Standards for Genetic Variation

Promoting the Use of Reference Identifiers for Genetic Markers in Agricultural Research

Marcela Karey Tello-Ruiz, PhD
Cold Spring Harbor Laboratory

AgBioData SGV

AgBioData Standards for Genetic Variation WG

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

Co-Chairs:

Marcela K. Tello-Ruiz
Timothee Cezard

AgBioData SGV Working Group Goals

AgBioData SGV

=> Bring together a community of agricultural GV data providers, biocurators & computer scientists to:

1. Improve FAIRness of Ag genotypic (and phenotypic) variant datasets for reuse

2. Promote interoperability and access to GV datasets

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

Challenges with (non)FAIR variation datasets

AgBioData SGV

Mandated by Funders & Encouraged by Journals

Only raw data is shared

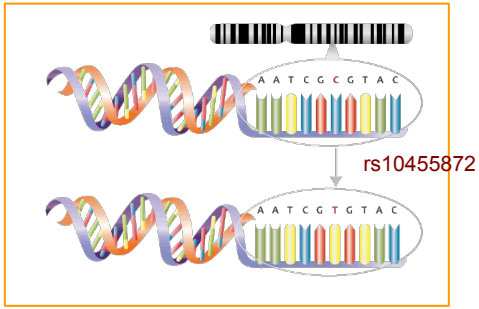
Dataset is shared as supplementary data or local DB

Dataset is shared in generalist FAIR data repository

Dataset uses standard identifiers & formats, and shared in specialized FAIR data repository

Not FAIR → FAIR

Power of using rsIDs

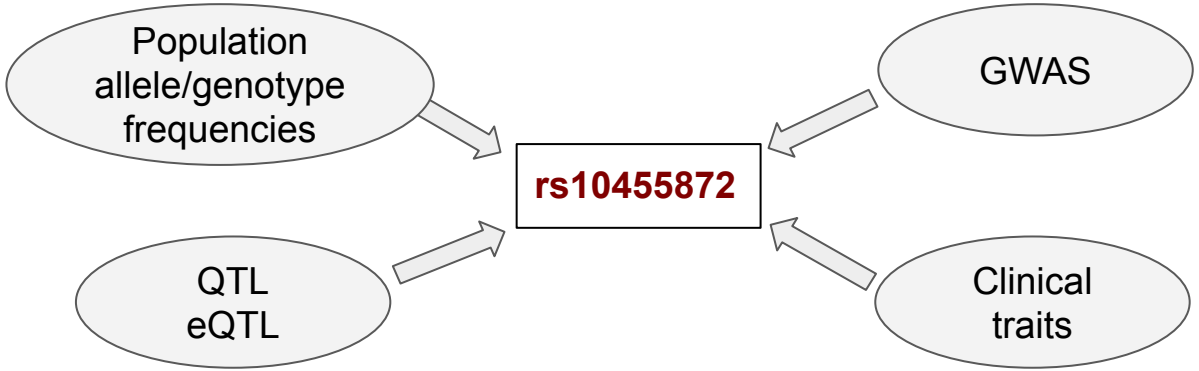


What is an rsID?

- Reference SNP cluster ID
- Identifies a variable genomic locus
- Globally unique, persistent accession
- Stable across genome assembly versions & crop varieties

