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

Ronald J. Nowling
Marquette University
Milwaukee, Wisconsin
ronald.nowling@marquette.edu

Scott J. Emrich
University of Tennessee
Knoxville, Tennessee
semrich@utk.edu

ABSTRACT

Principal Component Analysis (PCA) of dense single nucleotide polymorphism (SNP) data has wide-ranging applications in populations genetics, including detection of chromosomal inversions. SNPs associated with each PC can be identified through single-SNP association tests performed between SNP genotypes and PC coordinates; this approach has several advantages over thresholding loading factors or sparse PCA methods.

Insect vector SNP data often have a high proportion of unknown (uncalled) genotypes, however, that cannot be reliably imputed and prevent the direct usage of association tests. Building on our previous work, we propose a novel method for adjusting the association tests to handle these unknown genotypes.

We demonstrate the utility of the method through two applications: detecting chromosomal inversions and characterizing differentiation processed captured by PCA. When applied to SNP data from the 2L and 2R chromosome arms of 34 karyotyped *Anopheles gambiae* and *Anopheles coluzzii* mosquitoes, our method clearly identifies the 2La, 2Rb, 2Rc, 2Rj, and 2Ru inversions. Using our method to identify SNP associated with 2L-PC3, we observed one of the two insecticide-resistance variants in the *Rdl* gene; our results suggests that the PC is capturing differentiation driven by insecticide usage.

CCS CONCEPTS

• **Computing methodologies** → **Principal component analysis**; • **Applied computing** → **Molecular evolution**; **Population genetics**; *Computational genomics*;

KEYWORDS

Insect vectors, population genetics, association tests

ACM Reference Format:
Ronald J. Nowling and Scott J. Emrich. 2018. Detecting Chromosomal Inversions from Dense SNPs by Combining PCA and Association Tests. In *ACM-BCB'18: 9th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, August 29-September 1, 2018, Washington, DC, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3233547.3233571

ACM-BCB'18, August 29-September 1, 2018, Washington, DC, USA

ACM ISBN 978-1-4503-5794-4/18/08…$15.00
https://doi.org/10.1145/3233547.3233571

1 INTRODUCTION

Principal Component Analysis (PCA) of SNPs is widely used in population genetics for visualizing the relationships between samples [20], correcting for stratification in genome-wide association studies [24], and with clustering to determine population structure [14]. While PCA is often treated as a one-way transformation, it is possible and sometimes useful to identify the SNPs associated with each principal component (PC). The key insight is that each PC clusters a set of features that are strongly correlated with one another but weakly correlated with features outside the set. For example, Paschou, et al. [22] demonstrated that PC-SNP correlations can be used to identify a small set of SNPs that can be used as effective markers to associate individuals with populations.

Although selecting variables by thresholding weights based on magnitude is frequently used in practice (e.g., [22, 27]), it can lead to misleading results [3]; loading factors vary based on the number and distribution of features since all PC vectors are normalized. Sparse PCA techniques [36] have been proposed that employ regularization techniques such as *lasso* [31] and *elasticnet* [35]. Such regularization techniques, by design, choose a subset of the features, preventing recovery of *all* of the variables associated with a given PC.

Single-SNP association tests, on the other hand, provide p-values that estimate significance consistently, irrespective of the number or distribution of features. Further, these tests recover all of the variables and without bias resulting from correlation with other SNPs. In this setup, tests are performed between each SNP and each PC coordinate. The genotype is used as the outcome variable, while the PC coordinate is used as the predictor.

In our previous work [21], we noted that insect vector data sets tend to have a large number of unknown (uncalled) genotypes and very small sample sizes. Unlike in human data sets, these unknown genotypes cannot be reliably imputed. Since the genotypes are used as the labels (outcome variables) here, unknown genotypes must be handled in order to perform association tests.

Previously, we proposed an adjusted likelihood-ratio test that uses an uninformative (uniform) prior over the unknown genotypes. When compared with F_{ST} or a standard likelihood-ratio test, our adjusted test is significantly better at avoiding the large number of false positives. In that case, we were able to up-sample the training set, while using the original data for the likelihood evaluations. Here, we need a different strategy: up-sample the data set, impute the unknown genotypes in a one-to-one relationship to the sample copies, and then re-weight the samples in the likelihood evaluation so that the estimated p-values are consistent with the original number of samples. Our method is implemented and made available through our open-source variant analysis toolkit Asaph.

To demonstrate the utility of this approach, we apply the method to identifying chromosomal inversions and characterizing differentiation processes in the malaria vectors *Anopheles gambiae* and *Anopheles coluzzii*. Chromosomal inversions are thought to play an important role in ecological adaptation by enabling the accumulation of beneficial alleles [8, 15]. For example, the 2La inversion in various species of the *Anopheles gambiae* complex has been associated with thermal tolerance of larvae [26], enhanced desiccation resistance in adult mosquitoes [9], and susceptibility to at least one species (*Plasmodium falciparum*) of malaria [25].

