

RESEARCH

Open Access



Sequencing the orthologs of human autosomal forensic short tandem repeats provides individual- and species-level identification in African great apes

Ettore Fedele^{1,2}, Jon H. Wetton^{1*} and Mark A. Jobling^{1*}

Abstract

Background Great apes are a global conservation concern, with anthropogenic pressures threatening their survival. Genetic analysis can be used to assess the effects of reduced population sizes and the effectiveness of conservation measures. In humans, autosomal short tandem repeats (aSTRs) are widely used in population genetics and for forensic individual identification and kinship testing. Traditionally, genotyping is length-based via capillary electrophoresis (CE), but there is an increasing move to direct analysis by massively parallel sequencing (MPS). An example is the ForenSeq DNA Signature Prep Kit, which amplifies multiple loci including 27 aSTRs, prior to sequencing via Illumina technology. Here we assess the applicability of this human-based kit in African great apes. We ask whether cross-species genotyping of the orthologs of these loci can provide both individual and (sub)species identification.

Results The ForenSeq kit was used to amplify and sequence aSTRs in 52 individuals (14 chimpanzees; 4 bonobos; 16 western lowland, 6 eastern lowland, and 12 mountain gorillas). The orthologs of 24/27 human aSTRs amplified across species, and a core set of thirteen loci could be genotyped in all individuals. Genotypes were individually and (sub)species identifying. Both allelic diversity and the power to discriminate (sub)species were greater when considering STR sequences rather than allele lengths. Comparing human and African great-ape STR sequences with an orangutan outgroup showed general conservation of repeat types and allele size ranges. Variation in repeat array structures and a weak relationship with the known phylogeny suggests stochastic origins of mutations giving rise to diverse imperfect repeat arrays. Interruptions within long repeat arrays in African great apes do not appear to reduce allelic diversity.

Conclusions Orthologs of most human aSTRs in the ForenSeq DNA Signature Prep Kit can be analysed in African great apes. Primer redesign would reduce observed variability in amplification across some loci. MPS of the orthologs of human loci provides better resolution for both individual and (sub)species identification in great apes than standard CE-based approaches, and has the further advantage that there is no need to limit the number and size ranges of analysed loci.

*Correspondence:

Jon H. Wetton
jw418@le.ac.uk
Mark A. Jobling
maj4@le.ac.uk

Full list of author information is available at the end of the article



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Keywords Gorilla, Chimpanzee, Bonobo, Short-tandem repeat (STR), Individual identification, Single nucleotide polymorphism (SNP), Massively parallel sequencing, Verogen ForenSeq™ DNA Signature Prep kit

Background

Habitat loss, disease, climate change and hunting are among the main drivers of localised and global extinctions [1]. As species become increasingly restricted to fragmented habitats it is necessary to assess their viability to support effective management decisions. Increasing global awareness has drawn attention towards the preservation of charismatic flagship species [2], among which the African great apes have been a focal interest: most of these species remain critically endangered throughout their home ranges [3] (Fig. 1). However, when threat status is measured merely on the basis of species decline and habitat degradation [4], it can neglect the biological and ecological impacts of shifts in population size and distribution [5]. As populations decline and inbreeding intensifies, heterozygosity falls [6]. In turn, reduced allelic diversity can affect the adaptive ability of the species and potentially lead to the emergence of genetic defects underpinned by recessive alleles [7].

As a response, DNA sequence-based approaches to assess population parameters now play an important role in implementing effective wildlife management and conservation policies [10]. Measuring polymorphism at sets of autosomal short tandem repeats (aSTRs) via capillary electrophoresis (CE) has been an important tool in population genetic analysis [11]. Because they assort independently at meiosis, sets of unlinked aSTRs also yield multilocus genotypes that are unique to individuals within a species: this forms the basis of human forensic identification technologies [12], and can be applied in forensic casework involving animals, for example in poaching or illegal trade cases [13]. Such genotypes also have the potential to distinguish between species and subspecies when allelic spectra are suitably differentiated and characteristic.

Because of the high levels of sequence similarity among great-ape genomes [14], PCR primers for aSTR markers developed in humans are expected to amplify their orthologs. Indeed, some STR multiplexes designed for human forensic analysis have been shown to have cross-species application for the analysis of orthologous loci in other great apes (e.g. [15, 16]). The underlying assumption is that amplicons generated at orthologous loci are generally commensurable across species [17]. However, this assumption is often incorrect; indeed, the presence of species-specific indels in flanking sequence together with different organisation and variability of STRs present difficulties with great-ape cross-species comparisons

[17]. In translating multiplexes designed in humans to other species, there is also a practical problem of interpretation, since allele size ranges for different loci (labelled with the same fluorescent dye) were designed to be non-overlapping in humans, but may well overlap in non-human primates.

These issues arise because of the nature of capillary electrophoresis, which assesses polymorphism by measuring the length of PCR fragments and converting this to an assumed number of repeat units within each allele. An alternative approach is multiplex massively parallel sequencing (MPS), in which the sequences of STRs, rather than their lengths, are analysed. This obviates the problem of size-range overlap, since it is the sequence itself that identifies the locus, and also permits larger numbers of STRs to be simultaneously analysed than is possible with length-based CE genotyping. Extensive concordance studies show agreement between the two analytical methods [18, 19]. MPS-based analysis is now becoming established in human forensic genetics. For example, the ForenSeq DNA Signature Prep Kit (Verogen) [20, 21] includes multiple autosomal, X- and Y-chromosomal STRs, as well as autosomal SNPs for individual identification.

This study aims to assess how the human-designed ForenSeq multiplex system performs in amplifying and sequencing autosomal STRs in a set of chimpanzees, bonobos and gorillas, and to ask if the orthologous loci are both individually identifying and can robustly distinguish groups at the species and subspecies levels. Sequencing across subspecies and species may also reveal aspects of the mutation processes of these widely used STRs across ~8 million years of primate evolution.

Results

We assembled a set of DNA samples from 52 non-human great-ape individuals (14 chimpanzees, 4 bonobos, 16 western lowland gorillas, 6 eastern lowland gorillas, and 12 mountain gorillas) for sequencing. Both prior information on some sampled individuals and later deductions from our own data using the software ML-Relate [22] (Table S1) indicate that the sample set contains close relatives within (sub)species, including some parent-offspring, full-sib and apparent half-sib pairs, though no mother-father-child trios. In describing the diversity of STR sequences and in considering identification at the individual and (sub)species levels, we retain all these individuals since they contribute new alleles to the dataset.

