Standards for Genetic Variation

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

Timothee Cezard EMBL - EBI

PAG 32

Challenges with (non)FAIR variation datasets

AgBioData SGV

Mandated by Funders

Only raw data is shared

Dataset is shared in paper's supplementary

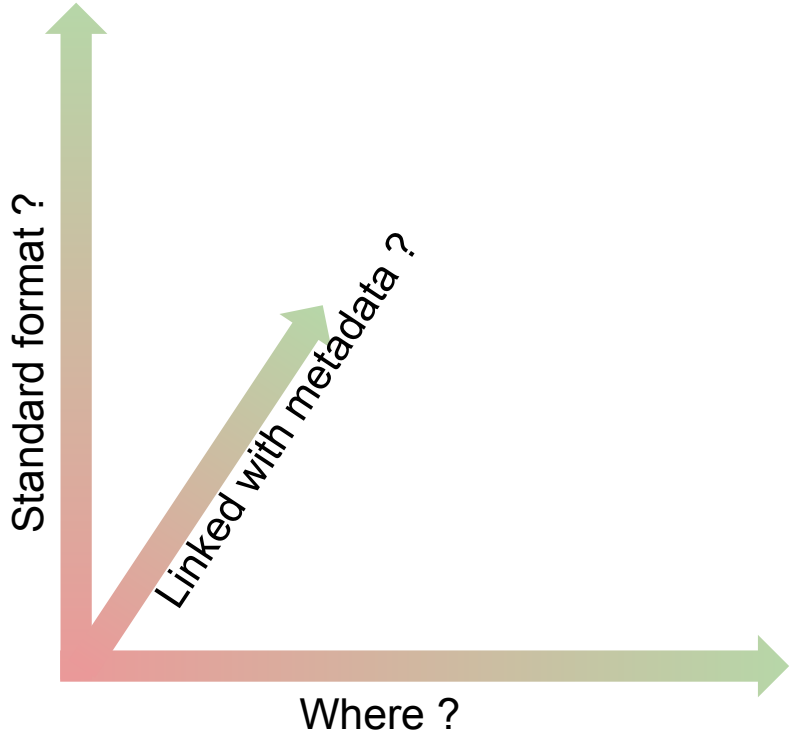
Dataset is share in generalist FAIR data repository

Dataset is share in specialist FAIR data repository

Not FAIR → FAIR

Challenges with (non)FAIR variation datasets

AgBioData SGV



Increasingly complex context

- New tools and technologies
- New assemblies being generated
- Complex metadata

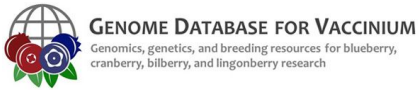
Standards for Genetic Variation WG

Co-Chairs:

- Marcela K. Tello-Ruiz
- Timothee Cezard

Most active members:

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



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

PAG32: AgBioData Workshop

Landscape analysis through curations

AgBioData SGV

Curation Decision tree

Reuse

Cannot Reuse

- GV data not in standardized format
- Missing reference genome assembly
- GV data not readily available (e.g., private FTP)
- Samples not using standard

Improving Interoperability via standards

AgBioData SGV

1. **Genomics variation dataset: VCF** Variant Call Format

[Specification](#) from GA4GH

1. Metadata: BioSamples

Variant detection: bcftools, GATK
Variant imputation: PLINK2
Validation: [vcf-validator](#)

