Gramene Diversity Module

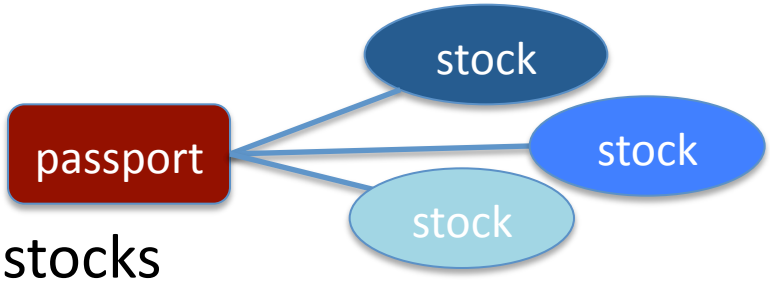
technical details

Genevieve DeClerck

10/5/2010, Gramene phone meeting

What is stored in Gramene Diversity

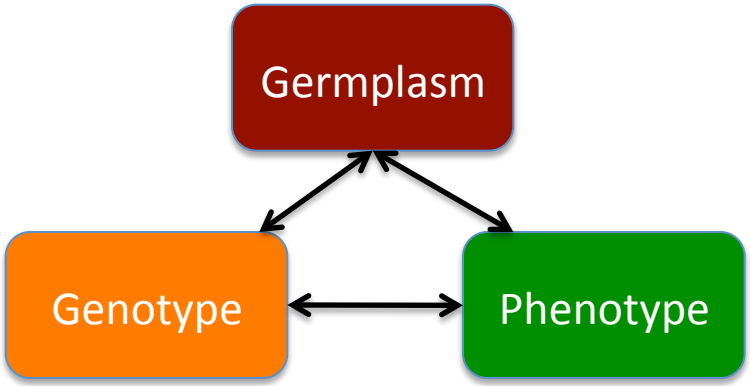
- **Germplasm**
  - passports, stocks



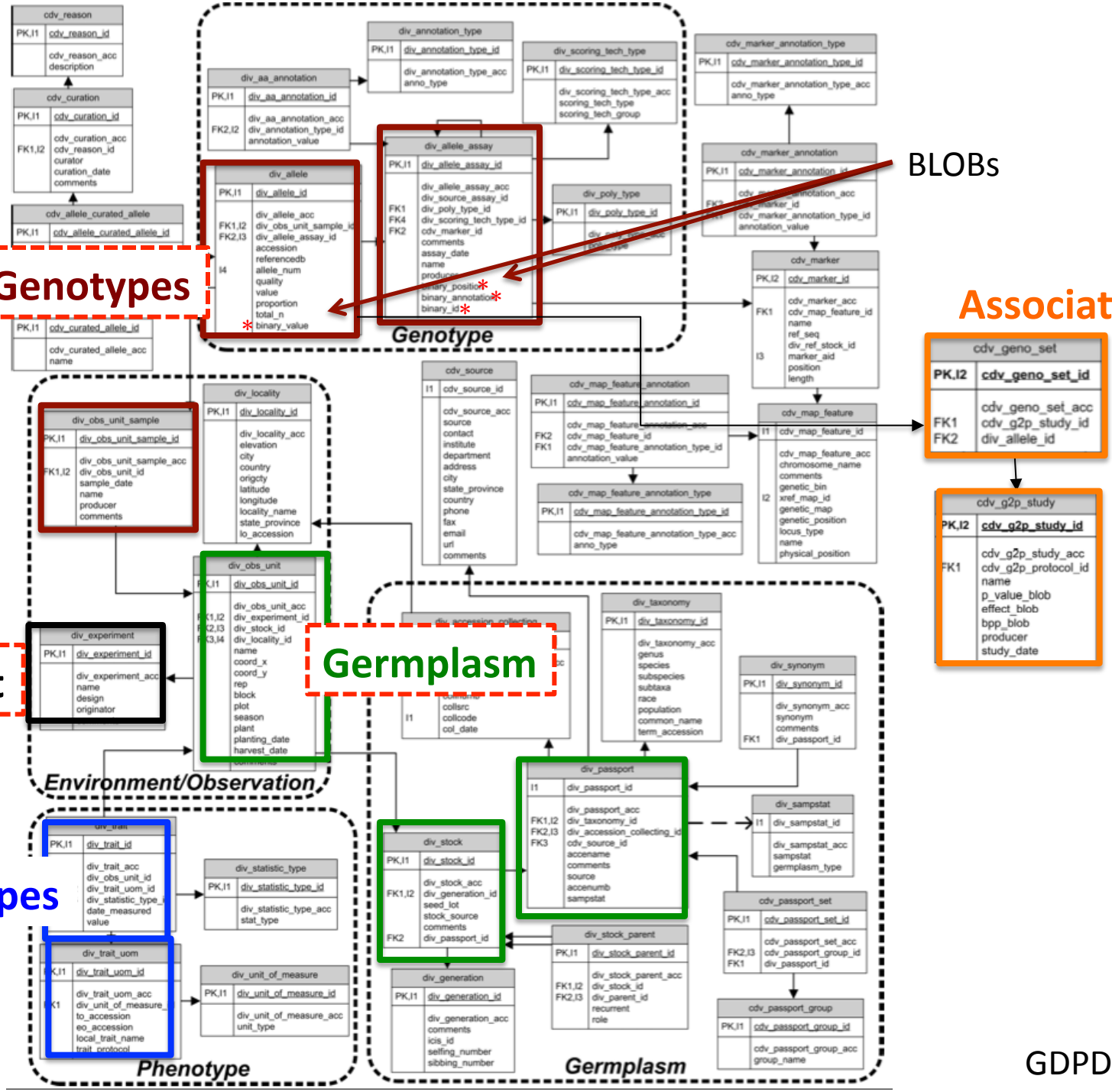
- **Genotype values**
  - SNPs, Insertion/Deletions

- **Phenotype measurements**

  - values (raw, mean, stddev), ontological terms, units of measure, field & replicate information
- **Association mapping data**
  - geno <-> pheno
- **Experiment information**
  - publication, meta data

