Supplementary Information for:

Emergence of a *Homo sapiens*-specific gene family and chromosome 16p11.2 CNV susceptibility

Xander Nuttle[1*], Giuliana Giannuzzi[2*], Michael H. Duyzend[1], Joshua G. Schraiber[1], Iñigo Narvaiza[3], Francesca Camponeschi[4], Simone Ciofi-Baffoni[4,5], Peter H. Sudmant[1,6], Osnat Penn[1], Giorgia Chiatante[7], Maika Malig[1], John Huddleston[1,8], Chris Benner[3], Holly A.F. Stessman[1], Maria C. N. Marchetto[3], Laura Denman[1], Lana Harshman[1], Carl Baker[1], Archana Raja[1,8], Kelsi Penewit[1], Nicolette Janke[1], W. Joyce Tang[9], Mario Ventura[7], Francesca Antonacci[7], Joshua M. Akey[1], Chris T. Amemiya[9], Lucia Banci[4,5], Fred H. Gage[3,10], Alexandre Reymond[2], and Evan E. Eichler[1,8]

Table of Contents

1. Sequencing and assembly of the chromosome 16p11.2 region

We generated high-quality reference sequences over the chromosome 16p11.2 region for orangutan, chimpanzee, and multiple human haplotypes by sequencing and assembling large-insert clones using a previously described strategy[6]. We examined clone paired-end sequence mapping data[40] and/or performed hybridization experiments to identify bacterial artificial chromosomes (BACs) likely harboring sequence from the chromosome 16p11.2 region. We utilized three BAC libraries constructed from human genomic material: CH17 (from the complete hydatidiform mole CHM1[41]), VMRC53 (from the HapMap female NA12878), and RP11 (from a male, the primary library sequenced as part of the Human Genome Project). We also used BAC libraries from chimpanzee (CH251, from a male named Clint) and orangutan (CH276, from a female named Susie). All candidate BACs were sequenced using a Nextera protocol[42] and massively parallel Illumina sequencing technology, and reads were mapped and analyzed as previously described[6,38]. This procedure allowed us to select tiling paths of clones spanning the region. This process was complicated by the presence of segmental duplications having high sequence identity within the chromosome 16p11.2 region and between chromosome 16p11.2 and other chromosome 16 loci. However, because the Nextera data provide sequence information across the entirety of the clone insert (~170 kbp in length) rather than merely at the ends, it was possible to distinguish truly overlapping clones from their allelic and paralogous counterparts and, thus, establish single-haplotype tiling paths.

BAC clones were sequenced using capillary sequencing or Pacific Biosciences (PacBio) single-molecule, real-time (SMRT) technology[31]. SMRT-sequenced clones were assembled using HGAP and error-corrected using Quiver to generate one complete single contig per clone as previously described[31]. Sequences from overlapping clones were assembled into larger haplotype contigs using Sequencher (Gene Codes Corporation). To ensure proper assembly, we assembled clones into the same contig only if they exhibited >99.9% sequence identity over their shared region of overlap. Due to the complexity of the 16p11.2 locus in chimpanzee and our discovery of a large inversion polymorphism, we performed additional rounds of genomic library colony hybridization, Nextera clone sequencing[6,38,42], and SMRT BAC sequencing and assembly[6,31] to fill gaps remaining after the initial set of chimpanzee clones was sequenced and assembled into contigs. Ten chimpanzee clones could not be fully resolved owing to large (>30 kbp), high-identity tandem duplications within them that could not be spanned by SMRT sequence read lengths (**Extended Data Fig. 1a** and **Table S1**). We encountered this same problem sequencing human haplotypes but were able to overcome it by sequencing shorter clones (~40 kbp) from a fosmid library (ABC12) corresponding to the same human individual (NA12878).

In total, we incorporated 106 clones into our final chromosome 16p11.2 contig assemblies, including 70 BACs sequenced using SMRT technology, 21 BACs previously sequenced using capillary technology, and 15 fosmids sequenced using SMRT technology (**Fig. 1**, **Extended Data Fig. 1**, and **Table S1**). A small gap in unique sequence in one contig (corresponding to the effectively haploid hydatidiform mole, CHM1[41]) was filled using consensus sequence from SMRT whole-genome sequencing (WGS) of the same genome[43]. To enable detailed analysis of the evolutionary history of chromosome 16p11.2, we also compiled or generated sequence data over loci paralogous to duplicated sequences within chromosome 16p11.2. Specifically, we assembled publicly available BAC sequences into contigs covering the chromosome 16p12.1 locus in human and chimpanzee (**Table S1**) and sequenced or collected sequence data from several BACs corresponding to ancestral paralogs of chromosome 16p11.2 duplicated sequences (**Table S1**). All contig sequences and assembled clone inserts are publicly available under the accession code provided at the end of the manuscript.

2. Structural variation

2.1 Segmental duplication analyses

Two approaches were used to annotate and characterize segmental duplications within each assembled haplotype. First, we identified all regions within our sequences homologous to known human segmental duplications by analyzing each sequence using DupMasker[44] (default settings). Second, we identified segmental duplications by applying a whole-genome assembly comparison (WGAC) pipeline[45] to a minimal genome assembly consisting of only our contig sequences, with each treated as a separate "chromosome". These analyses revealed that the chromosome 16p11.2 locus in human and chimpanzee, but not orangutan, has acquired more than 1 Mbp of duplicated sequences originating from over a dozen ancestral genomic loci (**Fig. 1a** and **Extended Data Fig. 1**). The most complex duplication architecture was observed in chimpanzee.

Three segmental duplications are of particular interest: i) a 30 kbp segment containing the genes *BOLA2*, *SLX1*, and *SULT1A3* (as well as a potential fusion gene *SLX1-SULT1A3*); ii) a 72 kbp segment including *NPIP* and the 3' end of *SMG1P*; and iii) a 45 kbp segment harboring the 5' end of *SMG1P* (**Extended Data Fig. 1**). These segmental duplications constitute the largest block of duplicated sequences found at multiple locations within the chromosome 16p11.2 BP1-BP5 region[8] in humans. They flank the autism critical region in direct orientation and, thus, are strong candidates for mediating nonallelic homologous recombination (NAHR) underlying recurrent microdeletions and microduplications[46]. In contrast, the largest segments of duplication within the chimpanzee haplotypes occur in inverted orientation and would promote recurrent inversions, consistent with the observed inversion polymorphism between the two chimpanzee haplotypes. Interestingly, both of these duplications harbor species-specific duplicated sequences—the 30 kbp block including *BOLA2* in humans and an ~160 kbp block originating from chromosome 16p12.1 in chimpanzee.

Comparative analysis of human haplotypes indicates structural variation affects both sides of the critical region, results in large (102 and 95 kbp) blocks of highly identical, directly oriented sequences adjacent to and flanking the critical region, and always involves the same 102 kbp unit, including *BOLA2*, *SLX1*, and *SULT1A3* (**Extended Data Fig. 1b**). These data, together with our *BOLA2* copy number estimates from WGS data (**Fig. 2b**), suggest that humans with extreme *BOLA2* copy numbers differ by >500 kbp as a result of copy number variation. The duplication architecture of the region predisposes it to NAHR both within BP4 and BP5, resulting in tandem expansions and contractions of the variable unit on each side of the critical region, and between BP4 and BP5, resulting in disease-associated microduplications and microdeletions (**Extended Data Fig. 5c**).

2.2 Inversion analysis

We visualized structural differences between sequenced haplotypes using Miropeats[47] and refined breakpoints by sequence alignment using Clustal 2.1[48]. A series of three-color fluorescence *in situ* hybridization (FISH) experiments (**Table S2** and section 9.1) were performed to validate the order and orientation of inversions (**Extended Data Fig. 2**). The results confirm accurate assembly of our contigs and imply extensive reorganization of the chromosome 16p11.2 region over the past ~15 million years of great ape evolution.

3. Evolutionary reconstruction

We put forward a model for the evolution of the chromosome 16p11.2 region in great apes (**Extended Data Figs. 3, 4** and **Table S3**) and detail the evidence supporting each hypothesized step below.

Evolution of chromosome 16p11.2 from the great ape ancestor to the human-chimpanzee ancestor (Steps 1-5)

Ancestral ape genome organization

Comparison of our assembled orangutan sequence contig with orthologous sequence from the mouse genome assembly (GRCm38/mm10 chr7:126110001-127410000) reveals identical gene order and orientation between these two species. Similar to orangutan, mouse sequence over this region is largely devoid of segmental duplications. Both observations indicate the orangutan organization likely represents the great ape ancestral state and the dramatic changes that restructured this region in human and chimpanzee occurred after divergence with orangutan. To determine the ancestral *BOLA2* locus, we examined the order and orientation of genes closest to *BOLA2* copies at BP4 and BP5 but not duplicated between the two loci in human. Considering these genes in our nonhuman primate contigs and in mouse, we observed conserved gene order synteny only between *BOLA2* and *CORO1A*. Because *CORO1A* is present at BP5 but not at BP4 in humans, we conclude that BP5 represents the ancestral locus.

3.1 Step 1: Expansion of the LCR16a segmental duplication

A striking feature of human and chimpanzee assembled sequence contigs is the abundance of ~20 kbp chromosome 16 low-copy repeat (LCR16a) sequences containing the *NPIP* family[7] (red triangles in **Fig. 1** and **Extended Data Figs. 1, 3, 4**). These *NPIP* core duplicons[49] are absent from the chromosome 16p11.2 locus in orangutan and mouse. Previous phylogenetic analyses[7,50] revealed that LCR16a expanded in the human-chimpanzee-gorilla common ancestor and showed that about two-thirds of all copies are orthologous among African great apes. Interestingly, all of the inversion breakpoints between our contig sequences map within regions of segmental duplications including inversely oriented *NPIP* core duplicons—consistent with their involvement in mediating most of the evolutionary inversions. Strikingly, the LCR16a segmental duplication is associated with 14 evolutionary events affecting chromosome 16p11.2 and/or genes therein (**Table S3**), not including its likely role in driving inversions. These events include duplications from another locus on chromosome 16 into chromosome 16p11.2 (Steps 1, 3, and 13), duplications within chromosome 16p11.2 (Steps 5, 7, 9, 10, 11, 14, 18, and 20), interlocus gene conversion (Step 6), and duplication from chromosome 16p11.2 to chromosome 17 (not included in our model schematics).

3.2 Step 2: Evolutionary inversions before human-chimpanzee divergence

For each unique gene-rich region (numbered 2-5 in **Fig. 1a** and **Extended Data Fig. 1a**) bracketed by segmental duplications, we determined the most likely number of inversions using the following logic. If the orientation of the region in a particular species is the same as the orientation of the orthologous region in the orangutan proxy for the great ape ancestor, the region either did not invert or must have inverted an even number of times along that species' lineage after divergence with orangutan. In contrast, if the orientation in the species of interest is opposite that in orangutan, the region must have inverted an odd number of times over the same evolutionary period. For each region of unique sequence, for both human and chimpanzee, we applied maximum parsimony, selecting the path requiring the fewest rearrangements to reconcile modern human and chimpanzee organizations with the ancestral great ape state. Our full evolutionary model (**Extended Data Figs. 3, 4** and **Table S3**) suggests a total of six evolutionary inversions, including two along the branch leading from the great ape ancestor to the human-chimpanzee ancestor, one specific to the human lineage, and three specific to the chimpanzee lineage.

The two distinct inversions occurring along the lineage between the great ape ancestor and the human-chimpanzee ancestor affect unique regions 2-4 (17 genes) and unique region 5 (30 genes including *BOLA2*, *SLX1*, and *SULT1A3*), respectively. Neither precise timing estimates for these events nor their order relative to each other or to most other events along the human-chimpanzee ancestral lineage can be inferred. Because the breakpoints of these inversion events map within regions of segmental duplications

including inversely oriented *NPIP* core duplicons, it appears likely that NAHR between inverted *NPIP* copies mediated these inversions. Thus, these inversions are displayed together in this step, predating the human-chimpanzee common ancestor, but occurring after the dispersal of *NPIP* across the chromosome 16p11.2 locus.

3.3 Step 3: Duplicative transposition between chromosome 16p12.1 and chromosome 16p11.2

A complex higher-order duplication block (~255 kbp) located between BP3 and BP4 in humans and at the orthologous chimpanzee position maps to two locations in these genomes: chromosome 16p11.2 and chromosome 16p12.1. Although partial fragments of this block are found elsewhere on chromosome 16, the identical structure (with respect to order and orientation of smaller segmental duplications) at these two positions implies that the components (individual duplicons) first evolved at one of these loci, followed by the larger cassette duplicating to the other locus. We estimated the timing of this duplication using a molecular clock approach, generating a phylogenetic tree incorporating sequence over a region of the cassette that is unique sequence in orangutan and, therefore, orthologous to both chimpanzee and human.

To estimate the evolutionary age of duplications into and within the chromosome 16p11.2 locus, we calibrated local molecular clocks based on sequence divergence of paralogs and orthologs and assumed divergence times of 6 mya between human and chimpanzee[10] and 15 mya between human or chimpanzee and orangutan[38]. Specifically, we generated multiple sequence alignments including relevant sequences from our contigs (**Table S1**), reference genome assemblies, and sequenced BACs (**Table S1**) using Clustal 2.1[48] and fixed alignment errors manually (including removing regions of poor alignment quality) using Jalview[51]. We built a series of neighbor-joining phylogenetic trees using MEGA5[52] with the complete deletion option, calculating genetic distances using the Kimura 2-parameter model with standard error estimates based on an interior branch test of phylogeny with 500 bootstrap replicates. For each tree, we performed several Tajima's relative rate tests to assess the validity of the molecular clock assumption.

