Agricultural sciences in the big data era: Genotype and Phenotype Data Standardization, Utilization and Integration

Cecilia H. Deng, Sushma Naithani, Sunita Kumari, Irene Cobo-Simón, Elsa H. Quezada-Rodríguez, Maria Skrabisova, Nick Gladman, Melanie J. Correll, Akeem Babatunde Sikiru, Olusola O Afuwape, Annarita Marrano, Ines Rebollo, Wentao Zhang, and **Sook Jung**

Genotype-Phenotype Working Group

Genotype and Phenotype Working Group

Sushma Naithani
Oregon State Univ.
Corvallis, OR, USA

Sook Jung,
Washington State Univ.
Pullman, WA, USA

Sunita Kumari
Cold Spring Harbor Laboratory, NY, USA

Elsa H Quezada
UNAM, Tizayuca
Mexico

Irene Cobo-Simón
Univ. of Connecticut,
Storrs, CT, USA

Nicholas Gladman,
Cold Spring Harbor Laboratory, NY, USA

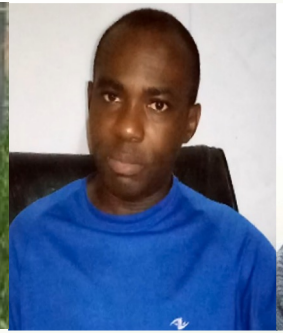
Annarita Marrano,
Phoenix Bioinformatics, USA

Melanie Correll
Univ. of Florida,
Gainesville, FL, USA

Maria Skrabisova
Palacký Univ.
Olomouc, Czech Republic

Cecilia H. Deng
The New Zealand Institute for Plant & Food Research Limited, New Zealand

Wentao Zhang
National Research Council Canada,



Olusola Afuwape
Univ. of Lagos,
Lagos, Nigeria

Akeem Sikiru
Federal Univ. of Agri. Zuru, Nigeria

Ines Rebollo
Univ. de la República, Uruguay

Overview of the presentation

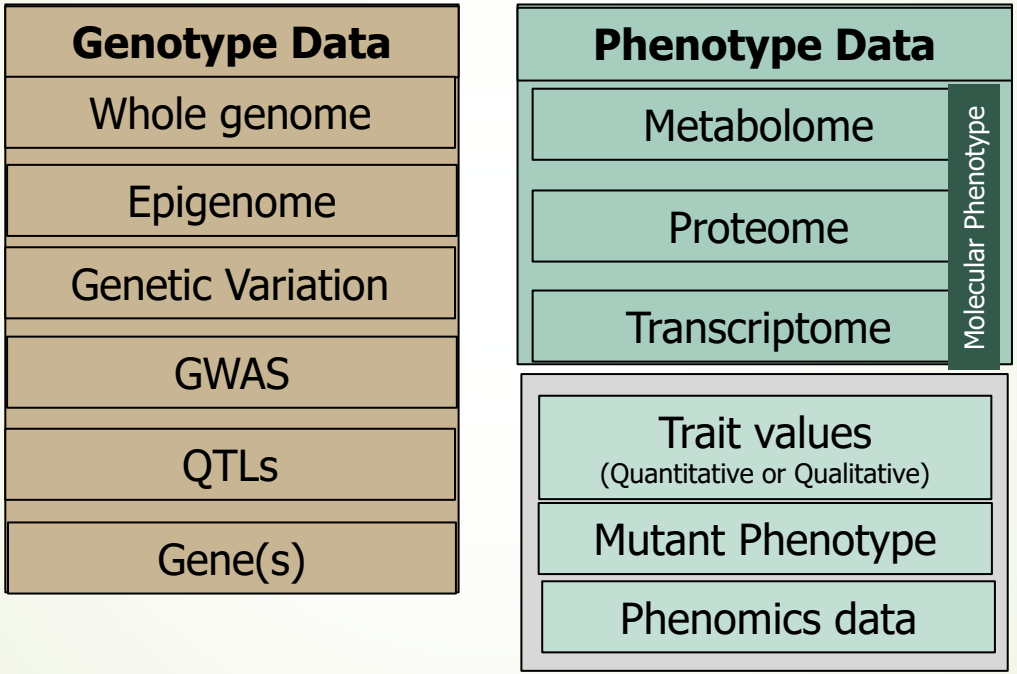
- Aims of the Genotype to Phenotype Working Group
- Accomplishment
  - What and How
- Recommendation
- Challenges our WG faced
- What went well in our WG

