Johannes Gutenberg-Universität Mainz

Fachbereich 10 – Biologie

Insights into Rapid Adaptation Patterns in Chironomus riparius through Advanced Bioinformatics

DISSERTATION

zur Erlangung des Grades

Doktor der Naturwissenschaften

am Fachbereich Biologie

der Johannes Gutenberg-Universität Mainz

Cosima Caliendo

geb. am 24.09.1992 in Bruchsal

1. Reviewer: Prof. Dr. Markus Pfenninger
2. Reviewer: Prof. Dr. Thomas Hankeln
Tag der Promotion 21.08.2025

Abstract

Understanding the genetic mechanisms underlying rapid adaptation remains a significant challenge in evolutionary biology. While populations can adapt to environmental changes within just a few generations, the genetic architecture behind these rapid responses is complex. Adaptive traits are often influenced by complex networks of interacting genes, each contributing small effects to the overall phenotype. This polygenic nature of adaptation creates substantial challenges for detecting and analyzing evolutionary change, as selection can act simultaneously on many genomic regions with subtle individual effects. Traditional methods struggle to capture these distributed patterns of selection, particularly during ongoing adaptation. This thesis presents a multi-faceted investigation combining methodological development, experimental evolution, and genomic analysis to examine rapid adaptation. First, I developed a novel computational approach combining unsupervised machine learning with a classic statistical test (OCSVM-FET) to detect adaptation patterns in sequencing data. Using simulated datasets, this method demonstrated superior performance in detecting selection signatures across a wide range of evolutionary scenarios, particularly for highly polygenic traits under ongoing selection.

The method was then applied to analyze a selection experiment on development time in the non-model organism *Chironomus riparius*. This experimental system revealed substantial phenotypic adaptation across multiple fitness-related traits over seven generations. More importantly, it provided an ideal test case for investigating the temporal dynamics of rapid adaptation in real populations.

Application of the OCSVM-FET approach to the experimental data revealed a novel two-phase adaptation process. The initial phase showed rapid phenotypic changes corresponding with selection on broadly shared metabolic pathways, while the subsequent phase demonstrated replicate-specific specialization in signaling pathways. Notably, despite minimal overlap in selected genomic positions between replicates, all populations converged on similar regulatory pathways, particularly in key cellular signaling networks. This work provides novel insights into the temporal dynamics of rapid adaptation

and demonstrates how populations can achieve similar phenotypic outcomes through distinct genetic trajectories while maintaining pathway-level convergence.

Zusammenfassung

Das Verständnis der genetischen Mechanismen, die einer schnellen Anpassung zugrunde liegen, bleibt eine bedeutende Herausforderung in der Evolutionsbiologie. Während Populationen sich innerhalb weniger Generationen an Umweltveränderungen anpassen können, ist die genetische Architektur hinter diesen schnellen Reaktionen komplex. Adaptive Merkmale werden oft durch komplexe Netzwerke interagierender Gene beeinflusst, die jeweils kleine Effekte zum Gesamtphänotyp beitragen. Diese polygene Natur der Anpassung schafft erhebliche Herausforderungen bei der Erkennung und Analyse evolutionärer Veränderungen, da Selektion gleichzeitig auf viele genomische Regionen mit subtilen individuellen Effekten wirken kann. Traditionelle Methoden haben Schwierigkeiten, diese verteilten Selektionsmuster zu erfassen, insbesondere während laufender Anpassungsprozesse. Diese Dissertation präsentiert eine vielschichtige Untersuchung, die methodische Entwicklung, experimentelle Evolution und genomische Analyse kombiniert, um schnelle Anpassung zu untersuchen. Zunächst entwickelte ich einen neuartigen computergestützten Ansatz, der unüberwachtes maschinelles Lernen mit einem klassischen statistischen Test (OCSVM-FET) kombiniert, um Anpassungsmuster in Sequenzierungsdaten zu erkennen. Unter Verwendung simulierter Datensätze zeigte diese Methode überlegene Leistung bei der Erkennung von Selektionssignaturen über ein breites Spektrum evolutionärer Szenarien, insbesondere für hochpolygene Merkmale unter laufender Selektion. Die Methode wurde dann angewendet, um ein Selektionsexperiment zur Entwicklungszeit im Nicht-Modellorganismus *Chironomus riparius* zu analysieren. Dieses experimentelle System offenbarte eine substanzielle phänotypische Anpassung über mehrere fitnessrelevante Merkmale im Verlauf von sieben Generationen. Noch wichtiger ist, dass es einen idealen Testfall für die Untersuchung der zeitlichen Dynamik schneller Anpassung in realen Populationen bot. Die Anwendung des OCSVM-FET-Ansatzes auf die experimentellen Daten offenbarte einen neuartigen zweiphasigen Anpassungsprozess. Die anfängliche Phase zeigte schnelle phänotypische Veränderungen, die mit der Selektion auf breit geteilte Stoffwechselwege korrespondierten, während die nachfolgende Phase eine replikatspezifische Spezial-

isierung in Signalwegen demonstrierte. Bemerkenswerterweise konvergierten alle Populationen trotz minimaler Überlappung in selektierten genomischen Positionen zwischen Replikaten auf ähnliche regulatorische Pfade, besonders in wichtigen zellulären Signalnetzwerken. Diese Arbeit liefert neuartige Einblicke in die zeitliche Dynamik schneller Anpassung und zeigt, wie Populationen ähnliche phänotypische Ergebnisse durch unterschiedliche genetische Trajektorien erreichen können, während sie eine Konvergenz auf Pfadebene beibehalten.

Acknowledgement

I would like to express my deepest gratitude to my supervisor, Prof. Dr. Susanne Gerber, for making this thesis possible. Her trust in my abilities and approach to research allowed me to develop independently as a scientist. Her supportive guidance and role as a female mentor in academia has been invaluable throughout this journey. I am profoundly grateful to Prof. Dr. Markus Pfenninger for introducing me to this fascinating research topic and entrusting me with the experimental work. His insightful guidance and expertise have significantly shaped my scientific thinking. The knowledge and skills I gained under his mentorship will undoubtedly prove valuable throughout my future career. Special thanks go to Prof. Dr. Thomas Hankeln for reviewing this thesis and for his trust in sharing his beloved research subject with me. His confidence in my abilities has been truly encouraging.

Beyond my supervisors, this work has been enriched by the invaluable contributions of several colleagues and friends. I am deeply grateful to those who dedicated their time and expertise to reviewing my thesis and manuscripts. Their thoughtful feedback, critical insights, and constructive suggestions have significantly enhanced the quality of this work. Namely, I'd like to mention here: Dr. Hristo Todorov, Dr. María Esther Nieto-Blázquez, Prof. Dr. Peter Tino, Dr. Daniele Zambon, Dr. Alejandro Ceron-Noriega, Dr. Pietro Verzelli, Dr. Barbara Feldmeyer, Dr. Nadine Erbeldinger and Dr. Stephan Weißbach. Big shout-out!

Beyond the academic support, I received an immeasurable amount of emotional and mental support during my doctoral journey that seems almost impossible to put into words. Hristo, thank you for your constant encouragement, uplifting words, and those essential after-work gatherings. Susanne K., thank you for being you! You are the most positive, supportive, and best partner in crime anyone could wish for. Your laughter brightens even the gloomiest days. Pauline, your support during the experiments meant incredibly much to me, you make problems seem easy and doable. Thanks for many 'feucht-fröhliche' evenings at your sweet home! Anna (aka the problemsolver) and Stephan (aka the softwareguru), you made this time truly unforgettable;

you were always there for me in good times and bad, and we grew together - may our festival group live forever. Nicolas, Stefan, Lioba, Vincent, Mohit, and Kanak - thank you for making work fun, leisure time enjoyable, and conversations meaningful. And to my reinforcement from Frankfurt: Burak, thank you for many good laughs and your sincere friendship. Maide, Shadi and Linda, I will deeply miss our safe-space girls' room conversations - thank you for your support! Special thanks to Bonn: Janko, Surbhit, Ulzii, and Kanaan for the cheerful moments in the Anthro building!

Finally, I want to express my deepest gratitude to family and friends who supported me through ups and downs outside university life. Nella and Paul, thanks for your emotional support and many funny series evenings. Steffen, you're the best friend I could imagine, thanks for your unconditional love and company throughout the years! Heartfelt thanks to Pietro - you've supported me so much, waited patiently, cooked delicious meals: you're the best. Finally, Thank you to my family, you made all of this possible - I love you.

Contents

Part I

Introduction

The development of methodologies to detect polygenic adaptation patterns presents significant challenges in evolutionary biology. This study originated from the application of unsupervised machine learning algorithms to investigate adaptive processes in *Chironomus riparius*. The inherent complexity of biological systems and the subtle nature of polygenic adaptation necessitated a refined analytical framework. Through systematic development and integration of biological context, this research established a comprehensive approach to investigating rapid adaptation mechanisms, synthesizing machine learning techniques with established population genetics methods.

Current Challenges in Detecting Polygenic Adaptation

Understanding how organisms adapt to environmental changes represents a fundamental scientific interest and has important societal implications for nature conservation biology, agriculture, and medicine [1]. Yet, despite its critical importance, unraveling the complex mechanisms of adaptation remains one of the main challenges in evolutionary biology. While traditional studies have focused on major-effect alleles, growing evidence suggests that many adaptive traits involve subtle changes across multiple genes - a process known as polygenic adaptation [2]. This mode of adaptation presents unique challenges for detection and analysis, particularly in experimental settings where rapid evolutionary responses are studied [3]. Rapid polygenic adaptation is a key phenomenon in evolutionary biology, allowing organisms to quickly adjust to changing environments [4]. This process plays a crucial role in species survival and diversification, particularly in the face of anthropogenic changes and climate fluctuations [5, 6]. Rapid adaptation can occur through various mechanisms, including selection on standing genetic variation, de novo mutations, and epigenetic changes [7]. The speed and extent of these adaptive polygenic changes can vary widely across taxa and environmental contexts, from subtle alterations in gene expression to dramatic morphological shifts [8]. For instance, studies have documented rapid adaptive responses in Galapagos finches during drought periods [9] or wing size adaptation in flies [10]. 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 populations [11, 12]. The advent of high-throughput sequencing technologies, particularly pool sequencing (Pool-Seq), has revolutionized our ability to track genetic changes in evolving populations [13]. However, current methodological approaches face significant limitations in detecting polygenic adaptation patterns. Therefore, traditional approaches to detecting selec-

tion signatures in population genomics have primarily relied on statistical methods such as Fisher's Exact Test (FET) or Cochran-Mantel-Haenszel test to identify significantly changed positions after selective events [14, 15, 16, 10, 17]. These methods are often implemented within specialized tools like PoPoolation2 [18], SNVer [19], or EM algorithms [20], primarily designed for identifying Single Nucleotide Polymorphisms (SNPs). While Chapter 3 provides detailed theoretical background on genetic diversity and heterozygosity, this section focuses on the practical challenges of these tools and the need for more sophisticated approaches.

PoPoolation2 employs FET to detect significant allele frequency changes (AFC), complemented by population genetic statistics such as Fst [21] and Tajima's D [22] to identify shifts in genetic diversity [14]. However, this approach faces several limitations. A critical issue is P-value inflation, where observed P-value distributions deviate from the expected uniform distribution under the null hypothesis, leading to increased Type I errors. Consequently, strict significance thresholds potentially miss important signals of polygenic adaptation [2].

While tools like PoPoolation2 offer flexibility in analysis scale, from predefined windows to base-by-base calculations, challenges remain in effectively detecting polygenic adaptation patterns. 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 and Tajima's D, while detailed, may not capture the coordinated shifts across multiple loci that characterize polygenic adaptation [2, 23]. While haplotype block information could potentially refine the selection of window sizes, its extraction from Pool-Sequencing data remains problematic due to the inherent challenges of this data type (as detailed in Chapter 2.3). Consequently, this approach is not yet fully optimized.

SNVer, on the other hand, works with initially established P-values for each pool, testing whether the minor allele frequency surpasses a specified threshold. Subsequently, it combines these individual pool P-values to derive an overall P-value using Simes methods, as described in [24]. This process enables the ranking of observed variants, with higher-ranking variants more likely to represent true positive SNPs [19]. However, the algorithm requires a

predetermined sequencing error rate, a task that can be challenging due to potential variability in sequencing error rates across positions.

The EM algorithm is applied to estimate minor allele frequencies using pooled sequencing data. It updates these estimations iteratively until convergence to a stable solution is reached. Once allele frequencies are determined, SNPs can be identified by comparing observed frequencies to those expected under a null model. Deviations from the null model may signal the presence of SNPs. However, estimation accuracy diminishes as the number of individuals per pool increases, likely due to increased pooling leading to greater information loss [20]. In this study, pools containing 100 individuals were utilized, potentially reducing sensitivity for detecting rare variants or small effect sizes. Additionally, population structure and relatedness among individuals can introduce complexities in association studies, potentially resulting in spurious associations if not appropriately addressed.

Finally, the above mentioned approaches mainly make assumptions about the data, which can lead to a high number of false positives and limit the ability to identify adaptive events. In contrast, this thesis elaborated on a classification approach that does not rely on prior assumptions about the data, offering a more flexible and nuanced method for identifying adaptive patterns.

Novel Approaches to Detecting Polygenic Adaptation

The limitations of traditional approaches in detecting polygenic adaptation patterns necessitate novel computational solutions. Machine learning methods, with their ability to identify complex patterns in high-dimensional data, offer promising alternatives for addressing these challenges. Recent developments in machine learning have shown increased potential for detecting complex patterns in genomic data [25]. In particular, unsupervised learning approaches have demonstrated success in identifying subtle patterns in several biological contexts [26, 27]. While various machine learning paradigms exist for genomic analyses, many present specific limitations for detecting polygenic adaptation patterns. Deep learning approaches, such as Neural Networks and Convolutional Neural Networks, require large training datasets [28] often unavailable in evolutionary studies. Clustering algorithms like K-means and hierarchical clustering make assumptions about the data structure [29] that may not hold for polygenic adaptation patterns. Dimensionality reduction techniques such as Principal Component Analysis, while widely used in population genetics, are limited to linear relationships [30] and may miss local patterns.

Given these considerations, this thesis explores a classification-based approach [31] for detecting polygenic adaptation patterns. Specifically, it enables the classification of allele frequency changes (AFC) into normal (non-anomalous) and abnormal (anomalous) classes, where the latter is hypothesized to signify components of polygenic adaptation patterns. Our approach employs two distinct classification paradigms: density-based and boundary-based classification. For the density-based classification approach, the Naive Bayesian Classifier (NBC) [32] was utilized, while for boundary-based classification, One-Class Support Vector Machines (OCSVM) [33] was employed. The fundamental distinction between these approaches lies in their underlying principles. NBC, as a density-based classifier, models separate probability distributions for anomalous and non-anomalous data, assigning classes based on posterior probabilities [34]. While it doesn't explicitly model complete density functions, it effectively uses density estimation principles through class-conditional probabilities. In contrast, OCSVM operates by projecting data into higher-dimensional space where anomalous and non-anomalous data can be separated by a boundary (linear, polynomial, or radial) [35, 33]. Each approach offers distinct advantages and limitations. NBC provides computational simplicity but requires careful model specification and may not transfer well between datasets. OCSVM, while computationally more intensive, offers greater flexibility through unsupervised learning and kernel-based approaches. In this thesis, the radial basis function (RBF) kernel was employed. Since combining multiple approaches often leads to improved precision [36], this study investigated hybrid methods linking both classifiers with FET, resulting in OCSVM-FET and NBC-FET approaches. This integration allowed me to leverage both the pattern recognition capabilities of machine learning and the statistical rigor of traditional hypothesis testing. The aim of combining these methods was to gain a more comprehensive understanding of the data, identifying both anomalous and significantly altered positions. This methodological development led to five approaches that were systematically tested and evaluated:

(a) Traditional Fisher's Exact Test (FET)

(b) Naive Bayesian Classifier (NBC)

(c) One-Class Support Vector Machine (OCSVM)

(d) Combined OCSVM-FET approach

(e) Combined NBC-FET approach

The selected techniques offer several key advantages that make them particularly suitable for detecting polygenic adaptation patterns in genomic data. Importantly, they operate effectively without requiring labeled training data [32, 37], a crucial advantage in evolutionary studies where ground truth data is often unavailable or difficult to obtain. NBC specifically is built upon clear statistical frameworks [34], enabling a straightforward biological interpretation of results - a critical feature for translating computational findings into meaningful biological insights. OCSVM, on the other side, learns patterns from the whole dataset [33] making it practical for studying polygenic adaptation across entire genomes rather than being limited to specific regions. Finally, these approaches demonstrated robust performance in the presence of noise and missing data in literature [36, 38], a common challenge in genomic sequences due to technical limitations in sequencing technologies and the inherent complexity of biological systems.

Validation Strategy and Application

The validation of computational approaches for detecting polygenic adaptation presents unique challenges, particularly given the subtle nature of the genetic changes involved. To address these challenges, a two-phase validation strategy was developed: first, a comprehensive evaluation using simulated data, and second, application to experimental Pool-Seq data from our *C. riparius* evolution experiment (Fig 1).

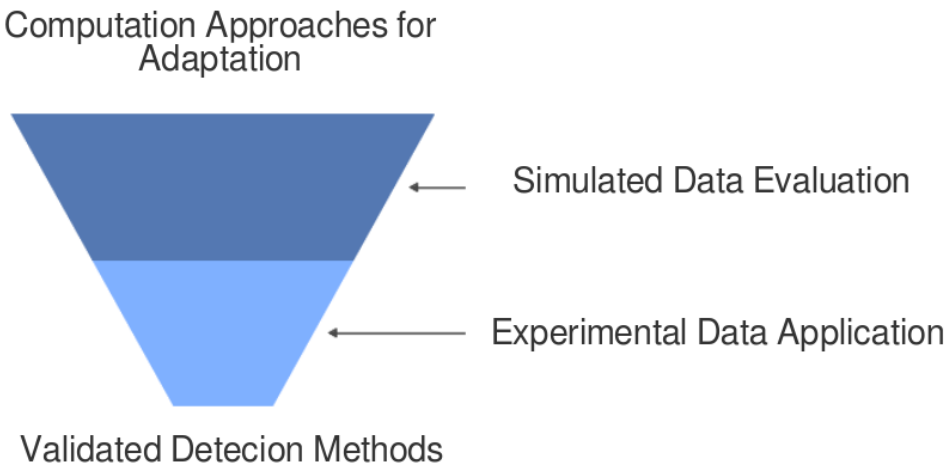

Fig. 1.: Validation Process for Computational Approach. Candidate methods include Fisher's Exact Test (FET), One Class Support Vector Machines (OCSVM), Naive Bayesian Classifier (NBC) and the combined approaches OCSVM-FET and NBC-FET

Simulation-Based Validation

Simulated data play a crucial role in method validation for population genomic analyses as it provides ground truth [39]. MimicrEE2 [40] was selected as simulation tool after careful comparison with alternatives such as SLiM [41] and forward-time simulators [42]. MimicrEE2 offers several advantages specifically relevant to our study, such as the ability to simulate Pool-Seq data directly, integration of selection on quantitative traits and comprehensive manuals. Simulated data provides several crucial advantages for validating our methods. Most importantly, they provided known ground truth data, enabling accurate assessment of method performance against verifiable outcomes. Additionally, simulations allow for control over evolutionary parameters, making it possible to systematically investigate how different selective pressures and demographic scenarios affect detection accuracy [39]. Furthermore, the simulation framework enabled testing of method robustness across multiple scenarios, providing insights into the reliability and limitations of our approaches under varying conditions.

To quantify method performance across simulated scenarios, standard classification metrics were employed: Area Under the Receiver Operating Characteristic Curve (AUC-ROCC) to assess overall classification performance, False Positive Rate (FPR) to measure method specificity, and accuracy to evaluate prediction quality [43, 44]. These metrics, computed across varying parameter combinations including selection coefficients and numbers of selected loci, enabled systematic evaluation of method robustness and identification of optimal operating conditions for different evolutionary scenarios.

Experimental-Based Validation through Evolution Studies

Following simulation-based validation, the best-performing approach was applied to an experimental Pool-Seq data set from *C. riparius* populations. The non-biting midge *Chironomus riparius* serves as an ideal model organism for validating our approach, because it demonstrates great capacity for rapid adaptation. Research has shown that these populations can undergo significant evolutionary changes within controlled laboratory settings [45]. This adaptability extends to the real world, with *C. riparius* populations exhibiting the ability to adjust their traits in response to seasonal variations in their environment [46]. This suggests that *C. riparius* is capable of responsive genetic mechanisms that allow it to rapidly adapt in dynamic environments. Yet, the specific genetic mechanisms underlying this adaptation remain elusive. Much

of the current understanding relies on simulations, which lack the complexities of real-world scenarios. This study aimed to bridge this knowledge gap by conducting a controlled laboratory experiment, applying selection pressure on four replicate *C. riparius* populations to investigate:

(a) The feasibility of accelerating developmental time (shortening emergence time).

(b) The number of generations required for rapid phenotypic adaptation to occur.

(c) The population trajectories across generations, revealing the dynamics of rapid adaptation at the genetic level.

Our experimental design employed a selection-pressure approach involving 4,000 non-biting midges *C. riparius*, divided into four replicate groups. These populations were subjected to emergence pressure by allowing only the first 50 hatching males and females per replicate group to reproduce. This design was carried through seven generations, with comprehensive Pool-Seq data collection and analysis at key generational timepoints. The experimental design aligns with and builds upon successful evolutionary studies across various model organisms. In *Drosophila melanogaster*, similar experimental timeframes (5-10 generations) have demonstrated significant adaptive responses to different temperature regimes [14] and desiccation stress [47]. Our choice of population size (1,000 individuals per replicate) falls within the optimal range identified in previous insect evolution studies, where populations of 500-2,000 individuals have shown effective adaptation while maintaining sufficient genetic diversity [48, 49]. The use of multiple replicate populations provides robust statistical power while allowing to investigate convergent and divergent evolutionary trajectories, a design feature successfully employed in multiple evolution studies [50]. We performed Pool-Sequencing, since this approach is used for genomic analysis in various studies, [13, 51, 52], offering high resolution while remaining cost-effective.

Application to Experimental Data

The transition from simulated to real data presents several challenges. Unlike simulated data, real experimental data lack known selection targets, making direct validation of results more complex. Additionally, natural populations possess complex demographic histories that can confound selection signals [53]. Technical noise inherent in sequencing data adds another layer of complexity, while potential epistatic interactions between loci create intricate

patterns [54] that may be difficult to distinguish from direct selection effects. To address these challenges, a two-threshold validation strategy was implemented (Fig 2). This approach combines a broader significance level to capture widespread polygenic adaptation patterns and a stricter threshold to identify strong candidate targets of selection. This dual-threshold approach aligns with recent understanding of polygenic adaptation, which suggests that adaptive responses often involve both subtle, widespread changes and stronger selective sweeps at key loci [14, 51]. Studies of experimental evolution have demonstrated that replicate populations under similar selection pressures often show a combination of convergent evolution at major-effect loci and divergent responses across the broader genome [55, 56].

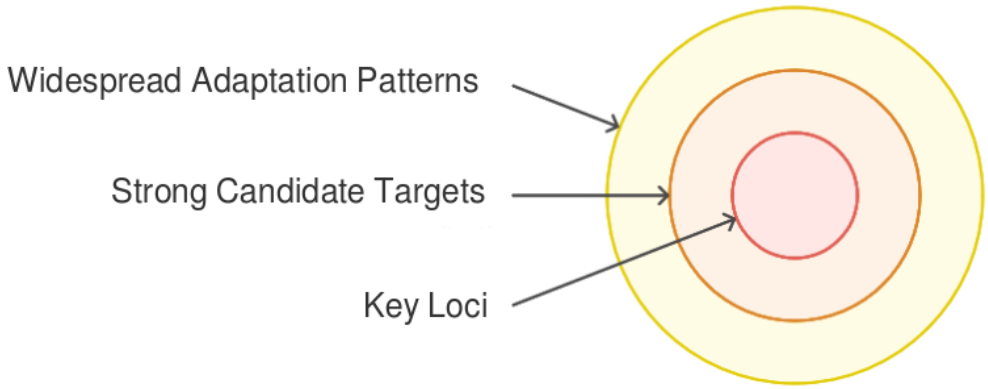

Fig. 2.: Polygenic Adaptation Strategy using validated computational approach (OCSVM-FET). In the first step, One Class Support Vector Machines (OCSVM) is applied to identify widespread adaptation patterns, where strong candidate targets are extracted setting Fisher's Exact Test (FET) threshold to a broad range. Candidate Key loci targeted by rapid polygenic adaptation are further narrowed down by setting a more stringent FET threshold.

Previous studies of emergence time selection in insects have identified several candidate pathways and genes. For instance, research in *Drosophila* has implicated two developmental pathways known to regulate growth of morphological structures: a limb-patterning pathway that specifies the location and shape of a structure, and the insulin pathway, which modulates trait growth in response to larval nutrition [57]. Studies in the same organism implicate several core RNA pathway genes, that show evolutionary patterns that may relate to developmental timing [58], yet the genes involved in insect developmental time are not fully understood. While these studies may provide a valuable framework for interpreting our results, the genetic architecture of emergence time in *C. riparius* may involve distinct pathways given

its aquatic larval phase and specific life history adaptations - which have not been investigated so far.

To investigate this complex trait architecture, we employed two analytical thresholds to capture different aspects of the adaptation process. The strict threshold analysis aimed to identify "strong-effect variants" - genomic regions showing immediate and substantial frequency changes, likely representing primary targets of selection and key regulatory elements. The broader threshold analysis captured "broad-effect variants" - regions showing more moderate frequency changes that might include both secondary targets and linked variants that accumulate changes over time. This dual approach allows us to examine both immediate and gradual selection responses, providing insight into how rapid adaptation proceeds through a combination of primary selection targets and broader genomic changes. Such an approach helps illuminate the temporal dynamics of polygenic adaptation while acknowledging the complex interplay between direct selection and linked genetic changes.

Through this systematic evaluation of both simulated and experimental data, this work aimed to develop a reliable approach for detecting and characterizing the full spectrum of polygenic adaptation patterns in evolving populations, from subtle polygenic effects to stronger selective signatures.

Thesis Structure

This thesis investigates rapid polygenic adaptation through three major themes that are systematically addressed throughout each chapter. The themes - methodological development, experimental evolution and practical application - are reflected in the organization of the background, materials and methods, and results chapters. Specifically, each chapter is subdivided into three corresponding sections: "Evaluating Methods for Polygenic Adaptation Detection", "Experimental Evolution: Selection on Emergence Time in *C. riparius*," and "OCSVM-FET Analysis of *C. riparius* Adaptation Patterns". This parallel structure allows readers to follow each aspect of the work independently through the thesis, from theoretical foundations to implementation and outcomes. The first theme addresses our experimental evolution study, documenting both phenotypic changes and genomic patterns in populations under selection for accelerated development. The second focuses on our methodological development, comparing traditional approaches with novel machine learning methods to establish a robust framework for detecting polygenic adaptation patterns. The third theme applies our validated methodology

to the experimental data, revealing insights into the genetic architecture of rapid adaptation.

Part II

Background

1 Evaluating Methods for Polygenic Adaptation Detection

This section explains two machine learning approaches for detecting genetic adaptation: One Class Support Vector Machines (OCSVM) and Naive Bayesian Classifier (NBC). NBC works by calculating probabilities based on Gaussian distributions to classify data points into three categories (non-anomalous and two types of anomalous data), while OCSVM uses kernel functions to separate normal from anomalous data points in higher-dimensional space. Both methods were fine-tuned using previously published data from *C. riparius* and validated using simulated data.

1.1 Candidate Approaches

Different methods were tested and compared for identifying polygenic patterns: FET, OCSVM, NBC, and their combinations (OCSVM-FET and NBC-FET). These methods were applied to both real allele frequency data and simulated data to evaluate their performance. Using simulated data sets makes it possible to set a ground truth on which several pipelines can be tested and evaluated. Although simulations might not perfectly represent the real-life situation, they can be a great operator for testing and comparison purpose. Currently, two widely used software tools providing the ability to simulate evolutionary scenarios: MimicrEE2 [40] and SLim 3 [41]. Although SLim3 comes with a vast set of parameters, a graphical user's interface (GUI) and a detailed user's manual, the tool requires use of the less known scripting language Eidos, which makes it harder to directly apply. Furthermore, settings the correct parameter values is non-trivial. For practical reasons, I therefore decided to employ MimicrEE2 in the current study. Another advantage of this program is that the output file can be directly processed with Popoolation. Furthermore, the user can input their own haplotype file. This feature

makes it possible to use sequencing data from the organism under examination, *C. riparius*. This strategy is much closer to the real life experiment than simulated data. Mimicree2 enables the replication of scenarios like rapid polygenic adaptation by using 'QT' mode (quantitative trait model) and variable selection coefficient for pronounced adaptive responses. Further parameters as recombination rate, heritability, number of generations to simulate and selection regime helps to optimize the simulation of an adaptive event.

1.2 Fisher's Exact Test

The Fisher's exact test, introduced by the researcher R. Fisher [59], is a statistical hypothesis testing method used to assess the presence of nonrandom associations between two categorical variables. It evaluates whether the observed distribution of the data could have occurred by chance, given fixed row and column in the contigency table. This is accomplished by examining all possible combinations of the data that would result in the same marginal totals, and then determining the probability of observing the actual data distribution. The test is grounded in the hyper-geometric distribution and yields an exact P-value for the observed data. The test is implemented for 2x2 contingency tables and can be generalized to larger contingency tables and other discrete data problems [60]. Its capacity to deliver exact P-values derived from the specific randomization process underlying the observed data makes it a useful tool for analyzing genetic data, particularly in situations where the assumptions of alternative tests may not be met. Additionally, the Fisher's exact test has been demonstrated to be a good exact test for comparing proportions in binomial experiments [61]. Unlike traditional hypothesis tests, which often rely on predefined statistical functions, exact test directly compute probabilities from the observed data distribution, increasing their robustness in small sample sizes. Therefore, Fisher's exact test assigns a probability to every possible arrangement of numbers in the contingency table under the null hypothesis, assuming that the numbers in each row and column were randomly sampled from the total pool of observations [62]. This probability is computed based on the hyper-geometric distribution.

The observed counts in a contingency table is denoted as follows:

	Column 1	Column 2	Row totals
Row 1	n_{11}	n_{12}	$n_{1\cdot}$
Row 2	n_{21}	n_{22}	$n_{2\cdot}$
Column totals	$n_{\cdot 1}$	$n_{\cdot 2}$	n

Tab. 1.1.: 2×2 Contingency Table Notation

Then the formula for calculating the probability of the observed arrangement of counts under the null hypothesis is:

$$P(n_{11}, n_{12}, n_{21}, n_{22}) = \frac{n_{1\cdot}! \cdot n_{2\cdot}! \cdot n_{\cdot 1}! \cdot n_{\cdot 2}!}{n! \cdot n_{11}! \cdot n_{12}! \cdot n_{21}! \cdot n_{22}!} \tag{1.1}$$

where n is the total sample size.

To calculate the P-value, one aggregates all configurations with a probability smaller (or more unlikely) than the observed configuration. Therefore, the P-value represents the probability of observing a configuration as unlikely (or more unlikely) under the null hypothesis.

Due to the fact that for any chosen P-value threshold α and m numbers of testings, we expect $mx\alpha$ test to pass, there is a certain False Discovery Rate (FDR) to expect. The Benjamini-Hochberg correction is a method used to adjust P-values obtained from multiple hypothesis tests to control the FDR [63]. The correction is applied as follows:

- Calculate P-value for each of the tests performed
- Rank the tests by their test statistics so that most significant is 1st etc.
- Move down the ranked list and reject null hypothesis for the r most significant tests, where the P-value is less than $\frac{r*\alpha}{m}$, where m is the number of tests and α is the FDR threshold

1.3 Naive Bayesian Classifier

Using Bayesian statistics offers a solid foundation for determining a set of classes and their descriptions that are most likely to explain a given data set [34, 64]. In the context of this analysis, I employed an NBC to distinguish

between normal (non-anomalous) and anomalous data points based on their allele frequency changes in a three-dimensional (3D) space.

The NBC is a probabilistic classifiers based on Bayes' theorem, where the fundamental idea involves modeling the probability density functions (PDFs) of data set distributions, in this case for non-anomalous and anomalous data. Therefore, each allele frequency serves as a feature, and separate PDFs for both normal and anomalous data points are modelled. By fitting appropriate probability distributions to the observed data, the likelihood of a given allele frequency belonging to either the normal or anomalous distribution can be estimated.

In a three-dimensional space, the Bayesian classifier tracks data patterns using two key components. First, it calculates the average position (mean) as a point in 3D space. Second, it measures how the data spreads out (variance) using a matrix that captures relationships between all three dimensions. This is more complex than two-dimensional analysis, where only single values for average position and spread are considered.

The derivation of the Bayes theorem can be found in the appendix. The following workflow and formulas for the NBC are described by [65], [32], [66]: We assume there are k classes, $C1, C2, ..., CkC_1, C_2, ..., C_k$. Each sample is represented by an n-dimensional vector, $X = \{x_1, x_2, ..., x_n\}$, signifying n measured values of n attributes $A_1, A_2, ..., A_n$, respectively.

