



AgBio Databases & FAIR Principles

Marcela Karey Tello-Ruiz, PhD

October 15th 2019



Data & Databases

- Proper **data management** is critical for scientific discovery and reproducibility
 - Acquiring, validating, storing, protecting, and processing *data* to ensure the accessibility, reliability, and timeliness of the *data* for its users
- **Databases** are physical entities to store and organize these data so that it can be easily accessed, *managed* & shared
- What are the **FAIR data principles** & why should you care?
- **Bioinformatics** resources for cereal crops



Structural & functional annotation via GO, SO (maize, rice) → **AgBase***

Ag ontologies → **AgroPortal***

Alfalfa Breeder's Toolbox

Arabidopsis Biological Resource Center

Animal QTLdb

AraPort

Bovine Genome Database

CassavaBase

Citrus Genome Database

Cool Season Food Legume Database

CottonGen

CyVerse*

Genome Database for Rosaceae

Genome Database for Vaccinium

Wheat, barley, oat → **GrainGenes**

Comparative genomics & pathways → **Gramene**

Germplasm resources → USDA-ARS **GRIN***

Hardwood Genomics

Hymenoptera Genome Database

i5K National Ag Library

Rice mutants → **KitBase**

Legume Information System

MaizeGDB

Maize Stock Center

MusaBase

National Animal Disease Center

PeanutBase

Plant ontologies → **Planteome***

PulseCrop

Solanaceae Genomics Network

SoyBase

SweetPotatoBase

Triticeae (wheat, barley) & oat → **T3**

TAIR

TreeGenes

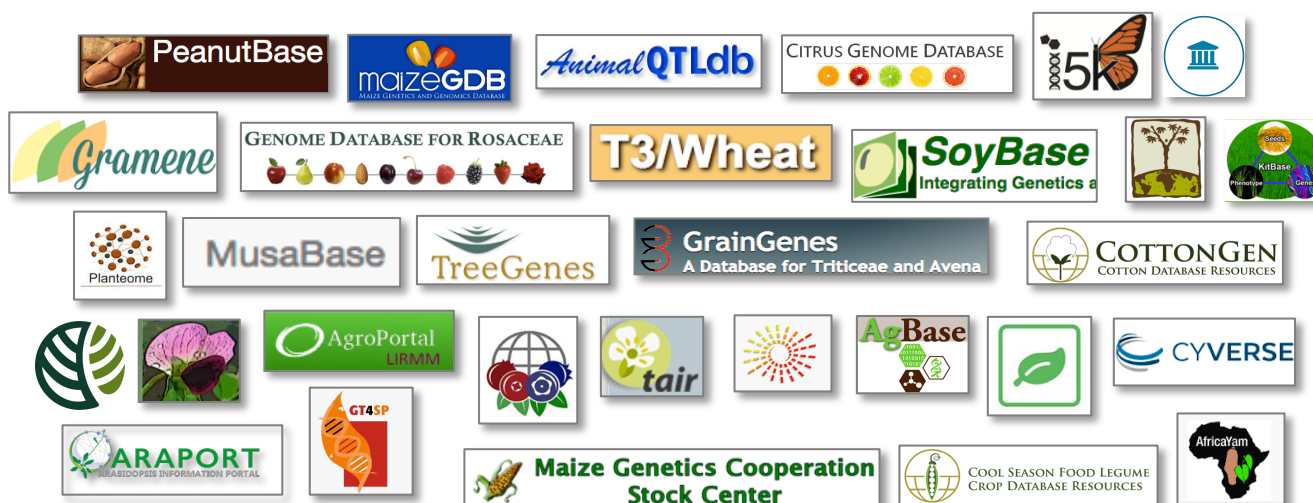
WheatIS

YamBase

The AgBioData Consortium

⇒ Improve handling of data in AgBio databases via collaboration, communication, and developing recommendations & standards.

Comprised of > 100 members from >30 databases



<https://www.agbiodata.org/databases>

Species-specific:

MaizeGDB

<https://www.maizegdb.org>

Pan-species:

PanMaize

<http://maize-pangenome-ensembl.gramene.org>

The screenshot shows the MaizeGDB website homepage. At the top, there is a navigation bar with links for Home, About, Community, Genome Browsers, Genomes, Tools, Data Centers, Search, and Feedback. Below the navigation bar, there is a search box and a 'Welcome to MaizeGDB!' message. The main content area is divided into several sections: 'Quick Links' with icons for Genome Browser, BLAST, qTeller, and MalzMine; 'Reference Assembly' with buttons for B73 ASSEMBLY, B73 ANNOTATION, and ALL GENOMES; 'Contribute data' with buttons for 'Contribute your data' and 'Make your data FAIR'; 'News' with several news items dated from 2019; 'Cooperation & Outreach' with social media icons; and 'Funding Sources' with logos for MNL, USAID, and the United States Department of Agriculture.

