Nucleic Acids Research, 2026, **54**, D1208–D1215
https://doi.org/10.1093/nar/gkaf1207
Advance access publication date: 8 December 2025
Database issue

Genome Variation Map: a platform for the analysis and integration of genomic variation

Xue Bai[1,2,†], **Haixia Xie**[1,2,3,†], **Hong Luo**[1,2,†], **Siwei Ren**[1,2,3,†], **Bixia Tang**[1,2], **Cuiping Li**[1,2], **Yi Wang**[1,2,3], **Bo Xu**[1,2], **Zishan Wu**[1,2,3], **Dongmei Tian**[1,2,*], **Shuhui Song**[1,2,3,*]

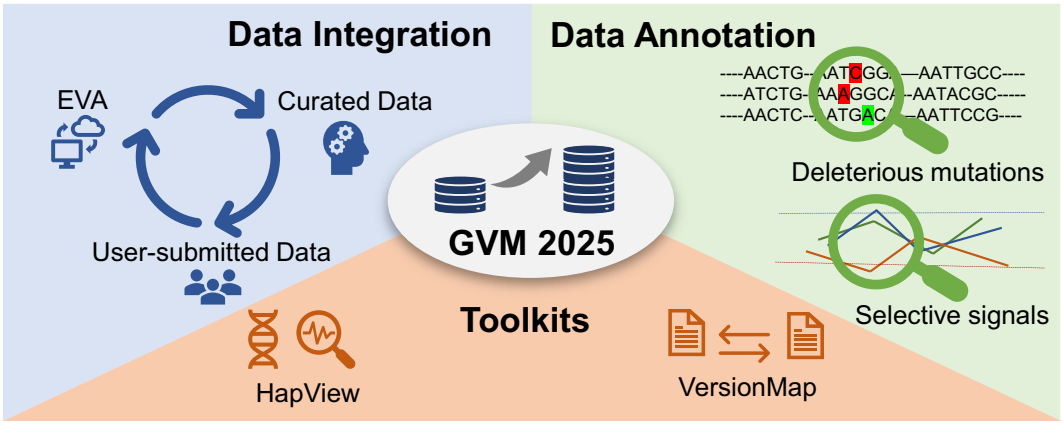
[1]National Genomics Data Center, China National Center for Bioinformation, Beijing 100101, China
[2]Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China
[3]University of Chinese Academy of Sciences, Beijing 100049, China

*To whom correspondence should be addressed. Email: songshh@big.ac.cn
Correspondence may also be addressed to Dongmei Tian. Email: tiandm@big.ac.cn
†The first four authors should be regarded as Joint First Authors.

Abstract

The Genome Variation Map (GVM; https://ngdc.cncb.ac.cn/gvm) serves as a public repository for genetic variation data. This version represents a major upgrade to GVM, integrating rigorously curated high-quality DNA resequencing data from a wide range of species and delivering significant enhancements over previous releases. Specifically, the current release encompasses ~2.14 billion standardized variants across 73 species and archives 898 genetic variation projects originating from 79 species, contributed by 180 organizations globally. Furthermore, GVM innovatively incorporates a public variant aggregation pipeline that consolidates in-house curated datasets, user-submitted data, and records from the European Variation Archive to generate comprehensive reference variant sets for each organism. A new module has also been introduced to provide deleterious variant annotations and population genetic selection signals based on population genetic analyses. Additionally, newly implemented online tools support population-specific haplotype analysis and cross-mapping between genome assemblies, serving as integral features of GVM. Collectively, this resource remains essential for archiving and utilizing genomic variation data and for catalyzing advancements in evolutionary biology, disease etiology research, and agricultural genomics.

Graphical abstract



Introduction

The analysis of genomic variation data holds paramount significance in evolutionary and population genetics research pertaining to human health and disease, breeding programs for plants and animals, as well as biodiversity conservation [1]. Since its establishment in 2017 [2], the Genome Variation Map (GVM; https://ngdc.cncb.ac.cn/gvm/) has evolved progressively as a cornerstone resource within the National Genomics Data Center and China National Center for Bioinformation [3]. Serving as a key complement to leading international genomic variation repositories [such as the European Variation Archive (EVA) [4] and dbSNP [5]], GVM has emerged as an indispensable public archive [6]. Currently serving tens of thousands of users across 110 nations and regions, it supports data curation, research implementation, and practical applications [7–10]. Notably, it is recognized

Received: September 15, 2025. **Revised:** October 11, 2025. **Accepted:** October 13, 2025

through joint endorsement by Cell Press and Springer Nature as a designated repository for genetic variation data (see https://www.cell.com/pb-assets/journals/research/cellpress/data/RecommendRepositories-1621989644133.pdf; https://www.springernature.com/gp/authors/research-data-policy/repositories-mandates/19540364)

