ON THE USE OF GENERATIVE MODELS FOR EVOLUTIONARY INFERENCE OF MALARIA VECTORS FROM GENOMIC DATA

Amelia Adibe Eneli[†,1], **Pui Chung Siu**[†,1], **Manolo F. Perez**[2,3] **Austin Burt**[3]

Matteo Fumagalli[1,4] **Sara Mathieson**[*,5]

ABSTRACT

Malaria in sub-Saharan Africa is transmitted by mosquitoes, in particular the *Anopheles gambiae* complex. Efforts to control the spread of malaria have often focused on these vectors, but relatively little is known about the relationships between populations and species in the *Anopheles* complex. Here, we first quantify the genetic structure of mosquito populations in sub-Saharan Africa using unsupervised machine learning. We then adapt and apply an innovative generative deep learning algorithm to infer the joint evolutionary history of populations sampled in Guinea and Burkina Faso, West Africa. We further develop a novel model selection approach and discover that an evolutionary model with migration fits this pair of populations better than a model without post-split migration. For the migration model, we find that our method outperforms earlier work based on summary statistics, especially in capturing population genetic differentiation. These findings demonstrate that machine learning and generative models are a valuable direction for future understanding of the evolution of malaria vectors, including the joint inference of demography and natural selection. Understanding changes in population size, migration patterns, and adaptation in hosts, vectors, and pathogens will assist malaria control interventions, with the ultimate goal of predicting nuanced outcomes from insecticide resistance to population collapse.

[1] School of Biological and Behavioural Sciences, Queen Mary University of London, Mile End Road, E1 4NS, London, United Kingdom
[2] Real Jardín Botánico, CSIC, 2 Pl. Murillo, 28014, Madrid, Spain
[3] Department of Life Sciences, Imperial College London, Silwood Park, SL5 7PY, Ascot, United Kingdom
[4] The Alan Turing Institute, 96 Euston Road, NW1 2DB, London, United Kingdom
[5] Department of Computer Science, Haverford College, 370 Lancaster Ave, Pennsylvania, 19041, USA

[†] These authors contributed equally.
[*] Corresponding author `smathieson@haverford.edu`

Keywords malaria parasite, demographic inference, generative adversarial networks, population genetics

1 Introduction

Malaria is a tropical disease that, in humans, is mainly caused by the *Plasmodium falciparum* parasite hosted by the female *Anopheles* mosquito. Infected female *Anopheles* mosquitoes introduce the parasites through their saliva into the bloodstream of the humans they bite [1]. In mild malaria cases, symptoms include fever, headaches, vomiting, and muscle aches, while in extreme cases, symptoms include jaundice, seizures, coma, and death [2]. The World Malaria Report 2024, released by the World Health Organization (WHO), indicates a rise in malaria cases globally, with an

estimated 263 million cases in 2023, up from 252 million in 2022 and 247 million in 2021 [3, 4]. According to the WHO, most malaria deaths have occurred in Africa among children under five years of age.

Malaria interventions and elimination efforts have frequently focused on vector control methods. These have been achieved through preventive measures and, where unsuccessful, curative therapies have been used to treat the disease [5]. Preventive measures include the use of rapid diagnostic tests, insect repellents, indoor residual insecticide spraying [6], insecticide-treated mosquito nets [7], and the release of the first vaccine in 2021 [8], which has now been rolled out in 19 countries in Africa [9]. Additionally, curative therapies include the use of antimalarial drugs such as chloroquine phosphate and artemisinin-based combination therapies [10]. More recently, gene drive vector control strategies are being developed that employ genetic engineering to modify the genomes of mosquito [11]. A field release has not yet been performed, but if successful it would suppress the number of female mosquitoes to promote population collapse.

Genomic surveillance of malaria vectors has gained importance with the emergence and spread of pesticide or insecticide resistance within and between mosquito populations. The extent to which resistance mutations can spread throughout the African continent is still under investigation [12]. Inference of adaptive and neutral demographic histories of mosquito populations using genomic data presents an opportunity to improve understanding. These inferences range from the estimation of past population sizes and migration rates to the identification of genetic variants under natural selection [13]. In addition to empirical analyses, mathematical models are often developed to infer the spread of resistance mutations [14] and to inform on a potential gene drive release [15]. A comprehensive understanding of the genetic history of *Anopheles* mosquitoes would inform robust surveillance and control measures against these vectors, the pathogens they spread, and would aid the prediction of their potential future demographic changes.

1.1 Genetic monitoring of malaria vectors: challenges and opportunities

Insecticide resistance in mosquitoes is a significant challenge for vector control strategies aimed at reducing mosquito-borne diseases in sub-Saharan Africa. In addition to genetic mechanisms, environmental and operational factors contribute to the development and spread of insecticide resistance in mosquito populations. Environmental factors include local climatic, ecological, agricultural, and urban conditions [16, 17], while operational factors refer to the repeated and improper use of insecticides [18]. The latter exerts strong selective pressure on mosquito populations, accelerating the development of resistance.

Genetic resistance to insecticides encompasses target-site, metabolic, cuticular, and behavioural mechanisms. Target-site resistance are changes in proteins targeted by insecticides. A key example is the knockdown mutation (kdr) in the voltage-gated sodium channel gene that confers resistance to pyrethroids and DDT. The kdr mutation occurs in two forms in African Anopheline mosquitoes: kdr-east (L1014S/L995S) and kdr-west (L1014F/L995F) [19, 20, 21]. Metabolic resistance involves changes in genes coding for detoxification enzymes, primarily cytochrome P450s, esterases, and glutathione S-transferases [22]. These enzymes can break down insecticides before they reach their target sites. Cuticular resistance refers to genetic changes in the mosquito cuticle that reduce insecticide penetration,

often associated with thickening of the cuticle or altered compositions [23]. Finally, behavioural resistance enables mosquitoes to avoid contact with insecticides, although this mechanism is still poorly understood [24, 25].

Effective resistance management requires integrated approaches, including rotating insecticides with different modes of action, using synergists to enhance insecticide efficacy, and exploring and employing alternative control methods such as biological interventions, e.g. fungi and bacteria [26]. Continuous monitoring and adaptive strategies are also essential to mitigate the impact of resistance on mosquito control efforts. In order to plan an effective control strategy, it is important to assess the relationship between genetic diversity in *Anopheles* mosquitoes and their dispersal rate and adaptive potential to environmental changes. This includes estimation of demographic parameters such as effective population size, population growth rates, and migration between populations. Genetic variation is also shaped by other geographical factors, including levels of urbanization and trade flows between cities. Understanding the interactions between demographic history, population structure, gene flow, and local selection pressures is critical to predict and manage the spread of insecticide resistance in mosquito populations.

The large-scale genomic resources generated by the MalariaGEN consortium have provided valuable information on the demographic history of *Anopheles*. Using $\partial a \partial i$ [27], coalescent-based approaches based on the site frequency spectrum, scientists inferred an expansion in populations located north of the Congo Basin that occurred between 7,000 and 25,000 years ago [28]. Such expansions could be related to the expansion of human populations as a result of agricultural advances [29]. A more recent signal of population bottlenecks, associated with the widespread use of insecticides, was also observed in some populations [28].

Other demographic inference methods, such as `SMC++` [30], have been applied to mosquito populations, specifically to the invasive species *Aedes aegypti* [31]. These results also indicate occurrences of population bottlenecks (of varying severity) and re-expansions in the recent past. Additionally, dispersal rates of malaria vectors have been inferred using neural networks [32, 33]. Due to the complexity of the *Anopheles* system and the ability of neural network-based methods to uncover subtle signals from noisy data, machine learning methods provide a way forward to understand their historical processes.

1.2 Generative models for population genetic data

Deep learning algorithms in population genetics have recently emerged as a powerful tool to address some of the most complex questions in the field [34]. Convolutional neural networks, in particular, have been used for inference of natural selection, changes in population size, variation of recombination rate, and time to the most recent common ancestor (TMRCA), among other processes [34]. Deep learning has also been shown to outperform traditional simulation-based methods such as Approximate Bayesian Computation (ABC) [35]. In the context of studies on malaria, machine learning has been deployed to detect genomic signatures of selection in vectors [36] and *Plasmodium* parasites [37].

Generative models have recently gained popularity, with the ability to create novel text, image, audio, and video from available data. For biological applications, generative models are an even more recently introduced technology, but they have already shown promising results. In population genetics specifically, simulated data are used extensively

for intuition, validation, comparison of methods, and more recently for training machine learning models. Therefore, developing methods to create more realistic simulated data is very desirable. Early generative models in population genetics have been based on generative adversarial networks (GANs) [38, 39, 40]. Yelmen and Jay provided a comprehensive review of generative models for population genetics [41].

GANs work by training two models in concert: a generator that creates synthetic examples and a discriminator that predicts whether examples are real or generated. Throughout the training process, both models ideally improve so that, in the end, the generator is creating synthetic examples that confuse the discriminator. In [39, 40, 42] the generator is a neural network, creating artificial genetic data with the same Single Nucleotide Polymorphism (SNP) patterns as the real data. This is useful for genomic privacy and downstream GWAS or similar polygenic trait analysis. In [38] the generator is based on an evolutionary model, and, during training, the goal is to estimate the parameters of this evolutionary model that create data that closely match the real data (from the perspective of the discriminator). One disadvantage of this method, called `pg-gan`, is that it produces a point estimate of demographic parameters. To better capture uncertainty, Gower *et al* [43] developed a variation that uses kernel density estimation (KDE) to estimate a distribution for each evolutionary parameter, by weighting parameter estimates by their discriminator score (i.e., how realistic are the resulting simulations).