AgBioData SGV

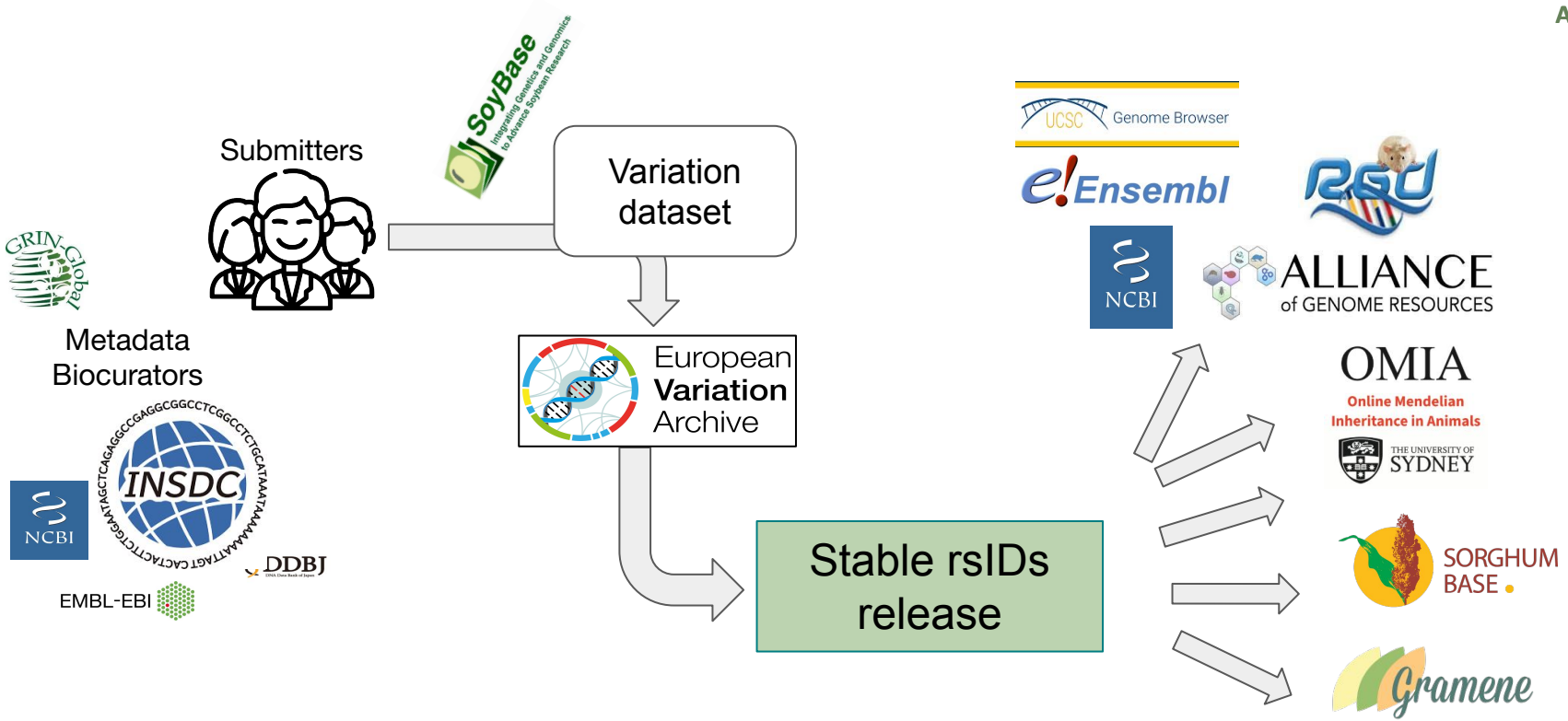
Several data types aggregated around a marker



Europe PMC

~ 1 M Publications linked to an rsID

Integration of rsIDs in downstream resources



AgBioData SGV

Data journey for GV datasets & recommended actions

AgBioData SGV

Request assembly to be submitted to INSDC

Encourage or broker SNP submission to EVA; promote using standard IDs & formats

Adopt rsIDs, germplasm IDs & controlled vocabularies

Integrate with other standardized data types & link to other DBs & repos

Other AgBio DBs provided overview & progress towards FAIRifying GV data for white paper

Not FAIR → FAIR

Promoting use of rsIDs – Gramene / SorghumBase

AgBioData SGV

Gramene PanGenome	Reference Crop (release #)	# SNPs (M)	# rsIDs (M)
SORGHUMBASE. Genomic resources for the sorghum research community 	Sorghum (R9)	78	41
MAIZE Gramene Comparative plant genomics focused on maize varieties 	Maize (R5)	220	79
ORYZA Gramene Comparative plant genomics focused on rice varieties 	Rice (R8)	66	27
GRAPEVINE Gramene Comparative plant genomics focused on grapevine varieties 	Grape (R4)	0.5	0.3

FAIR Standards for Agricultural GV Data

AgBioData SGV

rsIDs
Reference cluster ID
EVA provides stable/unique identifiers for non-human markers

Germplasm IDs
Provided by major germplasm repo

rs123
rs456
rs789

PI276837
IS12661

VCF
Variant Call Format

```
##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
```

GWAs, QTLs
Associate ontology terms with rsIDs

Ontology terms
Controlled vocabulary to describe sorghum traits associated with rsIDs

