Letter

Sequencing human–gibbon breakpoints of synteny reveals mosaic new insertions at rearrangement sites

Santhosh Girirajan,[1,4] Lin Chen,[1,4] Tina Graves,[2] Tomas Marques-Bonet,[1] Mario Ventura,[3] Catrina Fronick,[2] Lucinda Fulton,[2] Mariano Rocchi,[3] Robert S. Fulton,[2] Richard K. Wilson,[2] Elaine R. Mardis,[2] and Evan E. Eichler[1,5]

[1]*Department of Genome Sciences, Howard Hughes Medical Institute, University of Washington School of Medicine, Seattle, Washington 98195, USA;* [2]*Genome Sequencing Center, Washington University, St. Louis, Missouri 63108, USA;* [3]*Department of Genetics and Microbiology, University of Bari, 70126 Bari, Italy*

The gibbon genome exhibits extensive karyotypic diversity with an increased rate of chromosomal rearrangements during evolution. In an effort to understand the mechanistic origin and implications of these rearrangement events, we sequenced 24 synteny breakpoint regions in the white-cheeked gibbon (*Nomascus leucogenys*, NLE) in the form of high-quality BAC insert sequences (4.2 Mbp). While there is a significant deficit of breakpoints in genes, we identified seven human gene structures involved in signaling pathways (*DEPDC4*, *GNG10*), phospholipid metabolism (*ENPP5*, *PLSCR2*), β-oxidation (*ECHI*), cellular structure and transport (*HEATR4*), and transcription (*ZNF461*), that have been disrupted in the NLE gibbon lineage. Notably, only three of these genes show the expected evolutionary signatures of pseudogenization. Sequence analysis of the breakpoints suggested both nonclassical nonhomologous end-joining (NHEJ) and replication-based mechanisms of rearrangement. A substantial number (11/24) of human–NLE gibbon breakpoints showed new insertions of gibbon-specific repeats and mosaic structures formed from disparate sequences including segmental duplications, LINE, SINE, and LTR elements. Analysis of these sites provides a model for a replication-dependent repair mechanism for double-strand breaks (DSBs) at rearrangement sites and insights into the structure and formation of primate segmental duplications at sites of genomic rearrangements during evolution.

[Supplemental material is available online at www.genome.org.]

Chromosomal evolution in primates has been investigated at several levels of resolution, including comparative chromosome banding (Yunis and Prakash 1982), gene mapping (Turleau et al. 1983), cross-species chromosomal painting (Jauch et al. 1992; Murphy et al. 2005), comparative genome hybridization painting (Carbone et al. 2006), and fluorescent in situ hybridization (FISH) (Wienberg 2005). In general, linkage groups, gene order, and function have remained relatively unchanged since the common catarrhine primate ancestor (Haig 1999). Recent studies have not only identified the role of segmental duplications in disease and evolution but have also supported a nonrandom "fragile-breakage" model for chromosomal rearrangements (Armengol et al. 2003; Pevzner and Tesler 2003; Bailey et al. 2004). Overall, ~40% of chromosomal rearrangements are associated with segmental duplications in mammals (Bailey and Eichler 2006). Segmental duplications are also a major impetus for the evolution of novel genes and gene functions by duplication and domain accretion (Eichler 2001; Samonte and Eichler 2002). However, in certain primate lineages, the position of chromosomal breaks and the evolutionary rate of rearrangements follow unpredictable patterns (O'Brien and Stanyon 1999) and the role of segmental duplications is not well established.

Gibbons, extant genera among the hominoids, show both anatomical and behavioral specializations. Compared with other apes, gibbons are small, slender, and agile, exhibit no sexual dimorphism, and have very long arms adapted for a spectacular arm swinging locomotion called "brachiation" (Clutton-Brock et al. 1977; Gebo 1996; Plavcan 2001; Usherwood and Bertram 2003). Gibbons have loud vocalizations and live in small monogamous families composed of a mated pair and offspring (Harcourt et al. 1981; Plavcan 2001; Dooley and Judge 2007). In contrast to other apes, which show limited chromosomal variation, gibbons (family Hylobatidae) exhibit rapid chromosomal evolution with a diverse karyotypic pattern among different species and subspecies (O'Brien and Stanyon 1999; Muller et al. 2003). Humans and gibbons are estimated to have separated from their common hominoid ancestor between 15 and 20 million years ago (mya) (Goodman 1999), and, subsequently, waves of synteny block rearrangements in the common gibbon ancestor (Hylobatidae) gave rise to four distinct gibbon genera with varying chromosomal numbers (Jauch et al. 1992; Muller et al. 2003). Furthermore, 84 of the 107 synteny breaks in gibbons, relative to humans, are specific to the gibbon lineage, inherited from the common gibbon ancestor, while the remainder (23/107) occurred in the common hominoid ancestor (Roberto et al. 2007). Interestingly, 14 of the 84 gibbon synteny breaks are specific to the white-cheeked gibbon (*Nomascus leucogenys*, NLE), suggesting increased chromosomal rearrangement in that gibbon lineage (Muller et al. 2003).

The orthologous chromosomal blocks between human and NLE gibbon were recently mapped by two studies using bacterial artificial chromosome (BAC) end sequencing or array painting and confirmed by FISH (Carbone et al. 2006; Roberto et al. 2007). The average breakpoint resolutions of these two studies were ~80 kbp and 200 kbp, respectively (Carbone et al. 2006; Roberto et al.

[4]These authors contributed equally to this work.
[5]Corresponding author.
E-mail eee@gs.washington.edu; fax (206) 221-5795.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.086041.108.

19:178–190 ©2009 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/09; www.genome.org

2007). At this level of resolution, molecular mechanisms causing synteny breaks were not clear; however, segmental duplications were estimated to be associated with 46% of the rearrangements (Carbone et al. 2006). Although the potential for disruption of several genes in the vicinity of the breaks was suggested, the effect of the breaks on the gene structures, per se, was not well defined at this resolution. Previously, sequencing of a subset of gibbon BAC clones revealed segmental duplications or interspersed repeats at the breakpoints, although a detailed analysis of these regions was not presented (Carbone et al. 2006). While karyotypic variations are implicated for anatomical and phenotypic differences between hominoid species (Ferguson-Smith and Trifonov 2007), a high-resolution comparative genomics approach is imperative to identify the underlying causative molecular event.

We performed a sequence-based assessment of human and white-cheeked gibbon synteny breaks (1) to determine the sequence architecture and genomic characteristics predisposing to synteny breaks and chromosomal instability in gibbons, and (2) to determine the extent of gene rearrangements, correlating these with signatures of molecular evolution. Since regions of chromosomal rearrangement are frequently enriched in complex repetitive structures that are sometimes difficult to resolve by whole-genome sequence assembly, we targeted large-insert gibbon BAC clones for complete high-quality sequence analysis. Our analysis has characterized a subset of human–gibbon breakpoints at the sequence level, provided insight into the mechanism of rearrangement, and identified genes that potentially contribute to the evolution of the gibbons.

Results

Sequence resolution of human–gibbon breakpoints

We previously mapped the position of gibbon rearrangements orthologous to human chromosomes (HSA) by BAC-end sequence mapping and FISH (Fig. 1A; Roberto et al. 2007). Based on the BAC-end sequencing data and FISH-derived framework of human and NLE gibbon maps, we selected 24 gibbon BACs that span the syntenic breaks on the human genome for complete insert sequencing (see Methods). Our target set included eight intrachromosomal and 16 interchromosomal gibbon rearrangements with respect to the human genome (Table 1). We purposefully biased against regions associated with segmental duplications (SDs) due to the inherent difficulties in resolving breakpoints within duplicated regions, ambiguity associated with experimentally validating these events by FISH, and difficulties in obtaining large-insert clones. As such, we anticipated that we would enrich for rearrangement events mediated by nonhomologous end-joining (NHEJ) as opposed to nonallelic homologous recombination (NAHR). Each of the 24 BACs was sequenced (generating ~4.2 Mbp of finished, high-quality NLE genomic sequence) and aligned to the human genome sequence assembly (Build 35; Fig. 1B). The NLE gibbon synteny blocks mapped unambiguously to orthologous regions on human chromosomes, consistent with the experimental FISH results (Table 1; Fig. 1A,B).

Breakpoint analysis

We compared orthologous human and gibbon genomic sequences using a modified *mirepeats* analysis (Parsons 1995) and a multiple sequence alignment analysis (ClustalW) (Higgins et al. 1996) to precisely identify the breakpoint or breakpoint interval for each event (see Methods). We manually curated all multiple sequence alignments and, due to the sequence heterogeneity and complexity of several breakpoints, we inspected regions flanking each of the breakpoints for orthology based on the analysis of high-quality alignments. The repeat content of both gibbon BACs and human orthologous regions was annotated using RepeatMasker (http://repeatmasker.org) and *DupMasker* (Jiang et al. 2008) (Supplemental Tables 1, 2). In addition, we examined the gibbon BAC sequences for the presence of lineage-specific gibbon duplications by identifying regions of excess read depth from available gibbon whole-genome shotgun (WGS) sequence data (Bailey et al. 2002).

A comparison of human and gibbon breakpoints revealed two distinct classes: class I (n = 9), where the two syntenic regions precisely abut the breakpoint, and class II (n = 15), where the breakpoint could only be assigned to a sequence interval (termed breakpoint interval) (Fig. 2; Supplemental File 1). Class II breakpoints typically included additional sequences, ranging in length from 9 bp to 20 kbp, that did not map to either human orthologous chromosomal region (Table 1; Supplemental File 1). Nine class II breakpoints contained intervals ranging between 9 bp and 669 bp that also included insertions of AT-rich repeats, LTR (Supplemental Fig. 1), and *Alu*Y repeat elements, and one breakpoint interval contained insertion sequences generated by a replication slippage event (Table 1; Supplemental File 1). The 669-bp insertion formed a "mosaic" structure consisting of a series of three LTR5 elements and L1 repeats interspersed with nonrepeat sequences (Supplemental Fig. 1). We found no significant difference in the distribution of class I and class II events (Supplemental Table 3) when considering rearrangement events that occurred early within the gibbon phylogeny or, more recently, within the *Nomascus* lineage (Misceo et al. 2008).

