# Evaluating the Effects of Trembling Aspen Genetics on Insect Communities

Stat 877 project presentation

Clay Morrow and Ting-Fung Ma

May 2, 2019

# Background

## Plant-insect interactions:

- Plants are known to influence insect community structure
- Plant Diversity is positively correlated with insect diversity.
  - effects of interspecific variation are well studied
  - effects of **intraspecific variation** are less well studied

## *Populus tremuloides* (Trembling Aspen):

- the most wide-spread tree species in North America
- one of the fastest-growing tree species
- incredible amounts of intraspecific variation
- food source for an incredible diversity of herbivores

# Genome-wide association analysis (GWA)

## Question: How do aspen genetics shape insect communities?

- Herbivorous insects are known to differ by aspen clone (genet) and by aspen traits (size, defense chemistry, nutrition, etc.).
- Is there evidence for genetic drivers other than these physical traits?
- **Are there genomic regions that affect multiple insects?**

## GWA of insect incidence on aspen common garden:

- 18 common insect species; present/absent (1/0)
- 1,414 trees from 437 genets
- 8 tree trait covariates
- 4 time periods (longitudinal)
- 114,420 SNP regions with 3 possible genotypes $\{0,1,2\}$

# Statistical Model

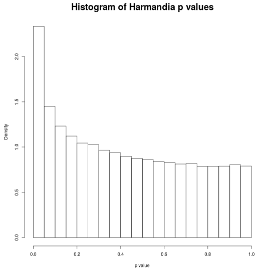
$$\mathsf{logit}(p_{ijk_g}) = \beta_0 + \alpha \mathsf{G}_g + x_{jk_g}^{\mathsf{T}}\beta + \varepsilon_{g(j)}$$

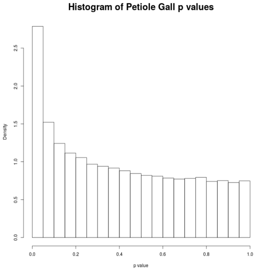
Where:

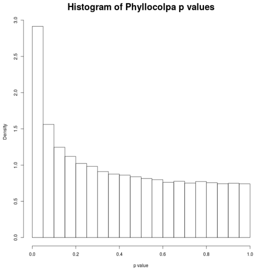
- $\mathsf{logit}(y) = \log\left(\frac{y}{1-y}\right)$
- $p_{ijk_g}$: probability of observing $> 0$ individual insects of species $i$ during survey event $j$ on tree $k$ of genet $g$
- $\mathsf{G}_g$: SNP-specific genotype of genet $g$
- $x_{jk_g}$: vector of observed tree trait covariates for tree $k$ during survey event $j$
- $\varepsilon_{g(j)}$: nested random effect of genet within survey event
- Storey's q values of coefficient $\alpha$ used to select significant SNPs (.05 cutoff)

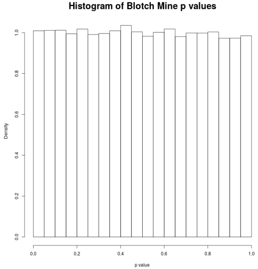
A total of 2,059,560 ($18$ insects × $114,420$ SNPs) GLMMs were fit.

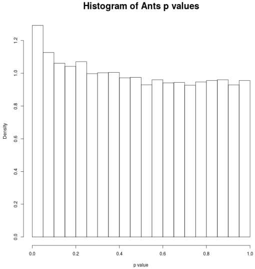
# Diagnostics: p value distributions

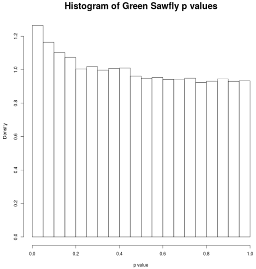

Histogram of Harmandia p values


Histogram of Petiole Gall p values

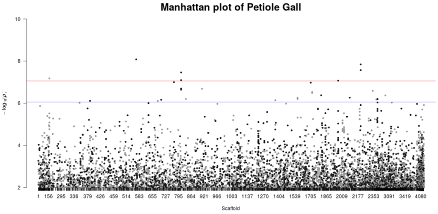

Histogram of Phyllocolpa p values

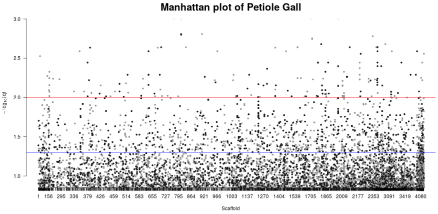

Histogram of Blotch Mine p values


Histogram of Ants p values


Histogram of Green Sawfly p values

# Diagnostics: Bonferroni vs. Storey's q


Manhattan plot of Petiole Gall


Manhattan plot of Petiole Gall

- Using q values (bottom) allows for discovery of many more potential associations than bonferroni p values (top).
- significant q values contain all significant bonferroni p values

