Sue Rhee

Carnegie Institution for Science

Plant Biology Department

Stanford, California

What can bio-ontologies do for us?

1. Increases searchability
   - Uncouples from the editorial style of authors, consistency across databases
2. Enables complex queries
   - Semantic web, ‘smart queries’, or simple queries with complex answers?
3. Enables quantitative comparison
   - Allows groupings, enrichment analysis, cross-species comparisons
4. Enables predictions
   - benchmarks for functional similarity measures, ‘omics’ data integration

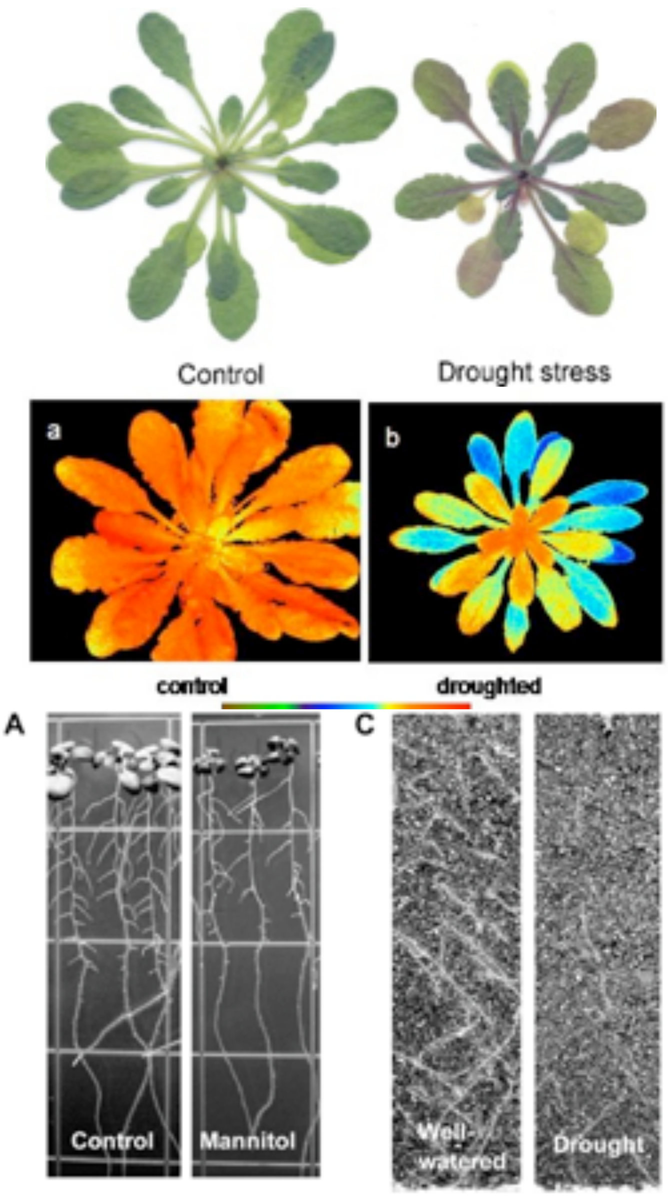
Questions and Goals

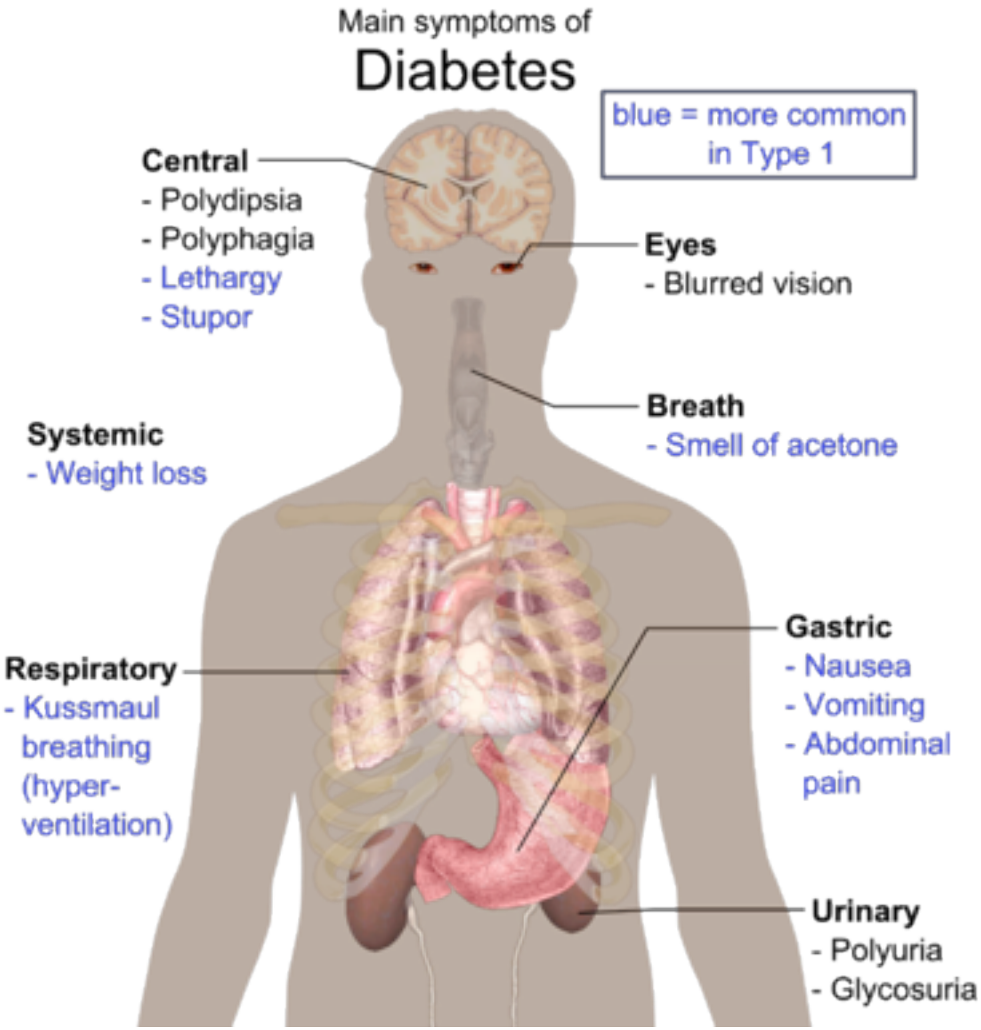
1. How do we **describe** and define biological processes and functions to allow comparison, modeling, and prediction?

2. How do we **find** all the genes that are involved in a biological process?

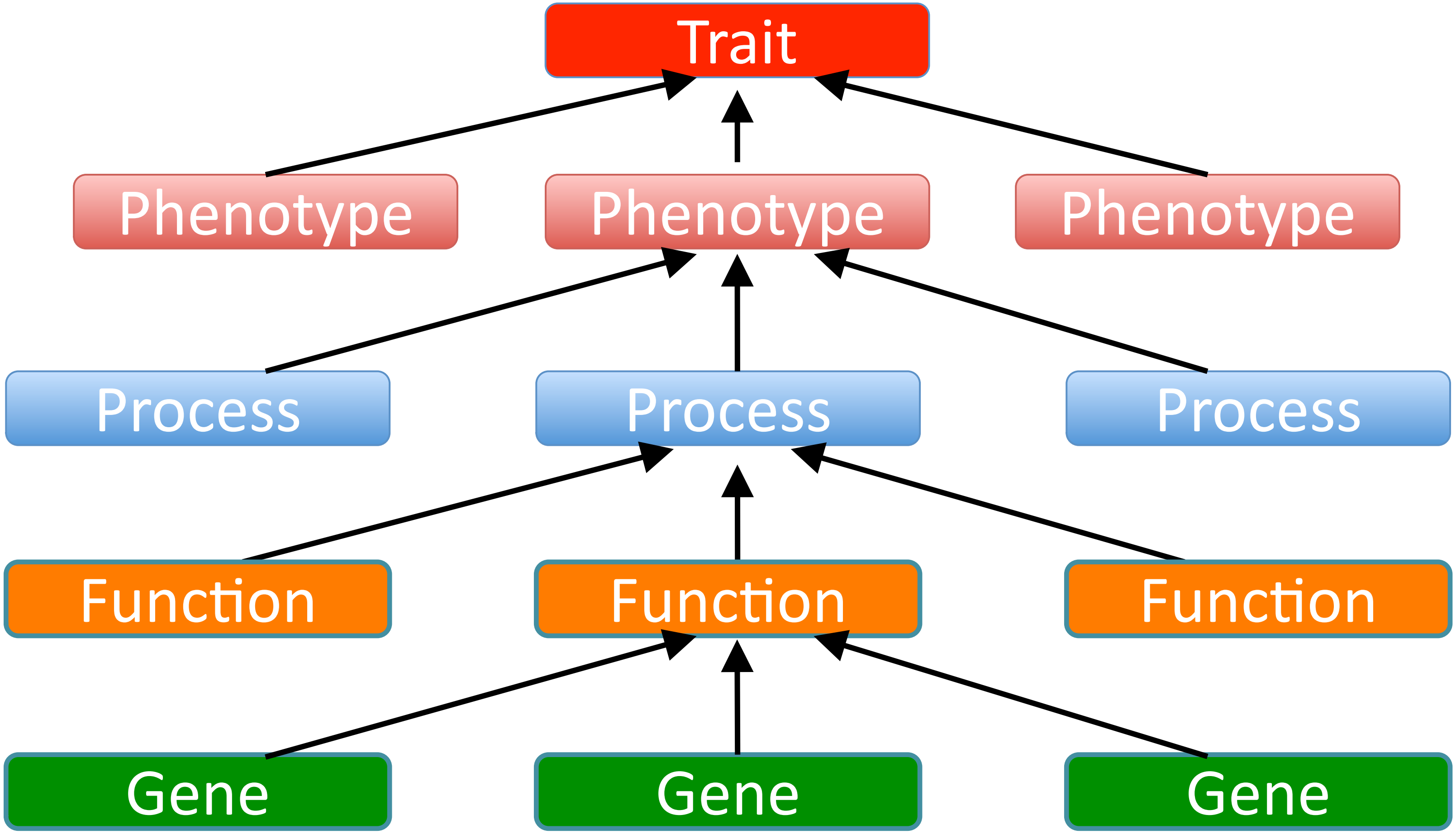
3. How do we **model** the functions, processes, and phenotypes to explain complex traits and predict phenotypes from genotypes?

