Supplementary Methods

Identification and annotation of genetic variants in rhesus macaque and human

Whole-genome sequencing data of 24 independent in-house macaque animals (Chen et al. 2015; Yang et al. 2015), together with seven public macaque animals (Fang, et al. 2011; Gokcumen, et al. 2013; Yan, et al. 2011) downloaded from SRA (SRA023856, SRA037810, and ERP002376), were aligned to the macaque reference genome (rheMac2) using BWA (0.7.10-r789) (Li and Durbin 2009). Alignments were merged into a single BAM file and marked for duplicates using Picard-1.118. Only non-duplicate reads were used for the downstream analyses. GenomeAnalysisTK-3.2-2 (GATK) was then used to realign indel and recalibrate base quality (DePristo et al. 2011; McKenna et al. 2010). We further performed variant calling by running HaplotypeCaller and variant refining by using variant quality score recalibration (VQSR). When running HaplotypeCaller, GVCF mode was used to generate an intermediate genomic gVCF for each sample and GenotypeGVCFs was used to pool all of them together to create the raw SNP and indel VCF.

The variants identified by GATK were filtered using VariantRecalibrator. VariantRecalibrator created a Gaussian mixture model to evaluate the quality of candidate variants, using the distribution of scores for high quality variants as the reference. As no reference variant panel is available for rhesus macaque, we firstly defined a high-quality macaque variant panel by performing GATK recommended hard filters and multiple other filtering steps to exclude potential false-positives, including variants 1) in tandem repeat region (homopolymer > 6bp or microsatellite >10bp); 2) in variant clusters with more than 4 variants within 25bp window; 3) with the alternative allele homozygous in all 31 samples; 4) with multiple alternative alleles. The variants defined by this stringent filtering steps were then used as the reference panel to evaluate other variants using VariantRecalibrator, with the threshold of the sensitivity above 99.0%.

The functions of these variants were then annotated using SnpEff (v4.1), according to Ensembl (release 79) (Cingolani et al. 2012). The haplotype structure of 31 animals were further imputed and phased using SHAPEIT (v2.r790) (Delaneau et al. 2013). The ancestral status of these variants was inferred following the Enredo-Pecan-Ortheus (EPO) pipeline (Paten et al. 2008), using sequences from 8 primate species (Human, Gorilla, Chimpanzee, Orangutan, Macaque, Olive baboon, Vervet-AGM, Marmoset).

Using a similar approach, we also profiled the polymorphism data in human populations, by re-analyzing whole genome sequencing data in 103 individuals from different sub-populations, archived with high sequencing coverage by the 1000 Genomes Project **(supplementary table S2)**. A list of 21,485,050 variants were identified across the human genome.

Hierarchical clustering analysis

For each autosomal polymorphism site, the genotype of each animal was coded as follows: 0 for homozygous reference alleles, 1 for heterozygotes, 2 for homozygous alternative alleles. The distance matrix of the genotype data was then calculated using R function "dist" with "euclidean" method, and subsequently plotted into dendrogram using R function "hclust" (**fig. 1B**).

Population genetic analyses in rhesus macaque and human

On the basis of the polymorphism data from the population of 31 macaque animals, we estimated the nucleotide diversity (π), population mutation rate (θw) and Tajima's D in sliding windows of 300 bps and 3 Mbps across the macaque genome, using VCFtools (v0.1.12b) and in-house scripts. Fu and Li's F^* were calculated using scripts implemented in Perl package (Bio::PopGen::Statistics). All of these calculations were performed using standard equations as follows:

(1) *π (Nucleotide Diversity):*

$$\pi = \sum_{ij} x_i x_j \pi_{ij} = 2 * \sum_{i=1}^{n} \sum_{j=1}^{i-1} x_i x_j \pi_{ij}$$

where x_i and x_j are the respective frequencies of the *i*th and *j*th sequences, π_{ij} is the number of nucleotide differences *per* nucleotide site between the *i*th and *j*th sequences, and *n* is the number of sequences in the sample.

(2) *θ (Watterson estimator)*:

$$\hat{\theta}_w = \frac{K}{a_n}$$

where *K* is the number of segregating sites in the sample and a_n is the $(n-1)$th harmonic number.

