Chestnut Chat
GWAS 101

David Kainer
Computational Biology, Biosciences Division, ORNL
kainerd@ornl.gov
GWAS 101

1. Overview of GWAS
2. What kind of traits can be analyzed?
3. Phenotypes and Genotypes
4. The three steps of GWAS
5. A case study in Eucalyptus
6. The importance of sample size
7. Population Structure
8. Interpreting results
GWAS according to the NIH

A genome-wide association study (GWAS) is an approach used in genetics research to associate specific genetic variations with particular diseases.

The method involves scanning the genomes from many different people and looking for genetic markers that can be used to predict the presence of a disease. Once such genetic markers are identified, they can be used to understand how genes contribute to the disease and develop better prevention and treatment strategies.

GWAS according to me

Phenotype = Genotype + Environment
GWAS for pathogen resistance in trees

**Poplar**

Association mapping, transcriptomics, and transient expression identify candidate genes mediating plant–pathogen interactions in a tree


* Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831; † Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR 97331; ‡Complex Carbohydrate Research Center, University of Georgia, Athens, GA 30602; §Joint Genome Institute, US Department of Energy, Walnut Creek, CA 94596; †′Department of Plant Pathology, North Dakota State University, Fargo, ND 58102; †‖HudsonAlpha Institute for Biotechnology, Huntsville, AL 35896; and †||Forest Engineering, Resources, and Management, Oregon State University, Corvallis, OR 97331

**Beech**

Genome-wide association study identifies a major gene for beech bark disease resistance in American beech (Fagus grandifolia Ehrh.)

Irina Cali, Jennifer Koch, David Carey, Charles Addo-Quaye, John E. Carson and David B. Neale

**Ash**

Genomic basis of European ash tree resistance to ash dieback fungus


**Eucalyptus**

A Genome-Wide Association Study for Resistance to the Insect Pest Leptocybe invasa in Eucalyptus grandis Reveals Genomic Regions and Positional Candidate Defense Genes

Lorraine Mhoswa, Marja M. O'Neill, Makobatjari M. Mphahlele, Caryl N. Oates, Kirt G. Payn, Bernard Slippers, Alexander A. Myburg and Sanushka Naidoo
GWAS for wood properties

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Species</th>
<th>Population</th>
<th>Sample size</th>
<th>No. of markers</th>
<th>Method</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Growth and wood properties</td>
<td>Eucalyptus globulus</td>
<td>Families and bulk collections</td>
<td>303</td>
<td>7,680 (DArT)</td>
<td>General linear model (GLM) and unified mixed model (UMM)</td>
<td>Cappe et al., 2013</td>
</tr>
<tr>
<td>Wood density, stiffness, microfibril angle, and ring width</td>
<td>Pinus glauca</td>
<td>Open-pollinated families</td>
<td>1694</td>
<td>7434 (SNPs)</td>
<td>Mixed linear model (MLM)</td>
<td>Lamara et al., 2016</td>
</tr>
<tr>
<td>16 wood chemistry/ultrastructure traits</td>
<td>Populus trichocarpa</td>
<td>Unrelated individuals</td>
<td>334</td>
<td>29,233 (SNPs)</td>
<td>GLM</td>
<td>Porth et al., 2013</td>
</tr>
<tr>
<td>Lignin percentage, Lignin S/G ratio, 5-carbon sugars, and 6-carbon sugars</td>
<td>Populus deltoides</td>
<td>Unrelated individuals</td>
<td>391</td>
<td>334,679 (consensus SNPs), 185,526 (Common SNPs), 76,804 (functional SNPs)</td>
<td>Single-variant and multiple-variant associations on GLM</td>
<td>Fahrenkrog et al., 2017</td>
</tr>
<tr>
<td>Basic wood density (BWD), bleached pulp, pulp yield (SPY), and pulp bleaching content</td>
<td>Eucalyptus grandis x Eucalyptus urophylla</td>
<td>Hybrid breeding population</td>
<td>768</td>
<td>24,806 (SNPs)</td>
<td>GWAS and regional heritability mapping</td>
<td>Resende et al., 2017</td>
</tr>
<tr>
<td>17 wood-quality traits</td>
<td>Norway spruce</td>
<td>Mother trees</td>
<td>517</td>
<td>178,101 (SNPs)</td>
<td>Multitocus LASSO penalized regression</td>
<td>Baison et al., 2018</td>
</tr>
<tr>
<td>Seven wood properties</td>
<td>Populus tomentosus</td>
<td>Unrelated individuals</td>
<td>436</td>
<td>5,482 (InDels)</td>
<td>MLM and Kemphtheon model</td>
<td>Gong et al., 2017</td>
</tr>
</tbody>
</table>

Diversity Array Technology (DArT) markers.

You can GWAS almost anything you can measure

Typical GWAS traits in trees

• Growth: Height, DBH, crown architecture, branching, leaf shape
• Wood properties: lignin, sugars, microfibril angle
• Metabolic: primary or secondary metabolite abundances
• Pathogens: Fungal/bacterial abundances
• Resistance: inoculation outcomes
• Adaptive: flowering time, leaf senescence
• Sustainability: water use efficiency, nitrogen use
• Gene Expression: Gene Transcript (RNAseq) abundances
Phenotypic variation

- Without variation there is little to explore
- Variation is the source material
- Is the phenotypic variation due to:
  - Genetic variation?
  - Environmental variation?
  - Both?

