Segmenting and Genotyping Large, Polymorphic Inversions

Ronald J. Nowling* Samuel H. Keyser† Alex R. Moran‡ John G. Peters§

Daniel Leskiewicz¶

Electrical Engineering and Computer Science
Milwaukee School of Engineering
Milwaukee, WI USA
Email: *nowling@msoe.edu, †keysers@msoe.edu, ‡morana@msoe.edu, §petersjg@msoe.edu, ¶leskiewiczd@msoe.edu

2023 IEEE International Conference on Electro Information Technology (eIT) | 978-1-6654-9376-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/EIT57321.2023.10187331

***Abstract*—Large, polymorphic inversions can contribute to population structure and enable mutually-exclusive adaptations to survive in the same population. Current methods for detecting inversions from single-nucleotide polymorphisms (SNPs) called from population genomics data require an experienced, human user to prepare the data and interpret the results. Ideally, these methods would be completely automated yet robust to allow usage by inexperienced users. Towards this goal, automated approaches for segmentation of inversions and inference of sample genotypes are introduced and evaluated on chromosomes from flies, mosquitoes, and prairie sunflowers.**

***Index Terms*—population genomics, inversions, segmentation, PCA, clustering**

I. Introduction

An inversion is a reversal of a subsequence in a larger sequence [1]. For example, the subsequence $[6..4]$ of the sequence $[1, 2, 3, 6, 5, 4, 7, 8, 9]$ is inverted relative to the sequence $[1..9]$. Regions of chromosomes, the large molecules of DNA in which genomes are physically organized, can also undergo inversions during meiosis [2]. A polymorphic inversion differs in orientations (standard or inverted) across individuals in the same species or population. Organisms may contain one (haploid), two (diploid), or more (polyploid) copies of each chromosome. Each copy may have a different orientation of the inversion. The inversion genotype refers to the combinations of these inversion orientations (e.g., homozygous inverted, homozygous standard, or heterozygous for a diploid genome).

Detecting inversions is of great interest to the genomics community. During meiosis, the parent's chromosomes can swap pieces (recombine), which is a way of sharing genetic changes with offspring and disrupting correlations (linkage) between adjacent nucleotides [2]. Inverted regions do not recombine with non-inverted regions, however. This enables each inversion orientation to accumulate and maintain private mutations. As a result, through inversions, a species may carry multiple, mutually-exclusive genetic mutations that are advantageous. For example, inversions in *Anopheles* mosquitoes have been associated with thermal tolerance [3], desiccation resistance [4]–[6], and susceptibility to infection by the malaria parasite [7]. The frequencies of inversions can vary geographically and by time of year in correlation with geospatial climate and seasonal patterns [8].

Inversions can be detected from variants (most commonly single-nucleotide polymorphisms or SNPs) called from population genomics sequencing data [9]–[11]. Most methods depend on principal component analysis (PCA), in which samples appear to cluster by inversion genotype [10], [12]. Population structure, selective sweeps, population bottlenecks, and other processes can also produce similar cluster structures in PCA that confound detection of inversions [13]–[18]. The spatial pattern of variants contributing to the cluster structure can be revealed through association testing and visualization [19]–[23]. Inversions can be differentiated from other phenomena by the presence of a square wave-like pattern in Manhattan plots [18], [22], [23].

Despite the ability of these methods to detect and localize inversions, the inversion analysis process is still largely manual and requires an experienced human operator to interpret noisy PCA and Manhattan plots. Subpopulations of data sets may need to be selected to achieve a clear inversion signal [18]. The presence of multiple, overlapping inversions can also result in noisy signals that make both the PCA and Manhattan plots difficult to interpret for a non-expert [18]. Ideally, the inversion analysis process would be completely automated to enable use by non-experts and in a high-throughput manner to the increasing number of publicly-available variant data sets.

A previous method called the window test for automated segmentation of inversion regions was introduced by [23]. The primary downfall of this method is that it assumes that each PC has a single, contiguous inversion region associated with it. In this paper, we validate our inversion analysis tool Asaph on three chromosomes from the prairie sunflower. We introduce alternative visualizations and segmentation models and evaluate them on nine chromosomes across three organisms (fly, mosquito, and prairie sunflowers). Lastly, we describe and evaluate an approach to automate inference of inversion genotypes using the regions identified by the segmentation

This work received financial support to RJN from the National Science Foundation (IIS #194727). Funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

models. These new methods are available in the latest release of Asaph.

II. Methods

A. Preparation of Data Sets

We used SNP data from three organisms (the fly *Drosophila melanogaster*, the closely-related malaria mosquitoes *Anopheles gambiae* and *coluzzii*, and the prairie sunflower *Helianthus petiolaris* var. var. *petiolaris*, see Table I). VCF files were retrieved from repositories reported in the original papers [24]–[26]. The VCF files were filtered to select biallelic SNPs with a minimum minor allele frequency of 5%. Sample inversion genotypes and ranges were retrieved from [24], [27], [28] (*Anopheles*), [25], [29] (*Drosophila*), and [26] (*Helianthus*). Scripts and meta data files are available at https://github.com/nowling-lab/inverson-test-set.

B. Feature Matrix Construction and Principal Component Analysis (PCA)

The biallelic variant data were read from a variant call format (VCF) file [30] provided by the user. VCF files store the $m \times n$ (samples by variants) matrix column-wise (one variant per line) in a text format. Subsampling was used to reduce memory requirements. Each line was read and parsed to produce two 1D Numpy arrays [31], [32] of reference and alternative allele counts with missing values set to 0. Strings of the form "chrom_position_allele" (e.g., "2L_5678_T" or "2L_5678_A") were created for each column. The strings were hashed using murmurhash3 [33] as implemented by the mmh3 library [34] to 32-bit signed integers. Pairs of `(abs(hash), count_vector)` for each allele were inserted into a k-sized min heap implemented using Python's `heapq` module. After processing the file, the heap contained the k variants with the smallest-magnitude hash values, also known as a bottom-k sketch [35]–[37]. The parameter k was estimated from the approximate inversion size and number of samples using the heuristic described in [23]. This process resulted in an $m \times k$ allele count matrix.

Principal component analysis (PCA) was performed using scikit-learn's PCA class [38]. The number of components to calculate was set to ten, while the rest of the parameters were left at their defaults. The samples' coordinates along each component were written to disk in a text file.

