



AgBioData SGV

# Towards Standards for Biocuration & Interoperability of Genetic Variation Data

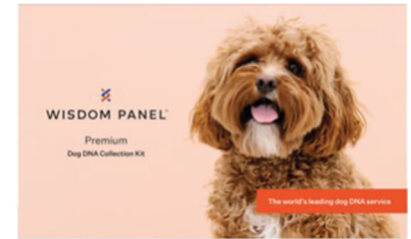
Marcela Karey Tello-Ruiz, PhD  
Cold Spring Harbor Laboratory

**Standards for Genetic Variation Working Group**  
AgBioData Consortium

# Genotyping at our Finger Tips - DIY SNP Kits

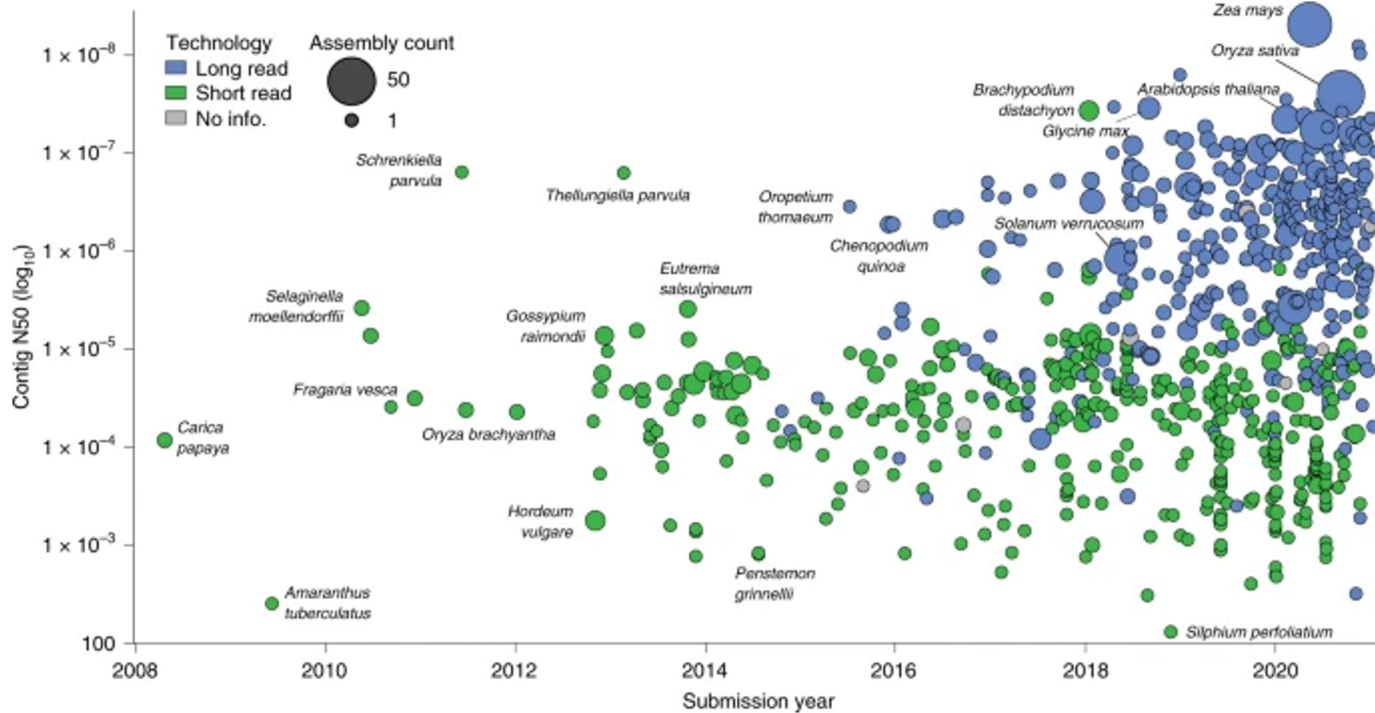


AgBioData SGV





# Increased number & quality of plant genome assemblies



# AgBioData Standards for Genetic Variation WG

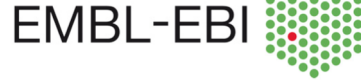


**Chair:** Doreen Ware

**Co-Chair:** Timothee Cezard

**Members:**

- Alexey Sokolov
- Andria Harkey
- Doreen Ware
- Emily Grau
- Kazim Wazir
- Kelly Vining
- Marcela K. Tello-Ruiz
- Mazdak Salavati
- Melanie Harrison



- Nahla Bassil
- Rajdeep S. Khangura
- Sarah Dyer
- Sebastian Beier
- Sharon Wei
- Shaun Clare
- Vivek Kumar
- Yogendra Khedikar



**Past members:**

- Tao-Ho Chang (Rice)

For more information, visit [https://www.agbiodata.org/working\\_groups/sgv](https://www.agbiodata.org/working_groups/sgv)







AgBioData SGV

# AgBioData SGV Working Group Goals

- Support the harmonization and adoption of standards for genetic variation (GV) data from various platforms in Plants & Animals
- Bring together a community of data providers, biocurators & computer scientists to promote interoperability and access to GV datasets

