

OPEN

# The draft genome of mandrill (*Mandrillus sphinx*): An Old World monkey

Ye Yin<sup>1,2,6\*</sup>, Ting Yang<sup>2,4,6</sup>, Huan Liu<sup>1,2,5,6</sup>, Ziheng Huang<sup>2,3</sup>, Yaolei Zhang<sup>2,3</sup>, Yue Song<sup>2,3</sup>, Wenliang Wang<sup>2,3</sup>, Xuanmin Guang<sup>2</sup>, Sunil Kumar Sahu<sup>2,3,5</sup> & Karsten Kristiansen<sup>1,2\*</sup>

Mandrill (*Mandrillus sphinx*) is a primate species, which belongs to the Old World monkey (*Cercopithecidae*) family. It is closely related to human, serving as a model for human health related research. However, the genetic studies on and genomic resources of mandrill are limited, especially in comparison to other primate species. Here we produced 284 Gb data, providing 96-fold coverage (considering the estimated genome size of 2.9 Gb), to construct a reference genome for the mandrill. The assembled draft genome was 2.79 Gb with contig N50 of 20.48 Kb and scaffold N50 of 3.56 Mb. We annotated the mandrill genome to find 43.83% repeat elements, as well as 21,906 protein-coding genes. The draft genome was of good quality with 98% gene annotation coverage by Benchmarking Universal Single-Copy Orthologs (BUSCO). Based on comparative genomic analyses of the Major Histocompatibility Complex (MHC) of the immune system in mandrill and human, we found that 17 genes in the mandrill that have been associated with disease phenotypes in human such as Lung cancer, cranial volume and asthma, barboled amino acids changing mutations. Gene family analyses revealed expansion of several genes, and several genes associated with stress environmental adaptation and innate immunity responses exhibited signatures of positive selection. In summary, we established the first draft genome of the mandrill of value for studies on evolution and human health.

Since the successful accomplishment of human genome project<sup>1</sup>, followed by continuous reduction in the cost of genome sequencing and dramatic increase in throughput of new sequencers, magnanimous primates genomic data are becoming available for both Old World monkeys, such as chimpanzee<sup>2</sup>, and New World monkeys, such as marmoset (*Callithrix jacchus*) (Table 1). Generally, the species being selected for genome sequencing must meet certain criteria including: (1) important evolutionary position within the phylogeny (for instance, chimpanzee, gibbon and orangutan); (2) biomedical relevance to human. For example, macaque and baboon were selected as they are often used to explore the genetic basis of human diseases<sup>3</sup>, and squirrel monkey has been a model for studies on neurobiology and infectious diseases. These genomic data resources provided deeper understanding on genome content, evolution as well as diversity. It enabled comparative analyses of human and other primates, and primates and other mammals.

*Mandrillus sphinx* (henceforth referred to as Mandrill) is a primate species living in Africa. It is a relatively ancient monkey species belonging to the Papionini tribe. Mandrills are mostly terrestrial but they are more arboreal than baboons<sup>4</sup>. They live in large, stable groups with the size as big as hundreds of individuals. The largest horde which has been verifiably observed contained more than 1,300 mandrills, which is the largest non-human aggregation ever documented. Mandrills have been used as experimental models for studies of human diseases, although chimpanzee and gorilla are more closely related to human<sup>5</sup>. Notably, mandrills are used in immune related research such as in studies of bacterial infections<sup>6,7</sup>, parasite infections<sup>8,9</sup> and viral diseases particularly Simian immunodeficiency viral (SIVs) infection<sup>10,11</sup>. But until now no genomic data on mandrill has been released impeding further scientific research.

In this study, we present a high-quality draft genome sequence of the mandrill using high throughput sequencing, making it a useful resource for future comparative genomic studies and studies related to human health.

<sup>1</sup>Laboratory of Genomics and Molecular Biomedicine, Department of Biology, University of Copenhagen, Copenhagen, 2100, Denmark. <sup>2</sup>BGI-Shenzhen, Shenzhen, 518083, China. <sup>3</sup>China National GeneBank, BGI-Shenzhen, Shenzhen, 518120, China. <sup>4</sup>Department of Biotechnology and Biomedicine, Technical University of Denmark, 2800 Kgs., Lyngby, Denmark. <sup>5</sup>State Key Laboratory of Agricultural Genomics, BGI-Shenzhen, Shenzhen, 518083, China. <sup>6</sup>These authors contributed equally: Ye Yin, Ting Yang and Huan Liu. \*email: [yinye@bgi.com](mailto:yinye@bgi.com); [kk@bio.ku.dk](mailto:kk@bio.ku.dk)

