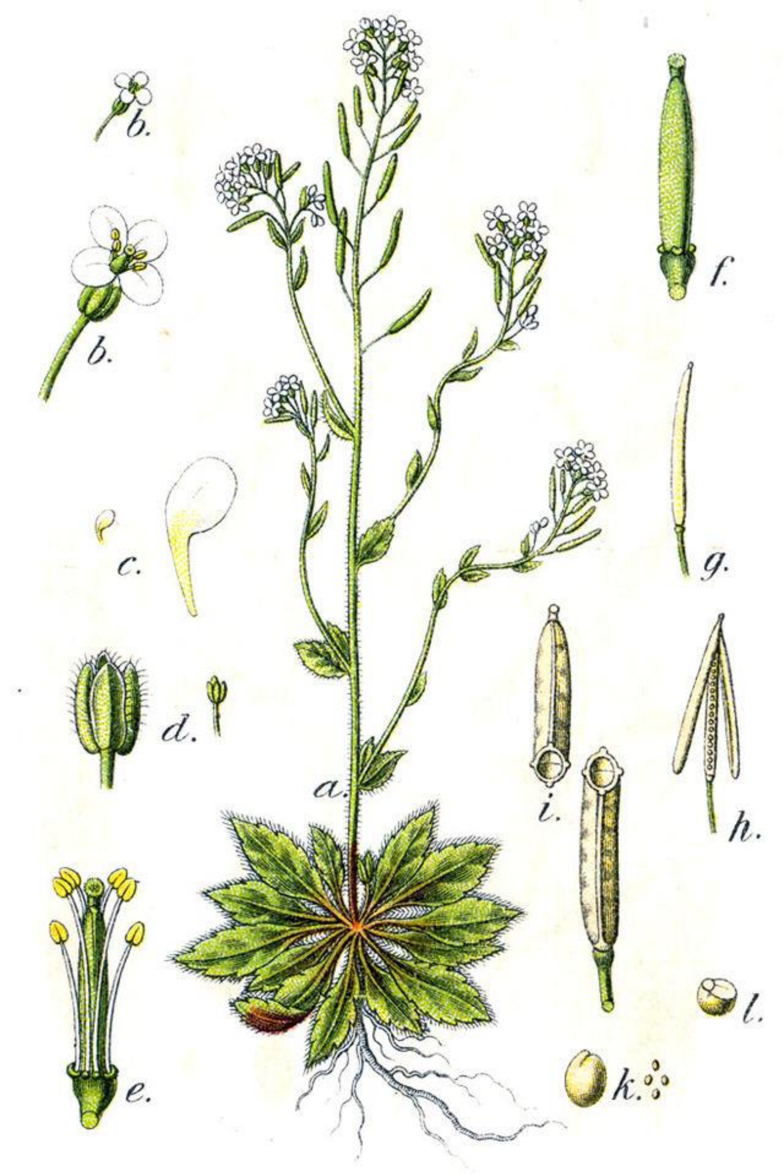


GWAS in a model organism: *Arabidopsis thaliana*

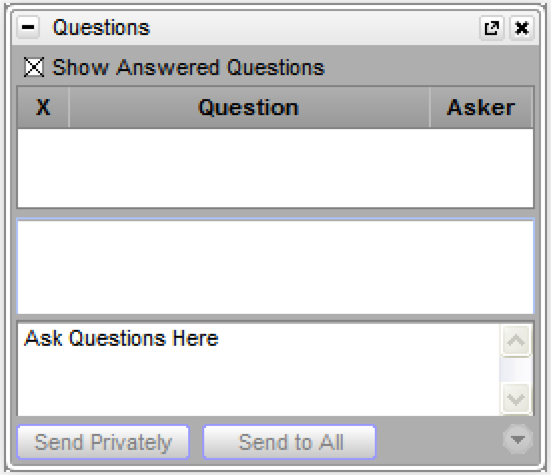
June 9, 2014

Ashley Hintz
Field Application Scientist

Golden Helix
Accelerating the Quest for Significance™

Questions during the presentation

Use the Questions pane in your GoToWebinar window



☒ Show Answered Questions

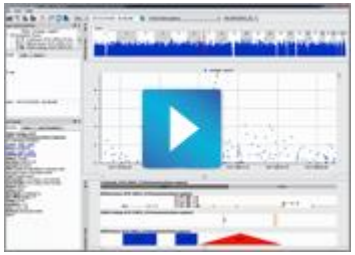
X	Question	Asker

Ask Questions Here

Golden Helix
Accelerating the Quest for Significance™

Introduction

- Combining topics of previous webcasts:
  - “Maximizing Public Data Sources for Sequencing and GWAS Studies”
  - “Back to Basics: Using GWAS to Drive Discovery for Complex Diseases”
  - “Mixed Models: How to Effectively Account for Inbreeding and Population Structure in GWAS”
- Additional Dimension: Non-human data from *A. thaliana*
- Combining these topics in a single webcast!

