RESEARCH ARTICLE SUMMARY

NONHUMAN GENOMICS

Sequence diversity analyses of an improved rhesus macaque genome enhance its biomedical utility

Wesley C. Warren* *et al.*

INTRODUCTION: The rhesus macaque (*Macaca mulatta*) is one of the most widely used nonhuman primate (NHP) models for studying human biology and disease. As a representative of the Old World monkey lineage, its genetic sequence is also critical for studies of primate evolution.

RATIONALE: Because of the central role of rhesus macaques in both biomedical research and primate adaptation, we sought to generate a new reference genome for this NHP in which most gaps were closed and most protein-coding genes were annotated. A more comprehensively annotated macaque genome and extensive sequencing of individual macaques from existing research populations enables the characterization of standing genetic variation. Understanding the extent of genetic variation among research populations under phenotypic surveillance will identify new models of human genetic disease and allow for the further development of NHP models for investigating aspects of genome function such as gene regulation.

RESULTS: We sequenced and assembled the genome of a female rhesus macaque of Indian origin using a multiplatform genomics approach that included long-read sequencing, extensive manual curation, and experimental validation. With the exception of humans, the resulting assembly is one of the most complete primate references to date, with 99.7% of the gaps now closed and >99% of the genes represented. We generated 6.5 million full-length transcripts and used these to create a comprehensive set of protein-encoding and noncoding gene models, including the identification of new macaque isoforms and gene candidates.

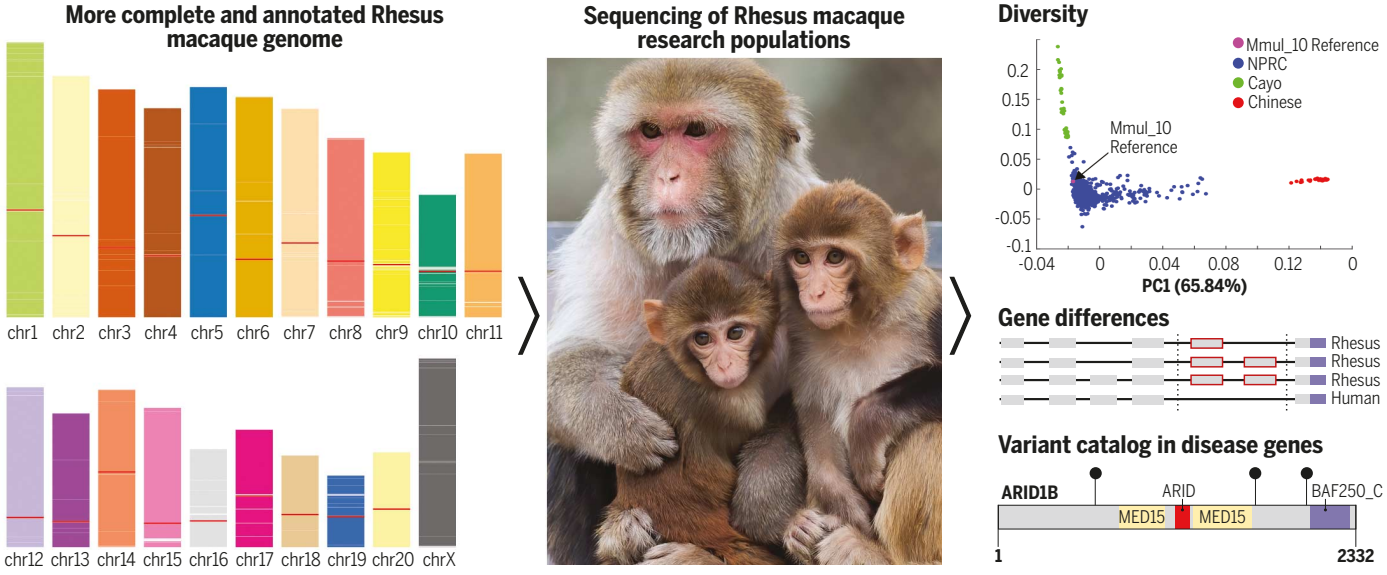
The more complete macaque genome overcomes many of the limitations of the previous assemblies. Segmental duplications are improved threefold, leading to the characterization of lineage-specific genes and gene families (e.g., *ZNF669*) that have expanded recently during evolution. Most full-length, active mobile elements have been resolved at the sequence level and are now integrated into the genome assembly instead of being fragmented and unassigned. In the case of LINEs, this has led to a reclassification of the order of appearance of active elements during Old World monkey evolution. Human-macaque gene comparisons identify a limited number of lineage-specific exon changes of potential functional effect, including the formation of isoforms that distinguish the two species.

We generated whole-genome sequence data for 850 rhesus macaques from captive U.S. research colonies and three wild-caught Chinese samples, including 133 previously published samples. We used these data to identify 85.7 million single-nucleotide variants (SNVs; 21.3 million singletons) in addition to 10.5 million indels, generating the most extensive collection of segregating genetic variants for any NHP species. We can now confirm that research rhesus macaques are more than twice as diverse per individual as humans, with the average macaque carrying 9.7 million SNVs, and used this variation to understand the genetic diversity of existing research populations. We also identified potentially deleterious mutations in macaque genes that are intolerant to mutation in humans. Such mutations segregating in rhesus macaque research centers offer the opportunity to develop new genetic models of disease.

CONCLUSION: This new macaque reference genome and the genetic characterization of research populations will substantially advance biomedical research and studies of primate genome evolution by providing an improved framework for more complete studies of genetic variation and its phenotypic consequence. ■

The list of author affiliations is available in the full article online.
*Corresponding author. Email: warrenwc@missouri.edu; jr13@bcm.edu (Jeffrey Rogers); eee@gs.washington.edu (Evan E. Eichler)
Cite this article as: W. C. Warren *et al.*, *Science* **370**, eabc6617 (2020). DOI: 10.1126/science.abc6617

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abc6617



Genetic diversity in the rhesus macaque. A more completely assembled and annotated macaque reference genome (left panel) coupled to sequencing of research populations (middle panel) provides a deep understanding of diversity, functional changes in gene models, and rare variants that may be used to develop better genetic models of disease (right panels). Photo credit: Kathy West.

RESEARCH ARTICLE

NONHUMAN GENOMICS

Sequence diversity analyses of an improved rhesus macaque genome enhance its biomedical utility

Wesley C. Warren[1,2,3]*, R. Alan Harris[4], Marina Haukness[5], Ian T. Fiddes[6], Shwetha C. Murali[7,8], Jason Fernandes[9], Philip C. Dishuck[7], Jessica M. Storer[10,11], Muthuswamy Raveendran[4], LaDeana W. Hillier[7], David Porubsky[7], Yafei Mao[7], David Gordon[7,8], Mitchell R. Vollger[7], Alexandra P. Lewis[7], Katherine M. Munson[7], Elizabeth DeVogelaere[5], Joel Armstrong[5], Mark Diekhans[5], Jerilyn A. Walker[10], Chad Tomlinson[12], Tina A. Graves-Lindsay[12], Milinn Kremitzki[12], Sofie R. Salama[9], Peter A. Audano[7], Merly Escalona[9], Nicholas W. Maurer[9], Francesca Antonacci[13], Ludovica Mercuri[13], Flavia A. M. Maggiolini[13], Claudia Rita Catacchio[13], Jason G. Underwood[14], David H. O'Connor[15], Ashley D. Sanders[16], Jan O. Korbel[16], Betsy Ferguson[17], H. Michael Kubisch[18], Louis Picker[19], Ned H. Kalin[20], Douglas Rosene[21], Jon Levine[22,23], David H. Abbott[23,24], Stanton B. Gray[25], Mar M. Sanchez[26,27], Zsofia A. Kovacs-Balint[26], Joseph W. Kemnitz[23,28], Sara M. Thomasy[29,30], Jeffrey A. Roberts[31], Erin L. Kinnally[31,32], John P. Capitanio[31,32], J. H. Pate Skene[33], Michael Platt[34], Shelley A. Cole[35], Richard E. Green[9], Mario Ventura[13], Roger W. Wiseman[15], Benedict Paten[5], Mark A. Batzer[10], Jeffrey Rogers[4]*, Evan E. Eichler[7,8]*

The rhesus macaque (*Macaca mulatta*) is the most widely studied nonhuman primate (NHP) in biomedical research. We present an updated reference genome assembly (Mmul_10, contig N50 = 46 Mbp) that increases the sequence contiguity 120-fold and annotate it using 6.5 million full-length transcripts, thus improving our understanding of gene content, isoform diversity, and repeat organization. With the improved assembly of segmental duplications, we discovered new lineage-specific genes and expanded gene families that are potentially informative in studies of evolution and disease susceptibility. Whole-genome sequencing (WGS) data from 853 rhesus macaques identified 85.7 million single-nucleotide variants (SNVs) and 10.5 million indel variants, including potentially damaging variants in genes associated with human autism and developmental delay, providing a framework for developing noninvasive NHP models of human disease.

A detailed understanding of nonhuman primate (NHP) genome evolution is key to recognizing the origins of human traits and identifying putative disease genes. Evolutionary analyses of a diverse range of NHP genomes spanning the breadth of primate phylogeny from great apes to prosimians have begun to uncover the genetic basis of this phenotypic and biochemical diversity. Comparisons among species reveal lineage-specific changes in retroelements, the death and birth of duplicated genes, including the segmental duplications (SDs) underlying them, and functionally relevant, sometimes deleterious, mutations in genes associated with human disease (*1*–*5*). Collectively, these studies are beginning to illuminate the history and new mechanisms of molecular and phenotypic adaptation.

Rhesus macaques (*Macaca mulatta*) play a specific and critical role in both evolutionary comparisons and biomedical research. Although the great apes (chimpanzees, bonobos, gorillas, and orangutans) are phylogenetically closer to humans, the rhesus macaque is an essential model for a wide variety of studies related to infectious disease, neurobiology, developmental psychology, and other elements of primate (including human) biological function (*6*). This significance is highlighted by the role that rhesus macaque models have played in our understanding of AIDS pathogenesis and prevention strategies such as preexposure prophylaxis regimes (*7*), the development of highly effective Ebola vaccines (*8*), and the notable results obtained by editing genes related to risk for autism (*9*).

