Standards for Genetic Variation
Promoting identifiers to improve FAIRness

Timothee Cezard EMBL - EBI

April 29th 2024

AgBioData Standards for Genetic Variation WG

Co-Chairs:

Marcela K. Tello-Ruiz
Timothe Cezard

Most active members:

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

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

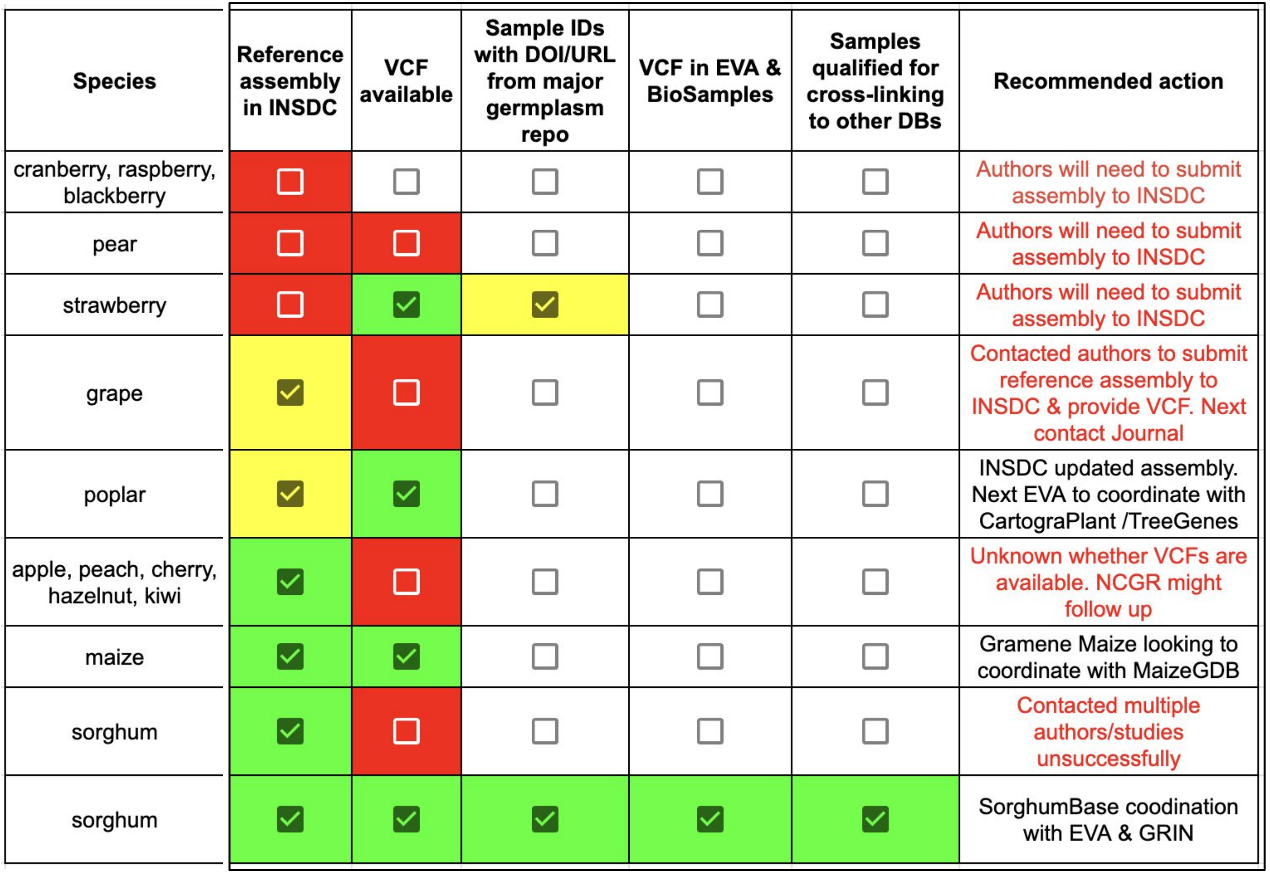
AgBioData SGV Working Group Goals

1. Improve availability and reusability of variation datasets

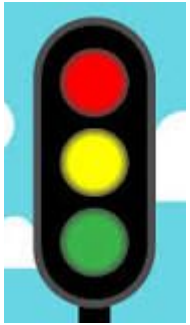
2. Bring together a community of data providers, biocurators & computer scientists to promote interoperability and access to GV datasets

