AgBioData SGV

From Fragmented Data to Shared Knowledge: Adopting rsIDs in Agricultural Genomics

Marcela Karey Tello-Ruiz, PhD
AgBioData Standards for Genetic Variation Working Group

2026 PAG AgBioData Workshop

AgBioData Standards for Genetic Variation WG

AgBioData SGV

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

SORGHUM BASE

USDA ARS Agricultural Research Service

SoyBase

European Variation Archive

EnsemblPlants

EMBL-EBI

GRIN-Global

elixir

Gramene

BREEDING Insight

TreeGenes

FAANG Functional Annotation of Animal Genomes

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

MaizeGDB Maize Genetics and Genomics Database

GrainGenes A Database for Triticeae and Avena

A Community-Led Path Toward FAIR Variant Data

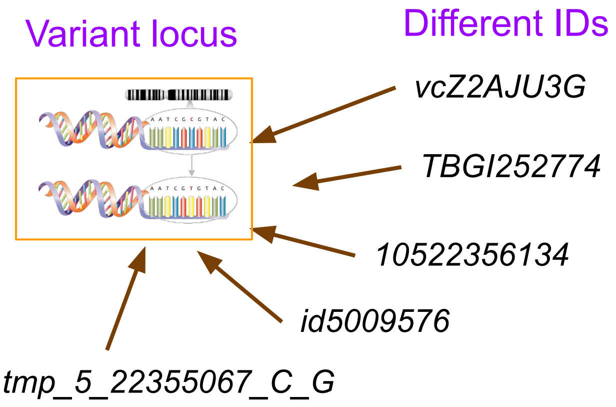
Goals:

- Assess existing standards
- Identify communities at different data-maturity stages
- Focus on promoting rsID adoption via EVA
- Provide practical FAIR guidance*

Standardizing Genetic Variation for FAIR Agricultural Genomics

- Genetic variation data are critical for **breeding, trait discovery & adaptation**
- **Agricultural genomics** generates vast amounts of SNP data that is hard to reuse & integrate

Why hard to reuse?



Variant locus

Different IDs

vcZ2AJU3G

TBGI252774

10522356134

id5009576

tmp_5_22355067_C_G

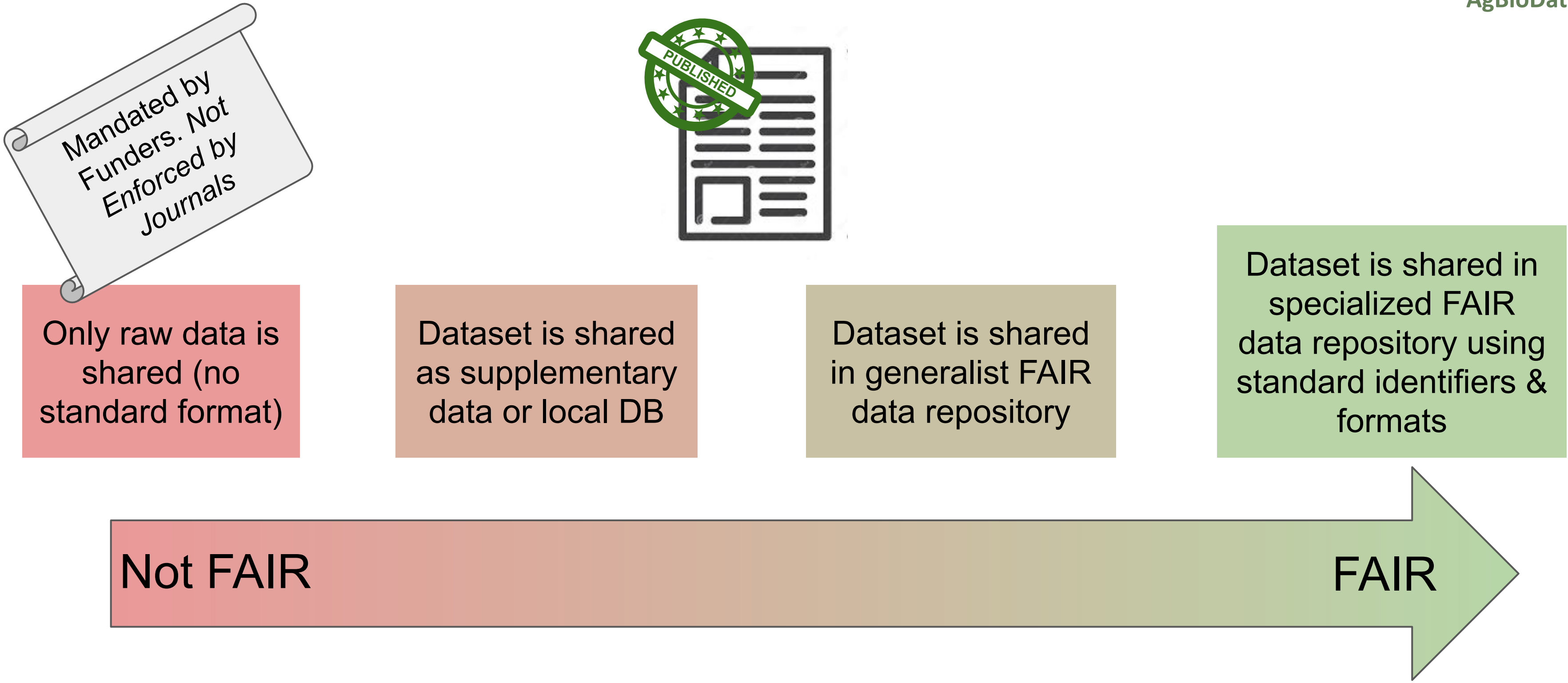
SNPs often labeled with:

- Project- or array-specific identifiers
- Assembly-dependent coordinates

Data spread across:

- Community databases
- Supplementary tables
- Commercial genotyping platforms

Challenges with (non)FAIR variation datasets



Mandated by Funders. Not Enforced by Journals

Only raw data is shared (no standard format)

Dataset is shared as supplementary data or local DB

Dataset is shared in generalist FAIR data repository

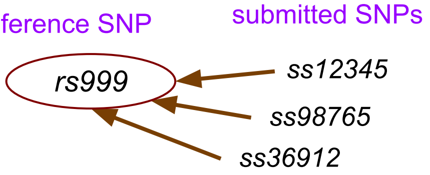
Dataset is shared in specialized FAIR data repository using standard identifiers & formats

Not FAIR

FAIR

How human genomics has tackled the SNP naming problem & why the problem is getting worse in agriculture?

The Power of Using rsIDs



reference SNP

submitted SNPs

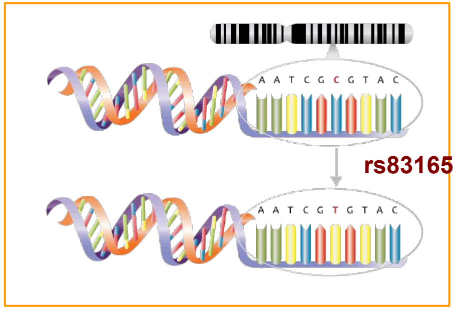
rs999

ss12345

ss98765

ss36912

