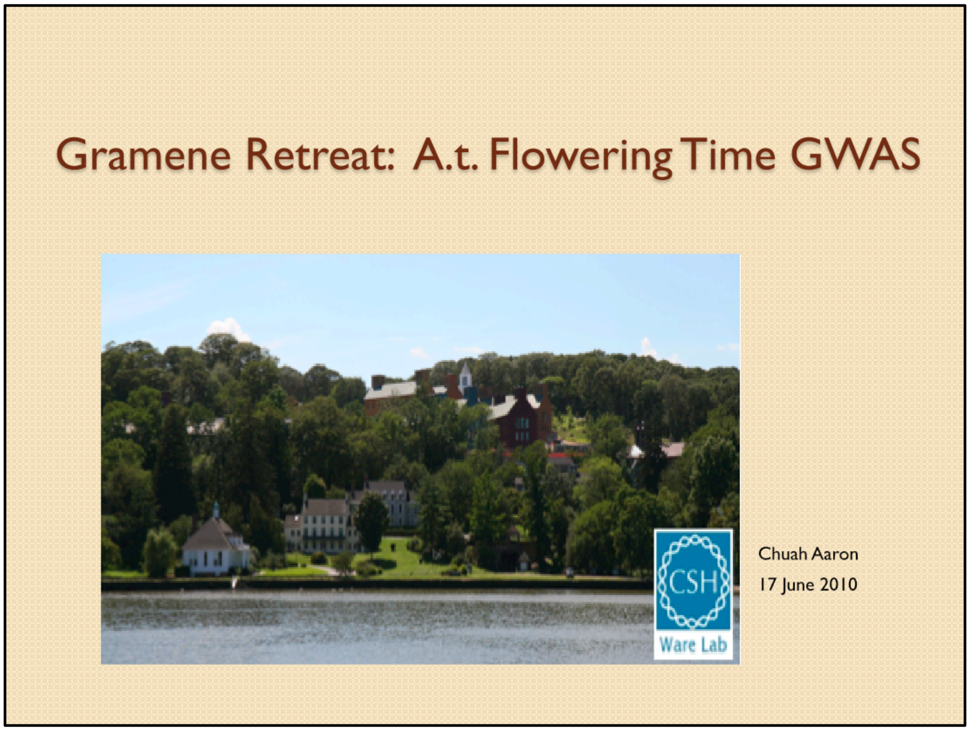
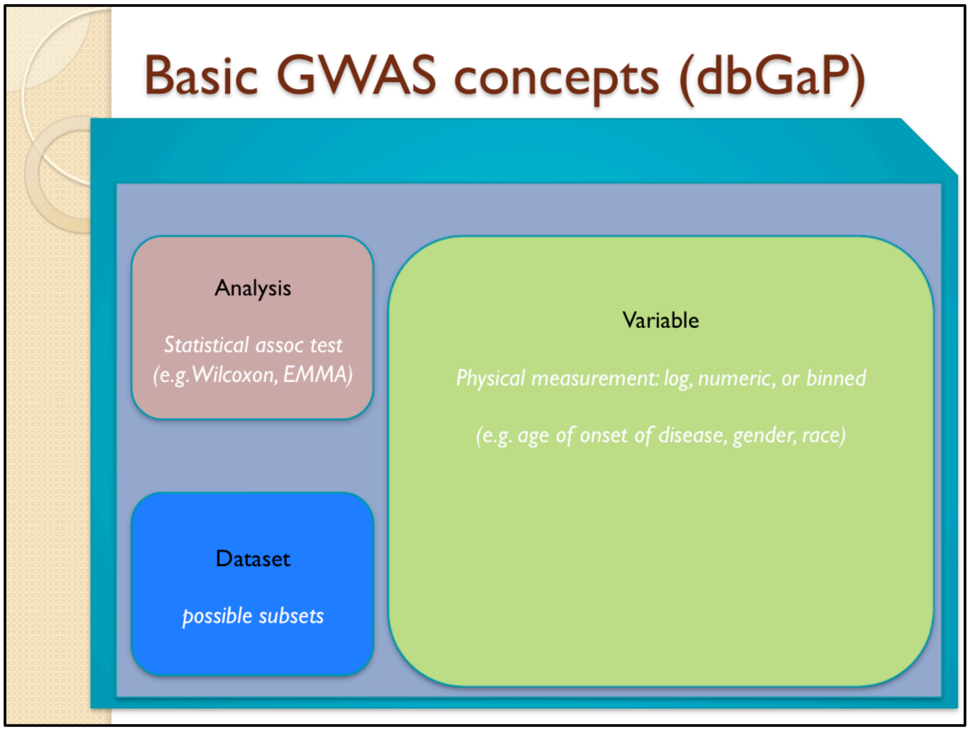


Gramene Retreat: A.t. Flowering Time GWAS

Chuah Aaron

17 June 2010



Basic GWAS concepts (dbGaP)

Analysis

Statistical assoc test (e.g. Wilcoxon, EMMA)

Dataset

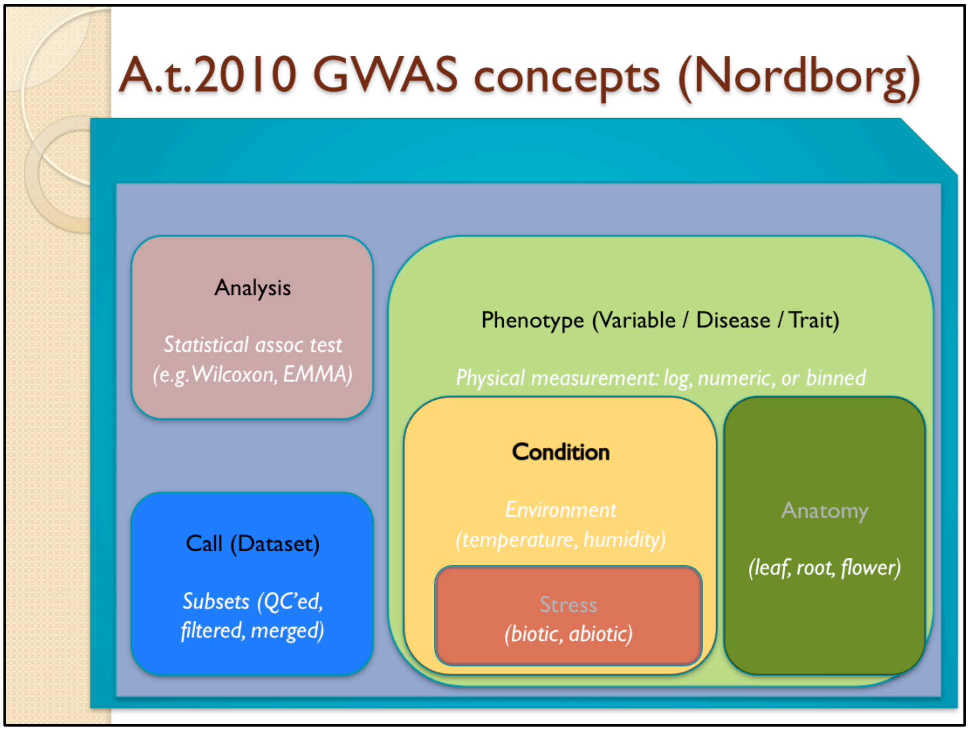
possible subsets

Variable

Physical measurement: log, numeric, or binned

(e.g. age of onset of disease, gender, race)

