Nucleic Acids Research, 2025, **53**, D1250–D1260
https://doi.org/10.1093/nar/gkae956
Advance access publication date: 29 October 2024
Database issue

VDGE: a data repository of variation database for gene-edited animals across multiple species

Wenwen Shi [1,†], Enhui Jin [2,3,4,†], Lu Fang [5,†], Yanling Sun [2,3,6,7], Zhuojing Fan [2,3], Junwei Zhu [2,3], Chengzhi Liang [4,5], Ya-Ping Zhang [8], Yong Q. Zhang [1,4,9,*], Guo-Dong Wang [4,8,*] and Wenming Zhao [2,3,4,*]

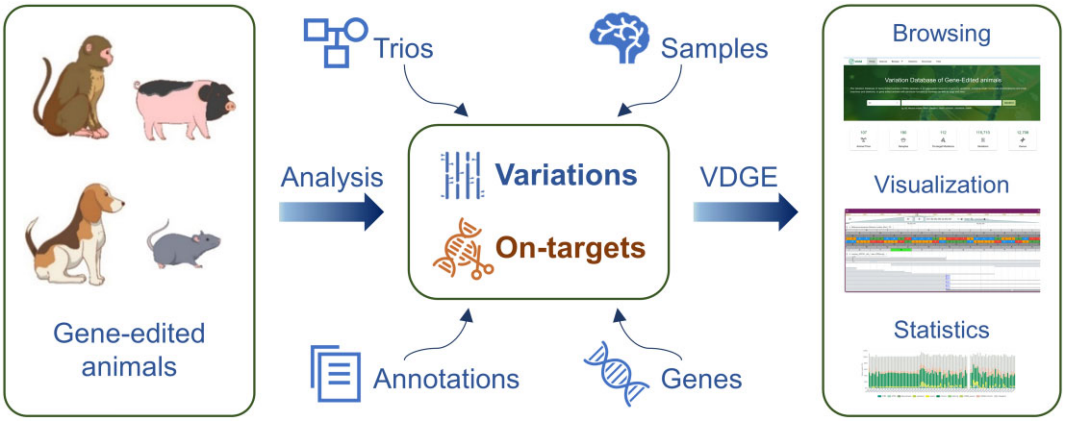
[1]State Key Laboratory of Molecular and Developmental Biology, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, No. 1 West Beichen Road, Chaoyang District, Beijing 100101, China
[2]National Genomics Data Center, China National Center for Bioinformation, No. 1 West Beichen Road, Chaoyang District, Beijing 100101, China
[3]Beijing Institute of Genomics, Chinese Academy of Sciences, No. 1 West Beichen Road, Chaoyang District, Beijing 100101, China
[4]University of Chinese Academy of Sciences, No.1 Yanqihu East Rd, Huairou District, Beijing 101408, China
[5]Key Laboratory of Seed Innovation, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, No. 1 West Beichen Road, Chaoyang District, Beijing 100101, China
[6]Lester and Sue Smith Breast Center, Baylor College of Medicine, One Baylor Plaza, Cambridge Street, Houston, TX 77030, USA
[7]Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Cambridge Street, Houston, TX 77030, USA
[8]Key Laboratory of Genetic Evolution and Animal Models, Yunnan Key Laboratory of Molecular Biology of Domestic Animals, Kunming Institute of Zoology, Chinese Academy of Sciences, No.17 Longxin Road, Panlong District, Kunming 650201, China
[9]School of Life Sciences, Hubei University, 368 Youyi Avenue, Wuchang District, Wuhan 430062, China

[*]To whom correspondence should be addressed. Tel: +86 1084097636; Fax: +86 1084097720; Email: zhaowm@big.ac.cn
Correspondence may also be addressed to Guo-Dong Wang. Email: wanggd@mail.kiz.ac.cn
Correspondence may also be addressed to Yong Q. Zhang. Email: yqzhang@genetics.ac.cn
[†]The first three authors should be regarded as Joint First Authors.

Abstract

Gene-edited animals are crucial for addressing fundamental questions in biology and medicine and hold promise for practical applications. In light of the rapid advancement of gene editing technologies over the past decade, a dramatically increased number of gene-edited animals have been generated. Genome editing at off-target sites can, however, introduce genomic variations, potentially leading to unintended functional consequences in these animals. So, there is an urgent need to systematically collect and collate these variations in gene-edited animals to aid data mining and integrative in-depth analyses. However, existing databases are currently insufficient to meet this need. Here, we present the Variation Database of Gene-Edited animals (VDGE, https://ngdc.cncb.ac.cn/vdge), the first open-access repository to present genomic variations and annotations in gene-edited animals, with a particular focus on larger animals such as monkeys. At present, VDGE houses 151 on-target mutations from 210 samples, and 115,710 variations identified from 107 gene-edited and wild-type animal trios through unified and standardized analysis and concurrently provides comprehensive annotation details for each variation, thus facilitating the assessment of their functional consequences and promoting mechanistic studies and practical applications for gene-edited animals.

Graphical abstract



Received: August 19, 2024. **Revised:** October 5, 2024. **Editorial Decision:** October 8, 2024. **Accepted:** October 10, 2024

Introduction

With the rapid advancement of the clustered regularly interspaced short palindromic repeat (CRISPR) gene editing technology (1,2), a dramatically increased number of gene-edited animals have been generated over the past decade. Large gene-edited animals such as monkeys, pigs and dogs are utilized increasingly in studies of human disease modelling, regenerative medicine and agricultural breeding (3–9). Gene editors can induce either DNA double-stranded breaks (DSBs) (2,10) or single-stranded breaks (SSBs) (11–14) at on- or off-target sites, potentially leading to genomic variations in gene-edited animals that differ from those in their parents and in non-gene-edited control animals (15–18). Off-target genome editing at unintended sites, extensively studied *in vitro* (19–22) and *in vivo* (23–26), poses one of the most significant challenges in both basic research and translational applications of gene editing (27,28).

This rising number of gene-edited animals being generated is accompanied by an escalating number of off-target events (15–18,29). The analysis of off-target effects is essential for phenotype analyses, safety assessments and the practical application of these gene-edited animals. From a practical perspective, there is an urgent need to understand the biological phenomena associated with variations caused by off-target gene editing, such as genomic locations, variation types (e.g. single-nucleotide variants [SNVs] or small insertions and deletions [INDELs]) and allele frequencies, and to assess their subsequent effects on phenotype and health. However, this information is scattered through different studies and various databases, making it difficult to gain clear insights into the off-target effects in these gene-edited animals. Towards an increasing understanding of the causes and mechanisms behind off-target effects (26,30,31), a variety of analytical methodologies have been employed to detect off-target events in such animals (15–18,29,32–36). Some studies have focused on examining algorithm-predicted off-target events in these animals (29,32) or comparing differences in genome variants between gene-edited and wild-type animals (29,33) without eliminating genetic variants inherited from parents. Other studies have identified whole-genome *de novo* variations based on deep sequencing data analysis of parent-offspring trios (8,34,35), which distinguish *de novo* variations from genetic background in these animals. Different analytical methods yield diverse results, which further increase the difficulty in achieving a comprehensive understanding of off-target effects. So, it is necessary to adopt unified and standardized methods for identifying genomic variations across different gene-edited animals. Deep whole-genome sequencing (WGS) data, especially when combined with parent-offspring trio analyses, provides a straightforward, universally applicable and more accurate approach to identifying whole-genome *de novo* variations in gene-edited animals.

Compared with a substantial increase in the generation of gene-edited animals and related omics data, integrated data resources and analyses for gene-editing associated variations remain scarce. A number of public variation databases have been established, such as dbSNP (37) and dbVar (38), the two major resources for archiving genome variations for humans; ClinVar (39), a public archive of human variations classified for diseases and drug responses; GVM (40,41) as a public repository of genome variations for a range of species and iDog (42) as an integrated resource recording variations for domestic and wild canids.

Most current variation databases focus on variations in humans or naturally born animals, with a notable absence of databases that focus on variations and their functional impacts in gene-edited animals. It is imperative to establish a publicly accessible platform for archiving, analysing and presenting genomic variations and their impacts in these animals. This process includes the collation of WGS data of gene-edited animals, the identification of whole-genome variations using standardized methods, the assessment of their functional impacts and the enabling of data sharing.

