

**CHIIMP: An automated high-throughput microsatellite genotyping approach reveals greater allelic diversity in wild chimpanzees**

Journal:	<i>Ecology and Evolution</i>
Manuscript ID	ECE-2018-05-00542.R1
Wiley - Manuscript type:	Original Research
Date Submitted by the Author:	n/a
Complete List of Authors:	<p>Barbian , Hannah ; University of Pennsylvania Perelman School of Medicine, Departments of Microbiology and Medicine  Connell, Andrew; University of Pennsylvania Perelman School of Medicine, Departments of Microbiology and Medicine  Avitto , Alexa ; University of Pennsylvania Perelman School of Medicine, Departments of Microbiology and Medicine  Russell , Ronnie ; University of Pennsylvania Perelman School of Medicine, Departments of Microbiology and Medicine  Smith , Andrew ; University of Pennsylvania Perelman School of Medicine, Departments of Microbiology and Medicine  Gundlapally, Madhurima; University of Pennsylvania Perelman School of Medicine, Departments of Microbiology and Medicine  Shazad, Alexander ; University of Pennsylvania Perelman School of Medicine, Departments of Microbiology and Medicine  Li , Yingying; University of Pennsylvania Perelman School of Medicine, Departments of Microbiology and Medicine  Bibollet-Ruche , Frederic; University of Pennsylvania Perelman School of Medicine, Departments of Microbiology and Medicine  Wroblewski, Emily; Washington University in St. Louis, Department of Anthropology  Mjungu, Deus; Gombe Stream Research Center  Lonsdorf , Elizabeth ; Franklin and Marshall College, Department of Psychology  Stewart , Fiona ; Liverpool John Moores University, School of Natural Sciences and Psychology  Piel , Alexander; Liverpool John Moores University, School of Natural Sciences and Psychology  Pusey, Anne; Duke University , Department of Evolutionary Anthropology  Sharp, Paul; University of Edinburgh, Institute of Evolutionary Biology and Centre for Immunity  Hahn, Beatrice; University of Pennsylvania Perelman School of Medicine, Departments of Microbiology and Medicine</p>
Category:	Genetics
Habitat:	Terrestrial

Organism:	Vertebrate
Approach:	Method Development
Abstract:	<p>Short tandem repeats (STRs), also known as microsatellites, are commonly used to non-invasively genotype wild-living endangered species, including African apes. Until recently, capillary electrophoresis has been the method of choice to determine the length of polymorphic STR loci. However, this technique is labor intensive, difficult to compare across platforms, and notoriously imprecise. Here we developed a MiSeq-based approach and tested its performance using previously genotyped fecal samples from long-term studied chimpanzees in Gombe National Park, Tanzania. Using data from eight microsatellite loci as a reference, we designed a bioinformatics platform that converts raw MiSeq reads into locus-specific files and automatically calls alleles after filtering stutter sequences and other PCR artifacts. Applying this method to the entire Gombe population, we confirmed previously reported genotypes, but also identified 31 new alleles that had been missed due to sequence differences and size homoplasy. The new genotypes, which increased the allelic diversity and heterozygosity in Gombe by 61% and 8%, respectively, were validated by replicate amplification and pedigree analyses. This demonstrated inheritance and resolved one case of an ambiguous paternity. Using both singleplex and multiplex locus amplification, we also genotyped fecal samples from chimpanzees in the Greater Mahale Ecosystem in Tanzania, demonstrating the utility of the MiSeq-based approach for genotyping non-habituated populations and performing comparative analyses across field sites. The new automated high-throughput analysis platform (available at <a href="https://github.com/ShawHahnLab/chiimp">https://github.com/ShawHahnLab/chiimp</a>) will allow biologists to more accurately and effectively determine wildlife population size and structure, and thus obtain information critical for conservation efforts.</p>

1  
2  
3 **1 CHIIMP: An automated high-throughput microsatellite genotyping**  
4  
5 **2 platform reveals greater allelic diversity in wild chimpanzees**  
6  
7 **3**  
8

9  
10 4 Hannah J. Barbian<sup>1</sup>, A. Jesse Connell<sup>1</sup>, Alexa N. Avitto<sup>1</sup>, Ronnie M. Russell<sup>1</sup>,  
11 5 Andrew G. Smith<sup>1</sup>, Madhurima S. Gundlapally<sup>1</sup>, Alexander L. Shazad<sup>1</sup>, Yingying Li<sup>1</sup>,  
12 6 Frederic Bibollet-Ruche<sup>1</sup>, Emily E. Wroblewski<sup>2</sup>, Deus Mjungu<sup>3</sup>, Elizabeth V. Lonsdorf<sup>4</sup>,  
13 7 Fiona A. Stewart<sup>5</sup>, Alexander K. Piel<sup>5</sup>, Anne E. Pusey<sup>6</sup>,  
14 8 Paul M. Sharp<sup>7</sup> and Beatrice H. Hahn<sup>1\*</sup>  
15  
16  
17  
18  
19  
20  
21

22 10 <sup>1</sup>Departments of Microbiology and Medicine, Perelman School of Medicine, University of  
23 11 Pennsylvania, Philadelphia, PA, USA

24 12 <sup>2</sup>Department of Anthropology, Washington University in St. Louis, St. Louis, MO, USA

25 13 <sup>3</sup>Gombe Stream Research Center, Kigoma, Tanzania

26 14 <sup>4</sup>Department of Psychology, Franklin and Marshall College, Lancaster, Pennsylvania

27 15 <sup>5</sup>School of Natural Sciences and Psychology, Liverpool John Moores University,  
28 16 Liverpool, United Kingdom

29 17 <sup>6</sup>Department of Evolutionary Anthropology, Duke University, Durham, North Carolina

30 18 <sup>7</sup>Institute of Evolutionary Biology and Centre for Immunity, Infection and Evolution,  
31 19 University of Edinburgh, Edinburgh EH9 3FL, United Kingdom

32 20  
33 21 Professor, Departments of Medicine and Microbiology  
34 22 University of Pennsylvania Perelman School of Medicine  
35 23 409 Johnson Pavilion  
36 24 3610 Hamilton Walk  
37 25 Philadelphia, PA 19104-6076  
38 26 USA  
39 27 bhahn@penmedicine.upenn.edu  
40  
41  
42  
43

44 28  
45 29 Running title: High throughput STR genotyping of chimpanzee  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

**Abstract**

Short tandem repeats (STRs), also known as microsatellites, are commonly used to non-invasively genotype wild-living endangered species, including African apes. Until recently, capillary electrophoresis has been the method of choice to determine the length of polymorphic STR loci. However, this technique is labor intensive, difficult to compare across platforms, and notoriously imprecise. Here we developed a MiSeq-based approach and tested its performance using previously genotyped fecal samples from long-term studied chimpanzees in Gombe National Park, Tanzania. Using data from eight microsatellite loci as a reference, we designed a bioinformatics platform that converts raw MiSeq reads into locus-specific files and automatically calls alleles after filtering stutter sequences and other PCR artifacts. Applying this method to the entire Gombe population, we confirmed previously reported genotypes, but also identified 31 new alleles that had been missed due to sequence differences and size homoplasy. The new genotypes, which increased the allelic diversity and heterozygosity in Gombe by 61% and 8%, respectively, were validated by replicate amplification and pedigree analyses. This demonstrated inheritance and resolved one case of an ambiguous paternity. Using both singleplex and multiplex locus amplification, we also genotyped fecal samples from chimpanzees in the Greater Mahale Ecosystem in Tanzania, demonstrating the utility of the MiSeq-based approach for genotyping non-habituated populations and performing comparative analyses across field sites. The new automated high-throughput analysis platform (available at <https://github.com/ShawHahnLab/chiimp>) will allow biologists to more accurately and effectively determine wildlife population size and structure, and thus obtain information critical for conservation efforts.

**Keywords:** high-throughput STR genotyping, length homoplasy, parentage analysis, short tandem repeats (STRs), *Pan troglodytes*

## 56 Introduction

57  
58 Microsatellites comprise short tandem repeats (STRs) of one to six base pairs, which are  
59 commonly used to profile DNA for a variety of applications ranging from cancer diagnosis  
60 to forensics (Bennett 2000; Ellegren 2004; Guichoux *et al.* 2011; Lynch & de la Chapelle  
61 2003). STR loci have a high mutation rate and vary in the number of their repeat motifs,  
62 due to slippage of the polymerase during DNA synthesis (Kelkar *et al.* 2010; Levinson &  
63 Gutman 1987). Because of their ubiquity, high allelic diversity and co-dominant  
64 inheritance, microsatellites are commonly used for individual identification, parentage  
65 analyses and population genetics (Balloux & Lugon-Moulin 2002; Jarne & Lagoda 1996;  
66 Queller *et al.* 1993; Selkoe & Toonen 2006). STR analysis can also be performed on  
67 samples containing little host DNA, such as hair and fecal samples, and has thus been the  
68 method of choice to genotype endangered primate species, which are typically sampled  
69 non-invasively (Constable *et al.* 1995; Constable *et al.* 2001; Morin *et al.* 1993; Taberlet *et*  
70 *al.* 1997). An accurate determination of the number, distribution, and population  
71 connectivity of wild primates is essential for designing effective conservation measures to  
72 protect these species under increasing anthropogenic threat from habitat loss, disease  
73 and hunting (Arandjelovic & Vigilant 2018). However, census and population genetics  
74 studies of wild apes have been impeded by difficulties of accurately and cost effectively  
75 genotype large numbers of non-invasively collected samples.

76       Until recently, the length of polymorphic STR loci has been determined by capillary  
77 electrophoresis, which compares the mobility of fluorescently labeled PCR products to a  
78 size standard of control fragments and thus yields only approximate results (e.g., a locus  
79 size of “167.5 bp”). Manual correction of such ambiguities can lead to arbitrary allele  
80 binning and inconsistent calls between experiments and/or investigators (Ewen *et al.*  
81 2000; Weeks *et al.* 2002). In addition, amplification of STR loci frequently generates PCR

1  
2  
3 82 artifacts, which are difficult to identify on electropherograms. These include stutter peaks,  
4  
5 83 which are usually one repeat shorter than the correct STR allele and derive from Taq  
6  
7 84 polymerase slippage (Hauge & Litt 1993; Shinde *et al.* 2003), split peaks which are  
8  
9 85 caused by inconsistent A-overhang addition (Schuelke 2000), and artifactual peaks, which  
10  
11 86 are the product of off-target amplification and/or unspecific fluorescent signaling (Ewen *et*  
12  
13 87 *al.* 2000; Fernando *et al.* 2001; Guichoux *et al.* 2011). Existing peak-calling software often  
14  
15 88 fails to differentiate erroneous from real peaks and frequently omits peaks of low height.  
16  
17 89 Automatically called peaks must therefore be corrected manually, which is labor intensive  
18  
19 90 and time consuming (Guichoux *et al.* 2011). Finally, multiplexing is restricted to only a few  
20  
21 91 fluorescent labels, thus limiting the number of loci that can be analyzed simultaneously.  
22  
23 92 As a consequence, capillary electrophoresis based STR genotyping is laborious,  
24  
25 93 notoriously imprecise, and generally not useful for large sample sets or data sharing  
26  
27 94 between different platforms and/or field sites (Pasqualotto *et al.* 2007).

28  
29 95 To improve the accuracy and throughput of STR genotyping, investigators have  
30  
31 96 begun to use next-generation sequencing technologies to characterize amplified  
32  
33 97 microsatellite loci. This approach is superior to capillary electrophoresis, since it yields  
34  
35 98 unambiguous allele lengths regardless of protocol or sequencing platform. In addition,  
36  
37 99 genotyping-by-sequencing (GBS) distinguishes alleles of the same size that contain  
38  
39 100 substitutions or differ in length by a single nucleotide (Adams *et al.* 2004). Although  
40  
41 101 initially developed for human forensics (Fordyce *et al.* 2011; Van Neste *et al.* 2012), GBS  
42  
43 102 technologies have recently been used to genotype wild animals, including Atlantic cod  
44  
45 103 (Vartia *et al.* 2016), brown bear (De Barba *et al.* 2017), boarfish (Farrell *et al.* 2016), and  
46  
47 104 muskrat (Darby *et al.* 2016). These studies demonstrated the utility of GBS for molecular  
48  
49 105 ecology applications (Darby *et al.* 2016; Farrell *et al.* 2016) and showed that even  
50  
51 106 samples containing small quantities of host DNA, such as dung and hair, can be used for  
52  
53 107 these analyses (De Barba *et al.* 2017). However, alleles were primarily called manually by  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 108 visual inspection of read length histograms (Darby *et al.* 2016; Farrell *et al.* 2016; Vartia *et*  
4  
5 109 *al.* 2016), and none of these studies have compared the performance of capillary  
6  
7 110 electrophoresis and high throughput sequencing side-by-side to validate and improve the  
8  
9 111 genotyping approach.

112 For nearly two decades, our group has been studying chimpanzees in Gombe  
113 National Park (Tanzania) to assess the long-term impact of simian immunodeficiency virus  
114 (SIVcpz) infection on this wild-living population (Keele *et al.* 2009; Rudicell *et al.* 2010;  
115 Santiago *et al.* 2003). To identify SIVcpz infected individuals, we developed non-invasive  
116 diagnostic assays that detect virus-specific antibodies and nucleic acids by analysis of  
117 fecal samples. To reliably monitor the spread of SIVcpz in all three Gombe communities,  
118 we verified the individual origin of each fecal sample by microsatellite analysis at eight  
119 polymorphic STR loci. Thus, most Gombe chimpanzees have been repeatedly genotyped,  
120 resulting in a consensus genotype that has been used for paternity and kinship  
121 determinations, immunogenetics, microbiome analyses and behavioral studies (Barbian *et*  
122 *al.* 2018; Keele *et al.* 2009; Moeller *et al.* 2016; Rudicell *et al.* 2010; Santiago *et al.* 2003;  
123 Walker *et al.* 2017; Wroblewski *et al.* 2015).

124 Here, we used these multiply confirmed reference microsatellites as a guide to  
125 develop and iteratively improve a MiSeq-based STR genotyping approach. To permit the  
126 direct comparison with previous capillary electrophoresis results, we determined the  
127 length of STR loci by sequencing PCR amplicons in their entirety, including both forward  
128 and reverse primers. We also developed a *Computational High-throughput Individual*  
129 *Identification through Microsatellite Profiling* (CHIIMP) pipeline that detects and filters  
130 erroneous alleles and automatically generates a number of downstream analyses, such  
131 as allele length histograms, alignments of allele sequences, contamination heatmaps and  
132 genotype comparisons. By directly comparing the new CHIIMP-derived genotypes to  
133 previously determined capillary electrophoresis results, we show that the new analysis

1  
2  
3 134 tools, which are not included in any of the previously published STR genotyping pipelines,  
4  
5 135 greatly improve the speed, cost and accuracy of allele determinations.  
6

7 136

8  
9 137

## 10 11 138 **Material and methods**

12  
13 139

### 14 140 *Chimpanzee fecal samples*

15  
16 141

17  
18  
19  
20 142 Fecal samples were collected from wild-living chimpanzees in Gombe National Park,  
21  
22 143 including members of the Mitumba, Kasekela and Kalande communities, as well as the  
23  
24 144 Greater Mahale Ecosystem (GME) in Tanzania as previously described (Keele *et al.* 2009;  
25  
26 145 Rudicell *et al.* 2010; Rudicell *et al.* 2011; Santiago *et al.* 2003). Habituated Gombe  
27  
28 146 chimpanzees have been under direct observation since the 1960s (Pusey *et al.* 2007; van  
29  
30 147 Lawick-Goodall 1968), with prospective fecal sampling and SIVcpz diagnostics initiated in  
31  
32 148 1999 (Keele *et al.* 2009; Rudicell *et al.* 2010). Long-term monitoring of non-habituated  
33  
34 149 chimpanzees in the GME began in 2008, with non-invasive SIVcpz screening  
35  
36 150 implemented in 2009 (Rudicell *et al.* 2011). Gombe and GME fecal samples were  
37  
38 151 collected 1:1 (vol/vol) in RNA*later* (Ambion), a high salt solution that preserves nucleic  
39  
40  
41 152 acids and allows storage and transport at room temperature. For individual identification,  
42  
43 153 samples were routinely subjected to mitochondrial, sex, and microsatellite analyses, with  
44  
45 154 up to eight STR loci characterized by capillary electrophoresis as described previously  
46  
47 155 (Keele *et al.* 2009; Rudicell *et al.* 2010; Rudicell *et al.* 2011). All fieldwork has been  
48  
49 156 approved by the Tanzania National Parks, the Tanzania Commission for Science and  
50  
51 157 Technology, the Tanzania Wildlife Research Institute, and has followed the American  
52  
53 158 Society of Primatologists' Principles for Ethical Treatment of Nonhuman Primates.  
54

55  
56 159  
57  
58  
59  
60



1  
2  
3 160 *Quantification of chimpanzee DNA*  
4

5 161

6  
7 162 Fecal DNA was extracted from 0.5 ml of homogenized fecal suspension using the  
8  
9 163 QIAamp DNA Stool Kit and the automated QIAcube system (Qiagen). Purified DNA was  
10  
11 164 eluted in 200 µl water and stored at -20 °C. Chimpanzee genomic DNA content was  
12  
13 165 determined using a previously described *c-myc* gene-based quantitative (q)PCR (Morin *et*  
14  
15 166 *al.* 2001). Briefly, 2 µl DNA extract was added to 1x High Fidelity PCR Buffer, 3.5 mM  
16  
17 167 MgSO<sub>4</sub>, 0.3 µM forward (5'-GCCAGAGGAGGAACGAGCT-3') and reverse (5'-  
18  
19 168 GGGCCTTTTCATTGTTTTCCA-3') qPCR primers, 0.2 µM of a FAM-labeled probe (FAM-  
20  
21 169 TGCCCTGCGTGACCAGATCC-BHQ1), 0.2 mM dNTPs, 1x ROX Reference Dye, and 0.5  
22  
23 170 U Platinum Taq DNA Polymerase High Fidelity (Invitrogen). Each sample was run in  
24  
25 171 triplicate on a 7900HT Fast Real-Time PCR System, together with human genomic DNA  
26  
27 172 standards of known concentration (the sequence of the particular *c-myc* amplicon is  
28  
29 173 identical between humans and chimpanzees). Negative “no-template” controls were  
30  
31 174 included in each run. Sequence Detection Systems version 2.3 software (Applied  
32  
33 175 Biosystems) was used to quantify the host DNA content of each sample. Since host DNA  
34  
35 176 concentrations differed, approximately half of the samples were extracted on more than  
36  
37 177 one occasion to generate enough material for all analyses.  
38  
39  
40

41 178

42  
43 179 *Amplification of STR loci*  
44

45 180

46  
47 181 Previous genotyping studies of Gombe and GME chimpanzees utilized eight STR loci  
48  
49 182 containing tetranucleotide repeats (Constable *et al.* 2001; Keele *et al.* 2009; Rudicell *et al.*  
50  
51 183 2011). These included D18s536 (also termed locus A), D4s243 (locus B), D10s676 (locus  
52  
53 184 C), D9s922 (locus D), D2s1326 (locus 1) D2s1333 (locus 2), D4s1627 (locus 3), and  
54  
55 185 D9s905 (locus 4) (Table S1). To facilitate MiSeq sequencing of the amplified loci, we  
56  
57  
58  
59  
60

1  
2  
3 186 added MiSeq-specific adapters to the 5' end of both the forward (5'-  
4  
5 187 TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG-3') and the reverse primer (5'-  
6  
7 188 GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG-3'), respectively. Individual STR  
8  
9 189 loci were amplified using 3 - 5  $\mu$ l fecal DNA extract, 2.5  $\mu$ l 10x AmpliTaq Gold Buffer, 1.75  
10  
11 190  $\mu$ l 25mM MgCl<sub>2</sub>, 1.5  $\mu$ l 10 mM dNTPs, 0.5  $\mu$ l 50  $\mu$ g/ml BSA, 1.5  $\mu$ l of 10 mM forward and  
12  
13 191 reverse primers, and 0.25  $\mu$ l AmpliTaq Gold polymerase (5U/ml; Applied Biosystems) in a  
14  
15 192 25  $\mu$ l reaction volume. Thermocycling was performed using an initial denaturation for 10  
16  
17 193 minutes at 94 °C, followed by 50 cycles of 30 seconds at 94 °C, 30 seconds at 54 °C, and  
18  
19 194 45 seconds at 72 °C, followed by a final extension of 10 minutes at 72 °C.

20  
21  
22 195 Testing the sensitivity of MiSeq derived allele detection, we found that individual  
23  
24 196 PCR reactions often produced only partial genotypes, while the combination of multiple  
25  
26 197 amplicons from the same DNA sample generally yielded a more complete set of alleles.  
27  
28 198 Consistent with previous studies (Morin *et al.* 2001), we also found that PCR amplification  
29  
30 199 of less than 25 pg of host DNA generally failed to amplify STR loci. For all genotyping  
31  
32 200 analyses, we thus included only DNA samples that contained more than 25 pg of  
33  
34 201 chimpanzee DNA, amplified each STR locus on three independent occasions, and  
35  
36 202 combined equal volumes of these replicate PCR reactions prior to MiSeq sequencing.

37  
38  
39 203 The eight STR loci were also amplified in one-step and two-step multiplex  
40  
41 204 reactions. To minimize primer-primer interactions, locus A, B, C and 3 primers were  
42  
43 205 combined at an even ratio in one pool, while locus D, 1, 2, 4 primers were similarly  
44  
45 206 combined in a second pool. Fecal DNA was then amplified in two (rather than eight)  
46  
47 207 different reactions, using the identical cycling conditions as for singleplex PCR. For two-  
48  
49 208 step multiplexing, 2  $\mu$ l of a 1:100 dilution of the one-step product were used as a template  
50  
51 209 for a second round of PCR in which each locus was amplified individually using the same  
52  
53 210 thermocycling conditions (Arandjelovic *et al.* 2009).

54  
55  
56 211

### 212 *Library preparation and MiSeq sequencing*

213

214 Following STR locus amplification, PCR products (individual or pooled) were diluted in  
215 nuclease-free sterile water (1:10) and subjected to two rounds of PCR to add Illumina  
216 barcodes and enrich for properly indexed DNA products as described (Iyer *et al.* 2017).  
217 The resulting libraries were pooled, purified with Ampure Beads (Beckman Coulter),  
218 quantified using a Qubit Fluorometer (Thermo Scientific) and TapeStation 2200 (Agilent),  
219 and diluted to a final DNA concentration of 4 nM (Iyer *et al.* 2017). A randomly fragmented  
220 (adapter ligated) control library of PhiX DNA (Illumina) was added to increase read length  
221 diversity to ensure cluster recognition on the flow-cell. Both PhiX control and STR  
222 amplicon libraries were adjusted to a final DNA concentration of 12 pM and mixed 1:1  
223 prior to loading onto the sequencing reagent cartridge. All STR amplicons were  
224 sequenced in one direction using v2 chemistry (500 cycle kits) without fragmentation. This  
225 increased the length of the STR loci that could be analyzed to ~400 bp (instead of 2 x 250  
226 paired-end reads). Although 500 cycles are the theoretical maximum of the sequencing  
227 kit, we observed diminishing data quality between 350-400 cycles. We thus selected 375  
228 forward and 51 reverse read cycles, using only the forward reads for analysis to preclude  
229 alignment artifacts of pairing reads in the repeat regions (the reverse reads were only  
230 used for MiSeq quality control). To maximize the number of amplicons sequenced per run,  
231 we used dual index multiplexing of samples.

232

### 233 *CHIIMP analysis pipeline*

234

235 Following MiSeq sequencing, read files were processed using standard methods. First,  
236 sample demultiplexing and FASTQ file generation was performed using the Illumina  
237 MiSeq Reporter software with default settings. Next, MiSeq adapter sequences were

1  
2  
3 238 trimmed using cutadapt (Martin 2011). The adapter trimmed forward reads from each read  
4  
5 239 pair, which covered the entire STR amplicon, were then imported into the R package,  
6  
7 240 which was used for all downstream analyses.  
8

9 241 The CHIIMP analysis pipeline generates multi-locus genotypes in three stages.  
10  
11 242 First, each MiSeq sequence file is processed into unique sequences with relevant  
12  
13 243 attributes, such as read counts, sequence length, and whether the sequence matches the  
14  
15 244 locus-specific forward primer, repeat motifs and length range. Sequences are also queried  
16  
17 245 for potential PCR artifacts, such as single nucleotide substitutions, indels, and stutter  
18  
19 246 sequences introduced by Taq polymerase and sequencing errors. These artifacts are  
20  
21 247 identified as comprising less than one third of the read counts of the corresponding allele.  
22  
23 248 The 33% threshold was selected because inspection of known heterozygous loci revealed  
24  
25 249 that all of the true second most frequent alleles contained more than that proportion of  
26  
27 250 reads. Finally, for each sample and locus the proportion of sequence reads of the total  
28  
29 251 read count is determined. At this stage, data are kept for all loci to ensure flexible  
30  
31 252 downstream processing, such as detecting cross-locus contamination.  
32  
33

34 253 The second stage removes all sequences that do not match the locus attributes,  
35  
36 254 such as the forward primer, repeat motif, and locus length, and/or contain likely PCR  
37  
38 255 artifacts. In addition, only sequences comprising a minimum fraction of the total number of  
39  
40 256 filtered reads (5%) are retained, and only loci with a total filtered read count above a  
41  
42 257 customizable per-sample read threshold (>500) are genotyped. Application of these filters  
43  
44 258 determines the sample zygosity; if only one sequence passes these filters, the locus is  
45  
46 259 reported as homozygous. However, if two or more sequences pass the filters, the two  
47  
48 260 most abundant are kept and the sample is reported as heterozygous. The output at this  
49  
50 261 stage includes a spreadsheet with the sequence content, read counts, sequence lengths,  
51  
52 262 as well as other relevant information such as whether the sequence contains the correct  
53  
54 263 repeat motif or was identified as a likely stutter sequence or other PCR artifact. Of note, all  
55  
56  
57  
58  
59  
60

1  
2  
3 264 filters and thresholds are customizable, with the above parameters representing the  
4  
5 265 default.

6  
7 266 In the final stage, genotypes are assembled for all samples and loci, with quality  
8  
9 267 control tables generated as output files (Fig. S1). First, a summary genotype table is  
10  
11 268 generated that lists sample designations for each row, STR loci for each column, and  
12  
13 269 unique allele identifiers for each cell (Fig. S1a). If specific allele codes are provided, the  
14  
15 270 summary table will include these designations. If an allele does not match previous  
16  
17 271 identifiers, the software will create a short name based on sequence length and content to  
18  
19 272 identify these new alleles (e.g., sample 4781, locus C, allele 2 in Fig. S1a). The similarity  
20  
21 273 of genotypes is also depicted in a heatmap (Fig. 1b), which groups closely related  
22  
23 274 genotypes (Peakall & Smouse 2006). In cases where genotypes of individuals are known,  
24  
25 275 the program links samples with the corresponding individuals (Fig. S1c). A heatmap  
26  
27 276 shows the extent of similarity of every sample with every known genotype, thus allowing  
28  
29 277 simple individual identification (Fig. S1d). The program also generates a set of tables that  
30  
31 278 flag alleles that require additional attention, such as loci where the stutter filter has been  
32  
33 279 invoked, where more than two sequences passed the filter, where a large proportion of  
34  
35 280 sequences was not contained in the identified alleles, and where homozygosity may  
36  
37 281 reflect allelic dropout (Fig. S1e). For each locus, the program creates a FASTA file of all  
38  
39 282 allele sequences and an image of their alignment (Fig. S1f) generated by the  
40  
41 283 Bioconductor's MSA package (Bodenhofer *et al.* 2015). In addition, a heatmap of  
42  
43 284 sequence counts that match the locus-specific forward primer for all samples and loci is  
44  
45 285 generated (Fig. S1g). For singleplex samples, this identifies sequences that match other  
46  
47 286 loci and thus highlights potential cross-locus contamination. For multiplexed samples, this  
48  
49 287 shows the read distribution across different loci. Finally, histograms that show sequence  
50  
51 288 length-frequency distributions are saved as image files (Fig. S1h). A summary file is  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 289 created that combines all key results (sequences, read counts, etc.) for alleles for all  
4  
5 290 samples and loci. This data output file is suitable for further analysis in R.  
6

7 291 The new analysis platform, termed *Computational High-throughput Individual*  
8  
9 292 *Identification through Microsatellite Profiling* or CHIIMP, has been designed to allow  
10  
11 293 customization of the number and sequence content of microsatellite loci to be analyzed.  
12  
13 294 Particular locus attributes such as the expected locus length range, primer sequences,  
14  
15 295 and repeat motif sequence can be specified in a simple text file. Thus, the software can be  
16  
17 296 readily adapted to additional microsatellite loci, as long as the respective amplicons fall  
18  
19 297 within the length limits of the particular sequence chemistry used. The software is also  
20  
21 298 suitable to analyze multiplexed samples, which contain reads from several loci but are  
22  
23 299 processed separately, again using the forward primers to select locus-specific reads. No  
24  
25 300 additional software is required other than providing a list of samples and loci prior to  
26  
27 301 analysis. CHIIMP is available at <https://github.com/ShawHahnLab/chiimp> and can be  
28  
29 302 installed on any Windows, Mac OS, or Linux computer with a standard installation of R  
30  
31 303 and RStudio in a single step. On Windows, a desktop shortcut to the analysis script is  
32  
33 304 provided. Dragging a simple text file containing analysis options onto the shortcut triggers  
34  
35 305 analysis with the selected options. In addition to the standalone program, all features can  
36  
37 306 also be used individually from within R. A comprehensive user guide including examples  
38  
39 307 of analysis options and locus attributes is provided with the software.  
40  
41  
42

43 308

45 309 *Error, diversity, and heterozygosity calculations*

47 310

49 311 Error rates for the MiSeq derived genotypes were calculated by determining the number of  
50  
51 312 allelic mismatches for each sample to the known genotype of the corresponding  
52  
53 313 chimpanzee (including allelic dropout, stutter sequences, PCR/sequencing artifacts, and  
54  
55 314 locus amplification failure) and by dividing the total number of alleles by the number of  
56  
57  
58  
59  
60

1  
2  
3 315 erroneous alleles (Broquet and Petit, 2004). The expected heterozygosity (also termed  
4  
5 316 gene diversity) for the sampled Gombe and GME chimpanzees was calculated from both  
6  
7 317 capillary electrophoresis and MiSeq based microsatellite data as described in  
8  
9 318 Charlesworth & Charlesworth 2010. Allelic diversity was calculated by summing the total  
10  
11 319 number of unique alleles in a population.  
12

13 320

## 14 321 **Results**

15 322

### 16 323 *Direct comparison of MiSeq and capillary electrophoresis based STR genotyping*

17 324

18 325 To compare the performance of MiSeq and capillary electrophoresis side-by-side, we  
19 326 selected samples from 19 Gombe chimpanzees, who were previously genotyped by  
20 327 capillary electrophoresis on multiple occasions (Keele *et al.* 2009; Rudicell *et al.* 2010;  
21 328 Santiago *et al.* 2003). Testing more recently collected fecal samples that had not yet been  
22 329 genotyped, we used the consensus of previous genotypes at eight STR loci as the  
23 330 benchmark to which all MiSeq derived data were compared (Table 1). Fecal DNA was  
24 331 extracted, confirmed to contain more than 25 pg of chimpanzee DNA per PCR aliquot,  
25 332 and amplified using the same STR primers and conditions, except for the presence of  
26 333 MiSeq adapters versus fluorescent labels. For MiSeq sequencing, three PCR replicates  
27 334 were pooled, while only a single replicate was analyzed by capillary electrophoresis using  
28 335 both automated and manual peak calling options. The latter was done because capillary  
29 336 electrophoresis analysis of pooled samples is compromised when allele peaks differ in  
30 337 relative height in independent PCR reactions.