Common name	Species name	Bases in contigs	Contig N50	Scaffold N50	Reference
Chimpanzee	<i>Pan troglodytes</i>	2.7 Gb	15.7 kb	8.6 Mb	<sup>38</sup>
Chimpanzee (updated)	<i>P. troglodytes</i>	2.9 Gb	50.7 kb	8.9 Mb	<sup>39</sup>
Bonobo	<i>Pan paniscus</i>	2.7 Gb	67 kb	9.6 Mb	<sup>40</sup>
Gorilla	<i>Gorilla gorilla</i>	2.7 Gb	11.8 kb	914 kb	<sup>41</sup>
Gorilla (updated)	<i>Gorilla gorilla</i>	2.8 Gb	9.6 Mb	23.1 Mb	<sup>42</sup>
Orang-utan	<i>Pongo abelii</i>	3.1 Gb	15.5 kb	739 kb	<sup>43</sup>
Indian rhesus macaque	<i>Macaca mulatta</i>	2.9 Gb	25.7 kb	24.3 Mb	<sup>44</sup>
Indian rhesus macaque (updated)	<i>M. mulatta</i>	3.1 Gb	107.2 kb	4.2 Mb	<a href="https://www.ncbi.nlm.nih.gov/assembly/GCA_000772875.3">https://www.ncbi.nlm.nih.gov/assembly/GCA_000772875.3</a> <sup>45</sup>
Chinese rhesus macaque	<i>M. mulatta</i>	2.8 Gb	11.9 kb	891 kb	<sup>46</sup>
Vietnamese cynomolgus macaque	<i>M. fascicularis</i>	2.9 Gb	12.5 kb	652 kb	<sup>46</sup>
baboons	<i>Papio baboons</i>	2.9 Gb	149.87 kb	140.35 Mb	<sup>47</sup>
Aye-aye	<i>D. madagascarensis</i>	3.0 Gb	NA	13.6 kb	<sup>48</sup>
Vervet	<i>C. aethiops</i>	2.8 Gb	90.4 kb	81.8 Mb	<sup>49</sup>
Gibbon	<i>Nomascus leucogenys</i>	2.8 Gb	35.1 kb	22.7 Mb	<sup>50</sup>
Marmoset	<i>Callithrix jacchus</i>	2.3 Gb	29 kb	6.7 Mb	<sup>51</sup>
Mouse lemur	<i>Microcebus murinus</i>	2.4 Gb	210.7 kb	108.2 Mb	<a href="https://www.ncbi.nlm.nih.gov/assembly/GCF_000165445.2">https://www.ncbi.nlm.nih.gov/assembly/GCF_000165445.2</a>
Pig-tailed macaque	<i>Macaca nemestrina</i>	2.8 Gb	106.9 kb	15.2 Mb	<a href="https://www.ncbi.nlm.nih.gov/assembly/GCF_000956065.1/#/st">https://www.ncbi.nlm.nih.gov/assembly/GCF_000956065.1/#/st</a>
Sifaka	<i>Propithecus coquereli</i>	2.1 Gb	28.1 kb	5.6 Mb	<a href="https://www.ncbi.nlm.nih.gov/assembly/GCF_000956105.1/#/st">https://www.ncbi.nlm.nih.gov/assembly/GCF_000956105.1/#/st</a>
Sooty mangabey	<i>Cercocebus atys</i>	2.8 Gb	112.9 kb	12.8 Mb	<a href="https://www.ncbi.nlm.nih.gov/assembly/GCF_000955945.1/">https://www.ncbi.nlm.nih.gov/assembly/GCF_000955945.1/</a>
Squirrel monkey	<i>Saimiri boliviensis</i>	2.5 Gb	38.8 kb	18.7 Mb	<a href="https://www.ncbi.nlm.nih.gov/assembly/GCF_000235385.1/#/def">https://www.ncbi.nlm.nih.gov/assembly/GCF_000235385.1/#/def</a>
Bushbaby	<i>Otolemur garnettii</i>	2.4 Gb	27.1 kb	13.9 Mb	<a href="https://www.ncbi.nlm.nih.gov/assembly/GCF_000181295.1/">https://www.ncbi.nlm.nih.gov/assembly/GCF_000181295.1/</a>
Mouse lemur	<i>Microcebus murinus</i>	2.4 Gb	182.9 kb	3.7 Mb	<a href="https://www.ncbi.nlm.nih.gov/assembly/GCF_000165445.1/">https://www.ncbi.nlm.nih.gov/assembly/GCF_000165445.1/</a>
Tarsier	<i>Tarsius syrichta</i>	3.4 Gb	38.2 kb	401 Mb	<sup>52</sup>

**Table 1.** Published primate genome sequences. (modified based on<sup>37</sup>).

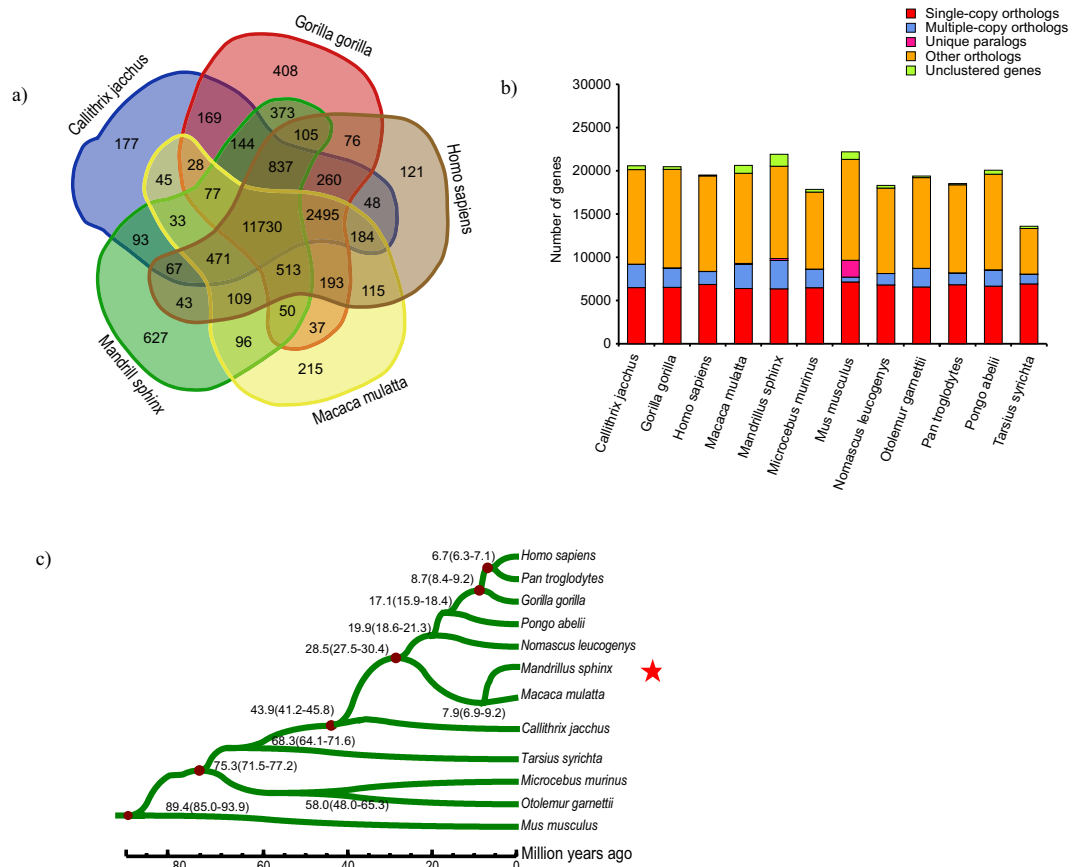
	Contig		Scaffold	
	Size (bp)	Number	Size (bp)	Number
N90	5,266	141,475	638,217	936
N80	9,025	101,618	1,303,160	634
N70	12,638	75,505	1,962,294	457
N60	16,336	56,061	2,730,696	332
N50	20,483	40,751	3,564,730	241
Longest	211,017	–	19,105,867	–
Total size	2,798,997,503	–	2,882,689,325	–
Total number (> = 100 bp)	–	455,069	–	215,140
Total number (> = 2 kb)	–	194,923	–	4,742