Here, we present a database named Variation Database of Gene-Edited animals (VDGE, https://ngdc.cncb.ac.cn/vdge), an open-access repository that curates and integrates genomic variation and annotation data of multiple gene-edited animal species. Currently, VDGE houses off-targeting analyses conducted through unified and standardized workflows for multiple animal species, including monkeys, pigs, dogs and mice, encompassing 151 on-target mutations from 210 samples, and 115,710 variations identified from 107 gene-edited and wild-type animal trios, with comprehensive annotation details for each variation. It provides a free, one-stop service for researchers involved in gene-edited animal studies for browsing, searching, visualizing and downloading information on variations.

Materials and methods

Data collection

The WGS data of gene-edited and wild-type animals were collected from published literature or generated by this group. The criteria used in the selection of data from literature published before 1 May 2024 included gene-edited or wild-type monkeys, pigs, dogs and mice and their parents with WGS data and the coverage of WGS data had to exceed 20×. A total of 70 WGS samples of monkeys, 65 WGS samples of dogs and 39 WGS samples of mice that met the aforementioned criteria were selected for further data processing as shown in Table 1 (8,15,17,34,43–46). These 174 WGS samples were derived from 107 animal trios, where each trio was composed of three samples: one individual with its two parents. Raw sequence data were downloaded from the Sequence Read Archive (SRA) (47) and the Genome Sequence Archive (GSA) (48), as seen in Table 1.

In addition to the gene-edited animals with trio data available, there is a vast array of those lacking trio information. Most related studies only present data on the on-target sites. We have collated a small number of these studies into the current database version, including 36 samples and 95 on-target mutations (49–58), as seen in Table 1. One such on-target gene is *myostatin* (*MSTN*) gene in pigs and dogs. MSTN is a negative regulator of skeletal muscle mass, which is closely related to the meat yield of agricultural animals (49,53).

The reference genomes used in this study included the rhesus monkey (*Macaca mulatta*: Mmul_10, GCF_003339765.1) (59), pig (*Sus scrofa*: Sscrofa11.1, GCF_000003025.6) (60), crab-eating macaque (*Macaca fascicularis*: MFA1912RKSv2, GCF_012559485.2) (61), dog (*Canis lupus familiaris*: Dog10K_Boxer_Tasha, GCF_000002285.5) (62) and mouse (*Mus musculus:* GRCm39, GCF_000001635.27) (63). Ref-

Table 1. Data summary of the VDGE database

Species	Gene editors	Animal trios	Samples	On-target mutations	SNVs	INDELs	Genes	WGS data source
Macaca mulatta	SpCas9	7	10	3	413	150	226	NCBI SRA
	NA	8	18	0	163	109	153	NCBI SRA
Macaca fascicularis	SpCas9	1	3	1	36	7	24	NCBI SRA
	BE4max	27	33	14	107,452	357	9,182	NGDC GSA
	NA	4	6	0	181	320	319	NGDC GSA
Sus scrofa	ZFN	0	4	4	0	0	1	-
	TALEN	0	2	4	0	0	1	-
	SpCas9	0	5	7	0	0	1	-
	BE3	0	5	41	0	0	3	-
Canis lupus familiaris	SpCas9	17	28	23	481	2,121	1,326	NGDC GSA
	NA	14	39	0	399	249	362	NGDC GSA
Mus musculus	SpCas9	7	20	31	201	90	124	NCBI SRA
	BE3	0	7	13	0	0	0	-
	BE4	9	9	10	1,925	193	988	NCBI SRA
	NA	13	21	0	578	285	395	NCBI SRA

Note: GSA, Genome Sequence Archive; INDEL, small insertion and deletion; NCBI, National Center for Biotechnology Information; NGDC, National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences; SRA, Sequence Read Archive; SNV, single-nucleotide variant; TALEN, transcription activator-like effector nuclease; WGS, whole-genome sequencing; ZFN, zinc finger nucleases; NA, not applicable as they are wild-type animals; -, no WGS data available.

erence genomes and corresponding annotation feature files were downloaded from the NCBI Genome (https://www.ncbi.nlm.nih.gov/datasets/genome) (64). Gene Ontology (GO) information of various species was downloaded from AmiGO 2 (https://amigo.geneontology.org/amigo) (65).

Data processing

Prior to detecting variations in gene-edited and wild-type animal trios, the genetic relationship between the offspring and their parents was first validated by analysing identity by descent (IBD) using PLINK1.9 software (66). Read sequences were then filtered using the fastp software (v 0.23.1) (67) with default parameters. The qualified short reads of all samples were mapped to reference genomes using the bwa-mem2 (v2.2.1) algorithm (68). Following the initial alignment, GATK (v4.2.6.1) (69) was used to sort the aligned BAM files and mark duplicate reads. Platypus (v0.8.1) (70) was employed to identify SNVs and INDELs from de-duplicated BAM files. We compared four variant calling tools, GATK (v 4.2.6.1) (69), Platypus (v0.8.1) (70), FreeBayes (v1.3.6) (71) and Strelka2 (v2.9.10) (72), and found that Platypus identified the highest number of known on-target mutations with a minimal occurrence of false positives compared to the other three tools. We, therefore, selected Platypus as the final variant calling tool.

The criteria for identifying *de novo* variations from variant call format (VCF) files of each animal trio called by Platypus were as follows: (i) The site filtering strategy filters variants at the site-level, taking variant quality by depth (QD), mapping quality (MQ), *P*-value for strand bias (SbPval), the number of forward reads (NF) and the number of reverse reads (NR) into consideration. (ii) The genotype quality (GQ) had to be no <40 for SNVs and no <5 for INDELs. (iii) For SNVs, the read depth was required to be between one-third and twice the average depth of each individual; for INDELs, the read depth was required to be between 5 and twice the average depth of each individual. (iv) The occurrence of *de novo* variations in the offspring was identified in accordance with the principles of Mendel's genetic law. This means that the *de novo* variation in the offspring was heterozygous (0/1), whereas the genotype at the identical locus from both parents was homozygous (0/0 or 1/1). (v) The number of reads supporting the alternative allele of the offspring in the parents had to be no >1. (vi) The variant allele frequency defined as alternative allele/alternative allele + reference allele in the offspring had to exceed 0.2. (vii) Any missing genotype mutations present in a sample were excluded.

To remove false positive mutations from the filtered *de novo* variations, IGV (v2.16.1) (73) and the tview module of Samtools (v1.6) (74) were employed to examine the reads associated with all *de novo* variations. This process was conducted on the cleaned BAM files. A maximum of one read carrying an alternative allele at the variant position in the parental alignment BAM files was permitted. The term alternative allele read was defined as a read that aligned with the corresponding alternative allele observed in the offspring.

Finally, annotated information including genic and intergenic regions, variant effects, variant consequences and genes were obtained by comparing the variations against NCBI resources (64) using ANNOVAR (75). Gene symbols and names were standardized to vertebrate gene nomenclature committee (VGNC) nomenclature (76). Gene groups were delineated based on HUGO Gene Nomenclature Committee (HGNC) resources (77). The GO terms were identified through the GO knowledge base (78).

Database implementation

The VDGE database was implemented using Spring Boot, a standalone Java application (https://spring.io/projects/spring-boot) as the backend framework. The frontend user interface was developed using Vue.js, an approachable and versatile framework for building web user interfaces (https://vuejs.org) and Quasar, an enterprise-ready cross-platform Vue.js framework (https://quasar.dev). All metadata were stored in MySQL, a free and popular relational database management system (https://www.mysql.com). For data visualization, JBrowse 2 (79) was used for secondary develop-

ment. The charts on the web page were constructed using Apache ECharts, an open-source JavaScript visualization library (https://echarts.apache.org).

