scientific data

Check for updates

OPEN

ARTICLE

Adoption of Standard Reference SNP Identifiers in Agricultural Genomics for Interoperability and Data Reuse

Marcela K. Tello-Ruiz[1,18] ✉, Timothee Cezard[2], Carson Andorf[3,4], Sonia Balyan[5], Nahla V. Bassil[6], Sebastian Beier[7], Jill M. Bushakra[6], Tao-Ho Chang[8], Kapeel Chogule[1], Irene Cobo-Simón[9], Sarah Dyer[2], Christine G. Elsik[10], Nicholas Gladman[1,11], Melanie Harrison[12], Jodi Humann[13], Catherine Kim[1], Vivek Kumar[1], Raja S. Nandety[14], Rex Nelson[4], Andrew Olson[1], Taner Z. Sen[15], Moira J. Sheehan[16,17], Sharon Wei[1] & Doreen Ware[1,11] ✉

Agricultural research has long faced challenges with data sharing, often relying on informal networks and requiring significant effort to clean and harmonize data. This hampers collaboration and limits data reuse. While FAIR (Findable, Accessible, Interoperable, and Reusable) principles are widely adopted in biomedical research, their uptake in agricultural genomics has lagged. The AgBioData Standards for Genetic Variation Working Group aims to close this gap by promoting FAIR data practices. We surveyed current standards for managing agricultural genetic variation and recommend adopting reference SNP identifiers (rsIDs) as a key step. We present examples from crop research communities with varying data maturity, including those without reference assemblies. Milestones include introducing nearly 220 million rsIDs to Gramene and pangenome databases, projecting rsIDs from reference to pangenome varieties in sorghum and maize, and developing an agricultural FAIR guide for rsID adoption. Better coordination among data producers, repositories, and breeding platforms is essential to improve interoperability, consistency, and accelerate genetic variant discovery for crop trait improvement.

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 11724, USA. [2]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK. [3]Department of Computer Science, Iowa State University, Ames, IA, 50011, USA. [4]United States Department of Agriculture, Agricultural Research Service (USDA-ARS), Corn Insects and Crop Genetics Research Unit, Ames, IA, 50011, USA. [5]Indian Biological Data Centre, Regional Centre for Biotechnology, Faridabad, Haryana, 121001, India. [6]USDA ARS National Clonal Germplasm Repository, Corvallis, OR, 97333, USA. [7]Institute of Bio- and Geosciences (IBG-4 Bioinformatics), CEPLAS, BIOSC, Forschungszentrum Jülich GmbH, Wilhelm Johnen Straße, Jülich, Germany. [8]Program in Plant Health Care, Academy of Circular Economy, National Chung Hsing University, Taichung, Taiwan. [9]Institute of Forest Science, National Center National Institute for Agricultural and Food Research and Technology, Spanish National Research Council (ICIFOR-INIA-CSIC), Madrid, Spain. [10]Division of Animal Sciences, University of Missouri, Columbia, MO, 65211, USA. [11]USDA ARS NEA, Plant Soil & Nutrition Laboratory Research Unit, Ithaca, NY, 14853, USA. [12]USDA ARS, Plant Genetic Resources Conservation Unit, Griffin, GA, 30223, USA. [13]Department of Horticulture, Washington State University, Pullman, WA, 99164, USA. [14]USDA ARS PA, Edward T. Schafer Agricultural Research Center, Fargo, ND, 58102, USA. [15]USDA ARS, Crop Improvement and Genetics Research, Albany, CA, 94710, USA. [16]Department of Plant Breeding and Genetics, Cornell University, Ithaca, NY, 14850, USA. [17]Breeding Insight, University of Florida - IFAS, Gainesville, FL, 32611, USA. [18]Present address: Phoenix Bioinformatics, 39899 Balentine Dr, Ste 200, Newark, CA, 94560, USA. ✉e-mail: marcela@phxbio.org; ware@cshl.edu

Introduction

Genetic variation underpins evolution and the improvement of crops and livestock, providing the raw material for selection and adaptation in both natural ecosystems and breeding programs. As the agricultural sector increasingly turns to genomics-assisted breeding to meet global food security challenges, the ability to effectively reuse and integrate genetic variation data across studies and platforms has become essential. The FAIR principles of data management promote standardization through the use of shared identifiers, file formats, and metadata schemas[1]. However, agricultural genomics still lags behind human genetics in adopting FAIR practices. Fragmented workflows, inconsistent standards, and limited cross-platform interoperability have led to siloed and poorly annotated datasets, impeding collaboration, reproducibility, and downstream breeding applications.

FAIRification, the process of applying FAIR principles, is particularly challenging in agriculture due to the diversity of species, data types, research goals, and community practices. To address these barriers, the AgBioData Consortium[2] formed the Standards for Genetic Variation Working Group (SGV WG), bringing together geneticists, breeders, bioinformaticians, and biocurators to improve the interoperability and reusability of genetic variation data. Descriptions of key data standards for biosamples, phenotypic traits, and genetic markers, in the context of FAIR management best practices are provided in Supplementary Information section 1.

A major obstacle in the FAIRification of agricultural data has been the limited and inconsistent use of globally recognized identifiers, particularly reference SNP cluster IDs (rsID, singular; rsIDs plural), for genetic markers. Introduced in 1998 by dbSNP[3], rsIDs are globally unique, assembly-independent identifiers that cluster genetic variants observed at the same genomic locus. They enable consistent referencing of variants across assemblies, databases, and studies, unlocking compatibility with a wide range of bioinformatics tools and analysis pipelines[4]. rsIDs are widely adopted in human genomics, where they form the backbone of genome-wide association studies (GWAS), clinical variant interpretation, and large-scale meta-analyses[5–7]. Their adoption in agriculture, however, has been slow.

This delay stems from key differences between agricultural and human genomics. While human genetics focuses on a single species and benefits from centralized funding and infrastructure, agricultural genetics spans numerous species, each with its own research community, data ecosystem, and breeding priorities. dbSNP initially assigned rsIDs to multiple species but discontinued non-human support in 2017. The European Variation Archive[8] (EVA, https://www.ebi.ac.uk/eva) assumed this role, becoming the largest short genetic variant database[9]. It currently holds over 3.4 billion variants across 288 species, with the integration of agricultural species gaining momentum in recent years (Fig. 1).

Another limiting factor has been the lack of high-quality, International Nucleotide Sequence Database Collaboration[10] (INSDC)-accessioned reference genomes in many crop species, a prerequisite for assigning rsIDs by the EVA. Historically, agricultural genomics communities relied heavily on internal IDs or project-specific marker names, often without consistent metadata or shared naming conventions, making it difficult to harmonize genetic variant data files or integrate data across studies. Breeding programs further prioritized short-term deployment over long-term data interoperability, contributing to fragmented data ecosystems.

Recent developments are beginning to shift this landscape. A growing number of high-quality plant and animal reference genomes have now been accessioned in the INSDC databases (https://www.insdc.org), meeting a key requirement for rsID assignment. Figure 1 outlines technology shifts (e.g., from short- to long-read sequencing) that have supported the development of non-human genomes, as well as key milestones in the evolution of rsID use, illustrating both the early success of human genomics and their delayed, but accelerating availability in agriculture. While rsIDs were introduced in 1998 and have been widely adopted in human genomics for over two decades, the scalable support for non-human organisms by EVA only began in recent years, leaving a multi-year gap in persistent identification for agricultural datasets. This delay underscores the historical fragmentation in agricultural genomics and the recent efforts to bridge these gaps.

Here, we describe coordinated, community-led efforts to close this gap, including practical guidance on data standardization, submission pipelines, and pioneering implementations that demonstrate the feasibility and benefits of large-scale rsID adoption in agriculture. rsIDs enable consistent tracking, comparative analyses, and long-term data sustainability. Embedding rsIDs in pangenome browsers, breeding databases, and annotation pipelines enables critical linkages between genotype and phenotype, powering more effective association studies and trait discovery. We are now at a tipping point, driven by improved reference genome availability, mature submission workflows, and increased community engagement. This article highlights the strategic importance of rsID adoption for agricultural research and proposes a roadmap for advancing standardization and interoperability. We call on researchers, professional societies, journals, funders, and data stewards to join in this effort to unlock the full potential of genetic variation data for crop and livestock improvement.

Results

Challenges for genetic variation data standardization. One of our primary targets has been to support the harmonization and adoption of standards for genetic variation data, and to promote interoperability and access to agricultural datasets. After surveying the agricultural research community on existing and anticipated genetic variation data sets, accessibility, data handling practices, interoperability, and usability, we identified challenges and proposed initiatives to FAIRify the data. Table S1 (see supplementary xlsx file) provides an overview of the main challenges identified over a three-year period through surveys, panel presentations, and breakout room discussions at focused workshops with other members of the AgBioData community. Chief among them was the need to promote the widespread adoption of standard rsIDs among agricultural communities.

