Long- read sequencing and de novo assembly of the cynomolgus macaque genome

Bai, Bing; Wang, Yi; Zhu, Ran; Zhang, Yaolei; Wang, Hong; Fan, Guangyi; Liu, Xin; Shi, Hong; Niu, Yuyu; Ji, Weizhi

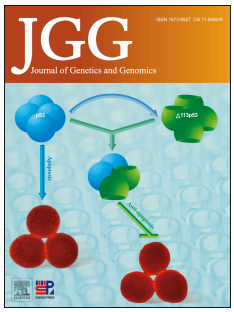
Published in:
Journal of Genetics and Genomics

Link to article, DOI:
10.1016/j.jgg.2021.12.013

Publication date:
2022

Document Version
Peer reviewed version

Journal Pre-proof



Long- read sequencing and *de novo* assembly of the cynomolgus macaque genome

Bing Bai, Yi Wang, Ran Zhu, Yaolei Zhang, Hong Wang, Guangyi Fan, Xin Liu, Hong Shi, Yuyu Niu, Weizhi Ji

PII: S1673-8527(22)00005-4

DOI: https://doi.org/10.1016/j.jgg.2021.12.013

Reference: JGG 1009

To appear in: *Journal of Genetics and Genomics*

Received Date: 24 November 2021

Revised Date: 24 December 2021

Accepted Date: 29 December 2021

Please cite this article as: Bai, B., Wang, Y., Zhu, R., Zhang, Y., Wang, H., Fan, G., Liu, X., Shi, H., Niu, Y., Ji, W., Long- read sequencing and *de novo* assembly of the cynomolgus macaque genome, *Journal of Genetics and Genomics*, https://doi.org/10.1016/j.jgg.2021.12.013.

Long- read sequencing and *de novo* assembly of the cynomolgus macaque genome

Cynomolgus macaques (*Macaca fascicularis*) belong to the Macaca genus. They are the most widespread nonhuman primate (NHP) and share a common ancestor with humans from about 25 million years ago (Kumar and Hedges, 1998). Because of their phylogenetic proximity to humans, macaques are an attractive NHP research model for a wide range of biomedical research (Yan et al., 2011).The extensive distribution of cynomolgus macaques throughout South East Asia (Shiina and Blancher, 2019) and its unique phenotypes including year-round reproductive capability, moderate size, complex behaviors, and similarity to humans in physiology and pathology (Ma et al., 2016) make it a superior option for biological studies. During recent decades, cynomolgus macaques have been used increasingly, accelerating progress in several research fields, including development and reproduction (Tan et al., 2021), stem cell therapy (Guo and Li, 2015), disease model construction and genome editing (Chen et al., 2017), immunotherapies (Iwasaki et al., 2019), and degenerative diseases and aging (Verdier et al., 2015).

A high-quality reference genome is the backbone of the effective application of cynomolgus macaques in biological research. Currently, the most frequently used reference genome is Macaca_fascicularis_5.0 (MF 5.0) which was sequenced and assembled through next-generation sequencing (NGS) methods (https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/364/345/GCF_000364345.1_Macaca_fascicularis_5.0/GCF_000364345.1_Macaca_fascicularis_5.0_genomic.fna.gz). However, the inherent technical properties of short-read sequencing leads to deficits in the contiguity and completeness of the MF 5.0 (contig N50: 0.08 Mb; total gap length: 142.98 Mb), consequently causing limits to its usages in downstream analyses. With long-read sequencing and multi-platform scaffolding strategy, it is possible to assemble high-quality genomes, resolve highly complex genomic regions, such as highly repetitive and GC-rich regions, and better support scientific goals (He et al., 2019).