Database content and usage

Overview of VDGE

VDGE is a comprehensive platform that presents genomic variations and their potential functional impacts in gene-edited animals, offering users a one-stop solution for data acquisition and analysis across multiple species (Figure 1). It adopts an integrated architecture comprising six pivotal modules including Species, Animal Trios, Samples, On-target Mutations, Variations and Genes. The Variations and On-target Mutations modules interface with other modules to catalogue every variation identified by a standardized analytical pipeline in both gene-edited and wild-type animal trios (see 'Materials and methods' section). In the current version, VDGE presents a total of 115,710 variations and 56 on-target mutations, identified from 174 samples of 107 animal trios across four distinct species (8,15,17,34,43–46) using the standardized analysis pipeline. VDGE also includes 95 on-target mutations derived from 36 gene-edited animal samples which lack trio deep sequencing data. This information was collected and curated manually from publicly available literature (49–58). Additionally, 12,708 genes associated with variations were incorporated into the database. Data statistical information is summarized in Table 1.

Species

The Species module organizes diverse data spanning different species, providing users with a comprehensive overview of the contents of the database. VDGE currently contains data for gene-edited animals from four large animal species including the rhesus monkey (*M. mulatta*), the crab-eating macaque (*M. fascicularis*), the pig (*S. scrofa*) and the domestic dog (*Canis lupus familiaris*), as well as the house mouse (*M. musculus*) as a classic animal model. The Species browsing page presents basic information for each species, including taxonomy ID, reference genome and genome size. It also incorporates concise statistical summaries of variations and related data including animal trios, samples, on-targets, variations and genes (Figure 2A), allowing users to navigate and access data resources for species of interest. For each species, VDGE offers a detailed page containing an overview and tabular displays of integrated data from other modules (Figure 2B).

Animal Trios and Samples

The Animal Trios and Samples modules provide detailed information on trios and samples from both gene-edited and wild-type animals. Currently, VDGE houses 15 *M. mulatta* trios, 32 *M. fascicularis* trios, 31 *Canis lupus familiaris* trios and 29 *M. musculus* trios (Table 1). Each trio includes three samples from father, mother and their offspring (80). Animal Trios module provides comprehensive data sets, which are derived from other modules, including on-targets, variations and genes. These data sets are closely linked to the offspring samples. Each animal trio is assigned a unique identifier (ID) prefixed with 'Tri', while each WGS sample is tagged with 'Samp'. These IDs serve as seamless navigational aids, enabling users to quickly access and explore data across diverse modules with ease and precision. The Animal Trios browsing page enables users to filter data of interest by species, gene editor and target gene, with the associated metadata presented in a tabular form. In addition, VDGE provides a detailed page for each animal trio, showing relevant information and tabular data from other modules. Similarly, the Samples browsing page permits the filtering of data based on sex, tissue, sequencing platform, sequencing depth and sequencing data access.

On-target Mutations and Variations

The On-target Mutations module, designated with the prefix 'On', archives on-target variations induced by various gene editors in gene-edited animals. In parallel, the Variations module, characterized by unique identifiers commencing with 'V', stores variations including *de novo* SNVs and INDELs occurring at unintended genomic sites in the offspring samples of each gene-edited or wild-type animal trio. A total of 115,710 variations were identified though a standardized analysis pipeline from 107 animal trios (Table 1), with the majority of SNVs were derived from cytosine base editor (CBE)-edited animal trios. It is important to note that not all gene-edited animals exhibit on-target editing, as exemplified by cases such as Tri012 and Tri049, where the gene editing may not always occur at the intended on-target site (45).

Both the On-target Mutations and Variations browsing pages empower users to quickly filter data of interest based on a range of criteria in the left panel, including species, gene editors, target genes and variation types (Figure 2C). The metadata, presented in a structured table format, ensures clarity and accessibility. VDGE goes a step further by offering dedicated, comprehensive pages for each on-target mutation and variation. These pages are replete with metadata, encompassing detail information such as variation type, position, reference and alternative allele, variant allele frequency, variant effect, associated gene and phenotype. To enhance the user experience, JBrowse visualization and related gene ontology information are seamlessly integrated, providing a holistic view of the data (Figure 2D).

Genes

To further evaluate the potential impacts of variations to gene-edited animals, each variation was annotated and subsequently mapped to the corresponding genes. All genes related to variations are included in the Genes module. A total of 12,708 genes were identified as being associated with all 115,710 variations in the database. Apparently, base editors (BEs) affect a much greater number of genes within an individual animal compared to *Streptococcus pyogenes* Cas9 (SpCas9) editors. The Genes page allows users to quickly filter data of interest based on species, gene symbol, position and gene type (e.g. coding or non-coding genes). By pinpointing the genes linked to variations, the impact of these genetic changes on the physiological status and health of gene-edited animals can be more readily evaluated.

Data browsing, retrieval and download

VDGE provides data statistics and corresponding module entrances on the homepage. The search box allows users to input any keyword to search through all the data in the database or select the appropriate module to quickly explore data of interest. For each module, VDGE provides user-friendly browsing, filtering and retrieval interfaces. Users can filter data based

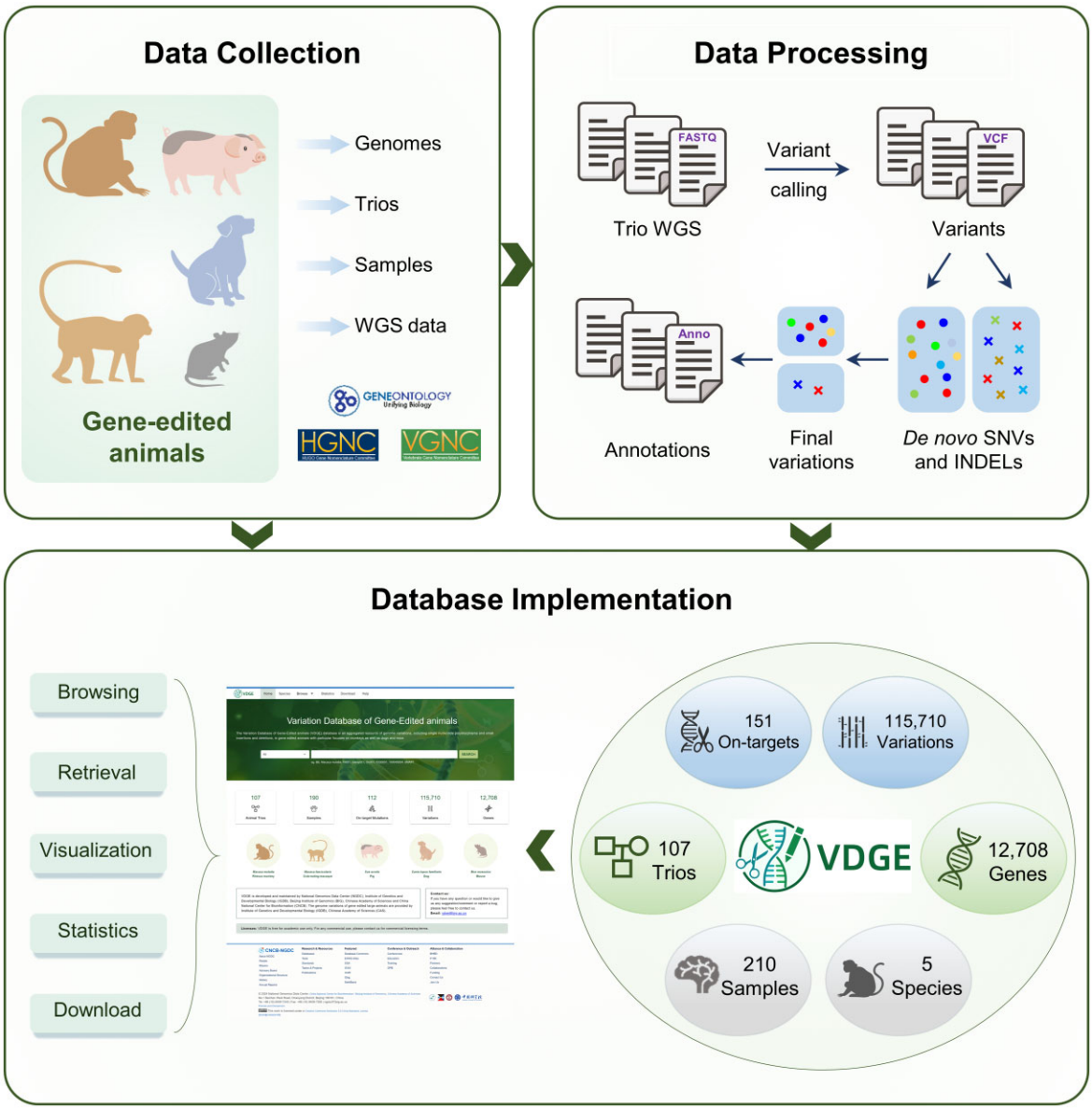


Figure 1. The construction pipeline of VDGE, including data collection, data processing and database implementation. INDEL, small insertion and deletion; SNV, single-nucleotide variant; WGS, whole-genome sequencing.

on common entries. The resulting filtered data are presented in a clear, tabular format, and each module is interconnected through IDs, allowing users to navigate to any other data of interest. To further enhance users' understanding, the Statistics module presents statistical metrics for select data, including variation number, gene region, SNV type, INDEL length and the distribution of variant allele frequency (Figure 2E). This statistical portrait facilitates the discovery of hidden data features.