# Results: Significant SNPs per Insect

| | insect description | SNPs | scaffolds |
|---|---|---|---|
| Green Aphids | free-feeding, specialist (salicaceae) | 1788 | 1020 |
| Petiole Gall | leaf-galling, specialist (populus) | 1432 | 788 |
| Phyllocolpa | leaf-rolling, specialist (salicaceae) | 1363 | 814 |
| Harmandia | leaf-galling, specialist (populus) | 174 | 130 |
| Smokey Aphids | free-feeding, specialist (populus) | 90 | 64 |
| Casebearer Moth | case-bearing, generalist | 7 | 3 |
| Lombardy Mine | leaf-mining, specialist (populus) | 1 | 1 |
| Cottonwood Leaf Mine | leaf-mining, specialist (salicaceae) | 1 | 1 |
| Leaf Edge Mine | leaf-mining, specialist (salicaceae) | 0 | 0 |
| Blotch Mine | leaf-mining, specialist (populus) | 0 | 0 |
| Weevil Mine | leaf-mining, specialist (salicacea) | 0 | 0 |
| Blackmine Beetle | leaf-mining, specialist (populus) | 0 | 0 |
| Leafhoppers | free-feeding, generalist | 0 | 0 |
| Ants | aphid-tending, non-herbivore | 0 | 0 |
| Pale Green Notodontid | free-feeding, specialist (populus) | 0 | 0 |
| Aspen Leaf Beetle | free-feeding, specialist (populus) | 0 | 0 |
| Green Sawfly | free-feeding, specialist (populus) | 0 | 0 |
| Cotton Scale | scale insect, generalist | 0 | 0 |

There are 4,768 unique significant SNPs.

# Results: Common Associations

- Of these 4,768 unique SNPs, There are **84** SNPs with multiple insect associations. These will be our SNPs of interest.
- 4 of these had associations with 3 insects and were contained in 1 gene:

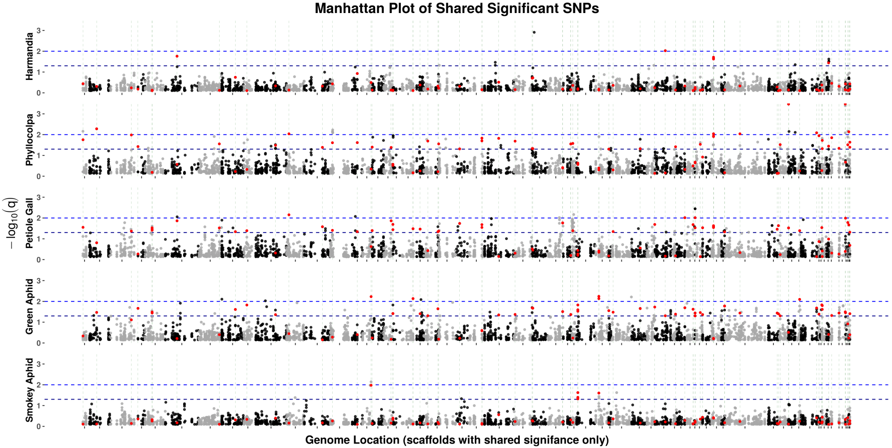
| SNP.name | qval.Harmandia | qval.Phyllocolpa | qval.Petiole.Gall |
|---|---|---|---|
| Potra002191:26587 | 0.0193784 | 0.0123563 | 0.0286529 |
| Potra002191:26642 | 0.0226734 | 0.0091460 | 0.0258001 |
| Potra002191:26644 | 0.0226734 | 0.0091460 | 0.0258001 |
| Potra002191:26650 | 0.0236877 | 0.0116280 | 0.0234547 |

