Main change from HDB

- Storing the hierarchical 'living stock' data in Stock table instead of stock and biotype (except specimen and clones) - Stock table, along with stock_relationship table, can store the relationships between stocks (eg. Stock groups and pedigree)

- Store the data of specimens and clones in 'stocksample' table and the specific factors of breeding data that give distinct phenotypic and genotypic value in gtexperiment of ptexperiment tables.

- Other breeding data schemas
  - GDPDM
  - ICIS (International Crop Information System)

Changes from HDB and current Chado

1. **HDB has been modified to make it more chado-ish and to fit to plant breeding data**
   - Modified: gtassay, gtexperiment,
   - Added : crossexperiment_stock, crossexperiment_project feature_gtassay, stocksample, stocksampleprop, ptexperiment
   - Deleted: biotype, biotype_organism, stock (HDB-type), individual, individual_biotype, individualprop, crossexperiment_individual, individual_relationship, specimen, gtexperiment_project, individual_image, specimen_image, individual_phenotype, individual_phenotype_project, stock_phenotype, stock_phenotype_project, biotype_phenotype, biotype_phenotype_project

2. **For other modules of chado, tables below are added**
   - Contactprop
   - Projectprop
   - Project_relationship
   - Stock_relationship_cvterm
   - Possible Modification of Phenotype Table: Deleting observable_id and attr_id – not included in the sql document since we can choose not to use those rows.

Example breeding data

- phenotype (genotype) data
  - Project - plant (Fuji) – 35 (sugar content)
  - Project - plant (Fuji) – property (planted in plot A or orchard B, treated with 100 g of fertilizer B, collected in Mar 2008, and assayed in Apr 2008) – 35 (sugar content)
- stock table – Fuji
- Stocksample table – Fuji-Orchard B-plot A
  - Stocksampleprop table records other details such as treatment
- ptexperiment
  - 'stocksample'
  - 'collection_date': Mar 2008
  - 'assay_date': Apr 2008
- The three rows above will be associated with distinct ptassay (sugar content assay) and distinct phenotype (35).

Another example breeding data

PlantYear	Orchard	*Code	**Rep	spring '04 Diam(mm)	fall '04 Diam(mm)	fall '05 Diam(mm)	Flower Number	Fruit Number	Mean wt
2004	A	13	2	.	18	36	77	17	232
2004	A	6	2	.	16	34	32	2	241
2004	A	3	2	.	18	37	60	24	92
2004	A	8	2	.	16	25	36	6	164
2004	A	12	2	.	19	27	46	27	250
2004	A	5	2	.	19	25	22	6	210

*stock; **plot number

- Users would like to search for all the stocks with tree diammeter over 30
- stocksample table can store the specific clone data - orchard, stock, and plot number.
- Ptexperiment table links the clone (Orchard A, Code 13, Rep 2), assay date (fall 05), ptassay (Diam), and the phenotype value (36).

1. stock
    - Stock (germplasm) named by breeders, researchers or germplasm centers
    - can be a population, an established cultivar/breeding line or an individual (eg. collected from wild or from a cross) but not likely to be clones that are propagated in breeding programs (presentation on Germplasm Ontology)
    - link to pedigree and passport data

2. stocksample
    - A clone of a tree that is planted in a certain plot in a breeding program
    - A specimen of an organism (wing of a fly)
    - Any treatment done for the stocksample (eg. Fertilizer) or plant date can be recorded in stocksampleprop table.
    - When stocksample data is not available, stock name may be stored as stocksample name

3. gtexperiment/ptexperiment
    - Along with collection date and assay date, stocksample is associated with a specific genotypic/phenotypic value and a specific phenotypic/genotypic assay

Breeding Data (apple)

Female x Male

↓

Test 1st seedlings (1st Phase)

↓

Pick 50 genotypes
and propagate (15 trees per genotyp

Test in three different trial sites
(5 trees per genotype in each site)
Controls are included (2nd Phase)

↓

Pick 15 genotypes
and propagate (100 trees per genotype)

Test in four different sites
(25 trees per genotype in each site)
Controls are included (3rd Phase)

