Supplementary file 4 to:

Whole genome sequencing of Tibetan macaque (*Macaca thibetana*) provides new insight into the macaque evolutionary history

Zhenxin Fan[1,#], Guang Zhao[2,#], Peng Li[1], Naoki Osada[3], Jinchuan Xing[4], Yong Yi[5], Lianming Du[1], Pedro Silva[6], Hongxing Wang[5], Ryuichi Sakate[7], Xiuyue Zhang[1], Huailiang Xu[8], Bisong Yue[2,*], Jing Li[1,*]

[1] *Key Laboratory of Bioresources and Ecoenvironment (Ministry of Education), College of Life Sciences, Sichuan University, Chengdu 610064, People's Republic of China*
[2] *Sichuan Key Laboratory of Conservation Biology on Endangered Wildlife, College of Life Sciences, Sichuan University, Chengdu, 610064, People's Republic of China*
[3] *Division of Evolutionary Genetics, Department of Population Genetics, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan*
[4] *Department of Genetics, Rutgers, the State University of New Jersey, Piscataway, New Jersey 08854, USA*
[5] *Experimental Animal Institute of Sichuan Academy of Medical Sciences & Sichuan Provincial People's Hospital, Chengdu 610212, People's Republic of China*
[6] *Research Center in Biodiversity and Genetic Resources, University of Porto (CIBIO-UP), Campus Agrário de Vairão, 4485-661 Vila do Conde, Portugal*
[7] *Laboratory of Rare Disease Biospecimen, Department of Disease Bioresources Research, National Institute of Biomedical Innovation, 7-6-8 Saito-asagi, Ibaraki, Osaka 567-0085, Japan*
[8] *College of Animal Science and Technology, Sichuan Agricultural University, Ya'an 625014, People's Republic of China*
[#] *Contributed equally*

Email: Zhenxin Fan: zxfan068@gmail.com; Guang Zhao: 910882735@qq.com; Peng Li: 574694862@qq.com; Naoki Osada: nosada@nig.ac.jp; Jinchuan Xing: xing@dls.rutgers.edu; Yong Yi: 524019163@qq.com; Lianming Du: adu220@126.com; Pedro Silva: p.manuel.silva@gmail.com; Hongxing Wang: whx6633@163.com; Ryuichi Sakate: rsakate@nibio.go.jp; Xiuyue Zhang: zhangxy317@126.com; Huailiang Xu: huailxu@yahoo.com; Bisong Yue: bsyue@scu.edu.cn; Jing Li: ljtjf@126.com

***Corresponding author**

J. Li (ljtjf@126.com) and B. Yue (bsyue@scu.edu.cn)

Sample identification

Prior to genome sequencing, two Tibetan macaque (TM) and one Stump-tailed macaque (*M. arctoides*; SM) specific *Alu* locus (table S18) (Li et al. 2009), and 896 bp of mitochondrial DNA fragment of NADH 4 and 5 (primers: 5'-CCTTGTAATCGTAGCCATCCTC-3'/5'-TGGGATGAATGTTATGGAGAAG-3'; fig. S7) were amplified to confirm the identification, since TM and SM were morphologically very similar.

Figure S7 showed that two TM specific *Alu* sites were present in TM whereas they were absent in SM. The SM specific *Alu* site was present in SM but was absent in TM. The neighbor-joining tree based on mitochondrial NADH 4 and 5 fragment showed that TM03 was clustered with another Tibetan macaque sample (fig. S8).

Table S18 *Alu* sites information for amplifying Tibetan macaque and Stump-tailed macaque

***Alu* sites**	**Primers (5'-3')**	**Note**
TM-Yb3-mb-23	TGGCCAGTTTGCTTATAAAGGT TCTACCAGGTTCTTTGCCAGAT	Tibetan macaque specific
TM-JH-24	TGTGGCCTTAAGTTTCAGCTCT CTTGGCCTCTAAGAGTCAGCAG	Tibetan macaque specific
Mab-B-23	TGAGTCTCTTTCCCATCCATCT TTCTAAAACAACCCCAAAAGGA	Stump-tailed macaque specific