- They are within 63 base pairs (LD/same gene)

Note that these are the 2 **leaf-galling** and 1 **leaf-rolling** species. They are tightly associated with leaves. However, the gene annotation is not very insightful:

| Potri Gene | %match | Descr. | avg allele. freq |
|---|---|---|---|
| Potri\|011G157300.1 | 98.97 | Uncharacterized oxidoreductase C663 | 0.136 |

# Results: Multiple-Manhattan Plot



# Results: Imputed Gene Function

The following is a few of our top (by % gene match) genes with annotations, imputed from congeneric *Populus trichocarpa*.

| Potri Gene | %match | Descr. | avg allele. freq |
|---|---:|---|---:|
| Potri\|006G267600.1 | 99.66 | protein phosphatase 2C | 0.110 |
| Potri\|013G036000.1 | 99.44 | tRNA-specific adenosine deaminase | 0.204 |
| Potri\|007G058500.1 | 99.02 | 1-deoxy-D-xylulose-5-phosphate synthase | 0.226 |
| Potri\|014G079900.1 | 99.01 | dependent malic enzyme | 0.112 |
| Potri\|003G045300.1 | 99.01 | multivesicular body protein | 0.143 |
| Potri\|011G157300.1 | 98.97 | Uncharacterized oxidoreductase C663 | 0.136 |
| Potri\|006G249800.1 | 98.63 | Probable boron transporter | 0.139 |
| Potri\|016G142100.1 | 98.60 | Unknown protein 1 | 0.150 |
| Potri\|001G369000.1 | 98.59 | Transmembrane emp24 domain-containing protein | 0.205 |
| Potri\|016G037900.1 | 98.40 | STRICTOSIDINE SYNTHASE-LIKE | 0.170 |
| Potri\|006G082900.1 | 98.38 | Kinesin-like protein | 0.218 |
| Potri\|010G014100.1 | 98.33 | repeat-containing protein | 0.141 |
| Potri\|015G112500.1 | 98.16 | hydrolase domain-containing | 0.126 |
| Potri\|014G156100.1 | 98.00 | CSC1-like protein | 0.137 |
| Potri\|001G113500.1 | 98.00 | domain-containing protein | 0.403 |
| Potri\|005G061600.1 | 97.99 | glucuronate:xylan alpha | 0.183 |
| Potri\|002G009100.1 | 97.98 | Bromodomain-containing protein | 0.230 |
| Potri\|010G164400.1 | 97.93 | uncharacterized protein LOC105116966 isoform X1 | 0.156 |
| Potri\|006G149900.1 | 97.81 | RNA-binding protein | 0.150 |
| Potri\|001G253900.1 | 97.81 | Protein TIC 62, chloroplastic | 0.222 |
| Potri\|001G253800.1 | 97.67 | Sugar transport protein | 0.211 |
| Potri\|010G164600.1 | 97.66 | Transmembrane emp24 domain-containing protein | 0.171 |

These annotations are vague and will require more research to determine which

# Discussion

## Conclusions

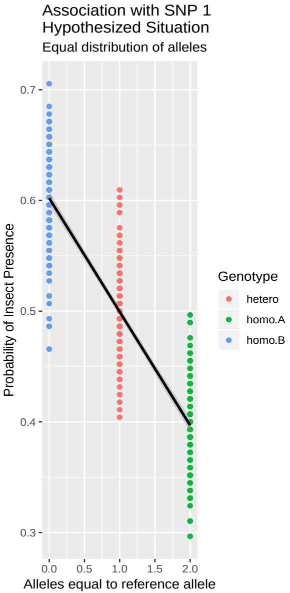
- Evidence of genomic regions that influence individual insects and communities
- These genomic regions are independent of the observed tree traits

