nature communications

Article

https://doi.org/10.1038/s41467-025-63747-x

Forward and reverse genomic screens enhance the understanding of phenotypic variation in a large Chinese rhesus macaque cohort

Received: 31 July 2024

Accepted: 27 August 2025

Published online: 30 September 2025

Check for updates

Bao-Lin Zhang[1,2,3,12], Yongxuan Chen[1,4,12], Yali Zhang[2,12], Yicheng Qiao[4,5], Yang Wu[6], Yi Zhang[1,2], Yizheng Lu[1,4], Xinran You[1], Yanling Li[2], Hong-Di Huang[2,4], Qiong Wang[2,7], Yijiang Li[2,7], Yun Wang[2,7], Wenxian Xiao[2,7], Hexian Duan[2,7], Ming-Hao Qiu[2], Nan-Hui Chen[2], Xiaomei Yu[2], Min-Min Yang[1], Longbao Lv[2,4,7], David N. Cooper[8], Ping Zheng[1,2,9,10], Yong-Gang Yao[1,2,4,9,10] ✉, Ning Liu[4,5] ✉, Jian-Hong Wang[2,9] ✉ & Dong-Dong Wu[1,2,10,11] ✉

Combining genotype and phenotype data promises to greatly increase the value of macaque as biomedical models for human disease. Here we launch the Macaque Biobank project by deeply sequencing 919 captive Chinese rhesus macaques (CRM) while assessing 52 phenotypic traits. Genomic analyses reveal the captive CRMs are a mixture of multiple wild sources and exhibit significantly lower mutational load than their Indian counterparts. We identify hundreds of loss-of-function variants linked to human inherited disease and drug targets, and at least seven exert significant effects on phenotypes using forward genomic screens. Genome-wide association analyses reveal 30 independent loci associated with phenotypic variations. Using reverse genomic approaches, we identify *DISC1* (p.Arg517Trp) as a genetic risk factor for neuropsychiatric disorders, with macaques carrying this deleterious allele exhibiting impairments in working memory and cortical architecture. This study demonstrates the potential of macaque cohorts for the investigation of genotype-phenotype relationships and exploring potential spontaneous models of human genetic disease.

Over the past decades, rhesus macaque (*Macaca mulatta*) bioresources have played a crucial role in deepening our understanding of human physiology, metabolism, reproduction, development, cognition, and pathology[1–3]. More recently, the importance of this species as an experimental model increased substantially during the COVID-19 pandemic, a dire public health crisis that urgently necessitated the recruitment of many animal models for vaccine testing and drug treatments[4]. However, this global pandemic also triggered, either directly or indirectly, a worldwide shortage of rhesus macaques for research[5,6]. In consequence, fully appreciating and efficiently utilizing macaque bioresources has become a major challenge currently faced by all biologists[7,8].

Effectively utilizing rhesus macaques as an experimental animal model benefits from the greater resolution of genetic variation and detailed phenotypic examination in parallel[9,10]. Additionally, insights into the genetic diversity of macaque populations will greatly assist in

A full list of affiliations appears at the end of the paper. ✉e-mail: yaoyg@mail.kiz.ac.cn; liuning@ibp.ac.cn; wangjh@mail.kiz.ac.cn; wudongdong@mail.kiz.ac.cn

the rational genetic management of research colonies[11]. Rhesus macaques are geographically widespread and consequently genetically diverse[12–14]. Three distinct lineages are nevertheless well recognized: Indian, Chinese and Indochinese[15]. Currently, the most significant macaque bioresource, macaque genotype and phenotype (mGAP)[16], primarily concentrates on Indian rhesus macaques (IRM), with only a limited number of samples being of Chinese origin. However, it is now clear that Chinese rhesus macaque (CRM) populations exhibit considerable genetic variations, potentially surpassing that of their Indian counterparts[17], and they vary markedly in traits such as body size, pelage, and other morphological characteristics[18,19]. To effectively monitor and preserve the diversity of CRM, and with an eye to utilizing them as biomedical experimental models, a national primate facility known as "National Research Facility of Phenotypic and Genetic Analyses of Model Animals (Primate Facility)" has been established at the Kunming Institute of Zoology (KIZ), Chinese Academy of Sciences (CAS)[20]. Thanks to more than 80 years of dedicated artificial breeding efforts since the 1960s, along with the occasional introduction of new monkeys into the colony, the population now exceeds 1800 CRMs, descended from a diverse range of wild ancestors. This invaluable bioresource not only offers an opportunity to explore the genetic variation that underlies observable phenotypic, physiological and behavioral differences between macaques, but the identification of functionally significant genetic variations will also enhance our understanding of existing models thereby paving the way for the discovery of novel genetic models for inherited human diseases.

Two complementary approaches, namely forward genomics and reverse genomics, can be utilized to achieve these goals. Forward genomics, a phenotype-driven strategy (i.e., genome-wide association study [GWAS]), starts with the measurement or observation of a phenotype and proceeds to the mapping of the causative loci or genes[21]. This method is particularly powerful in deciphering the molecular mechanisms underlying natural phenotypic variation, in those cases where we have no prior knowledge of the genes involved in the biological process. Conversely, reverse genomics is a gene-driven approach that involves identifying mutations in specific genes of interest, followed by phenotypic assessment[22]. Whereas reverse genetic studies tend to be more straightforward and shorter in duration by comparison with forward genetic studies, they can be hampered by challenges such as inefficient gene knockdown or genetic background effects[23,24]. Until now, both approaches have been successfully applied to a number of model organisms, including mouse[25,26], zebrafish[27], *Drosophila*[28], and *Arabidopsis*[23].

Accordingly, we have launched the Macaque Biobank (MB) project, with the aim of capturing a wide range of phenotypic and omics data across large numbers of individual macaques. In the initial phase, we densely genotyped 919 CRMs and assessed 52 phenotypic traits that were collected from the colony of KIZ. We first explored the ancestry, genetic diversity and sequence variations present in this cohort. Next, we performed forward genomic screens to identify the genetic variants responsible for specific phenotypes. Finally, we employed reverse genomic screens, focusing mainly on neurological disease genes, to examine the phenotypic consequences arising from specific mutations. Overall, the MB introduced here promises to serve as an invaluable resource for the study of the genotype-phenotype relevance of macaques to molecular medicine, as well as for the discovery of new spontaneous models of human genetic diseases.

Results

Genetic ancestry and status of the CRM cohort

The initial dataset comprised 919 captive CRM individuals that were sequenced to a high mean depth (~30.47X) (Supplementary Data 1) and 80 wild CRM samples[18] with moderate genomic coverage (~11.71X). After applying a series of sample and variant quality controls (see Methods), we obtained a total of 84,480,388 high-quality sequence variants across 961 individuals, including 74,752,163 single-nucleotide variants (SNVs) and 9,728,225 insertions or deletions (Indels) (Fig. 1a). This corresponds to an average of one variant per 35 base-pairs (bp) genomic DNA. Nearly 59% of these variants occurred at low allele frequencies ($AF < 0.01$) whereas approximately 8.0% were classified as very common ($AF > 0.05$). The comparison of variant dataset with the largest mGAP cohort (v2.2)[16] revealed that more than 62 million of the SNVs and Indels (73.94%, Fig. 1a) were newly identified, despite the much smaller sample size of our cohort compared to that of the mGAP project[16] (961 *vs.* 2,425). This is perhaps not surprising given that the reference genome per se is an Indian-origin lineage, which is phylogenetically distinct from the CRM[29]. Nevertheless, we cannot exclude another possibility that our CRM cohort may possess higher levels of genetic diversity compared to the mGAP cohort[16], which is evident from the results presented below.

We traced the genetic ancestry of the CRM cohort by incorporating samples from diverse geographical regions of China alongside samples from India. The PCA results show a clear separation between the CRMs and the IRMs (Fig. 1b), thereby corroborating the marked genetic divergence of these two geographically separated subpopulations[18,30]. Within the Chinese samples, the captive CRMs were indistinguishable from the wild population, irrespective of whether or not the Indian-origin samples were excluded. Such pronounced admixture between captive CRM samples and the wild population was further corroborated in FRAPPE[31] -inferred ancestral clusters (Fig. 1c), implying that the captive CRMs are likely an admixture of multiple wild sources, aligning with the maintenance history of the cohort. The combination of multiple genetic ancestries introduces increased nucleotide variation into the recipient population. As expected, we found that the captive CRMs showed the highest genetic diversity (mean $\pi = 0.0016$), which is comparable to that of the wild population (average $\pi = 0.0015$) and 1.7-fold higher than the mGAP cohort[16] (average $\pi = 0.0001$) (Fig. 1d). The observation of slightly lower genetic diversity among wild individuals than captive CRMs was likely caused by their lower sequencing depth ($R^2 = 0.61$, p-value = 6.065e-09, Supplementary Fig. 1). This notwithstanding, the mutational load pattern indicated that both the captive CRMs and the wild population carried significantly fewer deleterious mutations (Fig. 1e) and homozygous loss-of-function (LoF) (Fig. 1f) than the mGAP cohort[16] (Mann–Whitney *U* test, p-value $< 2.2 \times 10^{-16}$). This pattern consistent with the anticipation that a more inbred population would logically exhibit a higher genetic load[32]. High genetic diversity and low genetic load are reliable indicators of a population's long-term viability[33,34]. These results imply that the genetic status of our captive CRMs compares favorably with the mGAP[16] samples, with a lower risk of inbreeding, germplasm degradation, and loss of genetic diversity.

Variant annotation and mutational profiling

We classified the variants into different categories based on their location and functional impact. As seen in human cohorts[35,36], the majority of the CRM variants were found in intergenic and intronic regions, accounting for 45.13% and 39.75%, respectively, whereas the variants located in coding and splicing regions made up 0.89% of the total (Fig. 2a and Supplementary Fig. 2a–c). The number of synonymous variants (~328 K) was slightly higher than the non-synonymous variants (~315 K); they together comprised 85.22% of the variants in coding and splice regions. The allele frequency distribution indicated that the non-synonymous and frameshift mutations, start/stop gains or losses, and splice site variants are more likely to be rare or singletons (Fig. 2b), reflecting the putative purifying selection acting on them.

