Rates and patterns of great ape retrotransposition

Fereydoun Hormozdiari[a], Miriam K. Konkel[b], Javier Prado-Martinez[c], Giorgia Chiatante[d], Irene Hernando Herraez[c], Jerilyn A. Walker[b], Benjamin Nelson[a], Can Alkan[e], Peter H. Sudmant[a], John Huddleston[a], Claudia R. Catacchio[d], Arthur Ko[a], Maika Malig[a], Carl Baker[a], Great Ape Genome Project[a,c,1], Tomas Marques-Bonet[c,f], Mario Ventura[d], Mark A. Batzer[b], and Evan E. Eichler[a,g,2]

[a]Department of Genome Sciences, University of Washington, Seattle, WA 98195; [b]Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803; [c]Institut de Biologia Evolutiva, Spanish National Research Council–Universitat Pompeu Fabra, Barcelona, 08003 Catalonia, Spain; [d]Department of Biology, University of Bari, 70126 Bari, Italy; [e]Department of Computer Engineering, Bilkent University, Ankara, 06800 Turkey; [f]Institució Catalana de Recerca i Estudis Avançats, Barcelona, 08010 Catalonia, Spain; and [g]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195

Contributed by Evan E. Eichler, June 15, 2013 (sent for review May 8, 2013)

We analyzed 83 fully sequenced great ape genomes for mobile element insertions, predicting a total of 49,452 fixed and polymorphic *Alu* and long interspersed element 1 (L1) insertions not present in the human reference assembly and assigning each retrotransposition event to a different time point during great ape evolution. We used these homoplasy-free markers to construct a mobile element insertions-based phylogeny of humans and great apes and demonstrate their differential power to discern ape subspecies and populations. Within this context, we find a good correlation between L1 diversity and single-nucleotide polymorphism heterozygosity (r^2 = 0.65) in contrast to *Alu* repeats, which show little correlation (r^2 = 0.07). We estimate that the "rate" of *Alu* retrotransposition has differed by a factor of 15-fold in these lineages. Humans, chimpanzees, and bonobos show the highest rates of *Alu* accumulation—the latter two since divergence 1.5 Mya. The L1 insertion rate, in contrast, has remained relatively constant, with rates differing by less than a factor of three. We conclude that *Alu* retrotransposition has been the most variable form of genetic variation during recent human–great ape evolution, with increases and decreases occurring over very short periods of evolutionary time.

genomics | genetic diversity | structural variation | retrotransposon

Mobile elements comprise ~50% of our genetic code. Among these, *Alu* (a primate-specific short interspersed element, SINE) and L1 repeats (a long interspersed element, LINE) are the most abundant (1, 2). Both elements propagated in the germ line as a result of target primed reverse transcription (TPRT) using an AP-endonuclease and reverse transcriptase activities encoded by L1 elements (3–5). These integrations—termed "mobile element insertions" (MEIs)—have the potential to disrupt genes, alter transcript expression and splicing, as well as promote genomic instability as a result of nonallelic homologous recombination (6–9). In addition, these MEIs are powerful phylogenetic (10–12) and population genetic markers (13–16) because they are generally regarded as homoplasy-free character states—i.e., precise excision is an exceedingly rare event and, as such, the ancestral and derived state can be unambiguously determined (17–20).

Critical to our understanding of MEI impact with respect to disease and evolution is a detailed assessment of changes in retrotransposition activity within different lineages (21). Genome sequencing comparisons have been used as one method to infer differences in activity between humans and great apes (22–26). There are several important limitations of previous studies. First, genome-wide assessments are generally incomplete because of their dependence on a single representative genome from each species, where consequently the fixed versus polymorphic status of most MEIs is not known. Second, published great ape genome assemblies vary considerably in quality and completeness. The gorilla genome, for example, was assembled primarily from Illumina sequencing data and consists of over 433,000 gaps. Many of the gaps over 100 bp in length (n = 192,481) map to MEIs and segmental duplications (25, 27). Finally, for those lineages for which there are rate estimates, these rates differ between some studies by more than a factor of two (24, 28, 29). Some of these discrepancies arise from methodological differences in discovery and limited genomic sampling (e.g., some experimental studies have focused on a relatively small number of MEIs) (22, 23). To date, there is no genome-wide synthesis of changes in rates, particularly as they relate to single-nucleotide substitution.

Here we present a genome-wide discovery and synthesis of differences in the accumulation of L1 and *Alu* elements during the course of human–great ape evolution. We leverage deep sequence data generated from 83 hominid genomes along with 10 additional human genomes in an attempt to maximize our understanding about the diversity of the different species and populations. Our results more than triple the number of known polymorphic MEIs in great apes, including the discovery of ancestry-informative markers and MEIs corresponding to regions of incomplete lineage sorting (ILS). Such ILS segments define regions where the gene genealogy differs from that of the species phylogeny due to rapid speciation or hybridization and are relatively rare, especially as defined by the MEI (19, 30, 31). The availability of single-nucleotide polymorphism (SNP) data (32) from the same individual genomes allows us to more accurately estimate rate changes in *Alu* and L1 retrotransposition in contrast to single-nucleotide accumulation and to compare the utility of these markers in reconstructing the evolutionary relationships of our species.

Results

To discover MEIs, we applied a read-pair mapping approach to 83 genomes sequenced for the Great Ape Diversity Project (32) as well as 10 diverse human genomes for a total of 93 genomes. Genomes were sequenced to an average depth of 23-fold sequence coverage from samples that included all ape species and representatives from 10 recognized subspecies (Table 1). We mapped the paired-end reads of these genomes to the human reference genome (GRCh36) using mrsFAST (33) and predicted *Alu* and L1 insertions using an extension of our previously described algorithm (34, 35). We discovered a total of 24,210 *Alu* and 25,242 L1 great ape insertions compared with the human reference genome. We estimate that these correspond to 13,600 new *Alu* and 17,000 new L1 insertions compared with previously published ape reference genomes (24–26, 28).