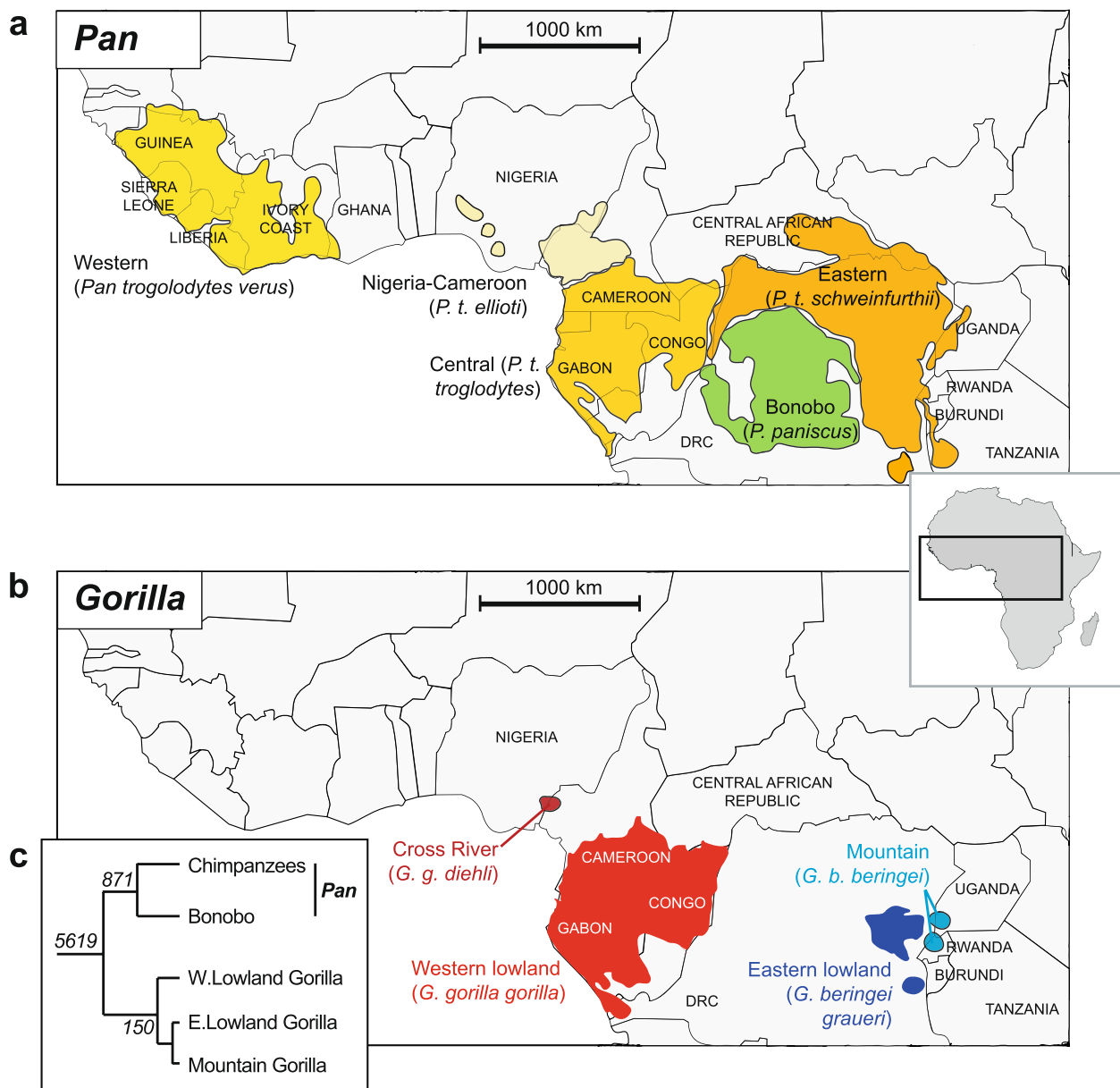


Fig. 1 *Pan* and *Gorilla* species and sub-species distributions, and phylogenetic relationships. Distributions of **a**) *Pan*, and **b**) *Gorilla*, adapted from [8]. **c** Phylogeny showing relationships between (sub)species, with classifications reflecting those used in this study. Italic numbers at nodes are split times in thousands of years, based on a mutation rate of 1×10^{-9} per bp per year [9]. Map adapted from *Africa just countries grayish.svg*, published on Wikimedia under a Creative Commons Attribution-Share Alike 4.0 International license. DRC: Democratic Republic of the Congo

When considering population structure, heterozygosity, inbreeding (F_{is}), and forensically-relevant diversity statistics we remove individuals such that there are no predicted relatives in the dataset, apart from in the highly inbred mountain gorillas, where we retain predicted half-sibs. Given that whole-genome sequencing [7] in this sub-species has shown chromosomes to be homozygous over > 38% of their lengths, ML-Relate's prediction of half-sib status here is likely to arise due to general close genetic similarity among individuals in the population.

Amplification of orthologs of human loci in the multiplex

The ForenSeq™ DNA Signature Prep Kit, designed to assess human DNA diversity, was used to amplify autosomal, X- and Y-STRs and autosomal SNP-containing loci (see Methods for details) in the set of 52 African great ape samples. Table S2 summarises amplification results across the entire set of 152 amplicons in the multiplex. Here, we focus on the 27 autosomal STRs (sequences given in Table S3), but also report the sequences of amplified X-STRs in Table S4. Twenty-one of the 52

samples are females, so Y-STR data are less extensive, and there is also a relatively high failure rate for amplifying orthologs of human loci, likely due to the elevated MSY mutation rate [23]. Two Y-STRs failed in all (sub)species (DYS481, DYS533), four failed in *Pan* (DYS19, DYS612, DYS385a,b, DYS448), and thirteen failed in *Gorilla* (DYS505, DYS570, DYS522, DYS437, DYS439, DYS389II, DYS438, DYS390, DYS643, Y-GATA-H4, DYS549, DYS392, DYF387S1). The amelogenin sex test loci [24] amplified in all individuals and gave results consistent with previously known sex (Table S1). We do not report sequence information for the human identity-informative SNP amplicons.

Of the 27 autosomal STRs targeted in the multiplex, two (D7S820 and D9S1122) failed to amplify in any individuals, and D5S818 amplifies only in gorillas. Although the actual primers used in the ForenSeq kit are not published, it is likely that they include or resemble well-established primers. Examination of 4-way species alignments around these loci (Figures S1-3) reveal that established primers for both D7S820 and D9S1122 lie across human-specific variants, which seems a possible explanation for failure to amplify. This is not so for D5S818, but there is a *Pan*-specific variant close to the 3' end of one established primer; if the ForenSeq primer terminates at this nucleotide, only humans and gorillas will amplify. Because D5S818 contains a low-diversity STR array with the structure $[AGAT]_{1-2}[AG]_{9-13}$, unlike the human ortholog which is a highly variable tetranucleotide repeat, $[AGAT]_{6-18}$, we do not consider it further here. Of the remaining 24 STRs, six (Fig. 2) could be analysed only in particular species, likely due to inter-specific sequence differences affecting primer sites. A set of 18 STRs amplifies in all species, but with some missing data in particular individuals. Missingness could be due to null alleles arising from sequence variants affecting primer sites, or to poor sequence quality (< 20 reads). Neglecting all STRs that show missing data leaves a 'core' set of thirteen STRs that were sequenced across all individuals; this set allows cross-species comparisons to be done.

Sequence diversity in autosomal STRs

By allowing variation within both the repeat array and flanking DNA to be observed (Fig. 3a), sequencing human autosomal STRs increases the observed allelic diversity [18, 25]. This is also the case in the great apes studied here (Fig. 3b-f; Table S5). Focusing on variation within the repeat array (since the lengths of flanking regions are not completely comparable between species) we see that STRs that show sequence variants are not well conserved across species. In humans, D12S391 shows by far the greatest increase in diversity due to

repeat array sequence variation [18, 25], but this feature is not observed in the great apes studied here. D2S1338 shows the greatest degree of repeat array sequence variants across (sub)species.

STR variant classes within and between (sub) species

To consider the sequence variation in the 18 cross-species amplifiable STRs in an evolutionary framework, we compared the *Pan* and *Gorilla* data to the predominant sequence structures of human orthologs (retrieved from STRBase.nist.gov and [18]). We included a single orangutan orthologous allele for the 17/18 loci where this could be identified, extracted from the orangutan (*Pongo abelii*) reference sequence (ponAbe3 assembly). Figure 4a summarises the STR structural categories observed; the range of allele structures for each locus is shown in a phylogenetic context in Fig. 4b-h and Figure S4.

Several loci (including D13S317, D19S433, TH01, TPOX and D16S539) show conserved features across the great apes, with perfect repeat arrays of the same repeat unit across all (sub)species examined, and similar repeat ranges (Fig. 4, Figure S4). We see no examples in which the major variable repeat unit differs in sequence between (sub)species, but among the remaining loci there is variation in structural types and little obvious relationship with the phylogeny, suggesting stochastic origins of mutations giving rise to diverse non-perfect repeat arrays. Repeat array length distributions are particularly well understood in humans because of very large sample sizes, whereas our great-ape sample sizes are small and may be highly unrepresentative. However, given this caveat, the number of repeats observed in all species fall within the range of human variation, with the exception of D13S317 (based on the lists given by STRBase.nist.gov and [18]).

Below, we summarise some features of structures for the 18 STRs that were amplifiable and sequenced across *Pan* and *Gorilla*. For several STRs (in particular D6S1043, D18S51, D19S433, PentaE and TH01), recorded human allele repeat number ranges are much wider than those seen in our sample of great apes. In fact, across all 18 STRs, there is only one case, D13S317 in western lowland gorilla, where the observed non-human primate allele size range exceeds that seen in humans. This may reflect the influence of ascertainment bias towards human STR variability for forensic use and the relatively large surveyed human sample sizes.

