Alu repeat discovery and characterization within human genomes

Fereydoun Hormozdiari¹,²,*, Can Alkan²,³,*, Mario Ventura²,⁴,*, Iman Hajirasouliha¹, Maika Malig², Faraz Hach¹, Deniz Yorukoglu¹, Phuong Dao¹, Marzieh Bakhshi¹, S. Cenk Sahinalp¹, Evan E. Eichler²,³

¹ School of Computing Science, Simon Fraser University, Burnaby, BC V5A 1S6, Canada
² Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA 98195, USA
³ Howard Hughes Medical Institute, Seattle, WA 98195, USA
⁴ Department of Genetics and Microbiology, University of Bari, 70126 Bari, Italy

* These authors contributed equally.

Corresponding author: Evan E. Eichler, Ph.D.
University of Washington School of Medicine
Howard Hughes Medical Institute
Foege S413C, 3720 15th Ave NE
Box 355065
Seattle, WA 98195
Phone: (206) 543-9526
E-mail: eee@gs.washington.edu
ABSTRACT

Human genomes are now being rapidly sequenced but not all forms of genetic variation are routinely characterized. In this study, we focus on Alu retrotransposition events and seek to characterize differences in the pattern of mobile insertion between individuals based on the analysis of eight human genomes sequenced using next-generation sequencing. Applying a rapid read-pair analysis algorithm, we discover 4,342 Alu insertions not found in the human reference genome and show that 98% of a selected subset (63/64) experimentally validate. Of these new insertions, 89% correspond to AluY elements suggesting they arose by retrotransposition. Eighty percent of the Alu insertions have not been previously reported and more novel events were detected in Africans when compared to non-African samples (76% vs. 69%). Using these data, we develop an experimental and computational screen to identify ancestry informative Alu retrotransposition events among different human populations.
INTRODUCTION

The discovery of the \textit{Alu} elements more than 30 years ago (Schmid and Deininger 1975; Houck et al. 1979) as \textasciitilde300 basepairs (bp) interspersed repeat sequences commonly found within the introns of genes (Deininger et al. 1981) prompted an active area of research to address the role of mobile elements in genome evolution and human disease (Batzer and Deininger 2002). More than one million \textit{Alu} retrotransposons comprise over 10\% of the human genome sequence (IHGSC 2001; Batzer and Deininger 2002; IHGSC 2004). They are partitioned into numerous subfamilies, which have been active at different time points during primate evolution (Price et al. 2004; Liu et al. 2009). Currently, \textasciitilde30 distinct categories of \textit{Alu} subfamilies are recognized (Mills et al. 2007) with \textit{AluYa5} and \textit{AluYb8} being most active in the human lineage (Carroll et al. 2001). \textit{Alu} retrotranspositions have numerous consequences leading to insertional mutations, gene conversion, recombination, alterations in gene expression, pseudogenization, structural variation and formation of segmental duplications (Batzer and Deininger 2002; Bailey et al. 2003; Jurka et al. 2004; Xing et al. 2009).

Traditional methods to detect \textit{Alu} insertion polymorphisms involve polymerase chain reaction (PCR) where putative polymorphic loci are genotyped one by one (Bamshad et al. 2003; Salem et al. 2003; Cordaux et al. 2007). Recently, PCR-based capture and high-throughput sequencing methods have been applied to quickly screen thousands of mobile element transposition events (Ewing and Kazazian 2010; Witherspoon et al. 2010). Although promising, these methods also require the design of appropriate PCR primers.
and are susceptible to cloning failures. Other methods to detect retrotransposons include paired-end and full fosmid sequencing (Kidd et al. 2008; Beck et al. 2010; Kidd et al. 2010), transposon insertion profiling by microarray (Huang et al. 2010), and restriction enzyme profiling followed by Sanger and Roche 454 sequencing (Iskow et al. 2010). Whole-genome shotgun sequencing (WGS) of different individuals (Levy et al. 2007; Bentley et al. 2008; Wang et al. 2008; Wheeler et al. 2008; McKernan et al. 2009) provides a resource to discover \textit{Alu} element insertions at a much higher scale and throughput. However, such findings are limited by the read length of the sequencing platform (Xing et al. 2009), and few studies have attempted to systematically discover these events at the individual genome level.

We recently described a computational method to discover mobile element insertions in genomes sequenced by paired-end next-generation sequencing (NGS) platforms (Hormozdiari et al. 2010). Based on our structural variation detection algorithm, \textit{VariationHunter} (Hormozdiari et al. 2009), our method follows the “repeat anchored mapping” approach (Kidd et al. 2008; Marques-Bonet et al. 2009) to effectively cluster paired-end reads where one end maps to an annotated repeat element and its mate maps to a position within the genome. We previously demonstrated the sensitivity and specificity of our algorithm by simulation, proving its detection power (Hormozdiari et al. 2010). Here, we apply this algorithm to construct \textit{Alu} retrotransposition maps from the genomes of eight human individuals sequenced with the Illumina platform. In addition, we also analyze one Yoruban trio from Ibadan, Nigeria, and describe the properties of parent-to-child \textit{Alu} transmission.
RESULTS

Discovery and Validation

We downloaded WGS data (http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi) from the genomes of eight human individuals generated using the Illumina paired-end sequencing technology (Table 1). We considered individuals from different populations, including three Yoruban individuals from Ibadan, Nigeria (YRI: NA18506, NA18507, and NA18508 (Bentley et al. 2008)), one Centre d’Etude du Polymorphisme Humain (CEPH) individual of European origin (Utah resident with ancestry from northwestern Europe, CEU: NA10851 (Park et al. 2010)), two Khoisan individuals from southern Africa (KB1 (Schuster et al. 2010) and HGDP01029 (Green et al. 2010)), one Han Chinese (YH (Wang et al. 2008)), and one Altaic Korean (AK1 (Kim et al. 2009)). The three Yoruban genomes constitute a parent-child trio providing us the opportunity to study transmission of Alu insertions (Table 1).

