Standardizing Biocuration of Genetic Variation Data to Promote FAIRification

Standards for Genetic Variation Working Group
AgBioData Consortium

Marcela Karey Tello-Ruiz, PhD
Cold Spring Harbor Laboratory
Outline

1. The SGV Working Group
2. Standards for Genetic Variation & Interoperability
3. Data Submission to the European Variation Archive
4. Challenges
5. Progress towards FAIRifying plant data sets
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 (engagement with Education & Sci. Literature WGs)
  o Participation at AgBioData annual workshop & PAG workshop
  o Data biocuration & coordination across participant resources
  o Promoting FAIRification of GV data & recruiting members at relevant events
  o Merging with Public Genetics Resources WG
  o Reporting to funders
AgBioData Standards for Genetic Variation WG

Co-Chairs:
Marcela K. Tello-Ruiz
Timothee Cezard

Most active members:
- Nahla Bassil
- Sebastian Beier
- Irene Cobo
- Sarah Dyer
- Osman Gutierrez
- Melanie Harrison
- Jodi Humann
- Rex Nelson
- Mazdak Salavati
- Moira Sheen
- Doreen Ware
- Sharon Wei

Full list at https://www.agbiodata.org/working_groups/sgv
Different databases are serving different purposes

Archival DBs

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

Community/species DBs

- Integration between genotypes and phenotypes
- Tailored feature/toolsets

INTEROPERABILITY
Lessons from human genetics

- **Trait & disease associations**
- **Genetic variants**
- **Biosamples**
- **Functional annotations**

Image taken from doi: 10.1161/ATVBAHA.120.315300
Standards for Genetic Variation – Interoperability

**Biosample IDs**
- BioSample ID (EVA requirement)
- Germplasm ID (genebanks
  - ICRIAT: IS 12661, GRIN: PI

**rsIDs**
- Reference cluster ID
- Stable/unique for a genetic locus.
- EVA provides ‘ss’ (submission) and ‘rs’ (ref) ids for non-human variants

**VCF**
- Variant Call Format
- Text file format with meta-info and data for a variant position in a genome sequence assembly at INSDC

**Traits**
- Controlled vocabularies for GWAS, QTLs, etc.

Phenos image modified from: 10.3390/agronomy12040838
Data submission to the European Variation Archive

Preparation
- Submission metadata
- VCF files

Validation
- Metadata validator
- VCF validator
- Assembly checker

Linking
- INSDC
- ENA
- European Variation Archive

Ingestion
- European Variation Archive

Slide courtesy of EVA
Major challenges associated with genetic variation

- All data could be reused
- Remapping to a newer assembly may result in reduced precision & data loss
- 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.)

=> Solution: Submit GV to EVA to get rsIDs, unique genetic variant identifiers
Challenges associated with biosamples

<table>
<thead>
<tr>
<th>Study 1</th>
<th>Study 2</th>
<th>Study 3</th>
<th>Other names</th>
</tr>
</thead>
<tbody>
<tr>
<td>PI534138</td>
<td>SC62C</td>
<td>SAP_PI534138.B064FABXX.2.F7</td>
<td>Matchikah SAP-416</td>
</tr>
<tr>
<td>PI534138</td>
<td>SC62C</td>
<td>SAP_PI534138.B064FABXX.2.G1</td>
<td></td>
</tr>
</tbody>
</table>

DB sample with multiple ids because of naming conventions

<table>
<thead>
<tr>
<th>Study 1 at NCBI</th>
<th>Study 2 at DB2</th>
<th>Study 1 at DB1</th>
</tr>
</thead>
<tbody>
<tr>
<td>EarlyHegari</td>
<td>Early_Hegari</td>
<td></td>
</tr>
<tr>
<td>IBC/E-38432</td>
<td>38432</td>
<td>IBC_E38432</td>
</tr>
<tr>
<td>Karper 669</td>
<td>Karper_669</td>
<td></td>
</tr>
<tr>
<td>Malisor 84-7</td>
<td>Malisor84-7</td>
<td></td>
</tr>
<tr>
<td>RTx7000</td>
<td>RTx7000</td>
<td></td>
</tr>
<tr>
<td><em>S. bicolor (PI226096)</em></td>
<td>S.bicolor.subsp.Verticilliflorum(PI226096)</td>
<td>PI226096</td>
</tr>
<tr>
<td><em>S. bicolor subsp. drummondii (PI330272)</em></td>
<td>S.bicolor.subsp.drummondii</td>
<td>PI330272</td>
</tr>
<tr>
<td><em>S. bicolor subsp. verticilliflorum (AusTRCF 317961)</em></td>
<td>S.arundinaceum</td>
<td>AusTRCF_317961</td>
</tr>
<tr>
<td><em>S. bicolor subsp. verticilliflorum (PI300119)</em></td>
<td>S.bicolor.subsp.Verticilliflorum(PI300119)</td>
<td>PI300119</td>
</tr>
<tr>
<td>Cherekit (IBC/E-460)</td>
<td>Cherekit(S)</td>
<td>Cherekit_ICB_E460</td>
</tr>
<tr>
<td>Kilo (IBC/E-382)</td>
<td>Kilo</td>
<td>Kilo_ICB_E382</td>
</tr>
<tr>
<td>Yik.solate (IBC/E-339)</td>
<td>Yik.solate</td>
<td>Yik_ICB_E339</td>
</tr>
<tr>
<td>Zengada (IBC/E-308)</td>
<td>Zangeda</td>
<td>Zengada_ICB_E308</td>
</tr>
</tbody>
</table>

