View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Explore Bristol Research

Frandsen, P., Fontsere, C., Nielsen, S. V., Hanghøj, K., Castejon-Fernandez, N., Lizano, E., Hughes, D., Hernandez-Rodriguez, J., Korneliussen, T. S., Carlsen, F., Siegismund, H. R., Mailund, T., Marques-Bonet, T., & Hvilsom, C. (2020). Targeted conservation genetics of the endangered chimpanzee. *Heredity*, *2020*. https://doi.org/10.1038/s41437-020-0313-0

Peer reviewed version

Link to published version (if available):
10.1038/s41437-020-0313-0

Link to publication record in Explore Bristol Research
PDF-document

This is the author accepted manuscript (AAM). The final published version (version of record) is available online via Springer Nature at https://www.nature.com/articles/s41437-020-0313-0 . Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research
General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available: http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/

1 Targeted conservation genetics of the endangered chimpanzee

2

3 Peter Frandsen*[1,2,†], Claudia Fontsere*[3], Svend Vendelbo Nielsen[4], Kristian Hanghøj[2],

4 Natalia Castejon-Fernandez[4], Esther Lizano[3], David Hughes[5,6], Jessica Hernandez-

5 Rodriguez[3], Thorfinn Korneliussen[2], Frands Carlsen[1], Hans Redlef Siegismund[2], Thomas

6 Mailund[4], Tomas Marques-Bonet[3,7,8,9], Christina Hvilsom[1].

7

8 1 Research and Conservation, Copenhagen Zoo, Roskildevej 38, 2000 Frederiksberg, Denmark.

9 2 Section for Computational and RNA Biology, Department of Biology, University of Copenhagen, Ole

10 Maaløes Vej 5, 2200 Copenhagen, Denmark.

11 3 Institute of Evolutionary Biology, (UPF-CSIC), PRBB, Dr. Aiguader 88, 08003, Barcelona, Spain.

12 4 Bioinformatics Research Center, Department of Mathematics, Aarhus University, C. F. Møllers Allé

13 8, 8000 Aarhus C, Denmark.

14 5 MRC Integrative Epidemiology Unit at Universit of Bristol, Bristol, BS8 2BN, UK.

15 6 Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, BS8 2BN, UK.

16 7 Catalan Institution of Research and Advanced Studies (ICREA), Passeig de Lluís Companys 23,

17 08010, Barcelona, Spain.

18 8 CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology

19 (BIST), Baldiri I Reixac 4, 08028 Barcelona, Spain.

20 9 Institut Català de Paleontologia Miquel Crusafant, Universitat Autònoma de Barcelona, Edifici ICTA-

21 ICP, c/ Columnes s/n, 08193 Cerdanyala del Vallès, Barcelona, Spain.

Running title: Targeted conservation genetics

22 * These authors contributed equally to this work

23

24 † Corresponding author:

25 Peter Frandsen, PhD

26 Research and Conservation

27 Copenhagen Zoo

28 Roskildevej 38

29 2000 Frederiksberg, Denmark

30 E-mail: pef@zoo.dk

31

32 Running title

33 Targeted conservation genetics

34

35 Keywords

36

37 Conservation genetics, *ex situ* breeding programmes, illegal wildlife trade, *in situ* conservation,

38 non-invasive sampling.

39

40 Word count

41 6 793

Running title: Targeted conservation genetics

42 ## Abstract

43

44 Populations of the common chimpanzee (*Pan troglodytes*) are in an impending risk of going
45 extinct in the wild as a consequence of damaging anthropogenic impact on their natural habitat
46 and illegal pet and bushmeat trade. Conservation management programmes for the chimpanzee
47 have been established outside their natural range (*ex situ*), and chimpanzees from these
48 programmes could potentially be used to supplement future conservation initiatives in the wild
49 (*in situ*). However, these programmes have often suffered from inadequate information about
50 the geographical origin and subspecies ancestry of the founders. Here, we present a newly
51 designed capture array with ~60 000 ancestry informative markers used to infer ancestry of
52 individual chimpanzees in *ex situ* populations and determine geographical origin of confiscated
53 sanctuary individuals. From a test panel of 167 chimpanzees with unknown origins or
54 subspecies labels, we identify 90 suitable non-admixed individuals in the European Association
55 of Zoos and Aquaria (EAZA) *Ex situ* Programme (EEP). Importantly, another 46 individuals
56 have been identified with admixed subspecies ancestries, which therefore over time, should be
57 naturally phased out of the breeding populations. With potential for future re-introduction to
58 the wild, we determine the geographical origin of 31 individuals that were confiscated from the
59 illegal trade and demonstrate the promises of using non-invasive sampling in future
60 conservation action plans. Collectively, our genomic approach provides an exemplar for *ex situ*
61 management of endangered species and offers an efficient tool in future *in situ* efforts to combat
62 the illegal wildlife trade.

65 ## Introduction

66 In an era of human-induced acceleration of species loss, often referred to as the sixth mass
67 extinction era (Ceballos *et al.*, 2015), conservation efforts to save endangered species are calling
68 for novel approaches to mitigate the ongoing extinction crisis.

69 Since the discovery of the common chimpanzee (*Pan troglodytes*), humans have
70 been drawn to this charismatic species. Despite our fascination, human activities have led to a
71 drastic decline in the population size of the chimpanzee. In the last two decades, chimpanzees
72 have been listed as 'Endangered' at the species level in the IUCN Red List, with one of the four
73 recognized subspecies, the western chimpanzee (*P. t. verus*) being listed as 'Critically
74 Endangered' in the latest assessment (Humle *et al.*, 2016). Human encroachment on the natural
75 range of the chimpanzee has further caused an intensified conflict between humans and
76 chimpanzees (Hockings *et al.*, 2015). One by-product of the human wildlife conflicts has been
77 a rise in opportunistic trafficking of chimpanzees, which, in recent years has become more
78 organized and systematic (Stiles *et al.*, 2013). Besides wildlife trade, other continuous threats
79 including habitat destruction, poaching for local consumption, and human linked disease
80 outbreaks has led to a drastic decline in the wild chimpanzee populations (Humle *et al.*, 2016).
81 Together, these threats emphasize the importance of a '*One Plan Approach*' conservation
82 programme linking *in situ* and *ex situ* efforts (Traylor-Holzer *et al.*, 2019) to prevent the
83 predicted extinction of chimpanzees within the current century (Estrada *et al.*, 2017).

84 Outside Africa, several regional chimpanzee conservation programmes exist, with
85 the largest being the European Association of Zoos and Aquaria (EAZA) *Ex situ* Programme
86 (henceforth EEP). The EEP targets the subspecies level and today, breeding programmes for
87 two of the four recognized subspecies, the western chimpanzee (*P. t. verus*) and the central

88 chimpanzee (*P. t. troglodytes*) have been established (Carlsen and de Jongh, 2018) The primary
89 aim of the EEP is to safeguard the survival of healthy self-sustaining populations targeting the
90 taxonomical level of subspecies (Carlsen and de Jongh, 2018). The extant EEP populations
91 consist of wild founders and descendants thereof. However, in times before high resolution
92 genetic technologies were available and even in its early development, knowledge of subspecies
93 labels and relatedness between founders were inaccurate and has led to admixture of subspecies
94 in the captive population (Hvilsom *et al.*, 2013). Early attempts to add a genetic layer to the
95 EEP management has confirmed that knowledge of subspecies ancestries, inbreeding and
96 relatedness estimates are instrumental to preserve genetic diversity in captive populations
97 (Hvilsom *et al.*, 2013). Yet, most recent attempts based on microsatellite markers (Hvilsom *et*
98 *al.*, 2013), did not have the necessary resolution or predictive power to disentangle several
99 generations of hybridizations in the EEP breeding population. Although we still do not know
100 its full extent, hybridization between neighbouring subspecies of chimpanzees has been shown
101 to occur in the wild (Hvilsom *et al.*, 2013; Prado-Martinez *et al.*, 2013; de Manuel *et al.*, 2016)
102 and therefore, it is not unlikely that some founders in the EEP harbour shared ancestries from
103 more than one subspecies. The current strategy in the EEP targets un-admixed breeding
104 individuals and with the current methods, it is impossible to tell if small admixture proportions
105 arose from an early *ex situ* hybridization event followed by several generations of backcrossing
106 or from a naturally admixed founder. Therefore, founders are potentially being wrongfully
107 excluded from the breeding programme due to their admixed ancestry.

