Phylogenetic signal of genomic repeat abundances can be distorted by random homoplasy: a case study from hominid primates

MARÍA MARTÍN-PECIÑA[1], FRANCISCO J. RUIZ-RUANO[1],
JUAN PEDRO M. CAMACHO1[*] AND STEVEN DODSWORTH[2,3*]

[1]*Departamento de Genética, Facultad de Ciencias, Universidad de Granada, 18071 Granada, Spain*
[2]*School of Biological and Chemical Sciences, Queen Mary University of London, London E1 4NS, UK*
[3]*School of Life Sciences, University of Bedfordshire, University Square, Luton LU1 3JU, UK*

Received 3 January 2018; revised 4 July 2018; accepted for publication 14 September 2018

The genomic abundance of different types of repetitive DNA elements contains a phylogenetic signal useful for inferring the evolutionary history of different groups of organisms. Here we test the reliability of this approach using the Hominidae family of primates, whose consensus phylogeny is well accepted. We used the software RepeatExplorer to identify the different repetitive DNA clusters and quantify their abundances. With these data, we performed phylogenetic analyses by maximum parsimony, including one, two or three individuals per species, technical replicates, and including or discarding two clusters of repetitive elements (i.e. a satellite DNA and an endogenous retrovirus) that generated random homoplasy, because they were abundant in *Pan* and *Gorilla* but almost absent in *Homo* and *Pongo*. The only phylogenetic tree congruent with the accepted topology for hominids, thus coinciding with that obtained from the mitogenomes of the same individuals, was the one built after filtering out the libraries for the two homoplasious clusters and using three individuals per species. Our results suggest some caution in the use of repeat abundance for phylogenetic studies, because some element abundances are homoplasious, which severely distorts the phylogenetic signal owing to their differential amplification among evolutionary lineages.

ADDITIONAL KEYWORDS: Hominidae – homoplasy – inter-individual variation – phylogenetics – repetitive DNA.

INTRODUCTION

With the rise of high-throughput sequencing technologies, there has been an intersection between previously disparate fields of cytogenetics/genomics and phylogenetics. There are many approaches that seek to use genome-scale data for phylogenetic inference (often termed 'phylogenomics') and usually aim to reduce the genome complexity to something manageable for phylogenetic purposes. Additionally, such data are very useful for characterizing repeats and other markers for efficiently producing cytogenetic probes. The simplest method is one of 'genome skimming' *sensu* Straub *et al.* (2012), whereby whole-genome shotgun sequencing is performed but at a very low depth of coverage (< 1× genome coverage and perhaps < 0.1×). These datasets consist primarily of those sequences that are in high abundance, either in the genome itself or within the organism; this includes, predominantly, the high-copy organellar genome sequences (plastome in plants or mitogenome) but also those sequences that are in high copy in the nuclear genome. Amongst these high-copy nuclear sequences are mainly repetitive elements, an array of different types of repeat sequences, which include satellite (tandem) repeats, and transposable elements (TEs), such as retroelements (class I TEs) and DNA transposons (class II TEs). Often these data are discarded by researchers focusing on phylogenetics with such datasets, who instead use only the reconstructed organellar genomes (e.g. Guschanski *et al.*, 2013; Richter *et al.*, 2015; Timmermans *et al.*, 2016; Ren *et al.*, 2017).

[*]Corresponding authors. E-mail: jpmcamac@ugr.es; steven.dodsworth@beds.ac.uk

The importance of repetitive DNA abundance as a marker for the phylogenetic history of species has been increasingly explored (e.g. Ricci *et al.*, 2013; Sveinsson *et al.*, 2013; Cai *et al.*, 2014). Several recent studies have shown that genomic repeat abundance, rather than the sequence itself, can be used as an informative character for phylogenetic inference (Novák *et al.*, 2014; Dodsworth *et al.*, 2015, 2016a, b; Usai *et al.*, 2017). Using a recently developed pipeline for *de novo* repeat analysis from low-coverage sequence data, RepeatExplorer (Novák *et al.*, 2010, 2013), a high number of clusters are generated, each representing a putatively homologous repeat family/class. Within each cluster or element, the sequence divergence is low, and although this can be used for fine-scale classification of element types, particularly retroelements (e.g. Piednoël *et al.*, 2013; Mascagni *et al.*, 2015; Harkess *et al.*, 2016; Tetreault & Ungerer, 2016), the sequence divergence is not sufficient to infer taxon relationships. However, the abundance of homologous repeats does differ and the abundance of elements is often indicative of evolutionary relatedness, i.e. phylogeny (e.g. in bananas, Novák *et al.*, 2014; angiosperms and *Drosophila*, Dodsworth *et al.*, 2015; and in poplars, Usai *et al.*, 2017). But in some cases this is not entirely clear-cut, owing to the activity of some elements, particularly those in high abundance, that are more reflective of recent activity or, perhaps, differential processes of elimination from the genome (Pons *et al.*, 2004; Ribeiro *et al.*, 2017; Ustyantsev *et al.*, 2017). This needs to be explored and tested in cases where the topology is 'known', such that particular element histories can be teased apart and their impact on overall phylogenetic signal investigated.

Here we decided to test the abundance of repeats as adequate phylogenetic characters, particularly exploring the homoplasy of repeats, using the hominids as a case study. This group was selected owing to the widely accepted phylogenetic hypothesis based on much previous research and genome-scale data. Specifically, we set out to answer the following questions in this study:

1. Is the phylogenetic signal of genomic repeat abundance reliable in the case of the hominids?
2. Do certain clusters/repeats with homoplasious abundances adversely affect the phylogenetic signal?
3. Is one individual per taxon enough to build a reliable phylogenetic tree from genomic repeat abundances?

MATERIAL AND METHODS

Data acquisition

We downloaded high-throughput sequence data from 15 National Center for Biotechnology Information (NCBI) short read archive (SRA) accessions, including Illumina reads (Illumina Inc., San Diego, CA, USA) from three individuals belonging to five of the well-known species of the Hominidae family of primates (Table 1): *Homo sapiens*, *Pan troglodytes*, *Pan paniscus*, *Gorilla gorilla* and *Pongo pygmaeus*. We also downloaded

Table 1. Taxon sampling of hominids from National Center for Biotechnology Information SRA accessions

Species	Phylogeny ID	SRA run ID	BioSample ID	Geographical origin
Homo sapiens	HSAP1	ERR068394	SAMN00263022	Iberian populations in Spain
	HSAP2	ERR050125	SAMN00014366	Iberian populations in Spain
	HSAP3	ERR050124	SAMN00014365	Iberian populations in Spain
Pan troglodytes	PTRO1	SRR748072	SAMN01920536	Gabon
	PTRO2	SRR748062	SAMN01920534	Gabon
	PTRO3	SRR748058	SAMN01920533	Gabon
Pan paniscus	PPAN1	SRR740802	SAMN01920509	Democratic Republic of Congo
	PPAN2	SRR740794	SAMN01920508	Democratic Republic of Congo
	PPAN3	SRR740768	SAMN01920506	Democratic Republic of Congo
Gorilla gorilla	GGOR1	SRR748092	SAMN01920490	Western lowland
	GGOR2	SRR748096	SAMN01920491	Western lowland
	GGOR3	SRR748097	SAMN01920492	Western lowland
Pongo pygmaeus	PPYG1	SRR748020	SAMN01920551	Bornean
	PPYG2	SRR748000	SAMN01920547	Bornean
	PPYG3	SRR748004	SAMN01920548	Bornean
Macaca mulatta	MMUL1	SRR1944168	SAMN03264679	Indian breed

SRA, short read archive.

