Co-option of the gibbon-specific *LAVA* retrotransposon in DNA repair pathways

Mariam Okhovat[1*], Kimberly A. Nevonen[1], Brett Davis[1], Pryce Michener[1†], Samantha Ward[1], Mark Milhaven[2], Lana Harshman[3,4], Ajuni Sohota[3,4], Rachel J. O'Neill[5,6], Nadav Ahituv[3,4], Krishna R. Veeramah[2], Lucia Carbone[1,7-9*]

1. Department of Medicine, Knight Cardiovascular Institute, Oregon Health and Science University, Portland, OR 97239, USA
2. Department of Ecology and Evolution/ Institute for Advance Computational Science, Stony Brook University, Stony Brook, NY 11794, USA
3. Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA 94158, USA
4. Institute for Human Genetics, University of California San Francisco, San Francisco, CA 94158, USA
5. Institute for Systems Genomics, University of Connecticut, Storrs, CT 06269, USA
6. Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT 06269, USA
7. Department of Molecular and Medical Genetics, Oregon Health and Science University, Portland, OR 97239, USA
8. Division of Genetics, Oregon National Primate Research Center, Beaverton, OR 97006, USA
9. Department of Informatics and Clinical Epidemiology, Oregon Health and Science University, Portland, OR 97239, USA

† Current affiliation: University of Massachusetts Medical School, Worcester, MA 01605, USA

*Corresponding authors

Abstract

Transposable elements (TEs) can shape gene regulation networks by being co-opted as enhancers. However, the contribution of lineage-specific TE insertions to recent adaptations remains poorly understood. Gibbons present a suitable model to study these contributions, as they have evolved many distinct traits, including heavily rearranged genomes and a novel TE called *LAVA*. The LAVA retrotransposon is still active in the gibbon genome and is thought to have contributed to evolution of gibbon-specific traits. In this study, we characterized fixed and polymorphic LAVA insertions across multiple gibbon genomes and found that 10% of all LAVA elements overlap chromatin states associated with enhancer function. Moreover, LAVA was enriched in multiple transcription factor motifs, was bound by the important lymphoid transcription factor PU.1, and was associated with higher levels of gene expression in *cis*. Interestingly, despite the highly similar genomic distribution and epigenetic characteristics of fixed and polymorphic LAVA, only fixed LAVA insertions showed strong signatures of positive selection, and were enriched near genes implicated in DNA repair. Altogether, our population genetics, epigenetics, and evolutionary analyses indicate that several LAVA insertions have been co-opted in the gibbon genome as *cis*-regulatory elements. Specifically, a subset of the fixed LAVA insertions appear to have been co-opted to enhance regulation of DNA repair genes, likely as an

adaptive mechanism to improve genome integrity in response to the genomic rearrangements occurring in the gibbon lineage.

Introduction

Transposable elements (TEs) comprise nearly half of mammalian genomes and provide a major source of genetic and epigenetic variation during evolution. While many studies have focused on the disruptive consequences of TE insertions, especially those that impact human health (*1*), growing evidence is revealing widespread presence of advantageous TE insertions across lineages (*2*). Depending on their impact on the host, TEs face different evolutionary fates. Most TE insertions are neutral and therefore drift randomly in the population, while disruptive insertions are actively selected against and removed from the population. The occasional adaptive TE insertions however, are favored by selection and may ultimately become evolutionarily incorporated in the host genome, a process known as “co-option” or “exaptation”. To date, several examples of co-opted TEs have been reported in vertebrates [Reviewed in (*3*)].

Many co-opted TEs function as *cis*-regulatory elements (i.e. enhancers) and are capable of modifying gene expression in a tissue- or time-specific manner (*4*). These regulatory TEs often contain transcription factor (TF) binding sites and their transposition in the genome can reshape entire gene regulatory networks by introducing similar regulatory modules nearby multiple genes. Since TE content varies drastically across lineages, co-option of lineage-specific regulatory TEs likely represents a major mechanism for rapid evolution of gene expression patterns within lineages (*5*, *6*). Indeed, a recent comparative study in primates demonstrated that nearly all human-specific regulatory elements overlapped TE sequences, and that most TE families enriched at *cis*-regulatory regions were relatively young and lineage-specific (*6*). These and other findings point to young lineage-specific regulatory TEs as a primary source for evolution of regulatory novelty in primates (*6*–*8*). However, due to technical challenges associated with studying recent TE insertions (e.g. low mappability), their contributions to the evolution of gene-regulatory adaptations remains poorly understood, especially in non-human primates.

Among primates, the endangered gibbons (*Hylobatidae*) present an attractive model for exploring functional contributions of a lineage-specific TE. Gibbons (or small apes) have an intriguing evolutionary history, and have evolved many unique traits [e.g. locomotion via brachiating, monogamy, etc. (*9*)]. Most notably, the gibbon lineage has experienced drastic genomic rearrangements since its divergence from the common Hominidae ancestor ~17 million years ago (mya) (*10*). These evolutionary rearrangements are not only evident through comparisons with great ape genomes, but also in the vastly different karyotypes of the four extant gibbon genera: *Nomascus* (2n=52)*, Hylobates* (2n=44)*, Hoolock* (2n=38) *and Siamang* (2n=50), which split only 5 mya. The factors leading to these evolutionary genome reorganizations are not fully understood, but a gibbon-specific retrotransposon, called *LAVA* (Fig. 1A), may have played a role (*11*).

The LAVA element is a non-autonomous composite retrotransposon consisting of portions of repeats found in most primate genomes (CT-rich, *Alu*-like, a truncated SVA element, and portions of *Alu*Sz and L1ME5 elements), but the fully assembled element is only found among gibbons (*12*, *13*). In the original analysis of the reference gibbon genome, which was derived from a northern white-cheeked gibbon (*Nomascus leucogenys*), nearly half of the 1,256 intact LAVA insertions found were located within or near genes, especially genes involved in regulation of cell cycle and chromosome segregation (*11*). Since some LAVA insertions are capable of terminating gene transcription prematurely, disruption of these cell cycle genes by LAVA may have caused the evolutionary genomic rearrangements in the gibbon lineage (*11*). However, LAVA's successful and ongoing propagation in the gibbon genome (*14*), in spite of its disruptive effects, indicates that it may have also provided adaptive functions. TEs like LAVA, that have high prevalence near genes, have a stronger propensity for adopting enhancer function and modulating adjacent gene expression (*15*). Thus, putative adaptive contributions from LAVA likely involved regulation of nearby genes and may have resulted in its co-option as a *cis*-regulatory element.

In this study, we characterized LAVA insertions across multiple gibbon genomes and used genomic, epigenetic, and evolutionary analyses to investigate the possible co-option of LAVA as a *cis*-regulatory element and its potential adaptive contributions to the evolution of the gibbon genome.

Results

Genome-wide identification of LAVA insertions in the gibbon lineage reveals genus-specific expansion patterns.

Distribution of LAVA insertions (Fig. 1A) is expected to vary across individuals, since the LAVA retrotransposon is still active in the gibbon genome (*14*). To this end, we characterized LAVA insertions in 23 whole genome sequencing (WGS) datasets obtained from blood of unrelated individuals across the four extant gibbon genera (Fig. 1B, Table S1). Considering the composite structure of LAVA (Fig. 1A), as well as the lower quality of the gibbon genome assembly compared to human, selecting the right software for identifying LAVA insertions was critical. We validated the suitability of the Mobile Element Locator Tool (MELT; (*16*)) through *in-silico* simulation analyses (Supplemental Text). Overall, MELT was able to predict the position of most ($\geq$75%) simulated LAVA indel sites within 10bp of their true position (Supplemental Text, Figs. S1A-B) and identify simulated LAVA indels with high sensitivity and specificity, when coverage was higher than 10X (Supplemental Text, Figs. S1C-D). Using MELT on our 23 gibbon WGS datasets, which all had >10X coverage, we initially identified 20,734 *de novo* LAVA insertions. These newly discovered LAVA insertions were combined with 1,118 full-length LAVA insertions previously identified on the assembled chromosomes of the gibbon reference genome (Nleu3.0; (*16*)). To minimize false positives, we filtered these 21,852 LAVA insertions based on strict quality, length and frequency criteria (see Materials and Methods), which reduced the total number of LAVA insertion sites to 5,490 high-confidence hits.

Upon inspection of genome-wide LAVA genotypes, we observed that the abundance of LAVA insertions was highly variable across genera (Fig. 1C), with individuals in the *Nomascus* and *Hoolock* genera carrying the smallest and largest average copies of LAVA per genome, respectively (*Nomascus*= 2,730.6 $\pm$ 83.4, *Hylobates*= 3,396.0 $\pm$ 143.7, *Siamang*= 3,595.0 $\pm$ 29.7, *Hoolock*= 4,452.3 $\pm$ 110.0; mean LAVA copies in diploid genome $\pm$ stdev)). Of the total 5,490 LAVA insertion sites identified across genera, 905 (16.5%) were homozygous for presence of LAVA in all 23 individuals, 1,317 (24.0%) contained a LAVA insertion at least once in each genus, and approximately half (52.6%) contained LAVA insertions exclusively in one genus (i.e. genus-specific). Unsupervised hierarchical clustering and

logarithmic principal component analysis (PCA) of LAVA genotypes showed that individuals of the same species and genus grouped together (Figs. 1D-E), indicating that our LAVA discovery and genotyping successfully captured genus-specific LAVA expansion patterns in the gibbon lineage.

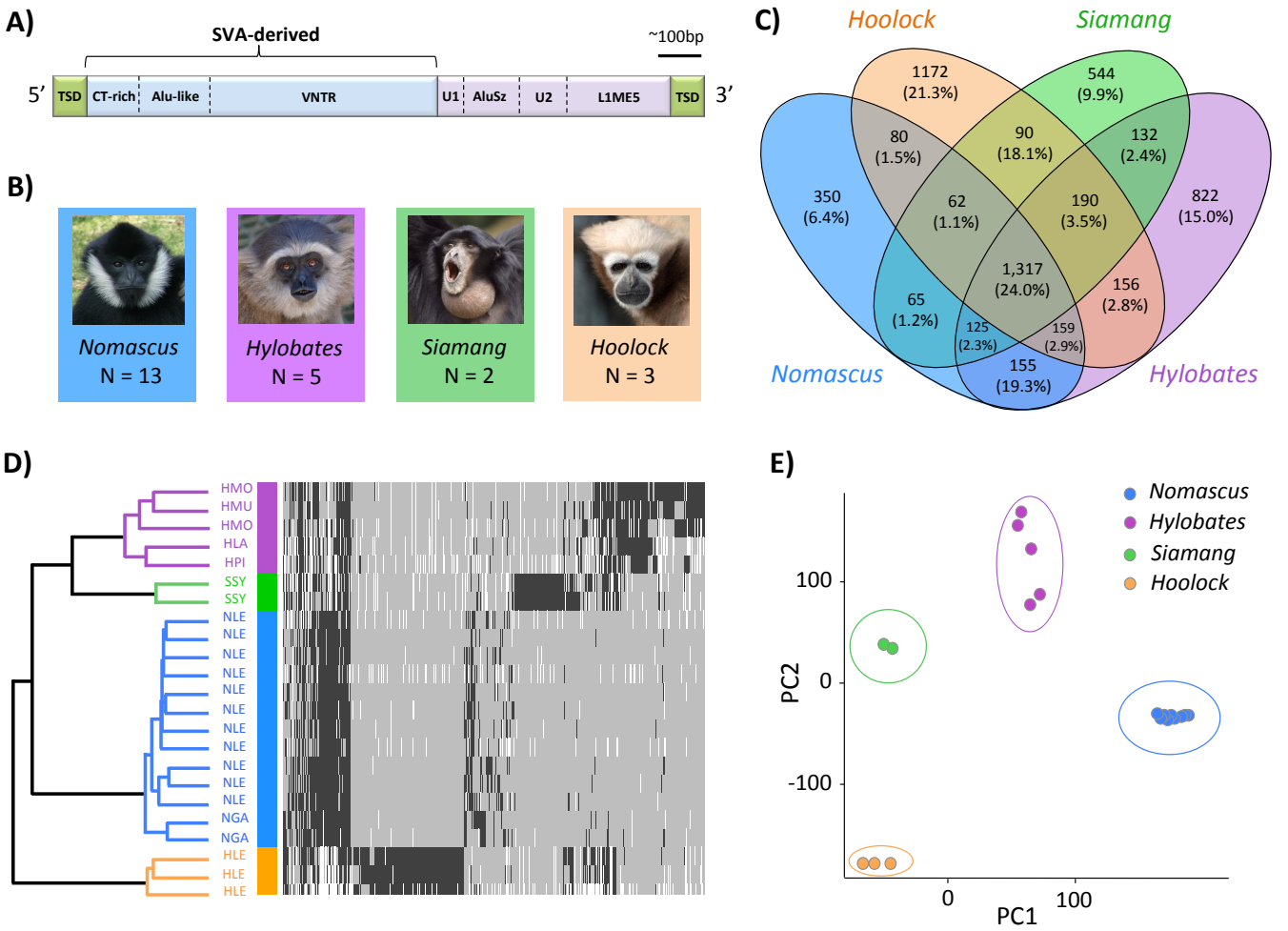


Figure 1. LAVA displays genus specific expansion patterns. A) A schematic representation of the composite LAVA element (TSD= target size duplication, VNTR= variable number tandem repeat, U= unique non-repetitive sequence). **B)** WGS data from four gibbon genera (*Nomascus (NLE= N. leucogenys, NGA= N. gabriellae), Hylobates (HLA= H. lar, HMO= H. moloch, HPI= H. pileatus), Siamang (SSY= S. symphalangus), and Hoolock (HLE= H. hoolock)*) was used to identify LAVA indels (N= sample size). **C)** A Venn diagram displays the distribution of LAVA insertions across genera. **D)** Unsupervised hierarchical clustering of individuals based on LAVA genotype is shown. In the heatmap, presence of LAVA is marked in dark gray and absence of LAVA is marked with light gray (missing data is shown in white). **E)** PC1 and PC2 from logarithmic Principal Component Analysis of LAVA genotypes is plotted.

The genomic distribution of LAVA is non-homogenous and enriched nearby repeats and genes

TEs often show non-random distribution in the genome, likely as a result of their propagation strategies (*17*, *18*). To investigate the distribution of LAVA elements within and across individuals, we focused only on LAVA insertion sites found in the *Nomascus leucogenys* (NLE), which is the same species used to build the gibbon reference genome (*11*). Of the total 2,266 LAVA insertion sites we found among NLE gibbons, 48% (1,095) were homozygous for presence of LAVA in all 11 NLE individuals and were thus called fixed-LAVA elements (Fig. 2A, Table S2). This group includes LAVA insertions that have reached high frequency or fixation either due to selection or random drift. The remaining LAVA insertions were present in lower frequencies (<95%) in our population and displayed presence/absence polymorphisms (polymorphic LAVA, referred to as poly-LAVA here on; Fig. 2A, Table S2).

Both fixed- and poly-LAVA insertion sites were non-homogenously distributed across chromosomes ($\chi_{\text{fixed-LAVA}}^{2(25, N=1095)}$, $p<0.001$; $\chi_{\text{poly-LAVA}}^{2(25, N=2266)}$, $p<0.001$; Figs. 2A, S2A) and often appeared in clusters within chromosomes (one-tailed permutation $p<0.001$, Supplemental Text), suggesting a nonrandom distribution likely influenced by genomic context. Consistent with observations for other retrotransposons (*19*), LAVA elements were generally enriched nearby repeats, specifically retroposons (permutation $q<0.05$; Fig. S2B). Notably, both fixed- and poly-LAVA insertions were often found near genes, with over half of LAVA insertions being found inside introns or within 3Kb of a gene (Fig. 2B). In line with this observation, there was a strong correlation between LAVA- and gene-density across chromosomes ($R^2_{\text{fixed-LAVA}}=0.23$, $p=0.008$ and $R^2_{\text{poly-LAVA}}=0.48$, $p=0.0006$; Fig. 2C). Hence, fixed- and poly-LAVA insertion sites appear to be equally abundant and similarly distributed in the NLE genome.