[https://www.agbiodata.org/working\\_groups/sgv](https://www.agbiodata.org/working_groups/sgv)



# Standards for Genetic Variation Working Group

- **Specific objectives:**

- Enable sharing of GV data to support agriculture
- Identify existing GV and technical barriers for data exchange
- Review technical standards for GV to support adoption
- Review GV workflows
- Engage community to support ingestion and usability of GV data into community and archival resources

- **Activities:**

- Regular monthly meetings (break July-September)
- Biocurators & smaller group meetings
- AgBioData annual workshop
- Community surveys
- Webinar “Biocurating Genetic Variation” (8 speakers)



AgBioData SGV

# AgBio Community Surveys

## Goals:

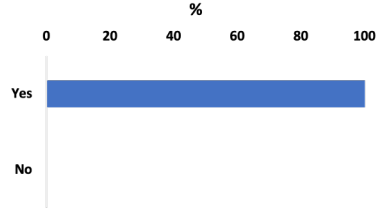
- Identify existing & anticipated GV data sets for agriculturally important species
- Identify challenges & propose solutions for data integration & interoperability
- Recruit WG members

# Live Poll - Feb 2022 (15 participants)

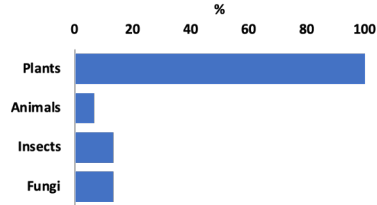


AgBioData SGV

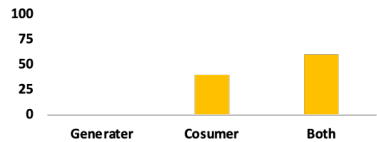
- Are you aware of/or working on genetic variation?



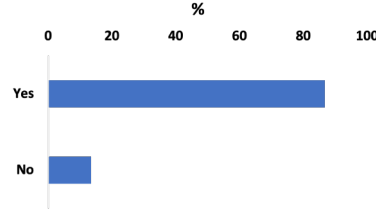
- What type of species are you working on?



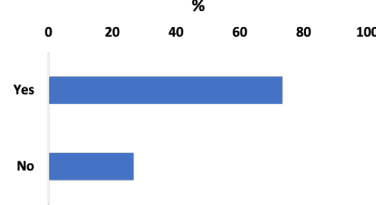
- Are you a data generator, consumer or both?



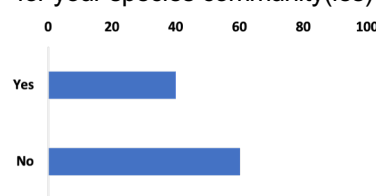
- Is there a community resource to host your GV data?



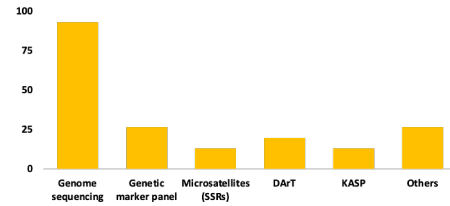
- Have you heard of the EVA (European Variation Archive)?



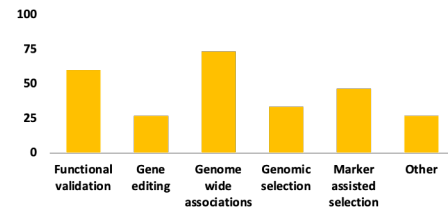
- Are there standards to name samples (i.e., standard identifiers) for your species community(ies)?



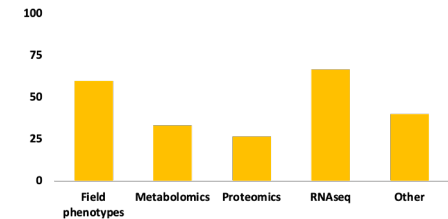
- If generating the data what type of technology are you using?



- What are you using the genetic variation information for?



- What other data types are you generating from the same germplasms/biosamples?