For phylogenetic trees where the molecular clock was supported, we estimated divergence times corresponding to duplication events of interest using the equation T = K/2R, where T is time (in millions of years), K is divergence (substitutions per site), and R is the substitution rate (substitutions per site per million years). It follows that the divergence times corresponding to duplication events of interest can be estimated without explicitly calculating the substitution rate if assumptions are made regarding divergence times between species. Specifically, the fraction of sequence divergence after the duplication event relative to the total divergence between duplicated sequences of interest and a single-copy orthologous outgroup sequence is equal to the fraction of time that elapsed after the duplication event relative to the total divergence time between the species having the duplicated sequences and the outgroup species. For trees where the molecular clock was not supported, we scaled sequence divergences to match the substitution rate of branches corresponding to the ancestral locus. These divergence corrections are noted in the text.

For this initial duplication from chromosome 16p12.1 to chromosome 16p11.2 including the 5' end of the gene *SMG1* (i.e., including *SMG1P*), we scaled duplicate branch lengths to the branch lengths for the ancestral *SMG1* locus (at chromosome 16p12.3), as the tree did not pass Tajima's relative rate test. We estimate that duplication of the cassette occurred ~8.9 mya. We cannot confidently infer the directionality of this duplication, so the chromosome 16p11.2 duplicons may in fact be older if chromosome 16p11.2 is the ancestral locus for the cassette. However, the fact that later duplication events have clearly transferred large pieces of sequence to chromosome 16p11.2 within the BP3-BP4 region or its chimpanzee counterpart (Steps 4, 13, and 16—twice from chromosome 16p12.1) suggests the direction of this duplication was most likely from chromosome 16p12.1 to chromosome 16p11.2.

3.4 Step 4: Duplicative transposition from chromosome 16q24.2 to chromosome 16p11.2

The second largest component of the complex block of duplications between BP3 and BP4 in humans is an ~175 kbp segment originating from chromosome 16q24.2. Tajima's relative rate test indicated that the chromosome 16p11.2 copy did not evolve at the same relative rate as the chromosome 16q24.2 copy. Adjusting the chromosome 16p11.2 branch length accordingly, we estimate this duplication event occurred ~7.5 mya, after the duplication of *SMG1P* from chromosome 16p12.1 to chromosome 16p11.2 but before the duplication of *SMG1P* within chromosome 16p11.2. Indeed, sequence analysis indicates that this segmental duplication disrupts the contiguity of the ~255 kbp *SMG1P* duplication block (Step 3), unequivocally making it a secondary event after the initial duplication of *SMG1P* into chromosome 16p11.2. Consistent with our timing estimate suggesting this duplication from chromosome 16q24.2 occurred near the time of human-African great ape speciation, FISH experiments (**Table S2**) show human and chimpanzee have this ~175 kbp segment duplicated between chromosome 16q24.2 and chromosome 16p11.2, while orangutan and gorilla lack the duplicate copy at chromosome 16p11.2 (data not shown).

3.5 Step 5: Duplicative transposition from BP4 to BP5 within chromosome 16p11.2

The duplication of an ~95 kbp segment that included *SMG1P* from BP4 to BP5 is the least confident step in our evolutionary reconstruction. Approximately 67 kbp of this segment at BP5 was subsequently subjected to interlocus gene conversion along the human lineage (see Step 6 for evidence), obscuring the phylogenetic signal of its duplication history. Phylogenetic analysis of sequence within the ~95 kbp segment not affected by conversion suggests this duplication occurred ~6.3 mya, around the time of human-chimpanzee speciation. Although there are at least seven *SMG1P* copies on human chromosome 16, BP4 and BP5 *SMG1P* copies share sequence with the ancestral *SMG1* locus that is not shared with other *SMG1P* duplicates. Furthermore, chromosome 16p11.2 *SMG1P* copies are more highly identical to one another than to the ancestral locus. These observations are consistent with the phylogenetic inference that the BP5 copy originated as a result of duplication from BP4.

Human-specific evolution of chromosome 16p11.2 (Steps 6-11)

3.6 Step 6: Complex interlocus gene conversion event between chromosome 16p12.1 and 16p11.2

The duplication architecture at BP5 in humans exhibits an unusual property. Except for the 30 kbp segment including *BOLA2*, most sequence duplicated between BP5 and BP4 is also duplicated between BP5 and chromosome 16p12.1. Strikingly, this duplication between BP5 and chromosome 16p12.1 does not exhibit uniform sequence identity across the alignment, instead showing >99% identity over most of the 72 kbp block and lower identity (<99%) over the remainder of the 72 kbp block and over the 45 kbp block. This sequence identity pattern suggests the duplications at BP5 are likely a product of more than one evolutionary event. A simple scenario involving a single duplication of the 72 kbp and 45 kbp blocks in concert from BP4 to BP5 fails to explain the non-uniform sequence identity with chromosome 16p12.1. Visualizing the patterns of sequence identity between the paralogous loci suggests a complex human-specific interlocus gene conversion involving BP5, BP4, and sequence from chromosome 16p12.1 (data not shown). Combining our sequence identity analysis with phylogenetic timing, we propose that BP5 acted as a conversion acceptor for 54 kbp of sequence from chromosome 16p12.1 and 36 kbp of sequence from BP4 ~1.8 mya.

3.7 Step 7: Duplicative transposition from BP2 to BP1 within chromosome 16p11.2

The high sequence identity (>99.6%) of an ~115 kbp duplication from BP2 to BP1 suggests this event likely occurred specifically along the human lineage (< ~1.8 mya).The resulting architecture rendered the interstitial region susceptible to NAHR-mediated inversion associated with asthma and obesity[53].

3.8 Step 8: Human ~450 kbp inversion polymorphism

The ~450 kbp inversion between BP1 and BP2 has been previously reported as an inversion polymorphism in humans. It was estimated to have occurred ~1.35 mya[53,54].

3.9 Step 9: Tandem 102 kbp segmental duplication at BP5

The duplication architecture of BP4 haplotypes including *BOLA2B* matches the structure of BP5 haplotypes having a tandem duplication of the variable 102 kbp unit (**Extended Data Fig. 1b**). In fact, the junction sequences between the end of the 72 kbp block and the start of the 30 kbp block (i.e., junctions between *SMG1P* and *BOLA2*) at BP4 are identical to the same junction sequence at BP5 in the H2 haplotype contig. Since the BP5 *BOLA2* paralog is ancestral, we hypothesize that the *Homo sapiens*-specific duplication of *BOLA2* across the critical region originated from a BP5 haplotype having a tandem duplication of the variable 102 kbp unit. To test this hypothesis, we performed a series of sequence alignments and sliding window sequence identity analyses. The results corroborate our proposed scenario for the formation of the *BOLA2B* paralog, suggesting tandem duplication of the 102 kbp unit at BP5 preceded the duplication including *BOLA2* and the junction sequence from BP5 to BP4 (**Extended Data Fig. 6**). Because we do not observe duplicate *BOLA2* copies in archaic hominins, this tandem duplication most likely occurred specifically in *Homo sapiens*. It is possible this tandem duplication occurred before the human-archaic hominin species split but is absent from archaic hominins due to incomplete lineage sorting. We have no evidence that this tandem duplication is polymorphic in archaic hominins, although sampling of such genomes is limited.

3.10 Step 10: Duplicative transposition of 95 kbp (including *BOLA2*) from BP5 to BP4 within chromosome 16p11.2

Analyses of sequence identity (**Extended Data Fig. 6a-c**) and informative sites extracted from a multiple sequence alignment (**Table S5**) allowed us to resolve the breakpoints of the *BOLA2* duplication from BP5 to BP4 within 1 kbp. We delineate an ~95 kbp region within *BOLA2B*-containing BP4 haplotypes corresponding to *Homo sapiens*-specific duplicate sequence having originated from BP5 (**Extended Data Fig. 6a**). The data suggest the duplication structure at BP4 in the H2 haplotype corresponds to the ancestral state of the BP4 locus (**Extended Data Fig. 6b**). The *BOLA2* duplication likely involved template switching between BP4 and BP5 during DNA replication, resulting in the duplication of 95 kbp from BP5 to BP4 along with the duplication of a 7 kbp segment within BP4 (**Extended Data Fig. 6d**).

To estimate the evolutionary timing of when the *BOLA2* duplication occurred, we generated a multiple sequence alignment spanning an ~21 kbp region within the 30 kbp block including *BOLA2*, *SLX1*, and *SULT1A3* using sequences from our contigs and from a gorilla clone containing orthologous sequence (CH277-206A9). Estimates of molecular divergences between human sequences were very low (**Table S6**), and phylogenetic analysis did not show human sequences at BP4 and BP5 forming distinct clades (**Fig. 2a** and **Extended Data Fig. 6e**). This result suggested either *BOLA2* duplicated very recently, such that the common ancestor of alleles at BP5 from our haplotype sequences is older than the *BOLA2* duplication event, or that interlocus gene conversion occurred between BP4 and BP5. To distinguish between these possibilities, we constructed a larger ~88 kbp alignment and phylogenetic tree using the full extent of sequence shared between all human copies. Here we observed sequences at BP4 and BP5 forming distinct clades (data not shown), suggesting interlocus gene conversion underlies the branching pattern in the original tree. Assuming a human-chimpanzee divergence time of 6 mya[10] and a constant substitution rate (**Table S7**), we estimate that *BOLA2* duplicated across the critical region ~282 kya, around the time when contemporary *Homo sapiens* emerged as a species[11] (**Fig. 2a** and **Extended Data Fig. 6e**). This estimate is consistent with our *BOLA2* copy number estimates in humans and archaic hominins (**Fig. 2b-c**).

We computed a 95% confidence interval for our *BOLA2* duplication timing estimate using branch length error estimates and the following procedure. First, for each branch in the tree, we set the branch length to a randomly selected value between the actual branch length minus the branch length error (or zero if that value is negative) and the actual branch length plus the branch length error, inclusive. Second, we calculated a timing estimate using the same calculations as for the original tree except with modified

branch length values. Third, we repeated the above two steps until one million modified trees and corresponding timing estimates were obtained. Fourth, we sorted the timing estimates and reported the 25,000th and 975,000th sorted timing estimate values as the 95% confidence interval: 361-209 kya.

3.11 Step 11: Polymorphic 102 kbp expansions and contractions at BP4 and BP5

Comparative sequence analyses of distinct human haplotypes delineate the nature and spatial extent of copy number variation within the chromosome 16p11.2 locus in humans and provide insight into the mechanism by which it occurs (**Extended Data Fig. 1b**, **Extended Data Fig. 5c**, and section 2.1). Variation in human genomes is largely restricted to the 102 kbp segmental duplication including *BOLA2* at both BP5 (*BOLA2A*) and BP4 (*BOLA2B*) (**Fig. 2c** and section 4). Thus, we incorporate this knowledge into the final step of our evolutionary model for humans, highlighting a likely ongoing process of tandem expansions and contractions including *BOLA2* at BP4 and BP5 via NAHR.

Chimpanzee-specific evolution of chromosome 16p11.2 (Steps 12-20)

3.12 Step 12: Chimpanzee-specific ~1.5 Mbp inversion

Our inversion analysis (section 3.2) suggests three inversions occurred specifically along the chimpanzee lineage. The largest such inversion included unique regions 3-5. The timing and order of this inversion relative to other chimpanzee-specific inversions cannot be inferred.

3.13 Step 13: Duplicative transposition from chromosome 16p12.1 to chromosome 16p11.2

The chimpanzee orthologs to human chromosome 16p11.2 BP4 and BP5 both contain chimpanzee-specific duplications originating from two separate locations at chromosome 16p12.1. The juxtapositions of these two chromosome 16p12.1 segments only at BP4 and BP5 implies that this duplication architecture first evolved at either BP4 or BP5, followed by the larger cassette duplicating to the other locus. The extent of contiguous sequence shared between BP4 and chromosome 16p12.1 is longer than that shared between BP5 and chromosome 16p12.1, and this extent includes junction sequence (between duplicons) not present at BP5 or at human BP4 (a proxy for chimpanzee BP4 prior to chimpanzee-specific duplications). These observations imply that the complex chimpanzee-specific duplication architecture first evolved at BP4. Phylogenetic analysis suggests the first chimpanzee-specific duplication (~170 kbp) from chromosome 16p12.1 to chromosome 16p11.2 BP4 occurred ~5.2 mya.

3.14 Step 14: Duplicative transposition within chromosome 16p11.2 into unique sequence

Comparison of our sequenced haplotype contigs (**Fig. 1a** and **Extended Data Fig. 1a**) reveals that a chimpanzee-specific ~130 kbp segment originating from BP2 duplicated to a region of unique sequence, separating regions 3 and 4. Phylogenetic analysis suggests this chimpanzee-specific duplication occurred ~4.8 mya. Because not all sequence at the duplicate locus is also present at BP2, it is likely the duplication block between unique regions 3 and 4 formed via multiple events.