Given long-standing research on *Anopheles gambiae* inversions, the 34 *Anopheles gambiae* and *Anopheles coluzzii* samples from [7] samples are karyotyped for the 2La inversion, which has the strongest PCA-determined signal [33], but not for known inversions on 2R (see [18]). Chromosome 3 has no known inversions in these species or other strong population-linked signals and can serve as a negative control. These characteristics and ongoing large-scale sequencing efforts across Africa make such SNP data ideal for developing finer-grained prediction.

Although a number of computational methods can uncover large-scale inversion breakpoints using either paired-end sequence data or long-range reads/contacts (e.g., [4, 5, 10, 28, 30, 34]) the most confident inversion detection is still performed using experimental karyotying. We focus on detecting potentially adaptive inversions using only dense single-nucleotide polymorphism (SNP) data. Previous methods rely on the calculation of linkage disequilibrium (LD) in windows to identify blocks of SNPs likely to belong to an inversion [1, 2, 29].

Alternatively, Ma, et al. observed that chromosomal inversions cause a "three-stripe pattern" to appear in PCA of SNPs [16, 17]. When applied to our data (see Section 3.1), however, the PCA projection plots alone are ambiguous. We demonstrate, using our method, that identifying the SNPs associated with the PCs enables more reliable detection of inversions. In Section 3.2, we validate that our method can successfully identify the SNPs associated with each PC using simulated data. When the p-values of the extracted SNPs are plotted along the chromosome arms, chromosomal inversions are clearly and unambiguously revealed. In Section 3.3, we confirm that our method is able to detect the 2La inversion and does not generate false positives on the 3 and X chromosomes using the 34 *Anopheles* samples. We then applied our the method to detecting poorly-characterized inversions on the 2R chromosome arm in Section 3.4.

Lastly, in Section 3.5, we applied our method to characterizing the differentiation captured by PCA through the associated SNPs. We tested associations between each PC and the species and geographic location labels for our samples. Differences captured by 2L-PC3 were not associated with any of our labels. We used our method to identify the SNPs associated with the PC. We observed that one of the SNPs is an insecticide-resistance mutation in the *Rdl* gene, suggesting that the PC captures differences driven by insecticide usage.

2 METHODS

2.1 Data Sets

For our validation on simulated data, We simulated 300 biallelic SNPs across a single artificial "chromosome" for 100 individuals to represent different karyotypes. The individuals were divided into four groups. By default, all of the SNPs chosen to be uncorrelated with populations labels and frequencies of 45% homozygous A, 45% homozygous T, and 10% heterozygous. For group 1, all of the the SNPs were uncorrelated. For group 2, SNPs 101 - 200 had frequencies of 95% homozygous A, 2.5% homozygous T, and 2.5% heterozygous. For group 4, SNPs 201-300 had frequencies of 90% homozygous A, 5% homozygous T, and 5% heterozygous. Lastly, SNPS 101 - 300 for group 3 had the frequencies of both groups 2 and 4.

For the analysis on real data, we used biallelic SNPs on the 2 (347,510 positions for 2L, 394,487 positions for 2R), 3 (346,391 positions on 3R), and X (72,003) chromosomes from 34 *Anopheles gambiae* and *Anopheles coluzzii* samples from [7]. Details on the sequencing and variant calling (including filtering) are given in [7].

2.2 Encoding SNP Genotypes

Assume that we have N samples with V positions with biallelic variants. Each position has a reference allele and an alternative allele, and at each position, each sample has one of three genotypes (homozygous reference, homozygous alternate, or heterogyzous).

We encode the variants as a feature matrix $\mathbf{X}$ with dimensions $N \times 3V$. If sample i has the homozygous reference genotype at position k, then we set $\mathbf{X}_{i,3k+1} = 1$. If sample i has the homozygous alternate genotype at position k, then we set $\mathbf{X}_{i,3k+2} = 1$. If sample i has the heterozygous genotype at position k, then we set $\mathbf{X}_{i,3k+3} = 1$. If the genotype of sample i is unknown at position k, then we do nothing.

2.3 Principal Component Analysis

Principal component analysis (PCA) of the feature matrix $\mathbf{X}$ produces a $3V \times P$ matrix $\mathbf{W}$ of principal components and a $N \times P$ matrix $\mathbf{T}$ of projected coordinates for the samples such that:

$$\mathbf{T} = \mathbf{X}\mathbf{W}$$

As directly computing PCA would involve computing a $3V \times 3V$ co-variance matrix, we used a randomized truncated SVD implementation from Scikit Learn [23]. Whitening was applied to the resulting PCs.

2.4 Single-SNP Association Tests

A single association test is performed for each combination of principal component (PC) j and SNP position k. Each sample i can have one of three possible genotypes for a given position k. We use a Logistic Regression model for each possible genotype such that each LR model predicts the probability that a sample i has a given genotype g; in other words, we use a one-versus-all scheme to implement multinomial Logistic Regression. The Logistic Regression models are written as [11]:

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

where $\mathbf{y}_{i,g}$ is binary indicator as to whether sample i has genotype g, $\mathbf{T}_{i,j}$ is the coordinate for a single sample i along PC j, β_1 is the weight, and β_0 is the intercept. We train the models using Stochastic Gradient Descent (SGD) and an L_2 penalty. (For the experiments in this paper, we performed 10,000 epochs of training for each model.)

