# Detecting Chromosomal Inversions from Dense SNPs by Combining PCA and Association Tests

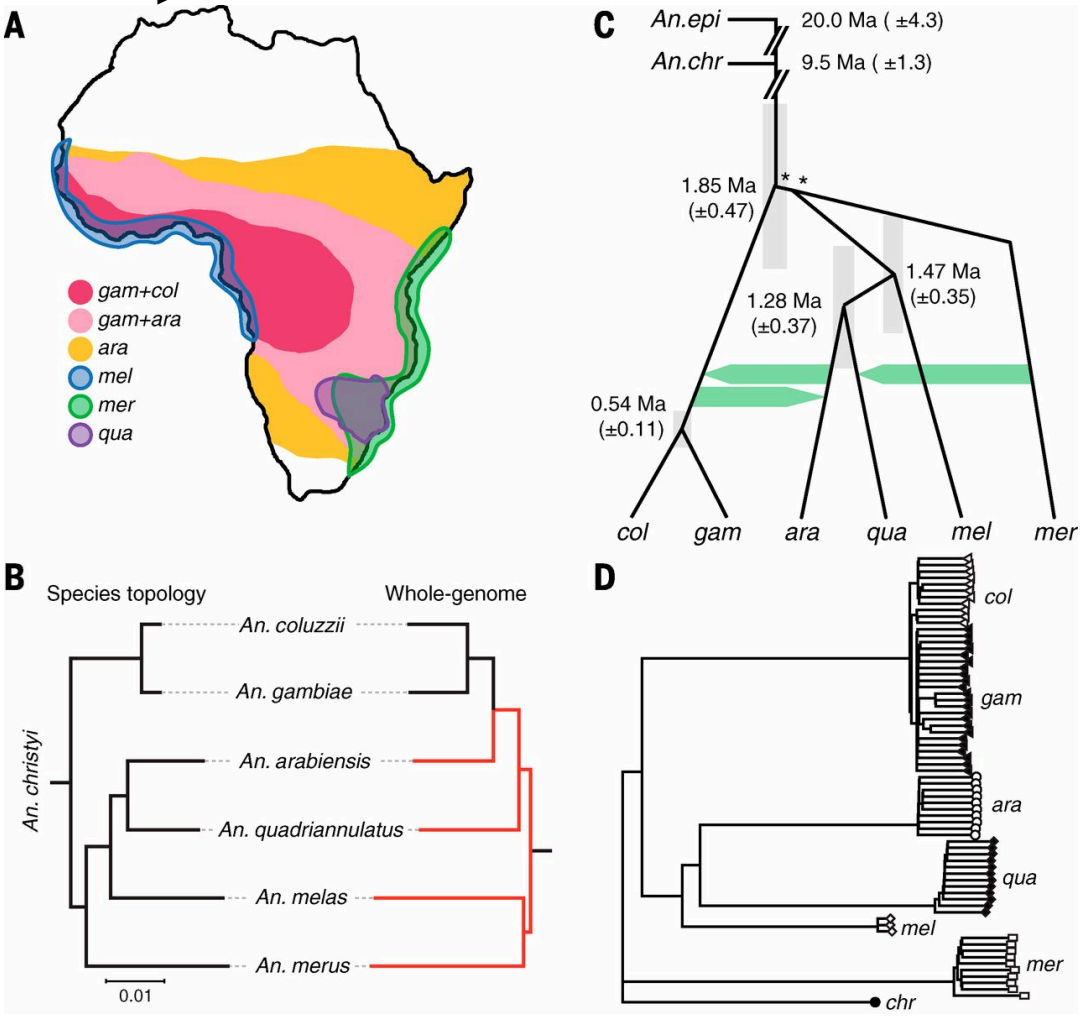
Ronald J. Nowling
Electrical Engineering and Computer Science
Milwaukee School of Engineering
nowling@msoe.edu

Scott Emrich
Electrical Engineering and Computer Science
University of Tennessee Knoxville
semrich@utk.edu