C. PC-SNP Association Testing

After performing PCA, the variant call format (VCF) file was re-read and each variant was tested for association with each of the principal components (PCs). The allele counts of the samples were tested against the samples' coordinates along a single principal component using a one-way analysis of variation (ANOVA) test as implemented by Scipy [39]. Samples' PC coordinates were partitioned into groups by the samples' genotypes (homozygous reference, homozygous alternative, and heterozygous). When the allele counts for a sample were missing, the sample was excluded from the test. The resulting p-values were written to a text file.

D. Spatial Visualizations of PC-SNP Associations

1) Manhattan Plots: In Fig. 1 and 2, Manhattan plots were generated for each PC by creating a scatter plot from the chromosome positions and $-\log_{10}$ transformed p-values. Each SNP is categorized as significant or not using a Bonferroni-corrected significance threshold of $0.01/\text{n\_variants}$. Dots for the significance SNPs were colored orange, while the insignificant SNPs were colored blue. Expected inversion boundaries were plotted as horizontal black lines.

2) Window Plots: In Fig. 1 and 2, an alternative visualization approach was introduced that plots the fraction of significant SNPs in non-overlapping windows. The chromosome was divided into non-overlapping windows (1 Mb for the prairie sunflower samples and 250 Kb for the other data sets). In each window, the fraction of significant SNPs to total SNPs in the window was calculated. The significance fractions were visualized by creating piecewise line plots (in purple) with horizontal segments for each window. Expected inversion boundaries were plotted as horizontal black lines. When segmented regions were available, the corresponding portions of the lines were plotted in orange.

E. Segmentation of Inversions

1) Window Test: The window test was introduced in [23] and is summarized here. The SNP p-values from the PC–SNP association tests (see above) were used as input to the algorithm. Each SNP was categorized as significant or not using a Bonferroni-corrected significance threshold of $0.01/\text{n\_variants}$. The chromosome was divided into non-overlapping windows (default window size of 10 Kb). The fraction of significant SNPs in each window was tested using a binomial test with the alternative hypothesis that the observed fraction of statistically significant SNPs was greater than expected. The expected probability of success (that a SNP is significant) was estimated as the fraction of statistically significant SNPs across the entire chromosome. In cases where a window had no SNPs or no significant SNPs, the p-value was estimated as 1.0. Windows were tested for significance using a Bonferroni-corrected significance threshold of $0.0001/\text{num\_windows}$. Lastly, the inversion ends were estimated from the centers of the leftmost and rightmost statistically significant windows. Note that this method assumed that there was only one contiguous inversion associated with each PC.

2) Gaussian Hidden Markov Model: The SNP p-values from the PC–SNP association tests (see above) were used as input to the algorithm. Each SNP was categorized as significant or not using a Bonferroni-corrected significance threshold of $0.01/\text{n\_variants}$. The chromosome was divided into non-overlapping windows (1 Mb for the prairie sunflowers and 250 Kb for the remaining samples). For each window, the fraction of significant SNPs was calculated. The window significance fractions were scaled to the range $[0, 1]$ within each chromosome by

$$\text{scaled} = \frac{\text{win\_sig\_frac} - \text{lower}}{\text{upper} - \text{lower}}$$

TABLE I
DETAILS OF DATA SETS USED IN THIS STUDY.

Species	Chromosome	Known Inversions (Ranges)	Number of Samples	Data Source
An. gambiae	2L	2La (20.5 - 42.2 Mb)	89	[24]
An. coluzzii	2R	2Rbc (19.0 Mb - 31.5 Mb); 2Rd (31.5 Mb - 42.4 Mb)	61	[24]
An. gambiae	2R	2Rb (19.0 - 26.8 Mb)	89	[24]
Drosophila	2L	*In(2L)t* (2.2 - 13.2 Mb)	198	[25]
Drosophila	2R	*In(2R)ns* (11.3 - 16.2 Mb)	198	[25]
Drosophila	3R	*In(3R)mo* (17.2 Mb - 24.9 Mb)	198	[25]
Helianthus	Pet05	Pet05.01 (154 - 186 Mb)	166	[26]
Helianthus	Pet09	Pet09.01 (105 - 141 Mb)	166	[26]
Helianthus	Pet11	Pet11.01 (3.0 - 65.0 Mb)	166	[26]

The windows were clustered into two clusters using the Gaussian Mixture Model (GMM) implementation in Scikit-Learn [38]. The GMM was trained with two components, diagonal covariance, k-means++ for initializing cluster centers, and ten initialization trials. All other parameters were left at their default values.

The windows were segmented using a 2-component Gaussian Hidden Markov Model (GHMM) with diagonal covariances using hmm-learn [40]. The means and covariances of the two GHMM components were initialized using the means and covariance parameters inferred by the GMM model. The transition matrix and starting probabilities were randomly initialized (`init_params=''st''`). The remaining parameters were left at their defaults. The GHMM model was fitted using the Baum-Welch algorithm. Hidden states for each window were then inferred by applying the model using the Viterbi algorithm. If the mean of state 0 was larger than that of state 1, the state labels were switched so that state 1 always indicated an inversion region.

3) Evaluation: The automated segmentation outputs were evaluated by calculating balanced accuracy, recall, and precision on a per-nucleotide position basis.

F. Automated Genotyping

1) Method: A method to perform automated genotyping was introduced. The VCF file was re-read for a third time. If segmentation coordinates were provided, only SNPs in the region were kept. A feature matrix was constructed and PCA performed as described above. Samples were clustered using the Gaussian Mixture Model (GMM) implementation in Scikit-Learn [38] from their coordinates along the first two PCs. Two cluster models, one with two components and the other with three components were trained. The GMMs were trained with two components, diagonal covariance, k-means++ for initializing cluster centers, and ten initialization trials. All other parameters were left at their default values. The best-fitting GMM model was chosen using the Davies-Bouldin index [41]. Cluster labels were evaluated against known sample genotypes using the adjusted Rand index [42].

2) Visualization: In Fig. 3, dots were plotted for the samples using their projected PCA coordinates; the dots were colored by their known genotypes. The GMM model was displayed by evaluating the probabilities on a grid and creating contour plots using Matplotlib's contour function [43].

G. Software Implementation and Availability

Asaph is available on GitHub (https://github.com/nowling-lab/asaph) under the open-source Apache Software License v2.0. Asaph is implemented in Python 3 and uses the Numpy [31], [32], Scipy [39], Matplotlib [43], Scikit-Learn [38], hmm-learn [40], and mmh3 [34] libraries. Documentation is provided in the form of tutorials available in the repository.