(3) *Tajima'D:*

$$D = \frac{\pi - \theta_w}{\sqrt{V(\pi - \theta_w)}}$$

(4) *Fu And Li D*:*

$$D^* = \frac{\frac{n}{n-1}\eta - a_n \eta_s}{\sqrt{u_{D^*}\eta + v_{D^*}\eta^2}}$$

$$a_n = \sum_{k=1}^{n-1} \frac{1}{k}$$

$$b_n = \sum_{k=1}^{n-1} \frac{1}{k^2}$$

$$\eta = \sum_{i}^{m} s_i$$

$$u_{D^*} = \frac{n}{n-1}\left(a_n - \frac{n}{n-1}\right) - v_{D^*}$$

$$c_n = 2\frac{na_n - 2(n-1)}{(n-1)(n-2)}$$

$$d_n = c_n + \frac{n-2}{(n-1)^2} + \frac{2}{n-1}\left(\frac{3}{2} - \frac{2a_{n+1} - 3}{n-2} - \frac{1}{n}\right)$$

$$v_{D^*} = \left[\left(\frac{n}{n-1}\right)^2 b_n + a_n^2 d_n - 2\frac{na_n(a_n+1)}{(n-1)^2}\right] / (a_n^2 + b_n)$$

n is the number of sequences in the population, s_i is equal to the number of different nucleotides minus one at site *i* among the *n* sequences, η_s is the number of singletons, which refer to nucleotide that appears only once at the site among the sequences in the

sample.

(5) *Fu And Li F*[*]

$$F^* = \frac{\pi_n - \frac{n-1}{n}\eta_s}{\sqrt{u_{F^*}\eta + v_{F^*}\eta^2}}$$

$$v_{F^*} = [d_n + \frac{2(n^2+n+3)}{9n(n-1)} - 2\frac{1}{n-1}(4b_n - 6 + \frac{8}{n})]/(a_n^2 + b_n)$$

$$u_{F^*} = [\frac{n}{n-1} + \frac{n+1}{3(n-1)} - 2\frac{2}{n(n-1)} + 2\frac{n+1}{(n-1)^2}(a_{n+1} - \frac{2n}{n+1})]/a_n - v_{F^*}$$

The definition of a_n, b_n , d_n , η and η_s can be seen in the equations of *Fu And Li D*[*].

For each gene of rhesus macaque and its ortholog in human, we estimated the polymorphism levels (θw and π) for different genomic regions (non-synonymous sites, synonymous sites, CDS, UTR, exon, intron) using DnaSP (v5.10) and customized scripts. For McDonald–Kreitman test, codon-based alignment between human and rhesus macaque for each gene was constructed to classify the divergent sites between the two species, as well as polymorphic sites in rhesus macaque, into synonymous and non-synonymous changes. The number of synonymous and non-synonymous changes was counted as described in R package (Popgenome). The McDonald–Kreitman test was then performed using Perl package (Bio::PopGen::Statistics). All related scripts were uploaded to GitHub (https://github.com/rhesusbase/PopGateway/).

Development of RhesusBase PopGateway

RhesusBase PopGateway was developed using Apache-based web development technologies, with the meta-data deposited following a MySQL relation schema. The PopGateway includes three major components. First, the Population Genetics Page was developed to provide gene-based annotations of population genetics annotations, in a comparative mode of human and rhesus macaque. Second, the RhesusBase Genome Browser was updated with 24 new tracks to display region-based population genetics annotations. Third, a mobile APP for iPhone was developed with Object-C to support offline retrieval of RhesusBase annotations. We further performed RhesusBase annual update on transcriptome and regulatory annotations, with new

NGS datasets processed according to previous pipeline (Zhang et al. 2013; Zhang et al. 2014).