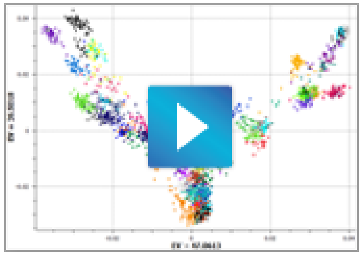


+

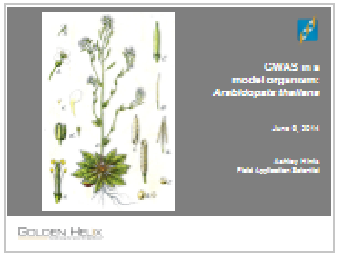
Sources of Public Data

- NCBI
  - dbGaP
  - GEO
  - SRA
- EGA
- HapMap Project
- 1000 Genomes Project
- Hardware vendors
- Software vendors
- All over the internet…

+



=



GWAS in a model organism: Arabidopsis thaliana

June 9, 2014

Ashley Hintz
Field Application Scientist

Agenda

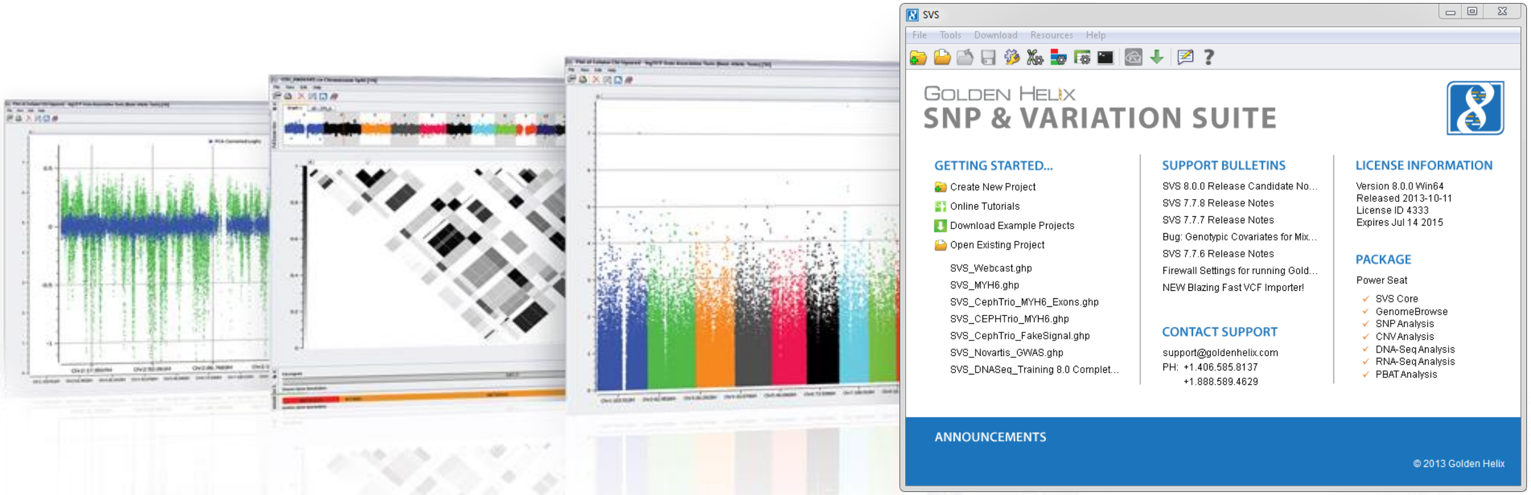
1 *A.* Thaliana Data Overview

2 Obtaining Public Data

3 SVS Demonstration

4 Conclusion and Questions

SNP & Variation Suite (SVS)



Golden Helix
SNP & VARIATION SUITE

GETTING STARTED...