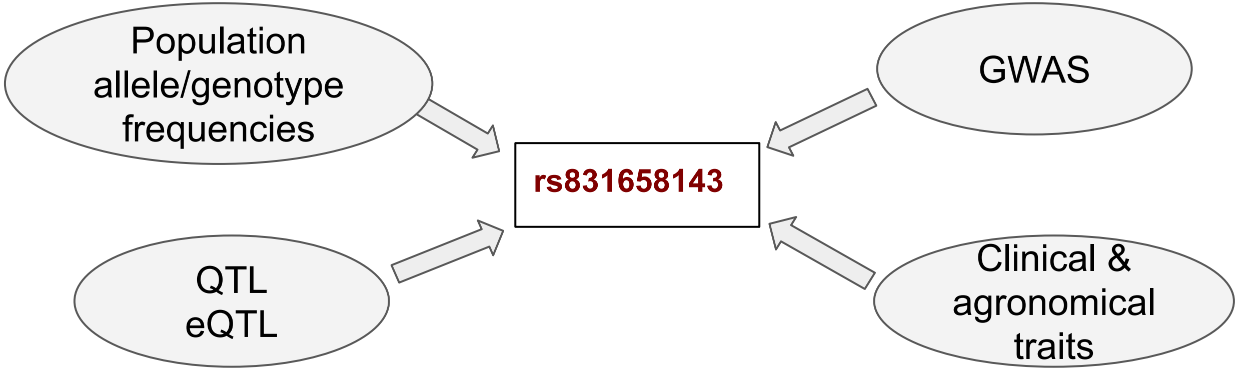
AgBioData SGV



What is an rsID?

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

Several data types aggregated around a marker



Population allele/genotype frequencies

GWAS

rs831658143

QTL eQTL

Clinical & agronomical traits

~ 1 M Publications linked to an rsID

Better Genomes, Bigger Integration Challenges

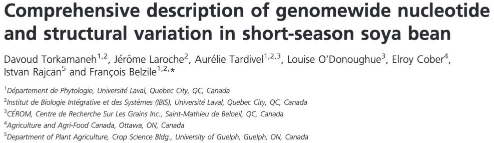
- **Long-read sequencing** → high-quality reference genomes
- **Pan-genomes** reveal:
  - Previously hidden variation
  - Reference-specific biases

Without stable identifiers:

- Variant IDs break across assemblies
- Data integration becomes harder

Promoting use of rsIDs – SoyBase

AgBioData SGV

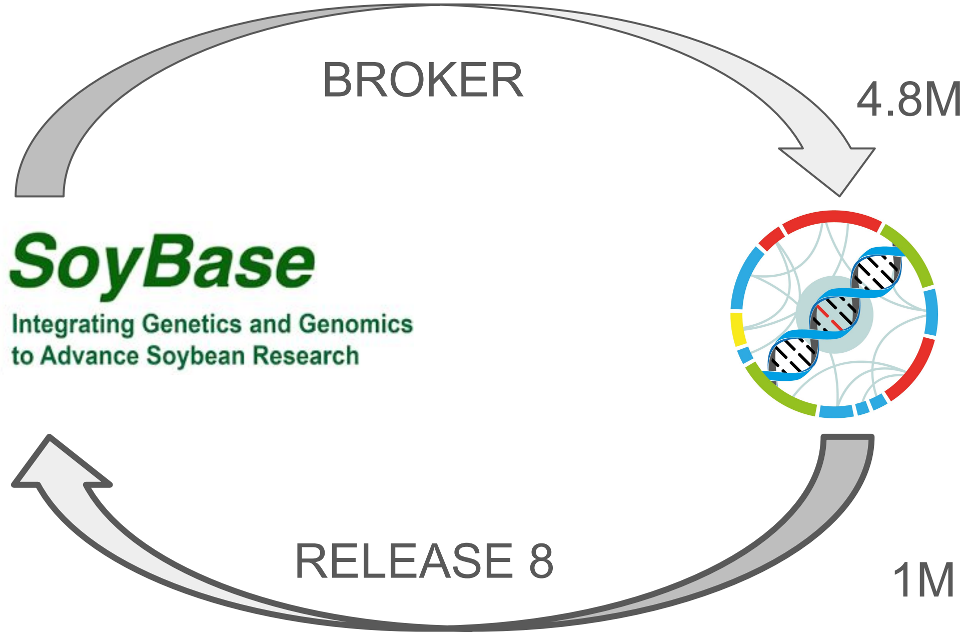


Comprehensive description of genomewide nucleotide and structural variation in short-season soya bean

