Challenges and Opportunities in Connecting Genotype Data to Phenotype Data

The AgBioData Consortium

https://www.agbiodata.org/

Sushma Naithani
Dept. of Botany and Plant Pathology
Oregon State University
sushma.naithani@oregonstate.edu

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

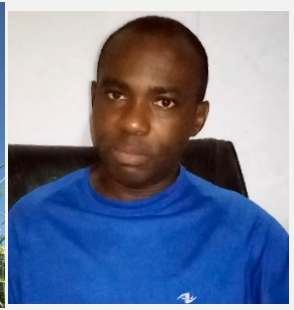
Nicholas Gladman,
Cold Spring Harbor
Laboratory, NY, USA

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

Challenges

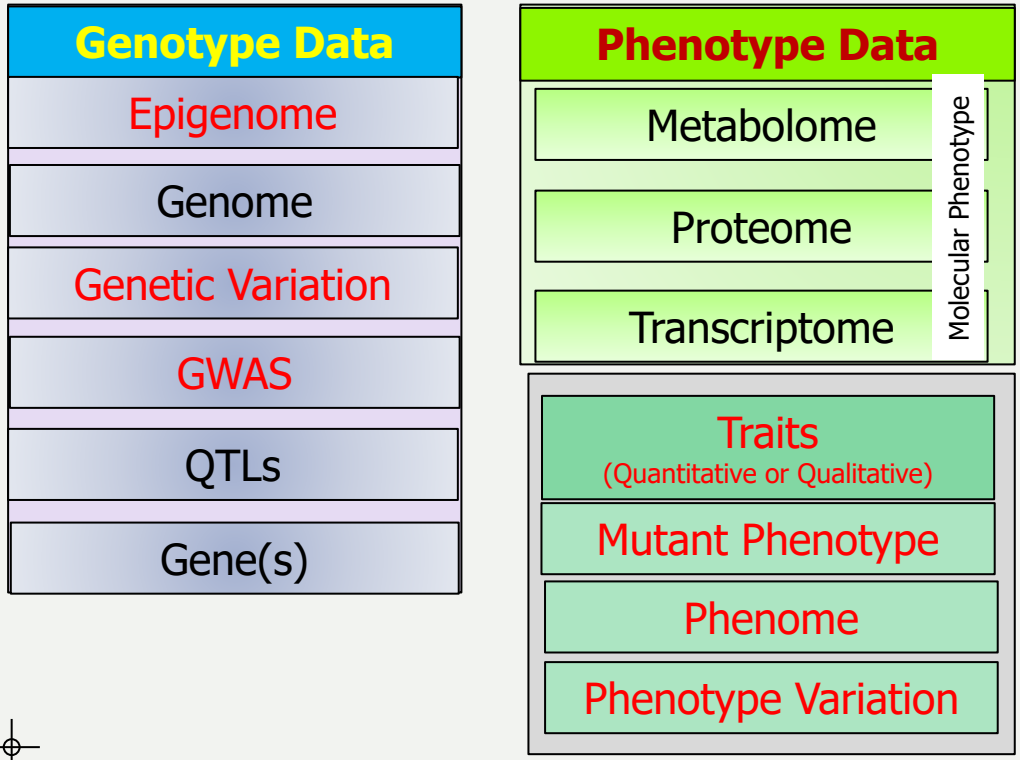
- **Data collection, sharing and integration issues :**
  - increased data size and complexity
  - diverse data formats
  - computer intensive analysis
  - quality control
  - data formats, metadata annotation
  - data storage and access
- **Limited data visualization and analysis tools**: to address a biological question, output formats
- **Data interoperability and integration:** genotype and phenotype data
- **Lack of Funding for Expert Biocuration:** limited high quality knowledge synthesis

The Genotype-Phenotype Working Group Aims & Goals

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

Aim 1 To Improve the Data Collection

Review of diverse data types, annotation, storage, and archiving in primary repositories

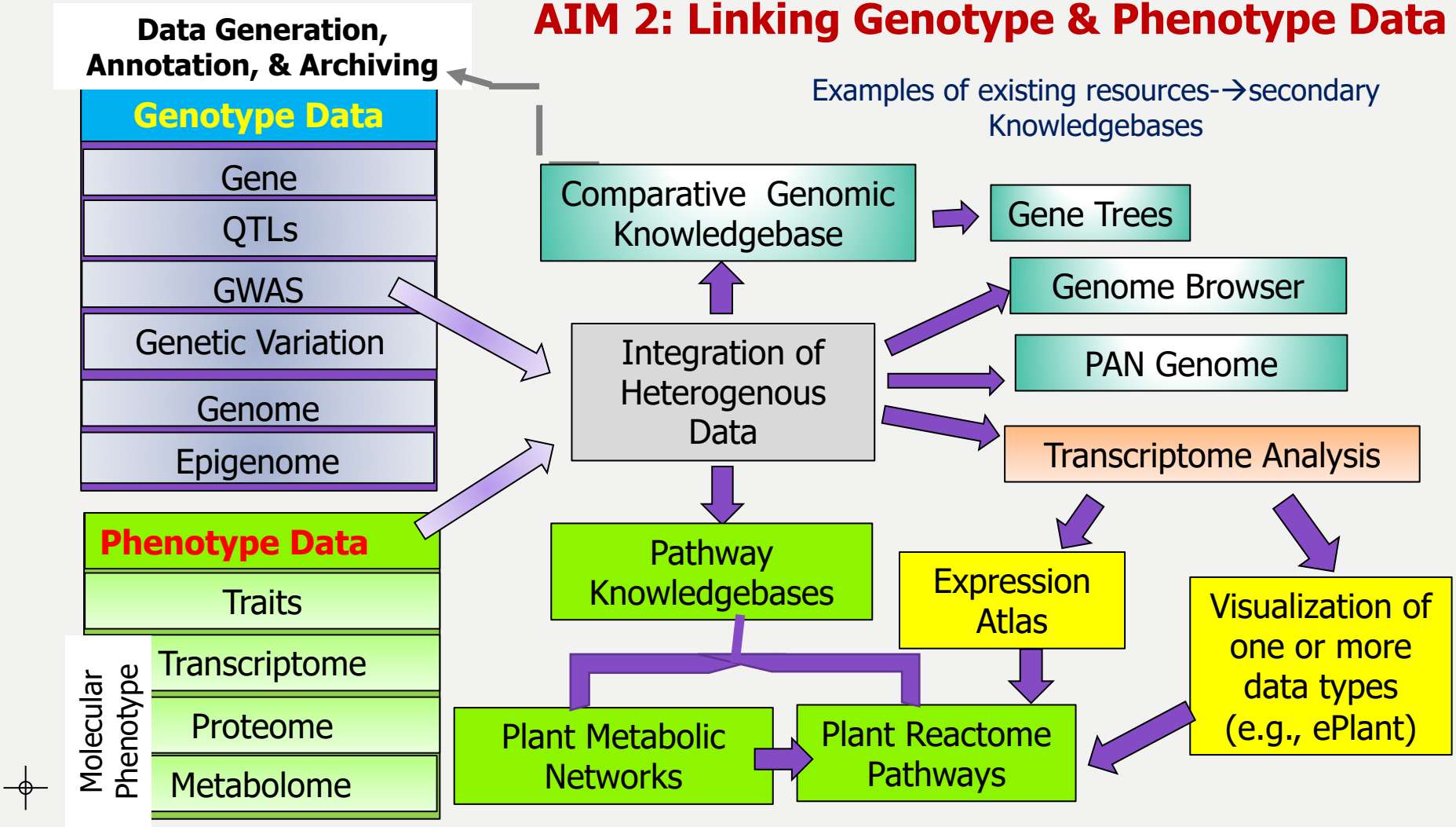


Data and metadata annotation: best practices are available for Genotype data and molecular phenotype data