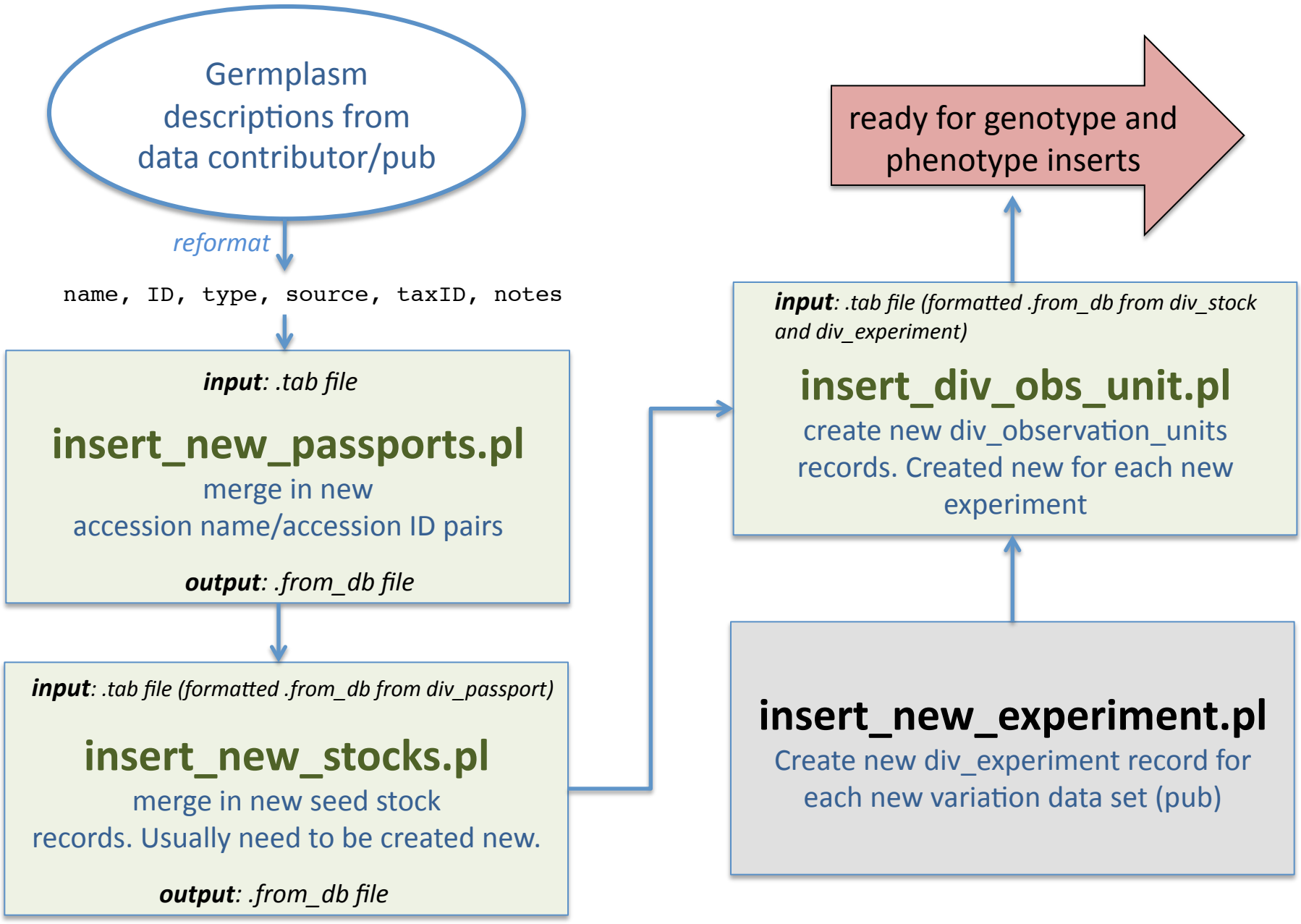


Diversity db central players



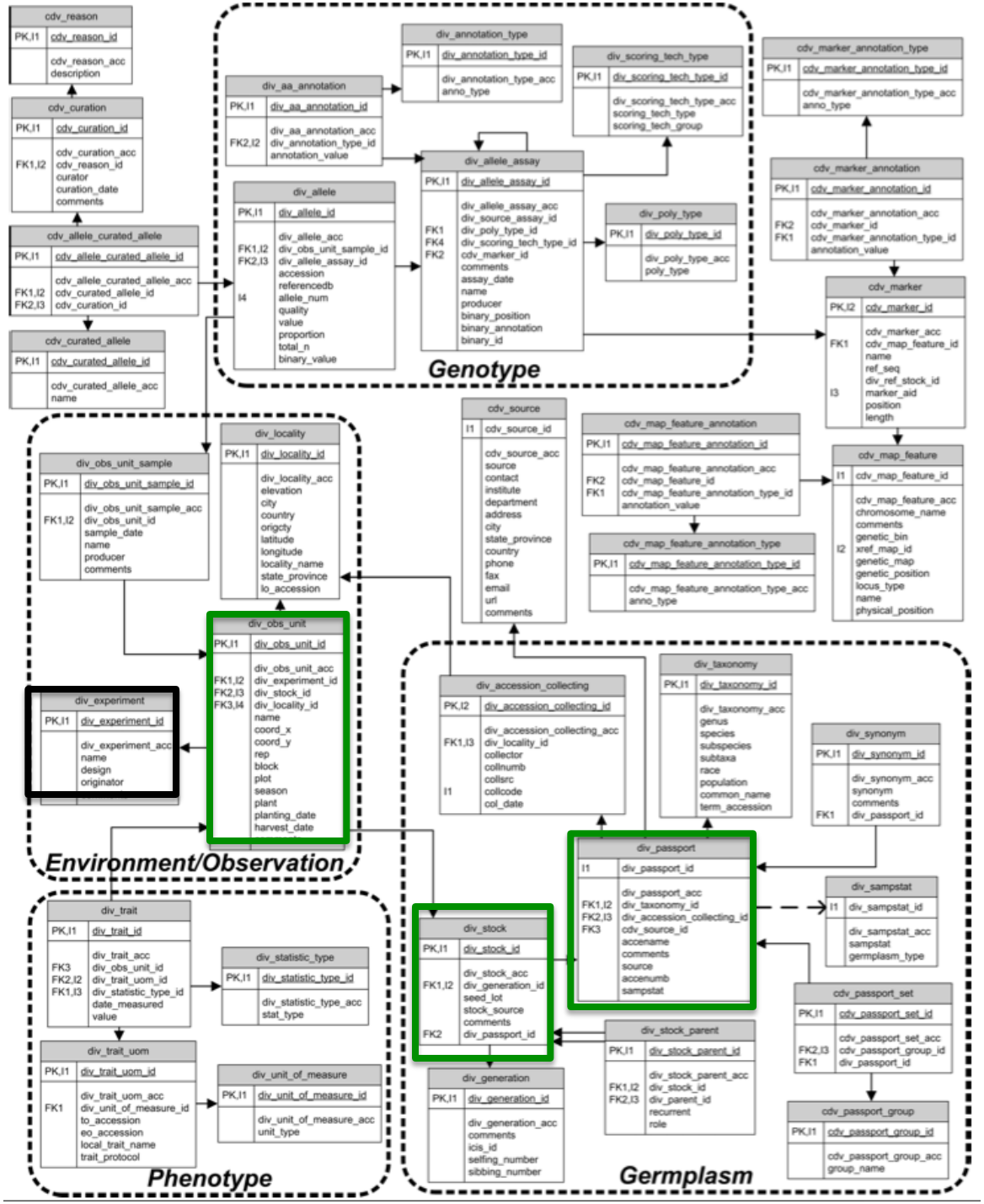
GDPDM4.2

Germplasm loading scripts