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

Rex Nelson

***~90% concordance** across studies despite differences in technology and SNP density*

AgBioData SGV

Promoting use of rsIDs – Gramene / SorghumBase

Gramene PanGenome	Reference Crop	# rsIDs* (M)
SORGHUMBASE Genomic resources for the sorghum research community	Sorghum	46
MAIZE Gramene Comparative plant genomics focused on maize varieties	Maize	79
ORYZA Gramene Comparative plant genomics focused on rice varieties	Rice	68
GRAPEVINE Gramene Comparative plant genomics focused on grapevine varieties	Grape	0.3
Gramene	Arabidopsis	26

Sharon Wei

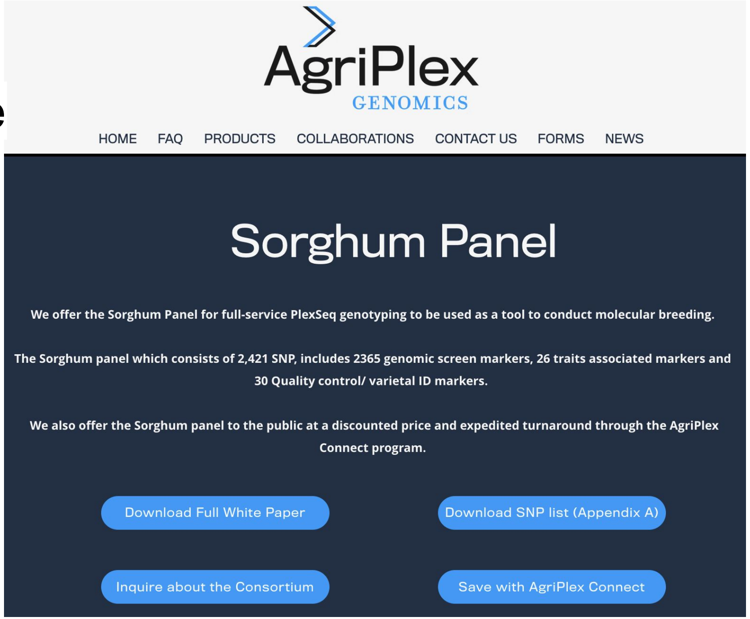
Marcela K. Tello-Ruiz

AgBioData SGV

Promoting use of rsIDs – Community DB / Industry

Collaborative development of community marker panels with rsIDs:

- Sorghum 2.4K SNPs (AgriPlex panel)
- EVA assigned rsIDs to 26 markers without one



AgriPlex GENOMICS

Sorghum Panel

We offer the Sorghum Panel for full-service PlexSeq genotyping to be used as a tool to conduct molecular breeding.

The Sorghum panel which consists of 2,421 SNP, includes 2365 genomic screen markers, 26 traits associated markers and 30 Quality control/ varietal ID markers.

We also offer the Sorghum panel to the public at a discounted price and expedited turnaround through the AgriPlex Connect program.

Download Full White Paper
Download SNP list (Appendix A)
Inquire about the Consortium
Save with AgriPlex Connect

Vivek Kumar

Marcela K. Tello

SORGHUMBASE.

European Variation Archive

Tim Cezard

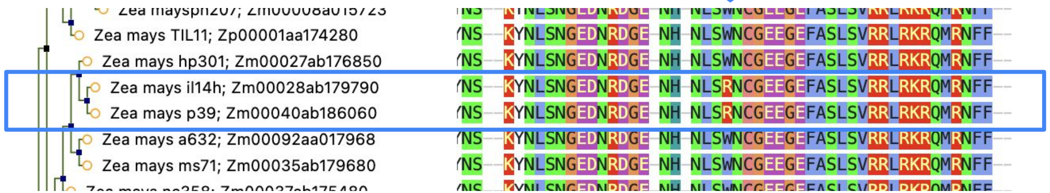
rsIDs provide stable identifiers for transversing Pan-Genomes

AgBioData SGV

rs831658143 (T/C)

