Assembly of polymorphic *Alu* repeat sequences from whole genome sequence data in diverse humans

Julia H. Wildschutte[1], Alayna Baron[1], Nicolette M. Diroff[1], and Jeffrey M. Kidd[1,2#]

[1]*University of Michigan Medical School, Department of Human Genetics, 1241 E. Catherine St., Ann Arbor, MA, 48109.*

[2] *University of Michigan Medical School, Department of Computational Medicine and Bioinformatics, 100 Washtenaw Ave., Ann Arbor, MA, 48109.*

[#] To whom correspondence should be sent: jmkidd@med.umich.edu

Jeffrey M. Kidd, Ph.D.

Department of Human Genetics

Department of Computational Medicine and Bioinformatics

University of Michigan Medical School

1241 E. Catherine St.

Ann Arbor, MI 48109

Office phone (734) 763-7083

Contributing Authors' e-mail: JHW: jwilds@med.umich.edu; AB: alaynaa@med.umich.edu ; ND: nmdiroff@umich.edu

Running title:

Assembly of Alu insertions from shotgun sequences

Keywords
Mobile element insertion, retroelement, reverse transcription, *Alu*, SINE, *de novo* assembly

Abstract

Alu insertions have contributed to >11% of the human genome. About ~30-35 *Alu* subfamilies remain actively mobile, and are recognized as major drivers of genetic variation and disease. Sophisticated computational methods permit identification of non-reference insertions based on specific signatures from whole genome sequencing data, but reporting of entire insertion sequences is limited. We build on existing methods and develop an approach that combines *Alu* detection and *de novo* assembly of WGS data to reconstruct the full sequence of insertion events. Using this approach, we generate a highly accurate call set of 1,614 completely assembled *Alu* variants from 53 samples from the Human Genome Diversity Project panel. Experimental validation of 30 sites shows 100% this method produces a highly accurate call set that accurately reconstructs insertion sequence. We utilize the reconstructed alternative insertion haplotypes to genotype 1,010 fully assembled insertions, obtaining >99% accuracy. We find evidence of insertion by non-classical mechanisms and observe 5' truncation in 16% of *Alu*Ya5 and *Alu*Yb8 insertions. The sites of truncation coincide with stem-loop structures and SRP9/14 binding sites in the *Alu* RNA, implicating L1 ORF2p pausing in the generation of 5' truncations.

Introduction

Mobile elements (MEs) are discrete fragments of nuclear DNA that are capable of copied movement to other chromosomal locations within the genome (Kazazian 2004). In humans, the ~300 bp *Alu* retroelements are the most successful and ubiquitous MEs, collectively amounting to >1.1 million genome copies and accounting for >11% of our nuclear genome (McPherson et al. 2001; Batzer and Deininger 2002). The vast majority of *Alu* insertions represent germline events that occurred millions of years ago and now exist as non-functional elements that are highly mutated and no longer capable of mobilization (McPherson et al. 2001). However, subsets of MEs, including *Alu* and its autonomous partner *L1Hs*, remain active and continue to contribute to new ME insertions (MEIs), resulting in genomic variation between individuals in the population (Mills et al. 2007).

The human *Alu* is represented by the *Alu*Y, *Alu*S, and *Alu*J lineages, which can be further stratified into more than ~35 subfamilies based on sequence diversity and diagnostic mutations (Jurka and Smith 1988; Mills et al. 2007). Most human *Alu* elements are from the youngest lineage, *Alu*Y, whose members have been most actively mobilized during primate evolution (Mills et al. 2007). Of these, the *Alu*Ya5 and *Alu*Yb8 subfamilies have contributed to the bulk of insertions in humans (Bennett et al. 2004; Wang et al. 2006; Xing et al. 2009; Stewart et al. 2011), although polymorphic insertions from >20 other *Alu*Y and >6 *Alu*S subfamilies have also been reported (Carroll et al. 2001; Mills et al. 2007). In contemporary humans, the retrotransposition of active *Alu* copies results in *de novo* germline insertions at a frequency of ~1:20 live births (Cordaux et al. 2006; Xing et al. 2009). Over 60 novel *Alu* insertions have been shown to cause mutations leading to disease, either as a direct consequence of insertional mutagenesis, or by providing a template of highly repetitive sequence that has since facilitated chromosomal rearrangements and structural variation (Batzer and Deininger 2002; Callinan and Batzer 2006; Sen et al. 2006; Hancks and Kazazian 2012; Solyom and Kazazian 2012). Thus, *Alu* insertions continue to shape the genomic landscape and are recognized as profound mediators of genomic structural variation.

Active copies of *Alu* are non-autonomous but contain an internal RNA Pol III promoter (Fuhrman et al. 1981). Mediated by L1 encoded enzymes, *Alu* transcripts are mobilized by a 'copy-and-paste' mechanism referred to as target primed reverse transcription (TPRT) (Dewannieux et al. 2003; Deininger 2011). Classical TPRT involves the reverse transcription of a single stranded *Alu* RNA to a double stranded DNA copy, during which two staggered single-stranded breaks are introduced in the target DNA of ~5 to ~25 bp that are later filled by cellular machinery. The resulting structure consists of a new *Alu* flanked by characteristic target site duplications (TSDs) and a poly-A tail of variable length. Together these serve as hallmarks of retrotransposition. Integration of the new copy is permanent; although *Alu* can be removed by otherwise encompassing deletions, there is no known mechanism for precise excision. Classical TPRT is responsible for the majority of *Alu* insertions, however a minority of insertions has undergone movement by detectable non-classical TPRT mechanisms (Callinan et al. 2005; Chen et al. 2007; Srikanta et al. 2009a).

The primary difficulty in identifying novel *Alu* insertion loci stems from the highly repeated nature of the element itself. Various approaches for large-scale analyses of *Alu* and other ME types have been developed to utilize next generation sequencing platforms. Scaled sequencing of targeted *Alu* junction libraries has permitted genome-wide detection, offered in techniques such as Transposon-Seq (Iskow et al. 2010) and ME-Scan (Witherspoon et al. 2010; Witherspoon et al. 2013). Such targeted methods offer high specificity and sensitivity, but are restricted by the primers used for detection and are generally subfamily-specific. Broader detection of *Alu* variant locations is possible using computational methods to detect insertions from whole genome sequence (WGS) paired reads, for example in 'anchored' mapping. This method seeks to identify discordant read pairs where one read maps uniquely to the reference (*i.e.*, the 'anchor') and its mate maps to the element type in query (Hormozdiari et al. 2011; Stewart et al. 2011; Keane et al. 2013) However, since only read pair data is considered, these methods are limited in the recovery of variant-genome junctions that might otherwise be captured from split read information. Specialized algorithms now offer improved breakpoint accuracy from WGS data by integrating both read pair and split read signals during variant detection (Lee et al. 2012; David et al. 2013;

Keane et al. 2013; Thung et al. 2014). Having been mostly applied to WGS exceeding 35-40x, these methods offer high sensitivity and accuracy but, as with all WGS read-based algorithms, are limited for lower coverage data. Also, these detection methods have been developed in the context of ME variant loci discovery and reporting. Assembly approaches have received increasing application to next-generation sequencing data for breakpoint identification. Tools such as TIGRA (Chen et al. 2014), a modification of the SGA assembler used in HYDRA-MULTI (Simpson and Durbin 2012; Malhotra et al. 2013), and the use of a *de brujin* graph-based approach in SVMerge (Wong et al. 2010) have been developed to assemble structural variant breakpoints from population scale or heterogeneous tumor sequencing studies. Assembly-based approaches have also lead to increased sensitivity and specificity for the detection of SNPs and small indels (Iqbal et al. 2012; Li 2012; Li et al. 2013; Narzisi et al. 2014; Rimmer et al. 2014).

Here, we utilize a classic overlap-layout-consensus assembly strategy applied to ME-insertion supporting reads to completely reconstruct and characterize *Alu* insertions. We apply this approach to pooled WGS data, from 53 individuals in 7 geographically diverse populations from the Human Genome Diversity Project (HGDP) panel (Cann et al. 2002; Martin et al. 2014), enabling a more comprehensive characterization of fully intact polymorphic *Alu* elements.. We present the analysis of 1,614 fully reconstituted *Alu* insertions from these samples, including breakpoint refinement and genotyping of 1,010 insertions, with >99% accuracy. These results provide a basis for future study of such MEIs in human disease and population variation.

Results

Detection and assembly of Alu variants

We attempted to assemble a high-quality collection of fully reconstituted non-reference polymorphic *Alu* variants utilizing WGS data from a subset of the HGDP collection (Figure 1). The data consisted of 2x101 bp paired-end libraries from 53 individuals across seven populations, with a median coverage of ~7x per genome. All samples were aligned to the human GRCh37/hg19 reference and processed as previously described (Martin et al. 2014). Using these data, we initially sought to identify non-reference *Alu* insertions based on supporting read-pair signatures, as reported by the program RetroSeq (Keane et al. 2013). RetroSeq infers candidate MEIs based on insertion-supporting read clusters identified from anchored mapping; candidate calls are gauged by an assigned quality metric that accounts for overall read depth, number of supporting reads per cluster, and the balance of support for 5' and 3' ends of each call. Because this approach is inherently limited by read presence, we anticipated insertions that were private to a single individual were likely to be missed given the range in coverage per sample. However, we reasoned that borrowing read information across samples would increase our ability to detect rarer insertions that were present in multiple samples, and pooled the data into a single merged BAM from all individuals, having an effective coverage of ~429x.

To identify candidate *Alu* insertions from these data, we applied RetroSeq to the merged BAM requiring minimum read support for a putative call site. To minimize false calls associated with reference elements, we removed any candidate call that mapped within 500 bp of an annotated *Alu* in the human reference sequence. After filtering, this resulted in 41,365 putative *Alu* insertions with a quality score 6 or higher, though we note that this candidate set from RetroSeq likely contained many false positives: only 8,867 had a score of 7 or 8. As with other MEI detection software, RetroSeq was designed to identify putative MEI variant locations based on WGS read signatures; to our knowledge there is no available software that further reports information for the actual inserted sequence. Consequently, validation and further sequence analysis requires examination of putative insertions on a per-site basis.

Making use of reads associated with candidate sites, we obtained full reconstruction of for as many individual insertion variants as possible, including the complete *Alu* sequence and contiguous flanking breakpoints for each site. In recent years, much effort in genome assembly has focused on the analysis of short-reads using approaches based on a *de brujin* graph (Li and Homer 2010; Miller et al. 2010; Lee and Tang 2012). However, we reasoned that given our data, a local assembly that could utilize an overlap-layout-consensus approach would take full advantage of 101 bp paired end reads. For these purposes we utilized the program CAP3 (Huang and Madan 1999), which was originally designed for the assembly of large-insert clones sequenced using capillary sequencing, but which has also been applied to *de novo* assembly of short read RNA-seq (Yang and Smith 2013) and metagenomic sequence data (Reddy et al. 2014). For each putative site called by RetroSeq, we retrieved read pairs that were 1) reported as supporting the candidate call by RetroSeq and 2) proximal soft-slipped reads that mapped within 200 bp of the predicted insertion site, requiring that the clipped portion was at least 20bp in length and had a mean quality of ≥20. We then performed 41,365 *de novo* assemblies with CAP3 using these read sets (Figure 1).

We ran CAP3 with parameters adjusted for joining smaller overlaps present in shorter reads (see Methods). CAP3 builds consensus-based contigs using constraints from read pair information and reports 'links' between assembled contigs based on those constraints. A best-case consensus combines all associated reads into a single contig; however, associated contigs may be linked but unmerged, for example from insufficient read overlap. Making use of this information, we merged linked contigs (separated by a string of 300 "N"s representing sequence gaps) to create a collection of assembled scaffolds for each candidate *Alu* insertion. We then subjected the resulting scaffolds to additional analyses as follows. We first identified assembled sequences that contained an *Alu* sequence using RepeatMasker (Smit et al. 1996-2010). Because we excluded any predicted insertion from our call set that was within 500 bp of any annotated *Alu* element, assembled scaffolds having *Alu*-matching sequence were

interpreted to represent the presence of a *bona-fide* non-reference insertion. We further required the presence of at least 30 bp corresponding to *Alu* sequence (requiring ≥90% nucleotide identity) and recovery of ≥30 bp of flanking non-gap sequence at one end, resulting in 2,971 candidate assemblies.

