AgBioData SGV

Standards for Genetic Variation

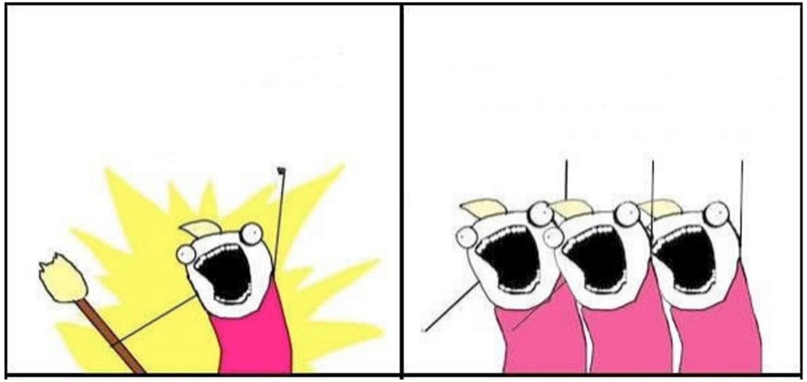
Current status, challenges & future directions

Marcela Karey Tello-Ruiz, PhD
Cold Spring Harbor Laboratory

May 1-2, 2023

AgBioData Workshop 2023

Outline



- Activities & outcomes
- Challenges
- Coming up with solutions…

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

GRIN-Global

USDA ARS Agricultural Research Service

EMBL-EBI

elixir

AgBioData SGV

SORGHUM BASE

SoyBase
Integrating Genetics and Genomics
to Advance Soybean Research

CartograPlant

Gramene

EnsemblPlants

April 13, 2023 meeting
New members always welcomed!

AgBioData Workshop 2023

AgBioData SGV Working Group Goals

1. Identify practical challenges associated with sharing and reusing genetic variation (GV) datasets

2. Bring together a community of data providers, biocurators & computer scientists to promote interoperability and access to GV datasets

3. Support the harmonization and adoption of standards for GV data from various platforms in Plants & Animals

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

AgBioData SGV

Activities & Outcomes

- Monthly meetings
- Biocurators discussions
- Surveys
- AgBio GV data & resources webinar (8 short talks)
- Pilot biocuration of GV sets

=> Identified critical *vs* desirable information from germplasm biocurators

=> Identified practical challenges associated with sharing and reusing GV datasets

=> Progress towards the FAIRification of GV sets (pilot studies)

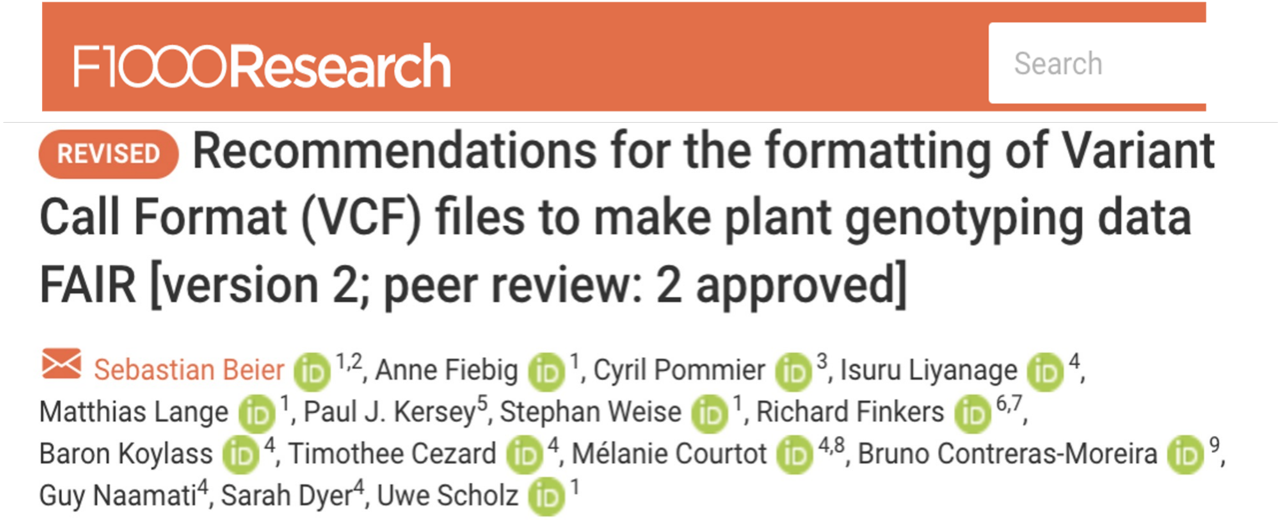
AgBioData SGV

Biocurators meetings

EU-FONDUE recommendations data standards for plants

AgBioData SGV

- 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 iits submission to EVA (and BioSamples)