The likelihood for the multinomial Logistic Regression model is given by [11]:

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

To perform the Likelihood-Ratio Test, two sets of Logistic Regression models are trained in total. The models in the alternative set contain additional independent variables (features) not in the null model. In our case, the set of null models only contain the intercept and thus, predicts the conditional class probabilities using the ratio of one class of samples to all samples. The weights (β_1, β_0) from the two models are used to compute the log likelihoods. The difference G between the two is calculated by:

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

The p-value for the difference in log likelihoods is calculated using the χ^2 distribution:

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

where df is the difference in the number of degrees of freedom (weights) between the two models.

2.5 Adjusted-Likelihood Ratio Test for Unknown Genotypes

It is common for genotypes to be unknown (uncalled). In our previous work [21], we proposed an adjusted likelihood-ratio test that assumes that unknown genotypes are distributed according to an uninformative (uniform) prior to avoid learning on the missing data. In that case, we were able to adjust the training set but use the original set of samples for the likelihood function calculation. While we previously used the genotype as the predictor and the population labels as the outcome variable, we use the genotype as the outcome variable and the PC coordinate as the predictor here, necessitating a different strategy.

To handle the unknown genotypes, we chose to deterministically upsample the samples, impute unknown genotypes, and then reweight the samples in the likelihood. In particular, if we have M genotypes, we create M copies of each sample. (In our case, $M = 3$ since we are working with biallelic SNPs with three genotypes.) If the genotype is known, the copies have the same genotype as the original. Otherwise, we make the conservative assumption that there is an uninformative (uniform) prior over the genotypes and impute the copies so that there is a one-to-one relationship between the copies and possible genotypes.

Since we increased the number of samples, we need to weight the samples so that the calculated p-values are consistent with the original number of samples. The modified likelihood function is then:

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

2.6 Asaph: an Open-Source Toolkit for Variant Analysis

We implemented our method in Asaph, our open-source toolkit for variant analysis. Asaph is implemented in Python using Numpy / Scipy [32], Matlotlib [12], and Scikit-Learn [23] and is available at https://github.com/rnowling/asaph under the Apache Public License v2.

3 RESULTS

3.1 PCA Projections Insufficient for Detecting Inversions

We performed PCA of the SNPs on the 2L chromosome arm of 34 *Anopheles gambiae* and *Anopheles coluzzii* samples from Burkina Faso, Cameroon, Mali, and Tanzania. The 2L chromosome arm contains a large inversion (2La), and these samples have been previously karyotyped for this inversion. Four PCs explained most of the variation on this arm. We next tested the obtained PCs against the known 2La karyotype labels (see Table 1). PC1 had a statistically-significant association with a p-value of 6.07×10^{-11}, while the three other PCs had p-values around or greater than 1.00×10^{-3}. We thus expect that PC1 captures most of the 2La inversion signal in these data. Based on the work of [16], we would expect to see a "three-stripe" pattern indicative of an inversion in the PCA projections for 2L. Instead, we found that the PCA projections were ambiguous (see Figure 1a), illustrating that it can be difficult to identify inversions from PCA projections alone.

We then performed PCA of SNPs on the 2R chromosome arm of the same 34 *Anopheles gambiae* and *Anopheles coluzzii* samples, which can contain up to four four smaller inversions: 2Rb, 2Rc, 2Rj, and 2Ru. As with arm 2L, four PCs explained most of the variation, and these PCA projections were also ambiguous with respect to a "three-stripe pattern" (see Figures 1c and 1d).

3.2 Validation on Simulated Data

To validate our method, we applied it to well-characterized simulated data (see Section 2.1). We simulated two orthogonal processes driving differentiation in 200 out of 300 variants along an artificial chromosome. The resulting individuals formed four clusters.

PCA analysis on these simulated SNPs indicated that the explained variance ratios of the first two PCs explained most of the variation (see Figure 2a) and that the four groups were clearly clustered in the PCA projection (see Figure 2b). When we performed association tests for PCs 1 and 2 and plotted the p-values along the artificial chromosome (see Figures 2c and 2d), PC1 captured variations in positions 101 to 200, while PC2 captured variations in positions 201 to 300, as expected.

3.3 Validation on Chromosome Arms with Known Inversion Karyotypes

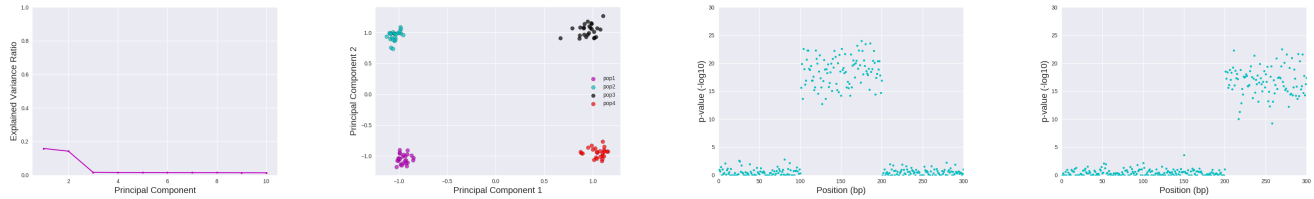
We applied our method to SNPs on the 2L chromosome arm of 34 *Anopheles gambiae* and *Anopheles coluzzii* samples from Burkina Faso, Cameroon, Mali, and Tanzania. Based on the association tests versus the known karyotype labels in Section 3.1, we expected a single strong PC underlying this inversion. For each of the four previously uncovered PCs (see above), we plotted the p-values of

Table 1: Association Tests Between Principal Components and 2La Inversion Karyotypes, Species, and Geographic Locations of Samples. We report *p*-values and accuracies.

