Cell

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

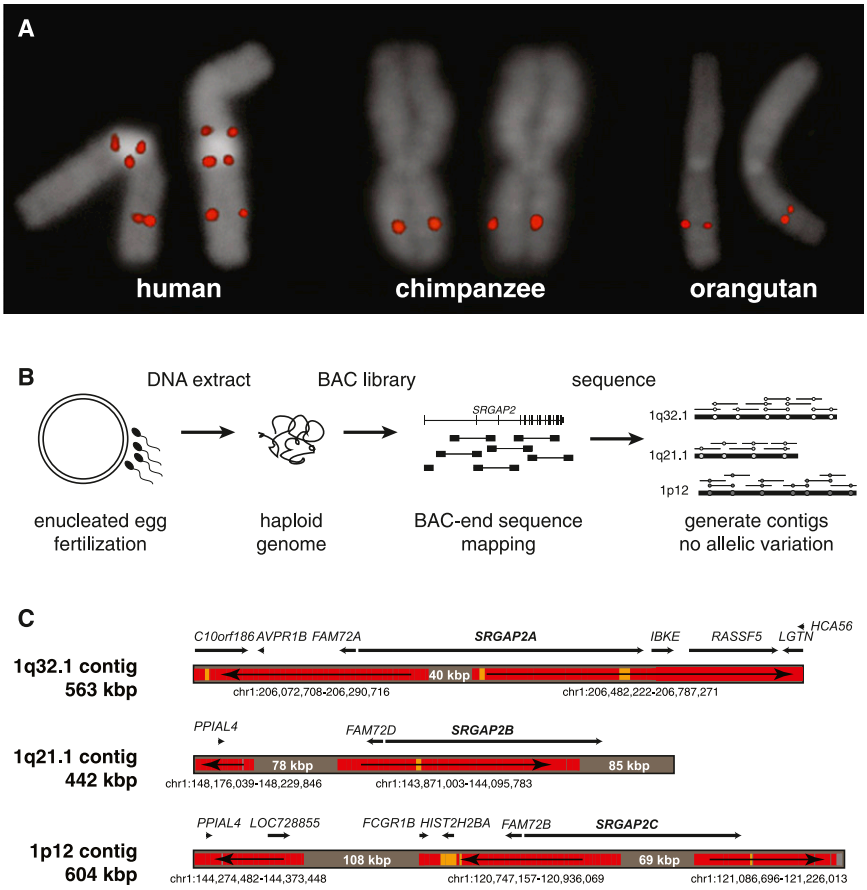

Figure 1. Genomic Characterization and Sequence Resolution of *SRGAP2* Loci

(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*

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

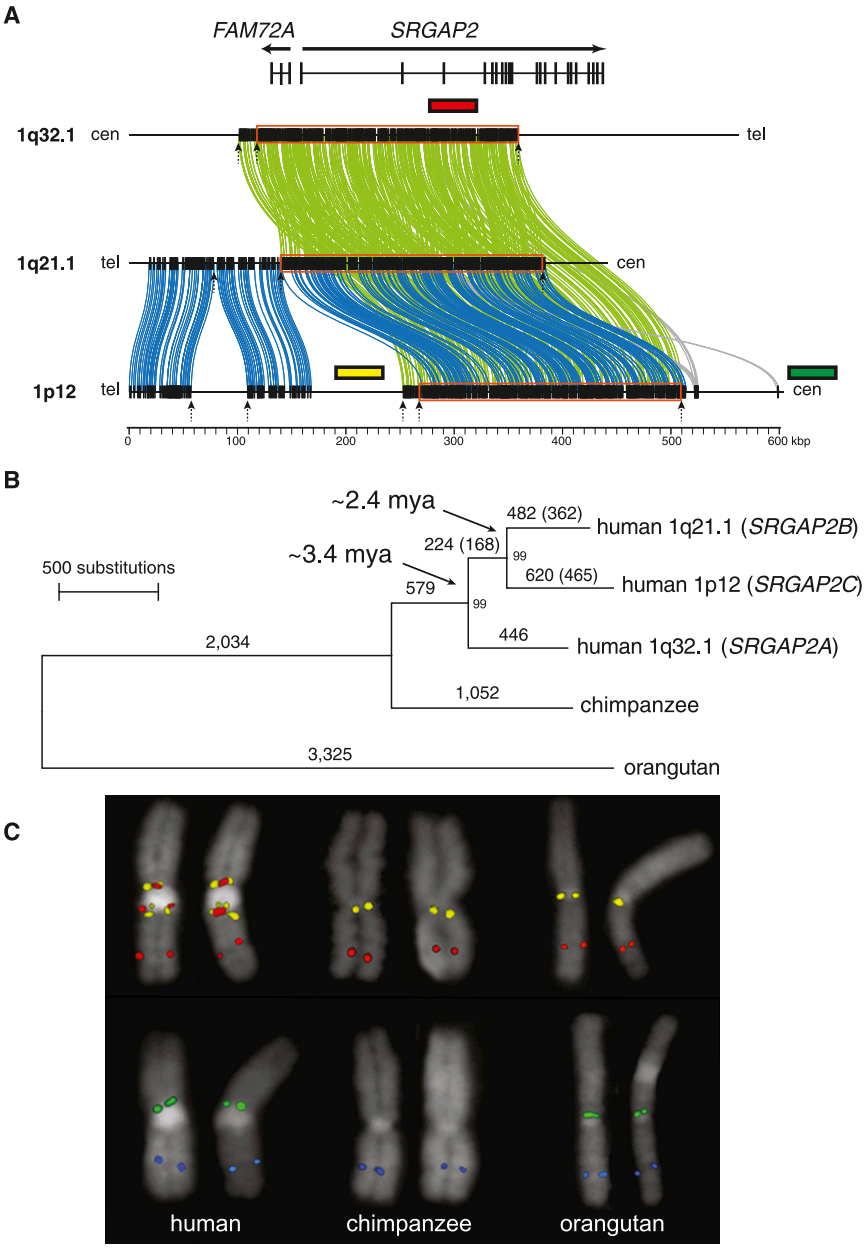

Figure 2. Evolutionary Characterization of *SRGAP2* Duplications

(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, $p = 1.32 \times 10^{-11}$, 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

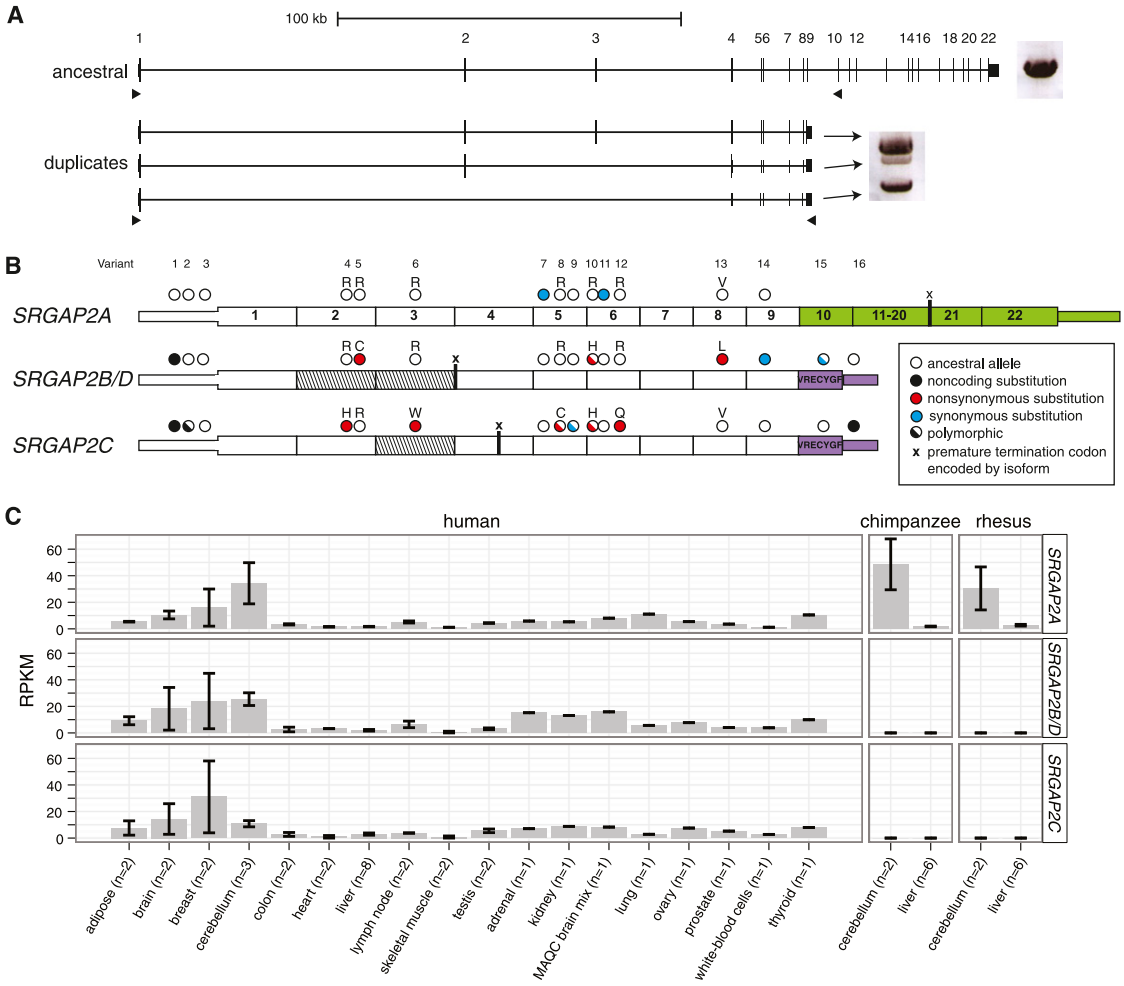

Figure 3. Paralog-Specific *SRGAP2* Gene Expression

(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* transcript 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;