3. Advocate for the increase use of standard format and identifiers for data and metadata

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

Variation Data journey identified

AgBioData SGV

1. Ra

- Large Curation effort to recreate the GV data ✗
- No check done prior to publication ✗

Mandated by Funders

2. Ge

- Variation data readily available ✓
- No check done prior to publication ✗

DRYAD

3. Ge

- Variation data readily available ✓
- Data / Metadata validation ✓

Gramene

SoyBase

ASE FOR ROSACEAE

Resources for Rosaceae Research Discovery and Crop Improvement

Standards for Genetic Variation – Interoperability

Genomics variation data

VCF

Variant Call Format

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

Metadata

INSDC

BioSample

- Name, Source, Location Date
- Germplasm ID (genebanks ICRISAT: IS 12661, GRIN: PI 276837)

vcf-validator

https://github.com/EBIvariation/vcf-validator

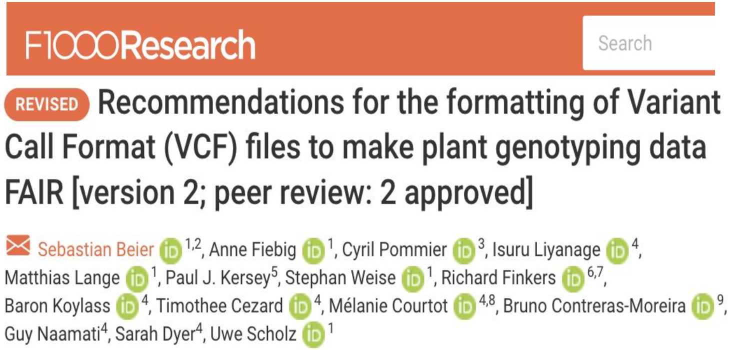
Community defined checklist

miappe

Recommendations for data standards for plants

AgBioData SGV

- Guidelines on FAIR handling of GV data published in 2022
  - How to create and format VCF files
  - Step by step guide on how best to validate and submit data to ENA, BioSamples and EVA

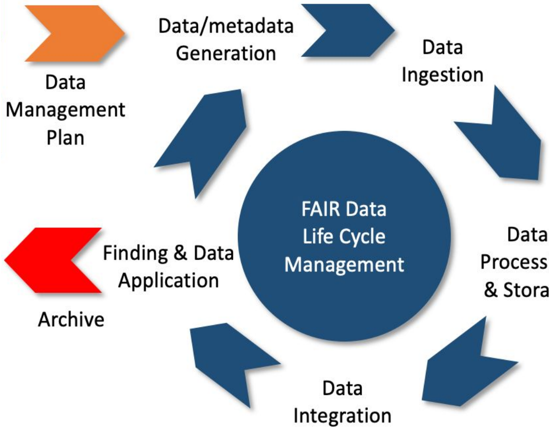


Recommendations for the formatting of Variant Call Format (VCF) files to make plant genotyping data FAIR [version 2; peer review: 2 approved]

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]

