An Alu insertion map of the Indian population: identification and analysis in 1021 genomes of the IndiGen project

P. Prakrithi1,‡, Khushboo Singhal1,2,†, Disha Sharma1,2, Abhinav Jain1,2, Rahul C. Bhoyar1, Mohamed Imran1,2, Vigneshwar Senthivel1,2, Mohit Kumar Divakar1,2, Anushree Mishra1, Vinod Scaria1,2, Sridhar Sivasubbu1,2 and Mitali Mukerji1,2,*

1CSIR Institute of Genomics and Integrative Biology, Mathura Road, New Delhi 110025, India and 2Academy of Scientific and Innovative Research (AcSIR), Ghaziabad 201002, Uttar Pradesh, India

Received April 15, 2021; Revised December 21, 2021; Editorial Decision January 17, 2022; Accepted January 25, 2022

ABSTRACT

Actively retrotransposing primate-specific Alu repeats display insertion-deletion (InDel) polymorphism through their insertion at new loci. In the global datasets, Indian populations remain under-represented and so do their Alu InDels. Here, we report the genomic landscape of Alu InDels from the recently released 1021 Indian Genomes (IndiGen) (available at https://clingen.igib.res.in/indigen). We identified 9239 polymorphic Alu insertions that include private (3831), rare (3974) and common (1434) insertions with an average of 770 insertions per individual. We achieved an 89% PCR validation of the predicted genotypes in 94 samples tested. About 60% of identified InDels are unique to IndiGen when compared to other global datasets; 23% of sites were shared with both SGDP and HGSVC; among these, 58% (1289 sites) were common polymorphisms in IndiGen. The insertions not only show a bias for genic regions, with a preference for introns but also for the associated genes showing enrichment for processes like cell morphogenesis and neurogenesis (P-value < 0.05). Approximately, 60% of InDels mapped to genes present in the OMIM database. Finally, we show that 558 InDels can serve as ancestry informative markers to segregate global populations. This study provides a valuable resource for baseline Alu InDels that would be useful in population genomics.

INTRODUCTION

Indian populations with their complex demographic history are extremely diverse. They contain thousands of endogamous sub-populations from different ethnic and linguistic lineages, with varying levels of admixture as well as social structure. There are four major linguistic lineages: Indo-European (IE), Dravidian (DR), Tibeto-Burman (TB) and Austro-Asiatic (AA) (1). Distinct geographical and climatic clines further contribute to this population diversity. So far, estimates of genetic diversity within India and its relatedness with global populations have been studied extensively using single nucleotide polymorphisms (SNPs) (1,2). Primate-specific Alu elements, present in more than a million copies in the human genome, also serve as informative markers for understanding the genetic diversity of populations (3–5). Since SNPs can occur due to replication errors, not all SNPs are identical by descent. Moreover, each Alu insertion creates a structural feature of approximately 300 bp, and therefore, they may be inherently more likely to have a practical consequence than an SNP. The younger subfamilies of Alu (AluY) are still retro-transpositionally active (6). AluYa5 is currently the most active Alu subfamily in the human lineage, followed by AluYb8, and many others including the four newly identified subfamilies termed as AluYb7a3, AluYb8b1, AluYa4a1 and AluYb10 (7,8). Active transposition of these into newer sites contributes to Alu insertion-deletion polymorphism (InDels) in the genome (9). Once retrotransposed, these are stable and define a biallelic locus based on their presence/absence at specific locations in the human genome (10). The absence of the Alu in the loci of interest is considered as the ancestral state and is regarded as the deletion allele (10). Polymorphic Alu el-
emements are therefore identical by descent and this feature makes them more informative compared to other genetic markers such as SNPs. Thus, *Alu* InDels are of enormous utility as ancestry informative markers in population genomics and association studies (11).

Notwithstanding this, identification of *Alu* insertion-deletion polymorphism has been tenuous as their size, repetitive nature and abundance make them challenging to discover and annotate even by high throughput genomic methods (12,13). Recently, next-generation genome sequencing with a higher depth of coverage in diverse populations has started yielding these polymorphic markers from different populations. One of the prominent resources for population genetic studies is the 1000 Genomes project that includes approximately 489 individuals related to India (Phase 3 release, last accessed 15 September 2021) (14). These samples have been sequenced with higher coverage in the Human Genome Structural Variation Consortium (HGSVC), a resource specific to structural variants (15). The South Asians or Indian populations represented in these datasets are mostly admixed populations and do not represent the entire genetic spectrum of India. Another attempt by the Simon Genome Diversity Project (SGDP) used 296 individuals representing 49 South Asians of which only 21 samples were from Indians (16).