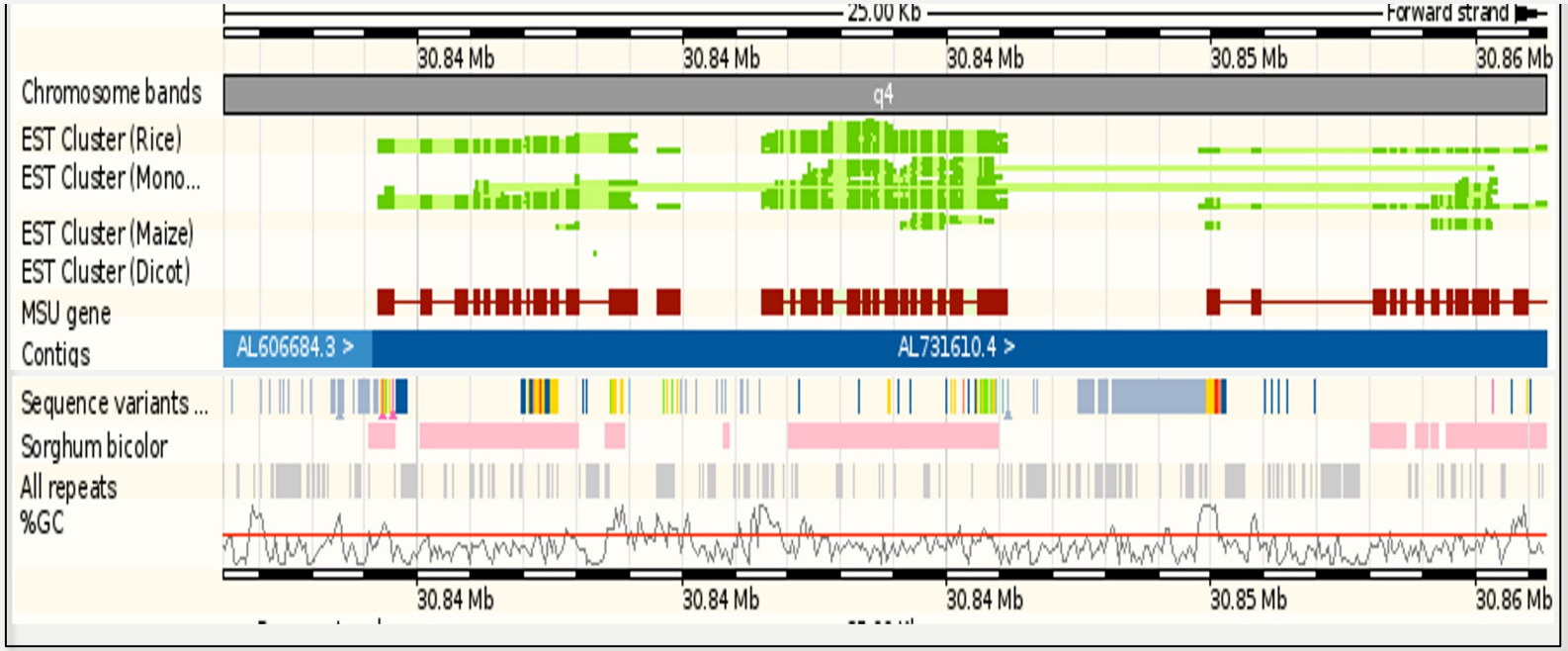
Ontologies: well-developed to describe both genotype and phenotype data of various types

Public Repositories: for storage, archiving & accessibility: Additional infrastructure is needed

AIM 2: Linking Genotype & Phenotype Data



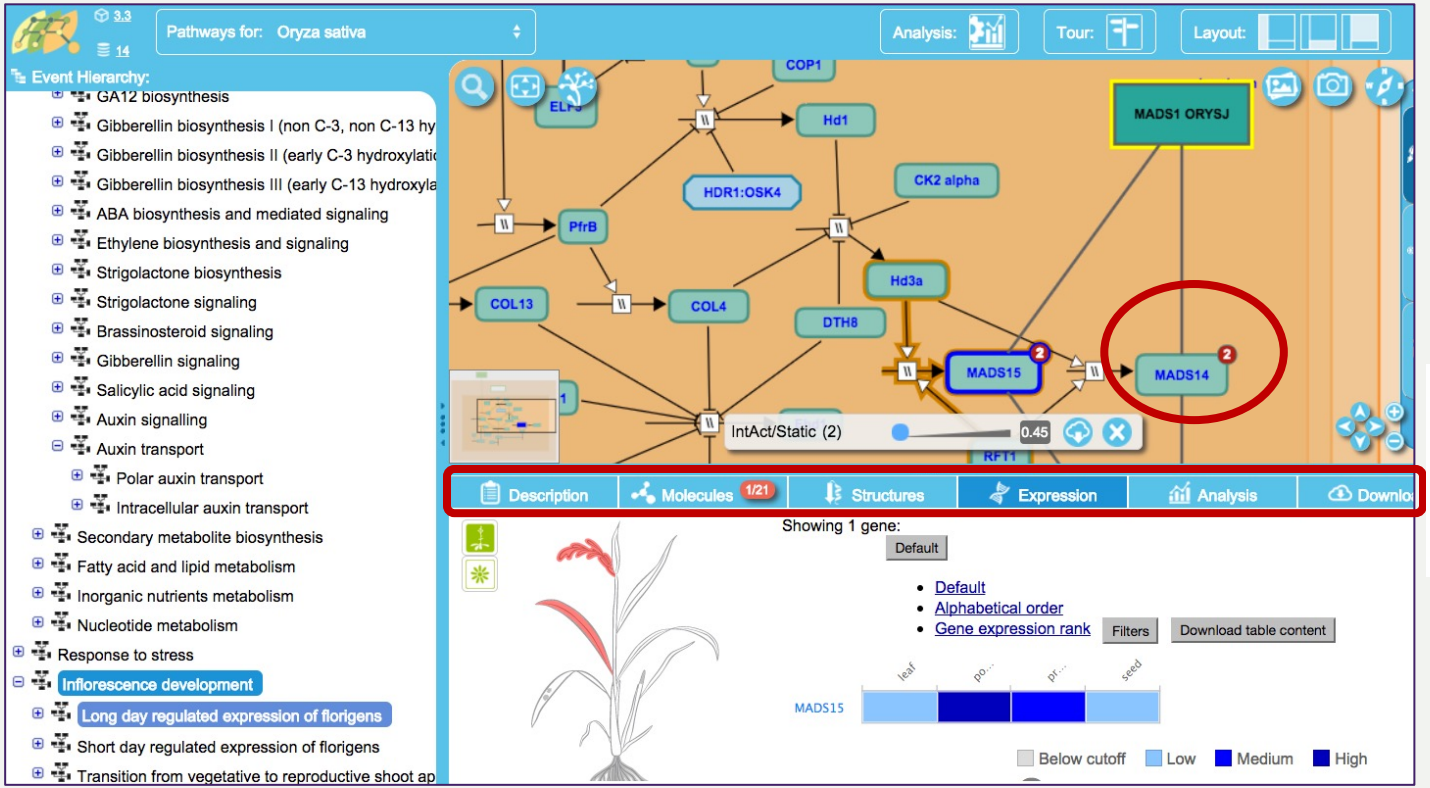
Integration of Heterogenous Data Types Gramene-Ensembl Genome Browsers



- Assembly structure and sequence
- Genes & transcripts
- Comparative alignments
- Baseline Annotation & Ontologies

- Genetic markers (SNPs, Indels, QTLs, SSRs)
- Repeats & transposable elements
- Regulatory & Epigenetic marks
- Transposon discovery

Integration of omics data in the Plant Reactome



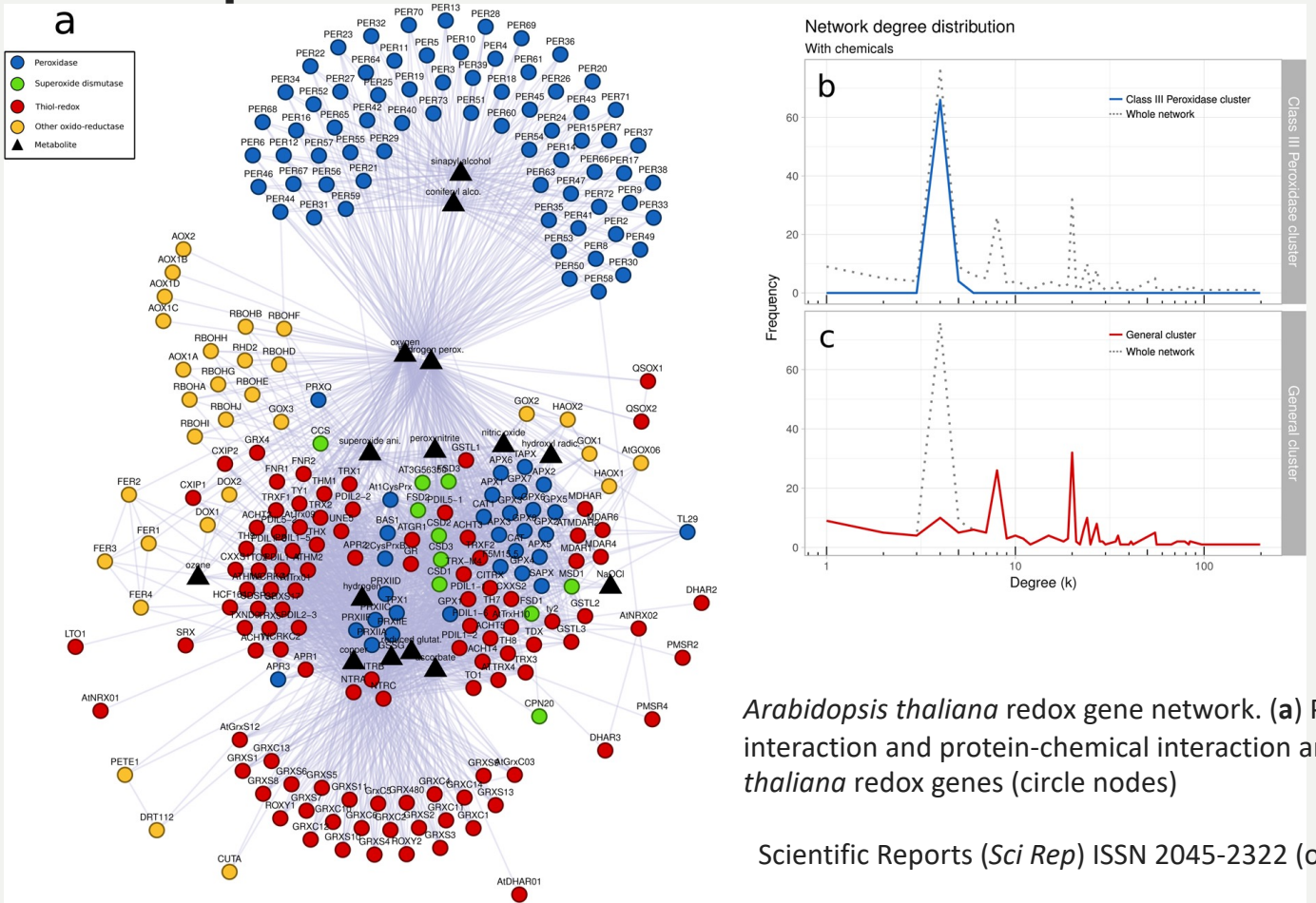
Links to the most updated information and contents from Gramene Gene Pages, UniProt, CHEBI, Expression Atlas, and PDB etc.

