AgBioData SGV

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

PAG31

Outline

AgBioData SGV

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

AgBioData SGV

- 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

AgBioData 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 (engagement with Education & Sci. Literature WGs)
  - Participation at AgBioData annual workshop & PAG workshop
  - Data biocuration & coordination across participant resources
  - Promoting FAIRification of GV data & recruiting members at relevant events
  - Merging with Public Genetics Resources WG
  - Reporting to funders

AgBioData Standards for Genetic Variation WG

AgBioData SGV

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

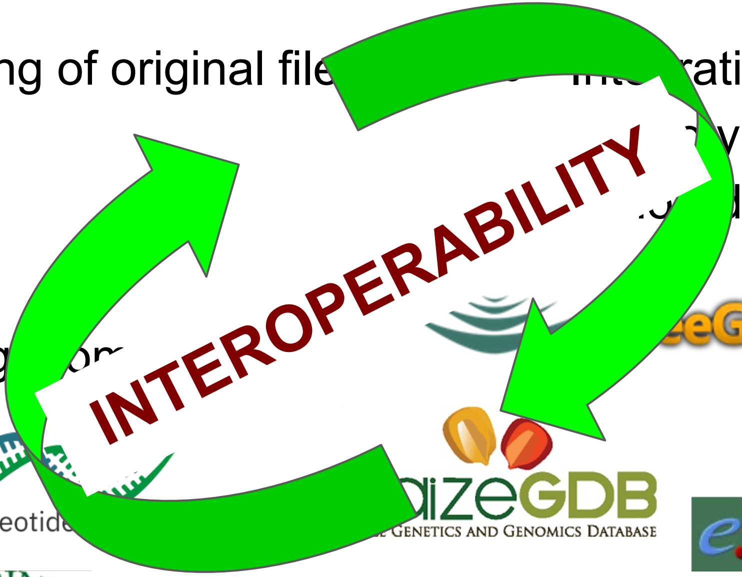
AgBioData SGV

Archival DBs

- Long-term archiving of original file
- Accessioning
  - Study
  - Samples
  - Variants
- Update to newer g