Gap#1: dbGaP is NCBI’s pheno/genotypic equivalent of dbSNP – it stores Variations, Documents and Analyses (Associations), which may not be enough



A.t.2010 GWAS concepts (Nordborg)

Analysis

Statistical assoc test (e.g. Wilcoxon, EMMA)

Call (Dataset)

Subsets (QC'ed, filtered, merged)

Phenotype (Variable / Disease / Trait)

Physical measurement: log, numeric, or binned

Condition

Environment (temperature, humidity)

Stress

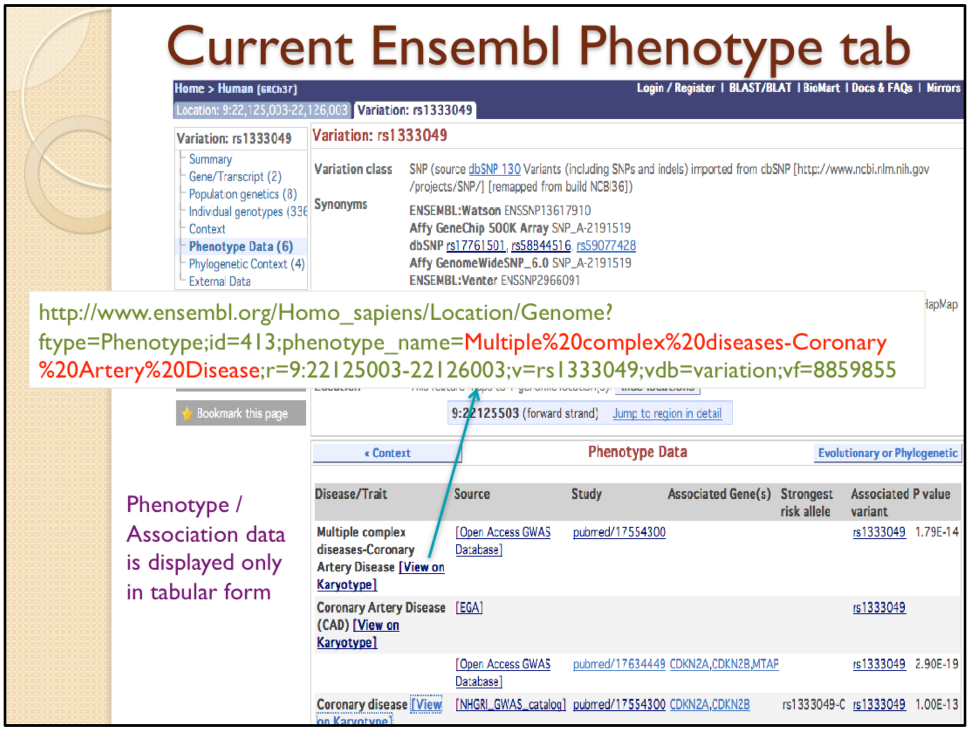
(biotic, abiotic)

Anatomy

(leaf, root, flower)

Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana inbred lines*

Susanna Atwell, Yu S. Huang, Bjarni J. Vilhjálmsson, Glenda Willems, Matthew Horton, Yan Li, Dazhe Meng, Alexander Platt, Aaron M. Tarone, Tina T. Hu, Rong Jiang, N. Wayan Muliyati, Xu Zhang, Muhammad Ali Amer, Ivan Baxter, Benjamin Brachi, Joanne Chory, Caroline Dean, Marilyne Debieu, Juliette de Meaux, Joseph R. Ecker, Nathalie Faure, Joel M. Kniskern, Jonathan D. G. Jones, Todd Michael et al.AffiliationsContributionsCorresponding authorNature 465, 627–631 (03 June 2010)



Current Ensembl Phenotype tab

Variation: rs1333049

Variation class SNP (source dbSNP 130 Variants (including SNPs and indels) imported from dbSNP [http://www.ncbi.nlm.nih.gov /projects/SNP/] [remapped from build NCBI36])

Synonyms
ENSEMBL:Watson ENSSNP13617910
Affy GeneChip 500K Array SNP_A-2191519
dbSNP rs17761501, rs58844516, rs59077428
Affy GenomeWideSNP_6.0 SNP_A-2191519
ENSEMBL:Venter ENSSNP2966091

http://www.ensembl.org/Homo_sapiens/Location/Genome?ftype=Phenotype;id=413;phenotype_name=Multiple%20complex%20diseases-Coronary%20Artery%20Disease;r=9:22125003-22126003;v=rs1333049;vdb=variation;vf=8859855

9:22125503 (forward strand) Jump to region in detail

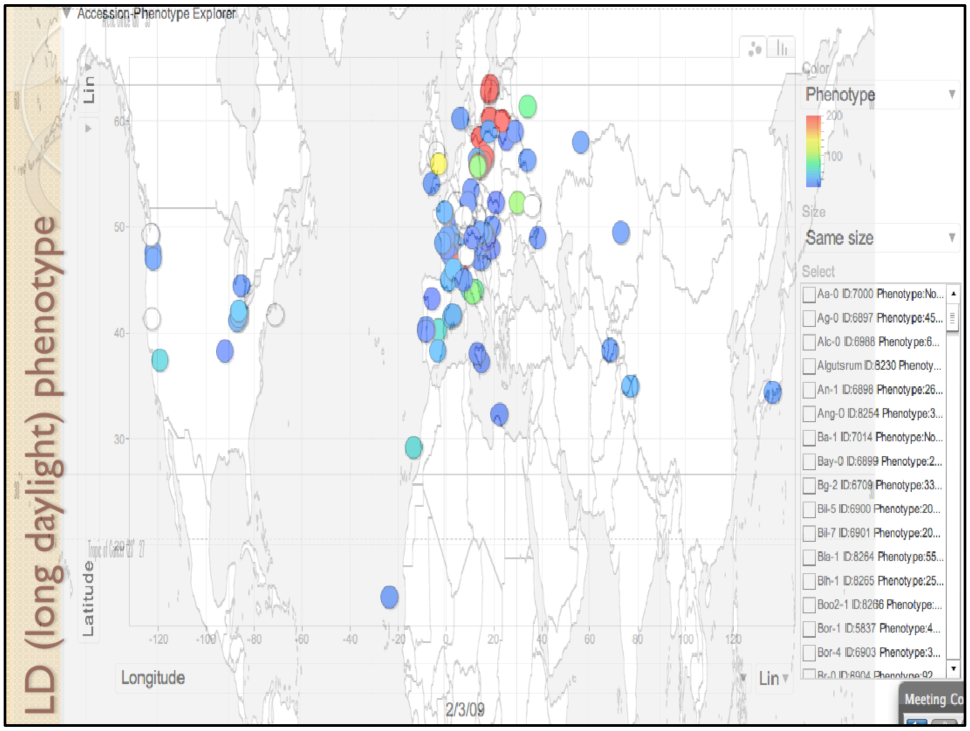
« Context | **Phenotype Data** | Evolutionary or Phylogenetic

