The effects of human selection on elite tomato germplasm and implications for genome-based selection

David Francis, Sung-Chur Sim , Heather Merk ,
Horticulture and Crop Science, The Ohio State University, Wooster, OH
Allen Van Deynze & Kevin Stoffel ,
Seed Biotechnology Center, University of California, Davis, CA
C. Robin Buell and John Hamilton ,
Biology, Michigan State University, East Lansing, MI
David Douches & Dan Zarka ,
Crop and Soil Sciences, Michigan State University, East Lansing, MI

Solanaceae Coordinated Agricultural Project

THE OHIO STATE UNIVERSITY OARDC

USDA United States Department of Agriculture

NIFA National Institute of Food and Agriculture

ASHS American Society for Horticultural Science

Breeding in a Genomics Era: State of the Art and New Opportunities

Objectives of the colloquium: bridge the gap between breeding, genotyping and MAS for vegetable and horticultural crops

Objectives of talk:

Review SolCAP sequence and genotyping resources

Population genetics of cultivated tomato

Analysis within breeding programs and implications for selection

USDA NIFA United States Department of Agriculture National Institute of Food and Agriculture

ASHS American Society for Horticultural Science

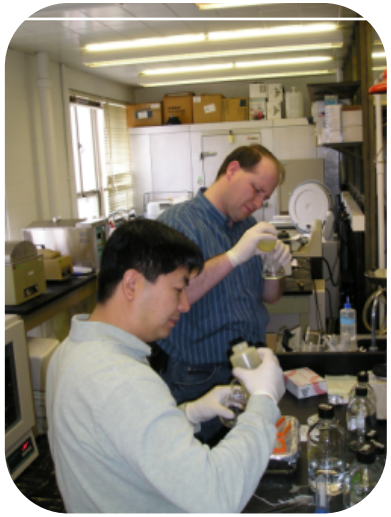
Sequence Resources for Tomato

Community resources

- H1706 genome (Processing)
- LA1589 genome (wild current tomato)
- TA496 EST (Processing)
- Micro-Tom EST (Novelty)

SolCAP resources (GAII cDNA sequence)

- OH068-6405 (fresh-market)
- FL7600 (fresh-market)
- NC84173 (fresh-market)
- OH9242 (processing)
- PI114490 (cherry tomato)
- PI128216 (wild current tomato)



Focus on cultivated types with direct relevance to breeding programs

Single Nucleotide (SNP) Discovery "Pipeline"

Align RNA-Seq reads to tomato scaffolds using Bowtie and call SNPs using SAMTools: **3,730,151 SNPs**

↓

Filter SNPs for read depth and density using the SAMTools utility samtools.pl varFilter: **113,493 SNPs**

↓

Filter SNPs for min SNP quality, consensus quality, mapping quality, and max percentage 3' bases at SNP positions: **86,556 SNPs**

↓

Create list of unique quality filtered SNP positions on tomato scaffolds **57,996 SNPs** (High Confidence SNPs)

Remove multi-allelic SNPs and SNPs within 50bp of an intron: **33,394 SNPs**

↓

Filter out SNPs with a BeadXpress score < 0.75 and an Infinium score < 0.9 : **27,238 SNPs** (Illumina compatible SNPs)

↓

Filter based on frequency among six accessions; add spatial criteria; cross validated w/ INRA and Trait Genetics

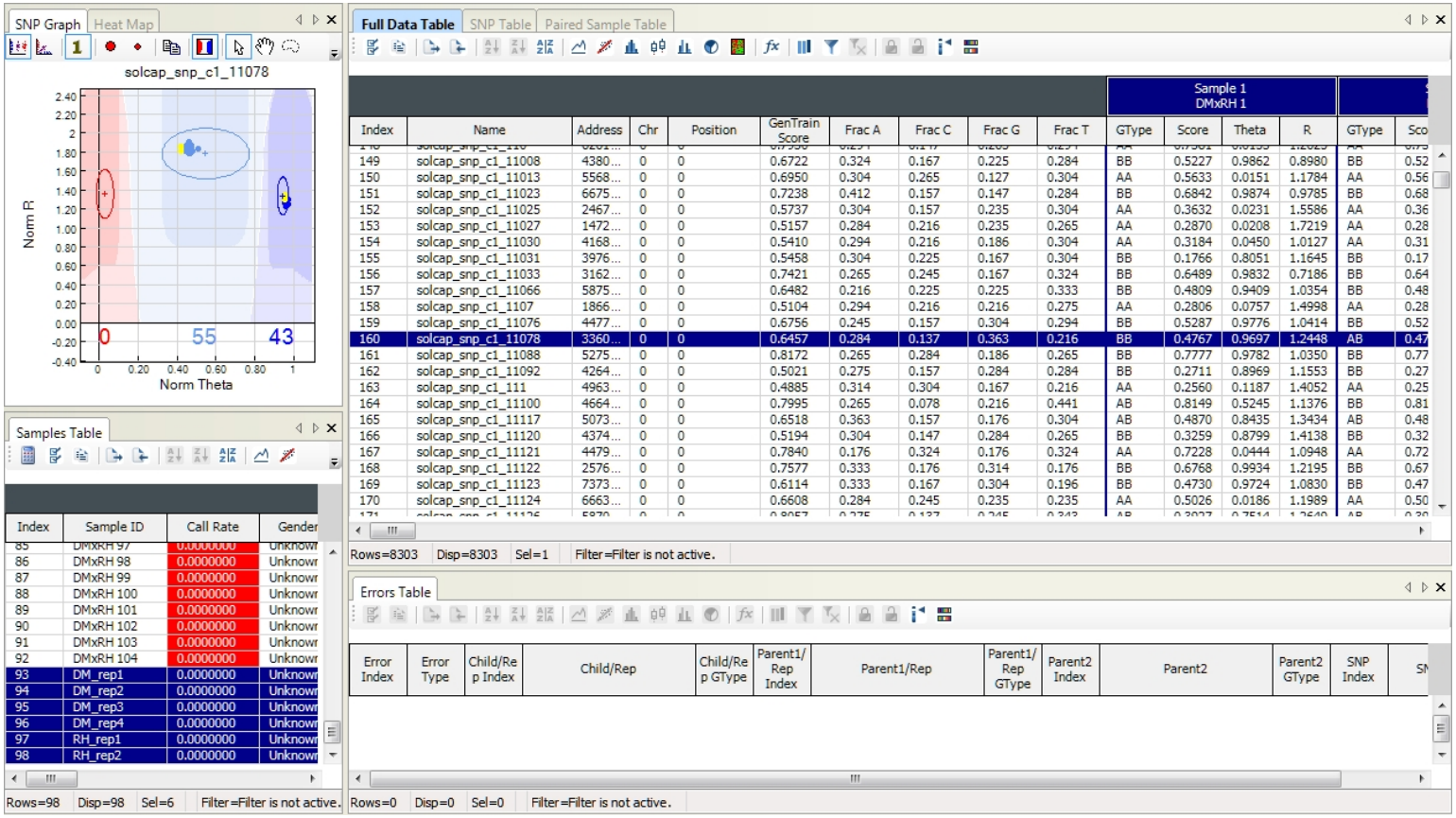
↓

8,784 SNPs (**7,720 SNPs** passed manufacture/ production QC and are "scoreable") illumina

Hamilton et al., The Plant Genome 2012 5:17-29

Tomato Infinium SNP Array

- 7,720 total & 6,167 SNPs in cultivated varieties



Sol CAP

TraitGenetics

INRA

illumina

Martin Ganal **Mathilde Causse** **Cindy Lawley**

Tomato Infinium SNP Array

- 7,720 total & 6,167 SNPs in cultivated varieties

SNP design/annotation file: **http://solcap.msu.edu/**

Cluster File: **http://www.extension.org/pages/61007**

*** Illumina orders through Dec. 2012**

Putting research into practice

Detail on sequencing, selection and validation of SNPs

Hamilton et al. 2012. Single Nucleotide Polymorphism Discovery in Cultivated Tomato via Sequencing by Synthesis. The Plant Genome 2012 5:17-29
https://www.crops.org/publications/tpg/articles/5/1/17

Sequence reads deposited into NCBI Short Read Archive (SRA)
Accession: SRX111862
Accession: SRX111861
Accession: SRX111859
Accession: SRX111858
Accession: SRX111857
Accession: SRX111853
Accession: SRX111850
Accession: SRX111849
Accession: SRX111848
Accession: SRX111845
Accession: SRX111558
Accession: SRX111557
Accession: SRX111556

How well do the SolCAP SNPs work?

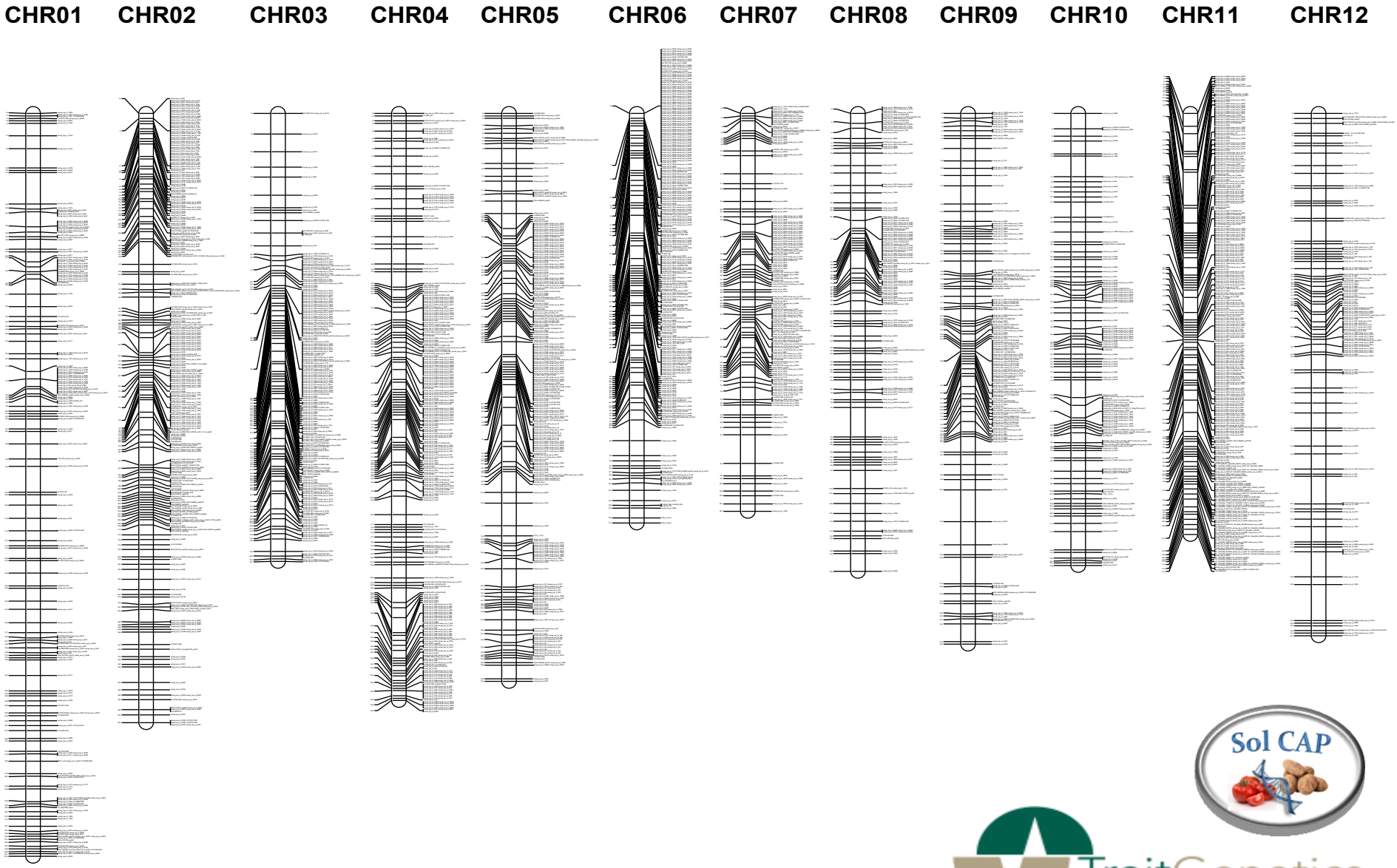
Application to mapping

Tomato Infinium SNP Array

- 3,503 PM (Sl x LA716)
- 4,491 SNPs (Sl x Spimp)