In this study, we performed *de novo* genome assembly to build a cynomolgus macaque reference genome (MaFaC) using long-read sequencing data and multiple scaffolding strategies. In detail, high-quality DNA was extracted from the peripheral blood of a male cynomolgus macaque that was chosen by us (Fig. S1A). At first, 95 long-read SMRT cells were sequenced (about 65-fold genome coverage, subreads N50 11 Kb) and contigs were assembled by using FALCON (Chin et al., 2016). To further improve the accuracy of the assembly, Illumina short-read data from four monkeys (about 60-fold genome coverage), which included the monkey used for long-read sequencing, were used to correct remaining errors in the sequence contigs using Pilon (Walker et al., 2014) (Table S1). Then, the Hi-C data (48.72-fold genome coverage) was used to link contigs to their respective chromosomes (Fig. S1B). Finally, we constructed a *de novo* chromosome-scale cynomolgus macaque genome assembly (MaFaC) with a total length of 3.06 Gb, contig N50 of 6.03 Mb, and scaffold N50 of 152.38 Mb (Table S2). The new assembly includes fewer fragments and higher contiguity than MF 5.0 (7,254 *vs.* 88,075, 11-fold reduction in contig numbers), and in terms of length, both contig N50 and scaffold N50 were improved by 75-fold and 1.7-fold, respectively (Figs. 1A and S1C; Table S2). MF 5.0 was built from a female monkey sample with no Y chromosome, our assembly makes up for this deficiency.

We then used MaFaC to close gaps in the MF 5.0 reference genome. In total, 53.4% of the N-gaps (76,119 gaps, 137.6 Mb) in the MF 5.0 chromosomes were filled by 73.6 Mb long-read sequences (Fig. S1C; Table S3). A total of 2094 gaps (36.1 Mb) out of the 45,359 defined gaps mapped to protein-coding regions (detailed information is given in Supplementary Data S1), 17,373, 6, and 25,886 gaps fell within intron, splicing, and intergenic regions, respectively (Fig. 1B). MaFaC not only improves the completeness of the genome but also provides more intact gene structures.

To validate the reliability and quality of the MaFaC assembly, we assessed it in the following aspects: sequence consensus with MF 5.0, sequencing quality,

mappability, availability for variants, orientation errors and resolution of high repetitive regions, for example, major histocompatibility locus (MHC). As shown in Table S4, the average sequence consensus between MaFaC and MF 5.0 reached 95%. Alignment of all chromosomes shows high consistency between these two assemblies (Figs. 1C and S2), which proves the high reliability of MaFaC. We mapped NGS data (about 50× depth) of the same cynomolgus macaque used for long-reads sequencing to the MaFaC assembly to calculate the base error rate in the MaFaC assembly. As shown in Table S5, the average base error rate of each chromosome is $1.55923\times e^{-5}$, indicating that the MaFaC assembly is of high quality. To evaluate the mappability and availability for variants, a set of five unrelated cynomolgus macaque NGS data (about 60× depth) (Wang et al., 2020) were mapped to MF 5.0 and MaFaC assemblies. Mapping rates were significantly increased for MaFaC compared to MF 5.0 (99.67% *vs*. 99.22%, respectively; $P < 0.001$, two-tailed paired *t*-test; Fig. 1D; Table. S6). Genome-wide variant calling shows similar numbers of nearly all mutant types including SNPs (Single Nucleotide Polymorphism), indels, and SVs (Structure Variant: deletions, insertions, and inversions as well as duplications) between the two sets of references, attesting to the reliability of MaFaC for variant detection (Table S7). Four largest inversions (3.7 Mb – 16 Mb) between MaFaC and MF 5.0 were picked to assess orientation errors in MaFaC. As shown in Fig.1E, all of these regions were assembled by reads longer than 10 Kb with average sequencing depth about 60×, which guaranteed the reliability and the high orientation accuracy of MaFaC. Then, both MaFaC and MF 5.0 were mapped to the human (GRCh38) MHC region, and three regions matched. Compared to MF 5.0, MaFaC correctly assembled an additional about 53 Kb regions in the second matched position covering four more MHC genes (Fig. 1F; Table S8), which illustrates that MaFaC is much better at resolving highly repetitive regions than MF 5.0.

For annotation, we combined the *de novo*, homology-based and transcriptome-based methods to predict protein-coding genes in the MaFaC. Our annotation landscape is similar to MF 5.0 and other primates (Table S9). We annotated 21,446