Applications of standard rsIDs in agricultural research. Standardized rsIDs enable consistent tracking and identification of specific SNPs across studies, platforms, and databases, with important applications in agricultural research. Genetic variation data undergoes a complex, multi-step life cycle, which herein we refer to

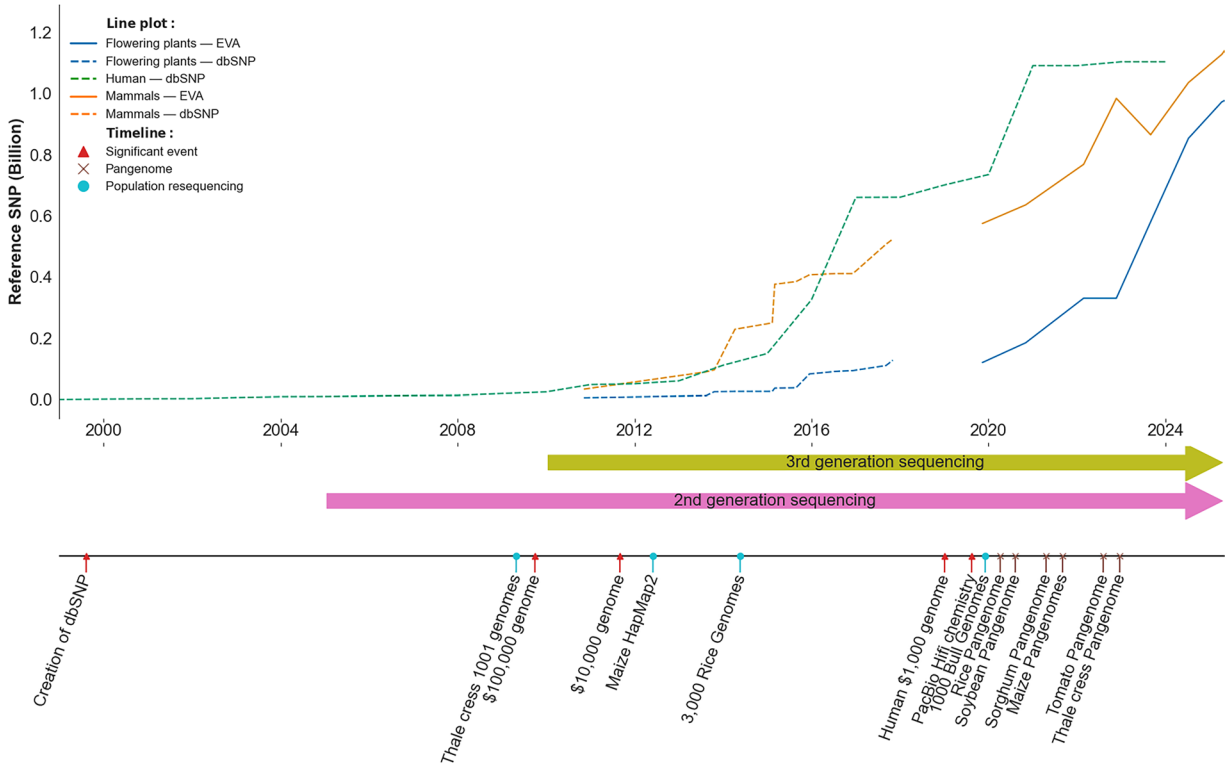

Fig. 1 Timeline of rsID availability alongside major sequencing and genomics milestones. The graph shows the number of rsIDs released by dbSNP and EVA over time for humans, mammals or flowering plants; the gap marks the transition of non-human variant curation from dbSNP to EVA between 2018–2020. The timeline at the bottom indicates key milestones and developments in the adoption of rsIDs.

as the data journey (Fig. 2). This journey progresses from initial variant detection and submission to centralized archives like the EVA, to rsID assignment and downstream applications such as integration into community databases, pangenome projections, and the development of marker assays. In the following sections, we describe the practical implementation of this framework by agricultural community databases, highlighting the specific results achieved and the FAIR-compliant resources developed at each stage of the data journey (Table 1 and Fig. 2).

Brokering the submission of community-generated variant datasets to the EVA. To quantify the impact of active curation on data availability, we brokered the submission of key datasets from the sorghum and soybean communities to the EVA (Supplementary Information section 2). In sorghum, the submission of three major studies[11–13] increased the number of rsIDs available to the community from 8 million in EVA release 3 (February 2022) to over 55 million in release 7 (April 2025). Similarly, SoyBase[14] (https://www.soybase.org) curators submitted over 30 million soybean SNPs to the EVA on behalf of data generators from three selected studies[15–17] aiming at broadly representing genomic variation across the species, raising the number of assigned rsIDs from 18 million to over 28 million. For communities such as raspberry and blackberry, where the lack of an INSDC-accessioned reference genome prevented immediate archiving, we established a support workflow to guide principal investigators through genome submission. This intervention initiates the prerequisite steps for future rsID generation in under-resourced crops. These results demonstrate that targeted brokering effectively populates central archives, creating a foundation of standardized identifiers that can be reused by the broader community.

Agricultural community databases adopt rsIDs to integrate variation data with traits, functional predictions, and population-level data. Submissions to the EVA must include validated VCF files with allele frequency or genotype data, along with metadata describing the samples. Once rsIDs are assigned, these identifiers act as anchors for cross-platform interoperability between the EVA, INSDC databases, and biosample/germplasm repositories, enabling downstream use by various agricultural research platforms (Figure S1, see Supplementary Information document).

Gramene Plants[18] (https://www.gramene.org) and its associated pangenome sites: SorghumBase[19] (https://www.sorghumbase.org), Gramene Oryza[20] (https://oryza.gramene.org), Gramene Maize (https://maize-pangenome.gramene.org), and Gramene Grapevine (https://vitis.gramene.org), pioneered the adoption of standardized rsIDs for genetic variants in five crops and plant model species, leveraging EVA releases 5 and 6. In total, nearly or almost 220 million custom variant IDs were replaced with stable rsIDs: 78.9 million in maize, 67.7 million in rice, 46.4 million in sorghum, 26.3 million in Arabidopsis, and 315,353 in grapevine. rsID

Category	Main Objective	Key Actions	Results and Resources
Submission	Support community submissions of curated SNP datasets to EVA	• Identified key studies in sorghum and soybean • SorghumBase and SoyBase contacted lead PIs to promote the importance of rsID adoption • EVA and SoyBase provided submission and brokering support to under-resourced research groups	• **Result**: Sorghum rsIDs in EVA increased from 8 M to 55 M • **Result**: Soybean rsIDs in EVA increased from 18 M to 28 M • **Resource**: EVA submission guideline for agricultural species
Database adoption	Integrate rsIDs with traits, germplasm, and functional predictions	• Replaced custom IDs with rsIDs in community databases (*e.g.*, MaizeGDB, SorghumBase, and SoyBase) • Used Ensembl VEP for variant annotation • Aligned sample metadata with available genotyping information	• **Result**:220 M rsIDs adopted across 5 species in Gramene pangenome databases • **Result**: Integrated GWAS/QTLs with germplasm and phenotypes • **Resource**: Enabled seed ordering via GRIN and other repositories
Pangenome projection	Ensure consistency across agricultural pangenomes for variant tracking	• Benchmarking with EBI remapping pipeline • SorghumBase projected rsIDs across sorghum cultivars • Gramene Maize applied approach in maize cultivars	• **Result**: Validated projection accuracy using functional variants (e.g., *Su1*) in sweet corn • **Resource**: Expanded pangenome tools in Gramene Maize
Commercial arrays	Extend rsIDs usage in commercial marker panels	• Developed streamlined rsIDs assignment protocol with EVA • Collaborated with genotyping service providers to foster rsID adoption • Matched array markers to rsIDs	• **Resource**: Assigned rsIDs to 2,421 SCMP and 3,490 DArTag markers • **Resource**: Shared rsIDs lists with 12 providers
Inter-operability	Enable cross-resource integration of variant data	• Integrating EVA rsIDs with Ensembl, BovineMine, MaizeGDB, Gramene, etc. • Linking variants with accessions in germplasm repositories	• **Resource**: Indexed accessions with putative loss-of-function (pLoF) variants across repositories (i.e., variants in Gramene databases linked to germplasm accessions in GRIN, IRRI, MaizeGDB, SorbMutDB) • **Resource**: Created "Germplasm" tab for seed ordering associated with pLoFs in Gramene's search interface

