Gene 675 (2018) 69–79

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Research paper

Evidence for positive selection on recent human transposable element insertions

Lavanya Rishishwar[a,b,c], Lu Wang[a,b], Jianrong Wang[d], Soojin V. Yi[a], Joseph Lachance[a], I. King Jordan[a,b,c,e,*]

[a] School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA 30332, USA
[b] PanAmerican Bioinformatics Institute, Cali, Valle del Cauca 760043, Colombia
[c] Applied Bioinformatics Laboratory, Atlanta, GA 30332, USA
[d] Department of Computational Mathematics, Science and Engineering, Michigan State University, East Lansing, MI 48824, USA
[e] BIOS Centro de Bioinformática y Biología Computacional, Manizales, Caldas 170001, Colombia

ARTICLE INFO

Keywords:
Natural selection
Evolution
Population genetics
Human genome
Population branch statistic
1000 genomes project

ABSTRACT

Insertional activity of transposable elements (TEs) has had a major impact on the human genome; approximately one-half to two-thirds of the genome sequence is likely to be derived from TE insertions. Several families of human TEs – primarily Alu, L1 and SVA – continue to actively transpose, thereby generating insertional polymorphisms among individual genomes. The impact that TE insertions have on their human hosts' fitness, and accordingly the role that natural selection plays in shaping patterns of TE polymorphisms among populations, have yet to be systematically evaluated using whole genome sequence data. We present here a population genomic study of the effects of natural selection on human genetic variation that results from the recent activity of TEs. We developed a genome-wide scan for selection on human TE polymorphisms and applied it to a dataset of 14,384 locus-specific TE insertions characterized for 1511 individuals from 15 populations. Our TE selection scan looks for anomalously high population-specific TE insertion allele frequencies that are consistent with the action of positive (adaptive) selection. To control for the effects of demographic history, we compared the observed patterns of population-specific TE insertion allele frequencies to a neutral evolutionary model generated using time forward simulation of TE insertion allele frequencies among human population groups. This approach uncovered seven cases of polymorphic TE insertions that appear to have increased in frequency within specific human populations owing to the effects of positive selection. Five of the seven putatively selected TE insertions map to tissue-specific enhancers, and two cases correspond to expression quantitative trait loci that are associated with inter-individual gene regulatory differences. This study represents the first report of recent, local adaptation acting on polymorphic human TEs.

1. Introduction

One of the major discoveries from the Human Genome Project was the extent to which the genome sequence was found to be derived from transposable element (TE) insertions (Lander et al., 2001); current estimates of the fraction of TE-derived sequences in the human genome are as high as 69% (de Koning et al., 2011). While the vast majority of these TE-derived sequences are inert remnants of ancient insertion events, several families of human TEs remain active. Ongoing TE insertional activity generates a large amount of human genetic variation in the form of population-specific structural variations. Until this time, it has not been possible to evaluate the genome-wide effects of natural selection on structural variations caused by human TE activity. We performed a comparative evolutionary analysis using a recently generated catalog of TE generated structural variations, for thousands of individuals across scores of human populations, to develop and apply a genome-wide test of natural selection on polymorphic human TE insertions.

The ubiquity and abundance of TE-derived sequences in eukaryotic genomes, such as our own, begs an explanation. For years, the selfish DNA theory was held as the gold-standard explanation for the genomic presence of TEs (Doolittle and Sapienza, 1980; Orgel and Crick, 1980).

Abbreviations: TE, Transposable Element; PolyTE, Polymorphic Transposable Element; LINE, Long Interspersed Element; SINE, Short Interspersed Element; LTR, Long Terminal Repeat; 1KGP, 1000 Genomes Project; PBS, Population Branch Statistic; LD, Linkage Disequilibrium; ChIP, Chromatin Immunoprecipitation; eQTL, Expression Quantitative Trait Loci

* Corresponding author at: 950 Atlantic Drive, Atlanta, GA 30332-0230, USA.
E-mail address: king.jordan@biology.gatech.edu (I.K. Jordan).

https://doi.org/10.1016/j.gene.2018.06.077
Received 20 June 2018; Accepted 24 June 2018
Available online 25 June 2018
0378-1119/

The selfish DNA theory posits that TEs are genomic parasites that provide no benefit for their hosts and exist simply by virtue of their ability to out-replicate the genomes in which they reside. This idea is based on the fact that since TEs replicate when they transpose, and are also inherited vertically across generations, they have a biased transmission rate compared to host genes that rely exclusively on vertical transmission for their propagation. It was even shown that TEs' replicative advantage meant that they could, in theory, persist and spread in the face of a selective cost to their host genome (Hickey, 1982).

The selfish DNA theory for TEs is closely linked to the notion that TE sequences should be either neutral genetic elements or subject to purifying selection. Given the fact that human TE activity entails the insertion of rather large pieces of DNA, ranging from several hundred to almost ten-thousand base pairs, it is entirely reasonable to expect TE insertions to often be deleterious for their hosts. There is in fact abundant evidence from studies of disease that human TE insertions can be highly deleterious, consistent with the expectations of the selfish DNA theory, since human TE insertions have been linked to a number of diseases including rare Mendelian diseases as well as more common chronic diseases such as cancer (Beck et al., 2011; Burns and Boeke, 2012; Solyom et al., 2012; Reilly et al., 2013; Chenais, 2015; Hancks and Kazazian, 2016; Payer et al., 2017; Wang et al., 2017a).

Nevertheless, in the years since the publication of the draft human genome sequence, there have been many other studies that have demonstrated how formerly selfish human TE sequences have been exapted (Bowen and Jordan, 2007), or domesticated (Miller et al., 1992), to play a functional role for their hosts. For the most part, these studies have uncovered a role for TE-derived sequences in the regulation of human genes (Feschotte, 2008). TE-derived sequences have been shown to contribute a wide variety of regulatory sequences, including promoters (Jordan et al., 2003; Marino-Ramirez et al., 2005; Conley et al., 2008), enhancers (Bejerano et al., 2006; Kunarso et al., 2010; Chuong et al., 2013; Notwell et al., 2015; Chuong et al., 2016), transcription terminators (Conley and Jordan, 2012) and several classes of small RNAs (Weber, 2006; Piriyapongsa et al., 2007; Kapusta et al., 2013). Human TEs also influence various aspects of chromatin structure throughout the genome (Lander et al., 2001; Pavlicek et al., 2001; Schmidt et al., 2012; Jacques et al., 2013; Sundaram et al., 2014; Roadmap Epigenomics et al., 2015).

It is important to note that all of the aforementioned studies on TE-derived regulatory sequences have dealt exclusively with relatively ancient TE insertions that are fixed among human populations. In other words, known examples of specific human TE-derived regulatory sequences will be found at the same genomic locations in any individual person. While such fixed TE-derived regulatory sequences are certainly functionally relevant, by definition they will not be a source of genetic regulatory variation between individuals; although, they may contribute to species-specific gene regulation. The fact that TE-derived regulatory sequences correspond to relatively ancient fixed TEs is not at all surprisingly when you consider that the vast majority of human TE sequences (~99.2%) correspond to ancient TE families that are no longer capable of transposition (Rishishwar et al., 2017). However, very recent developments in genomics and bioinformatics are just beginning to enable systematic, genome-scale surveys of human polymorphic TEs (polyTEs) with insertion site locations that vary among individuals. The 1000 Genomes Project (1KGP) in particular has resulted in a collection of 16,192 polyTE genotypes characterized for 2504 individuals from 26 global populations (Genomes Project et al., 2015; Sudmant et al., 2015). Analysis of this data set has the potential to yield novel insights regarding the role of natural selection in shaping human TE genetic variation.

There is evidence of adaptive evolution of polyTEs in *Drosophila* (Gonzalez et al., 2008; Gonzalez et al., 2009; Gonzalez et al., 2010) along with studies that show the regulatory potential of polyTEs in mice (Rebollo et al., 2011). However, at this time there is only tentative evidence to suggest that human polyTEs have been subject to positive (adaptive) selection (Kuhn et al., 2014). Here, we utilized the recently released 1KGP polyTE data in order to evaluate the role that natural selection has played in shaping this understudied, but potentially impactful, source of human genetic variation. In particular, we were interested to measure the effect of natural selection on human TE genetic variation along with the potential connection between polyTE selection and genome regulation. To do so, we performed genome-wide comparative analyses on the polyTE insertion allele frequencies within and between major human population groups (Supplementary Fig. 1). This allowed us to evaluate the effect of both negative (purifying) and positive (adaptive) selection on polyTE genetic variation. We developed a novel approach utilizing the population branch statistic (PBS) test to detect positive selection on polyTE insertions. This approach incorporates recently established population genetic parameters with extensive evolutionary simulation of polyTE allele frequencies, in order to detect cases of polyTE insertions that have been swept to high allele frequencies in specific human populations. Our analysis supports the pervasive action of negative selection on human TE polymorphisms. Moreover, we were able to demonstrate, for the first time, signatures of population-specific positive selection on polymorphic TE insertions in the human genome.

2. Results and discussion

2.1. Characterization of human polymorphic transposable elements (polyTEs)

There are three main families of active human TEs that generate insertion polymorphisms among individual human genomes (Ray and Batzer, 2011): L1 (Kazazian et al., 1988; Brouha et al., 2003), Alu (Batzer and Deininger, 1991; Batzer et al., 1991) and SVA (Ostertag et al., 2003; Wang et al., 2005). L1 stands for Long Interspersed Element-1, or LINE1, and L1s are ~6 kb long, autonomous, non-LTR (long terminal repeat) retrotransposons (Burton et al., 1986; Fanning and Singer, 1987; Moran et al., 1996). Alu and SVA are non-autonomous, non-LTR retrotransposons that are retrotransposed in *trans* via the L1 transposition machinery (Dewannieux et al., 2003; Salem et al., 2003). Alus are short interspersed elements (SINEs) that are ~300 bp long (Schmid and Deininger, 1975; Ullu and Tschudi, 1984), whereas SVA are composite elements made up of SINE, VNTR (Variable number tandem repeat) (Ono et al., 1987; Shen et al., 1994) and Alu elements and can vary from ~100–1600 bp in length (Sudmant et al., 2015).

