University of Natural Resources
and Life Sciences, Vienna

Department of Sustainable
Agricultural Systems
Division of Livestock Sciences

Transposon evolution in South American camelids

Christian Andres Ramos Uria

Supervisor
Assoc. Prof. Priv.-Doz. Dr. Gábor Mészáros

Co - Supervisor
Dr. Mirte Bosse

Vienna, 06.06.2021

Statutory Declaration

I hereby declare that I am the sole author of this work. No assistance other than that which is permitted has been used. Ideas and quotes taken directly or indirectly from other sources are identified as such. This written work has not yet been submitted in any part.

X 

Christian Andres Ramos Uria

Abstract

Transposable elements are genetic sequences that can replicate within a genome and are present in all domains of life. They act as a source of genetic variation, and their dynamics depend on the balance between the mechanisms of transposition and suppression; this balance depends on demography, recombination, selection, and the environment. This genetic variation can be adaptive. Domestication affects and relies on genetic diversity, and therefore could harness the genetic variation generated by transposons. South American camelids have 4 species: two are domesticated, and there is a wild counterpart for each, which makes them ideal to explore the role of transposons in domestication. Hence, we hypothesized that domesticated South American camelids had more recent transposition events than their wild counterparts. Using publicly available Illumina reads, we tested that hypothesis by measuring the relative age of transposons in domesticated species (llama and alpaca) and comparing it with their wild counterparts (guanaco and vicuña, respectively). The resulting pattern depends on the transposon family. LINEs and SINEs age can be explained by phylogenetic relationships: *Llama sp.* have younger copies than *Vicugna sp.* In contrast, LTR patterns are species and transposon specific. None of the families showed a pattern that would be congruent with a difference in transposition rates attributable to domestication. Finally, we also found evidence of genetic exchange of repetitive elements between llama un alpaca, either a consequence of their admixture or horizontal transfer.

Keywords: South American camelids, transposons, domestication.

Zusammenfassung

Transposons sind genetische Sequenzen, die sich innerhalb eines Genoms replizieren können und in allen Lebensbereichen vorkommen. Sie erhöhen die genetische Variation, und ihre Dynamik hängt vom Gleichgewicht zwischen den Mechanismen der Transposition und des Verlustes ab; dieses Gleichgewicht hängt von Demografie, Rekombination, Selektion und der Umgebung ab. Diese genetische Variation kann adaptiv sein. Südamerikanische Kameliden haben 4 Arten: zwei sind domestiziert, und für jede gibt es ein wildes Gegenstück. Deswegen sind sie ideal, um die Rolle von Transposons bei der Domestikation zu erkunden. Daher stellten wir die Hypothese auf, dass es bei der Domestikation südamerikanischer Kameliden eine Rolle spielen könnte. Mit Illumina-Reads haben wir diese Hypothese getestet, indem wir das relative Alter von Transposons bei domestizierten Arten (Lama und Alpaka) gemessen und mit ihren wilden Gegenstücken (Guanako und Vicugna) verglichen haben. Jeder Transposon-Familie verhält sich anders. Das Alter von LINEs und SINEs kann durch phylogenetische Beziehungen erklärt werden: *Llama sp.* haben jüngere Exemplare als *Vicugna sp.* Im Gegensatz dazu sind LTRs Art- und Transposon spezifisch. Kein Transposon unterstützt unsere Hypothese. Schließlich fanden wir Hinweise auf einen genetischen Austausch repetitiver DNA zwischen Lama und Alpaka, entweder als Folge ihrer Introgression oder horizontalen Transfer.

Schlüsselwörter: südamerikanischer Kameliden, Transposons, Domestikation.

Acknowledgements

There are many people who had an influence on my development as a person and scientist; and who contributed to this thesis.

Thanks to the EMABG staff, for giving me the opportunity to be a part of this program. Particularly, thanks to Nathalie Schwaiger, whose constant help was crucial to easing my arrival and stay in a previously unknown place.

To Dirkjan Schokker and Mario Calus, who guided me through my first steps into the world of genomics. Thanks for believing in my ideas, even before I knew if they could work.

To Gábor Mészáros, my main supervisor. His insightful feedback was never late, nor early, but precisely when I needed it. And to my co-supervisor, Mirte Bosse, whose opinions solidified the study.

To Volga Iñiguez, my beloved mentor. Her advice is a beacon guiding my life, inside and outside academia. To Kazuya Naoki, whose teachings provided a solid background that made me a better scientist. And Luis Pacheco, who dared me to explore fields beyond my comfort zone, enriching my imagination.

To every friend who listened to my random scientific ramblings. Your tolerance was very much needed.

I am grateful for all the support my family provides. Your care is felt despite the distance. And particularly to my grandpa, now among the stars.

Table of contents

Introduction

Transposable element (TE) is an umbrella term that encompasses different DNA sequences that can replicate (increase copy number) within a host genome, and subsequently be transmitted across generations (Bourgeois & Boissinot, 2019; Wells & Feschotte, 2020). They have been found in every domain of life (Makałowski et al., 2019). Likewise, they have different origins rather than a single phylogenetic origin (Piégu et al., 2015); some TEs share evolutionary origins with virophages (Campbell et al., 2017; Koonin et al., 2015), or retroviruses (Platt et al., 2018), while others derive from tRNA (Platt et al., 2018). TEs evolutionary origins is still an active research area. And the prevalent horizontal transfer coupled with the gain and loss of modular motifs results in a phylogenetic network rather than a tree (Koonin & Krupovic, 2017).

TEs are classified into two different groups, depending on the mechanism of replication within the genome: class I transposons have an RNA intermediate, whereas class II transposons have a DNA intermediate (Makałowski et al., 2019). Class I transposons, also called copy and paste, transcribe the sequence into RNA to then use a reverse polymerase (Makałowski et al., 2019; Wells & Feschotte, 2020); Class II get excised and then reinserted elsewhere (Makałowski et al., 2019; Wells & Feschotte, 2020). Within each class, TEs are further classified by the mechanism of transposition (Piégu et al., 2015). In practice, new members are added to a TE family according to the 80-80-80 rule, which states that a TE belongs to a family if it has a length > 80 bp and shares at least 80% of sequence identity over 80% of its length (Wells & Feschotte, 2020). This rule only reflects phylogeny if the sequences evolved neutrally after a single burst (Wells & Feschotte, 2020).

Unlike in other taxa, mammals have fewer TE families with very high copy numbers (Le Rouzic & Deceliere, 2005) and occupy between 30% and 50% of the host's genomes (Platt et al., 2018). The few copies of class II transposons that remain are inactive (in most species), and of the class I transposons, LINEs and SINEs are the most abundant, followed by LTRs (Platt et al., 2018). Both LINEs and SINEs are non-LTR TEs, meaning that they do not have flanking long terminal repeats (Figure 1). LINEs codify two proteins, ORF1 and ORF2 (Figure 1), that catalyze the reverse

transcription and transposition of the LINE RNA, whereas SINEs are parasite RNAs that highjack the LINE machinery to transpose (Makałowski et al., 2019; Wells & Feschotte, 2020). LTRs also use a reverse-polymerase and are characterized by flanking long terminal repeats (Figure 1) (Makałowski et al., 2019).

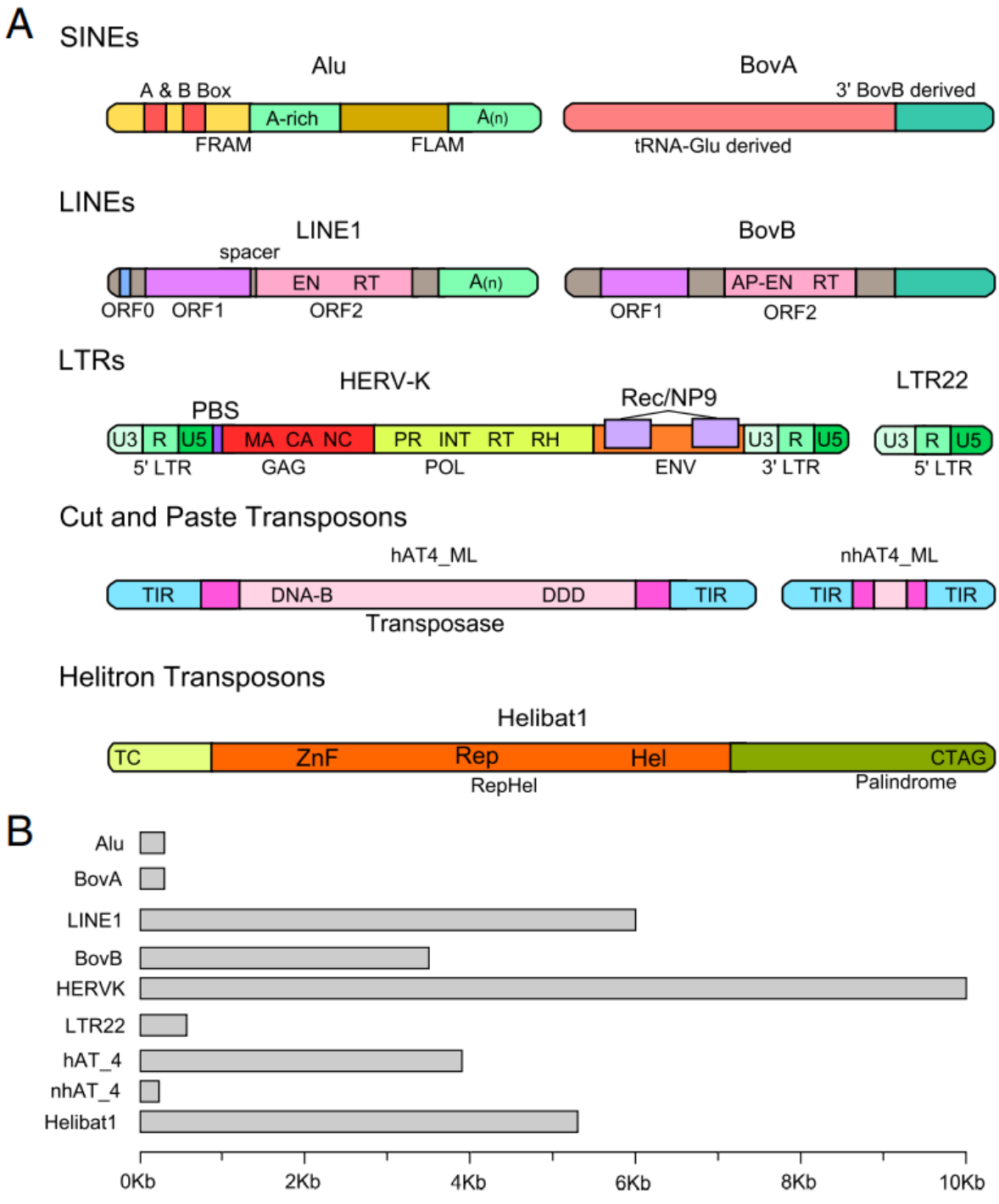


Figure 1. Mammalian TEs (Platt et al., 2018). A) Structure of common mammalian TEs. SINEs: Short Interspersed Nuclear Elements; LINEs: Long Interspersed Nuclear Elements; LTRs: Long Terminal Repeat (retrotransposons). B) Common mammalian transposons drawn to scale.

It was originally expected that genomic TE content could be explained by the effective population size, and that it would correlate with genome size, but the full explanation is more complex (Bourgeois & Boissinot, 2019; Makałowski et al., 2019; Wells & Feschotte, 2020). TE microevolution proved to be a depend on the selection-drift balance, population demography, the environment and the regulation of transposition and deletion events (Figure 2) (Guio & González, 2019). TE abundance is controlled by the balance between rates of transposition, fixation and deletion (Wells & Feschotte, 2020); so to keep them in check, they have to be (self)inhibited or they can be kept in check via negative selection (Le Rouzic & Deceliere, 2005). The rates of transposition and deletion need not to be in equilibrium, which can cause bursts of transposition (Bourgeois & Boissinot, 2019; Le Rouzic & Deceliere, 2005). The effect of selection depends on the population size, but also on TE allele frequencies and recombination: TEs that cause non-homogeneous recombination are more deleterious, and that effect is mitigated as the frequency of TEs in a population increases and depends on the TE copy number in the genome (Bourgeois & Boissinot, 2019). This negative selection and mechanisms that allow for preferential transposition to specific regions in a genome generate non-random spatial TE distributions (Sultana et al., 2017). TE spread on the long run also depends on horizontal transmission events and sex (Bourgeois & Boissinot, 2019; Gilbert & Feschotte, 2018).