doi: 10.12688/f1000research.109080.2

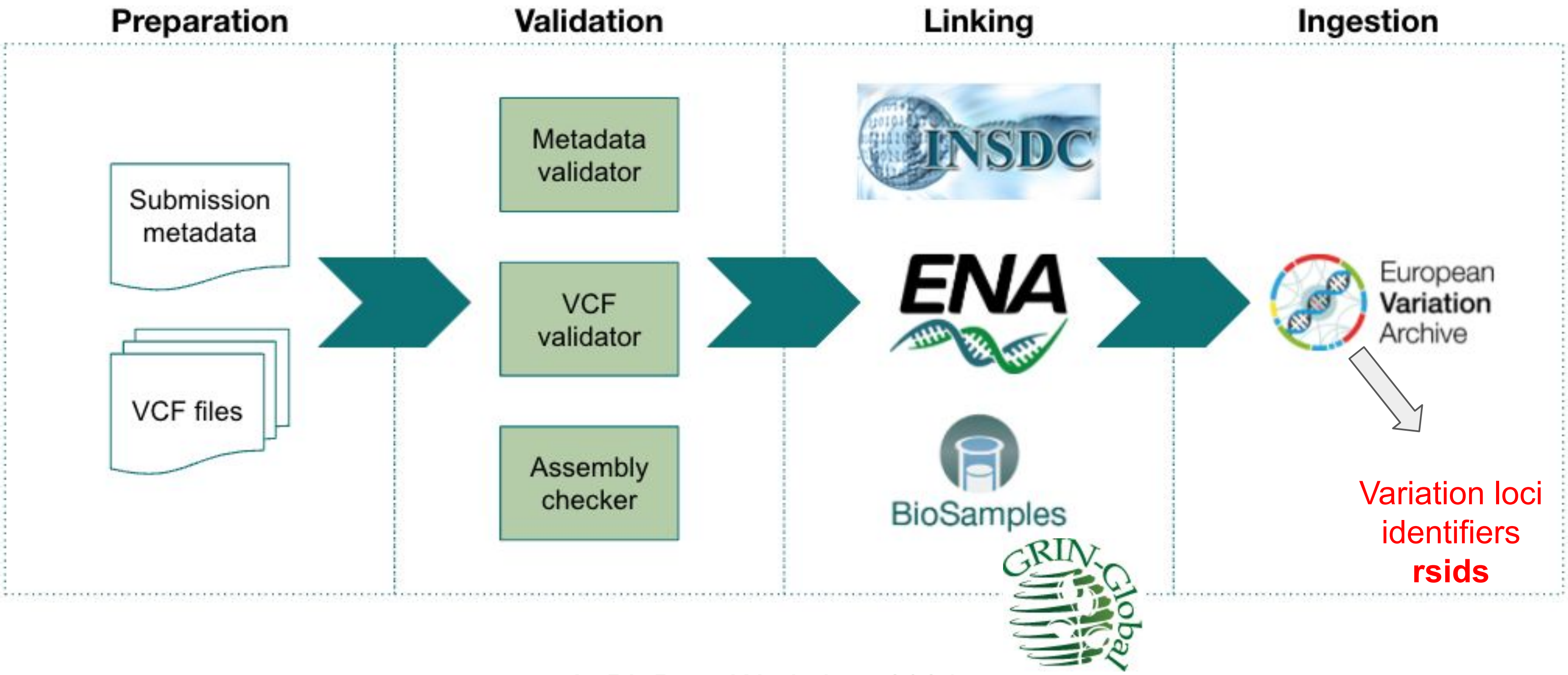
Plant genomic and genetic variation data submission



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

Data submission - European Variation Archive

AgBioData SGV

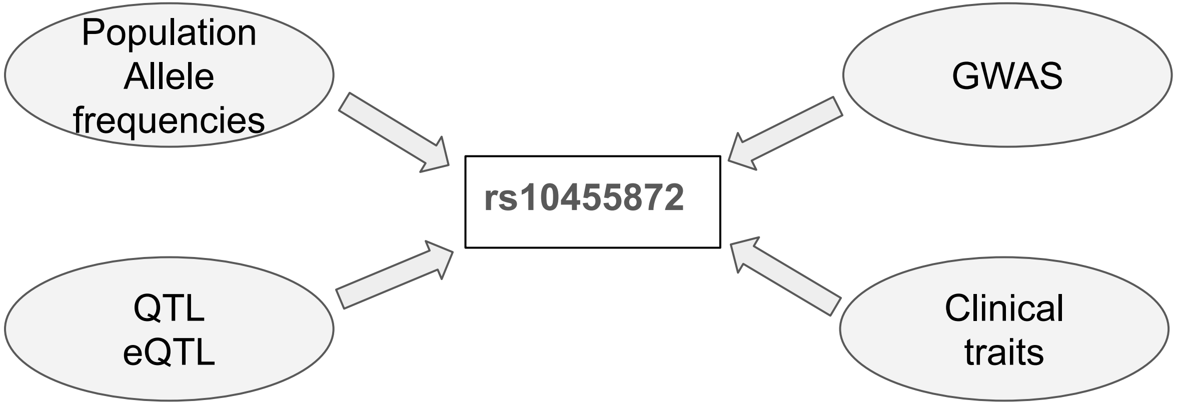


Variation loci identifiers **rsids**

AgBioData Workshop 2024

Variant identifiers

- Globally unique long term accessions
- Identify Variable loci on a genome
- Stable across genome assembly version