Illumina read data from a *Macaca mulatta* individual to be used as an outgroup for phylogenetic analyses.

In order to avoid data biases based on different sequencing protocols, all reads used in this study were chosen because they had been obtained on the same sequencing platform (Illumina HiSeq 2000), thus yielding reads of 100 bp in length, except for the *M. mulatta* library, in which the Illumina read length was 101 bp. Chimpanzee, bonobo, gorilla and orangutan data were acquired from wild-born individuals sequenced within the same SRA BioProject (PRJNA189439; IBE CSIC-Universitat Pompeu Fabra; Prado-Martínez *et al.*, 2013), whereas human short reads belong to the 1000 Genomes Project Phase 3 (PRJNA262923).

Mitochondrial genome assembly, phylogenetic analysis and filtering

A total of 5 000 000 100/101 bp raw Illumina read pairs were randomly selected using the SeqTK software (https://github.com/lh3/seqtk) from each library downloaded from the SRA and were used for mitochondrial genome assembly with MITObim v.1.8 (Hahn *et al.*, 2013). The mitochondrial genomes used as reference for assembly are indicated in Table 2 and were downloaded from NCBI GenBank reference sequences. Genome annotation was performed in GENEIOUS v.4.8.5 (Biomatters Ltd, Auckland, New Zealand) by aligning with the reference mitochondrial genome of each species. To verify its phylogenetic identity, a phylogenetic tree was built based on maximum parsimony (MP) analysis of a global alignment of the whole newly assembled mitochondrial genome of each individual included in this study. The Tree Analysis Using New Technology (TNT) software for Linux 64 (no taxon limit), updated version of 11 December 2013 (Goloboff *et al.*, 2008), was used for phylogenetic reconstruction, using implicit enumeration. Before subsequent analyses of repetitive DNA abundance, all Illumina libraries were filtered out for mitochondrial DNA with the software DeconSeq v.0.4.3 (Schmieder & Edwards, 2011), using as reference the mitochondrial genome for each species shown in Table 2.

Table 2. Mitochondrial genome reference sequences

Taxa	NCBI reference sequence accession
Homo sapiens	NC_012920.1
Pan troglodytes	NC_001643.1
Pan paniscus	NC_001644.1
Gorilla gorilla	NC_001645.1
Pongo pygmaeus	NC_001646.1
Macaca mulatta	NC_005943.1

NCBI, National Center for Biotechnology Information.

Preparation of read data for repeat analyses

The SRA files were unpacked into FASTQ using the FASTQ-DUMP tool from the SRA Toolkit. Low-quality reads in FASTQ files were discarded using Trimmomatic (Bolger *et al.*, 2014) by removing adapters and selecting read pairs with all their nucleotides with Q (Phred quality score) > 30, using the options 'ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:30 MINLEN:[100/101]'.

All samples were assumed to have a genome size of ~3.5 Gbp, based on data available in the Animal Genome Size Database, which showed only slight variation in genome size between the species used in this study (3.47–3.85 Gbp; http://www.genomesize.com/ last accessed 12 November 2016), which is considered appropriate for this type of study (Dodsworth *et al.*, 2016a). Each accession was then sampled for 0.6% of the genome by randomly subsampling each Illumina dataset. This resulted in 200 000 reads per sample from all Hominidae accessions, randomly selected with SeqTK and then converted into FASTA format.

Selected reads from each sample were labelled with a unique five-character prefix, making a total combined dataset of 1 200 000 reads for datasets of one individual per species, 2 200 000 reads for datasets of two individuals per species and 3 200 000 reads for the global dataset including all individual samples. Specifically, we prepared three different datasets of one individual (library or sample) per species plus *M. mulatta* as an outgroup (six operational taxonomic units [OTUs] per dataset), three different datasets of two biological individuals per species plus *M. mulatta* as an outgroup (11 OTUs per dataset) and one dataset grouping together all libraries representing three biological individuals per species, making a total of 16 OTUs for phylogenetic analysis, as shown in Table 3.

RepeatExplorer clustering of samples

Clustering of Illumina reads was performed using the RepeatExplorer (RE) pipeline, implemented in a GALAXY server environment running locally in the University of Granada. RepeatExplorer clustering was used to identify genomic repeat clusters within each dataset, with default settings (minimum overlap = 55, cluster size threshold for detailed analysis = 0.01%, and the 'all reads are paired' option selected). For additional details about the clustering algorithm see Nóvak *et al.* (2010, 2013). For further identification of repeat clusters, we used a custom repeat database of all primate repetitive DNA annotations included in RepBase (Bao *et al.*, 2015; http://www.girinst.org/repbase/ last accessed 20 November 2016). Following Dodsworth *et al.* (2016a), we used the 1000 most abundant repeat clusters,

Table 3. Read sampling for repetitive DNA clustering and phylogenetic analyses

Dataset	Individuals per species	Phylogeny ID						Total reads
		HSAP	PTRO	PPAN	GGOR	PPYG	MMUL	
1	1	HSAP1	PTRO1	PPAN1	GGOR1	PPYG1	MMUL1	1 200 000
2	1	HSAP2	PTRO2	PPAN2	GGOR2	PPYG2	MMUL1	1 200 000
3	1	HSAP3	PTRO3	PPAN3	GGOR3	PPYG3	MMUL1	1 200 000
4	2	HSAP1	PTRO1	PPAN1	GGOR1	PPYG1	MMUL1	2 200 000
		HSAP2	PTRO2	PPAN2	GGOR2	PPYG2		
5	2	HSAP1	PTRO1	PPAN1	GGOR1	PPYG1	MMUL1	2 200 000
		HSAP3	PTRO3	PPAN3	GGOR3	PPYG3		
6	2	HSAP2	PTRO2	PPAN2	GGOR2	PPYG2	MMUL1	2 200 000
		HSAP3	PTRO3	PPAN3	GGOR3	PPYG3		
7	3	HSAP1	PTRO1	PPAN1	GGOR1	PPYG1	MMUL1	3 200 000
		HSAP2	PTRO2	PPAN2	GGOR2	PPYG2		
		HSAP3	PTRO3	PPAN3	GGOR3	PPYG3		

because they represented a sufficient proportion of the genome for phylogenetic analyses. Read counts per cluster and sample information obtained from RE can be found in figshare under the accession https://figshare.com/s/c2ccda047dd502890dcb

Phylogenetic analysis of clusters

The 1000 most abundant clusters of each dataset were used to create the data matrices for phylogenetic inference. TNT software was chosen for phylogenetic analyses under the maximum parsimony principle (Goloboff & Mattoni, 2006; Goloboff *et al.*, 2008). Cluster abundances were used as input (continuous characters). To make the cluster abundance values suitable as input for the TNT software, we divided all abundances by a factor calculated by dividing the abundance of the most abundant cluster by 65, so that all data would fall within the range 0–65 (with up to three decimals) as needed for analysis of continuous characters with TNT. Further transformations (e.g. cubed root) were checked but provided no improvement on the factorial transformation. Implicit enumeration (branch and bound) tree searches were used for datasets in this study owing to the small number of taxa in each dataset. Resampling was performed using 10 000 replicates, and symmetrical resampling was done by a modification of the standard bootstrap (Goloboff *et al.*, 2003). FigTree v.1.4.3 (http://tree.bio.ed.ac.uk/) was used for graphical view and representation of phylogenetic trees.

