Supplemental Information

1. Sequencing techniques for unambiguous assembly of difficult genomic regions

Recently, Nurk and colleagues reported the first *de novo* sequencing of a complete human genome using ultra-long Oxford Nanopore Technologies (ONT) sequencing reads (Nurk et al. 2022). This human genome build (T2T-CHM13) includes five new chromosome arms representing an additional 8% of the genome that had not been previously sequenced due to technical limitations. Nanopore sequencing enables intact DNA fragments as long as 4 megabase pairs (Mb) to be sequenced in their entirety (Genome assembly 2022). This allows repetitive genomic blocks to be bridged by unambiguously assembling ultra-long DNA reads into chromosome-level scaffolds (Jain et al. 2018; Nurk et al. 2022). While individual ONT sequencing reads have a higher per-base error rate than other sequencing platforms, the ultra-long scaffolds can be combined with higher-accuracy sequencing for error correction. Traditionally researchers used Illumina short-read sequencing for this purpose, but recent improvements in Pacific Biosciences (PacBio) methods for HiFi circular consensus sequencing routinely produce high-accuracy reads of >10 kilobase pairs (kb) (Wenger et al. 2019; Nurk et al. 2020). This provides improved error correction versus short reads through increased confidence in unique mapping across the scaffold assembly with an ability to distinguish between even repetitive sequence elements. Researchers are beginning to use PacBio HiFi sequencing for gap filling and sequence polishing of ONT scaffolds to create unprecedentedly complete high-quality assemblies (Kakuk et al. 2021; Nurk et al. 2022; Wang et al. 2022).

2. Macaque MHC nomenclature

As with human HLA class I and class II genes, nomenclature and maintenance of databases of macaque MHC class I and class II alleles is complicated. The ImmunoPolymorphism Non-Human Primates MHC Database (IPD-MHC NHP) was established for this purpose (https://www.ebi.ac.uk/ipd/mhc/group/NHP/) (Maccari et al. 2017; de Groot et al. 2020). A full, detailed nomenclature report for nonhuman primate (NHP) MHC alleles is released periodically;

the latest was created in 2019 (de Groot et al. 2020). Here we present a few highlights from this report that are particularly relevant to the cynomolgus macaque MHC class I and class II names throughout this paper.

NHP allele nomenclature largely follows the rules established by the World Health Organization Nomenclature Committee for Factors of the HLA System wherever possible (Marsh et al. 2010). Each allele name consists of up to six parts with different field separators in the following style:

1-2*3:4:5:6

*Mafa-B*098:05:01:01*

Part 1 is the species identifier, derived from a four-letter abbreviation of the species' scientific name. The first two letters are from the genus, and the last two letters are from the species – *Macaca fascicularis* is thus *Mafa*. Part 1 is separated from part 2 with a hyphen. Part 2 is the gene designation. Part 2 is separated from part 3 with an asterisk. Parts 3 through 6 are separated from each other with colons and provide increasingly specific differences between alleles. Part 3 is the allele group; any alleles with the same allele group designation have a high degree of sequence homology to each other. Part 4 defines distinct proteins – *Mafa-B*098:05* differs from *Mafa-B*098:01* in one or more amino acids. Part 5 is used to show any synonymous substitutions within the coding region, and part 6 is used to show any differences in a non-coding region. Not all NHP alleles will contain all six parts since different NHP alleles have been identified by different groups from different nucleic acid templates (RNA/cDNA versus genomic DNA) using different primers. An allele defined only from RNA/cDNA template can only contain up to five parts since no non-coding sequence is defined. Any alleles identified from gDNA should contain all six parts.