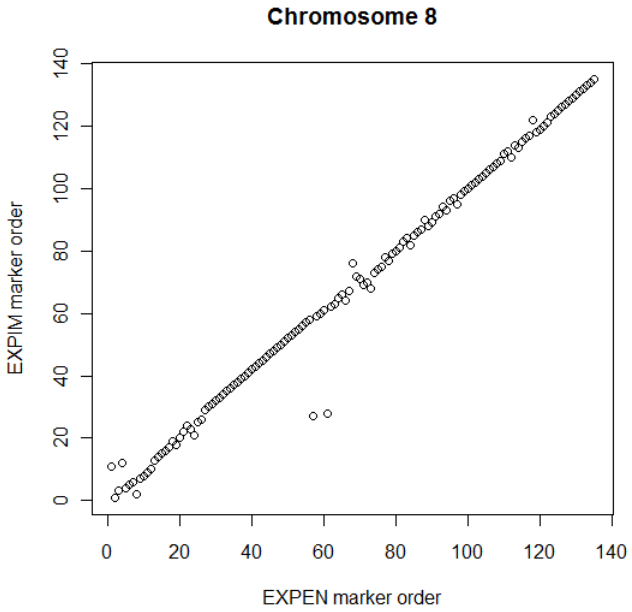
TraitGenetics

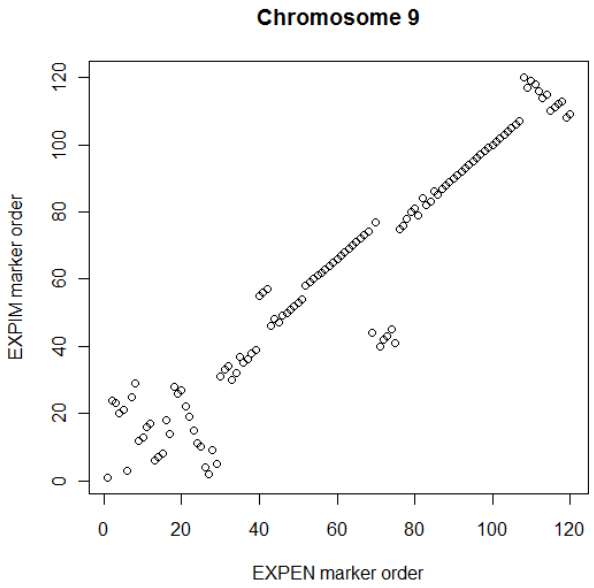
Tomato genetic map EXPIM 2012 using 4,491 SNP markers

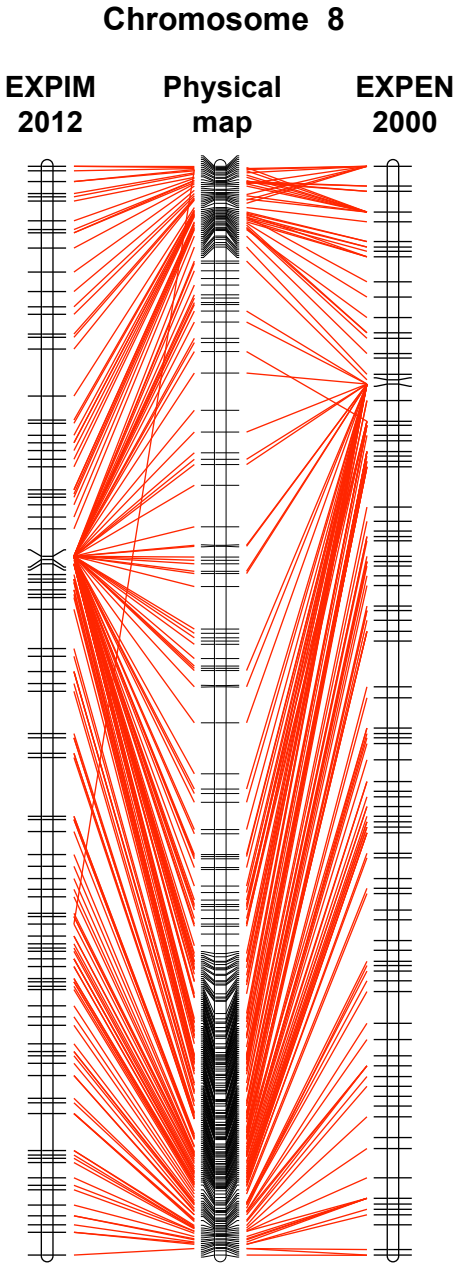


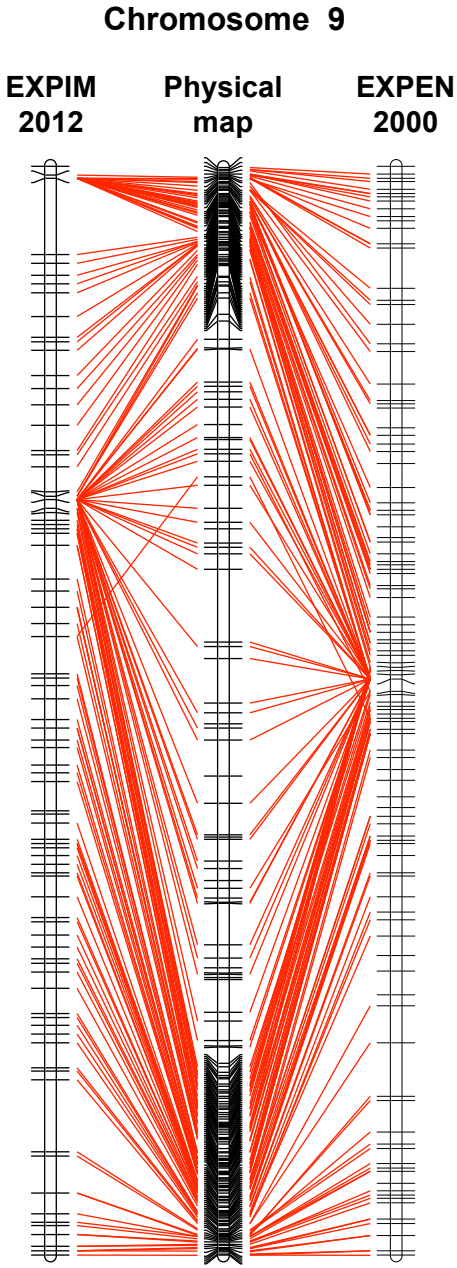
Sol CAP

TraitGenetics



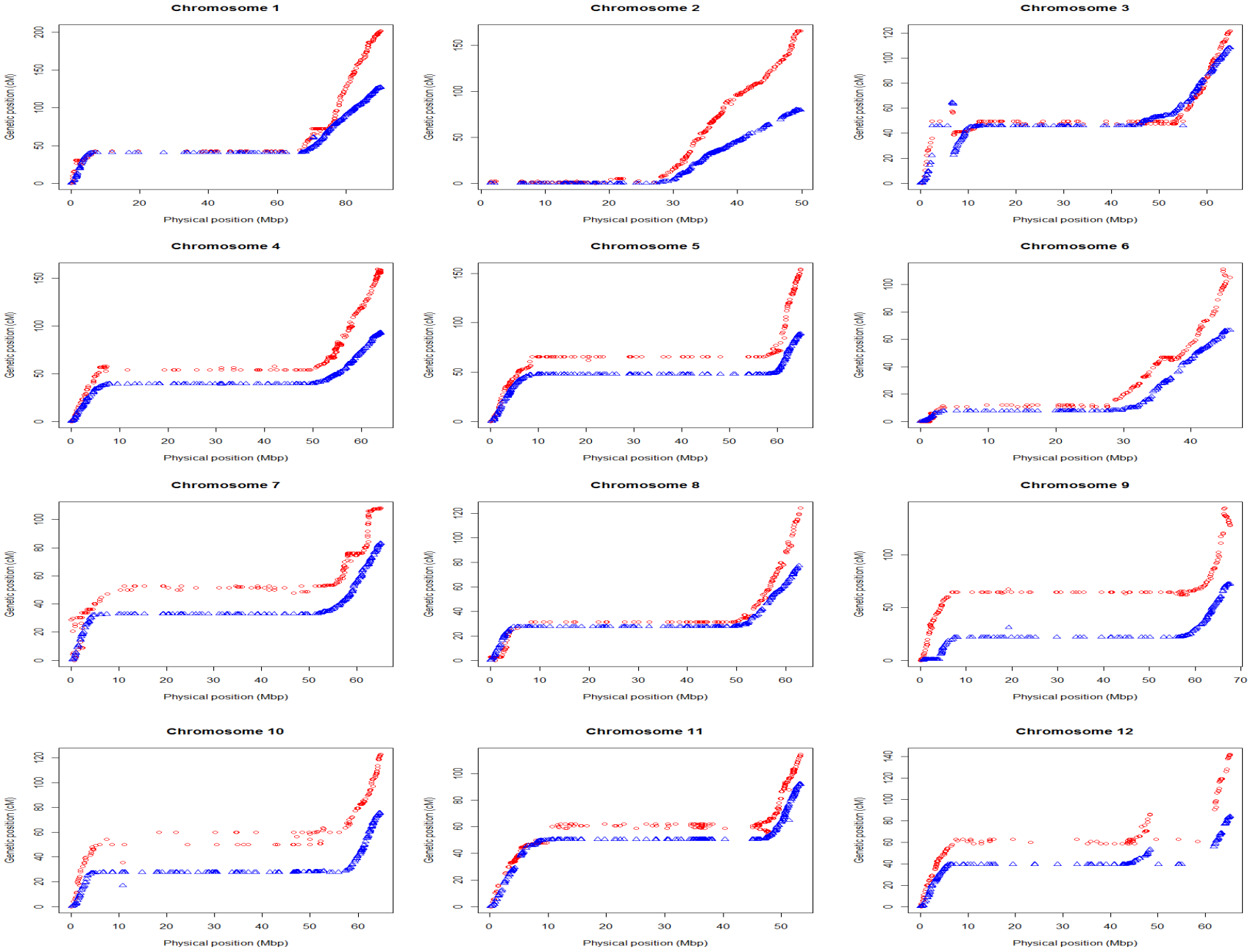






Map quality and potential inversions

Comparison of genetic maps with sequence assembly



Details:

Sim et al., 2012. Development of a Large SNP Genotyping Array and Generation of High-Density Genetic Maps in Tomato. PLoS ONE 7(7): e40563. doi:10.1371/journal.pone.0040563

http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0040563

TraitGenetics

How well do the SolCAP SNPs work?
Application to genotyping in cultivated tomato

Accessions

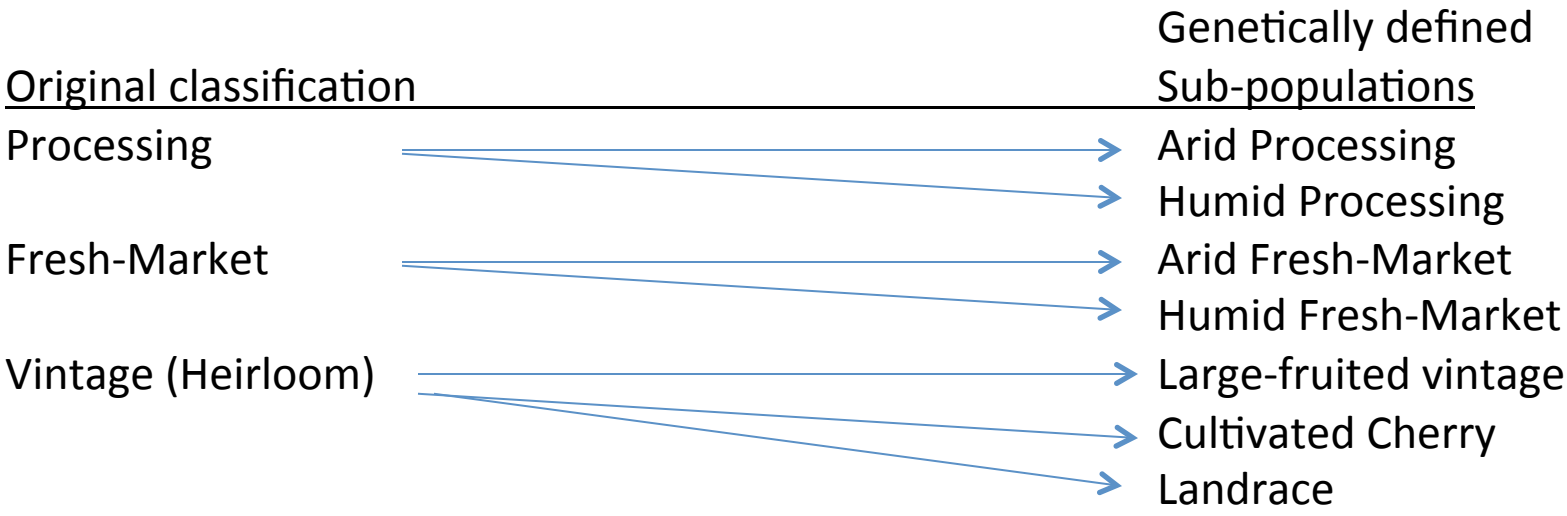
- 144 Fresh Market
- 144 Processing
- 2 Greenhouse
- 40 land races
- 3 unimproved
- 48 vintage (heirloom)
- 48 *cerasiformae*
- 64 wild species
- 11 hybrids

Parents of key resource Populations

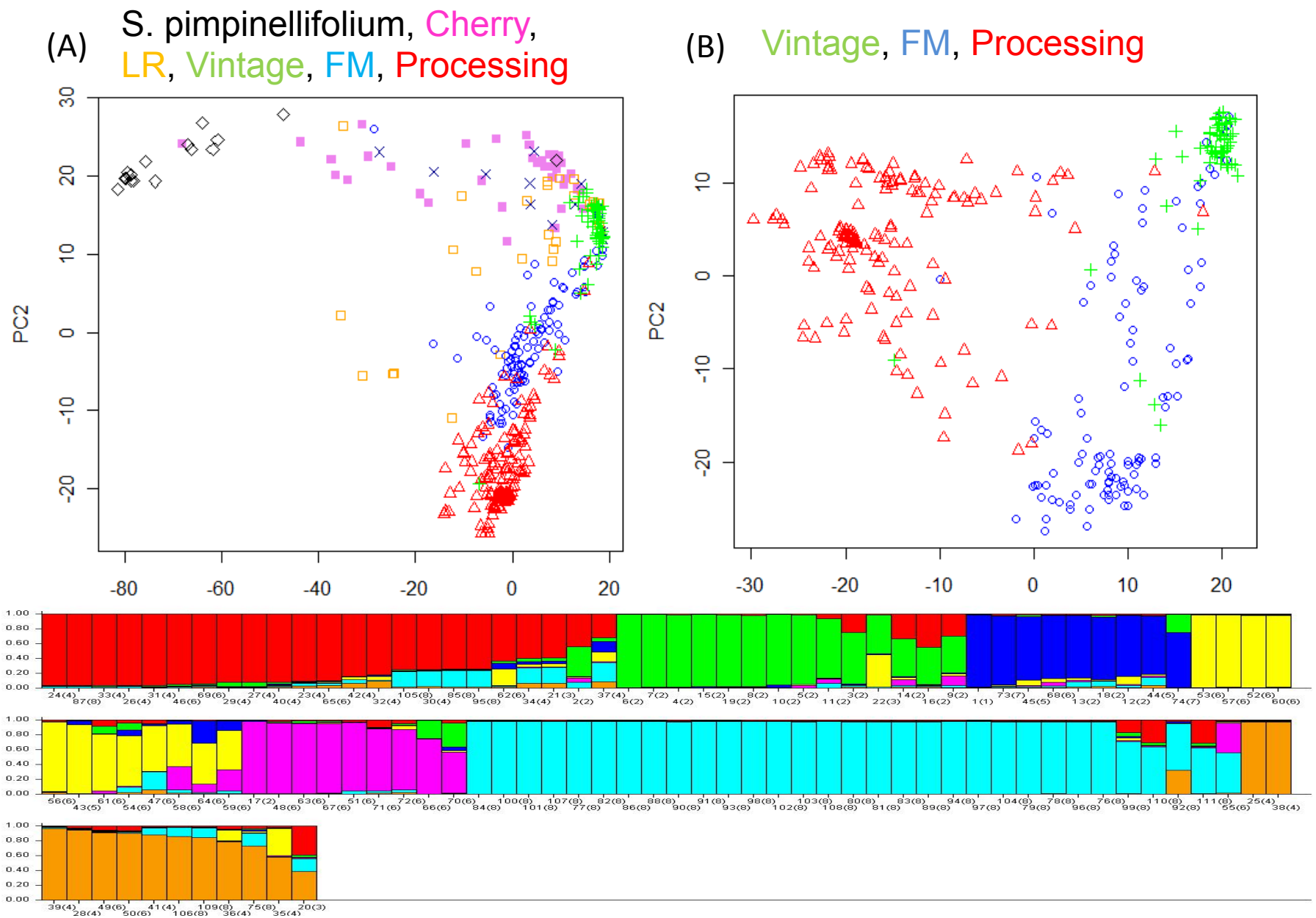
- M82 x LA716
- Hunt 100 x LA407
- NC 23E X LA1269
- Sun1642 x LA1589
- OH 88119 x PI128216
- OH 8245 x PI365914
- (Fla7600 xPI114490) xOH9242
- OH 88119 x Ha 7998
- OH 9242 x OH 8245
- OH 9241 x OH 8245

Population analysis

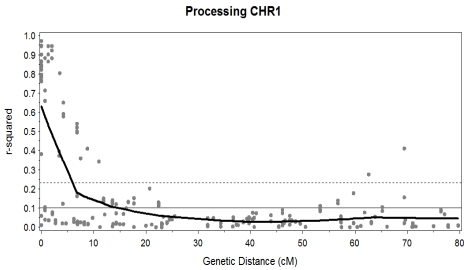
386 cultivated varieties

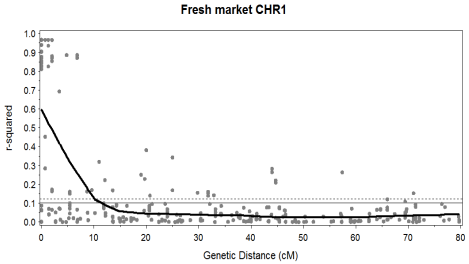


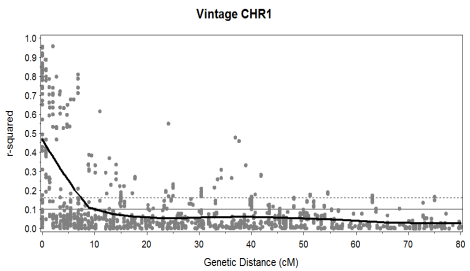
Sub-populations defined by PCA and model-based clustering (STRUCTURE)

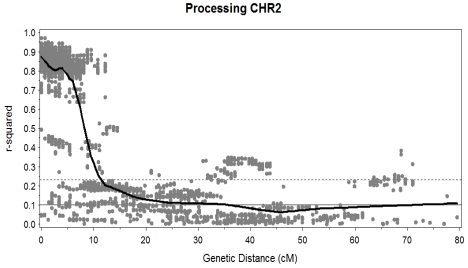


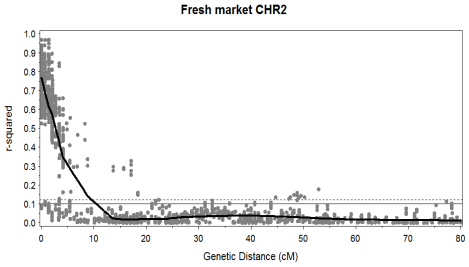
LD across each chromosome for processing, fresh market, and vintage varieties (n = 386 cvs).