Each candidate assembly was then subjected to a more comprehensive breakpoint analysis. For these purposes, we adapted a procedure developed for the analysis of structural variant breakpoints represented in finished fosmid clone sequences (Kidd et al. 2010a) to utilize our locally assembled scaffolds (also see Methods). A total of 1,614 *Alu*-containing contigs were reconstructed, each having the complete associated insertion and at least one breakpoint with ≥30 bp of mapped flanking sequence (Table S1, Figure S1). Overall, the assembled insertions ranged in size from 66 bp to 681 bp (median of 310), with predicted TSDs up to 102 bp in length (median of 14 bp). Five loci were predicted to have *Alu* insertions associated with deletion of pre-insertion sequence relative to the hg19 assembly, indicating potential non-classical insertion mechanisms (Zingler et al. 2005; Srikanta et al. 2009a). From the full set, 1,010 *Alu* insertions had both breakpoints flanked by at least 100 bp of non-gap assembled sequences. These 1,010 insertions were deemed to be of the highest quality and were suitable for subsequent *in silico* genotyping (see below).

Validation of insertion calls

From our assembled set of 1,614 non-reference insertions, we selected 30 sites for experimental validation by PCR and sequencing, biased in favor of sites with unusual breakpoint characteristics (*i.e.*, 0-3 bp TSDs, TSDs larger than 25bp, and sites with corresponding target site deletions) (Table S2). In validation, our specific goals were to demonstrate 1) the presence of the *Alu* at that chromosomal location, 2) agreement between the assembled sequence with the cognate validated sequence for the insertion, and 3) contiguous sequence of each insertion with its mapped flanking regions. We obtained Sanger sequence of the insertion allele for each of the 30 assembled sites in up to two individuals predicted to have the insertion, utilizing samples that were homozygous for the insertion when possible. Sequencing was performed with primers situated both upstream and downstream of the insertion in order to account for uncertainty introduced from polymerase slippage at the poly-A tails.

For all 30/30 tested sites, we confirmed the presence of an *Alu* insertion in the tested sample. Subsequent analysis of the corresponding nucleotide alignments verified that the nucleotide sequence of each *Alu* recovered in CAP3 assembly was in complete agreement with the corresponding Sanger traces. Examples for three representative insertions are highlighted in Figure 2, illustrating the recovery of mapped *Alu*-containing contigs, breakpoint estimations at those sites, and alignment of the deduced nucleotide sequence to the CAP3 assembly. Detailed alignments corresponding to individual insertions, including visualized trace information, are provided in Figure S2.

We assessed *in silico* breakpoint estimations for each validated CAP3 assembly in comparison to the *Alu*-genome junctions as obtained from Sanger reads. Six of the 30 validated sites were predicted to have breakpoints within 100 bp of a gap in the CAP3 assembly; sequence comparison to the CAP3 assembly revealed correctly predicted breakpoints for just 1 of these 6 insertions, justifying their exclusion from subsequent *in silico* genotyping (described further below). Exact breakpoint and TSD sequences for the remaining 5 insertions were determined from Sanger sequencing (Figure S2).

Overall, 21/30 insertions had target sites that precisely agreed with the corresponding assembly. Representative examples are shown in Figure 2; properties for all validated sites are summarized in Table S3. Of the 23 validated sites that were later utilized for genotyping, just 3 were found to have incorrect breakpoints, in each case due to the absence of target site sequence adjacent to the *Alu*-poly-A tail in the CAP3 assembly. For an additional site (insertion at chr2:123330649), comparison to the Sanger traces revealed a longer inferred TSD than predicted (16 bp *vs*. 13 bp), resulting from a nucleotide change within the assembled poly-A stretch, thus altering the inferred target duplication by 3 bp. The remaining 20 sites (83.3% of the 24) correctly matched the breakpoints recovered from the CAP3 assembly. These included insertions with unusual breakpoint characteristics, for example, a full-length insertion having a validated TSD of 46bp (chr18:74638702). We also correctly recovered insertions with evidence of non-classical insertion mechanisms, including three sites that were correctly predicted to have short target site

deletions relative to the pre-insertion allele ranging in size from 1 to 6 bp (at positions chr6:164161904, chr11:26601646, and chr12:73056650) (see also Figure 2B), and at least 93 elements with evidence of 5' truncation events (described further below). We were able to completely reconstruct, with correct breakpoints, insertions that were within other repetitive sequence classes (Figure 2C). Finally, we note one insertion, located at chr11:35425392 for which the recovered CAP3 contig was found to be in complete agreement with corresponding Sanger reads, however our automated identification of the TSD was 'miscalled' due to the presence of concomitant variation at this site relative to the hg19 assembly (Figure S3).

Comparison of calls with other discovery sets

We then compared our assembled calls for overlap with two recent studies that utilized comparable methods for discovery of non-reference MEI events using mapping of anchored read pairs (Stewart et al. 2011; Hormozdiari et al. 2013) and/or MEI-informative split reads (Stewart et al. 2011). For these purposes, we converted reported coordinates to the GRCh37/hg19 assembly using liftOver and restricted analysis to non-reference insertion events only. We then determined the overlap of predicted *Alu* insertions from each data set, permitting a 200 bp window (Figure 3). The majority of our events overlap with these previously described insertions, with ~88% of our calls (1,430/1,614) found in at least one existing call set, and ~65% of insertions (1,064/1,614) common to all three call sets. This is in contrast to *Alu* events that were shared between the two previous studies that, respectively, had ~50% overlap in the three way comparison, indicating a relatively high proportion of overlap for our assembled call set. This is also reflected in the number of calls found to be unique in our analysis compared to either existing call set. Repeating this analysis with the 8,515 best scoring candidates (RetroSeq quality score of 8, without regard to assembly; see Methods), only slightly elevates overlap between call sets, with 1,833 of our calls shared in at least one previous study and 1,305 of events identified in all studies, again suggesting a high level of false positives prior to assembly.

Characteristics of assembled Alu insertions

Given the accuracy of our assemblies, we sought to more comprehensively characterize our set of reconstructed *Alu* insertions. Previous studies of full-length polymorphic elements have been mostly limited to insertions taken from an assembled reference genome (Lander et al. 2001; Wang et al. 2006), examination of trace archive data (Bennett et al. 2004), or from insertions having been captured in relatively long read data (Stewart et al. 2011). By making use of contemporary WGS data in *de novo* assembly, the insertion sequence itself is accurately reconstructed for analysis. Thus, utilizing our assembled contigs, we readily extracted the corresponding 1,614 *Alu* nucleotide sequences and characterized each in terms of subfamily distributions and properties.

Based on sequence divergence from *Alu* subfamily consensus sequences obtained from the most recent RepBase update (Jurka et al. 2005), we were able to assign 1,452 (90%) of our insertions to one of 30 subfamilies (Table 1). We found 162 elements that were equally diverged from more than one subfamily consensus and could not be conclusively classified. Insertions from *Alu*Y subfamilies made up >99% of all assigned calls, with *Alu*Ya5 and Yb8 collectively representing more than half (62.7%) of the set. This observation was expected, given that *Alu*Y insertions have contributed to nearly all *Alu* genomic variation in humans, with *Alu*Ya5 and Yb8 being the most active subfamilies (Mills et al. 2007). Also as expected, insertions derived from *Alu*S and *Alu*J subfamilies were a minority, together representing less than 1% of calls that could be assigned to a subfamily. In general these data are similar to previous analyses of representative intact polymorphic *Alu* in humans (Bennett et al. 2004; Wang et al. 2006; Xing et al. 2009; Stewart et al. 2011).

To assess the length distribution of non-reference *Alu* variants from our call set, we focused on insertions assembled from the *Alu*Ya5 and *Alu*Yb8 subfamilies. We reasoned that analysis of these particular subfamilies should provide the most informative resource for comparison given their representation as the majority of identified variants. We further limited analysis to those *Alu* that were suitable for genotyping, as insertions that do not meet our criteria for genotyping may erroneously appear

to be truncated due to an incomplete breakpoint assembly. This resulted in an analysis set of 351 *Alu*Ya5 and 215 *Alu*Yb8 insertions. Based on nucleotide alignments of the assembled insertions against their respective consensus, we examined the collective coverage of assembled elements, per subfamily, in comparison to the nucleotide positions relative to their respective consensus (Figure 4A and B).

We observed that 84.9% of *Alu*Ya5 (298/351) and 81.4% *Alu*Yb8 (175/215) variants were full-length, or within at least 5 bp of being full-length, consistent with previous reports of the genome-wide distribution of full-length *Alu* (Zingler et al. 2005; Chen et al. 2007). Comparing the length distribution of all insertions revealed a detectable minority of 5' truncations that were present in both subfamilies and exhibited a similar distribution of the apparent truncation point (Figure 4). More specifically, a subset of insertions from either subfamily was truncated ~8-45 bp from the consensus start (9.9% or 35/351 *Alu*Ya5, and 13.4% or 29/215 *Alu*Yb8 insertions), and a second subset was truncated ~55-171bp from the consensus start (5.1% or 18/351 *Alu*Ya5, and 5% or 11/215 *Alu*Yb8) (Table 2). Besides having apparent 5' truncations, all but two of these assembled insertions displayed characteristics of 'standard' *Alu*, including flanking TSDs and a poly-A tail of variable length (insertions at chr13:86166445 and chr11:26601646; also see Table S1). We note the observed distribution is similar to that from two previous analyses of 10,062 reference human *Alu* (as extracted from NCBI build33) (Zingler et al. 2005), and of 1,402 intact polymorphic *Alu* from the then-current dbRIP (Chen et al. 2007); aspects of both are addressed further in the Discussion.

L1 and *Alu* insertions that are truncated but otherwise standard are thought to arise from non-classical TPRT (Zingler et al. 2005; Chen et al. 2007; Srikanta et al. 2009a; Srikanta et al. 2009b). For example, one mechanism thought to contribute to 5' truncations is a microhomology-mediated pairing of nucleotides at the genomic target 5' end with the nascent *Alu* mRNA, resulting in premature completion of TPRT (Chen et al. 2007), and in turn leaving a detectable signature. We manually examined each three way alignment of the 53 *Alu*Ya5 and 40 *Alu*Yb8 assembled 5' truncation events for such evidence, specifically searching for nucleotides at the 5' break that were shared with the respective *Alu* consensus at that position (Chen et al. 2007). We observed a subset of insertions with detected microhomology, with 40.9% of truncation having 1 bp of matching sequence, and 15.1% of all truncations with ≥2 bp shared at the 5' break (details are summarized in Table S4), though we note limitations of interpreting a single shared nucleotide as a 'true' instance of microhomology. Given this observation, the data indicate premature TPRT may account for a subset of the truncated insertions.

Insertion breakpoint distribution

We analyzed the distribution of assembled insertions relative to genes based on Gencode v19 annotations (Harrow et al. 2012). Of the 1,614 assembled insertions, 865 (~53.5%) were found within genes, of which 643 (~39.8%) were located within protein coding genes. Although these values are slightly higher than has been reported in previous analysis (Xing et al. 2009; Stewart et al. 2011), these values are lower than expected based on random permutations of our data (924, or ~57.2% expected within all gene regions and 688, or ~42.6% in protein coding genes). Just 10 insertions (~0.61% of all calls) were found within exons, all of which were located in untranslated regions and therefore would not be predicted to disrupt coding sequence. This value is much lower than expected based on random simulations (50 sites, or ~3.1 %; $p<0.02$), indicating potential selection against retrotransposition into exons and other coding sequence, and consistent with previous studies indicating exonic depletion of *Alu* (Xing et al. 2009; Hormozdiari et al. 2011; Stewart et al. 2011; Witherspoon et al. 2013).