Recently, the IndiGen project has provided the whole genome sequences of over 1021 individuals from different geographical locations of India (17). The high depth of coverage of these genomes allowed us to explore the *Alu* insertion-deletion (InDels) landscape in these populations. We compared IndiGen *Alu* InDels with those reported in some of the publicly available datasets (HGSVC and SGDP) and also studied their patterns within genomes. We report a total of 9239 polymorphic *Alu* insertions in Indian genomes out of which 60% are unique to IndiGen. These include 1434 common insertions with frequency ≥5%, and the remaining fraction includes private and rare insertions (with frequency <5%). We could experimentally validate 84 of the 94 predicted genotypes. The polymorphic insertions show significant bias for genic regions and are significantly enriched in cell morphogenesis and neurogenesis processes. Approximately 60% also map to loci implicated in Mendelian diseases. A set of 558 *Alu* insertions are ancestry informative and can distinguish world populations based on their genetic affinities. This study provides an enormous resource for genome-wide *Alu* InDels in the Indian population that would be useful in population genomics, disease associations as well as functional genomics studies.

**MATERIALS AND METHODS**

**Study population and datasets**

*IndiGen*. Raw BAM files were obtained from the whole genome sequencing of 1021 young, healthy, unrelated Indian individuals sequenced as a part of the IndiGen study (17). The samples were sequenced on Illumina NovaSeq 6000 platform (Illumina Inc. San Diego, CA, USA) and data were generated as 150 × 2 bp paired-end reads with ~25–30× coverage and were mapped to the human genome build GRCh38/hg38.

*Global datasets*. *Alu* insertions were also retrieved from 296 samples of Simons Genome Diversity Project (SGDP) that houses data on seven major world populations which we obtained on 25 June 2021 on request from the lead author (16) and 3202 samples of the Human Genome Structural Variation Consortium (HGSVC last accessed on 15 September 2021) (15). The latter includes high coverage genotypes of the 2504 samples from the 1000 Genomes Project Phase 3 release. These were used for comparison with our IndiGen dataset (15). Both these datasets had comparable coverage (~30×) as IndiGen.

*Pipeline for identification of Alu insertions*. The MELT (Mobile Element Locator Tool version 2.1.5, last accessed on 15 September 2021) (18) was used to detect the polymorphic *Alu* InDels as it has been earlier shown to outperform other tools in terms of accuracy, sensitivity, scalability and runtime (19) and had also been used in global Genome diversity projects (1000 Genomes Phase 3, HGSVC and SGDP) to identify *Alu* insertions. BAM files were used in the MELT- SPLIT pipeline for the identification of polymorphic *Alus* and private insertions. The identified sites were annotated with the prefix *Alu* _indigen* _Alu_ with a bash script. The chromosome-wise count, the distribution of sites within genes (in exonic, intronic, UTRs, upstream (up to 5 kb before a gene start site), and downstream regions (up to 5 kb after a gene end site) were analyzed and plotted using R (20).

*Quality checking and filtering*. To obtain a high-quality data set for downstream analyses and to avoid false positives and missing genotypes, the raw MELT calls were filtered stringently. First, the sites with no calls (ac0 flag by MELT) and those with >10% missing genotype calls were removed. Then, sites with a PASS flag by MELT, a flanking target-site duplication (TSD) defined by a MELT ASSESS score of five and in Hardy–Weinberg equilibrium (HWE) in the population were retained. HWE analysis was carried out with PLINK v1.9 (21). Sites that were either in (i) low complexity regions, (ii) not genotyped in >25% samples (s25), (iii) did not have enough supporting discordant mapped reads, (iv) without a genotyped allele (Allele count 0 filter [ac0] which were removed in the first filtering step), (v) biased reads only from one end i.e. 3′ or 5′ of the predicted insertion site (rSD) and (vi) split discordant filter (hDP) are not marked with the PASS flag and were removed (Supplementary Figure S1).

*Analysis of the identified insertions*. Variant Effect Predictor (VEP version 104; GRCh38/hg38) (22) was used to annotate the identified *Alu* insertions for their location in the genome. For selecting the consequence of a variant insertion in a gene, the results were filtered based on ‘one selected consequence per variant’ criteria in VEP. MELT annotations were used for assessing subfamily distribution. The numbers of polymorphic *Alu* InDels and their density for chromosomal regions split into 10 MB contiguous bins, i.e. percentage of *Alu* insertions occupying 10 MB regions of each chromosome, were calculated.
with customized R scripts. The correlation analyses of genic Alu insertion density with GC content, intron density, gene density for 1 MB chromosomal regions, and the number of insertions with intron lengths and gene lengths were performed using R scripts. The annotations of GC content, genes, and introns were downloaded from the UCSC Table browser Gencodev36 human genome build GRCh38/hg38.