Given a sample X, the classifier predicts that X belongs to the class with the highest probability conditioned on X. In other words, X is predicted to belong to class C_i if and only if

$$P(C_i|X) > P(C_j|X) \text{ for } 1 \leq j \leq m, \text{ where } j \neq i \tag{1.2}$$

Thus, we identify the class that maximizes $P(C_i|X)$. The class C_i for which $P(C_i|X)$ is maximized is called the maximum posterior hypothesis. According to Bayes' theorem:

$$P(C_i|X) = \frac{P(X|C_i) \cdot P(C_i)}{P(X)} \tag{1.3}$$

Evaluating $P(C_i|X)$ for data sets with numerous attributes can be computationally resource-intensive. To compensate for this, the naive assumption of class conditional independence is employed. This assumption implies that the attribute values are conditionally independent of one another given the class label of the sample. Mathematically, this translates to:

$$P(X|C_i) \approx \prod_{k=1}^{n} P(x_k|C_i) \tag{1.4}$$

When attribute A_k is continuous-valued, it is common to assume that the values follow a Gaussian distribution with a mean μ and a standard deviation σ, defined as:

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{1.5}$$

in one-dimensional space. In 2D the mean μ becomes a multi-dimensional vector and the deviation σ becomes a $n \times n$ covariance matrix Σ :

$$g(x; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \tag{1.6}$$

To predict the class label of X, $P(X|C_i) \times P(C_i)$ is evaluated for each class C_i). The classifier predicts that the class label of X is C_i if and only if it maximizes $P(X|C_i) \times P(C_i)$

Different settings for μ and Σ will vary the modeling of the data, as displayed with Toy Data Set in Fig. 1.1

The majority of the data is assmued to be non-anomalous characterized by small allele frequency changes (AFC). In contrast, anomalous data is characterized by allele frequencies that shift from low values before an adaptive event to high frequencies or vice versa. In a scatter plot, comparing allele $\text{frequency}_{\text{before}}$ against allele $\text{frequency}_{\text{after}}$, the anomalous data points are present in the upper left corner or the bottom right corner. Fig. 1.2 shows the theoretical PDF models for the non-anomalous data and the two PDF models of anomalous data. Therefore, three classes were employed: one class for the

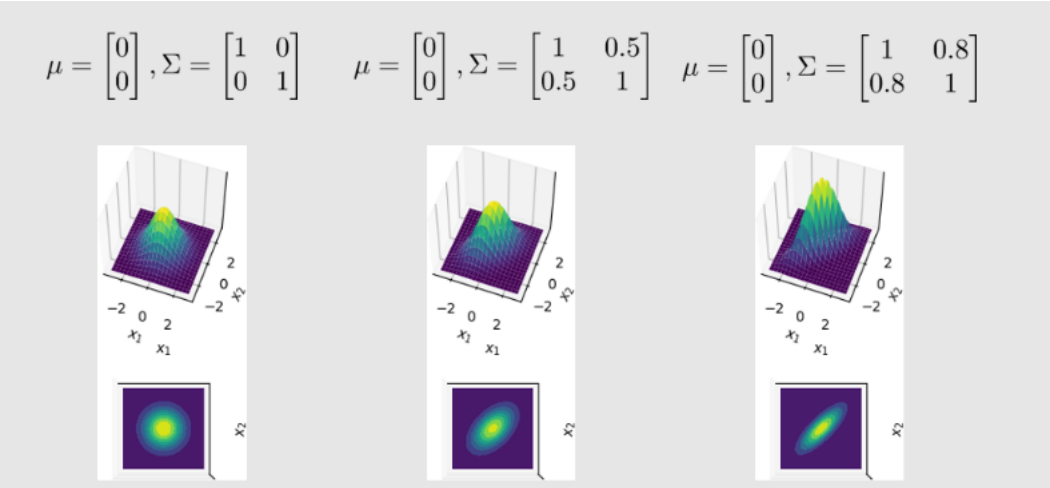

Fig. 1.1.: Example illustration of the effects of different parameter settings for the covariance matrix. The modelling of the data changes depending on the values of the covariance parameters. Code for figures are taken from [@67]

non-anomalous data, a second class for the AFC from high to low and a third class for the AFC from low to high.

To fit the parameters for μ and Σ to real-life data, in this thesis the already published cold-snap data set [68] was used (see section 1.6). To ascertain the accuracy of the fine-tuned parameters, simulated data was used (see section 1.1).

1.4 One Class Support Vector Machines

The OCSVM algorithm is an unsupervised machine learning technique for collective anomaly detection [69]. The algorithm is described as performing best in finding collective anomalies in an unsupervised manner compared to alternative approaches [70, 36]. In the current study, the algorithm was utilized to detect unusual patterns within AFC data that deviate from the expected distribution (anomalies) with the aim of narrowing down the genetic targets of rapid adaptation processes. OCSVM is a variation of an One-Class Classifier (OCC) algorithm [33], categorized as boundary-based . In theory, the decision weather to classify a given data point as non-anomalous or anomalous depends on two instances [71]:

- a parameter to calculate the distance of a sample to the target class (i.e. anomaly class)

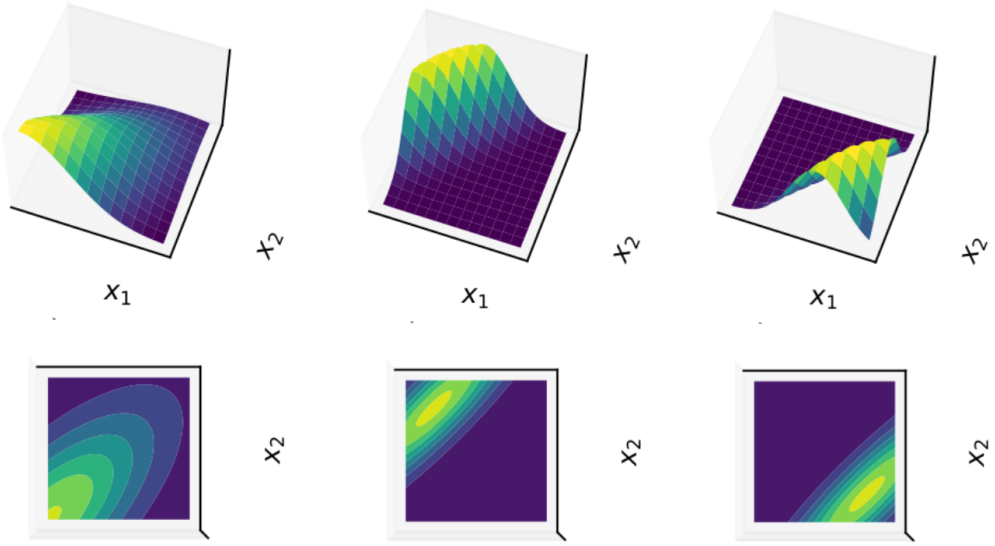

Fig. 1.2.: Schematic illustration of the three probability density function in 3D (above) and 2D (below), modelling A) non-anomalous data characterized by small AFC B) anomalous data in the upper left corner, characterized by high AFC and C) in the down right corner in a scatter plot, characterized by high AFC before the adaptive event

- a threshold limit to compare the distance and accept or reject the data point as non-anomaly

The algorithm uses kernel functions, such as polynomial kernel function or the radial basis function (RBF) kernel, to perform the so-called "kernel trick". This enables the algorithm to implicitly compute dot products in a higher-dimensional space without explicitly transforming the data. In this way, a function can be calculated that separates the data into two classes, as schematically shown in Fig. 1.3.

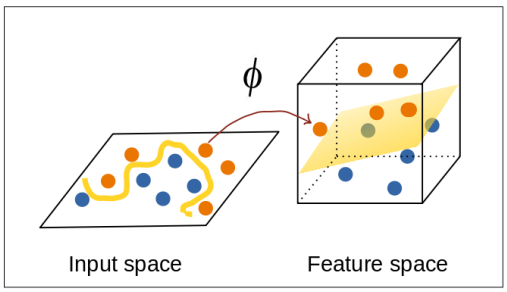

Fig. 1.3.: Schematic illustration of OCSVM's kernel trick: data are projected into a higher dimension where they are separable by a function, image adapted from [72]

Therefore, the core of the OCSVM algorithm involves constructing a hyper-sphere in the higher-dimensional space. This hyper-sphere serves as the decision boundary, separating normal data points from potential anomalies. During the optimization process, the algorithm aims to maximize the margin around the hyper-sphere while minimizing the number of anomalies (to avoid false positive results) [33].

The optimization objective involves the minimization of the sum of two terms: The term that characterize the hyper-sphere (indirect margin maximization) and the term that models the number of anomalies. The latter essentially controls the trade-off between maximizing the margin and allowing some data points to be classified to the data points outside the boundary (i.e. anomaly class) The balance between these objectives is controlled by the parameters ν [35], [73].

The mathematical formulation in its basic essence can be broken down to 4 major steps, plus the "kernel trick" step:

1. Hyper-sphere Construction in a higher dimensional space: In this section, ϕ is used to represent the feature space images in input space of the corresponding patterns. Thus $\phi(x_i)$ represents the data points $x_1, x_2, ..., x_i$ into higher dimensional space

$$x_i := \phi(x_i) \tag{1.7}$$

$$\|\phi(x) - \phi(c)\|^2 = r^2 \tag{1.8}$$

$\phi(c)$ is the center of the hyper-sphere. The center is determined by the algorithm as a complex relationship involving the support vector machines, their weight and the kernel functions' transformation. r denotes the radius of the hyper-sphere.

2. Decision Function:

$$f(x) = (w, \phi(x)) - p \tag{1.9}$$

where w is the weight vector and p is a threshold.

3.1 Optimization Objective 1: Maximize Margin

$$min(_w, \epsilon, p)\frac{1}{2}\|w\|^2 \tag{1.10}$$

w must be optimized to find the minimum via the derivative given in equation [1.10]. The minimization of this equation yields in an indirect maximization of the margin. The parameter ϵ_i denotes the slack variable for point i that measures the deviation of data point from the decision boundary, i.e. allows data points to lie on the other side of the decision boundary.

3.2 Optimization Objective 2: Minimize number of support vector machines outside the margin (i.e. anomalies)

$$\frac{1}{\nu_n}\sum_{i=1}^{n}\epsilon_i - p \tag{1.11}$$

ν is a parameter that controls the upper bound on the fraction of margin errors (i.e. regularization/trade-off parameter) and is one of the two major parameters to set in the OCSVM algorithm. Higher values for ν allow for more slack (miss-classification) resulting potentially in a wider margin and permitting more anomalies - the algorithm becomes more tolerant for outliers. Conversely, as ν decreases, the algorithm becomes less tolerant to anomalies, leading to a more narrow margin and potentially higher precision. The parameter n denotes the number of training samples. The sum of the two optimization objectives is displayed by equation[1.12]

$$min(_w, \epsilon, p)\frac{1}{2}\|w\|^2 + \frac{1}{\nu_n}\sum_{i=1}^{n}\epsilon_i - p \tag{1.12}$$

subject to

$$f(x_i) >= p - \epsilon_i, \epsilon >= 0, i = 1, 2, ..., n \tag{1.13}$$

Data points with $\epsilon_i > p$ are considered to be anomalies.

Kernel Trick, RBF: dot product in higher dimensional space is approximated by

$$k(x, x_i) = exp(-\frac{|x - x_i|^2}{2\sigma^2}) \tag{1.14}$$

σ is a hyper-parameter that controls the spread of the gaussian function. x and x_i are input vectors, $||x\breve{}x_i||^2$ is the squared Euclidean distance between the two vectors.

Kernel Trick, Polynomial: dot product in higher dimensional space is approximated by

$$k(x, x_i) = (\gamma(x, x_i) + r)^d \tag{1.15}$$

γ is a scaling factor and r denotes a coefficient representing the constant term in the polynomial. d is the degree of the polynomial function.

The goal is to find the weight vector w, the threshold p, and the slack variable ϵ_i that minimize the objective function (also often referred to as loss function or cost function) while satisfying constraints.

The deduction from the theoretical to the mathematical objective can be stated by the following procedure: Let the function $g()$ be defined as follows

$$g(x) = w^T \phi(x) - p \tag{1.16}$$

Then equation [1.16] shows the decision function that OCSVM uses in order to identify non-anomalous points

$$f(x) = \text{sgn}(g(x)) \tag{1.17}$$

sgn stands for sign-function, which takes a real number x as input and returns

-1 if x is negative (indicating anomaly),

0 if x is exactly 0 and

1 if x is positive (indicating non-anomaly) .

Given equation[1.12] , subject to

$$w^T \phi(x_i) >= p - \epsilon_i \tag{1.18}$$

the loss function can be stated by the distance to the decision boundary. The decision boundary is defined as

$$g(x) = 0 \tag{1.19}$$

Then the distance of any arbitrary data point to the decision boundary can be computed as

$$d(x) = \frac{|g(x)|}{|w|} \tag{1.20}$$

So the distance that the algorithm attempts to maximize can be obtained by plugging the origin into the equation yielding p/|w|, or minimizing

$$\frac{|w|^2}{2} - p \tag{1.21}$$

As we can rewrite equation [1.12] to

$$min(_w, \epsilon, p)\frac{\|w\|^2}{2} - p + \frac{1}{\nu_n}\sum_{i=1}^{n}\epsilon_i \tag{1.22}$$

$$\tag{1.23}$$

the second part of the objective [1.22] is the minimizing the slack variable ϵ_i for all points, controlled by $\frac{1}{\nu_n}$, therefore the trade-off parameter nu.

The second most important parameter to set using OCSVM algorithm, besides ν, is γ. This parameter is the kernel coefficient. In the context of the RBF kernel, γ controls the spread of the Gaussian function and is indirectly implicated in equation [1.14] that often can be also found as being defined as:

$$k(x, x_i) = exp(-\gamma * |x - x_i|^2) \tag{1.24}$$

where γ corresponds to $\frac{1}{2\sigma}$. In context of the polynomial kernel, the influence of the parameter $gamma$ becomes quite obvious from equation [1.15] as being the scaling factor of two vectors x, x_1 Therefore, γ determines the influence of each training example on the decision boundary. A larger value of γ will cause a more narrow curve, focusing more on individual data, whereas a smaller value will cause a wider curve of the decision boundary. Algorithm 1 shows the process of the OCSVM in pseudo code.

Algorithm 1 One-Class SVM Anomaly Detection

Require: Scaled allele frequency data

Initialize OC SVM model with specified parameters (kernel, gamma, nu, degree, cache size)
Train the model on the scaled allele frequency data:
    Find the optimal hyperplane that best separates the data
Predict anomalies and non-anomalies:
for each data Point: **do**
    classify it as either an anomaly or non-anomaly
    Identify indices of anomalies and non-anomalies
    Extract values corresponding to anomalies and non-anomalies
end for
Save anomaly and non-anomaly data to separate files

Return Anomalies.csv, NoAnomalies.csv

1.5 Overlap with Fisher's Exact Test

Next to the two candidate approaches OCSVM and NBC, a combined approach that integrates the binary classification results from OCSVM and NBC with the statistical hypothesis testing of FET was developed. This method aimed to leverage the strengths of both machine learning and traditional statistical methods. Several studies have demonstrated the benefits of integrating multiple analytical approaches in genomic analysis. For instance, studies have successfully combined supervised learning with statistical testing to identify genetic variants [74, 75]. Similarly, Sarno et al. showed that overlapping results from different methodological approaches can increase the confidence in detected anomalies [76].

1.6 Parameter Fine-Tuning

The crucial parameters μ and Σ for NBC as well as μ and γ for OCSVM were tuned on real-life and previously published data sets by Pfenninger et. al. [68]. This data, derived from a natural experiment on *C. riparius*, captured genome-wide allele frequency changes before and after a cold snap event, indicating a potential rapid, polygenic adaptation in the population. The study involved DNA extraction from larval head capsules, followed by rigorous quality assessment and concentration measurement. Subsequent steps included

whole-genome pool sequencing with trimming and quality control. Mapped reads were utilized for SNP calling and allele frequency estimation. Selected SNP loci were pinpointed based on significant changes in allele frequencies. To gain insights into the long-term selection regime, Tajima's D was calculated. The behavior of candidate loci over time was evaluated, contrasting their correlation with randomly selected SNPs.

In this paper, 19 genetic markers were identified exhibiting substantial shifts in allele frequencies before and after the cold snap, signifying potential selection events. Some of these markers were found in close genomic proximity. The increasing prevalence of certain alleles at these markers was associated with specific genetic patterns. Accounting for these patterns, 10 independent genetic loci were identified as being influenced by selection. These genetic loci were used as true positives to fine tune needed parameters for the analysis strategies employed in this thesis. The estimated parameter settings were then tested for accuracy and precision on simulated data (see section 1.1)

2 Experimental Evolution: Selection on Emergence Time in *C. riparius*

This section introduces rapid phenotypic adaptation in evolutionary biology, focusing on the midge species *Chironomus riparius* as a study organism for investigating quick evolutionary responses. The research aims to examine the acceleration of developmental time through a controlled selection-pressure experiment using 4,000 midges divided into 4 replicate groups. The text also covers fundamental evolutionary theory, *C. riparius* biology, and pool-sequencing methodology as a cost-effective tool for studying genetic variation in populations.

2.1 Evolutionary Theory and Adaptation

Evolution, the change in heritable characteristics of biological populations over successive generations, is fundamentally driven by selection pressures in the environment [77]. Selection pressure refers to any factor in an organism's environment that impacts its survival and reproductive success, thereby influencing the frequency of certain traits in subsequent generations [78]. In natural populations, selection pressures can arise from various sources, including changes in climate, resource availability, predation, or competition. These various types of pressure create differential reproductive success among individuals with different phenotypes, leading to the process of natural selection. Over time, this process can result in populations becoming better adapted to their environments, as beneficial traits become more prevalent.

The study of rapid adaptation has gained increasing importance in evolutionary biology. While evolution was traditionally viewed as a slow, gradual process, recent research has shown that significant adaptive changes can occur within just a few generations, especially in response to strong selection pressures [79]. This phenomenon, often referred to as contemporary

evolution or rapid adaptive evolution, has been observed in various species responding to human-induced environmental changes, including climate change and pollution [80]. Understanding rapid adaptation is crucial for several reasons. First, it provides insights into the mechanisms of evolutionary change and the limits of adaptive potential in natural populations [7]. Second, it has important implications for conservation biology and the management of ecosystems in the face of global environmental change [5]. Finally, studying rapid adaptation can inform predictions about how species might respond to future environmental challenges, which is particularly relevant in the context of ongoing climate change and habitat alteration [1]. By focusing on rapid adaptation in controlled experimental settings, such as subjecting midge populations to specific selection pressures, valuable insights into the dynamics of evolutionary processes and the genetic basis of adaptation can be gained. This approach allows for the direct observation of evolutionary change in real-time, providing a powerful tool for testing evolutionary theories and understanding the mechanisms underlying adaptive responses in natural populations.

2.2 Chironomus riparius as a study organism

Non-biting midges (Chironomidae, Diptera) are aquatic insects found across various water bodies throughout the holarctic [81]. These short-lived adults, which appear mostly in swarms above water surfaces, play essential roles in freshwater ecosystems. By acting as bacterial carriers, they influence benthic communities and nutrient cycling [82]. In the larval stage, for example, the organism stays at the bottom of water bodies, where it can feed on organic material and bacteria and thus nourish the basis for benthic food webs [83]. Moreover, the species builds a crucial part of the diet for aquatic as well as terrestrial species. For example, Chironomidae are the most important food for some birds (e.g. mallard ducklings) and other invertebrates, fish and amphibians [84].

Chironomus riparius (*C. riparius*), commonly known as the harlequin fly, is a species of non-biting midge that has gained prominence as a valuable non-model organism in ecological and evolutionary studies. The life cycle of *C. riparius* comprises four larval stages, a brief pupal stage, and the adult midge, as shown in Figure 2.1. Larval and pupal stages are generally bound to aquatic habitats, while the imago lives aerial and mostly located near to their

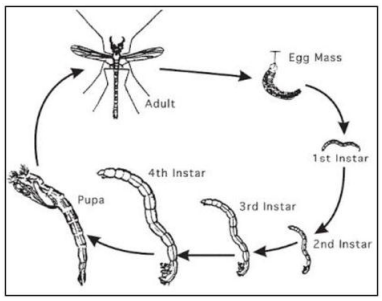

Fig. 2.1.: Chironomid life cycle, taken from [85]

emergence habitats. Adult midges form large mating swarms, with females typically producing a single egg mass containing several hundred eggs a few days after hatching. Larvae hatch shortly after, and the entire life cycle can be completed within four weeks [82]. These species are typically multivoltine, with up to 15 generations per year in Europe [86]. The species' large demographic and effective population size (>1,000,000) [87] combined with the significant number of offspring produced per breeding pair (400–800) [46] facilitate rapid adaptation. This was confirmed by several research studies on temperature, which have revealed local and temporal adaptations among seasons [87, 45]. Several characteristics make *C. riparius* ideal for laboratory evolutionary studies [88]: ease of culture, sensitivity to toxicants and environmental changes, a short life cycle and sufficient tissue mass for genetic analyses [89].

C. riparius possesses a diploid chromosome number of 2n = 8, with a genome size estimated at approximately 200 Mb [90]. This relatively compact genome, combined with the species' polytene chromosomes, facilitates genetic studies and the identification of genomic regions under selection [91]. The genetic diversity of *C. riparius* populations has been extensively studied, revealing significant variation both within and between populations. Nowak et al. [92] demonstrated high levels of genetic diversity in natural populations using microsatellite markers, suggesting a substantial pool of standing genetic variation that could facilitate rapid adaptation. This diversity is further evidenced by the species' ability to form distinct ecotypes adapted to various environmental conditions, including different thermal regimes and pollutant exposures [93].

Since genomic resources for *C. riparius* have become more and more available, *C. riparius* is becoming popular as a model organism for molecular ecology. Therefore, *C. riparius* has been utilized in various research contexts,

particularly in ecotoxicological studies due to its sensitivity to environmental pollutants. Vogt et al. [94] used *C. riparius* to investigate the effects of endocrine-disrupting chemicals on development and reproduction, demonstrating the species' utility in assessing anthropogenic impacts on freshwater ecosystems. In evolutionary biology, *C. riparius* has proven valuable for studying rapid adaptation. Nemec et al.[95] conducted a multi-generation experiment examining the species' adaptive responses to temperature changes, observing significant shifts in life-history traits over relatively few generations. This work highlighted the potential of *C. riparius* for investigating rapid evolutionary responses to climate change. The species has also been used to study the genomic basis of adaptation. Pfenninger et al. [68] employed next-generation sequencing techniques to identify genomic regions associated with thermal adaptation in *C. riparius*, providing insights into the molecular mechanisms underlying rapid environmental adaptation.

2.3 Pool-sequencing: A Powerful Tool for Population Genomics

Pool-sequencing (Pool-seq) has emerged as a cost-effective and efficient approach for studying genetic variation in populations. This method, which involves sequencing pooled DNA samples from multiple individuals within a population rather than sequencing each individual separately, has revolutionized the field of population genomics [13]. In Pool-seq, equal amounts of DNA from multiple individuals are combined into a single pool before library preparation and sequencing. The resulting sequencing data represents the allele frequencies in the population rather than individual genotypes. This approach allows researchers to estimate allele frequencies across the genome for a large number of individuals at a fraction of the cost of individual sequencing [96] The applications of Pool-seq are wide-ranging in genetics and evolutionary biology. In population genetics, it is used to study genetic diversity, population structure, and demographic history [97]. Evolutionary biologists employ Pool-seq in experimental evolution studies to track allele frequency changes over time [48]. It has also found utility in quantitative trait locus (QTL) mapping to identify genomic regions associated with specific traits [98], and in some cases, for genome-wide association studies (GWAS) to identify genetic variants associated with diseases or traits [99]. The advantages of Pool-seq are numerous and significant. Its cost-effectiveness allows

for the analysis of more individuals or populations within a given budget, greatly expanding the scope of many studies [13]. The reduced computational requirements, as individual genotypes are not generated, make data analysis less intensive [100]. Pool-seq also improves the detection of rare alleles in a population, as pooling can increase the likelihood of capturing these less common variants [101]. Furthermore, the pooling of multiple individuals can help mitigate the effects of sequencing errors on allele frequency estimates [96].

However, Pool-seq has its limitations. The loss of individual-level information is a significant drawback, as Pool-seq does not provide individual genotypes, limiting certain types of analyses [13]. Detecting structural variants such as large insertions, deletions, or rearrangements can be challenging with Pool-seq data [100]. There is also potential for allele frequency estimation bias if there are unequal DNA contributions from individuals in the pool [102]. Additionally, Pool-seq typically does not provide information about linkage between variants or haplotype structure. When compared to individual sequencing, Pool-seq offers a balance between cost, sample size, and genomic information, making it particularly useful for studies focusing on population-level allele frequencies and genetic diversity. While individual sequencing provides more detailed genetic information, including genotypes and haplotypes for each sample, it is significantly more expensive and computationally intensive for large-scale population studies [13]. The choice between Pool-seq and individual sequencing ultimately depends on the specific research questions, budget constraints, and required resolution of genetic information. For studies requiring individual-level data or focusing on rare variants, individual sequencing may still be preferable despite its higher cost [100]. Nonetheless, Pool-seq has become an invaluable tool in population genomics, offering a cost-effective approach to study genetic variation across large numbers of individuals. While it has limitations, its advantages make it particularly suited for many population-level studies, including those investigating rapid adaptation and evolution.

3 OCSVM-FET Analysis of *C. riparius* Adaptation Patterns

This chapter summaries polygenic adaptation through integrated quantitative and population genetics perspectives. It explores how populations maintain genetic diversity during rapid adaptation using classical population genetic measures (F_{ST}, π, θ, Tajima's D) while considering key factors like effective population size, selection strength, and genetic redundancy. The background introduces concepts of genetic redundancy in polygenic adaptation and how different analytical approaches can reveal both subtle and strong selection responses.

Genetic architecture of rapid adaptation

Understanding the genetic architecture underlying such rapid adaptation remains a central challenge in evolutionary biology, particularly for complex polygenic traits. The study of polygenic adaptation has emerged as a crucial field in evolutionary biology, offering critical insights into how organisms can swiftly respond to environmental challenges through the coordinated action of multiple genes. This complexity stems from the intricate genetic architecture of adaptive traits, where numerous loci contribute to phenotypic variation, making the detection and characterization of adaptive changes particularly challenging [14]. Understanding these architectural patterns is essential for explaining how populations can simultaneously maintain genetic diversity while adapting to novel conditions. Moreover, the significance of polygenic adaptation extends beyond theoretical interest, as many ecologically relevant traits exhibit polygenic inheritance. This understanding is particularly vital in the context of rapid environmental change, where it can inform conservation strategies and biodiversity management in shifting environments [103]. Such insights become increasingly critical as populations face unprecedented challenges from climate change and habitat alterations, where rapid adaptation may be essential for survival. The study of adaptive processes has historically been approached through two distinct perspectives. While quantitative genetics examines phenotypic changes, molecular population genetics analyzes genomic signatures at selected and linked loci. The

extensive findings from genome-wide association studies (GWAS) revealing the highly polygenic nature of traits, combined with the central role of the infinitesimal model in quantitative genetics, led to the development of the 'omnigenic' model. This model proposes that thousands of genes outside core pathways significantly contribute to heritability, providing a concrete mechanistic foundation for the abstract infinitesimal framework [104]. Although the omnigenic model suggests a broad genetic basis for trait, detecting such effects remains challenging with traditional selection tools.

In contrast, population genetics has traditionally characterized adaptation through substantial frequency changes and swift fixation of unconditionally beneficial alleles [105, 106]. This view suggests that phenotypic adaptation occurs through independent selective sweeps driven by single-locus mutations. Recent research efforts have aimed to bridge these divergent perspectives [2, 2, 104, 103], recognizing the need for a unified framework that accommodates both major single-gene changes and subtle coordinated shifts across multiple loci. To reconcile these perspectives, Barghi et al. [14] proposed an integrated framework emphasizing genetic redundancy as a key feature of polygenic adaptation, potentially explaining both parallel and heterogeneous adaptation patterns. Their empirical study have demonstrated that the combination of genetic redundancy and quantitative trait concepts offers a more comprehensive explanation of experimental results than explanations based purely on independent selective sweeps. This genetic redundancy, characterized by an abundance of beneficial variants [107, 108, 109], suggest that that populations evolving toward the same fitness optimum may exhibit non-parallel genomic changes.

To address these challenges in detecting subtle allele frequency changes across multiple loci, a novel approach combining One-Class Support Vector Machines with Fisher's Exact Test (OCSVM-FET) were applied for detecting polygenic adaptation patterns in the Pool-Sequencing data from 4 replicated *Chironomus riparius* populations across 7 time points. By applying different stringent FET-thresholds for selected alleles, the current study distinguishes between strong and subtle selection responses in allele frequency dynamics. These data sets were subsequently analysed regarding genetic and nucleotide diversity, differentiation, neutrality and eventually gene- and KEGG Pathways analysis was conducted.

3.1 Population Genetic Measures: Heterozygosity and Neutrality

The population genetic measures - F_{ST}, Tajima's π, Watterson's θ, and Tajima's D - provide complementary insights into the genetic diversity, population structure, and evolutionary history of species. For this study, these measures were used in combination, offering a deeper understanding of the complex dynamics of rapid adaptation in the conducted experiment.

F_{ST} and Population Structure The F_{ST} parameter, also called F-statistic or fixation-index, is a fundamental measure in population genetics and was introduced by Sewall Wright [110]. F_{ST} quantifies the degree of genetic differentiation among subpopulations, providing insights into population structure and gene flow [111]. Mathematically, F_{ST} is defined as the proportion of total genetic variance contained in a subpopulation relative to the total genetic variance [110]. While F_{ST} has become a cornerstone in population genetics studies, a variety of methods have been developed to measure F_{ST}, such as Nei [112], Weir and Cockerham [113], and Hudson [114]). The application of the methodological variations has been discussed in several papers and reviews [115, 116, 111]. Therefore the classic F_{ST} calculation e.g. is used for genetic data with a moderate number of loci and individuals, whereas Weir and Cockerham's F_{ST} is mostly used for genetic data with varying population sizes and sampling intensities. Hudson's F_{ST} is another popular method for estimating population differentiation. It is particularly useful for large data-sets with many loci and individuals. The latter was used in this study for exploring the polygenic adaptation pattern given as output by the OCSVM-FET approach. Hudon's F_{ST} calculation comes with the Popoolation2 [18] software utilized in this study. F_{ST} is generally calculated with the following formula:

$$F_{ST} = 1 - \frac{H_w}{H_b} \tag{3.1}$$

where H_w is the average number of differences between sequences sampled within subpopulations, and H_b is the average number of differences between sequences sampled between subpopulations

The interpretation of Hudson's F_{ST} is consistent with other F_{ST} measures: values range from 0 (no differentiation between populations) to 1 (complete

differentiation). However, it's important to note that the maximum value of F_{ST} can be less than 1 for loci with high heterozygosity [117]).

Nucleotide Diversity Measures Several measures have been developed to quantify nucleotide diversity within populations, each providing unique insights into evolutionary processes, whereas this work highlights especially Tajima's π and Watterson's θ. Tajima's π, introduced by Fumio Tajima [118], quantifies the average nucleotide differences per site between randomly selected DNA sequences. Tajima's π is sensitive to both low and intermediate frequency variants. Watterson's θ, developed by Watterson [119], is an estimate based on the number of segregating sites in a sample of DNA sequences. Watterson's θ is particularly sensitive to low-frequency variants and is often used in conjunction with Tajima's π to infer demographic history or selection.

While π directly measures nucleotide differences, θ etsimates variation based on segregating sites, both contributing complementary insights into genetic diversity. Tajima's D combines both measures to detect deviations from neutrality. In summary, Tajima's π and Watterson's θ provide complementary information about genetic diversity, while Tajima's D helps to interpret these measures in the context of population genetic

Tajima's D: A Test of Neutrality Tajima's D, introduced by Fumio Tajima [118], is a statistical test used to distinguish between a DNA sequence evolving neutrally and one evolving under a non-random process, including directional selection, balancing selection, or demographic changes. Tajima's D is calculated using the formula [120]:

$$D = \frac{\pi - \theta}{\sqrt{V_D}} \tag{3.2}$$

where:

- π is the average number of nucleotide differences between sequences,
- θ is Watterson's estimator of genetic diversity,
- V_D is the variance of $\pi - \theta$

Under neutral evolution, π and θ are expected to be equal, resulting in a Tajima's D value of zero. Deviations from zero can indicate various evolutionary scenarios:

Negative Tajima's D values suggest an excess of low frequency polymorphisms, which indicates directional selection or population expansion. Positive Tajima's D values suggest an excess of intermediate frequency polymorphisms, which suggests balancing selection or a population bottleneck [22, 121].

In practical applications, Tajima's D has been widely used to infer demographic history and selection pressures in various species. For example, in a study of the Eurasian Collared Dove, negative Tajima's D values suggested population expansion or positive selection [122]. Conversely, in a study of *Dendrolimus kikuchii*, non-significant Tajima's D values, combined with other evidence, supported the hypothesis of stable population sizes in southern China [123].

3.2 Factors Influencing Genetic Adaptation and Diversity

The process of genetic adaptation in populations is shaped by a complex interplay of various factors. Understanding these elements is crucial for interpreting patterns of genetic diversity and predicting evolutionary trajectories. This section explores briefly key factors that influence genetic adaptation and diversity in populations.

Effective Population Size (N_e) Effective population size (N_e) is a fundamental concept in population genetics, representing the size of an idealized population that would experience the same rate of genetic drift as the observed population [124]. N_e significantly influences the rate of evolution and the effectiveness of selection [125]. In small populations (low N_e), genetic drift plays a more prominent role, potentially overwhelming weak selection pressures. Conversely, in large populations (high N_e), selection can act more efficiently on even slightly beneficial alleles [126]. The effective population size is typically smaller than the census population size due to various factors including unequal sex ratios, non-random mating, fluctuating population sizes, and overlapping generations [125].

Heritability Heritability, the proportion of phenotypic variance attributable to genetic factors, plays a crucial role in determining the response to selection

[127]. Traits with high heritability are more likely to respond rapidly to selection pressures, facilitating adaptation. However, it is important to note that heritability is not a fixed property of a trait but can vary across environments and populations [128]. Understanding the heritability of key traits is essential for predicting the potential for adaptation in changing environments.

Selection Strength The strength of selection is a key determinant of adaptive outcomes. Strong selection can drive rapid changes in allele frequencies, potentially leading to selective sweeps [129]. However, very strong selection can also deplete genetic variation, potentially limiting long-term adaptive potential. Weak selection, while less dramatic in its immediate effects, can still lead to significant adaptation over time, especially in large populations where genetic drift is less dominant [130].

Genetic Drift Genetic drift, the random changes in allele frequencies due to sampling effects in finite populations, is a crucial factor in shaping genetic diversity [131]. In small populations, drift can lead to the fixation or loss of alleles independently of their selective value, potentially counteracting the effects of weak selection. The interplay between drift and selection is particularly important in understanding the fate of slightly deleterious mutations and the evolution of genome complexity [132].