The Genotype-Phenotype Working Group Goals and Aims

- To improve the data collection and data sharing
- To facilitate linking genotype and phenotype data
- To promote data interoperability and re-use

✓Such a big goal!!!

What Types of Data are we even talking about?



So, what can we do?

- Find out the current status
  - Where the data is stored
  - What data and metadata are kept
  - How are they currently integrated
  - How can they be re-used
  - What are the limitation
  - Then we could come up with recommendations!
- This could be a white paper!

Accomplishment: paper submitted to Oxford Database!

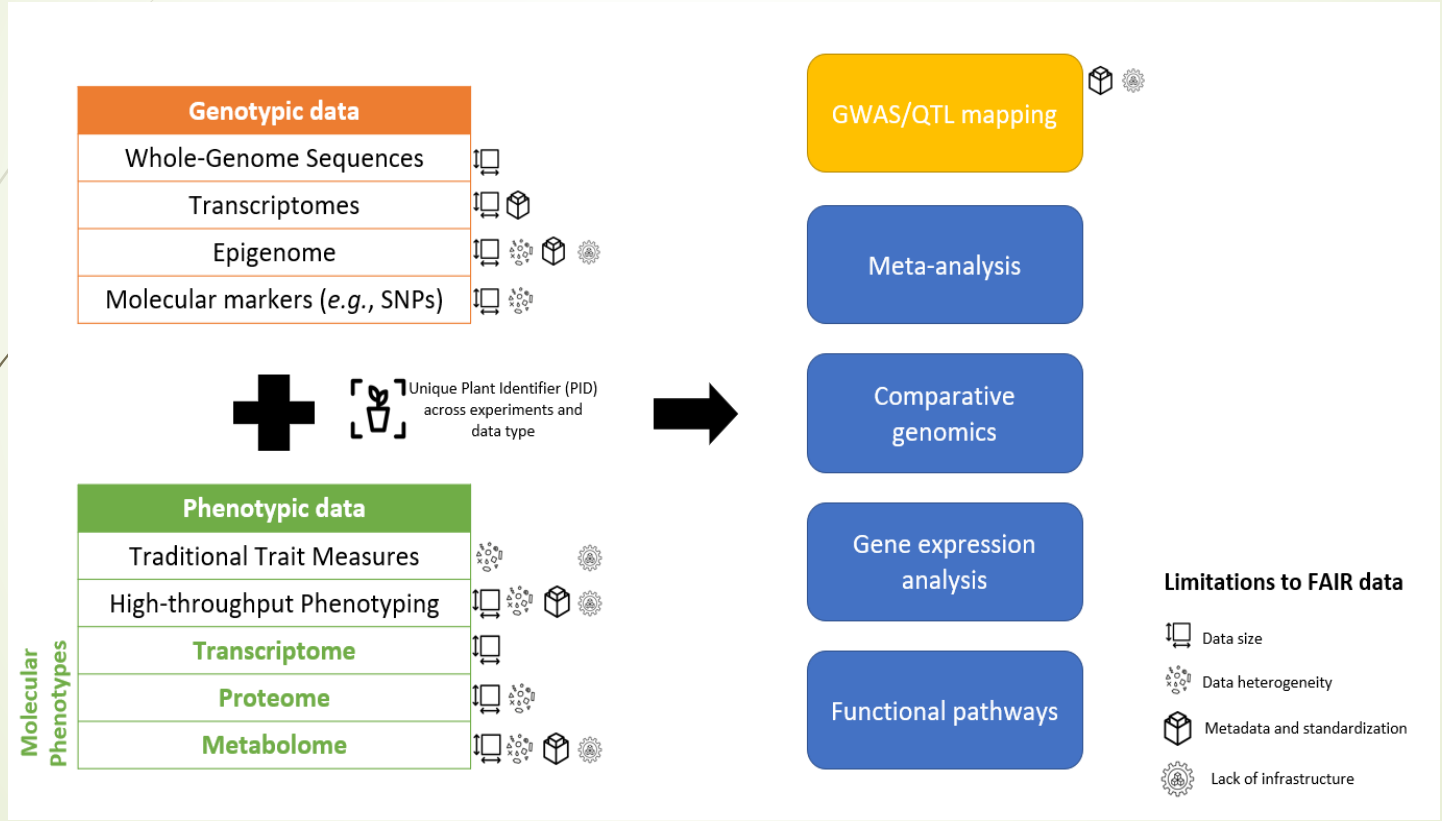
Agricultural sciences in the big data era: Genotype and Phenotype Data Standardization, Utilization and Integration