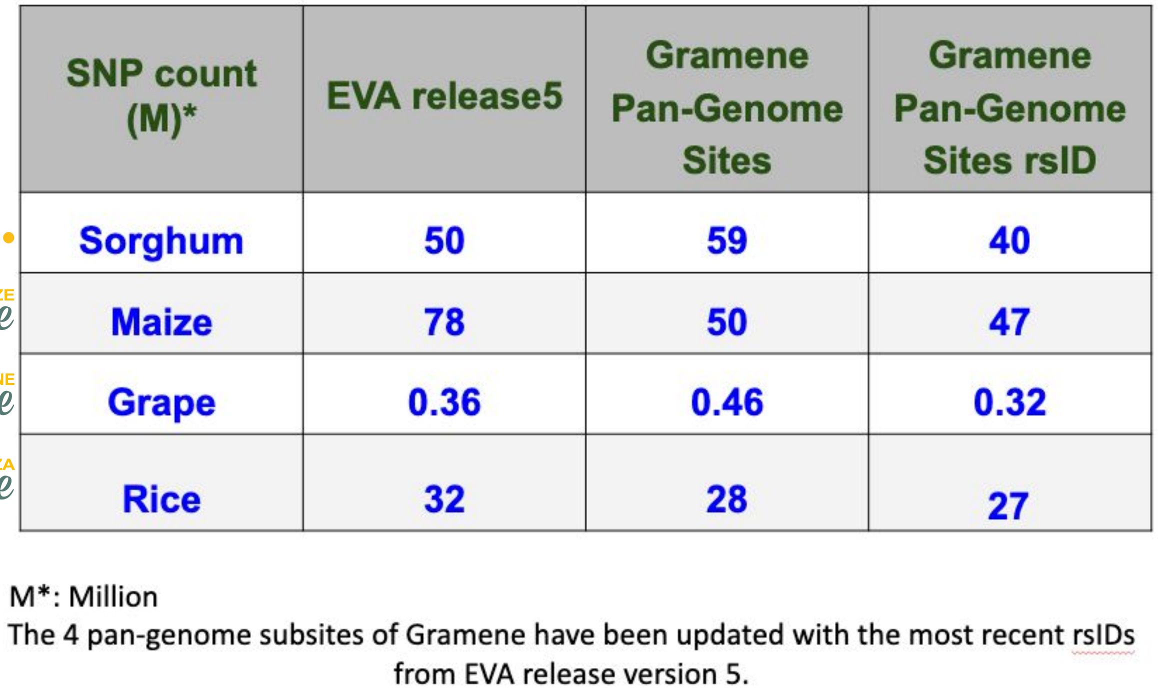
~ 1 M Publications Link to a RS ids

Integration of RSids (non-human)

Integrated with multiple resources

- Ensembl
- UCSC
- NCBI genome data viewer
- Alliance of Genome Resources

Promoting use of RSids - Gramene / SorghumBase



SNP count (M)*	EVA release5	Gramene Pan-Genome Sites	Gramene Pan-Genome Sites rsID
Sorghum	50	59	40
Maize	78	50	47
Grape	0.36	0.46	0.32
Rice	32	28	27

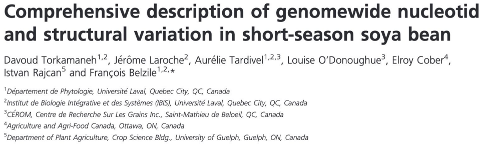
M*: Million

The 4 pan-genome subsites of Gramene have been updated with the most recent rsIDs from EVA release version 5.