# *An. gambie* and *An. coluzzii*

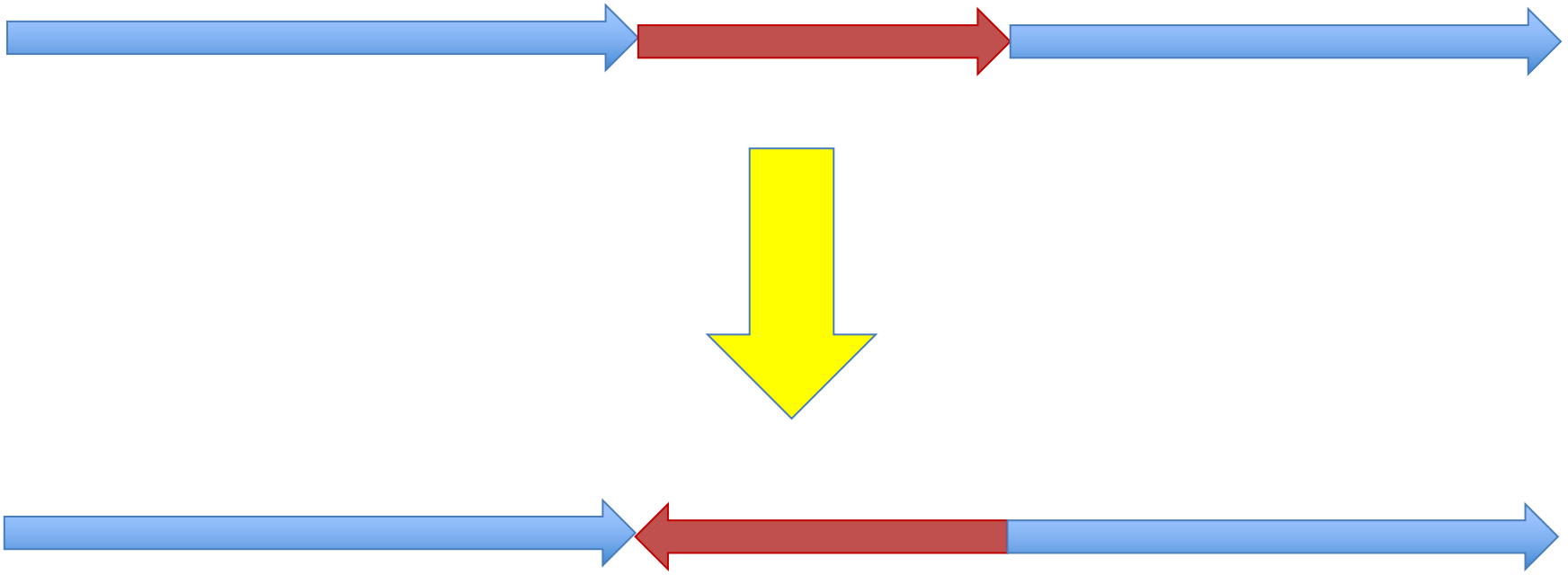
Vectorbase.org / James Gathany

# Population Genetics

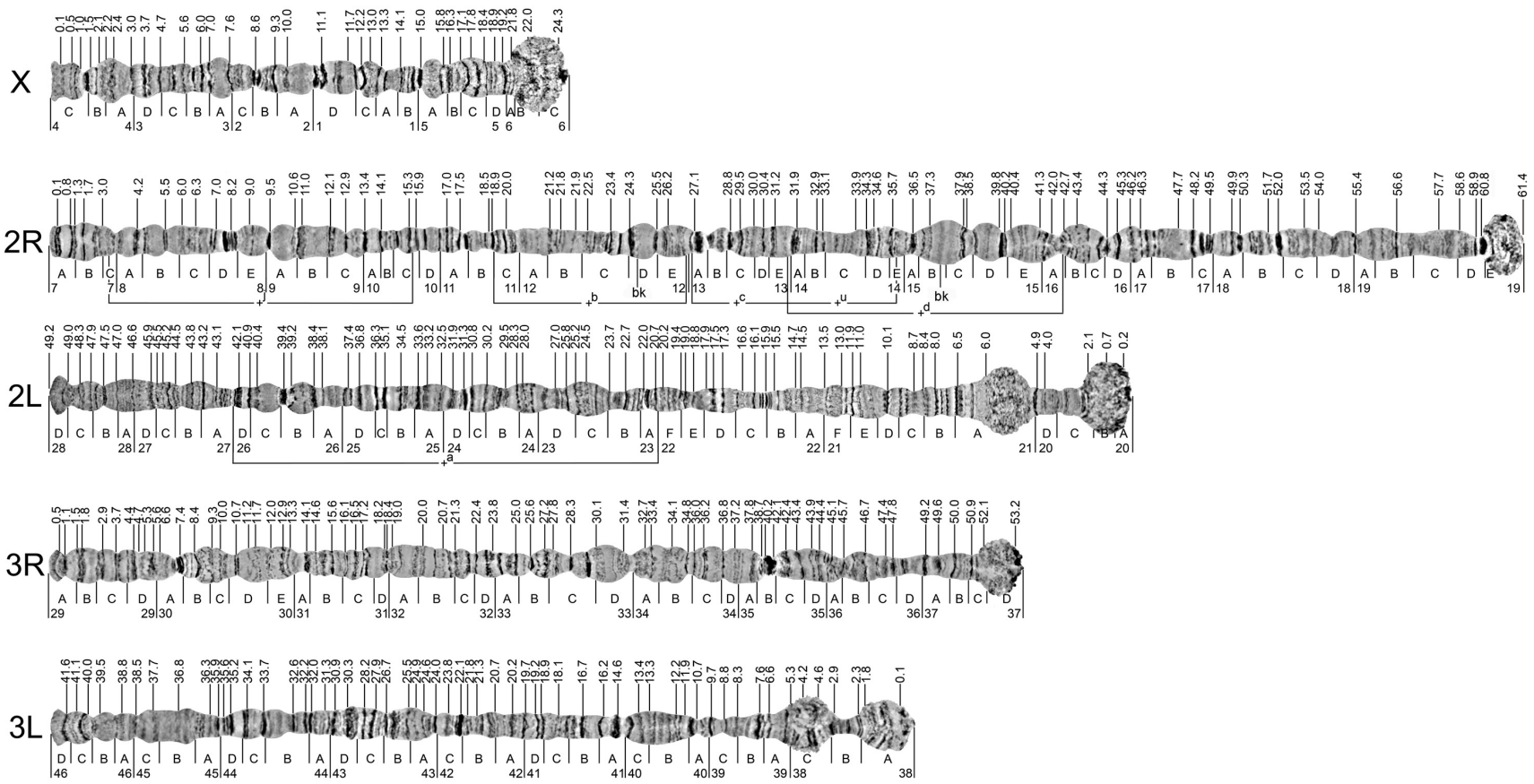



Fontaine, et al. *Science*. 2015.