Convergent Evolution and Phenotypic Plasticity Convergent evolution, where similar phenotypes evolve independently in different lineages, highlights the role of environmental pressures in shaping adaptation [133]. This phenomenon can complicate the interpretation of genetic data, as similar phenotypes may arise from different genetic mechanisms. Additionally, phenotypic plasticity, the ability of a single genotype to produce multiple phenotypes in response to environmental conditions, can influence the process of genetic adaptation [134]. High plasticity may buffer populations against selective pressures, potentially slowing genetic adaptation but also providing time for beneficial mutations to arise.

Part III

Material and Methods

4 Evaluating Methods for Polygenic Adaptation Detection

Python, R and Bash scripts as well as pipelines can be found on GitHub **https://github.com/CoCaliendo/Polygenic-Adaption-Pattern**

4.1 Simulation

To assess the performance of the different approaches, a series of simulations were conducted using the tool MimicrEE2 [40]. 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. The simulation pipeline began with the generation of a customized haplotype input file using individual sequencing data from 66 *C. riparius* specimens [46]. I utilized the non-model organism pipeline for GATK4 [135, 136] to create haplotypes from individual sequencing fastq files. The pipeline for conducting the desired haplotype-file from individual sequencing data can be found on GitHub[1]. Using this haplotype file, we conducted a total of 20 distinct simulation setups, each replicated 10 times for robustness. These 20 setups 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 allows 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.

[1] Scripts and pipeline details available on `Github : CoCaliendo/Polygenic-Adaption-Pattern`

This variation helps understanding how the number of genes involved in adaptation affects detection methods. (Thus ranging from oligogenic over mildly to highly polygenic)

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 quantities × 10 replicates) (see Fig. 1) Using 'qt' mode, I modeled a scenario of strong selection by specifying a range of effect sizes for the loci under selection. I used a custom Python script to generate a selected-loci file, which specifies the loci under selection and in-cooperates 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 frequencies provided by the effect-file. In this work, allele frequencies starts ranging from 0.01 to 0.3 were choosen. Another custom script was used to generate a recombination map. I allowed the recombination rate to vary between 0.1 and 4 centimorgans, encompassing a range of possible recombination rates that may reflect the natural variation observed in *C. riparius* populations. Another key parameter was heritability, which was set to 0.8, indicating a strong genetic component in the trait under selection. For the selection regime, a truncating selection model was applied with a factor of 0.5, i.e. removing individuals with phenotypic values below the 50th percentile. To simulate the desired time under selection, generation steps were set at 10, 20, 40 and 60. The output was generated in sync-file format to facilitate comparison with real-life experimental data. The R package poolSeq was utilized [137] for subsequent analysis of allele frequencies across generations (R version 4.3.2). This comprehensive simulation design allowed me 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, I aim to capture a range of evolutionary scenarios, providing a robust test for the analytical methods in identifying polygenic adaptation patterns

For the quantitative trait mode (setting 'qt'), specific parameter settings were employed to accurately model the evolutionary dynamics:

```
#!/usr/bin/env java
java -jar mim2-v206.jar qt/
--haplotypes-g0 haplotype-file.mimhap/
--recombination-rate recomb\_rate.txt/
--effect-size selected\_loci.txt/
--heritability 0.8/
```

```
--snapshots 10,20,40,60 /
--replicate-runs 10 /
--output-sync output.sync.gz /
--selection-regime truncating\_05.txt /
--threads 6
```

Listing 4.1: Parameter Setting for MimicrEE2 embedded in a java script.

The parameter *–effect-size* identifies the loci under selection, driving the adaptation process. To generate the file, a python script provided by MimicrEE2 was used and customized accordingly:

```
#!/usr/bin/env python
python pick-SNPs-QTL.py/
--mimhap haplotype-file.mimhap/
--n 100/
--f 0.15/
--effect-file effect\_file.txt > selected\_loci\_output.txt
```

Listing 4.2: python script to draw 100 position randomly that will undergo selection.

The script generates a file with three columns, that gives information about the positions, additive effect and heterozygous additive effect. The starting allele frequency for the selected loci was set to 0.15 (–f 0.15) In natural populations, the initial frequency of advantageous alleles can vary. Setting the starting allele frequency to 0.15 may reflect a scenario where the advantageous alleles are moderately common but not overwhelmingly dominant in the population.

The effect size, specified trough *–effect-file* 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: First, for each locus under selection, an effect size is randomly chosen from the specified range (in this 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 some 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. A high effect size in range of 1.5 to 3.5 was choosen, allowing to observe significant changes within a relatively small number of generations, which is computationally efficient and helps in detecting selection signals more clearly. The parameter *–selection-regime* determines whether phenotypic values are truncated or continuous. In this case it was set to a truncation factor of 0.5,

indicating that individuals with phenotypic values below the 50th percentile are removed from the population.

The *–recombination-rate* determines the recombination rate, influencing genetic diversity. MimicrEE2 provides a recombination map, generated by a customized Python script to provide centimorgan distances between genomic positions. By allowing the centimorgan distance to vary between 0.1 and 4, the simulation encompasses a range of possible recombination rates that may reflect the natural variation observed in *C. riparius* populations.

```
#!/usr/bin/env python
python simple-recombination-map.py/
--mimhap haplotype-file.mimhap --rr 4 > pre-recomb-rate.txt
```

Listing 4.3: python script to annotate recombination rate in centimorgan distance.

```
#!/usr/bin/env bash
awk 'BEGIN { srand() } \{ print \$1, rand() * (4 - 0.1) + 0.1 \}'/
pre-recomb-rate.txt > recombination-rate.txt
```

Listing 4.4: bash command to select random recombination rate between 0.1 - 4 (in centimorgan distance).

The parameter *–snapshots* indicates the number of generations, for which simulation runs are conducted. This parameter was set to 10,20,40,60 to generate the respective output for these generations, similar to the real-life experiment conducted as described in 5.1. The parameter *–replicate-runs* defines the number of replicates per generation. This parameter was set to 10 as proposed from guidelines [40]. Output format, specified as *–output-sync* was set to sync-file. For further analysis, the R package *poolSeq* [137] was used. To speed up the process, parameter *–threads*, which determines the number of kernels to use from operating system, was set to 6.

For an atomized script, looping over different setting of select loci, see Github[2]

4.2 Fisher's Exact Test

The allele frequencies across all simulated generations were acquired utilizing the R packages *poolSeq* with R version 4.3.2 (2023-10-31). Fisher's exact tests were conducted to compare the initial and final generations, yielding corresponding p-values [137], which were corrected for False Discovery Rate

[2] Scripts and pipeline details available on Github:CoCaliendo/Polygenic-Adaption-Pattern

using Benjamini-Hochberg correction. After p-values were annotated to each position, a significance level cut-off of <0.001 was chosen, taken from the paper of the already published data [68], which was used for parameter fine tuning (see 4.6).

4.3 Naive Bayes Classifier

The mathematical function described in formula (3.6, 1.3) were put into the following code function:

```
#!/usr/bin/env python
python nbc_function.py/

def multivariate_distri(pos, mu, Sigma):
    """Return the multivariate Gaussian distribution on array pos."""

    n = mu.shape[0]
    Sigma_det = np.linalg.det(Sigma)
    Sigma_inv = np.linalg.inv(Sigma)
    N = np.sqrt((2*np.pi)**n * Sigma_det)
    # This einsum call calculates (x-mu)T.Sigma-1.(x-mu) in a vectorized
    # way across all the input variables.
    fac = np.einsum('...k,kl,...l->...', pos-mu, Sigma_inv, pos-mu)

    return np.exp(-fac / 2) / N
```

Listing 4.5: python script function for NBC.

The parameters for the covariance matrix and the mean were tuned onto already published data as described in 1.6. Therefore, first the PDF of the normal, non-anomalous data was modelled, using Python 3.8.13 and resulting in the following 2D arrays:

```
#!/usr/bin/env python
python nbc_parameter_tuning.py/
import numpy as np
import pandas as pd

# Define the parameters for the non-anomalous distribution
Sigma1 = np.array([[0.25, 0.18], [0.18, 0.25]])
mu1 = np.array([0., 0.])

# Define the parameters for the anomalous distribution 1
Sigma2 = np.array([[0.2, 0.19], [0.19, 0.2]])
```

```
mu2 = np.array([0.1, 0.8])

# Define the parameters for the anomalous distribution 2
Sigma3 = np.array([[0.2, 0.19], [0.19, 0.2]])
mu3 = np.array([0.8, 0.1])

# Create multivariate  distributions
mvn1 = multivariate_distri(pos, mean=mu1, cov=Sigma1)
mvn2 = multivariate_distri(pos, mean=mu2, cov=Sigma2)
mvn3 = multivariate_distri(pos, mean=mu3, cov=Sigma3)
```

Listing 4.6: python script to model NBC with published data set.

These settings were applied respectively to each of the simulated data sets for the initial allele frequency ('af0') as well the allele frequency of generation 10('af10'), 20('af20'), 40('af40') and 60('af60').

```
#!/usr/bin/env python
python nbc_parameter_tuning.py/
import numpy as np
import pandas as pd
from scipy.stats import multivariate_normal

# Calculate probability density values for each data point
# under specific distribution
pdf_values1 = mvn1.pdf(df[['af0', 'af10']])
pdf_values2 = mvn2.pdf(df[['af0', 'af10']])
pdf_values3 = mvn3.pdf(df[['af0', 'af10']])

# Assign each data point to the distribution with the
# highest probability density
df['cluster'] = np.argmax(np.column_stack(
                (pdf_values1, pdf_values2, pdf_values3)), axis=1) + 1
```

Listing 4.7: python script to model NBC with simulated data set.

4.4 One Classifier Support Vector Machines

The One Class Support Vector Machines (OCSVM) algorithm [33] was employed in this study to identify potential signatures of polygenic adaptation in allele frequency data, using scikit-learn [138]. This machine learning approach is particularly suited for detecting subtle, coordinated changes across multiple loci that are characteristic of polygenic adaptation. OCSVM algorithm was used with the Radial Basis Function ('RBF') kernel, aiming to divide the data

into anomalous and non-anomalous categories in a higher dimensional space. This kernel function measures the similarity between two data points and helps define the boundaries between them. In this study, radial basis function (RBF) kernel was choosen 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 "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. The idea is, that this function maps 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 For mathemathical background on the kernel function see 1.4 Specifically, γ and ν were crucial parameters to be choosen. 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 the 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. These parameters were fine-tuned (ν = 0.01, γ = 0.05) using previously published data on *C. riparius* [46], ensuring the model was optimized for detecting genetic anomalies in this species (see 4.6.

Allele frequencies were acquired using the R packages poolSeq [137] with R (version 4.3.2) and were used as input for the algorithm. The algorithm generates a list where each anomalous position is labelled as -1, and non-anomalous data is labelled as 1. Loci flagged as anomalous (-1) by the algorithm are candidates for being part of the polygenic response to selection in the experimental populations.

```
#!/usr/bin/env python
python ocsvm_poly.py/
from sklearn.svm import OneClassSVM
from sklearn.preprocessing import StandardScaler
```

```
5
6       #start OCSVM with dataframe df
7       #normalize allele frequencies
8       sX=pp.StandardScaler(copy=True)
9       scale = sX.fit_transform(df[['af0', 'af10']])
10      df_scale = pd.DataFrame(scale, columns = ['af0', 'af10'])
11
12      #parameters for svm
13      svm = OneClassSVM(kernel='poly', nu=0.01,  gamma=0.05,
14      degree = 3, coef0 = 0.0, cache_size=1000)
15
16      # fit the model with the dataset and get the prediction data
17      by using the fit() and predict() method
18      svm.fit(df_scale)
19      pred = svm.predict(df_scale)
20
21      #get scores
22      scores = svm.score_samples(df_scale)
```

Listing 4.8: python script OCSVM polynomial kernel.

For OCSVM with a RBF kernel [138], γ were set to 0.05 and ν was set to 0.01. Allele frequencies of the initial (af0) and resepctively last (af10, af20, af40, af60) simulated generations were extracted using Popoolation2 [40]

4.5 Combination with Fisher's Exact Test

In this study, a combined approach was developed that integrates the binary classification results from OCSVM and NBC with the statistical hypothesis testing of FET. This method aimed to leverage the strengths of both machine learning and traditional statistical approaches. For each simulated dataset, I first applied OCSVM and NBC independently to classify positions as anomalous or non-anomalous. Concurrently, FET was preformed on the same datasets. The p-values from FET were corrected for multiple testing using the Benjamini-Hochberg method. A customized Python script (version 3.8.13)[3] was used to merge these results for each position in the genome and subsequently filter the data, retaining only those positions classified as anomalous by OCSVM or NBC that also had a FET-derived adjusted p-value < 0.001. This process resulted in a set of "significant anomalies" - positions that were flagged as

[3] Scripts and pipeline details available on Github:CoCaliendo/Polygenic-Adaption-Pattern

anomalies by the machine learning approaches and also showed statistically significant changes according to FET.

4.6 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 [46]. 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 [46]. 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 and area under curve 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. For the 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 grid search aimed to maximize performance across accuracy, false positive rate (FPR), and area under the curve (AUC). These metrics were chosen to balance the algorithm's ability to correctly identify true positives while minimizing false detection, 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 allows 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 expec-

tation that polygenic adaptation often involves multiple, unlinked genomic regions. This approach allows to capture the diversity of allele frequency changes associated with complex adaptive processes.

4.7 Performance Metrics

Performance of the different approaches (FET, OCSVM, NBC, OCSVM-FET, NBC-FET) was evaluated using metrics including false positive rate (FPR), area under curve (AUC), and accuracy with a customized python script (version 3.8.13, module: 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 ability of discrimination 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 a customized python script (version 3.8.13, module: matplotlib).

False Positive Rate (FPR): measures the proportion of data points that were falsely identified as positive by the algorithm. It is calculated as the ratio of false positives to the sum of true negatives and false positives

$$FPR = \frac{FalsePositives}{FalsePositives + TrueNegatives} \tag{4.1}$$

Area Under Curve (AUC): 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 ability of discrimination of the algorithm

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

$$Accuracy = \frac{NumberOfCorrectPredictions}{TotalNumberOfPredictions} \tag{4.2}$$

5 Experimental Evolution: Selection on Emergence Time in *C. riparius*

In this chapter, the experimental set-up of the Evolve and Re-sequence experiment is described. To ensure the success of the experiment, the collected data was statistically analysed. Additionally this chapter gives an overview to the bioinformatic preprocessing of Pool-Sequencing data.

5.1 Experimental Design

Aiming to set a population of *C. riparius* under selection pressure regarding faster developmental time (i.e. emergence), 24 egg clutches were randomly selected from the population described in 5.2 to establish the starting population for the experiment. A mixed pool was made and a total of 4.000 larvae were randomly drawn and grouped into sets of 50, which were distributed across 80 trays. (refer to fig. 5.1) (see 5.3 for tray setup).

Test trays were divided into four groups, each consisting of 20 trays accordingly, constituting four replicate groups - namely replicate blue, replicate red, replicate green, and replicate gold. Replicate trays were set into a climate chamber under constant temperature of 20 degree Celsius, 60% relative humidity, constant aeration and a light-dark cycle of 18:6. Trays were randomly shifted every 3-5 days to avoid effects of different light exposure. Feeding routines were done each morning adhering to a predetermined schedule outlined in the supplementary material (Appendix, A) using finely ground fish food (e.g. Tetramin Flakes).

To introduce selection pressure on emergence time, only the first 50 emerged females and first 50 emerged males were collected with a vacuum trap. This was done for each replicate separately. The collected early emerging individuals were put into a replication cage, where they were able to produce

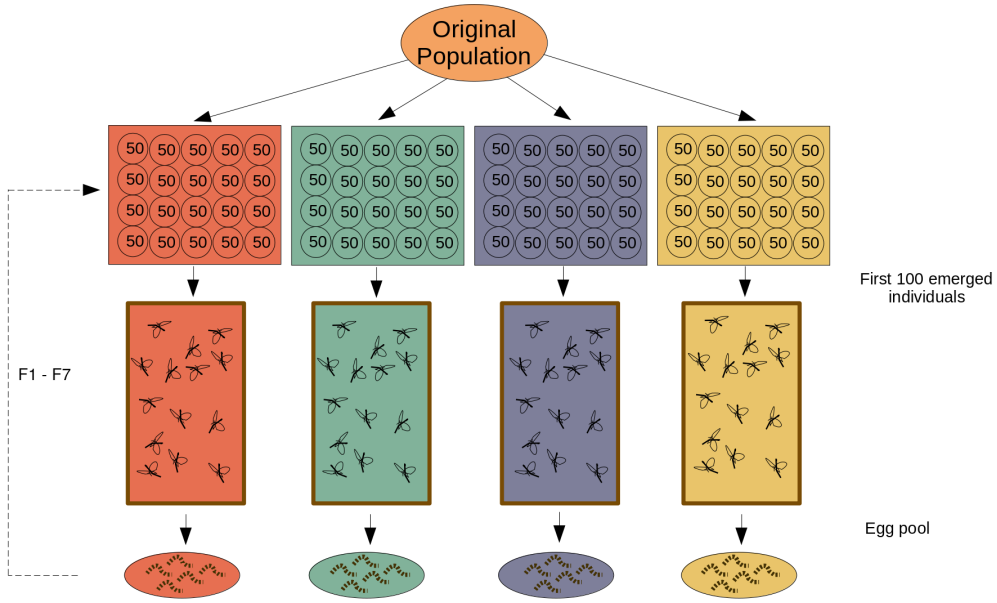

Fig. 5.1.: Scheme of experimental setup: Eggs from the original group were pooled and divided into four replicates: red, green, blue, and gold. The first 100 midges emerging, including 50 females and 50 males, were put into replication cages for mating. Eggs for the next generation were collected for four days. This process of choosing fast developing, i.e. fast emerging, midgets was repeated for seven generations.

offspring for the next generation. Each replication cage featured a glass bowl containing medium, which provided a suitable environment for females to lay their egg clutches. Additionally, cages contained a cup holding a solution of sugar, along with a wooden stick placed inside to absorb the liquid safely. During the initial three days of egg laying, egg clutches from all four replicates were collected and stored in medium. These eggs were refrigerated at 8 degree Celsius for a period ranging from one to three days. On the fourth day, the eggs were transferred to a climatic chamber set at 20 degree Celsius, where they typically hatched within a span of three days. Hatched larvae were pooled together and sorted as described in section 5.4, resulting in a total of 4,000 larvae as the start of the next generation, see Fig. 5.1. The first generation therefore forms the ancestral generation. The procedure was then carried out for a further 6 generations. Each emerged individual was meticulously collected, labeled, and preserved in ethanol at a temperature of -4 degrees Celsius.

5.2 Test organism

The population of *C. riparius* used in this study was sourced from a native population collected over a span of three years (2016-2019) from a small river located at Hasselbach, Hessen, Germany (50.167562°N, 9.083542°E). This population has been maintained for several generations as an in-house laboratory culture at an average temperature of 20-23 degree Celsius [139]. Maintaining genetic diversity in the laboratory population required periodic supplementation with wild-caught individuals to counteract random genetic drift. The culture conditions for the stock population adhere to a modified approach based on the procedure outlined in OECD guideline N°219, which was previously published by Foucault et al. [88]. The cultures were reared under a long day light regime (16:8 L:D) and fed daily with 0.4 g finely grounded fish food (e.g. Tetramin Flakes).

5.3 Tray Setup

Glass bowls were used as trays (Ø 20 × 10 cm), filled with a 1.5 cm sediment layer of washed, pH neutral, commercial sand and 1.250 L of medium, consisting of deionized water adjusted to a conductivity of 550 μS/cm with aquarium sea salt (e.g. TropicMarin) and a basic pH around 8 [88]. Preventing emerged individuals from escaping, a hair net was stretched over each tray. To ensure proper ventilation of the medium, an aeration system was devised using hoses. These hoses extended into the medium through the net attached to the trays. Compensating for water evaporation in the test vessels was achieved by introducing demineralized water.

5.4 Larvae Drawing

Each egg clutch was collected from the replication cages and stored in a single 3 mL well within a 6-well plate of medium. After the egg clutches had undergone a period of at least one day and maximum of three days in a refrigerator at 8 degree Celsius, they were incubated on the fourth day at a temperature of 20 degrees Celsius for three days to facilitate synchronized hatching. Subsequently, enclosed larvae from different clutches were pooled and 12 egg

clutches were selected for each case. These chosen clutches exhibited uniform larval size and had a hatching rate of around 90%. This ensured that all larvae reached the same developmental stage within narrow time frame. Pooled and randomly drawn larvae were distributed across the 80 trays, segregated by their respective replicate group. As the present experiment maintained a consistent count of females as the foundation for the subsequent generation, with only the initially deposited egg clutches being taken into account, fertility was determined based on the overall number of collected egg clutches.

5.5 Statistics

During the experiment, daily records were maintained for the count and sex of emerging individuals per tray to calculate survival, sex ratio, mean developmental time and median developmental time (EmT50). For each generation and replicate produced eggropes were collected, number of eggs were counted under the microscope. Fertility was accessed using the method outlined by Foucault et al. [88], which involved quantifying the count of egg ropes where at least 50% of the larvae successfully hatched, divided by the number of females. Subsequently, survival rates, EmT50 (median time for emergence) of females, and fertility values were combined using a simplified Euler–Lotka computation to evaluate the population growth rate (PGR). The latter measurement integrates each parameter while considering their respective contributions to the population's dynamics, following the approach by Nemec et al. [95]. Therefore PGR incorporates specifically the following parameters: mortality rate (f), EmT50 (the time taken for 50% emergence) of females (g), female fraction (h), average number of eggs per egg mass (i), and the number of fertile egg masses produced per female (j) and plug in to the formular [94]

$$PGR = \left[(j \times i \times h) \times \left(1 - \frac{f}{100} \right) \right]^{1/g} \tag{5.1}$$

To test for the rapid adaptation, mean emergence, EmT50, survival, fertility and PGR were statistically tested. Each generation was tested against the initial (1st) generation. Two approaches were utilized: Frequentist's approach

by using Mann-Whitney-U-Test Test, and Bayesian Inference Statistic. Mann-Whitney-U-Test is testing the hypothesis if the middle ranks of both groups are the same. For this Python 3.8.13 with the 'mannwhitneyu' function provided by scipy.stats was used. To investigate effective sizes, Cohen's D was used by calculating the mean difference between the two groups, and then dividing the result by the pooled standard deviation:

$$\text{Cohen's } D = \frac{(M_2 - M_1)}{SD_{\text{pooled}}} \tag{5.2}$$

For Bayesian Inference Statistics, PyMC3 was utilized, which is a tool for Bayesian statistical analysis and probabilistic machine learning, written in the Python programming language. It employs Markov-Chain-Monte-Carlo (MCMC) and variational inference algorithms to explore the posterior distribution of model parameters. As for the prior knowledge, expected performance range was used to set a uniform prior distribution, which was 0-26 (days) for the mean emergence and EmT50 and and 0-1.2 (units) for PGR.

```
#!/usr/bin/env python
python bayesian.py/
import pymc as pm

# loop thorugh each generation per replicate
for i in range(1, 8)
# Mean
# Define the PyMC3 model
with pm.Model() as model:
    # Define prior distribution for mean emergence time for each generation
    mu = [pm.Uniform(f'mu_{i}', lower=0, upper=26) for i in range(7)]

    # Define likelihood function for each generation
    obs = [pm.Normal(f'obs_{i}', mu=mu[i], sigma=1,
           observed=mean_emergence_times[i]) for i in range(7)]

    # Perform MCMC sampling
    trace = pm.sample(draws=2000, tune=1000)

#EmT50
with pm.Model() as model:
    # Define prior distribution for EmT50 for each generation
    mu = [pm.Uniform(f'mu_{i}', lower=0, upper=26) for i in range(7)]

    # Define likelihood function for each generation
    obs = [pm.Normal(f'obs_{i}', mu=mu[i], sigma=1,
           observed=emt50_values[i]) for i in range(7)]
```

```
28
29      # Perform MCMC sampling
30      trace = pm.sample(draws=2000, tune=1000)
31
32  # PGR
33  # Define the PyMC3 model
34  with pm.Model() as model:
35      # Define prior distribution for mean PGR for each generation
36      mu_pgr = [pm.Uniform(f'mu_pgr_{gen_idx}', lower=0, upper=1.2)
37         for gen_idx, _ in enumerate(generation.unique())]
38      # Prior for standard deviation of PGR
39      sigma_pgr = pm.HalfCauchy('sigma_pgr', beta=10)
40
41      # Define likelihood function for PGR
42      obs = [pm.Normal(f'obs_pgr_{i}', mu=mu_pgr[i], sigma=sigma_pgr,
43             observed=pgr_values[i]) for i in range(7)]
44
45      # Perform MCMC sampling
46      trace= pm.sample(draws=2000, tune=1000)
```

Listing 5.1: python script for bayesian statistics analysis.

5.6 Data Pre-processing

For each generation, each emerged midge was meticulously collected, labeled, and preserved in ethanol at a temperature of -4 degrees celsius. To prepare the DNA for subsequent Pool-Sequencing, all female individuals per generation were pooled and 100 individuals were drawn. For the DNA extraction, 2 legs were taken from each individual and processed following the *DNeasy Blood & Tissue Kit* instructions by Qiagen (see manual in appendix). DNA concentration for each samples was measured with a Qubit fluorimeter (Invitrogen). Whole-genome pool-sequencing was conducted on an Illumina MiSeq platform, generating 250bp paired-end reads. Subsequently, reads underwent a two-step process involving trimming and quality control using fastQC [140]. The trimming process was facilitated by Trimmomatic [141].

Trimmed reads were aligned to the *C. riparius* reference genome (version 3) [142] utilizing the BWA mem algorithm [143]. A subsequent filtration step was implemented to remove duplicated and unmapped reads. Additionally, only reads with 1) mapping quality above 20 were kept and 2) with read coverage between 15 and 70. For each replicate and each generation a merged

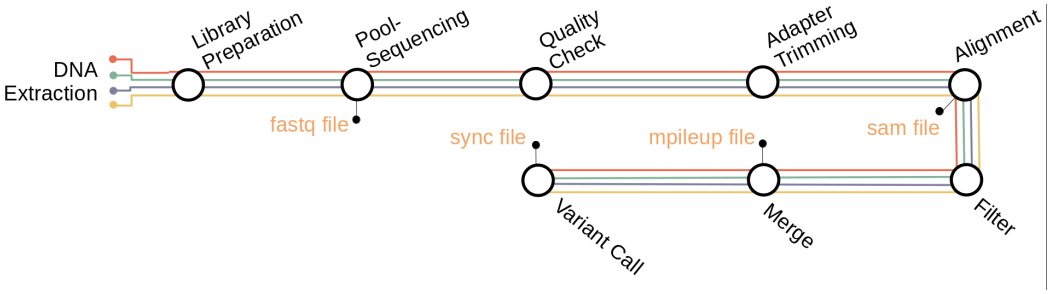

Fig. 5.2.: Flow chart of bioinformatic pipeline: DNA Extraction of 7 generations of the respective replicates underwent short-read library preparation and Pool-Sequencing (Illumina). Quality check of the fastq files were done by FastQC and for adapter trimming Trimmomatic was used. Sam files were generated by BWA-MEM alignment to the reference genome, subsequent converting to bam files, filtering and merging all files into one mpileup file was done by using samtools. For obtaining a synchronized file, data was processed with Popoolation2.

file was generated using samtool's function mpileup [144]. The merged file contains for each position information about the respective reference base, the bases from reads aligned to that position, quality scores associated with the reads and mapping qualities indicating how well each read aligns to the reference. By using Popoolation2's implementation mpilup2snyc.jar [145, 18] a synchronized file was created. Its primary function is to prepare data for downstream analysis, specifically for comparing allele frequencies between populations, by employing filters based on quality and synchronized data for populations. For the filtering, it removes low-quality reads or bases from the analysis to ensure reliable data. To enable comparison of allele frequencies, mpileup2sync.jar combines and summarizes the mpileup information for different populations into a single, synchronized format. To assert allele frequencies for each repliacte and generation, the synchronzied file was processed with R (version 4.2.1) using the library PoolSeq [137]. The allele frequencies form the start point of any further bioinformatic analysis. All scripts used for this pre-processing pipline can be found in appendix.

6 OCSVM-FET Analysis of *C. riparius* Adaptation Patterns

Python, R and Bash scripts as well as pipelines can be found on GitHub $https://github.com/CoCaliendo/Analysis_AdaptationPattern$

6.1 Post-Processing

Starting with obtained allele frequencies from the pre-procession steps (see section 5.6), a combined approach consisting of Fisher's Exact Test (FET) and One-Class Support Vector Machines with a radial kernel (OCSVM radial) was applied to identify candidate SNP loci potentially under selection for early emergence. OCSVM radial was utilized with a customized python script (version 3.8.13 , module: scikit-learn) with the settings γ = 0.05 and ν = 0.013. Allele frequencies served as input, and OCSVM radial classified loci as either anomalous (-1) or non-anomalous (1). Anomalous positions were overlapped with positions denoted as being significant according to FET. FET were conducted to compare the ancestral (first) and last (seventh) generations, yielding corresponding p-values using poolSeq with R (version 4.3.2) [137], which were corrected for False Discovery Rate using Benjamini-Hochberg correction. To identify both, broad-effect variants and strong-effect variants, two different cut-off were applied and analyzed:

1. **Broad-effect variants** To specify loci where the proportions of allele frequencies changed more than expected due to chance alone, neutral simulations were used to compute false discovery rate q-values < 0.001.

2. **Strong-effect variants** This group involved setting a very stringent cut-off by selecting the lower 0.0001% of the corrected p-values. This approach identified a smaller, more highly significant set of candidate SNP loci (red: 9.66, blue: 7.76, gold: 6.67, green: 8.51).

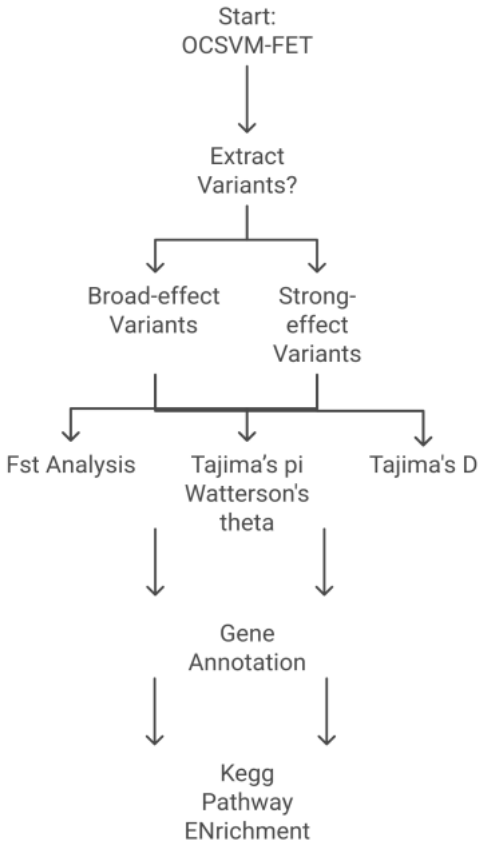

Fig. 6.1.: Flow chart of bioinformatic pipeline: After Pre-processing steps, obtained allele frequencies were used as input for the OCSVM-FET approach. Different thresholds for FET were applied to filter for broad-effect variants and strong-effect variants. The resulting significant anomalies were used for further examination: F-Statistics (F_{ST}), Tajima's π, Watterson's θ and Tajima's D were calculated by using Popoolation2. Genes were annotated via InterProScan and Pfam database and KEGG Pathway enrichment was performed by WebGeStalt.

6.2 Genetic diversity analysis of candidate variants

To understand the long-term selection regime at the identified loci and their linked sites, we used four key population genetic parameters: Tajima's D, Tajima's π, Watterson's θ, and F_{ST}. For all genetic parameters, a non-overlapping sliding windowsize of 1kb was chosen, reflecting the species' short average linkage disequilibrium of less than 150 bp [12]

Selection regime, Tajima's D Tajima's D was calculated to summarize the site frequency spectrum and infer the selection regime over the long term [146]. We used Popoolation1 [18] to compute Tajima's D for all non-overlapping

1 kb windows across the genome. The values of Tajima's D in all windows were compared to the windows containing broad-effect variants/strong-effect variants. This analysis helps distinguish between neutral evolution, purifying selection, and balancing selection. Parameters: –fastq-type sanger –measure d –min-count 3 –min-coverage 15 –max-coverage 50 –min-covered-fraction 0.5 –pool-size 100 –window-size 1000 –step-size 1000

Nucleotide diversity, Tajima's π We calculated Tajima's π for all non-overlapping 1kb windows using Popoolation1 [145]. This measure provides insights into the level of nucleotide diversity within populations and can indicate recent selective events [22]. Computed π in all windows were compared to the windows containing broad-effect variants/strong-effect variants. Parameters: –fastq-type sanger –measure pi –min-count 3 –min-coverage 15 –max-coverage 50 –min-covered-fraction 0.5 –pool-size 100 –window-size 1000 –step-size 1000

Population mutation rate, Watterson's θ Watterson's θ was calculated to estimate the population mutation rate based on the number of segregating sites within each 1kb window, using Popoolation1 [145]. This parameter offers a different perspective on genetic variation, particularly sensitive to rare alleles [119]. Calculated θ in all windows were compared to the windows containing broad-effect variants/strong-effect variants. Parameters: –fastq-type sanger –measure theta –min-count 3 –min-coverage 15 –max-coverage 50 –min-covered-fraction 0.5 –pool-size 100 –window-size 1000 –step-size 1000

Genetic Differentiation, Fixation Index F_{ST} To measure genetic differentiation and loss of heterozygosity, we calculated the Fixation Index (F_{ST}) using Popoolation2 [18]. Pairwise F_{ST} calculation provides information on population structure and can indicate local adaptation [111]. We performed two analyses: a) Within-replicate F_{ST} to assess differences between replicate populations. b) Pairwise F_{ST} between consecutive generations to track changes over time.

6.3 Statistical Analysis

Statistical analyses were conducted using Python version 3.8.13 (scipy.stats module). We employed the Mann-Whitney U test to assess differences in Tajima's π and Watterson's θ between ancestral and last generations, using Cohen's D to measure effect size. For pairwise F_{ST} analysis between replicate populations, we utilized the Kruskal-Wallis test with eta-squared as the effect

size measure. Effect sizes were interpreted following standard thresholds: Cohen's D values of 0.2, 0.5, and 0.8 indicated small, medium, and large effects, respectively, while eta-squared values were considered small at 0.01, medium at 0.06, and large at or above 0.14 [147]

6.4 Wright-Fisher Simulation Analysis

Neutral evolution scenarios were simulated using the Wright-Fisher model implemented in the poolSeq R package [137]. Effective population size (N_e) was estimated using poolSeq's estimateNe function, yielding values of 17,000 for red, blue, and gold replicates, and 22,000 for the green replicate, representing conservative estimates [87]. Comparisons between experimental and simulated data employed chi-square tests (poolSeq's chi.sq.test function), with p-values adjusted using the Benjamini-Hochberg correction and subsequently negative log-transformed.