In 2007, the first whole-genome analysis of the rhesus macaque revealed both fundamental genetic similarities to and interesting differences from the human genome (*2*). By combining our new rhesus reference assembly with the high-quality genomes now available for the great apes (*5*), more extensive reconstruction of human genome evolution is possible, such as gene structural changes that are specific to humans and apes. Early analyses suggested that macaques showed reduced SD content and complexity compared with humans, although a final determination required a higher-quality genome assembly (*2*). Similarly, initial analyses reported an expanded and more complex major histocompatibility complex (MHC) loci in macaques, but the organization of such loci has been difficult to resolve (*10*). Finally, macaques provide valuable models for diseases or processes that would not be adequately modeled in rodents (*11*–*13*). Naturally occurring variation, which is higher among macaques (*3*, *4*), has been leveraged to develop improved genetic models of Mendelian disorders (*14*, *15*), complex disease (*16*, *17*), and a hereditary form of cancer (*11*). To improve our understanding of rhesus macaque genetic diversity and its future translational implications, we annotated this new macaque reference by extensively characterizing genomic variation among 853 Indian- and Chinese-origin rhesus macaques from U.S. research colonies. Therefore, this work provides a roadmap for naturally occurring mutations and disease models.

[1]Department of Animal Sciences, Bond Life Sciences Center, University of Missouri, Columbia, MO 65211, USA. [2]Department of Surgery, School of Medicine, University of Missouri, Columbia, MO 65211, USA. [3]Institute of Data Science and Informatics, University of Missouri, Columbia, MO 65211, USA. [4]Human Genome Sequencing Center, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA. [5]Computational Genomics Laboratory, University of California–Santa Cruz, Santa Cruz, CA 95064, USA. [6]Inscripta Inc., Boulder, CO 80301, USA. [7]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA. [8]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA. [9]Department of Biomolecular Engineering, University of California–Santa Cruz, Santa Cruz, CA 95064, USA. [10]Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, USA. [11]Institue for Systems Biology, Seattle, WA 98109, USA. [12]McDonnell Genome Institute, Washington University, St. Louis, MO 63108, USA. [13]Department of Biology, University of Bari 'Aldo Moro', 70125 Bari, Italy. [14]Pacific Biosciences of California, Seattle, WA 94025, USA. [15]Department of Pathology and Laboratory Medicine, Wisconsin National Primate Research Center, University of Wisconsin–Madison, Madison, WI 53711, USA. [16]European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany. [17]Division of Genetics, Oregon National Primate Research Center, Oregon Health and Science University, Beaverton, OR 97006, USA. [18]Tulane National Primate Research Center, Covington, LA 70433, USA. [19]Oregon National Primate Research Center and Vaccine and Gene Therapy Institute, Oregon Health Sciences University, Beaverton, OR 97006, USA. [20]Department of Psychiatry, University of Wisconsin School of Medicine and Public Health, Madison, WI 53719, USA. [21]Department of Anatomy and Neurobiology, Boston University School of Medicine, Boston, MA 02118, USA. [22]Department of Neuroscience, University of Wisconsin, Madison, WI 53175, USA. [23]Wisconsin National Primate Research Center, University of Wisconsin, Madison, WI 53171, USA. [24]Department of Obstetrics and Gynecology, Wisconsin National Primate Research Center, University of Wisconsin, Madison, WI 53715, USA. [25]The University of Texas MD Anderson Cancer Center, Michale E. Keeling Center for Comparative Medicine and Research, Bastrop, TX 78602, USA. [26]Yerkes National Primate Research Center, Atlanta, GA 30329, USA. [27]Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, Atlanta, GA 30329, USA. [28]Department of Cell and Regenerative Biology, University of Wisconsin, Madison, WI 53706, USA. [29]Department of Surgical and Radiological Sciences, School of Veterinary Medicine, University of California–Davis, Davis, CA 95616, USA. [30]Department of Ophthalmology and Vision Science, School of Medicine, University of California–Davis, Davis, CA 95817, USA. [31]California National Primate Research Center, Davis, CA 95616, USA. [32]Department of Psychology, University of California, Davis, CA 95616, USA. [33]Department of Neurobiology, Duke University School of Medicine, Durham, NC 27710, USA. [34]Department of Neuroscience, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA. [35]Population Health Program, Texas Biomedical Research Institute and Southwest National Primate Research Center, San Antonio, TX 78227, USA.
*Corresponding author. Email: warrenwc@missouri.edu (W.C.W.); jr13@bcm.edu (J.R.); eee@gs.washington.edu (E.E.E.)

Results

Sequencing and assembly

We generated long-read sequence data (~66-fold single-molecule, real-time sequence coverage) and assembled the genome of a female Indian-origin rhesus macaque (2.9 Gbp) using a series of genomic methods followed by extensive manual curation (*18*). This reference (Mmul_10, GenBank accession number GCA_003339765.3) consists of 20 autosomes and the X chromosome; for completeness, we added a previous bacterial artificial chromosome (BAC)–based representation of the Y chromosome (*19*). Only 3.3% (97 Mbp) of the assembled macaque genome remains unassigned to scaffolds and is highly repetitive (>85% repeat-masked bases >10 kbp in length). In total, the new macaque reference is represented by 2979 scaffolds with contig and scaffold N50 lengths of 46 and 82 Mbp, respectively (Table 1). This resulted in a highly contiguous and accurate assembly (Fig. 1) owing to extensive manual curation, gap filling, and unique sequence directionality assessments (*18*). Overall contiguity increased by 120- and 5.6-fold compared with the previous Indian (Mmul_8.0.1) and Chinese (rheMacS) (*20*) rhesus macaque assemblies, respectively (Table 1). The underlying raw sequence data and other data are available at the National Center for Biotechnology Information (NCBI) (table S1).

Quality assessment

We performed a series of analyses to evaluate the quality of the Mmul_10 reference assembly. For example, we assessed accuracy and contiguity using macaque BAC-end sequence data estimating values of 99.7 and 92.5%, respectively, consistent with allelic variation among individual macaques. Detailed analyses of even complex regions such as the MHC genes (*10*) and the killer immunoglobulin–like receptor (KIR) gene families (*21*) showed that most of these loci were assembled accurately [tables S2 and S3 and (*18*)]. Using orthogonal sequencing datasets from the same sample, we estimate an overall assembly sequence accuracy of one error every 20,000 bp.

More than 99.7% of the gaps present in the previous Indian-origin rhesus macaque genome assembly are now closed. Eleven of the 20 macaque autosomes are now represented as two scaffolds separated only by an unassembled centromere. Although there is high overall synteny with Mmul_8.0.1, rheMacS, and macFas_5.0 [chromosome 3 (Fig. 1B and fig. S1)], there are 39 large orientation errors identified in the original Mmul_8.0.1 assembly that our assembly was able to correct (*22*) [chromosomes 13 (fig. S2) and 2 (fig. S3)]. To develop a more accurate assembly for this species, we investigated other potential orientation issues by generating Strand-seq data from an unrelated macaque (*18*), compared it with the two Indian macaque assemblies, and identified "homozygous inversions" as potential errors in orientation. In total, we detected 82 (130,115,998 bp) potentially misoriented regions in Mmul_8.0.1, in contrast to 13 (3,800,615 bp) in the new reference assembly (Fig. 1C and tables S4 and S5). These data indicate that the number of misoriented genes has been reduced from 4.83% to only 0.13% in Mmul_10. Many of these unresolved regions, not unexpectedly, map to structurally diverse and complex immune gene families such as the MHC and KIR regions (in the latter, a short homozygous inversion is predicted) (fig. S4). The Strand-seq analysis also detected one chimeric scaffold where a terminal part of chromosomal scaffold CM014356.1 belongs to the beginning of chromosomal scaffold CM014355.1 (fig. S5); this has now been corrected.

Gene annotation

We assessed the completeness of gene annotation by applying benchmarking universal single-copy ortholog (BUSCO) scores, which measure the representation of highly conserved mammalian genes (*23*). We found that 99.6% of BUSCO genes were annotated, with only 0.1% missing—an improvement over the Chinese macaque assembly (rheMacS), in which 1.8% are missing (*20*) (table S6). Analyzing a curated set of 6422 *M. mulatta* RefSeq transcripts (NCBI) showed an average coverage of 99.8%, confirming the high degree of completeness. Predicted protein-coding gene comparisons within each annotation pipeline, NCBI and Ensembl, for rhesus macaque and humans showed consistency (table S7). In rhesus macaques, NCBI and Ensembl produced similar outcomes: 21,121 and 21,748 genes, respectively (table S7). There was a substantial improvement in noncoding RNAs (ncRNAs) identification in Mmul_10, with 8720 new ncRNAs (table S7), especially long noncoding RNAs (lncRNAs) [tables S8 and S9 and (*18*)]. Many of the missing genes map to more complex regions of the genome or are duplicated genes that have expanded or are contracted differentially compared with those of humans (tables S10 and S11).

To further define macaque-specific transcript and protein isoform diversity, we generated and sequenced 6.5 million full-length cDNAs from macaque brain, induced pluripotent stem cell lines, and testes (table S12). We applied the Comparative Annotation Toolkit (CAT) (*24*) to annotate 17,838 protein-coding and 31,873 noncoding macaque genes. The set includes 83,692 protein-coding isoforms, of which 5353 were identified from Iso-Seq data as potentially new isoforms (fig. S6). A total of 980 genes have frameshifting indels disrupting all isoforms; RNA-sequencing–based cleanup reduced the proportion of isoforms with frameshifting indels from 16.5 to 4.9%. The final CAT gene set includes 80,248 protein-coding transcripts aligned in a 1-1 fashion, with the remaining 3444 paralogous isoforms being resolved with alignment metrics. We note that 550 gene structures were split over multiple contigs (table S13), and 827 protein-coding genes show evidence of being part of gene families that exhibit reduced copy number in this assembly relative to that of humans, with 473 of those showing a 2-1 relationship and 295 being 3-1 in humans compared with macaques (table S10). By contrast, 967 protein-coding genes showed evidence of gene family expansion in macaques, with 711 copied once and 116 copied twice in rhesus compared with humans (table S11).