**Table 2.** Summary of the mandrill genome assembly.

## Results

**Genome sequencing and assembly.** Whole genome sequencing of mandrill yielded 426.72 Gb of raw sequence data (142× considering the estimated genome size of 2.90 Gb). After filtering, clean reads amounting to 289.55 Gb were obtained for genome assembly, with ~73× from paired-end libraries and ~26× from mate-pair libraries (Table S1). 212.84 Gb data were used for k-mer analysis, which resulted in the distribution of depth-frequency (Fig. S1), with a secondary peak at half of the major peak coverage of ~31×. The genome size of mandrill was estimated to be 2.90 Gb with notable heterozygosity. All of the clean data were used to generate the draft genome assembly, followed by gap filling. The size of the assembled genome was 2.88 Gb (covering 99.31% of the estimated genome size). The contig N50 was 20.48 kb with the longest contig being 211.02 kb, and the scaffold N50 was 3.56 Mb with the longest scaffold being 19.10 Mb. In total, 634 of the longest scaffolds made up more than 80% of the whole genome (Table 2).

**Genome annotation.** By a combination of *de novo* and homology-based methods, about 42.22% of the assembled mandrill genome were identified as transposable elements (TEs) (Table S2). The Long Interspersed Nuclear Elements (LINEs) made up less of the mandrill genome (~17%) compared to the human genome (~21%), while the Short Interspersed Nuclear Elements (SINEs) representation was similar (~12%). Notably the *Alu* elements made up similar proportion (10%~11%) (Table S3), reflecting that *Alu* elements are conserved within the primate genomes as previously described<sup>12</sup>.



**Figure 1.** Phylogeny and evolution of the mandrill's gene families. **(a)** The Venn diagram of the gene families of human (*Homo sapiens*), macaque (*Macaca mulatta*), gorilla (*Gorilla gorilla*), marmoset (*Callithrix jacchus*) and mandrill. **(b)** Comparison of orthologous genes among 11 primates and mouse. **(c)** The maximum-likelihood phylogenetic tree based on the 4-fold degenerate sites of 5,133 single-copy gene families in the 12 species.

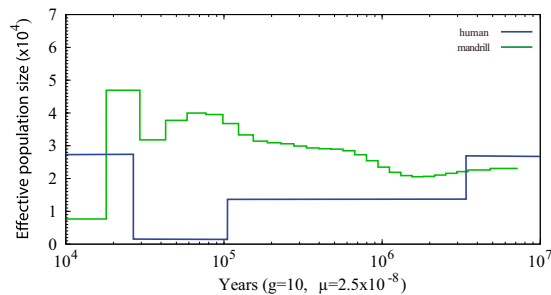
The gene set of the mandrill genome contains 21,906 protein-coding genes (Table S4). The average gene length was 39,087 base pair (bp) with the average intron length of 5,785 bp. A total of 21,622 (98.70%) of the predicted genes were functionally annotated (Table S5). Three types of ncRNAs were annotated in the mandrill genome, including tRNAs, rRNAs, and snRNAs. In total, 4,278 short noncoding RNA sequences were identified in the mandrill genome (Table S6).

The quality of mandrill genome and gene completeness were assessed by conducting the Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis<sup>13</sup>. 98% of BUSCOs were completely detected in the assembled genome (2981: complete and single-copy; 170: complete and duplicated) among 3,023 tested BUSCOs. The numbers of fragmented and missing BUSCOs were 28 and 14, respectively (Table S7).

**Gene family and phylogenetic analysis.** For mandrill, we identified 15,368 gene families with 1,387 genes that could not be clustered among 12 species, and 87 families were found to be unique (Table S8). These unique gene families were significantly enriched in the functional annotation with GO:0006412 of translation (GO level: BP,  $P = 6.29e-33$ ), and GO:0003735 of a structural constituent of ribosome (GO level, BP,  $P = 6.29e-33$ ) (Table S9). Compared to human (*Homo sapiens*), macaque (*Macaca mulatta*), gorilla (*Gorilla gorilla*) and marmoset (*Callithrix jacchus*), 627 gene families, with 1,293 genes, were found to be unique in the mandrill and 5,133 single-copy orthologous genes were found to be shared among all the 12 species (Fig. 1a).

The maximum-likelihood phylogenetic tree (Fig. 1b) of the 12 species using 5,133 single-copy genes indicated that mandrill is located in the same clade as macaque and this clade diverged from the human clade about 28.5 (27.5–30.4) million years ago (MYA) while the divergence time between *Cercopithecoidea* and *Hominoidea* was estimated to be 26.66 (24.29–28.95) MYA using mitochondrial genome sequences<sup>14</sup>. Mandrill was estimated to split from macaque about 7.9 (6.9–9.2) MYA, which is different from the previous estimate of 6.6 (6.0–8.0) MYA<sup>15</sup>.