We also performed a reciprocal analysis identifying insertions in the human genome that corresponded to precise "deletions" among the great apes. We identified 11,770 *Alu* and 8,428 L1 elements, assigning these to different branch points during

Author contributions: F.H., T.M.-B., M.A.B., and E.E.E. designed research; F.H., M.K.K., J.P.-M., G.C., I.H.H., J.A.W., C.R.C., M.M., C.B., T.M.-B., M.V., and E.E.E. performed research; F.H., M.K.K., J.P.-M., G.C., B.N., C.A., P.H.S., J.H., A.K., M.V., M.A.B., and E.E.E. analyzed data; and F.H., M.V., M.A.B., and E.E.E. wrote the paper.

The authors declare no conflict of interest.

Data deposition: Human mobile element insertions (MEIs) can be accessed from http://eichlerlab.gs.washington.edu/greatape-MEI/.

[1]A complete list of the Great Ape Genome Project can be found in the Supporting Information.

[2]To whom correspondence should be addressed. E-mail: eee@gs.washington.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1310914110/-/DCSupplemental.

Table 1. Summary of mobile element insertions in great ape genomes

Species/subspecies	*N*	Fold coverage	Nonreference *Alu* insertion statistics (vs. GRCh36)			Nonreference L1 insertion statistics (vs. GRCh36)		
			Discovery	Genotyping		Discovery	Genotyping	
			Total no. Ins	Fix	Polymorphic	Total no. Ins	Fix	Polymorphic
Pan	35	1,028	11,157	—	—	7,215	—	—
Pan paniscus	12	444	4,229	2,607	1,540	3,067	1,877	1,082
Pan troglodytes	23	584	8,715	1,971	6,403	6,066	1,292	4,578
Pan troglodytes troglodytes	4	130	5,570	2,807	2,763	3,326	1,900	1,426
Pan troglodytes ellioti	8	98	5,341	3,242	1,779	3,951	2,169	1,638
Pan troglodytes verus	5	131	4,129	2,864	1,150	2,958	1,842	1,041
Pan troglodytes schweinfurthii	6	225	5,607	3,002	1,875	3,697	2,146	1,116
*Homo sapiens**	10	158	2,932	127	2,805	448	35	413
Gorilla	35	830	8,809	3,309	5,127	5,059	1,711	2,937
Gorilla gorilla gorilla	32	753	8,445	3,228	4,791	4,686	1,708	2,563
Gorilla beringei graueri	2	53	5,382	—	—	2,748	—	—
Gorilla gorilla diehli	1	24	4,491	—	—	2,295	—	—
Pongo	13	446	1,739	974	765	13,410	5,800	7,610
Pongo abelii	6	217	1,666	1,267	399	11,378	7,235	4,143
Pongo pygmaeus	7	229	1,571	1,066	505	10,797	6,935	3,862
Total	93	2462	24,210	10,399	13,175	25,242	13,368	11,417

Discovery was based on the analysis of 93 genomes; genotyping status of fixed versus polymorphic was restricted to 72 genomes with the highest coverage and best insertion size distributions. Cov, coverage; Ins, insertions.

*A total of 7,041 *Alu* insertions and 1,488 L1 insertions in human lineage that also exist in GRCh36.

human evolution based on their presence or absence within different great ape lineages. Note that some of the samples used to predict MEIs have lower coverage sequencing and, thus, for most of the analysis where accurate genotyping is critical, we limited the analysis to 72 genomes with the highest coverage and best insert-size distributions (*SI Appendix*, Table S2). Our analysis of these 72 samples identified 187 MEI events (43 *Alu* and 144 L1 events) that were inconsistent with the great ape phylogeny (e.g., shared between human and gorilla but not chimpanzee). A total of 84% (157/187) of these loci were also flanked by SNPs of a similar phylogeny, confirming that most arose as a result of ILS (Dataset S1). Comparing our results with regions identified by ILS in these same genomes (32), we determined the average length of an ILS segment harboring an MEI marker to be ∼7 kbp in length. Because we sequenced multiple genomes from each species and subspecies (with the exception of the Cross River gorilla), we also classified MEIs as fixed (i.e., not polymorphic) if they were predicted to be seen in all of the samples of a species or subspecies with more than 90% probability, assuming a false genotyping rate of 10% (Table 1). We note that false genotyping rates of less than 10% for MEIs and structural variation discovery is relatively standard using high-throughput technologies (34, 36, 37). Increased genotyping error arises from the mapping of discordant short sequence reads to common repeats that exist at multiple locations in the genome and is exacerbated by cross-species mappings, which are required due to the lower quality of non-human reference genomes. To minimize potential genotyping biases, we, once again, restricted the assignment of fixed versus polymorphic status to those samples ($n = 72$) with the highest sequence coverage and the best insert-size distributions.