6.5 Gene Annotation

Gene annotation was performed on the identified broad-effect and strong-effect variant positions from both generation 4 and 7. For the initial annotation, tbg-tool [148] was used to identify variants located within exonic regions of genes. This direct positional analysis focused specifically on coding regions, without considering intronic regions or potential distant regulatory elements. Gene families were identified using interproscan with the pfam database [149, 150]. Overlap between replicate candidate loci was visualized using Venn diagrams (R version 4.3.2, package: venn [151]). Pathway enrichment analysis was conducted using WebGestalt 2024 (Web-based Gene Set Analysis Toolkit) [152] for KEGG pathway analysis. Due to its comprehensive annotation, Drosophila melanogaster was used as the model organism, with *C. riparius* as the reference genome. A p-value cutoff of 0.001 was applied, with enrichment ratio indicating pathway over-representation relative to the *C. riparius* reference genome background.

Part IV

Results

7 Evaluating Methods for Polygenic Adaptation Detection

Our analysis revealed that polygenic adaptation can be effectively detected using a combined machine learning approach. The method performs best when examining intermediate stages of adaptation, where selection has had sufficient time to act but before genetic changes plateau. This timing aligns with theoretical predictions about polygenic selection. Multiple test scenarios demonstrated that the method is particularly effective at detecting selection when spread across few hundreds of loci, reflecting realistic patterns of complex trait adaptation. The observed patterns of allele frequency changes matched empirical data from natural populations, validating the biological relevance of the approach.

7.1 Parameter Tuning

The effectiveness of both OCSVM and NBC approaches in identifying patterns of polygenic adaptation heavily depends on their parameter settings. To determine the optimal configuration, extensive parameter optimization was conducted using established allele frequency data from *C. riparius* populations.

The OCSVM optimization focused on two critical parameters: ν and γ. The ν parameter determines the balance between support vector quantity and training error tolerance, which affects how the algorithm responds to potential outliers. The γ parameter establishes the reach of a training example's influence in the feature space, where smaller values indicate broader influence and larger values denote more localized effects. Systematic evaluation covered ν values spanning from 0.001 to 1 (specifically: 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, as shown in Fig. 7.2.

The performance assessment, illustrated in Figure 7.1, evaluated three key metrics: False Positive Rate (FPR), Area Under the Curve (AUC), and Accuracy across various parameter combinations. The most effective configuration emerged with ν=0.01 and γ=0.05, achieving an impressive combination of low FPR (0.013), high AUC (0.993), and remarkable accuracy (0.987).

The NBC parameter optimization presented additional complexity due to its requirement for three distinct model approximations: one representing the primary, non-anomalous data distribution (characterized by mu1 and sigma1) and two describing anomalous data patterns (defined by mu2/mu3 and sigma2/sigma3). A key technical challenge in this optimization process was maintaining valid positive semi-definite (PSD) matrices for covariance calculations, a fundamental mathematical requirement. The range of tested mu and sigma combinations is depicted in Figure 5b.

Through systematic evaluation, NBC demonstrated optimal performance with Setting 1, utilizing the following parameters: mu1=[0.0, 0.0], sigma1=[0.15, 0.1], mu2/mu3=[0.1, 0.9], sigma2/sigma3=[0.25, 0.2]. This configuration yielded exceptional results with a minimal FPR of 0.004, superior AUC of 0.998, and high accuracy of 0.996.

When evaluating the highest-performing configurations of both approaches (Fig. 7.3), the analysis revealed that while both methods achieved remarkable accuracy and AUC scores, the NBC approach demonstrated marginally better FPR performance. However, the OCSVM exhibited greater stability across different parameter configurations, suggesting enhanced robustness to parameter adjustments.

The optimized parameters identified through this rigorous testing were subsequently implemented in the primary analysis of simulated data, ensuring maximum sensitivity in detecting subtle genomic signatures associated with rapid adaptation.

7.2 Application on simulated data

To validate the methodological approaches against known ground truth data, simulated datasets were generated using Mimicree2 (version mim2-v206), based on actual *C. riparius* allele frequency distributions.

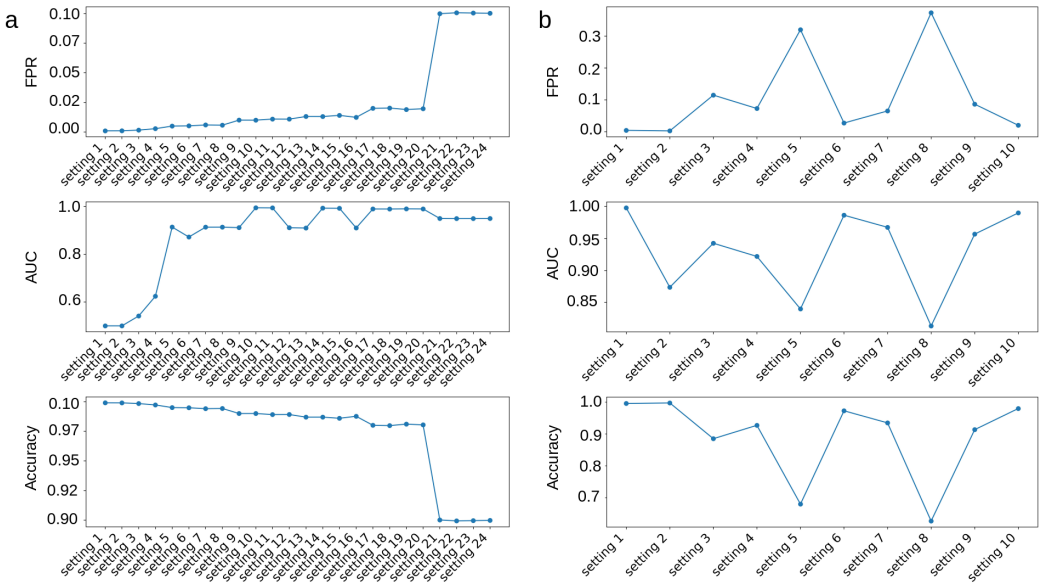

Fig. 7.1.: Parameter optimization for OCSVM and NBC algorithms. (a) Performance metrics (False Positive Rate, Area Under the Curve, and Accuracy) for various parameter settings of the OCSVM radial kernel. (b) Corresponding metrics for different NBC parameter configurations.

Setting	*nu*	*gamma*	FPR	AUC	ACCURACY
Setting 1	0.001	0.01	0.001	0.500	0.999
Setting 2	0.001	0.05	0.001	0.499	0.999
Setting 3	0.001	0.5	0.002	0.541	0.998
Setting 4	0.001	1	0.003	0.624	0.997
Setting 5	0.005	0.01	0.005	0.914	0.995
Setting 6	0.005	0.05	0.005	0.872	0.995
Setting 7	0.005	0.5	0.006	0.914	0.994
Setting 8	0.005	1	0.006	0.914	0.994
Setting 9	0.01	0.01	0.010	0.912	0.990
Setting 10	0.01	0.05	0.010	0.995	0.990
Setting 11	0.01	0.5	0.011	0.995	0.989
Setting 12	0.01	1	0.011	0.911	0.989
Setting 13	0.013	0.01	0.013	0.910	0.987
Setting 14	0.013	0.05	0.013	0.993	0.987
Setting 15	0.013	0.5	0.014	0.993	0.986
Setting 16	0.013	1	0.012	0.910	0.988
Setting 17	0.02	0.01	0.020	0.990	0.980
Setting 18	0.02	0.05	0.020	0.990	0.980
Setting 19	0.02	0.5	0.019	0.991	0.981
Setting 20	0.02	1	0.020	0.990	0.980
Setting 21	0.1	0.01	0.100	0.950	0.900
Setting 22	0.1	0.05	0.101	0.950	0.899
Setting 23	0.1	0.5	0.100	0.950	0.900
Setting 24	0.1	1	0.100	0.950	0.900

Fig. 7.2.: Parameter optimization for OCSVM algorithm. Tested parameter settings (ν and γ) for the OCSVM algorithm with their respective performance metrics.

Setting	*mu1*	*sigma1*	*mu2/mu3*	*sigma2/sigma3*	FPR	AUC	ACCURACY
Setting 1	0.0, 0.0	0.15, 0.1	0.1, 0.9	0.25, 0.2	0.004	0.998	0.996
Setting 2	0.0, 0.0	0.25, 0.18	0.2, 0.7	0.2, 0.199	0.003	0.874	0.997
Setting 3	0.0, 0.0	0.15, 0.1	0.3, 0.4	0.35, 0.25	0.115	0.943	0.885
Setting 4	0.0, 0.0	0.2, 0.1	0.3, 0.5	0.4, 0.2	0.073	0.922	0.927
Setting 5	0.0, 0.0	0.3, 0.1	0.3, 0.6	0.25, 0.2	0.320	0.840	0.680
Setting 6	0.0, 0.0	0.25, 0.2	0.3, 0.8	0.15, 0.1	0.027	0.986	0.973
Setting 7	0.0, 0.0	0.2, 0.15	0.35, 0.6	0.3, 0.2	0.065	0.967	0.935
Setting 8	0.0, 0.0	0.3, 0.2	0.5, 0.4	0.35, 0.3	0.374	0.813	0.626
Setting 9	0.0, 0.0	0.25, 0.2	0.6, 0.3	0.2, 0.15	0.087	0.957	0.913
Setting 10	0.0, 0.0	0.15, 0.1	0.7, 0.2	0.25, 0.2	0.020	0.990	0.980

Fig. 7.3.: Parameter optimization for NBC algorithm. Tested parameter settings for NBC, showing μ and σ values for three approximations along with their performance metrics

The comprehensive analysis examined five distinct approaches (FET, OCSVM, OCSVM-FET, NBC, and NBC-FET) across multiple generational timepoints (10, 20, 40, and 60). The results indicated that generation 40 provided the most favorable conditions for detecting loci under selection (Figure 7.4).

The FPR analysis revealed a consistent decline from generation 10 to 40 across all approaches (Figure 7.4a), suggesting increasing reliability over time. Notably, while NBC and OCSVM exhibited a slight uptick in FPR at generation 60, other approaches maintained their downward trajectory. This trend corresponded with AUC performance (Figure 7.4b), where all methods except FET achieved peak effectiveness at generation 40. NBC maintained robust performance (AUC>60%) even at generation 60, while other approaches showed decreased effectiveness, approaching 50%. FET consistently demonstrated performance levels indicative of random chance, while OCSVM-FET emerged as the superior method, particularly at generation 40. Accuracy measurements (Figure 7.4c) showed progressive improvement across all approaches as generations advanced, stabilizing between generations 40 and 60. The OCSVM-FET approach achieved peak accuracy at generation 40, followed by FET and NBC.

The collective analysis of these metrics established generation 40 as the optimal timepoint for all approaches, considering both accuracy and AUC performance. Although three of the four approaches reached their lowest FPR values in generation 60, generation 40 demonstrated the most consistent performance across all metrics.

An in-depth examination of allele frequency distributions across generational progression (Figure 7.5) and corresponding phenotypic expressions provided

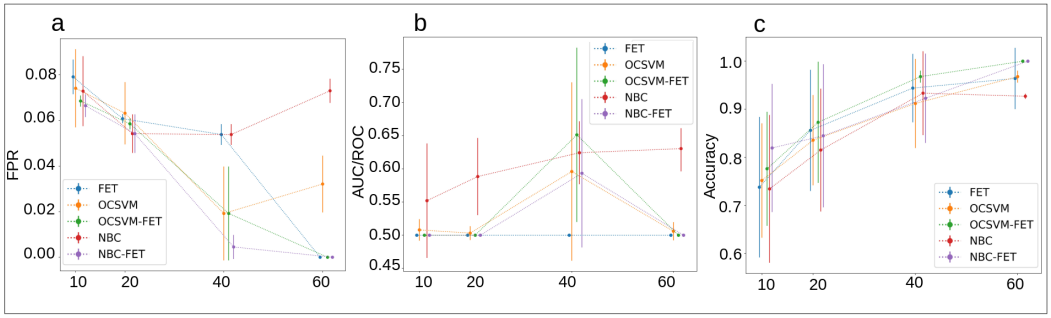

Fig. 7.4.: Performance comparison of five approaches for detecting loci under selection across different generations. The approaches evaluated are FET, OCSVM, OCSVM-FET, NBC, and NBC-FET. Performance metrics shown are (a) FPR, (b) AUC/ROC, and (c) Accuracy. Data points represent mean values across simulations, with error bars indicating standard deviation. Generations tested: 10, 20, 40, and 60.

further insight into these outcomes. The allele frequencies of SNPs identified under selection (depicted in red) exhibited increasing divergence from background frequencies (shown in blue) over successive generations, with particularly pronounced differentiation emerging around generation 40 (Figure 7.5a). Notably, this distribution pattern showed remarkable similarity to the real-life data utilized in the parameter optimization process (Figure 7.5b). This parallel between simulated data at generation 40 and empirical observations helps explain the enhanced performance of the optimized algorithms used in this study. These findings highlight the critical role of temporal dynamics in studying rapid adaptation and demonstrate this methodology's particular strength in identifying selection signatures that have achieved sufficient differentiation without reaching complete fixation. While the approach shows optimal effectiveness in scenarios matching the training data characteristics, it consistently outperforms conventional methods like FET across diverse conditions.

Analysis of phenotypic trajectories revealed distinct adaptive patterns across scenarios with varying numbers of selected loci over the 60-generation period (Fig. 7.6). Simulations incorporating fewer loci (10 and 50) demonstrated rapid initial phenotypic advancement, achieving higher early-generation values but quickly reaching plateau states. In contrast, scenarios with larger numbers of loci (250 and 500) exhibited more gradual initial progression but maintained steady increases throughout the study period, ultimately achieving superior phenotypic values.

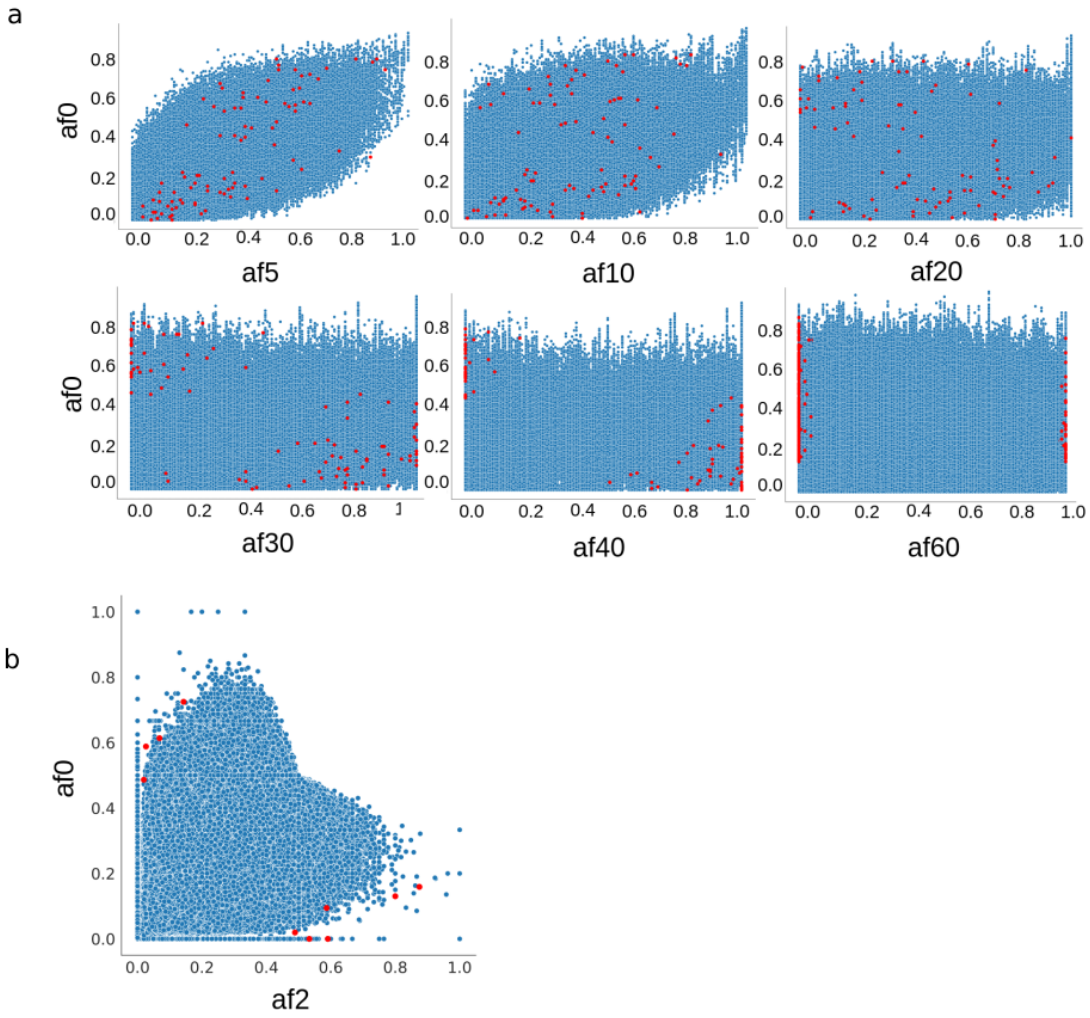

Fig. 7.5.: Allele frequency distributions across generations. (a) Simulated data showing allele frequencies (AF) at generations 5, 10, 20, 30, 40, and 60 plotted against initial allele frequencies (af0). Red dots represent selected SNPs, while blue dots represent background SNPs. (b) Real-life data used for parameter tuning, showing allele frequencies at generation 2 plotted against initial frequencies (af0). The distribution in generation 40 of the simulated data most closely resembles the real-life data.

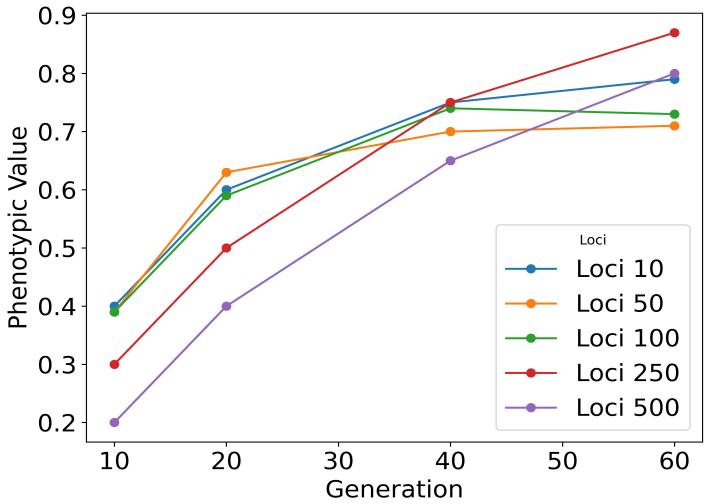

Fig. 7.6.: Phenotypic values over generations for different numbers of selected loci. The plot shows the change in phenotypic values across 60 generations for simulations with 10, 50, 100, 250, and 500 loci under selection. Each line represents a different number of selected loci, with measurements taken at generations 10, 20, 40, and 60. The y-axis represents the calculated phenotypic value, while the x-axis shows the generation number.

The 250-loci simulation revealed a particularly interesting pattern: despite showing the lowest phenotypic value at generation 10, it demonstrated the most pronounced increase between generations 20 and 40, ultimately achieving the highest phenotypic value by generation 60. This pattern correlates strongly with our observation of peak algorithm performance at generation 40 for the 250-loci scenario, indicating that this time-point provides an optimal balance between the strength of selection pressure and the time required for adaptive differentiation to become detectable.

The intermediate case of 100 loci exhibited characteristics bridging the extremes, showing moderately rapid initial advancement followed by more gradual progression.

7.3 OCSVM-FET shows Superior Performance Across Loci Numbers

Following the identification of generation 40 as the optimal timepoint, a detailed evaluation assessed methodological performance across different

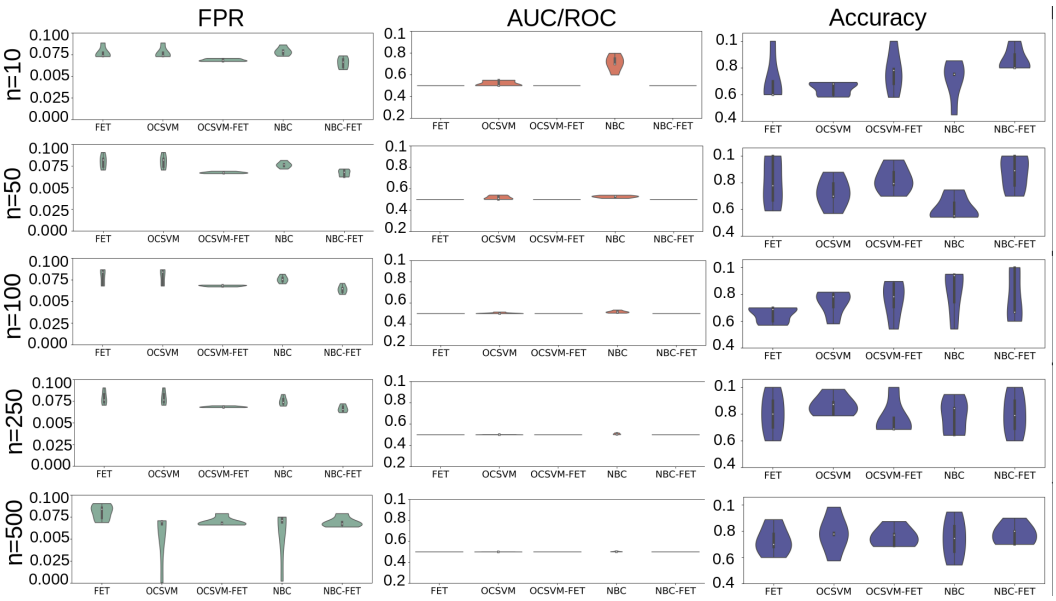

Fig. 7.7.: Comparative analysis of detection approaches across varying numbers of loci under selection at generation 10. The graph displays FPR (first column graphs, green), AUC/ROC (second column graphs, red), and Accuracy (third column graphs, blue for FET, OCSVM, OCSVM-FET, NBC, and NBC-FET. Each row represents the number of loci under selection (n = 10, 50, 100, 250, 500). The y-axis represents the values of the comparative metrics of the respective column

quantities of loci under selection (n = 10, 50, 100, 250, and 500) (Figure 7.9). FPR analysis revealed consistent patterns, with FET and NBC maintaining approximately 0.06 across all loci configurations. NBC-FET demonstrated stable performance with a low FPR of roughly 0.01, while OCSVM and OCSVM-FET exhibited slightly elevated values at n=100, but otherwise maintained approximate values of 0.02.

The AUC performance exhibited considerable variation among methodologies. FET consistently demonstrated values approximating 50% across all loci configurations, suggesting performance equivalent to random chance. OCSVM and OCSVM-FET displayed enhanced performance correlating with increasing loci numbers, reaching peak efficiency at n=250 (OCSVM >70% AUC, OCSVM-FET >80% AUC), followed by a modest decline at n=500 (OCSVM <65% AUC, OCSVM-FET <70% AUC). NBC achieved optimal performance at n=50 (65% AUC), showing gradual decline to approximately 60% as n increased. In contrast, NBC-FET demonstrated progressive improvement with increasing n, stabilizing at n=100 with consistent performance above 60% (see Fig. 7.9).

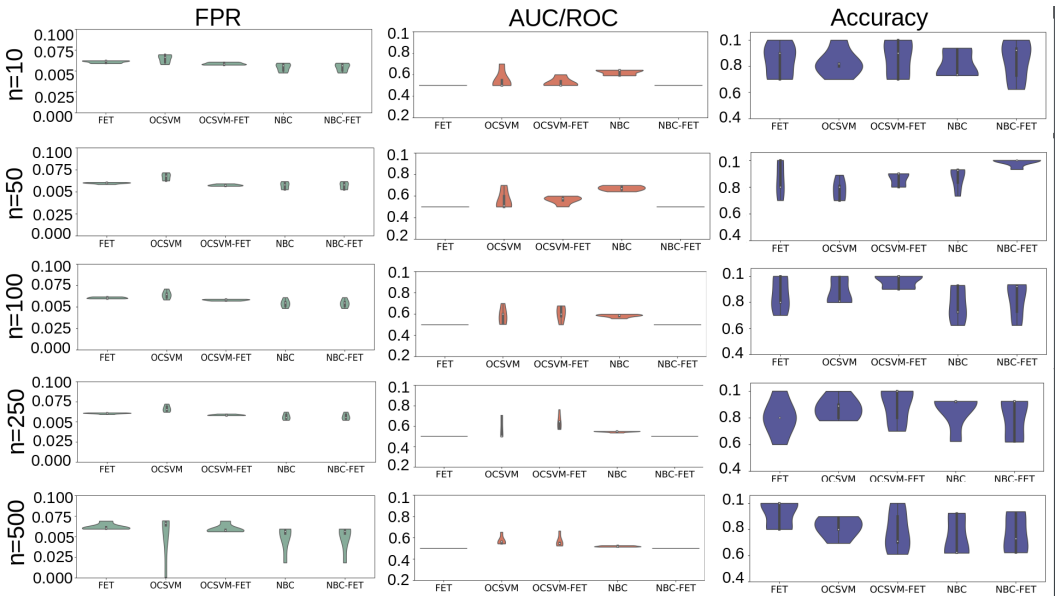

Fig. 7.8.: Comparative analysis of detection approaches across varying numbers of loci under selection at generation 20. The graph displays FPR (first column graphs, green), AUC/ROC (second column graphs, red), and Accuracy (third column graphs, blue for FET, OCSVM, OCSVM-FET, NBC, and NBC-FET. Each row represents the number of loci under selection (n = 10, 50, 100, 250, 500). The y-axis represents the values of the comparative metrics of the respective column

Accuracy assessment revealed that OCSVM-FET achieved remarkable results with n=250 selected loci, approaching 99% accuracy. This method showed generally increasing accuracy with n, though experiencing a slight reduction to 80% at n=500. NBC-FET maintained robust accuracy (approximately 90%) across n=10, 50, 100, and 150, with minor degradation at n=500. FET demonstrated consistent performance around 80% accuracy regardless of n value. Both OCSVM and NBC exhibited steady accuracy improvements correlating with increasing n, reaching maximum effectiveness at n=250 before showing slight decline.

While generation 40 emerged as the primary focus due to peak performance, additional evaluations at generations 10, 20, and 60 were conducted to assess the stability of detection signals over time and establish a broader understanding of temporal dynamics in polygenic adaptation.

In the early phase (generation 10, Fig. 7.7), all methodologies demonstrated relatively modest performance across loci configurations. AUC/ROC values generally ranged between 50-60%, indicating marginally better than random performance. Accuracy measurements showed substantial variability across

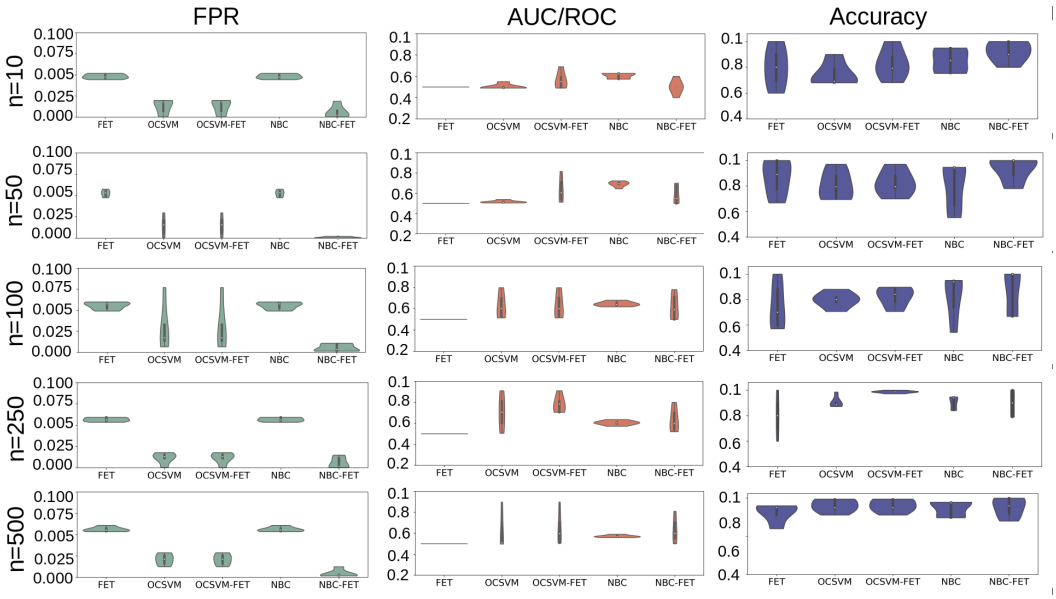

Fig. 7.9.: Comparative analysis of detection approaches across varying numbers of loci under selection at generation 40. The graph displays FPR (first column graphs, green), AUC/ROC (second column graphs, red), and Accuracy (third column graphs, blue for FET, OCSVM, OCSVM-FET, NBC, and NBC-FET. Each row represents the number of loci under selection (n = 10, 50, 100, 250, 500). The y-axis represents the values of the comparative metrics of the respective column

methods and loci numbers. By generation 20 (Fig. 7.8), notable improvements emerged, particularly in scenarios with higher loci numbers (n=250, 500). OCSVM and OCSVM-FET began showing enhanced AUC/ROC values, especially for n=250 and n=500, though not reaching the levels observed at generation 40. At generation 60 (Fig. 7.10), most methods exhibited performance plateaus or slight declines compared to generation 40, particularly evident in scenarios with higher loci numbers (n=250, 500). This trend was most pronounced in AUC/ROC and Accuracy measurements for OCSVM and OCSVM-FET. FPR remained relatively stable across generations for most methodologies.

7.4 Optimal Framework Performance

Extensive analysis identified the OCSVM-FET approach at generation 40 with 250 selected loci as the most effective framework for detecting polygenic adaptation patterns (Fig. 7.11). Under these conditions, OCSVM-FET achieved:

- Minimal False Positive Rate,

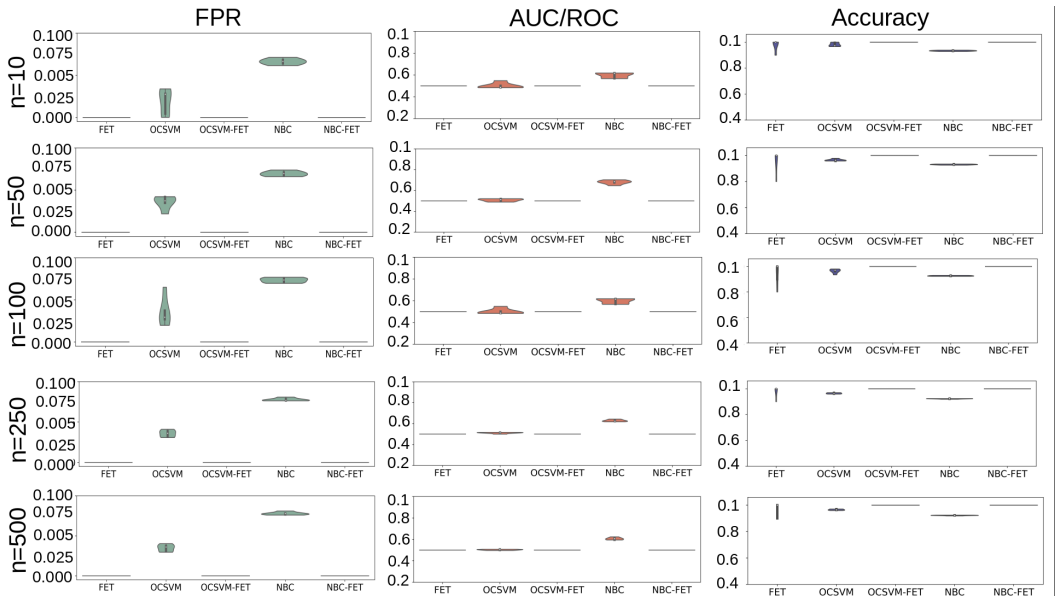

Fig. 7.10.: Comparative analysis of detection approaches across varying numbers of loci under selection at generation 60. The graph displays FPR (first column graphs, green), AUC/ROC (second column graphs, red), and Accuracy (third column graphs, blue for FET, OCSVM, OCSVM-FET, NBC, and NBC-FET. Each row represents the number of loci under selection (n = 10, 50, 100, 250, 500). The y-axis represents the values of the comparative metrics of the respective column

- AUC values surpassing 80% ,
- Exceptional accuracy approaching 99%

This particular parameter combination provided optimal balance between detection sensitivity and specificity, establishing OCSVM-FET as a robust tool for identifying polygenic adaptation signatures in genomic data.