Some loci lack variant structures, and show straightforward patterns of variation in perfect arrays across the phylogeny. An example is TH01 (Fig. 4b), which is a simple, perfect array of AATG repeats in humans, and the same across *Pan* and *Gorilla*, albeit with narrower repeat number ranges (and invariant in bonobos). The

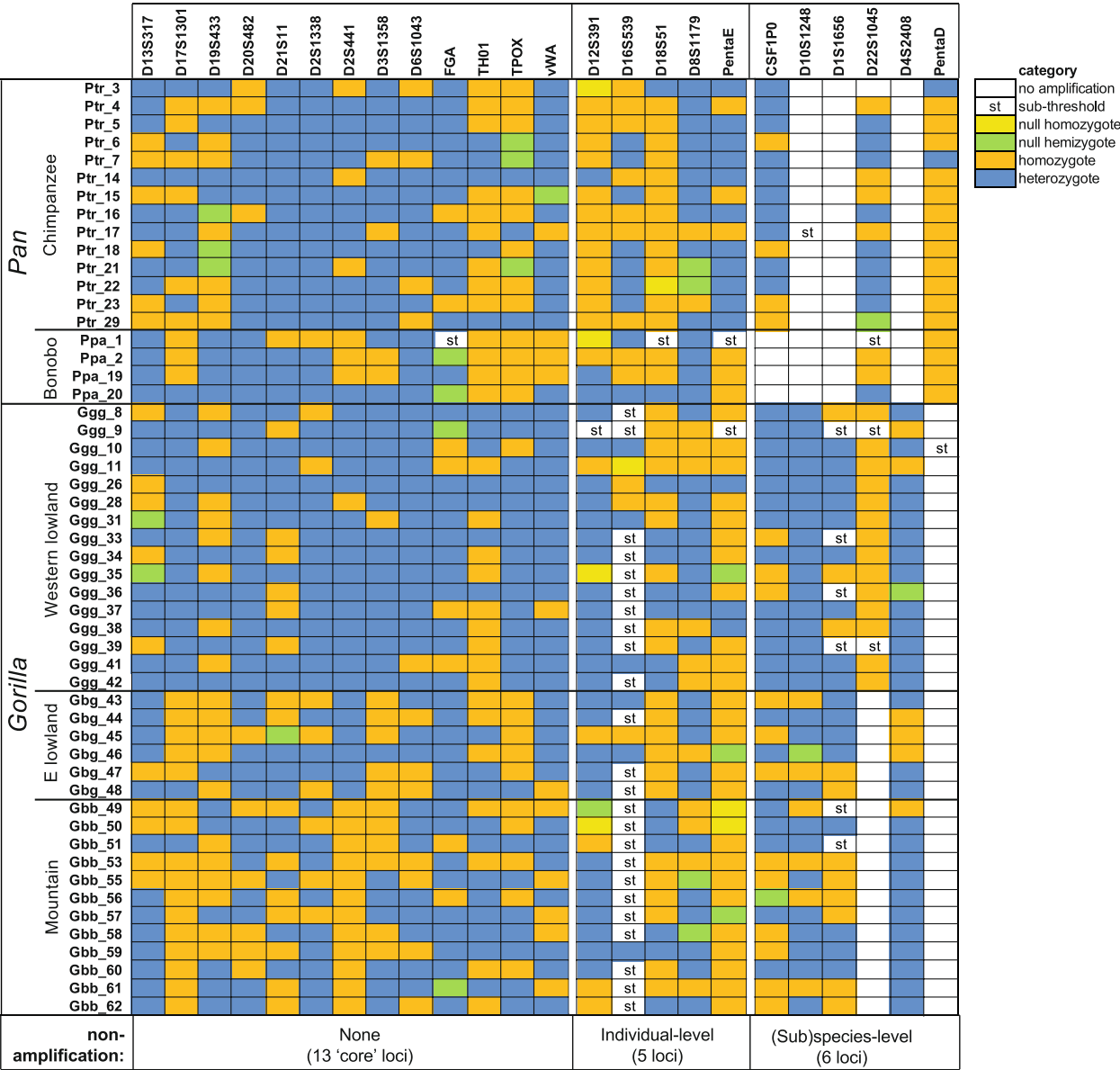


Fig. 2 Summary of amplification behaviours of autosomal STRs across individuals. For each STR and each great-ape individual, amplification behaviour is summarised, as indicated in the key to the right. Distinction between categories is based on sequence read-depth analysis. STRs are organised into three groups reflecting the amplifiability and degree of data completeness as indicated below the figure

(See figure on next page.)

Fig. 3 Counts of distinguishable alleles in each (sub)species by STR locus, and per-locus increment due to sequence variants. The observed numbers of length variants among individuals are shown as grey bars, and the number of additional alleles resulting from sequence variation within and flanking the repeat array are shown in white and black respectively. STRs are organised into three groups as in Fig. 2, and shown below the figure. **a** Human [25] **b** Chimpanzee; **c** Bonobo; **d** Western lowland gorilla; **e** Eastern lowland gorilla; **f** Mountain gorilla. Note that, although repeat array sequence variation is comparable across species, flanking sequence variation is not strictly comparable because the amount of sequence considered in different species varies somewhat

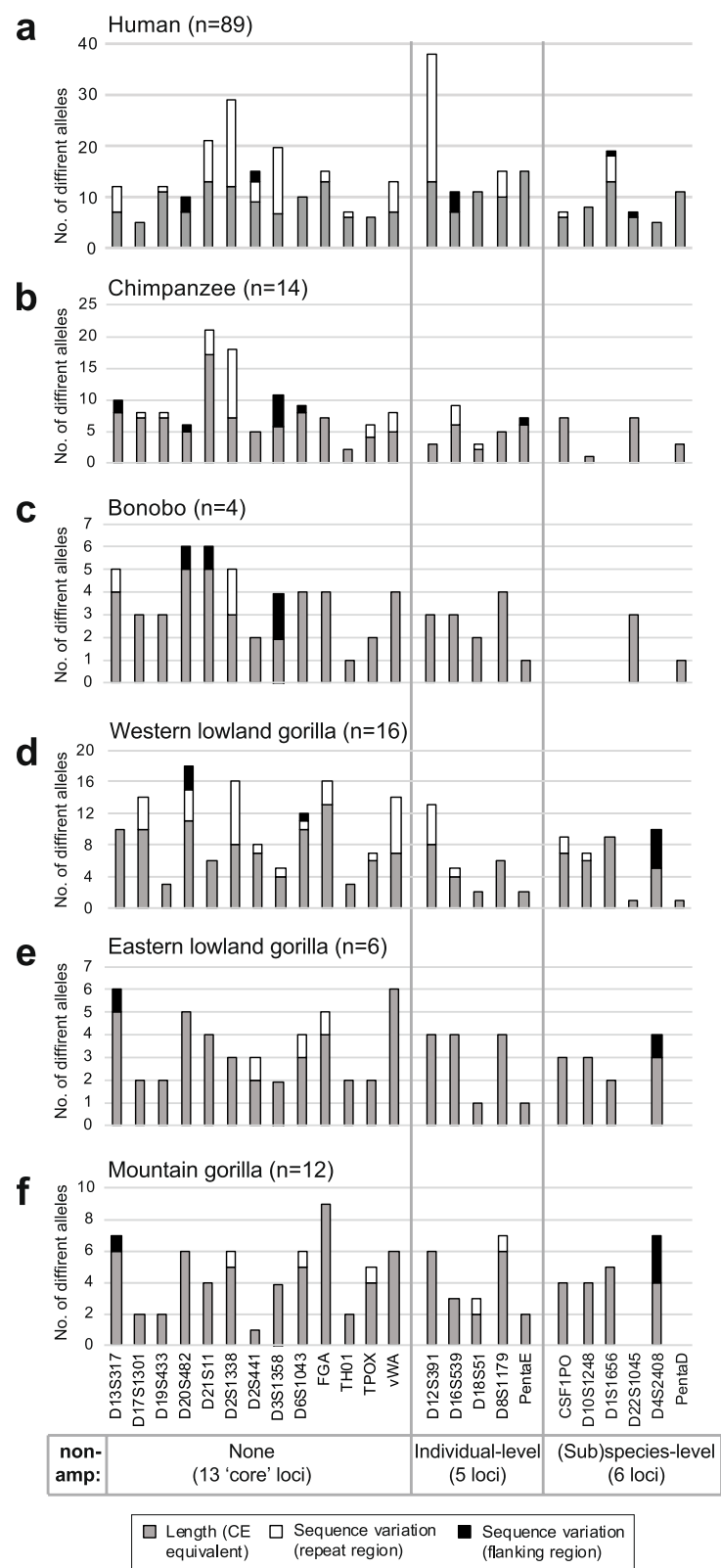


Fig. 3 (See legend on previous page.)

orangutan allele is very short and interrupted, and unlikely to be variable. Similarly simple features are seen at PentaE (Fig. 4c), D18S51, and D19S433 (Figure S1). Two of the human loci, D2S1338 and D12S391, are compound in humans with two variable blocks of different repeat types. These features are conserved: D2S1338 (Fig. 4d) shows similar structure and approximate array length ranges in humans, *Pan* and *Gorilla*, as a compound and polymorphic $[GGAA]_n[GGCA]_m$ STR. Surprisingly, the orangutan allele here comprises short arrays of different repeat units (AGGG and AGG). D12S391 (Fig. 4e) shows variable arrays of AGAT and AGAC repeats, and in orangutan is a simple perfect array of just one of these repeat types, AGAT.

There is little evidence of novel repeat arrays arising and expanding in particular species. One exception is D21S11 (Fig. 4f), which in all species shows one or more arrays of TCTA repeats, but in humans also includes a highly variable array of TCTG repeats that is not seen in any other species. The other example is at D12S391 (Fig. 4e), where (as well as the AGAT and AGAC arrays mentioned above) an array of AGGT repeats is specific to *Gorilla*, and polymorphic in western lowland and mountain gorillas.

