Network-Based Discovery: Seeing the Forest for the Trees

Adventures with 82,000 Phenotypes

Dan Jacobson
# Experimental Data Types

- **Natural Variation**
  - Genome Wide Association Studies
  - Millions of SNPs
  - ~82,000 Phenotypes
    - Morphology/Phenology
    - Molecular
  - 344 billion statistical tests to date
  - 4 trillion statistical tests later this year

- **Best BESC Lines** (*Populus* and *Switchgrass*)
  - Natural Variants
  - Transgenic Lines

- **All publically available Genomes**

- **Differential/Time Series Expression Studies**
  - Small number of variables
  - Hundreds of variables
  - 95,000 variables

- **Microbiomes & Metagenomes**
  - Natural Communities
  - Constructed Communities

- **Omens**
  - Genomics, Transcriptomics, Proteomics, Metabolomics, Microbiomes (and meta-omes of all types)

- **Systems Biology Approach**
  - Combining datasets across omics layers, sample sets, and species
<table>
<thead>
<tr>
<th>Gene</th>
<th>baseMean</th>
<th>log2FoldChange</th>
<th>lfcSE</th>
<th>stat</th>
<th>pvalue</th>
<th>padj</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pavir.Aa00004</td>
<td>23.03260874</td>
<td>-0.772176419</td>
<td>0.235718754</td>
<td>-3.275837864</td>
<td>0.00105349</td>
<td>0.036650136</td>
</tr>
<tr>
<td>Pavir.Aa00007</td>
<td>3.617339133</td>
<td>-3.277187207</td>
<td>0.925328577</td>
<td>-3.541647029</td>
<td>0.000397637</td>
<td>0.016344905</td>
</tr>
<tr>
<td>Pavir.Aa00018</td>
<td>2.194595376</td>
<td>-1.375554763</td>
<td>0.421360908</td>
<td>-3.264552399</td>
<td>0.001096372</td>
<td>0.037862673</td>
</tr>
<tr>
<td>Pavir.Aa00110</td>
<td>432.2298561</td>
<td>-0.920355344</td>
<td>0.087912301</td>
<td>-10.46901667</td>
<td>1.20E-25</td>
<td>1.26E-22</td>
</tr>
<tr>
<td>Pavir.Aa00136</td>
<td>1.76644122</td>
<td>-7.964343955</td>
<td>1.643802037</td>
<td>-4.84507488</td>
<td>1.27E-06</td>
<td>0.000109099</td>
</tr>
<tr>
<td>Pavir.Aa00161</td>
<td>63.5108945</td>
<td>1.524126268</td>
<td>0.377624869</td>
<td>4.03608552</td>
<td>5.44E-05</td>
<td>0.002965915</td>
</tr>
<tr>
<td>Pavir.Aa00164</td>
<td>86.61299946</td>
<td>1.970704034</td>
<td>0.235133135</td>
<td>8.381226395</td>
<td>5.24E-17</td>
<td>2.16E-14</td>
</tr>
<tr>
<td>Pavir.Aa00168</td>
<td>45.57577197</td>
<td>-2.776318341</td>
<td>0.3350917</td>
<td>-8.285249514</td>
<td>1.18E-16</td>
<td>4.66E-14</td>
</tr>
<tr>
<td>Pavir.Aa00180</td>
<td>7.78468493</td>
<td>1.72469973</td>
<td>0.269294597</td>
<td>6.405571227</td>
<td>1.50E-10</td>
<td>2.64E-08</td>
</tr>
<tr>
<td>Pavir.Aa00185</td>
<td>15.77390176</td>
<td>-3.03656463</td>
<td>0.739522148</td>
<td>-4.106117228</td>
<td>4.02E-05</td>
<td>0.00228249</td>
</tr>
<tr>
<td>Pavir.Aa00190</td>
<td>246.4158349</td>
<td>0.749398201</td>
<td>0.130879565</td>
<td>5.725861023</td>
<td>1.03E-08</td>
<td>1.35E-06</td>
</tr>
<tr>
<td>Pavir.Aa00201</td>
<td>194.2868719</td>
<td>0.55688662</td>
<td>0.146656817</td>
<td>3.797209232</td>
<td>0.000146334</td>
<td>0.007032352</td>
</tr>
<tr>
<td>Pavir.Aa00210</td>
<td>71.8661413</td>
<td>-0.945676165</td>
<td>0.223959112</td>
<td>-4.222539364</td>
<td>2.42E-05</td>
<td>0.001454015</td>
</tr>
<tr>
<td>Pavir.Aa00213</td>
<td>45.08826603</td>
<td>-2.821545181</td>
<td>0.381707372</td>
