Population Genetics Analysis without the Population Labels

RJ Nowling[1], JL Abrudan[2], and SJ Emrich[3]

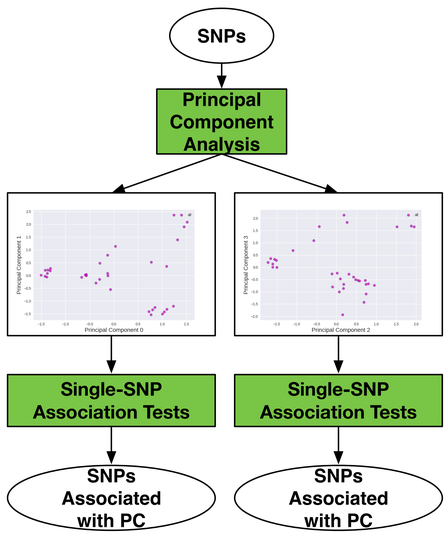
[1]Mathematics, Statistics, and Computer Science, Marquette University [2]Genome Science and Precision Medicine Center, Medical College of Wisconsin [3]Electrical Engineering and Computer Science, University of Tennessee-Knoxville

Problem

Insect population genetics analysis often starts with visualizing relationships between samples using Principal Component Analysis (PCA). Next, populations are identified via clustering (either using uncovered variants or PCA-projected coordinates). And lastly, variants of interest are identified by performing pairwise tests between pairs of populations. In cases where insect populations interbreed, however, it is difficult to clearly define the populations and their memberships and hinder analyses.

Our Method

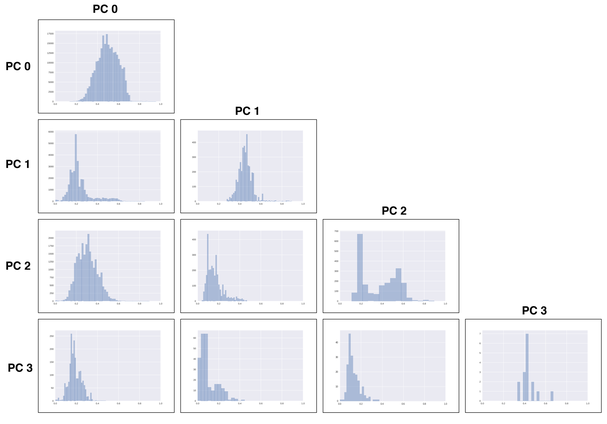
We propose a new approach emphasizing *unsupervised* learning (no population labels). We use PCA to identify the latent variables that explain variation. We then apply single-SNP association tests between the samples' PC coordinates and variant genotypes to identify interesting SNPs.



Our Hypothesis

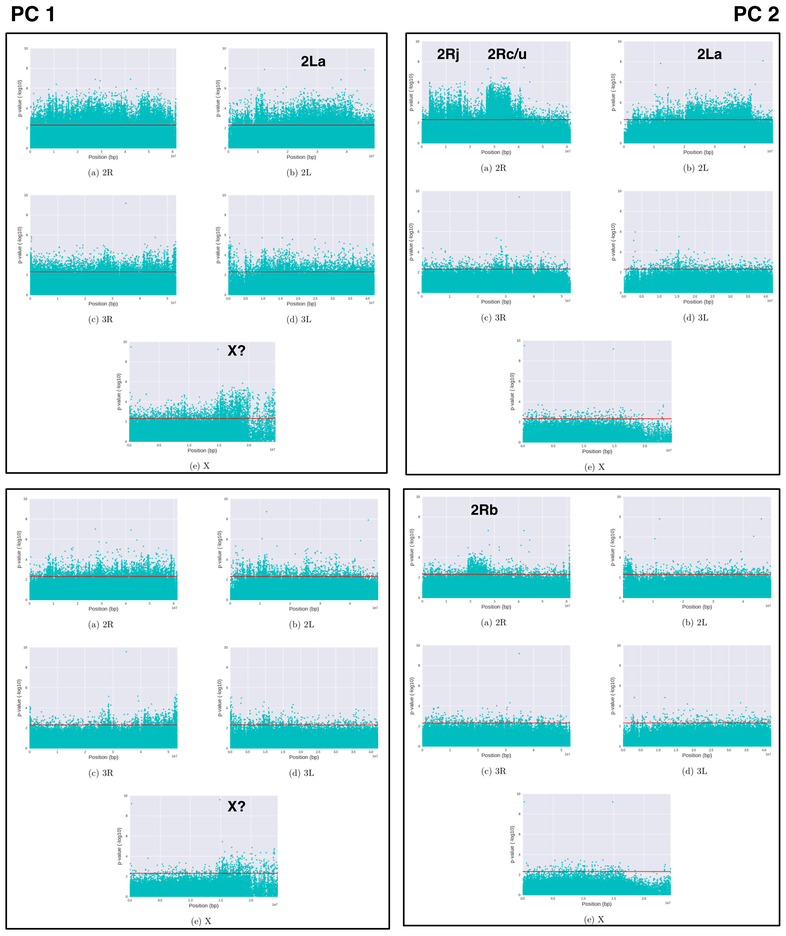
Conceptually, each PC captures a "cluster" of variants whose genotypes are strongly correlated with each other but weakly correlated with other variants. We hypothesize that the variants in each cluster are likely affected by the same, single process, either directly or indirectly. Thus, the PCs capture and isolate the effects of different underlying processes.

