Enhancing Detection of Polygenic Adaptation: A Comparative Study of Machine Learning and Statistical Approaches Using Simulated Evolve-and-Resequence Data

Cosima Caliendo[1]*, Susanne Gerber[1], Markus Pfenninger[2,3,4]

[1]Institute of Human Genetics; University Medical Center; Johannes Gutenberg University Mainz, Germany [2]Departument of Molecular Ecology, Senckenberg Biodiversity and Climate Research Centre, Frankfurt am Main, Germany [3]LOEWE Centre for Translational Biodiversity Genomics, Senckenberg Biodiversity and Climate Research Centre, Frankfurt am Main, Germany [4]Institute for Molecular and Organismic Evolution, Johannes Gutenberg University, Mainz, Germany

*corresponding author: caliendo@uni-mainz.de (CC)

Abstract

Detecting signals of polygenic adaptation remains a significant challenge in evolutionary biology, as traditional methods often struggle to identify the associated subtle, multi-locus allele-frequency shifts. Here, we introduced and tested several novel approaches combining machine learning techniques with traditional statistical tests to detect polygenic adaptation patterns. We implemented a Naive Bayesian Classifier (NBC) and One-Class Support Vector Machines (OCSVM), and compared their performance against the Fisher's Exact Test (FET). Furthermore, we combined machine learning and statistical models (OCSVM-FET and NBC-FET), resulting in 5 competing approaches. Using a simulated data set based on empirical evolve-and-resequencing *Chironomus riparius* genomic data, we evaluated methods across evolutionary scenarios, varying in generations and numbers of loci under selection. Our results demonstrate that the combined OCSVM-FET approach consistently outperformed competing methods, achieving the lowest false positive rate, highest area under the curve, and high accuracy. The performance peak aligned with the late dynamic phase of adaptation, highlighting the method's sensitivity to ongoing selective processes and thus for experimental approaches. Furthermore, we emphasize the critical role of parameter tuning, balancing biological assumptions with methodological rigor. Our approach thus offers a powerful tool for detecting polygenic adaptation in pool sequencing data particularly from evolve-and-resequence experiments.

Author's Summary

Organisms often adapt to environmental changes through polygenic adaptation - a process where multiple genes collectively contribute to evolutionary change. However, detecting these small shifts spread across multiple genes has been a persistent challenge for researchers. We developed new computational methods that combine machine learning with traditional statistical approaches to better detect these subtle genetic changes. Using data from a laboratory evolution experiment with the freshwater midge *Chironomus riparius*, we tested five different approaches to identify genes under selection. Our results showed that combining the machine learning technique One Class Support Vector Machines with a traditional statistical test (Fisher's Exact Test) was particularly effective at identifying genes involved in adaptation. This combined approach excelled specifically in detecting ongoing adaptive changes while avoiding false positives. Our method provides a reliable tool for researchers studying evolutionary adaptation, particularly in laboratory evolution experiments where populations are tracked over multiple generations. This advancement facilitates improved understanding of how organisms adapt to new environments, which is increasingly important in the context of rapid environmental changes.

Introduction

Rapid phenotypic adaptation is a key phenomenon in nature, allowing organisms to quickly adjust to changing environments [1]. This process plays a crucial role in species survival and diversification, particularly in the face of anthropogenic changes and climate fluctuations [2][3]. Rapid adaptation can occur through various mechanisms, including selection on standing genetic variation, *de novo* mutations, and epigenetic changes [4]. While some adaptive traits are controlled by single genes with large effects, many phenotypic changes involve selection acting on multiple genes simultaneously (polygenic adaptation) [5]. This polygenic nature of adaptation often makes it challenging to detect and measure the individual contributions of genes under selection [6]. The speed and extent of these adaptive phenotypic changes can vary widely across taxa and environmental contexts, from subtle alterations in gene expression to dramatic morphological shifts [7]. For instance, studies have documented rapid adaptive responses in Galápagos finches during drought periods [8] or wing size adaption in flies [9]. Beyond these discrete adaptive events, populations can also exhibit so-called adaptive tracking, where they continuously evolve in response to fluctuating environmental conditions, as demonstrated in both natural and experimental *Drosophila* and *Chironomus* populations [10] [11]. Understanding the genetic basis of rapid adaptation is not only of fundamental scientific interest but also has important societal implications for nature conservation biology, agriculture, climate change and medicine [12].

In recent years, the advent of high-throughput sequencing technologies has enabled researchers to generate vast amounts of genomic data, presenting both opportunities and challenges for detecting signatures of adaptation [13]. These technologies have revolutionized our ability to study genetic variation at unprecedented scales, from single nucleotide polymorphisms (SNPs) to structural variants across entire genomes [14]. This wealth of data has led to the development of new statistical methods and bioinformatic tools for identifying genomic regions under selection [15]. However, the sheer volume and complexity of genomic data also present significant analytical challenges, particularly in distinguishing true signals of molecular adaptation from background noise [16].

Pool sequencing (Pool-Seq) has emerged as a cost-effective approach for population-level studies, offering a means to efficiently survey genetic variation across large numbers of individuals [17][18]. This method involves sequencing pooled DNA samples from multiple individuals within a population, providing estimates of allele frequencies at a fraction of the cost of individual sequencing [20]. Pool-Seq has been successfully applied in various contexts, e.g. studying local adaptation in *Drosophila* [19], in particular in experimental settings. However, it also introduces complexities in data analysis, particularly in identifying haplotype blocks and rare variants [20].

While our understanding of rapid phenotypic adaptation has grown, the underlying molecular dynamics and genomic architectures driving these adaptations in many species remain an active area of research. Significant progress has been made in model organisms like *Drosophila* and *Chironomus*, providing valuable insights into the genetic

basis of adaptation [21][22]. These studies have revealed complex patterns of polygenic adaptation, where multiple loci contribute to adaptive traits [23].

Despite these advances, significant challenges remain in the field. First, the broader application of findings from model organisms to diverse taxa remains difficult, as genetic architectures and selective pressures can vary substantially across species. Second, while some methods have been developed to detect both hard and soft selective sweeps in individual sequencing data [24] [25], traditional approaches for detecting selection have largely focused on identifying strong selective sweeps or large-effect loci [26], which represent only a probably minor subset of adaptive processes. Furthermore, these existing methods, including those for soft sweeps, are primarily designed for individual sequencing data and cannot be directly applied to Pool-Seq data, which is increasingly used in population genomics studies. These approaches often fail to capture polygenic adaptation, where selection acts on many loci simultaneously, each contributing small effects to the phenotype [27]. This limitation is particularly problematic for time series data, where subtle but coordinated allele frequency shifts across multiple loci may indicate important adaptive responses. Therefore, there is a pressing need for novel data analysis approaches that can effectively detect and characterize polygenic selection, especially in temporally sampled genomic data.

To date, most researchers used traditional statistical approaches such as Fisher's Exact Test (FET) or it's extension, Cochran–Mantel–Haenszel test, to identify significantly changed allele frequencies after a selective period [28][21][29][30][9] in combination with several tools like PoPoolation2 [31], SNVer [32], or expectation-maximization (EM) algorithms [33] that are mainly capable of identifying candidate SNPs in Pool-Seq data. While these approaches have proven to be useful, yet they come with certain limitations.

For instance, tools like PoPoolation2 [31] offer flexibility in analysis scale, from predefined windows to base-by-base calculations. However, effectively detecting polygenic adaptation patterns is still challenging. Large windows may overlook subtle changes, while narrow windows or base-by-base analyses might overemphasize small fluctuations. In the context of polygenic adaptation, where multiple loci of small effect contribute to a trait, the optimal scale of analysis is not always clear. Base-by-base comparisons of metrics like Fst, while detailed, may not capture the coordinated shifts across multiple loci that characterize polygenic adaptation [6][34]. Furthermore, statistical tests like FET almost always lead to a *p-value inflation*, which occurs when the observed distribution of p-values from a statistical test deviates strongly from the expected uniform distribution under the null hypothesis, with an excess of low p-values. This mainly leads to an increased false-positive rate (Type I errors) and thus to an overestimation of the significance of genetic association. Choosing a stringent significance thresholds for FET, on the other side, may bias results towards the identification of hard sweeps or punctual mutations, potentially missing important polygenic adaptation signals characteristic. Some tools, such as SNVer require a predetermined sequencing error rate [32]. However, genomic-position specific variations may lead to poor estimates of such error rates, thereby introducing additional uncertainty. EM algorithms for estimating allele frequencies face reduced accuracy as the number of individuals per pool

increases, potentially compromising the detection of rare variants or small effect sizes. Lastly, many statistical approaches make assumptions about the underlying data distribution, which may not always hold true for empirical genomic data [34].

To address these challenges, we propose a novel combination of machine learning techniques and statistical tests. Our approach uses binary classifiers, specifically One-Class Support Vector Machines (OCSVM) [35] and Naive Bayesian Classifiers (NBC) [36], to analyse genomic data with regard to the identification of selection traces. The application of machine learning techniques offers several advantages in genomic analysis. OCSVM has proven effective at identifying collective, unusual patterns across diverse fields, from detecting anomalous heart rhythms in ECG data [37] to identifying environmental anomalies in atmospheric measurements [38]. As an unsupervised learning method, OCSVM's ability to detect anomalous patterns without requiring pre-labelled training data makes it particularly promising for identifying adaptive loci in allele frequency changes, where the adaptive variants are, by definition, initially unknown. Both OCSVM and NBC demonstrate flexibility in pattern recognition, capable of capturing complex, non-linear relationships in the data, making them well-suited to detect subtle, polygenic adaptation patterns that might be missed by linear statistical methods. These classifiers are also suited for handling high-dimensional data, a common characteristic of genetic datasets, potentially uncovering patterns that are not apparent in univariate analyses. Contrary to window-based methods, these classifiers can analyse data across the genome without relying on arbitrary window sizes, potentially capturing both localized and broad-scale patterns of adaptation. NBC provides an additional advantage by offering a probabilistic framework for classification, aligning well with the stochastic nature of evolutionary processes and allowing for nuanced interpretation of results. By combining these classifiers with traditional statistical tests (FET) in our OCSVM-FET and NBC-FET approaches, we leveraged the strengths of both machine learning and classical statistical methods, providing a framework with enhanced performance.

Therefore, in this study, we explore and compare five distinct approaches and their combinations: FET, NBC,OCSVM, OCSVM-FET and NBC-FET. To evaluate and compare strategies in the presence of a known ground truth, we used simulated data. Our simulations explored various evolutionary scenarios, modifying in the number of generations, loci under selection, and selection coefficients. As a proof of concept, real genomic data from an Evolve and Resequence experiment with *Chironomus riparius* [39] was processed, a model organism known for its rapid adaptation capabilities [40].