<td>-7.391906442</td>
<td>1.45E-13</td>
<td>3.90E-11</td>
</tr>
<tr>
<td>Pavir.Aa00219</td>
<td>82.09354863</td>
<td>2.652283666</td>
<td>0.48092843</td>
<td>5.514923839</td>
<td>3.49E-08</td>
<td>4.08E-06</td>
</tr>
<tr>
<td>Pavir.Aa00237</td>
<td>48.01170214</td>
<td>1.765138681</td>
<td>0.318940668</td>
<td>5.534379463</td>
<td>3.12E-08</td>
<td>3.70E-06</td>
</tr>
<tr>
<td>Pavir.Aa00238</td>
<td>4.900020424</td>
<td>-6.641133503</td>
<td>1.55203963</td>
<td>-4.278971603</td>
<td>1.88E-05</td>
<td>0.001166295</td>
</tr>
<tr>
<td>Pavir.Aa00240</td>
<td>3.536707907</td>
<td>-2.288869563</td>
<td>0.396004267</td>
<td>-5.77991361</td>
<td>7.47E-09</td>
<td>1.01E-06</td>
</tr>
<tr>
<td>Pavir.Aa00245</td>
<td>100.2653536</td>
<td>0.851939179</td>
<td>0.154407276</td>
<td>5.517480799</td>
<td>3.44E-08</td>
<td>4.03E-06</td>
</tr>
<tr>
<td>Pavir.Aa00246</td>
<td>74.76890191</td>
<td>0.900755926</td>
<td>0.267107154</td>
<td>3.372263199</td>
<td>0.000745529</td>
<td>0.027702451</td>
</tr>
<tr>
<td>Pavir.Aa00246</td>
<td>129.7507991</td>
<td>1.878568856</td>
<td>0.195429139</td>
<td>9.612532015</td>
<td>7.08E-22</td>
<td>5.19E-19</td>
</tr>
<tr>
<td>Pavir.Aa00246</td>
<td>0.855875118</td>
<td>-3.952874961</td>
<td>1.177355482</td>
<td>-3.357418402</td>
<td>0.00078674</td>
<td>0.028956754</td>
</tr>
<tr>
<td>Pavir.Aa02517</td>
<td>239.8175815</td>
<td>3.424148863</td>
<td>0.634311687</td>
<td>5.398211843</td>
<td>6.73E-08</td>
<td>7.46E-06</td>
</tr>
<tr>
<td>Pavir.Aa02526</td>
<td>20.12897762</td>
<td>-1.829988585</td>
<td>0.51374501</td>
<td>-3.56207357</td>
<td>0.000367937</td>
<td>0.015318345</td>
</tr>
<tr>
<td>Pavir.Aa02574</td>
<td>1.957536218</td>
<td>-5.978272647</td>
<td>1.222823914</td>
<td>-4.88907208</td>
<td>1.01E-06</td>
<td>8.89E-05</td>
</tr>
<tr>
<td>Pavir.Aa02621</td>
<td>0.909365395</td>
<td>-6.53529993</td>
<td>1.672432432</td>
<td>-3.907661562</td>
<td>9.32E-05</td>
<td>0.004726253</td>
</tr>
<tr>
<td>Pavir.Aa02666</td>
<td>26.27692122</td>
<td>0.691682664</td>
<td>0.195671446</td>
<td>3.534918755</td>
<td>0.000407901</td>
<td>0.01668753</td>
</tr>
<tr>
<td>Pavir.Aa02688</td>
<td>20.64051337</td>
<td>1.419916888</td>
<td>0.31120505</td>
<td>4.563880767</td>
<td>5.02E-06</td>
<td>0.000367199</td>
</tr>
<tr>
<td>Pavir.Aa02777</td>
<td>32.70837314</td>
<td>0.824566433</td>
<td>0.256714392</td>
<td>3.211999243</td>
<td>0.00131847</td>
<td>0.044226251</td>
</tr>
<tr>
<td>Pavir.Aa02799</td>
<td>5.953157198</td>
<td>1.635139531</td>
<td>0.489562315</td>
<td>3.340002856</td>
<td>0.000387775</td>
<td>0.030512025</td>
</tr>
<tr>
<td>Pavir.Aa02841</td>
<td>4.061306867</td>
<td>-1.69398357</td>
<td>0.345840001</td>
<td>-4.898171305</td>
<td>9.67E-07</td>
<td>8.51E-05</td>
</tr>
<tr>
<td>Pavir.Aa03067</td>
<td>7.20334301</td>
<td>-6.09679446</td>
<td>1.535018046</td>
<td>-3.971806374</td>
<td>7.13E-05</td>
<td>0.003773958</td>
</tr>
</tbody>
</table>
Populus: Linking genes to complex phenotypes

1000+ Populus Genomes

Reference Assembly  Meta Assembly

Core genome  Pan genome

SNP and INDEL detection

Genome Variants  Physical and Molecular Phenotype Data
GWAS visually

- GWAS tries to uncover links between **genetic basis of the disease**
- Which set of SNPs explain the phenotype?