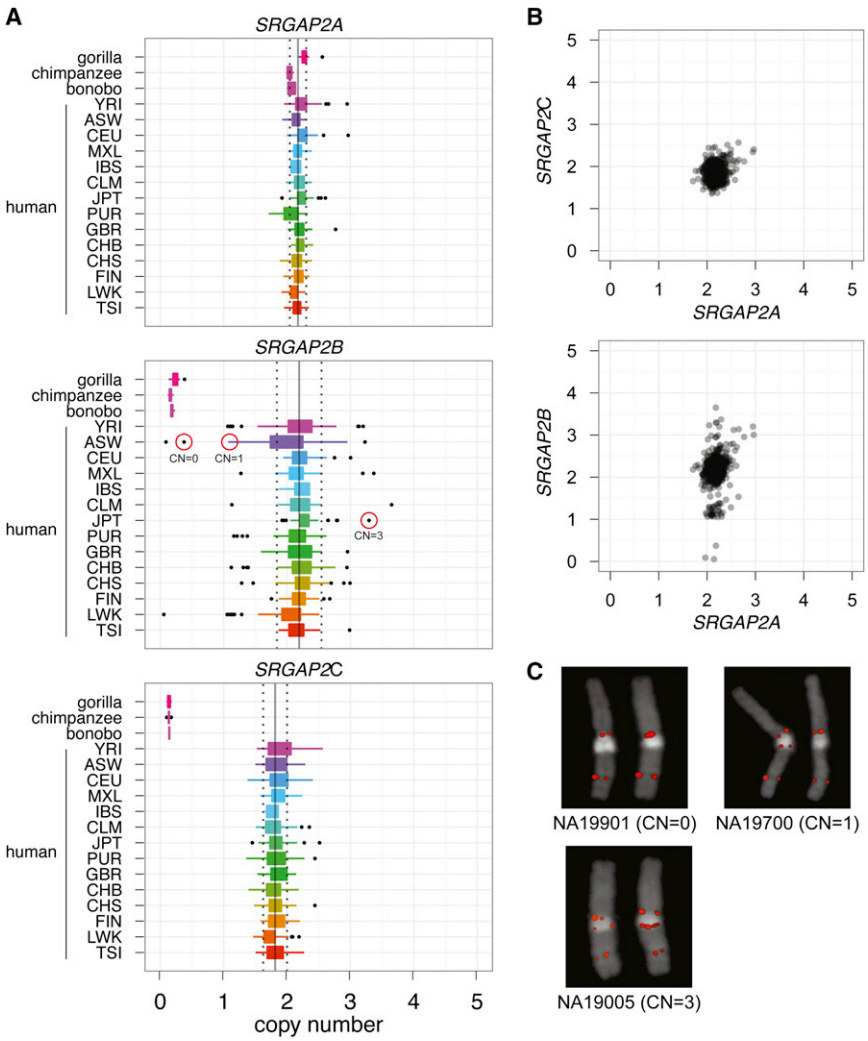

Figure 4. *SRGAP2* Copy Number Diversity in Human Populations

(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 homozygotes 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 additionally 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

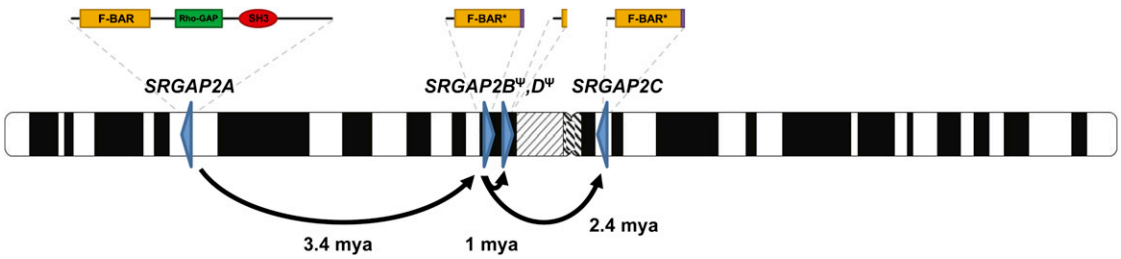

Figure 5. Model for *SRGAP2* Evolution

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]).

SRGAP2 Transcript Analysis

Total RNA was isolated using Trizol reagent (Invitrogen) and the RNeasy Mini Kit (QIAGEN) from SH-SY5Y neuronal cell line. Total RNA was analyzed from human fetal brain (collected from spontaneously aborted fetuses, 50–60 pooled samples, 20–33 weeks of development; ClonTech S2437) as well as a single fetal (R1244035, BioChain) and adult brain sample (M1234035, BioChain) (see Extended Experimental Procedures for details regarding RT-PCR, cDNA cloning, and sequencing). We also analyzed RNA-Seq data from 17 different human tissues (Illumina's Human BodyMap 2.0), seven human cell lines (Wang et al., 2008), and both chimpanzee and macaque cerebellum and liver tissues (Blekhman et al., 2010). Briefly, RNA-Seq data sets were mapped to the human reference genome (NCBI36/hg18) and our described *SRGAP2* contigs. Expression levels for specific paralogs were calculated in units of RPKM (reads per kilobase of exon model per million mapped reads) (Liu et al., 2011) with transcribed PSVs, which allowed RNA-Seq data to be unambiguously assigned to a specific paralog.

Paralog-Specific Copy Number Genotyping

CNVs in cases with intellectual disability and controls for *SRGAP2A* were identified from previously published array CGH data and SNP microarray data, respectively (Cooper et al., 2011). Copy number estimates of specific *SRGAP2* paralogs by using SUNs were determined using previously described methods (Sudmant et al., 2010). Custom qPCR assays were performed in triplicate using variants specific to each *SRGAP2* paralogous locus (see Extended Experimental Procedures for a description of variant detection and primer sequences). Validations of deletions and duplications, as well as identification of CNVs in the autism cohorts and some controls, were performed by array CGH using custom microarrays (Agilent) and a HapMap individual (NA18507) as a reference.

ACCESSION NUMBERS

The GenBank accession numbers for the sequences reported in this paper are listed in Table S5.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, five figures, and six tables and can be found with this article online at doi:10.1016/j.cell.2012.03.033.

ACKNOWLEDGMENTS

We thank B. Coe for assistance in CNV analysis and the 1000 Genomes Project for access to sequence data of the *SRGAP2* loci. For DNA samples used in paralog-specific CNV screening and detailed phenotypic information of patients, we would like to thank C. Romano, M. Fichera, J. Gécz, B. de Vries, R. Bernier, the Simons Foundation, Autism Speaks, the National Institute of Mental Health, and the ClinSeq Project. We acknowledge C. Baker, L. Vives, and J. Huddleston for technical assistance, T. Brown for manuscript editing, and the laboratory of S. Fields for use of their Roche LC480. We also thank J. Akey, T. Marques-Bonet, A. Andres, S. Girirajan, and K. Meltz Steinberg for helpful discussion, as well as the laboratory of F. Polleux for comments and kindly sharing human RNA samples for expression studies. The BAC clones from the complete hydatidiform mole were derived from a cell line created by U. Surti. M.Y.D. is supported by U.S. National Institutes of Health (NIH) Ruth L. Kirchstein National Research Service Award (NRSA) Fellowship (1F32HD071698-01). X.N. is supported by an NIH NRSA Genome Training Grant to the University of Washington (2T32HG000035-16). P.H.S. is a Howard Hughes Medical Institute International Student Research Fellow. This work was supported by NIH Grants HG002385 and GM058815. E.E.E. is an investigator of the Howard Hughes Medical Institute. J.A.R. and L.G.S. are employees of Signature Genomic Laboratories, a subsidiary of PerkinElmer, Inc. E.E.E. is on the scientific advisory boards for Pacific Biosciences, Inc. and SynapDx Corp.

Received: December 8, 2011
Revised: February 17, 2012
Accepted: March 1, 2012
Published online: May 3, 2012

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.

EXTENDED EXPERIMENTAL PROCEDURES

FISH Experiments

FISH experiments were used to detect *SRGAP2* paralogous regions (probes 1 and 9, see Table S1), resolve the chromosomal orientation of contigs (probes 2–7), infer the evolutionary order of duplication events (probes 1, 2, 7, and 8), and validate copy number polymorphism at the *SRGAP2B* locus (probe 1). Experiments were performed using clones obtained from a G248 fosmid library (Kidd et al., 2010), directly labeled by nick-translation with Cy3-dUTP (Perkin-Elmer), Cy5-dUTP (PerkinElmer), and fluorescein-dUTP (Enzo), as previously described (Antonacci et al., 2010) with minor modifications. Briefly, 300 ng of labeled probe were used for the FISH experiments; hybridization was performed at 37°C in 2xSSC, 50% (v/v) formamide, 10% (w/v) dextran sulfate, and 3 μg sonicated salmon sperm DNA, in a volume of 10 μl. Posthybridization washing was at 60°C in 0.1xSSC (three times, high stringency). Nuclei were simultaneously DAPI stained. Digital images were obtained using a Leica DMRXA2 epifluorescence microscope equipped with a cooled CCD camera (Princeton Instruments). DAPI, Cy3, Cy5, and fluorescein fluorescence signals, detected with specific filters, were recorded separately as gray-scale images. Pseudocoloring and merging of images were performed using Adobe Photoshop software.

We used a series of FISH assays to determine the chromosomal orientation (Figure S1) and evolutionary order (Figure 2C) of *SRGAP2* duplications. Because our contigs did not map concordantly to the human reference genome and spanned multiple gaps, we could not confidently infer the chromosomal orientation of our contigs based on comparison to the reference sequence. We performed three-color interphase FISH assays to resolve this. These assays made use of two probes mapping within our sequenced contig (or extensions of our contig based on contiguous reference sequence) and one probe mapping outside of our contig, closer to the telomere. FISH analyses with probes targeting regions upstream of the *SRGAP2B* paralog indicated that this sequence is extensively locally duplicated (see yellow probe in Figure 2C—even though the region it targets is deleted in our *SRGAP2B* contig, this sequence is still present in two haploid copies at chromosome 1q21.1). Because FISH analysis could not resolve the chromosomal orientation of *SRGAP2B* or *SRGAP2D*, we instead utilized an anchored contig spanning the entire 1q21.1 region recently generated at The Genome Institute at Washington University School of Medicine (T.G. and R.W., unpublished data). Comparison of our *SRGAP2B* contig showed that this paralog is transcribed toward the centromere, whereas the *SRGAP2D* orientation remains uncertain, as the contig containing this paralog has not yet been anchored.

Generation of Paralog-Specific Sequence Contigs

Due to the highly identical nature of sequences within segmental duplications, many genes embedded within duplicated regions are not properly represented in the human reference genome (Sharp et al., 2005). To generate and distinguish sequences corresponding to *SRGAP2* paralogs, we leveraged a large-insert BAC library (CHORI-17) generated from a well-characterized complete hydatidiform mole cell line (CHM1hTERT), a resource developed to resolve paralogous regions of the genome (http://www.genome.gov/Pages/Research/Sequencing/BACLibrary/HydatidiformMoleBAC021203.pdf). Clones were selected for sequencing based on mapping of BAC-end sequences to partial copies of *SRGAP2* within the human reference (Tuzun et al., 2005). Selected clones included both concordant and discordant clones as well as clones with only one end mapped to one of the *SRGAP2* paralogous regions (*SRGAP2A*, chr1:204,582,823-204,704,406; *SRGAP2B*, chr1:142,625,200-142,805,783; *SRGAP2C*, chr1:120,637,333-120,832,584; NCBI36/hg18). *SRGAP2*-containing discordant clones were validated by PCR amplification using primers targeting either intron 1 or intron 2 of *SRGAP2*. Additional clones from the CH17-BAC library were also obtained by mapping efforts at The Genome Institute at Washington University School of Medicine directed to improving the quality of the human reference genome (T.G. and R.W., unpublished data).

Clone inserts were completely sequenced using a hierarchical clone-based strategy with high-quality capillary fluorescent-based sequencing. This entailed the construction of genomic libraries, sequencing of paired-end shotgun libraries, and assembly of inserts into a finished sequencing contig for 22 distinct CH17 clones [see Table S5 with clone IDs (ordered as they assemble into contigs) and GenBank accessions]. We used sequence quality standards for sequencing and assembly commensurate to that applied to human genome reference (estimated 1 error in 100,000 bases) (Schmutz et al., 2004).