A trait exhibits multiple phenotypes.

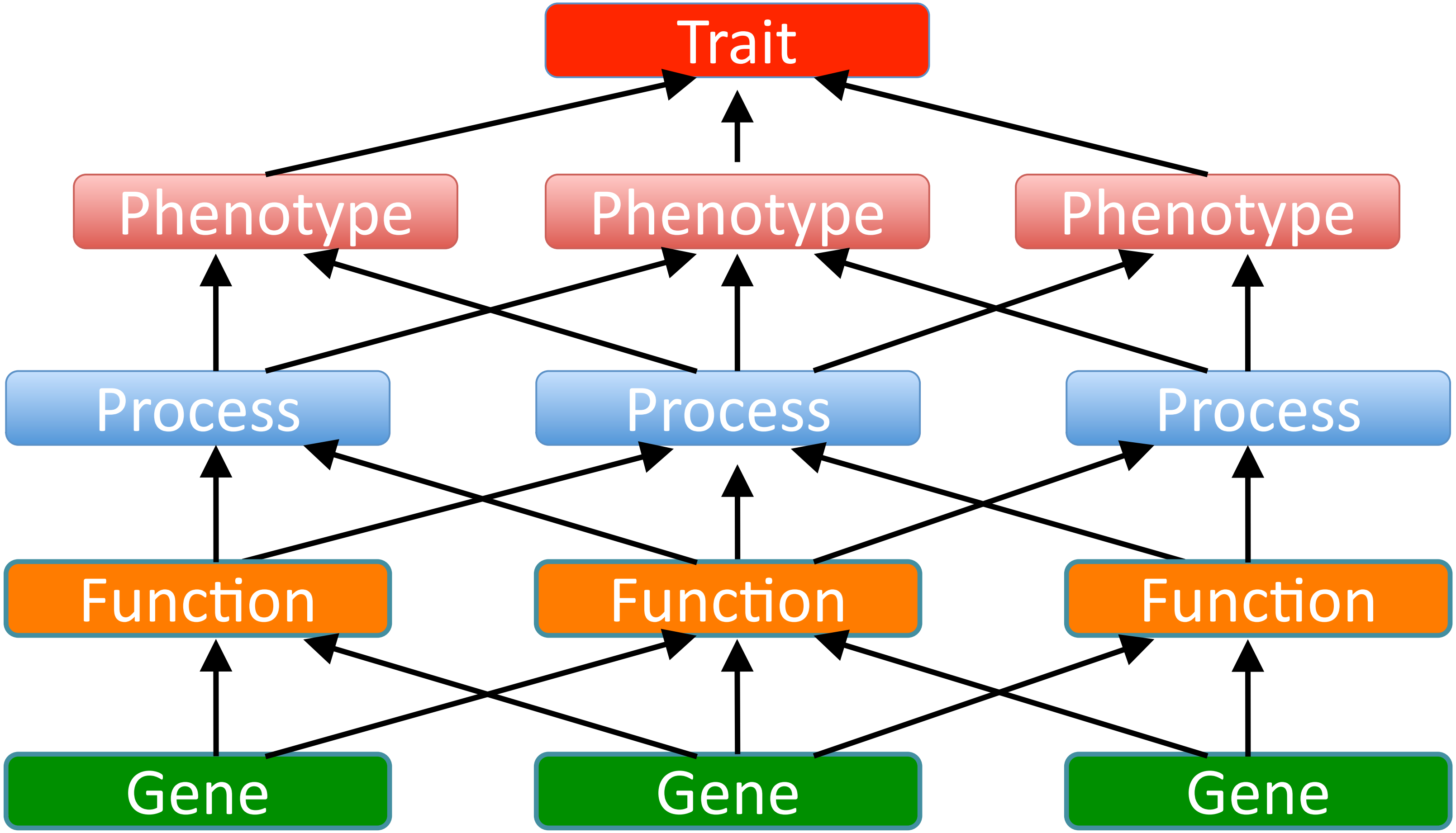




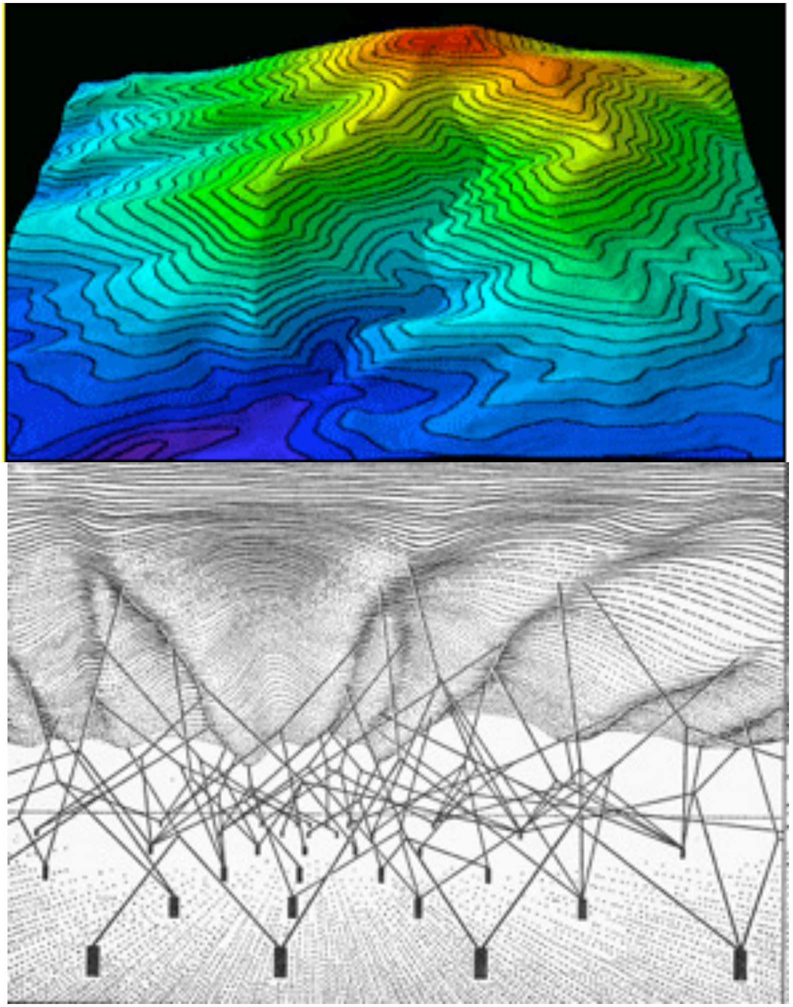
What are the molecular mechanisms that control a trait?



What are the molecular mechanisms that control a trait?



Phenotypes are manifested by underlying molecular networks.



CH Waddington (1957) The Strategy of the Genes. George Allen & Unwin Publishing.

Questions and Goals

1. How do we describe and define biological processes and functions to allow comparison, modeling, and prediction?

2. How do we find all the genes that are involved in a biological process?

3. How do we model the functions, processes, and phenotypes to explain complex traits and predict phenotypes from genotypes?

Bio-Ontologies

A hierarchy of terms each with a precise definition, identifier, and relationship to other terms

Gene Ontology

- Molecular Function

  What a product does at a biochemical level

- Biological Process

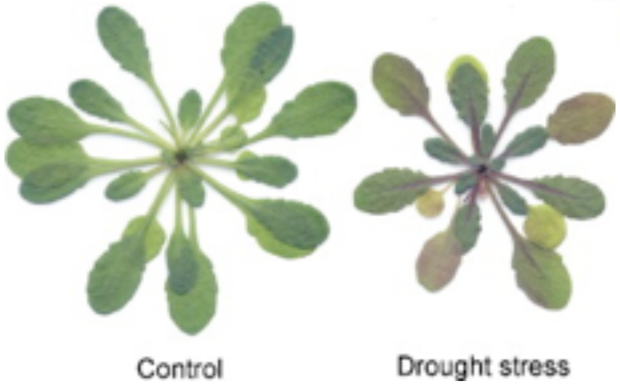
  Biological goal, accomplished via one or more ordered assemblies of molecular functions

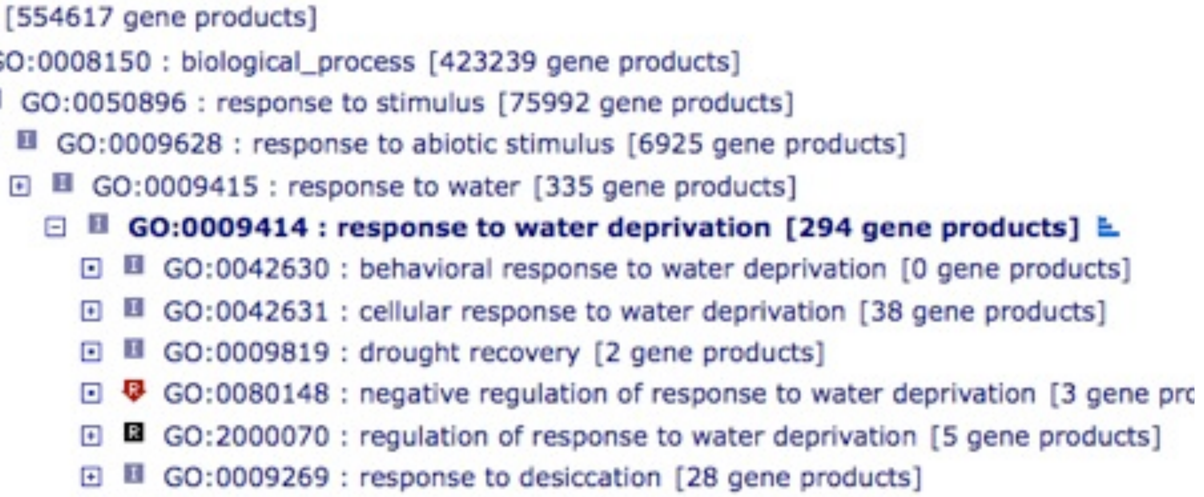
- Cellular Component

  Where in the cell a product is located

Gene Ontology Consortium (geneontology.org)