A total of 708 (~43.8%) of our assembled insertions were located within repetitive sequence. The majority these insertions were found within other retrotransposon-derived elements (459, or ~28.4%, were in LINEs and 124, or ~7.6% in LTRs), and in DNA transposons (69, or ~4.2%); 22 insertions were found in minor or unknown repetitive classes. This distribution is also consistent with that observed in previous survey of non-reference *Alu* insertions (Stewart et al. 2011). Since we excluded any candidate call that was near an annotated *Alu* prior to assembly, no insertion from our callset was recovered within any existing *Alu*, though a handful of insertions were observed within non-*Alu* SINE classes (*e.g.*, from the *Mir*, FLAM, or FRAM groups). We compared these data to randomized values based on simulated

placement of insertions, excluding regions in the hg19 reference assembly that mapped to gaps or are annotated as *Alu*, observing no significant difference relative to random placement. Based on separate simulations permitting placement within annotated *Alu* elements, we estimate that 10.5% of random insertions would be near annotated *Alu*s, and hence excluded. Assuming that the placement of real insertions is random, we estimate that ~189 sites are missing from the set of 1,614 insertions due to their colocalization with reference *Alu* elements.

Genotyping

We identified a subset of 1,010 insertions that had sites with both breakpoints at least 100 bp away from an assembly gap that were suitable for genotyping using Illumina sequencing reads. For each site, we recreated the reference and insertion haplotypes based on 600 bp flanking the inferred insertion site. For each sample, we then remapped Illumina read-pairs that mapped to each reference location against both reconstructed sequences using bwa (see Materials and Methods). We then determined genotype likelihoods based on the mapping of reads to each alternative allele, with error probabilities as indicated by the read mapping quality (Li 2011). Of note, read pairs that map equally well to the reference and insertion sequences will have a MAPQ of 0 and are uninformative for establishing the genotypes. Final genotypes were determined for each sample based on the resulting genotype likelihoods (Table S5). For sites on the autosomes and the pseudo-autosomal regions of chrX, genotypes were obtained after LD-aware refinement using BEAGLE v3 (Browning and Browning 2009), in a procedure that also included SNP genotypes as previously described (DePristo et al. 2011). We compared the inferred genotypes for 11 autosomal sites with PCR-based genotyping across 10 samples, and found a total concordance rate of 99% (109/110) (Figure 5A and B; predicted genotypes are in Table S6 for direct comparison). The only error among the tested calls occurred when the inferred genotype was homozygous for the insertion allele, while PCR genotyping indicates that the site is heterozygous (chr10:19550721; HGDP00476). Finally, we performed a Principle Component Analysis (PCA) of the autosomal genotypes across all 53 samples (Patterson et al. 2006). As expected, individual samples largely cluster together by population with the first PC separating African from non-African samples (Figure 5C). This result further confirms the high accuracy of the inferred genotypes.

Discussion

We utilized whole genome NGS data to fully reconstitute a set of 1,614 non-reference *Alu* insertions from a subset of 53 genetically diverse individuals in 7 global populations from the HGDP (Cann et al. 2002; Martin et al. 2014). Experimental interrogation of 30 sites confirmed the presence of a non-reference *Alu* insertion at that site. For 1,010 insertions that had at least 100 bp of flanking assembled sequence, we obtained a high level of breakpoint accuracy, having perfect agreement at 87%, or 20/23 sites tested. We confirmed the presence of several insertions with aberrant assembled breakpoint characteristics, including insertions for which the TSD was absent (for example at chr17:46617220), the TSD was of extreme length (chr18:74638702), and insertions with deleted sequence relative to the hg19 reference (chr6:164161904, chr11:26601646, and chr12:73056650; also see Figure S2).

Examination of individual sites indicated that elements located near edges of assembled contigs (*e.g.*, excluding the complete TSD length) were more likely to have incompletely assembled breakpoints. Further inspection of individual reads supporting the assembled contig indicated that this was due to aberrant joining or incomplete TSD capture of reads that covered the poly-A tract (also refer to insertions at chr12:99227704, chr22:26997608, and chrX:5781742). In each case, the correct sequence of the *Alu* insertion itself was still obtained. For example, our assembly of the Y Alu Polymorphic element (YAP) (Hammer 1994) located at chrY:21611993 contained an incomplete TSD. Capillary sequencing in sample HGDP00213 revealed the correct 11 bp TSD (5' AAAGAAATATA), and confirmed the presence of YAP-specific nucleotide markers (at bases 64, 207, 243, and 268 relative to the *Alu*Y8b consensus), as predicted by our CAP3 assembly and consistent with previous reports (Figure S2) (Hammer 1994). We additionally note that even when alleles are fully reconstructed, interpretation of the variant may not be clear. An insertion at chr11:35425392 is illustrative of this complexity. At this site, our automated identification of the variant breakpoints was inaccurate due to the presence of concomitant variation at this site relative to the hg19 reference, as revealed by sequencing in other individuals without the insertion to better reconstruct the structure of the pre-insertion allele (Figure S3). Notably, the CAP3 assembled insertion and proximal genomic sequence was found to be in complete agreement with corresponding Sanger reads, despite the presence of this surrounding structural variation relative to the reference build.

By comparing of our full set of reconstituted *Alu* insertions to existing *Alu* discoveries from Stewart *et al* (Stewart et al. 2011) and Hormozdiari *et al* (Hormozdiari et al. 2011) having been catalogued through similar read signatures we determined 184, or about 11.4%, were novel to our study (Figure 3). That most of our identified insertions overlap with existing datasets indicates a very low false positive rate, further supported by our validation of 100% of sites tested, and our genotyping accuracy of >99%. This also suggests that our current catalogue of polymorphic *Alu* variants within the population is still incomplete. In contrast, the numbers of predicted polymorphic MEIs in these previous studies was much higher, with more than 2,100 novel insertions reported from 185 individuals from the 1000GP Pilot (Stewart et al. 2011), and nearly the same number of novel sites from just 8 samples (Hormozdiari et al. 2011), suggesting a much higher false discovery rate of such approaches when applied to low coverage data..

The availability of assembled insertion alleles results in a more precise call set and permits subsequent analysis of element properties. By remapping Illumina read-pairs to reference and alternative haplotypes, we performed direct *in silico* genotyping, achieving an estimated 99% genotype concordance (109 of 110 genotypes analyzed). The vast majority of elements exhibited properties consistent with classical retrotransposition, including being full length and the presence of a TSD and poly-A tail. Examination of the length distribution of the reconstructed *Alu*Ya5 and Yb8 insertions revealed that 93 (~16.4%) of this subset had evidence of having been 5' truncated, despite appearing otherwise standard, indicating insertion by potential non-classical TPRT mechanisms. We also observed evidence of at least two groups of this subset, respectively truncated ~30 to 50 bp and ~160 to 180 bp from the canonical 5' edge (Table 2 and Figure 4).

These data are consistent with a previous manual curation of 1,402 intact polymorphic *Alu* from dbRIP that characterized full-length elements available at the time (Chen et al. 2007). In that study the authors identified 115 elements (~8.2%) with apparent 5' truncations ~8-45 bp from the *Alu* start (~8.2%) and 89 elements had ~55-171 bp truncations (6.3%) (Chen et al. 2007). The authors proposed a model of microhomology-mediated nucleotide pairing of the 5' end of the genomic strand with the *Alu* RNA, having observed 41.2% events with nucleotides at the 5' break shared with the *Alu* consensus at that position. However the majority of the truncations were supported by a single shared base; considering ≥ 2 bp accounted for 16.7% of their observed events. We searched our own data corresponding to all 5' truncation events, and observed insertions with similar levels of putative microhomology: 15.1% had at least 2 shared bases at the 5' edge, and 40.9% of insertions shared 1 base; although tentatively considered to represent true cases of microhomology, this is greater than 25% of sites expected at random. One other study reported similar instances of *Alu* truncation events (1,005/10,062, or ~10.5%), but found little to no statistical support for base overlap at the 5' breaks (~29% 1 bp; ~13% ≥ 2 bp) (Zingler et al. 2005). Given that the 5' *Alu* end is particularly GC rich, this suggests such a 'mis'-pairing during TPRT would account for a minority of observed truncations.

In support of this idea, we examined the nick site for truncations with and without putative signatures of microhomology and found no difference; with both classes containing the canonical L1 ORF2 protein (ORF2p) nick site, 5' T_4/A_2 (the '/' indicating the site of cleavage) (Feng et al. 1996; Cost and Boeke 1998). We note that secondary structure of the Alu RNA may drive the non-random distribution of 5' truncation points. The truncation points at ~45bp and ~180 from the Alu start are coincident with a predicted hairpin structure in the folded RNA (Bennett et al. 2008). The *Alu* RNA is reverse transcribed by the L1 encoded ORF2p, which pauses at sites of RNA secondary structure such as poly-purine tracts and stem-loops (Piskareva and Schmatchenko 2006). Additionally, both truncation regions are located directly 3' to predicted SRP9/14 binding locations (Weichenrieder et al. 2000; Deininger 2011). Although SRP9/14 binding is necessary for efficient retrotransposition, the younger *Alu*S and *Alu*Y subfamilies contain nucleotide substitutions that reduce SRP9/14 binding affinity, suggesting that efficient displacement of bound SRP9/14 is important for the successful propagation of these elements (Mills et al. 2007; Bennett et al. 2008). This suggests that the characteristic location of 5' truncations may be a consequence of ORF2p pausing and disengaging from the *Alu* RNA during reverse transcription. Regardless, the data indicate non-classical TPRT mechanisms may account for a subset of the truncated insertions, although alternative mechanisms cannot be ruled out (Callinan et al. 2005; Srikanta et al. 2009a; Srikanta et al. 2009b).

Although our assembled results are of high quality, our discovery process suffers from the same limitations that are common to other studies utilizing NGS. For example, because of the variability in coverage across samples, we may be missing rare sites present in only one or a small number of the analyzed samples. Additionally, the inability to confidently map 100bp sequencing reads into some genomic regions may bias the discovery set away from insertions near other repetitive elements. However, 43.8% of the assembled insertions are located within other repetitive elements, a similar distribution as reported previously (Stewart et al. 2011). A notable exception is insertions near existing *Alu* elements in the reference sequence due to their explicit removal from our call set. Based on random permutations we estimate that up to 10.5% of insertions may have been missed due this filter. As sequencing studies utilize longer reads and higher depths these problems will be ameliorated. Such changes will also increase the feasibility of assembly-based approaches, permitting the direct reconstruction of full insertion sequences and other structural variant breakpoints and ultimately contributing to a more a more complete picture of all types of genomic variation.

Methods

Samples

We analyzed whole genome, 2x101 bp Illumina read sequence data from a subset of the HGDP including 53 samples were from 7 populations: Cambodia (HGDP00711, HGDP00712, HGDP00713, HGDP00715, HGDP00716, HGDP00719, HGDP00720, HGDP00721), Pathan (HGDP00213, HGDP00222, HGDP00232, HGDP00237, HGDP00239, HGDP00243, HGDP00247, HGDP00258), Yakut (HGDP00948, HGDP00950, HGDP00955, HGDP00959, HGDP00960, HGDP00963, HGDP00964, HGDP00967), Maya (HGDP00854, HGDP00855, HGDP00856, HGDP00857, HGDP00858, HGDP00860, HGDP00868, HGDP00877), Mbuti Pygmy (HGDP00449, HGDP00456, HGDP00462, HGDP00471, HGDP00474, HGDP00476, HGDP01081), Mozabite (HGDP01258, HGDP01259, HGDP01262, HGDP01264, HGDP01267, HGDP01274, HGDP01275, HGDP01277), and San (HGDP00987, HGDP00991, HGDP00992, HGDP01029, HGDP01032, HGDP01036). WGS was processed using BWA, GATK (McKenna et al. 2010) and Picard (http://picard.sourceforge.net), as described previously (Martin et al. 2014) and available at the Sequence Read Archive under accession SRP036155. Final datasets are ~7x coverage per sample.