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATGCAGTT</td>
<td>control</td>
</tr>
<tr>
<td>TTGCAGTT</td>
<td>control</td>
</tr>
<tr>
<td>CTGCAAGTT</td>
<td>control</td>
</tr>
<tr>
<td>ATGCGGTT</td>
<td>case</td>
</tr>
<tr>
<td>TTGCGGTT</td>
<td>case</td>
</tr>
<tr>
<td>CTGCCGTT</td>
<td>case</td>
</tr>
</tbody>
</table>

SNP
SNP Vectors

Phenotype Vectors
**Populus: Linking genes to complex phenotypes**

- 1000+ *Populus* Genomes
  - Reference Assembly
  - Meta Assembly
  - Core genome
  - Pan genome
    - SNP and INDEL detection
    - Genome Variants
      - Co-evolution Networks
      - Genome Wide Association Networks
      - GWAS-Co-evolution Networks
      - Co-expression Networks
      - Networks/Genes involved in plant-microbial interfaces
    - Physical and Molecular Phenotype Data
  
- Genome variant correlation calculation

*U.S. Department of Energy*
Network Theory: The Practical

- Networks can be used to represent biological systems
  - **Nodes**
    - Represent any object (genes, SNPs, proteins, metabolites, species, microbiomes, etc.)
  - **Edges**
    - Represent a relationship between two nodes (correlation, co-occurrence, physical contact, etc.)
    - Relationships can be quantitative (represented by the thickness of the line)

- Visualization

- Mathematical Structure
  - Allows to be computed upon
  - Millions of nodes
  - Trillions of edges

Species Co-occurrence Network
Co-evolution Networks: SNP Correlations

• Identify SNPs which have become fixed relative to one another
  – Linkage Disequilibrium
  – Selective pressure
  – Potentially co-evolving
Big data and high performance computing reveals the underlying biological signatures

- **8 million SNPs**, 1084 *Populus* Genomes
  - Arrayed in common gardens

- **$1.28 \times 10^{14}$** allele-specific SNP correlations calculated

- 367 million significant correlations

- Results modeled as a co-evolution network

- Network topology reflects the underlying biology
  - Genes under the same or similar selective pressures tend to co-evolve – which is reflected in SNP correlations and therefore, network topology

- Next: Increase to **50 million SNPs**
  - 10 million CPU hours
Big data and high performance computing reveals the underlying biological signatures

- **8 million SNPs**, 1084 *Populus* Genomes
  - Arrayed in common gardens

- **1.28 x 10^{14}** allele-specific SNP correlations calculated

- 367 million significant correlations
- Results modeled as a co-evolution network
- Network topology reflects the underlying biology
  - Genes under the same or similar selective pressures tend to co-evolve – which is reflected in SNP correlations and therefore, network topology

- Next: Increase to **50 million SNPs**
  - 10 million CPU hours
Big data and high performance computing reveals the underlying biological signatures

- **8 million SNPs**, 1084 *Populus* Genomes
  - Arrayed in common gardens
- **$1.28 \times 10^{14}$** allele-specific SNP correlations calculated
- 367 million significant correlations
- Results modeled as a co-evolution network

**Network topology reflects the underlying biology**
- Genes under the same or similar selective pressures tend to co-evolve
  - which is reflected in SNP correlations and therefore, network topology

- Next: Increase to **50 million SNPs**
  - 10 million CPU hours
Targets of Interest: Gene Co-evolution Networks
Big data and high performance computing reveals the underlying biological signatures

- **8 million SNPs**, 1084 *Populus* Genomes
  - Arrayed in common gardens
- **1.28 x 10^{14}** allele-specific SNP correlations calculated
- 367 million significant correlations
- Results modeled as a co-evolution network
- Network topology reflects the underlying biology
  - Genes under the same or similar selective pressures tend to co-evolve – which is reflected in SNP correlations and therefore, network topology
- Next: Increase to **50 million SNPs**
  - 10 million CPU hours

10 million CPU hours

- **1141 years** on a single computer
  - 1000 CPUs
- **1.2 days**
  - 350,000 CPUs
Infrastructure

• This is being achieved as a collaboration between Biosciences, and the Oak Ridge Leadership Computing Facility (OLCF) and the Compute & Data Environment for Science (CADES)
  – Large clusters
    • 300,000 CPUs
    • 18,000 GPUs
  – Preprocessing platform
    • Petabytes of storage
  – Large Memory Platforms
    • SGI 3 Tbyte Memory
  – Map Reduce/Hadoop Systems
    • Cray Urika XA