Gene Ontology –Biological Process




⊞ all : all [554617 gene products]
⊞ I GO:0008150 : biological_process [423239 gene products]
⊞ I GO:0050896 : response to stimulus [75992 gene products]
⊞ I GO:0009628 : response to abiotic stimulus [6925 gene products]
⊞ I GO:0009415 : response to water [335 gene products]
⊟ I **GO:0009414 : response to water deprivation [294 gene products]**
⊞ I GO:0042630 : behavioral response to water deprivation [0 gene products]
⊞ I GO:0042631 : cellular response to water deprivation [38 gene products]
⊞ I GO:0009819 : drought recovery [2 gene products]
⊞ R GO:0080148 : negative regulation of response to water deprivation [3 gene prc
⊞ R GO:2000070 : regulation of response to water deprivation [5 gene products]
⊞ I GO:0009269 : response to desiccation [28 gene products]

- Response to water deprivation (GO:0009414)
  - Photosynthesis (GO:0015979)
  - Anthocyanin biosynthesis (GO:0009718)
  - Stomatal closure (GO:0090332)
  - Leaf development (GO:0048366)
  - Root development (GO:0048364)

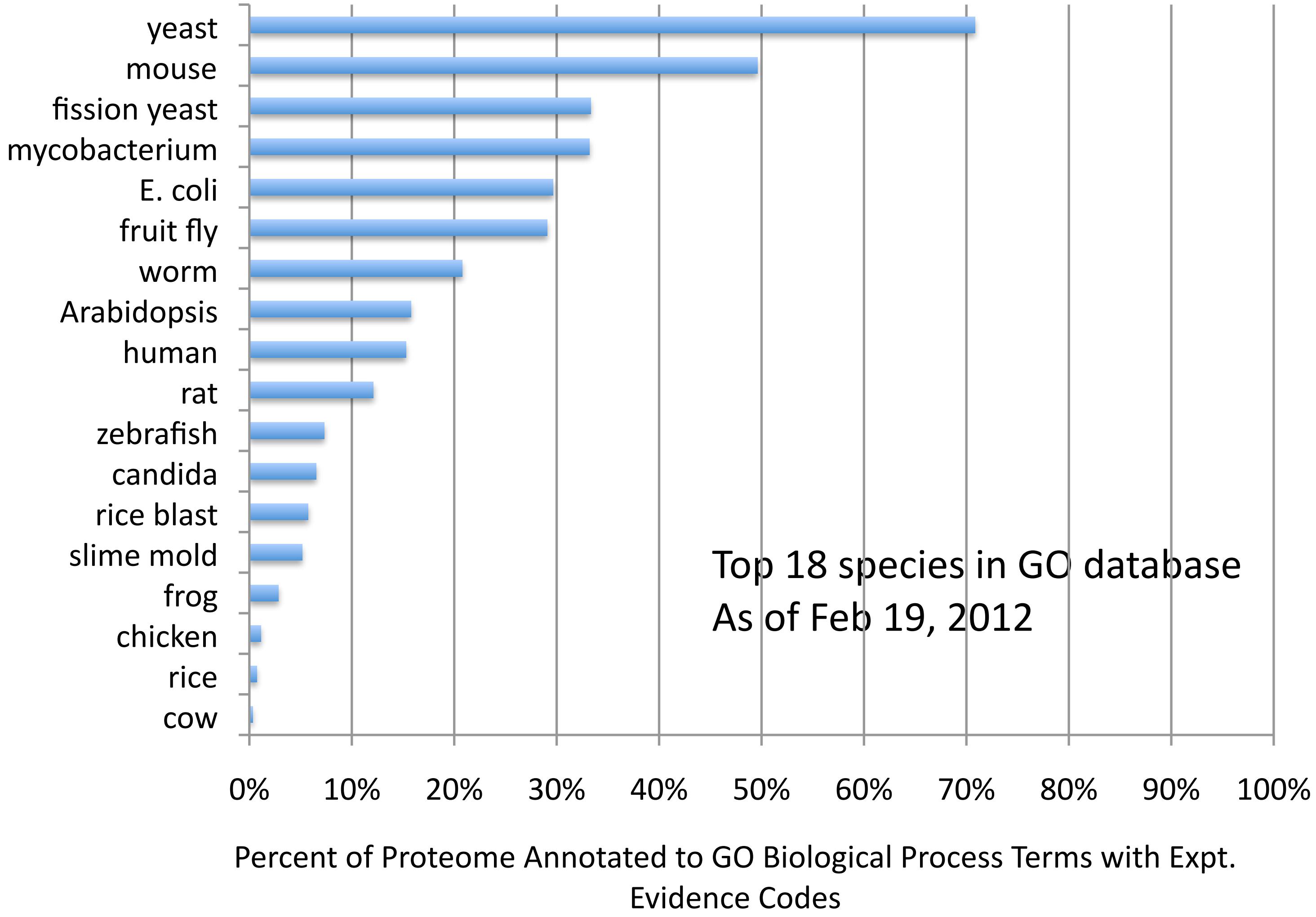
Questions and Goals

1. How do we describe and define processes and functions to allow comparison, modeling, and prediction?

2. How do we find all the genes that are involved in a biological process?

3. How do we model the functions, processes, and phenotypes to explain complex traits and predict phenotypes from genotypes?

We still don't know what most genes are doing in most organisms.

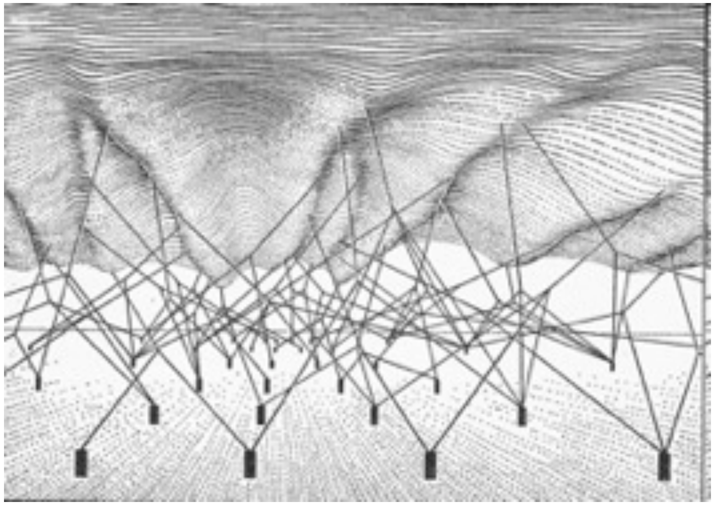


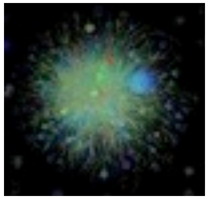
Top 18 species in GO database
As of Feb 19, 2012

QuickGo

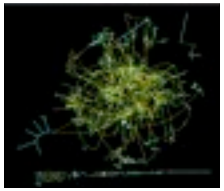
Friday, February 24, 2012

Reconstruction of Biological Networks

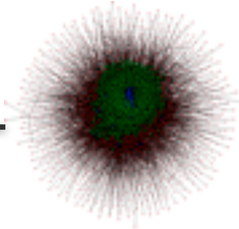




Gene co-function network (AraNet)
Lee et al (2010) Nat. Biotech.
Huang et al (2011) Nat. Protocols
Chae e al (2012) Curr. Op. Plant Biol.



Metabolic network (PlantCyc, AraCyc, etc.)
Zhang et al (2010) Plant Physiol.



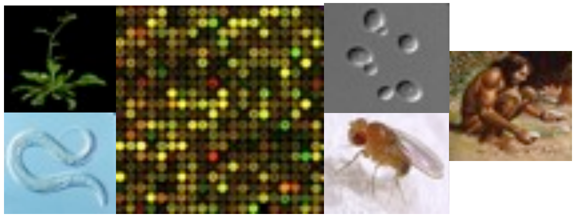
Protein-interaction network (membrane/signaling proteins)
Lalonde et al (2010) Frontiers in Plant Physiol.

AraNet: A Genome-Wide Co-function Network for Arabidopsis

Gold-standard data

- ~5000 genes annotated to biological processes with experimental support
- ~65000 links between genes annotated to be involved in the same process
- generated by TAIR curators

