Ecology and Evolution Open Access

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

Journal:	*Ecology and Evolution*
Manuscript ID	ECE-2018-05-00542.R1
Wiley - Manuscript type:	Original Research
Date Submitted by the Author:	n/a
Complete List of Authors:	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
Category:	Genetics
Habitat:	Terrestrial

Organism:	Vertebrate
Approach:	Method Development
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.

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

Hannah J. Barbian[1], A. Jesse Connell[1], Alexa N. Avitto[1], Ronnie M. Russell[1], Andrew G. Smith[1], Madhurima S. Gundlapally[1], Alexander L. Shazad[1], Yingying Li[1], Frederic Bibollet-Ruche[1], Emily E. Wroblewski[2], Deus Mjungu[3], Elizabeth V. Lonsdorf[4], Fiona A. Stewart[5], Alexander K. Piel[5], Anne E. Pusey[6], Paul M. Sharp[7] and Beatrice H. Hahn[1*]

[1]Departments of Microbiology and Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

[2]Department of Anthropology, Washington University in St. Louis, St. Louis, MO, USA

[3]Gombe Stream Research Center, Kigoma, Tanzania

[4]Department of Psychology, Franklin and Marshall College, Lancaster, Pennsylvania

[5]School of Natural Sciences and Psychology, Liverpool John Moores University, Liverpool, United Kingdom

[6]Department of Evolutionary Anthropology, Duke University, Durham, North Carolina

[7]Institute of Evolutionary Biology and Centre for Immunity, Infection and Evolution, University of Edinburgh, Edinburgh EH9 3FL, United Kingdom

Professor, Departments of Medicine and Microbiology
University of Pennsylvania Perelman School of Medicine
409 Johnson Pavilion
3610 Hamilton Walk
Philadelphia, PA 19104-6076
USA
bhahn@pennmedicine.upenn.edu

Running title: High throughput STR genotyping of chimpanzee

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*

Introduction

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

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

artifacts, which are difficult to identify on electropherograms. These include stutter peaks, which are usually one repeat shorter than the correct STR allele and derive from Taq polymerase slippage (Hauge & Litt 1993; Shinde *et al.* 2003), split peaks which are caused by inconsistent A-overhang addition (Schuelke 2000), and artifactual peaks, which are the product of off-target amplification and/or unspecific fluorescent signaling (Ewen *et al.* 2000; Fernando *et al.* 2001; Guichoux *et al.* 2011). Existing peak-calling software often fails to differentiate erroneous from real peaks and frequently omits peaks of low height. Automatically called peaks must therefore be corrected manually, which is labor intensive and time consuming (Guichoux *et al.* 2011). Finally, multiplexing is restricted to only a few fluorescent labels, thus limiting the number of loci that can be analyzed simultaneously. As a consequence, capillary electrophoresis based STR genotyping is laborious, notoriously imprecise, and generally not useful for large sample sets or data sharing between different platforms and/or field sites (Pasqualotto *et al.* 2007).

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

visual inspection of read length histograms (Darby *et al.* 2016; Farrell *et al.* 2016; Vartia *et al.* 2016), and none of these studies have compared the performance of capillary electrophoresis and high throughput sequencing side-by-side to validate and improve the genotyping approach.

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

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

tools, which are not included in any of the previously published STR genotyping pipelines, greatly improve the speed, cost and accuracy of allele determinations.

Material and methods

Chimpanzee fecal samples

Fecal samples were collected from wild-living chimpanzees in Gombe National Park, including members of the Mitumba, Kasekela and Kalande communities, as well as the Greater Mahale Ecosystem (GME) in Tanzania as previously described (Keele *et al.* 2009; Rudicell *et al.* 2010; Rudicell *et al.* 2011; Santiago *et al.* 2003). Habituated Gombe chimpanzees have been under direct observation since the 1960s (Pusey *et al.* 2007; van Lawick-Goodall 1968), with prospective fecal sampling and SIVcpz diagnostics initiated in 1999 (Keele *et al.* 2009; Rudicell *et al.* 2010). Long-term monitoring of non-habituated chimpanzees in the GME began in 2008, with non-invasive SIVcpz screening implemented in 2009 (Rudicell *et al.* 2011). Gombe and GME fecal samples were collected 1:1 (vol/vol) in RNA*later* (Ambion), a high salt solution that preserves nucleic acids and allows storage and transport at room temperature. For individual identification, samples were routinely subjected to mitochondrial, sex, and microsatellite analyses, with up to eight STR loci characterized by capillary electrophoresis as described previously (Keele *et al.* 2009; Rudicell *et al.* 2010; Rudicell *et al.* 2011). All fieldwork has been approved by the Tanzania National Parks, the Tanzania Commission for Science and Technology, the Tanzania Wildlife Research Institute, and has followed the American Society of Primatologists' Principles for Ethical Treatment of Nonhuman Primates.

Quantification of chimpanzee DNA

Fecal DNA was extracted from 0.5 ml of homogenized fecal suspension using the QIAamp DNA Stool Kit and the automated QIAcube system (Qiagen). Purified DNA was eluted in 200 µl water and stored at -20 °C. Chimpanzee genomic DNA content was determined using a previously described *c-myc* gene-based quantitative (q)PCR (Morin *et al.* 2001). Briefly, 2 µl DNA extract was added to 1x High Fidelity PCR Buffer, 3.5 mM $MgSO_4$, 0.3 µM forward (5'-GCCAGAGGAGGAACGAGCT-3') and reverse (5'-GGGCCTTTTCATTGTTTTCCA-3') qPCR primers, 0.2 µM of a FAM-labeled probe (FAM-TGCCCTGCGTGACCAGATCC-BHQ1), 0.2 mM dNTPs, 1x ROX Reference Dye, and 0.5 U Platinum Taq DNA Polymerase High Fidelity (Invitrogen). Each sample was run in triplicate on a 7900HT Fast Real-Time PCR System, together with human genomic DNA standards of known concentration (the sequence of the particular *c-myc* amplicon is identical between humans and chimpanzees). Negative "no-template" controls were included in each run. Sequence Detection Systems version 2.3 software (Applied Biosystems) was used to quantify the host DNA content of each sample. Since host DNA concentrations differed, approximately half of the samples were extracted on more than one occasion to generate enough material for all analyses.

Amplification of STR loci

Previous genotyping studies of Gombe and GME chimpanzees utilized eight STR loci containing tetranucleotide repeats (Constable *et al.* 2001; Keele *et al.* 2009; Rudicell *et al.* 2011). These included D18s536 (also termed locus A), D4s243 (locus B), D10s676 (locus C), D9s922 (locus D), D2s1326 (locus 1) D2s1333 (locus 2), D4s1627 (locus 3), and D9s905 (locus 4) (Table S1). To facilitate MiSeq sequencing of the amplified loci, we

added MiSeq-specific adapters to the 5' end of both the forward (5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG-3') and the reverse primer (5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG-3'), respectively. Individual STR loci were amplified using 3 - 5 µl fecal DNA extract, 2.5 µl 10x AmpliTaq Gold Buffer, 1.75 µl 25mM $MgCl_2$, 1.5 µl 10 mM dNTPs, 0.5 µl 50 µg/ml BSA, 1.5 µl of 10 mM forward and reverse primers, and 0.25 µl AmpliTaq Gold polymerase (5U/ml; Applied Biosystems) in a 25 µl reaction volume. Thermocycling was performed using an initial denaturation for 10 minutes at 94 °C, followed by 50 cycles of 30 seconds at 94 °C, 30 seconds at 54 °C, and 45 seconds at 72 °C, followed by a final extension of 10 minutes at 72 °C.

Testing the sensitivity of MiSeq derived allele detection, we found that individual PCR reactions often produced only partial genotypes, while the combination of multiple amplicons from the same DNA sample generally yielded a more complete set of alleles. Consistent with previous studies (Morin *et al.* 2001), we also found that PCR amplification of less than 25 pg of host DNA generally failed to amplify STR loci. For all genotyping analyses, we thus included only DNA samples that contained more than 25 pg of chimpanzee DNA, amplified each STR locus on three independent occasions, and combined equal volumes of these replicate PCR reactions prior to MiSeq sequencing.

The eight STR loci were also amplified in one-step and two-step multiplex reactions. To minimize primer-primer interactions, locus A, B, C and 3 primers were combined at an even ratio in one pool, while locus D, 1, 2, 4 primers were similarly combined in a second pool. Fecal DNA was then amplified in two (rather than eight) different reactions, using the identical cycling conditions as for singleplex PCR. For two-step multiplexing, 2 µl of a 1:100 dilution of the one-step product were used as a template for a second round of PCR in which each locus was amplified individually using the same thermocycling conditions (Arandjelovic *et al.* 2009).

Library preparation and MiSeq sequencing

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

CHIIMP analysis pipeline

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

trimmed using cutadapt (Martin 2011). The adapter trimmed forward reads from each read pair, which covered the entire STR amplicon, were then imported into the R package, which was used for all downstream analyses.

The CHIIMP analysis pipeline generates multi-locus genotypes in three stages. First, each MiSeq sequence file is processed into unique sequences with relevant attributes, such as read counts, sequence length, and whether the sequence matches the locus-specific forward primer, repeat motifs and length range. Sequences are also queried for potential PCR artifacts, such as single nucleotide substitutions, indels, and stutter sequences introduced by Taq polymerase and sequencing errors. These artifacts are identified as comprising less than one third of the read counts of the corresponding allele. The 33% threshold was selected because inspection of known heterozygous loci revealed that all of the true second most frequent alleles contained more than that proportion of reads. Finally, for each sample and locus the proportion of sequence reads of the total read count is determined. At this stage, data are kept for all loci to ensure flexible downstream processing, such as detecting cross-locus contamination.

The second stage removes all sequences that do not match the locus attributes, such as the forward primer, repeat motif, and locus length, and/or contain likely PCR artifacts. In addition, only sequences comprising a minimum fraction of the total number of filtered reads (5%) are retained, and only loci with a total filtered read count above a customizable per-sample read threshold (>500) are genotyped. Application of these filters determines the sample zygosity; if only one sequence passes these filters, the locus is reported as homozygous. However, if two or more sequences pass the filters, the two most abundant are kept and the sample is reported as heterozygous. The output at this stage includes a spreadsheet with the sequence content, read counts, sequence lengths, as well as other relevant information such as whether the sequence contains the correct repeat motif or was identified as a likely stutter sequence or other PCR artifact. Of note, all

filters and thresholds are customizable, with the above parameters representing the default.

In the final stage, genotypes are assembled for all samples and loci, with quality control tables generated as output files (Fig. S1). First, a summary genotype table is generated that lists sample designations for each row, STR loci for each column, and unique allele identifiers for each cell (Fig. S1a). If specific allele codes are provided, the summary table will include these designations. If an allele does not match previous identifiers, the software will create a short name based on sequence length and content to identify these new alleles (e.g., sample 4781, locus C, allele 2 in Fig. S1a). The similarity of genotypes is also depicted in a heatmap (Fig. 1b), which groups closely related genotypes (Peakall & Smouse 2006). In cases where genotypes of individuals are known, the program links samples with the corresponding individuals (Fig. S1c). A heatmap shows the extent of similarity of every sample with every known genotype, thus allowing simple individual identification (Fig. S1d). The program also generates a set of tables that flag alleles that require additional attention, such as loci where the stutter filter has been invoked, where more than two sequences passed the filter, where a large proportion of sequences was not contained in the identified alleles, and where homozygosity may reflect allelic dropout (Fig. S1e). For each locus, the program creates a FASTA file of all allele sequences and an image of their alignment (Fig. S1f) generated by the Bioconductor's MSA package (Bodenhofer *et al.* 2015). In addition, a heatmap of sequence counts that match the locus-specific forward primer for all samples and loci is generated (Fig. S1g). For singleplex samples, this identifies sequences that match other loci and thus highlights potential cross-locus contamination. For multiplexed samples, this shows the read distribution across different loci. Finally, histograms that show sequence length-frequency distributions are saved as image files (Fig. S1h). A summary file is

created that combines all key results (sequences, read counts, etc.) for alleles for all samples and loci. This data output file is suitable for further analysis in R.

The new analysis platform, termed *Computational High-throughput Individual Identification through Microsatellite Profiling* or CHIIMP, has been designed to allow customization of the number and sequence content of microsatellite loci to be analyzed. Particular locus attributes such as the expected locus length range, primer sequences, and repeat motif sequence can be specified in a simple text file. Thus, the software can be readily adapted to additional microsatellite loci, as long as the respective amplicons fall within the length limits of the particular sequence chemistry used. The software is also suitable to analyze multiplexed samples, which contain reads from several loci but are processed separately, again using the forward primers to select locus-specific reads. No additional software is required other than providing a list of samples and loci prior to analysis. CHIIMP is available at https://github.com/ShawHahnLab/chiimp and can be installed on any Windows, Mac OS, or Linux computer with a standard installation of R and RStudio in a single step. On Windows, a desktop shortcut to the analysis script is provided. Dragging a simple text file containing analysis options onto the shortcut triggers analysis with the selected options. In addition to the standalone program, all features can also be used individually from within R. A comprehensive user guide including examples of analysis options and locus attributes is provided with the software.

Error, diversity, and heterozygosity calculations

Error rates for the MiSeq derived genotypes were calculated by determining the number of allelic mismatches for each sample to the known genotype of the corresponding chimpanzee (including allelic dropout, stutter sequences, PCR/sequencing artifacts, and locus amplification failure) and by dividing the total number of alleles by the number of

erroneous alleles (Broquet and Petit, 2004). The expected heterozygosity (also termed gene diversity) for the sampled Gombe and GME chimpanzees was calculated from both capillary electrophoresis and MiSeq based microsatellite data as described in Charlesworth & Charlesworth 2010. Allelic diversity was calculated by summing the total number of unique alleles in a population.

Results

Direct comparison of MiSeq and capillary electrophoresis based STR genotyping

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

Using the consensus genotype of the corresponding chimpanzees for reference (Table 1), we found that MiSeq genotyping reduced the number of allelic dropouts by more than half (Table 2). This was due, at least in part, to the pooling of PCR replicates,

which increased the number of alleles that were detected. However, MiSeq genotyping was also more accurate than the traditional method, which could not differentiate off-target amplifications (Tables 1 and 2). In addition, stutter peaks were completely eliminated by the CHIIMP analysis pipleline, which was not the case for the automated capillary electrophoresis method. Although manual peak calling also eliminated stutter peaks, this was considerably more time consuming than the MiSeq approach. For the 19 samples, conventional peak calling and allele binning took two hours, while reviewing the bioinformatics outputs took minutes. Most importantly, MiSeq genotyping identified eight heterozygous loci that were scored as homozygous by capillary electrophoresis because of a failure to resolve minor sequence and length (1bp) differences (Fig. 1). These sequence variants were readily identified in the read histograms (Fig. 1a) and their frequency identified in sequence alignments of the entire locus (Fig. 1b and c). Inspection of allele lengths across all loci revealed that 24% of all MiSeq derived alleles did not differ by multiples of four, indicating frequent nucleotide insertions and deletions in the tetranucleotide repeats (Fig. 1b and c).

MiSeq genotyping uncovers increased allelic diversity and heterozygosity

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

per locus. Nearly half of all previously reported alleles had closely related length or sequence variants (Table 4).

Although the great majority of the newly identified alleles were found in multiple individuals, we wanted to validate their authenticity by demonstrating inheritance. Since paternity and kinship relationships are known for most Gombe chimpanzees, we were able to trace the majority of the newly identified allelic variants from parents to their offspring. For example, Locus 3 includes four alleles that are identical in size (234 bp) but differ by up to three substitutions and two single nucleotide insertions and deletions (Fig. 2a). Alleles 234-a, 234-b, 234-c, and 234-d were found in 80, 25, 10 and 4 chimpanzees, respectively, including several parent-offspring triads (Fig. 2b). Overall, we were able to document inheritance for 25 (81%) of the 31 new alleles. For the remaining 6 existing pedigree information was insufficient, and their existence was thus confirmed by sequencing at least two independent PCR amplicons (Table 4).

The newly identified alleles revealed that over a quarter of genotypes at loci previously assigned as homozygous (60 of a total of 228) were in fact heterozygous (Table S2). This increased allelic diversity resolved one case of an ambiguous paternity determination. Using the standard eight STR loci, we were previously unable to identify the father of one infant (Google) because two candidate males (Faustino and Londo) had the identical genotype at all eight STR loci (Walker *et al.* 2017). Using the new genotypes, we were able to exclude Londo and confirm Faustino as a father by revealing differences at one locus (Fig. 2c). Although Faustino was identified as the correct father at the time by genotyping 10 additional loci using capillary electrophoresis (Walker *et al.* 2017), this would not have been necessary had the increased allelic diversity been known. Thus, MiSeq genotyping revealed much greater allelic and microsatellite gene diversity in Gombe than previously appreciated, thus increasing the analytical potential of the existing STR loci.

MiSeq genotyping based individual identification

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

STR genotyping of multiplexed samples

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

the exception of the Issa community, are not habituated. As a result, fecal collections, sample transport and storage are logistically more difficult, which can result in reduced amounts of collected material and/or partially degraded host DNA. To test the suitability of MiSeq genotyping for such samples, we selected 12 previously characterized chimpanzee fecal specimens from the Issa Valley (Rudicell *et al.* 2011) and re-genotyped them using both singleplex and multiplex locus amplification. Singleplex PCR was performed as in Gombe, while multiplex PCR was carried out in two steps as previously described (Arandjelovic *et al.* 2009). First, PCR primers for 4 loci were pooled and used to amplify fecal DNA in two (rather than eight) reactions (one-step multiplex product). Second, aliquots of this first round PCR were then used in a second round of PCR to amplify each of the 8 STR loci separately (two-step multiplex product). Three pooled replicates of both one-step and two-step multiplexed products were sequenced and compared to the previously determined genotypes (Table S3). Although the overall amplification efficiency was lower than originally reported (most likely due to repeat freezing and thawing of the 7-8 year old samples), one-step multiplexing performed as well as singleplex PCR, but used only a quarter of the fecal DNA (Table 5). Two-step multiplexing detected slightly more alleles, but not surprisingly, also resulted in an increased number of stutter sequences and other PCR artifacts. Thus, one-step multiplexing required less starting material and was also more cost efficient because the combined loci were sequenced in a single MiSeq run (and were subsequently de-multiplexed bioinformatically).

MiSeq genotyping also allowed us to compare the allelic diversity in Gombe and the GME. Fig. 4 depicts such an analysis for locus B and D, highlighting alleles that were only found in GME chimpanzees. Comparing all eight STR loci, we found ten alleles in only 12 GME chimpanzees that were absent from the 123 genotyped Gombe individuals, six of which represented alleles previously missed in the GME due to sequence and length differences. Although the mean expected heterozygosity value for the GME

chimpanzees (0.743) was lower than that for Gombe (0.812), this is likely due to the small sample size and the fact that all 12 individuals were sampled at a single location in Issa Valley (Rudicell *et al.* 2011). Additional samples from more diverse locations in the GME are needed to compare the genetic diversity of this population to that of Gombe and other field sites.