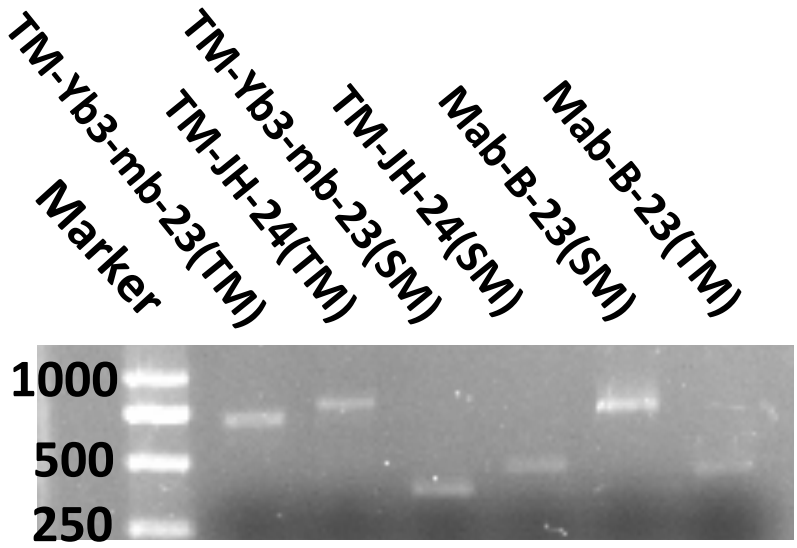

Figure S7 Agarose gel electrophoresis of PCR products of Tibetan macaque and Stump-tailed macaque specific *Alu* sites. *Alu* sites names and sample names are shown. TM: Tibetan macaque; SM: Stump-tailed macaque

