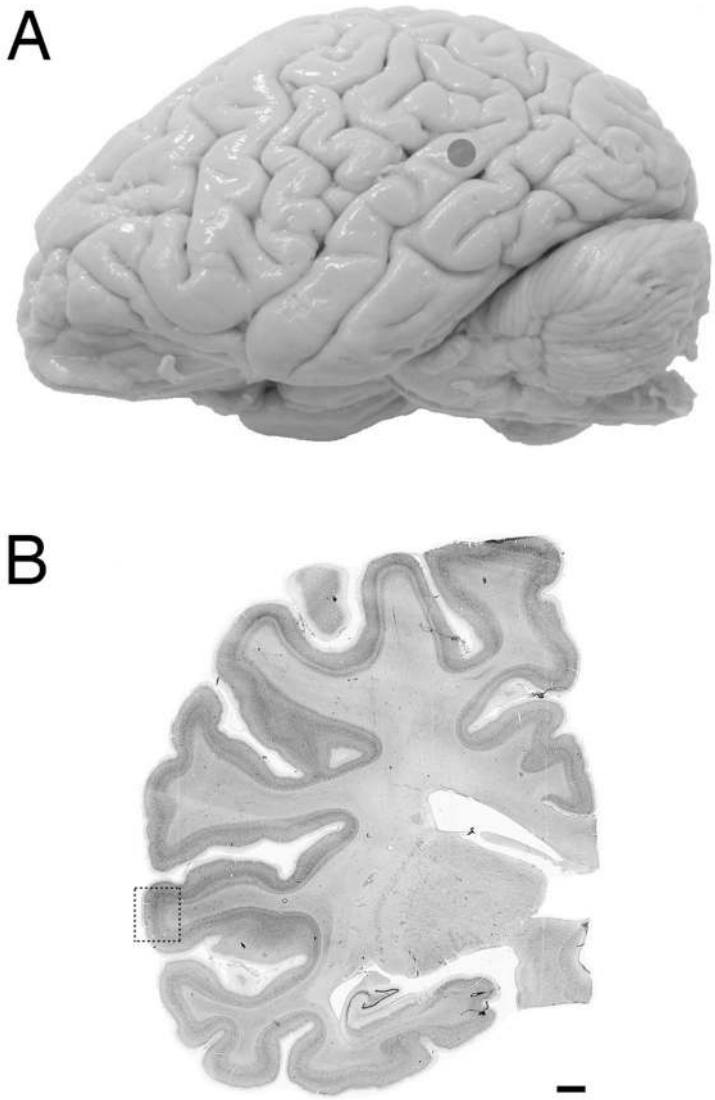


Fig. 1.

Location of the gorilla planum temporale. (A) Left lateral view of a western lowland gorilla brain showing the location where tissue was sampled from the planum temporale (dot) in the specimen used in our study. This immersion-fixed whole brain is from a 50 year old female gorilla, and was provided by the Hogue Zoo (Salt Lake City, Utah). (B) A Nissl-stained histological section through the hemisphere of another gorilla brain specimen showing the location of the planum temporale in a coronal plane (box). Both brains shown here are from individuals other than the gorilla used for RNA sequencing in our study. Scale bar = 0.75 mm for A and 1 mm for B.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

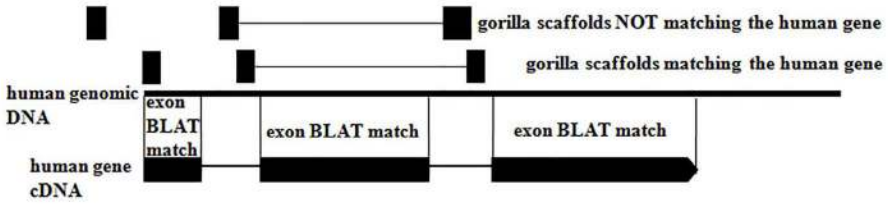


Fig. 2.