Large-scale 'omics' data



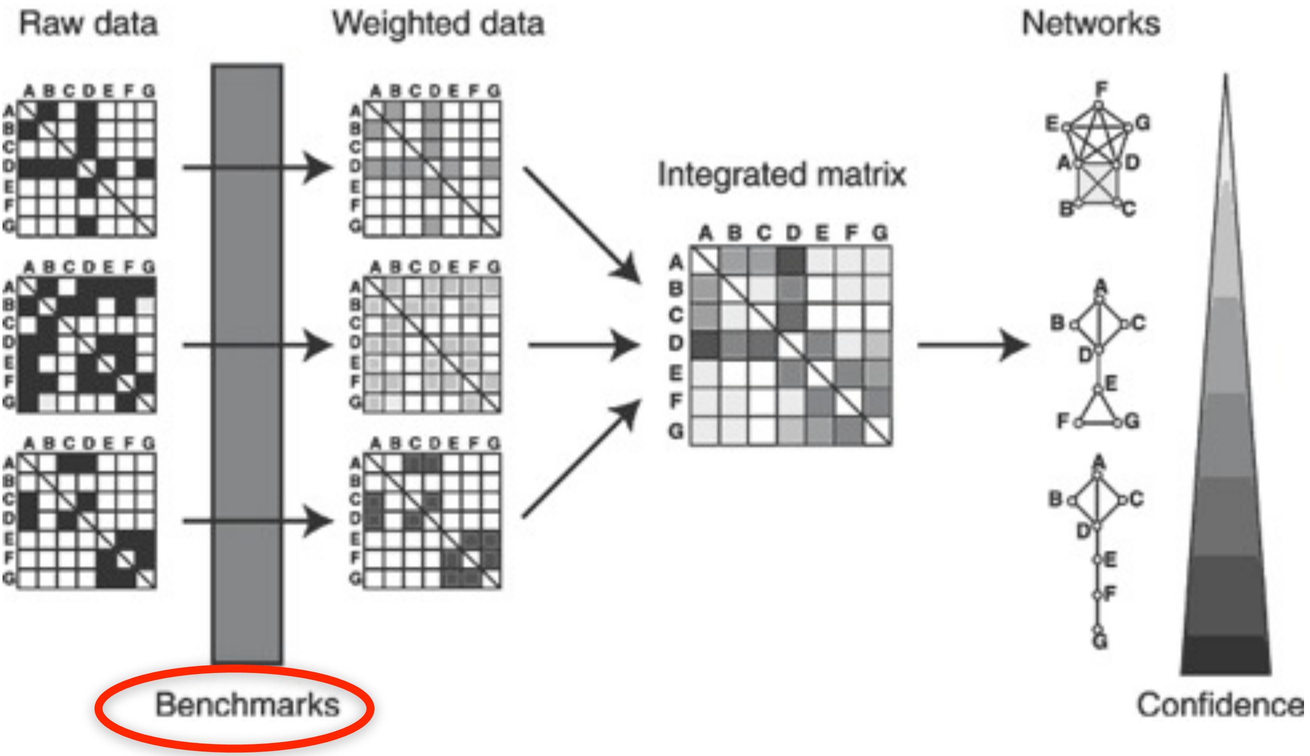
- Co-expression of Arabidopsis gene pairs or homolog pairs
- protein domain co-occurrence of gene pairs
- shared phylogenetic profile of homolog pairs
- genomic proximity of homolog pairs
- protein interactions of homolog pairs
- genetic interactions of homolog pairs
- co-citation of homolog pairs
- ~50 million data points

General Methodology of Building a Co-function Network from Large-scale Data

gene-expression
(e.g. microarrays)

protein interaction
(e.g. yeast 2-hybrid)

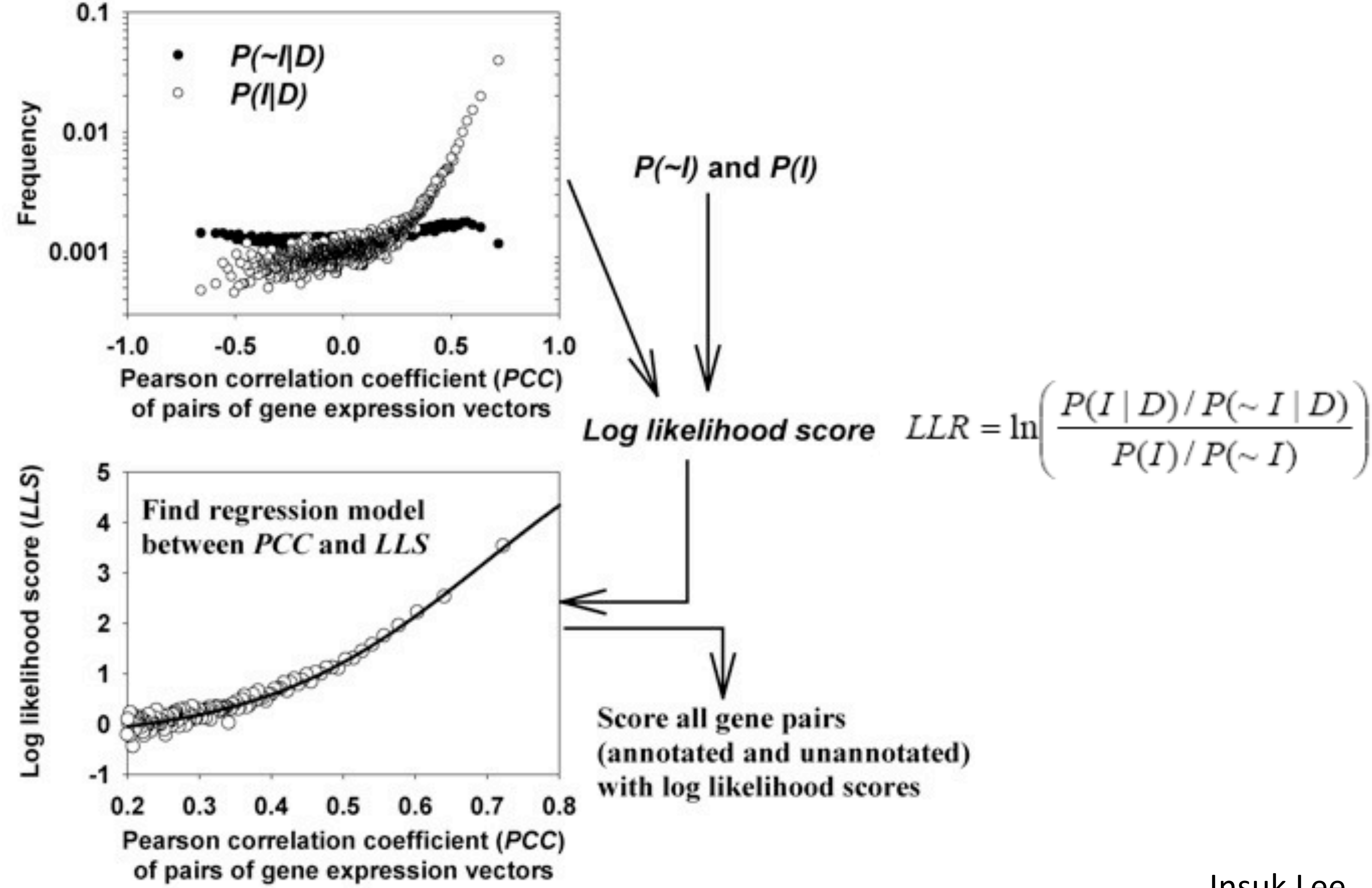
genetic interaction



Experimentally derived GO annotations from TAIR

Fraser & Marcotte (2006)
Nat Genetics

Calculating Functional Similarity of Gene Pairs



P(~I) and *P(I)*

Log likelihood score

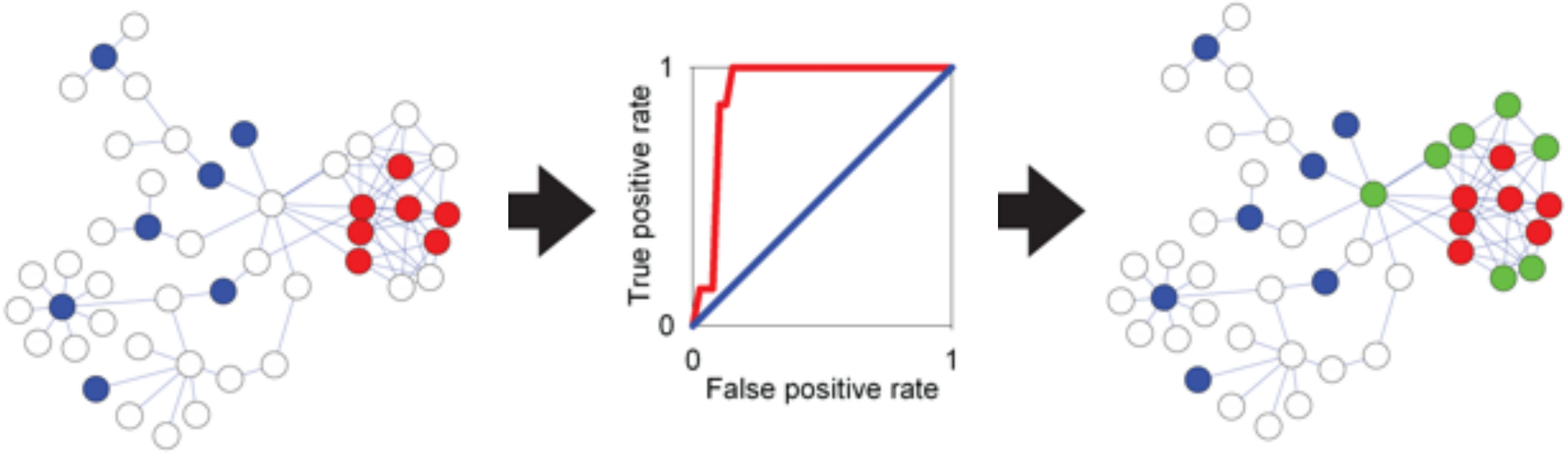
$$LLR = \ln\left(\frac{P(I \mid D) / P(\sim I \mid D)}{P(I) / P(\sim I)}\right)$$