	2La	Species	Burkina Faso	Cameroon	Mali
2L-PC1	2.09×10^{-14} / 100.0%	1.29×10^{-2} / 79.4%	3.20×10^{-2} / 79.4%	1.52×10^{-8} / 94.1%	5.42×10^{-9} / 97.1%
2L-PC2	1.19×10^{-3} / 52.9%	4.86×10^{-9} / 91.2%	6.59×10^{-1} / 79.4%	4.16×10^{-2} / 52.9%	3.29×10^{-1} / 76.5%
2L-PC3	1.48×10^{-1} / 44.1%	3.24×10^{-4} / 76.5%	4.17×10^{-1} / 79.4%	8.31×10^{-2} / 70.1%	6.76×10^{-1} / 76.5%
2L-PC4	4.49×10^{-2} / 52.9%	2.63×10^{-1} / 61.8%	5.86×10^{-8} / 94.1%	1.00×10^{0} / 52.9%	1.38×10^{-4} / 88.2%
2R-PC1		8.23×10^{-11} / 100.0%	8.65×10^{-1} / 79.4%	1.59×10^{-6} / 85.3%	1.77×10^{-9} / 100.0%
2R-PC2		3.41×10^{-1} / 67.6%	2.27×10^{-2} / 73.5%	4.60×10^{-1} / 29.4%	1.39×10^{-9} / 100.0%
2R-PC3		3.05×10^{-4} / 73.5%	4.58×10^{-1} / 79.4%	6.96×10^{-3} / 67.6%	2.56×10^{-1} / 76.5%
2R-PC4		5.67×10^{-1} / 67.6%	3.12×10^{-7} / 94.1%	4.08×10^{-2} / 73.5%	1.00×10^{0} / 76.5%
3R-PC1		1.17×10^{-5} / 79.4%	2.53×10^{-1} / 79.4%	6.47×10^{-8} / 88.2%	1.58×10^{-1} / 73.5%
3R-PC2		2.19×10^{-4} / 76.5%	9.70×10^{-2} / 79.4%	2.59×10^{-1} / 55.9%	1.41×10^{-9} / 100.0%
3R-PC3		2.36×10^{-1} / 64.7%	7.92×10^{-3} / 85.3%	2.37×10^{-1} / 64.7%	2.25×10^{-3} / 76.5%
3R-PC4		5.67×10^{-3} / 67.6%	1.08×10^{-4} / 91.2%	4.13×10^{-3} / 70.6%	1.68×10^{-1} / 73.5%
X-PC1		9.12×10^{-7} / 88.2%	4.77×10^{-1} / 79.4%	9.19×10^{-8} / 79.4%	3.14×10^{-2} / 67.6%
X-PC2		1.22×10^{-3} / 79.4%	2.07×10^{-1} / 79.4%	5.69×10^{-1} / 35.9%	3.65×10^{-9} / 100.0%
X-PC3		6.07×10^{-3} / 64.7%	4.08×10^{-1} / 79.4%	4.59×10^{-2} / 64.7%	1.60×10^{-4} / 82.4%
X-PC4		2.58×10^{-1} / 76.5%	1.69×10^{-7} / 97.1%	1.11×10^{-2} / 79.4%	8.43×10^{-1} / 67.5%

(a) 2L, PCs 1 and 2 **(b) 2L, PCs 3 and 4** **(c) 2R, PCs 1 and 2** **(d) 2R, PCs 3 and 4**

Figure 1: PCA Projections of SNPs on the 2L and 2R Chromosome Arms of 34 *Anopheles* samples