Community/species DBs

- Inte ration between genotypes and
  ypes
  d feature/toolsets



European Variation Archive

ENA
European Nucleotide

BioSamples

GRIN-Global

eeGenes

BREEDBASE

SORGHUM BASE

aizeGDB
GENETICS AND GENOMICS DATABASE

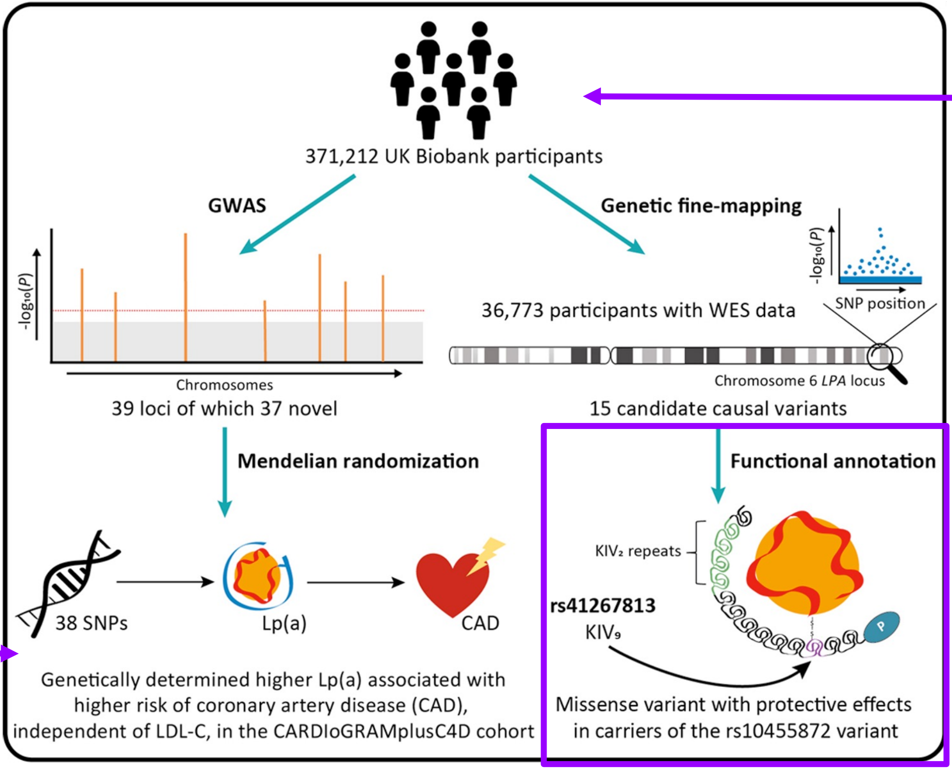
EnsemblPlants

Gramene

GENOME DATABASE FOR VACCINIUM
Genomics, genetics, and breeding resources for blueberry, cranberry, bilberry, and lingonberry research

SoyBase

Lessons from human genetics



371,212 UK Biobank participants
GWAS
Genetic fine-mapping
-log₁₀(P)
Chromosomes
39 loci of which 37 novel
-log₁₀(P)
SNP position
36,773 participants with WES data
Chromosome 6 LPA locus
15 candidate causal variants
Mendelian randomization
Functional annotation
38 SNPs
Lp(a)
CAD
KIV₂ repeats
rs41267813
KIV₉
P
Genetically determined higher Lp(a) associated with higher risk of coronary artery disease (CAD), independent of LDL-C, in the CARDIoGRAMplusC4D cohort
Missense variant with protective effects in carriers of the rs10455872 variant

Biosamples

Trait & disease associations

Functional annotations

rsID (missense) associated with protective trait in disease