We next examined the mutational constraint on different genes and pathways. To calibrate the number of mutations resulting from mapping to a distant reference genome (Indian rhesus macaque) and accounting for local sequence features (e.g., gene length), we utilized

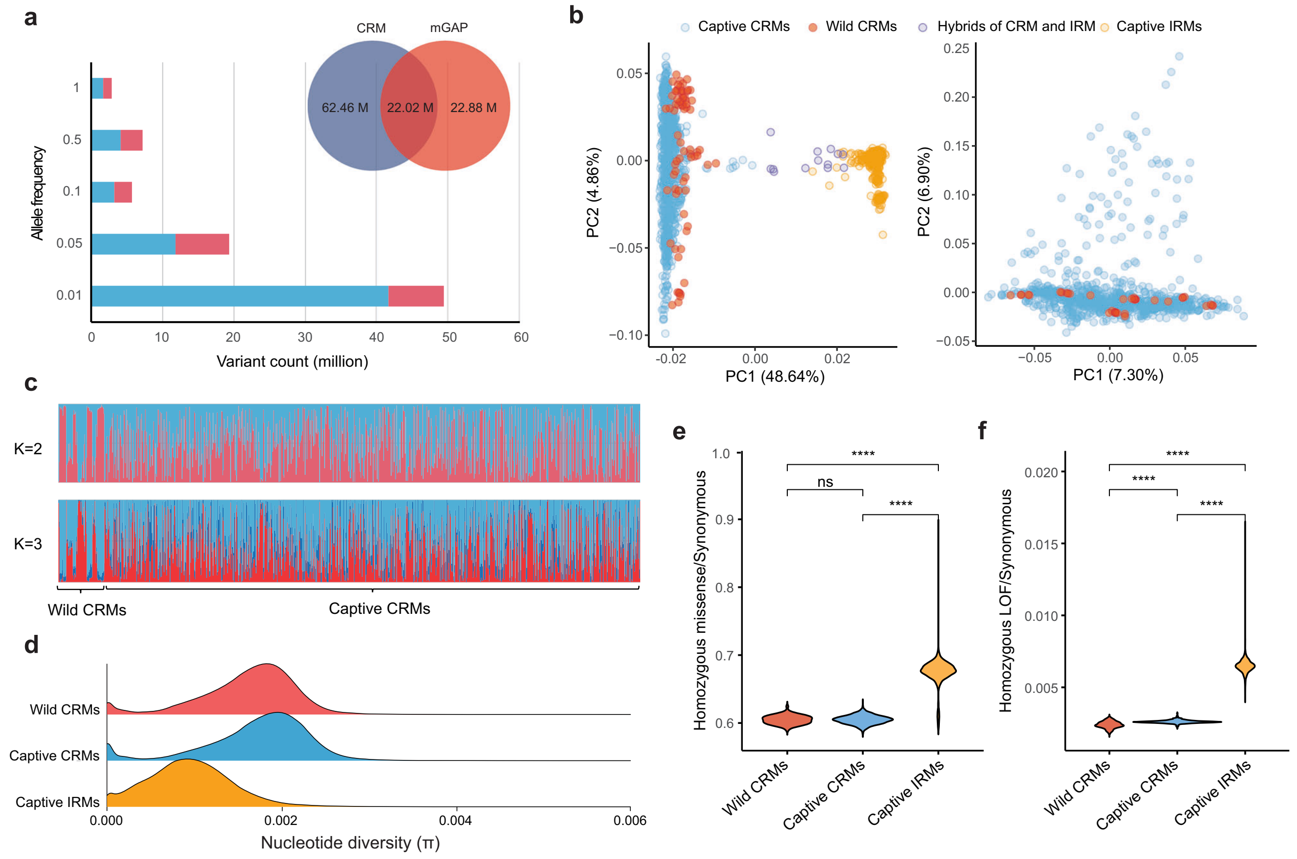

Fig. 1 | Genetic ancestry and status of the CRM cohort. **a** The Venn diagram depicting the number of variants (SNVs + Indels) detected in our CRM cohort compared to the mGAP project[16], and the stacked bar chart displaying the overlap (red) and novel variants (sky blue) finding in the CRM cohort. Variants were categorized on the basis of allele frequency (AF). **b** Genetic ancestry inferred from PCA analyses. Left: principle component plot for combined genotype data of CRM and mGAP cohorts. This combined dataset includes captive CRMs, wild CRMs, captive IRMs, and the hybrids of CRM and IRM. Right: principle component plot for CRM cohort only, which includes captive CRMs and wild CRMs. The figure in parenthesis represents the proportion of genetic variance explained by the eigenvector. **c** Mixed ancestry inferred from FRAPPE results (K = 2–3). **d** Genetic diversity (π) of captive CRMs compared to wild CRMs and captive IRMs. Mutational load pattern of captive CRMs compared to wild CRMs ($n = 961$) and captive IRMs ($n = 714$), (**e**) depicting the ratio of the number of homozygous derived missense variants to homozygous derived synonymous variants, and (**f**) showing the number of derived homozygous LoF variants to homozygous derived synonymous variants. *P*-values were estimated by two-tailed Mann–Whitney *U* test. ns not significant; ****, p-value < 2.2×10^{-16}. Source data are provided as a Source Data file.

the number of synonymous variations as a control baseline[37]. Specifically, we computed the ratio of non-synonymous to synonymous substitutions (*nsyn/syn*) for each gene. After controlling for the false discovery rate (FDR), our results showed that the most evolutionarily constrained pathways (involving genes with no observed non-synonymous mutations) were related to core biological processes, e.g., ribosome, spliceosome and proteasome components (adjusted *p*-value < 0.05, Fig. 2d), consistent with previous findings in human cohorts[37,38]. By contrast, the immune-related pathways, such as the chemokine signaling pathway, cytokine–cytokine receptor interactions, viral protein interactions with cytokine and cytokine receptors, were among the least constrained pathways (*nsyn/syn* > 4). Interestingly, several neurodegeneration pathways, such as those evident in amyotrophic lateral sclerosis (ALS), Parkinson's disease (PD), Huntington disease (HD), and Alzheimer's disease (AD), were also found to be markedly conserved (adjusted *p*-value < 0.05), implying their functional importance and strong purifying selection in rhesus macaques. It is reasonable to suppose that these categories of conserved pathways are also less tolerant to deleterious mutation.

Loss of function (LoF) variants and association with phenotypes

LoF variants, including nonsense, frameshift, or canonical splice-site mutations, are of particular interest as they have the potential to severely disrupt the functionality of protein-coding genes, thereby could serve as naturally occurring gene knockouts to explore gene function[39]. However, LoF variants are known to have a high false-positive rate due to various factors, including incomplete and imperfect genome annotation, occurrence on non-canonical transcripts or within the last 5% of the transcript[40,41]. To increase the probability of a given variant being accurately annotated as a predicted loss-of-function (pLoF) mutation, we applied a set of filtering strategies to the raw LoF variants derived from the SnpEff prediction[42] (see Methods for detail). In total, we identified 4,166 high-confidence pLoF variants across 2746 genes (Supplementary Data 2), where at least one copy of the gene was predicted to be inactivated based on both rhesus macaque and human genome annotations. Of these, the majority (83.08%) were found to be rare (MAF < 0.01) and only 5.61% of the pLoF variants were very common (MAF > 0.05). On average, each individual macaque carried 97 pLoF variants, similar to the numbers found in human genomes[39,40].

The very common pLoF alleles are likely to be LoF-tolerant because they are less constrained by purifying selection. We observed a significant enrichment in olfactory receptors among these alleles (adjusted *p*-value = 1.809×10^{-4}, Supplementary Table 1), consistent with the findings of previous studies[37,38,43]. It is intriguing to find that seven mouse essential genes (*PPP1R15B*, *IFT52*, *CYP1A2*, *ETV2*, *NFASC*, *SLC2A9*, *PLRG1*) and two human essential genes (*MAK16*, *PLRG1*) were tolerant to biallelic inactivation in CRMs (Supplementary Data 2). For

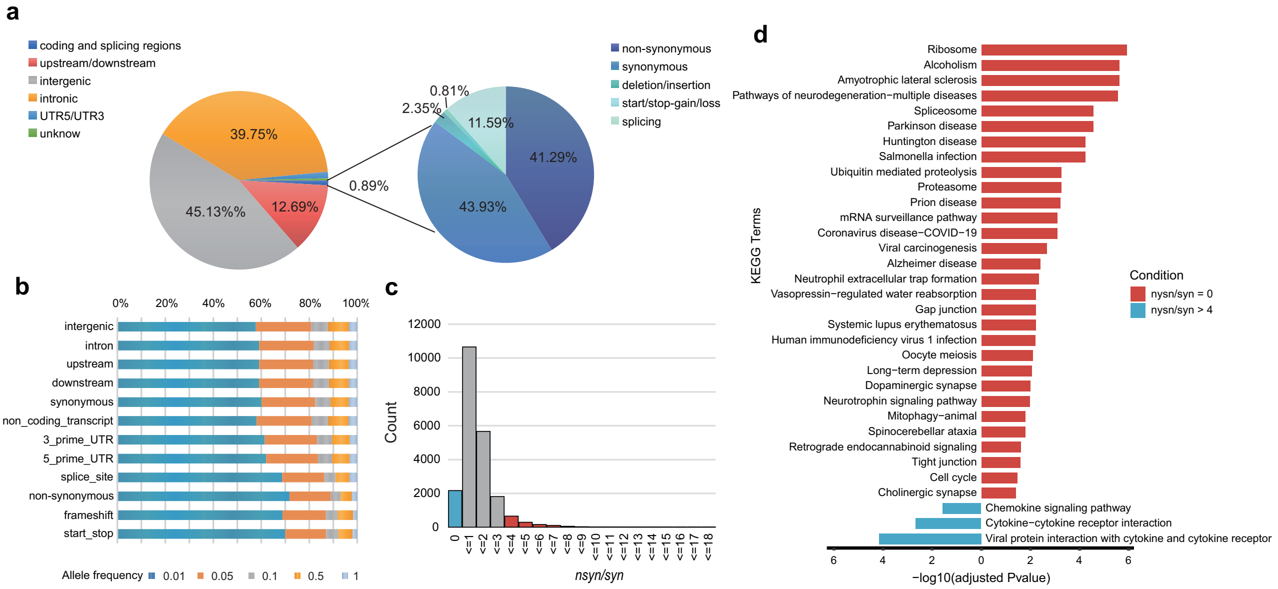

Fig. 2 | Statistics of variant annotation and mutational profile. a Pie chart displaying the proportion of functional annotation for all variants (*left*) and only variants in coding and splicing regions (*right*). **b** Variant type in coding and splicing regions categorized on the basis of allele frequency (AF). 0.01, AF ≤ 0.01; 0.05, AF ≤ 0.05; 0.5, AF ≤ 0.5; 1, AF ≤ 1. **c** Distribution of *nsyn/syn* ratio for all genes. We referred that the genes with no observed non-synonymous mutations (*nsyn/syn* = 0) in the cohort were evolutionarily conserved, whereas genes with *nsyn/syn* > 4 were less constrained. The numbers of these two categories were filled with different colors. **d** The bar plot displaying the significant KEGG pathways that were enriched for the conserved genes (*nsyn/syn* = 0; 0dnds, red) and less constrained genes (*nsyn/syn* > 4; 4dnds, blue). *P*-values were corrected by Benjamini−Hochberg algorithm. Source data are provided as a Source Data file.

example, the *PLRG1* gene, which encodes a core component of the cell division cycle 5-like (*CDC5L*) complex, is crucial for both mouse embryonic and human cells in terms of their viability[44,45]. Analysis of the transcriptome data confirmed that the splice acceptor mutation in *PLRG1* (c.10-2_10-1insA) observed in CRMs likely influences the fusion of exon1 and exon2, resulting in the transcript of *PLRG1* with a loss of the exon1 fragment (Supplementary Fig. 3). However, our observations suggest that the homozygous knockout of this gene does not result in severe consequences or a disease state in CRMs, probably the evolutionary change of gene essentiality across species[46] or a compensation effect from gene family members[47]. By contrast, rare pLoF alleles (MAF < 0.01) are expected to be less tolerated and likely associated with a strong functional effect. We found a strong depletion of homozygosity among rare pLoF variants, with only 78 (2.29%) of the variants being homozygous. These genes were significantly enriched for metabolic pathways, such as arachidonic acid metabolism, glycerophospholipid metabolism, and glycerolipid metabolism (adjusted *p*-value < 0.05, Supplementary Table 2). Interestingly, we identified 338 genes as potential drug targets within the high-quality pLoF catalog (Fig. 3b and Supplementary Data 2). These genes exhibited varying degrees of gene loss, which could potentially lead to interindividual differences in pharmacological efficacy. Consequently, the compilation of high-confidence LoF variants could serve as a key resource to guide the selection of suitable “druggable” targets, and it would be rewarding to have a primary screening for these druggable targets in CRMs for selecting the proper individuals for the pharmacological evaluations.