Schematic of gorilla transcriptome scaffold matching to human genes. To identify gorilla scaffolds that, when BLAT-mapped to the human genome, match exonic sequences of known human genes (protein-coding from RefSeq NM; noncoding-RNA from RefSeq NR and from the Jia et al. 2010 dataset), we interrogated block-level BLAT mappings of each scaffold for block (exon) overlap with blocks of any known human genes along the hg18 assembly. Partial or complete block-level overlaps between gorilla-scaffold-to-human-genome and human-gene-exon-to-human-genome mappings were scored. Gorilla scaffolds residing outside of or wholly intronic to any known human genes were not annotated with any gene matches.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

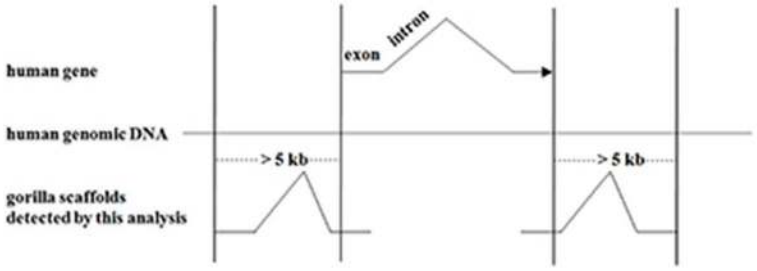


Fig. 3.

Human-gorilla gene structure differences at 5' and 3' termini of conserved orthologous genes. Gene structure differences were identified from transcript-to-genome alignments in which both gorilla and human transcripts were uniquely mapped to the human genome. All differences involving protein-coding genes were manually annotated using the UCSC platform as described in the text.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

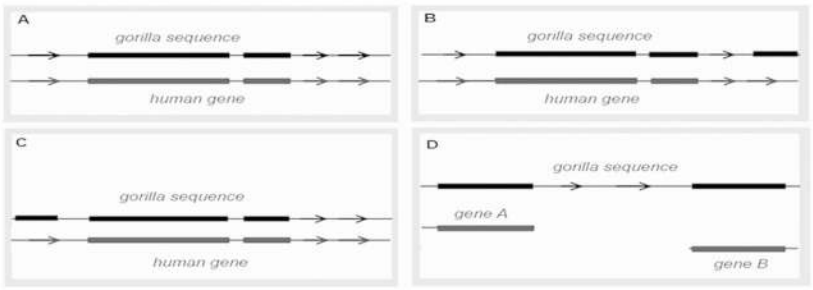


Fig. 4.

Annotation categories of aligned gorilla sequences relative to the human genome. Thick boxes are exons, thin boxes are introns and intergenic sequences, arrows indicate the direction of transcription. In each frame, the bottom horizontal track represents the UCSC Genome Browser human gene tracks whereas the top tracks represent the aligned gorilla transcripts. These are hypothetical examples. A. An uninteresting match: the gorilla sequence directly matches the human sequence and shares the same splice structure. B. Example of a 3'-end extended sequence, with an additional exon present in the gorilla sequence, relative to the human. C. Example of a 5'-end extended sequence, with an additional exon present in the gorilla sequence, relative to the human. D. Example of an intergenic splicing event: gene A and gene B are two human gene matches that are not connected by any human transcripts in the public databases, but are connected by an intergenically spliced gorilla transcript.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