TEs are a source of genomic variation, and therefore they have an adaptive potential as well (Percharde et al., 2020). Because of the continuous arms race between TE suppression and escape (Platt et al., 2018), mechanisms that arose to control TE proliferation have been exapted into more general roles, like DNA methylation and KRAB Zink Fingers (KZNFs) (Branco & Chuong, 2020). Likewise, the mechanisms that allow TEs to interact with the cellular machinery can be exapted to interfere with other selfish genetic elements (Jangam et al., 2017). TEs often affect gene expression

during stressful situations (Branco & Chuong, 2020); but also, if environmental factors affect the expression of genes that control the expression of mobile elements, then the environment can trigger a mobile element mediated rewiring of transcriptional networks (Shapiro, 2017). Hence, stressful conditions can trigger bursts of transpositions which can (but do not need to be) be adaptive (Bourgeois & Boissinot, 2019). Also, because new TEs are only transmitted vertically through the germline, they need to be expressed early in the embryogenesis (Percharde et al., 2020). Therefore, some TEs play a role early in the embryogenesis (Branco & Chuong, 2020), and in combination with their suppressors, provide robustness and flexibility (Percharde et al., 2020). Considering all of that, it has been proposed that TEs should be considered symbionts rather than only parasites (Percharde et al., 2020).

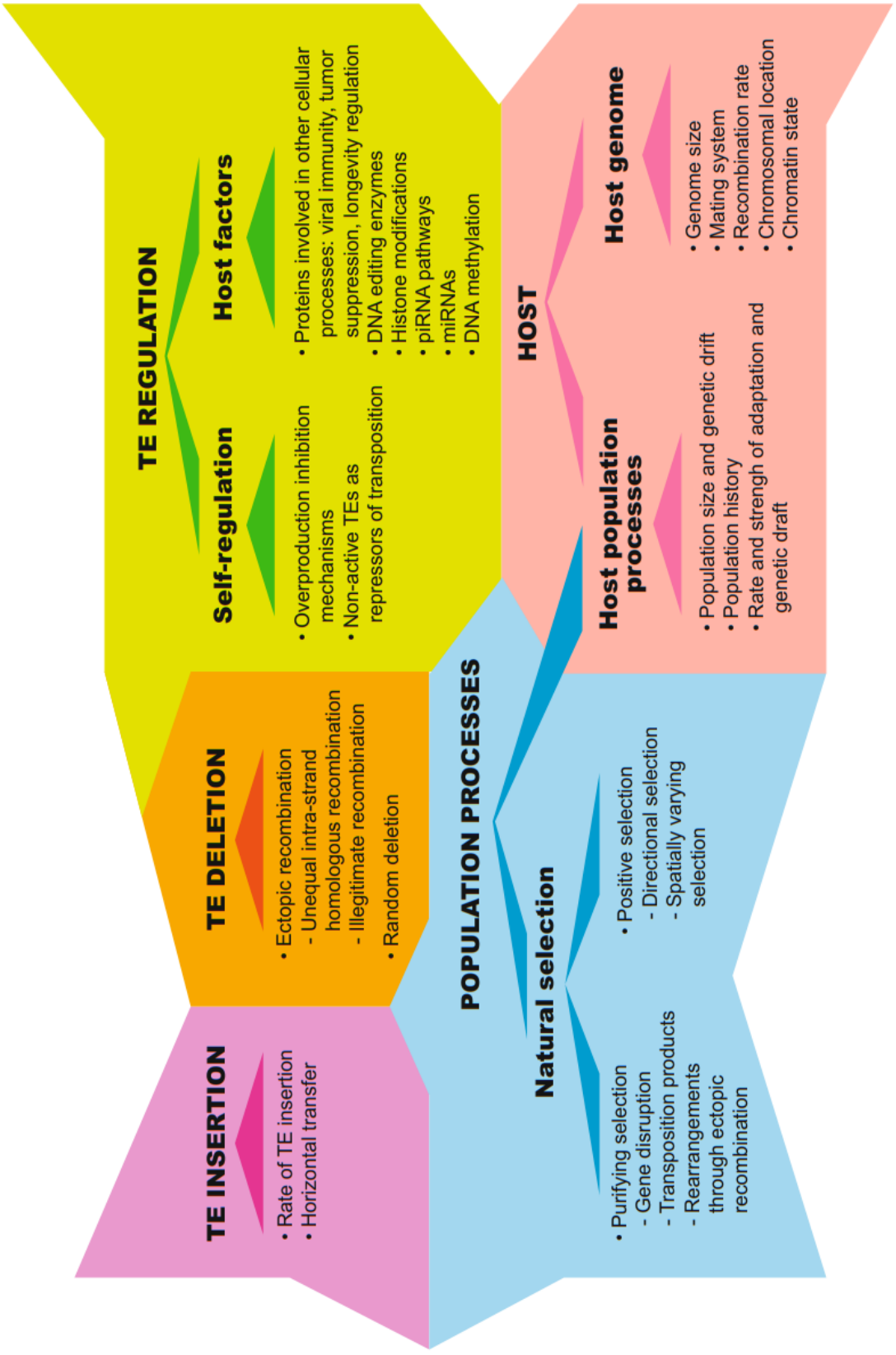


Figure 2. Processes that affect the spread and diversity of TEs (Guio & González, 2019).

Given this capacity of TEs to connect environment and life history traits with evolution and adaptation, they could play a role in the process of domestication. According to the expended evolutionary synthesis, niche construction leads to transmission of genetic, behavioural, ecologic, and cultural information, leading to periods of stasis and change, enabled by phenotypic plasticity (Zeder, 2017). If TEs play a role in adaptation during domestication, or if the transposition rate increases because of the domestication process, we would expect increased transposition events in domesticated species.

Therefore, we wanted to test that comparing domesticated species with their wild counterparts, and South American camelids (SACs) are ideal for that. They belong to the Camelidae family, together with old world camels (Wheeler, 2012). The four surviving species are divided into two genera: *Llama* and *Vicugna*. Each genus has a wild species and a domesticated counterpart: *Llama glama* (Llama) was domesticated from a subspecies of *L. guanicoe* (Guanaco), and *Vicugna pacos* (Alpaca) was domesticated from *V. vicugna* (Vicuña) (Fan et al., 2020; Wheeler, 2012). Whereas vicuñas occupy areas of high elevation in the high Andes, guanacos descend more, reaching the Pacific shore (Wheeler, 2012). Additionally, there is ongoing introgression between llamas and alpacas (Fan et al., 2020; Wheeler, 2012).

In this thesis, we compared the relative transposon age between domesticated South American camelids (llamas and alpacas) and their wild counterparts (guanacos and vicuñas, respectively). Younger transposons would imply an increased transposition rate. Additionally, we tested if the recent admixture between the domesticated species lead to genetic exchange of TEs.

Materials and methods

DNA samples

We used the samples from Fan et al. (2020), available at the NCBI with the project accession number PRJNA612032. Briefly, seven individuals per each species (llama, guanaco, vicuña and alpaca) were sampled, throughout all their geographical range. The resulting Illumina paired end reads had a depth between 16 and 22-fold. One *V. vicugna* sample was corrupted, hence we leaved it out from the analysis.

Transposon detection and annotation

We used two pipelines to detect and annotate repetitive elements: RepeatExplorer2 (Figure 3) and dnaPipeTE (Figure 4). Both use low coverage (less than 1) samples as an input, because only repeated elements will have enough reads to be assembled. The inverse of the coverage defines how many repeats need to be present in the genome for the repeat to be detected: for instance, a transposon with two copies can be detected with a coverage of 0.5, and if it had four copies it would need a minimum coverage of 0.25. RepeatExplorer2 works by assembling graphs, where each read is a node, and two reads share a link if one can be aligned with the other (Novák et al., 2013). Repeated elements will form clusters, and then the reads from each cluster are assembled into contigs. Because we know the sample each read belongs to, it is possible to make comparisons of the relative abundances of reads per cluster. In contrast, dnaPipeTE relies on Trinity (Grabherr et al., 2013), an algorithm developed to assemble mRNA from proteins with alternative splicing. Trinity assembles a graph with k-mers of a given length as nodes, and two k-mers are connected if the sequences are the same, displaced by one nucleotide. Paths in the network represent alternative possible assemblies, and they are weighted according to the number of reads that support them. The pipeline runs Trinity on the down-sampled reads from a genome iteratively, to detect repeated elements that could have been lost due to the random sampling. Besides the annotated contigs and the quantified abundances of each repeated element, the pipeline also estimates the distance from each read to the contig it maps to, and the resulting distribution gives a relative measure of the age of the repeated elements.

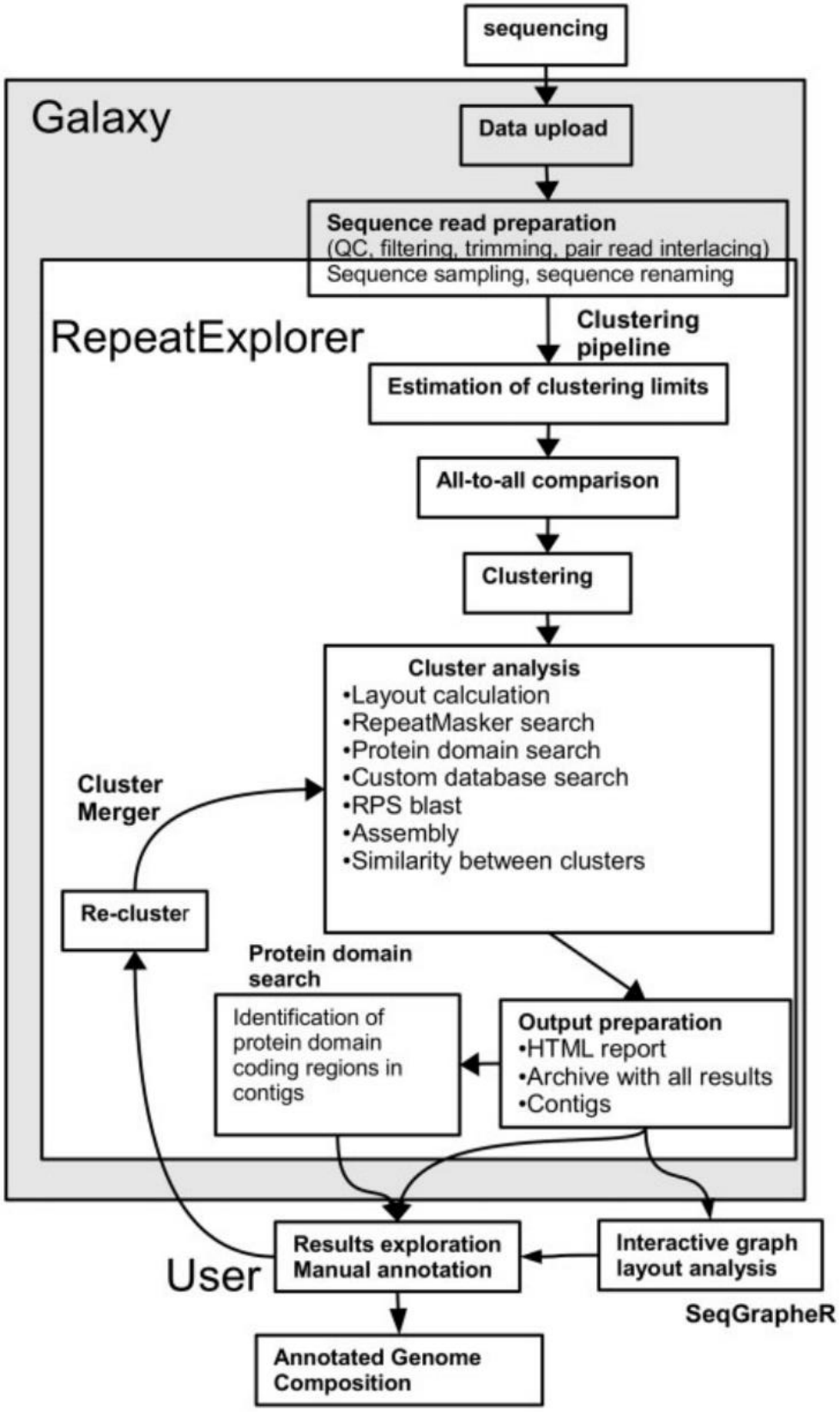


Figure 3. RepeatExplorer pipeline workflow (Novák et al., 2013).

