




# Genomic Copy Number Variation Study of Nine *Macaca* Species Provides New Insights into Their Genetic Divergence, Adaptation, and Biomedical Application

Jing Li<sup>1,2</sup>, Zhenxin Fan <sup>1,3</sup>, Feichen Shen<sup>4</sup>, Amanda L. Pendleton<sup>4</sup>, Yang Song<sup>1</sup>, Jinchuan Xing <sup>5</sup>, Bisong Yue<sup>1</sup>, Jeffrey M. Kidd<sup>4,\*</sup>, and Jing Li <sup>1,3,\*</sup>

<sup>1</sup>Key Laboratory of Bio-Resources and Eco-Environment (Ministry of Education), College of Life Sciences, Sichuan University, Chengdu, Sichuan, China

<sup>2</sup>Institute of Animal Genetics and Breeding, College of Animal Science and Technology, Sichuan Agricultural University, Chengdu, Sichuan, China

<sup>3</sup>Sichuan Key Laboratory of Conservation Biology on Endangered Wildlife, College of Life Sciences, Sichuan University, Chengdu, Sichuan, China

<sup>4</sup>Department of Human Genetics, Medical School, University of Michigan

<sup>5</sup>Department of Genetics and the Human Genetics Institute of New Jersey, Rutgers, The State University of New Jersey, Piscataway

\*Corresponding authors: E-mails: ljtf@126.com; jmkidd@umich.edu.

Accepted: 19 September 2020

## Abstract

Copy number variation (CNV) can promote phenotypic diversification and adaptive evolution. However, the genomic architecture of CNVs among *Macaca* species remains scarcely reported, and the roles of CNVs in adaptation and evolution of macaques have not been well addressed. Here, we identified and characterized 1,479 genome-wide hetero-specific CNVs across nine *Macaca* species with bioinformatic methods, along with 26 CNV-dense regions and dozens of lineage-specific CNVs. The genes intersecting CNVs were overrepresented in nutritional metabolism, xenobiotics/drug metabolism, and immune-related pathways. Population-level transcriptome data showed that nearly 46% of CNV genes were differentially expressed across populations and also mainly consisted of metabolic and immune-related genes, which implied the role of CNVs in environmental adaptation of *Macaca*. Several CNVs overlapping drug metabolism genes were verified with genomic quantitative polymerase chain reaction, suggesting that these macaques may have different drug metabolism features. The CNV-dense regions, including 15 first reported here, represent unstable genomic segments in macaques where biological innovation may evolve. Twelve gains and 40 losses specific to the Barbary macaque contain genes with essential roles in energy homeostasis and immunity defense, inferring the genetic basis of its unique distribution in North Africa. Our study not only elucidated the genetic diversity across *Macaca* species from the perspective of structural variation but also provided suggestive evidence for the role of CNVs in adaptation and genome evolution. Additionally, our findings provide new insights into the application of diverse macaques to drug study.

**Key words:** macaque, structural variation, genetic diversity, adaptive evolution, drug metabolism.

## Significance

Copy number variation (CNV) plays an important role in the adaptation and evolution of mammals. However, CNV in *Macaca* species has not been thoroughly studied so far, which hinders the understanding of genome features, adaptive evolution, and biomedical application of macaques. Here, we identified and characterized genome-wide interspecific CNVs among nine *Macaca* species, demonstrating that these macaques mainly diverge from one another in nutritional metabolism, drug metabolism, and immune-related pathways from the perspective of structural variation. Our findings provide not only suggestive evidence for the role of CNVs in the adaptation and evolution of *Macaca* species but also new insights into the biomedical application of these nonhuman primates.

© The Author(s) 2020. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

## Introduction

Copy number variations (CNVs) represent a major form of structural genetic variation. CNVs are segments of DNA that have variable copy numbers (CNs) within a species or a lineage (Feuk et al. 2006; Freeman et al. 2006). Widely spread over the genome, CNVs typically range from 1 kb to 5 Mb in the human genome (Redon et al. 2006). A growing amount of evidence suggests that CNVs are associated with local adaptation via phenotypic trait variations (Kondrashov 2012). CNVs may impact variance in expression levels of specific genes via gene dosage effects, positional effects, gene splits, gene fusions, and unmasking of recessive alleles (Lupski and Stankiewicz 2005; Henrichsen et al. 2009; Hou et al. 2012), leading to different phenotypes and susceptibility to diseases (Gökçümen and Lee 2009; Iskow et al. 2012). Additionally, they can modify genome architecture and lead to additional structural variants that promote genome evolution and speciation (Perry et al. 2008; Conrad et al. 2010).

CNV features have been studied in many animal taxa, including birds (chicken [Wang et al. 2010; Jia et al. 2013] and zebra finch [Völker et al. 2010]) and mammals (mice and rats [Cahan et al. 2009; Charchar et al. 2010], dogs and wolves [Nicholas et al. 2011; Alvarez and Akey 2012], pigs [Chen et al. 2012; Paudel et al. 2015], sheep [Yang et al. 2018], cattle [Keel et al. 2016], yaks [Goshu et al. 2019], horse [Jun et al. 2014], and great apes [Marques-Bonet et al. 2009; Gazave et al. 2011; Oetjens et al. 2016]). In particular, CNV studies in humans and other great apes have uncovered the role of CNVs in various phenotypic traits, diseases, and primate evolution (Stranger et al. 2007; Itsara et al. 2009; Zhang et al. 2009; Ventura et al. 2011; Almal and Padh 2012; Sudmant et al. 2013). Compared with the great ape lineage, the study of CNV in *Macaca* is rather limited. Prior work has mainly focused either on conspecific variations of the rhesus (Lee et al. 2008; Gokcumen et al. 2011) and cynomolgus macaques (Gschwind et al. 2017) or on particular gene families in macaques (Degenhardt et al. 2009; Uno et al. 2010; Ottolini et al. 2014), leaving genomic CNVs across *Macaca* species scarcely reported and the biological significance of CNVs in ecological and evolutionary processes of macaques unsolved.

The genus *Macaca* (Primates: Cercopithecidae) is a group of diverse Catarrhini that contains 23 species (Solari and Baker 2006; Li et al. 2015) which have diverged over a short evolutionary timespan. Macaques are the most widely distributed nonhuman primates (NHPs), occupying various habitat types in Asia along with Northern Africa and Southern Europe (Solari and Baker 2006). They have diverged genetically, morphologically, and behaviorally to adapt to a wide range of environmental conditions (Thierry 2007; Roos and Zinner 2015). Therefore, this genus has experienced both rapid speciation and adaptive radiation (Jiang et al. 2016). To date, the general phylogenetic relationships among macaques have

been well addressed, and the seven-species-group phylogeny (Zinner et al. 2013; Roos et al. 2014) based on molecular evidence is widely accepted.

Macaques are also important animal models in a wide range of biomedical research including drug development studies. They display very high similarity to humans in development, immunology, pathology, and behavior (Haus et al. 2014). Besides rhesus (*Macaca mulatta*) and cynomolgus (*Macaca fascicularis*) macaques, southern pig-tailed (*Macaca nemestrina*), Barbary (*Macaca sylvanus*), Tibetan (*Macaca thibetana*), and Japanese (*Macaca fuscata*) macaques are increasingly used as NHP models in biomedical research (Hatzioannou et al. 2009; Pouladi et al. 2013; Zhang et al. 2017). The genetic backgrounds of different macaques may strongly affect the results of biomedical studies. For example, because of a species-specific insertion of a retrotransposon in the *TRIM5* gene, the southern pig-tailed macaque can be infected by HIV-1, whereas rhesus and cynomolgus macaques cannot (Brennan et al. 2008; Newman et al. 2008). However, detailed exploration of the impact of using different *Macaca* species on research outcomes is limited. Despite many genome-wide single-nucleotide variant (SNV) studies elucidating the evolutionary history, population genetics, or intraspecific genetic diversity of macaques (Fang et al. 2011; Yan et al. 2011; Higashino et al. 2012; Fan, Zhao, et al. 2014; Zhong et al. 2016; Fan et al. 2018; Liu et al. 2018), the genetic divergence among *Macaca* species has not been thoroughly surveyed, especially from the perspective of structural variation and CNV.

Recently, easier accessibility of next-generation sequencing (NGS) data of *Macaca* species combined with well-developed CNV detection approaches allow for comprehensive characterization of genome-wide CNVs in macaques. NGS-based methods have become a popular strategy for CNV detection. Compared with array-based approaches like single-nucleotide polymorphism (SNP) and comparative genomic hybridization arrays (Carter 2007; Li and Olivier 2013), the NGS-based approach has higher resolution, more accurate estimation of CNs and breakpoints, and better capability to identify novel CNVs (Meyerson et al. 2010; Alkan et al. 2011). Various software tools, such as Breakdancer (Fan, Abbott, et al. 2014), and pipelines such as fastCN, which calculates CNs with a read depth (RD)-based approach (Pendleton et al. 2018), have been developed to detect CNVs using NGS data.

In this study, we used a RD-based method to identify interspecific CNVs and shared duplications genome-wide across nine species of macaques and further analyzed the distribution patterns and potential functions of these CNVs with bioinformatic approaches. In addition to the two flagship species in *Macaca*, rhesus and cynomolgus macaques, our sample set also includes the only non-Asian species, Barbary macaque (*M. sylvanus*) and several less studied macaque species. Altogether, this systematic study presents the first comprehensive map of genome-wide interspecific CNV in *Macaca*.

**Table 1**  
Information on Genome Data in This Study

Scientific Names	Sample Identifier(s)	GenBank Accession(s)	Sequencing Platform(s)	# Reads	Genome Depth	Total Usable Base Pairs	Sex	Sample Origin(s)	Source(s)
<i>M. mulatta mulatta</i>	Mmul_8	—	Illumina	20,100,000	5.1×	—	Female	Washington National Primate Research Center	Zimin et al. (2014)
<i>M. mulatta lasiota</i>	CR	SRA023856	Illumina	3,299,851,568	45.65×	2,264,143,011	Female	Yunnan, China	Yan et al. (2011)
<i>M. fascicularis</i>	CE	SRA023855	Illumina	3,299,851,568	43.96×	2,245,482,535	Female	Vietnam	Yan et al. (2011)
<i>M. arctoides</i>	SM	SRX1470574	Illumina	1,001,034,260	34.55×	2,280,352,231	Female	Southwestern China	Fan et al. (2018)
<i>M. thibetana</i>	TM	SRP032525	Illumina	1,275,012,390	36.92×	2,281,638,762	Female	Sichuan, China	Fan, Zhao, et al. (2014)
<i>M. nemestrina</i>	PM	SRX1022644	Illumina	770,413,198	25.59×	2,246,079,419	Female	Washington National Primate Research Center	Baylor College of Medicine
<i>M. fuscata</i>	JM	SRR11921216	Illumina	2,258,829,541	83.85×	2,271,704,290	Female	Kyoto Primate Research Center	Fan ZX, Zhou AB, Xing JC, Hey J, Osada N, Melnick DJ, Yue BS, Li J. (unpublished data)
<i>M. cyclops</i>	TwM	SRR11921217	Illumina	2,279,695,913	24.66×	2,279,695,913	Female	Kyoto Primate Research Center	Fan ZX, Zhou AB, Xing JC, Hey J, Osada N, Melnick DJ, Yue BS, Li J. (unpublished data)
<i>M. sylvanus</i>	BM	SRR11921218, SRR11927939–SRR11927943	Illumina	2,226,490,341	45.91×	2,226,490,341	Female	Columbia University	In this study
<i>M. silenus</i>	LM	SRR11921219, SRR11927944–SRR11927948	Illumina	2,241,953,780	46.49×	2,241,953,780	Female	Columbia University	In this study

We aimed at not only obtaining a better understanding of the genetic diversity and biomedical application of these *Macaca* species but also providing new insights into the roles of CNVs in genetic diversity, environmental adaptation, and the evolution of these animals.