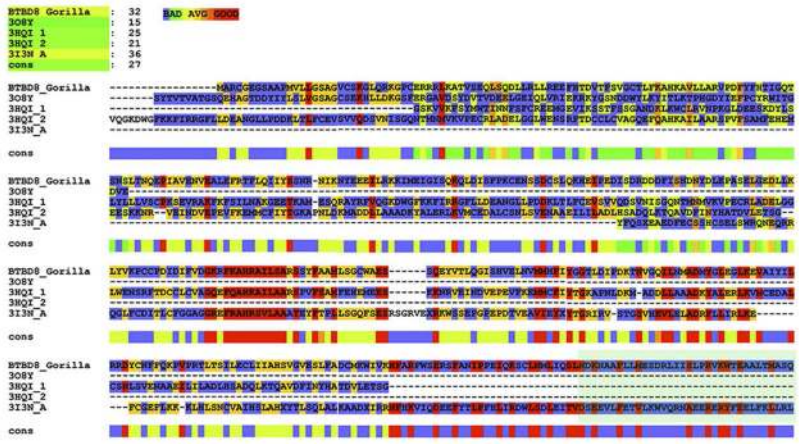


Fig. 5.

A multiple alignment of gorilla BTBD8 with the sequences of the proteins whose structures were used as templates to build the homology model.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

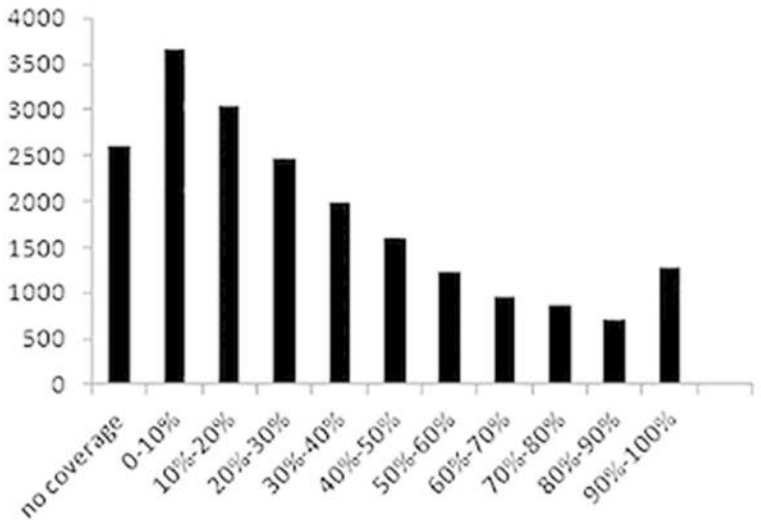


Fig. 6.

Frequency distribution histogram of known human protein-coding genes' open reading frame coverage by assembled gorilla transcripts. The x-axis is the percentage of open reading frame coverage by gorilla transcripts, within each bin. The y-axis is the number of human genes in that bin. These results reside in Supplementary Dataset 12.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