Non-reference *Alu* discovery

For identification of *Alu* insertions from our whole genome sequence data, we performed anchored read pair mapping using RetroSeq (Keane et al. 2013) to identify non-reference variants relative to the GRCh37/hg19 genome assembly. To putatively call candidate loci from sites having discordant mapping read pairs, the 'discover' phase was run on individual BAM files for each sample to identify pairs with one read mapping uniquely to the reference genome and its pair to an *Alu* consensus or to an annotated *Alu* present in the reference. A FASTA file of the available *Alu* consensus sequences was obtained from RepBase (Jurka et al. 2005) and reference *Alu* elements were identified using existing RepeatMasker (Smit et al. 1996-2010) annotations. Next, candidate insertion calls were identified using the RetroSeq 'call' phase. For this analysis, we combined the supporting read information discovered in each individual and ran RetroSeq 'call' on a combined BAM consisting of all samples. In the 'call' phase, we required a minimum read support of 2 supporting read pairs per call (-reads). A maximum read-depth of 1000 (-depth; default is 200) was utilized for regions surrounding each call in order to accommodate the increased coverage of the merged BAM. Finally, any output call within 500 bp of an annotated *Alu* insertion was excluded using the bedtools window command (Quinlan 2014) and RepeatMasker hg19 reference annotations (Smit et al. 1996-2010). Unless otherwise noted, any other RetroSeq options were run at the default settings. Final *Alu* calls having met the further criteria of a filter tag FL=6,7,or 8 were selected for subsequent analysis.

Assembly of non-reference *Alu* elements

De novo assembly of insertion-supporting reads for each candidate insertion was performed use the CAP3 assembler that is based on the overlap-layout-consensus algorithm (Huang and Madan 1999). *Alu* insertion-supporting read pairs were extracted from the BAM file each sample for each candidate site. We defined a window of 200 bp around the predicted breakpoint for each site, and extracted 1) read-pairs reported to support insertion at that site based on RetroSeq "discover" output, and 2) read-pairs with a soft-clipped segment at least 20bp in length with a mean quality >=Q20. We performed *de novo* assembly for all extracted reads per site, using parameters chosen to account for shorter matches expected from 101 bp reads (c 25 -j 31 -o 16 -s 251 -z 1 -c 10). CAP3 also utilizes read-pair information to report scaffolds of contigs linked together without an assembled overlap. We merged such contigs together, separated by 300 'N' characters to represent sequence gaps in the assembly. The resulting contigs and scaffolds were analyzed using RepeatMasker (Smit et al. 1996-2010) to identify *Alu*-containing contigs. Of these 2,971 candidate assembled sites were identified that contain an *Alu* element (≥30 bp match) and at least 30bp of flanking non-gap sequence in our assemblies.

Breakpoint determination

For recovery of exact breakpoints for assembled *Alu* variants, we utilized a multiple alignment-based approach similar to an approach previously described (Kidd et al. 2010a; Kidd et al. 2010b). First, orientation of candidate insertion sequences relative to the reference genome was determined using BLAT (Kent 2002). Next, candidate breakpoints were identified using *miropeats* (Parsons 1995) (-s 80) followed by a semi-automated parsing process. In turn, a global alignment was obtained for sequences from the two insertion breakpoints to the corresponding segment on the reference genome using *stretcher* (Rice et al. 2000) with default parameters. This results in pairwise alignments for two sequences aligned independently against a common third sequence. A coherent 3-way alignment was then created from the two pair-wise alignments by inserting gaps into the alignment as appropriate. Alignment columns were scored as having either a match among all three sequences ('*'), a match between the left insertion breakpoint and the reference ('1'), a match between the right insertion breakpoint and the reference ('2'), or sequence mismatch among all three sequences ('N'). We then computed a cumulative alignment score across the left and right breakpoint sequence, with matches between the target sequence and the genome sequence ('1' or '*' for the left breakpoint and '2' or '*' for the right breakpoint) resulting in a score of +1, a sequence mismatch among all three sequences a score of -1, and a match among the reference and the other breakpoint a score of -3. The same procedure was applied to the right breakpoint, except the score was tabulated from right to left across the 3-way alignment. The breakpoint and corresponding sequence on the reference was then taken as the position where the maximum cumulative score was reached. Overlapping sequence coordinates on the reference allele indicate the extent of TSDs. Of note, 1) TSDs are defined from the alignment itself, without regard to the transposable element boundaries, and 2) due to the described scoring scheme, a small degree of divergence among putative TSDs is permitted, resulting in longer TSDs for some sites than when 100% identity is required. Visualizations of the resulting aligned sequences with breakpoint annotations were constructed and subjected to manual review. When necessary, the sequences extracted for breakpoint alignment were adjusted and the alignment and scoring scheme described above was repeated until a final curated set of 1,614 assembled insertions was obtained. Information pertaining to insertion site (locus siteID and determined breakpoint), reconstituted sequence, subfamily, and predicted TSD is summarized in Table S1 for each site.

Sub-family assignment and analysis

A multiple sequence alignment was constructed of *Alu* sequences extracted from the set of 1,614 assembled insertions (Table S1) as well as 43 *Alu* consensus sequences obtained from RepBase using MUSCLE v3.8.31 (Edgar 2004) run with default parameters. Poly-A tail regions were trimmed from the resulting multiple sequence alignment and the proportion of sequence differences between each element and each sub-family consensus were tabulated. Alignments for the 1,010 sites suitable for genotyping were utilized to assess the extent of the recovered element length relative to the subfamily consensus.

Validation

A subset of 30 assembled *Alu* insertions were validated by Sanger sequencing. For primer design, we extracted ~500bp in either direction of each insertion from the hg19 reference within the UCSC Genome Browser (http://genome.ucsc.edu/). Locus-specific primer sets were designed flanking each insertion to ensure amplification across the predicted breakpoints, and the subsequent mapping of those sequenced products uniquely to the hg19 reference, thus permitting comparison of amplified fragment sizes and subsequent estimations of false detection rates. Chromosomal coordinates for each insertion were considered based on unique mapping of CAP3 locally assembled contigs of supporting reads and subsequent breakpoint analysis to the hg19 build. Extracted sequence was masked using RepeatMasker, and primers designed to include ~150bp to ~200bp in either direction of the predicted insertion, avoiding masked sequence when possible. Each primer set was analyzed by *in silico* PCR and BLAT searches to the hg19 reference in the UCSC Genome Browser to ensure site-specific, unique target amplification predictions overlapping each breakpoint, and product size predictions. All primers were designed using

Primer3v.0.4.0 (Koressaar and Remm 2007) (http://bioinfo.ut.ee/primer3-0.4.0/) and purchased from IDT. Table S2 summarizes information for loci examined, primers, and samples analyzed for each site.

All PCRs were performed with ~50ng of genomic DNA as template along with 1.5-2.5 µM Mg^{++}, 200µM dNTPs, 0.2 µM each primer, and 2.5 U Platinum Taq Polymerase (Invitrogen). All reactions were carried out under conditions of 2 min denaturation at 95 °C; 35 x cycles of [95 °C 30 sec, 55 °C – 59 °C 30 sec, 72 °C 2 min]; and a final extension at 72°C for 10 min. For each PCR reaction, 10uL were analyzed by electrophoresis in 1% agarose in 1 x TBE. Products from at least one positive PCR reactions per primer set were sequenced to confirm amplification of the desired product and its unique mapping to the hg19 reference. When possible, PCR products from an individual that was homozygous for the insertion was sequenced; in the case no homozygous insertion was observed the insertion-supporting fragment was gel-extracted (Qiagen), and the products eluted in water and subjected to sequencing. For each site, sequenced products for the pre-insertion and non-reference (*i.e*, *Alu* containing) alleles were then aligned to the corresponding reference allele and CAP3 assembled contigs of supporting reads. to ensure site-specific amplification, presence of the *Alu* insertion and TSDs, and to confirm agreement in nucleotide sequence between the PCR-validated and assembled *Alu* insertion. Individual trace alignments for the 30 validated sites are in Figure S2.

Genotyping

We performed *in silico* genotyping by mapping relevant reads to a representation of the complete insertion and reference alleles for each site. The reference allele consisted of 600 bp of sequence upstream and downstream of the start and end of any inferred TSD extracted from the hg19 reference. Based on the aligned breakpoints, insertion alleles were created by replacing the appropriate portion of this sequence with insertion sequence, accounting for inferred TSDs or target site deletions. For each site, these insertion and reference alleles constituted the target genome for mapping of reads. A BWA index was constructed from each (bwa version 0.5.9). Mapping and analysis was performed separately for each sample and each site. We extracted read-pairs with at least one read having an original mapping within the coordinates of the targeted reference allele with a MAPQ $\geq$ 20. The extracted read-pairs were then aligned to the site reference and alternative sequences using bwa aln and bwa sampe (version 0.5.9). We then calculated genotype likelihoods based on the number of read pairs mapping to the insertion or reference alleles, considering the resulting MAPQ values as error probabilities as previously described (Li 2011). Read-pairs with equal mappings between reference and insertion sequences have a MAPQ of 0 and do not contribute.

Genotypes were obtained from the resulting raw genotype likelihoods using one of two approaches. For sites on the autosomes and the pseudoautosomal region of the X chromosome, genotype likelihoods for *Alu* insertions were processed, along with previously calculated SNP genotypes using LD-aware refinement using Beagle 3.3.2 (with options maxlr=5000, niteration=10, nsamples=30, maxwindow=2000) (Browning and Browning 2007). For sites on the X chromosome, genotypes were obtained using a ploidy-aware EM algorithm that utilized the genotype likelihoods and assumed Hardy-Weinberg Equilibrium across all 53 samples. Principal component analysis was performed on the resulting autosomal genotypes using the smartpca program from the EIGENSOFT package (Patterson et al. 2006). Predicted genotypes for all 1,010 sites are provided in Table S5.

Genotype Validation

In order to validate *in silico* genotyping and permit estimation of genotyping accuracy, 11 randomly selected insertion loci were screened from a panel of ten individuals utilizing gel band assays, for a total of 110 predicted genotypes (Table S6). Locus-specific primer sets flanking each insertion locus were designed to ensure site-specific, unique target amplification predictions overlapping each breakpoint (also see Validation section above). Primer pairs per locus were then used in PCR amplification of each sample in the panel, and the products were analyzed for predicted shifting patterns following electrophoresis. All PCRs were performed with a template of 0.25ng genomic DNA, in cycling conditions of 2min at 95°C; 35x [95°C 30 sec, 55°C-59°C 30 sec, 72°C 1 min], and a final 72°C extension of 3 min.

10uL were analyzed by electrophoresis through 1.2% agarose in 1xTBE. Results were interpreted by banding patterns that supported either the unoccupied or *Alu*-containing allele, based on predicted band sizes and shifts from *in silico* PCR and information of the assembled insertion size per site. Samples utilized for sequence and PCR genotyping validations are indicated respectively in Table S6 and Figure S2.

Comparison with previous studies

Insertion site coordinates from (Hormozdiari et al. 2011; Stewart et al. 2011) were converted to GRCh37/hg19 coordinates using liftOver. Coordinates from the resulting datasets were then separately intersected with our call set using bed tools, permitting a window of 200 bp.

Data access

The *Alu* sequence data file from this study will be submitted to the NCBI Database of genomic structural variation (dbVar; http://www.ncbi.nlm.nih.gov/dbvar/) and is also available in Supplemental Table S1.

Acknowledgments

We thank Sarah Emery and Dorina Twigg for technical advice and helpful discussions, Ryan Mills for meaningful input and critical reading of the manuscript, John Moran for advice on RNA secondary structure, and Amanda Pendleton for further discussion and editorial comments. This work was supported by the National Institutes of Health Research Grant 1DP5OD009154 to J.M.K.

Authors' contributions

JHW and JMK designed the study. JHW, AB and NMB performed all necessary PCR, sequencing, and sequence-based analysis. JHW and JMK were responsible for all other data analysis involving all HGDP samples. JHW and JMK wrote the paper. All authors have read and approved the final manuscript.

References

Batzer MA, Deininger PL. 2002. Alu repeats and human genomic diversity. *Nature reviews Genetics* **3**(5): 370-379.

Bennett EA, Coleman LE, Tsui C, Pittard WS, Devine SE. 2004. Natural genetic variation caused by transposable elements in humans. *Genetics* **168**(2): 933-951.

Bennett EA, Keller H, Mills RE, Schmidt S, Moran JV, Weichenrieder O, Devine SE. 2008. Active Alu retrotransposons in the human genome. *Genome research* **18**(12): 1875-1883.

Browning BL, Browning SR. 2009. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *American journal of human genetics* **84**(2): 210-223.

Browning SR, Browning BL. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American journal of human genetics* **81**(5): 1084-1097.

Callinan PA, Batzer MA. 2006. Retrotransposable elements and human disease. *Genome dynamics* **1**: 104-115.