Filtering of disturbing clusters

After the first RE clustering, we found some clusters for satellite DNA and an endogenous retrovirus (ERV) that were abundant in chimpanzee, bonobo and gorilla but were absent in human and orangutan libraries. We identified these clusters by means of a Python script (https://github.com/mmarpe/phyl_rep_hominidae/blob/master/sel_clusters.py) that helped us to locate those clusters that had < 25 reads in *Homo* and *Pongo* but that were abundant in the rest of the hominid species. The identity of these clusters was confirmed by the RepeatExplorer annotation and further characterized by means of sequence homology search using BLASTn (Altschul *et al.*, 1990) and CENSOR (Kohany *et al.*, 2006) tools.

To test the effect of these clusters on the phylogenies built with the abundance of repeats, we performed two sets of phylogenetic analyses, one using unfiltered libraries and the other using libraries previously filtered out for these particular clusters. Filtering was performed by DeconSeq software against the CL3 satellite consensus sequence (X74280.1 and X74281.1 GenBank accessions; Royle *et al.*, 1994) and against the CERV1_INT, the internal sequence for the endogenous retrovirus (Skaletsky *et al.*, 2004) included in RepBase.

Combinations of one or two individuals per species

Using a custom script, written in Python (https://github.com/mmarpe/phyl_rep_hominidae/blob/master/sample_mix.py), we phylogenetically analysed all possible combinations of one or two individuals per taxon (243 phylogenetic trees each), with abundances obtained from a global RE run of all libraries involved in this study after the above filtering of clusters. The 1000 most abundant clusters of each combination were phylogenetically analysed by means of MP implemented using TNT software as described previously. From the

1000 top abundant cluster data obtained from the RE of all three individuals per species (all samples included in this paper) after filtering, this script constructs all possible cluster abundance datasets for all different abundance data combinations of two individuals per species or one single individual per species without sample repetitions; later, it generates the trees derived from each dataset using the same parameters described above for the TNT software, and finally, transforms the tree files from .nex format to .pdf format using FigTree to make their visualization more accessible.

The 243 trees produced from these combinations were grouped together in a file and, using Consense v.3.695 included in the PHYLIP package (Felsenstein, 1989, 2005), we obtained the consensus tree for two individual per species cluster abundances combinations and for one individual per species combinations. This consensus tree consists of groups that occur as often as possible in the data through implementation of the majority rule (extended) method (Margush & McMorris, 1981).

RESULTS

Mitogenome phylogenetic tree

In order to check the integrity and reliability of the libraries used, we assembled the full mitochondrial DNA sequence in each individual library, using MITObim v.1.8, and built a mitochondrial phylogeny by means of MP (Fig. 1). This showed the absence of mis-tagging or sample confusion, because it coincided with the universally accepted Hominidae phylogeny (Roos & Zinner, 2017).

Phylogenetic analyses using unfiltered datasets

The first set of RE clustering and phylogenetic analyses was performed using the datasets indicated in Table 3. None of the phylogenies obtained (Fig. 2) reflected the universally accepted phylogeny for the Hominidae family confirmed by the mitogenome phylogeny depicted below (Fig. 1). In all cases, *Homo sapiens* appeared in a basal position in the phylogeny and sometimes forming a clade with *Pongo pygmaeus* (Fig. 2C–F). Given that we noticed that the topology of most trees shown in Figure 2 supported the hypothesis of a *Pan/Gorilla* clade, we searched for clusters showing extremely high abundance similarity between humans and orangutans, which could be responsible for the observed phylogenetic distortion. For this purpose, we searched for clusters showing < 25 reads in *Homo* and *Pongo* but showing higher abundance in *Pan* and *Gorilla*, using a custom script.

Phylogenetic analyses using filtered datasets

We found two repetitive DNA elements, which were practically absent in *Homo sapiens* and *Pongo pygmaeus* (< 25 reads) but were abundant in *Pan* and *Gorilla* (Fig. 3A). These clusters were identified as a subterminal satellite repeat and an endogenous retrovirus (Figs 3B, C). The repeat unit of the CL3 satellite is 32 bp long; it was isolated from the chimpanzee genome, found to be even more abundant in gorillas, but not detected in humans or orangutans (Royle *et al.*, 1994). The endogenous retrovirus, CERV1/PTERV1, was found by means of the analysis of bacterial artificial chromosome chimpanzee genome sequences. It is integrated in the germline of African great ape and Old World monkey species but is absent from human and Asian ape genomes (Yohn *et al.*, 2005; Polavarapu *et al.*, 2006).

To evaluate the possible effect of these two repeats on the phylogenetic signal, we filtered these repeats out of all libraries and performed a new batch of

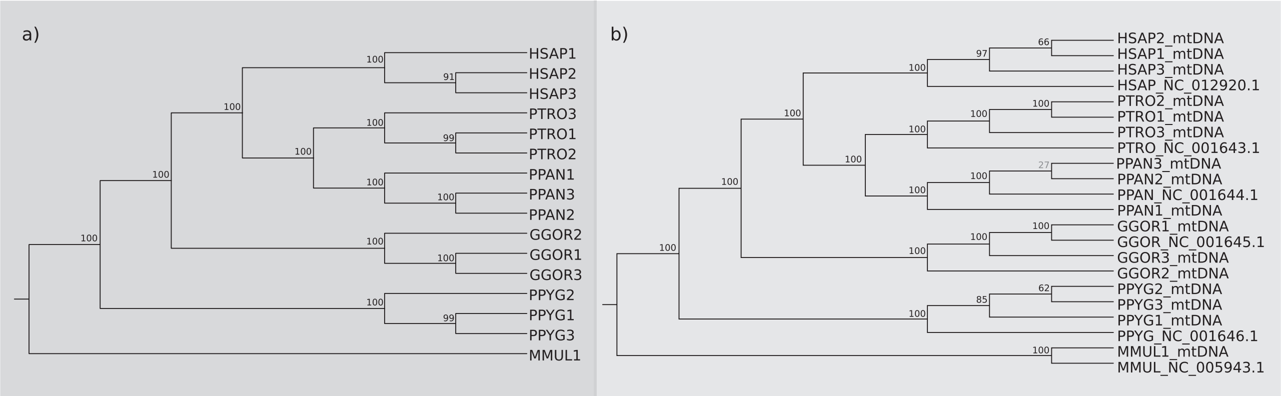

Figure 1. Mitochondrial phylogeny of all samples (libraries) used in the present study (A) and with the reference mitogenome from each species used in the sample mitochondrial DNA assembly (B). In each case, the trees represent the well-known {*Pongo* [*Gorilla* (*Pan* + *Homo*)]} topology. Bootstrap support of each node is specified on the tree (values < 50 in light grey indicate less robust nodes).