To ensure data integrity for researchers to reuse, VDGE provides metadata and sequence data downloads. Users can click the Download button at the top right of each data table to download all metadata. For sequence data downloads, VDGE provides a separate Download module, which organizes all sequencing data, mapping data, SNV and INDEL data on a per-sample basis. Click on the link allows users to download the desired data from the provided FTP address.

An example of using VDGE

The CBEs are a major class of BEs, which are engineered fusions of Cas9 and a cytidine deaminase enzyme that retain the ability to be programmed with a guide RNA and mediate the direct conversion of cytidine to uridine, effecting a C to T (or G to A) substitution (11). These CBEs have been reported to induce substantial genome-wide off-target SNVs in mouse embryos and rice (16,30).

The information about BE-edited animal trios can be found in the Species module, by navigating to the Species button, which is prominently featured in the homepage's navigation bar (Figure 3A). A fourth-generation base editor, BE4 (81), was used to target the *Wap* gene in *Mus musculus* (17). Ten on-targets and 2,118 variations were identified from nine mouse trios using the standardized analysis pipeline. These variations were mapped to 988 genes (Figure 3B). The number of SNVs in each trio was found to be in the hundreds, which is consis-

Figure 2. Screenshots of VDGE. (**A**) Data summary in the Species module. (**B**) Detailed page for the *Macaca mulatta* species. (**C**) Variations module. (**D**) Details of the variation V0000003, including JBrowse visualization and related gene ontology information. (**E**) Statistics module.

tent with the findings of a previous study indicating that CBE treatment results in hundreds of genome-wide *de novo* SNVs in mouse embryos (16).

In parallel, BE4max, a variant of fourth-generation BEs (82), was employed to target the *LMNA* gene in *Macaca fascicularis* (45). Fourteen on-targets and 107,809 variations were identified from 27 monkey trios using the standardized analysis pipeline. These variations were mapped to 9,182 genes (Figure 3C). Navigating to the LMNA detail page and clicking on the Trio ID, information of a single trio can be viewed and monkey Tri010 was selected for further exploration. The detailed page for Tri010 displays the data source and information on the gene-edited animal and its parents. It shows that BE4max-mediated gene editing resulted in 9,069 d*e novo* variations in 3,407 genes in the gene-edited monkey. Related information for each sample can be viewed via external links (Figure 3D).

By clicking Browse button located in the upper right corner, the Variation browsing page is selected (Figure 3E). The filter bar on the left side of the page supports further filtering and viewing of the variations based on variation type, position, variant allele frequency, gene region and variant effect. Among the 9,069 d*e novo* variations, there are 9,060 SNVs and 9 INDELs, with the number of SNVs far exceeding that of INDELs (Figure 3E). On the Tri010 page, the Gene browsing page can be selected. The filter bar on the left side of the page allows for further filtering and viewing of the 3,407 genes related to the variations (Figure 3F). There are 2,770 protein-coding genes among all of the 3,407 genes. Variations in some of these genes may produce functional consequences for the gene-edited animals. For example, the variation V028639, which is mapped to the *caspase 2* (*CASP2*) gene, is a C to T point mutation (Figure 3G). The Variant Effect column shows that V028639 is a nonsynonymous SNV, resulting in the conversion of amino

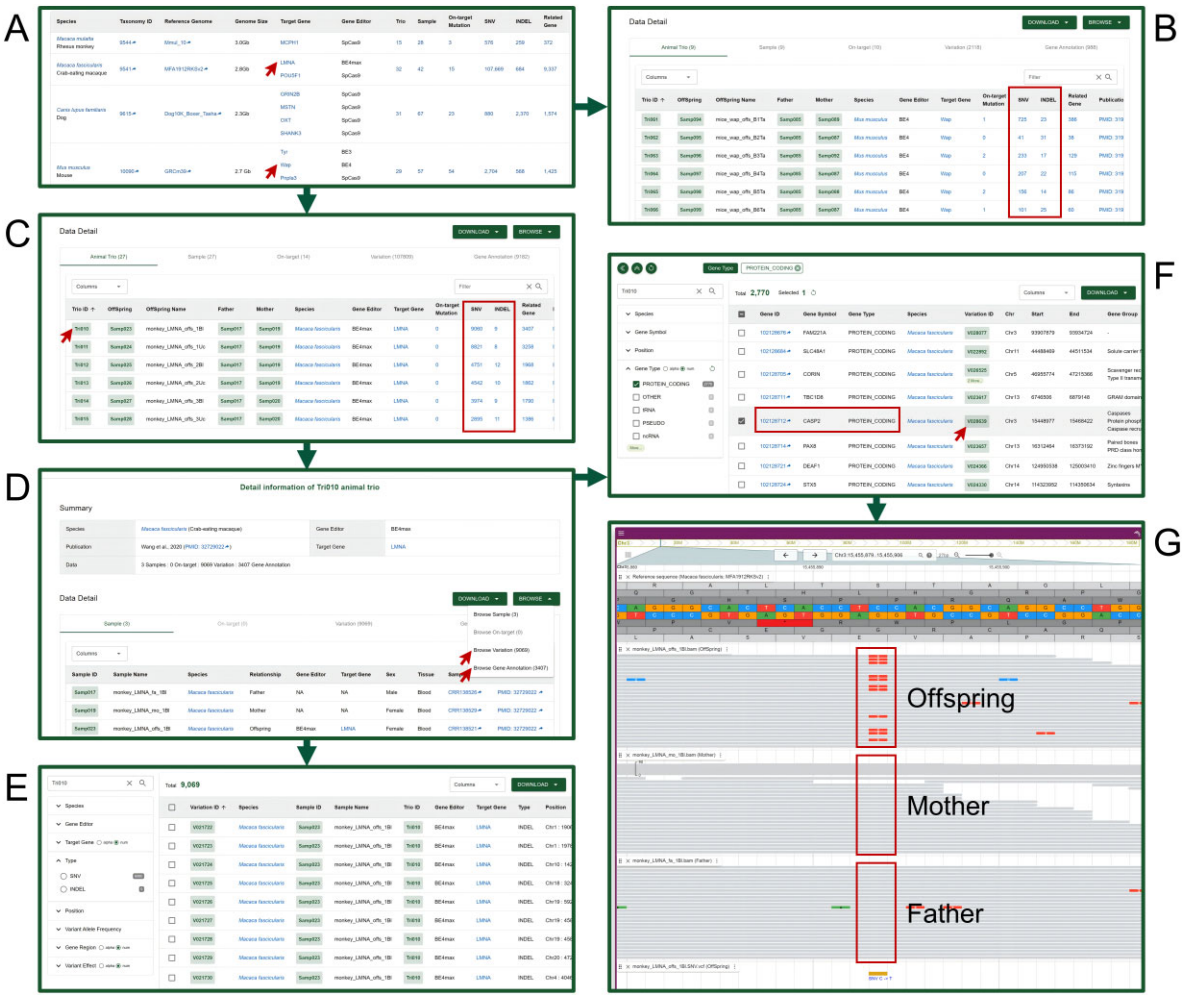


Figure 3. VDGE usage example. (**A**) Summarized panel for multiple species. (**B**) Mice trios edited by BE4. (**C**) Monkey trios edited by BE4max. The boxes indicate the number of SNVs and INDELs in each animal trio. (**D**) Detailed page for Tri010. Clicking on Browse Variation and Browse Gene Annotation buttons (indicated in D by arrows) will jump to Variations page (**E**) and annotated Genes page (**F**) of Tri010, respectively. (**G**) JBrowse visualization of *de novo* variation V028639 that is mapped to *caspase 2* (*CASP2*) gene (indicated by box). Clicking the buttons indicated by arrows in A, C, D and F will proceed to the subsequent page.

acids from glycine (G) to glutamic acid (E), as displayed in the Consequence column of the Variations module. Caspase 2 is a member of the caspases protein family, which is involved in cell death mediated by apoptosis, pyroptosis, necroptosis and autophagy (83). The functional disruption of Caspase 2 may affect the physiological status and health of the gene-edited animal (the offspring of the Tri010).