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

elixir

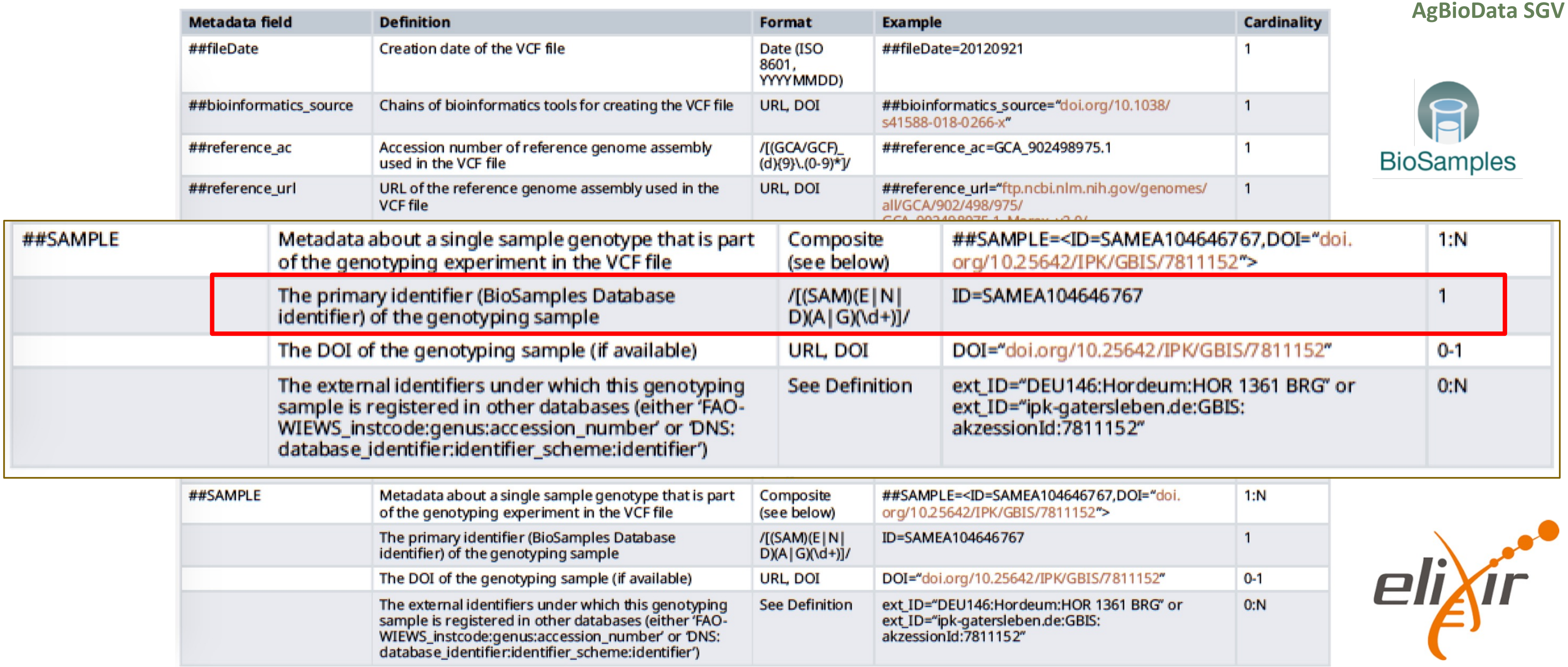
doi: 10.12688/f1000research.109080.2

Summary of recommendations for plant metadata formatting

AgBioData SGV

BioSamples

Table 1. Summary of recommendations for 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/s41588-018-0266-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_Morex_v2.0/	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