Find regression model between *PCC* and *LLS*

Score all gene pairs (annotated and unannotated) with log likelihood scores

Insuk Lee

Predicting New Genes Based on Known Genes

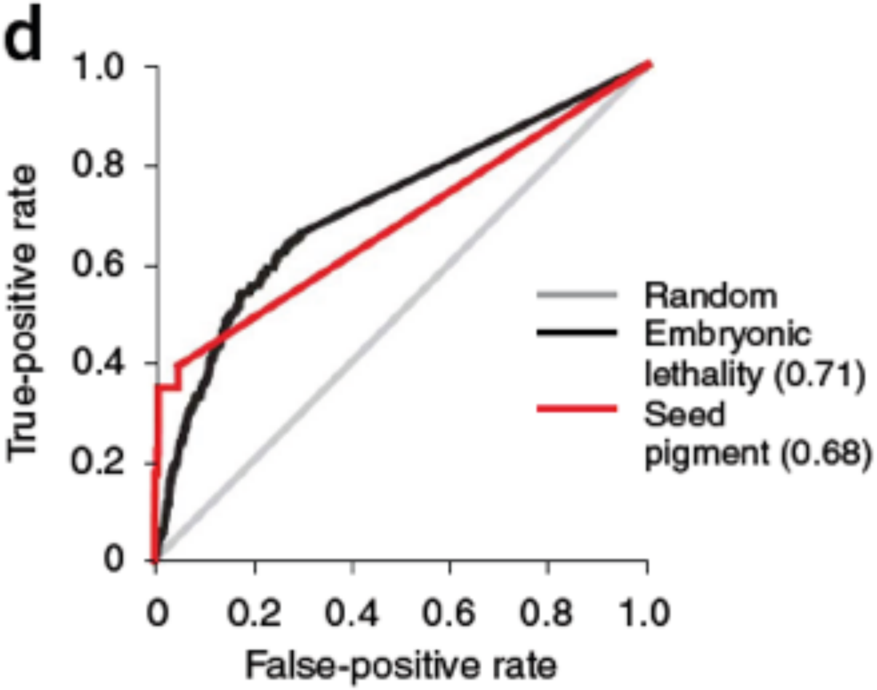


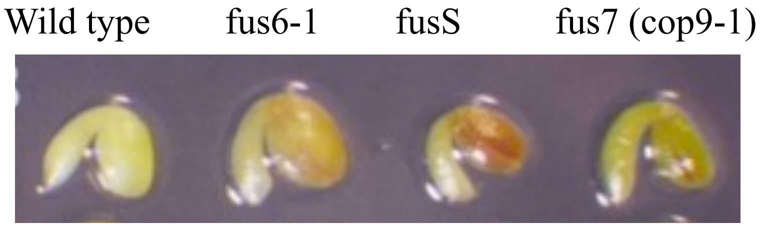
1. Rank all the genes based on connectivity to the bait genes

2. Repeat the calculations using randomized bait genes

🔴 Genes known to be involved in the same pathway
🔵 Same bait genes in randomized AraNet
🟢 New genes that might be associated with the pathway

Known Seed Pigmentation Defective Genes

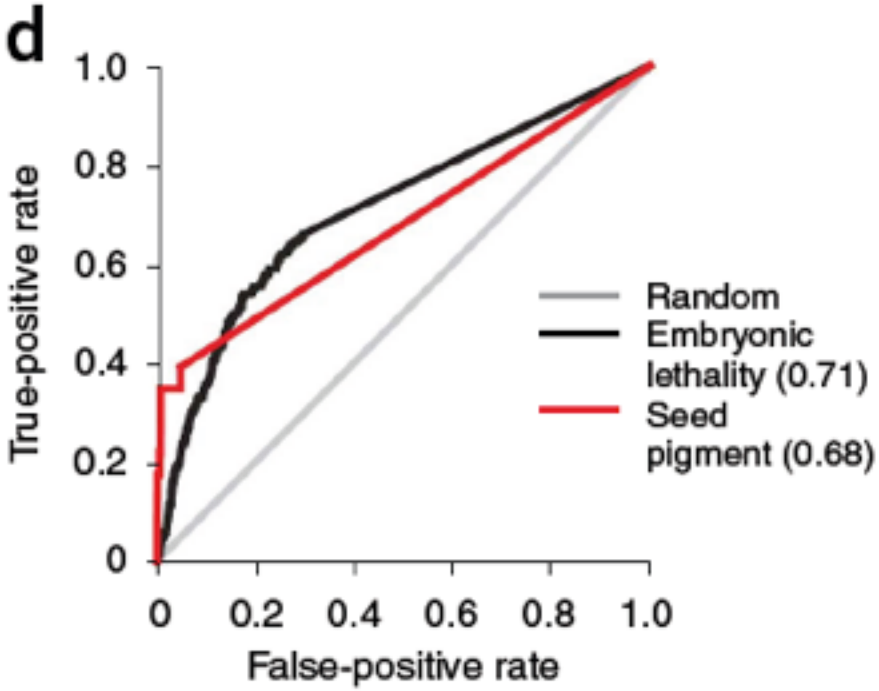


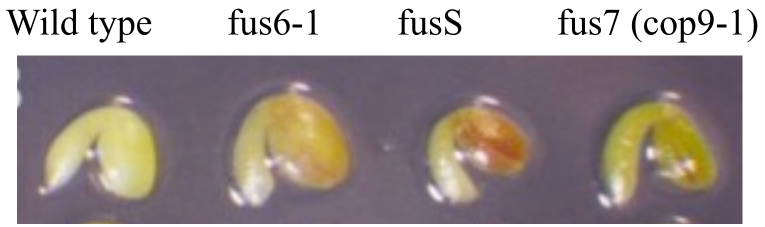


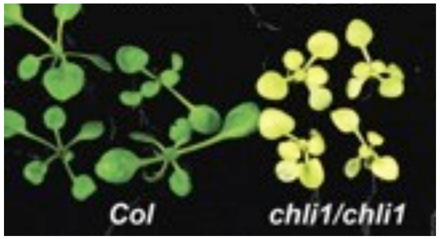
- 23 confirmed seed embryo pigmentation genes.

www.seedgenes.org (Meinke Lab)

Known Seed Pigmentation Defective Genes







- 23 confirmed seed embryo pigmentation genes.

www.seedgenes.org (Meinke Lab)

Screening for Seed Embryo Pigmentation Genes

23 confirmed seed embryo pigmentation genes.

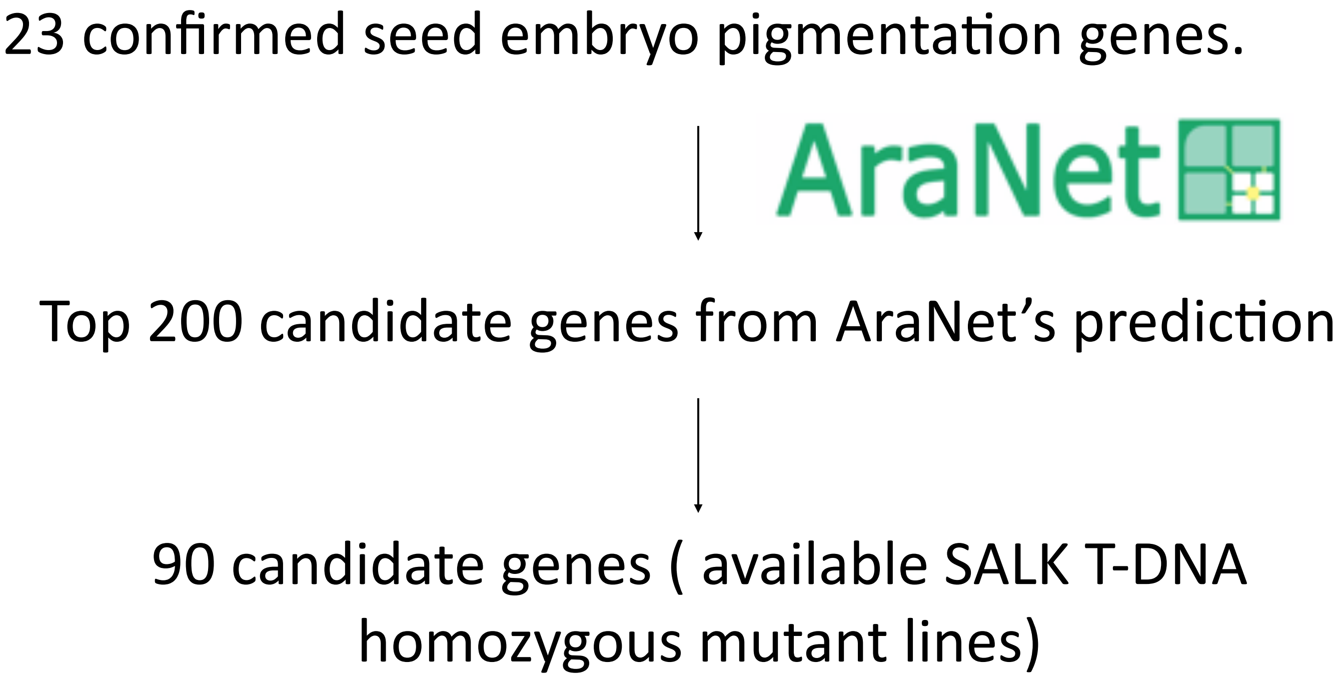
Screening for Seed Embryo Pigmentation Genes

23 confirmed seed embryo pigmentation genes.