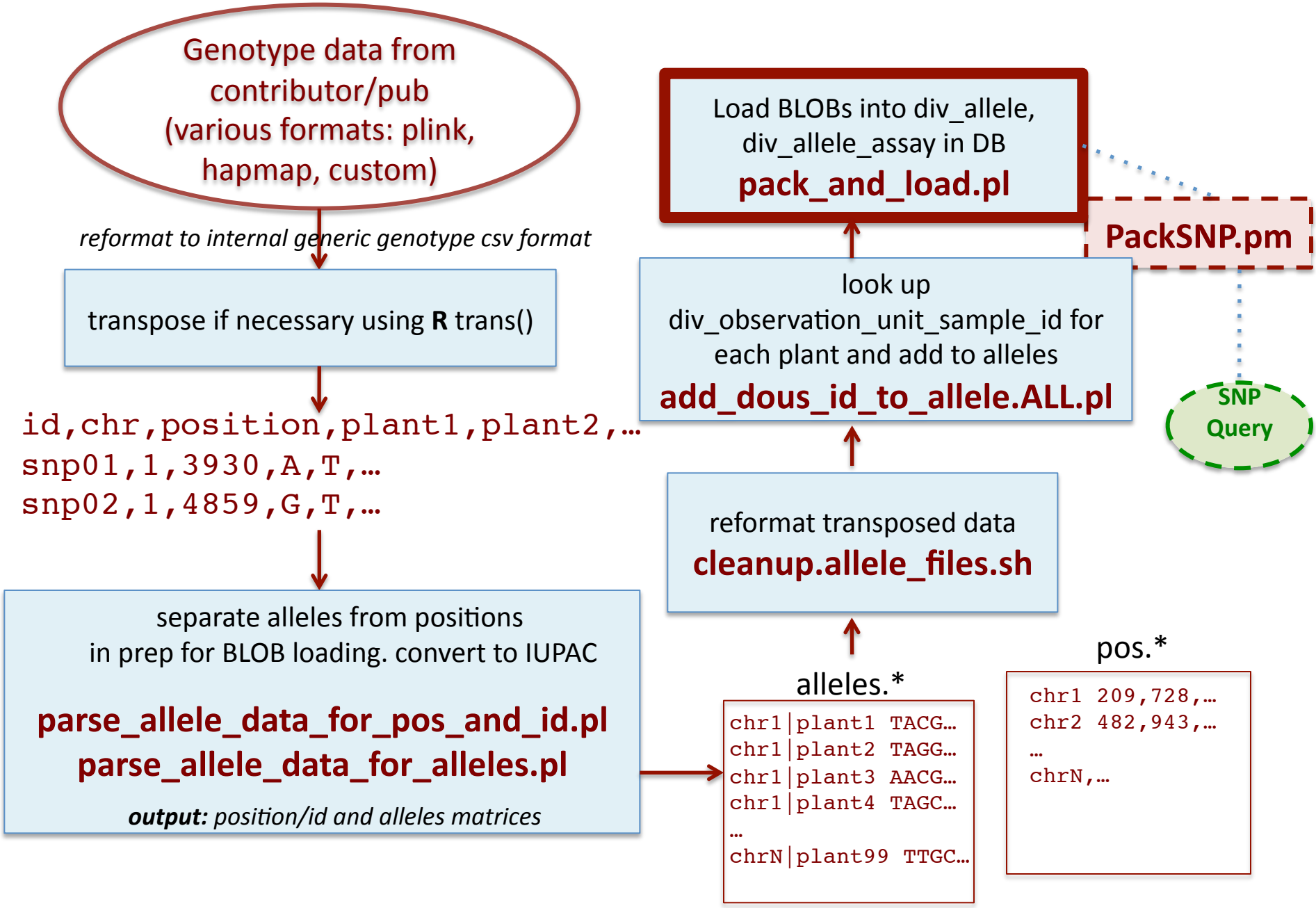
Experiment data

Germplasm data

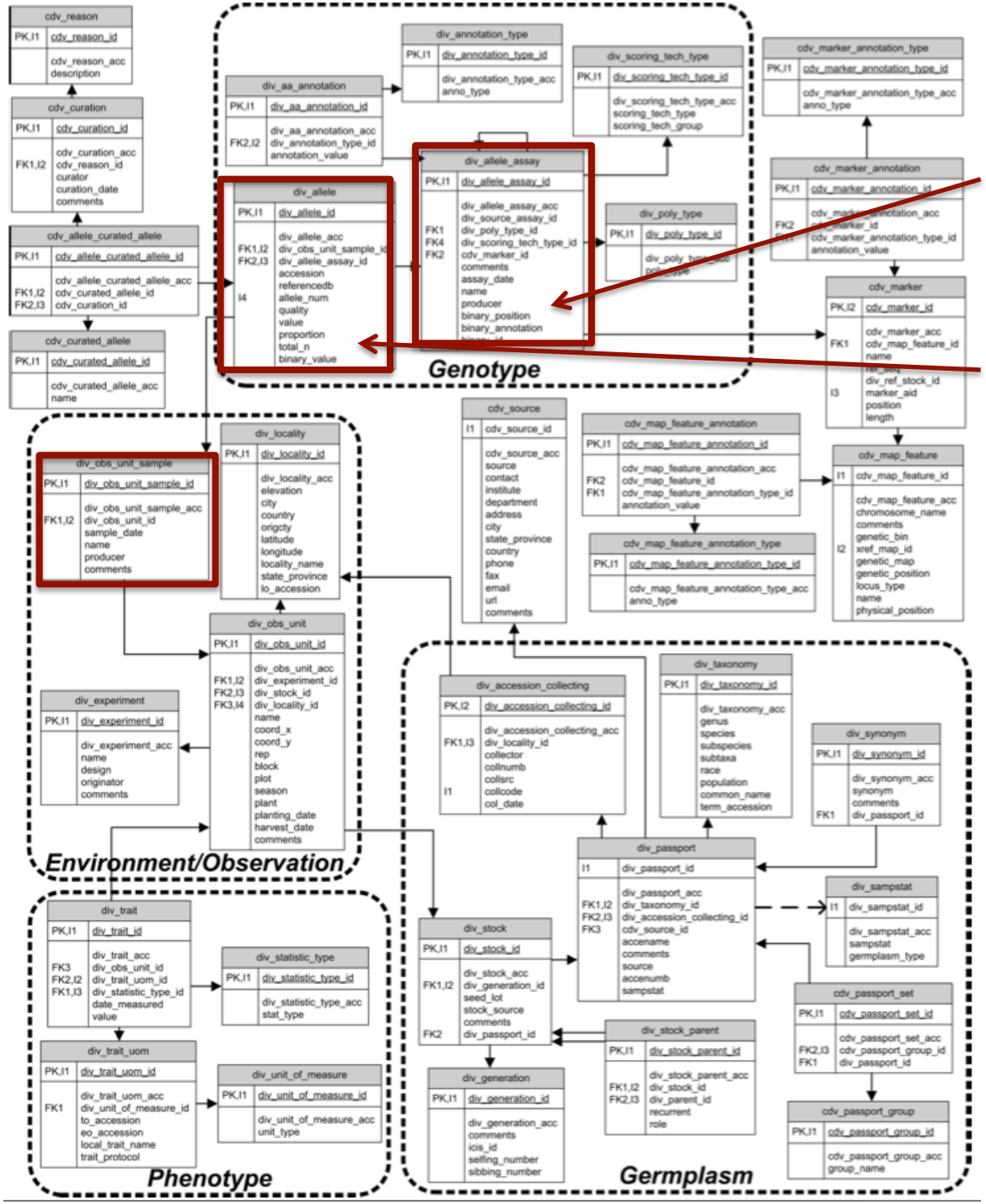


GDPDM

Genotype loading scripts



Genotype data



GDPDM