# What is an Inversion?



# *Anopheles* Cytogenetic Map




George, et al. *Insect Mol Bio*. 2010.

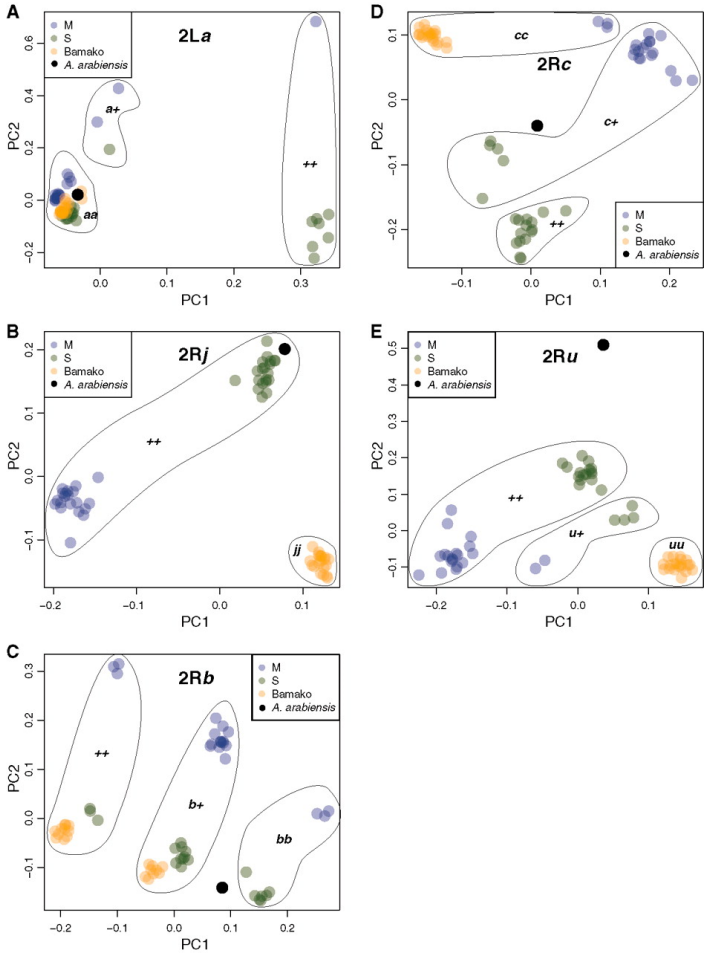
# Importance of Inversions

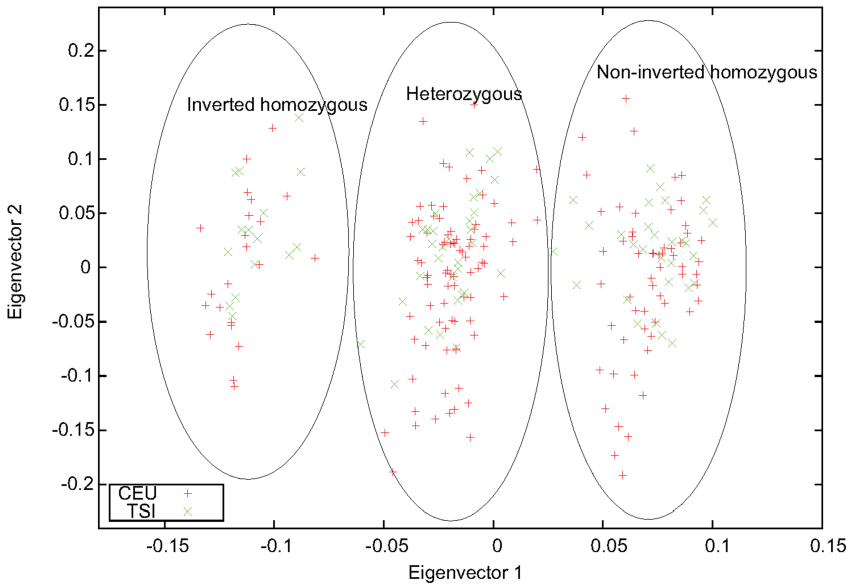
- Thought to play an important role in ecological adaptation by enabling the accumulation of beneficial alleles

  Fuller, et al. Bioarxiv. 2017.; Love, et al. *Mol. Ecol.* 2016.

- 2La in *Anopheles gambiae*
  - Thermal tolerance of larvae
    Rocca, et al. *Malaria Journal*. 2009.
  - Enhanced desiccation resistance
    Gray, et al. *Malaria Journal*. 2009.
  - Susceptibility to malaria parasite species
    Riehle, *Elife*. 2016.