## Materials and Methods

### Genome Data

Whole resequenced genomes of four macaques were produced in-house, including the Japanese (*Macaca fuscata*, JM), Taiwanese (*Macaca cyclopis*, TwM), Barbary (*M. sylvanus*, BM), and lion-tailed (*Macaca silenus*, LM) macaques, as shown in [table 1](#). These sequence data were combined with public genome data of the Chinese rhesus (*Macaca mulatta lasiotea*, CR) (Yan et al. 2011), cynomolgus (*M. fascicularis*, CE) (Yan et al. 2011), Tibetan (*M. thibetana*, TM) (Fan, Zhao, et al. 2014), stump-tailed (*Macaca arctoides*, SM) (Fan et al. 2018), and southern pig-tailed (*Macaca nemestrina*, PM) macaques. A genome data set of nine individuals was analyzed, involving nine species and representing six out of seven-species groups in *Macaca* ([table 1](#)). The sequence data for each sample varied from 25× (TwM) to nearly 84× (JM) coverage, allowing sufficient power to detect CNVs.

### CN Estimation Using Short-Read Data

As *Macaca* species are phylogenetically close, it is reasonable to estimate interspecific CNVs by mapping the short-read sequences from various macaques to the reference genome of the Indian rhesus macaque (*Macaca mulatta mulatta*; Mmul\_8) (Zimin et al. 2014). Prior to mapping, we employed FastQC (v0.11.8) (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc>) to perform quality control checks on raw sequences, then used Trimmomatic (v0.36) (Bolger et al. 2014) to filter and trim the reads. The cleaned reads were aligned to the reference genome using BWA mem (Li and Durbin 2009).

The program fastCN was designed to efficiently estimate genome CN from short-read data utilizing RD information (<https://github.com/KiddLab/fastCN>) (Pendleton et al. 2018). Two steps were implemented to estimate CNs with the fastCN pipeline. First, GC correction was performed using custom-defined control regions to remove the GC bias introduced by polymerase chain reaction (PCR) during library preparation and sequencing. Due to the lack of known control regions across macaques, we implemented an iterative process to retrieve effective control regions, which correspond to a CN of two in these diploid genomes. The initial control regions were defined as autosomal genomic regions excluding segments masked by RepeatMasker and Tandem Repeat Finder (Benson 1999), overrepresented 50mers, assembly gaps, and an additional 36 bp flanking each masked segment. RD was converted to estimated CN based on a set of control

regions. This calculation was performed in windows that contain an equal number (1,000 bp, 1 kb) of unmasked, nongap positions. As a result, although each window contains the same number of interrogated positions, the actual size of the windows along the genome is variable and individual windows may span assembly gaps. Using these initial data, we then defined a revised set of control regions which appeared to have a fixed CN. Specifically, we optimized the controls for each sample as segments where RD fell into the full width at half maximum of the RD distribution of all 1-kb windows ([supplementary fig. S1, Supplementary Material](#) online), which were likely to be regions with a CN of two. We then repeated the GC normalization and CN estimation procedure using the revised control regions for all the samples.

Second, GC-corrected per-bp depths were converted to mean depths in windows containing 1 kb of unmasked sequence. As described above, the windows differ in genomic length, but each window contains 1,000 unmasked, nongap positions. We estimated genome-wide CNs based on window depths by using a correction factor calculated from the average RD of the control regions. The calculation function is as follows:

$$CF = RD_{ctl}/2$$

$$CN = RD/CF'$$

where CF stands for the correction factor, RD represents the read depth of specific genomic window, and  $RD_{ctl}$  is the mean read depth of the control regions. Unplaced contigs were merged as a single “chrUn” in data processing to decrease the CPU time.

### CNV and Shared Duplication Identification

Due to the absence of multiple individuals of the same species, we used the maximum copy number difference ( $CND_m$ ) to define interspecific CNV, which is the difference between the maximum CN ( $CN_{max}$ ) and the minimum CN ( $CN_{min}$ ) of the samples:

$$CND_m = CN_{max} - CN_{min}$$

Theoretically, duplications are regions with CN of at least three copies, deletions are segments with CN of one (heterozygous deletion) or zero (homozygous deletion), and CNVs are bins where the CN difference is equal to or greater than one copy for any two samples, which means  $CND_m \geq 1$  among the nine samples. To correct for noise in the CNs, we checked the modal value of  $CND$  ( $\sim 0.6$ ) which should approximate zero, and thus set the  $CND_m$  threshold for CNV as 1.6 ( $1 + 0.6$ ) copies ([supplementary fig. S2, Supplementary Material](#) online). Duplications were defined as windows with 2.7 ( $3 - 0.6/2$ ) or more copies and deletions were defined as bins with 1.3 ( $1 + 0.6/2$ ) or less copies among the nine samples. Duplications shared by all macaques are

considered to be fixed in this genus, which are of research importance. To reduce false positives, we only kept CNVs or shared duplications no shorter than 3 kb. After merging consecutive 1-kb windows and calculating the mean CNs for merged windows, we filtered out these failing the thresholds or that were shorter than three windows (~3 kb). CNVs on chrUn were excluded from subsequent analyses. We employed the UCSC genome browser to examine the CN patterns using custom track files.

### CNV-Dense Region Detection

We surveyed CNV density across the genome using a sliding window of 10 Mb with a custom python script. CNV density was defined as the CNV count in each window. According to the count distribution, bins with ten or more CNVs were empirically considered to be a CNV-dense region. LiftOver (<https://genome.ucsc.edu/cgi-bin/hgLiftOver>) was used to convert the coordinates to match the human reference genome hg19. We investigated if the CNV-dense regions overlapped with CNV hotspots shared by human, chimpanzee, and rhesus macaque identified in Gokcumen et al. (2011).

### Gene-Based Annotation and Functional Enrichment Analyses

To delineate the functional impact of CNVs, we performed gene annotation and enrichment analyses. Gene-based annotation was implemented with “bedtools window” (Quinlan and Hall 2010). Because CNVs can regulate the expression and function of adjacent genes, we set the intersecting window size between CNVs and genes as 5 kb. Annotations of gene models in rhesus macaque genome were obtained from Ensembl (<http://ftp.ensembl.org/pub/release-92>). The gene ontology (GO) and KEGG enrichment analyses were performed with standalone KOBAS 3.0 (Xie et al. 2011) on genes that intersected CNVs or shared duplications with 5-kb window allowance. The background gene set contained all Ensembl genes in the rhesus macaque. We chose “Fisher’s exact test” and “Benjamini and Hochberg (1995)” as the statistical and FDR correction methods, respectively. Small terms with five or less genes were dropped from our analyses. Functional enrichment with g:Profiler (Reimand et al. 2007) was also conducted on genes intersecting with CNV-dense regions. Hierarchical filtering was set as “best per parent group”. The size of the functional categories ranged from 5 to 2,000 genes. Benjamini–Hochberg FDR was employed to calculate significant threshold.

### Permutation Test

To inspect if there was any positional bias of the CNVs or shared duplications, we calculated empirical significance by performing 1,000 genome-wide permutations. We shuffled both the locations of the CNVs or shared duplications and

locations of genes with “bedtools shuffle” (Quinlan and Hall 2010) to examine the following three factors: 1) the number of CNVs or shared duplications intersecting with genes, 2) the number of genes intersecting with CNVs or shared duplications, and 3) the lengths of intersecting genes. The shuffling tested the following hypotheses, 1) if the genes were overlapped with CNVs or shared duplications incidentally, 2) if large genes were more likely to emerge in the enriched pathways, and 3) if the enriched CNV intersecting genes tended to emerge together. *P* values were defined as the possibility of the observation in the distribution of the permutation data.

### Lineage-Specific CNV Screening

Lineage-specific CNVs were screened to investigate the evolutionary features or adaptive characteristics of macaques. We utilized Picard (v1.98; <http://broadinstitute.github.io/picard/>) and GATK (v3.2) (DePristo et al. 2011) to identify the genome-wide interspecific SNVs of the nine species. After hard filtration suggested by the GATK website (QualByDepth [QD] < 2.0; QUAL < 30.0; FisherStrand [FS] > 60.0; RMSMappingQuality [MQ] < 40.0; StrandOddsRatio [SOR] > 4.0; MQRankSum < -12.5; ReadPosRankSum < -8.0), the SNVs were thinned to 500,000 sites with PLINK (v1.07) (Purcell et al. 2007) to estimate a phylogenetic tree using the Neighbor-Joining (NJ) method by SNPhylo (Lee et al. 2014). Bootstrap replicates ( $n = 1,000$ ) were employed to assess branch support. Clade-specific duplication was defined as  $CN \geq 2.7$  for a clade and around two copies ( $1.7 \leq CN \leq 2.3$ ) for others. Correspondingly, a lineage-specific CNV deletion was called when  $CN \leq 1.3$  for a lineage but  $1.7 \leq CN \leq 2.3$  for others.

### Genomic Quantitative Polymerase Chain Reaction Validation

To validate the CNVs in drug metabolism genes including *CYP2C76*, *UGT2B33*, *UGT1A1*, *GSTM5*, and *GSTM1*, real-time quantitative polymerase chain reaction (qPCR) was conducted on genomic DNA. Primers were designed with Primer3Plus (<https://primer3plus.com/cgi-bin/dev/primer3-plus.cgi>) for the CNVs and a diploid internal control, part of *RPP30* with no CN alteration among macaques (supplementary fig. S3, Supplementary Material online). Primer information is available in supplementary table S1, Supplementary Material online. The fidelity of the primers was checked in silico.

Blood samples of Chinese rhesus (CR-AB, CR-OB), cynomolgus (CE-3), Tibetan (TM-4), stump-tailed (SM-2), and Japanese (JM-5) macaques were collected from Chengdu Zoo and Hengshu Bio-Technology Company. Due to lack of Indian rhesus macaque, Chinese rhesus macaque (CR-AB) was used as calibrator for all qPCR experiments. As southern pig-tailed macaque is not distributed in China, we used northern pig-tailed macaque (*Macaca leonine*, PM-6, feces,

Chengdu Zoo) instead, which is the phylogenetically closest species to the southern pig-tailed macaque in China, and both belong to *silenus* group (Zinner et al. 2013; Roos et al. 2014). Genomic DNA was extracted using TIANGEN Genomic DNA Kit (TIANGEN, Peking, China). All samples were obtained in accordance with Chinese regulations for the implementation of protection of terrestrial wild animals (State Council Decree [1992] No.13), and all laboratory work was approved by the Guidelines for Care and Use of Laboratory Animals and the Ethics Committee of Sichuan University (Chengdu, China). Throughout the procedure, care was taken to ensure animal welfare for all monkeys.

By referring to previous studies (Jung et al. 2013; Wang et al. 2018), relative quantification with  $\Delta\Delta\text{CT}$  method was employed, and CN of each target was calculated as  $2 \times 2^{-\Delta\Delta\text{CT}}$ . Genomic qPCR was performed using the real-time qPCR system as recommended by the manufacturer's instruction. In brief, a 10  $\mu\text{l}$  of reaction mixture contained 10 ng of genomic DNA, 1 $\times$  Taq SYBRGreen qPCR Mix (Innovagene, Changsha, China), and 5 pmol of each primer. Thermal cycling conditions consisted of one cycle of 3 min at 94 °C followed by 40 cycles of 20 s at 94 °C, 40 s at 60 °C, and 20 s at 72 °C. All qPCR experiments were triplicate.

### Differential Expression Analysis of CNV Genes Based on Population Transcriptome Data

To explore if CNVs affected gene expression and further biological functions, we investigated the expression levels of CNVs intersecting genes (CNVGs) whose CNs were distinct ( $\text{CND} \geq 1.6$ ) between the Chinese rhesus macaque (CR) and the Tibetan macaque (HT) using the expression matrices in Yan (2019), which studied the blood transcriptomes of 28 Chinese rhesus macaques and 24 Tibetan macaques. For the CNVGs with detectable expression, differentially expressed genes (DEGs) were identified using threshold of  $P < 0.05$  and  $q < 0.05$ . To test if the percentage of differentially expressed CNVGs was significant, we randomly resampled the same number of CNVs between the Chinese rhesus and Tibetan macaques from all interspecific CNVs 1,000 times and compared the observation with the percentages of differentially expressed CNVGs in the resampled data. We also reviewed the main functions of these DEGs to infer the biological impact of these CNVs.