We computationally predicted novel Alu insertion loci using an algorithm that analyzes short paired-end sequence data (Hormozdiari et al. 2010). Briefly, we mapped the WGS data using mrFAST (Alkan et al. 2009) to the reference genome (National Center for Biotechnology Information (NCBI) build 36) and identified all discordant read pairs. We then realigned such reads to both the reference genome and a database of Alu consensus sequences using a modified version of mrsFAST (Hach et al. 2010). We applied VariationHunter-2 (Hormozdiari et al. 2010) to predict Alu insertions in the sequenced samples dynamically adjusting the minimum read support as a function of sequence and physical coverage of each analyzed genome (Table 1). We adjusted the number of paired-end reads supporting each Alu insertion using the genome sequenced at the
deepest coverage (NA18506) and downsampling to the observed coverage, thereby defining the minimum support for each genome sequence (Table 1). We achieved this by first identifying the clusters of paired reads supporting Alu insertion sites in the NA18506 genome and then calculating expected number of Alu clusters for a given minimum support with different depth-of-coverage resampling. We used these values to estimate the expected number of Alu insertions in other genomes by adjusting for different read depths and minimum paired-end read support cut off values.

In total, we predicted 2,451 novel Alu insertions not present in the human reference genome for the YRI trio sequence data (Supplemental Fig. S1) and a total of 4,342 Alu insertions in the entire set (Supplemental Fig. S2 and Supplemental Table S1). The chromosomal distribution patterns are shown in the context of parent-child trio (Fig. 1) and for individual genomes (Fig. 2). We find that only 13.2% (571/4,342) of these loci have been previously reported in the database of retrotransposon insertion polymorphism (dbRIP) (Wang et al. 2006). If we include two additional, recently published surveys (Iskow et al. 2010; Witherspoon et al. 2010), we find that 79.0% (3,432/4,342) of our calls are novel. Of the Alu integration sites, 33.1% (1,437/4,342) mapped within genes as opposed to the expected 37.3% of the genome based on the most current (RefSeq) gene definition (downloaded from University of California, Santa Cruz (UCSC) Genome Browser on May 20, 2010). This represents a significant (p<0.001) depletion based on simulation confirming potential selection and preferential integration within gene-poor regions of the human genome (Fig. 3). We identified 31 Alu elements that retrotransposed within an exon, of which three are predicted to disrupt a coding sequence (Fig. 4; Supplemental Tables S2 and S3).
We experimentally validated a set of Alu insertion predictions from seven of the eight genomes using PCR. We selected 29 sites from the YRI trio where integrations had occurred in relatively unique genomic regions facilitating PCR primer design. All 29 sites and the transmission genotypes within the trio were validated by PCR (Table 2 and Fig. 1B). We then tested rare Alu insertions that were predicted to be specific to an individual targeting the genomes of NA10851, AK1, KB1, and YH. We performed PCR on 10 selected sites from each of these four genomes. We removed five of the 40 PCR assays from consideration due to amplification failure, and only 25 of the remaining 35 sites confirmed the predicted Alu insertion event in the original target genome. We re-examined the 10 sites that did not initially validate by designing a second PCR assay. For the second assay, we selected oligonucleotides that map further from the predicted integration site and validated 9 out of 10 of these sites (Table 2 and Fig. 2B). Our combined results suggest excellent sensitivity (63/64) but also suggest caution in interpreting the map location precision based strictly on in silico mapping (detailed results of the PCR experiments can be seen in Supplemental Tables S4 and S5).

Since the novel Alu insertions we detected could, in principle, represent Alu deletions from the reference genome as opposed to newly retrotransposed events, we attempted to assign each Alu insertion to a particular subfamily based on the presence of diagnostic sequence variants. If the events detected were predominantly new retrotranspositions, we would predict that the AluY subfamily would predominate—the only known active Alu subfamily (Batzer et al. 1995). To perform this classification, we compared Alu anchored
sequence reads at each site to consensus sequences corresponding to each of the 31 defined \textit{Alu} subfamilies (\textit{RepBase} (Jurka et al. 2005)) and used the minimum genetic distance to assign the \textit{Alu} to a particular subfamily. In some cases where we could not distinguish between two subfamilies (i.e. they were equally divergent with respect to a consensus), both were reported (e.g. an insertion reported as \textit{AluYa5}/\textit{AluYa8} indicates that we were not able to distinguish between these two subfamilies for that particular insertion). In addition, we were not able to confidently assign 106 \textit{Alu} insertions to subfamilies, and we report such insertions as ALU (\textit{Supplemental Table S1}).

Out of the 4,236 of \textit{Alu} insertions that we could classify, 3,785 (89.3\%) of them belonged to the \textit{AluY} subfamily with the two most active members (\textit{Ya5} and \textit{Yb8}) ranked at the top. We classified 397 \textit{Alu} insertions as \textit{AluS} (~9\%) and 54 as \textit{AluJ} (~1\%). These may represent potential deletions in the reference genome or integrations that arose by endonuclease-independent pathways as opposed to new retrotransposition events driven by target-site primed reverse transcription (see \textbf{Fig. 5} for the distribution of different \textit{Alu} insertion subfamilies).

\textbf{Familial Transmission and Genotyping}

We focused on the parent-child trio (NA18506, NA18507 and NA18058) to assess the transmission characteristics of the novel \textit{Alu} insertions. We performed two sets of experiments. First, we treated each genome individually and then compared the genomes for unique and shared \textit{Alu} retrotransposons. We found no significant difference between non-transmitted and transmitted \textit{Alu} elements from either parent, although slightly more transmissions were predicted from the mother due to increased sequence coverage and X
chromosome transmissions from the mother to the proband (NA18506) (Fig. 6A). In the second experiment, we pooled all sequence data from the trio providing approximately 50-fold sequence coverage for each haplotype (Table 1). As expected, the analysis led to an increased sensitivity for transmitted \textit{Alu} insertions, significantly reducing the number of potential \textit{de novo} candidates. In this analysis, we identified only seven (0.4\%) potential \textit{de novo} insertions out of 1,720 total insertions predicted in the proband (Fig. 6B).