Table 1. Summary of Results and Resources Advancing rsID Adoption and Interoperability in Agricultural Genomics.

adoption enabled consistent variant tracking across assemblies and streamlining germplasm and phenotyping data integration to support marker-assisted breeding. Functional consequences of variants were predicted using the Ensembl Variant Effect Predictor (VEP)[21] incorporating algorithms such as SIFT[22]. These analyses identified putative truncating variants (PTVs) including putative loss-of-function (pLOF) alleles across species. A key outcome was the implementation of a Germplasm Tab in SorghumBase[19], Gramene Oryza[20], and Gramene Maize (Supplementary Information section 3.1). This feature allows users to cross-reference predicted functional variants with the germplasm accessions carrying the corresponding alleles. Researchers can then order seeds directly through instances of the Germplasm Resource Information Network (GRIN)-Global, such as the International Rice Research Institute (IRRI, see https://www.irri.org/genesys-rice#/a/0385b84d-cad1-4067-80fa-16e01984d5a5). Interoperability with other germplasm resources was also implemented between sorghum pLoF variants and SorbMutDB[23] (https://www.depts.ttu.edu/igcast/SorbMutDB.php), and integrating Gramene Maize variants with MaizeGDB's SNPVersity[24]. This process revealed critical biocuration challenges, such as inconsistent cultivar spelling and synonyms, underscoring the parallel need for globally unique identifiers for biosamples.

Building on this interoperable framework, SorghumBase replaced temporary IDs, with permanent rsIDs, and integrated these standardized variants with GWAS phenotypes. Sorghum QTLs were further linked to corresponding entries in the OZ Sorghum QTL Atlas[25] (https://aussorgm.org.au/sorghum-qtl-atlas), representing the third stage of the data journey (Fig. 1c). This integration supports phenotype prediction, and enhances trait-based marker discovery (Figure S2, see Supplementary Information document). Similarly, MaizeGDB[26] (https://maizegdb.org; see Supplementary Information section 3.2), is in the second stage of the data journey (Fig. 2b), extending its variant-processing pipeline to assign stable rsIDs at the point of data intake, using EVA release 7. By integrating rsIDs with standard germplasm identifiers, MaizeGDB is enabling direct linking to and ordering from major repositories like the USDA-ARS National Clonal Germplasm System (NPGS) through the GRIN-Global database (https://npgsweb.ars-grin.gov/gringlobal/search), the International Maize and Wheat Improvement Center (CIMMYT, https://www.cimmyt.org/work/seed-request), and the Maize Genetics Cooperation Stock Center (https://maizecoopsc.org/about-our-collection). Standardizing phenotypic variation is crucial for integrating it with genetic variant data from GWAS and QTLs. For example, MaizeGDB hosts over 300 traits associated with more than 40,000 genomic positions, curated from three major GWAS datasets[27–29]. Where rsIDs were not initially assigned, they were added for cases where a GWAS SNP mapped to the same position as a variant existing in the EVA release. This integrated data is presented as genome browser tracks and tables on gene model pages.

Other community databases like TreeGenes[30] and CartograPlant[31] for forest trees, and GrainGenes[32] and the T3/Wheat Triticeae Toolbox[33] for small grain cereals, are in earlier phases of adopting rsIDs and implementing similar practices (Supplementary Information sections 3.3 and 3.4).

Together, these examples illustrate how rsID adoption across agricultural community databases enhances the integration of genetic variation with functional predictions, trait data, and germplasm accessions. This infrastructure forms a critical foundation for reproducible, interoperable, and FAIR-aligned data practices in agricultural research.

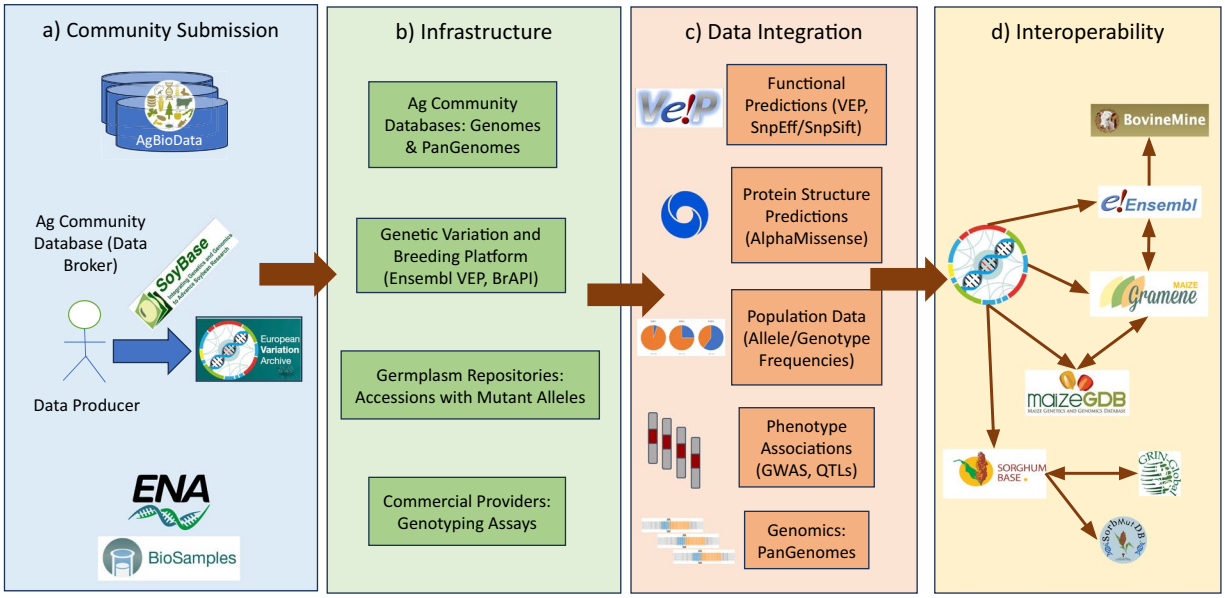

Fig. 2 The agricultural genetic variation data journey enabled by rsID adoption. The diagram illustrates the lifecycle of standardized variant data across four stages: (**a**) Community Submission: Data producers submit variant datasets to agricultural community databases (acting as data brokers, e.g., SoyBase), which facilitate validation and submission to the European Variation Archive (EVA) for the assignment of persistent rsIDs. (**b**) Infrastructure: This standardized data is ingested into a foundational infrastructure, including community genome databases, pangenome collections, and commercial genotyping assays. (**c**) Data Integration: rsIDs serve as stable anchors for downstream analysis, enabling the integration of variants with functional predictions (VEP, SnpEff), protein structure models (AlphaMissense), population allele frequencies, and phenotypic associations (GWAS/QTLs). (**d**) Interoperability: The final stage represents a network where rsIDs link distinct bioinformatic resources (e.g., Ensembl, Gramene, MaizeGDB, BovineMine) with germplasm repositories (e.g., GRIN-Global, SorbMutDB), allowing researchers to trace specific variants from genomic data directly to physical seed accessions.

Pioneering rsIDs assignment across agricultural pangenomes. As more agricultural pangenomes are sequenced, calling variants separately for each cultivar becomes increasingly impractical. A more scalable solution is to project existing rsIDs onto related cultivar assemblies. This approach is especially important for ensuring consistency in functional studies and gene editing applications. For example, in sorghum, the reference genome BTx623 is not commonly used in molecular systems such as CRISPR, whereas the cultivar RTx430 is preferred due to its higher transformation and regeneration efficiency[34–37]. Projecting rsIDs across these genomes ensures that variant data remains usable and interoperable across experimental platforms.

SorghumBase and Gramene Maize tested this strategy using the EBI Variation variant-remapping pipeline (see Methods) in sorghum[19] and maize (Fig. 3). According to EBI's documentation, the pipeline had worked well between different assemblies of the same genome or closely related genomes with SNPs and short Indels, but it had not been tested with larger or more complex variants. This approach was refined in maize, where gene-centric subsets of rsIDs within 1.5 Kb flanking regions of the reference maize B73 gene models were projected onto the genomes of 25 maize genomes from the Nested Association Mapping (NAM) panel and made available as genome tracks in Gramene Maize (see https://maize-pangenome.gramene.org/News?section=Release%205). This method enabled accurate rsIDs mapping within and across genomes of closely related accessions, offering significant potential to accelerate breeding efforts. Of note, MaizeGDB has also started propagating rsIDs across maize pangenomes.