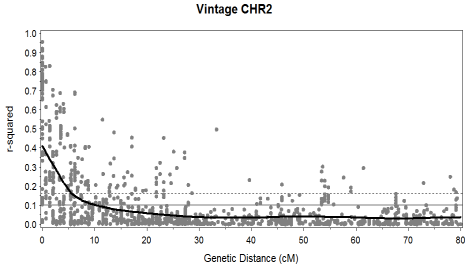


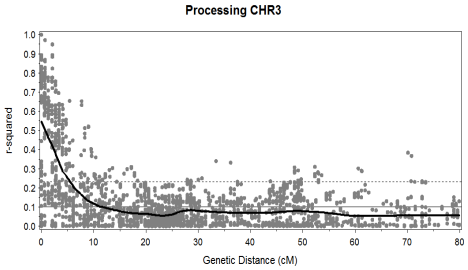


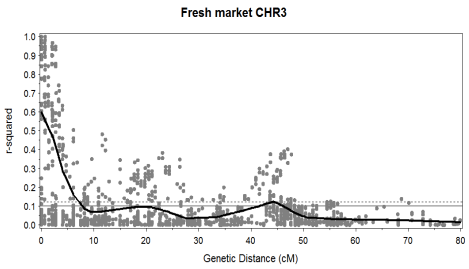


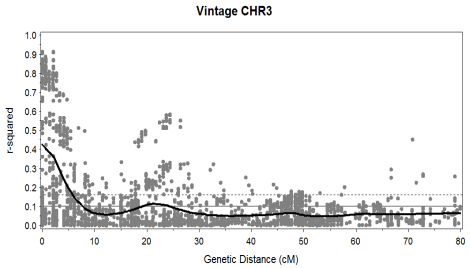


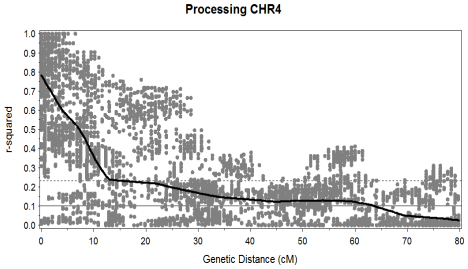


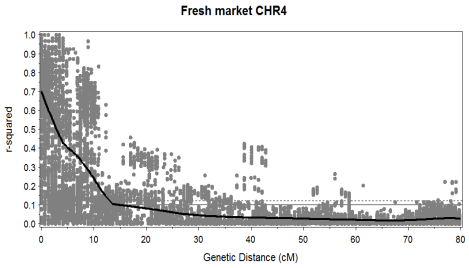


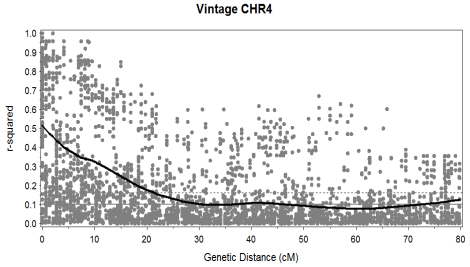








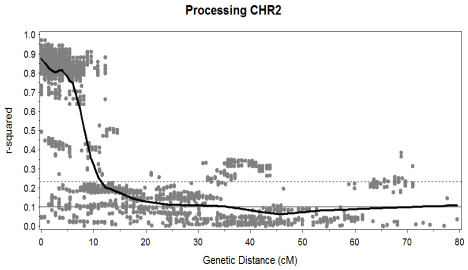


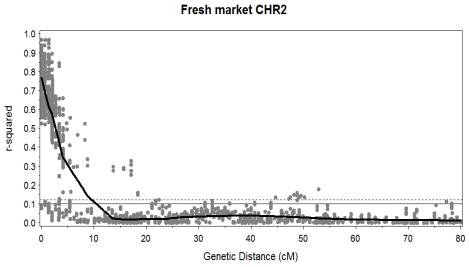


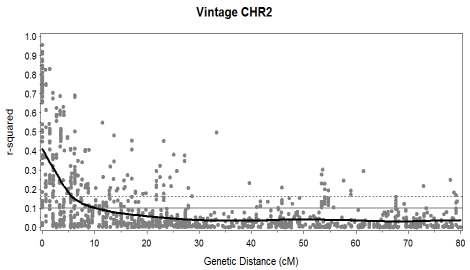
Note: The vintage group here includes Landrace and cherry type varieties.

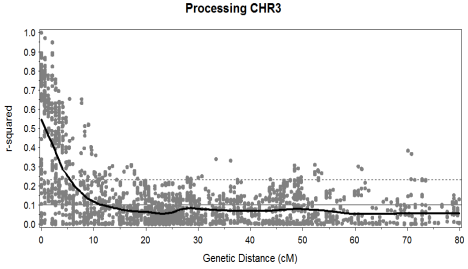
LD across each chromosome for processing, fresh market, and vintage varieties (n = 386 cvs).

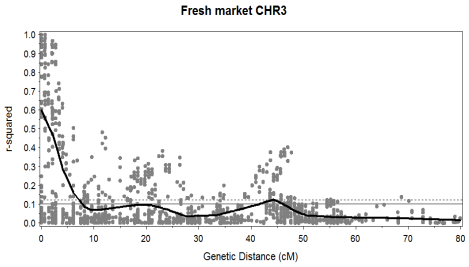
LD decays over cM (2-10) = recombination is limiting
LD decay varies chromosome-chromosome
LD decay reveals signals of sub-population structure

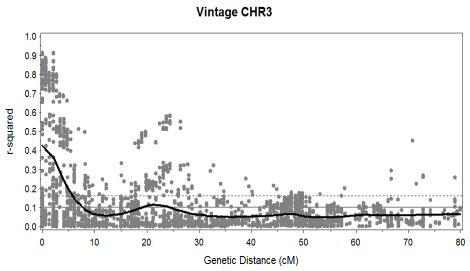


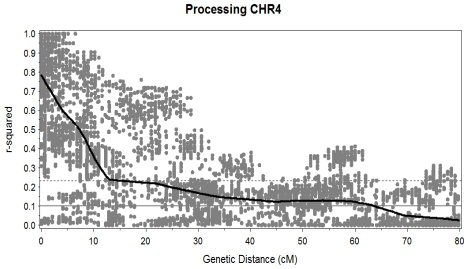


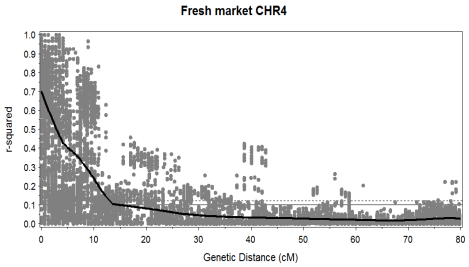


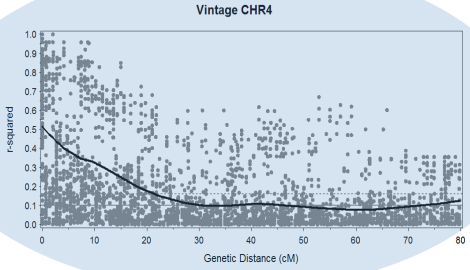








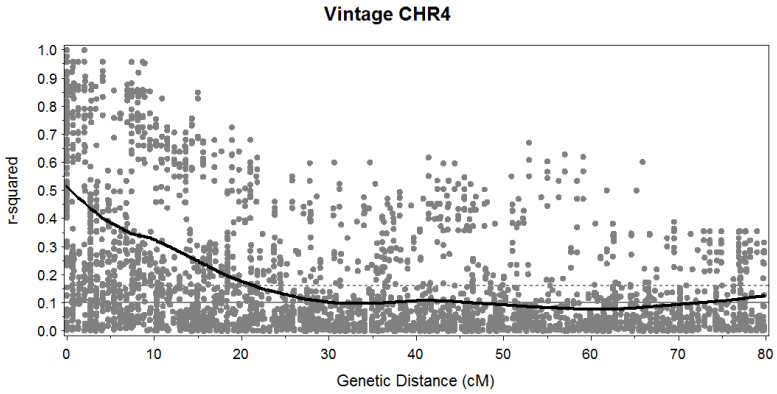




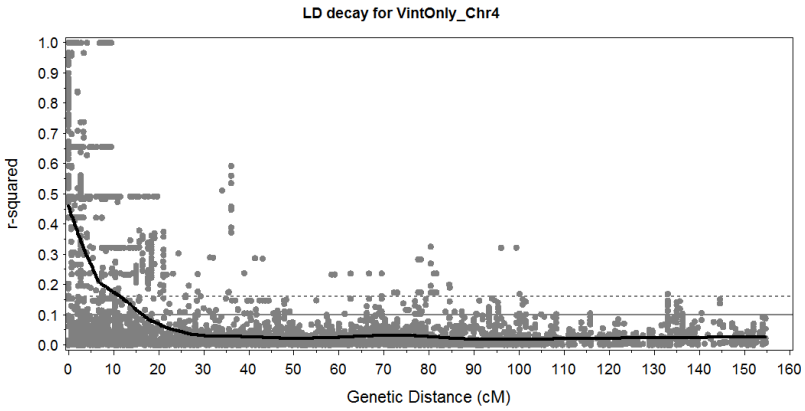
Note: The vintage group here includes Landrace and cherry type varieties.

Signals of structure in data and analysis driven identification of distinct sub-populations within "core" collections

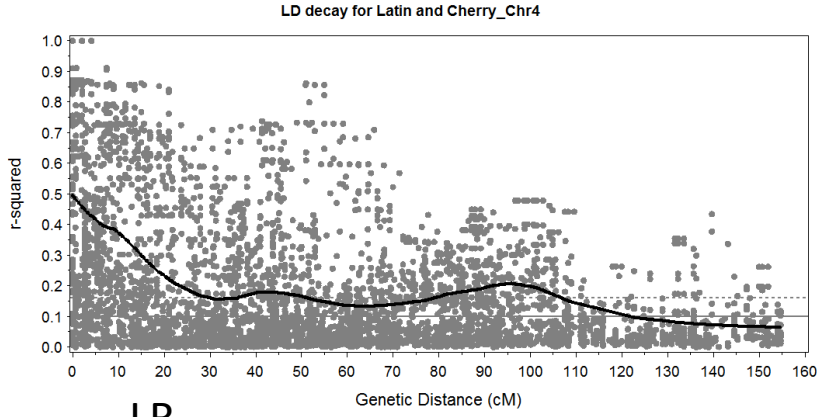
Vintage Varieties

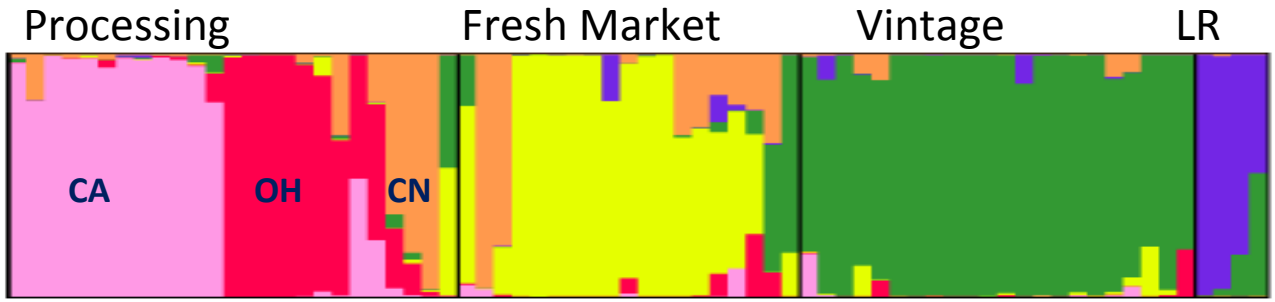


N. American and European vintage



Latin American Landrace (includes large fruited and cherry accessions)