Results

Parameter Optimization on template data

To ensure optimal performance of our algorithms in detecting polygenic adaptation patterns, we conducted extensive parameter tuning using previously published real-life

data (allele frequency distribution) from *C. riparius*. This process involved a comprehensive grid search for both the OCSVM with a radial kernel and the NBC.

For the OCSVM algorithm, we focused on optimizing two key parameters: ν and γ. While the ν parameter influences the algorithm's sensitivity to outliers, the γ parameter controls the decision boundary (for more details see Material and Methods). We tested ν values ranging from 0.001 to 1 (0.001, 0.005, 0.01, 0.013, 0.02, 0.1, 1) and γ values of 0.01, 0.05, 0.5, and 1.

Figure 1a illustrates the performance metrics - False Positive Rate (FPR), Area Under the Curve (AUC), and Accuracy - for various OCSVM parameter combinations. The optimal performance was achieved with ν=0.01 and γ=0.05, yielding the best balance between a low FPR (0.013), high AUC (0.993), and high Accuracy (0.987).

For the NBC, the parameter tuning process was more complex due to the need to approximate three models: one for the majority, non-anomalous data (represented by mu1 and sigma1) and two for the anomalous data (represented by mu2/mu3 and sigma2/sigma3). The challenge in this process was ensuring that the covariance matrices remained valid positive semi-definite matrices, a mathematical requirement for covariance matrices. Thus, we tested 10 different combinations of mu and sigma values, as shown in Figure 1b.

The NBC parameter search revealed that Setting 1 (mu1=[0.0, 0.0], sigma1=[0.15, 0.1], mu2/mu3=[0.1, 0.9], sigma2/sigma3=[0.25, 0.2]) provided the best overall performance with a low FPR of 0.004, high AUC of 0.998, and high accuracy of 0.996.

Comparing the top-performing settings for both algorithms (Tab 1a and 1b), we observed that while both methods achieved high accuracy and AUC, the NBC demonstrated a slightly lower FPR than the OCSVM. However, the OCSVM showed more consistent performance across different parameter settings, suggesting it might be more robust to parameter variations (for detailed information see Supplementary Tab S1).

These optimized parameters were subsequently used in our main analysis to detect polygenic adaptation patterns in simulated data. This ensured that our algorithms were fine-tuned for maximum effectiveness in identifying subtle genomic changes associated with rapid adaptation.

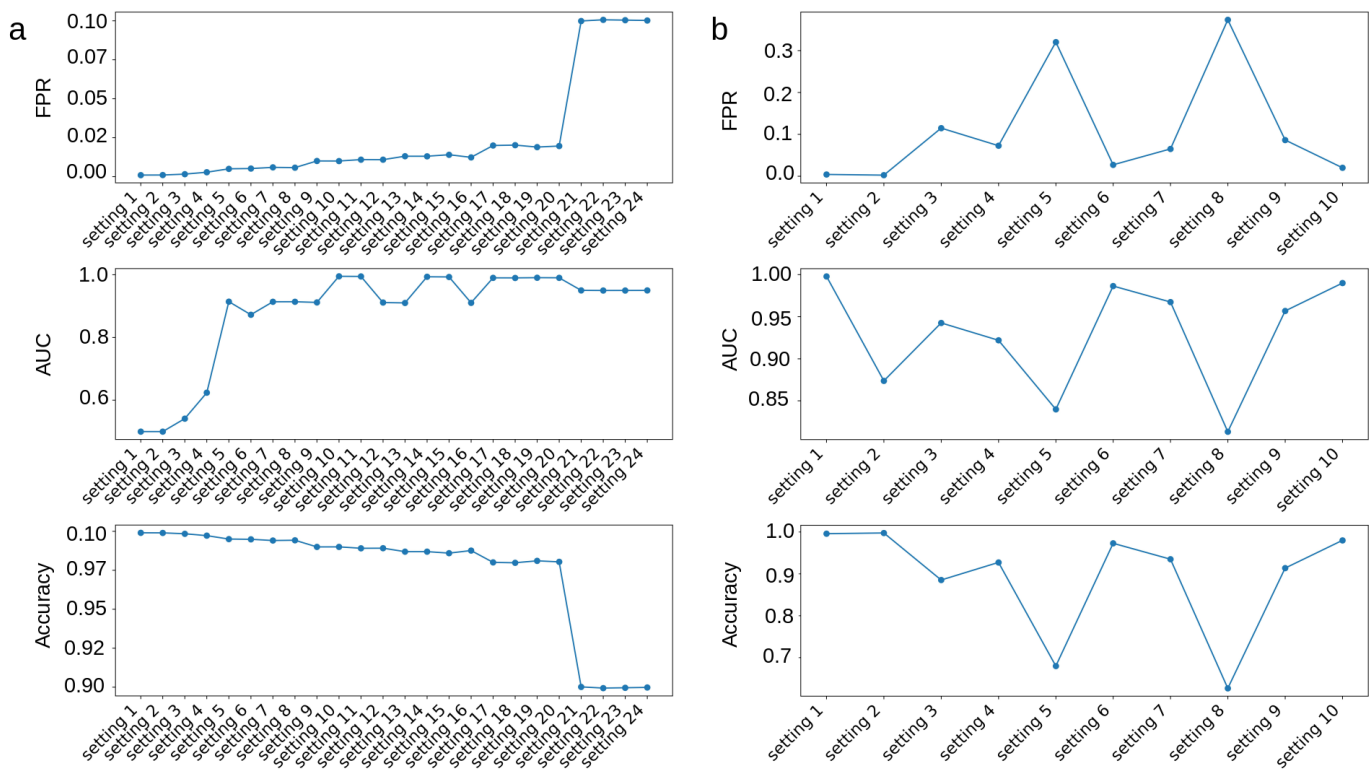

Fig 1. Parameter optimization for One Class Support Vector Machines (OCSVM) and Naive Bayesian Classifier (NBC) algorithms. (a) Performance metrics (False Positive Rate, Area Under the Curve, and Accuracy) for various parameter settings of the OCSVM with radial kernel. (b) Corresponding metrics for different NBC parameter configurations.

Tab 1. Top 5 parameters settings for One Class Support Vector Machines (OCSVM) and Naive Bayesian Classifier (NBC)

a

Setting	*nu*	*gamma*	FPR	AUC	ACCURACY
10	0.01	0.05	0.010	0.995	0.990
11	0.01	0.5	0.011	0.995	0.989
14	0.013	0.05	0.013	0.993	0.987
15	0.013	0.5	0.014	0.993	0.986
17	0.2	0.01	0.020	0.990	0.980

b

Seeting	nu_1	$sigma_1$	mu_2/mu_3	$sigma_2/sigma_3$	FPR	AUC	AUCCURAY
1	0.0, 0.0	0.15, 0.1	0.1, 0.9	0.25, 0.2	0.004	0.998	0.996
10	0.0, 0.0	0.15, 0.1	0.7, 0.2	0.25, 0.2	0.020	0.990	0.980
6	0.0, 0.0	0.15, 0.2	0.3, 0.8	0.15, 0.1	0.027	0.986	0.973
7	0.0, 0.0	0.2, 0.15	0.35, 0.6	0.3, 0.2	0.065	0.967	0.935
4	0.0, 0.0	0.2, 0.1	0.3, 0.5	0.4, 0.2	0.073	0.922	0.927

(a) Top 5 parameter settings (ν and γ) for the OCSVM algorithm with their respective performance metrics. (b) Top 5 parameter settings for NBC, showing mu and sigma values for three approximations along with their performance metrics.

Application on simulated data

To evaluate the performance of competing methods to detect patterns of polygenic adaptation where the ground truth is known, we utilized simulated data generated with Mimicree2 based on real *C. riparius* allele frequencies.
Our analysis of different approaches (FET, OCSVM, OCSVM-FET, NBC, and NBC-FET) across various generations (10, 20, 40, and 60) revealed that generation 40 provides the optimal balance for detecting loci under selection (Fig 2).

FPR showed a downward trend from generation 10 to generation 40 for all approaches (Fig 2a). NBC and OCSVM, showed a slight increase in generation 60, whereas the remaining approaches reached their lowest values in generation 60. AUC/ROC (Fig 2b) indicated that all approaches, except FET, reached peak performance at generation 40. Notably, NBC maintained fairly good performance (AUC>60%) even at generation 60, while other approaches declined to approximately 50%. FET consistently showed values around 50%, indicating random performance, while OCSVM-FET demonstrated the best overall results, particularly at generation 40. Accuracy (Fig 2c) showed a consistent improvement for all approaches with increasing generations, reaching a plateau-like signature from generation 40 to generation 60. OCSVM-FET achieved the highest accuracy at generation 40, followed by FET and NBC.

In summary, for the tested evolutionary scenarios, generation 40 emerged as the optimal scenario for all approaches, considering both accuracy and AUC. While 3 (OCSVM, OCSVM-FET and NBC-FET) of the 5 approaches reached their lowest values of FPR in generation 60, generation 40 shows comparable low values for these 3 approaches and additionally the lowest values for the remaining 2 approaches (FET, NBC).

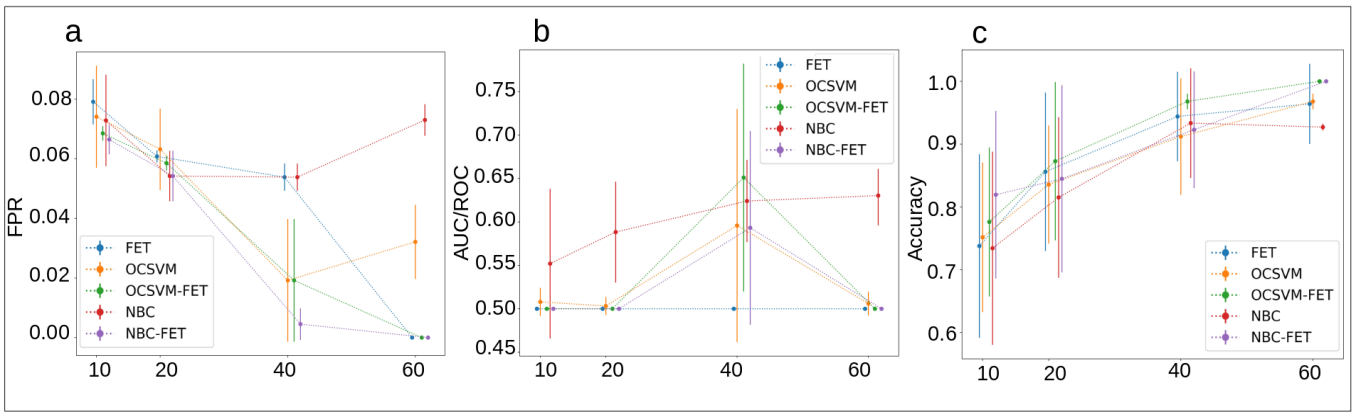

Fig 2. Performance comparison of five approaches for detecting loci under selection across different generations. The approaches evaluated are Fisher's Exact Test (FET), One Class Support Vector Machines (OCSVM), Naive Bayesian Classifier (NBC) and the combination OCSVM-FET, NBC-FET. Performance metrics shown are (a) false positive rates (FPR), (b) area under curve/receiver operating characteristic (AUC/ROC), and (c) Accuracy. Data points represent mean values across simulations, with error bars indicating standard deviation. Generations tested: 10, 20, 40, and 60.