Standardizing commercial genotyping arrays facilitates long-term variant data reuse. SNP genotyping arrays are powerful tools widely used in agricultural research and breeding. They support molecular genetics studies by facilitating the discovery of QTLs, GWAS analyses, and inference of kinship and parentage. In genomic and molecular breeding, these arrays facilitate genomic prediction, genomic selection, and marker-assisted selection to improve breeding efficiency. They also play a key role in cultivar development by accelerating the release of new varieties, identifying valuable genetic markers in existing germplasm, and selecting germplasm suited to specific growing environments. Additionally, SNP arrays are critical for intellectual property protection through the verification of clonal purity and confirmation of parent-offspring relationships.

rsIDs allow for better tracking of variants across genome assembly versions by mapping, tracking, merging and deprecating identifiers, which would assist genotyping platform upgrades. In collaboration with Gramene and SorghumBase, the EVA developed a protocol to engage commercial genotyping providers in generating rsIDs for species with incomplete coverage of rsIDs in EVA. Contact was established with 12 genotyping service providers and breeding companies to advocate for the use of rsIDs as a variation identifier in their arrays.

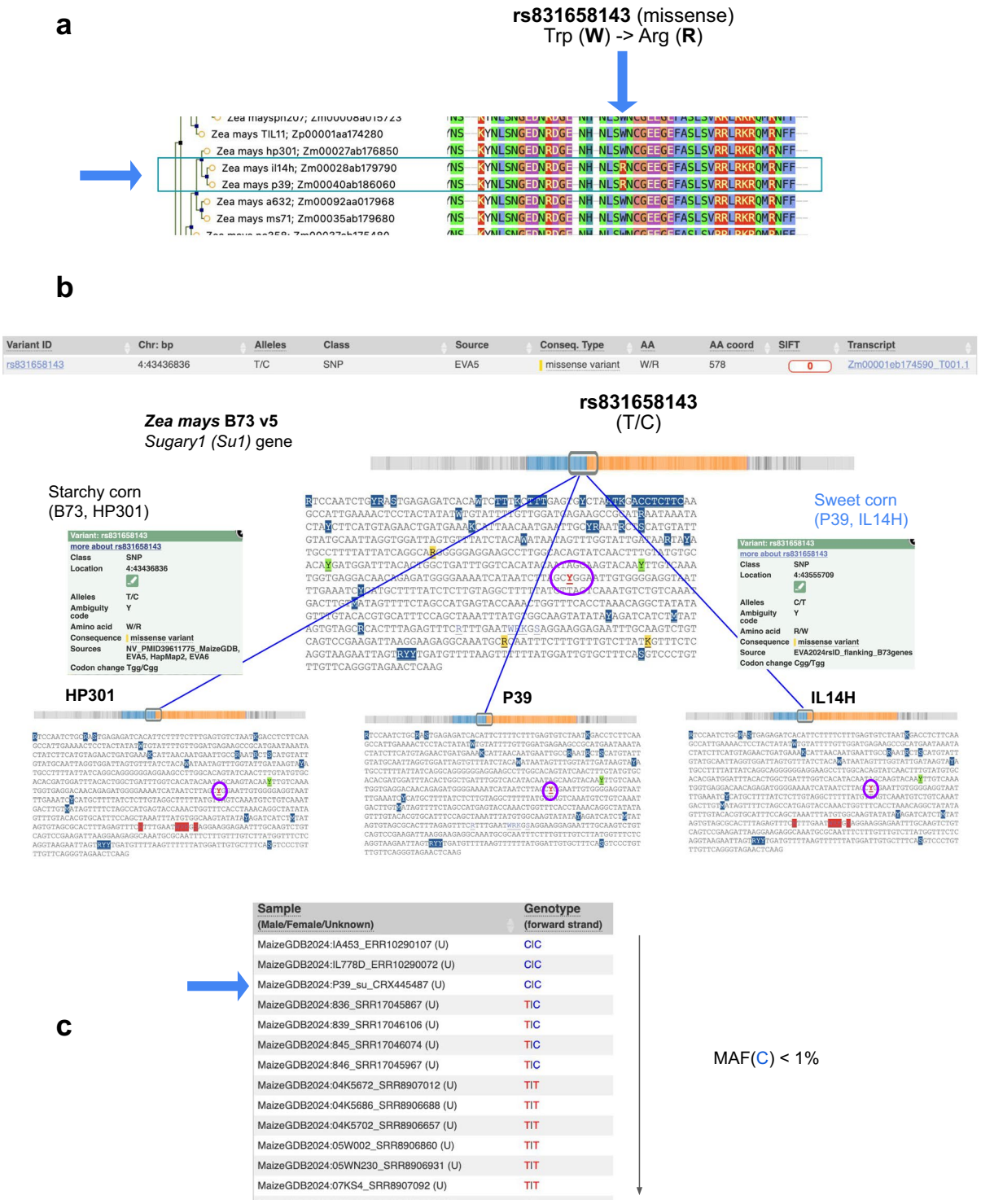

Fig. 3 rsID-mediated projection of a functional *Sugary1 (Su1)* variant identifies sweet corn mutants in the Gramene Maize pangenome browser. (**a**) Sweet corn arises from mutations in starch biosynthesis pathway genes, such as *Su1*, that increase sugar content in the endosperm at the expense of starch[62–64]. The missense variant rs831658143 (alleles T/C) is located at amino acid residue 578 towards the 3′ end of a highly conserved N-terminal glycoside hydrolase domain in *Su1*(blue box), which plays a role in converting starchy corn into sweet corn[65]. Shown is a zoomed-in view of the amino acid alignment for the *Su1* gene family tree highlighting the tryptophan-to-arginine substitution at residue 578 (W578R) that was previously reported[62,63,66]. The alternative allele C at this variant locus (blue arrow) is observed in two sweet corn varieties: P39 and IL14H. The gene tree was constructed using 60 plant genomes including 54 distinct maize varieties. (**b**) Flanking sequence context for rs831658143 in the reference maize B73 genome and three other accessions (HP301, P39, and IL14H). Sweet corn varieties carry the alternative allele C encoding arginine, whereas starchy corn varieties retain the reference T allele encoding tryptophan. The variant position is highlighted in red (purple circle), with nearby variants indicated by colored letters and IUPAC ambiguity codes. In the reference B73 genome,

this position is homozygous for the T allele. The W-to-R substitution constitutes a missense mutation with a SIFT score of zero, indicating a high likelihood of deleterious impact. (**c**) The Gramene Maize browser displays genotypes for 1,438 accessions at rs831658143, including three homozygous sweet corn varieties (IA453[67], IL778D, and P39_su), and four heterozygous landraces (835, 836, 839, and 846). This corresponds to a minor allele frequency (MAF) of <1%, indicating that fewer than 1% of accessions bear the C allele[24].

The protocol involves reaching out to these providers to explain the benefits of incorporating rsIDs into their workflows and offering a streamlined process through which EVA can assign rsIDs to valid markers not currently covered, without requiring full submission through the archive's standard procedure (https://www.ebi.ac.uk/eva/?Submit-Data). We applied this protocol to a new Sorghum Community Marker Panel (SCMP), a mid-density array developed by the sorghum research community in collaboration with AgriPlex[38]. The SCMP consists of 2,421 markers that were curated from existing resources including the DArTag genotyping service (https://excellenceinbreeding.org/toolbox/services/sorghum-mid-density-genotyping-services) offered by Diversity Arrays Technology Ltd for the Consultative Group on International Agricultural Research (CGIAR), a community Kompetitive Allele Specific PCR (KASP) panel (https://excellenceinbreeding.org/module3/kasp), quality control markers, and trait-associated markers suggested by breeders. Of the 2,421 SCMP variants, 2,395 were matched to existing rsIDs in EVA release 5. The remaining 26 variants received new rsIDs without having to wait for the next release. All 2,421 rsIDs (Table S2, see supplementary xlsx file) are now included in EVA release 7 for Sorghum bicolor (NCBI assembly GCA_000003195.3). Tailored for the US sorghum community, SCMP is well suited for marker-assisted selection, genetic purity testing and germplasm quality control.

Stable identifiers enable resource interoperability. The EVA plays a central role in enabling data integration across resources. For example, Ensembl[39] (http://www.ensembl.org; see Supplementary Information section 3.5) imports variation data from several public sources, with data for agricultural species primarily coming from the EVA. Its new beta infrastructure at https://beta.ensembl.org (Figure S3, see Supplementary Information document) directly incorporates EVA-submitted datasets, supporting 55 species, 27 of which are agriculturally relevant (12 crops, 10 livestock, and 5 aquaculture), as of release 7. This evolution strengthens the connection between agricultural variant data and Ensembl's broader annotation and visualization ecosystem, enhancing accessibility and comparative analysis capabilities.