In the mandrill lineage, there were 797 expanded and 3,982 contracted gene families (Fig. S2). Expanded gene families were found to be significantly enriched in functions related to biosynthetic processes, structural constituents of ribosomes, nucleosomal DNA binding, G-protein coupled receptor activity, olfactory receptor activity, glucose catabolic process, peptidyl-prolyl isomerization, and electron transport chain pathway (Table S10). In mandrill, peptidylprolyl isomerase A (PPIA) was significantly expanded (GO:0003755,  $P = 3.60E-89$ ). The PPIA belongs to the peptidyl-prolyl cis-trans isomerase (PPIase) family which catalyzes the cis-trans isomerization, folding of the newly synthesized protein, and regulates many biological processes including inflammation and



**Figure 2.** The demographic changes of mandrill compared to human. The x-axis indicates the time (from left to right indicates recent to ancient), and the y-axis indicates the estimated effective population size.

apoptosis, and has even been reported to play a role in cerebral hypoxia-ischemia. In a stressed environment with the presence of reactive oxygen species, cells secrete PPIA to induce an inflammatory response and mitigate tissue injury. The peroxiredoxin-6 (PRDX6) family, which can reduce peroxides and protect against oxidative injury in relation to metabolism, was also significantly expanded (GO:0051920,  $P = 0.000641$ ).

In total, 657 Positively Selected Genes (PSGs) were identified with significant enrichment in molecular functions related to kinase activity, transferase activity, and phosphotransferase activity (Table S11). The functions of these PSGs were further investigated, with 34 genes being found to be innate immunity response-related genes based on an InnateDB search. Interactions between these genes were predicted by STRING: functional protein association networks (<http://string-db.org/cgi>). As shown in Fig. S3, the genes *STAT1*, *IL5*, *IL1R1*, *ATG5*, *CREB1*, *DICER1*, *PIK3R1* which play important roles in the immune system, are strongly associated with stress resistance and wound healing. Besides that, PSGs related to innate immunity responses were found to be enriched in terms of GO:0080134: regulation of response to stress (GO level: BP,  $P = 1.59e-05$ ), GO:0006955: immune response (GO level: BP,  $P = 4.11e-05$ ) and KEGG:4640: hematopoietic cell lineage ( $P = 6.83e-05$ ) (Table S12).

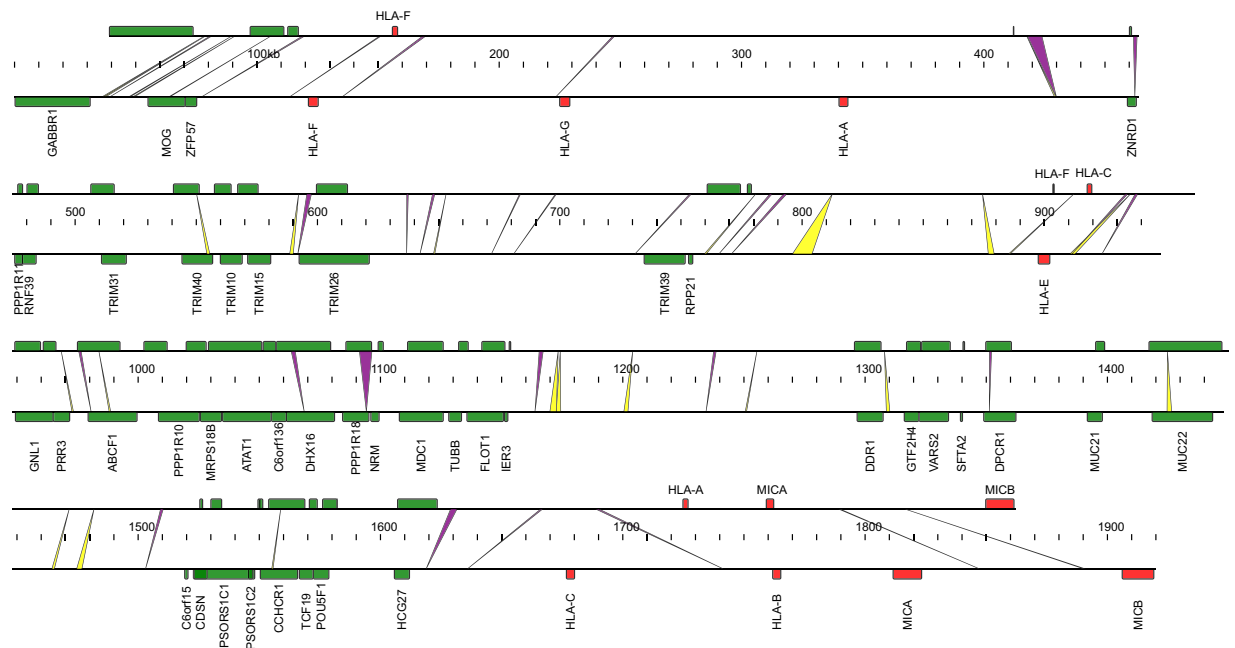
The demographic history of a species reflects historical population variation, and therefore information on the genome would be important. We inferred a noticeable population bottleneck in the demographic history of the mandrill (Fig. 2). Around 28 thousand years (kys) ago, the mandrill population went through a sharp expansion, followed by a noticeable bottleneck from a peak of 61,000 and 47,000 to ~6,500 around 17 kys ago. The expansion of the population size coincided with the increase of the human population. This might indicate that the climate change was suitable for mammal population expansion, while the recent bottleneck of mandrill populations is different from the recent increase of the human population.