Six breakpoints contained larger insertion sequences ranging from >1 kbp up to 20 kbp in length (Table 1; Supplemental File 1). Three of these corresponded to LINE elements (one case with an L1P insertion [1.1 kbp] and two cases with L1PA4 elements [8 kbp and 5.5 kbp]) (Fig. 3A,B; Supplemental File 1). Of note, one of these breakpoints contained three L1PA4 elements arranged in tandem in gibbons but was absent in the corresponding syntenic region in humans (Fig. 3A). While the 1.1-kbp interval consisted of a single L1P element, the 8-kbp and 5-kbp intervals both consisted of a combination of L1PA4, L1MA3, simple repeats, or nonrepeat sequences (Fig. 3A,B; Supplemental File 1). No target site duplications (TSDs) were associated with these elements (Supplemental Table 4), suggesting an endonuclease-independent retrotransposition process (Morrish et al. 2007).

Although we biased our initial selection against segmental duplications, we found that one-third (8/24) of the sequenced gibbon BACs contained segmental duplications flanking the breakpoint intervals, ~58% (135/234 kbp) of which occurred specifically within the gibbon lineage (Supplemental Table 5). We identified two breakpoint intervals that were themselves novel gibbon SDs (20 kbp and 4.3 kbp in length) (Fig. 4A,B) and spanned the breakpoint interval. Both SDs were also mosaic in their organization. For example, our sequence analysis of the 20-kbp SD showed that it mapped to multiple locations on human chromosome 17. It consisted of three major segments: a 5.9-kbp fragment, containing the gene structures for *CCL3*, *CCL3L1*, and a previously identified "core" duplicon (partial duplications of the *TBC1D3B* and *TBC1D3C* genes) on chr17q12 (Jiang et al. 2007; Sharp et al. 2008), a 12.6-kbp segment mapping to the *KRT17* gene on chr17q21.2, and an overlapping 7.4-kbp segment that lacked

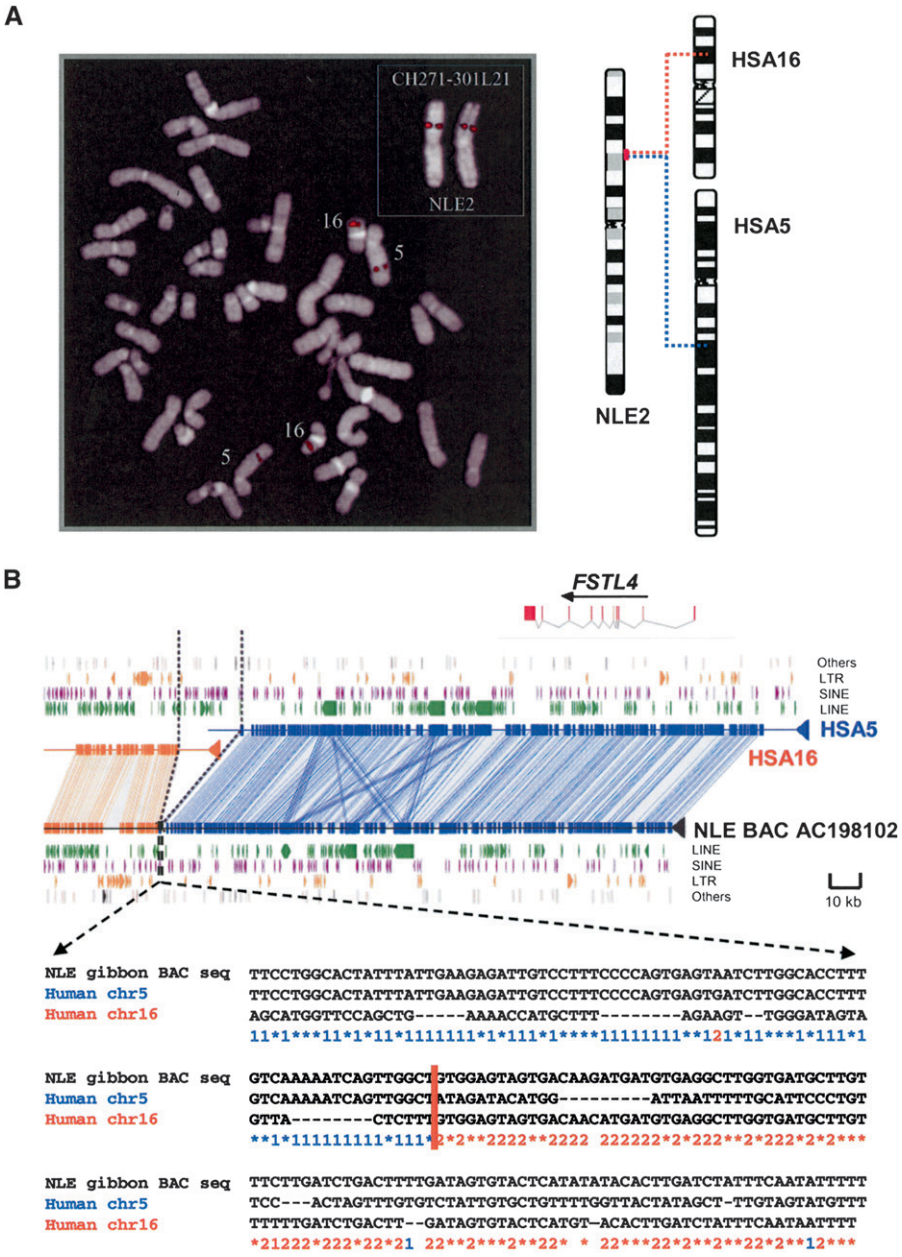

Figure 1. (*A*) Identification of gibbon BAC clones at the breakpoint of synteny. All BAC clones were experimentally validated by FISH as described previously (Roberto et al. 2007). In this example, a gibbon BAC clone spanning the breakpoint shows a single signal on gibbon chromosome 2 (NLE 2), but FISH mapped to human shows two signals on chromosomes 5 and 16, identifying an interchromosomal rearrangement (as represented by the chromosomal ideogram). (*B*) Sequence architecture at human–NLE gibbon synteny breaks. (*Top* panel) *Miropeats* analysis of the gibbon BAC, CH271-301L21 (AC198102), when compared with segments of human chromosome 5 (132461336–132644892, blue) and chromosome 16 (73369800–73421145, orange). Representative repeat elements, LINE, SINEs, LTRs, segmental duplications, and genes mapping to the synteny blocks with arrows denoting transcriptional orientation are also shown based on human genome annotation. (*Bottom* panel) Three-way ClustalW alignment between human and NLE gibbon sequences at the breakpoint with 1 (blue) denoting a sequence identity with the human chromosome 5 segment and 2 (orange) indicating sequence identity with the human chromosome 16 segment. The figure shows a class I breakpoint where the human breakpoints abut precisely at the point of fusion on the gibbon chromosome.

genes (Fig. 4A). The second duplication at a gibbon breakpoint was smaller in size, a 4.3-kbp SD insertion. It shared high sequence identity (>95% identity, >1 kbp) to two sequences located 72 kbp and 64.5 kbp upstream of the translocation on chromosome 3 (Fig. 4B), possibly as a result of skipping of templates during replication (Fig. 4B; Lee et al. 2007; Smith et al. 2007; Payen et al. 2008). In both cases, the SDs mapped at the junctions of interchromosomal translocation fusion points (in gibbon) but were formed from template sequences located on only one of the two chromosomes involved in the translocation process.

The final class II breakpoint carried a 1.2-kbp insertion that was a "hodgepodge" of LINE, SINE, and LTR elements (Table 1; Supplemental Fig. 2; Supplemental File 1). BLAST analysis showed this breakpoint interval sequence did not map en bloc to either the human or the macaque genome, indicating that this particular constellation of sequence elements formed within the gibbon lineage. Similarly, our sequence analysis showed the divergence estimates of the LINE insertions and both SD insertions to be consistent with events that had occurred specifically within the gibbon lineage (Supplemental Tables 2, 3). Irrespective of their mechanism of origin, these data argue that human–gibbon synteny breaks are particularly receptive for the accumulation of additional retrotransposons and segmental duplications.

To explore a possible common mechanism for synteny breaks, we further analyzed breakpoint regions for enriched sequence motifs (Supplemental File 1; see Methods). We identified short stretches of 2–6 bp of microhomology in 50% (12/24) of the breakpoint regions from both classes (Supplemental File 1), suggesting a nonclassical NHEJ mechanism for synteny breaks (Yan et al. 2007). Such microhomology motifs have, for example, been associated with template switching double-strand break (DSB) repair (Lee et al. 2007; Smith et al. 2007). Also, previously described sequence motifs associated with DSBs and recombination hotspots (Abeysinghe et al. 2003) were identified in the region flanking the breaks (Supplemental File 1). Finally, several orthologous NLE breakpoint regions in humans mapped within known regions of human copy number variation and structural variation (see Methods; Table 2; Supplemental Table 6).

Table 1. Human–NLE gibbon synteny breakpoint description

Accession #	NLE BP1	NLE BP2	BI (bp)	HSA1	HSA1 BP1	HSA2	HSA2 BP2	Class	Major events at BI
AC198096.2	55260	55930	669	chr12	99160177	chr19	58419994	Class 2	LTR insertion[a]
AC198097.2	139560	139561	0	chr7	2426891	chr2	150197264	Class 1	
AC198098.1	188452	193945	5492	chrX	34109189	chrX	62959716	Class 2	L1PA4 insertion[a]
AC198099.1	80766	80921	154	chr20	16576948	chr7	79700383	Class 2	AT repeats insertion
AC198100.1	117646	117647	0	chr9	111500125	chr9	22288616	Class 1	
AC198101.2	112428	112429	0	chr8	62850942	chr8	99136636	Class 1	
AC198102.2	35647	35648	0	chr16	73411602	chr5	132471261	Class 1	
AC198103.2	68442	76575	8132	chr2	169062862	chr6	46244494	Class 2	L1PA4 insertion[a]
AC198144.2	104171	104191	19	chr5	75704508	chr16	19419764	Class 2	
AC198146.2	79913	79915	0	chr3	19801897	chr8	19972258	Class 1	
AC198147.2	104878	104897	18	chr1	54949528	chr1	209419093	Class 2	Replication slippage
AC198148.2	83464	87782	4317	chr12	45891115	chr3	147669371	Class 2	4.3-kbp SD insertion[a]
AC198149.2	133643	133657	13	chr2	27838990	chr17	59312954	Class 2	
AC198150.2	149046	151272	210	chr14	30985847	chr14	73091550	Class 2	AluY insertion
AC198151.2	63094	63102	0	chr10	52084834	chr10	89181767	Class 1	
AC198152.2	85902	85904	0	chr1	52267836	chr1	177890931	Class 1	
AC198153.2	19121	39159	20037	chr17	61632684	chr2	73522945	Class 2	20-kbp SD insertion[a]
AC198154.2	47396	48527	1130	chr19	44013676	chr7	22873425	Class 2	L1P insertion
AC198155.2	65597	66831	1233	chr17	77869736	chr2	99381310	Class 2	*hodgepodge* insertion[a]
AC198183.2	27239	27553	313	chr4	140726707	chr22	31041712	Class 2	LTR insertion
AC198526.1	177364	177374	9	chr3	131200589	chr3	15139105	Class 2	
AC198875.2	128267	128271	0	chr12	63567432	chr19	41824918	Class 1	
AC198944.2	144167	144169	0	chr9	30938803	chr6	27133088	Class 1	
AC198945.2	178362	178449	86	chr10	23997347	chr4	110641976	Class 2	AT repeat insertion

Shaded rows represent intrachromosomal rearrangements. (NLE) *Nomascus leucogenys*; (HSA) human chromosome; (BAC) bacterial artificial chromosome; (BI) breakpoint interval; (BP) breakpoint; (SD) segmental duplication.
[a]Mosaic insertions.