III. RESULTS

A. A New Visualization Approach

Multiple factors (e.g., inversions, population structure, selective sweeps, and population bottlenecks) can cause clustering patterns in principal component analysis. Manhattan plots are a useful tool for visualizing the spatial distribution of SNP associations used in genome-wide association tests. When used to visualize PC-SNP associations, inversions can be distinguished from other processes by a square-wave pattern.

As implemented in Asaph, Manhattan plots were previously evaluated against PCA scatter plots for detecting inversions on chromosomes with and without inversions from the fly *Drosophila melanogaster* and the closely-related malaria mosquitoes *Anopheles gambiae* and *coluzzii* [18], [22]. Here, we extended the validation of Asaph to include three chromosomes from the prairie sunflower *Helianthus petiolaris* var. var. *petiolaris*. Manhattan plots for three *Drosophila*, three *Anopheles*, and three *Helianthus* chromosomes are shown in Fig. 1 (first row). In all nine cases, the Manhattan plots display the expected square-wave patterns indicative of inversions.

We also tried an alternative approach for visualizing the signal along the chromosome. Chromosomes were divided into non-overlapping, equally-sized windows (250 Kb for *Drosophila* and *Anopheles*, 1 Mb for *Helianthus*). The fraction of SNPs in each window with significant associations was calculated. The window fractions across the chromosome were plotted in purple (see Fig.1, second row).

The visualizations generated from plotting the windows' significant SNP fractions were remarkably clear compared to the Manhattan plots (Fig.1). The noisy background signal in the Manhattan plots was nearly absent in the significance fraction plots. As expected, the signals for the *Anopheles* and *Helianthus* chromosomes aligned closely with the expected inversions; significance fractions were enriched in and sharply

declined at the borders to substantially lower levels outside of the inversion region. Notably, the signals for the *Helianthus* pet09 and pet11 chromosomes indicated small additional regions of low significance within the expected chromosome region which might impact segmentation. Both types of plots indicated a break in the pet09.01 inversion region, while a similar break in pet05.01 was only visible in the significance fraction plots. For the *Drosophila* chromosomes, the signals declined slowly rather than sharply at the boundaries. The slow decline is indicative of recombination, expected since these lines were heavily inbred, which can cause the inversion boundaries to shift among samples.

B. Alternative Segmentation Approach with Gaussian Hidden Markov Models (GHMMs)

The current "window test" segmentation demonstrated high segmentation accuracies when previously evaluated on the *Anopheles gambiae* 2L and 2R and *Drosophila melanogaster* 2L and 2R chromosomes [23]. The window test's primary limitation was that it assumed a single, contiguous inversion per PC.

Here, we expanded the evaluation to the *Anopheles coluzzii* 2R, *Drosophila* 3R, and *Helianthus* pet05, pet09, and pet11 chromosomes (see Table II). The segmented regions for the *Anopheles gambiae* 2L (99.8% balanced accuracy) and 2R (99.6%), *Drosophila* 2R (97.5%), and *Helianthus* Pet09 (98.3%) and Pet11 (99.3%) agreed remarkably well with the expected inversion boundaries. SNPs on the right-hand end of the 2Rbc inversion in *Anopheles coluzzii* show substantially reduced associations which caused the segmentation algorithm to miss 3 Mb of the total 12.5 Mb expected region. The presence of significant SNPs outside the expected inversion regions for *Drosophila* 2L and 3R led to segmentation of larger regions than expected.

Using the significant SNP fractions by window as input, we developed an alternative segmentation approach using clustering and Gaussian Hidden Markov Models (GHMMs). A 2-component Gaussian Mixture Model (GMM) with diagonal covariances was used to partition the windows into 2 clusters by their fraction of significant SNPs. The inferred means and covariances from the GMM were used to initialize the means and covariances of a 2-component Gaussian Hidden Markov Model (GHMM) with diagonal covariances. The remaining GHMM parameters (transition matrix and starting probabilities) were inferred using the Baum-Welch algorithm. The trained GHMM was applied to the signal data to determine the most probable hidden state for each window using the Viterbi algorithm and label each window as state 0 or 1. If the average for state 0 was higher than state 1, the labels were flipped so that state 1 indicates an inversion region and state 0 indicates no evidence of an inversion.

The significance fraction plots from Fig. 1 (bottom row) were reproduced but with the windows inferred to be inversions marked in orange (see Fig. 2, window test in the top row, GHMM model in the bottom row). The two segmentation methods give largely similar results except in a few cases. The two segmentation models demonstrated high accuracies for the *Anopheles* and *Helianthus* pet11 chromosomes and expectedly lower accuracies for the *Drosophila* inversions.

The sunflowers had regions within pet05.01 and pet09.01 with zero or very few significant SNPs. These regions were marked as not-inverted. Similarly, the original authors treated the empty regions as a break in the inversion [26]. Biologically, it is not clear whether this was correct or not; more investigation will be needed. The window test segmented these inversions as large contiguous units, while the GHMM model generated two segmented regions for each inversion. Similarly, the window test did not capture the lower significance region on the right-hand side of the *Anopheles coluzzii* 2Rbc inversion, while the GHMM segmentation did.

The GHMM model identified several small regions of interest outside of the inversions on pet05, pet11, and *Anopheles gambiae* 2R. These regions are not likely to be inversions; the regions are small and have sloped sides rather than square-wave patterns. The regions may be caused by selective sweeps or other processes and warrant further study.

C. Segmentation Enables Accurate Automated Genotyping

Here, we introduce an approach for automatically inferring inversion genotypes using clustering. Love, et al. [44] demonstrated that *Anopheles* inversion genotypes can be inferred more accurately when PCA is performed on SNPs in the known inversion region compared rather than the entire chromosome, especially if there are multiple inversions present. Clustering was compared using all SNPs on the chromosomes versus those in the segmented regions identified by the two segmentation methods.

First, PCA was performed on all of the SNPs or those in the specified region(s). The samples were then clustered using the coordinates along the first two PCs. Depending on whether two or three genotypes are present among the samples, there may be two or three clusters. Two Gaussian Mixture Models (GMMs) with two and three components, respectively, were fitted. The models were evaluated using the unsupervised Davies-Bouldin score for cluster structure; the model with the lowest score was used to infer genotypes and label the samples.