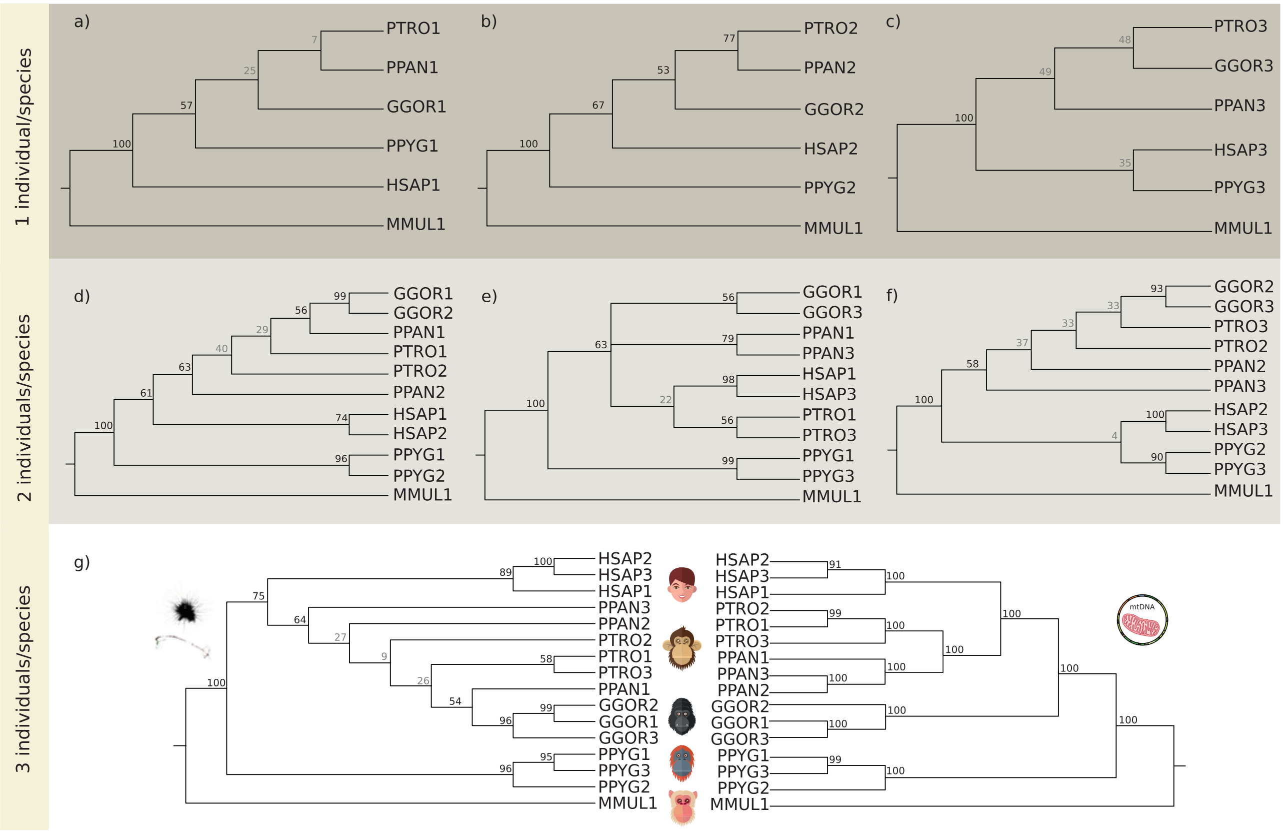

Figure 2. Genomic repeat phylogenies of one (A–C), two (D–F) and all samples (G) after RepeatExplorer clustering of unfiltered libraries. Bootstrap support of each node is specified on the tree (values < 50 in light grey indicate less robust nodes). Note that none of the trees matches the topology of the mitochondrial DNA tree. Even with three individuals per species (G), the tree reconstructed using repetitive element abundances (on the left) placed *Homo* as the ancestor of *Pan* and *Gorilla*, in strong disagreement with the mitochondrial DNA tree (and current accepted placement).

phylogenetic analyses on the same datasets described in Table 3, following the same protocol after filtering. As shown in Figure 3C, the endogenous retrovirus was partially clustered in CL140 (cluster graphs of full ERVs should have a circular shape). We found some other clusters containing part of this ERV, but they were less abundant and they were not discarded after the use of the script for filtering ERV reads. We decided to include these small clusters in subsequent phylogenetic analyses, because their presence did not influence the phylogenetic signal of the dataset as a whole. In addition, homoplasious clusters were filtered out of libraries using full reference sequences from RepBase, which means that the number of retained reads matching those repetitive elements is very low after filtering.

The phylogenies obtained (Fig. 4) failed to show the previous close relationship between *Homo* and *Pongo*, indicating that the discarded repeats were responsible for the distortion of the phylogenetic signal shown in the first set of analyses. In fact, the tree built with three individuals per species yielded a tree (Fig. 4G) with essentially the same topology as the mitogenome tree, albeit with low node support in places.

This result demonstrates that some repeats can generate 'random homoplasy' by differential amplification among different evolutionary lineages. In the present datasets, a satellite DNA and a retrovirus became highly abundant in the *Pan* and *Gorilla* lineages, whereas they did not prosper in the *Homo* and *Pongo* lineages, for which reason the two latter species showed a homoplasious rather than real phylogenetic relationship. This might present a serious problem for using the abundance of repeats for phylogenetic analysis in groups not as well known as the hominids.

One or two individuals per species can yield poor phylogenetic trees

As shown in Figure 4G, the phylogeny built with three individuals per species was very similar to that obtained with the mitogenomes, when the two