108 The scenario outlined above, is by no means exclusive to captive management of
109 chimpanzees but extends to practically any *ex situ* management programme of populations
110 based on wild born founders with a taxonomical subdivision. When morphology alone is
111 insufficient in taxonomical delimitation between subspecies or the targeted conservation units,

112 genetic resources becomes increasingly important. Yet, the choice of genetic resource is not
113 always trivial. In response to a growing availability of different types of genetic resources with
114 widely different applications, several studies have tried to develop guidelines based on the
115 management requirements (see e.g. Grueber *et al.*, 2019; Norman *et al.*, 2019).

116 As described, the complexities in EEP management of chimpanzees requires a
117 new rigorous solution as previous attempts using either mitochondrial DNA, or microsatellites
118 have proven insufficient. With a genome-wide set of ancestry informative markers, we predict
119 that it will be possible to obtain the desired depth of predictive power to infer ancestries in the
120 present and previous generations and classify individuals with shared ancestries as either
121 descendants of admixed founders or *ex situ* hybrids. This could provide the foundation of a
122 possible reassessment of the current management strategies under the EEP and in turn, allow
123 for inclusion of wild born hybrids in the breeding programme if these are found to resemble the
124 diversity of the species in the wild.

125 In their natural range, chimpanzees have become a commodity and organized
126 illegal trade poses a serious threat to the species. Over the period from 2005-2011 a reported
127 minimum of 643 chimpanzees were harvested from the wild for illegal trade activities (Stiles
128 *et al.*, 2013). However, extrapolations suggest that twenty times as many individuals have
129 become victims of the illegal wildlife trade in that relatively short time span (Stiles *et al.*, 2013).
130 While most of the captured individuals are sold as bushmeat, a considerable number of mostly
131 juvenile chimpanzees end up in the illegal pet trade. When conservation authorities confiscate
132 illegally kept chimpanzees, they are placed at wildlife sanctuaries, often arbitrarily based on
133 availability of space and proximity to the confiscation site. Whilst some of the rescued
134 chimpanzees require specialised lifetime care, others may be successfully reintroduced into
135 their natural habitats after extensive preparation (Beck *et al.*, 2007). For chimpanzees destined

137 to lifetime care, proper management planning requires knowledge about relatedness among
138 sanctuary chimpanzees in order to set up family groups. In cases, were chimpanzees are suitable
139 for reintroduction, knowledge of geographical origin is essential as several studies have shown
140 lineage-specific adaptations in all four subspecies in their respective geographical ranges (e.g.
141 Nye *et al.*, 2018). In the first complete geo-referenced genomic map of the chimpanzee, de
142 Manuel *et al.* (2016) portrayed a strong correlation between geographical origin and genetic
143 diversity, where the former can be inferred solely based on the latter. Employing genetic testing
144 at the site of confiscation (e.g. airports, transport hubs) would enable conservation authorities
145 to infer geographical origin of confiscated individuals and with time, strive to facilitate a return
146 of these individuals to a protected area in the region where they were captured. Alternatively,
147 confiscated chimpanzees can be sent to a neighbouring sanctuary with housing capacity, where
148 specialized care and rehabilitation can be provided, and if possible, future reintroduction can
149 be planned. Genetic testing at an early stage of confiscation also has the potential to understand
150 and help break trafficking routes and enable CITES authorities to track and enforce law control
151 in situations where chimpanzees are housed in disreputable zoos and entertainment facilities.
152 However, to be a practical tool in conservation, the genetic test needs to maximise the inference
153 accuracy, require very little investment, and pose as little risk to animal health as possible.
154 These requirements limit our choice of applicable data types. With a novel SNP array design
155 where the level of genetic information is only surpassed by costly whole genome sequencing,
156 we argue that our approach constitutes the most cost-efficient option for conservation
157 management in situations where funding is often scarce and demands for rigorous solutions are
158 high.

159 Using a selected panel of 59 800 targeted ancestry informative markers, we
160 demonstrate the ability to infer robust estimates of ancestry in several generations of the EEP

162 chimpanzee breeding population. We further show how this set of ancestry informative markers
163 can be used to determine geographical origin of confiscated individuals and demonstrate how
164 these methodologies can readily be applied to using non-invasive sampling. In combination,
165 these methods harbour great potential for future global management plans for the chimpanzee
166 and provides an important exemplar for management of endangered species in general.

167 ## Materials and Methods

168 Samples

169 A total of 179 chimpanzee samples were collected and analysed in the present study (Suppl.
170 File S1 SequencingStatistics.xlsx). For the purpose of cross-validation between sequencing
171 batches and to test our methodology on non-invasive hair sampling, a number of individuals
172 were sequenced in duplicates and triplicates, which lead to 167 unique individuals. 136 from
173 the EEP population housed in 47 different European zoos and primate rescue and rehabilitation
174 centres (Table S2), and 31 from eight sanctuaries across Africa (Table S3). To form a reference
175 panel, we complemented the genotypes of EEP and sanctuary chimpanzees with whole genome
176 data from 58 geo-referenced wild-born chimpanzees, representing the four chimpanzee
177 subspecies, and additionally, one known admixed individual (*Ptv-Donald*) and one known
178 descendant of wild born individuals (*Ptv-Clint*) (Prado-Martinez *et al.*, 2013; de Manuel *et al.*,
179 2016).

180 DNA extraction and library preparation

181 DNA was extracted using a standard phenol-chloroform protocol. Samples were quantified with
182 a Qubit 2.0 fluorometer, Qubit® dsDNA BR Assay Kit (Thermo Fisher Scientific). DNA library
183 preparation was carried out in three batches. For the first batch (24 samples) and the second
184 batch (63 samples), extracted DNA was sheared with a Covaris S2 ultrasonicator using the
185 recommended fragmentation settings to obtain a 350 bp insert size. For the third batch (92
186 samples) DNA was sheared using the recommended settings of Covaris S2 to obtain 200 bp
187 insert size. The first batch of 24 libraries (with 6 more samples not used in this study) were
188 prepared using 1.5 μg of DNA and the TruSeq DNA HT Sample Prep Kit (Illumina), following

189 manufacturer's instructions and 14 cycles of PCR amplification. The second batch of 63
190 samples (with 17 more samples not used in this study) were processed using 500 ng of starting
191 DNA and following the custom dual-indexed protocol described by Kircher *et al.* (2012) and
192 12 cycles of PCR were done for indexing and amplification. The remaining 92 samples (with 2
193 more samples not used in this study) were processed using 200 ng of starting DNA following
194 the BEST protocol (Carøe *et al.*, 2018) with minor modifications (initial reaction volume was
195 incremented up to 50 µl to accommodate a larger amount of starting DNA and 10 cycles of
196 PCR amplification). For this third batch, we used inline barcoded short adapters with the same
197 seven nucleotide barcodes at the P5 and P7 adapters. Clean-ups were done using homemade
198 SPRI beads (Rohland and Reich, 2012). Libraries were eluted in 25 µl of ddH_2O and quantified
199 with an Agilent 2100 Bioanalyzer using a DNA 1000 assay kit.