All 50 genotypes are stored in Stock

Each 15 x 50 tree is stored in stocksample table.
Each stocksample at a certain collection/assay dates is associated with specific ptassay and the phenotypic value

Each 15 x 100 is stored in stocksample table.
Each stocksample at a certain collection/assay dates is associated with specific ptassay and the phenotypic value

Breeding Data (apple)

Female x Male

↓

Test 1st seedlings (one of them WSU123)

↓ Pick 50 genotypes and propagate (15 trees per genotype)

Test in three different trial sites
(5 trees per genotype in each site)
Controls are included

↓ Pick 15 genotypes and propagate (100 trees per genotype)

Test in four different sites
(25 trees per genotype in each site)
Controls are included

stocksample entries

1. ID_1: Wsu123 @ nursery
2. ID_2: Wsu123 @ site 1 @plot2

ptexperiment

1. ID_1 collected @1/1/2008 , tested on 2/1/2008 @1st phase with ptassay1 and phenotype value 2
2. ID_2 collected @1/1/2009 , tested on 2/1/2009 @2nd phase with ptassay1 and phenotype value 3

↓

Stock entry
WSU123

QTL (tomato data from SGN)

General info on QTL experiment is stored in 'project' table

Sausage (F) x LA1589 (M)

F1

107 F2 lines ← Phenotypes are evaluated

107 F2 lines

1. All F2s are stored in Stock

2. Each 107 F2 line is stored in stocksample and associated with a specific ptassay and phenotype through ptexperiment

When there is no replication in the experiment, in other words one genotype (stock entry) is tested once in a project, stocksample.name can be the stock name.

3. When 107 F2 line is planted again those may not be stored in the stocksample again, but associated with new values through different project_id in ptassayexperiment table

Genotype/phenotype diversity data (Data from GRIN)