Tgg (W) → Cgg (R)
starchy sweet

Sweet corn

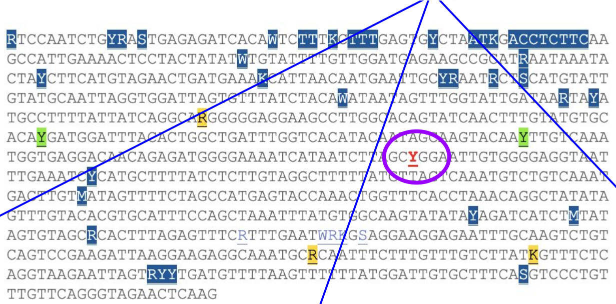


Zea maysph207; Zm00008a015723
Zea mays TIL11; Zp00001aa174280 — NS KYNLSNGEDNRDGE NH NLSWNCGEEGEFASLSVRRLRKRQMRNFF
Zea mays hp301; Zm00027ab176850 — NS KYNLSNGEDNRDGE NH NLSWNCGEEGEFASLSVRRLRKRQMRNFF
Zea mays il14h; Zm00028ab179790 — NS KYNLSNGEDNRDGE NH NLSRNCGEEGEFASLSVRRLRKRQMRNFF
Zea mays p39; Zm00040ab186060 — NS KYNLSNGEDNRDGE NH NLSRNCGEEGEFASLSVRRLRKRQMRNFF
Zea mays a632; Zm00092aa017968 — NS KYNLSNGEDNRDGE NH NLSWNCGEEGEFASLSVRRLRKRQMRNFF
Zea mays ms71; Zm00035ab179680 — NS KYNLSNGEDNRDGE NH NLSWNCGEEGEFASLSVRRLRKRQMRNFF
Zea mays oh43; Zm00037ab175480

MAF <1%

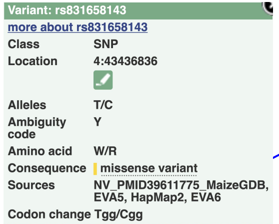
***Zea mays* B73 v5**
Sugary1 (Su1) gene

rs831658143 (T/C)



```
RTCCAATCTGYRASTGAGAGATCACAWTCTTTKCTTTGAGTGYCTAATKGACCTCTTCAA
GCCATTGAAAACTCCTACTATATWTGTATTTTGTTGGATGAGAAGCCGCATRAATAAATA
CTAYCTTCATGTAGAACTGATGAAAKCATTAACAATGAATTGCYRAATRCTSCATGTATT
GTATGCAATTAGGTGGATTAGTGTTTATCTACAWATAATAGTTTGGTATTGATAARTAYA
TGCCTTTTATTATCAGGCARGGGGGAGGAAGCCTTGGCACAGTATCAACTTTGTATGTGC
ACAYGATGGATTTACACTGGCTGATTTGGTCACATACAATAGCAAGTACAAYTTGTCAAA
TGGTGAGGACAACAGAGATGGGGAAAATCATAATCTTAGCYGGAATTGTGGGAGGTAAT
TTGAAATCYCATGCTTTTATCTCTTGTAGGCTTTTTATGTTAGTCAAATGTCTGTCAAAT
GACTTGTMATAGTTTTCTAGCCATGAGTACCAAACTGGTTTCACCTAAACAGGCTATATA
GTTTGTACACGTGCATTTCCAGCTAAATTTATGTGGCAAGTATATAYAGATCATCTMTAT
AGTGTAGCRCACTTTAGAGTTTCRTTTGAATWRKGSAGGAAGGAGAATTTGCAAGTCTGT
CAGTCCGAAGATTAAGGAAGAGGCAAATGCRCAATTTCTTTGTTTGTCTTATKGTTTCTC
AGGTAAGAATTAGTRYYTGATGTTTTAAGTTTTTTATGGATTGTGCTTTCASGTCCCTGT
TTGTTCAGGGTAGAACTCAAG
```

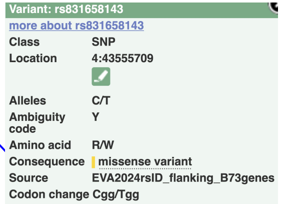
Starchy corn
(B73, HP301)