## Results

### Duplicated Regions across the Nine Species

We estimated genomic CN from NGS data of nine *Macaca* species based on the rhesus macaque reference genome (Mmul\_8) using the fastCN algorithm (Pendleton et al. 2018). To improve the performance of CN estimation, we created an iterative process to identify control regions for GC normalization of the observed sequencing depth as

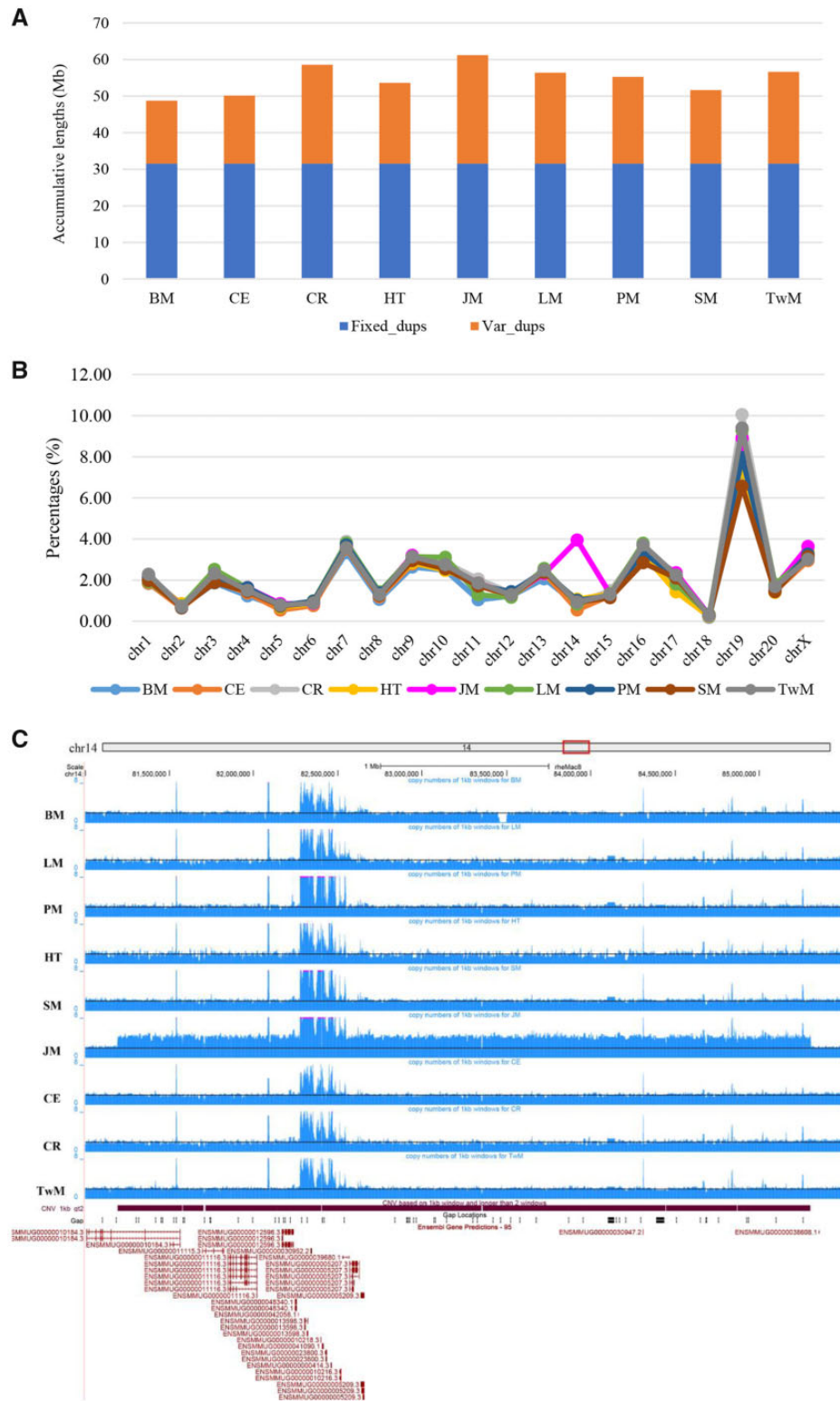
described in Materials and Methods. The improved controls lead to an effective GC correction for all species (supplementary fig. S4, Supplementary Material online).

Using the estimated CNs, gains and losses that spanned three or more 1-kb windows were detected on a per sample basis for the nine species. There were 2,183 (*M. fascicularis*, CE) to 2,686 (*M. thibetana*, HT) gains identified per sample. The cumulative lengths of duplications across all chromosomes for each sample are shown in figure 1A. The genomic distribution of CN gains shows a highly uneven pattern. For example, chromosome 19 harbors the largest proportion of duplications, varying from 6.42% for the Tibetan macaque to 10.36% for the Chinese rhesus macaque. The lowest percentage of duplications was found on chromosome 18, fluctuating from 0.04% for the Chinese rhesus, Japanese, and crab-eating macaques to 0.11% for the lion-tailed macaque. Additionally, we identified an excess of duplications on chromosome 14 in the Japanese macaque. This excess was driven by a 4-Mb event that was absent in all other samples (fig. 1B and 1C). This large duplication overlapped with 17 genes, including five genes in the *TRIM* family (*TRIM49*, putative *TRIM49B*, *TRIM51*, putative *TRIM64B*, and *TRIM77*) and three genes associated with the nervous system (*GRM5*, *FOLH1*, and *NAALAD2*), and also harbored a small shared duplication intersecting a homolog to human *TRIM64*.

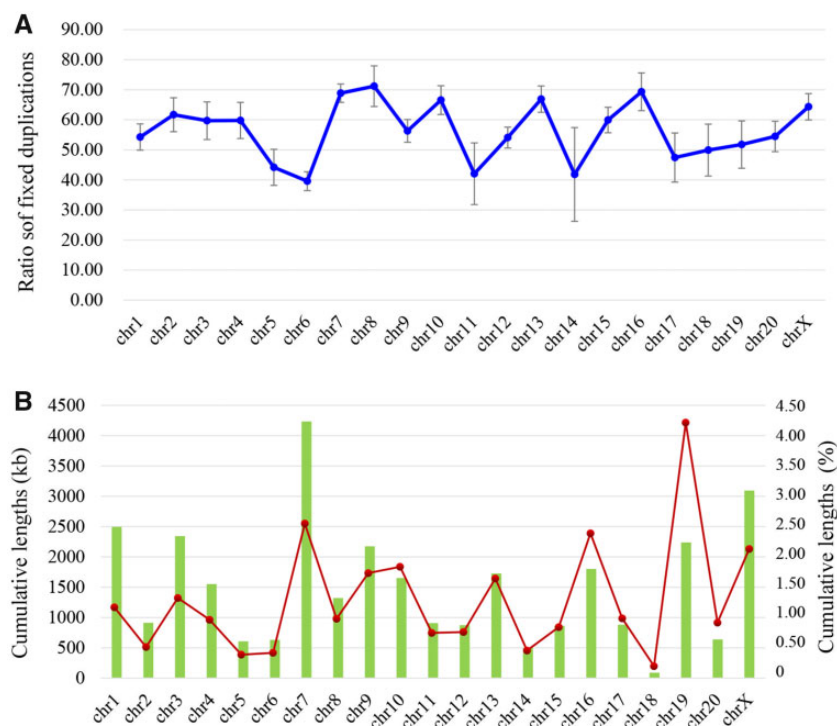
### Shared Duplications

We searched for regions that are duplicated in all analyzed macaques, regardless of the estimated CN. Although each species was represented by one individual, the shared duplicated regions likely represent genomic regions expanded across *Macaca*, indicating the common genomic features of this group. In total, 1,560 duplications were shared by all assessed macaques (fig 2 and supplementary fig. S5, Supplementary Material online). Shared duplications on chromosome 7 were longer than these on other chromosomes. Chromosome 19 displayed the highest abundance, whereas chromosome 18 held the lowest percentage of shared duplications (fig. 2B). Due to their across-genus distribution, these CN gains likely resulted from duplications that occurred in the last common ancestor of these species.

We observed that 1,166 shared duplications and their 5-kb flanking sequences intersected 1,656 Ensembl genes annotated in the rhesus macaque genome. To assess the genomic location patterns of these duplicates, we conducted permutation tests by randomly shuffling the coordinates ( $n = 1,000$  shuffles) of the duplications or the Ensembl genes and found that shared duplicates are significantly enriched in genic regions ( $P = 0.001$ ) and that more genes intersected with shared duplications than expected by chance ( $P = 0.001$ ). We also noted that the lengths of these genes were shorter than expected ( $P = 0.001$ ) (supplementary figs. S6 and S7, Supplementary Material online). This was opposite to the



**Fig. 1**—Genomic patterns of duplications across the nine macaque species. (A) Cumulative lengths of duplications detected in three or more 1-kb windows across all chromosomes for each sample. (B) Proportion of cumulative duplication lengths for each chromosome. (C) Copy number (blue histograms) across the large duplication on chromosome 14 of Japanese macaque is visualized in UCSC genome browser relative to other macaques (species symbols on left) and in the context of Ensembl gene models (red). Copy number was estimated in windows containing 1 kb of nongap, nonmasked sequence. As a result, the genomic span of individual windows is variable and may include positions annotated as assembly gaps.



**Fig. 2**—Cumulative lengths of the shared duplications detected in three or more 1-kb windows on each chromosome in the nine *Macaca* species. (A) Average ratios of the cumulative lengths of shared duplications to the cumulative length of duplications per chromosome across the nine samples. Error bars represent the standard deviations of the ratios among the nine species. (B) The cumulative lengths (green bars) of the shared duplications per chromosome and the percentage (red line) of shared duplications on each chromosome in terms of length.

expectation that CNVs would overlap with long genes due to random chance, suggesting that the enrichment may be driven by gene clustering, given that clustered genes are usually short in length.

Functional enrichment analysis of the 1,656 intersecting genes using KOBAS 3.0 ( $P \leq 0.01$ , [supplementary table S2, Supplementary Material](#) online) showed that the majority of the enriched pathways were metabolic pathways, including metabolism of steroid hormone, xenobiotics, retinol, pentose and glucuronate, starch and sucrose, aldarate, and porphyrin. The enriched categories also contained many ribosome-related terms, coincident with the finding of a CNV study of horse (Doan et al. 2012) that 11.9% of ribosomal RNA genes in horse were affected by CNVs, ranking the first among all kinds of genes. To explore if these functional terms were enriched incidentally due to colocalization of genes belonging a shared biological pathway, permutation tests were conducted again by carrying out enrichment analyses using shuffled duplication data sets or shuffled gene sets. Permutations showed that the observed count of enriched pathways was >90% of the pathway counts found in the permutations and that the count of significantly enriched GO categories was larger than that found in >95% of the permutations ([supplementary figs. S8 and S9, Supplementary Material](#) online). This suggests the observed results reflect true enrichment signals rather than spurious hits due to random chance.

### Genome-Wide Interspecific CNVs

A total of 1,479 regions were identified as CNVs variable across the nine *Macaca* species with lengths of three or more consecutive 1-kb windows (fig. 3 and [supplementary data](#) set S1 and [fig. S10, Supplementary Material](#) online), including 1,106 gains and 451 losses. Of these, 78 were complex CNVs containing both gains and losses. Gains were ~2.5-fold more common than losses and displayed comparatively larger average sizes. These CNVs totaled 39.7 Mb, or 1.41% of the genome. The individual lengths of CNVs ranged from 3,001 to 1,086,528 bp with a mean and median of 26,857 and 12,007 bp, respectively. The majority of identified CNVs were relatively small, as ~70% were between 3 and 20 kb. The count comparison of CNVs with different lengths ( $\geq 1$ -kb,  $\geq 3 \times 1$ -kb, and  $\geq 10 \times 1$ -kb windows) is shown in [supplementary table S3, Supplementary Material](#) online.