=> Solution: Use standard germplasm identifiers (BioSample / Genebank IDs)
Recommendations for data standards for plants

- FAIRification of Plant Genotyping Data (& linking it to Phenotyping)
- First guidelines on FAIR handling of GV data published in 2022
- Provide a checklist to classify and validate the data to support its submission to EVA (and BioSamples)

doi: 10.12688/f1000research.109080.2
Recommendations from Biocurator Meetings

Additional Suggestions for Plant Samples Metadata associated with VCFs

- **BioSamples entries:**
  - Require primary external identifier from major germplasm repository (e.g., GRIN, CGIAR, IPK, CNGB) with doi/url
  - Recommend including inventory or local number & identifier for the specific plant/genotype used in the study
Technical challenges revealed through biocuration

- Missing reference genome assembly
- Reference genome not registered at INSDC
- GV data not readily available (e.g., private FTP)
- GV data not in standardized format (e.g., VCF)
  - Non-standard format at community DB (e.g., tabular output .xls)
  - No format conversion method provided
  - Only precursor sequencing reads provided
### FAIRifying public plant GV data sets

<table>
<thead>
<tr>
<th>Species</th>
<th>Reference assembly in INSDC</th>
<th>VCF available</th>
<th>Sample IDs with DOI/URL from major germplasm repo</th>
<th>VCF in EVA &amp; BioSamples</th>
<th>Samples qualified for cross-linking to other DBs</th>
<th>Recommended action</th>
</tr>
</thead>
<tbody>
<tr>
<td>cranberry, raspberry, blackberry</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>Authors will need to submit assembly to INSDC</td>
</tr>
<tr>
<td>pear</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>Authors will need to submit assembly to INSDC</td>
</tr>
<tr>
<td>strawberry</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>Authors will need to submit assembly to INSDC</td>
</tr>
<tr>
<td>grape</td>
<td>☑</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>Contacted authors to submit reference assembly to INSDC &amp; provide VCF. Next contact Journal</td>
</tr>
<tr>
<td>poplar</td>
<td>☑</td>
<td>☑</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>INSDC updated assembly. Next EVA to coordinate with CartograPlant /TreeGenes</td>
</tr>
<tr>
<td>apple, peach, cherry, hazelnut, kiwi</td>
<td>☑</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>Unknown whether VCFs are available. NCGR might follow up</td>
</tr>
<tr>
<td>maize</td>
<td>☑</td>
<td>☑</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>Gramene Maize looking to coordinate with MaizeGDB</td>
</tr>
<tr>
<td>sorghum</td>
<td>☑</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>Contacted multiple authors/studies unsuccessfully</td>
</tr>
<tr>
<td>sorghum</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>SorghumBase coordination with EVA &amp; GRIN</td>
</tr>
</tbody>
</table>
Working towards solutions

- Assembly submissions to INSDC
  - Education & training
  - Elixir cookbook recipe

- Standard file format
  - Converter tools (e.g., excel => VCF)

- Data sharing
  - Minimum standards
  - File validation (community DBs effort)
  - Journals
  - Funding agencies

- BioSamples with germplasm IDs + sample doi/url
  - FAANG project extension
    - Experimental, metadata & bioinformatics standards
    - Reuse tools
Summary of Outcomes

- FAIRifying pilot studies (replaced tmp SNP IDs with rsIDs):
  - SorghumBase & Gramene: 41M sorghum rsIDs
  - Gramene Vitis: 0.3M grape rsIDs
- Standardized germplasm identifiers
  - Gramene, SorghumBase
- Recruited 14 new members
- Discussed synergy with Education & Sci Lit WGs
- Merged with Public Genetic Resources WG

Gramene Workshop
Tuesday, Jan. 16
Palm 8, 4 pm
Future work

● Ensure relevant reference assemblies registered at INSDC by active participation of WG members to:
  ○ Promote data submission to EVA
  ○ Lower barrier for biocuration through training, SOPs, etc.
  ○ Convert historical data into current reference assembly

● Biosamples metadata biocuration hackathon
  ○ Cross-link accessions to germplasm repositories
  ○ Cross-link passport data (germplasm synonyms)
  ○ Index widely used population panels
THANKS - Join our working group, chat with us…!

AgBioData Booth #422
Sunday Opening, Monday & Tuesday lunch

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