STR mutation processes are generally thought of as rapid compared to single-nucleotide changes in non-repetitive DNA, and (unless there has been recent gene flow) we might therefore expect little identity-by-descent in the features of repeat arrays over the several million years of primate evolution. However, this is not so, and the distribution of structures identical by descent appears to be non-uniform across the great apes. There are no examples of distinctive *Pan*-specific derived features in any of the 18 STRs analysed. However, the picture is different in *Gorilla*. For D12S391 (Fig. 4e), vWA (Fig. 4g), D2S441, D16S539, FGA, and TPOX (Figure S4), all gorilla (sub)species studied carry more than one allele structure, and these are shared among western lowland, eastern lowland and mountain gorillas (which have an estimated divergence time of ~ 150 KYA; Fig. 1c). Only one locus, D8S1179 (Figure S1), shows distinctive structural features restricted to the two eastern subspecies.

Considerations of STR array evolution based on human diversity and pedigree data have shown that interrupting a long perfect repeat array with a variant repeat or indel leads to a marked reduction of mutation rate [26] and consequent lower allelic diversity. However, in both *Pan* and *Gorilla* there are several allele structures featuring polymorphic arrays separated by interruptions (variant repeats, or insertions). In most of these cases, other variant structures in the same (sub)species are short and perfect, and these are shared across species suggesting they may be ancestral. This raises the possibility that the long interrupted alleles might arise via a non-slippage-like process, but larger sample sizes would be needed to address this. Chimpanzee shows this phenomenon at D21S11 (Fig. 4f) and D17S1301, while it is seen in *Gorilla* at D20S482 (Fig. 4h), D13S317, D17S1301, and FGA (Figure S4).

Within-(sub)species variability of multilocus STR genotypes

Within (sub)-species, all individuals (including related individuals; Table S1) are distinguishable by their STR genotypes, and this is true for both CE-equivalent and sequence-based allele designations.

After removing related individuals (Table S1) we assessed observed vs expected heterozygosity for the tested loci (Table S6); following Bonferroni correction, only one locus in one species (D16S539 in chimpanzee), shows a significant deviation from expectation. We estimated F_{is} as a measure of inbreeding (Table S7). Following Bonferroni correction, significant positive F_{is} values are seen for three loci (D16S539, D19S433, TPOX) in chimpanzee, two (D13S317, D16S539) in western lowland gorilla, and one (D8S1179) in eastern lowland gorilla. As shown in Fig. 2, all except one of these (TPOX) show evidence of null alleles or low read-depth in the relevant (sub)species, suggesting that the F_{is} results reflect amplification issues rather than evidence of inbreeding. Forensic statistics derived from the data are given in Table S8, and Table 1 presents the combined random match probabilities (RMPs) in each (sub)species. The values obtained strongly reflect the sample sizes, which in turn influence the mean number of alleles observed

(See figure on next page.)

Fig. 4 Summary of STR structures across (sub)species, and examples of inter- and intra-specific structural variation. **a** For each (sub)species and each locus, the structural class of the STR is summarised as indicated in the key to the right. In cases where two classes are both present at high frequencies, the two classes are given as a split cell in the table. Human structures are taken from the predominant observed class listed at STRBase.nist.gov. Orthologous orangutan (Pab: *Pongo abelii*) alleles are based on the reference sequence. Hsa: *Homo sapiens*; Ptr: *Pan troglodytes*; Ppa: *P. paniscus*; Ggg: *Gorilla gorilla gorilla*; Gbg: *G. beringei graueri*; Gbb: *G. b. beringei*. **b–h** Examples of variation across (sub)species, phylogenetically arranged, for seven STRs (see Figure S2 for further examples). Human structures are from STRBase.nist.gov and [18]. In each case, tetra- or trinucleotide repeat motifs are indicated by boxes coloured according to the keys below. Ranges of repeat numbers within variable arrays are indicated. Where more than one structural class is observed within a *Pan* or *Gorilla* (sub)species, pie-charts indicate their proportions

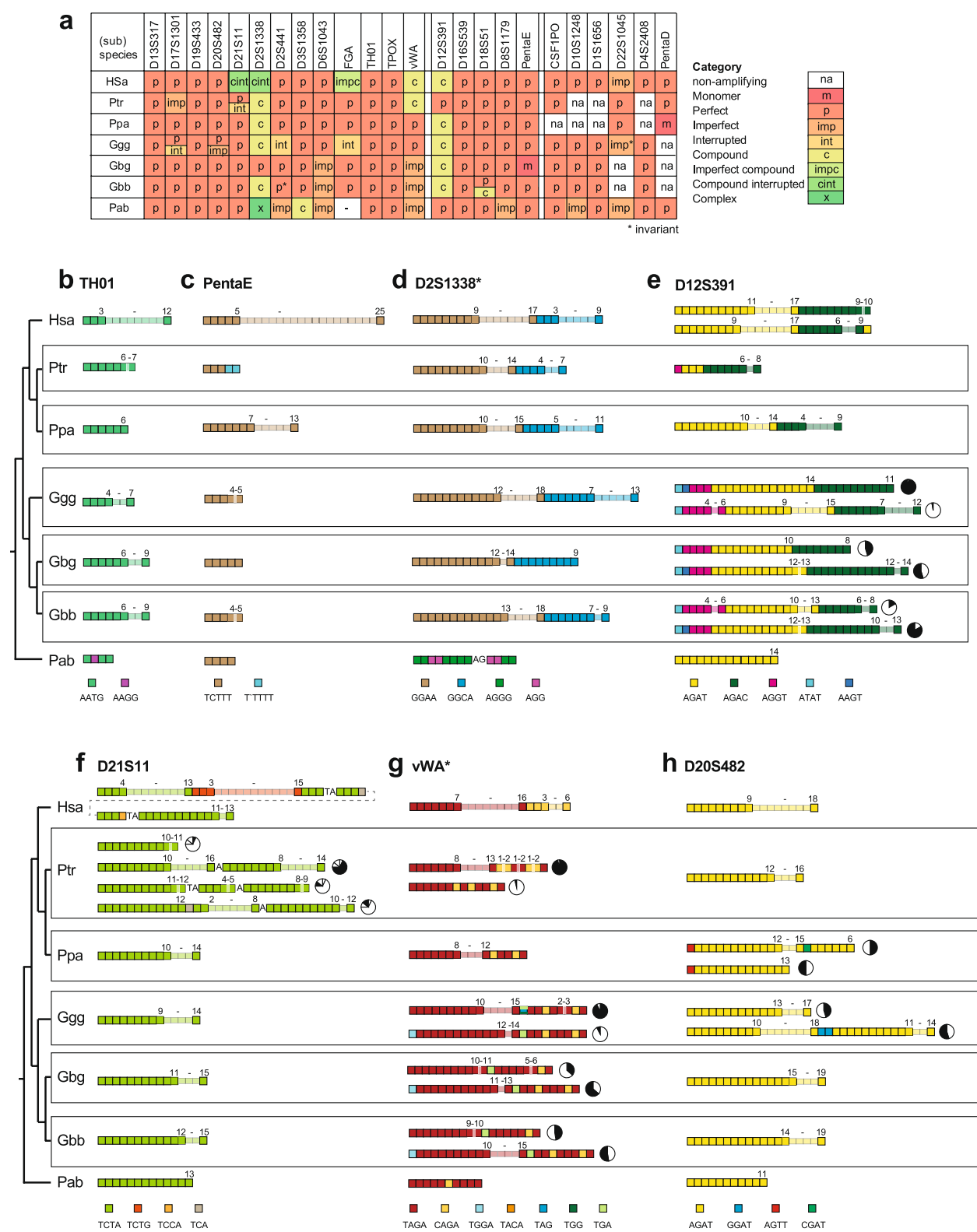


Table 1 Observed per genotype combined RMPs for different great ape (sub)species

(sub) species	Mean no. chromo-somes	No. loci	Mean no. alleles/locus		Combined RMP	
			CE	MPS	CE	MPS
Ppa	6.6	20	3.0	3.1	6.91E-08	5.15E-08
Ptr	20.9	22	5.8	7.0	7.78E-15	2.17E-15
Ggg	22.6	24	6.0	7.3	1.46E-17	1.61E-18
Gbg	10.8	22	3.0	3.2	6.21E-10	5.70E-10
Gbb	15.7	22	4.1	4.3	3.96E-12	8.53E-13

See Table S8 for per-locus details

per locus. RMPs are in all cases lower for MPS than CE allele designations, and in the range 10^{-8} to 10^{-18} . Any comparison with human RMPs, where sample sizes and numbers of observed alleles are much larger, is not very meaningful. For example, the 24 loci analysable in western lowland gorillas give respective RMPs for CE- and MPS-based designations of 1.49×10^{-27} and 1.98×10^{-30} in a sample of 89 Saudi Arabian humans [25].