### Functional Annotation of CNVs

Using 5-kb intersecting windows, 854 out of 1,479 CNVs overlapped with 1,420 Ensembl genes. Specifically, 727 (65.73%) duplications intersected with 1,287 genes, and 164 (36.36%) deletions encompassed 302 genes ([supplementary fig. S11, Supplementary Material](#) online). In total, 52.81% of CNVs are directly located in genic regions, which is concordant with the study of Lee et al. (2008) where 55%





**Fig. 3**—Genomic distribution of all interspecific CNVs (detected in three or more 1-kb windows) across the nine *Macaca* species. The blue rectangles represent duplication CNVs and the red rectangles represent deletion CNVs.

(68/124) CNVs identified in rhesus macaque were genic. Genes overlapping with CNVs and their 5-kb flanking regions (CNVGs) can be separated into three categories: duplicated genes overlapped by CNV gains, deleted genes intersecting with CNV losses, and mixtures where genes colocalize with loci harboring both gains and losses.

Permutation tests highlighted the role of gene clustering in this enrichment. Although the observed CNVs did not overlap with a gene more often than expected by chance ( $P = 0.142$ , [supplementary fig. S12A](#), [Supplementary Material](#) online), the total number of genes that intersected with a CNVs was greater than expected ( $P = 0.001$ , [supplementary fig. S12B](#), [Supplementary Material](#) online), and the length of the intersecting genes was significantly shorter than expected ( $P = 0.001$ , [supplementary fig. S12C](#), [Supplementary Material](#) online). The observed intersection counts were substantially different from those found when the positions of genes were randomly shuffled ([supplementary fig. S13](#), [Supplementary Material](#) online). Thus, gene clustering, that is, the nonuniform placement of genes along the genome, partially accounts for the increased number of genes that overlap with CNVs. Furthermore, we observed that CNVGs were significantly shorter than expected ([supplementary figs. S12 and S13](#), [Supplementary Material](#) online). This may reflect a real genome feature or represent a bias due to the

comparatively low quality of the macaque genome assembly or gene model annotation.

We propose two hypotheses for the origin of the “bias” in CNV gene length: 1) the rhesus macaque reference genome is incompletely assembled, and/or 2) the gene models in CNV regions may be inaccurate. To investigate these assumptions, we compared not only the protein coding gene lengths in control regions and CNV regions but also the quality of gene models from rhesus (*Mmul\_8*), chimpanzee (*Pan\_troglodytes-2*), and human (*GRCh38*) reference genomes. The genomic lengths of all protein coding genes in the macaque genome were similar to that in the chimpanzee genome, but shorter than those in the human genome ([supplementary table S4a](#), [Supplementary Material](#) online), suggesting that assembly quality may affect our results. The comparison between control and CNV regions also indicates quality of gene models in CNVs may contribute to the observation ([supplementary table S4b](#), [Supplementary Material](#) online).

We again performed enrichment analyses with KOBAS 3.0 and subsequent permutation tests for CNVs. CNVGs were generally overrepresented ( $P \leq 0.01$ ) in three main categories: nutritional metabolism, xenobiotics/drug metabolism, and immune-related pathways, but some enrichments lacked strong support from the corrected  $P$  value ([table 2](#) and

**Table 2**

Enrichment Outputs of Genes Intersecting the CNVs and Their 5-kb Flanking Sequences Using KOBAS 3.0: (A) Enriched KEGG Pathways and (B) Enriched GO Terms (Only Exhibiting the Highest Category in the Tree for GO Terms Containing Exactly the Same Genes)

Term	ID	Input No.	Background No.	P Value	Corrected P Value	Gene Symbols <sup>a</sup>	CNV ID
(A) Enriched KEGG pathways							
Pentose and glu- curonate interconversions	mcc00040	6	21	1.9E-05	0.034	ALDH3A2, LOC706528, UGT2B33, UGT1A1, UGT2B15, ENSMMUG00000012355	chr16-19414797-19421121, chr3-160661084-160869232, chr5-65506980-65623157, chr5-65628860-65729639, chr12-116232453-116238344, chr5-65466947-65505089
Ascorbate and aldarate metabolism	mcc00053	5	14	4.0E-05	0.037	UGT2B15, ALDH3A2, UGT1A1, UGT2B33, ENSMMUG00000012355	chr5-65628860-65729639, chr16-19414797-19421121, chr12-116232453-116238344, chr5-65506980-65623157, chr5-65466947-65505089
Viral myocarditis	mcc05416	7	42	7.8E-05	0.047	ABL2, MAMU-DRB1, LOC106992470, SGCD, CASP9, MAMU-DRB1, ACTB	chr1-188324769-188338450, chr4-33338105-33352404, chr4-33355720-33386474, chr4-33406118-33412221, chr4-30203434-30413162, chr6-154795623-154806159, chr1-14308663-14373529, chr4-33928581-33933207, chr3-39508480-39512271
Retinol metabolism	mcc00830	7	45	1.1E-04	0.052	UGT2B33, CYP2C76, LOC713738, UGT1A1, UGT2B15, ALDH1A2, ENSMMUG00000012355	chr5-65506980-65623157, chr5-65628860-65729639, chr9-90297250-90344108, chr11-55695239-55754415, chr12-116232453-116238344, chr7-34700918-34706034, chr5-65466947-65505089
Chemical carcinogenesis	mcc05204	7	55	3.5E-04	0.013	GSTM5, UGT2B33, CYP2C76, UGT1A1, UGT2B15, CYP2A23, ENSMMUG00000012355	chr1-110697591-110747857, chr5-65506980-65623157, chr5-65628860-65729639, chr9-90297250-90344108, chr12-116232453-116238344, chr19-36763842-36769753, chr5-65466947-65505089
Antigen processing and presentation	mcc04612	6	44	6.7E-04	0.15	KIR2DL4, MAMU-DRB1, LOC106992470, KLRC3, KLRC1, MAMU-DRB1	chr19-50037741-50073879, chr4-33338105-33352404, chr4-33355720-33386474, chr4-33406118-33412221, chr4-30203434-30413162, chr11-10736952-10744681, chr11-10746926-10788644, chr4-33928581-33933207
Porphyryn and chlorophyll metabolism	mcc00860	5	29	7.4E-04	0.15	UGT2B15, UGT1A1, UGT2B33, ENSMMUG00000012355, FXN	chr5-65628860-65729639, chr12-116232453-116238344, chr5-65506980-65623157, chr5-65466947-65505089, 85260241
Drug metabolism: cytochrome P450	mcc00982	6	48	0.001	0.17	GSTM5, UGT2B33, CYP2C76, UGT1A1, UGT2B15, ENSMMUG00000012355	chr1-110697591-110747857, chr5-65506980-65623157, chr5-65628860-65729639, chr9-90297250-90344108, chr12-116232453-116238344, chr5-65466947-65505089
Type I diabetes mellitus	mcc04940	5	32	0.0011	0.17	MAMU-DRB1, HSPD1, MAMU-DRB1, LOC106992470, LOC693438	chr4-33928581-33933207, chr2-148896350-148910750, chr4-33338105-33352404, chr4-33355720-33386474, chr4-33406118-33412221, chr4-30203434-30413162, chr5-105546464-105549989
Metabolism of xenobiotics by cy- tochrome P450	mcc00980	6	49	0.0011	0.17	GSTM5, UGT2B33, UGT1A1, UGT2B15, CYP2A23, ENSMMUG00000012355	chr1-110697591-110747857, chr5-65506980-65623157, chr5-65628860-65729639, chr12-116232453-116238344, chr19-36763842-36769753, chr5-65466947-65505089
Starch and sucrose metabolism	mcc00500	5	36	0.0018	0.25	UGT2B15, UGT1A1, UGT2B33, ENSMMUG00000012355, AMY2B	chr5-65628860-65729639, chr12-116232453-116238344, chr5-65506980-65623157, chr5-65466947-65505089, 104421047
RNA degradation	mcc03018	6	64	0.0039	0.46	HSPD1, LOC693438, PARN, PABPC1, EXOSC1, PFKF	chr2-148896350-148910750, chr5-105546464-105549989, chr20-14709561-14721562, chr8-99416856-99422405, chr9-92935279-92941800, chr9-2908518-2937562, chr9-2980923-3000708

Drug metabolism: other enzymes	mcc000983	4	32	0.0072	0.46	<i>UGT2B15</i> , <i>UGT1A1</i> , <i>UGT2B33</i> , <i>ENSMIMUG00000012355</i>	chr5-65628860-65729639, chr12-116232453-116238344, chr5-65506980-65623157, chr5-65466947-65505089
(B) Enriched GO terms	GO:0015020	3	5	5.1E-04	0.13	<i>UGT1A1</i> , <i>UGT2B33</i> , <i>UGT2B15</i>	chr12-116232453-116238344, chr5-65506980-65623157, chr5-65628860-65729639
Glucuronosyltransferase activity	GO:0008194	3	14	0.0053	0.46	<i>UGT1A1</i> , <i>UGT2B33</i> , <i>UGT2B15</i>	chr12-116232453-116238344, chr5-65506980-65623157, chr5-65628860-65729639
UDP-glycosyltransferase activity	GO:0071103	3	15	0.0063	0.46	<i>HIMGB3</i> , <i>NCAPD2</i> , <i>H2BFWT</i>	chrX-144618359-144621948, chr11-6773682-6777531, chrX-97761715-97768667
DNA conformation change	GO:0009812	2	5	0.009	0.46	<i>UGT2B33</i> , <i>UGT2B15</i>	chr5-65506980-65623157, chr5-65628860-65729639
Flavonoid metabolic process							

NOTE.—The term description and ID are provided along with the number of genes identified near our CNVs (input) compared with the total Ensembl gene set of the rhesus macaque (background). The raw and corrected *P* values are indicated. Input gene names with enrichment signals ( $P \leq 0.01$ ) can be found in the last second column.

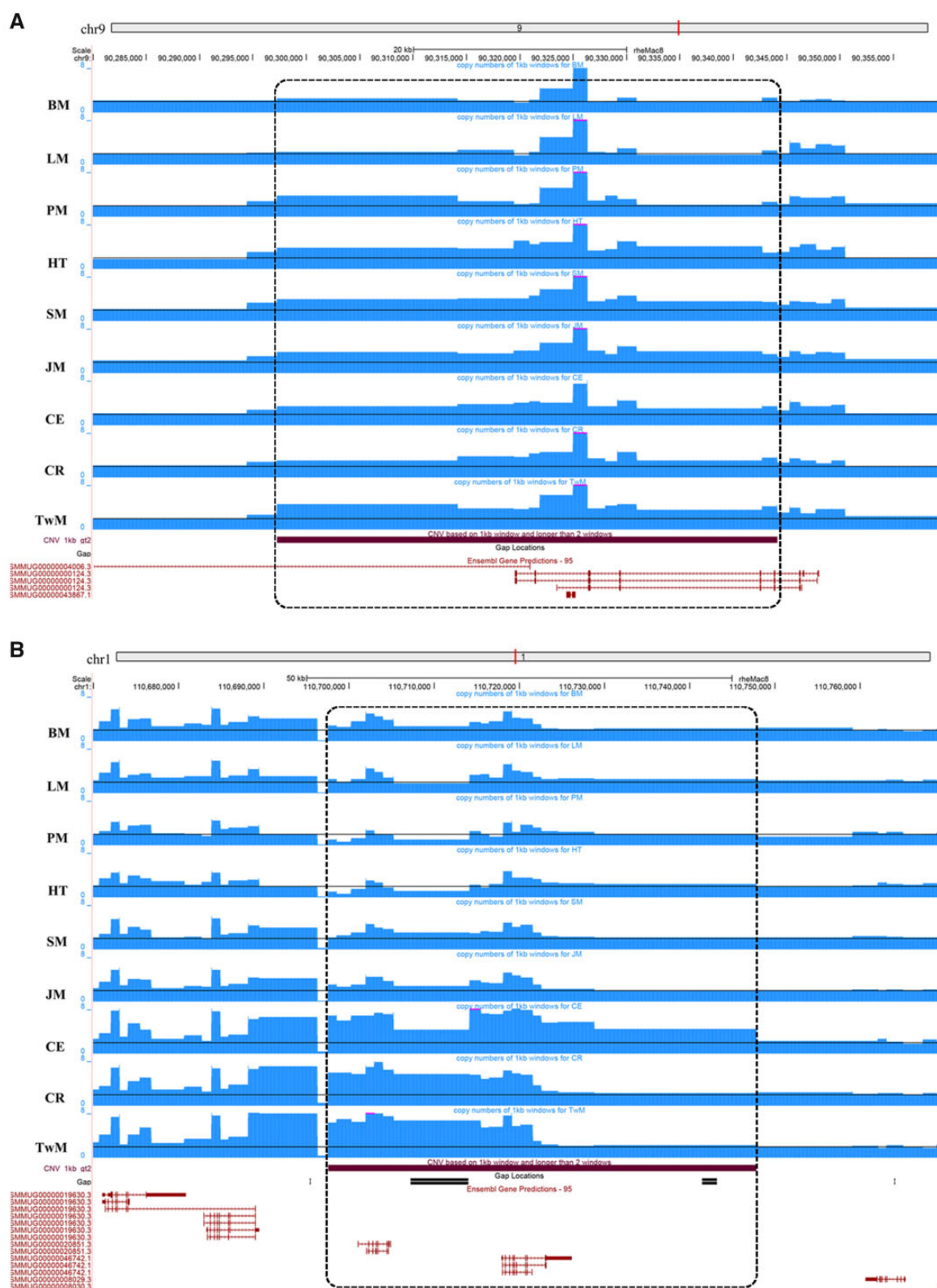
\*The novel genes without gene symbols are indicated with Ensembl gene IDs.