Experimental validation of identified insertions

We carried out experiment validation of a set of polymorphic Alu insertions chosen based on their frequency group. For each of these sites, we selected six different IndiGen samples, two each of homozygous insertion (Ins/Ins), deletion (Del/Del) and heterozygous (Ins/Del) genotypes that were identified from the genome analysis. (Supplementary Table S1). The sample sets would therefore vary based on the locus studied. We designed primers flanking the site of insertions such that an amplified product without the Alu insertion would give a product of ~200–300 bp and with an Alu insertion that of ~500–600 bp (Figure 2B). Primers were designed using NCBI primer blast from ~200 bp upstream and downstream DNA sequence of each target insertion site (Supplementary Table S2). Polymerase Chain Reaction were performed using oligos synthesized by Eurofins with ~20 ng genomic DNA in a 10 µl volume reaction using Taq DNA polymerase (GeNeI, Cat no. MME23L). The reaction was carried out on Veriti™ 96-Well Thermal CyclerGreen (Cat no: 4375786). The cycling conditions were: 3 min at 95°C, (30 s at 95°C, 30 s at 55°C (except for InDel15446 Ta for which was 57°C), 30 s at 72°C) X30 cycles, 3 min at 72°C. Insertion amplicons were confirmed using Sanger sequencing. Briefly, PCR products were cleaned up using SureExtract PCR/Gel Extraction Kit (Genetix Biotech Asia Pvt. Ltd., NP-36107) as per the manufacturer’s protocol before Sanger sequencing (ABI 3130/3730) using BigDye Terminator v3.1 (ABI, Thermo Scientific, California, USA) chemistry. For Sanger sequencing, the products were purified using the PEG purification method (https://openwetware.org/wiki/PEG_purification_of_PCR_products), and the reactions were set with either forward or the reverse primer. The cycling conditions were: 3 minutes at 95°C, (10 s at 95°C, 10 s at 55°C, 4 min at 60°C) X40 cycles. UCSC Blat was done to confirm the position of the sequenced amplicon using the FASTA files generated by Chromas 2.6.5 and the presence of the Alu insertion was confirmed using rmblast of RepeatMasker v3.0 with default parameters.

Comparison with global datasets

Novel Alu InDels in the IndiGen samples were discovered through comparisons with the HGSCV and SGDP datasets. Many Mobile Element Insertions (MEIs) discovered in the two datasets had identical positions. To account for the positional differences contributed by Target Site Duplication (TSD) length and the respective Alu coordinates assigned by different MELT versions, we allowed small windows of positional tolerance (up to ±50 bp). The overlap was substantially increased for up to ±20 bp, especially with HGSCV data, and hence this cut-off was used to compare the positions of Alus in these datasets. (Supplementary Figure S2)

Population genetics

We wanted to ascertain the utility of the common Alu InDel polymorphisms for population genomics studies. To identify the minimum number of insertions required to differentiate among the populations, we carried out PCA analysis with Alu insertions. PCA analysis with Plink (v1.07) (21) using the genotype data of the sites shared between the IndiGen, SGDP and HGSCV was performed. FST analysis was carried out with VCFTOOLS using the Weir-Cockerham estimator. FST values for the insertions were calculated across the major ancestral groups, i.e. Europeans, East Asians, Africans and South Asians (IndiGen included), and PCA analysis was also performed with the top (75%, 50%, 25% and 10%) differentiating Alu InDels.

Pathway enrichment analysis

The TopGene (TopFun) (https://toppgene.cchmc.org/enrichment.jsp) (23) was used to perform molecular function and biological processes analysis of the genes with Alu InDels. P < 0.05 was set as the threshold value. Pathways and processes that crossed significance cut-off of q values FDR B&Y < 0.05 are reported.

RESULTS

Identification of polymorphic Alu insertions in IndiGen

We identified 22 109 potential Alu insertions from the whole genome sequencing data of 1021 individuals using the MELT-SPLIT pipeline. After the stringent quality filtering steps (detailed in the Materials and Methods section; Supplementary Figure S1), 9239 polymorphic Alu insertions were retained with an average of 770 insertions per individual. About 90% of the insertions were >250 bp (Supplementary Figure S3a) which implies that the majority of them are full-length insertions mediated by retrotransposition events. Also, the target site duplication (TSD) length distribution of Alu insertions varied from 1 to 29 bp; however, the mean TSD length was around 15 bp (Supplementary Figure S3b). In general, the number of identified insertions was observed to be proportional to the size of the chromosomes with chromosome 2 having maximum insertions and chromosome Y the least (Figure 1A). However, the density of the insertions did not correlate with the chromosome size. On average, there were ~27 polymorphic insertions per 10 MB region of the autosomes. The density of Alu insertions was highest in chromosome 4 and lowest in chromosome Y (Figure 1B).
insertions. As anticipated, 99.3% of insertions were of the AluY family, followed by 0.6% from AluS and with the minimal representation from the oldest family AluJ (0.1%) (Figure 1C, inset). Approximately 45% of AluY insertions are contributed by the subfamilies AluYa5, AluYa4 and AluYb8 (Figure 1C and Supplementary Figure S4a). AluSz contributed the maximum number of insertions in the AluS (Supplementary Figure S4b). Only seven AluJo insertions were identified.