Genetic variants

AgBioData SGV

Image taken from doi: 10.1161/ATVBAHA.120.315300

Standards for Genetic Variation – Interoperability

AgBioData SGV

rsIDs

Reference cl

Stable/unique for a

EVA provides 'ss'

and 'rs' (ref) ids fo

variant

rs1234
rs4567
rs7890
…

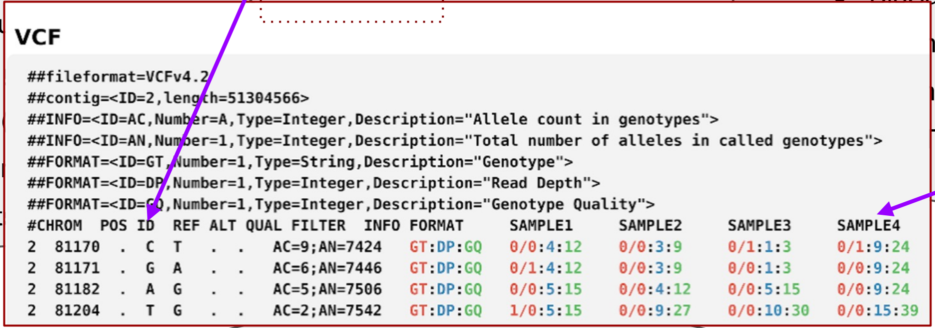
Biosample IDs

- BioSample ID (EVA

ent)

plasm ID (genebanks

: IS 12661, GRIN: PI

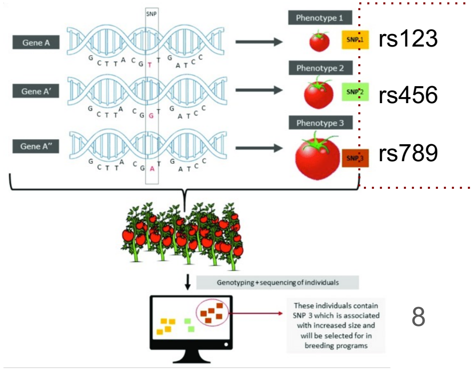


VCF

```
##fileformat=VCFv4.2
##contig=<ID=2,length=51304566>
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
#CHROM  POS  ID  REF ALT QUAL FILTER  INFO FORMAT     SAMPLE1    SAMPLE2    SAMPLE3     SAMPLE4
2  81170  .  C  T   .   .   AC=9;AN=7424  GT:DP:GQ   0/0:4:12   0/0:3:9    0/1:1:3     0/1:9:24
2  81171  .  G  A   .   .   AC=6;AN=7446  GT:DP:GQ   0/1:4:12   0/0:3:9    0/0:1:3     0/0:9:24
2  81182  .  A  G   .   .   AC=5;AN=7506  GT:DP:GQ   0/0:5:15   0/0:4:12   0/0:5:15    0/0:9:24
2  81204  .  T  G   .   .   AC=2;AN=7542  GT:DP:GQ   1/0:5:15   0/0:9:27   0/0:10:30   0/0:15:39
```

Traits

Controlled vocabularies
for GWAS, QTLs, etc.