# PCA

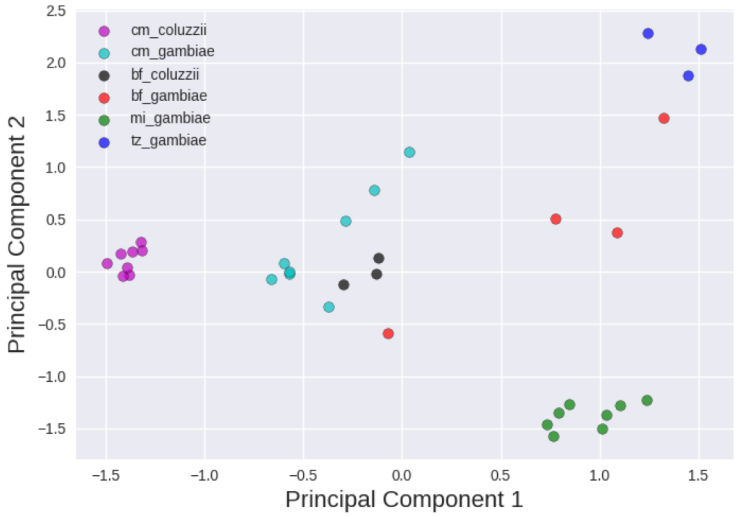


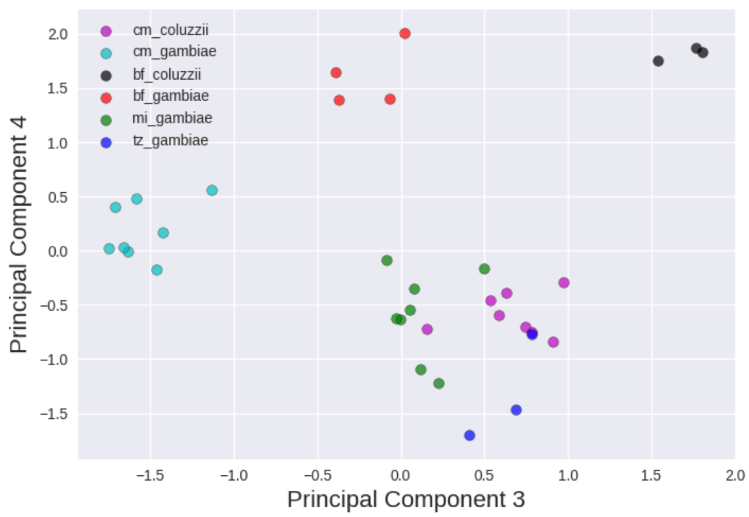




Neafsey, et al. *Science*. 2010.

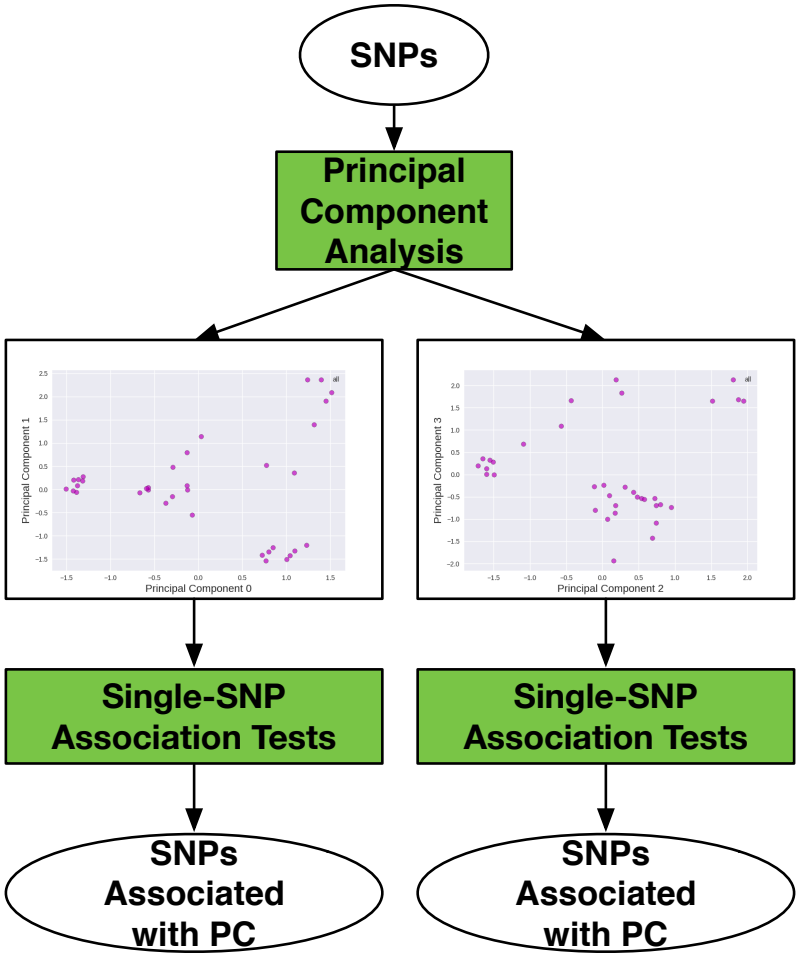
Ma and Amos. *PLOS ONE*. 2012.