Next Project: Exploring Higher Order Combinatorial Complexity

150 million CPU hours

• 17,000 years on a single computer
• 17 days
  – Titan
GWAS Networks

- Statistically significant associations between multiple SNPs (in genes) and metabolites measured across a population of 1000 different *Populus* genotypes
- GWAS Networks provide context for the associations that are significant for more than one gene.
The use of GWAS networks led us to conceive of and develop a new method: 

GWAS-Profile Correlation
Preliminary: GWAS-Profile Correlation

- GWAS profile
  - List of phenotypes associated with each SNP
  - List of SNPs associated with each phenotype

- Correlate SNPs based on GWAS profiles
- Connect SNPs with common phenotype associations
GWAS-Profile Correlation Methods

- Proportional Similarity Index between GWAS profiles

<table>
<thead>
<tr>
<th></th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5</th>
<th>P6</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNP1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>SNP2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
GWAS-Profile Correlation Methods

• Proportional Similarity Index between GWAS profiles

\[
C(X, Y) = \frac{2 \sum_i \min(x_i, y_y)}{\sum_i (x_i + y_i)}
\]
GWAS Profile Correlation Methods

SNPs

<table>
<thead>
<tr>
<th>Phenotypes</th>
<th>S₁</th>
<th>S₂</th>
<th>S₃</th>
<th>...</th>
<th>S₈ million</th>
</tr>
</thead>
<tbody>
<tr>
<td>P₁</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P₂</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P₄₃₆</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
GWAS Profile Correlation Methods

SNPs

S1  S2  S3  ...  S8 million

Phenotypes

P1

P2

P436
GWAS Profile Correlation Methods

SNPs

S₁ S₂ S₃ \ldots S₈ million

Phenotypes

P₁ P₂ \ldots P₄36
GWAS Profile Correlation Methods

SNPs

\[ S_1 \quad S_2 \quad S_3 \quad \ldots \quad S_{8 \text{ million}} \]

Phenotypes

\[ P_1 \quad P_2 \quad \ldots \quad P_{436} \]
GWAS Profile Correlation Methods

SNPs

$S_1$ $S_2$ $S_3$ \ldots $S_{8\text{ million}}$

Phenotypes

$P_1$

$P_2$

$P_{436}$
GWAS Profile Correlation Methods

SNPs

\[ S_1 \quad S_2 \quad S_3 \quad \ldots \quad S_{8\text{ million}} \]

Phenotypes

\[ P_1 \]

\[ P_2 \]

\[ \ldots \]

\[ P_{436} \]
GWAS Profile Network

Connected components: SNPs on chromosome 19 that are within genes
GWAS Profile Network

- Mapped SNPs to genes
- Bridges connected components together

Legend:
- Gene
- SNP
GWAS-profile Network

Co-evolution Network

Legend:
- Gene
- SNP
GWAS-profile Network

Co-evolution Network

Legend:
- Gene
- SNP
SNP Cluster 1

SNP Cluster 2

Legend:
- Gen
- SNP

Potri.019G054600

Potri.019G054700
Lignin Focused Network
Seed Subnetwork Creation

1. Selected metabolites associated with lignin-annotated genes

2. Selected Caffeic acid metabolites and their respective flowers

3. Selected 2-hop SNP correlation neighbours

4. Added in co-expression edges
Nodes
- Metabolites
- Gene
- Lignin gene

Edges
- GWAS
- SNP-correlation
- Negative co-ex
- Positive co-ex

5-O-caffeoyl-quinate
Nodes
- Metabolites
- Gene
- Lignin gene

Edges
- GWAS
- SNP-correlation
- Negative co-ex
- Positive co-ex

Nodes:
- 5-O-caffeoyl-quinate
- 4-O-caffeoyl-quinate
Nodes
- Metabolites
- Gene
- Lignin gene

Edges
- GWAS
- SNP-correlation
- Negative co-ex
- Positive co-ex

Nodes:
- 5-O-caffeoyl-quinate
- 4-O-caffeoyl-quinate
- 3-O-caffeoyl-quinate
  - caffeoyl conjugate
Nodes
- Metabolites
- Gene
- Lignin gene

Edges
- GWAS
- SNP-correlation
- Negative co-ex
- Positive co-ex

- 5-O-caffeoyl-quinate
- 4-O-caffeoyl-quinate
- 3-O-caffeoyl-quinate caffeoyl conjugate

