Evolution of Human-Specific Neural *SRGAP2* Genes by Incomplete Segmental Duplication

Megan Y. Dennis,[1,8] Xander Nuttle,[1,8] Peter H. Sudmant,[1] Francesca Antonacci,[1] Tina A. Graves,[3] Mikhail Nefedov,[4] Jill A. Rosenfeld,[5] Saba Sajjadian,[1] Maika Malig,[1] Holland Kotkiewicz,[3] Cynthia J. Curry,[6] Susan Shafer,[7] Lisa G. Shaffer,[5] Pieter J. de Jong,[4] Richard K. Wilson,[3] and Evan E. Eichler[1,2,*]

[1]Department of Genome Sciences
[2]Howard Hughes Medical Institute
University of Washington School of Medicine, Seattle, WA 98195, USA
[3]The Genome Institute at Washington University, Washington University School of Medicine, St. Louis, MO 63110, USA
[4]Children's Hospital Oakland Research Institute, Children's Hospital and Research Center at Oakland, Oakland, CA 94609, USA
[5]Signature Genomic Laboratories, PerkinElmer, Inc., Spokane, WA 99207, USA
[6]Genetic Medicine Central California, University of California, San Francisco, Fresno, CA 93701, USA
[7]Carle Clinic Association, Urbana, IL 61801, USA
[8]These authors contributed equally to this work
*Correspondence: eee@gs.washington.edu
DOI 10.1016/j.cell.2012.03.033

SUMMARY

Gene duplication is an important source of phenotypic change and adaptive evolution. We leverage a haploid hydatidiform mole to identify highly identical sequences missing from the reference genome, confirming that the cortical development gene Slit-Robo Rho GTPase-activating protein 2 (*SRGAP2*) duplicated three times exclusively in humans. We show that the promoter and first nine exons of *SRGAP2* duplicated from 1q32.1 (*SRGAP2A*) to 1q21.1 (*SRGAP2B*) ~3.4 million years ago (mya). Two larger duplications later copied *SRGAP2B* to chromosome 1p12 (*SRGAP2C*) and to proximal 1q21.1 (*SRGAP2D*) ~2.4 and ~1 mya, respectively. Sequence and expression analyses show that *SRGAP2C* is the most likely duplicate to encode a functional protein and is among the most fixed human-specific duplicate genes. Our data suggest a mechanism where incomplete duplication created a novel gene function—antagonizing parental *SRGAP2* function—immediately "at birth" 2–3 mya, which is a time corresponding to the transition from *Australopithecus* to *Homo* and the beginning of neocortex expansion.

INTRODUCTION

Several genes have been implicated as being important in specifying unique aspects of evolution along the human lineage. These include genes involved with the development of language (*FOXP2*) (Enard et al., 2002), changes in the musculature of the jaw (*MYH16*) (Stedman et al., 2004), and limb and digit specializations (*HACNS1*) (Prabhakar et al., 2008). Despite these intriguing candidates, the bulk of the morphological and behavioral adaptations unique to the human lineage remains genetically unexplained. Not all genes, however, have been amenable to standard genetic analyses. This is particularly true for genes embedded within recently duplicated sequences (Bailey et al., 2002), which are frequently missing or misassembled from the reference genome (Eichler, 2001). Genes residing in these complex regions are important to consider for three reasons: (1) duplicated genes have been recognized as a primary source of evolutionary innovation (Lynch and Katju, 2004; Ohno, 1970); (2) the human and great-ape lineages have experienced a surge of genomic duplications over the last 10 million years (Marques-Bonet et al., 2009); and (3) human-specific duplications are significantly enriched in genes important in neurodevelopmental processes (Fortna et al., 2004; Sudmant et al., 2010).

Among these human-specific duplicated genes, *SRGAP2* was recently shown to be important in cortical development (Guerrier et al., 2009; Guo and Bao, 2010). The gene encodes a highly conserved protein expressed early in development when it acts as a regulator of neuronal migration and differentiation by inducing filopodia formation, branching of neurons, and neurite outgrowth. Analysis of the human reference genome revealed that *SRGAP2* was misassembled and that most of its duplicate copies were not yet sequenced or characterized. We developed an approach by using genomic material devoid of allelic variation (from a complete hydatidiform mole [Kajii and Ohama, 1977]) to completely sequence and characterize the missing loci corresponding to this human-specific gene family. These data allowed us to reconstruct the complex evolutionary history of this gene family since humans diverged from nonhuman primates (~6 million years ago [mya]; Patterson et al., 2006), understand the potential of these loci to generate functional transcripts, and assay the extent of human genetic variation. We put forward
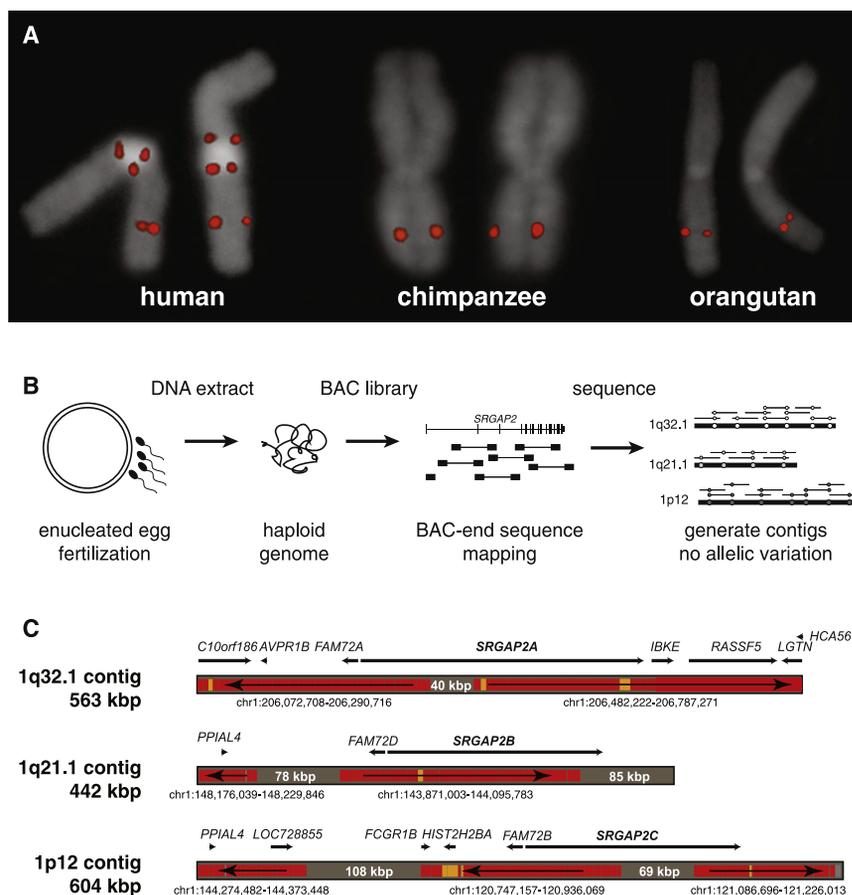
Figure 1. Genomic Characterization and Sequence Resolution of *SRGAP2* Loci

(A) FISH analysis shows three distinct copies of *SRGAP2* on metaphase human chromosome 1, compared to a single copy in chimpanzee and orangutan (see Figure 2A for location of FISH probe; Figure S1 and Table S1 for details of additional FISH assays).

(B) *SRGAP2* genomic loci were sequenced and assembled using a BAC library (CHORI-17) created from human haploid genomic source material (complete hydatidiform mole). The absence of allelic variation allowed paralogous sequences to be resolved with high confidence based on near-perfect sequence identity overlap (>99.9%).

(C) Regions highly identical to the reference genome (GRCh37/hg19) are colored in red (identity = 99.8%–100%) and orange (99.6%–99.8%), whereas regions completely absent from the current assembly are shaded gray (with region sizes indicated). Arrows show the orientation of the reference genome sequence with respect to the contigs (e.g., a left directional arrow indicates the reverse strand). Overall, this indicates that even the ancestral (*SRGAP2A*) gene locus was missing sequence data, misassembled, and incorrectly orientated over 400 kbp of the current high-quality reference assembly. Genomic coordinates correspond to the representative human reference region with corresponding genes within these regions mapped along each contig.

a model for gene evolution in which incomplete segmental duplication creates derivative copies that antagonize the ancestral function.

