From Genomes to Phenomes to Breeding

Lukas Mueller
Boyce Thompson Institute
http://solgenomics.net/
Tomato Genome

- BAC by BAC approach 2004-2009 (~1500 BACs)
- Whole genome shotgun of entire tomato genome, started in 2009
- Technologies used: 454, Solexa (Syngenta), SOLiD
- 454 data assembled using Newbler
- Homopolymer correction using Solexa data
- Integration of BAC sequences
- Ordering and orientation of scaffolds based on
Why Gh13, a begomivirus-resistant inbred?

- Begomoviruses are a major threat in subtropical and tropical countries
- Gh13 highly resistant to monopartite and bipartite begomoviruses
- Gh13 used in association studies of molecular marker with resistance
  - F3 family experiments
  - RIL population available
- SolCAP SNP analysis available
- NSF requires that seed be available for distribution
Origin OF Gh13
TYLCV virus resistance: HUJI
(Vidavski and Czosnek, 1994)

L. hirsutum

6 years

Ih902 x S line, FAVI 9

S. Habrochaites
LA1777 & LA0386

Hybrids sent to Guatemala
(1998)
Gh13 inbred: known introgressions

**Ty3** ï chromosome 6, introgressed from wild species (*S. chilense?*)

**I2** ï chromosome 11 (*S. pimpinellifolium*)

Other introgressions from *S. habrochaites.*
+ Other species?
Gh13 inbred: known introgressions

**Ty3** ⅈ chromosome 6, introgressed from wild species (*S. chilense*?)

**I2** ⅈ chromosome 11 (*S. pimpinellifolium*)

Other introgressions from *S. habrochaites*. + Other species?

Disease resistance alleles can be found in wild species

Need to find introgression **regions**

And define introgression **contents**

F1, F2, BC . . .
Gh13 inbred: Whole genome sequencing

- Illumina HiSeq 2000
- One lane paired-ends = 20X tomato genome coverage
- Cost in 2012 : 2,400$
Gh13 inbred: Whole genome sequencing

- Illumina HiSeq 2000
- One lane paired-ends = 20X tomato genome coverage
- Cost in 2012 : 2,400$

**Output:** high number of reads
Relatively simple to align to a reference genome.

**Challenges:**
- Low coverage regions
- Regions different from Heinz1706
Gh13 inbred: Whole genome sequencing

Assembly: Heinz1706 is the reference genome

- Alignment (BWA)
  - SNP calling (Samtools)
    - Plot SNP density
  - Plot gaps
- De-novo assembly (SOAP)
  - Align scaffolds (MUMMER)
  - Close gaps
- Select regions For PCR
Heinz1706 is the reference genome assembly. Gh13 SNPs and Illumina reads aligned to Heinz are shown in the diagram. De-novo assembly is also indicated.
Gh13 inbred: SNP distribution

Hypothesis: SNPs are denser in introgression regions.

Chromosome 1

Chromosome 6
Gh13 inbred: SNP distribution

Hypothesis: SNPs are denser in introgression regions.

Chromosome 1

Chromosome 6

~50kb, Chr 1

~50kb, Chr 6
SNP distribution: Gh13, *S. pimpinellifolium*
SNP distribution: Gh13, *S. pimpinellifolium*

*S. pimpinellifolium* introgressions?
SNP distribution: Gh13, S. pimpinellifolium

Chromosome 6

30.6Mb - 34Mb

Verlaan et al, 2011
SNP distribution: Gh13, S. pimpinellifolium

Chromosome 6

30.6Mb - 34Mb

Verlaan et al, 2011
PCR design: Gh13 chr. 6 and 11

Hypothesis:

1. SNP non-peak regions are closest to Heinz1706
2. SNP Peak regions come from wild introgressions
PCR design: Gh13 chr. 6

1. SNP non-peak regions are closest to Heinz1706

Gh13 closest to Heinz, Yellow Pear, S. pimpinellifolium

Similar results for all other non-peak markers!
(trees built with MEGA, maximum likelihood 500 bootstrap replicates)
PCR design: Gh13 chr. 6

2. SNP Peak regions come from wild introgressions

Gh13 closest to *S. chilense*. Purple Russian, Yellow Pear, Heinz, *S. galapagense* cluster together.

Similar results for all other SNP-peak markers!
SNP distribution: Gh13, *S. pimpinellifolium*

*S. pimpinellifolium* introgressions?
PCR design: Gh13 chr. 11 - SNPs shared with *S. pimpinellifolium*

Peak regions: Gh13 = *S. pimpinellifolium* (different from other assayed wild species)

Non-peak regions: Gh13 = *S. lycopersicum* (different from other assayed wild species)
Gh13 wild introgressions – more to explore

Another wild species?