200 Target Capture Design

201 We performed a target capture enrichment experiment using baits synthesized by Agilent
202 Technologies. We targeted 59 800 autosomal sites that were ancestry informative markers and
203 designed using the panTro4 genome. Marker selection was done using published chimpanzee
204 genomes (Prado-Martinez *et al.*, 2013) and by applying a sparse PCA method on 10 Mbp bins
205 of the genomes (Lee *et al.*, 2012). Variant sites were then weighted to identify the most
206 informative markers for the first two principal components (PCs) and 200 AIMs were extracted
207 per segment. The genome was binned to have an unbiased and evenly distributed sampling of
208 the genome and to have enough resolution to provide percent ancestry in highly admixed
209 individuals.

210 For target enrichment hybridization, libraries were pooled equimolarly based on
211 a library prep method to obtain a total of 19 pools (see Supporting Information for a detailed

212 description of the targeted enrichment hybridization). PCR amplification product was cleaned
213 up using our homemade SPRI beads (Rohland and Reich, 2012). Each enriched sample was
214 then quantified on a NanoDrop, BioAnalyzer and then sequenced.

215 Fastq filtering and mapping

216 Libraries were sequenced on five lanes of a HiSeq 2500 ultra-high-throughput sequencing
217 system, one lane for 24 chimpanzee samples, two lanes for 63 chimpanzee samples and two
218 lanes for the remaining 92 samples. Inline barcoded libraries captured in the same pool (92 from
219 Batch 3) were de-multiplexed using Sabre software v. 1.0 (https://github.com/najoshi/sabre).
220 Prior to mapping, paired-end reads were filtered to remove PCR duplicates using
221 FASTUNIQ v. 1.1 (Xu *et al.*, 2012) and adaptors (*Illuminaclip*) and low quality first five bases
222 in a read (*Slidingwindow:5:20*) were trimmed using TRIMMOMATIC v. 0.36 (Bolger *et al.*,
223 2014). Overlapping reads were merged with a minimum overlap of 10 bp and minimum length
224 of final read to 50 bp, using PEAR v. 0.9.6 (Zhang *et al.*, 2014). Then, reads were mapped using
225 BWA v. 0.7.12 (Li and Durbin, 2009) to the Hg19 reference genome (GRCh37, Feb.2009
226 (GCA_000001405.1)). PCR Duplicates were removed using PICARDTOOLS v. 1.95
227 (http://broadinstitute.github.io/picard/) with the *MarkDuplicates* option. Further filtering of the
228 reads was done to discard secondary alignments and reads with mapping quality lower than 30
229 using SAMTOOLS v. 1.5 (Li *et al.*, 2009). We then filtered for the targeted space (4 bp around
230 the selected SNP) using BEDTOOLS intersect v. 2.16.2 (Quinlan and Hall, 2010).
231 The total aligned reads were calculated by dividing the number of uniquely
232 mapped reads (the remaining reads after removing duplicates) by the number of production
233 reads. The on-target aligned reads were calculated by dividing the target filtered reads by the
234 production reads. Then, the total coverage was calculated by dividing aligned bases by the

235 length of the assembly (Hg19) and the target effective coverage dividing the on-target bases by
236 the targeted genomic space. Finally, the enrichment factor of the capture performance was
237 calculated by taking the ratio between the on-target reads by total mapped reads over the target
238 size by genome size.

239 Variant calling

240 Variant discovery was performed using GATK '*Unified Genotyper*' (DePristo *et al.*, 2011) for
241 each sample independently with the following parameters *-out_mode EMIT_ALL_SITES -*
242 stand_call_conf 5.0 *-stand_emit_conf* 5.0 *-A BaseCounts -A GCContent -A*
243 *RMSMappingQuality -A BaseQualityRankSumTest*. Genotypes from each sample were
244 combined in a single VCF using GATK `*CombineVariants*' (DePristo *et al.*, 2011) with -
245 *genotypeMergeOptions UNIQUIFY –excludeNonVariant* parameters. We also included the
246 genotype information of available whole genome data of aforementioned 58 wild-born geo-
247 referenced chimpanzees and *Ptv-Donald* and *Ptv-Clint* (Prado-Martinez *et al.*, 2013; de Manuel
248 *et al.*, 2016). Unless differently stated in separate analysis, the variants with a depth of coverage
249 less than 3, a quality score less than 30 (QUAL<30), minor allele frequency of 0.005 and a
250 missingness rate of > 60 % were removed using VCFTOOLS v. 0.1.12 (Danecek *et al.*, 2011).
251 We only kept the genotypes that were inside the target space by using the -bed option in
252 VCFTOOLS v. 0.1.12 (Danecek *et al.*, 2011).
253

254 Ancestry inference and inbreeding

255 We inferred proportions of shared ancestries in two approaches. First, to detect underlying
256 genetic structure with a reduction of the dimensionality in the data, we performed a principle
257 component analysis (PCA) using EIGENSOFT v. 6.1.3. (Price *et al.*, 2006). All samples were

258 included without pruning of sites in linkage disequilibrium or minor allele frequencies, in order
259 to avoid exclusion of fixed sites between populations. Analyses on shared ancestry in *ex situ*
260 and sanctuary populations were done with reference to the genetic structure in the wild born
261 individuals with ADMIXTURE v. 1.2 (Alexander *et al.*, 2009). To avoid any bias introduced
262 from a joint analysis with related individuals, each of the 167 unique individuals from the EEP
263 and sanctuary populations were analysed separately one by one against a reference panel of all
264 wild born individuals. After applying a minor allele frequency filter (--maf 0.05) in PLINK v.
265 1.07 (Purcell *et al.*, 2007) to exclude sites polymorphic in only one individual, a set of 45 542
266 sites where kept for analysis. Each analysis of ADMIXTURE v. 1.2 (Alexander *et al.*, 2009)
267 was iterated 100 times under an EM optimization algorithm and termination criteria of a log-
268 likelihood increase of 10^{-5} between iterations. A value of K=4 was chosen to obtain clusters in
269 line with the four recognized subspecies of chimpanzees. To assess convergence, the 100
270 iterations were evaluated to ensure that iterations did not differ by more than one log-likelihood
271 value.
272 For each of the individuals with admixture coefficients >0.99, we applied
273 NGSRELATEv2 (Hanghøj *et al.*, 2019) to estimate pairwise relatedness and individual
274 inbreeding coefficients based on population allele frequencies from each of the inferred
275 admixture clusters, after excluding minor allele frequencies (MAF) <0.05 (see Supplementary
276 Information).
277
278 *Hybrid classification*
279 To further explore the ancestry sharing in the EEP and sanctuary individuals and to be able to
280 differentiate shared ancestry originating from the founding individuals and EEP hybrids, we
281 developed a hidden Markov model (available on GitHub

282 http://gihub.com/svendvn/ImmediateAncestry) to allow for an inference of the posterior
283 proportion of ancestries in the three immediate previous generations. Additionally, we estimate
284 where these immediate ancestors belong in the pedigree. For full documentation of the model,
285 see Supplementary Information.
286
287 *Re-assignment of geographical origin*
288 We applied the methodology of ORIGEN (CRAN R package https://cran.r-
289 project.org/web/packages/OriGen/index.html) as described by Rañola, Novembre, & Lange
290 (2014), to re-assign the geographical origin of confiscated sanctuary individuals. We applied
291 the *FitOriGenModelFindUnknowns* parameter to the 1 690 highest ranked informative markers
292 to assign individual geographical origin onto the allele frequency surface, inferred from the
293 wild born reference panel.
294
295 *Non-invasive sampling*
296 To test our targeted capture approach on non-invasively collected hair samples, we sequenced
297 three individuals where we had both blood samples, whole genome reference data and hair
298 samples. Hair samples were capture sequenced using the same methodology as described above
299 for blood samples, except we added a pre-treatment step in the DNA extraction of hair samples
300 to enhance lysis of keratin. Shared ancestry and geo-graphical origin was analysed as described
301 above.