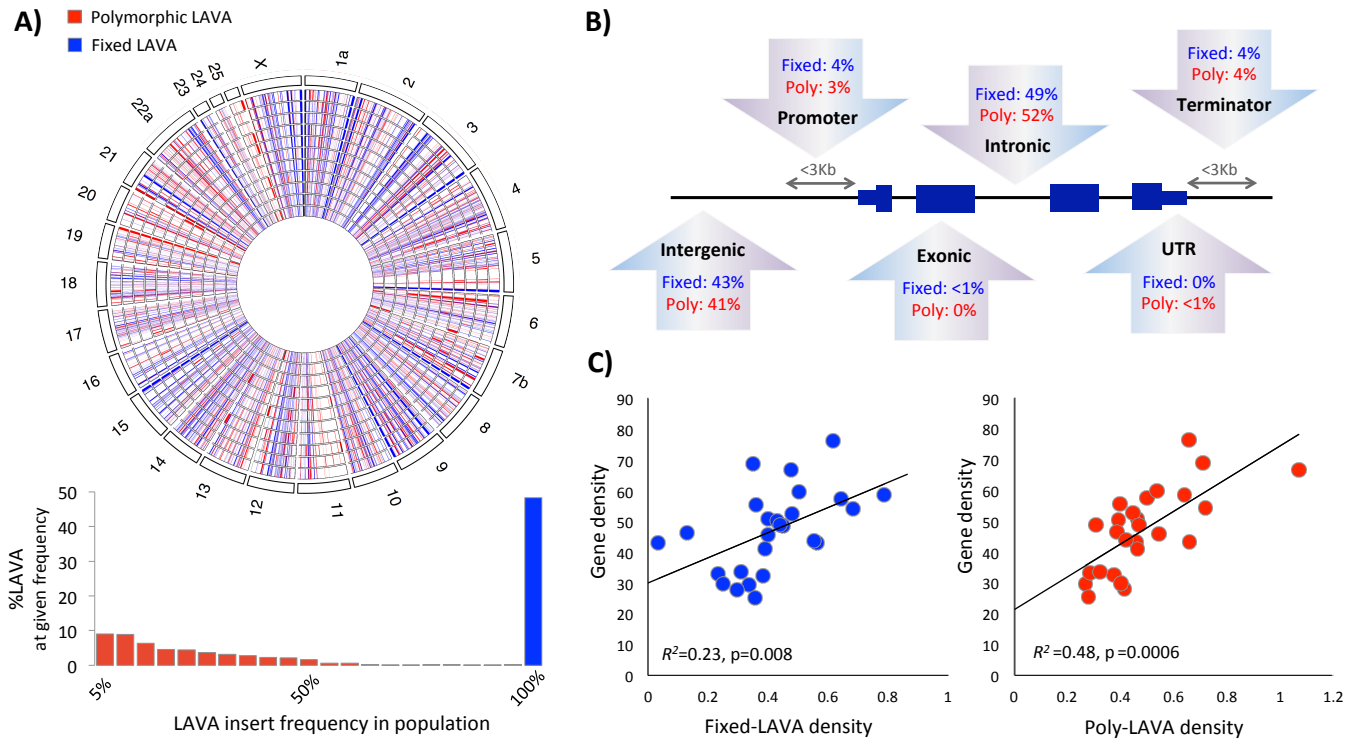


Figure 2) Fixed and poly-LAVA insertions have similar genomic distribution. A) LAVA insertions present at 100% frequency in the population were classified as fixed (blue) and the rest were classified as polymorphic (red). Circos plots of fixed- and poly-LAVA insertions demonstrate the non-homogenous distribution of LAVA insertions across chromosomes. **B)** Percentages of fixed- and poly-LAVA insertions found across different regions of the genome, with respect to genes, are shown. **C)** Correlations between gene- and LAVA-density per chromosome are shown for fixed (blue) and poly-LAVA (red).

LAVA elements overlap chromatin signatures of enhancer activity

Many co-opted TEs adopt enhancer functions in the host genome (*4*, *15*, *20*). To investigate if LAVA elements display biochemical hallmarks of enhancer activity, we performed chromatin immunoprecipitation sequencing (ChIP-seq) against three activating (H3K4me1, H3K27ac, and H3K4me3) and two repressing histone marks (H3K27me3 and H3K9me3), on EBV-transformed lymphoblastoid cell lines (LCL) that we established from three unrelated NLE individuals (Table S4, Supplemental Text). We combined replicate ChIP-seq data after finding high correlation across biological replicates (Pearson correlation coefficient 75-96%; Fig. S3). Using ChromHMM (*21*), we assigned nine chromatin states based on the combination of signals from different histone marks (Fig. 3A). As expected, the fixed- and poly-LAVA appeared silenced, and were most highly enriched in constitutive "Heterochromatin" and "Polycomb-Repressed" chromatin states, respectively (Fig. 3A). However, we also found modest enrichment of chromatin states associated with active, poised, or bivalent enhancer activity in LAVA elements (Fig. 3A-B). In total, 2% of both fixed and poly-LAVA showed overlap with active enhancer state, and another 7% and 8% had overlap with poised/bivalent enhancer states in fixed and poly-LAVA, respectively (Figs. 3B-C).

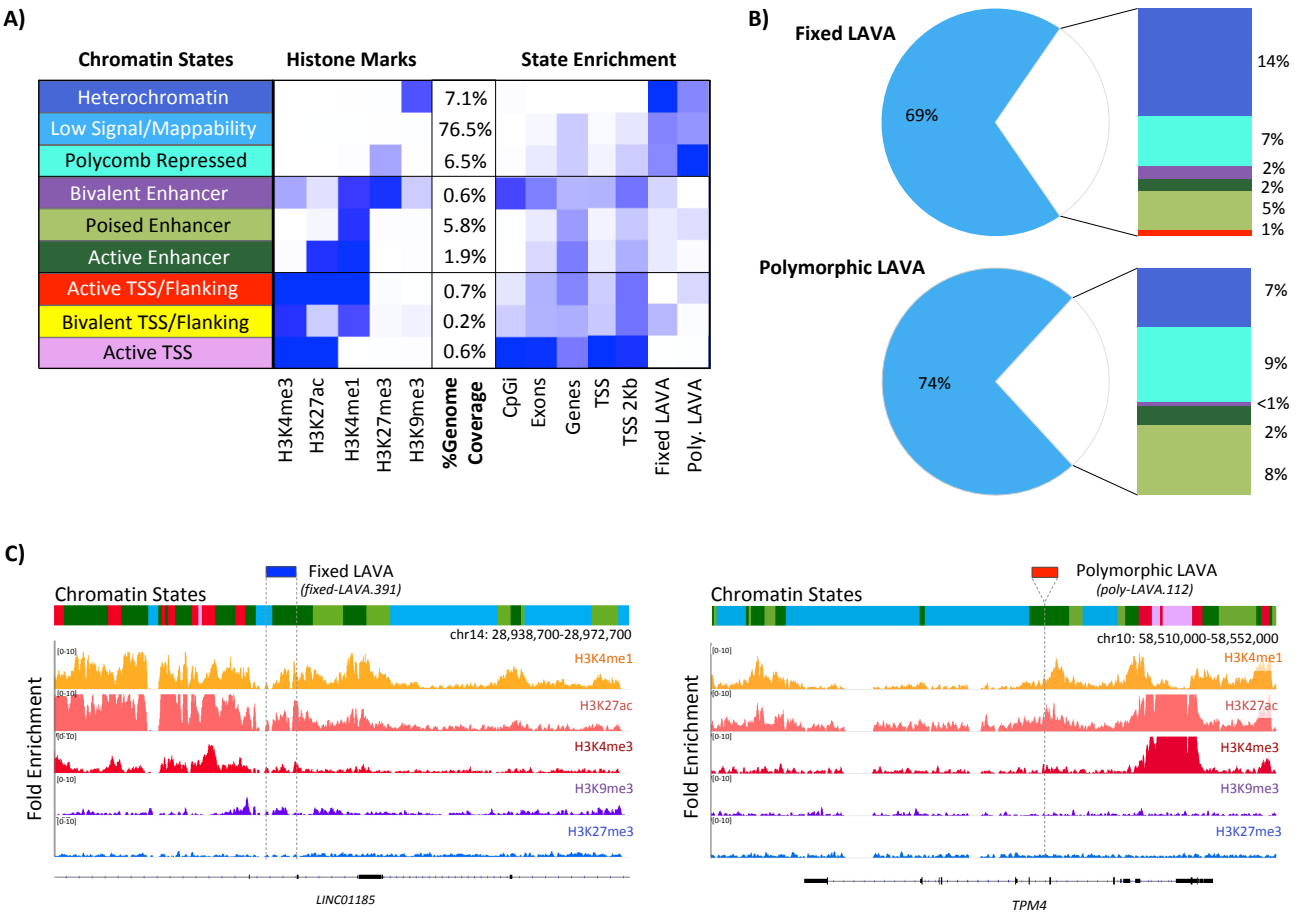


Figure 3. A subset of LAVA elements display enhancer chromatin states. A) Nine chromatin states were identified based on ChIP-seq signal from five different histone marks. Each state's percent genome coverage and fold enrichment within genetic features are shown in a heatmap. **B)** Pie charts depict percentages of fixed (*top*) and poly-LAVA insertions (*bottom*) that have at least 1bp overlap with each chromatin state. Colors correspond to chromatin states in (A). **C)** ChIP-seq fold-enrichment tracks shown for a fixed- (*left*) and poly-LAVA (*right*) that overlap "active enhancer" chromatin state (dark green). A single dashed line marks the poly-LAVA insert site, since this element is absent from in the reference genome.

LAVA elements contain multiple transcription factor binding motifs and bind to PU.1

Regulatory TEs contribute roughly 20% of all transcription factor (TF) binding sites in mammalian genomes (22). We identified a conservative list of 6 TFs whose motifs were significantly overrepresented in LAVA sequences and were predicted to bind LAVA with high affinity [q<0.05; (*23*)]. These TFs were PU.1 (encoded by *SPI1*), STAT3, SRF, SOX10, SOX17, and ZNF143 (Table 1).

PU.1, whose recognition motif is among the most highly enriched in LAVA (Table S3), is an important TF in development of B-lymphoid cells and has been linked to regulation of cell cycle duration and genome mutability (*24*). To directly investigate PU.1 binding to LAVA, we performed ChIP-seq against PU.1 on two gibbon LCLs (Table S4). ChIP-seq peaks were significantly enriched for the consensus PU.1 recognition motif (Fig. 4A, top) and co-occurred with active histone marks (Fig. 4A, bottom). We found a total of 16 repeat families significantly enriched in gibbon PU.1 ChIP-seq samples relative to input, and 12 families that were significantly under-represented (q<0.05; Fig. 4B, left). Of note, the most significantly enriched repeat family was SVA (log fold change= 0.4, q=1.52e-59; Fig. 4B, left), which should be equated to enrichment of LAVA, because almost all SVA repeats in the gibbon genome appear as part of a composite LAVA element (*25*, *26*). Although most PU.1 binding motifs within LAVA sequences were clustered in the SVA subunit, we did not detect enrichment of the SVA repeat family in human PU.1 ChIP-seq data (log fold change= -0.03, q=0.8; Fig. 4B, right) obtained from ENCODE (*27*, *28*, *28*). Given that the human genome contains thousands of SVA insertions (*25*), but no LAVA, we concluded that the widespread binding of PU.1 to SVA repeat family in gibbons appears to be specific to the LAVA element.

To identify specific putative LAVA enhancers that bind PU.1, we searched for PU.1 ChIP-seq peaks overlapping LAVA insertions. These events are expected to be highly elusive due to removal of multi-mapping repetitive short reads during ChIP-seq peak calling. Nevertheless, we found three overlaps, with two PU.1 peaks overlapping fixed-LAVA elements (fixed-LAVA.1257 and fixed-LAVA.1087) and one overlapping a polymorphic LAVA insertion site (poly-LAVA.3254). Fixed-LAVA.1257 displayed bivalent enhancer chromatin state (Fig. 4C) and poly-LAVA.3254 overlapped active enhancer chromatin state, while no histone signal was found at fixed-LAVA.1087.

TRAP rank	HOMER rank	TF	Consensus Motif
2	41	PU.1 (Spi1)	
7	122	STAT3	
13	119	SRF	
21	69	SOX10	
27	161	SOX17	
28	123	ZNF143	

Table 1. Transcription factor (TF) motifs with significant enrichment in LAVA sequences (q <0.05).

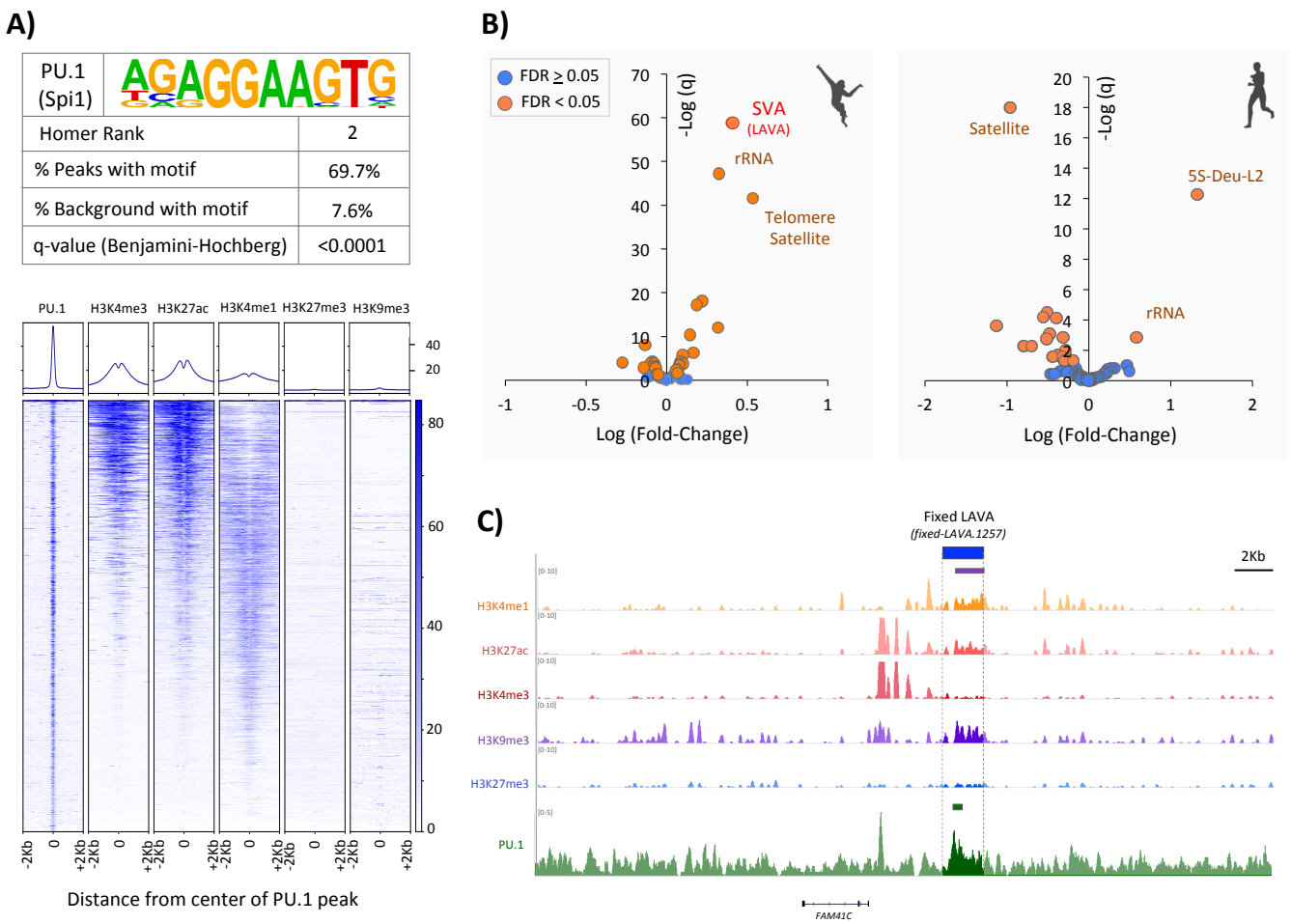


Figure 4. The PU.1 transcription factor binds to LAVA. A) PU.1 ChIP-seq peaks are enriched for the consensus PU.1 recognition site (top) and co-localize with epigenetic signatures of active chromatin (bottom). **B)** Volcano plots demonstrate repeat families significantly enriched (q<0.05) in PU.1 ChIP-seq from gibbon (left) and human (right). **C)** Example of a fixed-LAVA (thick blue bar) overlapping a bivalent enhancer chromatin state (thin purple bar) and a PU.1 ChIP-seq peak (green bar).