New exon adaptation in macaques

CAT predicted 2880 new transcripts that did not arise from any previously annotated transcript in the input human annotation. Of these, 2812 maintained open reading frames. From this set, we manually curated 84 new exons

Table 1. Macaque (*M. mulatta*) genome assembly comparisons.

Assembly version	N50 contig [Mbp]	Total size [Mbp]	Total contigs	Chromosome gaps*	Unplaced bases [Mbp]	Protein-coding genes	Missing genes [%]†
Mmul_8.0.1	0.107	2835	348,579	62,231	82	21,574	3.8
rheMacS_1.0‡	8	3031	4743	2862	82	20,389	1.8
Mmul_10	46	2936	3182	203	97	21,121	0.4

*Spanned gaps as designated by NCBI assembly file format. †Missing genes based on BUSCO analysis of 4104 total mammalian conserved genes and He *et al*. (*20*). ‡Ab initio gene predictions according to He *et al*. (*20*).

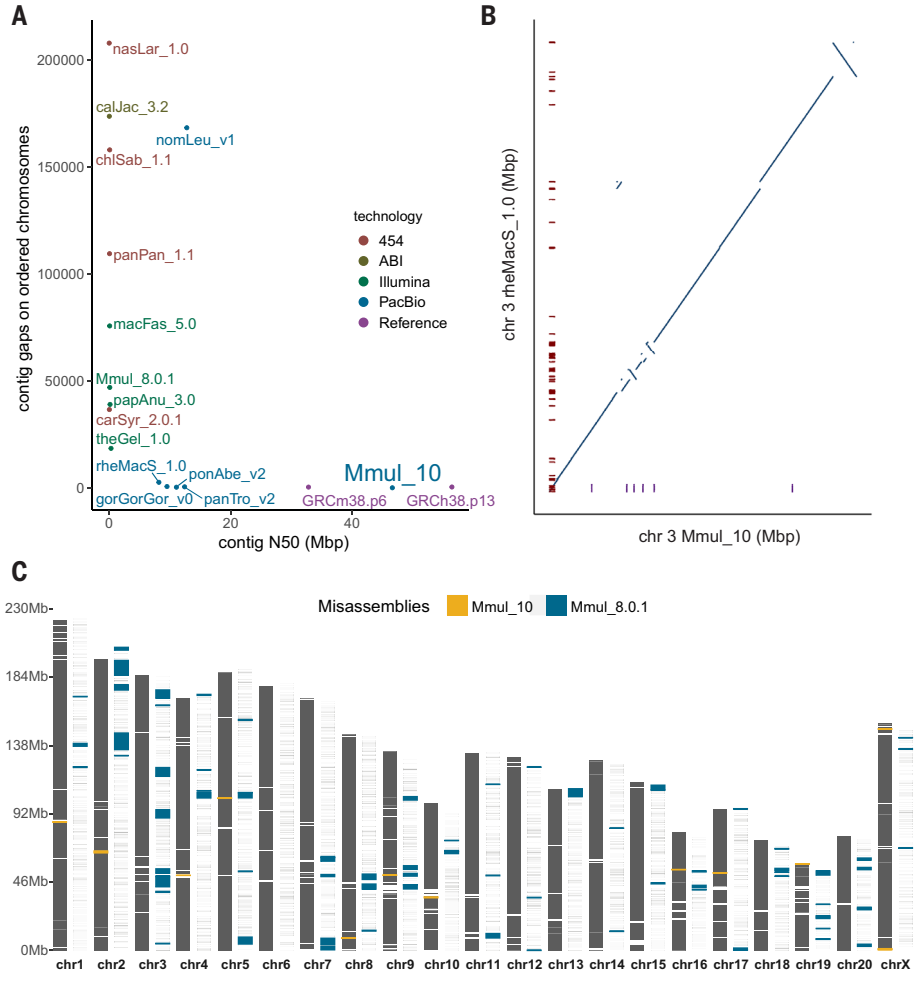


Fig. 1. Rhesus macaque genome assembly quality and contiguity. (**A**) Number of gaps and contig N50 lengths compared among mammalian genomes and color-coded based on sequencing technology. The contiguity of macaque (Mmul_10) is comparable to human (GRCh38) and mouse (GRCm38.p6) reference genomes. (**B**) Number of gaps (red ticks) compared against a synteny plot of Chinese (rheMacS) and Indian (Mmul_10) macaque chromosome 3 assemblies. (**C**) Comparison of potential orientation misassemblies based on Strand-seq analysis (*18*). Mmul_10 shows far fewer (yellow; n = 13) inversions compared with an earlier macaque assembly, Mmul_8.0.1 (blue; n = 82), predicting bases that are 34× less misoriented; 99.7% of gaps (white rectangles) in the earlier assembly are now closed.

with Iso-Seq support (Fig. 2 and table S14). We searched the translated protein sequence in the Pfam 32.0 database (*25*) using this set and found extensive homology to notable protein families. For example, three transcripts (Rhesus_T0212625, Rhesus_T0212626, and Rhesus_T0212627) correspond to a new gene model (Rhesus_G0055137 on chromosome 9: 95,447,150 to 95,611,700) that shares homology with the human CYP2C18 protein and has abundant Iso-Seq transcript support across multiple macaque tissues (Fig. 2A). The predicted mRNA sequence for tropoelastin (*ELN*) has two identifiable exons near the C terminus of the protein, with Iso-Seq support from six tissues for three alternatively spliced isoforms of *ELN*. These exons are shared to the base of the mammalian tree, but are not found in humans, great apes, or lesser apes, suggesting a specific loss of these exons in the hominoid lineage (Fig. 2B).

We explored additional exon adaptations that were potentially specific to rhesus macaques. The first is a deletion of 64 bases in the macaque genome affecting *MYO3A* (fig. S7A). This deletion leads to a new isoform of *MYO3A* that has altered exon structure and is supported by Iso-Seq data from multiple tissues. Second is an isoform of *GAS8* in which a new exon leads to a frameshift in downstream exons, creating a premature stop codon that is alternatively spliced as confirmed by Iso-Seq (fig. S7B). The third example is a rhesus-specific 6250-bp insertion in *DCHS2* that introduces a new exon to one of the *DCHS2* isoforms (fig. S7C). Once again, this new *DCHS2* exon is supported by Iso-Seq reads from testes tissue, where it is alternatively spliced. However, the Iso-Seq transcripts do not support the original CAT gene annotation predictions as a whole. Rather than a predicted exon-skipping event, the new exon appears to correspond to an alternate first exon of the gene. Additional experimental work will be required to determine the functional impact of these genic differences.

SD analyses

We analyzed Mmul_10 for recent duplications (*18*) and identified 111.5 Mbp of assembled SDs (≥1 kbp and ≥90% sequence identity). In principle, this represents a >3-fold improvement compared with the analysis of the first macaque assembly (*2*), in which only 32 Mbp were characterized as SDs. Despite this improvement, 54% of the duplicated base pairs remain unlocalized (figs. S8 and S9). Nevertheless, most of those assigned to a chromosome are clustered or distributed interchromosomally among pericentromeric and subtelomeric regions (fig. S10). Only 8% (755/9475) of the pairwise alignments are intrachromosomal and separated by at least 1 Mbp (fig. S10), for example, the collapsed SD of *NXF2* on the X chromosome (fig. S11).

We identified 276 regions of collapsed duplications (*26*) corresponding to 9.1 Mbp of the genome (table S15) and estimate that these correspond to 41.8 Mbp of SDs not yet properly integrated into the macaque assembly. Fluorescence in situ hybridization (FISH) analyses classified most of these (74%) as pericentromeric based on signals mapping to either side of the centromere. To resolve the sequence of these collapses, we applied a graph-based approach (*26*, *27*) to resolve SDs based on clustering and assembling reads using diagnostic paralogous sequence variants. This method resolved 168 of the 276 collapses into 531 distinct contigs representing 19.8 Mbp of SD sequence (contig N50 = 37.4 kbp). Among these are highly accurate sequence contigs corresponding to recently expanded rhesus macaque gene families, including MHC, olfactory receptor, and zinc finger genes. We have deposited these contigs into GenBank under BioProject PRJNA662298 as a resource that may be used to improve gene annotation. For example, we identified nine assembly collapses (20 to 92 kbp) corresponding to *ZNF669* genes in the Mmul_10 assembly (table S15). Segmental Duplication Assembler (SDA) assembled 53 contigs from these collapses, generating an additional 1.9 Mbp of assembled sequence (N50 = 36.5 kbp), and identified three contigs in which *ZNF669* Iso-Seq mapped with higher identity than the original Mmul_10 assembly (Fig. 3, A and B). Translation of the full-length cDNA confirmed open reading frames and duplicated gene models that had been missed by our initial annotation of the genome (Fig. 3C). FISH analysis of a large-insert BAC corresponding to one of the loci confirmed a