3.15 Step 15: Chimpanzee-specific ~215 kbp inversion

The chimpanzee-specific duplication that separated unique regions 3 and 4 included *NPIP*, resulting in unique region 3 becoming flanked by inversely oriented *NPIP* repeats. NAHR between these inverted *NPIP* copies likely mediated the ~215 kbp inversion of unique region 3 found exclusively in chimpanzees.

3.16 Step 16: Duplicative transposition from chromosome 16p12.1 to chromosome 16p11.2

The second chimpanzee-specific duplication from chromosome 16p12.1 to chromosome 16p11.2 BP4 included ~160 kbp of unique sequence. Tajima's relative rate test indicated that the chromosome 16p11.2 copies did not evolve at the same relative rate as the chromosome 16p12.1 copy. Adjusting the chromosome 16p11.2 branch lengths accordingly, we estimate this duplication event occurred ~4.6 mya.

3.17 Step 17: Duplicative transposition of sequence to chromosome 16p11.2

The start of the chimpanzee C1 haplotype contig consists of >50 kbp of duplicated sequence not found in orthologous locations in human or orangutan, implying that this sequence resulted from a duplicative transposition event specific to the chimpanzee lineage.

3.18 Step 18: Duplicative transposition from BP4 to BP5 within chromosome 16p11.2

The largest (>420 kbp) chimpanzee-specific duplication transposed sequence from BP4 to BP5, resulting in large blocks of inversely oriented duplicated sequences flanking unique regions 3 and 4 in chimpanzee (**Fig. 1a** and **Extended Data Fig. 1a**). Phylogenetic analysis suggests this duplication event occurred ~630 kya.

3.19 Step 19: Chimpanzee >1 Mbp inversion polymorphism

We discovered a large (>1 Mbp) inversion polymorphism in chimpanzee from our haplotype sequencing (**Extended Data Fig. 1a**) and subsequently confirmed this inversion in a different chimpanzee individual via FISH (**Extended Data Fig. 2b**). We mapped the breakpoints of this inversion to the second chimpanzee duplication from chromosome 16p12.1 (data not shown), implying the inversion occurred via NAHR between the large chimpanzee-specific inversely oriented duplication blocks at BP4 and BP5.

3.20 Step 20: Polymorphic tandem expansions including *NPIP*

Duplication analyses of our chimpanzee haplotype contigs revealed a polymorphic ~80 kbp tandem duplication at BP2 including *NPIP* and *EIF3C*, as well as several ~20 kbp duplications including only *NPIP*.

4. Copy number genotyping

4.1 Overview

We employed three complementary methods to genotype *BOLA2* copy number in a total of 2,824 humans[32,33], 3 archaic humans[13,14], 1 Neanderthal[2], 1 Denisovan[3], 14 bonobos[34], 23 chimpanzees[34], 32 gorillas[34], and 17 orangutans[34]. Genotypes for all samples are provided in **Table S8**, and **Table S9** shows population summary statistics. First, we estimated aggregate *BOLA2* copy number, along with *SLX1* copy number and *SULT1A3* copy number, using a previously described approach based on WGS read depth[12]. Second, we inferred aggregate and paralog-specific *BOLA2* copy number by examining relative WGS read depth over genetic markers distinguishing regions of interest. Third, we targeted a subset of these genetic markers as well as polymorphisms for molecular inversion probe (MIP) capture. Using massively parallel sequencing, we determined aggregate and paralog-specific *BOLA2* copy number using relative read-depth analysis[24] as above. Although only a subset of samples were genotyped using multiple methods (**Table S8**), distributions of aggregate *BOLA2* copy number estimates were similar among the methods. This suggests that results from each method reliably reflect the distribution of aggregate *BOLA2* copy number genotypes in the human population, except for high-copy estimates based on WGS read depth which showed the lowest validation rate. We detail each method below, with particular focus on the second approach which has not been previously described.

4.2 Aggregate *BOLA2*, *SLX1*, and *SULT1A3* copy number genotyping using WGS read depth

We genotyped aggregate copy number for all genes within the 30 kbp segment duplicated between BP5 and BP4 using a WGS read-depth method as previously described[12]. We measured sequence read depth over an ~5 kbp region that extends from just beyond *CORO1A* to *BOLA2* (GRCh37 chr16:30200398-30205627). To test the accuracy of these estimates, we performed FISH on cell lines derived from individuals having predicted *BOLA2* copy numbers from three to nine using fosmid probes overlapping

part of the 30 kbp segment (including *BOLA2*) as well as unique sequence adjacent to this block at BP5 (**Table S2**). FISH analysis validated individuals having *BOLA2* copy number as low as three and as high as eight (**Extended Data Fig. 7b**) but also showed some discrepancies (3 of 7 genotypes were discordant for higher copy number). Higher copy numbers are more difficult to discern because tandem *BOLA2* copies are only 102 kbp apart and such FISH signals cannot always be discriminated on interphase nuclei. Low sequence coverage for samples from the 1000 Genomes Project[32] is another source of error. We have previously shown that the accuracy of copy number estimates based on WGS read depth correlates positively with increasing sequencing coverage[12]. As a result, we consider copy number estimates from 236 genomes from the Simons Genome Diversity Project[33] to be more accurate than those from 1000 Genomes Project genomes. With one exception, all FISH estimates were concordant with WGS read-depth estimates within 1 diploid copy number.

Since *SLX1* and *SULT1A3* were also part of the 102 kbp *Homo sapiens*-specific duplication involving *BOLA2*, we reasoned that they may also be genes duplicated specifically in our species. We assessed their copy number using a genomic segment (~4 kbp) corresponding to *SLX1* (GRCh37 chr16:30205164-30208887) and an ~11 kbp segment including *SULT1A3* (GRCh37 chr16:30210549-30221660). Only *BOLA2* shows convincing evidence of being a *Homo sapiens*-specific duplicated gene (**Fig. 2b**). We identify a few humans with potentially two copies of *SLX1*, and multiple nonhuman primates potentially have three *SLX1* copies. We cannot definitively exclude the possibility that *SLX1* is duplicated specifically in *Homo sapiens* due to the imprecision associated with genotyping such a small genomic segment using this method. In contrast, all nonhuman primates were estimated to have three or more copies of *SULT1A3*. *SULT1A3* is unambiguously duplicated in nonhuman primates—a finding confirmed by the identification of a chimpanzee BAC clone from chromosome 17 harboring a duplicate *SULT1A3* paralog (RP43-175I2).

4.3 *BOLA2* paralog-specific copy number (PSCN) genotyping

We also genotyped *BOLA2* paralog-specific copy number (PSCN) by adapting previously described strategies[12,24] based on sequencing read depth over informative paralogous sequence variants. Informative genetic markers have three properties: i) they distinguish sequences of interest (chromosome 16p11.2 BP4 and BP5) from paralogous sequences elsewhere in the genome; ii) they distinguish different copies (e.g., BP4 from BP5 copies); and iii) they vary predictably with differences in paralog-specific *BOLA2* copy number. Sequencing reads containing such markers can be unambiguously assigned to particular copies of duplication blocks at chromosome 16p11.2, allowing quantification and inference of *BOLA2* PSCN.

Using our high-quality human haplotype sequences, we identified 70-mer markers strictly meeting the criteria above only within the 72 kbp block (**Extended Data Fig. 1**). Copy number of the 72 kbp block varies in concert with *BOLA2* copy number. The 72 kbp block and the 30 kbp block (which includes *BOLA2*) together constitute the 102 kbp variable unit. Thus, aggregate *BOLA2*, *BOLA2A* (BP5), and *BOLA2B* (BP4) copy number can be deduced from knowledge of total and BP5 72 kbp block copy number.

BP4 72 kbp blocks from *BOLA2B*-containing haplotypes are comprised of both ancestral BP4 sequence and sequence derived from the 95 kbp duplication containing *BOLA2* from BP5 (**Extended Data Fig. 6a-d**). This hybrid architecture effectively divides the 72 kbp block into three sectors (**Extended Data Fig. 7c**): i) a 59 kbp sector with markers distinguishing the most telomeric BP4 copy (B markers, blue box) from all other copies (R markers, red boxes); ii) a 7 kbp sector with markers distinguishing all BP4 copies (green boxes) from all BP5 copies (BP5 markers, orange boxes); and iii) a 6 kbp sector with markers distinguishing the most centromeric BP4 copy (purple box) from all other copies (yellow boxes). B markers do not vary in copy number in our sequenced human haplotypes—each haplotype had exactly one complete set of these markers. Thus, diploid genomes are expected to contain two complete sets of B markers. To assess this expectation, we estimated copy number over the most telomeric BP4 59 kbp

sector (i.e., the region harboring B markers) from WGS read depth over unique 30-mers in GRCh37 using WGS data from 236 humans from the Simons Genome Diversity Project sequenced to high coverage[33]. Over 96% of individuals were genotyped as having two diploid copies of this region, suggesting that nearly all genomes contain two complete sets of B markers.

We leveraged WGS sequencing data and all 59 kbp sector markers to determine aggregate 72 kbp block copy number as follows. We extracted all reads containing a 70-mer perfectly matching a 59 kbp sector marker and counted the number of reads corresponding to each such marker. For every marker pair (a B marker and its R marker counterpart), we plotted marker-specific read count relative frequencies, calculated as the number of reads corresponding to the marker of interest (B or R) divided by the number of reads corresponding to both markers. Because there are generally two complete sets of B markers in a diploid genome, B marker-specific read count frequencies have values around 2/N, where N is the total number of 72 kbp blocks. Thus, we used this reasoning to infer the total number of 72 kbp blocks from B marker-specific read count frequency data.

Once the total number of 72 kbp blocks (N) is known, the number of *BOLA2* copies equals N - 2 because there is parity between the number of sets of R markers (N - 2) and the aggregate number of copies of *BOLA2*. Performing the same analysis of WGS data using markers within the 7 kbp sector allows inference of BP5 72 kbp block copy number. BP5 marker-specific read count frequencies have values around C/N, where C is the number of BP5 72 kbp blocks (and by parity also the number of *BOLA2A* copies) and N is the total number of 72 kbp blocks, as above. We applied this inferential strategy to determine diploid aggregate *BOLA2* copy number and diploid *BOLA2A* copy number, with the difference between these values corresponding to diploid *BOLA2B* copy number.

We evaluated this approach by applying it to genotype *BOLA2* PSCN for NA12878 using high-coverage WGS data[55,56], plotting marker-specific read count relative frequencies (**Extended Data Fig. 7d**) and inferring diploid copy number of the 72 kbp block and of BP5 72 kbp blocks by manual inspection. With knowledge of these copy numbers, we employed the reasoning above to infer diploid aggregate *BOLA2, BOLA2A,* and *BOLA2B* copy numbers. Importantly, we know this individual has three diploid copies of *BOLA2A* and one diploid copy of *BOLA2B* because this individual was the source of genomic material for the BAC and fosmid libraries utilized in generating our H1 and H2 human haplotype sequences. The results show that, using this approach, we can accurately infer paralog-specific *BOLA2* copy number. We also successfully inferred *BOLA2* PSCN for the effectively haploid hydatidiform mole CHM1[41], the source of the BAC library used for generating our H3 haplotype sequence (data not shown). Finally, we accurately inferred PSCN of the 45 kbp block using a similar approach (**Extended Data Fig. 7d**) founded on the largely copy number invariant nature of the 45 kbp block. WGS read depth analysis[12] of 236 humans[33] over unique 30-mers in GRCh37 suggested BP4 and BP5 45 kbp blocks are each nearly always diploid copy number two (>98% of individuals examined) in humans.

We employed this WGS analysis strategy to estimate *BOLA2* PSCN by manually inspecting marker-specific read count frequency plots for 236 humans[33], three archaic humans[13,14], a Neanderthal[2], and a Denisovan[3]. We excluded all WGS data from the 1000 Genomes Project[32], as the sequencing coverage for corresponding genomes is too low for accurate PSCN estimation using this method. Nevertheless, using this approach, we were able to estimate high-confidence *BOLA2* PSCN genotypes for 94% of humans (222 of 236) as well as for the Neanderthal and the Denisovan (**Fig. 2c**).

For the remaining 14 humans (a subset of individuals of African ethnicity) and for all three archaic human genome sequence datasets, we could not confidently infer *BOLA2* PSCN. B marker read count frequencies were not uniform across the spatial extent of the 59 kbp sector for the 14 humans of African descent, a signature implying either i) tandem 102 kbp expansion or contraction involving NAHR between BP4 72 kbp blocks, ii) interlocus gene conversion, or iii) structural variation affecting part of the

59 kbp sector, in each case resulting in some B markers no longer existing at two diploid copies and/or non-uniform aggregate copy number over the spatial extent of the 72 kbp block. As for the archaic humans, the marker-specific read count frequency data were too noisy to discriminate between alternative possible genotypes, likely in part due to the requirement for perfect matching between sequencing reads and 70-mer markers. Despite these challenges, the data clearly show *BOLA2* is duplicated in all humans, including archaic humans.