Linked Data Hub

- Data extensions provided by collaborators at EBI

Interactors (IntAct) and baseline expression data

Analysis of *Arabidopsis thaliana* Redox Gene Network Indicates Evolutionary Expansion of Class III Peroxidase in Plants

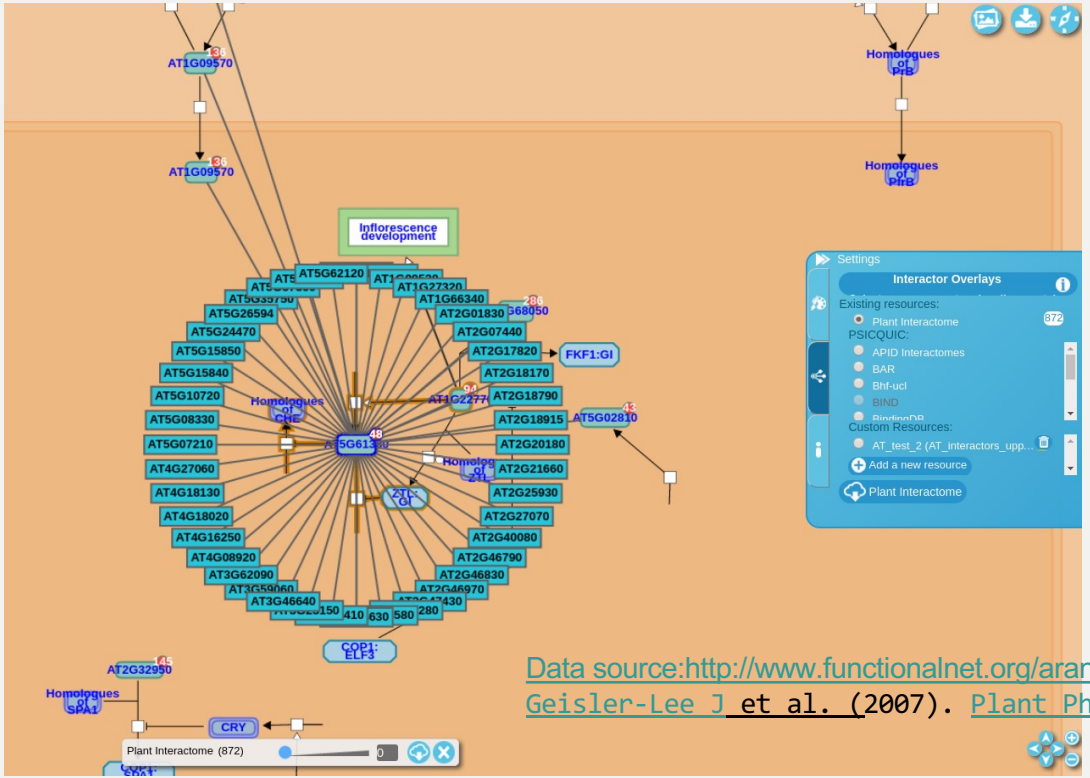


Arabidopsis thaliana redox gene network. (**a**) Protein-protein interaction and protein-chemical interaction among 246 *A. thaliana* redox genes (circle nodes)

Scientific Reports (*Sci Rep*) ISSN 2045-2322 (online)

Displaying gene-gene interaction data on Plant Reactome pathways

Synthesizing gene-gene interactions information



Users can also upload their own data

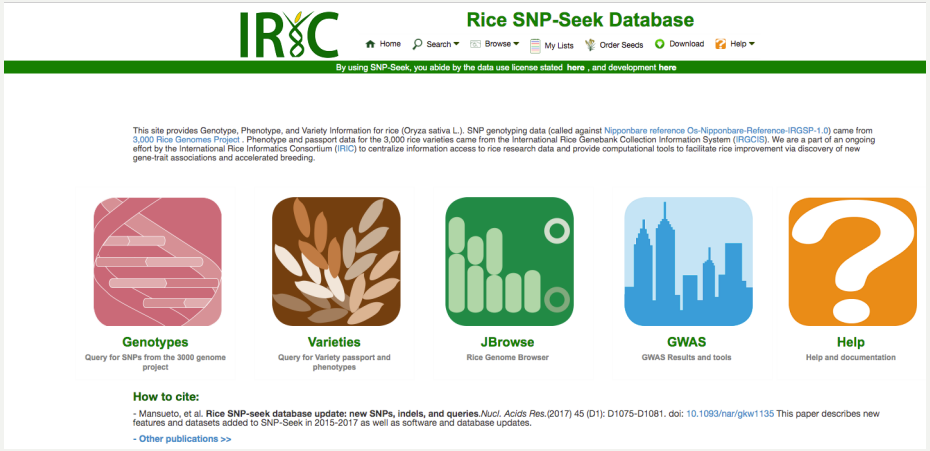
Users can send us the data from their favorite species

Data source:http://www.functionalnet.org/aranet/

Geisler-Lee J et al. (2007). Plant Physiol. 145(2):317-29

AIM 2: Linking Genotype & Phenotype Data

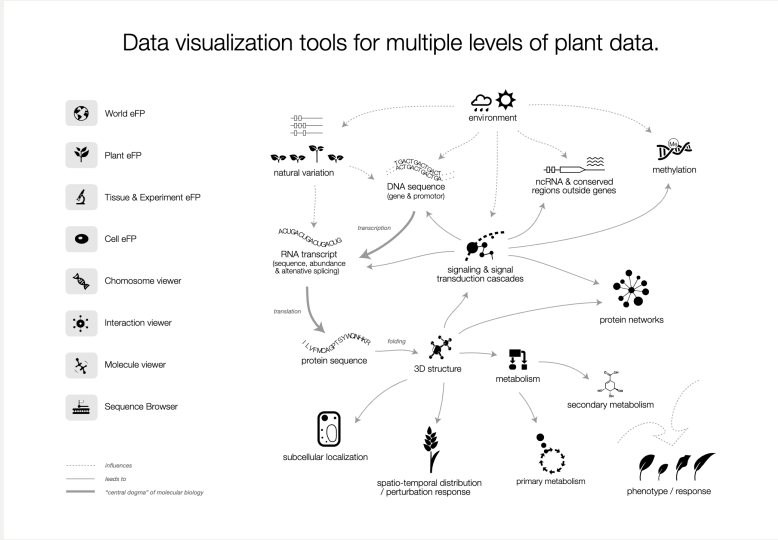
Examples of existing resources-→secondary Knowledgebases