Discussion

Over the past two decades, microsatellite analyses have been an integral part of studies of wild chimpanzees, providing insight into their evolution, population genetics, behavior, disease association and social structure (Barbian *et al.* 2018; Becquet *et al.* 2007; Keele *et al.* 2009; Langergraber *et al.* 2007; Moeller *et al.* 2016; Rudicell *et al.* 2010; Santiago *et al.* 2003; Vigilant *et al.* 2001; Walker *et al.* 2017; Wroblewski *et al.* 2015). However, traditional genotyping methods are cumbersome, imprecise and investigator/platform dependent, due to the use of capillary electrophoresis to determine the length of STR loci. Here, we report a high-throughput MiSeq-based approach, which represents a marked improvement, because it is faster, more accurate and able to detect the full extent of allelic diversity in a population. Moreover, it includes a new analysis platform, CHIIMP, which not only automates the conversion of raw MiSeq data into multi-locus genotypes, but also implements a number of quality control measures that improve genotyping accuracy (Fig. 5). Of note, CHIIMP has been designed for maximal customization. While analysis of pedigreed fecal samples from chimpanzees allowed rigorous validation, the pipeline is not limited to a particular species or sample type.

Improved accuracy of MiSeq based genotyping

Sequence-based genotyping methods not only determine the length of STR loci, but also reveal their sequence content, and thus have the potential to detect a greater number of distinct alleles than capillary electrophoresis. Indeed, such genotyping of Atlantic cod and muskrats revealed high proportions of cryptic alleles, ranging from 32% to 44% (Darby *et al.* 2016; Vartia *et al.* 2016). In light of these data, our discovery of 38% new alleles (31 of 82) in Gombe is not surprising (Table 3). However, this finding suggests that existing STR data vastly underestimate the diversity of microsatellite sequences in wild chimpanzees, not only in Gombe but also in other populations. New alleles were found for all loci, with some comprising twice as many variants as previously observed (Table 3), which will undoubtedly add to the statistical power of future analyses. However, any new allele will have to be examined carefully by repeat amplification and sequencing, unless it can be validated by pedigree analysis. In our dataset, a minor fraction of "new" alleles were found to represent PCR and/or sequencing artifacts that exceeded the 33% threshold for heterozygous alleles. Repeat amplification of these alleles resolved all sequencing artifacts.

Comparison of the MiSeq data to validated reference genotypes also allowed us to assess the error rate of the new approach. After implementation of all filters, CHIIMP eliminated 98% of stutter sequences and 100% of off-target amplicons. Among the samples tested, true alleles, allelic dropouts and false alleles were detected with a frequency of 96%, 7%, and 0%, respectively. These data are comparable to MiSeq derived genotyping results for wild-living brown bears, where true alleles, allelic dropouts and false alleles were detected with a frequency of 93%, 0.4% and 0.05% for tissues, and 80%, 14% and 1% for fecal samples, respectively (De Barba *et al.* 2017). Although our overall error rate of 3.3% is slightly higher than the 2.1% error rate reported for a MiSeq genotyping study of laboratory raised (pedigreed) fish (Zhan *et al.* 2017), this is not surprising since the latter study examined freshly extracted tissue DNA.

Since non-invasively collected samples frequently contain diluted and/or degraded host DNA, they are genotyped using multiple PCR reactions to guard against the selective loss of alleles (allelic dropout). Loci are only considered homozygous if they can be confirmed in multiple PCR reactions (Morin *et al.* 2001; Taberlet *et al.* 1996). Capillary electrophoresis requires that these replicates are run independently to distinguish true alleles from non-specific signal, often more than tripling the amount of time and effort required to genotype a single sample. In contrast, MiSeq genotyping can be performed after combining the products of multiple PCR reactions. Although the quality of the input DNA remains the same, MiSeq genotyping of pooled PCR replicates reduces the frequency of allelic dropout and thus renders the resulting genotypes more accurate. However, amplicon pooling foregoes data from repeat analyses, which are used by some as a measure of DNA quality and/or data reliability (Taberlet *et al.* 1996)

Once MiSeq data files are imported into the CHIIMP platform, the program calls alleles automatically, thus saving days of hands-on work. While automated allele calling has been reported previously (De Barba *et al.* 2017; Suez *et al.* 2016; Zhan *et al.* 2017), CHIIMP includes downstream analyses, such as alignments of allele sequences or flagging loci that may contain contaminants, which provide important additional quality control measures. In contrast to previous studies, CHIIMP also retains non-repeat regions (Suez *et al.* 2016), which can contribute to allelic diversity, and does not require the presence of stutter sequences for allele calling, which may not be sufficiently abundant under conditions of low coverage (De Barba *et al.* 2017). Finally, CHIIMP reports both allele length and sequence content, and is thus designed to detect minor length and sequence differences by including sequence-specific allele names and generating locus-specific sequence alignments (Figs. 4 and S1). To guide subsequent analyses, we have also added features that flag potentially problematic alleles and standardize allele naming.

CHIIMP thus represents the most comprehensive analysis platform yet to ensure the accuracy of MiSeq-based genotyping results.

Multiplexing improves MiSeq genotyping efficiency and reduces cost

The Illumina MiSeq v2 500 sequencing kit has an output of ~25 million reads per run, thus allowing the multiplexing of many samples, the number of which depends on the desired read depth. Comparing read depths per allele, we found that a cut-off of 500 reads yielded the most accurate results for our dataset. This value is higher than the 50 read cut-off used previously to genotype laboratory raised fish (Zhan *et al.* 2017). However, the latter study used high quality tissues rather than fecal samples for analysis. To determine the sources of allele-calling errors, we did not multiplex samples from Gombe chimpanzees. However, we tested multiplexing using samples from GME chimpanzees and confirmed that this approach yields accurate results. Although primer incompatibilities allowed the combination of only four loci, this number can be significantly increased with additional primer design. For example, a recent study genotyped bear fecal DNA by multiplexing 14 loci (De Barba *et al.* 2017). Pooling amplicons from multiple loci after singleplex PCR can circumvent the need for specialized primer design, as the maximum number of pooled loci for any given sample is limited only by the desired read depth. Moreover, barcoding of individual samples allows their combination in sequencing reactions, thus further increasing sequencing efficiency and throughput (Farrell *et al.* 2016).

MiSeq genotyping is expensive, but these costs decrease with sample numbers. Capillary electrophoresis is undoubtedly cheaper when only a small number of samples has to be analyzed; however, MiSeq sequencing becomes increasingly more cost-effective with multiplexing and analyses of pooled replicates (Darby *et al.* 2016). The costs of MiSeq genotyping three replicates of 96 samples multiplexed at four loci would roughly

be equivalent to analyzing the same multiplexed samples via capillary electrophoresis, because the latter cannot analyze pooled replicates. While this estimate only considers genotyping supplies, labor to manually analyze samples is not included. In addition, the improved accuracy has downstream cost advantages since fewer repeat analyses would have to be performed.

Effective sharing of MiSeq genotyping data

A direct comparison of MiSeq and capillary electrophoresis derived alleles revealed consistent length differences of one to three nucleotides, the number of which were locus specific (Tables 1 and S3). For Gombe samples, locus 3 alleles derived by capillary electrophoresis were always three nucleotides longer than the corresponding MiSeq alleles (Table 1). However, for the GME samples, the same alleles were all one nucleotide shorter than the MiSeq alleles (Table S3). This is as expected since the capillary electrophoresis data were generated on different platforms. However, this also means that a simple conversion of existing capillary electrophoresis to MiSeq data will generally not be possible. In contrast, MiSeq genotyping generates unambiguous alleles that can be compared across multiple studies and field sites (Fig. 4). In the future, it will thus be possible to compare STR genotypes across different chimpanzee populations, such as those in Gombe and the GME, since the use of different sequencing equipment will no longer confound these analyses.

Versatility of the CHIIMP analysis platform

To increase its utility, we designed the CHIIMP analysis platform to be versatile. STR locus attributes, such as the expected length range, primer sequences, and repeat motifs,

as well as all thresholds for allele calling can be customized. For example, analysis of loci with dinucleotide repeats may require a lower threshold for stutter peaks, since these are more susceptible to polymerase slippage (Guichoux *et al.* 2011). Similarly, locus length ranges can be expanded or contracted, depending on the rate of off-target amplicons. The CHIIMP analysis pipeline also includes tools that facilitate iterative improvements for new applications (Fig. S1). For example, the program provides a heatmap that indicates the number of unique sequences that pass all filters. If that number is too high, thresholds can be adjusted to remove stutter peaks, off-target amplicons, and/or PCR errors. In addition, the distribution of loci is visualized, which can be used to reveal contamination in singleplexed samples or identify poorly performing primers in multiplexed samples (Fig. S1g). For potentially problematic alleles, CHIIMP generates histograms that provide information concerning their length and relative abundance. All of these tools can be used to adapt the platform to additional STR loci and/or host species.

The length of STR loci suitable for sequence-based genotyping depends on the sequencing chemistry. We used Illumina v2 technology, which has maximum read lengths of 500 nucleotides. MiSeq sequences are most often generated using paired-end reads, with a maximum read length of 250 nucleotides in each direction. For STR genotyping, locus sequences must span the repeat motif region, since assembly of shorter reads could result in misalignments. To accommodate loci of greater than 250 bp length, we opted to only use forward reads for analysis. Although the sequencing kit could theoretically accommodate fragments of up to 500 nucleotides, we found that the quality of reads (Q scores) decreased significantly after 400 cycles. Given that the longest locus in our panel spanned 357 nucleotides, we used 375 cycles in the forward direction. Illumina v3 sequencing chemistry has a 600-cycle limit, which may accommodate loci of up to 500 bp, but this would have to be determined experimentally. The majority of microsatellite loci are shorter than this length.

As STR genotyping transitions from capillary electrophoresis to sequence based approaches, it will be necessary to standardize allele nomenclature, as has already been suggested for human forensics (Gelardi *et al.* 2014; Parson *et al.* 2016). At a minimum, allele names will have to incorporate the length and unique sequence content for each allele (Darby *et al.* 2016). In our study, we added an alphabetical identifier (-a, -b, -c, etc.) to differentiate identically sized alleles that differed in their sequence (Table S2). Since it is impossible to capture all allele attributes in a single name, it may become necessary to establish databases that link allele identifiers to their respective sequences. CHIIMP is designed to allow users to supply a spreadsheet of allele names and sequences, and thus guarantees consistent nomenclature across experiments. As MiSeq genotyping is adapted to additional projects, standardized allele designations will become necessary to ensure consistent nomenclature across studies.

Conclusions

Genetic study of wild primates and other endangered species has been shown to provide more accurate information concerning the size, structure, distribution and dynamics of populations than observational studies. However, genotyping can be prohibitively expensive given the large numbers of samples that are required for such analyses. The MiSeq based genotyping platform provides a new approach that drastically reduces time and labor, while providing more accurate and informative genotypes compared to capillary electrophoresis. This will allow much faster and more streamlined analysis of samples that are necessary for censusing and monitoring of non-habituated populations in addition to revealing previously inaccessible allelic diversity. The CHIIMP platform has been designed to be adaptable to additional loci and/or species. This allows the study of group

membership, dispersal, gene flow, and association patterns for a multitude of wildlife species with broad conservation and biological implications.

Author contributions

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

Acknowledgements

We thank the Jane Goodall Institute field staff at the Gombe Stream Research Centre as well as field assistants from the Greater Mahale Ecosystem Research and Conservation Project (GMERC) for collecting chimpanzee observational data as well as fecal samples; the Tanzania Commission for Science and Technology (COSTECH), the Tanzania Wildlife Research Institute (TAWIRI), and the Tanzania National Parks Association (TANAPA) for their support and permission to conduct research in Gombe and the GME. This work was supported by grants from the National Institutes of Health, USA (R01 AI 091595, R37 AI 050529, R01 AI 120810, P30 AI 045008), the National Science Foundation (IOS-1052693, IOS-1457260), the Jane Goodall Institute, and the University of California at San Diego/Salk Center for Academic Research and Training in Anthropogeny (CARTA). H.J.B.

and R.M.R. were funded by training grants (T32 AI 055400 and T32 AI 007632, respectively). The authors declare no competing financial interests.

Data accessibility

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

References

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

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

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

Balloux F, Lugon-Moulin N (2002) The estimation of population differentiation with microsatellite markers. *Molecular Ecology* **11**, 155-165.

Barbian HJ, Li Y, Ramirez M*, et al.* (2018) Destabilization of the gut microbiome marks the end-stage of simian immunodeficiency virus infection in wild chimpanzees. *American Journal of Primatology* **80**, 10.1002/ajp.22515.

Becquet C, Patterson N, Stone AC, Przeworski M, Reich D (2007) Genetic structure of chimpanzee populations. *PLOS Genetics* **3**, e66.

Bennett P (2000) Demystified …: Microsatellites. *Molecular Pathology* **53**, 177-183.

Bodenhofer U, Bonatesta E, Horejš-Kainrath C, Hochreiter S (2015) msa: an R package for multiple sequence alignment. *Bioinformatics* **31**, 3997-3999.

Broquet T, Petit E (2004) Quantifying genotyping errors in noninvasive population genetics. *Molecular Ecology* **13**, 3601-3608.

Charlesworth B, Charlesworth D (2010) *Elements of evolutionary genetics*. Roberts and Co. Publishers. 734 p.

Constable JJ, Packer C, Collins DA, Pusey AE (1995) Nuclear DNA from primate dung. *Nature* **373**, 393.

Constable JL, Ashley MV, Goodall J, Pusey AE (2001) Noninvasive paternity assignment in Gombe chimpanzees. *Molecular Ecology* **10**, 1279-1300.

Darby BJ, Erickson SF, Hervey SD, Ellis-Felege SN (2016) Digital fragment analysis of short tandem repeats by high-throughput amplicon sequencing. *Ecology and Evolution* **6**, 4502-4512.

De Barba M, Miquel C, Lobréaux S*, et al.* (2017) High-throughput microsatellite genotyping in ecology: improved accuracy, efficiency, standardization and success with low-quantity and degraded DNA. *Molecular Ecology Resources.* **17**, 492-507.

Ellegren H (2004) Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetics* **5**, 435.

Ewen KR, Bahlo M, Treloar SA*, et al.* (2000) Identification and analysis of error types in high-throughput genotyping. *American Journal of Human Genetics* **67**, 727-736.

Farrell ED, Carlsson JEL, Carlsson J (2016) Next Gen Pop Gen: implementing a high-throughput approach to population genetics in boarfish (Capros aper). *Royal Society Open Science* **3**, 160651.

Fernando P, Evans BJ, Morales JC, Melnick DJ (2001) Electrophoresis artefacts — a previously unrecognized cause of error in microsatellite analysis. *Molecular Ecology Notes* **1**, 325-328.

Fordyce SL, Ávila-Arcos MC, Rockenbauer E*, et al.* (2011) High-throughput sequencing of core STR loci for forensic genetic investigations using the Roche Genome Sequencer FLX platform. *BioTechniques* **51**, 127-133.

Gelardi C, Rockenbauer E, Dalsgaard S, Børsting C, Morling N (2014) Second generation sequencing of three STRs D3S1358, D12S391 and D21S11 in Danes and a new nomenclature for sequenced STR alleles. *Forensic Science International: Genetics* **12**, 38-41.

Guichoux E, Lagache L, Wagner S*, et al.* (2011) Current trends in microsatellite genotyping. *Molecular Ecology Resources* **11**, 591-611.

Hauge XY, Litt M (1993) A study of the origin of 'shadow bands' seen when typing dinucleotide repeat polymorphisms by the PCR. *Human Molecular Genetics.* **2**, 411-415.

Iyer SS, Bibollet-Ruche F, Sherrill-Mix S*, et al.* (2017) Resistance to type 1 interferons is a major determinant of HIV-1 transmission fitness. *Proceedings of the National Academy of Sciences* **114**: E590-E599.

Jarne P, Lagoda PJL (1996) Microsatellites, from molecules to populations and back. *Trends in Ecology & Evolution* **11**, 424-429.

Keele BF, Jones JH, Terio KA*, et al.* (2009) Increased mortality and AIDS-like immunopathology in wild chimpanzees infected with SIVcpz. *Nature* **460**, 515-519.

Kelkar YD, Strubczewski N, Hile SE*, et al.* (2010) What is a microsatellite: a computational and experimental definition based upon repeat mutational behavior at A/T and GT/AC repeats. *Genome Biology and Evolution* **2**, 620-635.

Langergraber KE, Mitani JC, Vigilant L (2007) The limited impact of kinship on cooperation in wild chimpanzees. *Proceedings of the National Academy of Sciences* **104**, 7786-7790.

Levinson G, Gutman GA (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Molecular biology and evolution.* **4**, 203-221.

Lynch HT, de la Chapelle A (2003) Hereditary Colorectal Cancer. *New England Journal of Medicine* **348**, 919-932.

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10-12.

Moeller AH, Foerster S, Wilson M*, et al.* (2016) Social behaviour promotes diversity in the chimpanzee gut microbiome. *American Association for the Advancement of Science* **2**, e1500997.

Moore J (1996) Savanna chimpanzees, referential models and the last common ancestor. *Great Ape Societies*. Cambridge University Press, 275-292.

Morin PA, Chambers KE, Boesch C, Vigilant L (2001) Quantitative polymerase chain reaction analysis of DNA from noninvasive samples for accurate microsatellite genotyping of wild chimpanzees (Pan troglodytes verus). *Molecular Ecology* **10**, 1835-1844.

Morin PA, Wallis J, Moore JJ, Chakraborty R, Woodruff DS (1993) Non-invasive sampling and DNA amplification for paternity exclusion, community structure, and phylogeography in wild chimpanzees. *Primates* **34**, 347-356.

751 Ogawa H, Idani Gi, Kanamori M (1999) Chimpanzee habitat in the savanna woodland,
752 Ugalla, Tanzania. *Primate Research* **15**, 135-146.
753 Parson W, Ballard D, Budowle B*, et al.* (2016) Massively parallel sequencing of forensic
754 STRs: Considerations of the DNA commission of the International Society for
755 Forensic Genetics (ISFG) on minimal nomenclature requirements. *Forensic*
756 *Science International: Genetics* **22**, 54-63.
757 Pasqualotto AC, Denning DW, Anderson MJ (2007) A cautionary tale: lack of consistency
758 in allele sizes between two laboratories for a published multilocus microsatellite
759 typing system. *Journal of Clinical Microbiology* **45**, 522-528.
760 Peakall ROD, Smouse PE (2006) genalex 6: genetic analysis in Excel. Population genetic
761 software for teaching and research. *Molecular Ecology Notes* **6**, 288-295.
762 Pusey AE, Pintea L, Wilson ML, Kamenya S, Goodall J (2007) The contribution of long-
763 term research at Gombe National Park to chimpanzee conservation. *Conservation*
764 *Biology* **21**, 623-634.
765 Queller DC, Strassmann JE, Hughes CR (1993) Microsatellites and kinship. *Trends in*
766 *Ecology & Evolution* **8**, 285-288.
767 Rudicell RS, Holland Jones J, Wroblewski EE*, et al.* (2010) Impact of simian
768 immunodeficiency virus infection on chimpanzee population dynamics. *PLoS*
769 *Pathogens* **6**, e1001116.
770 Rudicell RS, Piel AK, Stewart F*, et al.* (2011) High prevalence of simian immunodeficiency
771 virus infection in a community of savanna chimpanzees. *Journal of Virology* **85**,
772 9918-9928.
773 Santiago ML, Lukasik M, Kamenya S*, et al.* (2003) Foci of endemic simian
774 immunodeficiency virus infection in wild-living eastern chimpanzees (*Pan*
775 *troglodytes schweinfurthii*). *Journal of Virology* **77**, 7545-7562.

Schoeninger MJ, Moore J, Sept JM (1999) Subsistence strategies of two “savanna” chimpanzee populations: The stable isotope evidence. *American Journal of Primatology* **49**, 297-314.

Schuelke M (2000) An economic method for the fluorescent labeling of PCR fragments. *Nature Biotechnology* **18**, 233.

Selkoe KA, Toonen RJ (2006) Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. *Ecology Letters* **9**, 615-629.