Stable identifiers are essential for connecting genetic, phenotypic, and germplasm data across diverse bioinformatics resources. BovineMine[40] (https://bovinemine.rnet.missouri.edu; Supplementary Information section 3.6), the data warehouse of the Bovine Genome Database, exemplifies this by integrating single-nucleotide variants and variant effect predictions from Ensembl[39], with QTL and GWAS data from AnimalQTLdb[41] (https://www.animalgenome.org/cgi-bin/QTLdb/index), all mapped to the latest bovine genome assembly. Trait associations are annotated using the Vertebrate Trait Ontology[42], while variants are indexed using rsIDs and include SNP array aliases to support flexible querying. Through InterMine[43] tools such as QueryBuilder (https://bovinemine.rnet.missouri.edu/bovinemine/customQuery.do) and Genomic Regions search (https://bovinemine.rnet.missouri.edu/bovinemine/genomicRegionSearch.do), users can retrieve, filter, and combine genomic, functional, and trait data, enabling cross-platform integration, reproducible analyses, and customized exploration of genotype-to-phenotype relationships in cattle.

In agricultural systems, interoperability also relies on consistent identifiers for germplasm accessions. Community databases like MaizeGDB, SorghumBase, and Gramene Oryza support cross-referencing with major germplasm repositories using standardized sample identifiers, such as PI and GSOR numbers from GRIN, and IS numbers from the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT, https://genebank.icrisat.org/IND/orderGermplasmDashboard). These identifiers allow researchers and breeders to easily locate and request seeds for accessions of interest.

SorghumBase, Gramene Maize and Gramene Oryza have further extended interoperability by implementing a Germplasm tab within gene search interfaces, linking gene-level variation data with stock center records (Figure S2, see Supplementary Information document). These interfaces index accessions carrying variants predicted by the Ensembl VEP to truncate protein function. A curated metadata layer maps population samples to standardized accession names across stock centers, including GRIN-Global, IRRI, and SorbMutDB, allowing researchers to order both natural accessions and ethyl methanesulfonate (EMS)-induced mutants bearing pLoF alleles.

FAIR Implementation Guide for Agricultural Genetic Variation Data. We have developed a FAIR Implementation Guide (see Supplementary Information annex) to standardize the sustainable submission, management, and use of variation datasets across diverse agricultural research contexts. This guide addresses critical questions such as how to verify whether existing variants already have assigned rsIDs in the EVA, how to submit new SNP data to EVA, including scenarios where an accessioned reference genome exists for the species in the INSDC, and how to submit one if it does not. It also provides guidance on how to ensure that the variant data is provided in a standardized, validated format such as VCF, including VCF metadata requirements, and the use of globally recognized sample identifiers from major germplasm repositories, such as BioSample or GRIN to ensure traceability and interoperability. The guide enables researchers in adopting the best practices to produce FAIR-compliant variant datasets that are reusable, interoperable and ready for integration in global genomic resources.

We have also advocated for the importance of adopting FAIR standards for genetic variant data, and particularly for the adoption of rsIDs via presentations at major conferences, agricultural workshops, and collaborator

meetings. Genetic variation data producers are encouraged to use standard formats like VCF (preferably using rsIDs when available), submitting key SNP data sets to the EVA to obtain rsIDs, and adopting standard sample identifiers associated with a major germplasm repositories like the NPGS, ICRISAT or the Genebank Information System of the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK-GBIS).

Discussion

SNPs are among the most important forms of genetic variation and are foundational to genomic research, including GWAS, marker-assisted selection, and population diversity analysis. Accurate SNP calling depends on high-quality read mapping, variant calling, and variant filtering, all of which benefit from improved reference genomes and bioinformatics tools.

Historically, many crop species (particularly clonally propagated or polyploid species) lacked high-quality, accessioned reference genomes due to challenges such as genome repetitiveness, polyploidy, and heterozygosity. Recent advances in long-read sequencing technologies[44] applied to the creation of telomere-to-telomere (T2T) assemblies have helped close longstanding genomic gaps, particularly in repetitive regions such as centromeres and telomeres. This has led to the discovery of novel sequences and previously undetectable SNPs, providing a more complete picture of genetic variation essential for understanding disease resistance, adaptation, and complex trait architecture, which facilitates its use in breeding programs. These technological improvements have also enabled the construction of pangenomes, which offer more accurate placement of short reads and improved variant calling across genetically diverse populations. These enhancements reduce false positives, recover novel SNPs, and ultimately improve the resolution of domestication history and population diversity studies, as demonstrated in humans[45] and goats[46].

Despite these advances, agricultural genomics has lagged behind human genomics in adopting standardized variant identifiers and shared formats. Human genetic databases use rsIDs as stable, assembly-independent references to specific variant loci. This consistency enables integration across genome builds, platforms, and publications, and supports large-scale meta-analysis and data reuse. By contrast, agricultural SNP databases have often been siloed and fragmented, leading to inconsistencies and reduced resource interoperability.

The use of rsIDs represents a critical step forward for agricultural genomics. They serve as persistent identifiers that allow researchers to unambiguously refer to the same variant across genome browsers, literature, phenotypic databases, and functional annotations. In human studies, rsIDs are central to GWAS and ClinVar[6] catalogs, genotyping arrays, and literature indexing tools such as PubMed (https://pubmed.ncbi.nlm.nih.gov) and LitVar[47]. Resources hosting SNP variant data and genotype-to-phenotype (G2P) associations for agricultural species, such as Genome Variation Map[48], AnimalQTLdb[41], Agricultural Animal Omics Database (http://animal.omics.pro) and related variation databases[49–52]), and OMIA[53], are examples of databases that integrate rsIDs. Through our efforts, rsID adoption and implementation has begun to spread to agricultural community resources like SorghumBase and BovineMine, enabling integration of variant data with germplasm resources and trait information. Other knowledgebases focused on variant cataloging for multiple agricultural species and that were not part of our group, like GWAS Atlas[54] and Animal-SNP Atlas[55], could greatly benefit from rsID adoption to support integration with agro-functional genomics and breeding pipelines. Despite these benefits, the adoption of rsIDs in newly emerging crops or livestock genomics initiatives remains challenging. Key bottlenecks include the need for coordination among community databases, breeding programs, and germplasm repositories; the availability of a sufficiently mature reference genome accessioned in the INSDC; limited submission capacity and familiarity with EVA workflows in smaller laboratories; and the retrospective harmonization of incompatible, pre-existing SNP naming schemes. Based on our experience, these challenges can be mitigated by seeding variant catalogs from existing genotyping panels or published GWAS datasets, establishing an initial set of ~10K–50 K high-confidence variants to demonstrate value, assessing whether rsIDs already exist in the EVA, and promoting lightweight tools and concise training materials. Even simple liftOver workflows and short submission guides, adapted from community resources such as the ELIXIR FAIR Cookbook (https://faircookbook.elixir-europe.org) and our FAIR User Guide (see Supplementary Information annex), could substantially lower barriers and accelerate rsID adoption in new agricultural genomics efforts.

Most agricultural SNP studies have relied on a single reference genome, but the adoption of pangenomic approaches has revealed previously missed variants and allowed for correction of misassigned SNPs. For example, improved mapping quality in pangenome contexts reduces spurious SNPs and improves detection of variants linked to economically important traits[56]. Some rsIDs initially assigned to orthologous loci can be revised as reference genomes improve, underscoring the value of having flexible but persistent identifiers.

In this study, we demonstrate that the use of standard rsIDs facilitates merging variant datasets across studies, even when using different reference assemblies or genotyping technologies. We also model how rsIDs support integration with GWAS trait data and improve coordination across databases cataloging pLoF variants. The result is a more interoperable, reusable data ecosystem that supports pangenomic comparisons, cross-platform meta-analyses, and more efficient discovery of genotype-phenotype relationships.

Critically, rsIDs enable compatibility with widely used tools for functional prediction (e.g., Ensembl VEP[21]), protein structure modeling (e.g., AlphaMissense[57]), and literature tracking (e.g., LitVar[47]). They also promote long-term dataset sustainability by enabling variant tracking across genome builds and pangenomes. This makes them indispensable for integration across breeding databases (e.g., BreedBase[58], DeltaBreed[59]), pangenome browsers (e.g., SorghumBase, Gramene), and germplasm repositories (e.g., NPGS, ICRISAT, IPK-GBIS). Also providing an anchor point for integrating multi-omics datasets (e.g., transcriptomics, epigenomics, metabolomics), facilitating systems-level analyses of trait regulation.