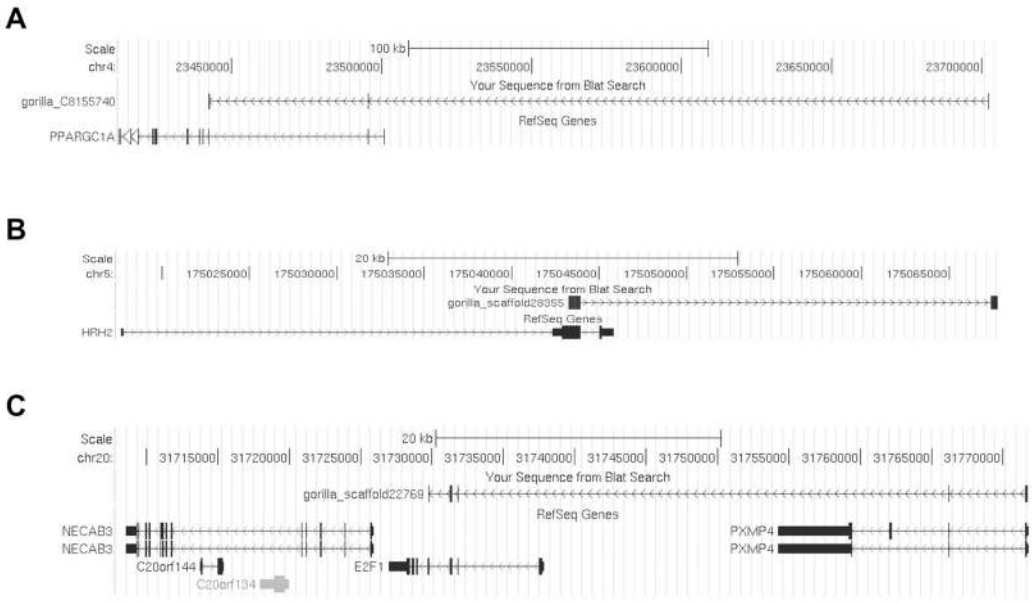


Fig. 7.

Examples of sequence matching using BLAT and the UCSC Genome Browser. All examples depict UCSC BLAT alignments of query sequences to the hg18 human genome assembly. The best-scoring BLAT result was viewed in each case. A. A unique 5’-end sequence relative to the human gene, as evidenced by the sequence extension and the direction of transcription. B. Similar to (A), a unique 3’-end sequence relative to the human gene. C. An intergenic splicing event.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

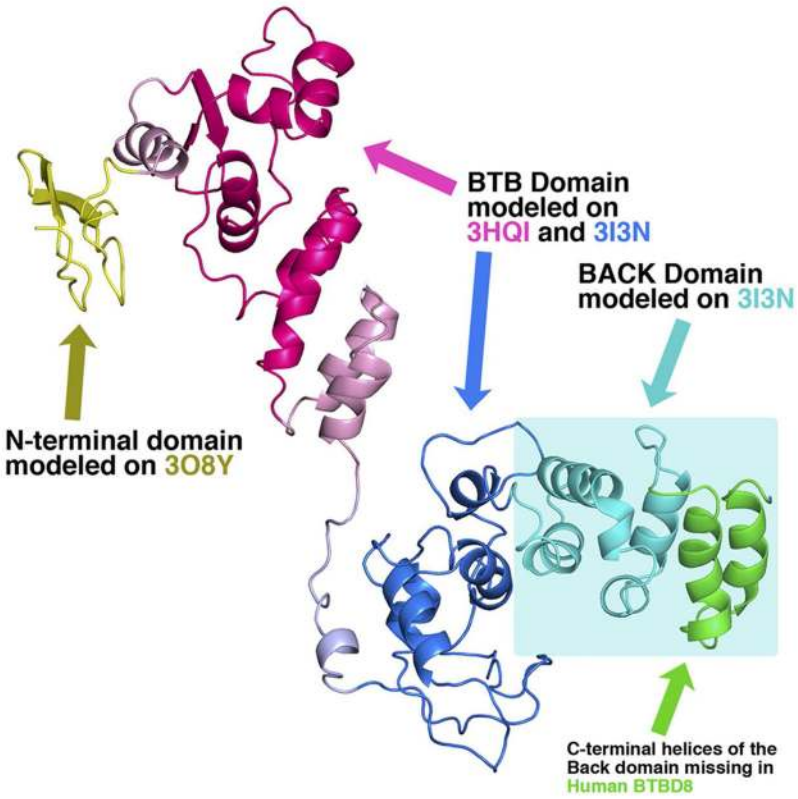


Fig. 8.

Ribbon representation of the full-atom model of gorilla BTBD8 protein. Upon an initial energy minimization of the side chains corresponding to the sequence of gorilla BTBD8 protein, the entire protein was relaxed to a stable conformation by means of a 1.2 ns molecular dynamics (MD) run. Four domains are visible in the model. The two central domains contain a BTB subdomain (colored in magenta and dark blue, respectively). The C-terminal domain (cyan box) has a BACK fold. The final two helices (colored in green) of the BACK domain are notably absent in the human protein, whose amino acid sequence is 30 residues shorter. Residues 1–41 of BTBD8 were modeled using residue 15–55 of the X-ray structure of 5-lipoxygenase (PDB 3O8Y) as template. Residues 42–185 of BTBD8 containing the first BTB domain (residue 58–157) were modeled using residues 184–326 of the MATH-BTB protein known as SPOP (PDB 3HQI). Residues 186–312 of BTBD8 containing a second BTB domain were modeled using residues 180–306 of SPOP and residues 77–198 of human klhl11 (Kelch-like protein 11; PDB 3I3N). Residues 313–408 of gorilla BTBD8 protein, which comprise the entire BACK domain, were modeled using residues 199–294 of human Klhl11.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1