Shinde D, Lai Y, Sun F, Arnheim N (2003) Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)(n) and (A/T)(n) microsatellites. *Nucleic Acids Research* **31**, 974-980.

Suez M, Behdenna A, Brouillet S*, et al.* (2016) MicNeSs: genotyping microsatellite loci from a collection of (NGS) reads. *Molecular Ecology Resources* **16**, 524-533.

Taberlet P, Camarra JJ, Griffin S*, et al.* (1997) Noninvasive genetic tracking of the endangered Pyrenean brown bear population. *Molecular Ecology* **6**, 869-876.

Taberlet P, Griffin S, Goossens B*, et al.* (1996) Reliable genotyping of samples with very low DNA quantities using PCR. *Nucleic Acids Research* **24**, 3189-3194.

van Lawick-Goodall J (1968) The behaviour of free-living chimpanzees in the Gombe Stream Reserve. *Animal Behaviour Monographs* **1**, 161-311.

Van Neste C, Van Nieuwerburgh F, Van Hoofstat D, Deforce D (2012) Forensic STR analysis using massive parallel sequencing. *Forensic Science International.* **6**, 810-818.

Vartia S, Villanueva-Cañas JL, Finarelli J*, et al.* (2016) A novel method of microsatellite genotyping-by-sequencing using individual combinatorial barcoding. *Royal Society Open Science* **3**, 150565.

Vigilant L, Hofreiter M, Siedel H, Boesch C (2001) Paternity and relatedness in wild chimpanzee communities. *Proceedings of the National Academy of Sciences* **98**, 12890-12895.

Walker KK, Rudicell RS, Li Y*, et al.* (2017) Chimpanzees breed with genetically dissimilar mates. *Royal Society Open Science* **4**, 160422.

Weeks DE, Conley YP, Ferrell RE, Mah TS, Gorin MB (2002) A tale of two genotypes: consistency between two high-throughput genotyping centers. *Genome Res* **12**, 430-435.

Wroblewski EE, Norman PJ, Guethlein LA*, et al.* (2015) Signature patterns of MHC diversity in three gombe communities of wild chimpanzees reflect fitness in reproduction and immune defense against SIVcpz. *PLoS Biology* **13**, e1002144.

Zhan L, Paterson IG, Fraser BA*, et al.* (2017) megasat: automated inference of microsatellite genotypes from sequence data. *Molecular Ecology Resources* **17**, 247-256.

Figure legends

Fig 1 MiSeq genotyping uncovers cryptic alleles. Eight polymorphic STR loci were amplified from the fecal DNA of 19 previously genotyped chimpanzees. (a) Histogram depicting the length (x-axis) and read count (y-axis) of unique sequences for one representative heterozygous locus that was previously determined to be homozygous by multiple capillary electrophoresis analyses (sample 4861, locus C, Table 1). The grey box highlights the expected locus size range. The horizontal line indicates the cutoff of 500 reads. Colored peaks indicate reads that passed the locus-specific filters (note that peaks can be comprised of identically sized reads that differ in their sequence content). Black reads were eliminated. Pink reads appear to be locus-specific, but did not pass the PCR

artifact filters. Red reads represent the true allele sequences (180 and 181 bp in lengths, respectively). (b, c) Alignment images of locus-specific allele sequences are shown for locus 1 (b) and locus C (c), respectively (the complete data set is shown in Table 1). Allele sequences are ordered by length (indicated in bp on the right), with the frequency with which they were found in different chimpanzees indicated on the left (the x-axis indicates the position within the alignment). Nucleotides are colored as shown, with gaps in the alignment shown in grey. The insets highlight alleles that differ in their sequence content and/or length. Nucleotide substitutions are colored; dashes indicate gaps that were introduced to optimize the alignment.

Fig. 2 MiSeq genotyping uncovers increased allelic diversity and heterozygosity. (a) Alignment of four locus 3 alleles that are of identical length (234 bp), but differ in sequence content. Nucleotide substitutions are colored; dashes indicate single nucleotide insertions and deletions (b) Mendelian inheritance of allele 234 for a group of related chimpanzees. Fathers and mothers are shown as squares and circles, respectively, with offspring connected by vertical lines. Both alleles are shown for each animal, with the four allelic variants highlighted in different colors. Individuals of unknown identity or genotype are left blank. (c) Increased allelic diversity resolves a previously ambiguous paternity determination. Two potential fathers with identical allele lengths (238 bp) can now be distinguished based on differences in allele sequence content (238-a and 238-b). Since the offspring is homozygous for allele 238-a, the male with allele 238-b can be excluded as a father.

Fig. 3 Individual identification based on MiSeq genotyping. (a-c) Genotypes of newly collected samples (top) are compared to the genotypes of known community members, with the closest match listed below (based on descending distance scores). Genotypes

that differ by fewer than four alleles are indicated in bold because they represent likely matches. Differences are highlighted in yellow. (d) Heatmap showing the relative similarity of sample genotypes (rows) with genotypes of known individuals (columns) based on distance scores. Dark red cells indicate likely matches.

Fig 4 Comparison of MiSeq genotypes across chimpanzee communities. Alignment images of locus-specific allele sequences are shown for chimpanzees from the GME and Gombe. Two representative loci (locus B on the left; locus D on the right) are shown for (a) 12 chimpanzees from the GME (Table S3), (b) 123 chimpanzees from Gombe (Table S2), and (c) a combination of both. Allele sequences are ordered by length (indicated in base pairs on the right), with the frequency with which they were found in different chimpanzees indicated on the left (the x-axis indicates the position within the alignment). Nucleotides are colored as indicated, with alignment gaps shown in grey. Arrows indicate alleles that are unique to the GME samples.

Fig. 5 MiSeq-based STR genotyping of wild chimpanzees. (a) Schematic representation of singleplex STR amplification and MiSeq sequencing of chimpanzee fecal DNA. (b) Schematic representation of the CHIIMPS analysis pipeline with decision tree and downstream data reports.

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

Hannah J. Barbian[1], A. Jesse Connell[1], Alexa N. Avitto[1], Ronnie M. Russell[1], Andrew G. Smith[1], Madhurima S. Gundlapally[1], Alexander L. Shazad[1], Yingying Li[1], Frederic Bibollet-Ruche[1], Emily E. Wroblewski[2], Deus Mjungu[3], Elizabeth V. Lonsdorf[4], Fiona A. Stewart[5], Alexander K. Piel[5], Anne E. Pusey[6], Paul M. Sharp[7] and Beatrice H. Hahn[1*]

[1]Departments of Microbiology and Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

[2]Department of Anthropology, Washington University in St. Louis, St. Louis, MO, USA

[3]Gombe Stream Research Center, Kigoma, Tanzania

[4]Department of Psychology, Franklin and Marshall College, Lancaster, Pennsylvania

[5]School of Natural Sciences and Psychology, Liverpool John Moores University, Liverpool, United Kingdom

[6]Department of Evolutionary Anthropology, Duke University, Durham, North Carolina

[7]Institute of Evolutionary Biology and Centre for Immunity, Infection and Evolution, University of Edinburgh, Edinburgh EH9 3FL, United Kingdom

Professor, Departments of Medicine and Microbiology
University of Pennsylvania Perelman School of Medicine
409 Johnson Pavilion
3610 Hamilton Walk
Philadelphia, PA 19104-6076
USA
bhahn@pennmedicine.upenn.edu

Running title: High throughput STR genotyping of chimpanzee

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 ~~previously characterized~~ 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 analyse~~i~~s. 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 ~~fully~~ 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*

Introduction

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

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

artifacts, which are difficult to identify on electropherograms. These include stutter peaks, which are usually one repeat shorter than the correct STR allele and derive from Taq polymerase slippage (Hauge & Litt 1993; Shinde *et al.* 2003), split peaks which are caused by inconsistent A-overhang addition (Schuelke 2000), and artifactual peaks, which are the product of off-target amplification and/or unspecific fluorescent signaling (Ewen *et al.* 2000; Fernando *et al.* 2001; Guichoux *et al.* 2011). Existing peak-calling software often fails to differentiate erroneous from real peaks and frequently omits peaks of low height. Automatically called peaks must therefore be corrected manually, which is labor intensive and time consuming (Guichoux *et al.* 2011). Finally, multiplexing is restricted to only a few fluorescent labels, thus limiting the number of loci that can be analyzed simultaneously. As a consequence, capillary electrophoresis based STR genotyping is laborious, notoriously imprecise, and generally not useful for large sample sets or data sharing between different platforms and/or field sites (Pasqualotto *et al.* 2007).

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

visual inspection of read length histograms (Darby *et al.* 2016; Farrell *et al.* 2016; Vartia *et al.* 2016), and none of these studies have compared the performance of capillary electrophoresis and high throughput sequencing ~~directly~~ side-by-side to validate and improve the genotyping approach.

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

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

tools, which are not included in any of the previously published STR genotyping pipelines, greatly improve the speed, cost and accuracy of allele determinations. ~~Using this method to genotype fecal samples from two previously studied chimpanzee populations, we tested whether CHIIMP is superior to capillary electrophoresis in terms of intensity of time and labor, and accuracy of results that can be compared across field sites.~~

Material and methods

Chimpanzee fecal samples

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

Commission for Science and Technology, the Tanzania Wildlife Research Institute, and has followed the American Society of Primatologists' Principles for Ethical Treatment of Nonhuman Primates.

Quantification of chimpanzee DNA

Fecal DNA was extracted from 0.5 ml of homogenized fecal suspension using the QIAamp DNA Stool Kit and the automated QIAcube system (Qiagen). Purified DNA was eluted in 200 µl water and stored at -20 °C. Chimpanzee genomic DNA content was determined using a previously described *c-myc* gene-based quantitative (q)PCR (Morin *et al.* 2001). Briefly, 2 µl DNA extract was added to 1x High Fidelity PCR Buffer, 3.5 mM $MgSO_4$, 0.3 µM forward (5'-GCCAGAGGAGGAACGAGCT-3') and reverse (5'-GGGCCTTTTCATTGTTTTCCA-3') qPCR primers, 0.2 µM of a FAM-labeled probe (FAM-TGCCCTGCGTGACCAGATCC-BHQ1), 0.2 mM dNTPs, 1x ROX Reference Dye, and 0.5 U Platinum Taq DNA Polymerase High Fidelity (Invitrogen). Each sample was run in triplicate on a 7900HT Fast Real-Time PCR System, together with human genomic DNA standards of known concentration (the sequence of the particular *c-myc* amplicon is identical between humans and chimpanzees). Negative "no-template" controls were included in each run. Sequence Detection Systems version 2.3 software (Applied Biosystems) was used to quantify the host DNA content of each sample. Since host DNA concentrations differed, approximately half of the samples were extracted on more than one occasion to generate enough material for all analyses.

Amplification of STR loci

Previous genotyping studies of Gombe and GME chimpanzees utilized eight STR loci containing tetranucleotide repeats (Constable *et al.* 2001; Keele *et al.* 2009; Rudicell *et al.* 2011). These included D18s536 (also termed locus A), D4s243 (locus B), D10s676 (locus C), D9s922 (locus D), D2s1326 (locus 1) D2s1333 (locus 2), D4s1627 (locus 3), and D9s905 (locus 4) (Table S1). To facilitate MiSeq sequencing of the amplified loci, we added MiSeq-specific adapters to the 5' end of both the forward (5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG-3') and the reverse primer (5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG-3'), respectively. Individual STR loci were amplified using 3 - 5 µl fecal DNA extract, 2.5 µl 10x AmpliTaq Gold Buffer, 1.75 µl 25mM $MgCl_2$, 1.5 µl 10 mM dNTPs, 0.5 µl 50 µg/ml BSA, 1.5 µl of 10 mM forward and reverse primers, and 0.25 µl AmpliTaq Gold polymerase (5U/ml; Applied Biosystems) in a 25 µl reaction volume. Thermocycling was performed using an initial denaturation for 10 minutes at 94 °C, followed by 50 cycles of 30 seconds at 94 °C, 30 seconds at 54 °C, and 45 seconds at 72 °C, followed by a final extension of 10 minutes at 72 °C.

Testing the sensitivity of MiSeq derived allele detection, we found that individual PCR reactions often produced only partial genotypes, while the combination of multiple amplicons from the same DNA sample generally yielded a more complete set of alleles. Consistent with previous studies (Morin *et al.* 2001), ~~Each fecal DNA sample was amplified on three different occasions for each STR locus, with the resulting products pooled at equivolumeual proportions prior to MiSeq sequencing.~~ we also found that PCR amplification of less than 25 pg of host DNA generally failed to amplify STR loci. For all genotyping analyses, we thus included only DNA samples that contained more than 25 pg of chimpanzee DNA, amplified each STR locus on three independent occasions, and combined equal volumes of these replicate PCR reactions prior to MiSeq sequencing.

The eight STR loci were also amplified in one-step and two-step multiplex reactions. To minimize primer-primer interactions, locus A, B, C and 3 primers were

combined at an even ratio in one pool, while locus D, 1, 2, 4 primers were similarly combined in a second pool. Fecal DNA was then amplified in two (rather than eight) different reactions, using the identical cycling conditions as for singleplex PCR. For two-step multiplexing, 2 µl of a 1:100 dilution of the one-step product were used as a template for a second round of PCR in which each locus was amplified individually using the same thermocycling conditions (Arandjelovic *et al.* 2009).

Library preparation and MiSeq sequencing

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

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

CHIIMP analysis pipeline

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

The CHIIMP analysis pipeline generates multi-locus genotypes in three stages. First, each MiSeq sequence file is processed into unique sequences with relevant attributes, such as read counts, sequence length, and whether the sequence matches the locus-specific forward primer, repeat motifs and length range. Sequences are also queried for potential PCR artifacts, such as single nucleotide substitutions, indels, and stutter sequences introduced by Taq polymerase and sequencing errors. These artifacts are identified as comprising less than one third of the read counts of the corresponding allele. The 33% threshold was selected because inspection of known heterozygous loci revealed that all of the true second most frequent alleles contained more than that proportion of reads. Finally, for each sample and locus the proportion of sequence reads of the total read count is determined. At this stage, data are kept for all loci to ensure flexible downstream processing, such as detecting cross-locus contamination.

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

filtered reads (5%) are retained, and only loci with a total filtered read count above a customizable per-sample read threshold (>500) are genotyped. Application of these filters determines the sample zygosity; if only one sequence passes these filters, the locus is reported as homozygous. However, if two or more sequences pass the filters, the two most abundant are kept and the sample is reported as heterozygous. The output at this stage includes a spreadsheet with the sequence content, read counts, sequence lengths, as well as other relevant information such as whether the sequence contains the correct repeat motif or was identified as a likely stutter sequence or other PCR artifact. Of note, all filters and thresholds are customizable, with the above parameters representing the default.

In the final stage, genotypes are assembled for all samples and loci, with quality control tables generated as output files (Fig. S1). First, a summary genotype table is generated that lists sample designations for each row, STR loci for each column, and unique allele identifiers for each cell (Fig. S1a). If specific allele codes are provided, the summary table will include these designations. If an allele does not match previous identifiers, the software will create a short name based on sequence length and content to identify these new alleles (e.g., sample 4781, locus C, allele 2 in Fig. S1a). The similarity of genotypes is also depicted in a heatmap (Fig. 1b), which groups closely related genotypes (Peakall & Smouse 2006). In cases where genotypes of individuals are known, the program links samples with the corresponding individuals (Fig. S1c). A heatmap shows the extent of similarity of every sample with every known genotype, thus allowing simple individual identification (Fig. S1d). The program also generates a set of tables that flag alleles that require additional attention, such as loci where the stutter filter has been invoked, where more than two sequences passed the filter, where a large proportion of sequences was not contained in the identified alleles, and where homozygosity may reflect allelic dropout (Fig. S1e). For each locus, the program creates a FASTA file of all

allele sequences and an image of their alignment (Fig. S1f) generated by the Bioconductor's MSA package (Bodenhofer *et al.* 2015). In addition, a heatmap of sequence counts that match the locus-specific forward primer for all samples and loci is generated (Fig. S1g). For singleplex samples, this identifies sequences that match other loci and thus highlights potential cross-locus contamination. For multiplexed samples, this shows the read distribution across different loci. Finally, histograms that show sequence length-frequency distributions are saved as image files (Fig. S1h). A summary file is created that combines all key results (sequences, read counts, etc.) for alleles for all samples and loci. This data output file is suitable for further analysis in R.

The new analysis platform, termed *Computational High-throughput Individual Identification through Microsatellite Profiling* or CHIIMP, has been designed to allow customization of the number and sequence content of microsatellite loci to be analyzed. Particular locus attributes such as the expected locus length range, primer sequences, and repeat motif sequence can be specified in a simple text file. Thus, the software can be readily adapted to additional microsatellite loci, as long as the respective amplicons fall within the length limits of the particular sequence chemistry used. The software is also suitable to analyze multiplexed samples, which contain reads from several loci but are processed separately, again using the forward primers to select locus-specific reads. No additional software is required other than providing a list of samples and loci prior to analysis. CHIIMP is available at https://github.com/ShawHahnLab/chiimp and can be installed on any Windows, Mac OS, or Linux computer with a standard installation of R and RStudio in a single step. On Windows, a desktop shortcut to the analysis script is provided. Dragging a simple text file containing analysis options onto the shortcut triggers analysis with the selected options. In addition to the standalone program, all features can also be used individually from within R. ~~Examples~~ A comprehensive user guide including examples of analysis options and locus attributes is~~are~~ provided with the software.

~~Statistical, E~~error, diversity, and heterozygosity calculations

~~Statistical analyses of *c-myc* concentration and allele detection were performed using the Mann-Whitney test and Prism version 5 software (GraphPad).~~ Error rates for the MiSeq derived genotypes were calculated by determining the number of allelic mismatches for each sample to the known genotype of the corresponding chimpanzee (including allelic dropout, stutter sequences, PCR/sequencing artifacts, and locus amplification failure) and by dividing the total number of alleles by the number of erroneous alleles (Broquet and Petit, 2004). The expected heterozygosity (also termed gene diversity) for the sampled Gombe and GME chimpanzees was calculated from both capillary electrophoresis and MiSeq based microsatellite data as described in Charlesworth & Charlesworth 2010. Allelic diversity was calculated by summing the total number of unique alleles in a population.

Results

~~Development of a MiSeq-based STR genotyping approach for wild chimpanzees~~

~~To take advantage of available microsatellite and kinship data from Gombe chimpanzees, we selected fecal samples from 24 individuals, who were previously genotyped by capillary electrophoresis on multiple occasions (Keele *et al.* 2009; Rudicell *et al.* 2010; Santiago *et al.* 2003). Their consensus genotype at eight STR loci served as the benchmark to which all MiSeq derived data were compared (Table S2). STR loci were amplified as in the past using the same PCR conditions, except for primers containing MiSeq adapters rather than fluorescent labels. To avoid read alignment artifacts across the repeat motif regions, amplification products were not fragmented and sequenced~~

using only the forward reads. This allowed us to utilize STR amplicons of up to 400 bp, which included all previously characterized STR loci (Table S1).

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

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

of reads (Fig. S3). To exclude Taq polymerase and sequencing errors, the same read count requirements were also imposed for alleles of identical lengths that differed in their nucleotide sequence or were one base pair shorter or longer. Implementation of these filters removed all erroneous alleles, except for a single stutter allele that fell just above the required threshold (Fig. 1d f, Fig. S3). The only other discrepancies from the reference genotype were amplification failures of either one (allelic drop out) or both alleles (orange and grey fields in Fig. 1f, respectively). Thus, comparison to the benchmark genotypes allowed us to iteratively improve the filtering process, such that off target amplicons, PCR and sequencing errors, and almost all stutter sequences were removed.