1. Markers: rs IDs

Improving Interoperability via standards

AgBioData SGV

1. Genomics variation dataset: VCF

Entity shared between NCBI, EBI, DDBJ

1. Metadata: BioSamples

Container for any metadata

1. Markers: rs IDs

Community defined checklist

Improving Interoperability via standards

AgBioData SGV

1. Genomics variation dataset: VCF

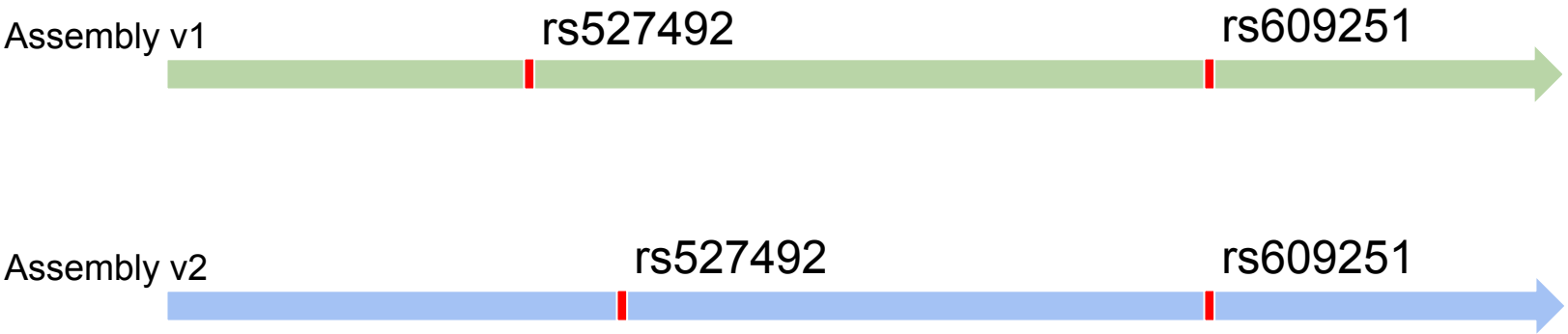
1. Metadata: BioSamples

1. **Markers: rs IDs**

Ref SNP identifiers (rs IDs)

Location on a genome associated with a variation

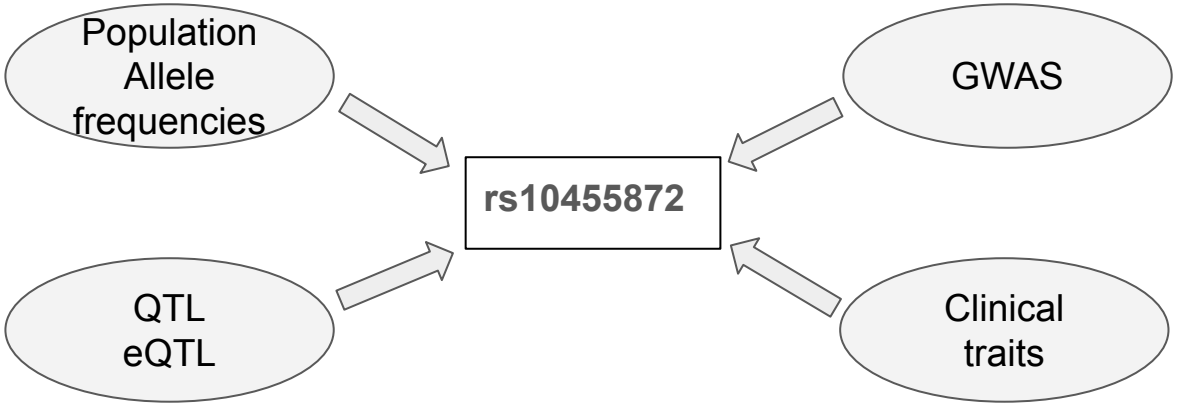
- Globally unique long term accessions
- Identify Variable loci on a genome
- Stable across genome assembly version