Slide courtesy of Marcela

Promoting use of RSids - Soybase

AgBioData SGV



Comprehensive description of genomewide nucleotide and structural variation in short-season soya bean

Davoud Torkamaneh[1,2], Jérôme Laroche[2], Aurélie Tardivel[1,2,3], Louise O'Donoughue[3], Elroy Cober[4], Istvan Rajcan[5] and François Belzile[1,2,*]

[1] *Département de Phytologie, Université Laval, Quebec City, QC, Canada*

[2] *Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Quebec City, QC, Canada*

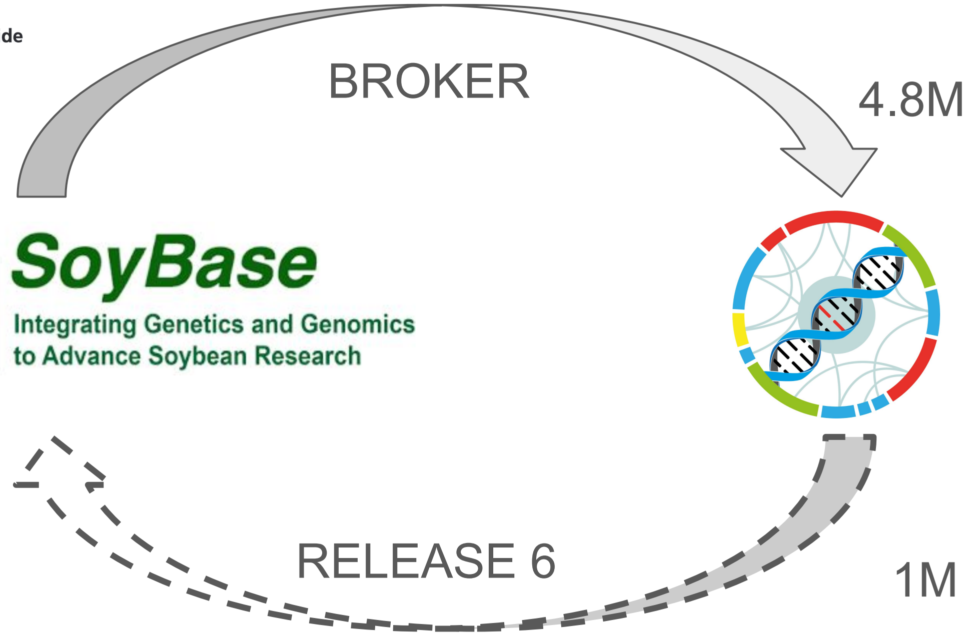
[3] *CÉROM, Centre de Recherche Sur Les Grains Inc., Saint-Mathieu de Beloeil, QC, Canada*

[4] *Agriculture and Agri-Food Canada, Ottawa, ON, Canada*

[5] *Department of Plant Agriculture, Crop Science Bldg., University of Guelph, Guelph, ON, Canada*