To further characterize the phenotypic consequences of the rare pLoF variants, we performed an association screen against 52 distinct phenotypes (Supplementary Tables 3 and 4). Association results surpassed the Bonferroni significance threshold (p-value = 2.83×10^{-5}, see “Methods”) for seven pLoF-trait pairs (Supplementary Table 5). The most significant association was a splice acceptor variant in *ANO10* (c.203-2 AG > G), which was related to the full-leg length (p-value = 8.97×10^{-6}). Compared to the non-carriers, *ANO10* (c.203-2 AG > G) heterozygotes displayed a significant reduction in full-leg length (Mann−Whitney *U* test, *p*-value = 0.0251, Fig. 3c). Notably, *ANO10* (c.203-2 AG > G) heterozygous carriers also exhibited a nominally significant reduction in full-arm length (Mann−Whitney *U* test, *p*-value = 0.0139, Fig. 3d), although the association test (p-value = 4.12×10^{-5}) did not surpass the level of significance required by Bonferroni correction, likely because the correction approach is highly conservative and would tend to “overcorrect” the variants in the context of a mild or small effect[48]. *ANO10* encodes a transmembrane protein that belongs to the transmembrane 16 family. Defects in this gene can cause ataxia, a neurological condition characterized by gait and balance impairment, upper limb coordination problems, as well as impairment of speech and eye movements[49,50]. However, to our knowledge, *ANO10* has never been reported to be associated with limb length. Similarly, we could identify a heterozygous splice acceptor mutation at *PRRC2B* (c.6379-2 A > G), which was predicted to play a role in embryonic development[51], was significantly associated with a higher body weight (p-value = 9.67×10^{-6}, Fig. 3c). If employing a less conservative association *p*-value threshold (e.g., 1×10^{-4}), we could identify another 13 associations that was aligned with the gene function (Fig. 3d). For instance, the carriers of a stop gain mutation in the *ATR* gene possess a smaller head length (Mann−Whitney *U* test, *p*-value = 0.0072). It has been suggested defects of this gene was a cause of Seckel syndrome 1, a syndrome characterized by severe intrauterine and postnatal growth retardation, microcephaly and mental retardation[52]. In addition, the heterozygous knock-out of *ALOX15*, which encodes an enzyme that acts on various polyunsaturated fatty acid substrates[53], was associated with lower high-density lipoprotein (HDL) and low-density lipoprotein (LDL) concentrations in serum of CRMs (*p*-value = 0.0042 and 0.0073, respectively).

Genome-wide association for 52 phenotypes in CRMs

The availability of multiple genomes coupled with phenotypic data also provides an unprecedented opportunity to investigate the genetic foundations of phenotypic variation in CRMs. To this end, we performed GWAS analyses for each quantified trait on the common variants (SNVs + Indels) with a mixed linear model by fitting relevant covariates, e.g., age, sex, genetic relationship, population structure (see Methods). The genomic control factor λ did not show any sign of inflation for all tests (λ < 1.03), suggesting that population structure has been well controlled. The resulting statistical power of a GWAS was 0.0046 (Supplementary Fig. 4) when assuming a mean heritability ($\hat{h}^2$) of 0.5 for the traits. This indicates that 0.46% of the causal variants

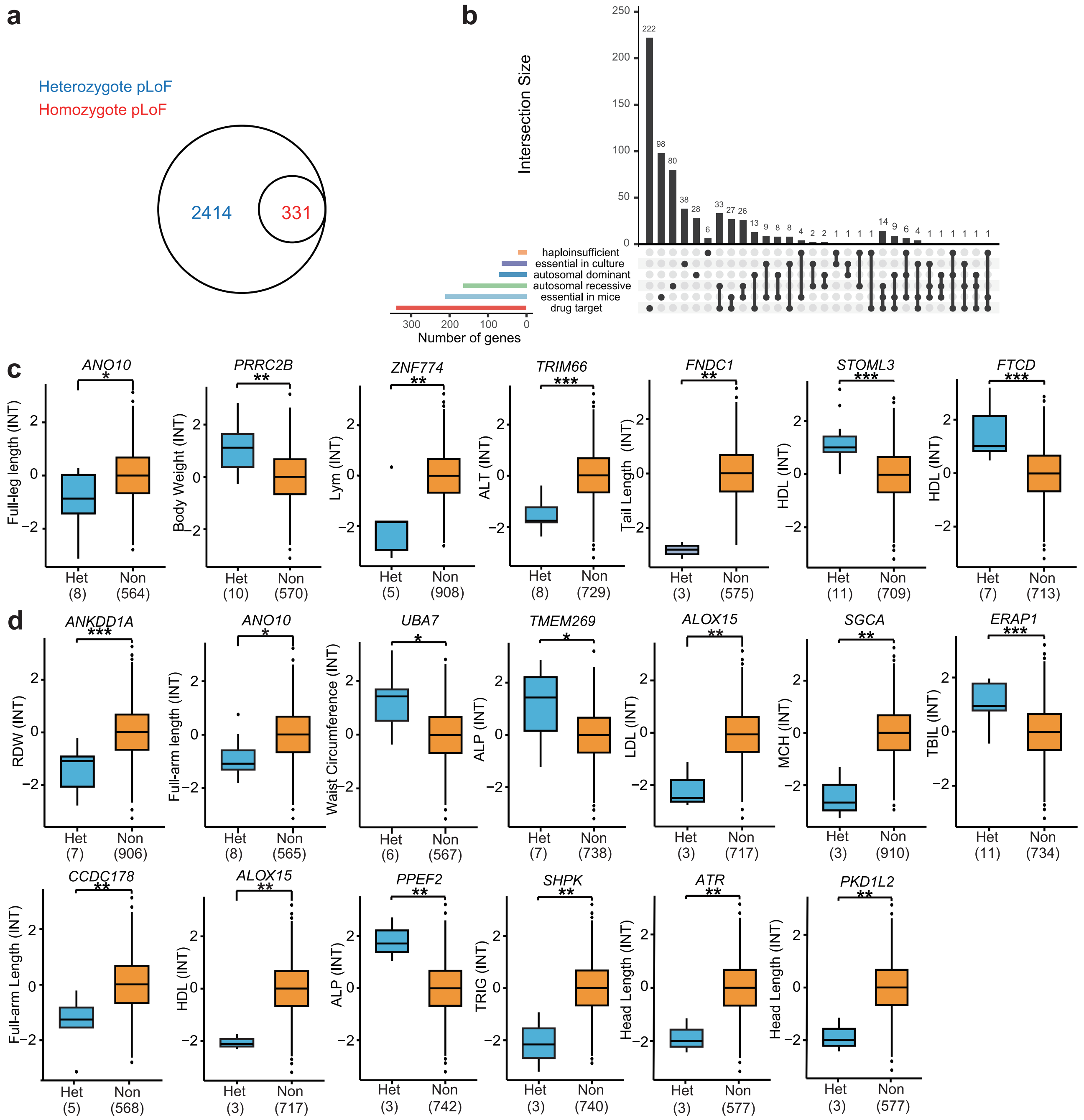

Fig. 3 | pLoF genes and their association with phenotypes. **a** Number of heterozygous (blue) and homozygous (red) pLoF genes in the CRM cohort. **b** UpSet plot depicting the intersection of pLoF genes with the following gene lists: haploinsufficient gene determined by the ClinGen Dosage Sensitivity Map (haploinsufficient), essential gene in multiple cultured cell lines (essential in cultured), essential gene for the viability of mice (essential in mice), OMIM disease genes of autosomal dominant and autosomal recessive, and drug target in DrugBank. The first five gene lists are available at https://github.com/macarthur-lab/gene_lists, corresponding to the "ClinGen haploinsufficient genes", "Essential in culture", "Essential in mice", "All dominant genes", and "All recessive genes" lists, respectively. The drug target gene were annotated by Metascape[108]. Details were provided in Supplementary Data 2. Associations of pLoF gene with the phenotypic trait that surpassed (**c**) the Bonferroni significance threshold (p-value = 2.83×10^{-5}), and (**d**) p-value of 1×10^{-4} based on a mixed linear model. These results are ordered according the significance of association p-value. Please refer Supplementary Table 5 for detail variant type and exact p-value for each gene. The values on the y-axis represent the trait that were separately normalized using inverse normal transformation (INT) and were adjusted for age and sex. Numbers in brackets (x-axis) indicate the sample size with the mutation. P-values were estimated by two-tailed Mann–Whitney U test. The center line of the boxplot represents the median, the box spans the interquartile range (IQR, 25th to 75th percentile), and the whiskers extend to the minimum and maximum values within 1.5 × IQR. Outliers beyond this range are shown as individual points. Asterisks denote the level of significance of the compared groups by a two-tailed Mann–Whitney test. "Het" denotes the group of heterozygous allele carriers. "Non" represents the non-carriers. *, p-value < 0.05; **, p-value < 0.01; ***, p-value < 0.001. Source data are provided as a Source Data file.