Between-(sub)species variability of STR genotypes

To compare multilocus STR genotypes for the 13 ‘core’ loci across (sub)species, we carried out cluster analysis using STRUCTURE and DAPC (discriminant analysis of principal components), both for data at the full sequence level and for CE-equivalent (length-based). In STRUCTURE analysis of CE-equivalent data (Figure S5a), the best-supported value of K is 2, in which *Pan* and *Gorilla* form two clusters. DAPC analysis reveals three clusters, with *Gorilla* divided into clusters corresponding to western and eastern species (Figure S5b), reflecting the behaviour of this method in minimising differences within, while maximising differences between, populations. In STRUCTURE analysis of sequence-level data, $K=4$ is best supported, differentiating clusters corresponding to bonobo, chimpanzee, western gorilla and eastern gorilla (Fig. 5a). DAPC analysis gives five clusters, separating out the two eastern gorilla subspecies (Fig. 5b). Sequence-based analysis therefore performs better in distinguishing between (sub)species. Given the sharing of repeat motif variation across *Gorilla* (sub)species (Fig. 4; Figure S1), it seems likely that the differences contributing to differentiation here reflect variation in the flanking sequences.

Discussion

Recent conservation initiatives have witnessed a considerable increase in the use of DNA testing for the implementation of effective wildlife conservation and management plans throughout the world. The current rate of biodiversity loss has prompted researchers to utilise markers that can be readily transferred

between species to facilitate the study of taxa in which allelic diversity is poorly characterised [27, 28]. In this context, aSTRs have been a dominant source of neutral genetic markers for a variety of applications, including individual identification, assessment of population diversity and structure, and evolutionary studies [29]. Cross-species amplification depends on the presence of flanking sequences that, despite sometimes long divergence times, are conserved across organisms, and is directly related to the phylogenetic distance between the source and the target species [30, 31]. This has enabled the exploitation of common sets of PCR primers to type orthologous aSTR loci via capillary electrophoresis (CE) for the study of non-model organisms [17, 29, 32–35]. Following CE, PCR fragment lengths are converted into numbers of repeats at STR regions to produce individual genotypes. Recent studies, however, have identified several caveats to this approach, especially when it is used in cross-species analyses. Firstly, owing to convergent mutations, repetitive regions that are identical by state (i.e. have the same length) may not be identical by descent [36], therefore estimates of differentiation across species can be inaccurate. Secondly, CE fails to distinguish indels occurring within STR flanking sequences from changes in the structure of the repetitive regions, compromising the assessment of the organisation and variability of STRs [17]). As a result, the underlying assumption, under which orthologous STRs are commensurable across species, is often incorrect.

In recent years, the advent of MPS has obviated these problems by allowing researchers to investigate the structures of STR alleles, in virtually unlimited numbers. Consequently, MPS tolerates size homoplasy and the occurrence of overlapping ranges between loci that arise when homologous primers are used to genotype different species, as both STR and flanking sequences may not be invariant across species [17, 29, 32–35]. Because MPS does not rely on length discrimination, primer pairs can be strategically designed to target shorter fragments and increase multiplexing capability, thus making this

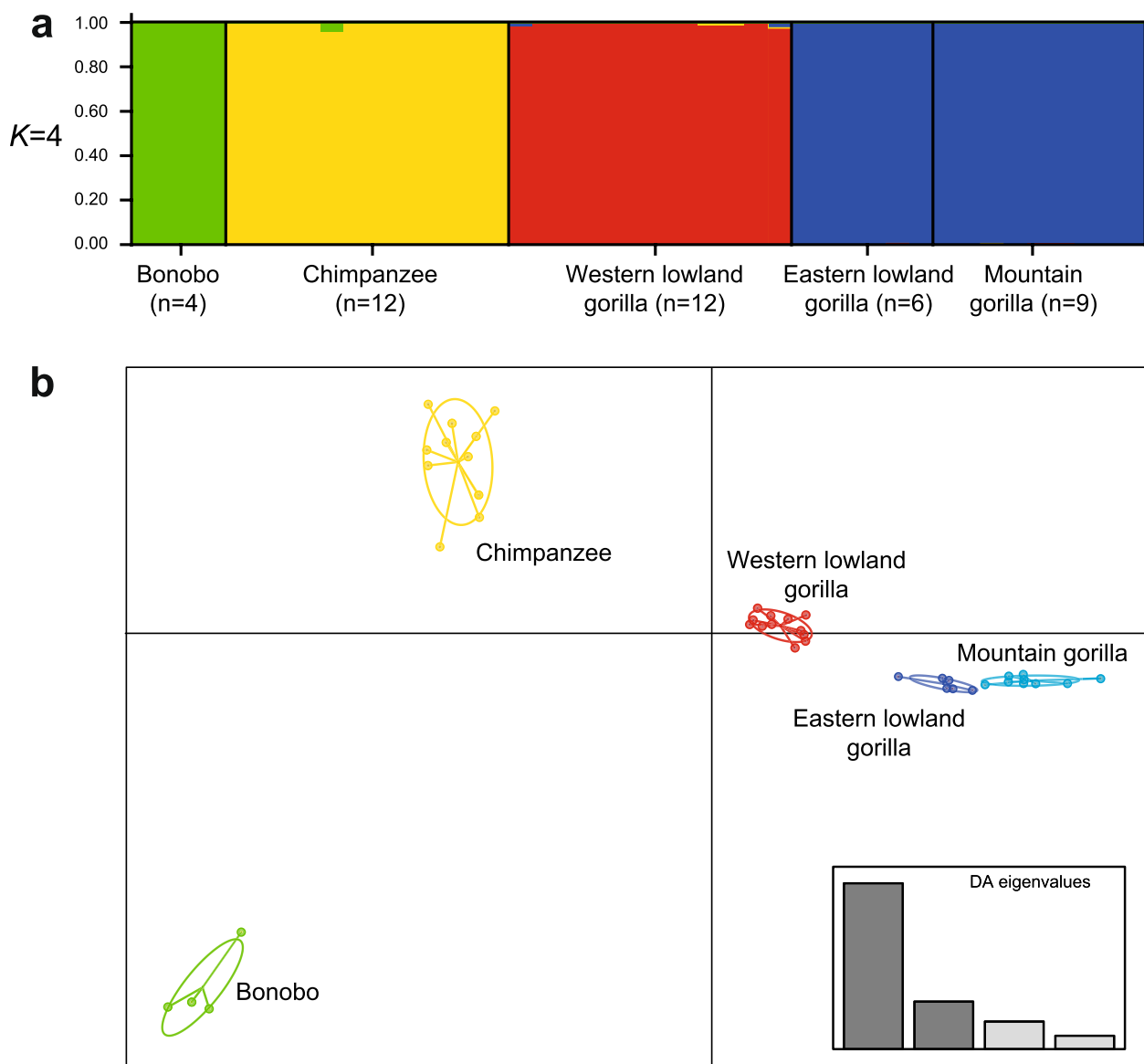


Fig. 5 Cluster analysis based on sequence-based autosomal STR genotypes. **a** Results based on STRUCTURE, for $K=4$; **b** Results based on DAPC analysis. Full sequence information was used here (both array and flanking sequence data). An analysis based on CE-equivalent data is given in Figure S2. Related individuals are removed for this analysis (see Table S1)

technology particularly suitable for the analysis of highly degraded DNA found in non-invasive samples. MPS has been used to sequence 46 STRs from faecal samples of the Iberian wolf [37]. In chimpanzees and bonobos, sequence-based analysis of multiple STRs from faeces has been carried out [38, 39] using a bioinformatic platform developed for calling alleles from Illumina MiSeq data [40].

Here, we applied the human-designed ForenSeq kit to amplify and sequence human loci of forensic interest in 52 DNA samples from chimpanzees, bonobos, and

gorillas, focusing on the results obtained for 27 autosomal STRs (aSTRs). As expected, given the low average sequence divergence between African great ape genomes (~1.3% between human and chimpanzee/bonobo [41]; ~1.75% between human and western lowland gorilla [42]), most of the aSTRs amplified successfully in most cases. Thirteen STRs could be genotyped in all individuals, and a further five showed only individual-level dropouts or sub-threshold amplification. The remaining nine either failed amplification altogether or failed in a particular species or genus. Failure to amplify is likely

due to sequence divergence in primer-binding sites; since the ForenSeq kit's primer sequences are proprietary and therefore not exactly known, this cannot be investigated definitively, but analysis of three loci supports the idea (Figures S1-3).