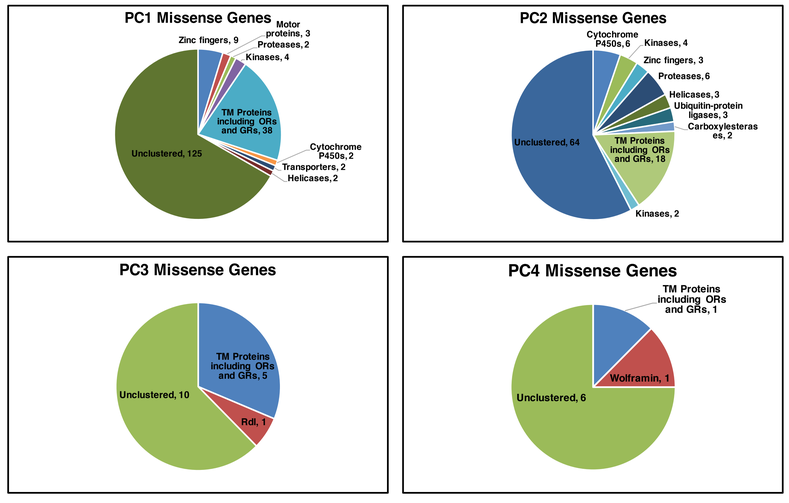
Below are associations calculated between pairs of variants using Cramer's V.



Identifying Inversions in *Anopheles* Samples

We applied our approach to variants from 34 previously-published samples of *Anopheles gambiae* and *Anopheles coluzzii*. We found that 4 PCs explain most of the variation. Plotting the *p*-values of the SNPs' associations with each PC along the chromosome arms, PCs 2 and 4 identified known inversions on 2L and 2R. We note that these samples have not been karyotyped for the 2R inversions.





Genes with Missense Mutations

A number of the variants associated with the PCs cause missense mutations. We used snpEff to annotate the variants in our data set and then identified missense variants with strong associations with each PC. We analyzed the resulting genes with DAVID.

The genes include a broad array of chemosensory receptors, cytochrome P450s, carboxylesterases, and various other enzymes. Mutations in cytochrome P450s and carboxylesterases have previously been associated with insecticide resistance, while chemosensory receptors are directly related to differences in mating, feeding, and habitation behavior.

PC3 had no associations with known geographic or species labels, but insecticide-resistance mutation in *Rdl* (AGAP006028) was strongly associated with PC3. We hypothesize that PC3 captured differentiation due to insecticide usage.

Although PC2 is associated with differences between Mali and other geographic locations, we noted that the association is stronger than the associations between PCs 1 and 4 versus differences in Cameroon versus other locations and Burkina Faso and other locations. Missense variants in the cytochrome P450 *CYP6P3* (AGAP002865), which has been shown to metabolize pyrethroids, were associated with PC2. Given previous work on differentiation in Mali driven by usage of insecticide-treated bed nets, we hypothesize that PC2 is capturing differences between Mali and other geographic locations driven by insecticide usage.

Conclusions

Our method provides a new way to identify the underlying processes driving variation and the affected SNPs. Compared with supervised approaches, our method allows us to identify differentiation even in cases where we do not have labels.

As one application of our method, we demonstrate how it can be used to detect chromosomal inversions. We performed PCA on the SNPs, identified SNPs associated with each PC, and then plotted the *p*-values along the chromosome arms. Along with the well-characterized 2La inversions, we identified poorly-characterized inversions on 2R for which our samples were not karyotyped.

We then evaluated associations between the PCs and our samples' geographic and species labels. PC3 had no associations with our known labels. By analyzing genes containing associated missense mutations, we discovered that one of the *Rdl* insecticide-resistance mutations was associated with PC3. This led us to hypothesize that PC3 captures differences due to insecticide usage.

Even in cases where labels are known, our method still provides valuable insights. PC2 has a strong association with differences between Mali and other locations. *CYP6P3* was one of the genes containing PC-associated missense variants, suggesting that differentiation in Mali may be driven by insecticide usage.

Our method has the potential to uncover previously-missed differentiation and improve our understanding of differentiation in insect vectors.

Acknowledgments

We gratefully acknowledge Nora Besansky, Michael Fontaine, Becca Love, Aaron Steele, and Paul Hickner for thoughtful discussions that provided the motivation for this effort. We would like to thank Yue (Shawn) Shen for help with validating the association tests.