To understand these results, we examined the allele frequency distributions across different generations (Fig 3), as well as phenotypic values. Allele frequencies of SNPs identified to be under selection (red dots) became increasingly distinct from the background (blue dots) as generations progressed, with a clear separation emerging

around generation 40 (Fig 3a). This pattern closely resembled the distribution observed in the real-life data used for parameter tuning (Fig 3b). The similarity between the simulated data at generation 40 and the real-life data explains the effectiveness of our tuned algorithms. This finding underscores the importance of considering the temporal dynamics of allele frequency changes when studying rapid adaptation, and suggests that our approach is particularly effective at detecting selection signatures that have had sufficient time to manifest but have not yet reached fixation.

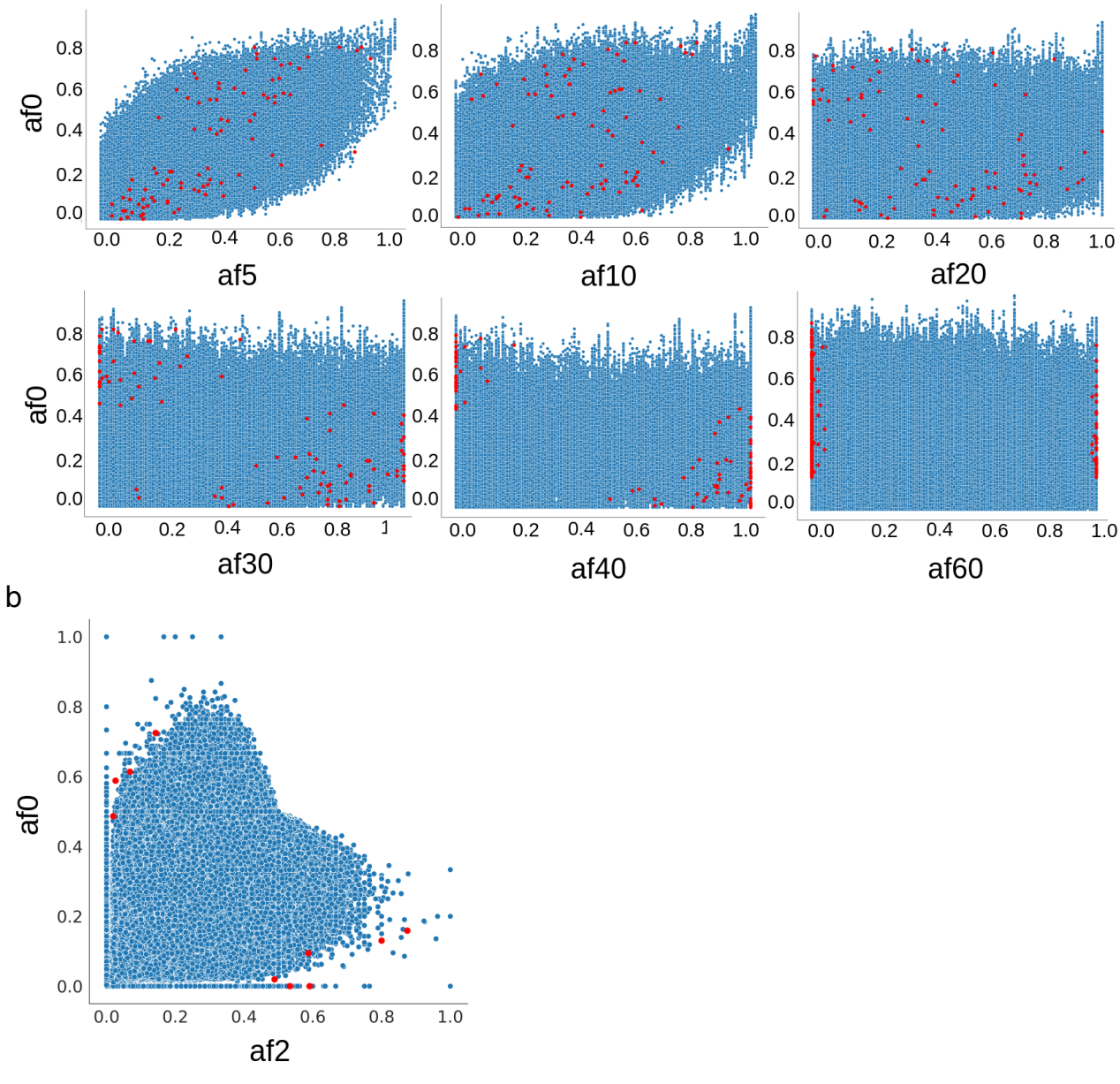

Fig 3. Allele frequency distributions across generations. (a) Simulated data showing allele frequencies (AF) at generations 5, 10, 20, 30, 40, and 60 plotted against initial allele frequencies (af0). Red dots represent selected SNPs, while blue dots represent background SNPs. (b) Real-life data used for parameter tuning, showing allele frequencies at generation 2 plotted against initial frequencies (af0). The distribution of anomalies identified loci in generation 40 of the simulated data most closely resembles the real-life data.

The phenotypic trajectories for different numbers of selected loci revealed distinct patterns of adaptation over 60 generations (Supplementary Fig S6). Simulations with fewer loci (10 and 50) showed rapid initial increases in phenotypic values, reaching higher values in early generations but plateauing more quickly. In contrast, simulations with more loci (250 and 500) displayed a slower initial increase but continued to rise steadily throughout the 60 generations, ultimately reaching higher phenotypic values.

The simulation with 250 loci exhibited a unique pattern, starting with the lowest phenotypic value at generation 10 but showing the steepest increase between generations 20 and 40. By generation 60, it achieves the highest phenotypic value among all scenarios. This pattern aligns with our observation of peak algorithm performance at 40 generations for 250 loci, suggesting that this represents an optimal balance between the number of loci under selection and the time allowed for adaptation (Supplementary Fig S6). The 100 loci scenario shows intermediate behaviour, with a relatively rapid initial increase followed by a more gradual rise, positioning it between the extremes of the fewer and greater loci scenarios (Supplementary Fig S6).

OCSVM-FET shows Superior Performance Across Loci Numbers

Using generation 40 as the optimal timepoint, we evaluated the performance of different approaches across varying numbers of loci under selection (n = 10, 50, 100, 250, and 500) (Fig 4). FPR analysis showed consistent results of approximately 0.06 for FET and NBC across all loci settings. NBC-FET maintained a low FPR of around 0.01, while OCSVM and OCSVM-FET showed slightly elevated values for n=100, but otherwise maintained values around 0.02.

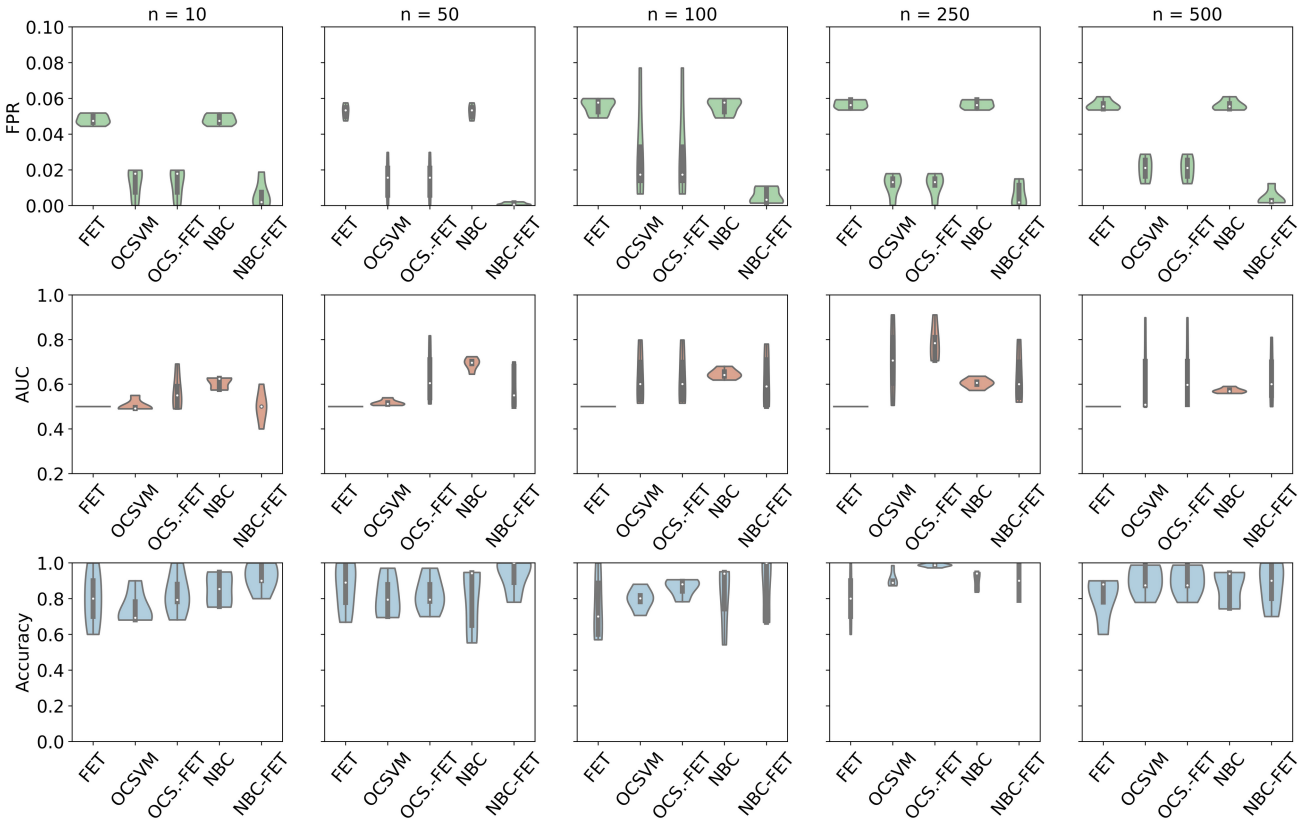

Fig 4. Comparative analysis of detection approaches across varying numbers of loci under selection at generation 40. The graph displays false positive rate (FPR) (first column graphs, green), area under curve/receiver operating characteristic (AUC/ROC) (second column graphs, red), and accuracy (third column graphs, blue for Fisher's Exact Test (FET), One Class Support Vector Machines (OCSVM), Naive Bayesian Classifier (NBC) and the combined approaches OCSVM-FET (OCS.-FET), NBC-FET. Each row represents the number of loci under selection (n = 10, 50, 100, 250, 500).