As the discriminator produces a probability (with closer to 1 meaning more "real" and closer to 0 meaning more "fake"), this trained neural network can be used to identify regions of the input alignment with features not modelled in simulation. For example, regions of real data with a very high discriminator score may be under some form of natural selection or have some unusual recombination or mutational features. In [44], the `pg-gan` discriminator is fine-tuned using a transfer learning approach to detect various forms of natural selection.

Our motivation for using `pg-gan` is to understand and quantify the demographic history of *Anopheles gambiae* mosquitoes. Specifically, we use `pg-gan` to detect population splits, effective population size changes, exponential growth, and migration rates. Other authors have deployed `pg-gan` for the inference of demography in mosquito populations [45]. As the method was originally developed for and applied to human genetic data, our first aim was to adapt `pg-gan` to take the nuances of mosquito data into account. We then propose and apply a novel method to compare competing historical scenarios and estimate demographic parameters using GANs. GANs are used to fit a series of demographic models, then a separate neural network model is designed to discriminate between them. This network can be used to identify the most probable demographic history for real mosquito data. Finally, we define future research directions on the development and use of generative models in population genetics for pathogens and disease-vectors. Our software, called `pg-gan-mosquito`, is open-source and available at `https://github.com/mathiesonlab/pg-gan-mosquito/`.

2 Materials and Methods

2.1 Genetic data and model exploration

The *Anopheles gambiae* complex is comprised of at least eight mosquito species, five of which are major malaria vectors: *An. gambiae*, *An. coluzzii*, *An. arabiensis*, *An. melus*, and *An. merus* [46]. To assess their genetic diversity, population structure and demographic history, we analysed genomic data of samples captured in 13 sub-Saharan African countries of two species of the complex, i.e. *Anopheles gambiae*, *Anopheles coluzzii*, and three types of hybrid strains from the open access Ag1000G Phase 2 dataset [47], for a total sample size of $n = 1142$ mosquitoes. Following suggested data filtering on chromosomal arm 3L [28], we extracted more than 10 million SNPs. This full data set was used to assess population structure [48] using Uniform Manifold Approximation and Projection (UMAP) [49].

To showcase the applicability of `pg-gan-mosquito` to this system, we focused on a pair of populations, GN (Guinea) and BF (Burkina Faso), due to their genetic similarity and geographic proximity. We retrieved haplotype data from 112 *An. gambiae* samples (31 from Guinea (GN) and 81 from Burkina Faso (BF)) from the MalariaGEN database, Phase 2, following the same data processing protocol from a previous study [28]. Specifically, we considered only biallelic variants in chromosomal arms 3R and 3L that passed quality control and were located in heterochromatin states [28]. Furthermore, we filtered out SNPs within the *Gste* gene region, as this locus is a known target of selection associated with insecticide resistance [28].

For our simulations, we assumed a mutation rate of 3.5×10^{-9} per site per generation, with 11 generations per year [28], and a recombination rate of 1.45×10^{-8} per site per generation [50, 51]. We sought to compare our demographic inferences with previously estimated models [28] that incorporated only information from the joint site frequency spectrum using the software $\partial a \partial i$ [27]. From the ensemble of such previous models, we consider those with a lower Akaike Information Criterion (AIC) score with or without bidirectional migration [28]. We refer to these demographic models as *mig* and *no-mig*, respectively.

2.2 Generative model

To assess GAN training, we should observe the following patterns in loss functions and discriminator accuracy. At the start of training, the discriminator should correctly identify real data from simulations, resulting in high accuracy for both types of data. The loss of the discriminator should therefore be low and the loss for the generator high (since its generated data is not fooling the discriminator). As training progresses, the generator improves and reduces its loss, and the discriminator is progressively having difficulties distinguishing real data from simulated data. During the stable competing phase the loss functions have more or less stabilized, although there may still be minor fluctuations in the parameter choices of the generator. Ideally, at the end of the training, the discriminator accuracy is around 50% for both the real and the simulated data.

Due to the differences between human and mosquito genetic data, we adapted `pg-gan` both in terms of hyperparameters and training regimes. This included increasing the number of SNPs per region from 36 to 72, changing the gradient

descent optimizer, modifying the discriminator pre-training, and increasing the sizes of the last hidden layers. Further training details can be found in the Supporting Information. Using this modified version of `pg-ga`, we fit two different demographic models. The first model [28](*no-mig*) begins with an ancestral population of size N_I. At time T_G the ancestral population can start to grow or contract (exponentially), with a final size of N_F right before the split. The population split occurs at T_S, with population sizes N_{I1} and N_{I2} for the two populations, respectively. Both populations can undergo exponential size changes until their final sizes, N_{F1} and N_{F2}. The second model (*mig*) has the same structure, but with bidirectional migration after the split (parameter M_G).

To assess whether simulated genetic data resemble real data, we compared their distributions of commonly used summary statistics of genetic diversity and differentiation [52], as previously done [38]. These summary statistics include pairwise heterozygosity (π), Watterson's θ, site frequency spectrum, inter-SNP distance, linkage disequilibrium (LD) decay, number of haplotypes, and a measure of population genetic differentiation (F_{ST}).

3 Results

After using UMAP to assess the genetic structure in the *Anopheles* complex [48], we confirm that the samples tend to cluster closely according to their species and country of origin (Figure S1). The mainland populations are tightly clustered in both groups with hybrid forms on the periphery of the *An. coluzzii* cluster. Likewise, *An. gambiae* samples from mainland populations form a large cluster with the different populations appearing distinctly. Samples from Gabon, Uganda, and the Mayotte island form separate and distant clusters. Interestingly, we obtain clusters which are more clearly defined than in previous studies [48], possibly due to the larger number of samples analysed herein and a different UMAP parametrisation. Finally, we confirm these patterns of population structure by visual inspection of pairwise mean genetic distances between samples (Figure S3). The East African samples of Kenya and Mayotte, together with the Central African samples from Angola and Gabon, appear to be genetically distinct populations with low within-population distances. The degree to which allele sharing varies with geographical distance shows only weak support for isolation by distance in West Africa [48], a pattern consistent with a complex scenario of recent divergence and/or gene flow.

To understand the joint demographic history and divergence of the GN (Guinea) population and BF (Burkina Faso) population, we fit two different demographic models using `pg-gan-mosquito` – *no-mig* (no migration after the population split) and *mig* (migration after the population split). Figures 1 and 2 show successful training runs for the *no-mig* and *mig* models, respectively. In both cases, we observe that the generator and discriminator losses are well-matched by the end of training. For the accuracies, initially the discriminator is easily able to distinguish real from simulated data, but during training it exhibits some difficulty. At the end of training, the *no-mig* model discriminator displays high, but not perfect, accuracy, indicating an appropriate level of confusion compatible with a successful training. Figure S2 shows an example of a failed training run, where learning stops leading to a plateau in the losses and accuracies.

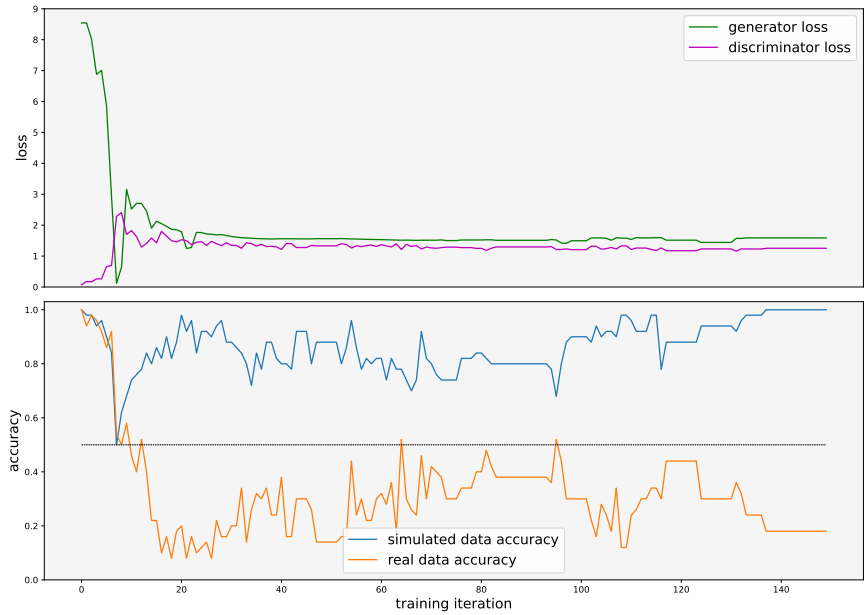

Figure 1: Training run with populations GN and BF under the *no-mig* model. The top panel shows the generator (green) and discriminator (pink) loss functions. The bottom panel shows the discriminator accuracy: blue for simulated data and orange for real data.