We validated the quality of our MEI predictions and genotypes by three independent analyses. First, we compared our predicted MEIs for the Western chimpanzee Clint to the chimpanzee reference genome (PanTro3), which was previously assembled from Sanger sequence data generated from the same donor (28). We predicted a total of 3,230 *Alu* and 2,317 L1 insertions based on paired-end read mapping of Clint Illumina data to the human reference (GRCh36). Among those that we could successfully cross reference using LiftOver (38), we found that 85.5% (2,396/2,802) and 84.5% (1,651/1,953) of our *Alu* and L1 read-pair predicted insertions, respectively, matched PanTro3 assembly insertions. Concordance rises to 90% and 91% for *Alu* and L1 insertions, respectively, if we exclude insertions that map to repetitive DNA, such as segmental duplications (*SI Appendix*, Fig. S3). Next, we randomly selected 13 *Alu* insertions and 9 L1 insertions and performed PCR on DNA from eight sequenced samples (three chimpanzees, one bonobo, three gorillas, and one orangutan). We observed a genotyping concordance of >95% (86/90) and 98% (>54/55) for *Alu* and L1 insertions, respectively (*SI Appendix*, Figs. S6–S8). Finally, we specifically selected MEIs that appeared ancestry informative (i.e., detected in one ape subspecies to the exclusion of others) and MEIs that showed ILS among chimpanzee, bonobo, gorilla, and human. We designed a total of 118 successful PCR assays with an overall validation rate of 96.6% (*SI Appendix*, Table S5). Of the validated ILS events, 47 were used to determine the precise breakpoint of the insertion. Using multiple alignments suggests that breakpoints are accurately predicted with an average interval of 32 bp (see *SI Appendix* for details). Sequencing revealed that 93.6% (44/47) of the sequenced insertions corresponded to younger subfamilies and carried target-site duplications diagnostic of recent retrotransposition events (*SI Appendix*, Figs. S19 and S20).

The map location of all MEIs was annotated on the human reference genome, including elements that mapped near or within exons of genes. A total of 17,468 MEIs mapped within the introns of genes (Fig. 1*A*). As expected, L1 insertions were significantly depleted in genic regions, whereas the *Alu* density was as expected by chance (39). We observed a strong bias against L1 and *Alu* insertions within protein-coding sequence ($P < 1e\text{-}40$; Fig. 1*B*). The reduction was also true for MEIs in untranslated regions (UTRs) but was less significant. In total, we identified only 10 MEIs that are predicted potentially to disrupt the protein-coding regions of genes, although a total of 160 MEIs intersect with genic UTRs (*SI Appendix*, Tables S3 and S4). Similarly, we observed a bias for both *Alu*s and L1s to be inserted in an antisense orientation when mapping within genes (40) (Fig. 1*C*). We also specifically analyzed the GC composition of *Alu* insertions due to the reported shift in GC bias (28). All lineage-specific *Alu* insertions show a stronger bias against GC-rich DNA compared with ancestral events (Fig. 1*D* and *SI Appendix*, Fig. S23). This transition to more GC-rich DNA occurs relatively gradually, with significant increases observed in the

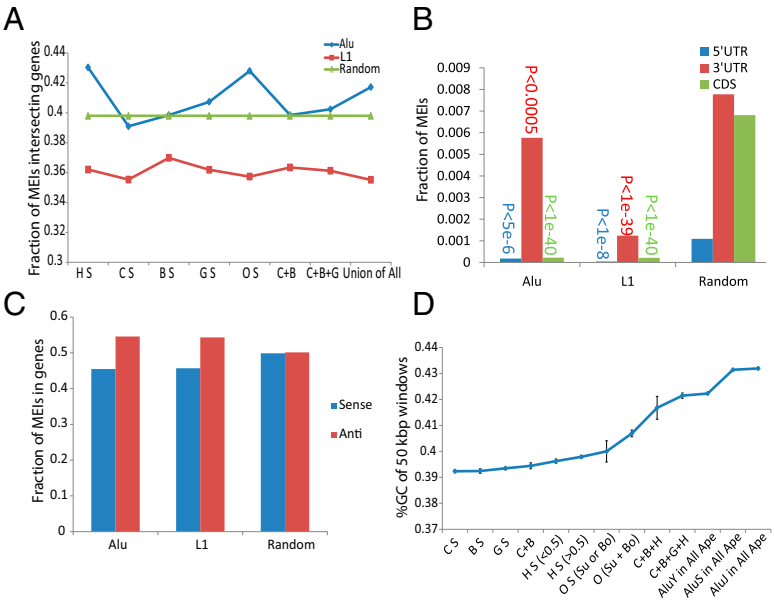

Fig. 1. Insertion bias. (*A*) Fraction of *Alu* (blue) and L1 (red) insertions that map within genic regions compared with a random distribution (green). Insertions are classified as species specific (S) or shared between chimpanzee (C), bonobo (B), human (H), gorilla (G), and orangutan (O). (*B*) Fraction of *Alu* and L1 insertions that map within protein-coding sequence (CDS) and untranslated regions (UTRs) of genes based on RefSeq. A significant bias is observed against each based on a random insertion model. (*C*) Significant bias is also observed against insertions in the sense orientation of gene transcription. (*D*) GC content of *Alu* insertion events classified by phylogenetic category and sorted by increasing GC content. Human-specific events are distinguished based on allele frequency (±50% frequency), whereas orangutan-specific insertions are grouped as shared or specific to one of the species (Su, Sumatran or Bo, Bornean).

chimpanzee–human ancestral and chimpanzee–human–gorilla ancestral lineages (6–8 Mya).

We investigated the correlation between SNP heterozygosity with *Alu* and L1 diversity for each subspecies. Since the same genomes have been completely sequenced and SNPs identified, this represents one of the first times that MEI and SNP diversity can be comprehensively compared in the same samples. We computed SNP heterozygosity (32) as the ratio of heterozygous SNPs over the length of the genome and compared it to the average *Alu* and L1 differences between samples within a given population or subspecies (allele sharing method) (22). Interestingly, we find almost no correlation between *Alu* insertion diversity in great apes with SNP heterozygosity ($r^2 = 0.07$, $P = 0.44$), whereas L1 insertion diversity has moderate correlation with SNP heterozygosity ($r^2 = 0.65$, $P = 0.0025$) (*SI Appendix*, Figs. S15 and S16).