31 338 Using the consensus genotype of the corresponding chimpanzees for reference  
32 339 (Table 1), we found that MiSeq genotyping reduced the number of allelic dropouts by  
33 340 more than half (Table 2). This was due, at least in part, to the pooling of PCR replicates,  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50

1  
2  
3 341 which increased the number of alleles that were detected. However, MiSeq genotyping  
4  
5 342 was also more accurate than the traditional method, which could not differentiate off-target  
6  
7 343 amplifications (Tables 1 and 2). In addition, stutter peaks were completely eliminated by  
8  
9 344 the CHIIMP analysis pipeline, which was not the case for the automated capillary  
10  
11 345 electrophoresis method. Although manual peak calling also eliminated stutter peaks, this  
12  
13 346 was considerably more time consuming than the MiSeq approach. For the 19 samples,  
14  
15 347 conventional peak calling and allele binning took two hours, while reviewing the  
16  
17 348 bioinformatics outputs took minutes. Most importantly, MiSeq genotyping identified eight  
18  
19 349 heterozygous loci that were scored as homozygous by capillary electrophoresis because  
20  
21 350 of a failure to resolve minor sequence and length (1bp) differences (Fig. 1). These  
22  
23 351 sequence variants were readily identified in the read histograms (Fig. 1a) and their  
24  
25 352 frequency identified in sequence alignments of the entire locus (Fig. 1b and c). Inspection  
26  
27 353 of allele lengths across all loci revealed that 24% of all MiSeq derived alleles did not differ  
28  
29 354 by multiples of four, indicating frequent nucleotide insertions and deletions in the  
30  
31 355 tetranucleotide repeats (Fig. 1b and c).  
32  
33  
34  
35  
36

356

### 357 *MiSeq genotyping uncovers increased allelic diversity and heterozygosity*

358

359 To examine the true extent of allelic diversity in Gombe, we selected fecal samples from  
360 123 chimpanzees, which included all currently living adults and juveniles, except for  
361 offspring born within the past three years, as well as 38 deceased individuals. All of these  
362 were previously genotyped by capillary electrophoresis on at least three occasions.  
363 Subjecting one representative fecal sample to MiSeq analysis, we confirmed 51 known  
364 alleles, but also detected 31 new alleles, which had previously gone unrecognized due to  
365 size (1 bp) or nucleotide sequence differences (Tables 3 and S2). Such cryptic alleles  
366 were detected for all eight STR loci, increasing allelic diversity by an average of 1.6 fold



1  
2  
3 367 per locus. Nearly half of all previously reported alleles had closely related length or  
4  
5 368 sequence variants (Table 4).  
6

7 369 Although the great majority of the newly identified alleles were found in multiple  
8  
9 370 individuals, we wanted to validate their authenticity by demonstrating inheritance. Since  
10  
11 371 paternity and kinship relationships are known for most Gombe chimpanzees, we were  
12  
13 372 able to trace the majority of the newly identified allelic variants from parents to their  
14  
15 373 offspring. For example, Locus 3 includes four alleles that are identical in size (234 bp) but  
16  
17 374 differ by up to three substitutions and two single nucleotide insertions and deletions (Fig.  
18  
19 375 2a). Alleles 234-a, 234-b, 234-c, and 234-d were found in 80, 25, 10 and 4 chimpanzees,  
20  
21 376 respectively, including several parent-offspring triads (Fig. 2b). Overall, we were able to  
22  
23 377 document inheritance for 25 (81%) of the 31 new alleles. For the remaining 6 existing  
24  
25 378 pedigree information was insufficient, and their existence was thus confirmed by  
26  
27 379 sequencing at least two independent PCR amplicons (Table 4).  
28  
29

30 380 The newly identified alleles revealed that over a quarter of genotypes at loci previously  
31  
32 381 assigned as homozygous (60 of a total of 228) were in fact heterozygous (Table S2). This  
33  
34 382 increased allelic diversity resolved one case of an ambiguous paternity determination.  
35  
36 383 Using the standard eight STR loci, we were previously unable to identify the father of one  
37  
38 384 infant (Google) because two candidate males (Faustino and Londo) had the identical  
39  
40 385 genotype at all eight STR loci (Walker *et al.* 2017). Using the new genotypes, we were  
41  
42 386 able to exclude Londo and confirm Faustino as a father by revealing differences at one  
43  
44 387 locus (Fig. 2c). Although Faustino was identified as the correct father at the time by  
45  
46 388 genotyping 10 additional loci using capillary electrophoresis (Walker *et al.* 2017), this  
47  
48 389 would not have been necessary had the increased allelic diversity been known. Thus,  
49  
50 390 MiSeq genotyping revealed much greater allelic and microsatellite gene diversity in  
51  
52 391 Gombe than previously appreciated, thus increasing the analytical potential of the existing  
53  
54 392 STR loci.  
55  
56  
57  
58  
59  
60

393

394 *MiSeq genotyping based individual identification*

395

396 Since chimpanzee communities are often studied longitudinally, we added an individual  
397 identification tool to the analysis platform. This tool compares the genotype of every new  
398 sample with all previously characterized genotypes and generates a distance score to  
399 indicate their relative similarity. For example, samples with a distance score of 0 match at  
400 all loci, while samples with a distance score of 2 differ by two alleles. We then used this  
401 approach to characterize the same 19 newly genotyped samples (Table 1) as well as 5  
402 samples from infants with unknown genotypes. To account for allelic dropout, a distance  
403 score of up to 3 was allowed. The results revealed accurate individual identification for all  
404 samples from previously genotyped chimpanzees. Of the 19 samples, 8 exhibited a  
405 perfect match across all loci (Fig. 3a), while 11 others had distance scores of 1-3, which  
406 were consistent with allelic dropout (Fig. 3b). However, 5 samples with distance scores of  
407 5-7 could not be assigned to known individuals (Fig. 3c), and a review of field notes  
408 revealed that they were all collected from new infants. A heatmap allowed the quick  
409 identification of very close (4821, 4807) and very distant (4566) matches (Fig. 3d). Thus,  
410 the individual identification tool detected previously determined genotypes with reasonable  
411 accuracy.

412

413 *STR genotyping of multiplexed samples*

414

415 Chimpanzees in the Greater Mahale Ecosystem in Tanzania occupy a large home range,  
416 live at low population densities, and face extreme seasonal changes (Moore 1996; Ogawa  
417 *et al.* 1999; Schoeninger *et al.* 1999). Thus, these “savanna chimpanzees” live under  
418 ecologically more challenging conditions than their forest-dwelling counterparts, and with

1  
2  
3 419 the exception of the Issa community, are not habituated. As a result, fecal collections,  
4  
5 420 sample transport and storage are logistically more difficult, which can result in reduced  
6  
7 421 amounts of collected material and/or partially degraded host DNA. To test the suitability of  
8  
9 422 MiSeq genotyping for such samples, we selected 12 previously characterized chimpanzee  
10  
11 423 fecal specimens from the Issa Valley (Rudicell *et al.* 2011) and re-genotyped them using  
12  
13 424 both singleplex and multiplex locus amplification. Singleplex PCR was performed as in  
14  
15 425 Gombe, while multiplex PCR was carried out in two steps as previously described  
16  
17 426 (Arandjelovic *et al.* 2009). First, PCR primers for 4 loci were pooled and used to amplify  
18  
19 427 fecal DNA in two (rather than eight) reactions (one-step multiplex product). Second,  
20  
21 428 aliquots of this first round PCR were then used in a second round of PCR to amplify each  
22  
23 429 of the 8 STR loci separately (two-step multiplex product). Three pooled replicates of both  
24  
25 430 one-step and two-step multiplexed products were sequenced and compared to the  
26  
27 431 previously determined genotypes (Table S3). Although the overall amplification efficiency  
28  
29 432 was lower than originally reported (most likely due to repeat freezing and thawing of the 7-  
30  
31 433 8 year old samples), one-step multiplexing performed as well as singleplex PCR, but used  
32  
33 434 only a quarter of the fecal DNA (Table 5). Two-step multiplexing detected slightly more  
34  
35 435 alleles, but not surprisingly, also resulted in an increased number of stutter sequences  
36  
37 436 and other PCR artifacts. Thus, one-step multiplexing required less starting material and  
38  
39 437 was also more cost efficient because the combined loci were sequenced in a single MiSeq  
40  
41 438 run (and were subsequently de-multiplexed bioinformatically).

42  
43  
44  
45 439 MiSeq genotyping also allowed us to compare the allelic diversity in Gombe and  
46  
47 440 the GME. Fig. 4 depicts such an analysis for locus B and D, highlighting alleles that were  
48  
49 441 only found in GME chimpanzees. Comparing all eight STR loci, we found ten alleles in  
50  
51 442 only 12 GME chimpanzees that were absent from the 123 genotyped Gombe individuals,  
52  
53 443 six of which represented alleles previously missed in the GME due to sequence and  
54  
55 444 length differences. Although the mean expected heterozygosity value for the GME  
56  
57  
58  
59  
60

1  
2  
3 445 chimpanzees (0.743) was lower than that for Gombe (0.812), this is likely due to the small  
4  
5 446 sample size and the fact that all 12 individuals were sampled at a single location in Issa  
6  
7 447 Valley (Rudicell *et al.* 2011). Additional samples from more diverse locations in the GME  
8  
9 448 are needed to compare the genetic diversity of this population to that of Gombe and other  
10  
11 449 field sites.

12  
13 450

## 14 451 **Discussion**

15  
16 452

17  
18 453 Over the past two decades, microsatellite analyses have been an integral part of studies  
19  
20 454 of wild chimpanzees, providing insight into their evolution, population genetics, behavior,  
21  
22 455 disease association and social structure (Barbian *et al.* 2018; Becquet *et al.* 2007; Keele  
23  
24 456 *et al.* 2009; Langergraber *et al.* 2007; Moeller *et al.* 2016; Rudicell *et al.* 2010; Santiago *et*  
25  
26 457 *al.* 2003; Vigilant *et al.* 2001; Walker *et al.* 2017; Wroblewski *et al.* 2015). However,  
27  
28 458 traditional genotyping methods are cumbersome, imprecise and investigator/platform  
29  
30 459 dependent, due to the use of capillary electrophoresis to determine the length of STR loci.  
31  
32 460 Here, we report a high-throughput MiSeq-based approach, which represents a marked  
33  
34 461 improvement, because it is faster, more accurate and able to detect the full extent of  
35  
36 462 allelic diversity in a population. Moreover, it includes a new analysis platform, CHIIMP,  
37  
38 463 which not only automates the conversion of raw MiSeq data into multi-locus genotypes,  
39  
40 464 but also implements a number of quality control measures that improve genotyping  
41  
42 465 accuracy (Fig. 5). Of note, CHIIMP has been designed for maximal customization. While  
43  
44 466 analysis of pedigreed fecal samples from chimpanzees allowed rigorous validation, the  
45  
46 467 pipeline is not limited to a particular species or sample type.

47  
48 468

49  
50 469 *Improved accuracy of MiSeq based genotyping*

51  
52 470

1  
2  
3 471 Sequence-based genotyping methods not only determine the length of STR loci, but also  
4  
5 472 reveal their sequence content, and thus have the potential to detect a greater number of  
6  
7 473 distinct alleles than capillary electrophoresis. Indeed, such genotyping of Atlantic cod and  
8  
9 474 muskrats revealed high proportions of cryptic alleles, ranging from 32% to 44% (Darby *et*  
10  
11 475 *al.* 2016; Vartia *et al.* 2016). In light of these data, our discovery of 38% new alleles (31 of  
12  
13 476 82) in Gombe is not surprising (Table 3). However, this finding suggests that existing STR  
14  
15 477 data vastly underestimate the diversity of microsatellite sequences in wild chimpanzees,  
16  
17 478 not only in Gombe but also in other populations. New alleles were found for all loci, with  
18  
19 479 some comprising twice as many variants as previously observed (Table 3), which will  
20  
21 480 undoubtedly add to the statistical power of future analyses. However, any new allele will  
22  
23 481 have to be examined carefully by repeat amplification and sequencing, unless it can be  
24  
25 482 validated by pedigree analysis. In our dataset, a minor fraction of “new” alleles were found  
26  
27 483 to represent PCR and/or sequencing artifacts that exceeded the 33% threshold for  
28  
29 484 heterozygous alleles. Repeat amplification of these alleles resolved all sequencing  
30  
31 485 artifacts.  
32  
33

34  
35 486 Comparison of the MiSeq data to validated reference genotypes also allowed us to  
36  
37 487 assess the error rate of the new approach. After implementation of all filters, CHIIMP  
38  
39 488 eliminated 98% of stutter sequences and 100% of off-target amplicons. Among the  
40  
41 489 samples tested, true alleles, allelic dropouts and false alleles were detected with a  
42  
43 490 frequency of 96%, 7%, and 0%, respectively. These data are comparable to MiSeq  
44  
45 491 derived genotyping results for wild-living brown bears, where true alleles, allelic dropouts  
46  
47 492 and false alleles were detected with a frequency of 93%, 0.4% and 0.05% for tissues, and  
48  
49 493 80%, 14% and 1% for fecal samples, respectively (De Barba *et al.* 2017). Although our  
50  
51 494 overall error rate of 3.3% is slightly higher than the 2.1% error rate reported for a MiSeq  
52  
53 495 genotyping study of laboratory raised (pedigreed) fish (Zhan *et al.* 2017), this is not  
54  
55 496 surprising since the latter study examined freshly extracted tissue DNA.  
56  
57  
58  
59  
60

1  
2  
3 497 Since non-invasively collected samples frequently contain diluted and/or degraded  
4  
5 498 host DNA, they are genotyped using multiple PCR reactions to guard against the selective  
6  
7 499 loss of alleles (allelic dropout). Loci are only considered homozygous if they can be  
8  
9 500 confirmed in multiple PCR reactions (Morin *et al.* 2001; Taberlet *et al.* 1996). Capillary  
10  
11 501 electrophoresis requires that these replicates are run independently to distinguish true  
12  
13 502 alleles from non-specific signal, often more than tripling the amount of time and effort  
14  
15 503 required to genotype a single sample. In contrast, MiSeq genotyping can be performed  
16  
17 504 after combining the products of multiple PCR reactions. Although the quality of the input  
18  
19 505 DNA remains the same, MiSeq genotyping of pooled PCR replicates reduces the  
20  
21 506 frequency of allelic dropout and thus renders the resulting genotypes more accurate.  
22  
23 507 However, amplicon pooling foregoes data from repeat analyses, which are used by some  
24  
25 508 as a measure of DNA quality and/or data reliability (Taberlet *et al.* 1996)

26  
27  
28 509 Once MiSeq data files are imported into the CHIIMP platform, the program calls  
29  
30 510 alleles automatically, thus saving days of hands-on work. While automated allele calling  
31  
32 511 has been reported previously (De Barba *et al.* 2017; Suez *et al.* 2016; Zhan *et al.* 2017),  
33  
34 512 CHIIMP includes downstream analyses, such as alignments of allele sequences or  
35  
36 513 flagging loci that may contain contaminants, which provide important additional quality  
37  
38 514 control measures. In contrast to previous studies, CHIIMP also retains non-repeat regions  
39  
40 515 (Suez *et al.* 2016), which can contribute to allelic diversity, and does not require the  
41  
42 516 presence of stutter sequences for allele calling, which may not be sufficiently abundant  
43  
44 517 under conditions of low coverage (De Barba *et al.* 2017). Finally, CHIIMP reports both  
45  
46 518 allele length and sequence content, and is thus designed to detect minor length and  
47  
48 519 sequence differences by including sequence-specific allele names and generating locus-  
49  
50 520 specific sequence alignments (Figs. 4 and S1). To guide subsequent analyses, we have  
51  
52 521 also added features that flag potentially problematic alleles and standardize allele naming.  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 522 CHIIMP thus represents the most comprehensive analysis platform yet to ensure the  
4  
5 523 accuracy of MiSeq-based genotyping results.  
6

7 524

9 525 *Multiplexing improves MiSeq genotyping efficiency and reduces cost*

11 526

13 527 The Illumina MiSeq v2 500 sequencing kit has an output of ~25 million reads per run, thus  
14  
15 528 allowing the multiplexing of many samples, the number of which depends on the desired  
16  
17 529 read depth. Comparing read depths per allele, we found that a cut-off of 500 reads yielded  
18  
19 530 the most accurate results for our dataset. This value is higher than the 50 read cut-off  
20  
21 531 used previously to genotype laboratory raised fish (Zhan *et al.* 2017). However, the latter  
22  
23 532 study used high quality tissues rather than fecal samples for analysis. To determine the  
24  
25 533 sources of allele-calling errors, we did not multiplex samples from Gombe chimpanzees.  
26  
27 534 However, we tested multiplexing using samples from GME chimpanzees and confirmed  
28  
29 535 that this approach yields accurate results. Although primer incompatibilities allowed the  
30  
31 536 combination of only four loci, this number can be significantly increased with additional  
32  
33 537 primer design. For example, a recent study genotyped bear fecal DNA by multiplexing 14  
34  
35 538 loci (De Barba *et al.* 2017). Pooling amplicons from multiple loci after singleplex PCR can  
36  
37 539 circumvent the need for specialized primer design, as the maximum number of pooled loci  
38  
39 540 for any given sample is limited only by the desired read depth. Moreover, barcoding of  
40  
41 541 individual samples allows their combination in sequencing reactions, thus further  
42  
43 542 increasing sequencing efficiency and throughput (Farrell *et al.* 2016).  
44  
45  
46

47 543 MiSeq genotyping is expensive, but these costs decrease with sample numbers.  
48  
49 544 Capillary electrophoresis is undoubtedly cheaper when only a small number of samples  
50  
51 545 has to be analyzed; however, MiSeq sequencing becomes increasingly more cost-  
52  
53 546 effective with multiplexing and analyses of pooled replicates (Darby *et al.* 2016). The costs  
54  
55 547 of MiSeq genotyping three replicates of 96 samples multiplexed at four loci would roughly  
56  
57  
58  
59  
60

1  
2  
3 548 be equivalent to analyzing the same multiplexed samples via capillary electrophoresis,  
4  
5 549 because the latter cannot analyze pooled replicates. While this estimate only considers  
6  
7 550 genotyping supplies, labor to manually analyze samples is not included. In addition, the  
8  
9 551 improved accuracy has downstream cost advantages since fewer repeat analyses would  
10  
11 552 have to be performed.  
12

13 553

14  
15  
16 554 *Effective sharing of MiSeq genotyping data*  
17

18 555

19  
20 556 A direct comparison of MiSeq and capillary electrophoresis derived alleles revealed  
21  
22 557 consistent length differences of one to three nucleotides, the number of which were locus  
23  
24 558 specific (Tables 1 and S3). For Gombe samples, locus 3 alleles derived by capillary  
25  
26 559 electrophoresis were always three nucleotides longer than the corresponding MiSeq  
27  
28 560 alleles (Table 1). However, for the GME samples, the same alleles were all one nucleotide  
29  
30 561 shorter than the MiSeq alleles (Table S3). This is as expected since the capillary  
31  
32 562 electrophoresis data were generated on different platforms. However, this also means that  
33  
34 563 a simple conversion of existing capillary electrophoresis to MiSeq data will generally not  
35  
36 564 be possible. In contrast, MiSeq genotyping generates unambiguous alleles that can be  
37  
38 565 compared across multiple studies and field sites (Fig. 4). In the future, it will thus be  
39  
40 566 possible to compare STR genotypes across different chimpanzee populations, such as  
41  
42 567 those in Gombe and the GME, since the use of different sequencing equipment will no  
43  
44 568 longer confound these analyses.  
45  
46

47 569

48  
49 570 *Versatility of the CHIIMP analysis platform*  
50

51 571

52  
53  
54 572 To increase its utility, we designed the CHIIMP analysis platform to be versatile. STR  
55  
56 573 locus attributes, such as the expected length range, primer sequences, and repeat motifs,  
57  
58  
59  
60



1  
2  
3 574 as well as all thresholds for allele calling can be customized. For example, analysis of loci  
4  
5 575 with dinucleotide repeats may require a lower threshold for stutter peaks, since these are  
6  
7 576 more susceptible to polymerase slippage (Guichoux *et al.* 2011). Similarly, locus length  
8  
9 577 ranges can be expanded or contracted, depending on the rate of off-target amplicons. The  
10  
11 578 CHIIMP analysis pipeline also includes tools that facilitate iterative improvements for new  
12  
13 579 applications (Fig. S1). For example, the program provides a heatmap that indicates the  
14  
15 580 number of unique sequences that pass all filters. If that number is too high, thresholds can  
16  
17 581 be adjusted to remove stutter peaks, off-target amplicons, and/or PCR errors. In addition,  
18  
19 582 the distribution of loci is visualized, which can be used to reveal contamination in  
20  
21 583 singleplexed samples or identify poorly performing primers in multiplexed samples (Fig.  
22  
23 584 S1g). For potentially problematic alleles, CHIIMP generates histograms that provide  
24  
25 585 information concerning their length and relative abundance. All of these tools can be used  
26  
27 586 to adapt the platform to additional STR loci and/or host species.

28  
29  
30 587         The length of STR loci suitable for sequence-based genotyping depends on the  
31  
32 588 sequencing chemistry. We used Illumina v2 technology, which has maximum read lengths  
33  
34 589 of 500 nucleotides. MiSeq sequences are most often generated using paired-end reads,  
35  
36 590 with a maximum read length of 250 nucleotides in each direction. For STR genotyping,  
37  
38 591 locus sequences must span the repeat motif region, since assembly of shorter reads  
39  
40 592 could result in misalignments. To accommodate loci of greater than 250 bp length, we  
41  
42 593 opted to only use forward reads for analysis. Although the sequencing kit could  
43  
44 594 theoretically accommodate fragments of up to 500 nucleotides, we found that the quality  
45  
46 595 of reads (Q scores) decreased significantly after 400 cycles. Given that the longest locus  
47  
48 596 in our panel spanned 357 nucleotides, we used 375 cycles in the forward direction.  
49  
50 597 Illumina v3 sequencing chemistry has a 600-cycle limit, which may accommodate loci of  
51  
52 598 up to 500 bp, but this would have to be determined experimentally. The majority of  
53  
54 599 microsatellite loci are shorter than this length.  
55  
56  
57  
58  
59  
60

1  
2  
3 600 As STR genotyping transitions from capillary electrophoresis to sequence based  
4  
5 601 approaches, it will be necessary to standardize allele nomenclature, as has already been  
6  
7 602 suggested for human forensics (Gelardi *et al.* 2014; Parson *et al.* 2016). At a minimum,  
8  
9 603 allele names will have to incorporate the length and unique sequence content for each  
10  
11 604 allele (Darby *et al.* 2016). In our study, we added an alphabetical identifier (-a, -b, -c, etc.)  
12  
13 605 to differentiate identically sized alleles that differed in their sequence (Table S2). Since it  
14  
15 606 is impossible to capture all allele attributes in a single name, it may become necessary to  
16  
17 607 establish databases that link allele identifiers to their respective sequences. CHIIMP is  
18  
19 608 designed to allow users to supply a spreadsheet of allele names and sequences, and thus  
20  
21 609 guarantees consistent nomenclature across experiments. As MiSeq genotyping is  
22  
23 610 adapted to additional projects, standardized allele designations will become necessary to  
24  
25 611 ensure consistent nomenclature across studies.  
26  
27  
28  
29

### 30 613 *Conclusions*

31  
32 614  
33  
34 615 Genetic study of wild primates and other endangered species has been shown to provide  
35  
36 616 more accurate information concerning the size, structure, distribution and dynamics of  
37  
38 617 populations than observational studies. However, genotyping can be prohibitively  
39  
40 618 expensive given the large numbers of samples that are required for such analyses. The  
41  
42 619 MiSeq based genotyping platform provides a new approach that drastically reduces time  
43  
44 620 and labor, while providing more accurate and informative genotypes compared to capillary  
45  
46 621 electrophoresis. This will allow much faster and more streamlined analysis of samples that  
47  
48 622 are necessary for censusing and monitoring of non-habituated populations in addition to  
49  
50 623 revealing previously inaccessible allelic diversity. The CHIIMP platform has been  
51  
52 624 designed to be adaptable to additional loci and/or species. This allows the study of group  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 625 membership, dispersal, gene flow, and association patterns for a multitude of wildlife  
4  
5 626 species with broad conservation and biological implications.  
6

7 627

8  
9 628 *Author contributions*

10  
11 629

12  
13 630 All authors contributed to the acquisition, analysis, and interpretation of the data. H.J.B.,  
14  
15 631 A.J.C. and B.H.H. conceived, planned and executed the study; H.J.B., A.N.A., R.M.R.,  
16  
17 632 M.S.G. and Y.L. performed STR locus amplifications and data analyses; H.J.B. and A.J.C.  
18  
19 633 developed the CHIIMP analysis pipeline; A.G.S., A.L.S., and F.B.R. optimized the MiSeq  
20  
21 634 sequencing approach; D.M., E.V.L, F.A.S., A.K.P., and A.E.P. conducted or supervised  
22  
23 635 field work; A.J.C., E.E.W, and P.M.S. performed allelic diversity and parentage analyses;  
24  
25 636 H.J.B., A.J.C., R.M.R. and B.H.H coordinated the contributions of all authors and wrote  
26  
27 637 the manuscript.  
28  
29

30 638

31  
32 639 *Acknowledgements*

33  
34 640

35  
36  
37 641 We thank the Jane Goodall Institute field staff at the Gombe Stream Research Centre as  
38  
39 642 well as field assistants from the Greater Mahale Ecosystem Research and Conservation  
40  
41 643 Project (GMERC) for collecting chimpanzee observational data as well as fecal samples;  
42  
43 644 the Tanzania Commission for Science and Technology (COSTECH), the Tanzania Wildlife  
44  
45 645 Research Institute (TAWIRI), and the Tanzania National Parks Association (TANAPA) for  
46  
47 646 their support and permission to conduct research in Gombe and the GME. This work was  
48  
49 647 supported by grants from the National Institutes of Health, USA (R01 AI 091595, R37 AI  
50  
51 648 050529, R01 AI 120810, P30 AI 045008), the National Science Foundation (IOS-1052693,  
52  
53 649 IOS-1457260), the Jane Goodall Institute, and the University of California at San  
54  
55 650 Diego/Salk Center for Academic Research and Training in Anthropogeny (CARTA). H.J.B.  
56  
57  
58  
59  
60

1  
2  
3 651 and R.M.R. were funded by training grants (T32 AI 055400 and T32 AI 007632,  
4  
5 652 respectively). The authors declare no competing financial interests.  
6

7 653

8  
9 654 *Data accessibility*

10 655

11  
12  
13 656 STR sequences are archived in the NCBI Sequence Read Archive (SRA) under  
14  
15 657 <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA434411>. Preprocessed sequence data,  
16  
17 658 analysis software and supporting R code are archived on Dryad doi: #####. Ongoing  
18  
19 659 Software Development, including supporting R code, is available on  
20  
21  
22 660 <https://github.com/ShawHahnLab/chiimp/releases/tag/0.1.0>  
23

24 661 .

25  
26 662

27  
28 663 *References*

29  
30 664

31  
32 665 Adams RI, Brown KM, Hamilton MB (2004) The impact of microsatellite electromorph size  
33  
34 666 homoplasy on multilocus population structure estimates in a tropical tree  
35  
36 667 (Corythophora alta) and an anadromous fish (Morone saxatilis). *Molecular Ecology*  
37  
38 668 **13**, 2579-2588.

39  
40 669 Arandjelovic M, Guschanski K, Schubert G, *et al.* (2009) Two-step multiplex polymerase  
41  
42 670 chain reaction improves the speed and accuracy of genotyping using DNA from  
43  
44 671 noninvasive and museum samples. *Molecular Ecology Resources* **9**, 28-36.

45  
46 672 Arandjelovic M, Vigilant L (2018) Non-invasive genetic censusing and monitoring of  
47  
48 673 primate populations. *American Journal of Primatology*, e22743.

49  
50 674 Balloux F, Lugon-Moulin N (2002) The estimation of population differentiation with  
51  
52 675 microsatellite markers. *Molecular Ecology* **11**, 155-165.  
53  
54  
55  
56  
57  
58  
59  
60

- 1  
2  
3 676 Barbian HJ, Li Y, Ramirez M, *et al.* (2018) Destabilization of the gut microbiome marks the  
4  
5 677 end-stage of simian immunodeficiency virus infection in wild chimpanzees.  
6  
7 678 *American Journal of Primatology* **80**, 10.1002/ajp.22515.  
8  
9 679 Becquet C, Patterson N, Stone AC, Przeworski M, Reich D (2007) Genetic structure of  
10  
11 680 chimpanzee populations. *PLOS Genetics* **3**, e66.  
12  
13 681 Bennett P (2000) Demystified ...: Microsatellites. *Molecular Pathology* **53**, 177-183.  
14  
15 682 Bodenhofer U, Bonatesta E, Horejš-Kainrath C, Hochreiter S (2015) msa: an R package  
16  
17 683 for multiple sequence alignment. *Bioinformatics* **31**, 3997-3999.  
18  
19 684 Broquet T, Petit E (2004) Quantifying genotyping errors in noninvasive population  
20  
21 685 genetics. *Molecular Ecology* **13**, 3601-3608.  
22  
23 686 Charlesworth B, Charlesworth D (2010) *Elements of evolutionary genetics*. Roberts and  
24  
25 687 Co. Publishers. 734 p.  
26  
27 688 Constable JJ, Packer C, Collins DA, Pusey AE (1995) Nuclear DNA from primate dung.  
28  
29 689 *Nature* **373**, 393.  
30  
31 690 Constable JL, Ashley MV, Goodall J, Pusey AE (2001) Noninvasive paternity assignment  
32  
33 691 in Gombe chimpanzees. *Molecular Ecology* **10**, 1279-1300.  
34  
35 692 Darby BJ, Erickson SF, Hervey SD, Ellis-Felege SN (2016) Digital fragment analysis of  
36  
37 693 short tandem repeats by high-throughput amplicon sequencing. *Ecology and*  
38  
39 694 *Evolution* **6**, 4502-4512.  
40  
41 695 De Barba M, Miquel C, Lobréaux S, *et al.* (2017) High-throughput microsatellite  
42  
43 696 genotyping in ecology: improved accuracy, efficiency, standardization and success  
44  
45 697 with low-quantity and degraded DNA. *Molecular Ecology Resources*. **17**, 492-507.  
46  
47 698 Ellegren H (2004) Microsatellites: simple sequences with complex evolution. *Nature*  
48  
49 699 *Reviews Genetics* **5**, 435.  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

- 1  
2  
3 700 Ewen KR, Bahlo M, Treloar SA, *et al.* (2000) Identification and analysis of error types in  
4  
5 701 high-throughput genotyping. *American Journal of Human Genetics* **67**, 727-736.  
6  
7 702 Farrell ED, Carlsson JEL, Carlsson J (2016) Next Gen Pop Gen: implementing a high-  
8  
9 703 throughput approach to population genetics in boarfish (*Capros aper*). *Royal*  
10  
11 704 *Society Open Science* **3**, 160651.  
12  
13 705 Fernando P, Evans BJ, Morales JC, Melnick DJ (2001) Electrophoresis artefacts — a  
14  
15 706 previously unrecognized cause of error in microsatellite analysis. *Molecular*  
16  
17 707 *Ecology Notes* **1**, 325-328.  
18  
19 708 Fordyce SL, Ávila-Arcos MC, Rockenbauer E, *et al.* (2011) High-throughput sequencing of  
20  
21 709 core STR loci for forensic genetic investigations using the Roche Genome  
22  
23 710 Sequencer FLX platform. *BioTechniques* **51**, 127-133.  
24  
25 711 Gelardi C, Rockenbauer E, Dalsgaard S, Børsting C, Morling N (2014) Second generation  
26  
27 712 sequencing of three STRs D3S1358, D12S391 and D21S11 in Danes and a new  
28  
29 713 nomenclature for sequenced STR alleles. *Forensic Science International: Genetics*  
30  
31 714 **12**, 38-41.  
32  
33 715 Guichoux E, Lagache L, Wagner S, *et al.* (2011) Current trends in microsatellite  
34  
35 716 genotyping. *Molecular Ecology Resources* **11**, 591-611.  
36  
37 717 Hauge XY, Litt M (1993) A study of the origin of 'shadow bands' seen when typing  
38  
39 718 dinucleotide repeat polymorphisms by the PCR. *Human Molecular Genetics*. **2**,  
40  
41 719 411-415.  
42  
43 720 Iyer SS, Bibollet-Ruche F, Sherrill-Mix S, *et al.* (2017) Resistance to type 1 interferons is a  
44  
45 721 major determinant of HIV-1 transmission fitness. *Proceedings of the National*  
46  
47 722 *Academy of Sciences* **114**: E590-E599.  
48  
49 723 Jarne P, Lagoda PJJ (1996) Microsatellites, from molecules to populations and back.  
50  
51 724 *Trends in Ecology & Evolution* **11**, 424-429.  
52  
53  
54  
55  
56  
57  
58  
59  
60

- 1  
2  
3 725 Keele BF, Jones JH, Terio KA, *et al.* (2009) Increased mortality and AIDS-like  
4  
5 726 immunopathology in wild chimpanzees infected with SIVcpz. *Nature* **460**, 515-519.  
6  
7 727 Kelkar YD, Strubczewski N, Hile SE, *et al.* (2010) What is a microsatellite: a computational  
8  
9 728 and experimental definition based upon repeat mutational behavior at A/T and  
10  
11 729 GT/AC repeats. *Genome Biology and Evolution* **2**, 620-635.  
12  
13 730 Langergraber KE, Mitani JC, Vigilant L (2007) The limited impact of kinship on  
14  
15 731 cooperation in wild chimpanzees. *Proceedings of the National Academy of*  
16  
17 732 *Sciences* **104**, 7786-7790.  
18  
19 733 Levinson G, Gutman GA (1987) Slipped-strand mispairing: a major mechanism for DNA  
20  
21 734 sequence evolution. *Molecular biology and evolution.* **4**, 203-221.  
22  
23 735 Lynch HT, de la Chapelle A (2003) Hereditary Colorectal Cancer. *New England Journal of*  
24  
25 736 *Medicine* **348**, 919-932.  
26  
27 737 Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing  
28  
29 738 reads. *EMBnet.journal* **17**, 10-12.  
30  
31 739 Moeller AH, Foerster S, Wilson M, *et al.* (2016) Social behaviour promotes diversity in the  
32  
33 740 chimpanzee gut microbiome. *American Association for the Advancement of*  
34  
35 741 *Science* **2**, e1500997.  
36  
37 742 Moore J (1996) Savanna chimpanzees, referential models and the last common ancestor.  
38  
39 743 *Great Ape Societies*. Cambridge University Press, 275-292.  
40  
41 744 Morin PA, Chambers KE, Boesch C, Vigilant L (2001) Quantitative polymerase chain  
42  
43 745 reaction analysis of DNA from noninvasive samples for accurate microsatellite  
44  
45 746 genotyping of wild chimpanzees (*Pan troglodytes verus*). *Molecular Ecology* **10**,  
46  
47 747 1835-1844.  
48  
49 748 Morin PA, Wallis J, Moore JJ, Chakraborty R, Woodruff DS (1993) Non-invasive sampling  
50  
51 749 and DNA amplification for paternity exclusion, community structure, and  
52  
53 750 phylogeography in wild chimpanzees. *Primates* **34**, 347-356.  
54  
55  
56  
57  
58  
59  
60