Identification of human-specific gene loss, and genomic regions under balancing selection

A list of human-specific gene loss events was collected from Wang *et al* (Wang, et al. 2006) **(supplementary table S3)**. The macaque orthologs of these human pseudogene were then retrieved using LiftOver with manual inspections, with the coding regions defined by the longest putative ORF. The ratios of the nucleotide diversity between non-synonymous sites to synonymous sites were then estimated for each coding region, as well as the pseudogene in human, in which the pseudo-non-synonymous and pseudo-synonymous sites were determined by codon-level alignment with proteins in rhesus macaque. The polymorphism levels were estimated by DnaSP (v5.10) and in-house scripts. Wilcoxon test was performed to test whether the nucleotide diversity between the two groups are significantly different (*p-value* cutoff of 0.05).

For *MAMU-DQA1* haplotypes, 20 single polymorphism sites with considerable frequency (minor allele in at least three individuals) were selected in the hierarchical clustering analyses. The hierarchical clustering chart of 62 haplotypes was constructed using R package "pheatmap". The reduced-median network was then built using NETWORK 4.6 to infer the evolutionary relationship of these haplotypes (Bandelt et al. 1995). Pairwise mismatch distribution of these haplotypes were further calculated by counting the number of differences between all pairs of the 62 haplotypes (Rogers and Harpending 1992). For regions under balance selection in human and chimpanzee (Leffler et al. 2013), the orthologous regions in rhesus macaque were identified using LiftOver with default parameters. The raggedness value for the candidate regions were then calculated using ARLECORE (v3.5.2.1) (Excoffier and Lischer 2010). Regions with raggedness value >0.03 were defined as macaque regions under balancing selection (Jobling et al. 2013).

Reference

Bandelt HJ, Forster P, Sykes BC, Richards MB. 1995. Mitochondrial portraits of human populations using median networks. Genetics 141: 743-753.
Chen JY, Shen QS, Zhou WZ, Peng J, He BZ, Li Y, Liu CJ, Luan X, Ding W, Li S, Chen C, Tan BC, Zhang YE, He A, Li CY. 2015. Emergence, Retention and Selection: A Trilogy of Origination for Functional De Novo Proteins from Ancestral LncRNAs in Primates. PLoS Genet 11: e1005391.
Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6: 80-92.
Delaneau O, Howie B, Cox AJ, Zagury JF, Marchini J. 2013. Haplotype estimation using sequencing reads. Am J Hum Genet 93: 687-696.
DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43: 491-498.
Excoffier L, Lischer HE. 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Molecular ecology resources 10: 564-567.
Jobling M, Hurles M, Tyler-Smith C. 2013. Human evolutionary genetics: origins, peoples & disease: Garland Science.
Leffler EM, Gao Z, Pfeifer S, Segurel L, Auton A, Venn O, Bowden R, Bontrop R, Wall JD, Sella G, Donnelly P, McVean G, Przeworski M. 2013. Multiple instances of ancient balancing selection shared between humans and chimpanzees. Science 339: 1578-1582.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754-1760.
McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20: 1297-1303.
Paten B, Herrero J, Fitzgerald S, Beal K, Flicek P, Holmes I, Birney E. 2008. Genome-wide nucleotide-level mammalian ancestor reconstruction. Genome Res 18: 1829-1843.
Rogers AR, Harpending H. 1992. Population growth makes waves in the distribution of pairwise genetic differences. Mol Biol Evol 9: 552-569.
Yang XZ, Chen JY, Liu CJ, Peng J, Wee YR, Han X, Wang C, Zhong X, Shen QS, Liu H, Cao H, Chen XW, Tan BC, Li CY. 2015. Selectively Constrained RNA Editing Regulation Crosstalks with piRNA Biogenesis in Primates. Mol Biol Evol.
Zhang SJ, Liu CJ, Shi M, Kong L, Chen JY, Zhou WZ, Zhu X, Yu P, Wang J, Yang X, Hou N, Ye Z, Zhang R, Xiao R, Zhang X, Li CY. 2013. RhesusBase: a knowledgebase for the monkey research community. Nucleic Acids Res 41: D892-905.
Zhang SJ, Liu CJ, Yu P, Zhong X, Chen JY, Yang X, Peng J, Yan S, Wang C, Zhu X, Xiong J, Zhang YE, Tan BC, Li CY. 2014. Evolutionary interrogation of human biology in well-annotated genomic framework of rhesus macaque. Mol Biol Evol 31: 1309-1324.