Recent advancements in sequencing technologies, coupled with reduced operational costs, have accelerated large-scale population genetics studies across diverse species. Correspondingly, we have implemented significant enhancements to the GVM repository, characterized by (i) substantial data expansion: the current iteration exhibits marked increases in species coverage (particularly crops, primates, and viruses), sample volumes, variant counts, and internationally submitted datasets compared to the 2021 release; (ii) integrated multisource variant consolidation: variants from disparate origins, including GVM-curated entries, EVA-collected data, and user submissions, have been systematically integrated to construct comprehensive reference variant sets for designated species; (iii) enriched functional annotations: implementation of deleterious variant scoring for key crop populations, supplemented by evolutionary selection signal inferences derived from population genetic analyses; and (iv) novel analytical utilities: deployment of specialized tools including a haplotype tool for population-specific haplotype frequency computation, variant-trait association modules, and the VersionMap genome coordinate conversion system. This enhanced iteration of GVM is positioned to substantially empower diverse research domains encompassing plant/animal breeding, human population genetics, viral evolutionary dynamics, and translational applications.

Materials and methods

The variant data integration workflow consolidated curated data and user-submitted data from GVM, along with variant data (RS Release 7) from the EVA. The merging process aligned the other two types of data to the curated dataset. Initially, each batch of user-submitted datasets underwent quality control and filtering: (i) datasets containing fewer than 1000 variants were excluded, as they were considered to represent only a few genomic regions; (ii) datasets generated by genotyping arrays were discarded if they could not be fully aligned with the reference genome version; and (iii) variants identified as LowQual in the VCF files, based on filtering criteria (QD <2.0 || FS >60.0 || SOR >5.5 || MQ <40.0 || MQRankSum <-12.5 || ReadPosRankSum <-8.0), were excluded. For each user-submitted dataset, a filtered VCF file was generated, with all samples consolidated into a single file. Subsequently, the quality-controlled user data and EVA data were converted to ensure consistency with the reference genome version used in the curated dataset, using the EVA remapping pipeline (https://github.com/EBIvariation/variant-remapping) [4]. Finally, data integration for each species was performed by updating allele information for existing curated variant sites and assigning unique database identifiers to novel variant sites.

To predict the functional effects of variants, variants in the GVM were annotated using Sorting Intolerant From Tolerant 4G (SIFT 4G) [11]. To create the SIFT 4G annotation database, UniRef90 (https://www.uniprot.org/ download date:15 October 2024) was used as a reference protein set. For each species, the SIFT 4G annotation database was constructed by integrating the UniRef90 protein dataset with the corresponding reference genome and GTF annotation file implemented through SIFT4G_Create_Genomic_DB. Based on these species-specific databases, variant effect prediction was carried out using SIFT4G_Annotator.jar, which calculates a SIFT score ranging from 0 to 1. Variants with a SIFT score <0.05 were classified as deleterious, while those with a score ≥ 0.05 were considered tolerated.

Selective sweep analysis was conducted to identify genomic regions potentially affected by positive selection during domestication and improvement. Three complementary population genetic statistics—Tajima's D [12], reduction of diversity (ROD) [13], and F_{st} [14]—were applied to detect selection signatures. Tajima's D was used to evaluate deviations from neutrality within populations, ROD quantified the relative reduction in nucleotide diversity in the cultivated population compared with the wild population, and F_{st} was used to assess population differentiation between cultivated and wild groups. All analyses were performed in the GVM using VCFtools [15]. Tajima's D was calculated with a 20-kb window size, while ROD and F_{st} were computed using a sliding window of 20 kb with a 2-kb step size. The top 5% of F_{st} and ROD values and the bottom 5% of Tajima's D values were defined as significance thresholds. Candidate regions identified from these analyses were subsequently integrated with gene annotation and variant effect prediction results to pinpoint loci potentially associated with agronomic and adaptive traits.

Updates and new features

Data updates

Over recent years, GVM has achieved substantial enhancements in data content, analytical capabilities, and user services (Table 1 and Fig. 1). Regarding curated data, standardized variant datasets were constructed through systematic retrieval of raw resequencing data from repositories including the Sequence Read Archive [16], European Nucleotide Archive [17], and Genome Sequence Archive [18], processed through a unified variant-calling framework [6]. The current release encompasses 73 diverse taxonomic groups—representing an 80% expansion—spanning vertebrates, plants, and viral pathogens (Supplementary Table S1). Notable additions include 17 newly integrated agronomically significant crops (nine staple food sources and eight commercial cultivars), providing foundational genomic variation resources to support agricultural innovation. This release also includes five new primate species, whose variation profiles offer a valuable comparative framework for human disease genetics, as well as four human coronaviruses and the monkeypox virus, which provide critical data for epidemic research and public health preparedness. The curated repository now harbors nearly twice the volume of samples and variants compared to the previous version, aggregating ~2.14 billion distinct genetic variants.

