Genomic Estimated Breeding Values and Genomic Predictions

• Motivation for Genomic Estimated Breeding Values
  – Example: The Problems with Male Selection

• Systems
  – Genotyping and Genetics System Parameters
  – Phenotyping and Trait Analyses
  – Genome-wide Association and Mixed Linear Modeling
  – Prediction Accuracy with Machine Learning (Artificial Intelligence)

• Results Examples:
  1. Sex determination genes
  2. Powdery mildew resistance genes
Time-line of variety development … ….too slow

<table>
<thead>
<tr>
<th>Year 0 - 1</th>
<th>Year 2 – 3</th>
<th>Year 4 – 6</th>
<th>Year 7 – 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection of parents</td>
<td>Single Hill Evaluation</td>
<td>Multi Hill Evaluation</td>
<td>Semi Commercial</td>
</tr>
<tr>
<td>Seedling Screening</td>
<td>Disease resistance</td>
<td>Agronomic traits</td>
<td>Confirmation on agronomic and chemical traits</td>
</tr>
<tr>
<td>Diseases</td>
<td>Chemical traits</td>
<td>Chemical traits</td>
<td>Extensive brewing tests</td>
</tr>
<tr>
<td>Marker screening</td>
<td>Maturity</td>
<td>Different environments</td>
<td></td>
</tr>
<tr>
<td>50 crosses, ≈25,000 seedlings</td>
<td>≈7,000 plants</td>
<td>100 plants, 20 plants</td>
<td>2 varieties</td>
</tr>
</tbody>
</table>
Breeding systems:
- Single Cross
- Select
- Idiotypes

Quantitative Genetics
- Estimated Breeding Values (EBV)

Molecular Quantitative Genetics
- GEBV

Calculate EBVs for Male Alpha

Alpha Acid h2 = 0.5
The male selection problem in hops

Quantitative Traits

- Probably caused by multiple loci
  - Interaction effects
  - Environment

If the mean trait value for individuals with marker state MM is different from the mean trait value of individuals with marker state mm (i.e., the marker is associated with the phenotype), then the marker is linked to a quantitative trait locus.
General and Mixed Linear Modeling of Associations Between Traits Variations and Genes Variations (alleles, markers)

Accounting for Random Effects: Mixed Linear Models

- "Cost" associated with estimating a parameter
- We are not interested in the value of the parameter, only the variance
- Q-K method (structured association)

\[ y = X\beta + S\alpha + Qv + Zu + e \]

**Fixed effects:**
- \( \beta \) Vector of fixed effects
- \( \alpha \) Vector of SNPs effects
- \( v \) Vector of subpopulation effects

**Random effects:**
- \( u \) Vector of kinship effects
- \( e \) Residuals

**Notation:**
- \( Q \) Matrix of population association (STRUCTURE)
- \( X, S, Z \) Incidence Matrices
Technologies for Breeding

Apollo
Whole Genome Sequence
~44,000 contigs

Molecular Selection Tools
Specific system components

Genomes
- 5
  - Teamaker 1.8 Gb
  - ShinsuWase 2.05 Gb
  - Apollo 2.28 Gb

SNPs
- 1,235,148

5,572,988,000 SNP calls

Germplasm
- 4396 cultivars
- 116 wilds
- 22 families

Transcriptomes
- 36 cultivars
- 5 tissues
- 39,000 genes

Traits
- (70,128)
  - Targeted Chemical (113)
  - Untargeted Chemical (70,000)
  - Morphological (15)
  - Disease (7)

Non-Mendelian inheritance of SNP markers reveals extensive chromosomal translocations in dioecious hops. bioRxiv

“5000 Genome Project”
3.9 x 10^{13} Gene:Trait Associations
Genomic Estimated Breeding Values and Genomic Predictions

- Motivation for Genomic Estimated Breeding Values
  - Example: The Problems with Male Selection

- Systems
  - Genotyping and Genetics System Parameters
  - Phenotyping and Trait Analyses
  - Genome-wide Association and Mixed Linear Modeling
  - Prediction Accuracy with Machine Learning (Artificial Intelligence)

- Results Examples:
  1. Sex determination genes
  2. Powdery mildew resistance genes
Phylogeny of domestic and wild hops
Genetic Distance Principle Component Analysis