https://bar.utoronto.ca/eplant/

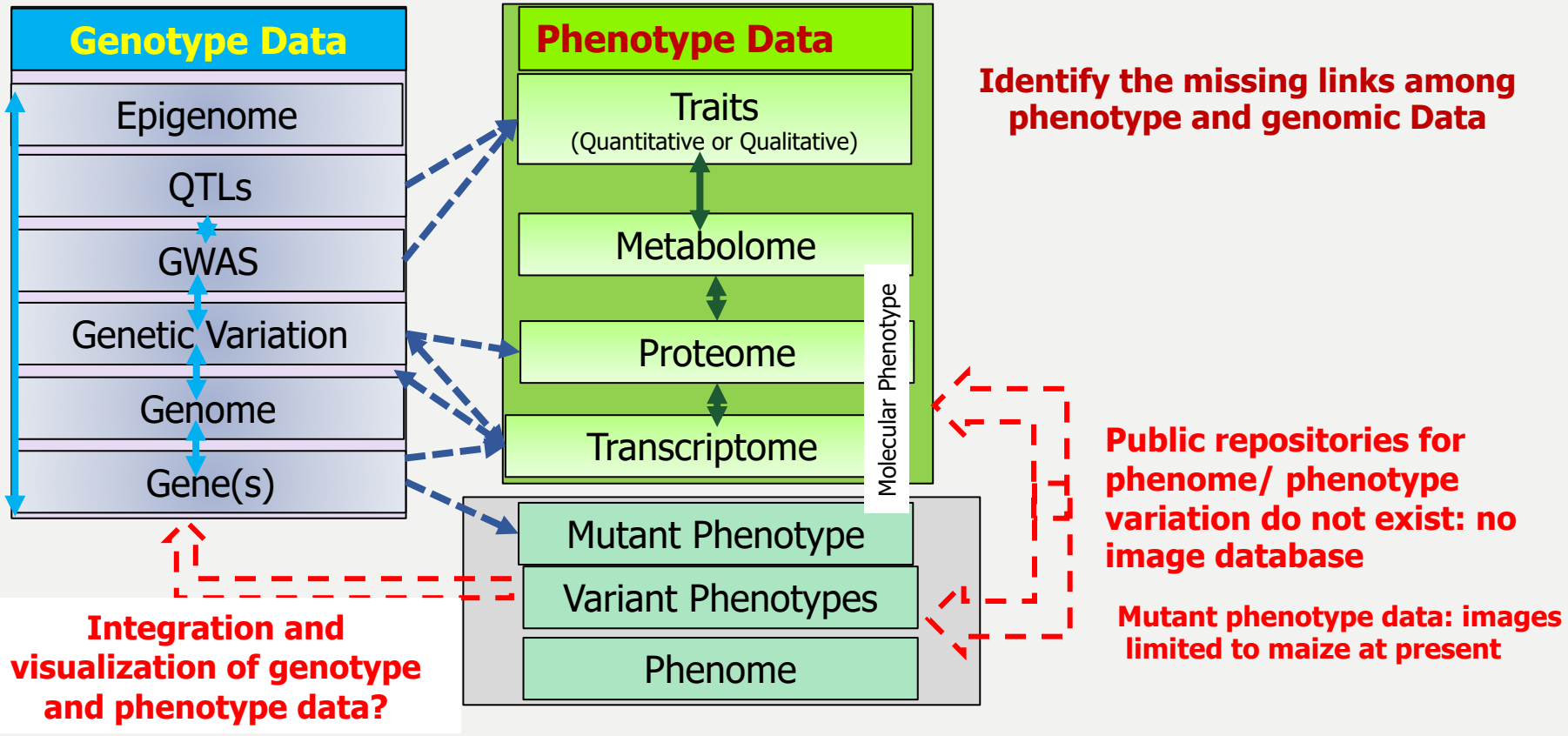
Genomic variation in 3,010 diverse accessions of Asian cultivated rice

- ~29 million single nucleotide polymorphisms,
- ~2.4 million small indels
- ~ 90,000 structural variations that contribute to within- and between-population variation.
- ~10,000 novel full-length protein-coding genes that have a high number of presence–absence variations.



Wensheng W. et al. (2018), *Nature* **557**:43–49

AIM 3: Limitations of data re-use and interoperability



How does genotype relate to phenotype?

Reverse Genetics Approach

One Gene

↓

One Phenotype

Mendelian Genetics

Gray, W. M. (2004). "Hormonal Regulation of Plant Growth and Development". *PLoS Biology* **2** (9): e311. DOI:10.1371/journal.pbio.0020311

Extension to Mendelian Genetics

Gene 1+ Gene 2-→ Multiple Phenotypes

Example:
Fruit color in squash

Gene A
increases orange

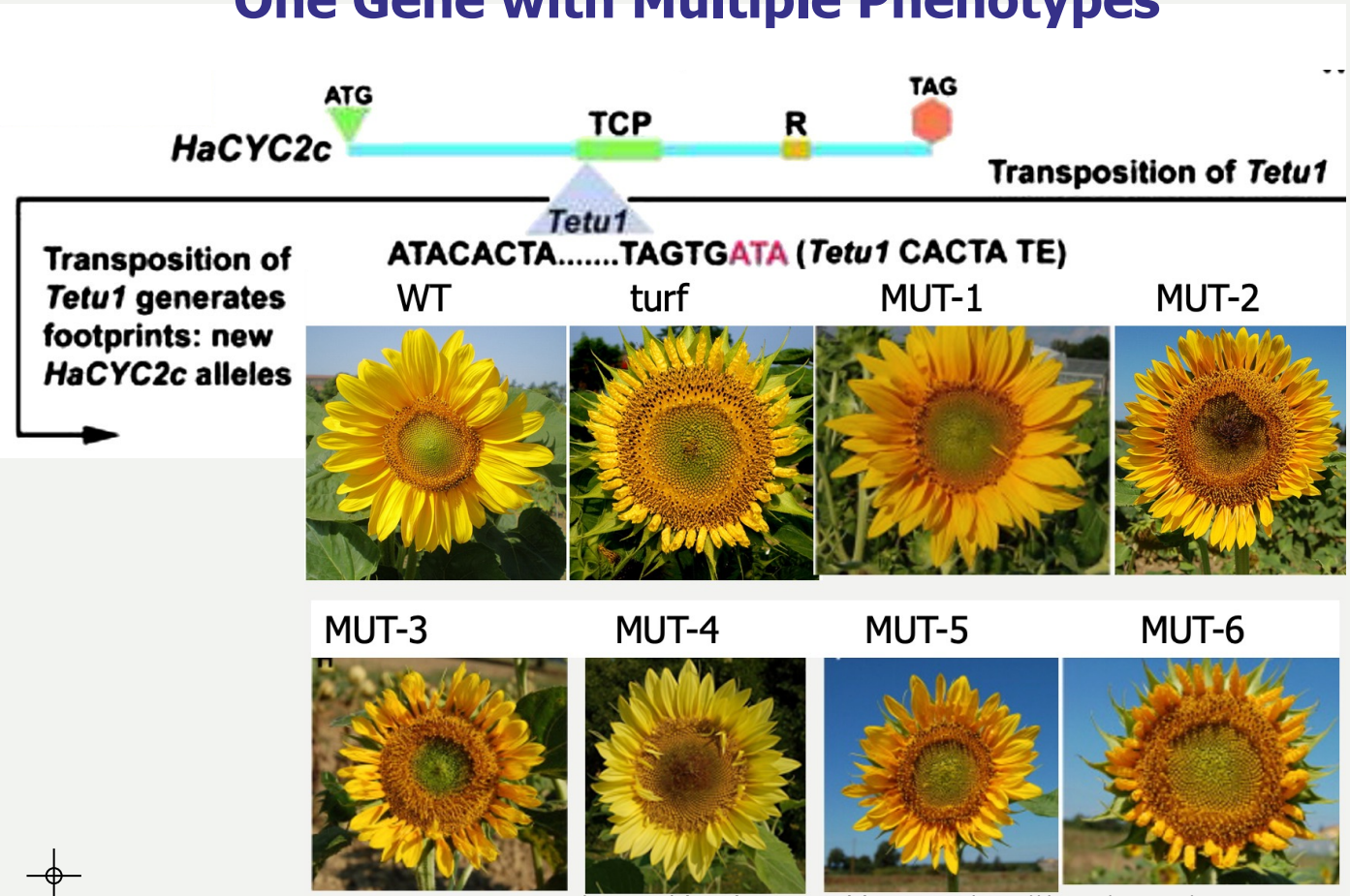
Gene B
increases yellow

Additional colors result from interaction of these two loci

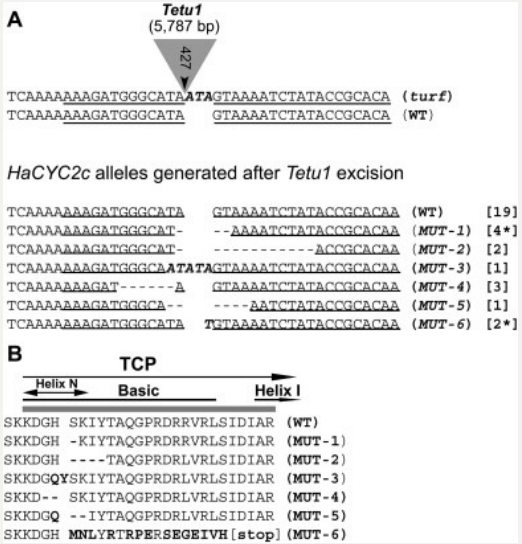
Immature cukes are green

How does genetic variation give rise to phenotypic variation?