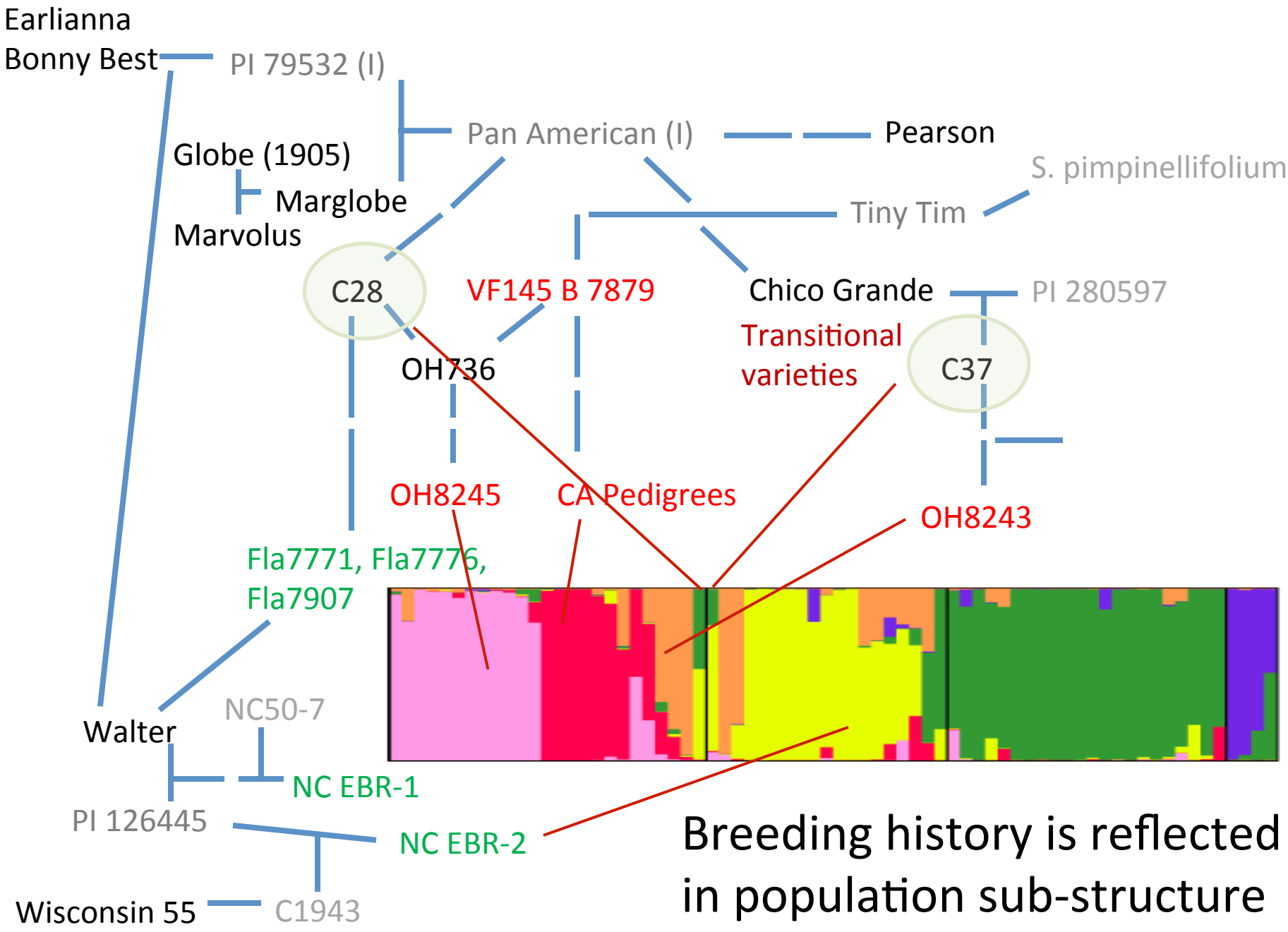
Sim, S-C., et al. 2011. *Heredity*. Vol. 17

Pairwise estimates of Fst (θ) between sub-populations

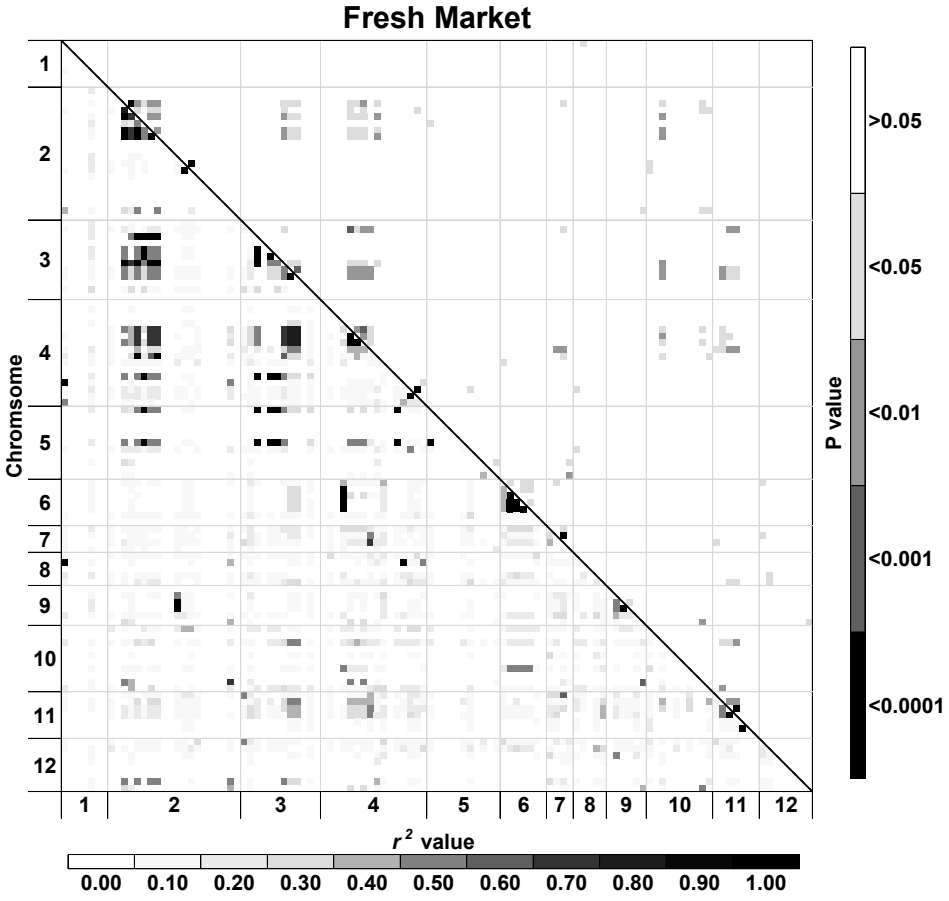
Sub-population	Proc		FM		Vintage	Cherry	Landrace	Wild cherry	Pimp
Processing (Proc)	0.00		0.29**		0.41**	0.27**	0.40**	0.38**	0.72**
Fresh marekt (FM)			0.00		0.27**	0.18**	0.28**	0.29**	0.72**
Vintage					0.00	0.13**	0.18**	0.20**	0.81**
Cultivated cherry (Cherry)						0.00	0.04^{NS}	0.05*	0.58**
Landrace							0.00	0.04^{NS}	0.64**
Wild cherry								0.00	0.57**
S. pimpinellifolium (Pimp)									0.00
Further division within sub-population	Proc 1	Proc 2	FM 1	FM 2	Vintage	Cherry	Landrace	Wild cherry	Pimp
Proc 1	0.00	0.27**	0.42**	0.40**	0.52**	0.34**	0.49**	0.44**	0.74**
Proc 2		0.00	0.52**	0.32**	0.47**	0.30**	0.45**	0.38**	0.75**
FM 1			0.00	0.32**	0.52**	0.34**	0.49**	0.42**	0.76**
FM 2				0.00	0.12**	0.10**	0.17**	0.19**	0.72**

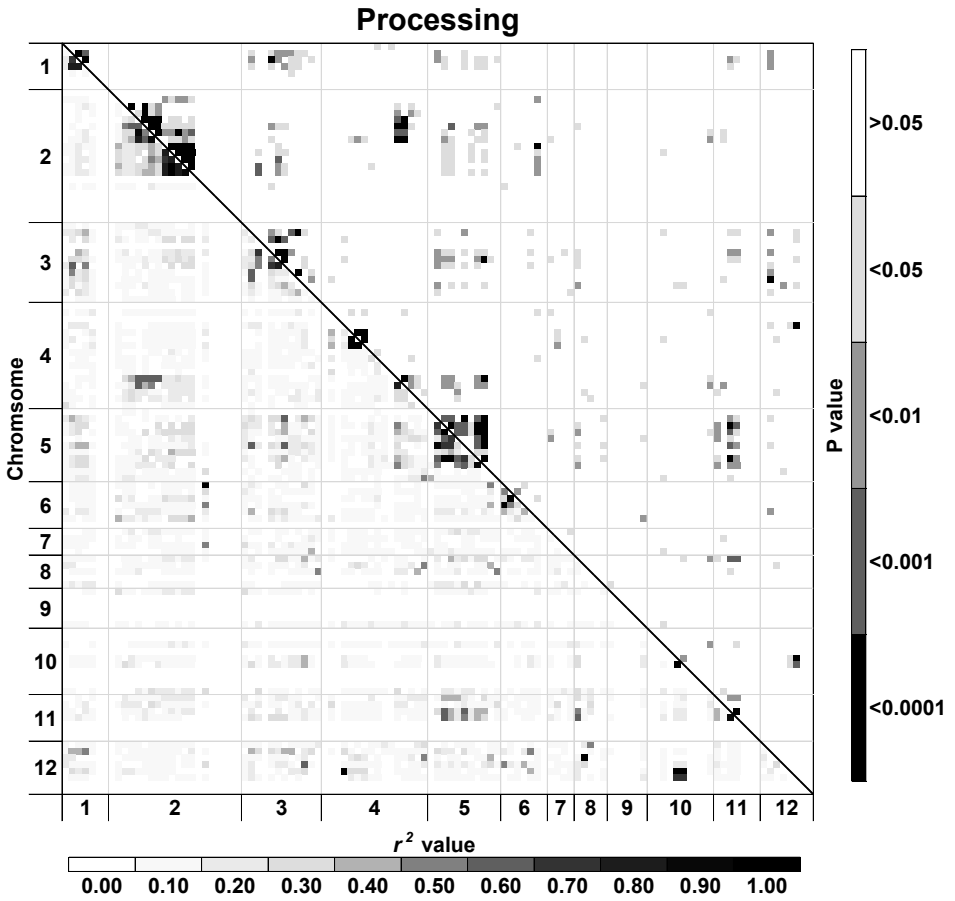
Multiple lines of evidence suggest that there are distinct sub-populations

PCA, STRUCTURE, F_{ST}


Earlianna
Bonny Best
PI 79532 (I)
Pan American (I)
Pearson
Globe (1905)
Marglobe
Marvolus
S. pimpinellifolium
Tiny Tim
C28
VF145 B 7879
Chico Grande
PI 280597
Transitional
varieties
OH736
C37
OH8245
CA Pedigrees
OH8243
Fla7771, Fla7776,
Fla7907
NC50-7
Walter
NC EBR-1
PI 126445
NC EBR-2
Wisconsin 55
C1943
Breeding history is reflected
in population sub-structure

Heat maps of LD show associations that are not due to physical linkage





Conclusion: Plant breeders select for combinations of genes
("oh, I knew that…" Dan Ariely, The Center for Advanced Hindsight)