RESULTS

Genome Sequencing

We confirmed that *SRGAP2* was specifically duplicated in the human lineage by fluorescent in situ hybridization (FISH) by using a probe corresponding to the human *SRGAP2* (spanning exon 3, Table S1). We identified three map locations on chromosome 1 (1q32.1, 1q21.1, and 1p12), as compared to a single chromosomal signal at 1q32.1 among other ape species (Figure 1A). An analysis of the segmental duplication content of 11 additional mammalian genomes (see Extended Experimental Procedures) showed no evidence of recent duplication in any lineage other than human and established 1q32.1 as the ancestral copy. FISH analysis of cell lines derived from humans of diverse ethnicity consistently showed a pattern of three distinct signals on each chromosome 1 corresponding to paralogs that were all incompletely sequenced in the human reference genome (GRCh37/hg19).

We reasoned that the recent nature of the duplications resulted in high-identity duplications with little genetic variation. As a result, allelic and paralogous copies became difficult to disentangle during genome assembly (Lander et al., 2001). To resolve the different genomic copies, we constructed a large-insert bacterial artificial chromosome (BAC) library from DNA derived from a complete hydatidiform mole (CHORI-17). Because a complete hydatidiform mole originates from the fertilization of an enucleated human oocyte with a single spermatozoon (Fan et al., 2002; Kajii and Ohama, 1977), the corresponding DNA represents a haploid, as opposed to a diploid, equivalent of the human genome (Figure 1B). We leveraged the absence of allelic variation to unambiguously distinguish *SRGAP2* copies despite their high sequence identity. We selected clones with homology to *SRGAP2* and subjected them to high-quality capillary-based sequencing, requiring >99.9% sequence identity of the overlap between sequenced inserts for assembly into the same contig.

We generated three sequence contigs corresponding to *SRGAP2* paralogs at 1q32.1 (562,704 bp; *SRGAP2A*), 1q21.1 (441,682 bp; *SRGAP2B*), and 1p12 (603,678 bp; *SRGAP2C*) (Figure 1C), generating over 1.6 Mbp of high-quality finished sequence. During the assembly process, we identified a single BAC clone (CH17-248H7) that harbored sequence for a *SRGAP2* paralog (exons 7–9) but did not share >99.9% identity with any of the three contigs, suggesting that a fourth *SRGAP2* duplicate existed (*SRGAP2D*). Upon this discovery, we repeated our FISH analysis using a probe mapping across exon 1 of *SRGAP2*

Cell

Table 1. Percent Sequence Divergence of *SRGAP2* Paralogs

	SRGAP2A	SRAGP2B	SRGAP2C	SRGAP2D
SRGAP2A	–	0. 015	0. 016	0.069
SRAGP2B	0.525	–	0. 014	0.038
SRGAP2C	0.584	0.451	–	0.065
SRGAP2D	0.452	0.136	0.400	–

Kimura two-parameter model of genetic distance computed as base substitutions per site (left diagonal) and standard error (right diagonal). Pairwise distances are computed across 244,200 sites representing the complete shared genomic region between *SRGAP2* paralogs. Values for *SRGAP2D* represent pairwise distances computed across 9,541 sites. As a reference, the genetic distance between *SRGAP2A* and its chimpanzee ortholog locus is 0.852 ± 0.019, whereas that of chimpanzee to human paralogs *SRGAP2B* and *SRGAP2C* (0.901 ± 0.019 and 0.960 ± 0.020) are consistent with the accelerated mutation rate for these chromosomal regions.

and discovered four distinct signals on chromosome 1, with *SRGAP2D* mapping proximally to *SRGAP2B* on chromosome 1q21.1 (Figure S1 available online, Table S1). The absence of this signal from the initial FISH assay (Figure 1A) suggested that a genomic region containing exon 3 was deleted from *SRGAP2D*.

The new local assemblies resolved the sequence and structure of three copies, adding 379,665 bp of new sequence completely absent from the human reference, including 40,233 bp within the ancestral *SRGAP2A* (Figure 1C). Additionally, we discovered 559,693 bp of sequence mapped incorrectly in orientation or chromosomal location within the human reference. Combined, we added or corrected more than 0.4% of the human chromosome 1 euchromatic sequence (Gregory et al., 2006). All finished sequence data, as well as the new human genome assemblies, have been deposited into GenBank and will be integrated into subsequent human genome reference assemblies (see Extended Experimental Procedures for accession numbers).

Comparisons between the three sequence contigs revealed large, interspersed segmental duplications of high-sequence identity (99%–99.5%) that were incomplete with respect to the ancestral locus (Table 1). We determined that the original duplication event (258,245 bp) encompassed the promoter, other *cis* regulatory elements, and the first nine exons of the 22-exon ancestral *SRGAP2A* (Figure 2A). Clusters of Alu repeat elements mapped precisely at the boundaries of this duplicated segment (Figure S2), confirming previous observations that Alu repeats are strongly associated with primate genomic duplications (Bailey et al., 2003; Zhou and Mishra, 2005). A larger, secondary duplication event (>515 kbp) was shared between the *SRGAP2B* (1q21.1) and *SRGAP2C* (1p12) loci and included the entirety of the original duplication, although the *SRGAP2B* locus was subjected to subsequent larger deletions (102.6 and 49.0 kbp) upstream of the gene (Figure S2). Using multicolor FISH assays, we determined that the ancestral *SRGAP2A* paralog at 1q32.1 is transcribed toward the telomere, whereas the duplicate paralogs *SRGAP2B* and *SRGAP2C* are oriented such that gene transcription would proceed toward the centromere (Figure S1).

Evolutionary History of *SRGAP2*

To reconstruct the evolution of the duplication events, we generated a multiple-sequence alignment for a 244.2 kbp region that is shared among the three contigs by using orthologous sequence from chimpanzee (build GGSC 2.1.3/panTro3) and orangutan (build WUGSC 2.0.2/ponAbe2) as outgroups (Figure 2B). Phylogenetic analysis provides strong support (>99%) for distinct duplication events occurring at different time points during human evolution. Notably, we find that the duplicated sequences have evolved much more rapidly (Tajima's relative rate test; p = 0.00001–0.0249) than the ancestral 1q32.1 locus (p = 0.5345). Mutation rates are known to vary significantly depending on chromosomal location and context (CSAC, 2005). Based on analysis of unique orthologous sequence adjacent to the *SRGAP2C* duplicate region, we determined that the distal 1p12 region shows a 20%–46% higher substitution rate when compared to 1q32.1. If we adjust for this difference, calibrating to the estimated 1q32.1 substitution rate, we predict that the initial duplication occurred ~3.4 mya and that the secondary event occurred ~2.4 mya. We note that estimates of molecular divergence between the paralogs are robust (e.g., 0.451 ± 0.014% substitutions per site between the *SRGAP2B* and *SRGAP2C* loci), owing to the large number of substitutions discovered in the high-quality sequence used in these comparisons (Table 1). Some uncertainty in our estimates comes from our correction factor for differing substitution rates, but most uncertainty arises from ambiguity in the evolutionary timing of the divergence of chimpanzee and human (estimated at ~6 mya) (Patterson et al., 2006). If we take into account previously reported human and chimpanzee divergence times ranging from ~5–7 mya, based on fossil records (Brunet et al., 2002, 2005; Vignaud et al., 2002) as well as recent genetic analyses (Patterson et al., 2006), we estimate that the initial duplication occurred 2.8–3.9 mya, followed by the secondary duplication at 2.0–2.8 mya. We also performed phylogenetic analysis of the 9,541 bp region shared among the *SRGAP2A–C* paralogs and the incompletely sequenced *SRGAP2D* and determined that this copy was derived from the *SRGAP2B* locus ~1 mya (0.4–1.3 mya assuming a 6 mya divergence time for human and chimpanzee). Using comparative FISH analysis and probes mapping outside of the original duplication (Figure 2C), we determined the likely order of events: the ancestral *SRGAP2A* region duplicated first to 1q21.1 (*SRGAP2B*), and later the 1q21.1 copy duplicated to chromosome 1p12 (*SRGAP2C*) and within 1q21.1 (*SRGAP2D*).