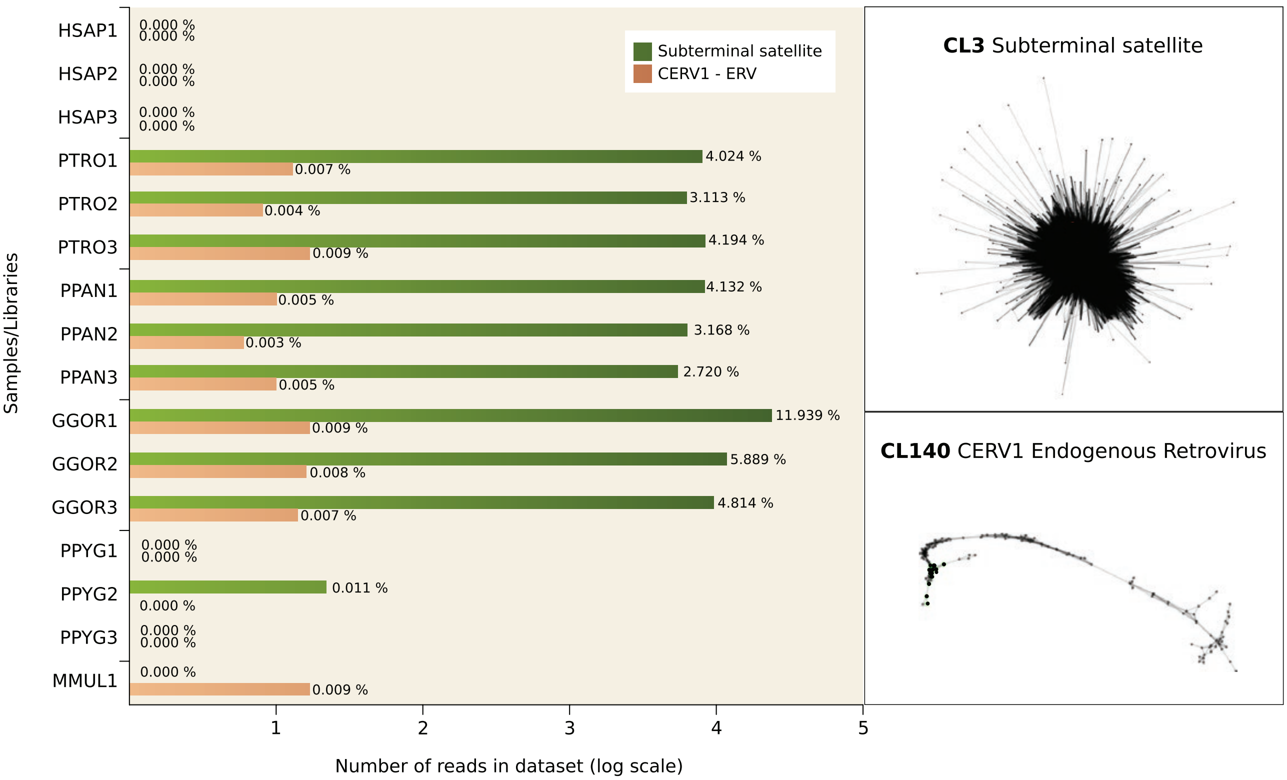

Figure 3. A, abundance of the CL3 subterminal satellite and the CERV1-ERV (CL140) retrovirus per individual. Number of reads as log-scaled bars and percentages shown next to bars indicate the proportion of each element per sample in the RepeatExplorer dataset. B, C, graph-clusters of the two homoplasious repeats, CL3 satellite and CL140 CERV1 retrovirus.

homoplasy-generating repeats were filtered out from the libraries. However, trees built with one or two individuals per species were still better than those performed by the unfiltered libraries, because *Pongo* was ancestral with respect to *Gorilla*, *Pan* and *Homo*, but they did not resolve properly the phylogenetic relationships between the three latter taxa (see Fig. 4A–F), because all these topologies show an unsolved *Homo*/*Pan*/*Gorilla* clade. According to the phylogenetic analysis of technical replicates (15 technical replicates, one for each biological sample used in this study, outgroup excluded), this issue of resolution might be attributable to inter-individual variation rather than sequencing bias (for technical replicates analysis, see Supporting Information, Tables S1 and S2).

To evaluate the effect of inter-individual (coincident with intraspecific in this case) variation in repeat abundance on phylogenetic reconstruction, we made all possible combinations of one or two individuals per species, chosen from the matrix of abundances obtained after RE clustering of the dataset including all three filtered libraries per species. We thus performed the phylogenetic inference for each combination, producing 243 trees for the combinations of one individual and 243 trees for the combinations of two individuals per species. The results showed that the consensus tree for the combinations of one individual per species did not reflect the phylogeny of the mitogenome (Fig. 5A), although 36 trees out of the set of 243 did. However, the consensus tree obtained from the combinations of two individuals clearly represented the phylogenetic relationships universally accepted for the Hominidae (Fig. 5B), although only 16 trees out of the 243 showed the resolved and accepted topology.

We conclude that the phylogenetic inference obtained from genomic repeat abundance is highly dependent on inter-individual variation, and the use of only one or two individuals per taxon may potentially lead, with high probability [(24– 36)/243 = 0.85 with $N = 1$ and (243 – 16)/243 = 0.93 with $N = 2$], to wrong phylogenetic inferences, at least in the case of the Hominidae family.

DISCUSSION

Phylogeny of Hominidae using repeat abundance

The phylogenetic relationships of the Hominidae family have been the object of study and great interest

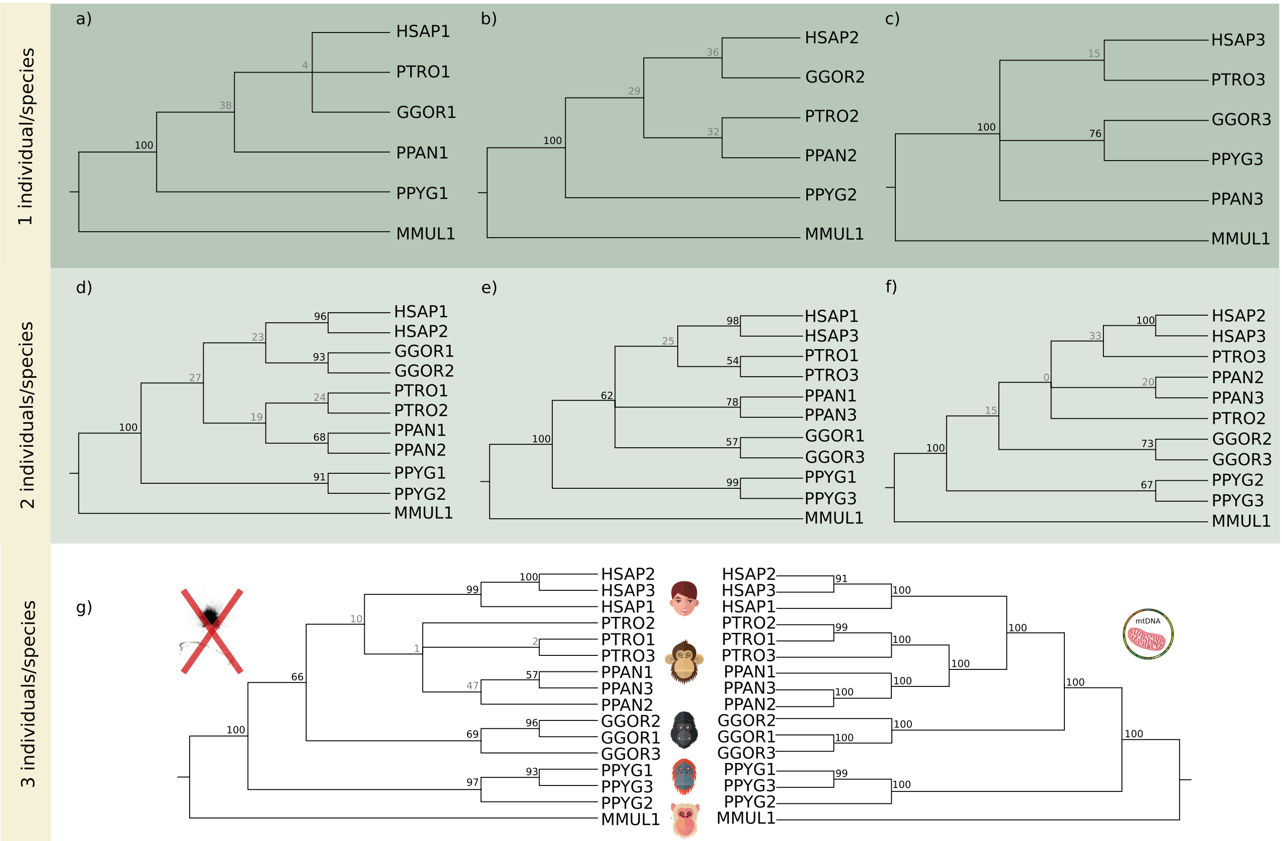

Figure 4. Genomic repeat phylogenies of one (A–C), two (D–F) and all samples (G) after RepeatExplorer clustering of libraries previously filtered out for a satellite DNA (CL3) and an endogenous retrovirus (CL140-CERV1) that have homoplasious abundance distributions. Bootstrap support for each node is specified on the tree (values < 50 in light grey indicate less robust nodes).

for the scientific community for a long time, and they have not been exempt from controversy (Holmquist *et al.*, 1988; Dean & Delson, 1992; Grehan & Schwartz, 2009, 2011). Currently, the {*Pongo* [*Gorilla* (*Pan* + *Homo*)]} evolutionary reconstruction is universally accepted and well established (Purvis, 1995; Arnason *et al.*, 2000; Arnold *et al.*, 2010; Perelman *et al.*, 2011; Popadin *et al.*, 2017); therefore, we believe that it is an appropriate model to test the method of phylogenetic estimation from the abundance of genomic repeats.