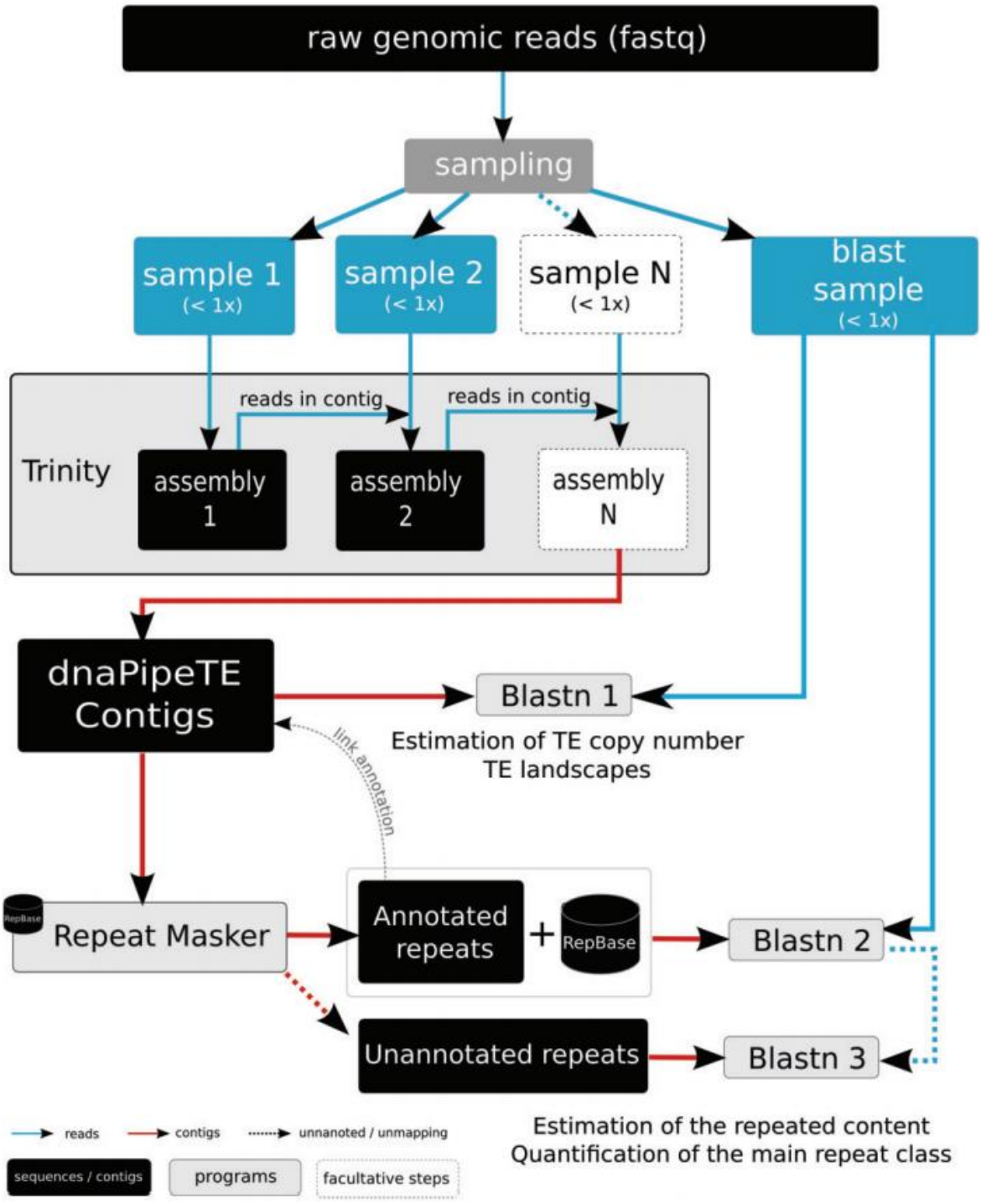


Figure 4. dnaPipeTE pipeline (Goubert et al., 2015).

We ran the RepeatExplorer2 pipeline (Figure 3) (Neumann et al., 2019; Novák et al., 2010, 2013, 2017), as explained in the second protocol of (Novák et al., 2020). Briefly, we took a random subset of reads per sample so that the final coverage is 0.1%. The same reads need to be selected in the forward and reverse reads. After uploading the

reads in fastq format to the RepeatExplorer galaxy server (Galaxy (cerit-sc.cz)), we performed a quality control with FastQC using default settings (Andrews, 2015) followed by trimming with the RepeatExplorer utilities with the default parameters except for 'end position' = 150. Then, the samples were concatenated, and we ran the pipeline in comparative mode and with the option of masking satellites turned on. Each sample, corresponding to a different individual, was treated as a separate genome (rather than pooling the reads of all samples per species). Our average coverage per sample was 0.0018x, for a genome of size = 2.6 Gb. The pipeline generates a tentative annotation that was checked manually before being plotted. Besides plotting the copy number per cluster and sample and the corresponding average copy number per species, we also did a PCoA using a Bray-Curtis dissimilarity matrix, using the R package Vegan (Oksanen et al., 2020).

We used dnaPipeTE version 1.3 (Figure 4) (Goubert et al., 2015), installed with Trinity version 2.5.1 (Grabherr et al., 2013), Tandem Repeats Finder version 4.09 (Benson, 1999) and RepeatMasker version 4.1.2 (RRID:SCR_012954; RepeatMasker Home Page) complemented with the RepeatMasker libraries, and using RMBlastn version 2.11.0, Blastn version 2.11.0 as search engines. Given that dnaPipeTE can only process single end reads, we used only the forward reads. We ran the pipeline on each sample with the following parameters: genome size = 2.6 Gb, coverage = 0.1, sample number = 2 and species = Camelidae.

Differences in relative copy numbers

All the subsequent analysis were carried out in R (R Core Team, 2020). We compared the estimated abundances between species using an ANOVA on each cluster. The null hypothesis is that all four species have the same copy number; and the alternative hypothesis states that at least one species has a different copy number. The equation can be written as:

$$y_{ij} = \beta_0 + \beta_i + \varepsilon_{ij}$$

Where the abundance of a cluster y_{ij} in the species i and sample j is explained by an intercept β_0 and a species-specific slope β_i. The error was denoted as ε_{ij}.

Differences in relative age

Additionally, we compared the relative distances from reads to contigs from dnaPipeTE as a proxy for relative age. Only four samples per species were used, given the high processing time required. We used gamlss (Rigby & Stasinopoulos, 2005) to run the following generalized linear model (one model per contig), with a logistic link and a zero-inflated beta distribution:

$$y_{ij} = g^{-1}(\beta_0 + \beta_i + \varepsilon_{ij})$$

Where y_{ij} is the proportion of divergence between a read j from species i and the contig it maps to; β_0 and β_i are the intercept and the slope of species i respectively; ε_{ij} is the error and g^{-1} is the logistic link. The models (with or without species as a factor) were ranked according to the corrected Akaike information criterion (AICc). We followed that with a post-hoc analysis —using emmeans (Lenth, 2021)— for the contigs where $2 + AICc_{full} \leq AICc_{null}$, meaning that the best model includes species as a factor.

Genetic exchange of transposons

If there has been genetic exchange of TEs between alpacas and llamas due to introgression (or, alternatively, horizontal transfer), the corresponding reads of the two species should be more similar than expected. That similarity can be measured as the link density in the RepeatExplorer networks. Hence, we clustered samples according to the link density. We used the ratio of expected to observed links between samples as a dissimilarity metric, as it should be independent of the number of reads per sample. The samples were clustered using UPGMA (unweighted pair group method with arithmetic mean). We made trees for the RepeatExplorer networks with significant p-values in the ANOVA test.

Results and discussion

Our study can be divided in three parts. First, we provided a description of the TE families present in SACs. Knowing that baseline, we explored if there are differences in copy number between and within species. Those differences do not imply a difference in the transposition rate, as a change in the rate of deletion can generate the same patterns. Second, we tested a possible connection between domestication and transposition rates by looking at the relative age of TEs. Older copies had more time to diverge neutrally, whereas newer copies are more self-similar. At last, we sought evidence of inter-species genetic exchange of TEs.

Transposon detection and annotation

Our first aim as to detect and annotate the TEs present in SACs, for which we used two different pipelines: RepeatExplorer and dnaPipeTE. The annotation of RepeatExplorer (Figure 5) shows 76 clusters -each cluster corresponds to a contig-. Most contigs have very low abundances, and the abundances do not change drastically across species (but clusters 2, 5, 6 and 7 do show noticeable differences in copy numbers). Many clusters, and particularly those with small counts, are not annotated. From the annotated clusters, satellites are the most abundant, followed by LINEs and then LTRs. Comparing to previous measurements, the alpaca genome VicPac3.1 shows LINEs as the most common TEs, followed by LTRs and then SINEs (Richardson et al., 2019).