S. pimpinellifolium?
We eat phenotypes.
Phenotypes

Phenotyping is hard

- Labor intensive, expensive
- Standardization of phenotypic measurements
- Ontology-based systems for databases
Ontologies

Many ontologies are currently developed:

  - [http://www.bioversityinternational.org/](http://www.bioversityinternational.org/)
- [http://plantontology.org](http://plantontology.org) & PATO

- Difficult to apply one ontology to all plants!
Crop Ontology Curation Tool

General Germplasm Ontology
- **FAQ/IPGRI Multi-Crop Passport Descriptor**
  - 88 terms
  - BIOVERSITY
  - FAO/IPGRI Multi-Crop Passport Descriptor
- **Germplasm**
  - 386 terms
  - SHRESTHA
  - germplasm
- **ICIS germplasm method**
  - 166 terms
  - SHRESTHA
  - ICIS germplasm methods
- **Identifier Test Ontology**
  - 58 terms
  - SHRESTHA
  - Test ontology for identifier functionality

Phenotype and Trait Ontology
- **Cassava**
  - 128 terms
  - BAKARE
  - Cassava Trait Ontology
- **Chickpea**
  - 253 terms
  - PRASAD
  - Chickpea Traits
- **Common Bean**
  - 437 terms
  - AGUIERRE

Location and Environmental Ontology
- **Country and Location**
  - 1118 terms
  - SHRESTHA
  - Describes official ISO 3166-1 alpha-2, alpha-3 and numeric country codes along with location names.
- **Crop Research**
  - 256 terms
  - SHRESTHA
  - Describes experimental design, environmental conditions and methods associated with the crop study/experiment/trial and their evaluation.

Plant Anatomy & Development Ontology
- **Musa Anatomy**
  - 149 terms
  - CHANNELIERE
  - Musa Anatomy
Accession: 313-100

Stock details

Organism: *Solanum lycopersicum*
Stock type: accession
Stock name: 313-100
Uniquename: 313-100
Description

Stock editors: Esther van der Knaap

Synonyms

Pedigree data

Additional information

Associated loci (0)

Experimental data

Related stocks

Accessions this accession is a member of

<table>
<thead>
<tr>
<th>Type</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>f2 population</td>
<td>QTL Tomato Sausage x LA1589 F2</td>
</tr>
</tbody>
</table>

Images (1)

Literature annotation (0)

Ontology annotation ()

Phenotype data

Experiment: phenotypes recorded for population QTL Tomato Sausage x LA1589 F2 by Esther van der Knaap
### Phenotype data

**Experiment:** phenotypes recorded for population QTL Tomato Sausage x LA1589 F2 by Esther van der Knaap

<table>
<thead>
<tr>
<th>Trait</th>
<th>Average</th>
<th>Min</th>
<th>Max</th>
<th>Lines/repeats</th>
</tr>
</thead>
<tbody>
<tr>
<td>distal angle macro 10%</td>
<td>151.30</td>
<td>151.30</td>
<td>151.30</td>
<td>1</td>
</tr>
<tr>
<td>distal angle macro 15%</td>
<td>131.60</td>
<td>131.60</td>
<td>131.60</td>
<td>1</td>
</tr>
<tr>
<td>distal angle macro 20%</td>
<td>108.55</td>
<td>108.55</td>
<td>108.55</td>
<td>1</td>
</tr>
<tr>
<td>distal angle micro 2%</td>
<td>173.07</td>
<td>173.07</td>
<td>173.07</td>
<td>1</td>
</tr>
<tr>
<td>distal angle micro 3%</td>
<td>168.79</td>
<td>168.79</td>
<td>168.79</td>
<td>1</td>
</tr>
<tr>
<td>distal angle micro 5%</td>
<td>167.66</td>
<td>167.66</td>
<td>167.66</td>
<td>1</td>
</tr>
<tr>
<td>distal eccentricity index</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>1</td>
</tr>
<tr>
<td>distal fruit end blockiness 10%</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>1</td>
</tr>
<tr>
<td>distal fruit end blockiness 20%</td>
<td>0.79</td>
<td>0.79</td>
<td>0.79</td>
<td>1</td>
</tr>
<tr>
<td>distal fruit end blockiness 30%</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>1</td>
</tr>
<tr>
<td>distal fruit end blockiness 5%</td>
<td>0.44</td>
<td>0.44</td>
<td>0.44</td>
<td>1</td>
</tr>
<tr>
<td>distal fruit end indentation</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1</td>
</tr>
<tr>
<td>distal fruit end protrusion</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1</td>
</tr>
<tr>
<td>eccentricity area index</td>
<td>0.09</td>
<td>0.09</td>
<td>0.09</td>
<td>1</td>
</tr>
<tr>
<td>fruit area</td>
<td>35566.63</td>
<td>35566.63</td>
<td>35566.63</td>
<td>1</td>
</tr>
<tr>
<td>fruit length mid-width</td>
<td>202.00</td>
<td>202.00</td>
<td>202.00</td>
<td>1</td>
</tr>
<tr>
<td>fruit longest length</td>
<td>204.25</td>
<td>204.25</td>
<td>204.25</td>
<td>1</td>
</tr>
<tr>
<td>fruit mid-height width</td>
<td>215.50</td>
<td>215.50</td>
<td>215.50</td>
<td>1</td>
</tr>
<tr>
<td>fruit perimeter</td>
<td>708.60</td>
<td>708.60</td>
<td>708.60</td>
<td>1</td>
</tr>
<tr>
<td>fruit shape circular</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>1</td>
</tr>
<tr>
<td>fruit shape eccentric</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
<td>1</td>
</tr>
<tr>
<td>fruit shape ellipsoid</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>1</td>
</tr>
<tr>
<td>fruit shape index external</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>1</td>
</tr>
<tr>
<td>fruit shape index external 2</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>1</td>
</tr>
</tbody>
</table>
Trait scoring