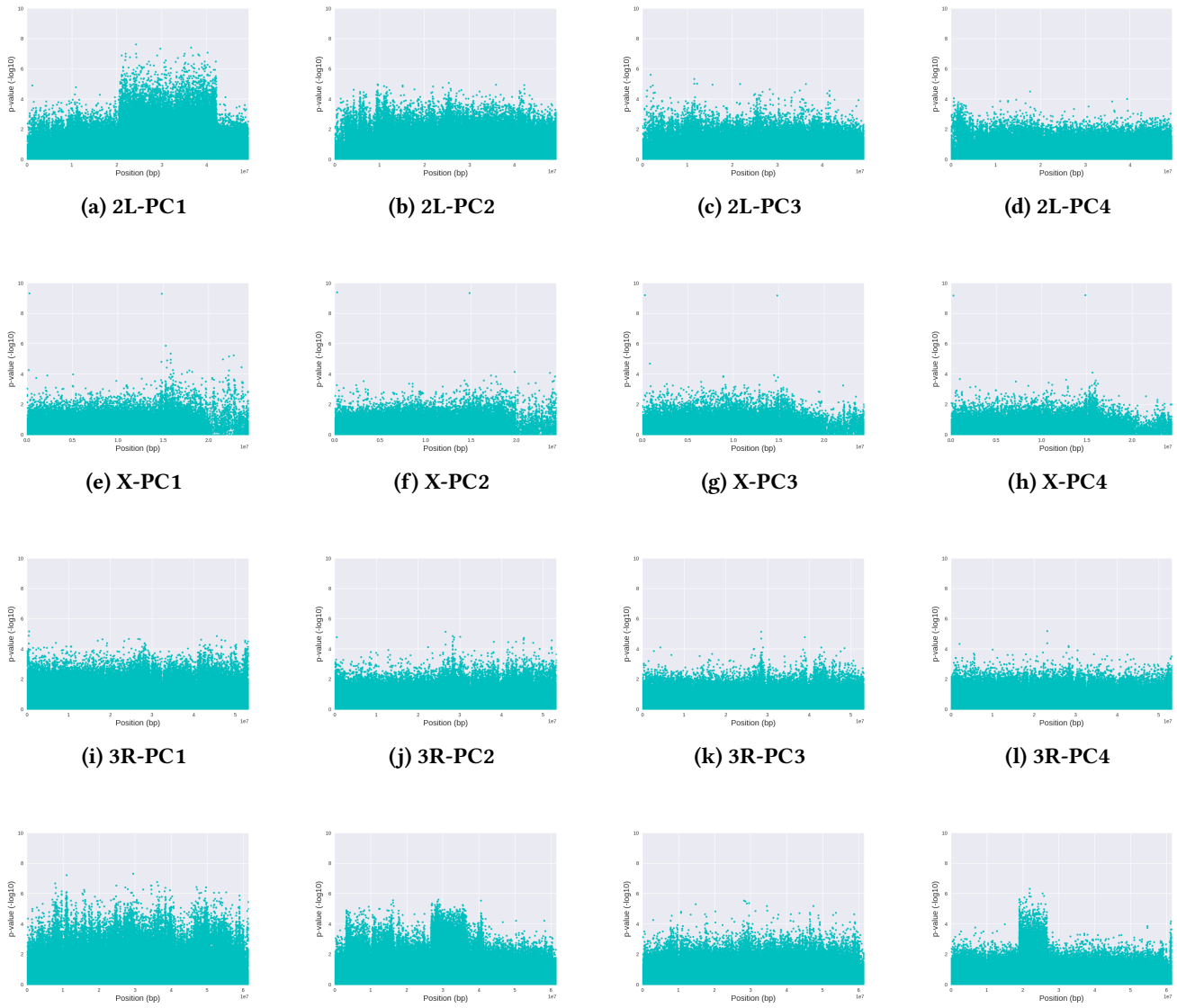


(a) Explained Variance Ratios **(b) PCA Projections** **(c) PC1** **(d) PC2**

Figure 2: Simulated Data for 100 Individuals Divided Among Four Groups

each SNP along the chromosome arms (see Figure 3a–d). The plots illustrated that, as expected, SNPs located in the 2La inversion region (20 – 42.5 Mbp) were strongly associated with PC1 but not the other PCs. For example, SNPs above a conservative cut off of 6 ($-\log_{10}$ of the probability) are located only in the known 2La interval. SNPs outside of that region were not as strongly associated with PC1. Our method is thus able to identify the 2La inversion without explicit karyotype information by exploiting the characteristics of PC1 of these SNP data.

To validate that our method does not produce false positives, we applied our method to SNPs on the X and 3R chromosome arms, which are not known to contain any inversions (see [7]). As expected, no inversions or inversion-like patterns were observed when the SNPs associated with the first four PCs were plotted along the chromosome arms (see Figures 3e–h and 3i–l)).



(a) 2L-PC1 (b) 2L-PC2 (c) 2L-PC3 (d) 2L-PC4

(e) X-PC1 (f) X-PC2 (g) X-PC3 (h) X-PC4

(i) 3R-PC1 (j) 3R-PC2 (k) 3R-PC3 (l) 3R-PC4

(m) 2R-PC1 (n) 2R-PC2 (o) 2R-PC3 (p) 2R-PC4

Figure 3: *p*-values of SNPs on the 2L (a–d), X (e–h), 3R (i–l), and 2R (m–p) Chromosome Arms of 34 *Anopheles* samples

3.4 Identifying 2R Inversions in Samples with Unknown Karyotypes

We applied our method to SNPs on the 2R chromosome arm of the same 34 *Anopheles gambiae* and *Anopheles coluzzii* samples. The 2R chromosome arm contains four smaller inversions (2Rb, 2Rc, 2Rj, and 2Ru). These samples have not been karyotyped for the 2R inversions and both the divergence and segregation differences are not as great as the 2La inversion for complex biological reasons (see [7]).