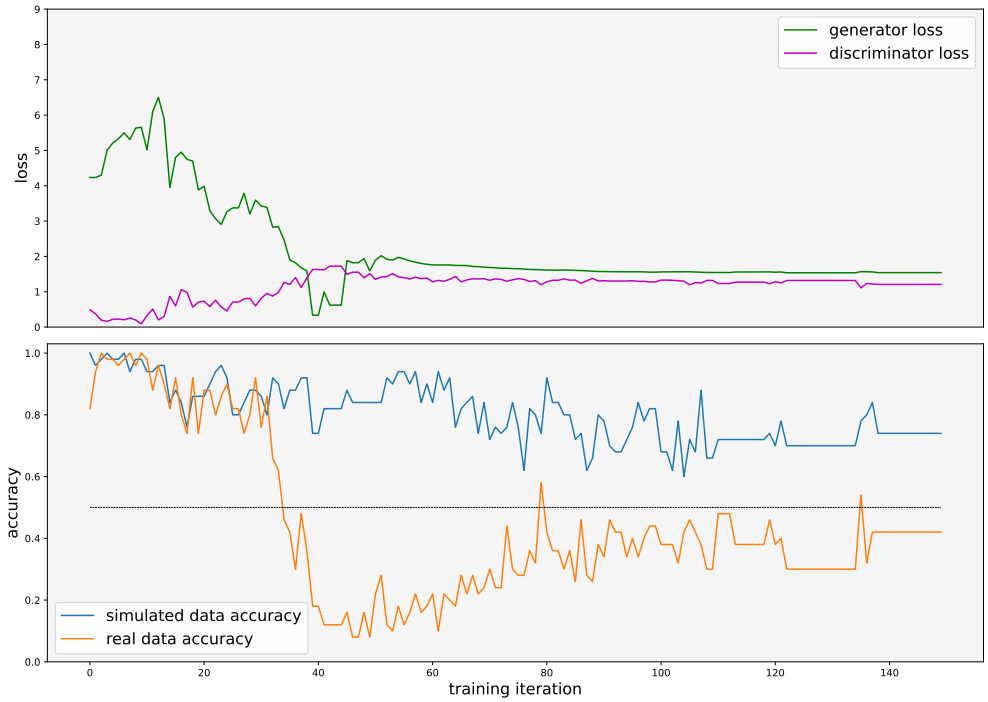

Figure 2: Training run with populations GN and BF under the *mig* model. The top panel shows the generator (green) and discriminator (pink) loss functions. The bottom panel shows the discriminator accuracy: blue for simulated data and orange for real data.

Table 1 shows point estimates for model parameters. For both models, we infer a very recent split time (T_S) with large recent population sizes (N_{F1} and N_{F2}).

NAME	description	baseline *no-mig*	`pg-gan-mosquito` *no-mig*	baseline *mig*	`pg-gan-mos`
N_I	initial ancestral size	420,646	579,516	415,254	
T_G	time of size change of ancestral population	89,506	75,249	93,341	
N_F	size of ancestral population before split	9,440,437	2,784,452	8,292,759	
T_S	time of split	2,245	3,519	11,637	
N_{I1}	initial size of population GN	18,328,570	117,287,221	2,635,696	
N_{I2}	initial size of population BF	42,062,652	62,771,288	2,748,423	
N_{F1}	final size of population GN	42,064,645	59,092,640	11,101,754	1
N_{F2}	final size of population BF	42,064,198	189,066,655	11,439,976	2
M_G	migration rate	n/a	n/a	20	

Table 1: Joint population model parameter inference from `pg-gan-mosquito`. Model parameter descriptions and baseline inference results are based on [28]. Time units are in generations, $M_G = 2N_f m$ where m is the bidirectional fractional migration rate per generation.

To visually assess the match between data simulated under the inferred parameters and real data, we generate distributions of various classical summary statistics. We also plot these distributions under the previously inferred demographic model (herein named *baseline*)[28]. To obtain each distribution, we simulate many regions under each set of parameters (for the simulated data) and sample random regions from the genome (for the real data). Note that SNPs for each region are a combined set of GN and BF data, and, when restricted to a single population, there are no segregating sites (i.e. 0 entry in the site frequency spectrum, SFS). Figures 3 and 4 show these distributions for the *no-mig* and *no-mig* model, respectively. We generally find a closer match between our results (green) and the real data (blue) vs. the *baseline* model (orange) and the real data.

To quantify how well summary statistics from simulated data match the real data, we compute Wasserstein distances between summary statistic distributions. These values are provided for both *baseline* model and our estimates using `pg-gan-mosquito` in Table 2 for the *no-mig* model, and Table 3 for the *mig* model. For the *no-mig* model we observe that the data simulated under the *baseline* model are generally closer to the real data. On the other hand, for the *mig* model, data simulated under the `pg-gan-mosquito` model are overall closer to the real data. This observation is especially true for F_{ST} which is poorly fitted by the *baseline* model, possibly due to migration rate being capped to its upper bound 20 [28]. We acknowledge that migration rates are notoriously difficult to estimate [43], and we anticipate some uncertainty in our estimates as well.

To assess which demographic model is a better fit to the data, we develop a novel machine learning-based model selection algorithm. We trained an additional neural network (with the same architecture of the `pg-gan-mosquito` discriminator) with data simulated under both models to perform a binary task. After training is complete, we feed regions of real data into the trained network to predict whether their histories are more similar to the *no-mig* or *mig* model. An accuracy curve is shown in Figure 6, where we plot the training and validation accuracy (both based on simulated data) throughout the training iterations. Both demographic models produce similar data, as demonstrated by the accuracy starting at around 50% and plateauing between 65-70%.

After running the GN/BF real data through the model selection discriminator, the fraction of real regions classified as *mig* model is 0.744, while the remaining 0.256 is classified as *no-mig*. These findings are in line with the observation

summary statistic	`pg-gan-mosquito` Wasserstein distance	baseline Wasserstein distance
GN minor allele count (SFS)	0.5825	**0.274**
GN inter-SNP distances	5.7212	**3.1334**
GN distance between SNPs	0.006	**0.0019**
GN number of haplotypes	3.547	**1.674**
GN pairwise heterozygosity (π)	1.1352	**0.2619**
GN Watterson	0.3976	**0.2493**
BF minor allele count (SFS)	0.5971	**0.4874**
BF inter-SNP distances	5.7212	**3.1334**
BF distance between SNPs	0.0026	**0.0008**
BF number of haplotypes	8.043	**5.654**
BF pairwise heterozygosity (π)	1.1361	**0.2492**
BF Watterson	0.0745	**0.0576**
GN/BF Hudson's F_{ST}	**0.0006**	0.0007

Table 2: Wasserstein distances for the *no-mig* model. The left-hand column shows common population genetic summary statistics, computed for each mosquito population separately (GN and BF) and for F_{ST} together. The middle column is the Wasserstein distance between the summary statistic distributions for the real data and data simulated under the `pg-gan-mosquito`-inferred demography. The last column is the Wasserstein distance between the summary statistic distributions for the real data and data simulated under the *b*aseline model. The smaller distance in each row is bolded, indicating the closer match to the real data.

summary statistic	`pg-gan-mosquito` Wasserstein distance	baseline Wasserstein distance
GN minor allele count (SFS)	**0.2676**	0.5402
GN inter-SNP distances	5.2128	**4.1924**
GN distance between SNPs	**0.0025**	0.0029
GN number of haplotypes	**2.157**	2.266
GN pairwise heterozygosity (π)	**0.2351**	0.2447
GN Watterson	0.3503	**0.1652**
BF minor allele count (SFS)	**0.3854**	1.6757
BF inter-SNP distances	5.2128	**4.1924**
BF distance between SNPs	**0.0012**	0.002
BF number of haplotypes	**5.037**	11.929
BF pairwise heterozygosity (π)	**0.224**	0.2332
BF Watterson	**0.0549**	0.9319
GN/BF Hudson's F_{ST}	**0.0005**	0.0164

Table 3: Wasserstein distances for the *mig* model. The left-hand column shows common population genetic summary statistics, computed for each mosquito population separately (GN and BF) and for F_{ST} together. The middle column is the Wasserstein distance between the summary statistic distributions for the real data and data simulated under the `pg-gan-mosquito`-inferred demography. The last column is the Wasserstein distance between the summary statistic distributions for the real data and data simulated under the *b*aseline model. The smaller distance in each row is bolded, indicating the closer match to the real data.