protein-coding genes in the MaFaC, slightly higher than that of MF 5.0 (20,757). MaFaC also has a longer average length compared to MF5.0 (1659.46 *vs*. 1525.74 bp, respectively). The improvement of gene annotation in the new genome could partly be attributed to the successful rescue of missing exons and genes located in closed gap regions (Table S3). Orthologous gene family comparison between MaFaC, MF 5.0, rhesus macaque, gorilla, chimpanzee, and human was implemented using OrthoMCL (Li et al., 2003). As shown in Fig. S3, all the species have similar gene family composition, indicating the reliability of our gene set. To further evaluate the integrity of gene annotation, we conducted benchmarking universal single-copy ortholog (BUSCO) analysis by BUSCO (v3.0.1) with the Mammalian odb9 set of 4104 genes. We identified 4062 BUSCO genes (99.0%) in the MaFaC annotation (Table S10). To analyze the repetitive sequences in MaFaC, we used a combination of *de novo* repeat prediction and homology-based search tools, RepeatMasker (Tarailo-Graovac and Chen, 2009), and Tandem Repeat Finder (Benson, 1999), and found that 46.25% (1.42 Gb) of the genome assembly contained repetitive DNA. Although the repetitive ratio of MaFaC is lower than in other primates, it is in the same range as MF 5.0 (Table S11).

Long-read sequencing data could increase SV detection sensitivity (He et al., 2019), which affects gene expression and phenotype and has become increasingly important in genetic and evolutionary studies (He et al., 2019; Sharp et al., 2006). Therefore, we aligned subreads of long-read PacBio data to the cynomolgus macaque reference genome (MF 5.0) to call SVs. In total, 32,028 SVs were detected and 22,772 retained after filtering (Fig. S4A). The final set contains 12,207 deletions, 368 duplications, 10,081 insertions, and 116 inversions with a medium length of 320,498, 298, and 3831, respectively (Fig. S4B and S4C). In total, SVs span 248 Mb and cover 8.1% of the genome. In addition, we found that 1.2% (257 SVs) map within exons, 44% (10,026 SVs) in introns, and the remaining 54.8% (12,489 SVs) in intergenic regions. Compared to an array of comparative genomic hybridization (Array CGH)-based SV set obtained from a previous study (Gschwind et al., 2017), only seven

records from the previous dataset (8656 SV in total) correspond to our SV set. In other words, 99.9% of our SV set is novel, which emphasizes the merits of long-read sequencing data for genome-wide SV detection. In subsequent genotyping analysis using the five unrelated cynomolgus macaque mentioned initially (see supplementary materials: Mappability), we found 7223 (31.71%) SVs were identified in these samples, of which 3692 (51.11%) are fixed and 3531 (48.89%) are polymorphic. Synonymous to previous long-read sequencing studies (Sharp et al., 2006; He et al., 2019), a vast majority of 9956 (98.76%) insertions were invalidated, highlighting the limitation of short-read sequencing in the detection of insertions. Our SV set enormously enriched the original cynomolgus macaque SV dataset which could be used in further primate genetic studies.

As a popular NHP model, cynomolgus macaques have been used repeatedly to construct genetically modified models of human diseases (Chen et al., 2017; Wang et al., 2020). The more intact genomic regions reconstructed by MaFaC provide a good opportunity to design a reliable gRNA library for each gene to support CRISPR experiments in the future. Thus, we designed CRISPR target sequences for each of the 21,446 genes in the MaFaC genome. In total, 5,241,674 gRNAs (20 bp) were designed with an average of about 244 for each gene (Fig. S5A). We selected gRNA with the highest-ranking score (Hsu et al., 2013) and analyzed their key features. We found that 3396 genes could be designed with at least one gRNA with no off-target site. However, more than half of the designed gRNAs associate with at least one possible off-target site (Fig. S5B). Although many gRNA were not perfect, they have high scores (Fig. S5C) as off-target effects will decrease or vanish because of high mismatches of possible off-target sites, which means they may be used in actual experiments. As for mismatches of possible off-target sites for each gene, about 25.6% of the gRNA have only one identical sequence (0 mismatches) in the genome (Fig. S5D), indicating that these gRNA have a high off-target effect. Taking more factors into account in addition to the ranking scores used above will further improve the selection of the best gRNA. The gRNA identified remains vital for future genomic

research. Furthermore, off-target assessment is a crucial step in evaluating the efficiency and accuracy of Cas9-editing (Luo et al., 2019). All mutations including SNPs, indels, and SVs in this study could be used as a highly confident spectrum to implement an off-target assessment. Additionally, we identified SNPs and indels of another five unrelated cynomolgus macaques (Wang et al., 2020) using both MaFaC and MF 5.0 as a reference to further enrich the cynomolgus macaque wild mutations spectrum.