We compared our results with the reference tree built by mitogenomes (Fig. 1), which agrees with the previously accepted topology for this group (chromosomal evidence, Seuánez, 1982; morphological data, Ciochon *et al.*, 1983; identity of the α and β haemoglobin sequences, Goodman *et al.*, 1983; using DNA–DNA hybridization values, Sibley & Ahlquist, 1984; mitochondrial DNA analyses, Hayasaka *et al.*, 1988; β-globin gene clusters study, Koop *et al.*, 1989). Our results show that there is phylogenetic signal present in repeat abundances, for the top 1000 most abundant repetitive elements in the hominid nuclear genomes (Figs 2–4). Generally, we recovered phylogenetic hypotheses close to the accepted tree topology indicated above. However, this was only after adding more than one individual per species and after filtering out two particular repeats that had high abundance but not in closely related taxa, therefore distorting the phylogenetic inference (Fig. 4). The most acceptable phylogeny was inferred when making a consensus of all possible combinations of two-taxon datasets (Fig. 5B) after RE clustering of three individuals per species and filtering out the two clusters causing homoplasy. Even then, some nodes are not well supported according to bootstrapping, which underlies a lack of phylogenetic signal of repeat abundances for some parts of the tree.

Inter-individual variation affects phylogenetic inference

The abundance of repetitive elements appears to show high variation between individuals, so that ideally two or more individuals per species should be used

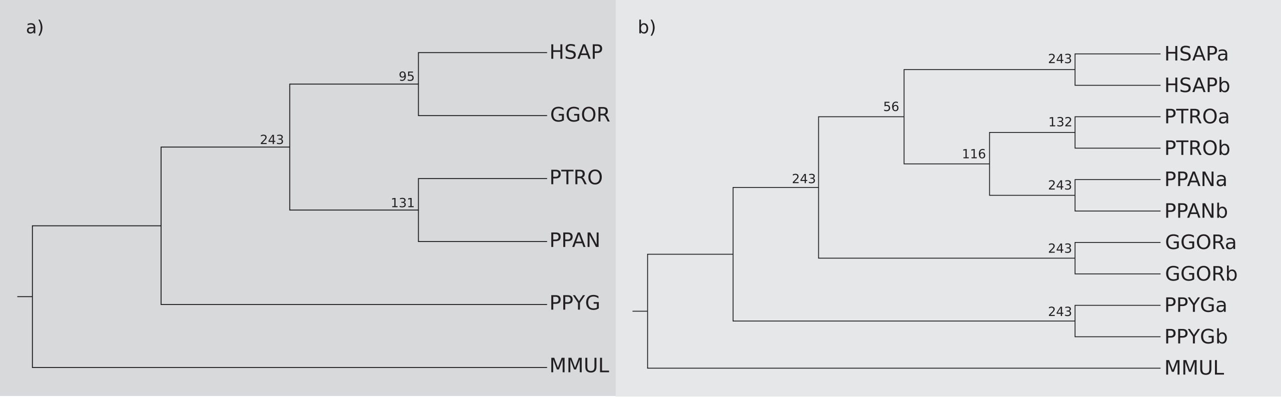

Figure 5. Consensus phylogenetic trees obtained from all possible combinations of one (A) and two (B) individuals per species (after filtering of the two homoplasious repeats). Numbers beside nodes indicate the number of trees, out of 243, that support the split. Note that the consensus tree built with two samples per taxon (B) shows a similar topology to the mitochondrial DNA tree shown in Figure 1, albeit with low support for two nodes.

for phylogenetic analysis based on repeat abundance (Fig. 4). The most unsatisfactory phylogenetic trees we generated were from the datasets that included only one individual per taxon (Figs 2–4), in which *Homo* is either misplaced or the tree is generally unresolved with respect to other hominids. This did not vastly improve even after filtering of clusters with homoplasious distributions (Fig. 4), suggesting that although this might eliminate the issue of (some) homoplasy, it does not negate the caveat of inter-individual variation in repeat abundance.

Homoplasious repeats obscure true phylogenetic signal

A phylogenetic hypothesis reflecting the currently accepted Hominidae phylogeny was obtained only using two or three individuals per taxon when their libraries were filtered out for the 'disturbing' clusters of repetitive DNA (Fig. 4G), a satellite DNA and an endogenous retrovirus, which showed differences in abundance between closely related species (e.g. *Homo* and *Pan*). These repetitive elements thus distorted the phylogenetic signal, yielding a falsely close relationship between *Homo* and *Pongo*. Removing these sequences from the libraries substantially improves the phylogenies obtained (Fig. 4). We believe this is a case of 'random homoplasy' generated by the chance amplification of the satellite DNA (and spread of the retrovirus) in *Pan* and *Gorilla* but not in *Homo*, which makes the latter more similar to *Pongo* in this respect. As Figure 3 shows, the homoplasious satellite DNA was the third repetitive element in order of decreasing abundance in *Pan* (2.7–4.2%) and *Gorilla* (4.8–11.9%), such that its influence on phylogenetic signal appears to be logical. However, the endogenous retrovirus was only the 140th most abundant cluster (0.003–0.009% in *Pan* and 0.007–0.009% in *Gorilla*), but the trees built that included this repeat failed to fit the accepted phylogeny even after filtering out the abundant satellite (data not shown). This poses a serious problem for phylogenetic reconstruction through this approach, because the phylogenetic signal can be distorted not only by the most abundant repeats but also by others that show much lower abundance in the genomes.

Methods of phylogenetic inference that handle continuous data adequately as phylogenetic characters are currently limited but could be improved upon (e.g. model-based solutions) and, in this case, would aid phylogenetic inference from repeat abundance data. The MP algorithm implemented in TNT is similar to ordinary MP, and therefore homoplasious repeats with large differences in abundance (such as those two identified for hominids) have an adversely large effect on tree length and therefore the most parsimonious phylogenetic tree that is reconstructed. This effect can sometimes be minimized by the use of different transformations on the data matrix, in order to make the abundances between zero and 65. For example, square root or other root transformations retain the abundance differences between taxa but minimize the overall abundance (length) differences for any particular cluster (character), as used, for example, by Dodsworth *et al.* (2016b) and tested in the present study (data not shown). However, these approaches do not alleviate the problem in the worst cases, such as the one shown in the present study for the Hominidae family, and it is advised that these clusters (repeats) are identified and removed from the dataset before phylogenetic inference. In cases without previous knowledge of phylogenetic relationships for the taxa involved, discarding every cluster showing large differential abundances that might be homoplasious, i.e. being absent or present in only two taxa, could be an option. We tried to do this for the present dataset, but it eliminated some clusters that were important for grouping the two *Pan* species, because they included repeats specific to that clade of two species (data not shown). More adequate model-based methods for inferring the phylogeny would also help to overcome the homoplasious nature of some repeat types, but these methods require further development. Therefore, the homoplasy problem might not be easy to solve, because repetitive DNA rarely shows a static path along the tree of life (Kuhn *et al.*, 2008; Feliciello *et al.*, 2014; Rojo *et al.*, 2015; Barghini *et al.*, 2015; Ferreira de Carvalho *et al.*, 2016).