Based on the gene structure of the ancestral *SRGAP2A*, sequence analysis predicts that *SRGAP2B* and *SRGAP2C* would produce transcripts maintaining an open-reading frame (ORF). These two duplicate copies, however, are predicted to produce a truncated form of *SRGAP2*, carrying nearly the entire F-BAR domain that lacks the final 49 amino acids (Figure 2A) (Guerrier et al., 2009). The ancestral SRGAP2 protein sequence is highly constrained based on our analysis of ten mammalian lineages. We find only a single amino acid change between human and mouse and no changes among nonhuman primates within the first nine exons of the *SRGAP2* orthologs. This is in stark contrast to the duplicate copies, which diverged from ancestral *SRGAP2A* less than 4 mya but have
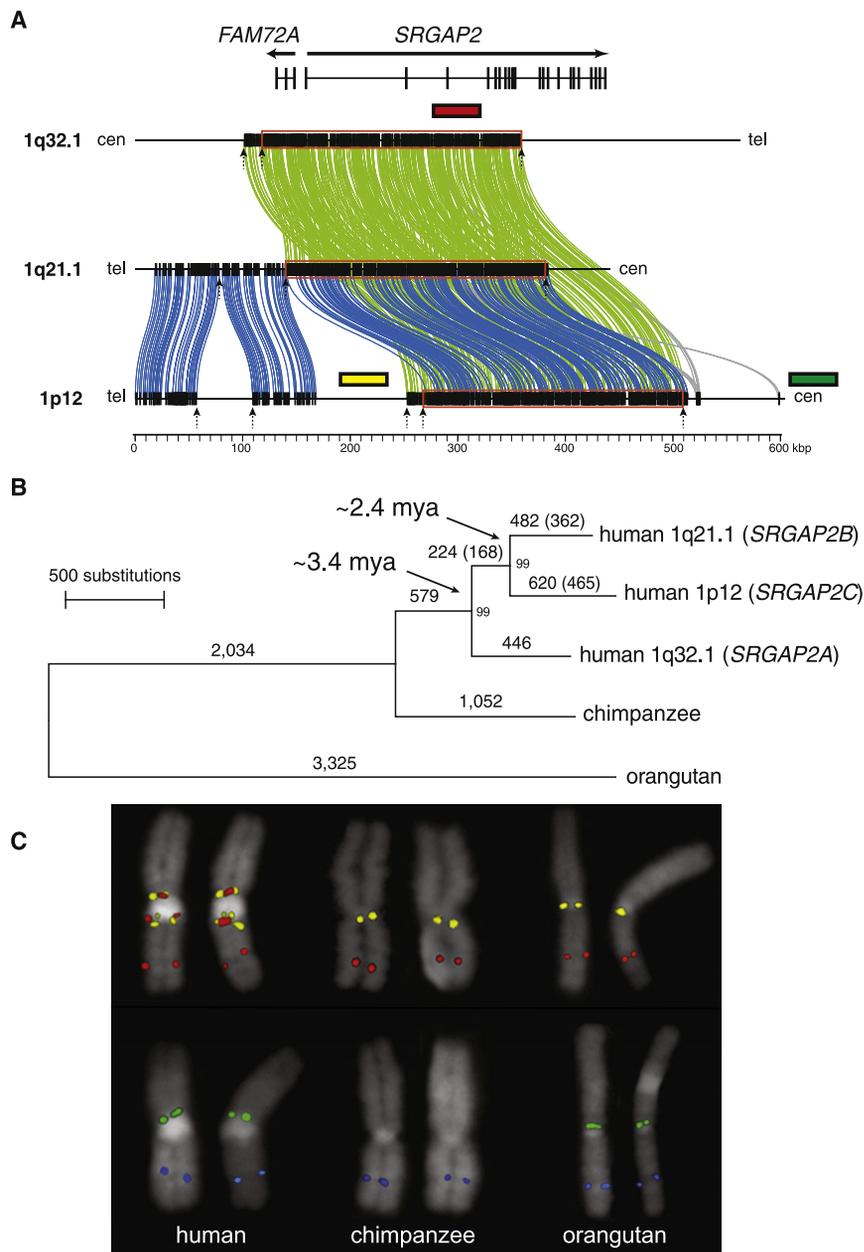
Figure 2. Evolutionary Characterization of *SRGAP2* Duplications

(A) A depiction of the gene structure of *SRGAP2* with respect to the three assembled contigs. Homologous segments are shown using Miropeats (Parsons, 1995) where green lines indicate nearly identical segments (s = 1,000) shared between *SRGAP2A* and the duplicate *SRGAP2* paralogs, and the blue lines delineate the larger (>515 kbp) extent of homology between *SRGAP2B* and *SRGAP2C*. The 244.2 kbp genomic region shared among all three contigs is highlighted (red box) with clusters of Alu repeats at the breakpoints (arrows). Also see Figure S2 for a detailed representation of Alu elements and segmental duplications across duplicated regions.

(B) An unrooted neighbor-joining tree was constructed based on a 244.2 kbp multiple sequence alignment of the three loci. Both 1p12 and 1q21.1 branches show accelerated rates of substitution (p = 0.00001 and p = 0.0249; Tajima's relative rate test). The actual (no parentheses) and adjusted (parentheses) number of substitutions for locus-specific acceleration is indicated above each branch along with the bootstrap support at each node. We estimate the timing assuming chimpanzee and human diverged 6 mya. Also see Table S2 for molecular evolution of the shared *SRGAP2* coding regions.

(C) FISH experiments on metaphase human chromosome 1, as well as the orthologous chimpanzee and orangutan chromosomes, were performed to discern the order of duplication events. Locations of probes with respect to the contigs are shown in (A). A probe (yellow) targeting the sequence adjacent to the original *SRGAP2* duplicate region hybridizes to 1q21.1 in chimpanzee and orangutan, suggesting that the original *SRGAP2* duplicate paralog maps to the region homologous with nonhuman primate 1q21.1. A probe (green) targeting the unique sequence on the p arm of chromosome 1 proximal to *SRGAP2C* hybridizes to the chromosome 1p arm in orangutan, refuting the possibility that *SRGAP2C* moved to the p arm via a simple pericentromeric inversion (Szamalek et al., 2006) and distinguishing the p arm from the genomic region at 1q21.1 where the original *SRGAP2* duplicate paralog maps. A probe (blue) was used to distinguish the chromosome 1q arm.

accumulated as many as seven amino acid replacements (five for *SRGAP2C* and two for *SRGAP2B*), compared to one synonymous change.

We used a likelihood ratio test (Yang, 2007) to evaluate differences in selective pressures acting on *SRGAP2* and found that the best model of selection allows an increased nonsynonymous (dN) to synonymous substitution (dS) ratio of the *SRGAP2* duplicate paralogs while maintaining purifying selection in the remaining lineages (compared with the fixed dN/dS model, p = 1.32 × 10^{-11}, Table S2). This difference is consistent with an increased substitution rate of the 1q21.1 and 1p12 chromosomal regions and a relaxation of selective pressure on the duplicate copies. Overall, this mechanism provides a means for rapid evolutionary

change of an otherwise constrained developmental gene (Lynch and Conery, 2000).