Using the MEIs as genetic markers, we constructed both neighbor-joining (41) and Unweighted Pair Group Method with Arithmetic Mean phylogenetic trees for humans and great apes and compared them to a phylogeny constructed from single-nucleotide variants generated from the same great ape genomes (Fig. 2 and *SI Appendix*, Fig. S10) (32). One of the advantages of a phylogenetic tree constructed using MEIs is that there are, in principle, no backward mutations or revertants; as such, the absence of an MEI insertion is ancestral and the presence of an insertion indicates identity by descent. For the purpose of this study, we limited this analysis to 72 genomes with the highest coverage and largest insert-size distributions to avoid potential ascertainment biases in discovery and ensure the most accurate genotype for each genome (*SI Appendix*, Table S2). The general topology of the human–great ape phylogeny is remarkably consistent; both *Alu* (Fig. 2*B*) and L1 (Fig. 2*C*) trees show 100% bootstrap support for the separation of all known great ape species, and there is strong bootstrap support for the separation of the four chimpanzee subspecies. However, the phylogenetic analysis also suggests that different MEIs confer different resolution, especially among terminal branches depending on the great ape population. Whereas *Alu* insertions robustly distinguish populations of human, chimpanzee, and gorilla, we were unable to discriminate the different species of orangutan based on an *Alu* insertion phylogeny. This is in sharp contrast to L1s, which not only clearly distinguish Bornean and Sumatran species but also suggest distinct subpopulations with 100% bootstrap support within this primate lineage.

A principal component analysis (PCA) provides insight into additional substructure within different ape populations (Fig. 3 and *SI Appendix*, Figs. S11–S14). The PCA analysis combining both *Alu* and L1 insertions clearly separates the four chimpanzee subspecies and identifies one Nigerian chimpanzee (Julie) as an outlier. Notably, the same individual was identified as an outlier in the PCA analysis using SNPs (*SI Appendix*, Fig. S13) (32). It is interesting that the first principal component (PC1) based on *Alu* insertions distinguishes two groups of chimpanzee: western-Nigerian from central-eastern. A similar result is seen for PCA from SNPs but not L1 insertions. Although there is no information available on the geographic origin of the bonobos, there is evidence of a reproducible clustering of individuals based on *Alu*, L1, and SNP PCA (32). For the gorilla, PC1 from both L1s and *Alu*s separates Eastern lowland gorillas from western lowland gorillas. Notably and also supported by the SNP data, a PCA analysis of western lowland gorilla samples using the combination of *Alu* and L1 insertions shows a gradient along PC2 consistent with their country of origin (i.e., Congo or Cameroon). All PCA analyses separate Sumatran and Bornean orangutans (with the exception of Kiki) along PC1. Additional substructure is observed for the Borneans along PC2 for both *Alu* and L1 insertions, which are not observed with SNP data.

The availability of multiple deeply sequenced ape genomes allowed us to unambiguously assign *Alu* and L1 insertions to terminal and ancestral branches along the human–great ape phylogeny. We used this information to estimate the rate of *Alu* and L1 insertion accumulation at different times during evolution. For the purpose of rate calculations, we limited our analysis to 10 genomes per species (human, chimpanzee, bonobo, gorilla, and orangutan) with the highest sequence quality to avoid potential artifacts that might arise from lower coverage. We computed the rate of insertion per lineage per million years (based on estimated species divergences) and normalized the number of average insertions per million SNPs for each branch of great ape evolution (Fig. 4 *A* and *B*). The latter has the advantage of eliminating the inherent uncertainty associated with species divergence and replacing it with a neutral genetic distance estimator. In addition, normalization of the MEI rates against SNP rates for each lineage gives us the ability to control for demographic effects (such as population size) in each lineage. Such demographic changes should affect MEI and single-nucleotide substitutions equally helping to eliminate skews that could be introduced by a simple ultrametric-based approach.

Our results quantify *Alu* and L1 accumulation across the great ape phylogeny, revealing radical differences in *Alu* and L1 insertion activity along each branch (Fig. 4). Changes in *Alu* activity appear to be most dramatic, differing by a factor of 15-fold over a few million years. The orangutan lineage and ancestral