One Gene with Multiple Phenotypes

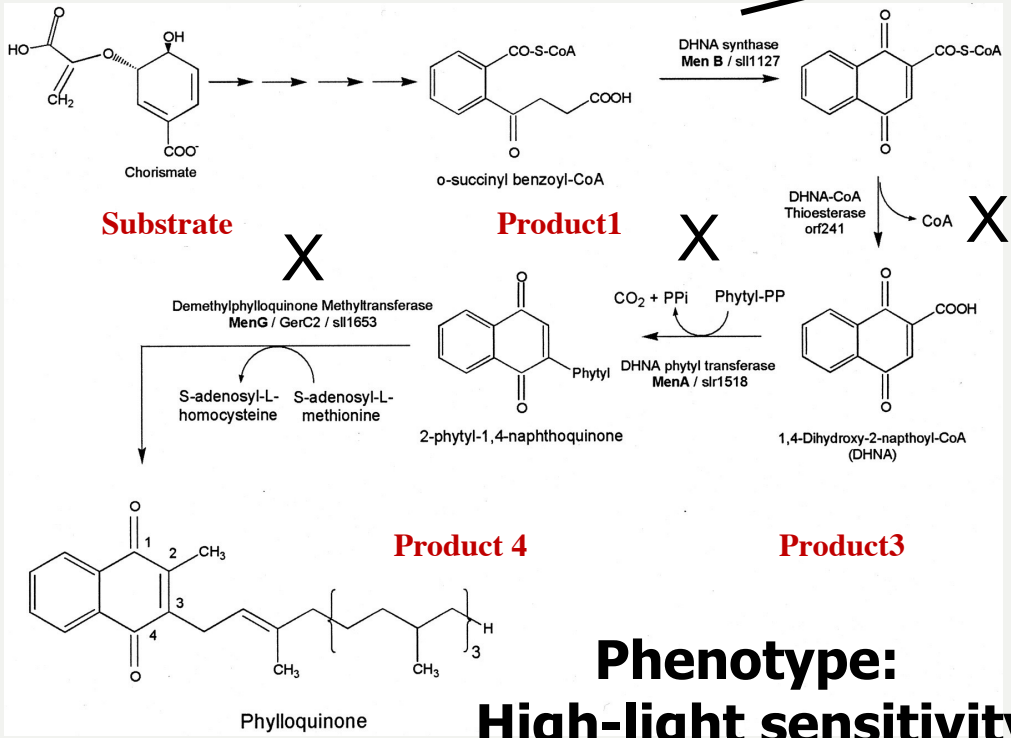


Example: Transposition Tetu1 transposon alters *HaCYC2c* expression and affects its TCP domain folding that gives rise to a range of flower phenotypes of ray flower corolla and sexual organs

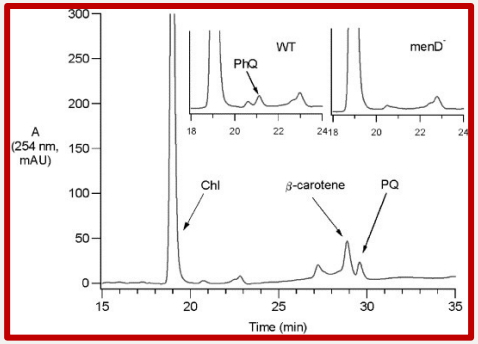


Fambrini *et al.* (2014) *Gene*, *549*(1), 198-207. https://doi.org/10.1016/j.gene.2014.07.018.

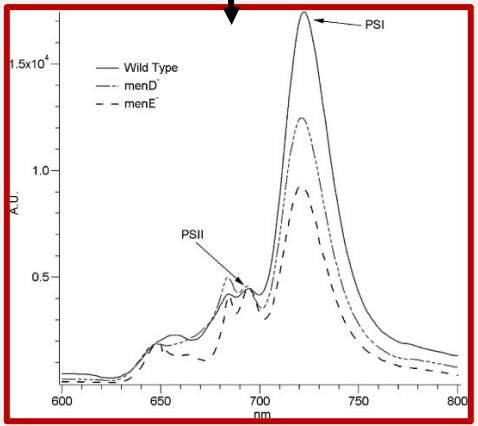
Multiple Genes→ One Phenotype



X



Molecular Phenotype:
Absence of Phylloquinone



Imbalance in PSI/PSII ratio

Physiological Phenotype:
no photosystem I activity

Phenotype:
High-light sensitivity

Johnson TW, Naithani S, et al. (2003). Biochim Biophys Acta. ;1557:67-76.
DOI: 10.1016/s0005-2728(02)00396-1

Complexities in Relating Phenotype to Genotype

(Exemption to Mendel's Law of one gene->one phenotype)

Perspectives

- **Cell specific expression**
- **Stage specific expression**
- **Environmental effect**

- Gene redundancy
- Multiple alleles → many phenotypes
- Penetrance or Expressivity
- Epistasis
- Lethal genes
- Linkage
- Pleiotropy

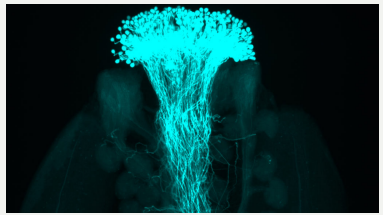
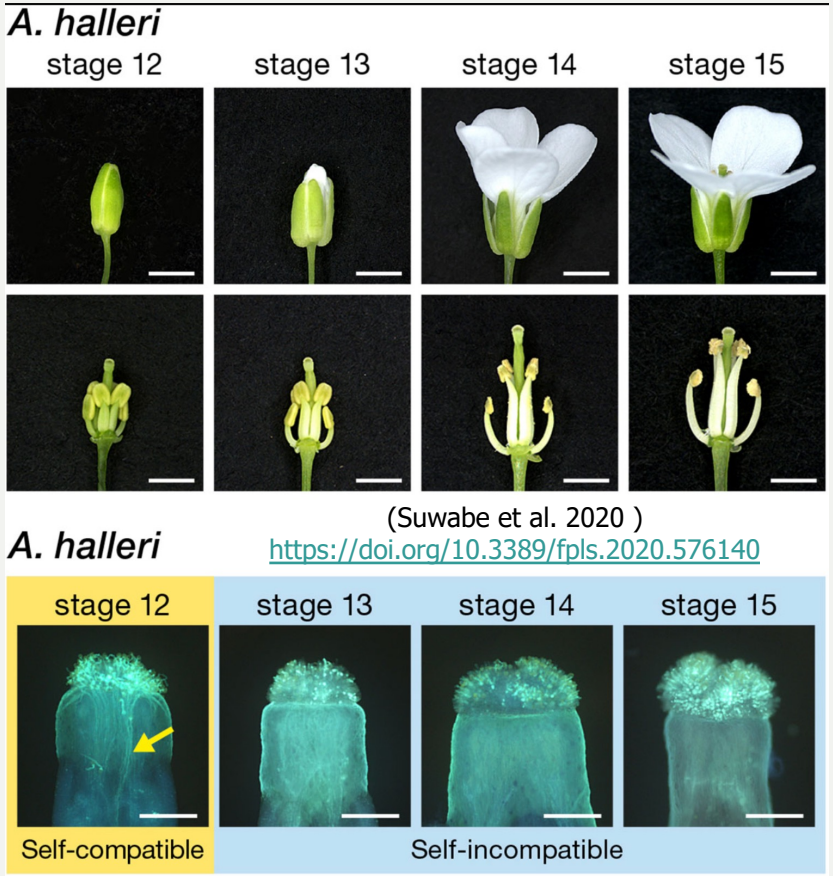
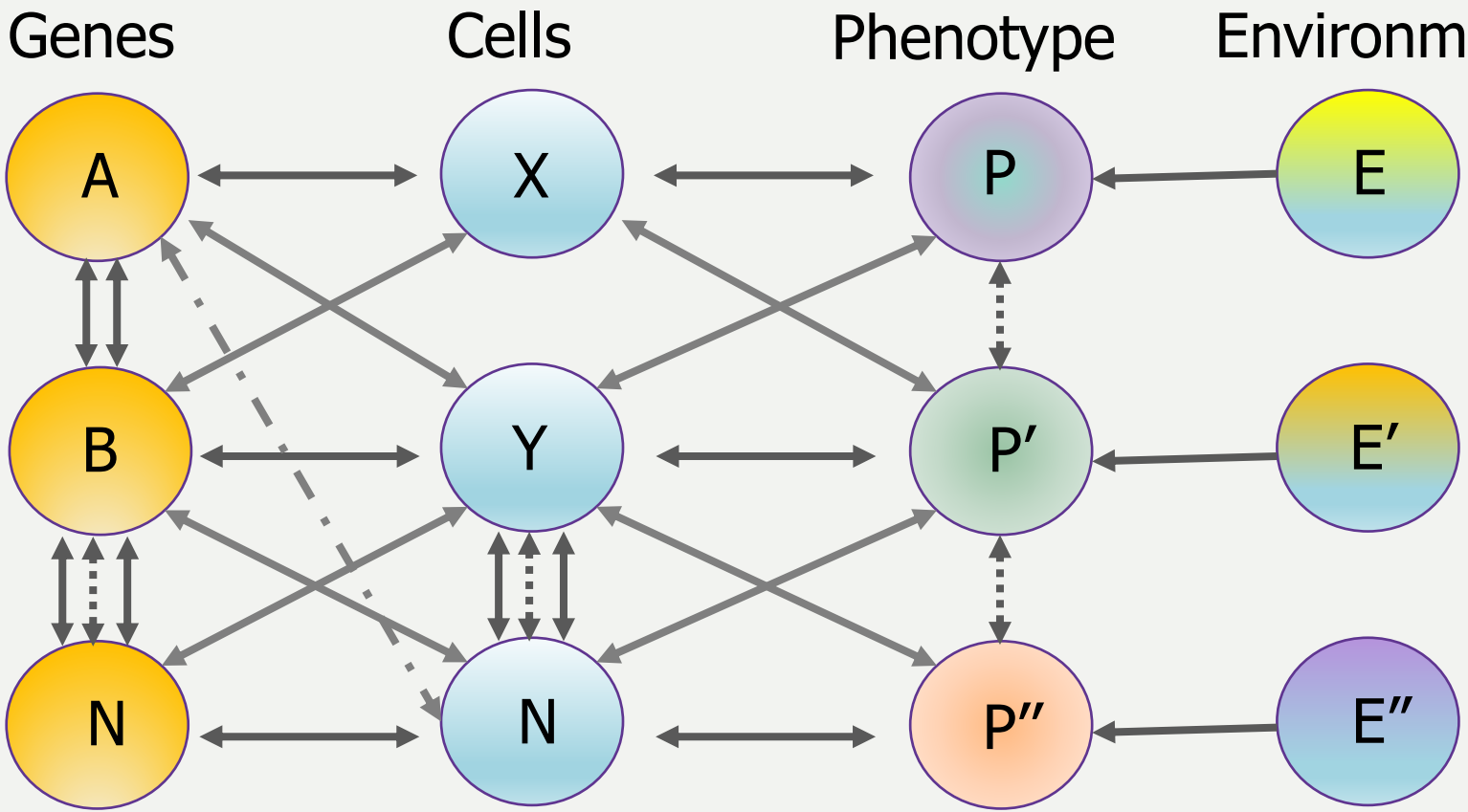