**Comparative analysis of the major histocompatibility complex (MHC) region.** The MHC region harbors a series of genes to assist cells in recognizing foreign substances. Based on the assembled mandrill genomes, a relatively intact MHC region was found on chromosome 4. In order to check the assembly quality, reads that mapped back to the MHC class I region showed good coverage and pair-end/mate-pair relationships (Fig. S4), supporting proper assembly quality of the MHC region. Since the MHC region is highly repetitive, a detailed repeat annotation was carried out for both mandrill and human MHC class I regions (from gene *GABBR1* to gene *MICB* in the direction from the telomere side to the centromere side) with the same parameter to find similar repeat content for the two species in this region (48.27% in mandrill compared to 51.03% in human) (Table S13). *HLA* genes are important for immune recognition; thus, mandrill *HLA* genes were further checked and compared to human *HLAs*. In the human MHC class I region, there were 50 genes in total including six *HLAs*, while in mandrill MHC class I region, only four *HLAs* were identified. By searching the entire genome outside the MHC region, another four *HLAs* were identified, making the total number of *HLAs* to be eight in mandrill. However, further inspection of the eight *HLAs* genes in mandrill revealed that five of them harbored start or stop codon changes, or frameshift mutations resulting in premature termination (Fig. 3), reflecting possible differences in immune responses between mandrill and human.

**3.8 Disease-related genomic features.** In order to gain further insight in disease-related mutations in the mandrill, we first performed a comparative genomic analysis of human with mandrill using *lastz*<sup>16</sup>, then identifying SNPs in the mandrill. In total, 557 SNPs in the coding regions were found which were distributed among 520 genes. We collected the information related to mutations from the HGMD database, and checked the presence/absence of such mutations in the mandrill genome. Based on this analysis, we found amino acid changes in 17 genes suggesting disease-related mutation (Table 3). Moreover, we found that some of the mutations are in the functional domains, which may strongly affect the function of these genes (Table S14). These mutations have been associated with disease phenotype in humans, in relation to lung cancer, cranial volume, and asthma.

## Discussion

Primates are well-studied mammals because of their evolutionary importance as well as their close relationship to human. Since the first publication of the gorilla genome in 2012, seven non-human primate genomes have become available and their genomic features comprehensively studied. Despite current progress in primate genomic studies, more genomic data for primate species are needed. Here, by utilizing the high throughput sequencing technologies, we established the draft genome for the mandrill, which is a valuable resource for



**Figure 3.** Synteny of MHC class I region between mandrill and human. Red boxes indicate the MHC genes. Green boxes indicate other protein-coding genes. Yellow and green indicate deletions and insertions in mandrill, respectively.

Gene name	Full name	Position	Wildtype AA	Mutation AA	Disease Description
<i>ALAD</i>	Aminolevulinatase Dehydratase	59	K	N	Amyotrophic lateral sclerosis
<i>CIITA</i>	Class II Major Histocompatibility Complex Transactivator	500	G	A	Multiple sclerosis
<i>CRB1</i>	Crumbs Cell Polarity Complex Component 1	959	G	S	Retinitis pigmentosa
<i>IL4R</i>	Interleukin 4 Receptor	75	I	V	Asthma, atopic
<i>MCPH1</i>	Microcephalin 1	761	A	V	Cranial volume
<i>NPHS2</i>	NPHS2 Stomatins Family Member, Podocin	192	I	V	Nephrotic syndrome
<i>TP53BP1</i>	Tumor Protein P53 Binding Protein 1	353	D	E	Lung cancer

**Table 3.** Genes with disease and its' mutation in mandrill.

primate and human diseases-related studies. The contig N50 of the genome was longer than 20 kb while the scaffold N50 reached 3.56 Mb, indicating good quality of genome assembly. In order to further improve the genome assembly, long reads sequencing might be applied to fill in the gaps of the assembly. Other than that, with further development of technologies like formaldehyde cross-linking and sequencing (Hi-C sequencing), the assembled scaffolds might be further anchored to the chromosomes. The repeat content and gene content in the mandrill were similar to other primate species. According to the phylogenetic tree established based on single-copy gene families, mandrill was found to diverge from the human clade about 28.5 MYA.

In our study, the genomic features of mandrill were comprehensively explored and compared with human. 797 gene families were expanded and enriched including families involved in G-protein coupled receptor activity, olfactory receptor activity, glucose catabolic processes, peptidyl-prolyl isomerization. Furthermore, 657 PSGs were identified in mandrill, 34 of which were found to be innate immunity response genes. Some of these might be involved in rapid initiation of the innate immune responses in mandrill<sup>17</sup>. Several key genes including *STAT1*, *IL5*, *IL1R1*, *ATG5*, *CREB1*, *DICER1*, *PIK3R1* are known to play important roles in the immune system and strongly associated with stress resistance and wound healing.

The mandrill is commonly used as a model for humans with both species possessing characteristic features, including genetic mechanisms underpinning the immune system, language ability, as well as the olfactory system. In terms of the immune system, the MHC regions were specifically analyzed in the mandrill genome. A remarkable synteny was found between the mandrill and the human MHC region with only 54 insertions and deletions (longer than 100 bp). A substantial expansion of olfactory receptor genes was found in mandrill compared to other species, indicating a unique olfactory systems in the mandrill to be further examined.

The assembly and analysis of the draft mandrill genome also emphasize the potential in establishing more genomes for primate species. Primates are an order of mammal species with ~about 16 families and ~500 species, which are all highly evolved animal species with special physiological and behavioral characteristics. Despite the

evolutionary importance and relatively simple genome content, reference genomes have only been established for ~20 species, and the genome assemblies varied in quality and continuity impeding further in depth analysis and applications. Thus, establishing draft genomes by new sequencing technologies for all primate species would be invaluable for evolutionary studies, conservation/preservation, as well as human genetic/disease research and applications.