Conclusions

Here we tested the abundance of repetitive elements as phylogenetic characters to infer the phylogenetic relationships of hominid primates, the family Hominidae. In general, we were able to recover a phylogenetic hypothesis close to the accepted topology, i.e. that which was recovered from much previous genomic sequence data. We discovered two important caveats when exploring this type of data, which should be borne in mind for future analyses of repeat abundances as phylogenetic characters: (1) individual variation in repeat abundance suggests that multiple samples per taxon should be included if at all possible; and (2) particular repeats can have highly homoplasious distributions such that they distort the phylogenetic signal in the overall dataset. We suggest that without a priori knowledge of the expected phylogenetic topology, researchers should be cautious and check for unusual signals yielded by repetitive elements irregularly distributed in the genomes of the tested organisms.

ACKNOWLEDGEMENTS

M.M.P was supported by a FPU fellowship (FPU13/01553) from the Spanish Ministerio de Educación, Cultura y Deporte. F.J.R.R. and J.P.M.C. were supported by the Spanish Secretaría de Estado de Investigación, Desarrollo e Innovación (CGL2015-70750-P), including FEDER funds, and F.J.R.R. was also supported by a Junta de Andalucía fellowship. S.D. was supported by a NERC studentship.

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* **215:** 403–410.

Arnason U, Gullberg A, Burguete AS, Janke A. 2000. Molecular estimates of primate divergences and new hypotheses for primate dispersal and the origin of modern humans. *Hereditas* **133:** 217–228.

Arnold C, Matthews LJ, Nunn CL. 2010. The 10kTrees website: a new online resource for primate phylogeny. *Evolutionary Anthropology* **19:** 114–118.

Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6:** 11.

Barghini E, Natali L, Giordani T, Cossu RM, Scalabrin S, Cattonaro F, Šimková H, Vrána J, Doležel J, Morgante M, Cavallini A. 2015. LTR retrotransposon dynamics in the evolution of the olive (*Olea europaea*) genome. *DNA Research* **22:** 91–100.

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

Cai Z, Liu H, He Q, Pu M, Chen J, Lai J, Li X, Jin W. 2014. Differential genome evolution and speciation of *Coix lacryma-jobi* L. and *Coix aquatica* Roxb. hybrid guangxi revealed by repetitive sequence analysis and fine karyotyping. *BMC Genomics* **15:** 1025.

Ciochon RL. 1983. Hominoid cladistics and the ancestry of modern apes and humans. In: Corruccini RS, Ciochon RL, ed. *New interpretations of ape and human ancestry*. New York: Academic Press, 783–843.

Dean D, Delson E. 1992. Palaeoanthropology. Second gorilla or third chimp? *Nature* **359:** 676–677.

Dodsworth S, Chase MW, Kelly LJ, Leitch IJ, Macas J, Novák P, Piednoël M, Weiss-Schneeweiss H, Leitch AR. 2015. Genomic repeat abundances contain phylogenetic signal. *Systematic Biology* **64:** 112–126.

Dodsworth S, Chase MW, Särkinen T, Knapp S, Leitch AR. 2016a. Using genomic repeats for phylogenomics: a case study in wild tomatoes (*Solanum* section *Lycopersicon*: Solanaceae). *Biological Journal Linnean Society* **117:** 96–105.

Dodsworth S, Jang TS, Struebig M, Chase MW, Weiss-Schneeweiss H, Leitch AR. 2016b. Genome-wide repeat dynamics reflect phylogenetic distance in closely related allotetraploid *Nicotiana* (Solanaceae). *Plant Systematics and Evolution* **33:** 1013–1020.

Feliciello I, Akrap I, Brajković J, Zlatar I, Ugarković Đ. 2014. Satellite DNA as a driver of population divergence in the red flour beetle *Tribolium castaneum*. *Genome Biology and Evolution* **7:** 228–239.

Felsenstein J. 1989. PHYLIP – Phylogeny Inference Package (Version 3.2). *Cladistics* **5:** 164–166.

Felsenstein J. 2005. *PHYLIP (Phylogeny Inference Package) version 3.6*. Distributed by the author. Seattle: Department of Genome Sciences, University of Washington.

Ferreira de Carvalho J, de Jager V, van Gurp TP, Wagemaker NC, Verhoeven KJ. 2016. Recent and dynamic transposable elements contribute to genomic divergence under asexuality. *BMC Genomics* **17:** 884.

Goloboff PA, Farris JS, Kallersjo M, Oxelman B, Ramirez MJ, Szumik CA. 2003. Improvements to resampling measures of group support. *Cladistics* **19:** 324–332.

Goloboff PA, Farris JS, Nixon KC. 2008. TNT, a free program for phylogenetic analysis. *Cladistics* **24:** 774–786.

Goloboff PA, Mattoni CI. 2006. Continuous characters analyzed as such. *Cladistics* **22:** 589–601.

Goodman M, Braunitzer G, Stangl A, Schrank B. 1983. Evidence on human origins from haemoglobins of African apes. *Nature* **303:** 546–548.

Grehan JR, Schwartz JH. 2009. Evolution of the second orangutan: phylogeny and biogeography of hominid origins. *Journal of Biogeography* **36:** 1823–1844.

Grehan JR, Schwartz JH. 2011. Evolution of human–ape relationships remains open for investigation. *Journal of Biogeography* **38:** 2397–2404.

Guschanski K, Krause J, Sawyer S, Valente LM, Bailey S, Finstermeier K, Sabin R, Gilissen E, Sonet G, Nagy ZT, Lenglet G, Mayer F, Savolainen V. 2013. Next-generation museomics disentangles one of the largest primate radiations. *Systematic Biology* **62:** 539–554.

Hahn C, Bachmann L, Chevreux B. 2013. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. *Nucleic Acids Research* **41:** e129.

Harkess A, Mercati F, Abbate L, McKain M, Pires JC, Sala T, Sunseri F, Falavigna A, Leebens-Mack J. 2016. Retrotransposon proliferation coincident with the evolution of dioecy in *Asparagus*. *G3* **6:** 2679–2685.

Hayasaka K, Gojobori T, Horai S. 1988. Molecular phylogeny and evolution of primate mitochondrial DNA. *Molecular Biology and Evolution* **5:** 626–644.

Holmquist R, Miyamoto MM, Goodman M. 1988. Higher-primate phylogeny—why can't we decide? *Molecular Biology and Evolution* **5:** 201–216.

Kohany O, Gentles AJ, Hankus L, Jurka J. 2006. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics* **7:** 474.

Koop BF, Tagle DA, Goodman M, Slightom JL. 1989. A molecular view of primate phylogeny and important systematic and evolutionary questions. *Molecular Biology and Evolution* **6:** 580–612.