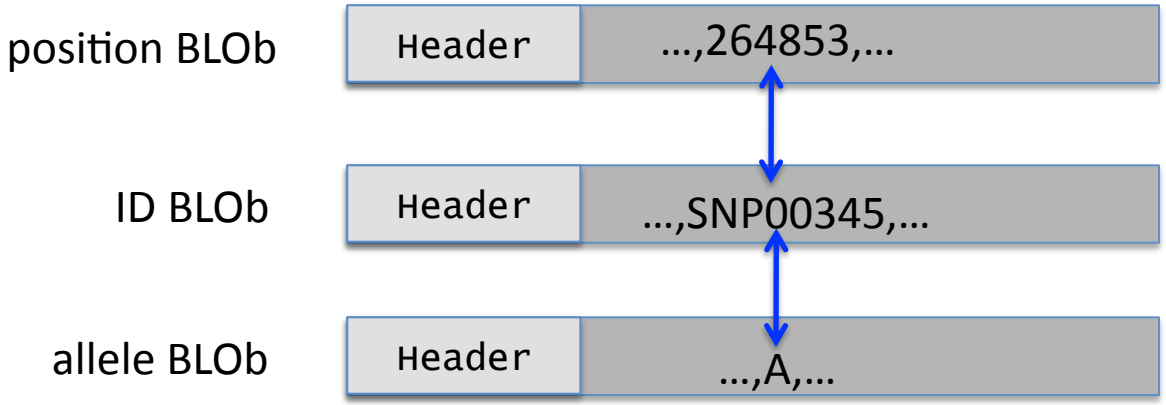
GDPDM BLOBs

Header

The BLOb header contains information about the data held in the Data part of the BLOb. e.g. chromsome#, genome version, plant accession (allele BLOb only), …

Data

The data in the a BLOb is ordered. The position of each binary data element relates to binary data positions in the other related BLObs.