SNP
Gene A
Gene A'
Gene A''
Phenotype 1
Phenotype 2
Phenotype 3
SNP 1
SNP 2
SNP 3
rs123
rs456
rs789
Genotyping + sequencing of individuals
These individuals contain SNP 3 which is associated with increased size and will be selected for in breeding programs

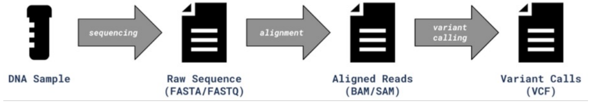
VCF

Variant Call Format

Text file format with meta-info and data for a

variant position in a *genome sequence assembly*

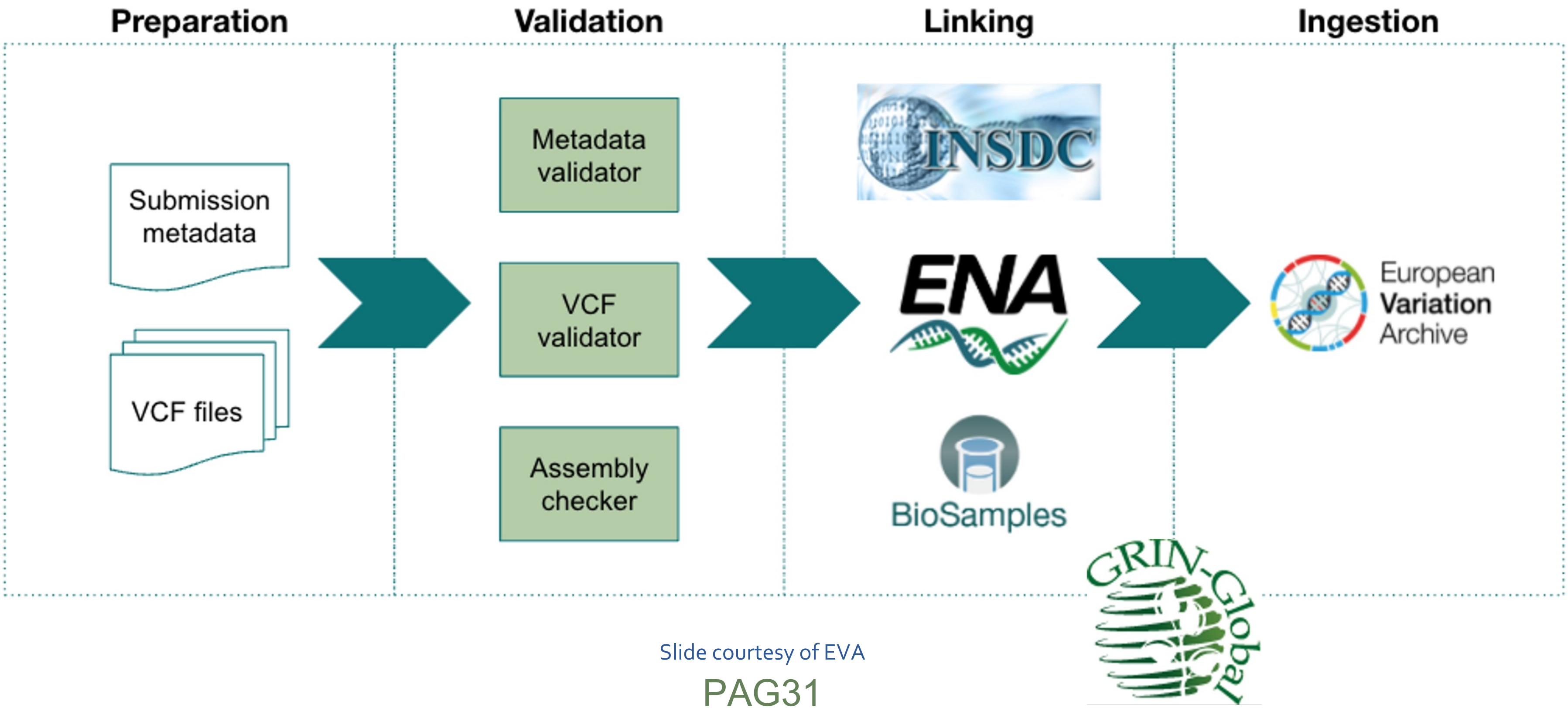
at INSDC



Phenos image modified from:
10.3390/agronomy12040838

Data submission to the European Variation Archive

AgBioData SGV



Slide courtesy of EVA

PAG31

Major challenges associated with genetic variation

AgBioData SGV

- 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.)

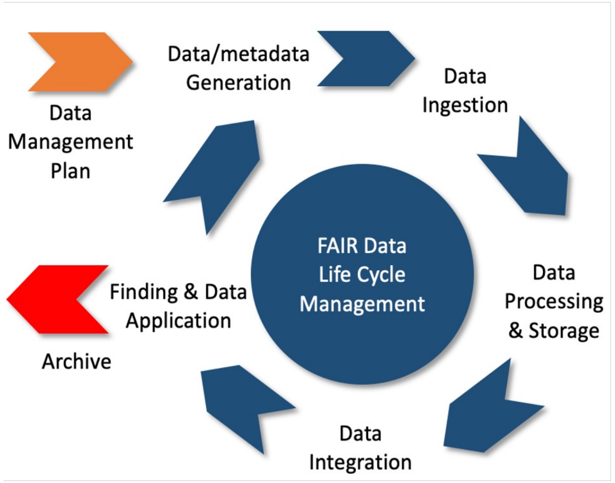
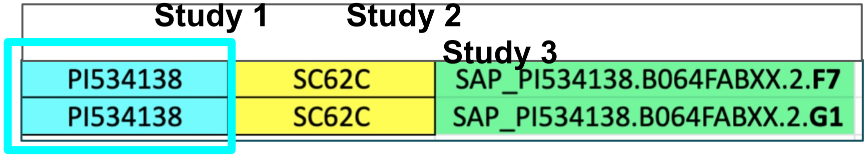

Data Management Plan
Data/metadata Generation
Data Ingestion
Data Processing & Storage
Data Integration
Finding & Data Application
Archive
FAIR Data Life Cycle Management