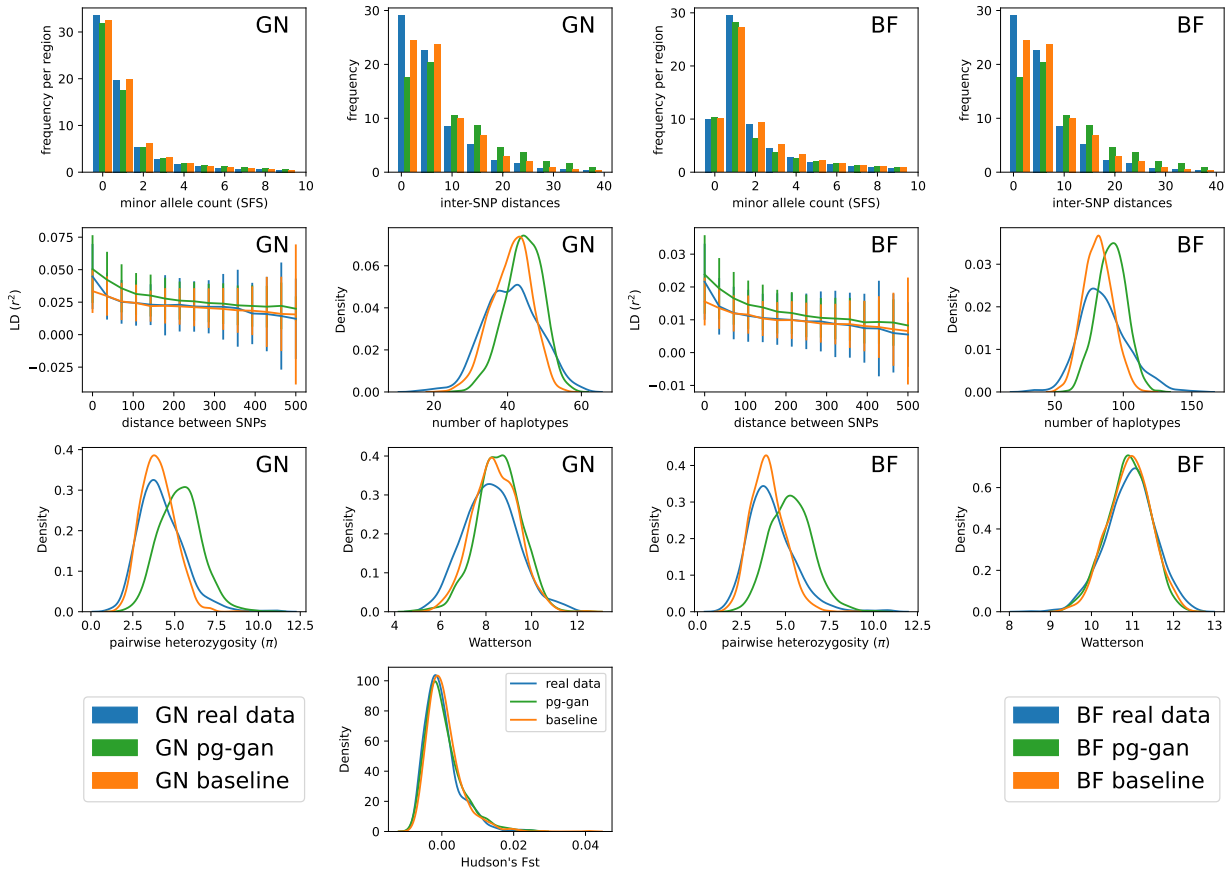

Figure 3: `pg-gan-mosquito` results for GN and BF populations, fitting a demographic history without migration (*no-mig*). Summary statistic distributions are shown for three datasets. Blue: real data from either the GN or BF population. Green: simulations under the parameters inferred by `pg-gan-mosquito` inference. Orange: simulations under the parameters inferred by $\partial a \partial i$ (baseline results from [28].

that these two populations are geographically close. Therefore, we conclude that a model with bidirectional migration explains the observed genetic data better than a model without migration.

One limitation of the current implementation of `pg-gan-mosquito` is that the input must take the form of biallelic SNPs. Although most variants in the *Anopheles* complex are biallelic, we found high levels of multiallelism – the presence of more than two alleles at a single genetic locus within a population. Although all populations show high levels of diversity on average, East African populations exhibit lower levels of genetic variation and rare alleles, possibly consistent with recent bottlenecks [48]. Hence, we sought to test whether multiallelism is more prevalent in some populations than in others. Figure 7 shows the geographic locations of analysed populations with their fractions of multiallelic SNPs.

On average, we find that approximately 23% of sites are triallelic and approximately 3% are tetrallelic. Notably, we observe that multiallelism is more prevalent in populations from the Central African region, i.e. Gabon, Equatorial Guinea, and Cameroon. However, samples from Guinea-Bissau, Mayotte, and Burkina Faso exhibit a lower proportion of tetrallelic sites, confirming previous results of lower genetic diversity in East Africa [48]. These contrasting patterns

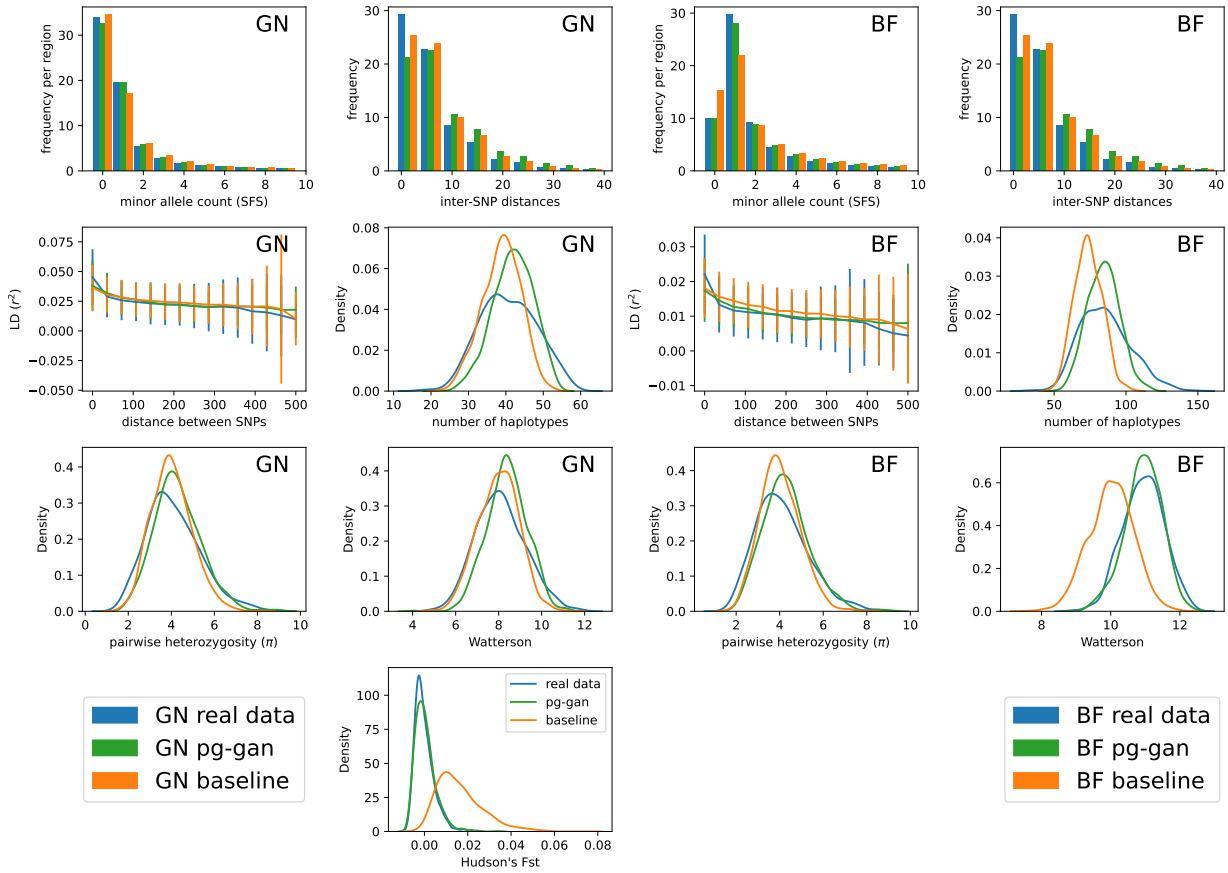

Figure 4: `pg-gan-mosquito` results for GN and BF populations, fitting a demographic history with migration. Summary statistic distributions are shown for three datasets. Blue: real data from either the GN or BF population. Green: simulations under the parameters inferred by `pg-gan-mosquito` inference. Orange: simulations under the parameters inferred by $\partial a \partial i$ (baseline results from [28]).

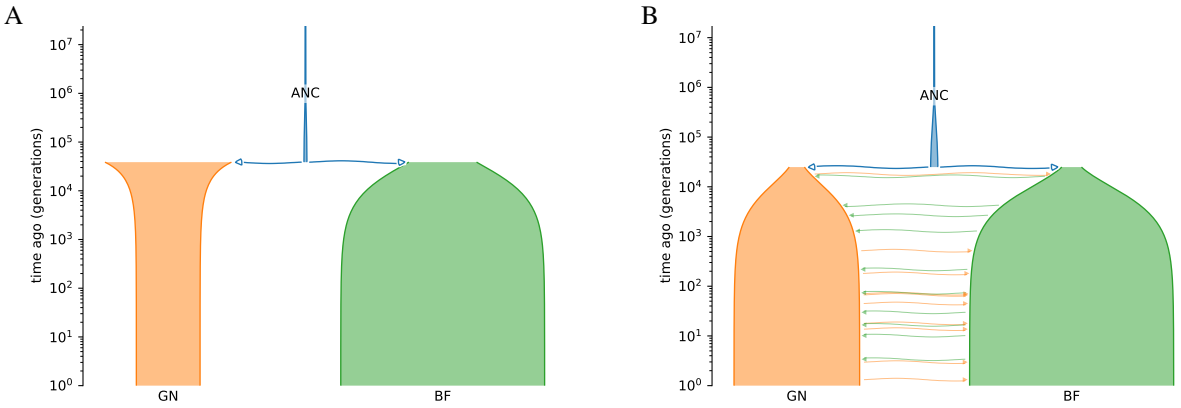

Figure 5: (left) `pg-gan-mosquito` inferred demography for GN and BF using a *no-mig* model. (right) `pg-gan-mosquito` inferred demography for GN and BF using a model with migration (the more likely history based on our model selection results).