Disease/Trait	Source	Study	Associated Gene(s)	Strongest risk allele	Associated variant	P value
Multiple complex diseases-Coronary Artery Disease [View on Karyotype]	[Open Access GWAS Database]	pubmed/17554300			rs1333049	1.79E-14
Coronary Artery Disease (CAD) [View on Karyotype]	[EGA]				rs1333049	
	[Open Access GWAS Database]	pubmed/17634449	CDKN2A,CDKN2B,MTAP		rs1333049	2.90E-19
Coronary disease [View on Karyotype]	[NHGRI_GWAS_catalog]	pubmed/17554300	CDKN2A,CDKN2B	rs1333049-C	rs1333049	1.00E-13

Phenotype / Association data is displayed only in tabular form

Gap #2: You've seen the verbosity of the Disease/Trait (Phenotype) naming yesterday http://uswest.ensembl.org/Homo_sapiens/Variation/Phenotype?source=dbSNP;v=rs420259

LD (long daylight) phenotype



Color: Phenotype

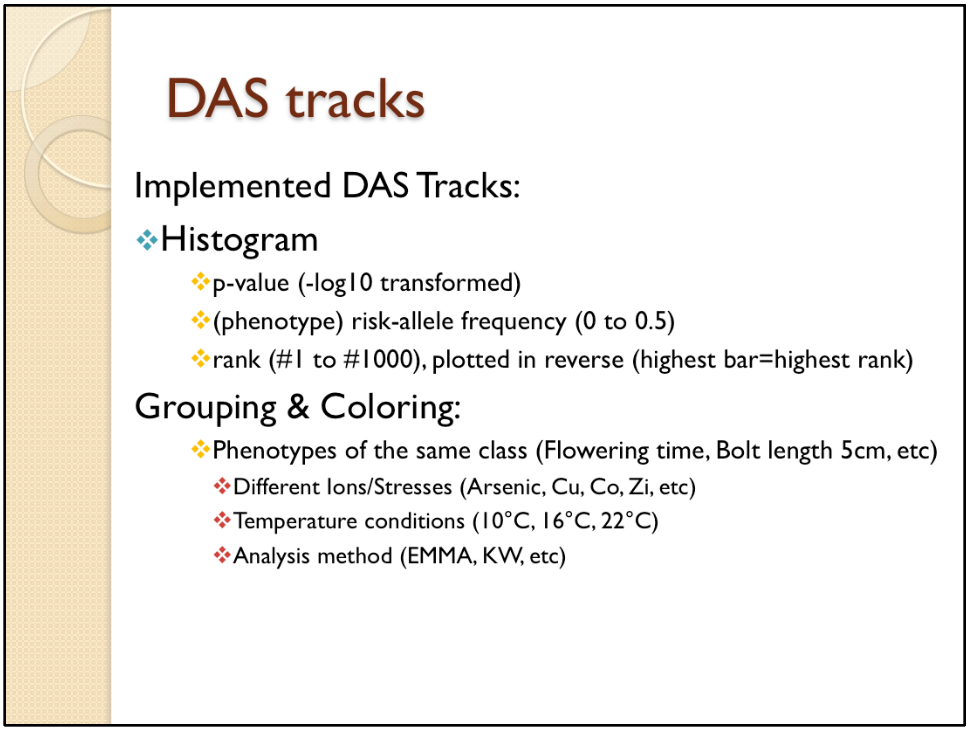
Size: Same size

Select

- □ Aa-0 ID:7000 Phenotype:No...
- □ Ag-0 ID:6897 Phenotype:45...
- □ Alc-0 ID:6988 Phenotype:6...
- □ Algutsrum ID:8230 Phenoty...
- □ An-1 ID:6898 Phenotype:26...
- □ Ang-0 ID:8254 Phenotype:3...
- □ Ba-1 ID:7014 Phenotype:No...
- □ Bay-0 ID:6899 Phenotype:2...
- □ Bg-2 ID:6709 Phenotype:33...
- □ Bil-5 ID:6900 Phenotype:20...
- □ Bil-7 ID:6901 Phenotype:20...
- □ Bla-1 ID:8264 Phenotype:55...
- □ Blh-1 ID:8265 Phenotype:25...
- □ Boo2-1 ID:8266 Phenotype:...
- □ Bor-1 ID:5837 Phenotype:4...
- □ Bor-4 ID:6903 Phenotype:3...
- □ Br-0 ID:6904 Phenotype:92

2/3/09

Gap #3: Nordborg's data contains much information, but without an appropriate visualization, it's hard to interpret