Our analysis indicates that *BOLA2A* copy number (standard deviation = 0.77) is more variable than *BOLA2B* copy number (standard deviation = 0.46). This differential variability is consistent with the nearly identical 102 kbp tandem duplications within BP5 being more likely to mediate NAHR than the directly oriented duplications of the 72 kbp block within BP4, which contain only a few small regions of near-perfect sequence identity. Our genotyping results also suggest that some individuals have more than two copies of *BOLA2B.* Prior to our WGS analyses, we developed a FISH assay for genotyping paralog-specific *BOLA2* copy number (**Table S2**). Interphase FISH on the first individual tested revealed both chromosome 16 homologs having *BOLA2* paralogs at BP4 and BP5. The fluorescence intensities showed evidence of one homolog harboring multiple copies of *BOLA2A* and the other including multiple copies of *BOLA2B* (**Extended Data Fig. 7a**). This finding provides experimental confirmation that individuals exist in the human population having more than two copies of *BOLA2B*.

The PSCN genotyping method has some limitations. It requires high sequencing coverage, and interlocus gene conversion events over markers critical for copy number inference may yield inaccurate PSCN genotypes if not detected and taken into account. Small focal deletions or duplications affecting *BOLA2* but not the 72 kbp block would be completely missed using our approach. Marker-specific read count frequencies become more difficult to distinguish as the aggregate copy number of the 72 kbp block increases. For example, distinguishing 2/9 from 2/10 is much more difficult than differentiating 2/6 from 2/7. Thus, higher *BOLA2* copy number estimates using this method are inherently less confident than lower *BOLA2* copy number estimates. Finally, even if successful, our method cannot determine the configuration of specific haplotypes, but only diploid PSCN genotypes. All these considerations also apply to the MIP method detailed in section 4.4 below, which benefits from higher sequence coverage (>20-fold)[24] but relies on fewer markers because not all useful markers can be successfully targeted for capture.

4.4 *BOLA2* PSCN genotyping using MIPs

To enable large-scale genotyping of *BOLA2* PSCN, we designed MIPs targeting 112-mer markers differentiating the same groups of 72 kbp blocks as in our WGS marker-specific read count frequency analyses (section 4.3). 112 bp informative k-mers were specifically selected because this represents the distance between MIP targeting arms. We used the same general parameters for MIP design as previously described[24], except the copy count threshold for arm hybridization sequences was 30 rather than 8 and no filtering was performed based on target G+C content. Only 20 MIPs were successfully designed meeting the above criteria. We also designed an additional 27 MIPs corresponding to polymorphic single-nucleotide variants to improve genotyping precision. We performed MIP capture, sequencing, and analysis as previously described[24] using single-molecule MIPs[35] (**Table S10**), mapping sequencing data to a minimal genome consisting of all BP4 and BP5 72 kbp blocks plus all paralogous sequences from GRCh37 outside chromosome 16p11.2. We inferred *BOLA2* PSCN as described above (**Extended Data Fig. 7e**) for 894 humans (**Extended Data Fig. 7f**). Importantly, we used this method to evaluate *BOLA2* copy number estimates for individuals at the extremes of the aggregate *BOLA2* copy number estimate distribution based on WGS read depth, confirming *BOLA2* copy number is at least three and at most eight among all humans examined.

In summary, we assessed copy number estimates at the extremes of the distribution for humans. We genotyped all individuals predicted to have three *BOLA2* copies or more than eight using our MIP

approach. The maximum *BOLA2* copy number validated was eight. The majority of the low-copy estimates (15/21) were experimentally validated as four copies as opposed to three. No humans were found to have fewer than three *BOLA2* copies. Neanderthal and Denisova were found to have two *BOLA2* copies based on examination of WGS marker-specific read count frequencies (section 4.3). Thus, the conclusion that *BOLA2* is duplicated specifically in *Homo sapiens* is robust to uncertainties in human and archaic hominin *BOLA2* copy number estimates based on the WGS read-depth method.

5. Population genetic analyses

5.1 Overview

Because time is necessary for a new mutation to reach high frequency, the near fixation of *BOLA2B* in humans contrasted with its young evolutionary age suggests a rapid increase in allele frequency. We modeled various evolutionary scenarios and assessed the joint probability of our observed genotype data for a single Neanderthal, a single Denisovan, 4 San individuals, and 55 Yorubans. For most scenarios, we also calculated the joint probability of observing *BOLA2B* absent from archaic hominins and present at observed or higher frequencies in humans. We performed five sets of simulations, each based on the same underlying model of human demographic history[16]: i) neutral evolution without specifying *BOLA2B* age, ii) neutral evolution assuming *BOLA2B* formed 282 kya (our point-estimate for *BOLA2B* age based on phylogenetic analysis), iii) neutral evolution assuming *BOLA2B* formed at a specified age varied from 200 kya to 2 mya, iv) evolution under positive selection assuming *BOLA2B* formed 282 kya, and v) neutral evolution assuming *BOLA2B* originated at a specified age (both 282 kya or 650 kya were considered) and allowing recurrent *BOLA2B* formation at a specified constant rate thereafter. All scripts used to perform these simulations are freely available via GitHub at https://github.com/xnuttle/BOLA2. These analyses suggest *BOLA2B* likely did not rise to high frequency in *Homo sapiens* under neutrality. Finally, we examined patterns of genetic variation within the unique chromosome 16p11.2 critical region, revealing a lack of archaic introgression, low diversity, and an excess of rare variants, observations consistent with possible selection.

5.2 Coalescent simulations

We adapted a published demographic model[16] for the *Homo* lineage to simulate the evolution of *BOLA2B*. Our adaptation includes the Neanderthal and Denisova species as well as the Yoruban and San human populations (**Extended Data Fig. 8a**). We modeled the duplication that formed *BOLA2B* as a point mutation that occurred once in history with no recurrent mutation or reversion, considering individuals having two or more *BOLA2B* copies as homozygous for the derived state, individuals having a single copy as heterozygous, and individuals lacking *BOLA2B* as homozygous for the ancestral state. Although it is possible that individuals having two *BOLA2B* copies could be effectively heterozygous by having both copies on the same homolog, never observing a human lacking *BOLA2B* and only rarely observing individuals with four *BOLA2B* copies suggests this possibility is remote. Under these assumptions, our PSCN genotype data indicate that 53 Yorubans are homozygotes for the derived state, 2 Yorubans are heterozygotes, 4 San individuals are homozygotes for the derived state, and both Neanderthal and Denisova are homozygotes for the ancestral state.

In our first set of simulations, we used the coalescent simulator *ms*[17] (implemented within *msms*[18]) to assess the joint probability of observing *BOLA2B* on at least 108 of 110 sampled Yoruban haplotypes and 8 of 8 San haplotypes while not observing *BOLA2B* in archaic hominins, conditional on its presence on at least one sampled human haplotype. We performed 10,000,000 simulations and generated a heat map (data not shown) to show how often each possible genotype outcome occurred in simulated data where *BOLA2B* was absent from archaic hominins, with each genotype outcome defined by the numbers of simulated Yoruban and San haplotypes having *BOLA2B*. Under this scenario modeling neutral evolution

without accounting for *BOLA2B* age, we find the observed genotype data improbable. Only 84,953 of 8,804,486 simulations where *BOLA2B* was present on at least one human haplotype had *BOLA2B* present exclusively in humans at allele frequencies as high as or higher than those inferred from our genotyping data given the assumptions above ($p < 0.0097$).

5.3 Assessing different evolutionary ages of *BOLA2B*

Our initial simulations did not make use of additional information we have regarding the evolution of *BOLA2B*, namely, its age estimated from our phylogenetic analyses. To condition on *BOLA2B* age, we performed simulations with *msms*[18], assuming that *BOLA2B* originated 282 kya and was not subject to selection. In no simulation out of 1,000,000 did *BOLA2B* reach even close to the observed frequencies in Yoruban and San populations (**Extended Data Fig. 8b**). In most simulations (999,531), *BOLA2B* was lost from the human species. Extending this analysis to 10,000,000 simulations yielded the same pattern: no simulation where *BOLA2B* was not lost (n = 4,788) had *BOLA2B* present at frequencies anywhere near the observed frequencies. The fact that the highest simulated *BOLA2B* frequencies were not close to the observed frequencies argues that *BOLA2B* having risen to high frequency in *Homo sapiens* within the last 282,000 years is very unlikely under neutrality.

To account for uncertainty surrounding the age of *BOLA2B*, we asked how old it would have to be to have risen to high frequency exclusively in *Homo sapiens* under neutral evolution. We again employed *msms*[18] and performed simulations excluding the possibility of selection, exploring a range of possible *BOLA2B* ages. We varied *BOLA2B* age from 200,000 years old to 2 million years old, performing 10,000,000 simulations for each age in this range at all 25,000 year increments. We calculated the relative likelihood of our genotype data for each age value as the proportion of simulations yielding the observed genotype data among simulations at that particular age value where *BOLA2B* was present only in humans divided by the maximum such proportion for any single age value, considering all age values examined. Regardless of the age considered, the observed genotype data were always unlikely, as were simulation outcomes with *BOLA2B* present at higher frequencies exclusively in humans. The most likely age for *BOLA2B* based on simulations alone, assuming neutral evolution, was 1.575 million years old, more than five times older than our age estimate determined from phylogenetic analysis. The highest proportion of simulations with *BOLA2B* present exclusively in humans at observed or higher frequencies among simulations where *BOLA2B* was present on at least one human haplotype occurred at a *BOLA2B* age of 1.75 mya (42 of 890 such simulations, $p < 0.0472$). Considering only *BOLA2B* ages equal to or younger than 700 kya (around the estimated divergence of modern humans from archaic hominins[2]), the highest same proportion occurred at age 650 kya (3 of 2,105 simulations, $p < 0.0015$). Importantly, no simulations assuming a *BOLA2B* age within the 95% confidence interval for our phylogenetic age estimate met the *BOLA2B* frequency criteria (presence exclusively in humans at observed or higher frequencies). Together, these results support our conclusion of non-neutral evolution for *BOLA2B* and indicate this conclusion is not dependent on the validity and precision of our estimate for the age of *BOLA2B*.

5.4 Estimating positive selection

Given that the observed high frequencies of *BOLA2B* in humans were improbable under neutrality (especially when age is taken into account), we asked if the data were better explained by a model where *BOLA2B* was driven to high frequency by positive selection. We again used *msms*[18] and incorporated *BOLA2B* age (282 kya), this time assuming a model of genic selection, with *BOLA2B* homozygotes assigned a relative fitness of $(1 + s)$, heterozygotes assigned a relative fitness of $(1 + s/2)$, and homozygotes lacking *BOLA2B* assigned a relative fitness of 1, where s is the selection coefficient. We initially explored a wide range of values for s using a small number of simulations to get a sense of selection strengths most consistent with the data. Eventually, we settled on a narrow range from $s = 0.0009$ to $s = 0.0024$ where we performed 10,000,000 simulations for each value of s and varied s by

increments of 0.0001. We calculated the relative likelihood of the observed San and Yoruban *BOLA2B* genotypes for each value of *s* as the proportion of simulations yielding the observed genotype data among simulations at that particular *s* value where *BOLA2B* was not lost divided by the maximum such proportion for any single value of *s*, considering all *s* values examined. We thus obtained a maximum likelihood estimate of the selection coefficient: $s = 0.0015$ (**Extended Data Fig. 8c**). Importantly, this is an estimate for the net value of *s*, (i.e., the sum of advantageous and deleterious effects). The positive effect of the *BOLA2B* duplication is likely even higher than suggested by this *s* value because it also has adverse consequences, conferring susceptibility to microdeletions and microduplications associated with disease.

Given that *BOLA2B* was present homozygously in all archaic humans examined, we also explored the strength of positive selection necessary to explain the observed genotype data if the period 282-45 kya were considered. For this analysis, we assumed our genotype data for contemporary humans instead corresponded to archaic San and Yoruban populations living ~45 kya. Performing simulations akin to those detailed above, we calculate a maximum likelihood estimate of $s = 0.0018$ under this scenario.

5.5 Modeling recurrent *BOLA2B* formation

Positive selection is not the only potential explanation for the rapid rise in *BOLA2B* frequency along the human lineage. Alternatively, recurrent duplicative transposition may have contributed to or even driven this evolutionary trajectory. To explore this possibility, we again used *msms*[18] to perform simulations under neutrality incorporating *BOLA2B* age and allowing for recurrent origins of *BOLA2B* along any lineage after its formation 282 kya along the human lineage. We initially explored a wide range of values for μ, the per-generation rate of recurrence, from $\mu = 1*10^{-3}$ to $\mu = 1*10^{-12}$, varying μ by orders of magnitude across this range and performing 10,000,000 simulations for each μ value. *BOLA2B* was present exclusively in human populations at frequencies as high as or higher than the observed frequencies only when μ was $1*10^{-4}$—at $\mu = 1*10^{-3}$, *BOLA2B* was always present on at least one simulated archaic hominin haplotype, and at lower recurrence rates, *BOLA2B* never reached the frequencies observed in humans.

Given these results, we settled on exploring a narrow range for the recurrence rate from $\mu = 0.00002$ to $\mu = 0.0005$, where we performed 10,000,000 simulations for each value of μ and varied μ by increments of 0.00002. We calculated the relative likelihood of the observed San and Yoruban *BOLA2B* genotypes for each value of μ as the proportion of simulations yielding the observed genotype data among simulations at that particular μ value where *BOLA2B* was present exclusively in humans divided by the maximum such proportion for any single value of μ, considering all μ values examined. We thus obtained a maximum likelihood estimate for the rate of recurrent *BOLA2B* formation most consistent with the *BOLA2B* frequency data under neutrality: $\mu = 0.00032$.