The impact of this approach is already evident. Our group's work facilitated the assignment of standardized rsIDs for 47 million sorghum and 10 million soybean variants in the EVA. To encourage widespread adoption of rsIDs by agricultural communities, we are providing a stepwise FAIR Genetic Variation User Guide (see

Supplementary Information annex) according to the maturity level of a community in the data journey. For example, for communities with an INSDC-accessioned reference genome and existing variant datasets, this might be achieved with simple tools such as bcftools annotate or custom scripts to retroactively assign rsIDs, provided the sequence and alleles match known entries in EVA. In sorghum, rsIDs enabled integration of genotypic data with phenotypic traits and germplasm collections, facilitating direct access to seed materials through platforms like GRIN-Global.

The need for rsID adoption is also seen in commercial genotyping platforms. For example, the strawberry community has a 50 K SNP genotyping array[60] that includes SNPs tied to multiple reference genomes, making it necessary to map SNP identifiers across platforms. The absence of rsIDs limited interoperability and backward compatibility with prior arrays, while incorporating rsIDs into such platforms would greatly enhance their long-term utility and data traceability.

Going forward, adoption of rsIDs must extend to autopolyploid crops such as alfalfa and blueberry, which pose additional challenges for variant representation. Special considerations are also needed for targeted-amplicon approaches (e.g., DArTag) and for integrating rsIDs with microhaplotype-based identifiers to support haplotype-level selection in breeding programs.

Realizing the full potential of rsID standardization across agricultural genomics requires a coordinated and sustained effort across the research community. While the work presented here demonstrates the feasibility and impact of integrating rsIDs in agricultural pangenome resources, widespread adoption will depend on a multi-pronged strategy involving data infrastructure, tool development, policy alignment, and community engagement.

First, data submission pipelines must be strengthened. Researchers are encouraged to submit high-quality, validated VCF files along with complete sample metadata to the EVA, ensuring that associated reference genomes are accessioned in the INSDC[10], such as ENA, NCBI, or DDBJ. Comprehensive guidance for these submissions is available through the ELIXIR FAIR Cookbook (e.g., https://w3id.org/faircookbook/FCB061) and the FAIR Implementation Guide we developed (see Supplementary Information annex), which emphasize best practices for data standardization and reproducibility.

Second, the ability to propagate rsIDs across multiple genome assemblies is critical for ensuring variant consistency across diverse germplasm, including landraces, breeding lines, and elite cultivars. Tools such as the EVA remapping workflow provide the technical foundation for this process, supporting alignment across the rapidly growing landscape of pangenomes.

Third, community databases and bioinformatics platforms must be updated to integrate rsIDs within both backend systems and user-facing tools. Embedding rsIDs within variant browsers, functional annotation pipelines, and germplasm access portals will enhance traceability, support G2P association studies, and streamline access to seeds and tissue materials linked to functionally relevant variants.

Equally important is policy-level advocacy. Encouraging journals, funding agencies, and breeding programs to recommend or mandate the use of rsIDs in publications and public data releases will help establish consistent norms. Such policy alignment would not only accelerate adoption but also improve data interoperability across disciplines and repositories. Beyond policy mandates, the data management systems employed by agricultural researchers and plant and animal breeders day-to-day currently do not support storage, display, or retrieval of rsIDs. This is in part due to the lack of annotated rsIDs for many crops, but also confounded by the absence of breeder-defined use cases that require or involve rsIDs. Without defined use cases, software developers cannot create visualizations, analytics, or other features in breeding data platforms to incorporate rsIDs. If and when breeders start to see the value of rsIDs in making more data-driven decisions, and concurrent with the more widespread annotation and adoption of rsIDs across crops, those breeders can make defined feature requests to integrate these data standards into their management software platforms and workflows.

Finally, ongoing community training and support are essential to bridging adoption gaps, particularly for under-resourced crop communities. Targeted outreach, technical documentation, and hands-on workshops will help lower barriers to entry and ensure that all researchers, regardless of crop system or region, can participate in and benefit from rsID-based variant tracking. Without coordinated community efforts, particularly bulk submissions of novel variants and standardized practices, the benefits of rsIDs will remain unrealized.

At the same time, challenges remain in ensuring accurate variant calling in highly repetitive regions, achieving representation of under-studied crops, and securing long-term funding for data infrastructure. Addressing these limitations will be key to realizing the full potential of rsID-based frameworks.

We also advocate for development of expert-curated frameworks, similar to the ACMG/AMP guidelines[61] in human genetics, for interpreting variant causality in crops. These would provide a consistent basis for linking genotypes to agriculturally important traits across species and breeding systems.

In summary, the delayed adoption of rsIDs in agricultural genomics has constrained progress toward data reuse, interoperability, and large-scale integration. But the field is now at a critical inflection point. With infrastructure in place, community momentum building, and successful implementation examples like maize, soybean, and sorghum, the opportunity is ripe to establish rsIDs as a universal standard for agricultural variation. This transition is not simply a technical upgrade, it is a foundational shift that will enable FAIR data principles, accelerate crop improvement, and link decades of genetic research through a shared, reusable framework.

Methods

Our methodology followed a three-stage process to standardize agricultural genetic variation: (1) assessing actionable community needs to define data standards, (2) developing technical pipelines to replace temporary identifiers in existing databases, and (3) establishing workflows to project these identifiers across pangenomes.

Community assessment and definition of standards. In order to identify specific barriers to interoperability, we convened a consistent group of agricultural bioinformatics experts including 5-8 major agricultural bioinformatics resources and a broader membership of approximately 35 researchers. We assessed current practices within the wider AgBioData community through anonymous surveys and workshops, gathering insights on the use of different data types, quality control measures, and the adoption of stable identifiers across member resources (Supplementary Information). These assessments identified two primary technical requirements: the need for standardized biosample identifiers and broader adoption of rsIDs. Based on these findings, we consulted with USDA-affiliated scientists and germplasm curators to review and refine recommendations for making agricultural genotyping data FAIR, with a focus on VCF formatting and metadata alignment. Together with a targeted review of published SNP data sets co-authored by members of the AgBioData Consortium, as well as the identification of unpublished SNP datasets and data standardization challenges through community presentations, this effort informed the development of a "Data Journey" model that illustrates stages of rsID adoption. Such adoption was subsequently extended to pangenomes, supported by community-developed tools and templates for INSDC submission and rsID generation, as described below. Workshop outcomes have been described in Table S1 (see supplementary xlsx file), and metadata guidelines are described in the FAIR User Guide to rsID adoption (see Supplementary Information annex).

Implementation of rsIDs in community databases. To facilitate the retroactive assignment of stable identifiers, we developed a custom Perl script *import_variant_rsid_all.pl* (https://github.com/warelab/gramene-ensembl/blob/e506aaf4001340e91351f0b620dffb9aea25651d/scripts/load-scripts/import_variant_rsid_all.pl) that processes VCF files from the EVA (releases 5 and 6) to update existing community databases (such as Gramene). The workflow involves downloading full rsID sets for target species (https://ftp.ebi.ac.uk/pub/databases/eva/rs_releases/), splitting them by chromosome for parallel processing, and matching them against local database records using genomic position and allele identity. When a match is identified, the script replaces the temporary internal identifier with the standard rsID and archives the original name in a synonym table to maintain backward compatibility; if no match is found, a new variation record is created.

Projection of rsIDs across agricultural pangenomes. We adapted the EBI variant remapping pipeline (https://github.com/EBIvariation/variant-remapping/blob/master/README.md) to extend rsID consistency across pangenome assemblies by projecting identifiers from a reference genome to target cultivar assemblies. This pipeline uses the reference assembly flanking sequences of the rsID to anchor it to the target assembly, implementing a tiered mapping strategy using increasing flanking sequence sizes (200 bp, 5 Kb and 50 Kb) based on benchmarking results from sorghum. In order to increase the specificity and reduce the computing cost, we partitioned the rsIDs into chromosome-specific subsets, and restricted mapping to the corresponding chromosomes.

Data availability

The remapped rsIDs in the Gramene Maize PanGenome are available in VCF at https://ftp.gramene.org/maize/release-current/variation/rsID_remapping/. The data sources for BovineMine are available at http://bovinegenome.org/bovinemine/dataCategories.do.

Code availability

Code from the EBI variant remapping pipeline is available at https://github.com/EBIvariation/variant-remapping/tree/master. Complementary custom scripts for rsIDs mapping, including *import_variant_rsid_all.pl*, are available from https://github.com/warelab/gramene-ensembl/blob/e506aaf4001340e91351f0b620dffb9aea25651d/scripts/load-scripts/.