## Methods

**Sample preparation.** We obtained 5 mL blood from the left jugular vein of an eighteen-year-old male mandrill from Beijing Zoo. The blood was collected in a plastic collection tube with 4% (w/v) sodium citrate, snap-frozen in liquid nitrogen and stored at  $-80^{\circ}\text{C}$ . Genomic DNA was extracted using the AXYGEM Blood and Tissue Extraction Kit according to the manufacturer's instructions. To assess the quality, the extracted DNA was subjected to electrophoresis in 2% agarose gel and stained with ethidium bromide. The DNA concentration was detected by Quant-iT™ PicoGreen® dsDNA Reagent and Kits (Thermo Fisher Scientific, USA) according to the manufacturer's protocol.

**Ethics statement.** Animal blood collection was approved by both Beijing Zoo and BGI-IRB. The study was further carried out based on the agreement between BGI and Beijing Zoo. Moreover, the utility was in accordance with guidelines from the China Council on Animal Care.

**Library establishment and sequencing.** The mandrill DNA was used for library formation, following a previously published protocol<sup>18</sup>. A total of seven libraries were constructed, and sequencing was carried out on the Illumina sequencer HiSeq2000. Of the seven libraries, three were short insert size libraries including insert sizes of 250 bp (sequenced to 150 bp at two ends), 500 bp and 800 bp (sequenced to 100 bp at two ends), respectively. The other four libraries were mate-pair libraries with insert sizes of 2 kb, 5 kb, 10 kb and 20 kb (sequenced to 90 bp at two ends). SOAPnuke was used to filter reads according to the criterions<sup>19</sup>: (i) reads with more than 10% Ns (ambiguous bases); (ii) reads with more than 40% of low-quality bases (quality score less than 10); (iii) reads contaminated by adaptor (adaptor matched 50% with no more than one base mismatch) as well as PCR duplicated reads (identical reads at both ends).

**Genome assembly.** In order to assess genome features, 17-mers (17 bp sub-sequences) were extracted and subjected to the K-mer analysis. Reads from 250 bp, 500 bp and 800 bp insert libraries were used for this analysis.

Then the genome was assembled by short-reads assembly software SOAPdenovo<sup>20</sup> using the filtered data (with parameter settings pregraph-K 35; contig -M 1; scaff). Gaps were filled using paired sequence data from 3 libraries (250, 500, and 800 bp) with -p 31 parameters by GapCloser.

**Transposable elements and repetitive DNA.** RepeatMasker v4.0.5<sup>21</sup> and Repeat-ProteinMask were used to scan the whole genome for known transposable elements in the RepBase library v20.04<sup>22</sup>. Then, RepeatMasker was applied again for identifying *de novo* repeats based on the custom TE library constructed by combining results of RepeatModeler v1.0.8 (RepeatModeler,RRID:SCR\_015027) and LTR\_FINDER v1.0.6<sup>23</sup>. Prediction of tandem repeats was also done using Tandem Repeat Finder v4.0.7<sup>24</sup> with the following setting: Match = 2, Mismatch = 7, Delta = 7, PM = 80, PI = 10, Minscore = 50, MaxPeriod = 2000.

**Protein-coding gene and non-coding RNAs annotation.** After masking known TE repeat elements, genes were predicted using three methods, including homolog based, evidence-based and *ab initio* prediction. For homolog based annotation, protein sequences of *Macaca mulatta* (Ensemble 73 release), *Pan troglodytes* (Ensemble 73 release), *Nomascus leucogenys*, *Pongo abelii* (Ensemble 73 release), *Gorilla gorilla* (Ensemble 73 release), and *Homo sapiens* (Ensemble 73 release) were aligned to the mandrill genome using BLAT<sup>25</sup>. Then GeneWise<sup>26</sup> (Version 2.2.0) was used for further precise alignment and gene structure prediction. For *ab initio* prediction, we employed AUGUSTUS<sup>27</sup> (Version 3.1) to predict gene models in the repeat masked genome. Finally, the gene prediction results were combined using GLEAN<sup>28</sup>. In order to identify the function of the final gene set, three databases (SwissProt, KEGG, and TrEMBL databases) were searched for best matches using BLASTP (version 2.2.26) with an E-value of  $1e-5$ . The InterProScan software<sup>29</sup>, which searches Pfam, PRINTS, ProDom and SMART databases for known motifs and domains, was also used for the gene function annotation.

To identify tRNAs, tRNAscan<sup>30</sup> was used. While for rRNAs identification, 757,441 rRNAs from the public domain were used to search against the genome with command -p blastn -e  $1e-5$ . To identify RNA genes and other non-coding RNA (ncRNA), Rfam database<sup>31</sup> was used to search against the genome.

**Gene family clustering.** Protein sequences of 11 species including *Gorilla gorilla*, *Homo sapiens*, *Macaca mulatta*, *Microcebus murinus*, *Nomascus leucogenys*, *Otolemur garnettii*, *Pan troglodytes*, *Pongo abelii*, *Tarsius syricht*, *Callithrix jacchus*, and *Mus musculus* were used to cluster the gene families. TreeFam (<http://www.treefam.org/>) was used to defined gene families in *Mandrillus sphinx*. Firstly, all-versus-all blastp was used with the e-value cutoff of  $1e-7$  for 12 species and then the possible blast matches were joined together by an in-house program. Thirdly, we removed genes with aligned proportion of less than 33% and converted bit score to percent score. Finally, hcluster\_sg (Version0.5.0, <https://pypi.python.org/pypi/hcluster>) was used to cluster genes into gene families. We selected single-copy orthologs as the set of genes that remained single copy in each species.