The Phase 3 release of the 1000 Genomes Project (1KGP) includes polymorphic transposable (polyTE) genotype calls for these three active TE families from 2504 individuals sampled across 26 populations world-wide (Genomes Project et al., 2015; Sudmant et al., 2015). The 26 populations from the 1KGP can be organized into five major regional population groups. The African, Asian, and European regional population groups consist of (relatively) non-admixed individuals, and polyTE genotypes from these groups were analyzed here for the purpose of measuring selection on polyTEs (Supplementary Fig. 2). We analyzed a total of 14,384 polyTE genotypes from 1511 individuals across 15 individual populations from these three regional population groups (Supplementary Table 1). PolyTE genotype calls from the three most actively transposing families of TEs were evaluated: Alu (11,216 or 78.0%), L1 (2421 or 16.8%) and SVA (747 or 5.2%). PolyTE genotype calls were used to calculate insertion allele frequencies within and between populations in order to measure the effects of natural selection on human genetic variation caused by recent TE activity (see Materials and Methods).

2.2. Negative selection on human polyTEs

Consistent with the results of our previous study on human polyTEs (Rishishwar et al., 2015), we found several lines of evidence in support of the action of negative (purifying) selection on recent human TE

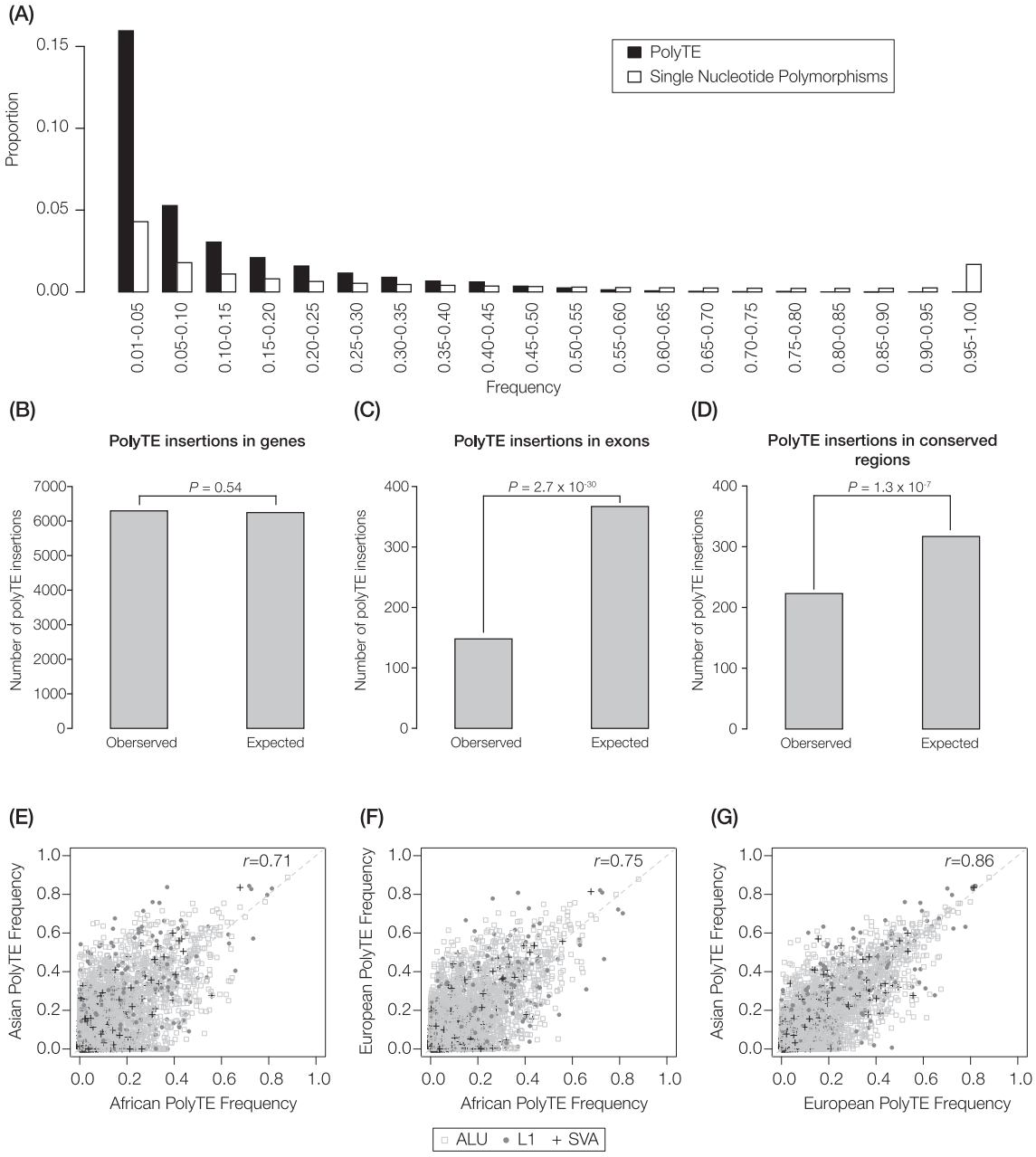

Fig. 1. Signatures of purifying selection on polyTE insertions. (A) Unfolded allele frequency spectrum for polyTE insertions (black bars) and SNPs (white bars). Derived allele frequencies for both polyTEs and biallelic intergenic SNPs are as reported by the 1KGP. Observed versus expected counts of polyTE insertions in genes (B), exons (C) and conserved regions (D). The significance of the differences between observed versus expected TE counts (Fisher's exact test *P*-values) are shown for each plot. (E–G) Correlations of polyTE insertion allele frequencies between regional population groups are shown for shared Alu, L1 and SVA insertions; Spearman correlation coefficients are shown as *r*-values.

insertions based on their allele frequency distributions and insertion site patterns. First, the majority of polyTE insertions show low derived allele frequencies; 11,658 (81.0%) polyTE loci exhibit average allele frequencies of < 5% across all three regional population groups, and 10,119 (70.3%) exhibit allele frequencies < 5% within each of the regional groups. Accordingly, polyTE insertions show a highly left-skewed allele frequency distribution with relatively fewer high frequency alleles than can be seen for biallelic intergenic single nucleotide polymorphisms (SNPs) (Fig. 1A). When the Alu, L1 and SVA polyTE families are considered separately, they all show similarly left skewed allele frequency distributions (Supplementary Fig. 3). This skew is consistent with purifying selection acting to keep polyTE insertions at low frequencies and/or neutrality of polyTE insertions. Second, polyTE insertions were less abundant than expected in functionally important regions such as exons and evolutionary conserved regions (Fig. 1B–D) as previously observed for fixed TE sequences (Cooper et al., 2005; Davydov et al., 2010). In order to ensure that this observation cannot be attributed to insertion bias, we considered the relative polyTE insertion frequencies in bins of very low (private) compared to mid-frequency insertions. Mid-frequency insertions show a greater difference between the observed and expected numbers of insertions in all three classes of functional regions, which is consistent with negative selection as opposed to polyTE insertion bias among the different regions (Supplementary Fig. 4). Third, the allele frequencies of polyTE insertions that are shared among regional population groups are both skewed towards low frequencies and highly correlated between groups (Fig. 1E–F and

Supplementary Fig. 5). These results can be also explained by the action of purifying selection maintaining polyTE insertion allele frequencies low in all populations. All of these observations are consistent with the known deleterious effects of human TE insertions (Beck et al., 2011; Burns and Boeke, 2012; Solyom et al., 2012; Reilly et al., 2013; Chenais, 2015; Hancks and Kazazian, 2016). Importantly, these results also provide support for the reliability of the polyTE genotype calls used to generate the allele frequencies analyzed here. It should be noted that since the ancestral state for polyTEs is the absence of an insertion (Rishishwar et al., 2015), polyTE insertion allele frequencies measured here correspond to derived alleles.

We further evaluated the evidence for negative selection on polyTE insertions by comparing the allele frequencies of full-length L1 polyTE insertions, which are expected to be most deleterious, compared to the smallest L1 polyTE insertions. Full-length L1 polymorphic insertions have a significantly lower mean allele frequency than the smaller L1 insertions, which can be taken as an additional line of evidence in support of the action of negative selection on polyTE insertions (Supplementary Fig. 6).

2.3. Positive selection on human polyTEs

Evaluation of polyTE allele frequency spectra for each of the three regional population groups separately suggested the possibility that some human polyTE insertions may have increased in frequency owing to the effects of positive selection. As mentioned above, the population group-specific polyTE allele frequency spectra are highly skewed to the low end of the distribution; however, there are a number of polyTE insertions, particularly in Asian and European populations, that appear at higher than expected frequencies (Fig. 2 and Supplementary Fig. 7). At the high end of the polyTE allele frequency spectrum, there is a shift whereby Asian and European polyTEs become relatively more frequent than African TEs.