AUC performance varied among approaches. FET consistently showed values around 50% across all loci settings, indicating random performance. OCSVM and OCSVM-FET demonstrated increasing performance with higher numbers of selected loci, peaking at n=250 (OCSVM >70% AUC, OCSVM-FET >80% AUC), followed by a slight decrease at n=500 (OCSVM <65% AUC, OCSVM-FET <70% AUC). NBC reached its optimal performance at n=50 (~65% AUC), gradually declining to around 60% with increasing number of loci. Conversely, NBC-FET showed improving performance with increasing n, plateauing at n=100 with values consistently above 60% (Fig 4).

Accuracy analysis revealed that the OCSVM-FET approach achieved optimal results with n=250 selected loci, reaching approximately 99% accuracy. Accuracy generally increased with n for this approach, with a slight decrease to 80% at n=500. NBC-FET maintained high accuracy (around 90%) for n=10, 50, 100, and 150, with a minor decrease at n=500. FET showed no distinct trend, maintaining around 80% accuracy across various n values. OCSVM and NBC exhibited a gradual increase in accuracy with increasing n, both peaking at n=250 before slightly decreasing.

While our main analysis focused on generation 40, we also examined the performance of our methods across generations 10, 20, and 60 to provide a comprehensive view of how the detection of polygenic adaptation signals evolves over time (Supplementary Figs S3-S5).

At the early stage of generation 10 (Fig S3), all methods showed relatively poor performance across all loci numbers. The AUC/ROC values were generally lower compared to later generations, with most methods hovering around 50-60%, indicating only slightly better than random performance. Accuracy was also generally lower, with high variability across methods and loci numbers. For generation 20 (Fig S4), performance improved for most methods compared to generation 10, particularly for scenarios with higher numbers of loci under selection (n=250, 500). OCSVM and OCSVM-FET began to show better AUC/ROC values, especially for n=250 and n=500, though not as high as observed at generation 40. Accuracy also improved, with OCSVM-FET and NBC-FET showing the most consistent performance across loci numbers. Interestingly, for generation 60 (Fig S5), performance for most methods slightly declined or plateaued compared to generation 40, particularly for scenarios with higher number of loci (n=250, 500). This trend was most noticeable in the AUC/ROC and accuracy measures for OCSVM and OCSVM-FET. FPR remained relatively stable across generations for most methods. This decline in performance at later generations might indicate that the strongest signals of polygenic adaptation become harder to detect as populations approach equilibrium or as genetic drift begins to play a more significant role.

Optimal Framework Performance

Our comprehensive analysis indicated that the OCSVM-FET approach at generation 40 with 250 selected loci provided the most robust framework for detecting polygenic adaptation patterns (Fig 5). Under these conditions, OCSVM-FET demonstrated consistently strong performance metrics: it achieved the lowest False Positive Rate, produced AUC values exceeding 80%, and maintained near-perfect accuracy of approximately 99%. This combination of parameters offered the best trade-off between sensitivity and specificity in identifying loci under selection, suggesting its potential as a powerful tool for detecting polygenic adaptation in genomic data.

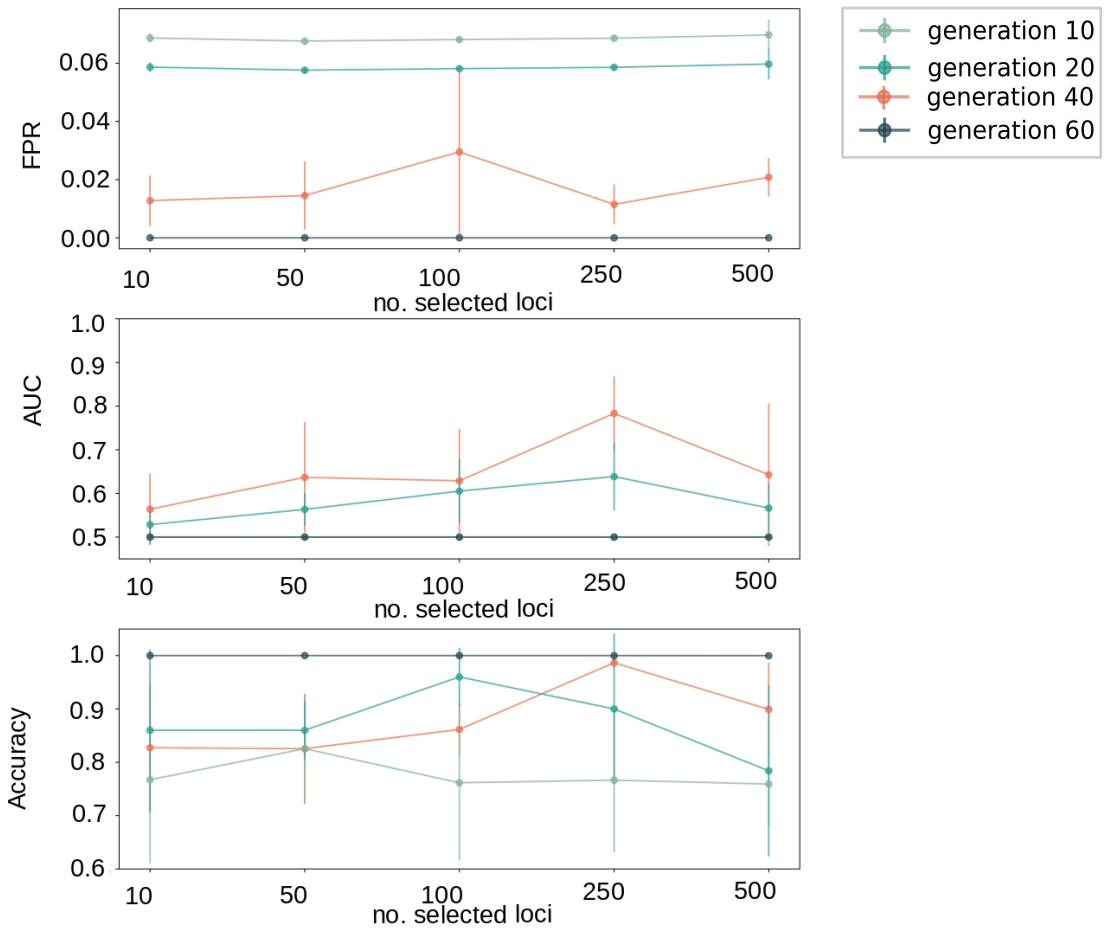

Fig 5. Performance metrics of the combined One Class Support Vector Machines – Fisher's Exact Test (OCSVM-FET) approach. OCSVM-FET performance metrics at generation 10, 20, 40 and 60 for different numbers of selected loci. The graph illustrates false positive rate (FPR), area under curve (AUC), and accuracy. The x-axis represents the number of loci under selection (n = 10, 50, 100, 250, 500). This figure highlights the optimal performance achieved with 250 selected loci, demonstrating the lowest FPR, highest AUC, and near-perfect accuracy.

Discussion

Our study introduced and tested novel approaches to investigate polygenic adaptation at the genetic level by combining standard statistical methods with novel machine learning

techniques. Our results indicate that combining classic statistics with information-theory-based methods leads to a significantly optimized detection of polygenic adaptation. By fitting parameters to real-life data and validating with simulations, we strike a balance between biological realism and methodological rigour, aligning with recent trends in population genetics that emphasize the incorporation of real genomic complexity into analytical methods [41]. Importantly, our study underscores the crucial role of parameter optimization in classification algorithms, highlighting both the applicability of our approach and its inherent limitations to study cases. This nuanced perspective reflects the complexity of adapting machine learning techniques to specific biological contexts, a key consideration in the evolving landscape of computational population genetics.

Comparative Performance and Temporal Dynamics of Adaptation

The combined OCSVM-FET method consistently outperformed Fisher's Exact Test across all tested scenarios. The peak in accuracy and AUC at 250 selected loci suggested an optimal balance between signal strength and noise in our simulated data. This pattern underscored the importance of considering the scale of polygenic adaptation when applying detection methods.

Our simulations revealed that the detection power of OCSVM-FET was particularly strong for highly polygenic architectures. The scenarios with 100 and 250 loci both exhibited pronounced phenotypic trajectories, characterized by steady increases between generations 20 and 40 (Supplementary Fig S6) These patterns indicated strong and consistent selection signals across many loci, providing robust patterns for the OCSVM method to detect. At generation 40, these populations were undergoing significant adaptive change, making the selection signal particularly pronounced and detectable, not yet dominated by genetic drift. With fewer loci (10 or 50), the simulated adaptive signal might have been too simple or lacked the complexity that our OCSVM method is designed to capture in polygenic traits. These scenarios showed rapid initial adaptation but quick plateauing, potentially making the selection signal too weak or homogeneous by generation 40. Conversely, with very high numbers of loci (500), the adaptive signal might became too diffuse or noisy, showing a slower, more gradual increase that might not provided a sufficiently strong signal at generation 40 for optimal detection. This time-step, however, provided enough genetic complexity to represent realistic polygenic traits while not being so complex as to obscure the adaptive signal [42]. This aligns with recent understanding in quantitative genetics that many adaptive traits are influenced by hundreds of loci rather than just a few major effect loci or thousands of minute-effect loci [23]. For instance, studies in humans have found hundreds of loci associated with height [43] and other quantitative traits. However, the exact number of loci involved in adaptation can vary widely depending on the trait and organism studied [44]. Moreover, OCSVM-FET showed robust results across various simulation settings. Even when performance slightly declined (e.g., at generation 60 or with more than 250 loci), OCSVM-FET still outperformed traditional methods. This robustness indicates its value in detecting ongoing selection across various scenarios of polygenic adaptation.

It is noteworthy to mention that the peak performance we observed is partly due to model optimization, which likely reflects the patterns in the data used to tune the model. This phenomenon is well-known when using unsupervised machine learning approaches in genomic contexts [45]. For this reason, we used real-life data for model-optimization to reflect real-life scenarios as closely as possible. While the true architecture of polygenic adaptation patterns remains complex, our findings provided deeper insights into the genomic scale at which such patterns may be most readily detectable, offering a valuable tool for pinpointing candidate SNPs in similar scenarios.