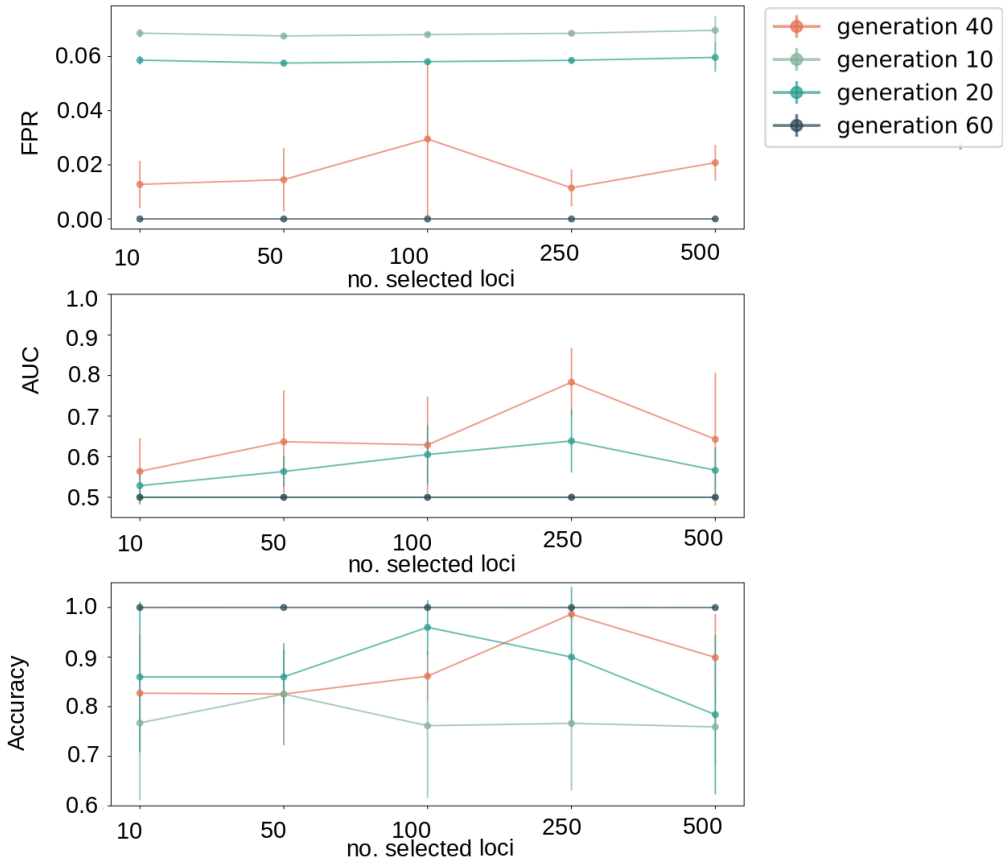

Fig. 7.11.: Performance metrics of the OCSVM-FET approach at generation 10, 20, 40 and 60 for different numbers of selected loci. The graph illustrates FPR, AUC, and Accuracy. The x-axis represents the number of loci under selection (n = 10, 50, 100, 250, 500). Figure highlights the optimal performance achieved with 250 selected loci, demonstrating the lowest FPR, highest AUC, and near-perfect accuracy.

8 Experimental Evolution: Selection on Emergence Time in *C. riparius*

The experimental evolution study demonstrated rapid adaptation of multiple fitness components in C. riparius. A distinct two-phase pattern emerged: initial rapid improvement in emergence timing followed by continued enhancement of survival and reproductive success. Replicate populations developed different adaptive strategies, revealing multiple viable paths to fitness optimization. While some populations achieved faster development times, others showed enhanced survival, demonstrating how similar selection pressures can lead to diverse but equally successful adaptive solutions. This divergence in adaptive strategies highlights the flexibility of rapid evolution, even under controlled laboratory conditions.

8.1 Data Evaluation Experiment

Adult Emergence Rate Across Generations

The experiment observed a significant overall decrease ($p = 1e-07$) in mean emergence time across generations. The initial generation displayed an average emergence time of 21.4 ± 1.35 days, which was reduced to 19.9 ± 1.06 days by the seventh generation. Notably, the fourth generation exhibited the shortest emergence time, averaging only 18.3 ± 0.97 days (Fig. 8.1B,F). This pattern of decline extended to the EmT50 metric (time required for 50% emergence) (see Fig. 8.1E).

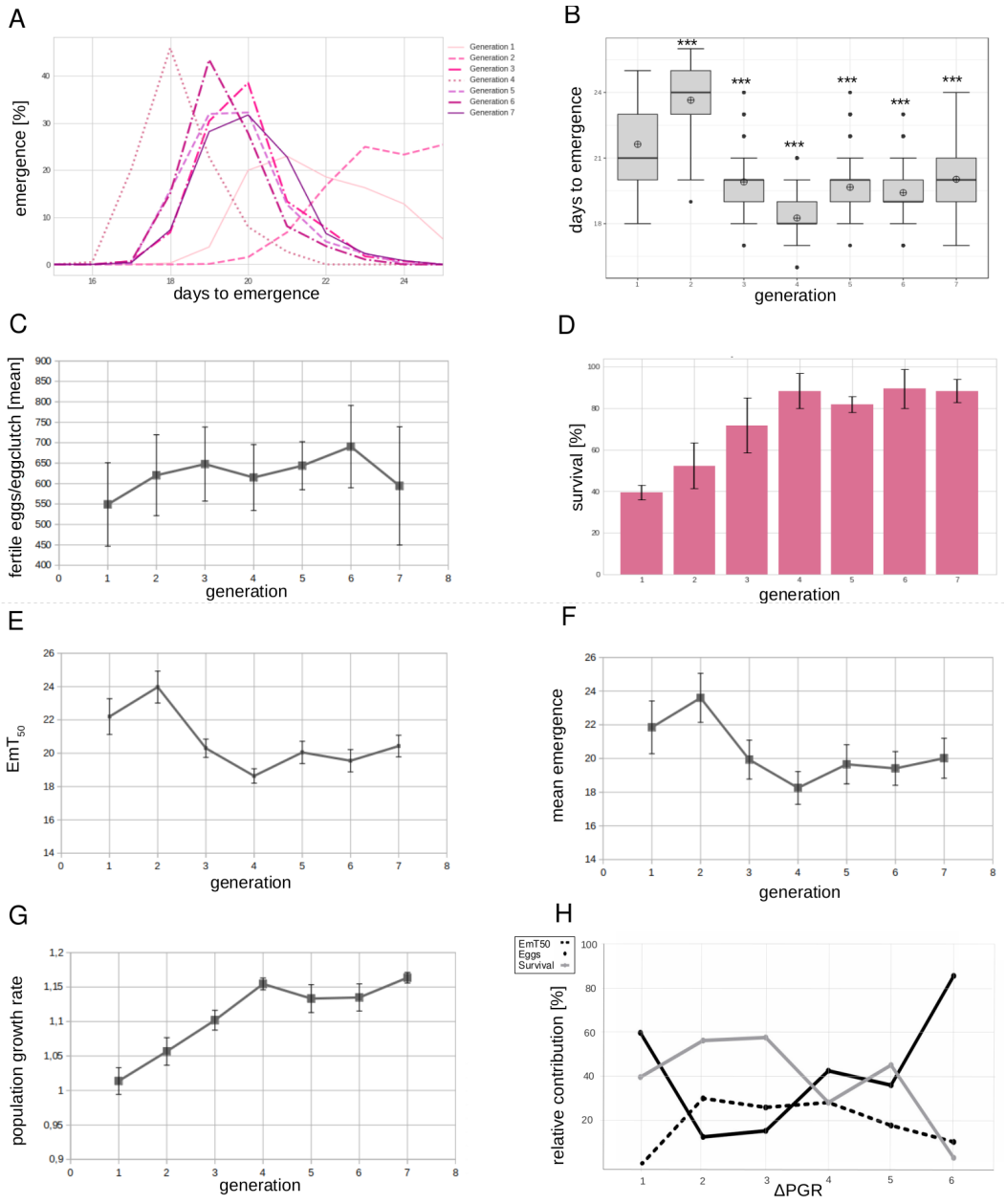

Fig. 8.1.: Overview of Emergence of all 4 replicates combined (Mean), regarding EmT50, Eggrate and PGR over time (combined replicates). The overview displays A) an overview of the distribution of emergence days per generation in percentage as a lined plot and B) the distribution of the emergence day per generation as boxplot. C) The fertility per generation is estimated from fertile egg ropes laid during the first three days of egg production. D) Survival of adult midges in percentage. E) EmT50 was calculated as the time point, when 50% of the population was ermerged as well as F) the mean emergence per generation. G) The Population growth rate (PGR) combines fertility, emergence and survival of the replicates. H) Contribution of the Parameters to delta PGR. Error bars denote for the standard deviation and asterisks indicate differences in comparison to the initial (first) generation ($p < 0,001$, Mann-Whitney U Test)

Each generation of each replicate was subjected to two independent analyses. The Mann-Whitney U test with a subsequent Cohen's D effect size test assessed classical frequentist statistics. Bayesian analysis was conducted in parallel. Figures 8.2 and 8.3 depict the results for emergence rate analysis, comparing each generation to the ancestral (first) generation. Replicate gold's second generation showed exceptional mean emergence (23.15 ± 0.992, A.10), exceeding the ancestral generation's probable range (19.188 - 22.970) (Appendix, A.13) with significant statistical deviation ($p = 1.7e-11$, $D = 2.82$, Tab. 8.1). All other replicates also demonstrated significant emergence increases versus the ancestral generation, with subsequent generations falling within their respective 94% HD intervals.

Mean emergence declined from the third generation, reaching its lowest point in generation four across all replicates (red: $p = 7.4e-1$, $D = 2.71$; blue:$p = 3.4e-11$, $D = 2.37$; green: $p = 2.9e-11$, $D = 4.75$; gold: $p = 4.3e-11$, $D = 2.39$, Tab. 8.1). Emergence rates increased and stabilized from generation four onward, maintaining levels above the ancestral generation. The seventh generation showed significant changes (red:$p = 4.8e-3$, blue:$p = 2.8e-3$, gold:$p = 4.7e-3$, green:$p = 3.6e-1$, Tab. 8.1) with small to medium effect sizes (red:$D = 0.60$, blue:$D = 0.26$, gold:$D = 0.10$, green:$D = 0.78$, Tab. 8.1). These values exceeded the Mean HDI while remaining within the 94% HD interval (Appendix, A.11).

EmT50 displayed similar trends (Fig. 8.2), with observed values consistently within the 94% HDI. All generations and replicates demonstrated significant differences with large effect sizes compared to the ancestral generation ($p < 0.001$, $D > 0.8$, Tab. 8.1).

Survival Across Generations

The overall percentage of emerged adult midgets demonstrated a significant upward trend ($p = 1.4e-06$). The first generation exhibited a 40 ± 1.75% emergence rate, whereas the final generation achieved an 88 ± 2.80% emergence rate. Adult survival also displayed a significant improvement ($p = 1.4e-06$) from the first generation to the fourth. This improvement plateaued afterwards, with emergence rates remaining stable until the seventh generation (see Fig. 8.1D, A.2).

As displayed in Fig. 8.5C, each replicate exhibited an increase in survival rate across generations, starting with survival around 25 - 30% and reaching

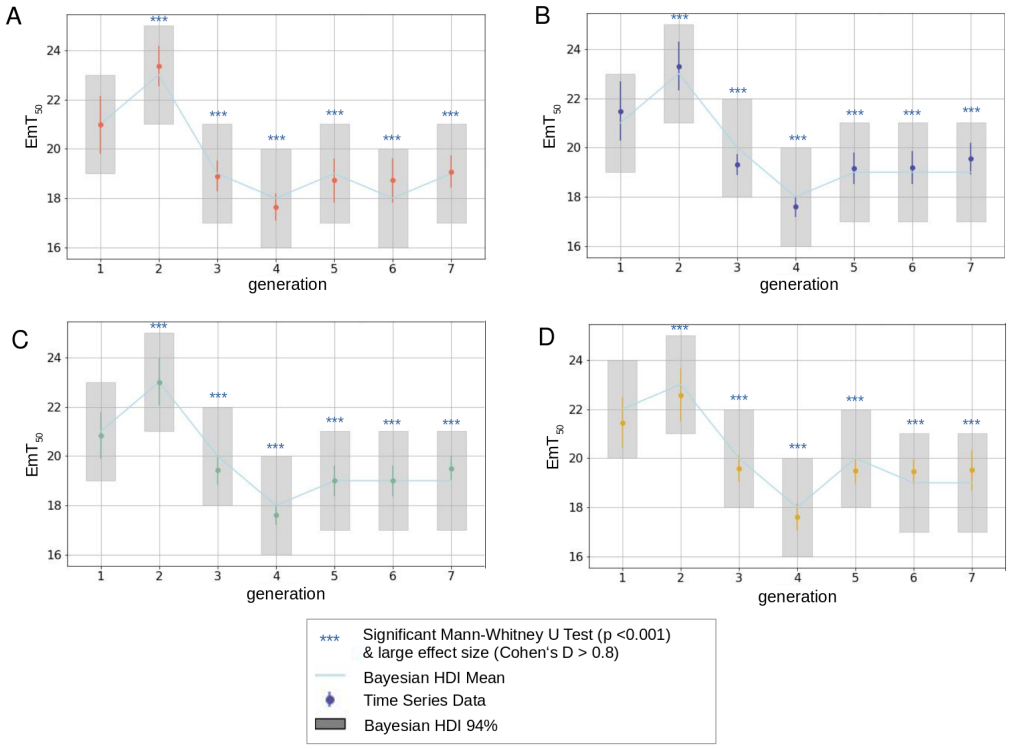

Fig. 8.2.: EmT50 for each generation and replicate. Blue asterisks denote statistically significant differences ($p < 0.001$) accompanied by a large effect size (Cohen's D > 0.8). The gray boxes represent the 94% Bayesian Highest Density Interval (HDI), indicating the range of credible values for the data based on the posterior distribution. Data points falling within their respective HDI boxes are consistent with the most probable range for the true value within this analysis. The gray line connecting the boxes represents the mean of the Bayesian HDI. EmT50 values are displayed as a dot with respective standard deviation in the respective color-code for the replicate. A) replicate red B) replicate blue C) replicate green and D) replicate gold

around 70 - 80% survival by generation seven. Notably, replicate red showed the fastest survival increase, reaching around 76.4% already at generation three. Replicate green showed the highest survival rate of all four replicates in generation four with nearly 88.4% of survival and plateaued around this value by generation seven (A.10).

Fertility and Sex Ratio Show Positive Shifts

The experiment revealed a significant overall increase ($p = 8.4e - 06$) in the number of fertile egg ropes laid by females across generations. The first generation produced the fewest egg ropes, while the last generation produced the most (see Fig. 8.1C, Appendix A.2). The sex ratio of adult females across all

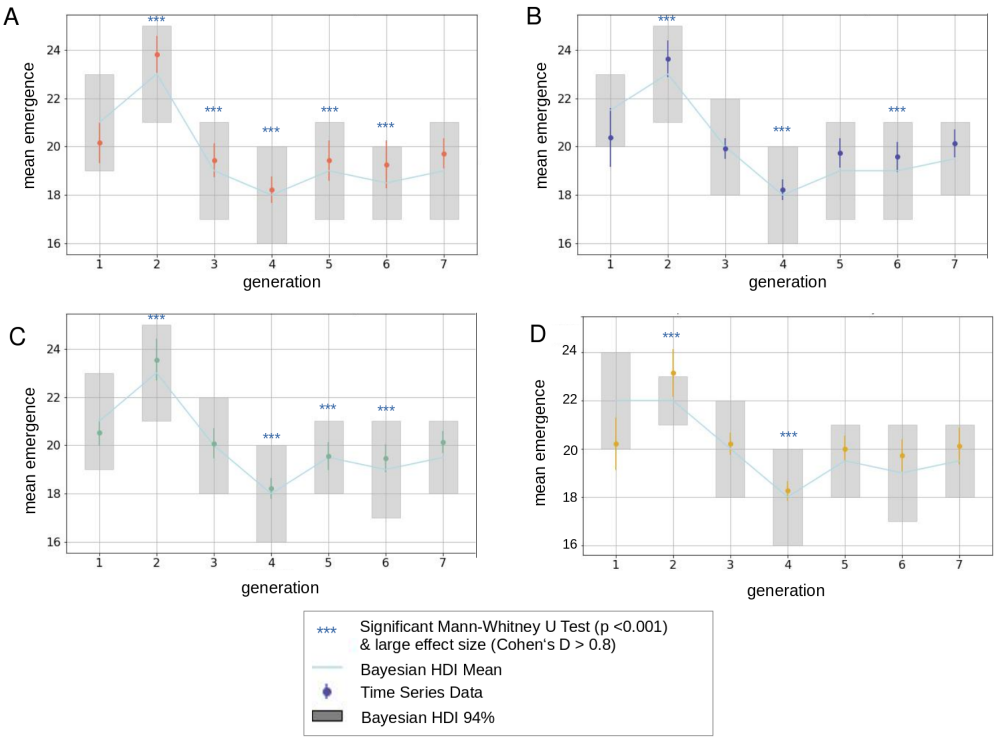

Fig. 8.3.: Mean Emergence Time for each generation and replicate. Blue asterisks denote statistically significant differences ($p < 0.001$) accompanied by a large effect size (Cohen's D > 0.8). The gray boxes represent the 94% Bayesian Highest Density Interval (HDI), indicating the range of credible values for the data based on the posterior distribution. Data points falling within their respective HDI boxes are consistent with the most probable range for the true value within this analysis. The gray line connecting the boxes represents the mean of the Bayesian HDI. Mean emergence values are displayed as a dot with respective standard deviation in the respective color-code for the replicate. A) replicate red B) replicate blue C) replicate green and D) replicate gold

generations remained statistically similar ($p = 0.154$), with values fluctuating between 0.4 and 0.5 without exceeding the upper limit (Appendix Fig. A.2. Fig.A.9).

As displayed by Fig. 8.5D, the fertility measured by fertile egg ropes per female varied across the replicates, all being significantly different than the ancestral generation ($p < 1.5e-4$, $D > 9.35$, Tab. 8.1). Replicates red and green consistently achieved the highest fertility across generations two to five. Conversely, gold displayed the lowest fertility in these generations, as well as in generation six. Notably, red surged to the absolute highest fertility level by generation seven. Overall, the data demonstrates a trend of red and green maintaining a fertility advantage over blue and gold. While overall

adult female sex ratio remained statistically consistent across generations, replicates red, blue, and gold exhibited a notable "X" pattern. These replicates began and ended with a male-dominant population, punctuated by a distinct dip in generation four where females outnumbered males. Replicate green, however, displayed a distinct trend. It maintained equal male-female ratios in generation two and exhibited a surge in females during generation three (see A.9)

Population Growth Rate Reflects Fitness Improvement

The population growth rate (PGR), a metric of overall replicate fitness, displayed a significant increase from the first generation to the fourth ($p = 0.0001$). This initial rise was primarily driven by the faster emergence rate. The PGR continued to increase, albeit at a slower pace, until the final generation ($p = 1.8e - 06$) (see Fig. 8.1G, Appendix Fig. A.3). Fertility became the key driver of this later growth (see Fig. 8.1H).

Two independent analyses were conducted for each replicate of each generation: a Mann-Whitney U test with Cohen's D effect size for frequentist statistics, and a parallel Bayesian analysis. The Population Growth Rate (PGR) most effectively captured the subtle differences between the four replicates (Fig. 8.4). Despite lower emergence rates in the second generation, all replicates demonstrated higher PGR values at time-step 7 compared to the ancestral generation, emphasizing the influence of fertility and survival (Fig. 8.1H). Replicate red reached its peak PGR at generation seven (1.194 ± 0.009, A.10). From generation two to four, PGR increased steadily before showing slight decreases in generations five and six. All effects were statistically significant with large effect sizes ($p < 0.001$, $D > 0.8$, 8.1) and fell within the 94% HD interval (A.11). Replicate blue's highest PGR occurred at generation six (1.153 ± 0.016, A.10), showing consistent increases from generation two before stabilizing from generation four onwards. Results were significant with large effect sizes ($p < 0.001$, $D > 0.8$, 8.1) and within the 94% HD interval (A.12). Replicate green peaked at generation seven (1.158 ± 0.005, A.10), notably falling outside the 94% HD interval. This deviation, combined with significant differences and large effect sizes ($p < 0.001$, $D > 0.8$, 8.1), represented a marked departure from Bayesian probability predictions. The PGR pattern mirrored Replicate red's trajectory (A.14). Replicate gold achieved its maximum PGR at generation seven (1.162 ± 0.007, A.10). Following the pattern of other replicates, values increased from generation two before declining in generations five

and six. From generation three onwards, results showed significance with large effect sizes ($p < 0.001$, $D > 0.8$, Tab. 8.1) within the 94% HD (Highest Density) interval (A.13).

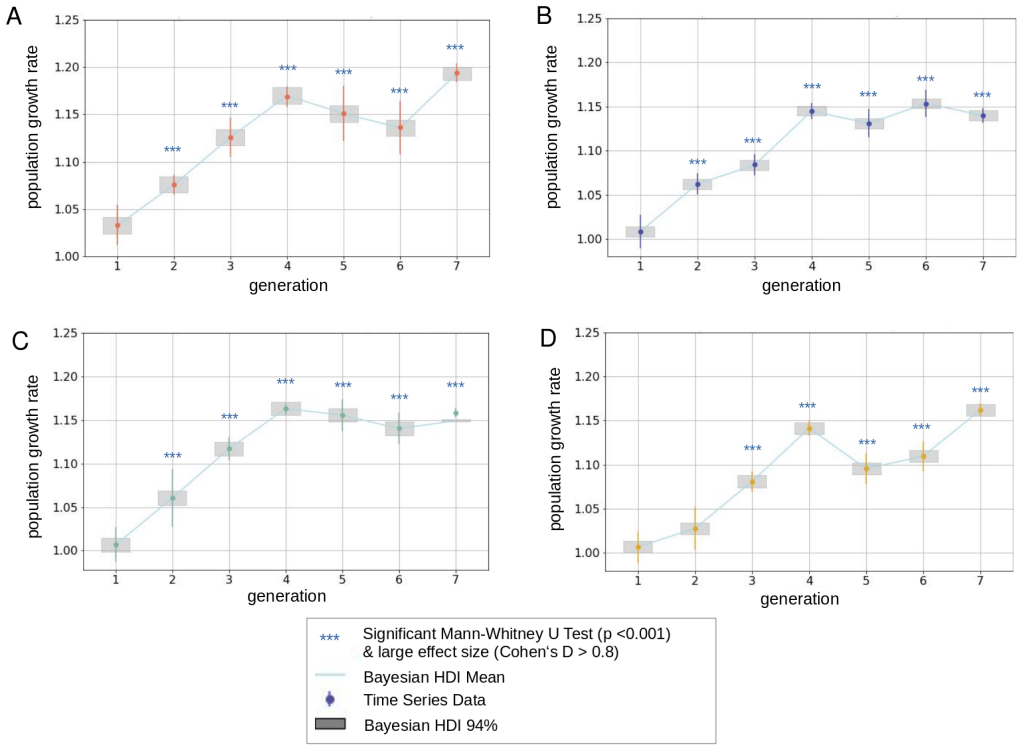

Fig. 8.4.: PGR for each generation and replicate. Blue asterisks denote statistically significant differences ($p < 0.001$) accompanied by a large effect size (Cohen's $D > 0.8$). The gray boxes represent the 94% Bayesian Highest Density Interval (HDI), indicating the range of credible values for the data based on the posterior distribution. Data points falling within their respective HDI boxes are consistent with the most probable range for the true value within this analysis. The gray line connecting the boxes represents the mean of the Bayesian HDI. PGR values are displayed as a dot with respective standard deviation in the respective color-code for the replicate. A) replicate red B) replicate blue C) replicate green and D) replicate gold

Comparison within replicates

A Kruskal-Wallis test evaluated inter-generational differences in emergence, survival, fertility, and PGR across replicates (Figure 8.5). The emergence patterns were largely consistent across replicates, with significant differences observed in EmT50 for generations five ($p = 0.002144$) and six ($p = 0.003867$), and mean emergence for generation three ($p = 0.001763$) (Tab. 8.2). Replicate red consistently demonstrated the fastest emergence rates, followed by Replicate green from generation four onward. Generation four notably showed uniformly low emergence rates with minimal inter-replicate variation.

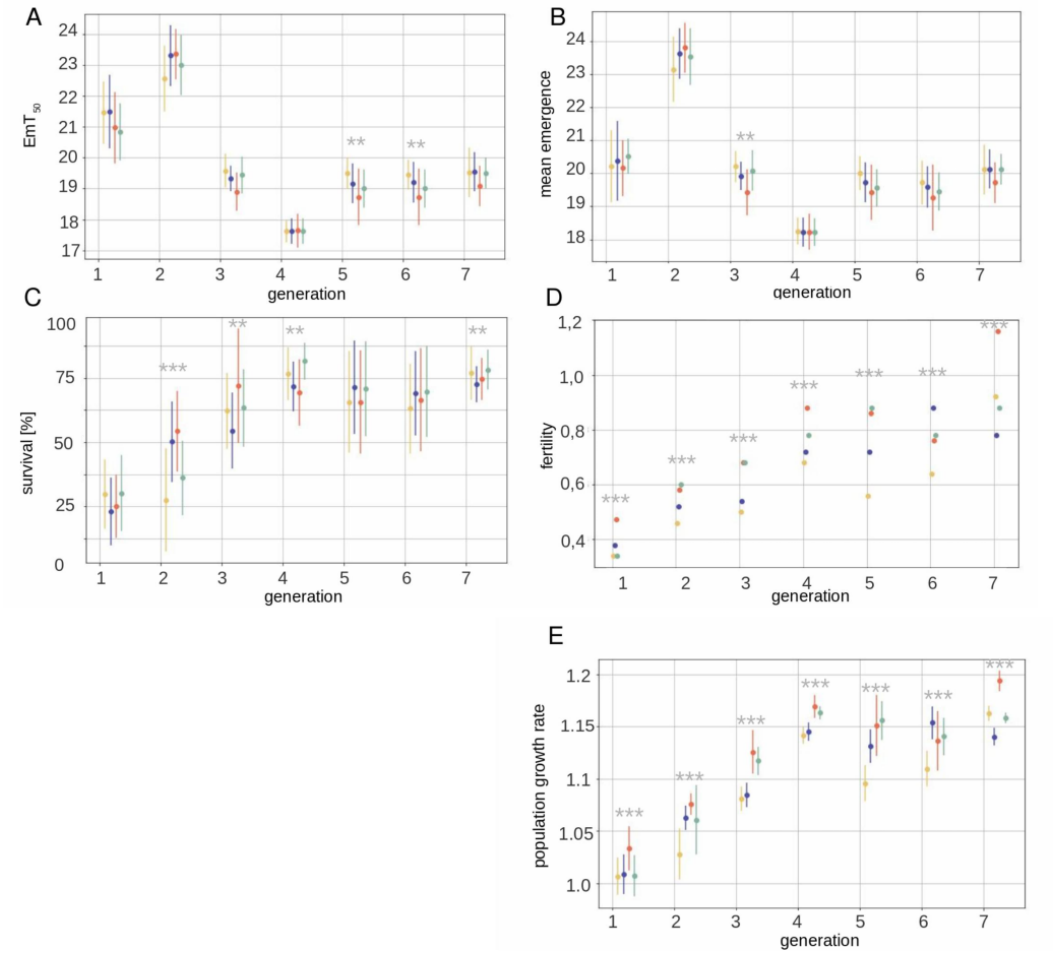

Fig. 8.5.: Overview of Emergence of all 4 replicates (separately) regarding EmT50, mean emergence, survival, fertility and PGR (population growth rate) over time. The respective replicates are color coded. Asterisks indicate significant differences within the 4 replicates; p-values were interfered from Kruskal-Wallis Test A) EmT50 B) mean emergence C) survival D) fertility and E) population growth rate (PGR). Statistical significance and effect size indicated by colored asterics: $**p < 0.005$, $***p < 0.001$; light grey asterics indicates small effect size (eta-squared)

Survival rates showed significant differences in generations two ($p =< 0.000009$), three ($p = 0.003991$), four ($p = 0.003907$), and seven ($p = 0.031493$). Fertility differences were significant across all generations ($p < 5e - 17$) (Tab. 8.2). From generation four onward, Replicate green maintained the highest survival rates. Replicates red and green demonstrated superior fertility throughout, except in generation six. Replicate gold consistently showed the lowest fertility until generation seven. PGR values differed significantly across replicates in all generations ($p < 4e - 05$) (Tab. 8.2). Replicate red achieved among the replicates the highest PGR in generations one through four and seven, while

Tab. 8.1.: Effect Size through Cohen's D as well as p-values obtained from performing Mann Whitney U Test (MWU) to the respective generation against the initial (first) generation

		EmT50		Mean		PGR		Survival		Fertility	
Generation	Replicate	MWU P Value	Cohen's D	MWU P Value	Cohen's D	MWU P Value	Cohen's D	MWU P Value	Cohen's D	MWU P Value	Cohen's D
1 vs. 2	gold	2.7e-5	1.07	1.7e-11	2.82	1.4e-2	0.99	3.7e-1	0.14	1.5e-4	inf
	blue	5.8e-9	1.66	7.1e-12	3.22	3.1e-11	3.42	4.0e-1	1.87	4.9e-13	inf
	red	1.3e-9	2.38	6.6e-12	4.53	7.1e-12	2.57	7.6e-1	2.07	4.9e-13	4.83
	green	1.5e-7	2.29	9.2e-12	4.20	1.1e-9	1.96	2.5e-2	0.40	4.9e-13	inf
1 vs. 3	gold	1.3e-8	2.32	4.9e-1	0.01	6.7e-12	4.92	4.4e-3	2.27	4.9e-13	inf
	blue	1.2e-10	2.43	4.6e-2	0.50	5.9e-12	4.82	2.9e-1	2.24	4.9e-13	inf
	red	5.1e-11	2.21	2.2e-4	0.95	5.6e-12	4.41	5.8e-8	2.60	4.9e-13	9.35
	green	2.1e-9	1.78	2.7e-1	0.74	5.6e-12	6.51	2.6e-4	2.22	4.9e-13	inf
1 vs. 4	gold	5.5e-12	5.06	4.3e-11	2.39	5.6e-12	9.68	3.7e-10	3.91	4.9e-13	inf
	blue	5.5e-12	4.32	3.4e-11	2.37	5.6e-12	9.22	6.4e-9	4.22	4.9e-13	inf
	red	5.5e-12	3.65	7.4e-11	2.71	5.6e-12	8.10	8.0e-7	3.49	4.9e-13	18.39
	green	5.5e-12	4.47	2.9e-11	4.75	5.6e-12	10.61	1.6e-11	4.39	4.9e-13	inf
1 vs. 5	gold	2.4e-9	2.46	1.5e-1	0.24	7.1e-12	5.06	4.5e-5	2.11	4.9e-13	inf
	blue	3.7e-10	2.43	5.5e-3	0.67	5.6e-12	7.01	6.3e-6	3.04	4.9e-13	inf
	red	2.2e-10	2.15	1.3e-4	0.88	1.0e-11	4.63	6.7e-8	2.44	4.9e-13	17.48
	green	5.8e-11	2.31	4.8e-4	1.72	5.6e-12	7.77	3.2e-8	2.43	4.9e-13	inf
1 vs. 6	gold	1.3e-9	2.52	5.0e-3	0.55	5.9e-12	5.89	5.8e-5	2.13	4.9e-13	inf
	blue	6.9e-10	2.37	9.1e-4	0.83	5.6e-12	8.38	7.3e-6	3.08	4.9e-13	inf
	red	2.2e-10	2.15	3.5e-5	0.97	2.5e-11	4.14	2.9e-8	2.48	4.9e-13	12.97
	green	5.8e-11	2.31	1.2e-4	1.90	5.6e-12	7.05	3.8e-8	2.42	4.9e-13	inf
1 vs. 7	gold	6.9e-8	2.12	4.7e-3	0.10	5.6e-12	11.37	1.0e-9	3.87	4.9e-13	inf
	blue	1.7e-8	2.03	2.8e-3	0.26	5.6e-12	9.02	5.3e-10	4.71	4.9e-13	inf
	red	1.9e-10	2.00	4.8e-3	0.60	5.6e-12	9.74	3.0e-10	4.71	4.9e-13	31.04
	green	2.2e-9	1.77	3.6e-1	0.78	5.6e-12	10.41	5.3e-11	4.05	4.9e-13	inf

Tab. 8.2.: Kruskal Wallis P Values to evaluate significant differences within the replicates

	Kruskal Wallis P Value				
Generation	EmT50	Mean	PGR	Survival	Fertility
1	0.131428	0.602947	4.215719e-05	0.235128	6.739990e-16
2	0.068423	0.157274	5.290410e-09	0.000009	5.029423e-17
3	0.006933	0.001763	1.697557e-10	0.003991	5.029423e-17
4	0.894780	0.952157	2.893843e-12	0.003907	5.029423e-17
5	0.002144	0.008559	8.647399e-11	0.479749	5.029423e-17
6	0.003867	0.024512	3.830948e-09	0.393494	5.029423e-17
7	0.111772	0.098725	4.421552e-14	0.031493	5.029423e-17

Replicate gold recorded the lowest values through generations one to six. This pattern aligns with fertility's identified role as a key PGR determinant (Fig 8.1H).