Cecilia H. Deng[1#*], Sushma Naithani[2#], Sunita Kumari[3#], Irene Cobo-Simón[4], Elsa H. Quezada-Rodríguez[5,6], Maria Skrabisova[7], Nick Gladman[3,8], Melanie J. Correll[9], Akeem Babatunde Sikiru[10], Olusola O Afuwape[11], Annarita Marrano[12], Ines Rebollo[13], Wentao Zhang[14], and Sook Jung[15#*]
On behalf of the Genotype-Phenotype Working Group, AgBioData

[1]The New Zealand Institute for Plant and Food Research Limited, Auckland, New Zealand
[2]Department of Botany and Plant Pathology, Oregon State University, Corvallis, Oregon, USA
[3]Cold Spring Harbor Laboratory, 1 Bungtown Rd, Cold Spring Harbor, New York, USA
[4] Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, USA
[5]Departamento de Producción Agrícola y Animal, Universidad Autónoma Metropolitana-Xochimilco, Ciudad de México, México
[6]Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, Ciudad de México, México.
[7]Department of Biochemistry, Faculty of Science, Palacky University, Olomouc, Czech Republic
[8]U.S. Department of Agriculture-Agricultural Research Service, NEA Robert W. Holley Center for Agriculture and Health, Cornell University, Ithaca, New York, USA.

Submitted in April 2023

Potential integrative analysis using genotype and phenotype data



Strategy for the white paper

- Introduction
- Main Content
- Conclusion

- Data type
  - Whole genome/transcriptome
  - Genetic variation
  - Phenotype (trait variation)
  - Phenomic
  - Proteomic
  - Metabolomic
  - GWAS/QTL

- For each data type
  - Brief Introduction of the data type and method
  - Where are the data submitted?
  - What metadata and data are submitted and are there minimum standards?
  - What type of studies reuse these data?
  - What is the limitation for the re-use?

Final Structure of the paper

1. Introduction
2. Genomics and Transcriptomics data
3. Phenotypes and Phenomics
4. Association mapping (GWAS) and linkage mapping (QTL)
5. Data reusability limitations and challenges
6. Recommendations