Figure 6: Model selection accuracy curve. We train a discriminator to distinguish data simulated under two models (*no-mig* and *mig*). The accuracy on training and validation data is shown over the training iterations. Although the training is successful, the final accuracy is not very high, indicating these demographic models produce similar data.

may indicate that populations have experienced distinct demographic changes, with recent colonisation events or bottlenecks eroding rare multiallelic variants.

4 Discussion

In this study, we refine a generative adversarial network algorithm (`pg-gan` [38]) for its use with genetic data from malaria vectors. We applied this demographic inference method to a pair of mosquito populations (GN from Guinea and BF from Burkina Faso) and fit two different evolutionary models, one with post-split migration and one without. We develop a model selection method based on discriminating between datasets generated under each model, with the goal of identifying which evolutionary history best fits the real data. Our results indicate that a model that includes post-split migration between the GN and BF is most appropriate, consistent with the close geographic proximity of these two populations. Furthermore, the parameters of our fitted demographic history produce simulated data that are closer to the real data than a previously reported result [28], measured using the Wasserstein distance metric.

There are many directions for future work in this area, including scaling machine learning-based demographic inference methods to more pairs of populations or groups of populations. At the moment, we perform data filtering to retain neutral variation only, but natural selection is still likely impacting our results. Future work could incorporate a distribution of fitness effects, as in previous studies aimed to infer demographic histories in the presence of selection [53, 54, 55]. Sites under natural selection could also be identified using post-hoc analyses of outlier regions that do

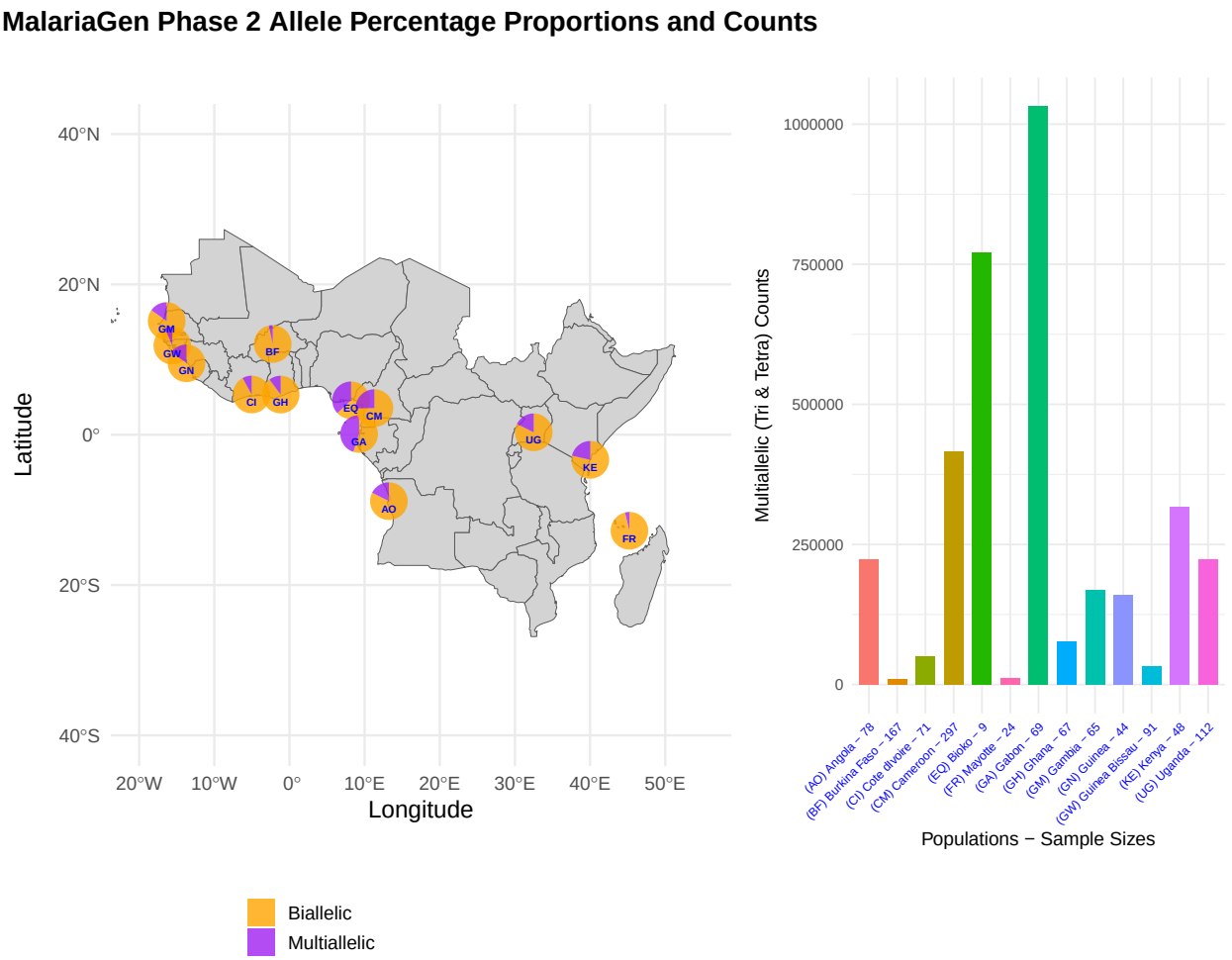

Figure 7: Ag1000G Phase 2 mosquito sample locations. Circle colours denote biallelic (orange) and multiallelic (purple) for triallelic and tetrallelic percentage proportions of each population dataset. The bar chart shows the multiallelic counts for each population.

not fit the inferred demographic history [44]. Domain adaptation and transfer learning could also be used to mitigate the source/target data mismatch in cases where models have been trained on different species/populations or datasets [56, 57]. These techniques could also reduce energy consumption by reducing simulation and training time.

Our exploration of genetic diversity in mosquito populations uncovers geographical patterns of multiallelism, with more than 25% variants being triallelic or tetrallelic on average. In particular, multiallelism seems to be highly prevalent in populations that have been sampled near water resources. Swampy vegetation, such as mangroves, may support the lifecycle development (i.e. eggs, larvae, pupae stages) of the mosquitoes as the water entities of such habitats are less disturbed by anthropogenic activities due to the pneumatophores of the surrounding mangrove vegetation. This, most likely coupled with other pollutants (e.g. agrochemicals) in the waters, could potentially promote or accelerate the development of insecticide resistant phenotypes [58, 59]. However, it remains unknown if insecticide resistance is directly correlated with multiallelism specifically (see [60, 61] for studies of insecticide resistance in mosquitoes from

different regions of Africa). Multiallelism is less pronounced in East and West African populations (relative to Central African populations), where bottlenecks or recent colonisation events may have reduced rare variation. Overall, future demographic analyses should include information on multiallelic genetic variants to further elucidate recent historical events that affected diversity levels.

Deep learning in population genetics is rapidly evolving with new architectures and implementations being proposed at a fast pace [62]. Here we used UMAP to understand the population structure of malaria vectors, but complementary approaches such as contrastive learning [63], hierarchical soft clustering [64], or other supervised non-linear approaches [65] could augment our analyses. Approaches comparing different machine learning methods could help us refine the set of evolutionary histories that explain the data well, leading to more robust conclusions about ongoing and future contact between populations.

We envisage that our new implementation, `pg-gan-moquito`, will be applicable to a wider range of species with high genetic diversity. By successfully modifying the original implementation, we demonstrate how, in general, generative models are a valuable approach for the demographic inferences of non-model species from genomic data. As recent studies showed that simulation-based approaches can be a suitable framework for the inference of evolutionary histories of malaria parasites [66], we argue that our study lays the foundation for future research directions on the joint inference of host, vectors, and pathogen co-evolution using deep learning and synthetic data. Thinking forward to intervention strategies, realistic models of mosquito dispersal and migration between populations could help us anticipate the trajectory and tempo of insecticide resistance. A fitted evolutionary history (such as what is produced by `pg-gan-mosquito`) provides a null model for selection scans, which could illuminate which genes are contributing to insecticide resistance. Finally, knowledge of relative effective population sizes in different geographic regions can help us predict the effectiveness of gene drive release strategies and the pace of population collapse.

Acknowledgments

SM is funded in part by a National Institutes of Health (NIH) grant R15HG011528. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This work was supported by a Natural Environment Research Council (NERC) NE/X009637/1 grant to MF in collaboration with SM and AB.

References

[1] Danny A Milner, Jr. Malaria pathogenesis. *Cold Spring Harb. Perspect. Med.*, 8(1):a025569, January 2018.

[2] Lynne S Garcia. Malaria. *Clin. Lab. Med.*, 30(1):93–129, March 2010.

[3] World Health Organization. World malaria report 2024. Web page, 2024.

[4] Jeanne Rini Poespoprodjo, Nicholas M Douglas, Daniel Ansong, Steven Kho, and Nicholas M Anstey. Malaria. *Lancet*, 402(10419):2328–2345, December 2023.