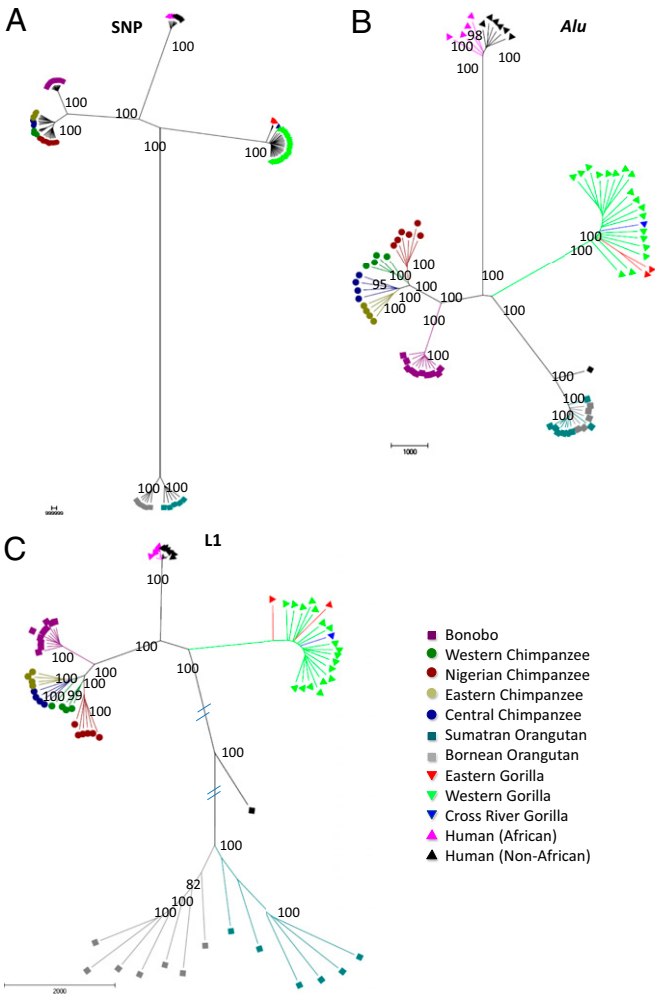

Fig. 2. Ape phylogeny. Neighbor-joining trees constructed from (*A*) SNP, (*B*) *Alu*, and (*C*) L1 insertions define the genetic relationship among species and subspecies of great apes. Only the bootstrap scores above 80% are shown. MEI neighbor-joining trees were constructed based on the presence of *Alu* or L1 elements and represent a subset ($n = 72$) of samples analyzed for SNPs (32).

human–chimpanzee branch show the lowest rate (27–45 *Alu*s per million SNPs), whereas the terminal African ape branches all show an increase (317–421 *Alu*s per million SNPs). Interestingly, the number of fixed *Alu* insertions in the chimpanzee lineage is reduced by 4-fold compared with those fixed on the human lineage (5,567 vs. 1,350). Most of these differences appear to be the result of far fewer insertion events before divergence of chimpanzee and bonobo.

In contrast to the *Alu* insertions, the accumulation of L1 elements has been more constant and clocklike with an average rate of 141 L1 insertions per million SNPs. Our analysis suggests a maximum rate difference of 2.7-fold. We observe the highest rate of accumulation in the common ancestor of human and African apes (241 L1 insertions per million SNPs) followed by the orangutan lineage, which approximates the ancestral rate (180 L1s per million SNPs). African great ape lineages, in general, show a continual decline with human terminal branch representing the nadir (90 L1s per million SNPs). In general, there is a weak negative correlation with *Alu* and L1 activity ($r = -0.409$, $P = 0.31$), which is slightly more prominent for terminal branches of the great ape phylogeny ($r = -0.5578$, $P = 0.32$) (*SI Appendix*, Figs. S21 and S22).

Discussion

Our analyses of deep genome sequence data from 83 great apes has provided one of the most comprehensive surveys of *Alu* and L1 genetic variation among any group of closely related mammalian species to date. Although our results more than triple the number of lineage-specific *Alu* and L1 insertions known for humans and great apes, the census is not yet complete even for the genomes we have sequenced. The short-read nature of next-generation sequencing limits our ability to map insertions near or within repetitive sequences (34, 37). Our comparison with the chimpanzee reference genome suggests a false negative of ∼30%, which reduces to 9% when excluding MEIs that map within segmental duplications and other common repeats. In contrast to other forms of structural variation and SNPs, *Alu*s and L1s do not appear to have been a potent force disrupting the protein-coding regions of genes. We have identified only 10 possible events in contrast to the more than 1,886 predicted loss-of-function mutations that arose and fixed as a result of insertion/deletion and single-nucleotide substitution during great ape evolution (32). We note that further examination reveals that three of our predicted protein disruptions actually map adjacent to an exon (based on chimpanzee and gorilla genome data), four are restricted to a single sample and are not fixed, and the final three are seen in genes in the orangutan which have undergone segmental duplication but are a single copy in humans. We conclude that L1 and *Alu* repeats have contributed minimally to gene loss by way of disrupting protein-encoding ORFs during evolution, likely because such events are deleterious and eliminated by purifying selection (42). Note, the relative number of MEI to SNP/indel gene-disruptive mutations observed between ape genomes is concordant with the ratio of disease-causing mutations reported for humans in the literature (i.e., a factor of ∼350- to 1,000-fold more SNP and indel mutations versus MEIs) (43, 44).

Both L1 and *Alu* markers recapitulate the generally accepted phylogeny of humans and great apes well, including strong evidence for separation of chimpanzee into four distinct subspecies. Because these markers are not subject to homoplasy (18), the root can be unequivocally identified in each tree unlike other forms of genetic variation. Our results, similar to data from SNP analysis (32), suggest a bipartite division of *Pan troglodytes* where western and Nigerian chimpanzees form one group and central and eastern chimpanzees represent another group. Interestingly, western chimpanzees show the greatest difference (both PC1 and PC2 distinguish this group based on PCA) from other chimpanzees, harbor the greatest number of ancestry-informative MEIs, and show the lowest genetic diversity—all consistent with a population that has experienced a strong bottleneck. Compared with single-nucleotide markers, we also observe significant distortions in branch length along the primate tree consistent with differences in MEI activity at different time points during evolution as well as the fact MEI polymorphisms sample deeper aspects of the species genealogy (16). Among pongids, for example, L1 insertions appear to provide additional resolving power to distinguish subpopulations of orangutan (but not chimpanzee or gorilla). Based on our limited sampling of several genomes, we have now defined 9,000 markers that are unique to a specific subspecies of chimpanzee or gorilla and over 40,000 markers specific to each species or lineage. These represent an important resource for testing of a much larger cohort of samples from different subspecies to define ancestry-informative markers for population genetics and conservation purposes.