↓ AraNet

Top 200 candidate genes from AraNet's prediction

Screening for Seed Embryo Pigmentation Genes



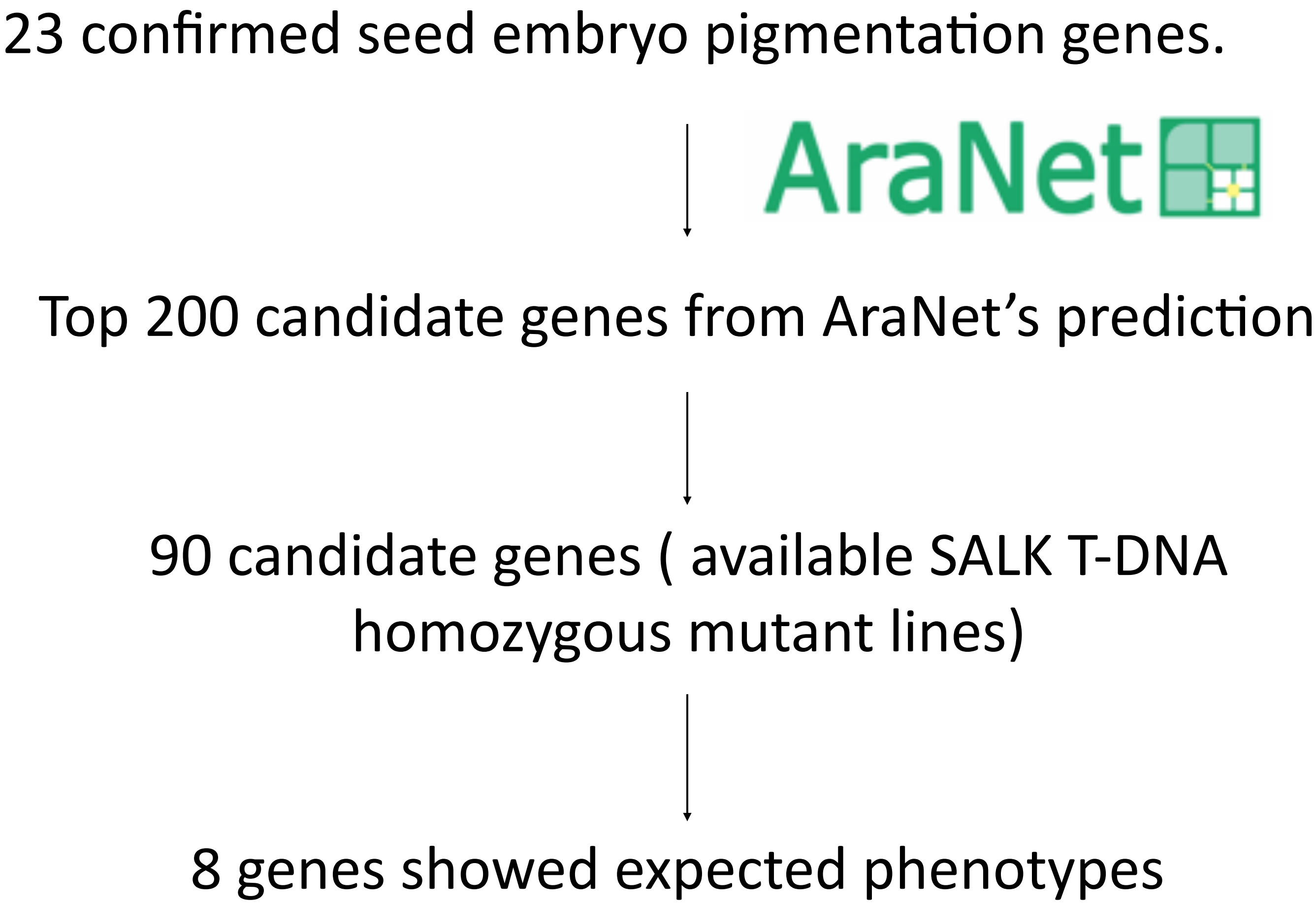
23 confirmed seed embryo pigmentation genes.

AraNet

Top 200 candidate genes from AraNet's prediction

90 candidate genes (available SALK T-DNA homozygous mutant lines)

Screening for Seed Embryo Pigmentation Genes



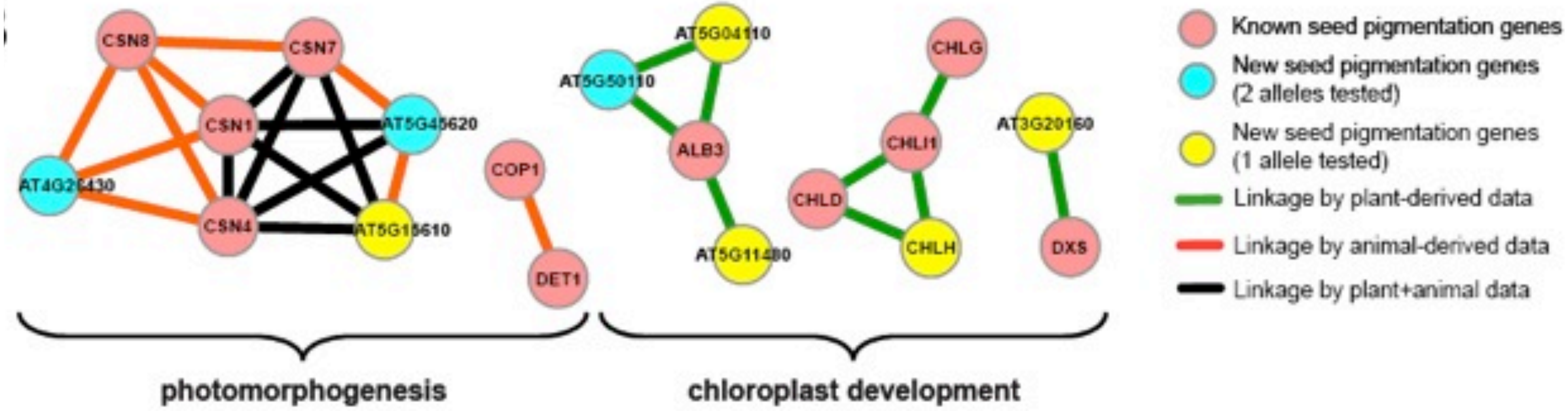
23 confirmed seed embryo pigmentation genes.

Top 200 candidate genes from AraNet's prediction

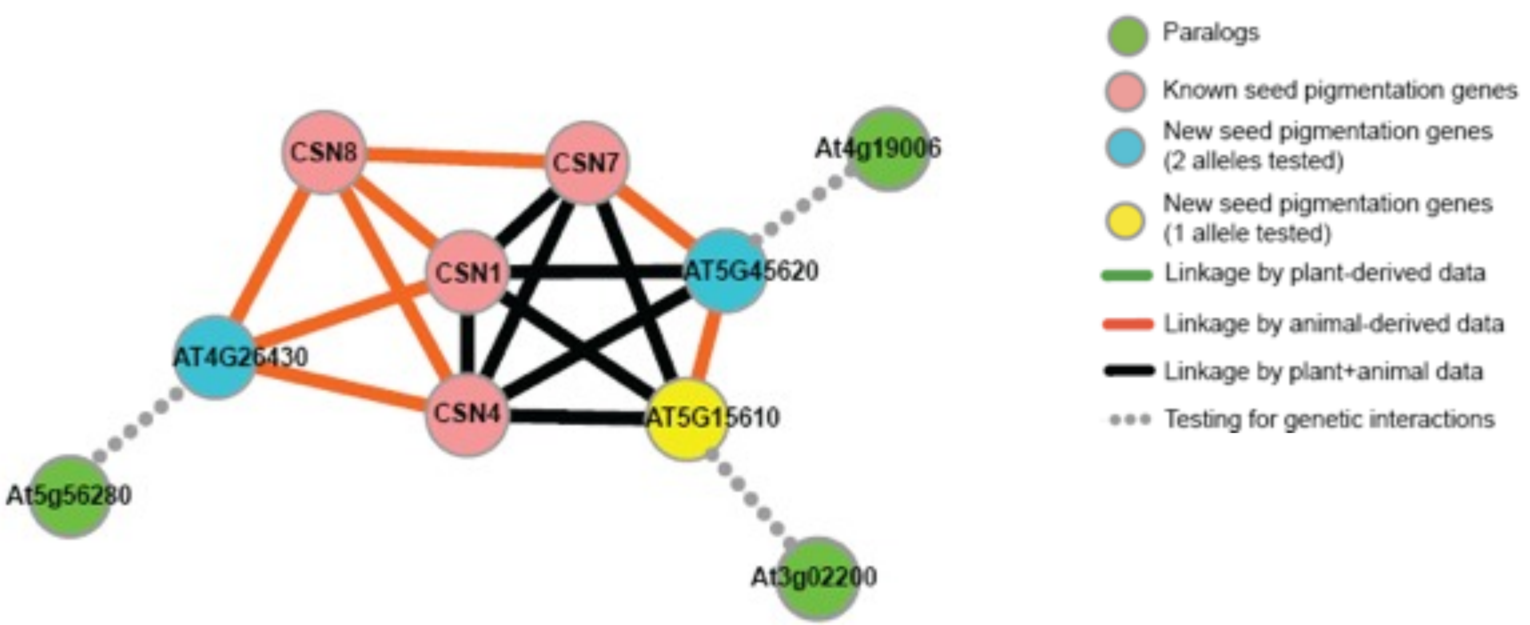
90 candidate genes (available SALK T-DNA homozygous mutant lines)

8 genes showed expected phenotypes

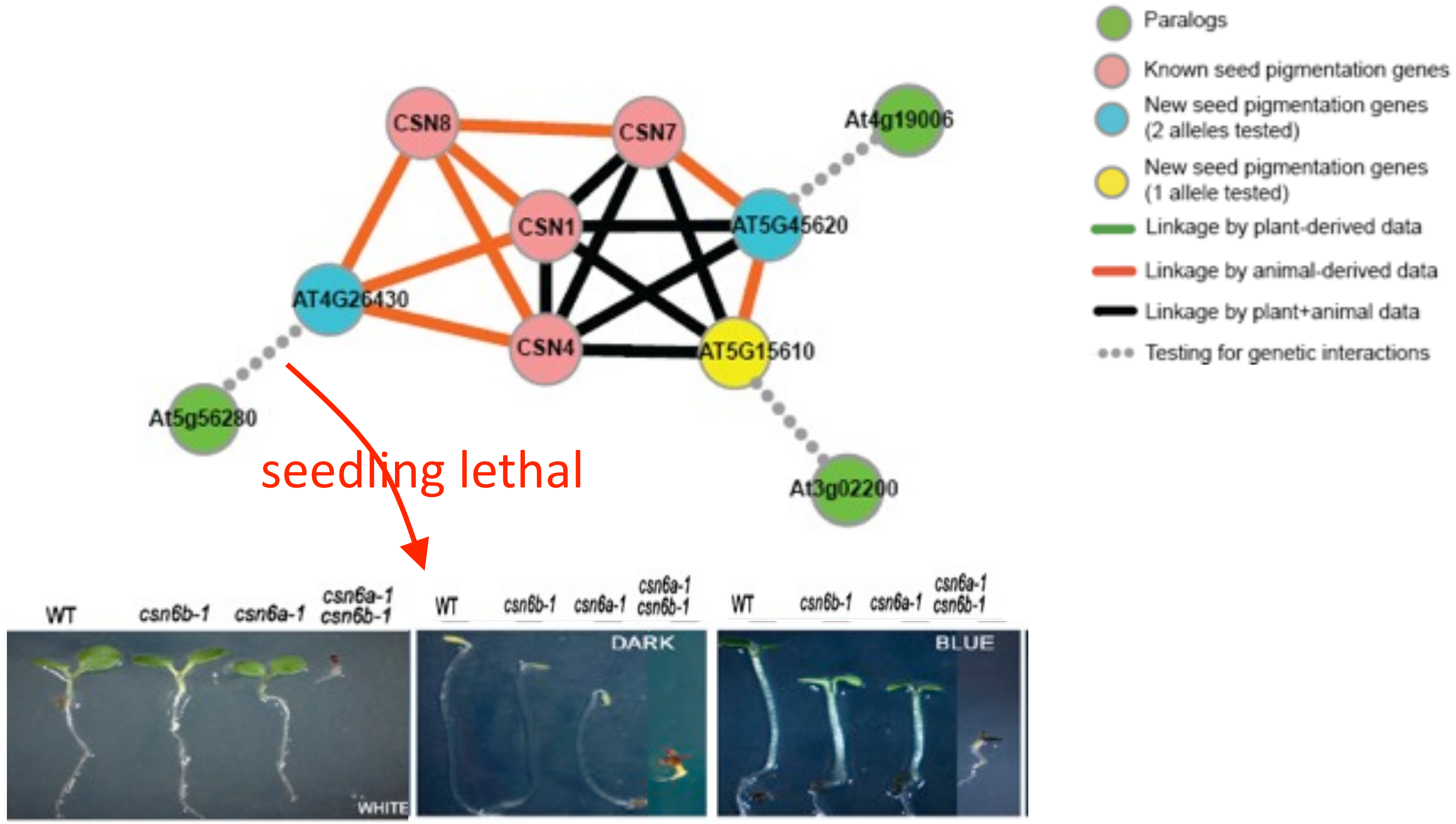
Eight seed pigmentation mutants categorize into five network components.