- Create New Project
- Online Tutorials
- Download Example Projects
- Open Existing Project

SVS_Webcast.ghp
SVS_MYH6.ghp
SVS_CephTrio_MYH6_Exons.ghp
SVS_CEPHTrio_MYH6.ghp
SVS_CephTrio_FakeSignal.ghp
SVS_Novartis_GWAS.ghp
SVS_DNASeq_Training 8.0 Complet...

SUPPORT BULLETINS

SVS 8.0.0 Release Candidate No...
SVS 7.7.8 Release Notes
SVS 7.7.7 Release Notes
Bug: Genotypic Covariates for Mix...
SVS 7.7.6 Release Notes
Firewall Settings for running Gold...
NEW Blazing Fast VCF Importer!

CONTACT SUPPORT

support@goldenhelix.com
PH: +1.406.585.8137
+1.888.589.4629

LICENSE INFORMATION

Version 8.0.0 Win64
Released 2013-10-11
License ID 4333
Expires Jul 14 2015

PACKAGE

Power Seat

- ✓ SVS Core
- ✓ GenomeBrowse
- ✓ SNP Analysis
- ✓ CNV Analysis
- ✓ DNA-Seq Analysis
- ✓ RNA-Seq Analysis
- ✓ PBAT Analysis

ANNOUNCEMENTS

© 2013 Golden Helix

Core Features

- Powerful Data Management
- Rich Visualizations
- Robust Statistics
- Flexible
- Easy-to-use

Applications

- Genotype Analysis
- DNA sequence analysis
- CNV Analysis
- RNA-seq differential expression
- Family Based Association

Agenda

1 *A. thaliana* Data Overview

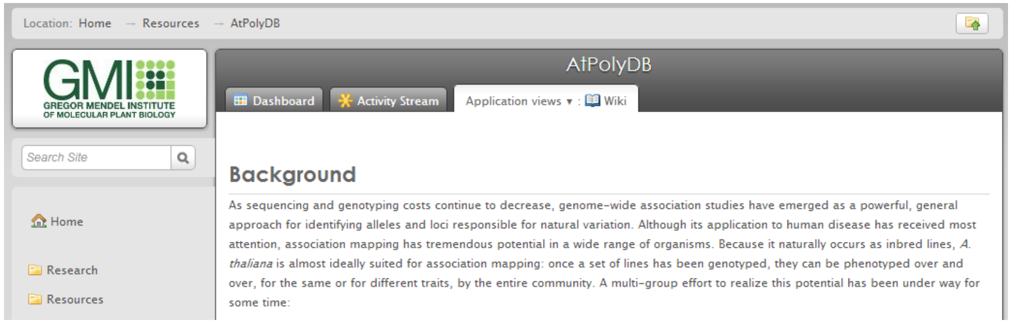
2 Obtaining Public Data

3 SVS Demonstration

4 Conclusion and Questions

Dataset Overview

- Downloaded from Gregor Mendel Institute for Molecular Plant Biology through AtPolyDB
- Custom Affymetrix 250K SNP chip
  - Genotyped 1307 samples covering 214,051 markers
- 107 different phenotypes were recorded from Atwell *et al.* 2010
  - Flowering, ionomics, defense and development
- Trait associated with virulence to *Pseudomonas*



Location: Home → Resources → AtPolyDB
GMI
GREGOR MENDEL INSTITUTE OF MOLECULAR PLANT BIOLOGY
Search Site
Home
Research
Resources
AtPolyDB
Dashboard
Activity Stream
Application views ▾ : Wiki

Background

As sequencing and genotyping costs continue to decrease, genome-wide association studies have emerged as a powerful, general approach for identifying alleles and loci responsible for natural variation. Although its application to human disease has received most attention, association mapping has tremendous potential in a wide range of organisms. Because it naturally occurs as inbred lines, *A. thaliana* is almost ideally suited for association mapping: once a set of lines has been genotyped, they can be phenotyped over and over, for the same or for different traits, by the entire community. A multi-group effort to realize this potential has been under way for some time:

www.pseudomonas-syringae.org

Agenda

1 *A. thaliana* Data Overview

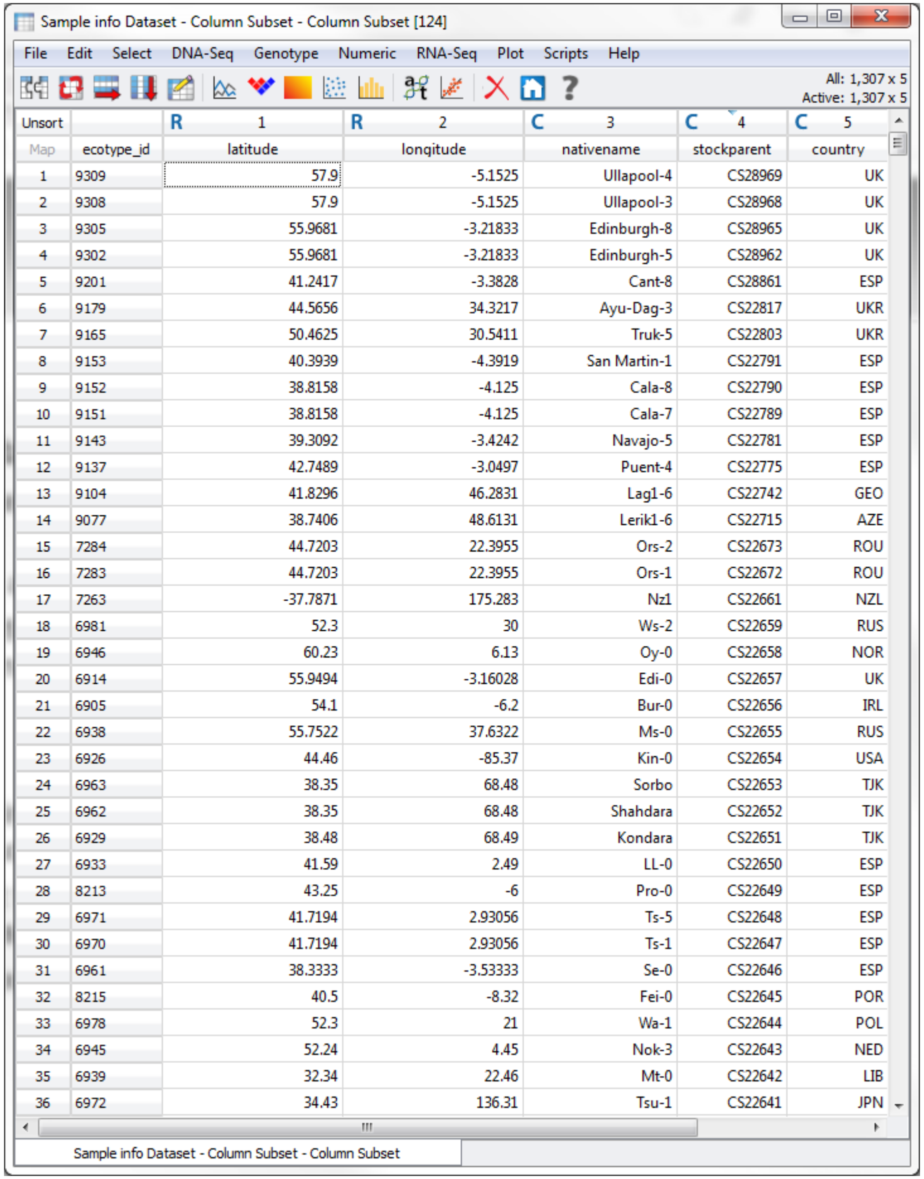
2 Obtaining Public Data

3 SVS Demonstration

4 Conclusion and Questions

Raw Data

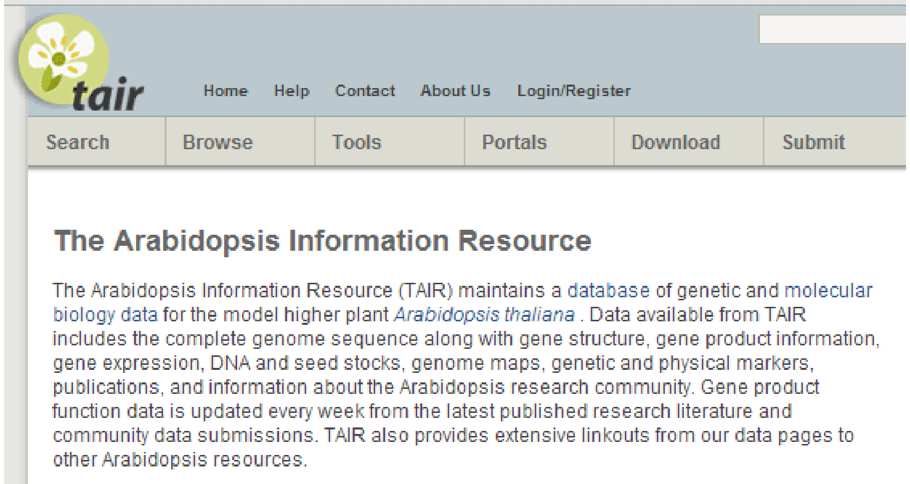
- Genotype Data
  - Downloaded as .CSV file containing haploid genotypes
  - Imported text file to SVS, used Python script to convert to homozygous diploid genotypes for analysis
- Sample Data
  - Text file detailing collection site of all 1307 samples, including GPS coordinates
  - Additional text file with phenotypes for 199 samples, as used in Atwell *et al.* 2010 GWAS paper
  - Both were imported to SVS for further analysis


Sample info Dataset - Column Subset - Column Subset [124]
File Edit Select DNA-Seq Genotype Numeric RNA-Seq Plot Scripts Help
All: 1,307 x 5
Active: 1,307 x 5

Unsort		R 1	R 2	C 3	C 4	C 5
Map	ecotype_id	latitude	longitude	nativename	stockparent	country
1	9309	57.9	-5.1525	Ullapool-4	CS28969	UK
2	9308	57.9	-5.1525	Ullapool-3	CS28968	UK
3	9305	55.9681	-3.21833	Edinburgh-8	CS28965	UK
4	9302	55.9681	-3.21833	Edinburgh-5	CS28962	UK
5	9201	41.2417	-3.3828	Cant-8	CS28861	ESP
6	9179	44.5656	34.3217	Ayu-Dag-3	CS22817	UKR