Patterns of distribution of InDels in the genome. Alu repeats have been earlier reported to have a non-random distribution (24). We, therefore, wanted to see whether the polymorphic Alu insertions also have a preference for specific genomic regions. To our surprise, 72.3% (6683) of Alu insertions were observed within the coding and regulatory regions (Figure 1D, inset). Among these, 79% of the sites were intronic, ~14% inserted in upstream and downstream regions and only 1% were exonic. A minor fraction (<1.5%) of the sites also mapped to the 3′UTRs and 5′UTRs (Figure 1D). We then assessed the correlation of different genomic features with the number and density of polymorphic Alu insertions (Figure 2A). Overall, polymorphic Alu insertions are positively correlated with the intron length although the extent of the correlation differs across chromosomes (Supplementary Figure S5). There was also a significant positive correlation (P-value = 2.2 × 10^{-16}) of Alu insertions with gene lengths across all chromosomes. Noteworthily, we observed a significant positive association between intron density and gene Alu density (P-value = 5.2 × 10^{-44}) but not between gene density and genic Alu density (P-value = 0.081) (Supplementary Table 3a). Chromosomes 9, 13–16 and 22 exhibited the highest correlations with GC content (Supplementary Table S3b). The Alu insertion density showed a significant negative association in the intergenic regions (data not shown). We also observed a similar pattern of correlation with the whole-genome fixed Alus that were retrieved from the UCSC table browser human genome build GRCh38/hg38 (Figure 2B). This biased distribution of Alu insertion-deletion polymorphism corroborates with their overall biased representations that have been reported from the first draft of human genome sequencing projects (24–26).

Experimental validation of polymorphic insertions
About 84 of the 94 (i.e. 89%) MELT predicted genotypes were validated through PCR and Sanger sequencing (Supplementary Table S4, see Materials and Methods). Since polymorphic Alu insertions are biallelic markers, there are three possible genotypes viz, homozygous insertion (Ins/Ins), heterozygous (Ins/Del) and homozygous deletions (Del/Del) (Figure 3A). We could validate all three genotypes for 17 out of 18 insertions selected, which in-
Figure 2. Correlation of Alus with GC content, gene density, gene length, intron length and intron density. (A) Polymorphic genic Alu InDels density identified in 1021 IndiGenomes. (B) Fixed Alus in the reference human genome retrieved from the UCSC genome browser GRCh38/hg38. * marks parameters where r values for all chromosomes are significant. For correlation with GC content and gene density, only a few chromosomes did not pass the significance cut-off as provided in Supplementary Table S3. Dotted lines connecting the different points is to show the trend across different chromosomes.

Included 10 common, 4 rare and 3 private insertions. However, 1 private insertion could not be validated. A representative image of different genotypes of a subset of 12 Alu insertions is shown in Figure 3B. Details of the represented Alu insertions are given in Table 1.

In all the cases RepeatMasker identified the presence of an AluY element in the amplicons as expected (Supplementary Figure S6a-i). The two private insertions validated i.e. InDel_15446 and InDel_5507 were located in intron 6 of the HDAC7 gene and 3′ UTR (exon 4/4) of TLR1 gene, respectively. The presence of Alu Y insertion in the sequenced amplicon from positions 88 to 399 bases in InDel_15446 was confirmed using Repeat Masker. (Figure 3B and Supplementary Figure S6g). InDel_5507 showed the presence of AluY element from 115–393 bases (Supplementary Figure S6h). Though most of the validated insertions were intronic, few were present in the UTRs as well. For instance, InDel_4893 is a rare variant present in the 3′ UTR (exon 3/3) of the PTX3 gene, with the AluY element present from 55 to 343 bases. This position of Alu overlaps with an enhancer element in UCSC implying the presence of Alu could have an impact on regulation (Supplementary Figure S6i). Overall, we could experimentally validate 91% of the polymorphic Alu insertions identified in our study.

Comparison with the global datasets
We compared IndiGen data with HGSVC and SGDP datasets and observed 60% (5570) of Alu insertions to be unique to the IndiGen data (novel) (Figure 4A). Approximately half of the shared polymorphic Alu insertions between IndiGen, HGSVC, and SGDP are common polymorphisms with minor allele frequency (MAF ≥5%) in all these datasets. The remaining sites reported are found in different frequencies across the populations compared. For example, 88 sites with a frequency >5% in IndiGen have comparatively lower frequencies (rare polymorphisms) in the other two global datasets. About 68 common insertions in IndiGen are found to be rare in the South Asians of HGSVC. A very few variants that were private to IndiGen overlapped with other datasets as well (214 with SGDP and 40 of these with HGSVC) and hence could not be called private insertions in a global perspective, but we refer to them as ‘private to IndiGen’ (Supplementary Table S5 and Figure S7). A summary of the three datasets is provided in Table 2.