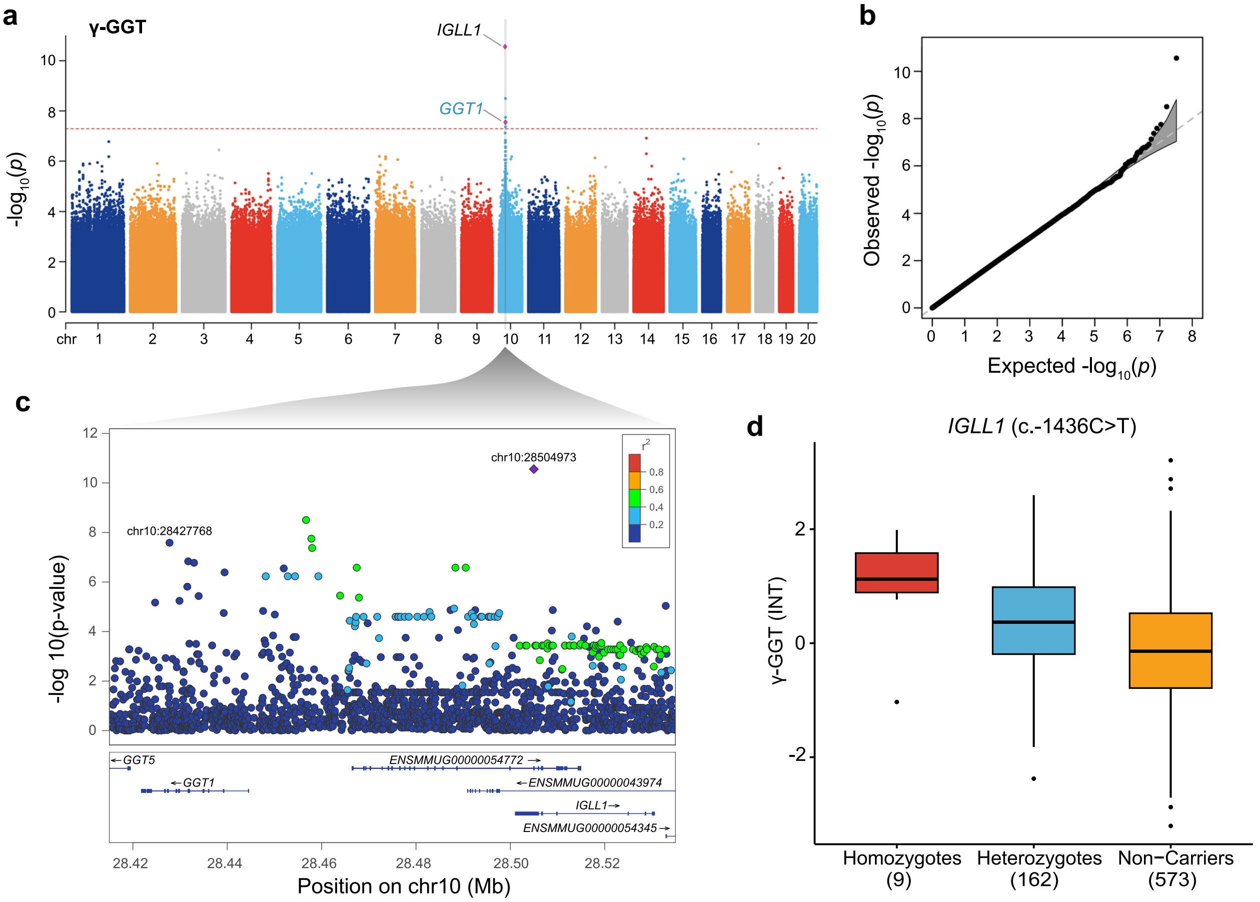

Fig. 4 | Illustrative examples of GWAS. a Manhattan plots showing the GWAS result for γ-GGT concentration in CRMs. The genetic loci that satisfied the genome-wide significance threshold of *p*-value $< 5.13 \times 10^{-8}$ (red dashed line) are presented. This threshold was estimated by using a uniform threshold of 1/n, where n is the effective number of independent variants. The locus of *GGT1* has been previously reported to be associated with γ-GGT concentration[60] in the human GWAS catalog and is highlighted in blue. **b** Q-Q plots corresponding to the Manhattan plot of γ-GGT. Gray shaded areas show 95% confidence intervals for the expected distributions. **c** LocusZoom plots for the two independent SNPs (chr10:28504973 and chr10:28427768) related to the γ-GGT concentration on chromosome 10. The purple diamond (chr10:28504973) represents the most significant SNP; all other variants are colored by their r^2 values. SNP positions and gene boundaries are based on the gene build of Mmul_10[79]. **d** Carriage of the most significant SNP (chr10:28504973), c.-1436C > T, in 5′-UTR region of *IGLL1*, was associated with a step increase of γ-GGT concentration. Phenotypic data were normalized using INT method. Numbers in brackets (x-axis) indicate the sample size with the mutation. The center line of the boxplot represents the median, the box spans the interquartile range (IQR, 25th to 75th percentile), and the whiskers extend to the minimum and maximum values within 1.5 × IQR. Outliers beyond this range are shown as individual points. The summary statistics of GWAS results can be download from Non-Human Primate BioBank database (https://nhpbiobank.kiz.ac.cn/Home/Download). Source data are provided as a Source Data file.

could be detected given a sample size of 875 (the number of captive CRMs possessed the phenotypic data). In total, we identified 44 variants associated with 16 phenotypic traits that passed the genome-wide significance threshold (*p*-value $= 5.13 \times 10^{-8}$). These variants were clumped into 30 independent loci across 18 chromosomes, explaining 3.36–5.97% of phenotypic variations (Supplementary Fig. 5 and Supplementary Table 6).

The annotation of these significant variants revealed six genes (*DCDC2C*, *TRIB1*, *EDIL3*, *GGT1*, *SHISA9*, *WWOX*) have been reported to be associated with specific human traits (Supplementary Table 6). For instance, the *EDIL3* gene, which encodes an integrin ligand, has been previously suggested to be related to human body mass index (BMI)[54]. In this study, we discovered that a downstream variant of this gene was significantly associated with a reduction in BMI in rhesus macaques (beta = −1.0737, *p*-value $= 6.30 \times 10^{-9}$, Supplementary Fig. 5c). We also observed associations of the *SHISA9* locus link to hip circumference[55](beta = −0.6693, *p*-value $= 3.69 \times 10^{-8}$), and the *WWOX* locus with body weight[56](beta= 0.3909, *p*-value $= 3.87 \times 10^{-8}$) (Supplementary Fig. 5j, p). Apart from these known associations, we identified 11 significant associations that had not previously been reported in the human GWAS catalog[57] (Supplementary Table 6). Of these, the most significant association was observed for a 5′-UTR variant at the *IGLL1* locus (c.-1436C > T), which was related to the serum gamma-glutamyl transpeptidase concentration (γ-GGT) level in CRMs (*p*-value $= 2.76 \times 10^{-11}$, Fig. 4a, b and Supplementary Table 6). This gene encodes an immunoglobulin lambda-like polypeptide 1 protein which plays an important role in B cell development[58]. In CRMs, the heterozygous and homozygous carriers exhibited a gradual increase in γ-GGT concentration as compared to non-carriers (Fig. 4d). Interrogation of human ENCODE databases[59] revealed that this signal region exhibited distinct active enhancer signatures in a range of human cell types (Supplementary Fig. 6). It is noteworthy that this peak also encompassed an independent locus of *GGT1* (*p*-value $= 2.59 \times 10^{-8}$), which has previously been reported to be associated with γ-GGT level in human[60]. However, regional association analysis indicated that these two variants were in weak linkage disequilibrium (LD) ($r^2 = 0.01$, Fig. 4c), suggesting they are being independently linked to the GGT level.

Reverse genetic screen identifies *DISC1* (p.Arg517Trp) as a genetic risk factor for neuropsychiatric disorders

The above classical forward genetic approaches enabled the identification of multiple genetic variants associated with the phenotypic variations in CRMs. It is intriguing to verify whether a distinct genotype can predict a specific phenotype. In a reverse genetic screen, we identified 3192 non-synonymous mutations across 2216 genes that were predicted to be deleterious based on the intersection results of SIFT4G[61] and PolyPhen-2[62] (Supplementary Data 3). We are particularly interested in the genes related to human neurological disorders (NDs) as these complex diseases are difficult to investigate using rodent models[3,63]. Non-human primates (NHPs) are not only phylogenetically close but they also share similar brain structure and function with humans, making them more suitable for the study of human NDs than other mammalian species[64]. Below, we highlight the case regarding the phenotypic consequences arising from a deleterious missense mutation in the *DISC1* (Disrupted-In-Schizophrenia 1) gene (p.Arg517Trp, c.1549 C > T, SIFT4G score = 0.01).

In this cohort, we identified eight CRMs that carried the *DISC1* p.Arg517Trp mutation in the homozygous state versus 725 non-carriers. These macaques included three adults (aged 5–7 years) and five elderly individuals (aged over 19 years). Given that aging could potentially affect the results obtained (e.g., working memory), we focused on the three adults and excluded the elderly monkeys from the behavioral and brain imaging experiments. We observed a significant reduction in neurological function in carriers of the risk allele (Trp) than 19 non-carriers (Arg, two-tailed *t*-test, p-value < 0.0001, Fig. 5d). This reduction was manifested by diminished limb reflexes, as well as a decreased response to pain and teasing. We further assessed the working memory under mild-stressful and non-stressful conditions, respectively. Our results showed that risk allele carriers consistently exhibited lower working memory performance with increasing delay lengths, and this pattern was particularly evident in the trials with 30 s delays (Fig. 5a, b). When a restraint stress was applied, the risk allele carriers displayed markedly more errors under these stressful conditions (two-tailed *t*-test, p-value = 0.0363, Fig. 5c). Since stress is a risk factor for psychiatric disorders associated with impaired prefrontal function[65,66], these data may help to explain why the deleterious missense mutation of *DISC1* increases the risk of psychiatric disorders.

Next, we carried out magnetic resonance imaging (MRI) to examine whether any cortical structure was altered in *DISC1* Trp carriers. Although we did not detect a significant reduction in gray matter volume and thickness (Supplementary Fig. 7), we observed an increase in gray matter surface area in the frontal lobe of the Trp risk allele carriers (p-value = 0.0338, Fig. 5e), particularly in the motor cortices of the caudal frontal lobe. Additionally, we detected a significant reduction of white matter volume in the temporal lobe (p-value = 0.0064, Fig. 5f) and a significant increase in ventricular volume (p-value = 0.0169, Fig. 5g). Further region-level results confirmed that the majority of changes in gray matter surface area and white matter volume were localized to the frontal lobe and temporal lobe, respectively (Supplementary Figs. 8, 9). We also collected resting-state functional magnetic resonance imaging (rs-fMRI) data. Although no significant group differences in functional connectivity (FC) were observed at the whole-brain or lobe levels, the lobar analysis revealed a trend toward reduced parietal-frontal FC and increased subcortical and frontal-frontal FC in Trp monkeys (Supplementary Fig. 11a, b). The network-based statistic (NBS) was further conducted across a range of primary thresholds (t = 3.0–3.4) to identify differences in functional connectivity between the Trp-bearing macaques and the Arg controls under the general anesthesia. As the primary threshold increased, a stable set of differing functional connectivity persisted (Supplementary Fig. 10), with the results at the median threshold (t = 3.2) presented in Fig. 5h,i. Among these findings, the majority of increased functional connectivity measures in Trp-bearing monkeys were localized within the frontal lobe (n = 11), while a subset was observed between the frontal lobe and subcortical regions (n = 7) (Fig. 5h). Additionally, we identified 27 connections that displayed a reduction in strength in the Trp-bearing macaques compared to controls, with the majority of these reductions occurring between the frontal lobe and parietal lobe (n = 13) (Fig. 5i). At the regional level, Trp monkeys showed altered functional connectivity density (FCD) in the SII, areas 24a/b prime, and the nucleus accumbens (Acb) (Supplementary Fig. 11c, d). Since the sample size in the current study is relatively small, especially considering the challenge in identifying group difference in resting state fMRI with only 3 Trp-bearing macaques, the functional connectivity results should be interpreted with caution. Further studies with larger sample sizes are needed to validate these findings.

Discussion

The macaque cohort presented here represents one of the most extensive sequencing studies so far performed in rhesus macaques, although our data have primarily been derived from the CRM population. This notwithstanding, we have for the first time incorporated a diverse array of phenotypic data from numerous macaque individuals. The current cohort comprises genomic data from 961 CRMs, supported by 52 hematological, biochemical and anthropometric measurements. Our preliminary analyses indicate that the captive CRMs are a mixture of animals from multiple wild sources, which was consistent with the introduction of wild animals into the colony to avoid potential inbreeding. Together they harbor over 62 million variants (74%) that were previously undetected in the mGAP project[16], thereby demonstrating the distinctness of the CRM and IRM lineages, which serves as a *caveat* for their use as nonhuman primate models. The higher nucleotide diversity in the CRM cohort was also supported, but our new data with its large sample size and high coverage genomic sequencing indicate that the captive CRMs carry a significantly lower genetic load, and hence are less susceptible to inbreeding compared to the mGAP individuals[16].

The relatively large sample size of the genomic data obtained enables us to assess the sensitivity of genes to functional variations in non-human primates, thereby enhancing our capacity to discover disease-related genes, especially in these monkeys with spontaneous diseases[67–69]. Our results corroborate previous findings performed on large human cohorts[37,38], indicating that genes implicated in core biological processes (e.g., ribosome, spliceosome and proteasome components) belong to the most constrained categories, whereas immune-related genes are the least constrained. Notably, we discovered that human orthologous genes associated with neurological disorders, such as ALS, PD, HD and AD, are also under strong selective constraints (Fig. 2d). This implies that these neural genes are of functional importance and have been conserved in rhesus macaques, making them less tolerant to LoF mutations or detrimental non-synonymous mutations. Our findings therefore provide compelling genetic evidence to support the use of rhesus macaques as a suitable model for studying neurological diseases[3,64,70].