SoyBase

Integrating Genetics and Genomics to Advance Soybean Research



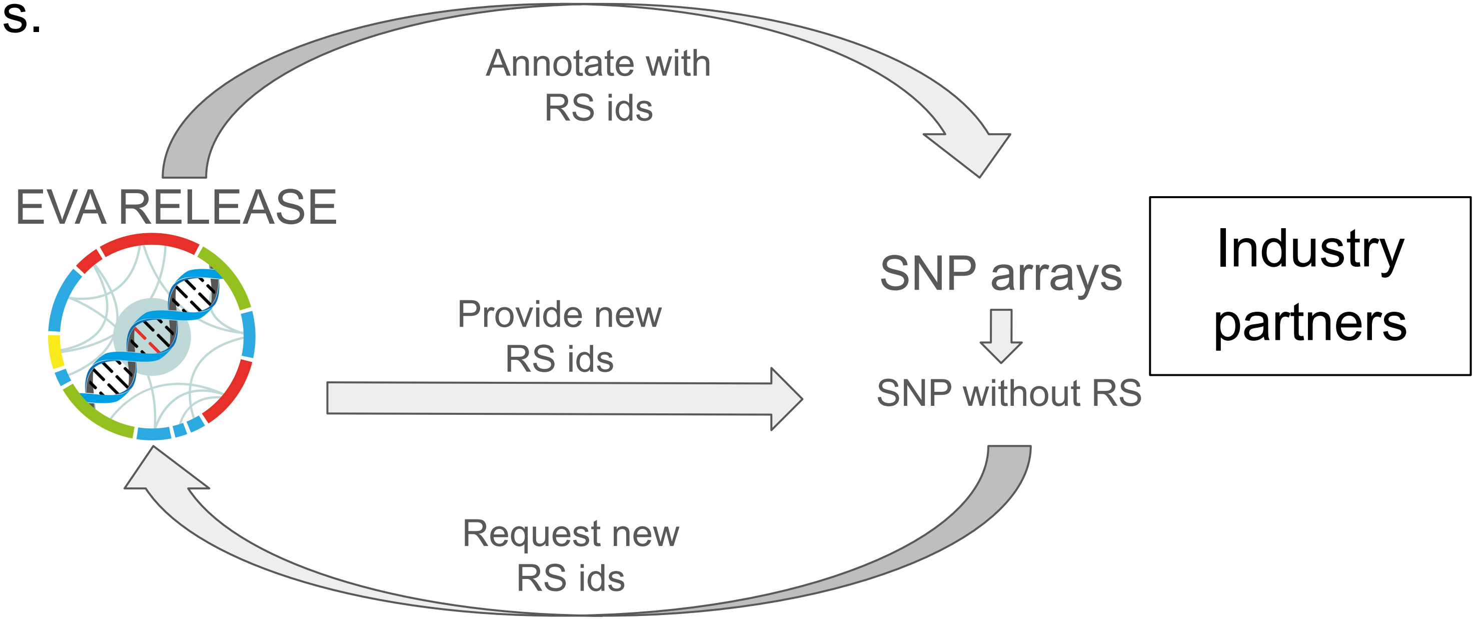
4.8M variants

1M new RSids

AgBioData Workshop 2024

Promoting use of RSids - Industry collaboration

Using RS ids in SNP panels would help users map to genomics coordinates.



Promoting use of RSids - Industry collaboration

Develop a community marker panel with RS ids:

- Sorghum 2.4K SNPs (AgriPlex)
- 26 Markers without RS ids were assigned one

AgriPlex GENOMICS

Leverage Group's contact to start discussion with Thermofisher

ThermoFisher SCIENTIFIC

Summary of Outcomes

- FAIRifying pilot studies
  - Identified data journeys
  - Highlight curation challenges
- Metadata: Standardized germplasm identifiers
- Promoting usages RSids
  - In community database
  - Data Aggregator
  - Industry partners

Writing White paper

Public Genomic Resources

Merged with SGV with complementary goals

- How to identify haplotypes ?
- Where should haplotype and/or allele databases be hosted for best accessibility and continuity ?
- How to describe identify Merged datasets ?