**Phylogenetic tree construction.** With gene family clusters defined, the four-fold degenerate (4D) sites of 5,133 single-copy orthologous among the 12 species were extracted for the phylogenetic tree construction. PhyML package<sup>32</sup> was used to build the phylogenetic tree with maximum-likelihood methods and GTR + gamma as an amino acid model (1,000 rapid bootstrap replicates conducted). Based on the phylogenetic tree, divergence

times of these species were estimated by MCMCTree (<http://abacus.gene.ucl.ac.uk/software/paml.html>) with the default parameters. To further calibrate the evolution time in the tree, six fossil dates collected from the TimeTree database (<http://www.timetree.org/>) were used, including the divergence time between *Mus musculus* and human to be 85–93 MYA<sup>33</sup>, divergent time between human and chimpanzee, gorilla, to be 6 MYA (with a range of 5–7)<sup>34</sup> and 9 MYA (range, 8–10)<sup>35</sup>.

With the gene family clustering result, contraction and expansion could be detected to investigate the dynamic evolutionary changes along the phylogenetic tree. According to the phylogenetic tree and divergence time, CAFÉ<sup>36</sup> was used for gene family contraction and expansion analysis.

Demography was estimated using the pairwise sequentially Markovian coalescent (PSMC) model with the following setting: -N25 -t15 -r5 -p “4 + 25\*2 + 4 + 6”. In order to scale the results to real-time, 10 years per generation and a neutral mutation rate of 2.5e-08 per generation were used.

**Gene family expansions/contractions and positively selected genes.** The selection pressure in mandrill was measured by comparing nonsynonymous (dN) and synonymous (dS) substitution rates on protein-encoding genes. This ratio would be equal to 1 if the whole coding sequence evolves neutrally. When dN/dS < 1, it's under constraint, and when dN/dS > 1 it should be under positive selection. PSGs was detected using models in the program package PAML version 3.14, neutral (M1 and M7) and selection (M2 and M8) models were used.

## Data availability

1. NCBI Sequence Read Archive SRP134063 (<https://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP134063>).
2. CNGB Nucleotide Sequence Archive CNP0000251.
3. NCBI Assembly GCA\_004802615.1 ([https://www.ncbi.nlm.nih.gov/assembly/GCA\\_004802615.1](https://www.ncbi.nlm.nih.gov/assembly/GCA_004802615.1)).

Received: 25 June 2019; Accepted: 2 December 2019;

Published online: 12 February 2020

## References

1. International Human Genome Sequencing, C. Initial sequencing and analysis of the human genome. *Nature* **409**, 860, <https://doi.org/10.1038/35057062>, <https://www.nature.com/articles/35057062supplementary-information> (2001).
2. Initial sequence of the chimpanzee genome and comparison with the human genome. *J. Nature* **437**, 69–87, <https://doi.org/10.1038/nature04072> (2005).
3. Cox, L. A. *et al.* Baboons as a model to study genetics and epigenetics of human disease. *ILAR Journal* **54**, 106–121 (2013).
4. Leigh, S. R., Setchell, J. M., Charpentier, M., Knapp, L. A. & Wickings, E. J. Canine tooth size and fitness in male mandrills (*Mandrillus sphinx*). *Journal of Human Evolution* **55**, 75–85 (2008).
5. Wilson, D. E. & Reeder, D. M. *Mammal species of the world: a taxonomic and geographic reference*. (JHU Press, 2005).
6. Zwick, L. S. *et al.* Paratuberculosis in a mandrill (*Papio sphinx*). *Journal of Veterinary Diagnostic Investigation* **14**, 326–328 (2002).
7. O'Rourke, J., Dixon, M., Jack, A., Enno, A. & Lee, A. Gastric B-cell mucosa-associated lymphoid tissue (MALT) lymphoma in an animal model of ‘*Helicobacter heilmannii*’ infection. *The Journal of Pathology* **203**, 896–903 (2004).
8. Setchell, J. M. *et al.* Parasite prevalence, abundance, and diversity in a semi-free-ranging colony of *Mandrillus sphinx*. *International Journal of Primatology* **28**, 1345–1362 (2007).
9. Ungeheuer, M. *et al.* Cellular responses to *Loa loa* experimental infection in mandrills (*Mandrillus sphinx*) vaccinated with irradiated infective larvae. *Parasite immunology* **22**, 173–184 (2000).
10. Roussel, M. *et al.* Modes of transmission of Simian T-lymphotropic Virus Type 1 in semi-captive mandrills (*Mandrillus sphinx*). *Veterinary Microbiology* **179**, 155–161 (2015).
11. Pandrea, I. *et al.* Simian immunodeficiency virus SIVagm. sab infection of Caribbean African green monkeys: a new model for the study of SIV pathogenesis in natural hosts. *Journal of Virology* **80**, 4858–4867 (2006).
12. Kriegs, J. O., Churakov, G., Jurka, J., Brosius, J. & Schmitz, J. Evolutionary history of 7SL RNA-derived SINEs in Supraprimates. *Trends Genet.* **23**, 158–161 (2007).
13. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
14. Raau, R. L., Sterner, K. N., Noviello, C. M., Stewart, C.-B. & Disotell, T. R. Catarrhine primate divergence dates estimated from complete mitochondrial genomes: concordance with fossil and nuclear DNA evidence. *Journal of Human Evolution* **48**, 237–257 (2005).
15. Steiper, M. E. & Young, N. M. Primate molecular divergence dates. *Phylogenetics and Evolution* **41**, 384–394 (2006).
16. Harris, R. S. Improved pairwise Alignment of genomic DNA. (2007).
17. Liu, W. *et al.* Cyclophilin A-regulated ubiquitination is critical for RIG-I-mediated antiviral immune responses. *Elife* **8**, 24425 (2017).
18. Li, R. *et al.* The sequence and de novo assembly of the giant panda genome. *Nature* **463**, 311 (2010).
19. Chen, Y. *et al.* SOAPnuke: A MapReduce Acceleration supported Software for integrated Quality Control and Preprocessing of High-Throughput Sequencing Data. *Gigascience* (2017).
20. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**, 18 (2012).
21. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics* Chapter 4, Unit 4.10 (2009).
22. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
23. Xu, Z. & Wang, H. LTR\_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
24. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**, 573 (1999).
25. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
26. Birney, E., Clamp, M. & Durbin, R. GeneWise and genomewise. *Genome Research* **14**, 988–995 (2004).
27. Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
28. Elsik, C. G. *et al.* Creating a honey bee consensus gene set. *Genome Biology* **8**, R13 (2007).
29. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Research* **32**, D138–D141 (2004).
30. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research* **25**, 955–964 (1997).
31. Gardner, P. P. *et al.* Rfam: updates to the RNA families database. *Nucleic Acids Research* **37**, D136–D140 (2008).
32. Guindon, S., Delsuc, F., Dufayard, J.-F. & Gascuel, O. Estimating maximum likelihood phylogenies with PhyML. *Bioinformatics for DNA Sequence Analysis*, 113–137 (2009).