Increasing the sensitivity of STR allele detection

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

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

~~their host DNA concentrations, with some containing no detectable chimpanzee DNA (<2.5 pg/ul) (Fig. S4a). Consistent with previous results (Morin *et al.* 2001), PCR amplification of less than 25 pg of host DNA generally failed to amplify STR loci (Fig. S4b), even when amplicons from three independent reactions were pooled. When all data from the 24 chimpanzee samples were compared, we found that not surprisingly, 67% of missing alleles were from PCR reactions that contained less than 25 pg of host DNA (samples 2, 5, 9, and 10 in Fig. 1e). After removing these samples, only 4% of loci failed to amplify and only 3% of loci exhibited allelic dropout. For all subsequent genotyping analyses, we thus pooled the products of three PCR replicates (each containing more than 25 pg of chimpanzee DNA) prior to MiSeq sequencing.~~

Direct comparison of MiSeq and capillary electrophoresis based STR genotyping

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

Using the consensus genotype of the corresponding chimpanzees for reference (Table ~~S3~~1), we found that MiSeq genotyping reduced the number of allelic dropouts by more than half (Table 2~~4~~). This was due, at least in part, to the pooling of PCR replicates, ~~which also eliminated amplification failure~~which increased the number of alleles that were detected~~, resulting in detection of at least one allele for each locus~~. However, MiSeq genotyping was also more accurate than the traditional method, which could not differentiate off-target amplifications (Tables 1 and ~~S3~~2). In addition, stutter peaks were completely eliminated by the CHIIMP analysis pipleline, which was not the case for the automated capillary electrophoresis method. Although manual peak calling also eliminated stutter peaks, this was considerably more time consuming than the MiSeq approach. For the 19 samples, conventional peak calling and allele binning took two hours, while reviewing the bioinformatics outputs took ~~approximately 5~~ minutes. Most importantly, MiSeq genotyping identified eight heterozygous loci that were scored as homozygous by capillary electrophoresis because of a failure to resolve minor sequence and length (1bp) differences (Fig. 1~~3~~). These sequence variants were readily identified in the read histograms (Fig. 1a) and their frequency identified in sequence alignments of the entire locus (Fig. 1b and c). Inspection of allele lengths across all loci revealed that 24% of all MiSeq derived alleles did not differ by multiples of four, indicating frequent nucleotide insertions and deletions in the tetranucleotide repeats (Fig. 1b and c).

MiSeq genotyping uncovers increased allelic diversity and heterozygosity

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

Subjecting one representative fecal sample to MiSeq analysis, we confirmed 51 known alleles, but also detected 31 new alleles, which had previously gone unrecognized due to ~~size homoplasy~~minor ~~size (1 nt~~bp) or nucleotide sequence differences (Tables 3~~2~~ and S2~~4~~). Such cryptic alleles were detected for all eight STR loci, increasing allelic diversity by an average of 1.6 fold per locus. Nearly half of all previously reported alleles had closely related length or sequence variants (Table 4~~3~~).

Although the great majority of the newly identified alleles were found in multiple individuals, we wanted to validate their authenticity by demonstrating inheritance. Since paternity and kinship relationships are known for most Gombe chimpanzees, we were able to trace the majority of the newly identified allelic variants from parents to their offspring. For example, Locus 3 includes four alleles that are identical in size (234 bp) but differ by up to three substitutions and two single nucleotide insertions and deletions (Fig. 2~~5~~4a). Alleles 234-a, 234-b, 234-c, and 234-d were found in 80, 25, 10 and 4 chimpanzees, respectively, including several parent-offspring triads (Fig. 2~~5~~4b). Overall, we were able to document inheritance for 25 (81%) of the 31 new alleles. For the remaining 6 existing pedigree information was insufficient, and their existence was thus confirmed by sequencing at least two independent PCR amplicons (Table 4~~3~~).

The newly identified alleles revealed that over a quarter of genotypes at loci previously assigned as homozygous (60 of a total of 228) were in fact heterozygous (Table S2~~4~~). This increased allelic diversity resolved one case of an ambiguous paternity determination. Using the standard eight STR loci, we were previously unable to identify the father of one infant (Google) because two candidate males (Faustino and Londo) had the identical genotype at all eight STR loci (Walker *et al.* 2017). Using the new genotypes, we were able to exclude Londo and confirm Faustino as a father by revealing differences at one locus (Fig. ~~5~~2~~4a and c~~). Although Faustino was identified as the correct father at the time by genotyping 10 additional loci using capillary electrophoresis (Walker *et al.*

2017), this would not have been necessary had the increased allelic diversity been known. Thus, MiSeq genotyping revealed much greater allelic and microsatellite gene diversity in Gombe than previously appreciated, thus increasing the analytical potential of the existing STR loci.

MiSeq genotyping based individual identification

Since chimpanzee communities are often studied longitudinally, we added an individual identification tool to the analysis platform. This tool compares the genotype of every new sample with all previously characterized genotypes and generates a distance score to indicate their relative similarity. For example, samples with a distance score of 0 match at all loci, while samples with a distance score of 2 differ by two alleles. We then used this approach to characterize the same 19 newly genotyped samples (Table 1~~S3~~) as well as 5 samples from infants with unknown genotypes. To account for allelic dropout, a distance score of up to 3 was allowed. The results revealed accurate individual identification for all samples from previously genotyped chimpanzees. Of the 19 samples, 8 exhibited a perfect match across all loci (Fig. 3~~5~~a), while 11 others had distance scores of 1-3, which were consistent with allelic dropout (Fig. 3~~5~~b). However, 5 samples with distance scores of 5-7 could not be assigned to known individuals (Fig. 3~~5~~c), and a review of field notes revealed that they were all collected from new infants. A heatmap allowed the quick identification of very close (4821, 4807) and very distant (4566) matches (Fig. ~~5d~~3d). Thus, the individual identification tool detected previously determined genotypes with reasonable accuracy.

STR genotyping of multiplexed samples

Chimpanzees in the Greater Mahale Ecosystem in Tanzania occupy a large home range, live at low population densities, and face extreme seasonal changes (Moore 1996; Ogawa *et al.* 1999; Schoeninger *et al.* 1999). Thus, these "savanna chimpanzees" live under ecologically more challenging conditions than their forest-dwelling counterparts, and with the exception of the Issa community, are not habituated. As a result, fecal collections, sample transport and storage are logistically more difficult, which can result in reduced amounts of collected material and/or partially degraded host DNA. To test the suitability of MiSeq genotyping for such samples, we selected 12 previously characterized chimpanzee fecal specimens from the Issa Valley (Rudicell *et al.* 2011) and re-genotyped them using both singleplex and multiplex locus amplification. Singleplex PCR was performed as in Gombe, while multiplex PCR was carried out in two steps as previously described (Arandjelovic *et al.* 2009). First, PCR primers for 4 loci were pooled and used to amplify fecal DNA in two (rather than eight) reactions (one-step multiplex product). Second, aliquots of this first round PCR were then used in a second round of PCR to amplify each of the 8 STR loci separately (two-step multiplex product). Three pooled replicates of both one-step and two-step multiplexed products were sequenced and compared to the previously determined genotypes (Table S3~~5~~). Although the overall amplification efficiency was lower than originally reported (most likely due to repeat freezing and thawing of the 7-8 year old samples), one-step multiplexing performed as well as singleplex PCR, but used only a quarter of the fecal DNA (Table 5~~4~~). Two-step multiplexing detected slightly more alleles, but not surprisingly, also resulted in an increased number of stutter sequences and other PCR artifacts. Thus, one-step multiplexing required less starting material and was also more cost efficient because the combined loci were sequenced in a single MiSeq run (and were subsequently de-multiplexed bioinformatically).

MiSeq genotyping also allowed us to compare the allelic diversity in Gombe and the GME. Fig. 4~~6~~ depicts such an analysis for locus B and D, highlighting alleles that were

only found in GME chimpanzees. Comparing all eight STR loci, we found ten alleles in only 12 GME chimpanzees that were absent from the 123 genotyped Gombe individuals, six of which represented alleles previously missed in the GME due to sequence and length differences. Although the mean expected heterozygosity value for the GME chimpanzees (0.743) was lower than that for Gombe (0.812), this is likely due to the small sample size and the fact that all 12 individuals were sampled at a single location in Issa Valley (Rudicell *et al.* 2011). Additional samples from more diverse locations in the GME are needed to compare the genetic diversity of this population to that of Gombe and other field sites.

Discussion

Over the past two decades, microsatellite analyses have been an integral part of studies of wild chimpanzees, providing insight into their evolution, population genetics, behavior, disease association and social structure (Barbian *et al.* 2018; Becquet *et al.* 2007; Keele *et al.* 2009; Langergraber *et al.* 2007; Moeller *et al.* 2016; Rudicell *et al.* 2010; Santiago *et al.* 2003; Vigilant *et al.* 2001; Walker *et al.* 2017; Wroblewski *et al.* 2015). However, ~~current~~ traditional genotyping methods are cumbersome, imprecise and investigator/platform dependent, due to the use of capillary electrophoresis to determine the length of STR loci. Here, we report a high-throughput MiSeq-based approach, which represents a marked improvement, because it is faster, more accurate and able to detect the full extent of allelic diversity in a population. Moreover, it includes a new analysis platform, CHIIMP, which not only automates the conversion of raw MiSeq data into multi-locus genotypes, but also implements a number of quality control measures that improve genotyping accuracy (Fig. ~~7~~5). Of note, CHIIMP has been designed for maximal ~~adaptability and~~ customization. While ~~our~~analysis of pedigreed ~~chimpanzee~~fecal samples

from chimpanzees ~~and genotypes~~ allowed rigorous validation, the ~~analysis~~ pipeline is not limited to a particular species or sample type.

Improved accuracy of MiSeq based genotyping

Sequence-based genotyping methods not only determine the length of STR loci, but also reveal their sequence content, and thus have the potential to detect a greater number of distinct alleles than capillary electrophoresis. Indeed, such genotyping of Atlantic cod and muskrats revealed high proportions of cryptic alleles, ranging from 32% to 44% (Darby *et al.* 2016; Vartia *et al.* 2016). In light of these data, our discovery of 38% new alleles (31 of 82) in Gombe is not surprising (Table ~~2~~3). However, this finding suggests that existing STR data vastly underestimate the diversity of microsatellite sequences in wild chimpanzees, not only in Gombe but also in other populations. New alleles were found for all loci, with some comprising twice as many variants as previously observed (Table 3~~2~~), which will undoubtedly add to the statistical power of future analyses. However, any new allele will have to be examined carefully by repeat amplification and sequencing, unless it can be validated by pedigree analysis. In our dataset, ~~8% of initially identified~~a minor fraction of "new" alleles were found to represent PCR and/or sequencing artifacts that exceeded the 33% threshold for heterozygous alleles~~, which prompted us to add a filter that specifically removed such erroneous alleles~~. ~~Since all PCR/sequencing artifacts occurred in combination with an allele of the same length, they were easily recognized~~Repeat amplification of these alleles ~~in question~~ resolved ~~all~~all ~~cases of~~ sequencing~~e~~ artifacts ~~that were called as alleles~~.

Comparison of the MiSeq data to validated reference genotypes also allowed us to assess the error rate of the new approach. After implementation of all filters, CHIIMP eliminated 98% of stutter sequences and 100% of off-target amplicons. Among the

samples tested, true alleles, allelic dropouts and false alleles were detected with a frequency of 96%, 7%, and 0%, respectively. These data are comparable to MiSeq derived genotyping results for wild-living brown bears, where true alleles, allelic dropouts and false alleles were detected with a frequency of 93%, 0.4% and 0.05% for tissues, and 80%, 14% and 1% for fecal samples, respectively (De Barba *et al.* 2017). Although our overall error rate of 3.3% is slightly higher than the 2.1% error rate reported for a MiSeq genotyping study of laboratory raised (pedigreed) fish (Zhan *et al.* 2017), this is not surprising since the latter study examined freshly extracted tissue DNA.

Since non-invasively collected samples frequently contain diluted and/or degraded host DNA, they are genotyped using multiple PCR reactions to guard against the selective loss of alleles (allelic dropout). Loci are only considered homozygous if they can be confirmed in multiple PCR reactions (Morin *et al.* 2001; Taberlet *et al.* 1996). Capillary electrophoresis requires that these replicates are run independently to distinguish true alleles from non-specific signal, often more than tripling the amount of time and effort required to genotype a single sample. In contrast, MiSeq genotyping can be performed after combining the products of multiple PCR reactions ~~(Fig. 2)~~. Although the quality of the input DNA remains the same, MiSeq genotyping of pooled PCR replicates reduces the frequency of allelic dropout and thus renders the resulting genotypes more accurate. ~~While~~However, amplicon pooling ~~replicate amplifications can maximize data output, it does~~~~can~~does result in the loss of the~~~~foregoes data from repeat analyses~~, which are used by some as a measure of DNA quality and/or data reliability (Taberlet *et al.* 1996)

Once MiSeq data files are imported into the CHIIMP platform, the program calls alleles automatically, thus saving days of hands-on work. While automated allele calling has been reported previously (De Barba *et al.* 2017; Suez *et al.* 2016; Zhan *et al.* 2017), CHIIMP includes downstream analyses, such as alignments of allele sequences or flagging loci that may contain contaminants, which provide important additional quality

control measures. In contrast to previous studies, CHIIMP also retains non-repeat regions~~,~~ (Suez *et al.* 2016), which can contribute to allelic diversity, and does not require the presence of stutter sequences for allele calling, which may not be sufficiently abundant under conditions of low coverage (De Barba *et al.* 2017~~; Suez *et al.* 2016~~). Finally, CHIIMP reports both allele length and sequence content, and is thus designed to detect minor length and sequence differences by including sequence-specific allele names and generating locus-specific sequence alignments (Figs. 4~~6~~ and S1). To guide subsequent analyses, we have also added features that flag potentially problematic alleles and standardize allele naming. CHIIMP thus represents the most comprehensive analysis platform yet to ensure the accuracy of MiSeq-based genotyping results.

Multiplexing improves MiSeq genotyping efficiency and reduces cost

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

number of pooled loci for any given sample is limited only by the desired read depth. Moreover, barcoding of individual samples allows their combination in sequencing reactions, thus further increasing sequencing efficiency and throughput (Farrell *et al.* 2016).

MiSeq genotyping is expensive, but these costs decrease with sample numbers. Capillary electrophoresis is undoubtedly cheaper when only a small number of samples has to be analyzed; however, MiSeq sequencing becomes increasingly more cost-effective with multiplexing and analyses of pooled replicates (Darby *et al.* 2016). The costs of MiSeq genotyping three replicates of 96 samples multiplexed at four loci would roughly be equivalent to analyzing the same multiplexed samples via capillary electrophoresis, because the latter cannot analyze pooled replicates. While this estimate only considers genotyping supplies, labor to manually analyze samples is not included. In addition, the improved accuracy has downstream cost advantages since fewer repeat analyses would have to be performed.

Effective sharing of MiSeq genotyping data

A direct comparison of MiSeq and capillary electrophoresis derived alleles revealed consistent length differences of one to three nucleotides, the number of which were locus specific (Tables 1 and S3~~S2 and S3~~). For Gombe samples, locus 3 alleles derived by capillary electrophoresis were always three nucleotides longer than the corresponding MiSeq alleles (Table 1~~s S2 and S3~~). However, for the GME samples, the same alleles were all one nucleotide shorter than the MiSeq alleles (Table S3~~5~~). This is as expected since the capillary electrophoresis data were generated on different platforms. However, this also means that a simple conversion of existing capillary electrophoresis to MiSeq data will generally not be possible. In contrast, MiSeq genotyping generates unambiguous

alleles that can be compared across multiple studies and field sites (Fig. ~~6~~4). In the future, it will thus be possible to compare STR genotypes across different chimpanzee populations, such as those in Gombe and the GME, since the use of different sequencing equipment will no longer confound these analyses.

Versatility of the CHIIMP analysis platform

To increase its utility, we designed the CHIIMP analysis platform to be versatile. STR locus attributes, such as the expected length range, primer sequences, and repeat motifs, as well as all thresholds for allele calling can be customized. For example, analysis of loci with dinucleotide repeats may require a lower threshold for stutter peaks, since these are more susceptible to polymerase slippage (Guichoux *et al.* 2011). Similarly, locus length ranges can be expanded or contracted, depending on the rate of off-target amplicons. The CHIIMP analysis pipeline also includes tools that facilitate iterative improvements for new applications (Fig. S1). For example, the program provides a heatmap that indicates the number of unique sequences that pass all filters. If that number is too high, thresholds can be adjusted to remove stutter peaks, off-target amplicons, and/or PCR errors. In addition, the distribution of loci is visualized, which can be used to reveal contamination in singleplexed samples or identify poorly performing primers in multiplexed samples (Fig. S1g). For potentially problematic alleles, CHIIMP generates histograms that provide information concerning their length and relative abundance. All of these tools can be used to adapt the platform to additional STR loci and/or host species.

The length of STR loci suitable for sequence-based genotyping depends on the sequencing chemistry. We used Illumina v2 technology, which has maximum read lengths of 500 nucleotides. MiSeq sequences are most often generated using paired-end reads, with a maximum read length of 250 nucleotides in each direction. For STR genotyping,

locus sequences must span the repeat motif region, since assembly of shorter reads could result in misalignments. To accommodate loci of greater than 250 bp length, we opted to only use forward reads for analysis. Although the sequencing kit could theoretically accommodate fragments of up to 500 nucleotides, we found that the quality of reads (Q scores) decreased significantly after 400 cycles. Given that the longest locus in our panel spanned 357 nucleotides, we used 375 cycles in the forward direction. Illumina v3 sequencing chemistry has a 600-cycle limit, which may accommodate loci of up to 500 bp, but this would have to be determined experimentally. The majority of microsatellite loci are shorter than this length.

As STR genotyping transitions from capillary electrophoresis to sequence based approaches, it will be necessary to standardize allele nomenclature, as has already been suggested for human forensics (Gelardi *et al.* 2014; Parson *et al.* 2016). At a minimum, allele names will have to incorporate the length and unique sequence content for each allele (Darby *et al.* 2016). In our study, we added an alphabetical identifier (-a, -b, -c, etc.) to differentiate identically sized alleles that differed in their sequence (Table S2). Since it is impossible to capture all allele attributes in a single name, it may become necessary to establish databases that link allele identifiers to their respective sequences. CHIIMP is designed to allow users to supply a spreadsheet of allele names and sequences, and thus guarantees consistent nomenclature across experiments. As MiSeq genotyping is adapted to additional projects, standardized allele designations will become necessary to ensure consistent nomenclature across studies.

Conclusions

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

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

Author contributions

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

Acknowledgements

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

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

Data accessibility

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

References

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

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

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

Balloux F, Lugon-Moulin N (2002) The estimation of population differentiation with microsatellite markers. *Molecular Ecology* **11**, 155-165.

Barbian HJ, Li Y, Ramirez M*, et al.* (2018) Destabilization of the gut microbiome marks the end-stage of simian immunodeficiency virus infection in wild chimpanzees. *American Journal of Primatology* **80**, 10.1002/ajp.22515.

Becquet C, Patterson N, Stone AC, Przeworski M, Reich D (2007) Genetic structure of chimpanzee populations. *PLOS Genetics* **3**, e66.

Bennett P (2000) Demystified …: Microsatellites. *Molecular Pathology* **53**, 177-183.

Bodenhofer U, Bonatesta E, Horejš-Kainrath C, Hochreiter S (2015) msa: an R package for multiple sequence alignment. *Bioinformatics* **31**, 3997-3999.

Broquet T, Petit E (2004) Quantifying genotyping errors in noninvasive population genetics. *Molecular Ecology* **13**, 3601-3608.