302 ## Results

303

304 *Capture sequencing and variant calling*

305 First we quantified and assessed the performance of our capture methodology in the selected
306 targeted space. We wanted to ensure sufficient representation of the targeted genomic regions
307 to reliably call the selected variants. In a total of five lanes of HiSeq2500 we obtained ~1 000
308 million production reads, and on average, each sample received five million reads. After
309 removing PCR duplicates and considering only primary alignments with a mapping quality
310 higher than 30, we obtained an average of 3.6 million mapped reads (74.31%) per sample
311 (Suppl. File S1). The average effective target coverage on the 59,800 autosomal SNPs was
312 21.69 X with 12.91% of on-target reads (four base pairs around the targeted SNP, Suppl. File
313 S1) which fulfilled our theoretical prediction of 20 X. In terms of capture performance, this last
314 statistic is an underestimate since the full length of the capture bait is 120 base pairs and in this
315 analysis, we only considered the four base pairs around the targeted SNP. Still, we considered
316 it to be more accurate since it is the true space where the informative SNP falls. Lastly, to
317 summarise the performance of the capture methodology, we computed the enrichment factor
318 that relates the number of aligned reads on the target space divided by the production reads,
319 with the size of the target space to the size of the whole genome. The resulting enrichment
320 factor of 89.31 X reasserts the advantages of capture to ensure enough coverage for genotyping
321 purposes (Suppl. File S1).

322 Considering all samples without overlap, we obtained a total of ~150 000
323 genotypes. However the average number of SNPs called per sample was 30 337 sites passing
324 the filtering steps (MAF 0.05 and max-missing 0.6, after we excluded samples '12103' and

325 ‘12349’ due to low coverage). The maximum number of SNPs called in one individual was 51
326 952 and the minimum was 10 783 (Figure S1). Among the variation found in western
327 chimpanzees, only a third of these were polymorphic in the western chimpanzee (Table S1),
328 yet, of the 46 260 polymorphic sites, 15 738 were private in the western chimpanzee (Figure
329 S2). For fixed sites, the western chimpanzee also had the highest number of private sites (Figure
330 S2). Among the four subspecies, the eastern chimpanzee had the highest total number of
331 polymorphic sites, followed by the central chimpanzee, Nigerian-Cameroon chimpanzee, and
332 western chimpanzee, respectively (Table S1).
333

334 Population structure, ancestry, and inbreeding

335 The major axes of variance in EEP and sanctuary individuals were explored with a principal
336 component analysis with reference to the panel of geo-referenced individuals with known
337 subspecies label from Prado-Martinez *et al.* (2013) and de Manuel *et al.* (2016). The first
338 principal component (PC1) explained 70.49 % of the variance in our data, separating the
339 western chimpanzees from the three other subspecies in the reference panel (Figure 1B). With
340 16.53 % of explained variance, PC2 separated the Nigerian-Cameroon chimpanzee, central
341 chimpanzee, and eastern chimpanzee.

342 The majority of the 167 tested individuals from the EEP and sanctuary
343 populations, clustered with either of the four reference populations, while a minor part of the
344 individuals scattered in between the defined populations (Figure 1B). The inferred ancestries
345 from the ADMIXTURE analysis conveyed the same patterns of genetic population structure
346 separating the geo-referenced individuals into four distinct clusters with varying degree of
347 ancestry sharing between geographically neighbouring subspecies (Figure 1C). With this as a
348 reference, we assigned the EEP and sanctuary individuals into groupings in terms of their

349 ancestry patterns of either non-admixed or hybrids with multiple components of ancestry. Of
350 the 167 tested individuals, 121 could be confidently assigned as non-admixed (admixture
351 proportion from one subspecies ≥0.99). All 31 sanctuary individuals were assigned to
352 subspecies level without evidence of admixture, where five clustered with the western
353 chimpanzee, one with the Nigerian-Cameroon chimpanzee, one with the central chimpanzee,
354 and 24 with the eastern chimpanzee. In the EEP population, we inferred the majority of the 90
355 non-admixed individuals to belong to the western chimpanzee (41), three with the Nigerian-
356 Cameroon chimpanzee, 25 with the central chimpanzee, and 21 with the eastern chimpanzee.
357 Of the remaining 46 EEP individuals, 38 were inferred to be hybrids with two ancestry
358 components while the last eight had three ancestry components.

359 Of all the individuals from the EEP, sanctuary, and the reference panel with
360 admixture coefficients >0.99, relatedness estimates were low (Figure S3-S6) while we
361 identified eight individuals with inbreeding coefficients above 0.2 (Figure 1D). Within these
362 eight individuals, all four subspecies were represented, as were wild and captive born
363 chimpanzees.

364

365 *Hybrid classification*

366 To explore ancestry patterns in the previous three generations, we ran our ancestry classification
367 model going back k = 3 generations and visualized the number of loci each ancestor in
368 generation k contributed to the ancestral informative part of the genome (see Supplementary
369 Information). In general, our method correctly estimated the expected ancestries of our
370 reference panel individuals (Figure 2A). Several eastern and Nigerian-Cameroonian
371 chimpanzee individuals were estimated to contain substantial ancestry components from the
372 mutually neighbouring central subspecies. The known hybrid *Ptv-Donald* (Prado-Martinez *et*

374 *al.*, 2013) was estimated by the method to be at least 1/8 central chimpanzee, yet the large
375 proportion of loci that were assigned to the central chimpanzee in the posterior distribution
376 might suggest that *Ptv-Donald* could be as much as 1/4 central chimpanzee.

377 Similar to the ancestries inferred with ADMIXTURE, our method classified a
378 large fraction of the EEP and sanctuary individuals to have ancestors from only one subspecies
379 in the last three generations (Figure 1C, Figure 2B, Figure 2C). In general, individuals inferred
380 to belong to the eastern chimpanzee had third generation ancestors of central chimpanzee
381 ancestry (Figure 2B, Figure 2C). Similarly, four inferred central chimpanzees in the EEP
382 population, showed small proportions of ancestry from the Nigeria-Cameroon chimpanzee.
383 Comparably, one sanctuary individual, *Edward*, was inferred here as a Nigeria-Cameroon
384 chimpanzee with a small proportion of central chimpanzee ancestry. However, performing
385 posterior correction by replacing the low central chimpanzee ancestor with another high
386 posterior Nigeria-Cameroon ancestor, would likely make a more accurate estimate. Among the
387 admixed EEP individuals, our model showed similar results to those obtained with
388 ADMIXTURE but as ancestry patterns became increasingly complex (more than two ancestral
389 subspecies) our inferred posterior proportions became increasingly uncertain (Figure 2B, Figure
390 S12). We further observed that in some cases, small deviating (possibly deep coalescing)
391 segments could have let the model to prefer configurations in the ancestry patterns to switch
392 halves (Figure 2C), while the correct configuration would probably be a simple case of
393 hybridization in the parent generation.

394

395 *Geo-localisation*

396 Based on an allele frequency surface map, built from our reference panel of wild born
397 individuals, we determined the geographical origin of all 31 sanctuary individuals. Generally,

398 the inferred probabilities of geographical origin gave accurate estimates (i.e. high probabilities
399 assigned to just one or a few adjacent grid cells) for all sanctuary individuals (Figure 3). Also,
400 all individuals assigned to the natural range of their inferred subspecies label. The majority of
401 our tested sanctuary individuals belonged to the eastern chimpanzee where the geographical
402 origins were inferred to six provinces along the eastern part of the natural range of the
403 subspecies. Seven of the eastern individuals had low probability estimates divided over a cluster
404 of adjacent grid cells, with the highest ranking cell assigned probability of less than 0.1. All
405 five western chimpanzee individuals were assigned to the same grid cell in the eastern limits of
406 their range. The single individual from the Nigeria-Cameroon chimpanzee was assigned to a
407 locality in Cameroon while the one central chimpanzee was assigned to the coastal region of
408 Gabon.
409
410 Non-invasive sampling
411 Expanding our targeted capture approach to non-invasively collected hair samples,
412 corroborated the results obtained with blood samples. ADMIXTURE estimates converged to
413 the same result in the two sample types for all tested individuals and geographical origin was
414 assigned to the same locality between samples (Figure 4, Figure S13-17). Compared to the
415 reference, ancestry estimates in our capture array approach did not always reveal the minor
416 components of shared ancestries found when including all variant sites in the genome (Figure
417 4).