- 1  
2  
3 751 Ogawa H, Idani Gi, Kanamori M (1999) Chimpanzee habitat in the savanna woodland,  
4  
5 752 Ugalla, Tanzania. *Primate Research* **15**, 135-146.  
6  
7 753 Parson W, Ballard D, Budowle B, *et al.* (2016) Massively parallel sequencing of forensic  
8  
9 754 STRs: Considerations of the DNA commission of the International Society for  
10  
11 755 Forensic Genetics (ISFG) on minimal nomenclature requirements. *Forensic*  
12  
13 756 *Science International: Genetics* **22**, 54-63.  
14  
15 757 Pasqualotto AC, Denning DW, Anderson MJ (2007) A cautionary tale: lack of consistency  
16  
17 758 in allele sizes between two laboratories for a published multilocus microsatellite  
18  
19 759 typing system. *Journal of Clinical Microbiology* **45**, 522-528.  
20  
21 760 Peakall ROD, Smouse PE (2006) genalex 6: genetic analysis in Excel. Population genetic  
22  
23 761 software for teaching and research. *Molecular Ecology Notes* **6**, 288-295.  
24  
25 762 Pusey AE, Pintea L, Wilson ML, Kamenya S, Goodall J (2007) The contribution of long-  
26  
27 763 term research at Gombe National Park to chimpanzee conservation. *Conservation*  
28  
29 764 *Biology* **21**, 623-634.  
30  
31 765 Queller DC, Strassmann JE, Hughes CR (1993) Microsatellites and kinship. *Trends in*  
32  
33 766 *Ecology & Evolution* **8**, 285-288.  
34  
35 767 Rudicell RS, Holland Jones J, Wroblewski EE, *et al.* (2010) Impact of simian  
36  
37 768 immunodeficiency virus infection on chimpanzee population dynamics. *PLoS*  
38  
39 769 *Pathogens* **6**, e1001116.  
40  
41 770 Rudicell RS, Piel AK, Stewart F, *et al.* (2011) High prevalence of simian immunodeficiency  
42  
43 771 virus infection in a community of savanna chimpanzees. *Journal of Virology* **85**,  
44  
45 772 9918-9928.  
46  
47 773 Santiago ML, Lukasik M, Kamenya S, *et al.* (2003) Foci of endemic simian  
48  
49 774 immunodeficiency virus infection in wild-living eastern chimpanzees (*Pan*  
50  
51 775 *trogodytes schweinfurthii*). *Journal of Virology* **77**, 7545-7562.  
52  
53  
54  
55  
56  
57  
58  
59  
60



- 1  
2  
3 776 Schoeninger MJ, Moore J, Sept JM (1999) Subsistence strategies of two “savanna”  
4  
5 777 chimpanzee populations: The stable isotope evidence. *American Journal of*  
6  
7 778 *Primatology* **49**, 297-314.  
8  
9 779 Schuelke M (2000) An economic method for the fluorescent labeling of PCR fragments.  
10  
11 780 *Nature Biotechnology* **18**, 233.  
12  
13 781 Selkoe KA, Toonen RJ (2006) Microsatellites for ecologists: a practical guide to using and  
14  
15 782 evaluating microsatellite markers. *Ecology Letters* **9**, 615-629.  
16  
17 783 Shinde D, Lai Y, Sun F, Arnheim N (2003) Taq DNA polymerase slippage mutation rates  
18  
19 784 measured by PCR and quasi-likelihood analysis: (CA/GT)(n) and (A/T)(n)  
20  
21 785 microsatellites. *Nucleic Acids Research* **31**, 974-980.  
22  
23 786 Suez M, Behdenna A, Brouillet S, *et al.* (2016) MicNeSs: genotyping microsatellite loci  
24  
25 787 from a collection of (NGS) reads. *Molecular Ecology Resources* **16**, 524-533.  
26  
27 788 Taberlet P, Camarra JJ, Griffin S, *et al.* (1997) Noninvasive genetic tracking of the  
28  
29 789 endangered Pyrenean brown bear population. *Molecular Ecology* **6**, 869-876.  
30  
31 790 Taberlet P, Griffin S, Goossens B, *et al.* (1996) Reliable genotyping of samples with very  
32  
33 791 low DNA quantities using PCR. *Nucleic Acids Research* **24**, 3189-3194.  
34  
35 792 van Lawick-Goodall J (1968) The behaviour of free-living chimpanzees in the Gombe  
36  
37 793 Stream Reserve. *Animal Behaviour Monographs* **1**, 161-311.  
38  
39 794 Van Neste C, Van Nieuwerburgh F, Van Hoofstat D, Deforce D (2012) Forensic STR  
40  
41 795 analysis using massive parallel sequencing. *Forensic Science International*. **6**,  
42  
43 796 810-818.  
44  
45 797 Vartia S, Villanueva-Cañas JL, Finarelli J, *et al.* (2016) A novel method of microsatellite  
46  
47 798 genotyping-by-sequencing using individual combinatorial barcoding. *Royal Society*  
48  
49 799 *Open Science* **3**, 150565.  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

- 1  
2  
3 800 Vigilant L, Hofreiter M, Siedel H, Boesch C (2001) Paternity and relatedness in wild  
4  
5 801 chimpanzee communities. *Proceedings of the National Academy of Sciences* **98**,  
6  
7 802 12890-12895.  
8  
9 803 Walker KK, Rudicell RS, Li Y, *et al.* (2017) Chimpanzees breed with genetically dissimilar  
10  
11 804 mates. *Royal Society Open Science* **4**, 160422.  
12  
13 805 Weeks DE, Conley YP, Ferrell RE, Mah TS, Gorin MB (2002) A tale of two genotypes:  
14  
15 806 consistency between two high-throughput genotyping centers. *Genome Res* **12**,  
16  
17 807 430-435.  
18  
19 808 Wroblewski EE, Norman PJ, Guethlein LA, *et al.* (2015) Signature patterns of MHC  
20  
21 809 diversity in three gombe communities of wild chimpanzees reflect fitness in  
22  
23 810 reproduction and immune defense against SIVcpz. *PLoS Biology* **13**, e1002144.  
24  
25 811 Zhan L, Paterson IG, Fraser BA, *et al.* (2017) megasat: automated inference of  
26  
27 812 microsatellite genotypes from sequence data. *Molecular Ecology Resources* **17**,  
28  
29 813 247-256.  
30  
31  
32  
33  
34

### 35 **Figure legends**

36  
37 816  
38  
39 817 **Fig 1** MiSeq genotyping uncovers cryptic alleles. Eight polymorphic STR loci were  
40  
41 818 amplified from the fecal DNA of 19 previously genotyped chimpanzees. (a) Histogram  
42  
43 819 depicting the length (x-axis) and read count (y-axis) of unique sequences for one  
44  
45 820 representative heterozygous locus that was previously determined to be homozygous by  
46  
47 821 multiple capillary electrophoresis analyses (sample 4861, locus C, Table 1). The grey box  
48  
49 822 highlights the expected locus size range. The horizontal line indicates the cutoff of 500  
50  
51 823 reads. Colored peaks indicate reads that passed the locus-specific filters (note that peaks  
52  
53 824 can be comprised of identically sized reads that differ in their sequence content). Black  
54  
55 825 reads were eliminated. Pink reads appear to be locus-specific, but did not pass the PCR  
56  
57  
58  
59  
60

1  
2  
3 826 artifact filters. Red reads represent the true allele sequences (180 and 181 bp in lengths,  
4  
5 827 respectively). (b, c) Alignment images of locus-specific allele sequences are shown for  
6  
7 828 locus 1 (b) and locus C (c), respectively (the complete data set is shown in Table 1). Allele  
8  
9 829 sequences are ordered by length (indicated in bp on the right), with the frequency with  
10  
11 830 which they were found in different chimpanzees indicated on the left (the x-axis indicates  
12  
13 831 the position within the alignment). Nucleotides are colored as shown, with gaps in the  
14  
15 832 alignment shown in grey. The insets highlight alleles that differ in their sequence content  
16  
17 833 and/or length. Nucleotide substitutions are colored; dashes indicate gaps that were  
18  
19 834 introduced to optimize the alignment.  
20  
21  
22  
23

24 836 **Fig. 2** MiSeq genotyping uncovers increased allelic diversity and heterozygosity. (a)  
25  
26 837 Alignment of four locus 3 alleles that are of identical length (234 bp), but differ in  
27  
28 838 sequence content. Nucleotide substitutions are colored; dashes indicate single nucleotide  
29  
30 839 insertions and deletions (b) Mendelian inheritance of allele 234 for a group of related  
31  
32 840 chimpanzees. Fathers and mothers are shown as squares and circles, respectively, with  
33  
34 841 offspring connected by vertical lines. Both alleles are shown for each animal, with the four  
35  
36 842 allelic variants highlighted in different colors. Individuals of unknown identity or genotype  
37  
38 843 are left blank. (c) Increased allelic diversity resolves a previously ambiguous paternity  
39  
40 844 determination. Two potential fathers with identical allele lengths (238 bp) can now be  
41  
42 845 distinguished based on differences in allele sequence content (238-a and 238-b). Since  
43  
44 846 the offspring is homozygous for allele 238-a, the male with allele 238-b can be excluded  
45  
46 847 as a father.  
47  
48  
49  
50

51 849 **Fig. 3** Individual identification based on MiSeq genotyping. (a-c) Genotypes of newly  
52  
53 850 collected samples (top) are compared to the genotypes of known community members,  
54  
55 851 with the closest match listed below (based on descending distance scores). Genotypes  
56  
57  
58  
59  
60

1  
2  
3 852 that differ by fewer than four alleles are indicated in bold because they represent likely  
4  
5 853 matches. Differences are highlighted in yellow. (d) Heatmap showing the relative similarity  
6  
7 854 of sample genotypes (rows) with genotypes of known individuals (columns) based on  
8  
9 855 distance scores. Dark red cells indicate likely matches.  
10

11  
12 856

13  
14 857 **Fig 4** Comparison of MiSeq genotypes across chimpanzee communities. Alignment  
15  
16 858 images of locus-specific allele sequences are shown for chimpanzees from the GME and  
17  
18 859 Gombe. Two representative loci (locus B on the left; locus D on the right) are shown for  
19  
20 860 (a) 12 chimpanzees from the GME (Table S3), (b) 123 chimpanzees from Gombe (Table  
21  
22 861 S2), and (c) a combination of both. Allele sequences are ordered by length (indicated in  
23  
24 862 base pairs on the right), with the frequency with which they were found in different  
25  
26 863 chimpanzees indicated on the left (the x-axis indicates the position within the alignment).  
27  
28 864 Nucleotides are colored as indicated, with alignment gaps shown in grey. Arrows indicate  
29  
30 865 alleles that are unique to the GME samples.  
31

32  
33 866

34  
35 867 **Fig. 5** MiSeq-based STR genotyping of wild chimpanzees. (a) Schematic representation  
36  
37 868 of singleplex STR amplification and MiSeq sequencing of chimpanzee fecal DNA. (b)  
38  
39 869 Schematic representation of the CHIIMPS analysis pipeline with decision tree and  
40  
41 870 downstream data reports.  
42

43  
44 871  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3  
4  
5  
6  
7  
8 1 | **CHIIMP: An fully automated high-throughput microsatellite genotyping**  
9 2 | **platform reveals greater allelic diversity in wild chimpanzees**  
10 3 |

11 4 | Hannah J. Barbian<sup>1</sup>, A. Jesse Connell<sup>1</sup>, Alexa N. Avitto<sup>1</sup>, Ronnie M. Russell<sup>1</sup>,  
12 5 | Andrew G. Smith<sup>1</sup>, Madhurima S. Gundlapally<sup>1</sup>, Alexander L. Shazad<sup>1</sup>, Yingying Li<sup>1</sup>,  
13 6 | Frederic Bibollet-Ruche<sup>1</sup>, Emily E. Wroblewski<sup>2</sup>, Deus Mjungu<sup>3</sup>, Elizabeth V. Lonsdorf<sup>4</sup>,  
14 7 | Fiona A. Stewart<sup>5</sup>, Alexander K. Piel<sup>5</sup>, Anne E. Pusey<sup>6</sup>,  
15 8 | Paul M. Sharp<sup>7</sup> and Beatrice H. Hahn<sup>1\*</sup>  
16 9 |

17 10 | <sup>1</sup>Departments of Microbiology and Medicine, Perelman School of Medicine, University of  
18 11 | Pennsylvania, Philadelphia, PA, USA

19 12 | <sup>2</sup>Department of Anthropology, Washington University in St. Louis, St. Louis, MO, USA

20 13 | <sup>3</sup>Gombe Stream Research Center, Kigoma, Tanzania

21 14 | <sup>4</sup>Department of Psychology, Franklin and Marshall College, Lancaster, Pennsylvania

22 15 | <sup>5</sup>School of Natural Sciences and Psychology, Liverpool John Moores University,  
23 16 | Liverpool, United Kingdom

24 17 | <sup>6</sup>Department of Evolutionary Anthropology, Duke University, Durham, North Carolina

25 18 | <sup>7</sup>Institute of Evolutionary Biology and Centre for Immunity, Infection and Evolution,  
26 19 | University of Edinburgh, Edinburgh EH9 3FL, United Kingdom

27 20 | Professor, Departments of Medicine and Microbiology  
28 21 | University of Pennsylvania Perelman School of Medicine  
29 22 | 409 Johnson Pavilion  
30 23 | 3610 Hamilton Walk  
31 24 | Philadelphia, PA 19104-6076  
32 25 | USA  
33 26 | bhahn@penmedicine.upenn.edu  
34 27 |

35 28 |  
36 29 | Running title: High throughput STR genotyping of chimpanzee  
37 30 |  
38 31 |  
39 32 |  
40 33 |  
41 34 |  
42 35 |  
43 36 |  
44 37 |  
45 38 |  
46 39 |  
47 40 |  
48 41 |  
49 42 |  
50 43 |  
51 44 |  
52 45 |  
53 46 |  
54 47 |  
55 48 |  
56 49 |  
57 50 |  
58 51 |  
59 52 |  
60 53 |

1  
2  
3  
4  
5  
6  
7  
8 **30 Abstract**

9  
10 31 Short tandem repeats (STRs), also known as microsatellites, are commonly used to non-  
11 32 invasively genotype wild-living endangered species, including African apes. Until recently,  
12 33 capillary electrophoresis has been the method of choice to determine the length of  
13 34 polymorphic STR loci. However, this technique is labor intensive, difficult to compare  
14 35 across platforms, and notoriously imprecise. Here we developed a MiSeq-based approach  
15 36 and tested its performance using previously genotyped fecal samples from long-term  
16 37 studied chimpanzees in Gombe National Park, Tanzania. Using data from eight ~~previously~~  
17 38 ~~characterized~~ microsatellite loci as a reference, we designed a bioinformatics platform that  
18 39 converts raw MiSeq reads into locus-specific files and automatically calls alleles after  
19 40 filtering stutter sequences and other PCR artifacts. Applying this method to the entire  
20 41 Gombe population, we confirmed previously reported genotypes, but also identified 31  
21 42 new alleles that had been missed due to sequence differences and size homoplasy. The  
22 43 new genotypes, which increased the allelic diversity and heterozygosity in Gombe by 61%  
23 44 and 8%, respectively, were validated by replicate amplification and pedigree analysis.  
24 45 This demonstrated inheritance and resolved one case of an ambiguous paternity. Using  
25 46 both singleplex and multiplex locus amplification, we also genotyped fecal samples from  
26 47 chimpanzees in the Greater Mahale Ecosystem in Tanzania, demonstrating the utility of  
27 48 the MiSeq-based approach for genotyping non-habituated populations and performing  
28 49 comparative analyses across field sites. The new ~~fully~~-automated high-throughput  
29 50 analysis platform (available at <https://github.com/ShawHahnLab/chiimp>) will allow  
30 51 biologists to more accurately and effectively determine wildlife population size and  
31 52 structure, and thus obtain information critical for conservation efforts.

32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52

53  
54 **Keywords:** high-throughput STR genotyping, length homoplasy, parentage analysis, short  
55 tandem repeats (STRs), *Pan troglodytes*

## 56 Introduction

57

58 Microsatellites comprise short tandem repeats (STRs) of one to six base pairs, which are  
59 commonly used to profile DNA for a variety of applications ranging from cancer diagnosis  
60 to forensics (Bennett 2000; Ellegren 2004; Guichoux *et al.* 2011; Lynch & de la Chapelle  
61 2003). STR loci have a high mutation rate and vary in the number of their repeat motifs,  
62 due to slippage of the polymerase during DNA synthesis (Kelkar *et al.* 2010; Levinson &  
63 Gutman 1987). Because of their ubiquity, high allelic diversity and co-dominant  
64 inheritance, microsatellites are commonly used for individual identification, parentage  
65 analyses and population genetics (Balloux & Lugon-Moulin 2002; Jarne & Lagoda 1996;  
66 Queller *et al.* 1993; Selkoe & Toonen 2006). STR analysis can also be performed on  
67 samples containing little host DNA, such as hair and fecal samples, and has thus been the  
68 method of choice to genotype endangered primate species, which are typically sampled  
69 non-invasively (Constable *et al.* 1995; Constable *et al.* 2001; Morin *et al.* 1993; Taberlet *et*  
70 *al.* 1997). An accurate determination of ~~wild primate the~~ numbers, distribution, and  
71 population connectivity of wild primates is essential for designing effective conservation  
72 measures to protect these species under increasing anthropogenic threat from habitat  
73 loss, disease and hunting (Arandjelovic & Vigilant 2018). However, census and population  
74 genetics studies of wild apes have been impeded by difficulties of accurately and cost  
75 effectively genotype large numbers of non-invasively collected samples.

76 Until recently, the length of polymorphic STR loci has been determined by capillary  
77 electrophoresis, which compares the mobility of fluorescently labeled PCR products to a  
78 size standard of control fragments and thus yields only approximate results (e.g., a locus  
79 size of “167.5 bp”). Manual correction of such ambiguities can lead to arbitrary allele  
80 binning and inconsistent calls between experiments and/or investigators (Ewen *et al.*  
81 2000; Weeks *et al.* 2002). In addition, amplification of STR loci frequently generates PCR

1  
2  
3  
4  
5  
6  
7  
8 82 artifacts, which are difficult to identify on electropherograms. These include stutter peaks,  
9  
10 83 which are usually one repeat shorter than the correct STR allele and derive from Taq  
11 84 polymerase slippage (Hauge & Litt 1993; Shinde *et al.* 2003), split peaks which are  
12 85 caused by inconsistent A-overhang addition (Schuelke 2000), and artifactual peaks, which  
13 86 are the product of off-target amplification and/or unspecific fluorescent signaling (Ewen *et*  
14 87 *al.* 2000; Fernando *et al.* 2001; Guichoux *et al.* 2011). Existing peak-calling software often  
15 88 fails to differentiate erroneous from real peaks and frequently omits peaks of low height.  
16 89 Automatically called peaks must therefore be corrected manually, which is labor intensive  
17 90 and time consuming (Guichoux *et al.* 2011). Finally, multiplexing is restricted to only a few  
18 91 fluorescent labels, thus limiting the number of loci that can be analyzed simultaneously.  
19 92 As a consequence, capillary electrophoresis based STR genotyping is laborious,  
20 93 notoriously imprecise, and generally not useful for large sample sets or data sharing  
21 94 between different platforms and/or field sites (Pasqualotto *et al.* 2007).

22 95 To improve the accuracy and throughput of STR genotyping, investigators have  
23 96 begun to use next-generation sequencing (NGS) technologies to characterize amplified  
24 97 microsatellite loci. This approach is superior to capillary electrophoresis, since it yields  
25 98 unambiguous allele lengths regardless of protocol or sequencing platform. In addition,  
26 99 genotyping-by-sequencing (GBS) distinguishes alleles of the same size that contain  
27 100 substitutions or differ in length by a single nucleotide (Adams *et al.* 2004). Although  
28 101 initially developed for human forensics (Fordyce *et al.* 2011; Van Neste *et al.* 2012), GBS  
29 102 technologies have recently been used to genotype wild animals, including Atlantic cod  
30 103 (Vartia *et al.* 2016), brown bear (De Barba *et al.* 2017), boarfish (Farrell *et al.* 2016), and  
31 104 muskrat (Darby *et al.* 2016). These studies demonstrated the utility of GBS for molecular  
32 105 ecology applications (Darby *et al.* 2016; Farrell *et al.* 2016) and showed that even  
33 106 samples containing small quantities of host DNA, such as dung and hair, can be used for  
34 107 these analyses (De Barba *et al.* 2017). However, alleles were primarily called manually by



1  
2  
3  
4  
5  
6  
7  
8 108 visual inspection of read length histograms (Darby *et al.* 2016; Farrell *et al.* 2016; Vartia *et*  
9  
10 109 *al.* 2016), and none of these studies have compared the performance of capillary  
11 110 electrophoresis and high throughput sequencing [directly side-by-side](#) to validate and  
12  
13 111 improve the genotyping approach.

14  
15 112 For nearly two decades, our group has been studying chimpanzees in Gombe  
16  
17 113 National Park (Tanzania) to assess the long-term impact of simian immunodeficiency virus  
18  
19 114 (SIVcpz) infection on this wild-living population (Keele *et al.* 2009; Rudicell *et al.* 2010;  
20  
21 115 Santiago *et al.* 2003). To identify SIVcpz infected individuals, we developed non-invasive  
22  
23 116 diagnostic assays that detect virus-specific antibodies and nucleic acids by analysis of  
24  
25 117 fecal samples. To reliably monitor the spread of SIVcpz in all three Gombe communities,  
26  
27 118 we verified the individual origin of each fecal sample by microsatellite analysis at eight  
28  
29 119 polymorphic STR loci. Thus, most Gombe chimpanzees have been repeatedly genotyped,  
30  
31 120 resulting in a consensus genotype that has been used for paternity and kinship  
32  
33 121 determinations, immunogenetics, microbiome analyses and behavioral studies (Barbian *et*  
34  
35 122 *al.* 2018; Keele *et al.* 2009; Moeller *et al.* 2016; Rudicell *et al.* 2010; Santiago *et al.* 2003;  
36  
37 123 Walker *et al.* 2017; Wroblewski *et al.* 2015).

38  
39 124 Here, we used these multiply confirmed reference microsatellites as a guide to  
40  
41 125 develop and iteratively improve a MiSeq-based STR genotyping approach. To permit the  
42  
43 126 direct comparison with previous capillary electrophoresis results, we determined the  
44  
45 127 length of STR loci by sequencing PCR amplicons in their entirety, including both forward  
46  
47 128 and reverse primers. We also developed a *Computational High-throughput Individual*  
48  
49 129 *Identification through Microsatellite Profiling* (CHIIMP) pipeline that detects and filters  
50  
51 130 erroneous alleles and [automatically](#) generates a number of downstream analyses, such  
52  
53 131 as allele length histograms, alignments of allele sequences, contamination heatmaps and  
54  
55 132 genotype comparisons. [By directly comparing the new CHIIMP-derived genotypes to](#)  
56  
57 133 [previously determined capillary electrophoresis results, we show that the new analysis](#)

1  
2  
3  
4  
5  
6  
7  
8 134 [tools, which are not included in any of the previously published STR genotyping pipelines,](#)  
9  
10 135 [greatly improve the speed, cost and accuracy of allele determinations. Using this method](#)  
11 136 [to genotype fecal samples from two previously studied chimpanzee populations, we tested](#)  
12 137 [whether CHIMP is superior to capillary electrophoresis in terms of intensity of time and](#)  
13 138 [labor, and accuracy of results that can be compared across field sites.](#)  
14  
15  
16  
17 139  
18  
19 140

## 141 **Material and methods**

### 142 143 *Chimpanzee fecal samples*

144  
145 Fecal samples were collected from wild-living chimpanzees in Gombe National Park,  
146 including members of the Mitumba, Kasekela and Kalande communities, as well as the  
147 Greater Mahale Ecosystem (GME) in Tanzania as previously described (Keele *et al.* 2009;  
148 Rudicell *et al.* 2010; Rudicell *et al.* 2011; Santiago *et al.* 2003). Habituated Gombe  
149 chimpanzees have been under direct observation since the 1960s (Pusey *et al.* 2007; van  
150 Lawick-Goodall 1968), with prospective fecal sampling and SIVcpz diagnostics initiated in  
151 1999 (Keele *et al.* 2009; Rudicell *et al.* 2010). Long-term monitoring of non-habituated  
152 chimpanzees in the ~~Greater Mahale Ecosystem (GME) in Tanzania~~ began in 2008, with  
153 non-invasive SIVcpz screening implemented in 2009 (Rudicell *et al.* 2011). Gombe and  
154 GME fecal samples were collected 1:1 (vol/vol) in RNA<sup>later</sup> (Ambion), a high salt solution  
155 that preserves nucleic acids and allows storage and transport at room temperature. For  
156 individual identification, samples were routinely subjected to mitochondrial, sex, and  
157 microsatellite analyses, with up to eight STR loci characterized by capillary  
158 electrophoresis as described previously (Keele *et al.* 2009; Rudicell *et al.* 2010; Rudicell *et*  
159 *al.* 2011). All fieldwork has been approved by the Tanzania National Parks, the Tanzania

1  
2  
3  
4  
5  
6  
7  
8 160 Commission for Science and Technology, the Tanzania Wildlife Research Institute, and  
9  
10 161 has followed the American Society of Primatologists' Principles for Ethical Treatment of  
11  
12 162 Nonhuman Primates.

13 163

14  
15 164 *Quantification of chimpanzee DNA*

16  
17 165

18 166 Fecal DNA was extracted from 0.5 ml of homogenized fecal suspension using the  
19  
20 167 QIAamp DNA Stool Kit and the automated QIAcube system (Qiagen). Purified DNA was  
21  
22 168 eluted in 200  $\mu$ l water and stored at -20 °C. Chimpanzee genomic DNA content was  
23  
24 169 determined using a previously described *c-myc* gene-based quantitative (q)PCR (Morin *et*  
25  
26 170 *al.* 2001). Briefly, 2  $\mu$ l DNA extract was added to 1x High Fidelity PCR Buffer, 3.5 mM  
27  
28 171  $MgSO_4$ , 0.3  $\mu$ M forward (5'-GCCAGAGGAGGAACGAGCT-3') and reverse (5'-  
29  
30 172 GGGCCTTTTCATTGTTTTCCA-3') qPCR primers, 0.2  $\mu$ M of a FAM-labeled probe (FAM-  
31  
32 173 TGCCCTGCGTGACCAGATCC-BHQ1), 0.2 mM dNTPs, 1x ROX Reference Dye, and 0.5  
33  
34 174 U Platinum Taq DNA Polymerase High Fidelity (Invitrogen). Each sample was run in  
35  
36 175 triplicate on a 7900HT Fast Real-Time PCR System, together with human genomic DNA  
37  
38 176 standards of known concentration (the sequence of the particular *c-myc* amplicon is  
39  
40 177 identical between humans and chimpanzees). Negative "no-template" controls were  
41  
42 178 included in each run. Sequence Detection Systems version 2.3 software (Applied  
43  
44 179 Biosystems) was used to quantify the host DNA content of each sample. Since host DNA  
45  
46 180 concentrations differed, approximately half of the samples were extracted on more than  
47  
48 181 one occasion to generate enough material for all analyses.

49  
50 182

51  
52 183 *Amplification of STR loci*

53  
54 184

1  
2  
3  
4  
5  
6  
7  
8 185 Previous genotyping studies of Gombe and GME chimpanzees utilized eight STR loci  
9  
10 186 containing tetranucleotide repeats (Constable *et al.* 2001; Keele *et al.* 2009; Rudicell *et al.*  
11 187 2011). These included D18s536 (also termed locus A), D4s243 (locus B), D10s676 (locus  
12 188 C), D9s922 (locus D), D2s1326 (locus 1) D2s1333 (locus 2), D4s1627 (locus 3), and  
13 189 D9s905 (locus 4) (Table S1). To facilitate MiSeq sequencing of the amplified loci, we  
14  
15 190 added MiSeq-specific adapters to the 5' end of both the forward (5'-  
16 191 TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG-3') and the reverse primer (5'-  
17 192 GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG-3'), respectively. Individual STR  
18 193 loci were amplified using 3 - 5 µl fecal DNA extract, 2.5 µl 10x AmpliTaq Gold Buffer, 1.75  
19 194 µl 25mM MgCl<sub>2</sub>, 1.5 µl 10 mM dNTPs, 0.5 µl 50 µg/ml BSA, 1.5 µl of 10 mM forward and  
20 195 reverse primers, and 0.25 µl AmpliTaq Gold polymerase (5U/ml; Applied Biosystems) in a  
21 196 25 µl reaction volume. Thermocycling was performed using an initial denaturation for 10  
22 197 minutes at 94 °C, followed by 50 cycles of 30 seconds at 94 °C, 30 seconds at 54 °C, and  
23 198 45 seconds at 72 °C, followed by a final extension of 10 minutes at 72 °C.

24  
25  
26  
27  
28  
29  
30  
31  
32 199 Testing the sensitivity of MiSeq derived allele detection, we found that individual  
33 200 PCR reactions often produced only partial genotypes, while the combination of multiple  
34 201 amplicons from the same DNA sample generally yielded a more complete set of alleles.  
35 202 Consistent with previous studies (Morin *et al.* 2001), Each fecal DNA sample was  
36 203 amplified on three different occasions for each STR locus, with the resulting products  
37 204 pooled at equal proportions prior to MiSeq sequencing. we also found that PCR  
38 205 amplification of less than 25 pg of host DNA generally failed to amplify STR loci. For all  
39 206 genotyping analyses, we thus included only DNA samples that contained more than 25 pg  
40 207 of chimpanzee DNA, amplified each STR locus on three independent occasions, and  
41 208 combined equal volumes of these replicate PCR reactions prior to MiSeq sequencing.

42  
43  
44  
45  
46  
47  
48  
49 209 The eight STR loci were also amplified in one-step and two-step multiplex  
50 210 reactions. To minimize primer-primer interactions, locus A, B, C and 3 primers were

1  
2  
3  
4  
5  
6  
7  
8 211 combined at an even ratio in one pool, while locus D, 1, 2, 4 primers were similarly  
9  
10 212 combined in a second pool. Fecal DNA was then amplified in two (rather than eight)  
11  
12 213 different reactions, using the identical cycling conditions as for singleplex PCR. For two-  
13  
14 214 step multiplexing, 2 µl of a 1:100 dilution of the one-step product were used as a template  
15  
16 215 for a second round of PCR in which each locus was amplified individually using the same  
17  
18 216 thermocycling conditions (Arandjelovic *et al.* 2009).  
19

#### 20 218 *Library preparation and MiSeq sequencing*

21  
22 219  
23  
24 220 Following STR locus amplification, PCR products (individual or pooled) were diluted in  
25  
26 221 nuclease-free sterile water (1:10) and subjected to two rounds of PCR to add Illumina  
27  
28 222 barcodes and enrich for properly indexed DNA products as described (Iyer *et al.* 2017).  
29  
30 223 The resulting libraries were pooled, purified with Ampure Beads (Beckman Coulter),  
31  
32 224 quantified using a Qubit Fluorometer (Thermo Scientific) and TapeStation 2200 (Agilent),  
33  
34 225 and diluted to a final DNA concentration of 4 nM (Iyer *et al.* 2017). A randomly fragmented  
35  
36 226 (adapter ligated) control library of PhiX DNA (Illumina) was added to increase read length  
37  
38 227 diversity to ensure cluster recognition on the flow-cell. Both PhiX control and STR  
39  
40 228 amplicon libraries were adjusted to a final DNA concentration of 12 pM and mixed 1:1  
41  
42 229 prior to loading onto the sequencing reagent cartridge. All STR amplicons were  
43  
44 230 sequenced in [one direction](#) using v2 chemistry (500 cycle kits) without fragmentation. This  
45  
46 231 increased the length of the STR loci that could be analyzed to ~400 bp (instead of 2 x 250  
47  
48 232 paired-end reads). Although 500 cycles [are](#) the theoretical maximum of the sequencing  
49  
50 233 kit, we observed diminishing data quality between 350-400 cycles. We thus selected 375  
51  
52 234 forward and 51 reverse read cycles, using only the forward reads for analysis to preclude  
53  
54 235 alignment artifacts of pairing reads in the repeat regions (the reverse reads were only  
55  
56  
57  
58  
59  
60

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

236 used for MiSeq quality control). To maximize the number of amplicons sequenced per run,  
237 we used dual index multiplexing of samples.

238

239 *CHIIMP analysis pipeline*

240

241 Following MiSeq sequencing, read files were processed using standard methods. First,  
242 sample demultiplexing and FASTQ file generation was performed using the Illumina  
243 MiSeq Reporter software with default settings. Next, MiSeq adapter sequences were  
244 trimmed using cutadapt (Martin 2011). The adapter trimmed forward reads from each read  
245 pair, which covered the entire STR amplicon, were then imported into the R package,  
246 which was used for all downstream analyses.

247 The CHIIMP analysis pipeline generates multi-locus genotypes in three stages.

248 First, each MiSeq sequence file is processed into unique sequences with relevant  
249 attributes, such as read counts, sequence length, and whether the sequence matches the  
250 locus-specific forward primer, repeat motifs and length range. Sequences are also queried  
251 for potential PCR artifacts, such as single nucleotide substitutions, indels, and stutter  
252 sequences introduced by Taq polymerase and sequencing errors. These artifacts are  
253 identified as comprising less than one third of the read counts of the corresponding allele.

254 The 33% threshold was selected because inspection of known heterozygous loci revealed  
255 that all of the true second most frequent alleles contained more than that proportion of  
256 reads. Finally, for each sample and locus the proportion of sequence reads of the total  
257 read count is determined. At this stage, data are kept for all loci to ensure flexible  
258 downstream processing, such as detecting cross-locus contamination.

259 The second stage removes all sequences that do not match the locus attributes,  
260 such as the forward primer, repeat motif, and locus length, and/or contain likely PCR  
261 artifacts. In addition, only sequences comprising a minimum fraction of the total number of

1  
2  
3  
4  
5  
6  
7  
8 262 filtered reads (5%) are retained, and only loci with a total filtered read count above a  
9  
10 263 customizable per-sample read threshold (>500) are genotyped. Application of these filters  
11  
12 264 determines the sample zygosity; if only one sequence passes these filters, the locus is  
13  
14 265 reported as homozygous. However, if two or more sequences pass the filters, the two  
15  
16 266 most abundant are kept and the sample is reported as heterozygous. The output at this  
17  
18 267 stage includes a spreadsheet with the sequence content, read counts, sequence lengths,  
19  
20 268 as well as other relevant information such as whether the sequence contains the correct  
21  
22 269 repeat motif or was identified as a likely stutter sequence or other PCR artifact. Of note, all  
23  
24 270 filters and thresholds are customizable, with the above parameters representing the  
25  
26 271 default.