Callinan PA, Wang J, Herke SW, Garber RK, Liang P, Batzer MA. 2005. Alu retrotransposition-mediated deletion. *Journal of molecular biology* **348**(4): 791-800.

Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A et al. 2002. A human genome diversity cell line panel. *Science* **296**(5566): 261-262.

Carroll ML, Roy-Engel AM, Nguyen SV, Salem AH, Vogel E, Vincent B, Myers J, Ahmad Z, Nguyen L, Sammarco M et al. 2001. Large-scale analysis of the Alu Ya5 and Yb8 subfamilies and their contribution to human genomic diversity. *Journal of molecular biology* **311**(1): 17-40.

Chen JM, Ferec C, Cooper DN. 2007. Mechanism of Alu integration into the human genome. *Genomic medicine* **1**(1-2): 9-17.

Chen K, Chen L, Fan X, Wallis J, Ding L, Weinstock G. 2014. TIGRA: a targeted iterative graph routing assembler for breakpoint assembly. *Genome research* **24**(2): 310-317.

Cordaux R, Hedges DJ, Herke SW, Batzer MA. 2006. Estimating the retrotransposition rate of human Alu elements. *Gene* **373**: 134-137.

Cost GJ, Boeke JD. 1998. Targeting of human retrotransposon integration is directed by the specificity of the L1 endonuclease for regions of unusual DNA structure. *Biochemistry* **37**(51): 18081-18093.

David M, Mustafa H, Brudno M. 2013. Detecting Alu insertions from high-throughput sequencing data. *Nucleic acids research* **41**(17): e169.

Deininger P. 2011. Alu elements: know the SINEs. *Genome biology* **12**(12): 236.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**(5): 491-498.

Dewannieux M, Esnault C, Heidmann T. 2003. LINE-mediated retrotransposition of marked Alu sequences. *Nature genetics* **35**(1): 41-48.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**(5): 1792-1797.

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.

Fuhrman SA, Deininger PL, LaPorte P, Friedmann T, Geiduschek EP. 1981. Analysis of transcription of the human Alu family ubiquitous repeating element by eukaryotic RNA polymerase III. *Nucleic acids research* **9**(23): 6439-6456.

Hammer MF. 1994. A recent insertion of an alu element on the Y chromosome is a useful marker for human population studies. *Molecular biology and evolution* **11**(5): 749-761.

Hancks DC, Kazazian HH, Jr. 2012. Active human retrotransposons: variation and disease. *Current opinion in genetics & development* **22**(3): 191-203.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research* **22**(9): 1760-1774.

Hormozdiari F, Alkan C, Ventura M, Hajirasouliha I, Malig M, Hach F, Yorukoglu D, Dao P, Bakhshi M, Sahinalp SC et al. 2011. Alu repeat discovery and characterization within human genomes. *Genome research* **21**(6): 840-849.

Hormozdiari F, Konkel MK, Prado-Martinez J, Chiatante G, Herraez IH, Walker JA, Nelson B, Alkan C, Sudmant PH, Huddleston J et al. 2013. Rates and patterns of great ape retrotransposition. *Proceedings of the National Academy of Sciences of the United States of America* **110**(33): 13457-13462.

Huang X, Madan A. 1999. CAP3: A DNA sequence assembly program. *Genome research* **9**(9): 868-877.

Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G. 2012. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature genetics* **44**(2): 226-232.

Iskow RC, McCabe MT, Mills RE, Torene S, Pittard WS, Neuwald AF, Van Meir EG, Vertino PM, Devine SE. 2010. Natural mutagenesis of human genomes by endogenous retrotransposons. *Cell* **141**(7): 1253-1261.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research* **110**(1-4): 462-467.

Jurka J, Smith T. 1988. A fundamental division in the Alu family of repeated sequences. *Proceedings of the National Academy of Sciences of the United States of America* **85**(13): 4775-4778.

Kazazian HH, Jr. 2004. Mobile elements: drivers of genome evolution. *Science* **303**(5664): 1626-1632.

Keane TM, Wong K, Adams DJ. 2013. RetroSeq: transposable element discovery from next-generation sequencing data. *Bioinformatics* **29**(3): 389-390.

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome research* **12**(4): 656-664.

Kidd JM, Graves T, Newman TL, Fulton R, Hayden HS, Malig M, Kallicki J, Kaul R, Wilson RK, Eichler EE. 2010a. A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell* **143**(5): 837-847.

Kidd JM, Sampas N, Antonacci F, Graves T, Fulton R, Hayden HS, Alkan C, Malig M, Ventura M, Giannuzzi G et al. 2010b. Characterization of missing human genome sequences and copy-number polymorphic insertions. *Nature methods* **7**(5): 365-371.

Koressaar T, Remm M. 2007. Enhancements and modifications of primer design program Primer3. *Bioinformatics* **23**(10): 1289-1291.

Lander ES Linton LM Birren B Nusbaum C Zody MC Baldwin J Devon K Dewar K Doyle M FitzHugh W et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**(6822): 860-921.

Lee E, Iskow R, Yang L, Gokcumen O, Haseley P, Luquette LJ, 3rd, Lohr JG, Harris CC, Ding L, Wilson RK et al. 2012. Landscape of somatic retrotransposition in human cancers. *Science* **337**(6097): 967-971.

Lee H, Tang H. 2012. Next-generation sequencing technologies and fragment assembly algorithms. *Methods in molecular biology* **855**: 155-174.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**(21): 2987-2993.

Li H. 2012. Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. *Bioinformatics* **28**(14): 1838-1844.

Li H, Homer N. 2010. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in bioinformatics* **11**(5): 473-483.

Li S, Li R, Li H, Lu J, Li Y, Bolund L, Schierup MH, Wang J. 2013. SOAPindel: efficient identification of indels from short paired reads. *Genome research* **23**(1): 195-200.

Malhotra A, Lindberg M, Faust GG, Leibowitz ML, Clark RA, Layer RM, Quinlan AR, Hall IM. 2013. Breakpoint profiling of 64 cancer genomes reveals numerous complex rearrangements spawned by homology-independent mechanisms. *Genome research* **23**(5): 762-776.

Martin AR, Costa HA, Lappalainen T, Henn BM, Kidd JM, Yee MC, Grubert F, Cann HM, Snyder M, Montgomery SB et al. 2014. Transcriptome sequencing from diverse human populations reveals differentiated regulatory architecture. *PLoS genetics* **10**(8): e1004549.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**(9): 1297-1303.

McPherson JD Marra M Hillier L Waterston RH Chinwalla A Wallis J Sekhon M Wylie K Mardis ER Wilson RK et al. 2001. A physical map of the human genome. *Nature* **409**(6822): 934-941.

Miller JR, Koren S, Sutton G. 2010. Assembly algorithms for next-generation sequencing data. *Genomics* **95**(6): 315-327.

Mills RE, Bennett EA, Iskow RC, Devine SE. 2007. Which transposable elements are active in the human genome? *Trends in genetics : TIG* **23**(4): 183-191.

Narzisi G, O'Rawe JA, Iossifov I, Fang H, Lee YH, Wang Z, Wu Y, Lyon GJ, Wigler M, Schatz MC. 2014. Accurate de novo and transmitted indel detection in exome-capture data using microassembly. *Nature methods* **11**(10): 1033-1036.

Parsons J. 1995. Miropeats: graphical DNA sequence comparisons. *Comput Appl Biosci* **11**: 615-619.

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. *PLoS genetics* **2**(12): e190.

Piskareva O, Schmatchenko V. 2006. DNA polymerization by the reverse transcriptase of the human L1 retrotransposon on its own template in vitro. *FEBS letters* **580**(2): 661-668.

Quinlan AR. 2014. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al]* **47**: 11 12 11-11 12 34.

Reddy RM, Mohammed MH, Mande SS. 2014. MetaCAA: A clustering-aided methodology for efficient assembly of metagenomic datasets. *Genomics* **103**(2-3): 161-168.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends in genetics : TIG* **16**(6): 276-277.
Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SR, Consortium WGS, Wilkie AO, McVean G, Lunter G. 2014. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature genetics* **46**(8): 912-918.
Sen SK, Han K, Wang J, Lee J, Wang H, Callinan PA, Dyer M, Cordaux R, Liang P, Batzer MA. 2006. Human genomic deletions mediated by recombination between Alu elements. *American journal of human genetics* **79**(1): 41-53.
Simpson JT, Durbin R. 2012. Efficient de novo assembly of large genomes using compressed data structures. *Genome research* **22**(3): 549-556.
Smit AF, Hubley R, Green P. 1996-2010. RepeatMasker Open.
Solyom S, Kazazian HH, Jr. 2012. Mobile elements in the human genome: implications for disease. *Genome medicine* **4**(2): 12.
Srikanta D, Sen SK, Conlin EM, Batzer MA. 2009a. Internal priming: an opportunistic pathway for L1 and Alu retrotransposition in hominins. *Gene* **448**(2): 233-241.
Srikanta D, Sen SK, Huang CT, Conlin EM, Rhodes RM, Batzer MA. 2009b. An alternative pathway for Alu retrotransposition suggests a role in DNA double-strand break repair. *Genomics* **93**(3): 205-212.
Stewart C, Kural D, Stromberg MP, Walker JA, Konkel MK, Stutz AM, Urban AE, Grubert F, Lam HY, Lee WP et al. 2011. A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS genetics* **7**(8): e1002236.
Thung D, de Ligt J, Vissers L, Steehouwer M, Kroon M, de Vries P, Slagboom EP, Ye K, Veltman JA, Hehir-Kwa JY. 2014. Mobster: accurate detection of mobile element insertions in next generation sequencing data. *Genome biology* **15**(10): 488.
Wang J, Song L, Gonder MK, Azrak S, Ray DA, Batzer MA, Tishkoff SA, Liang P. 2006. Whole genome computational comparative genomics: A fruitful approach for ascertaining Alu insertion polymorphisms. *Gene* **365**: 11-20.
Weichenrieder O, Wild K, Strub K, Cusack S. 2000. Structure and assembly of the Alu domain of the mammalian signal recognition particle. *Nature* **408**(6809): 167-173.
Witherspoon DJ, Xing J, Zhang Y, Watkins WS, Batzer MA, Jorde LB. 2010. Mobile element scanning (ME-Scan) by targeted high-throughput sequencing. *BMC genomics* **11**: 410.
Witherspoon DJ, Zhang Y, Xing J, Watkins WS, Ha H, Batzer MA, Jorde LB. 2013. Mobile element scanning (ME-Scan) identifies thousands of novel Alu insertions in diverse human populations. *Genome research* **23**(7): 1170-1181.
Wong K, Keane TM, Stalker J, Adams DJ. 2010. Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. *Genome biology* **11**(12): R128.
Xing J, Zhang Y, Han K, Salem AH, Sen SK, Huff CD, Zhou Q, Kirkness EF, Levy S, Batzer MA et al. 2009. Mobile elements create structural variation: analysis of a complete human genome. *Genome research* **19**(9): 1516-1526.
Yang Y, Smith SA. 2013. Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. *BMC genomics* **14**: 328.
Zingler N, Willhoeft U, Brose HP, Schoder V, Jahns T, Hanschmann KM, Morrish TA, Lower J, Schumann GG. 2005. Analysis of 5' junctions of human LINE-1 and Alu retrotransposons suggests an alternative model for 5'-end attachment requiring microhomology-mediated end-joining. *Genome research* **15**(6): 780-789.

Figure Legends

Figure 1. Strategy for detection and assembly of non-reference *Alu* insertions.
Approach for reconstruction of non-reference *Alu* insertions from WGS data. 1). WGS in aligned BAM format from 53 samples were merged to a single BAM file, from which anchored mapping was implemented to identify clusters of *Alu*-supporting read pairs using the RetroSeq program by Keane *et al*. 2). Supporting read pairs and intersecting split reads were extracted for each candidate site, and 3). Subjected to a *de novo* assembly using the CAP3 overlap-layout assembler 4). Assembled *Alu*-containing contigs were then mapped to the reference genome for recovery to verify chromosomal coordinates and uniqueness of the call. 5). Breakpoints and putative TSDs from each contig were computationally predicted from a three way alignment to determine overlap of the assembled upstream and downstream flanks with the pre-insertion site from the hg19 reference.