Gene content analysis

We identified seven human gene orthologs whose protein-coding sequences were disrupted by the rearrangement in gibbons (Table 3). These included genes involved in G-protein-coupled receptor signaling pathways (*DEPDC4* and *GNG10* (LOC552891), phospholipid metabolism including sphingomyelin hydrolysis (*ENPP5*) and transport (*PLSCR2*), peroxisomal β-oxidation (*ECH1*), cell structure organization (*HEATR4*), and ovarian and testicular functions (*ZNF461* [also known as *GIOT-1*]) in humans (Mi et al. 2005). To test for the enrichment of genes at synteny breaks, we simulated a random distribution of breakpoints to the human genome assembly, excluding segmental duplications due to our initial bias in selecting against these regions for sequence analysis in the gibbon. The number of breakpoints mapping within human RefSeq coordinates was used to estimate an empirical *P*-value (n = 100 permutations). Compared with the random simulation (expected = 19, standard deviation = 3.5), the rate of gene disruption observed in 24 gibbon breakpoints was significantly lower (observed = 7), indicating that gibbon rearrangement breakpoints are biased against gene disruptions (P = 0.02) (see Methods; Supplemental Fig. 3; Supplemental Table 7).

Interestingly, we found that 33% (8/24) of the BAC clones sequenced contained clusters of tandemly duplicated genes mapping within 50 kbp of the breakpoint, including the growth hormone cluster, KRAB-containing zinc finger genes (*ZNF677*, *ZNF483*, *ZNF512*, *ZNF567*, and *ZNF382*), vomeronasal type 1 receptors (*VN1R2* and *VN1R4*), phospholipase scramblase (*PLSCR1* and *PLSCR2*), and acyl-CoA thioesterases (*ACOT1* and *ACOT2*) (Supplemental Figs. 4, 5; Supplemental Data). In some cases, paralogous genes (based on human gene annotation) were disrupted in gibbons (Supplemental Fig. 5A). For example, one breakpoint mapped to the 5′UTR of the somatotropin hormone, *GH2*, predicting a disruption of transcription due to uncoupling of the promoter from its coding sequence—an observation that was also reported by Carbone and colleagues (2006). Sequence analysis of the other gene family members, *CSH1*, *CSH2*, and *CSHL1* within the gibbon, demonstrated numerous sequence variations, including obliteration of the start codon and point mutations in the sequence coding for the signal peptide domain of the proteins (Supplemental File 2). Similarly, the human paralogous gene, *ACOT1*, may be disrupted by the gibbon rearrangements, as SIM4 analysis predicted only the *ACOT2* gene in gibbons (see Methods; Supplemental Fig. 5B).

We investigated whether the gibbon rearrangement events coincided with changes in the evolutionary pressure of genes mapping at the breakpoints or distal to the breakpoints. For this purpose, we performed a maximum-likelihood evolutionary analysis using Phylogenetic Analysis by Maximum Likelihood (PAML) to calculate d_N/d_S ratios (ω) (see Methods) (Yang 1997). First, we reconstituted a complete gibbon gene model based on the BAC sequence and the available gibbon whole-genome shotgun sequence (for the portion of the gene that was not represented within the BAC clone) (Table 4). Next, we created a multiple sequence alignment of the coding sequence from available genome sequence data and generated a phylogenetic gene tree with a minimum of five orthologous genes from various primate and mammalian lineages (Supplemental File 3). It should be noted that the latter approach in the case of duplicated genes is suboptimal as it is impossible to accurately distinguish paralogous genes from WGS read data. Thus, more rigorous tests of selection within the human and gibbon lineages are not possible until a high-quality sequence of all duplicated gene family members has been generated.

Three genes disrupted in the protein-coding sequences clearly showed a relaxation of selection pressure within the gibbon

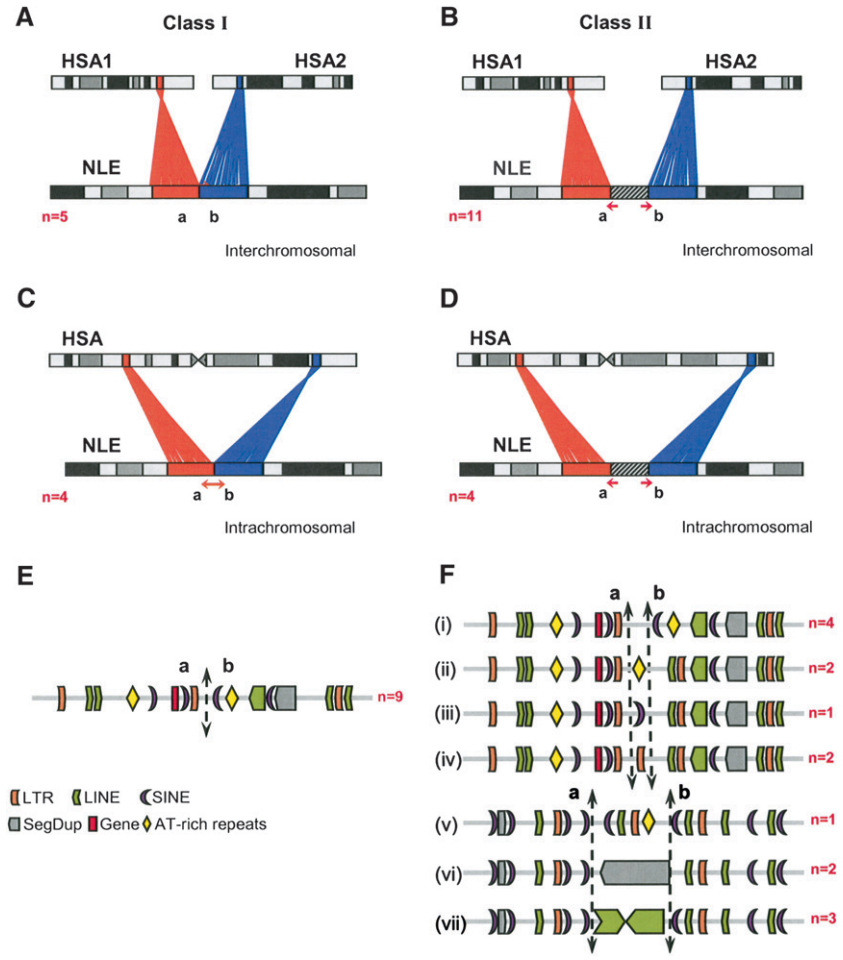

Figure 2. Class I and class II breakpoints. The schematic shows the types of rearrangements identified by high-resolution sequence analysis: Class I and class II breakpoints causing inter- (*A*,*B*) or intrachromosomal (*C*,*D*) rearrangements are shown. Based on the sequence context, the number (*n*) of different human–gibbon breakpoints identified from both categories (*E*,*F*) are also shown. Note that the class II breakpoints contain: (*i*) nonrepeat sequences, (*ii*) AT-rich repeats, (*iii*) SINEs (*AluY* element), (*iv*) LTR insertions, (*v*) a "hodgepodge" of repeats, (*vi*) segmental duplications, and (*vii*) LINE-1 elements. The diagram is not to scale.

lineage; namely, *DEPDC4* (ω = 1.31), *HEATR4* (ω = 1.03), and *GNG10* (ω = 0.927), consistent with pseudogenization as a result of the rearrangement (ω ~21 for gibbon branch in the phylogeny; Table 4). Two additional gibbon gene models showed the presence of multiple nonsense mutations despite d_N/d_S ratios suggesting purifying selection (ω < 1) (Fig. 5; Table 4; Supplemental File 3); namely, *ECH1* (ω = 0.25 and 0.18) and *ZNF461* (ω = 0.13 and 0.0001). A comparison using a free codon-substitution model for neutral (ω = 1) or conserved (ω = 0.5) evolution in the gibbon branch for all analyzed genes suggested a significantly conserved evolution for *ECH1*, *ZNF461*, and *GNG10* (LOC552891) (see Methods; Supplemental Table 7). Coding sequences for *PLSCR2* and *ENPP5* were not available (in the current gibbon WGS assembly) for evolutionary analysis. As expected, analysis of genes distal to the breakpoints demonstrated signatures of purifying selection (Table 4; Supplemental Table 8; Supplemental File 3).

Discussion

Gibbons are known to have a rapid rate of chromosomal evolution among the hominoids, mainly involving large-scale rearrangements and rapid karyotypic divergence (Muller et al. 2003; Carbone et al. 2006; Roberto et al. 2007). In contrast to human and great ape segmental duplications, where ~70% of all large-scale evolutionary rearrangements map to regions of segmental duplication (Kehrer-Sawatzki and Cooper 2008), initial studies of the gibbon reported that only 46% of gibbon breakpoints mapped to sites of segmental duplication in the human lineage (Carbone et al. 2006). BAC sequence analysis of a smaller subset identified segmental duplications or interspersed repeats at most breakpoints; however, two clones in this initial study also showed evidence of "micro-rearrangements" containing disparate repeat sequences derived from various human chromosomal locations (Carbone et al. 2006). These initial data from Carbone and colleagues hinted at potential alternate mechanisms of rearrangements, although the number of sites and the extent of sequence analysis were limited. In this study, we expanded upon earlier work (Carbone et al. 2006; Roberto et al. 2007) to present single-base-pair resolution of 24 human–gibbon breakpoints of synteny within the context of 4.2 Mbp of high-quality gibbon BAC sequence.