Genes near LAVA display expression patterns that differ from the rest of the genome

To investigate LAVA's effects on expression of nearby genes, we generated RNA-seq data from nine available NLE gibbon LCLs and investigated global patterns of gene expression near LAVA. We considered genes with depth normalized read counts (reads per million, CPM) higher than 0.5 in at least two of the nine gibbons, to be "actively expressed". Using this proxy, 72% of genes associated with fixed-LAVA (448 out of 620) and 69% of genes associated with poly-LAVA (478 out of 694) were considered actively expressed. These proportions were not significantly different from each other (two tailed Fisher's exact test, p=0.58), but both were significantly higher than the 39% of genes (15,715 out of 40,504) that were actively expressed genome-wide (two-tailed chi-square test with Yates correction, p<0.0001). Moreover, permutation analysis showed that among actively expressed genes, those located near fixed- and poly-LAVA had significantly higher mean expression compared to the null distribution in the whole-genome (two-tailed permutation test, p<0.0001; Figs. 5A, left). Of note, mean-normalized inter-individual variability in gene expression (i.e. coefficient of variation, CV) of genes nearby fixed-, but not poly-LAVA, was significantly lower than the null distribution of gene expression CV genome-wide (two-tail permutation tests, $p_{fixed\text{-}LAVA}$=0.0001 and $p_{poly\text{-}LAVA}$ =0.13; Fig. 5A, right). Overall, these observations indicated that genes nearby LAVAs, especially fixed-LAVA insertions, display patterns of expression that are broadly different from the rest of the genome.

Next, we took advantage of the presence/absence of poly-LAVA elements and examined correlation between genotype at individual LAVA insertions and expression level of genes within 1Mb. Despite being underpowered due to our small sample-size, we found two poly-LAVA insertions associated with significant increase in expression of a nearby gene (q<0.05; Fig. 5A). One of these genes was *TRPV1* (transient receptor potential cation channel subfamily V member 1), whose expression was significantly higher in an individual with a LAVA insertion ~300Kb downstream (p=8.38e-07, q=0.01; Fig. 5A, top). In humans, *TRPV1* is highly expressed in the central nervous system and has a human-specific SVA insertion hypothesized to have contributed to the evolution of human-specific behaviors (*29*). The other gene was the secreted frizzled related protein (*SFRP1*). This gene, which is a Wnt antagonist implicated in cell-cycle regulation and senescence (Elzi et al. 2012, Zhou et al. 1998),

was more highly expressed in when a LAVA insertion was present ~800Kb upstream (p=3.8e-06, q=0.04; Fig. 5A, bottom). Combined, these findings supported LAVA's putative role as an enhancer of gene expression.

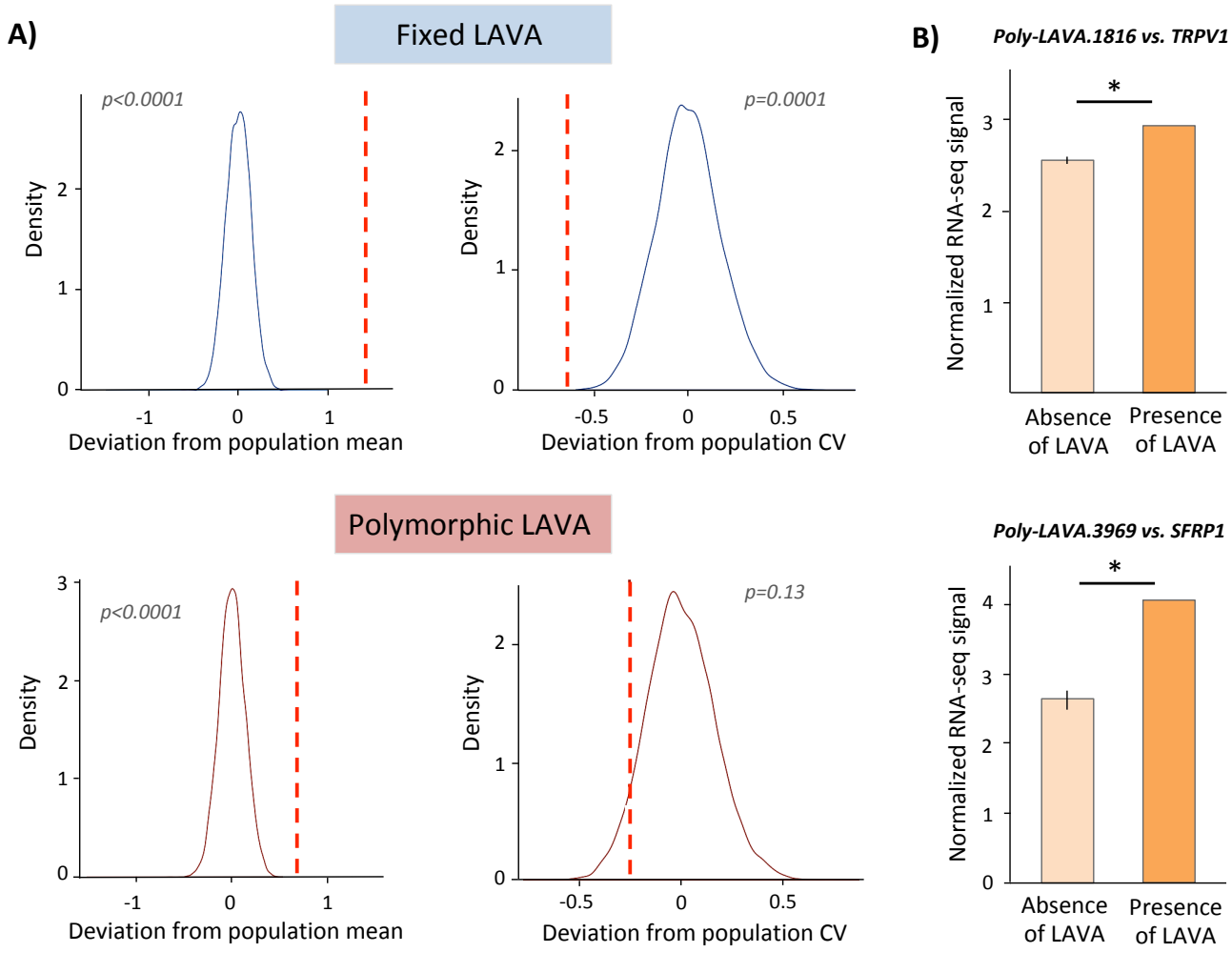


Figure 5. LAVA is associated with increased expression of genes in *cis*. A) The null distribution of mean (*left*) and coefficient of variation (CV) of gene expression (*right*) in shown against the observed mean and CV (red dashed line) for genes nearby fixed- (*top*) and poly-LAVA (*bottom*). **B)** Normalized expression of *TRPV1* (*top*) and *SFRP1* (*bottom*) are shown against genotype at poly-LAVA.1816 and poly-LAVA.3969, respectively. Bar heights represent mean, and error bars are standard error (*= q<0.05).

Fixed, but not polymorphic, LAVA insertions show strong signatures of selection

To test if LAVA has been co-opted in the gibbon lineage, we investigated signatures of selection at fixed- and poly-LAVA inserts, individually and collectively. We measured Tajima's D (*30*) in 10Kb windows directly flanking both sides of LAVA insertion sites, since measuring selection within the LAVA elements is not possible due to the absence of many LAVA insertions from the reference genome, as well as ambiguities in sequences of LAVA elements present in the reference genome. We found that 20 of the 808 poly-LAVA (2.5%) and 35 of the 734 fixed-LAVA (4.7%) elements included in our analysis demonstrated Tajima's D values that were more negative than any data point simulated under neutrality ($p<0.0001$). Among the fixed-LAVA, the most noteworthy were a ~300Kb cluster of 4 LAVA insertions on chr18 which overlapped a major dip in Tajima's D and displayed some of the lowest Tajima's D values in the whole genome (Fig. S4, Table S5).

While signatures of selection were identified at individual LAVA elements, a more striking pattern was observed when fixed- and poly-LAVA were examined collectively. The average of the averages Tajima's D for all poly-LAVA was -1.138, which was not significantly different from random collections of matching genomic loci (mean= -1.156, 1% and 99% percentile= -1.120 and -1.109, respectively; two-sided permutation $p=0.813$; Supplemental Text). In contrast however, the average Tajima's D around all fixed LAVA elements was -1.247, which was more negative than any randomly selected set of similar regions (mean= -1.153, 1% and 99% percentile= -1.191 and -1.106, respectively; two-sided permutation $p<0.0001$; Supplemental Text). Furthermore, fixed-LAVA that overlapped enhancer chromatin states had significantly lower Tajima's D than randomly selected fixed-LAVA, further supporting functionality of putative LAVA enhancers (two-sided permutation $p=0.02$; Supplemental Text).

Fixed LAVA insertions are enriched in DNA repair pathways

TEs can affect the regulation of nearby genes and even rewire entire functional gene networks (*15*). Since both fixed- and poly-LAVA were strongly associated with gibbon genes, we examined overrepresentation of gene ontology (GO) terms in genes within 3Kb of LAVA elements. For these analysis we included all LAVA insertions, regardless of their overlap with enhancer chromatin states

because LAVA might influence nearby gene expression via other mechanisms, also the absence of enhancer chromatin states could be due to low mappability of LAVA or tissue- and time-specificity of LAVA enhancers. We analyzed genes adjacent to fixed- and poly-LAVA separately, as these LAVA insertions appear to have experienced different evolutionary trajectories. While we found no significant GO term enrichment for genes nearby poly-LAVA, genes nearby fixed-LAVA insertions displayed significant enrichment ($q<0.1$, two-tail permutation $p<0.001$) for 15 biological functions related to DNA repair (e.g. base-excision repair, double- and single-strand break repair, and DNA synthesis in DNA repair) and 5 cellular components important in cell cycle (e.g. pericentriolar material and spindle-pole centrosome; Fig. 6A, Table S6). Of the fixed-LAVA associated with significant GO terms, 89% (51 out of 57) were also homozygous in all other individuals examined in this study from other gibbon genera (Table S3), suggesting their insertion and fixation predate the genera split and are shared among all gibbons. Among the 25 fixed-LAVA inserted near DNA repair genes, the three near *PDS5A, RAD9A* and *SETD2* overlapped enhancer chromatin states, with the latter two also having some of the most significantly negative Tajima's D values detected around fixed-LAVA elements (Tajima's D= -1.932 and -1.913 respectively, and $p<0.001$; Table S6). Furthermore, the mean Tajima's D for all fixed-LAVA associated with DNA repair genes was -1.371, which was significantly smaller than the null distribution of fixed-LAVA located near genes (two-tailed permutation $p=0.02$, Fig. 6B; Supplemental Text).

Notably, the protein products of several of the DNA repair genes associated with fixed-LAVA (Table S6) are shown to interact closely with each other within DNA repair networks. For example, MSH2 and MSH6 dimerize to form a complex that detects mismatched DNA bases (*31*), and SIRT1 directly deacteylates WRN to initiate WRN-mediated cellular DNA repair response (*32*). In general, many of the DNA repair genes found near fixed-LAVA insertions are implicated in either activation of DNA repair processes [e.g. *SETD2* in DNA double-break repair (*33*)], or cell senescence in response to DNA damage [e.g. cell cycle arrest by *RAD9A* in response to DNA damage, (*34*)], which are both crucial in maintaining genome integrity. Therefore, co-option of LAVA near these genes may represent an adaptive mechanism to improve genome integrity in the face of the many genomic breakages and rearrangements that occurred during gibbon evolution.

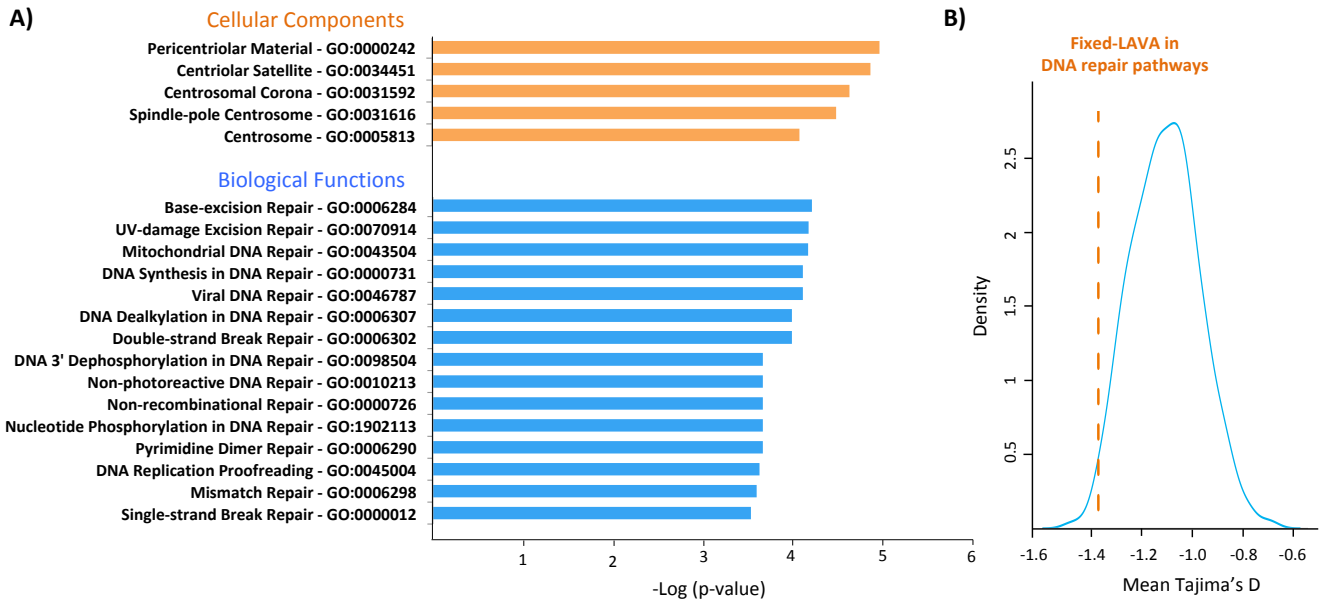


Figure 6. Fixed-LAVA insertions are enriched near DNA repair genes. A) Significant gene ontology terms associated with fixed-LAVA are demonstrated. **B)** Mean Tajima's D of fixed-LAVA in DNA repair pathways (orange dashed line) is show against the null distribution for other genic fixed-LAVA (blue curve).

Discussion

In this study we characterized genome-wide insertions of the gibbon-specific LAVA retrotransposon across gibbon genera (Figs. 1A-B). We found compelling genetic, epigenetic, and evolutionary evidence suggesting that LAVA was co-opted as a *cis*-regulatory element and may have modulated regulation of DNA repair gene networks, likely in response to genomic rearrangements during gibbon evolution. This is the first study to include whole-genome sequences from several (>20) individuals of these endangered small apes, paralleling previous efforts for the great apes (*35*).

Nearly half of all LAVA insertions characterized in this study appeared to be genus-specific and LAVA inserts were highly variable across gibbon genera (Fig. 1C). The smallest numbers of LAVA insertions were found in the *Nomascus* genus and the highest numbers were found in the *Hoolock,* the same genus in which LAVA was first discovered and found to make up long centromeric expansions (*12*, *36*). These findings reveal that the LAVA element has undergone genus-specific evolutionary trajectories following the split of the four genera ~5 mya (*11*). Indeed, unsupervised hierarchical clustering of individuals based on LAVA insertion genotypes sorted gibbons by genera, in a pattern resembling a potential LAVA-based gibbon phylogeny (Fig. 1D). While this tree did recapitulate some of the published phylogenies obtained from gibbon mitochondrial DNA (*37*, *38*), it does not match more recent phylogenies obtained by us and others based on nuclear genomes (*39*, *40*). Therefore, we believe that our tree should be interpreted merely as a validation of our LAVA identification process, rather than a true gibbon phylogeny, which still remains elusive.