Charlesworth B, Charlesworth D (2010) *Elements of evolutionary genetics*. Roberts and Co. Publishers. 734 p.

Constable JJ, Packer C, Collins DA, Pusey AE (1995) Nuclear DNA from primate dung. *Nature* **373**, 393.

Constable JL, Ashley MV, Goodall J, Pusey AE (2001) Noninvasive paternity assignment in Gombe chimpanzees. *Molecular Ecology* **10**, 1279-1300.

Darby BJ, Erickson SF, Hervey SD, Ellis-Felege SN (2016) Digital fragment analysis of short tandem repeats by high-throughput amplicon sequencing. *Ecology and Evolution* **6**, 4502-4512.

De Barba M, Miquel C, Lobréaux S, *et al.* (2017) High-throughput microsatellite genotyping in ecology: improved accuracy, efficiency, standardization and success with low-quantity and degraded DNA. *Molecular Ecology Resources.* **17**, 492-507.

Ellegren H (2004) Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetics* **5**, 435.

Ewen KR, Bahlo M, Treloar SA, *et al.* (2000) Identification and analysis of error types in high-throughput genotyping. *American Journal of Human Genetics* **67**, 727-736.

Farrell ED, Carlsson JEL, Carlsson J (2016) Next Gen Pop Gen: implementing a high-throughput approach to population genetics in boarfish (Capros aper). *Royal Society Open Science* **3**, 160651.

Fernando P, Evans BJ, Morales JC, Melnick DJ (2001) Electrophoresis artefacts — a previously unrecognized cause of error in microsatellite analysis. *Molecular Ecology Notes* **1**, 325-328.

Fordyce SL, Ávila-Arcos MC, Rockenbauer E, *et al.* (2011) High-throughput sequencing of core STR loci for forensic genetic investigations using the Roche Genome Sequencer FLX platform. *BioTechniques* **51**, 127-133.

Gelardi C, Rockenbauer E, Dalsgaard S, Børsting C, Morling N (2014) Second generation sequencing of three STRs D3S1358, D12S391 and D21S11 in Danes and a new nomenclature for sequenced STR alleles. *Forensic Science International: Genetics* **12**, 38-41.

Guichoux E, Lagache L, Wagner S, *et al.* (2011) Current trends in microsatellite genotyping. *Molecular Ecology Resources* **11**, 591-611.

Hauge XY, Litt M (1993) A study of the origin of 'shadow bands' seen when typing dinucleotide repeat polymorphisms by the PCR. *Human Molecular Genetics.* **2**, 411-415.

Iyer SS, Bibollet-Ruche F, Sherrill-Mix S*, et al.* (2017) Resistance to type 1 interferons is a major determinant of HIV-1 transmission fitness. *Proceedings of the National Academy of Sciences* **114**: E590-E599.

Jarne P, Lagoda PJL (1996) Microsatellites, from molecules to populations and back. *Trends in Ecology & Evolution* **11**, 424-429.

Keele BF, Jones JH, Terio KA*, et al.* (2009) Increased mortality and AIDS-like immunopathology in wild chimpanzees infected with SIVcpz. *Nature* **460**, 515-519.

Kelkar YD, Strubczewski N, Hile SE*, et al.* (2010) What is a microsatellite: a computational and experimental definition based upon repeat mutational behavior at A/T and GT/AC repeats. *Genome Biology and Evolution* **2**, 620-635.

Langergraber KE, Mitani JC, Vigilant L (2007) The limited impact of kinship on cooperation in wild chimpanzees. *Proceedings of the National Academy of Sciences* **104**, 7786-7790.

Levinson G, Gutman GA (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Molecular biology and evolution*. **4**, 203-221.

Lynch HT, de la Chapelle A (2003) Hereditary Colorectal Cancer. *New England Journal of Medicine* **348**, 919-932.

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10-12.

Moeller AH, Foerster S, Wilson M*, et al.* (2016) Social behaviour promotes diversity in the chimpanzee gut microbiome. *American Association for the Advancement of Science* **2**, e1500997.

Moore J (1996) Savanna chimpanzees, referential models and the last common ancestor. *Great Ape Societies*. Cambridge University Press, 275-292.

Morin PA, Chambers KE, Boesch C, Vigilant L (2001) Quantitative polymerase chain reaction analysis of DNA from noninvasive samples for accurate microsatellite genotyping of wild chimpanzees (Pan troglodytes verus). *Molecular Ecology* **10**, 1835-1844.

Morin PA, Wallis J, Moore JJ, Chakraborty R, Woodruff DS (1993) Non-invasive sampling and DNA amplification for paternity exclusion, community structure, and phylogeography in wild chimpanzees. *Primates* **34**, 347-356.

Ogawa H, Idani Gi, Kanamori M (1999) Chimpanzee habitat in the savanna woodland, Ugalla, Tanzania. *Primate Research* **15**, 135-146.

Parson W, Ballard D, Budowle B*, et al.* (2016) Massively parallel sequencing of forensic STRs: Considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements. *Forensic Science International: Genetics* **22**, 54-63.

Pasqualotto AC, Denning DW, Anderson MJ (2007) A cautionary tale: lack of consistency in allele sizes between two laboratories for a published multilocus microsatellite typing system. *Journal of Clinical Microbiology* **45**, 522-528.

Peakall ROD, Smouse PE (2006) genalex 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes* **6**, 288-295.

Pusey AE, Pintea L, Wilson ML, Kamenya S, Goodall J (2007) The contribution of long-term research at Gombe National Park to chimpanzee conservation. *Conservation Biology* **21**, 623-634.

Queller DC, Strassmann JE, Hughes CR (1993) Microsatellites and kinship. *Trends in Ecology & Evolution* **8**, 285-288.

Rudicell RS, Holland Jones J, Wroblewski EE*, et al.* (2010) Impact of simian immunodeficiency virus infection on chimpanzee population dynamics. *PLoS Pathogens* **6**, e1001116.

Rudicell RS, Piel AK, Stewart F*, et al.* (2011) High prevalence of simian immunodeficiency virus infection in a community of savanna chimpanzees. *Journal of Virology* **85**, 9918-9928.

Santiago ML, Lukasik M, Kamenya S*, et al.* (2003) Foci of endemic simian immunodeficiency virus infection in wild-living eastern chimpanzees (*Pan troglodytes schweinfurthii*). *Journal of Virology* **77**, 7545-7562.

Schoeninger MJ, Moore J, Sept JM (1999) Subsistence strategies of two "savanna" chimpanzee populations: The stable isotope evidence. *American Journal of Primatology* **49**, 297-314.

Schuelke M (2000) An economic method for the fluorescent labeling of PCR fragments. *Nature Biotechnology* **18**, 233.

Selkoe KA, Toonen RJ (2006) Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. *Ecology Letters* **9**, 615-629.

Shinde D, Lai Y, Sun F, Arnheim N (2003) Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)(n) and (A/T)(n) microsatellites. *Nucleic Acids Research* **31**, 974-980.

Suez M, Behdenna A, Brouillet S*, et al.* (2016) MicNeSs: genotyping microsatellite loci from a collection of (NGS) reads. *Molecular Ecology Resources* **16**, 524-533.

Taberlet P, Camarra JJ, Griffin S*, et al.* (1997) Noninvasive genetic tracking of the endangered Pyrenean brown bear population. *Molecular Ecology* **6**, 869-876.

Taberlet P, Griffin S, Goossens B*, et al.* (1996) Reliable genotyping of samples with very low DNA quantities using PCR. *Nucleic Acids Research* **24**, 3189-3194.

van Lawick-Goodall J (1968) The behaviour of free-living chimpanzees in the Gombe Stream Reserve. *Animal Behaviour Monographs* **1**, 161-311.

Van Neste C, Van Nieuwerburgh F, Van Hoofstat D, Deforce D (2012) Forensic STR analysis using massive parallel sequencing. *Forensic Science International.* **6**, 810-818.

Vartia S, Villanueva-Cañas JL, Finarelli J*, et al.* (2016) A novel method of microsatellite genotyping-by-sequencing using individual combinatorial barcoding. *Royal Society Open Science* **3**, 150565.

Vigilant L, Hofreiter M, Siedel H, Boesch C (2001) Paternity and relatedness in wild chimpanzee communities. *Proceedings of the National Academy of Sciences* **98**, 12890-12895.

Walker KK, Rudicell RS, Li Y*, et al.* (2017) Chimpanzees breed with genetically dissimilar mates. *Royal Society Open Science* **4**, 160422.

Weeks DE, Conley YP, Ferrell RE, Mah TS, Gorin MB (2002) A tale of two genotypes: consistency between two high-throughput genotyping centers. *Genome Res* **12**, 430-435.

Wroblewski EE, Norman PJ, Guethlein LA*, et al.* (2015) Signature patterns of MHC diversity in three gombe communities of wild chimpanzees reflect fitness in reproduction and immune defense against SIVcpz. *PLoS Biology* **13**, e1002144.

Zhan L, Paterson IG, Fraser BA*, et al.* (2017) megasat: automated inference of microsatellite genotypes from sequence data. *Molecular Ecology Resources* **17**, 247-256.

Figure legends

Fig 13 MiSeq genotyping uncovers cryptic alleles. Eight polymorphic STR loci were amplified from the fecal DNA of 19 previously genotyped chimpanzees. (a) Histogram depicting the length (x-axis) and read count (y-axis) of unique sequences for one representative heterozygous locus that was previously determined to be homozygous by multiple capillary electrophoresis analyses (sample 4861, locus C, Table 1). The grey box highlights the expected locus size range. The horizontal line indicates the cutoff of 500 reads. Colored peaks indicate reads that passed the locus-specific filters (note that peaks can be comprised of identically sized reads that differ in their sequence content). Black reads were eliminated. Pink reads appear to be locus-specific, but did not pass the PCR artifact filters. Red reads represent the true allele sequences (180 and 181 bp in lengths, respectively). (ba, cb) Alignment images of locus-specific allele sequences are shown for locus 1 (ab) and locus C (bc), respectively (the complete data set is shown in Table 1S3). Allele sequences are ordered by length (indicated in bp on the right), with the frequency with which they were found in different chimpanzees indicated on the left (the x-axis indicates the position within the alignment). Nucleotides are colored as shown, with gaps in the alignment shown in grey. The insets highlight alleles that differ in their sequence content and/or length. Nucleotide substitutions are colored; dashes indicate gaps that were introduced to optimize the alignment.

Fig. 24 MiSeq genotyping uncovers increased allelic diversity and heterozygosity. (a) Alignment of four locus 3 alleles that are of identical length (234 bp), but differ in sequence content. Nucleotide substitutions are colored; dashes indicate single nucleotide insertions and deletions (b) Mendelian inheritance of allele 234 for a group of related chimpanzees. Fathers and mothers are shown as squares and circles, respectively, with offspring connected by vertical lines. Both alleles are shown for each animal, with the four allelic variants highlighted in different colors. Individuals of unknown identity or genotype

are left blank. (c) Increased allelic diversity resolves a previously ambiguous paternity determination. Two potential fathers with identical allele lengths (238 bp) can now be distinguished based on differences in allele sequence content (238-a and 238-b). Since the offspring is homozygous for allele 238-a, the male with allele 238-b can be excluded as a father.

Fig. 3~~5~~ Individual identification based on MiSeq genotyping. (a-c) Genotypes of newly collected samples (top) are compared to the genotypes of known community members, with the closest match listed below (based on descending distance scores). Genotypes that differ by fewer than four alleles are indicated in bold because they represent likely matches. Differences are highlighted in yellow. (d) Heatmap showing the relative similarity of sample genotypes (rows) with genotypes of known individuals (columns) based on distance scores. Dark red cells indicate likely matches.

Fig 4~~6~~ Comparison of MiSeq genotypes across chimpanzee communities. Alignment images of locus-specific allele sequences are shown for chimpanzees from the GME and Gombe. Two representative loci (locus B on the left; locus D on the right) are shown for (a) 12 chimpanzees from the GME (Table S3~~5~~), (b) 123 chimpanzees from Gombe (Table S2~~4~~), and (c) a combination of both. Allele sequences are ordered by length (indicated in base pairs on the right), with the frequency with which they were found in different chimpanzees indicated on the left (the x-axis indicates the position within the alignment). Nucleotides are colored as indicated, with alignment gaps shown in grey. Arrows indicate alleles that are unique to the GME samples.

Fig. 5~~7~~ MiSeq-based STR genotyping of wild chimpanzees. (a) Schematic representation of singleplex STR amplification and MiSeq sequencing of chimpanzee fecal DNA. (b)

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

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 electrophores," 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.

Reviewer: 2

In this ms, Barbian et al. describe a HTS approach and bioinformatics pipeline for microsatellite genotyping and demonstrate its application to genotype scat samples from wild chimpanzees. HTS microsatellite genotyping has recently come into the scene and holds great potential for wildlife studies. I believe continued development and testing is important at these early stages to favor wide applicability of the method, both in terms of laboratory and sequencing protocols and of bioinformatics analysis that need to be implemented to allow wider routine utilization.
In this perspective, the main contributions of this study are:
- the development of a pipeline that is flexible and that presents useful features for evaluating and reporting the data, such as tools to visualize alleles and non-alleles sequences, genotype similarity, match to known genotypes for individual ID, flag alleles and loci to be re-checked and possible contamination, allele alignment, etc. This is all very important and need to be incorporated into pipelines for routine applications of HTS STR genotyping methods also by non-bioinformatichians. This is the main contribution of this work in my view.
- testing for pooling replicates and the potential benefits of doing that (although authors failed to fully justify its application, see later comments)
- revealing the amount and impact of hidden allelic diversity on gene diversity estimation
- implementing standardized allele codes
However, the study is poorly presented, is too long, and is very hard to follow and track things done to their respective results. In particular:
- much of what is written is already known or has already been pointed out and addressed in other recent studies. Should focus more on the novelty and additional contribution of their approach, this will also shorten the paper
- the paper is not well structured. Much of what is currently in the Results should be in the Methods. They have done several things performing different sets of experiment with different sample sets, but it is very difficult to evaluate what was actually done because it's all lost between methods and results.
Therefore, in my opinion the ms can be considered for publication only after major restructuring and re-focusing.

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

Specific comments:

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

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

L294: statistical analysis and tests for what? need to be explicit. Maybe I missed it, but I did not see reference to statistical tests in the results

The supplementary figure that required statistical analysis has been removed.

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

This has now been explained and cited.

L300: but you also report allelic diversity in this study

Allelic diversity calculations are now defined in the methods.

L306: this section should be combined or followed directly the "CHIMP analysis pipeline" section as the pipeline was implemented within the platform

This has been done.

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

This has been done.

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

This is now explained in the methods.

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

This has been rephrased.

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

This has been rephrased to "increased of the number of alleles that were detected" instead of "eliminated amplification failure".

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

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

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

L576-581: Also other pipelines retain non-repeat regions, and report allele length and sequence and assign allele names…

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

Minor comments:
L148: use only the abbreviation since you have written it out already

This has been changed.

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

This has been added.

L196: do you mean equimolar or equivolume proportions?

This has been clarified.

L224-226: move to results

This has been done.

L285: no need for a separate section for this

This has been incorporated into the previous section.

L304: add "in" before Charlesworth and Charlesworth 2010

This has been corrected.

L483: clarify a bit "ecologically challenging conditions"

This has been done.

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

This has been done.

L602: add De Barba et al. 2017

Samples were not barcoded in De Barba et al. 2017 so the citation is not appropriate.