Phenotype = Genotype + Environment
genetic variation

An individual will have:

- or
- or

Which **may or may not affect their phenotype** (e.g. +/- for quantitative trait, or increased resistance to a disease)
genetic variation in a population
Population Genetic and Phenotypic variation

1 effective variant

frequency

phenotype

simple
Population Genetic and Phenotypic variation

Many effective variants

simple

complex
Population Genetic and Phenotypic variation

We observe this...

We don’t know what goes in here!!!
Genome-wide variation

Across the entire genome, there may be millions of variant locations (e.g. SNPs) that have alleles in a population.

Q: Which ones have an effect on a trait?
A: GWAS
How a GWAS works: quantitative trait

Take a population of mostly un-related individuals. Measure a phenotype that varies (e.g. Height)
How a GWAS works: quantitative trait

1. Take a population of mostly un-related individuals. Measure a phenotype that varies (e.g. Height)

2. Sequence the DNA of each of them. Find positions in the genome where the individuals vary (e.g. SNPs)
How a GWAS works: quantitative trait

1. Take a population of mostly un-related individuals. Measure a phenotype that varies (e.g., Height)

2. Sequence the DNA of each of them. Find positions in the genome where the individuals vary (e.g., SNPs)

3. Test each SNP to see if the alleles significantly correlate with the phenotypic variation. E.g. Does having more copies of the allele equate to a significant linear increase in height?
How a GWAS works: binary trait

Take a population of mostly un-related individuals. Split them into two groups (e.g. resistant / susceptible)

Sequence the DNA of each of them. Find positions in the genome where the individuals vary (e.g. SNPs)

Test each SNP to see if its alleles have significantly different frequency in one group compared to the other
Three main steps for GWAS

1. Phenotyping
2. Genotyping
3. Modelling
Three main steps for GWAS

1. Phenotyping
   - Field sites
   - Sampling time
   - Assay costs
   - Manpower

2. Genotyping
   - Library prep costs
   - Lab time
   - Sequencing costs

3. Modelling
   - Statistical power
   - Computational power

How Many Accessions?

Interpretation
GWAS for oil traits in *Eucalyptus polybractea*
Phenotyping

Sampled 480 trees

Extract oil in ethanol

GC-MS

West Wyalong

Metabolite profiles

<table>
<thead>
<tr>
<th></th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>Mn</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>0</td>
<td>300</td>
<td>12</td>
<td>65</td>
</tr>
<tr>
<td>S2</td>
<td>20</td>
<td>500</td>
<td>90</td>
<td>0</td>
</tr>
<tr>
<td>S3</td>
<td>2</td>
<td>50</td>
<td>250</td>
<td>0</td>
</tr>
<tr>
<td>S480</td>
<td>120</td>
<td>100</td>
<td>15</td>
<td>0</td>
</tr>
</tbody>
</table>
Phenotyping
Phenotyping pitfall – environmental variance
Phenotyping

• Minimize environmental variance
• Randomization is necessary to avoid batch effects or environmental trends
• Accuracy is hugely important
• consistency
Genotyping

Genotyping

West Wyalong

Sampled 480 trees

Snap freeze with liquid Nitrogen

Extract DNA

Genotype matrix

<table>
<thead>
<tr>
<th>Genotype matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>snp1</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>S1</td>
</tr>
<tr>
<td>S2</td>
</tr>
<tr>
<td>S3</td>
</tr>
<tr>
<td>S480</td>
</tr>
</tbody>
</table>
Genotyping

Snap freeze with liquid Nitrogen

Extract DNA

Sampled 480 trees

West Wyalong

Genotype matrix

<table>
<thead>
<tr>
<th></th>
<th>snp1</th>
<th>snp2</th>
<th>snp3</th>
<th>snpP</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>A/A</td>
<td>C/T</td>
<td>A/T</td>
<td>G/G</td>
</tr>
<tr>
<td>S2</td>
<td>A/C</td>
<td>T/T</td>
<td>A/A</td>
<td>G/G</td>
</tr>
<tr>
<td>S3</td>
<td>A/A</td>
<td>C/C</td>
<td>A/T</td>
<td>G/G</td>
</tr>
<tr>
<td>Sm</td>
<td>A/A</td>
<td>C/C</td>
<td>A/T</td>
<td>T/T</td>
</tr>
</tbody>
</table>

Detect variants
2 Genotyping

Reference genome

A

Fastq reads

A/G

A/A
Genotyping pitfall: More accessions or more sequencing depth?