CSN8
CSN7
CSN1
AT5G45620
AT4G26430
CSN4
AT5G15610
COP1
DET1
AT5G04110
AT5G50110
ALB3
AT5G11480
CHLG
CHLI1
CHLD
CHLH
AT3G20160
DXS
Known seed pigmentation genes
New seed pigmentation genes (2 alleles tested)
New seed pigmentation genes (1 allele tested)
Linkage by plant-derived data
Linkage by animal-derived data
Linkage by plant+animal data
photomorphogenesis
chloroplast development

Example of Overlapping Genetic Function between Duplicated Genes


CSN8
CSN7
At4g19006
CSN1
AT5G45620
AT4G26430
CSN4
AT5G15610
At5g56280
At3g02200
Paralogs
Known seed pigmentation genes
New seed pigmentation genes (2 alleles tested)
New seed pigmentation genes (1 allele tested)
Linkage by plant-derived data
Linkage by animal-derived data
Linkage by plant+animal data
Testing for genetic interactions

Example of Overlapping Genetic Function between Duplicated Genes

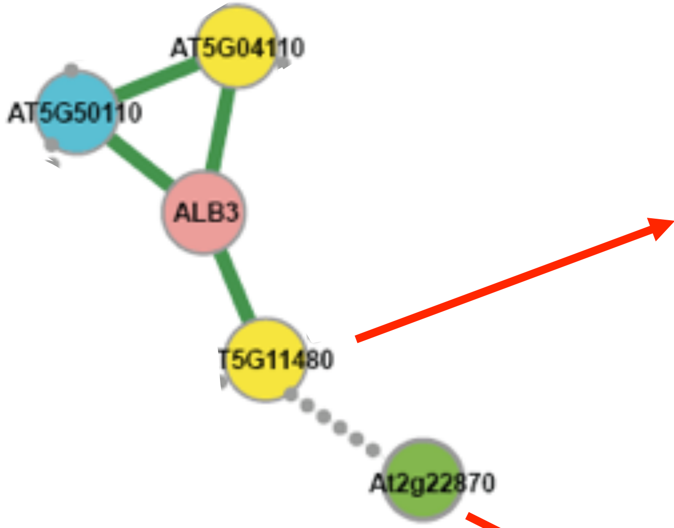

CSN8
CSN7
At4g19006
CSN1
AT5G45620
AT4G26430
CSN4
AT5G15610
At5g56280
At3g02200
Paralogs
Known seed pigmentation genes
New seed pigmentation genes (2 alleles tested)
New seed pigmentation genes (1 allele tested)
Linkage by plant-derived data
Linkage by animal-derived data
Linkage by plant+animal data
Testing for genetic interactions
WT
csn6b-1
csn6a-1
csn6a-1 csn6b-1
WHITE
WT
csn6b-1
csn6a-1
csn6a-1 csn6b-1
DARK
WT
csn6b-1
csn6a-1
csn6a-1 csn6b-1
BLUE

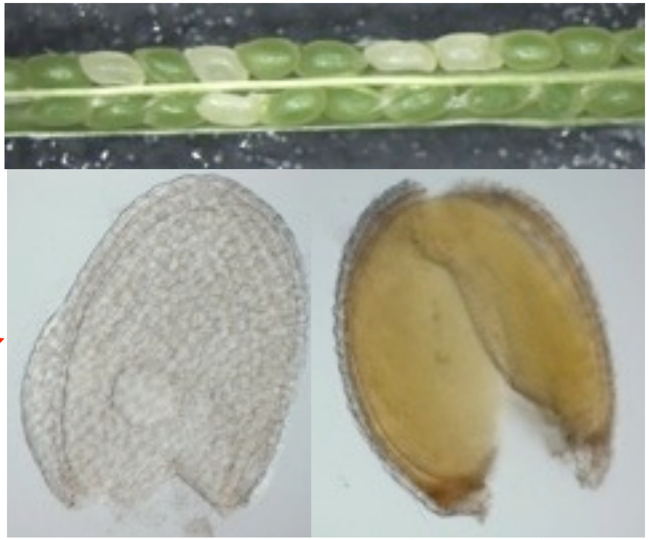
seedling lethal

Gusmaroli et al (2007) The Plant Cell, Vol. 19: 564–581

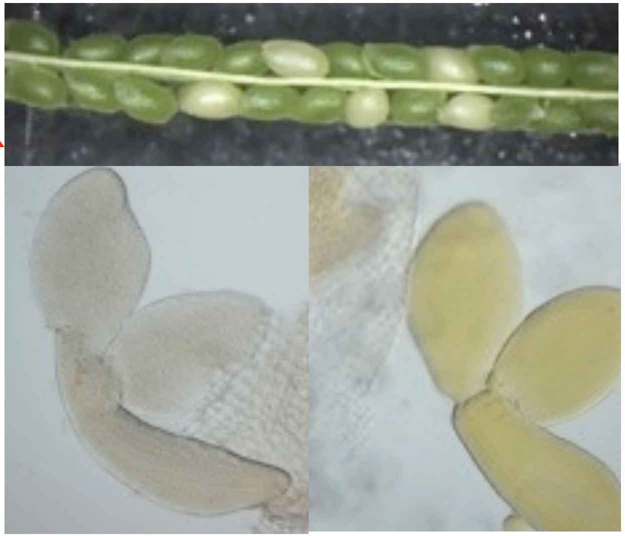
At5G11480 and its paralog are required for

Chloroplast development Subnetwork





-/- -/+ or +/+



-/- -/+ or +/+

● Paralogs

● Known seed pigmentation genes

● New seed pigmentation genes (2 alleles tested)

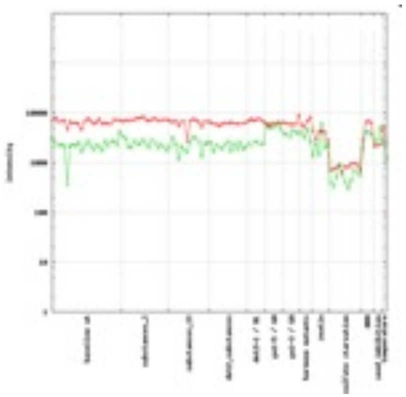
● New seed pigmentation genes (1 allele tested)

▬ Linkage by plant-derived data

▬ Linkage by animal-derived data

▬ Linkage by plant+animal data

••• Testing for genetic interactions



AtGenExpress Vis Tool –Weigel lab

Hye-In Nam

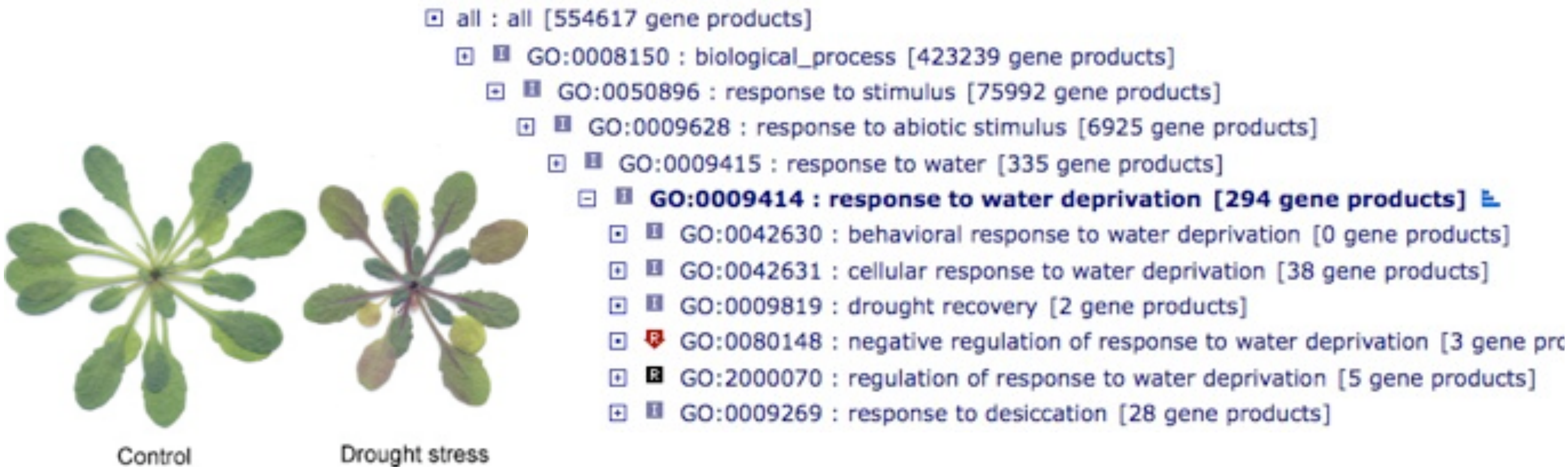
Questions and Goals

1. How do we describe and define processes and functions to allow comparison, modeling, and prediction?

2. How do we find all genes that are involved in a biological process?

3. How do we model the functions, processes, and phenotypes to explain complex traits and predict phenotypes from genotypes?