We attempted to validate the seven potential \textit{de novo} insertions but this proved difficult due to the repetitive nature of sequence flanking the insertion. Despite multiple attempts, we could not design a successful assay for two of the seven predicted events (Supplemental Table S6, gray rows). We tested the remaining five sites using PCR; two mapped to relatively unique sequence and in both cases the insertions were not only validated in the child but also in one of the parents, thus, were not true \textit{de novo} events. For the remaining three putative \textit{Alu} insertions, which were embedded within repetitive DNA, we developed two independent PCR assays: one where primers were selected in unique regions to create a larger PCR amplicon (for predictions in chromosomes 17 and Y) and the second with one oligonucleotide mapping within the predicted \textit{Alu} integrant (chromosome 1; Supplemental Table S6) and the other oligonucleotide mapping within repetitive flanking DNA. We applied both assay designs to test the insertions in chromosomes 17 and Y, while the second was applicable only to the \textit{Alu} insertion prediction on chromosome 1. For the chromosome 17 insertion, both the father (NA18507) and child (NA18506) showed the presence of the \textit{Alu} insertion, while for the prediction in chromosome 1, all three (NA18507, NA18508 and NA18506) showed the
presence of an *Alu* insertion. The PCR assay for the Y chromosome insertion generated multiple amplification bands due to the presence of both common repeats and duplicated sequence, so we could not draw any conclusions for this locus. In summary, of the seven initial *de novo* predictions, four were confirmed as *Alu* insertions in the child but found to be transmitted from one of the parents, and the remaining three could not be tested or interpreted.

We have also studied the parent-child trio for homozygous versus heterozygous insertions. Based on our analysis of the parent-child trio, we categorized all NA18506 *Alu* insertions as homozygous or heterozygous. To extract this information we developed a simple classifier for genotyping *Alu* insertions that considers two features: 1) the number of concordant paired-end mappings that span the loci of the predicted insertion (y-axis) and 2) the number of discordant paired-end reads that support an *Alu* insertion (x-axis). Our analysis shows that heterozygous and homozygous genotypes are accurately classified using this simple classifier (Methods). We experimentally tested the genotyping results of 29 previously validated *Alu* insertions in the YRI trio using PCR, where 28 of the 29 insertion polymorphisms were correctly genotyped (Supplemental Fig. S3). The only locus incorrectly genotyped in NA18506 (chr13: 78,169,592–78,169,605) was also the only locus incorrectly genotyped in the NA18508 and NA18507 genomes (Supplemental Fig. S3). One possible explanation may be that the region is enriched for long terminal repeat (LTR) elements and long interspersed nuclear elements (LINEs) in the flanking region confounding detection and validation.
Genome Comparisons and Population Stratification

We compared the extent of shared Alu insertion polymorphisms among the analyzed genomes in this study (Fig. 7). Based on our limited sample size of eight genomes, we found that ~50% of these novel Alu insertions were observed in two or more individuals suggesting an allele frequency greater than 10% (Fig. 7B). Due the non-uniformity in sequence coverage, this is likely an underestimate as a result of false negatives. We, therefore, repeated this analysis limiting it to four unrelated genomes each representative of a different human population, namely YH (Han Chinese), NA18506 (YRI), AK1 (Korean) and NA10851 (CEU). Of the Alu insertions, 4% (137/3,446) were shared among all four genomes but were not present in the reference genome (NCBI Build 36). Considering the diversity of the sampled genomes, we conclude that these 137 loci are common to most humans and the reference genome likely represents a rare polymorphism (Fig. 7A). We have also reported pairwise comparisons of the number of shared Alu insertions among the eight genomes (Supplemental Table S7). Note that although we find less common Alu insertions between AK1 and YH than we find between YH and NA10851, we believe this is only an artefact of lower sequence coverage in the AK1 genome compared to the other two genomes. As expected, the YRI genome shows greater genetic diversity. Approximately 59% (1,016/1,720) of the Alu insertions predicted in NA18506 are unique when compared to the other genomes, where the proportions of unique Alu integration loci in other genomes range between 37–45%. We identify 10–15% more Alu integrations in the YRI samples even after controlling for differences in sequence coverage (Table 1). In addition, fewer YRI insertions were previously reported in dbRIP (~16% (388/2,451) of YRI insertions vs. ~20% (488/2,430)
of non-African insertions). When we also compare with other recently published sets of novel Alu insertions (Iskow et al. 2010; Witherspoon et al. 2010) in addition to dbRIP, we find ~24% (589/2,451) of YRI insertions and ~30% (744/2,430) of non-African insertions were previously reported.

To assess the allele frequency distribution of the YRI Alu insertions, we selected 10 of the original 29 validated sites for which we had PCR genotyping assays and examined their allele frequency distribution more broadly among a panel of 30 individuals (10 Europeans, 10 Asians and 10 Africans). These insertions showed considerable allele frequency variation among the three populations but, as expected due to their discovery in YRI individuals, showed higher allele frequency among African samples (Table 3 and Supplemental Table S8). Three sites (alu18507_9, alu18508_6 and alu18507_11) showed the greatest enrichment among the YRI and were further tested on a larger set of human DNA samples. We genotyped 1,058 individuals from 52 populations included in the Human Genome Diversity Panel (HGDP) (Supplemental Table S9). Two loci (alu18507_11 (Supplemental Fig. S4A) and alu18508_6 (Supplemental Fig. S4B)) were largely specific to sub-Saharan populations with alu18508_6 being relatively specific to individuals of western and southern African descent (22–25% allele frequency among the Bantu, Biaka and Yoruba). In contrast, analysis of alu18507_9 showed a wider and somewhat unusual population distribution outside of Africa (Fig. 8). This allele is common among African populations (average allele frequency 37%) becoming the major allele among the Yorubans from Nigeria, and Mandenka from Senegal (54% and 58% allele frequencies). However, it is almost nonexistent among Asian populations (0.03%
allele frequency) but it is common in both European and Amerindian populations (37% allele frequency). The Sardinians, Mayans and Adygei of the Russian Caucasus show the highest non-African allele frequencies of 42%, 44% and 48%, respectively. Based on the worldwide distribution, we conclude that this insertion is ancient, predating human migrations, but has been essentially eliminated from eastern Asian populations possibly as a result of founder effect and genetic drift.