7	9165	50.4625	30.5411	Truk-5	CS22803	UKR
8	9153	40.3939	-4.3919	San Martin-1	CS22791	ESP
9	9152	38.8158	-4.125	Cala-8	CS22790	ESP
10	9151	38.8158	-4.125	Cala-7	CS22789	ESP
11	9143	39.3092	-3.4242	Navajo-5	CS22781	ESP
12	9137	42.7489	-3.0497	Puent-4	CS22775	ESP
13	9104	41.8296	46.2831	Lag1-6	CS22742	GEO
14	9077	38.7406	48.6131	Lerik1-6	CS22715	AZE
15	7284	44.7203	22.3955	Ors-2	CS22673	ROU
16	7283	44.7203	22.3955	Ors-1	CS22672	ROU
17	7263	-37.7871	175.283	Nz1	CS22661	NZL
18	6981	52.3	30	Ws-2	CS22659	RUS
19	6946	60.23	6.13	Oy-0	CS22658	NOR
20	6914	55.9494	-3.16028	Edi-0	CS22657	UK
21	6905	54.1	-6.2	Bur-0	CS22656	IRL
22	6938	55.7522	37.6322	Ms-0	CS22655	RUS
23	6926	44.46	-85.37	Kin-0	CS22654	USA
24	6963	38.35	68.48	Sorbo	CS22653	TJK
25	6962	38.35	68.48	Shahdara	CS22652	TJK
26	6929	38.48	68.49	Kondara	CS22651	TJK
27	6933	41.59	2.49	LL-0	CS22650	ESP
28	8213	43.25	-6	Pro-0	CS22649	ESP
29	6971	41.7194	2.93056	Ts-5	CS22648	ESP
30	6970	41.7194	2.93056	Ts-1	CS22647	ESP
31	6961	38.3333	-3.53333	Se-0	CS22646	ESP
32	8215	40.5	-8.32	Fei-0	CS22645	POR
33	6978	52.3	21	Wa-1	CS22644	POL
34	6945	52.24	4.45	Nok-3	CS22643	NED
35	6939	32.34	22.46	Mt-0	CS22642	LIB
36	6972	34.43	136.31	Tsu-1	CS22641	JPN