The most striking finding was the presence of additional sequences for ~40% of the gibbon sites of translocation, suggesting a more complex rearrangement mechanism than simply nonallelic homologous recombination or nonhomologous end joining. The largest (1–20 kbp) of these insertion sequences consisted of various classes of repetitive DNA including segmental duplications and L1 repeats. Detailed sequence analyses of these new insertions reveals two important features. First, we note that in the case of L1 elements, we observed no target-site duplications, suggesting that they did not originate as a result of typical endonuclease-mediated retrotransposition (Morrish et al. 2007). Second, in many cases the new insertion sequences are mosaic structures composed of disparate common repeats or duplicated sequences (Figs. 3, 4; Supplemental Fig. 2) that originate upstream of the rearrangement breakpoint.

At least two different mutational mechanisms are consistent with these observations. Since microhomology was observed in 50% of the human–NLE gibbon breaks (Supplemental Fig. 6), one possibility may be a microhomology-mediated end-joining (MMEJ) mechanism, recently reported as a nonclassical NHEJ mechanism for translocations in mammals (Yan et al. 2007). Sequence microhomology and site-specific recombinogenic sequences in the vicinity of the breakpoints have been associated with translocations

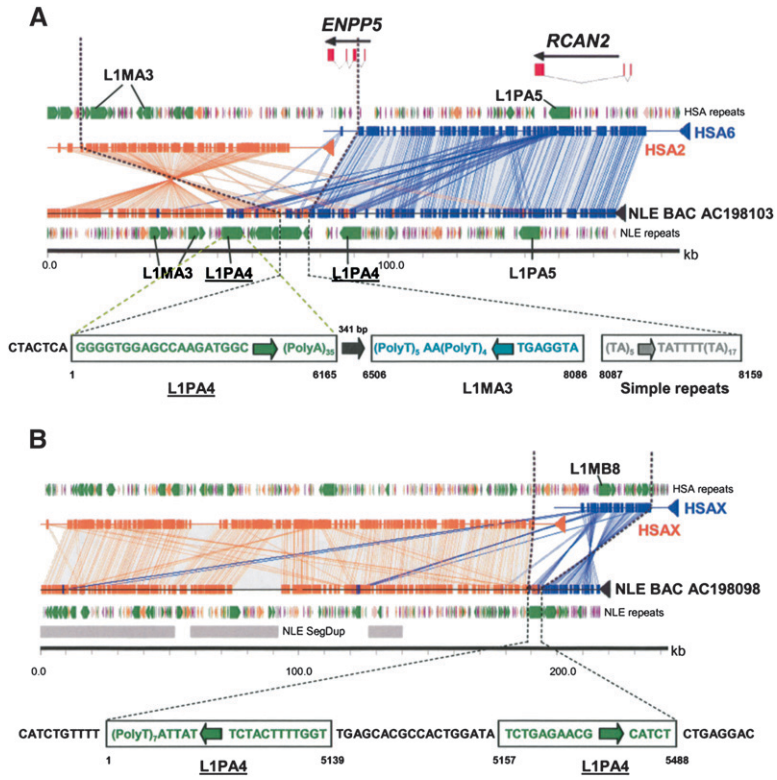

Figure 3. L1PA4 repeat insertions at the breakpoints. Human–gibbon pairwise alignment by *miropeats* is shown. The NLE gibbon-specific segmental duplications are also remarkable. LINE-1 elements, L1PA4 (green block arrows), and L1MA3 (dark green arrows) in the vicinity of human–gibbon synteny breaks are shown. There are three L1PA4 elements (underlined) in panel *A* and one in panel *B*. Note that the L1PA4 elements are specific to the NLE gibbon chromosomal segment. (Black dotted lines) Extent of breakpoint intervals and sequence structure of each repeat. The directions of the arrows denote orientation of the LINEs, and the numbers denote the sequence extent.

in evolutionary rearrangements and cancer (Kehrer-Sawatzki et al. 2002; Abeysinghe et al. 2003; Wei et al. 2003). We identified sequence motifs (e.g., topoisomerase II and translin sites) consistent with DSB and repair mechanisms generating overhangs at several human–NLE gibbon breakpoints (Negrini et al. 1993; Kanoe et al. 1999; Wei et al. 2003). We propose that these overhangs may have been repaired by an "error-prone" mechanism, creating some of the smaller breakpoint intervals (Fig. 6A).

Both the microhomology and, more importantly, the mosaic architecture of the larger breakpoint intervals are also consistent with more recently proposed replication-based mechanisms such as FoSTeS (*fo*rk *s*talling *te*mplate *s*witching) (Lee et al. 2007) and MMIR (microhomology/microsatellite-induced replication) (Payen et al. 2008). Template switching as a result of multiple rounds of strand invasion from DSB sites generated by stalled or collapsed replication forks to ectopic sites could, in principle, explain some of the events we have observed (see "gap-fill model," Fig. 6B) (McVey et al. 2004; Lee et al. 2007; Smith et al. 2007). Repeat-rich sequences frequently serve as preferred templates because of their tendency to interfere with replication fork progression, leading to the formation mosaic structures at the point of rearrangement (Figs. 3, 4, 6B; Supplemental Fig. 7; Kehrer-Sawatzki and Cooper 2008; Payen et al. 2008). A remarkable example was the presence of a 4.2-kbp gibbon-specific segmental duplication mapping precisely at the translocation fusion point between chromosomes 3 and 12. Sequence analysis revealed that this segmental duplication actually consisted of duplicatively transposed sequences mapping 72 kbp and 64.5 kbp further upstream of the point of fusion on chromosome 3.

Although we have clearly biased against homology-based events, such insertions of mosaic structures have not yet been described at sites of rearrangement between humans and the African great apes, most of which have now been characterized at the molecular level (Kehrer-Sawatzki and Cooper 2007, 2008). Do these results provide any insight into the apparent increased tempo of large-scale rearrangements in the gibbon lineages? There are a few important facts. First, computational analyses of the human genome based on percent sequence identity suggest a burst of segmental duplications in the African great-ape lineage when compared with other apes (Cheng et al. 2005; Bailey and Eichler 2006). Second, most large-scale chromosomal rearrangements in humans and African great apes are intrachromosomal as opposed to interchromosomal translocations (Kehrer-Sawatzki and Cooper 2007, 2008). Third, 65%–70% of all great ape chromosomal rearrangements were associated with large blocks of segmental duplication (Cheng et al. 2005; Kehrer-Sawatzki and Cooper 2007), although the number appears to be lower in gibbons (46%) (Carbone et al. 2006). One possibility may be that a paucity of segmental duplications in ancestral gibbon genomes channeled rearrangement pathways away from NAHR, favoring these alternate mechanisms (e.g., MMIR, FoSTeS, break-induced replication). We speculate that the overall "rate of rearrangement" is largely constant among all ape genomes but that fewer SDs drive fewer homology-mediated events and, consequently, nonhomology-based mechanisms contribute more significantly to large-scale chromosomal rearrangements in gibbons. Many SD-mediated events have occurred among great apes, but because of the predominance of interspersed duplication blocks within close proximity along a chromosome, a large number of these African-ape events are below the level of cytogenetic resolution and instead are observed as an abundance of smaller structural variant events (Feuk et al. 2005; Newman et al. 2005).

In this model, intrachromosomal segmental duplications essentially "resolve" larger chromosomal rearrangements in the African great ape/human genomes (Kehrer-Sawatzki and Cooper 2007). Moreover, given that NAHR events are often associated with breakpoint reuse (Bailey et al. 2004; Murphy et al. 2005; Zody et al. 2008), at a constant rearrangement rate, the great apes would show apparently fewer structural changes, due to recurrent rearrangements involving the same chromosomal segments.