Trade-offs between methods

Our analysis revealed significant differences in the performance among the methods we tested. Fisher's Exact Test, while widely used, mainly showed the highest FPR, lowest AUC, and in most cases, the lowest accuracy compared to other methods. This suggests that more sophisticated approaches like OCSVM and NBC are more suitable for detecting polygenic adaptation patterns. Notably, combining these methods with FET further enhanced the results, with OCSVM-FET demonstrated the best overall performance, characterized by consistently low FPR and high accuracy and AUC for specific settings.

The performance of OCSVM was highly dependent on parameter settings, necessitating optimal tuning for the specific scenario under investigation. Grid searches using comparable known positive scenarios can help identify optimal solutions [46]. The challenge lies in balancing the trade-off between model complexity and generalization ability, a common issue in machine learning applications to genomic data [41].

Furthermore, OCSVM in general can be computationally intensive, with performance heavily dependent on parameter settings and data architecture. The time complexity of OCSVM is generally O(n^2) to O(n^3), where n is the number of training samples. This can make predictions about time expenditure challenging, especially for large genomic datasets.

In contrast, NBC generally offers relatively low computational expenditure and can be performed on most machines due to the quick calculation of probability density functions. The time complexity of NBC is typically O(nd), where n is the number of training samples and d is the number of features, making it more scalable for large datasets.

While grid searches can optimize parameters for both OCSVM and NBC, the latter requires a deeper understanding of the underlying calculations. A key challenge in NBC parameter optimization is ensuring that covariance matrices remain valid positive semi-definite, a mathematical requirement crucial for the method's validity [47]. This constraint may necessitate considering alternative fine-tuning methods for NBC.

When computational resources are limited, investing time in careful parameter optimization for NBC may be worthwhile. However, if computational power is not a constraint, OCSVM would be the preferred choice due to its superior overall performance and more straightforward parameter tuning procedure.

Importance of parameter optimization

Our study demonstrated that parameter optimization is crucial for achieving optimal results with machine learning algorithms, particularly in complex biological applications like detecting polygenic adaptation. Therefore, careful parameter optimization for both OCSVM and NBC methods is an important pre-requisite for optimal model performance.

In the context of OCSVM, parameter selection presented unique challenges due to its unsupervised nature. As Amer et al. [48] point out, the performance of OCSVM is highly sensitive to the choice of parameters, particularly the kernel function and the regularization parameter. This sensitivity is especially pronounced in high-dimensional biological data, where the balance between model complexity and generalization ability is crucial [35].

In our study, we employed real-life data for parameter optimization to capture the true complexity of genetic variation and selection patterns that may not be fully replicated in simulations. Libbrecht and Noble [49] highlight that in genomic applications, the effectiveness of machine learning models heavily depends on their ability to capture the specific patterns and complexities of biological data. Our approach of using real-life data from *C. riparius* for initial parameter optimization aligned with this principle, ensuring that our models were calibrated to biologically relevant patterns of polygenic adaptation.

However, this approach also presents challenges. The application of machine learning in bioinformatics requires a delicate balance between leveraging domain-specific knowledge and avoiding overfitting to specific datasets [50]. Our method, tuned on real-life data, showed improved performance on similar simulated scenarios, demonstrating its potential for detecting polygenic adaptation patterns in comparable evolutionary contexts. This highlights the importance of careful parameter selection and the potential for developing more adaptive algorithms that can be adjusted to different evolutionary scenarios.

In conclusion, while our study demonstrates the power of well-tuned machine learning approaches in detecting polygenic adaptation, it also highlights the ongoing challenge of balancing specificity and generalizability in parameter tuning. This balance is crucial for developing methods that are both powerful and broadly applicable in the study of complex evolutionary processes.

Future Directions

While our results provide valuable insights into detecting polygenic adaptation, it is important to consider the limitations and broader implications of our findings. Future studies could explore how our method performs across a broader range of genetic architectures and selection intensities. Additionally, while we optimized our model using real-life data to reflect realistic scenarios, this approach may introduce biases towards certain types of polygenic architectures. Further validation across diverse empirical datasets would strengthen the generalizability of our findings.

Our results underscore the importance of considering the temporal dynamics of adaptation when applying detection methods in natural populations, where the timing of adaptive

events is often unknown. The model's peak performance at generation 40 might be due to optimization reflecting patterns in the training data or specific setting to the simulations, that mostly reflected the polygenic adaptation pattern observed in *C. riparius*. Future work could explore ways to make the model more generalizable across different stages of adaptation, perhaps by incorporating time-series data in the training process.

Moreover, future research could focus on advanced parameter tuning techniques, such as Bayesian optimization [51], to improve the efficiency of tuning for both OCSVM and NBC. For instance, Chapelle et al. [48] propose methods for automatically tuning multiple parameters in SVMs, which could be adapted for OCSVM in the context of polygenic adaptation detection. Additionally, incorporating techniques to enhance model robustness and transferability across different evolutionary scenarios could further improve the applicability of machine learning methods in diverse biological contexts.

It is important to note that while OCSVM effectively identifies loci showing unusual patterns of allele frequency change, the method does not directly provide functional characterisation of these selected loci. To fully leverage the power of this approach, future work could focus on developing complementary methods to elucidate the biological impact of the identified loci. Such methods should examine the genomic locations and known functional annotations of identified loci, analyse the magnitude and direction of allele frequency changes in these regions, investigate potential relationships among selected loci such as clustering in specific genomic regions, and explore associations between identified loci and specific traits or environmental variables. By integrating these analyses, researchers could transform the "shortlist" of interesting loci provided by OCSVM into a more comprehensive understanding of the genetic basis of adaptation. This multi-faceted approach would not only identify potential targets of selection but also provide insights into the functional and evolutionary significance of these genomic regions.

In conclusion, while our OCSVM-based method represents a significant advance in detecting polygenic adaptation, there remains ample room for refinement and expansion. Future research directions should focus on enhancing the model's generalizability, improving parameter optimization, and developing complementary approaches to characterize identified loci. These advancements will contribute to a more nuanced and comprehensive understanding of polygenic adaptation in both simulated and natural populations.

Material and Methods

Simulations

We selected MimicrEE2 (version mim2-v206) [52] as our simulation software due to its compatibility with downstream analysis tools like Popoolation2 (version popoolation2_1201) and its ability to incorporate user-provided haplotype files. This feature allowed us to use sequencing data from our organism of interest, *Chironomus riparius*, ensuring a close approximation to real-life experiments compared to purely simulated data.

MimicrEE2's quantitative trait (QT) mode was employed to simulate rapid polygenic adaptation. This mode first computes phenotypic values for each individual based on the effect sizes of the SNPs and environmental variance, then performs truncating selection, culling individuals with the most extreme phenotypic values. This approach allows for the simulation of complex adaptive events under various evolutionary scenarios.

Our simulation pipeline started with the generation of a haplotype input file using individual whole genome sequencing data from 66 *C. riparius* individuals, that were conducted during the study of Foucault et. al. [22]. We utilized the non-model organism pipeline for GATK4 [53] (version 4.5.0.0) to create haplotypes from individual sequencing fastq files. These empirical haplotype data served as the ancestral generation for subsequent simulations, providing a biologically relevant starting point specific for *C. riparius*.

Using this haplotype file, we conducted a total of 20 distinct simulation setups, which were designed to explore two key variables:

1. Time under selection: We simulated four different periods under selection - 10, 20, 40, and 60 generations. This range allowed us to observe how adaptation patterns developed over different time scales.
2. Number of loci under selection: We varied the number of loci experiencing selection, testing five different quantities - 10, 50, 100, 250, and 500. This variation helps us understand how the number of genes involved in adaptation affects detection methods, ranging from oligogenic over mildly to highly polygenic adaptation.

For each combination of generation time and number of selected loci, we ran 10 replicates to account for stochastic variations in evolutionary processes. This design resulted in a total of 200 simulation runs (4 generation times × 5 loci numbers × 10 replicates) (Tab 2)

Using *qt* mode, we modelled a scenario of strong selection by specifying a range of effect sizes for the loci under selection. The effect size represents the magnitude of influence a particular allele has on the phenotypic value of an individual. MimicrEE2 uses these effect sizes in the following way: Firstly, for each locus under selection, an effect size is randomly chosen from our specified range (in our case: 1.5 to 3.5). Secondly, the phenotypic value of an individual is calculated as the sum of the products of each locus's effect size and its allele frequency, plus random environmental noise. Lastly, the initial allele frequencies provided together with the effect size are used as the starting point for the simulation, representing the genetic makeup of the population at generation 0. We chose a high effect size in range of 1.5 to 3.5 allowing us to observe significant changes within a relatively small number of generations, which is computationally efficient and helps in detecting selection signals more clearly.

We generated a selected loci file using Python (version 3.8.13) and the library pandas (version 2.0.3) [54], which specifies the loci under selection and incorporates the effect sizes of the loci contributing to a quantitative trait. The script randomly selected 10, 50, 100, 250, and 500 positions with a starting allele frequency provided by the effect-file. In our case, we chose allele frequencies ranging from 0.01 to 0.3.

The recombination map was generated using Python and the library pandas. We allowed the recombination rate to vary between 0.1 and 4 centimorgans, encompassing a range of possible recombination rates that most likely reflect the natural variation observed in *C. riparius* populations [40].

Another key parameter was heritability, which was set to 0.8, indicating a strong genetic component in the trait under selection. A truncating selection model was applied with a factor of 0.5, i.e. removing individuals with phenotypic values below the 50th percentile.

The output was stored in sync-file format to facilitate comparison with real-life experimental data. We utilized the R package poolSeq for subsequent analysis of allele frequencies across generations.

This simulation design allowed us to model rapid adaptation under strong selection pressure. By incorporating real genetic data from *C. riparius* and varying key parameters such as selection coefficient and recombination rate, we aimed to capture a range of evolutionary scenarios, providing a robust test for our analytical methods in identifying polygenic adaptation patterns (parameters: java -jar mim2-v206.jar/ qt/ --haplotypes-g0 haplotype-file.mimhap/ --recombination-rate recombinationrate-file/ --effect-size (--n 10, 50, 100, 250, 500/ --f 0.15/ --effect-file)/ --heritability 0.8/ --snapshots 10,20,40,60 / --replicate-runs 10 / --selection-regime 0.5 / --threads 6 --output-sync output.sync.gz /).

Table 2. Schematic construction of simulated data set by Mimicree2.