The *SRGAP2* coding exons were distinguished in all clones, and variants specific to each paralog were identified. Each BAC was assigned to a specific duplicated region by performing BLAST sequence similarity searches of *SRGAP2* exons against each BAC clone sequence. Coding sequence substitutions served as features that distinguished clones originating from different *SRGAP2* paralogs. We used these paralog-specific clones as a query in a BLAST search of the human HTGS (high-throughput genomic sequence) database to identify clones from the hydatidiform mole BAC library (CH17). Clones with >99.9% sequence identity could be confidently inferred to have originated from the paralog containing that particular substitution (as opposed to a sequence identity of <99.5% observed for clones mapping to other *SRGAP2* loci). Clones thus inferred to have originated from the same paralog containing overlapping sequence were combined into contigs, and the entire process was repeated iteratively using the clones at the ends to extend overlaps. The iterations were carried out until no more hydatidiform mole clone sequences could be confidently inferred to have originated from the paralog represented by the contig. The contigs were assembled using Sequencher 4.9. Alignment quality was ensured by manual inspection and editing when necessary. This method allowed us to ultimately generate three single-haplotype contigs containing the sequences of three *SRGAP2* paralogs and their flanking sequences.

Comparison of Sequences from *SRGAP2* Contigs with the Human Reference Genome

We utilized pairwise BLAST (Altschul et al., 1990) to identify missing or mismapped regions within the human reference genome. Comparing shared sequence between contigs revealed high sequence identity between *SRGAP2* paralogs (99%–99.75%) with the *SRGAP2B* and *SRGAP2D* paralogs as the most highly identical. Based on these identities, we performed pairwise BLAST alignments of our complete contigs against the human reference genome (GRCh37/hg19) and identified extended contiguous regions of the reference (at least 5,000 bp in length) having >99.6% sequence identity to a particular contig. These regions were mapped back to their respective contigs allowing us to identify any sequences missing or mismapped within the reference.

Breakpoint Analysis of *SRGAP2* Duplicated Regions

Pairwise BLAST was used to identify extended contiguous regions of high sequence identity between every pair of paralogous sequence contigs. We observe that the original duplication event (258,245 bp) encompassed the promoter and first nine exons of *SRGAP2*. A larger, second duplication event (>515 kbp) originated from the daughter *SRGAP2* paralog and included the entire original duplicated sequence. Subsequently, two large deletions (102,605 bp and 48,969 bp) occurred upstream of *SRGAP2B*. The nucleotide sequences at the ends of these regions were recorded. The contigs were then aligned at these sequences using Sequencher 4.9 and manually checked, revealing breakpoints between *SRGAP2* paralogs and breakpoints due to other structural rearrangements at a high resolution (in most cases, single-nucleotide resolution). The local sequences surrounding these breakpoints were then assessed for the presence of repetitive elements using RepeatMasker (Tarailo-Graovac and Chen, 2009) and via a BLAST search against a database of human Alu repeats. To gain a better understanding of the content within the deleted regions upstream of *SRGAP2B*, we assessed the genes and potential regulatory elements residing within these deleted regions. Notably, only portions of the deleted regions are represented in the most current human reference (smaller deletion, chr1:144,275,483-144,312,889; larger deletion, chr1:120,872,120-120,936,069; 1p12-contig region represented in GRCh37/hg19). The smaller deletion (49 kbp) resides 195 kbp upstream of *SRGAP2* and contains two uncharacterized genes. The larger deletion (103 kbp) resides 34 kbp upstream of *SRGAP2*, just downstream of *FAM72*. The deleted region would have contained paralogs to *HIST72H2BA* and *FCGR1B* in addition to putative regulatory elements predicted by conserved transcription-factor binding site predictions and ChIP-Seq data (using "Regulation" tracks within the UCSC Genome Browser). This deletion does not affect any obvious promoter elements of *SRGAP2* that would reside directly upstream of the transcription start site.

SRGAP2 Duplication Timing Using a Chromosome 1q32.1 Molecular Clock

We created a multiple-species alignment (MSA) (ClustalW; Thompson et al., 2002) of the 244 kbp *SRGAP2* genomic region shared across the three human loci, chimpanzee, and orangutan orthologous regions. The chimpanzee (October 2010) and orangutan (July 2007) orthologous sequences were identified using BLAT in the UCSC Genome Browser with the 1q32.1 (*SRGAP2A*) human sequence as the query. We manually inspected the alignment using the Jalview (Waterhouse et al., 2009) editor and manually corrected alignment errors. We repeated the multiple sequence alignment using nonhuman primate orthologous segments, inspecting the alignment for errors each time. The final alignment was contiguous for human and chimpanzee sequences, with 12 gaps within the orangutan sequence spanning 12,639 bp of sequence [most gaps were small; two large gaps account for 9,146 bp and 1,523 bp of sequence]. From this MSA, we constructed an unrooted phylogenetic tree using the neighbor-joining method (Saitou and Nei, 1987) [MEGA (Tamura et al., 2011); complete-deletion option]. Genetic distances were computed using 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]. We noticed that the branch lengths of *SRGAP2B* and *SRGAP2C* were considerably longer (>30%) than *SRGAP2A* suggesting that the rates of substitution at the chromosomal regions 1q21.1 and 1p12 were higher than 1q32.1. Using Tajima's relative rate test (MEGA), we determined that *SRGAP2A* evolved at the same rate as orthologous counterparts in chimpanzee and orangutan ($p = 0.5345$) while both *SRGAP2B* and *SRGAP2C* evolved at an accelerated rate ($p = 0.0001$-0.0249). Using the genetic distance established for human *SRGAP2A*, we applied a correction factor to the average branch length leading to *SRGAP2B* and *SRGAP2C* in effect forcing the substitution rate of these branches to equal that of chromosome 1q32.1. In turn, we used a chimpanzee divergence time of 6 mya, noting that estimates range from ~5–7 mya since the human and chimpanzee split, based on fossil records (Brunet et al., 2005; Brunet et al., 2002; Vignaud et al., 2002) as well as recent genetic estimates (Patterson et al., 2006), to estimate the timing of the duplication events.

The phylogenetic tree with genetic distances represented as percent of substitutions per total number of aligned sites (244,200 bp) and the standard errors:

(((((((human_*SRGAP2B*: 0.197 ± 0.0102, human_*SRGAP2C*: 0.254 ± 0.0062),:0.092 ± 0.0038), human_*SRGAP2A*:0.237 ± 0.0057),: 0.183 ± 0.0053), chimpanzee: 0.431 ± 0.0077),: 0.833 ± 0.0175), orangutan: 1.36 ± 0.0175).

To account for the increased substitution rate along the *SRGAP2B* and *SRGAP2C* branches (while conservatively leaving the standard error uncorrected), we calculated and applied a correction factor of 0.75:

$D_{SRGAP2B/C} = ½(0.197+0.254)+0.092 = 0.318$
$D_{SRGAP2A} = 0.237$
Correction factor = 0.237/0.318 = 0.75

Corrected phylogenetic tree:

(((((((human_*SRGAP2B*: 0.148 ± 0.0102, human_*SRGAP2C*: 0.191 ± 0.0062),: 0.069 ± 0.0038), human_*SRGAP2A*:0.237 ± 0.0057),:0.183 ± 0.0053), chimpanzee: 0.431 ± 0.0077),:0.833 ± 0.0175), orangutan: 1.36 ± 0.0175).

To estimate the evolutionary timing of the duplication events, we used R = D/2T:

$\text{Rate} = [D_{\text{chimpanzee/human},SRGAP2A}]/2T = (0.237+0.183+0.431)/2T = 0.426/T$
$\text{Rate}_{T = 6\text{mya}} = 0.0709\%$ substitutions/site/mya
$\text{Rate}_{T = 5\text{mya}} = 0.0852\%$ substitutions/site/mya
$\text{Rate}_{T = 7\text{mya}} = 0.0609\%$ substitutions/site/mya

From this, we estimated the timing of the initial *SRGAP2* duplication event:

$T = D_{\text{human},SRGAP2A} / R_{T = 6\text{mya}} = (0.237)/(0.0709) = 3.4$ mya
$T_{\text{lower}} = D_{\text{human},SRGAP2A} / R_{T = 5\text{mya}} = (0.237)/(0.0852) = 2.8$ mya
$T_{\text{upper}} = D_{\text{human},SRGAP2A} / R_{T = 7\text{mya}} = (0.237)/(0.0609) = 3.9$ mya

Likewise, we estimated the timing of the secondary duplication event:

$T_{\text{lower}} = D_{\text{human},SRGAP2B/\text{human},SRGAP2C} / 2R_{T = 6\text{mya}} = (0.148+0.191)/(2*0.0709) = 2.4$ mya
$T_{\text{lower}} = D_{\text{human},SRGAP2B/\text{human},SRGAP2C} / 2R_{T = 5\text{mya}} = (0.148+0.191)/(2*0.0852) = 2.0$ mya
$T_{\text{upper}} = D_{\text{human},SRGAP2B/\text{human},SRGAP2C} / 2R_{T = 7\text{mya}} = (0.148+0.191)/(2*0.0609) = 2.8$ mya

There is error in these divergence rate estimates, but this is minor compared to the inherent error in the chimpanzee-human divergence time estimates. To convey this, we have reported the standard errors of genetic distances in Table 1 of the main text.

Molecular Evolution of 1p12 Genomic Region Distal to *SRGAP2C*

As a control, we estimated the genomic substitution rate for the chromosomal 1p12 region. Specifically, we obtained a 50 kbp MSA using the ENSEMBL genome browser including orthologous sequences from human, chimpanzee, gorilla, orangutan, and rhesus macaque (GRCh37/hg19; chr1:120,193,477-120,253,477). This region maps approximately 2 Mbp distal to the *SRGAP2* duplication region based on the human reference sequence. Tajima's relative rate tests indicated that the sequences were evolving at a constant rate (Bonferroni corrected p = 0.159-1.0). Creating a neighbor-joining phylogenetic tree (as described above) and assuming a human-chimpanzee divergence time of 6 mya, we estimated a chromosome 1p12 substitution rate of $9.38 \pm 1.07\text{x}10^{-4}$ substitutions per site per million years (assuming a human-chimpanzee split of 6 mya), which is ~30% higher than that of 1q32.1 ($7.09 \pm 0.183\text{x}10^{-4}$ substitutions per site per million years) and consistent with our locus-specific correction factor. Using the standard error of the 1q32.1 and 1p12 estimates, we calculated a range of percent differences between rates.

Molecular Evolution of *SRGAP2D* Genomic Region

Based on partial sequence of clone CH17-248H7, we constructed a smaller 9.5 kbp *SRGAP2* MSA (ClustalW) including sequence from *SRGAP2D*. The tree topology (99% bootstrap support) and sequence identity comparisons both strongly suggest *SRGAP2D* arose via a duplication of the *SRGAP2B* paralog. The increase in substitution rates across the 1p12 and 1q21.1 regions is not evident in our analysis of this much smaller genomic region. Based on an assumption that the rate is indeed accelerated on these duplicate branches, we calculated the timing of this third duplication utilizing the same correction factor as before. The upper and lower estimates of timing for this duplication were estimated using standard error of branch lengths.

Molecular Evolution of the *SRGAP2* Coding Region

We assessed the level of conservation of *SRGAP2* across mammals by estimating its rate of protein evolution in a mammalian phylogeny. Specifically, we assessed the ratio (dN/dS) of nonsynonymous substitutions per nonsynonymous site (dN) to synonymous substitutions per synonymous site (dS). Because purifying selection acts to eliminate protein-coding changes, dN/dS decreases with stronger purifying selection. Alternatively, dN/dS increases with relaxed constraint and/or positive selection.