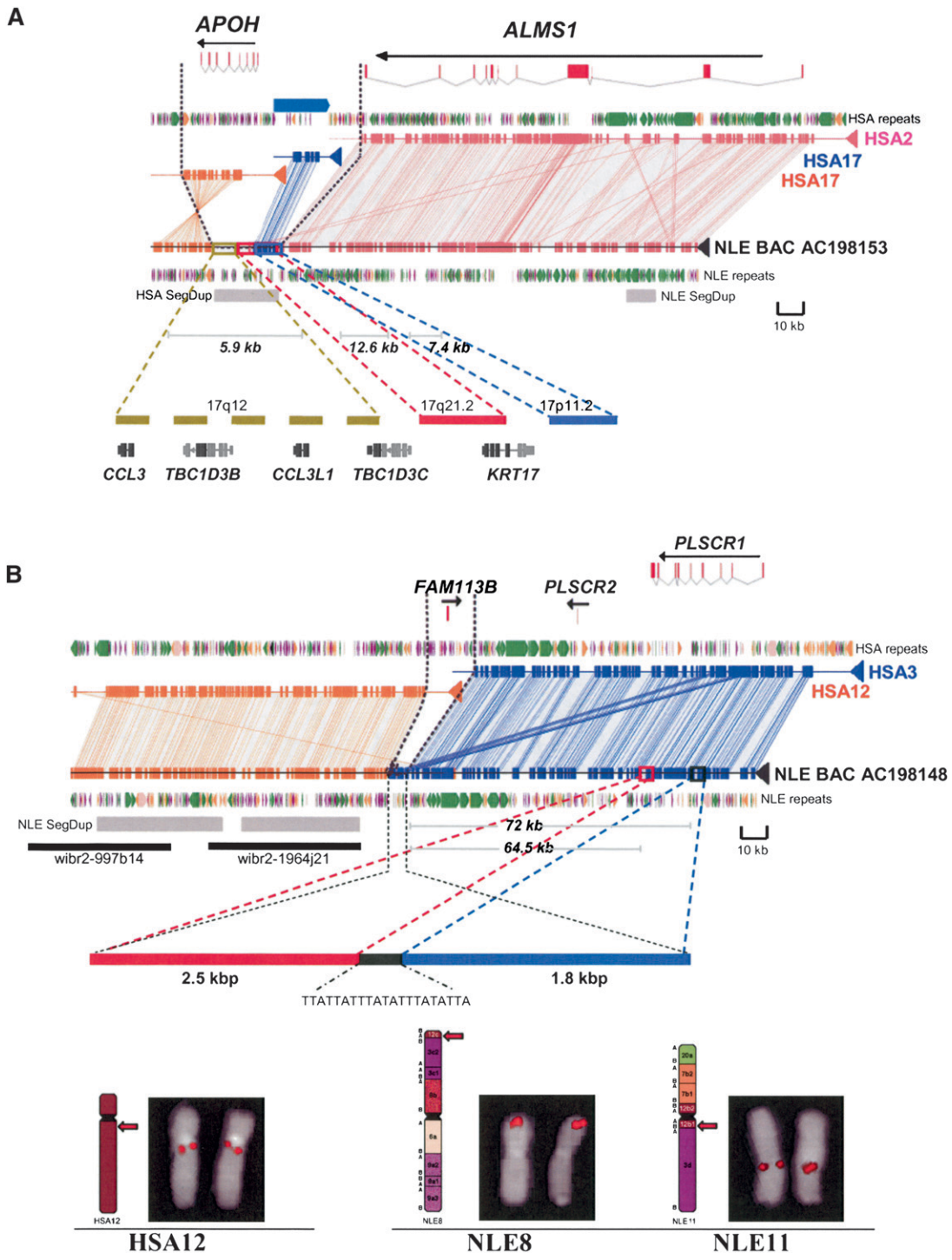

Figure 4. Segmental duplication insertions at the breakpoints. Alignments between the NLE BAC sequences and human chromosomes are shown. These breakpoints belong to class II category. (*A*) Note the insertion of an ~20 kbp segmental duplication (gray box) at the breakpoint. The sequence interval maps to several regions on human chromosome 17, some of which are depicted (solid colored bars). The length of each insertion segment, encompassing gene structures, and karyotypic mapping location are also shown. Gene fragments that do not map to the breakpoint sequences are shown in gray. (*B*) Insertion of a 4.3-kbp sequence at the breakpoint. Please note that the NLE gibbon BAC is in the reverse orientation. A schematic depicting the arrangement of a 4.3-kbp sequence block at the breakpoint derived from ~2.5-kbp and 1.8-kbp blocks located ~72 kbp and 64.5 kbp upstream, respectively, are also shown. The location of human fosmid probes (black bar), wibr2-1964j21 (chr12: 45810892–45850262) and wibr2-997b14 (chr12: 45855081–45893396), used to map the NLE-specific segmental duplication, is also shown. (*Bottom* panel) Representative comparative FISH signal pattern on human (HSA) and gibbon (NLE) chromosomes using a human fosmid (wibr2-1964j21) probe mapping to segmental duplications ~8 kbp downstream from the breakpoint (see Roberto et al. [2007] for FISH methods). Both the fosmids showed signals on NLE8 (12c) and NLE11 (12b1), displaying the presence of duplications on both translocated chromosomes. Syntenic blocks between human and gibbon chromosomes are reported diagrammatically on the *left* side of NLE chromosomes, according to Roberto et al. (2007).

Table 2. Sequence architecture of gibbon BACs containing human–NLE gibbon synteny breaks

Accession #	HSA1	Genes	HSA2	Genes	Overlap with SV/CNV sites[a,b]	Flanking repeat architecture
AC198096.2	chr12	*DEPDC4*	chr19			Segmental duplication
AC198097.2	chr7	*IQCE*	chr2		Recombination hotspot (HSA1)	SINE
AC198098.1	chrX		chrX			Segmental duplication
AC198099.1	chr20		chr7		Recombination hotspot (HSA2)	LINE, AT-rich repeats
AC198100.1	chr9	*GNG10* (LOC552891)	chr9			Segmental duplication
AC198101.2	chr8		chr8			SINE
AC198102.2	chr16		chr5		ASD CNV[c] (HSA1), recombination hotspot (HSA2)	LINE, SINE
AC198103.2	chr2		chr6	*ENPP5*		LINE
AC198144.2	chr5		chr16	*TMC5*		SINE
AC198146.2	chr3		chr8		Recombination hotspot (HSA1)	Simple repeats, LTR
AC198147.2	chr1	*FLVCR1*	chr1			SINE
AC198148.2	chr12		chr3	*PLSCR2*		Segmental duplication
AC198149.2	chr2		chr17	*GH2*		LINE, SINE
AC198150.2	chr14	*HEATR4, C14Orf126*	chr14	*ACOT1,2*	Fosmid SV map,[d] CNP1087 (both on HSA2)	LTR
AC198151.2	chr10		chr10		Autism CNV[e] (HSA1)	
AC198152.2	chr1		chr1	*BTF3L4*		
AC198153.2	chr17		chr2	*ALMS1*	CNP1218 (HSA1)	Segmental duplication
AC198154.2	chr19	*ECH1*	chr7		Recombination hotspot (HSA2)	SINE
AC198155.2	chr17	*CD7*	chr2		Recombination hotspot (HSA2)	Segmental duplication
AC198183.2	chr4		chr22			Segmental duplication
AC198526.1	chr3		chr3		CNP268 (HSA2)	SINE
AC198875.2	chr12		chr19	*ZNF461*	Recombination hotspot (HSA1)	SINE
AC198944.2	chr9	*DZIP1*	chr6		CNP779 (HSA1)	Segmental duplication
AC198945.2	chr10		chr4			LINE, AT-rich repeats

(HSA) Human chromosome.
[a]Recombination hotspot location obtained from the UCSC Genome Browser culled from the HapMap Phase I data and Perlegen data (Hinds et al. 2005).
[b]Copy number polymorphism map from Genome Browser SV database.
[c]Marshall et al. (2008).
[d]Kidd et al. (2008).
[e]Sebat et al. (2007).

However, gibbons with fewer SDs would tend to have more distinct structural changes, although with the same effective number of events. In this regard, it is interesting that we previously noted no apparent increase in smaller rearrangements in gibbon despite the nearly fourfold increase in gross chromosomal rearrangement events when compared with the African great apes (Roberto et al. 2007). High-quality sequence of many more breakpoints within ape lineages will be necessary to fully address this model.

Although the precise mechanism(s) underlying these events is not yet understood, it is clear that segmental duplications are intimately associated with large-scale chromosomal rearrangements. Even when we bias against SD regions such as in this study, the association resurfaces. Bailey et al. (2004) proposed that the association between segmental duplications and large-scale genomic rearrangements is not entirely causative. In our study, eight breakpoints mapped within 100 kbp of a previously characterized segmental duplication. Since no homology was detected at corresponding chromosomal positions of the rearrangement (Supplemental Data; Table 2), we exclude the possibility of homology-mediated (or NAHR) events. In four cases (Supplemental Table 5; Supplemental File 4), we identified gibbon-specific segmental duplications mapping distal to (within 50 kbp) gibbon fusion breakpoints. One example is the gibbon-specific segmental duplication mapping ~8 kbp downstream from the HSA3 and HSA12 translocation breakpoint (Fig. 4B). FISH analysis using human fosmid probes showed signals on both translocated chro-

Table 3. Genes disrupted at human–NLE gibbon synteny breaks

Genes	Location	Breakpoint	Description	Function
DEPDC4	12q23.1	Exon 1–2 (5)[a]	DEP (disheveled, Egl-10, pleckstrin) domain containing 4	G-protein-coupled membrane receptor
GNG10 (LOC552891)[a]	9q31.3	Exon 2 (3)	Guanine nucleotide binding protein, gamma 10	Heteromeric G-protein involved in neurohormonal pathways
ENPP5	6p12.3	Exon 1–2 (4)	Ectonucleotide pyrophosphatase/ phosphodiesterase 5	Hydrolysis of dietary sphingomyelin
PLSCR2	3q24	Exon 1–2 (9)	Phospholipid scramblase protein 2	Phospholipid metabolism
HEATR4	14q24.3	Exon 1–2 (17)	Heat repeat containing 4	Cytoskeletal organization, cellular transport
ECH1	19q13.2	Exon 2–3 (10)	Peroxisomal enoyl-coenzyme A hydratase	β-oxidation of fatty acids in peroxisomes
ZNF461	19q13.12	Exon 5–6 (6)	Gonadotropin inducible ovarian transcription repressor	LH and FSH-mediated folliculogenesis

Numbers in parentheses represent total exons.
[a]*GNG10* (LOC552891) is an alternative splice variant of *GNG10*.

Table 4. Evolutionary analysis of genes in the vicinity of human–NLE gibbon breakpoints

Gene #ID	ω (d_N/d_S)	d_N	No. of nonsynonymous substitutions	d_S	No. of synonymous substitutions
Genes disrupted at human–NLE gibbon synteny breakpoints					
DEPDC4 (BAC)	1.3174	0.0232	9.2	0.0176	2.1
ECH1 (BAC)	0.2581	0.0226	4.3	0.0877	6.3
ECH1 (reads)	0.1849	0.0222	6.2	0.12	5.9
ECH1 (Union)	0.2267	0.0222	10.4	0.0979	12.1
ZNF461 (BAC)	0.1346	0.0078	1	0.0578	3.2
ZNF461 (reads)	0.0001	0	0	0.0145	1.1
ZNF461 (Union)	0.1532	0.0047	2	0.0306	4.2
GNG10 (BAC)	Not available				
GNG10 (reads)	0.927	0.0333	4.3	0.036	2
HEATR4 (reads)	1.0342	0.0183	6.6	0.0177	2.1
GNG10 (LOC552891) (BAC)	0.0001	0	0	0.1594	4.4
GNG10 (LOC552891) (reads)	Not available				
ENPP5	Not available				
PLSCR2	Not available				
Genes distal to the breakpoints (within 2 kbp of breakpoint window)					
ACOT1	0.3589	0.0072	4.1	0.02	4.2
ALMS1	0.739	0.0126	32.9	0.0171	14.8
CD7	0.2303	0.016	5.2	0.0696	5.4
CSH2	0.4694	0.0394	13.8	0.0839	8.9
GH2	0.8034	0.0246	8.6	0.0307	3.1
PLSCR1	0.5751	0.0907	26.3	0.1578	13.4
TMC5	0.4201	0.0094	11.4	0.0224	9.6

Coding sequences were retrieved from either the gibbon BACs or gibbon whole-genome shotgun (WGS) reads.

mosomes (Fig. 4B); however, no direct involvement of SD was evident in this chromosomal rearrangement due to the absence of its homologous counterpart on the other side (HSA3) of the breakpoint. Similarly, when we reanalyzed the 11 gibbon BACs at breakpoints reported by Carbone and colleagues (Carbone et al. 2006) using our analytical pipeline (Supplemental Tables 9, 10; Supplemental File 4), we identified at least five breakpoints that contain segmental duplications. None of these, however, show evidence of homologous sequence at both corresponding regions in the human genome arguing, once again, against nonallelic homologous recombination between ancestral segmental duplications.