Notably, no recurrence rate examined resulted in more than 10,470 simulations (out of 10 million where *BOLA2B* was present on at least one human haplotype) with *BOLA2B* absent in Neanderthal and Denisova and present at observed or higher frequencies in humans ($p = 0.001047$, at $\mu = 0.00020$). This result indicates the observed genotype data or higher *BOLA2B* frequencies exclusively in humans are very unlikely to have arisen under neutrality even with recurrent *BOLA2B* formation. Finally, to account for uncertainty surrounding *BOLA2B* age, we performed simulations allowing recurrence and supposing *BOLA2B* arose 650 kya rather than 282 kya. Exploring the same narrow range of recurrence rates as above, the maximum likelihood value for the recurrence rate under this scenario was $\mu = 0.00012$. At most 61,107 of 10,000,000 simulations (at $\mu = 0.00006$) where *BOLA2B* was present on at least one human haplotype met the *BOLA2B* frequency criteria ($p = 0.0061107$), suggesting the conclusion that recurrence unlikely explains the observed data is robust to uncertainty in *BOLA2B* age.

5.6 Population genetic analyses of the chromosome 16p11.2 critical region

If the *BOLA2B* duplication were driven to high frequency through positive selection, it follows that linked genetic variation should show patterns consistent with this scenario. We specifically examined the chromosome 16p11.2 critical region, as this unique sequence between BP4 and BP5 is free from potential confounding factors associated with duplicated sequences such as interlocus gene conversion. First, we considered whether there is any evidence of Neanderthal or Denisovan introgression in the region, since if the region were selected together with *BOLA2B* after divergence from archaic hominins, introgressed alleles would likely have been lost. Second, we examined heterozygosity within the critical region, as reduction in heterozygosity is sometimes used as a metric to detect regions potentially under selection. Third, we examined Tajima's D, a statistic that tests an excess of rare variants, which is expected under scenarios of positive selection, population growth, or a combination of both.

To investigate archaic introgression, we examined recently published data from 503 Europeans, 504 East Asians, 489 South Asians and 35 Melanesians[19]. Introgression was reported over 50 kbp sliding genomic windows, with a step size of 10 kbp and various filters for callable regions applied as described[19]. In total, 236,780 bp within the critical region were callable for introgression. We observed introgressed haplotypes segregating at low frequencies at the chromosome 16p11.2 locus, but these never occurred over the critical region. Thus, although there are introgressed haplotypes at chromosome 16p11.2, the unique region flanked by *BOLA2* duplications corresponds to a *Homo sapiens*-specific haplotype. There is no evidence of any introgression in non-African populations across this region, an observation consistent with possible selection.

To examine heterozygosity, we used SNV data from 2,500 samples from the 1000 Genomes Project Phase 3 release[11]. We restricted our analysis to biallelic sites and computed heterozygosity at each SNV genome-wide as the number of called heterozygotes at the SNV divided by the number of callable genotypes. We then calculated the average heterozygosity of the ~550 kbp critical region, as well as ~550 kbp regions of unique sequence telomeric to BP1 and centromeric to BP5, as the sum of the heterozygosity values for all SNVs within the region under consideration divided by the total number of SNVs within the region. Next, we generated a distribution of average heterozygosity values for ~550 kbp regions genome-wide by analyzing 100,000 ~550 kbp regions drawn randomly from the autosomes (with replacement, excluding gaps and regions containing segmental duplications, where SNV genotype calls are often inaccurate). Comparing the ~550 kbp critical region to others genome-wide revealed that its average heterozygosity lies in the bottom 2.6% of the distribution (empirical $p = 0.02561$, **Extended Data Fig. 8d**). Even considering only ~550 kbp windows including 27 or more genes (as the region of interest includes 27 genes), the region of interest lies in the bottom 15th percentile, a striking observation given that most such windows contain more than 27 genes and thus based on gene content may be expected to show lower average heterozygosity. Remarkably, both regions flanking *BOLA2A* showed significantly reduced average heterozygosity (centromeric flanking sequence empirical $p = 0.0018$), whereas the region telomeric to BP1 exhibited average heterozygosity much closer to the genome-wide mean (empirical $p = 0.61158$). Thus, we conclude that the unique region between BP4 and BP5 has an average heterozygosity that is among the lowest compared with average heterozygosity values for similarly sized regions of the autosomal genome. This finding is consistent with potential selection.

Finally, to examine Tajima's D, we leveraged phased SNV genotypes from the same 2,500 individuals described above[11] together with VCFtools[57], a program that can compute Tajima's D from phased SNV data reported in standard variant call format. Specifically, we calculated Tajima's D for the ~550 kbp critical region, for the two ~550 kbp flanking regions described above, and for 2,987 non-overlapping ~550 kbp windows from the autosomes (excluding gaps and regions containing segmental duplications and including only windows with at least one polymorphic site). We found that the critical region falls in the bottom 2.7% of windows genome-wide (**Extended Data Fig. 8e**) and that the centromeric flanking

region is also a low outlier. These data, together with the introgression and heterozygosity analyses above, provide additional support for the possibility of selection having operated on the *BOLA2B* duplication.

6. *BOLA2* mRNA and protein characterization and expression

6.1 *BOLA2* RNA expression in human tissues and the discovery of *Homo sapiens*-specific fusion transcripts

We searched for transcripts mapping to *BOLA2* loci by analyzing annotated mRNA and expressed sequence tags (UCSC Genome Browser using the reference genome GRCh37). We identified two distinct sets of transcripts: a set consistent with the canonical *BOLA2* model and a set suggesting fusion transcripts between *BOLA2* and *SMG1P*. We designed two PCR assays to amplify both canonical and fusion *BOLA2* transcripts. Oligonucleotides for PCR amplification are provided below:

BOLA2_forward: TAGAGCAGGTAGACGCCGAAA
BOLA2_reverse: AATTTAATGGCTGTGCAGATCCC
BOLA2_fusion_forward: GAACAAGCTCTCGGGGACTATC
BOLA2_fusion_reverse: GTGATTCTGCAGACATGTTGACA

We prepared cDNA from total RNA (Clontech) from fetal brain (obtained from spontaneously aborted fetuses, ages 20-33 weeks), brain, cerebellum, heart, skeletal muscle, spleen, testis, and thymus using the Transcriptor High Fidelity cDNA Synthesis Kit (Roche) and oligo-dT primers, following the manufacturer's instructions (section 9.2). We then performed long-range PCR with each primer set on cDNA from each tissue using the Expand Long Template PCR System (Roche). All reactions yielded products of the expected sizes (**Extended Data fig. 9a-b**) except for the reaction including skeletal muscle cDNA and fusion *BOLA2* primers, indicating that canonical and fusion *BOLA2* isoforms are widely expressed.

We cloned and sequenced PCR products from reactions including brain cDNA. Specifically, we cloned products from both canonical and fusion *BOLA2* reactions into pCR Blunt II TOPO using the Zero Blunt TOPO PCR Cloning Kit (Thermo Fisher), following the manufacturer's instructions. We sequenced five clones from each reaction using capillary sequencing and M13 primers provided with the cloning kit. All sequenced products derived from *BOLA2* transcripts. The five products from the canonical *BOLA2* reaction matched the predicted *BOLA2* exon structure, including an open reading frame (ORF) spanning 456 bp predicted to encode a 152 amino acid (aa) protein. Sequenced products from the fusion transcript reveal two isoforms both including the first two exons from canonical *BOLA2*, skipping the final canonical exon, and including additional exons from *SMG1P* (**Fig. 3c**). One of these fusion isoforms (two sequenced products) terminates the *BOLA2* ORF shortly after the fusion junction (*BOLA2T*), such that if translated, only 5 aa would derive from SMG1P. Intriguingly, the other fusion isoform (three sequenced products) maintains the *BOLA2* ORF across the junction (*BOLA2F*), putatively encoding a predicted protein with 164 amino acids from SMG1P. Note that *SMG1P* is a paralog of *SMG1*—a phosphatidylinositol 3-kinase-related kinase involved in nonsense mediated decay and thought to be important for early embryogenesis[58].

To validate these transcripts and assess their expression in a wider panel of human tissues, we analyzed RNA-seq data from GTEx[59]. Specifically, we quantified expression for all three *BOLA2* transcripts in human tissues by running Sailfish[60] (version 0.6.3) with the default parameters and k = 20 on each of the GTEx RNA-seq samples (dbGaP version phs000424.v3.p1). We mapped reads to all GRCh38/hg38 RefSeq transcripts, replacing RefSeq transcripts for *BOLA2* with our three *BOLA2* transcript models, and quantified expression in units of transcripts per million (TPM, **Extended Data Fig. 9c**). These results

corroborate our RT-PCR data and confirm widespread mRNA expression for all *BOLA2* isoforms in human tissues, including muscle.

The genomic architecture juxtaposing *BOLA2* and *SMG1P* exists at two locations at chromosome 16p11.2: at the *BOLA2B* locus at BP4 and at junctions between tandem 102 kbp duplications at BP5 (**Extended Data Fig. 1b**). In both cases, this architecture evolved as a result of duplication events that occurred specifically in *Homo sapiens*. Thus, we predict that fusion *BOLA2* transcripts are *Homo sapiens*-specific. To test this hypothesis, we analyzed RNA-seq data from induced pluripotent stem cells (iPSCs) from human, chimpanzee, and bonobo[21]. We leveraged unique k-mers[12] distinguishing the exon 2-3 junction, which differs between each *BOLA2* isoform (**Fig. 3c**). For each isoform, we generated a list of all distinguishing 70-mers over the exon 2-3 junction including at least 15 bp from each exon. We quantified the relative expression of *BOLA2* isoforms in iPSCs from different species by assessing the reads assigned to each *BOLA2* isoform for each RNA-seq dataset.

We did not observe fusion isoforms in chimpanzee or bonobo, confirming the *Homo sapiens*-specific nature of fusion transcripts. No substitutions exist between human and chimpanzee or between human and bonobo over the last 55 bp of exon 2, so this result is not an artifact of generating our lists of 70-mers using human transcript sequence data. Read counts for different isoforms within each experiment provide insight into the relative expression of different isoforms. These data suggest that the expression of both fusion isoforms in human is approximately equal but ~3- to 10-fold lower than the expression of the canonical *BOLA2* transcript.

6.2 Correlation of *BOLA2* copy number with *BOLA2* RNA expression

We assessed whether the copy number of *BOLA2* is correlated with its expression at the mRNA level. We used genes designated as *BOLA2* (ENSG00000183336.7) and *BOLA2B* (ENSG00000169627.7) and expression quantifications[20] (PEER-normalized RPKM) from lymphoblastoid cell lines (LCLs) derived from 366 individuals having *BOLA2* copy number estimates (**Table S11**). Note that for this section and section 6.3, *BOLA2B* refers to the specific gene above and not to the *BOLA2* paralog at BP4. We computed Pearson's correlations between copy number and expression values using the R software environment[61]. We found significant correlations between *BOLA2* copy number and both *BOLA2* mRNA expression and *BOLA2B* mRNA expression (**Fig. 3a**), regardless of which method was used for genotyping copy number (for 330 WGS read-depth-based estimates, $R = 0.34$, $p = 1.7e\text{-}10$ and $R = 0.31$, $p = 1.2e\text{-}08$, for *BOLA2* and *BOLA2B*, respectively; for 366 MIP-based estimates, $R = 0.36$, $p = 2.1e\text{-}12$ and $R = 0.35$, $p = 3.5e\text{-}12$, for *BOLA2* and *BOLA2B*, respectively). Better correlations were observed with MIP-based copy number estimates, consistent with their generally higher accuracy compared to WGS read-depth-based estimates likely due to low sequencing coverage for 1000 Genomes Project samples[32] (section 4.2).

We next assessed whether the copy number change at the *BOLA2* ancestral centromeric locus (BP5) and at the human-specific telomeric locus (BP4) differently contribute to the expression variation. Using the R software environment[61], we compared two linear models that describe *BOLA2* and *BOLA2B* expression: i) accounting only for total *BOLA2* copy number (MIP-based estimates) and ii) accounting for the centromeric (BP5) and telomeric (BP4) *BOLA2* copy numbers (**Table S12**). The models suggest that the overall *BOLA2* copy number drives the variation of *BOLA2*/*2B* RNA expression and both the BP5 and BP4 copy numbers have an effect with a similar magnitude. As noted, the BP5 copy number variable is more statistically robust, likely because BP5 shows greater copy number polymorphism. We compared *BOLA2* copy number and expression within human population groups to control for potential differences in genetic background. The correlation is generally consistent in the different populations, with some, like the Toscani and British showing higher correlations than others such as Yorubans (data not shown).

6.3 Genome-wide correlation of *BOLA2* copy number with gene expression

We also assessed if the expression of other genes besides *BOLA2* and *BOLA2B* correlated with *BOLA2* WGS read-depth-based copy number in LCLs. We computed Pearson's correlations between *BOLA2* copy number and gene expression (Geuvadis dataset[20]) and corrected p-values for multiple testing using the Benjamini-Hochberg method. At a false discovery rate (FDR) <0.05, we identified *BOLA2*, *BOLA2B*, *SLX1A*, *SLX1B*, *SLX1A-SULT1A3*, *SULT1A1*, and *SULT1A2* as having RNA expression correlated with *BOLA2* copy number. Notably, these genes either map to the 102 kbp copy number variant segment at BP4 and BP5 or have paralogs mapping within this unit. This result is consistent with the 102 kbp unit of copy number variation described above (section 2.1) and indicates that the copy number change affects the expression level of genes embedded in the variable segment but does not more broadly perturb the transcriptome, at least in LCLs. We found no correlation between *BOLA2* copy number and the expression of the 29 genes mapping to the 550 kbp critical region[62]. Similar results were obtained using the MIP-based copy number estimates (data not shown).