The cluster labels were compared with the known inversion genotypes for the samples using the adjusted Rand index (higher is better, see Table III). Clustering using the segmented regions from the GHMMs and window tests either equaled (4) or outperformed (5) clustering with the entire set of SNPs for all nine data sets. Notably, clustering on all of the SNPs missed one of the three genotypes in *Anopheles coluzzii* 2Rbc and the *Helianthus* inversions (see Fig. 3, top row). Clustering using SNPs from the segmented regions (either model) recovered all three genotypes (see Fig. 3, middle and bottom rows). Love, et al.'s observation was confirmed to apply to *Drosophila* and *Helianthus* as well.

The two segmentation methods produced the same or similar results across all of the data sets (see Table III, Fig. 3). Both methods accurately inferred the genotypes for the *Anopheles*,

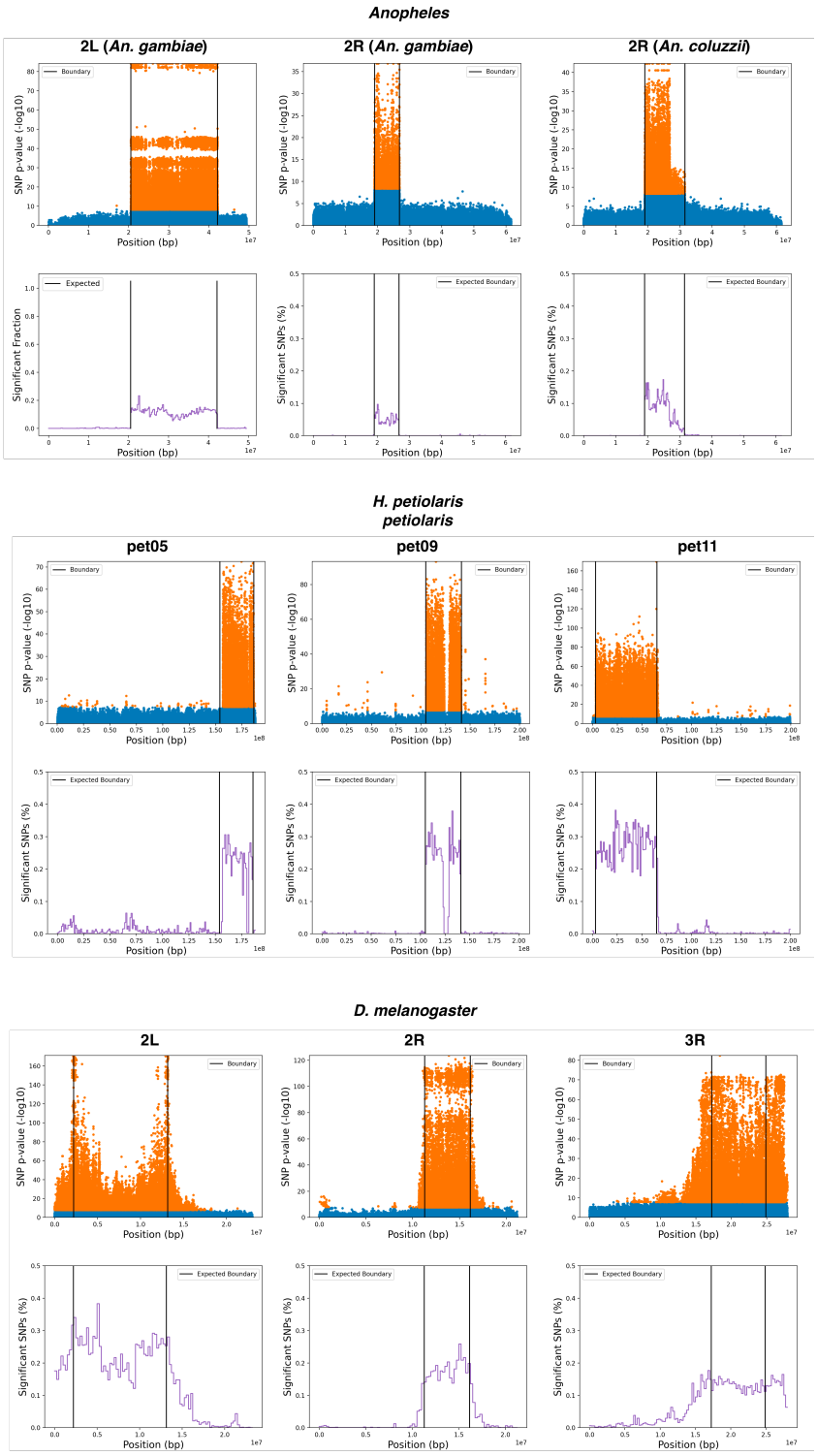

Fig. 1. **SNP Associations with Coordinates along the First Principal Component.** Manhattan plots (first row) and window significance plots (second row) for nine chromosomes from the fly *Drosophila melanogaster*, the closely-related malaria mosquitoes *Anopheles gambiae* and *coluzzii*, and the prairie sunflower *Helianthus petiolaris* var. var. *petiolaris* are shown. For the Manhattan plots, each point represents a single SNP and was colored according to statistical significance (orange for significant, blue if not). To generate the window plots, the chromosomes were divided into non-overlapping, equally-sized windows (250 kb for *Drosophila* and *Anopheles*, 1 Mb for *Helianthus*). The fraction of SNPs with significant associations with coordinates along the first principal component was calculated for each window and plotted in purple. The known inversion boundaries were indicated with black horizontal lines.