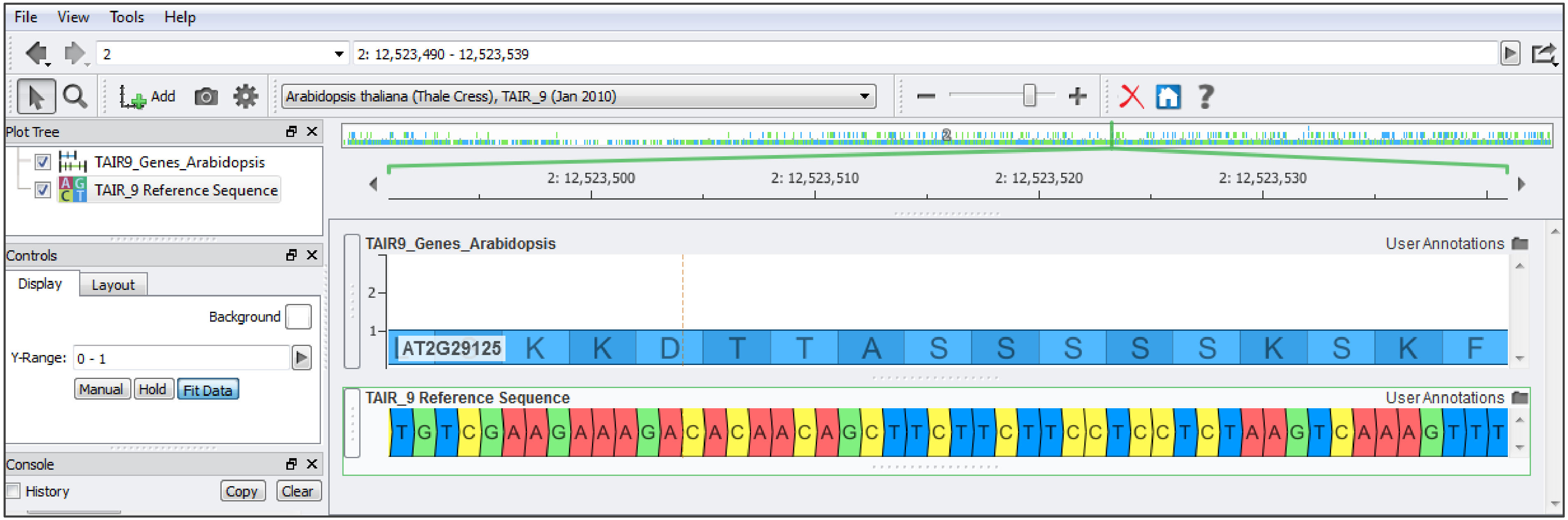
TAIR_9 Assembly Annotations

- Downloaded from Arabidopsis.org
- Reference sequence in FASTA format
- Gene annotations in GFF3 format
- Both converted to SVS native TSF format using tools in SVS



The Arabidopsis Information Resource

The Arabidopsis Information Resource (TAIR) maintains a database of genetic and molecular biology data for the model higher plant *Arabidopsis thaliana* . Data available from TAIR includes the complete genome sequence along with gene structure, gene product information, gene expression, DNA and seed stocks, genome maps, genetic and physical markers, publications, and information about the Arabidopsis research community. Gene product function data is updated every week from the latest published research literature and community data submissions. TAIR also provides extensive linkouts from our data pages to other Arabidopsis resources.



2: 12,523,490 - 12,523,539

Arabidopsis thaliana (Thale Cress), TAIR_9 (Jan 2010)

Y-Range: 0 - 1

Agenda

1 *A. thaliana* Data Overview

2 Obtaining Public Data

3 SVS Demonstration

4 Conclusion and Questions

Demonstration Agenda

1 Quality Assurance Filters

2 Principle Components Analysis

3 Genotype Association Testing

4 EMMAX

5 Visualizations

GOLDEN HELIX
SNP & VARIATION SUITE 8

[Demonstration]

Agenda

1 *A. thaliana* Data Overview

2 Obtaining Public Data

3 SVS Demonstration

4 Conclusion and Questions

Conclusion

(Atwell et al. 2010)