Annotation of gorilla RNAseq data with unique transcript ends and no human EST support

Unless otherwise specified, each gorilla transcript assembly has no EST or cDNA support in humans for the unique splice junction, has a unique 5'- or 3'-end relative to all known transcripts of the orthologous gene in humans, and has canonical splice junctions (splice donor GT, splice acceptor AG). Gene name refers to human RefSeq.

Gene symbol	Annotation	Splice Site Conservation Notes (genomic sequence, NOT gene structure)	ORF Properties
BTBD8	Matches gene BTBD8 **(kelch/ankyrin domains; most abundant human expression is in fetal brain; nuclear by GO)**; unique 3' exon located downstream from BTBD8 partially **overlaps KIAA1107 exon 1 (intergenic splicing to 5'UTR but not the CDS of KIAA1107)**, but contains unique sequence prior to KIAA1107 as well.	Intron preceding the unique last exon: Both GT and AG conserved in all UCSC vertebrates. The GT is used in human BTBD8 intragenic alternative splicing in breast cancer. The AG is never used in human.	**Gorilla: unique C-terminus (34 aa).** Human has a different C-terminus not alignable to gorilla. Gorilla C-terminus has BLASTP hits to BTB domains in vertebrates, hinting at a human-specific loss of this exon or at cryptic BTB ORF-carrying potential of KIAA1107's 5'UTR.
GBF1	Golgi-localized GDP/GTP exchange factor, essential for Golgi assembly. IMPORTANT HOST FACTOR FOR VIRAL REPLICATION. Gorilla-specific first exon is cis-antisense to PITX3 and is contributes a GBF1 5'UTR plus 6 aa. **bioGPS highest human expression: pineal, prostate.**	The GT- splice donor site following the gorilla-specific first exon is found only in placental mammals and opossum.	**Gorilla has unique N-terminus (6 aa)** in-frame to and upstream of the human initiator M (aa #7 in gorilla).
PPP1R3C	**Protein phosphatase 1, regulatory (inhibitor) subunit 3C, hypoxia-response element. Positive correlation with glycogen accumulation.** Unique TSS and two 5'exons not used in human. **bioGPS highest human expression: trachea, tongue, skeletal muscle.**	First intron relies on primate-specific GT and AG. Second intron has a conserved though not pan-mammalian GT, and splices into the conserved AG of the canonical PPP1R3C sole intron.	Misses the first 5 aa of the human protein. AA 6 (M) of human is the predicted initiator M in gorilla.
RNF17	**Spermatogenesis-associated ring-finger / tudor domain.** Gorilla-specific 3'-end exon is in an intron of the neighboring gene CENPJ in the antisense orientation. Gorila transcript lacks the last two exons of human RNF17. **bioGPS human expression: mostly testis.**	Gorilla-specific last exon uses the conserved splice donor of the penultimate human intron, and a unique splice acceptor within a LINE1 insertion.	**Gorilla-specific splice into the LINE1 generates an 11-aa unique C-terminus** distinct from the human one.
CATSPER2	**Voltage-gated Ca++ entry channel, male infertility gene, sperm-associated 2. Functions in sperm motility.** Gorilla-specific first exon usage is from adjacent CATSPER2P1 unprocessed pseudogene (segmental duplication): **intergenic splicing. bioGPS human expression: testis, retina, pineal gland, cerebellum.**	Uses conserved CATSPER2P1 **first-intron** splice donor and conserved CATSPER2 **first-intron** splice acceptor.	Aligns completely with CATSPER2 (AA1-AA55) with 3 substitutions but no gorilla-specific protein sequence otherwise. CATSPER2P1 contribution is 5'UTR only.