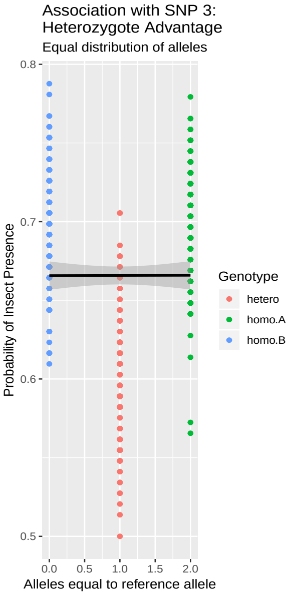
## Future Directions

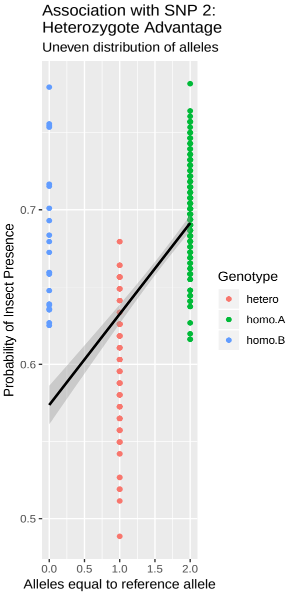
- More thoroughly research gene annotations from congeneric *Populus trichocarpa* and model *Arabidopsis thaliana* to check biological function where possible
- Compare with other methods (multi-SNP, BLUP association, GBlups, etc)
  - A similar study on this garden found only 2 insect associations with $< 12$ total significant SNPs using BLUP association and a more lenient inclusion cutoff ($q < 0.1$) in 2015
- Include more tree traits and environmental variables and interactions

# Extra Slides

## Extra Slides: Potential Issues (Simulated Data)







## BLUPs: Linear Mixed Model (LMM) Definition

Consider the following model for some response $Y$ on observation $j$:

$$Y_j = \mu + X_j\beta + Z_j\tau + \varepsilon$$

- $E(Y_j = X_j\beta)$
- $X_j$ and $Z_j$ are vectors of independent variables for the $j$th observation
- $\beta$ is an unknown vector of regression parameters (fixed effects)
- $\tau$ is an unknown vector of **random effects**
  - $E(\tau) = 0$ and $Var(\tau) = G$; $G = Cov(\tau_m, \tau_k)$ for all $m$ and $k$
- $\varepsilon$ is an unknown vector of random errors
  - $E(\varepsilon) = 0$ and $Var(\varepsilon) = R$

## LMM: Parameter Estimation

$$Y_j = \mu + X_j\beta + Z_j\tau + \varepsilon$$

- $\tau \sim N(0, G)$, $\varepsilon \sim N(0, R)$, and $Cov(\tau, \varepsilon) = 0$
- Find coefficients $\hat{\beta}$ and $\hat{\tau}$ such that they:
  - minimize variance prediction error $Var(Y_j - \hat{Y}_j)$
  - are constrained by $E(Y_j - \hat{Y}_j) = 0$ $(Bias(\hat{\theta}) = E(\hat{\theta}) - \theta)$
- $\hat{\beta}$ is a vector of best linear unbiased estimators (BLUEs) of fixed effects
- $\hat{\tau}$ is a vector of best linear unbiased predictors (BLUPs) of fixed effects
- conditional variances (covariance matrices for $\beta$ and $\varepsilon$) are often unknown and are estimated as a nuisance parameter with Bayesian EM algorithm (`nlme`,`lme4`, etc.)

## Take-Home Message

BLUPs are constrained coefficienct estimates of random effects.

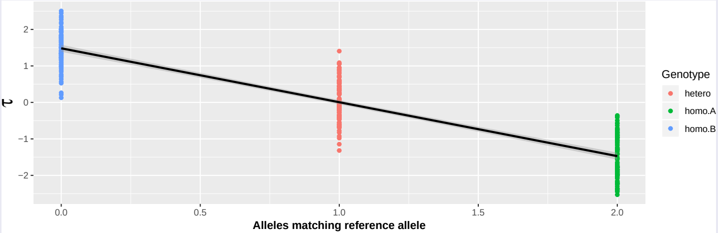
## Modern GWA Models

- First fit LMM for a trait of interest (Size) with $\beta_i$ fixed effects and $\tau_g$ random effects:

$$\text{Size} = \mu + \beta_1(\text{PGs}) + \beta_2(\text{CTs}) + \cdots + \beta_n(\text{Trait}_n) + \tau_g(\text{Genet}_g) + \varepsilon$$

- Then conduct a linear association analysis between the size-specific BLUPs $\tau$ and the Genotype of at SNP $s$ of each genet $g$:

$$\tau_g = \mu + \alpha(\text{Genotype}_{s,g}) + \varepsilon$$


τ
Alleles matching reference allele
Genotype
● hetero
● homo.A
● homo.B