# PCA of Our Samples



# Combining PCA and Association Tests



# Likelihood-Ratio Test

1. Fit Logistic Regression model

$$P_g(\mathbf{y}_{i,g}) = \frac{1}{1 + \exp(-\beta_1 \mathbf{T}_{i,j} + \beta_0)} \quad (1)$$

2. Compute Likelihoods for Null and Alternative Models

$$L(\beta, \beta_0 | \mathbf{T}, \mathbf{y}) = \prod_{i=1}^{N} \prod_{g} P(\mathbf{y}_{i,g} | \mathbf{T}_{i,j})^{\mathbf{y}_{i,g}} \quad (2)$$

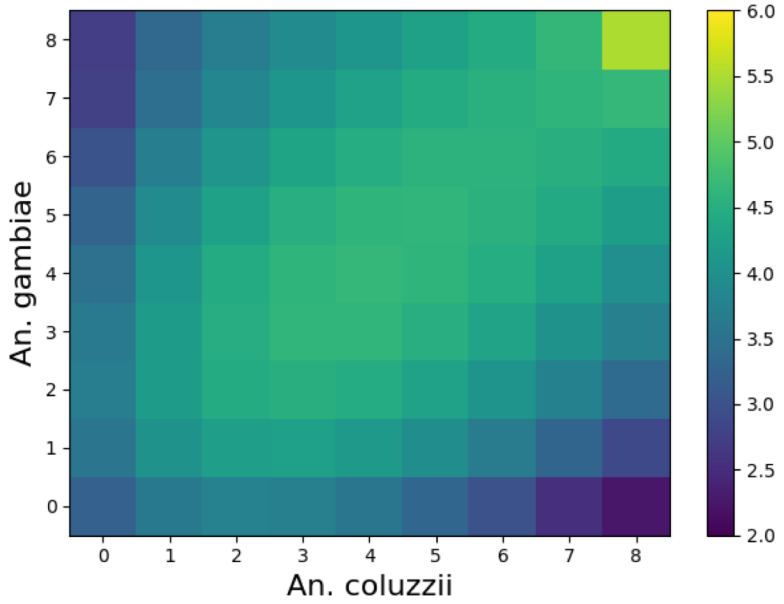
3. Compute G statistic

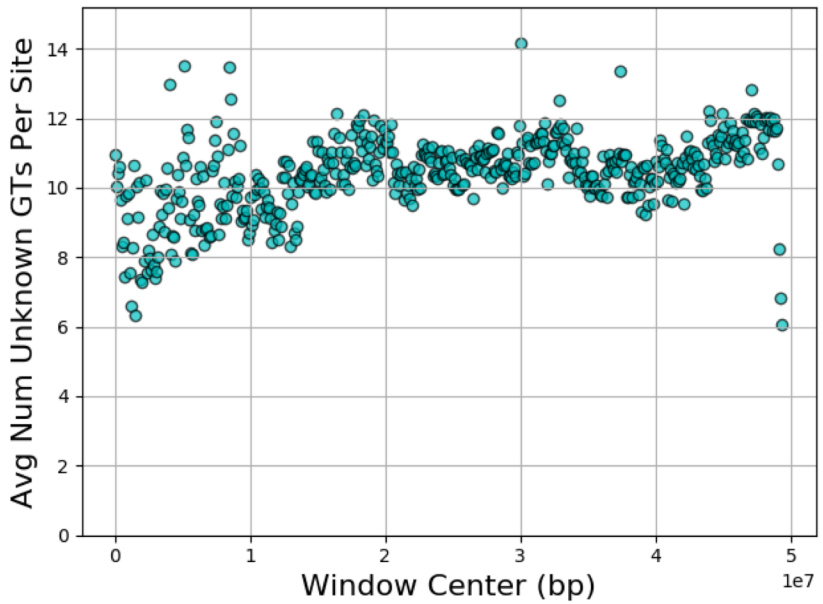
$$G = 2(\log L(\beta_1, \beta_0 | \mathbf{T}, \mathbf{y}) - \log L(\beta_0 | \mathbf{y})) \quad (3)$$

4. Estimate p-value

$$p = P[\chi^2(df) > G] \quad (4)$$

# Unknown Genotypes




Nowling and Emrich. BICOB 2018.

# Handling Unknown Genotypes

| SNP | PC Coordinate |
|---|---|
| A/A | -3.0 |
| T/T | 2.0 |
| X/X | 1.75 |
| A/T | 0.0 |

# Handling Unknown Genotypes

| SNP | PC Coordinate |
|---|---|
| A/A | -3.0 |
| A/A | -3.0 |
| A/A | -3.0 |
| T/T | 2.0 |
| T/T | 2.0 |
| T/T | 2.0 |
| X/X | 1.75 |
| X/X | 1.75 |
| X/X | 1.75 |
| A/T | 0.0 |
| A/T | 0.0 |
| A/T | 0.0 |

# Handling Unknown Genotypes

| SNP | PC Coordinate |
|---|---|
| A/A | -3.0 |
| A/A | -3.0 |
| A/A | -3.0 |
| T/T | 2.0 |
| T/T | 2.0 |
| T/T | 2.0 |
| A/A | 1.75 |
| T/T | 1.75 |
| A/T | 1.75 |
| A/T | 0.0 |
| A/T | 0.0 |
| A/T | 0.0 |

# Likelihood-Ratio Test

1. Fit Logistic Regression model

$$P_g(\mathbf{y}_{i,g}) = \frac{1}{1 + \exp(-\beta_1 \mathbf{T}_{i,j} + \beta_0)} \quad (1)$$