Utility as Ancestry Informative Markers (AIMs)
Polymorphic Alu elements have been used as ancestry informative markers in population genetic studies (3). We wanted
Figure 3. Validation of polymorphic Alu InDels identified in 1021 Indigen samples (A) Schematic of validation approach for selected polymorphic loci. PCR primers marked with red arrows are designed flanking the site of Alu insertions leading to expected amplicons of different sizes with and without Alu insertions; the three possible genotypes are also shown (C) Representative gel electrophoresis image of the three genotypes: Ins/Ins (single amplicon at ~600 bp), Ins/Del (two amplicons; insertion at ~600 bp and deletion at ~300 bp) and Del/Del (single band at ~300 bp) for loci listed in Table 1. The band at ~850 bp in InDel_15446 is non-specific.

Table 1. Details of polymorphic Alu insertions that are represented in Figure 3B.

| S.No. | Gene   | ID          | Expected amplicon size (bp) | Alu Size (bp) | Frequency Group |
|-------|--------|-------------|-----------------------------|---------------|----------------|
| 1     | FRAS1  | InDel_5797  | No Insertion                | 298           | 579            | 281            | Common         |
| 2     | LRRK2  | InDel_15374 | With insertion              | 281           | 558            | 277            | Common         |
| 3     | SLC30A49 | InDel_5542 |                             | 254           | 535            | 281            | Common         |
| 4     | SEMA6D | InDel_17932 |                             | 253           | 533            | 280            | Common         |
| 5     | COL4A2 | InDel_16968 |                             | 245           | 526            | 281            | Common         |
| 6     | NBAS_1 | InDel_1848  |                             | 300           | 581            | 281            | Common         |
| 7     | PTPRN2 | InDel_10992 |                             | 254           | 533            | 279            | Common         |
| 8     | XDH    | InDel_1977  |                             | 220           | 501            | 281            | Common         |
| 9     | CSMD1  | InDel_11019 |                             | 204           | 484            | 280            | Rare           |
| 10    | ITPR1  | InDel_3672  |                             | 280           | 561            | 281            | Rare           |
| 11    | IL17RD | InDel_4112  |                             | 298           | 579            | 281            | Rare           |
| 12    | HDAC7  | InDel_15446 |                             | 254           | 534            | 280            | Private        |

Table 2. Summary statistics from IndiGen, HGSVC and SGDP datasets

| Parameters                                      | IndiGen  | HGSVC                  | SGDP                   |
|------------------------------------------------|----------|------------------------|------------------------|
| Sample size                                     | 1021     | 3202 (687 of South Asian ancestry) | 296 (49 of South Asian ancestry) |
| Coverage                                       | 25–30×   | 30×                    | 30×                    |
| Total QC filtered Alu insertions                | 9239     | 9331                   | 11 661                 |
| MAF ≥ 5% (common polymorphisms)                 | 1434     | 3546                   | 1941                   |
| Average insertion sites per individual          | 614      | 1705                   | 835                    |
to ascertain the utility of the common Alu InDel polymorphisms for population genomics studies. PCA analysis using Plink (v1.07) (21) using the genotype data of the 24% sites (2232 sites) shared between IndiGen, HGSCV and SGDP revealed their proximity to the South Asian populations of the latter two global datasets (Figure 4B). Only a small percentage of IndiGen (71 samples) that were of the Tibeto-Burman ancestry was closer to the East Asians as expected (data for the analysis with SNPs are unpublished. Similar results were observed with Alus as AIMs, data not shown). Further, to identify the minimum number of insertions required to differentiate between the populations we carried out PCA analysis with Alu insertions, all insertions as well as top 75%, 50%, 25% and 10% insertions with high FST values. A minimum number of 223 Alu insertions could cluster different populations. However, 558 insertions i.e. top 25% with high FST showed results as good as all shared insertions (Figure 4C). About 58.5% of the shared sites in the three datasets were observed with a common frequency (MAF ≥ 5%) in IndiGen. For each set of Alus sorted based on their FST values, about 50% of the insertions were found to be of common frequency in IndiGen (MAF ≥ 5%) (Supplementary Figure S7 and Table S6).