OSTN	**Osteocrin. Functions: osteoblast differentiation; induces insulin resistance; binds natriuretic peptide receptor. bioGPS: equally low expression everywhere.**	The GT after the gorilla-specific exon is only conserved in humans, chimpanzee, and gorilla.	ORF identical to human. Gorilla upstream exon is entirely 5'UTR.
HRH2	**H2 histamine receptor; mediates gastric secretion.** Extends 3' sequence. Unique gorilla last coding exon contains ORF C-terminus and stop, and maps to the genome well after the last exon of human HRH2, which is not used in gorilla. **bioGPS: mostly cd34+, leukemia.**	Splice donor GT is shared with the human last intron. Gorilla splice acceptor AG is conserved in all mammals but is not used by any human transcripts.	Unique C-termini: 38 aa in humans, **63 aa in gorilla** (no alignment between the two, they map 23 kb apart on the human genome).
CWC15	**Spliceosome-associated protein homolog.** Gorilla unique first exon is cis-antisense to a histone demethylase gene (KDM4DL). **bioGPS: expressed everywhere.**	Intron 1: GG (but 1 bp away from GT used by an EST) – AG (pan-mammalian). Intron 2: GT (primate-specific) – AG (normal CWC15 intron 2 splice acceptor).	No ORF found; however, the ORF is expected to be identical to human CWC15 because gorilla-specific splicing is all in the 5'UTR.
APOH	**Apolipoprotein H: lipoprotein metabolism, up in type II diabetes. bioGPS: liver-specific.**	Unique first exon splice donor is primate-specific. Intra-5'UTR splice acceptor is conserved in most mammals.	Gorilla read stops short of the human ATG.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Gene symbol	Annotation	Splice Site Conservation Notes (genomic sequence, NOT gene structure)	ORF Properties
TTLL13	**tubulin tyrosine ligase-like, polyglutamylase.** 5'exon is not supported by any human cDNA/EST data. Splice sites appear non-canonical (but cannot be resolved due to splice alignment ambiguity). **bioGPS: very low expression everywhere.**	Unique first exon splice donor is primate-specific.	**At least 12 aa of gorilla-specific N-terminal sequence.**
E2F1	**Intergenic splicing** of E2F1 (Rb-associated transcription factor) and PXMP4 (a peroxisomal membrane protein). No human EST support (although 3 human ESTs support intergenic splicing between PXMP4 and another nearby gene). **bioGPS (E2F1): mostly early erythroid. (PXMP4): mostly lung and liver.**	Intergenic splice event uses a conserved splice donor GT in PXMP4 and a conserved splice acceptor AG in E2F1.	Predicted ORF is out of frame with human sequence of either protein.
RFPL1-AS1	Assigned by splice site orientation to the negative strand of the genome, so matches endogenous antisense of RFPL1, not the RFPL1 gene. 5'-end of gorilla is cis-antisense to neurofilament-H gene **(while 3'-end is cis-antisense to RFPL1). bioGPS: BRAIN-SPECIFIC**	Splice donor of first exon is canonical GT and highly conserved throughout placental mammals; second exon splice acceptor is canonical AG but appears to reside in a primate-specific sequence.	Not applicable; long non-protein-coding RNA gene.
PPARGC1A	**peroxisome proliferator-activated receptor gamma, coactivator 1 alpha. Regulates mitochondrial biogenesis; associated with glucose tolerance, BMI, insulin sensitivity. bioGPS: salivary gland, thyroid**	All splice sites are canonical and highly conserved.	Splices into human exon 2 (human ATG is in human exon 1). **At least 8 aa of gorilla-specific N-terminal sequence.**