N accessions

- High sequencing depth per sample
- Higher genotype accuracy
- Lower genetic diversity
- Lower statistical power

2N accessions

- Low depth per sample
- Lower genotype accuracy
- Higher genetic diversity
- Higher statistical power

Or...
modeling

<table>
<thead>
<tr>
<th></th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>Mn</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>0</td>
<td>300</td>
<td>12</td>
<td>65</td>
</tr>
<tr>
<td>S2</td>
<td>20</td>
<td>500</td>
<td>90</td>
<td>0</td>
</tr>
<tr>
<td>S3</td>
<td>2</td>
<td>50</td>
<td>250</td>
<td>0</td>
</tr>
<tr>
<td>S480</td>
<td>120</td>
<td>100</td>
<td>15</td>
<td>0</td>
</tr>
</tbody>
</table>

**Genotype matrix**

<table>
<thead>
<tr>
<th></th>
<th>snp1</th>
<th>snp2</th>
<th>snp3</th>
<th>snpP</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>A/A</td>
<td>C/T</td>
<td>A/T</td>
<td>G/G</td>
</tr>
<tr>
<td>S2</td>
<td>A/C</td>
<td>T/T</td>
<td>A/A</td>
<td>G/G</td>
</tr>
<tr>
<td>S3</td>
<td>A/A</td>
<td>C/C</td>
<td>A/T</td>
<td>G/G</td>
</tr>
<tr>
<td>S480</td>
<td>A/A</td>
<td>C/C</td>
<td>A/T</td>
<td>T/T</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>snp1</th>
<th>snp2</th>
<th>snp3</th>
<th>snpP</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>S2</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>S480</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>
modeling

Genotype matrix

<table>
<thead>
<tr>
<th></th>
<th>snp1</th>
<th>snp2</th>
<th>snp3</th>
<th>snpP</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>A/A</td>
<td>C/T</td>
<td>A/T</td>
<td>G/G</td>
</tr>
<tr>
<td>S2</td>
<td>A/C</td>
<td>T/T</td>
<td>A/A</td>
<td>G/G</td>
</tr>
<tr>
<td>S3</td>
<td>A/A</td>
<td>C/C</td>
<td>A/T</td>
<td>G/G</td>
</tr>
<tr>
<td>Sm</td>
<td>A/A</td>
<td>C/C</td>
<td>A/T</td>
<td>T/T</td>
</tr>
</tbody>
</table>

Filtering:
- Remove rare SNPs
- Remove SNPs that are likely to be erroneous

metabolite matrix

<table>
<thead>
<tr>
<th></th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>Mn</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>0</td>
<td>300</td>
<td>12</td>
<td>65</td>
</tr>
<tr>
<td>S2</td>
<td>20</td>
<td>500</td>
<td>90</td>
<td>0</td>
</tr>
<tr>
<td>S3</td>
<td>2</td>
<td>50</td>
<td>250</td>
<td>0</td>
</tr>
<tr>
<td>S480</td>
<td>120</td>
<td>100</td>
<td>15</td>
<td>0</td>
</tr>
</tbody>
</table>
3 modeling

**Linear model:**
\[ Y = u + b_1 x_1 + e \]

**P-value**

**Table 1:** Metabolite matrix

<table>
<thead>
<tr>
<th></th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>Mn</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>0</td>
<td>300</td>
<td>12</td>
<td>65</td>
</tr>
<tr>
<td>S2</td>
<td>20</td>
<td>500</td>
<td>90</td>
<td>0</td>
</tr>
<tr>
<td>S3</td>
<td>2</td>
<td>50</td>
<td>250</td>
<td>0</td>
</tr>
<tr>
<td>S480</td>
<td>120</td>
<td>100</td>
<td>15</td>
<td>0</td>
</tr>
</tbody>
</table>

**Table 2:** Genotype matrix

<table>
<thead>
<tr>
<th></th>
<th>snp1</th>
<th>snp2</th>
<th>snp3</th>
<th>snpP</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>A/A</td>
<td>C/T</td>
<td>A/T</td>
<td>G/G</td>
</tr>
<tr>
<td>S2</td>
<td>A/C</td>
<td>T/T</td>
<td>A/A</td>
<td>G/G</td>
</tr>
<tr>
<td>S3</td>
<td>A/A</td>
<td>C/C</td>
<td>A/T</td>
<td>G/G</td>
</tr>
<tr>
<td>Sm</td>
<td>A/A</td>
<td>C/C</td>
<td>A/T</td>
<td>T/T</td>
</tr>
</tbody>
</table>

**Table 3 (simplified):** Genotype matrix

<table>
<thead>
<tr>
<th></th>
<th>snp1</th>
<th>snp2</th>
<th>snp3</th>
<th>snpP</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>S2</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Sm</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>
Problem with our model

\[ Y = u + b_i x_i + e \]

**Population structure**
- Samples are often genetically related to each other (even if you think they aren’t!), which means they are not independent and can cause false positive SNP associations.
Problem with our model

\[ Y = u + b_i x_i + k + e \]

**Population structure**
- Samples are often genetically related to each other (even if you think they aren't!), which means they are not independent and can cause false positive SNP associations.
Results: Manhattan plot

sesquiterpenes

Multiple testing correction threshold
Results
Candidate genes!
Next Steps

• Validation !!
• Multi-omics integration
• Try other models
• Build the story
Thank You

- Carsten Kulheim
- William Foley
- Carlos Bustos
- Amanda Padovan
- Daniel Jacobson
- Jerry Tuskan