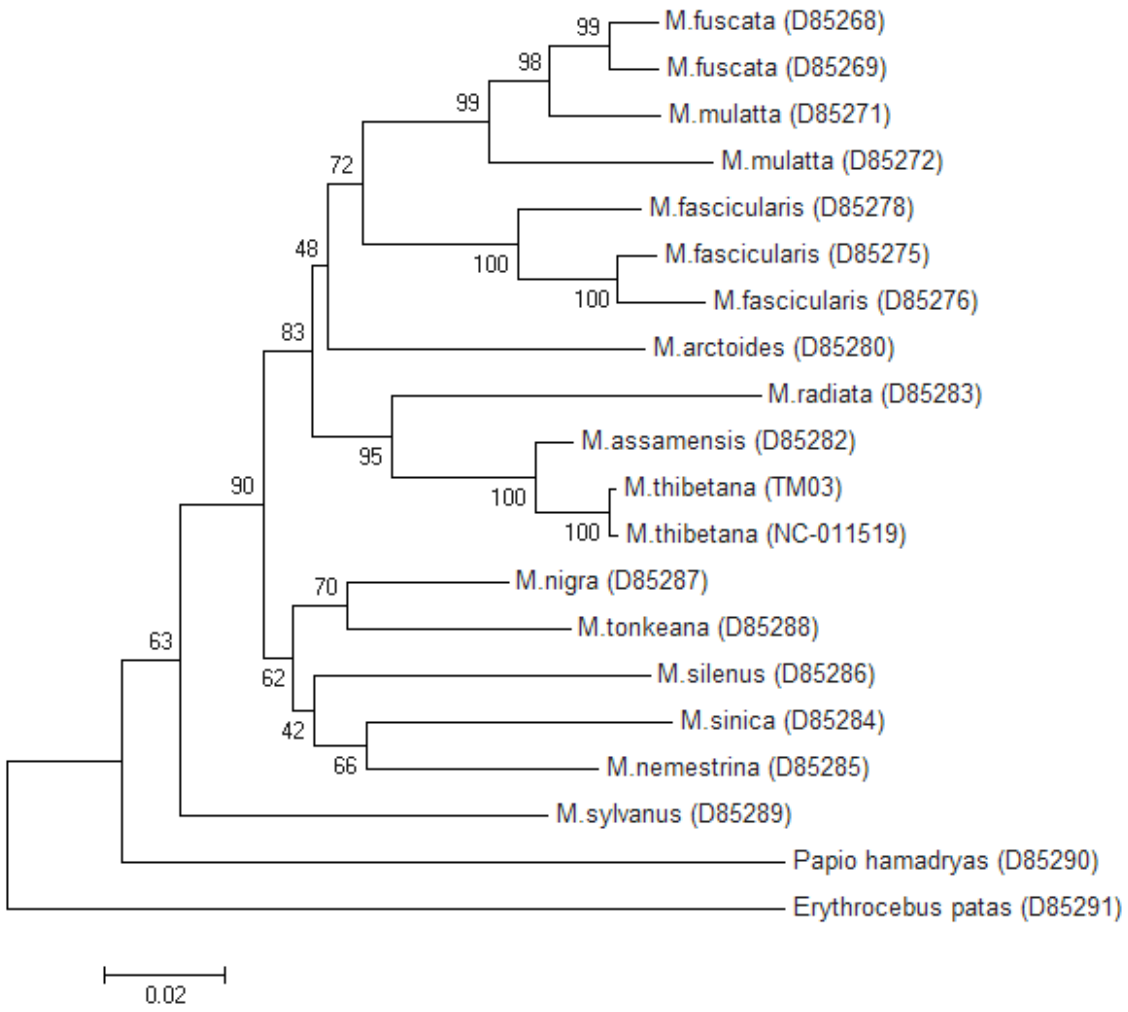

Figure S8 Neighbor joining tree based on the mitochondrial NADH 4 and 5 fragments of Tibetan macaque and other macaques. The individuals used to perform the whole genome sequencing (TM03) was clustered with another Tibetan macaque (NCBI ID: NC_011519)

Reference

Li J, Han K, Xing J, Kim HS, Rogers J, Ryder OA, Disotell T, Yue BS, Batzera MA. 2009. Phylogeny of the macaques (Cercopithecidae: *Macaca*) based on *Alu* elements. *Gene* 448:242–249.

Genotyping pipeline

The overview of our genotyping pipeline was showed in fig.1a at main text. Here, we described the details of several important steps.

S1. Local realignment

The false SNPs could be detected in regions where repeated alignment errors occur across overlapping reads because short read alignment algorithms work on each read independently. In order to reduce the false positive SNPs, we used GATK IndelRealigner (DePristo et al. 2011) to perform local multiple alignment. In total, there were three steps. First, identify suspicious intervals that may require realignment, then performing local realignment within these intervals, and then fix the mate pairing lost during the local realignment process.

Example command lines were:

a. Interval detection

java -Xmx9g -jar GenomeAnalysisTK.jar -T RealignerTargetCreator --read_filter BadCigar -R reference_genome.fa -I input_file.bam -o output.intervals

b. Local realignment

java -Xmx9g –jar GenomeAnalysisTK.jar -T IndelRealigner –R reference_genome.fa -I input_file.bam -targetIntervals output.intervals -o output_realign.bam

c. Fix the mate pair information

java -jar -Xmx9g FixMateInformation.jar INPUT=output_realign.bam OUTPUT=output_realign_fixed.bam SORT_ORDER=coordinate VALIDATION_STRINGENCY=LENIENT

S2. Base quality recalibration

Quality scores assigned to individual base calls only reflected confidence in the specified nucleotide, but they were may be weakly correlated with the actual probabilities of erroneous base calls (Freedman et al. 2014). Therefore, it is necessary to standardize quality scores across sequencing runs and libraries. Here, we performed empirical quality score recalibration using GATK. Three steps were involved: 1) since there was no dbSNP data set for Tibetan macaque, we genotype in the same way as below (see **S3. SNP and Indel calling**) to liberally define a SNP dataset that are excluded in the subsequent steps; 2) for the rest sites, creating the table for the frequency of base calls that are correct v.s. incorrect as a function of covariates reflecting features of the underlying sequence context stratified by library/sequencing run; 3) using the genome-wide empirical error rates conditional on each unique covariate set to replace the original quality scores.

Example command lines were:

a. Create recalibration table

java -jar GenomeAnalysisTK.jar -R reference_genome.fa -T CountCovariates -l INFO -cov ReadGroupCovariate -cov CycleCovariate -cov DinucCovariate -cov QualityScoreCovariate --default_platform Illumina -I input_file.bam --knownSites:VCF Recalibration_Input.vcf -recalFile output_retable.csv --solid_recal_mode SET_Q_ZERO --solid_nocall_strategy LEAVE_READ_UNRECALIBRATED

b. Generate recalibrated bam files

java -jar -Xmx9g GenomeAnalysisTK.jar -R reference_genome.fa -l INFO -T TableRecalibration --default_platform Illumina -I input_file.bam -o Recal_output.bam -recalFile output_retable.csv --doNotWriteOriginalQuals --solid_recal_mode SET_Q_ZERO --solid_nocall_strategy PURGE_READ

S3. SNP and Indel calling

We used the GATK Unified Genotyper (DePristo et al. 2011) to call genotypes for all the samples. Our main goal of this study is the evolutionary history of Tibetan macaque rather than population genomics, thus we here genotyped each sample separately.