The observed shift at the high end of the polyTE allele frequency spectrum could be due to positive selection or it could be due to genetic drift coupled with founder effects after population bottle necks, i.e. demographic history. We developed and applied a modified version of the population branch statistic (PBS) test (Jordan et al., 2001; Shriver et al., 2004; Yi et al., 2010) to distinguish between neutral evolution of polyTE insertions (i.e. genetic drift) versus population group-specific increases in polyTE allele frequencies that can be attributed to positive selection. We chose this method given its demonstrated power to detect recent positive selection of SNPs in human populations (Yi et al., 2010). The PBS test measures population-specific divergence levels by converting pairwise F_{ST} values into population-specific branch lengths, and we adopted this method by computing the F_{ST} values from polyTE allele frequencies as described in the Materials and Methods. Deviations from neutrality are detected as sites that have undergone accelerated evolution as revealed by extreme population-specific branch length values.

Fig. 3A shows the genome-wide polyTE PBS tree with average branch lengths for each regional population group. On average, shared polyTE insertions show higher branch lengths in Africa, compared to Asia and Europe, consistent with African populations being ancestral to the more recently diverged Asian and European populations. PBS branch length distributions for each regional population group are highly skewed; the vast majority of polyTEs have low PBS values for all three populations (Fig. 3B), consistent with strong purifying selection.

We further evaluated all of the PBS trees in an effort to look for rare cases of positively selected polyTEs. To do this, the observed population group-specific branch lengths from the polyTE PBS trees were compared to a null distribution of branch lengths generated via human population genetic model simulations as described in the Materials and Methods (Fig. 4). This population genetic model allowed us to control for the effects of demographic history, i.e. to ensure that any signal of positive selection that we detect could not be explained by genetic drift. For the first iteration of this simulation, the population size parameter values for the population genetic model that we used were previously estimated for a large set of neutrally evolving loci from whole genome sequences using rigorous Bayesian inference (Gronau et al., 2011); as such, they provide a realistic null model for human population divergence (Fig. 4A). The set of simulated PBS branch lengths was compared to the observed set in order to look for statistically significant outliers that represent putative positively selected polyTEs (Supplementary Fig. 8).

The population genetic simulation generates simulated ancestral and extant polyTE allele frequencies that are highly correlated (Fig. 4B–C), as can be expected for the null model and consistent with observed between population correlations seen for observed polyTE frequencies (Fig. 1E–G and Supplementary Fig. 5). In addition, we observe that simulations starting from the same ancestral polyTE insertion allele frequencies can generate moderately large extant allele frequency differences (Fig. 4D). Both of these results underscore the conservative nature of the evolutionary simulation approach we used to generate a null distribution of PBS branch lengths under a scenario of genetic drift.

Comparison of the observed versus simulated sets of PBS branch lengths yielded a set of 163 polyTE insertions (1.13% of the full set) that appear to have increased in frequency in the Asian or European population groups based on positive selection (Supplementary Table 2). Although there were some polyTEs with relatively high PBS branch lengths for Africa, none were significant given the null model. Among the 163 putatively selected polyTE insertions, only 79 have frequencies > 10% and only 14 have frequencies > 25% in either of the Asian or European populations. We did not find any significant enrichment or depletion of any particular TE family type or specific genomic region, indicating that the mechanistic basis of positive selection on TE insertion is highly diverse.

We repeated the same population genetic simulation a second and third time, changing the population size parameter values by ± 50% to yield wide population size ranges that would be expected to maximize the effects of drift on our data. The results of the three simulations are

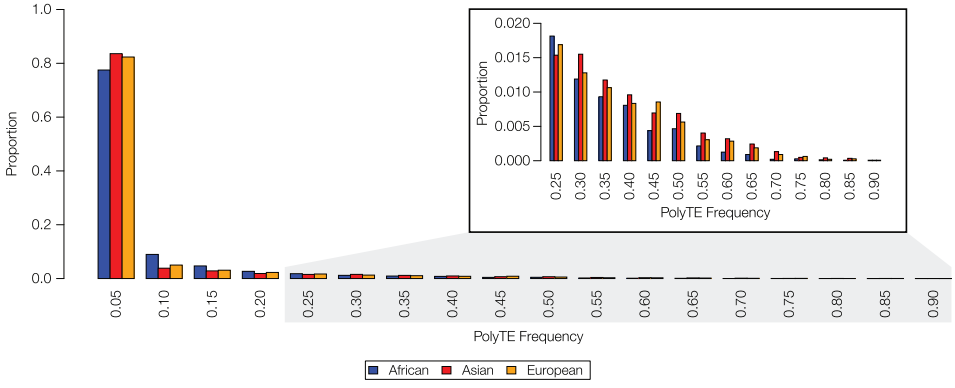

Fig. 2. Unfolded allele frequency spectrum for polyTE insertions from African (blue), Asian (red) and European (gold) population groups. The inset expands the higher range of the allele frequency spectrum (≥0.25). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

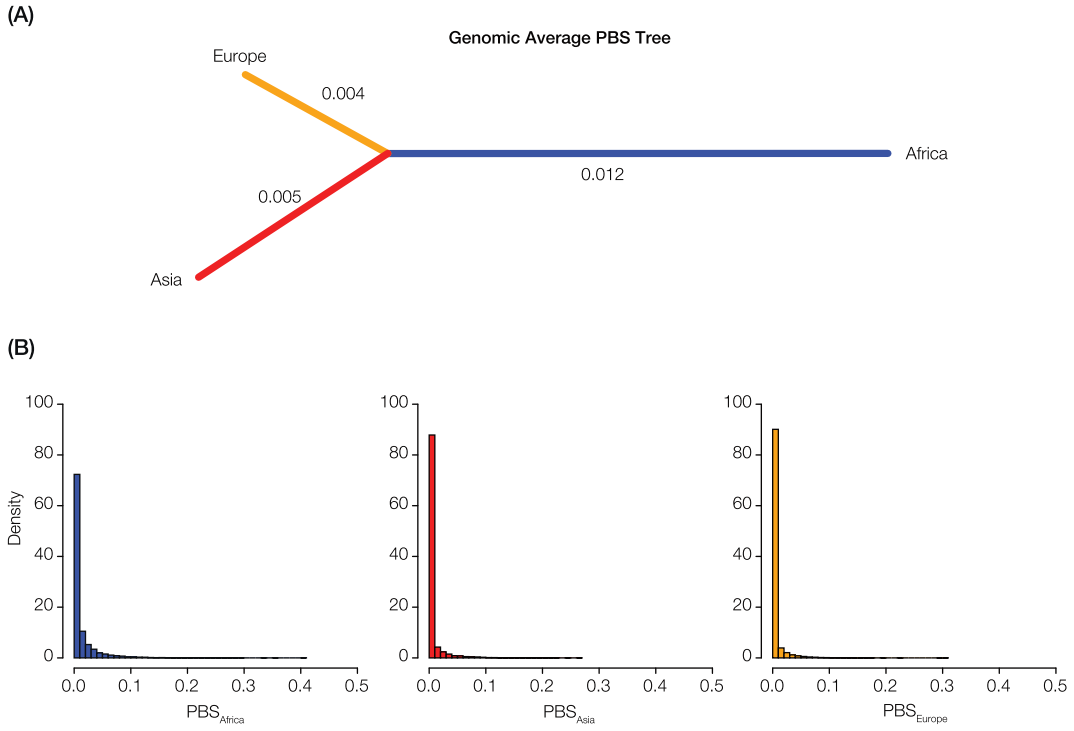

Fig. 3. Overview of the population branch statistic (PBS) test metric used to detect positive selection on polyTE insertions. (A) Tree constructed with branch lengths from the genomic averages of regional group-specific PBS values. (B) Histograms showing the PBS branch length distributions for the African (blue), Asian (red) and European (gold) population groups. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

highly consistent and support the notion that the extreme, population-specific polyTE allele frequencies uncovered here can be attributed to positive selection, as opposed to drift alone (Supplementary Fig. 9).

To better understand the mechanism of natural selection on polyTEs, we narrowed our focus to a more limited set of insertions for which multiple lines of evidence support the action of positive selection as well as some potential functional (regulatory) significance. To do so, we searched for putatively selected polyTE loci that were found at anomalously high frequencies within a single population group and were also co-located within genes and/or functionally important genomic regulatory elements. A list of seven positively selected polyTEs that fit these criteria is shown in Table 1, and a number of examples from this table are described further in the next section. We searched for orthogonal sources of validation for these to seven examples of positively selected TEs in an effort to provide additional support for their presence. We find multiple sources of independent evidence that support the presence and genomic locations for six out of seven of these selected polyTEs (Supplementary Table 3).

We performed an analysis of the patterns of linkage disequilibrium (LD) and nucleotide diversity in the genomic regions surrounding the top seven high confidence positively selected TEs to search for additional evidence for selection on polyTE linked SNPs. We found evidence for positive selection on SNPs linked to the selected polyTEs, based on literature searches and the distributions/values of three different positive selection test statistics: (1) the difference of derived allele frequency ($|DDAF| > 0.2$), (2) the integrated haplotype score ($|iHS| > 1.5$), and (3) the cross-population composite likelihood ratio ($XPCLR > 5$). The approach that we used for this analysis and our results are reported in the revised Supplementary material section entitled “Effects of selection on genomic regions with positively selected polyTEs” and Supplementary Table 4.