L662: this was also pointed out already in previous wildlife HTS STR studies

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

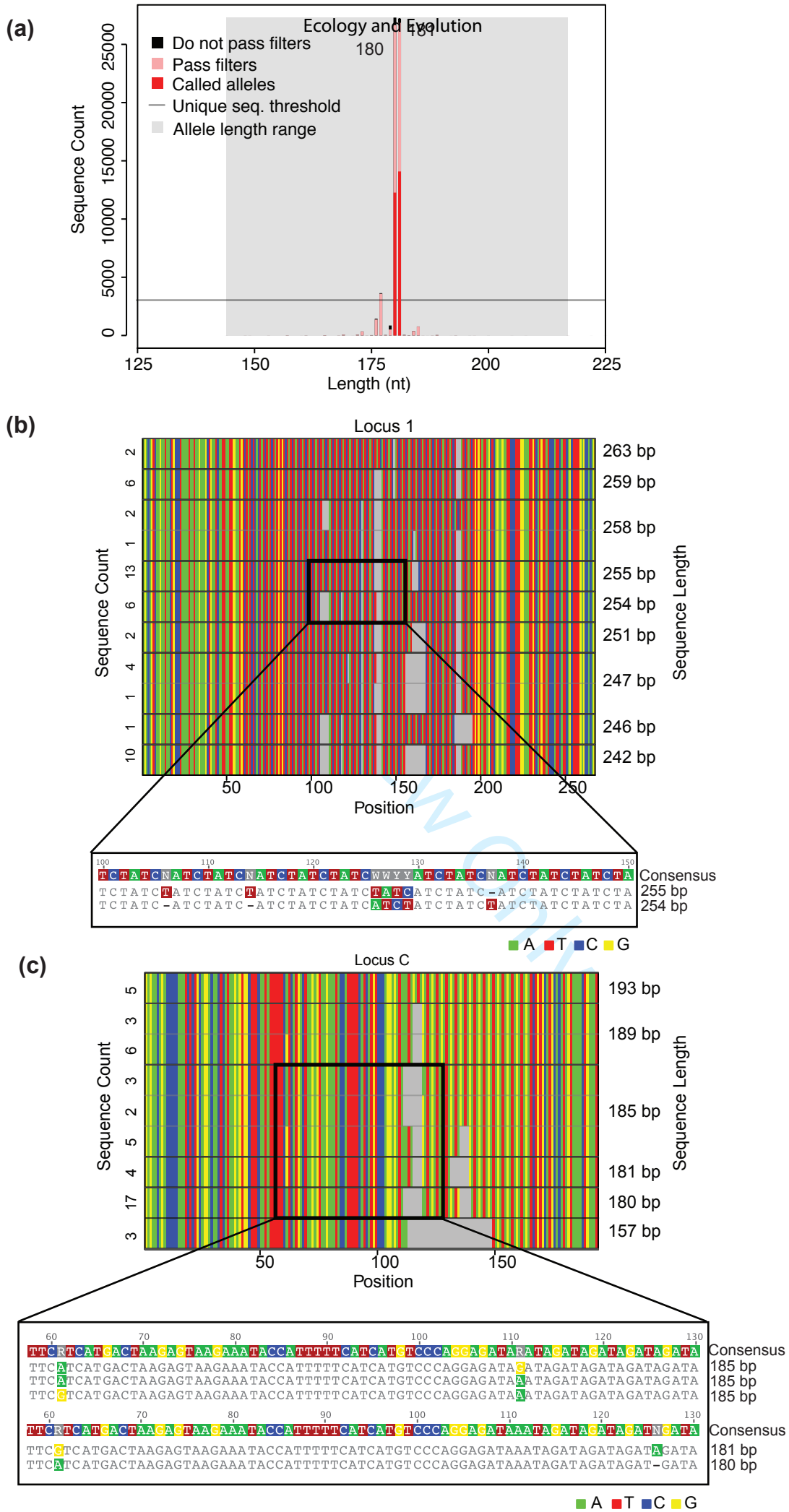


Fig. 1

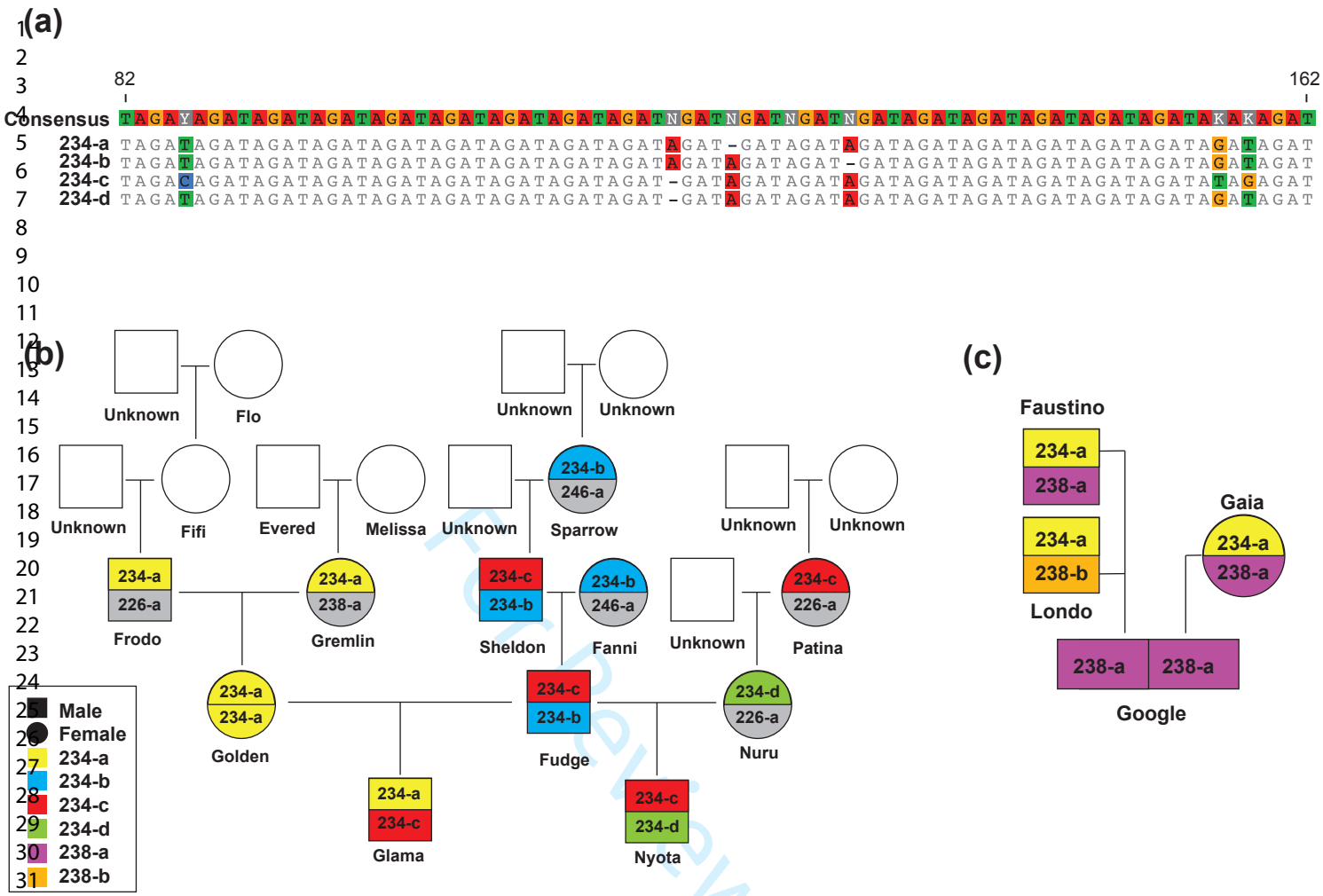


Fig. 2

(a)

Sample 4821															Distance Score	Name
157-a	173-a	231-a	235-a	185-b	189-a	285-a	300-a	255-a	255-a	302-a	322-a	234-b	234-d	294-a	295-a		Sample 4821
157-a	**173-a**	**231-a**	**235-a**	**185-b**	**189-a**	**285-a**	**300-a**	**255-a**	**255-a**	**302-a**	**322-a**	**234-b**	**234-d**	**294-a**	**295-a**	**0**	**Darbee**

(b)

Sample 4807															Distance Score	Name
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		
141-a	161-a	196-a	231-a	181-a	185-b	305-a	305-a	246-a	247-a	322-a	322-a	234-a	234-a	295-a	295-a		Sample 4807
141-a	**161-a**	**196-a**	**231-a**	**181-a**	**185-b**	**301-a**	**305-a**	**246-a**	**247-a**	**322-a**	**322-a**	**234-a**	**234-a**	**295-a**	**295-a**	**1**	**Losa**

(c)

Sample 4566															Distance Score	Name
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		
141-a	153-a	203-a	235-a	180-a	185-a	297-c	301-a	247-b	259-a	322-a	322-a	222-a	234-a	278-a	295-a		Sample 4566
141-a	153-a	203-a	203-a	185-a	185-a	285-a	297-c	247-b	259-a	322-a	322-a	222-a	234-b	286-a	294-a	6	Bahati
141-a	153-a	200-a	203-a	185-a	185-a	285-a	297-c	247-b	259-a	310-a	322-a	234-b	234-a	286-a	295-a	6	Bima
141-a	153-a	203-a	235-a	180-a	189-a	285-a	297-a	247-b	254-a	302-a	322-a	230-a	234-a	286-a	295-a	6	Bibi
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	7	Ferdinand
153-a	173-a	196-a	235-a	180-a	189-b	297-b	301-a	242-a	259-a	318-a	322-a	222-a	234-a	278-a	282-a	7	Ipo
141-a	157-a	203-a	203-a	180-a	185-a	297-c	297-a	242-a	259-a	310-a	322-a	234-a	242-a	278-a	294-a	7	Vanilla

(d)

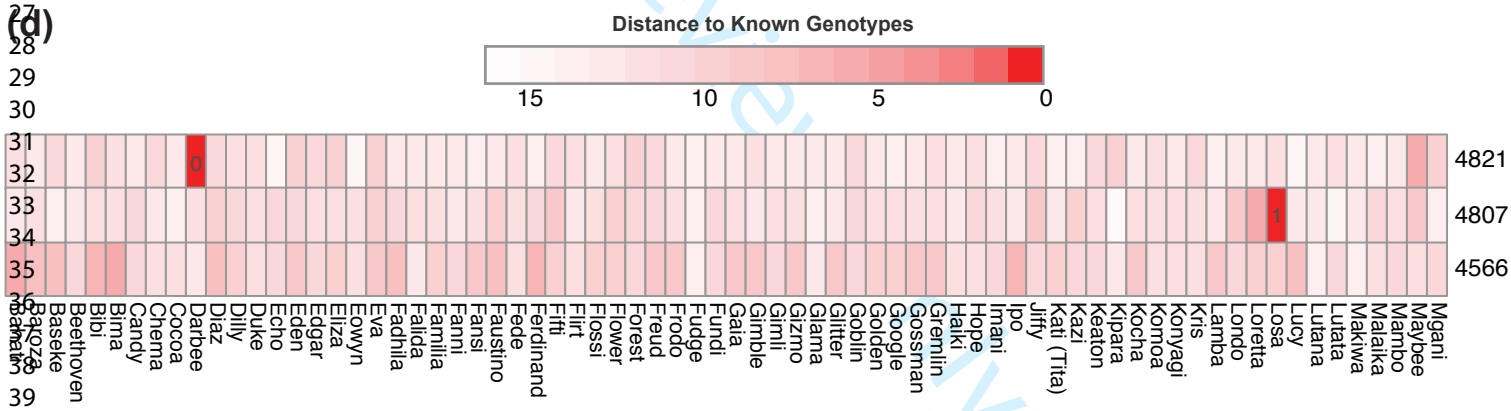


Fig. 3

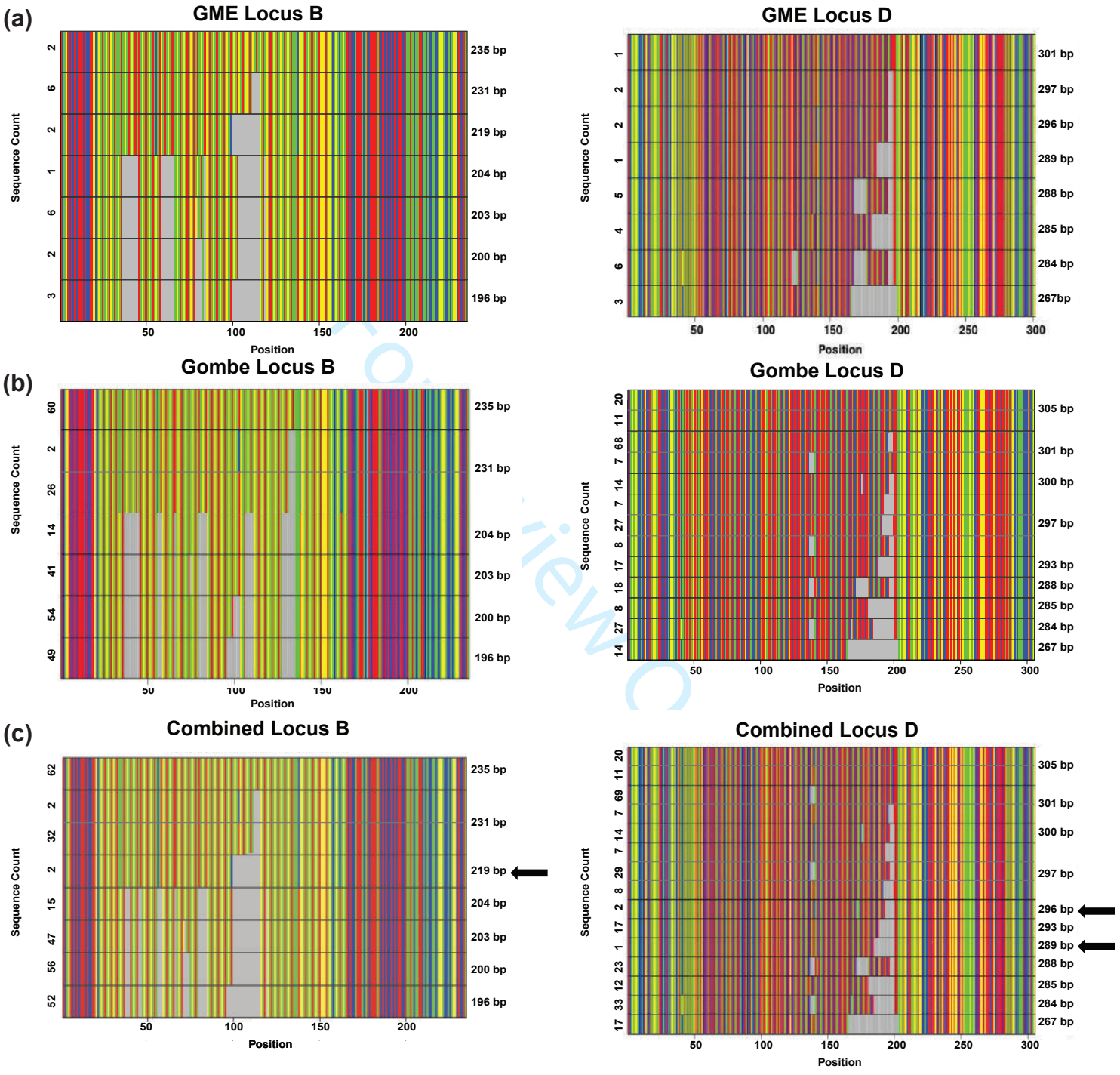
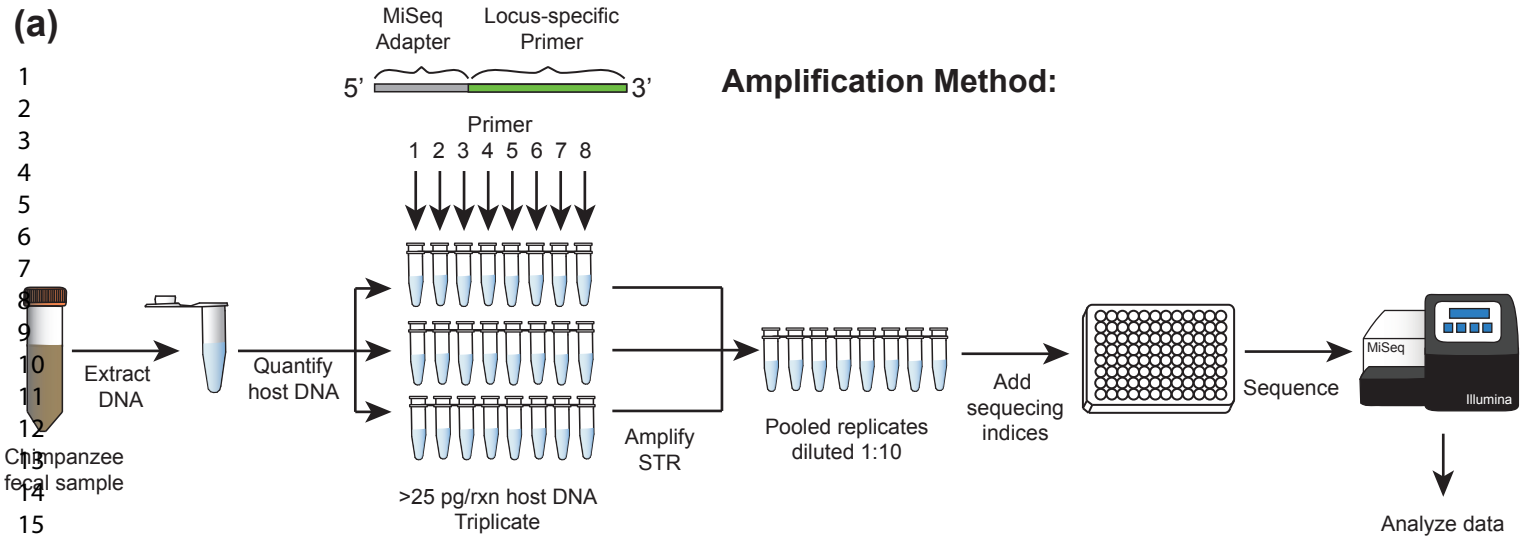


Fig. 4



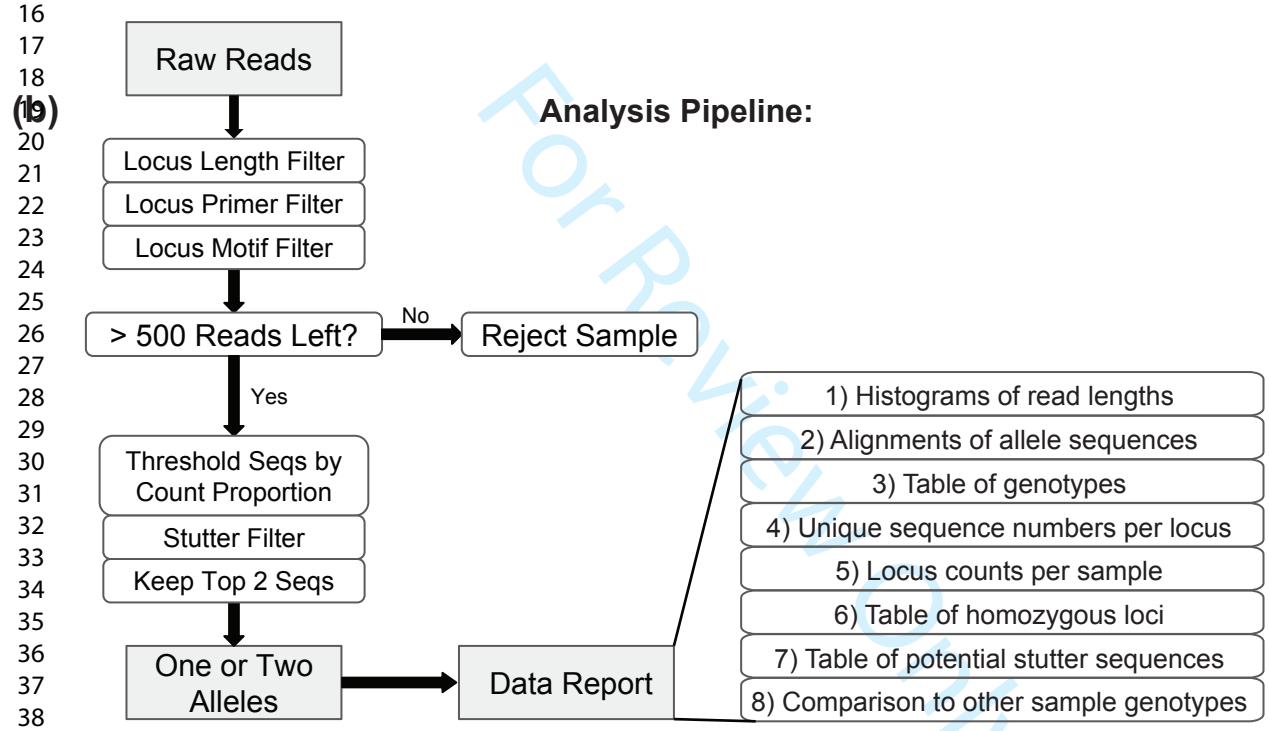


Fig. 5

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[†]	141	161	204	235	182	190	286	306	244	256	302	318	237	241	279	295
	MiSeq[‡]	141	161	203	235	180	189	284	305	242	255	302	318	234	238	278	295
	CE-manual[§]	161	161	204	235	182	190	306	306	256	256	318	318	237	237	279	279
	CE-auto[¶]	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

4844	CE-consensus	153	161	200	204	182	190	268	286	260	264	310	322	241	249	287	295
	MiSeq	153	161	200	203	180	180	267	284	259	263	310	322	238	246	286	295
	CE-manual	153	161	200	204	182	190	268	286	260	264	310	322	241	249	287	295
	CE-auto	153	161	200	204	182	190	268	286	260	264	310	322	241	249	287	287
4845	CE-consensus	141	153	196	200	182	190	268	294	244	264	318	322	237	241	287	295
	MiSeq	141	153	196	200	181	189	267	293	242	263	318	322	234	238	286	295
	CE-manual	141	153	196	200	182	190	268	294	244	264	318	322	237	241	287	287
	CE-auto	141	153	200	200	182	190	268	294	244	264	318	322	237	241	287	287
4850	CE-consensus	141	161	204	204	182	186	286	294	244	256	302	318	237	249	279	287
	MiSeq	141	161	203	203	180	185	284	293	242	255	302	318	234	246	278	286
	CE-manual	141	161	204	204	182	186	286	294	244	256	302	318	237	249	279	287
	CE-auto	141	161	204	204	182	186	286	286	244	256	302	318	237	249	279	287
4859	CE-consensus	141	173	200	204	186	194	298	302	248	248	302	310	237	241	279	295
	MiSeq	141	173	200	204	185	193	301	301	247	247	302	310	234	238	278	294
	CE-manual	141	173	200	204	186	194	302	302	248	248	302	310	237	241	279	295
	CE-auto	141	173	200	204	186	194	298	302	244	248	302	310	237	241	279	279
4861	CE-consensus	157	161	196	196	182	182	286	294	244	260	318	326	229	241	287	295
	MiSeq	161	161	196	196	180	181	284	293	242	259	318	326	226	238	286	294
	CE-manual	161	161	196	196	182	182	294	294					229	241	287	295
	CE-auto	161	198	196	196	162	182	294	294					229	241	287	295

[†]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.

[‡]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.

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

[¶]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.

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

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

[†]Peaks were called automatically using software.
[‡]Peaks were called manually.
[§]The percentage of erroneous alleles was calculated for 152 loci by comparing the newly derived results to the reference genotypes (Table 1).
[¶]Locus alleles do not match the locus primer and/or motif sequence.
[Ł]Hands-on analysis time included allele length calling, binning and individual identification.