The screenshot shows the Gramene website homepage. At the top, there is a navigation bar with links for Tools, Help, Feedback, and Gramene Search. Below the navigation bar, there is a search box and a 'Login/Register' link. The main content area is divided into several sections: 'Preliminary NAM genome assembly and Gene Annotations' with a list of species and links; 'Favourite genomes' with a list of species and links; 'All genomes' with a dropdown menu; and 'Comparative analyses' with a paragraph of text.

Clade-specific:

GrainGenes

<https://wheat.pw.usda.gov/GG3>

Comparative multi-species:

Gramene

<http://www.gramene.org>

The screenshot shows the GrainGenes website homepage. At the top, there is a navigation bar with links for Home, GrainGenes Tools, Query Data Types, Resources, Collaborations, and About. Below the navigation bar, there is a search box and a 'Feedback' link. The main content area is divided into several sections: 'Search' with a dropdown menu; 'Community Services' with a list of services; 'Species Portals on GrainGenes' with a list of species; 'Upcoming Events' with a list of events; 'Quick Links' with icons for Browse GrainGenes, Genome Browsers, and Clap; 'Hot Topics' with a link to 'Wheat 10+ Genomes Project Release'; 'IWGSC RefSeq v2.0 now available'; 'GrainGenes-T3 Shared Genome Browsers'; and 'Durum Wheat (cv. Svevo) RefSeq Release 1.0 at GrainGenes'.

The screenshot shows the Gramene website homepage. At the top, there is a navigation bar with links for Genome Browser, Plant Reactome, Tools, BLAST, Track Hub Registry, Bulk Downloads, Plant Expression ATLAS, Gramene Mart, Outreach and Training, and Archive. Below the navigation bar, there is a search box and a 'Search for genes, species, pathways, ontology terms, domains...' link. The main content area is divided into several sections: 'Gramene Portals' with a list of portals; 'Genome Browser' with a link to 'Browse genomes with annotations, variation and comparative tools'; 'Plant Reactome' with a link to 'Browse and analyze metabolic and regulatory pathways'; 'Tools' with a link to 'Tools for processing both our data and yours'; 'BLAST' with a link to 'Query our genomes with a DNA or protein sequence'; 'Track Hub Registry' with a link to 'A global centralised collection of publicly accessible track hubs'; 'Bulk Downloads' with a link to 'FTP download of our data'; 'Plant Expression ATLAS' with a link to 'Browse plant expression results at EBI ATLAS'; 'Gramene Mart' with a link to 'An advanced query interface powered by BioMart'; 'Outreach and Training' with a link to 'Educational resources and webinars'; and 'Archive' with a link to 'Legacy tools and data (markers, Cyp pathways, etc)'.

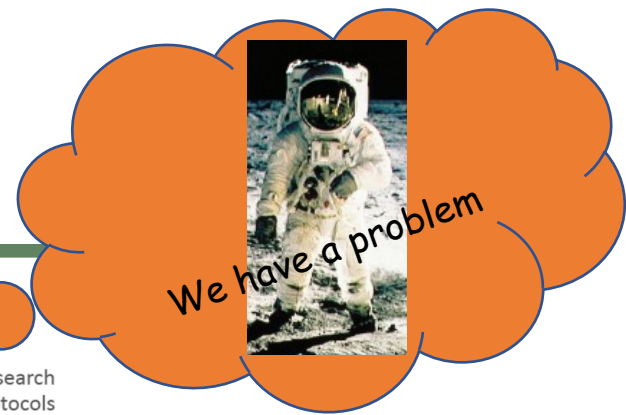
Maize
Sorghum
Rice (12)
Wheat (4)
Barley
Fox millet

agbiodata.org



AgBioData

Toward enhanced genomics, genetics, and breeding research outcomes through standardization of practices and protocols across agricultural databases



Home About Us - Databases - Working Groups - Meetings - Forums - Contact us

To cite AgBioData, please use: Harper, L. et al. (2018) [AgBioData consortium recommendations for sustainable genomics and genetics databases for agriculture](#). Database, 2018 1-32.

Welcome to AgBioData
AgBioData is a consortium of agricultural biological databases and associated resources working together to ensure standards and best practices for acquisition, display and retrieval of genomic, genetic and breeding data.