- Caffeoyl-quercetin-glycosides
Nodes
- Metabolites
- Gene
- Lignin gene

Edges
- GWAS
- SNP-correlation
- Negative co-ex
- Positive co-ex
- GWAS profile
• Peroxins – mentioned as being involved in lignin depolymerization in fungi
• Involved in Peroxisome biogenesis
• Glycosyltransferases involved in lignin biosynthesis
• Overexpression in tobacco lead to increased lignin content
Some Cytochromes involved in production of lignin monomers
“Maize Homologs of Hydroxycinnamoyltransferase, a Key Enzyme in Lignin Biosynthesis, Bind the Nucleotide Binding Leucine-Rich Repeat Rp1 Proteins to Modulate the Defense Response.”
“Maize Homologs of Hydroxycinnamoyltransferase, a Key Enzyme in Lignin Biosynthesis, Bind the Nucleotide Binding Leucine-Rich Repeat Rp1 Proteins to Modulate the Defense Response.”
“Maize Homologs of Hydroxycinnamoyltransferase, a Key Enzyme in Lignin Biosynthesis, Bind the Nucleotide Binding Leucine-Rich Repeat Rp1 Proteins to Modulate the Defense Response.”
Catharanthus roseus RLK1 - regulate cell wall function
- Cullins mediate responses to jasmonic acid.
- Cell wall damage-induced lignin biosynthesis is regulated by a reactive oxygen species- and jasmonic acid-dependent process in Arabidopsis.
Nodes
- Metabolites
- Wet chem phenotype

Edges
- GWAS
- SNP-correlation
- Negative co-ex
- Positive co-ex
- GWAS profile

Lignin monomers (Syringyl/Guaiacyl)
Nodes
- Metabolites
- Wet chem phenotype

Edges
- GWAS
- SNP-correlation
- Negative co-ex
- Positive co-ex
- GWAS profile

Alpha cellulose/hemicellulose

Alpha cellulose/hemicellulose

- Alpha cellulose/hemicellulose
- Nodes
- Metabolites
- Wet chem phenotype
- Gene
- Lignin gene
- Peroxin
- Glycosyltransferase/hydrolase
- Cytochrome
- Leucine Rich Repeat
- Defense/disease /wound
- Catharanthus roseus-like
- RLK1
- Cullin
Lignin Peaks
4-vinyl phenol; p-coumaric acid

Nodes
- Metabolites
- Wet chem phenotype
- pyMBMS phenotype
- Gene
- Lignin gene
- Peroxin
- Glycosyltransferase/hydrolase
- Cytochrome
- Leucine Rich Repeat
- Defense/disease /wound
- Catharanthus roseus-like
- RLK1
- Cullin

Edges
- GWAS
- SNP-correlation
- Negative co-ex
- Positive co-ex
- GWAS profile

Lignin Peaks
p-vinylguaiacol, comaryl alcohol, higher molecular weight methoxy phenols, Ethylguaiacol, homovanillin, coniferyl alcohol, coniferyl aldehyde
4-O-caffeoyl-quinate

DUF1118
4-O-caffeoyl-quinate

DUF1118

hydroxycinnamoyl transferase
4-O-caffeoyl-quinate

DUF1118

hydroxycinnamoyl transferase

Lignin pyMBMS peaks
DUF1118 eQTN: Expression Phenotypes
WRKY Transcription Factor

DUF1118 eQTN: Expression Phenotypes
WRKY Transcription Factor
HB Transcription Factor

DUF1118 eQTN:
Expression Phenotypes
WRKY Transcription Factor
HB Transcription Factor
Vesicle Transport

DUF1118 eQTN:
Expression Phenotypes
WRKY Transcription Factor
HB Transcription Factor
Vesicle Transport

DUF1118 eQTN: Expression Phenotypes

Signalling receptor kinases leucine rich repeat
Abiotic stress
Protein Kinase
Transcription Factor
Translational Regulation
Protein Degradation
Creating new, expanded hypotheses
Galacturonosyltransferase (GAUTs)

• GAUTs are key enzymes involved in cell wall (pectin and xylan containing polymers) synthesis (target genes in BESC)
  – Debra Mohnen, Center for Complex Carbohydrates Research, University of Georgia
GAUT11 pyMBMS Cluster
pyMBMS Annotations

- M/z 132: Methylbenzofuran(?) / 1-hydroxy-2-propanone acetate, xylan product
- M/z 154: 2,6-dimethoxyphenol (syringol)
- M/z 228: Chrysene, benz[a]anthracene, triphenylene, benzol[c]phenantherene
- M/z 180: Coniferyl alcohol, vinylsyringol, syringylethene ::: syringyl lignin; guaiacyl lignin, positive correlated to root lignin in populus
- 10 m/z values with no known annotation (m/z: 45, 90, 153, 230, 248, 258, 262, 290, 302, 393)
GAUT11 pyMBMS Cluster