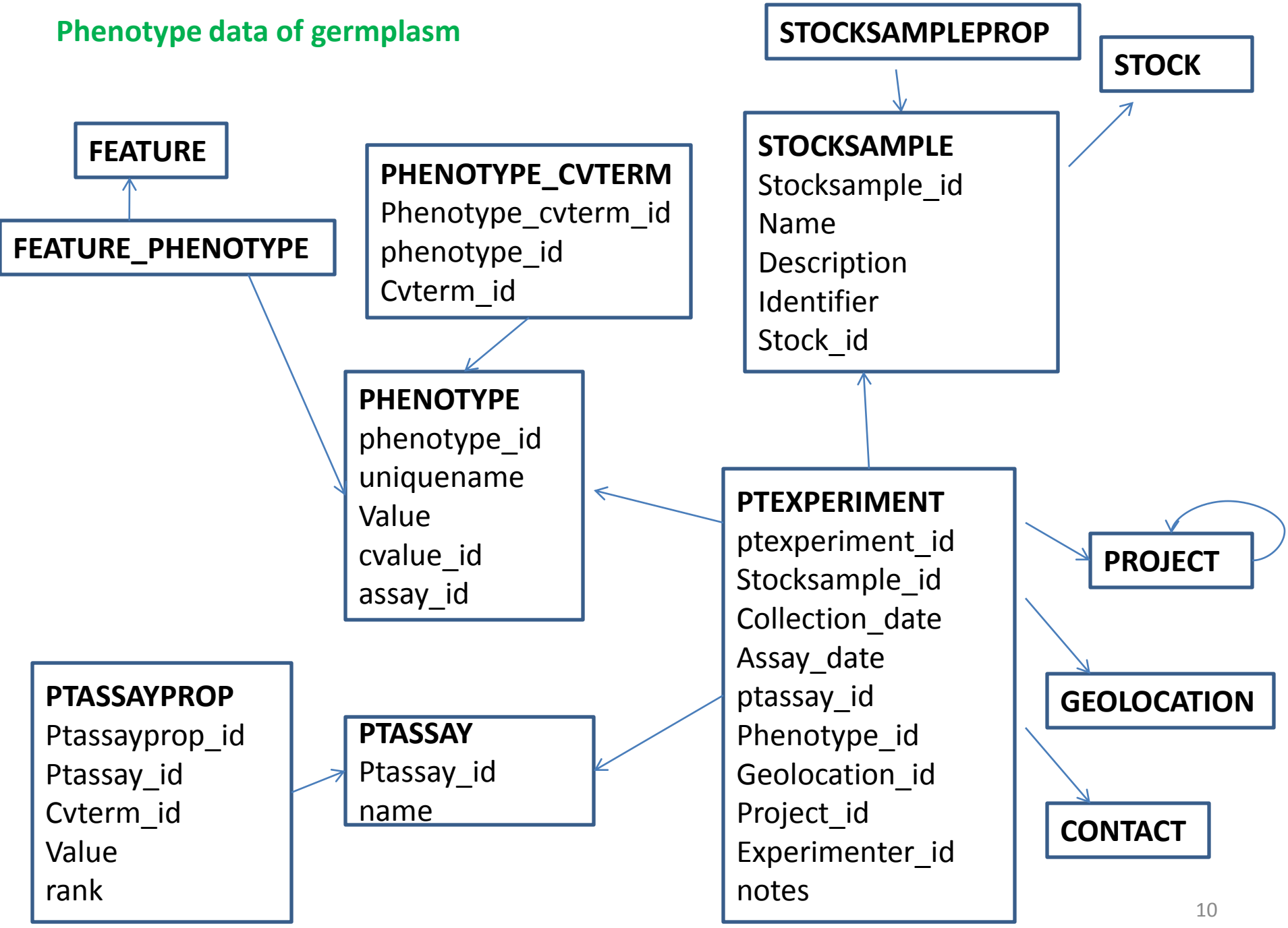
Observations for accession PI 590184

Molecular Data

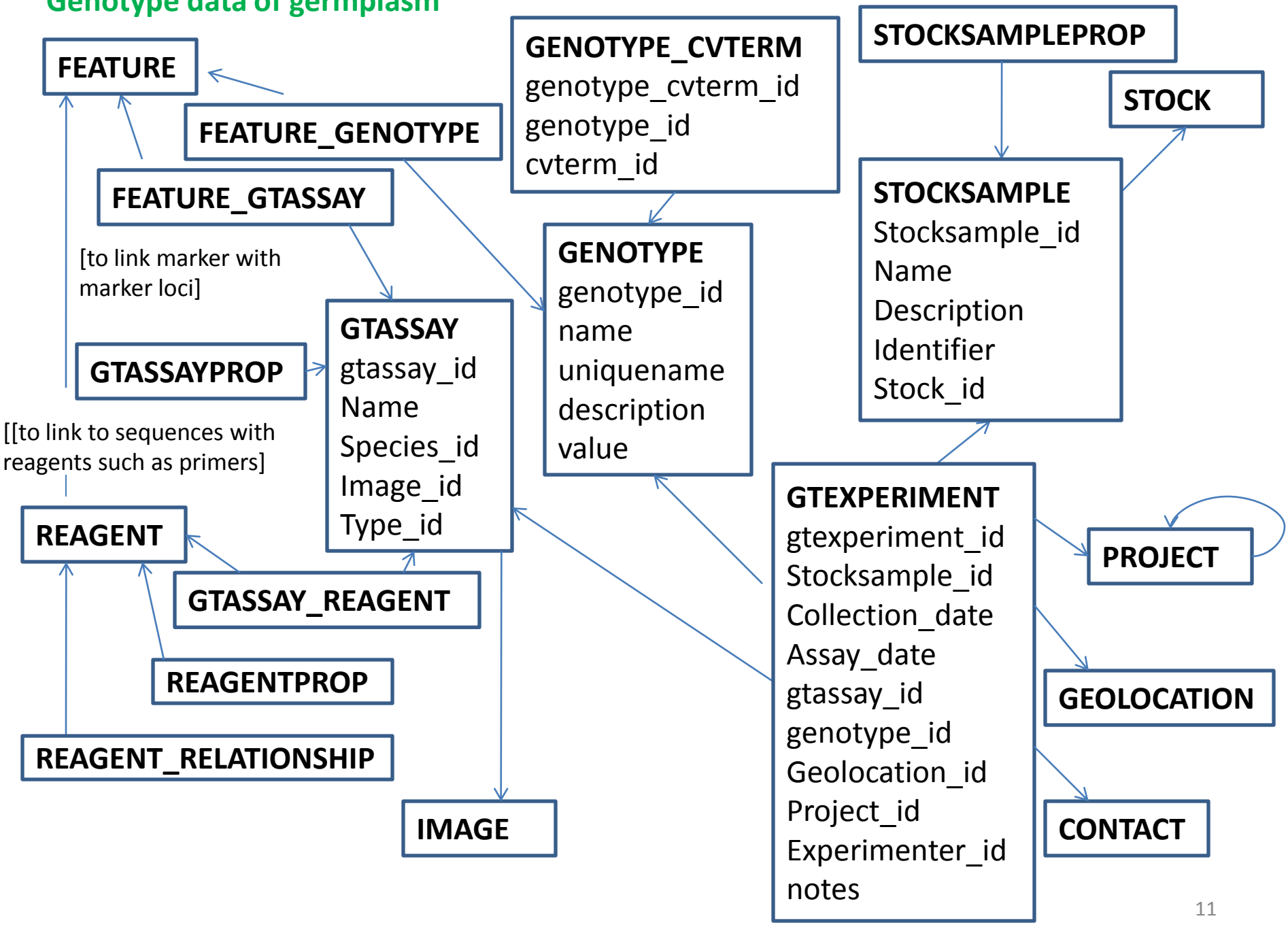
Poly Type	Marker	Value	Evaluation	Study Type	Inventory ID
Microsatellite (data in base pairs)	GD100A	230 230	2009MALUSDOMESTICADIVERSITY		PI 590184 .17 PL
Microsatellite (data in base pairs)	GD100A	240 240	2009MALUSDOMESTICADIVERSITY		PI 590184 .03 PL
Microsatellite (data in base pairs)	GD103A	84 110	2009MALUSDOMESTICADIVERSITY		PI 590184 .03 PL
Microsatellite (data in base pairs)	GD103A	84 110	2009MALUSDOMESTICADIVERSITY		PI 590184 .17 PL