Specifically, we created an MSA of *SRGAP2* coding exons 1 through 9 (shared between all human paralogs except *SRGAP2D* and encoding 452 amino acids) by extracting the exonic regions from the three contigs and using BLAT with exons 1–9 of human *SRGAP2A* as a reference to infer the exons in orthologous sequence from other species [chimpanzee (CGSC 2.1.3/panTro3), orangutan (WUGSC 2.0.2/ponAbe2), macaque (MGSC Merged 1.0/rheMac2), marmoset (WUGSC 3.2/calJac3), and dog (Broad/canFam2)]. We also obtained the mRNA sequence of the *SRGAP2* mouse ortholog (GenBank accession BC158055) and used this sequence to infer the rat coding sequence using BLAT against the rat genomic build (Baylor 3.4/rn4). A multiple-sequence alignment was created using ClustalW and Jalview (as before).

This alignment was used in conjunction with the codeml program (part of PAML 4; Yang, 2007) to test various models of selection. We estimated the overall dN/dS for the complete tree and compared likelihoods for models that allowed:

(1) free dN/dS for each branch (i.e., lineage heterogeneity);
(2) a primate-specific dN/dS;
(3) a human-specific dN/dS; and
(4) a duplicate-specific dN/dS.

Additionally, we performed tests aimed to detect site-specific signatures of positive selection across phylogeny (branch models):

(1) model 1a (neutral) versus model 2 (positive selection);
(2) model 7 (neutral) versus model 8 (with dN/dS >1); and
(3) model 8a (with dN/dS = 1) versus model 8 (with dN/dS >1).

Segmental Duplication Analysis of *SRGAP2*

To gain insight into the segmental duplication landscape around the *SRGAP2* paralogs, we calculated the percentage of bases classified as duplicated in the cytological bands 1q32.1, 1q21.1, 1p12, and 1p11.2 using the whole-genome assembly comparison (WGAC) on GRCh37/hg19 (Bailey et al., 2001). Note that *SRGAP2C* maps at the border of 1p12 and 1p11.2 in the reference genome (GRCh37/hg19). The results of this analysis show that chromosomal regions 1q21.1 and 1p11.2 are highly duplicated (63.9% and 69.4% of bases in these regions are classified as duplicated, respectively), whereas 1q32.1 and 1p12 (2.7% and 4.7% duplicated, respectively) are not.

We also sought to assess the duplication status of *SRGAP2* in mammals including chimpanzee, gorilla, orangutan, macaque, marmoset, gibbon, cow, dog, elephant, mouse, and rat. For chimpanzee, gorilla, and orangutan, we generated copy number heatmaps for several individuals using our recently described approach (Sudmant et al., 2010) and found no evidence of duplication or copy number variation at *SRGAP2*. Furthermore, characterizing segmental duplications by performing WSSD (Bailey et al., 2002) for human, chimpanzee, gorilla, and orangutan validated this result. For macaque, marmoset, gibbon, cow, dog, elephant, and mouse, we performed WSSD against the corresponding 1q32.1 *SRGAP2* region in each species and again showed no evidence of segmental duplication at this locus. Finally, for rat we examined the segmental duplication WGAC track at the *SRGAP2* locus—again, this region was not duplicated. These results show that *SRGAP2* is duplicated specifically in the human lineage.

Characterization of *SRGAP2* Copy Number Variation Using Sequencing Data

We analyzed high-throughput Illumina shotgun sequence data from 661 individuals from 14 diverse human populations sequenced in Phase 2 of the 1000 Genomes Project in addition to nonhuman *Homo* species, Neanderthal, and Denisova. We applied our copy number genotyping method (Sudmant et al., 2010) to determine aggregate copy number for 1000 bp windows of unmasked sequence spanning the *SRGAP2A*, *SRGAP2B*, and *SRGAP2C* loci (Figures S4A–4C). In marked contrast to other nonhuman primates, duplications encompassing the promoter and first nine exons of *SRGAP2A* (the ancestral *SRGAP2* locus) were present in all *Homo* species analyzed. Additionally, we noticed that two regions (~95 kbp and ~25 kbp) of *SRGAP2* duplicated sequence consistently showed a predicted copy number of eight total diploid copies in most, but not all individuals (see Figures S1A and S4A–4C). Using the aggregate copy number heatmap data, we identified individuals having predicted 0–4 copies of these two regions corresponding to *SRGAP2D*. From this analysis, we observe that *SRGAP2D* includes a large internal deletion including exons 2 and 3 (~115 kbp), which affects all copies examined to date, and is copy number polymorphic in the general population.

The homologous region shared among our three *SRGAP2* contigs was analyzed for variants that could distinguish different *SRGAP2* paralogs. Such singly unique nucleotide (SUN) identifiers are defined as single base-pair variants that are fixed in the population for a particular paralog (Sudmant et al., 2010). Identifying and characterizing these variants is a critical first step in developing our genotyping assays based on SUN k-mers (SUNKs) and paralog-specific quantitative PCR (qPCR; details of the primer design and protocol are below). SUNKs are defined as 30-mers that specifically tag a region of the genome and thus can be used in conjunction with short-read sequencing data to genotype highly identical paralogs (Sudmant et al., 2010). Briefly, the *SRGAP2* SUNK map was generated by dividing each of the newly constructed *SRGAP2* contigs into its constituent overlapping 30-mers and mapping these k-mers back to the human reference genome (NCBI36/hg18) and to each of the new contigs using the mrsFAST mapper (Hach et al., 2010). Sequence reads represented in both the contigs and the human reference (allelic regions) were masked in the reference so as to eliminate mapping to duplicate regions. Contig-specific SUNKs were then defined as 30-mers that only mapped to one contig, specifically tagging a particular paralogous locus.

Read-depth-based copy number estimates were then generated considering only these SUNKs, ensuring copy number estimates would be paralog-specific. Across all 661 individuals examined, *SRGAP2A* and *SRGAP2C* were fixed at two copies with the exception of 11 individuals who exhibited possible *SRGAP2A* duplications. Further analysis and qPCR-based copy number genotyping of the unique portion of the *SRGAP2A*, however, indicated all of these individuals have two copies of this paralog. To be certain, we also tested HapMap individuals at the lower and upper tails of the *SRGAP2C* copy number distribution using qPCR and validated that all of these individuals have two copies of this paralog. Alternatively, *SRGAP2B* appeared to be copy number polymorphic among individuals. We observed 3 homozygous and 33 heterozygous *SRGAP2B* deletions, as well as 49 individuals with three copies and 1 with four copies. A subset of individuals with *SRGAP2B* deletions and duplications were validated using FISH (described above, using probe 1, see Table S1) and a custom Agilent array targeting sequence from the *SRGAP2* contigs. From this analysis, we show

the deletions and duplications affect exons 2 and 3 of *SRGAP2B*. No copies of *SRGAP2B* and *SRGAP2C* were observed among any of the nonhuman primates analyzed (including 34 nonhuman primates consisting of 4 bonobos, 7 chimpanzees, 11 gorillas, and 12 Bornean and Sumatran orangutans [T. Marques-Bonet, personal communication]).

SRGAP2B Hardy-Weinberg Equilibrium Analysis

SRGAP2B showed copy number polymorphism in all human populations, leading us to believe it is likely a nonfunctional pseudogene. To more formally explore this possibility, we calculated whether copy number allele frequencies for this paralog are in Hardy-Weinberg proportions. If selection were operating on copy number at this locus, we would expect to find the *SRGAP2B* copy number allele frequencies inconsistent with Hardy-Weinberg equilibrium. Thus, if we find the *SRGAP2B* copy number allele frequencies at Hardy-Weinberg proportions, we can rule out selection on *SRGAP2B* copy number. The absence of selection on copy number would be consistent with the lack of an important functional role for this paralog. To determine whether the *SRGAP2B* copy number allele frequencies are at Hardy-Weinberg proportions in humans, we performed the following analysis (shown here for the Yoruban in Ibadan, Nigeria, but applied separately to all populations):

p: deletion allele frequency
q: normal allele frequency
r: duplication allele frequency

The observed counts of individuals having each *SRGAP2B* diploid copy number (CN) state are:

CN 0 (genotype PP): 0 individuals
CN 1 (genotype PQ): 6 individuals
CN 2 (genotype QQ + genotype PR): 49 individuals
CN 3 (genotype QR): 11 individuals
CN 4 (genotype RR): 0 individuals

We also assume all individuals with predicted copy number 2 have two normal alleles—this assumption will lead us to underestimate (though likely not by much) allele frequencies p and r:

estimate for p = (2*0 homozygous dels+6 heterozygous dels)/(66*2 chromosomes) = 0.045
estimate for q = (6 het dels+2*49 homozygous normal+11 heterozygous dups)/(66*2 chromosomes) = 0.871
estimate for r = (11 heterozygous dups+2*0 homozygous dup)/(66*2 chromosomes) = 0.083
$p^2+q^2+r^2+2pq+2pr+2qr = 1$, as required

Expected counts:

CN 0: $p^2*66 = 0.136$ individuals
CN 1: $2pq*66 = 5.227$ individuals
CN 2: $(q^2+2pr)*66 = 50.595$ individuals
CN 3: $2qr*66 = 9.583$ individuals
CN 4: $r^2*66 = 0.458$ individuals

Use these counts to calculate a chi-square statistic: $\text{sum}((\text{observed-expected})^2/\text{expected}) = 0.969$.

The resulting p-value for the chi-square value above with two degrees of freedom is 0.6161, meaning the Hardy-Weinberg equilibrium is not rejected when considering the Yoruban population. Performing the same calculations on other populations, we obtain strong evidence that *SRGAP2B* is at Hardy-Weinberg proportions in 13 of 14 populations considered. The low p-value in the remaining population (Columbian in Medellin, Colombia) likely reflects genotyping error for a single individual rather than a meaningful biological departure from Hardy-Weinberg equilibrium. If we assume this individual (copy number estimate = 3.65) has a *SRGAP2B* genotype of copy number 3 rather than copy number 4, the p-value for this last population becomes 0.98. Thus, these data are consistent with *SRGAP2B* paralog segregating at Hardy-Weinberg equilibrium in humans as expected for a nonfunctional pseudogene.

Paralog-Specific Genotyping Using qPCR

Using the 244.2 kbp alignment between our 1q32.1, 1q21.1, and 1p12 contigs, we designed paralog-specific primers for genotyping copy number. These primers had to pass several criteria: (1) they were deemed acceptable by the primer design software Primer3 (http://frodo.wi.mit.edu/primer3/); (2) they were found to theoretically amplify only one *SRGAP2* paralog; (3) all sequences in the NCBI HTGS database (http://www.ncbi.nlm.nih.gov/HTGS) containing the targeted region of their corresponding targeted paralog had the specificity-conferring variants (i.e., no evidence suggested these variants are not fixed in the population); (4) all sequences in the HTGS database containing the targeted region of non-targeted paralogs lacked the specificity-conferring variants; and (5) they could not yield more than one product as determined by the *In-Silico* PCR tool of the UCSC Genome Browser using the human reference (GRCh37/hg19). See Table S6 for primer sequences.