Concurrently, user-submitted data has manifested exponential augmentation, with the GVM currently hosting 95.55 terabytes of data encompassing 898 distinct projects. These include genetic variants from 764 837 samples across 79 species, contributed by 180 organizations from around the world. The submitted datasets cover diverse research areas, including population cohort analyses and complex disease genetics

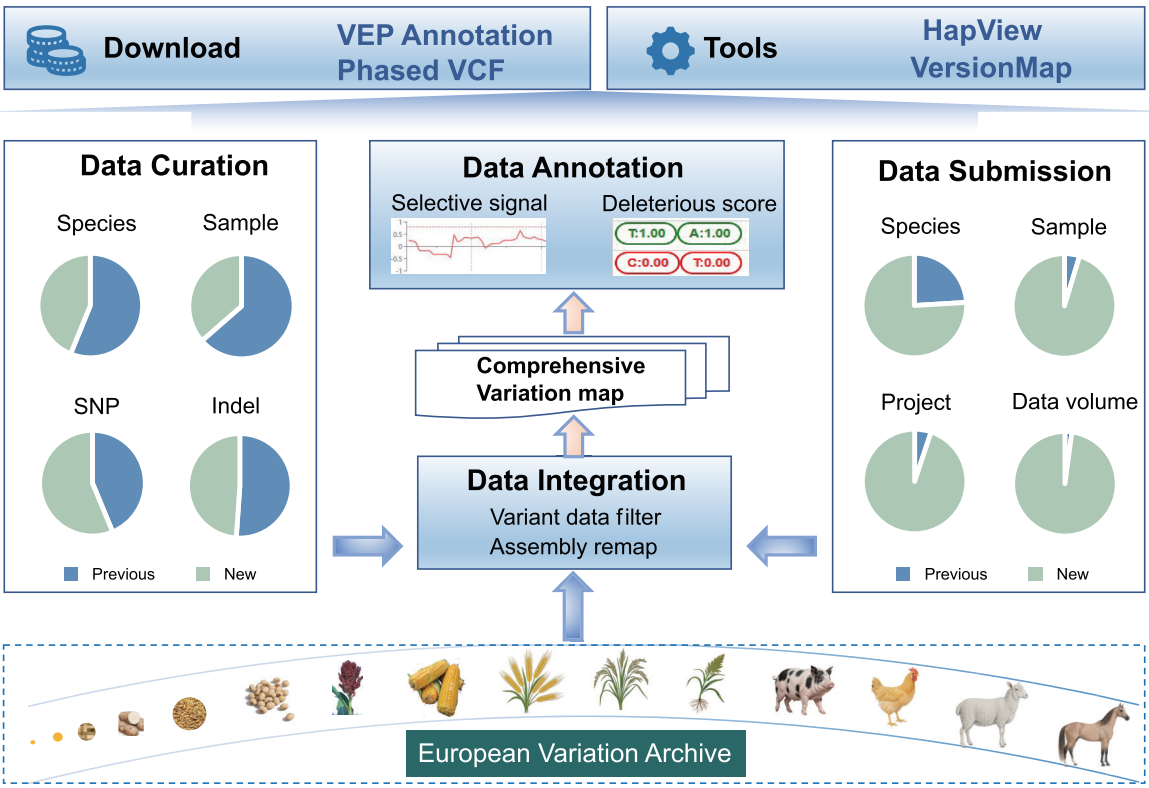


Figure 1. Overview of the GVM update. The sections 'Data Curation' and 'Data Submission' depict the growth of data volume in the data curation and submission modules compared to the previous version. The four sections 'Data Annotation', 'Data Integration', 'Download' and 'Tools' highlight the new features implemented in this update.

Table 1. Statistics and comparison between the two versions of GVM

	GVM (2021)	GVM (updated)
Curated data and information		
Species	41	73
Projects	202	437
Samples	64 819	101 967
Number of SNPs	818 769 204	1 871 449 856
Number of indels	141 640 247	277 154 506
G2P knowledge	260 393	466 100
Public variant data integration	NA	Available
Deleterious variant information	NA	SIFT
Population selective signal	NA	F_{st} /Tajima' *D*/ROD
Tools		
HapView (haplotype calculating)	NA	Available
VersionMap (cross map of assembly)	NA	Available
Submission data and information		
Submitted species	19	79
Submitted samples	35 961	764 837
Submitted projects	46	898
Submitted data volume	2.03 TB	95.55 TB
Submitted organizations	31	180
Support published papers	33	252
Usage		
Visitors	12 119	68 800