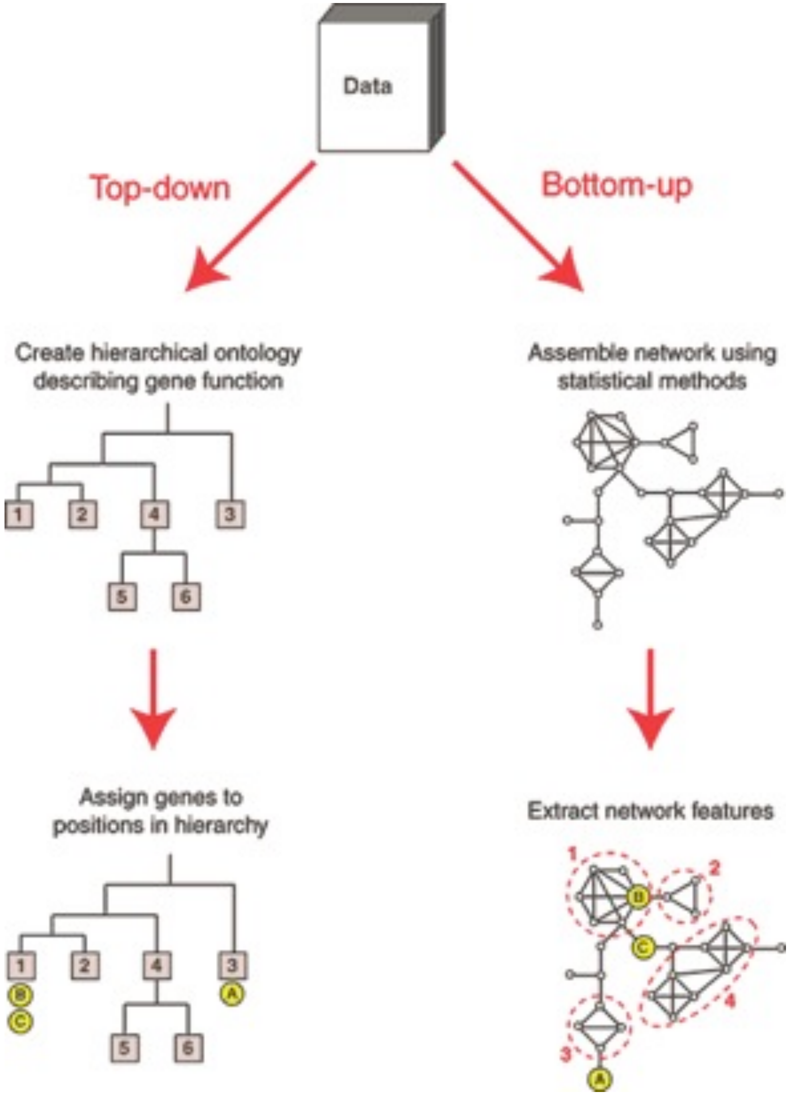
Gene Ontology –Biological Process



```
⊡ all : all [554617 gene products]
  ⊞ I GO:0008150 : biological_process [423239 gene products]
    ⊞ I GO:0050896 : response to stimulus [75992 gene products]
      ⊞ I GO:0009628 : response to abiotic stimulus [6925 gene products]
        ⊞ I GO:0009415 : response to water [335 gene products]
          ⊟ I GO:0009414 : response to water deprivation [294 gene products]
            ⊞ I GO:0042630 : behavioral response to water deprivation [0 gene products]
            ⊞ I GO:0042631 : cellular response to water deprivation [38 gene products]
            ⊞ I GO:0009819 : drought recovery [2 gene products]
            ⊞ R GO:0080148 : negative regulation of response to water deprivation [3 gene prc
            ⊞ R GO:2000070 : regulation of response to water deprivation [5 gene products]
            ⊞ I GO:0009269 : response to desiccation [28 gene products]
```

- Response to water deprivation (GO:0009414)
  - Photosynthesis (GO:0015979)
  - Anthocyanin biosynthesis (GO:0009718)
  - Stomatal closure (GO:0090332)
  - Leaf development (GO:0048366)
  - Root development (GO:0048364)

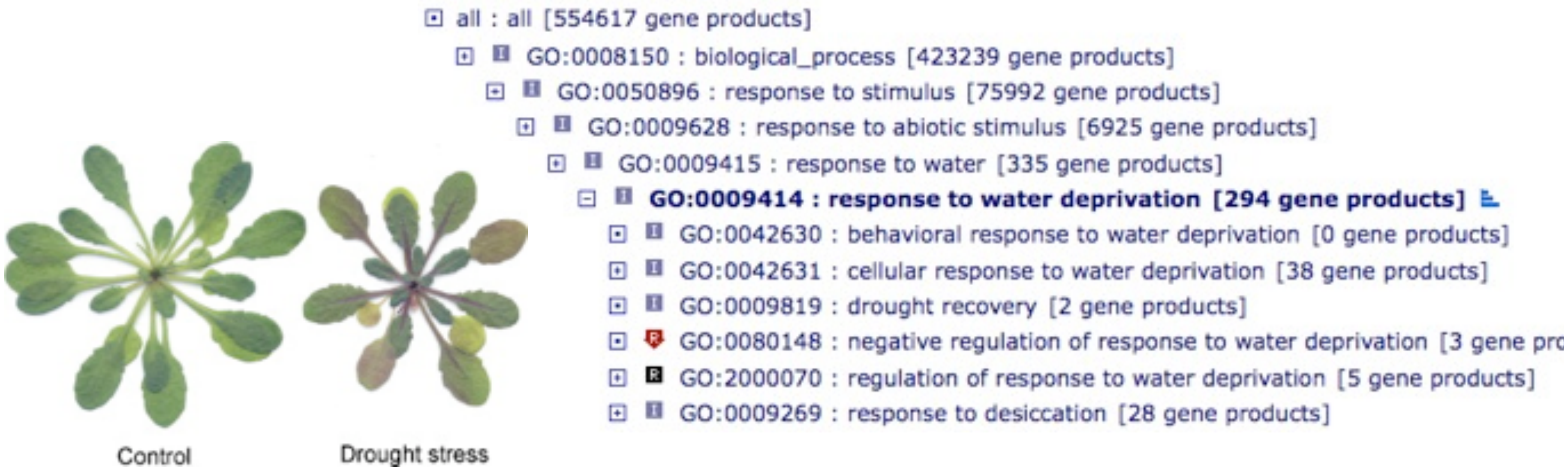
Two Approaches in Creating Biological Process Pathways/Networks



Data
Top-down
Bottom-up
Create hierarchical ontology describing gene function
Assemble network using statistical methods
1
2
4
3
5
6
Assign genes to positions in hierarchy
Extract network features
B
C
A
1
2
3
4

Fraser & Marcotte (2006)
Nat Genetics

Gene Ontology –Biological Process



```
⊡ all : all [554617 gene products]
  ⊡ GO:0008150 : biological_process [423239 gene products]
    ⊡ GO:0050896 : response to stimulus [75992 gene products]
      ⊡ GO:0009628 : response to abiotic stimulus [6925 gene products]
        ⊡ GO:0009415 : response to water [335 gene products]
          ⊟ GO:0009414 : response to water deprivation [294 gene products]
            ⊡ GO:0042630 : behavioral response to water deprivation [0 gene products]
            ⊡ GO:0042631 : cellular response to water deprivation [38 gene products]
            ⊡ GO:0009819 : drought recovery [2 gene products]
            ⊡ GO:0080148 : negative regulation of response to water deprivation [3 gene prc
            ⊡ R GO:2000070 : regulation of response to water deprivation [5 gene products]
            ⊡ GO:0009269 : response to desiccation [28 gene products]
```

- Response to water deprivation (GO:0009414)
  - Photosynthesis (GO:0015979)
  - Anthocyanin biosynthesis (GO:0009718)
  - Stomatal closure (GO:0090332)
  - Leaf development (GO:0048366)
  - Root development (GO:0048364)

Questions and Goals

1. How do we **describe** and define processes and functions to allow comparison, modeling, and prediction?

Ontologies and ontology-based annotations

1. How do we **find** all genes that are involved in a biological process?

Genome-wide co-function networks and experimental validations

1. How do we **model** the functions, processes, and phenotypes to explain complex traits and predict phenotypes from genotypes?

Networks of biological processes and establishment of causalities

Acknowledgements

Current:

Lee Chae (postdoc)
Taehyong Kim (postdoc)
Flavia Bossi (postdoc)
Meng Xu (postdoc)
Peifen Zhang (curator)
Kate Dreher (curator)
Hye-In Nam (RA)
Damian Priamurskiy (intern)
Tam Tran (intern)

Recently Former:

Jin Chen (postdoc)
Chang-hun You (postdoc)
Kun He (postdoc)
Bindu Ambaru (RA)
Pranjali Thakkar (intern)
Nathaniel Leu (intern)
Julian Huang (intern)
Purva Thakkar (intern)

Collaborators:

Insuk Lee, Yonsei University
Edward Marcotte, University of Texas at Austin
Wolf Frommer, Carnegie
Basil Nikolau, Iowa State University
Jin Chen, Michigan State University
Jiajie Peng, Michigan State University

tair Eva Huala

Gene Ontology Consortium

National Science Foundation
Carnegie Institution for Science