Power of using rs IDs

AgBioData SGV

Natural candidate for aggregating several data around marker

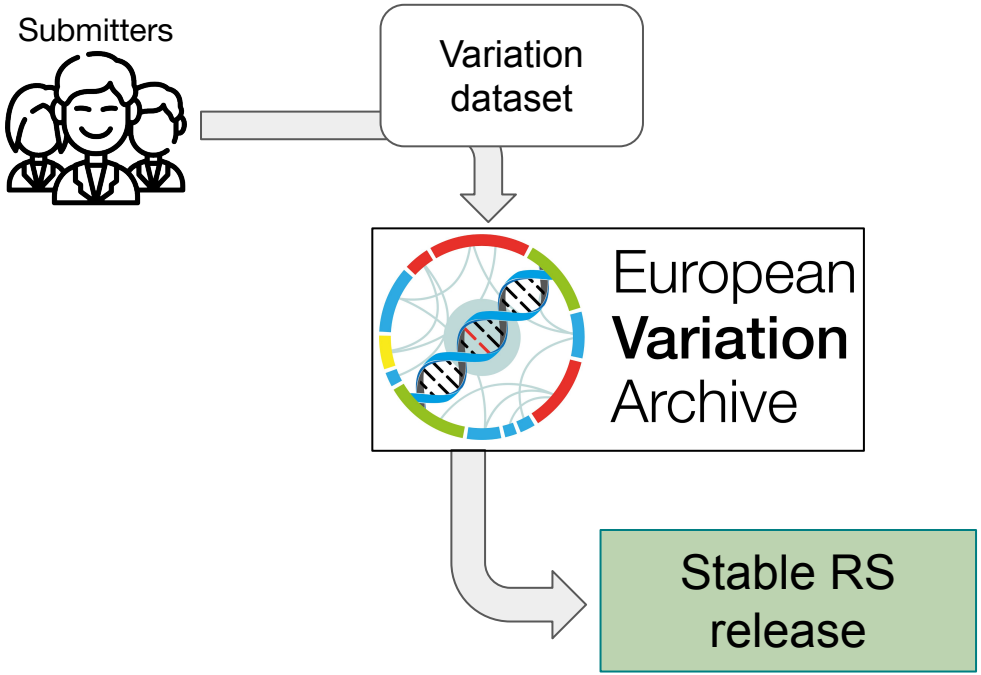


Use in publication to describe makers

Europe PMC

~ 1 M Publications Link to a RS ids

How to generate rs IDs ?



Promoting use of RSids - Gramene / SorghumBase

AgBioData SGV

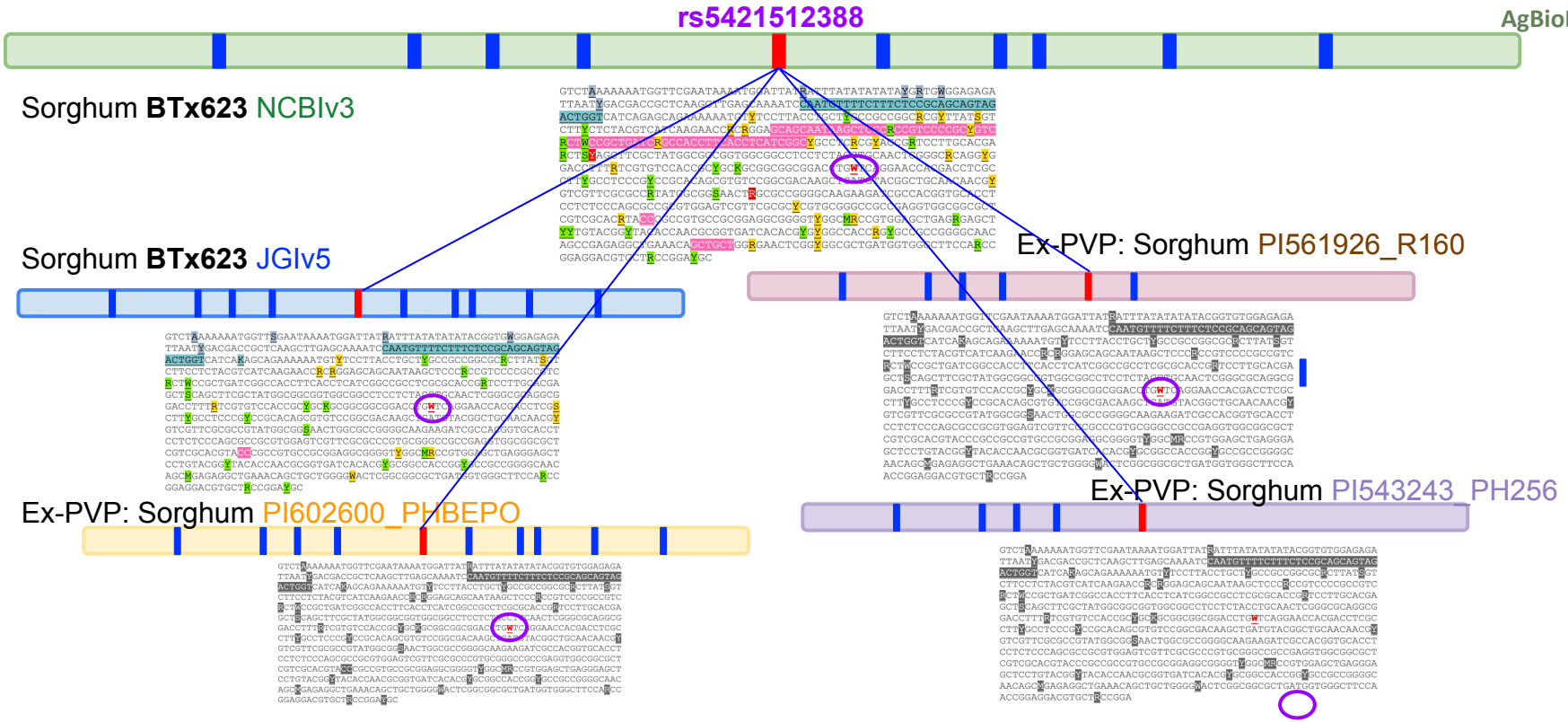
SNP count (M)*	EVA release5	Gramene Pan-Genome Sites	Gramene Pan-Genome Sites rsID
Sorghum	50	59	40
Maize	78	50	47
Grape	0.36	0.46	0.32
Rice	32	28	27