SRGAP2 mRNA Expression and Paralog Gene Structure

We assayed for expression of *SRGAP2* paralogs by designing specific reverse-transcriptase PCR (RT-PCR) assays that distinguish the duplicate paralogs from the ancestral copy based on the presence of a duplicate-specific 3′ untranslated region (UTR) present in a previously sequenced cDNA mapping to the *SRGAP2C* locus (GenBank accession BC112927). A total of 96 transcripts were sequenced from RNA derived from the SH-SY5Y neuronal cell line, pooled fetal brain, a single fetal brain, and a single adult brain (Figure 3A and Table S3). Comparing
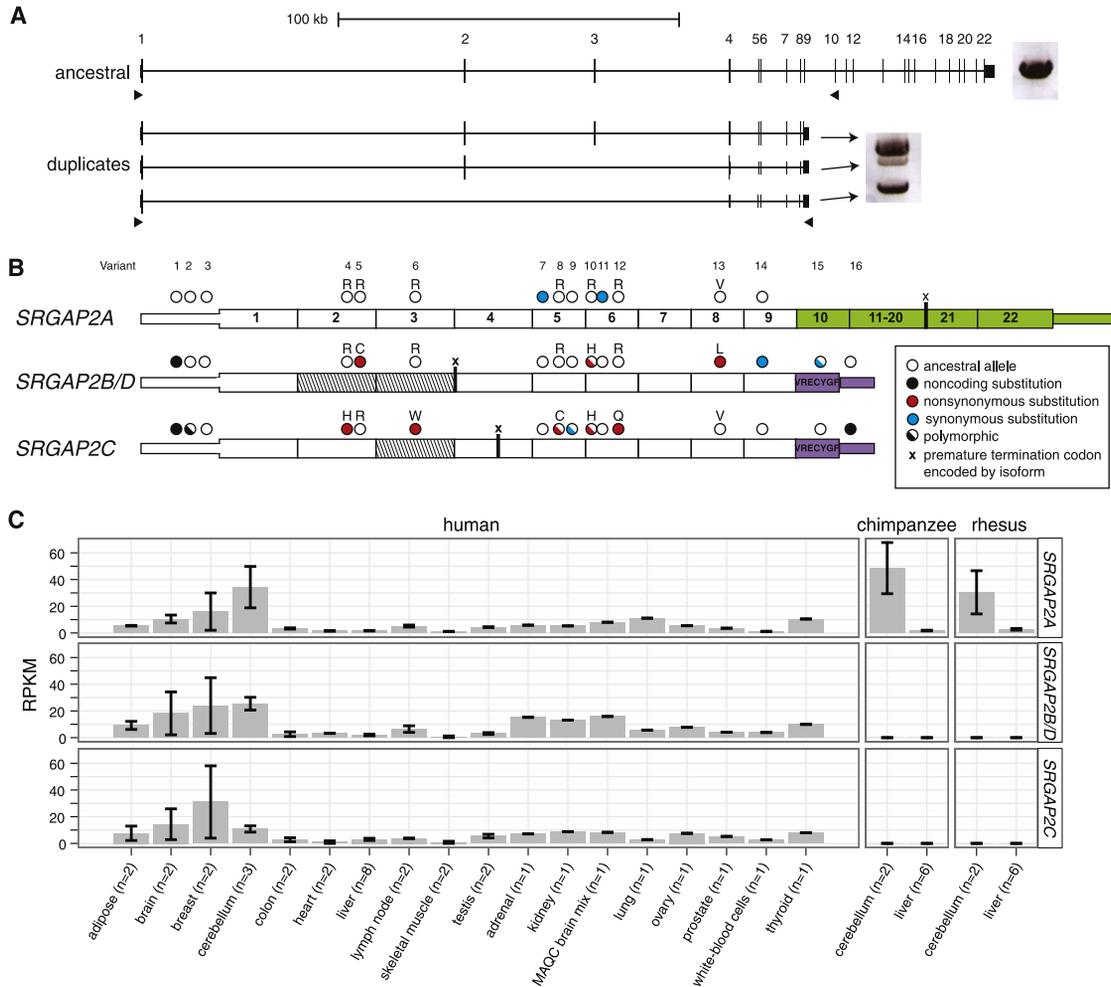
Figure 3. Paralog-Specific *SRGAP2* Gene Expression

(A) Long-range RT-PCR products from pooled fetal brain RNA are shown next to the gene models. A single band was amplified from the ancestral paralog, whereas three bands were amplified from duplicate paralogs by using primers (black triangles) designed to target alternative isoforms. Ninety-six cDNA transcripts were cloned and sequenced.

(B) Fixed paralog-specific variants were used to assign transcripts to respective genomic loci, allowing both polymorphic and fixed putative amino acid changes to be deduced. Exonic sequence specific to the ancestral copy (*SRGAP2A*; green) and the duplicate loci (*SRGAP2B/C/D*; purple) are shown. The locations of stop codons encoded by isoforms missing exons are represented with an "x." Exons missing from transcripts are indicated (diagonal lines) and likely correspond to the genomic deletion within *SRGAP2D* in the case of the exon 2 and 3 deleted isoform.

(C) Paralog-specific expression profiling was performed by using RNA-Seq data mapped to unique sequence identifiers. The mean RPKM of each *SRGAP2* paralog is shown for a variety of primate tissue types, with error bars representing ±SEM. The specificity of next-generation sequence data and the determination of single base-pair differences between the copies were necessary to tease apart the expression profiles of these virtually identical copies. Chimpanzee and macaque RNA-Seq data affirm the specificity of this assay. Also see Figure S3 and Table S3 for additional expression results.

genomic and cDNA sequences, we assigned the transcripts to their respective copies and identified the exon/intron structure, alternative splice forms, as well as fixed and polymorphic paralog-specific variants (PSVs) (Figure 3B). We found that all *SRGAP2* paralogs are transcribed, though at different relative proportions. We identified transcripts containing exons 1 through 9 that map specifically to *SRGAP2C* (n = 47) and *SRGAP2B* or *SRGAP2D* (n = 4). Using capillary sequencing of these transcripts and focusing our analysis on two fixed PSVs, we show that relative expression of the *SRGAP2B/D* transcript is markedly low (14%–25% and 30%–72% of *SRGAPC* tran-

script abundance in fetal and adult brain, respectively) (Figure S3). The most abundant duplicate transcript is expressed from *SRGAP2C* and predicts an ORF that would encode a truncated SRGAP2 protein (458 amino acids), including a partial F-BAR domain (Guerrier et al., 2009) and seven unique residues at the carboxyl terminus.

We also observed numerous transcripts and putative splice isoforms that are unlikely to encode functional proteins. The most abundant of these map to *SRGAP2B/D* (n = 31) missing exons 2 and 3 and result in a transcript that would encode a premature truncated protein (23 amino acids). These transcripts

Table 2. *SRGAP2A* and *SRGAP2C* Copy Number Variation Genotyping of Cases and Controls

Genotype Method	Size Resolution	Cohort[a]	Total Genotyped	Deletions	Duplications
SRGAP2A[b]					
Custom array CGH platforms	>50 kbp	intellectual disability (Signature Genomics) (Cooper et al., 2011)	15,767	3	3
SNP arrays	>50 kbp	controls (Cooper et al., 2011)	8,329	none	none
qPCR[c]	n/a	intellectual disability	1,602	none	none
		controls (NIMH and ClinSeq)	1,794	none	none
Illumina sequencing	>100 kbp	controls (1000 Genomes Project)	661	none	none
SRGAP2C[d]					
qPCR[e]	n/a	intellectual disability	1,602	none	1
		controls (NIMH and ClinSeq[g])	1,794	none	1
Custom array CGH[f]	>300 kbp	idiopathic autism (SSC)	2,294	none	2
		familial autism (AGRE)	579	none	none
		controls (NIMH and ClinSeq[g])	580	none	none
Illumina sequencing	>100 kbp	controls (1000 Genomes Project)	661	none	none

All detected deletions and duplications of *SRGAP2A* and *SRGAP2C* were >1 Mbp and include additional genes. Data from the Cooper et al. (2011) study could not be used to assess CNVs for *SRGAP2C*, as there was insufficient probe coverage on the microarrays used in those studies. See also Figure S4 and Table S4 for details of CNV breakpoints, phenotypes, and inheritance status.
[a]Abbreviations: SSC, Simons Simplex Collection (Fischbach and Lord, 2010); AGRE, Autism Genetic Resource Exchange (Geschwind et al., 2001); NIMH, National Institute of Mental Health (https://www.nimhgenetics.org/available_data/controls/); ClinSeq, Clinical Sequencing Pilot Project (Biesecker et al., 2009).
[b]Cases, n = 17,369; Controls, n = 10,784.
[c]The assay targeted intron 11 of *SRGAP2A*.
[d]Cases, n = 4,475; Controls, n = 2,662.
[e]Two assays were used targeting introns 6 and 7 of *SRGAP2C*, respectively.
[f]Using probes targeting the chromosome 1p11.2 region proximal to *SRGAP2C*, we identified duplications and determined that a subset of them extended into *SRGAP2C* by using qPCR assays. Notably, all duplications of *SRGAP2C* identified from the qPCR assay alone extended into the 1p11.2 proximal region and would have been detected using this same method.
[g]ClinSeq controls (n = 373) were screened both with array CGH and qPCR assays.

are consistent with our genomic sequence analysis, indicating that *SRGAP2D* has acquired a 115 kbp deletion including exons 2 and 3 (described later). Moreover, our analysis suggests that this transcript may be subjected to nonsense-mediated decay.