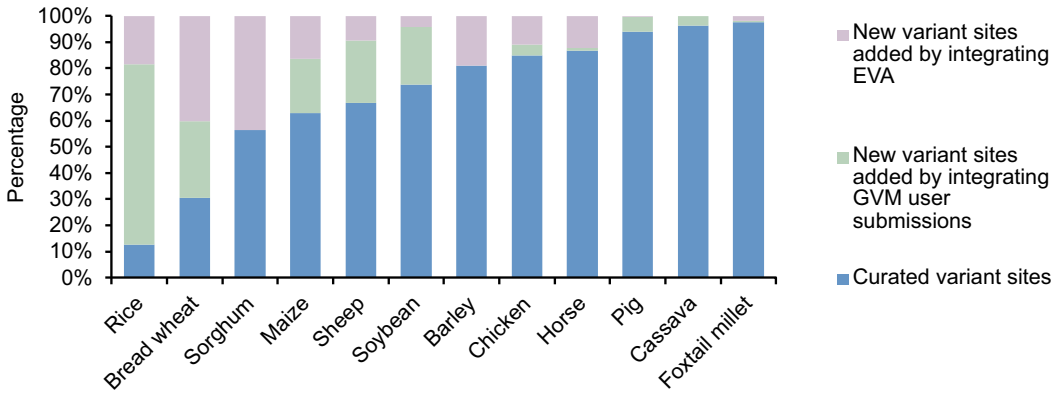
[19–21], disease mechanism research [22, 23], plant and animal adaptive evolution and breeding [24, 25], and microbial–host interactions [26, 27]. To date, these resources have supported 252 peer-reviewed publications.

Data integration

GVM and EVA both serve as key repositories for genomic variation in plant and animal species. However, variant information remains fragmented due to the lack of a systematic data exchange framework. This makes it imperative to integrate comprehensive variation datasets to construct a high-quality, unified variation atlas. To address this gap, we developed a data integration workflow that systematically combines curated and user-submitted data from GVM with variant records from EVA into a consolidated reference variant resource. Building on this reference dataset, research outputs from multiple institutions are standardized through unified genomic coordinates and nomenclature across all variant sites. The dataset establishes cross-referenced with RS identifiers from EVA and dbSNP, thereby enhancing interoperability. Variant credibility is further strengthened by integrating multiple independent sources of evidence. Through a unified platform, all integrated information is accessible, enabling users to retrieve comprehensive variant profiles within any genomic region via a single query—dramatically improving retrieval efficiency. In summary, this reference variant dataset facilitates global sharing and validation of genetic variation data among researchers worldwide.

To date, integration across eight plant and four animal species has yielded fold increases in aggregate variant loci ranging from 1.02 to 7.90 (Fig. 2). Notably, *Oryza sativa* (rice) and *Triticum aestivum* (wheat) manifested the most significant augmentations, reflecting their importance in agricultural genomics and the rapid pace of related research. These results demonstrate that the integration strategy significantly

A



B

All	VarID	Position	Alleles	MAF	Consequence Type|Effect	Gene	dbSNP
☐	osa123085081	chr1:1267	A/C	-:-	NMD_transcript_variant|MODIFIER;	Os01g0100100	rs54581192
☐	osa19130995	chr1:1268	C/T	T:0.0628	NMD_transcript_variant|MODIFIER;	Os01g0100100	
☐	osa123085082	chr1:1269	A/C	-:-	NMD_transcript_variant|MODIFIER;	Os01g0100100	rs56367930
☐	osa123085083	chr1:1270	G/A	-:-	NMD_transcript_variant|MODIFIER;	Os01g0100100	rs54581191
☐	osa123085084	chr1:1272	A/C	-:-	NMD_transcript_variant|MODIFIER;	Os01g0100100	rs56367931
☐	osa19130996	chr1:1273	G/T	T:0.0020	NMD_transcript_variant|MODIFIER;	Os01g0100100	
☐	osa19130997	chr1:1274	G/A	A:0.0040	NMD_transcript_variant|MODIFIER;	Os01g0100100	
☑	osa9	chr1:1277	T/C	C:0.1254	upstream_gene_variant|MODIFIER;	Os01g0100100	rs87752593
☐	osa19130998	chr1:1278	G/A	A:0.0423	NMD_transcript_variant|MODIFIER;	Os01g0100100	rs87795273
☐	osa19130999	chr1:1279	C/T	T:0.0141	NMD_transcript_variant|MODIFIER;	Os01g0100100	

① ② ③

View Genotype Download

GVM Dataset	Population	Sample #	Position:chr1:1277	Individuals Genotype
GVM_curated	Indica	2567	T:0.82; C:0.18;	show
GVM_curated	japonica	1499	T:0.97; C:0.03;	show
GVM_curated	Other	635	T:0.87; C:0.13;	show
GVM_curated	wild	451	T:0.73; C:0.27;	show
GVM000573	Wild,weedy and cultivated accessions	263	T:0.3502; C:0.6498;	-
GVM000882	Japonica rice varieties	411	T:1.0000;	-
GVM000285	Rice SNPs for convergent evolution	2433	T:0.6734; C:0.3266;	-

Figure 2. Data integration statistics and an example. (**A**) Statistical summary of variant proportions after integration. (**B**) Example of integrated data from rice. When a specific variant is selected and "view genotype" is clicked, EVA RS identifiers are shown in the last column (labeled as "1"), and allele frequency information for both the curated population dataset (labeled as "2") and user-submitted datasets, if available (labeled as "3") will be displayed.

improves the comprehensiveness of variant information, facilitating downstream applications such as the assessment of mutation load from deleterious variants.