Figure 2. Sequence analysis of assembled non-reference *Alu* insertions.
Breakpoints were determined based on alignment the 5' and 3' edge of each insertion sequence with the corresponding sequence from the hg19 reference. *Miropeats* annotation, aligned breakpoints, and Sanger sequences are shown for three representative insertions. A. Insertion at chr7:46102164 of 297 bp *Alu* with breakpoint overlap of 14 bp. *Upper left*: Alignment and breakpoint prediction of the assembled contig to hg19. Aligned breaks are shown in blue or red (leftmost or rightmost aligned nucleotides, respectively); the bracket indicates the *Alu* location in the contig relative to hg19. Repetitive elements in the reference and assembled contig are colored: LINEs, green; SINEs, purple; LTRs, orange; DNA elements, pink. *Lower left*: 3-way alignment of *Alu*-flanking assembled stretches to hg19. A '1' or '2' indicates nucleotides aligned between the assembled contig and hg19 reference upstream or downstream of the *Alu* junction. A '*' indicates positions with the same base. Terminal nucleotides of the left and right breaks are colored as above; the black bar shows contig overlap. *Right*: Alignment of the assembled contig with Sanger sequence data to the hg19 empty allele and subfamily consensus for that insertion. Blue and red bars indicate left and right breaks; shading shows assembled and validated base changes from the subfamily consensus. B. Insertion at chr8:120800779 of 221 bp with a 14 bp TSD. C. Insertion at chr11:26601646 of 310 bp with a target site deletion of 3 bp ('Δ' in the alignment). All breakpoint and alignment indications are as described in panel A.

Figure 3. Comparison of events detected in this study with other sequenced-based discovery studies.
The Venn diagram indicates overlap of *Alu* insertions found in this study (based on 53 individuals) with non-reference *Alu* insertions reported by Stewart *et al* (based on 185 samples) and Hormozdiari *et al* (8 samples). Coordinates were converted to hg19 and overlap assessed with bedtools requiring that insertions were within 200 bp of each other.

Figure 4. Length distribution of assembled AluYa5 and AluYb8 insertions.
A. Scaled representation of the *Alu*Ya5 and *Alu*Yb8 consensus and element properties. The *Alu* is comprised of two arms (*left*, blue; *right*, grey) joined by an A-rich region and having a 3' poly-A tail. The A and B boxes indicate promoter regions. A 31bp insertion distinguishes the arms; the *Alu*Yb8 has an extreme 3' 7 bp insertion relative to Ya5; the sequences are otherwise structurally

conserved. Bases involved SRP/14 binding sites are shown by the gold bar. B. The size distribution of 351 *Alu*Ya5 and 215 *Alu*Yb8 assembled insertions relative to the respective subfamily consensus. The number of assembled insertions containing an aligned nucleotide is shown against the corresponding position in the consensus.

Figure 5. Genotyping of a subset of non-reference *Alu* insertions.
Genotype validation was performed for 11 sites across 10 individuals. A. Strategy for primer design and allele detection. A single primer set was used for genotyping each locus, designed to target within 250bp of the assembled insertion coordinates relative to hg19. B. Genotyping from PCR screens and band scoring. Banding patterns supporting the unoccupied or *Alu*-containing allele were assessed following locus-specific PCR; predicted band sizes were estimated by *in silico* PCR analysis and mapped *Alu* coordinates per site. The chromosomal location of each *Alu* is indicated at left. A '+' or '-' shows the relative position of each allele. Sample information is provided for population (above) and for each individual (below). C. Principle Component Analysis PCA was performed on genotype matrix for 1,010 autosomal sites genotypes across 53 populations. A projection of the samples onto the first two Principal Components is shown.

Figure S1.
Summary of assembled insertions compared to the hg19 reference for all sites, depiction as in Figure 2. Information including the siteID, insertion coordinate, insertion size, predicted TSD, and genotyping are provided for each of 1,1614 insertions.

Figure S2. Trace alignments of validated sites
Nucleotide alignments of Sanger sequencing results with the corresponding CAP3 assemblies for each validated site. Trace alignments are shown both upstream and downstream the insertion for all validated sites. Each alignment is labeled as in Figure 2

Figure S3. Comparison of recovered alleles at chr11:35425392.
The sequence from the hg19 reference genome and the CAP3 assembled contig are compared for an insertion on chromosome 11. Extent of matching sequence between haplotypes is in yellow. The location of an indel coincident with the insertion site is depicted in blue; putative TSDs are underlined; the *Alu* insertion is in purple. Sequence present in the insertion haplotype but absent from the reference is in red. The green text corresponds to the extent of similar sequence (89.5% similarity) flanking the *Alu* insertion.

Tables

Table 1. Classification of assembled *Alu* variants.

Subfamily	Count	% Total	% Assigned
***Alu*Y**			
*Alu*Y	87	5.4%	6.0%
*Alu*Ya1	1	0.1%	0.1%
*Alu*Ya4	83	5.1%	5.7%
*Alu*Ya5	537	33.3%	37.0%
*Alu*Ya8	5	0.3%	0.3%
*Alu*Yb3a1	5	0.3%	0.3%
*Alu*Yb8	374	23.2%	25.8%
*Alu*Yb9	50	3.1%	3.4%
*Alu*Yc1	107	6.6%	7.4%
*Alu*Yc2	12	0.7%	0.8%
*Alu*Yc3	1	0.1%	0.1%
*Alu*Yd8	10	0.6%	0.7%
*Alu*Ye5	68	4.2%	4.7%
*Alu*Yf1	7	0.4%	0.5%
*Alu*Yg6	50	3.1%	3.4%
*Alu*Yh3	1	0.1%	0.1%
*Alu*Yh7	5	0.3%	0.3%
*Alu*Yi6	16	1.0%	1.1%
*Alu*Yi6_4d	7	0.4%	0.5%
*Alu*Yj4	2	0.1%	0.1%
*Alu*Yk11	4	0.2%	0.3%
*Alu*Yk12	5	0.3%	0.3%
*Alu*Yk13	5	0.3%	0.3%
***Alu*S**			
*Alu*Sc	1	0.1%	0.1%
*Alu*Sc8	1	0.1%	0.1%
*Alu*Sp	1	0.1%	0.1%
*Alu*Sq	2	0.1%	0.1%
*Alu*Sq2	1	0.1%	0.1%
*Alu*Sx3	2	0.1%	0.1%
***Alu*J**			
*Alu*Jb	2	0.1%	0.1%
Unclassified	162	10.0%	
Total	**1614**	100%	

Table 2. Truncation analysis of *Alu* variants.

Subfamily	Start	Count	% Total
***Alu*Ya5**			
	1 - 5 bp	298	84.9%
	8 - 45 bp	35	9.9%
	57 - 166 bp	18	5.1%
***Alu*Yb8**			
	1 - 5 bp	175	81.4%
	8 - 45 bp	29	13.4%
	57 - 166 bp	11	5.0%

Figure 1

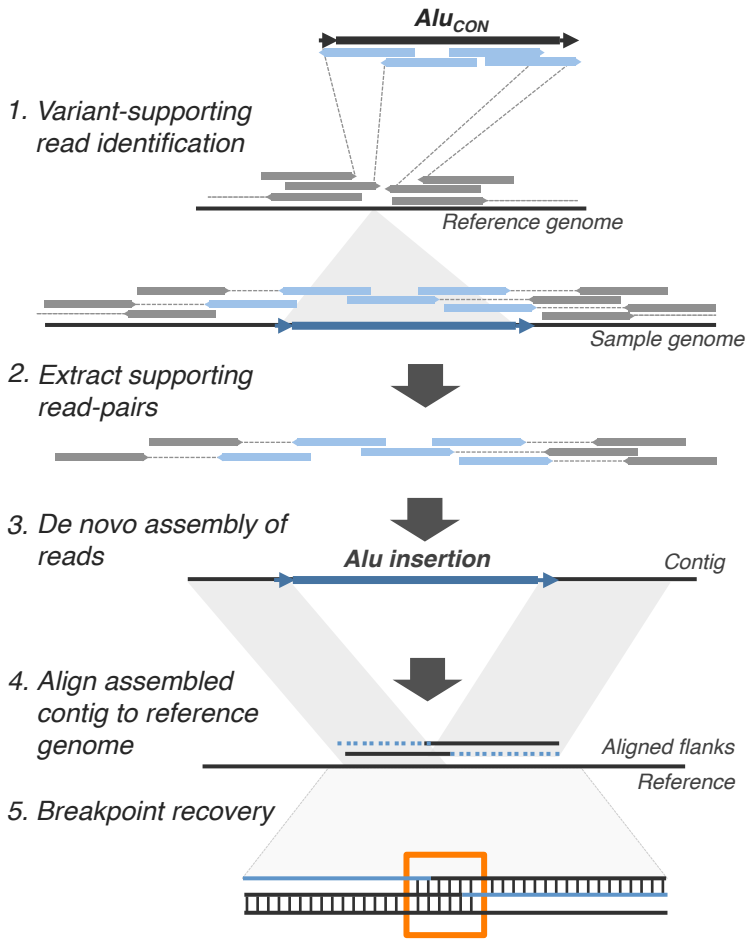
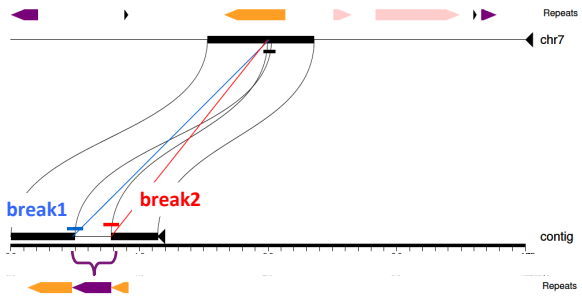

Alu$_{CON}$
1. Variant-supporting read identification
Reference genome
Sample genome
2. Extract supporting read-pairs
3. De novo assembly of reads
Alu insertion
Contig
4. Align assembled contig to reference genome
Aligned flanks
Reference
5. Breakpoint recovery

Figure 2

A.

Insertion site: chr7:46102164



siteID: chr7_46102169 Scafftig3 +
chr7:46102135-46102223
TSD size: 14 INS size: 297