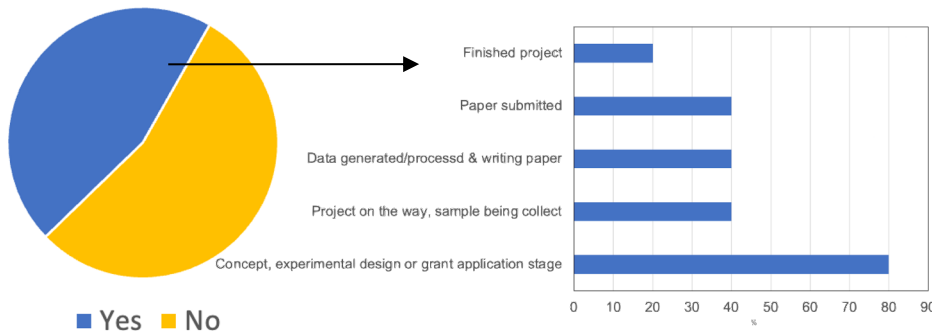


# Survey - Feb. 2022 (11 participants)

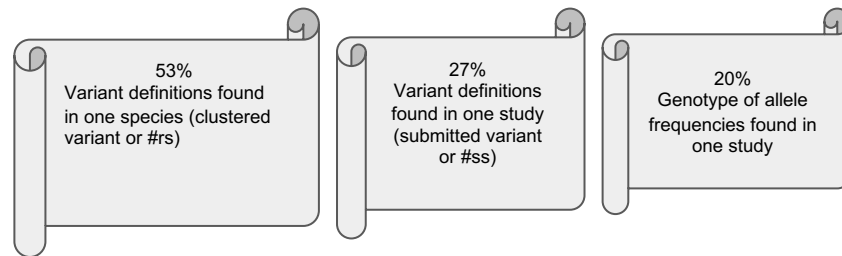


AgBioData SGV

## Generate or process variation data that could be submitted to EVA



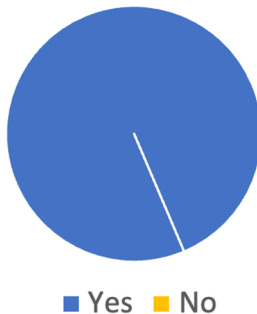
## Interest in specific data types



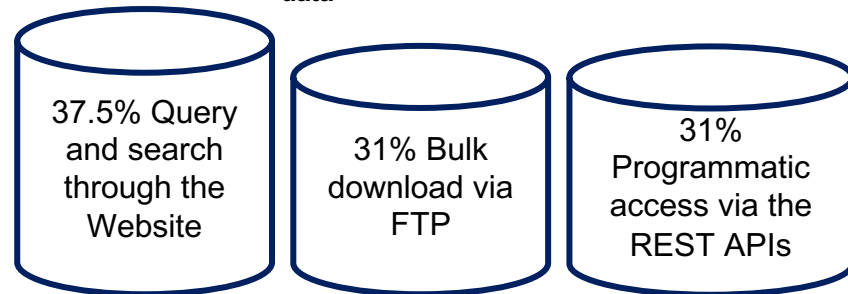
## Interest in a Biocuration Workshop to support processing and submitting variation data?



## Interest in using data already served by EVA?



## Preference to access data

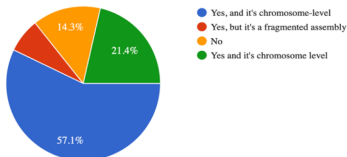




# Preliminary Survey - Jan. 2023 (14 responses)

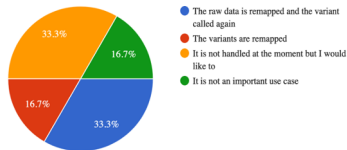
3. Does your species of interest usually have a reference assembly?

14 responses



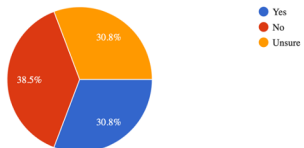
13. If the reference genome changes, how do you handle the update?

12 responses



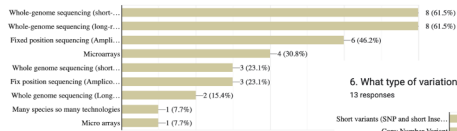
14. Are there stable variant identifiers associated with the variation data you hold?

13 responses



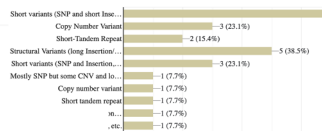
5. What type of technologies are used to generate the variation data?

13 responses



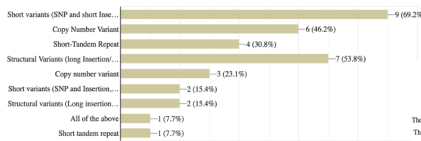
6. What type of variation data are you storing?

13 responses



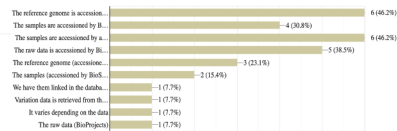
7. What type of variation data are you interested in storing in the future?

13 responses



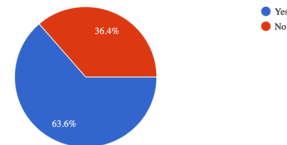
10. Is the variation data linked to other accessioned data?

13 responses



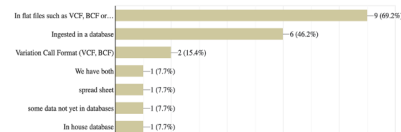
15. Are you interested in participating in the SGV working group?

11 responses



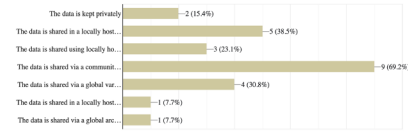
8. How do you keep your variation data?

13 responses



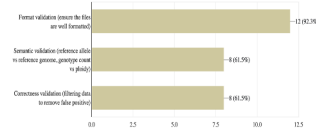
9. How do you share or plan to share your variation data?