Kuhn GC, Sene FM, Moreira-Filho O, Schwarzacher T, Heslop-Harrison JS. 2008. Sequence analysis, chromosomal distribution and long-range organization show that rapid turnover of new and old pBuM satellite DNA repeats leads to different patterns of variation in seven species of the *Drosophila buzzatii* cluster. *Chromosome Research* **16:** 307–324.

Margush T, McMorris FR. 1981. Consensus *n*-trees. *Bulletin of Mathematical Biology* **43:** 239–244.

Mascagni F, Barghini E, Giordani T, Rieseberg LH, Cavallini A, Natali L. 2015. Repetitive DNA and plant domestication: variation in copy number and proximity to genes of LTR-retrotransposons among wild and cultivated sunflower (*Helianthus annuus*) genotypes. *Genome Biology and Evolution* **7:** 3368–3382.

Novák P, Hřibová E, Neumann P, Koblížková A, Doležel J, Macas J. 2014. Genome-wide analysis of repeat diversity across the family Musaceae. *PLoS One* **9:** e98918.

Novák P, Neumann P, Macas J. 2010. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics* **11:** 378.

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:** 792–793.

Perelman P, Johnson WE, Roos C, Seuánez HN, Horvath JE, Moreira MA, Kessing B, Pontius J, Roelke M, Rumpler Y, Schneider MP, Silva A, O'Brien SJ, Pecon-Slattery J. 2011. A molecular phylogeny of living primates. *PLoS Genetics* **7:** e1001342.

Piednoël M, Carrete-Vega G, Renner SS. 2013. Characterization of the LTR retrotransposon repertoire of a plant clade of six diploid and one tetraploid species. *The Plant Journal* **75:** 699–709.

Polavarapu N, Bowen NJ, McDonald JF. 2006. Identification, characterization and comparative genomics of chimpanzee endogenous retroviruses. *Genome Biology* **7:** R51.

Pons J, Bruvo B, Petitpierre E, Plohl M, Ugarkovic D, Juan C. 2004. Complex structural features of satellite DNA sequences in the genus *Pimelia* (Coleoptera: Tenebrionidae): random differential amplification from a common 'satellite DNA library'. *Heredity* **92:** 418–427.

Popadin K, Gunbin K, Peshkin L, Annis S, Kraytsberg G, Markuzon N, Ackermann RR, Khrapko K. 2017. Mitochondrial pseudogenes suggest repeated interspecies hybridization in hominid evolution. *BioRxiv* 134502.

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

Purvis A. 1995. A composite estimate of primate phylogeny. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **348:** 405–421.

Ren Z, Harris AJ, Dikow RB, Ma E, Zhong Y, Wen J. 2017. Another look at the phylogenetic relationships and intercontinental biogeography of eastern Asian North American *Rhus* gall aphids (Hemiptera: Aphididae: Eriosomatinae): evidence from mitogenome sequences via genome skimming. *Molecular Phylogenetics and Evolution* **117:** 102–110.

Ribeiro T, Dos Santos KG, Richard MM, Sévignac M, Thareau V, Geffroy V, Pedrosa-Harand A. 2017. Evolutionary dynamics of satellite DNA repeats from *Phaseolus* beans. *Protoplasma* **254:** 791–801.

Ricci M, Luchetti A, Bonandin L, Mantovani B. 2013. Random DNA libraries from three species of the stick insect genus *Bacillus* (Insecta: Phasmida): repetitive DNA characterization and first observation of polyneopteran MITEs. *Genome* **56:** 729–735.

Richter S, Schwarz F, Hering L, Böggemann M, Bleidorn C. 2015. The utility of genome skimming for phylogenomic analyses as demonstrated for glycerid relationships (Annelida, Glyceridae). *Genome Biology and Evolution* **7:** 3443–3462.

Rojo V, Martínez-Lage A, Giovannotti M, González-Tizón AM, Nisi Cerioni P, Caputo Barucchi V, Galán P, Olmo E, Naveira H. 2015. Evolutionary dynamics of two satellite DNA families in rock lizards of the genus *Iberolacerta* (Squamata, Lacertidae): different histories but common traits. *Chromosome Research* **23:** 441–461.

Roos C, Zinner D. 2017. Primate phylogeny. In: Fuentes A, ed. *The international encyclopedia of primatology*. Hoboken: Wiley-Blackwell.

Royle NJ, Baird DM, Jeffreys AJ. 1994. A subterminal satellite located adjacent to telomeres in chimpanzees is absent from the human genome. *Nature Genetics* **6:** 52–56.

Schmieder R, Edwards R. 2011. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS One* **6:** e17288.

Seuánez HN. 1982. Chromosome banding and primate phylogeny. Inaugural Address. In: Chiarelli AB, Corruccini RS, eds. *Advanced views in primate biology. Proceedings in life sciences*. Berlin, Heidelberg: Springer.

Sibley CG, Ahlquist JE. 1984. The phylogeny of the hominoid primates, as indicated by DNA-DNA hybridization. *Journal of Molecular Evolution* **20:** 2–15.

Skaletsky H, Hughes JF, Page DC. 2004. Consensus sequence of an endogenous retrovirus CERV1. *Repbase Reports* **4:** 189.

Straub SC, Parks M, Weitemier K, Fishbein M, Cronn RC, Liston A. 2012. Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. *American Journal of Botany* **99:** 349–364.

Sveinsson S, Gill N, Kane NC, Cronk Q. 2013. Transposon fingerprinting using low coverage whole genome shotgun sequencing in cacao (*Theobroma cacao* L.) and related species. *BMC Genomics* **14:** 502.

Tetreault HM, Ungerer MC. 2016. Long terminal repeat retrotransposon content in eight diploid sunflower species inferred from next-generation sequence data. *G3* **6:** 2299–2308.

Timmermans MJTN, Lees DC, Thompson MJ, Sáfián S, Brattström O. 2016. Mitogenomics of 'Old World Acraea' butterflies reveals a highly divergent 'Bematistes'. *Molecular Phylogenetics and Evolution* **97:** 233–241.

Usai G, Mascagni F, Natali L, Giordani T, Cavallini A. 2017. Comparative genome-wide analysis of repetitive DNA in the genus *Populus* L. *Tree Genetics & Genomes* **13:** 96.

Ustyantsev K, Blinov A, Smyshlyaev G. 2017. Convergence of retrotransposons in oomycetes and plants. *Mobile DNA* **8:** 4.

Yohn CT, Jiang Z, McGrath SD, Hayden KE, Khaitovich P, Johnson ME, Eichler MY, McPherson JD, Zhao S, Pääbo S, Eichler EE. 2005. Lineage-specific expansions of retroviral insertions within the genomes of African great apes but not humans and orangutans. *PLoS Biology* **3:** e110.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Table S1. Sampling of technical replicates from NCBI short read archives.
Table S2. Reads in clusters after RepeatExplorer clustering of two technical replicates per biological sample.

SHARED DATA

Icons appearing in the phylogenetic trees were freely downloaded from www.freepik.es and designed by their own Website developers. All data matrices, raw RE cluster abundances and processed input matrices, and phylogenetic trees built in the present study can be found in figshare under the accession URL https://figshare.com/s/c2ccda047dd502890dcb. All the scripts used are available from https://github.com/mmarpe/phyl_rep_hominidae.