We estimated the allele frequency and extent of stratification among a subset of the newly discovered Alu integrations by examining sequence data from the 1000 Genomes Project Pilot 1 (1000GP 2010) (n = 179 individuals). We only used those genomes sequenced with paired-end Illumina technology, thus we computationally genotyped a total of 129 human genomes. We selected of 201 Alu insertions mapping to unambiguous locations on chromosome 1 based on our analysis of the initial eight genomes in this study. Next, we assayed these 201 loci within the 1000 Genomes Project (1000GP) sequence data by measuring the proportion of discordant (supportive of insertion) and concordant (supportive of null event) read data for each Alu integration locus as a surrogate for allele frequency. For this experiment, we pooled all paired-end genomic sequence data within each 1000GP population (43 YRI, 36 CEU, and 50 Asian (ASN) genomes)) and mapped reads to regions flanking the predicted insertion breakpoint. Paired-end sequence reads with one end mapping to an Alu consensus sequence and another mapping to the flanking sequence delineated the Alu insertion allele, while concordant paired-end sequences spanning the integration site and consistent with the reference genome defined the null allele. From these data, we estimated the median allele
frequency for these $Alu$ loci at 45% suggesting that these insertions are common in the general population (**Supplemental Table S10**). We predict that 10.5% (21/201) of the insertion polymorphisms on chromosome 1 are significantly stratified (Fst>0.2), with the majority (18/21) showing increased allele frequency in the YRI when compared to either the ASN or CEU populations (**Table 4**).

**DISCUSSION**

The methods we developed provide a sensitive and systematic approach to discover and genotype $Alu$ retrotransposon genetic variation in the human species using next-generation paired-end sequencing data. We have identified 4,342 novel $Alu$ insertions of which 79% are novel when compared to dbRIP (Wang et al. 2006) and other recent discovery efforts (Iskow et al. 2010; Witherspoon et al. 2010). Of the new insertions, 89% belong to the active $AluY$ subfamily suggesting that the majority arose as a result of retrotransposition as opposed to deletion or other template-directed repair processes (Batzer et al. 1995). Our analysis of eight genomes of diverse ethnicity has essentially doubled the number of $Alu$ polymorphisms currently within dbRIP providing a rich resource for future characterization.

Many aspects of the population genetics of $Alu$ insertion polymorphism are reminiscent of single nucleotide polymorphisms (SNPs). We observe greater diversity among Africans when compared to non-Africans with 10–15% more new insertions being predicted among the former. Concomitantly, a slightly larger fraction of African $Alu$ insertions are novel. While distributed throughout the genome, $Alu$ integrations are
significantly depleted within the exons and introns of genes (p<0.001) suggesting purifying selection and/or integration bias. *Alu* insertions have been shown to play an important role in creating disease alleles (Deininger and Batzer 1999). Construction of a catalogue of *Alu* repeat insertions of various allele frequencies among “unaffected” individuals is an important first step in the future discovery of pathogenic variants among patient genomes.

Our analysis of a parent-child trio shows that *de novo* *Alu* insertion events are rare and exceedingly difficult to detect and confirm. Even with high sequence coverage, we failed to find any validated *de novo* *Alu* insertions; all of the candidates represented false negatives transmitted from one of the two parents. A much larger number of trio genomes will need to be assayed before an accurate estimate of the germline mutation rate for *Alu* retrotransposition can be claimed.

The reduced level of overlap between the Han and Korean sample is influenced by differences in sequence coverage and data quality as opposed to genetic relatedness. For example, the total number of *Alu* insertions predicted for European is significantly greater than the Korean (1282 vs. 909). Similarly, both the quality and physical coverage of the Korean sample was significantly less (see Table 1, 160X physical coverage for CEU NA10851 vs. 49X coverage for Korean AK1).”

Finally, we present a rapid method for the discovery of population differentiated *Alu* insertion polymorphism. We estimate that ~10% of the *Alu* insertions we report are
stratified (Fst>0.2) between human populations. This is similar to what has been observed for SNPs discovered in the same samples (1000GP 2010). The discovery of ancestry informative *Alu* insertion polymorphisms, however, offers several advantages over traditional microsatellite and SNP markers for exploring population history. *Alu* insertions are by-and-large considered to be stable markers, unlikely to revert through precise deletion and, therefore, are homoplasy-free character states (Batzer et al. 1995). As a result, the *Alu* insertion represents the derived allele and all individuals carrying a particular insertion share identity by descent (IBD) for that locus. Genetic typing of ~100 *Alu* polymorphisms has proven useful for the unambiguous determination of the continental origin of DNA samples stripped of information (Bamshad et al. 2003). The discovery of modest numbers of geographic restricted *Alu* polymorphisms will facilitate further genetic analysis such as subgroup affiliation within population groups and have immediate application to forensics (Ray et al. 2005). Thousands of such markers that can be easily tested by PCR should be available in the very near future as human genomes become routinely sequenced.