Received: 14 October 2025; Accepted: 1 April 2026;
Published online: 16 April 2026

References

1. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018, https://doi.org/10.1038/sdata.2016.18 (2016).
2. Harper, L. *et al.* AgBioData consortium recommendations for sustainable genomics and genetics databases for agriculture. *Database* **2018**, bay088, https://doi.org/10.1093/database/bay088 (2018).
3. Sherry, S. T., Ward, M. & Sirotkin, K. dbSNP — database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res.* **9**, 677–679, https://doi.org/10.1101/gr.9.8.677 (1999).
4. Phan, L. *et al.* The evolution of dbSNP: 25 years of impact in genomic research. *Nucleic Acids Res.* **53**, D925–D931, https://doi.org/10.1093/nar/gkae977 (2025).
5. Fokkema, I. F. A. C. *et al.* LOVD v.2.0: the next generation in gene variant databases. *Hum. Mutat.* **32**, 557–563, https://doi.org/10.1002/humu.21438 (2011).
6. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067, https://doi.org/10.1093/nar/gkx1153 (2018).
7. Stenson, P. D. *et al.* The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum. Genet.* **133**, 1–9, https://doi.org/10.1007/s00439-013-1358-4 (2014).
8. Cezard, T. *et al.* The European Variation Archive: a FAIR resource of genomic variation for all species. *Nucleic Acids Res.* **50**, D1216–D1220, https://doi.org/10.1093/nar/gkab960 (2022).
9. Van Buren, S. L., Brown, C. T., Szmatoła, T. & Finno, C. J. A comparative review of short genetic variant databases across humans and animal species. *Brief. Bioinform.* **26**, bbaf356, https://doi.org/10.1093/bib/bbaf356 (2025).
10. Karsch-Mizrachi, I. *et al.* The international nucleotide sequence database collaboration (INSDC): enhancing global participation. *Nucleic Acids Res.* **53**, D62–D66, https://doi.org/10.1093/nar/gkae1058 (2025).

11. Kumar, N., Boatwright, J. L., Boyles, R. E., Brenton, Z. W. & Kresovich, S. Identification of pleiotropic loci mediating structural and non-structural carbohydrate accumulation within the sorghum bioenergy association panel using high-throughput markers. *Front. Plant Sci.* **15**, 1356619, https://doi.org/10.3389/fpls.2024.1356619 (2024).
12. Cruet-Burgos, C., Morris, G. P. & Rhodes, D. H. Characterization of grain carotenoids in global sorghum germplasm to guide genomics-assisted breeding strategies. *BMC Plant Biol.* **23**, 165, https://doi.org/10.1186/s12870-023-04143-9 (2023).
13. Boatwright, J. L. *et al*. Sorghum Association Panel whole-genome sequencing establishes cornerstone resource for dissecting genomic diversity. *Plant J.* **111**, 888–904, https://doi.org/10.1111/tpj.15853 (2022).
14. Brown, A. V. *et al*. A new decade and new data at SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Res.* **49**, D1496–D1501, https://doi.org/10.1093/nar/gkaa1107 (2021).
15. Zhang, H. *et al*. Development of a versatile resource for post-genomic research through consolidating and characterizing 1500 diverse wild and cultivated soybean genomes. *BMC Genomics* **23**, 250, https://doi.org/10.1186/s12864-022-08326-w (2022).
16. Torkamaneh, D. *et al*. Soybean (*Glycine max*) Haplotype Map (GmHapMap): a universal resource for soybean translational and functional genomics. *Plant Biotechnol. J.* **19**, 324–334, https://doi.org/10.1111/pbi.13466 (2021).
17. Torkamaneh, D. *et al*. Comprehensive description of genomewide nucleotide and structural variation in short-season soya bean. *Plant Biotechnol. J.* **16**, 749–759, https://doi.org/10.1111/pbi.12825 (2018).
18. Olson, A. *et al*. Gramene 2025: expanded comparative genomics and pathway resources, integrated search, and pangenome portals for crop research. *Nucleic Acids Res.* **54**, D1720–D1732, https://doi.org/10.1093/nar/gkaf1260 (2026).
19. Gladman, N. *et al*. SorghumBase: a knowledgebase for sorghum genomics, phenomics, and stakeholder engagement. *Genetics*, https://doi.org/10.1093/genetics/iyaf266 (2026).
20. Wei, S. *et al*. GrameneOryza: a comprehensive resource for *Oryza* genomes, genetic variation, and functional data. *Database* **2025**, baaf021, https://doi.org/10.1093/database/baaf021 (2025).
21. McLaren, W. *et al*. The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122, https://doi.org/10.1186/s13059-016-0974-4 (2016).
22. Ng, P. C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res.* **11**, 863–874, https://doi.org/10.1101/gr.176601 (2001).
23. Jiao, Y. *et al*. A large sequenced mutant library — valuable reverse genetic resource that covers 98% of sorghum genes. *Plant J.* **117**, 1543–1557, https://doi.org/10.1111/tpj.16582 (2024).
24. Andorf, C. M., Ross-Ibarra, J., Seetharam, A. S., Hufford, M. B. & Woodhouse, M. R. A unified VCF dataset from nearly 1,500 diverse maize accessions and resources to explore the genomic landscape of maize. *G3 (Bethesda)* **15**, jkae281, https://doi.org/10.1093/g3journal/jkae281 (2025).
25. Mace, E. *et al*. The Sorghum QTL Atlas: a powerful tool for trait dissection, comparative genomics and crop improvement. *Theor. Appl. Genet.* **132**, 751–766, https://doi.org/10.1007/s00122-018-3212-5 (2019).
26. Woodhouse, M. R. *et al*. A pangenomic approach to genome databases using maize as a model system. *BMC Plant Biol.* **21**, 385, https://doi.org/10.1186/s12870-021-03173-5 (2021).
27. Li, C. *et al*. Genomic insights into historical improvement of heterotic groups during modern hybrid maize breeding. *Nat. Plants* **8**, 750–763, https://doi.org/10.1038/s41477-022-01190-2 (2022).
28. Tian, D. *et al*. GWAS Atlas: a curated resource of genome-wide variant–trait associations in plants and animals. *Nucleic Acids Res.* **48**, D927–D932, https://doi.org/10.1093/nar/gkz828 (2020).
29. Wallace, J. G. *et al*. Association mapping across numerous traits reveals patterns of functional variation in maize. *PLoS Genet.* **10**, e1004845, https://doi.org/10.1371/journal.pgen.1004845 (2014).
30. Falk, T. *et al*. Growing and cultivating the forest genomics database, TreeGenes. *Database* **2018**, bay084, https://doi.org/10.1093/database/bay084 (2018).
31. Lind, B. *et al*. CartograPlant: bridging genomic, phenotypic, and environmental data to advance plant resilience and eco-evolutionary insight. *Genetics,* https://doi.org/10.1093/genetics/iyag060 (2026).
32. Yao, E. *et al*. GrainGenes: genetics, genomes, and pangenomes. *Genetics.* https://doi.org/10.1093/genetics/iyaf270 (2025).
33. Blake, V. C. *et al*. The Triticeae Toolbox: combining phenotype and genotype data to advance small-grains breeding. *Plant Genome* **9**, plantgenome2014.12.0099, https://doi.org/10.3835/plantgenome2014.12.0099 (2016).
34. Che, P. *et al*. Developing a flexible, high-efficiency Agrobacterium-mediated sorghum transformation system with broad application. *Plant Biotechnol. J.* **16**, 1388–1395, https://doi.org/10.1111/pbi.12879 (2018).
35. Liu, G., Li, J. & Godwin, I. D. Genome editing by CRISPR/Cas9 in sorghum through biolistic bombardment. *Methods Mol. Biol.* **1931**, 169–183, https://doi.org/10.1007/978-1-4939-9039-9_12 (2019).
36. Brant, E. J., Baloglu, M. C., Parikh, A. & Altpeter, F. CRISPR/Cas9 mediated targeted mutagenesis of *LIGULELESS-1* in sorghum provides a rapidly scorable phenotype by altering leaf inclination angle. *Biotechnol. J.* **16**, e2100237, https://doi.org/10.1002/biot.202100237 (2021).
37. Massel, K. *et al*. CRISPR-knockout of β-kafirin in sorghum does not recapitulate the grain quality of natural mutants. *Planta* **257**, 8, https://doi.org/10.1007/s00425-022-04038-3 (2022).
38. Kumar, V. *et al*. Development and evaluation of a cost-effective, mid-density SNP array as a sorghum community genotyping resource. Preprint at *bioRxiv*. https://doi.org/10.64898/2026.02.20.706663 (2026).
39. Dyer, S. C. *et al*. Ensembl 2025. *Nucleic Acids Res.* **53**, D948–D957, https://doi.org/10.1093/nar/gkae1071 (2025).
40. Kambal, S. *et al*. Bovine Genome Database: new curated collection of selective sweeps in bovine populations across the world. *Nucleic Acids Res.* **54**, D949–D957, https://doi.org/10.1093/nar/gkaf1214 (2026).
41. Hu, Z.-L., Park, C. A. & Reecy, J. M. Bringing the Animal QTLdb and CorrDB into the future: meeting new challenges and providing updated services. *Nucleic Acids Res.* **50**, D956–D961, https://doi.org/10.1093/nar/gkab1116 (2022).
42. Park, C. A. *et al*. The Vertebrate Trait Ontology: a controlled vocabulary for the annotation of trait data across species. *J. Biomed. Semant.* **4**, 13, https://doi.org/10.1186/2041-1480-4-13 (2013).
43. Smith, R. N. *et al*. InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. *Bioinformatics* **28**, 3163–3165, https://doi.org/10.1093/bioinformatics/bts577 (2012).
44. Wenger, A. M. *et al*. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162, https://doi.org/10.1038/s41587-019-0217-9 (2019).
45. Ameur, A. *et al*. De novo assembly of two Swedish genomes reveals missing segments from the human GRCh38 reference and improves variant calling of population-scale sequencing data. *Genes* **9**, 486, https://doi.org/10.3390/genes9100486 (2018).
46. Li, R. *et al*. Towards the complete goat pangenome by recovering missing genomic segments from the reference genome. *Front. Genet.* **10**, 1169, https://doi.org/10.3389/fgene.2019.01169 (2019).
47. Allot, A. *et al*. LitVar: a semantic search engine for linking genomic variant data in PubMed and PMC. *Nucleic Acids Res.* **46**, W530–W536, https://doi.org/10.1093/nar/gky355 (2018).
48. Li, C. *et al*. Genome Variation Map: a worldwide collection of genome variations across multiple species. *Nucleic Acids Res.* **49**, D1186–D1191, https://doi.org/10.1093/nar/gkaa1005 (2021).
49. Chen, N. *et al*. BGVD: an integrated database for bovine sequencing variations and selective signatures. *Genomics Proteomics Bioinformatics* **18**, 186–193, https://doi.org/10.1016/j.gpb.2019.03.007 (2020).
50. Fu, W. *et al*. RGD v2.0: a major update of the ruminant functional and evolutionary genomics database. *Nucleic Acids Res.* **50**, D1091–D1099, https://doi.org/10.1093/nar/gkab887 (2022).