Employing a reverse genomic approach, we successfully demonstrated a case arising from a deleterious missense mutation in the macaque *DISC1* gene (p.Arg517Trp, Fig. 5), a well-recognized risk gene for several types of human neuropsychiatric disorder. This gene encodes a multi-compartmentalised protein that functions as a scaffold hub, interacting with numerous partners involved in brain development and disease processes. Defects in *DISC1* have been reported to be associated with impaired working memory[71]. Anatomical changes mostly involve cortical abnormalities, including the prefrontal cortex as this area plays an important role in executive functions and working memory[72]. In this study, our data collectively indicated that the macaques carrying the risk allele of *DISC1* p.Arg517Trp exhibited alterations in cortical architecture and functional connectivity (Fig. 5e–i), which may ultimately contribute to the observed neurological deficits and impairment of working memory (Fig. 5a–d). As working memory impairment is a contributing symptom to most neuropsychiatric disorders linked to *DISC1* mutations, these promising

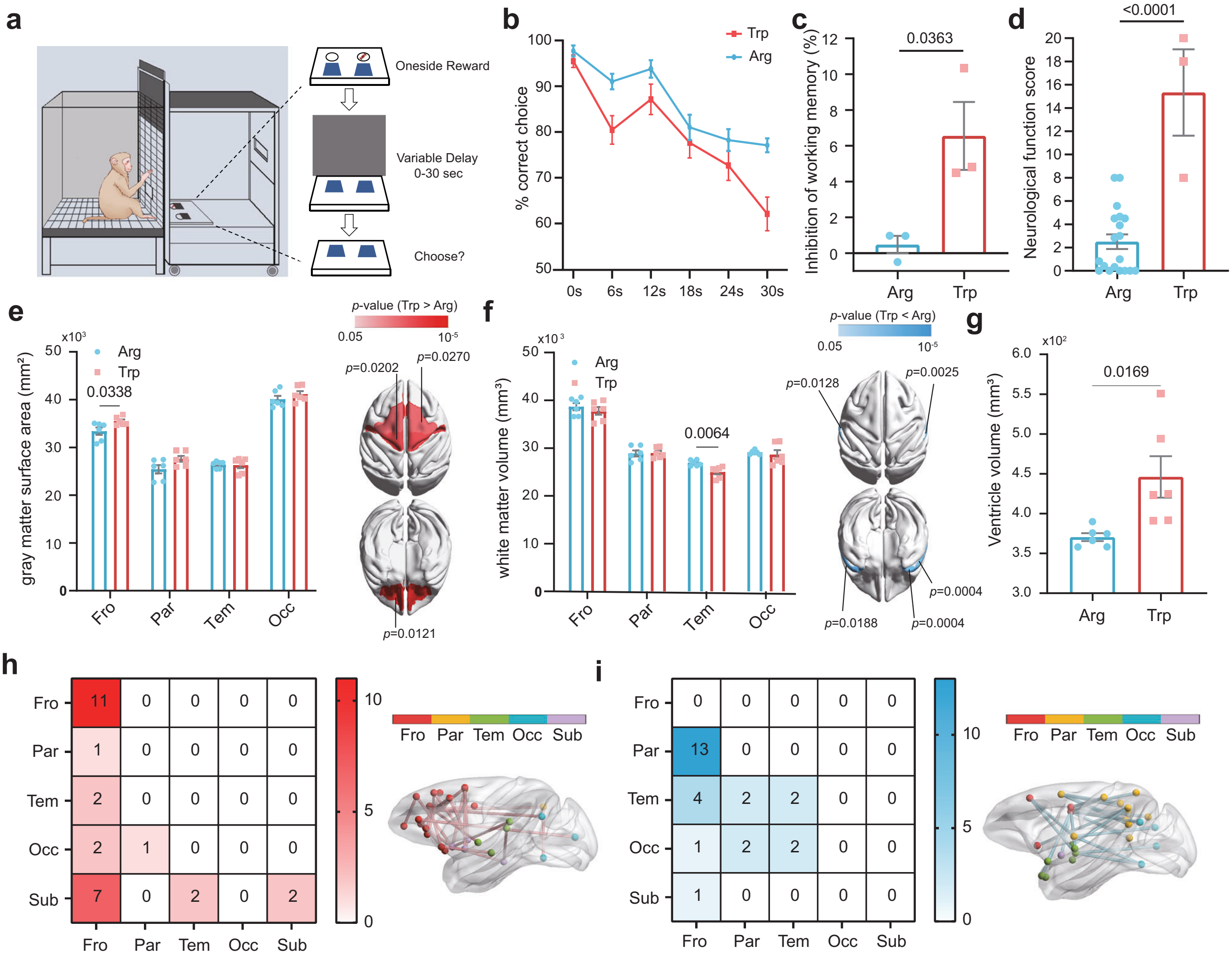

Fig. 5 | Phenotypic consequences arising from a deleterious missense mutation in *DISC1* (p.Arg517Trp). **a** Schematic diagram of spatial working memory test using the Wisconsin General Test Apparatus (WGTA), where the macaque was allowed to choose food (e.g., peanut) from one of the two covered wells with various time delays. Macaque drawings by Mu-Ru Zhou; full painting by Hong-Di Huang. **b** Performance (percentage of correct choice) of Trp-bearing macaques ($n = 3$, red) and Arg controls ($n = 3$, blue) across six tested delays (0 s, 6 s, 12 s, 18 s, 24 s, 30 s). Data here and below are presented as the mean ± SEM. **c** Inhibition of working memory (ratio of errors score) caused by restraint stress between the Trp-bearing macaques ($n = 3$) and Arg controls ($n = 3$). *P*-value in working memory examinations was estimated by two-tailed *t*-test. **d** Neurological function scores performed for 3 adults of homozygous carriers (Trp) vs 19 non-carriers (Arg). *P*-values were estimated by two-tailed unpaired *t*-test. **e** Quantification of brain structure differences in macaques with 3 Trp carriers compared to 3 Arg controls that involved in working memory examinations. Left: gray matter surface area of four lobes in Trp-bearing macaques and Arg controls. Right: visualization of frontal brain regions showing significant differences between Trp-bearing macaques and Arg controls on the mid-gray surfaces of the macaque template. Red indicates the surface area of Trp-bearing macaques is larger than Arg controls, blue is the opposite. Fro, frontal lobe; Par, parietal lobe; Tem, temporal lobe; Occ, occipital lobe. **f** White matter volume of the four lobes (left) and visualization of temporal brain regions showing significant differences (right) between the 3 Trp carriers and 3 Arg controls. **g** Ventricle volume of the whole-brain between the Trp carriers ($n = 3$) and Arg controls ($n = 3$). **e**–**g** Quantitative data in Arg and Trp groups are presented as means ± SEM, with data collected from the two hemispheres of each monkey. Generalized linear mixed models (GLMMs) were used to estimate the statistical significance. Number of pairs of brain areas exhibiting increased (**h**) and decreased (**i**) functional connectivity within and across lobes and subcortical area at the median threshold (t = 3.2), respectively, in Trp macaques compared to Arg controls. Detailed information is presented in the right panels adjacent to the matrices. Red, yellow, green, blue and purple dots represent regions of frontal (Fro), parietal (Par), temporal (Tem), occipital lobe (Occ) and subcortical (Sub) regions, respectively. *, $p < 0.05$; **, $p < 0.01$; ****, $p < 0.0001$.

data provide a remarkable bridge across human and macaque species, albeit the number of macaques with *DISC1* p.Arg517Trp were relatively small. The naturally occurrence of disease in captive macaques provided a unique resource for establishing non-human primate models for human diseases. With the genome information affiliated prediction, it would be interesting to monitor the onset of spontaneous diseases in these macaques with the pLoF and deleterious missense mutation.

Although we have identified hundreds of pLoF variants and missense variants that were matched to known human diseases and drug target genes. However, as demonstrated in previous studies[73,74], we cannot be certain that these functional mutations will be associated with an increased susceptibility to certain inherited diseases. Distinguishing the disease-causing mutations from benign genetic variation is challenging and problematic for organisms like macaques as there are few sources of genomic annotations, relative to human or mouse. Detailed phenotypic data offer a promising approach for understanding these functional mutations, as exemplified by the phenotypic consequences observed in *ANO10*, *PRRC2B*, *ATR* and *ALOX15* (Fig. 3c, d and Supplementary Table 5) in this study and human biobank[75,76]. The integration of extensive phenotyping data in the future will enhance the accuracy and reliability of predicting the significance of genetic mutations in macaque genome[20,77].

Despite these significant observations, several limitations deserve attention. First, although we have applied a series of filtering strategies as well as liftover to human coordinate to utilize human data sources, any given mutation annotated as pLoF may not truly lead to loss of protein function. Therefore, experimental validation such as reverse-transcription PCR of transcript and/or western blotting of protein will ultimately be required in order to address this issue. A second limitation is reduced statistical power to establish unambiguous genotype–phenotype correlations if the pLoF is observed in only one or two participants, a similar issue also seen in GWAS analyses. The limited sample size also restricts the statistical power of our findings in brain structure and resting state fMRI under sedation. This could be improved if larger sample sizes were employed in the future. Currently, we have started the breeding of Trp-bearing macaques to expand the mutant colony. Finally, our analysis was limited to readily available phenotypes; in future analyses, a standardized clinical phenotyping protocol would be desirable for each participant.

In short, we provided a large-scale genome dataset for CRMs, which serve as an invaluable resource for the study of the genotype-phenotype of macaques and for potential usage of precision medicine. This resource can also guide the selection of appropriate models for experimental and pharmaceutical tests, facilitating the discovery of new genetic models for human disease research, and further improving and refining the rational genetic management of macaque colonies.

Methods

Sample collection and sequencing

We enrolled a total of 919 Chinese rhesus macaques (Supplementary Data 1) that were housed in KIZ, for genomic sequencing during their annual physical checks (normally September or October, outside the breeding season) since 2021. The initial cohort comprised 293 males and 626 females, aged from 3 to 30 years. To ensure that our blood collection did not adversely affect the safety of the monkeys, we extracted a 3–5 ml peripheral blood sample from each individual using conventional intravenous sampling method. One half of each blood sample was used for hematological trait examination while the other half was used for genomic DNA extraction using the QIAGEN® extraction kit. After DNA quality assessment, libraries were prepared following the standard protocol of the DNBseq platform and sequenced to a target depth of ~30× per individual, generating about 90 GB sequencing data. All samples were collected in accordance with the policy of the Institutional Animal Care and Use Committee (IACUC) of KIZ, CAS (Approval ID: IACUC-PE-2022-11-003 and IACUC-PE-2024-11-002), which conforms to the regulatory standards for the human care and treatment of animals in research.

Phenotypic data collection

Hematological trait examination was performed using a hematology analyzer (Mindray, BC-5000Vet, China), which recorded 21 standard sets of blood cell traits. We also obtained a number of biochemical and anthropometric body measurements (summarized in Supplementary Tables 3, 4) during the following year (2022). Prior to biochemical testing, participant animals fasted overnight or at least 6 h prior to the peripheral blood sample being drawn, and the blood was centrifuged within 60 min of venipuncture. The serum samples were subsequently used to measure the biochemical traits via an automated autoanalyser (Dimension EXL200). For anthropometric body measurements, all individual animals received an intramuscular injection of 5 mg/kg ketamine to ensure sedation on the operating table while the various measurements were being obtained. We took 11 body measurements as well as the body weight for each animal. These measurements were taken following the standardized procedures as described in Supplementary Table 3.