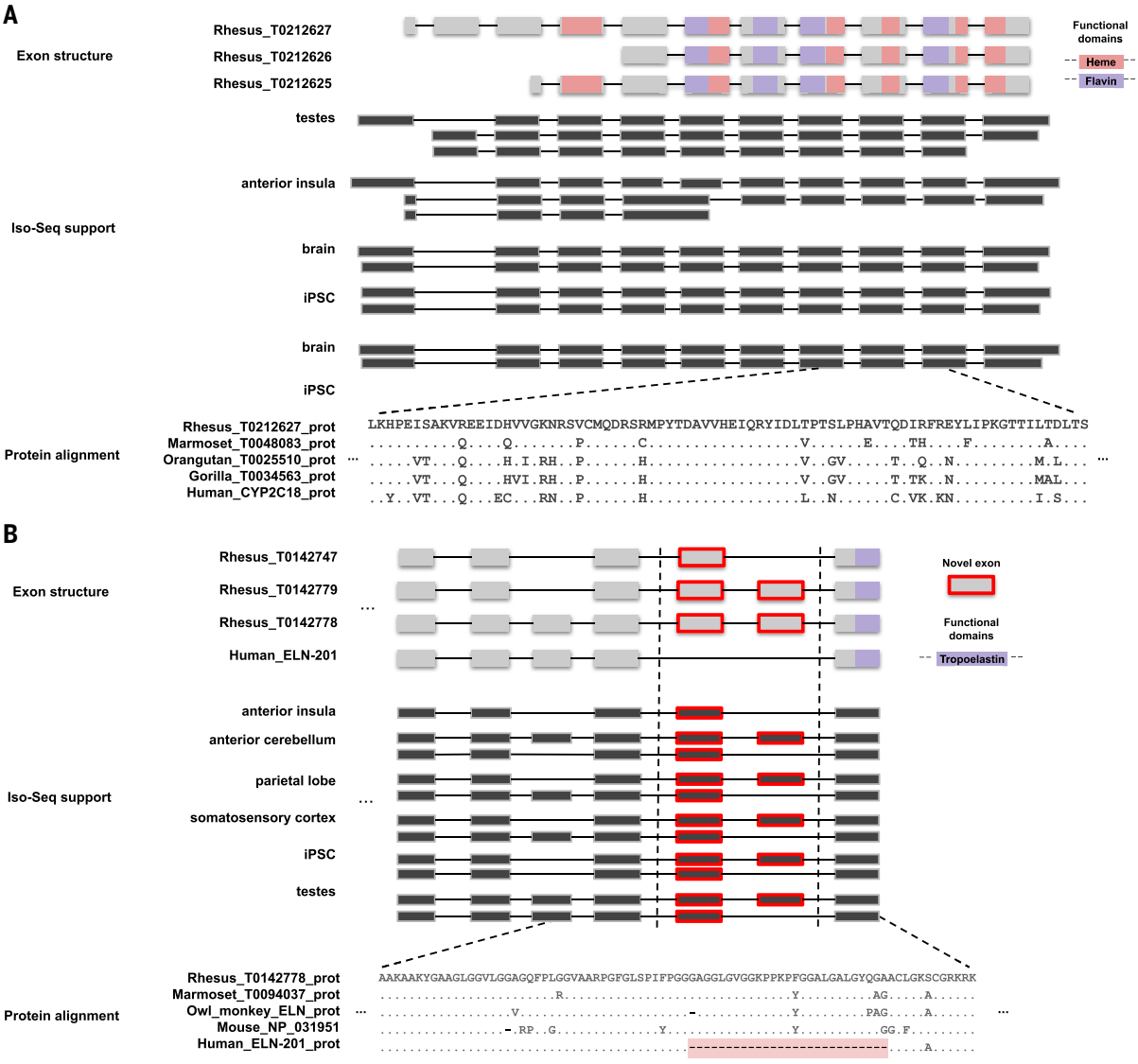


Fig. 2. New genes and gene models. (**A**) New gene model with homology to the cytochrome p450 protein family predicted by the AugustusPB mode of CAT. The gene structure and protein domain architecture of three isoforms are shown (top). The predictions arose from supporting Iso-Seq reads from five tissues (middle). Orthologous new genes are also predicted in marmoset, orangutan, and gorilla assemblies; a protein alignment (bottom) of those genes along with a human CYP2C18 protein is shown. (**B**) Two macaque isoforms in *ELN* (tropoelastin) are predicted by the AugustusPB mode of CAT and are supported by macaque Iso-Seq data but differ from human by two exons. The gene structure and functional domains for the last seven exons of this gene are shown (top), along with a comparison with a human transcript model. These two protein-encoding exons are also observed in marmoset, owl monkey, and mouse, but not in apes, as a result of an ape-specific deletion (bottom) that changed the gene structure of tropoelastin.

cluster of *ZNF669* genes mapping to chromosome 6 as well as several additional duplicated loci of this gene family distributed throughout the macaque genome (Fig. 3D).

Repetitive sequence analyses

Overall, fewer mobile element insertions were identified in Mmul_10 compared with Mmul_8.0.1 (*18*) (table S16), including lineage-specific elements across various retrotransposon classes (table S16). Despite these reductions, subfamily network analyses for both *Alu* (fig. S12) and LINE1 (L1) elements show increases in the number and connectivity of younger subfamilies, particularly full-length L1 (Fig. 4, A and B). There was an increase in full-length potentially active L1 elements (n = 6892 versus n = 4380) in Mmul_10 compared with Mmul_8.0.1 (table S16). Full-length L1 elements are less fragmented and more likely to be assigned to chromosomes as opposed to being mapped to unlocalized contigs (Fig. 4C). Similarly, the new assembly moves 8291 unlocalized *Alu* elements to specific chromosomal assignments (Fig. 4D). Thirty-three percent of potentially full-length endogenous repeat elements (long terminal repeat > 7 kbp in length) now map to different chromosomal locations in the newer assembly (fig. S13), consistent with improvements in the sequence resolution and integration of longer and potentially active repeat elements.

Recurrent deletions in full-length L1RS elements

Given the better representation of full-length repeats, we searched for systematic changes in the 5′ untranslated region (UTR) structure of the primate-specific L1PA subfamily from which L1 rhesus-specific (L1RS) retroelements originate (*28*). The 5′ UTRs of these selfish elements are often targeted by host factors, e.g., KRAB zinc finger (*KZNF*) proteins, which

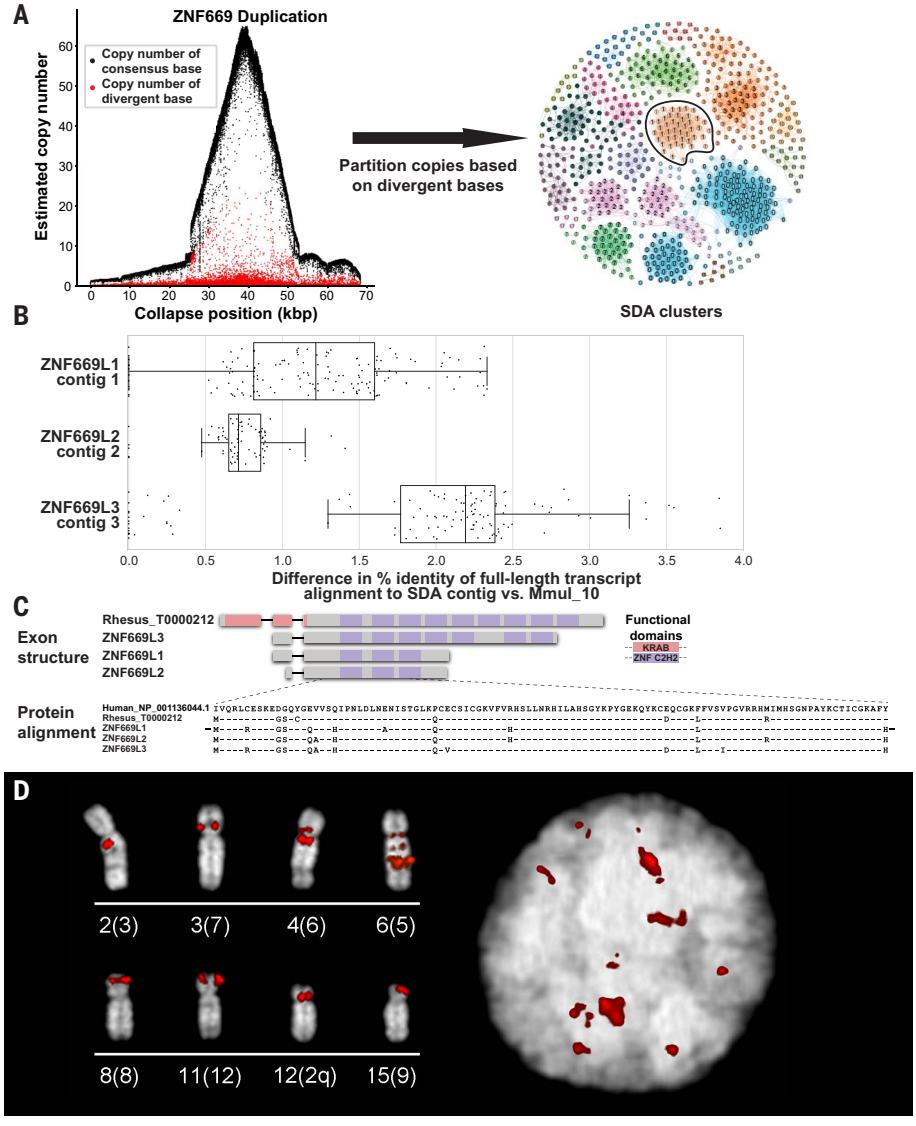


Fig. 3. Macaque *ZNF669* gene family expansion. (**A**) A 68-kbp region of collapsed assembly corresponding to the *ZNF669* gene family as indicated by the excess read depth and increased number of paralogous sequence variants (PSVs, red dots) that are diverged compared with the consensus sequence (black dots). The highly identical copies were thus unresolved in Mmul_10 and predicted to be present in ~50 copies in macaque (left). SDA partitioned the long reads into 19 distinct paralog clusters (colored and numbered) based on shared PSVs and assembled these clusters into 18 contigs. Vertices reflect individual PSVs and edges represent long-read sequences that contain both of the connected PSVs (right). SDA partitioned and assembled the remaining *ZNF669* collapses into 35 additional contigs. The outlined PSV cluster corresponds to contig 2 in (B). (**B**) Mapping of full length non-chimeric (FLNC) transcripts shows that they align better to SDA-resolved contigs than the original assembly. (**C**) Annotation of these genes shows that these three contigs encode a highly expanded *ZNF669* gene family in which there are FLNC data supporting complete open reading frames that differ by only a few amino acids. (**D**) FISH with BAC CH250-540H16 as a probe corresponding to a *ZNF669* locus demonstrates interchromosomal duplications (red) on interphase nucleus (right) and metaphase chromosomes (left), labeled by chromosome (human syntenic chromosome in parentheses).

repress their transcription and as a result are differentially expanded across primate lineages (*29*). We mapped 20,541 full-length L1RS elements to the human L1PA5 consensus sequence (UCSC Repeat Browser) and identified specific coverage drops that accrue and persist in subsets of L1RS elements (*28*). Using these deletion patterns and subfamily designations, we propose an order for the evolution of these 5′ UTR deletion events (Fig. 5A) and suggest a reclassification of L1RS nomenclature. Although older families are typically assigned a higher number, the appearance of coverage drops establishes that the L1RS16 family predates the L1RS21 family (Fig. 5B). Furthermore, the results indicate that after the human–rhesus macaque divergence, at least three different regions of the L1RS 5′ UTR were altered. A comparative analysis of these L1RS elements in the genomes of other Old World monkeys (OWMs: rhesus macaques, crab-eating macaques, baboons, and golden snub-nosed monkeys, with humans as an outgroup) supports this adaptive model (Fig. 5B) because all OWM species display changes at these positions; however, the actual sequence changes and size of the region varies, suggesting that these events recurred or have been refined at different points along the OWM phylogenetic tree (fig. S14). Two of these three sites (sites 1 and 3) also overlap changes observed in active, human-specific L1 elements, supporting the hypothesis that the deletions result from independent recurrent parallel evolution in the primates. The remaining site (site 2) is restricted to OWMs because no coverage drop is observed at this site in young full-length L1 human elements. The existence of a deletion at this site in all OWMs suggests that a repressive factor was present in the OWM common ancestor and highlights the distinctiveness of L1RS2 transcript diversity (Fig. 5C).