Discussion and future plans

As a revolutionary biotechnology, gene editing holds immense promise and generates great expectations for its potential to transform various fields of biological research and application. Many gene editing therapies are currently in clinical trials and the FDA has approved a CRISPR-based therapy called Casgevy to treat sickle cell disease and β-thalassemia (84,85). Gene-edited animals are also the subject of a variety of fundamental biological studies and potential commercial applications (4,86–88).

The safety of gene editing is a significant concern for both researchers and the public (89). Off-target events in gene editing are rare in some cases (34,35) but not in others (16,30). Different gene editors produce varying off-target effects in the same cell line and the same gene editor can exhibit different off-target effects in different cell types (26). Even an identical gene-editing system can lead to different off-target events in different individuals. So, it is necessary to conduct a detailed off-target analysis for each individual gene-edited animal. In this study, we focused on the analysis of *de novo* variations in gene-edited animals with trio WGS data. It should be noted that not all *de novo* variations are caused by off-target effects of gene editing; some of them could be spontaneous mutations, including *de novo* germline mutations and somatic mutations (90).

The VDGE database is the first repository to present genome variations and annotations in gene-edited animals, with a particular focus on larger animals that hold a great application value. VDGE exhibits the following key characteristics: (i) VDGE offers a user-friendly platform that facilitates the exploration of genomic variation and annotation information across multiple gene-edited animal species, with efficient data browsing, retrieval and downloading capabilities, making it a one-stop resource for researchers seeking information on variations and offering a valuable dataset for

the study of gene-edited animals. (ii) VDGE houses complete genomic variations for each animal trio by implementing a standardized analysis pipeline that leverages deep WGS data and parent-offspring trio analysis. (iii) VDGE provides an extensive dataset of variation-related information by integrating species, animal trios and on-target mutations, as well as annotation details such as variant type, genomic position, alternative allele and variant allele frequency, gene, gene ontology and potential functional consequence. These integrated data facilitate in-depth phenotype analysis, safety evaluations and translational studies for gene-edited animals. We recommend users to choose a computer browser to ensure the best user experience, even though our website is also compatible with mobile browsers.

In the future, VDGE will be maintained and updated by curating and integrating more on-target and variation information from gene-edited animals, particularly those lacking trio deep sequencing data. The range of organisms will be broadened by incorporating livestock, including cattle (86), sheep (91) and rabbits (92), as well as model organisms such as zebrafish (93), to meet the diverse needs of agricultural applications and scientific research. We also plan to enrich VDGE by incorporating variations identified by additional methodologies such as Cas-OFFinder (29), GUIDE-seq (15,23) and SITE-seq (18,21) to ensure a more comprehensive and accurate representation of variations present in each gene-edited animal. Moreover, a web server would be set up for the analysis pipeline used in this study. Once the new trio WGS data of gene-edited animals have been generated, the web server would facilitate the identification of variations. Furthermore, an upload functionality will be added to VDGE to allow users to submit on-target or variation information directly. A standardized data processing procedure will be implemented to ensure the consistency and accuracy of the data submitted by users and curated by VDGE staff.

Data availability

VDGE is available online for free at https://ngdc.cncb.ac.cn/vdge.

Acknowledgements

We thank Dr. Liang Wu and Dr. Tingting Chen for assistance in data collection and collation. We thank Dr. Huijuan Xu and Dr. Haiyang Hao for providing dog samples. We thank the National Genomics Data Center of China for providing the system operational environment.

Funding

Scientific and Technological Innovation 2030-Major Project [2021ZD0203900 to Y.P.Z., W.W.S. and G.D.W.]; National Key Research and Development Program of China [2023YFC2605700 to W.M.Z.; 2019YFA0707100 to Y.Q.Z., G.D.W.]; Strategic Priority Research Program of Chinese Academy of Sciences [XDB38050300 to W.M.Z.]; Yunnan Fundamental Research Projects [202201AV070011 to G.D.W.]; Spring City Plan: the High-level Talent Promotion and Training Project of Kunming [2022SCP001 to Y.Q.Z. and G.D.W.]. Funding for open access charge: Scientific and Technological Innovation 2030-Major Project.

Conflict of interest statement

The authors declare that they have no conflict of interest.

References

1. Jinek,M., Chylinski,K., Fonfara,I., Hauer,M., Doudna,J.A. and Charpentier,E. (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, **337**, 816–821.
2. Cong,L., Ran,F.A., Cox,D., Lin,S.L., Barretto,R., Habib,N., Hsu,P.D., Wu,X.B., Jiang,W.Y., Marraffini,L.A., *et al.* (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science*, **339**, 819–823.
3. Niu,Y.Y., Shen,B., Cui,Y.Q., Chen,Y.C., Wang,J.Y., Wang,L., Kang,Y., Zhao,X.Y., Si,W., Li,W., *et al.* (2014) Generation of gene-modified cynomolgus monkey via Cas9/RNA-mediated gene targeting in one-cell embryos. *Cell*, **156**, 836–843.
4. Whitworth,K.M., Rowland,R.R.R., Ewen,C.L., Trible,B.R., Kerrigan,M.A., Cino-Ozuna,A.G., Samuel,M.S., Lightner,J.E., McLaren,D.G., Mileham,A.J., *et al.* (2016) Gene-edited pigs are protected from porcine reproductive and respiratory syndrome virus. *Nat. Biotechnol.*, **34**, 20–22.
5. Barrangou,R. and Doudna,J.A. (2016) Applications of CRISPR technologies in research and beyond. *Nat. Biotechnol.*, **34**, 933–941.
6. Yan,S., Tu,Z., Liu,Z., Fan,N., Yang,H., Yang,S., Yang,W., Zhao,Y., Ouyang,Z., Lai,C., *et al.* (2018) A huntingtin knock-in pig model recapitulates features of selective neurodegeneration in Huntington's disease. *Cell*, **173**, 989–1002.
7. Zhao,J.G., Lai,L.X., Ji,W.Z. and Zhou,Q. (2019) Genome editing in large animals: current status and future prospects. *Natl. Sci. Rev.*, **6**, 402–420.
8. Tian,R., Li,Y., Zhao,H., Lyu,W., Zhao,J., Wang,X., Lu,H., Xu,H., Ren,W., Tan,Q.Q., *et al.* (2023) Modeling SHANK3-associated autism spectrum disorder in Beagle dogs via CRISPR/Cas9 gene editing. *Mol. Psychiatr.*, **28**, 3739–3750.
9. Pacesa,M., Pelea,O. and Jinek,M. (2024) Past, present, and future of CRISPR genome editing technologies. *Cell*, **187**, 1076–1100.
10. Hilton,I.B. and Gersbach,C.A. (2015) Enabling functional genomics with genome engineering. *Genome Res.*, **25**, 1442–1455.
11. Komor,A.C., Kim,Y.B., Packer,M.S., Zuris,J.A. and Liu,D.R. (2016) Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature*, **533**, 420–424.
12. Gaudelli,N.M., Komor,A.C., Rees,H.A., Packer,M.S., Badran,A.H., Bryson,D.I. and Liu,D.R. (2017) Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature*, **551**, 464–471.
13. Anzalone,A.V., Randolph,P.B., Davis,J.R., Sousa,A.A., Koblan,L.W., Levy,J.M., Chen,P.J., Wilson,C., Newby,G.A., Raguram,A., *et al.* (2019) Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature*, **576**, 149–157.
14. Anzalone,A.V., Koblan,L.W. and Liu,D.R. (2020) Genome editing with CRISPR-Cas nucleases, base editors, transposases and prime editors. *Nat. Biotechnol.*, **38**, 824–844.
15. Anderson,K.R., Haeussler,M., Watanabe,C., Janakiraman,V., Lund,J., Modrusan,Z., Stinson,J., Bei,Q.X., Buechler,A., Yu,C., *et al.* (2018) CRISPR off-target analysis in genetically engineered rats and mice. *Nat. Methods*, **15**, 512–514.
16. Zuo,E., Sun,Y., Wei,W., Yuan,T., Ying,W., Sun,H., Yuan,L., Steinmetz,L.M., Li,Y. and Yang,H. (2019) Cytosine base editor generates substantial off-target single-nucleotide variants in mouse embryos. *Science*, **364**, 289–292.
17. Lee,H.K., Smith,H.E., Liu,C., Willi,M. and Hennighausen,L. (2020) Cytosine base editor 4 but not adenine base editor generates off-target mutations in mouse embryos. *Commun. Biol.*, **3**, 19.
18. Burger,B.T., Beaton,B.P., Campbell,M.A., Brett,B.T., Rohrer,M.S., Plummer,S., Barnes,D., Jiang,K., Naswa,S., Lange,J., *et al.* (2024)