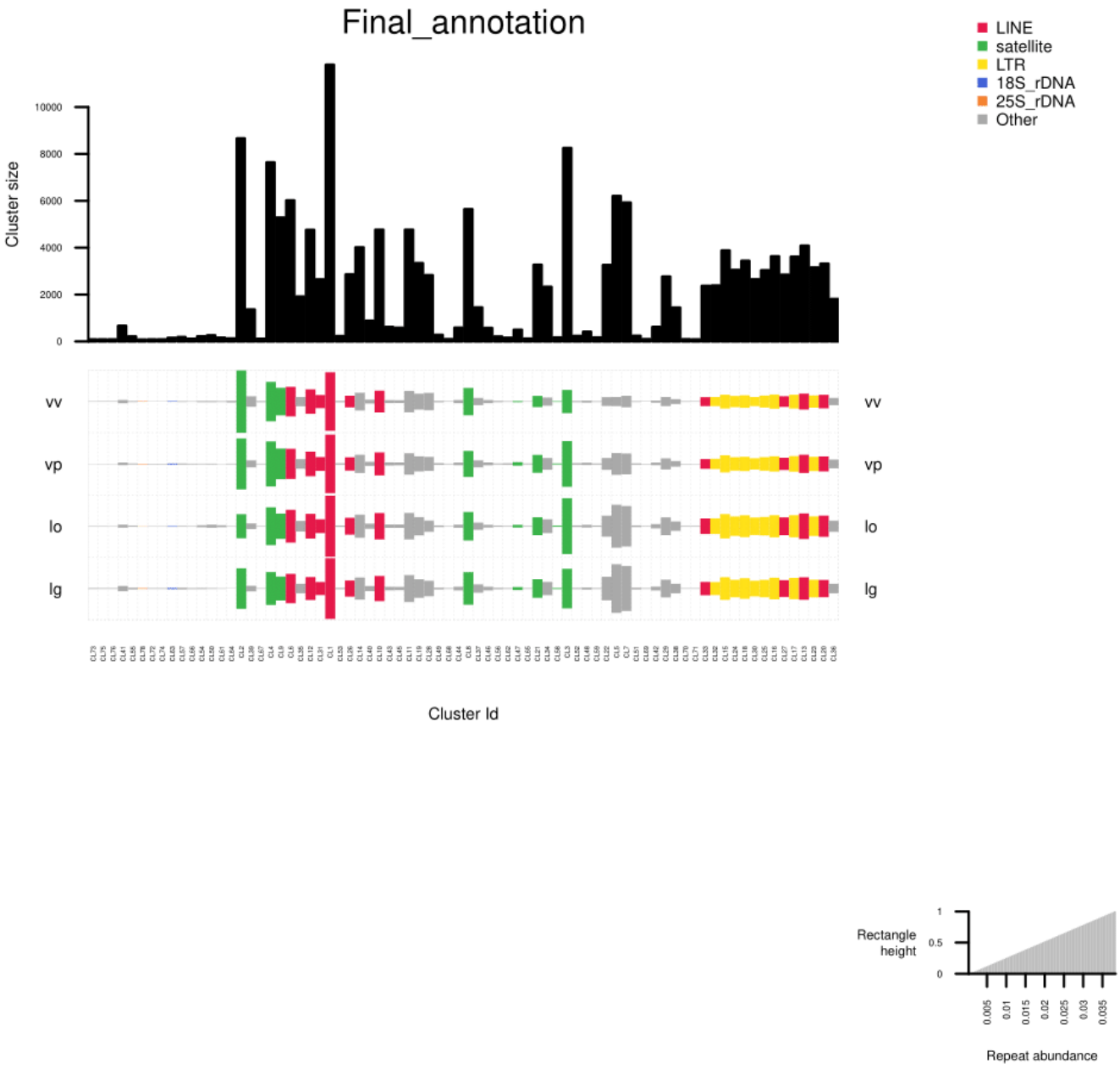
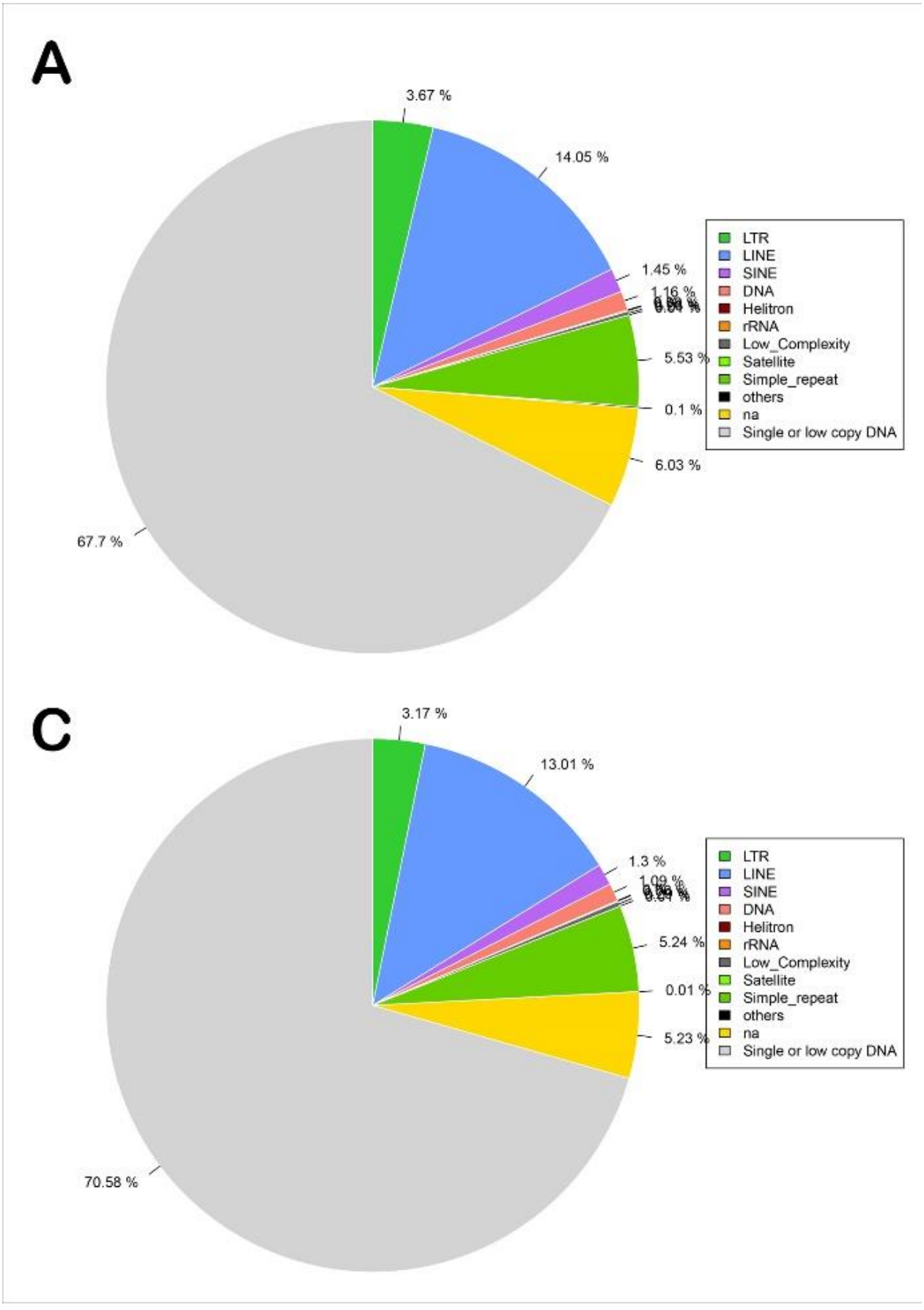
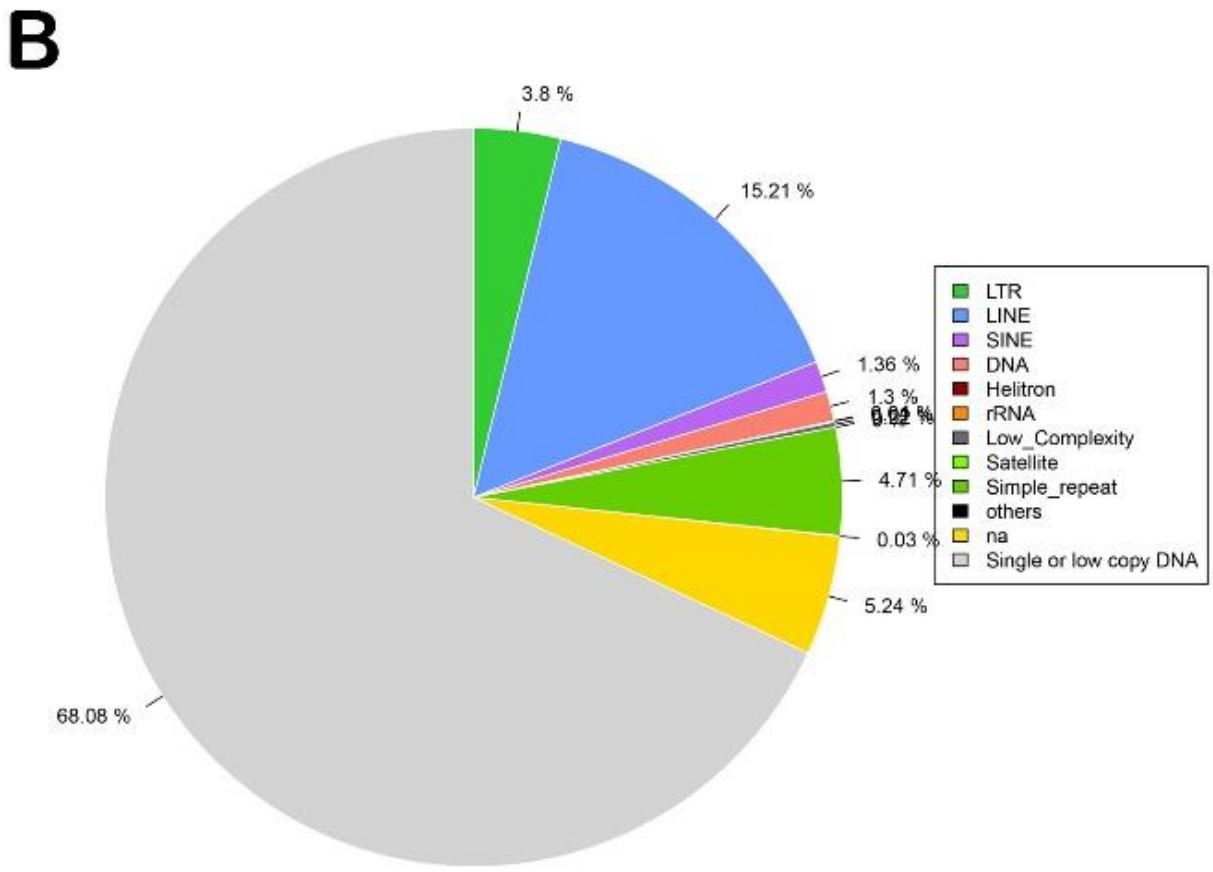


Figure 5. RepeatExplorer annotation per species. Top: Average number of reads (cluster size) per cluster. Each bar is a different cluster (the cluster id is at the bottom). Middle: Average number of reads per cluster per species. The bars correspond to clusters (the same as above), and the colours represent the annotation. Each row is one species. Labels: vv = vicuña (*V. vicugna*); vp = alpaca (*V. pacos*); lg = llama (*L. glama*); lo = guanaco (*L. guanicoe*). Bottom: A scale that maps repeat abundance (in the middle plot) with estimated proportion in the genome.

The dnaPipeTE annotation (Figure 6) also shows congruent proportions across species. LINEs are the most abundant TEs, followed by LTRs and simple repeats. In contrast to RepeatExplorer, there are almost no satellites detected. SINEs are also present in a small proportion, whereas they were not detected before. There is some ribosomal DNA detected by both pipelines, in small quantities. The overall trend of LINEs >> LTRs > SINEs coincides with the VicPac3.1 genome (Richardson et al., 2019); which also coincides with the proportions of the Bactrian and dromedary camels (Khalkhali-Evrigh et al., 2019; Zare, 2021). Whereas RepeatExplorer estimates that the proportion of the genome occupied by repetitive elements is smaller than 0.3 (Figure S1), dnaPipeTE estimates a proportion around 0.3 (Figures 6, S2). That difference can be explained by the variation in coverage: the higher coverage used in dnaPipeTE allows it to detect repeated elements with lower copy numbers. Previous estimations for Alpacas include 32.1% (Wu et al., 2014) and 33.5% (Richardson et al., 2019), in the same range as our results (Figure 6B). The proportions of TEs in old world camelids are also estimated to be around 30% (Khalkhali-Evrigh et al., 2019; Wu et al., 2014; Zare, 2021). Therefore, TE abundance is a conserved trait across all camelids.


A
3.67 %
14.05 %
1.45 %
1.16 %
5.53 %
0.1 %
6.03 %
67.7 %
LTR
LINE
SINE
DNA
Helitron
rRNA
Low_Complexity
Satellite
Simple_repeat
others
na
Single or low copy DNA
C
3.17 %
13.01 %
1.3 %
5.24 %
0.01 %
5.23 %
70.58 %
LTR
LINE
SINE
DNA
Helitron
rRNA
Low_Complexity
Satellite
Simple_repeat
others
na
Single or low copy DNA


B
3.8 %
15.21 %
1.36 %
1.3 %
4.71 %
0.03 %
5.24 %
68.08 %
LTR
LINE
SINE
DNA
Helitron
rRNA
Low_Complexity
Satellite
Simple_repeat
others
na
Single or low copy DNA

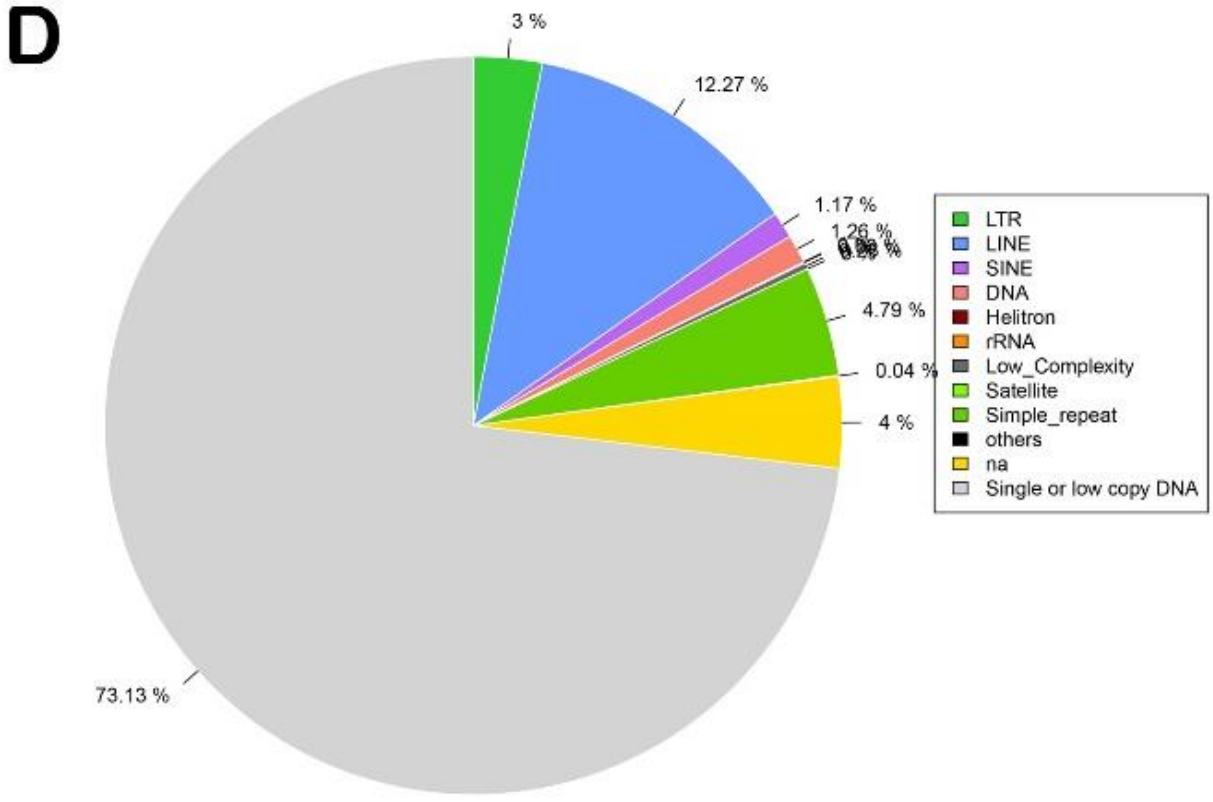

D
3 %
12.27 %
1.17 %
4.79 %
0.04 %
4 %
73.13 %
LTR
LINE
SINE
DNA
Helitron
rRNA
Low_Complexity
Satellite
Simple_repeat
others
na
Single or low copy DNA