Generation 10	10 loci	10 replicates
	50 loci	10 replicates
	100 loci	10 replicates
	250 loci	“
	500 loci	“
Generation 20	10 loci	“
	50 loci	“
	100 loci	“
	250 loci	“
	500 loci	“
Generation 40	10 loci	“
	50 loci	“
	100 loci	“
	250 loci	“
	500 loci	“
Generation 60	10 loci	“
	50 loci	“
	100 loci	“
	250 loci	“
	500 loci	“

For each generation 10, 20, 40 and 60 and each selected loci setting (n= 10, 50, 100, 250, 500) 10 replicate pools were created, originating from 66 individual sequencing files [22].

Phenotypic value calculation

While MimicrEE2 does not directly give phenotypic values, we derived these values from the simulation outputs using a custom Python script (version 3.8.13). The phenotypic values were calculated using the allele frequencies from the MimicrEE2 output and the effect sizes provided as input to the algorithm.

The calculation of phenotypic values was performed as follows:

1. Effect sizes were randomly sampled from the input effect size file for each locus under selection, allowing for replacement to match the number of loci in each scenario.
2. The phenotypic value for each generation was computed using the formula:
   Phenotypic Value = Genetic Component + Environmental Component
   Where:
   - $\text{Genetic Component} = \sum_i effectsize_i \cdot allelFrequency_i$ for all loci i
   - Environmental Component = random noise drawn from a normal distribution
3. The environmental component was calculated to maintain a heritability of 0.8, as specified in the simulation parameters.
4. This process was repeated for each scenario (10, 50, 100, 250, and 500 loci) using the corresponding allele frequency files generated by MimicrEE2

Methods for detecting polygenic adaptation

Fisher's Exact Test (FET)

Fisher's exact tests were conducted to compare the ancestral (first) with all other generations (10, 20, 40, and 60), yielding corresponding p-values using poolSeq (version 1.15.4, [55]) with R (version 4.3.2) [56]. Results were corrected for multiple comparisons using the Benjamini-Hochberg correction using the stats R package (version 4.5.0). A significance level cut-off of adjusted p-value < 0.001 was chosen, as proposed by Pfenninger et al. [39] (see section *Parameter Fine Tuning)*.

The One Class Support Vector Machines (OCSVM)

The OCSVM algorithm [35] was employed in this study to identify potential signatures of polygenic adaptation in our allele frequency data, using Python (version 3.8.13) and scikit-learn (version 1.1.3) [40]. The OCSVM algorithm was used with the Radial Basis Function (RBF) kernel, aiming to divide the data into anomalous and non-anomalous categories in higher dimensional space.
OCSVM is designed to detect anomalies or outliers [35]. Thus, we used the algorithm to identify select loci in the allele frequency change (AFC) data of ancestral and last (adapted) generation. The algorithm learns to separate the data into 'normal' and

'anomalous' patterns by finding a decision boundary that encloses the majority of data points, while simultaneously identifying outliers during the training process. In the context of our study, these anomalies represent loci that have undergone unusual changes in allele frequency between the ancestral and adapted generations. Such changes likely indicate selection pressures acting on these loci, either directly or through linkage to adaptive variants.

The algorithm constructs a boundary in a high-dimensional space that encloses most of the "normal" data points. Data points outside this boundary are classified as anomalies. The shape and properties of this boundary are determined by the chosen kernel function and the algorithm's parameters.

A kernel function measures the similarity between two data points and helps define the boundaries between them. In this study, we chose RBF kernel to capture non-linear relationships in the data. The RBF kernel can be thought of as a similarity measure that decreases with distance: data points that are close in the original space are considered more similar than those that are far apart.

The so-called kernel trick is a key concept that allows the OCSVM to operate effectively in higher dimensions. The kernel trick adds new features to each data point by applying a transformation function thereby mapping the original data to a higher-dimensional space. This is implicitly computed: instead of calculating these new features directly (which could be computationally expensive or even infeasible), the kernel trick computes the dot products between pairs of data points in this higher-dimensional space. This allows the algorithm to find patterns and separations that might not be visible in the original data space (Fig 6).

The RBF kernel is mathematically expressed as:

$$k(x, x_i) = \exp(-gamma|x - x_i|^2)$$

Where γ (gamma) is a crucial parameter that influences how the algorithm learns from the data:

- γ (gamma): This parameter determines the influence of each training example on the decision boundary. A larger γ value causes the influence of each point to be more localized, potentially capturing more complex patterns but risking overfitting. A smaller γ value allows points to have a broader influence, potentially capturing more general trends but risking underfitting.

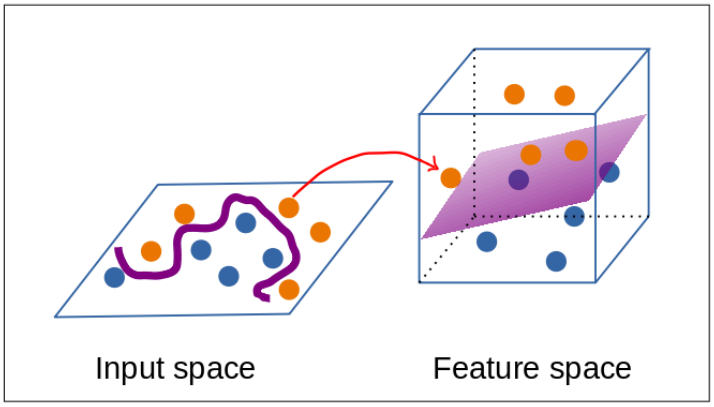

Fig 6. Schematic illustration of One Class Support Vector Machine's *kernel trick.* The so-called kernel trick is used by the OCSVM algorithm to compute dot products between pairs of data points in higher-dimensional space

Another crucial parameter is ν (nu):

- ν (nu): This parameter balances between maximizing the margin (the space around the decision boundary) and minimizing the number of anomalies. It essentially controls how strict the algorithm is in classifying points as anomalies. A smaller ν value results in fewer anomalies being identified, while a larger ν value allows for more anomalies.

Therefore, the γ parameter influences how the algorithm detects unusual allele frequency changes. A smaller γ value allows the algorithm to detect broader, more general trends in allele frequency shifts, which might be more appropriate for detecting polygenic adaptation involving many loci with small effects. A larger γ value might be more suitable for detecting more localized, stronger selection signals. The ν parameter essentially controls the sensitivity of our anomaly detection. A smaller ν value results in a more conservative approach, identifying only the most unusual allele frequency changes, while a larger ν value allows for the detection of more subtle shifts that might be relevant in polygenic adaptation [57]. These parameters were fine-tuned ($\nu = 0.01$, $\gamma = 0.05$) using previously published data on *C. riparius* [39], ensuring the model was optimized for detecting genetic anomalies in this species (see section 2.6 Parameter Fine Tuning).

The OCSVM algorithm the constructs a hyperplane in the higher-dimensional space that serves as the decision boundary. During the optimization process, the algorithm aims to maximize the margin around the hyperplane while minimizing the number of anomalies (see Fig.2).

Allele frequencies, acquired using the R packages poolSeq, were used as input for the algorithm. The algorithm generates a list where each anomalous position is labelled as -1, and non-anomalous data are labelled as 1. Loci flagged as anomalous (-1) by the algorithm are candidates for being part of the polygenic response to selection in our experimental populations. By applying the OCSVM algorithm to our allele frequency data, we aimed to identify the set of loci that collectively contribute to the adaptive response in our experimental populations, providing insights into the genomic architecture of polygenic adaptation.

Naive Bayesian Classifier (NBC)

We employed a NBC algorithm [36] to distinguish between normal (non-anomalous) and anomalous data points based on their allele frequency changes in two-dimensional (2D) space. The NBC is a probabilistic classifier based on Bayes' theorem, where the fundamental idea involves modelling the probability density functions of data set distributions, in this case for non-anomalous and anomalous data. Each data point is represented by two features: the initial allele frequency (generation 0) and the allele frequency at a later generation. Anomalous data represent in this context the candidate SNPs, whereas non-anomalous data represents the positions not targeted by selection (Fig 3a). The basic assumption is, that the allele frequencies of candidate SNPs increase over time because they are beneficial for the adaptive process (Fig 3b) or decline because there are disadvantageous for the adaptive process (Fig 3c). Anomalous data, therefore, show high allele frequency changes (AFC) between the ancestral and the last (adapted) generation. The classifier uses three probability distributions to model these patterns: one for neutral variation and two for potentially selected variants (increasing or decreasing frequencies). The mathematical function modelling these areas of interest is described in the supplement and was implemented using Python (version 3.8.13) and the library scipy (version 3.8.13) and the library scipy (version 1.7.3) [54].

The key parameters for our NBC implementation are the mean vector (μ) and covariance matrix (Σ) for each class (see Supplement for detailed information). Two classes approximate the anomalous, whereas the third class approximates the non-anomalous, background data (Supplementary Fig S1). These parameters were fine-tuned using previously published data on *C. riparius* [39], ensuring our model was optimized for detecting genetic anomalies in this species. The classifier models separate probability density functions for non-anomalous and anomalous data points. By fitting appropriate probability distributions to the observed data, we estimated the likelihood of a given allele frequency belonging to either the normal or an anomalous distribution. In practice, three classes were used: one for non-anomalous data (characterized by small allele frequency changes) and two for anomalous data (one for changes from low to high frequency, and vice versa).

Overlap with Fisher's Exact Test

We developed a combined approach that integrates the results from 3 independent analyses: the binary classification results from OCSVM and NBC, along with the statistical hypothesis testing of FET. Using Python and the library pandas, we merged the previously obtained results from all three methods by matching genomic positions. This integration process identified overlapping positions – those flagged as anomalous by both approaches (OCSVM and NBC) and simultaneously showing statistical significance in FET. This combined approach leverages the strengths of both machine learning and traditional

statistical testing to identify and narrow down candidate positions targeted by polygenic adaptation.

Parameter Fine Tuning

The key parameters μ and Σ for NBC, as well as ν and γ for OCSVM were optimized using real-life and previously published datasets by Pfenninger et al. [39]. This data, originating from an experiment on *C. riparius*, captured genome-wide allele frequency changes before and after a cold snap event, suggesting a potential rapid, polygenic adaptation in the population. DNA extraction, sequencing, trimming, mapping, filtering of reads and SNP calling was performed as described in [39]. To infer long-term selection regimes, Tajima's D was calculated. The temporal behavior of candidate loci was compared to randomly selected SNPs to assess their correlation. In this study, several genetic markers were identified with significant allele frequency shifts before and after the cold snap, indicating potential selection events. These loci were used as true positives to fine-tune the parameters for the introduced methods NBC and OCSVM. The optimized parameter settings were then evaluated for accuracy, false positive rate (FPR) and area under curve (AUC) using simulated data.