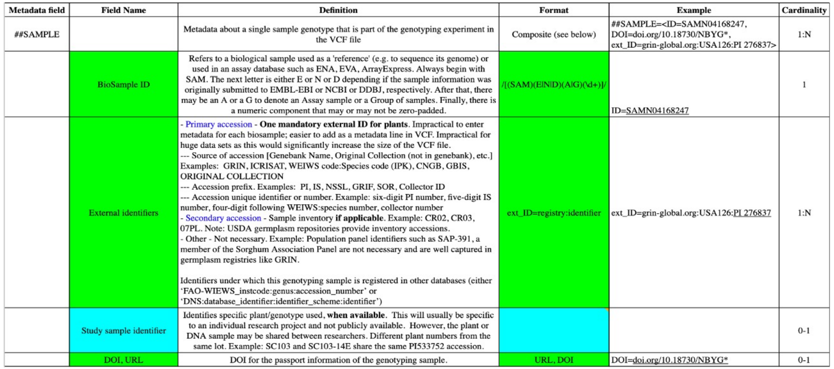
Metadata field	Definition	Format	Example	Cardinality
##SAMPLE	Metadata about a single sample genotype that is part of the genotyping experiment in the VCF file	Composite (see below)	##SAMPLE=<ID=SAMEA104646767,DOI="doi.org/10.25642/IPK/GBIS/7811152">	1:N
	The primary identifier (BioSamples Database identifier) 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

F1000Research

Brier et al (2022). F1000Research. doi: 10.12688/f1000research.109080.2

Outcomes from Biocurators meetings

Additional Suggestions for Plant Samples Metadata associated with VCFs



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 a 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)(\d+)]/	ID=SAMN04168247	1
	External identifiers	- 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 [Genebank Name, Original Collection (not in genebank), 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-391, a member of the Sorghum Association Panel are not necessary and are well captured in germplasm registries like GRIN. 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')	ext_ID=registry:identifier	ext_ID=grin-global.org:USA126:PI 276837	1:N
	Study sample identifier	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 SC103-14E share the same PI533752 accession.			0-1
	DOI, URL	DOI for the passport information of the genotyping sample.	URL, DOI	DOI=doi.org/10.18730/NBYG*	0-1

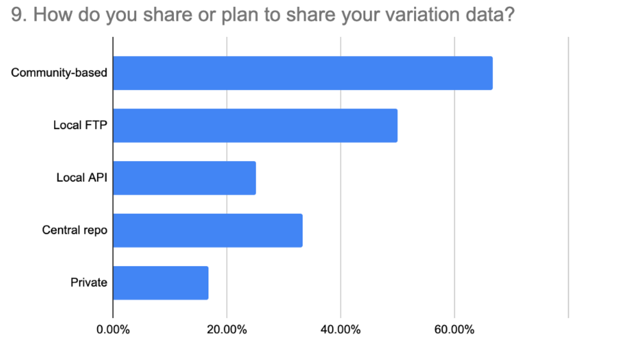
=> BioSamples entries:

- 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

Surveys

AgBioData Survey - Jan. 2023 (14 responses)

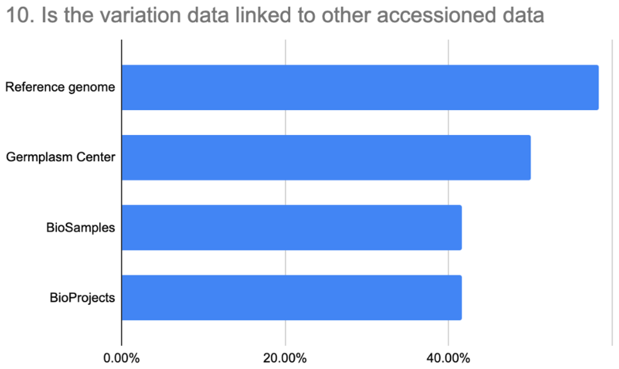
AgBioData SGV



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

Almost all datasets are publicly available but formats and methods for sharing are very diverse

Only 33% deposit to central database like EVA



10. Is the variation data linked to other accessioned data

Cross-linking to important metadata does not always happen

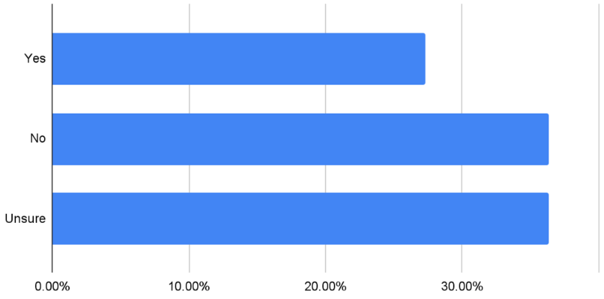
40% do not link to the reference genome

Detailed survey results

AgBioData Survey - Jan. 2023 (14 responses)