User Login

Username or e-mail *

Password *

[Create new account](#)
[Request new password](#)

Participating Databases and Resources

Gramene Search for genes, species, pathways, ontology terms, domains... 2076020 genes in 53 genomes

Gramene Portals

- Genome Browser**: Browse genomes with annotations, variation and comparative tools
- Tools**: Tools for processing both our data and yours
- BLAST**: Query our genomes with a DNA or protein sequence
- Track Hub Registry**: A global centralised collection of publicly accessible track hubs
- Bulk Downloads**: FTP download of our data
- Plant Reactome**: Browse and analyze metabolic and regulatory pathways
- Plant Expression ATLAS**: Browse plant expression results at EBI ATLAS
- Gramene Mart**: An advanced query interface powered by BiMart
- Outreach and Training**: Educational resources and webinars
- Archive**: Legacy tools and data (markers, Cyt pathways, etc)

Latest News

- Scholarships available for PI1 faculty - [Missouri Agricultural Experiment Station](#), Nov 21st, 2018
- Gramene will be at the Plant Biology Conference 2018 - [Come visit us at the AgBioData booth in Montreal!](#) Fri, 22 Jun 2018
- Scholarship DNA-interactive with third parties Thu, 14 Jun 2018
- Mining Maize with Gramene - [Free Webinar](#) May 22, 2018 @ 2 pm EDT Wed, 16 May 2018
- The Gramene Databases build 57 is out with a new snapshot genomes now! Thu, 03 May 2018
- Gramene Workshop at the 2018 Maize Genetics Meeting in Saint-Malo, France Tue, 06 Mar 2018
- Gramene webinar Feb 27, 2018: Plant Reaction pathway updates and new features Mon, 19 Feb 2018
- Gramene release # 56b: updates to Plant Reaction Thu, 01 Feb 2018
- The Gramene Databases build 56 is out with 8 new plant genomes!

News & Events

- [November AgBioData Conference Call](#)
Discussion about the future of agricultural GGB databases ...
Posted: 10/02/2019
- [October AgBioData Conference Call](#)
Join us for our October AgBioData conference ...
Posted: 09/12/2019
- [Recordings of conference calls available](#)
Don't worry if you missed a conference call ...
Posted: 09/05/2019
- [September AgBioData Conference Call](#)
Join us for our September AgBioData conference ...
Posted: 08/20/2019
- [August AgBioData Conference Call](#)
Join us for our August AgBioData conference ...
Posted: 08/03/2019
- [June AgBioData Conference Call](#)
Join us for our June AgBioData conference ...
Posted: 05/09/2019
- [May AgBioData Conference Call](#)
Join us for our April AgBioData conference ...
Posted: 04/07/2019

[More](#)

Tweets by @AgBioData

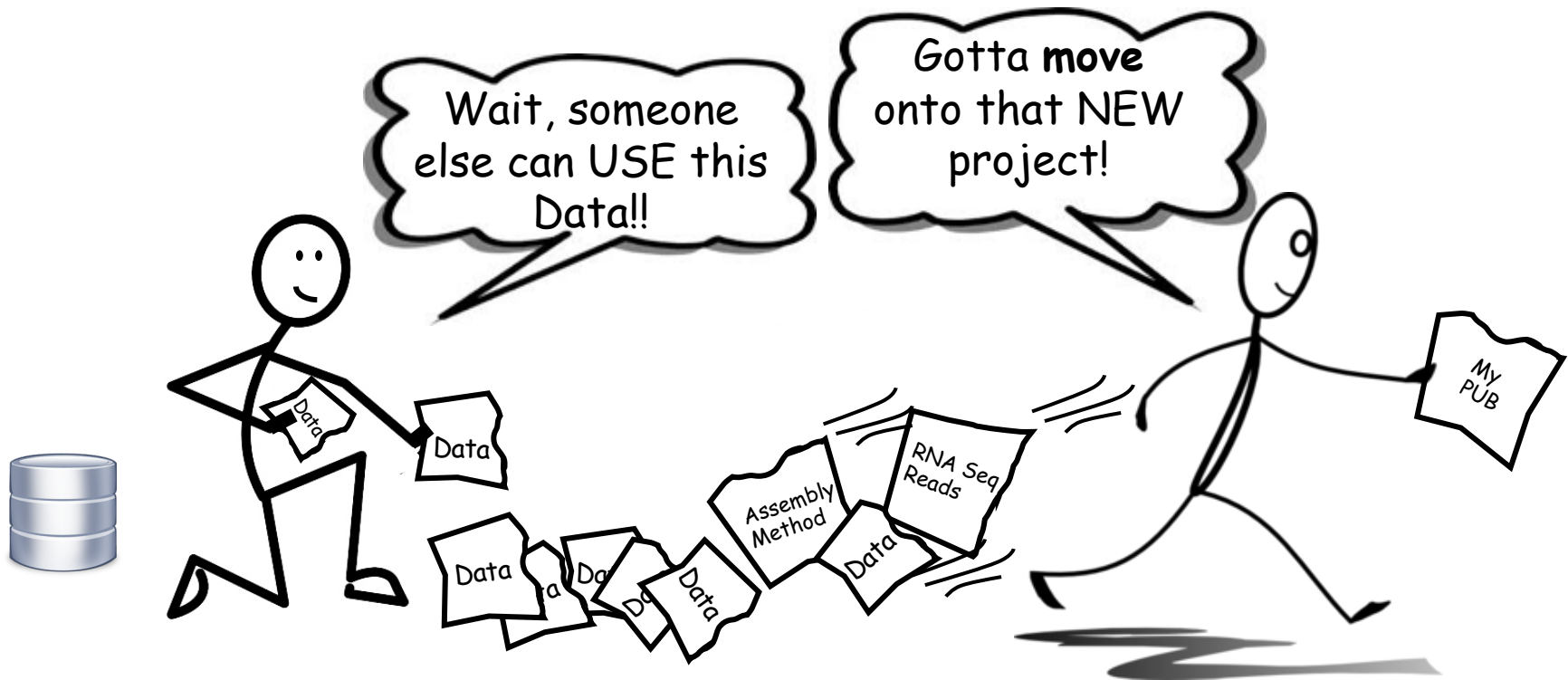
[#DataSharing](#) and [#datamanagement](#)
Snafuyoutu.be/66oNv_DJuPc