Figure 6. Estimated genome proportion of repeated elements, using dnaPipeTE. We show a single sample per species. A) vicuña (*V. vicugna*); B) alpaca (*V. pacos*); C = llama (*L. glama*); D) guanaco (*L. guanicoe*)

Differences in relative copy numbers

Whereas most LINE and LTR clusters appear to have roughly the same copy number across samples and species, some satellites are more variable (Figure 7). If samples are ordered according to the similarity of copy numbers, they do not necessarily align by species or genus (Figure 7). Hence, we used an ANOVA to find clusters where the variation in copy number can be explained by the species (Table S1). 29 out of 76 clusters have a p-value smaller than 0.01. Given that we are comparing samples with low coverage, the differences could be caused by the stochastic sampling of reads. Therefore, we plotted the number of reads per cluster against the p-values (Figure 8). Most of the clusters with low p-values also have high counts, suggesting that the differences are not an artifact of the low coverage (Figure 8). Most clusters with low p-values are LTRs or are not annotated, and a few are LINEs or satellites. The within-species variation of satellites (Figure 7) could override the between-species variation. Lastly, we did a Non-metric Multi-dimensional Scaling (NMDS) to see if the species have characteristic quantities of copy numbers, when considering all the clusters at the same time. There is a clear division between the two genera (Figure 9), as well as between species within a genus. Two guanaco samples cluster more closely with llama samples. That could be explained by a mislabelling of the samples: ADMIXTURE analysis of the same samples showed some guanacos that had a majority of llama ancestry (Fan et al., 2020), however we cannot confirm if those individuals correspond to ours, as they used a different ID.

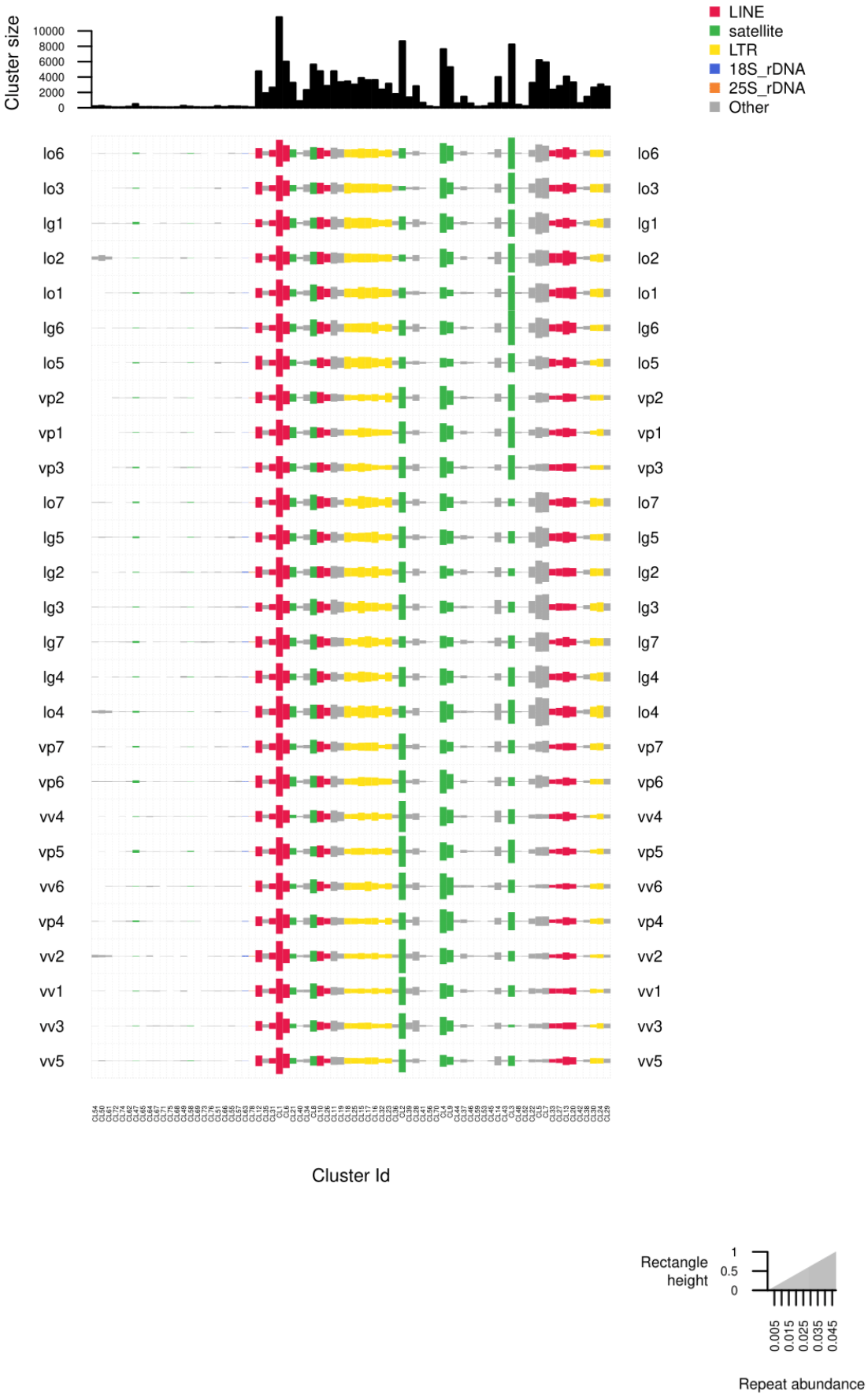

Cluster size
10000
8000
6000
4000
2000
0
LINE
satellite
LTR
18S_rDNA
25S_rDNA
Other
lo6
lo3
lg1
lo2
lo1
lg6
lo5
vp2
vp1
vp3
lo7
lg5
lg2
lg3
lg7
lg4
lo4
vp7
vp6
vv4
vp5
vv6
vp4
vv2
vv1
vv3
vv5
Cluster Id
Rectangle height
1
0.5
0
0.005
0.015
0.025
0.035
0.045
Repeat abundance

Figure 7. RepeatExplorer annotation per sample. Top: Number of reads (cluster size) per cluster. Each bar is a different cluster (the cluster id is at the bottom). Middle: Number of reads per cluster per individual. The bars correspond to clusters (the same as above), and the colours represent the annotation. Grey represents non annotated clusters. The rows represent individuals and are arranged according to similarity. Labels: vv = vicuña (*V. vicugna*); vp = alpaca (*V. pacos*); lg = llama (*L. glama*); lo = guanaco (*L. guanicoe*). Bottom: A scale that maps repeat abundance (in the middle plot) with estimated proportion in the genome.

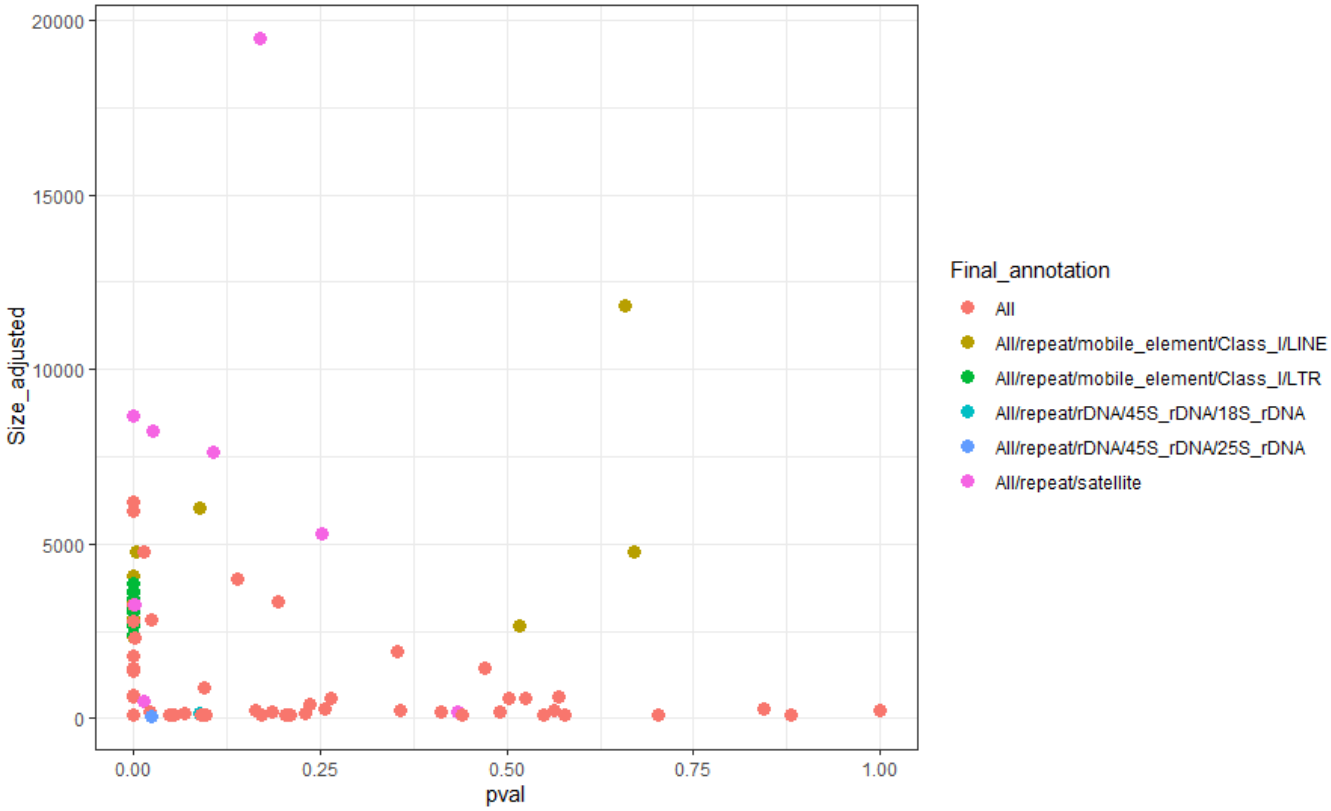


Figure 8. ANOVA p-values per cluster, against total number of reads per cluster. The colour represents the final annotation.