8.2 Bioinformatic Pre-Processing of Pool-Sequencing Data

Prior to analyzing selection signatures across generations, the sequencing data underwent comprehensive preprocessing. For each of the four replicates (red, blue, green, and gold), seven generations of pooled sequencing data were processed, resulting in 28 datasets. The preprocessing pipeline involved converting raw fastq files to mapped and filtered bam files. This critical step ensures data quality and reliability for downstream analyses. The effective-

A

```
105588540 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 secondary
1385156 + 0 supplementary
0 + 0 duplicates
102314484 + 0 mapped (96.90% : N/A)
104203384 + 0 paired in sequencing
52101692 + 0 read1
52101692 + 0 read2
96047182 + 0 properly paired (92.17% : N/A)
100321120 + 0 with itself and mate mapped
608208 + 0 singletons (0.58% : N/A)
3534150 + 0 with mate mapped to a different chr
1736660 + 0 with mate mapped to a different chr (mapQ>=5)
```

B

```
71142664 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 secondary
253308 + 0 supplementary
0 + 0 duplicates
71142664 + 0 mapped (100.00% : N/A)
70889356 + 0 paired in sequencing
35448831 + 0 read1
35440525 + 0 read2
70889356 + 0 properly paired (100.00% : N/A)
70889356 + 0 with itself and mate mapped
0 + 0 singletons (0.00% : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

Fig. 8.6.: Output of samtool's function flagstat for a random chosen bam files (3rd generation, replicate red) A) before filtering steps and B) after filtering steps. Based on the FLAG field values, flagstat classifies reads into various categories. There are typically around 13 categories, including: Mapped reads (primary and secondary), Properly paired reads, Singletons (unpaired reads), Reads failing quality control (QC), Duplicate reads

ness of the filtering process was evaluated by comparing the read statistics before and after filtering.

Comparing the created bam files before and after the filtering steps indicates and overall read reduction. As one representative bam file in Figure 8.6 shows, the total number of reads dropped from ca. 105 millions to ca. 71 millions. This indicates that the filtering process removed approximately 34% of the reads. Furthermore, the filtering steps had an impact on read categories; therefore the percentage of mapped reads increased from 96.90% to 100%. This suggests the filtering likely targeted unmapped reads or reads with low mapping quality. The number of paired-end reads (read1 & read2) and singletons dropped, suggesting the filter targets specifically unpaired reads or low-quality pairs. The percentage of properly paired reads increased to 100%, which shows that all remaining paired reads have their mates mapped to the same chromosome. This could be because the filter removed improperly paired reads or reads where one mate failed the filter. Categories like secondary, supplementary, and duplicates remained mostly unchanged, suggesting the filter likely focused on primary mapped reads.

9 OCSVM-FET Analysis of *C. riparius* Adaptation Patterns

The genetic analysis revealed a two-phase pattern of adaptation, characterized by distinct early and late selection signatures. The initial phase showed consistent selection on fundamental metabolic and regulatory pathways across all populations. Later adaptation demonstrated a shift toward specialized signaling pathways, with populations achieving similar functional outcomes through different genetic solutions. This pattern supports the concept of genetic redundancy in adaptation, where multiple molecular paths can lead to similar phenotypic improvements. The convergence at the pathway level, despite divergent genetic changes, illustrates how populations can navigate different routes through the adaptive landscape while reaching comparable functional outcomes.

9.1 Detection of significant genomic positions

The analysis of the phenotypic adaptation in section 8.1 revealed a complex pattern — rapid initial adaptation until generation 4 followed by phenotypic stabilization through generation 7, despite continued fitness improvement. The investigation of the genetic basis of this pattern began with comprehensive analyses of generation 7, employing a dual approach combining One-Class Support Vector Machines (OCSVM) with Fisher's Exact Test (FET). This strategy identified two distinct categories of selection signatures: broad-effect variants captured by less stringent FET filtering and strong-effect variants detected through strict FET filtering.

To understand how the adaptive process unfolded over time, annotated genes and pathways were investigated in generation 4, when emergence time reached its minimum. This focused analysis of functional elements in generation 4 provided insight into the biological mechanisms driving initial rapid adaptation (detailed results presented in section 9.3).

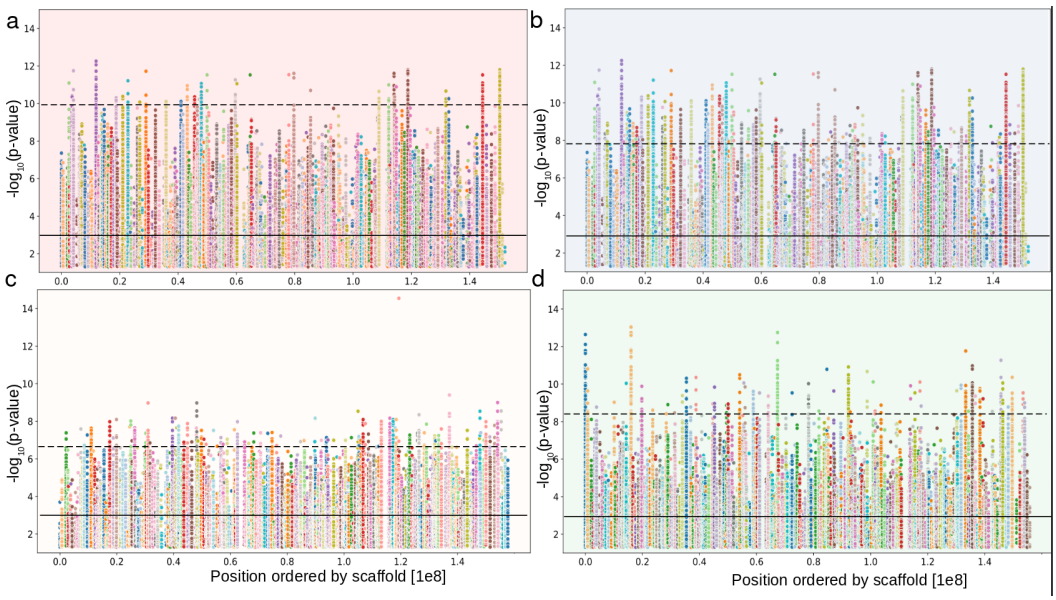

Fig. 9.1.: Manhattan plots displaying variants across scaffolds for four replicates. Reseptive plots shows replicate (a) red (b) blue (c) gold and (d) green. The x-axis shows positions ordered by scaffolds, y-axis represents the negative logarithm of p-values (-log(p-value)). The line represents the cut-off for broad-effect variants (-log(3) = 0.001), dashed line indicates the 0.0001% tail threshold of corrected p-values, serving as cut-off for strong-effect variants varying by replicates: red: 9.66, blue: 7.76, gold: 6.67, green: 8.51. Vertical chimney patterns indicate regions of high statistical significance across multiple adjacent positions within scaffolds.

The power of combining OCSVM with FET lies in its ability to detect coordinated allele frequency changes that might be missed by traditional statistical approaches alone. This initial OCSVM-FET screening with $p < 0.001$ likely captures genomic positions involved in polygenic adaptation, where multiple loci contribute collectively to adaptive traits. A more stringent threshold (0.0001% tail cut-off) was applied to refine the analysis, identifying genomic regions under stronger selection pressure, potentially representing direct targets of selection (hard sweeps) rather than background polygenic effects Therefore, FET analysis with a cutoff of $p < 0.001$ identified $7.25 \pm 0.94\%$ of input genomic positions as significant. OCSVM radial analysis refined this by detecting $0.61 \pm 0.26\%$ of positions as anomalous regions. The integration of both approaches identified broad-effect variants ($FET < 0.001$) in $0.041 \pm 0.004\%$ of positions across replicates, while the most stringent filtering (0.0001% tail of lowest p-values) further isolated a subset of $0.004 \pm 0.0007\%$ positions showing the strongest signatures of selection (Fig. 9.2).

The Manhattan plots (Fig. 9.1) reveal the distribution of these selection signatures across scaffolds, with distinctive chimney patterns indicating regions of high statistical significance. Each replicate showed unique threshold patterns for reduced significant anomalies, with cutoff values with set at 9.66 (red), 7.76 (blue), 6.67 (gold), and 8.51 (green). This analysis revealed both shared and unique adaptive responses: while three genomic positions overlapped among all replicates in the significant anomalies data set, the reduced significant anomalies showed no overlap, suggesting distinct evolutionary trajectories despite similar phenotypic outcomes (Fig. 9.6a, Fig. 9.6 c).

To analyze the dynamics of the identified variants, allele frequency trajectories were tracked over time across all four replicates and compared observed changes against neutral expectations from Wright-Fisher simulations. Figure 9.3 presents these trajectories for the red replicate's strong-effect variants as a representative example, while the corresponding data for the remaining replicates, which displayed similar trends, are provided in the Supplementary Material (Fig. A.15, A.16, A.17). The observed trajectories exhibit an overall linear trend, characterized by either an increasing or decreasing pattern over time. To distinguish selected variants from neutral changes, the empirical -log10(p-values) of allele frequency changes was compared with those obtained from neutral simulations. For example, in the red replicate, variants exceeding the simulated 95% threshold (-log10(p) > 2.071017) were considered candidates for selection, as this threshold corresponds to approximately the <95% quantile in the empirical distribution (empirical 95%= 3.036521), indicating an excess of high-frequency changes beyond neutral expectations. Tables for simulated and observed data's quantiles can be found in the Appendix (Tab. A.2).

9.2 Contrasting signatures of broad-effect and strong-effect variants in genetic diversity

To characterize selection modes during adaptation, genetic diversity (Tajima's π, Watterson's θ, Tajima's D) was analyzed across three datasets: the total genomic dataset, broad-effect variants and strong-effect variants. Analysis of the total genomic dataset revealed significant reductions in genetic diversity between generations 1 and 7 ($p < 0.001$, medium effect size). Reagrding strong-effect variants, the green replicate exhibited the most pronounced

	Replicate	Count	[%]	[%] Mean	[%] Std
Input Genomic Variants	Red	9717640	100.00		
	Blue	9415583	100.00		
	Gold	9224118	100.00		
	Green	9546162	100.00		
Fisher's Exact Test <0.001	Red	812232	8.36	7.25	0.94
	Blue	703256	7.47		
	Gold	562852	6.10		
	Green	674977	7.07		
OCSVM	Red	81410	0.84	0.61	0.26
	Blue	79256	0.84		
	Gold	35118	0.38		
	Green	37698	0.40		
Broad-effect variants	Red	4197	0.043	0.04	0.004
	Blue	4165	0.044		
	Gold	3410	0.037		
	Green	3636	0.038		
Strong-effect variants	Red	485	0.005	0.0040	0.0007
	Blue	363	0.0038		
	Gold	372	0.004		
	Green	321	0.0034		

	Replicate	Count	[%]	[%] Mean	[%] Std		Replicate	Count	[%]	[%] Mean	[%] Std
Annotated Proteins	Red	1376	32.79	35.64	1.95	Annotated unique Gene Families	Red	767	55.74	60.45	4.20
	Blue	1546	37.12				Blue	971	62.81		
	Gold	1248	36.60				Gold	811	64.98		
	Green	1311	36.06				Green	764	58.28		
Annotated Proteins	Red	89	18.35	18.64	5.05	Annotated unique Gene Families	Red	66	74.16	85.21	17.35
	Blue	47	12.95				Blue	47	100.00		
	Gold	67	18.01				Gold	67	100.00		
	Green	81	25.23				Green	54	66.67		

Fig. 9.2.: Summary of FET, OCSVM, OCSVM-FET and protein annotations across four replicates. Overview of genomic positions, FET, OCSVM, OCSVM-FET: broad- and strong effect threshold and annotations for each replicate (red, blue, gold, green) input genomic positions: total number of positions analysed; Fisher's Exact Test (FET) < 0.001: positions identified as significant by FET; OCSVM: regions identified by the OCSVM algorithm; broad-effect variants: determined by FET with corrected p-value < 0.001; strong-effect variants: top 0.0001% smallest FET-corrected p-values; annotated proteins: number and percentage of variants annotated to proteins; annotated unique gene families: number and percentage of annotated proteins assigned to unique gene families using InterProScan and the Pfam database. Counts, percentages, means, and standard deviations are provided where applicable. Annotations were performed using tbg tools and matched to gene families

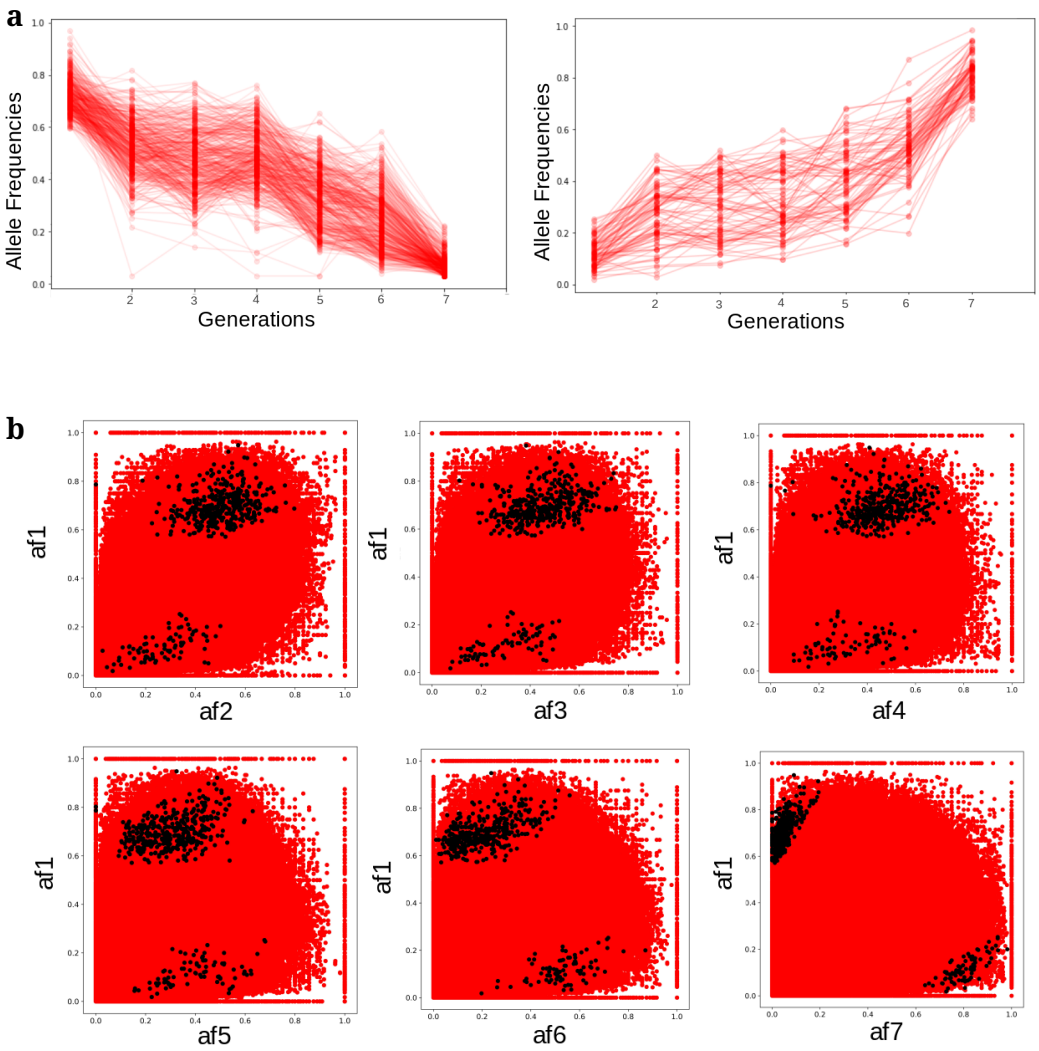

Fig. 9.3.: Allele Frequency (AF) trajectories over all generations of strong-effect variants in replicate red. (a) The left panel shows AF that appeared high in the ancestral generation and increased over time, the right panel shows the opposite cases. (b) Allele Frequency ancestral (af1) against following generations (af*), with strong-effect variants highlighted as black dots.

changes, with π decreasing from 0.024 ± 0.079 to 0.003 ± 0.0044 and θ declining from 0.0214 ± 0.006 to 0.0028 ± 0.0036 ($p < 0.001$, $Cohen'sD > 0.8$). Red and blue replicates showed similar significant reductions in π and θ ($p < 0.001$, $Cohen'sD > 0.8$), while gold displayed smaller, non-significant changes (Fig 9.4a,b) Tajima's D values, which measure the difference between average pairwise nucleotide differences and segregating sites, decreased significantly in red, gold, and green replicates from generation 1 to 7 ($p < 0.001$, $Cohen'sD > 0.8$). These replicates reached negative values in generation 7, indicating an excess of rare alleles. The gold replicate showed only slight decreases, maintaining positive values throughout (Fig 9.4c). All data can be found in Appendix (Tab. A.6. The broad-effect variants maintained higher nucleotide diversity (π and θ) compared to both the strong-effect variants and whole genomic datasets across generations. All replicates showed significant decreases in θ ($p < 0.001$, Cohen's $D > 0.2$, Tab 2 Supplement), with similar trends in π, except for the blue replicate which remained stable ($p < 0.005$, Cohen's $D > 0.2$, Fig. 9.4a,b). Tajima's D decreased in red and gold replicates ($p < 0.001$, Cohen's $D > 0.8$) but increased in blue and green replicates ($p < 0.001$, Cohen's $D > 0.8$), while maintaining positive values that suggest a lack of rare alleles (Fig 9.4c). All data can be found in Appendix (Tab. A.4.

9.3 Distinct genetic paths lead to convergent adaptation

Population differentiation analysis demonstrated increasing genetic divergence across generations despite similar phenotypic outcomes. Pairwise F_{ST} values increased from 0.027 ± 0.016 in early generations to 0.057 ± 0.041 by generation 7, signifying substantial heterozygosity loss (Fig. 9.5). Among replicates, blue and green consistently exhibited higher pairwise F_{ST} values compared to gold and red. Statistical analysis revealed significant differentiation among replicates with medium effect size in generations 2, 3, and 5 ($p < 0.001$, $eta - squared > 0.06$) relative to generation 1, while remaining generations showed significant differences with low effect size ($p < 0.001$, $eta - squared < 0.01$) (Tab. A.8). These patterns suggest divergent evolutionary trajectories despite pathway-level convergent adaptive outcomes. Gene family analysis further supported this pattern of distinct genetic solutions. Broad-effect variants showed substantial numbers of unique gene families across replicates (red: 767, blue: 971, gold: 811, green: 764), with 210 gene

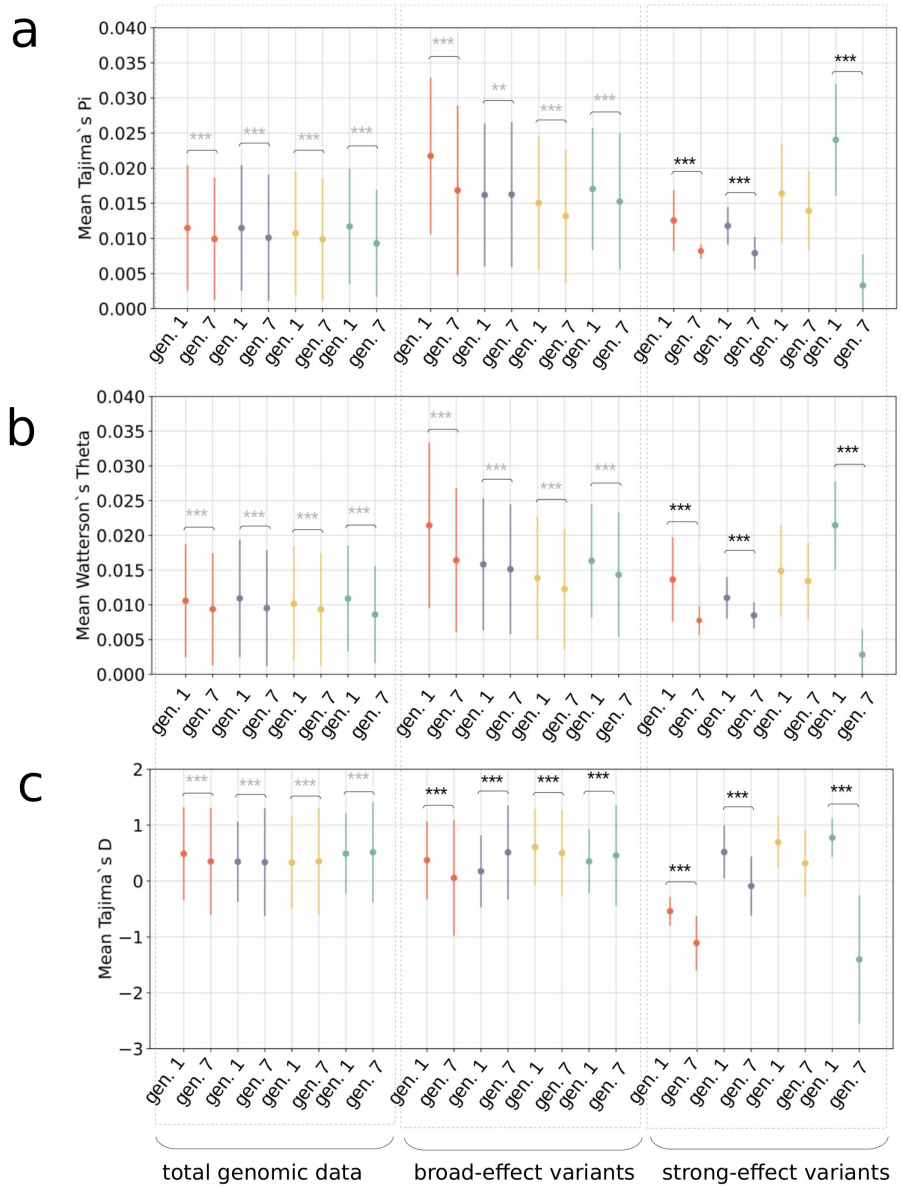

Fig. 9.4.: Analysis of nucleotide diversity. Comparison of (a) Tajima's D, (b) Tajima's π and (c) Watterson's θ across three groups comparing total genomic dataset, broad-effect variants and strong-effect variants. Each panel displays mean values (dots) and standard deviations (error bars) for each replicate (color-coded) and group. Generation 1 (gen.1) was compared to generation 7 (gen.7), with statistical significance and effect size indicated by colored asterics: $**p < 0.005$, $***p < 0.001$; colors: light grey= small effect, dark grey = medium effect, black = large effect.

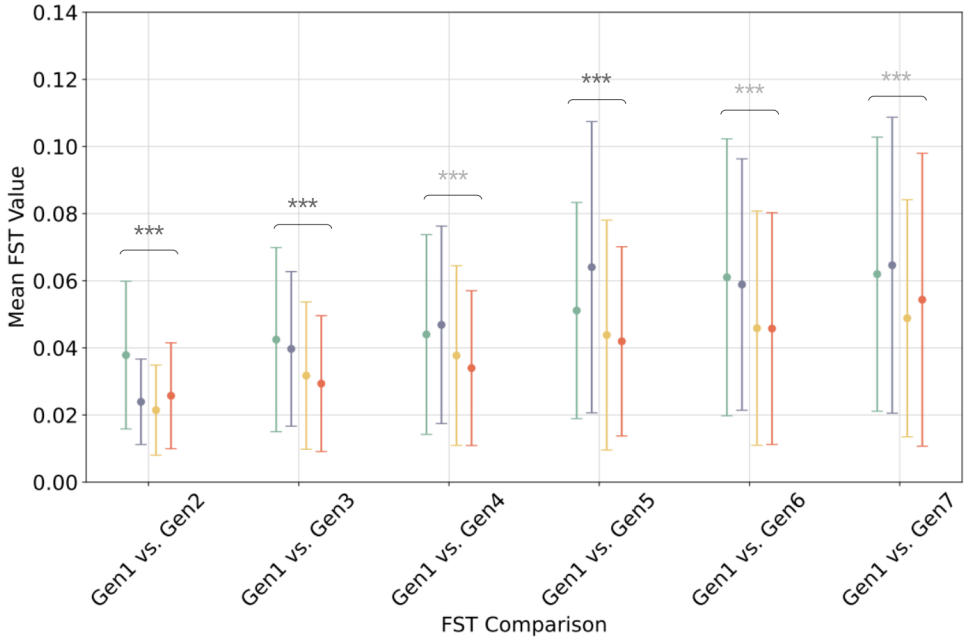

Fig. 9.5.: F_{ST} between the ancient and the respective next generation in non-overlapping 1kb window-size. To test for differences of the replicates among each other, Kruskal-Wallis Test with eta-squared effect size was calculated. Significance and effect size are represented by shaded asterics: $**p < 0.005$, $***p < 0.001$; colors represent effect sizes: light grey = small, grey = medium, black= large.

families shared among all populations (Fig. 9.6 b). This overlap in broader selection targets suggests common genetic pathways involved in initial adaptation. However, strong-effect variants revealed striking replicate-specific patterns, with highly unique positions per population (blue: 46, green: 81, gold: 66, red: 87) and minimal overlap (Fig. 9.6 d).

Having established the selection signatures and their genetic targets in generation 7, the next analysis investigated how these patterns emerged over time. Generation 4 represented a critical transition point where emergence time reached its optimum, even as populations continued evolving. The dual threshold approach applied to generation 4 identified both broad- and strong-effect variants, revealing the genetic basis of this early adaptation phase. In generation 4, broad-effect variants showed enrichment in fundamental regulatory and structural elements across all replicates, including immunoglobulin domains, zinc-related proteins, and calcium binding gene families (Fig 9.8 a). Strong-effect variants contained remarkably few annotations (2-6 proteins per replicate), suggesting initial selection had not yet crystallized into defined adaptive pathways (Fig 9.8 b). Pathway analysis provided further insight

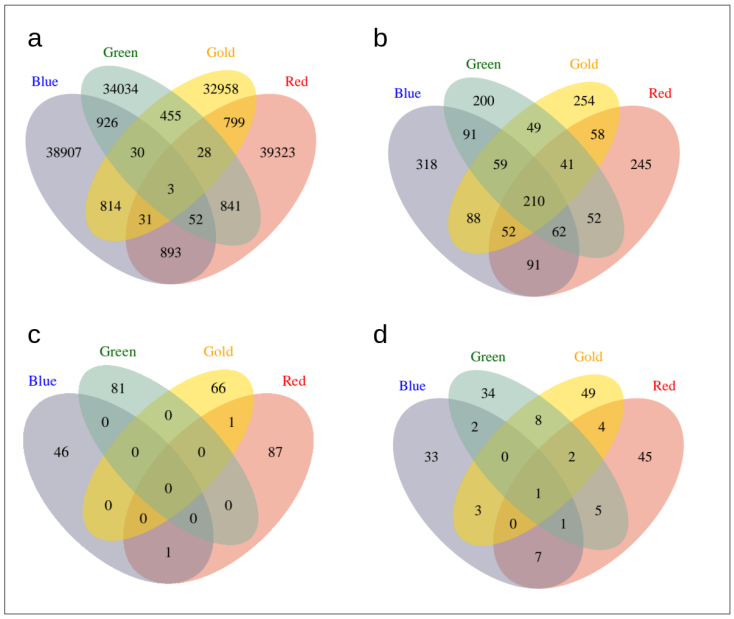

Fig. 9.6.: Venn Diagrams Comparing Overlapping Elements Across Four Replicate Populations. Figure shows four Venn diagrams (a-d) illustrating the overlap of elements among four replicate populations, represented by the colors blue, green, gold, and red. a) broad-effect variants: Overlap of positions across 4 replicates b) broad-effect variants: Overlap of annotated gene families across 4 replicates c) strong-effect variants: Overlap of positions across 4 replicates d) strong-effect variants: Overlap of annotated gene families across 4 replicates

into this temporal evolution. Generation 4 showed consistent enrichment in metabolic pathways across all replicates (Fig 9.9 a), particularly in retinol metabolism and ABC transporters. Additional metabolic pathways included porphyrin metabolism, cytochrome p450 metabolism, and drug metabolism, suggesting broad optimization of metabolic processes during early adaptation. By generation 7, while metabolic pathways remained enriched mainly in the broad data set (Fig 9.9 b), a striking shift toward signaling pathway specialization emerged in the narrow data set (Fig 9.10). All replicates showed enrichment in key signaling pathways, including mTOR, Wnt, Toll/Imd, and MAPK signaling. This transition from predominantly metabolic to signaling pathway enrichment suggests a shift from broad metabolic optimization to refined regulatory control. Though, pathways are shared amongst replicates, the analysis on gene family level reveals different strategies to the similar pathways: while in replicate red Fibronectin type III and Alpha amylase domains are mainly represented, replicate blue showed enriched Sulfotransferase domains. While gold developed increased Calcium-binding EGF domains, green showed increased Leucine-rich repeats and Trypsin domains (Fig 9.7).

Therefore, the reduced significant anomalies data set in generation 7 revealed additionally replicate-specific specialization in distinct signaling pathways, including Hippo, Hedgehog, and FoxO signaling. This temporal comparison reveals a two-phase adaptive process: initial broad selection on major regulatory elements and metabolic pathways, followed by replicate-specific refinement through specialized signaling pathways. This pattern aligns with the phenotypic observations of rapid early adaptation followed by stabilization and continued fitness improvement through alternate mechanisms.

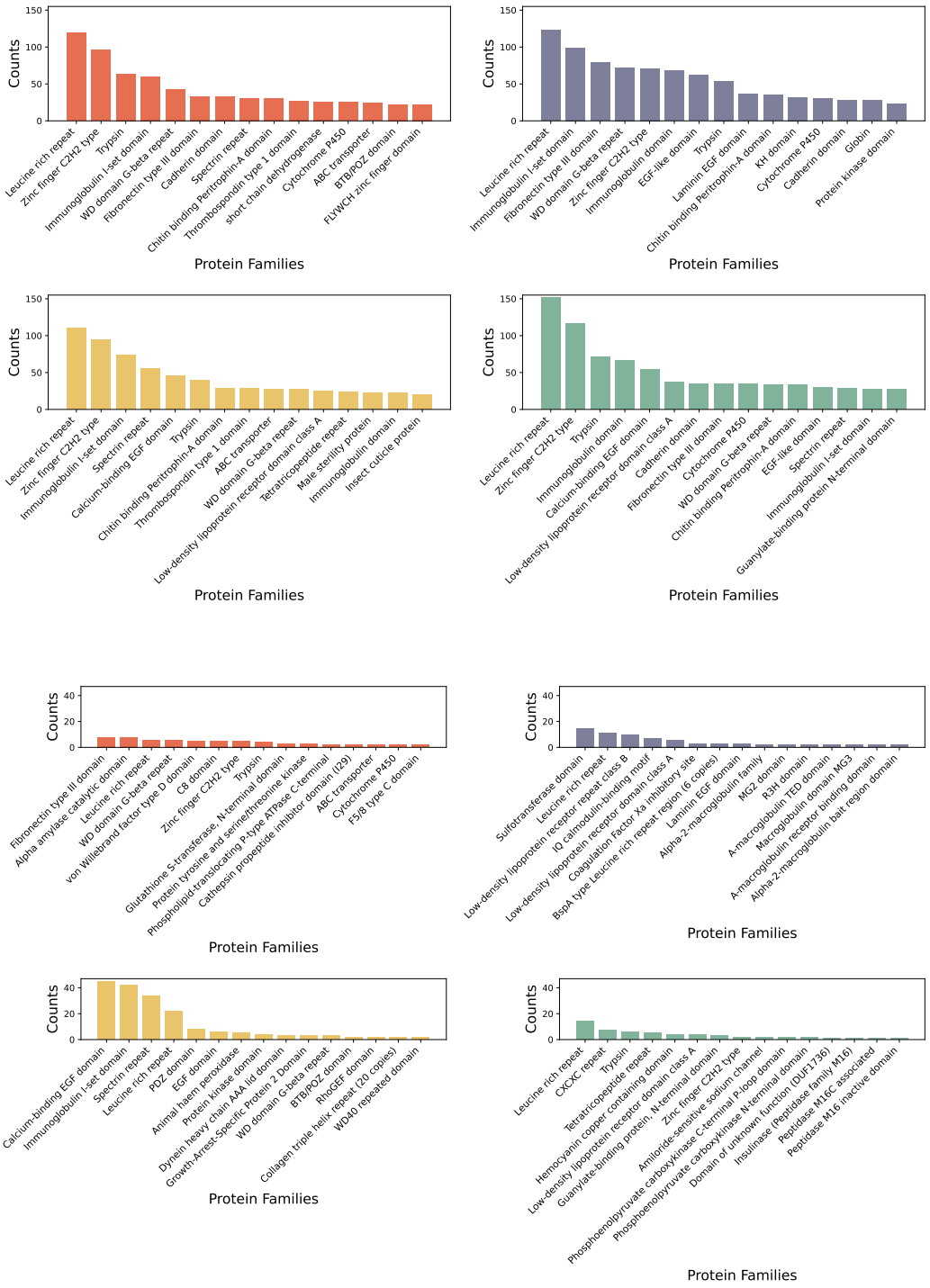

Fig. 9.7.: Top 15 annotated gene families across four replicates of *C. riparius* under artificial selection in generation 7. Figure displays the most frequently occurring gene families identified in four replicates (red, green, blue, and gold) subjected to artificial selection, identified in [a] borad-effect variants and [b] strong-effect variants in generation 7. Gene families were annotated using InterProScan with the Pfam database. The x-axis shows the gene family names, while the y-axis indicates the count (frequency) of each gene family. Gene families are sorted in descending order of frequency.

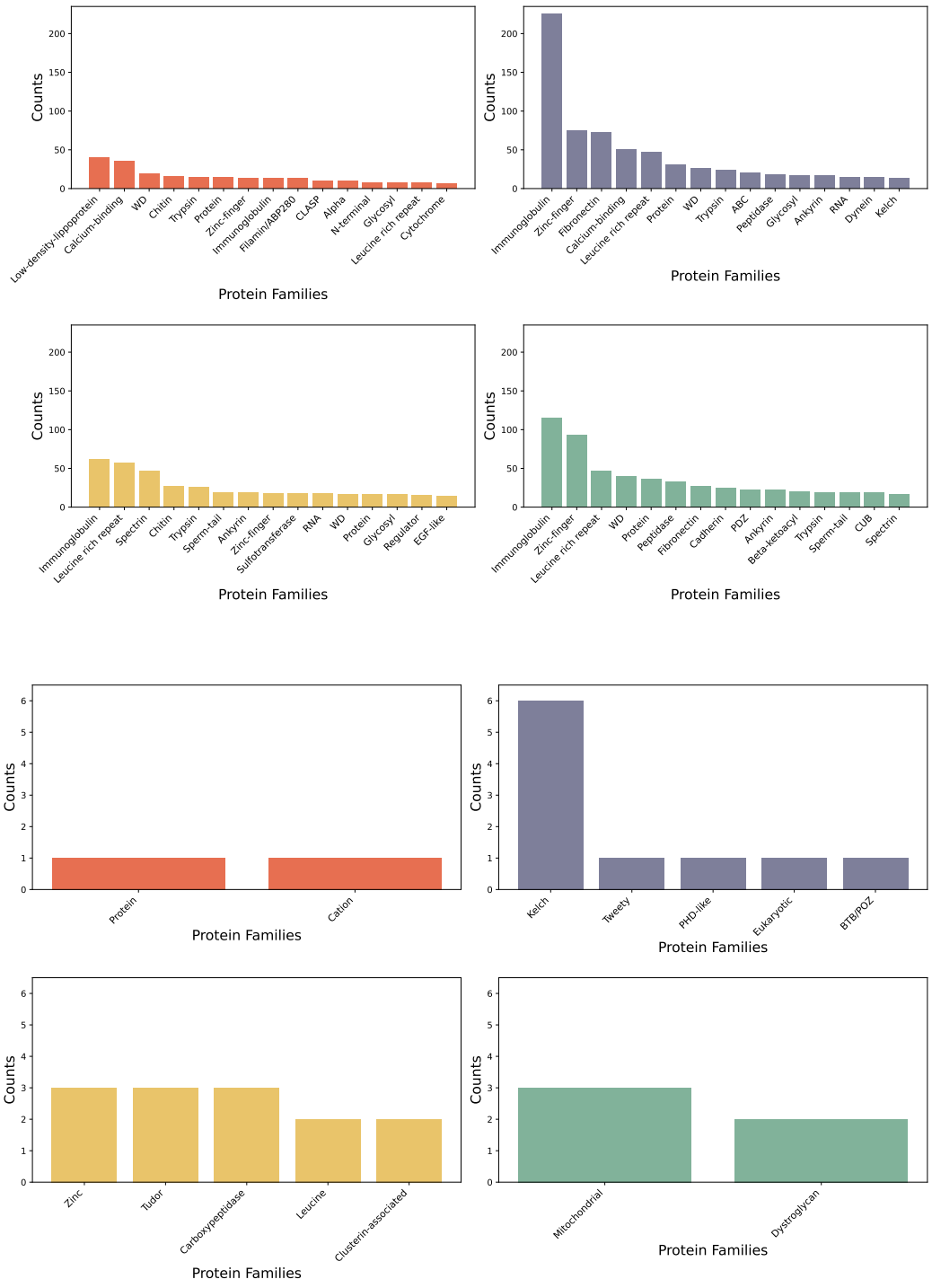

Fig. 9.8.: Top 15 annotated gene families across four replicates of *C. riparius* under artificial selection in generation 4. Figure displays the most frequently occurring gene families identified in four replicates (red, green, blue, and gold) subjected to artificial selection, identified in [a] broad-effect variants and [b]strong-effect variants in generation 4. Gene families were annotated using InterProScan with the Pfam database. The x-axis shows the gene family names, while the y-axis indicates the count (frequency) of each gene family. Gene families are sorted in descending order of frequency.

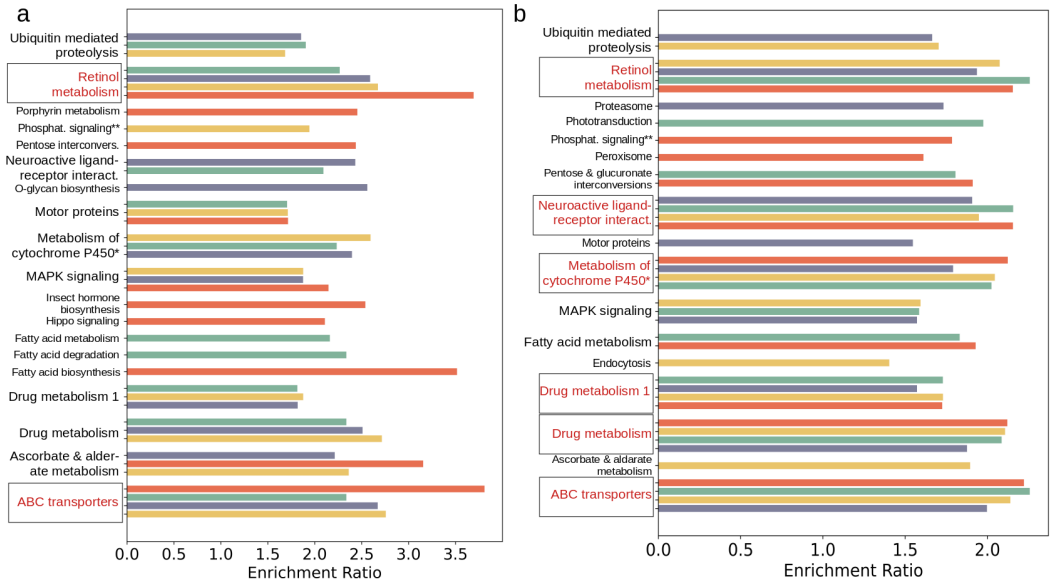

Fig. 9.9.: Enriched KEGG Pathways in braod-effect variants across selection generations. Enriched KEGG pathways ($FDR < 0.01$) identified in (a) generation 4 and (b) generation 7 across four replicates (red, green, blue, gold) under artificial selection. Enrichment ratio (x-axis) indicates pathway overrepresentation relative to background. Pathways present in all replicates are highlighted in red boxes. Pathway annotation performed using WebGeStalt.

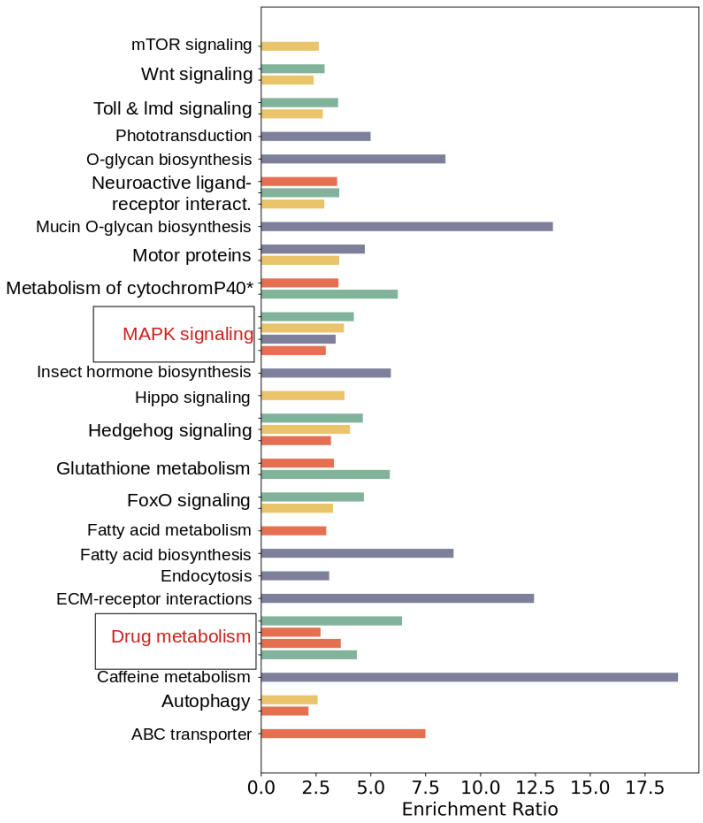

Fig. 9.10.: KEGG pathway enrichment in strong-effect variants. Enriched KEGG pathways ($FDR < 0.01$) identified in strong-effect variants from generation 7 across four replicates (red, green, blue, gold). Enrichment ratio (x-axis) shows pathway over-representation relative to background. Red boxes indicate pathways present in all replicates. Analysis performed using WebGeStalt. No pathways were ernichred using the the FDR threshold (< 0.01).

Part V

Discussion

This study provides comprehensive insights into rapid polygenic adaptation through three complementary approaches: (1) the development and validation of a novel computational method combining machine learning with traditional statistics for detecting polygenic adaptation, (2) an experimental evolution study tracking phenotypic changes in *C. riparius* under selection for accelerated development, and (3) the application of novel computational method to experimental data, revealing complex patterns of convergent and divergent adaptation. Together, these approaches illuminate the mechanisms and complexity of rapid adaptation at both phenotypic and genetic levels.

9.4 Methodological Advances in Detecting Polygenic Adaptation

This study introduced and tested novel approaches to investigate polygenic adaptation at the genetic level by combining standard statistical methods with novel machine learning techniques. The 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, this approach strikes 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 [25]. Importantly, this study underscores the crucial role of parameter optimization in classification algorithms, highlighting both the applicability of this 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 Analysis and Temporal Dynamics of Adaptation

The integrated OCSVM-FET methodology demonstrated superior performance compared to standalone Fisher's Exact Test across various experimental conditions. The observation of peak accuracy and AUC values at 250 selected loci suggests an optimal equilibrium between signal intensity and background noise in the simulated datasets. This observation emphasizes the necessity of considering the magnitude of polygenic adaptation when selecting detection methodologies. Simulation analyses revealed that OCSVM exhibits particularly robust detection capabilities in highly polygenic architectures.

The scenarios examining 100 and 250 loci demonstrated notable phenotypic progression, characterized by consistent increases between generations 20 and 40. These patterns indicate robust selection signals across multiple loci, generating clear patterns detectable by the OCSVM methodology. The 40th generation represents a period of significant adaptive transformation, resulting in particularly pronounced and detectable selection signals, not yet influenced by genetic drift. Scenarios with fewer loci (10 or 50) potentially generate selection signals that lack the complexity necessary for optimal OCSVM detection in polygenic traits. These cases exhibit rapid initial adaptation followed by quick stabilization, potentially resulting in insufficient or homogeneous selection signals by generation 40. Conversely, scenarios with very high loci numbers (500) may produce overly diffuse or noisy adaptive signals, demonstrating gradual increases that might not generate sufficiently strong signals at generation 40 for optimal detection. This temporal point provides adequate genetic complexity to represent realistic polygenic traits while maintaining clear adaptive signals [39]. This aligns with current understanding in quantitative genetics regarding the influence of hundreds of loci on adaptive traits, rather than either a few major-effect loci or thousands of minor-effect loci [104]. For example, research in human genetics has identified hundreds of loci associated with height [153] and other quantitative characteristics, though the precise number of loci involved in adaptation varies significantly across traits and organisms [154]. Furthermore, OCSVM-FET exhibited robust performance across diverse simulation parameters. Even in scenarios where performance showed slight degradation (such as at generation 60 or with loci counts exceeding 250), OCSVM-FET maintained superiority over conventional methodologies. This consistent performance indicates its utility in detecting ongoing selection across various polygenic adaptation scenarios. It is important to note that the observed peak performance partially results from model optimization, likely reflecting patterns within the optimization data set. This phenomenon is well-documented in unsupervised machine learning applications within genomic contexts [155]. Consequently, empirical data was utilized for model optimization to accurately reflect real-world scenarios. While the true architecture of polygenic adaptation patterns remains complex, these findings provide enhanced understanding of the genomic scale at which such patterns are most readily detectable, offering valuable tools for identifying candidate SNPs in comparable scenarios.

Trade-offs between methods

The comparative analysis revealed distinct performance variations among the tested methodologies. Traditional Fisher's Exact Test demonstrated the highest false positive rate, lowest AUC, and generally inferior accuracy compared to alternative approaches. This suggests that more sophisticated methodologies such as OCSVM and NBC are better suited for detecting polygenic adaptation patterns. The integration of these methods with FET further enhanced performance metrics, with OCSVM-FET exhibiting optimal overall performance, characterized by consistently low false positive rates and superior accuracy and AUC under specific conditions. OCSVM performance demonstrates significant parameter sensitivity, requiring careful optimization for each specific application scenario. Optimal solutions can be identified through grid searches using comparable known positive scenarios [156]. A key challenge lies in achieving an appropriate balance between model complexity and generalization capacity, a common consideration in machine learning applications to genomic data [25]. Moreover, OCSVM exhibits considerable computational intensity, with performance heavily dependent on parameter configurations and data architecture. The method's time complexity ranges from O(n^2) to O(n^3), where n represents the training sample count. This characteristic makes prediction of computational requirements challenging, particularly for extensive genomic data sets. In contrast, NBC presents relatively modest computational requirements and can be executed on standard computing systems due to its efficient probability density function calculations. NBC typically exhibits O(nd) time complexity, where n represents training sample count and d represents feature count, enabling superior scalability for large data sets. While grid searches can optimize parameters for both methodologies, NBC requires deeper understanding of underlying calculations. A significant challenge in NBC parameter optimization involves maintaining valid positive semi-definite covariance matrices, a mathematical requirement essential for methodological validity [34]. This constraint may necessitate consideration of alternative fine-tuning approaches for NBC. In resource-limited computational environments, investing time in careful NBC parameter optimization may prove beneficial. However, given sufficient computational resources, OCSVM represents the preferred methodology due to its superior overall performance and more straightforward parameter optimization procedures.

Parameter Optimization Significance

This study underscores the crucial role of parameter optimization in achieving optimal performance with machine learning algorithms, particularly in complex biological applications such as polygenic adaptation detection. Parameter optimization represents a critical prerequisite for both OCSVM and NBC methodologies to achieve optimal model performance. In the OCSVM context, parameter selection presents unique challenges due to its unsupervised nature. As noted by Amer et al. [157], OCSVM performance demonstrates high sensitivity to parameter selection, particularly regarding kernel function and regularization parameters. This sensitivity becomes particularly pronounced when dealing with high-dimensional biological data, where balancing model complexity and generalization capability is crucial [33]. The study employed empirical data for optimization to capture authentic genetic variation and selection patterns that might not be fully replicable in simulations. Libbrecht and Noble [46] emphasize that genomic applications of machine learning models heavily depend on their capacity to capture specific biological data patterns and complexities. The approach of utilizing empirical data from *C. riparius* for initial parameter optimization aligns with this principle, ensuring model calibration to biologically relevant polygenic adaptation patterns. This approach, however, presents specific challenges. The application of machine learning in bioinformatics requires careful balance between leveraging domain-specific knowledge and avoiding dataset-specific overfitting [155]. The methodology, optimized using empirical data, demonstrated enhanced performance on similar simulated scenarios, indicating its potential for detecting polygenic adaptation patterns in comparable evolutionary contexts. Future research directions could explore more sophisticated parameter optimization strategies. For instance, Chapelle et al. [158] propose automated multi-parameter optimization methods for SVMs, which could be adapted for OCSVM in polygenic adaptation detection contexts. Additionally, incorporating techniques to enhance model robustness and transferability across diverse evolutionary scenarios could further improve the applicability of machine learning methods in various biological contexts.