2. Compute Likelihoods for Null and Alternative Models

$$L(\beta, \beta_0 | \mathbf{T}, \mathbf{y}) = \prod_{i=1}^{N} \prod_{g} P(\mathbf{y}_{i,g} | \mathbf{T}_{i,j})^{\mathbf{y}_{i,g}/M} \quad (5)$$

3. Compute G statistic

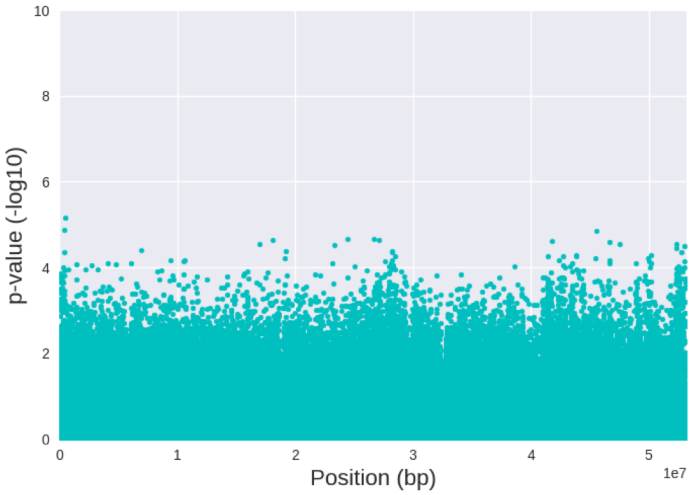
$$G = 2(\log L(\beta_1, \beta_0 | \mathbf{T}, \mathbf{y}) - \log L(\beta_0 | \mathbf{y})) \quad (3)$$