We focused our investigation of LAVA co-option on the northern white-cheeked gibbon (*Nomascus leucogenys*, NLE), as this is currently the only species with a reference genome (*11*). We found that fixed and polymorphic LAVA (poly-LAVA) insertions had highly similar genomic distributions, with enrichment near genes (Fig. 2B). Moreover, around 10% of both poly- and fixed-LAVA insertions overlapped enhancer-related chromatin states (Fig. 3). This proportion is likely to be an underestimation, due to the tissue- and time- specific activity of TE-derived enhancers (*4*) and the low mappability of short-read sequencing data for young repetitive elements like LAVA (*41*). In line with the epigenetic evidence for LAVA's putative enhancer function, we also found enrichment of multiple transcription factor (TF) motifs

in LAVA sequences (Tables 1 and S3). Among these TFs, we validated binding of LAVA by PU.1, which is an important and highly expressed TF in lymphocytes (Fig. 4A-B). Interestingly, although clusters of PU.1 motifs were identified in the SVA subunit of LAVA, we did not detect widespread PU.1 binding to SVA in human (Fig. 4B). Hence, LAVA appears to have evolved some unique genetic, epigenetic and functional features in the gibbon lineage since deriving from the SVA element about 17 mya (*13*, *14*).

Consistent with LAVA's putative enhancer activity, genes near LAVA, especially fixed-LAVA elements, collectively had higher mean expression compared to genes randomly selected from the rest of the genome. Furthermore, we identified two significant correlations between a poly-LAVA genotype and expression of genes in *cis*, and notably, in both cases, the presence of LAVA was associated with higher gene expression (Fig. 5). These findings suggested that putative regulatory LAVA elements generally increase expression of nearby genes, perhaps by providing binding sites for activating TFs. This is in line with our observation that LAVA is bound by PU.1, which is a TF shown to activate gene transcription when clusters of its binding motifs are available nearby a target gene (*42*). It should be noted however, that the nature of our gene expression analyses are correlative and do not infer causality. Therefore, higher expression of genes near LAVA may, at least partially, be caused by other factors such as preferential insertion of LAVA near highly active genes. Future genome-editing studies that directly manipulate LAVA insertions can assess LAVA's functional role in the regulation of adjacent genes.

Biochemical activity of TEs might merely reflect their selfish strategies for survival and propagation (*18*). Thus, to demonstrate that LAVA was co-opted, we investigated signals of selection at insertion sites. Remarkably, we found strong and significant signatures of positive selection collectively around fixed-LAVA, indicating that natural selection has favored many of these insertions and likely contributed to their fixation in the gibbon genome. Of note, among fixed-LAVA, those with putative enhancer function showed even stronger signatures of selection, further supporting their functionality and co-option. Unlike fixed-LAVA, significant positive selection was not detected collectively at poly-LAVA insertions, suggesting that most of these insertions are either not functional or are too young to show signatures of selection via our methodology. Another important implication of this observation is that the selection

signature detected around fixed-LAVA insertions is not merely a result of background selection due to their vicinity to genes, as poly- and fixed-LAVAs are equally enriched near genes (Fig. 2B). Together, these findings indicated that many fixed-LAVA insertions, especially those with putative enhancer function, have been adaptive and favored by natural selection. This may explain why LAVA has thrived in the gibbon lineage, in spite of its capability to disrupt gene transcription (*11*).

By investigating genes associated with fixed- and poly-LAVA, we recapitulated LAVA's previously described (*11*) association with cell cycle and chromosome segregation genes (i.e. enrichment of cellular components of cell division; Table S6, Fig. 6A). However, by characterizing new LAVA insertions across multiple individuals and classifying them based on frequency in the population, we were also able to unravel a novel association between LAVA and genes implicated in DNA repair pathways. Several lines of evidence corroborated LAVA's role in DNA repair pathways. First, fixed- but not poly-LAVA insertions were enriched near DNA repair genes (Fig. 6A). Notably, most of these insertions also appeared to be fixed in the other three gibbon genera, suggesting that these insertions occurred and became fixed in the common ancestor (Table S6). Second, we found a significantly stronger signature of positive selection at fixed-LAVA insertions adjacent to DNA repair genes, compared to fixed-LAVA located near other genes (Fig. 6B), supporting LAVA's adaptive role in these pathways. Third, some of the TFs predicted to bind the LAVA element, such as STAT3 and ZNF143, have been widely implicated in regulation of DNA repair networks in response to DNA damage (*43*, *44*). Thus, LAVA insertions may have increased transcription of DNA repair genes by providing new binding sites for these TFs, and subsequently promoted stability and integrity of the genome (*45*).

By integrating genetic, epigenetic and evolutionary data from the largest gibbon datasets available to date, we provide strong evidence that some LAVA insertions were co-opted in the gibbon genome to enhance DNA repair and improve genome integrity, likely as a way to mitigate the high number of genomic rearrangements experienced by all genera in this lineage. Like most studies focused on repetitive sequences, our analyses were challenged by a few technical limitations, such as low mappability and sequence ambiguity. In addition, studying an endangered species limited our access to a larger sample size and multiple tissues. As computational tools for studying TEs improve (*46*), and as

pluripotent stem cells (iPSC) from gibbon provide access to currently unavailable tissues, we should be able to surmount these roadblocks. Insights from this study, and future studies focused on lineage-specific TEs, will advance our understanding of how these elements contribute to evolutionary novelty and lineage-specific adaptations.

Materials and Methods

Genome-wide identification and genotyping of LAVA insertions/deletions

Genomic DNA extracted from blood of 23 unrelated gibbons across the four extant genera (*Nomascus*= 13, *Hylobates*= 5, *Hoolock*= 3, *Siamang*= 2) was used to construct whole genome sequencing (WGS) libraries as described before (*11*). All gibbon WGS data were aligned to the gibbon genome reference (Nleu3.0) using BWA (*47*). MELT v2.1.3 (*16*) was used to predict *de novo* LAVA insertions and deletions (indels) from each of the 23 WGS alignments similar to our simulation analyses (see Supplemental Text). We characterized all LAVA inserts based on information such as genomic location, length, and insertion position relative to closest gene [intergenic, exonic, intronic, promoter (<3Kb upstream of gene) or terminator (<3Kb downstream of gene)] (Table S2). The initial set of LAVA predictions was next filtered to remove: 1) low quality inserts, as assessed by MELT, 2) single copy inserts in the population, 3) inserts shorter than 290bp (the minimum length required to discriminate a composite LAVA from its non-SVA subunits), and 4) inserts found on unplaced Nleu3.0 contigs. We also generated binary LAVA genotype profiles for all gibbon individuals (heterozygous or homozygous LAVA insertion=1, homozygous absence of LAVA= 0) and performed hierarchical clustering using the *hclust* function with the ward D2 method in R and visualized the results with *heatmap.2* function in the *ggplot2* package. Logistic principal component analysis was carried out using the *logisticPCA* package in R (k=2 and m=4).

Characterization of the genomic context of LAVA insertion sites

All downstream analyses were performed on LAVA insertions identified in the 11 *Nomascus leucogenys* (NLE) gibbons. LAVA insertions found in two copies in all NLE individuals were called fixed-

LAVA, while the rest were called polymorphic LAVA (poly-LAVA). Due to sequence ambiguity and absence of many LAVA insertions from the reference gibbon genome, LAVA sequence polymorphism was not considered in our characterizations. To test whether the number of observed fixed- and poly-LAVA insertions per NLE chromosome deviated significantly from random distribution based on ungapped chromosome length, we performed a two-tailed chi-squared test using the Graphpad tool (https://www.graphpad.com/quickcalcs/contingency1.cfm). To test whether LAVA insertions were located closer to each other than expected by random chance (i.e. clustering), we used a permutation approach similar to (*48*) (Supplemental Text). Permutation p-values were corrected for multiple testing using the Benjamini-Hochberg procedure (*49*) in R. Next, we used the TEanalysis tool (available at (*51*)) with 1,000 permutations to test significant over/under-representation of repeats within 1Kb of fixed and polymorphic LAVA elements. Lastly, we used custom R scripts to perform linear regression between LAVA and gene density across chromosomes [density=(count in chromosome)/(Mb ungapped chromosome length)].

Chromatin immunoprecipitation sequencing (ChIP-seq) and characterization of chromatin states

ChIP-seq, library preparation and sequencing were performed on gibbon EBV transformed lymphoblastoid cell lines (LCLs) from three NLE individuals, as previously described (*51*) and outlined in the Supplemental Text. Raw reads were QC'd with FastQC (*51*) and since all libraries displayed high quality and complexity, trimming was not performed. All reads were aligned to Nleu3.0 using BWA (*47*), and low-quality and multi-mapping read alignments (MAPQ<30) were removed. Histone ChIP qualities were assessed by examining the Pearson correlation of ChIP-seq signal across biological replicates and histone marks (Fig. S3). Next, ChromHMM tool (*21*) was used to identify and characterize 9 chromatin states based on the histone ChIP-seq alignments. Intersection of LAVA insertions with chromatin states was performed using BEDtools (*52*).

Transcription factor motif enrichment and PU.1 binding to LAVA

We extracted sequences of all 1,118 LAVA insertions present in the assembled chromosomes of Nleu3.0. Motif enrichment analyses were performed once with the Homer suite (*53*), and once using the

Transcription factor Affinity Prediction [TRAP; (*54*)] web tool with jasper_vertebrates matrix and human promoters background. To meet TRAPs length restrictions, we removed 24 (of the total 1,118) LAVA sequences that were longer than 3Kb. P-values from both approaches were corrected using the Benjamini-Hochberg method (*49*). To reduce false positives, only significant TF motif enrichments ($q<0.05$) that agreed between the two methods were considered (Table S3). PU.1 (encoded by *Spi1*) was selected for ChIP-seq validation on two biological replicates of gibbon LCLs as described in the Supplemental Text. Findings in gibbon LCL were compared to public human LCL PU.1 ChIP-seq data from ENCODE [GEO accessions GSM803531 and GSM803398; (*27*, *28*, *28*)]

Characterization of gene expression patterns near LAVA

RNA-seq gene count data was collected from 9 available NLE LCLs and normalized as described in Supplemental Text. The linear regression model from Matrix-eQTL (*55*) was used to test association of binary genotypes at poly-LAVA with expression of genes within 1Mb. Next, we used the GraphPad tool (https://www.graphpad.com/quickcalcs/contingency1.cfm) to perform a two-tailed chi-square with Yates correction and compare the proportion of "active genes" (genes for which at least two individuals have >0.5 counts per million, CPM) nearby fixed- or poly- LAVA to the rest of the genome. Lastly, we used custom R scripts to perform permutation tests and compare mean and variability (coefficient of variation) of gene expression near fixed- and poly-LAVA to randomly selected genes in the rest of the genome (Supplemental Text).

Assessing selection around LAVA insertion sites

We used ANGSD (*56*) to estimate folded allele frequency spectra and Tajima's D in 10 NLE gibbons (excluding the individual used to construct the reference gibbon genome). LAVA elements were filtered based on mean coverage in the WGS datasets to avoid the potential affect of copy number variation. The X chromosome and the LAVA elements found on it were also excluded from downstream analysis. We used a diffusion approximation approach via δaδi (*57*) to fit a 2-epoch model with a population expansion in the recent past to a set of putatively neutral loci. Based on this model, for each

LAVA element we performed 10,000 coalescent simulations via ms (*58*) to compare the observed Tajima's D in 20Kb windows centered at individual LAVA insert sites to the expected distribution under neutrality and generate a p-value (Supplemental Text). To examine signals of selection collectively at poly- and fixed-LAVA, we calculated the average Tajima's D of 10Kb windows upstream and downstream flanking each LAVA element, and averaged those values once across all fixed-LAVA and once across poly-LAVA elements. We then compared the average of averages Tajima's D for fixed- and poly-LAVA to 10,000 sets of randomly sampled regions from across the genome that matched various properties of the LAVA elements and generated an empirical p-value (Supplemental Text).

Gene ontology (GO) analysis of genes nearby fixed- and poly-LAVA

To test over-enrichment of gene ontology pathways potentially affected by LAVA, we used Enrichr (*59*) with GO_Biological_Process_2017_7b, GO_Cellular_Component_2017_7b and GO_Molecular_Function_2017_7b libraries for the nearest genes within $\leq$ 3Kb of all polymorphic and fixed LAVA elements. Significance of all GO terms that had $p<0.05$ and $q<0.1$ were validated using two different *post-hoc* permutation tests, to ensure that our GO enrichments were not biased by the gibbon gene annotations (see Supplemental Text).

Lastly, we used permutation analysis to compare mean Tajima's D around fixed LAVA near DNA repair genes, to the null distribution of mean Tajima's D at other genic fixed-LAVA (Supplemental Text).

Acknowledgements

Authors would like to express their gratitude to the zoos (San Antonio Zoo and Aquarium, Point Defiance Zoo and Aquarium, Oregon Zoo, Gladys Porter Zoo and Los Angeles Zoo) and the staff at the Gibbon Conservation Center (Santa Clarita, CA), especially the director, Ms. Gabriella Skollar, who have provided us with opportunistic gibbon samples. The authors would also like to thank Drs. Jeff Wall and Michael Hammer for their invaluable contributions to WGS data collection, Dr. Eugene Gardner for providing assistance in optimizing the MELT pipeline, Dr. Jessica Minnier for guidance in RNA-seq analysis, Dr. R. Alan Harris, Patty Langasek and Christopher Klocke for their help with data analysis, and

members of the Carbone and Chavez lab for valuable feedback on the research. Authors would like to acknowledge the ENCODE Consortium and Dr. Myers at HudsonAlpha Institute for Biotechnology, who generated the human PU.1 ChIP-seq data. Gibbon PU.1 ChIP-seq assays were performed by the Epigenetics Consortium at Knight Cardiovascular Institute of Oregon Health and Science University (OHSU). All ChIP-seq libraries were sequenced at the OHSU Massively Parallel Sequencing Shared Resource (MPSSR) and the Genomics and Cell Characterization Core Facility (GC3F) at University of Oregon. High-throughput data analyses were performed on the Exacloud super computer cluster at OHSU. This work was financially supported by a grant awarded to L.C. from the Leakey foundation. L.C. and R.O.N are also supported by the National Science Foundation (1613856), L.C. and N.A. are supported by the National Human Genome Research Institute (R01HG010333) and L.C. is supported by the NIH/OD P51 OD011092 to the Oregon National Primate Research Center. Authors declare no conflict of interest.

References

1. M. K. Konkel, M. A. Batzer, A mobile threat to genome stability: The impact of non-LTR retrotransposons upon the human genome. *Semin. Cancer Biol.* **20**, 211–221 (2010).

2. L. Schrader, J. Schmitz, The impact of transposable elements in adaptive evolution. *Molecular Ecology*. **28**, 1537–1549 (2019).

3. G. Bourque, Transposable elements in gene regulation and in the evolution of vertebrate genomes. *Current Opinion in Genetics & Development*. **19**, 607–612 (2009).

4. M. Trizzino, A. Kapusta, C. D. Brown, Transposable elements generate regulatory novelty in a tissue-specific fashion. *BMC Genomics*. **19**, 468 (2018).

5. I. A. Warren, M. Naville, D. Chalopin, P. Levin, C. S. Berger, D. Galiana, J.-N. Volff, Evolutionary impact of transposable elements on genomic diversity and lineage-specific innovation in vertebrates. *Chromosome Res.* **23**, 505–531 (2015).

6. M. Trizzino, Y. Park, M. Holsbach-Beltrame, K. Aracena, K. Mika, M. Caliskan, G. H. Perry, V. J. Lynch, C. D. Brown, Transposable elements are the primary source of novelty in primate gene regulation. *Genome Res*. **27**, 1623–1633 (2017).

7. D. Blanco-Melo, R. J. Gifford, P. D. Bieniasz, Co-option of an endogenous retrovirus envelope for host defense in hominid ancestors. *eLife*. **6**, e22519 (2017).

8. A. F. Gombart, T. Saito, H. P. Koeffler, Exaptation of an ancient Alu short interspersed element provides a highly conserved vitamin D-mediated innate immune response in humans and primates. *BMC Genomics*. **10**, 321 (2009).