No properly described datasets whether deposited or supplemental, and the lead author may have ... returned to China ... and maybe named Sam Lee

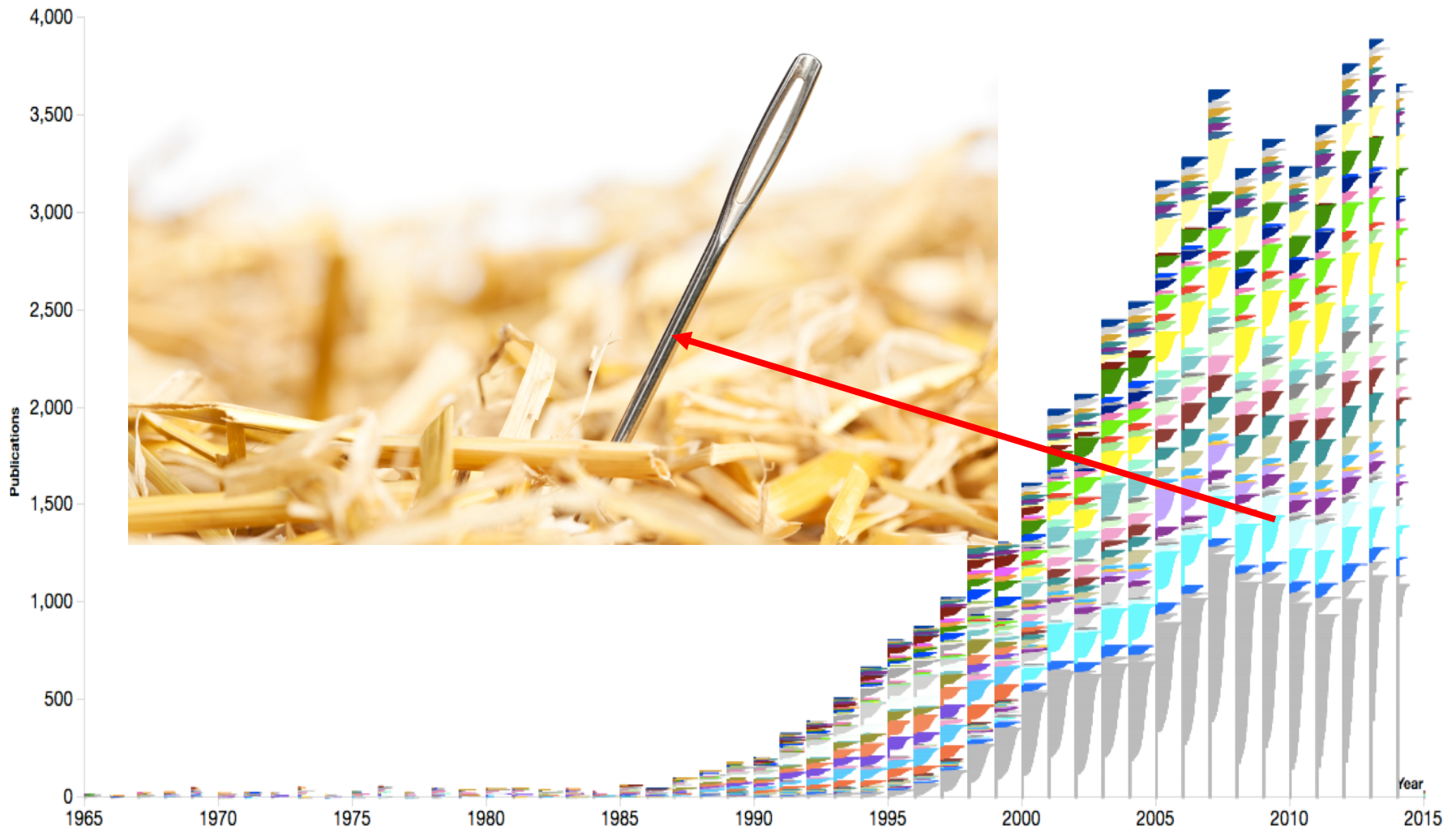
[YouTube](#) @YouTube

Data Sharing and Management
Snafu in 3 Short Acts

Databases TODAY...