51. Fu, W. *et al*. GGVD: a goat genome variation database for tracking the dynamic evolutionary process of selective signatures and ancient introgressions. *J. Genet. Genomics* **48**, 248–256, https://doi.org/10.1016/j.jgg.2021.03.003 (2021).
52. Fu, W. *et al*. Galbase: a comprehensive repository for integrating chicken multi-omics data. *BMC Genomics* **23**, 364, https://doi.org/10.1186/s12864-022-08598-2 (2022).
53. Nicholas, F. & Tammen, I. Online Mendelian Inheritance in Animals (OMIA). University of Sydney, https://doi.org/10.25910/2AMR-PV70 (1995).
54. Liu, X. *et al*. GWAS Atlas: an updated knowledgebase integrating more curated associations in plants and animals. *Nucleic Acids Res.* **51**, D969–D976, https://doi.org/10.1093/nar/gkac924 (2023).
55. Gao, Y. *et al*. Animal-SNPAtlas: a comprehensive SNP database for multiple animals. *Nucleic Acids Res.* **51**, D816–D826, https://doi.org/10.1093/nar/gkac954 (2023).
56. Kehr, B. *et al*. Diversity in non-repetitive human sequences not found in the reference genome. *Nat. Genet.* **49**, 588–593, https://doi.org/10.1038/ng.3801 (2017).
57. Cheng, J. *et al*. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492, https://doi.org/10.1126/science.adg7492 (2023).
58. Morales, N. *et al*. Breedbase: a digital ecosystem for modern plant breeding. *G3 (Bethesda)* **12**, jkac078, https://doi.org/10.1093/g3journal/jkac078 (2022).
59. Yarnes, S. C. *et al*. DeltaBreed: a BrAPI-centric breeding data information system. *PLoS ONE* **20**, e0324104, https://doi.org/10.1371/journal.pone.0324104 (2025).
60. Hardigan, M. A. *et al*. Genome synteny has been conserved among the octoploid progenitors of cultivated strawberry over millions of years of evolution. *Front. Plant Sci.* **10**, 1789, https://doi.org/10.3389/fpls.2019.01789 (2019).
61. Richards, S. *et al*. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424, https://doi.org/10.1038/gim.2015.30 (2015).
62. Dinges, J. R., Colleoni, C., Myers, A. M. & James, M. G. Molecular structure of three mutations at the maize *sugary1* locus and their allele-specific phenotypic effects. *Plant Physiol.* **125**, 1406–1418, https://doi.org/10.1104/pp.125.3.1406 (2001).
63. Whitt, S. R., Wilson, L. M., Tenaillon, M. I., Gaut, B. S. & Buckler, E. S. Genetic diversity and selection in the maize starch pathway. *Proc. Natl Acad. Sci. USA* **99**, 12959–12962, https://doi.org/10.1073/pnas.202476999 (2002).
64. Chhabra, R. *et al*. Allelic variation in *sugary1* gene affecting kernel sweetness among diverse-mutant and -wild-type maize inbreds. *Mol. Genet. Genomics* **296**, 1085–1102, https://doi.org/10.1007/s00438-021-01807-9 (2021).
65. Finegan, C. *et al*. Genetic perturbation of the starch biosynthesis in maize endosperm reveals sugar-responsive gene networks. *Front. Plant Sci.* **12**, 800326, https://doi.org/10.3389/fpls.2021.800326 (2021).
66. Tracy, W. F., Whitt, S. R. & Buckler, E. S. Recurrent mutation and genome evolution: example of *Sugary1* and the origin of sweet maize. *Crop Sci.* **46**, S49–S54, https://doi.org/10.2135/cropsci2006-03-0149tpg (2006).
67. Hu, Y. *et al*. Genome assembly and population genomic analysis provide insights into the evolution of modern sweet corn. *Nat. Commun.* **12**, 1227, https://doi.org/10.1038/s41467-021-21380-4 (2021).

Acknowledgements

We thank Peter Van Buren and the HPC team at CSHL for technical assistance, the AgBioData Consortium members and agricultural communities we serve for stimulating discussions, the Breeding Insight team and USDA-ARS breeders for input on breeding community data standardization (MJS), and Melania Ruiz for editorial support (MKT). The AgBioData SGV Working Group and MKT acknowledge support from the National Science Foundation - Research Coordination Network to the AgBioData Consortium [2126334]. Additional support was provided by the USDA Agricultural Research Service [8062-21000-051-000D to DW, 5030-21000-072-000D to CA, 2072-21000-059-000D to NB] and the Elzar High Performance Computing facility at Cold Spring Harbor Laboratory via the National Institutes of Health (NIH S10 OD0286321-01). SBe was supported by the German Federal Ministry of Research, Technology and Space (BMFTR) in the frame of the German Network for Bioinformatics Infrastructure (de.NBI), and ELIXIR, the European infrastructure for life science data. TC was supported by the Wellcome Trust [228142/Z/23/Z]. SD was supported by the European Molecular Biology Laboratory (EMBL). MJS was supported through Breeding Insight (RRID: SCR_026645), a USDA-ARS initiative previously hosted by Cornell University under Cooperative Agreements (8062-21000-043-004-A, 8062-21000-052-002-A, and 8062-21000-052-003-A) and currently hosted at the University of Florida, Gainesville (under 8062-21000-052-020-A).

Author contributions

Conceptualization: M.T., T.C., M.S., S.Be., D.W., S.D., T.C., J.H.; data curation: M.T., M.S., C.A., A.O., R.N.; formal analysis: M.T., S.W., C.K., K.C.; funding acquisition: M.T., C.A., D.W.; investigation: M.T., D.W.; methodology: M.T., M.S., S.Be., S.Ba., D.W., S.W., V.K., A.O., R.S.N., T.S., C.K., K.C.; project administration: D.W.; resources: S.Be., C.A., D.W., S.D., N.G.; supervision: M.T., T.C., D.W.; validation: S.W.; visualization: M.T., S.W.; writing – original draft: M.T., T.C., M.S., S.Be., C.A., D.W., S.D., V.K., R.N., I.C., R.S.N., T.S.; writing – review and editing: M.T., T.C., M.S., S.Be., C.A., S.Ba., D.W., S.W., S.D., C.G.E., T.C., J.H., A.O., J.B., N.B., M.H., I.C., R.S.N., T.S., N.G.

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/s41597-026-07208-0.

Correspondence and requests for materials should be addressed to M.K.T.-R. or D.W.