9. C. Cunningham, A. Mootnick, Gibbons. *Current Biology*. **19**, R543–R544 (2009).

10. L. Carbone, G. M. Vessere, B. F. H. ten Hallers, B. Zhu, K. Osoegawa, A. Mootnick, A. Kofler, J. Wienberg, J. Rogers, S. Humphray, C. Scott, R. A. Harris, A. Milosavljevic, P. J. de Jong, A High-Resolution Map of Synteny Disruptions in Gibbon and Human Genomes. *PLOS Genetics*. **2**, e223 (2006).

11. L. Carbone, R. Alan Harris, S. Gnerre, K. R. Veeramah, B. Lorente-Galdos, J. Huddleston, T. J. Meyer, J. Herrero, C. Roos, B. Aken, F. Anaclerio, N. Archidiacono, C. Baker, D. Barrell, M. A. Batzer, K. Beal, A. Blancher, C. L. Bohrson, M. Brameier, M. S. Campbell, O. Capozzi, C. Casola, G. Chiatante, A. Cree, A. Damert, P. J. de Jong, L. Dumas, M. Fernandez-Callejo, P. Flicek, N. V. Fuchs, I. Gut, M. Gut, M. W. Hahn, J. Hernandez-Rodriguez, L. W. Hillier, R. Hubley, B. Ianc, Z. Izsvák, N. G. Jablonski, L. M. Johnstone, A. Karimpour-Fard, M. K. Konkel, D. Kostka, N. H. Lazar, S. L. Lee, L. R. Lewis, Y. Liu, D. P. Locke, S. Mallick, F. L. Mendez, M. Muffato, L. V. Nazareth, K. A. Nevonen, M. O'Bleness, C. Ochis, D. T. Odom, K. S. Pollard, J. Quilez, D. Reich, M. Rocchi, G. G. Schumann, S. Searle, J. M. Sikela, G. Skollar, A. Smit, K. Sonmez, B. ten Hallers, E. Terhune, G. W. C. Thomas, B. Ullmer, M. Ventura, J. A. Walker, J. D. Wall, L. Walter, M. C. Ward, S. J. Wheelan, C. W. Whelan, S. White, L. J. Wilhelm, A. E. Woerner, M. Yandell, B. Zhu, M. F. Hammer, T. Marques-Bonet, E. E. Eichler, L. Fulton, C. Fronick, D. M. Muzny, W. C. Warren, K. C. Worley, J. Rogers, R. K. Wilson, R. A. Gibbs, Gibbon genome and the fast karyotype evolution of small apes. *Nature*. **513**, 195–201 (2014).

12. L. Carbone, R. A. Harris, A. R. Mootnick, A. Milosavljevic, D. I. K. Martin, M. Rocchi, O. Capozzi, N. Archidiacono, M. K. Konkel, J. A. Walker, M. A. Batzer, P. J. de Jong, Centromere remodeling in Hoolock leuconedys (Hylobatidae) by a new transposable element unique to the gibbons. *Genome Biol Evol*. **4**, 648–658 (2012).

13. B. Ianc, C. Ochis, R. Persch, O. Popescu, A. Damert, Hominoid Composite Non-LTR Retrotransposons—Variety, Assembly, Evolution, and Structural Determinants of Mobilization. *Mol Biol Evol*. **31**, 2847–2864 (2014).

14. T. J. Meyer, U. Held, K. A. Nevonen, S. Klawitter, T. Pirzer, L. Carbone, G. G. Schumann, The Flow of the Gibbon LAVA Element Is Facilitated by the LINE-1 Retrotransposition Machinery. *Genome Biol Evol*. **8**, 3209–3225 (2016).

15. E. B. Chuong, N. C. Elde, C. Feschotte, Regulatory activities of transposable elements: from conflicts to benefits. *Nat Rev Genet*. **18**, 71–86 (2017).

16. E. J. Gardner, V. K. Lam, D. N. Harris, N. T. Chuang, E. C. Scott, W. S. Pittard, R. E. Mills, S. E. Devine, The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome Res*. **27**, 1916–1929 (2017).

17. T. Sultana, A. Zamborlini, G. Cristofari, P. Lesage, Integration site selection by retroviruses and transposable elements in eukaryotes. *Nature Reviews Genetics*. **18**, 292–308 (2017).

18. G. Bourque, K. H. Burns, M. Gehring, V. Gorbunova, A. Seluanov, M. Hammell, M. Imbeault, Z. Izsvák, H. L. Levin, T. S. Macfarlan, D. L. Mager, C. Feschotte, Ten things you should know about transposable elements. *Genome Biology*. **19**, 199 (2018).

19. C. Gao, M. Xiao, X. Ren, A. Hayward, J. Yin, L. Wu, D. Fu, J. Li, Characterization and functional annotation of nested transposable elements in eukaryotic genomes. *Genomics*. **100**, 222–230 (2012).

20. D. Venuto, G. Bourque, Identifying co-opted transposable elements using comparative epigenomics. *Development, Growth & Differentiation*. **60**, 53–62 (2018).

21. J. Ernst, M. Kellis, Chromatin-state discovery and genome annotation with ChromHMM. *Nature Protocols*. **12**, 2478–2492 (2017).

22. G. Bourque, B. Leong, V. B. Vega, X. Chen, Y. L. Lee, K. G. Srinivasan, J.-L. Chew, Y. Ruan, C.-L. Wei, H. H. Ng, E. T. Liu, Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res*. **18**, 1752–1762 (2008).

23. T. Manke, H. G. Roider, M. Vingron, Statistical Modeling of Transcription Factor Binding Affinities Predicts Regulatory Interactions. *PLoS Comput Biol*. **4**, e1000039, (2008).

24. P. Rimmelé, J. Komatsu, P. Hupé, C. Roulin, E. Barillot, M. Dutreix, E. Conseiller, A. Bensimon, F. Moreau-Gachelin, C. Guillouf, Spi-1/PU.1 Oncogene Accelerates DNA Replication Fork Elongation and Promotes Genetic Instability in the Absence of DNA Breakage. *Cancer Res*. **70**, 6757–6766 (2010).

25. H. Wang, J. Xing, D. Grover, D. J. Hedges, K. Han, J. A. Walker, M. A. Batzer, SVA Elements: A Hominid-specific Retroposon Family. *Journal of Molecular Biology*. **354**, 994–1007 (2005).

26. A. Damert, Phylogenomic analysis reveals splicing as a mechanism of parallel evolution of non-canonical SVAs in hominine primates. *Mobile DNA*. **9**, 30 (2018).

27. ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature*. **489**, 57–74 (2012).

28. C. A. Davis, B. C. Hitz, C. A. Sloan, E. T. Chan, J. M. Davidson, I. Gabdank, J. A. Hilton, K. Jain, U. K. Baymuradov, A. K. Narayanan, K. C. Onate, K. Graham, S. R. Miyasato, T. R. Dreszer, J. S. Strattan, O. Jolanki, F. Y. Tanaka, J. M. Cherry, The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* **46**, D794–D801 (2018).

29. O. Gianfrancesco, V. J. Bubb, J. P. Quinn, SVA retrotransposons as potential modulators of neuropeptide gene expression. *Neuropeptides*. **64**, 3–7 (2017).

30. F. Tajima, Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*. **123**, 585–595 (1989).

31. T. Jascur, C. R. Boland, Structure and function of the components of the human DNA mismatch repair system. *International Journal of Cancer*. **119**, 2030–2035 (2006).

32. K. Li, A. Casta, R. Wang, E. Lozada, W. Fan, S. Kane, Q. Ge, W. Gu, D. Orren, J. Luo, Regulation of WRN protein cellular localization and enzymatic activities by SIRT1-mediated deacetylation. *J. Biol. Chem.* **283**, 7590–7598 (2008).

33. S. Carvalho, A. C. Vítor, S. C. Sridhara, F. B. Martins, A. C. Raposo, J. M. Desterro, J. Ferreira, S. F. de Almeida, SETD2 is required for DNA double-strand break repair and activation of the p53-mediated checkpoint. *eLife*. **3**, e02482 (2014).

34. H. B. Lieberman, Rad9, an evolutionarily conserved gene with multiple functions for preserving genomic Integrity. *Journal of Cellular Biochemistry*. **97**, 690–697 (2006).

35. J. Prado-Martinez, P. H. Sudmant, J. M. Kidd, H. Li, J. L. Kelley, B. Lorente-Galdos, K. R. Veeramah, A. E. Woerner, T. D. O'Connor, G. Santpere, A. Cagan, C. Theunert, F. Casals, H. Laayouni, K. Munch, A. Hobolth, A. E. Halager, M. Malig, J. Hernandez-Rodriguez, I. Hernando-Herraez, K. Prüfer, M. Pybus, L. Johnstone, M. Lachmann, C. Alkan, D. Twigg, N. Petit, C. Baker, F. Hormozdiari, M. Fernandez-Callejo, M. Dabad, M. L. Wilson, L. Stevison, C. Camprubí, T. Carvalho, A. Ruiz-Herrera, L. Vives, M. Mele, T. Abello, I. Kondova, R. E. Bontrop, A. Pusey, F. Lankester, J. A. Kiyang, R. A. Bergl, E. Lonsdorf, S. Myers, M. Ventura, P. Gagneux, D. Comas, H. Siegismund, J. Blanc, L. Agueda-Calpena, M. Gut, L. Fulton, S. A. Tishkoff, J. C. Mullikin, R. K. Wilson, I. G. Gut, M. K. Gonder, O. A. Ryder, B. H. Hahn, A. Navarro, J. M. Akey, J. Bertranpetit, D. Reich, T. Mailund, M. H. Schierup, C. Hvilsom, A. M. Andrés, J. D. Wall, C. D. Bustamante, M. F. Hammer, E. E. Eichler, T. Marques-Bonet, Great ape genetic diversity and population history. *Nature*. **499**, 471–475 (2013).

36. T. Hara, Y. Hirai, I. Jahan, H. Hirai, A. Koga, Tandem repeat sequences evolutionarily related to SVA-type retrotransposons are expanded in the centromere region of the western hoolock gibbon, a small ape. *J. Hum. Genet.* **57**, 760–765 (2012).

37. Y.-C. Chan, C. Roos, M. Inoue-Murayama, E. Inoue, C.-C. Shih, K. J.-C. Pei, L. Vigilant, Mitochondrial Genome Sequences Effectively Reveal the Phylogeny of Hylobates Gibbons. *PLOS ONE*. **5**, e14419 (2010).

38. K. Matsudaira, T. Ishida, Phylogenetic relationships and divergence dates of the whole mitochondrial genome sequences among three gibbon genera. *Molecular Phylogenetics and Evolution*. **55**, 454–459 (2010).

39. K. R. Veeramah, A. E. Woerner, L. Johnstone, I. Gut, M. Gut, T. Marques-Bonet, L. Carbone, J. D. Wall, M. F. Hammer, Examining phylogenetic relationships among gibbon genera using whole

genome sequence data using an approximate bayesian computation approach. *Genetics*. **200**, 295–308 (2015).

40. C.-M. Shi, Z. Yang, Coalescent-Based Analyses of Genomic Sequence Data Provide a Robust Resolution of Phylogenetic Relationships among Major Groups of Gibbons. *Mol. Biol. Evol.* **35**, 159–179 (2018).

41. A. D. Ewing, Transposable element detection from whole genome sequence data. *Mob DNA*. **6**, 24 (2015).

42. M. Ridinger-Saison, V. Boeva, P. Rimmelé, I. Kulakovskiy, I. Gallais, B. Levavasseur, C. Paccard, P. Legoix-Né, F. Morlé, A. Nicolas, P. Hupé, E. Barillot, F. Moreau-Gachelin, C. Guillouf, Spi-1/PU.1 activates transcription through clustered DNA occupancy in erythroleukemia. *Nucleic Acids Res*. **40**, 8927–8941 (2012).

43. S. P. Barry, P. A. Townsend, R. A. Knight, T. M. Scarabelli, D. S. Latchman, A. Stephanou, STAT3 modulates the DNA damage response pathway. *Int J Exp Pathol*. **91**, 506–514 (2010).

44. H. Ishiguchi, H. Izumi, T. Torigoe, Y. Yoshida, H. Kubota, S. Tsuji, K. Kohno, ZNF143 activates gene expression in response to DNA damage and binds to cisplatin-modified DNA. *Int. J. Cancer*. **111**, 900–909 (2004).

45. L. A. Mathews, S. M. Cabarcas, E. M. Hurt, X. Zhang, E. M. Jaffee, W. L. Farrar, Increased expression of DNA repair genes in invasive human pancreatic cancer cells. *Pancreas*. **40**, 730–739 (2011).

46. P. Goerner-Potvin, G. Bourque, Computational tools to unmask transposable elements. *Nature Reviews Genetics*. **19**, 688-704 (2018).

47. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. **25**, 1754–1760 (2009).

48. D. Kostka, A. K. Holloway, K. S. Pollard, Developmental Loci Harbor Clusters of Accelerated Regions That Evolved Independently in Ape Lineages. *Mol Biol Evol*. **35**, 2034–2045 (2018).

49. Y. H. Y Benjamini, Controlling the false fiscovery rate: a practical and powerful approach to multiple testing. *J. Royal Statist. Soc., Series B*. **57**, 289–300 (1995).

50. N. H. Lazar, K. A. Nevonen, B. O'Connell, C. McCann, R. J. O'Neill, R. E. Green, T. J. Meyer, M. Okhovat, L. Carbone, Epigenetic maintenance of topological domains in the highly rearranged gibbon genome. *Genome Res.*, **28**, 983-997 (2018).

51. S. Andrews, FastQC: A quality control tool for high throughput sequence data. (2010) (available at http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

52. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. **26**, 841–842 (2010).

53. S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, C. K. Glass, Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell*. **38**, 576–589 (2010).

54. M. Thomas-Chollier, A. Hufton, M. Heinig, S. O'Keeffe, N. E. Masri, H. G. Roider, T. Manke, M. Vingron, Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data and regulatory SNPs. *Nature Protocols*. **6**, 1860–1869 (2011).

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

56. T. S. Korneliussen, A. Albrechtsen, R. Nielsen, ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics*. **15**, 356 (2014).

57. R. N. Gutenkunst, R. D. Hernandez-Rodriguez, S. H. Williamson, C. D. Bustamante, Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. PloS Genet. **5**, e1000695 (2009).

58. R. R. Hudson, Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics*. **18**, 337–338 (2002).

59. M. V. Kuleshov, M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann, M. G. McDermott, C. D. Monteiro, G. W. Gundersen, A. Ma'ayan, Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res*. **44**, W90–W97 (2016).

Supplementary Information for

Co-option of the gibbon-specific *LAVA* retrotransposon in DNA repair pathways

Mariam Okhovat[1*], Kimberly A. Nevonen[1], Brett Davis[1], Pryce Michener[1†], Samantha Ward[1], Mark Milhaven[2], Lana Harshman[3,4], Ajuni Sohota[3,4], Rachel J. O'Neill[5,6], Nadav Ahituv[3,4], Krishna R. Veeramah[2], Lucia Carbone[1,7-9*]

10. Department of Medicine, Knight Cardiovascular Institute, Oregon Health and Science University, Portland, OR 97239, USA
11. Department of Ecology and Evolution/ Institute for Advance Computational Science, Stony Brook University, Stony Brook, NY 11794, USA
12. Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA 94158, USA
13. Institute for Human Genetics, University of California San Francisco, San Francisco, CA 94158, USA
14. Institute for Systems Genomics, University of Connecticut, Storrs, CT 06269, USA
15. Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT 06269, USA
16. Department of Molecular and Medical Genetics, Oregon Health and Science University, Portland, OR 97239, USA
17. Division of Genetics, Oregon National Primate Research Center, Beaverton, OR 97006, USA
18. Department of Informatics and Clinical Epidemiology, Oregon Health and Science University, Portland, OR 97239, USA

† Current affiliation: University of Massachusetts Medical School, Worcester, MA 01605, USA

*Corresponding authors

Supplemental Methods

In silico validation of the MELT pipeline

To test the ability of MELT (*1*) to identify LAVA insertions and deletions (indels) in whole-genome sequencing (WGS) datasets from different gibbon individuals, we *in silico* introduced LAVA insertions and deletions (indels) in the current gibbon genome reference (Nleu3.0) to simulate 9 WGS datasets. Briefly, we removed all short (<1Mb) and unplaced contigs from Nleu3.0. Next, we generated *in silico* LAVA insertions by randomly inserting sequences of different full-length (1.5-2Kb) LAVA into Nelu3.0 using SVsim (available at https://github.com/GregoryFaust/SVsim). *In silico* deletions of LAVA elements previously annotated in Nleu 3.0 were generated with BEDtools (*2*) and custom bash scripts. Using this approach we generated three artificial genomes, each containing 100 different insertions and 20 different deletions (mock genomes 1, 2 and 3) and then used wgsim (available at https://github.com/lh3/wgsim) to simulate 100bp paired-end illumina reads (fragment length of 250bp, mutation rate=0.01 and indel fraction=0.01) from each of them (30X coverage,~421 million read pairs). Simulated reads were then aligned to Nleu3.0 using bwa (*3*) with default paired-end settings. After using SAMtools (*4*) to binarize and sort alignments, each alignment was down-sampled to 20X and 10X coverage using Picard Tools (available at http://broadinstitute.github.io/picard). We then used MELT v2.1.3 (*1*) with default settings to predict *de novo* LAVA insertion and deletion (indel) on each of the 9 simulated WGS datasets.