Photo by Jing Yuan, Ph.D. student, Kessler Lab (https://ag.purdue.edu/stories/pollination-a-classic-tale-of-romance-love-and-death/)



(Suwabe et al. 2020)
https://doi.org/10.3389/fpls.2020.576140

Phenotypes are the outcomes of complex interactions (G(s) +E)



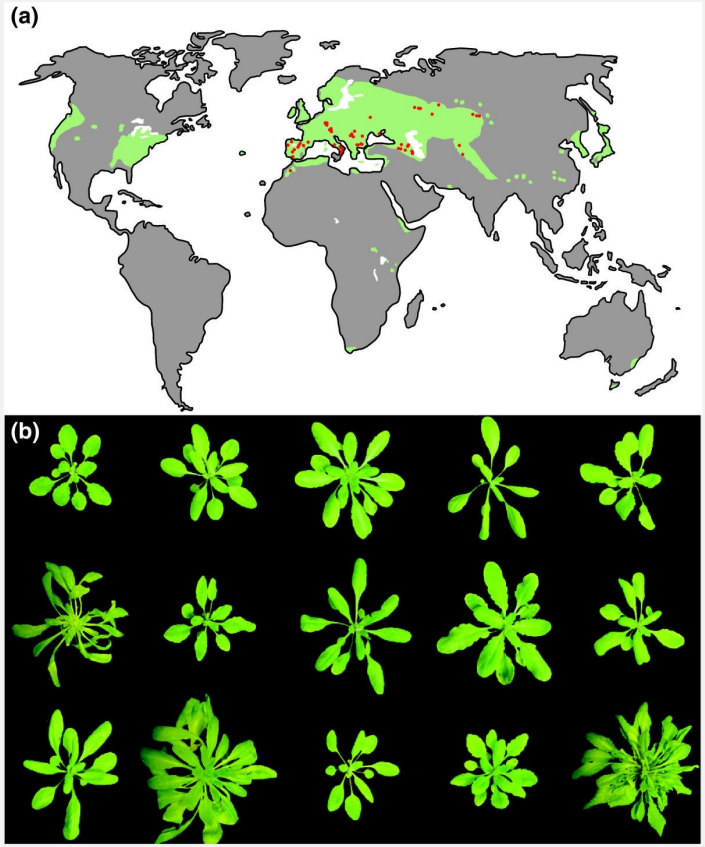
Phenotype -> Genotype

Phenomics Approach & GWAS

For understanding

natural variability and biodiversity

Linking phenotype to genotype in genome-wide association studies (GWAS): Forward Genetics Approach

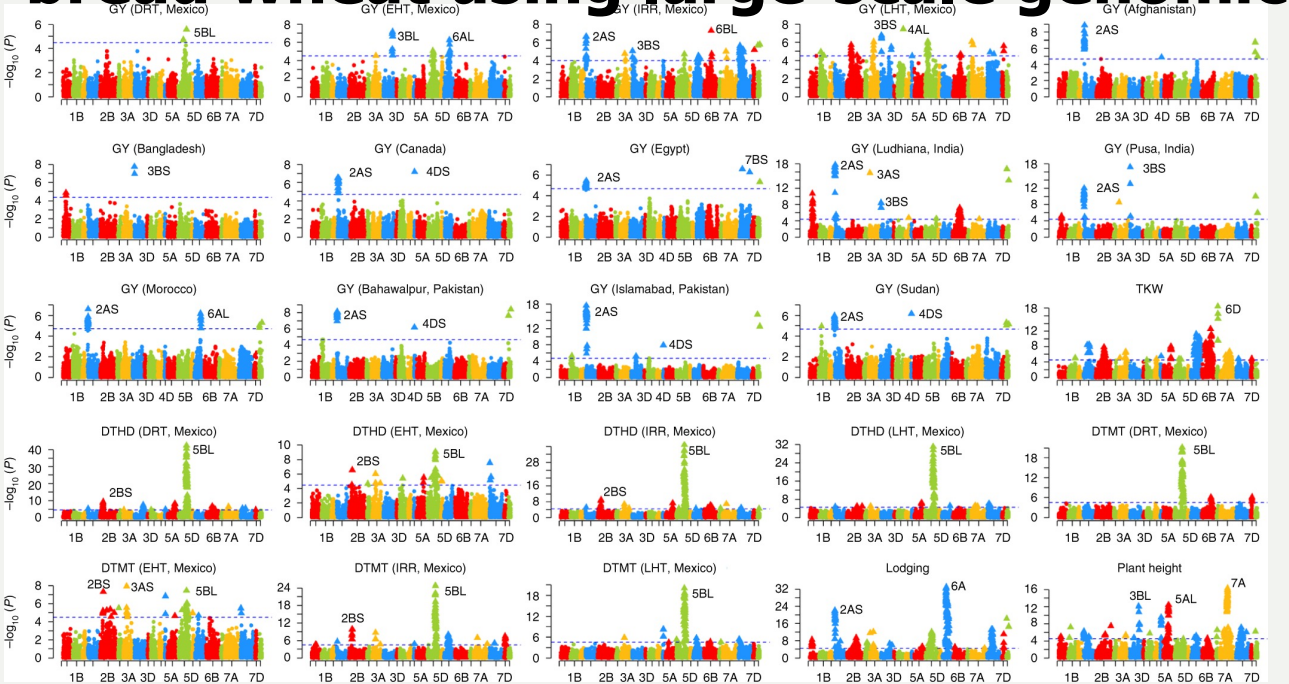


Natural habitat of various accessions of *A. thaliana*

Vegetative rosettes illustrating genetically determined variation in morphology among *A. thaliana* accessions

Weigel and Mott (2009): **Genome Biol**. 2009;10(5):107. doi: 10.1186/gb-2009-10-5-107

Improving grain yield, stress resilience and quality of bread wheat using large-scale genomics