Generation of a commercial-scale founder population of porcine reproductive and respiratory syndrome virus resistant pigs using CRISPR-Cas. *CRISPR J*, **7**, 12–28.
19. Fu,Y.F., Foden,J.A., Khayter,C., Maeder,M.L., Reyon,D., Joung,J.K. and Sander,J.D. (2013) High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat. Biotechnol.*, **31**, 822–826.
20. Kim,D., Bae,S., Park,J., Kim,E., Kim,S., Yu,H.R., Hwang,J., Kim,J.-I. and Kim,J.-S. (2015) Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. *Nat. Methods*, **12**, 237–243.
21. Cameron,P., Fuller,C.K., Donohoue,P.D., Jones,B.N., Thompson,M.S., Carter,M.M., Gradia,S., Vidal,B., Garner,E., Slorach,E.M., *et al.* (2017) Mapping the genomic landscape of CRISPR-Cas9 cleavage. *Nat. Methods*, **14**, 600–606.
22. Tsai,S.Q., Nguyen,N.T., Malagon-Lopez,J., Topkar,V.V., Aryee,M.J. and Joung,J.K. (2017) CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR Cas9 nuclease off-targets. *Nat. Methods*, **14**, 607–614.
23. Tsai,S.Q., Zheng,Z., Nguyen,N.T., Liebers,M., Topkar,V.V., Thapar,V., Wyvekens,N., Khayter,C., Iafrate,A.J., Le,L.P., *et al.* (2015) GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat. Biotechnol.*, **33**, 187–197.
24. Wienert,B., Wyman,S.K., Richardson,C.D., Yeh,C.D., Akcakaya,P., Porritt,M.J., Morlock,M., Vu,J.T., Kazane,K.R., Watry,H.L., *et al.* (2019) Unbiased detection of CRISPR off-targets *in vivo* using DISCOVER-Seq. *Science*, **364**, 286–289.
25. Zou,R.S., Liu,Y., Gaido,O.E.R., Konig,M.F., Mog,B.J., Shen,L.L., Aviles-Vazquez,F., Marin-Gonzalez,A. and Ha,T. (2023) Improving the sensitivity of in vivo CRISPR off-target detection with DISCOVER-Seq. *Nat. Methods*, **20**, 706–713.
26. Zhu,M., Xu,R.D., Yuan,J.S., Wang,J.C., Ren,X.Y., Cong,T.T., You,Y.X., Ju,A.J., Xu,L.C., Wang,H.M., *et al.* (2024) Tracking-seq reveals the heterogeneity of off-target effects in CRISPR-Cas9-mediated genome editing. *Nat. Biotechnol.*, https://doi.org/10.1038/s41587-024-02307-y.
27. Doudna,J.A. (2020) The promise and challenge of therapeutic genome editing. *Nature*, **578**, 229–236.
28. Tsuchida,C.A., Brandes,N., Bueno,R., Trinidad,M., Mazumder,T., Yu,B., Hwang,B., Chang,C., Liu,J., Sun,Y., *et al.* (2023) Mitigation of chromosome loss in clinical CRISPR-Cas9-engineered T cells. *Cell*, **186**, 4567–4582.
29. Li,B., Zhao,H., Tu,Z.C., Yang,W.L., Han,R., Wang,L., Luo,X.P., Pan,M.T., Chen,X.S., Zhang,J.W., *et al.* (2023) CHD8 mutations increase gliogenesis to enlarge brain size in the nonhuman primate. *Cell Discov.*, **9**, 27.
30. Jin,S., Zong,Y., Gao,Q., Zhu,Z., Wang,Y., Qin,P., Liang,C., Wang,D., Qiu,J.-L., Zhang,F., *et al.* (2019) Cytosine, but not adenine, base editors induce genome-wide off-target mutations in rice. *Science*, **364**, 292–295.
31. Schep,R., Brinkman,E.K., Leemans,C., Vergara,X., van der Weide,R.H., Morris,B., van Schaik,T., Manzo,S.G., Peric-Hupkes,D., van den Berg,J., *et al.* (2021) Impact of chromatin context on Cas9-induced DNA double-strand break repair pathway balance. *Mol. Cell*, **81**, 2216–2230.
32. Zhang,W.Q., Wan,H.F., Feng,G.H., Qu,J., Wang,J.Q., Jing,Y.B., Ren,R.T., Liu,Z.P., Zhang,L.L., Chen,Z.G., *et al.* (2018) SIRT6 deficiency results in developmental retardation cynomolgus monkeys. *Nature*, **560**, 661–665.
33. Wang,X.M., Liang,Y.H., Zhao,J.P., Li,Y., Gou,S.X., Zheng,M., Zhou,J.J., Zhang,Q.J., Mi,J.D. and Lai,L.X. (2021) Generation of permanent neonatal diabetes mellitus dogs with glucokinase point mutations through base editing. *Cell Discov.*, **7**, 92.
34. Luo,X., He,Y., Zhang,C., He,X., Yan,L., Li,M., Hu,T., Hu,Y., Jiang,J., Meng,X., *et al.* (2019) Trio deep-sequencing does not reveal unexpected off-target and on-target mutations in Cas9-edited rhesus monkeys. *Nat. Commun.*, **10**, 5525.
35. Iyer,V., Boroviak,K., Thomas,M., Doe,B., Riva,L., Ryder,E. and Adams,D.J. (2018) No unexpected CRISPR-Cas9 off-target activity revealed by trio sequencing of gene-edited mice. *PLoS Genet.*, **14**, e1007503.
36. Xu,H.J., Hao,H.Y., Wang,S.R., Liu,X.R., Lyu,W., Zuo,Z.T., Zhuo,Y., Mi,J.D., Zhang,Y.Q., Tian,R., *et al.* (2023) A dog carrying mutations in AVP-NPII exhibits key features of central diabetes insipidus. *J. Genet. Genomics*, **50**, 280–283.
37. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
38. Lappalainen,I., Lopez,J., Skipper,L., Hefferon,T., Spalding,J.D., Garner,J., Chen,C., Maguire,M., Corbett,M., Zhou,G., *et al.* (2013) dbVar and DGVa: public archives for genomic structural variation. *Nucleic Acids Res.*, **41**, D936–D941.
39. Landrum,M.J., Lee,J.M., Benson,M., Brown,G., Chao,C., Chitipiralla,S., Gu,B., Hart,J., Hoffman,D., Hoover,J., *et al.* (2016) ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.*, **44**, D862–D868.
40. Li,C.P., Tian,D.M., Tang,B.X., Liu,X.N., Teng,X.F., Zhao,W.M., Zhang,Z. and Song,S.H. (2021) Genome Variation Map: a worldwide collection of genome variations across multiple species. *Nucleic Acids Res.*, **49**, D1186–D1191.
41. Bao,Y.M., Zhang,Z., Zhao,W.M., Xiao,J.F., He,S.M., Zhang,G.Q., Li,Y.X., Zhao,G.P., Chen,R.S., Bu,C.F., *et al.* (2024) Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2024. *Nucleic Acids Res.*, **52**, D18–D32.
42. Tang,B., Zhou,Q., Dong,L., Li,W., Zhang,X., Lan,L., Zhai,S., Xiao,J., Zhang,Z., Bao,Y., *et al.* (2019) iDog: an integrated resource for domestic dogs and wild canids. *Nucleic Acids Res.*, **47**, D793–d800.
43. Cui,Y.Q., Niu,Y.Y., Zhou,J.K., Chen,Y.C., Cheng,Y.W., Li,S.G., Ai,Z.Y., Chu,C., Wang,H., Zheng,B., *et al.* (2018) Generation of a precise Oct4-hrGFP knockin cynomolgus monkey model via CRISPR/Cas9-assisted homologous recombination. *Cell Res.*, **28**, 383–386.
44. Wang,R.J., Thomas,G.W.C., Raveendran,M., Harris,R.A., Doddapaneni,H., Muzny,D.M., Capitanio,J.P., Radivojac,P., Rogers,J. and Hahn,M.W. (2020) Paternal age in rhesus macaques is positively associated with germline mutation accumulation but not with measures of offspring sociability. *Genome Res.*, **30**, 826–834.
45. Wang,F., Zhang,W.Q., Yang,Q.Y., Kang,Y., Fan,Y.L., Wei,J.K., Liu,Z.P., Dai,S.X., Li,H., Li,Z.F., *et al.* (2020) Generation of a Hutchinson-Gilford progeria syndrome monkey model by base editing. *Protein Cell*, **11**, 809–824.
46. Zhang,S.J., Ma,J., Riera,M., Besenbacher,S., Niskanen,J.E., Salokorpi,N., Hundi,S., Hytönen,M.K., Zhou,T., Li,G.-M., *et al.* (2024) Determinants of *de novo* mutations in extended pedigrees of 43 dog breeds. bioRxiv doi: https://doi.org/10.1101/2024.06.04.596747, 05 June 2024, preprint: not peer reviewed.
47. Katz,K., Shutov,O., Lapoint,R., Kimelman,M., Brister,J.R. and O'Sullivan,C. (2022) The Sequence Read Archive: a decade more of explosive growth. *Nucleic Acids Res.*, **50**, D387–D390.
48. Chen,T., Chen,X., Zhang,S., Zhu,J., Tang,B., Wang,A., Dong,L., Zhang,Z., Yu,C., Sun,Y., *et al.* (2021) The Genome Sequence Archive family: toward explosive data growth and diverse data types. *Genom. Proteomics Bioinform.*, **19**, 578–583.
49. Qian,L.L., Tang,M.X., Yang,J.Z., Wang,Q.Q., Cai,C.B., Jiang,S.W., Li,H.G., Jiang,K., Gao,P.F., Ma,D.Z., *et al.* (2015) Targeted mutations in by zinc-finger nucleases result in double-muscled phenotype in Meishan pigs. *Sci. Rep.*, **5**, 14435.
50. Bi,H., Xie,S., Cai,C., Qian,L., Jiang,S., Xiao,G., Li,B., Li,X. and Cui,W. (2020) Frameshift mutation in myostatin gene by zinc-finger nucleases results in a significant increase in muscle mass in Meishan sows. *Czech J. Anim. Sci.*, **65**, 182–191.
51. Rao,S., Fujimura,T., Matsunari,H., Sakuma,T., Nakano,K., Watanabe,M., Asano,Y., Kitagawa,E., Yamamoto,T. and Nagashima,H. (2016) Efficient modification of the myostatin gene