[5] Margaret A Phillips, Jeremy N Burrows, Christine Manyando, Rob Hooft van Huijsduijnen, Wesley C Van Voorhis, and Timothy N C Wells. Malaria. *Nat. Rev. Dis. Primers*, 3:17050, August 2017.

[6] Zachary Munn, Jennifer C Stone, Timothy Hugh Barker, Carrie Price, Danielle Pollock, Alinune Nathanael Kabaghe, John E Gimnig, and Jennifer C Stevenson. Residual insecticide surface treatment for preventing malaria: a systematic review protocol. *Syst. Rev.*, 12(1):89, June 2023.

[7] Joseph Pryce, Marty Richardson, and Christian Lengeler. Insecticide-treated nets for preventing malaria. *Cochrane Database Syst. Rev.*, 11:CD000363, November 2018.

[8] Heidi Ledford. Malaria vaccine booster prolongs protection. *Nature*, September 2022.

[9] World Health Organization. Malaria vaccines (rts,s and r21). Web page, 2025.

[10] Jair L Siqueira-Neto, Kathryn J Wicht, Kelly Chibale, Jeremy N Burrows, David A Fidock, and Elizabeth A Winzeler. Antimalarial drug discovery: progress and approaches. *Nat. Rev. Drug Discov.*, 22(10):807–826, October 2023.

[11] Tony Nolan. Control of malaria-transmitting mosquitoes using gene drives. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 376(1818):20190803, February 2021.

[12] Penelope A Hancock, Amy Lynd, Antoinette Wiebe, Maria Devine, John Essandoh, Francis Wat'senga, Emile Z Manzambi, Fiacre Agossa, Martin J Donnelly, David Weetman, and Catherine L Moyes. Modelling spatiotemporal trends in the frequency of genetic mutations conferring insecticide target-site resistance in african mosquito malaria vector species. *BMC Biol.*, 20(1):46, February 2022.

[13] Xiaoheng Cheng and Matthias Steinrücken. Population genomic scans for natural selection and demography. *Annu. Rev. Genet.*, 58(1):319–339, November 2024.

[14] Penelope A Hancock, Amy Lynd, Antoinette Wiebe, Maria Devine, John Essandoh, Francis Wat'senga, Emile Z Manzambi, Fiacre Agossa, Martin J Donnelly, David Weetman, and Catherine L Moyes. Modelling spatiotemporal trends in the frequency of genetic mutations conferring insecticide target-site resistance in african mosquito malaria vector species. *BMC Biol.*, 20(1):46, February 2022.

[15] John B Connolly, Austin Burt, George Christophides, Abdoulaye Diabate, Tibebu Habtewold, Penelope A Hancock, Anthony A James, Jonathan K Kayondo, Dickson Wilson Lwetoijera, Alphaxard Manjurano, Andrew R McKemey, Michael R Santos, Nikolai Windbichler, and Filippo Randazzo. Considerations for first field trials of low-threshold gene drive for malaria vector control. *Malar. J.*, 23(1):156, May 2024.

[16] Theresia Estomih Nkya, Idir Akhouayri, William Kisinza, and Jean-Philippe David. Impact of environment on mosquito response to pyrethroid insecticides: facts, evidences and prospects. *Insect biochemistry and molecular biology*, 43(4):407–416, 2013.

[17] Henry F Owusu, Nakul Chitnis, and Pie Müller. Insecticide susceptibility of anopheles mosquitoes changes in response to variations in the larval environment. *Scientific reports*, 7(1):3667, 2017.

[18] Neil Philip Hobbs, David Weetman, and Ian Michael Hastings. Insecticide resistance management strategies for public health control of mosquitoes exhibiting polygenic resistance: A comparison of sequences, rotations, and mixtures. *Evolutionary Applications*, 16(4):936–959, 2023.

[19] Linda Grigoraki, Ruth Cowlishaw, Tony Nolan, Martin Donnelly, Gareth Lycett, and Hilary Ranson. CRISPR/Cas9 modified an. gambiae carrying kdr mutation L1014F functionally validate its contribution in insecticide resistance and combined effect with metabolic enzymes. *PLoS Genet.*, 17(7):e1009556, July 2021.

[20] Sharon Mwagira-Maina, Steven Runo, Lucy Wachira, Stanley Kitur, Sarah Nyasende, Brigid Kemei, Eric Ochomo, Damaris Matoke-Muhia, Charles Mbogo, and Luna Kamau. Genetic markers associated with insecticide resistance and resting behaviour in anopheles gambiae mosquitoes in selected sites in kenya. *Malar. J.*, 20(1):461, December 2021.

[21] Pierre Fongho Suh, Emmanuel Elanga-Ndille, Magellan Tchouakui, Maurice Marcel Sandeu, Darus Tagne, Charles Wondji, and Cyrille Ndo. Impact of insecticide resistance on malaria vector competence: a literature review. *Malar. J.*, 22(1):19, January 2023.

[22] Eunice Oluwatobiloba Adedeji, Olubanke Olujoke Ogunlana, Segun Fatumo, Thomas Beder, Yvonne Ajamma, Rainer Koenig, and Ezekiel Adebiyi. Anopheles metabolic proteins in malaria transmission, prevention and control: a review. *Parasit. Vectors*, 13(1):465, September 2020.

[23] Or Wood, S Hanrahan, M Coetzee, Ll Koekemoer, and Bd Brooke. Cuticle thickening associated with pyrethroid resistance in the major malaria vector anopheles funestus. *Parasit. Vectors*, 3(1):67, August 2010.

[24] Michelle L Gatton, Nakul Chitnis, Thomas Churcher, Martin J Donnelly, Azra C Ghani, H Charles J Godfray, Fred Gould, Ian Hastings, John Marshall, Hilary Ranson, Mark Rowland, Jeff Shaman, and Steve W Lindsay. The importance of mosquito behavioural adaptations to malaria control in africa. *Evolution*, 67(4):1218–1230, April 2013.

[25] David Carrasco, Thierry Lefèvre, Nicolas Moiroux, Cédric Pennetier, Fabrice Chandre, and Anna Cohuet. Behavioural adaptations of mosquito vectors to insecticide control. *Current opinion in insect science*, 34:48–54, 2019.

[26] Layla Kamareddine. The biological control of the malaria vector. *Toxins*, 4(9):748–767, 2012.

[27] Ryan N Gutenkunst, Ryan D Hernandez, Scott H Williamson, and Carlos D Bustamante. Inferring the joint demographic history of multiple populations from multidimensional snp frequency data. *PLoS genetics*, 5(10):e1000695, 2009.

[28] Anopheles gambiae 1000 Genomes Consortium et al. Genetic diversity of the african malaria vector anopheles gambiae. *Nature*, 552(7683):96, 2017.

[29] Sen Li, Carina Schlebusch, and Mattias Jakobsson. Genetic variation reveals large-scale population expansion and migration during the expansion of bantu-speaking peoples. *Proceedings of the Royal Society B: Biological Sciences*, 281(1793):20141448, 2014.

[30] Jonathan Terhorst, John A Kamm, and Yun S Song. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature genetics*, 49(2):303–309, 2017.

[31] Tyler V Kent, Daniel R Schrider, and Daniel R Matute. Demographic history, genetic load, and the efficacy of selection in the globally invasive mosquito aedes aegypti. *Genome Biology and Evolution*, 17(4):evaf066, 2025.

[32] Chris CR Smith, Silas Tittes, Peter L Ralph, and Andrew D Kern. Dispersal inference from population genetic variation using a convolutional neural network. *Genetics*, 224(2):iyad068, 2023.

[33] Chris CR Smith and Andrew D Kern. dispersenn2: a neural network for estimating dispersal distance from georeferenced polymorphism data. *BMC bioinformatics*, 24(1):385, 2023.

[34] Kevin Korfmann, Oscar E Gaggiotti, and Matteo Fumagalli. Deep learning in population genetics. *Genome Biol. Evol.*, 15(2), February 2023.

[35] Théophile Sanchez, Jean Cury, Guillaume Charpiat, and Flora Jay. Deep learning for population size history inference: Design, comparison and combination with approximate bayesian computation. *Molecular Ecology Resources*, 21(8):2645–2660, 2021.

[36] Alexander T Xue, Daniel R Schrider, Andrew D Kern, and Ag1000g Consortium. Discovery of ongoing selective sweeps within anopheles mosquito populations using deep learning. *Mol. Biol. Evol.*, 38(3):1168–1183, March 2021.

[37] Wouter Deelder, Ernest Diez Benavente, Jody Phelan, Emilia Manko, Susana Campino, Luigi Palla, and Taane G Clark. Using deep learning to identify recent positive selection in malaria parasite sequence data. *Malaria journal*, 20(1):270, 2021.

[38] Zhanpeng Wang, Jiaping Wang, Michael Kourakos, Nhung Hoang, Hyong Hark Lee, Iain Mathieson, and Sara Mathieson. Automatic inference of demographic parameters using generative adversarial networks. *Molecular ecology resources*, 21(8):2689–2705, 2021.