supplementary table S5, Supplementary Material online). Enrichment outputs of duplicated genes were quite similar to that of all CNVGs, mainly because copy gains outnumbered copy losses. Deleted genes were enriched in “olfactory transduction” and disease pathways including “Viral myocarditis” (mcc05416), “Asthma” (mcc05310), and “Graft-versus-host disease” (mcc05332) (supplementary table S6, Supplementary Material online). These pathways were in accordance with the enriched GO terms related to signaling receptor activity. Results from randomized permutations based on locations of both CNVs and genes showed that there were 90% more enriched KEGG pathways and GO terms than expected by chance (supplementary figs. S14 and S15, Supplementary Material online), confirming the functional enrichment in these CNVGs. Additionally, we note that enriched genes (ribosome genes, *HLA* loci, and olfactory genes) tended to be clustered in the genome (Younger et al. 2001; Ishii et al. 2006).

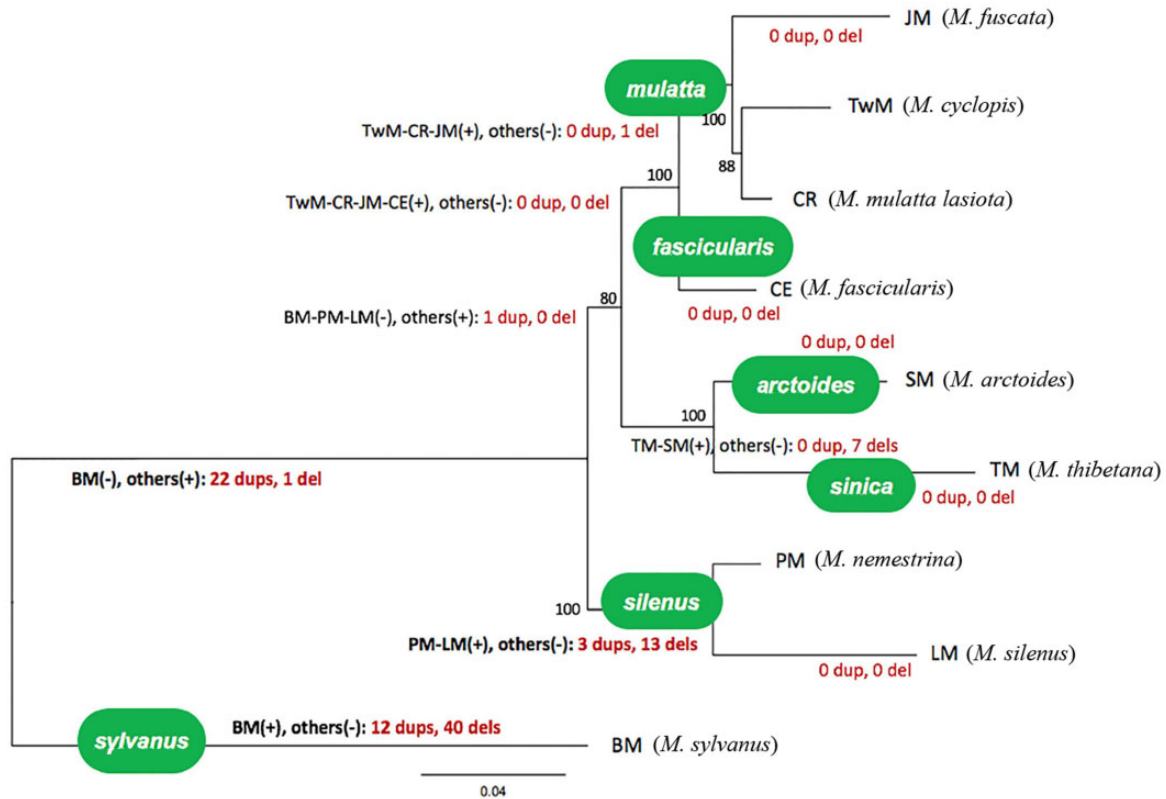
Drug metabolism genes were highly enriched among the CNV intersecting genes, including *CYP2C76*, *UGT2B33*, *UGT2B15*, *UGT1A1*, and *GSTM5*. Notably, the CN patterns of *CYP2C76* fit the expectations for *Macaca* species under the evolutionary phylogeny reconstructed from genomic SNVs with high bootstrap support values (supplementary fig. S16, Supplementary Material online). In detail, the species that diverged early in the phylogenetic tree (BM, LM, and PM) maintained two copies of *CYP2C76*, whereas others had around four copies. We observed distinct CN difference across macaques in this region on the UCSC genome browser, with a short shared duplication (~1 kb, 8–9 copies) embedded in the CNV (fig. 4A). We also uncovered 26 apparent SNVs in the CNV intersecting *CYP2C76* which were heterozygous in all macaques except for BM, LM, and PM, whose genotypes were homozygous at each locus, validating this CNV with SNV genotyping. The CN pattern at *GSTM5* was generally consistent with the phylogenetic topology of this genus as well. The clade of CE, CR, and TwM, representing recently diverged macaques, had 5–6 copies of *GSTM5*, but only 2–3 copies were found in other species (fig. 4B).

### CNV-Dense Regions

We detected 26 CNV-dense regions (containing 479 CNVs), which displayed at least ten CNVs per 10-Mb segment. These CNV-dense regions were distributed across the genome, and chromosome 7 harbored four such regions, ranking the first among all chromosomes. Fifteen of the 26 CNV-dense regions are first reported here, whereas the remaining 11 regions overlap with human, chimpanzee, and rhesus macaque-shared CNV regions identified by Gokcumen et al. (2011), which are very likely to be CNV hotspots in both great apes and Catarrhini. Several important immunity-related genes, such as *HLA*, *HCG9*, *DEFA*, and *DEFB*, were located in the overlapping segments, along with members of the



**Fig. 4**—Copy number patterns of *CYP2C76* and *GSTM5* across the nine *Macaca* species. (A) *CYP2C76* (chr9: 90,280,000–90,360,000) and (B) *GSTM5* (chr1: 110,670,000–110,770,000). The black baselines in the tracks indicate copy number of two, and CNV regions are indicated with black dashed box. As in figure 1, copy number was estimated in windows containing 1 kb of nongap, nonmasked sequence.



**Fig. 5**—Lineage-specific interspecific CNVs are displayed on the branches of the NJ tree of nine *Macaca* species, which was generated by SNPhylo based on thinned genomic SNVs (500k sites). Bootstrap values are at each node, as determined by 1,000 bootstraps. Species groups defined by Zinner et al. (2013) and Roos et al. (2014) are labeled in green ovals on the tree.

most polymorphic CNV-enriched gene families: olfactory receptor (*OR*), *TRIM*, and *ZNF*. Studies involving more species are needed to investigate this pattern further.

Functional enrichment analyses performed with g:Profiler showed that the 722 genes intersecting CNV-dense regions were enriched in immune function, as suggested by the most significantly enriched pathway “antigen processing and presentation” (*mcc04612*,  $P = 1.81E-05$ ) and enriched GO terms including immunoglobulin production (GO:0002377,  $P = 4.4E-08$ ), antigen processing and presentation of peptide antigen via MHC class I (GO:0002474,  $P = 1.08E-06$ ), MHC protein complex (GO:0042611,  $P = 1.9E-08$ ), and peptide antigen binding (GO:0042605,  $P = 0.00229$ ).

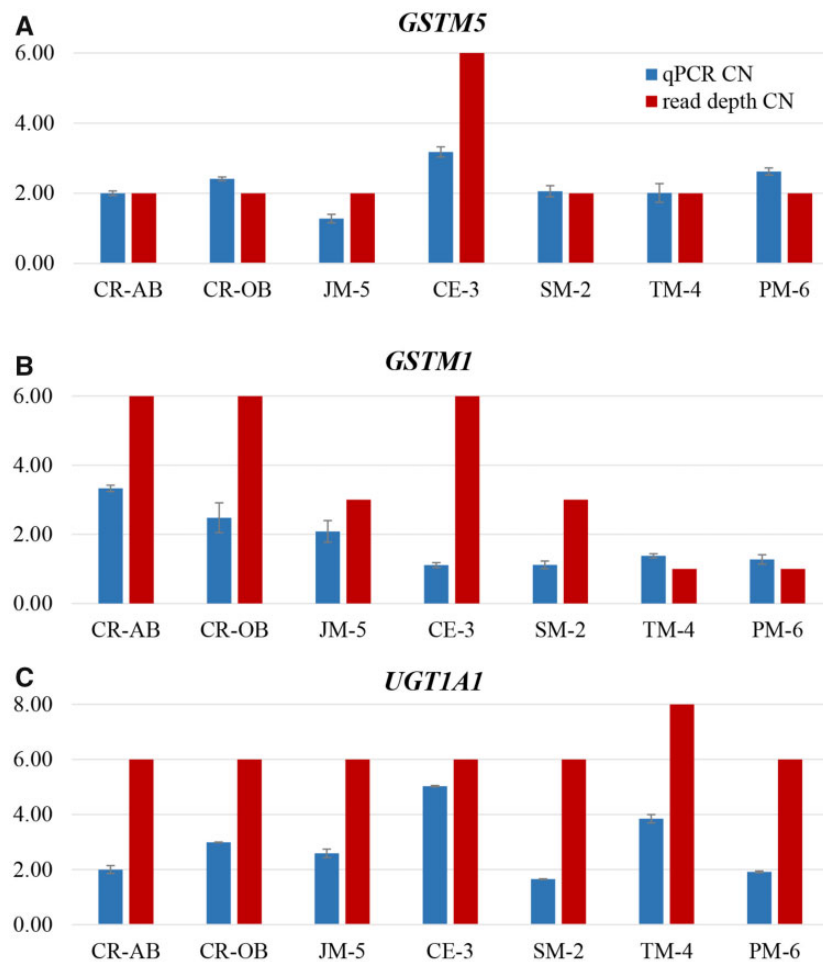
### Lineage-Specific CNVs

To address evolutionary issues of these sibling species, we reconstructed the *Macaca* phylogeny with SNPhylo using thinned genomic SNVs with the NJ method and surveyed the lineage-specific CNVs according to the resulting topology (fig. 5). Six of the seven-species groups defined by Zinner et al. (2013) and Roos et al. (2014) in *Macaca* are included in this study. We did not identify any CNV specific to group *mulatta*

(CR, TwM, and JM), *fascicularis* (CE), *sinica* (TM), or *arctoides* (SM). This absence may reflect the very short species divergence time in the four species groups.