Using diagnostic PSVs to distinguish copies, we interrogated the expression of specific *SRGAP2* paralogs in various human and nonhuman primate tissues using RT-PCR (Figure S3) and RNA-Seq data (Figure 3C). The tissue profile reveals that the paralogs show similar broad patterns of expression, including expression in the developing human fetal brain concurrently with *SRGAP2A*. We observe higher expression in multiple regions of the human cortex and cerebellum when compared to other tissues including lung, kidney, and testis. As expected, we did not detect expression of the duplicate copies in any of the nonhuman-primate-derived tissues.

SRGAP2 Copy Number Variation

Because *SRGAP2* has been shown to play an important role in brain development, we initially focused on the ancestral *SRGAP2A* gene by examining a large cohort of pediatric cases with developmental delay (1,602 individuals tested using a quantitative PCR [qPCR]) assay specifically targeting *SRGAP2A* and 15,767 individuals reported by Cooper et al. [2011]) for potential copy number variation. We identified six large (>1 Mbp) copy

number variants (CNVs), including three deletions of the ancestral 1q32.1 region (Table 2), with no similar large CNVs observed among 10,123 controls. Because the CNVs are large and encompass multiple candidate genes, this observation does not prove pathogenicity of dosage imbalance of *SRGAP2A*. We note, however, that in one patient the proximal breakpoint maps within the first intron of *SRGAP2A*, potentially disrupting the gene (Figure S4 and Table S4). The patient is a ten-year-old child with a history of seizures, attention deficit disorder, and learning disabilities. An MRI of this patient also indicates several brain malformations, including hypoplasia of the posterior body of the corpus callosum. Recently, a de novo-balanced translocation t(1;9)(q32;q13) breaking within intron six of *SRGAP2A* was reported in a five-year-old girl who was diagnosed with West syndrome and exhibited epileptic seizures, intellectual disability, cortical atrophy, and a thin corpus callosum (Saitsu et al., 2011). Although much more work needs to be done, the neurological phenotypes observed in these two cases are consistent with neuronal migration deficits implicated in forms of developmental delay and epileptic encephalopathies (Saitsu et al., 2011).

We next focused on assessing copy number variation of each *SRGAP2* paralog in the human population. This is particularly challenging because most recently duplicated genes are typically highly copy number polymorphic (Sharp et al., 2005;
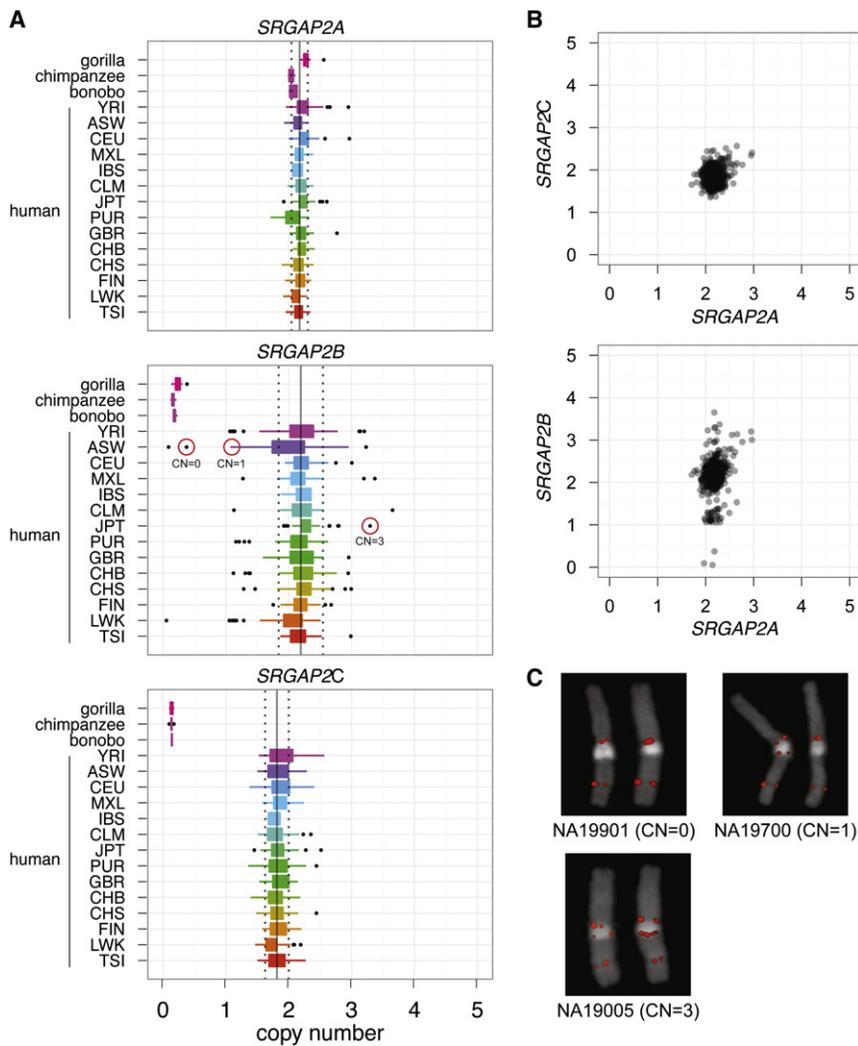
Figure 4. *SRGAP2* Copy Number Diversity in Human Populations

(A) Diploid copy number estimates of *SRGAP2* paralogs for 661 sequenced human genomes from 14 distinct populations (1000 Genomes Project) and from nonhuman primates are graphically represented as boxplots (the box contains the 25th to 75th percentile of the distribution, and the black dots represent outliers). The solid vertical and dashed lines represent the median copy number estimate and ±SD, respectively, of each paralog across all populations.

(B) *SRGAP2A* and *SRGAP2C* paralogs clearly are fixed at a copy number of two, while *SRGAP2B* is polymorphic showing four distinct copy number states. Note, we also detect polymorphism for *SRGAP2D* and have identified individuals homozygously deleted for this paralog.

(C) FISH validation of three HapMap individuals genotyped for *SRGAP2B* (circled in red in part [A]). All samples falling at the lower and upper tails of copy number distributions for all three paralogs were experimentally genotyped by using a paralog-specific qPCR assay; in all cases, *SRGAP2A* and *SRGAP2C* were validated as diploid copy number two. Also refer to Figure S5.

Sudmant et al., 2010), and experimental assays for accurately predicting copy number are problematic. For this purpose, we took advantage of diagnostic singly unique nucleotide (SUN) identifiers (n = 3,535) determined using our high-quality sequence of the three loci (see above). We mapped genome-sequencing data from 661 human individuals corresponding to 14 populations (1000 Genomes Project) and estimated the diploid copy number for each paralog by measuring read depth to these SUNs (Figure 4A) (Sudmant et al., 2010).

We find that both the ancestral *SRGAP2A* and the derived *SRGAP2C* copy are fixed at diploid copy number two across all humans assayed. In contrast, the *SRGAP2B* and *SRGAP2D* copies varied from 0–4 copies among the individuals tested (Figures 4B–4C). Importantly, we identified three individuals that are homozygously deleted for *SRGAP2B*. Notably, we also identified normal individuals that were homozygously deleted for *SRGAP2D*, the granddaughter copy with an acquired internal deletion of exons 2 and 3 (see Figure S5 for characterization of this internal deletion). We prepared cDNA from lymphoblastoid cells corresponding to one of these *SRGAP2B*-deletion homozy-

gotes and observed no full-length *SRGAP2B* transcript by RT-PCR, which is in contrast to samples carrying the paralog (Figure S3). Because the frequency of homozygotes is consistent with Hardy-Weinberg Equilibrium expectation and these individuals are representatives of the sample populations, the discovery of *SRGAP2B*-homozygous deletions in a "normal" population argues against a critical functional role of this copy in brain development. We addition-ally applied our method to 34 nonhuman primates and the Denisova and Neanderthal genomes (Green et al., 2010; Reich et al., 2010) and found that, consistent with our sequence-based estimations of the timing of the duplication events, *SRGAP2B*, *SRGAP2C*, and *SRGAP2D* copies are absent from all assayed nonhuman great apes yet are present in both the Neanderthal and Denisova genomes. We conclude that no new *SRGAP2* duplications have occurred since *Homo sapiens* and *Homo neanderthalensis* diverged about 1 mya.