Functional impact of Alu insertions

About 6683 (72.3% of total insertions) of the genic insertions mapped to 4209 genes, 60% of which are present in genes reported in the OMIM database (27) implying the likely importance of these insertions in Mendelian diseases (Supplementary Figure S8). Toppfun (23) biological pathway analysis with q-value FDR B&Y < 0.05 of all the genes revealed significant enrichment of biological processes like cell morphogenesis, cell adhesion, nervous system development, axonogenesis and synaptic transmission (Supplementary Table S7). Since Alu elements have been implicated in neurodevelopment and neurological diseases (28), we wanted to see how many genes in our data are implicated in neurological diseases. For this, we intersected (Venny) our gene list with the NDDVD database that has 289 genes associated with 37 different neurodegenerative diseases. We found that 62 out of 289 genes (i.e. 21%) have Alu insertions in them. 24 of those genes have been implicated in Alzheimer's disease (Supplementary Figure S9a, b).

DISCUSSION

Polymorphic Alu insertions arise due to recent retrotransposition events in the human genome. Alu insertion/deletion polymorphisms have been of enormous utility in population genomics studies (as they are one of the most informative markers for inferring ancestry), forensic applications and disease association studies (3,29,30). In this study, we report the genomic landscape of the polymorphic Alu insertions in 1021 Indian individuals. There were 9239 polymorphic Alus with an average of 770 insertions per individual. Earlier studies have reported that the average number of polymorphic Alu insertions per individual vary from 1283 (31) to 1574 (32). Most of the insertions observed were of full length as shown by their size distribution and variable target site duplication length, suggesting that they could be transposed by canonical L1 transposase activity (33,34).

We observe that ~99% of the polymorphic insertions are contributed by the most retrotranspositionally active AluY subfamily followed by AluS and very few from AluJ. Though older Alu subfamilies AluS and AluJ elements have been presumed to be inactive for the past 35 million years, there are reports of some of them being active in the human genome. For instance, Bennett et al. report four insertions from ancient AluS subfamilies two of them were intact Alus and two were fragmented copies (31). In another study done by Mills et al., 3.3% of total Alu insertions were found to be from the AluS subfamily in a comparison of human and chimpanzee retrotransposon insertions (35). The Alu elements harbor a large number of regulatory motifs and retrotransposition of these elements could provide novel regulatory sites (36–39). The insertions predominantly map to the coding and regulatory regions compared to intergenic regions. Of these, nearly 79% were in the introns and their densities significantly correlated with the length of the introns as well as genes (P-value < 2.2 × 10^{-16})
These patterns are consistent with the overall distribution of fixed Alu repeats in the genome. It remains to be explored whether the propensity of new Alu insertions in the genic region is driven by sites created by pre-existing Alus or is favored due to epigenetic differences in the vicinity of expressing genes. Insertions in genic regions could potentially alter the regulatory networks that are enriched in sites that could affect the expression of genes and transcripts through altered methylation, expression, editing, splicing, localization, etc (40–43). Whether polymorphic Alu insertions and private insertions also harbor these regulatory sites remains an aspect for future investigation. New insertions have also been implicated in many genetic diseases (44–48). In a study by Payer et al., 809 polymorphic Alu elements have been mapped to 1159 loci implicated in disease risk by genome-wide association study (GWAS) (P-value < 10^{-8}). About 44 of these Alu elements were observed to be in high linkage disequilibrium (r^2 > 0.7) with the trait-associated SNPs (49).

The patterns of the frequency distribution of Alu insertion-deletion polymorphisms vary across populations and they prove as good markers for studying population structure and evolution (10,50–55). In our study, the frequencies of Alu insertions varied over a wide range from 0.04% to 99.90% suggesting different time scales of their insertions. The proportions of private, rare, and common insertions were 41.4%, 43.01% and 15.52%, respectively. Compared with the global datasets 60% of Alu insertions were observed to be unique to the IndiGen data highlighting the utility of these Alu insertions for understanding the genetic structure of the Indian population. From a set of 2222 insertions that were shared among IndiGen, HGSVC and SGDP, 223 insertions were sufficient to cluster the different world populations. In a study by Rishishwar et al., on 2504 individuals from the 1000 Genomes project, among the 16192 loci of genome-wide polyTEs (polymorphic Alu, L1 and SVA) polymorphic Alus showed the highest levels of resolution for human evolutionary relationships, ascribed to their higher diversity and numbers (3).

We could achieve 89% validation of the selected polymorphic Alu insertions using the PCR and Sanger Sequencing. Though many insertions were present in the intronic regions few of them were also present in the UTRs. InDel_4893 present in the 3′UTR of PTX3 gene has an overlapping enhancer element. The presence of polymorphic Alu in the enhancer region could lead to differential regulation of gene expression under the control of that enhancer and thus different outcomes. The impact of the presence of Alu elements in such regulatory regions would need detailed experimental studies. Alu insertions identified in our study are significantly enriched in genes with roles in the neurogenesis process. Considering that insertions in the genes can have an impact on gene expression and thereby its function, these insertions could have an impact on the neurological pathways, which would need further detailed validation studies (28). This further adds to the increasing body of evidence of involvement of Alus in neurological diseases primarily by altering mitochondrial functions (56). Alu insertions have been reported to contribute to large-scale structural variations and genome rearrangements in many diseases (28,57–59).