Publications are increasing exponentially



<http://bar.utoronto.ca/50YearsOfArabidopsis/>

What can YOU do to make your data more **visible**, **persistent** and **reusable** to better **support research**?

Tips adapted from Lisa Harper (Maize Genetics Conference 2019)

Data LifeCycle

Making data

F indable,
A ccessible,
I nteroperable, and
R e-usable



Make your Data **FAIR**



Findable



Accessible



Interoperable



Reusable

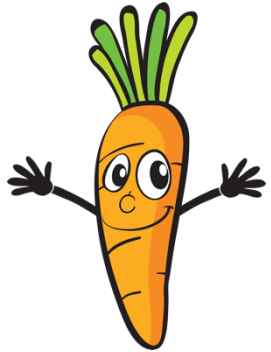
- **Findable** means data is human and machine readable and attached to persistent identifiers
- **Accessible** means data can be found and retrieved by humans and machines using standard formats
- **Interoperable** means data can be exchanged and used between systems.
- **Reusable** means data can be used by others

Wilkinson, et al., (2016) The FAIR Guiding Principles for scientific data management and stewardship.

<https://www.nature.com/articles/sdata201618>

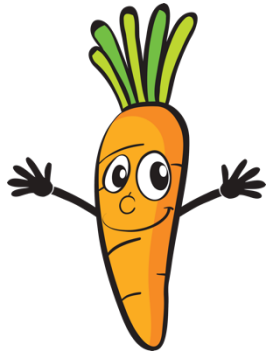
Credit: Melissa Haendel

What's in it for YOU?



More citations of YOUR work, increasing your visibility in the research community.

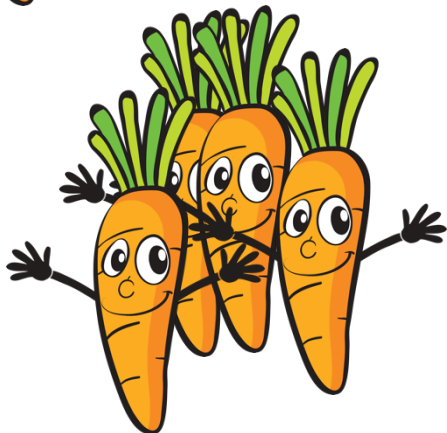
Easily comply with journal and funding requirements.



Less time spent fulfilling requests for data.



100% RECYCLABLE



We all benefit from data sharing --
higher rated Grants and Publications
(and hopefully promotion)

How to Make Your Published Data FAIR

- Use standard formats
- Supply complete **metadata**
- Embrace ontologies
- Use persistent and unambiguous identifiers
- Put your data in a long-term stable repository
- Cite, share freely and encourage others



Findable



Accessible



Interoperable



Reusable

1. Understand “Machine Readable”

This **IMAGE** of a table is NOT Machine Readable
(Published March 12, 2019 in PLOS one)

Species	Ensembl Plants Id's	Gene		cDNA		CDS	
		Length in bp (chr. no.)	Similarity (%)	Length in bp	Similarity (%)	Length in bp	Similarity (%)
Monocots							
<i>Z. mays</i>	Zm00001d043442_T001	3522 (3)	100	2191	100	1719	100
<i>T. aestivum</i> -A sub-genome	TraesCS3A02G274300	3732 (3)	65.98	2196	78.44	1683	87.19
- B sub-genome	TraesCS3B02G308000.1	3748 (3)	63.95	2210	77.82	1683	86.71
-D sub-genome	TraesCS3D02G273500.1	3718 (3)	65.42	2196	79.33	1683	86.77
<i>T. urartu</i>	TRIUR3_28465-T1	2088 (3)	75.81	1404	80.84	1404	80.84
<i>Ae. tauschii</i>	EMT16146	3172 (3)	72.06	1524	85.31	1524	87.53
<i>O. sativa</i>	OS01T0746400-00	3109 (1)	72	1710	87.66	1710	89.05
<i>B. distachyon</i>	BRAD12G49670.1	3577 (2)	68	2140	80.13	1713	86.23
<i>S. bicolor</i>	Sb03g034400.1	3545 (3)	75.91	1867	91.48	1740	94.76
Dicots							
<i>A. thaliana</i>	AT4G32810.1/Max4	3265 (4)	52.8	2026	56.59	1713	61.99
<i>G. max</i>	GLYMA06G09000.2	3891 (6)	52.72	2049	60.53	1692	65.88
<i>V. vinifera</i>	VIT_04s0008g03380.t01	2823 (4)	58.83	1782	64.65	1641	67.59
<i>S. lycopersicum</i>	Solyc08g066650.2.1	3075 (8)	52.36	1907	57.96	1674	61.56
<i>T. cacao</i>	EOY29749 (TCM_037195)	3680 (9)	52.04	2051	58	1680	63.75
<i>P. trichocarpa</i>	POPTR_0006s25490.1	3983 (6)	54.84	1674	63.79	1674	64.84
<i>P. persica</i>	EMJ23585 (PRUPE_ppa006042mg)	2893 (1)	56.11	2436	58.41	1296	67.28
<i>M. truncatula</i>	AES73861 (MTR_3g109610)	3439 (3)	52.31	2057	57.35	1698	63.28