Genomics and Transcriptomics data

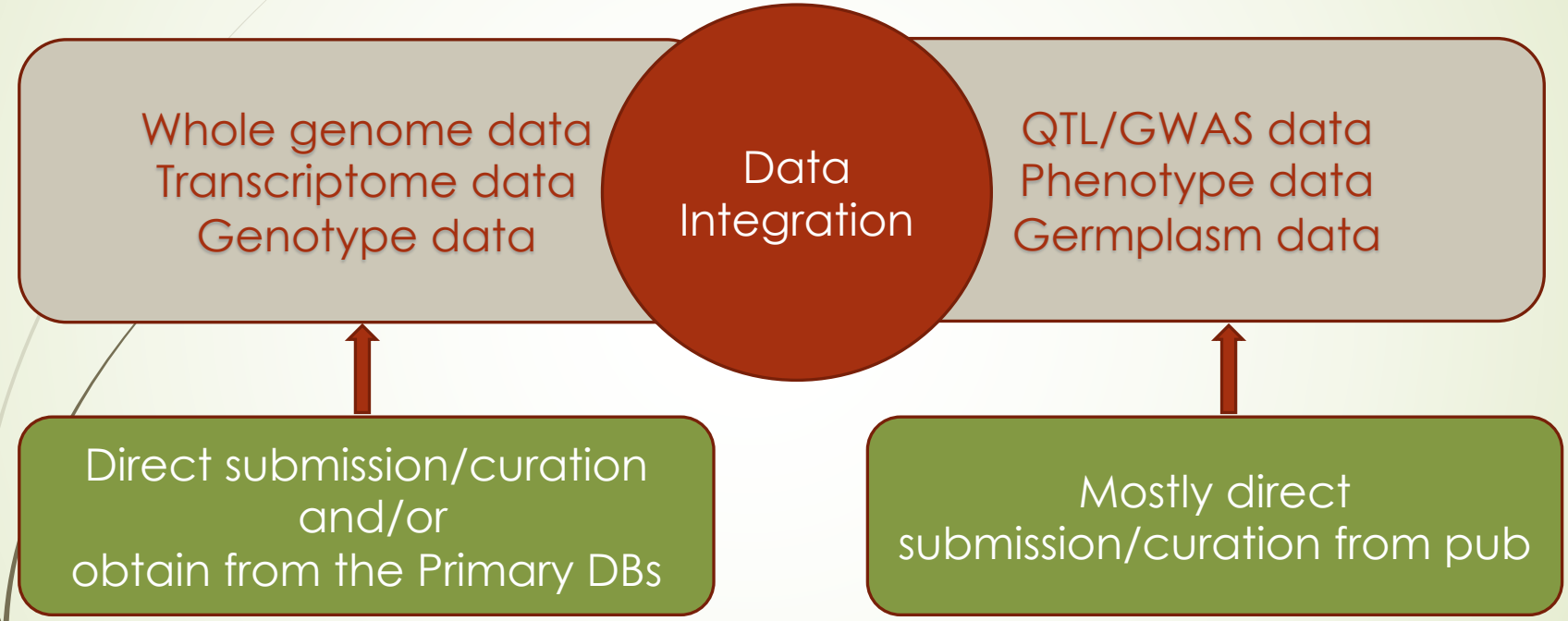
- Introduction on technologies
  - Whole genome and transcriptome sequencing
  - Genotyping
- Public repositories and metadata requirements
  - Primary Databases
  - Crop/clad Community GGB Databases
- Uses and Applications

Database name	NCBI	DRA	ENA	GSA	IBDC	AGDR †	DRYAD ¥	Zenodo ‡, ¥	FigShare
Genome sequence data	+	+	+	+	+	+	+	+	+
WGS annotations	+	?	?	?	?	?	?	?	+
Genotyping data	+	?	?	?	?	?	?	?	+
Transcriptome sequence data	+	+	+	?	?	?	+	+	+
fq.gz	+	+	+	+	+	+	+	+	+
BAM	+	+	+	+	+	+	+	+	+
SFF	+	+	+	+	+	-	+	+	+
HDF	+	+	+	+	+	-	+	+	+
VCF	+	+	+	?	?	?	+	+	+
INSDC-Source	+	+	+	a	b	c	d	e	f

Table 1. A list of public repositories for genomic, genotyping and transcriptome data that are active, maintained and updated.