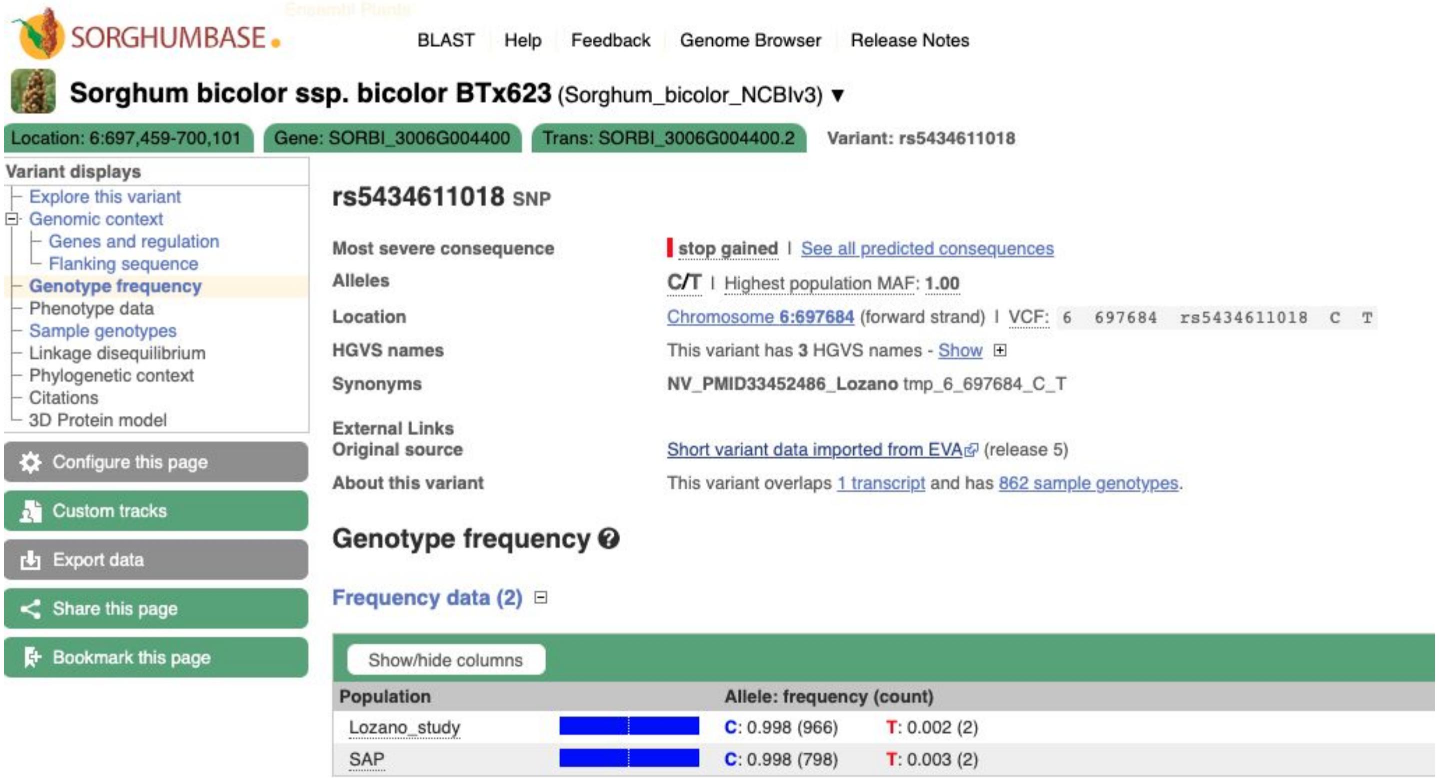
AgBioData SGV

Breakout Group Questions

1. Data submission:
   a. What are (if any) the barriers to submitting variation data to a central or community database?
   b. Have you ever submitted to a central database (NCBI, ENA, EVA)
2. Metadata:
   a. What metadata would be required for you to make the genomics variation reusable?
3. RS ids:
   a. Were you aware of rs ids before the presentation?
   b. Do you think rs ids would be useful for you?

Thanks!

RSids examples



Sorghum bicolor ssp. bicolor BTx623 (Sorghum_bicolor_NCBIv3) ▼

Location: 6:697,459-700,101 | Gene: SORBI_3006G004400 | Trans: SORBI_3006G004400.2 | Variant: rs5434611018

rs5434611018 SNP

Most severe consequence	stop gained I See all predicted consequences
Alleles	C/T I Highest population MAF: 1.00
Location	Chromosome 6:697684 (forward strand) I VCF: 6 697684 rs5434611018 C T
HGVS names	This variant has 3 HGVS names - Show ⊞
Synonyms	NV_PMID33452486_Lozano tmp_6_697684_C_T
External Links Original source	Short variant data imported from EVA (release 5)
About this variant	This variant overlaps 1 transcript and has 862 sample genotypes.