<https://doi.org/10.1371/journal.pone.0213531.t001>

1. Understand “Machine Readable”. Using Standard Formats (SNP example)

SNP: A chromosome number and genome position, an alternative allele relative to the allele in the reference genome used, and the genotypes of the lines tested.

CHROM	POS	REF	ALT	Line 1	Line 2
Chr01					A
Chr03					C
Chr10					U

**Use the File format
STANDARD
for your data type**

**VCF: Variant Call Format
is the STANDARD**

CHROM	POS	REF	ALT	Line 1	Line 2
Gm01	12345	A	C	0/0	0/0
Gm03	67891	C	T	0/1	0/0
Gm10	23456	G	T	1/1	./.

Note: PDF images are not findable or accessible.
Use tables but beware of Excel (e.g., hidden characters)

1. Understand “Machine Readable”

This Bar Code and QR Code are Machine Readable
but NOT human readable



Wilkinson et al (2016).
FAIR Data Principles. *Nature*

2. Put Data in the right repository

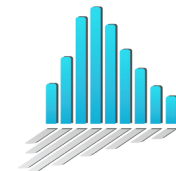
Data goes into Data repositories:

- Nucleotide sequences



Plant SNPs →
European
Variation Archive

- Protein/Proteomics/Metabolomics



- General Repositories



Get a DOI

- Not sure?



Contact YFD



3. Supply Complete Metadata

Metadata = Information about the data

Data (Phenotype):

Plant is 170 cm tall

Metadata:

Species = xxx

Germplasm = xxx

Age = xxx

Dev. Stage = xxx

Field location = xxx

Environment = xxx

Measurement

method = xxx

Data without Metadata is NOT FAIR

Use ontologies!

Same word, different meanings

Cell



Different words, same concept

Berenjena

Eggplant

Aubergine



Ontologies to the rescue!

Controlled Vocabulary
Hierarchy of terms & explicit relationships
among them, understood by computers

3. Supply Complete Metadata

Metadata = Information about the data

- Supply enough Metadata so that your experiment can be accurately reproduced
- Use community Standards for your data type. Examples:
 - MIxS: Minimal Information about any Sequence
 - MIAPPE: Minimum Information About a Plant Phenotyping Experiment
- Use Ontologies (GO, PO, TO, EO)
- Don't Know? Ask us!



MIxS



MIAPPE

Budget TIME to provide Metadata

5. Check Standards Set Forth by Species Communities – Let Curators Know About Your Work

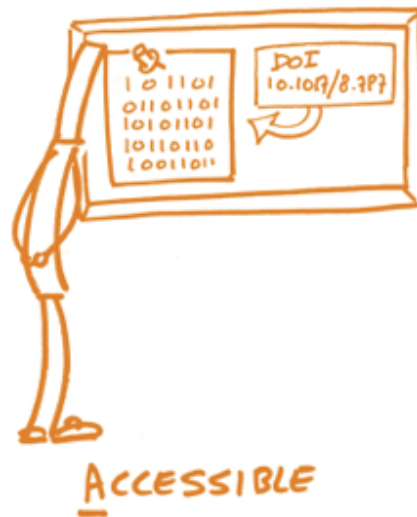
- ▶ Don't RE-name genes! Check out gene nomenclature standards for your species
- ▶ If you have published on a gene or genes – Let Curators Know
- ▶ If you have a dataset that will be useful to others – Let your Community DB Know

The screenshot shows the MaizeGDB website. At the top left is the MaizeGDB logo with the tagline 'MAIZE GENETICS AND GENOMICS DATABASE'. Below the logo is a navigation bar with 'Users', 'Genomes', 'Tools', and 'Data Centers'. A search bar contains 'all data'. The main content area features a 'Reference Assembly' section with three icons: 'B73 ASSEMBLY', 'B73 ANNOTATION', and 'ALL GENOMES'. Below these are buttons for 'Report assembly error' and 'Report gene model error', and a link for 'B73 RefGen_v4 is now available'. A red arrow points from the 'Report assembly error' button to the 'qTeller' logo. Other logos for 'qTeller', 'MaizeMine', and a news item dated '11, March 2019' are also visible.