Data annotation

Deleterious mutations and selection signals are critically important in population genetics [28–31], breeding [13, 32], and evolutionary biology [33, 34]. To support research in these areas, GVM provides functional annotations by predicting deleterious mutations and detecting selection signals using population genetic analyses (see the "Materials and methods" section). This enables researchers to identify and utilize biologically relevant genetic variants more effectively. Currently, GVM encompasses deleteriousness annotations for 27 species, covering 47 435 958 distinct variants (Fig. 3A). Each variant is assigned a deleteriousness label and a quantitative score. In parallel, the platform systematically collates and stratifies species population datasets to compute fundamental population genetic parameters—including F_{st}, ROD, and Tajima's D—thereby providing selection signal annotations across eight species (Fig. 3B).

By integrating outputs from multiple statistical frameworks, GVM delivers consolidated annotation of selective pressures. For example, analysis of the *Npp.07B002710.1* gene in *Saccharum* species [35] concurrently displays selection signal intensity across genomic segments along with deleteriousness scores of constituent variants (Fig. 3C). Candidate regions under selection are identified by comparing population genetic statistics with defined thresholds, and potentially

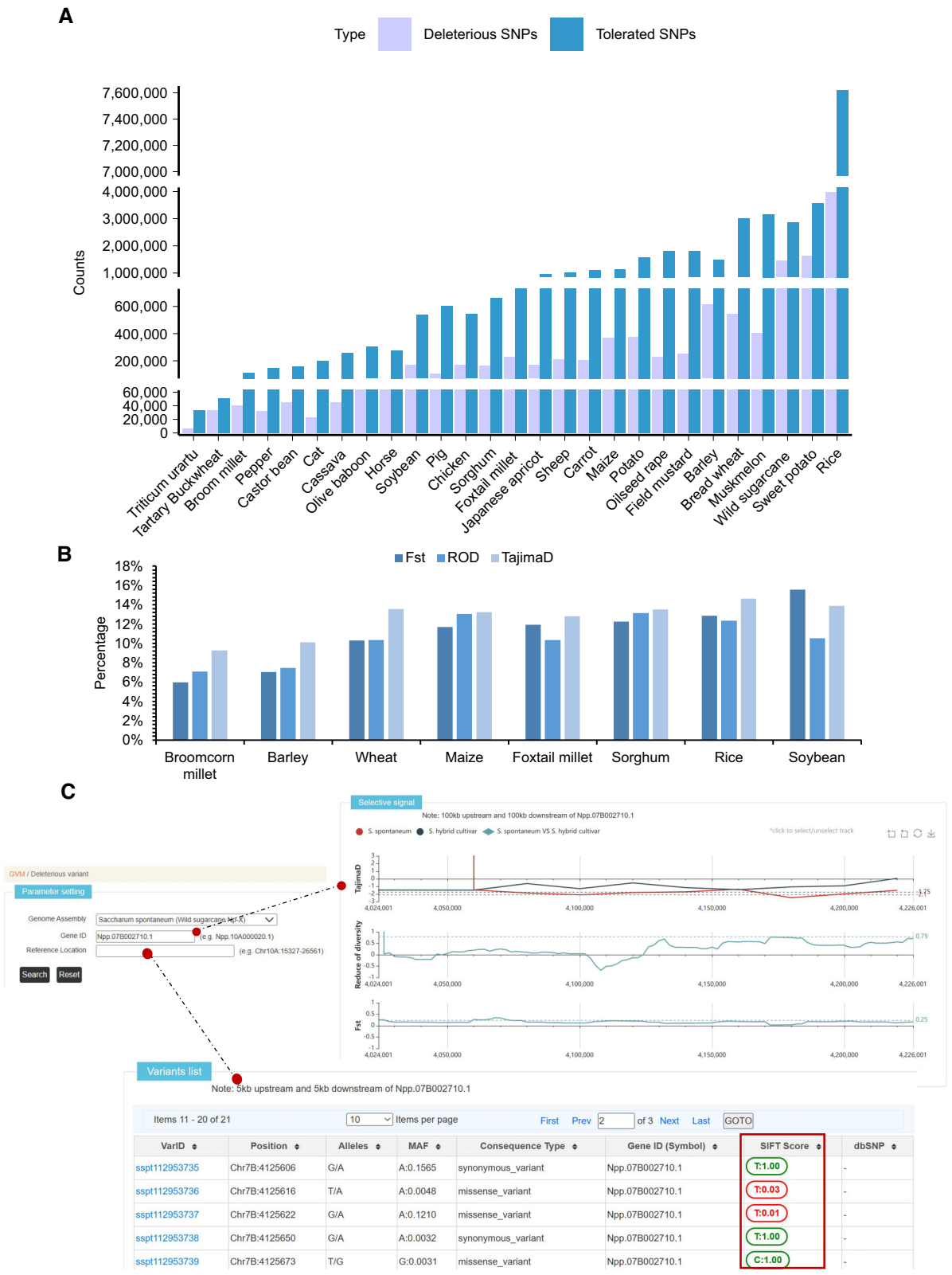


Figure 3. Overview of data annotation. (**A**) Summary statistics of deleterious mutations across annotated species. (**B**) Proportion of genomic regions under selection. (**C**) Example view of deleterious mutation and selection signal annotated for Npp.07B002710.1. The box highlights the deleterious annotation, where score less than 0.05 is defined as deleterious variant, otherwise it is defined as tolerated variant.

deleterious mutations are highlighted using color-coded indicators. Collectively, these functional annotations provide an indispensable resource for identifying deleterious variants, detecting constrained regions, and improving the functional interpretation of genetic polymorphism.

Data download

GVM provides open access to all publicly available variants, which are downloadable in both VCF and FASTA formats from https://ngdc.cncb.ac.cn/gvm/download. The brief VCF file contains genomic position, reference and alternative alleles, and variant quality, while the FASTA file provides 50 nt flanking sequences for each variant. The detailed VCF file contains the genotype information for all samples, which is more useful for users to conduct in-depth GWAS functional analysis. In response to user feedback, we have newly added the VEP annotation, which provides functional consequence prediction for each variant, and the phased VCF files, which offer filtered and phased variation data for enhanced usability.

Toolkits