The qPCR experiments were performed using the Roche LightCycler 480 with a primer set targeting the albumin gene (*ALB*, known to be at diploid copy number 2) as a control. Each reaction contained 5.0 μl Roche SYBR Green Master I, 0.2 μl of each primer (10 μM), 4 μl genomic DNA (2.5 ng/μL), and 0.6 μl PCR quality water. Cycling conditions included a hot start at 95°C for 5 min, followed by 40 cycles of melting at 95°C for 15 s, annealing primers at 58°C for 20 s, elongating products at 72°C for 20 s, and concluding with a melting curve from 50°C to 90°C. All qPCR reactions were run in three technical replicates. Cycles-at-threshold (C_t) values were calculated using the second derivative maximum method (Zhao and Fernald, 2005). Raw cycles-to-threshold data were converted into copy number estimates using the delta-delta C_t method using an individual with known copy number 2 at *SRGAP2A*, *SRGAP2B*, *SRGAP2C*, and *SRGAP2D* as the control (NA12878).

SRGAP2A and *SRGAP2C* Copy Number Variant Detection in Cases and Controls

We utilized previously existing data sets as well as targeted qPCR and array CGH to assess CNVs of *SRGAP2A* (1q32.1) and *SRGAP2C* (1p12), respectively. For *SRGAP2A*, we made use of previously identified CNVs reported in the Cooper et al. (2011) study to identify deletions and duplications in a cohort of 15,767 children with developmental delay (including intellectual disability and autism spectrum disorder) screened using array CGH. Four of the six CNVs were validated using a custom Agilent microarray. The remaining two CNVs were very large (>20 Mbp) and likely real. From the same study, we assessed data from 8,329 controls screened using SNP arrays. Notably, due to inefficient probe coverage across the *SRGAP2C* segmental duplication and flanking region, we were not confident in our ascertainment of CNVs of the 1p12 region.

For both *SRGAP2A* and *SRGAP2C* loci, we utilized paralog-specific qPCR assays (see Table S6 for primer sequences) using experimental procedures described above. Specifically, 1,602 children with intellectual disability and 1,794 controls (comprised of NIMH [https://www.nimhgenetics.org/available_data/controls/] and ClinSeq [Biesecker et al., 2009] individuals) were screened with qPCR assays targeting intron 12 of *SRGAP2A* and intron 7 of *SRGAP2C*, respectively. As a control, we used a qPCR assay for *ALB* (fixed at copy number 2). No deletions or duplications were identified for *SRGAP2A* using this assay. To validate a small subset of individuals showing a deletion or duplication of *SRGAP2C*, we performed a second qPCR screen specific to *SRGAP2C* intron 6. From this, we validated only one case and one control with a duplication.

Specifically for *SRGAP2C*, we also used array CGH data from a custom Agilent microarray to assess whether the 1p11.2 region proximal to the segmental duplication (chr1-120,843,952-121,057,437, NCBI36/hg18; 36 probes) was deleted or duplicated in a cohort of children with sporadic (n = 2,294, Simons Simplex Collection [Fischbach and Lord, 2010]) and familial (n = 579, Autism Genetic Resource Exchange [Geschwind et al., 2001]) autism spectrum disorder. Additionally, we screened 580 controls on the same microarray. We detected no deletions and a small number of duplications in both cases and controls (~0.1%). Individuals with a detected duplication were assayed using the *SRGAP2C*-specific qPCR assay described above, with only two sporadic autism probands showing the duplication extending across *SRGAP2C*. All *SRGAP2C* duplications were validated using an Agilent custom array targeting our *SRGAP2* contig sequences (much of which is missing from the human reference genome) in an attempt to identify any breakpoints.

Database Search of *SRGAP2* Transcripts

The sequences of all transcripts mapping to *SRGAP2* paralogs (*SRGAP2A* at 1q32.1, *SRGAP2B* at 1q21.1, and *SRGAP2C* at 1p12) in the GRCh37/hg19 reference sequence available through the UCSC Genome Browser were downloaded and analyzed for potential alternative splice variants or novel *SRGAP2* transcripts that could be expressed from the duplicate paralogs (i.e., transcripts containing only exons 1 through 9). We identified full-length *SRGAP2A* transcripts containing 22 coding exons (GenBank accessions AB007925, BC132872, BC144343, BC132874, BC150646) as well as several additional transcripts mapping to the *SRGAP2A* locus (GenBank accessions AK057565, AK294060, AK311111, BC063527, DQ786311, AK000885, AK091814, AK293335, AK295845, BC041635, DQ786257). Some *SRGAP2A* transcripts showed alternative splice forms excluding the first three nucleotides in exon 7, resulting in an in-frame removal of an amino acid. We also discovered a single transcript, cloned from a breast cancer cell line, truncated at exon 9 and including 1,373 additional nucleotides from intron 9 and a polyA tail (GenBank accession BC112927). We determined that the sequence of the transcript matches our *SRGAP2C* 1p12 contig (and is not a *SRGAP2A* truncated transcript) and predicts an open-reading frame (ORF) that, if translated, encodes a truncated SRGAP2 protein (458 amino acids), including a partial F-BAR domain (Guerrier et al., 2009) with seven unique residues at the carboxy terminus.

Sequencing of *SRGAP2* Transcripts

We performed long-range PCR (Expand Long Template, Roche) on cDNA (generated using the Roche High Fidelity cDNA Synthesis Kit, oligo(dT) primers) from SH-SY5Y neuronal cell line total RNA, human fetal brain total RNA (collected from spontaneously aborted fetuses, 20-33 weeks, ClonTech S2437; approximately 50-60 fetuses pooled), single human adult brain mRNA (BioChain, M1234035), and single human fetal brain total RNA (BioChain, R1244035-50). We amplified a single PCR band at the expected size for *SRGAP2A*-specific primers (including exon 10). Alternatively, multiple PCR bands (2-3) are amplified using the duplicate-specific *SRGAP2* primers (including the intron 9 extension). We PCR purified and cloned these fragments into the pCR4.0 vector (Invitrogen) and digested with EcoRI (NEB) to validate PCR insert sizes in vectors. We sequenced clones using primers spanning across the *SRGAP2* region (ABI3630 Genetic Analyzer). Primer sequences are shown in Table S6. Additional primers were used to sequence clones containing the *SRGAP2* transcripts (available upon request).

Transcripts were assigned to their respective paralogs based on diagnostic sequence variants from genome sequencing. We did not detect any splice or sequence variants within exons 1 through 10 of the *SRGAP2A* ancestral paralog (n = 11 transcripts). Alternatively, we detected three splice and numerous sequence variants of the *SRGAP2* duplicate-derived transcripts (n = 85 transcripts) based on diagnostic sequence differences. We identified "full-length" duplicate transcripts containing exons 1 through 9 mapping to both *SRGAP2B* (n = 4) and *SRGAP2C* (n = 47). One rare splice isoform (n = 2), which removed exon 3 and mapped to *SRGAP2C*, encodes a truncated 98-residue protein, including 11 unique amino acids at the carboxy terminus. The other splice isoform (n = 31), which removed exons 2 and 3, mapped to *SRGAP2D*. Although we do not have finished sequence for *SRGAP2D*, the sequence and structure of this transcript is consistent with *SRGAP2D* transcription, as this paralog harbors a large genomic deletion containing exons 2 and 3. From this analysis, we were able to assign 16 polymorphic and fixed PSVs to the *SRGAP2* paralogs.

In order to determine the relative abundance of "full-length" transcripts (i.e., including both exons 2 and 3) expressed from *SRGAP2B* and *SRGAP2C*, we performed long-range PCR amplification using primers specific to exon 3 and the intron 9 (duplicate-specific) extension, respectively, from cDNA (prepared using oligo(dT) primers) generated from human adult brain, fetal brain, and lymphoblastoid cell lines. These same cell lines had been genotyped for the *SRGAP2B* polymorphism. We performed capillary sequencing using the same primers used to amplify the cDNA and compared the chromatograms of two coding PSVs (variants 12 and 13). We show that relative expression of the *SRGAP2B* transcript is lower than *SRGAP2C* expression by quantifying peak heights of chromatogram plots for each nucleotide base corresponding to a PSV. As a control, we detected no *SRGAP2B* transcript from a HapMap lymphoblastoid cell line genotyped as copy number 0 for *SRGAP2B*.

Nonsense-Mediated Decay of the *SRGAP2D* Isoform

The *SRGAP2D* paralog produces a transcript missing exons 2 and 3 resulting in a premature stop codon at the exon 1 and 4 junction. We predict that this transcript will undergo nonsense-mediated decay (NMD); to test this, we assessed the abundance of this transcript in HapMap lymphoblastoid cell lines in normal and NMD-blocked conditions [i.e., in the presence of emetine, which blocks translation and NMD (Noensie and Dietz, 2001)]. Specifically, we incubated HapMap EBV-transformed lymphoblastoid cells ($1x10^7$) in 10 ml of media (see ATCC guidelines) with or without 100 μg/ml emetine dihydrochloride hydrate (Sigma, E2375) for 7 hr at 37°C. We immediately isolated total RNA from each treated and untreated cell line using Trizol (Invitrogen) and the RNeasy Mini Kit (QIAGEN). cDNA was prepared from 3 μg of total RNA using the Transcription High Fidelity cDNA Synthesis Kit (Roche) and random hexamer primers. qPCR was performed following the same protocol described earlier with 10 ng of cDNA and primers specific to primers mapping to: (1) the *SRGAP2D* exon 1 and 4 junction and within exon 5; (2) *SRGAP2A*-specific exons 21 and 22; and (3) exons of the housekeeping gene *GAPDH*. See Table S6 for primer sequences.

First, to determine whether the cDNA isoform excluding exons 2 and 3 ("deletion isoform") is derived from the *SRGAP2D* or *SRGAP2B* locus (e.g., from a copy having the internal deletion polymorphism), we assessed overall levels of the deletion isoform in cell lines with *SRGAP2B* copy numbers of 0, 1, and 2, respectively. We found no difference in levels of transcript across cell lines indicating that the deletion isoform is likely derived from *SRGAP2D*. Second, by blocking NMD in these cells we discovered a significant 1.1- to 1.6-fold increase of the *SRGAP2D* aberrant transcript compared to the full-length *SRGAP2A* transcript abundance (0.9- to 1.2-fold change). Overall, this indicates that NMD may be acting on the *SRGAP2D* transcript, though to a modest degree. Notably, performing this assay using lymphoblastoid cells rather than neuronal cells may limit the biological relevance of this result.

SRGAP2 Tissue-Specific Expression Analysis Using Paralog-Specific qPCR

From the transcript analysis of the *SRGAP2* paralogs, we designed paralog-specific RT-PCR primers to assess the expression of specific *SRGAP2* paralogs. In designing *SRGAP2C*-specific primers, we leveraged PSV-12 within exon 7. Similar attempts to design *SRGAP2B*-specific primers for PSVs-5, −13, and −16 resulted in non-specific amplification of both alleles. We found that the ancestral *SRGAP2A*, in addition to the *SRGAP2B* and *SRGAP2C* duplicates, were expressed in a variety of human brain tissues [using pooled total RNA (ClonTech)] including total fetal brain (ID: 636526), adult brain (ID: 636530), cerebellum (ID: 636535), frontal lobe (ID: 636563), and temporal lobe (ID: 636564). As expected, we did not detect expression of the duplicate copies in any of the nonhuman primate-derived tissues.