DAS tracks

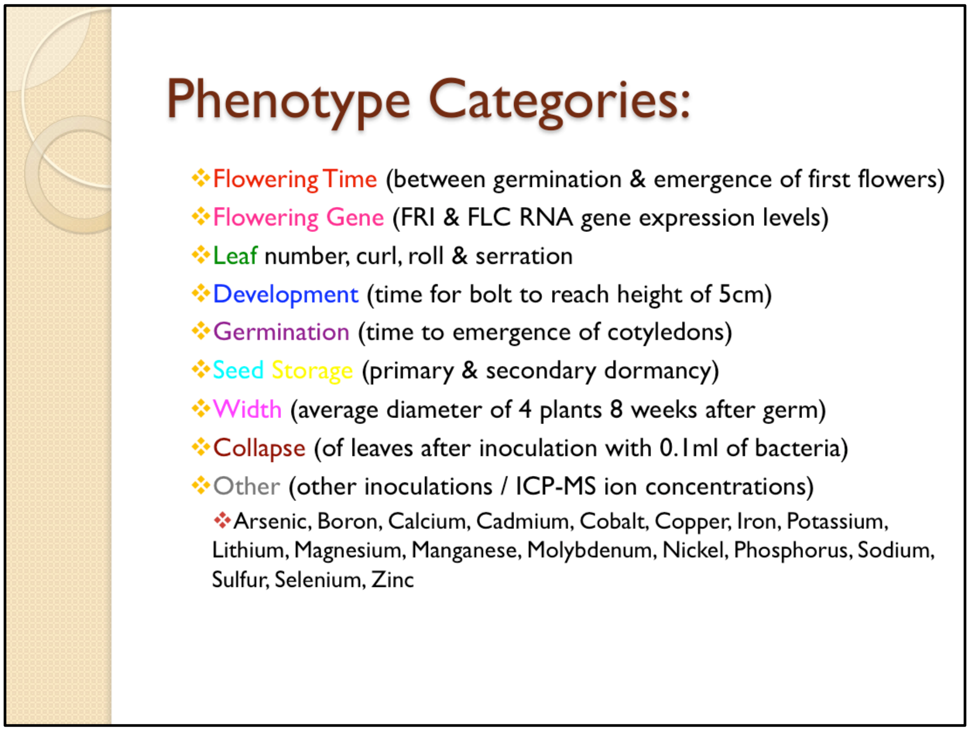
Implemented DAS Tracks:

- Histogram
  - p-value (-log10 transformed)
  - (phenotype) risk-allele frequency (0 to 0.5)
  - rank (#1 to #1000), plotted in reverse (highest bar=highest rank)

Grouping & Coloring:

- Phenotypes of the same class (Flowering time, Bolt length 5cm, etc)
  - Different Ions/Stresses (Arsenic, Cu, Co, Zi, etc)
  - Temperature conditions (10°C, 16°C, 22°C)
  - Analysis method (EMMA, KW, etc)

http://uswest.ensembl.org/Homo_sapiens/Variation/Phenotype?source=dbSNP;v=rs420259

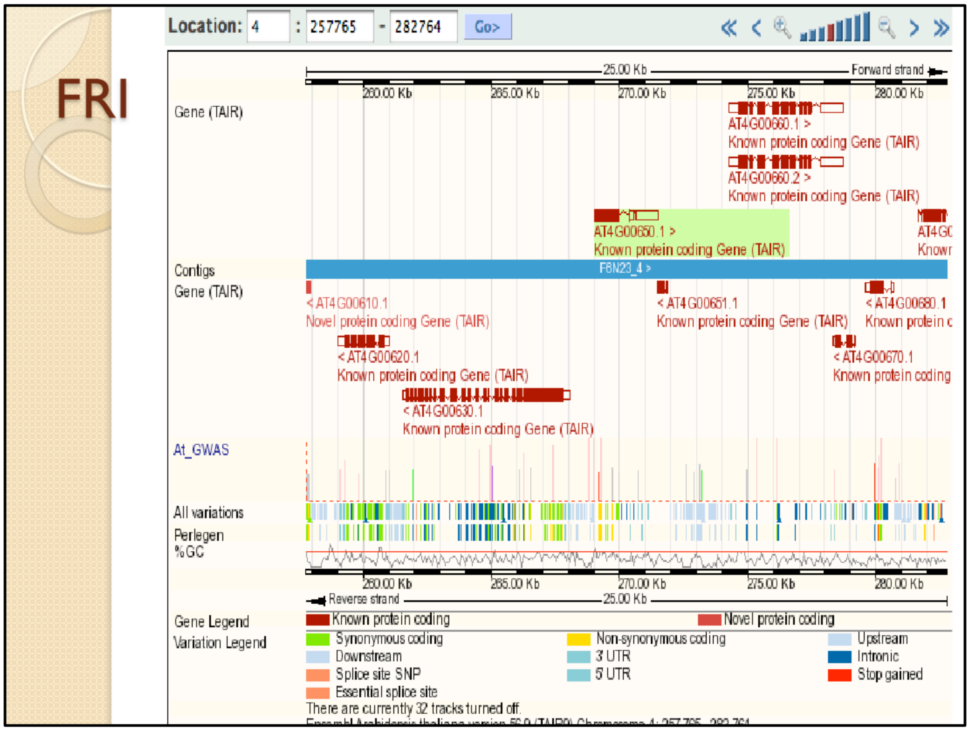


Phenotype Categories:

- Flowering Time (between germination & emergence of first flowers)
- Flowering Gene (FRI & FLC RNA gene expression levels)
- Leaf number, curl, roll & serration
- Development (time for bolt to reach height of 5cm)
- Germination (time to emergence of cotyledons)
- Seed Storage (primary & secondary dormancy)
- Width (average diameter of 4 plants 8 weeks after germ)
- Collapse (of leaves after inoculation with 0.1ml of bacteria)
- Other (other inoculations / ICP-MS ion concentrations)
  - Arsenic, Boron, Calcium, Cadmium, Cobalt, Copper, Iron, Potassium, Lithium, Magnesium, Manganese, Molybdenum, Nickel, Phosphorus, Sodium, Sulfur, Selenium, Zinc

http://uswest.ensembl.org/Homo_sapiens/Variation/Phenotype?source=dbSNP;v=rs420259

FRI

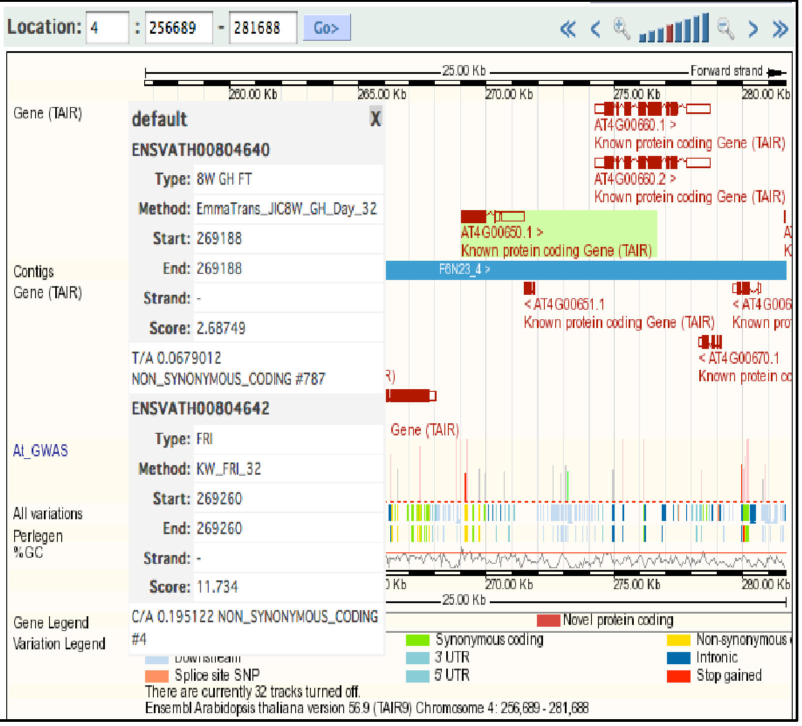


Location: 4 : 257765 - 282764 Go>

There are currently 32 tracks turned off.