Although it is common to observe a functional progenitor duplicated gene fixed in copy number, the discovery that a gene as recently evolved as *SRGAP2C* is fixed at a diploid copy number state is striking. When compared to the 23 genes duplicated specifically in the human lineage, we previously found that *SRGAP2* is among the six least copy number polymorphic gene families under a naive analysis that does not distinguish paralogs (Sudmant et al., 2010). When we extend this analysis to human-specific duplicates for which complete sequence is available and limit our analysis solely to those genes (n = 23), we find that *SRGAP2C* is the least copy number variable gene
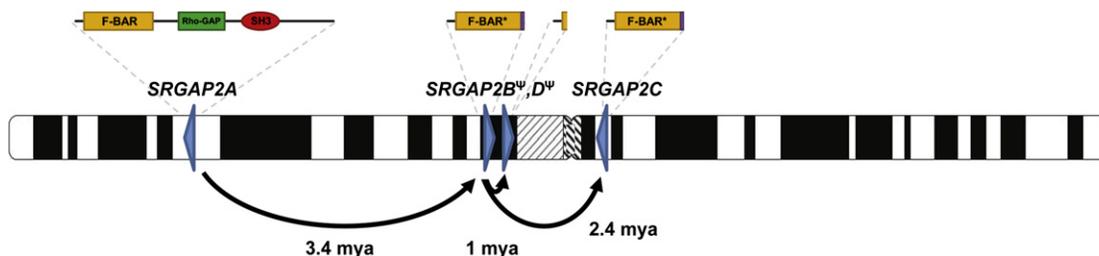
Figure 5. Model for *SRGAP2* Evolution
Schematic depicts location and orientation (blue triangles) of *SRGAP2* paralogs on human chromosome 1 with putative protein products indicated above each based on cDNA sequencing. Asterisks indicate a 49 amino acid truncation of the F-BAR domain. Note that the orientation of *SRGAP2D* remains uncertain, as the contig containing this paralog has not yet been anchored. Arrows trace the evolutionary history of *SRGAP2* duplication events. Copy number polymorphism and expression analyses suggest both paralogs at 1q21.1 (*SRGAP2B* and *SRGAP2D*) are pseudogenes, whereas the 1q32.1 (*SRGAP2A*) and 1p12 (*SRGAP2C*) paralogs are likely to encode functional proteins.

duplicate. Using qPCR assays that specifically assess copy number variation of *SRGAP2C*, we investigated this experimentally and found one individual harboring an ∼1 Mbp duplication containing numerous genes in an additional set of 1,794 controls (Table 2 and Figure S4). Applying this same assay to patients with intellectual disability and/or autism spectrum disorder (n = 4,475), we identified three additional individuals carrying large duplications of this locus. Strikingly, in our cumulative analysis of 7,137 individuals (cases and controls), we detected no deletions of *SRGAP2C*. In total, our combined analyses indicate that both *SRGAP2A* and *SRGAP2C* copies are nearly fixed at a copy number of two in all human populations assayed, with rare deletions and duplications observed only in cases with intellectual disability for *SRGAP2A* (p = 0.055, Fisher's exact test) and rare duplications observed at a frequency of ∼0.06% for *SRGAP2C*.

DISCUSSION

SRGAP2 has been highly conserved over mammalian evolution, and human is the only lineage wherein gene duplications have occurred. Our analysis indicates that the duplications spread across 80 Mbp of chromosome 1 at a time corresponding to the transition from *Australopithecus* to *Homo* (Figure 5). This included an initial large interspersed duplication (258 kbp) from chromosome 1q32.1 to 1q21.1, creating *SRGAP2B* ∼3.4 mya. The initial duplication was followed by larger (>515 kbp), secondary duplications of the 1q21.1 locus, creating *SRGAP2C* and *SRGAP2D* (∼2.4 and 1 mya, respectively). Consistent with these timing estimates, archaic *Homo* species, including Neanderthal and Denisova, carry these *SRGAP2* paralogs (Figure S5). It is intriguing that the general timing of the potentially functional copies, *SRGAP2B* and *SRGAP2C*, corresponds to the emergence of the genus *Homo* from *Australopithecus* (2–3 mya). This period of human evolution has been associated with the expansion of the neocortex and the use of stone tools, as well as dramatic changes in behavior and culture (Jobling et al., 2004).

Our analysis provides insight into one mechanism by which gene duplicates evolve. We find that the initial genomic duplication of *SRGAP2* was incomplete, encompassing the promoter and first nine exons of a 22 exon gene. Because SRGAP2 has

been shown to homodimerize via its F-BAR domain (Guerrier et al., 2009), we propose that incomplete segmental duplication of the gene ∼3.4 mya created an antagonistic functional state. In fact, functional evidence suggests that these partial *SRGAP2* copies produce protein with a nearly complete F-BAR domain but are missing other functional domains. These copies also heterodimerize with the full-length SRGAP2, creating a de facto dominant negative interaction equivalent to a knockdown of the ancestral copy (Charrier et al., 2012 [this issue of *Cell*]). The large size of the segmental duplication included the putative *cis* regulatory machinery of this gene and ensured that the duplicate genes would be developmentally coexpressed with the parental copy. Experimental analyses indicate (Guerrier et al., 2009; Charrier et al., 2012) that if the segmental duplication had been slightly larger (i.e., included exon 10), such antagonism would not be possible.

The incomplete nature of the segmental duplication was, therefore, ideal to establish this new function by virtue of its structure, which arose at the time of its "birth." This model of gene duplication that involves an "instantaneous" dominant negative function at birth stands in stark contrast to the favored model that involves duplication of a complete gene followed by the gradual accumulation of adaptive mutational events leading toward subfunctionalization or neofunctionalization (Lynch and Katju, 2004). We suggest that *SRGAP2C* ultimately assumed the antagonistic function of the *SRGAP2B* duplicate, which shows evidence of pseudogenization in contemporary humans. Although all four *SRGAP2* paralogs show evidence of transcription, it is unlikely that the two copies at 1q21.1 are now functional for several reasons. *SRGAP2B* has a markedly reduced expression in human brain compared to *SRGAP2C*. Likewise, the transcripts produced by *SRGAP2D* lack two internal exons, leading to a premature termination codon. Therefore, this copy is unlikely to produce a functional protein. Both *SRGAP2B* and *SRGAP2D* are highly copy number polymorphic, with normal individuals identified that completely lack these paralogs. This argues that if there is a phenotypic consequence to their complete deletion, it is likely to be relatively minor.

In stark contrast, both the *SRGAP2A* (progenitor) and *SRGAP2C* (granddaughter) paralogs are nearly fixed at a diploid state based on our analysis of 28,153 and 7,137 human DNA

samples, respectively. If we assume that the original *SRGAP2B* function was acquired by *SRGAP2C*, there is a possibility that both paralogs were functional at some point during human evolution. It is interesting that the comparison of the >515 kbp of duplicated sequence shared between *SRGAP2B* and *SRGAP2C* indicates that *SRGAP2B* has been subjected to large upstream deletions (103 kbp and 49 kbp in size), whereas *SRGAP2C* has not. Thus, the genomic instability of the *SRGAP2B* locus and its reduced expression in the contemporary human brain imply that the 1q21.1 locus may have been a suboptimal environment for gene transcription. The duplication event that yielded *SRGAP2C* ~2.4 mya may have provided a means of escape, transporting this truncated gene to a much more stable genomic environment for robust, long-term expression. One cannot, of course, definitively exclude the possibility that *SRGAP2B* and *SRGAP2D* transcripts may still confer some function (Charrier et al., 2012), perhaps via transcript regulation, but the finding of apparently normal individuals completely missing these duplicate copies would suggest that they are not critical for normal development.

We have identified larger deletions of the ancestral locus, *SRGAP2A*, only among children with developmental delay. Although the deletion intervals are large and other genes contributing to the disease phenotype cannot be excluded at this time, the absence of structural variation in the normal population and the discovery of a de novo translocation (Saitsu et al., 2011), as well as a second patient with a duplication breakpoint mapping within *SRGAP2*, provide some evidence of its role in brain development. In this light, the fixation of the duplicated *SRGAP2C* is especially noteworthy. *SRGAP2C* was found to be the least copy number polymorphic of all human-specific duplicate genes, despite the fact that it is embedded in a complex region prone to nonallelic homologous recombination. Our data, thus, point to two functional *SRGAP2* copies at 1p12 and 1q32.1, consistent with experimental characterization (Charrier et al., 2012). Based on these data, we propose more systematic screening of these genes for mutations in children with developmental delay and brain malformations that include West Syndrome, agenesis of the corpus callosum, and epileptic encephalopathies. This will be particularly challenging because most commercial SNP microarrays have failed to include probes from these duplicated regions, and reads from next-generation sequencing platforms are typically too short to assign to a specific paralog (Eichler et al., 2010). Nevertheless, final proof of the functional significance of these genes will rest on the discovery of disruptive mutations associated with human phenotypes.