in porcine somatic cells and generation of knockout piglets. *Mol. Reprod. Dev.*, **83**, 61–70.
52. Kang,J.D., Kim,S., Zhu,H.Y., Jin,L., Guo,Q., Li,X.C., Zhang,Y.C., Xing,X.X., Xuan,M.F., Zhang,G.L., *et al.* (2017) Generation of cloned adult muscular pigs with myostatin gene mutation by genetic engineering. *RSC Adv.*, **7**, 12541–12549.
53. Tanihara,F., Takemoto,T., Kitagawa,E., Rao,S.B., Do,L.T.K., Onishi,A., Yamashita,Y., Kosugi,C., Suzuki,H., Sembon,S., *et al.* (2016) Somatic cell reprogramming-free generation of genetically modified pigs. *Sci. Adv.*, **2**, e1600803.
54. Hua,Z., Xu,K., Xiao,W., Shu,C., Li,N., Li,K., Gu,H., Zhu,Z., Zhang,L., Ren,H., *et al.* (2023) Dual single guide RNAs-mediating deletion of mature myostatin peptide results in concomitant muscle fibre hyperplasia and adipocyte hypotrophy in pigs. *Biochem. Biophys. Res. Commun.*, **673**, 145–152.
55. Song,R.G., Wang,Y., Zheng,Q.T., Yao,J., Cao,C.W., Wang,Y.F. and Zhao,J.G. (2022) One-step base editing in multiple genes by direct embryo injection for pig trait improvement. *Sci. China-Life Sci.*, **65**, 739–752.
56. Zou,Q.J., Wang,X.M., Liu,Y.Z., Ouyang,Z., Long,H.B., Wei,S., Xin,J.G., Zhao,B.T., Lai,S.S., Shen,J., *et al.* (2015) Generation of gene-target dogs using CRISPR/Cas9 system. *J. Mol. Cell Biol.*, **7**, 580–583.
57. Mizuno,S., Tra Thi Huong,D., Kato,K., Mizuno-Iijima,S., Tanimoto,Y., Daitoku,Y., Hoshino,Y., Ikawa,M., Takahashi,S., Sugiyama,F., *et al.* (2014) Simple generation of albino C57BL/6J mice with G291T mutation in the tyrosinase gene by the CRISPR/Cas9 system. *Mamm. Genome*, **25**, 327–334.
58. Kim,K., Ryu,S.M., Kim,S.T., Baek,G., Kim,D., Lim,K., Chung,E., Kim,S. and Kim,J.S. (2017) Highly efficient RNA-guided base editing in mouse embryos. *Nat. Biotechnol.*, **35**, 435–437.
59. Warren,W.C., Harris,R.A., Haukness,M., Fiddes,I.T., Murali,S.C., Fernandes,J., Dishuck,P.C., Storer,J.M., Raveendran,M., Hillier,L.W., *et al.* (2020) Sequence diversity analyses of an improved rhesus macaque genome enhance its biomedical utility. *Science*, **370**, eabc6617.
60. Warr,A., Affara,N., Aken,B., Beiki,H., Bickhart,D.M., Billis,K., Chow,W., Eory,L., Finlayson,H.A., Flicek,P., *et al.* (2020) An improved pig reference genome sequence to enable pig genetics and genomics research. *GigaScience*, **9**, giaa051.
61. Jayakumar,V., Nishimura,O., Kadota,M., Hirose,N., Sano,H., Murakawa,Y., Yamamoto,Y., Nakaya,M., Tsukiyama,T., Seita,Y., *et al.* (2021) Chromosomal-scale *de novo* genome assemblies of Cynomolgus Macaque and Common Marmoset. *Sci. Data*, **8**, 159.
62. Jagannathan,V., Hitte,C., Kidd,J.M., Masterson,P., Murphy,T.D., Emery,S., Davis,B., Buckley,R.M., Liu,Y.-H., Zhang,X.-Q., *et al.* (2021) Dog10K_Boxer_Tasha_1.0: a long-read assembly of the dog reference genome. *Genes*, **12**, 847.
63. Church,D.M., Schneider,V.A., Graves,T., Auger,K., Cunningham,F., Bouk,N., Chen,H.C., Agarwala,R., McLaren,W.M., Ritchie,G.R., *et al.* (2011) Modernizing reference genome assemblies. *PLoS. Biol.*, **9**, e1001091.
64. Sayers,E.W., Bolton,E.E., Brister,J.R., Canese,K., Chan,J., Comeau,D.C., Farrell,C.M., Feldgarden,M., Fine,A.M., Funk,K., *et al.* (2023) Database resources of the National Center for Biotechnology Information in 2023. *Nucleic Acids Res.*, **51**, D29–D38.
65. Carbon,S., Douglass,E., Good,B.M., Unni,D.R., Harris,N.L., Mungall,C.J., Basu,S., Chisholm,R.L., Dodson,R.J., Hartline,E., *et al.* (2021) The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.*, **49**, D325–D334.
66. Chang,C.C., Chow,C.C., Tellier,L.C.A.M., Vattikuti,S., Purcell,S.M. and Lee,J.J. (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, **4**, 7.
67. Chen,S.F., Zhou,Y.Q., Chen,Y.R. and Gu,J. (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **34**, 884–890.
68. Vasimuddin,M., Misra,S., Li,H. and Aluru,S. (2019) Efficient architecture-aware acceleration of BWA-MEM for multicore systems. *Int. Parall. Distrib. P.*, **33**, 314–324.
69. McKenna,A., Hanna,M., Banks,E., Sivachenko,A., Cibulskis,K., Kernytsky,A., Garimella,K., Altshuler,D., Gabriel,S., Daly,M., *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.
70. Rimmer,A., Phan,H., Mathieson,I., Iqbal,Z., Twigg,S.R.F., Wilkie,A.O.M., McVean,G., Lunter,G. and WGS500 Consortium (2014) Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.*, **46**, 912–918.
71. Garrison,E. and Marth,G. (2012) Haplotype-based variant detection from short-read sequencing. arXiv doi: https://doi.org/10.48550/arXiv.1207.3907, 17 July 2012, preprint: not peer reviewed.
72. Kim,S., Scheffler,K., Halpern,A.L., Bekritsky,M.A., Noh,E., Källberg,M., Chen,X.Y., Kim,Y., Beyter,D., Krusche,P., *et al.* (2018) Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods*, **15**, 591–594.
73. Robinson,J.T., Thorvaldsdottir,H., Turner,D. and Mesirov,J.P. (2023) igv.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). *Bioinformatics*, **39**, btac830.
74. Danecek,P., Bonfield,J.K., Liddle,J., Marshall,J., Ohan,V., Pollard,M.O., Whitwham,A., Keane,T., McCarthy,S.A., Davies,R.M., *et al.* (2021) Twelve years of SAMtools and BCFtools. *Gigascience*, **10**, giab008.
75. Wang,K., Li,M.Y. and Hakonarson,H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.
76. Jones,T.E.M., Yates,B., Braschi,B., Gray,K., Tweedie,S., Seal,R.L. and Bruford,E.A. (2023) The VGNC: expanding standardized vertebrate gene nomenclature. *Genome Biol.*, **24**, 115.
77. Seal,R.L., Braschi,B., Gray,K., Jones,T.E.M., Tweedie,S., Haim-Vilmovsky,L. and Bruford,E.A. (2023) Genenames.org: the HGNC resources in 2023. *Nucleic Acids Res.*, **51**, D1003–D1009.
78. Aleksander,S.A., Balhoff,J., Carbon,S., Cherry,J.M., Drabkin,H.J., Ebert,D., Feuermann,M., Gaudet,P., Harris,N.L., Hill,D.P., *et al.* (2023) The Gene Ontology knowledgebase in 2023. *Genetics*, **224**, iyad031.
79. Diesh,C., Stevens,G.J., Xie,P., Martinez,T.D.J., Hershberg,E.A., Leung,A., Guo,E., Dider,S., Zhang,J., Bridge,C., *et al.* (2023) JBrowse 2: a modular genome browser with views of synteny and structural variation. *Genome Biol.*, **24**, 74.
80. Bergeron,L.A., Besenbacher,S., Zheng,J., Li,P., Bertelsen,M.F., Quintard,B., Hoffman,J.I., Li,Z., St. Leger,J., Shao,C., *et al.* (2023) Evolution of the germline mutation rate across vertebrates. *Nature*, **615**, 285–291.
81. Komor,A.C., Zhao,K.T., Packer,M.S., Gaudelli,N.M., Waterbury,A.L., Koblan,L.W., Kim,Y.B., Badran,A.H. and Liu,D.R. (2017) Improved base excision repair inhibition and bacteriophage Mu Gam protein yields C:g-to-T: a base editors with higher efficiency and product purity. *Sci. Adv.*, **3**, eaao4774.
82. Koblan,L.W., Doman,J.L., Wilson,C., Levy,J.M., Tay,T., Newby,G.A., Maianti,J.P., Raguram,A. and Liu,D.R. (2018) Improving cytidine and adenine base editors by expression optimization and ancestral reconstruction. *Nat. Biotechnol.*, **36**, 843–846.
83. Shalini,S., Dorstyn,L., Dawar,S. and Kumar,S. (2015) Old, new and emerging functions of caspases. *Cell Death Differ.*, **22**, 526–539.
84. Frangoul,H., Altshuler,D., Cappellini,M.D., Chen,Y.S., Domm,J., Eustace,B.K., Foell,J., de la Fuente,J., Grupp,S., Handgretinger,R., *et al.* (2021) CRISPR-Cas9 gene editing for sickle cell disease and β-thalassemia. *N. Engl. J. Med.*, **384**, 252–260.
85. Badwal,A.K. and Singh,S. (2024) A comprehensive review on the current status of CRISPR based clinical trials for rare diseases. *Int. J. Biol. Macromol.*, **277**, 134097.