NHP gene designations are assigned largely based on homology to HLA genes – *Mafa-B* is orthologous to *HLA-B*, etc. However, there are some NHP-specific designations used in this paper. First, since the macaque MHC classical class I genes are duplicated, gene designations contain Arabic numerals wherever possible. This is best observed in the MHC class I A allele designations – *Mafa-A1*063:02:01:01*, *Mafa-A2*05:11:01:01*, and *Mafa-A4*01:01:01* are defined to be different alleles assigned to different genes (*Mafa-A1*, *Mafa-A2*, and *Mafa-A4*) on the same haplotype. The specific configurations of genes for other duplicated regions, however, are not as well-defined. This is best observed in the MHC class I B region, where a subset of gene names (*Mafa-B11L*01:06:01:01N*, *Mafa-B17*01:05:01:01*, etc.) contain Arabic numerals based on sequence homology to *MHC-B* genes defined from the well-characterized bacterial artificial chromosome library CHORI-250 (Daza-Vamenta et al. 2004; Shiina et al. 2006). The remaining *Mafa-B* genes do not currently contain Arabic numeral gene designators, though the names will likely be refined in the future based on studies of full macaque MHC haplotypes like the one presented in this paper. Macaque class II *MHC-DRB* genes likewise contain a mixture of genes with and without Arabic numeral designations. Since the human *HLA-DRB* gene is also duplicated, there are definitions for *HLA-DRB1* through *HLA-DRB9* genes. When macaques show significant similarity to one of those *HLA-DRB* genes, they are assigned as such (*Mafa-DRB1*10:02:01:01*, *Mafa-DRB6*01:09:01:01*, etc.). However, when a macaque *MHC-DRB* allele is yet to be assigned to a specific gene, there is no Arabic numeral, and the allele group is preceded by a *W* for "workshop" designation (*Mafa-DRB*W049:01:01:01*).

A few other points about macaque nomenclature: macaques have *MHC-AG* and *MHC-I* genes that are not present in humans. As covered in more detail in this paper, the *MHC-G* genes in macaques are pseudogenes, with their functionality taken over by *MHC-AG* genes. The *MHC-I* genes are an oligomorphic *B*-like locus located within the MHC class I B region (on the MCM M3 haplotype, the third *MHC-B*-like coding gene from the telomeric end of the MHC class I B region is one such allele – *Mafa-I*01:10:01:02*) (Urvater et al. 2000). An "N" suffix is used occasionally to indicate a null allele characterized by an early stop codon, and “Ps” can be used as a prefix or suffix to indicate a presumed (but sometimes not officially confirmed) pseudogene.

3. Assessment of the region of homozygosity flanking the MHC in cy0333

To determine the boundaries of the homozygous region in cy0333 that flank our ~5.2 Mb assembly of the MHC M3 haplotype (OP204634), we analyzed SNP patterns in whole genome sequence datasets from a cohort of 18 MCM that were described previously (Ericsen et al. 2014). Illumina short-read data from cy0333 and 17 additional MCM (NCBI BioProject: PRJNA257343) were mapped against the Macaca_fascicularis_5.0 (GCF_000364345) assembly, and sequence variants were called with GATK version 3.3. The resulting gVCFs were merged into a single VCF and annotated with SnpEff. This VCF (16485.baylor-01.mafa5.ann.vcf.gz) is available at https://go.wisc.edu/svvlra and with the Supplemental Scripts. SNP patterns in this VCF were manually inspected with the Integrative Genomics Viewer version 2.11.4 and this analysis was restricted to SNPs with genotype quality scores of 99.

Examination of the SNP patterns for the *GABBR1*, *MOG*, and *ZFP57* genes at the telomeric boundary of the MHC region revealed the expected profiles for this MCM whole genome sequence cohort that includes six samples homozygous for the MHC M3 haplotype (cy0333, cy0334, cy0335, cy0336, cy0337, and cy0329), six samples homozygous for the MHC M1 haplotype (cy0320, cy0321, cy0322, cy0323, cy0324, cy0325) and six samples heterozygous for MHC M1 and M3 haplotypes (cy0326, cy0327, cy0328, cy0330, cy0331, and cy0332). Analysis of SNP patterns in the whole genome data for cy0333 demonstrated that the homozygous region of Chromosome 4 extends 4.7 Mb upstream of the 3’UTR of the *GABBR1* gene (Supplemental Table S3). The last homozygous SNP for cy0333 lies in the interval between the *TDP2* and *KIAA0319* genes at position 145,936,430 on Chromosome 4. A pair of heterozygous SNPs appear 319 and 350 bp telomeric to this position on Chromosome 4. In the centromeric direction from *KIFC1*, the last homozygous SNP was detected in the *ENSMFAG00000016485* gene at position 134,954,427 on Chromosome 4; this corresponds to an interval of 2.4 Mb downstream of the *KIFC1* gene. Taken together, these observations indicate that the region of homozygosity in cy0333 that contains the MHC region is at least 11.0 Mb. Given the highly collapsed nature of the *Mafa-A* and *Mafa-B* genomic regions of the Macaca_fascicularis_5.0 assembly (Fig. 3), the actual region of homozygosity in cy0333 is likely to be at least 1.3 Mb greater than the 11 Mb estimate derived from this SNP analysis.