Dong Zhang, Nicholi J. Pitra, Mark C. Coles, Edward S. Buckler, Paul D. Matthews

Linkage Groups in Families

a

b

C

247 LGs

144 LGs

265 LGs

144 LGs
Linkage groups in correlation projections

SD markers convergence
Linkage groups “sharing”

Dong Zhang, Nicholi J. Pitra, Mark C. Coles, Edward S. Buckler, Paul D. Matthews
Genomic Estimated Breeding Values and Genomic Predictions

• Motivation for Genomic Estimated Breeding Values
  – Example: The Problems with Male Selection

• Systems
  – Genotyping and Genetics System Parameters
  – Phenotyping and Trait Analyses
    – Genome-wide Association and Mixed Linear Modeling
    – Prediction Accuracy with Machine Learning (Artificial Intelligence)

• Results Examples:
  1. Sex determination genes
  2. Powdery mildew resistance genes
Loading the MLM

1. Making a Reference Genome
   - DNA Sequencing
   - Apollo
   - Raw Reads
   - Clean and Trimmed Reads
   - Contigs
   - Genome Assembly (reference contigs)

2. Generating GBS Markers
   - Extract DNA
   - Cut and Sequence Genomic DNA

3. Genome Based Selection
   - Mapping Reads to Reference
   - Single Nucleotide Polymorphism Detection
   - SNP Matrix
   - Matrix of Relatedness
   - Phenotype
   - Mixed Line Modeling
   - $y = X\beta + Zu + e$
   - $\kappa^2 = \frac{var(A)}{var(P)}$
   - Apply the Marker

Apply the Marker

Analyze Results and Advance Breeding Program

R = resistant
S = susceptible

Plant 1
- Susceptible
- Plant 1 yes
- Plant 1 no

Plant 2
- Resistant
- Plant 2 yes
- Plant 2 no

Plant 3
- Susceptible
- Plant 3 yes
- Plant 3 no

Specific Cut Sites
Genomic Estimated Breeding Values and Genomic Predictions

• Motivation for Genomic Estimated Breeding Values
  – Example: The Problems with Male Selection

• Systems
  – Genotyping and Genetics System Parameters
  – Phenotyping and Trait Analyses
  – Genome-wide Association and Mixed Linear Modeling
    – Prediction Accuracy with Machine Learning (Artificial Intelligence)

• Results Examples:
  1. Sex determination genes
  2. Powdery mildew resistance genes
MLM and rrBLUP (GEBV) for sex$\text{Sex}$

REFERENCES

Matthews PD, Coles MC, Pitra NJ, Next Generation Sequencing for a Plant of Great Tradition: Application of NGS to SNP Detection and Validation in Hops (Humulus lupulus L.), 2013, Monatsschrift für Brauwissenschaft, 66:8


Candidate Genes for Male Flower Development Found

1. Glucose-regulated protein 94 (GRP94)-like protein on scaffold LD152823 that is known in Arabidopsis affecting shoot apical meristems, floral meristems and pollen tube elongation

2. Squamosa-like protein, identified on scaffold LD147778, has essential roles in vegetative phase change and flower development in multiple plants
80% of breeding effort: 
Breeding for disease resistance

- Plant microbial diseases
  - Powder mildew (*Podosphaera macularis*)
  - Downy mildew (*Pseudoperonospora humuli*)
  - Viruses and viroids (stunt viroid)
Powdery Mildew GEBVs

<table>
<thead>
<tr>
<th>Gene</th>
<th>Source</th>
<th>Status in USA</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1, R3, Rb</td>
<td>Zenith</td>
<td>Tolerance</td>
</tr>
<tr>
<td>R2</td>
<td>Wye Target</td>
<td>Resistance</td>
</tr>
<tr>
<td>R4</td>
<td>Serebrianka</td>
<td>Tolerance</td>
</tr>
<tr>
<td>R5</td>
<td>Early Choice</td>
<td>Tolerance</td>
</tr>
<tr>
<td>R6</td>
<td>Nugget</td>
<td>Broken</td>
</tr>
<tr>
<td>19058mR6</td>
<td></td>
<td>Broken</td>
</tr>
</tbody>
</table>

Might stacked resistance genes confer durable resistance/tolerance? 

A set of 1224 half-siblings were used a nested association panel across six known powdery mildew resistance genes.
Nested Powdery Mildew Mapping Families

Digital metrics of powdery mildew disease index lesion area/leaf area used as quantitative trait.