Gene: Frigida

FRI



Location: 4 : 256689 - 281688 Go>

default

ENSVATH00804640

Type:	8W GH FT
Method:	EmmaTrans_JIC8W_GH_Day_32
Start:	269188
End:	269188
Strand:	-
Score:	2.68749

T/A 0.0679012 NON_SYNONYMOUS_CODING #787

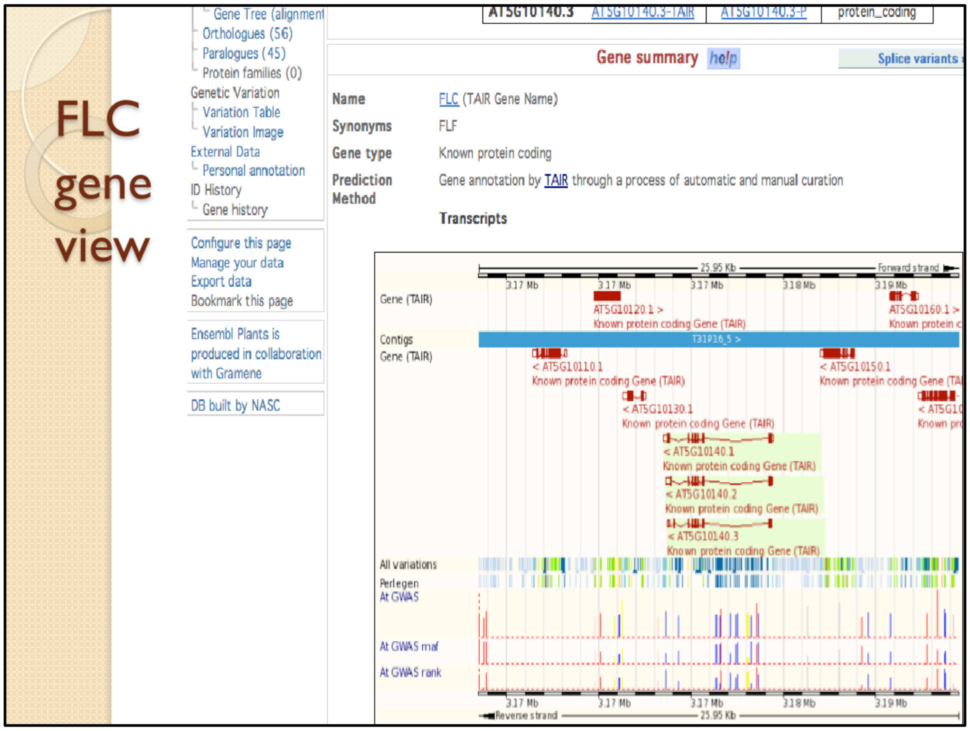
ENSVATH00804642

Type:	FRI
Method:	KW_FRI_32
Start:	269260
End:	269260
Strand:	-
Score:	11.734

C/A 0.195122 NON_SYNONYMOUS_CODING #4

There are currently 32 tracks turned off.

Ensembl Arabidopsis thaliana version 56.9 (TAIR9) Chromosome 4: 256,689 - 281,688



FLC gene view

- Gene Tree (alignment
- Orthologues (56)
- Paralogues (45)
- Protein families (0)

Genetic Variation

- Variation Table
- Variation Image

External Data

- Personal annotation

ID History

- Gene history

Configure this page
Manage your data
Export data
Bookmark this page

Ensembl Plants is produced in collaboration with Gramene

DB built by NASC

AT5G10140.3	AT5G10140.3-TAIR	AT5G10140.3-P	protein_coding

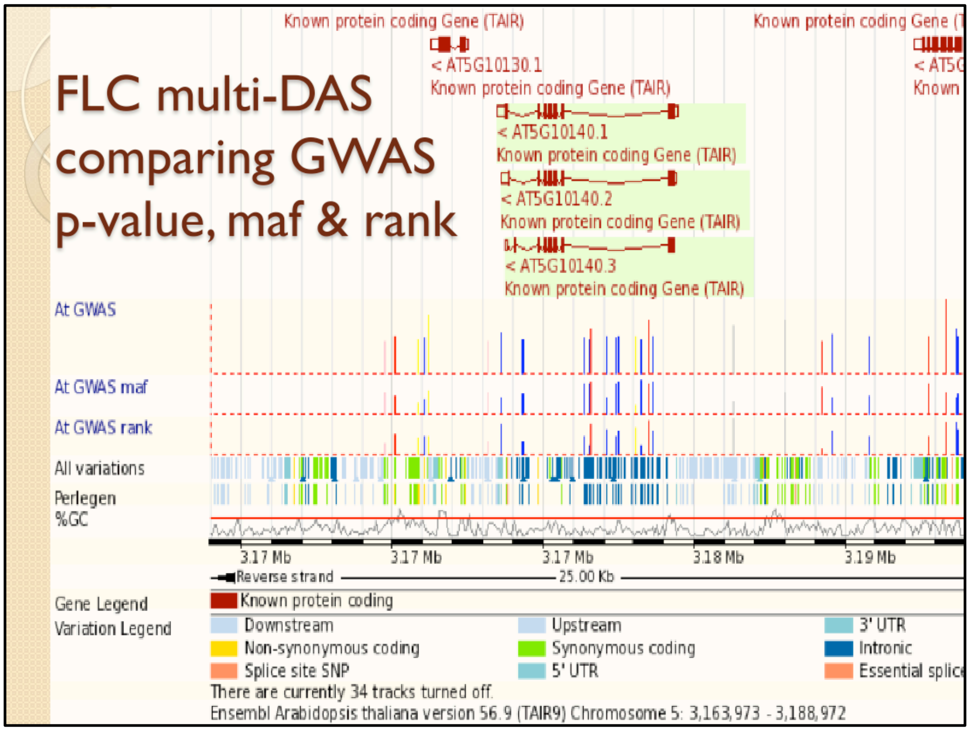
Gene summary help

Splice variants

Name	FLC (TAIR Gene Name)
Synonyms	FLF
Gene type	Known protein coding
Prediction Method	Gene annotation by TAIR through a process of automatic and manual curation

Transcripts

Flowering Locus C



FLC multi-DAS comparing GWAS p-value, maf & rank

There are currently 34 tracks turned off.