4. Estimate p-value

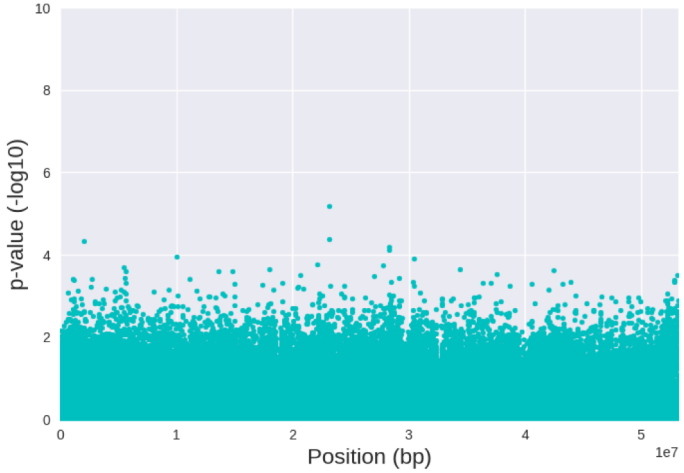
$$p = P[\chi^2(df) > G] \quad (4)$$

# No Inversions

Cameroon
3R-PC1

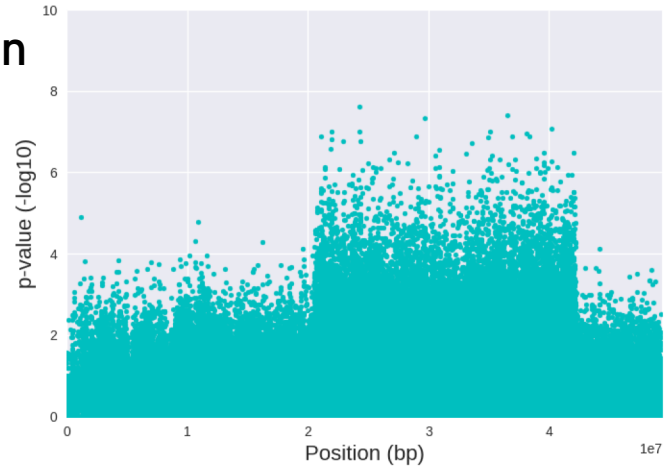



Burkina Faso
3R-PC4

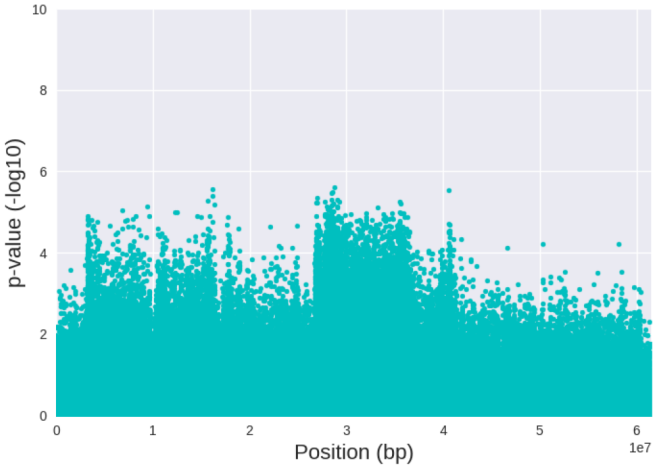


# Detecting Inversions

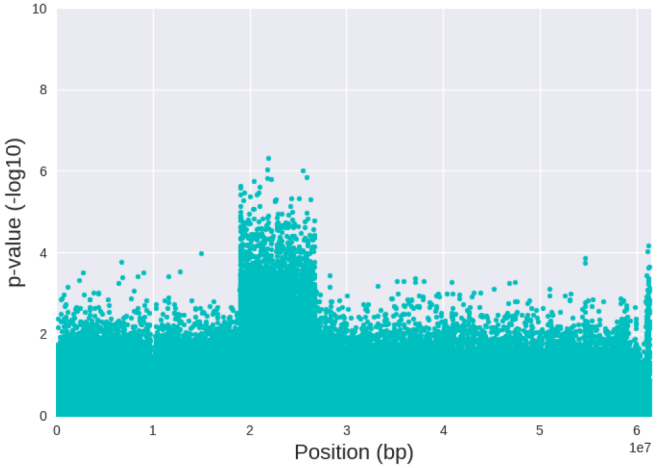
Cameroon
2L-PC1



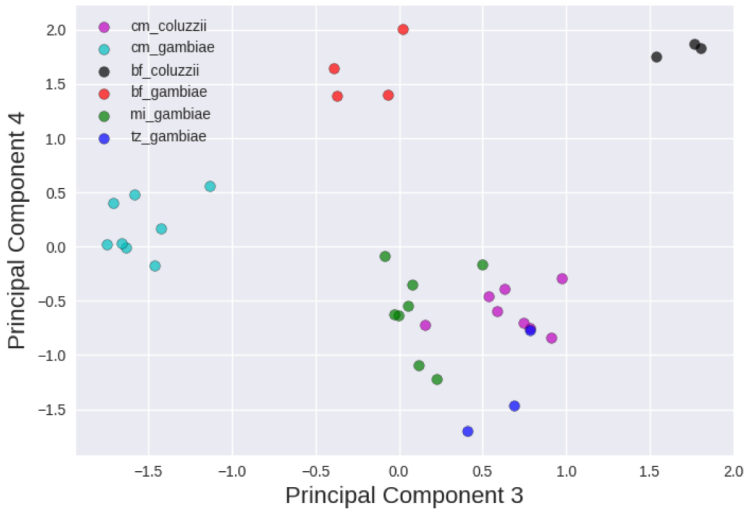

Mali
2R-PC2

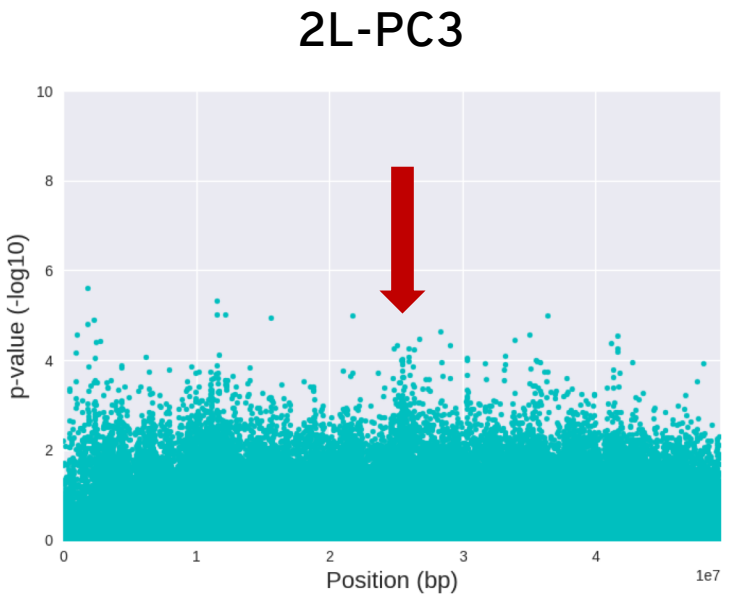



Burkina
Faso
2R-PC4



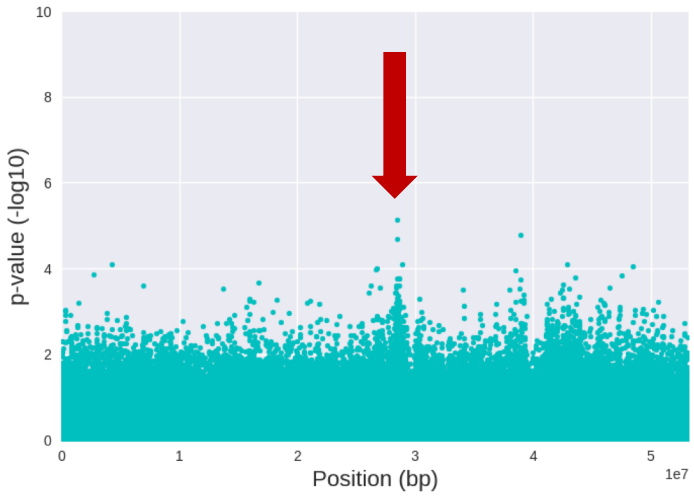
# Resistance to Dieldrin (*Rdl*) SNPs



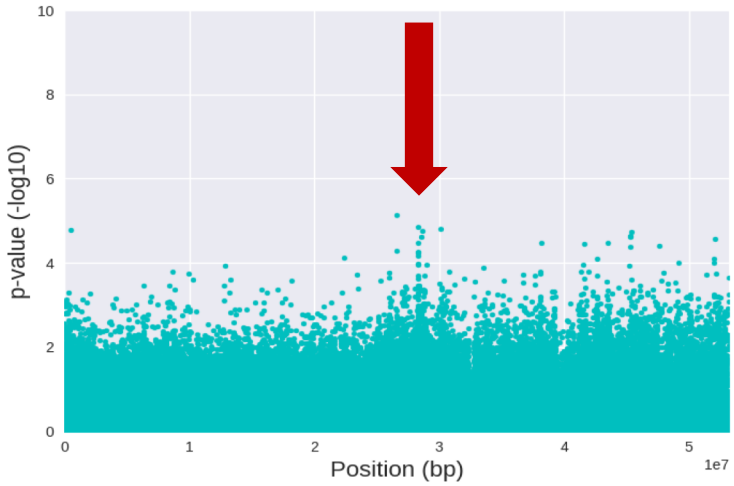



# Sweeps

3R-PC3



