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
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
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
Standards for Genetic Variation Working Group

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

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

- **Are you aware of/or working on genetic variation?**
- **What type of species are you working on?**
  - Plants
  - Animals
  - Insects
  - Fungi
- **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)?**
- **If generating the data what type of technology are you using?**
  - Genome sequencing
  - Genetic marker panel
  - Microarrays
  - DArT
  - KASP
  - Others
- **What are you using the genetic variation information for?**
  - Functional validation
  - Gene editing
  - Genome-wide associations
  - Genomic selection
  - Marker assisted selection
  - Other
- **Are there standards to name samples (i.e., standard identifiers) for your species community(ies)?**
- **What other data types are you generating from the same germplasms/biosamples?**
  - Field phenotypes
  - Metabolomics
  - Proteomics
  - RNAseq
  - Other
**Survey - Feb. 2022 (11 participants)**

Generate or process variation data that could be submitted to EVA

- **Yes**
- **No**

**Interest in specific data types**

- 53% Variant definitions found in one species (clustered variant or #rs)
- 27% Variant definitions found in one study (submitted variant or #ss)
- 20% Genotype of allele frequencies found in one study

**Preference to access data**

- 37.5% Query and search through the Website
- 31% Bulk download via FTP
- 31% Programmatic access via the REST APIs
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
Challenges associated with genetic variation

- 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
Central repositories for genetic variation

EVA issues long-term IDs for non-human variants

Since Sept. 2017

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

Published: 28 October 2021

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

- **Preparation**
  - Submission metadata
  - VCF files

- **Validation**
  - Metadata validator
  - VCF validator
  - Assembly checker

- **Linking**
  - INSDC
  - ENA
  - BioSamples

- **Ingestion**
  - European Variation Archive
Genetic variation data - Standard file format

Variant Call Format (VCF)

Header lines + Sample IDs
Data lines + Genotypes

Meta-info lines
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 easier management of datasets.

Please email all questions and feedback to eva-helpdesk@ebi.ac.uk

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

<table>
<thead>
<tr>
<th>Worksheet</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Submitter Details</td>
<td>This sheet captures the credentials of the submitter.</td>
</tr>
<tr>
<td>Project</td>
<td>The objective of this sheet is to gather general information about the Project including submitter, submitting centre, collaborators and publications.</td>
</tr>
<tr>
<td>Sample</td>
<td>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 are from a sample(s)” sections of the Sample(s) worksheet to have them registered at BioSamples.</td>
</tr>
<tr>
<td>Analysis</td>
<td>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 experimental details important to note; one project can have multiple associated analyses.</td>
</tr>
<tr>
<td>Files</td>
<td>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 their relevant files.</td>
</tr>
</tbody>
</table>

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/)
EU-FONDUCE recommendations data standards for plants

- FONDUCE: 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

doi: 10.12688/f1000research.109080.2
Summary of recommendations for plant metadata formatting

| 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/s41588-018-0266x" | 1           |
| reference_ac   | Accession number of reference genome assembly used in the VCF file        | ((GA|GC)(A|G)(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/902498975/ GCA_902498975.1 | 1           |

Table 1. Summary of recommendations for metadata formatting.

---

**The primary identifier (BioSamples Database identifier) of the genotyping sample**

```
/([S][M][A][E]|[E]|N|D)[A]|G\{d+\})/ID=SAMEA104646767
```

1:N

**The DOI of the genotyping sample (if available)**

```
URL, DOI
```

DOIs=doi.org/10.25642/IPK/GBIS/7811152

0-1

**The external identifiers under which this genotyping sample is registered in other databases (either FAO-VIEWS, Instdcode: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
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).
Different databases are serving different purpose

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

Community/species databases
- Integration between phenotypes and genotypes
- Tailored feature/toolsets

Slide courtesy of EVA
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

Scan QR Code to take our Survey

https://www.agbiodata.org/working_groups/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