Image credit: FAIRToolkit

=> Solution: Submit GV to EVA to get rsIDs, unique genetic variant identifiers

Challenges associated with biosamples

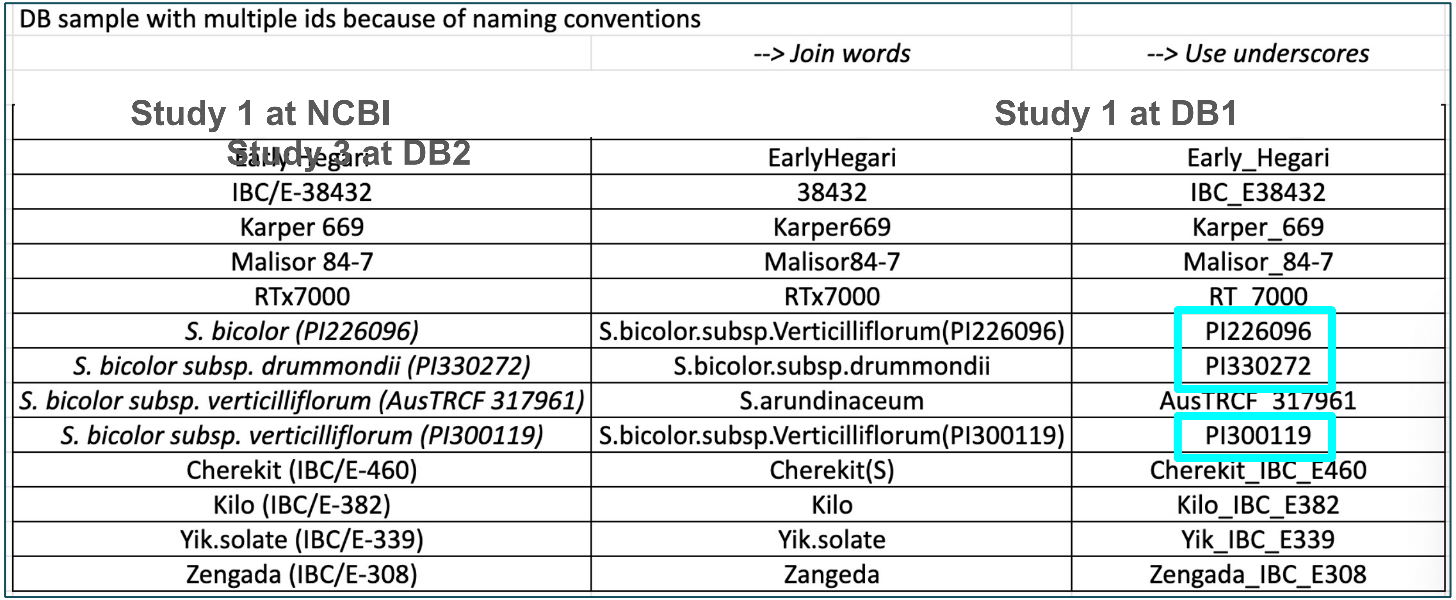
AgBioData SGV



Study 1	Study 2	Study 3
PI534138	SC62C	SAP_PI534138.B064FABXX.2.**F7**
PI534138	SC62C	SAP_PI534138.B064FABXX.2.**G1**

Other names

Matchikah	SAP-416



DB sample with multiple ids because of naming conventions

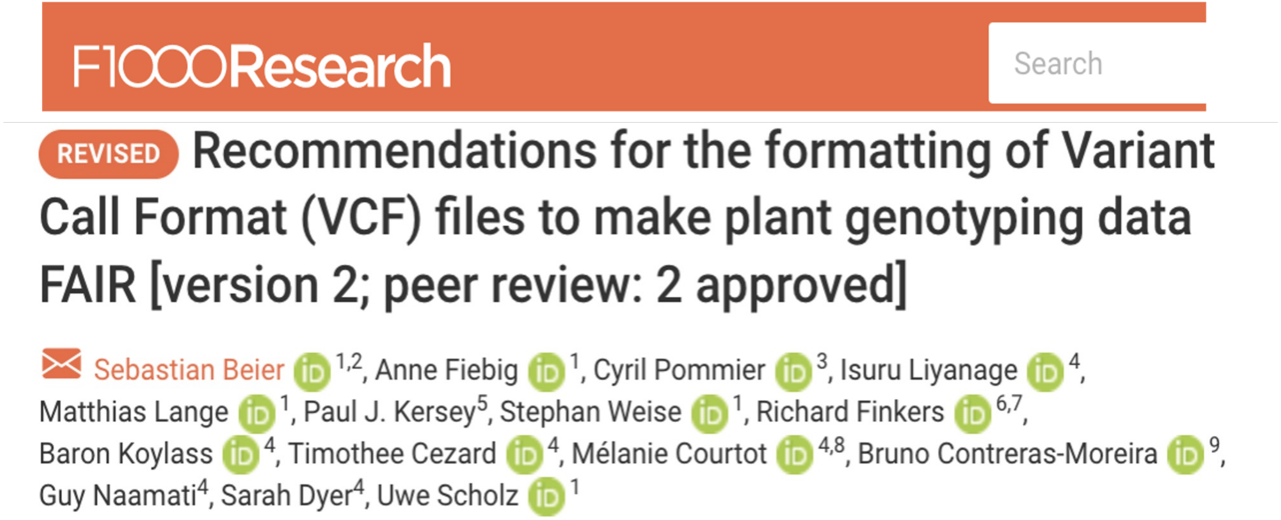
Study 1 at NCBI	Study 3 at DB2	Study 1 at DB1
	--> Join words	*--> Use underscores*
Early-Hegari	EarlyHegari	Early_Hegari
IBC/E-38432	38432	IBC_E38432
Karper 669	Karper669	Karper_669
Malisor 84-7	Malisor84-7	Malisor_84-7
RTx7000	RTx7000	RT_7000
S. bicolor (PI226096)	S.bicolor.subsp.Verticilliflorum(PI226096)	PI226096
S. bicolor subsp. drummondii (PI330272)	S.bicolor.subsp.drummondii	PI330272
S. bicolor subsp. verticilliflorum (AusTRCF 317961)	S.arundinaceum	AusTRCF_317961
S. bicolor subsp. verticilliflorum (PI300119)	S.bicolor.subsp.Verticilliflorum(PI300119)	PI300119
Cherekit (IBC/E-460)	Cherekit(S)	Cherekit_IBC_E460
Kilo (IBC/E-382)	Kilo	Kilo_IBC_E382
Yik.solate (IBC/E-339)	Yik.solate	Yik_IBC_E339
Zengada (IBC/E-308)	Zangeda	Zengada_IBC_E308