Supplementary Table 1: with metadata description
Supplementary Table 2: A list of sequence specific data resources.

Crop community GGB databases



Species/Crop	Database	Database URL
Arabidopsis	TAIR	https://www.arabidopsis.org/
Cassava	CassavaBase	https://www.cassavabase.org/
Citrus	Citrus Genome Database	https://www.citrusgenomedb.org/
Citrus / Diaphorina citri/ Ca. Liberibacter asiaticus	Citrus Greening	https://www.citrusgreening.org/
Cotton	CottonGen	https://www.cottongen.org/
Cucurbit	Cucurbit Genomics	http://cucurbitgenomics.org/
Forest trees	TreeGenes	https://treegenesdb.org
	Hardwood Genomics	http://www.hardwoodgenomics.org/
Grains	GrainGenes	https://wheat.pw.usda.gov
	Gramene	https://www.gramene.org/
	SorghumBase	https://www.sorghumbase.org/
	Triticeae toolbox, T3	https://wheat.triticetoolbox.org/
	WheatIS	https://wheatis.org
	KitBase	http://kitbase.ucdavis.edu/
Legumes	KnowPulse	https://knowpulse.usask.ca/
	Legume Information System	https://www.legumeinfo.org/
	PeanutBase	https://peanutbase.org
Pulses	Pulse Crop Database	https://www.pulsedb.org/
	Soybase	https://www.soybase.org/
Maize	MaizeGDB	https://maizegdb.org/
Musa	MusaBase	https://www.musabase.org/
Rosaceae	Genome Database for Rosaceae	https://www.rosaceae.org/
Solanaceae	Sol Genomics	https://solgenomics.net/
Sweet Potato	SweetPotatoBase	https://www.sweetpotatobase.org/
Vaccinium	Genome Database for Vaccinium	https://www.vaccinium.org/
Yam	YamBase	https://www.yambase.org/

Table 2. List of Crop/clad Community GGB Databases that **integrate various types of data including whole genome data, genotype, phenotype, QTL, GWAS, and germplasm data**.

Supplementary Table 3. A list of public crop community databases with **data types, metadata, submission format, and URL for data submission**

Comparative genomic database used by multiple communities		
A comparative genomic database for ~300 plant species	Phytozome	https://phytozome-next.jgi.doe.gov/
A comparative genomic database hosting 118 genomes from models, crops, fruits, vegetables, etc.	Gramene	https://www.gramene.org/
Others	AgBase	https://agbase.arizona.edu/
	Bio-Analytic Resource	https://bar.utoronto.ca/

Phenotypes and Phenomics

- Data types, Repositories, and Knowledge Bases
  - Phenotype
    - Trait variation data
    - Phenomics data
    - Proteomics data
    - Metabolomics data
  - Association mapping (GWAS) and linkage mapping (QTL)
- Phenotype data formats, standards and metadata