2.4. Examples of positively selected human polyTEs

One of the most promising candidates for positive selection is a polymorphic L1 insertion located on the short arm of chr1 at position 75,192,907 (Fig. 5A). This L1 is inserted within the second intron of the crystalline zeta gene (*CRYZ*, also known as Quinone Reductase or *QR*) and co-located with a liver enhancer element (Roadmap Epigenomics et al., 2015). This polymorphic L1 insertion is found in all 26 populations sampled as part of the 1KGP; however, it is found at low frequencies in the African (5%) and Asian (1%) population groups. There was a striking increase in the allele frequency of this insertion along the European lineage, and it is currently found at an average allele of 47% in European populations (Fig. 5B). When these polyTE allele frequencies are used to calculate the F_{ST} values that underlie the PBS test, the European-specific branch on the PBS tree is extremely long compared with the African and Asian branches (Fig. 5C). Comparison of this observed polymorphic L1 PBS tree to the set of simulated tress, with similar average branch lengths, yields an FDR q-value of 0.019 (Table 1). Consistent with the potential regulatory effects of this insertion, expression quantitative trait loci (eQTL) analysis shows that the presence of this specific L1 in European individuals is significantly associated with lower expression of the *CRYZ* gene in B-lymphoblastoid cell lines (Fig. 5D); although, it is formally possible that the eQTL result is due to a linked variant that reduces *CRYZ* expression.

Another strong candidate for positive selection is a polymorphic Alu insertion at chr16 position 75,655,176 (Fig. 6A). This Alu insertion is found within the second intron of the Adenosine Deaminase, tRNA Specific 1 gene (*ADAT1*) and is co-located with enhancer elements predicted to have activity in numerous cell lines analyzed by The Roadmap Epigenomics project (Roadmap Epigenomics et al., 2015). This polymorphic Alu insertion is also found in all human populations surveyed by the 1KGP. It is seen at low frequencies in African (4%) and European (5%) population groups and far higher frequency in the Asian population group (44%) (Fig. 6B). Accordingly, the Asian-specific branch on the PBS tree is far longer than the African or European branches (Fig. 6C), and comparison with simulated trees yields an FDR q-value of 0.04. This is a clear case of a marked increase in polyTE allele frequency that cannot be readily explained by genetic drift. In addition,

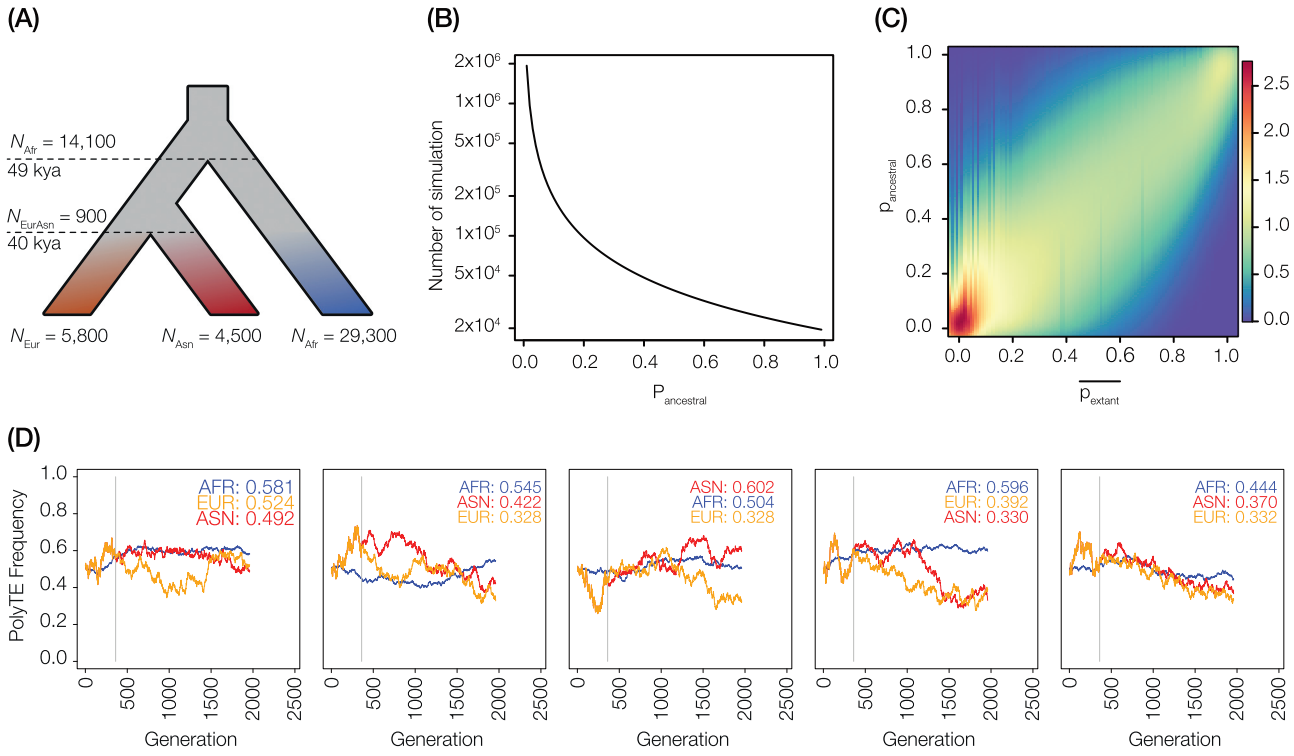

Fig. 4. Evolutionary modelling of polyTE insertion allele frequencies. Evolutionary modelling was used to generate a null distribution of PBS values for the purpose of detecting positive selection on polyTE insertions. (A) Scheme of the population genetic model and parameters used to simulate polyTE insertion allele frequencies. The population genetic model consists of the tree shown, the effective population sizes (N) at each node of the tree and the number of thousands of years ago (kya) that correspond to the two population splits in the tree: (i) out of Africa and (ii) Europe-Asia. (B) The number of model simulations run (y-axis) is plotted for each initial ancestral polyTE insertion frequency ($P_{ancestral}$ on the x-axis), ranging from 0.01 to 0.99. (C) Density scatter plot comparing the population genetic model's ancestral polyTE insertion frequencies ($P_{ancestral}$ on the y-axis) to the mean of the extant polyTE insertion frequencies (P_{extant} on the x-axis) for the three simulated population groups. (D) Five examples of model simulations run with initial polyTE frequencies of 0.5. The plots show the polyTE insertion frequency dynamics across generations for each evolutionary model run. The final (extant) polyTE frequency values of each evolutionary model run are shown for each population group: African (AFR-blue), Asia (ASN-red) and European (EUR-gold). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the location of the insertion is suggestive of regulatory function; although, a lack of gene expression data from matched Asian samples does not allow us to directly assess the association of the insertion with changes in *ADAT1* expression.

The candidate with the strongest PBS-based evidence of positive selection is a polymorphic Alu insertion at chr4 position 43,399,986 (Supplementary Fig. 10). This Alu element is inserted in an intergenic region and does not overlap with any known functional (regulatory) elements. Nevertheless, its relative allele frequencies leave little doubt as to the role for positive selection in shaping its population-specific patterns of variation. This polymorphic Alu insertion is found in all of the regional population groups with an allele frequency of at least 5% in

Table 1
List of high confidence positively selected polyTEs.

Chr[a]	Position[b]	Family[c]	p_{AFR}[d]	p_{ASN}[d]	p_{EUR}[d]	PBS[e]	q-Value[f]	Cont[g]	Gene[h]	Enh[i]	TFBS[j]	eQTL[k]
1	75,192,907	L1	0.05	0.01	0.47	0.29	0.019	EUR	CRYZ	Yes		CRYZ
1	169,442,974	ALU	0.03	0.35	0.02	0.19	0.046	ASN	SLC19A2	Yes		
4	43,399,986	ALU	0.08	0.16	0.61	0.31	0.033	EUR				
11	10,042,452	L1	0.01	0.16	0.01	0.08	0.039	ASN	SBF2	Yes		
14	88,415,499	L1	0.02	0.10	0.00	0.05	0.033	ASN	GALC	Yes		GPR65
16	75,655,176	ALU	0.05	0.44	0.04	0.24	0.040	ASN	ADAT1	Yes	Yes	
17	44,153,977	SVA	0.00	0.00	0.21	0.13	0.031	EUR	KANSL1			KANSL1

[a] Chromosome.
[b] Base position in the hg19 human genome reference assembly.
[c] PolyTE family.
[d] PolyTE insertion frequency in the African, Asian and European population groups.
[e] PBS branch length value for the polyTE from its selected regional group.
[f] FDR corrected q-value for PBS selection test based on the population genetic simulation (Fig. 4).
[g] Regional population group in which the polyTE is selected.
[h] Gene name in which the selected polyTE insertion is located.
[i] Selected polyTE insertion located in an enhancer.
[j] Selected polyTE insertion located in a transcription factor binding site (TFBS).
[k] Target gene name for which the selected polyTE insertion is an eQTL.