**METHODS**

We predicted the *Alu* insertion polymorphisms using VariationHunter (Hormozdiari et al. 2010). VariationHunter discovers the mobile element insertions based on a maximum parsimony structural variation discovery algorithm (Hormozdiari et al. 2009). In the first step, the algorithm clusters the discordant paired-end reads that support the insertion of an *Alu* element. Next, VariationHunter selects the minimum number of such clusters (mobile element insertions) that cover all paired-end reads (Hormozdiari et al. 2010). For
the YRI trio, we first pooled all discordant reads and then applied the VariationHunter algorithm on the combined set of read mappings. This pooling strategy takes advantage of a priori information that the variation between individuals within a trio should be limited.

Genotyping Classifier

We use a simple linear separator classifier based on two features to genotype the Alu insertions predicted in the donor genome. The first feature is the total number of discordant paired-end reads that support the Alu insertion, and the second feature is the total number of concordant paired-end read mappings that span the insertion locus (if the concordant paired-end read has multiple mappings, a mapping location with the least edit distance is considered). Note, if we assume no paired-end read is mapped incorrectly, we expect that for a heterozygous Alu insertion the total number of paired-end reads spanning the insertion locus will be almost half of the total paired-end reads that support the Alu insertion. On the other hand, if we assume the insertion is homozygous, we then expect that the total number of concordant paired-end mappings that span the locus to be almost zero. We use a two-dimensional space to represent these insertions, where the x-axis is the number of discordant paired-end reads supporting the Alu insertion, and the y-axis is the number of concordant paired-end reads that span the locus. In this two-dimensional space, the heterozygous insertions should fall close to the line \( y = 1/2x \) and the homozygous insertions should be close to \( y = 0 \). Thus, we can easily classify
insertions based on their distance to these two lines. This is equivalent to using the line \( y = 1/4x \) as the separator between these two classes.

**PCR**

We designed PCR primers approximately 75 bp proximal and distal to the predicted \( Alu \) insertion breakpoint. In this way, if there are no \( Alu \) insertions at the tested site, we expected to see an amplification product of roughly 150 bp. In the case where we observed a \( \sim 450 \) bp fragment (150 bp + 300 bp for the \( Alu \) element), we considered the prediction as validated. We only tested the loci that were not spanned by other repetitive elements and did not intersect with segmental duplications to facilitate reliable primer design (Supplemental Tables S4 and S5).

**ACKNOWLEDGMENTS**

We thank Farhad Hormozdiari for improvements in the mrFAST aligner, and T. Brown for manuscript preparation assistance. This work was supported, in part, by Natural Sciences and Engineering Research Council of Canada (NSERC), Genome BC grants to S.C.S and NIH grants HG004120 and HG005209 to E.E.E and NSERC Alexander Graham Bell Canada Graduate Scholarships (CSG-D) to F.H and I.H. E.E.E. is an Investigator of the Howard Hughes Medical Institute. E.E.E. is on the scientific advisory board for Pacific Biosciences.
Table 1. Summary of the analyzed human genomes.

<table>
<thead>
<tr>
<th>Individual</th>
<th>Population*</th>
<th>#reads (M)illion</th>
<th>Read Length (bp)</th>
<th>Insert size (bp)</th>
<th>Seq. Coverage</th>
<th>Phys. Coverage</th>
<th>Min Support</th>
<th>Expected**</th>
<th># Alu</th>
<th>dbRIP+Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA18506</td>
<td>YRI</td>
<td>3444M</td>
<td>35</td>
<td>222</td>
<td>40.1x</td>
<td>255x</td>
<td>6</td>
<td>1500</td>
<td>1720</td>
<td>294</td>
</tr>
<tr>
<td>NA18507</td>
<td>YRI</td>
<td>2261M</td>
<td>36-41</td>
<td>208</td>
<td>27.1x</td>
<td>157x</td>
<td>6</td>
<td>1400</td>
<td>1579</td>
<td>292</td>
</tr>
<tr>
<td>NA18508</td>
<td>YRI</td>
<td>3175M</td>
<td>35</td>
<td>203</td>
<td>37x</td>
<td>214x</td>
<td>6</td>
<td>1460</td>
<td>1744</td>
<td>310</td>
</tr>
<tr>
<td>NA10851</td>
<td>CEU</td>
<td>1309M</td>
<td>36-101</td>
<td>367</td>
<td>22x</td>
<td>160x</td>
<td>5</td>
<td>1330</td>
<td>1282</td>
<td>370</td>
</tr>
<tr>
<td>AK1</td>
<td>Korean</td>
<td>1430M</td>
<td>36-106</td>
<td>132-384</td>
<td>22.5x</td>
<td>49x</td>
<td>2</td>
<td>1435</td>
<td>909</td>
<td>225</td>
</tr>
<tr>
<td>YH</td>
<td>Han Chinese</td>
<td>979M</td>
<td>35</td>
<td>135-440</td>
<td>11.4x</td>
<td>27x</td>
<td>3</td>
<td>1326</td>
<td>1160</td>
<td>307</td>
</tr>
<tr>
<td>KB1</td>
<td>Khoisan</td>
<td>842M</td>
<td>36-76</td>
<td>181</td>
<td>21x</td>
<td>25x</td>
<td>2</td>
<td>1330</td>
<td>457</td>
<td>92</td>
</tr>
<tr>
<td>HGD01029</td>
<td>Khoisan</td>
<td>161M</td>
<td>76</td>
<td>150-300</td>
<td>4x</td>
<td>12x</td>
<td>2</td>
<td>477</td>
<td>307</td>
<td>60</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>&gt; 13,601M</td>
<td>185x</td>
<td>9158</td>
<td>4342</td>
<td>571</td>
<td>910</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*YRI=Yoruba, CEU=CEPH. Min. Support=Minimum number of Alu-anchored read-pairs. **Expected based on subsampling of high coverage NA18506 genome, at specified physical coverage. dbRIP+Others = number of Alu insertions previously reported in dbRIP (Wang et al. 2006), Iskow et al. 2010 and Witherspoon et al. 2010 that intersect our predictions. The predicted numbers of Alu integrations in the YRI trio are derived from the pooled experiment. NA18506 and NA18508 genomes are sequenced by Illumina and released for public use.
Table 2. PCR validation results.