- Qualitative field scores were corrected for infection hotspots.
- Among six nested families, all showed expected Mendelian segregation ratios.
Powdery Mildew Resistance Genes, R1, R2, R6 mapped

GLM for R1 (F1: 106 maternal LGs)

GLM for R2 (F1: 247 maternal LGs)

GLM for R6 (F1: 144 maternal LGs)
Powdery Mildew Genes on same Linkage Group
Prediction Accuracies

Ensemble (voting) method, combing different classifier, might be useful!
Conclusions:

- GEBVs could be a new evaluation tool for hop breeding potential, especially in male selection.
- An efficient, high density molecular marker system has been qualified and validated for hops.
- Segregation distortion in hops is extensive and real.
- GEBVs have been developed for sex and powdery mildew tolerance.
- Candidate genes have been identified for sex and PMT.
## Advance and sustainable bitter acids and yield

<table>
<thead>
<tr>
<th>Variety</th>
<th>Galena</th>
<th>Super Galena</th>
<th>07270</th>
<th>Zeus</th>
<th>Apollo</th>
<th>Bravo</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Alpha acids % w/w</strong></td>
<td>10.0 - 13.5</td>
<td>13.0 - 16.0</td>
<td>18-20</td>
<td>12.0 - 16.5</td>
<td>15.0 - 19.0</td>
<td>14.0 - 17.0</td>
</tr>
<tr>
<td><strong>Beta acids % w/w</strong></td>
<td>7.0 - 9.0</td>
<td>8.0 - 10.0</td>
<td>4.5-6.0</td>
<td>4.0 - 6.0</td>
<td>5.5 - 8.0</td>
<td>3.0 - 5.0</td>
</tr>
<tr>
<td><strong>CoH % w/w of α-acids</strong></td>
<td>35 - 40</td>
<td>35 - 40</td>
<td>27-29</td>
<td>27 - 35</td>
<td>24 - 28</td>
<td>29 - 34</td>
</tr>
<tr>
<td><strong>Total Oil ml/100g</strong></td>
<td>0.9 - 1.2</td>
<td>1.5 - 2.5</td>
<td>3.0</td>
<td>1.0 - 2.0</td>
<td>1.5 - 2.5</td>
<td>1.6 - 2.4</td>
</tr>
<tr>
<td><strong>Stability</strong></td>
<td>75 - 80%</td>
<td>75 - 80%</td>
<td>85%</td>
<td>50 - 60%</td>
<td>80 - 90%</td>
<td>60 - 70%</td>
</tr>
<tr>
<td><strong>Powdery Mildew</strong></td>
<td>Susceptible</td>
<td>Resistant</td>
<td>Resistant</td>
<td>Susceptible</td>
<td>Resistant</td>
<td>Resistant</td>
</tr>
<tr>
<td><strong>Yield lbs/acre</strong></td>
<td>1,600 - 2,220</td>
<td>2,500 - 2,800</td>
<td>2,500-3,400</td>
<td>2,400 - 3,000</td>
<td>2,600 - 3,000</td>
<td>2,700 - 3,100</td>
</tr>
</tbody>
</table>
### Advanced sustainable aroma and flavor

<table>
<thead>
<tr>
<th>Variety</th>
<th>Casc.</th>
<th>Calypso</th>
<th>Cent.</th>
<th>Lemondrop</th>
<th>EUREKA!</th>
<th>Denali</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Alpha acids % w/w</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.5-7.0</td>
<td>13.0 - 16.0</td>
<td>9.5-11.5</td>
<td>4.5-6.5</td>
<td>15-18</td>
<td>15-18</td>
<td></td>
</tr>
<tr>
<td><strong>Beta acids % w/w</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.5-7.0</td>
<td>8.0 - 10.0</td>
<td>3.5-4.5</td>
<td>4.0-6.0</td>
<td>5.0-6.0</td>
<td>4.0-5.0</td>
<td></td>
</tr>
<tr>
<td><strong>CoH % w/w of α-acids</strong></td>
<td>33-40</td>
<td>35 - 40</td>
<td>29-30</td>
<td>30-33</td>
<td>45-50</td>
<td>22-26</td>
</tr>
<tr>
<td><strong>Total Oil ml/100g</strong></td>
<td>0.8-1.5</td>
<td>1.5 - 2.5</td>
<td>1.5-2.3</td>
<td>1.5-2.0</td>
<td>1.5-2.0</td>
<td>4.0</td>
</tr>
<tr>
<td><strong>Stability</strong></td>
<td>48%</td>
<td>75 - 80%</td>
<td>45-55%</td>
<td>65%</td>
<td>80%</td>
<td>80%</td>
</tr>
<tr>
<td><strong>Powdery Mildew</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tolerant</td>
<td>Susceptible</td>
<td>Tolerant</td>
<td>Tolerant</td>
<td>Resistant</td>
<td>Tolerant</td>
<td></td>
</tr>
<tr>
<td><strong>Yield lbs/acre</strong></td>
<td>1600-2200</td>
<td>2,500 - 2,800</td>
<td>1,700-2,000</td>
<td>2,000 – 2,800</td>
<td>2,400 – 2,800</td>
<td>2,600 - 3,200</td>
</tr>
</tbody>
</table>
Matthews’s Lab
- Mark Coles, chemoanalytics, DNA
- Tiffany Pitra, sensory evaluation, administration, pathology
- Nicholi Pitra, Genomics, computing
- Rachel Jones, greenhouse, tissue culture