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 SorghumBase

rsIDs: stable identifiers to transverse Pan-Genomes



AgBioData SGV

Slide courtesy of SorghumBase

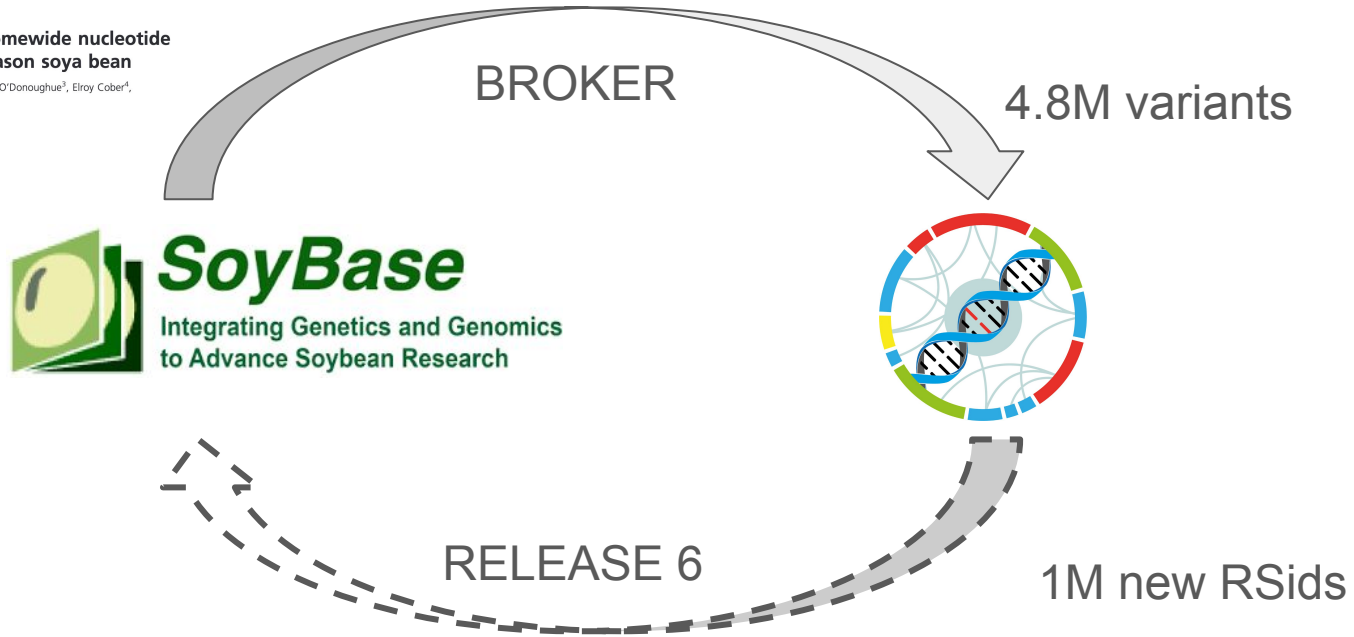
Promoting use of RSids - Soybase

AgBioData SGV

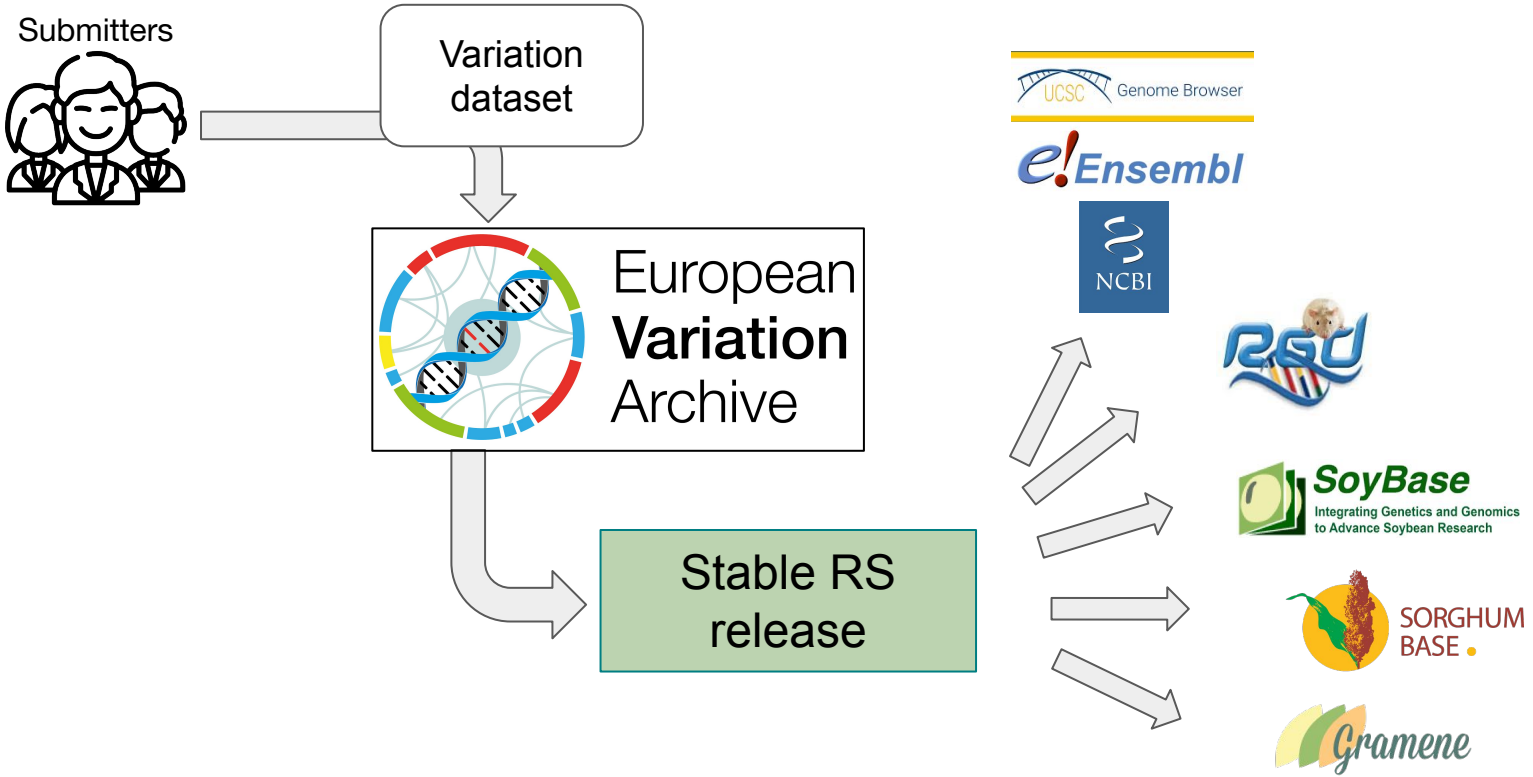
Comprehensive description of genomewide nucleotide and structural variation in short-season soya bean

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]CEROM, 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



Integration of rs IDs in downstream resources



Promoting use of rs IDs

AgBioData SGV

Workshop for AgBioData communities to highlight markers associated with rs ids.