86. Carlson,D.F., Lancto,C.A., Zang,B., Kim,E.S., Walton,M., Oldeschulte,D., Seabury,C., Sonstegard,T.S. and Fahrenkrug,S.C. (2016) Production of hornless dairy cattle from genome-edited cell lines. *Nat. Biotechnol.*, **34**, 479–481.
87. Harrison,C. (2022) CRISPR beef cattle get FDA green light. *Nat. Biotechnol.*, **40**, 448–448.
88. Qin,Y., Li,S., Li,X.J. and Yang,S. (2022) CRISPR-based genome-editing tools for Huntington's disease research and therapy. *Neurosci. Bull.*, **38**, 1397–1408.
89. Ledford,H. (2023) Is CRISPR safe? Genome editing gets its first FDA scrutiny. *Nature*, **623**, 234–235.
90. Noyes,M.D., Harvey,W.T., Porubsky,D., Sulovari,A., Li,R., Rose,N.R., Audano,P.A., Munson,K.M., Lewis,A.P., Hoekzema,K., *et al.* (2022) Familial long-read sequencing increases yield of *de novo* mutations. *Am. J. Hum. Genet.*, **109**, 631–646.
91. Wang,X.L., Liu,J., Niu,Y.Y., Li,Y., Zhou,S.W., Li,C., Ma,B.H., Kou,Q.F., Petersen,B., Sonstegard,T., *et al.* (2018) Low incidence of SNVs and indels in trio genomes of Cas9-mediated multiplex edited sheep. *BMC Genomics [Electronic Resource]*, **19**, 397.
92. Song,Y., Yuan,L., Wang,Y., Chen,M., Deng,J., Lv,Q., Sui,T., Li,Z. and Lai,L. (2016) Efficient dual sgRNA-directed large gene deletion in rabbit with CRISPR/Cas9 system. *Cell. Mol. Life Sci.*, **73**, 2959–2968.
93. Hoijer,I., Emmanouilidou,A., Ostlund,R., van Schendel,R., Bozorgpana,S., Tijsterman,M., Feuk,L., Gyllensten,U., den Hoed,M. and Ameur,A. (2022) CRISPR-Cas9 induces large structural variants at on-target and off-target sites in vivo that segregate across generations. *Nat. Commun.*, **13**, 627.

Received: August 19, 2024. **Revised:** October 5, 2024. **Editorial Decision:** October 8, 2024. **Accepted:** October 10, 2024