Considering India’s vast population, including more samples would give us even better insight into the landscape of polymorphic Alus. In summary, this study for the first time provides a spectrum of genome-wide active Alu insertions in the Indian population some of which are shared and a majority of them being novel. This baseline resource would be of enormous utility in understanding population structure as well as identifying new disease risk loci that might be specific to the Indian population.

CONCLUSION

Polymorphic Alu insertions can influence genome structure and function and serve as ancestry informative markers. With the recent release of IndiGen data from 1021 individuals, it has now become possible to study the genomic landscape of Alu InDel polymorphisms in the Indian population. This study from IndiGen adds to the repertoire of Alu InDel polymorphisms to the global dataset and enriches it in terms of diversity. This would be of enormous utility in population genomics and assignment of ancestry in association studies. Variability in the presence and expression of Alu insertion-deletion polymorphisms could confer population-specific differences in phenotypes and diseases.

DATA AVAILABILITY

The Alu InDelS unique to IndiGenomes have been submitted to the NCBI Variation Submission Portal, dbVAR with submission ID: nstd215 and also made available for download on the IndiGenomes database website http://clingen.igib.res.in/indigen/.

The codes used for InDel identification and analysis can be found on GitHub at https://github.com/Prakrithi-P/ALU_IndiGen.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We are grateful to W Scott Watkins (Department of Human Genetics, University of Utah) for sharing the VCF file containing Alu insertion elements discovered in the SGDP with us for our comparison study. We would like to acknowledge Vineet Jha (Persistent LABS, Persistent Systems Ltd., Pune, Maharashtra, India) for his suggestions in pipeline standardization and help in data retrieval, Anjali Bajaj for her help in getting IndiGen samples for experimental validation, Pooja Sharma for her help in setting up the sequencing run, and Kavita Pandhare for uploading the polymorphic Alu data on the IndiGen website. We are thankful to Aniket Bhattacharya (Child Health Institute and Robert Wood Johnson Medical School, Rutgers University, NJ 08901, USA) for providing a critical review of the manuscript.

Author’s contribution: P.P. did the pipeline standardization, annotations, data analysis and visualization. K.S. did functional enrichment, experimental validation and helped in improving data visualization. D.S. ran the MELT-SPLIT
pipeline and generated the Alu VCF file from raw BAM files. M.M., K.S. and P.P. wrote the manuscript. R.C.B., A.J., M.I., V.Sc, M.K.D. and A.M. were involved in sample collection, sequencing and IndiGen data processing. V.S. and S.S.B. supervised the IndiGen project. M.M. designed, conceptualized and supervised the overall study.

**FUNDING**

Council of Scientific and Industrial Research, India [MLP1809/MLP2021, GAP2026 to P.P., K.S.]; Intel Research Fellowship (to D.S.).

Conflict of interest statement. None Declared.

**REFERENCES**

1. Indian Genome Variation Consortium (2008) Genetic landscape of the people of india: a canvas for disease gene exploration. *J. Genet.*, 87, 3–20.

2. Xing, J., Watkins, W.S., Hu, Y., Huff, C.D., Sabo, A., Muzny, D.M., Bamshad, M.J., Gibbs, R.A., Jorde, L.B. and Yu, F. (2010) Genetic diversity in India and the inference of Eurasian population expansion. *Genome Biol.*, 11, R113.

3. Rishishwar, L., Tellez Villa, C.E. and Jordan, I.K. (2015) Transposable element polymorphisms recapitulate human evolution. *Mob. DNA*, 6, 21.

4. Lander, E.S., Linton, L.M., Birren, B., Nussbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. et al. (2001) Initial sequencing and analysis of the human genome. *Nature*, 409, 860–921.

5. Watkins, W.S., Rogers, A.R., Ostler, C.T., Wooding, S., Bamshad, M.J., Brassington, A.-M.E., Carroll, M.L., Nguyen, S.V., Walker, J.A., Prasad, B.V.R. et al. (2003) Genetic variation among world populations: inferences from 100 alu insertion polymorphisms. *Genome Res.*, 13, 1607–1618.

6. Bennett, E.A., Keller, H., Mills, R.E., Schmidt, S., Moran, J.V., Weiunchneider, O. and Devine, S.E. (2008) Active alu retrotransposons in the human genome. *Genome Res.*, 18, 1875–1883.

7. Konkel, M.K., Walker, J.A., Hotard, A.B., Ranck, M.C., Fontenot, C.C., Storer, J., Stewart, C., Marth, G.T. and the 1000 Genomes Consortium and Batzer, M.A. (2015) Sequence analysis and characterization of active human alu subfamilies based on the 1000 genomes pilot project. *Genome Biol. Evol.*, 7, 2608–2622.