Because several different conservative post-genotyping filters were applied later, we set both standard minimum confidence thresholds to zero here.

Example command line was:

java -jar -Xmx10g GenomeAnalysisTK.jar -R reference_genome.fa -T UnifiedGenotyper -l INFO --genotyping_mode DISCOVERY --output_mode EMIT_ALL_CONFIDENT_SITES -I input_file.bam --min_base_quality_score 20 --standard_min_confidence_threshold_for_emitting 0.0 --standard_min_confidence_threshold_for_calling 0.0 -A GCContent -o output.vcf -metrics output.metrics -dt NONE

Since we only had short insert size libraries, we only genotyped the small indels within GATK.

Example command line was:

java -jar -Xmx8g GenomeAnalysisTK.jar -R reference_genome.fa -T UnifiedGenotyper -l INFO --genotyping_mode DISCOVERY --output_mode EMIT_ALL_CONFIDENT_SITES -I input_file.bam -glm INDEL --indel_heterozygosity 0.000125 --min_indel_count_for_genotyping 5 --min_base_quality_score 20 --standard_min_confidence_threshold_for_emitting 0.0 --standard_min_confidence_threshold_for_calling 0.0 -A GCContent -o output.vcf -metrics output.metrics -dt NONE

Reference

DePristo MA, Banks E, Poplin R, et al. (13 co-authors). 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet.* 43:491-498.

Freedman AH, Gronau I, Schweizer RM, et al. (30 co-authors) 2014. Genome Sequencing Highlights the Dynamic Early History of Dogs. *PLoS Genet.* 10: e1004016.

Post-genotype filters

After having the genotype calls from GATK, we applied several conservative data quality filters to control the data quality again. There were two levels of filters: genome filters (GF), which are based on the reference genome's features and polymorphism across samples and sample filters (SF), which are based on the genotype calls of each sample. We described the details of the filters below.

S1. Genome filters

S1.1. Triallelic sites (MA). Triallelic sites were more prone to genotyping errors (Freedman et al. 2014). Besides, such site only contains a small fraction of the genome (0.01%) in Tibetan macaque. Thus we filter out all the triallelic sites.

S1.2. Copy Number Variants (CNV). Misalignment is possible when short reads mapped to the places of reference where contain novel CNVs. It can lead to false positive SNPs. To minimize this type of misalignment, we applied a set of CNV regions to filter out from downstream analyses. Since we did not detect CNVs in this study by ourselves, we used previously discovered CNVs reported in reference genome (Lee et al. 2008; Gokcumen et al. 2011). We calculated the SNP rate within and outside the CNVs. Within CNVs, the SNP rate (raw rate, without other filters) was 0.01, whereas the SNP rate was 0.006 outside CNVs.

S1.3. CpG. Mutation rate at CpG sites is higher than non-CpG sites (Hodgkinson and Eyre-Walker 2011), so that regions enriched for CpGs may display elevated diversity and/or divergence leading to outliers in window-based analyses. We flagged any sites that even one of the samples fell within a CpG dinucleotide.

S2. Sample filters

S2.1. Proximity to Indel (DL). Short reads are prone to misalignment near indels (Freedman et al. 2014) and the local realignment around indels in our genotyping pipeline

may not fully fix this problem. Therefore, to minimize the potential source of bias, for each sample we excluded any SNPs near indels (5bp, either up or downstream).

S2.2. Genotype Quality (GQ). Genotype quality is the phred-scaled probabilities ($10*\log_{10}$(P[error]).), which represent the genotype calls do not match the true genotype. Hard genotype quality thresholds work well with high coverage (>20 X), although it may cause underestimate of heterozygotes in low or moderate coverage genomes (Nielsen et al. 2011). All the samples in our study, except CR2, were sequenced at >20 X. Moreover, the distribution of genotype quality showed that large proportion of SNP sites have GQ > 20 (TM: 95.88%; CR1: 97.03%; CR2: 75.38%; CE1: 96.28%; CE2: 88.53%). Therefore, we chose a hard minimum GQ threshold of 20 (P[error]=0.01).

S2.3. Depth of Coverage (DP).