Genetic diversity among macaque research populations

The U.S. research colonies of rhesus macaques were founded primarily with animals imported from India decades ago, although a much smaller number of Chinese-origin rhesus macaques have been added to some colonies over time. It is not possible in all cases to trace the exact geographic origins of the U.S. research population, but sufficient information is available to identify many animals as derived from either Indian- or Chinese-origin founders. We generated WGS data for 850 rhesus macaques from captive U.S. research colonies and three wild-caught Chinese samples, including 133 previously published samples (*3*). Most of the samples (n = 810) were designated as being of Indian origin, and the remaining individuals were of Chinese or suspected admixed origin. Most samples were sequenced to at least 20-fold coverage (n = 764; average 33.69-fold), with the remaining (n = 89) sequenced to an average of 8.58-fold for the detection of SNVs and indels (figs. S15 and S16). SNVs and indels were identified based on mapping reads to Mmul_10 (*18*), with an overall 2.12 transition to transversion ratio consistent with prior studies (*3*). We identified 85.7 million SNVs, including 21.3 million singletons in addition to 10.5 million indels (Table 2), creating the most extensive collection of segregating genetic variants for any NHP species (table S17). By comparison, a recent study of 929 human genomes from 54 diverse global populations, sequenced to average 35-fold coverage, identified

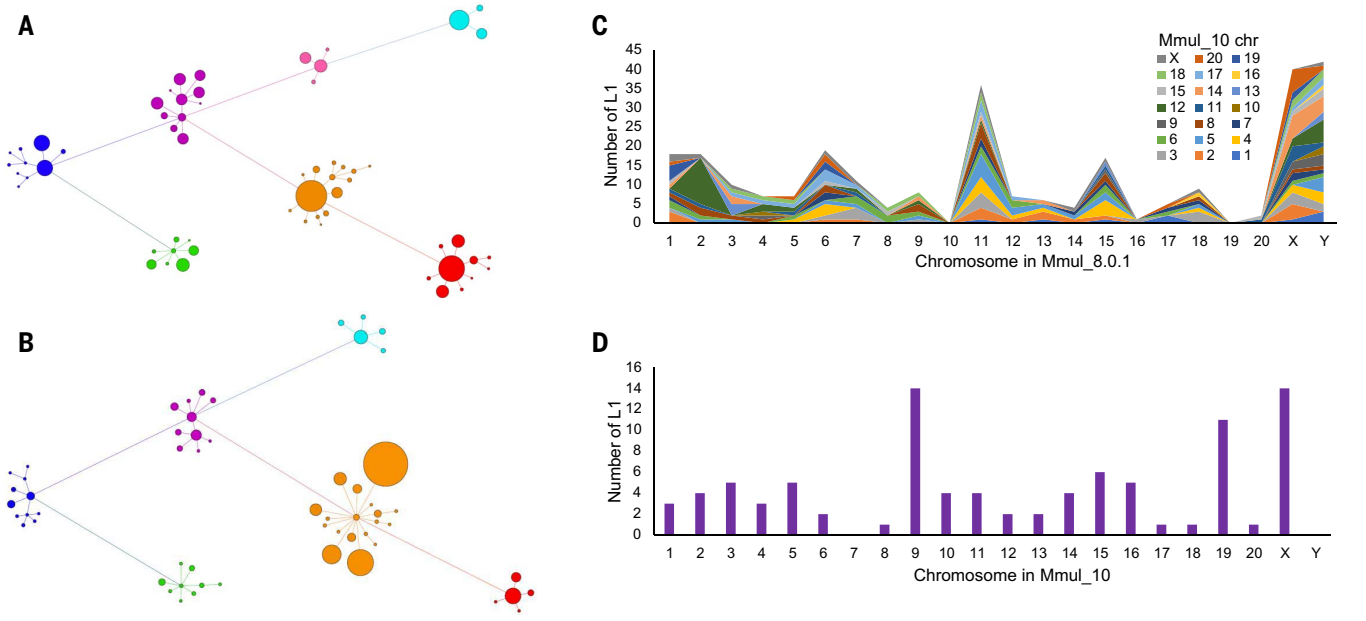


Fig. 4. Full-length LINE1 analyses. A LINE1 subfamily network analysis comparing an earlier macaque assembly, Mmul_8.0.1 (61 subfamilies) (**A**) with the new assembly, Mmul_10 (58 subfamilies) (**B**) (*18*). Related subfamilies are connected by lines and clustered by color: L1RS37 is shown in purple; L1PA7/8, blue; L1PA6, green; L1RS36, pink; L1RS2, red; L1RS10/16/21, orange; and L1RS25, teal. The size of each node corresponds to the relative number of LINE1 elements. There is an increase in annotated younger elements (orange), although the number of subfamilies has decreased the L1RS36 cluster as a result of reassignment based on a higher-quality assembly. (**C**) Plot depicting the number of full-length L1 elements (*y*-axis) that have been assigned to a new chromosome in Mmul_10 (key) compared with Mmul_8.0.1 (*x*-axis). (**D**) A similar analysis depicting the number of full-length LINE1 elements previously unplaced (n = 92) but now assigned to a chromosomal location in Mmul_10.

just 67.3 million SNVs (*30*). Thus, the research rhesus macaques were more than twice as diverse per individual as humans, with the average macaque carrying 9.7 million SNVs.

A principal component analysis (PCA) of the SNV genotypes readily discriminated between the Chinese and Indian rhesus macaques (Fig. 6A). Thirty-one individuals that were initially identified as being of Indian origin showed some degree of Chinese rhesus macaque admixture, although the extent of admixture varied considerably (Fig. 6A). The free-ranging Cayo Santiago rhesus macaque population (Caribbean Primate Research Center) showed a gradient of variation with respect to other Indian rhesus macaques (PC2), which is likely a consequence of a genetic bottleneck since its initial founding in 1938 on the Puerto Rican Island of Cayo Santiago (*31*). Consistent with this observation, the Cayo macaque population showed lower heterozygosity and larger runs of homozygosity compared with other National Primate Research Center (NPRC) populations (figs. S17 to S19). A preliminary analysis showed that linkage disequilibrium decays more rapidly among unrelated Indian rhesus macaques than in a subset of the human African population, but at greater physical distance (>50 kbp) the macaques retain higher linkage disequilibrium (fig. S20).

We repeated the PCA excluding both the Cayo and Chinese macaque populations (Fig. 6B). In this analysis, macaque genetic variation from most primate research centers was indistinguishable, with the exception of Oregon National Primate Center (ONPRC), Yerkes Primate Research Center (YPRC), and California National Primate Research Centers (CNPRC), for which subsets of individuals appear to be genetically distinct (Fig. 6B). A population structure analysis showed that the CNPRC macaques had a somewhat greater admixture with Chinese macaques than other populations on the basis of the genomes analyzed here (figs. S21 and S22). In addition, within some of the research populations (e.g., YNPRC, Cayo, etc.), genetically distinct subgroups of animals were identified with some consistent substructure.

Analysis of variants for functional changes

Macaques are important models of human genetic disease (*3*, *4*), so we investigated both common and rare variants from our samples that affect the sequences of protein-coding genes. In total, we identified 85.7 million SNVs, of which 3.2 million were multiallelic and 21.3 million were singletons, as well as 10.5 million indels (Table 2). For protein-coding sequences, we found 790,377 SNVs and 33,823 indels (Table 2). On the basis of Ensembl Variant Effect Predictor (VEP) annotations, we identified 408,496 missense and 20,400 likely gene-disruptive (LGD) mutations (table S18 and Table 2). Of all variant classes, the LGD mutations occurred at the lowest frequency (0.001 to 0.01), consistent with a more deleterious effect on phenotype (Fig. 6C). An assessment of homozygous-damaging mutations shows considerable overlap among the NPRC colonies for more common variants, whereas a smaller subset was specific to each (fig. S23). We generated a summary distribution of all missense variant counts per gene and normalized by the gene length as defined by the number of protein-coding bases in the gene, excluding introns and UTRs (fig. S24).