6.4 BOLA2 protein definition and anti-BOLA2 antibody validation

As we found correlation between *BOLA2* copy number and expression at the RNA level, we explored if the former also correlated with protein level changes. We tested three different antibodies to detect BOLA2 protein: a goat polyclonal antibody raised against a peptide mapping near the C-terminus of human BOLA2 (Sc-163747, Santa Cruz Biotechnology), a mouse polyclonal antibody raised against the human full-length BOLA2 (amino acids 1-152, ab169481, Abcam), and a rabbit polyclonal antibody raised against the N-terminal amino acids 1-50 of mouse BOLA2 (ab105534, Abcam). We detected a consistent band at a molecular weight of 10 kDa with the three different antibodies in human LCL whole-protein lysates. None of these three antibodies reacted to a band at the predicted molecular weight of 17 kDa.

To further assess the specificity of the antibodies, we cloned and overexpressed the coding sequence (CDS) of the annotated 17 kDa BOLA2 in HeLa cells. We detected two bands at ~10 and 17 kDa much more abundant in transfected cells than in control non-transfected cells (**Extended Data Fig. 9e**). These results show that two forms can be translated from the *BOLA2* CDS and are recognized by the antibodies—a 17 kDa form corresponding to the entire predicted CDS and a shorter one that migrates at the size of the above-mentioned antigenic band. This indicates that the 10 kDa band observed in LCL lysate is specific and corresponds to a BOLA2 form shorter than the annotated 17 kDa, 152 residue protein.

To investigate the origin of the different proteins expressed from the *BOLA2* CDS, we predicted translation start sites (TSS) in the human *BOLA2* mRNA sequence using atgpr[63] and NetStart[64]. Both tools reported two possible TSS at position 514 and 712, corresponding to translated proteins of 152 (17 kDa) and 86 (10 kDa) residues, respectively. We retrieved available CAGE data (5' cap analysis gene expression)[65], ribosome profiling data (GWIPS-viz)[66], and tandem mass spectrometry (MS/MS) peptide data (PeptideAtlas)[67] for human *BOLA2*. CAGE data showed a main RNA 5' cap site in the first exon of *BOLA2*, suggesting that the annotated mRNA does not correspond to the most abundant transcript. The Ribo-seq profile of *BOLA2A* (BP5) and *BOLA2B* (BP4) showed few reads mapping to the sequence encoding for the N-terminal additional part of the hypothetical 17 kDa form. Finally, almost all peptides from MS/MS spectra mapped to the sequence common to the two forms (38 of 40 peptides), as we identified only two peptides derived from the additional N-terminal 66 residue segment of the 17 kDa protein, both from placenta. Four different peptides that begin with the start methionine of the 10 kDa BOLA2 and that could not be the products of trypsin digestion demonstrate that the methionine codon 67 (position 712) is likely used as a TSS, consistent with the predicted Kozak sequences.

Overall, these data strongly suggest that the bulk of human *BOLA2* transcription and TSS are different from current annotations. These results also demonstrate that the 10 kDa band observed in LCL lysates using the three antibodies corresponds to the 10 kDa primary protein form of BOLA2.

To further assess the possible functional relevance of the annotated 17 kDa BOLA2 form, we constructed a multiple sequence alignment including predicted BOLA2 protein sequences from human, chimpanzee, gorilla, orangutan, macaque, marmoset, mouse, rat, dog, and cow. The first methionine of the 10 kDa form is present in all mammals, whereas the potential first methionine of the 17 kDa form is present only in primates, with the exception of orangutan.

We analyzed the evolutionary conservation of the sequence encoding for the 10 kDa protein form in primates, as well as the conservation across primates of the additional N-terminal portion of the putative 17 kDa form. We leveraged our multiple sequence protein alignment and employed a maximum likelihood framework to model different evolutionary scenarios using PAML[68]. The likelihood ratio test was used to assess the significance of different values of omega between the single parameter model (where omega is free to vary but remains constant across all branches of the phylogenetic tree) and the neutral model (where omega is set to 1 for all branches). The single parameter model provides a significantly better fit to the data than the neutral model for the 10 kDa protein sequence, suggesting it is evolving under negative constraint (omega = 0.1899, p-value = 0.007). In contrast, the additional N-terminal portion seems to be neutrally evolving, as the best single parameter model (omega = 0.8975) does not provide a significantly better fit to the data than the neutral model (p-value = 0.48). This result suggests the 17 kDa BOLA2 form is most likely evolving neutrally, and even if present probably lacks functional importance.

6.5 *BOLA2* phylogeny

We constructed multiple sequence alignments for both *BOLA2* CDS and BOLA2 protein sequence from different mammals using Clustal Omega[69]. We leveraged the former alignment to build a maximum likelihood phylogenetic tree using MEGA6[70] (data not shown).

6.6 Correlation of *BOLA2* copy number with BOLA2 protein expression

We assessed whether *BOLA2* copy number correlates with its protein expression level (10 kDa band). We analyzed whole-protein lysates from a panel of 38 human LCLs that are part of HapMap and the 1000 Genomes Project (Coriell Institute) from individuals having BOLA2 copy number estimates (**Table S13**). All LCLs were from individuals of European ancestry, except NA18907 who is of Yoruban ancestry. The human samples, together with one chimpanzee and one orangutan LCL, were analyzed by Western blotting in three parallel SDS-PAGE gels (data not shown). We quantified the BOLA2 band (antibody from Santa Cruz Biotechnology) using densitometry (Bio1D software) and normalized using actin densities. After removing the variation in BOLA2 actin-normalized densities due to the gel factor, we analyzed the correlation of BOLA2 densities with the copy number estimates. Similar to the RNA expression, we detected significant Pearson's correlations of BOLA2 protein level with copy number (**Fig. 3b**), suggesting that *BOLA2* copy number variation affects the quantity of BOLA2 protein in the cell. In chimpanzee and especially in orangutan LCL lysates, we detected a lower amount of BOLA2 compared to human cell lines, regardless of human *BOLA2* copy number.

To better understand the contribution of the copy number change at the centromeric and telomeric sides to the BOLA2 protein expression level, we computed and compared four different linear models that respectively describe BOLA2 protein expression level based on: i) the total *BOLA2* copy number (MIP-based estimates); ii) the centromeric (BP5) copy number; iii) the telomeric (BP4) copy number; and iv) both the centromeric and telomeric copy numbers as two independent variables (**Table S14**). Different from the RNA expression, the centromeric copy number variation explains 41% of the protein expression

variation and the telomeric copy number does not contribute significantly to expression variation. Comparing these models with those that describe the RNA expression level, we note that the lack of contribution from the less variable telomeric (BP4) copy number might relate to a power issue, having here far fewer samples with telomeric copy number differing from 2 than in the RNA dataset.

6.7 Ribosome profiling analysis

We analyzed previously published ribosome profiling datasets generated from the HEK293 human embryonic kidney cell line[71] and from LCLs from 30 individuals: five Europeans, two Asians, and 23 Yorubans[72]. We again leveraged unique kmers[12] distinguishing the exon 2-3 junctions of *BOLA2*, *BOLA2F*, and *BOLA2T* to count reads unambiguously corresponding to each isoform from each sequencing experiment. Specifically, we queried the sequencing data for reads containing the following kmers, each including 9 bp from exon 2 (shared between all isoforms) and 8 bp from exon 3 (unique to each isoform): GAGACACAGGCTGGTGA (*BOLA2*), GAGACACAGCTTGGATC (*BOLA2F*), and GAGACACAGGTTCTGTA (*BOLA2T*).

We identified at least one read spanning the exon 2-3 junction from *BOLA2* in all 51 independent ribosome profiling experiments, with 45 experiments having at least 10 reads derived from *BOLA2*. Despite the lower mRNA expression of *BOLA2F* and *BOLA2T* compared to *BOLA2*, we also identified at least one read corresponding to these isoforms in 24 and 45 experiments, respectively, with at least 10 reads from *BOLA2F* observed in 1 experiment and at least 10 reads from *BOLA2T* in 9 experiments (**Table S15**). Although the mere presence of ribosome profiling reads along a transcript does not definitively imply its translation[71,73], the majority of such reads signify active ribosomes[74]. Thus, detecting reads corresponding to *BOLA2F* and *BOLA2T* in ribosome profiling data provides evidence supporting their translation.

6.8 Chimpanzee and human iPSC and RNA sequencing analysis

6.8.1 Overview

Previous studies reported differences in the expression of *BOLA2* between human and nonhuman primate iPSCs[21]. In light of our new findings, we aimed to quantify the levels of different *BOLA2* isoforms in human, chimpanzee and bonobo iPSCs. We reanalyzed previously described RNA-seq data[21] generated from human embryonic stem cells (ESCs) HUES6 and H9; human iPSC lines WT-33, ADRC-40, WT-126 and WT9, chimpanzee iPSC lines PR00818 and PR01209, and bonobo iPSC lines AG05253 and PR01086. Moreover, to investigate the expression levels of *BOLA2* during neuronal development, we differentiated human and chimpanzee iPSCs into neural progenitor cells (NPCs) and neurons. We designed a specific bioinformatics analysis to quantify the expression of different *BOLA2* isoforms from RNA-seq.

6.8.2 Cell lines

Human ESC line H9 is from Wisconsin (WiCell Research Institute, Inc.). According to Thomson *et al.*[75], "Fresh or frozen cleavage stage human embryos, produced by *in vitro* fertilization (IVF) for clinical purposes, were donated by individuals after informed consent and after institutional review board approval. Embryos were cultured to the blastocyst stage, 14 inner cell masses were isolated, and five ES cell lines originating from five separate embryos were derived, essentially as described for nonhuman primate ES cells." Human ESC line HUES6 is from Harvard (Harvard Stem Cell Institute) and described by Cowan *et al.*[76]. This line was obtained from frozen cleavage- and blastocyst-stage human embryos, produced by IVF for clinical purposes, after obtaining written informed consent and approval by a Harvard institutional review board. Human iPSC lines (WT-33, ADRC-40, WT-126 and WT9) and chimpanzee iPS cell lines (PR00818 and PR01209) have been previously described[21,77].

6.8.3 Cell culture and neuronal differentiation

Established iPSC colonies were kept in feeder-free conditions and passed using mechanical dissociation. To obtain NPCs from human and chimpanzee iPSCs, embryoid bodies (EBs) were formed by mechanical dissociation of iPSC clusters and plated into low-adherence dishes in DMEM/F12 plus N2 and B27 (Invitrogen) medium in the presence of Noggin (R&D) for forebrain induction for approximately 7 days. Then, floating EBs were plated onto poly-ornithine/laminin (Sigma)-coated dishes in DMEM/F12 plus N2 and B27 (Invitrogen) with addition of Noggin. Rosettes were visible to collect after 7 days. Rosettes were then dissociated with accutase (Chemicon) and plated again onto coated dishes in DMEM/F12 plus N2 and B27 and 10ng/ml of FGF2 (R&D). Homogeneous populations of NPCs were achieved after 1-2 passages with accutase in the same conditions. To obtain mature neurons, NPCs were cultured with DMEM/F12 plus N2 and B27 with addition of 1ug/ml of Laminin, BDNF (20 ng/ml), GDNF (20 ng/ml) and cyclic AMP (500 ug/ml) for 8 weeks. The full transcriptomic and functional characterization of primate NPCs and neurons will be described elsewhere (Marchetto *et al*., manuscript in preparation). The use of chimpanzee and bonobo fibroblasts was approved by the US Fish and Wildlife Service, under permit MA206206. Protocols describing the use of iPSCs and human ESCs were previously approved by the University of California, San Diego (UCSD), the Salk Institute Institutional Review Board, and the Embryonic Stem Cell Research Oversight Committee[21].

6.8.4 RNA extraction, RNA libraries, deep sequencing, and data analysis

For RNA library generation and deep sequencing, total cellular RNA was extracted from ~1-5 x10^6 cells using the RNeasy Protect Mini kit or RNeasy Plus kit (Qiagen). RNA-seq datasets derived from iPSCs were previously described[21]. For RNA library generation from human and chimpanzee NPCs and eight-week-old neurons, PolyA^+ RNA was fragmented and prepared into sequencing libraries using the Illumina TruSeq RNA sample preparation kit. NPC-derived sequencing libraries were analyzed on an Illumina HiSeq 2000 sequencer at the UCSD Biomedical Genomics Laboratory (BIOGEM). cDNA libraries were prepared from two human and two chimpanzee NPC lines (two clones each) derived from human WT-33 and ADRC-40 iPSC lines and chimpanzee PR00818 and PR01209 iPSC lines, respectively. Libraries were sequenced using single-end 100 bp reads at a depth of 15–30 million reads per library. Sequencing libraries derived from eight-week-old neurons were analyzed on an Illumina HiSeq 2500 sequencer at the Salk Next Generation Sequencing Core. cDNA libraries were prepared from two human (WT-33 and ADRC-40) and two chimpanzee (PR00818 and PR01209) neurons, one clone each. Libraries were sequenced using paired-end 125 bp reads at a depth of 15–30 million reads per library.