S2.3.1. Excess Depth of Coverage for all sites. Extremely high depth of coverage relative to the genome-wide average likely indicates misalignment of reads generated from paralogous positions in the genome. Indeed, excess depth of coverage is a typical metric used to define CNV regions, but CNV filtering alone will fail to detect finer-resolution CNV signatures (Freedman et al. 2014). Therefore, we conservatively filtered all sites if their depth of coverage exceeded twice the mean depth of coverage of each sample.

S2.3.2. Minimum Depth of Coverage for non-variant sites. Since only the very old version of GATK gave the GQ value for non-variant sites and the version we are using does not, thus we did not have GQ filters for non-variant sites. Instead, we used minimum depth of coverage as one of the filters for non-variant sites. Here, we set the minimum threshold as eight.

S2.4. Clustered SNPs (DV). Within any sample, we excluded all SNPs that within 5 bp of another SNP.

S2.5. Mapping quality (MQ). MQ is related to "uniqueness." The bigger the gap between the best alignment's score and the second-best alignment's score, the more unique the best alignment, and the higher its mapping quality should be (Li et al. 2008). The recent bonobo genome paper used 30 (P[error]=0.001) as threshold (Prüfer et al.

2012), thus here we also applied MQ >= 30 as one of the sample filters for all the samples.

S3. Combination of filters

For different types of analyses, we used different combination of GF and SF filters. For analyses involving estimation of genome-wide patterns of diversity, we used GF1 and SF. For quantifying the number of Tibetan macaque specific variants, and for analysis of functional regions, we used GF2 and SF.

The combination of the filters were:

S3.1. Non-variant sites:

GF1: CNV, CpG

GF2: CNV

SF: DP >= 8, DP <= (2 times mean coverage), MQ

S3.2. SNV sites:

GF1: MA, CNV, CpG

GF2: MA, CNV

SF: GQ >= 20, DP <= (2 times mean coverage), DL, DV, MQ

Reference

Freedman AH, Gronau I, Schweizer RM, et al. (30 co-authors) 2014. Genome Sequencing Highlights the Dynamic Early History of Dogs. *PLoS Genet.* 10: e1004016.

Gokcumen O, Babb PL, Iskow RC, et al. (11 co-authors). 2011. Refinement of primate copy number variation hotspots identifies candidate genomic regions evolving under positive selection. *Genome Biology* 12:R52.

Hodgkinson A, Eyre-Walker A. 2011. Variation in the mutation rate across mammalian genomes. *Nat Rev Genet.* 12:756-766.

Lee AS, Gutiérrez-Arcelus M, Perry GH, Vallender EJ, Johnson WE, Miller GM, Korbel JO, Lee C. 2008. Analysis of copy number variation in the rhesus macaque genome identifies candidate loci for evolutionary and human disease studies. *Hum Mol Genet.* 17:1127-1136.

Li H, Ruan J, Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* 18:1851-8.

Nielsen R, Paul JS, Albrechtsen A, Song YS. 2011. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet.* 12:443-451.

Prüfer K, Munch K, Hellmann I, et al. (41 co-authors). 2012. The bonobo genome compared with the chimpanzee and human genomes. *Nature* 486:527-531.

PCR re-sequencing of some SNVs

We performed PCR re-sequencing of 14 genes in 18 unrelated Tibetan macaque (TM) individuals from Sichuan province to check the allele frequency of 29 TM specific homozygous non-synonymous variants (TSHNVs). Individuals were chosen to represent Tibetan macaque populations from Mabian, Emei, Ganluo and Pengzhou counties in Sichuan Province, China. Based on the records, none of the individuals shared grandparents (fig. S9). The detail information of genes and TSHNVs, and the summary of the PCR re-sequencing result can be found in table S15 and table S16 (Supplementary file 3), respectively. The information of primers we used to amplify the fragments of the 14 genes was listed in table S17 (Supplementary file 3). The details of three genes were discussed in main text, and the rest genes were discussed below.