Matt Robbins

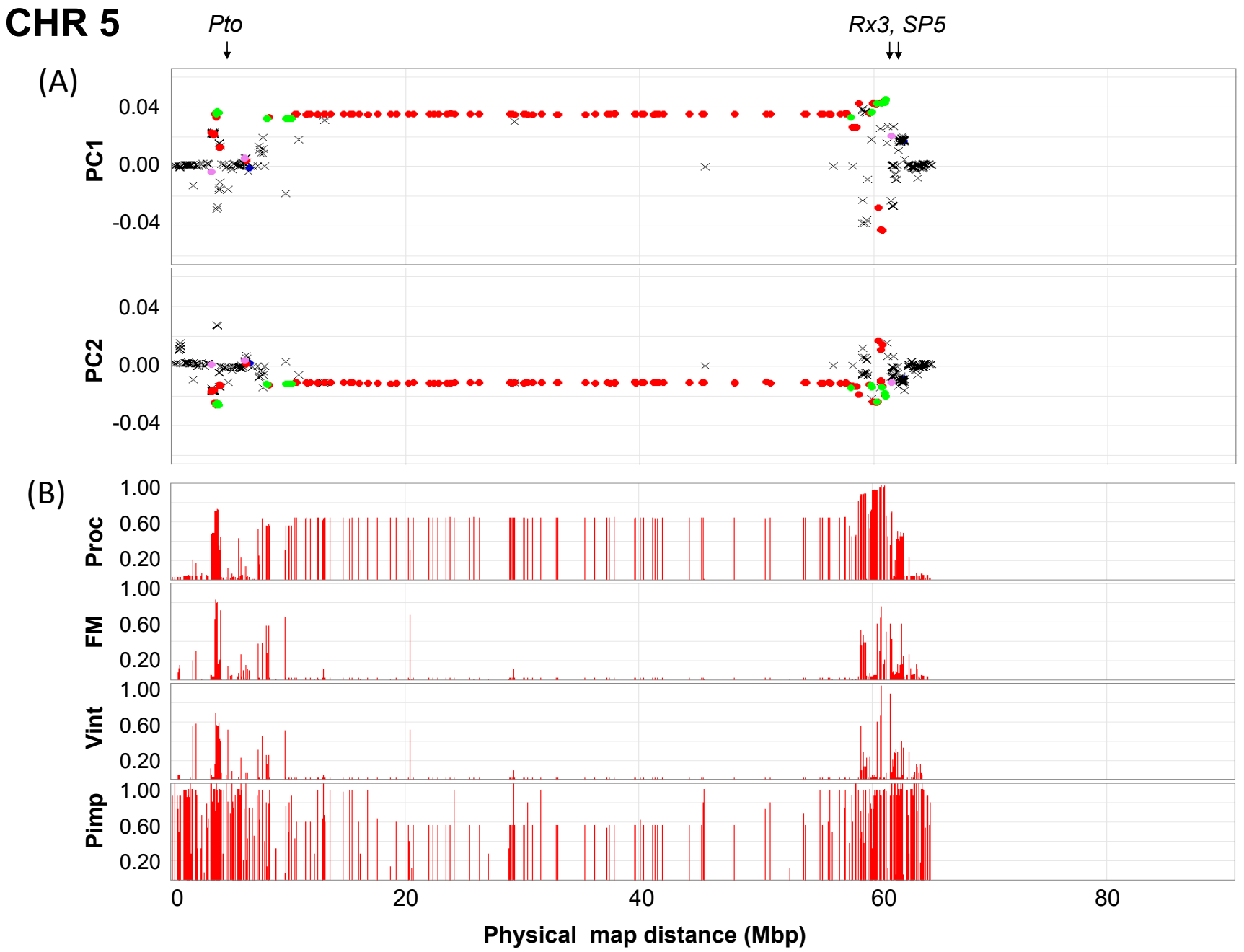
Can we detect where selection has occurred in the genome?

Approach requires comparison (e.g. Fresh-Market vs Processing; Arid vs Humid; etc…).

1) SNPs with high positive/negative PCA loadings (regions of the genome that explain observed variation)

2) F_{ST} outlier: LOSITAN (LOoking for Selection In a TANgled dataset): compares F_{ST} and expected He under neutral expectations. Antao et al., 2008 *BMC Bioinformatics*, **9**:323

3) Differences in allele frequencies


CHR 5
Pto
Rx3, SP5
(A)
PC1
0.04
0.00
-0.04
PC2
0.04
0.00
-0.04
(B)
Proc
FM
Vint
Pimp
1.00
0.60
0.20
0
20
40
60
80
Physical map distance (Mbp)

What genes might be under selection?

Limitation: insufficient recombination to identify candidate genes (though there are plenty!)

Fresh-Market VS Processing

Plant habit (internode length Sp5)	Chromosome 5
Disease resistance (PTO)	Chromosome 5
Disease resistance (Mi, TY, Cf2, Cf5) Chromosome 6	
Disease resistance (I, I3, Ph-1)	Chromosome 7
Fruit size and shape (lc, ovate, SUN, fasciated)	Chrom. 2, 7, 11
Disease resitance (Tm2^2, Frl, Sw5, Ph-3)	Chromosome 9
Disease resistance (I2, Rx-4, Xv3)	Chromosome 11
????	Chromosome 4

CA processing VS OH Processing

Disease resistance (Mi, I2, Rx-3, Rx-4)	Ch. 5, 6, 11




How can we apply these tools and this information to crop improvement?

Can sub-population data based on inbred lines predict hybrid performance?

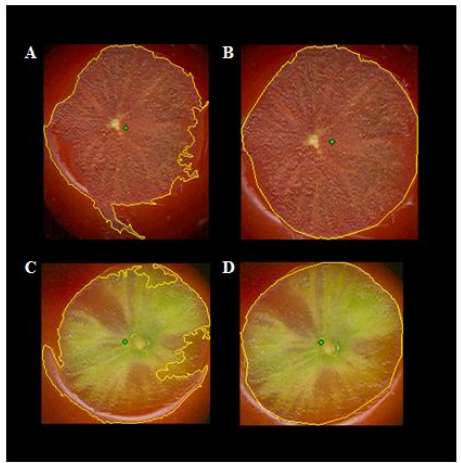
Can we detect associations between traits of economic importance and SNPs?

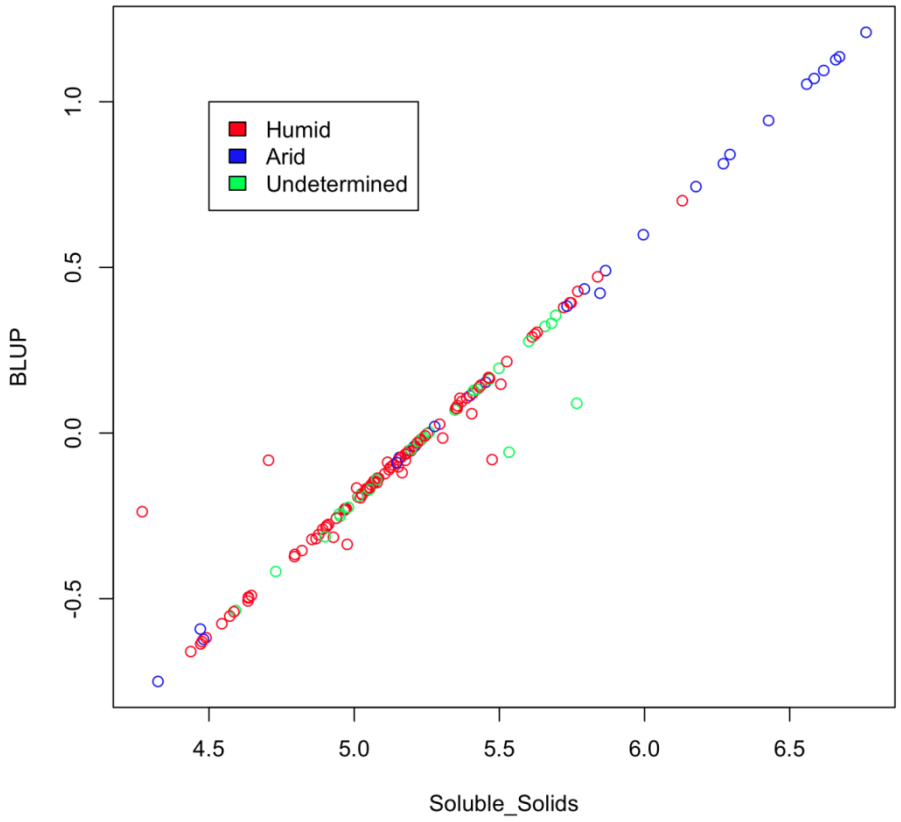
Traits:

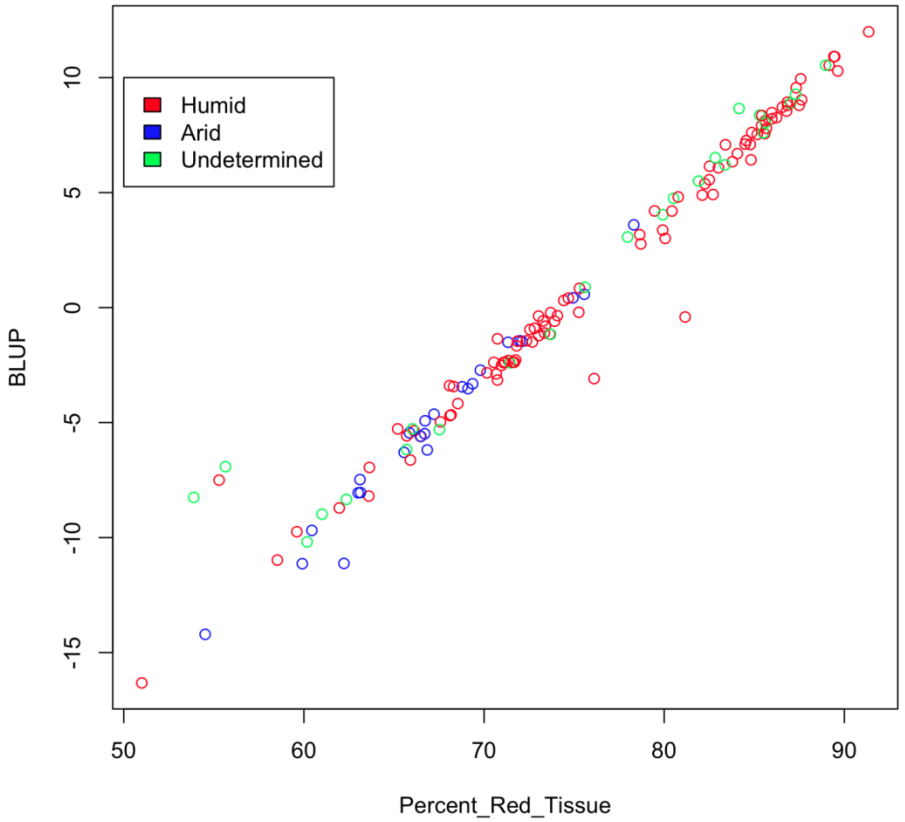
Total traits measured: 52

Reduced to 22 most informative

Yield (total and marketable)
Color and Color uniformity
BRIX
pH
Titratable acid
Fruit Size and Shape