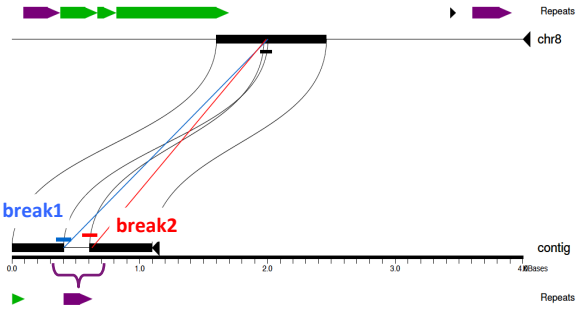
overlap

```
left   G-----------------AGAGAGCAAGAAAGACA------ACCCAAAACTG-------AAGTCTGTTTT
chrom  G-----------------AGAGAGCAAGAAAGACA------ACCCAAAACTG-------AAGTCTGTTTT
right  CGCCCGCCTCGGCCTCCCAAAGTGCTGGGATTACAGGCGTGAGCCACCGCGCCCGGCCCAAGTCTGTTTT
       11111111111111111*1**11*1*11***111111*1***111*111111111**********

left   TTTTTTTTTTTTTTTTTGAGACGGAGTCTCGCTCTGTCGCCCAGGCTGGAGTGCAGTGGCGGGATCTCGGC
chrom  TTTATAACCTCTTTTTG----GGAGT------------GACCACTCATCA----------------CTTGGG
right  TTTATAACCTCTTTTTG----GGAGT------------GACCACTCATCA----------------CATGGG
       ***2*2222*2******2222*****22222222222*2***22*222*222222222222222*12**2
```

```
                    10        20        30        40        50        60        70        80        90
                    CCACATGGTGCCTTTGCACAGGATTCCCTCACAGCATGGCAGCTGGCTTCCCCAGAGCTAGTGATCCAAGAGAGAGCAAGAAAGACAACCAA
contig.seq(1>1142)  CCACATGGTGCCTTTGCACAGGATTCCCTCACAGCATGGCAGCTGGCTTCCCCAGAGCTAGTGATCCAAGAGAGAGCAAGAAAGACAACCAA
forward(1>392)      CCACATGGTGCCTTTGCACAGGATTCCCTCACAGCATGGCAGCTGGCTTCCCCAGAGCTAGTGATCCAAGAGAGAGCAAGAAAGACAACCAA
Hg19.seq(1>831)     ccacatggtgcctttgcacaggatttccctcacagcatggcagctggcttccccagagctagtgatccaagagagagcaagaaagacaacccaa

                   100       110       120       130       140       150       160       170       180
                   AACTGaagtctgttttttttTTTTTTTTTTTGAGACGGAGTCTCGCTCTGTCGCCCAGGCTGGAGTGCAGTGGCGGGATCTCGGCTCACTGC
contig.seq(1>1142) AACTGAAGTCTGTTTTTTTTTTTTTTTTTTTTGAGACGGAGTCTCGCTCTGTCGCCCAGGCTGGAGTGCAGTGGCGGGATCTCGGCTCACTGC
forward(1>392)     AACTGAAGTCTGTTTTTTTTTTTTTTTTTTTT
Hg19.seq(1>831)    aactg----------------------------------------------------------------------------------
reverse(3>491)     ------------------------------TTTTTTTTTTTGAGACGGAGTCTCGCTCTGTCGCCCAGGCTGGAGTGCAGTGGCGGGATCTCGGCTCACTGC
AluYa5.seq(1>282)  ------------------------------------------------tggagcggagtctcgctctgtcgcccaggctggagtgcagtggcgggatctcggctcactgc

                  190       200       210       220       230       240       250       260       270
                  AAGCTCCGCCTCCCGGGTTCACGCCATTCTCCTGCCTCAGCCTCCCAAGTAGCTGGGACTACAGGCGCCCGCCACTACGCCCGGCTAATTTTT
contig.seq(1>1142) AAGCTCCGCCTCCCGGGTTCACGCCATTCTCCTGCCTCAGCCTCCCAAGTAGCTGGGACTACAGGCGCCCGCCACTACGCCCGGCTAATTTTT
Hg19.seq(1>831)    ------------------------------------------------------------------------------------------
reverse(3>491)     AAGCTCCGCCTCCCGGGTTCACGCCATTCTCCTGCCTCAGCCTCCCAAGTAGCTGGGACTACAGGCGCCCGCCACTACGCCCGGCTAATTTTT
AluYa5.seq(1>282)  aagctccgcctcccgggttcacgccattctcctgcctcagcctcccaagtagctgggactacaggcgcccgccactacgcccggctaattttt

                  280       290       300       310       320       330       340       350       360       370
                  TGTATTTTTAGTAGAGACGGGGTTTCACCGTTTTtAGCCGGGATGGTCTCGATCTCCTGACCTCGTGATCCGCCCGCCTCGGCCTCCCAAAG
contig.seq(1>1142) TGTATTTTTAGTAGAGACGGGGTTTCACCGTTTTTAGCCGGGATGGTCTCGATCTCCTGACCTCGTGATCCGCCCGCCTCGGCCTCCCAAAG
Hg19.seq(1>831)    ------------------------------------------------------------------------------------------
reverse(3>491)     TGTATTTTTAGTAGAGACGGGGTTTCACCGTTTTTAGCCGGGATGGTCTCGATCTCCTGACCTCGTGATCCGCCCGCCTCGGCCTCCCAAAG
AluYa5.seq(1>282)  tgtattttagtagagacggggtttcaccgtttt--agccgggatggtctcgatctcctgacctcgtgatccgcccgcctcggcctcccaaag

                  380       390       400       410       420       430       440       450       460
                  TGCTGGGATTACAGGCGTGAGCCACCGCGCCCGGCCcAAGTCTGTTTTTTTATAACCTCTTTTTGGGAGTGACCACTCATCACWTGGGCCATT
contig.seq(1>1142) TGCTGGGATTACAGGCGTGAGCCACCGCGCCCGGCCCAAGTCTGTTTTTTTATAACCTCTTTTTGGGAGTGACCACTCATCACATGGGCCATT
Hg19.seq(1>831)    ------------------------------------------aagtctgtttttttataacctcttttggggagtgaccactcatcactggccatt
reverse(3>491)     TGCTGGGATTACAGGCGTGAGCCACCGCGCCCGGCCCAAGTCTGTTTTTTTATAACCTCTTTTTGGGAGTGACCACTCATCACATGGGCCATT
AluYa5.seq(1>282)  tgctgggattacaggcgtgagccaccgcgcccggcc
```

B.

Insertion site: chr8:120800779



chr8:120800733-120800820
TSD size: 14 INS size: 221

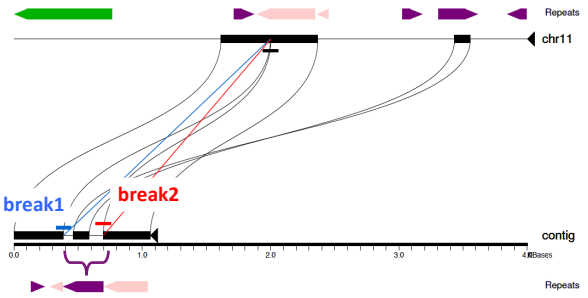
```
left   A--CCAATAT-----------------------------GTATGCTTATGTGTACAGAGAAATATAAATC
chrom  A--CCAATAT-----------------------------GTATGCTTATGTGTACAGAGAAATATAAATC
right  GAGCCGAGATTGCGCCACTGCACTCCAGCCTGGGCGACAGCGAGACTCCGTCTCAAAAAAAAAAAAAAA
       111**1*1**11111111111111111111111111111*111*11*11**1*11*1*1***1*1***11

left   AAGATATAAAATATTACACAACCATCCTGGCTAACACGGTGAAACCCCGTCTCTACTAAAAATACAAAAA
chrom  AAGATATAAAATATTACACAA-----------AA-----TGA------GTATCTGG-----AGACCTAGA
right  AAAAAAAAAAATATTACACAA-----------AA-----TGA------GTATCTGG-----AGACCTAGA
       **1*1*1**************22222222222**22222***222222**2***2222222*2**22*2*
```

overlap

```
                    10        20        30        40        50        60        70        80        90
                    GTATGATTGAAGGATCCATCTAACTCCTGCACTGTCACTTTAAGTGGTCCCTTTTAAAAAAAAAACGTACTGTTCAGAGATGTGATTTAAAAA
contig.seq(1>1097)  GTATGATTGAAGGATCCATCTAACTCCTGCACTGTCACTTTAAGTGGTCCCTTTTAAAAAAAAAACGTACTGTTCAGAGATGTGATTTAAAAA
Hg19.seq(1>862)     GTATGATTGAAGGATCCATCTAACTCCTGCACTGTCACTTTAAGTGGTCCCTTTTAAAAAAAAAACGTACTGTTCAGAGATGTGATTTAAAAA
forward(20>372)     ------------------------------ACTCCTGCACTGTCACTTTAAGTGGTCCCTTTTAAAAAAAAAACGTACTGTTCAGAGATGTGATTTAAAAA

                   100       110       120       130       140       150       160       170       180
                   ATACCAATATGTATGCTTATGTGTACAGAGAAATATAAATCAAGATATAAAATTATTACACAACCATCCTGGCTAACACGGTGAAACCCCGTCT
contig.seq(1>1097) ATACCAATATGTATGCTTATGTGTACAGAGAAATATAAATCAAGATATAAAATTATTACACAACCATCCTGGCTAACACGGTGAAACCCCGTCT
Hg19.seq(1>862)    ATACCAATATGTATGCTTATGTGTACAGAGAAATATAAATCAAGATATAAAATTATTACACAA----------------------------
forward(20>372)    ATACCAATATGTATGCTTATGTGTACAGAGAAATATAAATCAAGATATAAAATTATTACACAACCATCCTGGCTAACACGGTGAAACCCCGTCT
AluYe5.seq(83>281) ------------------------------------------------------------ccatcctggctaacacggtgaaaccccgtct

                  190       200       210       220       230       240       250       260       270
                  CTACTAAAAATACAAAAAATTAGCCGGGCGAGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATGGCGTGAACCC
contig.seq(1>1097) CTACTAAAAATACAAAAAATTAGCCGGGCGAGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATGGCGTGAACCC
Hg19.seq(1>862)    ------------------------------------------------------------------------------------------
forward(20>372)    CTACTAAAAATACAAAAAATTAGCCGGGCGAGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATGGCGTGAACCC
AluYe5.seq(83>281) ctactaaaaatacaaaaaattagccgggcgaggtggcgggcgcctgtagtcccagctactcgggaggctgaggcaggagaatggcgtgaaccc

                  280       290       300       310       320       330       340       350       360       370
                  CAGGGGGCGGAGCCTGCAGTGAGCCGAGATTGCGCCACTGCACTCCAGCCTGGGCGACAGCGAGACTCCGTCTCAaaaaaaaaaaaaaaaaaa
contig.seq(1>1097) CAGGGGGCGGAGCCTGCAGTGAGCCGAGATTGCGCCACTGCACTCCAGCCTGGGCGACAGCGAGACTCCGTCTCAAAAAAAAAAAAAAAAAAA
Hg19.seq(1>862)    ------------------------------------------------------------------------------------------
forward(20>372)    CAGGGGGCGGAGCCTGCAGTGAGCCGAGATTGCGCCACTGCACTCCAGCCTGGGCGACAGCGAGACTCCGTCTCAAAAAAAAAAAAAAAAAAA
reverse(12>209)    ------------------------------------------------------------------------AAAAAAAAAAAAAAAAAA
AluYe5.seq(83>281) cggggggcggagcctgcagtgagccgagatcgcgccactgcactccagcctgggcgacagcgagactccgtctca

                  380       390       400       410       420       430       440       450       460
                  AAAaaaaatattacacaaAATGAGTATCTGGAGACCTAGATAGGCAATTGGTGCCAAAGAGGGCAACTGGAAGATTCTGAGTCTCCAGAAATC
contig.seq(1>1097) AAAAAAAATATTACACAAAATGAGTATCTGGAGACCTAGATAGGCAATTGGTGCCAAAGAGGGCAACTGGAAGATTCTGAGTCTCCAGAAATC
Hg19.seq(1>862)    ------------------AATGAGTATCTGGAGACCTAGATAGGCAATTGGTGCCAAAGAGGGCAACTGGAAGATTCTGAGTCCCCAGAAATC
forward(20>372)    AAA
reverse(12>209)    AAAAAAAAATATTACACAAAATGAGTATCTGGAGACCTAGATAGGCAATTGGTGCCAAAGAGGGCAACTGGAAGATTCTGAGTCTCCAGAAATC
```

C.