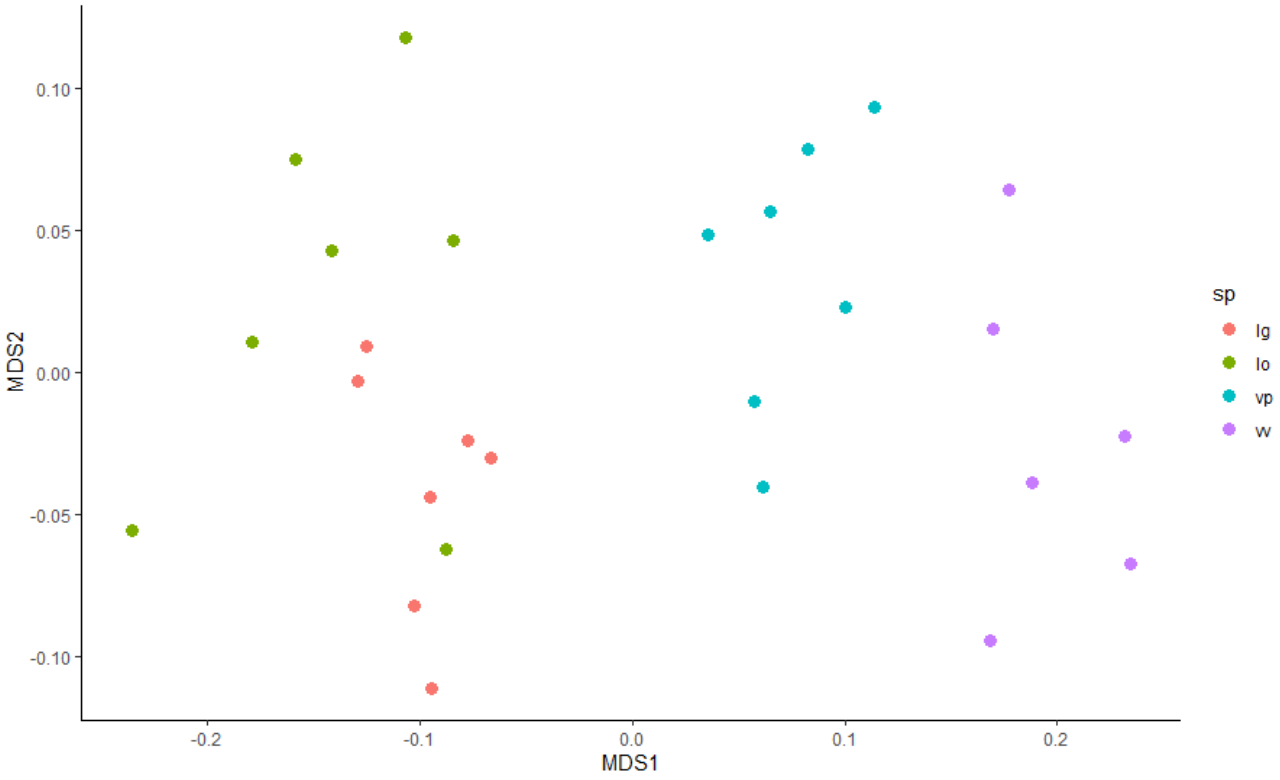


Figure 9. PCoA on cluster counts. Each colour represents a different species. Labels: vv = vicuña (*V. vicugna*); vp = alpaca (*V. pacos*); lg = llama (*L. glama*); lo = guanaco (*L. guanicoe*).

Differences in relative age

After events of transposition, TEs that evolve neutrally accumulate mutations, hence older transposons will have more divergence between copies compared to recent transposons. We compared this divergence between species, to see if some TEs are younger, and if that can be connected to the evolutionary history of SACs, particularly to domestication. The overall divergence distribution has the same shape for all TE families across species (Figure 10), suggesting that most duplication events predate the genus split. We ran a glm on each family to detect the more recent differences (Table 1). Most TEs show inter-species differences in the relative age. We followed that with a post-hock to find out which species have younger TEs. For most LTRs, LINEs and SINE, the genus explains most of the difference: *Llama spp.* have smaller slopes than *Vicugna spp.* (Figure 11), meaning that llama and guanaco have younger transposons. Within each genus, guanacos and alpacas have younger transposons

than llamas and vicuñas, respectively (Figure 11). Therefore, domestication was not the main process driving TE evolution in SACs. Looking at each TE family, all LINEs and SINEs have the same pattern -with the exception of LINE/CR1-, whereas LTRs vary more. This suggests that the rate of transposition is controlled in a way that is mainly dependent on the mechanism of transposition. SINEs have the same pattern as LINEs because they hijack the machinery of LINEs to transpose (Wells & Feschotte, 2020). The differing patterns of relative age in LTRs (Figure 11) and the differences in LTR copy numbers (Figure 8) suggest that different LTRs retain various levels of activity, but are controlled in different ways, or affected by different processes. Selection and population dynamics can also interact with the rates of transposition, and ideally rates of transposition should be estimated with selection signatures simultaneously (Bourgeois & Boissinot, 2019). Those interactions can only be unraveled with an accurate mapping of the TEs, which would allow the reconstruction of an allele frequency spectrum (Bourgeois & Boissinot, 2019). We suggest that specific to combinations of species and LTRs could reflect environmental effects on TE regulation.

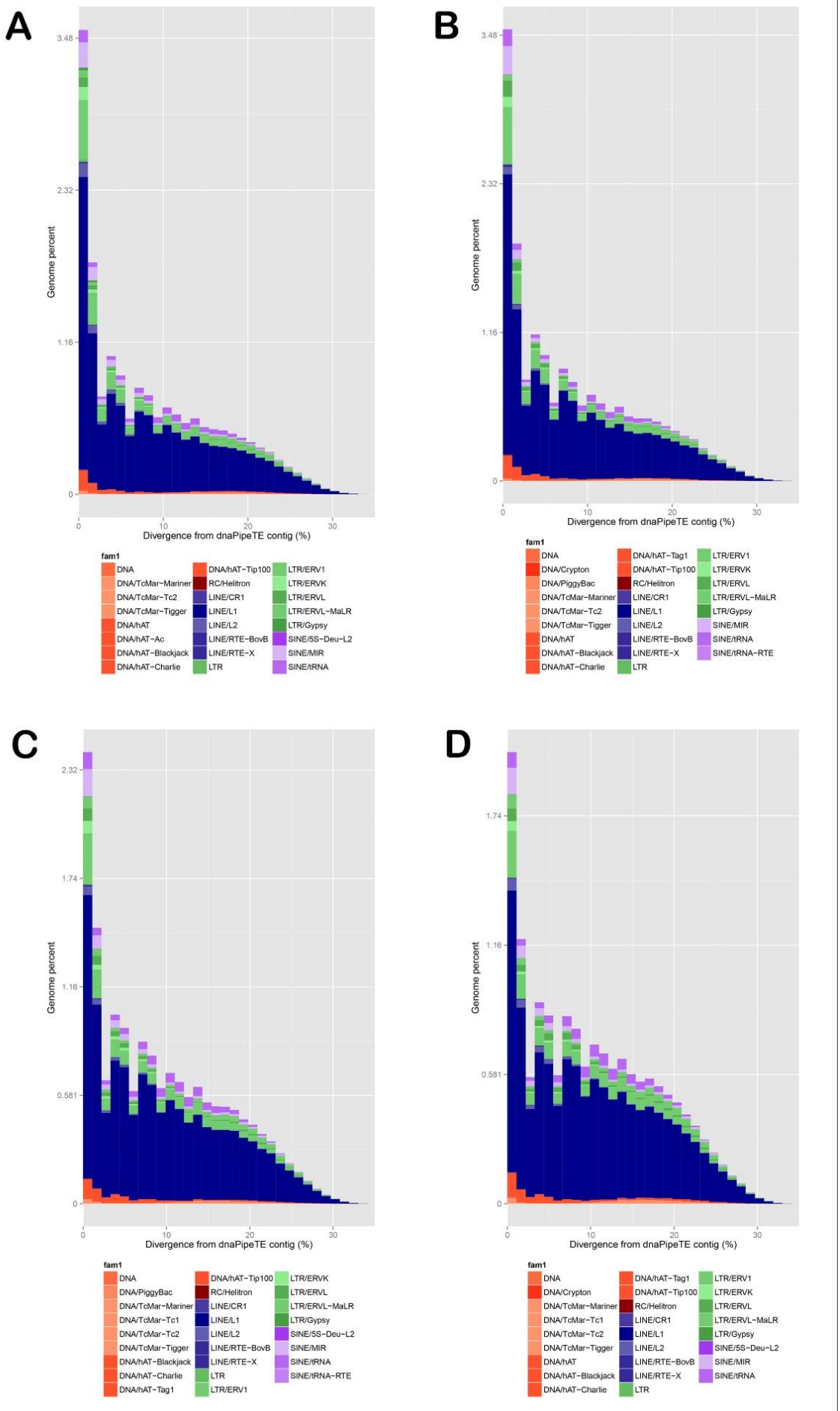

A
B
C
D
Genome percent
Divergence from dnaPipeTE contig (%)
fam1
DNA
DNA/Crypton
DNA/PiggyBac
DNA/TcMar-Mariner
DNA/TcMar-Tc1
DNA/TcMar-Tc2
DNA/TcMar-Tigger
DNA/hAT
DNA/hAT-Ac
DNA/hAT-Blackjack
DNA/hAT-Charlie
DNA/hAT-Tag1
DNA/hAT-Tip100
RC/Helitron
LINE/CR1
LINE/L1
LINE/L2
LINE/RTE-BovB
LINE/RTE-X
LTR
LTR/ERV1
LTR/ERVK
LTR/ERVL
LTR/ERVL-MaLR
LTR/Gypsy
SINE/5S-Deu-L2
SINE/MIR
SINE/tRNA
SINE/tRNA-RTE

Figure 10. dnaPipeTE landscape plots. For each read in a sample, the x axis is the proportion of differences when compared to the contig it maps to. The y axis is the number of reads with that specific proportions (a histogram). The corresponding annotation is plotted at the left of each histogram. Only four samples -one per species- are shown. A) llama (*L. glama*); B) guanaco (*L. guanicoe*); C) alpaca (*V. pacos*); D) vicuña (*V. vicugna*).

Table 1. GLMs on the relative age of repetitive elements across species. In all cases, the full model only includes species as a factor. Delta is calculated as AICc(null) - AICc(full). Hence, a Delta > 2 implies that the full model is the best, and a Delta < -2 implies that the null model is the best.

Annotation	**AICc full**	**AICc null**	**Delta**
DNA	-266324.9	-265258.2	1066.7
LINE/CR1	-4038	-4033.9	4.1
LINE/L1	-6009419	-5986986	22433
LINE/L2	-135661.9	-135318.6	343.3
LINE/RTE-BovB	-243.4	-248.2	-4.8
LINE/RTE-X	-23563	-23396.3	166.7
LTR	-9168	-9161.4	6.6
LTR/ERV1	-799169.6	-795053.9	4115.7
LTR/ERVK	-70537.4	-70303.2	234.2
LTR/ERVL-MaLR	-196114.7	-195497.3	617.4
LTR/ERVL	-155787	-155141.3	645.7
LTR/Gypsy	-8851.1	-8755.6	95.5
Other	-1350.4	-1351.1	-0.7
SINE	-658679.2	-656624.6	2054.6

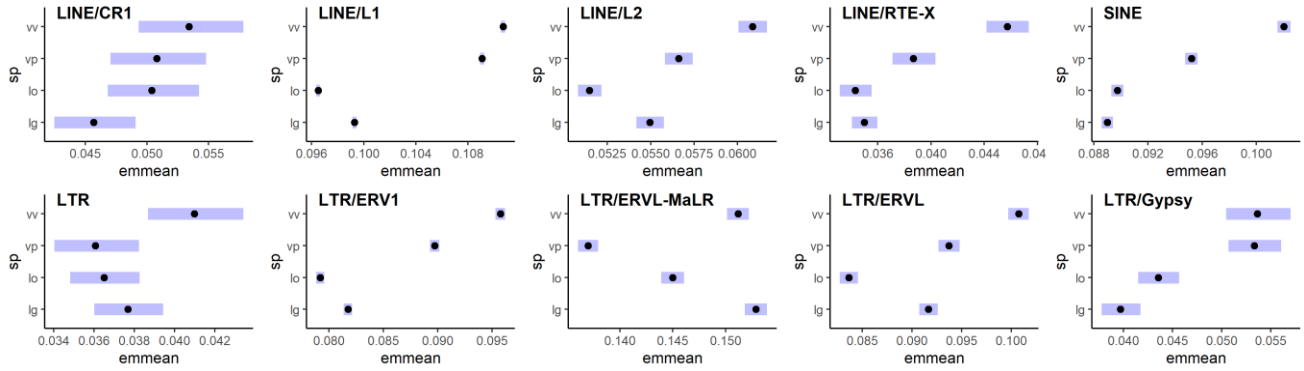


Figure 11. Estimated slopes of significant repeated elements. Y axis: species. X axis: estimated marginal means (dots), and their standard error (bars). Species: vv = vicuña (*V. vicugna*); vp = alpaca (*V. pacos*); lg = llama (*L. glama*); lo = guanaco (*L. guanicoe*).