In this study, we generated a high-quality genome assembly of cynomolgus macaque. Compared to the old reference genome (MF 5.0), MaFaC improved the aspect of contiguity and intactness in both genic and non-genic regions. We also identified a large number of SVs, the vast majority of which were not characterized in a previous study (Gschwind et al., 2017), which demonstrates the superiority of long-read sequencing in whole-genome SV detection and the rarity of cynomolgus macaque genomic studies. In addition, we provided high-quality gRNAs for each gene to facilitate the usage of cynomolgus macaques in Cas9 genome editing. All mutants called in our study are a valuable resource and of great applicable values for studies including primate evolution and genetics as well as genome editing. Detail information of data availability can be found in supplementary materials. In further studies, we will improve assembly and annotation accuracy to further support the usage of cynomolgus macaque as an NHP model for biomedical studies.

Conflict of interest

The authors declare that they have no competing interests.

Acknowledgments

This research was supported by the Natural Science Foundation of Yunnan Province (202001BC070001, 202102AA100053), the National Natural Science Foundation of

China General Program (No. 32070580). We also thank Newton O. Otecko for proof reading.

Reference

Benson, G., 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27, 573–580.

Chen, Y., Yu, J., Niu, Y., Qin, D., Liu, H., Li, G., Hu, Y., Wang, J., Lu, Y., Kang, Y., et al., 2017. Modeling rett syndrome using TALEN-Edited MECP2 mutant cynomolgus monkeys. Cell 169, 945–955.

Chin, C.S., Peluso, P., Sedlazeck, F.J., Nattestad, M., Concepcion, G.T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., et al. 2016. Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods 13, 1050-1054.

Gschwind, A.R., Singh, A., Certa, U., Reymond, A., Heckel, T., 2017. Diversity and regulatory impact of copy number variation in the primate Macaca fascicularis. BMC Genomics 18, 144.

Guo, X., Li, X.J., 2015. Targeted genome editing in primate embryos. Cell Res. 25, 767–768.

He, Y., Luo, X., Zhou, B., Hu, T., Meng, X., Audano, P.A., Kronenberg, Z.N., Eichler, E.E., Jin, J., Guo, Y., et al., 2019. Long-read assembly of the Chinese rhesus macaque genome and identification of ape-specific structural variants. Nat. Commun. 10, 4233.

Hsu, P.D., Scott, D.A., Weinstein, J.A., Ran, F.A., Konermann, S., Agarwala, V., Li, Y., Fine, E.J., Wu, X., Shalem, O., et al., 2013. DNA targeting specificity of RNA-guided Cas9 nucleases. Nat. Biotechnol. 31, 827–832.

Iwasaki, K., Uno, Y., Utoh, M., Yamazaki, H., 2019. Importance of cynomolgus

monkeys in development of monoclonal antibody drugs. Drug Metab. Pharmacokinet. 34, 55–63.

Kumar, S., Hedges, S.B., 1998. A molecular timescale for vertebrate evolution. Nature 392, 917–920.

Li, L., Stoeckert, C.J., Jr., Roos, D.S., 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13, 2178–2189.

Luo, X., He, Y., Zhang, C., He, X., Yan, L., Li, M., Hu, T., Hu, Y., Jiang, J., Meng, X., et al., 2019. Trio deep-sequencing does not reveal unexpected off-target and on-target mutations in Cas9-edited rhesus monkeys. Nat. Commun. 10, 5525.

Ma, Y., Li, J., Wang, G., Ke, Q., Qiu, S., Gao, L., Wan, H., Zhou, Y., Xiang, A.P., Huang, Q., et al., 2016. Efficient production of cynomolgus monkeys with a toolbox of enhanced assisted reproductive technologies. Sci. Rep. 6, 25888.

Sharp, A.J., Cheng, Z., Eichler, E.E., 2006. Structural variation of the human genome. Annu. Rev. Genomics Hum. Genet. 7, 407–442.

Shiina, T., Blancher, A., 2019. The cynomolgus macaque MHC polymorphism in experimental medicine. Cells 8.

Tan, T., Wu, J., Si, C., Dai, S., Zhang, Y., Sun, N., Zhang, E., Shao, H., Si, W., Yang, P., et al., 2021. Chimeric contribution of human extended pluripotent stem cells to monkey embryos ex vivo. Cell 184, 2020–2032 e2014.