13 responses



12. What kind of validation and quality control is performed on the variation data?

13 responses





# Biocurating Ag Genetic Variation

1. [GDR](#) (CottonGen, GDV, CGD, PCD) - Sook Jung
2. [BreedBase](#) (SGN, CassavaBase, YamBase, SweetPotatoBase, MusaBase) - Lukas Mueller
3. [MaizeGDB](#) - Carson Andorf by proxy
4. [NCGR Corvallis](#) - Nahla Bassil
5. [TreeGenes](#) - Emily Grau
6. [TAIR](#) - Tanya Berardini/Leonore Reiser
7. [InterMine](#) (MaizeMine, Bovine Genome Database, FAANGMine, Hymenoptera Genome DB) - Chris Elsik by proxy
8. [Gramene / Ensembl Plants](#) & [SorghumBase](#) - Marcela K. Tello-Ruiz



# Biocurating Ag Genetic Variation



AgBioData SGV



### Search Genotype

SNP Genotype | SSR Genotype

This page provides easy access to the datasets that are mentioned to be available from GDR in publications. Data are integrated in GDR and there are various ways to search for SNP genotype. To search for SNP genotype data only for cultivars and to click the question mark next to 'Dataset' to view the details of the dataset. [Text to](#)

Search SNP Genotype is a page where users can search for the SNP genotype data search for SSR Genotype.

Dataset: Any

Species: Any

Genom: Any

Chr/Scaffold: Any - between and bp

Gene Model: +/- bp

Search | Reset

### Publication datasets

This page provides easy access to the datasets that are mentioned to be available from GDR in publications. Data are integrated in GDR and there are various ways to search for SNP genotype. To search for SNP genotype data only for cultivars and to click the question mark next to 'Dataset' to view the details of the dataset. [Text to](#)

Links to publication pages where data can be accessed and additional links to search/Browse pages to access other data available below.

GRIN accession number	Publication	Access Data
HGDR1192	Reinert CM, Flores M, Casan V, de Silva-Linger C, Hottelinger M, Byles T, Rowlandson Z and Pascoe C (2003) Multi-environment genomic prediction for candidate alleles derived by wheat (Triticum aestivum, Linn., Plant. Physiol. 133:804-814)	genotypic data
HGDR1191	Alvarado JM, delacort E, Lator J, Davel CE, Durand C, Morano R, Ordoñez M, Albert DG (2003) Pathogen reconstruction for tropical apple cultivars using single nucleotide polymorphism array data. Plants, People, Planet. DOI	genotypic data (to be made available in a public repository)
HGDR1190	Fan Z, Tamara DM, Hoppa SJ, Zeller P, Farina R, Ballew CR, Fatta MM, Amadio RL, Liu M, Qin Y, Lamb S, Williams MA. A multi-ethnic research network: identifying flavor genes and their regulatory elements. The New phytologist 2003 Aug 05	pub (not), data available from pub (not) genome page
HGDR1189	Hein Oudek, Iva Cordeiro, Sebastião Franco Oliveira, Karlastrava S. Alvarez, Mônica A. Rocha-Castelo, Nóbilio A. Almeida-Costa, Marlene Werneck, Anne D'Angelo, Anna Sarmento, Peter A. Dowling and evaluation of an Arabidopsis SNP chip array for sorghum. Submitted to TSG	SNP array data
HGDR1188	Costain Serna, PhD María de los Angeles Bagán Quirós, José Mateo, Patricia Morales, Miguel Gallo de los Rios, Fernando Sánchez. Genome-wide colour variability in European pear (Pyrus) using high-throughput sequencing. Horticulture Research. DOI:10.1186/s13007-015-0050-111	SNP data (to be available in Market Search)
HGDR1187	Karlsson et al. Identification of novel genetic regions associated with resistance to European canker in apple. Submitted to BMC Plant Biology	SNP genotype data Haplotype data Phenotypic data QTL data

### MaizeMine v1.5

An integrated data warehouse for **MaizeGDB**. Previous releases: **MaizeMine v1.4**, **MaizeMine v1.3**.

Home | MyMine | Templates | Lists | QueryBuilder | Regions | Data Sources | Help | API

Contact Us | Log In

Search: e.g., Zm00001eb000020 GO

#### Quick Search

Search MaizeMine. Enter names, identifiers or keywords for genes, proteins, pathways, ontology terms, authors, etc. (e.g., Zm00001eb000020, 100037783, GR2b, Zm00001eb000020\_T002, Zm00001eb000020\_P002, NM\_00111367.2, Zm00001d023210, GRMZM2C109674, 708A6\_MAIZE, PZE010000203).

e.g., Zm00001eb000020

SEARCH

#### Quick List

Enter a list of identifiers.