Genotype frequency

Frequency data (2) ⊟

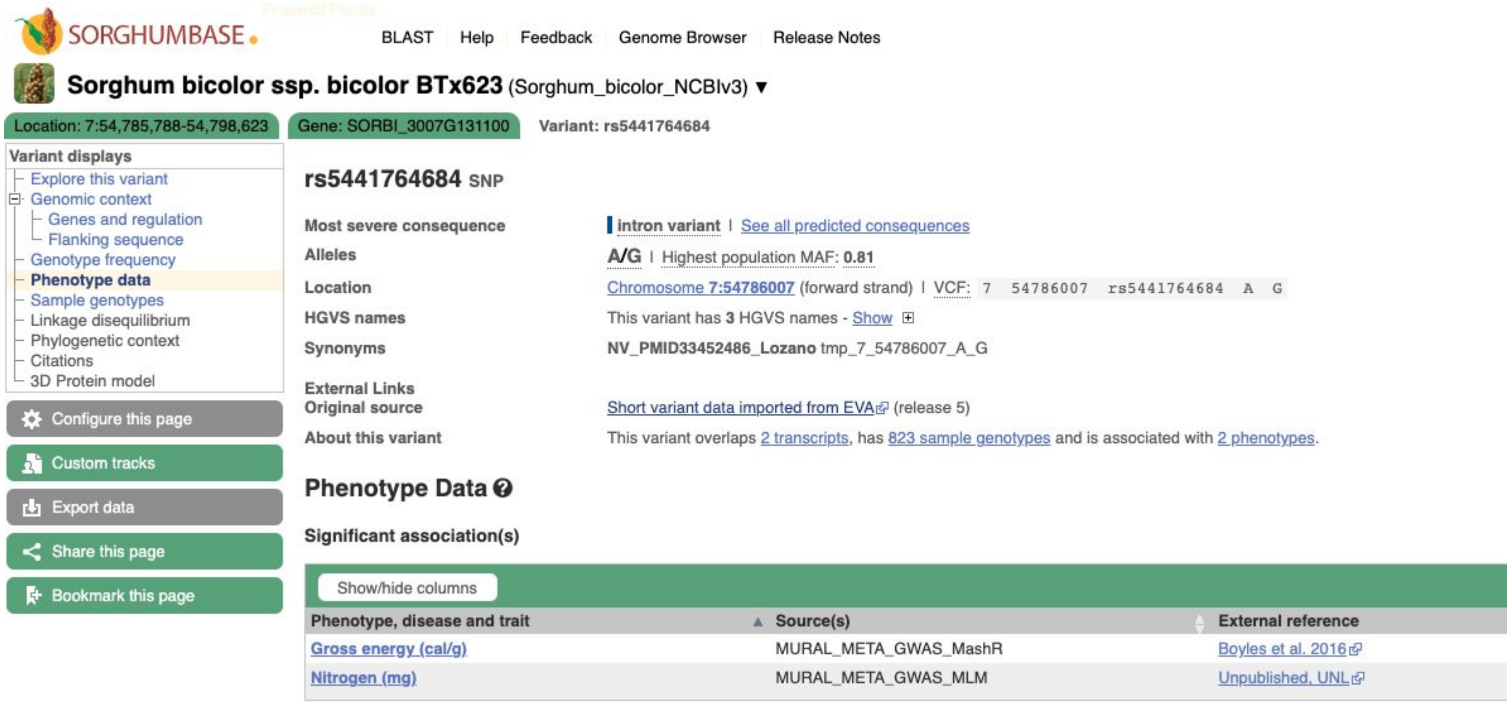
Show/hide columns

Population		Allele: frequency (count)	
Lozano_study		C: 0.998 (966)	T: 0.002 (2)
SAP		C: 0.998 (798)	T: 0.003 (2)

Link

RSids examples

AgBioData SGV



Sorghum bicolor ssp. bicolor BTx623 (Sorghum_bicolor_NCBIv3) ▼

Variant: rs5441764684

rs5441764684 SNP

Most severe consequence	intron variant | See all predicted consequences
Alleles	A/G | Highest population MAF: **0.81**
Location	Chromosome 7:54786007 (forward strand) | VCF: `7 54786007 rs5441764684 A G`
HGVS names	This variant has **3** HGVS names - Show ⊞
Synonyms	**NV_PMID33452486_Lozano** tmp_7_54786007_A_G
External Links	
Original source	Short variant data imported from EVA (release 5)
About this variant	This variant overlaps 2 transcripts, has 823 sample genotypes and is associated with 2 phenotypes.

Phenotype Data ?

Significant association(s)

Phenotype, disease and trait ▲	Source(s)	External reference
Gross energy (cal/g)	MURAL_META_GWAS_MashR	Boyles et al. 2016
Nitrogen (mg)	MURAL_META_GWAS_MLM	Unpublished, UNL

Link