To meet the growing need for online analytical capabilities, GVM has launched a new Toolkits module. The first component, the HapView portal, offers a modular analytical framework for population-specific haplotype investigation. This tool supports dual input sources: curated GVM repositories or user-uploaded VCF/phenotype files. A customizable quality control framework ensures analytical reliability, permitting filtration based on three critical parameters: missing data rates, heterozygosity indices, and minor allele frequencies. Beyond frequency calculations, HapView conducts haplotype–trait correlation analyses and enables comparative assessments against public GVM datasets. Results are presented through concise tabular summaries and dynamic visualization modules, with comprehensive outputs available for download, thereby expediting linkage disequilibrium analyses and preliminary functional haplotype screening. The second tool, VersionMap, facilitates online genomic coordinate translation across different assembly versions. Built on validated algorithms including CrossMap [36] and UCSC's liftOver [37], it delivers precise genomic coordinate translation for both VCF and BED formats. This function is essential for harmonizing multisource, temporally heterogeneous genomic datasets, substantially enhancing research reproducibility, and cross-study comparability.

Conclusion and perspectives

In conclusion, as a vital public repository for genetic variation data, GVM continues to evolve, providing researchers worldwide with comprehensive data submission, management, and sharing. This latest version of GVM significantly expands its data volume, including both curator-refined and user-submitted data, greatly broadening species coverage and deepening population-level insights. It has newly launched a public variant data integration pipeline, aiming to establish more comprehensive reference variation maps for each species. Another highlight of GVM 3.0 is the addition of new modules focusing on deleterious mutations and population selection signals, substantially enriching functional annotations of genetic variants. Additionally, this version introduces two practical analysis tools that support population haplotype analysis and genomic coordinate conversion, enhancing efficiency in data interrogation and application. These upgrades, along with enhanced data download functionality, considerably strengthen the platform's capacity for downstream analysis and promote wider data reuse.

Looking forward, GVM remains committed to the principle of open sharing and will continue to enhance its functionality. In the era of big data in genetics, coupled with advances in artificial intelligence (AI), there is tremendous potential to accelerate genetic research and uncover novel biological insights. In this context, the availability of AI-ready datasets is of paramount importance. GVM plans to enhance metadata organization, refine data cleaning and curation strategies, and establish stronger links with phenotypic data to construct high-quality, AI-compatible datasets. Moreover, GVM will strengthen its focus on functional variant annotations, with efforts to extend bioinformatic methods to enable identification and interpretation of noncoding deleterious mutations. Meanwhile, GVM will broaden the scope of functional annotations by integrating associations with other genomic and molecular datasets, like gene expression data from GEN [38] and epigenomic modifications from Methbank [39], to enhance the variant interpretation. In summary, as GVM continues to advance, it will sustain its role as an essential international genetic resource, supporting research in precision medicine, intelligent breeding, and public health safety.

Acknowledgements

We thank our colleagues, students, and users for reporting bugs and sending comments.