418 ## Discussion

419 As an exemplar for conservation genetics of endangered species, we have designed a novel
420 capture array that targets identified ancestry informative markers across the genomes of 24 wild
421 born chimpanzees (Prado et al., 2013) and the PanTro4 reference genome. Acknowledging that
422 the selected ancestry markers were derived from a relatively limited set of genomes, which
423 could potentially introduce an ascertainment bias towards specific subspecies, we confirmed
424 that our design has the power to correctly identify the subspecies of an extended panel of newly
425 sequenced chimpanzee genomes (de Manuel *et al.*, 2016) (Figure 1). Based on this proof of
426 concept, we sequenced 167 chimpanzees from the EEP and sanctuary populations and analysed
427 subspecies ancestries and geographical origin. We further show how this approach can be
428 extended to non-invasive samples with robust results.
429

430 Ancestry of the ex situ population

431 In our test panel of 167 chimpanzees, 136 were from the EEP population housed at 47 European
432 zoos and rehabilitation centres. Based on information on disembarkation or place of capture,
433 we know that the majority of chimpanzees who founded the current EEP population came from
434 West Africa. In accordance to this, a majority of the 90 non-admixed individuals could be
435 assigned to the western chimpanzee (Figure 1C). Our findings confirm that for the western
436 chimpanzee, early efforts of the EEP that sought to identify a core group of non-admixed
437 western chimpanzees using mitochondrial DNA (Jepsen and Carlsen *unpublished*) and
438 microsatellites (Hvilsom *et al.*, 2013), have been momentarily successful. Yet, using similar
439 methodologies, previous attempts have only managed to identify a small group of central
440 chimpanzees since the breeding effort for this subspecies was established (Carlsen and de

441 Jongh, 2018). Here, we identify 25 central chimpanzee individuals in the EEP population that
442 show no evidence of shared ancestry with other subspecies (Figure 1C), and hence from a
443 genetic viewpoint, would qualify as a suitable bolster to the current breeding population.
444 Similarly, the 21 inferred non-admixed eastern chimpanzee individuals could form the crucial
445 starting point from where a separate breeding effort could be established under the EEP. In
446 contrast to this, of our tested 136 EEP individuals, only three could be assigned to the Nigerian-
447 Cameroonian subspecies (Figure 1C) and in general, of the four subspecies, the Nigeria-
448 Cameroon chimpanzee is by far the least represented in the EEP population (Carlsen and de
449 Jongh, 2018). Yet, with our targeted capture approach, it will now be feasible to scan the
450 remaining EEP population (~1 000 housed individuals) for additional non-admixed chimpanzee
451 individuals in order to explore the possibilities of creating separate breeding populations for the
452 two remaining subspecies.

453 Still, with a presumed small EEP population of eastern and Nigerian-
454 Cameroonian chimpanzees, it might prove difficult to avoid inbreeding, although our estimates
455 suggests, that high inbreeding coefficients are not exclusive to these particular subspecies. In
456 fact, individuals with inbreeding coefficients in the range of 0.2-0.4 were found in each of the
457 four subspecies and includes both wild and captive born individuals (Figure 1D). It is therefore
458 difficult to establish whether the amount of inbreeding in EEP individuals are a consequence of
459 breeding among closely related individuals or whether it stems from inbred founders. In a few
460 cases, like individual '*14073*', we know from reliable pedigree information, that this individual
461 is the offspring of two full-siblings (Carlsen and de Jongh, 2018). For the large majority of the
462 EEP population, this knowledge is not available or are associated with high levels of
463 uncertainties. Together with accurate ancestry inferences, genetically-based inbreeding

464 estimates will be of high importance in management of the breeding population as will other
465 factors such as age, fecundity, behaviour, and housing capacities.

466 Of our 136 tested EEP individuals, 46 were inferred to be of hybrid origin (Figure
467 1C). In terms of distinguishing founder individuals with shared ancestry components (wild born
468 hybrids) from *ex situ* hybrids, our ancestry analyses show that the majority of our inferred
469 hybrids are between non-neighbouring populations in the wild (e. g. between the western
470 chimpanzee and either of the three other subspecies) and are therefore most likely the result of
471 hybridization in the EEP breeding population. From a management standpoint, these should
472 eventually be phased out of the breeding programme. Yet, some known hybrids have been
473 allowed to breed under the current management. This has been done with the purpose to
474 maintain population numbers in an interim period while the populations reach their target size
475 and also to allow experienced females to pass on up-bringing behaviour to young individuals
476 in the housed groups. To explore the extent of wild born hybrids in the EEP and the possibility
477 of including these in the breeding efforts, we developed a new method for hybrid classification
478 that can trace ancestry patterns three generations back. This could possibly allow us to
479 distinguish between hybrids bred in captivity and wild born hybrids, where the latter could be
480 included in breeding programmes, as they represent natural processes in the wild. However,
481 two key requirements to such an inclusion are a better understanding of the extent of
482 hybridization in the wild and an EEP management decision on what a suitable admixture
483 threshold would be.

484 As validation for the hybrid classification model (see also Supplementary.
485 Information), our method infers the known hybrid background of *Ptv-Donald* to have received
486 at least 12.5 percent of its ancestry from the central chimpanzee, which is in the range of what
487 was previously estimated using whole genome sequencing data (Prado-Martinez *et al.*, 2013).

488 Yet, in the EEP population, only a few of the inferred hybrids fit with the expectations of
489 ancestry patterns in wild born hybrids. The majority of the inferred hybrids include a western
490 chimpanzee ancestry component (Figure 2B), which is highly unlikely to occur in the wild due
491 to the vast geographical distance to any neighbouring subspecies (Figure 1A). Of the eight
492 inferred hybrids with adjacent distribution ranges, one central/Nigerian-Cameroonian and
493 seven central/eastern hybrids (Figure 2B), we know from studbook information that all eight
494 individuals were captive born (Carlsen and de Jongh, 2018) (Table S2). The only cases where
495 our model might have picked up remnants from natural hybridizations are the ancestry
496 components of central chimpanzee in what we inferred to be non-admixed eastern chimpanzees
497 using ADMIXTURE (Figure 1C, Figure 2B). However, this could likely be due to a general
498 limitation of our model to separate these two subspecies due to their evolutionary close
499 relationship and history of allele sharing (Prado-Martinez *et al.*, 2013; de Manuel *et al.*, 2016).
500 Although we did not identify any wild born hybrids in the tested set of individuals, our model
501 predictions will be highly useful in terms of pinpointing the timing of admixture and help to
502 illuminate blanks in the studbook regarding possible sires.
503

504 *Sanctuary ancestry and geographical origin*

505 In contrast to the predominance of western chimpanzee individuals in the EEP population, the
506 majority of the tested sanctuary individuals are inferred to belong to the eastern chimpanzee.
507 Of the 31 tested individuals, we only find four that can be assigned to the western chimpanzee
508 and a single individual from each of the Nigeria-Cameroon chimpanzee and the central
509 chimpanzee (Figure 1C). When exploring ancestry patterns in the last three generations, we
510 obtained similar results as in the EEP population, where small posterior proportions of central
511 chimpanzee were found in individuals of the eastern chimpanzee (Figure 2C). This is most

512 likely due to the limitations of our model when it comes to distinguishing shared alleles between
513 these two subspecies, and we do not infer any geographical origin close to possible contact
514 zones between the two subspecies (Figure 3).