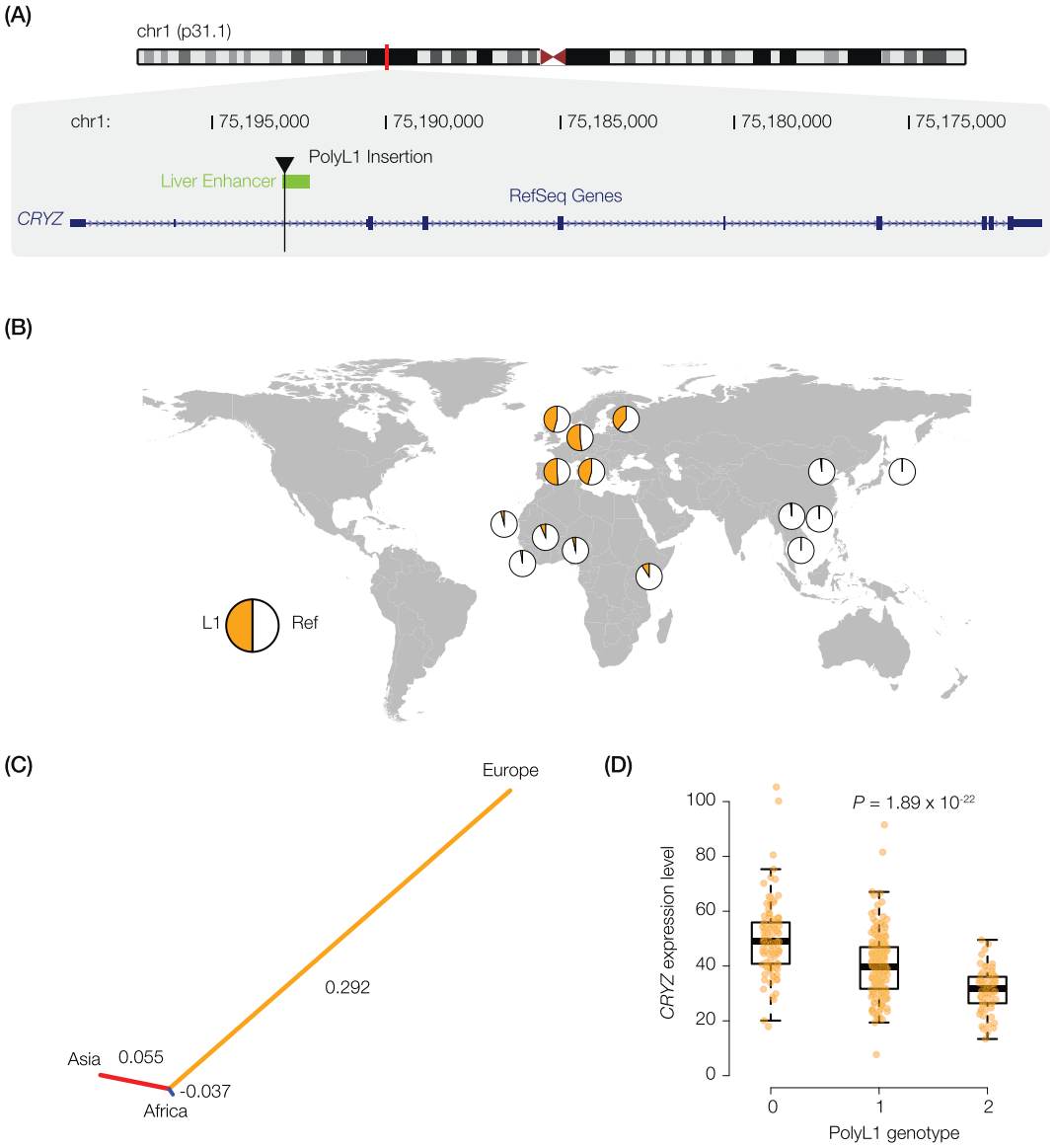

Fig. 5. Positively selected polyL1 insertion in the *CRYZ* gene. (A) Chromosome 1 ideogram showing the location (red bar) of the *CRYZ* gene on the short arm of chromosome 1 along with a *CRYZ* gene model showing the location of the polyL1 insertion and its co-located liver enhancer element (green bar). (B) Frequencies of the European selected polyL1 insertion (gold in the pie charts) for the individual populations studied here from Africa, Asia and Europe. (C) Tree with branch lengths scaled to the population group-specific PBS values (shown for each branch). (D) *CRYZ* expression level distributions are shown for European individuals that have 0, 1 or 2 copies of the selected polyL1 insertion. The significance of the differences in expression among individuals for the three different polyL1 insertion genotypes between is shown (linear additive model *P*-value). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

each of the 26 1KGP populations. It shows the highest population-specific frequency for any of the putatively selected insertions, with 61% average frequency in the European populations compared to 8% and 16% in the African and Asian population groups, respectively. Accordingly, it also has the highest PBS test statistic value for any of the high confidence positively selected polyTEs shown in Table 1.

In addition to the three examples described above, putatively selected polyTE insertions in four other genes – *SLC19A2*, *SBF2*, *GALC* and *KANSL1* – also showed strong composite signals of positive selection. Putatively selected polyTE insertions in the first three genes (*SLC19A2*, *SBF2* and *GALC*) overlap with enhancer elements while insertions in the last two genes (*GALC* and *KANSL1*) were found to behave as eQTLs to *GPR65* and *KANSL1* genes, respectively (Table 1).

2.5. Conclusions

We analyzed the population genetic variation caused by human TE activity in an effort to understand how natural selection acts on TE polymorphisms. This study represents the first genome-wide scan for selection on human polymorphic TE sequences. The majority of human polyTE insertions are found at low allele frequencies, within and between populations, and appear to evolve via negative (purifying) selection, with others increasing to moderate allele frequencies via genetic drift or positive selection. A small, but not insubstantial, minority of polyTE insertions show patterns of allele frequencies that are consistent with, albeit not definitive proof of, population-specific positive selection. If these polyTE insertions have in fact been subject to positive selection, this suggests that they play some functional role for their host genomes. We have recently reported evidence for polyTE effects on human gene regulation (Wang et al., 2017a; Wang et al., 2017b), which

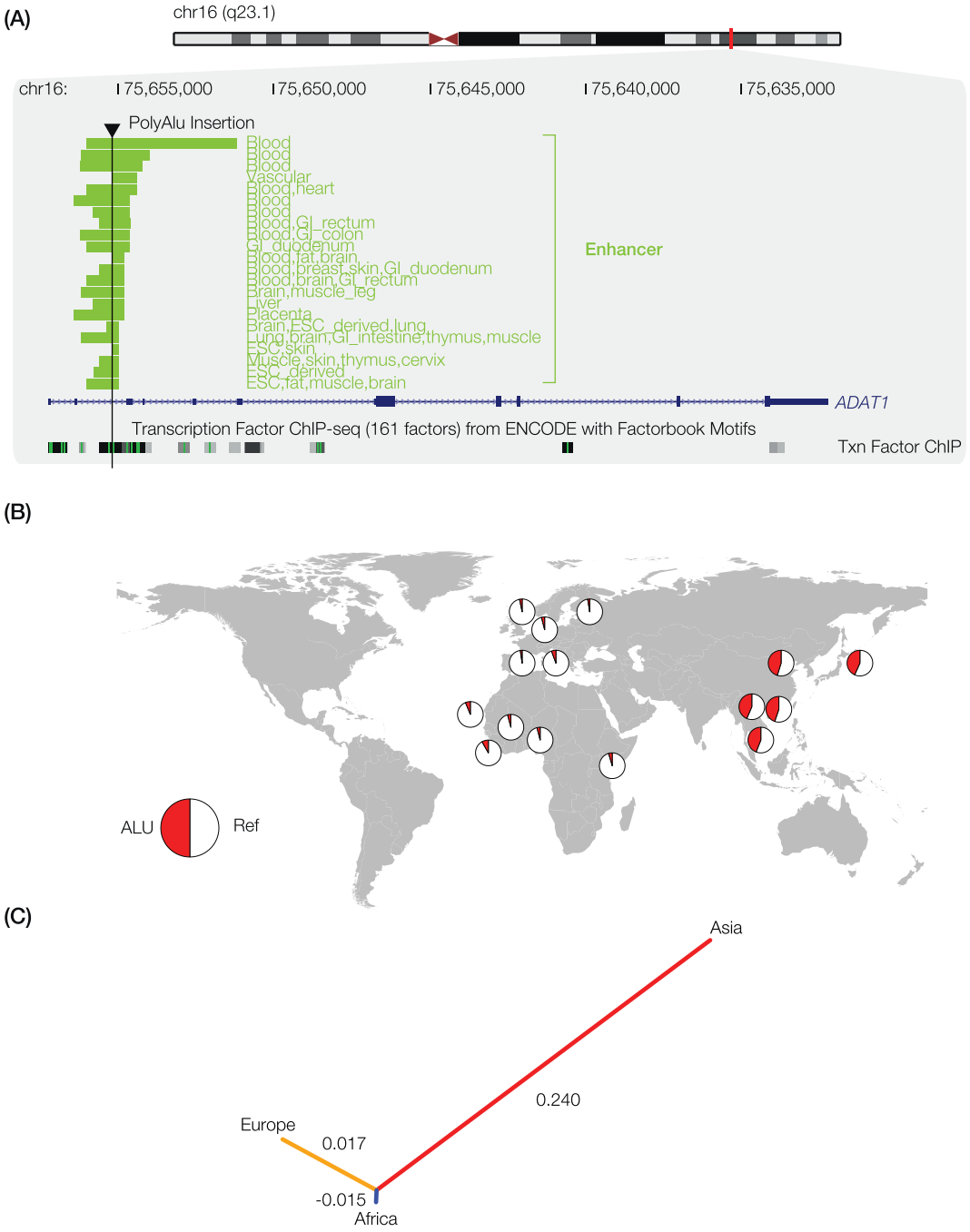

Fig. 6. Positively selected polyAlu insertion in the ADAT1 gene. (A) Chromosome 16 ideogram showing the location (red bar) of the ADAT1 gene on the long arm of chromosome 16. The location of the polyAlu insertion in the ADAT1 gene model is shown along with co-located enhancer elements, from a number of different tissues, and transcription factor binding sites. (B) Frequencies of the Asian selected polyAlu insertion (red in the pie charts) for the individual populations studied here from Africa, Asia and Europe. (C) Tree with branch lengths scaled to the population group-specific PBS values (shown for each branch). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

could be one mechanism that leads to positive selection of polyTEs. Indeed, a number of the positively selected TEs uncovered here have functional features that are consistent with a role in human gene regulation. These results indicate that the exaptation of human TE sequences, which was previously limited to relatively ancient and fixed TE sequences, can also occur for more recently active polyTEs with insertion sites that vary among individuals within and between populations.