Finally, we emphasize that much of the genomic sequence corresponding to the ancestral and duplicate gene copies was missing or misassembled in the current human reference genome. In this study, we sequenced, corrected, and annotated ~0.4% of the euchromatin of chromosome 1 more than 6 years after the "finished" human genome was declared (IHGSC, 2004). This was possible because the clone-based resource we developed using a complete hydatidiform mole essentially provides a haploid version of the human genome. Because this resource is devoid of allelic variation, we can rapidly distinguish even highly identical duplicate genes, thus providing a clear path forward for the characterization of other complex duplicated regions. It is worthwhile noting that we ensured the hydatidiform mole primary cell line (CHM1hTERT) we used did not contain any large CNVs that could confound our analysis (Fan et al., 2002). It is especially intriguing that *SRGAP2* is only one of several human-specific duplicate genes missing or incompletely assembled in the human genome (Sudmant et al., 2010). A number of remaining genes (e.g., *GPRIN2*, *GTF2IRD2*, and *HYDIN*) in this category have been implicated in neurodevelopment, neurite outgrowth, and behavior (Brunetti-Pierri et al., 2008; Chen et al., 1999; Dai et al., 2009). Additionally, human-specific protein-coding genes derived de novo from noncoding DNA merit further exploration (Wu et al., 2011). We propose that these uncharacterized human-specific genes constitute important pieces in the puzzle underlying the genetic basis of human brain evolution.

EXPERIMENTAL PROCEDURES

Fluorescent In Situ Hybridization

Metaphase spreads were prepared from lymphoblastoid human cell lines (NA12878, NA19317, NA20334, NA19901, NA19700, and NA19005; Coriell Cell Repository, Camden, NJ), a chimpanzee cell line (Douglas, provided by Dr. Mariano Rocchi), and an orangutan cell line (PR01109, a.k.a. Susie; Coriell Cell Repository, Camden, NJ). FISH experiments were performed using fosmid clones (Extended Experimental Procedures) as described previously (Antonacci et al., 2010).

Cloning Using a Complete Hydatidiform Mole Library

A large-insert BAC library (CHORI-17) was generated from a well-characterized complete hydatidiform mole primary cell culture (CHM1hTERT) using a modified protocol (Osoegawa et al., 1998) (http://bacpac.chori.org/library.php?id=231). To ensure the quality of CHM1hTERT, a karyotype analysis and extensive SNP genotyping with 1,494 SNP markers (Fan et al., 2002) and array comparative genomic hybridization (CGH) using the NimbleGen 2.1 M whole-genome array were performed. We generated paired-end sequences (n = 169,022) by using Sanger dideoxy methods, and we mapped sequence reads to the human reference genome. This provided a haplotype-resolved tiling path of clones for selection and sequencing (Kidd et al., 2008).

Sequencing and Assembly

We selected BAC clones with at least one sequenced end mapping to a *SRGAP2* region in the human reference genome and completely sequenced and assembled the insert (see Extended Experimental Procedures for detailed clone order, sequence assembly, and annotation). Inserts overlapping with >99.9% sequence identity were assembled into distinct contigs corresponding to *SRGAP2* loci at 1q32.1, 1q21.1, and 1p12.

Phylogenetic Analysis

We created a 244.2 kbp multiple sequence alignment from three completely sequenced *SRGAP2* genomic loci (ClustalW; Thompson et al., 2002) and constructed an unrooted phylogenetic tree (MEGA; Tamura et al., 2011) by using the neighbor-joining method (Saitou and Nei, 1987) with the complete-deletion option. Genetic distances were computed with the Kimura two-parameter method (Kimura, 1980) with standard error estimates (an interior branch test of phylogeny [Dopazo, 1994; Rzhetsky and Nei, 1994]; n = 500 bootstrap replicates). For the incompletely sequenced *SRGAP2D* paralog and the 1p12 chromosomal distal region, we created phylogenetic trees by using a 9.5 kbp and 50 kbp multiple species alignment, respectively (see Extended Experimental Procedures for details). The orthologous *SRGAP2* exons were extracted from different mammalian reference genomes without segmental duplications and were used to test various models of selection using a maximum-likelihood framework (codemL; PAML statistical software package [Yang, 2007]).

REFERENCES

Antonacci, F., Kidd, J.M., Marques-Bonet, T., Teague, B., Ventura, M., Girirajan, S., Alkan, C., Campbell, C.D., Vives, L., Malig, M., et al. (2010). A large and complex structural polymorphism at 16p12.1 underlies microdeletion disease risk. Nat. Genet. 42, 745–750.

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

Bailey, J.A., Liu, G., and Eichler, E.E. (2003). An Alu transposition model for the origin and expansion of human segmental duplications. Am. J. Hum. Genet. 73, 823–834.

Biesecker, L.G., Mullikin, J.C., Facio, F.M., Turner, C., Cherukuri, P.F., Blakesley, R.W., Bouffard, G.G., Chines, P.S., Cruz, P., Hansen, N.F., et al; NISC Comparative Sequencing Program. (2009). The ClinSeq Project: piloting large-scale genome sequencing for research in genomic medicine. Genome Res. 19, 1665–1674.

Blekhman, R., Marioni, J.C., Zumbo, P., Stephens, M., and Gilad, Y. (2010). Sex-specific and lineage-specific alternative splicing in primates. Genome Res. 20, 180–189.

Brunet, M., Guy, F., Pilbeam, D., Mackaye, H.T., Likius, A., Ahounta, D., Beauvilain, A., Blondel, C., Bocherens, H., Boisserie, J.R., et al. (2002). A new hominid from the Upper Miocene of Chad, Central Africa. Nature 418, 145–151.

Brunet, M., Guy, F., Pilbeam, D., Lieberman, D.E., Likius, A., Mackaye, H.T., Ponce de León, M.S., Zollikofer, C.P., and Vignaud, P. (2005). New material of the earliest hominid from the Upper Miocene of Chad. Nature 434, 752–755.

Brunetti-Pierri, N., Berg, J.S., Scaglia, F., Belmont, J., Bacino, C.A., Sahoo, T., Lalani, S.R., Graham, B., Lee, B., Shinawi, M., et al. (2008). Recurrent reciprocal 1q21.1 deletions and duplications associated with microcephaly or macrocephaly and developmental and behavioral abnormalities. Nat. Genet. 40, 1466–1471.

Charrier, C., Joshi, K., Coutinho-Budd, J., Kim, J.-E., Lambert, N., de Marchena, J., Jin, W.-L., Vanderhaeghen, P., Ghosh, A., Sassa, T., and Polleux, F. (2012). Inhibition of SRGAP2 function by its human-specific paralogs induces neoteny during spine maturation. Cell 149. Published online May 3, 2012. 10.1016/j.cell.2012.03.034.

Chen, L.T., Gilman, A.G., and Kozasa, T. (1999). A candidate target for G protein action in brain. J. Biol. Chem. 274, 26931–26938.

Cooper, G.M., Coe, B.P., Girirajan, S., Rosenfeld, J.A., Vu, T.H., Baker, C., Williams, C., Stalker, H., Hamid, R., Hannig, V., et al. (2011). A copy number variation morbidity map of developmental delay. Nat. Genet. 43, 838–846.

CSAC (Chimpanzee Sequencing and Analysis Consortium). (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437, 69–87.

Dai, L., Bellugi, U., Chen, X.N., Pulst-Korenberg, A.M., Järvinen-Pasley, A., Tirosh-Wagner, T., Eis, P.S., Graham, J., Mills, D., Searcy, Y., and Korenberg, J.R. (2009). Is it Williams syndrome? GTF2IRD1 implicated in visual-spatial construction and GTF2I in sociability revealed by high resolution arrays. Am. J. Med. Genet. A. 149A, 302–314.