These data clearly reinforce the strong association between segmental duplications and chromosomal rearrangements (O'Brien and Stanyon 1999; Armengol et al. 2003; Bailey et al. 2004) and imply that regions of rearrangement may, in fact, also be the source of new duplications (Kehrer-Sawatzki et al. 2002; Ranz et al. 2007). These data support an alternative model associating segmental duplication and rearrangements reinforcing that DSBs can generate segmental duplications (Koszul et al. 2004; Smith et al. 2007; Kim et al. 2008). Our model extends these observations to include both translocations as well as inversions. As mentioned, one possibility may be that the rearrangement regions could also serve as preferential templates for subsequent or concurrent strand invasion of other regions during replication-dependent repair, spawning de novo segmental duplications at other sites (Koszul et al. 2004). This view is further supported by our observation of the 20-kbp segmental duplication block mapping to a core duplicon on chromosome 17. Thus, regions of genome rearrangement may, in fact, promote the formation of segmental duplications at other regions of the genome, as opposed to these being the cause of evolutionary rearrangements.

From the genic perspective, our analysis supports the more general observation that structural variation occurs preferentially near or within duplicated genes (Locke et al. 2006; Redon et al. 2006; Kidd et al. 2008). The functional redundancy conferred by such duplicated genes might make these rearrangements more tolerable in an evolving species as opposed to disruptions of unique, single-copy genes. The growth hormone gene cluster, for example, is specific to the primate lineage and originated from a single ancestral *GH* gene by duplications. It comprises paralogous growth hormone genes (*GH1*, pituitary, and *GH2*, placental) and two chorionic somatomammotropin genes (*CSH1* and *CSH2*) (Barsh et al. 1983). The *CSH1* gene duplicated further to yield a chorionic somatomammotropin gene (*CSHL1*) that later became a pseudogene by inactivation (Misra-Press et al. 1994). Likewise, the *ACOT* gene cluster is variable in copy number between species. This protein family regulates intracellular levels of lipids by hydrolysis of acyl CoAs to free fatty acids and CoASH with localizations in the cytosol (*ACOT1*) and mitochondria (*ACOT2*). While the human *ACOT* cluster is composed of *ACOT1*, *ACOT2*, and *ACOT4*, the mouse cluster contains six paralogous genes (*Acot1–Acot6*). Similarly, the vomeronasal receptors have undergone a steady evolutionary decline from mouse to humans, with gradual inactivation of pheromone sensation genes, *VN1R2* and *VN1R4*, since the divergence of the Old World monkeys and the hominoids, ~23 mya (Zhang and Webb 2003). These examples highlight both the variability in copy number and functional diversity for these genes, making them preferred targets for large-scale rearrangement events. Recently, Dumas et al. identified a high rate of lineage-specific gene duplication in gibbons (Dumas et al. 2007). Our preliminary analysis of the gibbon genome does not support this observation. Among the segmental duplications that we identified at the breakpoints, we were unable to find any overlap between genes in our analysis and the ones identified by Dumas and colleagues.

Three genes disrupted by rearrangement in gibbon showed signatures of selection consistent with pseudogenization. While it is tempting to speculate that some of these gene losses may have contributed to morphological and behavioral specialization in the

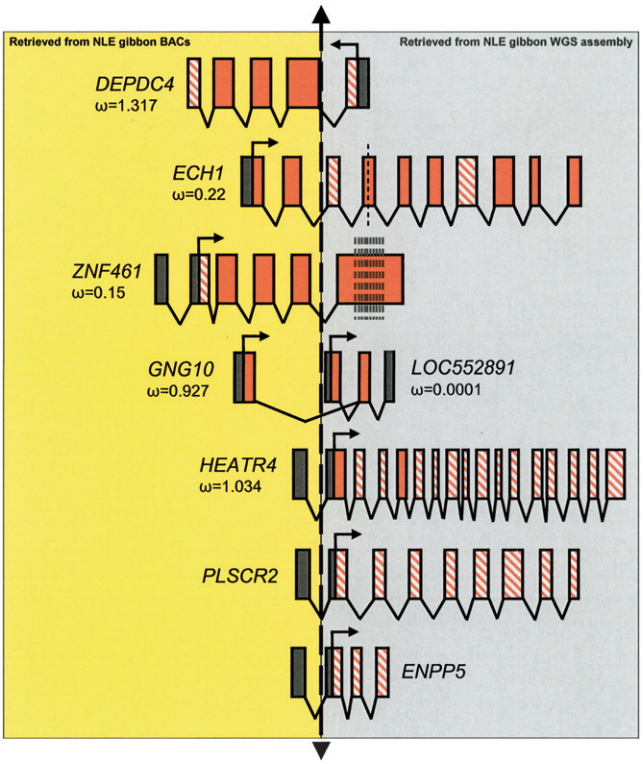

Figure 5. Gene disruptions at synteny breaks. The schematic shows the seven genes mapping to the breakpoints (dashed vertical arrow). One part of the gene is contained in the BAC (yellow region) and the other part is lost due to synteny break (gray region). Both gene parts were reconstructed either from the gibbon BAC sequences or contigs assembled from the gibbon WGS reads (see text). Coding exons (orange, completely retrieved sequences; stripes, missing sequences in gibbons) and noncoding exons (black) are depicted. (Black arrows) Transcriptional orientation. The d_N/d_S ratios (ω) and number of synonymous and nonsynonymous substitutions calculated for the available gene fragments (orange) are also shown. Vertical dashed lines on the exons indicate location of stop codons. The figure is not to scale.

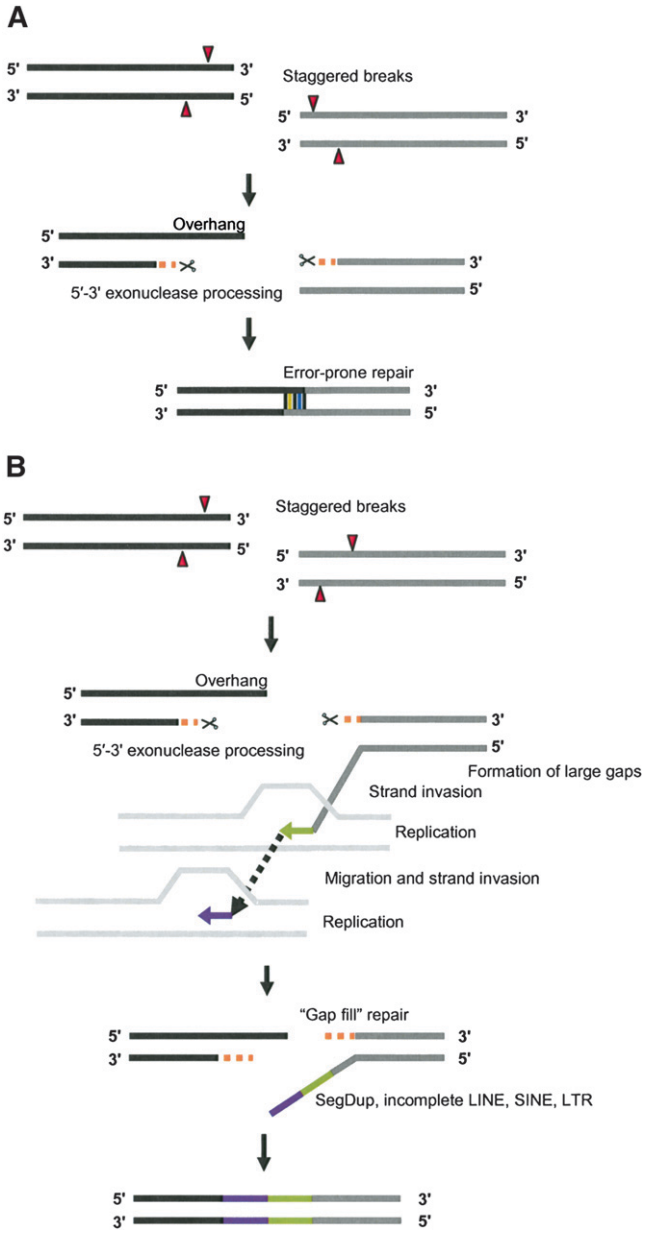

Figure 6. Models for human–NLE gibbon rearrangements. (*A*) An error-prone repair mechanism for smaller breakpoint intervals (<20 bp). DNA strands from two chromosomes (black and gray bars) are shown. Staggered double-strand breaks are processed by 5′–3′ exonuclease, creating overhangs. These overhangs are filled by an error-prone repair mechanism, creating shorter insertions. (*B*) "Gap-fill" model for larger breakpoint intervals. Large gaps are generated by double-strand breaks (due to possible collapsed or stalled replication forks) at rearrangement sites. These staggered breaks are processed by exonucleases to generate long 3′ overhangs. Replication is initiated by strand invasion to repair the gap. However, likely due to low processivity of the replication-dependent repair process (McVey et al. 2004), only smaller-length sequence stretches are synthesized. Consequently, a series of strand invasion, replication, and uncoupling of the replication machinery is necessary to fill the large gap. Thus, a less-efficient replication-based repair process generates a mosaic of incomplete repetitive elements at the larger breakpoint intervals.

gibbon lineage, further functional characterization of the genes and their impact on biochemical pathways and developmental lineages will be required. Our analysis, however, provides some interesting candidates for further investigation (i.e., loss of the growth hormone genes associated with lack of sexual dimorphism in the gibbon). Interestingly, not all genes appear to be dead as a result of rearrangement. Our preliminary analysis of two genes, *ECH1* and *ZNF461*, suggests a model of purifying selection. While the functional implications of these results are unclear, our results raise the intriguing possibility that a gene broken by a rearrangement event may not be doomed to pseudogenization, and the underlying coding sequences may be exapted for other functions in the organism.

Methods

Gibbon BAC sequencing

Twenty-four bacterial artificial chromosomes (BACs) were chosen from the white-cheeked gibbon, *Nomascus leucogenys*/NLE, BAC

library, CHORI-271, based on unambiguous signals with FISH (Roberto et al. 2007). The BACs were then subjected to whole-genome shotgun sequencing to at least sixfold sequence redundancy and assembled to completion at the Genome Sequencing Center, Washington University, St. Louis, Missouri.

The accession numbers of the BACs are as follows: AC198096.2, AC198097.2, AC198098.1, AC198099.1, AC198100.1, AC198101.2, AC198102.2, AC198103.2, AC198144.2, AC198146.2, AC198147.2, AC198148.2, AC198149.2, AC198150.2, AC198151.2, AC198152.2, AC198153.2, AC198154.2, AC198155.2, AC198526.1, AC198183.2, AC198944.2, AC198945.2, and AC198875.2 (Supplemental Data).