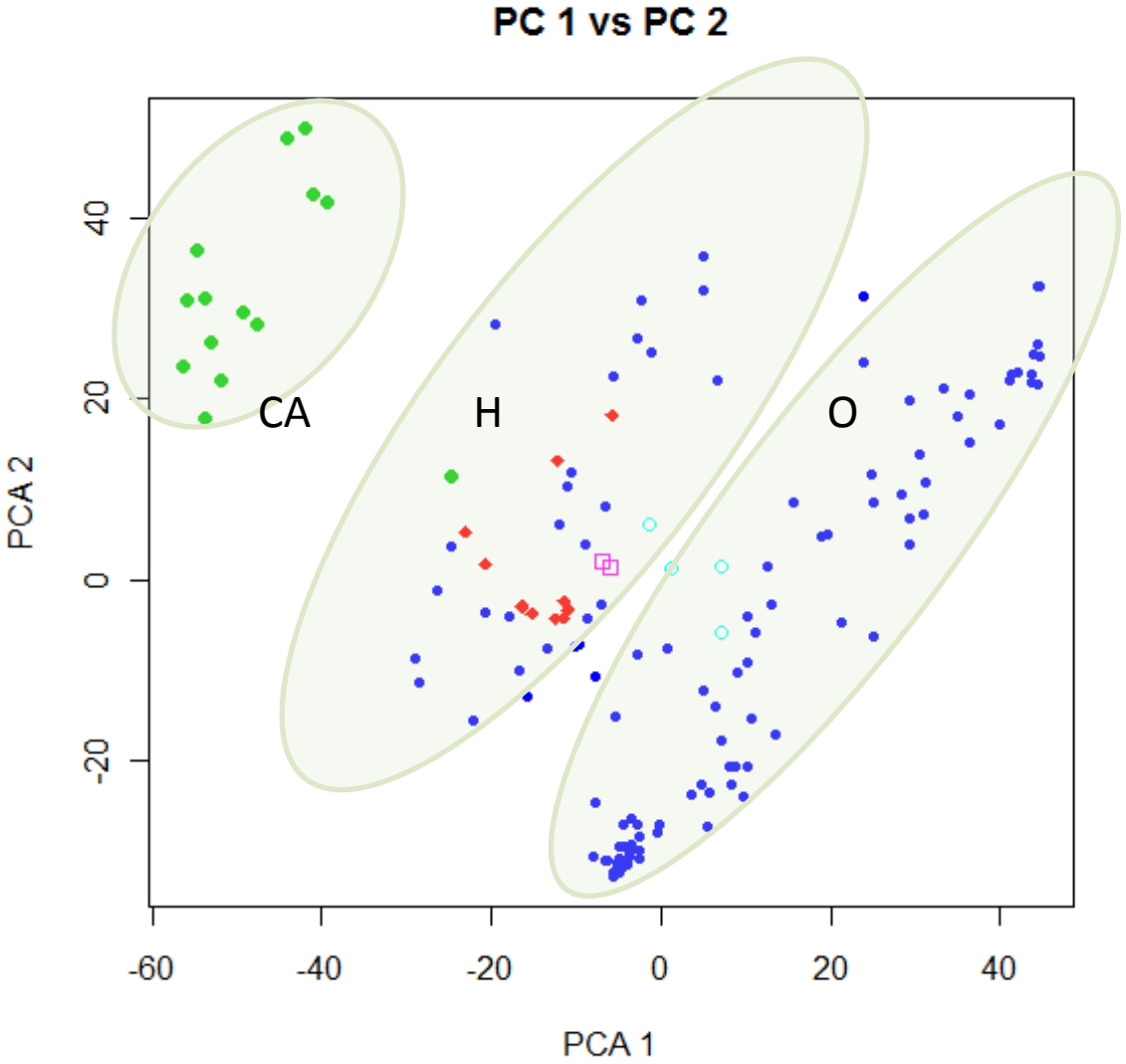




- SolCAP processing tomato collection.
- Comparison of average value with Best Linear Unbiased Predictor (BLUP); color coded by origin of accessions.
- Germplasm from different regional breeding programs contains significant variation for traits of value.

Substructure in Processing collection

California
Ohio
Undeterimined
Ridgetown College, Ontario
Oregon



Genotyping strategies (balancing information, cost, time)

- Genotyping by sequencing
  reduced representation ($50/sample)
- Genotyping using the Infinium array ($100 sample)
- Optimized pools of 384 SNPs for community mapping projects
  (BeadXpress and Kbio platforms)
  http://www.extension.org/pages/61007/

Process:
Select SNPs based on Polymorphic information content (PIC) in target germplasm pools (Processing, Fresh-Market)

Select SNPs based on genetic map position

Fill-in based on physical position

Does Population Genetic Data Help Predict F1 Performance?

Genome Wide Variation

Significance due to hybrid vs inbred comparisons

Little evidence for heterotic groups

Evidence for role of adaptation (difference between Mkt and Total yield)

Class of Cross	Normalized Yield	Marketable Yield	Total Yield	
OxO	1.05	40.26	49.81	
OxH	1.02	37.14	49.48	}
CAxCA	1.00	33.63	47.32	
OxCA	0.97	34.41	46.22	} RIL
HxCA	0.94	33.90	44.83	
HxH	0.98	35.33	43.78	
iO	0.88	33.66	43.40	
iH	0.90	34.18	42.37	
iCA	0.81	25.72	39.00	
P	<.0001	0.001	0.001	
LSD 0.05	0.116	5.821	6.05	

Predicting F1 performance

Single loci

Genotype	Normalized Yield	Marketable Yield	Total Yield
AT	1.01	36.52	47.53
AA	0.98	35.43	47.08
TT	0.93	33.58	44.04
P	0.008	0.067	0.016
LSD 0.05	0.069	3.469	3.53

Association approaches are somewhat limited by structure;
Confirmation using RILs

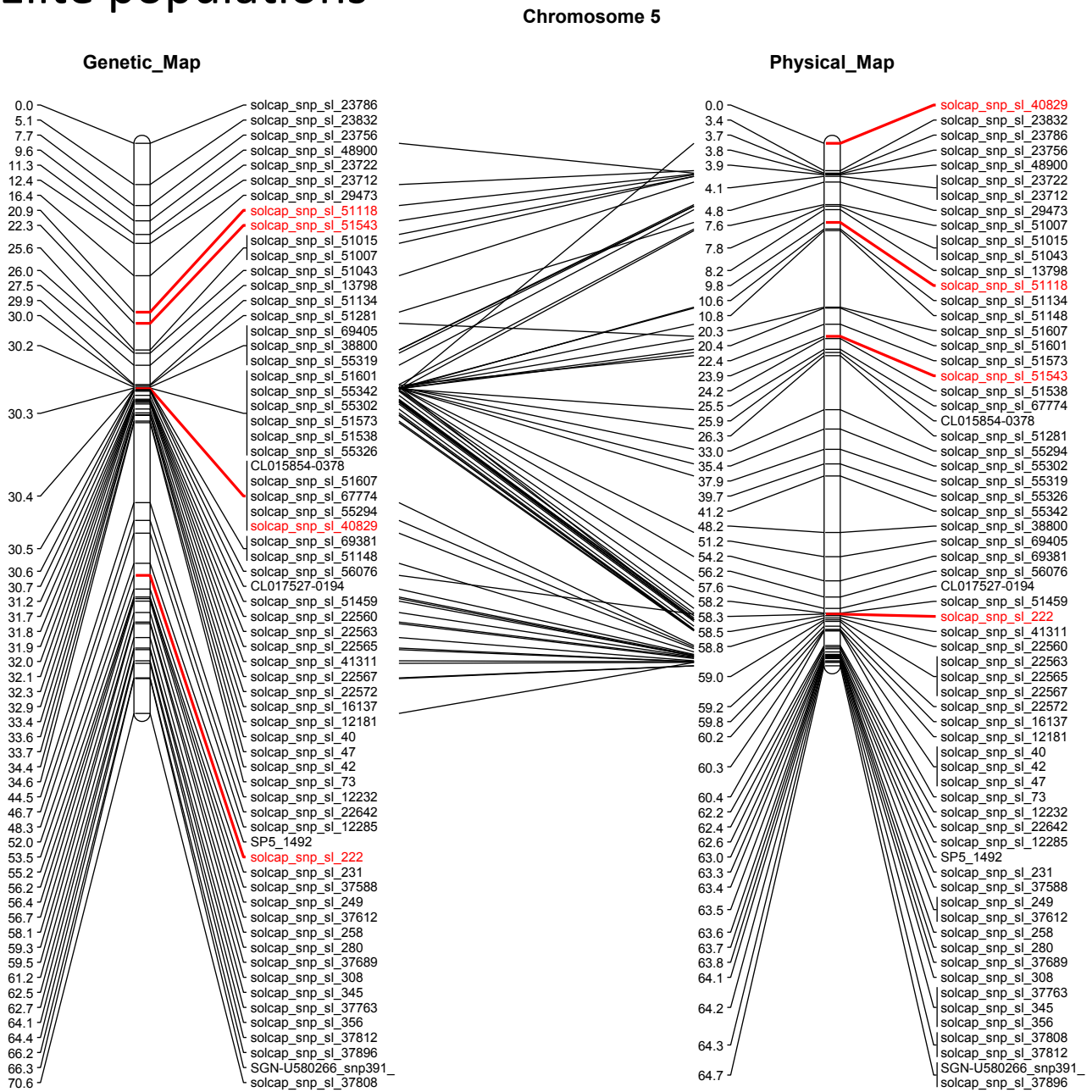
Association mapping in nested RIL
AxB; CXB; AXD
(OH2641 x OH987034; OH7814 x OH987034; OH2641 x OH981136)

288 progeny
Augmented Experimental design
http://www.extension.org/pages/60430/

Mapping in Elite x Elite populations

>95% correspondence between physical map and recombination map

(recombination is limiting; cannot rule out small rearrangements or SNPs in small gene family)



Association Analysis (Unified Mixed Model)

```
%macro Mol(mark);
proc mixed covtest data = three;
class &mark gen;
model T1 = Q1 Q2 Q3 &mark / solution;
random gen / type = lin(1),data=KIN;
%mend;

        %Mol(SL144);
        %Mol(CT10737I);
        %Mol(CT20244I);
        %Mol(SL10525);
        %Mol(SL10526);
         etc…
```

Trait BLUP

Q Matrix (STRUCTURE or PCA)

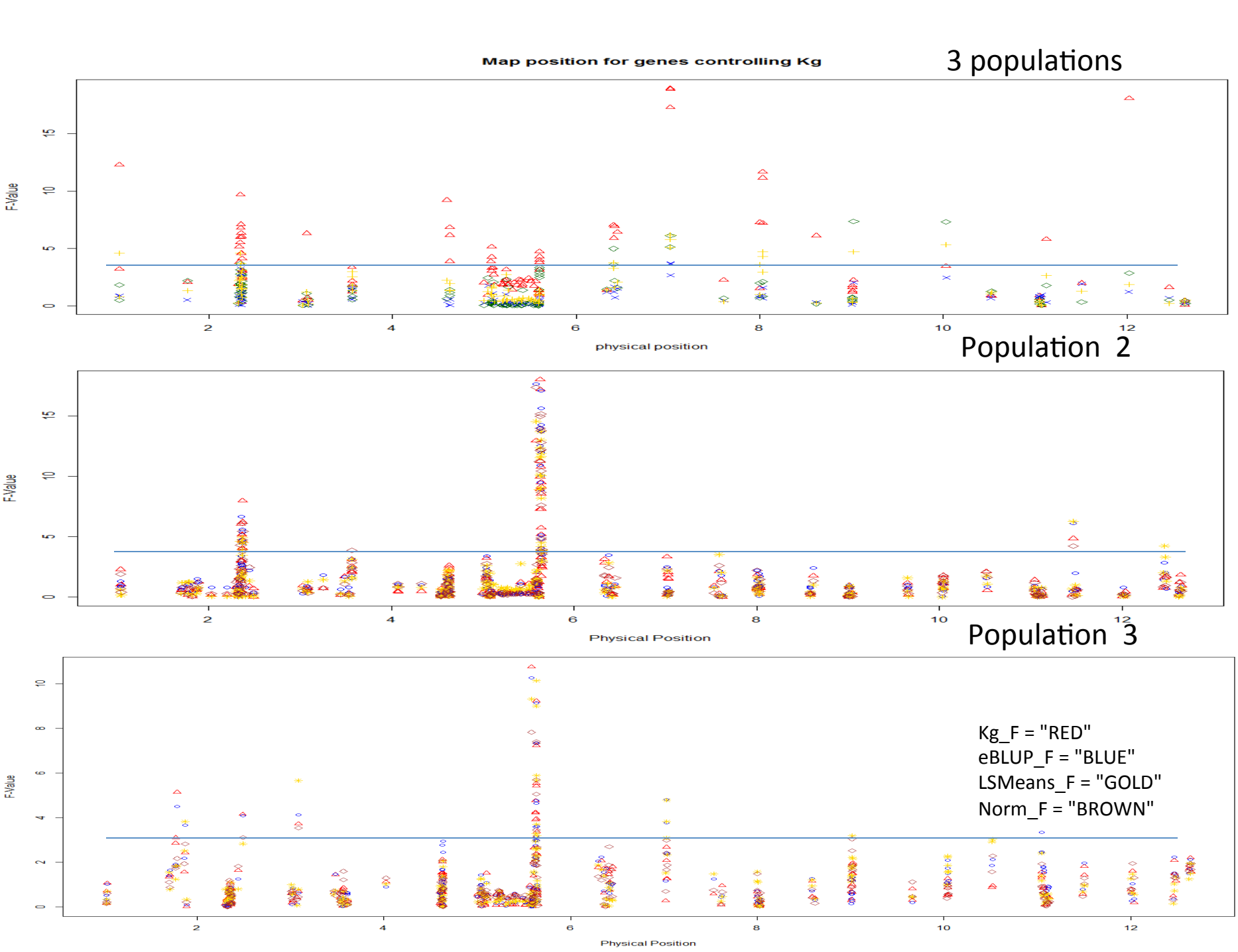
Marker

Kinship matrix (diagonal)

Models can estimate the contribution of STRUCTURE and Kinship to the trait and Marker-Trait linkage…

F-test for significance = $N(1-2r)^2g^2$

N = population size; r = recombination distance (marker to QTL; g^2 = proportion of variance explained by QTL


Map position for genes controlling Kg
3 populations
F-Value
physical position
Population 2
F-Value
Physical Position
Population 3
F-Value
Physical Position
Kg_F = "RED"
eBLUP_F = "BLUE"
LSMeans_F = "GOLD"
Norm_F = "BROWN"

Strong associations with yield on chromosome 2 and 5.