Nucleotide-binding oligomerization domain containing 2 (NOD2) gene and Excision repair cross-complementing rodent repair deficiency, complementation group 1 (ERCC1) gene were only within some eye problems categories in the GO result (e.g. NOD2: inflammatory abnormality of the eye and abnormality of the choroid; ERCC1: Aplasia/Hypoplasia affecting the eye; Supplementary file 3, table S15). NOD2 gene was a member of the Nod1/Apaf-1 family and encoded a protein with two caspase recruitment domains and six leucine-rich repeats, which played a role in the immune response to intracellular bacterial lipopolysaccharides. Mutations in NOD2 were associated with inflammatory bowel disease 1 (Hugot et al. 2001), Blau syndrome (Miceli-Richard et al. 2001) and early-onset Sarcoidosis (Kanazawa et al. 2005). Knockout mice exhibited abnormal immune system morphology and physiology and increased susceptibility to induced colitis (Supplementary file 3, table S15). ERCC1 gene encoded a protein involved in nucleotide excision repair and interstrand crosslink repair of DNA. It interacts with ERCC4 to form an endonuclease that excises the DNA for subsequent repair, thus ERCC1 is required for stabilizing and enhancing ERCC4 activity (Gregg et al. 2011; Kashiyama et al. 2013). Mutations in ERCC1 may cause cerebrooculofacioskeletal syndrome 4 in humans (Jaspers et al. 2007). In TM, we found two TSHNVs and one

TSHNV at NOD2 and ERCC1 gene, respectively. PCR re-sequencing confirmed that all the three TSHNVs were completely fixed in TM (fig. 5; Supplementary file 3, table S16).

Class II, major histocompatibility complex, transactivator (CIITA) gene encoded a protein with an acidic transcriptional activation domain, 4 leucine-rich repeats and a GTP binding domain, which acted as a positive regulator of class II major histocompatibility complex gene transcription (Harton and Ting, 2000). Mutations in this gene have been reported to be associated with bare lymphocyte syndrome type II and increased susceptibility to rheumatoid arthritis (Bontron et al. 1997). Knockout mice were immunologically abnormal with extremely little MHC class II expression, and greatly reduced serum IgG (Supplementary file 3, table S15). In TM, there were five non-synonymous variants but only one of them was TSHNV. PCR re-sequencing in 14 individuals found that four of them were heterozygous, whereas the rest samples are all fixed (allele frequency: 85.71%; Supplementary file 3, table S16).

Complement component 4 binding protein, alpha (C4BPA) gene encoded a protein belonged to a superfamily composed predominantly of tandemly arrayed short consensus repeats of approximately 60 amino acids. Previous study found a mutation (Arg240His) in C4BPA gene that was associated with atypical hemolytic uremic syndrome (Blom et al. 2008). In TM, there were six non-synonymous variants. Two TSHNVs were identified, and PCR re-sequencing found that one of them was completely fixed in 17 TM, whereas another one (chr01: 163228962) had low allele frequency (52.78%; Supplementary file 3, table S16).

Cyclic nucleotide gated channel beta 1 (CNGB1) gene was within both diabetes mellitus (HP:0000819) and eye problems categories in our GO result (Supplementary file 1, table S10). CNGB1 gene encoded a protein that formed the heterotetrameric rod photoreceptor cyclic nucleotide-gated (CNG) channel. It has been reported that the defects and mutations in CNGB1 can cause Retinitis pigmentosa type 45 in humans (Bareil et al. 2001). In CNGB1, there were four non-synonymous variants and one of them was TSHNV. The re-sequencing showed that only two individuals were heterozygous whereas the rest 12 samples all carried the fixed allele (allele frequency: 92.86%; Supplementary file 3, table S16).

IGF1R binds insulin-like growth factor, which plays a critical role in transformation events. It has been described that the defects in IGF1R gene would cause postnatal growth failure (Abuzzahab et al. 2003). In TM, two non-synonymous variants were found and one of them was TSHNV. Sanger sequences exhibited this TSHNV completely fixed in 14 TM (fig. 5; Supplementary file 3, table S16).

Mediterranean fever (MEFV) gene encoded a protein that was an important modulator of innate immunity. Mutations in MEFV gene were associated with Mediterranean fever, a hereditary periodic fever syndrome (International FMF Consortium 1997). Mouse knockout studies showed that homozygous null mice develop normally but show increased susceptibility to infection (Supplementary file 3, table S15). However, the two TSHNVs at MEFV gene were completely fixed (fig. 5; Supplementary file 3, table S16).

Membrane-spanning 4-domains, subfamily A, member 1 (MS4A1) gene encoded a member of the membrane-spanning 4A (MS4A) family that shared structural similarity and amino acid sequence homology. This gene encoded a B-lymphocyte surface molecule, which played a role in the development and differentiation of B-cells into plasma cells. A girl with common variable immunodeficiency-5 was identified with a homozygous mutation at intron 5 of the MS4A1 gene (Kuijpers et al. 2010). In TM, two non-synonymous variants were observed, and one of them was TSHNV. PCR re-sequencing showed that this TSHNV was completely fixed in all the individuals (fig. 5; Supplementary file 3, table S16).