In conclusion, while this investigation demonstrates the effectiveness of well-optimized machine learning approaches in detecting polygenic adaptation, it also highlights the ongoing challenge of balancing specificity and generalizability in parameter optimization. This balance remains crucial for developing methodologies that are both powerful and broadly applicable in studying complex evolutionary processes.

9.5 Rapid Phenotypic Adaptation Under Experimental Evolution

The experimental evolution study with *C. riparius* revealed adaptive potential, demonstrating the species' capacity for rapid evolutionary response under strong selective pressure. Within 7 generations, 4 replicate populations achieved a substantial reduction in development time of approximately two days, highlighting the speed and magnitude of possible adaptation.

Temporal Dynamics of Adaptation

The temporal dynamics of adaptation in *C. riparius* revealed a complex and non-linear trajectory, providing valuable insights into how organisms navigate fitness optimization under strong selective pressures. Initially, populations experienced reduced fitness during the first two generations, demonstrated through decreased survival rates and increased emergence times. Such temporary fitness decreases during early experimental generations are well-documented in multi-generational studies [159, 160] and likely reflect the initial stress of adapting to laboratory conditions.

A pivotal turning point occurred in the fourth generation, where populations achieved their lowest mean emergence values of the experiment. Interestingly, after this point, emergence times stabilized despite continued selection pressure. This apparent plateau effect initially raised questions about the limits of adaptive potential in development time. However, further investigation revealed a more nuanced picture: while emergence times stabilized, population fitness continued to improve through enhanced adult survival and early fertility, suggesting a shift in the focus of adaptation rather than a true evolutionary standstill. The stabilization of emergence times despite ongoing selection pressure pointed to potential environmental constraints on further adaptation. One crucial factor appeared to be the synchronizing effect of photoperiod, which potentially masked continued selection on developmental time by imposing developmental pauses during dark periods [161]. This hypothesis was confirmed through a follow-up experiment in generation eight, where exposure to constant light conditions revealed previously masked adaptive potential [139]. The removal of the standard 16:8 light-dark cycle led to striking divergent responses among replicates, unveiling at least two distinct adaptive strategies.

The red and blue replicates demonstrated one possible adaptive strategy, achieving further decreases in mean emergence time and shifting from a typical bell-shaped emergence distribution to a waveform pattern with peak emergence on the second day. However, this accelerated development came at a clear cost, manifesting in consistently lower survival rates. In contrast, the gold and green replicates exhibited an alternative strategy, maintaining their bell-shaped emergence pattern with peak emergence on the third day while demonstrating higher survival rates. These divergent responses provide compelling evidence for the evolution of distinct circadian phenotypes, each representing different trade-offs between development speed and survival. This trade-off between rapid development and survival aligns well with fundamental life-history theory, which predicts that faster development often comes at the cost of other fitness components [162]. Similar trade-offs have been observed in other insects, such as *Drosophila*, where selection for rapid development has been associated with decreased adult size and fecundity [163]. The emergence of these distinct phenotypes demonstrates that the original photoperiod regime may have masked underlying genetic variation in development time, potentially by imposing synchronized developmental pauses during dark periods [161].

Replicate-Specific Adaptation Trajectories

Despite experiencing identical selection pressures, the four experimental replicates developed markedly different characteristics over the course of selection. This divergence was particularly evident in the contrasting patterns between the red replicate, which consistently achieved the fastest emergence times, and the green replicate, which demonstrated superior survival rates. Such distinct performance profiles suggest the existence of multiple viable evolutionary strategies, each representing different solutions to the same selective challenge [164, 165].

The emergence of these alternative adaptive strategies aligns with theoretical predictions about the structure of adaptive landscapes. The classical adaptive landscape theory posits that evolutionary dynamics can be visualized as a landscape where peaks represent high fitness and valleys represent low fitness. This framework, initially proposed by Sewall Wright, suggests that populations navigate this landscape through mutations and natural selection, adapting to their environments over time [166]. Rather than converging on a single optimal solution, populations appear to have discovered different peaks in the adaptive landscape, each representing a unique combination of trait

values that confer similar overall fitness benefits. This phenomenon of parallel evolution leading to different adaptive outcomes has been documented in various other experimental systems, from bacteria to vertebrates. A review by Carroll et al. [167] therefore discusses multiple case studies, illustrating how parallel evolution can manifest in diverse contexts

The case of the gold replicate provides particularly interesting insights into the factors that can constrain adaptive potential. Despite experiencing identical selection pressures, this replicate consistently showed poorer performance compared to others across multiple fitness parameters. This pattern might reflect the consequences of genetic drift or founder effects during the establishment of replicates [168]. If the initial sampling resulted in a less favorable combination of alleles or reduced genetic variation in the gold replicate, this could have limited its capacity to respond to selection effectively. Alternatively, this replicate might have become trapped on a suboptimal fitness peak, unable to access more favorable genetic combinations due to epistatic constraints or the necessity of crossing fitness valleys [169].

The divergent evolutionary trajectories observed among replicates also have important implications for understanding adaptation in natural populations. They demonstrate that even under controlled laboratory conditions with strong directional selection, populations can evolve different solutions to the same selective challenges. This suggests that natural populations, facing more complex and variable selective pressures, might harbor even greater potential for diverse adaptive strategies. Furthermore, these findings emphasize the importance of maintaining multiple populations in conservation efforts, as different populations might preserve distinct adaptive potential even when seemingly redundant [170].

Fitness Component Integration

The integration and sequential improvement of different fitness components throughout the selection experiment revealed intriguing patterns about the organization and evolution of life-history traits in *C. riparius*. Observations suggest a hierarchical structure in how different fitness components respond to selection, with emergence time showing priority in the adaptive response. This temporal sequence of adaptation, where development time optimization preceded improvements in survival and fertility, provides valuable insights into the architecture of life-history evolution. Thus, life-history theory is a framework that seeks to explain how natural selection and other evolution-

ary forces shape organisms to optimize their survival and reproduction in response to environmental challenges [171].

The prioritization of emergence time as an early target of adaptation suggests its potential role as a 'master trait' in *C. riparius* life history. The observed reduction in development time from 21.4 to 19.9 days, representing a 7% improvement, likely triggered cascading effects through other fitness components. Such master traits, which exert broad influence over multiple aspects of an organism's life history, have been documented in insect systems [172]. For example, development time has been shown to influence adult body size, reproductive capacity, and lifetime fitness in numerous insects species [173]. The hierarchical nature of the adaptive response may reflect several underlying mechanisms. First, the genetic architecture of these traits likely plays a crucial role, with emergence time potentially having a simpler genetic basis allowing for more rapid evolution [174]. Second, varying degrees of evolutionary constraint may affect different traits, with some components being more constrained by physiological or developmental requirements than others [175]. Finally, the relative contributions of different traits to overall fitness under laboratory conditions may have shaped the sequence of adaptation, with emergence time potentially having the most immediate impact on reproductive success in this experimental setting.

The integration of these various fitness components is particularly well illustrated by the population growth rate (PGR) patterns observed across generations. The PGR served as a comprehensive metric of adaptation, capturing the combined effects of multiple fitness parameters and their interactions, as survival, fertility and mean emergence. The highly significant differences between replicates across all generations ($p < 4e - 05$) demonstrate how sensitive this integrated measure is to subtle variations in adaptive strategies. Moreover, the observed two-phase improvement pattern in PGR may provides evidence for the temporal structure of adaptation, where different fitness components contribute to overall population performance at different stages. This temporal pattern of adaptation is consistent with findings in other biological systems, such as in subalpine plants, where researchers observed similar multi-year variations in population growth rates that reflected different phases of adaptive response [176]. Perhaps the most remarkable phenotypic adaptation was the improvement in adult survival rates, which more than doubled from 40% to 88% over the course of the experiment. These demographic changes likely have significant consequences for population dynamics and evolutionary potential. First, the greatly improved survival rates

contribute to enhanced population stability and resilience [177], providing a larger effective population size that better buffers against environmental fluctuations [178]. Second, the maintenance of a larger breeding population potentially supports the retention of greater genetic diversity, which is crucial for continued adaptive potential [179]. Finally, the combination of improved survival and enhanced fertility creates positive feedback loops in the demographic structure, potentially accelerating future adaptive responses, as an experimental evolution studies with spider mites showed [180].

9.6 Complex Patterns of Convergent and Divergent Adaptation

The combined analysis of phenotypic and genetic changes revealed a complex adaptation process. The initial rapid decrease in emergence time (until generation 4) corresponded to selection on broadly shared regulatory elements and structural proteins. The subsequent plateau phase, rather than indicating a halt in adaptation, represented a period of genetic refinement where populations explored different optimization strategies, as evidenced by moderate increasing pairwise F_{ST} values and divergent gene family patterns. This demonstrates how polygenic adaptation can proceed through an initial targeting of major effect genes followed by population-specific fine-tuning, ultimately achieving similar phenotypic outcomes through different genetic architectures.

Genetic diversity patterns in broad-effect and strong-effect variants

The analysis of genetic diversity patterns across replicate populations revealed complex dynamics during rapid adaptation to the applied selection regime on emergence time. While using the terms 'broad-effect' and 'strong-effect' variants throughout this study, these likely represent different temporal patterns of selection response. Strong-effect variants may represent primary targets of selection showing immediate frequency changes, while broad-effect variants could include both secondary targets and linked variants that change frequency more gradually over time. The broad-effect variants exhibited slightly elevated genetic diversity (both π and θ) compared to genome-wide patterns in both generations 1 and 7, suggesting adaptation primarily occurred through standing genetic variation, where multiple alleles of small effect contribute to the phenotype [7]. This maintenance of diversity, even under strong selection,

indicates that multiple adaptive variants can persist within populations while still achieving directional selection for earlier emergence time. Interestingly, Tajima's D patterns showed distinct trajectories among replicates. Red and gold replicate populations exhibited decreasing values from generation 1 to 7, while replicates blue and green showed increased D values, though remaining positive. This divergent patter is consistent with genetic redundancy in pathway functions. The persistent positive Tajima's D values in blue and green replicates might indicate the maintenance of multiple allelic variants at intermediate frequencies [181]. However, it is important to note that positive Tajima's D values can also result from recent population contractions or moderate bottlenecks [182], which could be a consequence of the experimental design. In contrast, the strong-effect variants showed reduced nucleotide diversity (both π and θ) in generation 7 compared to genome-wide patterns in three of four replicates, with gold showing a distinct pattern. This reduction in diversity suggests regions under stronger selection, potentially representing crucial regulatory or functional elements where specific key regulation variants were favored [183]. Lower diversity suggests these variants are more critical and therefore less redundant.

The disparity between broad-effect and strong-effect variant patterns suggested a hierarchical model of adaptation, where some genomic regions maintain flexibility through standing variation while others undergo more targeted selection [23]. For instance, research indicates that standing genetic variation facilitates rapid adaptation to environmental changes, such as ocean acidification, by providing a reservoir of pre-existing alleles that can be selected for under new conditions [184]. Investigations into songbird populations e.g. revealed that pre-existing genetic variants were predominant in local adaptation, emphasizing the importance of standing variation in shaping evolutionary potential [185]. The balance between maintenance of variation and selective optimization was shown to be crucial for successful rapid adaptation, allowing populations to both respond to immediate selective pressures and maintain adaptive potential for future challenges [105, 186].

However, it is important to consider alternative explanations for these observations. The maintenance of diversity could be partly due to linkage disequilibrium causing neutral variants to hitchhike with selected alleles [187]. Further research, including analysis of linkage disequilibrium and more detailed temporal sampling, could help disentangle these potential confounding factors from the adaptive processes we aim to study.

Convergent Phenotypes Through Genetic Redundancy

The average pairwise F_{ST} value between the populations of ancestral generation and last generation ranged from around 0.03 to around 0.06, suggesting a low to moderate genetic differentiation during adaptation [188]. This genetic divergence, occurring alongside phenotypic convergence, reveals how populations can traverse distinct regions of the adaptive landscape while achieving similar functional outcomes. While all replicates originated from the same source population, random sampling effects during replicate establishment could have created subtle but important differences in their starting genetic composition. This variation in genetic starting points could have predisposed different replicates toward particular adaptive strategies. Stochastic effects during early generations likely played a crucial role in determining subsequent evolutionary trajectories. Early random fluctuations in allele frequencies, particularly during the initial periods of adaptation stress, might have pushed replicates onto different evolutionary paths [189, 190, 191]. This concept of historical contingency, where early events can have lasting effects on evolutionary outcomes, has been demonstrated in various experimental evolution studies [192, 193, 194]. The substantial overlap in affected gene families, despite the minimal number of shared specific loci, indicates that adaptation may be more predictable at the functional level than at the genetic level [195].

Furthermore, the accumulation of genetic differences between replicates while maintaining similar phenotypic trajectories indicates that rapid adaptation does not necessarily constrain future evolutionary potential. For instance, a study on the Australian wildflower *Senecio lautus* found that replicate populations evolved along the same phenotypic trajectory despite differing genetic changes, indicating that rapid adaptation does not necessarily limit future evolutionary potential. This suggests that maintaining pathway functionality may be more critical than preserving specific genetic variants, which has implications for conservation strategies [196]. These results emphasize the importance of examining adaptation across multiple biological levels of organization.

Sequential Optimization: A two-phase model of adaptation

In this work, a distinct two-phase pattern in emergence time adaptation was investigated, with corresponding signatures at both phenotypic and genetic levels. The data suggests an adaptive model where initial rapid phenotypic

changes operate through broad metabolic adjustments, followed by a refinement phase characterized by regulatory fine-tuning. The initial rapid decrease in emergence time through generation 4 was characterized by selection on broadly shared elements, followed by a plateau phase that appeared to represent a period of genetic refinement rather than adaptive stasis. This temporal pattern provides new insights into how populations optimize complex life-history traits under strong selection.

During this initial phase, adaptation was primarily associated with broad metabolic pathways, particularly retinol metabolism and ABC transporters. This pattern of initial metabolic adaptation has been observed in multiple systems, from bacterial antibiotic resistance to thermal adaptation in fish [197, 198]. The absence of enriched pathways in strong-effect variants during this phase supports theoretical predictions that early adaptation often operates through subtle changes across many loci rather than strong selection on specific genes [199, 200].

The subsequent phase (generations 4-7) revealed a shift from metabolic to signaling pathway optimization, suggesting a transition from general physiological adjustment to fine-tuned regulatory control. Research indicates that metabolic enzyme loci can influence signaling pathways, suggesting that initial broad metabolic changes can lead to more refined regulatory control over time [201, 202]. The convergence of replicates on shared pathways while maintaining distinct genetic architectures suggests that the two-phase pattern could represent a general feature of rapid adaptation rather than a population-specific response. However, it is important to note that the adaptive architecture may depend more on the ancestry of the founder population than the specific selection regime, as demonstrated by Otte et al. [203]. This two-phase model has important implications for understanding rapid adaptation in natural populations. The continued improvement in population growth rate during phenotypic plateau suggests that apparent stasis can mask ongoing genetic optimization, a phenomenon also observed in long-term evolution experiments, particularly with *Escherichia coli* [204, 205, 206]. In such studies, researchers observed sustained improvements in fitness, often measured as growth rates, across tens of thousands of generations, even when visible phenotypic changes appeared to have plateaued. These enhancements are suggested to result from fine-tuning genetic adaptations [207]. These findings highlight the importance of considering both phenotypic and genetic changes when studying adaptation trajectories.

Molecular pathways in emergence time regulation: from metabolism to signalling pathways

The pathway analysis revealed an intricate network of molecular mechanisms involved in emergence time adaptation. The prominent enrichment of retinol metabolism pathways in both generation 4 and 7 across all replicates suggests a fundamental role in emergence time adaptation. Retinoids are crucial regulators of developmental timing in insects, influencing both metamorphosis and emergence through their effects on hormone synthesis and molecular signalling [208]. Studies in mammals furthermore showed the precise regulation of vitamin A metabolism and its conversion to retinoic acid is critical for development, requiring coordinated action of multiple enzymes and regulatory pathways [209].

This analysis also revealed enrichment of ABC transporters across both generations, highlighting their consistent role in hormonal regulation during development. These transporters mediate ecdysone and juvenile hormone (JH) transport, with ecdysone promoting metamorphosis while JH acts as its antagonist. JH regulates developmental timing through Kr-h1 expression, which suppresses steroidogenic enzyme transcription in the prothoracic gland, thereby modulating ecdysone levels [210]. This hormone interplay also influences body size and growth duration through ecdysone synthesis control [211], demonstrating the integrated nature of hormonal regulation in insect development and metamorphosis [188].

By generation 7, additional pathways emerged, indicating further refinement of the adaptive response. The cytochrome P450 pathways, enriched in all replicates of generation 7, may indicate a regulatory network similar to that seen in mammals for RA metabolism [209]. Additionally, previous studies in Drosophila melanogaster have shown that P450 genes are crucial for proper developmental timing, with variations in expression affecting pupation and emergence times [212].

The MAPK signalling pathway, enriched in all replicates' strong-effect datasets of generation 7, is known to integrate environmental signals with developmental timing in insects [213, 214]. Similarly, mTOR signalling, found in some replicates, couples nutritional status with developmental progression, potentially optimizing emergence timing in response to resource availability [215] and was shown to be involved in regulating developmental processes in honey bees [216]. The WNT and hedgehog signaling pathways, also identified in later generations, are fundamental regulators of insect development, where

for example Wnt/β-catenin signaling is required for posterior elongation in insects, indicating its fundamental role in development [217] and Hedgehog signaling was shown to be essential for growth and pattern formation during wing development in insects [218].

The enrichment of neuroactive ligand-receptor interaction pathways suggests adaptation in neuronal signaling, which could influence emergence behavior. In insects, the timing of emergence is partly controlled by circadian rhythms and environmental sensing, processes that, for sintance, rely heavily on neurotransmitter signaling [219]. Additionally, Butt [220] showed its specific involvement in regulating insect behaviors and physiological responses, particularly those related to circadian rhythms. The involvement of FoxO signaling is particularly interesting, as this pathway acts as a master regulator integrating insulin signaling and juvenile hormone pathways [221]. Through its regulation of juvenile hormone degradation, FoxO plays a critical role in controlling growth and development in insects [222], potentially allowing for more robust emergence timing under varying conditions.

These molecular changes suggest adaptation occurred through optimization of both housekeeping metabolism and its underlying regulatory networks. The involvement of drug metabolism pathways might reflect broader changes in xenobiotic handling capacity, potentially enhancing stress tolerance during development [223]. This multilevel regulation could explain the continued improvement in population growth rate even after emergence time stabilized [224], as these pathways also influence general fitness traits beyond developmental timing [225].

9.7 Limitations and Future Directions

Several limitations and opportunities for future research emerge from this study. For instance, relatively short experimental duration of seven generations may have captured only initial adaptive responses. Longer-term studies could reveal whether the observed plateaus in development time represent stable endpoints or temporary stasis [226]. Additionally, Pool-Sequencing Data itself is a limiting factor, as it is not displaying the genomic insight on an individual's level. Therefore, haplotype block identification as well as recombination rate analysis is not yet easy to apply on Pool-Seq data and is still in the stage of development.

The here presented OCSVM-FET method, while effective, requires careful parameter tuning. Future work should explore automated parameter optimization approaches [158] to enhance the method's accessibility and reproducibility. The method's validation across diverse empirical data sets remains crucial, particularly given that parameter optimization in NBC significantly impacts results - as shown in yeast studies where reducing from 12 to 5 replicate populations eliminated detection of candidate regions [48]. The 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. 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 the OCSVM-based method represents a significant advance in detecting polygenic adaptation, there remains room for refinement and expansion. The dual-threshold approach was designed to capture different aspects of selection. The broad-effect variant threshold ($p < 0.001$) was specifically chosen to detect subtle allele frequency changes that might be biologically relevant in polygenic adaptation, complementing a stringent threshold analysis that focused on the strongest selection signals. Comparison with neutral simulations suggests that future studies might benefit from using a more conservative threshold (>99%) for the broad-effect variants to more definitively distinguish

selection from drift. Furthermore, the incorporation of linkage disequilibrium patterns and haplotype structure analysis would help distinguish primary selection targets from hitchhiking variants. Additionally, the recent completion of an improved reference genome for *C. riparius* offers opportunities to refine these findings. While the current analysis provides valuable insights into the temporal dynamics of adaptation, these additional analyses would further clarify the mechanisms of polygenic selection. Additionally, annotation databases beyond PFAM, and investigation of extended genomic windows around candidate positions would provide more comprehensive insights, since regulatory elements can be in distance of 100kb to the selected locus [227].

Future research should particularly investigate biological mechanisms underlying rapid adaptation. The role of epigenetic mechanisms deserves special attention [228], as recent work suggests their importance in facilitating rapid adaptation. The potential role of epistatic interactions in shaping these divergent adaptive pathways warrants particular attention. Complex interactions between genes could create different local fitness optima, making certain evolutionary trajectories more accessible depending on which combinations of alleles became established early in the selection process. For instance, research has shown that epistatic interactions can significantly affect the fitness effects of mutations, thereby shaping the paths available for adaptation [229]. Similarly, understanding the molecular basis of photoperiod-development time interactions [230] and temporal dynamics of gene expression (RNA-Sequencing data) changes during adaptation would provide valuable insights into adaptation mechanisms. Method development should explore the approach's performance across different genetic architectures and selection intensities. 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. Implementation of advanced parameter tuning techniques, such as Bayesian optimization [49], could enhance both OCSVM and NBC implementations. Additionally, developing complementary methods for functional characterization of identified loci would strengthen biological interpretations of results.

9.8 Conclusions