Candidate genes exist in these regions.

Recombination limits our ability to prove role of candidate genes.

Nest step: use of SNP resources for selection.

Alternative to MAS: Genome Wide Selection

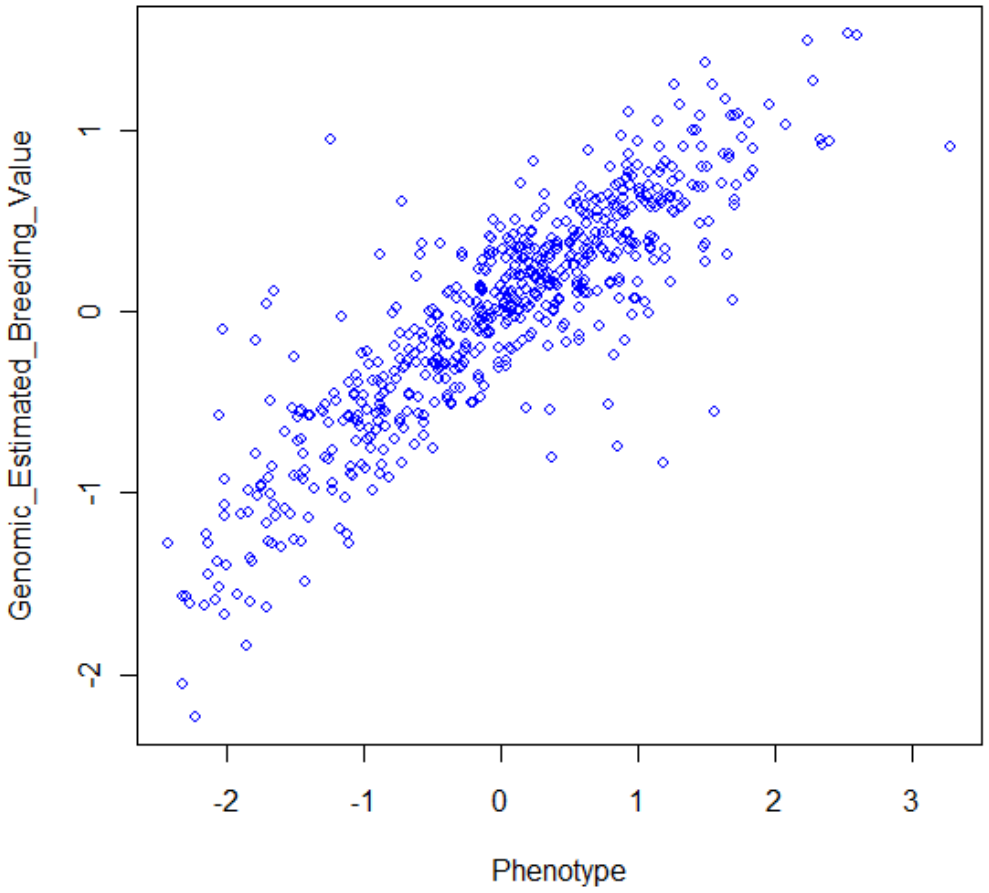
Populations:

- SolCAP population
- Ten year breeding evaluation
  - Ohio and others
- Nested RIL

Prediction of performance without evidence of statistically significant association.

- Single markers
- Haplotypes

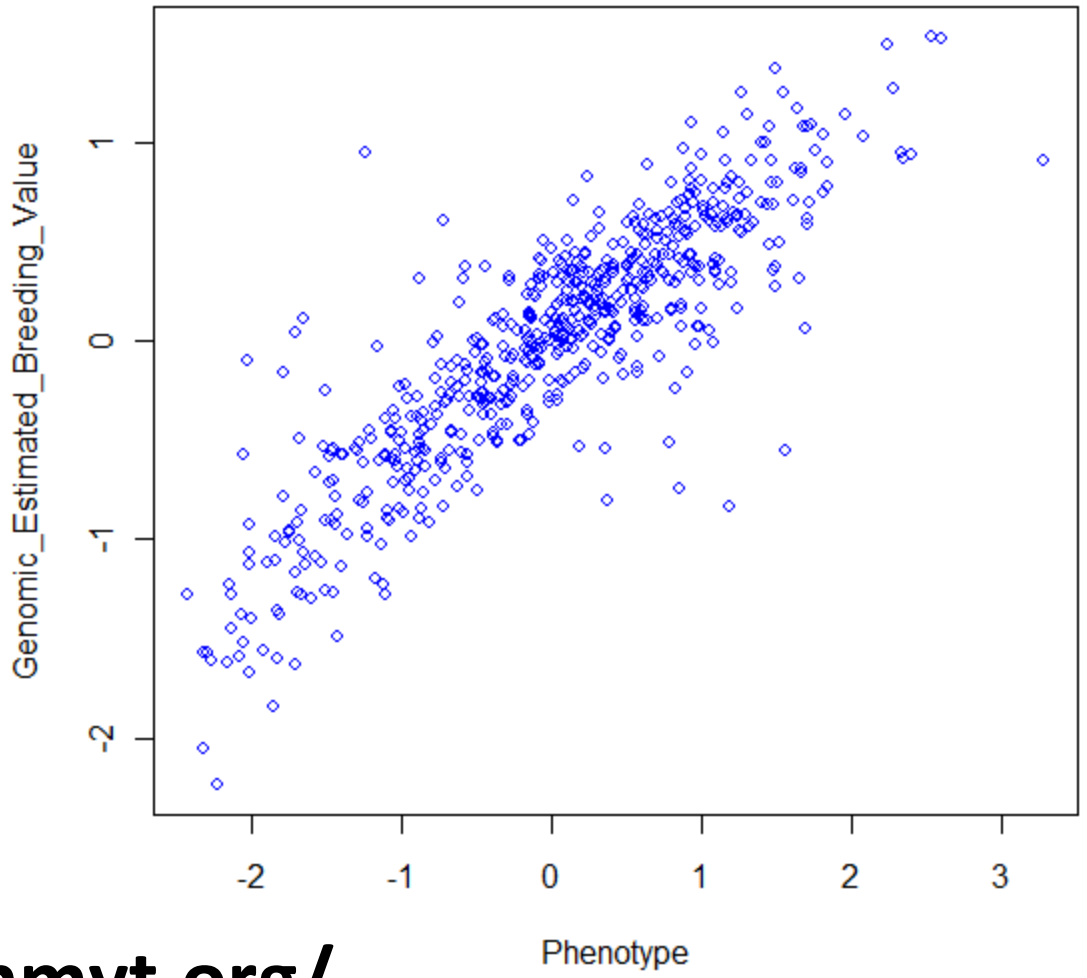
GWS – based on sum of breeding values estimated for all markers



Predicted Gen. Value relative to BLUP of Phenotype (BLR package)

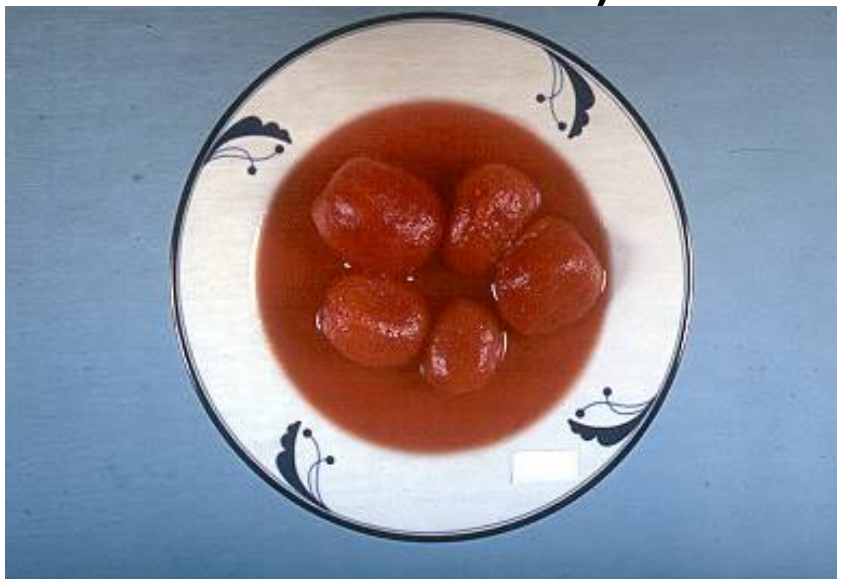
GEBV = $\sum_{i=1}^{n} X_i g_i$

Marker	Value
0	-0.32
1	1.23
0	0.40
1	-0.75
1	0.86
Sum	1.42



http://genomics.cimmyt.org/

How do we model multiple traits? (yield, quality and disease resistance)



Can we place all traits on the same scale e.g. economic merit?

What happens when the market does not value traits?

Summary:

- SolCAP Infinium SNP Array is providing a powerful tool for mapping and population level analysis
- Multiple lines of evidence suggest that there are distinct sub-populations within cultivated tomato; breeding history is reflected in population sub-structure.
- In cultivated tomato, LD decays over 2-10 cM. Recombination is limiting.
- Plant breeders select for combinations of genes.
- We can detect regions of the genome that are under selection during breeding. Insufficient recombination to identify candidate genes
- Germplasm from different regional breeding programs contains significant variation for traits of value.
- SolCAP SNPs permit mapping and association analysis within elite x elite populations (strong associations on chromosome 2 and 5).
- Limitation to GWS is establishing appropriate trait models.

Acknowledgments

Collaborators, OSU

Heather Merk
Sung-Chur Sim
Matt Robbins
Troy Aldrich

Collaborators, MSU

David Douches
C Robin Buell
John Hamilton
Dan Zarka
Kelly Zarka

Collaborators, Cornell

Walter de Jong
Lucas Mueller
Joyce van Eck
Naama Menda

Collaborators, UCD

Allen Van Deynze
Kevin Stoffel

Collaborators, INRA

Mathilde Causse

Industry Collaborators

Cindy Lawley, Illumina
Martin Ganal, Trait Genetics

Collaborators, CAU

Wencai Yang
Hui Wang

Collaborators, UIB

Hipolito Medrano
Pep Cifre
Josefina Bota
Miquel Angel Conesa

Funding

USDA/AFRI

This project is supported by the Agriculture and Food Research Initiative of USDA's National Institute of Food and Agriculture.