#### Gene

e.g., Zm00001eb000110, Zm00001eb000120, Zm00001eb000130, Zm00001eb000140, Zm00001eb000150, Zm00001eb000160, Zm00001eb000170

advanced ANALYZE

#### About v1.5 and Templates

Variant Legend:

- Missense variant
- Synonymous variant
- 3 prime UTR variant
- Upstream gene variant
- Intergenic variant
- Splice acceptor variant
- Splice region variant
- 5 prime UTR variant
- Intron variant
- Downstream gene variant

### BREEDBASE

Search | Manage | Analyze | About

Notice: This is a demo site only. If you would like to use Breedbase, please contact lam87@corn



Accession	Title	Tags	Species	Plant Count	Phenotypes Assessed	Phenotypic Measures	Genotype Count	ATON	ENTIRE GENE SET	ALIAS AND DEXREF	COMMUNITY
TGDR001	Association genetics of traits controlling lignin and cellulose biosynthesis in black cottonwood ( <i>Populus trichocarpa</i> , Salicaceae) secondary xylem.	TPPSc Phenotype Genotype	<i>Populus trichocarpa</i>	448	3	1335	391552				
TGDR002	Genetic Variation in <i>Quercus acutissima</i> Carruth., in Traditional Japanese Rural Forests and Agricultural Landscapes, Revealed by Chloroplast Microsatellite Markers	TPPSc Approximate Coordinates Genotype	<i>Quercus acutissima</i>	2152			12912				

### CHADO

### CartograPlant

### maizeGDB

Site | Allelic Chr Position | Gene (T01) | Type

10045	G/C/C	1	10045	GRMZM2C0888220	exon
10097	C/G	1	10097	GRMZM2C0888220	exon
10108	C	1	10108	GRMZM2C0928526	exon
10218	C	1	10218	GRMZM2C0888220	exon
10280	G	1	10280	GRMZM2C0928526	exon
10285	T	1	10285	GRMZM2C0888220	exon
10295	T	1	10295	GRMZM2C0928526	exon
10590	T/A	1	10590	IGR	IGR
10638	G/T	1	10638	IGR	IGR
83374	A/G	1	83374	View	IGR
83379	A	1	83379	View	IGR
109078	A	1	109078	View	IGR
111650	T	1	111650	GRMZM2C0931344	exon
111651	A	1	111651	GRMZM2C0931344	exon
111666	G	1	111666	GRMZM2C0931344	exon
111683	C	1	111683	GRMZM2C0931344	exon
111696	A	1	111696	GRMZM2C0931344	exon
111745	G	1	111745	GRMZM2C0931344	exon
111758	C	1	111758	GRMZM2C0931344	exon
128373	G/T	1	128373	View	IGR

# Challenges associated with genetic variation



AgBioData SGV

- All data has a lifecycle. It can become stale & could be reused
- Different versions of an assembly (quality & stability)
  - Remapping to a newer assembly may result in reduced precision & data loss
  - Raw data vs processed data
  - Availability & quality of data sets for clustering
- Moving from a single reference to a PanGenome
- Improvements in assays and algorithms to determine GV (GBS, WGS, etc.)
- Converting from SSRs to SNPs
- Integration between studies (new studies, meta-analyses, etc.)
  - Sample identifiers

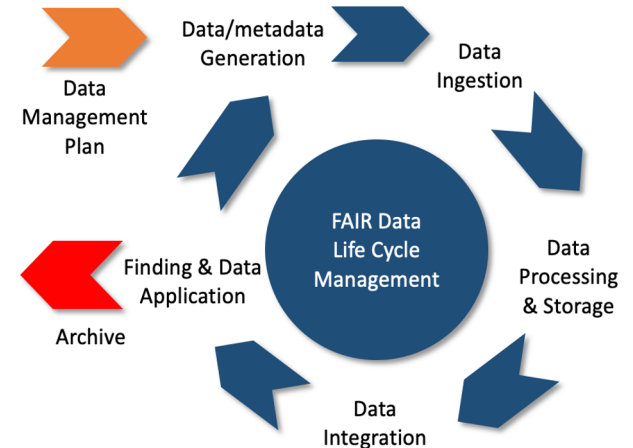


Image credit: FAIRToolkit



# Central repositories for genetic variation

## EVA issues long-term IDs for non-human variants

Since Sept.  
2017



The European Variation Archive: freely available data on genetic variation

### Summary

- New agreement between the NCBI and EMBL-EBI shares responsibility for managing data from genetic variation experiments worldwide.
- From September 2017, EMBL-EBI's European Variation Archive (EVA) will issue locus accession numbers (Reference SNP, rs#) for all non-human

## The European Variation Archive: a FAIR resource of genomic variation for all species

Timothe Cezard, Fiona Cunningham, Sarah E Hunt, Baron Koylass, Nitin Kumar, Gary Saunders, April Shen, Andres F Silva, Kirill Tsukanov, Sundararaman Venkataraman ... [Show more](#)

*Nucleic Acids Research*, Volume 50, Issue D1, 7 January 2022, Pages D1216–D1220,