1. PI590184 is stored in stock table
2. Two inventory IDs are stored stocksample table and linked to the respective genotype values through gtexperiment table.

Phenotype data of germplasm

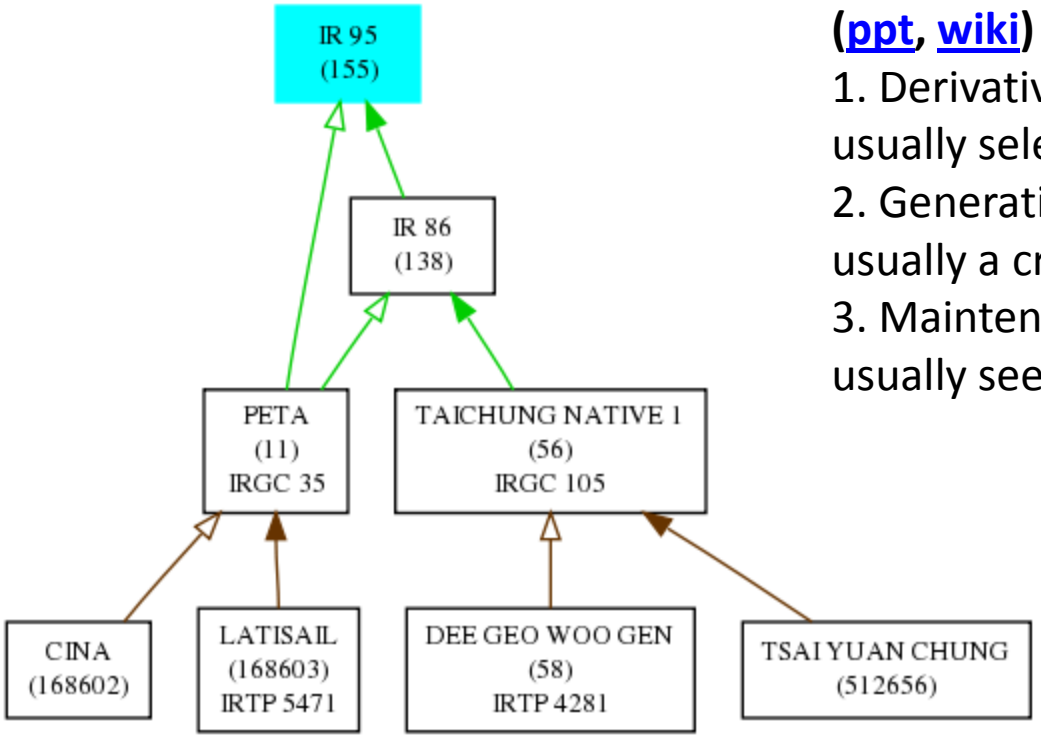


Genotype data of germplasm



Pedigree Data

(rice data generated from ICIS pedigree viewer)



Germplasm is created by methods below (ppt, wiki)

1. Derivative (to decrease genetic variation - usually selection)
2. Generative (to increase genetic variation - usually a cross)
3. Maintenance (to maintain genetic variation - usually seed increase or import)

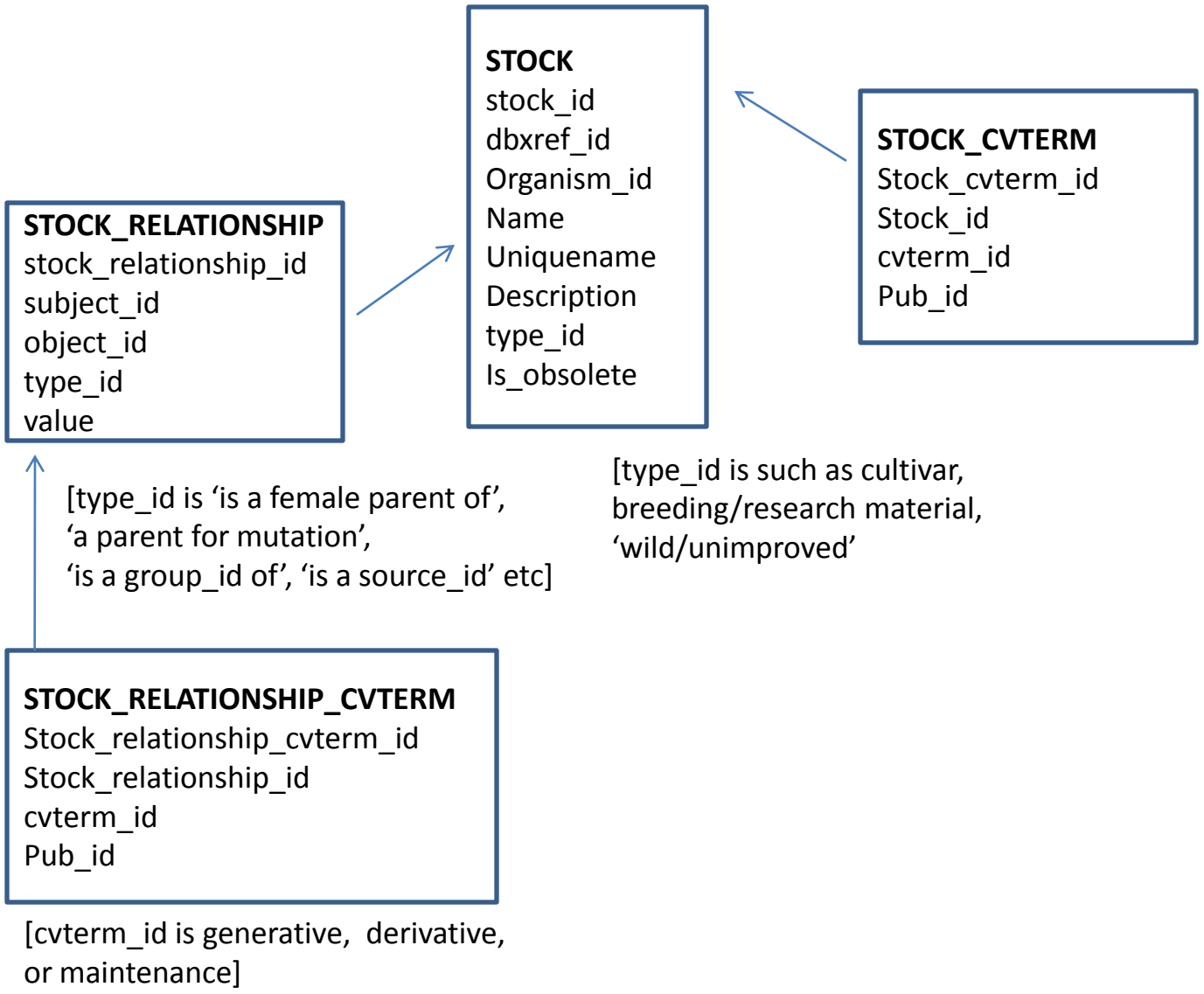
- We need to differentiate method 2 from 1 & 3 to display pedigree
- We need to track down the original parent