We first assessed the precision of MELT's position predictions by comparing predicted vs. true positions of the simulated LAVA indel loci. For both deletions and insertions, the 10X coverage simulation datasets had the lowest percentage of LAVAs predicted at the exact true genomic position (58.3 ± 4.7% of insertions, 10 ± 5% of deletions; mean ± stdev; Fig. S1A). This percentage improved by increasing the coverage to 20X (65.7 ± 5.9% of insertions, 13.3 ± 7.6% of deletions; mean ± stdev), but further improvements were minimal after increasing the coverage from 20X to 30X (66 ± 6.1% of insertions and 13.3 ± 7.6% of deletions; mean ± stdev; Fig. S1A). By allowing LAVA positions to be considered "correct" if they were predicted within 10bp of the true simulated indel site, the percentage of correct positions increased drastically across all coverages (insertions at 10, 20 and 30X: 92 ± 1.7%, 96 + 0.0% and 96.7 ± 0.6% (mean ± stdev) and deletions at 10, 20 and 30X: 70 ± 13.2%, 81.7± 7.6% and 81.7± 7.6% (mean ± stdev)). However, the effect of coverage remained broadly the same, with noticeable improvement obtained by increasing the coverage from 10X to 20X, but not from 20X to 30X (Fig. S1A). Interestingly, increasing the margin of error for predicted position to -/+ 100bp, or -/+1Kb did not drastically improve the percentage of correctly predicted LAVA indels (Fig. S1A), suggesting that MELT detected most LAVA within 10bp from their true indel site.

In order to formally measure specificity (true negative rate) and sensitivity (true positive rate) of our MELT pipeline in identifying LAVAs across the three WGS coverages (10X, 20X and 30X), we allowed for a 10bp margin of error between the predicted genomic position and the true LAVA indel site and calculated false positive, false negative, and true positive LAVA predictions. Based on these calculations, we observed that the average specificity for predicting LAVA insertions decreased slightly with increase in coverage, but remained high overall (>98%, Fig. S1B, left). Average specificity of prediction LAVA deletions remained 100% across all coverages (Fig. S1B, right). MELT's average sensitivity in predicting LAVA insertions was high across coverages and increased with higher coverage (>92%, Fig. S1B, left). Similarly, average sensitivity for detecting LAVA deletions increased from an average 70% in 10X WGS datasets to an average of 81.7% in 20 and 30X coverage simulations (Fig. S1B, right).

Assessing clustering of LAVAs on chromosomes

We used an approach similar to Kostka et al. (*5*), to determine whether fixed and polymorphic LAVA insert sites appeared in clusters along chromosomes. Briefly, we computed the median of distances between each LAVA insertion and its next nearest LAVA insert site. Next, we used BEDtools shuffle (*2*) to randomly distribute the LAVA insertions within their original chromosome 1,000 times, while avoiding assembly gaps. After each random shuffle, we calculated the median of the shortest distances between LAVA insertion sites. The proportion of the shuffled median distances that was smaller or equal to the true median value was our empirical p-value.

Establishment of gibbon EBV-transformed lymphoblastoid cell lines

Whole blood from gibbons was collected opportunistically in sodium heparin tubes during routine check-ups at Zoos or the Gibbon Conservation Center (Santa Clarita, CA). We isolated lymphocytes from the blood using Ficoll-Paque PLUS (GE Healthcare). Next, we transformed 3-9x10^6 lymphocytes with Epstein Barr Virus from the marmoset cell line B95-8 (ATCC CRL-1612), using a standard protocol. We incubated the cells with EBV for 2 hours at 37°C then diluted with RPMI-1640 (Corning cellgro) supplemented with 10% FBS (Hyclone), 1X MEM Non-essential Amino Acids Solution (Corning cellgro), 1mM Sodium pyruvate (Corning cellgro) 1% Pen-Strep (Corning cellgro) and 2mM L-glutamine (Hyclone). Lastly, we allowed the cells to grow undisturbed for 10-12 days. Once signs of transformation were observed, we fed the cells with the same supplemented RPMI-1640.

Chromatin immunoprecipitation (ChIP) and library preparation

We fixed 5x10^6 cells per ChIP assay, with 1% formaldehyde for 5min on ice and then quenched fixation by adding glycine to a final concentration of 0.1M. After washing the cells twice with cold 1X PBS, we lysed the fixed cells at a concentration of 3x10^6cells/100ul in Lysis buffer (0.1% SDS, 0.5% Triton X-100, 20mM Tris-HCl pH=8.0, 150mM NaCl, 1x Proteinase inhibitor (Roche)) for 5min on ice. Using the Bioruptor Pico sonicator (Diagenode), we sheared lysates in 1.5mL tubes with 7 cycles (30sec on/off). We spun the sonicated lysates at max speed for 10min at 4°C to remove cell debris and then rotated them for 2hr at 4°C with 20ul ChIP-grade Protein A/G Magnetic Beads (Pierce) to preclear the lysate.

From each chromatin preparation, we set aside a 1% volume aliquot as chromatin input, and the rest we divided equally into each histone mark ChIP reactions. We diluted the reactions to $1.6x10^6$cells/100ul with Lysis buffer and then added the corresponding antibodies in the following amounts: 2ug H3K4me1 (ab8895, Abcam), 1ug H3K4me3 (ab8580, Abcam), 1ug H3K27ac (ab4729, Abcam), 2ug H3K27me3 (39155, Active Motif) and 2ug H3K9me3 (ab9263, Abcam), and let reactions rotate overnight at 4°C. Next, we added 20ul Pierce ChIP-grade Protein A/G Magnetic Beads (Pierce) and let samples rotate for 2hrs at 4°C. We then washed the beads consecutively (rotating at 4°C) with TBST buffer (3 times for 15min), Lysis buffer (once for 1hr), 1xTE pH=8 (once for 1hr, then once for 10min). Following washes, we eluted chromatin in 250ul fresh Elution buffer (1% SDS, 0.1M NaHCO3). We incubated the ChIP and input samples overnight at 65°C in presence of NaCl to reverse crosslinks and then digested them consecutively with 8ug RNAse A (30 minutes at 37°C) and 80ug Proteinase K (2 hour at 55°C). Lastly, we purified the samples using traditional phenol:chloroform extraction and ethanol precipitation. We quantified all samples using the Qubit dsDNA High Sensitivity kit (Thermo Fisher Scientific).

ChIP against PU.1 was carried out similar to the histone ChIP procedure described above, with the following differences: $10x10^6$ fixed cells were used per ChIP assay, 6ul of the PU.1 antibody (MA5-15064, ThermoFisher Scientific) was used, fixed cells were only sonicated for 5 cycles (30sec on/off).

All sequencing libraries were generated using the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs) for 1-5ng of starting material and without size selection. The Qubit dsDNA High Sensitivity kit (Thermo Fisher Scientific) and Agilent Bioanalyzer 2100 were used to QC the libraries. Libraries were sequenced on the SE 75bp Illumina NextSeq or the SE 100bp Illumina HiSeq 2500 platform at the Massively Parallel Sequencing Shared Resource (MPSSR) at Oregon Health Science University, and on the PE 2x100 Illumina HiSeq 400 platform at the Genomics and Cell Characterization Core Facility (GC3F) at the University of Oregon.

PU.1 ChIP-seq analysis and assessing LAVA binding

Raw reads from in-house gibbon and public human PU.1 ChIP-seq (GEO no.: GSM803531 and GSM803398) were QC'd using FastQC (*6*) and trimmed with Trimmomatic (*7*). We then used bowtie2 (*8*) to align processed reads from gibbon and human ChIP-seq replicates to Nleu3.0 and Hg38, respectively.

RepEnrich2 (*9*) and edgeR (*10*) were next used to calculate number of aligned reads to each repeat family and to test enrichment of repeat families in PU.1 ChIP-seq data compared to input. Briefly, we removed all low-complexity and simple-repeat annotations from the RepeatMasker annotations of each genome. The modified RepeatMasker annotations were used by RepEnrich2 (*9*) along with the ChIP and input alignments to calculate the amount of unique and multi-mapping read alignment to repeats in each dataset. Next, edgeR (*10*) was used to normalize read counts based on library size (counts per million, CPM) and to test significant enrichment of all repeat families in PU.1 samples relative to input (FDR <0.05).

Lastly, we aligned the ChIP-seq data to the corresponding genomes using BWA (Li and Durbin 2009), removed low quality and multi-mapping read alignments using samtools (*4*) and used MACS2 (*11*) to predict significant peaks (q<0.01) relative to appropriate inputs. We determined the correlation between the biological replicates using deepTools2 (*12*) and since the replicates displayed high correlation (Pearson correlation coefficience=0.82) we pooled the gibbon PU.1 peaks across the two replicates. We intersected PU.1 peaks and LAVA inserts using BEDtools (*2*) by requiring the PU.1 peak to encompass the entire length of LAVA, or vice versa. For *de novo* LAVA inserts not annotated in the reference gibbon genome, we used the coordinate of their predicted insert sites for intersection.

RNA-seq data collection and normalization

Total RNA was extracted from the 9 EBV-transformed lymphoblastoid cell lines (LCLs) available for NLE individuals using the RNeasy Mini kit (Qiagen) and RNA integrity scores were assessed on the Bioanalyzer with the RNA 6000 Pico kit (Agilent Genomics). RNA-seq libraries were generated with the TruSeq kit and sequenced to obtain ~40 million SE 100bp reads/sample. Reads were trimmed with Trimmomatic (*7*) and aligned to Nleu3.0 using STAR (*13*). We counted high-quality uniquely mapping reads (MAPQ>25) aligning to each gene and then normalized our gene counts in multiple steps. First, we used edgeR (*10*) to remove genes for which fewer than two individuals had >0.5 read counts per million (CPM). Next, we used the Conditional Quantile Normalization package [CQN; (*14*)] to normalize gene counts based on gene length, GC content and RNA-seq library size. The normalized read counts were

regularized log (rlog) transformed using DEseq2 (*15*) to stabilize the variance, and lastly, genes located on unassembled contigs in the genome were excluded.

Gene expression near LAVA vs. the whole genome

We used custom R scripts and a two-tailed permutation approach to compare the overall mean and variation of expression of genes associated with fixed (n=448) and polymorphic LAVA (n=478) to rest of the genome. Briefly, for each gene associated with fixed and polymorphic LAVA we calculated the mean and inter-individual variability of expression in relation to the mean (i.e. coefficient of variation, CV) among our 9 NLE individuals. Then, we randomly selected the same number of genes across the genome (n=478 for comparisons with polymorphic LAVA and n= for 448 for fixed LAVA). We repeated the random selection of genes 10,000 times and each time we calculated the average mean and CV for the group of randomly selected genes. Our empirical p-value was the proportion of times (out of 10,000) that the mean and CV of randomly selected genes exceeded that of the true genes associated with fixed- and poly-LAVA.

Based on the direction of difference in mean and CV between genes nearby fixed- and poly-LAVA, we next used one-tailed permutation tests to compare gene expression patterns between genes associated with fixed- and poly-LAVAs. Briefly, we calculated the difference in mean and CV of expression of genes nearby fixed and poly-LAVA, then we randomly assigned 478 and 448 of all genes nearby LAVA to be associated with fixed and poly-LAVA classification. We repeated this random classification 10,000 times and each time we calculated the difference in mean and CV of the two classes of genes. The proportion of times these differences were greater than the true difference between fixed and poly-LAVA, was our empirical p-value.

Allele frequency spectrum and summary statistic calculation

We used ANGSD (*16*) on WGS data to estimate an allele frequency spectrum (AFS) for 10 *Nomascus leucogenys* (NLE), excluding the individual that was used to construct the reference genome. The estimated AFS was folded based on Nleu3.0 and individual genotype likelihoods calculated using the

GATK approach (-GL 2) and assuming Hardy-Weinberg equilibrium (-doSaf 1). We also estimated AFS separately for the following region types: 10Kb either side of fixed-LAVA, 10Kb either side of poly-LAVA, zero-fold and four-fold degenerate sites based on Ensembl genes annotations (identified using SnpEff (*17*), ~12,000 putative non-genic loci identified in Veeramah et al. (*18*) (total of 124,073,560bp), and ~34,000 putative 1Kb neutral loci lifted over from Hg18 from Gronau et al. (*19*) (total of 35,587,610bp). The genome-wide AFS was then used by ANGSD as the basis for estimating genome-wide summary statistics for the same set of 10 individuals. Summary statistics [e.g. ΘW, Θπ and Tajima's D (*20*)] were calculated in windows of 10Kb and a step size of 1bp.

Modeling gibbon population demography

We used ∂a∂i (*21*) to estimate demographic parameters from the AFS of neutral loci estimated using ANGSD. We tested the likelihood of both a standard neutral model (snm) and a one-step size change (2-epoch) model (Figure S5). For the latter, the free parameters were the relative size change from the ancestral population effective population size (nu) and the time of this size change in coalescent units (t). Likelihoods between the simulated and observed data were assessed using the scaled multinomial method, with θ estimated post fitting the best model. Likelihoods for different combinations of nu and t were first examined along a two dimensional grid with nu ranging from log(nu)=-3 to log(nu)=3 in steps of 0.1 and t ranging from 0 to 2.0. The best estimate from this grid was then used as a starting point to fit the data using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) optimizer. The grid sizes for extrapolating the approximate solution to the partial differential equation were 40, 50 and 60.

Testing significance of selection signal around LAVA

To limit the potential impact of copy number variation or segmental duplications in our analysis, we calculated the average per site WGS coverage for each LAVA element across 10 NLE individuals (excluding the individual used to construct the reference genome). Given that the average coverage across chromosome 2 was 258X, we performed our downstream evolutionary analysis only on LAVA elements with total coverage of 200-300X. LAVA elements on chrX were also filtered out. These

filtrations reduced the 1096 fixed-LAVA to 734, and the 1,172 poly-LAVA to 808 elements. Amongst the 734 fixed-LAVA elements included in our downstream analysis, 13 and 18 had both upstream and downstream sequences with Tajima's D values that were within the 5% most negative elements compared to A) all other 10Kb loci genome-wide with the same number of callable sites and B) all 10Kb loci with at least 25% of callable sites. 11 LAVA elements were considered significant using both measures, and 19 (2.6%) were identified in total (Table S5). All these loci had Tajima's D values <-2.0, compared to a genome wide estimate of -0.95, suggestive of positive selection occurring either side of (and presumably within) these LAVA elements. Particularly noteworthy amongst the fixed LAVA elements were a cluster of four elements spanning ~300Kb on chr18 from 30,466,077-30,797,766 that overlap a major dip in Tajima's D (Fig. S4). Some of the lowest Tajima's D values in the whole genome were found in this region. In comparison, 8 and 11 (with the all the former being found in the latter) out of the 808 LAVA elements (1.3%, approximately half the proportion of fixed LAVA elements) polymorphic in the NLE samples were significant outliers for Tajima's D based on the same criteria as above (Table S5).