Our analysis suggests that the accumulation of MEIs has varied substantially on different branches of the human–great ape phylogenetic tree. Our estimated rates differ from some previous studies (24, 26, 29) (as we report significantly more *Alu* and L1 insertions for chimpanzee, bonobo, and orangutan) but are consistent with others, such as those reporting a 2.2-fold excess of *Alu* insertions along the human branch compared with chimpanzee (22, 23) or a 1.22-fold excess of *Alu* insertions comparing bonobo to chimpanzee (24). Overall, *Alu* repeats show the most dramatic changes over the shortest time intervals with rates of accumulation differing by 15-fold and varying significantly among all branches ($P = 1.03 \times 10^{-16}$). By comparison, differences in L1 accumulation are much more modest and gradual (two- to threefold). Not surprisingly, there is a good correlation between L1 diversity and SNP heterozygosity for each branch, whereas no such correlation exists for

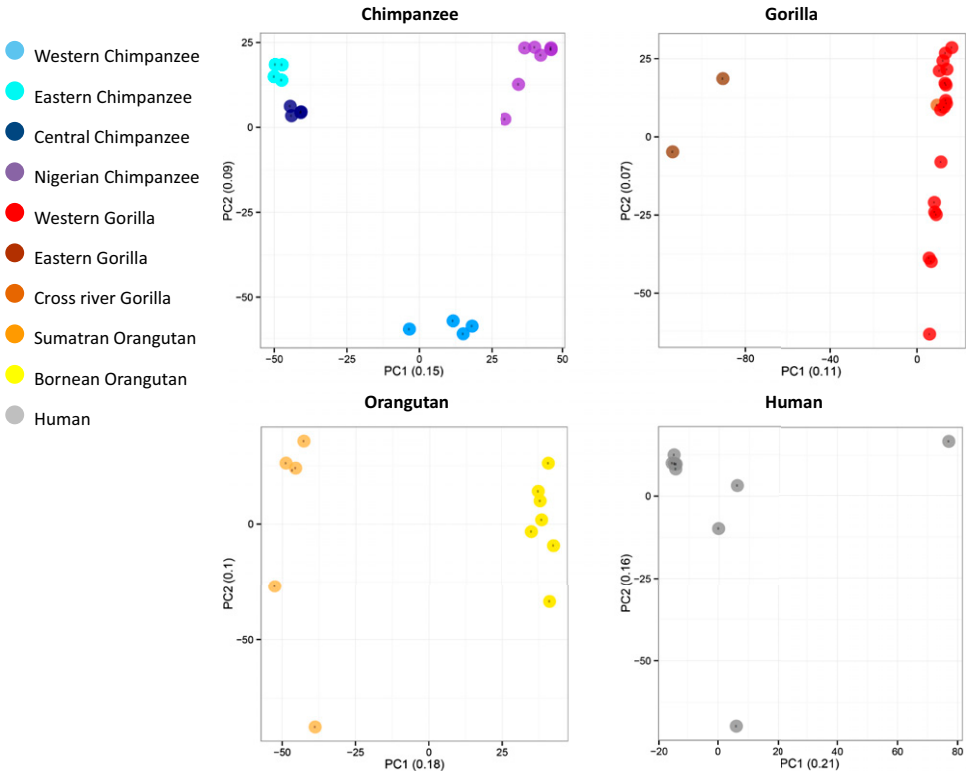

Fig. 3. PCA. Principal component analysis using merged *Alu* and L1 insertions events on GRCh36 is depicted for chimpanzee, gorilla, orangutan, and human. Names indicate individual genomes sequenced.

*Alu*s. Our results suggest that humans, chimpanzees, and bonobos all experienced an increase of *Alu* accumulation (independently on both branches) compared with the African ancestral branch or the orangutan lineages. The human lineage shows the most notable decline in L1 accumulation in contrast to chimpanzees where L1 activity apparently doubled after divergence from bonobos and chimpanzees but before the divergence of chimpanzee subspecies. L1 activity is also significantly higher in the orangutan, African great ape, and human–chimpanzee ancestral lineages ($P = 1.67 \times 10^{-6}$). Rates of *Alu* accumulation generally reciprocate those of L1s —i.e., when L1 rates are high, *Alu* accumulation is low (compare, for example, the orangutan and human lineage). This inverse relationship is thought to arise, in part, from the competition of SINEs and LINEs for the same reverse transcription machinery, although the effect of this interaction on the rate of insertion is both controversial and not well understood (43, 44).

Our analysis suggests a more complicated model—one in which *Alu* activity shows much more volatility in contrast to either L1 retrotransposition or single base pair substitution over the course of great ape evolution. One possibility may be that *Alu* activity is more susceptible to mutations that significantly dampen or improve their ability to retrotranspose or overcome cellular control mechanisms. This has been postulated to explain the sudden rise of *Alu* Ya5 and Yb8 events on the human lineage since separation from chimpanzee (28) as well as other bursts of activity along the primate lineage (45). We investigated this possibility by attempting to classify the various lineage-specific and ancestral MEIs into subfamilies based on a reanalysis of the short-read sequence data. Our results revealed that the resurgence of *Alu* Y mobile elements was driven by distinct subfamilies in human and chimpanzee. As expected, *Alu* Ya5 and Yb8 predominate in the human lineage, whereas *Alu* Yc1 and Yc2 predominate in both the bonobo and chimpanzee lineages. In contrast, when we perform this analysis for L1 retroposons, we find that both the human and chimpanzee lineages show fewer distinctive subfamilies with L1PA2 being most abundant followed by L1-Hs and L1-Ptr in the human and chimpanzee lineages, respectively. Whereas additional high-quality sequence for these nonreference insertions is required for a more detailed analysis, these data are consistent with L1 subtype and activity being more constant during (at least) recent primate evolution, whereas *Alu*