<table>
<thead>
<tr>
<th>Individual</th>
<th>PCR assays</th>
<th>Predicted Alu insertions</th>
<th>False positives</th>
<th>False negatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA18506</td>
<td>69</td>
<td>31</td>
<td>0 (0%)</td>
<td>1 (2%)</td>
</tr>
<tr>
<td>NA18507</td>
<td>69</td>
<td>30</td>
<td>1 (3%)</td>
<td>4 (10%)</td>
</tr>
<tr>
<td>NA18508</td>
<td>69</td>
<td>32</td>
<td>1 (3%)</td>
<td>1 (2%)</td>
</tr>
<tr>
<td>YH</td>
<td>40</td>
<td>10</td>
<td>0 (0%)</td>
<td>4 (13%)</td>
</tr>
<tr>
<td>AK1</td>
<td>40</td>
<td>10</td>
<td>0 (0%)</td>
<td>1 (3%)</td>
</tr>
<tr>
<td>KB1</td>
<td>40</td>
<td>10</td>
<td>0 (0%)</td>
<td>4 (13%)</td>
</tr>
<tr>
<td>NA10851</td>
<td>40</td>
<td>10</td>
<td>1 (10%)</td>
<td>2 (6%)</td>
</tr>
<tr>
<td>Total</td>
<td>367</td>
<td>103</td>
<td>3 (2%)</td>
<td>17 (6%)</td>
</tr>
</tbody>
</table>

False negatives are calculated as the number of loci that were predicted to be specific to another individual, yet PCR showed Alu insertion in the specified genome.
Table 3. *Alu* genotyping results.

<table>
<thead>
<tr>
<th>Alu Insertion Loci</th>
<th>YRI (n=10)</th>
<th>CEU (n=10)</th>
<th>CHB/JPT (n=10)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AA</td>
<td>Aa</td>
<td>aa</td>
</tr>
<tr>
<td>alu18507_5</td>
<td>1</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>alu18508_5</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>alu18507_9</td>
<td>3</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>alu18507_7</td>
<td>6</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>alu18508_6</td>
<td>6</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>alu18508_7</td>
<td>9</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>alu18507_8</td>
<td>2</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>alu18507_10</td>
<td>1</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>alu18507_11</td>
<td>8</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>alu18507_12</td>
<td>4</td>
<td>6</td>
<td>0</td>
</tr>
</tbody>
</table>

30 individuals were genotyped from YRI, CEPH and CHB/JPT HapMap populations. aa=homozygous *Alu* insertion; AA=no *Alu* insertion), f(A) and f(a) =allele frequencies of null and *Alu* insertion respectively.
Table 4. Most population stratified \(Alu\) insertions in chromosome 1.