515 For western and Nigerian-Cameroonian chimpanzees, we obtained high
516 probabilities in the assigned origins but with little spatial resolution. Essentially, all five western
517 chimpanzee individuals assign to the same grid cell. As de Manuel *et al.* (2016) have previously
518 shown, population structure inferred in the western and Nigerian-Cameroonian populations,
519 may not offer enough resolution to provide fine scale determination of geographical origin. To
520 improve origin estimates in these populations, it is crucial to obtain a better representation of
521 georeferenced samples across their distribution ranges. This has been achieved for most of the
522 central and eastern chimpanzee ranges, but with only one central chimpanzee individual
523 (*Doris*), we cannot fully evaluate the prediction power and resolution for this subspecies.
524 Nevertheless, the estimated geographical origin of *Doris* is very close to the reported
525 confiscation site (Table S3), which gives us some assurance that future efforts to determine
526 origins in the central chimpanzee will be possible. With a larger set of individuals from the
527 eastern chimpanzee, we can start to appreciate the full potential of the method. The 24 analysed
528 individuals can be assigned to geographical origins in six localities along the eastern edge of
529 the distribution range of the eastern chimpanzee, where the majority originates from two
530 locations in the northern and southern regions of the Democratic Republic of Congo (DRC)
531 (Figure 3). First of all, this might tell us that these regions are heavily affected by poaching and
532 illegal trafficking, although the abundance of confiscation sites might also be biased by the
533 locality of contributing sanctuaries. Only further testing of individuals from sanctuaries across
534 the species range will allow us to assess regional threat levels. However, with the inferred
535 origins of the eastern chimpanzee individuals all along the eastern edge of the range, we can

536 conclude that the threats are not confined to only two regions for this subspecies but are
537 distributed across the eastern boarders of the DRC.

538 When comparing the inferred geographical origins with the reported confiscation
539 sites for all our tested sanctuary individuals (Table S3), it becomes apparent that the trafficking
540 routes generally operate within a relatively local scale. Overall, we see that most of the tested
541 individuals originate from locations that are within close proximity to where they have been
542 confiscated, though with two notable exceptions, *Louise* and *Edward*. *Louise* was confiscated
543 in Moscow, Russia and inferred to have originated from West Africa, while *Edward* was
544 confiscated in Nairobi Airport, Kenya with inferred origin in Cameroon. This confirms that the
545 illegal trade of wild chimpanzees spans beyond country borders and the African continent as
546 reported in Stiles *et al.* (2013). Both individuals are now housed in sanctuaries where
547 specialized care can be provided, yet, in these cases, both individuals have been placed in
548 sanctuaries far from their geographical origin and possibly within mixed subspecies groups
549 (other individuals from these sanctuaries have been assigned to different subspecies). Without
550 proper knowledge of their ancestry, sanctuaries might face the same challenges as we have seen
551 in the EEP population, with admixture of subspecies as a result of (unintended) breeding.
552 Genetic testing at an early stage could help to ameliorate these challenges and as we have
553 shown, our genomic approach extents to non-invasive sampling (Figure 4), making these
554 methods both an accurate and practical tool in conservation efforts to help combat the illegal
555 trade of chimpanzees.

556 We further predict that this approach will be self-empowering as sampling gaps
557 in the distribution range of the chimpanzee are continuously covered and DNA extraction
558 methods for non-invasive samples improve. This will significantly advance our predictive

559 power of geographical origin and provide valuable insight to shared ancestries in natural
560 populations with positive knock-on effects to hybrid assessment in the *ex situ* populations.
561
562 Our capture array approach of targeting ancestry informative markers offers a
563 standardized and cost-effective method that accurately guides *ex situ* and *in situ* conservation
564 management programmes. At the current rate of decline, chimpanzees are predicted to go
565 extinct within the current century (Estrada *et al.*, 2017). Conservation efforts might therefore,
566 in a foreseeable future, be obligated to supplement wild populations with individuals from the
567 *ex situ* populations as a last resort to prevent them from going extinct. Should it come to this,
568 our approach facilitates the safeguarding of genetically self-sustainable populations that will
569 have preserved a genetic profile that resembles their wild counterparts.
570 The current extinction crisis however, extends well beyond chimpanzees and the
571 demand for molecular genetics to help guide future population management programmes is
572 immense, ranging across the taxonomical scale of birds, reptiles, amphibians, and mammals.
573 For the latter alone, more than ten EEP genetic projects are underway and globally, regional
574 zoo associations are undertaking molecular genetic studies for which the present study serves
575 as an important blueprint for linking *in situ* and *ex situ* conservation efforts.

576 Acknowledgements

577 The authors would like to thank all institutions who provided samples for this study: Stichting
578 AAP Amersfoort Zoo, Antwerp Zoo, Zoo Delle Star ad Aprilia, Royal Burgers' Zoo Arnhem,
579 Monde Sauvage Safari, Badoca Safari Park, Barcelona Zoo, Zoo Parc de Beauval, Bioparc
580 Valencia, Borås Zoo, Parco Natura Viva, Edinburgh Zoo, Ölands Zoo, Plättli Zoo, Givskud
581 Zoo/Zootopia, Safaripark Beekse Bergen, Hodonin Zoo, Xanthus Zoo, Kolmården Zoo,
582 Krakow Zoo, Kristiansand Zoo, Lagos Zoo, Le Pal Zoo, Leipzig Zoo, Liberec Zoo, Lisbon
583 Zoo, Madrid Zoo, Magdeburg Zoo, Olmense Zoo, Zoo Osnabrück, African Safari, Plaisance
584 du Touch, Plzen Zoo, Centro de Rescate de Primates Rainfer, Zoological Center Tel Aviv -
585 Ramat Gan, Safari Ravenna, Zoo di Roma, Leintal Zoo, Serengeti-park Hodenhagen, Reserve
586 Africaine de Sigean, Wilhelma Zoo, Loro Parque Zoo, Touroparc, Twycross Zoo, AAP
587 Primadomus, Warsaw Zoo, Schwabenpark, Zagreb Zoo, Cente International de Recherches
588 Médicales de Franceville, Chimfunshi Wildlife Orphanage, Jeunes Animaux Confisques au
589 Katanga, Ngamaba Island Chimpanzee Sanctuary, Tacugama Chimpanzee Sanctuary, The
590 Chimpanzee Conservation Center. We further wish to thank Tom de Jongh and Lisbeth
591 Borbye for valuable language editing and comments to the manuscript and Abigail Ramsøe
592 for early developmental stages of the hybrid classification model.

593 PF is supported by the Innovation Fund Denmark doctoral fellowship programme
594 and the Candys Foundation. CF is supported by "la Caixa" doctoral fellowship programme.
595 TMB is supported by BFU2017-86471-P (MINECO/FEDER, UE), U01 MH106874 grant,
596 Howard Hughes International Early Career, Obra Social "La Caixa" and Secretaria
597 d'Universitats i Recerca and CERCA Programme del Departament d'Economia i Coneixement
598 de la Generalitat de Catalunya (GRC 2017 SGR 880)

600 ## Conflict of Interest

601 The authors declare no conflicts of interest.

602

603 ## Data Archiving

604 Data will be archived at a publicly accessible repository (Dryad) as a VCF file containing all
605 samples included in the analyses.