33. Huchon, D. *et al.* Multiple molecular evidences for a living mammalian fossil. *Proceedings of the National Academy of Sciences* **104**, 7495–7499 (2007).
34. Glazko, G. V. & Nei, M. Estimation of divergence times for major lineages of primate species. *Molecular Biology and Evolution* **20**, 424–434 (2003).
35. Schrago, C. & Voloch, C. The precision of the hominid timescale estimated by relaxed clock methods. *Journal of Evolutionary Biology* **26**, 746–755 (2013).
36. De Bie, T., Cristianini, N., Demuth, J. P. & Hahn, M. W. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22**, 1269–1271 (2006).
37. Rogers, J. & Gibbs, R. A. Comparative primate genomics: emerging patterns of genome content and dynamics. *Nat. Rev. Genet.* **15**, 347–359, <https://doi.org/10.1038/nrg3707>, <http://www.nature.com/nrg/journal/v15/n5/abs/nrg3707.html#supplementary-information> (2014).
38. and Analysis ConsortiumThe Chimpanzee, S. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87, [http://www.nature.com/nature/journal/v437/n7055/supinfo/nature04072\\_S1.html](http://www.nature.com/nature/journal/v437/n7055/supinfo/nature04072_S1.html) (2005).
39. (National Center for Biotechnology Information [online], 2011).
40. Prufer, K. *et al.* The bonobo genome compared with the chimpanzee and human genomes. *Nature* **486**, 527–531, <http://www.nature.com/nature/journal/v486/n7404/abs/nature11128.html#supplementary-information> (2012).
41. Scally, A. *et al.* Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**, 169–175, <http://www.nature.com/nature/journal/v483/n7388/abs/nature10842.html#supplementary-information> (2012).
42. Gordon, D. *et al.* Long-read sequence assembly of the gorilla genome. *Science* **352**, <https://doi.org/10.1126/science.aae0344> (2016).
43. Locke, D. P. *et al.* Comparative and demographic analysis of orang-utan genomes. *Nature* **469**, 529–533, <http://www.nature.com/nature/journal/v469/n7331/abs/10.1038-nature09687-unlocked.html#supplementary-information> (2011).
44. Gibbs, R. A. *et al.* Evolutionary and Biomedical Insights from the Rhesus Macaque Genome. *Science* **316**, 222–234, <https://doi.org/10.1126/science.1139247> (2007).
45. Zimin, A. V. *et al.* A new rhesus macaque assembly and annotation for next-generation sequencing analyses. *Biol. Direct* **9**, 20, <https://doi.org/10.1186/1745-6150-9-20> (2014).
46. Yan, G. *et al.* Genome sequencing and comparison of two nonhuman primate animal models, the cynomolgus and Chinese rhesus macaques. *Nat Biotech.* **29**, 1019–1023, <http://www.nature.com/nbt/journal/v29/n11/abs/nbt.1992.html#supplementary-information> (2011).
47. Rogers, J. *et al.* The comparative genomics and complex population history of Papio baboons. *Science Advances* **5**, eaau6947 (2019).
48. Perry, G. H. *et al.* A Genome Sequence Resource for the Aye-Aye (*Daubentonia madagascariensis*), a Nocturnal Lemur from Madagascar. *Genome Biol. Evol.* **4**, 126–135, <https://doi.org/10.1093/gbe/evr132> (2012).
49. Warren, W. C. *et al.* The genome of the vervet (*Chlorocebus aethiops sabaeus*). *Genome Res.* <https://doi.org/10.1101/gr.192922.115> (2015).
50. Carbone, L. *et al.* Gibbon genome and the fast karyotype evolution of small apes. *Nature* **513**, 195–201, <https://doi.org/10.1038/nature13679>, <http://www.nature.com/nature/journal/v513/n7517/abs/nature13679.html#supplementary-information> (2014).
51. The Marmoset Genome, S. & Analysis, C. The common marmoset genome provides insight into primate biology and evolution. *Nat. Genet.* **46**, 850–857, <https://doi.org/10.1038/ng.3042>, <http://www.nature.com/ng/journal/v46/n8/abs/ng.3042.html#supplementary-information> (2014).
52. Schmitz, J. *et al.* Genome sequence of the basal haplorrhine primate *Tarsius syrichta* reveals unusual insertions. *Nature Communications* **7**, 12997, <https://doi.org/10.1038/ncomms12997> (2016).

## Acknowledgements

This work was supported by the grants of Basic Research Program, the State Key Laboratory of Agricultural Genomics (No. 2011DQ782025), and Guangdong Provincial Key Laboratory of Genome Read and Write (No. 2017B030301011).

## Author contributions

Y.Y. and K.K. conceived the project and H.L. collected the blood samples. Y.Y. and T.Y. performed the data analyses and Z.H., Y.Z., S.Y., W.W. and X.G. conducted other statistical analyses. T.Y. and S.K.S. wrote the first version of the manuscript and revised the manuscript together with K.K. All authors read and approved the final manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41598-020-59110-3>.

**Correspondence** and requests for materials should be addressed to Y.Y. or K.K.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



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

© The Author(s) 2020