Spatiotemporal *SRGAP2* Expression Using RNA-Seq Data

A subset of the SUNKs we identified (described above) were embedded in coding and UTR sequence of the *SRGAP2* paralogous transcripts, providing a unique opportunity to assess paralog-specific expression patterns using RNA-Seq data. Sequence from four different studies was analyzed encompassing 17 different human tissues (Illumina's Human BodyMap 2.0), 7 human cell lines (Wang et al., 2008), and both chimpanzee and macaque cerebellum and liver tissues (Blekhman et al., 2010). Briefly, RNA-Seq data sets were mapped to the human reference genome (NCBI36/hg18) in addition to the newly sequenced *SRGAP2* contigs and expression levels over genes were calculated in units of RPKM (Liu et al., 2011). Among human tissues, we found that the *SRGAP2* paralogs were most highly expressed in whole-brain, cerebellum, and breast tissues, with the whole-brain and cerebellum samples showing the tightest expression levels with least variability between biological replicates. *SRGAP2A* (the ancestral paralog) shows similar expression levels in the cerebellum of chimpanzees and macaques compared to humans with little to no expression in the liver of humans, chimpanzees, or macaques. As expected, no signature of *SRGAP2* duplicate transcripts was detected in

chimpanzees or macaques. Within the cerebellum, we observed *SRGAP2A* and *SRGAP2B/D* transcripts to be the most abundant and *SRGAP2C* transcripts to be the least abundant, though this analysis does not account for potential alternative splice isoforms.

SUPPLEMENTAL REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Bailey, J.A., Yavor, A.M., Massa, H.F., Trask, B.J., and Eichler, E.E. (2001). Segmental duplications: organization and impact within the current human genome project assembly. Genome Res. *11*, 1005–1017.

Hach, F., Hormozdiari, F., Alkan, C., Hormozdiari, F., Birol, I., Eichler, E.E., and Sahinalp, S.C. (2010). mrsFAST: a cache-oblivious algorithm for short-read mapping. Nat. Methods *7*, 576–577.

Kidd, J.M., Graves, T., Newman, T.L., Fulton, R., Hayden, H.S., Malig, M., Kallicki, J., Kaul, R., Wilson, R.K., and Eichler, E.E. (2010). A human genome structural variation sequencing resource reveals insights into mutational mechanisms. Cell *143*, 837–847.

Noensie, E.N., and Dietz, H.C. (2001). A strategy for disease gene identification through nonsense-mediated mRNA decay inhibition. Nat. Biotechnol. *19*, 434–439.

Schmutz, J., Wheeler, J., Grimwood, J., Dickson, M., Yang, J., Caoile, C., Bajorek, E., Black, S., Chan, Y.M., Denys, M., et al. (2004). Quality assessment of the human genome sequence. Nature *429*, 365–368.

Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. In Current Protocols in Bioinformatics. (New York: John Wiley & Sons), 4.10.1–4.10.14.

Tuzun, E., Sharp, A.J., Bailey, J.A., Kaul, R., Morrison, V.A., Pertz, L.M., Haugen, E., Hayden, H., Albertson, D., Pinkel, D., et al. (2005). Fine-scale structural variation of the human genome. Nat. Genet. *37*, 727–732.

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.

Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M., and Barton, G.J. (2009). Jalview Version 2 - a multiple sequence alignment editor and analysis workbench. Bioinformatics *25*, 1189–1191.

Zhao, S., and Fernald, R.D. (2005). Comprehensive algorithm for quantitative real-time polymerase chain reaction. J. Comput. Biol. *12*, 1047–1064.

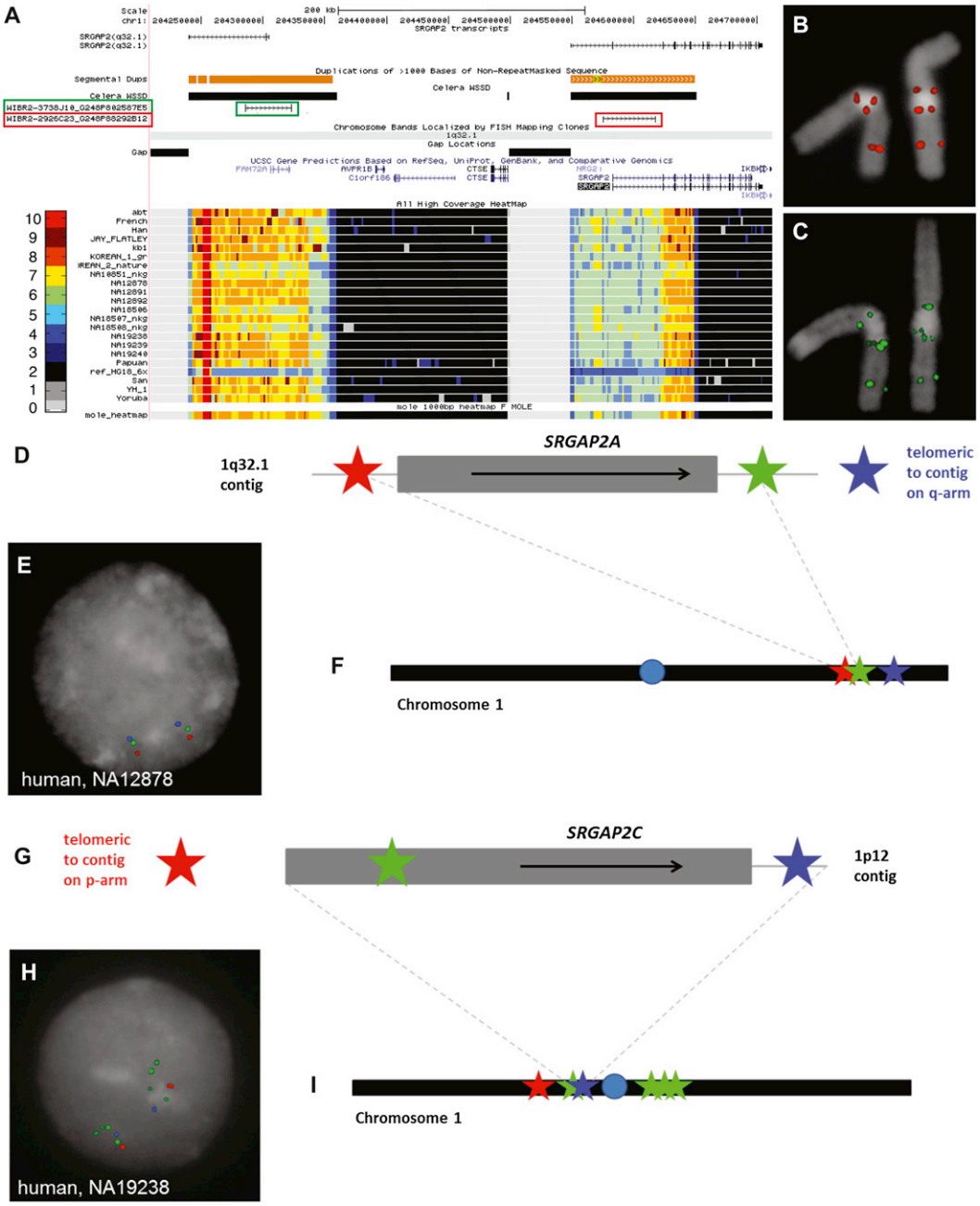

Figure S1. FISH Experiments Reveal a Fourth *SRGAP2* Paralog and Specify *SRGAP2* Paralog Orientations, Related to Figure 1

(A–C) The UCSC genome browser snapshot of the *SRGAP2A* locus in NCBI36/hg18 shows the *SRGAP2A* transcript, segmental duplications, fosmid clones used for FISH experiments, gaps in the reference genome, gene predictions, and copy number heatmaps for 22 human genomes sequenced at high coverage (including the complete hydatidiform mole genome used in this study) as well as for an "Illuminized" NCBI36/hg18 reference sequence. Colors indicate copy number predictions based on short-read sequence read-depth (Sudmant et al., 2010). *SRGAP2A* is incomplete and partially inverted in the reference genome NCBI36/hg18. FISH analysis using a probe spanning from intron 2 to intron 3 (B) detects three *SRGAP2* paralogs on metaphase human chromosome 1. However, FISH using a probe spanning from upstream of *SRGAP2* to intron 1 (C) detects four *SRGAP2* paralogs on metaphase human chromosome 1. These results are consistent with the heatmap data in (A) and suggest the existence of a fourth human *SRGAP2* paralog having an internal deletion of at least exon 3.

(D–I) Depictions of probe locations, FISH images of interphase human chromosome 1, and schematics showing the results of the experiments to resolve the orientation of the *SRGAP2A* (D–F) and *SRGAP2C* (G–I) paralogs. Colored stars indicate relative locations of the corresponding FISH probes with regard to our *SRGAP2* contigs (gray boxes). Thin gray lines indicate extensions of our contigs based on contiguous reference sequence. Extensive local duplication upstream of the *SRGAP2B* and *SRGAP2D* paralogs (see yellow probe in Figure 2C) prevented accurate determination of their orientation using FISH. Using an anchored contig spanning the entire human 1q21.1 region recently generated at The Genome Institute at Washington University School of Medicine (unpublished data), we determined that *SRGAP2B* is oriented such that gene transcription would proceed toward the centromere.

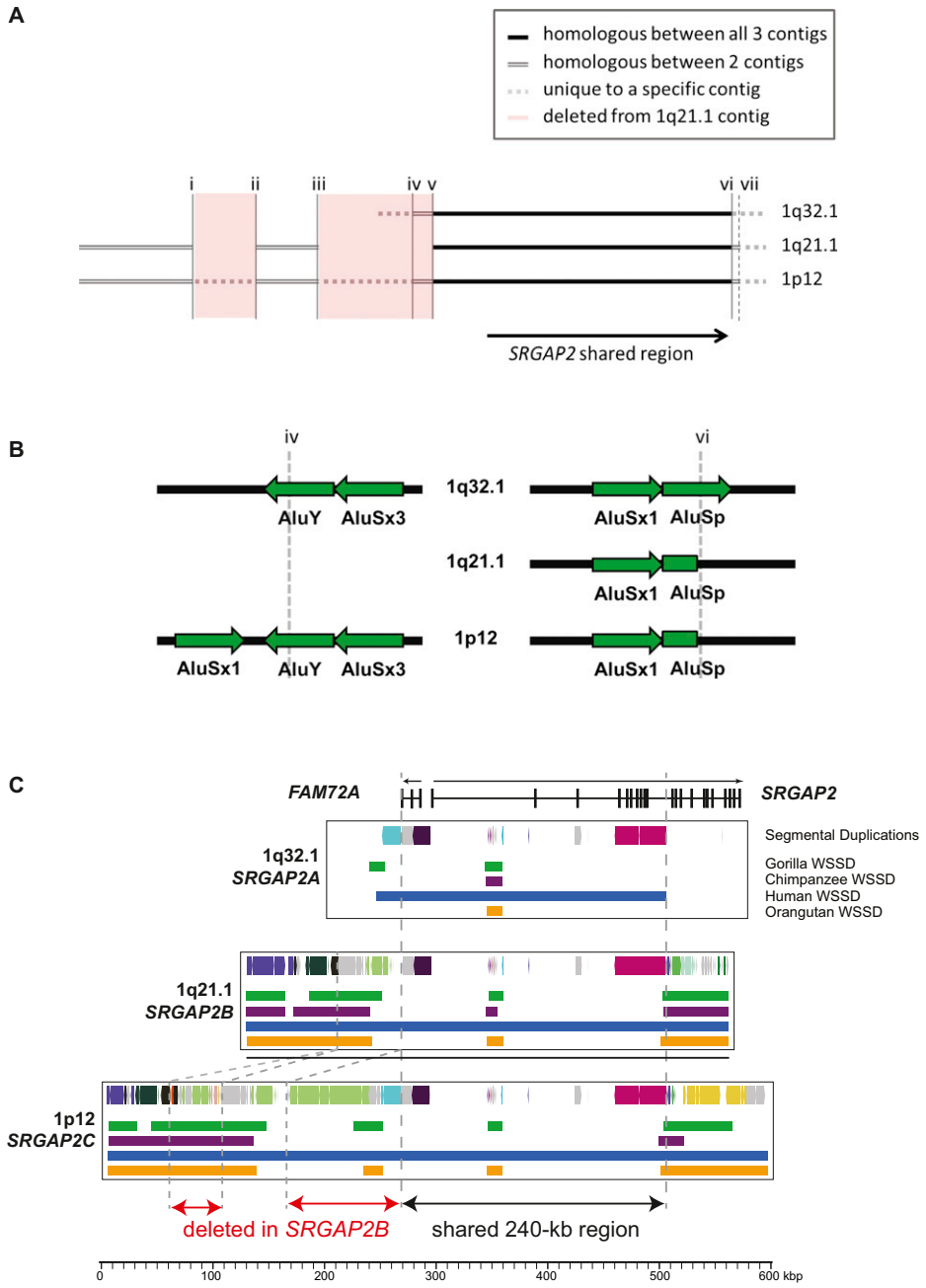

Figure S2. Breakpoint and Duplication Analyses of *SRGAP2* Contigs, Related to Figure 2