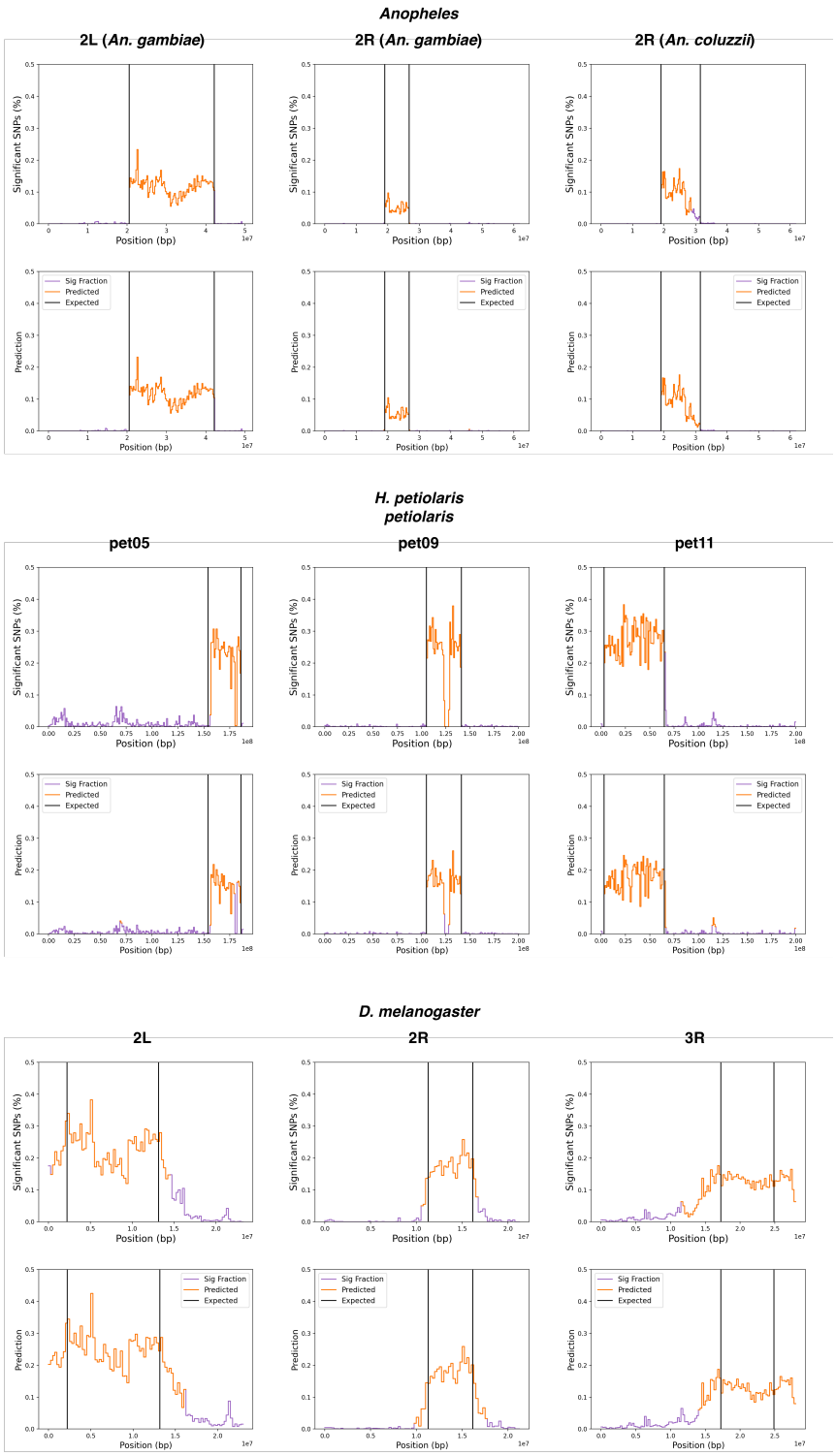

Fig. 2. **Segmentation Results.** Window significance plots with segmented regions in orange for nine chromosomes from the fly *Drosophila melanogaster*, the closely-related malaria mosquitoes *Anopheles gambiae* and *coluzzii*, and the prairie sunflower *Helianthus petiolaris* var. var. *petiolaris* are shown. The first row shows segmentation results from the window test model, while the second row shows segmentation results from the GHMM model. To generate the window plots, the chromosomes were divided into non-overlapping, equally-sized windows (250 kb for *Drosophila* and *Anopheles*, 1 Mb for *Helianthus*). The fraction of SNPs with significant associations with coordinates along the first principal component was calculated for each window and plotted in purple. Regions marked as inverted by the segmentation models are marked in orange. The known inversion boundaries were indicated with black horizontal lines.

TABLE II

EVALUATION OF SEGMENTATION RESULTS. THE REGIONS SEGMENTED BY THE WINDOW TEST AND GHMM MODELS FOR NINE CHROMOSOMES FROM THE FLY DROSOPHILA MELANOGASTER, THE CLOSELY-RELATED MALARIA MOSQUITOES ANOPHELES GAMBIAE AND COLUZZII, AND THE PRAIRIE SUNFLOWER HELIANTHUS PETIOLARIS VAR. VAR. PETIOLARIS WERE EVALUATED. SEGMENTATION RESULTS WERE USED TO MARK INDIVIDUAL NUCLEOTIDE POSITIONS AS PREDICTED TO BE INVERTED OR NOT. THE PREDICTIONS WERE COMPARED AGAINST THE KNOWN INVERSION BOUNDARIES USING BALANCED ACCURACY, RECALL, AND PRECISION.

Inversion	Expected Region (Mb)	Method	Predicted Region (Mb)	Balanced Accuracy	Recall	Precision
An. gambiae 2La	20.50 - 42.20	Window Test	20.55 – 42.16	99.8%	99.6%	100%
		GHMM	20.50 – 42.25	99.9%	100%	99.8%
An. coluzzii 2Rbc	19.00 - 31.50	Window Test	19.03 – 28.50	87.9%	75.8%	100%
		GHMM	19.00 – 31.50	100%	100%	100%
An. gambiae 2Rb	19.00 - 26.80	Window Test	19.03 – 26.77	99.6%	99.2%	100%
		GHMM	18.75 – 27.00; 45.75 – 46.00	99.3%	100%	91.8%
Drosophila In(2L)t	2.20 - 13.20	Window Test	0.47 – 14.30	88.2%	100%	79.5%
		GHMM	0.00 - 16.25	78.1%	100%	67.7%
Drosophila In(2R)ns	11.30 - 16.20	Window Test	10.68 – 16.55	97.5%	100%	83.5%
		GHMM	9.75 - 17.75	90.5%	100%	61.3%
Drosophila In(3R)mo	17.20 - 24.90	Window Test	11.74 – 27.76	79.4%	100%	48.1%
		GHMM	14.00 - 28.00	84.7%	100%	55.4%
Helianthus Pet05.01	154.00 - 186.00	Window Test	157.00 – 185.70	94.8%	89.7%	100%
		GHMM	69.00 - 71.00; 156.00 - 180.00; 182.00 - 186.00	93.1%	87.5%	93.3%
Helianthus Pet09.01	105.00 - 141.00	Window Test	105.38 – 140.15	98.3%	96.6%	100%
		GHMM	105.00 - 124.00; 128.00 - 141.00	94.4%	88.9%	100%
Helianthus Pet11.01	3.00 - 65.00	Window Test	3.59 – 65.67	99.3%	99.0%	98.9%
		GHMM	3.00 - 67.00; 114.00 - 118.00	97.8%	100%	91.2%

Helianthus, and *Drosophila In(2L)t* inversions but not for the remaining two *Drosophila* inversions.