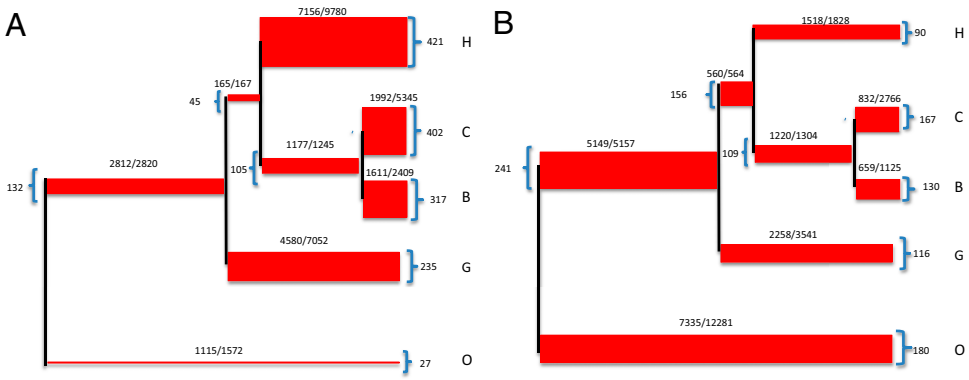

Fig. 4. Insertion rates. (*A*) Statistics of *Alu* insertions on each branch of speciation. The two numbers on each branch represent the average number of insertions and total number of insertions. The number near the bracket represents normalized average number of insertions over one million SNPs on each branch. The length of each branch is proportional to the SNP divergence in each branch and the width of each branch is proportional to the normalized average number of insertions over SNP divergence (rate of insertions). (*B*) Statistics of L1 insertions on each branch.

activity is more dependent upon changes in the competitive potential of the source elements that arise within specific lineages.

Materials and Methods

Data and Samples. All of the raw sequencing data were generated by the Great Ape Genome Project (32) and have been deposited into the Sequence Read Archive (SRA accession no. PRJNA189439/SRP018689). Mobile element insertions (MEIs) predicted for human and other great apes can be accessed online at http://eichlerlab.gs.washington.edu/greatape-MEI/.

MEI Discovery. MEI discovery is based on a VariationHunter paired-end mapping (PEM) strategy, as described previously (34, 35, 46). We used two different approaches to classify recent MEIs within the human reference as well as MEIs that do not exist within the human reference genome (GRCh36). For nonreference MEIs, we considered all PEM where one end maps to the reference genome and the other maps to a consensus set of *Alu* and L1 elements (RepeatMasker). For characterization of *Alu* and L1 insertions that do exist in the reference genome, we used the discordant read mappings from great ape genomes to identify a deletion that precisely specified an *Alu* or L1 in a specific genome or lineage. See *SI Appendix* for additional details.

Polymorphism Analysis. For each lineage we can calculate the likelihood of an event being fixed given the number of samples we have seen in the insertion. We define an insertion to be fixed if its likelihood of it being fix is >90% (with assumption of genotyping error of 10%). More formally for a lineage with n samples we first find the largest k such that for an event which is seen in k or more samples, the following equation holds

$$P(\text{seen } in \geq k | \text{fixed}) \approx \sum_{i=k}^{n} \binom{n}{i} 0.9^i 0.1^{n-i} > 0.9.$$

Then we assume any insertion which is seen in more than k samples of this lineage to be fixed insertion in that lineage.

PCR Validation. For genotyping accuracy calculation we designed PCR primers ~220 bp proximal and distal to the predicted *Alu* and L1 insertion breakpoints. We expected to see an amplification product of ~440 bp. In cases where we observed ~740-bp fragments (440 + 300 bp for *Alu*), we considered the prediction as validated. For L1s the increased fragment size can vary and the value used is the length of insertion predicted. The details of PCR validation for the ILS set are explained in the *SI Appendix*.

ACKNOWLEDGMENTS. We thank all those who generously provided the samples and sequence data for the Great Ape Genome Diversity Project, as well as the members of the consortium, especially Mikkel Schierup and Thomas Mailund for early access to incomplete lineage sorting SNP data. This work was supported by National Institutes of Health (NIH) Grant HG002385 (to E.E.E.), NIH R01 Grant GM59290 (to M.A.B.), an European Research Council Starting Grant (260372) (to T.M.-B.), and Ministerio de Ciencia e Innovacion (MICINN - Spain) BFU2011-28549 (to T.M.-B.). P.H.S. is supported by a Howard Hughes International Student Fellowship; E.E.E. is an Investigator of the Howard Hughes Medical Institute; and T.M.-B. is a Research Investigator (Institut Catala d'Estudis i Recerca Avancats de la Generalitat de Catalunya).