Ensembl Arabidopsis thaliana version 56.9 (TAIR9) Chromosome 5: 3,163,973 - 3,188,972

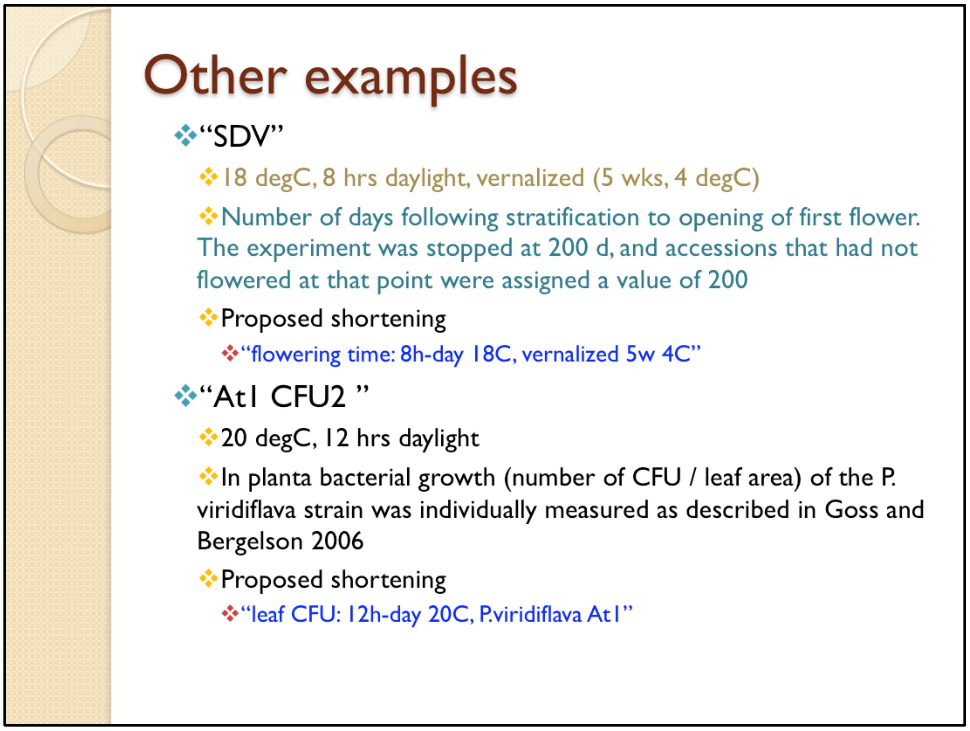
Note low rank of red line near 3.19M despite p-value close to the rest (and higher maf than most)

Primarily Red (FT), Pink (FT-gxp) & Blue (Development)

Phenotype description

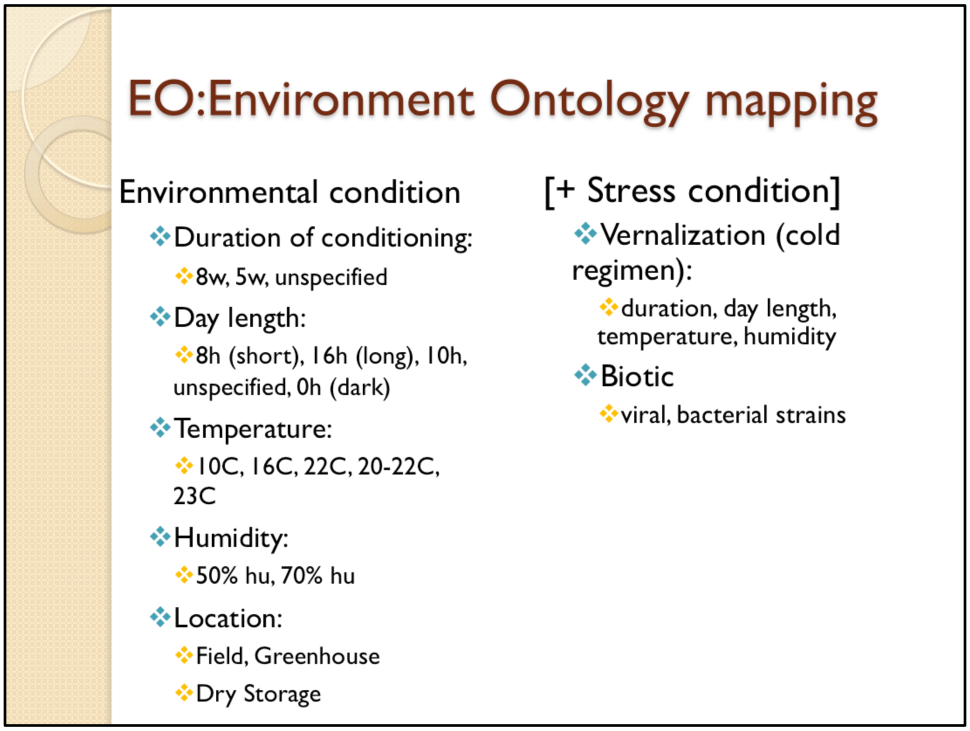
255 chars isn't enough to capture Nordborg's "growth_condition" & "phenotype_scoring" fields

- e.g. phenotype "avrPphB"
  - 20degC, 12 hrs daylight
  - Following inoculation of two leaves per plant with 0.1 ml of 10 -8 cfu/ml bacteria in 10 mM MgSO4 buffer using a blunt-tipped syringe, leaf collapse was scored at 20 hrs and again at 24 hrs after inoculation. A positive score at either time point was deemed a hypersensitive response
  - Proposed shortening
    - "hypersensitivity: 12h-day 20C, P.syringae AvrPphB"



Other examples

- “SDV”
  - 18 degC, 8 hrs daylight, vernalized (5 wks, 4 degC)
  - Number of days following stratification to opening of first flower. The experiment was stopped at 200 d, and accessions that had not flowered at that point were assigned a value of 200
  - Proposed shortening
    - “flowering time: 8h-day 18C, vernalized 5w 4C”
- “At1 CFU2 ”
  - 20 degC, 12 hrs daylight
  - In planta bacterial growth (number of CFU / leaf area) of the P. viridiflava strain was individually measured as described in Goss and Bergelson 2006
  - Proposed shortening
    - “leaf CFU: 12h-day 20C, P.viridiflava At1”