GRIN-Global

=> 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 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]**

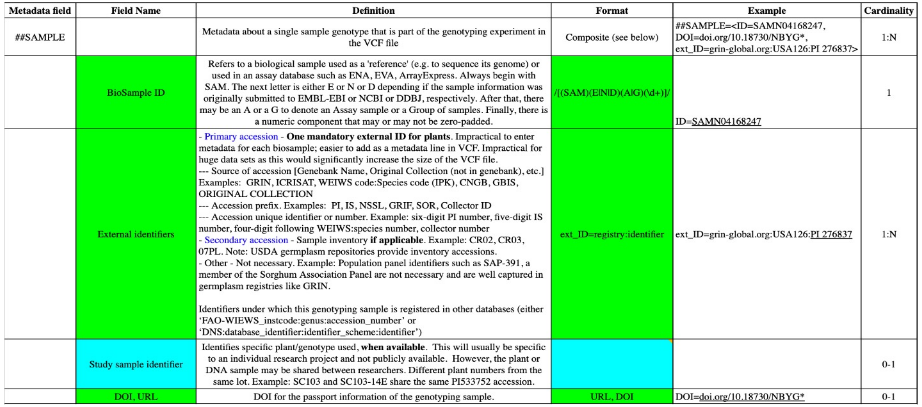
Sebastian Beier[1,2], Anne Fiebig[1], Cyril Pommier[3], Isuru Liyanage[4], Matthias Lange[1], Paul J. Kersey[5], Stephan Weise[1], Richard Finkers[6,7], Baron Koylass[4], Timothee Cezard[4], Mélanie Courtot[4,8], Bruno Contreras-Moreira[9], Guy Naamati[4], Sarah Dyer[4], Uwe Scholz[1]

elixir

doi: 10.12688/f1000research.109080.2

Recommendations from Biocurator 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:

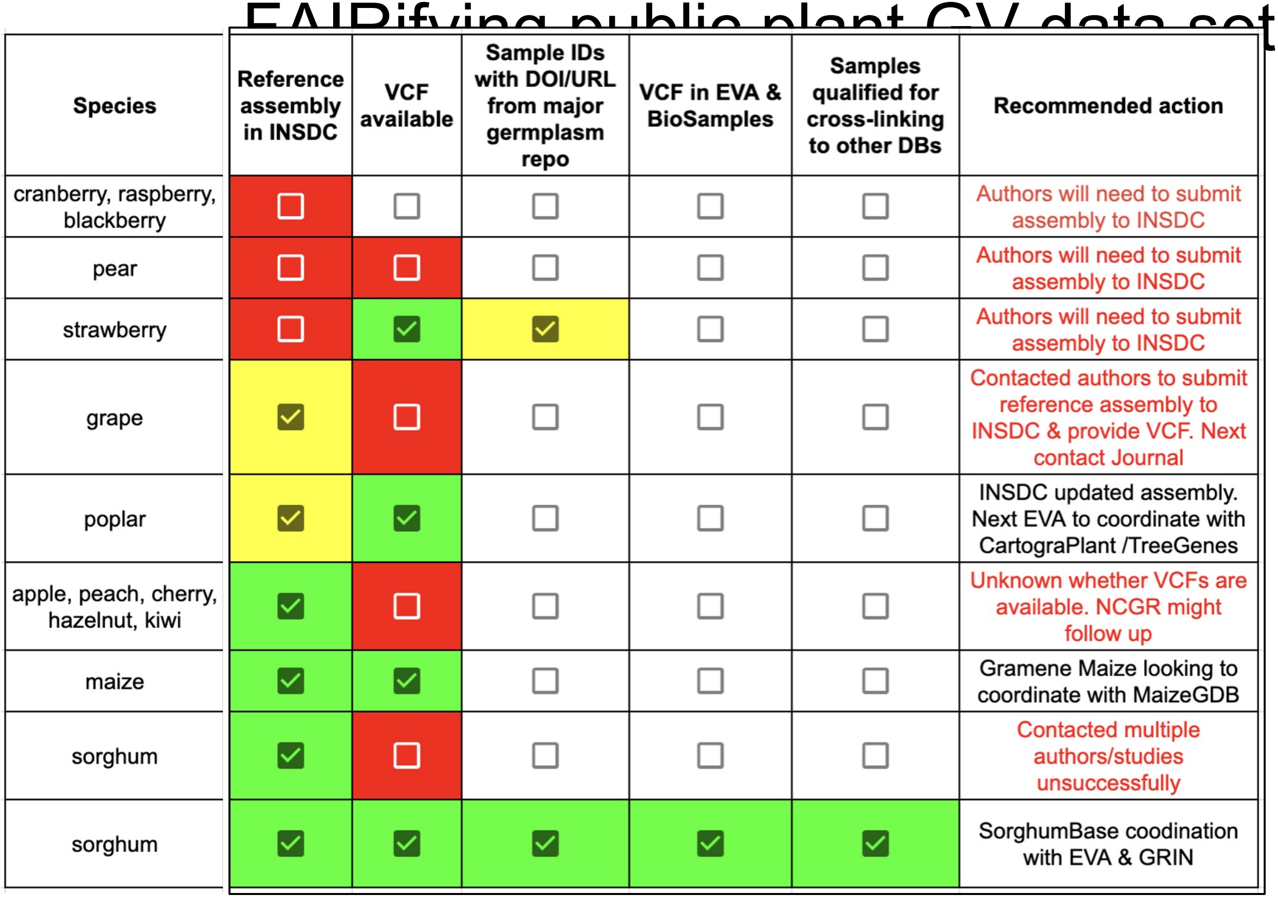
- Require primary external identifier from major germplasm repository (e.g., GRIN, CGIAR, IPK, CNGB) with doi/url
- Recommend including inventory or local number & identifier for the specific plant/genotype used in the study

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

AgBioData SGV

FAIRifying public plant GV data sets



Species	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
cranberry, raspberry, blackberry	🟥 □	□	□	□	□	Authors will need to submit assembly to INSDC
pear	🟥 □	🟥 □	□	□	□	Authors will need to submit assembly to INSDC
strawberry	🟥 □	🟩 ☑	🟨 ☑	□	□	Authors will need to submit assembly to INSDC
grape	🟨 ☑	🟥 □	□	□	□	Contacted authors to submit reference assembly to INSDC & provide VCF. Next contact Journal
poplar	🟨 ☑	🟩 ☑	□	□	□	INSDC updated assembly. Next EVA to coordinate with CartograPlant /TreeGenes
apple, peach, cherry, hazelnut, kiwi	🟩 ☑	🟥 □	□	□	□	Unknown whether VCFs are available. NCGR might follow up
maize	🟩 ☑	🟩 ☑	□	□	□	Gramene Maize looking to coordinate with MaizeGDB
sorghum	🟩 ☑	🟥 □	□	□	□	Contacted multiple authors/studies unsuccessfully
sorghum	🟩 ☑	🟩 ☑	🟩 ☑	🟩 ☑	🟩 ☑	SorghumBase coodination with EVA & GRIN

STOP

…

GO

Working towards solutions

AgBioData SGV

- 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

FAANG
Functional Annotation of Animal Genomes

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 SGV

AgBioData Booth #422

Sunday Opening, Monday & Tuesday lunch

USDA ars Agricultural Research Service

GRIN-Global

SORGHUM BASE

Gramene

EMBL-EBI

elixir

GENOME DATABASE FOR VACCINIUM
Genomics, genetics, and breeding resources for blueberry, cranberry, bilberry, and lingonberry research

SoyBase

CITRUS GENOME DATABASE
Resources for citrus genomics, genetics, breeding and disease research

PULSE CROP DATABASE
Genomic, Genetic, and Breeding Resources for Pulse Crop Improvement

FAANG
Functional Annotation of Animal Genomes

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

BREEDING Insight