Agronomy
- Roger Jeske, agronomist
- Ann Petro, propagation, field collections
- Danny Hallman, Paul Merritt, Tom Newhouse, growers

Past Hopsteiner-funded postdocs and graduate students:
- Dr. Lina Maloukh ILVO – ILVO, Belgium
- Dr. Adam Kavalier – CUNY
- Dr. Shi-Biao Wu- CUNY
- Nicholi Pitra – UNIowa
- Jared Koellling – UNIowa
- Jana Naegel – NRC – Canada
- Dr. Shaun Clark – NRC – Canada
- Alex Feiner – Martin Luther U, IPB-Germany

Collaborating PIs:
- Dr. Edward Buckler 4th, IGD, Cornell
- Dr. Arne Heyerrick – UGhent – Belgium
- Dr. Edward Kennelly – CUNY
- Dr. Dwight Kincaid - CUNY
- Dr. Jonathan Page – NRC- Canada
- Dr. Fred Stevens, L. Pauling Institute
- Dr. Axel Schwekendiek, UNIowa
- Dr. Ryan Weil, Emory U.
- Dr. Ludger Wessjohann – IPB – Germany
- Dr. Oliver Yu, Danforth Center
- Dr. John Henning, ARS-USDA
And Illumina

Crop Improvement Program

Current Hopsteiner Fellows:
- Taylan Morcol, CUNY, Kennelly Lab
- Dr. Dong Zhang, Cornell, Buckler Lab
- Jenna Kahn, Harvey Mudd College
- Aurélie Muntzel, Binghamton, SUNY

Contact: pmatthews@hopsteiner.com
Thank you very much.

Simon H. Steiner, Hopfen GmbH

Steiner Hops Limited

S.S. Steiner, Inc.

Steiner Asia Limited
Next Generation Sequencing provides Massive Molecular Marker Data

3-5 million reads of 64 nucleotides per plant

Genomic DNA is sheared into fragments and size selected, then separated into single strands.

Universal adaptor sequences are ligated to the target pool of DNA.

To select for particular areas of the genome, DNA is captured by complementary fragments of DNA or RNA on fixed arrays (shown) or on beads in solution.

The DNA fragments are washed over an array or incubated with microscopic beads such that one DNA molecule is anchored by its adaptor on a single bead or away from other fragments on an array.

DNA is amplified with the net result that clusters of cloned fragments are fixed in distinct areas of the array or on separate beads.
Hopsteiner Breeding Program

Paul Matthews, Ph.D.
Time-line of variety development … ….too slow

Year 0 - 1
- Selection of parents
- Seedling Screening
- Diseases
- Marker screening

Year 2 – 3
- Single Hill Evaluation
  - Disease resistance
  - Chemical traits
  - Maturity

Year 4 – 6
- Multi Hill Evaluation
  - Agronomic traits
  - Chemical traits
  - Different environments
  - 1st brewing trials

Year 7 – 10
- Semi Commercial
  - Confirmation on agronomic and chemical traits
  - Extensive brewing tests

- 50 crosses, ≈25,000 seedlings
- ≈7,000 plants
- 2 varieties
- 100 plants, 20 plants
Motivation for molecular marker development

- What are markers good for?
  - Trueness-to-type determination
  - Parent determination (whose the daddy?)
  - Variety rights protection
  - Marker-assisted selection
Molecular Marker Development