Variant: rs831658143

more about rs831658143

Class	SNP
Location	4:43436836
Alleles	T/C
Ambiguity code	Y
Amino acid	W/R
Consequence	missense variant
Sources	NV_PMID39611775_MaizeGDB, EVA5, HapMap2, EVA6
Codon change	Tgg/Cgg



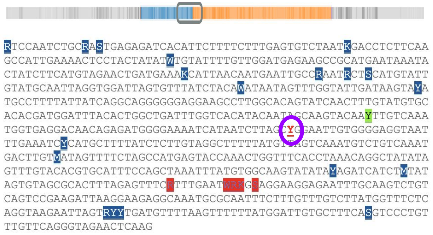
Variant: rs831658143

more about rs831658143

Class	SNP
Location	4:43555709
Alleles	C/T
Ambiguity code	Y
Amino acid	R/W
Consequence	missense variant
Source	EVA2024rsID_flanking_B73genes
Codon change	Cgg/Tgg

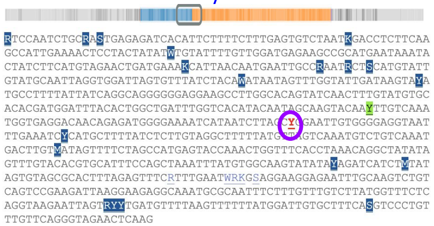
Sweet corn
(P39, IL14H)

HP301



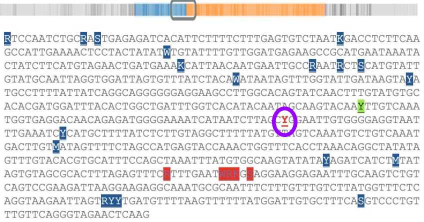
```
RTCCAATCTGCRASTGAGAGATCACATTCTTTTCTTTGAGTGTCTAATKGACCTCTTCAA
GCCATTGAAAACTCCTACTATATWTGTATTTTGTTGGATGAGAAGCCGCATGAATAAATA
CTATCTTCATGTAGAACTGATGAAAKCATTAACAATGAATTGCCRAATRCTSCATGTATT
GTATGCAATTAGGTGGATTAGTGTTTATCTACAWATAATAGTTTGGTATTGATAAGTAYA
TGCCTTTTATTATCAGGCAGGGGGGAGGAAGCCTTGGCACAGTATCAACTTTGTATGTGC
ACACGATGGATTTACACTGGCTGATTTGGTCACATACAATAGCAAGTACAAYTTGTCAAA
TGGTGAGGACAACAGAGATGGGGAAAATCATAATCTTAGCYGGAATTGTGGGAGGTAAT
TTGAAATCYCATGCTTTTATCTCTTGTAGGCTTTTTATGTTAGTCAAATGTCTGTCAAAT
GACTTGTMATAGTTTTCTAGCCATGAGTACCAAACTGGTTTCACCTAAACAGGCTATATA
GTTTGTACACGTGCATTTCCAGCTAAATTTATGTGGCAAGTATATAYAGATCATCTMTAT
AGTGTAGCGCACTTTAGAGTTTCRTTTGAATWRKGSAGGAAGGAGAATTTGCAAGTCTGT
CAGTCCGAAGATTAAGGAAGAGGCAAATGCGCAATTTCTTTGTTTGTCTTATGGTTTCTC
AGGTAAGAATTAGTRYYTGATGTTTTAAGTTTTTTATGGATTGTGCTTTCASGTCCCTGT
TTGTTCAGGGTAGAACTCAAG
```