1. Lander ES, et al.; International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409(6822):860–921.
2. Moran JV, et al. (1996) High frequency retrotransposition in cultured mammalian cells. *Cell* 87(5):917–927.
3. Feng Q, Moran JV, Kazazian HH, Jr., Boeke JD (1996) Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* 87(5): 905–916.
4. Kazazian HH, Jr., Goodier JL (2002) LINE drive: Retrotransposition and genome instability. *Cell* 110(3):277–280.
5. Dewannieux M, Esnault C, Heidmann T (2003) LINE-mediated retrotransposition of marked Alu sequences. *Nat Genet* 35(1):41–48.
6. Kazazian HH, Jr. (2004) Mobile elements: Drivers of genome evolution. *Science* 303 (5664):1626–1632.
7. Cordaux R, Batzer MA (2009) The impact of retrotransposons on human genome evolution. *Nat Rev Genet* 10(10):691–703.
8. Han K, et al. (2008) L1 recombination-associated deletions generate human genomic variation. *Proc Natl Acad Sci USA* 105(49):19366–19371.
9. Kazazian HH, et al. (1988) Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature* 332(6160): 164–166.
10. Roos C, Schmitz J, Zischler H (2004) Primate jumping genes elucidate strepsirrhine phylogeny. *Proc Natl Acad Sci USA* 101(29):10650–10654.
11. Schmitz J, Ohme M, Zischler H (2001) SINE insertions in cladistic analyses and the phylogenetic affiliations of Tarsius bancanus to other primates. *Genetics* 157(2): 777–784.
12. Murata S, Takasaki N, Saitoh M, Okada N (1993) Determination of the phylogenetic relationships among Pacific salmonids by using short interspersed elements (SINEs) as temporal landmarks of evolution. *Proc Natl Acad Sci USA* 90(15):6995–6999.
13. Watkins WS, et al. (2003) Genetic variation among world populations: Inferences from 100 Alu insertion polymorphisms. *Genome Res* 13(7):1607–1618.
14. Stoneking M, et al. (1997) Alu insertion polymorphisms and human evolution: Evidence for a larger population size in Africa. *Genome Res* 7(11):1061–1071.
15. Batzer MA, et al. (1994) African origin of human-specific polymorphic Alu insertions. *Proc Natl Acad Sci USA* 91(25):12288–12292.
16. Huff CD, Xing J, Rogers AR, Witherspoon D, Jorde LB (2010) Mobile elements reveal small population size in the ancient ancestors of Homo sapiens. *Proc Natl Acad Sci USA* 107(5):2147–2152.
17. Batzer MA, Deininger PL (2002) Alu repeats and human genomic diversity. *Nat Rev Genet* 3(5):370–379.
18. Xing J, et al. (2005) A mobile element based phylogeny of Old World monkeys. *Mol Phylogenet Evol* 37(3):872–880.
19. Ray DA, Xing J, Salem AH, Batzer MA (2006) SINEs of a nearly perfect character. *Syst Biol* 55(6):928–935.
20. van de Lagemaat LN, Gagnier L, Medstrand P, Mager DL (2005) Genomic deletions and precise removal of transposable elements mediated by short identical DNA segments in primates. *Genome Res* 15(9):1243–1249.
21. Deininger PL, Batzer MA (1999) Alu repeats and human disease. *Mol Genet Metab* 67(3):183–193.
22. Hedges DJ, et al. (2004) Differential alu mobilization and polymorphism among the human and chimpanzee lineages. *Genome Res* 14(6):1068–1075.
23. Liu G, et al.; NISC Comparative Sequencing Program (2003) Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. *Genome Res* 13(3):358–368.
24. Prüfer K, et al. (2012) The bonobo genome compared with the chimpanzee and human genomes. *Nature* 486(7404):527–531.
25. Ventura M, et al. (2011) Gorilla genome structural variation reveals evolutionary parallelisms with chimpanzee. *Genome Res* 21(10):1640–1649.
26. Locke DP, et al. (2011) Comparative and demographic analysis of orang-utan genomes. *Nature* 469(7331):529–533.
27. Scally A, et al. (2012) Insights into hominid evolution from the gorilla genome sequence. *Nature* 483(7388):169–175.
28. Mikkelsen TS, et al. (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437(7055):69–87.
29. Mills RE, et al. (2006) Recently mobilized transposons in the human and chimpanzee genomes. *Am J Hum Genet* 78(4):671–679.
30. Salem AH, et al. (2003) Alu elements and hominid phylogenetics. *Proc Natl Acad Sci USA* 100(22):12787–12791.
31. Churakov G, et al. (2009) Mosaic retroposon insertion patterns in placental mammals. *Genome Res* 19(5):868–875.
32. Prado-Martinez J, et al. (2013) Great ape genetic diversity and population history. *Nature*, 499:471–475.
33. Hach F, et al. (2010) mrsFAST: A cache-oblivious algorithm for short-read mapping. *Nat Methods* 7(8):576–577.
34. Hormozdiari F, et al. (2011) Alu repeat discovery and characterization within human genomes. *Genome Res* 21(6):840–849.
35. Hormozdiari F, et al. (2010) Next-generation VariationHunter: Combinatorial algorithms for transposon insertion discovery. *Bioinformatics* 26(12):i350–i357.
36. Mills RE, et al.; 1000 Genomes Project (2011) Mapping copy number variation by population-scale genome sequencing. *Nature* 470(7332):59–65.
37. Stewart C, et al.; 1000 Genomes Project (2011) A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genet* 7(8):e1002236.
38. Hinrichs AS, et al. (2006) The UCSC genome browser database: Update 2006. *Nucleic Acids Res* 34(Database issue, suppl 1)D590–D598.
39. Deininger PL, Moran JV, Batzer MA, Kazazian HH, Jr. (2003) Mobile elements and mammalian genome evolution. *Curr Opin Genet Dev* 13(6):651–658.
40. Medstrand P, van de Lagemaat LN, Mager DL (2002) Retroelement distributions in the human genome: Variations associated with age and proximity to genes. *Genome Res* 12(10):1483–1495.
41. Felsenstein J (2004) *Inferring Phylogenies* (Sinauer, Sunderland, MA), Vol 2.
42. Witherspoon DJ, et al. (2013) Mobile Element Scanning (ME-Scan) identifies thousands of novel Alu insertions in diverse human populations. *Genome Res* 23:1170–1181.
43. Ohshima K, et al. (2003) Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. *Genome Biol* 4(11):R74.
44. Wagstaff BJ, Kroutter EN, Derbes RS, Belancio VP, Roy-Engel AM (2013) Molecular reconstruction of extinct LINE-1 elements and their interaction with nonautonomous elements. *Mol Biol Evol* 30(1):88–99.
45. Roy-Engel AM (2012) LINEs, SINEs and other retroelements: Do birds of a feather flock together? *Front Biosci* 17:1345–1361.
46. Hormozdiari F, Alkan C, Eichler EE, Sahinalp SC (2009) Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res* 19(7):1270–1278.