To assess significance of Tajima's D values around each LAVA element we used ms (*22*) to perform neutral simulations under the one-size change demographic model inferred from ∂a∂i (*21*). Using a 2-epoch model with a population expansion in the recent past provided a much better fit to the data compared to a simple standard neutral model (Fig. S5). Estimates of the relative size change (nu) and time of this size change (t) were very similar for both data sets (Table S5) and are suggestive of a ~4 fold increase in population size ~50,000 generations ago assuming a per generation mutation rate per site of $1x10^{-8}$. However, we note that for our purposes we do not care about the exact demographic model, only that it can suitably replicate the observed AFS. In this case, simulating under the best-fit 2-epoch model gave an expected Tajima's D that matched or was very close to the observed data (-0.816 vs. -0.813 for the Veeramah et al. (*18*) data, and -0.821 vs. 0.821 for the Gronau et al. (*19*) data; Table S5). Given the model inferred above, we then performed 10,000 coalescent simulations for each LAVA element to represent the expected distribution of Tajima's D at a 20Kb window around LAVA under neutrality. For the fixed-LAVA elements, 35 loci (4.7%) demonstrated Tajima's D values more extremely negative than any simulated data point ($p<0.0001$), while 20 loci (2.5%) showed such extreme values for the

polymorphic LAVA elements. All 30 LAVA elements (fixed and polymorphic) identified as having unusually low Tajima's D value in the empirical distributions were also found to be more extreme than could be simulated under neutrality, except for one fixed loci that had a coalescent p-value of 0.0001 (i.e. 1 out of 10,0000 simulations produced a more extreme value).

Next, to assess the collective Tajima's D signal separately around fixed- and poly-LAVA, we averaged the average Tajima's D of the upstream and downstream window surrounding each LAVA element across all fixed- and poly-LAVA elements (i.e. an average of averages). Then, we randomly sampled pairs of 10Kb loci (734 pairs for fixed-LAVA and 808 pairs for poly-LAVA) from across the genome making sure windows within a pair were separated by the same physical distance as the corresponding LAVA elements. For each random set of window pairs, we calculated the average of the averages Tajima's D for the upstream and downstream windows, and repeated this process 10,000 times. An empirical p-value was calculated by comparing the average of averages Tajima's D for fixed and poly-LAVA to the corresponding estimated null distribution. We found that the average of the averages Tajima's D value for the fixed LAVA elements was -1.247. This was more negative than any of the 10,000 randomly generated sets, which had a mean of -1.153 (p-value<0.0001, 1% and 99% percentile of -1.191 and -1.106, respectively). For the poly-LAVA elements the averages Tajima's D was -1.138, which produced an empirical p-value of 0.813 based on the 10,000 randomly generated sets of genome-wide loci (mean -1.156, 1% and 99% percentile of -1.120 and -1.109 respectively), suggesting that this class of LAVA element is not behaving much differently from other loci across the genome.

Similarly, using the coalescent framework described above, we were unable to generate average Tajima's D values across 734 simulated loci that were as negative as those observed for the fixed LAVA elements (average Tajima's D for 734 fixed LAVA elements = -1.038, smallest Tajima's D generated from 10,000 simulations = -0.857). However, we note that we were also unable to generate as negative average Tajima's D values via neutral simulations as that observed for polymorphic LAVA elements (average Tajima's D for 808 polymorphic LAVA elements= -0.951, smallest Tajima's D generated from 10,000 simulations = -0.857). This may either be due to weak positive selection acting at polymorphic loci, or inability of our proposed neutral model to fully capture neutrality in this data.

Assessing genetic diversity around LAVA elements

Examining diversity via the AFS at LAVA elements compared to other categories of data, we found that fixed LAVA elements had a higher proportion of singletons compared to the genome-wide AFS and the two categories of putative neutral loci, as well as even four-fold degenerate sites. However, we note that the last category had a large relative excess of doubleton mutations (Fig. S6A). Only zero-fold degenerate sites had a larger proportion of singletons, which is probably the result of purifying selection and the presence of nearly neutral variants, similar to a pattern observed in humans (*23*, *24*), as this class of sites has much lower levels of diversity overall (~50-30%; Fig. S6B). Fixed LAVA sites also had a lower Tajima's D compared to the genome-wide and neutral sites (Fig. S6C). The zero-fold and four-fold sites showed lower Tajima's D compared to fixed-LAVA, because of the excess singletons and doubletons in these categories. The complex AFS (extreme doubletons) and diversity (very high ϴW and ϴπ, but relatively normal heterozygosity patterns at the level of individual genomes) patterns observed at four-fold sites are likely the result of linkage to the zero-fold sites, though more thorough population modeling would be required to better understand these patterns (*25*). Again, polymorphic LAVA elements were unremarkable compared to genome-wide patterns, although they did have greater proportions of singletons and lower Tajima's D than the neutral loci, consistent with the coalescent simulation analysis reported above.

Assessing selection at putative fixed-LAVA enhancers

We used permutation analysis to compare selection signal (i.e. Tajima's D value) at putative fixed-LAVA enhancers, against other fixed-LAVA. Briefly, we identified fixed LAVA that were 1) included in our evolutionary analysis and 2) overlapped with bivalent, poised or active enhancer chromatin states. Of these 60 putatively functional fixed-LAVA elements, 42 were located within 3Kb of genes (i.e. promoter, intronic or terminator). We averaged the Tajima's D values around 20Kb of all putative fixed-LAVA enhancers. Next, we selected random sets of 60 fixed-LAVA 1000 times, each time making sure that 42 of the selected LAVA were located nearby genes. This precaution was made to control for the potential

effects of background selection caused by proximity of putative functional LAVA to genes. An empirical p-value was generated by comparing the mean of putative fixed-LAVA enhancers, to the distribution of the randomly selected sets of fixed-LAVA.

Testing significance of association with gene ontology terms

To ensure that the significant enrichment of fixed LAVA elements near DNA repair gene ontology (GO) terms was not caused by biases or errors in the gibbon gene annotations, we used two separate permutation approaches to calculate empirical p-values for the significant GO terms identified by Enrichr (*26*). First, we recorded the number of genes found nearby LAVA for each of the significant GO terms. In the first permutation approach, we randomly shuffled the positions of all fixed-LAVAs across the whole genome and recorded the genes within 3Kb of the shuffled LAVA positions. In the second permutation approach, we randomly shuffled positions of LAVAs that were within 3Kb of genes, while restricting their new random positions to be >3Kb of another gene. This latter approach was more conservative and accounted for effects of gene length (i.e. longer genes are more likely to have LAVA inserts) and LAVA's overall tendency to be inserted nearby genes. For both approaches we then identified genes nearby the shuffled LAVA and used Enrichr (*26*) to obtain the GO terms associated with them. We then repeated each shuffling process 1000 times and each time recorded the total number of genes associated with each GO term of interest. We then obtained an empirical p-value by comparing the true gene counts for each significant GO term to the null distribution of GO term counts. Using these two permutation tests, we were not able to randomly generate as much enrichment as we had found with the true LAVA positions for any of the significant GO terms, suggesting that all the GO term enrichments identified by Enrichr were still significant after accounting for imperfections in the gibbon gene annotations (empirical p-value <0.001).

Examining selection at fixed-LAVA near DNA repair genes

To test whether the signals of selection at DNA-repair fixed-LAVA were significantly different from expectation, we identified the LAVA associated with the significant DNA repair GO terms above. Then,

we averaged the Tajima's D within 20Kb of these fixed-LAVA and compared it to the null distribution of the mean Tajima's D of the same number of randomly selected fixed-LAVA located nearby genes. We intentionally limited our 1,000 random selections to only select from the subset of fixed-LAVA that are <3Kb of genes, to make sure our results were not biased by the fact that all DNA-repair fixed-LAVA were nearby genes.

References

1. E. J. Gardner, V. K. Lam, D. N. Harris, N. T. Chuang, E. C. Scott, W. S. Pittard, R. E. Mills, S. E. Devine, The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome Res*. **27**, 1916–1929 (2017).

2. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. **26**, 841–842 (2010).

3. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. **25**, 1754–1760 (2009).

4. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. **25**, 2078–2079 (2009).

5. D. Kostka, A. K. Holloway, K. S. Pollard, Developmental Loci Harbor Clusters of Accelerated Regions That Evolved Independently in Ape Lineages. *Mol Biol Evol*. **35**, 2034–2045 (2018).

6. S. Andrews, FastQC: A quality control tool for high throughput sequence data. (2010) (available at http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

7. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. **30**, 2114–2120 (2014).

8. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat Meth*. **9**, 357–359 (2012).

9. S. W. Criscione, Y. Zhang, W. Thompson, J. M. Sedivy, N. Neretti, Transcriptional landscape of repetitive elements in normal and cancer human cells. *BMC Genomics*. **15** (2014), doi:10.1186/1471-2164-15-583.

10. M. D. Robinson, D. J. McCarthy, G. K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. **26**, 139–140 (2010).

11. Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, X. S. Liu, Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*. **9**, R137 (2008).

12. F. Ramírez, F. Dündar, S. Diehl, B. A. Grüning, T. Manke, deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res*. **42**, W187–W191 (2014).

13. A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T. R. Gingeras, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. **29**, 15–21 (2013).

14. K. D. Hansen, R. A. Irizarry, Z. Wu, Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics*. **13**, 204–216 (2012).

15. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. **15,** 550 (2014).

16. T. S. Korneliussen, A. Albrechtsen, R. Nielsen, ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics*. **15**, 356 (2014).

17. P. Cingolani, A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, D. M. Ruden, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*. **6**, 80–92 (2012).

18. K. R. Veeramah, A. E. Woerner, L. Johnstone, I. Gut, M. Gut, T. Marques-Bonet, L. Carbone, J. D. Wall, M. F. Hammer, Examining phylogenetic relationships among gibbon genera using whole genome sequence data using an approximate bayesian computation approach. *Genetics*. **200**, 295–308 (2015).

19. I. Gronau, M. J. Hubisz, B. Gulko, C. G. Danko, A. Siepel, Bayesian inference of ancient human demography from individual genome sequences. *Nature Genetics*. **43**, 1031–1034 (2011).

20. F. Tajima, Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*. **123**, 585–595 (1989).

21. R. N. Gutenkunst, R. D. Hernandez-Rodriguez, S. H. Williamson, C. D. Bustamante, Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. PloS Genet. **5**, e1000695 (2009).

22. R. R. Hudson, Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics*. **18**, 337–338 (2002).

23. K. R. Veeramah, R. N. Gutenkunst, A. E. Woerner, J. C. Watkins, M. F. Hammer, Evidence for increased levels of positive and negative selection on the X chromosome versus autosomes in humans. *Mol. Biol. Evol.* **31**, 2267–2282 (2014).

24. C. Hvilsom, Y. Qian, T. Bataillon, Y. Li, T. Mailund, B. Sallé, F. Carlsen, R. Li, H. Zheng, T. Jiang, H. Jiang, X. Jin, K. Munch, A. Hobolth, H. R. Siegismund, J. Wang, M. H. Schierup, Extensive X-linked adaptive evolution in central chimpanzees. *PNAS*. **109**, 2054–2059 (2012).

25. P. W. Messer, D. A. Petrov, Frequent adaptation and the McDonald–Kreitman test. *PNAS*. **110**, 8615–8620 (2013).

26. M. V. Kuleshov, M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann, M. G. McDermott, C. D. Monteiro, G. W. Gundersen, A. Ma'ayan, Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res*. **44**, W90–W97 (2016).

Supplemental Figures

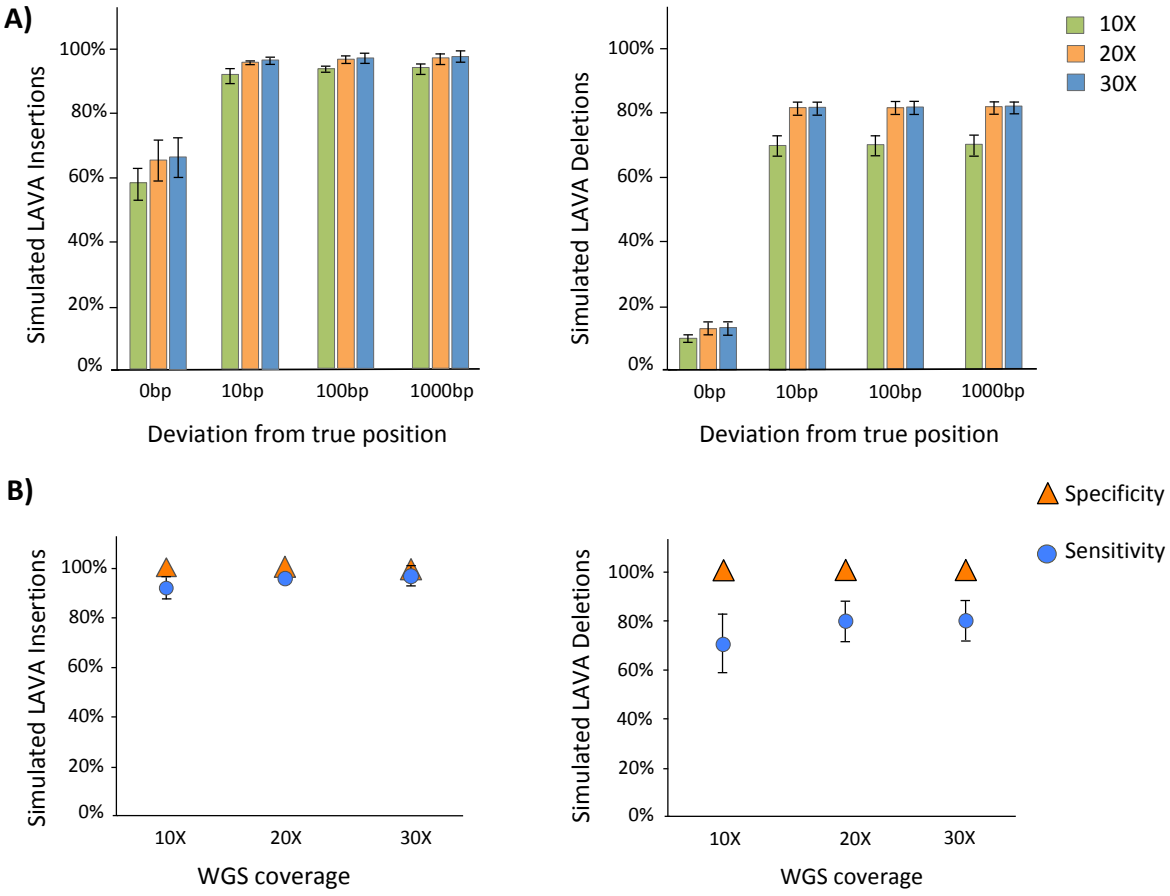


Figure S1. MELT predicts LAVA indels with high precision and accuracy. A) Percent of *in silico* LAVA insertions (*left*) and deletions (*right*) identified within various distances of their true positions are demonstrated. Bars represent mean of three mock WGS datasets, and error bars represent standard errors. **B)** Percent specificity and sensitivity of LAVA insertion (*left*) and deletion (*right*) predictions by MELT across three coverages. Data points represent mean values across the three mock WGS datasets and error bars represent standard errors.