To illustrate the biological potential of macaque genetic diversity, we identified naturally occurring macaque mutations in orthologs of human genes implicated in autism and developmental delay (*32*, *33*). In humans, de novo deleterious mutations in these genes are thought to be dominant and have a large effect, but mouse models often do not recapitulate the complexity of neurobehavioral features of human disease (*32*). We considered all missense and LGD mutations such as frameshift, stop, and splice-site mutations that would disrupt the protein-coding sequence. We further classified the genes on the basis of their intolerance to gene-disruptive mutation (pLI or the probability of being loss-of-function intolerant) (Fig. 6D and figs. S25 and S26). The pLI score has been widely adopted and is based on the number of observed versus expected protein-truncating

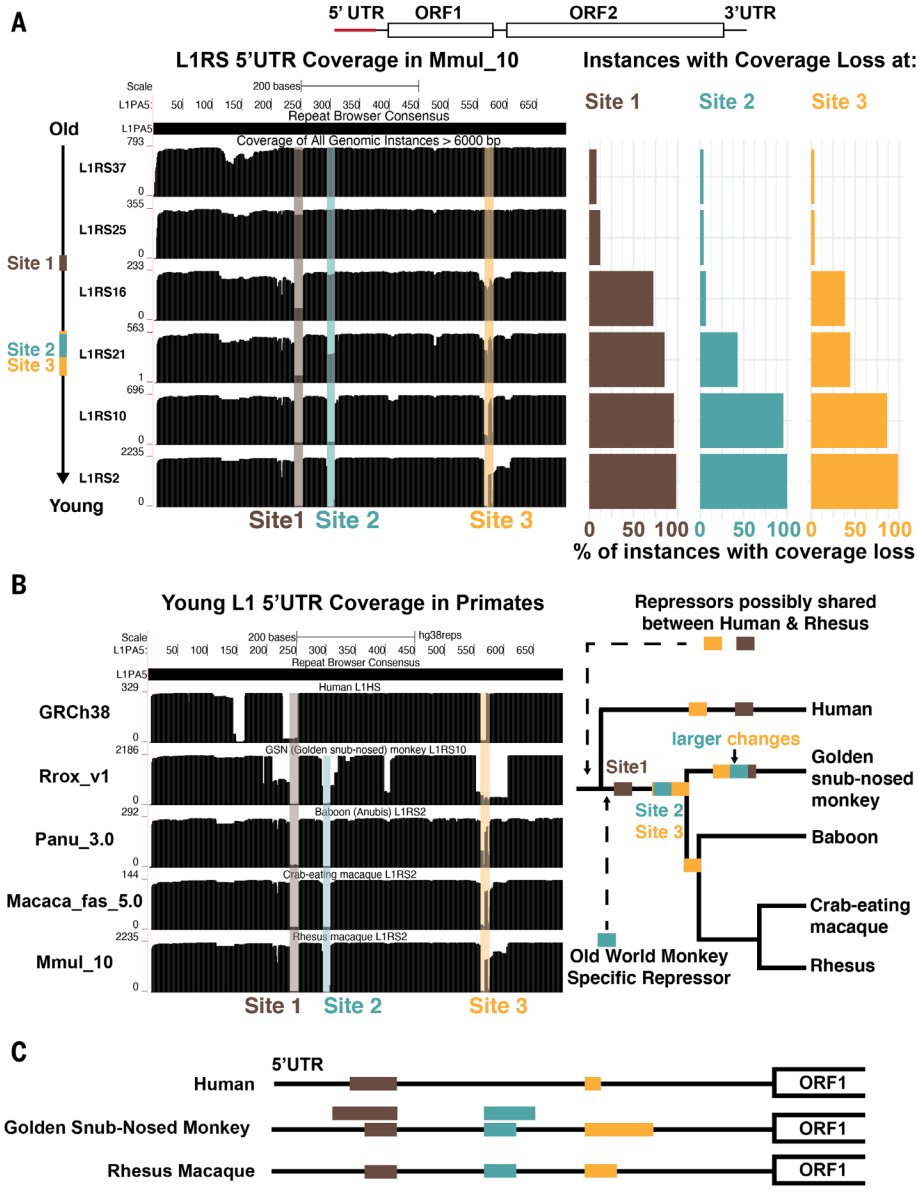


Fig. 5. Evolution of L1RS elements. (**A**) (Left) All full-length L1RS elements (>6000 nt, top schematic) were grouped by families and mapped to a consensus version of L1PA5 (the ancestral LINE1 element from which they derive) with the first 700 nt (red) of the 5′ UTR analyzed further. Site 1 (brown) experiences a coverage drop that is found in most L1RS16 and younger families. Coverage drops at site 2 (blue) and site 3 (yellow) occur in the L1RS21 family at nearly the same time. (Right) Percentage of individual instances that do not map to the L1PA5 consensuses for each L1RS family. Coverage drops are not found in old L1RS elements but found in nearly all young elements, suggesting a fitness advantage for the changes at each site. (**B**) (Left) All full-length elements (>6000 nt) of the youngest L1RS families in four OWM genomes [L1RS10 in Rrox_v1/rhiRox1 (golden snub-nosed monkey) and L1RS2 in Panu_3.0/papAnu4 (baboon), Macaca_fascicularis_5.0/macFas5 (crab-eating macaque), and Mmul_10/rheMac10] were aligned to the L1PA5 consensus to generate coverage plots. The youngest human L1 (L1HS) was also aligned to L1PA5 as an outgroup. Drops in coverage (sites 1, 2, and 3) were seen in OWMs, although golden snub-nosed monkeys (Rrox_v1/rhiRox1) displayed distinct patterns from other OWMs, suggesting convergent but distinct changes in the 5′ UTR, possibly to escape repressive elements. (Right) Evolutionary model for shared and convergent changes in L1RS elements. Site 1 changes are shared among all OWMs, whereas site 2 and 3 changes experience similar but not exact changes in Rrox_v1/rhiRox1 compared with other OWMs. Coverage drops at sites 1 and 3 are also observed in humans, whereas site 2 changes are OWM specific. (**C**) Schematic of sites 1, 2, and 3 (brown, blue, and yellow, respectively) changes on the L1 5′ UTR in representative lineages: humans, golden snub-nosed monkeys, and rhesus macaques. Rhesus macaques and golden snub-nosed monkeys have identical coverage drops at sites 1 and 2 that arose in the OWM common ancestor; golden snub-nosed monkeys also experience larger changes (larger bars) spanning these sites that most likely occurred after the *Colobinae* divergence because they are not observed in rhesus macaques. Humans experience a specific coverage drop at site 1 larger than those of rhesus macaques but smaller than those of the large golden snub-nosed monkey. All three species experience specific changes resulting in differing length elements at site 3.

variants in a population. The closer the pLI score is to 1, the more intolerant to variation the gene is predicted to be. Revisiting gene annotation (based on the Iso-Seq resource) shows that although some of these correspond to changes in gene annotation between species, other changes represent viable candidates to establish new models of neurodevelopmental disease (e.g., severe missense mutations in *ARID1B*). Neurodevelopmental delay genes are significantly depleted for missense variants compared with all genes (Wilcoxon rank sum test; P = 8.883e-41; fig. S25A). Unexpectedly, we identified nine genes with candidate deleterious mutations in macaques that were intolerant to mutation in humans and in which de novo mutations associated with neurodevelopmental disorders (NDDs) (table S19). Homozygous LGD variants segregating in rhesus macaque research centers offer even more opportunities for exploring their biological relevance among NHPs (table S20).

Finally, structural variation differences also present an opportunity to develop new rhesus macaque disease models and enhance our understanding of fixed and polymorphic changes between species. For example, we identified 301,000 insertions and 241,000 deletions shared between the Indian- and Chinese-origin macaques (fig. S27 and tables S21 and S22). Of these 542,000 insertion or deletion events, only a small fraction (1.68%) are predicted to affect genes (table S22). Among the 87,227 structural variants specific to either subspecies, we predict that 1614 may affect genes, but validation in addition to genotyping and genome assembly of more individuals (especially wild caught) will be required to establish fixed differences between the Chinese and Indian macaque genomes.

Discussion

The rhesus macaque is arguably the most important NHP for biomedical research and is a key species in the study of primate evolution. The resources that we developed and present here will substantially advance both areas by providing new biological insights and an

improved framework for future gene-based disease research. In this new assembly, we have corrected indel errors and misassemblies, properly represented inverted sequences, and flagged several remaining orientation issues. The various orthogonal technologies used in the data production process, coupled with detailed manual curation, led to substantially improved gene annotation, eliminated 99.7% of gaps, and reduced misorientations compared with the previous Indian-origin macaque assembly (*2*).

This new reference genome identifies previously unknown genes and genes with an intron-exon structure that differs compared with that of humans, such as the 3.1-kbp Alu-mediated deletion that removed two protein-encoding exons from tropoelastin (*ELN*) in the human and ape orthologs (Fig. 2). Similar to finishing efforts for the mouse and human assemblies (*34*, *35*), most of the new macaque and OWM genes occur among duplicated gene families (Fig. 3). Indeed, our estimate of recent SDs has nearly doubled for this assembly (to ~130 Mbp) and is beginning to approximate that of humans (5 to 6%) (*36*, *37*). Unlike humans, however, in whom a large fraction of the SDs are interspersed along chromosomal arms, most macaque duplications appear to be either clustered or pericentromeric, bracketing the centromeres of chromosomes. Similar to SDs, repeat content, especially for full-length elements, has facilitated comparative analyses identifying recurrent deletion events in the 5′ UTR of L1 elements in human and multiple OWM lineages as a potential adaptive response helping to evade KRAB-ZNF repression. It is tempting to speculate that expansion of OWM (including rhesus macaque) ZNF genes such as *ZNF669* are part of an ongoing arms race to suppress new rhesus L1 subfamilies (*38*). Although the assembly that we generated is superior by most metrics, we recognize that centromeres, acrocentric regions, and some of the largest SDs remain unresolved or unassigned.

Table 2. Summary of macaque genetic variation.

Type	Total variants	Singletons	Multiallelic variants
All SNVs*	85,721,160	21,270,272	3,160,393
All indels	10,501,197	2,956,760	1,875,826
Protein-coding SNVs	790,377	222,766	23,335
Protein-coding indels	33,823	13,227	3663
LGD SNVs†	20,400	6570	1026
LGD indels‡	26,774	10,330	3051

*SNV and indel protein-coding classifications based on VEP. †LGD SNVs. ‡LGD indels defined as VEP consequences splice_acceptor_variant, splice_donor_variant, stop_gained, or stop_lost, start_lost.