EO:Environment Ontology mapping

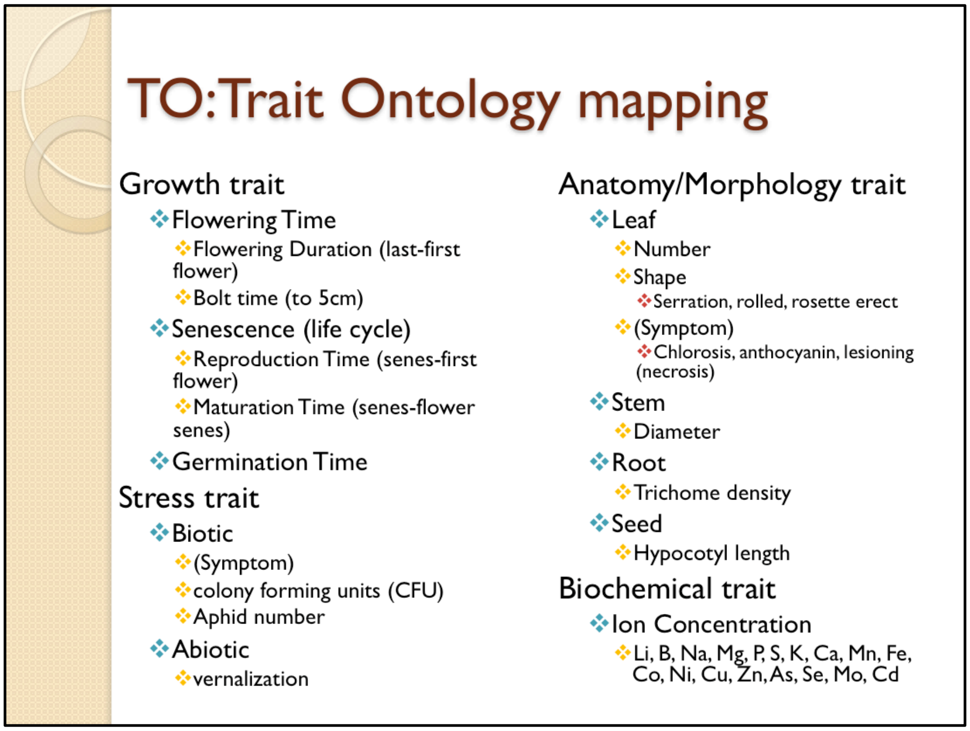
Environmental condition

- Duration of conditioning:
  - 8w, 5w, unspecified
- Day length:
  - 8h (short), 16h (long), 10h, unspecified, 0h (dark)
- Temperature:
  - 10C, 16C, 22C, 20-22C, 23C
- Humidity:
  - 50% hu, 70% hu
- Location:
  - Field, Greenhouse
  - Dry Storage

[+ Stress condition]

- Vernalization (cold regimen):
  - duration, day length, temperature, humidity
- Biotic
  - viral, bacterial strains

http://uswest.ensembl.org/Homo_sapiens/Variation/Phenotype?source=dbSNP;v=rs420259



TO:Trait Ontology mapping

Growth trait

- Flowering Time
  - Flowering Duration (last-first flower)
  - Bolt time (to 5cm)
- Senescence (life cycle)
  - Reproduction Time (senes-first flower)
  - Maturation Time (senes-flower senes)
- Germination Time

Stress trait

- Biotic
  - (Symptom)
  - colony forming units (CFU)
  - Aphid number
- Abiotic
  - vernalization

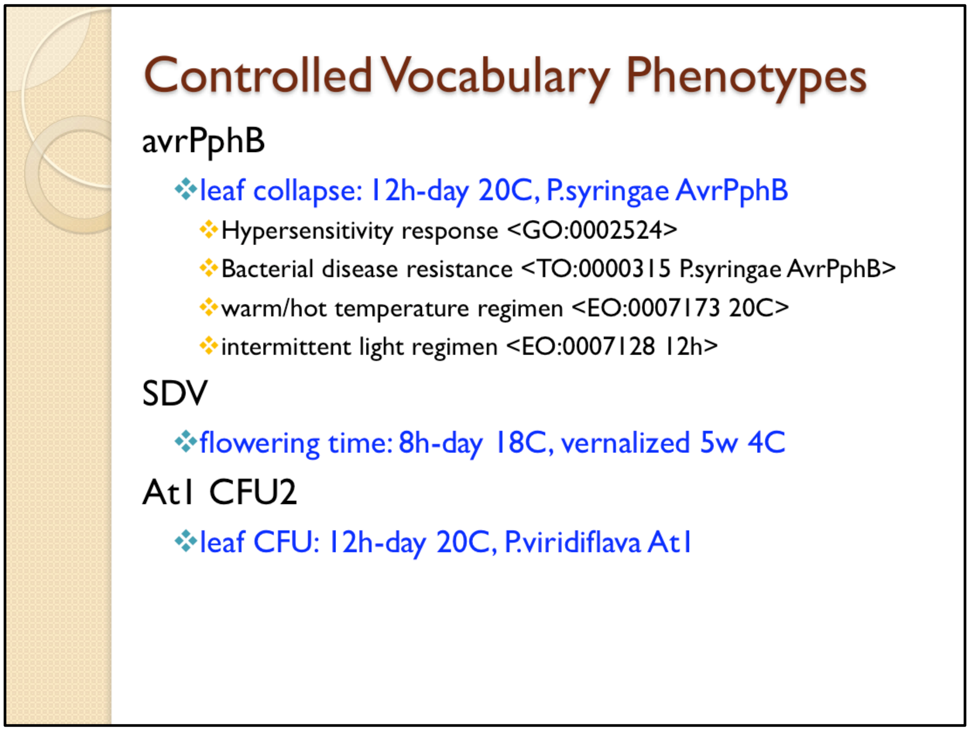
Anatomy/Morphology trait

- Leaf
  - Number
  - Shape
    - Serration, rolled, rosette erect
  - (Symptom)
    - Chlorosis, anthocyanin, lesioning (necrosis)
- Stem
  - Diameter
- Root
  - Trichome density
- Seed
  - Hypocotyl length

Biochemical trait

- Ion Concentration
  - Li, B, Na, Mg, P, S, K, Ca, Mn, Fe, Co, Ni, Cu, Zn, As, Se, Mo, Cd

http://uswest.ensembl.org/Homo_sapiens/Variation/Phenotype?source=dbSNP;v=rs420259



Controlled Vocabulary Phenotypes

avrPphB

- leaf collapse: 12h-day 20C, P.syringae AvrPphB
  - Hypersensitivity response <GO:0002524>
  - Bacterial disease resistance <TO:0000315 P.syringae AvrPphB>
  - warm/hot temperature regimen <EO:0007173 20C>
  - intermittent light regimen <EO:0007128 12h>