<https://doi.org/10.1093/nar/gkab960>

**Published:** 28 October 2021 **Article history** ▾

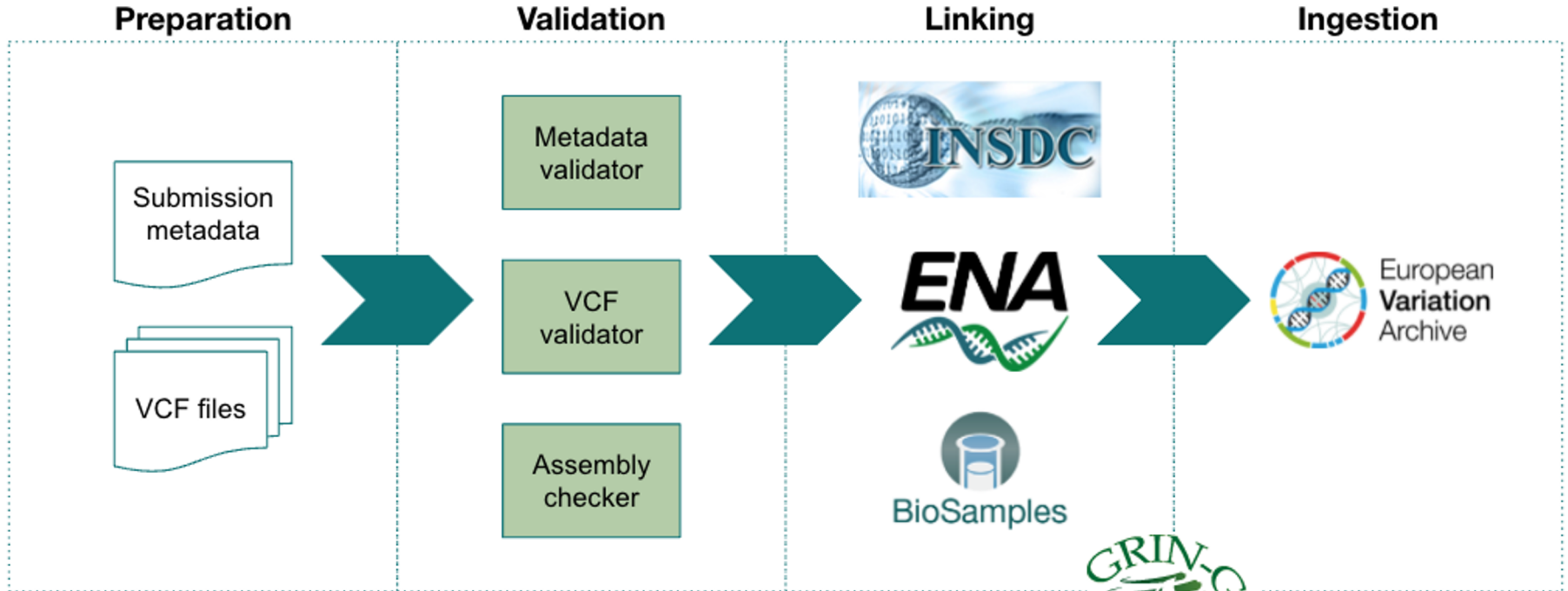
 PDF  Split View  Cite  Permissions  Share ▾

### Abstract

The European Variation Archive (EVA; <https://www.ebi.ac.uk/eva/>) is a resource for sharing all types of genetic variation data (SNPs, indels, and structural variants) for all species. The EVA was created in 2014 to provide FAIR access to genetic variation data and has since grown to be a primary resource for genomic variants hosting >3 billion records. The EVA and dbSNP have established a compatible global system to assign unique identifiers to all submitted genetic variants. The EVA is active within the Global Alliance of Genomics and Health (GA4GH), maintaining, contributing and implementing standards such as VCF, Refget and Variant Representation Specification (VRS). In this article, we describe the submission and permanent accessioning services along with the different ways the data can be retrieved by the scientific community.



# Submission process through EVA



Slide courtesy of EVA





# Genetic variation data - Standard file format

## Variant Call Format (VCF)

```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
```

```
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
```

Meta-info lines

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002	NA00003
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0 0:48:1:51,51	1 0:48:8:51,51	1/1:43:5:.,.
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0 0:49:3:58,50	0 1:3:5:65,3	0/0:41:3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1 2:21:6:23,27	2 1:2:0:18,2	2/2:35:4
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0 0:54:7:56,60	0 0:48:4:51,51	0/0:61:2
20	1234567	microsat1	GTC	G,GTCT	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3

Header lines + Sample IDs  
Data lines + Genotypes



# Genetic variation metadata standards

## EVA metadata submission template

*V1.1.4 August 2020*

The aim of this sheet is to facilitate effective completion of this template.

The minimum information required to be completed in this template in order for data to be submitted to EVA is: Submitter details, project information, at least data on 1 sample, analysis details, and file entries. However, we encourage our users to submit as much meta-data as possible. Increased metadata creates much greater visibility of your data and research in our search and analysis platforms. Additionally, such information allows for e

Please email all questions and feedback to [eva-helpdesk@ebi.ac.uk](mailto:eva-helpdesk@ebi.ac.uk)

This template is grouped into four sections, split into worksheets. Each worksheet is preceded by an "HELP" sheet which provides more information and instructions for each column.

Worksheet	Explanation
Submitter Details	This sheet captures the credentials of the submitter.
Project	The objective of this sheet is to gather general information about the Project including submitter, submitting centre, collaborators and publications.
Sample	Projects consist of analyses that are run on samples. We accept sample information in the form of BioSample, ENA or EGA accession(s). We also accept BioSamples sampleset accessions. If your samples a sample(s)" sections of the Sample(s) worksheet to have them registered at BioSamples.
Analysis	For EVA, each analysis is one vcf file, plus an unlimited number of ancillary files. This sheet allows EVA to link vcf files to a project and to other EVA analyses. Additionally, this worksheet contains experimen Important to note; one project can have multiple associated analyses.
Files	Filenames and associated checking data associated with this EVA submission should be entered into this worksheet. Each file should be linked to one, or more, analysis. We accept VCF files along with thei

Each worksheet contains a number of fields -

Completion of the remaining highlighted in **BOLD** is **REQUIRED**. **GREEN** indicates **EITHER/OR** requirement.