While the original study identified 19 SNPs as candidates for selection, these were consolidated into 10 independent loci due to linkage on the same scaffolds [39]. For our parameter fitting process, we used these 10 independent loci as true positives, as they represent distinct genomic regions potentially affected by selection and are likely physically unlinked.

To optimize OCSVM parameters, we performed a grid search over ν values ranging from 0.001 to 1 and γ values from 0.01 to 1. For NBC, we explored various combinations of μ and Σ values, ensuring the resulting covariance matrices remained valid positive semi-definite. The settings for these parameters can be found in the Supplementary Fig S2. The grid search aimed to maximize performance across accuracy, FPR and AUC. These metrics were chosen to balance the algorithm's ability to correctly identify true positives while minimizing false detections, crucial for detecting subtle signals of polygenic adaptation.

The optimized parameters were then applied to simulated data mimicking various evolutionary scenarios, including different generations (10, 20, 40, 60) and numbers of loci under selection (10, 50, 100, 250, 500). This allowed us to assess how well the parameters, tuned on real data, performed across a range of potential adaptive scenarios.

Using the 10 independent loci as true positives for parameter tuning aligns with the expectation that polygenic adaptation often involves multiple, unlinked genomic regions. This approach allowed us to capture the diversity of allele frequency changes associated with complex adaptive processes.

Performance Metrics

Figure 7 provides a flowchart of the bioinformatic pipeline, illustrating the process from data simulation though the application of various approaches to the final comparison of results.

Performance of the different approaches (FET, OCSVM, NBC, OCSVM-FET, NBC-FET) was evaluated using metrics including FPR, AUC, and accuracy with Python and scikit-learn. These metrics were chosen to be the most informative ones regarding a highly imbalanced dataset. FPR, therefore, measures the proportion of data points that were wrongly identified as positive by the algorithm. It is calculated as the ratio of false positives to the sum of true negatives and false positives. The AUC metric represents the area under the Receiver Operating Characteristic (ROC) curve, which is a graphical representation of the true positive rate against the false positive rate at various threshold settings. A higher AUC value indicates better discrimination ability of the algorithm. Lastly, accuracy measures the overall correctness of the algorithm's predictions and is calculated as the ratio of the total number of correctly classified samples to the total number of samples in the data set. Plots were produced using Python (version 3.8.13) and matplotlib (version 3.5.2) [58].

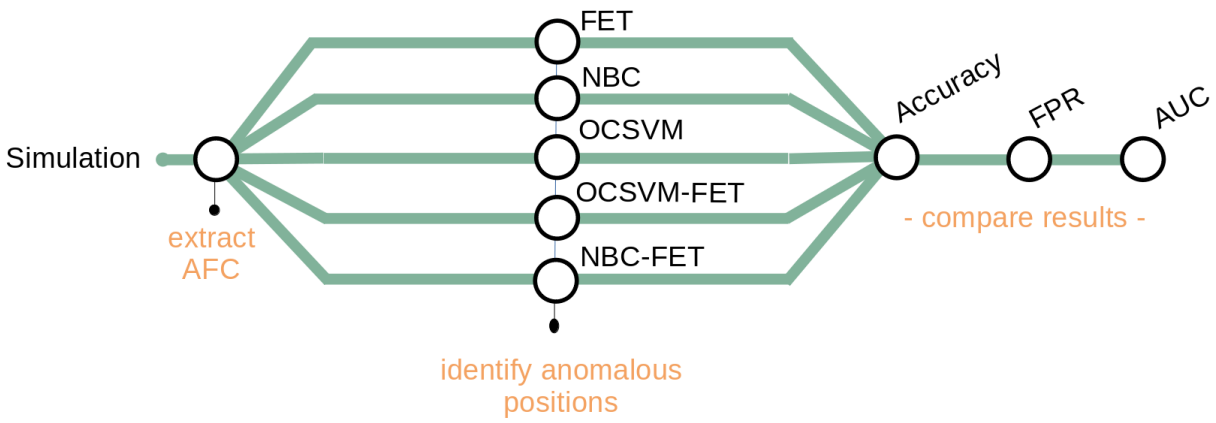

Fig 7. Flow chart of bioinformatic pipeline. Simulated data was generated by using the software MimicrEE2 and allele frequency changes were extracted. To identify anomalous positions, several approaches were tested: Fisher's Exact Test (FET), Naive Bayesian Classifier (NBC), One Class Support Vector Machines (OCSVM), as well as the combined approaches OCSVM-FET and NBC-FET. All approaches were compared based on accuracy, false positive rate (FPR) and area under curve (AUC).

Acknowledgements

We gratefully acknowledge the financial support provided by the Emergent AI Center through the Carl-Zeiss-Stiftung. We extend our sincere thanks to Prof. Dr. Peter Tino, Dr. Pietro Verzelli, and Dr. Daniele Zambone for their valuable insights and guidance. We thank Dr. Hristo Todorov, Dr. Maria Esther Nieto Blázques, Dr. Barbara Feldmeyer and Stephan Weißbach for their careful review of the manuscript. We also acknowledge Dr. Quentin Foucault for conducting the experiments and providing pool-Seq data for parameter tuning.

Code availability

Python-, R- and Bash-scripts generated and used throughout this study are available at Github: https://github.com/CoCaliendo/Polygenic-Adaption-Pattern

References

[1] P. H. Harvey and M. D. Pagel, "The Comparative Method in Evolutionary Biology," *Comp. Method Evol. Biol.*, May 1991, doi: 10.1093/OSO/9780198546412.001.0001.

[2] A. A. Hoffmann and C. M. Sgró, "Climate change and evolutionary adaptation," *Nature*, vol. 470, no. 7335, pp. 479–485, 2011, doi: 10.1038/nature09670.

[3] S. J. Franks and A. A. Hoffmann, "Genetics of climate change adaptation," *Annu. Rev. Genet.*, vol. 46, pp. 185–208, 2012, doi: 10.1146/ANNUREV-GENET-110711-155511.

[4] R. D. H. Barrett and D. Schluter, "Adaptation from standing genetic variation," *Trends Ecol. Evol.*, vol. 23, no. 1, pp. 38–44, Jan. 2008, doi: 10.1016/J.TREE.2007.09.008.

[5] K. Jain and W. Stephan, "Modes of rapid polygenic adaptation," *Mol. Biol. Evol.*, vol. 34, no. 12, pp. 3169–3175, 2017, doi: 10.1093/molbev/msx240.

[6] J. K. Pritchard and A. Di Rienzo, "Adaptation – not by sweeps alone," *Nat. Rev. Genet.*, vol. 11, no. 10, p. 665, Oct. 2010, doi: 10.1038/NRG2880.

[7] C. K. Ghalambor, J. K. McKay, S. P. Carroll, and D. N. Reznick, "Adaptive versus non-adaptive phenotypic plasticity and the potential for contemporary adaptation in new environments," *Funct. Ecol.*, vol. 21, no. 3, pp. 394–407, Jun. 2007, doi: 10.1111/J.1365-2435.2007.01283.X.

[8] P. R. Grant and B. R. Grant, "Unpredictable evolution in a 30-year study of Darwin's finches," *Science*, vol. 296, no. 5568, pp. 707–711, Apr. 2002, doi: 10.1126/SCIENCE.1070315.

[9] K. Pelletier *et al.*, "Polygenic architecture of adaptation to a high-altitude environment for Drosophila melanogaster wing shape and size," *bioRxiv*, p. 2024.02.08.579525, Feb. 2024, doi: 10.1101/2024.02.08.579525.

[10] S. M. Rudman *et al.*, "Direct observation of adaptive tracking on ecological time scales in Drosophila," *Science*, vol. 375, no. 6586, Mar. 2022, doi: 10.1126/SCIENCE.ABJ7484.

[11] M. Pfenninger and Q. Foucault, “Population Genomic Time Series Data of a Natural Population Suggests Adaptive Tracking of Fluctuating Environmental Changes,” *Integr. Comp. Biol.*, vol. 62, no. 6, pp. 1812–1826, Dec. 2022, doi: 10.1093/ICB/ICAC098.

[12] J. Merilä and A. P. Hendry, “Climate change, adaptation, and phenotypic plasticity: the problem and the evidence,” *Evol. Appl.*, vol. 7, no. 1, pp. 1–14, Jan. 2014, doi: 10.1111/EVA.12137.

[13] J. Stapley *et al.*, “Adaptation genomics: the next generation,” *Trends Ecol. Evol.*, vol. 25, no. 12, pp. 705–712, Dec. 2010, doi: 10.1016/J.TREE.2010.09.002.

[14] M. Mahmoud, N. Gobet, D. I. Cruz-Dávalos, N. Mounier, C. Dessimoz, and F. J. Sedlazeck, “Structural variant calling: the long and the short of it,” *Genome Biol.*, vol. 20, no. 1, Nov. 2019, doi: 10.1186/S13059-019-1828-7.

[15] J. J. Vitti, S. R. Grossman, and P. C. Sabeti, “Detecting natural selection in genomic data,” *Annu. Rev. Genet.*, vol. 47, pp. 97–120, 2013, doi: 10.1146/ANNUREV-GENET-111212-133526.

[16] J. M. Akey, “Constructing genomic maps of positive selection in humans: where do we go from here?,” *Genome Res.*, vol. 19, no. 5, pp. 711–722, May 2009, doi: 10.1101/GR.086652.108.

[17] C. Schlötterer, R. Tobler, R. Kofler, and V. Nolte, “Sequencing pools of individuals - mining genome-wide polymorphism data without big funding,” *Nat. Rev. Genet.*, vol. 15, no. 11, pp. 749–763, Nov. 2014, doi: 10.1038/NRG3803.

[18] A. Futschik and C. Schlötterer, “The next generation of molecular markers from massively parallel sequencing of pooled DNA samples,” *Genetics*, vol. 186, no. 1, pp. 207–218, Sep. 2010, doi: 10.1534/GENETICS.110.114397.

[19] P. Orozco-Terwengel, M. Kapun, V. Nolte, R. Kofler, T. Flatt, and C. Schlätterer, “Adaptation of Drosophila to a novel laboratory environment reveals temporally heterogeneous trajectories of selected alleles,” *Mol. Ecol.*, vol. 21, no. 20, pp. 4931–4941, Oct. 2012, doi: 10.1111/J.1365-294X.2012.05673.X.

[20] X. Chen, J. B. Listman, F. J. Slack, J. Gelernter, and H. Zhao, “Biases and Errors on Allele Frequency Estimation and Disease Association Tests of Next-Generation Sequencing of Pooled Samples,” *Genet. Epidemiol.*, vol. 36, no. 6, pp. 549–560, Sep. 2012, doi: 10.1002/GEPI.21648.