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

Locus	Number of alleles[†]			Gene diversity[‡]	
	CE[§]	MiSeq	Cryptic[¶]	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

[†]Number of alleles at eight STR loci determined for 123 Gombe chimpanzees (Table S2).
[‡]Nine individuals were excluded from heterozygosity calculations because they had incomplete CE genotypes.
[§]CE, capillary electrophoresis
[¶]Alleles newly discovered by MiSeq genotyping.

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

Locus	Cryptic allele[†]	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[‡]	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[‡]	1			NA
2	326-b	6	1			NA
3	226-b	1[‡]		2		NA
3	230-b	1[‡]	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

[†]Cryptic alleles were compared to the most abundant allele of the same or similar length.
[‡]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[†]	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	

[†]Of a total of 192 alleles analyzed for 12 GME chimpanzees.

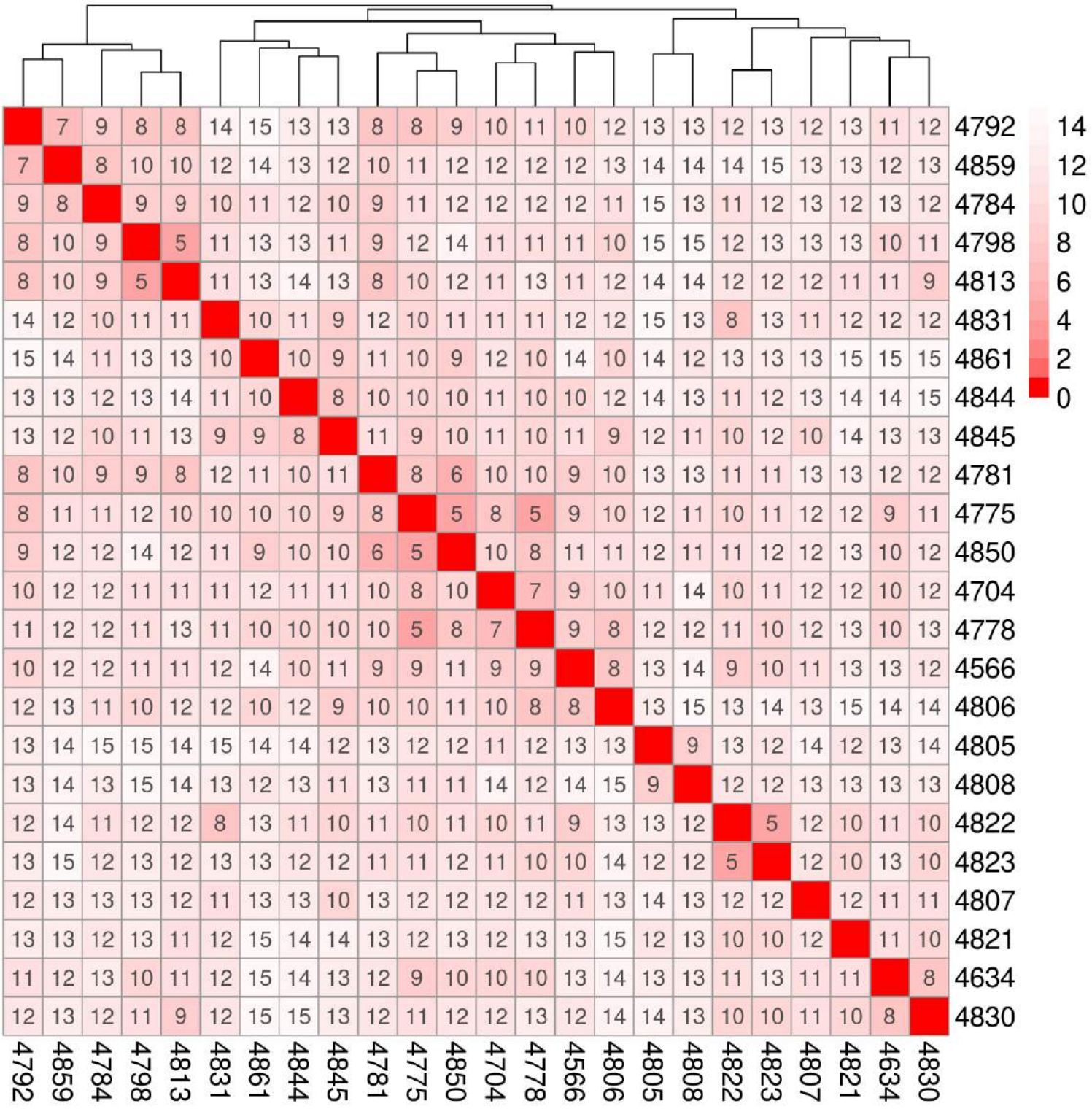
(a) Genotype Summary

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

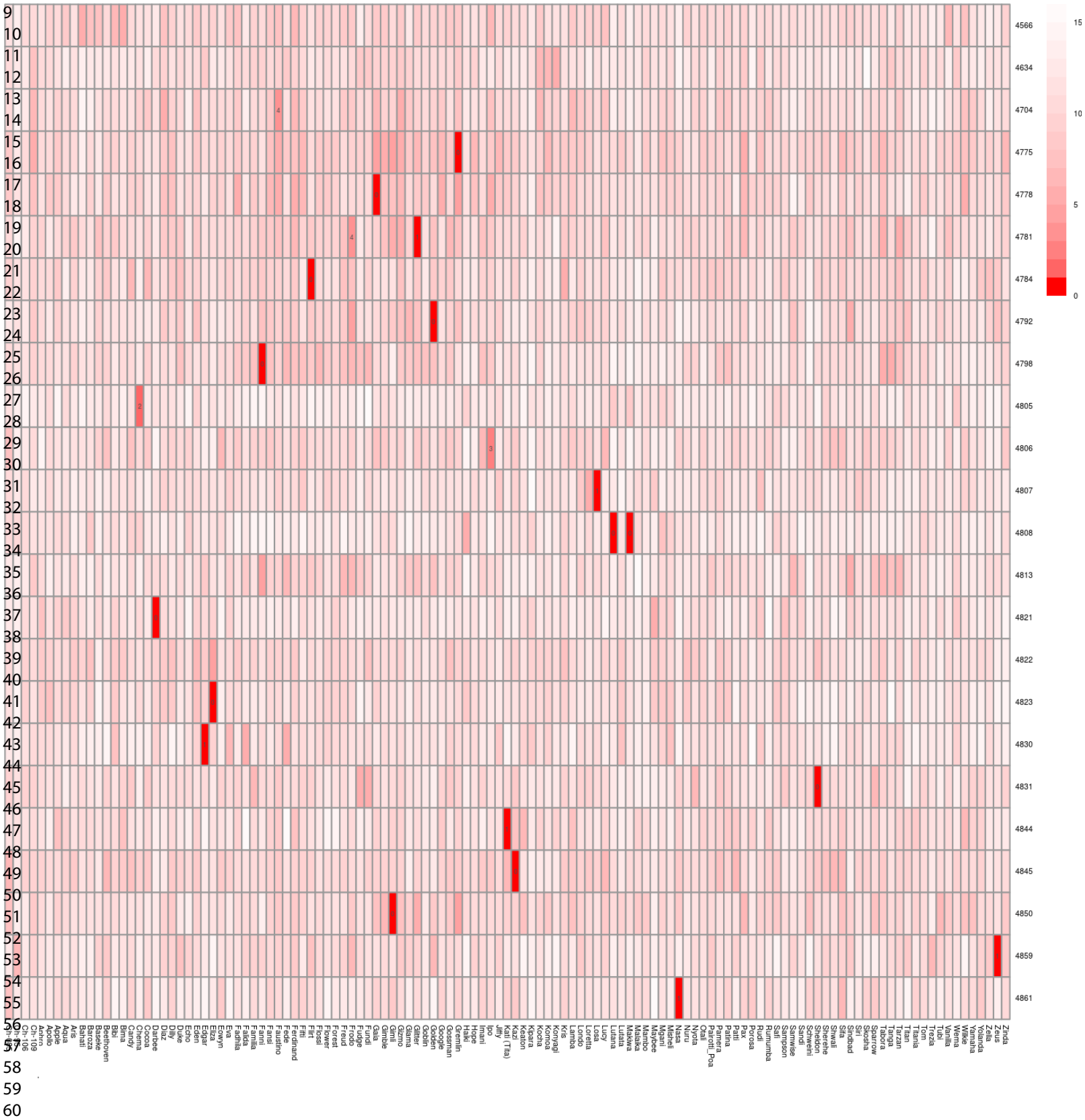


(c) Identification with Known Genotypes

Loci: A-D; Samples 4566-4781

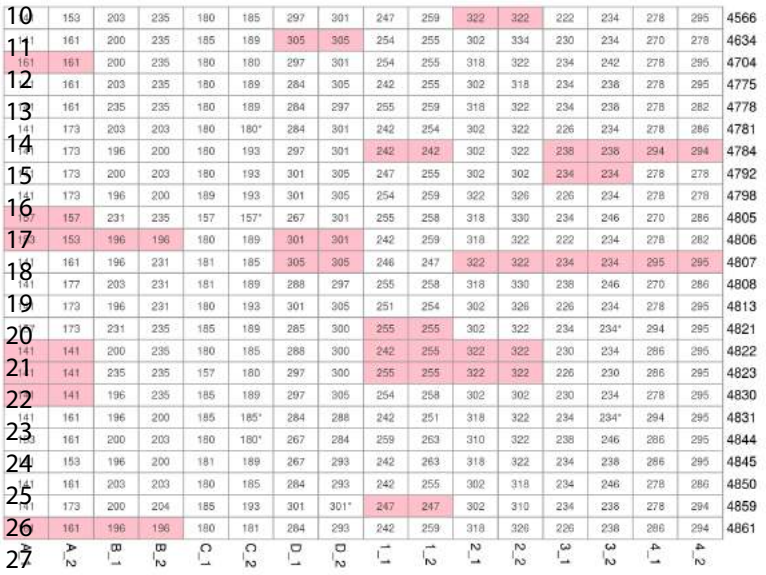
A_1	A_2	B_1	B_2	C_1	C_2	D_1	D_2	Distance	Name
Sample 4566									
141-a	153-a	203-a	235-a	180-a	185-a	297-c	301-a		
141-a	153-a	203-a	203-a	185-a	185-a	285-a	297-c	6	Bahati
141-a	153-a	200-a	203-a	185-a	185-a	285-a	297-c	6	Bima
141-a	153-a	203-a	235-a	185-a	189-a	285-a	297-a	7	Bibi
141-a	161-a	200-a	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
Sample 4634									
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
Sample 4704									
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
Sample 4775									
141-a	161-a	203-a	235-a	180-a	189-b	284-a	305-b		
141-a	**161-a**	**203-a**	**235-a**	**180-a**	**189-b**	**284-a**	**305-b**	**0**	**Gremlin**
Sample 4778									
141-a	161-a	235-a	235-a	180-a	189-b	284-a	297-b		
141-a	**161-a**	**235-a**	**235-a**	**180-a**	**189-b**	**284-a**	**297-b**	**0**	**Gaia**
Sample 4781									
141-a	173-a	203-a	203-a	180-a	180-fdd1c6	284-a	301-a		
141-a	**173-a**	**203-a**	**203-a**	**180-a**	**189-b**	**284-a**	**301-a**	**1**	**Glitter**

(d) Distance Matrix for Known Genotypes

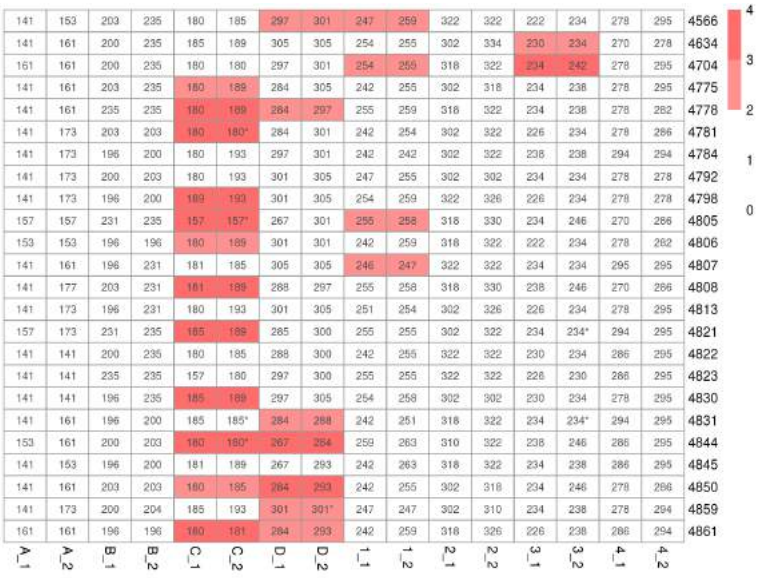


(e) Flagged Values

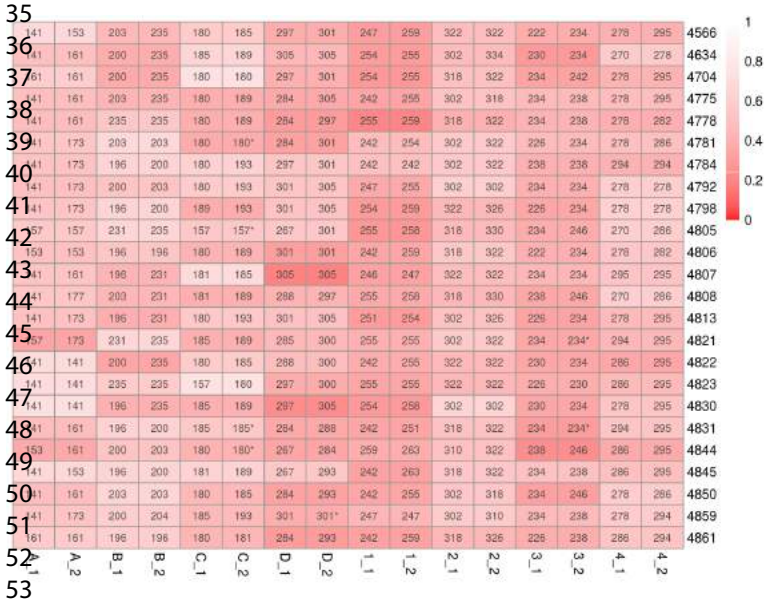
Loci subjected to stutter filter



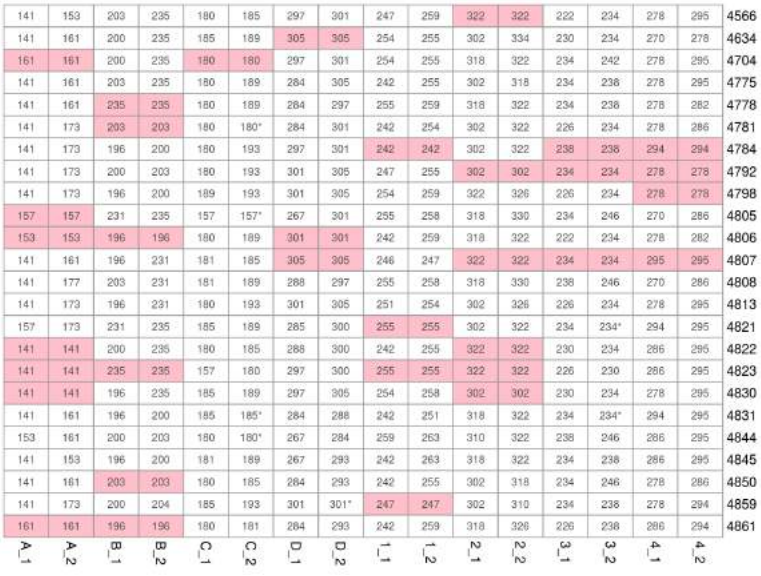
Loci with more than two prominent sequences