27 272 In the final stage, genotypes are assembled for all samples and loci, with quality  
28  
29 273 control tables generated as output files (Fig. S1). First, a summary genotype table is  
30  
31 274 generated that lists sample designations for each row, STR loci for each column, and  
32  
33 275 unique allele identifiers for each cell (Fig. S1a). If specific allele codes are provided, the  
34  
35 276 summary table will include these designations. If an allele does not match previous  
36  
37 277 identifiers, the software will create a short name based on sequence length and content to  
38  
39 278 identify these new alleles (e.g., sample 4781, locus C, allele 2 in Fig. S1a). The similarity  
40  
41 279 of genotypes is also depicted in a heatmap (Fig. 1b), which groups closely related  
42  
43 280 genotypes (Peakall & Smouse 2006). In cases where genotypes of individuals are known,  
44  
45 281 the program links samples with the corresponding individuals (Fig. S1c). A heatmap  
46  
47 282 shows the extent of similarity of every sample with every known genotype, thus allowing  
48  
49 283 simple individual identification (Fig. S1d). The program also generates a set of tables that  
50  
51 284 flag alleles that require additional attention, such as loci where the stutter filter has been  
52  
53 285 invoked, where more than two sequences passed the filter, where a large proportion of  
54  
55 286 sequences was not contained in the identified alleles, and where homozygosity may  
56  
57 287 reflect allelic dropout (Fig. S1e). For each locus, the program creates a FASTA file of all

1  
2  
3  
4  
5  
6  
7  
8 288 allele sequences and an image of their alignment (Fig. S1f) generated by the  
9  
10 289 Bioconductor's MSA package (Bodenhofer *et al.* 2015). In addition, a heatmap of  
11  
12 290 sequence counts that match the locus-specific forward primer for all samples and loci is  
13  
14 291 generated (Fig. S1g). For singleplex samples, this identifies sequences that match other  
15  
16 292 loci and thus highlights potential cross-locus contamination. For multiplexed samples, this  
17  
18 293 shows the read distribution across [different](#) loci. Finally, histograms that show sequence  
19  
20 294 length-frequency distributions are saved as image files (Fig. S1h). A summary file is  
21  
22 295 created that combines all key results (sequences, read counts, etc.) for alleles for all  
23  
24 296 samples and loci. This data output file is suitable for further analysis in R.

25  
26 297 The new analysis platform, termed *Computational High-throughput Individual*  
27  
28 298 *Identification through Microsatellite Profiling* or CHIIMP, has been designed to allow  
29  
30 299 customization of the number and sequence content of microsatellite loci to be analyzed.  
31  
32 300 Particular locus attributes such as the expected locus length range, primer sequences,  
33  
34 301 and repeat motif sequence can be specified in a simple text file. Thus, the software can be  
35  
36 302 readily adapted to additional microsatellite loci, as long as the respective amplicons fall  
37  
38 303 within the length limits of the particular sequence chemistry used. The software is also  
39  
40 304 suitable to analyze multiplexed samples, which contain reads from several loci but are  
41  
42 305 processed separately, again using the forward primers to select locus-specific reads. No  
43  
44 306 additional software is required other than providing a list of samples and loci prior to  
45  
46 307 analysis. CHIIMP is available at <https://github.com/ShawHahnLab/chimp> and can be  
47  
48 308 installed on any Windows, Mac OS, or Linux computer with a standard installation of R  
49  
50 309 and RStudio in a single step. On Windows, a desktop shortcut to the analysis script is  
51  
52 310 provided. Dragging a simple text file containing analysis options onto the shortcut triggers  
53  
54 311 analysis with the selected options. In addition to the standalone program, all features can  
55  
56 312 also be used individually from within R. [Examples-A comprehensive user guide including](#)  
57  
58 313 [examples](#) of analysis options and locus attributes [is](#) provided with the software.  
59  
60



1  
2  
3  
4  
5  
6  
7  
8 3149  
10 315 ~~Statistical, Error, diversity, and heterozygosity calculations~~11  
12 31613 317 ~~Statistical analyses of c-myc concentration and allele detection were performed using the~~14  
15 318 ~~Mann-Whitney test and Prism version 5 software (GraphPad).~~ Error rates for the MiSeq16  
17 319 derived genotypes were calculated by determining the number of allelic mismatches for18  
19 320 each sample to the known genotype of the corresponding chimpanzee (including allelic20  
21 321 dropout, stutter sequences, PCR/sequencing artifacts, and locus amplification failure) and22  
23 322 by dividing the total number of alleles by the number of erroneous alleles (Broquet and24  
25 323 Petit, 2004). The expected heterozygosity (also termed gene diversity) for the sampled26  
27 324 Gombe and GME chimpanzees was calculated from both capillary electrophoresis and28  
29 325 MiSeq based microsatellite data as described in Charlesworth & Charlesworth 2010.30  
31 326 Allelic diversity was calculated by summing the total number of unique alleles in a32  
33 327 population.34  
35 32836  
37 329 **Results**38  
39 330 ~~Development of a MiSeq-based STR genotyping approach for wild chimpanzees~~40  
41 33142  
43 332 ~~To take advantage of available microsatellite and kinship data from Gombe chimpanzees,~~44  
45 333 ~~we selected fecal samples from 24 individuals, who were previously genotyped by~~46  
47 334 ~~capillary electrophoresis on multiple occasions (Keele et al. 2009; Rudicell et al. 2010;~~48  
49 335 ~~Santiago et al. 2003). Their consensus genotype at eight STR loci served as the~~50  
51 336 ~~benchmark to which all MiSeq derived data were compared (Table S2). STR loci were~~52  
53 337 ~~amplified as in the past using the same PCR conditions, except for primers containing~~54  
55 338 ~~MiSeq adapters rather than fluorescent labels. To avoid read alignment artifacts across~~56  
57 339 ~~the repeat motif regions, amplification products were not fragmented and sequenced~~

1  
2  
3  
4  
5  
6  
7  
8 340 ~~using only the forward reads. This allowed us to utilize STR amplicons of up to 400 bp,~~  
9  
10 341 ~~which included all previously characterized STR loci (Table S1).~~

11 342 ~~Following MiSeq sequencing, read counts of identical sequences were tallied, and~~  
12 343 ~~entries that had fewer than 500 reads or fell outside the expected locus range were~~  
13 344 ~~excluded (Fig. 1a). Despite these filters, we found that a large number of loci (27%) still~~  
14 345 ~~yielded more than two prominent sequences (Fig. S2). For example, locus 2 of sample 10~~  
15 346 ~~yielded 6 prominent read peaks within the expected size range (grey area in Fig. 1a), only~~  
16 347 ~~two of which were of the correct length based on the reference genotypes (red bars in Fig.~~  
17 348 ~~1a; Table S2). Two other peaks differed from these two by four nucleotides each (green~~  
18 349 ~~bars in Fig. 1a), suggesting that they represented stutter peaks, while the remaining two~~  
19 350 ~~were distinct from all canonical locus 2 alleles (blue bars in Fig. 1a, blue cells in Table~~  
20 351 ~~S2). Since these peaks had lower read counts than the true alleles, we tested whether~~  
21 352 ~~reporting only the most frequent sequences for each locus would yield the correct~~  
22 353 ~~genotype. While this was the case for some loci (Table S2), a large number of alleles still~~  
23 354 ~~failed to match the reference (blue and green fields in Fig. 1b and Table S2). An alignment~~  
24 355 ~~of their sequences with those of true alleles revealed the absence of locus-specific repeat~~  
25 356 ~~motifs, indicating that they represented off-target amplifications (indicated by asterisks in~~  
26 357 ~~Fig. 1c). Of a total of 384 alleles, 16 (4%) represented such PCR artefacts (blue fields in~~  
27 358 ~~Fig. 1b), while an additional 21 (5%) represented likely stutter peaks (green fields in Fig.~~  
28 359 ~~1b).~~

29 360 ~~To eliminate incorrect alleles, we added additional filtering steps. First, we required~~  
30 361 ~~that sequences had to contain at least three of the locus-specific repeat motifs. Second,~~  
31 362 ~~sequences that were 4 bp shorter than other sequences had to comprise at least 33% of~~  
32 363 ~~the number of reads of the longer fragment to be counted as a true allele. This threshold~~  
33 364 ~~was selected because inspection of known heterozygous loci revealed that all of the true~~  
34 365 ~~second most frequent alleles that were 4 bp shorter contained more than that proportion~~

1  
2  
3  
4  
5  
6  
7  
8 366 ~~of reads (Fig. S3). To exclude Taq polymerase and sequencing errors, the same read~~  
9  
10 367 ~~count requirements were also imposed for alleles of identical lengths that differed in their~~  
11  
12 368 ~~nucleotide sequence or were one base pair shorter or longer. Implementation of these~~  
13  
14 369 ~~filters removed all erroneous alleles, except for a single stutter allele that fell just above~~  
15  
16 370 ~~the required threshold (Fig. 1d f, Fig. S3). The only other discrepancies from the reference~~  
17  
18 371 ~~genotype were amplification failures of either one (allelic drop out) or both alleles (orange~~  
19  
20 372 ~~and grey fields in Fig. 1f, respectively). Thus, comparison to the benchmark genotypes~~  
21  
22 373 ~~allowed us to iteratively improve the filtering process, such that off target amplicons, PCR~~  
23  
24 374 ~~and sequencing errors, and almost all stutter sequences were removed.~~

#### 375 376 *Increasing the sensitivity of STR allele detection*

377  
378 ~~We next tested whether combining amplification products from replicate PCR reactions~~  
379 ~~prior to MiSeq sequencing would increase the sensitivity of allele detection. Using the~~  
380 ~~same 24 fecal samples, we amplified STR loci from two additional aliquots of the same~~  
381 ~~DNA, pooled all three replicates, and then sequenced the products individually and as a~~  
382 ~~pool. The results showed that individual replicates often produced only partial genotypes,~~  
383 ~~while combining all three PCR reactions generated a full set of alleles (Fig. 2a). This was~~  
384 ~~also true for the remainder of the 24 fecal samples, where the pooling of replicates prior to~~  
385 ~~sequencing consistently yielded more complete genotypes than the average of individual~~  
386 ~~replicates (Fig. 2b). Replicate pooling reduced allele detection failures by 41% and allelic~~  
387 ~~dropouts by 43%, and yielded at least one correct allele for 36% of all loci that failed to~~  
388 ~~amplify in single replicates (Fig. 2b).~~

389 ~~We also asked whether the fecal samples that yielded only partial genotypes had~~  
390 ~~low chimpanzee DNA concentrations. Using a previously described *c-myc* based qPCR~~  
391 ~~assay (Morin *et al.* 2001), we found that the 24 chimpanzee samples differed markedly in~~

1  
2  
3  
4  
5  
6  
7  
8 392 ~~their host DNA concentrations, with some containing no detectable chimpanzee DNA~~  
9  
10 393 ~~(<2.5 pg/ul) (Fig. S4a). Consistent with previous results (Morin et al. 2001), PCR~~  
11 394 ~~amplification of less than 25 pg of host DNA generally failed to amplify STR loci (Fig.~~  
12  
13 395 ~~S4b), even when amplicons from three independent reactions were pooled. When all data~~  
14  
15 396 ~~from the 24 chimpanzee samples were compared, we found that not surprisingly, 67% of~~  
16  
17 397 ~~missing alleles were from PCR reactions that contained less than 25 pg of host DNA~~  
18 398 ~~(samples 2, 5, 9, and 10 in Fig. 1e). After removing these samples, only 4% of loci failed~~  
19  
20 399 ~~to amplify and only 3% of loci exhibited allelic dropout. For all subsequent genotyping~~  
21  
22 400 ~~analyses, we thus pooled the products of three PCR replicates (each containing more~~  
23  
24 401 ~~than 25 pg of chimpanzee DNA) prior to MiSeq sequencing.~~

402

#### 403 *Direct comparison of MiSeq and capillary electrophoresis based STR genotyping*

404

405 To compare the performance of MiSeq and capillary electrophoresis side-by-side, we  
406 selected samples from 19 Gombe chimpanzees, who were previously genotyped by  
407 capillary electrophoresis on multiple occasions (Keele et al. 2009; Rudicell et al. 2010;  
408 Santiago et al. 2003). Testing more recently collected fecal samples that had not yet been  
409 genotyped, we used the consensus genotype of previous genotypes at eight STR loci as  
410 the benchmark to which all MiSeq derived data were compared (new-Table 1). Fecal DNA  
411 was extracted, confirmed to contain more than 25 pg of sufficient amounts of chimpanzee  
412 DNA per PCR aliquot, and amplified using the same STR primers and conditions, except  
413 for the presence of MiSeq adapters versus fluorescent labels. For MiSeq sequencing,  
414 three PCR replicates were pooled, while only a single replicate was analyzed by capillary  
415 electrophoresis using both automated and manual peak calling options. The latter was  
416 done because capillary electrophoresis analysis of pooled samples is compromised when  
417 allele peaks differ in relative height in independent PCR reactions.

1  
2  
3  
4  
5  
6  
7  
8 418 | \_\_\_\_\_ Using the consensus genotype of the corresponding chimpanzees for reference  
9  
10 419 | (Table S31), we found that MiSeq genotyping reduced the number of allelic dropouts by  
11  
12 420 | more than half (Table 24). This was due, at least in part, to the pooling of PCR replicates,  
13  
14 421 | ~~which also eliminated amplification failure~~ which increased the number of alleles that were  
15  
16 422 | ~~detected, resulting in detection of at least one allele for each locus~~. However, MiSeq  
17  
18 423 | genotyping was also more accurate than the traditional method, which could not  
19  
20 424 | differentiate off-target amplifications (Tables 1 and S32). In addition, stutter peaks were  
21  
22 425 | completely eliminated by the CHIIMP analysis pipeline, which was not the case for the  
23  
24 426 | automated capillary electrophoresis method. Although manual peak calling also eliminated  
25  
26 427 | stutter peaks, this was considerably more time consuming than the MiSeq approach. For  
27  
28 428 | the 19 samples, conventional peak calling and allele binning took two hours, while  
29  
30 429 | reviewing the bioinformatics outputs took ~~approximately 5~~ minutes. Most importantly,  
31  
32 430 | MiSeq genotyping identified eight heterozygous loci that were scored as homozygous by  
33  
34 431 | capillary electrophoresis because of a failure to resolve minor sequence and length (1bp)  
35  
36 432 | differences (Fig. 13). These sequence variants were readily identified in the read  
37  
38 433 | histograms (Fig. 1a) and their frequency identified in sequence alignments of the entire  
39  
40 434 | locus (Fig. 1b and c). Inspection of allele lengths across all loci revealed that 24% of all  
41  
42 435 | MiSeq derived alleles did not differ by multiples of four, indicating frequent nucleotide  
43  
44 436 | insertions and deletions in the tetranucleotide repeats (Fig. 1b and c).

437

438 *MiSeq genotyping uncovers increased allelic diversity and heterozygosity*

439

440 To examine the true extent of allelic diversity in Gombe, we selected fecal samples from  
441 123 chimpanzees, which included all currently living adults and juveniles, except for  
442 offspring born within the past three years, as well as 38 deceased individuals. All of these  
443 were previously genotyped by capillary electrophoresis on at least three occasions.

1  
2  
3  
4  
5  
6  
7  
8 444 Subjecting one representative fecal sample to MiSeq analysis, we confirmed 51 known  
9  
10 445 alleles, but also detected 31 new alleles, which had previously gone unrecognized due to  
11  
12 446 ~~size homoplasy~~ ~~minor size (1 ntbp)~~ or nucleotide sequence differences (Tables [32](#) and  
13  
14 447 [S24](#)). Such cryptic alleles were detected for all eight STR loci, increasing allelic diversity  
15  
16 448 by an average of 1.6 fold per locus. Nearly half of all previously reported alleles had  
17  
18 449 closely related length or sequence variants (Table [43](#)).

18 450 Although the great majority of the newly identified alleles were found in multiple  
19  
20 451 individuals, we wanted to validate their authenticity by demonstrating inheritance. Since  
21  
22 452 paternity and kinship relationships are known for most Gombe chimpanzees, we were  
23  
24 453 able to trace the majority of the newly identified allelic variants from parents to their  
25  
26 454 offspring. For example, Locus 3 includes four alleles that are identical in size (234 bp) but  
27  
28 455 differ by up to three substitutions and two single nucleotide insertions and deletions (Fig.  
29  
30 456 [254a](#)). Alleles 234-a, 234-b, 234-c, and 234-d were found in 80, 25, 10 and 4  
31  
32 457 chimpanzees, respectively, including several parent-offspring triads (Fig. [254b](#)). Overall,  
33  
34 458 we were able to document inheritance for 25 (81%) of the 31 new alleles. For the  
35  
36 459 remaining 6 existing pedigree information was insufficient, and their existence was thus  
37  
38 460 confirmed by sequencing at least two independent PCR amplicons (Table [43](#)).

37 461 The newly identified alleles revealed that over a quarter of genotypes at loci previously  
38  
39 462 assigned as homozygous (60 of a total of 228) were in fact heterozygous (Table [S24](#)).  
40  
41 463 This increased allelic diversity resolved one case of an ambiguous paternity  
42  
43 464 determination. Using the standard eight STR loci, we were previously unable to identify  
44  
45 465 the father of one infant (Google) because two candidate males (Faustino and Londo) had  
46  
47 466 the identical genotype at all eight STR loci (Walker *et al.* 2017). Using the new genotypes,  
48  
49 467 we were able to exclude Londo and confirm Faustino as a father by revealing differences  
50  
51 468 at one locus (Fig. [524a and c](#)). Although Faustino was identified as the correct father at  
52  
53 469 the time by genotyping 10 additional loci using capillary electrophoresis (Walker *et al.*

1  
2  
3  
4  
5  
6  
7  
8 470 2017), this would not have been necessary had the increased allelic diversity been known.  
9  
10 471 Thus, MiSeq genotyping revealed much greater allelic and microsatellite gene diversity in  
11 472 Gombe than previously appreciated, thus increasing the analytical potential of the existing  
12  
13 473 STR loci.  
14

15 474

#### 16 475 *MiSeq genotyping based individual identification*

17  
18 476

19  
20 477 Since chimpanzee communities are often studied longitudinally, we added an individual  
21 478 identification tool to the analysis platform. This tool compares the genotype of every new  
22 479 sample with all previously characterized genotypes and generates a distance score to  
23 480 indicate their relative similarity. For example, samples with a distance score of 0 match at  
24 481 all loci, while samples with a distance score of 2 differ by two alleles. We then used this  
25 482 approach to characterize the [same](#) 19 newly genotyped samples (Table [1S3](#)) as well as 5  
26 483 samples from infants with unknown genotypes. To account for allelic dropout, a distance  
27 484 score of up to 3 was allowed. The results revealed accurate individual identification for all  
28 485 samples from previously genotyped chimpanzees. Of the 19 samples, 8 exhibited a  
29 486 perfect match across all loci (Fig. [35a](#)), while 11 others had distance scores of 1-3, which  
30 487 were consistent with allelic dropout (Fig. [35b](#)). However, 5 samples with distance scores  
31 488 of 5-7 could not be assigned to known individuals (Fig. [35c](#)), and a review of field notes  
32 489 revealed that they were all collected from new infants. A heatmap allowed the quick  
33 490 identification of very close (4821, 4807) and very distant (4566) matches (Fig. [5d3d](#)).  
34 491 Thus, the individual identification tool detected previously determined genotypes with  
35 492 reasonable accuracy.  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48

49 493

#### 49 494 *STR genotyping of multiplexed samples*

50  
51 495  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3  
4  
5  
6  
7  
8 496 Chimpanzees in the Greater Mahale Ecosystem in Tanzania [occupy a large home range](#),  
9  
10 497 [live at low population densities, and face extreme seasonal changes](#) (Moore 1996; Ogawa  
11 498 *et al.* 1999; Schoeninger *et al.* 1999). [Thus, these “savanna chimpanzees” live](#) under  
12 499 ecologically more challenging conditions than their forest-dwelling counterparts, and with  
13 500 the exception of the Issa community, are not habituated. As a result, fecal collections,  
14 501 sample transport and storage are logistically more difficult, which can result in reduced  
15 502 amounts of collected material and/or partially degraded host DNA. To test the suitability of  
16 503 MiSeq genotyping for such samples, we selected 12 previously characterized chimpanzee  
17 504 fecal specimens from the Issa Valley (Rudicell *et al.* 2011) and re-genotyped them using  
18 505 both singleplex and multiplex locus amplification. Singleplex PCR was performed as in  
19 506 Gombe, while multiplex PCR was carried out in two steps as previously described  
20 507 (Arandjelovic *et al.* 2009). First, PCR primers for 4 loci were pooled and used to amplify  
21 508 fecal DNA in two (rather than eight) reactions (one-step multiplex product). Second,  
22 509 aliquots of this first round PCR were then used in a second round of PCR to amplify each  
23 510 of the 8 STR loci separately (two-step multiplex product). Three pooled replicates of both  
24 511 one-step and two-step multiplexed products were sequenced and compared to the  
25 512 previously determined genotypes (Table S35). Although the overall amplification efficiency  
26 513 was lower than originally reported (most likely due to repeat freezing and thawing of the [7-](#)  
27 514 [8 year old](#) samples), one-step multiplexing performed as well as singleplex PCR, but used  
28 515 only a quarter of the fecal DNA (Table 54). Two-step multiplexing detected slightly more  
29 516 alleles, but not surprisingly, also resulted in an increased number of stutter sequences  
30 517 and other PCR artifacts. Thus, one-step multiplexing required less starting material and  
31 518 was also more cost efficient because the combined loci were sequenced in a single MiSeq  
32 519 run (and were subsequently de-multiplexed bioinformatically).

33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49 520 MiSeq genotyping also allowed us to compare the allelic diversity in Gombe and  
50  
51 521 the GME. Fig. 46 depicts such an analysis for locus B and D, highlighting alleles that were  
52  
53



1  
2  
3  
4  
5  
6  
7  
8 522 only found in GME chimpanzees. Comparing all eight STR loci, we found ten alleles in  
9  
10 523 only 12 GME chimpanzees that were absent from the 123 genotyped Gombe individuals,  
11  
12 524 six of which represented alleles previously missed in the GME due to sequence and  
13  
14 525 length differences. Although the mean expected heterozygosity value for the GME  
15  
16 526 chimpanzees (0.743) was lower than that for Gombe (0.812), this is likely due to the small  
17  
18 527 sample size and the fact that all 12 individuals were sampled at a single location in Issa  
19  
20 528 Valley (Rudicell *et al.* 2011). Additional samples from more diverse locations in the GME  
21  
22 529 are needed to compare the genetic diversity of this population to that of Gombe and other  
23  
24 530 field sites.

531

## 532 Discussion

533

534 Over the past two decades, microsatellite analyses have been an integral part of studies  
535 of wild chimpanzees, providing insight into their evolution, population genetics, behavior,  
536 disease association and social structure (Barbian *et al.* 2018; Becquet *et al.* 2007; Keele  
537 *et al.* 2009; Langergraber *et al.* 2007; Moeller *et al.* 2016; Rudicell *et al.* 2010; Santiago *et*  
538 *al.* 2003; Vigilant *et al.* 2001; Walker *et al.* 2017; Wroblewski *et al.* 2015). However,  
539 ~~current~~ traditional genotyping methods are cumbersome, imprecise and  
540 investigator/platform dependent, due to the use of capillary electrophoresis to determine  
541 the length of STR loci. Here, we report a high-throughput MiSeq-based approach, which  
542 represents a marked improvement, because it is faster, more accurate and able to detect  
543 the full extent of allelic diversity in a population. Moreover, it includes a new analysis  
544 platform, CHIIMP, which not only automates the conversion of raw MiSeq data into multi-  
545 locus genotypes, but also implements a number of quality control measures that improve  
546 genotyping accuracy (Fig. 75). Of note, CHIIMP has been designed for maximal  
547 adaptability and customization. While our analysis of pedigreed chimpanzee fecal samples

1  
2  
3  
4  
5  
6  
7  
8 548 ~~from chimpanzees and genotypes allowed rigorous validation, the analysis pipeline is not~~  
9 ~~limited to a particular species or sample type.~~  
10

11 550

12  
13 551 *Improved accuracy of MiSeq based genotyping*

14  
15 552

16 553 Sequence-based genotyping methods not only determine the length of STR loci, but also  
17 554 reveal their sequence content, and thus have the potential to detect a greater number of  
18 555 distinct alleles than capillary electrophoresis. Indeed, such genotyping of Atlantic cod and  
19 556 muskrats revealed high proportions of cryptic alleles, ranging from 32% to 44% (Darby *et al.*  
20 557 *2016*; Vartia *et al.* 2016). In light of these data, our discovery of 38% new alleles (31 of  
21 558 82) in Gombe is not surprising (Table 23). However, this finding suggests that existing  
22 559 STR data vastly underestimate the diversity of microsatellite sequences in wild  
23 560 chimpanzees, not only in Gombe but also in other populations. New alleles were found for  
24 561 all loci, with some comprising twice as many variants as previously observed (Table 32),  
25 562 which will undoubtedly add to the statistical power of future analyses. However, any new  
26 563 allele will have to be examined carefully by repeat amplification and sequencing, unless it  
27 564 can be validated by pedigree analysis. In our dataset, ~~8% of initially identified a minor~~  
28 565 ~~fraction of~~ “new” alleles were found to represent PCR and/or sequencing artifacts ~~that~~  
29 566 ~~exceeded the 33% threshold for heterozygous alleles, which prompted us to add a filter~~  
30 567 ~~that specifically removed such erroneous alleles. Since all PCR/sequencing artifacts~~  
31 568 ~~occurred in combination with an allele of the same length, they were easily~~  
32 569 ~~recognized~~ Repeat amplification of these alleles in question resolved all cases of  
33 570 ~~sequencing artifacts that were called as alleles.~~

34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47 571 Comparison of the MiSeq data to validated reference genotypes also allowed us to  
48 572 assess the error rate of the new approach. After implementation of all filters, CHIIMP  
49 573 eliminated 98% of stutter sequences and 100% of off-target amplicons. Among the  
50  
51  
52

1  
2  
3  
4  
5  
6  
7  
8 574 samples tested, true alleles, allelic dropouts and false alleles were detected with a  
9  
10 575 frequency of 96%, 7%, and 0%, respectively. These data are comparable to MiSeq  
11  
12 576 derived genotyping results for wild-living brown bears, where true alleles, allelic dropouts  
13  
14 577 and false alleles were detected with a frequency of 93%, 0.4% and 0.05% for tissues, and  
15  
16 578 80%, 14% and 1% for fecal samples, respectively (De Barba *et al.* 2017). Although our  
17  
18 579 overall error rate of 3.3% is slightly higher than the 2.1% error rate reported for a MiSeq  
19  
20 580 genotyping study of laboratory raised (pedigreed) fish (Zhan *et al.* 2017), this is not  
21  
22 581 surprising since the latter study examined freshly extracted tissue DNA.

23  
24 582 Since non-invasively collected samples frequently contain diluted and/or degraded  
25  
26 583 host DNA, they are genotyped using multiple PCR reactions to guard against the selective  
27  
28 584 loss of alleles (allelic dropout). Loci are only considered homozygous if they can be  
29  
30 585 confirmed in multiple PCR reactions (Morin *et al.* 2001; Taberlet *et al.* 1996). Capillary  
31  
32 586 electrophoresis requires that these replicates are run independently to distinguish true  
33  
34 587 alleles from non-specific signal, often more than tripling the amount of time and effort  
35  
36 588 required to genotype a single sample. In contrast, MiSeq genotyping can be performed  
37  
38 589 after combining the products of multiple PCR reactions (Fig. 2). Although the quality of the  
39  
40 590 input DNA remains the same, MiSeq genotyping of pooled PCR replicates reduces the  
41  
42 591 frequency of allelic dropout and thus renders the resulting genotypes more accurate.  
43  
44 592 ~~While however, amplicon pooling replicate amplifications can maximize data output, it~~  
45  
46 593 ~~does not result in the loss of the foregoing data from repeat analyses, which are~~  
47  
48 594 ~~used by some as a measure of DNA quality and/or data reliability (Taberlet *et al.* 1996)~~

49  
50 595 Once MiSeq data files are imported into the CHIIMP platform, the program calls  
51  
52 596 alleles automatically, thus saving days of hands-on work. While automated allele calling  
53  
54 597 has been reported previously (De Barba *et al.* 2017; Suez *et al.* 2016; Zhan *et al.* 2017),  
55  
56 598 CHIIMP includes downstream analyses, such as alignments of allele sequences or  
57  
58 599 flagging loci that may contain contaminants, which provide important additional quality  
59  
60

1  
2  
3  
4  
5  
6  
7  
8 600 | control measures. In contrast to previous studies, CHIIMP also retains non-repeat regions,  
9  
10 601 | [\(Suez et al. 2016\)](#), which can contribute to allelic diversity, and does not require the  
11  
12 602 | presence of stutter sequences for allele calling, which may not be sufficiently abundant  
13  
14 603 | under conditions of low coverage (De Barba et al. 2017; ~~Suez et al. 2016~~). Finally,  
15  
16 604 | CHIIMP reports both allele length and sequence content, and is thus designed to detect  
17  
18 605 | minor length and sequence differences by including sequence-specific allele names and  
19  
20 606 | generating locus-specific sequence alignments (Figs. [46](#) and S1). To guide subsequent  
21  
22 607 | analyses, we have also added features that flag potentially problematic alleles and  
23  
24 608 | standardize allele naming. CHIIMP thus represents the most comprehensive analysis  
25  
26 609 | platform yet to ensure the accuracy of MiSeq-based genotyping results.

610

#### 611 *Multiplexing improves MiSeq genotyping efficiency and reduces cost*

612

613 The Illumina MiSeq v2 500 sequencing kit has an output of ~25 million reads per run, thus  
614 allowing the multiplexing of many samples, the number of which depends on the desired  
615 read depth. Comparing read depths per allele, we found that a cut-off of 500 reads yielded  
616 | the most accurate results for our dataset. This value is higher than ~~a~~ [the 50 read](#) cut-off  
617 | used previously to genotype laboratory raised fish (Zhan et al. 2017). However, the latter  
618 | study used high quality tissues rather than fecal samples for analysis. To determine the  
619 | sources of allele-calling errors, we did not multiplex samples from Gombe chimpanzees.  
620 | However, we tested multiplexing using samples from ~~the~~ [GME chimpanzees](#) and  
621 | confirmed that this approach yields accurate results. Although primer incompatibilities  
622 | allowed the combination of only four loci, this number can be significantly increased with  
623 | additional primer design. For example, a recent study genotyped bear fecal DNA by  
624 | multiplexing 14 loci (De Barba et al. 2017). [Pooling amplicons from multiple loci after](#)  
625 | [singleplex PCR can circumvent the need for specialized primer design, as the maximum](#)

1  
2  
3  
4  
5  
6  
7  
8 626 | number of pooled loci for any given sample is limited only by the desired read depth.

9  
10 627 | Moreover, barcoding of individual samples allows their combination in sequencing  
11  
12 628 | reactions, thus further increasing sequencing efficiency and throughput (Farrell *et al.*  
13  
14 629 | 2016).

15 630 | MiSeq genotyping is expensive, but these costs decrease with sample numbers.  
16  
17 631 | Capillary electrophoresis is undoubtedly cheaper when only a small number of samples  
18  
19 632 | has to be analyzed; however, MiSeq sequencing becomes increasingly more cost-  
20  
21 633 | effective with multiplexing and analyses of pooled replicates (Darby *et al.* 2016). The costs  
22  
23 634 | of MiSeq genotyping three replicates of 96 samples multiplexed at four loci would roughly  
24  
25 635 | be equivalent to analyzing the same multiplexed samples via capillary electrophoresis,  
26  
27 636 | because the latter cannot analyze pooled replicates. While this estimate only considers  
28  
29 637 | genotyping supplies, labor to manually analyze samples is not included. In addition, the  
30  
31 638 | improved accuracy has downstream cost advantages since fewer repeat analyses would  
32  
33 639 | have to be performed.

34

34 641 | *Effective sharing of MiSeq genotyping data*

35  
36 642 |

37 643 | A direct comparison of MiSeq and capillary electrophoresis derived alleles revealed  
38  
39 644 | consistent length differences of one to three nucleotides, the number of which were locus  
40  
41 645 | specific (Tables 1 and S3S2 and S3). For Gombe samples, locus 3 alleles derived by  
42  
43 646 | capillary electrophoresis were always three nucleotides longer than the corresponding  
44  
45 647 | MiSeq alleles (Table 1s S2 and S3). However, for the GME samples, the same alleles  
46  
47 648 | were all one nucleotide shorter than the MiSeq alleles (Table S35). This is as expected  
48  
49 649 | since the capillary electrophoresis data were generated on different platforms. However,  
50  
51 650 | this also means that a simple conversion of existing capillary electrophoresis to MiSeq  
52  
53 651 | data will generally not be possible. In contrast, MiSeq genotyping generates unambiguous

1  
2  
3  
4  
5  
6  
7  
8 652 | alleles that can be compared across multiple studies and field sites (Fig. 64). In the future,  
9  
10 653 | it will thus be possible to compare STR genotypes across different chimpanzee  
11  
12 654 | populations, such as those in Gombe and the GME, since the use of different sequencing  
13  
14 655 | equipment will no longer confound these analyses.  
15

16 656

#### 17 657 *Versatility of the CHIIMP analysis platform*

18 658

19  
20 659 | To increase its utility, we designed the CHIIMP analysis platform to be versatile. STR  
21  
22 660 | locus attributes, such as the expected length range, primer sequences, and repeat motifs,  
23  
24 661 | as well as all thresholds for allele calling can be customized. For example, analysis of loci  
25  
26 662 | with dinucleotide repeats may require a lower threshold for stutter peaks, since these are  
27  
28 663 | more susceptible to polymerase slippage (Guichoux *et al.* 2011). Similarly, locus length  
29  
30 664 | ranges can be expanded or contracted, depending on the rate of off-target amplicons. The  
31  
32 665 | CHIIMP analysis pipeline also includes tools that facilitate iterative improvements for new  
33  
34 666 | applications (Fig. S1). For example, the program provides a heatmap that indicates the  
35  
36 667 | number of unique sequences that pass all filters. If that number is too high, thresholds can  
37  
38 668 | be adjusted to remove stutter peaks, off-target amplicons, and/or PCR errors. In addition,  
39  
40 669 | the distribution of loci is visualized, which can be used to reveal contamination in  
41  
42 670 | singleplexed samples or identify poorly performing primers in multiplexed samples (Fig.  
43  
44 671 | S1g). For potentially problematic alleles, CHIIMP generates histograms that provide  
45  
46 672 | information concerning their length and relative abundance. All of these tools can be used  
47  
48 673 | to adapt the platform to additional STR loci and/or host species.