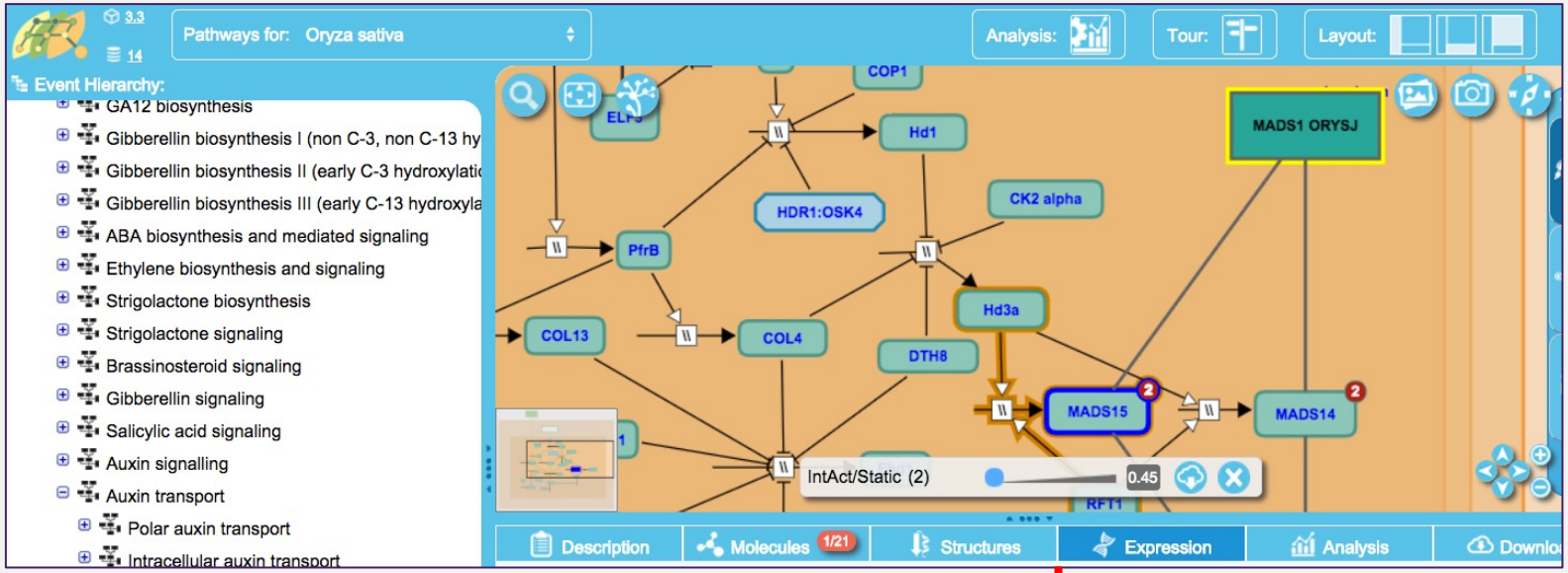


Manhattan plots showing marker–trait associations for GY and agronomic traits from a genome-wide association mapping study. The 2AS chromosomal region was significantly associated with grain yield in several environments, whereas the 5BL and 2BS chromosomal regions were associated with days to heading and maturity in several environments.

Nature Genetics **volume 51**, pages1530–1539(2019)

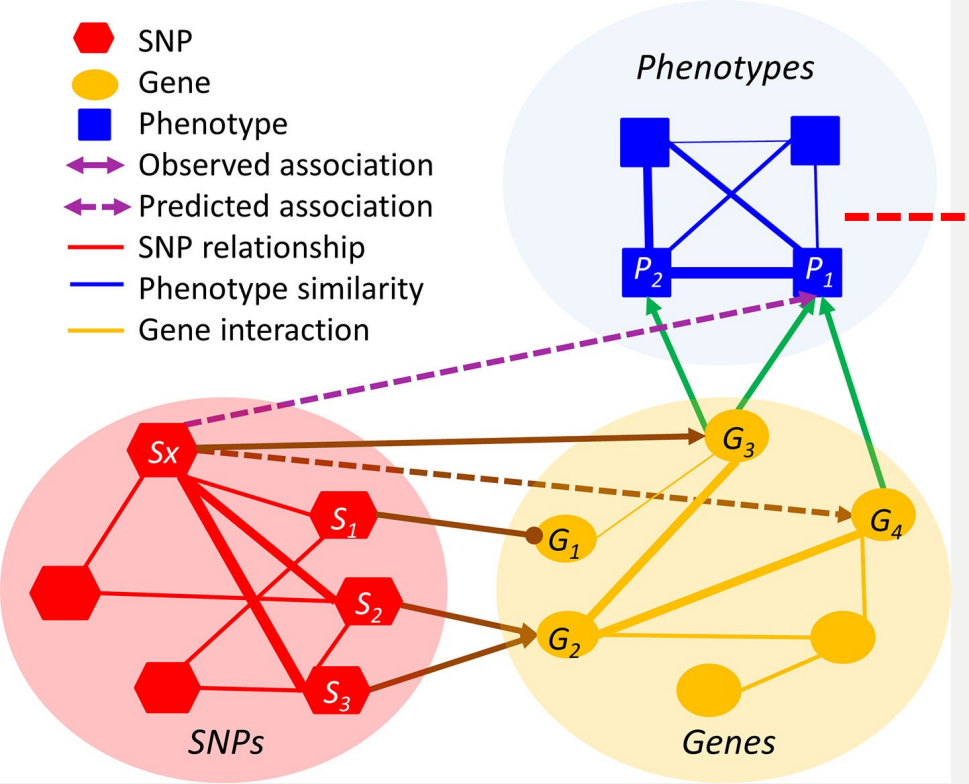
Linking Phenotype with existing platforms

Plant Pathways would be very useful



Missing:
Images of phenotypes

Linking genotype to Phenotype Data with gene pages of Ensembl and other crop databases like MaizeDB



Missing: Images of phenotypes

Mutant phenotype data; limited to maize at present

MaizeDIG: Maize Database of Images and Genomes

Aim 3: To Review Data Interoperability & Re-use

Data Formatting, Annotation, and Storage

Data generation & collection
- Genome Sequence
- Transcriptomes
- Proteomes
- Genotyping/GWAS
- Phenotype
- Phenome

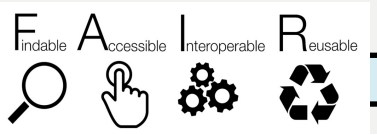
Primary Repositories
- Cyberinfrastructure
- Storage
- Annotation
- Metadata

Secondary Knowledgebases

e.g., Crop databases, Phytozome, Gramene-Ensembl, PMN, Reactome
- Annotation enrichment
- Integration of Genotype & Phenotype Data
- Analysis & Visualization
- Biocuration
- Knowledge synthesis
- Building analysis tools

Bottleneck in Omics data

Data Impacts
- Hypothesis
- Shortlist candidate genes & pathways
- Translational research



Modified figure from Naithani et al. (2019), Database (Oxford), Volume 2019, bay146, https://doi.org/10.1093/database/bay146

Some Thoughts …….

- Linking genotype and phenotype data is one of the greatest challenges in research and development

- Why it is so difficult to map phenotypes to genotype data?

Potential solutions:

- Standardized data formats
- Data annotations using ontologies, enriched metadata
- Automated flows for data management and mappings

Input machine readable--→ output machine readable

Not useful for human mind: the way it works, thinks and creates new

How to we integrate the diverse data so it is easy to understand ?

The Biocuration by experts could help in Knowledge synthesis

Additional investment in data visualization platforms will be of tremendous value to biologists