[21] N. Barghi *et al.*, “Genetic redundancy fuels polygenic adaptation in Drosophila,” *PLOS Biol.*, vol. 17, no. 2, p. e3000128, Feb. 2019, doi: 10.1371/JOURNAL.PBIO.3000128.

[22] Q. Foucault, A. Wieser, A. M. Waldvogel, B. Feldmeyer, and M. Pfenninger, “Rapid adaptation to high temperatures in Chironomus riparius,” *Ecol. Evol.*, vol. 8, no. 24, pp. 12780–12789, Dec. 2018, doi: 10.1002/ECE3.4706.

[23] E. A. Boyle, Y. I. Li, and J. K. Pritchard, “An expanded view of complex traits: from polygenic to omnigenic,” *Cell*, vol. 169, no. 7, p. 1177, Jun. 2017, doi: 10.1016/J.CELL.2017.05.038.

[24] S. Zhao, L. Chi, M. Fu, and H. Chen, “HaploSweep: Detecting and Distinguishing Recent Soft and Hard Selective Sweeps through Haplotype Structure,” *Mol. Biol. Evol.*, vol. 41, no. 10, 2024, doi: 10.1093/molbev/msae192.

[25] C. D. Huber, M. DeGiorgio, I. Hellmann, and R. Nielsen, “Detecting recent selective sweeps while controlling for mutation rate and background selection,” *Mol. Ecol.*, vol. 25, no. 1, pp. 142–156, 2016, doi: 10.1111/mec.13351.

[26] P. Pavlidis and N. Alachiotis, “A survey of methods and tools to detect recent and strong positive selection,” *J. Biol. Res.*, vol. 24, no. 1, pp. 1–17, 2017, doi: 10.1186/s40709-017-0064-0.

[27] J. K. Pritchard, J. K. Pickrell, and G. Coop, “The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation,” *Curr. Biol.*, vol. 20, no. 4, Feb. 2010, doi: 10.1016/J.CUB.2009.11.055.

[28] G. Arnqvist and | Ahmed Sayadi, “A possible genomic footprint of polygenic adaptation on population divergence in seed beetles?,” *Ecol. Evol.*, vol. 12, 2022, doi: 10.1002/ece3.9440.

[29] R. S. Brennan *et al.*, “Experimental evolution reveals the synergistic genomic mechanisms of adaptation to ocean warming and acidification in a marine copepod,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 119, no. 38, Sep. 2022, doi: 10.1073/PNAS.2201521119/FORMAT/EPUB.

[30] B. A. Gould, Y. Chen, and D. B. Lowry, “Pooled ecotype sequencing reveals candidate genetic mechanisms for adaptive differentiation and reproductive isolation,” *Mol. Ecol.*, vol. 26, no. 1, pp. 163–177, Jan. 2017, doi: 10.1111/MEC.13881.

[31] R. Kofler, R. V. Pandey, and C. Schlötterer, “PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq),” *Bioinformatics*, vol. 27, no. 24, pp. 3435–3436, Dec. 2011, doi: 10.1093/BIOINFORMATICS/BTR589.

[32] Z. Wei, W. Wang, P. Hu, G. J. Lyon, and H. Hakonarson, “SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data,” *Nucleic Acids Res.*, vol. 39, no. 19, Oct. 2011, doi: 10.1093/NAR/GKR599.

[33] Q. Chen and F. Sun, “A unified approach for allele frequency estimation, SNP detection and association studies based on pooled sequencing data using EM algorithms,” *BMC Genomics*, vol. 14 Suppl 1, no. Suppl 1, Jan. 2013, doi: 10.1186/1471-2164-14-S1-S1.

[34] N. Barghi, J. Hermisson, and C. Schlötterer, “Polygenic adaptation: a unifying framework to understand positive selection,” *Nat. Rev. Genet. 2020 2112*, vol. 21, no. 12, pp. 769–781, Jun. 2020, doi: 10.1038/s41576-020-0250-z.

[35] A. Ben-Hur, C. S. Ong, S. Sonnenburg, B. Schölkopf, and G. Rätsch, “Support Vector Machines and Kernels for Computational Biology,” *PLOS Comput. Biol.*, vol. 4, no. 10, p. e1000173, 2008, doi: 10.1371/JOURNAL.PCBI.1000173.

[36] P. Langley and W. Iba, “An Analysis Probabilistic Approaches to Induction”, Accessed: Oct. 09, 2024. [Online]. Available: www.aaai.org

[37] C. C. Lin and C. M. Yang, “Heartbeat Classification Using Normalized RR Intervals and Morphological Features,” *Math. Probl. Eng.*, vol. 2014, no. 1, p. 712474, Jan. 2014, doi: 10.1155/2014/712474.

[38] F. Harrou, A. Dairi, Y. Sun, and F. Kadri, “Detecting Abnormal Ozone Measurements with a Deep Learning-Based Strategy,” *IEEE Sens. J.*, vol. 18, no. 17, pp. 7222–7232, Sep. 2018, doi: 10.1109/JSEN.2018.2852001.

[39] M. Pfenninger, Q. Foucault, A. M. Waldvogel, and B. Feldmeyer, “Selective effects of a short transient environmental fluctuation on a natural population,” *Mol. Ecol.*, vol. 32, no. 2, pp. 335–349, Jan. 2023, doi: 10.1111/MEC.16748.

[40] H. Schmidt, S. L. Hellmann, A. M. Waldvogel, B. Feldmeyer, T. Hankeln, and M. Pfenninger, “A High-Quality Genome Assembly from Short and Long Reads for the Non-biting Midge Chironomus riparius (Diptera),” *G3 Genes|Genomes|Genetics*, vol. 10, no. 4, pp. 1151–1157, Apr. 2020, doi: 10.1534/G3.119.400710.

[41] D. R. Schrider and A. D. Kern, “Supervised Machine Learning for Population Genetics: A New Paradigm,” *Trends Genet.*, vol. 34, no. 4, pp. 301–312, Apr. 2018, doi: 10.1016/J.TIG.2017.12.005.

[42] W. Stephan, “Signatures of positive selection: From selective sweeps at individual loci to subtle allele frequency changes in polygenic adaptation,” *Mol. Ecol.*, vol. 25, no. 1, pp. 79–88, Jan. 2016, doi: 10.1111/mec.13288.

[43] A. R. Wood *et al.*, “Defining the role of common variation in the genomic and biological architecture of adult human height,” *Nat. Genet. 2014 4611*, vol. 46, no. 11, pp. 1173–1186, Oct. 2014, doi: 10.1038/ng.3097.

[44] M. V. Rockman, “The QTN program and the alleles that matter for evolution: all that’s gold does not glitter,” *Evolution*, vol. 66, no. 1, pp. 1–17, Jan. 2012, doi: 10.1111/J.1558-5646.2011.01486.X.

[45] P. Larrañaga *et al.*, “Machine learning in bioinformatics,” *Brief. Bioinform.*, vol. 7, no. 1, pp. 86–112, Mar. 2006, doi: 10.1093/BIB/BBK007.

[46] J. Bergstra, J. B. Ca, and Y. B. Ca, “Random Search for Hyper-Parameter Optimization Yoshua Bengio,” *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, 2012, Accessed: Sep. 12, 2024. [Online]. Available: http://scikit-learn.sourceforge.net.

[47] S. Nanga *et al.*, “Review of Dimension Reduction Methods,” *J. Data Anal. Inf. Process.*, vol. 9, no. 3, pp. 189–231, Jul. 2021, doi: 10.4236/JDAIP.2021.93013.

[48] M. Amer, M. Goldstein, and dfkide Slim Abdennadher, “Enhancing One-class Support Vector Machines for Unsupervised Anomaly Detection,” 2013.

[49] M. W. Libbrecht and W. S. Noble, “Machine learning applications in genetics and genomics,” *Nat. Rev. Genet.*, vol. 16, no. 6, pp. 321–332, May 2015, doi: 10.1038/NRG3920.

[50] P. Larrañaga *et al.*, “Machine learning in bioinformatics,” *Brief. Bioinform.*, vol. 7, no. 1, pp. 86–112, Mar. 2006, doi: 10.1093/BIB/BBK007.

[51] I. Roman, J. Ceberio, A. Mendiburu, and J. A. Lozano, “Bayesian optimization for parameter tuning in evolutionary algorithms,” *2016 IEEE Congr. Evol. Comput. CEC 2016*, pp. 4839–4845, Nov. 2016, doi: 10.1109/CEC.2016.7744410.

[52] C. Vlachos and R. Kofler, “MimicrEE2: Genome-wide forward simulations of Evolve and Resequencing studies,” *PLOS Comput. Biol.*, vol. 14, no. 8, p. e1006413, Aug. 2018, doi: 10.1371/JOURNAL.PCBI.1006413.

[53] R. Poplin *et al.*, “Scaling accurate genetic variant discovery to tens of thousands of samples,” *bioRxiv*, p. 201178, Jul. 2018, doi: 10.1101/201178.

[54] W. McKinney, “Data Structures for Statistical Computing in Python,” *Proc. 9th Python Sci. Conf.*, pp. 56–61, 2010, doi: 10.25080/MAJORA-92BF1922-00A.

[55] T. Taus, A. Futschik, and C. Schlötterer, “Quantifying Selection with Pool-Seq Time Series Data,” *Mol. Biol. Evol.*, vol. 34, no. 11, pp. 3023–3034, Nov. 2017, doi: 10.1093/MOLBEV/MSX225.

[56] T. Taus, A. Futschik, and C. Schlötterer, “Quantifying Selection with Pool-Seq Time Series Data,” *Mol. Biol. Evol.*, vol. 34, no. 11, pp. 3023–3034, Nov. 2017, doi: 10.1093/MOLBEV/MSX225.

[57] “On the optimal parameter choice for /spl nu/-support vector machines | IEEE Journals & Magazine | IEEE Xplore.” https://ieeexplore.ieee.org/document/1233901 (accessed Nov. 05, 2024).

[58] J. D. Hunter, “Matplotlib: A 2D graphics environment,” *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90–95, 2007, doi: 10.1109/MCSE.2007.55.

[59] I. Wickramasinghe and H. Kalutarage, “Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation.,” *Soft Comput.*, vol. 25, no. 3, pp. 2277–2293, Sep. 2020, doi: 10.1007/S00500-020-05297-6.

[60] K. Kaushik *et al.*, “Multinomial Naive Bayesian Classifier Framework for Systematic Analysis of Smart IoT Devices,” *Sensors (Basel).*, vol. 22, no. 19, p. 7318, Oct. 2022, doi: 10.3390/S22197318.

[61] R. R. Yager, “An extension of the naive Bayesian classifier,” *Inf. Sci. (Ny).*, vol. 176, no. 5, pp. 577–588, Mar. 2006, doi: 10.1016/J.INS.2004.12.006.