49  
50 674 | The length of STR loci suitable for sequence-based genotyping depends on the  
51  
52 675 | sequencing chemistry. We used Illumina v2 technology, which has maximum read lengths  
53  
54 676 | of 500 nucleotides. MiSeq sequences are most often generated using paired-end reads,  
55  
56 677 | with a maximum read length of 250 nucleotides in each direction. For STR genotyping,  
57  
58  
59  
60

1  
2  
3  
4  
5  
6  
7  
8 678 locus sequences must span the repeat motif region, since assembly of shorter reads  
9  
10 679 could result in misalignments. To accommodate loci of greater than 250 bp length, we  
11  
12 680 opted to only use forward reads for analysis. Although the sequencing kit could  
13  
14 681 theoretically accommodate fragments of up to 500 nucleotides, we found that the quality  
15  
16 682 of reads (Q scores) decreased significantly after 400 cycles. Given that the longest locus  
17  
18 683 in our panel spanned 357 nucleotides, we used 375 cycles in the forward direction.  
19  
20 684 Illumina v3 sequencing chemistry has a 600-cycle limit, which may accommodate loci of  
21  
22 685 up to 500 bp, but this would have to be determined experimentally. The majority of  
23  
24 686 microsatellite loci are shorter than this length.

25  
26 687 As STR genotyping transitions from capillary electrophoresis to sequence based  
27  
28 688 approaches, it will be necessary to standardize allele nomenclature, as has already been  
29  
30 689 suggested for human forensics (Gelardi *et al.* 2014; Parson *et al.* 2016). At a minimum,  
31  
32 690 allele names will have to incorporate the length and unique sequence content for each  
33  
34 691 allele (Darby *et al.* 2016). In our study, we added an alphabetical identifier (-a, -b, -c, etc.)  
35  
36 692 to differentiate identically sized alleles that differed in their sequence ([Table S2](#)). Since it  
37  
38 693 is impossible to capture all allele attributes in a single name, it may become necessary to  
39  
40 694 establish databases that link allele identifiers to their respective sequences. CHIIMP is  
41  
42 695 designed to allow users to supply a spreadsheet of allele names and sequences, and thus  
43  
44 696 guarantees consistent nomenclature across experiments. As MiSeq genotyping is  
45  
46 697 adapted to additional projects, standardized allele designations will become necessary to  
47  
48 698 ensure consistent nomenclature across studies.

49 699

## 50 700 *Conclusions*

51 701

52 702 Genetic study of wild primates and other endangered species has been shown to provide  
53  
54 703 more accurate information concerning the size, structure, distribution and dynamics of

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

704 populations than observational studies. However, genotyping can be prohibitively  
705 expensive given the large numbers of samples that are required for such analyses. The  
706 MiSeq based genotyping platform provides a new approach that drastically reduces time  
707 and labor, while providing more accurate and informative genotypes compared to capillary  
708 electrophoresis. This will allow much faster and more streamlined analysis of samples that  
709 are necessary for censusing and monitoring of non-habituated populations in addition to  
710 revealing previously inaccessible allelic diversity. The CHIIMP platform has been  
711 designed to be adaptable to additional loci and/or species. This allows the study of group  
712 membership, dispersal, gene flow, and association patterns for a multitude of wildlife  
713 species with broad conservation and biological implications.

714

#### 715 *Author contributions*

716

717 All authors contributed to the acquisition, analysis, and interpretation of the data. H.J.B.,  
718 A.J.C. and B.H.H. conceived, planned and executed the study; H.J.B., A.N.A., R.M.R.,  
719 M.S.G. and Y.L. performed STR locus amplifications and data analyses; H.J.B. and A.J.C.  
720 developed the CHIIMP analysis pipeline; A.G.S., A.L.S., and F.B.R. optimized the MiSeq  
721 sequencing approach; D.M., E.V.L, F.A.S., A.K.P., and A.E.P. conducted or supervised  
722 field work; A.J.C., E.E.W, and P.M.S. performed allelic diversity and parentage analyses;  
723 H.J.B., A.J.C., R.M.R. and B.H.H coordinated the contributions of all authors and wrote  
724 the manuscript.

725

#### 726 *Acknowledgements*

727

728 We thank the Jane Goodall Institute field staff at the Gombe Stream Research Centre as  
729 well as field assistants from the Greater Mahale Ecosystem Research and Conservation



730 Project (GMERC) for collecting chimpanzee observational data as well as fecal samples;  
 731 the Tanzania Commission for Science and Technology (COSTECH), the Tanzania Wildlife  
 732 Research Institute (TAWIRI), and the Tanzania National Parks Association (TANAPA) for  
 733 their support and permission to conduct research in Gombe and the GME. This work was  
 734 supported by grants from the National Institutes of Health, USA (R01 AI 091595, R37 AI  
 735 050529, R01 AI 120810, P30 AI 045008), the National Science Foundation (IOS-1052693,  
 736 IOS-1457260), the Jane Goodall Institute, and the University of California at San  
 737 Diego/Salk Center for Academic Research and Training in Anthropogeny (CARTA). H.J.B.  
 738 and R.M.R. were funded by training grants (T32 AI 055400 and T32 AI 007632,  
 739 respectively). The authors declare no competing financial interests.

740

741 *Data accessibility*

742

743 STR sequences are archived in the NCBI Sequence Read Archive (SRA) under  
 744 <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA434411>. Preprocessed sequence data,  
 745 analysis software and supporting R code are archived on Dryad doi: #####. Ongoing  
 746 Software Development, including supporting R code, is available on  
 747 <https://github.com/ShawHahnLab/chiimp/releases/tag/0.1.0>  
 748 <https://github.com/ShawHahnLab/chiimp>.

749

750 *References*

751

752 Adams RI, Brown KM, Hamilton MB (2004) The impact of microsatellite electromorph size  
 753 homoplasy on multilocus population structure estimates in a tropical tree  
 754 (Corythophora alta) and an anadromous fish (Morone saxatilis). *Molecular Ecology*  
 755 **13**, 2579-2588.

Formatted: Justified, Line spacing: Double

Formatted: Font: (Default) +Body (Cambria), Underline color: Auto, Font color: Auto

Formatted: Font: 11 pt

Formatted: Font: (Default) +Body (Cambria), Font color: Auto

Field Code Changed

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

- 756 Arandjelovic M, Guschanski K, Schubert G, *et al.* (2009) Two-step multiplex polymerase  
757 chain reaction improves the speed and accuracy of genotyping using DNA from  
758 noninvasive and museum samples. *Molecular Ecology Resources* **9**, 28-36.
- 759 Arandjelovic M, Vigilant L (2018) Non-invasive genetic censusing and monitoring of  
760 primate populations. *American Journal of Primatology*, e22743.
- 761 Balloux F, Lugon-Moulin N (2002) The estimation of population differentiation with  
762 microsatellite markers. *Molecular Ecology* **11**, 155-165.
- 763 Barbian HJ, Li Y, Ramirez M, *et al.* (2018) Destabilization of the gut microbiome marks the  
764 end-stage of simian immunodeficiency virus infection in wild chimpanzees.  
765 *American Journal of Primatology* **80**, 10.1002/ajp.22515.
- 766 Becquet C, Patterson N, Stone AC, Przeworski M, Reich D (2007) Genetic structure of  
767 chimpanzee populations. *PLOS Genetics* **3**, e66.
- 768 Bennett P (2000) Demystified ....: Microsatellites. *Molecular Pathology* **53**, 177-183.
- 769 Bodenhofer U, Bonatesta E, Horejš-Kainrath C, Hochreiter S (2015) msa: an R package  
770 for multiple sequence alignment. *Bioinformatics* **31**, 3997-3999.
- 771 Broquet T, Petit E (2004) Quantifying genotyping errors in noninvasive population  
772 genetics. *Molecular Ecology* **13**, 3601-3608.
- 773 Charlesworth B, Charlesworth D (2010) *Elements of evolutionary genetics*. Roberts and  
774 Co. Publishers. 734 p.
- 775 Constable JJ, Packer C, Collins DA, Pusey AE (1995) Nuclear DNA from primate dung.  
776 *Nature* **373**, 393.
- 777 Constable JL, Ashley MV, Goodall J, Pusey AE (2001) Noninvasive paternity assignment  
778 in Gombe chimpanzees. *Molecular Ecology* **10**, 1279-1300.

- 1  
2  
3  
4  
5  
6  
7  
8 779 Darby BJ, Erickson SF, Hervey SD, Ellis-Felege SN (2016) Digital fragment analysis of  
9  
10 780 short tandem repeats by high-throughput amplicon sequencing. *Ecology and*  
11  
12 781 *Evolution* **6**, 4502-4512.
- 13 782 De Barba M, Miquel C, Lobréaux S, *et al.* (2017) High-throughput microsatellite  
14  
15 783 genotyping in ecology: improved accuracy, efficiency, standardization and success  
16  
17 784 with low-quantity and degraded DNA. *Molecular Ecology Resources*. **17**, 492-507.
- 18 785 Ellegren H (2004) Microsatellites: simple sequences with complex evolution. *Nature*  
19  
20 786 *Reviews Genetics* **5**, 435.
- 21  
22 787 Ewen KR, Bahlo M, Treloar SA, *et al.* (2000) Identification and analysis of error types in  
23  
24 788 high-throughput genotyping. *American Journal of Human Genetics* **67**, 727-736.
- 25 789 Farrell ED, Carlsson JEL, Carlsson J (2016) Next Gen Pop Gen: implementing a high-  
26  
27 790 throughput approach to population genetics in boarfish (*Capros aper*). *Royal*  
28  
29 791 *Society Open Science* **3**, 160651.
- 30 792 Fernando P, Evans BJ, Morales JC, Melnick DJ (2001) Electrophoresis artefacts — a  
31  
32 793 previously unrecognized cause of error in microsatellite analysis. *Molecular*  
33  
34 794 *Ecology Notes* **1**, 325-328.
- 35 795 Fordyce SL, Ávila-Arcos MC, Rockenbauer E, *et al.* (2011) High-throughput sequencing of  
36  
37 796 core STR loci for forensic genetic investigations using the Roche Genome  
38  
39 797 Sequencer FLX platform. *BioTechniques* **51**, 127-133.
- 40  
41 798 Gelardi C, Rockenbauer E, Dalsgaard S, Børsting C, Morling N (2014) Second generation  
42  
43 799 sequencing of three STRs D3S1358, D12S391 and D21S11 in Danes and a new  
44  
45 800 nomenclature for sequenced STR alleles. *Forensic Science International: Genetics*  
46  
47 801 **12**, 38-41.
- 48 802 Guichoux E, Lagache L, Wagner S, *et al.* (2011) Current trends in microsatellite  
49  
50 803 genotyping. *Molecular Ecology Resources* **11**, 591-611.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

- 804 Hauge XY, Litt M (1993) A study of the origin of 'shadow bands' seen when typing  
805 dinucleotide repeat polymorphisms by the PCR. *Human Molecular Genetics*. **2**,  
806 411-415.
- 807 Iyer SS, Bibollet-Ruche F, Sherrill-Mix S, *et al.* (2017) Resistance to type 1 interferons is a  
808 major determinant of HIV-1 transmission fitness. *Proceedings of the National  
809 Academy of Sciences* **114**: E590-E599.
- 810 Jarne P, Lagoda P (1996) Microsatellites, from molecules to populations and back.  
811 *Trends in Ecology & Evolution* **11**, 424-429.
- 812 Keele BF, Jones JH, Terio KA, *et al.* (2009) Increased mortality and AIDS-like  
813 immunopathology in wild chimpanzees infected with SIVcpz. *Nature* **460**, 515-519.
- 814 Kelkar YD, Strubczewski N, Hile SE, *et al.* (2010) What is a microsatellite: a computational  
815 and experimental definition based upon repeat mutational behavior at A/T and  
816 GT/AC repeats. *Genome Biology and Evolution* **2**, 620-635.
- 817 Langergraber KE, Mitani JC, Vigilant L (2007) The limited impact of kinship on  
818 cooperation in wild chimpanzees. *Proceedings of the National Academy of  
819 Sciences* **104**, 7786-7790.
- 820 Levinson G, Gutman GA (1987) Slipped-strand mispairing: a major mechanism for DNA  
821 sequence evolution. *Molecular biology and evolution*. **4**, 203-221.
- 822 Lynch HT, de la Chapelle A (2003) Hereditary Colorectal Cancer. *New England Journal of  
823 Medicine* **348**, 919-932.
- 824 Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing  
825 reads. *EMBnet journal* **17**, 10-12.
- 826 Moeller AH, Foerster S, Wilson M, *et al.* (2016) Social behaviour promotes diversity in the  
827 chimpanzee gut microbiome. *American Association for the Advancement of  
828 Science* **2**, e1500997.

- 1  
2  
3  
4  
5  
6  
7  
8 829 Moore J (1996) Savanna chimpanzees, referential models and the last common ancestor.  
9  
10 830 *Great Ape Societies*. Cambridge University Press, 275-292.
- 11 831 Morin PA, Chambers KE, Boesch C, Vigilant L (2001) Quantitative polymerase chain  
12 832 reaction analysis of DNA from noninvasive samples for accurate microsatellite  
13 833 genotyping of wild chimpanzees (*Pan troglodytes verus*). *Molecular Ecology* **10**,  
14 834 1835-1844.
- 15 835 Morin PA, Wallis J, Moore JJ, Chakraborty R, Woodruff DS (1993) Non-invasive sampling  
16 836 and DNA amplification for paternity exclusion, community structure, and  
17 837 phylogeography in wild chimpanzees. *Primates* **34**, 347-356.
- 18 838 Ogawa H, Idani Gi, Kanamori M (1999) Chimpanzee habitat in the savanna woodland,  
19 839 Ugalla, Tanzania. *Primate Research* **15**, 135-146.
- 20 840 Parson W, Ballard D, Budowle B, *et al.* (2016) Massively parallel sequencing of forensic  
21 841 STRs: Considerations of the DNA commission of the International Society for  
22 842 Forensic Genetics (ISFG) on minimal nomenclature requirements. *Forensic*  
23 843 *Science International: Genetics* **22**, 54-63.
- 24 844 Pasqualotto AC, Denning DW, Anderson MJ (2007) A cautionary tale: lack of consistency  
25 845 in allele sizes between two laboratories for a published multilocus microsatellite  
26 846 typing system. *Journal of Clinical Microbiology* **45**, 522-528.
- 27 847 Peakall ROD, Smouse PE (2006) genalex 6: genetic analysis in Excel. Population genetic  
28 848 software for teaching and research. *Molecular Ecology Notes* **6**, 288-295.
- 29 849 Pusey AE, Pintea L, Wilson ML, Kamenya S, Goodall J (2007) The contribution of long-  
30 850 term research at Gombe National Park to chimpanzee conservation. *Conservation*  
31 851 *Biology* **21**, 623-634.
- 32 852 Queller DC, Strassmann JE, Hughes CR (1993) Microsatellites and kinship. *Trends in*  
33 853 *Ecology & Evolution* **8**, 285-288.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

- 854 Rudicell RS, Holland Jones J, Wroblewski EE, *et al.* (2010) Impact of simian  
855 immunodeficiency virus infection on chimpanzee population dynamics. *PLoS*  
856 *Pathogens* **6**, e1001116.
- 857 Rudicell RS, Piel AK, Stewart F, *et al.* (2011) High prevalence of simian immunodeficiency  
858 virus infection in a community of savanna chimpanzees. *Journal of Virology* **85**,  
859 9918-9928.
- 860 Santiago ML, Lukasik M, Kamenya S, *et al.* (2003) Foci of endemic simian  
861 immunodeficiency virus infection in wild-living eastern chimpanzees (*Pan*  
862 *troglodytes schweinfurthii*). *Journal of Virology* **77**, 7545-7562.
- 863 Schoeninger MJ, Moore J, Sept JM (1999) Subsistence strategies of two "savanna"  
864 chimpanzee populations: The stable isotope evidence. *American Journal of*  
865 *Primatology* **49**, 297-314.
- 866 Schuelke M (2000) An economic method for the fluorescent labeling of PCR fragments.  
867 *Nature Biotechnology* **18**, 233.
- 868 Selkoe KA, Toonen RJ (2006) Microsatellites for ecologists: a practical guide to using and  
869 evaluating microsatellite markers. *Ecology Letters* **9**, 615-629.
- 870 Shinde D, Lai Y, Sun F, Arnheim N (2003) Taq DNA polymerase slippage mutation rates  
871 measured by PCR and quasi-likelihood analysis: (CA/GT)(n) and (A/T)(n)  
872 microsatellites. *Nucleic Acids Research* **31**, 974-980.
- 873 Suez M, Behdenna A, Brouillet S, *et al.* (2016) MicNeSs: genotyping microsatellite loci  
874 from a collection of (NGS) reads. *Molecular Ecology Resources* **16**, 524-533.
- 875 Taberlet P, Camarra JJ, Griffin S, *et al.* (1997) Noninvasive genetic tracking of the  
876 endangered Pyrenean brown bear population. *Molecular Ecology* **6**, 869-876.
- 877 Taberlet P, Griffin S, Goossens B, *et al.* (1996) Reliable genotyping of samples with very  
878 low DNA quantities using PCR. *Nucleic Acids Research* **24**, 3189-3194.

- 1  
2  
3  
4  
5  
6  
7  
8 879 van Lawick-Goodall J (1968) The behaviour of free-living chimpanzees in the Gombe  
9 Stream Reserve. *Animal Behaviour Monographs* **1**, 161-311.  
10 880  
11 881 Van Neste C, Van Nieuwerburgh F, Van Hoofstat D, Deforce D (2012) Forensic STR  
12 analysis using massive parallel sequencing. *Forensic Science International*. **6**,  
13 882 810-818.  
14 883  
15 884 Vartia S, Villanueva-Cañas JL, Finarelli J, *et al.* (2016) A novel method of microsatellite  
16 genotyping-by-sequencing using individual combinatorial barcoding. *Royal Society*  
17 *Open Science* **3**, 150565.  
18 885  
19 886  
20 887 Vigilant L, Hofreiter M, Siedel H, Boesch C (2001) Paternity and relatedness in wild  
21 chimpanzee communities. *Proceedings of the National Academy of Sciences* **98**,  
22 888 12890-12895.  
23 889  
24 890 Walker KK, Rudicell RS, Li Y, *et al.* (2017) Chimpanzees breed with genetically dissimilar  
25 mates. *Royal Society Open Science* **4**, 160422.  
26 891  
27 892 Weeks DE, Conley YP, Ferrell RE, Mah TS, Gorin MB (2002) A tale of two genotypes:  
28 consistency between two high-throughput genotyping centers. *Genome Res* **12**,  
29 893 430-435.  
30 894  
31 895 Wroblewski EE, Norman PJ, Guethlein LA, *et al.* (2015) Signature patterns of MHC  
32 diversity in three gombe communities of wild chimpanzees reflect fitness in  
33 reproduction and immune defense against SIVcpz. *PLoS Biology* **13**, e1002144.  
34 896  
35 897  
36 898 Zhan L, Paterson IG, Fraser BA, *et al.* (2017) megasat: automated inference of  
37 microsatellite genotypes from sequence data. *Molecular Ecology Resources* **17**,  
38 899 247-256.  
39 900  
40 901  
41 902  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

902 **Figure legends**

1  
2  
3  
4  
5  
6  
7  
8 904 **Fig 13** MiSeq genotyping uncovers cryptic alleles. Eight polymorphic STR loci were  
9  
10 905 amplified from the fecal DNA of 19 previously genotyped chimpanzees. (a) Histogram  
11 906 depicting the length (x-axis) and read count (y-axis) of unique sequences for one  
12 907 representative heterozygous locus that was previously determined to be homozygous by  
13 908 multiple capillary electrophoresis analyses (sample 4861, locus C, Table 1). The grey box  
14 909 highlights the expected locus size range. The horizontal line indicates the cutoff of 500  
15 910 reads. Colored peaks indicate reads that passed the locus-specific filters (note that peaks  
16 911 can be comprised of identically sized reads that differ in their sequence content). Black  
17 912 reads were eliminated. Pink reads appear to be locus-specific, but did not pass the PCR  
18 913 artifact filters. Red reads represent the true allele sequences (180 and 181 bp in lengths,  
19 914 respectively). (b, c) Alignment images of locus-specific allele sequences are shown for  
20 915 locus 1 (ab) and locus C (bc), respectively (the complete data set is shown in Table 1S3).  
21 916 Allele sequences are ordered by length (indicated in bp on the right), with the frequency  
22 917 with which they were found in different chimpanzees indicated on the left (the x-axis  
23 918 indicates the position within the alignment). Nucleotides are colored as shown, with gaps  
24 919 in the alignment shown in grey. The insets highlight alleles that differ in their sequence  
25 920 content and/or length. Nucleotide substitutions are colored; dashes indicate gaps that  
26 921 were introduced to optimize the alignment.

27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39 922  
40 923 **Fig. 24** MiSeq genotyping uncovers increased allelic diversity and heterozygosity. (a)  
41 924 Alignment of four locus 3 alleles that are of identical length (234 bp), but differ in  
42 925 sequence content. Nucleotide substitutions are colored; dashes indicate single nucleotide  
43 926 insertions and deletions (b) Mendelian inheritance of allele 234 for a group of related  
44 927 chimpanzees. Fathers and mothers are shown as squares and circles, respectively, with  
45 928 offspring connected by vertical lines. Both alleles are shown for each animal, with the four  
46 929 allelic variants highlighted in different colors. Individuals of unknown identity or genotype



1  
2  
3  
4  
5  
6  
7  
8 930 are left blank. (c) Increased allelic diversity resolves a previously ambiguous paternity  
9  
10 931 determination. Two potential fathers with identical allele lengths (238 bp) can now be  
11  
12 932 distinguished based on differences in allele sequence content (238-a and 238-b). Since  
13  
14 933 the offspring is homozygous for allele 238-a, the male with allele 238-b can be excluded  
15  
16 934 as a father.

17 935

18 936 | **Fig. 35** Individual identification based on MiSeq genotyping. (a-c) Genotypes of newly  
19  
20 937 collected samples (top) are compared to the genotypes of known community members,  
21  
22 938 with the closest match listed below (based on descending distance scores). Genotypes  
23  
24 939 that differ by fewer than four alleles are indicated in bold because they represent likely  
25  
26 940 matches. Differences are highlighted in yellow. (d) Heatmap showing the relative similarity  
27  
28 941 of sample genotypes (rows) with genotypes of known individuals (columns) based on  
29  
30 942 distance scores. Dark red cells indicate likely matches.

31 943

32 944 | **Fig 46** Comparison of MiSeq genotypes across chimpanzee communities. Alignment  
33  
34 945 images of locus-specific allele sequences are shown for chimpanzees from the GME and  
35  
36 946 Gombe. Two representative loci (locus B on the left; locus D on the right) are shown for  
37  
38 947 (a) 12 chimpanzees from the GME (Table S35), (b) 123 chimpanzees from Gombe (Table  
39  
40 948 S24), and (c) a combination of both. Allele sequences are ordered by length (indicated in  
41  
42 949 base pairs on the right), with the frequency with which they were found in different  
43  
44 950 chimpanzees indicated on the left (the x-axis indicates the position within the alignment).  
45  
46 951 Nucleotides are colored as indicated, with alignment gaps shown in grey. Arrows indicate  
47  
48 952 alleles that are unique to the GME samples.

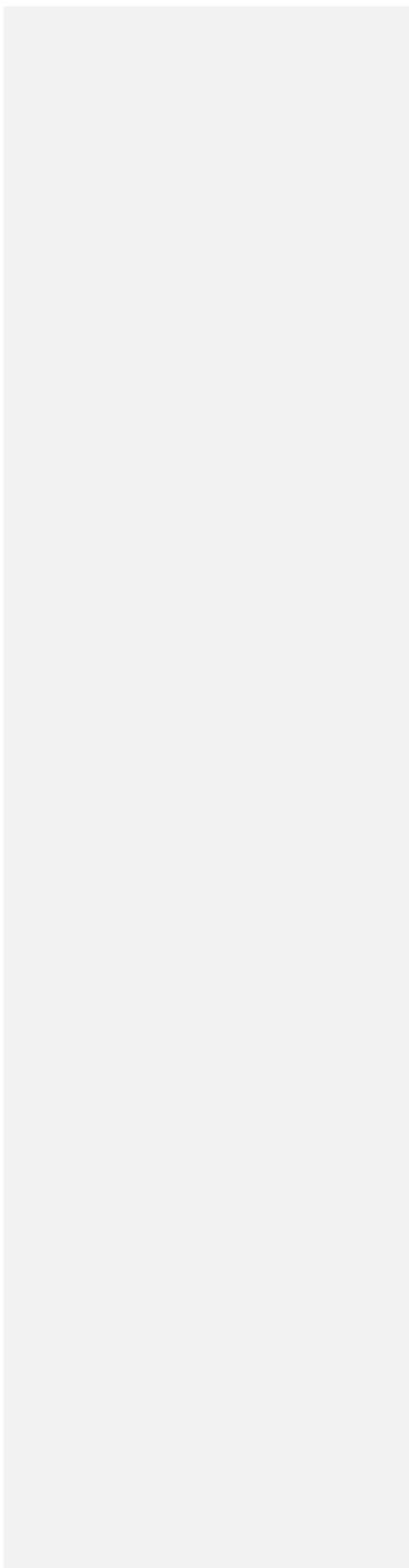
49 953

49 954 | **Fig. 57** MiSeq-based STR genotyping of wild chimpanzees. (a) Schematic representation  
50  
51 955 of singleplex STR amplification and MiSeq sequencing of chimpanzee fecal DNA. (b)

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

956 Schematic representation of the CHIIMPS analysis pipeline with decision tree and  
957 downstream data reports.  
958

For Review Only



## Response to Reviewers:

### Reviewer: 1

#### Comments to the Author

The authors test a high-throughput sequencing approach to genotype microsatellites from chimpanzees, and also present their software package that automates this allele-calling algorithm. The general genotyping-by-sequencing approach for microsatellites has already been tested and shown (as the authors note), but there is still value in seeing additional applications in different species and tissue types (especially the fecal samples used here, which are notoriously difficult to genotype). However, the authors make a few over confident claims that might not be supported. For example, line 109, that "none of these [previous] studies have compared the performance of capillary electrophoresis," is likely not entirely true. It's possible they made these tests, without presenting them. (Although, Fig. 2 from Darby et al. seems to make this direct comparison).

**None of the published genotyping papers have compared the performance of capillary electrophoresis and high throughput sequencing directly to validate the accuracy of their approach. If those tests were made, but not presented, we cannot know about them. Figure 2 from Darby et al. does not make this comparison.**

Also, their claim that the analysis time for MiSeq genotyping is only 5 minutes is probably a bit of an underestimate. I don't disagree that this method probably saves some time over allele calling for capillary electrophoresis, but it would certainly take much longer than 5 minutes just to assemble the necessary files to run CHIIMP, let alone the time (probably hours) that would be necessary to check the results and proof any discrepancies.

**The exact analyses included in the time estimate (allele length calling, binning, and individual identification) are specified in the legend to Table 2. Nonetheless, we have revised our statement.**

Finally, I appreciate that this algorithm is an advance of some of the previous attempts at genotyping microsatellites by sequencing, but it's still not entirely "fully" automated, but rather semi-automated, and I would think still requires some manual checking to be fully confident.

**We have changed the title from "a fully automated... platform" to "an automated... platform."**

Additionally, the software program is rather complex to implement and I'm not convinced it would be an easily adaptable solution for someone else working with this type of data (or at least not necessarily less work than it might take to develop one's own algorithm).

**Our program design makes very few assumptions with respect to species, tissue type, or sequencing strategy, and leaves many relevant analysis choices up to the user without requiring modification of the source code. We describe these choices in detail in a supplied user guide. Our latest improvements to the program have also allowed for more flexibility in the data input and analysis output. We have added a statement to clarify these points in the text.**

1  
2  
3 Reviewer: 2  
4

5 In this ms, Barbian et al. describe a HTS approach and bioinformatics pipeline for microsatellite  
6 genotyping and demonstrate its application to genotype scat samples from wild chimpanzees.  
7 HTS microsatellite genotyping has recently come into the scene and holds great potential for  
8 wildlife studies. I believe continued development and testing is important at these early stages  
9 to favor wide applicability of the method, both in terms of laboratory and sequencing protocols  
10 and of bioinformatics analysis that need to be implemented to allow wider routine utilization.

11 In this perspective, the main contributions of this study are:

- 12 - the development of a pipeline that is flexible and that presents useful features for evaluating  
13 and reporting the data, such as tools to visualize alleles and non-alleles sequences, genotype  
14 similarity, match to known genotypes for individual ID, flag alleles and loci to be re-checked and  
15 possible contamination, allele alignment, etc. This is all very important and need to be  
16 incorporated into pipelines for routine applications of HTS STR genotyping methods also by  
17 non-bioinformaticians. This is the main contribution of this work in my view.  
18 - testing for pooling replicates and the potential benefits of doing that (although authors failed to  
19 fully justify its application, see later comments)  
20 - revealing the amount and impact of hidden allelic diversity on gene diversity estimation  
21 - implementing standardized allele codes

22 However, the study is poorly presented, is too long, and is very hard to follow and track things  
23 done to their respective results. In particular:

- 24 - much of what is written is already known or has already been pointed out and addressed in  
25 other recent studies. Should focus more on the novelty and additional contribution of their  
26 approach, this will also shorten the paper  
27 - the paper is not well structured. Much of what is currently in the Results should be in the  
28 Methods. They have done several things performing different sets of experiment with different  
29 sample sets, but it is very difficult to evaluate what was actually done because it's all lost  
30 between methods and results.