Insertion site: chr11:26601646



chr11:26601587-26601709
TSD size: -3 INS size: 310

overlap

```
left   ATACAAAATGAGGTGCAAAAGTAATGAGAATGGAAAAAATCACAAATTTATAAATTTTTTTTAAA---ACTC
chrom  ATACAAAATGAGGTGCAAAAGTAATGAGAATGGAAAAAATCACAAATTTATAAATTTTTTTAAA---ACTC
right  TCTCC---TGACCT-CATGATCCACCCGCCTCGGCC---TCCCAAAGTGCTGGGATTACAGGCGTGACTC
       111*1111***11*1**11*111*111*11*1*111111**1****1*11*1111**111NNN11122*2

left   T----TTTCTTTTTCTTTTTCTTTTTTTTTTTTTTTTGAGACGGAGTCTCGCTCTGT
chrom  CAAAATAACATACACAAAAAGCAAAATCCAGTAGGTTAAAAATATTTTTAATCAACT
right  CAAAATAACATACACAAAAAGCAAAATCCAGTAGGTTAAAAATATTTTTAATCAACT
       22222*22*2*222*222222222222*2222*222*2*2*222*2*2*2222222*
```

```
                    10        20        30        40        50        60        70        80        90
                    TGAGAAGAGATATACAATTTTGGAAGCCCCTTTTAAGAAAAAGATTACAAAAATACCAATACAAAATGAGGTGCAAAAGTAATGAGAATGGAA
forward(10>219)     TGAGAAGAGATATACAATTTTGGAAGCCCCTTTTAAGAAAAAGATTACAAAAATACCAATACAAAATGAGGTGCAAAAGTAATGAGAATGGAA
contig.seq(1>1063)  TGAGAAGAGATATACAATTTTGGAAGCCCCTTTTAAGAAAAAGATTACAAAAATACCAATACAAAATGAGGTGCAAAAGTAATGAGAATGGAA
Hg19.seq(1>756)     TGAGAAGAGAtatacaattttggaagccccttttaagaaaaagattacaaaaataccaatacaaaatgaggtgcaaaagtaatgagaatggaa

                   100       110       120       130       140       150       160       170       180
                   AAAATCACAAATTTATAAATTTTTtttttttttttt-cttttctctttctctttttTTTTTTTTtTGAGACGGAGTCTCGCTCTGTCGCCCAGG
forward(10>219)    AAAATCACAAATTTATAAATTTTTTTTTTTTTTTTTTNCTTTTNCTTTTNCTTTTTTTTTTTT
contig.seq(1>1063) AAAATCACAAATTTATAAATTTTTTTTTTTTTTTTTT-CTTTTCTTTTTCTTTTTTTTTTTTTTTTGAGACGGAGTCTCGCTCTGTCGCCCAGG
Hg19.seq(1>756)    aaaatcacaaatttataaattttt----------------------------------------------------------------
reverse(14>518)    ------------------------------------------------------------TTTTTTTTTGAGACGGAGTCTCGCTCTGTCGCCCAGG
AluYb8.seq(18>289) ------------------------------------------------------------------------tgagacggagtctcgctctgtcgcccagg

                  190       200       210       220       230       240       250       260       270
                  CCGGACTGCGGACTGCAGTGGCGCAATCTCGGCTCACTGCAAGCTCCGCCTCCCGGGTTCACGCCATTCTCCTGCCTCAGCCTCCCGAGTAGC
contig.seq(1>1063) CCGGACTGCGGACTGCAGTGGCGCAATCTCGGCTCACTGCAAGCTCCGCCTCCCGGGTTCACGCCATTCTCCTGCCTCAGCCTCCCGAGTAGC
Hg19.seq(1>756)    ------------------------------------------------------------------------------------------
reverse(14>518)    CCGGACTGCGGACTGCAGTGGCGCAATCTCGGCTCACTGCAAGCTCCGCCTCCCGGGTTCACGCCATTCTCCTGCCTCAGCCTCCCGAGTAGC
AluYb8.seq(18>289) ccggactgcggactgcagtggcgcaatctcggctcactgcaagctccgcttcccgggttcacgccattctcctgcctcagcctcccgagtagc

                  280       290       300       310       320       330       340       350       360       370
                  TGGGACTACAGGCGCCCGCCACCGCGCCCGGCTAATTTTTGTATTTTTAGTAGAGACGGGGTTTCACCTTGTTAGCCAGGATGGTCTCGATC
contig.seq(1>1063) TGGGACTACAGGCGCCCGCCACCGCGCCCGGCTAATTTTTGTATTTTTAGTAGAGACGGGGTTTCACCTTGTTAGCCAGGATGGTCTCGATC
Hg19.seq(1>756)    ------------------------------------------------------------------------------------------
reverse(14>518)    TGGGACTACAGGCGCCCGCCACCGCGCCCGGCTAATTTTTGTATTTTTAGTAGAGACGGGGTTTCACCTTGTTAGCCAGGATGGTCTCGATC
AluYb8.seq(18>289) tgggactacaggcgcccgccaccgcgcccggctaatttttgtatttttagtagagacggggtttcaccttgttagccaggatggtctcgatc

                  380       390       400       410       420       430       440       450       460
                  TCCTGACCTCATGATCCACCCGCCTCGGCCTCCCAAAGTGCTGGGATTACAGGCGTG---ACTCCAAAATAACATACACAAAAAGCAAAATCC
contig.seq(1>1063) TCCTGACCTCATGATCCACCCGCCTCGGCCTCCCAAAGTGCTGGGATTACAGGCGTG---ACTCCAAAATAACATACACAAAAAGCAAAATCC
Hg19.seq(1>756)    ------------------------------------------------------aaaactccaaaataacatacacaaaaaagcaaaatcc
reverse(14>518)    TCCTGACCTCATGATCCACCCGCCTCGGCCTCCCAAAGTGCTGGGATTACAGGCGTG---ACTCCAAAATAACATACACAAAAAGCAAAATCC
AluYb8.seq(18>289) tcctgacctcatgatccacccgcctcggcctcccaaagtgctgggattacaggcgtg
```

Δ

Figure 3

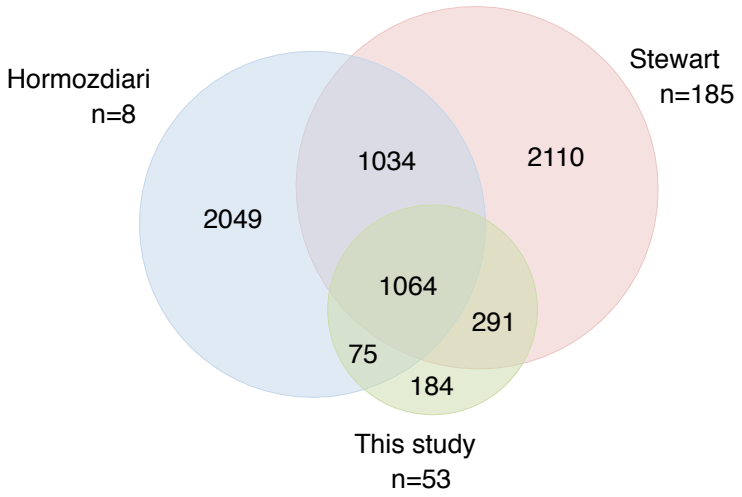

Hormozdiari
n=8
Stewart
n=185
This study
n=53
2049
1034
2110
1064
291
75
184

Figure 4

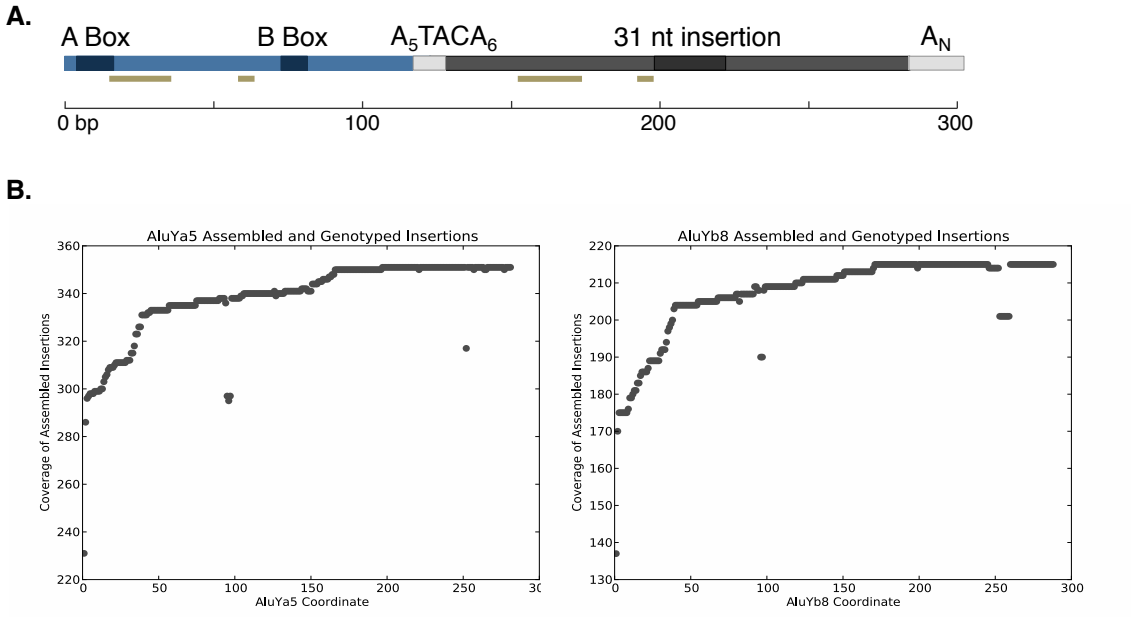
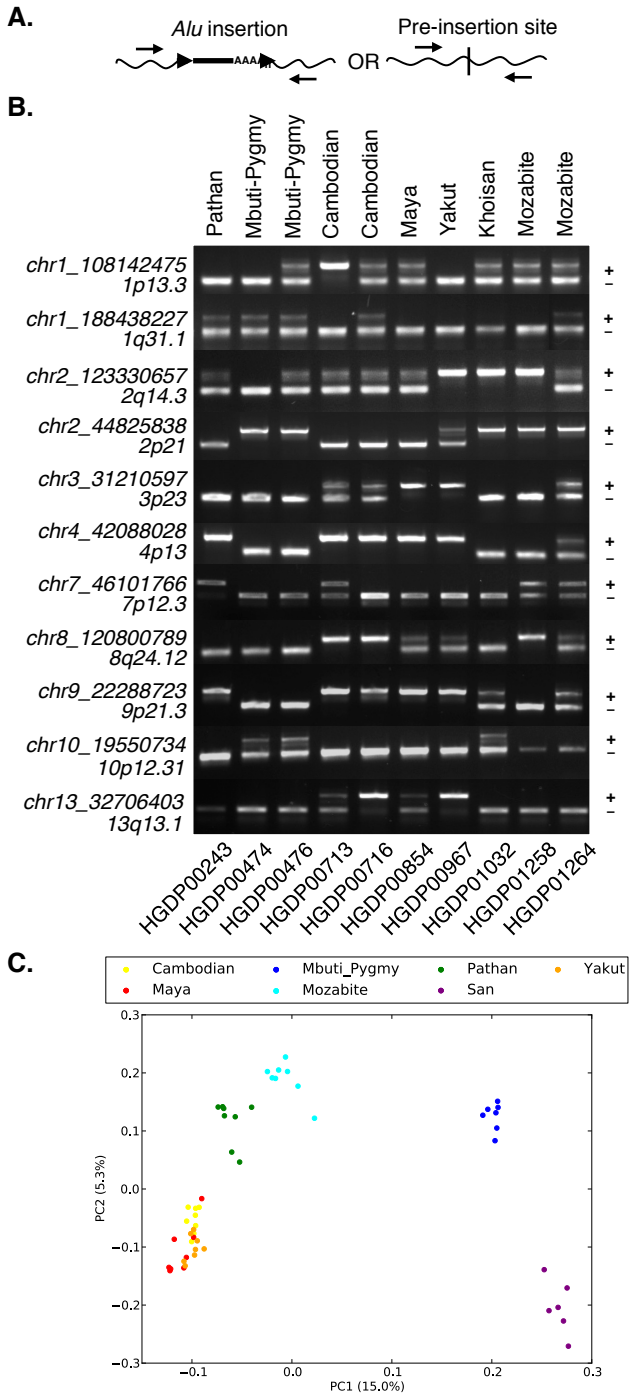

A.
A Box
B Box
A5TACA6
31 nt insertion
AN
0 bp
100
200
300
B.
AluYa5 Assembled and Genotyped Insertions
Coverage of Assembled Insertions
AluYa5 Coordinate
AluYb8 Assembled and Genotyped Insertions
Coverage of Assembled Insertions
AluYb8 Coordinate

Figure 5


A.
Alu insertion
AAAA
OR
Pre-insertion site
B.
Pathan
Mbuti-Pygmy
Mbuti-Pygmy
Cambodian
Cambodian
Maya
Yakut
Khoisan
Mozabite
Mozabite
chr1_108142475
1p13.3
chr1_188438227
1q31.1
chr2_123330657
2q14.3
chr2_44825838
2p21
chr3_31210597
3p23
chr4_42088028
4p13
chr7_46101766
7p12.3
chr8_120800789
8q24.12
chr9_22288723
9p21.3
chr10_19550734
10p12.31
chr13_32706403
13q13.1
+
–
HGDP00243
HGDP00474
HGDP00476
HGDP00713
HGDP00716
HGDP00854
HGDP00967
HGDP01032
HGDP01258
HGDP01264
C.
Cambodian
Mbuti_Pygmy
Pathan
Yakut
Maya
Mozabite
San
PC2 (5.3%)
PC1 (15.0%)
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3