Author contributions: Dongmei Tian (Data curation [equal], Formal analysis [equal], Software [equal], Supervision [equal], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Shuhui Song (Conceptualization [equal], Funding acquisition [equal], Project administration [equal], Supervision [equal], Writing—review & editing [equal]), Xue Bai (Data curation [equal], Formal analysis [equal], Visualization [equal], Writing—original draft [equal]), Haixia Xie (Data curation [equal], Formal analysis [equal], Visualization [equal], Writing—original draft [equal]), Hong Luo (Data curation [equal], Writing—original draft [equal], Writing—review & editing [equal]), Siwei Ren (Data curation [equal], Formal analysis [equal], Writing—original draft [equal]), BixiaTang (Software [equal]), Cuiping Li (Data curation [equal]), Yi Wang (Data curation [equal]), Bo Xu (Software [equal]), and Zishan Wu (Data curation [equal])

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

This work was supported by the Biological Breeding—National Science and Technology Major Project (2022ZD04017), the Informatization Plan of Chinese Academy of Sciences (CAS-WX2024GC-0602), the Strategic

Priority Research Program of the Chinese Academy of Sciences (XDA0450102, XDA0460405), National Key Research and Development Program of China (2025YFF1207901), and National Natural Science Foundation of China (32270718, 92374201). Funding to pay the Open Access publication charges for this article was provided by the research grant (2022ZD04017) from the Biological Breeding—National Science and Technology Major Project.

Data availability

GVM is freely accessible at https://ngdc.cncb.ac.cn/gvm/.