Sequence alignment and annotation

Gibbon BAC sequences were initially compared with human genome sequence using BLAST sequence similarity searches and *mir-opeats* (Altschul et al. 1990; Parsons 1995) to identify potential breakpoint intervals. Analysis for repeats on finished gibbon BAC sequences was performed using RepeatMasker, and segmental duplications (>94% identical, ≥10 kbp size) were detected using the whole-genome shotgun sequence detection (WSSD) strategy for gibbon (Bailey et al. 2002; Chen 2004). Human genomic coordinates corresponding to gibbon SDs (identified by WSSD mapped against the gibbon WGS clones) were intersected with human, chimp, orangutan, and macaque segmental duplication (T. Marques-Bonet and E.E. Eichler, unpubl.) to detect gibbon-specific SDs. Sequences homologous to known human SDs were detected on both syntenic human chromosomes and the gibbon BACs using *DupMasker* (Jiang et al. 2008). The sequences corresponding to syntenic regions on the human chromosomes and the gibbon BACs were aligned using ClustalW (Higgins et al. 1996). The exact sequence breaks in the alignments between gibbon and human sequences were identified as breakpoints or breakpoint intervals. To estimate the evolutionary age of various classes of repeats, sequence divergence from consensus repeat sequences was computed for each of the repeat elements mapping within and flanking the breakpoints.

Breakpoint analyses

Sequences around the breakpoints were compared with sequence motifs associated with DSBs, recombination, and chromosomal rearrangement, allowing for up to 2-bp mismatches. Sequences ±15 bp around the breaks were searched for previously reported 5–9-mer recombination hotspot sequences (Myers et al. 2005), topoisomerase consensus binding sites, topoIIv ([A/G]N[T/C]NNCNNG[T/C]NG[G/T]TN[T/C]N[T/C]) (Spitzner and Muller 1988), topoIId (GTN[T/A]A[C/T]ATTNATNN[A/G]) (Sander and Hsieh 1985), topoIIi ([T/C][T/C]CNTA[C/G][C/G]CC[T/G][T/C][T/C]TNNC) (Kas and Laemmli 1992), and translin recognition sites (ATGCAG and GCCC[A/T][G/C][G/C][A/T]) (Aoki et al. 1995) on both strands using C-program-based K-mer finder and BLAST (Altschul et al. 1990). A homology of >75% is considered a strong binding/cleavage site (Spitzner and Muller 1988). Sequence motifs identified in cancer-associated rearrangements were also compared with sequences near the human gibbon synteny breaks (Abeysinghe et al. 2003).

The significance of breakpoints within genes (human Refseq) and within human recombination hotspots was determined by simulation. Breakpoints were randomly distributed to the human genome assembly (Build 35), and the number of breakpoints mapping within human RefSeq coordinates and within human recombination hotspots (HapMap Phase II and Perlegen data [Hinds et al. 2005]) was used to estimate an empirical *P*-value (n = 100 permutations). For gene break simulation, segmental duplications were excluded from the human genome sequences duplications due to our initial bias in selecting against these regions for gibbon BAC sequence analysis.

Evolutionary gene analyses

To determine the gene structure, human cDNA sequences and gibbon BAC sequences were aligned using ClustalW. Exon–intron boundaries were determined using the SIM4 program (Higgins et al. 1996; Florea et al. 1998). Functional annotations for each of the genes were derived from www.pantherdb.org (Mi et al. 2005). The analysis of the evolution of the coding sequence was done by maximum likelihood using PAML (Yang 1997). The ratio d_N/d_S (ω), which compares the rate of nonsynonymous substitutions against the rate of synonymous substitutions, was used as a measure of evolutionary constraint. If a gene is under no selection (neutrality), it tends to have d_N/d_S close to 1 since the ratio of fixation of synonymous and nonsynonymous mutation will be the same. However, in a situation where the gene has a strong functional role, this ratio will tend to be <1 since the nonsynonymous mutation would tend to be removed from the population because of the disturbing effect on the functional protein. Finally, positive selection (adaptive evolution) acting continuously upon the gene generates a d_N/d_S ratio >1 as the new nonsynonymous substitutions acquired will be fixed more rapidly than the almost neutral synonymous substitutions.

To perform the evolutionary analysis on the coding sequences, we first retrieved the best orthologous sequences using the Ensembl predictions for as many eutherian species as possible (ranging from five to eight species using human, chimpanzee, orangutan, gibbon, macaque, lemur, mouse, and dog). A multiple-sequence alignment was then applied (using the translated amino acids as a unit for the alignments) and back-translating into DNA sequences. All the alignments were manually curated, and regions poorly aligned were removed (although this is a conservative measure against rapid evolution, we removed particular segments that were poorly aligned in more than one species). For the gibbon sequences containing stop codons, we used the longer translatable frame in order to study the amino acid evolution of the remaining part of the gene. We then used a codon-substitution branch model (CODEML) (Yang and Nielsen 2002). First, a free codon-substitution model (in which every branch of the tree is allowed to have different d_N/d_S) was applied to the accepted phylogeny for the species to estimate the evolutionary pressures at different times during the evolution of these genes. Then, in order to have a statistical significance to gibbon-specific estimations, different evolutionary situations were modeled and compared with the initial free model. Then, we compared a codon-substitution model for the branch leading to gibbons to a neutral evolution (ω = 1) or a conserved evolution (ω = 0.5) model. Likelihood ratio tests were performed using a χ^2 distribution with as many degrees of freedom as differences of parameters in the model to estimate the significance of the comparison (Yang and Nielsen 2002).

Acknowledgments

We thank Drs. Can Alkan, Zhaoshi Jiang, and Ze Cheng for assistance with bioinformatics. We also thank Michelle O'Laughlin and Laura Courtney for help with BAC insert sequencing and Jeffrey Kidd for critical reading of the manuscript. This work was supported, in part, by an NIH grant HG002385 to E.E.E. T.M.-B. is supported by a Marie Curie fellowship. E.E.E. is an investigator of the Howard Hughes Medical Institute.

References

Abeysinghe, S.S., Chuzhanova, N., Krawczak, M., Ball, E.V., and Cooper, D.N. 2003. Translocation and gross deletion breakpoints in human inherited disease and cancer I: Nucleotide composition and recombination-associated motifs. *Hum. Mutat.* **22:** 229–244.

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

Aoki, K., Suzuki, K., Sugano, T., Tasaka, T., Nakahara, K., Kuge, O., Omori, A., and Kasai, M. 1995. A novel gene, *Translin*, encodes a recombination hotspot binding protein associated with chromosomal translocations. *Nat. Genet.* **10:** 167–174.

Armengol, L., Pujana, M.A., Cheung, J., Scherer, S.W., and Estivill, X. 2003. Enrichment of segmental duplications in regions of breaks of synteny between the human and mouse genomes suggest their involvement in evolutionary rearrangements. *Hum. Mol. Genet.* **12:** 2201–2208.

Bailey, J.A. and Eichler, E.E. 2006. Primate segmental duplications: Crucibles of evolution, diversity and disease. *Nat. Rev. Genet.* **7:** 552–564.

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

Bailey, J.A., Baertsch, R., Kent, W.J., Haussler, D., and Eichler, E.E. 2004. Hotspots of mammalian chromosomal evolution. *Genome Biol.* **5:** R23. http://genomebiology.com/2004/5/4/R23.

Barsh, G.S., Seeburg, P.H., and Gelinas, R.E. 1983. The human growth hormone gene family: Structure and evolution of the chromosomal locus. *Nucleic Acids Res.* **11:** 3939–3958.

Carbone, L., Vessere, G.M., ten Hallers, B.F., Zhu, B., Osoegawa, K., Mootnick, A., Kofler, A., Wienberg, J., Rogers, J., Humphray, S., et al. 2006. A high-resolution map of synteny disruptions in gibbon and human genomes. *PLoS Genet.* **2:** e223. doi: 10.1371/journal.pgen.0020223.

Chen, N. 2004. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* **4:** Unit 4.10.

Cheng, Z., Ventura, M., She, X., Khaitovich, P., Graves, T., Osoegawa, K., Church, D., DeJong, P., Wilson, R.K., Paabo, S., et al. 2005. A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* **437:** 88–93.

Clutton-Brock, T.H., Harvey, P.H., and Rudder, B. 1977. Sexual dimorphism, socionomic sex ratio and body weight in primates. *Nature* **269:** 797–800.

Dooley, H. and Judge, D. 2007. Vocal responses of captive gibbon groups to a mate change in a pair of white-cheeked gibbons (*Nomascus leucogenys*). *Folia Primatol. (Basel)* **78:** 228–239.

Dumas, L., Kim, Y.H., Karimpour-Fard, A., Cox, M., Hopkins, J., Pollack, J.R., and Sikela, J.M. 2007. Gene copy number variation spanning 60 million years of human and primate evolution. *Genome Res.* **17:** 1266–1277.

Eichler, E.E. 2001. Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends Genet.* **17:** 661–669.

Ferguson-Smith, M.A. and Trifonov, V. 2007. Mammalian karyotype evolution. *Nat. Rev. Genet.* **8:** 950–962.

Feuk, L., MacDonald, J.R., Tang, T., Carson, A.R., Li, M., Rao, G., Khaja, R., and Scherer, S.W. 2005. Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. *PLoS Genet.* **1:** e56. doi: 10.1371/journal.pgen.0010056.

Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M., and Miller, W. 1998. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* **8:** 967–974.

Gebo, D.L. 1996. Climbing, brachiation, and terrestrial quadrupedalism: Historical precursors of hominid bipedalism. *Am. J. Phys. Anthropol.* **101:** 55–92.

Goodman, M. 1999. The genomic record of Humankind's evolutionary roots. *Am. J. Hum. Genet.* **64:** 31–39.

Haig, D. 1999. A brief history of human autosomes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **354:** 1447–1470.

Harcourt, A.H., Harvey, P.H., Larson, S.G., and Short, R.V. 1981. Testis weight, body weight and breeding system in primates. *Nature* **293:** 55–57.

Higgins, D.G., Thompson, J.D., and Gibson, T.J. 1996. Using CLUSTAL for multiple sequence alignments. *Methods Enzymol.* **266:** 383–402.

Hinds, D.A., Stuve, L.L., Nilsen, G.B., Halperin, E., Eskin, E., Ballinger, D.G., Frazer, K.A., and Cox, D.R. 2005. Whole-genome patterns of common DNA variation in three human populations. *Science* **307:** 1072–1079.

Jauch, A., Wienberg, J., Stanyon, R., Arnold, N., Tofanelli, S., Ishida, T., and Cremer, T. 1992. Reconstruction of genomic rearrangements in great apes and gibbons by chromosome painting. *Proc. Natl. Acad. Sci.* **89:** 8611–8615.