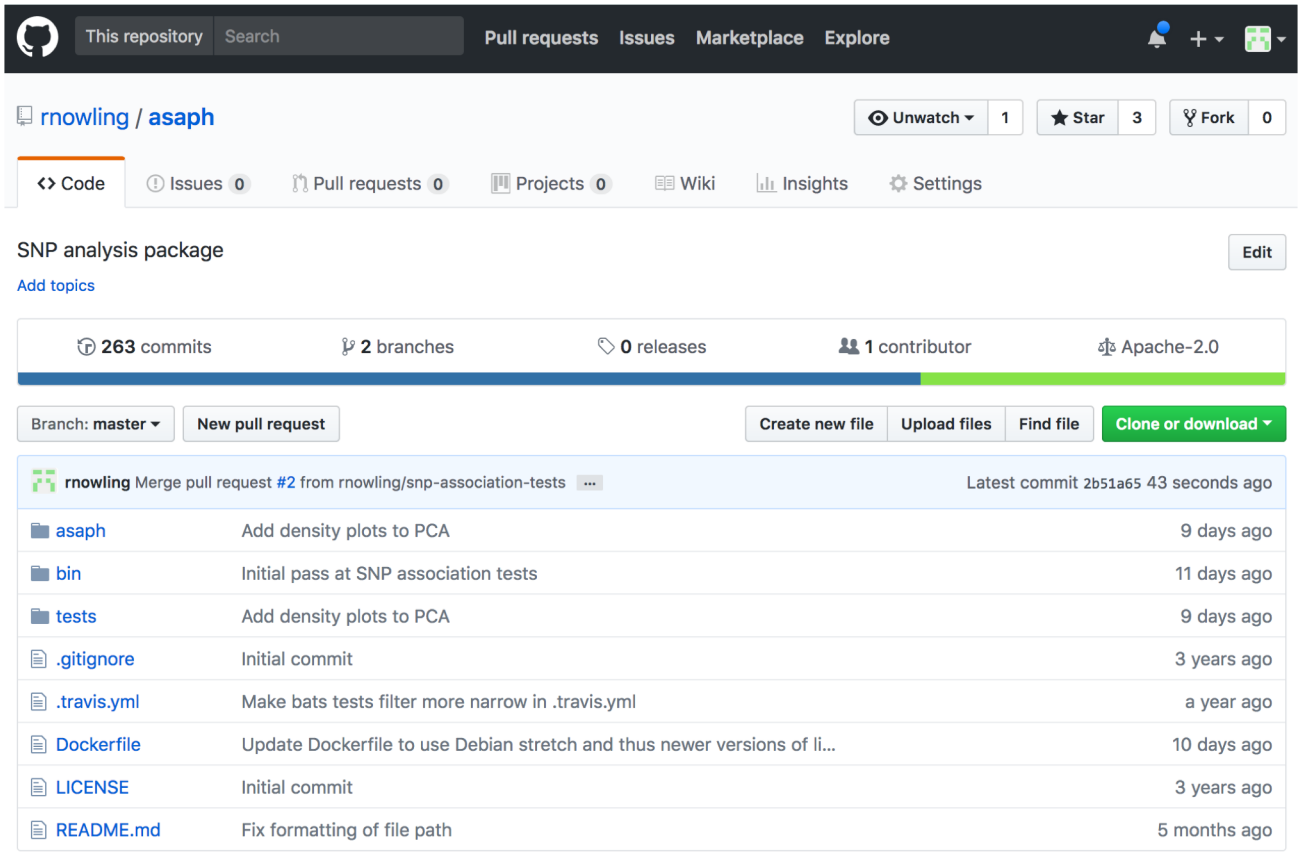

Mali
3R-PC2



# Conclusion

- Inversions of significant biological interest
- PCA of dense SNPs captures inversions but hard to identify from projection plots alone
- Inversion identification enhanced using association tests and Manhattan plots
- Beyond inversions, PCA captures signature of differentiation due to insecticide usage and resistance

# Asaph




https://github.com/rnowling/asaph

# Questions?

- Acknowledgements

  Nora Besansky, Michael Fontaine, Rebecca Love, Aaron Steele, Yue (Shawn) Shen

- Milwaukee School of Engineering is Hiring!

  www.msoe.edu