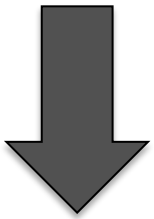
BLOB = Binary Large Object (binary DB data type)

Genotype data in database compressed by several orders of magnitude

For example,

100,000 SNPs x 100 maize lines

[**100,000** position data points and **100,000,000** allele call data points]



Compress data into GDPDM BLOBs

100,000 SNPs x 10 chromosomes = **10** position database records
100 maize lines x 10 chromosomes **= 1000** allele call database records

100,100,000 records → 1010 records

The number of db records is reduced by 10^6

This *vastly reduces query time* when mining large genotype and germplasm datasets in GDPDM

Diversity docs

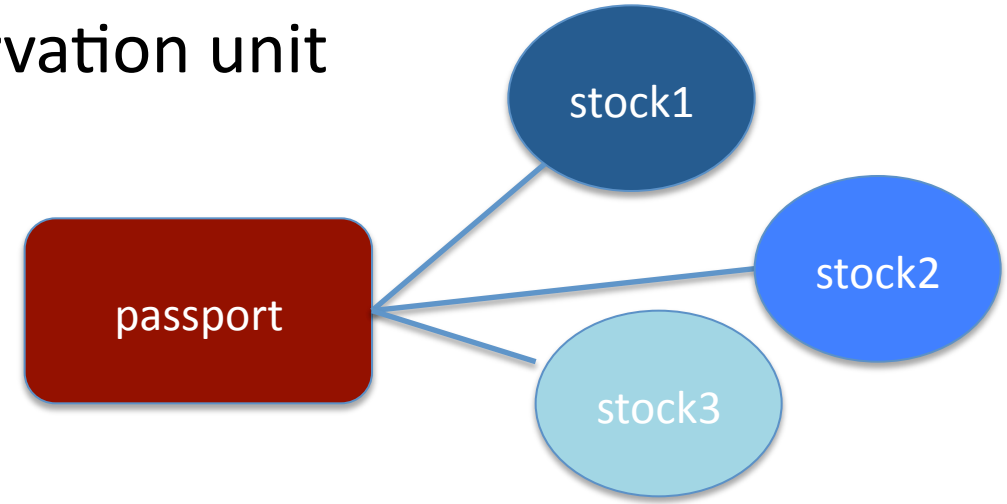
- Gwiki (http://gwiki.gramene.org/…)
  - Main
    - Category:Diversity
  - DB Loading
    - Insert_Germplasm_Data_into_GDPDM
    - Insert_Genotype_Data_in_GDPDM
    - IUPAC_nucleotide_codes
  - Data plan for release #32
    - Build32_Planning#Diversity

To do

- Release 32 (data):
  - Finish loading arabi Atwell data set (finalize pheno, insert associations)
  - Phenotype data from Karthik
  - Rice “1536” pheno data (2 traits)
  - Sorghum, wheat data – better visibility
  - Briana Gross data (Wash U)
  - Data from EnsemblGenomes
    - http://gwiki.gramene.org/Diversity/EnsVar_Databases
- Association data GDPDM pipeline
- Interface – phenotype, association data
- Consolidate/refine code, add to /usr/local/gramene/lib
- Continue documentation on Gwiki

Germplasm

- **passport**: unique seed bank accession
  - unique pair: *accession name* & *accession ID*
- **stock**: instance of a passport used in a diversity experiment
  - seed lot, source
  - links to observation unit



Packing SNP data in GDPDM

Example: allele, ID, and position BLOBs for 3 maize lines (B73, B79, CML103), on chromosome 9