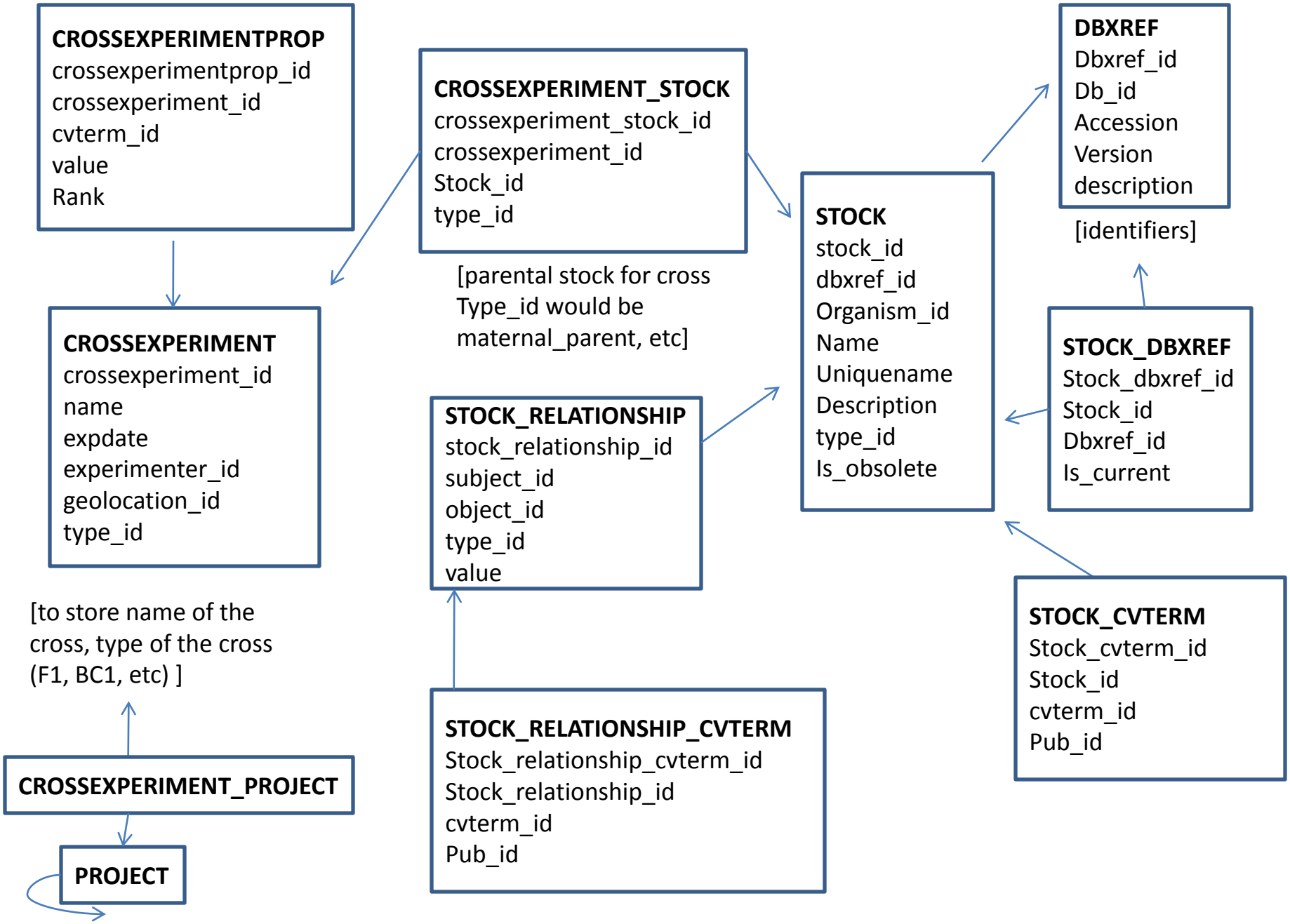
This pedigree shows only parentage, empty arrow is female parent and filled arrow is male parent. Green is for generative methods and brown is for maintenance. PETA went through maintenance step(s) and CINA and LATISAIL is the parents of the cross before those maintenance step(s).