Variant calling and filtration

To explore the genetic ancestry of our sequenced individuals, we additionally included 80 wild CRMs[18] in our cohort. We followed the Genome Analysis Toolkit (GATK) best practices pipeline[78] to call the variants. Briefly, raw sequence reads were mapped to the reference genome of IRM (Mmul_10)[79] using BWA-MEM v0.7.17-r11198[80] with default parameters. Sambamba[81] was used to remove multiple aligned, duplicated and unaligned reads. We first obtained the GVCF file for each sample using the HaplotypeCaller function in GATK version 4.1[82]. Then joint calling was performed to generate 'raw' variant data via the GenotypeGVCFs function. We used the following hard quality filter criteria (QD < 2.0 | | QUAL < 50.0 | | FS > 60.0 | | MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0) for SNPs filtering, and (QD < 2.0 | | QUAL < 50.0 | | FS > 200.0 | | MQ < 40.0 || ReadPosRankSum < -20.0) for Indels filtering, respectively, as suggested by the pipelines. After this, the filtered variant call files were merged together for subsequent quality control.

Variant level quality control. To reduce false positive calls, we removed SNPs occurring in a cluster (more than three SNPs within 10 bp) using the VariantFiltration function in GATK (--cluster-size 3 --cluster-window-size 10) because these tightly spaced SNPs are more likely to result from read mis-alignment. In addition, variants located within 6 bp of predicted indels, presenting in fewer than 80% of individuals, and the approximate read depth exceeded 97.5% or lower than 2.5% of the quantile distribution, were also filtered using BCFtools v1.9[83]. Triallelic alleles were further filtered out in the population genetic analyses (e.g., PCA, STRUCTURE).

Sample level quality control. For quality control of samples, we first removed duplicate samples (number = 8) with kinship coefficient >0.35 based on the estimations from KING software[84]. Then we removed samples (number = 26) with an excess of heterozygosity calls (inbreeding coefficient<− 0.1) or outlier number of SNPs (>17,000,000) which roughly equal three standard deviations of the mean. We also examined whether the self-reported information on gender could be verified by the "check-sex" option implemented in PLINK software (v1.90b6.9)[85]. This procedure identified two samples with discrepancy of sex identity which subsequently removed in downstream analyses. Finally, having removed samples (number = 2) with high missingness (>0.05), we retained 961 samples in the final cohort.

Variant annotations

Identification of loss-of-function variants. The effects of filtered variants were annotated and classified by software SnpEff version 4.3[42] based on the latest rhesus macaque gene build (Mmul_10)[79]. The putative loss-of-function (LoF) annotations, e.g., stop gains, stop losses, start losses, frameshifts, splice-disrupting mutations, were extracted and filtered using the accompanying software of SnpSift[42]. We retained those LoF variants that were predicted to affect more than 50% of transcripts (LOF[*].PERC > 0.5) and where the nonsense-mediated mRNA decay (NMD tag) occurred within more than half of the transcripts. The LoF variants located within the last 5% of the length of the transcript were filtered out using in-house Perl scripts. These steps led to 12,012 LoF variants retained. Despite these filtering strategies, LoF variants are known to be enriched for annotation artefacts, e.g., exons flanked by non-canonical splice sites or incomplete transcripts[40,41]. We utilized LOFTEE[37], a plugin of Ensembl Variant Effect Predictor (VEP)[62], to filter out the aforementioned LoFs. As LOFTEE is currently only available for the human genome, we utilized the LiftOver function in Picard (v2.23.9) (http://broadinstitute.github.io/picard) to transfer the variants in macaque (Mmul_10) position to the human genome (hg38) based on the overchain file download from the UCSC database. Only the successfully transferred (9136, 76%) and labeled high-confidence (HC) LoF variants were then considered as predicted LoF variants (pLoFs, $n = 4166$) in the following analyses.

Inferring the pathogenicity of missense variants. We used the software of SIFT4G[61] to predict the deleteriousness of missense variants. Prior to this step, a custom database was built with the genomic annotation file of Mmul_10[79]. The scores of SIFT4G range from 0 to 1, and SNPs are predicted to be deleterious if the score is <0.05 and tolerated if the score is ≥0.05. We also utilized human genome annotation to further infer the potential pathogenicity of missense variants detected in the macaque genome. Again, the LiftOver tool in Picard (v2.23.9) (http://broadinstitute.github.io/picard) was used to transfer the variants in the macaque (Mmul_10[79]) position to the corresponding human coordinates (hg38) based on the overchain file download from the UCSC database. Then, the functional impact of amino acid substitutions was predicted by SIFT and PolyPhen-2 implemented in VEP[62].

Function enrichment analyses

The web-server g:Profiler[86] was used to explore whether specific types of biological function were over-represented among the discovered genes. The species *Macaca mulatta* (Rhesus macaque) was selected as the background organism. *P*-values were adjusted by means of the Benjamini–Hochberg correction algorithm and the terms with false discovery rate (FDR) q < 0.05 were deemed to be significant.

Analyses of genetic ancestries

We performed principal component analysis (PCA) in software GCTA (v. 1.94.0)[87] to infer the genetic ancestries of the sequenced rhesus macaques. Two sample sets were used: one included the Indian-origin rhesus macaque (mGAP v2.2)[16] in our cohort and one without. The variant data from the mGAP project were filtered in the same manner and then merged with our cohort via BCFtools software[83]. For each sample set, we restricted our analyses to bi-allelic SNPs on autosomes and common variants with MAF above 1%. We further reduced the number of sites by applying a linkage disequilibrium (LD) pruning filter using PLINK v1.90b6.9 (–indep-pairwise 50 5 0.1)[85]. We also used Frappe 1.1 (EM algorithm)[31] to infer the individual ancestries. The postulated number of ancestral clusters (K) was set to range from 2 to 6, and the maximum number of EM iterations was set to 10,000.

Analyses of genetic diversity and genetic load

The level of nucleotide diversity (π) was estimated in a 50-kb sliding-window size with no step using VCFtools (v0.1.17)[88]. However, estimating genetic load is challenging without information on the fitness effects of deleterious mutations. An alternative approach is to estimate changes in mutational load (i.e., number of deleterious mutations)[89]. For the CRM cohort in this study and the mGAP cohort[16], we calculated the ratio of the number of derived homozygous LoF variants to homozygous derived synonymous variants, as well as the number of homozygous derived missense variants to homozygous derived synonymous variants for each individual, respectively, based on the annotation of SnpEff results (version 4.3)[42]. Since no ancestral allele information is available for macaques, we followed the example of a previous study in adopting the minor allele as the derived allele[90].

Association analyses with rare pLoFs

From the list of high confidence rare LoF mutations identified above, we sought to determine whether any of the pLoF variants was associated with phenotypic trait variation. We employed a mixed linear model-based association analysis (GCTA-MLMA)[91,92] for each pLoF–trait pairing. Quantitative traits were inverse normalized and age, sex and the first four ancestral clusters of FRAPPE[31] results were used as covariates. To reduce the likelihood of false positives, we only considered the pLoF–trait pairs in which there were at least three LOF alleles genotyped, yielding 1767 (2373) pLoF–trait pairs for analysis. After Bonferroni correction, we considered 2.83×10^{-5} (0.05/1767) as a threshold of significance.

Phenotype data processing and GWAS analyses

In order to focus on determinants of variation in the general population rather than on specific diseases, each quantitative trait was filtered those data over seven standard deviations of the mean value prior to GWAS analysis. Subsequently, the filtered trait data were standardized by rank-based inverse normal transformation (INT) using in-house R scripts. Genotype data were further filtered to exclude variants with a missing genotype rate greater than 0.02, minor allele frequency (MAF) less than 0.01, and deviation from Hardy-Weinberg equilibrium smaller than 1×10^{-6}, leaving 32,588,339 autosomal alleles for downstream analysis. After that, GWAS analyses were performed using the mixed linear model with the option of leaving one chromosome out (--mlma-loco) implemented in GCTA software[87] for each quantitative trait. This GCTA-LOCO approach[91] provides a more robust association estimate by employing a genetic relatedness matrix (GRM) to account for genomic relationships, and the Leave One Chromosome Out (LOCO) method to control for proximal contamination[93]. The data were adjusted for covariates including age, sex, and the first four ancestral clusters from FRAPPE[31] results. We further employed a deep neural network of DeepNull[94] to model and account for potential non-linear or interactive effects among phenotypic data and their covariates. This method allows one to control for type I errors while enhancing phenotypic prediction[94]. The genome-wide significance thresholds (5.13×10^{-8}) were determined using a uniform threshold of 1/n, where n is the effective number of independent variants calculated using the Genetic type 1 Error Calculator (v.0.2)[95]. This method utilized a divide-and-conquer algorithm to speed up the calculation of correlations between the genetic markers and ultimately gave the effective number of independent markers (Me) that in weak LD. The proportion of variance in the phenotype explained by a given SNP (PVE) was estimated using the formula from Shim et al.[96].

Analyze statistical power of GWAS

The statistical power of GWAS was determined by the non-centrality parameter (NCP) of the χ^2 test statistic[97], i.e., $NCP = n2f(1-f)b^2/(1-2f(1-f)b^2)$, where b is the per allele effect size, f is the minor allele frequency, and n is the sample size. We fixed n to 875 which equals the number of captive macaques that possess phenotypic data and sampling f from uniform distribution between 0.01 and 0.5. The effect size b was drawn from $N(0, h^2/m)$, where h^2 represents the mean estimated heritability across traits (e.g. 0.5) and m (the number of causal variants) was assumed to be 100. To estimate power, we ran 10,000 simulations and calculated the proportion of tests surpassing the genome-wide significance threshold of 5.13×10^{-8}.

Behavioral and brain imaging experimentation on CRMs with *DISC1* mutation p.Arg517Trp

Animals. We identified 3 adult samples (ages 5–7 years, two male and one female) and 5 elderly samples (ages >19 years, all female) harboring the homozygous missense mutations (p.Arg517Trp) in the cohort. Considering the old age of some of the monkeys, and our inability to eliminate the potential influence of aging on the results obtained, we performed the behavioral and brain imaging detection specifically on the three younger adult samples. All animal experimental procedures were approved by the Institutional Animal Care and Use Committee (IACUC) of KIZ, CAS (IACUC-PE-2022-07-001).

Behavioral experiments. We first estimated the neurological function of 3 homozygous carriers *vs.* 19 non-carriers using a neurological deficit score developed in our previous study (Supplementary Table 7)[98]. This scoring system assigned points to three aspects of neurological function: the motor system (16 points), skeletal muscle coordination (9 points) and the sensory system (25 points), totaling a maximum of 50 points. A score of 0 indicated normal behaviors whereas higher scores reflected neurological deficits. Next, we performed a spatial working memory test using the WGTA (Wisconsin General Test Apparatus) that modified from our previous studies[99,100]. Considering the significant amount of time required for the training and experimental stages, we selected three non-carriers, who were of similar age and gender as the controls. Briefly, the macaque was allowed to choose food (e.g., peanut) from one of the two covered wells with six time delays (0 s, 6 s, 12 s, 18 s, 24 s, 30 s; Fig. 5a). The delays were semi-randomly distributed over the trials with totaling 36 trials conducted in one session. We performed one session per day for each macaque and 10 sessions were performed. To investigate the spatial working memory under stress, restraint stress was performed by fixing the macaque in a narrow space in its home cage for 30 min, then working memory was tested immediately after the stress. The next session was conducted after a recovery interval of at least three days when the macaque attained the average performance level without stress. Three trials were performed for each macaque under stress. The inhibition of working memory was obtained using the formula of ((Pre - Post stress)/(Pre + Post stress)) × 100. Differences of the behavioral performance were estimated by unpaired *t*-test.