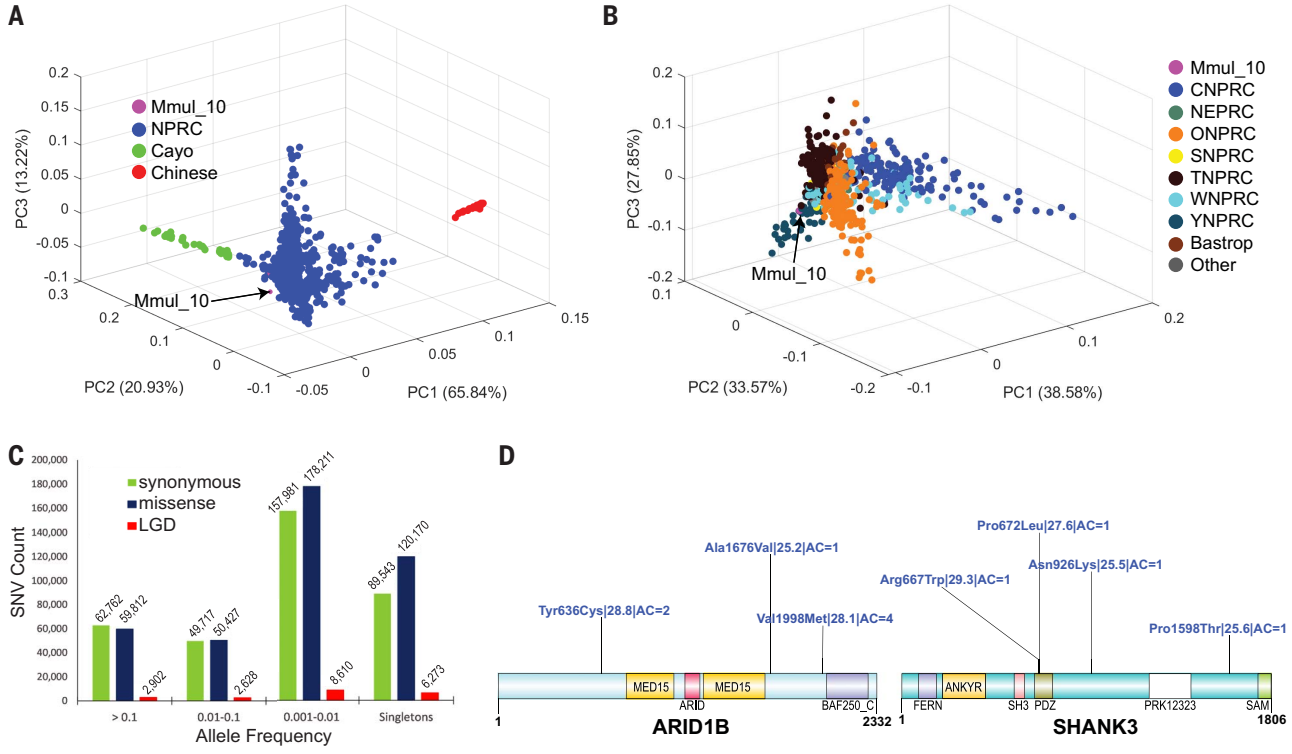


Fig. 6. Rhesus macaque population structure and developing macaque models of disease. (**A**) 3D PCA based on SNVs filtered for missing call rates >0.05 or major allele frequency (MAF) <0.1 from sequencing 853 macaque genomes shows clear separation of Chinese (PC1) genomes (red) and a gradient for Cayo macaques (green) with respect to other Indian macaques (PC2). (**B**) PCA excluding Chinese and Cayo populations comparing 771 macaques from different NPRCs. The Cattell-Nelson-Gorsuch (CNG) screen test retained the top three principal components in both PCAs and the percent variance explained calculations are based on those three components. (**C**) Allele frequency distribution of LGD mutations, including splice acceptor, splice donor, stop-gain, stop-loss, and start-loss variants (red) and missense (blue) variants compared with synonymous changes (green). (**D**) Genes implicated in human NDDs showing naturally occurring putatively damaging variants in macaque orthologs. A schematic of damaging missense (blue) variants (CADD ≥25) for NDD genes: *MBD5*, *ARID1B*, and *SHANK3*. For each variant, we indicate the amino acid change (CADD score) allele count. All potentially deleterious mutations are low frequency.

Using this new macaque reference, we have established an extensive SNV resource that will facilitate future genetic analyses of biomedical research colonies. Our catalog of thousands of naturally occurring common missense variants and identification of other rare macaque mutations may help in the discovery of new models of disease, such as those implicated in autism and human NDDs (*32*, *33*). These naturally occurring mutations provide an opportunity to develop noninvasive models of human disease without the expense of CRISPR engineering of embryos (*9*). These models may be particularly powerful in relation to phenotypes that are not readily reproduced in nonprimate knockout models (*12*) and for evaluating the effect of genetic variation on the efficacy of treatments before human trials.

Methods summary

A single female rhesus macaque of Indian origin (AG07107) was sequenced, assembled, and manually curated with a physical map, proximity ligation, and Strand-seq sequence data. Full-length cDNA was prepared from various tissue sources and used to annotate the genome assembly with three independent gene annotation pipelines: NCBI, Ensembl, and CAT. A whole-genome analysis comparison coupled with read-depth assessment was used to estimate the proportion of SDs. We used interphase and metaphase FISH on a female rhesus macaque lymphoblast cell line to validate SD order and orientations. All repeat elements were quantified using RepeatMasker for comparisons among rhesus macaque assemblies. To estimate rhesus macaque genetic diversity, we sequenced 853 animals across nine U.S. rhesus macaque research colonies using standard Illumina sequencing instruments. We used a compendium of best practices to call sequence variants for SNVs, insertions, deletions, and structural variants. To classify the potential impact of SNVs, we aligned all to the human genome and focused our analysis on those that cause loss of function by stop-gain, start-lost, splice-donor, or splice-acceptor base changes and missense variants that alter amino acid coding of the protein. A subset of these SNVs was compared with human genes known to harbor rare de novo deleterious variants that have been implicated in human NDDs.

REFERENCES AND NOTES

1. J. A. Bailey, E. E. Eichler, Primate segmental duplications: Crucibles of evolution, diversity and disease. *Nat. Rev. Genet.* **7**, 552–564 (2006). doi: 10.1038/nrg1895; pmid: 16770338
2. R. A. Gibbs *et al*., Evolutionary and biomedical insights from the rhesus macaque genome. *Science* **316**, 222–234 (2007). doi: 10.1126/science.1139247; pmid: 17431167
3. C. Xue *et al*., The population genomics of rhesus macaques (*Macaca mulatta*) based on whole-genome sequences. *Genome Res.* **26**, 1651–1662 (2016). doi: 10.1101/gr.204255.116; pmid: 27934697
4. B. N. Bimber *et al*., Whole genome sequencing predicts novel human disease models in rhesus macaques. *Genomics* **109**, 214–220 (2017). doi: 10.1016/j.ygeno.2017.04.001; pmid: 28438488
5. Z. N. Kronenberg *et al*., High-resolution comparative analysis of great ape genomes. *Science* **360**, eaar634 (2018). doi: 10.1126/science.aar6343; pmid: 29880660
6. J. Rogers, R. A. Gibbs, Comparative primate genomics: Emerging patterns of genome content and dynamics. *Nat. Rev. Genet.* **15**, 347–359 (2014). doi: 10.1038/nrg3707; pmid: 24709753
7. K. K. A. Van Rompay, Tackling HIV and AIDS: Contributions by non-human primate models. *Lab Anim. (NY)* **46**, 259–270 (2017). doi: 10.1038/laban.1279; pmid: 28530684
8. H. Feldmann, F. Feldmann, A. Marzi, Ebola: Lessons on vaccine development. *Annu. Rev. Microbiol.* **72**, 423–446 (2018). doi: 10.1146/annurev-micro-090817-062414; pmid: 30200851
9. Y. Zhou *et al*., Atypical behaviour and connectivity in SHANK3-mutant macaques. *Nature* **570**, 326–331 (2019). doi: 10.1038/s41586-019-1278-0; pmid: 31189958
10. R. Daza-Vamenta, G. Glusman, L. Rowen, B. Guthrie, D. E. Geraghty, Genetic divergence of the rhesus macaque major histocompatibility complex. *Genome Res.* **14**, 1501–1515 (2004). doi: 10.1101/gr.2134504; pmid: 15289473
11. B. K. Dray *et al*., Mismatch repair gene mutations lead to lynch syndrome colorectal cancer in rhesus macaques. *Genes Cancer* **9**, 142–152 (2018). doi: 10.18632/genesandcancer.170; pmid: 30108684
12. B. Y. Liao, J. Zhang, Null mutations in human and mouse orthologs frequently result in different phenotypes. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 6987–6992 (2008). doi: 10.1073/pnas.0800387105; pmid: 18458337
13. J. Seok *et al*., Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 3507–3512 (2013). doi: 10.1073/pnas.1222878110; pmid: 23401516
14. S. M. Peterson *et al*., Bardet-Biedl Syndrome in rhesus macaques: A nonhuman primate model of retinitis pigmentosa. *Exp. Eye Res.* **189**, 107825 (2019). doi: 10.1016/j.exer.2019.107825; pmid: 31589838
15. A. Moshiri *et al*., A nonhuman primate model of inherited retinal disease. *J. Clin. Invest.* **129**, 863–874 (2019). doi: 10.1172/JCI123980; pmid: 30667376
16. J. Rogers *et al*., CRHR1 genotypes, neural circuits and the diathesis for anxiety and depression. *Mol. Psychiatry* **18**, 700–707 (2013). doi: 10.1038/mp.2012.152; pmid: 23147386
17. D. H. Abbott, J. Rogers, D. A. Dumesic, J. E. Levine, Naturally occurring and experimentally induced rhesus macaque models for polycystic ovary syndrome: Translational gateways to clinical application. *Med. Sci. (Basel)* **7**, 107 (2019). doi: 10.3390/medsci7120107; pmid: 31783681
18. See the supplementary materials.
19. J. F. Hughes *et al*., Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. *Nature* **483**, 82–86 (2012). doi: 10.1038/nature10843; pmid: 22367542
20. Y. He *et al*., Long-read assembly of the Chinese rhesus macaque genome and identification of ape-specific structural variants. *Nat. Commun.* **10**, 4233 (2019). doi: 10.1038/s41467-019-12174-w; pmid: 31530812
21. J. G. Sambrook *et al*., Single haplotype analysis demonstrates rapid evolution of the killer immunoglobulin-like receptor (KIR) loci in primates. *Genome Res.* **15**, 25–35 (2005). doi: 10.1101/gr.2381205; pmid: 15632087
22. C. R. Catacchio *et al*., Inversion variants in human and primate genomes. *Genome Res.* **28**, 910–920 (2018). doi: 10.1101/gr.234831.118; pmid: 29776991
23. M. Seppey, M. Manni, E. M. Zdobnov, BUSCO: Assessing genome assembly and annotation completeness. *Methods Mol. Biol.* **1962**, 227–245 (2019). doi: 10.1007/978-1-4939-9173-0_14; pmid: 31020564
24. I. T. Fiddes *et al*., Comparative Annotation Toolkit (CAT)-simultaneous clade and personal genome annotation. *Genome Res.* **28**, 1029–1038 (2018). doi: 10.1101/gr.233460.117; pmid: 29884752
25. R. D. Finn *et al*., The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Res.* **44** (D1), D279–D285 (2016). doi: 10.1093/nar/gkv1344; pmid: 26673716
26. M. R. Vollger *et al*., Long-read sequence and assembly of segmental duplications. *Nat. Methods* **16**, 88–94 (2019). doi: 10.1038/s41592-018-0236-3; pmid: 30559433
27. M. J. Chaisson, S. Mukherjee, S. Kannan, E. E. Eichler, Resolving multicopy duplications *de novo* using polyploid phasing. *Res Comput Mol Biol* **10229**, 117–133 (2017). doi: 10.1007/978-3-319-56970-3_8; pmid: 28808695
28. J. D. Fernandes *et al*., The UCSC repeat browser allows discovery and visualization of evolutionary conflict across repeat families. *Mob. DNA* **11**, 13 (2020). doi: 10.1186/s13100-020-00208-w; pmid: 32266012
29. M. Imbeault, P. Y. Helleboid, D. Trono, KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* **543**, 550–554 (2017). doi: 10.1038/nature21683; pmid: 28273063
30. A. Bergström *et al*., Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**, eaay5012 (2020). doi: 10.1126/science.aay5012; pmid: 32193295
31. A. Widdig *et al*., Low incidence of inbreeding in a long-lived primate population isolated for 75 years. *Behav. Ecol. Sociobiol.* **71**, 18 (2017). doi: 10.1007/s00265-016-2236-6; pmid: 28018027
32. B. P. Coe *et al*., Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. *Nat. Genet.* **51**, 106–116 (2019). doi: 10.1038/s41588-018-0288-4; pmid: 30559488
33. F. K. Satterstrom *et al*., Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell* **180**, 568–584.e23 (2020). doi: 10.1016/j.cell.2019.12.036; pmid: 31981491
34. D. M. Church *et al*., Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLOS Biol.* **7**, e1000112 (2009). doi: 10.1371/journal.pbio.1000112; pmid: 19468303
35. International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004). doi: 10.1038/nature03001; pmid: 15496913
36. J. A. Bailey, A. M. Yavor, H. F. Massa, B. J. Trask, E. E. Eichler, Segmental duplications: Organization and impact within the current human genome project assembly. *Genome Res.* **11**, 1005–1017 (2001). doi: 10.1101/gr.GR-1871R; pmid: 11381028
37. J. A. Bailey *et al*., Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002). doi: 10.1126/science.1072047; pmid: 12169732
38. J. D. Fernandes *et al*., KRAB zinc finger proteins coordinate across evolutionary time scales to battle retroelements. bioRxiv 429563 [Preprint]. 27 September 2018. doi: 10.1101/429563