Even so, PC2 captured the 2Rc, 2Rj, and 2Ru inversions with relatively higher significance on this arm (but lower than 2La; see Figure 3n), while PC4 captured the 2Rb inversion (see Figure 3p). It appears that the other PCs did not show any evidence of the inversions (see Figures 3m and 3o).

3.5 Characterizing PCs with No Associations to Known Labels

Beyond inversions, we applied our method to characterizing the differentiation captured by PCA from the associated SNPs. We began by testing associations between the PCs, species, and geographic locations (Burkina Faso, Cameroon, and Mali) given their importance in prior analyses [15].

2L-PC1 had some association with Cameroon versus non-Cameroon samples, while 2L-PC1, 2R-PCs 1 and 2, 3R-PC2, and X-PC2 were all

strongly associated with Mali versus non-Mali samples. Likewise, 2L-PC4, 2R-PC4, and X-PC4 were strongly associated with Burkina versus non-Burkina samples, while 3R-PC4 was moderately associated. 2R-PC1 was strongly associated with differences between the species, while 2L-PC2 and X-PC1 were moderately associated.

Our independent chromosome arm observations are consistent between arms as well as prior work. For example, the first two PCs were often associated with Mali, while the fourth PC was usually associated with Burkina Faso. Although the class imbalance (16 Cameroon, 8 Mali, 7 Burkina Faso, and 3 Tanzania samples) in the data may be affecting the amount of variance explained by each PC, and thus their orderings, this result is also consistent with the large amount of standing variation and limited gene flow observed previously [19].

Notably, the third PC on all arms was only weakly associated with the tested labels, suggesting an alternative process is driving these underlying differences between samples. We used our method to identify SNPs associated with the PCs on 2L. The A296G insecticide-resistance variant ([6, 13], located at position 25,429,236 on 2L) in the *Rdl* gene (AGAP006028) was associated with 2L-PC3 (p-value of 8.24×10^{-5} versus a Bonferroni-corrected significance level of $0.01/(34-1) = 3.03 \times 10^{-4}$) but not the other PCs. Insecticides have previously been shown to be a strong force driving differentiation [7, 18]. We therefore hypothesize that at least 2L-PC3, but possibly all of the third PCs of the chromosome arms, is (are) associated with differences due to exposure and resistance to the insecticide dieldrin.

4 DISCUSSION AND CONCLUSION

Principal Component Analysis (PCA) of SNPs is widely used in population genetics for visualizing the relationships between samples [20], correcting for stratification in genome-wide association studies [24], and with clustering to determine population structure [14]. We proposed a novel method for identifying SNPs associated with each PC using single-SNP association tests. Previously, association tests would have been problematic due to unknown (uncalled) genotypes; we workaround this limitation by introducing an adjusted likelihood-ratio test.

To demonstrate the utility of this method, we first applied it to detecting chromosomal inversions from dense SNP data. Chromosomal inversions have been found to play a significant role in the adaptation of *Anopheles* species. Although previous work [16] found that inversions cause a distinctive "three-stripe pattern," that work used human variant data with over 1,000 samples. Insect-focused studies tend to have much smaller sample sizes, and we suspect using only 34 samples is a significant contributing factor to the difficulty in using PCA projections to detect inversions in our data (see Section 3.1).

Our new method, however, combines PCA and single-SNP association tests to help overcome this limitation. In Section 3.2, we validated that our method works as expected on simulated data. We then applied our method to SNPs from 34 previously published samples. When the p-values of SNPs associated with each arm 2L PC were plotted along the chromosome, the 2La inversion was clearly and unambiguously captured by 2L-PC1 (see Section 3.3). When applied to the 3 and X chromosomes, our method did not detect any inversions, as expected.

We then applied our method to SNPs on the 2R chromosome arm (see Section 3.4). Note that neither the karyotypes nor their frequencies for the four smaller inversions on 2R are known for these samples. Even so, the four 2R inversions were clearly and unambiguously captured, although with relatively lower SNP-based significance relative to 2La. When combined geographic location associations (see Section 3.5), we hypothesize that karyotypes of the 2Rc, 2Rj, and 2Ru inversions are associated with Mali versus non-Mali samples and 2Ru inversion are associated with Burkina Faso versus non-Burkina Faso samples.

Our method has several advantages and disadvantages with respect to detecting inversions. We have shown that our method is more sensitive and accurate than PC projections alone. Although we expect that long reads (Oxford Nanopore, PacBio) may have a role in computational karyotyping, existing methods require completely assembled breakpoints. Our method is less useful at present for data sets with incomplete assemblies—we rely on the assembly to determine the spatial relationships of the SNPs—but we also have shown conservative p-value cutoffs can clearly distinguish inversion regions. Applying this method to other insects with well-characterized inversions like members of the *Drosophila* genus is left for future work. Finally, although our current implementation does not provide a mechanism for karyotyping samples *in silico*, preliminary results indicate that we can do so by clustering samples along PCs (data not shown).