Dopazo, J. (1994). Estimating errors and confidence intervals for branch lengths in phylogenetic trees by a bootstrap approach. J. Mol. Evol. 38, 300–304.

Eichler, E.E. (2001). Segmental duplications: what's missing, misassigned, and misassembled—and should we care? Genome Res. 11, 653–656.

Eichler, E.E., Flint, J., Gibson, G., Kong, A., Leal, S.M., Moore, J.H., and Nadeau, J.H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. Nat. Rev. Genet. 11, 446–450.

Enard, W., Przeworski, M., Fisher, S.E., Lai, C.S., Wiebe, V., Kitano, T., Monaco, A.P., and Pääbo, S. (2002). Molecular evolution of *FOXP2*, a gene involved in speech and language. Nature *418*, 869–872.

Fan, J.B., Surti, U., Taillon-Miller, P., Hsie, L., Kennedy, G.C., Hoffner, L., Ryder, T., Mutch, D.G., and Kwok, P.Y. (2002). Paternal origins of complete hydatidiform moles proven by whole genome single-nucleotide polymorphism haplotyping. Genomics *79*, 58–62.

Fischbach, G.D., and Lord, C. (2010). The Simons Simplex Collection: a resource for identification of autism genetic risk factors. Neuron *68*, 192–195.

Fortna, A., Kim, Y., MacLaren, E., Marshall, K., Hahn, G., Meltesen, L., Brenton, M., Hink, R., Burgers, S., Hernandez-Boussard, T., et al. (2004). Lineage-specific gene duplication and loss in human and great ape evolution. PLoS Biol. *2*, E207.

Geschwind, D.H., Sowinski, J., Lord, C., Iversen, P., Shestack, J., Jones, P., Ducat, L., and Spence, S.J.; AGRE Steering Committee. (2001). The autism genetic resource exchange: a resource for the study of autism and related neuropsychiatric conditions. Am. J. Hum. Genet. *69*, 463–466.

Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H., et al. (2010). A draft sequence of the Neandertal genome. Science *328*, 710–722.

Gregory, S.G., Barlow, K.F., McLay, K.E., Kaul, R., Swarbreck, D., Dunham, A., Scott, C.E., Howe, K.L., Woodfine, K., Spencer, C.C., et al. (2006). The DNA sequence and biological annotation of human chromosome 1. Nature *441*, 315–321.

Guerrier, S., Coutinho-Budd, J., Sassa, T., Gresset, A., Jordan, N.V., Chen, K., Jin, W.L., Frost, A., and Polleux, F. (2009). The F-BAR domain of srGAP2 induces membrane protrusions required for neuronal migration and morphogenesis. Cell *138*, 990–1004.

Guo, S., and Bao, S. (2010). srGAP2 arginine methylation regulates cell migration and cell spreading through promoting dimerization. J. Biol. Chem. *285*, 35133–35141.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al; International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

IHGSC (International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. Nature *431*, 931–945.

Jobling, M., Hurles, M., and Tyler-Smith, C. (2004). Human Evolutionary Genomics (New York: Garland Science).

Kajii, T., and Ohama, K. (1977). Androgenetic origin of hydatidiform mole. Nature *268*, 633–634.

Kidd, J.M., Cooper, G.M., Donahue, W.F., Hayden, H.S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., et al. (2008). Mapping and sequencing of structural variation from eight human genomes. Nature *453*, 56–64.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. *16*, 111–120.

Liu, S., Lin, L., Jiang, P., Wang, D., and Xing, Y. (2011). A comparison of RNA-Seq and high-density exon array for detecting differential gene expression between closely related species. Nucleic Acids Res. *39*, 578–588.

Lynch, M., and Conery, J.S. (2000). The evolutionary fate and consequences of duplicate genes. Science *290*, 1151–1155.

Lynch, M., and Katju, V. (2004). The altered evolutionary trajectories of gene duplicates. Trends Genet. *20*, 544–549.

Marques-Bonet, T., Kidd, J.M., Ventura, M., Graves, T.A., Cheng, Z., Hillier, L.W., Jiang, Z., Baker, C., Malfavon-Borja, R., Fulton, L.A., et al. (2009). A burst of segmental duplications in the genome of the African great ape ancestor. Nature *457*, 877–881.

Ohno, S. (1970). Evolution by Gene Duplication (New York: Springer-Verlag).

Osoegawa, K., Woon, P.Y., Zhao, B., Frengen, E., Tateno, M., Catanese, J.J., and de Jong, P.J. (1998). An improved approach for construction of bacterial artificial chromosome libraries. Genomics *52*, 1–8.

Parsons, J.D. (1995). Miropeats: graphical DNA sequence comparisons. Comput. Appl. Biosci. *11*, 615–619.

Patterson, N., Richter, D.J., Gnerre, S., Lander, E.S., and Reich, D. (2006). Genetic evidence for complex speciation of humans and chimpanzees. Nature *441*, 1103–1108.

Prabhakar, S., Visel, A., Akiyama, J.A., Shoukry, M., Lewis, K.D., Holt, A., Plajzer-Frick, I., Morrison, H., Fitzpatrick, D.R., Afzal, V., et al. (2008). Human-specific gain of function in a developmental enhancer. Science *321*, 1346–1350.

Reich, D., Green, R.E., Kircher, M., Krause, J., Patterson, N., Durand, E.Y., Viola, B., Briggs, A.W., Stenzel, U., Johnson, P.L., et al. (2010). Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature *468*, 1053–1060.

Rzhetsky, A., and Nei, M. (1994). METREE: a program package for inferring and testing minimum-evolution trees. Comput. Appl. Biosci. *10*, 409–412.

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

Saitsu, H., Osaka, H., Sugiyama, S., Kurosawa, K., Mizuguchi, T., Nishiyama, K., Nishimura, A., Tsurusaki, Y., Doi, H., Miyake, N., et al. (2011). Early infantile epileptic encephalopathy associated with the disrupted gene encoding Slit-Robo Rho GTPase activating protein 2 (SRGAP2). Am. J. Med. Genet. *158A*, 199–205.

Sharp, A.J., Locke, D.P., McGrath, S.D., Cheng, Z., Bailey, J.A., Vallente, R.U., Pertz, L.M., Clark, R.A., Schwartz, S., Segraves, R., et al. (2005). Segmental duplications and copy-number variation in the human genome. Am. J. Hum. Genet. *77*, 78–88.

Stedman, H.H., Kozyak, B.W., Nelson, A., Thesier, D.M., Su, L.T., Low, D.W., Bridges, C.R., Shrager, J.B., Minugh-Purvis, N., and Mitchell, M.A. (2004). Myosin gene mutation correlates with anatomical changes in the human lineage. Nature *428*, 415–418.

Sudmant, P.H., Kitzman, J.O., Antonacci, F., Alkan, C., Malig, M., Tsalenko, A., Sampas, N., Bruhn, L., Shendure, J., and Eichler, E.E.; 1000 Genomes Project. (2010). Diversity of human copy number variation and multicopy genes. Science *330*, 641–646.

Szamalek, J.M., Goidts, V., Cooper, D.N., Hameister, H., and Kehrer-Sawatzki, H. (2006). Characterization of the human lineage-specific pericentric inversion that distinguishes human chromosome 1 from the homologous chromosomes of the great apes. Hum. Genet. *120*, 126–138.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. *28*, 2731–2739.

Thompson, J.D., Gibson, T.J., and Higgins, D.G. (2002). Multiple sequence alignment using ClustalW and ClustalX. In Current Protocols in Bioinformatics (New York: John Wiley & Sons), 2.3.1–2.3.22.

Vignaud, P., Duringer, P., Mackaye, H.T., Likius, A., Blondel, C., Boisserie, J.R., De Bonis, L., Eisenmann, V., Etienne, M.E., Geraads, D., et al. (2002). Geology and palaeontology of the Upper Miocene Toros-Menalla hominid locality, Chad. Nature *418*, 152–155.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470–476.

Wu, D.D., Irwin, D.M., and Zhang, Y.P. (2011). *De novo* origin of human protein-coding genes. PLoS Genet. *7*, e1002379.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. *24*, 1586–1591.

Zhou, Y., and Mishra, B. (2005). Quantifying the mechanisms for segmental duplications in mammalian genomes by statistical analysis and modeling. Proc. Natl. Acad. Sci. USA *102*, 4051–4056.