8. Ahmed, M., Li, W. and Liang, P. (2013) Identification of three new alu subfamilies by source tracking of recently integrated alu yb elements. *Mob. DNA*, 4, 25.

9. Cordaux, R. and Batzer, M.A. (2009) The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.*, 10, 691–703.

10. Stoneking, M., Fontius, J.J., Clifford, S.L., Soodyall, H., Arcto, S.S., Saha, N., Jenkins, F., Tahir, M.A., Deininger, P.L. and Batzer, M.A. (1997) Alu insertion polymorphisms and human evolution: evidence for a larger population size in Africa. *Genome Res.*, 7, 1061–1071.

11. Witherspoon, D.J., Marcheni, E., Watkins, W.S., Ostler, C.T., Wooding, S.P., Anders, B.A., Fowlkes, J.D., Boissinot, S., Furano, A.V., Ray, D.A. et al. (2006) Human population genetic structure and diversity inferred from polymorphic L1/LINE-1 and alu insertions. *Hum. Hered.*, 62, 30–46.

12. Tattini, L., D’Aurizio, R. and Magi, A. (2015) Detection of genomic structural variants from next-generation sequencing data. *Front. Biophys. Biotechnol.*, 3, 92.

13. Mahmoud, M., Obet, N., Cruz-Díazolos, D., Mounier, N., Dessimoi, C. and Sedlazeck, F.J. (2019) Structural variant calling: the long of it and the short of it. *Genome Biol.*, 20, 246.

14. Auton, A., Abecasis, G.R., Altshuler, D.M., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E. et al. (2015) A global reference for human genetic variation. *Nature*, 526, 68–74.

15. Chaisson, M.J.P., Sanders, A.D., Zhao, X., Malhotra, A., Porubský, D., Rausch, T., Gardner, E.J., Rodriguez, O.L., Guo, L., Collins, R.L. et al. (2019) Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.*, 10, 1784.
38. Chen,L.-L. and Yang,L. (2017) ALU ternative regulation for gene expression. Trends Cell Biol., 27, 480–490.

39. Yang,Z., Zhou,D., Li,H., Cai,X., Liu,W., Chang,H., Li,M. and Xiao,X. (2020) The genome-wide risk alleles for psychiatric disorders at 3p21.1 show convergent effects on mRNA expression, cognitive function, and mushroom dendritic spine. Mol. Psychiatry, 25, 48–66.

40. Saeliw,T., Tangsuwansri,C., Thongkorn,S., Chonchaiya,W., Levanon,K., Eisenberg,E., Rechavi,G. and Levanon,E.Y. (2005) Comparative genomics and comparative transcriptome profiling analyses reveal novel epigenetic regulatory networks associated with autism spectrum disorder. Mol. Autism, 9, 27.

41. Xiang,S., Liu,Z., Zhang,B., Zhou,J., Zhu,B.-D., Ji,J. and Deng,d,D., Nishimura,D.Y., Swiderski,R.E., Searby,C.C., Berg,E.M., Saeliw,T., Tangsuwansri,C., Thongkorn,S., Chonchaiya,W., Levanon,K., Eisenberg,E., Rechavi,G. and Levanon,E.Y. (2005) Comparative genomics and comparative transcriptome profiling analyses reveal novel epigenetic regulatory networks associated with autism spectrum disorder. Mol. Autism, 9, 27.

42. Payer,L.M., Steranka,J.P., Ardeljan,D., Walker,J., Fitzgerald,K.C., Calabresi,P.A., Cooper,T.A. and Burns,K.H. (2019) Alu insertion variants alter mRNA splicing. Nucleic Acids Res., 47, 421–431.

43. Levanon,K., Eisenberg,E., Rechavi,G. and Levanon,E.Y. (2005) Letter from the editor: adenosine-to-inosine RNA editing in alu repeats in the human genome. EMBO Rep., 6, 831–835.

44. Nishimura,D.Y., Swiderski,R.E., Searby,C.C., Berg,E.M., Ferguson,A.L., Hennekam,R., Merin,S., Weleber,R.G., Biesecker,L.G., Stone,E.M. et al. (2005) Comparative genomics and gene expression analysis identifies BBS9, a new bardet-biedl syndrome gene. Genet. Med., 831–835.

45. Fan,H.-H., Zheng,J., Huang,S.-S., Guo,Q., Liang,Y.-Z., Sun,Y., Zhu,J.-H. and Zhang,X. (2020) A Novel Antisense Alu Insertion/Deletion Polymorphism of ALDH1A1 Modulates Risk of Parkinson’s Disease Social Science Research Network. Rochester, NY.

46. Jahic,A., Erichsen,A.K., Deufel,T., Tallaksen,C.M. and Beetz,C. (2016) A polymorphic alu insertion that mediates distinct disease-associated deletions. Eur. J. Hum. Genet., 24, 1371–1374.