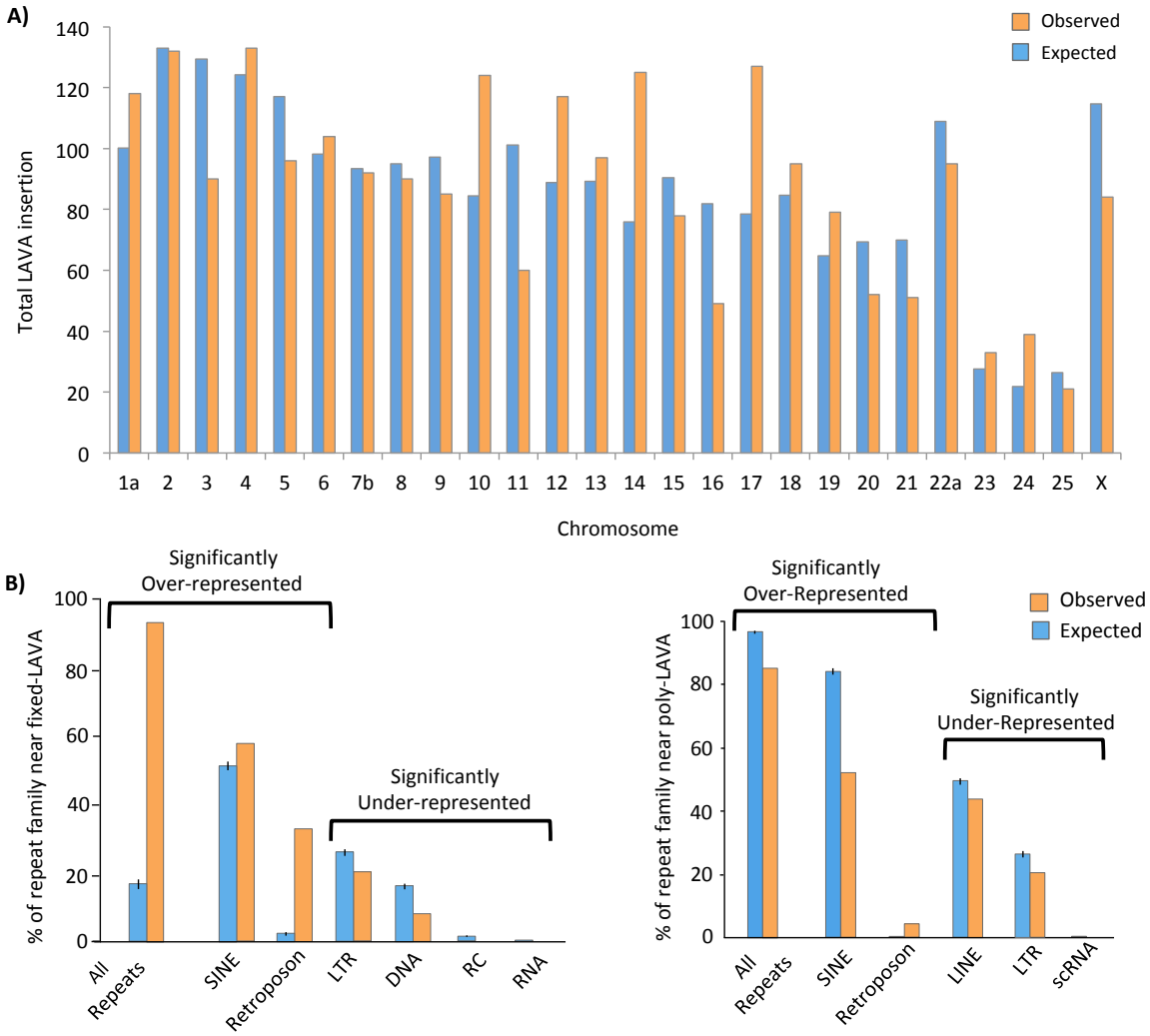


Figure S2. Distribution of LAVA across the gibbon genome. A) The number of expected and observed LAVA insertions is depicted per chromosome. **B)** Significantly over- and under-represented repeat families near fixed- (*left*) and poly-LAVA (*right*) are demonstrated (q<0.05). Error bars represent standard error.

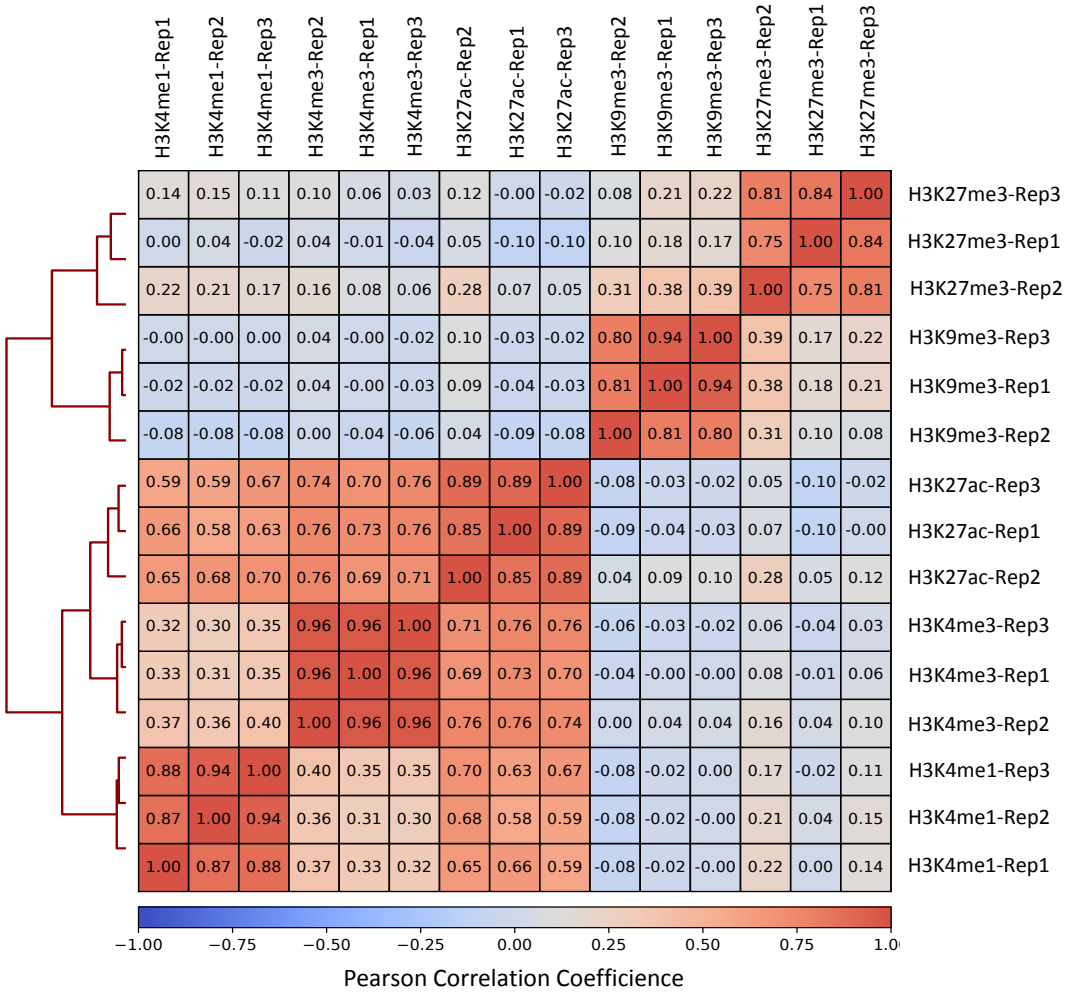


Figure S3. Histone ChIP-seq signals correlate across replicates. A heatmap demonstrates correlation between all histone ChIP-seq data sets. Cell shades reflect the Pearson Correlation Coefficient for each pair-wise comparison, with the exact values printed within each cell.

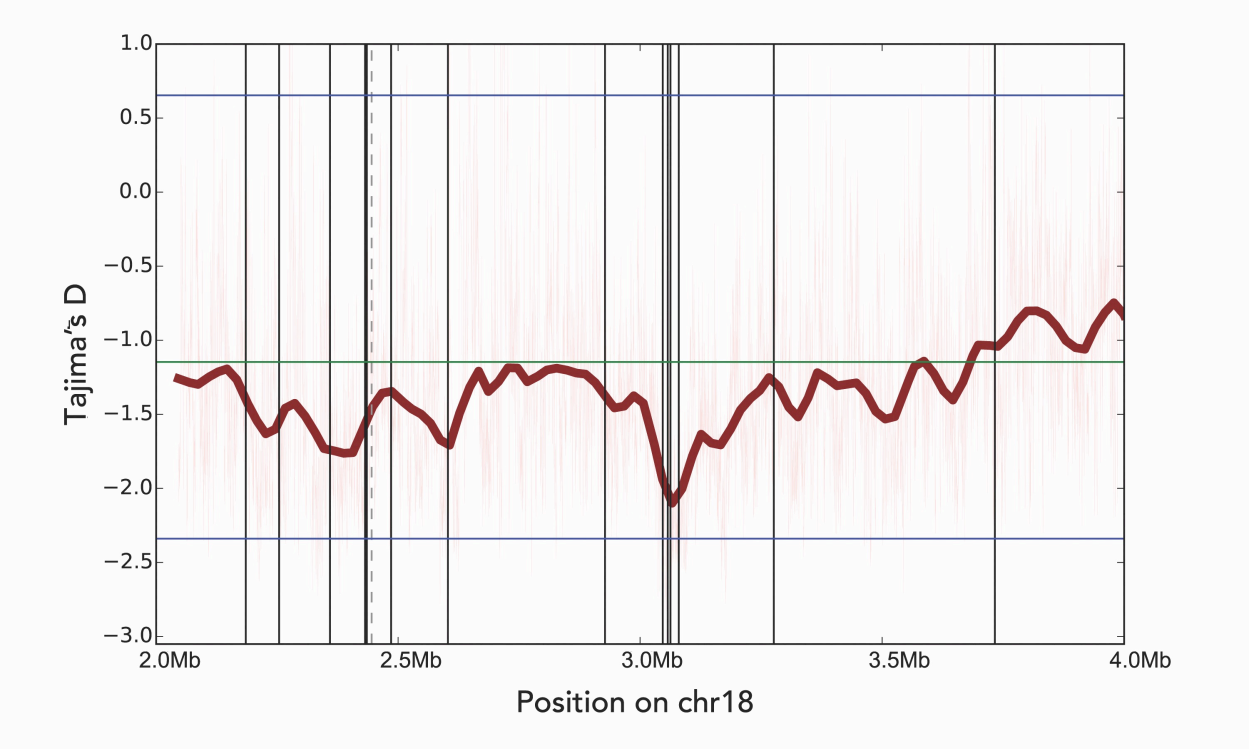


Figure S4. A cluster of four fixed LAVA elements display a strong Tajima's D dip.

Tajima's D values on chr18:2,000,000-4,000,000 are shown. Light red lines are Tajima's D of individual 10Kb windows, dark red line is lowess smoothing curve of these windows. Vertical black and grey dashed lines are positions of LAVA elements fixed- and poly-NLE, respectively. Blue horizontal lines are the 1% and 99% percentile of Tajima's D across the whole genome, and line green is the mean across the genome.

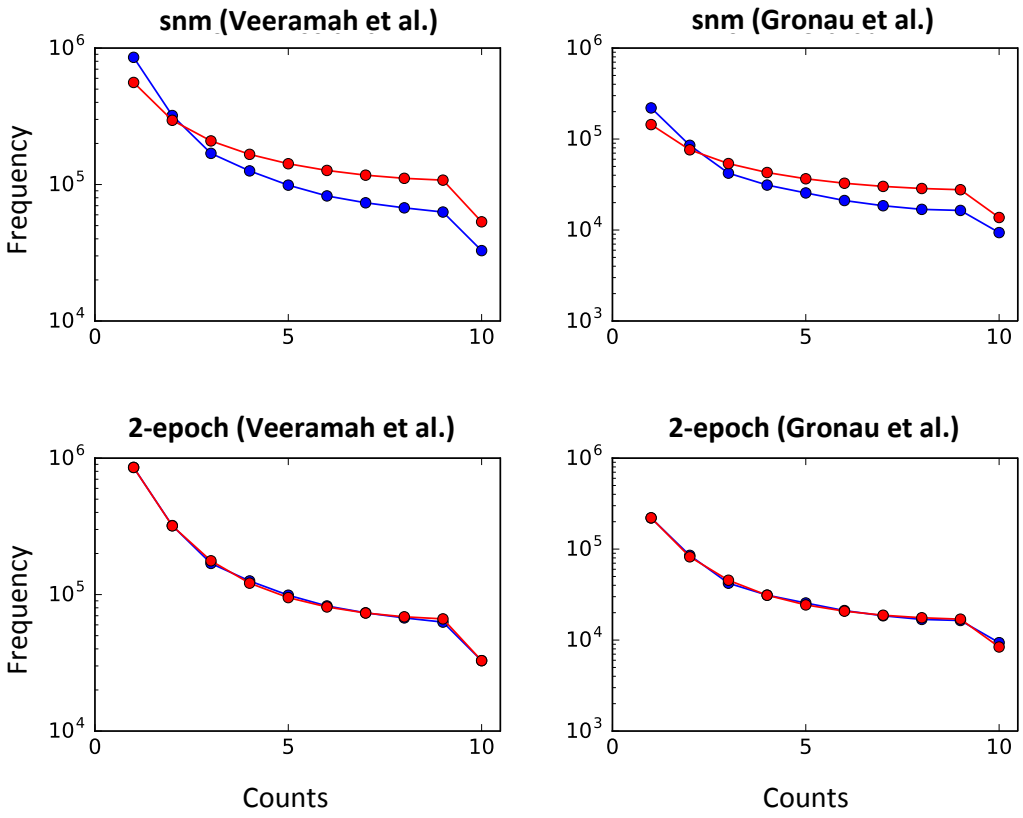
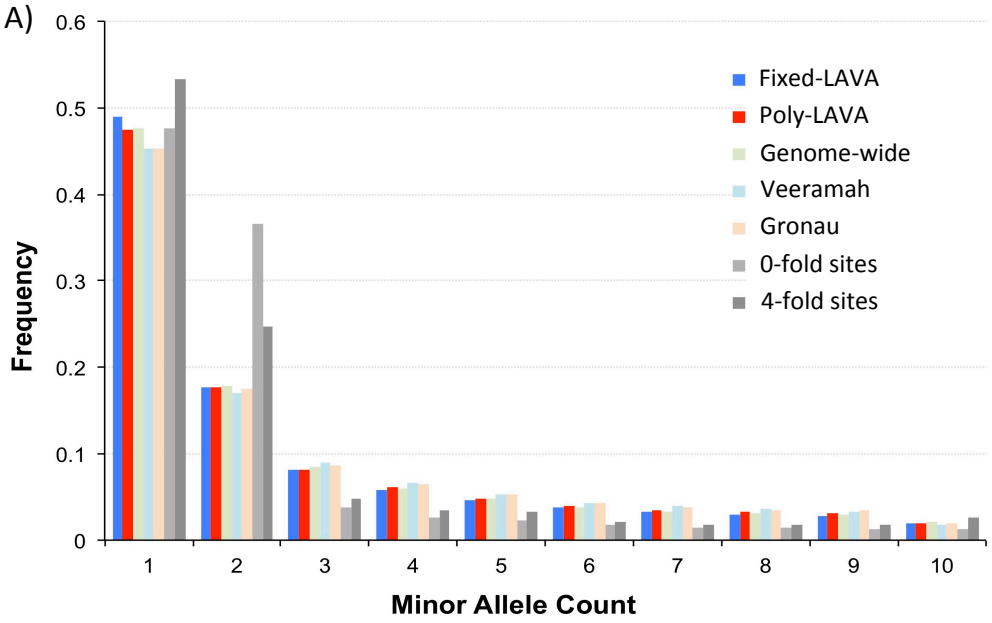
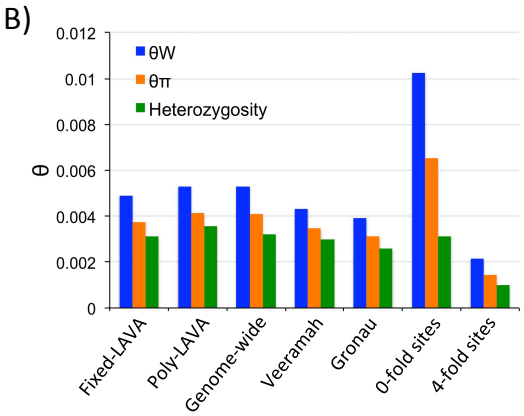


Figure S5. Comparison of standard neutral and best-fit 2-epoch models. Comparison of standard neutral model (snm) and best-fit 2 epoch model from ∂a∂i against folded AFS based on non-genic loci from Veeramah et al. (*18*) and Gronau et al. (*19*).





C)

Category	Tajima's D
4-fold sites	-1.518
0-fold sites	-1.388
LAVA_fixed	-1.008
LAVA_polymorphic	-0.931
Genome-wide	-0.953
Gronau	-0.821
Veeramah	-0.816

Figure S6. Comparing genetic diversity at LAVA elements to other sites. A) Folded AFS are shown considering segregating sites for various categories of the genome. **B)** Estimates of theta for various categories of the genome. "Heterozygosity" is based on the average proportion of heterozygous sites called in the three high coverage (>40x) genomes. **C)** Estimates of Tajima's D for various genomic categories. Abbreviations are as follows: Genome-wide= all callable sites in the genome, Veeramah and Gronau= neutral loci identified in Veeramah et al. (*18*) and Gronau et al. (*19*), respectively.