Category	Databases	URLs
Species-specific mutant collections	Database of image and genome (MaizeDIG)	https://maizedig.maizegdb.org/
	Mutant Variety Database	https://nucleus.iaea.org/sites/mvd/SitePages/Home.aspx
	Plant Genome Editing Database	http://plantcrispr.org/cgi-bin/crispr/index.cgi
	RIKEN Arabidopsis Genome Encyclopedia (RARGE)	http://rarge-v2.psc.riken.jp/line
	TOMATOMA	https://tomatoma.nbrp.jp/index.jsp
	Plant Editosome	https://ngdc.cncb.ac.cn/ped/
Traits and QTL	Gramene QTL	https://archive.gramene.org/qtl/
	Wheatqtl	http://www.wheatqtldb.net/
	GLOPNET	http://bio.mq.edu.au/~iwright/glopian.htm
	TRY database	https://www.try-db.org/TryWeb/Home.php
	Ecological Flora of the Britain and Ireland	http://ecoflora.org.uk/
	BIOPOP	http://www.landeco.uni-oldenburg.de/Projects/biopop/main.htm
	FloraWeb	https://www.floraweb.de/
	USDA GRIN	https://www.ars-grin.gov/
	BiolFlor	https://wiki.ufz.de/biolflor/index.jsp
	LEDA	https://uol.de/en/landeco/research/leda
	USDA PLANTS	https://plants.usda.gov/home
	BROT	https://www.uv.es/jgpausas/brot.htm
	AusTraits	https://austraits.org/
	Community Databases in Table 2 and Supplementary Table 3	
Phenomics	GnpIS	https://urgi.versailles.inra.fr/gnpis
	PGP Repository	https://edal-pgp.ipk-gatersleben.de/
	Cartograplant	https://cartograplant.org/
	AgData commons Plants & Crops:	https://data.nal.usda.gov/ag-data-commons-hierarchy/plants-crops
	PathoPlant	http://www.pathoplant.de/
	PncStress	http://bis.zju.edu.cn/pncstress/
	Indian Crop Phenome DB (ICPD)	https://jbdc.rcb.res.in/icpd/
Gene Expression	Ozone Stress Responsive Gene Database	https://www.osrgd.com
	EBI-Plant Expression Atlas	https://www.ebi.ac.uk/gxa/plant/experiments
	CoNeKT	https://conekt.sbs.ntu.edu.sg/
Protein, peptides and proteomes	Expath	http://expath.itps.ncku.edu.tw/
	Proteome Xchange	https://www.proteomexchange.org
	Plant Proteome Database	http://ppdb.tc.cornell.edu/
	PlantMWpIDB	https://plantmwpidb.com/
	Heat Shock Proteins database	http://hsfdb.bio2db.com/
	WallProtDB	https://www.polebio.lrsv.ups-tlse.fr/WallProtDB/
	Aramemnon	http://aramemnon.botanik.uni-koeln.de/
	PhosPhAt	https://phosphat.uni-hohenheim.de/db.html
	Database of Phospho-sites in Plants	http://dbppt.biocuckoo.org/browse.php
	Plant Protein Phosphorylation Database	https://www.p3db.org/home
	qPTMplants	http://qptmplants.omicsbio.info/
	Plant PTM viewer	https://www.psb.ugent.be/webtools/ptm-viewer/
	PlaPPISite	http://zzdlab.com/plappisite/index.ph
	M. truncatula Small Secreted Peptide Database	https://mtsspdb.zhaolab.org/database
	PlantPepDB	http://14.139.61.8/PlantPepDB/index.php
	Arabidopsis PeptideAtlas	http://www.peptideatlas.org/builds/arabidopsis/
	Indian Structural Data Archive	https://isda.rcb.ac.in/
Metabolites, biochemical, and small chemical entities	Antimicrobial plant peptides (PhytAMP)	http://phytamp.pfba-lab-tun.org/main.php
	PubChem	https://pubchem.ncbi.nlm.nih.gov
	ChEBI	https://www.ebi.ac.uk/chebi
	Metabolomics Workbench	https://www.metabolomicsworkbench.org
Secondary Knowledgebase	MetaboLights	https://www.ebi.ac.uk/metabolights/index
	PoDP	https://pairedomicsdata.bioinformatics.nl/
	Plant Reactome pathway knowledgebase	https://plantreactome.gramene.org
	MetaCyc	https://metacyc.org
	PMN	https://plantcyc.org/data
	KEGG pathways	https://www.genome.jp/kegg/pathway.html
	PlantPathMarks (PPMdb)	http://ppmdb.easyomics.org/
	The Bio-Analytic Resource (BAR)	https://bar.utoronto.ca
	The protein-protein interaction database for Maize (PPIM)	https://mai.fudan.edu.cn/ppim/

Table 3. List of public repositories, databases and secondary knowledgebases host or integrate various types of **phenotypes, phenomics and molecular phenotype data**.