P39



```
RTCCAATCTGCRASTGAGAGATCACATTCTTTTCTTTGAGTGTCTAATKGACCTCTTCAA
GCCATTGAAAACTCCTACTATATWTGTATTTTGTTGGATGAGAAGCCGCATGAATAAATA
CTATCTTCATGTAGAACTGATGAAAKCATTAACAATGAATTGCCRAATRCTSCATGTATT
GTATGCAATTAGGTGGATTAGTGTTTATCTACAWATAATAGTTTGGTATTGATAAGTAYA
TGCCTTTTATTATCAGGCAGGGGGGAGGAAGCCTTGGCACAGTATCAACTTTGTATGTGC
ACACGATGGATTTACACTGGCTGATTTGGTCACATACAATAGCAAGTACAAYTTGTCAAA
TGGTGAGGACAACAGAGATGGGGAAAATCATAATCTTAGCYGGAATTGTGGGAGGTAAT
TTGAAATCYCATGCTTTTATCTCTTGTAGGCTTTTTATGTTAGTCAAATGTCTGTCAAAT
GACTTGTMATAGTTTTCTAGCCATGAGTACCAAACTGGTTTCACCTAAACAGGCTATATA
GTTTGTACACGTGCATTTCCAGCTAAATTTATGTGGCAAGTATATAYAGATCATCTMTAT
AGTGTAGCGCACTTTAGAGTTTCRTTTGAATWRKGSAGGAAGGAGAATTTGCAAGTCTGT
CAGTCCGAAGATTAAGGAAGAGGCAAATGCGCAATTTCTTTGTTTGTCTTATGGTTTCTC
AGGTAAGAATTAGTRYYTGATGTTTTAAGTTTTTTATGGATTGTGCTTTCASGTCCCTGT
TTGTTCAGGGTAGAACTCAAG
```

IL14H



```
RTCCAATCTGCRASTGAGAGATCACATTCTTTTCTTTGAGTGTCTAATKGACCTCTTCAA
GCCATTGAAAACTCCTACTATATWTGTATTTTGTTGGATGAGAAGCCGCATGAATAAATA
CTATCTTCATGTAGAACTGATGAAAKCATTAACAATGAATTGCCRAATRCTSCATGTATT
GTATGCAATTAGGTGGATTAGTGTTTATCTACAWATAATAGTTTGGTATTGATAAGTAYA
TGCCTTTTATTATCAGGCAGGGGGGAGGAAGCCTTGGCACAGTATCAACTTTGTATGTGC
ACACGATGGATTTACACTGGCTGATTTGGTCACATACAATAGCAAGTACAAYTTGTCAAA
TGGTGAGGACAACAGAGATGGGGAAAATCATAATCTTAGCYGGAATTGTGGGAGGTAAT
TTGAAATCYCATGCTTTTATCTCTTGTAGGCTTTTTATGTTAGTCAAATGTCTGTCAAAT
GACTTGTMATAGTTTTCTAGCCATGAGTACCAAACTGGTTTCACCTAAACAGGCTATATA
GTTTGTACACGTGCATTTCCAGCTAAATTTATGTGGCAAGTATATAYAGATCATCTMTAT
AGTGTAGCGCACTTTAGAGTTTCRTTTGAATWRKGSAGGAAGGAGAATTTGCAAGTCTGT
CAGTCCGAAGATTAAGGAAGAGGCAAATGCGCAATTTCTTTGTTTGTCTTATGGTTTCTC
AGGTAAGAATTAGTRYYTGATGTTTTAAGTTTTTTATGGATTGTGCTTTCASGTCCCTGT
TTGTTCAGGGTAGAACTCAAG
```

Tello-Ruiz et al, *unpublished*

* W578R previously reported 3X

Data journey for GV datasets & recommended actions

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

INSDC

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

maizeGDB
MAIZE GENETICS AND GENOMICS DATABASE

Gramene

SORGHUM BASE

OMIA
Online Mendelian Inheritance in Animals
THE UNIVERSITY OF SYDNEY

Not FAIR

FAIR

A Tipping Point for Interoperable Agricultural Genomics

Anchoring variants to rsIDs unlocks integration with:

- Functional annotation tools
- GWAS, QTLs, and multi-omics data
- Breeding databases and germplasm repositories

IMPROVES

- Traceability
- Genotype-to-phenotype discovery
- Long-term reuse of genomic resources

What’s next?

- Continue to raise community awareness of the benefits of rsIDs
- More AgBioData community databases adopting rsIDs
- More commercial SNP array providers using rsIDs
- rsID mapping to more pangenomes

Thanks!