Our results show that MPS analysis of STR alleles can provide accurate individual and sub-species identification. As was observed previously in species including humans [18, 25], chimpanzees [40] and muskrats [43], our analysis reveals higher diversity of STR alleles than traditional length-based genotyping – though this is not a universal finding, as demonstrated by a study in Vancouver Island marmots [44]. In our study, STR structures show evidence of allele stability over long evolutionary times and reveal unexpectedly high levels of IBD across shared gorilla alleles (the only exception being D8S1179 in eastern gorilla subspecies, which reflects the short divergence time). Contrary to what has been reported in human pedigrees [26], we found that long interrupted alleles share a high degree of polymorphism across species: one speculative explanation for this is possible differences in mutation processes between species, but there may be other explanations. In the future, increasing whole-genome sequence data at the population and pedigree level and the application of genome-wide STR calling tools (e.g. HipSTR [45], LobSTR [46]) should illuminate these questions further.

Despite the advantages of MPS, the widespread adoption of high throughput sequence-based STR typing for wildlife conservation purposes is still hindered by high start-up costs (e.g. for equipment and reagents), labour-intensive sample preparation, and steep learning curves associated with MPS data analysis [47–49]. Additionally, the lack of well-established research facilities in biodiverse countries means that biological samples must be shipped to sites where sequencing can be performed [50]. Stringent international restrictions on the export of endangered species biological samples further contribute to increasing the cost and time of sequencing, de facto limiting the feasibility of DNA testing for wildlife conservation purposes [51, 52].

Nevertheless, recent technological advances have circumvented these issues by greatly reducing the cost for the acquisition of sequencing and laboratory equipment, with positive repercussions for the implementation of wildlife conservation genomics initiatives [47, 48]. In this regard, the commercialisation of portable nanopore sequencing devices by the company Oxford Nanopore Technologies promises to revolutionise the field of molecular ecology by permitting in situ analysis of DNA samples [50, 53–56]. The shift from a laboratory-centralised workflow to on-site DNA analysis overcomes the fundamental challenge of transporting biological material

to a site where sequencing can be performed [50]. While only few studies to date have assessed the applicability of the ONT MinION device for sequencing forensic STRs [57–60], recent findings suggest that STR panels can be compatible with ONT sequencing platforms [47], which opens up new opportunities in the field of wildlife forensics and conservation genetics.

Conclusions

Our results indicate that MPS via a human-designed kit represents an effective method for the analysis of orthologous aSTR loci in non-human great ape species, and it provides reliable identification of individual and (sub) species. Comparison with standard length-based allele definitions shows higher observed allelic diversity and improved (sub)species discrimination.

Methods

DNA samples and data

DNA samples were from a variety of sources including laboratory collections, detailed in Table S1. For chimpanzees, subspecies definition was sometimes unclear, and where it was defined, sample sizes for individual subspecies were small: we therefore considered chimpanzees at the species level. By contrast, gorilla samples were better defined, with at least six individuals in each of three of the four known subspecies, and therefore gorillas were considered at this level. As a result, our comparison groups were five in number: chimpanzee—*Pan troglodytes* ($n=14$), bonobo—*P. paniscus* ($n=4$), western lowland gorilla—*Gorilla gorilla gorilla* ($n=16$), eastern lowland gorilla—*G. beringei graueri* ($n=6$) and mountain gorilla—*G. b. beringei* ($n=12$ [7, 61]). To provide comparative information on the same set of loci in humans we used a published dataset based on analysis of the ForenSeq™ DNA Signature Prep Kit in 89 unrelated Saudi Arabian human males [25], as well as information from STRBase.nist.gov and [18].

Library preparation and sequencing

DNA samples were quantified using the Qubit™ Fluorometer with the Qubit™ dsDNA HS (High Sensitivity) Assay Kit for double-stranded DNA (dsDNA). Sequencing libraries were prepared with the human-based ForenSeq™ DNA Signature Prep Kit according to the manufacturer's recommendations (Verogen®, San Diego, CA, USA). Primer mix A was used to target 58 STRs (27 autosomal STRs, 7 X-STRs and 24 Y-STRs) and 94 identity-informative SNPs (iiSNPs) from 1 ng of template DNA. Details of all loci are available at <https://verogen.com/wp-content/uploads/2022/01/forenseq-dna-signature-prep-reference-guide-PCR1-vd2018005-d.pdf>. Steps for library preparation include amplifying, indexing,

purifying, normalising and pooling, prior to sequencing on an Illumina MiSeq FGx, all of which were performed in accordance with the manufacturer's recommended protocols.

Sequence data analysis

Quality-checked FASTQ files were generated using Trimmomatic v.0.36 [62] for adapter sequence and poor-quality base trimming using the Linux terminal. The threshold for minimum read length was set at 50 bp.

Analysis of human data using the DNA Signature Kit is usually undertaken using the ForenSeq™ Universal Analysis Software (UAS), but for the non-human analysis done here the software FDSTools [63] was employed. This is laborious, but has the advantage that tailored anchor, flanking and repeat-array sequences can be designed, hence obviating the need for a reliable reference genome, which is still lacking for *Gorilla beringei* and most *Pan* sub-species. In order to develop library files for variant calling for *Pan* and *Gorilla*, trimmed bam files were visualised and aligned with the human reference (GRCh38/hg38) using the Integrative Genomic Viewer (IGV) [64] allowing the identification of suitable flanking sequences as anchors [63]. Considering the kit chemistry, which produces short and unreliable second reads, the 5' anchor was set close to the 5' end of the repeat array of each locus, so as to maximise the coverage for each marker. Flanking sequences to be added to the final version of the library file were obtained through repeated runs of FDSTools.

If there are null (non-amplifying) alleles at a given STR locus, these may exist in a heterozygous state, and it then becomes necessary to distinguish between such heterozygotes and true non-null homozygotes, in which two identical alleles are amplified. This was done using a sequence read-depth approach, normalised against known heterozygote calls, since a true homozygote's read-depth should be equal to the sum of two heterozygous alleles (following a previous approach used for duplicated Y-STR alleles [65]). A threshold of ≥ 20 reads per locus was set to call alleles.

Orthologous sequences around three STRs for human, chimpanzee, bonobo and gorilla were retrieved from Multiz alignments [66] within the UCSC Genome Browser ([/genome-euro.ucsc.edu/](http://genome-euro.ucsc.edu/)) and secondary alignments generated using Clustal Omega [67].

Population, forensic and statistical analysis

STRAF [68] was used to calculate forensic statistics, including genotype count (N), allele count based on sequence (N_{all}), observed and expected heterozygosity (H_{obs} and H_{exp}), polymorphism information content (PIC), match probability (PM), power of discrimination

(PD), power of exclusion (PE), and typical paternity index (TPI).

Clustering of genetically similar individuals was investigated using both STRUCTURE [69], and discriminant analysis of principal components (DAPC). As different species are present in our data set, we applied STRUCTURE v.2.3.4 excluding admixture, carrying out five independent runs iterated for 150,000 Markov Chain Monte Carlo (MCMC) repetitions including 50,000 as burn-in for $K=1-10$. The output was analysed using the ΔK method for the detection of the optimal number of clusters [70], using STRUCTURE HARVESTER v0.694 [71].

DAPC was conducted using the package adegenet (version 2.1–3) [72] implemented in R version 3.6.3 [73]. For DAPC, the function `find.clusters()` was used to determine the optimal cluster number without prior information, and the Bayesian information criterion (BIC) was used to identify abrupt changes in fit models for successive runs of increasing k-means clustering with $K=1-8$. The number of PCs to retain was cross-validated using the function `xvalDapc()` with 50 repetitions in order to avoid overfitting.

ML-Relate [22] was used to screen the sample set for closely related individuals within (sub)species. Based on this, together with some prior information (Table S1) some individuals were removed for some analyses, as described in the first paragraph of the Results section.

In considering STR repeat arrays across species, we consider four basic types: perfect (an uninterrupted array of a single repeat type, e.g. $[\text{GATA}]_n$); interrupted (two or more arrays of the same repeat type interrupted by non-repeat material, e.g. $[\text{GATA}]_n\text{NNNNN}[\text{GATA}]_m$); imperfect (two or more arrays of the same repeat type interrupted by repeat-derived material, e.g. $[\text{GATA}]_n\text{GACA}[\text{GATA}]_m$ or $[\text{GATA}]_n\text{GAT}[\text{GATA}]_m$); compound (two or more variable arrays of different repeat types of the same length, e.g. $[\text{GATA}]_n[\text{GACA}]_m$). We also include two hybrid categories, compound interrupted (two or more variable arrays of different repeat types of the same length, interrupted by non-repeat material, e.g. $[\text{GATA}]_n\text{NNNNN}[\text{GACA}]_m$), and imperfect compound (two or more arrays of different repeat types of the same length, interrupted by repeat-derived material, e.g. $[\text{GATA}]_n\text{AATA}[\text{GACA}]_m$).