In Germplasm Ontology

Group_ID: Identification code for last instance of germplasm produced by a generative process (cross) from which the current instance was derived.
Source_ID: Identification code for the immediate source of a derived or preserved germplasm. (Not valid for generated (by cross) germplasm.

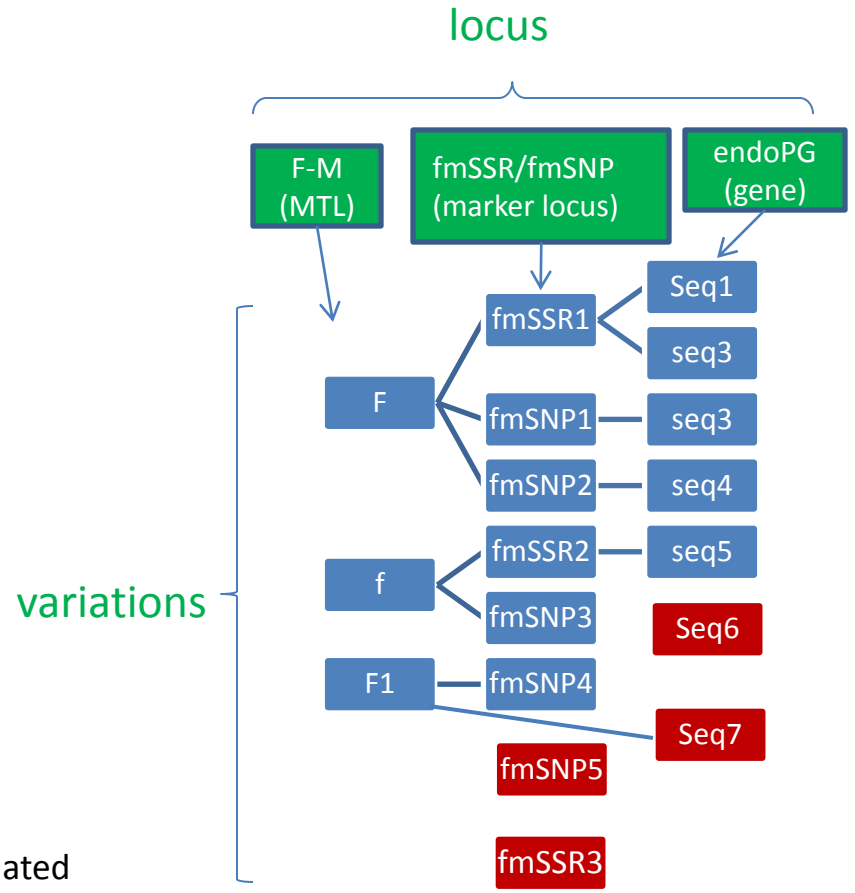
eg. When a Gala tree (Gala), generated by a cross, is imported to a WSU breeder, it will get a new germplasm_id (Gala_wsu), and the id of the original gala (Gala) would be group_id. When WSU gala is imported to a MSU breeder, the source_id would be Gala_wsu, and group_id would be Gala.