- GAUT11
- Benzoyl conjugate
- Benzoyl glycoside
Gene Annotations Associated with pyMBMS Cluster

• Regulation
  – Transcription Factors (22)

• Signaling
  – G-proteins
  – Calcium
  – Receptor kinases
  – Hormone: ABA
  – Phosphotases

• Transport
  – Golgi
  – Vessicle
  – Protein Secretory

• Metabolism
  – dehydroquinate/shikimate
  – tryptophan synthase
  – GABA transaminase
  – FA elongation
  – Trehalose
  – Simple Phenols
  – Berberine bridge enzymes, reticuline oxidases, troponine reductases

• Cell Wall
  – UDP-glucosyl transferase 89B1 (UGT89B1) Beta 1,3 glucan hydrolases
  – Glucan endo-1,3-beta-glucosidase
  – Cytochrome P450
  – Cell wall proteins: RGP
  – Cell wall cellulose synthesis: COBRA
What about our friend DUF1118?
GAUT11 pyMBMS Cluster

DUF1118

GAUT11

Benzoyl conjugate

Benzoyl glycoside
4-O-caffeoyl-quinate

DUF1118

hydroxycinnamoyl transferase

GAUT 11

Lignin pyMBMS peaks
Plant-Microbial Interfaces

- Cross talk between Lignin and Cellulose/Hemicellulose
- Cross talk between Cell Wall and Cell Membranes (sphingolipids)
- Affects of Lignin, Cellulose and Sphingolips on Plant-Microbe interactions
Plant-Microbial Interfaces

• Lectin Receptor-like Kinases (LRKs)
  – Microbe recognition, essential for colonization

• Small Secreted Proteins (SSPs)
  – Some pass out of the plant cell and into microbial cells to presumably regulate microbial gene expression

• Higher Order Salicylates
  – Help shape the plant’s microbiome
    • Toxic to some microbes
    • Carbon source for others

Question: Do LRKs, SSPs and Salicylates work in an integrated fashion?
Salicylate-related metabolite

SSP

LRK

Biotic stress/disease/wounding

Ankyrin repeat protein

Cell wall related

Wall-associated kinase
Testable hypotheses: Putative models unfold

- Transport Across the Cell Membrane
- Microbe Recognition (Receptor Kinase)
- Signal Transduction (Protein Kinases)
- Regulation of Salicylate Biosynthesis
- Small Secreted Proteins
- Transcriptional Regulation
- Activate/Moderate Defense Response
- Regulate Microbes
Salicylate associated microbiomes

Transport Across the Cell Membrane

Microbe Recognition (Receptor Kinase)

Regulate Microbes

Regulation of Salicylate Biosynthesis

Signal Transduction (Protein Kinases)

Small Secreted Proteins

Transcriptional Regulation

Activate/Moderate Defense Response
Signal Transduction (Protein Kinases)

Regulation of Salicylate Biosynthesis

Transport Across the Cell Membrane

Regulate Microbes

Small Secreted Proteins

Transcriptional Regulation

Activate/Moderate Defense Response

Salicylate associated microbiomes

OTUs
Salicin
Benzoyl-populin
HCH-populin
Populin isomer
Galloyl-populin
Background on Viruses

Baltimore Classification

I. dsDNA virus (*Pycodnaviridae*)
II. ssDNA virus (*Geminiviridae*, *Nanoviridae*)
III. dsRNA virus (*Reoviridae*, *Partitiviridae*)
IV. ssRNA(+) virus (*Bromoviridae*, *Comoviridae*)
V. ssRNA(-) virus (*Rhabdoviridae*, *Bunyaviridae*)
VI. ssRNA(RT) virus (*Pseudoviridae*)

- Viruses don’t have marker genes
- Viruses are difficult to cultivate

Image From:

http://www.ncbi.nlm.nih.gov/genome/browse/
**Aim 1: Develop tools that will be used to identify diverse taxonomic composition, abundance and structure**

1. **Enrichment Strategy 1**
   - Develop tools that will be used to identify diverse taxonomic composition, abundance and structure.
   - Aim: Develop tools for identifying diverse taxonomic composition, abundance, and structure.
   - Method: Use tools to identify diverse taxonomic composition, abundance, and structure.

2. **Enrichment Strategy 2**
   - Used to identify diverse taxonomic composition, abundance, and structure.
   - Method: Use tools to identify diverse taxonomic composition, abundance, and structure.

3. **Enrichment Strategy 3**
   - Used to identify diverse taxonomic composition, abundance, and structure.
   - Method: Use tools to identify diverse taxonomic composition, abundance, and structure.