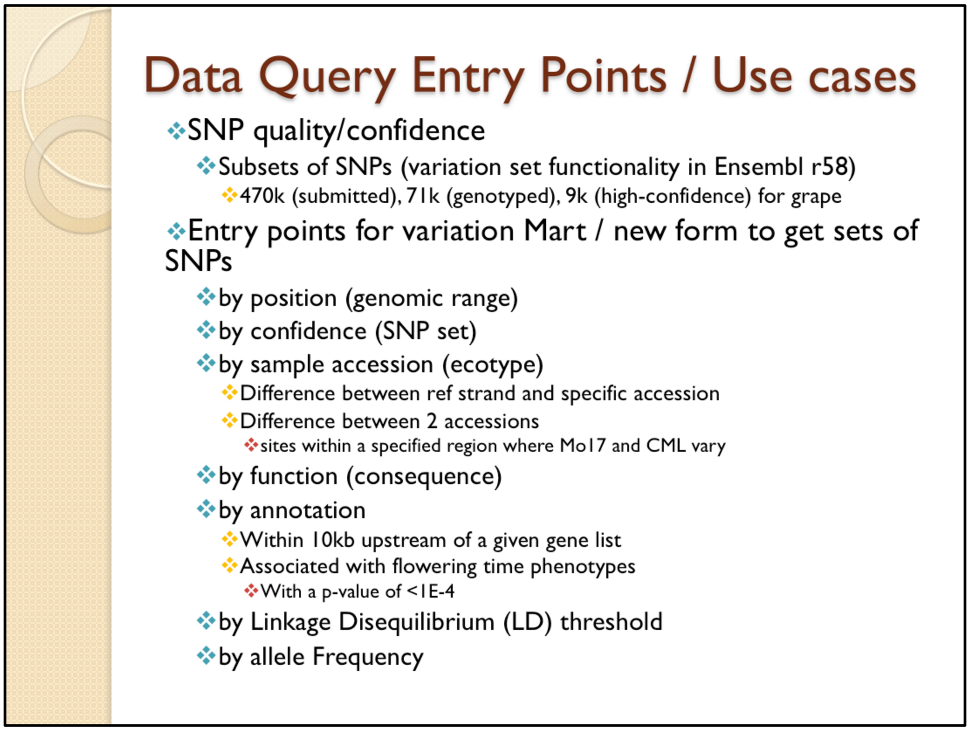
SDV

- flowering time: 8h-day 18C, vernalized 5w 4C

At1 CFU2

- leaf CFU: 12h-day 20C, P.viridiflava At1

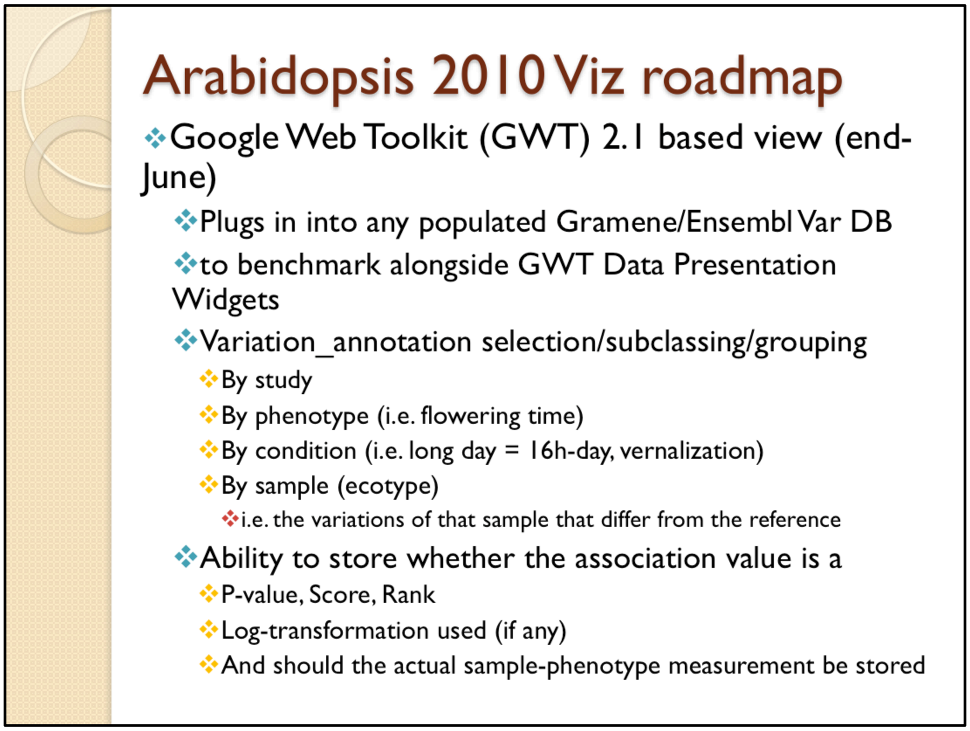
To send to Josh to work out the first pass…
AvrPphB is an avirulence (Avr) protein from the plant pathogen Pseudomonas syringae that can trigger a disease-resistance response



Data Query Entry Points / Use cases

- SNP quality/confidence
  - Subsets of SNPs (variation set functionality in Ensembl r58)
    - 470k (submitted), 71k (genotyped), 9k (high-confidence) for grape
- Entry points for variation Mart / new form to get sets of SNPs
  - by position (genomic range)
  - by confidence (SNP set)
  - by sample accession (ecotype)
    - Difference between ref strand and specific accession
    - Difference between 2 accessions
      - sites within a specified region where Mo17 and CML vary
  - by function (consequence)
  - by annotation
    - Within 10kb upstream of a given gene list
    - Associated with flowering time phenotypes
      - With a p-value of <1E-4
  - by Linkage Disequilibrium (LD) threshold
  - by allele Frequency

Gap #4: more entry points needed



Arabidopsis 2010 Viz roadmap

- Google Web Toolkit (GWT) 2.1 based view (end-June)
  - Plugs in into any populated Gramene/Ensembl Var DB
  - to benchmark alongside GWT Data Presentation Widgets
  - Variation_annotation selection/subclassing/grouping
    - By study
    - By phenotype (i.e. flowering time)
    - By condition (i.e. long day = 16h-day, vernalization)
    - By sample (ecotype)
      - i.e. the variations of that sample that differ from the reference
  - Ability to store whether the association value is a
    - P-value, Score, Rank
    - Log-transformation used (if any)
    - And should the actual sample-phenotype measurement be stored

Pulling from GDPDM?