Brain imaging. Magnetic resonance imaging (MRI) and resting state functional MRI (rs-fMRI) data were acquired with a 3.0 T UMR790 MRI scanner (United Imaging, Shanghai, China) at KIZ. T1-weighted images were acquired using a 3D T1-weighted fast spoiled gradient echo (gre_fsp) sequence (voxel size = 0.5 mm isotropic, TE = 5.6 ms, TR = 13.01 ms, flip angle: 8°) and T2-weighted images were acquired using a fse_mx sequence (voxel size = 0.5 mm isotropic, TE = 396.48 ms, TR = 3400 ms, flip angle: 59°) by using a 12-channel head coil. The structural data were processed using Analysis of Functional NeuroImages software (AFNI)[101], FMRIB Software Library (FSL)[102], Advanced Normalization Tools (ANTs)[103] and FreeSurfer[104] (see details in Supplementary materials). Rs-fMRI images were collected using an echo planar imaging (EPI) sequence (voxel size = 1.5 mm isotropic, TE = 29 ms, TR = 1700 ms, flip angle: 80°). During rs-fMRI scanning, macaques were placed under the general anesthesia, similar to structural imaging, to alleviate stress and minimize motion artifacts. Note that resting-state functional activity is an inherent characteristic of the brain, observed in both humans and macaques, even under anesthesia[105,106]. The rs-fMRI data preprocessing was performed using the workflow outlined in a previous study[107] (see details in Supplementary materials).

Quantification and statistical analysis

Mann–Whitney U test was used to compare the phenotype difference between the pLoF allele carriers and non-carriers. Two-tailed Student's *t*-test were used to determine the significance of behavioral difference between *DISC1* (p.Arg517Trp) carriers and controls. Structural difference at the global, lobe, and region levels were conducted under Generalized Linear Mixed Models (GLMMs), using Hemisphere as the random factor, and all structural data were corrected with the intracranial volume of the corresponding hemisphere. Other statistical analyses can be found in the relevant sections of the method details, also have given in figure legends and supplementary tables.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data needed to evaluate the conclusions in the paper are present in either the paper and/or the Supplementary Materials. The raw whole genomic sequencing data generated in this study have been deposited in the Genome Sequence Archive (GSA) of National Genomics Data Center under accession number CRA014717. The raw phenotypic data and the summary statistics of GWAS results can be download from Non-Human Primate BioBank database (https://nhpbiobank.kiz.ac.cn/Home/Download). Source data are provided with this paper.

Code availability

No specific custom codes were developed in this study. All commands and pipelines used for data analyses were conducted according to the manuals or protocols provided by the corresponding software development team, which are described in detail in the Methods section. Default parameters were employed if no detailed parameters were mentioned for the software used in this study.

References

1. Chiou, K. L. et al. Rhesus macaques as a tractable physiological model of human ageing. *Philos. Trans. R. Soc. B* **375**, 20190612 (2020).
2. Gardner, M. B. & Luciw, P. A. Macaque models of human infectious disease. *ILAR J.* **49**, 220–255 (2008).
3. Pan, M. T., Zhang, H., Li, X. J. & Guo, X. Y. Genetically modified non-human primate models for research on neurodegenerative diseases. *Zool. Res.* **45**, 263–274 (2024).
4. Yuan, L. Z. et al. SARS-CoV-2 infection and disease outcomes in non-human primate models: advances and implications. *Emerg. Microbes Infect.* **10**, 1881–1889 (2021).
5. Tian, C. Y. China is facing serious experimental monkey shortage during the COVID-19 lockdown. *J. Med. Primatol.* **50**, 225–227 (2021).
6. Reardon, S. Giant monkey facility could ease U.S. shortage. *Science* **383**, 803–804 (2024).
7. Rogers, J. Genomic resources for rhesus macaques (*Macaca mulatta*). *Mamm. Genome* **33**, 91–99 (2022).
8. Wu, D. D. et al. Initiation of the primate genome project. *Zool. Res* **43**, 147–149 (2022).
9. Sanchez-Roige, S. & Palmer, A. A. Emerging phenotyping strategies will advance our understanding of psychiatric genetics. *Nat. Neurosci.* **23**, 475–480 (2020).
10. Timpson, N. J., Greenwood, C. M. T., Soranzo, N., Lawson, D. J. & Richards, J. B. Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nat. Rev. Genet.* **19**, 110–124 (2018).
11. Brekke, T. D., Steele, K. A. & Mulley, J. F. Inbred or outbred? Genetic diversity in laboratory rodent colonies. *G3-Genes Genomes Genet.* **8**, 679–686 (2018).
12. Tosi, A. J., Morales, J. C. & Melnick, D. J. Paternal, maternal, and biparental molecular markers provide unique windows onto the evolutionary history of macaque monkeys. *Evolution* **57**, 1419–1435 (2003).
13. Roos, C. & Zinner, D. The nonhuman primate in nonclinical drug development and safety assessment. in *Diversity and evolutionary history of macaques with special focus on Macaca mulatta and Macaca fascicularis* (eds. Joerg, B., Sven, K., Emanuel, S. & Gerhard, F.W.) 3-16 (Elsevier, 2015).
14. Morales, J. C. & Melnick, D. J. Phylogenetic relationships of the macaques (Cercopithecidae: *Macaca*), as revealed by high resolution restriction site mapping of mitochondrial ribosomal genes. *J. Hum. Evol.* **34**, 1–23 (1998).
15. Srikulnath, K., Ahmad, S. F., Panthum, T. & Malaivijitnond, S. Importance of Thai macaque bioresources for biological research and human health. *J. Med. Primatol.* **51**, 62–72 (2022).
16. Bimber, B. N., Yan, M. Y., Peterson, S. M. & Ferguson, B. mGAP: the macaque genotype and phenotype resource, a framework for accessing and interpreting macaque variant data, and identifying new models of human disease. *BMC Genomics* **20**, 176 (2019).
17. Xue, C. et al. The population genomics of rhesus macaques (*Macaca mulatta*) based on whole-genome sequences. *Genome Res.* **26**, 1651–1662 (2016).

18. Liu, Z. J. et al. Population genomics of wild Chinese rhesus macaques reveals a dynamic demographic history and local adaptation, with implications for biomedical research. *Gigascience* **7**, giy106 (2018).
19. Wu, R. F. et al. Landscape genomics analysis provides insights into future climate change-driven risk in rhesus macaque. *Sci. Total Environ.* **899**, 165746 (2023).
20. Yao, Y. G. Towards the peak: The 10-year journey of the National Research Facility for Phenotypic and Genetic Analysis of Model Animals (Primate Facility) and a call for international collaboration in non-human primate research. *Zool. Res.* **43**, 237–240 (2022).
21. Tarantino, L. M. & Eisener-Dorman, A. F. Forward genetic approaches to understanding complex behaviors. *Curr. Top. Behav. Neurosci.* **12**, 25–58 (2012).
22. Argmann, C.A., Dierich, A. & Auwerx, J. Uses of forward and reverse genetics in mice to study gene function. *Curr. Protoc. Mol. Biol*. Chapter **29**, Unit 29A 1 (2006).
23. Alonso, J. M. & Ecker, J. R. Moving forward in reverse: genetic technologies to enable genome-wide phenomic screens in *Arabidopsis*. *Nat. Rev. Genet.* **7**, 524–536 (2006).
24. Lehner, B. Genotype to phenotype: lessons from model organisms for human genetics. *Nat. Rev. Genet.* **14**, 168–178 (2013).
25. Takahashi, J. S., Pinto, L. H. & Vitaterna, M. H. Forward and reverse genetic approaches to behavior in the mouse. *Science* **264**, 1724–1733 (1994).
26. Adams, D. J. & van der Weyden, L. Contemporary approaches for modifying the mouse genome. *Physiol. Genomics* **34**, 225–238 (2008).
27. Lawson, N. D. & Wolfe, S. A. Forward and reverse genetic approaches for the analysis of vertebrate development in the zebrafish. *Developmental Cell* **21**, 48–64 (2011).
28. Adams, M. D. & Sekelsky, J. J. From sequence to phenotype: reverse genetics in *Drosophila melanogaster*. *Nat. Rev. Genet.* **3**, 189–198 (2002).
29. He, Y. et al. Long-read assembly of the Chinese rhesus macaque genome and identification of ape-specific structural variants. *Nat. Commun.* **10**, 4233 (2019).
30. Yan, G. et al. Genome sequencing and comparison of two nonhuman primate animal models, the cynomolgus and Chinese rhesus macaques. *Nat. Biotechnol.* **29**, 1019–1023 (2011).
31. Tang, H., Peng, J., Wang, P. & Risch, N. J. Estimation of individual admixture: analytical and study design considerations. *Genet. Epidemiol.* **28**, 289–301 (2005).
32. Barrett, S. C. & Charlesworth, D. Effects of a change in the level of inbreeding on the genetic load. *Nature* **352**, 522–524 (1991).
33. Kardos, M. et al. The crucial role of genome-wide genetic variation in conservation. *Proc. Natl. Acad. Sci. USA* **118**, e2104642118 (2021).
34. Bertorelle, G. et al. Genetic load: genomic estimates and applications in non-model animals. *Nat. Rev. Genet.* **23**, 492–503 (2022).
35. Halldorsson, B. V. et al. The sequences of 150,119 genomes in the UK Biobank. *Nature* **607**, 732–740 (2022).
36. Cong, P. K. et al. Genomic analyses of 10,376 individuals in the Westlake BioBank for Chinese (WBBC) pilot project. *Nat. Commun.* **13**, 2939 (2022).
37. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
38. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
39. MacArthur, D. G. & Tyler-Smith, C. Loss-of-function variants in the genomes of healthy humans. *Hum. Mol. Genet.* **19**, R125–R130 (2010).
40. MacArthur, D. G. et al. A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828 (2012).
41. Saleheen, D. et al. Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. *Nature* **544**, 235–242 (2017).
42. Cingolani, P. et al. 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 (2012).
43. Steux, C. & Szpiech, Z. A. The maintenance of deleteriousv variation in wild Chinese rhesus macaques. *Genome Biol. Evol.* **16**, evae115 (2024).
44. Blake, J. A. et al. Mouse Genome Database (MGD): Knowledgebase for mouse-human comparative biology. *Nucleic Acids Res* **49**, D981–D987 (2021).
45. Kleinridders, A. et al. PLRG1 Is an Essential Regulator of Cell Proliferation and Apoptosis during Vertebrate Development and Tissue Homeostasis. *Mol. Cell. Biol.* **29**, 3173–3185 (2009).
46. Liao, B. Y. & Zhang, J. Null mutations in human and mouse orthologs frequently result in different phenotypes. *Proc. Natl. Acad. Sci. USA* **105**, 6987–6992 (2008).
47. Xu, L. et al. Loss of RIG-I leads to a functional replacement with MDA5 in the Chinese tree shrew. *Proc. Natl. Acad. Sci. USA* **113**, 10950–10955 (2016).
48. Duggal, P., Gillanders, E. M., Holmes, T. N. & Bailey-Wilson, J. E. Establishing an adjusted p-value threshold to control the family-wide type 1 error in genome wide association studies. *BMC Genomics* **9**, 516 (2008).
49. Vermeer, S. et al. Targeted next-generation sequencing of a 12.5 Mb homozygous region reveals *ANO10* mutations in patients with autosomal-recessive cerebellar ataxia. *Am. J. Hum. Genet.* **87**, 813–819 (2010).
50. Nanetti, L. et al. ANO10 mutational screening in recessive ataxia: genetic findings and refinement of the clinical phenotype. *J. Neurol.* **266**, 378–385 (2019).
51. Jacobo-Baca, G. et al. Proteomic profile of preeclampsia in the first trimester of pregnancy. *J. Matern-Fetal Neo. M.* **35**, 3446–3452 (2022).
52. Alderton, G. K. et al. Seckel syndrome exhibits cellular features demonstrating defects in the ATR-signalling pathway. *Hum. Mol. Genet.* **13**, 3127–3138 (2004).
53. Benatzy, Y., Palmer, M. A. & Brune, B. Arachidonate 15-lipoxygenase type B: Regulation, function, and its role in pathophysiology. *Front. Pharmacol.* **13**, 1042420 (2022).
54. Huang, J. et al. Genomics and phenomics of body mass index reveals a complex disease network. *Nat. Commun.* **13**, 7973 (2022).
55. Tachmazidou, I. et al. Whole-genome sequencing coupled to imputation discovers genetic signals for anthropometric traits. *Am. J. Hum. Genet.* **100**, 865–884 (2017).
56. Li, L. et al. Interactions between genetic variants and environmental risk factors are associated with the severity of pelvic organ prolapse. *Menopause* **30**, 621–628 (2023).
57. Sollis, E. et al. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res* **51**, D977–D985 (2023).
58. Thompson, E. C. et al. Ikaros DNA-binding proteins as integral components of B cell developmental-stage-specific regulatory circuits. *Immunity* **26**, 335–344 (2007).
59. Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
60. Whitfield, J. B. et al. Biomarker and genomic risk factors for liver function test abnormality in hazardous drinkers. *Alcohol. Clin. Exp. Res.* **43**, 473–482 (2019).
61. Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M. & Ng, P. C. SIFT missense predictions for genomes. *Nat. Protoc.* **11**, 1–9 (2016).
62. McLaren, W. et al. The ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).