Proportion of allele-matching reads

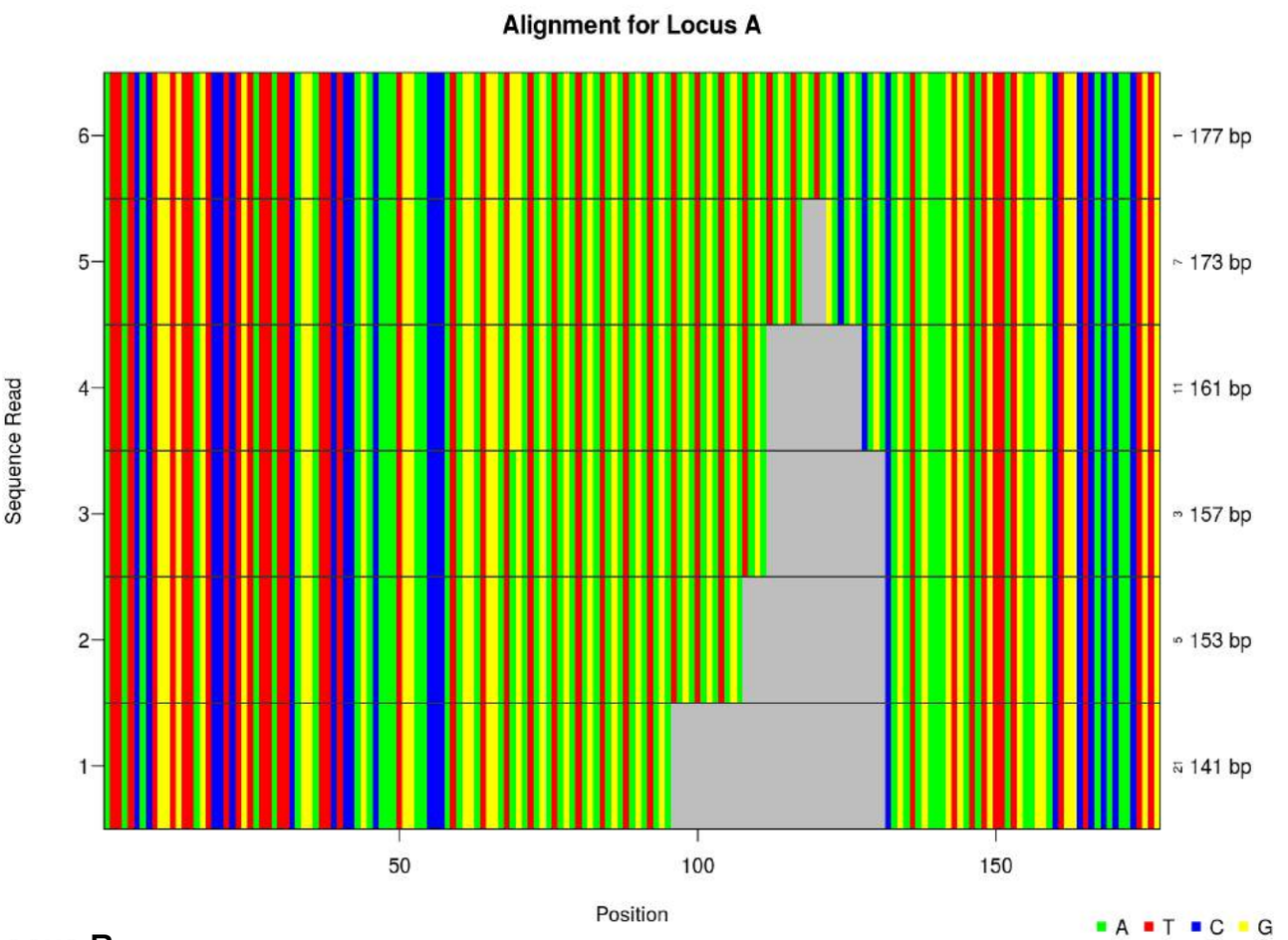


Loci with possible allelic dropout

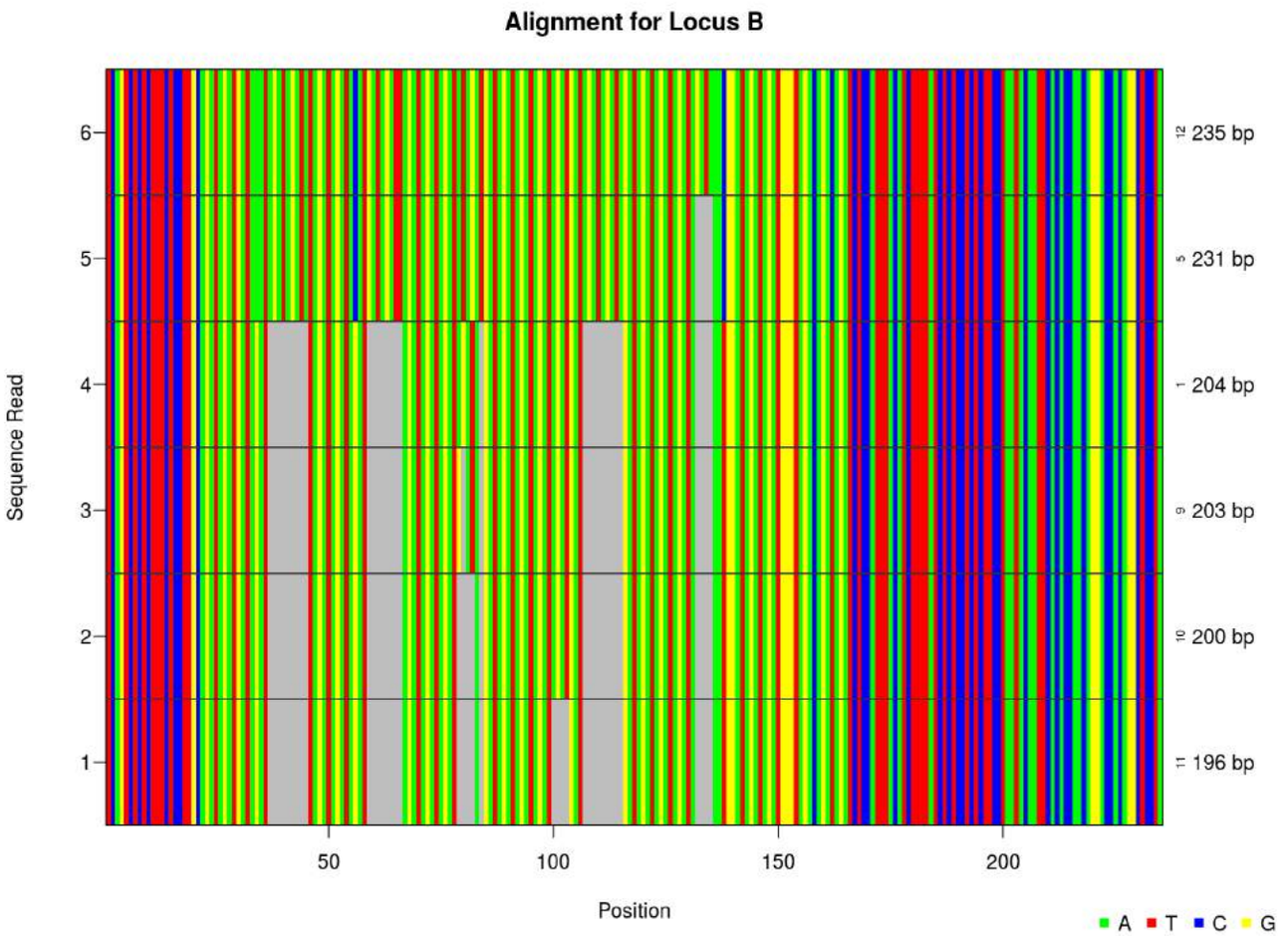


(f) Allele Alignments per Locus

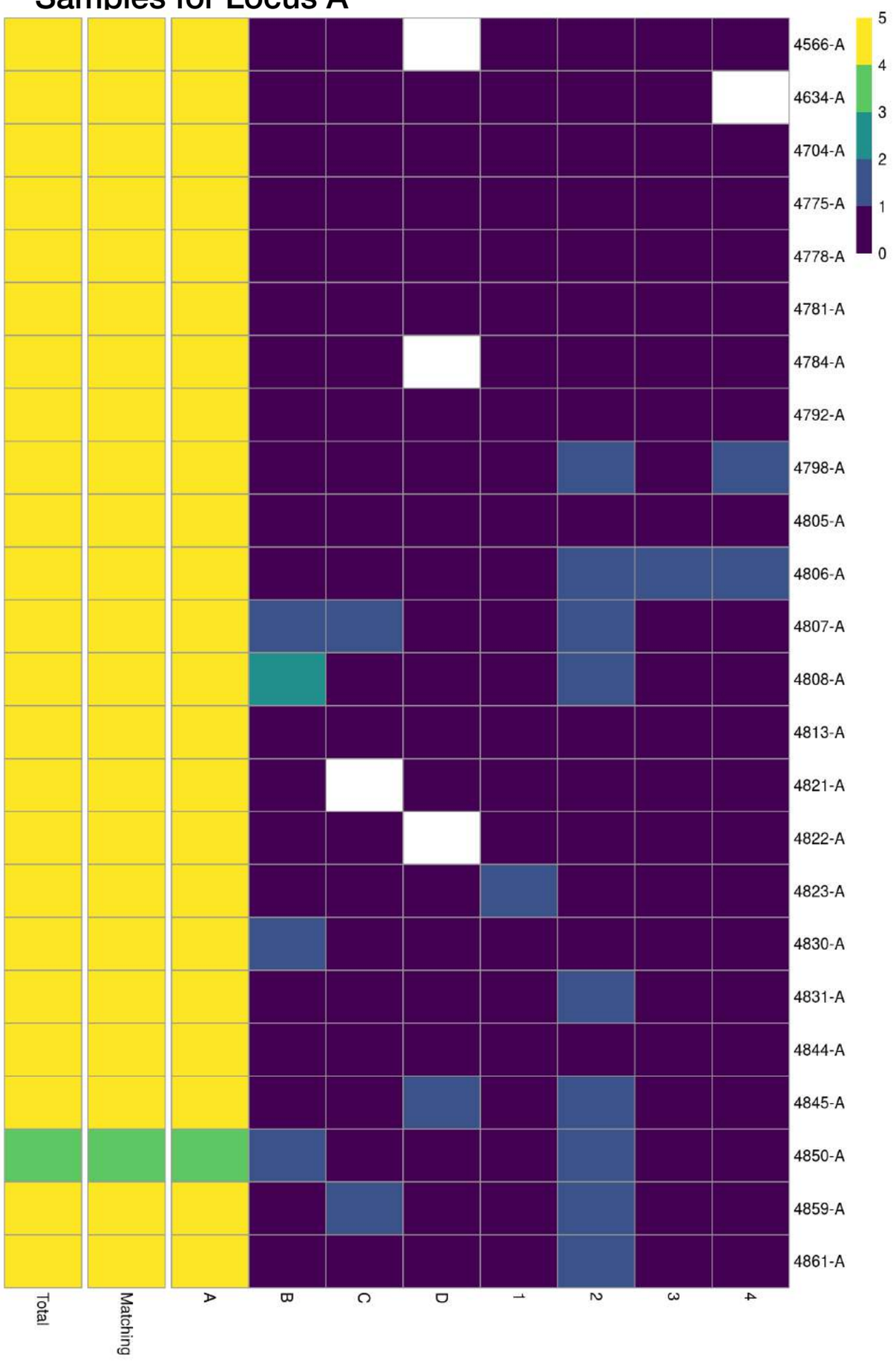
Locus A



Locus B

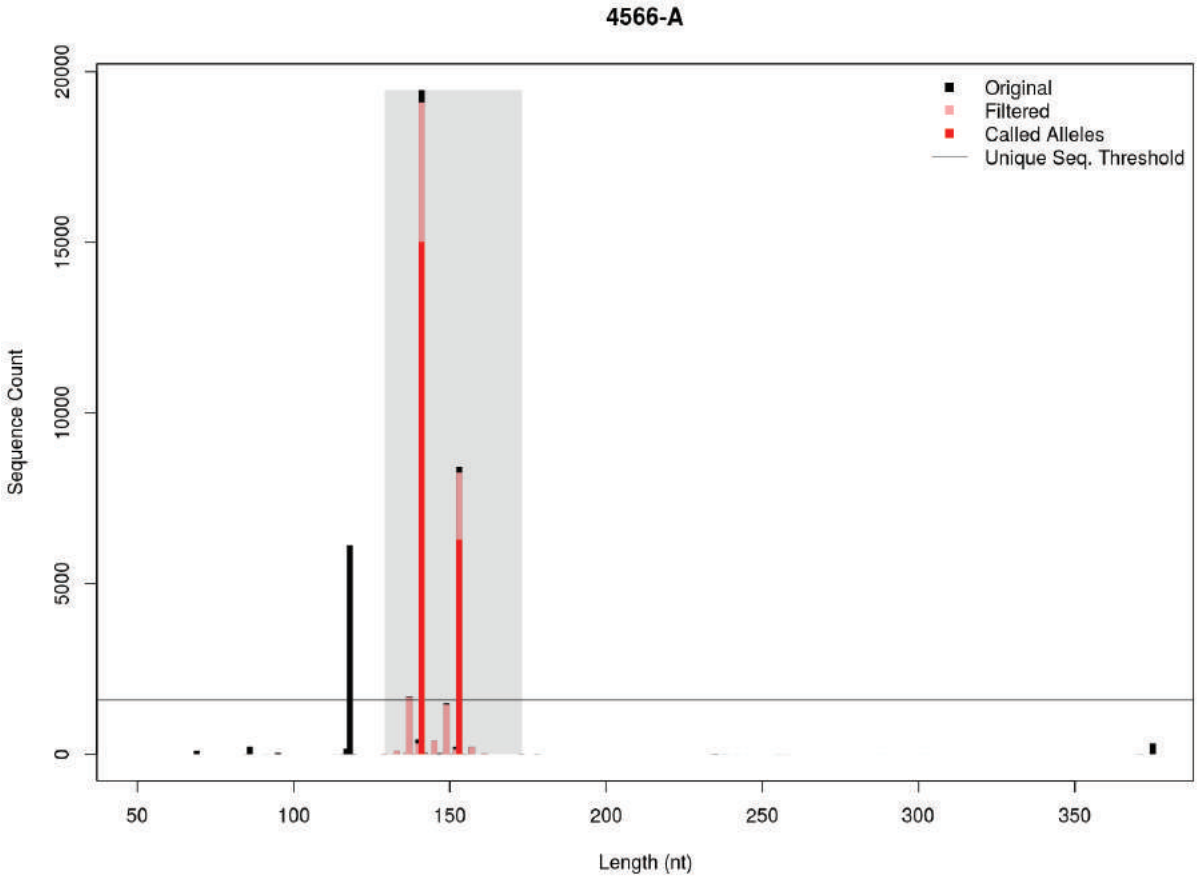


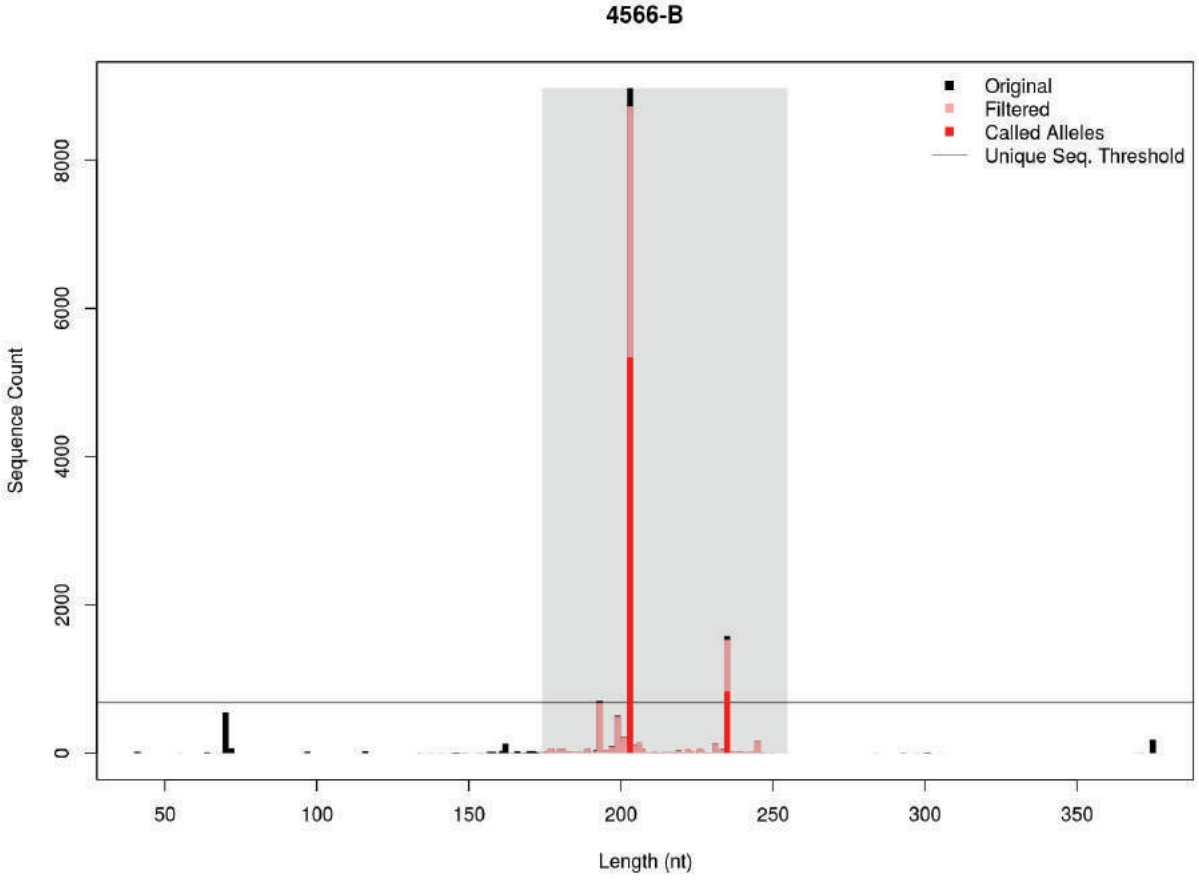
(g) Counts per Locus



(h) Histograms

Sample 4566; Locus A, 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

genotypes of known community members, with the closest match listed below (ordered by descending distance scores). Genotypes that differ by fewer than four alleles are indicated in bold because they represent potential matches. (d) Heatmap showing the relative similarity of sample genotypes (rows) with genotypes of known individuals (columns) based on distance scores. Dark red cells indicate likely matches (Lutana and Makiwa represent the same individual). (e) Quality control tables highlighting loci where stutter sequences have been filtered, where more than two sequences pass the filter (with darker cells indicating more sequences), where a large proportion of reads is not contained in the identified alleles (light red indicates very low level of non-locus reads, indicating absence of contamination; dark red requires further scrutiny), and where homozygosity may reflect allelic dropout. (f) Alignments of allele sequences. Two representative images for locus A and B are shown. Allele sequences are ordered by length (indicated in base pairs on the right), with the frequency with which they were found in different chimpanzees indicated on the left (the x-axis indicates the position within the alignment). Nucleotides are colored as indicated, with gaps in the alignment shown in grey. (g) Heatmap of sequence counts that match the locus-specific forward primer. A representative analysis is shown for 24 chimpanzee samples amplified at locus A. The first column shows the total number of reads. The second column shows the matching reads. The remaining columns show the reads matching each locus (the scale bar indicates log increases; white cells indicate no reads). For singleplex samples, this identifies sequences that match other loci and thus highlights potential cross-locus contamination. For multiplexed samples, this shows the read distribution across loci. (h) Histograms depicting sequence length-frequency distributions saved as image files. Representative histograms are shown for locus A and B of one sample. Note that peaks can be comprised of identically sized reads that differ in their sequence content and can thus contain different colors. Black highlights reads that did not match the locus

length or repeat motif filter. Pink highlights reads that appear to be locus-specific, but did not pass the PCR artifact filters (these are useful for identifying stutter sequences). Only red reads represent true allele sequences. The horizontal line indicates the minimum read cutoff for unique sequences. Histograms from each sample and locus are saved as separate image files.

Table S1. STR loci used for MiSeq genotyping

Locus	Code	Forward primer	Forward primer sequence†	Reverse primer	Reverse primer sequence†	Size range (bp)‡
D18S536	A	HUM05262	5'-ATTATCACTGGTGTTAGTCCTCTG-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'-ATTTCAGTTTTACTATGTGCATGC-3'	154-210
D9S922	D	HUM09025	5'-TCAGAGGACCACTGCCTAAG-3'	HUM09026	5'-CTGATGGGATTTGTGCCTAT-3'	260-308
D2S1326	1	HUM09373	5'-AGACAGTCAAGAATAACTGCCC-3'	HUM09374	5'-CTGTGGCTCAAAAGCTGAAT-3'	166-234
D2S1333	2	HUM12880	5'-CTTTGTCTCCCCAGTTGCTA-3'	HUM12881	5'-TCTGTCATAAACCGTCTGCA-3'	269-357
D4S1627	3	HUM05068	5'-AGCATTAGCATTTGTCCTGG-3'	HUM05069	5'-GACTAACCTGACTCCCCCTC-3'	202-254
D9S905	4	HUM07339	5'-GTGGGAAAATTGGCCTAAGT-3'	HUM07340	5'-CTTCTGAGCCTCACACCTGT-3'	257-298

†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.

‡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.[†]	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

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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
4000	12/27/14	Flirt	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
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

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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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

[†]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	CE - historical[†]	139	159	230	234	175	183	286	298	231	259	321	321	221	221	281	293
(Ch-44)	Singleplex[‡]	141	161					284	284	231	259			222	222	282	295
	One-step multiplex[§]	141	161			177	185	284	296	231	259			222	222		
	Two-step multiplex[¶]	141	141	181	181	177	185	284	284	231	231	322	322	222	222	282	294
TZ060	CE - historical	139	159	202	202	175	183	286	290	231	263	337	337	221	221	269	293
(Ch-58)	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	CE - historical	139	175	202	230	183	183	286	290	251	255	317	321	221	233	281	293
(Ch-19)	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	CE - historical	171	175	194	202	175	175	286	290	231	259	317	317	221	221	293	293
(Ch-51)	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	CE - historical	159	159	202	234	183	191	298	302	251	255	329	337	221	221	293	293
(Ch-26)	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	CE - historical	151	159	198	202	175	183	270	286	231	255	321	325	221	221	293	293
(Ch-54)	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	CE - historical	159	175	218	230	175	183	286	298	251	262	317	325	221	233	285	293
(Ch-4)	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	CE - historical	159	171	194	202	175	179	270	286	251	263	325	325	221	221	281	285
(Ch-48)	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	CE - historical	159	171	230	230	183	183	290	298	251	255	309	337	221	221	269	269
(Ch-53)	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	CE - historical	139	171	230	234	175	179	286	290	263	263	321	325	221	225	285	293
(Ch-18)	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

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

[†]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.
[‡]Singleplex: genotype of the same samples generated by amplifying and MiSeq sequencing each locus individually.
[§]One-step multiplex: genotype of the same sample generated by amplifying and MiSeq sequencing two pools of four loci (one-step PCR).
[¶]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.