NOP10 ribonucleoprotein (NOP10) gene is a member of the H/ACA snoRNPs (small nucleolar ribonucleoproteins) gene family, which appeared in the diabetes mellitus (HP:0000819) in our GO result (Supplementary file 3, table S15). The mutations in NOP10 were associated with autosomal recessive dyskeratosis congenita-1 (DKCB1) in human (Walne et al. 2007). In TM, there were three non-synonymous variants in NOP10 gene, and one variant belonged to TSHNV. Sanger sequences of 18 TM individuals showed that the TSHNV was completely fixed (fig. 5; Supplementary file 3, table S16).

Toll-like receptor 4 (TLR4) gene was a member of the Toll-like receptor (TLR) family, which played a fundamental role in pathogen recognition and activation of innate

immunity. Macular degeneration (Zareparsi et al. 2005) and susceptibility to colorectal cancer (BoraskaJelavic et al. 2006) were associated with TLR4. The genes in TLR family were highly conserved from Drosophila to humans (http://www.genecards.org/].). However, there were 12 variants at TLR4 in TM, and six non-synonymous variants were TSHNVs. PCR re-sequencing data found that four TSHNVs were completely fixed, and the remaining two still had very high allele frequency (fig. 5; Supplementary file 3, table S16).

TREM2 gene encoded a membrane protein that formed a receptor-signaling complex with the TYRO protein tyrosine kinase binding protein, which was important in immune response. Defects in TREM2 were a cause of polycystic lipomembranous osteodysplasia with sclerosing leukoencephalopathy (PLOSL) in human (Klünemann et al. 2005). One TSHNV was observed at TREM2 in TM, and it was completely fixed in all individuals (fig. 5; Supplementary file 3, table S16).

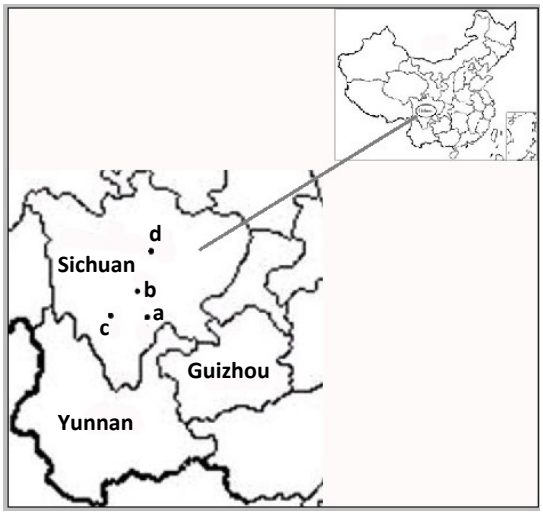

Figure S9 Sampling sites of Tibetan macaque. a: Mabian county; b: Emei city; c: Ganluo county; d: Pengzhou city. The samples TM03, TM05, and TM09 to TM18 were from Mabian county (three different groups). TM01 and TM 02 were from Ganluo county. TM04 and TM06 were from Emei city. TM 07 and TM 08 were from Pengzhou city.

Reference

Abuzzahab MJ, Schneider A, Goddard A, et al. (15 co-authors). 2003. IGF-I receptor mutations resulting in intrauterine and postnatal growth retardation. *New Engl J Med.* 349:2211-2222.

Bareil C, Hamel CP, Delague V, Arnaud B, Demaille J, Claustres M. 2001. Segregation of a mutation in CNGB1 encoding the beta-subunit of the rod cGMP-gated channel in a family with autosomal recessive retinitis pigmentosa. *Hum Genet.* 108:328-334.

Blom AM, Bergström F, Edey M, et al. (15 co-authors). 2008. A novel non-synonymous polymorphism (p.arg240his) in C4b-binding protein is associated with atypical hemolytic uremic syndrome and leads to impaired alternative pathway cofactor activity. *J Immunol.* 180:6385-6391.

Bontron S, Steimle V, Ucla C, Eibl MM, Mach B. 1997. Two novel mutations in the MHC class II transactivator CIITA in a second patient from MHC class II deficiency complementation group A. *Hum Genet.* 99:541-546.

BoraskaJelavic T, Barisic M, DrmicHofman I, Boraska V, Vrdoljak E, Peruzovic M, Hozo I, Puljiz Z, Terjic J. 2006. Microsatelite GT polymorphism in the toll-like receptor 2 is associated with colorectal cancer. *Clin Genet.* 70:156-160.