31 Therefore, in my opinion the ms can be considered for publication only after major restructuring  
32 and re-focusing.  
33  
34  
35

36 **We agree with the reviewer and have shortened and restructured our manuscript in**  
37 **accordance to his/her recommendations. Specifically, we have deleted the first two**  
38 **sections of the Results, which removed five figures and two tables and now focuses the**  
39 **manuscript on what is new and unique to this study.**  
40

41 Specific comments:

42  
43 L107-111: this it not accurate, as Suez et al. and Vartia et al. also compared to some extent to  
44 capillary electrophoresis, and Suez et al., Zhan et al., De Barba et al. developed automated  
45 pipelines.  
46

47 **While other studies have included capillary electrophoresis data, none have used these**  
48 **results to “to validate and improve the genotyping approach” as done in our manuscript.**  
49 **Also, the statement that studies of wild animals have largely used manual genotyping**  
50 **methods is correct. Automated pipelines (using mostly model species) including the**  
51 **studies listed by the reviewer are reviewed in the Discussion.**  
52

53  
54 L294: statistical analysis and tests for what? need to be explicit. Maybe I missed it, but I did not  
55 see reference to statistical tests in the results  
56  
57  
58  
59  
60

1  
2  
3 **The supplementary figure that required statistical analysis has been removed.**  
4

5 L296-297: need to explain better how error rates were calculated, giving the actual formula or  
6 referring to previously published methods that has become the standard in STR genotyping. As  
7 described it is not completely clear how estimates were derived.  
8

9 **This has now been explained and cited.**  
10

11 L300: but you also report allelic diversity in this study  
12

13 **Allelic diversity calculations are now defined in the methods.**  
14

15 L306: this section should be combined or followed directly the “CHIMP analysis pipeline” section  
16 as the pipeline was implemented within the platform  
17

18 **This has been done.**  
19

20  
21 Results: As currently written the Results are a mixture of methods and results. All  
22 methodological description, should be moved to the Methods section, including information on  
23 sample sizes, method refinement with the inclusion of additional filtering steps, optimization  
24 steps that lead to the final protocol, etc..., and all this need to be described in a clearer way.  
25 Results should then be reported following the order of the methods section for ease of reading.  
26

27 **This has been done.**  
28

29 L380-381: again not clear how these % were estimated  
30

31 **This is now explained in the methods.**  
32

33 L403-405: This is not a direct comparison of CE with the Miseq performance. In order to be a  
34 direct comparison, replicates should not have been pooled for Miseq. I suggest rewording this  
35 part  
36

37 **This has been rephrased.**  
38

39 L410: cannot say that. Pooling of PCR replicates does not eliminate amplification failure of the  
40 single PCR amplifications; it is simply a result of the additive signal from each independent  
41 replicate. As written, this is misleading.  
42

43 **This has been rephrased to “increased of the number of alleles that were detected”**  
44 **instead of “eliminated amplification failure”.**  
45

46 L569-570: as before, I think this has to be rephrased or elaborated further. I agree that the  
47 ability of pooling replicates is an advantage, but the cumulative signal obtained by pooling  
48 comes at the expense of losing information from independent replicates, which is used for  
49 determining genotype reliability (through repeatability). The cumulative signal in the pooling  
50 derives from the sum of all read counts of all replicates, therefore allowing more easily to pass  
51 thresholds of allele detection and resulting in reduced allelic dropout, but read counts may  
52 come, as an extreme example, from only one replicate that over-amplified compared to others.  
53 In addition without replicates we lose the information about DNA quality. In the example in Fig.  
54 2, all replicates present missing loci, normally I would discard that sample for low DNA  
55  
56  
57  
58  
59  
60

1  
2  
3 quality/quantity, but you still get a full genotype by pooling. This full genotype however, could  
4 still present errors that may remain undetected (for example ADO at the 3 homozygous loci). I  
5 think there is value in pooling, but it's application for reliable genotyping requires further testing.  
6

7 **We agree that pooling of triplicate amplifications could result in losing information from**  
8 **independent replicates. We have now included this caveat into the discussion.**  
9

10 L576-581: Also other pipelines retain non-repeat regions, and report allele length and sequence  
11 and assign allele names...  
12

13 **No published pipeline currently contains all of the features included in CHIIMP. To clarify**  
14 **this in the text, we are now including the respective citation immediately following each**  
15 **statement.**  
16

17 Minor comments:

18 L148: use only the abbreviation since you have written it out already  
19  
20

21 **This has been changed.**  
22

23 L150: could be useful to specify the approximate age of samples, i.e. if only fresh samples or  
24 also older samples were collected.  
25

26 **This has been added.**  
27

28 L196: do you mean equimolar or equivolume proportions?  
29

30 **This has been clarified.**  
31

32 L224-226: move to results  
33

34 **This has been done.**  
35

36 L285: no need for a separate section for this  
37  
38

39 **This has been incorporated into the previous section.**  
40

41 L304: add "in" before Charlesworth and Charlesworth 2010  
42

43 **This has been corrected.**  
44

45 L483: clarify a bit "ecologically challenging conditions"  
46  
47

48 **This has been done.**  
49

50 L524: I would replace "current" with "traditional" as the transition is already happening  
51

52 **This has been done.**  
53

54 L602: add De Barba et al. 2017  
55

56 **Samples were not barcoded in De Barba et al. 2017 so the citation is not appropriate.**  
57  
58  
59  
60

1  
2  
3  
4 L662: this was also pointed out already in previous wildlife HTS STR studies  
5

6 **While this was briefly mentioned in another article, we believe it is highly relevant to this**  
7 **study and therefore worth inclusion in the discussion.**  
8  
9

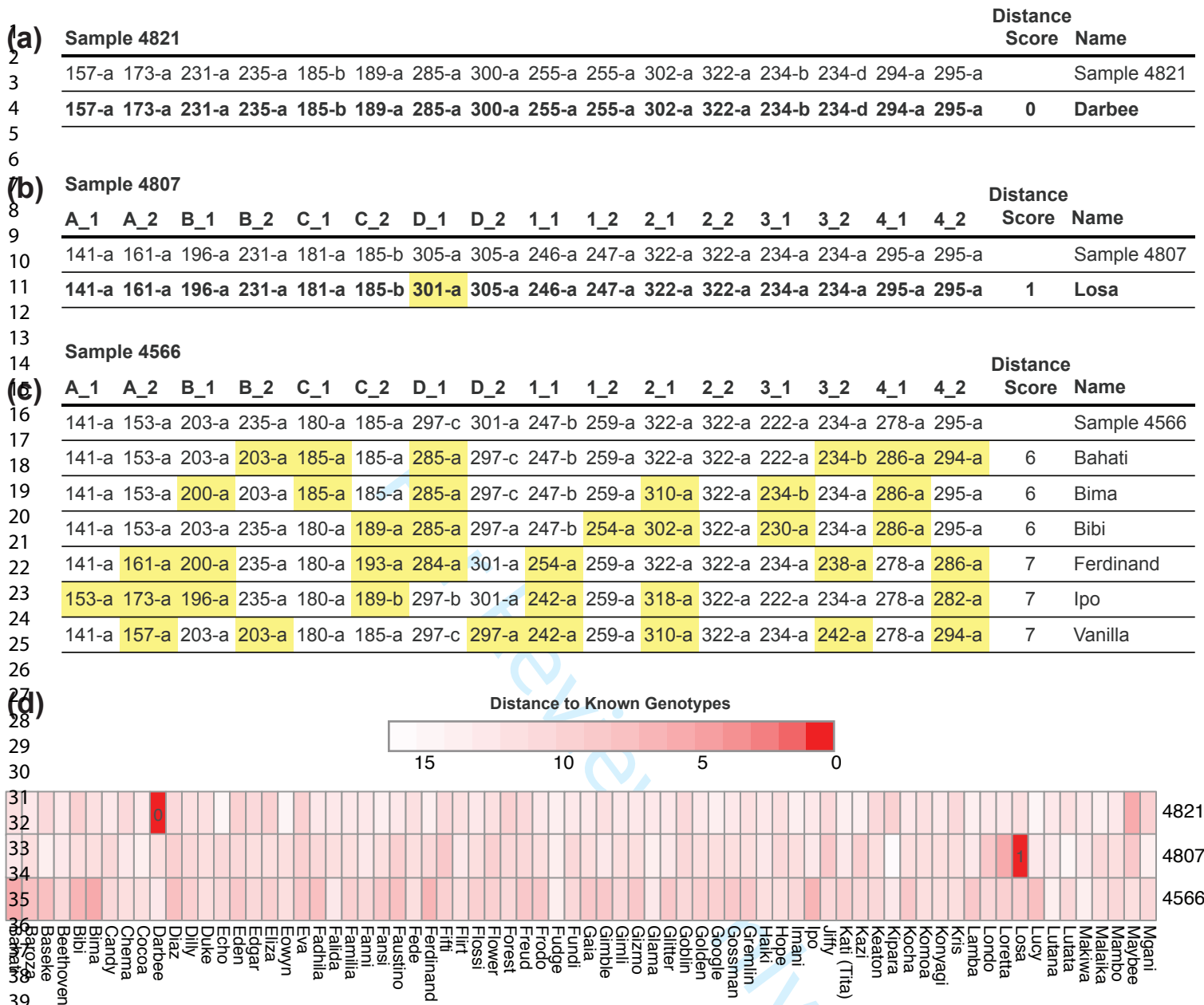
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

For Review Only

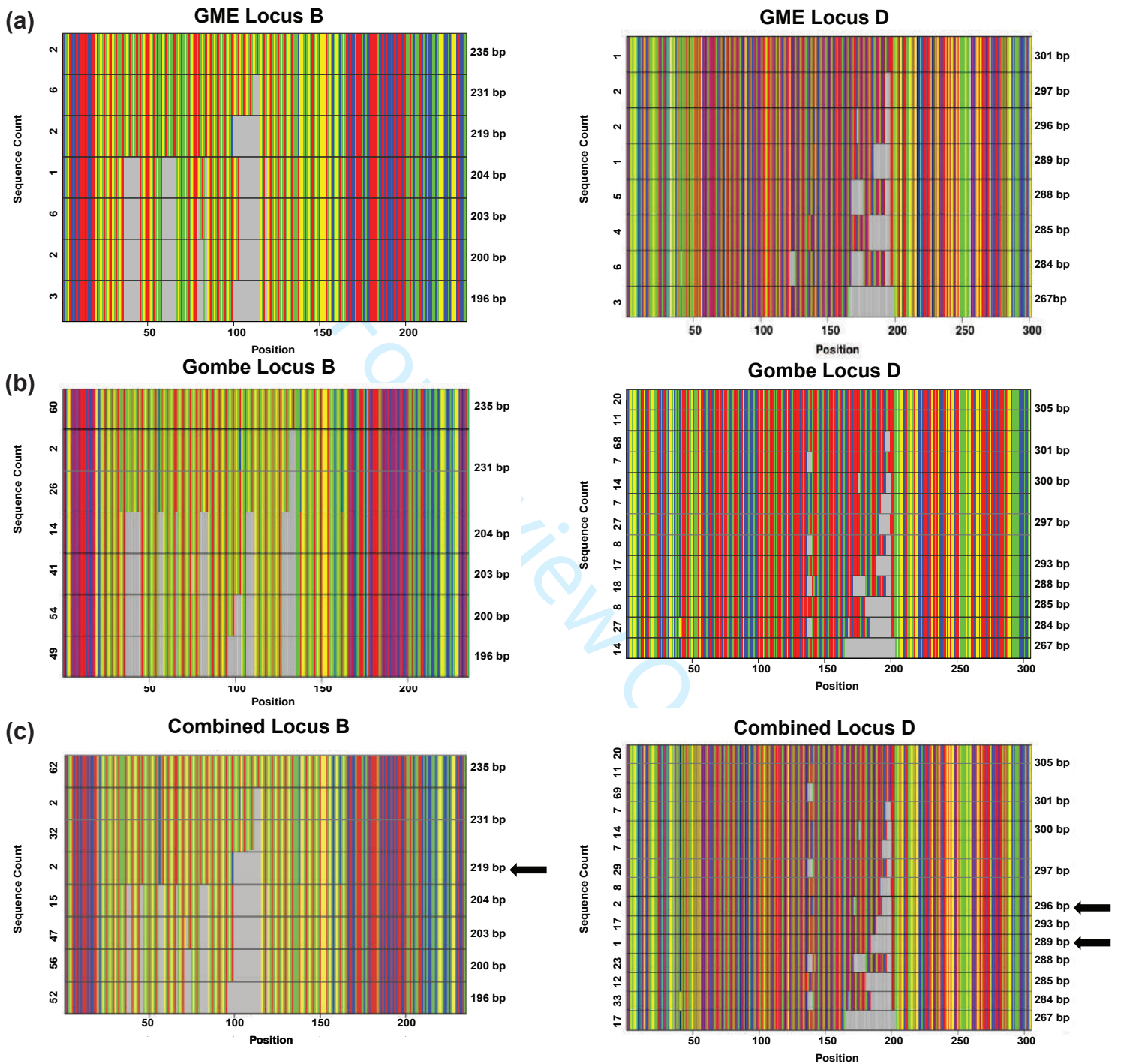






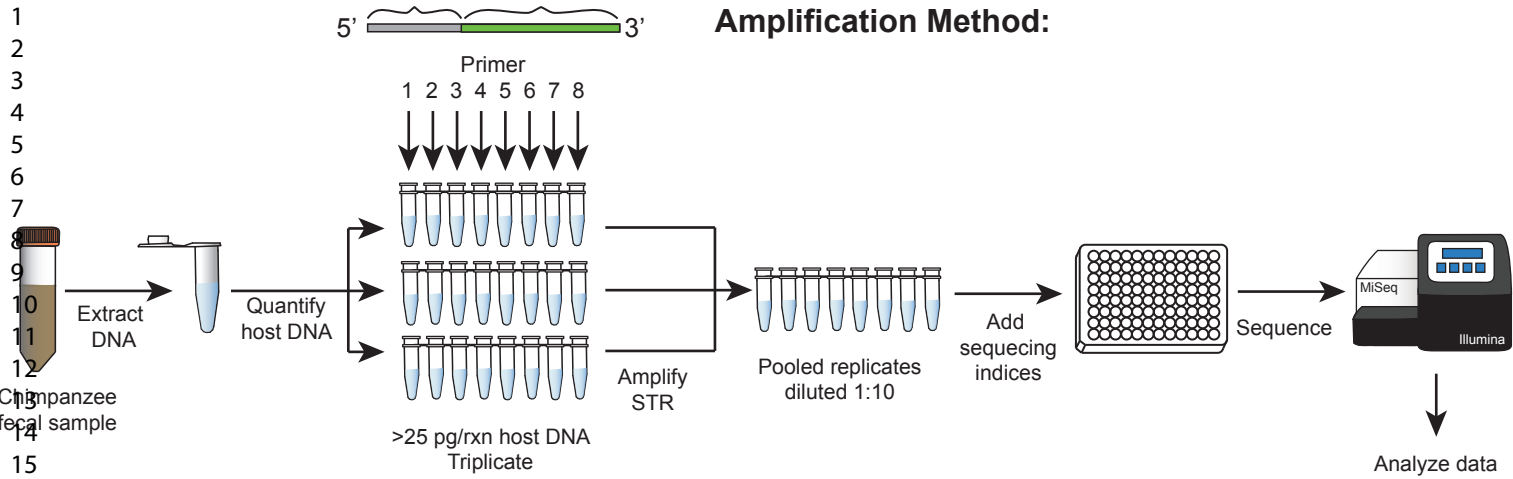


**Fig. 3**



**Fig. 4**

(a)



(b)

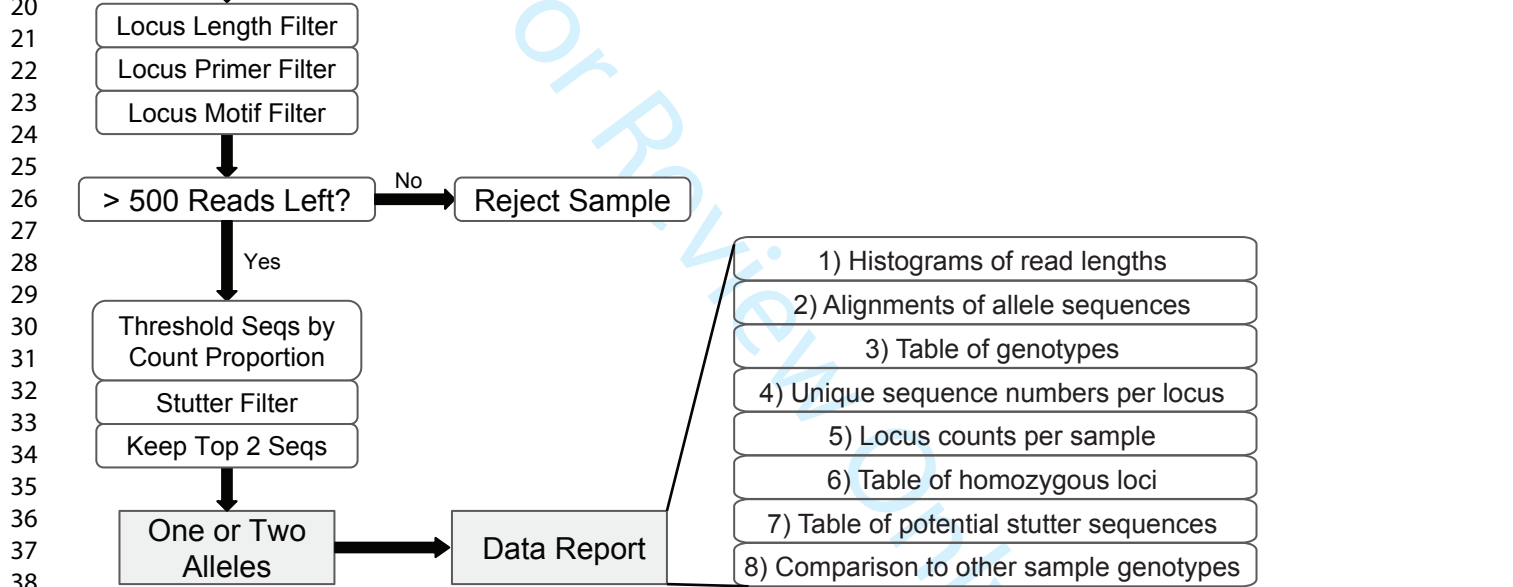


Fig. 5

**Table 1.** Comparison of capillary electrophoresis and MiSeq based genotyping results

Sample	Method	A-1	A-2	B-1	B-2	C-1	C-2	D-1	D-2	1-1	1-2	2-1	2-2	3-1	3-2	4-1	4-2
4775	CE-consensus <sup>†</sup>	141	161	204	235	182	190	286	306	244	256	302	318	237	241	279	295
	MiSeq <sup>‡</sup>	141	161	203	235	180	189	284	305	242	255	302	318	234	238	278	295
	CE-manual <sup>§</sup>	161	161	204	235	182	190	306	306	256	256	318	318	237	237	279	279
	CE-auto <sup>¶</sup>	141	161	204	235	182	190	302	306	252	256	314	318	233	237	279	279
4778	CE-consensus	141	161	235	235	182	190	286	298	256	260	318	322	237	241	279	283
	MiSeq	141	161	235	235	180	189	284	297	255	259	318	322	234	238	278	282
	CE-manual	141	161	235	235	182	190	285	298	256	256	318	322	237	241	279	283
	CE-auto	141	161	235	235	182	190	285	285	256	260	314	318	237	241	279	283
4781	CE-consensus	141	173	204	204	182	190	286	302	244	256	302	322	229	237	279	287
	MiSeq	141	173	203	203	180	180	284	301	242	254	302	322	226	234	278	286
	CE-manual	141	173	203	203	182	190	285	302	243	255	302	322	229	237	279	287
	CE-auto	141	173	203	203	182	190	285	286	243	255	302	322	229	237	279	287
4784	CE-consensus	141	173	196	200	182	194	298	302	244	244	302	322	241	241	295	295
	MiSeq	141	173	196	200	180	193	297	301	242	242	302	322	238	238	294	294
	CE-manual	173	173	196	200	182	194	302	302	243	243	302	322	241	241	295	295
	CE-auto	141	173	196	200	182	194	298	302	243	243	302	322	241	241	295	295
4792	CE-consensus	141	173	200	204	182	194	302	306	248	256	302	302	237	237	279	279
	MiSeq	141	173	200	203	180	193	301	305	247	255	302	302	234	234	278	278
	CE-manual	141	173	200	204	182	194	302	306	248	256	302	302	237	237	279	279
	CE-auto	141	173	200	200	182	194	302	306	248	256	302	302	233	237	279	279
4798	CE-consensus	141	173	196	200	190	194	302	306	256	260	322	326	229	237	279	279
	MiSeq	141	173	196	200	189	193	301	305	254	259	322	326	226	234	278	278
	CE-manual	141	173	196	200	190	194	302	306	256	260	322	326	229	237	279	279
	CE-auto	141	173	196	196	190	194	302	306	256	260	322	326	229	237	279	279
4805	CE-consensus	157	157	231	235	157	190	268	302	256	260	318	330	237	249	271	287
	MiSeq	157	157	231	235	157	157	267	301	255	258	318	330	234	246	270	286
	CE-manual	157	157	231	235	190	190	268	302	256	259	318	330	237	249	271	287
	CE-auto	154	157	235	235	183	190	268	302	256	259	318	330	237	249	287	287
4806	CE-consensus	153	173	196	235	182	190	298	302	244	260	318	322	225	237	279	283
	MiSeq	153	153	196	196	180	189	301	301	242	259	318	322	222	234	278	282
	CE-manual	153	173	196	196	182	194	298	302	244	260	318	322	227	237	283	283
	CE-auto	153	173	196	196	182	190	298	302	244	260	318	322	227	237	283	283
4807	CE-consensus	141	161	196	231	182	186	302	306	248	248	322	322	237	237	295	295
	MiSeq	141	161	196	231	181	185	305	305	246	247	322	322	234	234	295	295
	CE-manual	141	161	231	231	182	186	306	306	248	248	322	322	237	237	295	295
	CE-auto	141	161			182	186	306	306	244	248	318	322	233	237	295	295
4808	CE-consensus	141	177	203	231	182	190	290	299	257	260	318	331	241	249	271	287
	MiSeq	141	177	203	231	181	189	288	297	255	258	318	330	238	246	270	286
	CE-manual	141	177	203	231	182	190	290	298	256	259	318	330	241	249	271	286
	CE-auto	141	177	203	231	182	190	290	298	256	259	318	330	241	249	286	286
4821	CE-consensus	157	173	231	235	186	190	286	302	256	256	302	322	237	237	295	295
	MiSeq	157	173	231	235	185	189	285	300	255	255	302	322	234	234	294	295
	CE-manual	157	173	231	235	186	190	286	302	256	256	302	322	237	237	295	295
	CE-auto	157	173	231	231	186	190	286	302	252	256	302	322	233	237	295	295
4823	CE-consensus	141	141	235	235	158	182	298	302	256	256	322	322	229	233	287	295
	MiSeq	141	141	235	235	157	180	297	300	255	255	322	322	226	230	286	295
	CE-manual	141	141	235	235	158	182	300	300	256	256	322	322	229	233	287	287
	CE-auto	141	141	235	235	158	182	300	300	252	256	318	322	229	233	287	287
4830	CE-consensus	141	141	196	235	186	190	298	306	256	260	302	302	233	237	279	295
	MiSeq	141	141	196	235	185	189	297	305	254	258	302	302	230	234	278	295
	CE-manual	141	141	196	235	186	190	298	306	256	260	302	302	233	237	279	295
	CE-auto	137	141	196	235	186	190	298	306	256	260	302	302	233	237	279	279
4831	CE-consensus	141	161	196	200	186	186	286	290	244	252	318	322	237	237	295	295
	MiSeq	141	161	196	200	185	185	284	288	242	251	318	322	234	234	294	295
	CE-manual	141	161	196	200	186	186	286	290	244	252	318	322	237	237	295	295
	CE-auto	161	161			142	186	286	290	244	252	318	322	233	237	291	295

1

2	4844	CE-consensus	153	161	200	204	182	190	268	286	260	264	310	322	241	249	287	295
3		MiSeq	153	161	200	203	180	180	267	284	259	263	310	322	238	246	286	295
4		CE-manual	153	161	200	204	182	190	268	286	260	264	310	322	241	249	287	295
5		CE-auto	153	161	200	204	182	190	268	286	260	264	310	322	241	249	287	287
6	4845	CE-consensus	141	153	196	200	182	190	268	294	244	264	318	322	237	241	287	295
7		MiSeq	141	153	196	200	181	189	267	293	242	263	318	322	234	238	286	295
8		CE-manual	141	153	196	200	182	190	268	294	244	264	318	322	237	241	287	287
9		CE-auto	141	153	200	200	182	190	268	294	244	264	318	322	237	241	287	287
10	4850	CE-consensus	141	161	204	204	182	186	286	294	244	256	302	318	237	249	279	287
11		MiSeq	141	161	203	203	180	185	284	293	242	255	302	318	234	246	278	286
12		CE-manual	141	161	204	204	182	186	286	294	244	256	302	318	237	249	279	287
13		CE-auto	141	161	204	204	182	186	286	286	244	256	302	318	237	249	279	287
14	4859	CE-consensus	141	173	200	204	186	194	298	302	248	248	302	310	237	241	279	295
15		MiSeq	141	173	200	204	185	193	301	301	247	247	302	310	234	238	278	294
16		CE-manual	141	173	200	204	186	194	302	302	248	248	302	310	237	241	279	295
17		CE-auto	141	173	200	204	186	194	298	302	244	248	302	310	237	241	279	279
18	4861	CE-consensus	157	161	196	196	182	182	286	294	244	260	318	326	229	241	287	295
19		MiSeq	161	161	196	196	180	181	284	293	242	259	318	326	226	238	286	294
20		CE-manual	161	161	196	196	182	182	294	294					229	241	287	295
21		CE-auto	161	198	196	196	162	182	294	294					229	241	287	295

<sup>†</sup>CE-consensus: consensus genotype generated previously by capillary electrophoresis (CE) for multiple fecal samples from the same individual. This CE consensus genotype served as the benchmark to which all next generation sequencing (MiSeq) derived genotypes were compared.

<sup>‡</sup>MiSeq: MiSeq derived genotype of a newly collected (within the past two years) sample from the same individual. Note that most MiSeq alleles differ in length from the CE reference alleles by a few nucleotides. These discrepancies are locus-specific, with alleles of locus 2 exhibiting no length differences and alleles of locus 3 consistently differing by 3 bp.

<sup>§</sup>CE-manual: Capillary electrophoresis derived genotype of a newly collected sample from the same individual using manual peak calling and allele binning.

<sup>¶</sup>CE-auto: Capillary electrophoresis derived genotype of a newly collected sample from the same individual using peak calling software and manual allele binning; blue cells indicate false alleles, green cells indicate stutter sequences, orange cells indicate allelic dropout and gray cells indicate lack of amplification.

33

34

35

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

56

57

58

59

60

**Table 2.** Erroneous allele calls by capillary electrophoresis and MiSeq genotyping methods

	Capillary electrophoresis (automatic) <sup>†</sup>	%	Capillary electrophoresis (manual) <sup>‡</sup>	%	High throughput MiSeq genotyping	%
Allelic dropout	28	18 <sup>§</sup>	21	14	10	7
Missing locus	4	3	2	1	0	0
False allele <sup>¶</sup>	3	2	1	1	0	0
PCR stutter	18	12	0	0	0	0
Analysis time <sup>‡</sup>	75 min		120 min		5 min	

<sup>†</sup>Peaks were called automatically using software.

<sup>‡</sup>Peaks were called manually.

<sup>§</sup>The percentage of erroneous alleles was calculated for 152 loci by comparing the newly derived results to the reference genotypes (Table 1).

<sup>¶</sup>Locus alleles do not match the locus primer and/or motif sequence.

<sup>‡</sup>Hands-on analysis time included allele length calling, binning and individual identification.

For Review Only

**Table 3.** Increased allelic and gene diversity as detected by MiSeq STR genotyping

Locus	Number of alleles <sup>†</sup>			Gene diversity <sup>‡</sup>	
	CE <sup>§</sup>	MiSeq	Cryptic <sup>¶</sup>	CE	MiSeq
A	6	7	1	0.74	0.74
B	5	7	2	0.79	0.81
C	5	10	5	0.70	0.83
D	7	13	6	0.80	0.88
1	9	16	7	0.80	0.86
2	7	9	2	0.75	0.75
3	7	14	7	0.71	0.83
4	5	6	1	0.72	0.80
Total/Mean	51	82	31	0.75	0.81

<sup>†</sup>Number of alleles at eight STR loci determined for 123 Gombe chimpanzees (Table S2).

<sup>‡</sup>Nine individuals were excluded from heterozygosity calculations because they had incomplete CE genotypes.

<sup>§</sup>CE, capillary electrophoresis

<sup>¶</sup>Alleles newly discovered by MiSeq genotyping.



**Table 4.** Allelic sequence and length differences uncovered by MiSeq-based genotyping

Locus	Cryptic allele <sup>†</sup>	Number of apes carrying allele	Substitutions (identical length)	Indels (identical length)	Indels (1bp length difference)	Mendelian inheritance
A	157-b	3	2			Yes
B	204-a	14			1	Yes
B	231-b	2	1			Yes
C	181-a	10			1	Yes
C	181-b	1 <sup>‡</sup>	1			NA
C	185-b	20	1			Yes
C	185-c	11	1			Yes
C	189-b	35	1			Yes
D	285-a	8	3		1	Yes
D	297-b	8		2		Yes
D	297-c	7	1			Yes
D	300-a	14	1		1	Yes
D	301-b	7	3			Yes
D	305-b	11	1			Yes
1	246-a	10			5	Yes
1	247-b	4		2		Yes
1	250-a	2			5	Yes
1	254-a	27			1	Yes
1	258-a	6			5	Yes
1	258-b	3			3	Yes
1	266-b	2		4		Yes
2	310-b	1 <sup>‡</sup>	1			NA
2	326-b	6	1			NA
3	226-b	1 <sup>‡</sup>		2		NA
3	230-b	1 <sup>‡</sup>	3			NA
3	234-b	25		2		Yes
3	234-c	10	3	2		Yes
3	234-d	4		2		Yes
3	238-b	3		2		NA
3	246-b	7		2		Yes
4	294-a	38	3	2		Yes

<sup>†</sup>Cryptic alleles were compared to the most abundant allele of the same or similar length.

<sup>‡</sup>Alleles found in only one chimpanzee were confirmed by repeat amplification and sequencing.

**Table 4.** MiSeq genotyping of singleplex and multiplex amplified STR loci

	Singleplex PCR		One-step multiplex PCR		Two-step multiplex PCR	
		%		%		%
Allele detection	130	68 <sup>†</sup>	130	68	147	77
Incorrect allele	1	0.5	1	0.5	4	2.1
PCR stutter	1	0.5	1	0.5	4	2.1
DNA Input	24 ul		6 ul		6 ul	

<sup>†</sup>Of a total of 192 alleles analyzed for 12 GME chimpanzees.

For Review Only

# (a) Genotype Summary

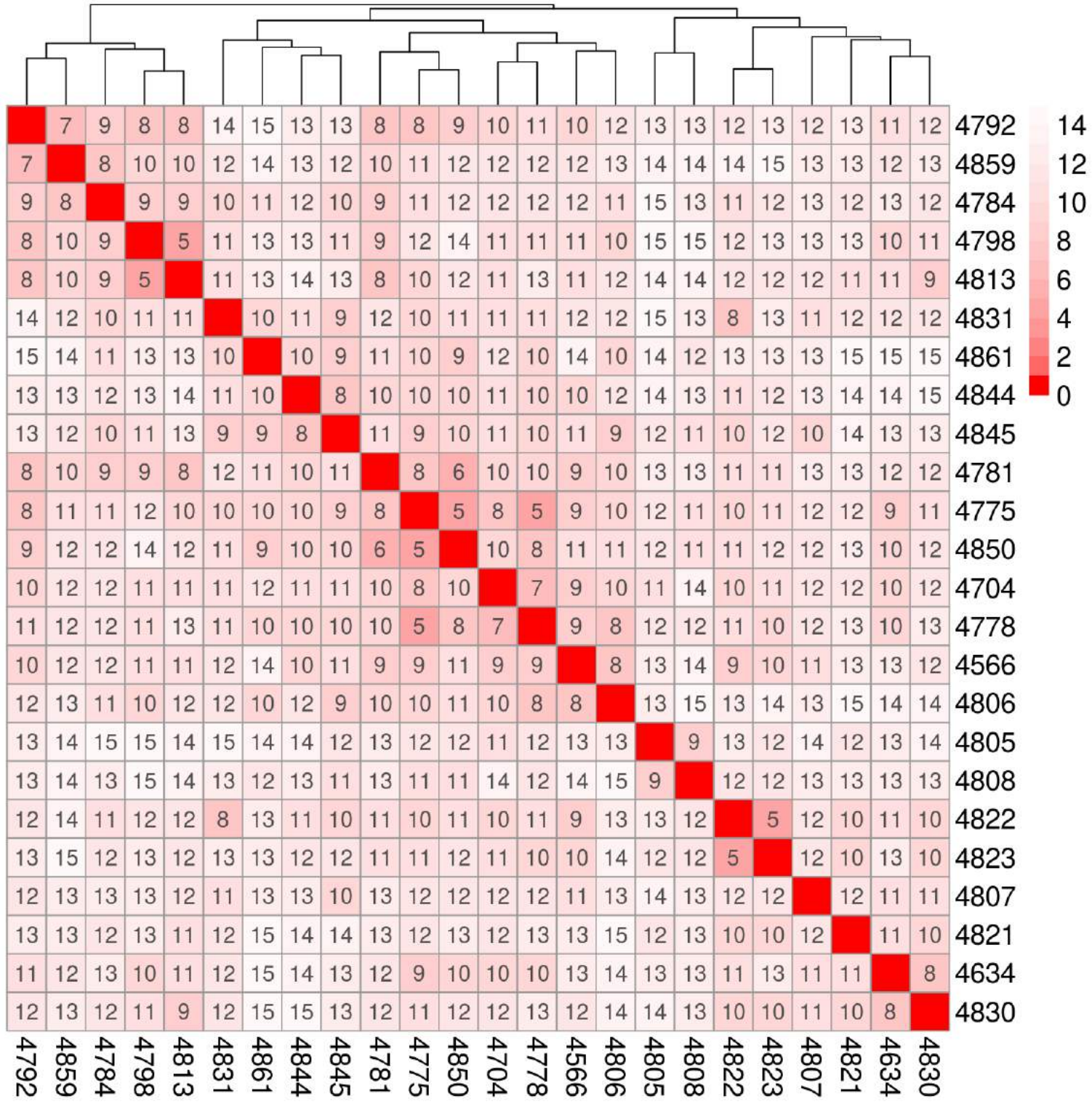
## Loci: A-D

Sample	A_1	A_2	B_1	B_2	C_1	C_2	D_1	D_2	Distance	Name
4566	141-a	153-a	203-a	235-a	180-a	185-a	297-c	301-a		
4634	141-a	161-a	200-a	235-a	185-b	189-b	305-a	305-a		
4704	161-a	161-a	200-a	235-a	180-a	180-a	297-b	301-a		
4775	141-a	161-a	203-a	235-a	180-a	189-b	284-a	305-b	0	Gremlin
4778	141-a	161-a	235-a	235-a	180-a	189-b	284-a	297-b	0	Gaia
4781	141-a	173-a	203-a	203-a	180-a	180-fdd1c6	284-a	301-a	1	Glitter
4784	141-a	173-a	196-a	200-a	180-a	193-a	297-a	301-a	0	Flirt
4792	141-a	173-a	200-a	203-a	180-a	193-a	301-a	305-b	0	Golden
4798	141-a	173-a	196-a	200-a	189-b	193-a	301-a	305-b	0	Fanni
4805	157-a	157-a	231-a	235-a	157-a	157-489d0d	267-a	301-a	1	Chema
4806	153-a	153-a	196-a	196-a	180-a	189-b	301-a	301-a	3	Ipo
4807	141-a	161-a	196-a	231-a	181-a	185-b	305-a	305-a	1	Losa
4808	141-a	177-a	203-a	231-a	181-a	189-a	288-a	297-a		
4813	141-a	173-a	196-a	231-a	180-a	193-a	301-a	305-b		
4821	157-a	173-a	231-a	235-a	185-b	189-a	285-a	300-a	0	Darbee
4822	141-a	141-a	200-a	235-a	180-a	185-a	288-a	300-a		
4823	141-a	141-a	235-a	235-a	157-a	180-a	297-a	300-a	0	Eliza
4830	141-a	141-a	196-a	235-a	185-b	189-a	297-a	305-a	1	Edgar
4831	141-a	161-a	196-a	200-a	185-a	185-c	284-a	288-a	0	Sheldon
4844	153-a	161-a	200-a	203-a	180-a	180-fdd1c6	267-a	284-a	1	Kati (Tita)
4845	141-a	153-a	196-a	200-a	181-a	189-b	267-a	293-a	0	Kazi
4850	141-a	161-a	203-a	203-a	180-a	185-a	284-a	293-a	0	Gimli
4859	141-a	173-a	200-a	204-a	185-c	193-a	301-b	301-a	1	Zeus
4861	161-a	161-a	196-a	196-a	180-a	181-a	284-a	293-a	2	Nasa