STOCK
stock_id
dbxref_id
Organism_id
Name
Uniquename
Description
type_id
Is_obsolete
STOCK_CVTERM
Stock_cvterm_id
Stock_id
cvterm_id
Pub_id
STOCK_RELATIONSHIP
stock_relationship_id
subject_id
object_id
type_id
value
[type_id is such as cultivar, breeding/research material, 'wild/unimproved'
[type_id is 'is a female parent of', 'a parent for mutation', 'is a group_id of', 'is a source_id' etc]
STOCK_RELATIONSHIP_CVTERM
Stock_relationship_cvterm_id
Stock_relationship_id
cvterm_id
Pub_id
[cvterm_id is generative, derivative, or maintenance]


CROSSEXPERIMENTPROP
crossexperimentprop_id
crossexperiment_id
cvterm_id
value
Rank
CROSSEXPERIMENT_STOCK
crossexperiment_stock_id
crossexperiment_id
Stock_id
type_id
DBXREF
Dbxref_id
Db_id
Accession
Version
description
[identifiers]
[parental stock for cross
Type_id would be
maternal_parent, etc]
STOCK
stock_id
dbxref_id
Organism_id
Name
Uniquename
Description
type_id
Is_obsolete
CROSSEXPERIMENT
crossexperiment_id
name
expdate
experimenter_id
geolocation_id
type_id
STOCK_DBXREF
Stock_dbxref_id
Stock_id
Dbxref_id
Is_current
STOCK_RELATIONSHIP
stock_relationship_id
subject_id
object_id
type_id
value
[to store name of the
cross, type of the cross
(F1, BC1, etc)]
STOCK_CVTERM
Stock_cvterm_id
Stock_id
cvterm_id
Pub_id
CROSSEXPERIMENT_PROJECT
PROJECT
STOCK_RELATIONSHIP_CVTERM
Stock_relationship_cvterm_id
Stock_relationship_id
cvterm_id
Pub_id

MTL (Mendelian Trait Loci), Genes, marker loci and their variation data

		Example
MTL		Freestone-Melting Flesh (F-M)
Functional alleles		F, f, and f1
Marker loci		fmSSR, fmSNP, etc
Marker loci alleles	w/o seq (SSR)	500 bp, 540 bp, etc
	w/ seq (SNP)	SNP sequences
Gene (gene product)		endoPG (endopolygalacturonase)
Sequence alleles		Nucleotide sequences for endoPG from NCBI



- One MTL can have more than one underlying genes/associated markers.
- Sequences that are associated with genes can be downloaded from NCBI and assembled (no need to make contigs or representatives) to make a gene set per species. The gene entry can have name and description (GO) and can be linked to multiple entries of sequences (most with associated cultivar data).

w/o known associated functional alleles and/or marker loci alleles

Storing MTL, Genes, marker loci and their variation data

Linked to stock

		Example	Feature* table (type**)	Phenotype table	Genotype table
MTL (and QTL)		Freestone-Melting Flesh	Name, PO/TO, position , and description (MTL/QTL)		
Functional alleles		F, f, and f1	Name (allele)	Name, PO, TO, and phenotypic value	
Marker loci		fmSSR, fmSNP, etc	Name and position (Marker locus)		
Marker loci alleles	w/o seq (SSR)	500 bp, 540 bp, etc			Name and description (500 bp, etc)
	w/ seq (SNP)	SNP sequences	Name and seq (sequence)		Name and description (C to T at 237)
Genes		endopolygalacturonase (endoPG)	Name and description (gene)		
Sequence alleles		Nucleotide sequences for for endoPG from NCBI	Name and seq (sequence)		Name (could be cultivar name + gene name?)

*feature, featureprop, featuremap, featurepos, featurerange, featiure_cvterm, etc

****For those non-SO, do we need to use feature_cvterm?**

- By feature_relationship table,
  - related MTL/QTL, genes, and marker loci can be linked
  - MTL and their functional alleles can be linked
- By feature_genotype table,
  - Marker Loci will be linked to marker variants
  - genes and sequence alleles can be linked
  - Functional alleles and the corresponding sequence alleles and marker variants can be linked
- By feature_phenotype table,
  - Functional alleles and their phenotypes can be linked

Do we need to store all the variation (even the marker variation without nucleotide sequences) in feature table?