Gene expression was calculated in TPM with Kallisto[36] (version 0.42.1) against a custom catalog of human transcripts, including all human RefSeq transcripts with the three *BOLA2* isoforms (**Table S16** and **Fig. 3d**). RefSeq isoforms nearly identical to the canonical *BOLA2* isoform (NM_001039182 and NM_001031827) were not included in the catalog of transcripts. Since the RNA-seq datasets are a mix of PE100 and SE100 reads, we quantified gene expression by using only the first read of PE100 sequencing.

For visualization and quantification of reads spanning *BOLA2* exon junctions in iPSCs, we generated a Sashimi plot[78] showing RNA-seq reads that aligned to the *BOLA2* locus using the Integrative Genomics Viewer (**Fig. 3c**). Reads from human iPSC lines WT-33 and ADRC-40 and chimpanzee iPSC lines PR00818 and PR01209 (two clones each) were mapped to the human (GRCh37) or chimpanzee (panTro4, CGSC 2.1.4) reference genomes using STAR with default parameters (version 2.2.0.c)[79]. All reads from human samples were mapped to the telomeric (BP4) *BOLA2-SMG1P* duplication in the human reference genome; chimpanzee reads were mapped to the chimpanzee reference genome.

6.9 *BOLA2* RNA expression in human, chimpanzee, and bonobo across a panel of tissues

To investigate whether *BOLA2* is differentially expressed between humans, chimpanzees, and bonobos more broadly, we quantified RNA expression in TPM using Kallisto[36] (see section 6.8.4 for details) and published RNA-seq data[22] from six adult tissues: kidney, liver, testis, brain, cerebellum, and heart (**Extended Data Fig. 9d**). Because this dataset consisted of sequenced fragments size-selected to be ~250-300 bp, quantification was performed using the arguments '-l 275' and '-s 35' (corresponding to the fragment length and standard deviation, respectively).

7. Susceptibility to recurrent chromosome 16p11.2 rearrangements

The organization (section 2.1) and high identity (**Extended Data Fig. 6a**) of *Homo sapiens*-specific duplicated sequences including *BOLA2* implicate them in predisposing chromosome 16p11.2 to recurrent rearrangements in humans associated with disease. We compared all directly oriented segmental duplications flanking the chromosome 16p11.2 autism critical region in human, chimpanzee, and orangutan to determine whether this susceptibility is specific to our species.

We identified directly oriented duplicated sequences flanking the autism critical region (**Table S4**) using a modified WGAC pipeline described above (section 2.1). We find that orangutan lacks directly oriented segmental duplications flanking the autism critical region, while chimpanzee possesses only small blocks of such duplicated sequence, no more than 50 kbp in size, having at most 98.6% average sequence identity (**Extended Data Fig. 5a**). In contrast, human haplotypes have directly oriented duplication blocks flanking the autism critical region that are at least 117 kbp in size and exhibit at least 98.8% average sequence identity. Restricting our analysis to human haplotypes having a duplicate *BOLA2B* paralog, the blocks of interest are at least 147 kbp in size and show at least 99.3% average sequence identity.

Long, identical stretches of sequence shared between duplications promote NAHR[102,103]. Such regions are abundant at BP4 and BP5 for human haplotypes having *BOLA2* copies at both loci (**Extended Data Fig. 6a**). For each contig sequence, we identified all tracts of perfect sequence identity at least 500 bp in size within the longest contiguous region of homology between directly oriented segmental duplications flanking the autism critical region (**Table S17** and **Extended Data Fig. 5b**). We selected 500 bp as a threshold since it appears to represent a minimal length for efficient processing for mammalian recombination machinery[80]. Neither orangutan nor chimpanzee possess any tracts meeting the criteria above, while such tracts were found in all human haplotypes, with the highest number and longest such tracts occurring in haplotypes including *BOLA2B* (**Table S18**). Long stretches of perfect sequence identity are exclusive to humans and most prevalent on haplotypes containing the *Homo sapiens*-specific *BOLA2* duplication at BP4. These findings corroborate the conclusion that the predisposition to recurrent, disease-associated rearrangements at chromosome 16p11.2 is specific to our species.

8. Microdeletion/microduplication breakpoint refinement

8.1 Overview

Chromosome 16p11.2 microdeletions and microduplications associated with autism and developmental delay arise via NAHR between directly oriented segmental duplications at BP4 and BP5 (**Extended Data Fig. 5c**). To evaluate the potential role of the *Homo sapiens*-specific duplication containing *BOLA2* in promoting instability at this locus, we localized breakpoints for 152 patients carrying a typical BP4-BP5 chromosome 16p11.2 rearrangement, corresponding to 72 independent microdeletions and 33

independent microduplications (**Table S19**). We utilized three methods to refine breakpoint locations: i) examination of WGS read depth at unique 30-mer locations in the human reference genome (GRCh37), ii) visualization of marker-specific WGS read count relative frequencies at positions informative for breakpoint mapping, and iii) analysis of marker-specific read count relative frequencies from sequencing data generated using a MIP assay targeting breakpoint-informative sites. We briefly detail each of these approaches below and show that, except for a few cases, they resolve breakpoints as mapping within the ~95 kbp interval corresponding to the *Homo sapiens*-specific duplication from BP5 to BP4.

8.2 Breakpoint refinement using normalized WGS read depth

We generated whole-genome shotgun sequence from three trios and three quads (21 genomes total), each including an initially identified proband having a *de novo* chromosome 16p11.2 BP4-BP5 microdeletion. Each genome was sequenced to an average coverage of at least 20-fold using the Illumina HiSeq platform, the Illumina NextSeq platform, or a combination of the two. Each sequence read was decomposed into 30-mers and mapped to the human reference genome GRCh37 using mrsFAST as previously described[12,38]. We generated copy number variation heat maps showing aggregate and PSCN across the chromosome 16p11.2 locus. Additionally, we computed read depth at all positions corresponding to unique 30-mer sequences in the reference genome GRCh37, normalized read-depth values based on overall genome sequence coverage, and visualized the normalized data using custom tracks uploaded to the UCSC Genome Browser (**Fig. 4a**) as previously described[12,38].

All six families showed the same pattern. Normalized read depth was about half of that observed in parents between the *Homo sapiens*-specific duplicated sequences but equal in probands and parents beyond the *Homo sapiens*-specific duplicated sequences. These observations refine all chromosome 16p11.2 microdeletion breakpoints examined using this method to an ~95 kbp interval containing *BOLA2* (**Table S19**).

8.3 Breakpoint refinement using marker-specific WGS read count frequencies

In a second approach, we utilized WGS data together with genetic markers identified for PSCN genotyping (section 4.3) to detect breakpoint signatures. Specifically, we expect a pattern of a reciprocal marker-specific copy number shift at the location of unequal crossover[24,37]. In this scenario, the BP4-BP5 recombinant duplication block formed by NAHR would include markers unique to BP4 before the breakpoint and markers unique to BP5 after the breakpoint. On the other hand, NAHR between sequences at BP4 and BP5 not distinguishable using our PSCN markers, i.e., NAHR between the *Homo sapiens*-specific duplication segments, should not produce a detectable signature. In this case, the BP4-BP5 recombinant would contain the same markers across its entirety as both of the original duplication blocks at BP4 and BP5 from which it derived (**Extended Data Fig. 10a**).

For each sequenced genome, we plotted marker-specific read count frequencies at each PSCN marker site (**Extended Data Fig. 10b**), performing the same analysis described in section 4.3. In no cases did we detect a reciprocal marker-specific copy number transition as would be expected if microdeletion breakpoints occurred outside of the *Homo sapiens*-specific duplication. Thus, these data corroborate the above results (section 8.2) that in all seven microdeletion patients (six independent rearrangements), breakpoints map within the ~95 kbp *Homo sapiens*-specific duplication (**Table S19**).

8.4 Breakpoint refinement using a MIP assay

We repurposed our paralog-specific *BOLA2* copy number MIP assay to refine microdeletion and microduplication breakpoint locations in a total of 152 individuals corresponding to 105 independent rearrangement events (**Table S19**). Specifically, we used the same MIP pool as in section 4.4 (**Table S10**), including the 47 MIPs described therein targeting markers within the 72 kbp block as well as 54 MIPs targeting markers across the 45 kbp block. These latter markers enable detection of breakpoints

occurring within the 45 kbp block. We performed MIP sequencing and analysis as above (section 4.4), except we mapped sequencing data to that minimal genome augmented with all 45 kbp blocks from our haplotype contigs and blocks of paralogy throughout the human genome (GRCh37) and developed an automated approach[24] to genotype paralog-specific copy number across the 45 kbp block. We plotted marker-specific read count frequencies (section 8.3) to determine whether breakpoints for each independent rearrangement map within or outside of the *Homo sapiens*-specific duplication and to define as precisely as possible intervals within which breakpoints occurred.

Breakpoints for at least 101 of 105 rearrangement events localize to the *Homo sapiens*-specific duplication (**Fig. 4b** and **Extended Data Fig. 10c**). In two cases, the breakpoints cannot be unambiguously resolved. In the remaining two, microdeletion breakpoints map outside the *Homo sapiens*-specific duplication, instead falling within the 45 kbp block. For these two, we narrow the putative breakpoint intervals to a 1.6 kbp region and a 22 kbp region (**Extended Data Fig. 10d**) within this block. Marker-specific read count frequency data over the 45 kbp block for these individuals indicate reciprocal transitions in BP4/BP5 PSCN. The signatures are consistent with these individuals having a total of three 45 kbp blocks at chromosome 16p11.2: two from the unaffected haplotype having BP4 or BP5 markers across their entire lengths, and one unequal crossover recombinant from the microdeletion haplotype. This recombinant has BP4 markers at its start and BP5 markers at its end. Interestingly, marker-specific read count frequency data for the sibling of one of these individuals (**Extended Data Fig. 10d**) reveal an ~22 kbp interval within the 45 kbp block showing both an increase in BP5 marker copy number and a decrease in BP4 marker copy number. We conclude that this region likely corresponds to an interlocus gene conversion event between BP5 and BP4, with the former serving as the conversion donor. This event was observed specifically in this family and was inferred to be present in the germline of the father (DNA not available for testing) based on its absence in the mother. Note that the start of the conversion region in the sibling maps to the same location as the PSCN transition in the proband. This interlocus gene conversion, thus, created a high sequence identity interval within the 45 kbp block predisposing this region to unequal crossover between BP4 and BP5 in this family.

Marker-specific read count frequency data for the remaining two rearrangements showed signatures consistent with breakpoints mapping within the 72 kbp block but outside of the *Homo sapiens*-specific duplication. However, these signatures were the same between patients from the two different families of interest and matched an interlocus gene conversion signature observed in some individuals lacking any chromosome 16p11.2 rearrangement. Thus, in these cases, the data are consistent with either atypical breakpoints or breakpoints within the *Homo sapiens*-specific duplication together with 72 kbp blocks affected by interlocus gene conversion.

9. Additional methods and analyses

9.1 Fluorescence *in situ* hybridization

FISH experiments (**Table S2**) were used to assay aggregate *BOLA2* copy number variation (**Extended Data Fig. 7b**), to show that such variation affects both BP4 and BP5 (**Extended Data Fig. 7a**), to compare chromosome 16p11.2 organization between human, chimpanzee, and orangutan (**Extended Data Fig. 2**), and to assess the duplication from chromosome 16q24.2 in human, chimpanzee, gorilla, and orangutan (data not shown).

Metaphase spreads were obtained from lymphoblast and fibroblast cell lines from human HapMap individuals HG01067, HG02314, NA12275, NA12878, NA19041, NA19091, and NA20127, as well as from chimpanzee (PTR5), gorilla (GGO5), and orangutan (PPY10, PPY13, and PPY16; Coriell Cell Repository, Camden, NJ). FISH experiments were performed using human fosmid clones (**Table S2**)

identified based on mapping clone paired-end sequence data[40] to the human reference genome GRCh37. Clones were directly labeled with Cy3-dUTP, Cy5-dUTP (GE Healthcare), or Fluorescein-dUTP (Invitrogen) by nick translation as previously described[38], with minor modifications. Two hundred ng of labeled probe were hybridized on metaphase spreads; hybridization was performed overnight at 37°C in 2xSSC, 50% (v/v) formamide, 10% (w/v) dextran sulphate, 5 mg COT1 DNA (Roche), and 3 mg sonicated salmon sperm DNA, in a volume of 10 mL. Post-hybridization washing was performed at 60°C in 0.1xSSC (three times, high stringency). Washes for interspecies hybridization experiments were performed at lower stringency: 37°C in 2xSSC, 50% formamide, followed by washes at 42°C in 2xSSC. 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.

9.2 RT-PCR

We amplified *BOLA2* transcripts using total RNA (Clontech) from a variety of human tissues (section 6.1). To prepare cDNA libraries, we used the Transcriptor High Fidelity cDNA Synthesis Kit (Roche). For each RNA sample, a template-primer mixture was prepared by combining 9.07 µL PCR-quality water, 1.33 µL total RNA (at a concentration at least 1 µg/µL), and 1 µL oligo dT primer. RNA secondary structures were denatured by incubating this mixture for 10 mins at 65°C, and afterwards the mixture was immediately cooled on ice. An RT-PCR master mix was prepared by combining 72 µL RT reaction buffer, 9 µL RNase inhibitor, 36 µL dNTP mix (10 mM), 18 µL DTT, and 19.8 µL reverse transcriptase. 8.6 µL of this master mix was added to each template-primer mixture. Reverse transcription was performed by incubating this 20 µL reaction for 30 min at 50°C, followed by a 5 min incubation at 85°C and cooling on ice.