[39] Burak Yelmen, Aurélien Decelle, Linda Ongaro, Davide Marnetto, Corentin Tallec, Francesco Montinaro, Cyril Furtlehner, Luca Pagani, and Flora Jay. Creating artificial human genomes using generative neural networks. *PLoS genetics*, 17(2):e1009303, 2021.

[40] Burak Yelmen, Aurélien Decelle, Leila Lea Boulos, Antoine Szatkownik, Cyril Furtlehner, Guillaume Charpiat, and Flora Jay. Deep convolutional and conditional neural networks for large-scale genomic data generation. *PLoS Computational Biology*, 19(10):e1011584, 2023.

[41] Burak Yelmen and Flora Jay. An overview of deep generative models in functional and evolutionary genomics. *Annual Review of Biomedical Data Science*, 6(1):173–189, 2023.

[42] Antoine Szatkownik, Cyril Furtlehner, Guillaume Charpiat, Burak Yelmen, and Flora Jay. Latent generative modeling of long genetic sequences with gans. *bioRxiv*, pages 2024–08, 2024.

[43] Graham Gower, Pablo Iáñez Picazo, Finn Lindgren, and Fernando Racimo. Inference of population genetics parameters using discriminator neural networks: an adversarial monte carlo approach. *bioRxiv*, pages 2023–04, 2023.

[44] Rebecca Riley, Iain Mathieson, and Sara Mathieson. Interpreting generative adversarial networks to infer natural selection from genetic data. *Genetics*, 226(4):iyae024, 2024.

[45] Scott T Small, Carlo Costantini, N'Fale Sagnon, Moussa W Guelbeogo, Scott J Emrich, Andrew D Kern, Michael C Fontaine, and Nora J Besansky. Standing genetic variation and chromosome differences drove rapid ecotype formation in a major malaria mosquito. *Proceedings of the National Academy of Sciences*, 120(11):e2219835120, 2023.

[46] Jacques Derek Charlwood. *The ecology of malaria vectors*. CRC Press, London, England, July 2019.

[47] The Anopheles gambiae 1000 Genomes Consortium. Ag1000g phase 2 ar1 data release. malariagen. http://www.malariagen.net/data-package/ag1000g-phase-2-ar1/. Accessed: 2024-03-01.

[48] Anopheles gambiae 1000 Genomes Consortium. Genome variation and population structure among 1142 mosquitoes of the african malaria vector species anopheles gambiae and anopheles coluzzii. *Genome Res.*, 30(10):1533–1546, October 2020.

[49] Alex Diaz-Papkovich, Luke Anderson-Trocmé, and Simon Gravel. A review of UMAP in population genetics. *J. Hum. Genet.*, 66(1):85–91, January 2021.

[50] Jeffrey R Adrion, Christopher B Cole, Noah Dukler, Jared G Galloway, Ariella L Gladstein, Graham Gower, Christopher C Kyriazis, Aaron P Ragsdale, Georgia Tsambos, Franz Baumdicker, et al. A community-maintained standard library of population genetic models. *elife*, 9:e54967, 2020.

[51] M Elise Lauterbur, Maria Izabel A Cavassim, Ariella L Gladstein, Graham Gower, Nathaniel S Pope, Georgia Tsambos, Jeffrey Adrion, Saurabh Belsare, Arjun Biddanda, Victoria Caudill, et al. Expanding the stdpopsim species catalog, and lessons learned for realistic genome simulations. *Elife*, 12:RP84874, 2023.

[52] Rasmus Nielsen. Molecular signatures of natural selection. *Annu. Rev. Genet.*, 39(1):197–218, 2005.

[53] Parul Johri, Kellen Riall, Hannes Becher, Laurent Excoffier, Brian Charlesworth, and Jeffrey D Jensen. The impact of purifying and background selection on the inference of population history: problems and prospects. *Molecular biology and evolution*, 38(7):2986–3003, 2021.

[54] Parul Johri, Susanne P Pfeifer, and Jeffrey D Jensen. Developing an evolutionary baseline model for humans: jointly inferring purifying selection with population history. *Molecular biology and evolution*, 40(5):msad100, 2023.

[55] Jacob I Marsh and Parul Johri. Biases in arg-based inference of historical population size in populations experiencing selection. *Molecular Biology and Evolution*, 41(7):msae118, 2024.

[56] Kerry Cobb and Megan L Smith. The reasonable effectiveness of domain adaptation for inference of introgression. January 2025.

[57] Sandipan Paul Arnab, Andre Luiz Campelo Dos Santos, Matteo Fumagalli, and Michael DeGiorgio. Efficient detection and characterization of targets of natural selection using transfer learning. *Mol. Biol. Evol.*, 42(5), April 2025.

[58] Stephanie L Richards, Brian D Byrd, Michael H Reiskind, and Avian V White. Assessing insecticide resistance in adult mosquitoes: Perspectives on current methods. *Environ. Health Insights*, 14:117863022095279, January 2020.

[59] Adrian C Williams and Lisa J Hill. Nicotinamide and demographic and disease transitions: Moderation is best. *International Journal of Tryptophan Research*, 12:1178646919855940, 2019.

[60] Chris S Clarkson, Alistair Miles, Nicholas J Harding, Andrias O O'Reilly, David Weetman, Dominic Kwiatkowski, Martin J Donnelly, and Anopheles gambiae 1000 Genomes Consortium. The genetic architecture of target-site resistance to pyrethroid insecticides in the african malaria vectors anopheles gambiae and anopheles coluzzii. *Molecular ecology*, 30(21):5303–5317, 2021.

[61] Vincent Corbel and Raphael N'Guessan. Distribution, mechanisms, impact and management of insecticide resistance in malaria vectors: a pragmatic review. In *Anopheles mosquitoes-New insights into malaria vectors*. IntechOpen, 2013.

[62] Sjoerd van den Belt and Nikolaos Alachiotis. Fast and accurate deep learning scans for signatures of natural selection in genomes using FASTER-NN. *Commun. Biol.*, 8(1):58, January 2025.

[63] Filip Thor and Carl Nettelblad. Dimensionality reduction of genetic data using contrastive learning. *bioRxiv*, 2025.

[64] Klara Elisabeth Burger, Solveig Klepper, Ulrike von Luxburg, and Franz Baumdicker. Inferring ancestry with the hierarchical soft clustering approach tanglegen. *Genome Res.*, 34(12):2244–2255, December 2024.

[65] Xinghu Qin, Charleston W K Chiang, and Oscar E Gaggiotti. KLFDAPC: a supervised machine learning approach for spatial genetic structure analysis. *Brief. Bioinform.*, 23(4), July 2022.

[66] Margaux J M Lefebvre, Fanny Degrugillier, Céline Arnathau, Gustavo A Fontecha, Oscar Noya, Sandrine Houzé, Carlo Severini, Bruno Pradines, Antoine Berry, Jean-François Trape, Fabian E Sáenz, Franck Prugnolle, Michael C Fontaine, and Virginie Rougeron. Genomic exploration of the journey of plasmodium vivax in latin america. *PLoS Pathog.*, 21(1):e1012811, January 2025.

Data Accessibility and Benefit-Sharing

Mosquito data for the project can be accessed through the MalariaGEN project `https://www.malariagen.net/`, Phase 2 data. Our software, `pg-gan-mosquito` is open-source and available on GitHub: `https://github.com/mathiesonlab/pg-gan-mosquito/`. We also provide instructions for how to use `pg-gan-mosquito`, benefitting researchers of other species who are interested in generative models, demographic inference, and evolutionary model selection.

Author contributions

Conceptualization: MF, SM. Data Curation: AE, PS, MF. Formal Analysis: AE, PS, SM. Funding Acquisition: MF, SM, AB. Investigation: AE, PS, MF, SM. Methodology: PS, MF, SM. Project Administration: MF. Resources: AB. Software: PS, MF, SM. Supervision: MF. Validation: SM. Visualization: AE. Writing – Original Draft Preparation: AE, MF, SM. Writing – Review & Editing: AE, PS, MP, MF, SM.

Supplementary Material

On the use of generative models for evolutionary inference of malaria vectors from genomic data

Amelia Adibe Eneli[†,1], Pui Chung Siu[†,1] Manolo F. Perez[2,3], Austin Burt[3], Matteo Fumagalli[1,4], Sara Mathieson[*,5]

[1] School of Biological and Behavioural Sciences, Queen Mary University of London, Mile End Road, E1 4NS, London, United Kingdom
[2] Real Jardín Botánico, CSIC, 2 Pl. Murillo, 28014, Madrid, Spain
[3] Department of Life Sciences, Imperial College London, Silwood Park, SL5 7PY, Ascot, United Kingdom
[4] The Alan Turing Institute, 96 Euston Road, NW1 2DB, London, United Kingdom
[5] Department of Computer Science, Haverford College, 370 Lancaster Ave, Pennsylvania, 19041, USA
[†] These authors contributed equally.
[*] Corresponding author `smathieson@haverford.edu`

Training modifications for mosquito genetic data

To adapt `pg-gan` for mosquito genetic data, we made several modifications to the hyper-parameters and training procedure. We used the `AdamW` optimizer instead of the regular `Adam` optimizer for more robust convergence. We changed the number of units in the first fully connected layer to 160. During pre-training (an initial phase of parameter exploration to make sure we are starting from a successful discriminator), we set the dropout rate to 0.5 and the learning rate to 1×10^{-3}, and during the main training loop we used dropout 0.8 and a learning rate of 25×10^{-6}. Since mosquitoes have a higher genetic variant density than humans, we increased the number of SNPs per region from 36 to 72. Since we have a different sample size in each population, we change the permutation-invariant collapsing function from `sum` to `mean`.