63. Park, J. E. & Silva, A. C. Generation of genetically engineered non-human primate models of brain function and neurological disorders. *Am. J. Primatol.* **81**, e22931 (2019).
64. Capitanio, J. P. & Emborg, M. E. Contributions of non-human primates to neuroscience research. *Lancet* **371**, 1126–1135 (2008).
65. Gamo, N. J. et al. Role of disrupted in schizophrenia 1 (*DISC1*) in stress-induced prefrontal cognitive dysfunction. *Transl. Psychiatry* **3**, e328 (2013).
66. Arnsten, A. F. Stress signalling pathways that impair prefrontal cortex structure and function. *Nat. Rev. Neurosci.* **10**, 410–422 (2009).
67. Li, H., Yao, Y. G. & Hu, X. T. Biological implications and limitations of a cynomolgus monkey with naturally occurring Parkinson's disease. *Zool. Res* **42**, 138–140 (2021).
68. Li, H. et al. A cynomolgus monkey with naturally occurring Parkinson's disease. *Natl. Sci. Rev.* **8**, nwaa292 (2021).
69. Sherman, L. S. et al. A novel non-human primate model of Pelizaeus-Merzbacher disease. *Neurobiol. Dis.* **158**, 105465 (2021).
70. Passingham, R. How good is the macaque monkey model of the human brain?. *Curr. Opin. Neurobiol.* **19**, 6–11 (2009).
71. Cannon, T. D. et al. Association of DISC1/TRAX haplotypes with schizophrenia, reduced prefrontal gray matter, and impaired short- and long-term memory. *Arch. Gen. Psychiatry.* **62**, 1205–1213 (2005).
72. Perlstein, W. M., Carter, C. S., Noll, D. C. & Cohen, J. D. Relation of prefrontal cortex dysfunction to working memory and symptoms in schizophrenia. *Am. J. Psychiatry* **158**, 1105–1113 (2001).
73. Li, J. et al. 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**, 3123 (2018).
74. Pritchard, J. K. Are rare variants responsible for susceptibility to complex diseases?. *Am. J. Hum. Genet.* **69**, 124–137 (2001).
75. Garg, M. et al. Disease prediction with multi-omics and biomarkers empowers case-control genetic discoveries in the UK Biobank. *Nat. Genet* **56**, 1821–1831 (2024).
76. Orliac, E. J. et al. Improving GWAS discovery and genomic prediction accuracy in biobank data. *Proc. Natl. Acad. Sci. USA* **119**, e2121279119 (2022).
77. Wong, A. K., Sealfon, R. S. G., Theesfeld, C. L. & Troyanskaya, O. G. Decoding disease: from genomes to networks to phenotypes. *Nat. Rev. Genet* **22**, 774–790 (2021).
78. Van der Auwera, G. A. et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* **43**, 11.10.1–11.10.33 (2013).
79. Warren, W. C. et al. Sequence diversity analyses of an improved rhesus macaque genome enhance its biomedical utility. *Science* **370**, eabc6617 (2020).
80. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
81. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).
82. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297–1303 (2010).
83. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
84. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
85. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
86. Reimand, J., Kull, M., Peterson, H., Hansen, J. & Vilo, J. g:Profiler-a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res* **35**, W193–W200 (2007).
87. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. Genome-wide complex trait analysis (GCTA): methods, data analyses, and interpretations. *Methods Mol. Biol.* **1019**, 215–236 (2013).
88. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
89. von Seth, J. et al. Genomic insights into the conservation status of the world's last remaining Sumatran rhinoceros populations. *Nat. Commun.* **12**, 2393 (2021).
90. Zhu, Q. et al. A genome-wide comparison of the functional properties of rare and common genetic variants in humans. *Am. J. Hum. Genet.* **88**, 458–468 (2011).
91. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.* **46**, 100–106 (2014).
92. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
93. Cheng, R., Parker, C. C., Abney, M. & Palmer, A. A. Practical considerations regarding the use of genotype and pedigree data to model relatedness in the context of genome-wide association studies. *G3 (Bethesda)* **3**, 1861–1867 (2013).
94. McCaw, Z. R. et al. DeepNull models non-linear covariate effects to improve phenotypic prediction and association power. *Nat. Commun.* **13**, 241 (2022).
95. Li, M. X., Yeung, J. M., Cherny, S. S. & Sham, P. C. Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Hum. Genet.* **131**, 747–756 (2012).
96. Shim, H. et al. A multivariate genome-wide association analysis of 10 LDL subfractions, and their response to statin treatment, in 1868 Caucasians. *PLoS One* **10**, e0120758 (2015).
97. Yang, J., Wray, N. R. & Visscher, P. M. Comparing apples and oranges: equating the power of case-control and quantitative trait association studies. *Genet Epidemiol.* **34**, 254–257 (2010).
98. Yang, L. et al. Extracellular vesicle-mediated delivery of circular RNA SCMH1 promotes functional recovery in rodent and nonhuman primate ischemic stroke models. *Circulation* **142**, 556–574 (2020).
99. Wang, J. H. et al. Interactive effects of morphine and dopaminergic compounds on spatial working memory in rhesus monkeys. *Neurosci. Bull.* **29**, 37–46 (2013).
100. Zhang, B. et al. Chronic phencyclidine treatment impairs spatial working memory in rhesus monkeys. *Psychopharmacol. (Berl.)* **236**, 2223–2232 (2019).
101. Cox, R. W. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* **29**, 162–173 (1996).
102. Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W. & Smith, S. M. Fsl. *Neuroimage* **62**, 782–790 (2012).
103. Avants, B. B., Tustison, N. & Song, G. Advanced normalization tools (ANTS). *Insight j.* **2**, 1–35 (2009).
104. Fischl, B. FreeSurfer. *Neuroimage* **62**, 774–781 (2012).
105. Vincent, J. L. et al. Intrinsic functional architecture in the anaesthetized monkey brain. *Nature* **447**, 83–86 (2007).
106. Larson-Prior, L. J. et al. Cortical network functional connectivity in the descent to sleep. *Proc. Natl. Acad. Sci. USA* **106**, 4489–7794 (2009).
107. Jo, H.J. et al. Effective preprocessing procedures virtually eliminate distance-dependent motion artifacts in resting state FMRI. *J. Appl. Math.* **2013**, 935154 (2013).
108. Zhou, Y. et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.* **10**, 1523 (2019).

Acknowledgements

We thank Amy F. T. Arnsten for invaluable input and discussions. We also thank staff members of the National Research Facility for Phenotypic & Genetic Analysis of Model Animals (Primate Facility) (https://cstr.cn/31137.02.NPRC) and the Core Technology Facility at Kunming Institute of Zoology, Chinese Academy of Sciences, for providing support in animal housing and care, MRI services. This work was funded by the National Key Research and Development Program of China (2022YFF0710900), the STI2030-Major Projects (2021ZD0200901, 2022ZD0205100, 2021ZD0203900, 2021ZD0200200, and 2021ZD0204200), Yunnan Province (202305AH340006), Yunnan Revitalization Talent Support Program Young Talent Project to B.L.Z., Yunnan Revitalization Talent Support Program Yunling Scholar Project to D.-D.W. and Y.-G.Y., the CAS Light of West China Program (xbzg-zdsys-202213) to D.-D.W, Technology Talent and Platform Plan of Yunnan Province (202305AF150160), and Academician Expert Workstation of Yunnan Kunming (YSZJGZZ-2022063). This work was also supported by the Biodiversity Data Center of KIZ, CAS.

Author contributions

D.-D.W., J.W., N.L., and Y.-G.Y. conceived and supervised the project; B.-L.Z., Y.C., Yi Zhang, Y. Lu, Yijiang Li, W.X., and H.D. performed anthropometric body measurements; B.-L.Z., Y.C., and Y. Wu performed genetic association analyses; Yali Zhang, Yanling Li, H.-D.H., and J.W. performed behavioral experiments and statistical analyses; Yali Zhang, Y.Q., M.-H.Q., N.-H.C., and N.L. performed brain imaging and statistical analyses; Q.W. and Y. Wang performed hematological and biochemical measurements; X. You, Yijiang Li, Y. Wang, X. Yu, M.-M.Y., and L.L. provided the blood sample; P.Z. provided the macaque monkey for behavioral and brain imaging test; B.-L.Z. and Y.C. performed the overall analysis; B.-L.Z and D.-D.W. wrote the original draft; Y. Wu, D.N.C., and Y.-G.Y. reviewed and edited the paper; all authors discussed the results and commented on the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-63747-x.

Correspondence and requests for materials should be addressed to Yong-Gang Yao, Ning Liu, Jian-Hong Wang or Dong-Dong Wu.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

[1]State Key Laboratory of Genetic Evolution & Animal Models, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan, China. [2]National Research Facility for Phenotypic & Genetic Analysis of Model Animals (Primate Facility) and National Resource Center for Non-Human Primates, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan, China. [3]Yunnan Key Laboratory of Biodiversity Information, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan, China. [4]College of Life Sciences, University of Chinese Academy of Sciences, Beijing, China. [5]State Key Laboratory of Cognitive Science and Mental Health, Institute of Biophysics, Chinese Academy of Sciences, Beijing, China. [6]Institute of Rare Diseases, West China Hospital of Sichuan University, Chengdu, China. [7]Laboratory Animal Center, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China. [8]Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff, UK. [9]Yunnan Key Laboratory of Animal Models and Human Disease Mechanisms, and Yunnan Engineering Center on Brain Disease Models, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan, China. [10]KIZ-CUHK Joint Laboratory of Bioresources and Molecular Research in Common Diseases, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan, China. [11]Kunming Natural History Museum of Zoology, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan, China. [12]These authors contributed equally: Bao-Lin Zhang, Yongxuan Chen, Yali Zhang. ✉e-mail: yaoyg@mail.kiz.ac.cn; liuning@ibp.ac.cn; wangjh@mail.kiz.ac.cn; wudongdong@mail.kiz.ac.cn