To demonstrate utility beyond inversions, we used our method to characterize 2L-PC3 from its associated SNPs. We tested associations between the PCs geographic location and species labels given their importance in prior analyses [15]. The third PCs had no strong or moderate associations with the geographic or species labels. When we analyzed the SNPs associated with the 2L PCs, however, we observed that an insecticide-resistance mutation in *Rdl* was strongly associated with 2L-PC3 alone. We hypothesized that 2L-PC3 might be capturing differentiation due to insecticide resistance and exposure. As illustrated by this example, we anticipate that our method will be widely useful beyond the original objective of detecting inversions.

ACKNOWLEDGMENT

The authors would like to thank Nora Besansky, Michael Fontaine, Becca Love, and Aaron Steele 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.

REFERENCES

[1] Alejandro Cáceres and Juan R González. 2015. Following the footprints of polymorphic inversions on SNP data: from detection to association tests. *Nucleic Acids Res.* 43, 8 (April 2015), e53.

[2] Alejandro Cáceres, Suzanne S Sindi, Benjamin J Raphael, Mario Cáceres, and Juan R González. 2012. Identification of polymorphic inversions from genotypes. *BMC Bioinformatics* 13 (Feb. 2012), 28.

[3] Jorge Cadima and Ian T Jolliffe. 1995. Loading and correlations in the interpretation of principle compenents. *J. Appl. Stat.* 22, 2 (Jan. 1995), 203–214.

[4] Ken Chen, John W Wallis, Michael D McLellan, David E Larson, Joelle M Kalicki, Craig S Pohl, Sean D McGrath, Michael C Wendl, Qunyuan Zhang, Devin P Locke, Xiaoqi Shi, Robert S Fulton, Timothy J Ley, Richard K Wilson, Li Ding, and Elaine R Mardis. 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* 6, 9 (Sept. 2009), 677–681.

[5] Russell B Corbett-Detig, Charis Cardeno, and Charles H Langley. 2012. Sequence-based detection and breakpoint assembly of polymorphic inversions. *Genetics* 192, 1 (Sept. 2012), 131–137.

[6] W. Du, T. Awolola, P. Howell, L. Koekemoer, B. Brooke, M. Benedict, M. Coetzee, and L. Zheng. 2005. Independent mutations in the Rdl locus confer dieldrin resistance to *Anopheles gambiae* and *An. arabiensis*. *Insect Mol Biol* 14, 2 (2005), 179–183. https://doi.org/10.1111/j.1365-2583.2004.00544.x

[7] M. C. Fontaine, J. B. Pease, A. Steele, R. M. Waterhouse, D. E. Neafsey, I. V. Sharakhov, X. Jiang, A. B. Hall, F. Catteruccia, E. Kakani, et al. 2015. Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science* 347, 6217 (2015). https://doi.org/10.1126/science.1258524 arXiv:http://www.sciencemag.org/content/347/6217/1258524.full.pdf

[8] Zachary Fuller, Christopher Leonard, Randee Young, Stephen Schaeffer, and Nitin Phadnis. 2017. The role of chromosomal inversions in speciation. (Nov. 2017), 211771 pages.

[9] Emilie M Gray, Kyle A C Rocca, Carlo Costantini, and Nora J Besansky. 2009. Inversion 2La is associated with enhanced desiccation resistance in Anopheles gambiae. *Malar. J.* 8 (Sept. 2009), 215.

[10] Fereydoun Hormozdiari, Can Alkan, Evan E Eichler, and S Cenk Sahinalp. 2009. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res.* 19, 7 (July 2009), 1270–1278.

[11] D. W. Hosmer Jr., S. Lemeshow, and R. X. Sturdivant. 2013. *Applied Logistic Regression* (3 ed.). Wiley.

[12] J. D. Hunter. 2007. Matplotlib: A 2D graphics environment. *Computing In Science & Engineering* 9, 3 (2007), 90–95.

[13] M. K. N. Lawniczak, S. J. Emrich, A. K. Holloway, A. P. Regier, M. Olson, B. White, S. Redmond, L. Fulton, E. Appelbaum, J. Godfrey, et al. 2010. Widespread Divergence Between Incipient *Anopheles gambiae* Species Revealed by Whole Genome Sequences. *Science* 330, 6003 (2010), 512–514. https://doi.org/10.1126/science.1195755 arXiv:http://www.sciencemag.org/content/330/6003/512.full.pdf

[14] Chih Lee, Ali Abdool, and Chun-Hsi Huang. 2009. PCA-based population structure inference with generic clustering algorithms. *BMC Bioinformatics* 10 Suppl 1 (Jan. 2009), S73.

[15] R Rebecca Love, Aaron M Steele, Mamadou B Coulibaly, Sékou F Traore, Scott J Emrich, Michael C Fontaine, and Nora J Besansky. 2016. Chromosomal inversions and ecotypic differentiation in Anopheles gambiae: the perspective from whole-genome sequencing. *Mol. Ecol.* 25, 23 (Dec. 2016), 5889–5906.

[16] Jianzhong Ma and Christopher I Amos. 2012. Investigation of inversion polymorphisms in the human genome using principal components analysis. *PLoS One* 7, 7 (July 2012), e40224.