ACKNOWLEDGMENTS

We thank the staff at the NPRCs involved in the preparation of biomaterials used for WGS; S. Peterson (OHSU) for preparation of mRNA used in the generation of Iso-Seq datasets; the UC Davis Primate Research Center for the rhesus macaque fetal brain material and specific tissues from developmentally staged samples used for Iso-Seq data production; the Baylor College of Medicine Human Genome Sequencing Center production teams, especially D. Muzny and H. Doddapaneni, for their work on macaque diversity sequencing; the Baylor Human Genome Sequencing Center, the Texas Advanced Computing Center, and Rice University for allowing us to use their computational resources for mapping and variant calling; and A. Pollen (UCSF) and S. Peterson (OHSU) for preparation of mRNA used in the generation of Iso-Seq datasets. The cell line MMU1 used for Strand-seq and FISH analyses was obtained from the Department of Comparative Genetics and Refinement, Biomedical Primate Research Centre (BPRC), Netherlands (courtesy of R. E. Bontrop). **Funding:** This work was supported in part by the National Institutes of Health (NIH grants HG002385 and 1U24HG009081 to E.E.E.; U01HG010961, U41HG010972, R01HG010485, 2U41HG007234, U01HL137183, 5U54HG007990, and 5T32HG008345-04 to B.P.; R01MH081884, R01MH046729, and P50MH100031 to N.H.K.; and R01GM59290 to M.A.B.) by a subagreement from the European Molecular Biology Laboratory with funds provided by agreement no. 2U41HG007234-05 from National Institute of Health, NHGRI. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of NIH, NHGRI, or European Molecular Biology Laboratory. Rhesus WGS from primate research centers was supported by NIH grant no. 5R24OD011173 to J.R.; grant no. R24OD021324 to B.F. that supported sequencing of the ONPRC cohort; grant no. P51-OD011106 that supported sequencing of the WNPRC cohort as well as support for D.O., J.R., N.K., D.A., and J.K.; CNPRC base grant no. P51OD011107 and BBA grant no. OD010962 to J.P.C.; YNPRC Pilot Research Project Program base grant no. P51OD011132 to Z.A.K-B.; TNPRC grants P51OD011104, U42OD024282, U42OD010568, and 1R01HG010329 to S.R.S.; grant no.

R01HG002939 to M.A.B. provided support for repeat analyses. E.E.E. is an investigator of the Howard Hughes Medical Institute.
Author contributions: L.W.H., D.G., C.T., T.A.G-L., M.K., M.E., N.W.M., S.S., R.E.G., and W.C.W. completed de novo assembly and its curation. P.C.D., D.G., M.R.V., A.P.L., P.A.A., and E.E.E. performed SD and genome quality assessment. A.D.S., F.A.M.M., C.R.C., F.A., D.P., and J.O.K. performed Strand-seq single-cell libraries construction and data analysis. L.M. and M.V. performed FISH and SD analyses. E.E.E., J.G.U., and K.M.M. generated Iso-Seq data and performed analyses. J.M.S., J.A.W., and M.A.B. conducted the repeat element analyses. J.F. and S.R.S. reconstructed the evolution of L1RS elements. D.H.O. and R.W.W. evaluated MHC and KIR regions. R.A.H., M.R., and J.R. generated the rhesus macaque sequence variation data. R.A.H., M.R., J.R., E.E.E., Y.M., and S.C.M. analyzed rhesus macaque sequence variation. M.H., I.T.F., J.A., M.D., E.D., and B.P. annotated and analyzed rhesus macaque gene evolution. B.F., H.M.K., L.P., N.H.K., D.R., D.H.A., S.B.G., M.M.S., Z.A.K-B., J.W.K., S.M.T., J.R., E.L.K., J.P.C., J.H.P.S., M.P., J.L., J.A.R., and S.A.C., provided rhesus macaque biomaterials for whole-genome and transcriptome sequencing. W.C.W., J.R., and E.E.E. supervised the project and wrote the manuscript. **Competing interests:** The authors declare no competing financial interests. **Data availability:** All data are available in the main text or the supplementary materials. The Mmul_10 genome assembly is available in the NCBI assembly archive under accession number GCF_003339765.1.

Supplemental Materials
science.sciencemag.org/content/370/6523/eabc6617/suppl/DC1
Materials and Methods
Supplemental Text
Figs. S1 to S30
References (*39*–*80*)
Tables S1 to S29
Database S1
MDAR Reproducibility Checklist
View/request a protocol for this paper from *Bio-protocol*.

7 May 2020; accepted 29 October 2020
10.1126/science.abc6617

Science

Sequence diversity analyses of an improved rhesus macaque genome enhance its biomedical utility

Wesley C. Warren, R. Alan Harris, Marina Haukness, Ian T. Fiddes, Shwetha C. Murali, Jason Fernandes, Philip C. Dishuck, Jessica M. Storer, Muthuswamy Raveendran, LaDeana W. Hillier, David Porubsky, Yafei Mao, David Gordon, Mitchell R. Vollger, Alexandra P. Lewis, Katherine M. Munson, Elizabeth DeVogelaere, Joel Armstrong, Mark Diekhans, Jerilyn A. Walker, Chad Tomlinson, Tina A. Graves-Lindsay, Milinn Kremitzki, Sofie R. Salama, Peter A. Audano, Merly Escalona, Nicholas W. Maurer, Francesca Antonacci, Ludovica Mercuri, Flavia A. M. Maggiolini, Claudia Rita Catacchio, Jason G. Underwood, David H. O'Connor, Ashley D. Sanders, Jan O. Korbel, Betsy Ferguson, H. Michael Kubisch, Louis Picker, Ned H. Kalin, Douglas Rosene, Jon Levine, David H. Abbott, Stanton B. Gray, Mar M. Sanchez, Zsofia A. Kovacs-Balint, Joseph W. Kemnitz, Sara M. Thomasy, Jeffrey A. Roberts, Erin L. Kinnally, John P. Capitanio, J. H. Pate Skene, Michael Platt, Shelley A. Cole, Richard E. Green, Mario Ventura, Roger W. Wiseman, Benedict Paten, Mark A. Batzer, Jeffrey Rogers and Evan E. Eichler

Science **370** (6523), eabc6617.
DOI: 10.1126/science.abc6617

A high-quality rhesus macaque genome

Genome technology has improved substantially since the first full organismal genomes were generated. Applying new technology, Warren *et al.* refined the genome of the rhesus macaque, a model nonhuman primate. Long-read technology and other recent advances in sequencing technology were applied to generate a genome with far fewer gaps and helped to refine the locations and numbers of repetitive elements. Furthermore, the authors performed resequencing among populations to identify the genetic variability of the rhesus macaque. Thus, a previously incomplete and inaccurate set of sequence information is now fully resolved, improving gene mapping for biomedical and comparative genetic studies.

Science, this issue p. eabc6617

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science* is a registered trademark of AAAS.