Gregg SQ, Robinson AR, Niedernhofer LJ. 2011. Physiological consequences of defects in ERCC1-XPF DNA repair endonuclease. *DNA Repair* 10:781-791.

Harton JA, Ting JPY. 2000. Class II transactivator: mastering the art of major histocompatibility complex expression. *Mol. Cell Biol.* 20:6185-6194.

Hugot JP, Chamaillard M, Zouali H, et al. (20 co-authors). 2001. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* 411:599-603.

International FMF Consortium. 1997. Ancient missense mutations in a new member of the RoRet gene family are likely to cause familial Mediterranean fever. *Cell* 90:797-807.

Jaspers NG, Raams A, Silengo MC, et al. (11 co-authors). 2007. First reported patient with human ERCC1 deficiency has cerebro-oculo-facio-skeletal syndrome with a mild defect in nucleotide excision repair and severe developmental failure. *Am J*

Hum Genet. 80:457-466.

Kanazawa N, Okafuji I, Kambe N, et al. (19 co-authors). 2005. Early-onset sarcoidosis and CARD15 mutations with constitutive nuclear factor-kappa-B activation: common genetic etiology with Blau syndrome. *Blood* 105:1195-1197.

Kashiyama K, Nakazawa Y, Pilz DT, et al. (26 co-authors). 2013. Malfunction of nuclease ERCC1-XPF results in diverse clinical manifestations and causes Cockayne syndrome, xeroderma pigmentosum, and Fanconi anemia. *Am J Hum Genet.* 92:807-819.

Klünemann HH, Ridha BH, Magy L, et al. (12 co-authors). 2005. The genetic causes of basal ganglia calcification, dementia, and bone cysts: DAP12 and TREM2. *Neurology* 64:1502-1507.

Kuijpers TW, Bende RJ, Baars PA, et al. (11 co-authors). 2010. CD20 deficiency in humans results in impaired T cell-independent antibody responses. *J Clin Investig.* 120:214-222.

Miceli-Richard C, Lesage S, Rybojad M, Prieur AM, Manouvrier-Hanu S, Häfner R, Chamaillard M, Zouali H, Thomas G, Hugot JP. 2001. CARD15 mutations in Blau syndrome. *Nat Genet.* 29:19-20.

Walne AJ, Vulliamy T, Marrone A, Beswick R, Kirwan M, Masunari Y, Al-Qurashi FH, Aljurf M, Dokal I. 2007. Genetic heterogeneity in autosomal recessive dyskeratosis congenita with one subtype due to mutations in the telomerase-associated protein NOP10. *Hum Mol Genet.* 16:1619-1629.

Zareparsi S, Buraczynska M, Branham KE, et al. (15 co-authors). 2005. Toll-like receptor 4 variant D299G is associated with susceptibility to age-related macular degeneration. *Hum Mol. Genet.*14:1449-1455.

Contiguous putative introgression regions (PIRs)

We checked whether there were some contiguous PIRs in the 100kb window test (with *P*-value 0.01). It turned out that only 20 contiguous PIRs were found in 13 chromosomes (chr01:91900000-92100000, chr01:141600000-141800000, chr02:85400000-85600000, chr02:96500000-96700000, chr04:28900000-29100000, chr05:34300000-34500000, chr05:38800000-39000000, chr06:47500000-47700000, chr09:5300000-5500000, chr09:37300000-37500000, chr11:11300000-11700000, chr11:35900000-36200000, chr13:131300000-131500000, chr14:36000000-36200000, chr17:35500000-35700000, chr18:1900000-2100000, chr18:24200000-24400000, chr19:32300000-32700000, chr19:33000000-33700000, chr20:30700000-31200000). Most of them (15 PIRs) are only 200kb. The longest one is 700kb (chr19: 33000000- 33700000). We then checked whether there were genes within them. Only one PIR in the chr02 (chr02:85400000-85600000) contains one full gene (LOC701151). But it is a predicted gene in rhesus macaque. There are other six genes within the PIRs, but all of them are only part of the gene within the PIRs, and half of them are also predicted genes (chr02:96500000-96700000: LOC717173; chr05:34300000-34500000: RFC1; chr09:5300000-5500000: AKR1C1; chr09:37300000-37500000: LOC700258; chr18:1900000-2100000: LOC722137; chr18:24200000-24400000: DSG3).