Jiang, Z., Tang, H., Ventura, M., Cardone, M.F., Marques-Bonet, T., She, X., Pevzner, P.A., and Eichler, E.E. 2007. Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nat. Genet.* **39:** 1361–1368.

Jiang, Z., Hubley, R., Smit, A., and Eichler, E.E. 2008. DupMasker: A tool for annotating primate segmental duplications. *Genome Res.* **18:** 1362–1368.

Kanoe, H., Nakayama, T., Hosaka, T., Murakami, H., Yamamoto, H., Nakashima, Y., Tsuboyama, T., Nakamura, T., Ron, D., Sasaki, M.S., et al. 1999. Characteristics of genomic breakpoints in TLS-CHOP translocations in liposarcomas suggest the involvement of Translin and topoisomerase II in the process of translocation. *Oncogene* **18:** 721–729.

Kas, E. and Laemmli, U.K. 1992. In vivo topoisomerase II cleavage of the *Drosophila* histone and satellite III repeats: DNA sequence and structural characteristics. *EMBO J.* **11:** 705–716.

Kehrer-Sawatzki, H. and Cooper, D.N. 2007. Structural divergence between the human and chimpanzee genomes. *Hum. Genet.* **120:** 759–778.

Kehrer-Sawatzki, H. and Cooper, D.N. 2008. Molecular mechanisms of chromosomal rearrangement during primate evolution. *Chromosome Res.* **16:** 41–56.

Kehrer-Sawatzki, H., Schreiner, B., Tanzer, S., Platzer, M., Muller, S., and Hameister, H. 2002. Molecular characterization of the pericentric inversion that causes differences between chimpanzee chromosome 19 and human chromosome 17. *Am. J. Hum. Genet.* **71:** 375–388.

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

Kim, P.M., Lam, H.Y., Urban, A.E., Korbel, J.O., Chen, X., Snyder, M., and Gerstein, M.B. 2008. Analysis of copy number variants and segmental duplications in the human genome: Evidence for a change in the process of formation in recent evolutionary history. *Genome Res* **18:** 1865–1874.

Koszul, R., Caburet, S., Dujon, B., and Fischer, G. 2004. Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments. *EMBO J.* **23:** 234–243.

Lee, J.A., Carvalho, C.M., and Lupski, J.R. 2007. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell* **131:** 1235–1247.

Locke, D.P., Sharp, A.J., McCarroll, S.A., McGrath, S.D., Newman, T.L., Cheng, Z., Schwartz, S., Albertson, D.G., Pinkel, D., Altshuler, D.M., et al. 2006. Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *Am. J. Hum. Genet.* **79:** 275–290.

Marshall, C.R., Noor, A., Vincent, J.B., Lionel, A.C., Feuk, L., Skaug, J., Shago, M., Moessner, R., Pinto, D., Ren, Y., et al. 2008. Structural variation of chromosomes in autism spectrum disorder. *Am. J. Hum. Genet.* **82:** 477–488.

McVey, M., Adams, M., Staeva-Vieira, E., and Sekelsky, J.J. 2004. Evidence for multiple cycles of strand invasion during repair of double-strand gaps in *Drosophila*. *Genetics* **167:** 699–705.

Mi, H., Lazareva-Ulitsky, B., Loo, R., Kejariwal, A., Vandergriff, J., Rabkin, S., Guo, N., Muruganujan, A., Doremieux, O., Campbell, M.J., et al. 2005. The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res.* **33:** D284–D288.

Misceo, D., Capozzi, O., Roberto, R., Dell'oglio, M.P., Rocchi, M., Stanyon, R., and Archidiacono, N. 2008. Tracking the complex flow of chromosome rearrangements from the Hominoidea Ancestor to extant *Hylobates* and *Nomascus* Gibbons by high-resolution, punctuated synteny mapping. *Genome Res.* **18:** 1530–1537.

Misra-Press, A., Cooke, N.E., and Liebhaber, S.A. 1994. Complex alternative splicing partially inactivates the human chorionic somatomammotropin-like (hCS-L) gene. *J. Biol. Chem.* **269:** 23220–23229.

Morrish, T.A., Garcia-Perez, J.L., Stamato, T.D., Taccioli, G.E., Sekiguchi, J., and Moran, J.V. 2007. Endonuclease-independent LINE-1 retrotransposition at mammalian telomeres. *Nature* **446:** 208–212.

Muller, S., Hollatz, M., and Wienberg, J. 2003. Chromosomal phylogeny and evolution of gibbons (Hylobatidae). *Hum. Genet.* **113:** 493–501.

Murphy, W.J., Larkin, D.M., Everts-van der Wind, A., Bourque, G., Tesler, G., Auvil, L., Beever, J.E., Chowdhary, B.P., Galibert, F., Gatzke, L., et al. 2005. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science* **309:** 613–617.

Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. 2005. A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310:** 321–324.

Negrini, M., Felix, C.A., Martin, C., Lange, B.J., Nakamura, T., Canaani, E., and Croce, C.M. 1993. Potential topoisomerase II DNA-binding sites at the breakpoints of a t(9;11) chromosome translocation in acute myeloid leukemia. *Cancer Res.* **53:** 4489–4492.

Newman, T.L., Tuzun, E., Morrison, V.A., Hayden, K.E., Ventura, M., McGrath, S.D., Rocchi, M., and Eichler, E.E. 2005. A genome-wide

survey of structural variation between human and chimpanzee. *Genome Res.* **15:** 1344–1356.
O'Brien, S.J. and Stanyon, R. 1999. Phylogenomics. Ancestral primate viewed. *Nature* **402:** 365–366.
Parsons, J.D. 1995. Miropeats: Graphical DNA sequence comparisons. *Comput. Appl. Biosci.* **11:** 615–619.
Payen, C., Koszul, R., Dujon, B., and Fischer, G. 2008. Segmental duplications arise from Pol32-dependent repair of broken forks through two alternative replication-based mechanisms. *PLoS Genet.* **4:** e1000175. doi: 10.1371/journal.pgen.1000175.
Pevzner, P. and Tesler, G. 2003. Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc. Natl. Acad. Sci.* **100:** 7672–7677.
Plavcan, J.M. 2001. Sexual dimorphism in primate evolution. *Am. J. Phys. Anthropol.* Suppl. **33:** 25–53.
Ranz, J.M., Maurin, D., Chan, Y.S., von Grotthuss, M., Hillier, L.W., Roote, J., Ashburner, M., and Bergman, C.M. 2007. Principles of genome evolution in the *Drosophila melanogaster* species group. *PLoS Biol.* **5:** e152. doi: 10.1371/journal.pbio.0050152.
Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., et al. 2006. Global variation in copy number in the human genome. *Nature* **444:** 444–454.
Roberto, R., Capozzi, O., Wilson, R.K., Mardis, E.R., Lomiento, M., Tuzun, E., Cheng, Z., Mootnick, A.R., Archidiacono, N., Rocchi, M., et al. 2007. Molecular refinement of gibbon genome rearrangements. *Genome Res.* **17:** 249–257.
Samonte, R.V. and Eichler, E.E. 2002. Segmental duplications and the evolution of the primate genome. *Nat. Rev. Genet.* **3:** 65–72.
Sander, M. and Hsieh, T.S. 1985. *Drosophila* topoisomerase II double-strand DNA cleavage: Analysis of DNA sequence homology at the cleavage site. *Nucleic Acids Res.* **13:** 1057–1072.
Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., et al. 2007. Strong association of de novo copy number mutations with autism. *Science* **316:** 445–449.
Sharp, A.J., Mefford, H.C., Li, K., Baker, C., Skinner, C., Stevenson, R.E., Schroer, R.J., Novara, F., De Gregori, M., Ciccone, R., et al. 2008. A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. *Nat. Genet.* **40:** 322–328.
Smith, C.E., Llorente, B., and Symington, L.S. 2007. Template switching during break-induced replication. *Nature* **447:** 102–105.
Spitzner, J.R. and Muller, M.T. 1988. A consensus sequence for cleavage by vertebrate DNA topoisomerase II. *Nucleic Acids Res.* **16:** 5533–5556.
Turleau, C., Creau-Goldberg, N., Cochet, C., and de Grouchy, J. 1983. Gene mapping of the gibbon. Its position in primate evolution. *Hum. Genet.* **64:** 65–72.
Usherwood, J.R. and Bertram, J.E. 2003. Understanding brachiation: Insight from a collisional perspective. *J. Exp. Biol.* **206:** 1631–1642.
Wei, Y., Sun, M., Nilsson, G., Dwight, T., Xie, Y., Wang, J., Hou, Y., Larsson, O., Larsson, C., and Zhu, X. 2003. Characteristic sequence motifs located at the genomic breakpoints of the translocation t(X;18) in synovial sarcomas. *Oncogene* **22:** 2215–2222.
Wienberg, J. 2005. Fluorescence in situ hybridization to chromosomes as a tool to understand human and primate genome evolution. *Cytogenet. Genome Res.* **108:** 139–160.
Yan, C.T., Boboila, C., Souza, E.K., Franco, S., Hickernell, T.R., Murphy, M., Gumaste, S., Geyer, M., Zarrin, A.A., Manis, J.P., et al. 2007. IgH class switching and translocations use a robust non-classical end-joining pathway. *Nature* **449:** 478–482.
Yang, Z. 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13:** 555–556.
Yang, Z. and Nielsen, R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* **19:** 908–917.
Yunis, J.J. and Prakash, O. 1982. The origin of man: A chromosomal pictorial legacy. *Science* **215:** 1525–1530.
Zhang, J. and Webb, D.M. 2003. Evolutionary deterioration of the vomeronasal pheromone transduction pathway in catarrhine primates. *Proc. Natl. Acad. Sci.* **100:** 8337–8341.
Zody, M.C., Jiang, Z., Fung, H.C., Antonacci, F., Hillier, L.W., Cardone, M.F., Graves, T.A., Kidd, J.M., Cheng, Z., Abouelleil, A., et al. 2008. Evolutionary toggling of the MAPT 17q21.31 inversion region. *Nat. Genet.* **40:** 1076–1083.

Received September 2, 2008; accepted in revised form November 17, 2008.

Sequencing human–gibbon breakpoints of synteny reveals mosaic new insertions at rearrangement sites

Santhosh Girirajan, Lin Chen, Tina Graves, et al.

Genome Res. 2009 19: 178-190 originally published online November 24, 2008
Access the most recent version at doi:10.1101/gr.086041.108