# Fig. S1

# (b) Inter-Sample Distance Matrix

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



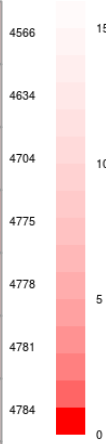
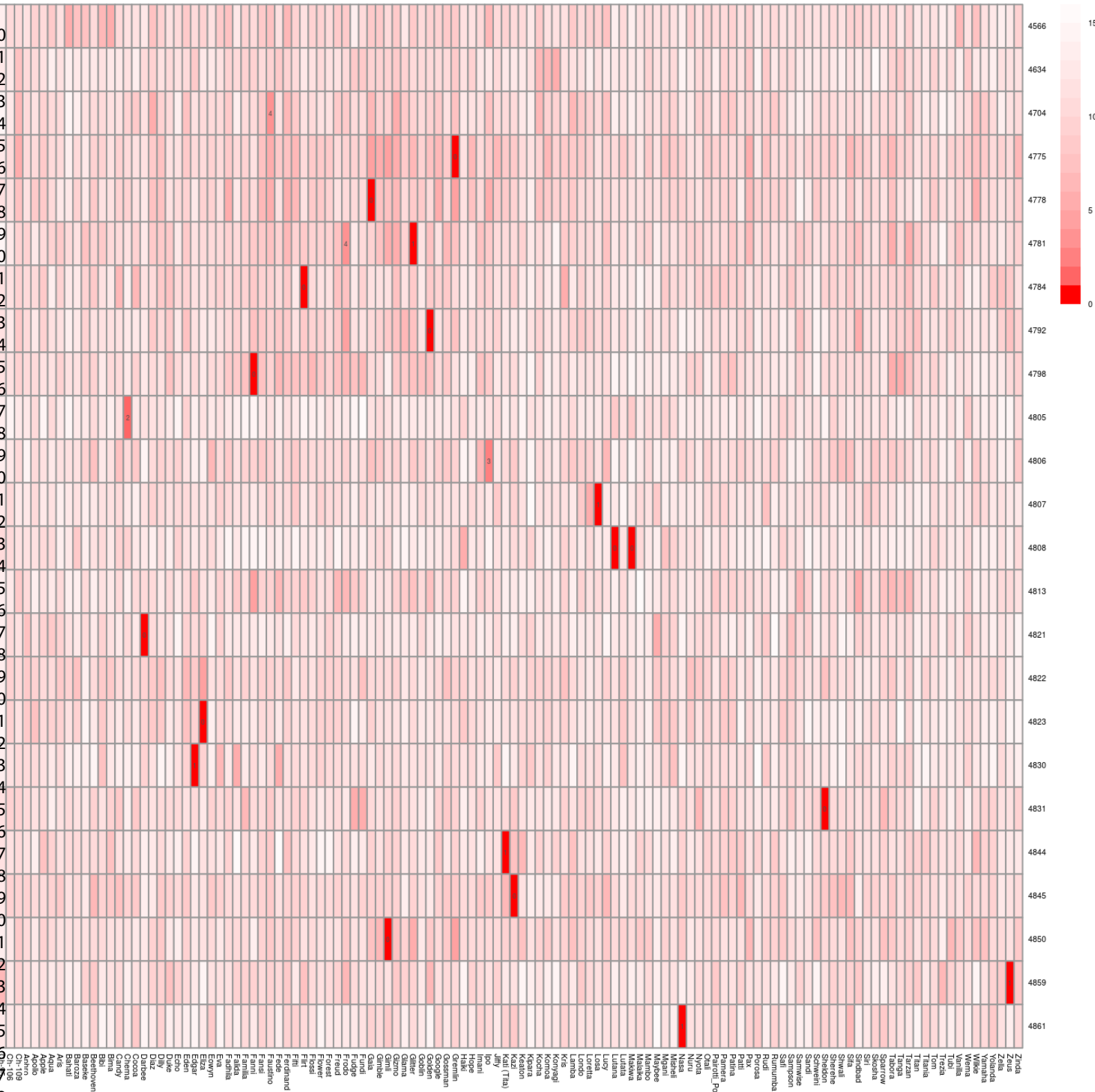
1  
2  
3  
4 **(c) Identification with Known Genotypes**  
5  
6

7 **Loci: A-D; Samples 4566-4781**  
8  
9

A_1	A_2	B_1	B_2	C_1	C_2	D_1	D_2	Distance	Name
<b>Sample 4566</b>									
141-a	153-a	203-a	235-a	180-a	185-a	297-c	301-a		
141-a	153-a	203-a	203-a	185-a	185-a	285-a	297-c	6	Bahati
141-a	153-a	200-a	203-a	185-a	185-a	285-a	297-c	6	Bima
141-a	153-a	203-a	235-a	185-a	189-a	285-a	297-a	7	Bibi
141-a	161-a	200-a	235-a	180-a	193-a	284-a	301-a	7	Ferdinand
153-a	173-a	196-a	235-a	180-a	189-b	297-b	301-a	7	Ipo
141-a	157-a	203-a	203-a	180-a	185-a	297-c	297-a	7	Vanilla
<b>Sample 4634</b>									
141-a	161-a	200-a	235-a	185-b	189-b	305-a	305-a		
161-a	177-a	200-a	235-a	180-a	189-b	305-a	305-a	6	Konyagi
141-a	161-a	200-a	235-a	180-a	180-a	301-a	305-a	7	Kocha
161-a	177-a	235-a	235-a	180-a	189-b	305-a	305-a	7	Komoa
<b>Sample 4704</b>									
161-a	161-a	200-a	235-a	180-a	180-a	297-b	301-a		
141-a	161-a	200-a	235-a	180-a	180-a	284-a	301-a	4	Faustino
<b>Sample 4775</b>									
141-a	161-a	203-a	235-a	180-a	189-b	284-a	305-b		
<b>141-a</b>	<b>161-a</b>	<b>203-a</b>	<b>235-a</b>	<b>180-a</b>	<b>189-b</b>	<b>284-a</b>	<b>305-b</b>	<b>0</b>	<b>Gremlin</b>
<b>Sample 4778</b>									
141-a	161-a	235-a	235-a	180-a	189-b	284-a	297-b		
<b>141-a</b>	<b>161-a</b>	<b>235-a</b>	<b>235-a</b>	<b>180-a</b>	<b>189-b</b>	<b>284-a</b>	<b>297-b</b>	<b>0</b>	<b>Gaia</b>
<b>Sample 4781</b>									
141-a	173-a	203-a	203-a	180-a	180-fdd1c6	284-a	301-a		
<b>141-a</b>	<b>173-a</b>	<b>203-a</b>	<b>203-a</b>	<b>180-a</b>	<b>189-b</b>	<b>284-a</b>	<b>301-a</b>	<b>1</b>	<b>Glitter</b>

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

# (d) Distance Matrix for Known Genotypes



# (e) Flagged Values

Loci subjected to stutter filter

10	153	203	235	180	185	297	301	247	259	322	322	222	234	278	295	4566	
11	161	200	235	185	189	305	305	254	255	302	334	230	234	270	278	4634	
12	161	200	235	180	180	297	301	254	255	318	322	234	242	278	295	4704	
13	161	203	235	180	189	284	305	242	255	302	318	234	238	278	295	4775	
14	161	235	235	180	189	284	297	255	259	318	322	234	238	278	282	4778	
15	173	203	203	180	180*	284	301	242	254	302	322	226	234	278	286	4781	
16	173	196	200	180	193	297	301	242	242	302	322	238	238	294	294	4784	
17	173	200	203	180	193	301	305	247	255	302	302	234	234	278	278	4792	
18	173	196	200	189	193	301	305	254	259	322	326	226	234	278	278	4798	
19	157	157	231	235	157*	157*	267	301	255	258	318	330	234	246	270	286	4805
20	153	153	196	196	180	189	301	301	242	259	318	322	222	234	278	282	4806
21	161	196	231	181	185	305	305	246	247	322	322	234	234	295	295	4807	
22	141	177	203	231	181	189	288	297	255	258	318	330	238	246	270	286	4808
23	173	196	231	180	193	301	305	251	254	302	326	226	234	278	295	4813	
24	173	231	235	185	189	285	300	255	255	302	322	234	234*	294	295	4821	
25	141	200	235	180	185	288	300	242	255	322	322	230	234	286	295	4822	
26	141	235	235	157	180	297	300	255	255	322	322	226	230	286	295	4823	
27	141	196	235	185	189	297	305	254	258	302	302	230	234	278	295	4830	
28	161	196	200	185	185*	284	288	242	251	318	322	234	234*	294	295	4831	
29	161	200	203	180	180*	267	284	259	263	310	322	238	246	286	295	4844	
30	153	196	200	181	189	267	293	242	263	318	322	234	238	286	295	4845	
31	161	203	203	180	185	284	293	242	255	302	318	234	246	278	286	4850	
32	173	200	204	185	193	301	301*	247	247	302	310	234	238	278	294	4859	
33	161	196	196	180	181	284	293	242	259	318	326	226	238	286	294	4861	

Loci with more than two prominent sequences

141	153	203	235	180	185	297	301	247	259	322	322	222	234	278	295	4566
141	161	200	235	185	189	305	305	254	255	302	334	230	234	270	278	4634
161	161	200	235	180	180	297	301	254	255	318	322	234	242	278	295	4704
141	161	203	235	180	189	284	305	242	255	302	318	234	238	278	295	4775
141	161	235	235	180	189	284	297	255	259	318	322	234	238	278	282	4778
141	173	203	203	180	180*	284	301	242	254	302	322	226	234	278	286	4781
141	173	196	200	180	193	297	301	242	242	302	322	238	238	294	294	4784
141	173	200	203	180	193	301	305	247	255	302	302	234	234	278	278	4792
141	173	196	200	189	193	301	305	254	259	322	326	226	234	278	278	4798
157	157	231	235	157*	157*	267	301	255	258	318	330	234	246	270	286	4805
153	153	196	196	180	189	301	301	242	259	318	322	222	234	278	282	4806
141	161	196	231	181	185	305	305	246	247	322	322	234	234	295	295	4807
141	177	203	231	181	189	288	297	255	258	318	330	238	246	270	286	4808
141	173	196	231	180	193	301	305	251	254	302	326	226	234	278	295	4813
157	173	231	235	185	189	285	300	255	255	302	322	234	234*	294	295	4821
141	141	200	235	180	185	288	300	242	255	322	322	230	234	286	295	4822
141	141	235	235	157	180	297	300	255	255	322	322	226	230	286	295	4823
141	141	196	235	185	189	297	305	254	258	302	302	230	234	278	295	4830
141	161	196	200	185	185*	284	288	242	251	318	322	234	234*	294	295	4831
153	161	200	203	180	180*	267	284	259	263	310	322	238	246	286	295	4844
141	153	196	200	181	189	267	293	242	263	318	322	234	238	286	295	4845
141	161	203	203	180	185	284	293	242	255	302	318	234	246	278	286	4850
141	173	200	204	185	193	301	301*	247	247	302	310	234	238	278	294	4859
161	161	196	196	180	181	284	293	242	259	318	326	226	238	286	294	4861

Proportion of allele-matching reads

35	141	153	203	235	180	185	297	301	247	259	322	322	222	234	278	295	4566
36	141	161	200	235	185	189	305	305	254	255	302	334	230	234	270	278	4634
37	161	161	200	235	180	180	297	301	254	255	318	322	234	242	278	295	4704
38	141	161	203	235	180	189	284	305	242	255	302	318	234	238	278	295	4775
39	141	161	235	235	180	189	284	297	255	259	318	322	234	238	278	282	4778
40	141	173	203	203	180	180*	284	301	242	254	302	322	226	234	278	286	4781
41	141	173	196	200	180	193	297	301	242	242	302	322	238	238	294	294	4784
42	141	173	200	203	180	193	301	305	247	255	302	302	234	234	278	278	4792
43	141	173	196	200	189	193	301	305	254	259	322	326	226	234	278	278	4798
44	157	157	231	235	157*	157*	267	301	255	258	318	330	234	246	270	286	4805
45	153	153	196	196	180	189	301	301	242	259	318	322	222	234	278	282	4806
46	141	161	196	231	181	185	305	305	246	247	322	322	234	234	295	295	4807
47	141	177	203	231	181	189	288	297	255	258	318	330	238	246	270	286	4808
48	141	173	196	231	180	193	301	305	251	254	302	326	226	234	278	295	4813
49	157	173	231	235	185	189	285	300	255	255	302	322	234	234*	294	295	4821
50	141	141	200	235	180	185	288	300	242	255	322	322	230	234	286	295	4822
51	141	141	235	235	157	180	297	300	255	255	322	322	226	230	286	295	4823
52	141	141	196	235	185	189	297	305	254	258	302	302	230	234	278	295	4830
53	141	161	196	200	185	185*	284	288	242	251	318	322	234	234*	294	295	4831
54	153	161	200	203	180	180*	267	284	259	263	310	322	238	246	286	295	4844
55	141	153	196	200	181	189	267	293	242	263	318	322	234	238	286	295	4845
56	141	161	203	203	180	185	284	293	242	255	302	318	234	246	278	286	4850
57	141	173	200	204	185	193	301	301*	247	247	302	310	234	238	278	294	4859
58	161	161	196	196	180	181	284	293	242	259	318	326	226	238	286	294	4861

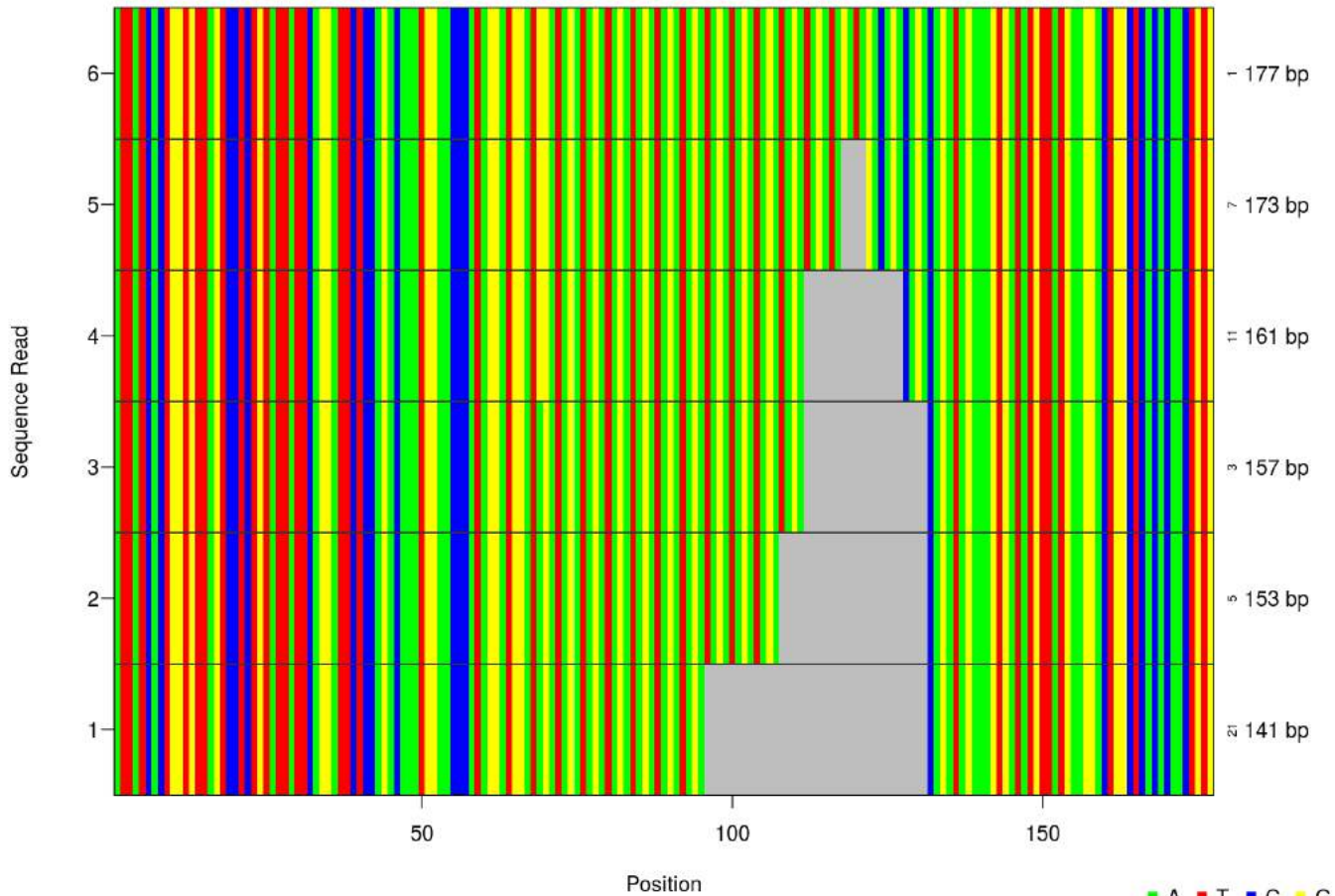
Loci with possible allelic dropout

141	153	203	235	180	185	297	301	247	259	322	322	222	234	278	295	4566
141	161	200	235	185	189	305	305	254	255	302	334	230	234	270	278	4634
161	161	200	235	180	180	297	301	254	255	318	322	234	242	278	295	4704
141	161	203	235	180	189	284	305	242	255	302	318	234	238	278	295	4775
141	161	235	235	180	189	284	297	255	259	318	322	234	238	278	282	4778
141	173	203	203	180	180*	284	301	242	254	302	322	226	234	278	286	4781
141	173	196	200	180	193	297	301	242	242	302	322	238	238	294	294	4784
141	173	200	203	180	193	301	305	247	255	302	302	234	234	278	278	4792
141	173	196	200	189	193	301	305	254	259	322	326	226	234	278	278	4798
157	157	231	235	157*	157*	267	301	255	258	318	330	234	246	270	286	4805
153	153	196	196	180	189	301	301	242	259	318	322	222	234	278	282	4

# (f) Allele Alignments per Locus

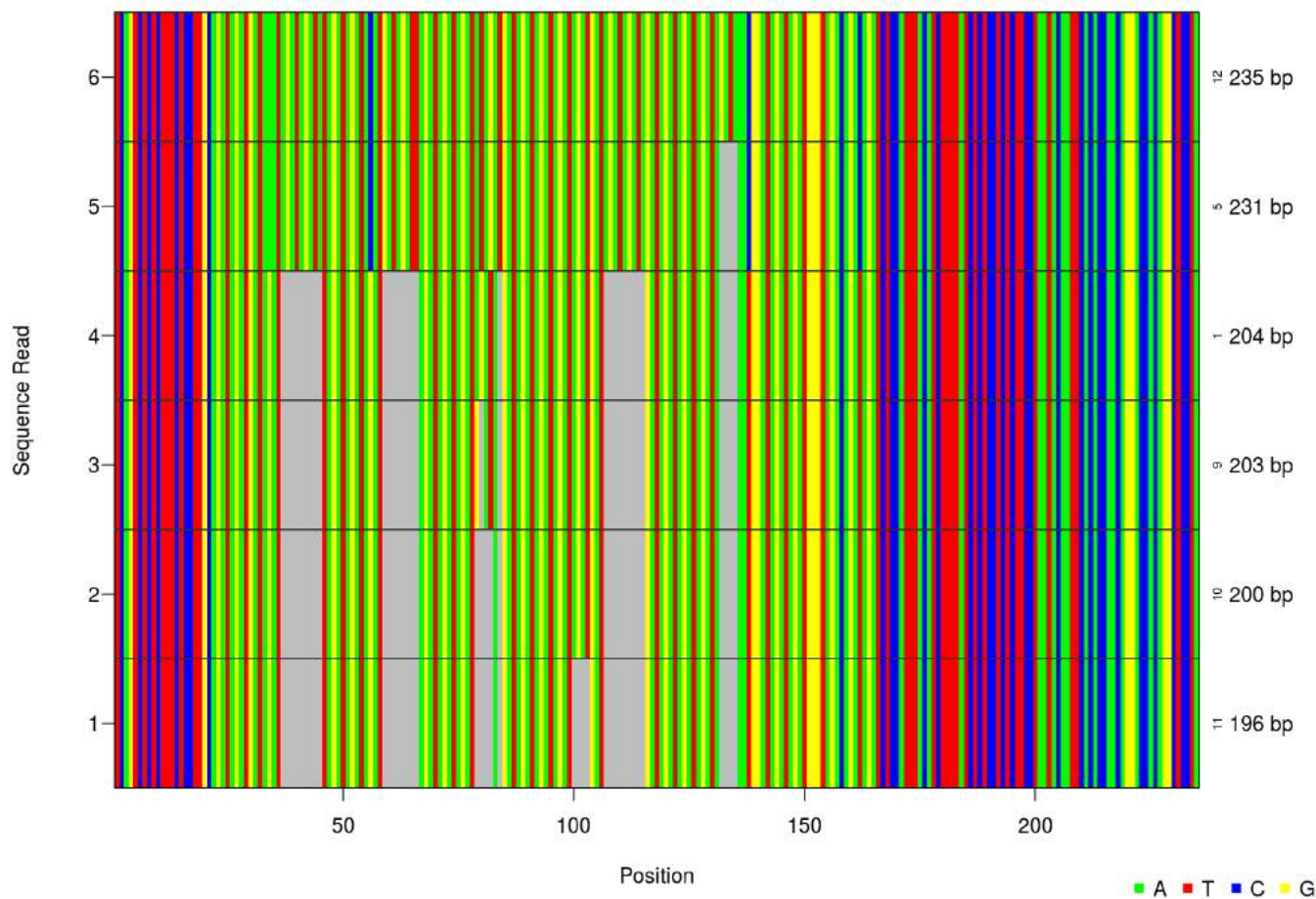
## Locus A

Alignment for Locus A



## Locus B

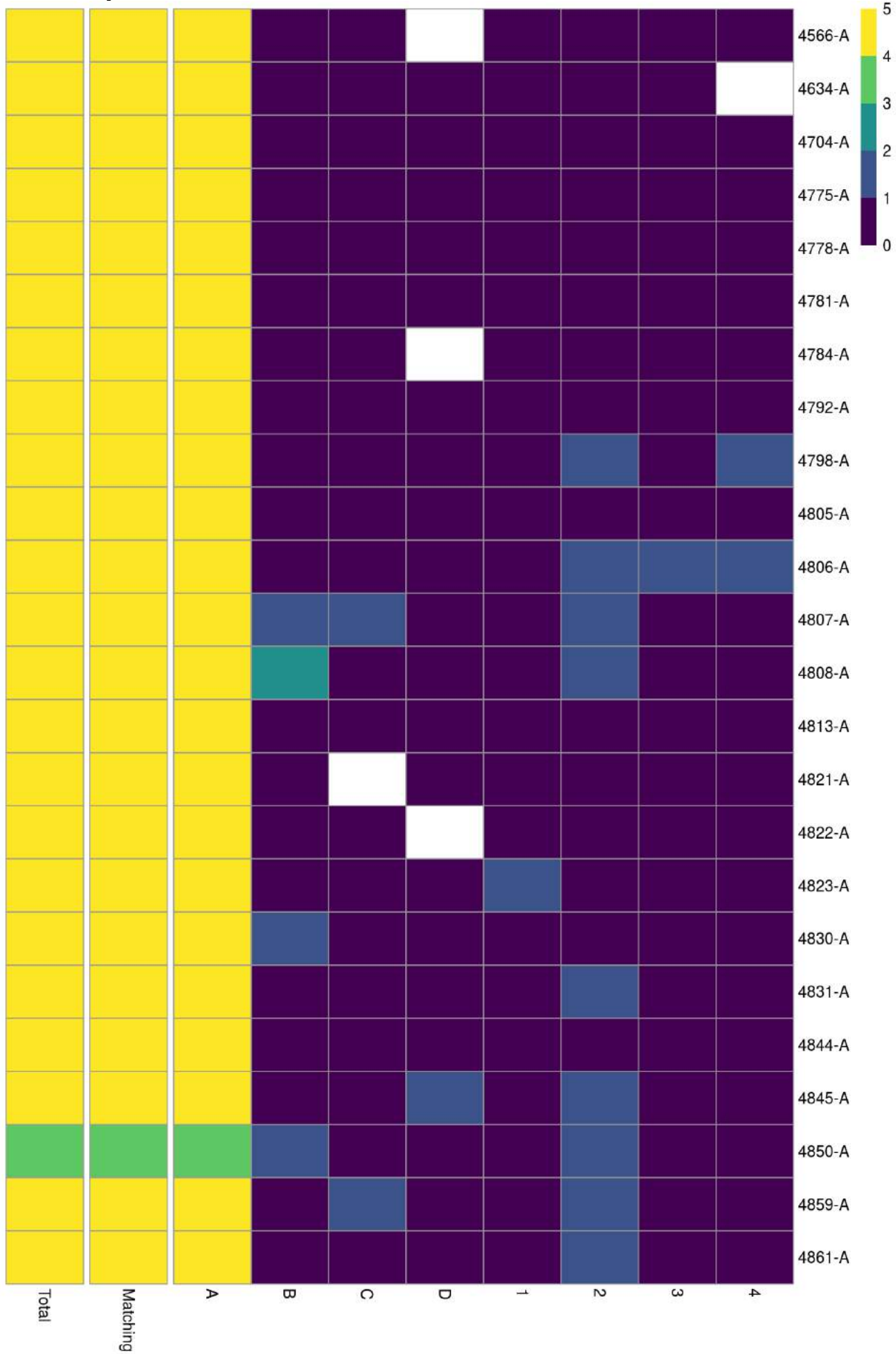
Alignment for Locus B





1  
2  
3  
4 (g) Counts per Locus  
5

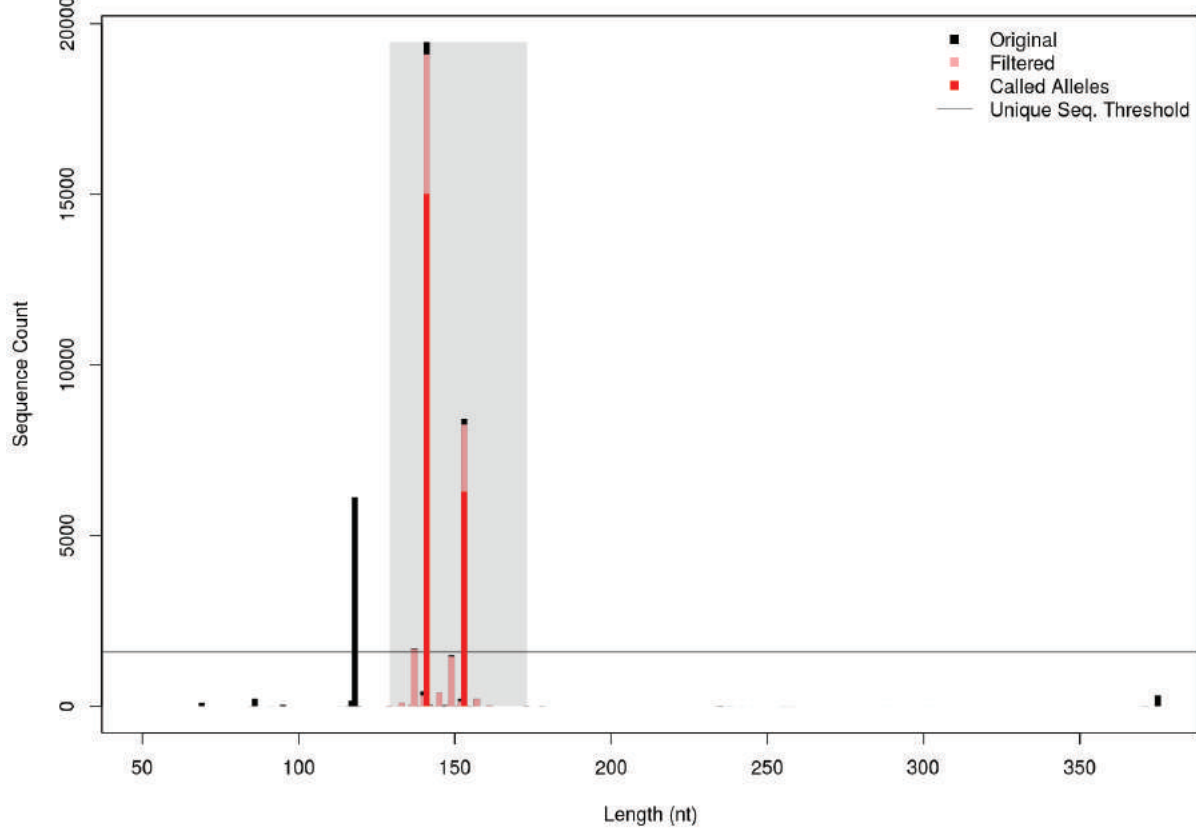
6 Samples for Locus A



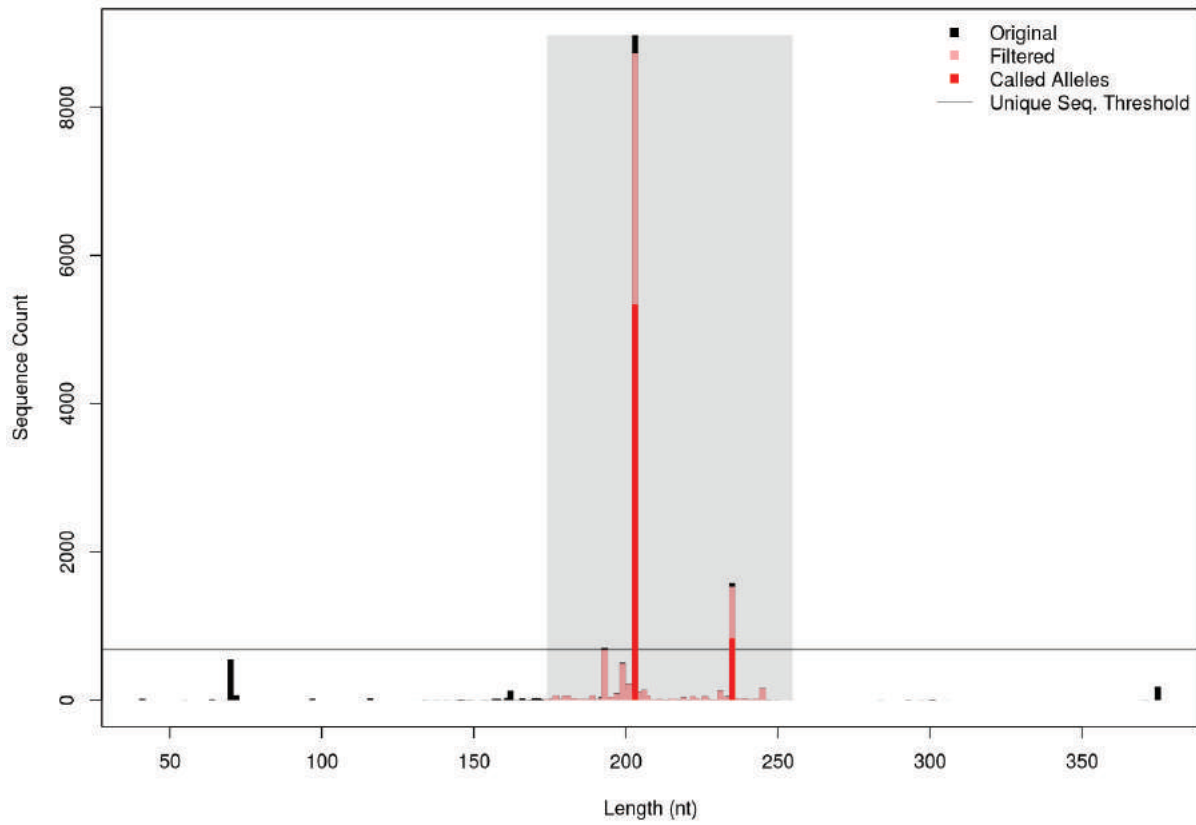
# (h) Histograms

## Sample 4566; Locus A, B

4566-A



4566-B



## Supplemental Information for:

### CHIIMP: An automated high-throughput microsatellite genotyping platform reveals greater allelic diversity in wild chimpanzees

Hannah J. Barbian, A. Jesse Connell, Alexa N. Avitto, Ronnie M. Russell, Andrew G. Smith, Madhurima S. Gundlapally, Alexander L. Shazad, Yingying Li, Frederic Bibollet-Ruche, Emily E. Wroblewski, Deus Mjungu, Elizabeth V. Lonsdorf, Fiona A. Stewart, Alexander K. Piel, Anne E. Pusey, Paul M. Sharp and Beatrice H. Hahn

**Fig. S1** The CHIIMP STR genotyping pipeline. Program outputs are shown for 24 chimpanzee fecal samples, genotyped at four polymorphic STR loci (A-D) corresponding to Table 1 and Fig. 3. (a) Summary genotype table listing sample designations for each row, STR loci for each column, and unique allele identifiers for each cell. Alleles are labeled by length, with a letter added (-a, -b, -c, etc.) to distinguish variants that differ in sequence content. Alleles that do not match previous identifiers receive a software-generated name to flag them as potentially new alleles (e.g., sample 4781, locus C, allele 2). (b) Distance matrix heatmap indicating the relative similarity of genotypes. Each cell contains a distance score, which is based on the number of allele mismatches between the respective samples (for 8 loci, the minimum is zero and the maximum is 16). Genotypes were clustered using the complete-linkage clustering algorithm in the hclust function of R (Hierarchical Clustering). More closely related genotypes are colored in darker red. Groups of closely related genotypes (top) may reveal close relatives (or in resampled communities, specimens from the same individual) (c) Individual identification based on genotyping. Genotypes of newly collected samples (top) are compared to the

1  
2  
3 genotypes of known community members, with the closest match listed below (ordered  
4 by descending distance scores). Genotypes that differ by fewer than four alleles are  
5 indicated in bold because they represent potential matches. (d) Heatmap showing the  
6 relative similarity of sample genotypes (rows) with genotypes of known individuals  
7 (columns) based on distance scores. Dark red cells indicate likely matches (Lutana and  
8 Makiwa represent the same individual). (e) Quality control tables highlighting loci where  
9 stutter sequences have been filtered, where more than two sequences pass the filter  
10 (with darker cells indicating more sequences), where a large proportion of reads is not  
11 contained in the identified alleles (light red indicates very low level of non-locus reads,  
12 indicating absence of contamination; dark red requires further scrutiny), and where  
13 homozygosity may reflect allelic dropout. (f) Alignments of allele sequences. Two  
14 representative images for locus A and B are shown. Allele sequences are ordered by  
15 length (indicated in base pairs on the right), with the frequency with which they were  
16 found in different chimpanzees indicated on the left (the x-axis indicates the position  
17 within the alignment). Nucleotides are colored as indicated, with gaps in the alignment  
18 shown in grey. (g) Heatmap of sequence counts that match the locus-specific forward  
19 primer. A representative analysis is shown for 24 chimpanzee samples amplified at locus  
20 A. The first column shows the total number of reads. The second column shows the  
21 matching reads. The remaining columns show the reads matching each locus (the scale  
22 bar indicates log increases; white cells indicate no reads). For singleplex samples, this  
23 identifies sequences that match other loci and thus highlights potential cross-locus  
24 contamination. For multiplexed samples, this shows the read distribution across  
25 loci. (h) Histograms depicting sequence length-frequency distributions saved as image  
26 files. Representative histograms are shown for locus A and B of one sample. Note that  
27 peaks can be comprised of identically sized reads that differ in their sequence content  
28 and can thus contain different colors. Black highlights reads that did not match the locus  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 length or repeat motif filter. Pink highlights reads that appear to be locus-specific, but  
4  
5 did not pass the PCR artifact filters (these are useful for identifying stutter  
6  
7 sequences). Only red reads represent true allele sequences. The horizontal line  
8  
9 indicates the minimum read cutoff for unique sequences. Histograms from each sample  
10  
11 and locus are saved as separate image files.  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

For Review Only

**Table S1.** STR loci used for MiSeq genotyping

Locus	Code	Forward primer	Forward primer sequence <sup>†</sup>	Reverse primer	Reverse primer sequence <sup>†</sup>	Size range (bp) <sup>‡</sup>
D18S536	A	HUM05262	5'-ATTATCACTGGTGTAGTCCTCTG-3'	HUM05263	5'-CACAGTTGTGTGAGCCAGTC-3'	127-183
D4S243	B	MGS02609	5'-TCAGTCTCTCTTTCTCCTTGCA-3'	MGS02610	5'-TAGGAGCCTGTGGTCCTGTT-3'	187-235
D10S676	C	HUM05148	5'-GAGAACAGACCCCCAAATCT-3'	HUM05149	5'-ATTTTCAGTTTTACTATGTGCATGC-3'	154-210
D9S922	D	HUM09025	5'-TCAGAGGACCACTGCCTAAG-3'	HUM09026	5'-CTGATGGGATTTGTGCCTAT-3'	260-308
D2S1326	1	HUM09373	5'-AGACAGTCAAGAATAACTGCCC-3'	HUM09374	5'-CTGTGGCTCAAAGCTGAAT-3'	166-234
D2S1333	2	HUM12880	5'-CTTTGTCTCCCCAGTTGCTA-3'	HUM12881	5'-TCTGTCATAAACCGTCTGCA-3'	269-357
D4S1627	3	HUM05068	5'-AGCATTAGCATTGTGCCTGG-3'	HUM05069	5'-GACTAACCTGACTCCCCCTC-3'	202-254
D9S905	4	HUM07339	5'-GTGGGAAAATTGGCCTAAGT-3'	HUM07340	5'-CTTCTGAGCCTCACACCTGT-3'	257-298

<sup>†</sup>STR loci were amplified as previously described (Keele *et al.* 2009b; Rudicell *et al.* 2010), except for the addition of MiSeq adapters at the 5' end of both forward (5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG-3') and reverse (5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG-3') primers.