3. Materials and methods

3.1. Polymorphic transposable element (polyTE) analysis

Genotype calls for 14,384 human polymorphic transposable element (polyTE) insertions were obtained from the Phase 3 data release of the 1000 Genomes Project (1KGP), with locations corresponding to build GRCh37/hg19 of the human genome reference sequence (Genomes Project et al., 2015; Sudmant et al., 2015). The insertion site locations of polyTEs in the 1KGP sample donors' genomes, along with their presence/absence genotypes, were characterized from next-

generation sequence data by the 1KGP Structural Variation Group using the computational tool MELT. The program MELT works by screening for discordant read mappings for short paired-end reads and split read mapping for longer reads. MELT's performance was previously benchmarked by its developers using an experimentally validated set of polyTEs characterized for a single 1KGP individual, and the polyTE genotype calls from MELT were found to be quite reliable (Sudmant et al., 2015). In addition, our own group independently benchmarked the performance of MELT and validated the accuracy of the human polyTE genotype calls that it generates (Rishishwar et al., 2016). In our hands, MELT showed 90.4% precision and 81.5% recall and was the top performer among 21 polyTE detection programs that were evaluated. We performed an additional series of controls to ensure that any differences observed for polyTEs among the population groups analyzed here cannot be attributed to population-specific biases in polyTE insertion detection and calling (see Supplementary material pp. 17–18; Supplementary Figs. 11 & 12; Supplementary Table 3).

PolyTE genotype calls report the presence or absence of insertions for members of three families of human polyTEs: Alu, L1 and SVA. For any given polyTE insertion site, individuals can be homozygous absent (0 insertions), heterozygous (1 insertion) or homozygous present (2 insertions). PolyTE genotype calls were taken for 1511 individuals from 15 populations corresponding to the 3 non-admixed regional (super) population groups: Africa, Asia and Europe (Supplementary Table 1 and Supplementary Fig. 2). For each polyTE insertion site, its polyTE allele frequency was calculated as the total number of TE insertions observed at that site (TE_i) normalized by the total number of chromosomes in the population under consideration ($2n$): $TE_i/2n$. PolyTE insertion site allele frequencies were calculated separately for all 15 individual population groups as well as for the 3 regional population groups.

The BEDTools (Quinlan, 2014) program was used to compare the locations of polyTE insertions to (1) the genomic coordinates of RefSeq genes (O'Leary et al., 2016) (transcription start to transcription stop site for each gene), (2) the locations of RefSeq gene exons, and (3) the locations of conserved genomic regions. Conserved genomic regions were characterized using GERP++ RS conservation scores (Davydov et al., 2010) taken from the UCSC Genome Browser (Speir et al., 2016), with GERP++ RS > 3 taken to represent conserved genomic regions. The observed counts of polyTE insertions for each of these three functional features – genes, exons and conserved regions – were compared to the expected counts, which were computed as the total number of polyTE insertions multiplied by the fraction of the genome occupied by each feature. The significance of the differences in the observed versus expected counts of polyTE insertions for each feature were calculated using Fisher's exact test. All statistical analyses and correlations were performed in R.

3.2. Population branch statistic (PBS) calculation

For each polyTE insertion, its regional population group-specific allele frequencies were used to calculate African, Asian and European population branch statistic (*PBS*) values. *PBS* values were calculated using pairwise polyTE insertion frequency F-statistics (F_{ST}) based on the approach previously used for SNPs (Yi et al., 2010):

$$F_{ST} = \frac{(H_T - H_S)}{H_T} \tag{2}$$

$$T = -\log(1 - F_{ST}) \tag{3}$$

$$PBS_{Africa} = \frac{(T^{AS} + T^{AE} - T^{SE})}{2};\ PBS_{Asia} = \frac{(T^{AS} + T^{SE} - T^{AE})}{2};$$
$$PBS_{Europe} = \frac{(T^{SE} + T^{AE} - T^{AS})}{2} \tag{4}$$

where,

H_S is the sample polyTE heterozygosity within each regional population group being compared.
H_T is the total polyTE heterozygosity for both regional population groups being compared.
T^{XY} is the polyTE divergence level for regional population groups X and Y being compared.
T^{AS}, T^{AE} and T^{SE} denote the polyTE divergence levels between all three pairs of regional groups compared: Africa-Asia (AS), Africa-Europe (AE) and Asia-Europe (SE).

3.3. Detection of positively selected polyTEs using PBS values and population genetic modelling

Observed polyTE insertion *PBS* values were compared to a null distribution of values generated via evolutionary modelling in order to detect positively selected polyTEs. A Wright-Fisher based human population genetic model with two population divergence events, yielding the three extant regional population groups analyzed here, was implemented for this purpose (Fig. 4). Model parameter values for – (1) the time elapsed since the population divergence events and (2) the effective population sizes – were taken from a previous report by Gronau et al. (2011). The population genetic model was used to simulate polyTE insertion frequency dynamics starting with ancestral frequencies (p) ranging from 0.01 to 0.99, incrementing by steps of 0.01. The number of simulations (s_i) for each ancestral frequency (p_i) was performed proportional to $1/p_i$ such that a total number of 10 million simulated instances of regional population group-specific extant polyTE frequencies were generated:

$$s_i = int\left(\frac{10{,}000{,}000}{p_i \times \sum_{0.01}^{0.99} p}\right) \tag{1}$$

PolyTE frequencies simulated in this way were then used to calculate simulated *PBS* values, in the same way as described in the previous section, and the simulated *PBS* values were used to form a null distribution for statistical testing. For the purposes of statistical testing, simulated and observed *PBS* trees with similar mean branch lengths were compared and the deviation of the observed versus simulated regional population group-specific branch lengths were calculated. Since this procedure entailed multiple statistical tests, false discovery rate (FDR) q-values were used to establish statistical significance.

3.4. Gene regulatory potential of selected polyTEs

The locations of polyTE insertions that show evidence for positive selection were compared to several classes of gene regulatory features and functional genomic data. Computationally inferred enhancer locations from 125 cell lines were obtained from The Roadmap Epigenomics Project (Roadmap Epigenomics et al., 2015), and transcription factor binding site locations were obtained from the UCSC Genome Browser Txn Factor ChIP track. The locations of enhancer elements were computationally inferred using the core 15-state model from five chromatin marks assayed for 128 epigenomes across 30 different cell types (Roadmap Epigenomics et al., 2015). Human gene expression levels for 358 individuals from four European 1KGP populations (CEU, FIN, TSI, GBR) were obtained from the RNA-seq analysis performed by the GUEDVADIS project (Lappalainen et al., 2013b; t Hoen et al., 2013)' (Lappalainen et al., 2013a). Individuals' polyTE genotypes were compared to their gene expression levels to identify expression quantitative trait loci (eQTL) that correspond to polyTE insertion sites using the program Matrix eQTL (Shabalin, 2012). Matrix eQTL was run using the additive linear (least squares) model with covariates for gender and population.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gene.2018.06.077.

Funding information

L.R. and I.K.J. were supported by the IHRC-Georgia Tech Applied Bioinformatics Laboratory (ABiL). L.W. was supported by the Georgia Tech Bioinformatics Graduate Program.

Author contributions

L.R., S.V.Y. and I.K.J. conceived and designed the study. J.L., S.V.Y. and I.K.J. supervised the project. L.R. computed the frequencies, test statistics and performed overlap analysis. L.R. and J.L. designed and performed evolutionary modelling and significance testing. L.R. and L.W. performed expression analysis. L.R. and J.W. performed enhancer overlap analysis. L.R. and I.K.J. wrote the manuscript. All authors read and approved the manuscript.

Competing financial interests

The authors declare no competing financial interests.