Tarailo-Graovac, M., Chen, N., 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr. Protoc. Bioinformatics *Chapter 4*, Unit 4 10.

Verdier, J.M., Acquatella, I., Lautier, C., Devau, G., Trouche, S., Lasbleiz, C., Mestre-Frances, N., 2015. Lessons from the analysis of nonhuman primates for understanding human aging and neurodegenerative diseases. Front Neurosci. 9, 64.

225 Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo,
226 C.A., Zeng, Q., Wortman, J., Young, S.K., Earl, A.M., 2014. Pilon: an integrated
227 tool for comprehensive microbial variant detection and genome assembly
228 improvement. PLoS One 9, e112963.

229 Wang, F., Zhang, W., Yang, Q., Kang, Y., Fan, Y., Wei, J., Liu, Z., Dai, S., Li, H., Li,
230 Z., et al. 2020. Generation of a Hutchinson-Gilford progeria syndrome monkey
231 model by base editing. Protein Cell 11, 809–824.

232 Yan, G., Zhang, G., Fang, X., Zhang, Y., Li, C., Ling, F., Cooper, D.N., Li, Q., Li, Y.,
233 van Gool, A.J., et al., 2011. Genome sequencing and comparison of two
234 nonhuman primate animal models, the cynomolgus and Chinese rhesus macaques.
235 Nat. Biotechnol. 29, 1019–1023.

Bing Bai[1], Yi Wang[1], Ran Zhu[1],

State Key Laboratory of Primate Biomedical Research, Institute of Primate Translational Medicine, Kunming University of Science and Technology, Kunming 650500, China

Yunnan Key Laboratory of Primate Biomedical Research, Kunming 650500, China

Yaolei Zhang[1],

BGI-Qingdao, BGI-Shenzhen, Qingdao 266555, China

Translational Immunology group, Department of Biotechnology and Biomedicine, Technical University of Denmark, Lyngby, Denmark

Hong Wang[1],

State Key Laboratory of Primate Biomedical Research, Institute of Primate Translational Medicine, Kunming University of Science and Technology, Kunming 650500, China

Yunnan Key Laboratory of Primate Biomedical Research, Kunming 650500, China

Guangyi Fan, Xin Liu,

BGI-Qingdao, BGI-Shenzhen, Qingdao 266555, China

BGI-Shenzhen, Shenzhen 518083, China

China National GeneBank, BGI-Shenzhen, Shenzhen 518120, China

Hong Shi,

State Key Laboratory of Primate Biomedical Research, Institute of Primate Translational Medicine, Kunming University of Science and Technology, Kunming 650500, China

Yunnan Key Laboratory of Primate Biomedical Research, Kunming 650500, China

Yuyu Niu,

State Key Laboratory of Primate Biomedical Research, Institute of Primate Translational Medicine, Kunming University of Science and Technology, Kunming 650500, China

Yunnan Key Laboratory of Primate Biomedical Research, Kunming 650500, China

Faculty of Life Science and Technology, Kunming University of Science and Technology, Kunming 650500, China

Weizhi Ji[*].

State Key Laboratory of Primate Biomedical Research, Institute of Primate Translational Medicine, Kunming University of Science and Technology, Kunming 650500, China

Yunnan Key Laboratory of Primate Biomedical Research, Kunming 650500, China

[*]*Corresponding author.*

Email address: wji@lpbr.cn (W. Ji)

[1] These authors contributed equally to this work.

Figures Legends

Fig. 1. Quality assessment of the MaFaC assembly. **A**: The contig length distribution in MF 5.0 and MaFaC. **B**: Distribution of gap lengths in MF 5.0 filled by long-reads of MaFaC and gene annotation results. **C**: Sequence consensus of chr1 in MF 5.0 and MaFaC assembly. **D**: Boxplot summary of the mapping rate of five short-read NGS data mapped to the MF 5.0 and MaFaC assemblies. A two-tailed paired *t* test was used for statistical assessment. ***, $P < 0.001$. **E**: Long-read coverage of four largest inversions between MaFaC and MF 5.0. **F**: Circos illustration of the alignment of human MHC regions and MF5.0 and MaFaC. RL, read length.