IV. DISCUSSION AND CONCLUSION

PCA has been a successful foundation for the development of several approaches for identifying and characterizing large, polymorphic inversions. PCA can capture a wide variety of phenomena that cause correlation between variants, however, such as populations prevented from intermating by geographic barriers, selective sweeps, and population bottlenecks. Using current methods, an experienced human observer is needed to prepare data to reveal a clean signal and needed to interpret the results of outputs such as Manhattan plots.

Our goal is to develop a completely automated yet reliable approach to identify and characterize inversion patterns. An automated approach would enable scaling the detection of inversions to the large number of population genetics data sets generated by the genomics community on a regular basis. Automation would also remove human bias in the interpretation, leading to more consistent results.

As a step towards this goal, we evaluated our existing visualization and segmentation methods on three additional chromosomes from the prairie sunflower *Helianthus petiolaris* var. var. *petiolaris*. The pet05.01, pet09.01, and pet11.01 inversions were correctly represented both in the Manhattan plots and identified by the window test segmentation model.

We also described and evaluated alternative visualization and segmentation methods on nine total chromosomes from the *Drosophila* fly, *Anopheles* mosquitoes, and *Helianthus*. The new visualization approach displays the fraction of significant SNPs in non-overlapping windows along the chromosome. This approach substantially reduces the amount of noise, making it easier to detect inversions from the square-wave pattern. The GHMM segmentation approach accurately identified the inversion boundaries in *Anopheles* and *Helianthus* with some reduced accuracy for the *Drosophila* samples due to recombination, which was expected.

When compared, the GHMM and window test segmentations produced similar results. The main advantage of the GHMM approach is that it does not assume a single, contiguous inversion region. The advantage of being able to detect multiple regions was observed when identifying the breaks in pet05.01 and pet09.01 and additional regions of potential interest on pet05, pet09, and the *Anopheles coluzzii* 2R chromosomes.

The automated segmentation approaches output coordinates rather than requiring the user to infer the coordinates from plots. We used the coordinates as the basis of a new automated genotyping method. Our method tests several GMM models to determine the number of genotypes and then labels each sample. Clustering only on SNPs in segmentation regions produced substantially better agreement with known sample genotypes than using all of the SNPs along the chromosome for multiple data sets. Through the combination of automated segmentation and clustering, the inversion regions and sample genotypes can be inferred accurately and robustly without user input.

Our work here represents a step towards achieving our

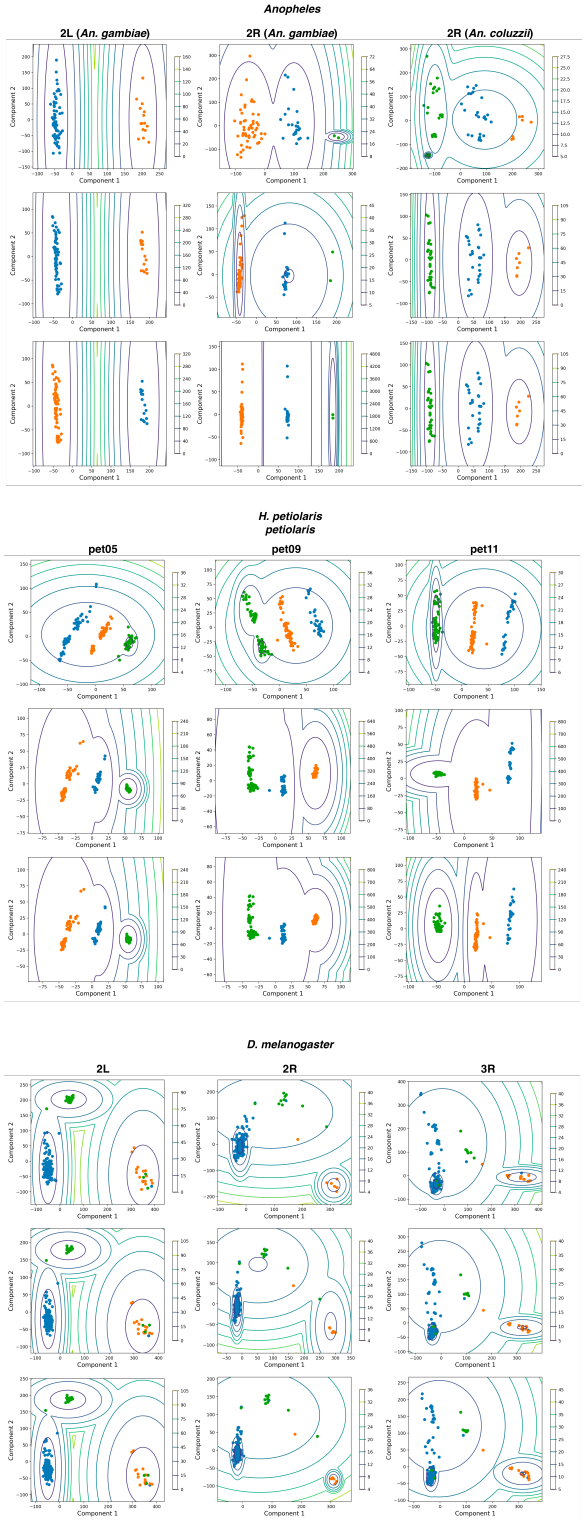

Fig. 3. **Genotype Inference Gaussian Mixture Models (GMMs).** Samples (one per dot) are plotted along the first two PC coordinates and colored by known inversion genotype. The predicted probabilities of the GMM models were evaluated along a grid and used to create contour plots.

TABLE III
EVALUATION OF THE GENOTYPE INFERENCE WITH GAUSSIAN MIXTURE MODELS (GMMS). CLUSTERING WAS PERFORMED USING SNPS FROM THE ENTIRE CHROMOSOME AND THOSE LOCATED WITHIN SEGMENTED REGIONS FROM THE GHMM AND WINDOW TEST MODELS. THE ADJUSTED RAND INDEX WAS USED TO EVALUATE THE CLUSTER STRUCTURE VERSUS THE KNOWN INVERSION GENOTYPES.