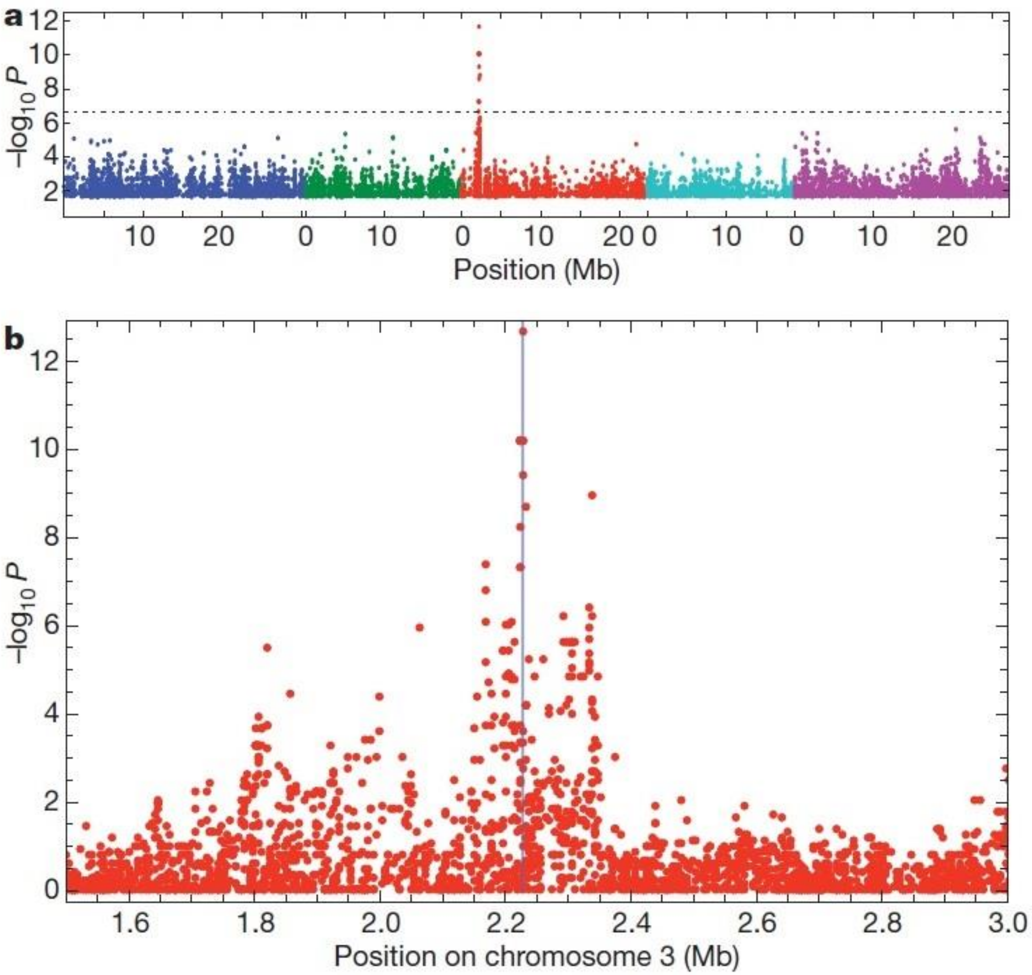
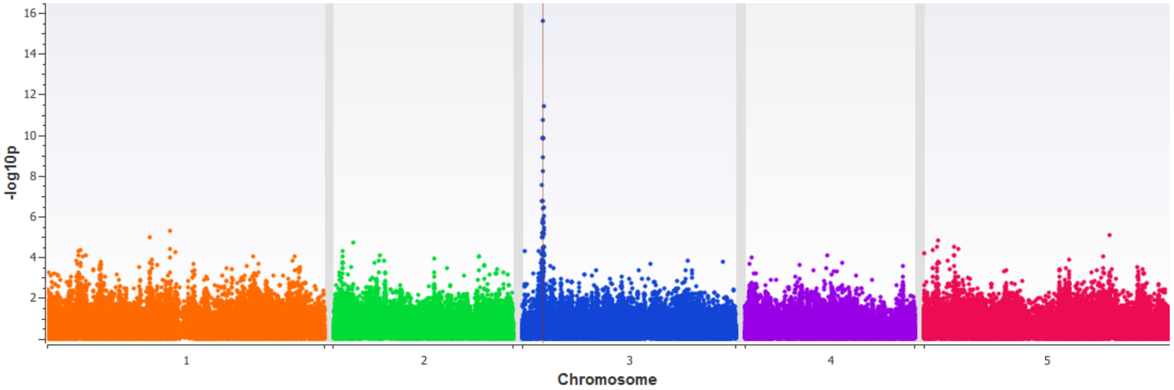
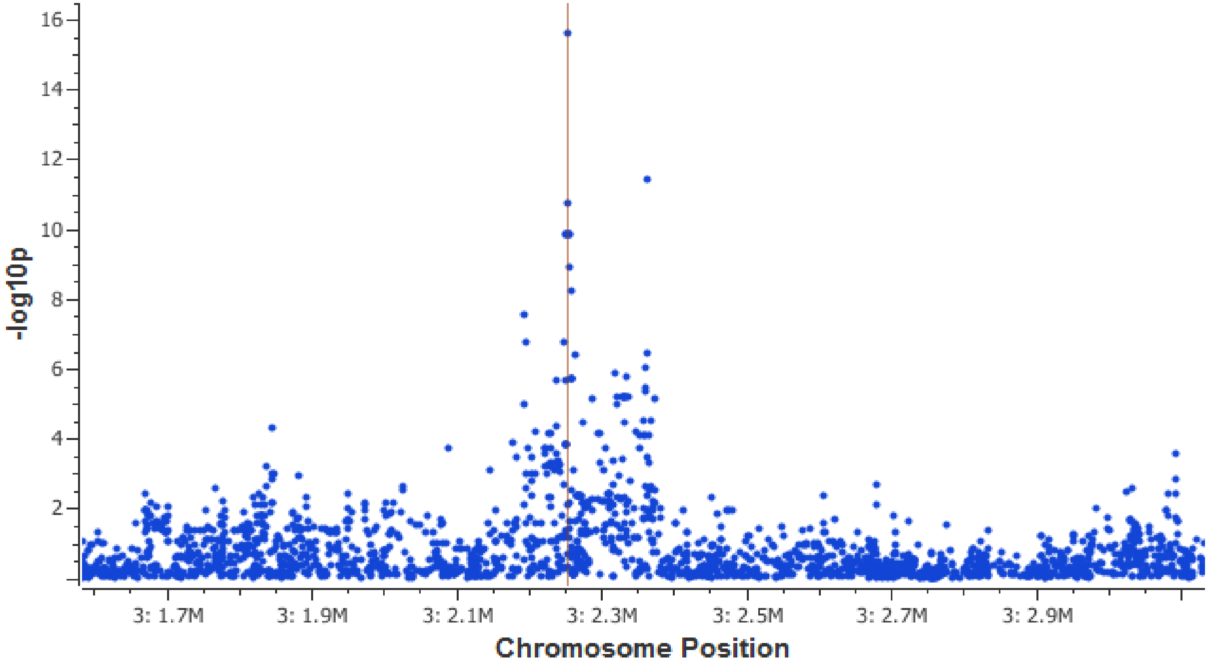