We observed that BM, the only member in *sylvanus*, formed the longest branch of the phylogenetic tree and there were 12 gains and 40 losses specific to this clade (fig. 5). It is intriguing that the number of specific loss events was more than twice the number of specific gain events. BM-specific (*sylvanus*-specific) CNVs overlapped with genes related to metabolism and immunity, including duplications in *PRKD1/PKD1* and *CD55*, and deletions in *ADIPOR2*, *TBX20*, and *SERINC5* (supplementary fig. S17, Supplementary Material online). These CNVs also intersect *LRP1B*, a gene whose deletion or downregulation significantly correlates with acquired chemotherapy resistance in high-grade serious cancers (Cowin et al. 2012).

Twenty-two duplications and one deletion were specific to the large group comprising all other species. These CNVs intersect with genes that are mainly involved in metabolism and immunity, too, such as *ACBD7*, *APOBEC3F*, and *IGHV7-4-1*. The *silenus* species group (composed of LM and PM) displayed 16 shared CNVs, including three duplications and 13 deletions. One of the *silenus*-specific duplications is just ~500



**Fig. 6**—Comparison of copy number patterns determined by genomic qPCR and bioinformatic analysis. Copy numbers from qPCR (blue) and bioinformatically estimated (red) approaches are provided for genes (A) *GSTM5*, (B) *GSTM1*, and (C) *UGT1A1*. The whiskers stand for standard errors of copy numbers estimated by independent technical replicates of qPCR experiments.

bp away from the gene *NPTX2*, a member of neuronal pentraxin family associated with neurological disorders and cancers.

#### Genomic qPCR Validation of CNVs

We attempted to validate the CNs of drug metabolism genes including *CYP2C76*, *UGT2B33*, *UGT1A1*, *GSTM5*, and *GSTM1* with qPCR conducted on genomic DNA from samples of the same species, but different than the resequenced individuals. Results of qPCR indicated that CNV detection based on RD was credible and also demonstrated that some of the identified CNVs were not fixed in the studied species. In detail, the CN distribution of *GSTM5* was consistent with that based on NGS data (fig. 6A). And CN patterns of *GSTM1* and *UGT1A1* were generally in accord with that obtained from NGS data except for the cynomolgus macaque (fig. 6B and C). Taking *UGT1A1* for example, Tibetan and cynomolgus macaques have more copies of this region than others according

to the qPCR results, whereas CNV detection with NGS data showed that Tibetan macaque had more copies than other species who displayed similar CNs. These inconsistencies could be due to outstanding genetic variation in the populations of cynomolgus macaques which was demonstrated by Li et al. (2018), Ling et al. (2011), and Satkoski Trask et al. (2013), because the qPCR sample originated from a breeding population in China of unclear geographical source; however, the resequenced individual came from Vietnam (Yan et al. 2011). Therefore, our results suggest that *GSTM1* and *UGT1A1* are CN polymorphic in cynomolgus macaque. Because of failure in primer design or PCR, we were unable to validate the CNV in *CYP2C76* and *UGT2B33*.

#### Differential Expression of the CNV Intersecting Genes

Based on the blood transcriptomes of 28 Chinese rhesus macaques and 24 Tibetan macaques (Yan 2019), we explored if the CNVs had an impact on gene expression. We discovered

that a considerable proportion of CNVGs were DEGs. In total, 370 CNVs showed distinct copy numbers ( $CND \geq 1.6$ ) between the Chinese rhesus macaque and the Tibetan macaque and intersected with 135 genes with quantifiable expression based on the transcriptome data. Approximately 46% (62/135) of these CNVGs were DEGs ( $P < 0.05$  and  $q < 0.05$ , [supplementary table S7, Supplementary Material](#) online). However, the ratio was not significantly different from the randomly resampled data ( $P = 0.065$ , [supplementary fig. S18, Supplementary Material](#) online). Genes playing important roles in metabolism (*APOL1*, *PDK3*, and *GLUD1*), immune function (*IL9R*, *LILRB1*, *LILRA2*, *MAMU-A*, and *MAMU-A3*) along with zinc finger genes were included in the DEGs.

## Discussion

This study represents the most extensive assessment of CNV across macaques to date, sampling six out of seven-species groups in *Macaca* and including some less studied species. Our results provide new evidence for the involvement of CNVs in the adaptation and evolution of macaques. In total, 1,479 CNVs, constituting 1.41% of the macaque genome, along with 1,560 duplications shared across species, 26 CNV-dense regions, and dozens of lineage-specific CNVs were identified. High coverage genome data and an improved CNV detection pipeline based on fastCN allowed for a higher-resolution map of CNVs across *Macaca* species. Although each species was only represented by single individual, our study identified interspecific genetic divergence in *Macaca* from the perspective of structural variation, in contrast to previous genomic CNV studies in rhesus and cynomolgus macaques that focused on intraspecific genetic polymorphism (Lee et al. 2008; Gokcumen et al. 2011). Function enrichment and expression-level analyses with transcriptome data from Yan (2019) suggest roles for CNV in environmental adaptation and genome evolution of *Macaca*, with implications for the usage of these NHPs in drug metabolism or diseases research.

### Characteristics of the CNV Regions

We uncovered the general CN patterns and chromosomal distributions of CNVs among *Macaca* species. For example, we found that chromosome 19, the shortest autosome, displayed the most significant enrichment of CNVs. This chromosome is also enriched for genes and microsatellites (Xu et al. 2016) in macaques. In genome-wide CNVs, we observed a higher number of gains (1,106) relative to losses (451). This imbalance was unlikely a bias derived from CNV detection methods, because identification of deletions is very robust in NGS approaches (Pinto et al. 2011; Pang et al. 2014) and previous CNV studies of rhesus macaque (Lee et al. 2008) and human (Sudmant et al. 2015) also detected more genomic gains than losses. Furthermore, genomes are more likely

tolerant of duplications than deletions which could result in loss of functions (Brewer et al. 1999) and are typically selected against (Zarrei et al. 2015). However, within a single lineage, gene copy losses generally dominated gains, demonstrating that CNV deletions are also related to the phylogenetic evolution and may be used as phylogenetic markers for *Macaca*.

Along with shared duplications, CNVs were distributed unevenly in the genomes of macaques. Greater than 52% of CNVs were located directly in genes, whereas only 34–36% of interspecific SNVs were genic for macaques (Li et al. 2018). This is in agreement with previous finding that CNVs were prone to occur in gene-rich regions (Conrad et al. 2010). CNVs and shared duplications overlapped more frequently with relatively short genes than expected, which may be a true trait of these regions attributed to gene clustering or a bias due to the quality of the reference genome. In addition, 26 CNV-dense regions were identified, with 15 regions specific to *Macaca* and 11 shared by human, chimpanzee, and macaques. According to a human study of such loci (Dumas et al. 2007), CNV-dense regions are prone to gene instability and are possible “gene nurseries” where new gene families may be emerging, facilitating biological innovation and rapid evolution of macaques. Genes overlapping the 11 potential CNV hotspots in primates included several ORs and immunity-related genes, such as *HLA*, *HCG9*, *DEFA*, and *DEFB*, which suggests that diversity in immunity represents a main evolutionary strategy in primates.

### The Possible Role of CNV in Adaptations of *Macaca*

Although some enriched GO or KEGG terms lacked significant support based on corrected  $P$  value, we do find that CNVs are functionally relevant, with a bias toward metabolism and immunity function. CNVGs were mainly enriched for nutritional metabolism, xenobiotics/drug metabolism, and immune-related pathways (table 2). Using expression data of Chinese rhesus and Tibetan macaques (Yan 2019), we found that differentially expressed CNVGs also mainly consisted of metabolic and immune-related genes (e.g., *APOL1* and *LILRB1*). The functional categories were not only partially overlapping with the enrichment outputs of all DEGs between the two species (Yan 2019) but also consistent with results from a comparative transcriptome study (Li et al. 2017) in which these expression differences were found to be mainly in the GO term of nutrient reservoir activity and KEGG subcategories including infectious diseases and immune system. Our results indicate that these monkeys are genetically divergent from one another in metabolism and immunity, agreeing with the conclusions of a SNV study of macaques (Li et al. 2018), and also indicate that CNVs may affect gene functions related to environmental responses such as metabolism and immune response. Given that *Macaca* species have different foraging habits (Srivastava 1999; Hanya et al. 2011), body sizes (Solari and Baker 2006), and immunity traits (Trichel

et al. 2002; De Vries et al. 2012), these findings may reflect adaptation to diverse habitats.

Biological processes influencing adaptation have been identified in CNVGs of many animals, including metabolic processes, stress response, and defense response. For example, CNVs in the  $\alpha$ -amylase gene facilitated adaptation to dietary starch consumption in both humans (Mandel and Breslin 2012) and dog breeds (Mandel and Breslin 2012; Arendt et al. 2016). CNV-overlapping genes related to drug detoxification and innate or adaptive immunity were overrepresented in human (Freeman et al. 2006; Almal and Padh 2012), pig (Wang et al. 2012; Paudel et al. 2013), dog (Nicholas et al. 2009), and cattle (Fadista et al. 2010; Hou et al. 2011), involving gene families like *CYP*, *ABC*, *HLA*, *MHC*, *BD*, *IL*, and *OR*, which were also present in the CNVGs identified in this study. This can be explained by a general model that phylogenetically stable genes have core functions in development and physiology, whereas unstable genes have accessory functions associated with unstable environmental interactions such as toxin and pathogen exposure (Thomas 2007).

However, it is worth noting that there are limitations for gene set enrichment analysis even in human where the majority of annotations is generated, and more uncertainty exists in species such as macaques which have less precise gene annotations. For example, some immune-related pathways may include poorly annotated genes containing immunoglobulin-like domains that are evolving fast and hence are subject to duplication, without truly being involved in immunity-related traits. Additionally, we found CNVs tended to overlap with gene clusters, necessitating another layer of cautiousness on the enrichment results. Therefore, more investigation is needed to elucidate the connection between the identified CNVs and adaptive differences between *Macaca* species.

### Implications for the Biomedical Application of *Macaca* Species

*Macaca* species have been extensively used as experimental models in drug discovery research and drug safety evaluation, including rhesus, crab-eating, and Barbary macaques (Zuber et al. 2002). Intriguingly, xenobiotics/drug metabolism was one of the most enriched biologic processes for the CNVGs in *Macaca*, suggesting that various macaques could react differently to drugs. Highly overrepresented CNVGs included *CYP2C76*, *UGT1A1*, *UGT2B33*, and *GSTM5*, which belong to three well-known drug-metabolizing enzyme families, *CYP*, *UGT*, and *GST*. *CYP2C76* and *GSTM5* expanded in recently diverged species, such as the crab-eating, Chinese rhesus, and Taiwanese macaques (fig. 4).

Drug metabolism genes can determine drug half-life (Linder et al. 1997; He et al. 2011). These highly polymorphic genes are thus important in pharmaceutical development (Linder et al. 1997; He et al. 2011). A previous study showed

that *CYP* genes in macaques were nearly identical to the orthologs in human (Uno et al. 2011). CNVs in these genes were also observed in human (He et al. 2011; Fuselli 2019), and individuals with more than two copies of *CYP2D6* wild-type alleles had elevated *CYP2D6* enzyme activity (Ingelman-Sundberg 2005). Additionally, we found that ~46% of the CNVGs with distinct copies in Chinese rhesus and Tibetan macaques were differentially expressed, suggesting that a large proportion of CNVGs have an altered expression level and may result in different phenotypes. Therefore, we propose that the Barbary, lion-tailed, crab-eating, and Chinese rhesus macaques may differ in drug metabolism of certain substrates due to CNV of drug metabolism-related genes. This CNV study, to some extent, provides theoretical basis for the selection of optimal NHP models for drug research and preclinical toxicology tests. Further functional studies of individuals CNVs are needed to fully address this issue.

Because CNV plays an essential role in phenotypes and diseases (Stranger et al. 2007; Zhang et al. 2009; Almal and Padh 2012), CNVs can affect the outcome and interpretation of biomedical studies in which various *Macaca* species are employed as NHP models of diseases, especially CNVs overlapping with genes related to immunity or diseases (table 2). Thus, genetic characterization of the macaques is recommended before their usage in biomedical research.