Ensembl

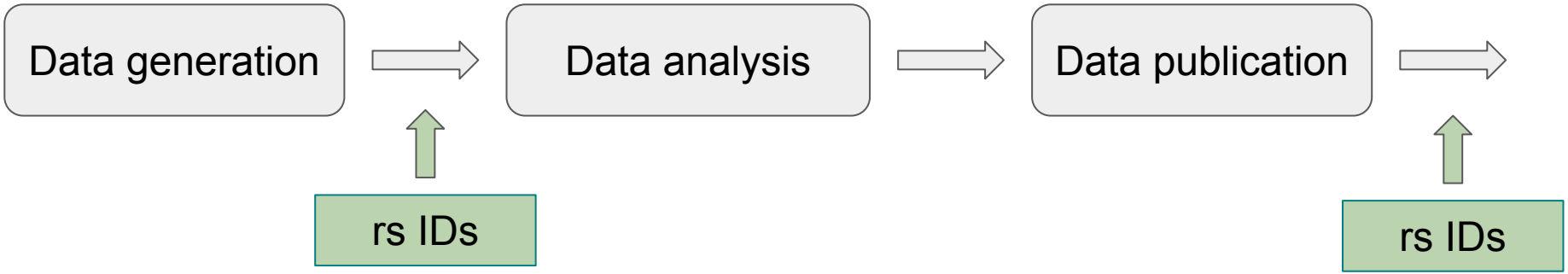
maizeGDB
Maize Genetics and Genomics Database

SORGHUMBASE

OMIA
Online Mendelian Inheritance in Animals
The University of Sydney

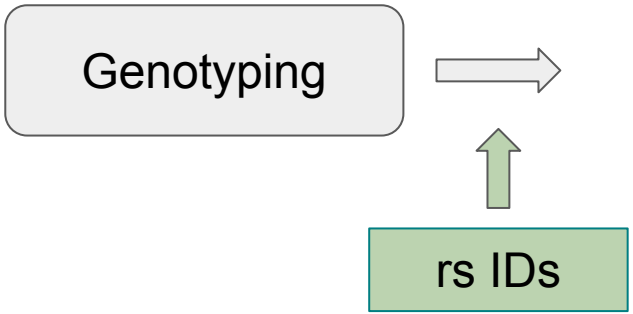
SoyBase
Integrating Genetics and Genomics to Advance Soybean Research

Promoting use of rs IDs



Promoting use of rs IDs

AgBioData SGV




```
marker1    rs102743256    A    T
marker2    rs398477430    C    A
marker3    rs214779348    G    C
...
```

Easier to convert to genomic coordinates

Improve interoperability with other DB downstream

Promoting use of rs IDs - Industry collaboration

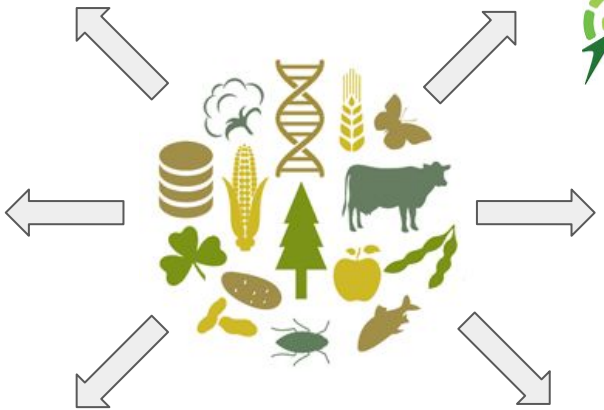
AgBioData SGV

Develop a community marker panel with RS ids:

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

Promoting use of rs IDs - Industry collaboration

AgBioData SGV



eurofins | Genomics

STANDARD BIOTOOLS

NEOGEN CORPORATION

DArT diversity arrays technology

GENALICE TECHNOLOGY FOR PEOPLE & SCIENCE

LGC

BIOSEARCH TECHNOLOGIES

TWIST BIOSCIENCE

interteks

illumina

ThermoFisher SCIENTIFIC

Conclusions & Future work

AgBioData SGV

- AgBioData Standard Genetic Variation meets monthly
- Recommendations for standard to be used
- Adoption of standards by community members
- Outreach internally, externally and to industry partners

⇒ Writing White paper

Come and talk to us

AgBioData SGV

Booth 207

EMBL-EBI

Workshop on Sunday at 8:00

European **Variation** Archive

Poster 345

Ensembl

Poster 108