Figure 2 | GWA analysis of hypersensitive response to the bacterial elicitor *AvrRpm1*.

Conclusion


-log10p
16
14
12
10
8
6
4
2
1
2
3
4
5
Chromosome


-log10p
16
14
12
10
8
6
4
2
3: 1.7M
3: 1.9M
3: 2.1M
3: 2.3M
3: 2.5M
3: 2.7M
3: 2.9M
Chromosome Position

Questions or more info:

- Email info@goldenhelix.com
- Request an evaluation of the software at www.goldenhelix.com

"Putting analytic power in the hands of the biologists who are thinking about the disease is tremendously valuable in terms of having things progress rapidly and fully being able to explore the data... That's what SVS gives you.

Peter K. Gregersen, MD
Feinstein Institute for Medical Research

10.30.13
Performing Small-N Sequencing Workflows: Approaches to Analyzing Trio NGS Data
Autumn Laughbaum, Biostatistician

09.10.13
Making NGS Data Analysis Clinically Practical: Repeatable and Time-Effective Workflows
Autumn Laughbaum, Biostatistician

Who is Golden Helix?

Founded in 1998, Golden Helix is known for helping genetic research groups working with large-scale DNA-sequencing or microarray data overcome the frustration and challenges of bioinformatic roadblocks: delayed projects, lack of quality findings, and low productivity. By empowering researchers with highly effective software tools, world-class support, and an array of complementary analytic services, we refute the notion that analysis has to be difficult or time consuming.

GOLDEN HELIX
Accelerating the Quest for Significance™