**Steps for Enrichment Strategy 1**

- **5000 x g for 10 minutes centrifugation**
- **0.22 μm filtration**
- **2 U. ul⁻¹ DNAse I** and/or **10 μg.mL⁻¹ RNAse A at 37°C for 1 hour**
- **Total nucleic acid extract (DNA, ssRNA, dsRNA) with 16.6% EtOH**
- **Repeat x 2: Wash Buffer (1 X STE 16% EtOH)**
- **dsRNA Flow-through (DNA, ssRNA)**
- **Elution Buffer (1 X STE)**
- **dsRNA Flow-through (purified dsRNA)**
- **RISC (guide strand) incorporated into RISC complex**
- **Size Selection**
- **Putative Viral Sequences**

**Tools used for Enrichment Strategy 1**

- **Cellulose I**
- **Dicer**: cleaves long dsRNA into short dsRNA
- **RISC + Argonaute + guide strand**: cleaves mRNA

**Enrichment Strategy 2**

- Enrichment Strategy 2
- (Okada et al 2015)

**Enrichment Strategy 3**

- Enrichment Strategy 3
- (Kreuze et al 2009)
Aim 1: Develop tools that will be used to identify diverse taxonomic composition, abundance and structure

- **Kraken**
- **Profile HMM**
- **Random Forest**

**Multiple Sequence Alignment**

- Similarity: Maximum scored root to leaf path

**Query Sequence**

- Map k-mer to LCA. Similarity: Maximum scored root to leaf path

Features = \{Profile HMM score, k-mer content\}
Aim 1: Develop tools that will be used to identify diverse taxonomic composition, abundance and structure

Risks

- Non-detection
- Low-abundance
- Low-diversity
- Bias in viral diversity
Microbes and Viruses Across the GWAS Population

Populus trichocarpa Genotypes
Receptor-Centric Network
Virus and Bacteria Receptor Intersection
Virus and Bacteria Intersection

Virus

Bacteria

Lectin Receptor Like Kinases
Virus and Bacteria Intersection

Transcription factor
Salicylic acid synthesis-degradation
Unknown function
Secondary metabolism: isoprenoids, terpenoids
Virus and Bacteria Intersection

Virus

- PTR2 family proton/oligopeptide symporter
- Sec14p-like phosphatidylinositol transfer family protein
- Unknown function
- Ankyrin repeat protein

Bacteria

- gwava_phenotype_1212
- Potri.019G079500.v3.0
- Potri.019G079600.v3.0
- Potri.019G107800.v3.0
- Potri.003G211800.v3.0
- Potri.003G211700.v3.0
- gwava_phenotype_1373
- Potri.010G174700.v3.0
- Potri.010G174600.v3.0
- Potri.072300.v3.0
- Potri.072400.v3.0

 PTR2 family proton/oligopeptide symporter, Sec14p-like phosphatidylinositol transfer family protein, Unknown function, Ankyrin repeat protein
Endophytic Microbiome/Viriome Community Structure

- Endophytic microbiomes from different *Populus* genotypes in common gardens
- OTU co-occurrence measured with Proportional Similarity Index
- Markov clustering applied to find modules of co-occurring OTUs
- Merged with a high threshold maximum spanning tree

Community Structure Network
(Markov clusters and MST)

- **Community profiles**
  - Can be used to design model communities
  - Each profile can be used to query the original microbiomes
    - Are their communities that are universal?
    - GWAS: What is it about those *Populus* genotypes that is driving community structure?
Focused Integrated Networks

• Abiotic stress
  – Drought/water stress
  – Aluminum
  – Cold stress
  – Heat stress

• Biotic Stress
  – Plant defense
  – Non-host defense

• Macronutrients
  – Phosphate starvation
  – Nitrogen fixation
  – Calcium

• Senescence

• Seed Development

• Specific Genes
  – GAUTs & GATLs
  – FPGS
  – Angustifolia
  – PolyQ Genes
  – Dwarf Genes
  – Photosynthesis
  – Auxin
  – CAZy Enzymes
  – Isoprenoids
  – Lignin
  – Laccases

• Cell Components
  – Cell wall
  – Golgi

• Gravitropism

• Plant-microbial interactions
  – Lectin-like Receptor Kinases
  – Small Signaling Proteins
  – Salicylates
Future Work
Phenologs

• Genes that, despite great evolutionary distance, are performing similar functions in trees and humans....
Disrupted ecosystem
Disrupted ecosystem
Mining for Phenologous Patterns

• Create GWAS Networks for *Populus* and Humans
• Find distant orthologs with the use of Hidden Markov Models
  – New method: HMMMCL
• Use orthologous relationships as a bridge between the two GWAS networks
• Extract subnetworks of interest
Neuroblastoma