Inversion	Entire Chromosome	GHMM	Window Test
An. gambiae 2La	1.00	1.00	1.00
An. coluzzii 2Rbc	1.00	1.00	0.97
An. gambiae 2Rb	0.63	1.00	1.00
Drosophila In(2L)t	0.97	0.97	0.97
Drosophila In(2R)ns	0.76	0.76	0.81
Drosophila In(3R)mo	0.41	0.40	0.33
Helianthus Pet05.01	0.40	1.00	1.00
Helianthus Pet09.01	0.33	1.00	0.98
Helianthus Pet11.01	0.73	1.00	1.00

ultimate goal of a completely automated and reliable workflow. That said, further work is needed. Our method assumes that the chromosome has at least one inversion region that must be separated from the non-inversion regions. Our method does not have the ability to determine when no inversion is present. Secondly, our method tests each PC independently – the user must choose which PCs to evaluate based on analysis of the SNP significance plots. We intend to tackle these two challenges in future work.

ACKNOWLEDGMENT

The authors would like to thank Jenica P. Abrudan for offering feedback on the manuscript.

REFERENCES

[1] D. Gusfield, *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*, 1st ed. Cambridge University Press, May 1997.
[2] T. Brown, *Genomes 4*, 4th ed. Garland Science, May 2017.
[3] K. A. C. Rocca, E. M. Gray, C. Costantini, and N. J. Besansky, "2la chromosomal inversion enhances thermal tolerance of anopheles gambiae larvae," *Malar. J.*, vol. 8, p. 147, Jul. 2009.
[4] E. M. Gray, K. A. C. Rocca, C. Costantini, and N. J. Besansky, "Inversion 2la is associated with enhanced desiccation resistance in anopheles gambiae," *Malar. J.*, vol. 8, p. 215, Sep. 2009.
[5] C. Fouet, E. Gray, N. J. Besansky, and C. Costantini, "Adaptation to aridity in the malaria mosquito anopheles gambiae: chromosomal inversion polymorphism and body size influence resistance to desiccation," *PLoS One*, vol. 7, no. 4, p. e34841, Apr. 2012.
[6] D. Ayala, S. Zhang, M. Chateau, C. Fouet, I. Morlais, C. Costantini, M. W. Hahn, and N. J. Besansky, "Association mapping desiccation resistance within chromosomal inversions in the african malaria vector anopheles gambiae," *Mol. Ecol.*, Sep. 2018.
[7] M. M. Riehle, T. Bukhari, A. Gneme, W. M. Guelbeogo, B. Coulibaly, A. Fofana, A. Pain, E. Bischoff, F. Renaud, A. H. Beavogui, S. F. Traore, N. Sagnon, and K. D. Vernick, "The anopheles gambiae 2la chromosome inversion is associated with susceptibility to plasmodium falciparumin in africa," *Elife*, vol. 6, Jun. 2017.
[8] D. Ayala, P. Acevedo, M. Pombi, I. Dia, D. Boccolini, C. Costantini, F. Simard, and D. Fontenille, "Chromosome inversions and ecological plasticity in the main african malaria mosquitoes," *Evolution*, vol. 71, no. 3, pp. 686–701, Mar. 2017.
[9] S. S. Sindi and B. J. Raphael, "Identification and frequency estimation of inversion polymorphisms from haplotype data," *J. Comput. Biol.*, vol. 17, no. 3, pp. 517–531, Mar. 2010.
[10] J. Ma and C. I. Amos, "Investigation of inversion polymorphisms in the human genome using principal components analysis," *PLoS One*, vol. 7, no. 7, p. e40224, Jul. 2012.
[11] H. Li and P. L. Ralph, "Local PCA shows how the effect of population structure differs along the genome," *Genetics*, Nov. 2018.
[12] A. Cáceres and J. R. González, "Following the footprints of polymorphic inversions on SNP data: from detection to association tests," *Nucleic Acids Res.*, vol. 43, no. 8, p. e53, Apr. 2015.
[13] J. Ma and C. I. Amos, "Principal components analysis of population admixture," *PLoS One*, vol. 7, no. 7, p. e40115, Jul. 2012.
[14] D. Reich, A. L. Price, and N. Patterson, "Principal component analysis of genetic data," *Nat. Genet.*, vol. 40, no. 5, pp. 491–492, May 2008.
[15] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, "Principal components analysis corrects for stratification in genome-wide association studies," *Nat. Genet.*, vol. 38, no. 8, pp. 904–909, Aug. 2006.
[16] N. Patterson, A. L. Price, and D. Reich, "Population structure and eigenanalysis," *PLoS Genet.*, vol. 2, no. 12, p. e190, Dec. 2006.
[17] G. McVean, "A genealogical interpretation of principal components analysis," *PLoS Genet.*, vol. 5, no. 10, p. e1000686, Oct. 2009.
[18] R. J. Nowling, K. R. Manke, and S. J. Emrich, "Detecting inversions with PCA in the presence of population structure," *PLoS One*, vol. 15, no. 10, p. e0240429, Oct. 2020.
[19] X. Zheng, D. Levine, J. Shen, S. M. Gogarten, C. Laurie, and B. S. Weir, "A high-performance computing toolset for relatedness and principal component analysis of SNP data," *Bioinformatics*, vol. 28, no. 24, pp. 3326–3328, Dec. 2012.
[20] F. Privé, K. Luu, B. J. Vilhjálmsson, and M. G. B. Blum, "Performing highly efficient genome scans for local adaptation with R package pcadapt version 4," *Mol. Biol. Evol.*, Apr. 2020.
[21] K. Luu, E. Bazin, and M. G. B. Blum, "pcadapt: an R package to perform genome scans for selection based on principal component analysis," *Mol. Ecol. Resour.*, vol. 17, no. 1, pp. 67–77, Sep. 2016.
[22] R. J. Nowling and S. J. Emrich, "Detecting chromosomal inversions from dense SNPs by combining PCA and association tests," in *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, ser. BCB '18. New York, NY, USA: Association for Computing Machinery, Aug. 2018, pp. 270–276.
[23] R. J. Nowling, F. Fallas-Moya, A. Sadovnik, S. Emrich, M. Aleck, D. Leskiewicz, and J. G. Peters, "Fast, low-memory detection and localization of large, polymorphic inversions from SNPs," *PeerJ*, vol. 10, p. e12831, Jan. 2022.
[24] Anopheles gambiae 1000 Genomes Consortium, "Genetic diversity of the african malaria vector anopheles gambiae," *Nature*, vol. 552, no. 7683, pp. 96–100, Dec. 2017.
[25] W. Huang, A. Massouras, Y. Inoue, J. Peiffer, M. Ràmia, A. M. Tarone, L. Turlapati, T. Zichner, D. Zhu, R. F. Lyman, M. M. Magwire, K. Blankenburg, M. A. Carbone, K. Chang, L. L. Ellis, S. Fernandez, Y. Han, G. Highnam, C. E. Hjelmen, J. R. Jack, M. Javaid, J. Jayaseelan, D. Kalra, S. Lee, L. Lewis, M. Munidasa, F. Ongeri, S. Patel, L. Perales, A. Perez, L. Pu, S. M. Rollmann, R. Ruth, N. Saada, C. Warner, A. Williams, Y.-Q. Wu, A. Yamamoto, Y. Zhang, Y. Zhu, R. R. H. Anholt, J. O. Korbel, D. Mittelman, D. M. Muzny, R. A. Gibbs, A. Barbadilla, J. S. Johnston, E. A. Stone, S. Richards, B. Deplancke, and T. F. C. Mackay, "Natural variation in genome architecture among 205 drosophila melanogaster genetic reference panel lines," *Genome Res.*, vol. 24, no. 7, pp. 1193–1208, Jul. 2014.
[26] M. Todesco, G. L. Owens, N. Bercovich, J.-S. Légaré, S. Soudi, D. O. Burge, K. Huang, K. L. Ostevik, E. B. M. Drummond, I. Imerovski,