Genetic exchange of transposons

The ongoing admixture between domesticated SACs could lead to between species transfer of TEs: whereas most TEs will be lost due to drift in a few generations, some will fixate. This would lead to a higher-than-expected similarity between TEs of llamas and alpacas. We used the RepeatExplorer networks to test this hypothesis: the more similar two samples, the more links they will share. Therefore, if we cluster individuals according to this (dis)similarity, mixed groups of llamas and alpacas would be evidence of genetic exchange of TEs. Most clusters do not show a pattern at all (results not shown), which implies that those TEs have not had enough time to diverge neutrally from each other. A few networks showed a division among genus and species, like in Figure 12 (A), and a satellite showed alpacas clustered with llamas (Figure 12 B). Therefore, at least some satellites are transferred from llamas to alpacas. There are two possible explanations for the transfer of only specific TE families: Old introgression events followed by loss of deleterious/neutral TEs (and transposition of active copies, if any); or horizontal gene transfer of active TEs. It should be possible to differentiate the two by comparing TE similarity and the overall segment similarity if the copies are mapped to a reference genome. The satellite network (Figure 13) has a ball-like topology, which is expected for short contigs that have repeated motifs (Novák et al., 2013). Each cluster shows a different behaviour, even when they have the same annotation, showing again that TE regulation depends on a combination of the

transposition mechanism and other processes. The probability of horizontal transmission is influenced by both the TE transposition mechanism and the host species (Bourgeois & Boissinot, 2019).

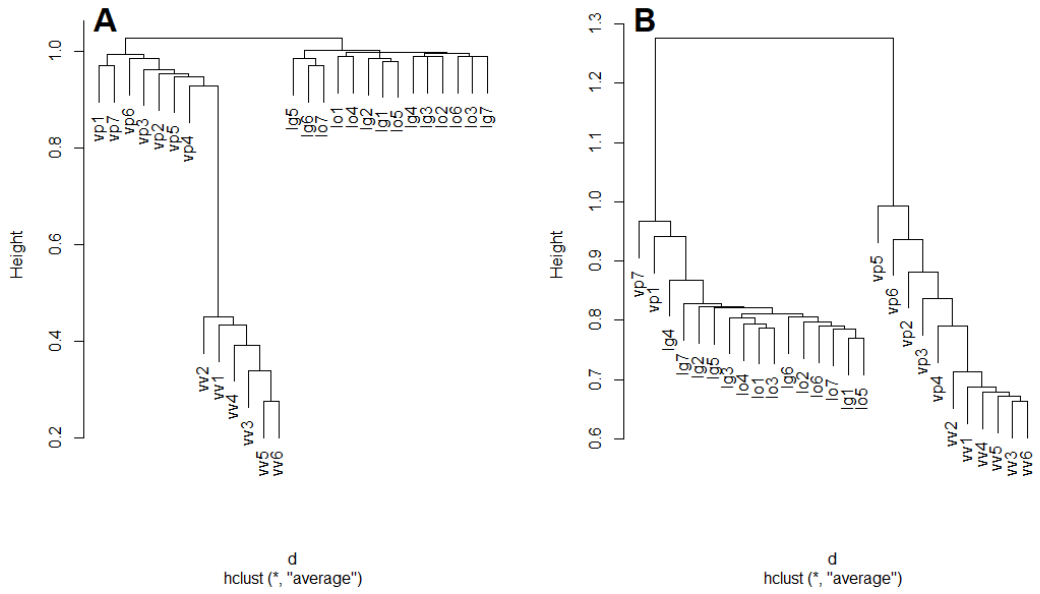


Figure 12. UPGMA tree of clusters 5 (A) and 2 (B). Each branch is one sample. Labels: vv = vicuña (*V. vicugna*); vp = alpaca (*V. pacos*); lg = llama (*L. glama*); lo = guanaco (*L. guanicoe*).

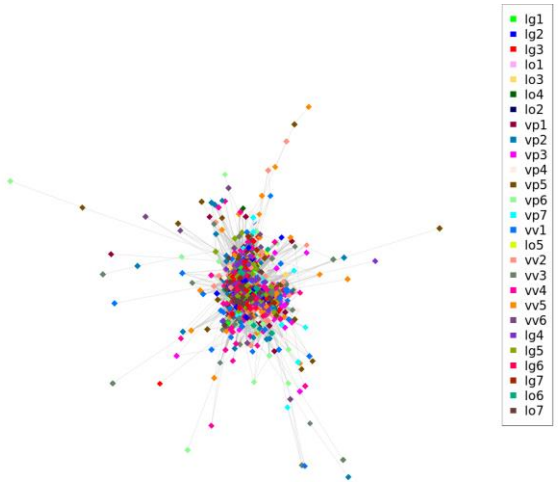


Figure 13. RepeatExplorer network of cluster 2. Each vertex is a read, and two reads are connected if they can be mutually aligned. Colors represent samples. Species: vv = vicuña (*V. vicugna*); vp = alpaca (*V. pacos*); lg = llama (*L. glama*); lo = guanaco (*L. guanicoe*).

Conclusions

TE composition and abundance in SACs is congruent with the composition of old-world camelids. The genome proportion covered by TEs is approximately 0.3 for all SACs, and the majority of TEs are LINEs, which are more common that SINEs and LTRs. Whereas LINE and SINE relative ages reflect the phylogenetic history of SACs, LTR recent evolution depends on a combination of species and TE family, and they show inter-species variation of copy numbers. Domestication is not a driver of global transposition rates, but in SACs it enabled genetic exchange of specific repetitive elements, either via introgression or horizontal transfer.

References

Andrews, S. (2015). *FastQC*. https://qubeshub.org/resources/fastqc

Benson, G. (1999). Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Research*, *27*(2), 573–580. https://doi.org/10.1093/nar/27.2.573

Bourgeois, Y., & Boissinot, S. (2019). On the population dynamics of junk: A review on the population genomics of transposable elements. *Genes*, *10*(6). https://doi.org/10.3390/genes10060419

Branco, M. R., & Chuong, E. B. (2020). Crossroads between transposons and gene regulation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1795), 2–5. https://doi.org/10.1098/rstb.2019.0330

Campbell, S., Aswad, A., & Katzourakis, A. (2017). Disentangling the origins of virophages and polintons. *Current Opinion in Virology*, *25*, 59–65. https://doi.org/10.1016/j.coviro.2017.07.011

Fan, R., Gu, Z., Guang, X., Marín, J. C., Varas, V., González, B. A., Wheeler, J. C., Hu, Y., Li, E., Sun, X., Yang, X., Zhang, C., Gao, W., He, J., Munch, K., Corbett-Detig, R., Barbato, M., Pan, S., Zhan, X., … Dong, C. (2020). Genomic analysis of the domestication and post-Spanish conquest evolution of the llama and alpaca. *Genome Biology*, *21*(1), 1–26. https://doi.org/10.1186/s13059-020-02080-6

Gilbert, C., & Feschotte, C. (2018). Horizontal acquisition of transposable elements and viral sequences: patterns and consequences. *Current Opinion in Genetics and Development*, *49*(October 2017), 15–24. https://doi.org/10.1016/j.gde.2018.02.007

Goubert, C., Modolo, L., Vieira, C., Moro, C. V., Mavingui, P., & Boulesteix, M. (2015). De novo assembly and annotation of the Asian tiger mosquito (Aedesalbopictus) repeatome with dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever mosquito (Aedes aegypti). *Genome Biology and Evolution*, *7*(4), 1192–1205. https://doi.org/10.1093/gbe/evv050

Grabherr, M. G. ., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I.,

Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., Palma, F. di, W., B., Friedman, N., & Regev, A. (2013). Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature Biotechnology*, *29*(7), 644–652. https://doi.org/10.1038/nbt.1883.Trinity

Guio, L., & González, J. (2019). New insights on the evolution of genome content: population dynamics of transposable elements in flies and humans. *Evolutionary Genomics*, 505–530.

Jangam, D., Feschotte, C., & Betrán, E. (2017). Transposable Element Domestication As an Adaptation to Evolutionary Conflicts. *Trends in Genetics*, *33*(11), 817–831. https://doi.org/10.1016/j.tig.2017.07.011

Khalkhali-Evrigh, R., Hedayat-Evrigh, N., Hafezian, S. H., Farhadi, A., & Bakhtiarizadeh, M. R. (2019). Genome-wide identification of microsatellites and transposable elements in the dromedary camel genome using whole-genome sequencing data. *Frontiers in Genetics*, *10*(JUL), 1–10. https://doi.org/10.3389/fgene.2019.00692

Koonin, E. V., & Krupovic, M. (2017). Polintons, virophages and transpovirons: a tangled web linking viruses, transposons and immunity. *Current Opinion in Virology*, *25*(June), 7–15. https://doi.org/10.1016/j.coviro.2017.06.008

Koonin, E. V., Krupovic, M., & Yutin, N. (2015). Evolution of double-stranded DNA viruses of eukaryotes: From bacteriophages to transposons to giant viruses. *Annals of the New York Academy of Sciences*, *1341*(1), 10–24. https://doi.org/10.1111/nyas.12728

Le Rouzic, A., & Deceliere, G. (2005). Models of the population genetics of transposable elements. *Genetical Research*, *85*(3), 171–181. https://doi.org/10.1017/S0016672305007585

Lenth, R. V. (2021). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. https://cran.r-project.org/package=emmeans

Makałowski, W., Gotea, V., Pande, A., & Makałowska, I. (2019). Transposable elements: Classification, identification, and their use as a tool for comparative

genomics. In *Evolutionary Genomics* (pp. 177–207). Springer.

Neumann, P., Novák, P., Hoštáková, N., & MacAs, J. (2019). Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. *Mobile DNA*, *10*(1), 1–17. https://doi.org/10.1186/s13100-018-0144-1

Novák, P., Neumann, P., & Macas, J. (2010). Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics*, *11*, 1–12. https://doi.org/10.1186/1471-2105-11-378

Novák, P., Neumann, P., & Macas, J. (2020). Global analysis of repetitive DNA from unassembled sequence reads using RepeatExplorer2. *Nature Protocols*, *15*(11), 3745–3776. https://doi.org/10.1038/s41596-020-0400-y

Novák, P., Neumann, P., Pech, J., Steinhaisl, J., & MacAs, J. (2013). RepeatExplorer: A Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*, *29*(6), 792–793. https://doi.org/10.1093/bioinformatics/btt054

Novák, P., Robledillo, L. Á., Koblížková, A., Vrbová, I., Neumann, P., & Macas, J. (2017). TAREAN: A computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nucleic Acids Research*, *45*(12). https://doi.org/10.1093/nar/gkx257

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., & Wagner, H. (2020). *vegan: Community Ecology Package*. https://cran.r-project.org/package=vegan

Percharde, M., Sultana, T., & Ramalho-Santos, M. (2020). What Doesn't Kill You Makes You Stronger: Transposons as Dual Players in Chromatin Regulation and Genomic Variation. *BioEssays*, *42*(4), 1–10. https://doi.org/10.1002/bies.201900232

Piégu, B., Bire, S., Arensburger, P., & Bigot, Y. (2015). A survey of transposable element classification systems - A call for a fundamental update to meet the challenge of their diversity and complexity. *Molecular Phylogenetics and*

Evolution, *86*, 90–109. https://doi.org/10.1016/j.ympev.2015.03.009

Platt, R. N., Vandewege, M. W., & Ray, D. A. (2018). Mammalian transposable elements and their impacts on genome evolution. *Chromosome Research*, *26*(1–2), 25–43. https://doi.org/10.1007/s10577-017-9570-z

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. https://www.r-project.org/

Richardson, M. F., Munyard, K., Croft, L. J., Allnutt, T. R., Jackling, F., Alshanbari, F., Jevit, M., Wright, G. A., Cransberg, R., Tibary, A., Perelman, P., Appleton, B., & Raudsepp, T. (2019). Chromosome-level alpaca reference genome VicPac3.1 improves genomic insight into the biology of new world camelids. *Frontiers in Genetics*, *10*(JUN), 1–15. https://doi.org/10.3389/fgene.2019.00586

Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape,(with discussion). *Applied Statistics*, *54*, 507–554.