The screenshot shows the Plant Reactome website. At the top is the Plant Reactome logo with the tagline 'Gramene Pathways'. Below the logo is a navigation bar with 'About', 'Content', 'Docs', 'Tools', 'Community', and 'Download'. The main content area features a search bar with the text 'Find Reactions, Proteins and Pathways' and a search input field containing 'e.g. YUC4, cytokinin, jasmonic'. A 'Go!' button is next to the search input. Below the search bar are three large green icons: a tree structure for 'Pathway Browser', a bar chart for 'Analyze Data', and a document icon for 'Documentation'. A 'Contact Us' button is in the bottom right corner.

The FAIR Data Principles

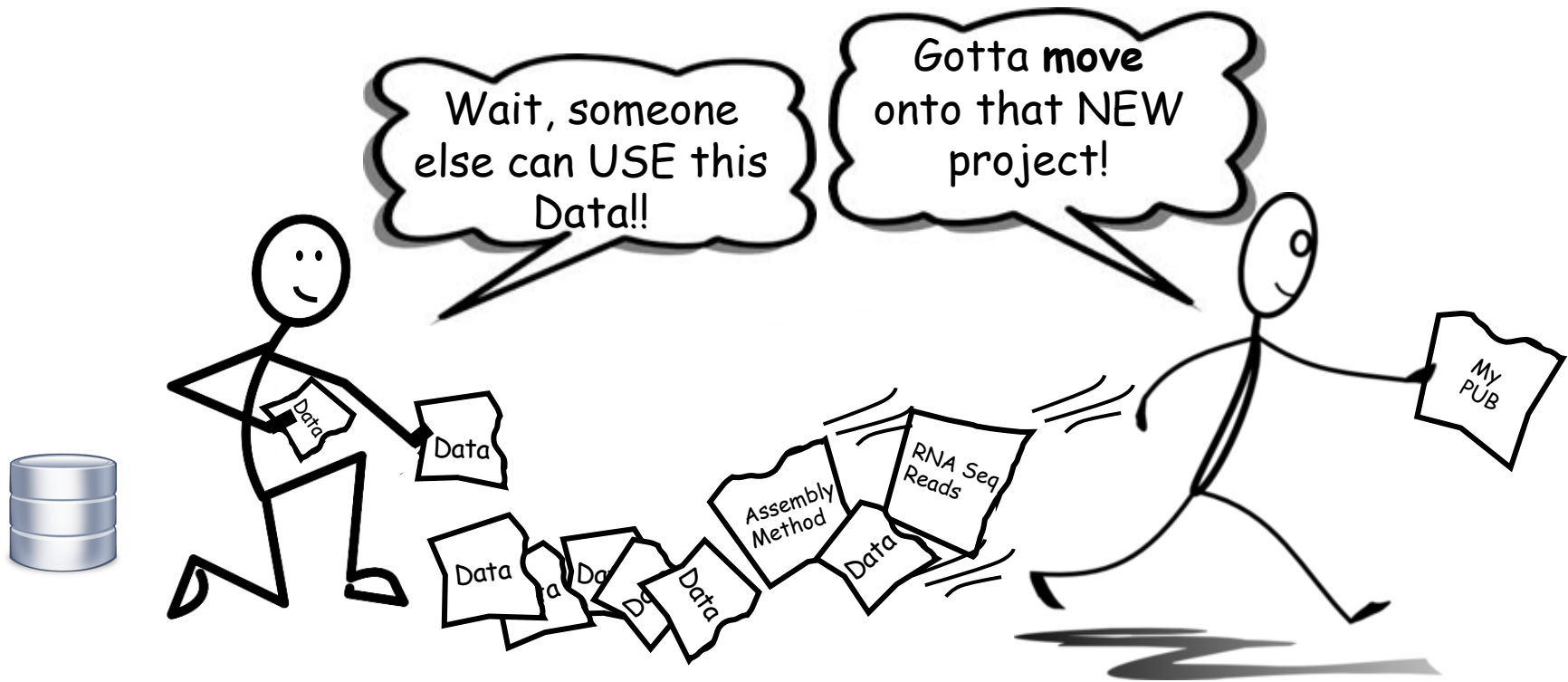
FAIR DATA PRINCIPLES



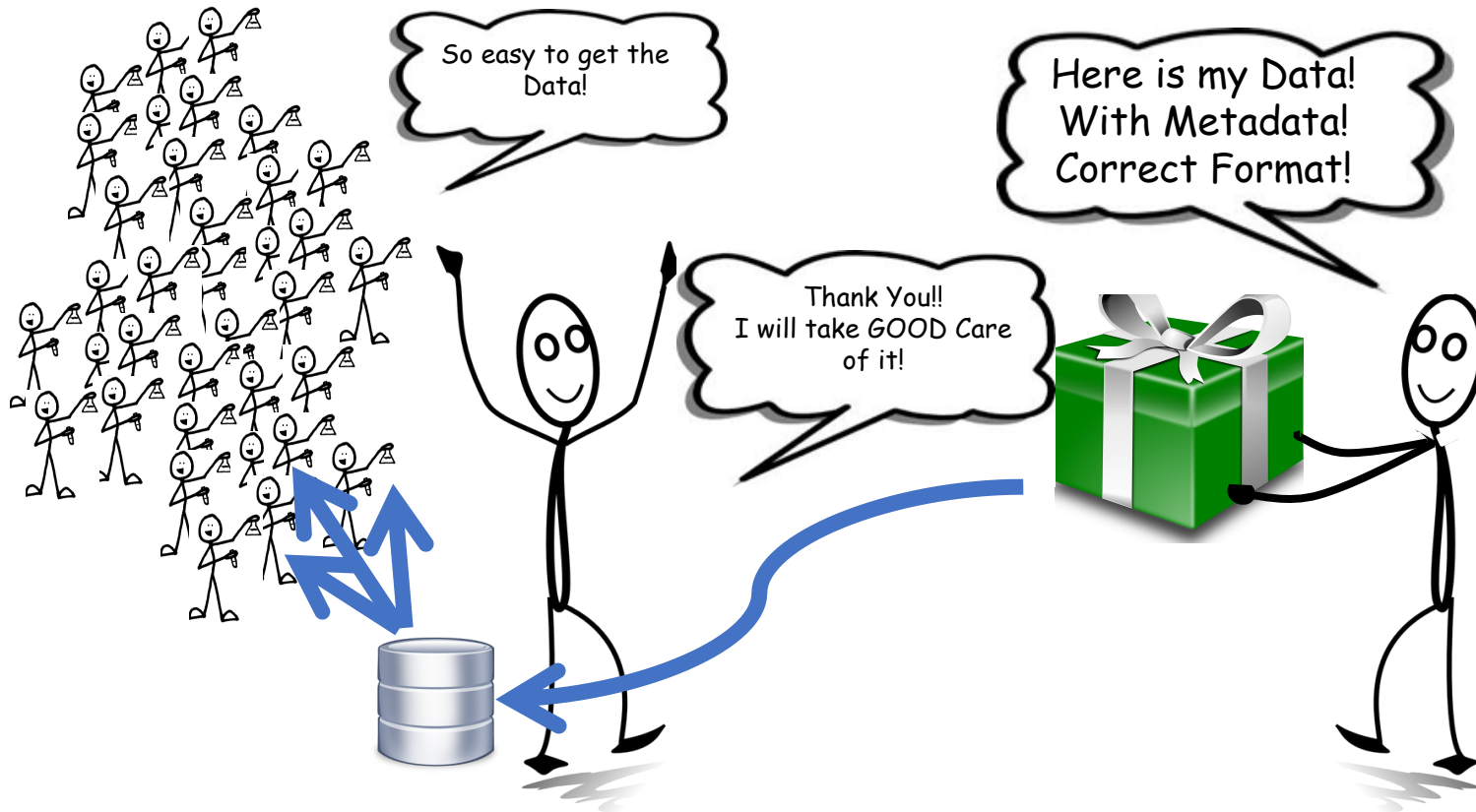
Good Data Stewardship

- Publish Data with the paper
- Describe Data to your fullest ability
- Use the right words to identify Data
- Deposit data in the right Data Repository
- Budget time for Data Management
- Don't think of it as YOUR data

Databases TODAY...



We Hope... Databases in FUTURE





Good data practices benefit everyone (and help you get funded)



The **Data Management Plan** is an integral part of an NSF grant proposal, which NSF will consider under Intellectual Merit or Broader Impacts.

The plan describes how the proposal will conform to NSF policy on the dissemination and sharing of research results, and may include:

- the types of **data**, samples, software, curriculum materials, etc.;
- the standards to be used for **data and metadata** format and content;
- policies for **access and sharing**;
- policies and provisions for **re-use, re-distribution**, & production of **derivatives**;
- plans for **archiving data**, samples, etc., and preserved access.

Making your Data **FAIR**



Findable



Accessible



Interoperable



Reusable

Plantae Webinar by Lisa Harper & Leonore Reiser:

Data Management 101 - Tips for making your published data more Findable, Accessible, Interoperable and Reusable

<https://community.plantae.org/video/4992321984424576936/plantae-seminar-data-management-101-tips-for-making-your-published-data-more-findable-accessible-interoperable-and-reusable-with-lisa-harper-and-leonore-reiser>

Thanks!



Lisa Harper



Leonore Reiser



Marcela K. Tello-Ruiz