References

1. Stange M, Barrett RDH, Hendry AP. The importance of genomic variation for biodiversity, ecosystems and people. *Nat Rev Genet* 2021;22:89–105.
2. Song S, Tian D, Li C *et al.* Genome Variation Map: a data repository of genome variations in BIG Data Center. *Nucleic Acids Res* 2018;46:D944–9. https://doi.org/10.1093/nar/gkx986
3. Members C-N, Partners. Database resources of the National Genomics Data Center, China National Center for Bioinformation in 2025. *Nucleic Acids Res* 2025;53:D30–44.
4. Cezard T Cunningham F, Hunt SE *et al.* The European Variation Archive: a FAIR resource of genomic variation for all species. *Nucleic Acids Res* 2022;50:D1216–20.
5. Phan L, Zhang H, Wang Q *et al.* The evolution of dbSNP: 25 years of impact in genomic research. *Nucleic Acids Res* 2024;53:D925–31. https://doi.org/10.1093/nar/gkae977
6. Li C, Tian D, Tang B *et al.* Genome Variation Map: a worldwide collection of genome variations across multiple species. *Nucleic Acids Res* 2021;49:D1186–91. https://doi.org/10.1093/nar/gkaa1005
7. Talenti A, Wilkinson T, Morrison L.J *et al.* The evolution and convergence of mutation spectra across mammals. *Commun Biol* 2025;8:763. https://doi.org/10.1038/s42003-025-08181-x
8. Van Buren SL, Brown CT, Szmatoła T *et al.* A comparative review of short genetic variant databases across humans and animal species. *Brief Bioinform* 2025;26:bbaf356. https://doi.org/10.1093/bib/bbaf356
9. Liu S, Li C, Wang H *et al.* Mapping regulatory variants controlling gene expression in drought response and tolerance in maize. *Genome Biol* 2020;21:163. https://doi.org/10.1186/s13059-020-02069-1
10. Liu X, Tian D, Li C *et al.* GWAS Atlas: an updated knowledgebase integrating more curated associations in plants and animals. *Nucleic Acids Res* 2023;51:D969–76. https://doi.org/10.1093/nar/gkac924
11. Vaser R Adusumalli S, Leng SN *et al.* SIFT missense predictions for genomes. *Nat Protoc* 2016;11:1–9. https://doi.org/10.1038/nprot.2015.123
12. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 1989;123:585–95. https://doi.org/10.1093/genetics/123.3.585
13. Xu X, Liu X, Ge S *et al.* Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat Biotechnol* 2012;30:105–111. https://doi.org/10.1038/nbt.2050.
14. Weir BS, Cockerham CC. Estimating *F*-statistics for the analysis of population structure. *Evolution* 1984;38:1358–70.
15. Danecek P, Auton A, Abecasis G *et al.* The variant call format and VCFtools. *Bioinformatics* 2011;27:2156–8. https://doi.org/10.1093/bioinformatics/btr330
16. Katz K, Shutov O, Lapoint R *et al.* The Sequence Read Archive: a decade more of explosive growth. *Nucleic Acids Res* 2022;50:D387–90. https://doi.org/10.1093/nar/gkab1053
17. Yuan D, Ahamed A, Burgin J *et al.* The European Nucleotide Archive in 2023. *Nucleic Acids Res* 2023;52:D92–7. https://doi.org/10.1093/nar/gkad1067
18. Wang Y, Song F, Zhu J *et al.* GSA: genome Sequence Archive. *Genomics Proteomics Bioinformatics* 2017;15:14–8. https://doi.org/10.1016/j.gpb.2017.01.001
19. Wu Z, Jiang Z, Li T *et al.* Structural variants in the Chinese population and their impact on phenotypes, diseases and population adaptation. *Nat Commun* 2021;12:6501. https://doi.org/10.1038/s41467-021-26856-x
20. Huang S, Liu S, Huang M *et al.* The Born in Guangzhou Cohort Study enables generational genetic discoveries. *Nature* 2024;626:565–73. https://doi.org/10.1038/s41586-023-06988-4
21. Guo J, Guo Q, Zhong T *et al.* Phenome-wide association study in 25,639 pregnant Chinese women reveals loci associated with maternal comorbidities and child health. *Cell Genom* 2024;4:100632. https://doi.org/10.1016/j.xgen.2024.100632
22. Zeng Y, Luo CL, Lin GW *et al.* Whole-exome sequencing association study reveals genetic effects on tumor microenvironment components in nasopharyngeal carcinoma. *J Clin Invest* 2025;135:e182768. https://doi.org/10.1172/JCI182768.
23. Lou S, Zhu G, Xing C *et al.* Transcriptome-wide association identifies KLC1 as a regulator of mitophagy in non-syndromic cleft lip with or without palate. *Imeta* 2024;3:e262. https://doi.org/10.1002/imt2.262
24. Yang C, Yan J, Jiang S *et al.* Resequencing 250 Soybean Accessions: new insights into genes associated with agronomic traits and genetic networks. *Genomics Proteomics Bioinformatics* 2022;20:29–41. https://doi.org/10.1016/j.gpb.2021.02.009
25. Liu Y, Du H, Li P *et al.* Pan-genome of wild and cultivated soybeans. *Cell* 2020;182:162–76. https://doi.org/10.1016/j.cell.2020.05.023.
26. Yang H, Wu J, Huang X *et al.* ABO genotype alters the gut microbiota by regulating GalNAc levels in pigs. *Nature* 2022;606:358–67. https://doi.org/10.1038/s41586-022-04769-z
27. Tan X, Xie H, Yu J *et al.* Host genetic determinants drive compartment-specific assembly of tea plant microbiomes. *Plant Biotechnol J* 2022;20:2174–86. https://doi.org/10.1111/pbi.13897
28. He F, Pasam R, Shi F *et al.* Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome. *Nat Genet* 2019;51:896–904. https://doi.org/10.1038/s41588-019-0382-2
29. Kim MS, Lozano R, Kim JH *et al.* The patterns of deleterious mutations during the domestication of soybean. *Nat Commun* 2021;12:97. https://doi.org/10.1038/s41467-020-20337-3
30. Lozano R, Gazave E, Dos Santos JPR *et al.* Comparative evolutionary genetics of deleterious load in sorghum and maize. *Nat Plants* 2021;7:17–24. https://doi.org/10.1038/s41477-020-00834-5
31. Ramu P, Esuma W, Kawuki R *et al.* Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. *Nat Genet* 2017;49:959–63. https://doi.org/10.1038/ng.3845
32. Li X, Chen X, Fang J *et al.* Whole-genome sequencing of a worldwide collection of sugarcane cultivars (*Saccharum* spp.) reveals the genetic basis of cultivar improvement. *Plant J* 2024;119:2151–67. https://doi.org/10.1111/tpj.16861
33. Dwivedi SL, Heslop-Harrison P, Spillane C *et al.* Evolutionary dynamics and adaptive benefits of deleterious mutations in crop gene pools. *Trends Plant Sci* 2023;28:685–97. https://doi.org/10.1016/j.tplants.2023.01.006
34. Yi H, Wang J, Dong S *et al.* Genomic signatures of inbreeding and mutation load in tree ferns. *Plant J* 2024;120:1522–35. https://doi.org/10.1111/tpj.17064
35. Luo H, Bai X, Wu Z *et al.* SugarcaneOmics: an integrative multi-omics platform for sugarcane research. *Plant Commun* 2025;101489. https://doi.org/10.1016/j.xplc.2025.101489

36. Zhao H, Sun Z, Wang J *et al.* CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* 2014;30:1006–7. https://doi.org/10.1093/bioinformatics/btt730
37. Kuhn RM, Haussler D, Kent WJ. The UCSC genome browser and associated tools. *Brief Bioinform* 2013;14:144–61. https://doi.org/10.1093/bib/bbs038
38. Zhang Y, Zou D, Zhu T *et al.* Gene Expression Nebulas (GEN): a comprehensive data portal integrating transcriptomic profiles across multiple species at both bulk and single-cell levels. *Nucleic Acids Res* 2022;50:D1016–24. https://doi.org/10.1093/nar/gkab878
39. Zhang M, Zong W, Zou D *et al.* . MethBank 4.0: an updated database of DNA methylation across a variety of species. *Nucleic Acids Res* 2023;51:D208–16. https://doi.org/10.1093/nar/gkac969

Received: September 15, 2025. **Revised:** October 11, 2025. **Accepted:** October 13, 2025