References

Batzer, M.A., Deininger, P.L., 1991. A human-specific subfamily of Alu sequences. Genomics 9, 481–487.
Batzer, M.A., Gudi, V.A., Mena, J.C., Foltz, D.W., Herrera, R.J., Deininger, P.L., 1991. Amplification dynamics of human-specific (HS) Alu family members. Nucleic Acids Res. 19, 3619–3623.
Beck, C.R., Garcia-Perez, J.L., Badge, R.M., Moran, J.V., 2011. LINE-1 elements in structural variation and disease. Annu. Rev. Genomics Hum. Genet. 12, 187–215.
Bejerano, G., Lowe, C.B., Ahituv, N., King, B., Siepel, A., Salama, S.R., Rubin, E.M., Kent, W.J., Haussler, D., 2006. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature 441, 87–90.
Bowen, N.J., Jordan, I.K., 2007. Exaptation of protein coding sequences from transposable elements. Genome Dyn. 3, 147–162.
Brouha, B., Schustak, J., Badge, R.M., Lutz-Prigge, S., Farley, A.H., Moran, J.V., Kazazian Jr., H.H., 2003. Hot L1s account for the bulk of retrotransposition in the human population. Proc. Natl. Acad. Sci. U. S. A. 100, 5280–5285.
Burns, K.H., Boeke, J.D., 2012. Human transposon tectonics. Cell 149, 740–752.
Burton, F.H., Loeb, D.D., Voliva, C.F., Martin, S.L., Edgell, M.H., Hutchison 3rd, C.A., 1986. Conservation throughout mammalia and extensive protein-encoding capacity of the highly repeated DNA long interspersed sequence one. J. Mol. Biol. 187, 291–304.
Chenais, B., 2015. Transposable elements in cancer and other human diseases. Curr. Cancer Drug Targets 15, 227–242.
Chuong, E.B., Rumi, M.A., Soares, M.J., Baker, J.C., 2013. Endogenous retroviruses function as species-specific enhancer elements in the placenta. Nat. Genet. 45, 325–329.
Chuong, E.B., Elde, N.C., Feschotte, C., 2016. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science 351, 1083–1087.
Conley, A.B., Jordan, I.K., 2012. Cell type-specific termination of transcription by transposable element sequences. Mob. DNA 3, 15.
Conley, A.B., Piriyapongsa, J., Jordan, I.K., 2008. Retroviral promoters in the human genome. Bioinformatics 24, 1563–1567.
Cooper, G.M., Stone, E.A., Asimenos, G., Program, N.C.S., Green, E.D., Batzoglou, S., Sidow, A., 2005. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 15, 901–913.
Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A., Batzoglou, S., 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput. Biol. 6, e1001025.
Dewannieux, M., Esnault, C., Heidmann, T., 2003. LINE-mediated retrotransposition of marked Alu sequences. Nat. Genet. 35, 41–48.
Doolittle, W.F., Sapienza, C., 1980. Selfish genes, the phenotype paradigm and genome evolution. Nature 284, 601–603.
Fanning, T.G., Singer, M.F., 1987. LINE-1: a mammalian transposable element. Biochim. Biophys. Acta 910, 203–212.
Feschotte, C., 2008. Transposable elements and the evolution of regulatory networks. Nat. Rev. Genet. 9, 397–405.
Genomes Project, C., Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., Abecasis, G.R., 2015. A global reference for human genetic variation. Nature 526, 68–74.
Gonzalez, J., Lenkov, K., Lipatov, M., Macpherson, J.M., Petrov, D.A., 2008. High rate of recent transposable element-induced adaptation in Drosophila melanogaster. PLoS Biol. 6, e251.
Gonzalez, J., Macpherson, J.M., Petrov, D.A., 2009. A recent adaptive transposable element insertion near highly conserved developmental loci in *Drosophila melanogaster*. Mol. Biol. Evol. 26, 1949–1961.
Gonzalez, J., Karasov, T.L., Messer, P.W., Petrov, D.A., 2010. Genome-wide patterns of adaptation to temperate environments associated with transposable elements in drosophila. PLoS Genet. 6, e1000905.
Gronau, I., Hubisz, M.J., Gulko, B., Danko, C.G., Siepel, A., 2011. Bayesian inference of ancient human demography from individual genome sequences. Nat. Genet. 43, 1031–1034.
Hancks, D.C., Kazazian Jr., H.H., 2016. Roles for retrotransposon insertions in human disease. Mob. DNA 7, 9.
Hickey, D.A., 1982. Selfish DNA: a sexually-transmitted nuclear parasite. Genetics 101, 519–531.
Jacques, P.E., Jeyakani, J., Bourque, G., 2013. The majority of primate-specific regulatory sequences are derived from transposable elements. PLoS Genet. 9, e1003504.
Jordan, I.K., Kondrashov, F.A., Rogozin, I.B., Tatusov, R.L., Wolf, Y.I., Koonin, E.V., 2001. Constant relative rate of protein evolution and detection of functional diversification among bacterial, archaeal and eukaryotic proteins. Genome Biol. 2 (RESEARCH0053).
Jordan, I.K., Rogozin, I.B., Glazko, G.V., Koonin, E.V., 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet. 19, 68–72.
Kapusta, A., Kronenberg, Z., Lynch, V.J., Zhuo, X.Y., Ramsay, L., Bourque, G., Yandell, M., Feschotte, C., 2013. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. PLoS Genet. 9.
Kazazian Jr., H.H., Wong, C., Youssoufian, H., Scott, A.F., Phillips, D.G., Antonarakis, S.E., 1988. Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. Nature 332, 164–166.
de Koning, A.P., Gu, W., Castoe, T.A., Batzer, M.A., Pollock, D.D., 2011. Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet. 7, e1002384.
Kuhn, A., Ong, Y.M., Cheng, C.Y., Wong, T.Y., Quake, S.R., Burkholder, W.F., 2014. Linkage disequilibrium and signatures of positive selection around LINE-1 retrotransposons in the human genome. Proc. Natl. Acad. Sci. U. S. A. 111, 8131–8136.
Kunarso, G., Chia, N.Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y.S., Ng, H.H., Bourque, G., 2010. Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat. Genet. 42, 631–634.
Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., et al., 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.
Lappalainen, T., Sammeth, M., Friedlander, M.R., t Hoen, P.A., Monlong, J., Rivas, M.A., Gonzalez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., Barann, M., Wieland, T., Greger, L., van Iterson, M., Almlof, J., Ribeca, P., Pulyakhina, I., Esser, D., Giger, T., Tikhonov, A., Sultan, M., Bertier, G., MacArthur, D.G., Lek, M., Lizano, E., Buermans, H.P., Padioleau, I., Schwarzmayr, T., Karlberg, O., Ongen, H., Kilpinen, H., Beltran, S., Gut, M., Kahlem, K., Amstislavskiy, V., Stegle, O., Pirinen, M., Montgomery, S.B., Donnelly, P., McCarthy, M.I., Flicek, P., Strom, T.M., Geuvadis, C., Lehrach, H., Schreiber, S., Sudbrak, R., Carracedo, A., Antonarakis, S.E., Hasler, R., Syvanen, A.C., van Ommen, G.J., Brazma, A., Meitinger, T., Rosenstiel, P., Guigo, R., Gut, I.G., Estivill, X., Dermitzakis, E.T., 2013a. Transcriptome and Genome Sequencing Uncovers Functional Variation in Humans.
Lappalainen, T., Sammeth, M., Friedlander, M.R., t Hoen, P.A., Monlong, J., Rivas, M.A., Gonzalez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., Barann, M., Wieland, T., Greger, L., van Iterson, M., Almlof, J., Ribeca, P., Pulyakhina, I., Esser, D., Giger, T., Tikhonov, A., Sultan, M., Bertier, G., MacArthur, D.G., Lek, M., Lizano, E., Buermans, H.P., Padioleau, I., Schwarzmayr, T., Karlberg, O., Ongen, H., Kilpinen, H., Beltran, S., Gut, M., Kahlem, K., Amstislavskiy, V., Stegle, O., Pirinen, M., Montgomery, S.B., Donnelly, P., McCarthy, M.I., Flicek, P., Strom, T.M., Geuvadis, C., Lehrach, H., Schreiber, S., Sudbrak, R., Carracedo, A., Antonarakis, S.E., Hasler, R., Syvanen, A.C., van Ommen, G.J., Brazma, A., Meitinger, T., Rosenstiel, P., Guigo, R., Gut, I.G., Estivill, X., Dermitzakis, E.T., 2013b. Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501, 506–511.
Marino-Ramirez, L., Lewis, K.C., Landsman, D., Jordan, I.K., 2005. Transposable elements donate lineage-specific regulatory sequences to host genomes. Cytogenet. Genome Res. 110, 333–341.
Miller, W.J., Hagemann, S., Reiter, E., Pinsker, W., 1992. P-element homologous sequences are tandemly repeated in the genome of *Drosophila guanche*. Proc. Natl. Acad. Sci. U. S. A. 89, 4018–4022.
Moran, J.V., Holmes, S.E., Naas, T.P., DeBerardinis, R.J., Boeke, J.D., Kazazian Jr., H.H., 1996. High frequency retrotransposition in cultured mammalian cells. Cell 87, 917–927.
Notwell, J.H., Chung, T., Heavner, W., Bejerano, G., 2015. A family of transposable elements co-opted into developmental enhancers in the mouse neocortex. Nat. Commun. 6, 6644.
O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C.M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V.S., Kodali,

V.K., Li, W., Maglott, D., Masterson, P., McGarvey, K.M., Murphy, M.R., O'Neill, K., Pujar, S., Rangwala, S.H., Rausch, D., Riddick, L.D., Schoch, C., Shkeda, A., Storz, S.S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R.E., Vatsan, A.R., Wallin, C., Webb, D., Wu, W., Landrum, M.J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T.D., Pruitt, K.D., 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 44, D733–D745.

Ono, M., Kawakami, M., Takezawa, T., 1987. A novel human nonviral retroposon derived from an endogenous retrovirus. Nucleic Acids Res. 15, 8725–8737.

Orgel, L.E., Crick, F.H., 1980. Selfish DNA: the ultimate parasite. Nature 284, 604–607.

Ostertag, E.M., Goodier, J.L., Zhang, Y., Kazazian Jr., H.H., 2003. SVA elements are nonautonomous retrotransposons that cause disease in humans. Am. J. Hum. Genet. 73, 1444–1451.

Pavlicek, A., Jabbari, K., Paces, J., Paces, V., Hejnar, J.V., Bernardi, G., 2001. Similar integration but different stability of Alus and LINEs in the human genome. Gene 276, 39–45.

Payer, L.M., Steranka, J.P., Yang, W.R., Kryatova, M., Medabalimi, S., Ardeljan, D., Liu, C., Boeke, J.D., Avramopoulos, D., Burns, K.H., 2017. Structural variants caused by Alu insertions are associated with risks for many human diseases. Proc. Natl. Acad. Sci. U. S. A. 114, E3984–E3992.

Piriyapongsa, J., Marino-Ramirez, L., Jordan, I.K., 2007. Origin and evolution of human microRNAs from transposable elements. Genetics 176, 1323–1337.

Quinlan, A.R., 2014. BEDTools: the Swiss-army tool for genome feature analysis. Curr. Protoc. Bioinformatics 47 (11), 121–134.