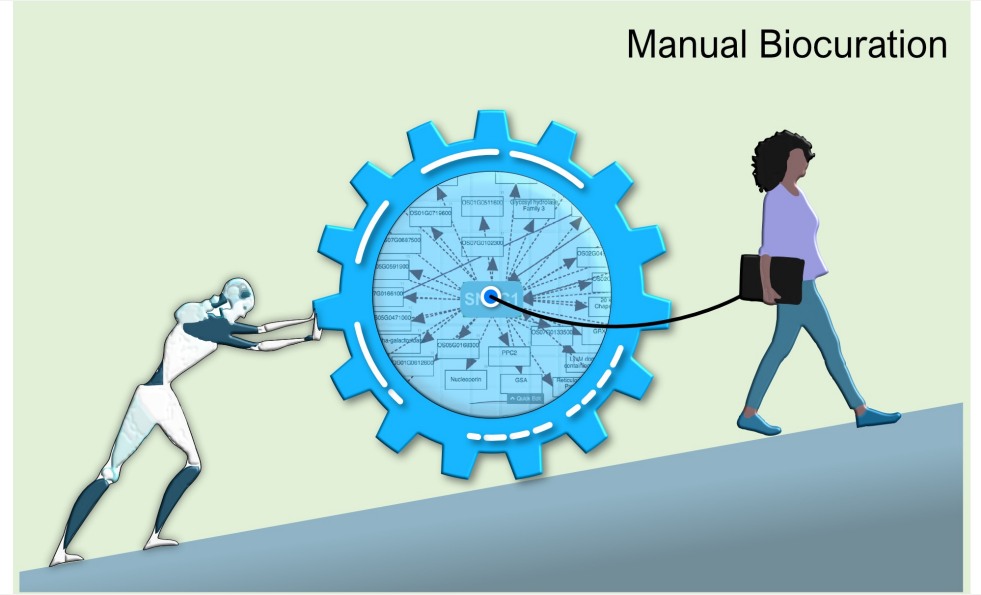
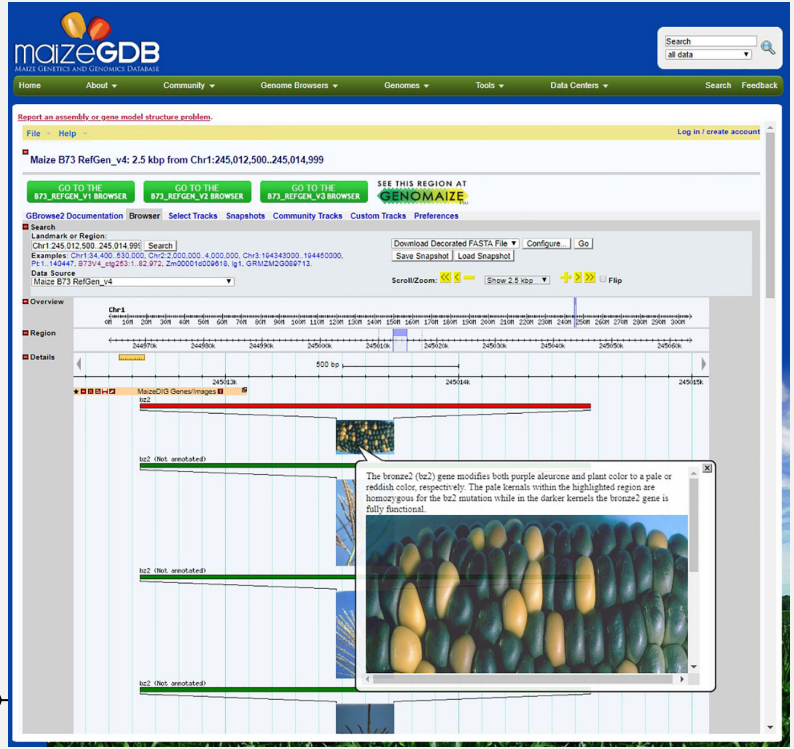
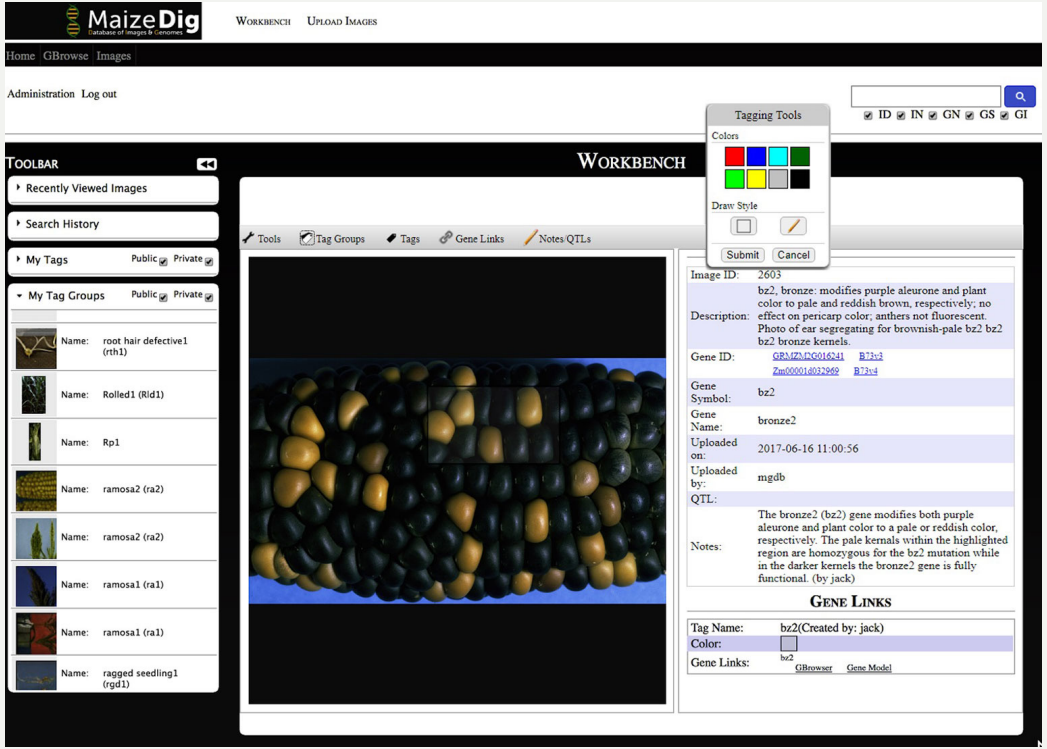


Illustration by **Dmitry Shevela**
https://www.instagram.com/scigrafik

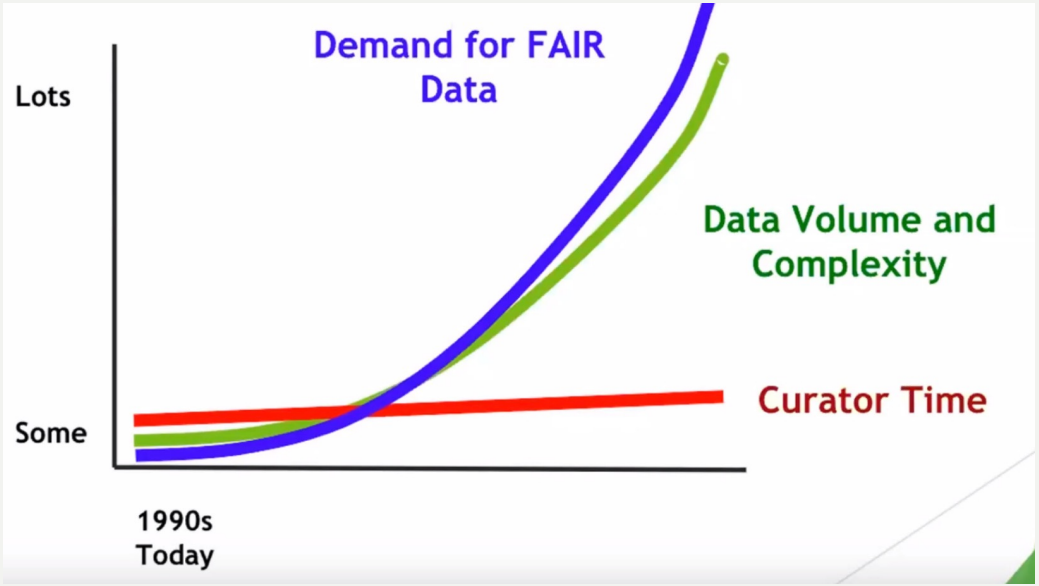
MaizeDIG: provides high-quality data on Genotype—to-phenotype through value added expert curation, with easy-to-understand visual format in a model species rich in genotype and phenotype data.

https://maizedig.maizegdb.org (Cho et al. 2019: https://doi.org/10.3389/fpls.2019.01050)





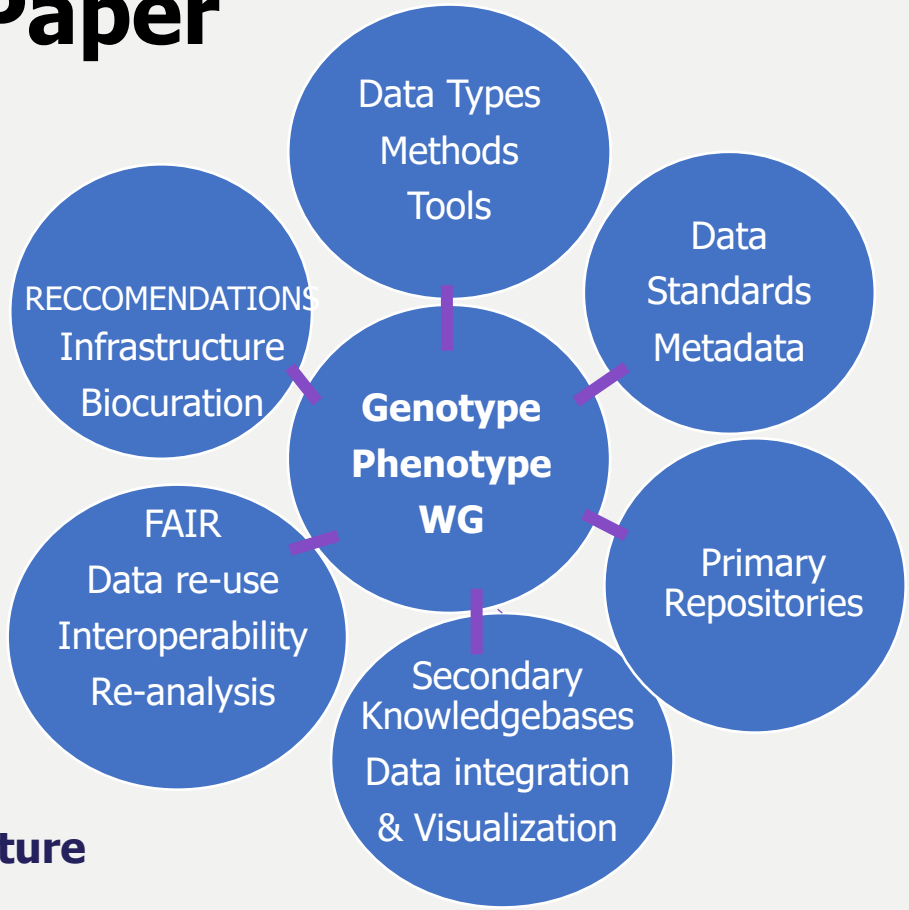
Needed : Increased Support for Biocuration



Needed: Sustainability of the Core Databases

Final Product: White Paper

- **Introduction**
- **Main Content**
  - + **Review the data types and method**
  - + **Primary repositories and secondary Knowledgebases**
  - + **Requirements of metadata and the minimum standards**
  - + **Examples of re-use and re-analysis**
  - + **Limitations: data re-use, FAIR policy, sustainability of public databases**
- **Recommendations: additional infrastructure & Biocuration needs**



Thanks!

The community is fundamental to advance
we have deluge of Omics data: help is appreciated

Acknowledgements

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

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

Past PC:
Darwin Campbell

The AgBioData consortium

Award Abstract # 2126334