K. Lande, M. A. Pascual-Robles, M. Nanavati, M. Jahani, W. Cheung, S. E. Staton, S. Muños, R. Nielsen, L. A. Donovan, J. M. Burke, S. Yeaman, and L. H. Rieseberg, "Massive haplotypes underlie ecotypic differentiation in sunflowers," *Nature*, vol. 584, no. 7822, pp. 602–607, Aug. 2020.

[27] R. B. Corbett-Detig, I. Said, M. Calzetta, M. Genetti, J. McBroome, N. W. Maurer, V. Petrarca, A. Della Torre, and N. J. Besansky, "Fine-Mapping complex inversion breakpoints and investigating somatic pairing in the anopheles gambiae species complex using Proximity-Ligation sequencing," *Genetics*, vol. 213, no. 4, pp. 1495–1511, Dec. 2019.

[28] N. F. Lobo, D. M. Sangaré, A. A. Regier, K. R. Reidenbach, D. A. Bretz, M. V. Sharakhova, S. J. Emrich, S. F. Traore, C. Costantini, N. J. Besansky, and F. H. Collins, "Breakpoint structure of the anopheles gambiae 2rb chromosomal inversion," *Malar. J.*, vol. 9, p. 293, Oct. 2010.

[29] T. F. C. Mackay, S. Richards, E. A. Stone, A. Barbadilla, J. F. Ayroles, D. Zhu, S. Casillas, Y. Han, M. M. Magwire, J. M. Cridland, M. F. Richardson, R. R. H. Anholt, M. Barrón, C. Bess, K. P. Blankenburg, M. A. Carbone, D. Castellano, L. Chaboub, L. Duncan, Z. Harris, M. Javaid, J. C. Jayaseelan, S. N. Jhangiani, K. W. Jordan, F. Lara, F. Lawrence, S. L. Lee, P. Librado, R. S. Linheiro, R. F. Lyman, A. J. Mackey, M. Munidasa, D. M. Muzny, L. Nazareth, I. Newsham, L. Perales, L.-L. Pu, C. Qu, M. Ràmia, J. G. Reid, S. M. Rollmann, J. Rozas, N. Saada, L. Turlapati, K. C. Worley, Y.-Q. Wu, A. Yamamoto, Y. Zhu, C. M. Bergman, K. R. Thornton, D. Mittelman, and R. A. Gibbs, "The drosophila melanogaster genetic reference panel," *Nature*, vol. 482, no. 7384, pp. 173–178, Feb. 2012.

[30] P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, and 1000 Genomes Project Analysis Group, "The variant call format and VCFtools," *Bioinformatics*, vol. 27, no. 15, pp. 2156–2158, Aug. 2011.

[31] S. v. d. Walt, S. C. Colbert, and G. Varoquaux, "The NumPy array: A structure for efficient numerical computation," *Comput. Sci. Eng.*, vol. 13, no. 2, pp. 22–30, Mar. 2011.

[32] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. Del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020.

[33] "smhasher: Automatically exported from code.google.com/p/smhasher."

[34] H. Senuma, "mmh3: Python wrapper for MurmurHash (MurmurHash3), a set of fast and robust hash functions."

[35] E. Cohen and H. Kaplan, "Summarizing data using bottom-k sketches," in *Proceedings of the twenty-sixth annual ACM symposium on Principles of distributed computing*, ser. PODC '07. New York, NY, USA: Association for Computing Machinery, Aug. 2007, pp. 225–234.

[36] ——, "Bottom-k sketches: better and more efficient estimation of aggregates," in *Proceedings of the 2007 ACM SIGMETRICS international conference on Measurement and modeling of computer systems*, ser. SIGMETRICS '07. New York, NY, USA: Association for Computing Machinery, Jun. 2007, pp. 353–354.

[37] ——, "Tighter estimation using bottom k sketches," *Proceedings VLDB Endowment*, vol. 1, no. 1, pp. 213–224, Aug. 2008.

[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[39] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: fundamental algorithms for scientific computing in python," *Nat. Methods*, vol. 17, no. 3, pp. 261–272, Mar. 2020.

[40] "hmmlearn — hmmlearn 0.2.8.post23+ge76c035 documentation," https://hmmlearn.readthedocs.io/en/latest/, accessed: 2023-1-31.

[41] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 1, no. 2, pp. 224–227, Feb. 1979.

[42] L. Hubert and P. Arabie, "Comparing partitions," *J. Classification*, vol. 2, no. 1, pp. 193–218, Dec. 1985.

[43] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Computing in Science Engineering*, vol. 9, no. 3, pp. 90–95, May 2007.

[44] R. R. Love, S. N. Redmond, M. Pombi, B. Caputo, V. Petrarca, A. Della Torre, Anopheles gambiae 1000 Genomes Consortium, and N. J. Besansky, "In silico karyotyping of chromosomally polymorphic malaria mosquitoes in the anopheles gambiae complex," *G3*, vol. 9, no. 10, pp. 3249–3262, Oct. 2019.