(A) Homologous regions shared between the 5′ ends of the *SRGAP2* contigs are indicated, as well as their relative position with respect to the first nine exons of *SRGAP2* shared between these three paralogs. The region shared by these three paralogs spans 244.2 kbp, with the total shared sequence length depicted here over 515 kbp. Vertical lines indicate breakpoints corresponding to duplication breakpoints (iv, vi, vii) or breakpoints from other structural rearrangements (i, ii, iii, v). All breakpoints were extremely well defined except for vii; this lower resolution (within a few hundred base pairs) is indicated by a dashed vertical line. The 5′ duplication breakpoint between the 1q21.1 and 1p12 paralogous regions lies beyond the edge of the contigs. Missing sequence in the 1q21.1 contig between iii and v resulted from a deletion in 1q21.1 rather than an insertion in 1p12 because part of this missing sequence (between iv and v) extends the region of homology with paralogous sequence at 1q32.1.

(B) Zoomed-in views of the initial duplication breakpoints are presented with repetitive elements highlighted. These elements were identified by using RepeatMasker (Tarailo-Graovac and Chen, 2009) on sequences surrounding the breakpoints—the best repeat subfamily matches are indicated. Breakpoints i, ii, and v also contained Alu elements at their boundaries. The remaining breakpoints lack repetitive features in their immediate surrounding sequences.

(C) A duplication analysis of the *SRGAP2* contigs using SegDupMasker and whole-genome shotgun sequence detection (WSSD) (Bailey et al., 2002) highlights the highly duplicated chromosomal environments flanking *SRGAP2* paralogs at 1q21.1 and 1p12 and affirms that the *SRGAP2* duplication is specific to the human lineage.

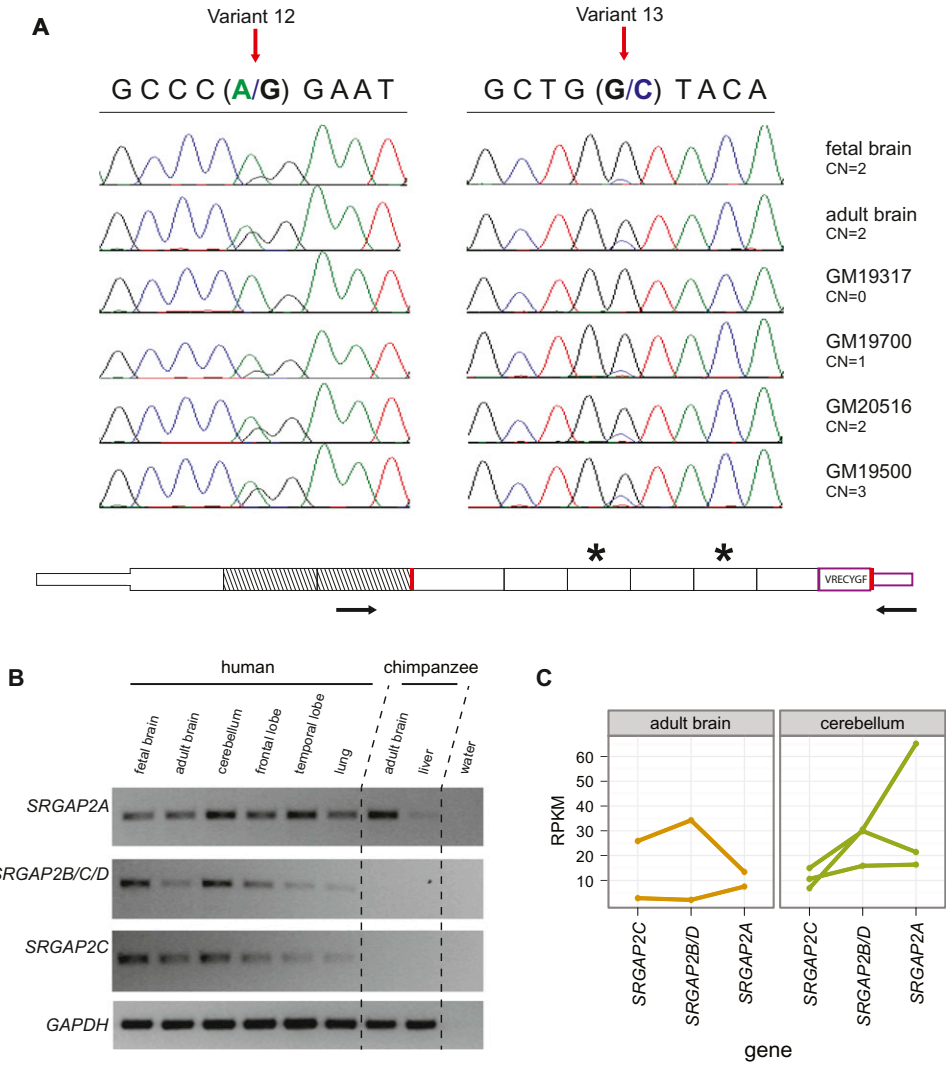

Figure S3. *SRGAP2* Gene Expression Analysis, Related to Figure 3

(A) Sequencing of *SRGAP2* "full-length" transcript from *SRGAP2B* and *SRGAP2C* revealed reduced expression of *SRGAP2B*. To perform this experiment, we used primers targeting transcripts containing exons 3 and the 3′ UTR (intron 9 extension), respectively, to avoid quantifying expression of the *SRGAP2D* transcript (with deleted exons 2 and 3). Pictured are chromatograms for coding paralog-specific variants (PSVs) 12 (*SRGAP2C*-allele = A and *SRGAP2B*-allele = G) and 13 (*SRGAP2C*-allele = G and *SRGAP2B*-allele = C) (see Figure 3B in the main text) from transcripts derived from human adult brain, fetal brain, and lymphoblastoid cells. Relative transcript abundances were determined by comparing the heights of the PSV peaks of the chromatograms.

(B) RT- PCR was performed using primers specific to the ancestral and duplicate *SRGAP2* paralogs and a housekeeping gene, *GAPDH*, using cDNA derived from human and chimpanzee tissues. The following paralogs were amplified based on the existence of specific exons or by utilizing PSVs including: *SRGAP2A* (exon 8 and exon 10); *SRGAP2B/C/D* (exon 8 and 3′ UTR extending into intron 9); and *SRGAP2C* (exon 6 containing PSV-12 and exon 7). For primer sequences used, refer the Extended Experimental Procedures.

(C) Individual RPKM estimates (Liu et al., 2011) allow quantification of expression *SRGAP2* paralogs from human adult brain and cerebellum. Tissue from two and three individuals was used to test expression in adult brain and cerebellum, respectively.

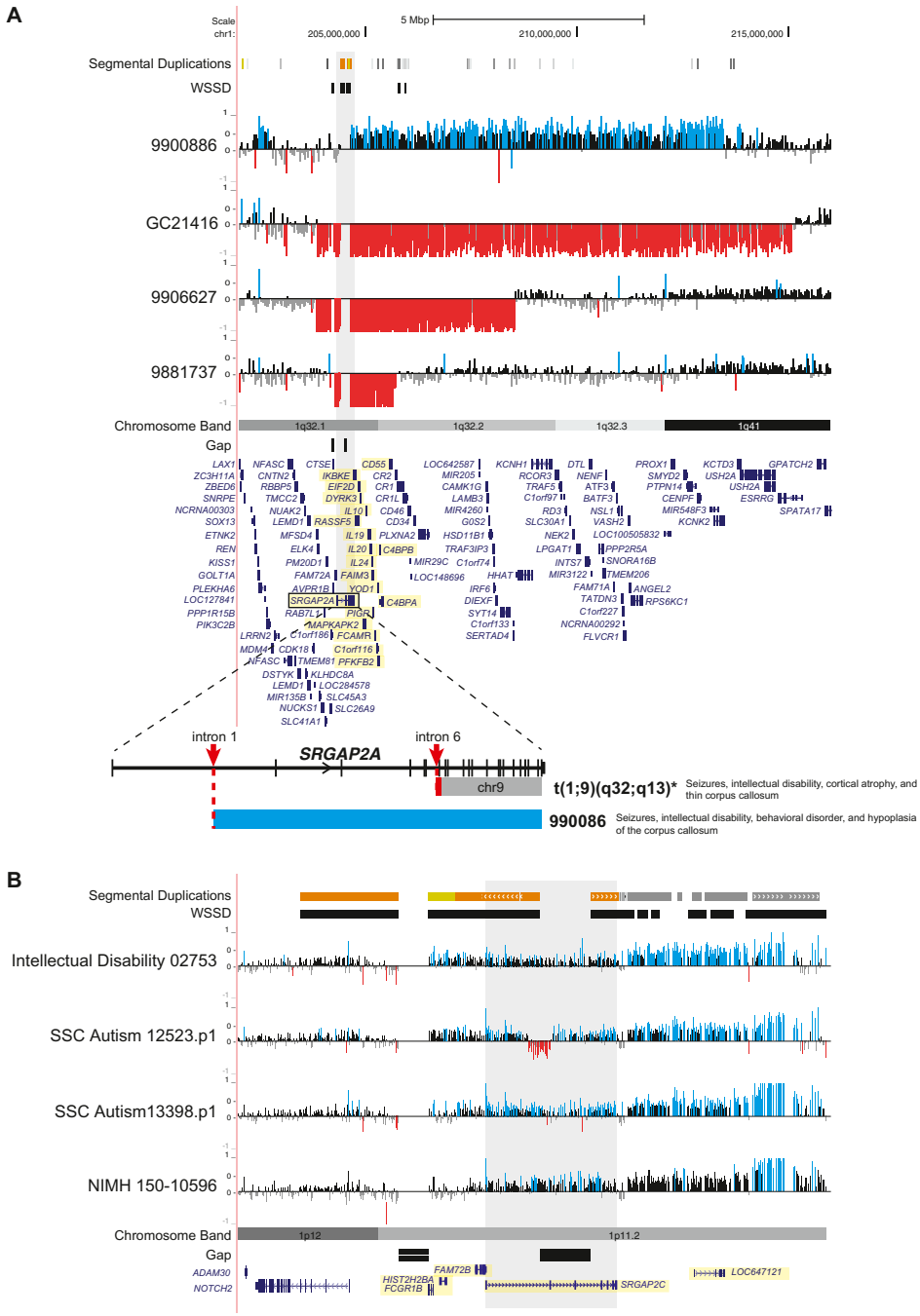

Figure S4. Large CNVs of *SRGAP2A* and *SRGAP2C* Detected in Children with Developmental Delay and Autism, Related to Table 2

Large (>1 Mbp) deletions (red) and duplications (blue) of *SRGAP2A* (A) and *SRGAP2C* (B) were confirmed by array CGH for seven children with developmental delay and autism spectrum disorder, as well as an adult control individual. Two duplications of *SRGAP2A* (>20 Mbp) are not shown. Blue (duplication) and red (deletion) histograms depict $\log_2$ relative hybridization signals. The genes within the smallest region of overlap are highlighted in yellow.