CO_324_0000079: Grain weight
CO_324_0000027: Flowering time
CO_324_0000250: Grain protein content

Crop Ontology
for agricultural data

Integration of rsIDs (non-human)

AgBioData SGV

Integrated with multiple resources:

- Ensembl
- UCSC
- NCBI genome data viewer
- Alliance of Genome Resources
- OMIA (animals)
- **Gramene** (maize, rice, grape)
- **SorghumBase** (sorghum)

Promoting rsIDs in SNP arrays – Industry collaboration

AgBioData SGV

eurofins | Genomics

STANDARD BIOTOOLS

NEOGEN CORPORATION

DArT diversity arrays technology

GENALICE TECHNOLOGY FOR PEOPLE & SCIENCE

LGC

BIOSEARCH TECHNOLOGIES

TWIST BIOSCIENCE

intertexs

intertek Total Quality. Assured.

BAYER

illumina

ThermoFisher SCIENTIFIC

AgriPlex GENOMICS

Promoting use of rs IDs - Industry collaboration

AgBioData SGV

Develop a community marker panel with rsIDs:

- Sorghum 2.4K SNPs (AgriPlex)
- 26 markers without rsIDs were assigned one

SORGHUMBASE

AgriPlex GENOMICS

European Variation Archive

Summary of Outcomes

AgBioData SGV

- FAIRifying pilot studies
  - Identified data journeys
  - Highlighted curation challenges
- Metadata
  - Additional recommendations for VCF
  - Standardized germplasm identifiers
- Interaction with other WGs
  - Public Genetic Resources (merged) & Education (publication => Global Bioinformatics Education Summit)
- Promoting adoption of rsIDs
  - Community databases
  - Data aggregator & cross-link to other DBs
  - Outreach via new training material
  - Industry partners

Writing white paper

AgBioData SGV

Thanks!

Join Our Breakout Room!

AgBioData SGV

1. Based on the data journey (outline of our manuscript), are there examples of use cases not represented?

2. Are there LLMs that could be leveraged for the adoption of GV standards?

3. Future directions… Leverage strategies from human genetics (e.g., ACMG/AMP guidelines for clinical variants)

Lessons learned from medical genetics: ACMG/AMP
Recommendations for Mendelian Variant Interpretation

AgBioData SGV

- Colossal effort 2 years, 100s participants, many surveys. Original publication 2015; several modifications thereafter, many disease-focused (expert panels).
- Describes process for guidelines, quantitative classification criteria.
- Proposes standard terminology to classify variants (e.g., pathogenic, benign, uncertain significance with numeric levels)
- Criteria using typical types of variant evidence (e.g., population data, computational data, functional data, segregation data)

	Benign		Pathogenic			
	Strong	Supporting	Supporting	Moderate	Strong	Very strong
Population data	MAF is too high for disorder BA1/BS1 **OR** observation in controls inconsistent with disease penetrance BS2			Absent in population databases PM2	Prevalence in affecteds statistically increased over controls PS4	
Computational and predictive data		Multiple lines of computational evidence suggest no impact on gene /gene product BP4 Missense in gene where only truncating cause disease BP1 Silent variant with non predicted splice impact BP7 In-frame indels in repeat w/out known function BP3	Multiple lines of computational evidence support a deleterious effect on the gene /gene product PP3	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before PM5 Protein length changing variant PM4	Same amino acid change as an established pathogenic variant PS1	Predicted null variant in a gene where LOF is a known mechanism of disease PVS1
Functional data	Well-established functional studies show no deleterious effect BS3		Missense in gene with low rate of benign missense variants and path. missenses common PP2	Mutational hot spot or well-studied functional domain without benign variation PM1	Well-established functional studies show a deleterious effect PS3	
Segregation data	Nonsegregation with disease BS4		Cosegregation with disease in multiple affected family members PP1	Increased segregation data →		
De novo data				De novo (without paternity & maternity confirmed) PM6	De novo (paternity and maternity confirmed) PS2	
Allelic data		Observed in *trans* with a dominant variant BP2 Observed in *cis* with a pathogenic variant BP2		For recessive disorders, detected in trans with a pathogenic variant PM3		
Other database		Reputable source w/out shared data = benign BP6	Reputable source = pathogenic PP5			
Other data		Found in case with an alternate cause BP5	Patient's phenotype or FH highly specific for gene PP4			

10.1038/gim.2015.30