Shapiro, J. A. (2017). Biological action in Read-Write genome evolution. *Interface Focus*, *7*(5). https://doi.org/10.1098/rsfs.2016.0115

Sultana, T., Zamborlini, A., Cristofari, G., & Lesage, P. (2017). Integration site selection by retroviruses and transposable elements in eukaryotes. *Nature Reviews Genetics*, *18*(5), 292–308. https://doi.org/10.1038/nrg.2017.7

Wells, J. N., & Feschotte, C. (2020). A Field Guide to Eukaryotic Transposable Elements. *Annual Review of Genetics*, *54*, 539–561. https://doi.org/10.1146/annurev-genet-040620-022145

Wheeler, J. C. (2012). South American camelids - past, present and future. *Journal of Camelid Science*, *5*, 1–24.

Wu, H., Guang, X., Al-Fageeh, M. B., Cao, J., Pan, S., Zhou, H., Zhang, L., Abutarboush, M. H., Xing, Y., Xie, Z., Alshanqeeti, A. S., Zhang, Y., Yao, Q., Al-Shomrani, B. M., Zhang, D., Li, J., Manee, M. M., Yang, Z., Yang, L., … Wang, J. (2014). Camelid genomes reveal evolution and adaptation to desert environments. *Nature Communications*, *5*. https://doi.org/10.1038/ncomms6188

Zare, N. (2021). Identification and investigation of transposable elements in the

Iranian bactrian camel genomes. *Agricultural Biotechnology Journal*, *12*(4), 43–60. https://doi.org/10.22103/JAB.2020.15642.1218

Zeder, M. A. (2017). Domestication as a model system for the extended evolutionary synthesis. *Interface Focus*, *7*(5). https://doi.org/10.1098/rsfs.2016.0133

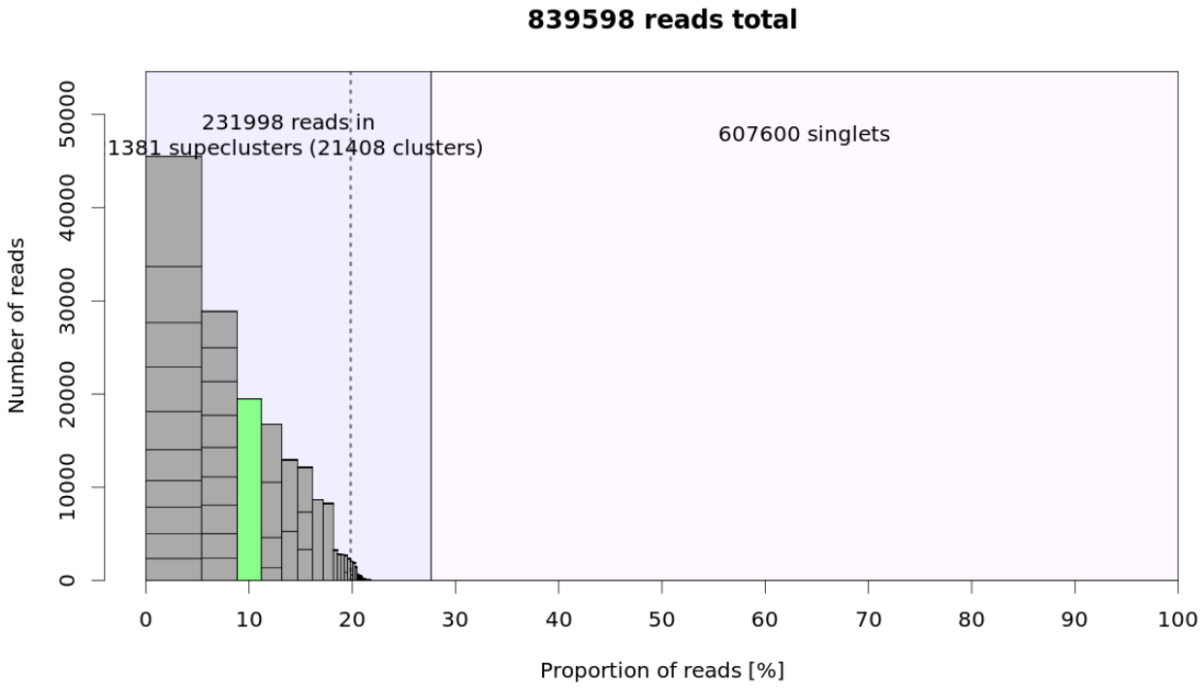


Figure S1. Clustering results of RepeatExplorer. The bars are superclusters, and the divisions within each bar mark each cluster. The y axis represents the number of reads of a (super)cluster, and the X axis is the proportion of reads that belongs to a cluster. The green bar is the cluster affected by the satellite filtering. The blue region delimits the reads that clustered, and thus are repeated across the genome, from the reads that are not repeated, in pink.

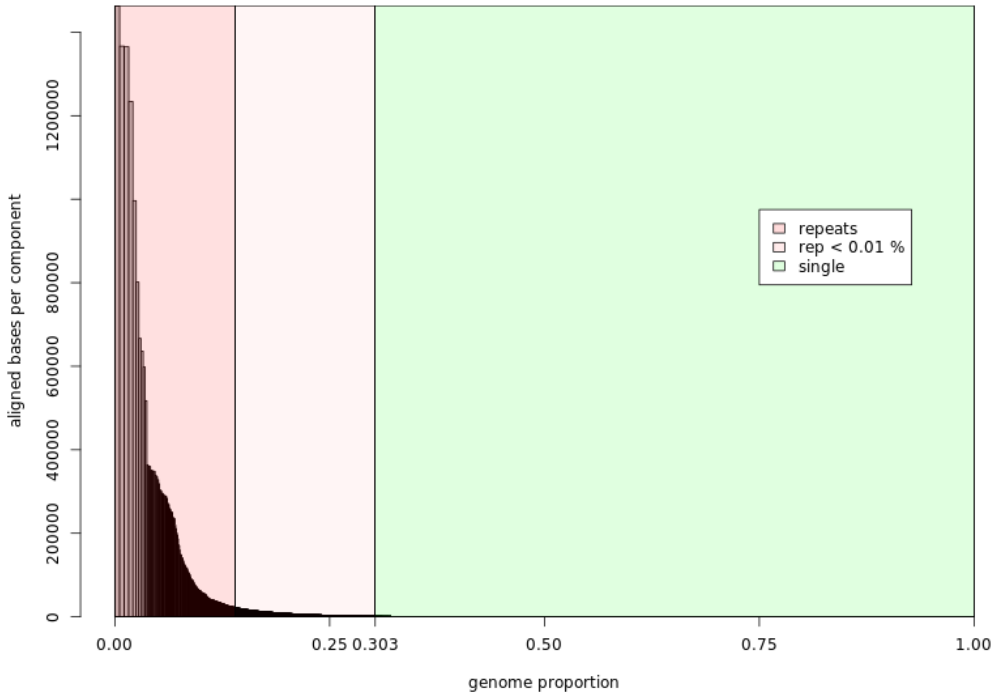


Figure S2. dnaPipeTE alignment results from a *L. glama* sample. Each bar represents a contig and its height corresponds to the number of bases that mapped to it. The x axis shows the genome proportion of each contig, and the colors delimit the fraction of the genome occupied by repetitive elements.

Table S1. p-values of ANOVA test. Rows were sorted according to the p-value. All tests have degrees of freedom = 3, 23.

Cluster	Supercluster	Size_adjusted	Final_annotation	F	pval
5	3	6210	All	59.24435	5.65E-11
7	3	5926	All	35.86198	7.68E-09
22	3	3263	All	28.93181	5.54E-08
38	14	1447	All	25.2529	1.85E-07
29	11	2770	All	20.37821	1.14E-06

18	2	3444	All/repeat/mobile_element/Class_I/LTR	18.28227	2.74E-06
32	2	2388	All/repeat/mobile_element/Class_I/LTR	17.05838	4.74E-06
15	2	3884	All/repeat/mobile_element/Class_I/LTR	16.85173	5.21E-06
2	6	8666	All/repeat/satellite	16.84392	5.23E-06
27	1	2846	All/repeat/mobile_element/Class_I/LINE	16.23198	6.95E-06
25	2	3038	All/repeat/mobile_element/Class_I/LTR	16.20502	7.04E-06
64	39	128	All	16.02813	7.66E-06
30	2	2657	All/repeat/mobile_element/Class_I/LTR	14.8674	1.35E-05
16	2	3634	All/repeat/mobile_element/Class_I/LTR	12.20212	5.56E-05
17	2	3619	All/repeat/mobile_element/Class_I/LTR	12.03185	6.13E-05
13	1	4088	All/repeat/mobile_element/Class_I/LINE	12.0227	6.16E-05
26	1	2861	All/repeat/mobile_element/Class_I/LINE	11.80147	6.99E-05
23	2	3159	All/repeat/mobile_element/Class_I/LTR	11.47988	8.43E-05
39	3	1365	All	11.38639	8.90E-05
24	2	3050	All/repeat/mobile_element/Class_I/LTR	10.54234	0.000148
33	1	2371	All/repeat/mobile_element/Class_I/LINE	10.50118	0.000152
20	1	3320	All/repeat/mobile_element/Class_I/LINE	8.66725	0.000497

42	14	623	All	7.947719	0.000819
36	12	1812	All	7.212204	0.001397
41	17	670	All	7.180058	0.001431
21	9	3271	All/repeat/satellite	6.86771	0.001809
34	13	2331	All	6.458583	0.002478
10	1	4771	All/repeat/mobile_element/Class_I/LINE	5.646647	0.004742
47	22	499	All/repeat/satellite	4.360901	0.014293
11	5	4770	All	4.355401	0.014364
57	32	192	All	3.852642	0.022729
78	53	83	All/repeat/rDNA/45S_rDNA/25S_rDNA	3.784102	0.024227
28	10	2827	All	3.757443	0.024838
3	7	8253	All/repeat/satellite	3.685434	0.026572
67	42	117	All	3.038709	0.049476
74	49	88	All	2.950416	0.053974
62	37	161	All	2.712428	0.068415
63	38	155	All/repeat/rDNA/45S_rDNA/18S_rDNA	2.450719	0.089166
6	1	6023	All/repeat/mobile_element/Class_I/LINE	2.43963	0.090181

72	47	90	All	2.421157	0.0919
40	12	883	All	2.393852	0.094503
73	48	89	All	2.355955	0.098247
4	4	7643	All/repeat/satellite	2.268727	0.107471
14	5	4018	All	2.015748	0.139764
51	26	241	All	1.867542	0.163274
8	8	19469	All/repeat/satellite	1.83128	0.169632
70	45	98	All	1.813499	0.172843
55	30	213	All	1.742026	0.186399
19	5	3342	All	1.703696	0.194118
66	41	117	All	1.652212	0.205009
76	51	88	All	1.627496	0.210459
61	36	162	All	1.545836	0.229547
48	23	416	All	1.515469	0.237089
9	4	5291	All/repeat/satellite	1.456919	0.252352
50	25	267	All	1.438627	0.257322
46	21	579	All	1.410948	0.265032

35	15	1918	All	1.140092	0.35382
52	27	235	All	1.130675	0.357386
56	31	211	All	0.996807	0.411979
58	33	175	All/repeat/satellite	0.948027	0.433774
65	40	122	All	0.934237	0.440128
37	16	1453	All	0.869085	0.471322
59	34	173	All	0.829928	0.491022
45	20	586	All	0.806787	0.503006
31	1	2652	All/repeat/mobile_element/Class_I/LINE	0.77984	0.517285
44	19	591	All	0.767015	0.524203
75	50	88	All	0.721532	0.549379
54	29	219	All	0.697354	0.563169
43	18	622	All	0.685337	0.570128
69	44	100	All	0.672176	0.577829
1	1	11808	All/repeat/mobile_element/Class_I/LINE	0.542065	0.658339
12	1	4761	All/repeat/mobile_element/Class_I/LINE	0.524921	0.669513
71	46	91	All	0.475568	0.702347

49	24	281	All	0.274394	0.843236
68	43	100	All	0.220668	0.88104
53	28	228	All	0.008201	0.998949