Ray, D.A., Batzer, M.A., 2011. Reading TE leaves: new approaches to the identification of transposable element insertions. Genome Res. 21, 813–820.

Rebollo, R., Karimi, M.M., Bilenky, M., Gagnier, L., Miceli-Royer, K., Zhang, Y., Goyal, P., Keane, T.M., Jones, S., Hirst, M., Lorincz, M.C., Mager, D.L., 2011. Retrotransposon-induced heterochromatin spreading in the mouse revealed by insertional polymorphisms. PLoS Genet. 7, e1002301.

Reilly, M.T., Faulkner, G.J., Dubnau, J., Ponomarev, I., Gage, F.H., 2013. The role of transposable elements in health and diseases of the central nervous system. J. Neurosci. 33, 17577–17586.

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

Rishishwar, L., Marino-Ramirez, L., Jordan, I.K., 2016. Benchmarking computational tools for polymorphic transposable element detection. Brief. Bioinform. 18 (6), 908–918. http://dx.doi.org/10.1093/bib/bbw072.. Pubmed: https://www.ncbi.nlm.nih.gov/pubmed/27524380.

Rishishwar, L., Wang, L., Clayton, E.A., Mariño-Ramírez, L., McDonald, J.F., Jordan, I.K., 2017. Population and clinical genetics of human transposable elements in the (post) genomic era. Mobile Genetic Elements 7 (1), 1–20. http://dx.doi.org/10.1080/2159256X.2017.1280116.. eCollection, Pubmed: https://www.ncbi.nlm.nih.gov/pubmed/28228978.

Roadmap Epigenomics, C., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., Amin, V., Whitaker, J.W., Schultz, M.D., Ward, L.D., Sarkar, A., Quon, G., Sandstrom, R.S., Eaton, M.L., Wu, Y.C., Pfenning, A.R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R.A., Shoresh, N., Epstein, C.B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R.D., Lister, R., Hong, C., Gascard, P., Mungall, A.J., Moore, R., Chuah, E., Tam, A., Canfield, T.K., Hansen, R.S., Kaul, R., Sabo, P.J., Bansal, M.S., Carles, A., Dixon, J.R., Farh, K.H., Feizi, S., Karlic, R., Kim, A.R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T.R., Neph, S.J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R.C., Siebenthall, K.T., Sinnott-Armstrong, N.A., Stevens, M., Thurman, R.E., Wu, J., Zhang, B., Zhou, X., Beaudet, A.E., Boyer, L.A., De Jager, P.L., Farnham, P.J., Fisher, S.J., Haussler, D., Jones, S.J., Li, W., Marra, M.A., McManus, M.T., Sunyaev, S., Thomson, J.A., Tlsty, T.D., Tsai, L.H., Wang, W., Waterland, R.A., Zhang, M.Q., Chadwick, L.H., Bernstein, B.E., Costello, J.F., Ecker, J.R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J.A., Wang, T., Kellis, M., 2015. Integrative analysis of 111 reference human epigenomes. Nature 518, 317–330.

Salem, A.H., Kilroy, G.E., Watkins, W.S., Jorde, L.B., Batzer, M.A., 2003. Recently integrated Alu elements and human genomic diversity. Mol. Biol. Evol. 20, 1349–1361.

Schmid, C.W., Deininger, P.L., 1975. Sequence organization of the human genome. Cell 6, 345–358.

Schmidt, D., Schwalie, P.C., Wilson, M.D., Ballester, B., Goncalves, A., Kutter, C., Brown, G.D., Marshall, A., Flicek, P., Odom, D.T., 2012. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. Cell 148, 335–348.

Shabalin, A.A., 2012. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics 28, 1353–1358.

Shen, L., Wu, L.C., Sanlioglu, S., Chen, R., Mendoza, A.R., Dangel, A.W., Carroll, M.C., Zipf, W.B., Yu, C.Y., 1994. Structure and genetics of the partially duplicated gene RP located immediately upstream of the complement C4A and the C4B genes in the HLA class III region. Molecular cloning, exon-intron structure, composite retroposon, and breakpoint of gene duplication. J. Biol. Chem. 269, 8466–8476.

Shriver, M.D., Kennedy, G.C., Parra, E.J., Lawson, H.A., Sonpar, V., Huang, J., Akey, J.M., Jones, K.W., 2004. The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. Hum. Genomics 1, 274–286.

Solyom, S., Ewing, A.D., Rahrmann, E.P., Doucet, T., Nelson, H.H., Burns, M.B., Harris, R.S., Sigmon, D.F., Casella, A., Erlanger, B., Wheelan, S., Upton, K.R., Shukla, R., Faulkner, G.J., Largaespada, D.A., Kazazian Jr., H.H., 2012. Extensive somatic L1 retrotransposition in colorectal tumors. Genome Res. 22, 2328–2338.

Speir, M.L., Zweig, A.S., Rosenbloom, K.R., Raney, B.J., Paten, B., Nejad, P., Lee, B.T., Learned, K., Karolchik, D., Hinrichs, A.S., Heitner, S., Harte, R.A., Haeussler, M., Guruvadoo, L., Fujita, P.A., Eisenhart, C., Diekhans, M., Clawson, H., Casper, J., Barber, G.P., Haussler, D., Kuhn, R.M., Kent, W.J., 2016. The UCSC genome browser database: 2016 update. Nucleic Acids Res. 44, D717–D725.

Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., Konkel, M.K., Malhotra, A., Stutz, A.M., Shi, X., Paolo Casale, F., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., Malig, M., Chaisson, M.J., Walter, K., Meiers, S., Kashin, S., Garrison, E., Auton, A., Lam, H.Y., Jasmine Mu, X., Alkan, C., Antaki, D., Bae, T., Cerveira, E., Chines, P., Chong, Z., Clarke, L., Dal, E., Ding, L., Emery, S., Fan, X., Gujral, M., Kahveci, F., Kidd, J.M., Kong, Y., Lameijer, E.W., McCarthy, S., Flicek, P., Gibbs, R.A., Marth, G., Mason, C.E., Menelaou, A., Muzny, D.M., Nelson, B.J., Noor, A., Parrish, N.F., Pendleton, M., Quitadamo, A., Raeder, B., Schadt, E.E., Romanovitch, M., Schlattl, A., Sebra, R., Shabalin, A.A., Untergasser, A., Walker, J.A., Wang, M., Yu, F., Zhang, C., Zhang, J., Zheng-Bradley, X., Zhou, W., Zichner, T., Sebat, J., Batzer, M.A., McCarroll, S.A., Genomes Project, C., Mills, R.E., Gerstein, M.B., Bashir, A., Stegle, O., Devine, S.E., Lee, C., Eichler, E.E., Korbel, J.O., 2015. An integrated map of structural variation in 2,504 human genomes. Nature 526, 75–81.

Sundaram, V., Cheng, Y., Ma, Z., Li, D., Xing, X., Edge, P., Snyder, M.P., Wang, T., 2014. Widespread contribution of transposable elements to the innovation of gene regulatory networks. Genome Res. 24, 1963–1976.

t Hoen, P.A., Friedlander, M.R., Almlof, J., Sammeth, M., Pulyakhina, I., Anvar, S.Y., Laros, J.F., Buermans, H.P., Karlberg, O., Brannvall, M., Consortium, G., den Dunnen, J.T., van Ommen, G.J., Gut, I.G., Guigo, R., Estivill, X., Syvanen, A.C., Dermitzakis, E.T., Lappalainen, T., 2013. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. Nat. Biotechnol. 31, 1015–1022.

Ullu, E., Tschudi, C., 1984. Alu sequences are processed 7SL RNA genes. Nature 312, 171–172.

Wang, H., Xing, J., Grover, D., Hedges, D.J., Han, K., Walker, J.A., Batzer, M.A., 2005. SVA elements: a hominid-specific retroposon family. J. Mol. Biol. 354, 994–1007.

Wang, L., Norris, E.T., Jordan, I.K., 2017a. Human retrotransposon insertion polymorphisms are associated with health and disease via gene regulatory phenotypes. Front. Microbiol. 8, 1418.

Wang, L., Rishishwar, L., Marino-Ramirez, L., Jordan, I.K., 2017b. Human population-specific gene expression and transcriptional network modification with polymorphic transposable elements. Nucleic Acids Res. 45, 2318–2328.

Weber, M.J., 2006. Mammalian small nucleolar RNAs are mobile genetic elements. PLoS Genet. 2, 1984–1997.

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z.X., Pool, J.E., Xu, X., Jiang, H., Vinckenbosch, N., Korneliussen, T.S., Zheng, H., Liu, T., He, W., Li, K., Luo, R., Nie, X., Wu, H., Zhao, M., Cao, H., Zou, J., Shan, Y., Li, S., Yang, Q., Asan, Ni, P., Tian, G., Xu, J., Liu, X., Jiang, T., Wu, R., Zhou, G., Tang, M., Qin, J., Wang, T., Feng, S., Li, G., Huasang, Luosang, J., Wang, W., Chen, F., Wang, Y., Zheng, X., Li, Z., Bianba, Z., Yang, G., Wang, X., Tang, S., Gao, G., Chen, Y., Luo, Z., Gusang, L., Cao, Z., Zhang, Q., Ouyang, W., Ren, X., Liang, H., Zheng, H., Huang, Y., Li, J., Bolund, L., Kristiansen, K., Li, Y., Zhang, Y., Zhang, X., Li, R., Li, S., Yang, H., Nielsen, R., Wang, J., Wang, J., 2010. Sequencing of 50 human exomes reveals adaptation to high altitude. Science 329, 75–78.