- Most common cancer in infants
- Arises from neural crest elements during development
Human: Neuroblastoma Phenotype
Poplar: Growth Phenotypes
Poplar: Budflush, Developmental Phenotype
Neuroblastoma

- Most common cancer in infants
- Arises from neural crest elements during development
- Intersects with several *Populus* phenotypes, including a growth and development related phenotypes.
- Many of the *Populus* genes associated with the intersecting phenotypes are involved in embryogenesis or embryo development
Questions to explore

- Are there analogous, fundamental molecular mechanisms for host-microbiome interactions for plant and humans?
- How much (if any) of the phytobiome is shared with the human microbiome?
3-way Networks: Application of Hypergraphs for Modelling Increased Complexity in Comparative Genomics

- More sophisticated models
  - Captures complex relationships
  - Developed new n-way similarity metric
  - Compares triplets of genomes instead of pairs

- Found relationships between genomes not found with two-way comparisons

- Good combination with gene family enrichment networks

\[ C_3(X, Y, Z) = \frac{3}{2} \sum_i \left( \min(X_i, Y_i) + \min(X_i, Z_i) + \min(Y_i, Z_i) - \min(X_i, Y_i, Z_i) \right) \frac{1}{\sum_i (X_i + Y_i + Z_i)} \]
Scaling up Computational Biology to Address Biological Complexity: High Performance Computing

• Genome Wide Association Studies
  – 6.4 billion statistical tests this year
  • 4.32 trillion statistical tests later this year

• SNP Correlation Networks
  – $1.2 \times 10^{14}$ SNP correlations
  • $1.2 \times 10^{15}$ SNP correlations later this year
  – All JGI re-sequencing projects
  – All microbial genomes with more than 50 Strains (100 species)

• Gene Family Networks
  – $6.25 \times 10^{16}$ Gene Comparisons to do later this year
Exascale Computing: 3-way Networks

- Able to capture complexity better
- Scale in an $n^3$ fashion
- 3-way genome comparisons of 50,000 genomes
  - $1.25 \times 10^{14}$ Genome Comparisons
- 3-way comparisons of 50 million SNPs
  - $1.6 \times 10^{23}$ SNP Comparisons
- 3-way gene family comparisons of 250 million genes
  - $1.6 \times 10^{25}$ Gene Comparisons
Exascale Computing: 4-way Networks

- Able to capture complexity better
- Scale in an $n^4$ fashion
- 4-way genome comparisons of 50,000 genomes
  - $6.25 \times 10^{18}$ Genome Comparisons
- 4-way comparisons of 50 million SNPs
  - $6.25 \times 10^{30}$ SNP Comparisons
- 4-way gene family comparisons of 250 million genes
  - $3.9 \times 10^{33}$ Gene Comparisons
Combinatorial Applications for Systems Biology Analysis
ORNL Supercomputing

• Scaling up to meet these needs is being achieved as a collaboration between Biosciences, Computer Sciences & Mathematics, OLCF and CADES
Analytical Chemistry

Microbiology

Plant Systems Biology

- 1084 Genomes
- Transcriptomes
- Phenotypes
- Target Genes

Computational Biology

- Statistics
- Networks
- Process Models
- Context

- Metabolomes
- Proteomes
- Compound IDs

Microbiology

- Automated Extraction
- Statistics
- Networks
- ID Priorities

Model Community Design

- OTU Data
- Genomes
- Metagenomes
- Metaomics
- Model Community Data

System-wide understanding of Complex Biological Systems
Acknowledgements

- Computational Biology Team last year

Dan Jacobson  Debbie Weighill  Carissa Bleker  Piet Jones
Acknowledgements

14 People in the group now
Acknowledgements

Wet lab Co-authors

- Jerry Tuskan
- Wellington Muchero
- Mark Davis
- Tim Tschaplinski

Computational Infrastructure

- Oak Ridge Leadership Computing Facility (OLCF)
- Compute and Data Environment for Science (CADES)
Funding

• DOE:
  – Bioenergy Science Center
  – Plant Microbial Interfaces
  – Correlotypes: Determining complex genotypic profiles responsible for complex phenotypes
  – An Experimental and Computational Framework for Directed Succession: Unraveling Cobamide Control of Microbial Community Assembly, Structure and Function
  – Using a multi-omics approach to unravel the complex control mechanisms limiting oleaginous synthesis in yeast and develop a new class of large-scale production organisms
  – Genome Wide Association of Viriome/Microbiome Analysis (GWAVA)
Expanding

• Hiring:
  – Post Doctoral Fellows
  – Post Masters
  – Post Bachelors
  – PhD Students