<table>
<thead>
<tr>
<th>Coordinates</th>
<th>YRI Alu-</th>
<th>YRI Alu+</th>
<th>P-YRI Alu+</th>
<th>CEU Alu-</th>
<th>CEU Alu+</th>
<th>P-CEU Alu+</th>
<th>ASN Alu-</th>
<th>ASN Alu+</th>
<th>P-ASN Alu+</th>
<th>Fst YRI-CEU</th>
<th>Fst YRI-ASN</th>
<th>Fst CEU-ASN</th>
<th>Fst All</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr1 : 82084201</td>
<td>66</td>
<td>55</td>
<td>0.455</td>
<td>64</td>
<td>46</td>
<td>0.418</td>
<td>83</td>
<td>1</td>
<td>0.012</td>
<td>0.001</td>
<td>0.287</td>
<td>0.270</td>
<td>0.219</td>
</tr>
<tr>
<td>chr1 : 78379219</td>
<td>126</td>
<td>44</td>
<td>0.259</td>
<td>87</td>
<td>38</td>
<td>0.304</td>
<td>36</td>
<td>139</td>
<td>0.794</td>
<td>0.003</td>
<td>0.287</td>
<td>0.242</td>
<td>0.253</td>
</tr>
<tr>
<td>chr1 : 104939772</td>
<td>103</td>
<td>13</td>
<td>0.112</td>
<td>78</td>
<td>15</td>
<td>0.161</td>
<td>34</td>
<td>91</td>
<td>0.728</td>
<td>0.005</td>
<td>0.382</td>
<td>0.312</td>
<td>0.362</td>
</tr>
<tr>
<td>chr1 : 230654842</td>
<td>100</td>
<td>70</td>
<td>0.412</td>
<td>107</td>
<td>39</td>
<td>0.267</td>
<td>57</td>
<td>158</td>
<td>0.735</td>
<td>0.023</td>
<td>0.107</td>
<td>0.214</td>
<td>0.156</td>
</tr>
<tr>
<td>chr1 : 237846343</td>
<td>86</td>
<td>76</td>
<td>0.469</td>
<td>86</td>
<td>27</td>
<td>0.239</td>
<td>98</td>
<td>7</td>
<td>0.067</td>
<td>0.057</td>
<td>0.213</td>
<td>0.060</td>
<td>0.155</td>
</tr>
<tr>
<td>chr1 : 215979705</td>
<td>66</td>
<td>0</td>
<td>0.000</td>
<td>81</td>
<td>20</td>
<td>0.198</td>
<td>46</td>
<td>52</td>
<td>0.531</td>
<td>0.118</td>
<td>0.343</td>
<td>0.113</td>
<td>0.269</td>
</tr>
<tr>
<td>chr1 : 36246808</td>
<td>173</td>
<td>30</td>
<td>0.148</td>
<td>84</td>
<td>75</td>
<td>0.472</td>
<td>200</td>
<td>0</td>
<td>0.000</td>
<td>0.125</td>
<td>0.085</td>
<td>0.341</td>
<td>0.247</td>
</tr>
<tr>
<td>chr1 : 74920230</td>
<td>164</td>
<td>53</td>
<td>0.244</td>
<td>164</td>
<td>0</td>
<td>0.000</td>
<td>41</td>
<td>159</td>
<td>0.795</td>
<td>0.129</td>
<td>0.303</td>
<td>0.618</td>
<td>0.475</td>
</tr>
<tr>
<td>chr1 : 103468666</td>
<td>135</td>
<td>3</td>
<td>0.022</td>
<td>71</td>
<td>40</td>
<td>0.360</td>
<td>46</td>
<td>131</td>
<td>0.740</td>
<td>0.196</td>
<td>0.531</td>
<td>0.144</td>
<td>0.389</td>
</tr>
<tr>
<td>chr1 : 10346977</td>
<td>35</td>
<td>20</td>
<td>0.364</td>
<td>51</td>
<td>0</td>
<td>0.000</td>
<td>33</td>
<td>3</td>
<td>0.057</td>
<td>0.207</td>
<td>0.146</td>
<td>0.025</td>
<td>0.201</td>
</tr>
<tr>
<td>chr1 : 8368722</td>
<td>104</td>
<td>37</td>
<td>0.262</td>
<td>41</td>
<td>106</td>
<td>0.721</td>
<td>21</td>
<td>191</td>
<td>0.901</td>
<td>0.209</td>
<td>0.425</td>
<td>0.054</td>
<td>0.327</td>
</tr>
<tr>
<td>chr1 : 77406503</td>
<td>80</td>
<td>96</td>
<td>0.545</td>
<td>121</td>
<td>14</td>
<td>0.104</td>
<td>136</td>
<td>10</td>
<td>0.068</td>
<td>0.215</td>
<td>0.275</td>
<td>0.004</td>
<td>0.263</td>
</tr>
<tr>
<td>chr1 : 224062607</td>
<td>161</td>
<td>10</td>
<td>0.058</td>
<td>71</td>
<td>60</td>
<td>0.458</td>
<td>99</td>
<td>98</td>
<td>0.497</td>
<td>0.217</td>
<td>0.230</td>
<td>0.002</td>
<td>0.177</td>
</tr>
<tr>
<td>chr1 : 213351508</td>
<td>76</td>
<td>111</td>
<td>0.594</td>
<td>107</td>
<td>17</td>
<td>0.137</td>
<td>177</td>
<td>0</td>
<td>0.000</td>
<td>0.218</td>
<td>0.439</td>
<td>0.084</td>
<td>0.371</td>
</tr>
<tr>
<td>chr1 : 3698783</td>
<td>91</td>
<td>15</td>
<td>0.142</td>
<td>60</td>
<td>92</td>
<td>0.605</td>
<td>52</td>
<td>77</td>
<td>0.597</td>
<td>0.234</td>
<td>0.217</td>
<td>0.000</td>
<td>0.189</td>
</tr>
<tr>
<td>chr1 : 57844993</td>
<td>111</td>
<td>40</td>
<td>0.265</td>
<td>31</td>
<td>110</td>
<td>0.780</td>
<td>75</td>
<td>142</td>
<td>0.654</td>
<td>0.263</td>
<td>0.151</td>
<td>0.019</td>
<td>0.187</td>
</tr>
<tr>
<td>chr1 : 230935384</td>
<td>17</td>
<td>2</td>
<td>0.105</td>
<td>4</td>
<td>6</td>
<td>0.600</td>
<td>10</td>
<td>9</td>
<td>0.474</td>
<td>0.274</td>
<td>0.159</td>
<td>0.016</td>
<td>0.177</td>
</tr>
<tr>
<td>chr1 : 230936320</td>
<td>91</td>
<td>61</td>
<td>0.401</td>
<td>10</td>
<td>160</td>
<td>0.941</td>
<td>13</td>
<td>179</td>
<td>0.932</td>
<td>0.317</td>
<td>0.325</td>
<td>0.000</td>
<td>0.346</td>
</tr>
<tr>
<td>chr1 : 212977226</td>
<td>44</td>
<td>106</td>
<td>0.707</td>
<td>92</td>
<td>9</td>
<td>0.089</td>
<td>93</td>
<td>0</td>
<td>0.000</td>
<td>0.387</td>
<td>0.564</td>
<td>0.053</td>
<td>0.523</td>
</tr>
<tr>
<td>chr1 : 102692647</td>
<td>57</td>
<td>60</td>
<td>0.512</td>
<td>85</td>
<td>11</td>
<td>0.114</td>
<td>89</td>
<td>10</td>
<td>0.101</td>
<td>0.177</td>
<td>0.204</td>
<td>0.000</td>
<td>0.200</td>
</tr>
<tr>
<td>chr1 : 64730235</td>
<td>110</td>
<td>26</td>
<td>0.191</td>
<td>11</td>
<td>82</td>
<td>0.882</td>
<td>22</td>
<td>133</td>
<td>0.858</td>
<td>0.473</td>
<td>0.447</td>
<td>0.001</td>
<td>0.443</td>
</tr>
</tbody>
</table>

* Number of read pairs mapped in concordance with the reference genome (Alu-) and discordant read pairs consistent with the insertion allele were determined for the YRI, CEU and ASN genomes from the pilot 1000 Genomes project dataset. The proportion of discordant reads was used to estimate allele frequency and calculate Fst. The 20 most stratified loci based on pairwise comparisons among the three populations are shown (Fst>0.2).
LEGENDS

Fig. 1. *Alu* insertions predicted in the Yoruban parent-child trio. (A) Novel *Alu* integration sites for four chromosomes are shown. *Alu* sequences transmitted from the father (green) are compared to those transmitted from the mother (red). Purple bars represent centromeres, and black bars denote *Alu* insertions shared by both parents and transmitted to the offspring. In these chromosomes five *de novo* insertions were predicted in the child (blue), and further marked with a red arrow. Note that the predicted *de novo* *Alu* insertion in distal location in chromosome Y is very close to a transmitted integration, and it is difficult to see at this resolution. (B) PCR validation results for three individuals for four different loci (GM18507_*Alu*5, GM18507_*Alu*7, GM18507_*Alu*9 and GM18508_*Alu*5).