If a successful training run is not observed (see Figure S2 for an example) we created several debugging strategies. During pre-training, the goal is to try random parameter combinations to find an initial configuration that produces simulated data that is easily distinguishable from real data. If the number of pre-training iterations is reached without reaching a sufficient discriminator accuracy, we can a) increase the parameter search space, b) increase the model capacity by adding more filters or hidden units, and/or c) lower the dropout rate (while being aware of over-fitting). If the discriminator quickly obtains a high accuracy, the dropout rate can be increased, or force pre-training to continue.

During training, if the discriminator is over-fitting (indicated by very high accuracy) we can reduce the number of training iterations. `AdamW` also reduces the magnitude of the weights which helps with over-fitting. If the discriminator is under-fitting, we increase the model capacity and/or lower dropout rate. Finally, if the discriminator accuracy remains high after pre-training, this could mean the evolutionary model is not expressive enough to create simulations matching the real data. A different or more complex evolutionary model is likely needed.

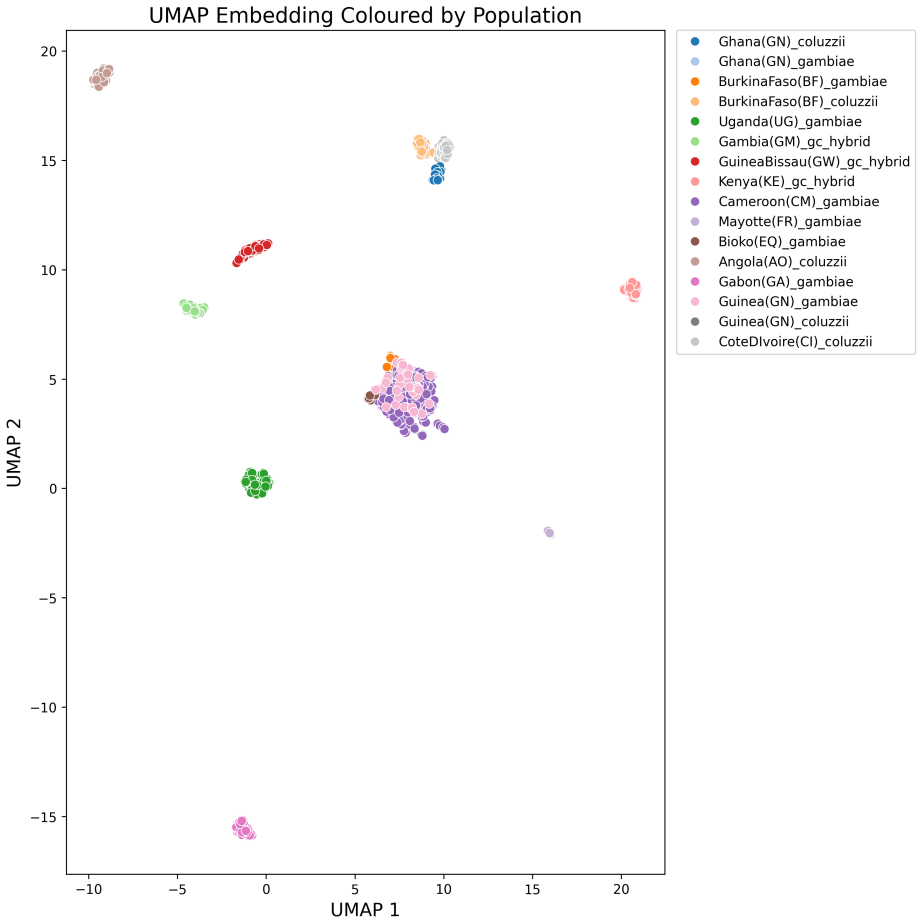

Figure S1: Population structure analysis of mosquitoes using the UMAP technique. Each mosquito represented by a marker. Biallelic SNP data from euchromatic regions of Chromosome arm 3L were projected onto two components denoted as UMAP 1 and UMAP 2.

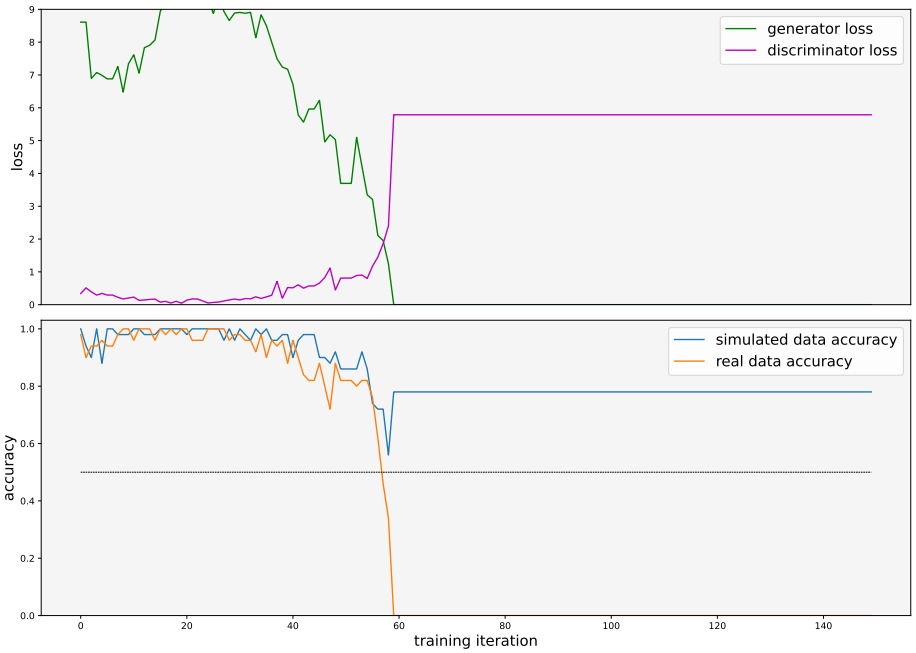

Figure S2: `pg-gan` training run with populations GN and BF under the *mig* model, where training was unsuccessful. The top panel shows the generator (green) and discriminator (pink) loss functions. The bottom panel shows the discriminator accuracy: blue for simulated data and orange for real data. Around training iteration 60, learning stops and the model weights do not change, leading to a plateau in both losses and both accuracies.

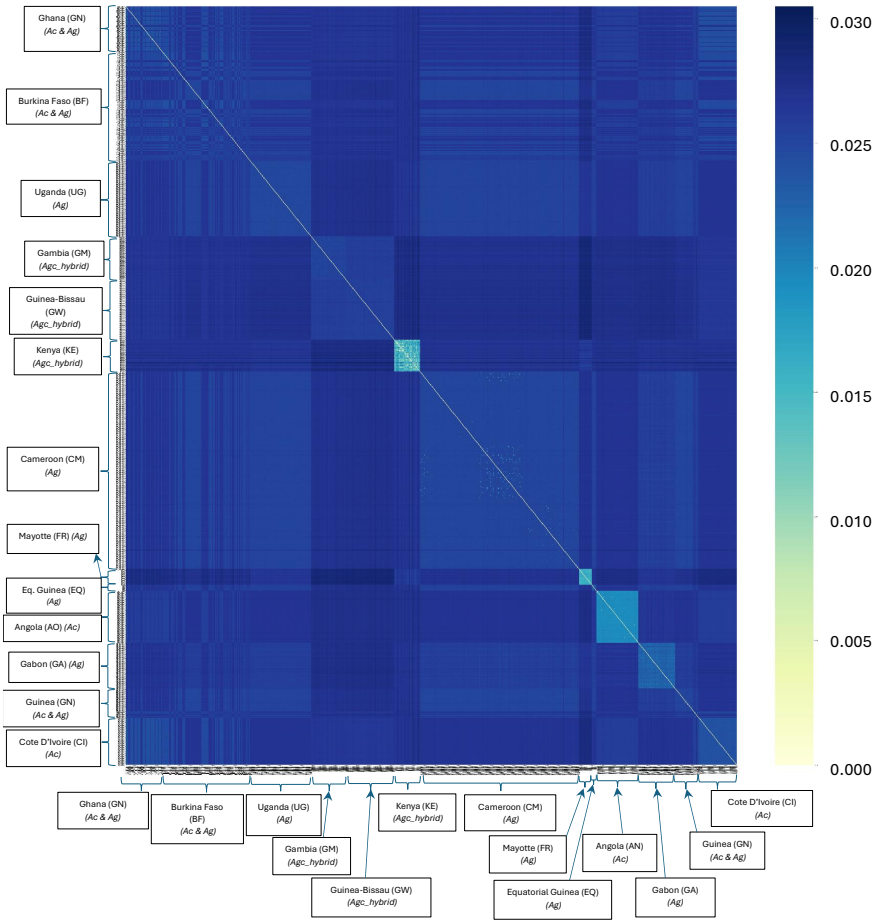

Figure S3: Heatmap analysis of genetic distances between sampled mosquito populations. Colour shades represent the degree of genetic similarities or differences between pairs of mosquito populations.