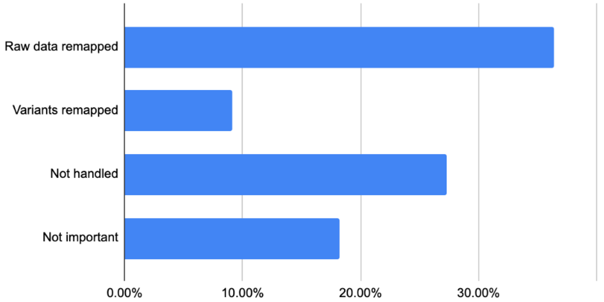
AgBioData SGV

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



Majority do not use stable variant identifiers (e.g., rsIDs)

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



When new genome is available, about one in three respondents remaps the raw data

About half do not update the datasets

[Detailed survey results]

AgBio GV Webinar

AgBio Genetic Variation Webinar

AgBioData SGV

Presentations from multiple resources:

1. GDR (CottonGen, GDV, CGD, PCD) - Sook Jung
2. BreedBase (SGN, Cassava/Yam/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, BGD, FAANGMine, Hymenoptera) - Chris Elsik by proxy
8. Gramene / Ensembl Plants & SorghumBase - Marcela K. Tello-Ruiz

AgBio Genetic Variation Webinar

Outcomes:

- Identified GV datasets in a wide range of Ag species
- Sampled diversity in data submission, formatting, processing, display/analysis tools, interoperability & use cases (Ag bioinformatic resources)
- Recruited new members

AgBioData SGV

Pilot studies

Pilot projects based on readiness of the communities

1. Species with high-quality reference assemblies in INSDC and GV data in AgBio community DBs
2. Species with high-quality reference assemblies & population variation data sets without resources to host large GV data sets
3. Species with high-quality reference assemblies that are developing new GV data sets

AgBioData SGV

Species	Study	Reference assembly in INSDC	VCF available	Sample IDs with DOI/URL from major germplasm repo	VCF in EVA & BioSamples	Samples qualified for cross-linking to other DBs	Recommended action
sorghum	Boatwright et al (2022)	☑	☑	☑	☑	☐	SorghumBase coodinating with EVA & GRIN
sorghum	Cuevas et al (2019) Ahn et al (2021) Cuevas & Prom (2020) Cuevas et al (2018) Ahn et al (2019)	☑	☐	☐	☐	☐	-
strawberry	Hardigan et al (2021)	☐	☑	☑	☐	☐	Authors will need to submit assembly to INSDC
apple, peach, cherry, hazelnut, kiwi		☑	☐	☐	☐	☐	-
pear, cranberry, raspberry, blackberry		☐	☐	☐	☐	☐	Focus on pear. EVA coordinating with GDR
cranberry, raspberry, blackberry		☐	☐	☐	☐	☐	Authors will need to submit assembly to INSDC
poplar	Zhang et al (2019)	☑	☑	☐	☐	☐	Ensembl Plants/Gramene updated assembly. EVA coordinates with CartograPlant /TreeGenes
grape	Dong et al (2023)	☑	☐	☐	☐	☐	PN40024 at ENA, not *V. sylvestris* (sequencing reads provided). Write to Journal. Gramene Vitis coordinates
maize	Grzybowski et al (2023)	☑	☑	☐	☐	☐	Gramene Maize to coordinate with MaizeGDB

AgBioData SGV

Outcomes Summary

- Reviewed guidelines & proposed additional recommendations to support adoption
- Identified existing GV datasets, workflows & technical barriers for data exchange

Challenges Revealed Through Biocuration

- Missing reference genome
- Reference genome not registered at INSDC
- Variation not readily available
  - Request from authors or private FTP
- Variation not in standardized format (VCF)
  - Non-standard format at community DB and no conversion method provided. Two resequencing studies:
    1. “in addition to raw and filtered SNP files, we are releasing GATK GenomicsDB datastores... Making the GATK GenomicsDB datastores from this project publicly available will allow researchers to generate VCF files for the region of interest and calculate accurate values of nucleotide diversity for this region in their population.. VCFs deposited in USDA Box & linked from community DB”
    2. Domesticated genome in ENA; public WGR reads for wild genome (no VCF).
  - Precursor sequencing reads or array tabular output (.xls)

Working towards solutions

- Assembly submissions to INSDC
  - Education & training (partnerships)
  - 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

FAANG
Functional Annotation of Animal Genomes

Thanks!

Breakout Group Questions

AgBioData SGV

1. Sharing raw sequencing or array genotyping data is common practice but not standard variation files (e.g., VCF). How do we address the challenges of standardized formatting and sharing genotyping data?

2. How to incentivize submission of sequence assembly data to INSDC databases?

3. Strategies for encouraging better metadata submission (e.g., external germplasm IDs & doi/url)