- Species specific mutant collections
- Traits and QTL
- Phenomics
- Gene expression
- Proteins, peptides, and proteomes
- Metabolites, biochemical, and small chemical entities
- Secondary Knowledgebase

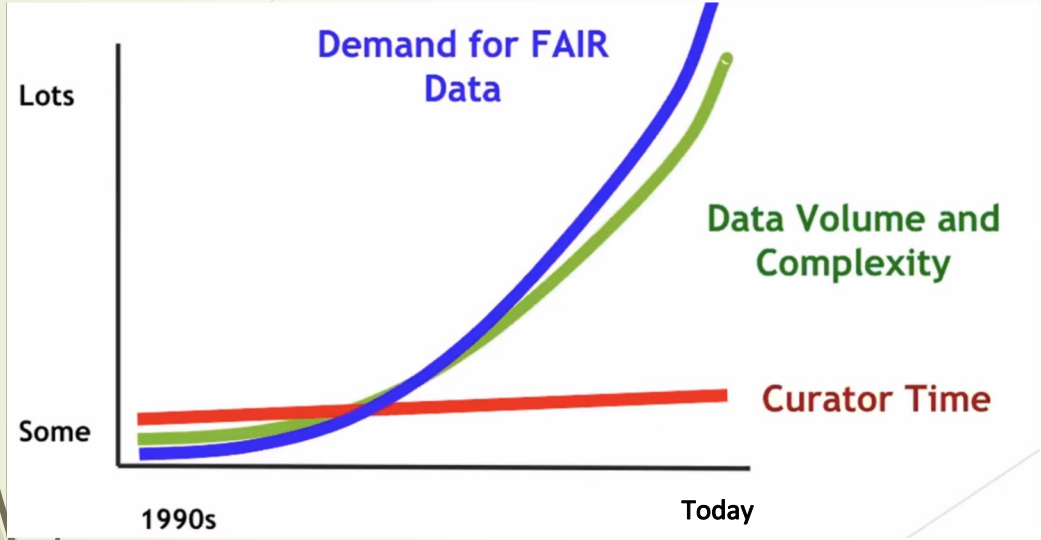
Data reusability limitations and challenges

- Challenges Associated with Data
  - Data diversity and data format heterogeneity
  - Data size, quality and versioning
  - Object identification
  - Metadata and data standardization
- Resources and Funding
- Implementation of FAIR data policy

Recommendations

1. Standardization of data collection protocols
2. Consistent data annotation
3. Data quality control
4. Data storage infrastructure, data management software and data curation tools
5. A concerted effort to make multi-omics data sets interoperable
   - **Community Crop Databases play an important role**

Needed : Increased Support for Biocuration!



Needed: Sustainability of the Core Databases

What challenges our WG faced?

- Too broad goals
  - Narrowed down the scope
- Multiple time zones
  - Sacrifices of some of the members (Thanks!)
- Everyone was already busy!
  - Division of work
  - Being flexible

What went well?

- Having wonderful people ☺
- Regular meeting at a fixed time
  - Kick Off Meeting Monday November 15th, 2021 12 PM EST
  - Bi-weekly meeting until ~ July 2022
  - Weekly meeting since ~ August 2022
  - Paper submitted on April 13, 2023
- Meeting reminders
- Having a finished product (white paper)

Acknowledgements

The AgBioData consortium

AgBioData SC members:
Jacqueline Campbell
Sunita Kumari
Sook Jung
Sushma Naithani
Monica Poelchau
Leonore Reiser
Meg Staton
Peter Harrison
John P. McNamara

Past AgBioData SC members:
Lisa Harper
Eva Huala
Marcela Tello-Ruiz
Laurel Cooper
Ethy Cannon
Ramona Walls
Dorrie Main

Past PC:
Darwin Campbell

Award # 2126334

External Advisory Committee
Robert W. Cottingham
Anne Kwitek
James Koltes
Marie-Angelique Laporte
Susanna Sansone