[17] Jianzhong Ma, Momiao Xiong, Ming You, Guillermina Lozano, and Christopher I Amos. 2014. Genome-wide association tests of inversions with application to psoriasis. *Hum. Genet.* 133, 8 (Aug. 2014), 967–974.

[18] Bradley J Main, Yoosook Lee, Travis C Collier, Laura C Norris, Katherine Brisco, Abdrahamane Fofana, Anthony J Cornel, and Gregory C Lanzaro. 2015. Complex genome evolution in Anopheles coluzzii associated with increased insecticide usage in Mali. *Mol. Ecol.* 24, 20 (Oct. 2015), 5145–5157.

[19] A. Miles, N. J Harding, G. Botta, C. Clarkson, T. Antao, K. Kozak, D. Schrider, A. Kern, S. Redmond, I. Sharakhov, et al. 2016. Natural diversity of the malaria vector *Anopheles gambiae*. (2016). https://doi.org/10.1101/096289

[20] D. E. Neafsey, M. K. N. Lawniczak, and D. J. Park. 2010. SNP genotyping defines complex gene-flow boundaries among African malaria vector mosquitoes. *Science* 2984 (2010).

[21] Ronald J Nowling and Scott J Emrich. 2018. Adjusted Likelihood-Ratio Test for Variants with Unknown Genotypes. In *10th International Conference on Bioinformatics and Computational Biology (BiCOB)*.

[22] Peristera Paschou, Elad Ziv, Esteban G Burchard, Shweta Choudhry, William Rodriguez-Cintron, Michael W Mahoney, and Petros Drineas. 2007. PCA-correlated SNPs for structure identification in worldwide human populations. *PLoS Genet.* 3, 9 (Sept. 2007), 1672–1686.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[24] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38, 8 (Aug. 2006), 904–909.

[25] Michelle M Riehle, Tullu Bukhari, Awa Gneme, Wamdaogo M Guelbeogo, Boubacar Coulibaly, Abdrahamane Fofana, Adrien Pain, Emmanuel Bischoff, Francois Renaud, Abdoul H Beavogui, Sekou F Traore, N'fale Sagnon, and Kenneth D Vernick. 2017. The Anopheles gambiae 2La chromosome inversion is associated with susceptibility to Plasmodium falciparumin in Africa. *Elife* 6 (June 2017).

[26] Kyle A C Rocca, Emilie M Gray, Carlo Costantini, and Nora J Besansky. 2009. 2La chromosomal inversion enhances thermal tolerance of Anopheles gambiae larvae. *Malar. J.* 8 (July 2009), 147.

[27] Joseph C Roden, Brandon W King, Diane Trout, Ali Mortazavi, Barbara J Wold, and Christopher E Hart. 2006. Mining gene expression data by interpreting principal components. *BMC Bioinformatics* 7 (April 2006), 194.

[28] Haojing Shao, Devika Ganesamoorthy, Tania Duarte, Minh Duc Cao, Clive Hoggart, and Lachlan Coin. 2017. npInv: accurate detection and genotyping of inversions mediated by non-allelic homologous recombination using long read sub-alignment. (Aug. 2017), 178103 pages.

[29] Suzanne S Sindi and Benjamin J Raphael. 2010. Identification and frequency estimation of inversion polymorphisms from haplotype data. *J. Comput. Biol.* 17, 3 (March 2010), 517–531.

[30] Toshifumi Suzuki, Yoshinori Tsurusaki, Mitsuko Nakashima, Noriko Miyake, Hirotomo Saitsu, Satoru Takeda, and Naomichi Matsumoto. 2014. Precise detection of chromosomal translocation or inversion breakpoints by whole-genome sequencing. *J. Hum. Genet.* 59, 12 (Dec. 2014), 649–654.

[31] Robert Tibshirani. 1996. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Series B Stat. Methodol.* 58, 1 (1996), 267–288.

[32] S. v. d. Walt, S. C. Colbert, and G. Varoquaux. 2011. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science & Engineering* 13, 2 (2011), 22–30. https://doi.org/10.1109/MCSE.2011.37

[33] David Weetman, Craig S Wilding, Daniel E Neafsey, Pie Müller, Eric Ochomo, Alison T Isaacs, Keith Steen, Emily J Rippon, John C Morgan, Henry D Mawejje, Daniel J Rigden, Loyce M Okedi, and Martin J Donnelly. 2018. Candidate-gene based GWAS identifies reproducible DNA markers for metabolic pyrethroid resistance from standing genetic variation in East African Anopheles gambiae. *Sci. Rep.* 8, 1 (Feb. 2018), 2920.

[34] S Zhu, S J Emrich, and D Z Chen. 2017. Inversion detection using PacBio long reads. 237–242.

[35] Hui Zou and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series B Stat. Methodol.* 67, 2 (April 2005), 301–320.

[36] Hui Zou, Trevor Hastie, and Robert Tibshirani. 2006. Sparse Principal Component Analysis. *J. Comput. Graph. Stat.* 15, 2 (June 2006), 265–286.