606 **REFERENCES**

607 Alexander DH, Novembre J, Lange K (2009). Fast Model-Based Estimation of Ancestry
608 in Unrelated Individuals. *Genome Res* **19**: 1655–1664.
609 Beck B, Walkup K, Rodriques M, Unwin S, Travis D, Stoinski T (2007). *Best Practice*
610 *Guidelines for the Re-introduction of Great Apes*. Gland, Switzerland.
611 Bolger AM, Lohse M, Usadel B (2014). Trimmomatic: a flexible trimmer for Illumina
612 sequence data. *Bioinformatics* **30**: 2114–2120.
613 Carlsen F, de Jongh T (2018). *European Studbook for chimpanzee (Pan troglodytes)*.
614 Copenhagen.
615 Carøe C, Gopalakrishnan S, Vinner L, Mak SST, Sinding MHS, Samaniego JA, *et al.*
616 (2018). Single-tube library preparation for degraded DNA. *Methods Ecol Evol* **9**:
617 410–419.
618 Ceballos G, Ehrlich PR, Barnosky AD, García A, Pringle RM, Palmer TM (2015).
619 Accelerated modern human–induced species losses: Entering the sixth mass
620 extinction. *Sci Adv* **1**: e1400253.
621 Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, *et al.* (2011). The
622 variant call format and VCFtools. *Bioinformatics* **27**: 2156–2158.
623 DePristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, *et al.* (2011). A
624 framework for variation discovery and genotyping using next-generation DNA
625 sequencing data. *Nat Genet* **43**: 491–8.
626 Estrada A, Garber PA, Rylands AB, Roos C, Fernandez-duque E, Fiore A Di, *et al.*
627 (2017). Impending extinction crisis of the world's primates: Why primates matter.
628 *Sci Adv* **3**.
629 Grueber CE, Fox S, McLennan EA, Gooley RM, Pemberton D, Hogg CJ, *et al.* (2019).

630 Complex problems need detailed solutions: Harnessing multiple data types to
631 inform genetic management in the wild. *Evol Appl* 12: 280–291.
632 Hanghøj K, Moltke I, Andersen PA, Manica A, Korneliussen TS (2019). Fast and
633 accurate relatedness estimation from high-throughput sequencing data in the
634 presence of inbreeding. 8: 1–9.
635 Hockings KJ, McLennan MR, Carvalho S, Ancrenaz M, Bobe R, Byrne RW, *et al.*
636 (2015). Apes in the Anthropocene: Flexibility and survival. *Trends Ecol Evol* 30:
637 215–222.
638 Humle T, Maisels F, Oates JF, Plumtre A, Williamson EA (2016). Pan troglodytes.
639 *IUCN Red List Threat Species*.
640 Hvilsom C, Frandsen P, Børsting C, Carlsen F, Sallé B, Simonsen BT, *et al.* (2013).
641 Understanding geographic origins and history of admixture among chimpanzees
642 in European zoos, with implications for future breeding programmes. *Heredity*
643 *(Edinb)* 110: 586–93.
644 IUCN (2015). IUCN Red List of Threatened Species. *Version 20153*:
645 www.iucnredlist.org.
646 Jepsen BI, Carlsen F Genetic identification of West African Chimpanzee, Pan
647 troglodytes verus, based on mitochondrial DNA analysis. *unpublished.*
648 Kircher M, Sawyer S, Meyer M (2012). Double indexing overcomes inaccuracies in
649 multiplex sequencing on the Illumina platform. *Nucleic Acids Res* 40.
650 Lee S, Epstein MP, Duncan R, Lin X (2012). Sparse principal component analysis for
651 identifying ancestry-informative markers in genome-wide association studies.
652 *Genet Epidemiol* 36: 293–302.
653 Li H, Durbin R (2009). Fast and accurate short read alignment with Burrows-Wheeler

654 transform. *Bioinformatics* 25: 1754–1760.
655 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, *et al.* (2009). The
656 Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078–2079.
657 de Manuel M, Kuhlwilm M, Frandsen P, Sousa VC, Desai T, Prado-Martinez J, *et al.*
658 (2016). Chimpanzee genomic diversity reveals ancient admixture with bonobos.
659 *Science* 354: 477–481.
660 Norman AJ, Putnam AS, Ivy JA (2019). Use of molecular data in zoo and aquarium
661 collection management: Benefits, challenges, and best practices. *Zoo Biol* 38:
662 106–118.
663 Nye J, Laayouni H, Kuhlwilm M, Mondal M, Marques-Bonet T, Bertranpetit J (2018).
664 Selection in the Introgressed Regions of the Chimpanzee Genome. *Genome Biol*
665 *Evol* 10: 1132–1138.
666 Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, *et al.*
667 (2013). Great ape genetic diversity and population history. *Nature* 499: 471–5.
668 Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006).
669 Principal components analysis corrects for stratification in genome-wide
670 association studies. *Nat Genet* 38: 904–909.
671 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, *et al.* (2007).
672 PLINK: A tool set for whole-genome association and population-based linkage
673 analyses. *Am J Hum Genet* 81: 559–575.
674 QGIS (2018). QGIS Geographic Information System.
675 Quinlan AR, Hall IM (2010). BEDTools: a flexible suite of utilities for comparing
676 genomic features. *Bioinformatics* 26: 841–2.
677 Rañola JM, Novembre J, Lange K (2014). Fast spatial ancestry via flexible allele

678 frequency surfaces. *Bioinformatics* 30: 2915–22.

679 Rohland N, Reich D (2012). Cost-effective, high-throughput DNA sequencing libraries
680 for multiplexed target capture. *Genome Res* 22: 939–946.

681 Stiles D, Redmond I, Cress D, Nellemann C, Formo RK (2013). *Stolen Apes - The Illicit*
682 *Trade in Chimpanzees, Gorillas, Bonobos and Orangutans*. Birkeland Trykkeri AS,
683 Norway.

684 Traylor-Holzer K, Leus K, Bauman K (2019). Integrated Collection Assessment and
685 Planning (ICAP) workshop: Helping zoos move toward the One Plan Approach.
686 *Zoo Biol* 38: 95–105.

687 Xu H, Luo X, Qian J, Pang X, Song J, Qian G, *et al.* (2012). FastUniq: A Fast De Novo
688 Duplicates Removal Tool for Paired Short Reads. *PLoS One* 7.

689 Zhang J, Kobert K, Flouri T, Stamatakis A (2014). PEAR: a fast and accurate Illumina
690 Paired-End reAd mergeR. *Bioinformatics* 30: 614–620.

691

692 Figure 1

693 Subspecies ancestry in wild and captive populations of chimpanzees. A) Geographical
694 distribution ranges of the four chimpanzee subspecies (IUCN, 2015; QGIS, 2018). B)
695 Population structure by principal component decomposition of sanctuary and the EAZA Ex situ
696 Programme (EEP) populations with reference to wild born individuals. C) Shared ancestry
697 inferences of sanctuary and EEP individuals summarised from individual ADMIXTURE
698 analysis against the reference panel of wild born individuals. Individuals from the reference
699 panel are labelled with a subspecies ancestry prefix and known sample name in previous
700 literature (Prado-Martinez *et al.*, 2013; de Manuel *et al.*, 2016), sanctuary individuals are
701 labelled with common sample name identifiers, and individuals from the EEP are labelled by
702 studbook number (Table S2, Table S3). D) Inbreeding coefficients for all individuals with
703 admixture proportions >0.99 in either of the four inferred clusters. Clusters are colour labelled
704 in accordance to A, B, and C.

705

706 Figure 2

707 Hybrid ancestry in A) the reference panel, B) the EEP population, and C) the sanctuary
708 population. The estimated posterior ancestries, θ is shown for the eight ancestors $k = 3$
709 generations back in time, for each individual in the three populations. The ancestors are ordered
710 according to the "unphased" pedigree in the bottom of the plot. The width of each rectangle
711 indicate the expected proportion of loci that are assigned to that ancestor (conditioned on the
712 estimate of θ). Small widths suggest deviations from the model and features that could be
713 improved by posterior correction.

714

715 Figure 3

716 Geographical origin estimates for sanctuary individuals. Based on the allele frequency surface
717 map of the reference panel, sanctuary individuals are assigned probabilities of geographical
718 origin, here summarized from individual estimates.