Completion of the remaining fields is optional, however please provide as much information as you can and avoid the use of non-ASCII characters in any fields.

An example of a completed template suitable for EVA submission is available at our website ([www.ebi.ac.uk/eva/](http://www.ebi.ac.uk/eva/))



# EU-FONDUE recommendations data standards for plants



AgBioData SGV

- FONDUE: FAIR-ification of Plant Genotyping Data and its linking to Phenotyping using ELIXIR Platforms
- First guidelines on FAIR handling of GV data published in 2022
- Support data submission to BioSamples & EVA by providing a checklist to classify and validate the data

F1000Research

Search

REVISED

Recommendations for the formatting of Variant Call Format (VCF) files to make plant genotyping data FAIR [version 2; peer review: 2 approved]



✉ Sebastian Beier <sup>1,2</sup>, Anne Fiebig <sup>1</sup>, Cyril Pommier <sup>3</sup>, Isuru Liyanage <sup>4</sup>, Matthias Lange <sup>1</sup>, Paul J. Kersey<sup>5</sup>, Stephan Weise <sup>1</sup>, Richard Finkers <sup>6,7</sup>, Baron Koylass <sup>4</sup>, Timothee Cezard <sup>4</sup>, Mélanie Courtot <sup>4,8</sup>, Bruno Contreras-Moreira <sup>9</sup>, Guy Naamati<sup>4</sup>, Sarah Dyer<sup>4</sup>, Uwe Scholz <sup>1</sup>



# Summary of recommendations for plant metadata formatting

Table 1. Summary of recommendations for metadata formatting.



AgBioData SGV



BioSamples

Metadata field	Definition	Format	Example	Cardinality
##fileDate	Creation date of the VCF file	Date (ISO 8601, YYYYMMDD)	##fileDate=20120921	1
##bioinformatics_source	Chains of bioinformatics tools for creating the VCF file	URL, DOI	##bioinformatics_source="doi.org/10.1038/s41588-018-0266-x"	1
##reference_ac	Accession number of reference genome assembly used in the VCF file	/[([GCA/GCF_(d)(9)\.(0-9)*]/	##reference_ac=GCA_902498975.1	1
##reference_url	URL of the reference genome assembly used in the VCF file	URL, DOI	##reference_url="ftp.ncbi.nlm.nih.gov/genomes/all/GCA/902/498/975/GCA_902498975.1_Musca domestica L. chr11"	1

##SAMPLE	Metadata about a single sample genotype that is part of the genotyping experiment in the VCF file	Composite (see below)	##SAMPLE=<ID=SAMEA104646767,DOI="doi.org/10.25642/IPK/GBIS/7811152">	1:N
	The primary identifier (BioSamples Database identifier) of the genotyping sample	/[([SAM)(E N D)(A G)(\d+)]/	ID=SAMEA104646767	1
	The DOI of the genotyping sample (if available)	URL, DOI	DOI="doi.org/10.25642/IPK/GBIS/7811152"	0-1
	The external identifiers under which this genotyping sample is registered in other databases (either 'FAO-WIEWS_instcode:genus:accession_number' or 'DNS:database_identifier:identifier_scheme:identifier')	See Definition	ext_ID="DEU146:Hordeum:HOR 1361 BRG" or ext_ID="ipk-gatersleben.de:GBIS:akzessionId:7811152"	0:N

##SAMPLE	Metadata about a single sample genotype that is part of the genotyping experiment in the VCF file	Composite (see below)	##SAMPLE=<ID=SAMEA104646767,DOI="doi.org/10.25642/IPK/GBIS/7811152">	1:N
	The primary identifier (BioSamples Database identifier) of the genotyping sample	/[([SAM)(E N D)(A G)(\d+)]/	ID=SAMEA104646767	1
	The DOI of the genotyping sample (if available)	URL, DOI	DOI="doi.org/10.25642/IPK/GBIS/7811152"	0-1
	The external identifiers under which this genotyping sample is registered in other databases (either 'FAO-WIEWS_instcode:genus:accession_number' or 'DNS:database_identifier:identifier_scheme:identifier')	See Definition	ext_ID="DEU146:Hordeum:HOR 1361 BRG" or ext_ID="ipk-gatersleben.de:GBIS:akzessionId:7811152"	0:N





# Suggestions for plant samples meta by our WG

- 1) Mandatory (1:N): Primary external identifier from major germplasm repository (e.g., GRIN, CGIAR, IPK, CNGB)
- 2) Recommended (0:N): Inventory or local number
- 3) Recommended (0:N): Identifier for the specific plant/genotype used in the study