- Use barcode tools
Tomato panel (~400 accessions)
Incl. Processing, fresh market, heirloom, wild relatives

Phenotyping for breeder traits:
- Tomato Analyzer
- Fruit shape
- Color
- PH
- Brix
- Vitamin C
- Lycopene
- sugars

Potato panel (~400 accessions)

Phenotyping:
- Specific gravity
- chip color after cold storage
- sucrose/glucose
- Skin texture
- tuber shape(l/w/h)
- eyedepth
- skincolor
- Flower color
- Flesh color
- growth habit
- total yield etc.

Genotyping (Illumina Infinium chip)

Genotyping (Illumina Infinium chip)
Genotyping by Sequencing (GBS)

- Developed by Buckler lab (Elshire, 2011)
- Full genome sequencing too expensive
- Reduce sequence space using restriction
- Use highly multiplexed NGS approach
Genotyping by Sequencing (GBS)

- Focuses NextGen sequencing power to ends of restriction fragments
- Scores both SNPs and presence/absence markers
Genotyping by Sequencing (GBS)

![Graph showing costs for differentplexing numbers (48-plex, 96-plex, 384-plex) with costs labeled as $33.00, $19.00, and $9.00 respectively. The graph indicates costs for sequencing, labor, and reagents & consumables.]
Storing genotypic data

- Challenge: Extremely voluminous
- 50,000 plants 20,000 markers = 1,000,000,000 datapoints
- Special techniques are needed to store data
  - Relational databases: Compress genotype data into strings
  - Non-relational databases: HDF5
Chado Natural Diversity Schema
Breeding technologies

- Phenotypic Selection
- Marker-Assisted Selection (MAS)
- Genomic Selection
Genomic Selection

- Remove phenotyping from line development

- Use markers to model genetic relatedness between lines.
  - Use relatedness estimates to make predictions

- Use markers as predictors in regression-type models
  - Use estimated marker effects to make selections
Genomic Selection

- Model Training Cycle
  - Phenotype (lines already genotyped)
  - Train prediction model
  - Advance lines informative for model improvement

- Updated Model
  - Genomic Selection

- Line Development Cycle
  - Genotype
  - New Germplasm
  - Make crosses and advance generations
  - Advance lines with highest GEBV
  - Test varieties and release
Integrate Breeding functions

- Store genotypes and phenotypes in the database
  - Calculation of GS models
  - Prediction of phenotypes
- Manage breeding process:
  - Crosses
  - Pedigree tracking
  - Field planting
  - Sample collection
  - Data collection
Breeder Tools

**Trials**

**Locations**

- unknown (0 plots)
- OSU-OARDC Fremont, OH (19739 plots)
- Tidewater. Plymouth, NC (0 plots)
- UofI R&E Center, Aberdeen, Idaho (404 plots)
- Campbell’s Soup Company (10930 plots)
- Mills River, North Carolina (3041 plots)
- Hutchinson Drive, Davis CA (8921 plots)
- University of Florida/ IFAS Gulf Coast Research & Education Center (5777 plots)

**Add new location**

**Crosses**

- Add new cross
- Upload cross file
- View all crosses

**Phenotypes**

- Upload
- Phenotype search

**Accessions & plots**

List of accessions:

Something wrong? Report a problem

SGN is supported by the NSF (#0115076), USDA CSREES and hosted at the Boyce Thompson Institute.
Conclusions

- Genome databases need to adapt to the needs of breeders
- Genomic technologies applicable to improvement of the breeding process
  - Genotyping by Sequencing
  - Genomic Selection
- Bioinformatics infrastructure required
  - Genome, phenome, & genotypic information, algorithms, breeder functions