Fig. 2. Chromosomal distribution of *Alu* insertion polymorphisms. (A) *Alu* integration sites are depicted for chromosomes 8 and 13 in the genomes of NA18507, NA18508, NA10851, YH, AK1, KB1 and HGDP01029. Purple bars represent the centromeres. (B) PCR genotyping assays are shown for four different loci (YH *Alu*2, KB *Alu*4, AK1 *Alu*1 and NA10851 *Alu*8) in the genomes of seven individuals.

Fig. 3. Gene overlap analysis. 1,437/ 4,342 (33%) of predicted *Alu* insertions map within a human gene as defined by RefSeq (black dot). The histogram shows the expected distribution of gene overlap based on 1,000 permutations.

Fig. 4. Gene disruptions. The locations of three novel insertions within the coding exons of *PRAMEF4* (chr1:12,864,273-12,864,302), *CHI3L2* (chr1:111,573,857-111,573,923) and *PARP4* (chr13:23,907,208-23,807,370) are shown. Unfilled black rectangles represent the exons (and parts of exons) in the UTR, where filled rectangles show protein
coding exons. (A) and (C) demonstrate the two predicted Alu insertions mapped within a coding region while (B) is an example of one Alu insertion in the UTR.

**Fig. 5. Alu frequency distribution among subfamilies.** We show the number of predicted Alu integrations in the eight genomes separated by the inferred Alu subfamilies. As expected, AluY class had the highest frequency, where AluJ was the rarest. The Alu classes (AluY, AluS, AluJ) are sorted from youngest to the oldest.

**Fig. 6. Improved specificity in detecting de novo insertions.** A three-way comparison of novel Alu insertion polymorphisms in the YRI trio: (A) when they are predicted separately and (B) when the reads from three individuals are pooled together. Pooled coverage reduces the number of false positive de novo events for further consideration.

**Fig. 7. Shared Alu polymorphisms.** (A) A Venn diagram of shared novel Alu insertions among four genomes. The African (NA18507) genome shows the greatest number of individual-specific Alu integration polymorphisms when compared to the Asian (AK1 and YH) and European (NA10851) genomes. This effect holds even after correcting for differences in genome coverage. (B) The histogram shows the number of Alu insertions that are unique and shared among two or more genomes.

**Fig. 8. Global genome-wide distribution of Alu18507_9 polymorphisms.** Allele frequency as determined by PCR of 1,054 samples from 52 HGDP populations. Insertion polymorphism frequency (black portion of the pie chart) for Alu18507_9 is depicted.
REFERENCES


A

chr8

NA18507
NA18508
NA10851
YH
AK1
KB1
HGDP01029

chr13

NA18507
NA18508
NA10851
YH
AK1
KB1
HGDP01029

B

YH Alu2

KB1 Alu4

AK1 Alu1

NA10851 Alu8

M: log2
1: YH
2: KB1
3: AK1
4: NA10851
5: GM18506
6: GM18507
7: GM18508
8: Negative Control
A. PRAMEF4

- Exon 1
- Exon 2
- Exon 3
- Exon 4

Coding (chr1:12864261-12864843)

AluYa5 (chr1:12864273-12864302)

B. CHI3L2

- Exon 1
- Exon 2
- Exon 3
- Exon 4
- Exon 5
- Exon 6
- Exon 7
- Exon 8
- Exon 9
- Exon 10
- Exon 11

Coding (chr1:111573855-111574026)

AluYa4 (chr1:111573857-111573923)

C. PARP4

- Exon 1
- Exon 2
- Exon 3
- Exon 4
- Exon 5
- Exon 6
- Exon 7
- Exon 8
- Exon 9
- Exon 10
- Exon 11

Coding (chr13:23906532-23907612)

ALU (chr13:23907208-23907370)
Alu repeat discovery and characterization within human genomes

Fereydoun Hormozdiari, Can Alkan, Mario ventura, et al.

Genome Res. published online December 3, 2010
Access the most recent version at doi:10.1101/gr.115956.110

Supplemental Material
http://genome.cshlp.org/content/suppl/2010/12/03/gr.115956.110.DC1
http://genome.cshlp.org/content/suppl/2011/01/13/gr.115956.110.DC2

Related Content
Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads
Gerton Lunter and Martin Goodson
Genome Res. June, 2011 21: 936-939 Low-coverage sequencing: Implications for design of complex trait association studies
Yun Li, Carlo Sidore, Hyun Min Kang, et al.
Genome Res. June, 2011 21: 940-951 SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples
Si Quang Le and Richard Durbin
Cornelis A. Albers, Gerton Lunter, Daniel G. MacArthur, et al.
Genome Res. June, 2011 21: 961-973 CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing
Alexej Abyzov, Alexander E. Urban, Michael Snyder, et al.
Genome Res. June, 2011 21: 974-984 Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans
Adam D. Ewing and Haig H. Kazazian, Jr.
David A. Ray and Mark A. Batzer

P<P Published online December 3, 2010 in advance of the print journal.

Accepted Manuscript
Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.

Advance online articles have been peer reviewed and accepted for publication but have not yet appeared in the paper journal (edited, typeset versions may be posted when available prior to final publication). Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include the digital object identifier (DOIs) and date of initial publication.

To subscribe to Genome Research go to:
http://genome.cshlp.org/subscriptions

Copyright © 2010, Cold Spring Harbor Laboratory Press
License

Email Alerting Service
Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here.

Advance online articles have been peer reviewed and accepted for publication but have not yet appeared in the paper journal (edited, typeset versions may be posted when available prior to final publication). Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include the digital object identifier (DOIs) and date of initial publication.

To subscribe to Genome Research go to:
http://genome.cshlp.org/subscriptions

Copyright © 2010, Cold Spring Harbor Laboratory Press