<sup>‡</sup>All previously selected loci fell within the size range of the sequencing chemistry (Illumina v2 chemistry, 500 cycle kit) and were thus sequenced without fragmentation using only the forward reads (<400 bp).

**Table S2.** MiSeq derived genotypes for Gombe chimpanzees at eight STR loci

Sample code	Date collected	Chimp ID	Com. <sup>†</sup>	A_1	A_2	B_1	B_2	C_1	C_2	D_1	D_2	1_1	1_2	2_1	2_2	3_1	3_2	4_1	4_2
168	6/8/02	Yolanda	KK	157-b	161-a	196-a	200-a	180-a	185-b	297-b	301-b	242-a	242-a	310-a	326-a	234-a	238-a	286-a	294-a
181	5/4/02	Beethoven	KK	141-a	157-a	200-a	204-a	189-a	189-a	301-a	301-a	242-a	259-a	318-a	322-a	234-a	238-a	286-a	294-a
218	8/3/02	Aqua	MT	157-a	173-a	203-a	235-a	181-a	189-b	293-a	297-c	255-a	259-a	310-a	322-a	226-a	234-a	286-a	295-a
240	8/23/02	Haiki	KL	141-a	177-a	203-a	235-a	181-b	181-a	293-a	297-a	231-a	255-a	322-a	330-a	230-a	246-b	270-a	286-a
247	5/28/02	Ch-085	KL	141-a	161-a	196-a	196-a	185-a	189-a	293-a	301-a	242-a	255-a	318-a	322-a	222-a	238-a	286-a	286-a
337	8/7/03	Skosha	KK	173-a	173-a	196-a	196-a	157-a	180-a	293-a	297-a	259-a	259-a	318-a	322-a	234-a	234-a	295-a	295-a
646	8/18/04	Goblin	KK	161-a	173-a	196-a	235-a	189-b	189-a	300-a	305-a	246-a	259-a	322-a	326-a	234-b	234-a	270-a	278-a
661	5/11/05	Malaika	KK	161-a	177-a	203-a	203-a	181-a	185-a	293-a	301-b	231-a	259-a	310-a	322-a	230-a	234-a	270-a	286-a
667	12/14/04	Patti	KK	153-a	173-a	196-a	200-a	189-b	189-a	267-a	301-a	242-a	263-a	302-a	310-a	234-a	246-a	270-a	295-a
715	5/11/05	Echo	KK	141-a	177-a	196-a	196-a	185-c	189-a	288-a	297-a	247-a	247-a	322-a	322-a	222-a	238-a	270-a	278-a
863	10/13/05	Sherehe	KK	157-a	173-a	200-a	231-a	185-b	189-a	301-a	301-a	242-a	259-a	302-a	318-a	234-a	238-a	286-a	295-a
981	11/5/05	Gimble	KK	141-a	161-a	203-a	235-a	189-b	189-a	300-a	301-a	242-a	259-a	302-a	318-a	226-b	234-a	278-a	286-a
1164	2/11/07	Candy	KK	141-a	153-a	196-a	231-a	180-a	189-a	288-a	301-a	242-a	247-a	318-a	322-a	238-a	238-a	270-a	294-a
1212	7/22/07	Cocoa	KK	153-a	157-a	196-a	200-a	180-a	180-a	288-a	301-a	242-a	254-a	318-a	322-a	226-a	238-a	294-a	294-a
1320	8/9/07	Titania	KK	141-a	157-a	200-a	231-a	180-a	189-a	284-a	300-a	242-a	259-a	310-a	326-b	222-a	246-a	286-a	286-a
1393	4/19/08	Gremlin	KK	141-a	161-a	203-a	235-a	180-a	189-a	284-a	305-a	242-a	255-a	302-a	318-a	234-a	238-a	278-a	295-a
1532	9/17/08	Patina	KL	141-a	161-a	200-a	204-a	180-a	189-a	297-a	301-a	254-a	259-a	322-a	326-b	226-a	234-c	286-a	295-a
1542	1/19/09	Ch-106	KL	141-a	177-a	196-a	203-a	181-a	189-a	297-c	300-a	235-a	255-a	310-a	322-a	230-a	234-c	294-a	295-a
1648	9/8/09	Sheldon	KK	141-a	161-a	196-a	200-a	185-b	185-c	284-a	288-a	242-a	251-a	318-a	322-a	234-c	234-b	294-a	295-a
1660	12/8/09	Lucy	MT	141-a	153-a	200-a	204-a	180-a	180-a	267-a	301-a	231-a	242-a	318-a	322-a	222-a	234-a	278-a	286-a
1703	9/27/09	Sandi	KK	157-a	173-a	231-a	231-a	185-b	189-a	284-a	301-a	242-a	266-a	302-a	302-a	234-b	234-a	295-a	295-a
1705	7/9/09	Ch-109	KL	141-a	161-a	200-a	200-a	180-a	189-a	285-a	301-a	254-a	255-a	302-a	318-a	234-c	238-a	278-a	295-a
1709	10/9/09	Kris	KK	141-a	141-a	196-a	231-a	180-a	189-a	288-a	301-a	242-a	255-a	302-a	322-a	238-a	238-a	286-a	294-a
1720	10/28/09	Darbee	MT	157-a	173-a	231-a	235-a	185-a	189-b	285-a	300-a	255-a	255-a	302-a	322-a	234-b	234-d	294-a	295-a
1740	1/3/10	Lutana	KL	141-a	177-a	203-a	231-a	181-a	189-b	288-a	297-a	255-a	258-b	318-a	330-a	238-a	246-b	270-a	286-a
1752	10/7/09	Frodo	KK	141-a	173-a	200-a	203-a	180-a	193-a	301-a	301-a	247-a	254-a	302-a	322-a	226-a	234-a	278-a	286-a
1827	2/6/10	Hope	KK	141-a	161-a	235-a	235-a	189-a	189-a	284-a	300-a	247-a	247-a	302-a	302-a	234-a	238-a	270-a	295-a
1903	4/14/10	Eva	MT	141-a	153-a	196-a	235-a	189-b	189-b	297-a	301-a	258-a	266-a	302-a	322-a	234-b	234-a	295-a	295-a
2142	10/2/10	Baroza	KK	141-a	141-a	203-a	231-a	185-b	189-a	288-a	297-c	255-a	259-a	322-a	322-a	222-a	238-a	286-a	286-a
2297	1/7/11	Safi	KK	141-a	153-a	196-a	235-a	157-a	185-a	297-b	297-a	242-a	255-a	318-a	322-a	242-a	246-a	270-a	286-a
2376	3/11/11	Google	KK	141-a	161-a	200-a	235-a	180-a	189-a	284-a	301-a	246-a	259-a	318-a	318-a	238-a	238-a	282-a	295-a
2445	4/21/11	Tubi	KK	141-a	157-a	203-a	203-a	185-a	189-a	267-a	293-a	247-a	255-a	302-a	310-a	234-a	246-a	294-a	294-a
2589	6/7/11	Londo	MT	141-a	161-a	196-a	200-a	180-a	189-b	301-a	301-a	247-a	247-a	302-a	318-a	234-a	238-b	295-a	295-a
2597	7/20/11	Ferdinand	KK	141-a	161-a	200-a	235-a	180-a	193-a	284-a	301-a	254-a	259-a	322-a	322-a	234-a	238-a	278-a	286-a
2641	8/11/11	Ch-095	KL	141-a	177-a	200-a	204-a	185-c	185-a	301-b	305-a	247-a	259-a	310-a	330-a	238-a	246-b	282-a	294-a
2665	9/1/11	Maybee	MT	157-a	173-a	231-a	235-a	185-a	189-a	300-a	300-a	242-a	255-a	322-a	322-a	230-a	234-d	295-a	295-a
2673	9/7/11	Apollo	KK	141-a	141-a	196-a	235-a	157-a	189-a	297-a	297-a	255-a	259-a	322-a	322-a	234-a	242-a	270-a	294-a
2736	10/7/11	Mambo	KK	161-a	161-a	203-a	235-a	185-a	189-a	301-b	301-a	242-a	259-a	302-a	310-a	234-a	234-a	270-a	286-a
2906	4/9/12	Familia	KK	141-a	161-a	200-a	200-a	185-c	189-a	284-a	301-a	242-a	259-a	322-a	322-a	234-b	234-b	278-a	294-a
2935	4/21/12	Bima	MT	141-a	153-a	200-a	203-a	185-b	185-b	285-a	297-c	247-b	259-a	310-a	322-a	234-b	234-a	286-a	295-a
3001	6/8/12	Aris	MT	173-a	173-a	200-a	235-a	189-b	189-a	293-a	301-a	242-a	250-a	302-a	322-a	226-a	230-a	286-a	295-a
3011	6/21/12	Titan	KK	173-a	173-a	200-a	203-a	189-a	193-a	267-a	301-a	254-a	263-a	302-a	310-a	234-a	234-a	270-a	286-a
3016	6/24/12	Yamaha	MT	157-b	161-a	200-a	203-a	180-a	185-a	297-b	301-b	242-a	246-a	318-a	326-a	234-a	246-a	294-a	295-a
3046	7/11/12	Zella	KK	141-a	141-a	196-a	204-a	180-a	185-a	288-a	305-a	247-a	255-a	302-a	310-a	238-a	238-a	294-a	294-a
3049	7/12/12	Zinda	KK	141-a	161-a	203-a	204-a	185-c	189-a	301-a	305-a	242-a	259-a	318-a	330-a	234-a	246-b	278-a	294-a

1	3096	8/12/12	Diaz	KK	141-a	161-a	204-a	235-a	180-a	180-a	300-a	301-a	254-a	255-a	322-a	322-a	234-a	234-a	278-a	294-a
2	3097	8/12/12	Sampson	KK	141-a	157-a	231-a	235-a	157-a	185-b	284-a	297-a	242-a	255-a	302-a	322-a	234-b	242-a	294-a	295-a
3	3162	9/11/12	Mgani	MT	141-a	157-a	235-a	235-a	180-a	189-b	297-a	301-a	255-a	258-b	302-a	318-a	238-a	238-a	286-a	295-a
4	3165	9/11/12	Loretta	MT	141-a	161-a	196-a	203-a	181-a	189-b	301-a	301-a	247-a	247-a	318-a	322-a	234-a	234-a	295-a	295-a
5	3171	9/17/12	Glama	KK	161-a	173-a	200-a	200-a	180-a	185-c	305-a	305-a	254-a	255-a	302-a	326-a	234-c	234-a	278-a	278-a
6	3224	10/15/12	Fede	MT	141-a	141-a	196-a	235-a	189-b	193-a	305-b	305-a	242-a	254-a	302-a	326-a	234-b	234-a	278-a	278-a
7	3250	10/24/12	Kipara	KK	153-a	157-a	203-a	235-a	185-c	189-b	288-a	297-a	255-a	266-a	302-a	302-a	230-a	234-b	286-a	286-a
8	3280	10/31/12	Fudge	KK	141-a	161-a	200-a	200-a	185-c	189-a	288-a	305-a	251-a	254-a	318-a	326-a	234-c	234-b	278-a	294-a
9	3333	12/5/12	Zeus	KK	141-a	173-a	200-a	204-a	185-c	193-a	301-a	301-b	247-a	247-a	302-a	310-a	234-a	238-a	278-a	294-a
10	3348	12/20/12	Edgar	MT	141-a	141-a	196-a	235-a	185-b	189-b	297-a	305-b	254-a	258-a	302-a	302-a	230-a	234-b	278-a	295-a
11	3380	1/2/13	Forest	MT	141-a	141-a	231-a	235-a	189-b	189-a	300-a	305-a	242-a	247-a	322-a	330-a	226-a	234-a	278-a	294-a
12	3390	12/18/12	Wilkie	KK	153-a	161-a	203-a	235-a	180-a	185-a	284-a	297-b	246-a	259-a	318-a	322-a	234-a	246-a	282-a	295-a
13	3402	1/15/13	Golden	KK	141-a	173-a	200-a	203-a	180-a	193-a	301-a	305-a	247-a	255-a	302-a	302-a	234-a	234-a	278-a	278-a
14	3452	5/19/13	Sindbad	KK	157-a	173-a	203-a	231-a	180-a	193-a	284-a	301-a	247-a	251-a	302-a	302-a	234-b	234-a	278-a	295-a
15	3453	6/7/13	Rudi	MT	161-a	173-a	231-a	235-a	185-a	189-a	301-a	305-b	242-a	246-a	302-a	322-a	230-a	234-a	278-a	295-a
16	3505	4/20/13	Nasa	KK	157-b	161-a	196-a	196-a	180-a	181-a	284-a	293-a	242-a	259-a	318-a	326-b	226-a	238-a	286-a	294-a
17	3524	6/15/13	Apple	MT	161-a	173-a	200-a	203-a	180-a	189-b	293-a	297-a	255-a	259-a	302-a	322-a	234-c	238-a	282-a	295-a
18	3533	6/25/13	Lutata	MT	141-a	141-a	204-a	235-a	180-a	189-b	297-a	301-a	231-a	258-a	302-a	318-a	222-a	230-a	278-a	286-a
19	3705	2/4/14	Tarzan	KK	141-a	173-a	200-a	200-a	180-a	189-b	267-a	301-a	242-a	254-a	302-a	322-a	226-a	246-a	278-a	295-a
20	3731	3/9/14	Freud	KK	141-a	157-a	196-a	200-a	180-a	193-a	300-a	301-a	247-b	254-a	302-a	322-a	226-a	234-a	278-a	294-a
21	3746	6/17/14	Gossman	KK	141-a	141-a	196-a	203-a	185-b	189-a	288-a	301-a	242-a	242-a	302-a	322-a	234-b	234-a	278-a	286-a
22	3785	7/8/14	Nuru	KK	141-a	141-a	204-a	235-a	189-b	189-a	288-a	301-a	242-a	259-a	322-a	322-a	226-a	234-d	286-a	286-a
23	3806	7/16/14	Chema	KK	157-a	157-a	231-a	235-a	157-a	189-a	267-a	301-a	255-a	258-a	318-a	330-a	234-a	246-b	270-a	286-a
24	3807	7/16/14	Duke	KK	141-a	173-a	200-a	231-b	189-a	193-a	300-a	301-a	242-a	254-a	310-a	322-a	234-a	234-a	270-a	294-a
25	3816	2/20/14	Fifti	KK	161-a	173-a	196-a	235-a	185-a	189-a	284-a	305-a	246-a	259-a	318-a	322-a	226-a	234-a	278-a	295-a
26	3824	7/24/14	Pax	KK	141-a	161-a	235-a	235-a	180-a	189-a	284-a	297-c	242-a	254-a	302-a	326-b	234-a	234-a	278-a	286-a
27	3836	8/3/14	Gaia	KK	141-a	161-a	235-a	235-a	180-a	189-a	284-a	297-b	255-a	259-a	318-a	322-a	234-a	238-a	278-a	282-a
28	3848	8/7/14	Schweini	KK	153-a	161-a	196-a	235-a	185-b	185-a	284-a	297-b	242-a	246-a	318-a	318-a	234-a	246-a	282-a	286-a
29	3859	8/11/14	Falida	MT	141-a	141-a	196-a	204-a	189-b	193-a	297-a	305-b	242-a	258-a	302-a	326-a	234-b	234-a	278-a	278-a
30	3861	8/11/14	Aphro	MT	141-a	173-a	200-a	235-a	189-b	189-b	285-a	293-a	250-a	255-a	302-a	322-a	226-a	234-c	286-a	295-a
31	3874	8/16/14	Dilly	KK	141-a	161-a	204-a	231-b	180-a	189-a	300-a	301-a	242-a	255-a	322-a	322-a	234-a	238-a	286-a	294-a
32	3879	8/17/14	Ipo	KK	153-a	173-a	196-a	235-a	180-a	189-a	297-b	301-a	242-a	259-a	318-a	322-a	222-a	234-a	278-a	282-a
33	3884	8/19/14	Eliza	KK	141-a	141-a	235-a	235-a	157-a	180-a	297-a	300-a	255-a	255-a	322-a	322-a	226-a	230-a	286-a	295-a
34	3885	8/20/14	Sparrow	KK	157-a	161-a	196-a	231-a	180-a	185-b	284-a	301-a	242-a	251-a	302-a	318-a	234-b	246-a	286-a	295-a
35	3903	8/27/14	Jiffy	KK	141-a	161-a	231-a	235-a	189-b	189-a	284-a	301-a	247-a	258-a	302-a	302-a	234-a	238-a	295-a	295-a
36	3908	8/27/14	Flossi	MT	141-a	161-a	204-a	235-a	189-b	193-a	297-a	305-a	242-a	259-a	322-a	326-a	226-a	234-a	278-a	278-a
37	3953	9/29/14	Pamera	KL	141-a	161-a	196-a	200-a	189-a	189-a	297-a	301-a	254-a	255-a	322-a	322-a	234-c	238-a	286-a	295-a
38	3955	9/29/14	Porosa	KL	161-a	173-a	231-a	231-a	180-a	180-a	285-a	293-a	259-a	266-b	310-a	322-a	226-a	238-a	282-a	294-a
39	3958	10/6/14	Vanilla	KK	141-a	157-a	203-a	203-a	180-a	185-b	297-c	297-a	242-a	259-a	310-a	322-a	234-a	242-a	278-a	294-a
40	3961	10/11/14	Losa	MT	141-a	161-a	196-a	231-a	181-a	185-a	301-a	305-b	246-a	247-a	322-a	322-a	234-a	234-a	295-a	295-a
41	3966	10/22/14	Gimli	KK	141-a	161-a	203-a	203-a	180-a	185-a	284-a	293-a	242-a	255-a	302-a	318-a	234-a	246-a	278-a	286-a
42	3973	11/1/14	Baseke	KK	153-a	173-a	203-a	203-a	185-b	193-a	285-a	301-a	254-a	259-a	302-a	322-a	222-a	234-a	270-a	294-a
43	3974	11/1/14	Bahati	KK	141-a	153-a	203-a	203-a	185-b	185-b	285-a	297-c	247-b	259-a	322-a	322-a	222-a	234-b	286-a	294-a
44	3978	11/5/14	Imani	KK	141-a	173-a	196-a	196-a	189-a	189-a	288-a	301-a	242-a	247-a	318-a	322-a	222-a	226-a	278-a	286-a
45	3993	12/18/14	Faustino	KK	141-a	161-a	200-a	235-a	180-a	180-a	284-a	301-a	246-a	254-a	318-a	322-a	234-a	238-a	278-a	295-a
46	3996	12/20/14	Eowyn	KK	161-a	177-a	196-a	196-a	189-a	189-a	288-a	301-a	242-a	247-a	318-a	322-a	222-a	238-a	278-a	286-a
47	4000	12/27/14	Fliirt	MT	141-a	173-a	196-a	200-a	180-a	193-a	297-a	301-a	242-a	242-a	302-a	322-a	238-a	238-a	294-a	294-a
48	4001	12/28/14	Wema	MT	141-a	157-a	231-a	235-a	189-b	189-a	267-a	293-a	254-a	255-a	302-a	310-a	226-a	234-a	270-a	295-a



1	4004	12/9/14	Tanga	KK	161-a 173-a 196-a 200-a 189-b 189-a 301-a 305-a 242-a 259-a 302-a 326-a 234-b 246-a 270-a 278-a
2	4063	1/21/15	Pairotti	KL	141-a 173-a 196-a 231-a 157-a 180-a 293-a 293-a 259-a 266-b 302-a 310-a 234-a 238-a 294-a 295-a
3	4109	7/12/15	Nyota	KK	141-a 161-a 200-a 235-a 185-c 189-b 288-a 301-a 242-a 254-a 322-a 326-a 234-c 234-d 286-a 294-a
4	4113	6/6/15	Samwise	KK	141-a 173-a 203-a 231-a 185-b 193-a 301-a 301-a 247-a 266-a 302-a 302-a 226-a 234-b 286-a 295-a
5	4118	5/30/15	Sifa	KK	141-a 161-a 196-a 196-a 180-a 189-a 293-a 301-a 242-a 259-a 302-a 310-a 234-a 238-a 286-a 295-a
6	4152	7/17/15	Fadhila	KK	141-a 161-a 200-a 235-a 180-a 189-a 284-a 301-a 259-a 259-a 322-a 322-a 234-b 234-a 278-a 282-a
8	4180	8/20/15	Otali	MT	153-a 161-a 235-a 235-a 180-a 185-a 301-a 301-a 242-a 247-a 302-a 322-a 230-a 238-a 270-a 295-a
9	4195	8/23/15	Komoa	MT	161-a 177-a 235-a 235-a 180-a 189-a 305-b 305-b 242-a 255-a 302-a 334-a 230-a 246-a 278-a 295-a
10	4196	8/22/15	Gizmo	KK	141-a 161-a 200-a 235-a 180-a 180-a 284-a 301-a 242-a 254-a 302-a 322-a 234-a 238-a 278-a 278-a
11	4204	8/23/15	Kocha	MT	141-a 161-a 200-a 235-a 180-a 180-a 301-a 305-b 255-a 259-a 322-a 334-a 238-b 246-a 270-a 278-a
12	4205	8/23/15	Misheli	MT	141-a 157-a 235-a 235-a 189-b 189-b 297-a 301-a 254-a 255-a 302-a 318-a 230-a 238-a 295-a 295-a
13	4206	8/11/15	Tabora	KK	141-a 173-a 196-a 235-a 180-a 189-a 284-a 305-a 242-a 254-a 322-a 326-a 234-b 234-a 278-a 286-a
14	4211	9/2/15	Makiwa	KL	141-a 177-a 203-a 231-a 181-a 189-b 288-a 297-a 255-a 258-b 318-a 330-a 238-a 246-b 270-a 286-a
15	4219	8/25/15	Fansi	MT	141-a 161-a 200-a 235-a 180-a 193-a 267-a 305-a 242-a 259-a 322-a 326-b 226-a 234-a 278-a 282-a
16	4220	8/25/15	Eden	MT	141-a 141-a 200-a 235-a 180-a 189-b 267-a 301-a 255-a 266-a 302-a 322-a 234-a 234-a 286-a 295-a
17	4223	8/26/15	Kati	KL	153-a 161-a 200-a 203-a 180-a 189-a 267-a 284-a 259-a 263-a 310-a 322-a 238-a 246-a 286-a 295-a
18	4234	9/26/15	Fundi	KK	141-a 161-a 196-a 200-a 185-b 189-a 288-a 305-a 242-a 254-a 322-a 322-a 234-b 234-b 278-a 294-a
19	4239	9/24/15	Fanni	KL	141-a 173-a 196-a 200-a 189-a 193-a 301-a 305-a 254-a 259-a 322-a 326-a 226-a 234-b 278-a 278-a
20	4247	9/14/15	Trezia	KK	141-a 141-a 200-a 204-a 185-c 185-a 301-b 305-a 247-a 259-a 310-a 330-a 238-a 246-b 282-a 294-a
21	4249	9/22/15	Rumumba	KK	153-a 161-a 200-a 200-a 180-a 180-a 267-a 297-a 231-a 259-a 318-a 326-b 234-a 238-b 278-a 286-a
22	4346	1/28/16	Shwali	KK	153-a 173-a 196-a 200-a 185-b 189-a 301-a 297-b 242-a 263-a 310-a 318-a 234-a 234-a 286-a 286-a
23	4443	5/6/16	Glitter	KK	141-a 173-a 203-a 203-a 180-a 189-a 284-a 301-a 242-a 254-a 302-a 322-a 226-a 234-a 278-a 286-a
24	4448	5/9/16	Siri	KK	141-a 157-a 196-a 231-a 189-a 189-a 284-a 297-a 259-a 266-a 302-a 322-a 234-a 234-a 270-a 295-a
25	4475	5/18/16	Flower	KL	141-a 173-a 235-a 235-a 189-b 189-a 305-b 305-a 242-a 246-a 322-a 322-a 230-a 234-a 278-a 278-a
26	4515	7/7/16	Bibi	MT	141-a 153-a 203-a 235-a 185-b 189-b 285-a 297-a 247-b 254-a 302-a 322-a 230-a 234-a 286-a 295-a
27	4517	7/17/16	Lamba	MT	153-a 161-a 200-a 203-a 180-a 180-a 267-a 267-a 231-a 255-a 318-a 322-a 222-a 234-a 278-a 286-a
28	4519	7/21/16	Konyagi	MT	161-a 177-a 200-a 235-a 180-a 189-a 305-b 305-b 255-a 259-a 310-b 334-a 230-b 246-a 270-a 295-a
29	4522	7/23/16	Tom	KK	141-a 161-a 196-a 231-a 189-a 189-a 288-a 301-a 255-a 259-a 302-a 322-a 238-a 246-a 270-a 294-a
30	4528	7/28/16	Keaton	KL	153-a 161-a 203-a 203-a 185-a 189-b 284-a 297-a 259-a 266-a 302-a 322-a 230-a 246-a 286-a 295-a
31	4534	7/19/16	Kazi	KL	141-a 153-a 196-a 200-a 181-a 189-a 267-a 293-a 242-a 263-a 318-a 322-a 234-a 238-a 286-a 295-a

<sup>†</sup>Gombe community: KK – Kasekela; MT – Mitumba; KL - Kalande

**Table S3.** MiSeq genotyping of GME chimpanzees using singleplex and multiplex locus amplification

Sample (Individual)	Method	A-1	A-2	B-1	B-2	C-1	C-2	D-1	D-2	1-1	1-2	2-1	2-2	3-1	3-2	4-1	4-2
TZ037 (Ch-44)	CE - historical <sup>†</sup>	139	159	230	234	175	183	286	298	231	259	321	321	221	221	281	293
	Singleplex <sup>‡</sup>	141	161					284	284	231	259			222	222	282	295
	One-step multiplex <sup>§</sup>	141	161			177	185	284	296	231	259			222	222		
	Two-step multiplex <sup>¶</sup>	141	141	181	181	177	185	284	284	231	231	322	322	222	222	282	294
TZ060 (Ch-58)	CE - historical	139	159	202	202	175	183	286	290	231	263	337	337	221	221	269	293
	Singleplex	141	161	203	204	177	185	284	288	231	231	338	338	222	222	270	295
	One-step multiplex	141	161	203	204	177	185	284	288	231	231			222	222		
	Two-step multiplex	141	141	203	204	177	185	284	288	231	231	334	338	222	222	270	295
TZ096 (Ch-19)	CE - historical	139	175	202	230	183	183	286	290	251	255	317	321	221	233	281	293
	Singleplex	141	141	203	231	185	185	285	288	251	255	318	322	222	234	282	295
	One-step multiplex	141	177	203	203	185	185	285	288	251	255			222	234		
	Two-step multiplex	141	141	203	203	185	185			251	255	318	322	222	234	282	295
TZ220 (Ch-51)	CE - historical	171	175	194	202	175	175	286	290	231	259	317	317	221	221	293	293
	Singleplex	173	177	196	203	177	177	285	288	231	258			222	222	295	295
	One-step multiplex	173	177	196	203	177	177	285	288	231	258			222	222		
	Two-step multiplex	169	173	196	203	177	177	288	288	231	231	318	318	222	222	295	295
TZ254 (Ch-26)	CE - historical	159	159	202	234	183	191	298	302	251	255	329	337	221	221	293	293
	Singleplex	161	161	203	235	185	193			251	255			222	222		
	One-step multiplex	161	161			185	193	297	301	251	255			222	222		
	Two-step multiplex	161	161	203	203	185	193			251	255	330	338	222	222	295	295
TZ259 (Ch-54)	CE - historical	151	159	198	202	175	183	270	286	231	255	321	325	221	221	293	293
	Singleplex	153	161	200	203	177	185	267	284	227	231			222	222		
	One-step multiplex	153	161	200	203	177	185	267	267	227	231			222	222		
	Two-step multiplex	153	153	200	203	177	185	267	267	227	231			222	222	295	295
TZ260 (Ch-4)	CE - historical	159	175	218	230	175	183	286	298	251	262	317	325	221	233	285	293
	Singleplex	161	177	219	219	177	185	284	296	251	262			222	234		
	One-step multiplex	161	177			177	181	284	296	251	262			222	234		
	Two-step multiplex	161	177	219	231	177	181	284	284	251	262	318	326	222	234	286	295
TZ263 (Ch-48)	CE - historical	159	171	194	202	175	179	270	286	251	263	325	325	221	221	281	285
	Singleplex	161	173	196	203			267	285	251	262			222	222		
	One-step multiplex	161	173	196	203	177	180	267	285	251	262			222	222		
	Two-step multiplex	161	161	196	203	177	180	267	285	251	262			222	222	282	286
TZ264 (Ch-53)	CE - historical	159	171	230	230	183	183	290	298	251	255	309	337	221	221	269	269
	Singleplex	161	173	231	231					251	255			222	222		
	One-step multiplex	161	161			185	185	288	297	251	255			222	222		
	Two-step multiplex	161	161	231	231	185	185							222	222	270	270
TZ271 (Ch-18)	CE - historical	139	171	230	234	175	179	286	290	263	263	321	325	221	225	285	293
	Singleplex			231	235	177	181			262	262			222	226		
	One-step multiplex	141	173	231	235	177	181	284	288	262	262			222	226		
	Two-step multiplex	141	173	231	235	177	181					322	326	222	226	286	295

1																			
2	TZ320	CE - historical	175	179	194	198	183	191	270	290	263	263	321	321	233	233	269	293	
3	(Ch-67)	Singleplex	141	177	196	200			267	289	262	262			234	234	270	295	
4		One-step multiplex	177	181	196	200	185	193	267	289	262	262			234	234			
5		Two-step multiplex	177	181	196	200	185	185	267	267	212	262			230	234	270	295	
6	TZ336	CE - historical	139	175	218	230	179	183	286	286	251	251	317	321	221	233	285	293	
7	(Ch-3)	Singleplex	141	177	219	231					251	251							
8		One-step multiplex	141	177	219	219	180	185	284	285	251	251			222	234			
9		Two-step multiplex	141	141	219	231	180	180	284	284					222	234	286	295	

<sup>†</sup>CE-historical: historical genotype generated previously by capillary electrophoresis (CE) analysis for samples from GME chimpanzees (Rudicell et al., 2011). All newly-derived genotypes are compared to this reference genotype.

<sup>‡</sup>Singleplex: genotype of the same samples generated by amplifying and MiSeq sequencing each locus individually.

<sup>§</sup>One-step multiplex: genotype of the same sample generated by amplifying and MiSeq sequencing two pools of four loci (one-step PCR).

<sup>¶</sup>Two-step multiplex: genotype of the same sample generated by using the one-step multiplex PCR product as the input for a second round PCR to then amplify each locus individually (two-step PCR). Blue cells indicate false alleles, green cells indicate stutter sequences, orange cells indicate allelic dropout, and gray cells indicate lack of amplification.

For Review Only

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60