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12862-024-02324-0>.

Supplementary Material 1.

Supplementary Material 2.

Acknowledgements

We thank Yahya Khubrani and Tunde Huszar for assistance. We gratefully acknowledge colleagues who contributed DNA samples, particularly Chris Tyler-Smith and Yali Xue, and Lisa Gillespie (Twycross Zoo). We thank NUCLEUS Genomic Services at the University of Leicester for access to Illumina MiSeq sequencing. This research used the ALICE High Performance Computing Facility at the University of Leicester for data analysis.

Authors' contributions

EF: Investigation, Methodology, Formal analysis, Visualisation. JHW: Conceptualisation, Supervision. MAJ: Conceptualisation, Supervision, Resources, Visualisation. All authors: Writing: original draft, reviewing and editing.

Funding

EF was supported by a PhD studentship from the Natural Environment Research Council CENTA doctoral training programme (grant no. NE/L002493/1).

Data availability

Genotype data generated during this study are included in this published article and its supplementary information files. Sequence data have been deposited in NCBI GenBank under accession numbers PQ397797 - PQ399638.

Declarations

Ethics approval and consent to participate

This research was approved by the University of Leicester's Animal Welfare and Ethical Review Body (ref.: AWERB/2021/159), and all great ape samples were taken by qualified veterinarians. All methods were performed in accordance with the relevant guidelines and regulations, and the study is reported in accordance with ARRIVE guidelines (<https://arriveguidelines.org>). Consent to participate is not relevant. For the purpose of open access, the author has applied a Creative Commons Attribution license (CC BY) to any Author Accepted Manuscript version arising from this submission.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

¹Department of Genetics, Genomics & Cancer Sciences, University of Leicester, University Road, Leicester LE1 7RH, UK. ²Current address: Faculty of Science & Engineering, Swansea University, Swansea, UK.

Received: 29 November 2023 Accepted: 17 October 2024

Published online: 31 October 2024

References

- Maxwell SL, Fuller RA, Brooks TM, Watson JE. Biodiversity: The ravages of guns, nets and bulldozers. *Nature*. 2016;536(7615):143–5.
- Williams PH, Burgess ND, Rahbek C. Flagship species, ecological complementarity and conserving the diversity of mammals and birds in sub-Saharan Africa. *Animal Conserv*. 2000;3:249–60.
- Kuhlwlilm M, de Manuel M, Nater A, Greminger MP, Krutzen M, Marques-Bonet T. Evolution and demography of the great apes. *Curr Opin Genet Dev*. 2016;41:124–9.
- Rivers MC, Brummitt NA, Ludhadha EN, Meagher TR. Do species conservation assessments capture genetic diversity? *Global Ecol Conserv*. 2014;2:81–7.
- Ouborg NJ, Pertoldi C, Loeschcke V, Bijlsma RK, Hedrick PW. Conservation genetics in transition to conservation genomics. *Trends Genet*. 2010;26(4):177–87.
- Keller LF, Waller DM. Inbreeding effects in wild populations. *Trends Ecol Evol*. 2002;17:230–41.
- Xue Y, Prado-Martinez J, Sudmant PH, Narasimhan V, Ayub Q, Szpak M, Frandsen P, Chen Y, Yngvadottir B, Cooper DN, et al. Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. *Science*. 2015;348(6231):242–5.
- Jobling MA, Hollox EJ, Hurler ME, Kivisild T, Tyler-Smith C. Human evolutionary genetics. 2nd ed. New York and London: Garland Science; 2014.
- Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G, et al. Great ape genetic diversity and population history. *Nature*. 2013;499(7459):471–5.
- Vigilant L, Guschanski K. Using genetics to understand the dynamics of wild primate populations. *Primates*. 2009;50(2):105–20.
- Allendorf FW. Genetics and the conservation of natural populations: allozymes to genomes. *Mol Ecol*. 2017;26(2):420–30.
- Jobling MA, Gill P. Encoded evidence: DNA in forensic analysis. *Nat Rev Genet*. 2004;5:739–51.
- Linacre A. *Animal Forensic Genetics* Genes (Basel). 2021;12(4):515.
- Wall JD. Great ape genomics. *ILAR J*. 2013;54(2):82–90.
- Thakur M, Chandra K, Sahajpal V, Samanta A, Sharma A, Mitra A. Functional validation of human-specific PowerPlex(R) 21 System (Promega, USA) in chimpanzee (*Pan troglodytes*). *BMC Res Notes*. 2018;11(1):695.
- Singh A, Sahajpal V, Thakur M, Sharma LK, Chandra K, Bhandari D, Sharma A. Applicability of human-specific STR systems, GlobalFiler PCR Amplification Kit, Investigator 24plex QS Kit, and PowerPlex(R) Fusion 6C in chimpanzee (*Pan troglodytes*). *BMC Res Notes*. 2021;14(1):212.
- Kwong M, Pemberton TJ. Sequence differences at orthologous microsatellites inflate estimates of human-chimpanzee differentiation. *BMC Genomics*. 2014;15:990.
- Gettings KB, Kiesler KM, Faith SA, Montano E, Baker CH, Young BA, Guerrieri RA, Vallone PM. Sequence variation of 22 autosomal STR loci detected by next generation sequencing. *Forensic Sci Int Genet*. 2016;21:15–21.
- Beasley J, Shorrock G, Neumann R, May CA, Wetton JH. Massively parallel sequencing and capillary electrophoresis of a novel panel of falcon STRs: Concordance with minisatellite DNA profiles from historical wildlife crime. *Forensic Sci Int Genet*. 2021;54:102550.
- Churchill JD, Schmedes SE, King JL, Budowle B. Evaluation of the Illumina® Beta Version ForenSeq DNA Signature Prep Kit for use in genetic profiling. *Forensic Sci Int Genet*. 2016;20:20–9.
- Just RS, Moreno LI, Smerick JB, Irwin JA. Performance and concordance of the ForenSeq system for autosomal and Y chromosome short tandem repeat sequencing of reference-type specimens. *Forensic Sci Int Genet*. 2017;28:1–9.
- Kalinowski ST, Wagner AP, Taper ML. ML-Relate: a computer program for maximum likelihood estimation of relatedness and relationship. *Mol Ecol Notes*. 2006;6:576–9.
- Makova KD, Pickett BD, Harris RS, Hartley GA, Cechova M, Pal K, Nurk S, Yoo D, Li Q, Hebbard P et al. The complete sequence and comparative analysis of ape sex chromosomes. *Nature*. 2024;630(8016):401–11.
- Sullivan KM, Mannucci A, Kimpton CP, Gill P. A rapid and quantitative DNA sex test: fluorescence-based PCR analysis of X-Y homologous gene amelogenin. *Biotechniques*. 1993;15(4):636–8 640–631.
- Khubrani YM, Hallast P, Jobling MA, Wetton JH. Massively parallel sequencing of autosomal STRs and identity-informative SNPs highlights consanguinity in Saudi Arabia. *Forensic Sci Int Genet*. 2019;43:102164.
- Sun JX, Helgason A, Masson G, Ebenesersdottir SS, Li H, Mallick S, Gnerre S, Patterson N, Kong A, Reich D, et al. A direct characterization of human mutation based on microsatellites. *Nat Genet*. 2012;44(10):1161–5.
- Barbara T, Palma-Silva C, Paggi GM, Bered F, Fay MF, Lexer C. Cross-species transfer of nuclear microsatellite markers: potential and limitations. *Mol Ecol*. 2007;16(18):3759–67.
- Maduna SN, Rossouw C, Roodt-Wilding R. Bester-van der Merwe AE: Microsatellite cross-species amplification and utility in southern African elasmobranchs: A valuable resource for fisheries management and conservation. *BMC Res Notes*. 2014;7:352.
- FitzSimmons NN, Moritz C, Moore SS. Conservation and dynamics of microsatellite loci over 300 million years of marine turtle evolution. *Mol Biol Evol*. 1995;12(3):432–40.
- Miles LG, Isberg SR, Glenn TC, Lance SL, Dalzell P, Thomson PC, Moran C. A genetic linkage map for the saltwater crocodile (*Crocodylus porosus*). *BMC Genomics*. 2009;10:339.
- Primmer CR, Painter JN, Koskinen MT, Palo JU, Merilä J. Factors affecting avian cross-species microsatellite amplification. *J Avian Biol*. 2005;36(4):348–60.