### BM-Specific/*sylvanus*-Specific CNVs

The Barbary macaque is the sole living member of a distinct and ancient species group in *Macaca*, *sylvanus* (Fa 2012), and is the only NHP indigenous to North Africa (Taub 1978). It lives for extended periods in snow-covered areas during winter, suffering from not only cold stress but also food shortage. CNVs specific to BM/*sylvanus* intersected with genes including *PRKD1/PKD1*, *ADIPOR2*, and *TBX20* (supplementary fig. S17, Supplementary Material online), they may have a role in the adaptation of BM to the harsh environment of its habitats. A recent study of pancreatic  $\beta$  cells found that protein PKD1 controlled the granule degradation in response to nutrient availability and concluded that switching from macroautophagy to insulin granule degradation using a PKD-dependent mechanism was important to keep insulin secretion low upon fasting (Goginashvili et al. 2015). Therefore, the BM-specific duplication located in the intron of *PKD1* may affect its expression and may enable BM to better control insulin secretion during starvation, aiding in winter survival.

A BM-specific deletion overlapped with the first intron of *ADIPOR2*, a gene that is highly conserved from yeast to human (Tang et al. 2005) and plays important roles in the regulation of glucose and lipid metabolism, inflammation, and oxidative stress. Targeted disruption of *ADIPOR2* decreased the activity of PPAR-alpha signaling pathways, affecting lipid metabolism and adaptive thermogenesis (Yamauchi et al. 2007). The partial deletion in the intron may change the



metabolic actions of glucose and lipid, and thermogenesis, probably via downregulation of its expression and then increasing the level of adiponectin. Park et al. (2011) found that long-term central infusion of adiponectin improves energy and glucose homeostasis by decreasing fat storage and suppressing hepatic gluconeogenesis without changing food intake, suggesting increased adiponectin leads to high level of glucose homeostasis. It is plausible that this CNV would benefit BM in food shortage and coldness during winter.

Partial deletion of *TBX20* was another event specific to *BM/sylvanus*. Sakabe et al. (2012) found in adult *TBX20*−/− hearts, additional genes involved in cardiovascular biology and energy metabolism were downregulated, whereas genes related to immune response and cell proliferation were upregulated. This deletion might lower energy metabolism requirements and increase immunity defenses in BM, which could be beneficial to the Barbary macaque in an environment where nutrient shortage is frequent. Further functional studies would help address these hypotheses.

### Challenges in CNV Study of Macaques

Several challenges remain for the CNV study of macaques. First, although a single individual is informative of interspecific divergence of *Macaca*, it is necessary to verify CNVs among *Macaca* species on a population scale. Some regions may be CN variable among individuals, as demonstrated by the cynomolgus macaque in the genomic qPCR validation. A population-scale CNV study of drug metabolism genes in macaques is highly desirable to assess the impact of such variation on biomedical studies. In addition to the gene set enrichment strategy, more robust evidence is required to clarify the role of CNVs in environmental adaptation of *Macaca* species. The interpretation of gene set enrichment is uncertain even in humans where gene annotations are superior to most species. These uncertainties are exasperated in divergent species like macaques.

In NGS-based methods, the accuracy and sensitivity of CNV identification depend heavily on the quality of the reference genome. Functional analysis of CNVs also calls for accurate gene models. The relatively short lengths of CNVGs suggested that inferior quality of the reference genome (Mmul\_8) had an effect on our study. Single molecule, real-time sequencing is a very promising way to improve the continuity of the reference genome with very long reads (McCarthy 2010; Roberts et al. 2013). Recently, the single-molecule assembly of Chinese rhesus macaque (He et al. 2019) has been reported. Along with an increasing number of intensive genome studies, this resource can undoubtedly improve CNV detection in macaques. In turn, CNV surveys can broaden our knowledge of *Macaca* genome variation.

### Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

### Acknowledgments

This work was supported by the State Key Program of National Natural Science Foundation of China (31530068), the National Natural Science Foundation of China (31770415), the Enterprises and Institutions Cooperation Project funded by Science and Technology Agency of Sichuan Province (18SYXHZ0019), the Fundamental Research Funds for the Central Universities (2012017yjsy153), and the China Scholarship Council. We sincerely thank Sarah Emery for language assistance in the manuscript draft and Chaochao Yan for providing the transcriptome data.

### Author Contributions

J.L., J.L. (first author), and B.Y. designed this study. J.L. (first author) performed data analysis and wrote the article. Z.F. participated in genomic data collection. F.S., A.L.P., and J.M.K. contributed to methodology. J.L. and J.M.K. revised the article. Y.S. and J.X. respectively participated in the data analysis and qPCR experiment design.

### Data Availability

The genome-wide copy number data are available in figshare at <http://doi.org/10.6084/m9.figshare.9900401>. The scripts for gene annotation, enrichment, and permutation can be found at [https://github.com/umscholj/Macaca\\_CNV/blob/master/annotation\\_enrichment\\_permutation](https://github.com/umscholj/Macaca_CNV/blob/master/annotation_enrichment_permutation). Other additional information is available in the [Supplementary Material](#) online.

### Literature Cited

- Alkan C, Coe BP, Eichler EE. 2011. Genome structural variation discovery and genotyping. *Nat Rev Genet.* 12(5):363–376.
- Almal SH, Padh H. 2012. Implications of gene copy-number variation in health and diseases. *J Hum Genet.* 57(1):6–13.
- Alvarez CE, Akey JM. 2012. Copy number variation in the domestic dog. *Mamm Genome* 23(1–2):144–163.
- Arendt M, Cairns K, Ballard J, Savolainen P, Axelsson E. 2016. Diet adaptation in dog reflects spread of prehistoric agriculture. *Heredity* 117(5):301–306.
- Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27(2):573–580.
- Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114–2120.
- Brennan G, Kozyrev Y, Hu S-L. 2008. TRIMCyp expression in Old World primates *Macaca nemestrina* and *Macaca fascicularis*. *Proc Natl Acad Sci U S A.* 105(9):3569–3574.
- Brewer C, Holloway S, Zawalynski P, Schinzel A, FitzPatrick D. 1999. A chromosomal duplication map of malformations: regions of suspected

- haplo-and triplolethality—and tolerance of segmental aneuploidy—in humans. *Am J Hum Genet.* 64(6):1702–1708.
- Cahan P, Li Y, Izumi M, Graubert TA. 2009. The impact of copy number variation on local gene expression in mouse hematopoietic stem and progenitor cells. *Nat Genet.* 41(4):430–437.
- Carter NP. 2007. Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat Genet.* 39(5):S16–S21.
- Charchar FJ, et al. 2010. Whole genome survey of copy number variation in the spontaneously hypertensive rat: relationship to quantitative trait loci, gene expression, and blood pressure. *Hypertension* 55(5):1231–1238.
- Chen C, et al. 2012. A comprehensive survey of copy number variation in 18 diverse pig populations and identification of candidate copy number variable genes associated with complex traits. *BMC Genomics.* 13(1):733.
- Conrad DF, et al. 2010. Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. *Nat Genet.* 42(5):385–391.
- Cowin PA, et al. 2012. *LRP1B* deletion in high-grade serous ovarian cancers is associated with acquired chemotherapy resistance to liposomal doxorubicin. *Cancer Res.* 72(16):4060–4073.
- De Vries RD, et al. 2012. Measles immune suppression: lessons from the macaque model. *PLoS Pathog.* 8(8):e1002885.
- Degenhardt JD, et al. 2009. Copy number variation of *CCL3-like* genes affects rate of progression to simian-AIDS in rhesus macaques (*Macaca mulatta*). *PLoS Genet.* 5(1):e1000346.
- Depristo MA, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet.* 43(5):491–498.
- Doan R, et al. 2012. Identification of copy number variants in horses. *Genome Res.* 22(5):899–907.
- Dumas L, et al. 2007. Gene copy number variation spanning 60 million years of human and primate evolution. *Genome Res.* 17(9):1266–1277.
- Fa JE. 2012. The Barbary macaque: a case study in conservation. New York: Plenum Press.
- Fadista J, Thomsen B, Holm L-E, Bendixen C. 2010. Copy number variation in the bovine genome. *BMC Genomics.* 11(1):284.
- Fan X, Abbott TE, Larson DB, Chen K. 2014. BreakDancer: identification of genomic structural variation from paired-end read mapping. *Curr Protoc Hum Genet.* 45:15.6.1–15.6.11.
- Fan Z, Zhao G, et al. 2014. Whole-genome sequencing of Tibetan macaque (*Macaca thibetana*) provides new insight into the macaque evolutionary history. *Mol Biol Evol.* 31(6):1475–1489.
- Fan Z, et al. 2018. Ancient hybridization and admixture in macaques (genus *Macaca*) inferred from whole genome sequences. *Mol Phylogenet Evol.* 127:376–386.
- Fang X, et al. 2011. Genome sequence and global sequence variation map with 5.5 million SNPs in Chinese rhesus macaque. *Genome Biol.* 12(7):R63.
- Feuk L, Carson AR, Scherer SW. 2006. Structural variation in the human genome. *Nat Rev Genet.* 7(2):85–97.
- Freeman JL, et al. 2006. Copy number variation: new insights in genome diversity. *Genome Res.* 16(8):949–961.
- Fuselli S. 2019. Beyond drugs: the evolution of genes involved in human response to medications. *Proc R Soc B* 286(1913):20191716.
- Gazave E, et al. 2011. Copy number variation analysis in the great apes reveals species-specific patterns of structural variation. *Genome Res.* 21(10):1626–1639.
- Goginashvili A, et al. 2015. Insulin secretory granules control autophagy in pancreatic  $\beta$  cells. *Science* 347(6224):878–882.
- Gökçümen Ö, Lee C. 2009. Copy number variants (CNVs) in primate species using array-based comparative genomic hybridization. *Methods* 49(1):18–25.
- Gokcumen O, et al. 2011. Refinement of primate copy number variation hotspots identifies candidate genomic regions evolving under positive selection. *Genome Biol.* 12(5):R52.
- Goshu HA, et al. 2019. Genomic copy number variation of the *CHKB* gene alters gene expression and affects growth traits of Chinese domestic yak (*Bos grunniens*) breeds. *Mol Genet Genomics.* 294(3):549–561.
- Gschwind AR, 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(1):144.
- Hanya G, et al. 2011. Dietary adaptations of temperate primates: comparisons of Japanese and Barbary macaques. *Primates* 52(2):187–198.
- Hatzioannou T, et al. 2009. A macaque model of HIV-1 infection. *Proc Natl Acad Sci U S A.* 106(11):4425–4429.
- Haus T, et al. 2014. Genome typing of nonhuman primate models: implications for biomedical research. *Trends Genet.* 30(11):482–487.
- He Y, Hoskins JM, McLeod HL. 2011. Copy number variants in pharmacogenetic genes. *Trends Mol Med.* 17(5):244–251.
- He Y, et al. 2019. Long-read assembly of the Chinese rhesus macaque genome and identification of ape-specific structural variants. *Nat Commun.* 10(1):4233.
- Henrichsen CN, Chaignat E, Reymond A. 2009. Copy number variants, diseases and gene expression. *Hum Mol Genet.* 18(R1):R1–R8.
- Higashino A, et al. 2012. Whole-genome sequencing and analysis of the Malaysian cynomolgus macaque (*Macaca fascicularis*) genome. *Genome Biol.* 13(7):R58.
- Hou Y, et al. 2011. Genomic characteristics of cattle copy number variations. *BMC Genomics.* 12(1):127.
- Hou Y, et al. 2012. Genomic regions showing copy number variations associate with resistance or susceptibility to gastrointestinal nematodes in Angus cattle. *Funct Integr Genomics.* 12(1):81–92.
- Ingelman-Sundberg M. 2005. Genetic polymorphisms of cytochrome *P450* 2D6 (CYP2D6): clinical consequences, evolutionary aspects and functional diversity. *Pharmacogenomics J.* 5(1):6–13.
- Ishii K, et al. 2006. Characteristics and clustering of human ribosomal protein genes. *BMC Genomics.* 7(1):37.
- Iskow RC, Gokcumen O, Lee C. 2012. Exploring the role of copy number variants in human adaptation. *Trends Genet.* 28(6):245–257.
- Itsara A, et al. 2009. Population analysis of large copy number variants and hotspots of human genetic disease. *Am J Hum Genet.* 84(2):148–161.
- Jia X, et al. 2013. Copy number variations identified in the chicken using a 60K SNP BeadChip. *Anim Genet.* 44(3):276–284.
- Jiang J, et al. 2016. Mitochondrial genome and nuclear markers provide new insight into the evolutionary history of macaques. *PLoS One* 11(5):e0154665.
- Jun J, et al. 2014. Whole genome sequence and analysis of the Marwari horse breed and its genetic origin. *BMC Genomics.* 15(Suppl 9):S4.
- Jung S-H, et al. 2013. De novo copy number variations in cloned dogs from the same nuclear donor. *BMC Genomics.* 14(1):863.
- Keel BN, Lindholm-Perry AK, Snelling WM. 2016. Evolutionary and functional features of copy number variation in the cattle genome. *Front Genet.* 7:207.
- Kondrashov FA. 2012. Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proc R Soc B Biol Sci.* 279(1749):5048–5057.
- Lee AS, et al. 2008. Analysis of copy number variation in the rhesus macaque genome identifies candidate loci for evolutionary and human disease studies. *Hum Mol Genet.* 17(8):1127–1136.
- Lee T-H, Guo H, Wang X, Kim C, Paterson AH. 2014. SNPPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. *BMC Genomics.* 15(1):162.
- Li C, Zhao C, Fan PF. 2015. White-cheeked macaque (*Macaca leucogenys*): a new macaque species from Medog, southeastern Tibet. *Am J Primatol.* 77(7):753–766.

- Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25(14):1754–1760.
- Li J, et al. 2018. Comparative genome-wide survey of single nucleotide variation uncovers the genetic diversity and potential biomedical applications among six *Macaca* species. *Int J Mol Sci.* 19(10):3123.
- Li P, et al. 2017. Generation and characterization of the blood transcriptome of *Macaca thibetana* and comparative analysis with *M. mulatta*. *Mol Biosyst.* 13(6):1121–1130.
- Li W, Olivier M. 2013. Current analysis platforms and methods for detecting copy number variation. *Physiol Genomics.* 45(1):1–16.
- Linder MW, Prough RA, Valdes R. 1997. Pharmacogenetics: a laboratory tool for optimizing therapeutic efficiency. *Clin Chem.* 43(2):254–266.
- Ling F, et al. 2011. Characterization of the major histocompatibility complex class II *DOB*, *DPB1*, and *DQB1* alleles in cynomolgus macaques of Vietnamese origin. *Immunogenetics* 63(3):155–166.
- Liu Z, et al. 2018. Population genomics of wild Chinese rhesus macaques reveals a dynamic demographic history and local adaptation, with implications for biomedical research. *GigaScience* 7(9):gij106.
- Lupski JR, Stankiewicz P. 2005. Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS Genet.* 1(6):e49.
- Mandel AL, Breslin PA. 2012. High endogenous salivary amylase activity is associated with improved glycemic homeostasis following starch ingestion in adults. *J Nutr.* 142(5):853–858.
- Marques-Bonet T, et al. 2009. A burst of segmental duplications in the genome of the African great ape ancestor. *Nature* 457(7231):877–881.
- McCarthy A. 2010. Third generation DNA sequencing: pacific biosciences' single molecule real time technology. *Chem Biol.* 17(7):675–676.
- Meyerson M, Gabriel S, Getz G. 2010. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet.* 11(10):685–696.
- Newman RM, et al. 2008. Evolution of a TRIM5–CypA splice isoform in old world monkeys. *PLoS Pathog.* 4(2):e1000003.
- Nicholas TJ, Baker C, Eichler EE, Akey JM. 2011. A high-resolution integrated map of copy number polymorphisms within and between breeds of the modern domesticated dog. *BMC Genomics.* 12(1):414.
- Nicholas TJ, et al. 2009. The genomic architecture of segmental duplications and associated copy number variants in dogs. *Genome Res.* 19(3):491–499.
- Oetjens MT, Shen F, Emery SB, Zou Z, Kidd JM. 2016. Y-chromosome structural diversity in the bonobo and chimpanzee lineages. *Genome Biol Evol.* 8(7):2231–2240.
- Ottolini B, et al. 2014. Evidence of convergent evolution in humans and macaques supports an adaptive role for copy number variation of the  $\beta$ -defensin-2 gene. *Genome Biol Evol.* 6(11):3025–3038.
- Pang AWC, MacDonald JR, Yuen RK, Hayes VM, Scherer SW. 2014. Performance of high-throughput sequencing for the discovery of genetic variation across the complete size spectrum. *G3 (Bethesda)* 4:63–65.
- Park S, Kim D, Kwon D, Yang H. 2011. Long-term central infusion of adiponectin improves energy and glucose homeostasis by decreasing fat storage and suppressing hepatic gluconeogenesis without changing food intake. *J Neuroendocrinol.* 23(8):687–698.
- Paudel Y, et al. 2013. Evolutionary dynamics of copy number variation in pig genomes in the context of adaptation and domestication. *BMC Genomics.* 14(1):449.
- Paudel Y, et al. 2015. Copy number variation in the speciation of pigs: a possible prominent role for olfactory receptors. *BMC Genomics.* 16(1):330.
- Pendleton AL, et al. 2018. Comparison of village dog and wolf genomes highlights the role of the neural crest in dog domestication. *BMC Biol.* 16(1):64.
- Perry GH, et al. 2008. Copy number variation and evolution in humans and chimpanzees. *Genome Res.* 18(11):1698–1710.
- Pinto D, et al. 2011. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nat Biotechnol.* 29(6):512–520.
- Pouladi MA, Morton AJ, Hayden MR. 2013. Choosing an animal model for the study of Huntington's disease. *Nat Rev Neurosci.* 14(10):708–721.
- Purcell S, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 81(3):559–575.
- Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6):841–842.
- Redon R, et al. 2006. Global variation in copy number in the human genome. *Nature* 444(7118):444–454.
- Reimand J, Kull M, Peterson H, Hansen J, Vilo J. 2007. g: profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.* 35(Suppl 2):W193–W200.
- Roberts RJ, Carneiro MO, Schatz MC. 2013. The advantages of SMRT sequencing. *Genome Biol.* 14(6):405.
- Roos C, Zinner D. 2015. Diversity and evolutionary history of macaques with special focus on *Macaca mulatta* and *Macaca fascicularis*. In: The nonhuman primate in nonclinical drug development and safety assessment. San Diego: Academic Press. p. 3–16.
- Roos C, et al. 2014. An updated taxonomy and conservation status review of Asian primates. *Asian Primates J.* 4(1):2–38.
- Sakabe NJ, et al. 2012. Dual transcriptional activator and repressor roles of TBX20 regulate adult cardiac structure and function. *Hum Mol Genet.* 21(10):2194–2204.
- Satkoski Trask JA, et al. 2013. Single-nucleotide polymorphisms reveal patterns of allele sharing across the species boundary between rhesus (*Macaca mulatta*) and cynomolgus (*M. fascicularis*) macaques. *Am J Primatol.* 75(2):135–144.
- Solari S, Baker RJ. 2006. Mammal species of the world, a taxonomic and geographic reference. *Mastozoool Neotrop.* 13:290–293.
- Srivastava A. 1999. Primates of Northeast India. Bikaner (India): Megadiversity Press.
- Stranger BE, et al. 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 315(5813):848–853.
- Sudmant PH, et al. 2013. Evolution and diversity of copy number variation in the great ape lineage. *Genome Res.* 23(9):1373–1382.
- Sudmant PH, et al. 2015. Global diversity, population stratification, and selection of human copy-number variation. *Science* 349(6253):aab3761.
- Tang YT, et al. 2005. PAQR proteins: a novel membrane receptor family defined by an ancient 7-transmembrane pass motif. *J Mol Evol.* 61(3):372–380.
- Taub DM. 1978. The Barbary macaque in North Africa. *Oryx* 14(3):245–253.
- Thierry B. 2007. Unity in diversity: lessons from macaque societies. *Evol Anthropol.* 16(6):224–238.
- Thomas JH. 2007. Rapid birth–death evolution specific to xenobiotic cytochrome P450 genes in vertebrates. *PLoS Genet.* 3(5):e67.
- Trichel A, et al. 2002. Species-specific variation in SIV disease progression between Chinese and Indian subspecies of rhesus macaque. *J Med Primatol.* 31(4–5):171–178.
- Uno Y, Iwasaki K, Yamazaki H, Nelson DR. 2011. Macaque cytochromes P450: nomenclature, transcript, gene, genomic structure, and function. *Drug Metab Rev.* 43(3):346–361.
- Uno Y, Uehara S, Kohara S, Murayama N, Yamazaki H. 2010. Cynomolgus monkey CYP2D44 newly identified in liver, metabolizes bufuralol and dextromethorphan. *Drug Metab Dispos.* 38(9):1486–1492.
- Ventura M, et al. 2011. Gorilla genome structural variation reveals evolutionary parallelisms with chimpanzee. *Genome Res.* 21(10):1640–1649.
- Völker M, et al. 2010. Copy number variation, chromosome rearrangement, and their association with recombination during avian evolution. *Genome Res.* 20(4):503–511.

- Wang J, et al. 2012. A genome-wide detection of copy number variations using SNP genotyping arrays in swine. *BMC Genomics*. 13(1):273.
- Wang S, et al. 2018. De novo sequence and copy number variants are strongly associated with Tourette disorder and implicate cell polarity in pathogenesis. *Cell Rep*. 24(13):3441–3454.e3412.
- Wang X, Nahashon S, Feaster TK, Bohannon-Stewart A, Adefope N. 2010. An initial map of chromosomal segmental copy number variations in the chicken. *BMC Genomics*. 11(1):351.
- Xie C, et al. 2011. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Res*. 39(Suppl 2):W316–W322.
- Xu Y, et al. 2016. Characterization of perfect microsatellite based on genome-wide and chromosome level in Rhesus monkey (*Macaca mulatta*). *Gene* 592(2):269–275.
- Yamauchi T, et al. 2007. Targeted disruption of AdipoR1 and AdipoR2 causes abrogation of adiponectin binding and metabolic actions. *Nat Med*. 13(3):332–339.
- Yan C. 2019. Differential expression analysis of genes in blood tissues of two macaques during the process of ageing. Chengdu (China): Sichuan University.
- Yan G, et al. 2011. Genome sequencing and comparison of two nonhuman primate animal models, the cynomolgus and Chinese rhesus macaques. *Nat Biotechnol*. 29(11):1019–1023.
- Yang L, et al. 2018. Diversity of copy number variation in a worldwide population of sheep. *Genomics*. 110(3):143–148.
- Younger RM, et al. 2001. Characterization of clustered MHC-linked olfactory receptor genes in human and mouse. *Genome Res*. 11(4):519–530.
- Zarrei M, MacDonald JR, Merico D, Scherer SW. 2015. A copy number variation map of the human genome. *Nat Rev Genet*. 16(3):172–183.
- Zhang F, Gu W, Hurles ME, Lupski JR. 2009. Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet*. 10(1):451–481.
- Zhang X, et al. 2017. Genetic characterization of a captive colony of pig-tailed macaques (*Macaca nemestrina*). *J Am Assoc Lab Anim Sci*. 56(4):390–395.
- Zhong X, et al. 2016. RhesusBase PopGateway: genome-wide population genetics atlas in rhesus macaque. *Mol Biol Evol*. 33(5):1370–1375.
- Zimin AV, et al. 2014. A new rhesus macaque assembly and annotation for next-generation sequencing analyses. *Biol Direct* 9(1):15.
- Zinner D, et al. 2013. Family Cercopithecidae (old world monkeys). In: *Handbook of the mammals of the world-primates*. Barcelona: Lynx Edicions.
- Zuber R, Anzenbacherová E, Anzenbacher P. 2002. Cytochromes P450 and experimental models of drug metabolism. *J Cell Mol Med*. 6(2):189–198.

**Associate editor:** David Enard