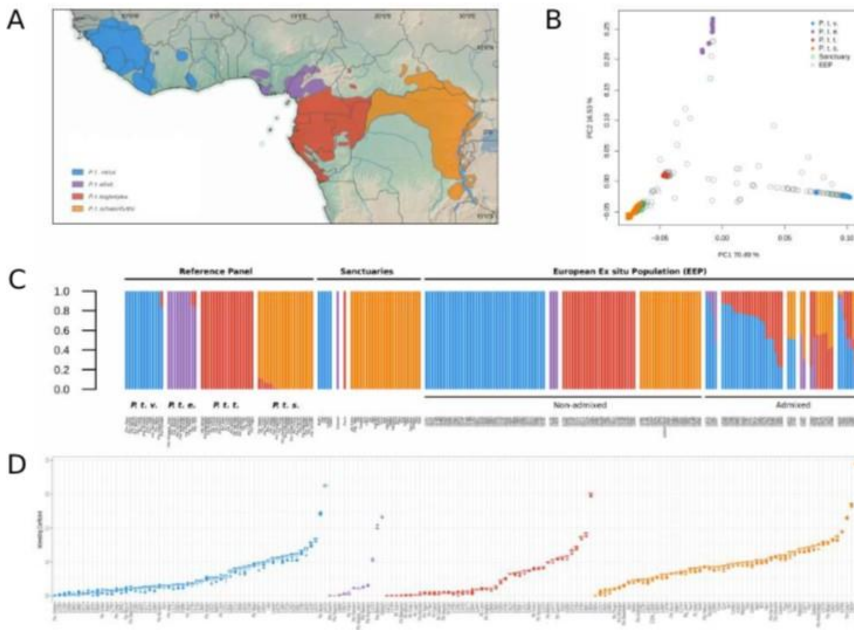
719

720 Figure 4

721 Ancestry and geographical origin estimates from non-invasive samples. A) Geographical origin
722 estimates from hair samples based on the allele frequency surface map of the reference panel,
723 tested individuals are assigned probabilities of geographical origin, here summarized from
724 individual estimates with comparison to blood samples (Figure S13-17). B) Shared ancestry
725 estimates for hair samples compared to whole genome reference data and capture sequenced
726 data from blood.

727

728 Figure 1

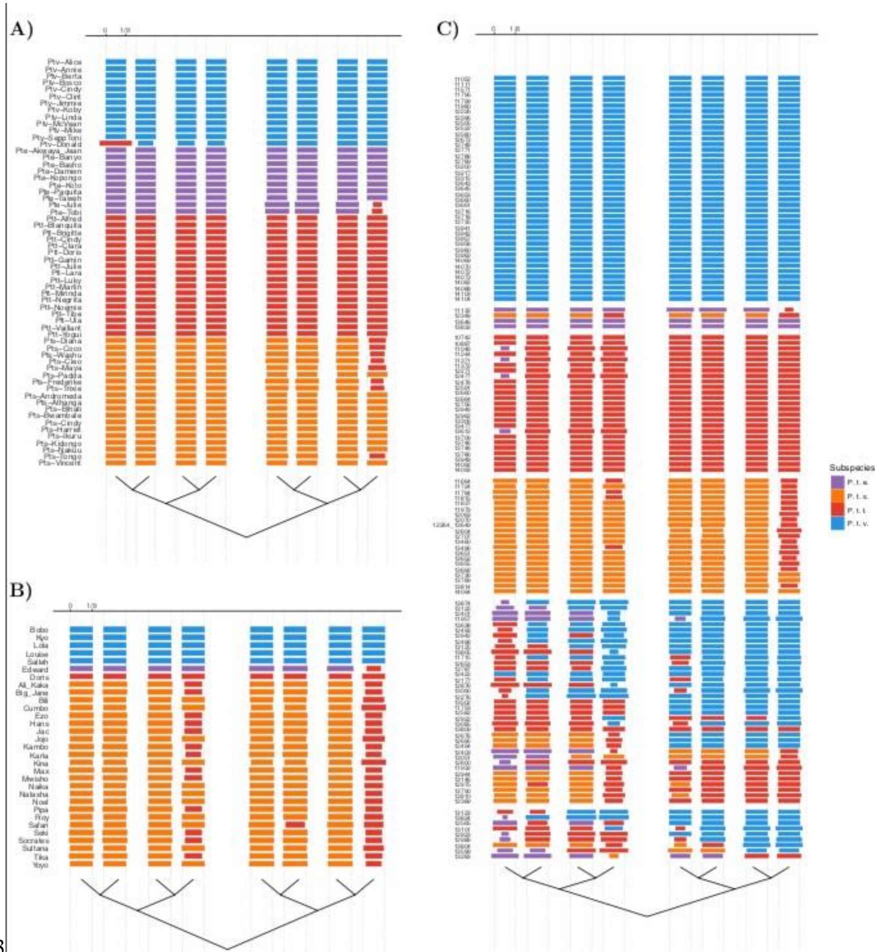


729

730

731

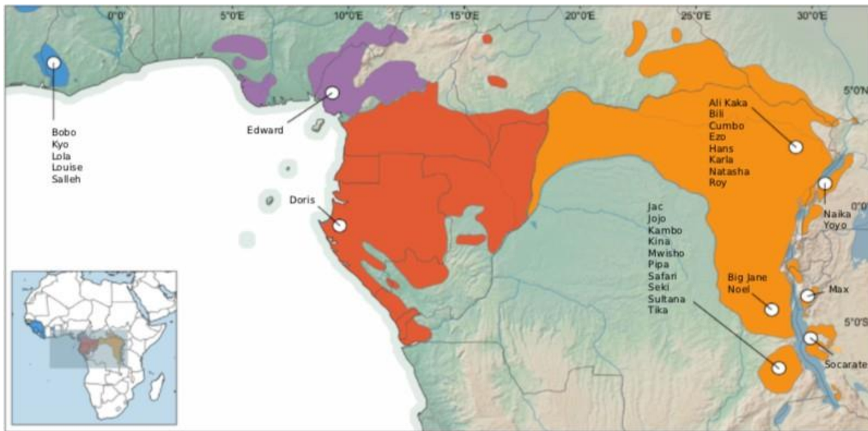
732 Figure 2



733

734

735 Figure 3

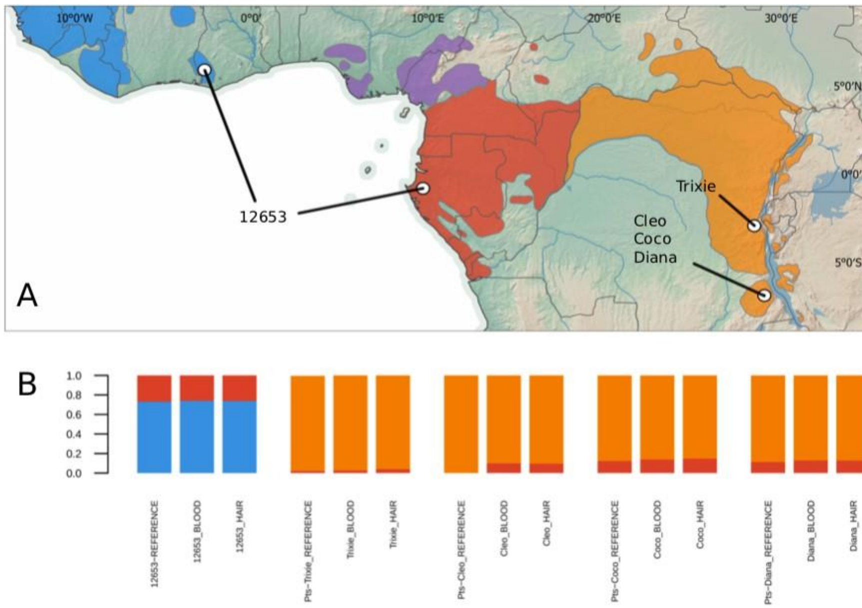


736

737

738

739 Figure 4



740