Using the Expand Long Template PCR System (Roche), we prepared two PCR master mixes, one including primers targeting the canonical *BOLA2* isoform and one including primers targeting fusion isoforms. Each master mix contained 50 µL 10x buffer 1, 17.5 µL dNTP mix (10 mM), 15 µL forward primer (10 µM), 15 µL reverse primer (10 µM), 355 µL PCR-quality water, and 7.5 µL enzyme mix. For each cDNA library, we set up a PCR reaction by combining 46 µL of the master mix with 4 µL of cDNA. Reactions were incubated at 94°C for 2 min, followed by 10 cycles at 94°C for 10 s, 55°C for 30 s, and 68°C for 3 min, followed by 25 cycles at 94°C for 10 s, 55°C for 30 s, and 68°C for 3 min + 20 additional seconds each cycle, followed by 68°C for 7 min and finally 4°C indefinitely.

9.3 Western blotting

Human LCLs were grown in RPMI-1640 medium (Gibco) supplemented with 15% fetal bovine serum and 1% antibiotics (penicillin and streptomycin). Cells were lysed in RIPA buffer (Millipore) supplemented with protease inhibitors. Protein lysates were run on an SDS-PAGE gel and transferred to a PVDF membrane. After blocking the membrane in PBS-T 0.05%, gelatin 1%, the primary antibodies were incubated overnight at 4°C. The membrane was then washed, incubated with the appropriate HRP-conjugated secondary antibody for 1 h at RT, washed and revealed.

9.4 CMV transient expression in HeLa cells

HeLa cells were grown in DMEM (Gibco) supplemented with 10% fetal bovine serum and 1% antibiotics (penicillin and streptomycin). Gateway PLUS shuttle clone for human *BOLA2* corresponding to the annotated 17 kDa protein CDS was obtained from tebu-bio (GC-Z3591), moved to pCS2plus destination vector by LR reaction for CMV transient and constitutive expression in cells, and transfected to HeLa cells using FuGENE reagent (Promega) in medium without antibiotics. After 24 h, cells were lysed in RIPA buffer and whole lysates were analyzed by Western blotting.

9.5 *BOLA2* 10 kDa Gateway cloning

The BOLA2 10 kDa CDS was amplified using attB sites flanked primers and *Pfu* DNA polymerase (Promega). The gel-purified PCR product was cloned in the pDONR221 vector (Invitrogen) through the BP reaction (Invitrogen). *BOLA2* CDS was moved to the pCS2plus destination vector for CMV constitutive expression through the LR reaction (Invitrogen).

9.6 Inversion density analysis

We evaluated the probability of large-scale inversions clustering in genomic space with a permutation test using 27 previously described inversions found in the human or chimpanzee lineages[81]. We first measured the density of observed inversions based on the distance between pairs of inversions on the same chromosome. To minimize error in pairwise comparisons caused by coarse breakpoint resolution, we measured the distance between inversion midpoints as calculated by the minimum start coordinate plus half of the range between start and end coordinates. Inversions that occurred alone on a chromosome were necessarily omitted from the distance calculation leaving 12 of the 27 original inversions. After visual confirmation that the distribution of observed distances was not normal, we selected the median as a summary statistic to compare observed and null distributions. To generate a null distribution for inversion distances, we randomly shuffled the coordinates for the full extent of all 27 published inversions across all GRCh37 genomic space for 1,000,000 iterations and calculated the median distance between midpoints for each iteration. We found that the observed median distance between previously published inversion midpoints (9.1 Mbp) was significantly smaller than expected based on the null distribution ($p = 0.003262$), which had a median of 29 Mbp. Most strikingly, null distribution median distances were never as small as or smaller than the observed 216 kbp median distance between chromosome 16p11.2 inversion midpoints ($p < 0.000001$).

Inversion breakpoints are typically associated with segmental duplications, telomeres, and centromeres. To test whether the density of inversions we observed in human and chimpanzee is significant with respect to the genomic space of these associated sequences, we ran an additional permutation test for 1,000,000 iterations where inversion breakpoints were shuffled across segmental duplications, telomeres (150 kbp from chromosome ends), and centromeres (5 Mbp on either side of annotated centromeres in GRCh37) to create the null distribution. We found that the observed median distance between inversions (9.1 Mbp) was significantly smaller than expected based on the null distribution ($p = 0.005602$) which had a median of ~27 Mbp. Again, null distribution median distances were never as small as or smaller than the observed 216 kbp median distance between chromosome 16p11.2 inversion midpoints ($p < 0.000001$).

9.7 Comparison of human reference genomes GRCh37 and GRCh38 over the chromosome 16p11.2 region and analysis of reference sequence accuracy

To assess potential impacts of using an older version of the human reference genome (GRCh37) for many of our analyses, we systematically compared the chromosome 16p11.2 region between GRCh37 (chr16:28,195,661-30,573,128) and the current human reference genome GRCh38 (chr16:28,184,340-30,561,807). These sequences were identical except for three single-nucleotide changes within unique sequence in the chromosome 16p11.2 critical region: GRCh37 chr16:29,791,561 T → GRCh38 chr16:29,780,240 G; GRCh37 chr16:29,905,677 T → GRCh38 chr16:29,894,356 C; and GRCh37 chr16:29,931,065 T → GRCh38 chr16:29,919,744 C. Thus, using GRCh37 rather than the current GRCh38 for several analyses did not affect our results and conclusions.

As part of our efforts to generate contiguous human haplotype sequences over the chromosome 16p11.2 region and in light of the extensive copy number variation within BP4 and BP5, we carefully characterized the accuracy of the human reference assemblies GRCh37 and GRCh38. Although the structures of BP4 and BP5 in these reference assemblies match those in our H1 contig, we found that both

reference sequences contain clones at BP5 incorrectly assembled, originating from two different RP11 haplotypes. RP11-455F5 (coming from the same haplotype as our $H4_C$ contig and corresponding to GRCh37 chr16:30100753-30202572 and GRCh38 chr16:30089432-30191251) and RP11-347C12 (coming from the other RP11 haplotype containing a tandem duplication of the 102 kbp variable segment at BP5 and corresponding to GRCh37 chr16:30202573-30383280 and GRCh38 chr16:30191252-30371959) were assembled to form the BP5 sequence in these reference genomes. We anticipate future human reference assemblies will improve this region by incorporating our high-quality, contiguous sequence data.

Supplementary References

40 Kidd, J. M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56-64, doi:10.1038/nature06862 (2008).

41 Dennis, M. Y. *et al.* Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication. *Cell* **149**, 912-922, doi:10.1016/j.cell.2012.03.033 (2012).

42 Adey, A. *et al.* Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome biology* **11**, R119, doi:10.1186/gb-2010-11-12-r119 (2010).

43 Chaisson, M. J. *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608-611, doi:10.1038/nature13907 (2015).

44 Jiang, Z., Hubley, R., Smit, A. & Eichler, E. E. DupMasker: a tool for annotating primate segmental duplications. *Genome research* **18**, 1362-1368, doi:10.1101/gr.078477.108 (2008).

45 Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J. & Eichler, E. E. Segmental duplications: organization and impact within the current human genome project assembly. *Genome research* **11**, 1005-1017, doi:10.1101/gr.187101 (2001).

46 Eichler, E. E. & Zimmerman, A. W. A hot spot of genetic instability in autism. *The New England journal of medicine* **358**, 737-739, doi:10.1056/NEJMe0708756 (2008).

47 Parsons, J. D. Miropeats: graphical DNA sequence comparisons. *Computer applications in the biosciences : CABIOS* **11**, 615-619 (1995).

48 Thompson, J. D., Gibson, T. J. & Higgins, D. G. Multiple sequence alignment using ClustalW and ClustalX. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* **Chapter 2**, Unit 2.3, doi:10.1002/0471250953.bi0203s00 (2002).

49 Jiang, Z. *et al.* Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nature genetics* **39**, 1361-1368, doi:10.1038/ng.2007.9 (2007).

50 Johnson, M. E. *et al.* Recurrent duplication-driven transposition of DNA during hominoid evolution. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 17626-17631, doi:10.1073/pnas.0605426103 (2006).

51 Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M. & Barton, G. J. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics (Oxford, England)* **25**, 1189-1191, doi:10.1093/bioinformatics/btp033 (2009).

52 Tamura, K. *et al.* MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular biology and evolution* **28**, 2731-2739, doi:10.1093/molbev/msr121 (2011).

53 Gonzalez, J. R. *et al.* A common 16p11.2 inversion underlies the joint susceptibility to asthma and obesity. *American journal of human genetics* **94**, 361-372, doi:10.1016/j.ajhg.2014.01.015 (2014).

54 Martin, J. *et al.* The sequence and analysis of duplication-rich human chromosome 16. *Nature* **432**, 988-994, doi:10.1038/nature03187 (2004).

55 Illumina, I. *Platinum Genomes*, <http://www.illumina.com/platinumgenomes/> (

56 Center, T. N. Y. G. *Unpublished data*.

57 Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics (Oxford, England)* **27**, 2156-2158, doi:10.1093/bioinformatics/btr330 (2011).

58 McIlwain, D. R. *et al.* Smg1 is required for embryogenesis and regulates diverse genes via alternative splicing coupled to nonsense-mediated mRNA decay. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 12186-12191, doi:10.1073/pnas.1007336107 (2010).

59 Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science (New York, N.Y.)* **348**, 648-660, doi:10.1126/science.1262110 (2015).
60 Patro, R., Mount, S. M. & Kingsford, C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nature biotechnology* **32**, 462-464, doi:10.1038/nbt.2862 (2014).
61 Computing, R. F. f. S. *R: A Language and Environment for Statistical Computing*, <http://www.R-project.org/> (2014).
62 Migliavacca, E. *et al.* A Potential Contributory Role for Ciliary Dysfunction in the 16p11.2 600 kb BP4-BP5 Pathology. *American journal of human genetics* **96**, 784-796, doi:10.1016/j.ajhg.2015.04.002 (2015).
63 Salamov, A. A., Nishikawa, T. & Swindells, M. B. Assessing protein coding region integrity in cDNA sequencing projects. *Bioinformatics (Oxford, England)* **14**, 384-390 (1998).
64 Pedersen, A. G. & Nielsen, H. Neural network prediction of translation initiation sites in eukaryotes: perspectives for EST and genome analysis. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology* **5**, 226-233 (1997).
65 Kodzius, R. *et al.* CAGE: cap analysis of gene expression. *Nature methods* **3**, 211-222, doi:10.1038/nmeth0306-211 (2006).
66 Michel, A. M. *et al.* GWIPS-viz: development of a ribo-seq genome browser. *Nucleic acids research* **42**, D859-864, doi:10.1093/nar/gkt1035 (2014).
67 Desiere, F. *et al.* The PeptideAtlas project. *Nucleic acids research* **34**, D655-658, doi:10.1093/nar/gkj040 (2006).
68 Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution* **24**, 1586-1591, doi:10.1093/molbev/msm088 (2007).
69 Sievers, F. & Higgins, D. G. Clustal omega. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* **48**, 3.13.11-13.13.16, doi:10.1002/0471250953.bi0313s48 (2014).
70 Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular biology and evolution* **30**, 2725-2729, doi:10.1093/molbev/mst197 (2013).
71 Calviello, L. *et al.* Detecting actively translated open reading frames in ribosome profiling data. *Nature methods* **13**, 165-170, doi:10.1038/nmeth.3688 (2016).
72 Cenik, C. *et al.* Integrative analysis of RNA, translation, and protein levels reveals distinct regulatory variation across humans. *Genome research* **25**, 1610-1621, doi:10.1101/gr.193342.115 (2015).
73 Guttman, M., Russell, P., Ingolia, N. T., Weissman, J. S. & Lander, E. S. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell* **154**, 240-251, doi:10.1016/j.cell.2013.06.009 (2013).
74 Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science (New York, N.Y.)* **324**, 218-223, doi:10.1126/science.1168978 (2009).
75 Thomson, J. A. *et al.* Embryonic stem cell lines derived from human blastocysts. *Science (New York, N.Y.)* **282**, 1145-1147 (1998).
76 Cowan, C. A. *et al.* Derivation of embryonic stem-cell lines from human blastocysts. *The New England journal of medicine* **350**, 1353-1356, doi:10.1056/NEJMsr040330 (2004).
77 Marchetto, M. C. *et al.* A model for neural development and treatment of Rett syndrome using human induced pluripotent stem cells. *Cell* **143**, 527-539, doi:10.1016/j.cell.2010.10.016 (2010).
78 Katz, Y. *et al.* Sashimi plots: Quantitative visualization of RNA sequencing read alignments. 1306.3466 (2013).

79 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).
80 Mundia, M. M., Desai, V., Magwood, A. C. & Baker, M. D. Nascent DNA synthesis during homologous recombination is synergistically promoted by the rad51 recombinase and DNA homology. *Genetics* **197**, 107-119, doi:10.1534/genetics.114.161455 (2014).
81 Antonacci, F. & Ventura, M. Personal communication. (2015).