(A) Below the array CGH data for *SRGAP2A* is an expanded view of proximal breakpoints (red arrows) mapping within *SRGAP2A* for two patients. The first is a de novo t(1;9)(q32;q13) translocation breakpoint (*described by Saitsu et al. [2011]) that maps within intron 6 and resulted in the deletion of exon 7 (red) and the remaining chromosome 1q-arm translocated to chromosome 9 (gray). The second is an 8.7 Mbp duplication (blue) breakpoint identified in this study that maps to *SRGAP2A* intron 1, assayed using a custom microarray targeting our *SRGAP2* contig sequences. Both patients show remarkable similarity in phenotype, including abnormalities of the corpus callosum, seizure, and intellectual disability.

(B) The depicted genomic region is a hybrid of the human reference (GRCh37/hg19) and missing sequence data generated from our *SRGAP2C* contigs. Note the genome assembly gap within the human reference extends across exons 2 and 3 of *SRGAP2C*. The deletion spanning exon 2 in the 12523.p1 autistic proband likely represents polymorphism of *SRGAP2B*.

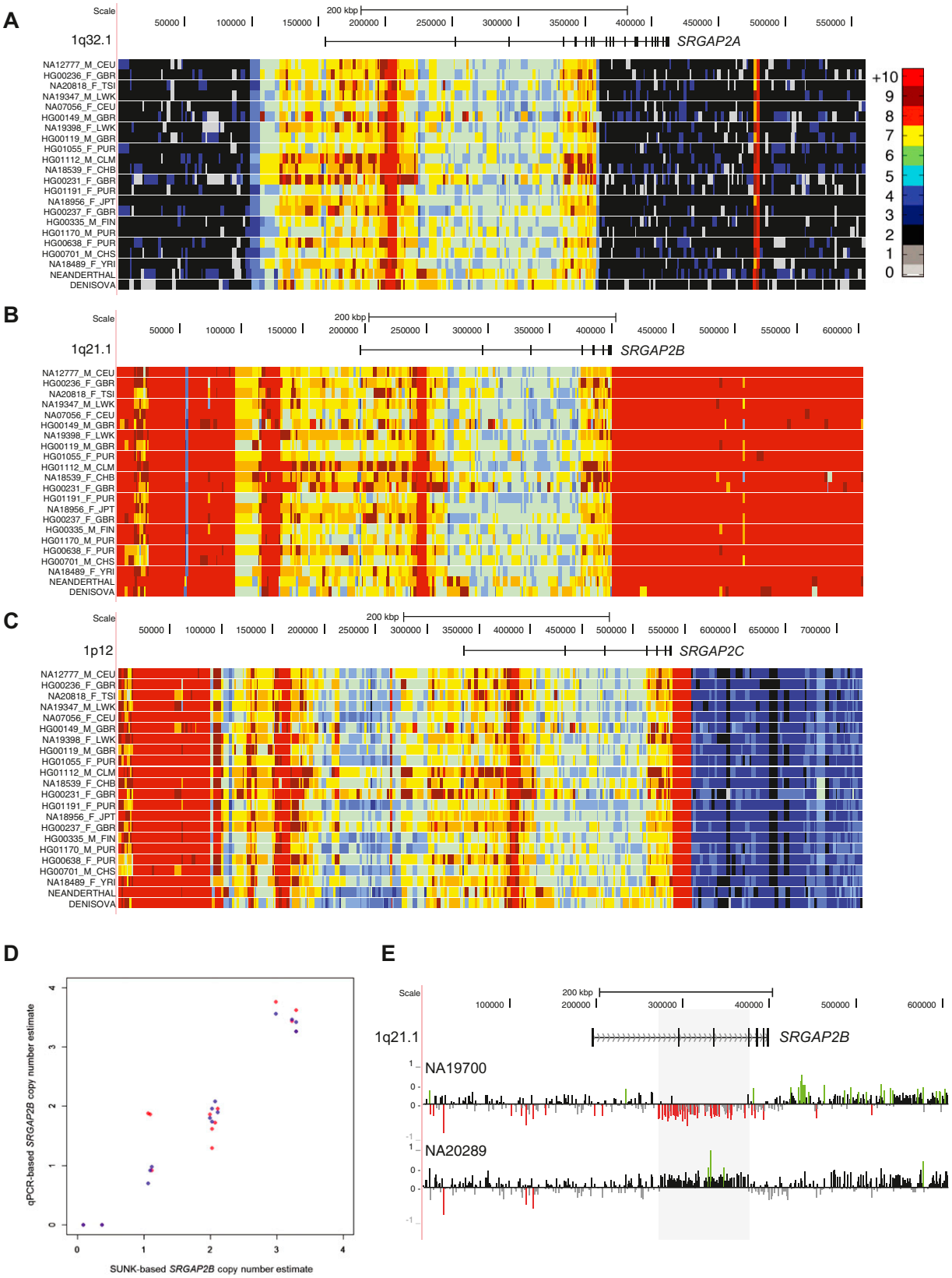
A
Scale
200 kbp
50000 100000 150000 200000 250000 300000 350000 400000 450000 500000 550000
1q32.1
SRGAP2A
NA12777_M_CEU
HG00236_F_GBR
NA20818_F_TSI
NA19347_M_LWK
NA07056_F_CEU
HG00149_M_GBR
NA19398_F_LWK
HG00119_M_GBR
HG01055_F_PUR
HG01112_M_CLM
NA18539_F_CHB
HG00231_F_GBR
HG01191_F_PUR
NA18956_F_JPT
HG00237_F_GBR
HG00335_M_FIN
HG01170_M_PUR
HG00638_F_PUR
HG00701_M_CHS
NA18489_F_YRI
NEANDERTHAL
DENISOVA
+10
9
8
7
6
5
4
3
2
1
0
B
Scale
200 kbp
50000 100000 150000 200000 250000 300000 350000 400000 450000 500000 550000 600000
1q21.1
SRGAP2B
NA12777_M_CEU
HG00236_F_GBR
NA20818_F_TSI
NA19347_M_LWK
NA07056_F_CEU
HG00149_M_GBR
NA19398_F_LWK
HG00119_M_GBR
HG01055_F_PUR
HG01112_M_CLM
NA18539_F_CHB
HG00231_F_GBR
HG01191_F_PUR
NA18956_F_JPT
HG00237_F_GBR
HG00335_M_FIN
HG01170_M_PUR
HG00638_F_PUR
HG00701_M_CHS
NA18489_F_YRI
NEANDERTHAL
DENISOVA
C
Scale
200 kbp
50000 100000 150000 200000 250000 300000 350000 400000 450000 500000 550000 600000 650000 700000
1p12
SRGAP2C
NA12777_M_CEU
HG00236_F_GBR
NA20818_F_TSI
NA19347_M_LWK
NA07056_F_CEU
HG00149_M_GBR
NA19398_F_LWK
HG00119_M_GBR
HG01055_F_PUR
HG01112_M_CLM
NA18539_F_CHB
HG00231_F_GBR
HG01191_F_PUR
NA18956_F_JPT
HG00237_F_GBR
HG00335_M_FIN
HG01170_M_PUR
HG00638_F_PUR
HG00701_M_CHS
NA18489_F_YRI
NEANDERTHAL
DENISOVA
D
qPCR-based SRGAP2B copy number estimate
SUNK-based SRGAP2B copy number estimate
0 1 2 3 4
E
Scale
200 kbp
100000 200000 300000 400000 500000 600000
1q21.1
SRGAP2B
NA19700
NA20289
1
0
-1

Figure S5. Copy Number Analysis of Next-Generation Sequencing Data, Related to Figure 4

Shown are heatmaps of aggregate copy number across the *SRGAP2* contigs using short-read sequences from human, Neanderthal, and Denisova genomes. Depicted are a representative sample of diverse human individuals from a total panel of 661 individuals from 14 populations (1000 Genomes Project), Neanderthal, and Denisova for the (A) 1q32.1 (*SRGAP2A*), (B) 1q21.1 (*SRGAP2B*), and (C) 1p12 (*SRGAP2C*) contigs. Gene models are shown at the top. Colors represent varying aggregate copy number predictions based on read-depth analysis of short-read sequencing data. In contrast to nonhuman primates, the first nine exons of *SRGAP2* in all the human samples analyzed are duplicated. Specifically, the genomic regions containing exons 1 as well as exons 4 through 9 are predicted as diploid copy number 7-8 while the region containing exon 2 and 3 are predicted as diploid copy number 5-6. From this analysis, we validated the existence of a fourth paralog lacking exons 2 and 3 (*SRGAP2D*). Additionally, the genomic regions flanking both sides of *SRGAP2B* at 1q21.1 show high copy number (>10 diploid copies) adding to the evidence that this region likely represented a non-ideal gene environment at the time of the initial duplication.

(D) Comparison of SUNK-based and qPCR-based copy number estimates for *SRGAP2B* in multiple human individuals shows a clear correlation. Each point corresponds to an ordered pair of *SRGAP2B* copy number estimates, with the abscissa being the SUNK-based estimate and the ordinate being the qPCR estimate. Red points are from one qPCR experiment, and blue points are from a replicate qPCR experiment. The qPCR results recapitulate the four clusters seen in our SUNK analysis, clusters corresponding to different copy number states for *SRGAP2B* paralog. The overall fit of a linear model to these points has an R^2 = 0.9087, indicating strong concordance between the two orthogonal copy number estimation methods. These data confirm that *SRGAP2B* paralog is indeed polymorphic in humans. Note, qPCR of *SRGAP2D* also confirmed that this paralog is polymorphic in humans (not shown).

(E) Array CGH of two HapMap individuals (NA19700 and NA20289) with a predicted deletion and duplication, respectively, of *SRGAP2B* was performed to validate polymorphism of this paralog. Blue (duplication) and red (deletion) histograms depict $\log_2$ relative hybridization signals mapped to our 1q21.1-sequenced contig (*SRGAP2B*). Both the deletion and duplication span the genomic region containing exons 2 and 3 of the paralog. The corresponding FISH experiment for NA19700 is depicted in Figure 4C.