- 600 Diversity Array Technology markers (2010)
  - DNA hybridization micro array (give citation)

- 1000 Genic Simple Sequence Repeat markers (2011)
  - Transcriptome mining (give citation)

- 300,000 Single Nucleotide Polymorphisms (2013)
  - Genotyping-by-sequencing (give example >)
GBS Revolution:
• 2-5 million makers per analysis & plant
• >1000 analyses completed
• 3 billion markers scored
• 64 bp gene tags around each marker

Mark Coles
Nicholi Pitra
GBS markers behave well in hierarchical cluster analysis of commonly known genotypes.

For all marker subsets:

1. Mantel’s tests show distance matrices are highly similar ($331 < Z > 356$, $p < 0.0001$, 5,000 permutations).
2. Cophenetic correlations coefficients are very high (>0.95).
3. Bootstrapping (10,000 subsets) gave very high accuracy.

Next Generation Sequencing for a Plant of Great Tradition: Application of NGS to SNP Detection and Validation in Hops (*Humulus lupulus* L.)

Application of next generation DNA sequencing technology to hops yielded an unprecedented, large number of novel single nucleotide polymorphisms (17, 128 SNPs). The markers were detected and then validated for use in genotyping and control of quality for hops. By using genotyping-by-sequencing (GBS) and a universal network-enabled analysis kit (UNEAK) designed for species with no “reference genome”, we generated a set of molecular markers with a genome-wide distribution. Validation of the markers was accomplished by observation of metrics of sequencing quality, by marker behavior in genetic segregation and by application to genetic distance and hierarchical cluster analyses across a set of commonly known cultivars. The SNPs were characterized by average read depth of 3.7 and a call rate across 178 diverse individuals of 0.82. Many SNP alleles segregated with near test cross ratios of 1 : 3 or 3 : 1 and intercross ratios of 0.50 among 103 full-siblings. Erroneous SNPs, with unusually high or low allele segregation ratios were detected at a rate of 4.1% and could be removed from further analyses. Filtering of SNPs for potentially higher quality was accomplished by selection of call rate thresholds above 0.5, 0.75 and 0.90 or, alternatively, by selection of markers with minimal segregation distortion. Genetic distance matrices and dendrograms for marker subgroups were similar as shown by Mantel’s Z-tests and cophenetic correlation coefficients. Bootstrapping generated an exceptionally well-supported tree for genetic relationships among the hop cultivars.

A set of 1224 half-siblings were used a nested association panel across six known powdery mildew resistance genes. Might stacked resistance genes confer durable resistance/tolerance?

<table>
<thead>
<tr>
<th>Gene</th>
<th>Source</th>
<th>Status in USA</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1, R3, Rb</td>
<td>Zenith</td>
<td>Tolerance</td>
</tr>
<tr>
<td>R2</td>
<td>Wye Target</td>
<td>Resistance</td>
</tr>
<tr>
<td>R4</td>
<td>Early Choice</td>
<td>Tolerance</td>
</tr>
<tr>
<td>R5</td>
<td>Cascade</td>
<td>Tolerance</td>
</tr>
<tr>
<td>R6</td>
<td>Nugget</td>
<td>Broken</td>
</tr>
<tr>
<td>19058mR6</td>
<td></td>
<td>Broken</td>
</tr>
<tr>
<td>Kazak 2000R</td>
<td>Kazak 2000</td>
<td>Resistant, HSR</td>
</tr>
</tbody>
</table>
Q-Q plot of MLM associations

Expected $-\log(P\text{-Value})$ vs. $-\log(P\text{-Value})$
Comparison of linkage degree across F1 populations

144 - 2.1 = 265_7.1 + 265_7.2

Dong Zhang
Morphological Traits

1. Plant Biomass
2. Cone Biomass
3. Bine Biomass
4. Average cone weight
5. Average cone shape
6. Average cone length
7. Average cone width
8. Average number of bracts/tioles
9. Anther gland count
10. Leaf gland density
11. Bract gland density
12. Bractiole gland density
13. Yield
14. Cone set
15. Branch angle
80% of breeding effort: Breeding for disease resistance

- Plant microbial diseases
  - Powder mildew (*Podosphaera macularis*)
  - Downy mildew (*Pseudoperonospora humuli*)
  - Viruses and viroids (stunt viroid)