32. Blanquer-Maumont A, Crouau-Roy B. Polymorphism, monomorphism, and sequences in conserved microsatellites in primate species. *J Mol Evol*. 1995;41(4):492–7.
33. Brohede J, Ellegren H. Microsatellite evolution: polarity of substitutions within repeats and neutrality of flanking sequences. *Proc Biol Sci*. 1999;266(1421):825–33.
34. Clisson I, Lathuillière M, Crouau-Roy B. Conservation and evolution of microsatellite loci in primate taxa. *Am J Primatol*. 2000;50(3):205–14.
35. Gugerli F, Brodbeck S, Holderegger R. Insertions–deletions in a microsatellite flanking region may be resolved by variation in stuttering patterns. *Plant Mol Biol Report*. 2008;26:255–62.
36. Estoup A, Jarne P, Cornuet JM. Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Mol Ecol*. 2002;11(9):1591–604.
37. Salado I, Fernandez-Gil A, Vila C, Leonard JA. Automated genotyping of microsatellite loci from feces with high throughput sequences. *PLoS ONE*. 2021;16(10):e0258906.
38. Bonnin N, Piel AK, Brown RP, Li Y, Connell AJ, Avitto AN, Boubli JP, Chitayat A, Giles J, Gundlapally MS, et al. Barriers to chimpanzee gene flow at the south-east edge of their distribution. *Mol Ecol*. 2023;32(14):3842–58.
39. Wroblewski EE, Guethlein LA, Anderson AG, Liu W, Li Y, Heisel SE, Connell AJ, Ndjanga JN, Bertolani P, Hart JA, et al. Malaria-driven adaptation of MHC class I in wild bonobo populations. *Nat Commun*. 2023;14(1):1033.
40. Barbian HJ, Connell AJ, Avitto AN, Russell RM, Smith AG, Gundlapally MS, Shazad AL, Li Y, Bibollet-Ruche F, Wroblewski EE, et al. CHIMP: An automated high-throughput microsatellite genotyping platform reveals greater allelic diversity in wild chimpanzees. *Ecol Evol*. 2018;8(16):7946–63.
41. Prufer K, Munch K, Hellmann I, Akagi K, Miller JR, Walenz B, Koren S, Sutton G, Kodira C, Winer R, et al. The bonobo genome compared with the chimpanzee and human genomes. *Nature*. 2012;486(7404):527–31.
42. Scally A, Dutheil JY, Hillier LM, Jordan GE, Goodhead I, Herrero J, Hobolth A, Lappalainen T, Mailund T, Marques-Bonet T, et al. Insights into hominid evolution from the gorilla genome sequence. *Nature*. 2012;483(7388):169–75.
43. Darby BJ, Erickson SF, Hervey SD, Ellis-Felege SN. Digital fragment analysis of short tandem repeats by high-throughput amplicon sequencing. *Ecol Evol*. 2016;6(13):4502–12.
44. Barrett KG, Amaral G, Elphinstone M, McArdie ML, Davis CS, Janes JK, Carnio J, Moehrenschrager A, Gorrell JC. Genetic management on the brink of extinction: sequencing microsatellites does not improve estimates of inbreeding in wild and captive Vancouver Island marmots (*Marmota vancouverensis*). *Conserv Genet*. 2022;23(2):417–28.
45. Willems T, Zielinski D, Yuan J, Gordon A, Gymrek M, Erlich Y. Genome-wide profiling of heritable and de novo STR variations. *Nat Methods*. 2017;14(6):590–2.
46. Gymrek M, Golan D, Rosset S, Erlich Y. IobSTR: A short tandem repeat profiler for personal genomes. *Genome Res*. 2012;22(6):1154–62.
47. Hall CL, Kesharwani RK, Phillips NR, Planz JV, Sedlazeck FJ, Zascavage RR. Accurate profiling of forensic autosomal STRs using the Oxford Nanopore Technologies MinION device. *Forensic Sci Int Genet*. 2022;56: 102629.
48. Hall CL, Zascavage RR, Sedlazeck FJ, Planz JV. Potential applications of nanopore sequencing for forensic analysis. *Forensic Sci Rev*. 2020;32(1):23–54.
49. Zascavage RR, Shewale SJ, Planz JV. Deep-sequencing technologies and potential applications in Forensic DNA testing. *Forensic Sci Rev*. 2013;25(1–2):79–105.
50. Pomerantz A, Peñañel N, Arteaga A, Bustamante L, Pichardo F, Coloma LA, Barrio-Amorós CL, Salazar-Valenzuela D, Prost S. Real-time DNA barcoding in a rainforest using nanopore sequencing: opportunities for rapid biodiversity assessments and local capacity building. *Gigascience*. 2018;7(4):gij033.
51. Fernandez F. The greatest impediment to the study of biodiversity in Colombia. *Caldasia*. 2011;33:iii–v.
52. Gilbert N. Biodiversity law could stymie research. *Nature*. 2010;463(7281):598.
53. Blanco MB, Greene LK, Rasambainarivo F, Toomey E, Williams RC, Andrianandrasana L, Larsen PA, Yoder AD. Next-generation technologies applied to age-old challenges in Madagascar. *Conserv Genet*. 2020;21:785–93.
54. Chang JJ, Ip YC, Ng CS, Huang D. Takeaways from Mobile DNA Barcoding with BentoLab and MinION. *Genes (Basel)*. 2020;11(10):1121.
55. Menegon M, Cantaloni C, Rodriguez-Prieto A, Centomo C, Abdelfattah A, Rossato M, Bernardi M, Xumerle L, Loader S, Delledonne M. On site DNA barcoding by nanopore sequencing. *PLoS ONE*. 2017;12(10): e0184741.
56. Pomerantz A, Sahlin K, Vasiljevic N, Seah A, Lim M, Humble E, Kennedy S, Krehenwinkel H, Winter S, Ogden R, et al. Rapid in situ identification of biological specimens via DNA amplicon sequencing using miniaturized laboratory equipment. *Nat Protoc*. 2022;17:1415–43.
57. Asogawa M, Ohno A, Nakagawa S, Ochiai E, Katahira Y, Sudo M, Osawa M, Sugisawa M, Imanishi T. Human short tandem repeat identification using a nanopore-based DNA sequencer: a pilot study. *J Hum Genet*. 2020;65(1):21–4.
58. Cornelis S, Gansemans Y, Deleye L, Deforce D, Van Nieuwerburgh F. Forensic SNP genotyping using nanopore MinION sequencing. *Sci Rep*. 2017;7:41759.
59. Ren ZL, Zhang JR, Zhang XM, Liu X, Lin YF, Bai H, Wang MC, Cheng F, Liu JD, Li P, et al. Forensic nanopore sequencing of STRs and SNPs using Verogen's ForenSeq DNA Signature Prep Kit and MinION. *Int J Legal Med*. 2021;135(5):1685–93.
60. Tytgat O, Gansemans Y, Weymaere J, Rubben K, Deforce D, Van Nieuwerburgh F. Nanopore sequencing of a forensic STR multiplex reveals loci suitable for single-contributor STR profiling. *Genes (Basel)*. 2020;11(4):381.
61. Pawar H, Rymbekova A, Cuadros-Espinoza S, Huang X, De Manuel M, Van der Valk T, Lobon I, Alvarez-Estape M, Haber M, Dolgova O, et al. Ghost admixture in eastern gorillas. *Nat Ecol Evol*. 2023;7:1503–14.
62. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114–20.
63. Hoogenboom J, van der Gaag KJ, de Leeuw RH, Sijen T, de Knijff P, Laros JF. FDSTools: A software package for analysis of massively parallel sequencing data with the ability to recognise and correct STR stutter and other PCR or sequencing noise. *Forensic Sci Int Genet*. 2017;27:27–40.
64. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. *Nat Biotechnol*. 2011;29(1):24–6.
65. Huszar TI, Jobling MA, Wetton JH. A phylogenetic framework facilitates Y-STR variant discovery and classification via massively parallel sequencing. *Forensic Sci Int Genet*. 2018;35:97–106.
66. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, et al. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res*. 2004;14(4):708–15.
67. Sievers F, Higgins DG. Clustal Omega, Accurate Alignment of Very Large Numbers of Sequences. In: *Multiple Sequence Alignment Methods*. Edited by Russell DJ. Totowa, NJ: Humana Press; 2014:105–116.
68. Gouy A, Zieger M. STRAF—A convenient online tool for STR data evaluation in forensic genetics. *Forensic Sci Int Genet*. 2017;30:148–51.
69. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics*. 2000;155:945–59.
70. Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol*. 2005;14(8):2611–20.
71. Earl DA, Volholdt BM. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv Genet Resour*. 2012;4:359–61.
72. Jombart T, Ahmed I. Adegenet 1.3–1: new tools for the analysis of genome-wide SNP data. *Bioinformatics*. 2011;27(21):3070–1.
73. R Core Team: R: A language and environment for statistical computing; <http://www.R-project.org/>. In: Vienna, Austria: R Foundation for Statistical Computing; 2014.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.