4. References for Supplemental Material

Daza-Vamenta R, Glusman G, Rowen L, Guthrie B, Geraghty DE. 2004. Genetic divergence of the rhesus macaque major histocompatibility complex. *Genome Res* **14**: 1501–1515.

de Groot NG, Otting N, Maccari G, Robinson J, Hammond JA, Blancher A, Lafont BAP, Guethlein LA, Wroblewski EE, Marsh SGE, et al. 2020. Nomenclature report 2019: major histocompatibility complex genes and alleles of Great and Small Ape and Old and New World monkey species. *Immunogenetics* **72**: 25–36.

Ericsen AJ, Starrett GJ, Greene JM, Lauck M, Raveendran M, Deiros DR, Mohns MS, Vince N, Cain BT, Pham NH, et al. 2014. Whole genome sequencing of SIV-infected macaques identifies candidate loci that may contribute to host control of virus replication. *Genome Biol* **15**: 478.

Genome assembly. 2022. *Oxford Nanopore Technologies*. https://nanoporetech.com/applications/investigation/assembly (Accessed September 30, 2022).

Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, et al. 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol* **36**: 338–345.

Kakuk B, Tombácz D, Balázs Z, Moldován N, Csabai Z, Torma G, Megyeri K, Snyder M, Boldogkői Z. 2021. Combined nanopore and single-molecule real-time sequencing survey of human betaherpesvirus 5 transcriptome. *Sci Rep* **11**: 14487.

Maccari G, Robinson J, Ballingall K, Guethlein LA, Grimholt U, Kaufman J, Ho C-S, de Groot NG, Flicek P, Bontrop RE, et al. 2017. IPD-MHC 2.0: an improved inter-species database for the study of the major histocompatibility complex. *Nucleic Acids Res* **45**: D860–D864.

Marsh SGE, Albert ED, Bodmer WF, Bontrop RE, Dupont B, Erlich HA, Fernández-Viña M, Geraghty DE, Holdsworth R, Hurley CK, et al. 2010. Nomenclature for factors of the HLA system, 2010. *Tissue Antigens* **75**: 291–455.

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al. 2022. The complete sequence of a human genome. *Science* **376**: 44–53.

Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, Miga KH, Eichler EE, Phillippy AM, Koren S. 2020. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res* **30**: 1291–1305.

Shiina T, Ota M, Shimizu S, Katsuyama Y, Hashimoto N, Takasu M, Anzai T, Kulski JK, Kikkawa E, Naruse T, et al. 2006. Rapid evolution of major histocompatibility complex class I genes in primates generates new disease alleles in humans via hitchhiking diversity. *Genetics* **173**: 1555–1570.

Urvater JA, Otting N, Loehrke JH, Rudersdorf R, Slukvin II, Piekarczyk MS, Golos TG, Hughes AL, Bontrop RE, Watkins DI. 2000. *Mamu-I*: A Novel Primate MHC Class I*B*-Related Locus with Unusually Low Variability. *The Journal of Immunology* **164**: 1386–1398. http://dx.doi.org/10.4049/jimmunol.164.3.1386.

Wang B, Yang X, Jia Y, Xu Y, Jia P, Dang N, Wang S, Xu T, Zhao X, Gao S, et al. 2022. High-quality Arabidopsis thaliana Genome Assembly with Nanopore and HiFi Long Reads. *Genomics Proteomics Bioinformatics* **20**: 4–13.

Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, et al. 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* **37**: 1155–1162.