Biocurators meetings

Metadata field	Field Name	Definition	Format	Example	Cardinality
#SAMPLE		Metadata about a single sample genotype that is part of the genotyping experiment in the VCF file	Composite (see below)	##SAMPLE=<ID=SAMN04168247, DOI=doi.org/10.18730/NBYG*, ext_ID=grin-global.org:USA126-PI 276837>	1:N
	BioSample ID	Refers to a biological sample used as a 'reference' (e.g. to sequence its genome) or used in an assay database such as ENA, EVA, ArrayExpress. Always begin with SAM. The next letter is either E or N or D depending if the sample information was originally submitted to EMBL-EBI or NCBI or DDBJ, respectively. After that, there may be an A or G to denote an Assay sample or a Group of samples. Finally, there is a numeric component that may or may not be zero-padded.	{(SAM E N D)(A G 0+)*}	ID=<SAMN04168247	1
	External identifier	- Primary accession - One mandatory external ID for plants. Impractical to enter metadata for each biosample; easier to add as a metadata line in VCF. Impractical for huge data sets as this would significantly increase the size of the VCF file. - Source of accession [Genbank Name, Original Collection (not in genbank), etc.] Examples: GRIN, ICRISAT, WEIWS code:Species code (IPK), CNGB, GBIS, ORIGINAL COLLECTION - Accession prefix. Examples: PI, IS, NSSL, GRIF, SOR, Collector ID - Accession unique identifier or number. Example: six-digit PI number, five-digit IS number, four-digit following WEIWS species number, collector number - Secondary accession - Sample inventory if applicable. Example: CR02, CR03, 07PL. Note: USDA germplasm repositories provide inventory accessions. - Other - Not necessary. Example: Population panel identifiers such as SAP_301, a member of the Sorghum Association Panel are not necessary and are well captured in germplasm registries like GRIN.	ext_ID=registry.identifier	ext_ID=grin-global.org:USA126-PI 276837	1:N
	Study sample identifier	Identifiers under which this genotyping sample is registered in other databases (either FAO-WIEWS_justcode:genus:accession_number* or DONS(database_identifier:identifier,scheme:identifier))			0-1
	DOI/URL	Identifies specific plant/genotype used, when available. This will usually be specific to an individual research project and not publicly available. However, the plant or DNA sample may be shared between researchers. Different plant numbers from the same lot. Example: SC103 and SC105-146 share the same P153755 accession.	URL_DOI	DOI=doi.org/10.18730/NBYG*	0-1
	DOI/URL	DOI for the passport information of the genotyping sample.	URL_DOI	DOI=doi.org/10.18730/NBYG*	0-1



AgBioData SGV

# FAANG guidelines for data submission



## Understanding the Genome to Phenome link in domesticated animals

The adoption and dissemination of metadata standards for animal genetic variants is relatively advanced. The FAANG online portal can manage metadata in the form of rule sets and provide tools for central validation, and links to public repositories (ENA, EVA).



AgBioData SGV

# Different databases are serving different purpose

## Central databases

- Long-term archiving of original files
- Accessioning
  - Study
  - Samples
  - Variants
- Update to newer genomes



**ENA**  
European Nucleotide Archive



## Community/species databases

- Integration between phenotypes and genotypes
- Tailored feature/toolsets





# Pilot projects based on readiness of the communities

1. **Species communities with high-quality reference assemblies in INSDC and GV data in other DBs** (e.g., MaizeGDB, GDR, SolGenomics, CassavaBase, TreeGenes)
  - Support interoperability with community resources
  - Demonstrate added value in an archival resource
  - Triage use cases
2. **Species with high-quality reference assemblies and population variation data sets without resources to host large GV data sets** (i.e., germplasm centers, GRIN)
  - Support for species where infrastructure is not available, capacity building
  - Promote submission of reference assemblies to INSDC
  - FAIR access to data
3. **Species developing GV data sets**
  - Standards for community & capacity building
  - FAIR access to data



# Summary of Outcomes

- **Surveys** - Identified existing GV datasets, workflows and technical barriers for data exchange
- **VCF & metadata for samples** - Reviewed guidelines & made additional recommendations to support adoption



## Next Steps

**Pilot projects** - Engaging community to support ingestion and usability of GV data into community & archival resources

- Recruiting WG members
- Recruiting new communities for pilot projects
- Lowering the barrier for generating metadata
- Training materials and virtual hackathons



# AgBioData survey

*AgBioData Standards for  
Genetic Variation*



AgBioData SGV

## Got SNPs?



*Scan QR Code to take  
our Survey*

[https://www.agbiodata.org/working\\_groups/sgv](https://www.agbiodata.org/working_groups/sgv)



AgBioData SGV

# Join our working group, take our survey, meet with us!



**Chair:** Doreen Ware

**Co-Chair:** Timothee Cezard

**Members:**

- Alexey Sokolov
- Andria Harkey
- Doreen Ware
- Emily Grau
- Kazim Wazir
- Kelly Vining
- Marcela K. Tello-Ruiz
- Mazdak Salavati
- Melanie Harrison



- Nahla Bassil
- Rajdeep S. Khangura
- Sarah Dyer
- Sebastian Beier
- Sharon Wei
- Shaun Clare
- Vivek Kumar



**Past members:**

- Tao-Ho Chang (Rice)



**AgBioData Booth #230**

Sunday Opening, Monday lunch, Tuesday PM

PAG30