This study introduces a powerful integrated approach to understanding rapid polygenic adaptation, combining experimental evolution, novel methodological development, and comprehensive genomic analysis. The development and successful application of the OCSVM-FET method addresses a critical gap in detecting subtle polygenic selection signatures in pool-sequencing data, enabling researchers to capture both strong and distributed selection effects that traditional methods often miss.

The findings of this work not only provide empirical support for key theoretical predictions in evolutionary biology but also reveal previously unrecognized complexities in rapid adaptation processes. The discovery of a two-phase adaptation pattern offers strong empirical evidence for theoretical models predicting an initial rapid response through large-effect variants followed by slower optimization through smaller-effect variants [199]. However, this results challenge simplistic views of adaptation by demonstrating that this process involves replicate-specific strong-effect variants operating alongside shared pathway-level responses. This suggests a more sophisticated model of adaptation where selection acts simultaneously across multiple scales of genetic organization - from individual variants to broader regulatory networks - rather than proceeding through a simple sequential progression. This pattern particularly validates recent theoretical frameworks emphasizing genetic redundancy and pathway-level organization in rapid evolutionary responses [23].

The maintenance of genetic diversity in broad-effect variants, even under strong directional selection, provides compelling empirical evidence for theoretical models that predict the preservation of standing genetic variation during polygenic adaptation [186]. This finding directly challenges classical selective sweep models while supporting modern theoretical frameworks incorporating genetic redundancy and multivariate selection. Furthermore, the continued improvement in population growth rate despite phenotypic stabilization strongly supports theoretical predictions about the importance of genetic variation and compensatory evolution in adaptive processes [231].

This finding challenges classical selective sweep models while supporting more recent theoretical frameworks that incorporate genetic redundancy and multivariate selection. Furthermore, the continued improvement in population growth rate despite phenotypic stabilization supports theoretical

predictions about the importance of cryptic genetic variation and compensatory evolution in adaptive processes [231].

Perhaps most significantly, the observation of pathway-level convergence despite divergent genetic trajectories addresses fundamental questions about evolutionary predictability. While specific genetic changes appear stochastic, the functional convergence at the pathway level suggests that certain aspects of adaptation may be more deterministic than previously recognized, particularly at higher levels of biological organization. This discovery has profound implications for understanding evolutionary processes across different scales and timeframes.

These findings have important applications for predicting evolutionary responses in rapidly changing environments, understanding the genetic basis of complex adaptive traits, and potentially informing conservation strategies for populations facing environmental challenges. Future research should expand this approach to more complex traits and diverse organisms, explore the time-dependent nature of polygenic adaptation in greater detail, and investigate how pathway-level selection interfaces with environmental complexity.

Bibliography

[1]Juha Merilä and Andrew P. Hendry. “Climate change, adaptation, and phenotypic plasticity: the problem and the evidence”. In: *Evolutionary Applications* 7.1 (Jan. 2014), 1–14 (cit. on pp. 3, 30).

[2]Jonathan K. Pritchard and Anna Di Rienzo. “Adaptation – not by sweeps alone”. In: *Nature Reviews Genetics* 11.10 (Sept. 2010), 665–667 (cit. on pp. 3, 4, 36).

[3]Neda Barghi, Raymond Tobler, Viola Nolte, and Christian Schlötterer. “Drosophila simulans: A Species with Improved Resolution in Evolve and Resequence Studies”. In: *G3 Genes|Genomes|Genetics* 7.7 (July 2017), 2337–2343 (cit. on p. 3).

[4]Paul H Harvey and Mark D Pagel. *The Comparative Method in Evolutionary Biology*. Oxford University PressOxford, May 1991 (cit. on p. 3).

[5]Ary A. Hoffmann and Carla M. Sgrò. “Climate change and evolutionary adaptation”. In: *Nature* 470.7335 (Feb. 2011), 479–485 (cit. on pp. 3, 30).

[6]Steven J. Franks and Ary A. Hoffmann. “Genetics of Climate Change Adaptation”. In: *Annual Review of Genetics* 46.1 (Dec. 2012), 185–208 (cit. on p. 3).

[7]R BARRETT and D SCHLUTER. “Adaptation from standing genetic variation”. In: *Trends in Ecology and Evolution* 23.1 (Jan. 2008), 38–44 (cit. on pp. 3, 30, 115).

[8]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”. In: *Functional Ecology* 21.3 (Apr. 2007), 394–407 (cit. on p. 3).

[9]Peter R. Grant and B. Rosemary Grant. “Unpredictable Evolution in a 30-Year Study of Darwin’s Finches”. In: *Science* 296.5568 (Apr. 2002), 707–711 (cit. on p. 3).

[10]Katie Pelletier, Megan Bilodeau, Isabella Pellizzari-Delano, et al. “Polygenic architecture of adaptation to a high-altitude environment forDrosophila melanogasterwing shape and size”. In: (Feb. 2024) (cit. on pp. 3, 4).

[11]Seth M. Rudman, Sharon I. Greenblum, Subhash Rajpurohit, et al. “Direct observation of adaptive tracking on ecological time scales in Drosophila”. In: *Science* 375.6586 (Mar. 2022) (cit. on p. 3).

[12]Markus Pfenninger and Quentin Foucault. “Population Genomic Time Series Data of a Natural Population Suggests Adaptive Tracking of Fluctuating Environmental Changes”. In: *Integrative and Comparative Biology* 62.6 (June 2022), 1812–1826 (cit. on pp. 3, 64).

[13]Christian Schlötterer, Raymond Tobler, Robert Kofler, and Viola Nolte. “Sequencing pools of individuals — mining genome-wide polymorphism data without big funding”. In: *Nature Reviews Genetics* 15.11 (Sept. 2014), 749–763 (cit. on pp. 3, 9, 32, 33).

[14]Neda Barghi, Raymond Tobler, Viola Nolte, et al. “Genetic redundancy fuels polygenic adaptation in Drosophila”. In: *PLOS Biology* 17.2 (Feb. 2019). Ed. by Greg Gibson, e3000128 (cit. on pp. 4, 9, 10, 35, 36).

[15]Billie A. Gould, Yani Chen, and David B. Lowry. “Pooled ecotype sequencing reveals candidate genetic mechanisms for adaptive differentiation and reproductive isolation”. In: *Molecular Ecology* 26.1 (Nov. 2016), 163–177 (cit. on p. 4).

[16]Reid S. Brennan, James A. deMayo, Hans G. Dam, et al. “Experimental evolution reveals the synergistic genomic mechanisms of adaptation to ocean warming and acidification in a marine copepod”. In: *Proceedings of the National Academy of Sciences* 119.38 (Sept. 2022) (cit. on p. 4).

[17]Göran Arnqvist and Ahmed Sayadi. “A possible genomic footprint of polygenic adaptation on population divergence in seed beetles?” In: *Ecology and Evolution* 12.10 (Oct. 2022) (cit. on p. 4).

[18]Robert Kofler, Ram Vinay Pandey, and Christian Schlötterer. “PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq)”. In: *Bioinformatics* 27.24 (Oct. 2011), pp. 3435–3436 (cit. on pp. 4, 37, 61, 64, 65).

[19]Zhi Wei, Wei Wang, Pingzhao Hu, Gholson J. Lyon, and Hakon Hakonarson. “SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data”. In: *Nucleic Acids Research* 39.19 (Aug. 2011), e132–e132 (cit. on p. 4).

[20]Quan Chen and Fengzhu Sun. “A unified approach for allele frequency estimation, SNP detection and association studies based on pooled sequencing data using EM algorithms”. In: *BMC Genomics* 14.Suppl 1 (2013), S1 (cit. on pp. 4, 5).

[21]Gaurav Bhatia, Nick Patterson, Sriram Sankararaman, and Alkes L. Price. “Estimating and interpreting FST: The impact of rare variants”. In: *Genome Research* (July 2013) (cit. on p. 4).

[22]F Tajima. “Statistical method for testing the neutral mutation hypothesis by DNA polymorphism.” In: *Genetics* (1989) (cit. on pp. 4, 39, 65).

[23]Neda Barghi, Joachim Hermisson, and Christian Schlötterer. “Polygenic adaptation: a unifying framework to understand positive selection”. In: *Nature Reviews Genetics* 21.12 (June 2020), 769–781 (cit. on pp. 4, 116, 123).

[24]Heng Li and Richard Durbin. “Fast and accurate long-read alignment with Burrows–Wheeler transform”. In: *Bioinformatics* 26.5 (Jan. 2010), 589–595 (cit. on p. 4).

[25]Daniel R. Schrider and Andrew D. Kern. “Supervised Machine Learning for Population Genetics: A New Paradigm”. In: *Trends in Genetics* 34.4 (Apr. 2018), 301–312 (cit. on pp. 5, 107, 109).

[26]Alexander Rives, Joshua Meier, Tom Sercu, et al. “Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences”. In: *Proceedings of the National Academy of Sciences* 118.15 (Apr. 2021) (cit. on p. 5).

[27]Kyle T David and Kenneth M Halanych. “Unsupervised Deep Learning Can Identify Protein Functional Groups from Unaligned Sequences”. In: *Genome Biology and Evolution* 15.5 (May 2023). Ed. by Brian Golding (cit. on p. 5).

[28]Christof Angermueller, Tanel Pärnamaa, Leopold Parts, and Oliver Stegle. “Deep learning for computational biology”. In: *Molecular Systems Biology* 12.7 (July 2016) (cit. on p. 5).

[29]Lai Wei, Weiming Zeng, and Hong Wang. “K-means clustering with manifold”. In: *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery*. IEEE, Aug. 2010, 2095–2099 (cit. on p. 5).

[30]Gil McVean. “A Genealogical Interpretation of Principal Components Analysis”. In: *PLoS Genetics* 5.10 (Oct. 2009). Ed. by Molly Przeworski, e1000686 (cit. on p. 5).

[31]R. Saravanan and Pothula Sujatha. “A State of Art Techniques on Machine Learning Algorithms: A Perspective of Supervised Learning Approaches in Data Classification”. In: *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE, June 2018, 945–949 (cit. on p. 6).

[32]Irina Rish. “An empirical study of the naive Bayes classifier”. In: *In IJCAI 2001 workshop on empirical methods in artificial intelligence* (2001) (cit. on pp. 6, 7, 18).

[33]Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. “Estimating the Support of a High-Dimensional Distribution”. In: *Neural Computation* 13.7 (July 2001), 1443–1471 (cit. on pp. 6, 7, 20, 22, 48, 110).

[34]Indika Wickramasinghe and Harsha Kalutarage. “Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation”. In: *Soft Computing* 25.3 (Sept. 2020), 2277–2293 (cit. on pp. 6, 7, 17, 109).

[35]Sebastien Lecomte, Regis Lengelle, Cedric Richard, Francois Capman, and Bertrand Ravera. “Abnormal events detection using unsupervised One-Class SVM - Application to audio surveillance and evaluation -”. In: *2011 8th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, Aug. 2011 (cit. on pp. 6, 22).

[36]Naeem Seliya, Azadeh Abdollah Zadeh, and Taghi M. Khoshgoftaar. “A literature review on one-class classification and its potential applications in big data”. In: *Journal of Big Data* 8.1 (Sept. 2021) (cit. on pp. 6, 7, 20).

[37]Efrem Heri Budiarto, Adhistya Erna Permanasari, and Silmi Fauziati. “Unsupervised Anomaly Detection Using K-Means, Local Outlier Factor and One Class SVM”. In: *2019 5th International Conference on Science and Technology (ICST)*. IEEE, July 2019, 1–5 (cit. on p. 7).

[38]Qiong Wei and Roland L. Dunbrack. “The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics”. In: *PLoS ONE* 8.7 (July 2013). Ed. by Iddo Friedberg, e67863 (cit. on p. 7).

[39]Sean Hoban, Giorgio Bertorelle, and Oscar E. Gaggiotti. “Computer simulations: tools for population and evolutionary genetics”. In: *Nature Reviews Genetics* 13.2 (Jan. 2012), 110–122 (cit. on p. 8).

[40]Christos Vlachos and Robert Kofler. “MimicrEE2: Genome-wide forward simulations of Evolve and Resequencing studies”. In: *PLoS computational biology* 14.8 (2018), e1006413 (cit. on pp. 8, 15, 43, 46, 50).

[41]Benjamin C Haller and Philipp W Messer. “Evolutionary Modeling in SLiM 3 for Beginners”. In: *Molecular Biology and Evolution* 36.5 (Dec. 2018). Ed. by Ryan Hernandez, 1101–1109 (cit. on pp. 8, 15).

[42]Rui Zhang, Chang Liu, Kai Yuan, et al. “AdmixSim 2: a forward-time simulator for modeling complex population admixture”. In: *BMC Bioinformatics* 22.1 (Oct. 2021) (cit. on p. 8).

[43]J A Hanley and B J McNeil. “The meaning and use of the area under a receiver operating characteristic (ROC) curve.” In: *Radiology* 143.1 (Apr. 1982), 29–36 (cit. on p. 8).

[44]David Powers. “Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation”. English. In: *Journal of Machine Learning Technologies* 2.1 (2011), pp. 37–63 (cit. on p. 8).

[45]Quentin Foucault, Andreas Wieser, Ann-Marie Waldvogel, Barbara Feldmeyer, and Markus Pfenninger. “Rapid adaptation to high temperatures in Chironomus riparius”. In: *Ecology and Evolution* 8.24 (Dec. 2018), 12780–12789 (cit. on pp. 8, 31).

[46]Markus Pfenninger and Quentin Foucault. “Genomic processes underlying rapid adaptation of a natural Chironomus riparius population to unintendedly applied experimental selection pressures”. In: *Molecular Ecology* 29.3 (2020), pp. 536–548 (cit. on pp. 8, 31, 43, 49, 51).

[47]Lin Kang, Dau Dayal Aggarwal, Eugenia Rashkovetsky, Abraham B. Korol, and Pawel Michalak. “Rapid genomic changes in Drosophila melanogaster adapting to desiccation stress in an experimental evolution system”. In: *BMC Genomics* 17.1 (Mar. 2016) (cit. on p. 9).

[48]Molly K. Burke, Joseph P. Dunham, Parvin Shahrestani, et al. “Genome-wide analysis of a long-term evolution experiment with Drosophila”. In: *Nature* 467.7315 (Sept. 2010), 587–590 (cit. on pp. 9, 32, 121).

[49]Anthony Long, Gianni Liti, Andrej Luptak, and Olivier Tenaillon. “Elucidating the molecular architecture of adaptation via evolve and resequence experiments”. In: *Nature Reviews Genetics* 16.10 (Sept. 2015), 567–582 (cit. on p. 9).

[50]Molly K. Burke. “Embracing Complexity: Yeast Evolution Experiments Featuring Standing Genetic Variation”. In: *Journal of Molecular Evolution* 91.3 (Feb. 2023), 281–292 (cit. on p. 9).

[51]Vince Buffalo and Graham Coop. “Estimating the genome-wide contribution of selection to temporal allele frequency change”. In: *Proceedings of the National Academy of Sciences* 117.34 (Aug. 2020), 20672–20680 (cit. on pp. 9, 10).

[52]Manolis Lirakis, Viola Nolte, and Christian Schlötterer. “Pool-GWAS on reproductive dormancy in Drosophila simulans suggests a polygenic architecture”. In: *G3 Genes|Genomes|Genetics* 12.3 (Feb. 2022). Ed. by R Anholt (cit. on p. 9).

[53]JUNRUI LI, HAIPENG LI, MATTIAS JAKOBSSON, et al. “Joint analysis of demography and selection in population genetics: where do we stand and where could we go?” In: *Molecular Ecology* 21.1 (Oct. 2011), 28–44 (cit. on p. 9).

[54]Trudy F. C. Mackay. “Epistasis and quantitative traits: using model organisms to study gene–gene interactions”. In: *Nature Reviews Genetics* 15.1 (Dec. 2013), 22–33 (cit. on p. 10).

[55]Philippa C Griffin, Sandra B Hangartner, Alexandre Fournier-Level, and Ary A Hoffmann. “Genomic Trajectories to Desiccation Resistance: Convergence and Divergence Among Replicate Selected Drosophila Lines”. In: *Genetics* 205.2 (Feb. 2017), 871–890 (cit. on p. 10).

[56]Renaud Kaeuffer, Catherine L. Peichel, Daniel I. Bolnick, and Andrew P. Hendry. “PARALLEL AND NONPARALLEL ASPECTS OF ECOLOGICAL, PHENOTYPIC, AND GENETIC DIVERGENCE ACROSS REPLICATE POPULATION PAIRS OF LAKE AND STREAM STICKLEBACK: PARALLEL AND NON-PARALLEL EVOLUTION”. In: *Evolution* 66.2 (Sept. 2011), 402–418 (cit. on p. 10).

[57]D J Emlen, Q Szafran, L S Corley, and I Dworkin. “Insulin signaling and limb-patterning: candidate pathways for the origin and evolutionary diversification of beetle ‘horns’”. In: *Heredity* 97.3 (July 2006), 179–191 (cit. on p. 10).

[58]Daniel Dowling, Thomas Pauli, Alexander Donath, et al. “Phylogenetic Origin and Diversification of RNAi Pathway Genes in Insects”. In: *Genome Biology and Evolution* (Jan. 2017), evw281 (cit. on p. 10).

[59]R. A. Fisher. “Statistical Methods for Research Workers”. In: *Breakthroughs in Statistics*. Springer New York, 1992 (cit. on p. 16).

[60]Karim F. Hirji, Shu-Jane Tan, and Robert M. Elashoff. “A quasi-exact test for comparing two binomial proportions”. In: *Statistics in Medicine* (1991) (cit. on p. 16).

[61]Telba Z. Ironyt and Carlos A.B. Pereirat. “Exact tests for equality of two proportions: fisher v. bayes”. In: *Journal of Statistical Computation and Simulation* 1–2 (Aug. 1986) (cit. on p. 16).

[62]M.-A. C. Bind and D. B. Rubin. “When possible, report a Fisher-exact P value and display its underlying null randomization distribution”. In: *Proceedings of the National Academy of Sciences* (2020) (cit. on p. 16).

[63]Y. Benjamini and Y. Hochberg. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1995) (cit. on p. 17).

[64]P. Langley and W. Iba. “An Analysis of Probabilistic Approaches to Induction”. In: (199X). Accessed: Oct. 09, 2024 (cit. on p. 17).

[65]Ronald R. Yager. “An extension of the naive Bayesian classifier”. In: *Information Sciences* 176.5 (Mar. 2006), 577–588 (cit. on p. 18).

[66]K. Ming Leung. *Naive Bayesian Classifier*. Abstract: A statistical classifier called Naive Bayesian classifier is discussed. This classifier is based on the Bayes’ Theorem and the maximum posteriori hypothesis. The naive assumption of class conditional independence is often made to reduce the computational cost. Directory: Table of Contents, Begin Article. Copyright c© 2007 mleung@poly.edu. Last Revision Date: November 28, 2007. Polytechnic University, Department of Computer Science / Finance and Risk Engineering. 2007 (cit. on p. 18).

[68]Markus Pfenninger, Quentin Foucault, Ann-Marie Waldvogel, and Barbara Feldmeyer. “Selective effects of a short transient environmental fluctuation on a natural population”. In: *Molecular Ecology* 32.2 (Nov. 2022), pp. 335–349 (cit. on pp. 20, 26, 32, 47).

[69]Asa Ben-Hur, Cheng Soon Ong, Sören Sonnenburg, Bernhard Schölkopf, and Gunnar Rätsch. “Support Vector Machines and Kernels for Computational Biology”. In: *PLoS Computational Biology* 4.10 (Oct. 2008). Ed. by Fran Lewitter, e1000173 (cit. on p. 20).

[70]Tommaso Zoppi, Andrea Ceccarelli, and Andrea Bondavalli. “On Algorithms Selection for Unsupervised Anomaly Detection”. In: (Dec. 2018) (cit. on p. 20).

[71]Anna M Bartkowiak. “Anomaly, novelty, one-class classification: a comprehensive introduction”. In: *International Journal of Computer Information Systems and Industrial Management Applications* 3 (2011), pp. 11–11 (cit. on p. 20).

[72]*What is a good resource for understanding One Class SVM for distribution esitmation? — quora.com*. `https://www.quora.com/What-is-a-good-resource-for-understanding-One-Class-SVM-for-distribution-esitmation`. [Accessed 15-03-2024] (cit. on p. 21).

[73]Arijit Maji and Indrajit Mukherjee. “An unsupervised one-class-classifier support vector machine to simultaneously monitor location and scale of multivariate quality characteristics”. In: *International Journal of Quality and Reliability Management* 40.2 (Dec. 2021), 419–454 (cit. on p. 22).

[74]Sebastian Okser, Tapio Pahikkala, Antti Airola, et al. “Regularized Machine Learning in the Genetic Prediction of Complex Traits”. In: *PLoS Genetics* 10.11 (Nov. 2014). Ed. by Nicholas J. Schork, e1004754 (cit. on p. 26).

[75]Minglie Li, Shusen Zhou, Tong Liu, et al. “TSVM: Transfer Support Vector Machine for Predicting MPRA Validated Regulatory Variants”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 21.3 (May 2024), 472–479 (cit. on p. 26).

[76]Riyanarto Sarno, Fernandes Sinaga, and Kelly Rossa Sungkono. “Anomaly detection in business processes using process mining and fuzzy association rule learning”. In: *Journal of Big Data* 7.1 (Jan. 2020) (cit. on p. 26).

[77]C. Darwin. *On the Origin of Species by Means of Natural Selection*. London: Murray, 1859 (cit. on p. 29).

[78]Abdessalem Abdessamad, Imen Dhib, Ghada Baraket, Mustapha Ksontini, and Amel Salhi-Hannachi. “Evaluation of Phenotypic Diversity by Use of Variable Analysis Multi of Various Populations of Oak Cork in Tunisia”. In: *Open Journal of Ecology* 04.14 (2014), 861–872 (cit. on p. 29).

[79]ANDREW P. HENDRY, THOMAS J. FARRUGIA, and MICHAEL T. KINNISON. “Human influences on rates of phenotypic change in wild animal populations”. In: *Molecular Ecology* 17.1 (July 2007), 20–29 (cit. on p. 29).

[80]Steven J. Franks, Sheina Sim, and Arthur E. Weis. “Rapid evolution of flowering time by an annual plant in response to a climate fluctuation”. In: *Proceedings of the National Academy of Sciences* 104.4 (Jan. 2007), 1278–1282 (cit. on p. 30).

[81]Strenzke K. “Die systematische und oekologische Differnzierung derGattung Chironomus”. In: (1957) (cit. on p. 30).

[82]Markus Pfenninger and Carsten Nowak. “Reproductive Isolation and Ecological Niche Partition among Larvae of the Morphologically Cryptic Sister Species Chironomus riparius and C. piger”. In: *PLoS ONE* 3.5 (May 2008). Ed. by Dennis Marinus Hansen, e2157 (cit. on pp. 30, 31).

[83]Gábor Horváth, Arnold Móra, Balázs Bernáth, and György Kriska. “Polarotaxis in non-biting midges: Female chironomids are attracted to horizontally polarized light”. In: *Physiology amp; Behavior* 104.5 (Oct. 2011), 1010–1015 (cit. on p. 30).

[84]Alexandre R.R. Péry, Raphael Mons, and Jeanne Garric. “Modelling of the life cycle of Chironomus species using an energy-based model”. In: *Chemosphere* 59.2 (Apr. 2005), 247–253 (cit. on p. 30).

[85]Zerguine Karima. “Chironomidae: Biology, Ecology and Systematics”. In: *The Wonders of Diptera - Characteristics, Diversity, and Significance for the World’s Ecosystems*. IntechOpen, Sept. 2021 (cit. on p. 31).

[86]Ann-Marie Oppold, João A. M. Pedrosa, Miklós Bálint, et al. “Support for the evolutionary speed hypothesis from intraspecific population genetic data in the non-biting midgeChironomus riparius”. In: *Proceedings of the Royal Society B: Biological Sciences* 283.1825 (Feb. 2016), p. 20152413 (cit. on p. 31).

[87]Ann-Marie Waldvogel, Andreas Wieser, Tilman Schell, et al. “The genomic footprint of climate adaptation in Chironomus riparius”. In: *Molecular Ecology* 27.6 (Mar. 2018), 1439–1456 (cit. on pp. 31, 66).

[88]Quentin Foucault, Andreas Wieser, Ann-Marie Waldvogel, and Markus Pfenninger. “Establishing laboratory cultures and performing ecological and evolutionary experiments with the emerging model species Chironomus riparius”. In: *Journal of Applied Entomology* 143.5 (2019), pp. 584–592 (cit. on pp. 31, 57, 58).

[89]Amanda Callaghan, Thomas C Fisher, Albania Grosso, Graham J Holloway, and Mark Crane. “Effect of temperature and pirimiphos methyl on biochemical biomarkers in Chironomus riparius Meigen”. In: *Ecotoxicology and environmental safety* 52.2 (2002), pp. 128–133 (cit. on p. 31).

[90]Hanno Schmidt, Bastian Greshake, Barbara Feldmeyer, Thomas Hankeln, and Markus Pfenninger. “Genomic basis of ecological niche divergence among cryptic sister species of non-biting midges”. In: *BMC Genomics* 14.1 (June 2013) (cit. on p. 31).

[91]Ann-Marie Oppold, Hanno Schmidt, Marcel Rose, et al. “Chironomus riparius (Diptera) genome sequencing reveals the impact of minisatellite transposable elements on population divergence”. In: *Molecular Ecology* 26.12 (Apr. 2017), 3256–3275 (cit. on p. 31).

[92]Carsten Nowak, Christian Vogt, Markus Pfenninger, et al. “Rapid genetic erosion in pollutant-exposed experimental chironomid populations”. In: *Environmental Pollution* 157.3 (Mar. 2009), 881–886 (cit. on p. 31).

[93]João A.M. Pedrosa, Berardino Cocchiararo, Tiago Verdelhos, et al. “Population genetic structure and hybridization patterns in the cryptic sister species Chironomus riparius and Chironomus piger across differentially polluted freshwater systems”. In: *Ecotoxicology and Environmental Safety* 141 (July 2017), 280–289 (cit. on p. 31).

[94]C Vogt, A Pupp, C Nowak, et al. “Interaction between genetic diversity and temperature stress on life-cycle parameters and genetic variability in midge Chironomus riparius populations”. In: *Climate Research* (2007) (cit. on pp. 32, 58).

[95]Sabrina Nemec, Simit Patel, Carsten Nowak, and Markus Pfenninger. “Evolutionary determinants of population differences in population growth rate × habitat temperature interactions in Chironomus riparius”. In: *Oecologia* 172.2 (Nov. 2012), pp. 585–594 (cit. on pp. 32, 58).

[96]Andreas Futschik and Christian Schloetterer. “The Next Generation of Molecular Markers From Massively Parallel Sequencing of Pooled DNA Samples”. In: *Genetics* 186.1 (Sept. 2010), 207–218 (cit. on pp. 32, 33).

[97]C Schlötterer, R Kofler, E Versace, R Tobler, and S U Franssen. “Combining experimental evolution with next-generation sequencing: a powerful tool to study adaptation from standing genetic variation”. In: *Heredity* 114.5 (Oct. 2014), 431–440 (cit. on p. 32).

[98]Ian M. Ehrenreich, Noorossadat Torabi, Yue Jia, et al. “Dissection of genetically complex traits with extremely large pools of yeast segregants”. In: *Nature* 464.7291 (Apr. 2010), 1039–1042 (cit. on p. 32).

[99]Héloise Bastide, Andrea Betancourt, Viola Nolte, et al. “A Genome-Wide, Fine-Scale Map of Natural Pigmentation Variation in Drosophila melanogaster”. In: *PLoS Genetics* 9.6 (June 2013). Ed. by Patricia Wittkopp, e1003534 (cit. on p. 32).

[100]Marco Fracassetti, Philippa C. Griffin, and Yvonne Willi. “Validation of Pooled Whole-Genome Re-Sequencing in Arabidopsis lyrata”. In: *PLOS ONE* 10.10 (Oct. 2015). Ed. by Ulrich Melcher, e0140462 (cit. on p. 33).

[101]Christian Rellstab, Stefan Zoller, Andrew Tedder, Felix Gugerli, and Martin C. Fischer. “Validation of SNP Allele Frequencies Determined by Pooled Next-Generation Sequencing in Natural Populations of a Non-Model Plant Species”. In: *PLoS ONE* 8.11 (Nov. 2013). Ed. by Gabriel AB Marais, e80422 (cit. on p. 33).

[102]Mathieu Gautier, Julien Foucaud, Karim Gharbi, et al. “Estimation of population allele frequencies from next-generation sequencing data: pool-versus individual-based genotyping”. In: *Molecular Ecology* 22.14 (June 2013), 3766–3779 (cit. on p. 33).

[103]Katalin Csilléry, Alejandra Rodríguez-Verdugo, Christian Rellstab, and Frédéric Guillaume. “Detecting the genomic signal of polygenic adaptation and the role of epistasis in evolution”. In: *Molecular Ecology* 27.3 (Feb. 2018), 606–612 (cit. on pp. 35, 36).

[104]Evan A. Boyle, Yang I. Li, and Jonathan K. Pritchard. “An Expanded View of Complex Traits: From Polygenic to Omnigenic”. In: *Cell* 169.7 (June 2017), 1177–1186 (cit. on pp. 36, 108).

[105]Jonathan K. Pritchard, Joseph K. Pickrell, and Graham Coop. “The Genetics of Human Adaptation: Hard Sweeps, Soft Sweeps, and Polygenic Adaptation”. In: *Current Biology* 20.4 (Feb. 2010), R208–R215 (cit. on pp. 36, 116).

[106]Kavita Jain and Wolfgang Stephan. “Rapid Adaptation of a Polygenic Trait After a Sudden Environmental Shift”. In: *Genetics* 206.1 (May 2017), 389–406 (cit. on p. 36).

[107]David B. Goldstein and Kent E. Holsinger. “MAINTENANCE OF POLYGENIC VARIATION IN SPATIALLY STRUCTURED POPULATIONS: ROLES FOR LOCAL MATING AND GENETIC REDUNDANCY”. In: *Evolution* 46.2 (Apr. 1992), 412–429 (cit. on p. 36).

[108]Martin A. Nowak, Maarten C. Boerlijst, Jonathan Cooke, and John Maynard Smith. “Evolution of genetic redundancy”. In: *Nature* 388.6638 (July 1997), 167–171 (cit. on p. 36).

[109]Sam Yeaman. “Local Adaptation by Alleles of Small Effect”. In: *The American Naturalist* 186.S1 (Oct. 2015), S74–S89 (cit. on p. 36).

[110]SEWALL WRIGHT. “THE GENETICAL STRUCTURE OF POPULATIONS”. In: *Annals of Eugenics* 15.1 (Jan. 1949), 323–354 (cit. on p. 37).

[111]Kent E. Holsinger and Bruce S. Weir. “Genetics in geographically structured populations: defining, estimating and interpreting FST”. In: *Nature Reviews Genetics* 10.9 (Sept. 2009), 639–650 (cit. on pp. 37, 65).

[112]Masatoshi Nei. “Analysis of Gene Diversity in Subdivided Populations”. In: *Proceedings of the National Academy of Sciences* 70.12 (Dec. 1973), 3321–3323 (cit. on p. 37).

[113]B. S. Weir and C. Clark Cockerham. “ESTIMATINGF-STATISTICS FOR THE ANALYSIS OF POPULATION STRUCTURE”. In: *Evolution* 38.6 (Nov. 1984), 1358–1370 (cit. on p. 37).

[114]R R Hudson, M Slatkin, and W P Maddison. “Estimation of levels of gene flow from DNA sequence data.” In: *Genetics* 132.2 (Oct. 1992), 583–589 (cit. on p. 37).

[115]Sankar Subramanian. “The Difference in the Proportions of Deleterious Variations within and between Populations Influences the Estimation of FST”. In: *Genes* 13.2 (Jan. 2022), p. 194 (cit. on p. 37).

[116]PATRICK G. MEIRMANS and PHILIP W. HEDRICK. “Assessing population structure: FST and related measures”. In: *Molecular Ecology Resources* 11.1 (Oct. 2010), 5–18 (cit. on p. 37).

[117]Mattias Jakobsson, Michael D Edge, and Noah A Rosenberg. “The Relationship Between FST and the Frequency of the Most Frequent Allele”. In: *Genetics* 193.2 (Feb. 2013), 515–528 (cit. on p. 38).

[118]F. Tajima. “Evolutionary relationship of DNA sequences in finite populations”. In: *Genetics* 105.2 (1983), pp. 437–460 (cit. on p. 38).

[119]G A Watterson. “On the number of segregating sites in genetical models without recombination”. en. In: *Theor. Popul. Biol.* 7.2 (1975), pp. 256–276 (cit. on pp. 38, 65).

[120]Alper Karagöl and Taner Karagöl. “An Evolutionary Statistics Toolkit for Simplified Sequence Analysis on Web with Client-Side Processing”. In: (Aug. 2024) (cit. on p. 38).

[121]Kevin Thornton. “Recombination and the Properties of Tajima’s D in the Context of Approximate-Likelihood Calculation”. In: *Genetics* 171.4 (Dec. 2005), 2143–2148 (cit. on p. 39).

[122]Zoltán Bagi, Evangelos Antonis Dimopoulos, Dimitrios Loukovitis, Cyril Eraud, and Szilvia Kusza. “MtDNA genetic diversity and structure of Eurasian Collared Dove (Streptopelia decaocto)”. In: *PLoS One* 13.3 (Mar. 2018), e0193935 (cit. on p. 39).

[123]A. B. Zhang, X. B. Kong, D. M. Li, and Y. Q. Liu. “DNA barcoding of Chinese Dendrolimus punctatus and Dendrolimus tabulaeformis (Lepidoptera: Lasiocampidae) based on the COI gene and ITS2 sequences”. In: *Molecular Ecology Resources* 14.1 (2014), pp. 122–133 (cit. on p. 39).

[124]Sewall Wright. “EVOLUTION IN MENDELIAN POPULATIONS”. In: *Genetics* 16.2 (Mar. 1931), 97–159 (cit. on p. 39).

[125]Brian Charlesworth. “Effective population size and patterns of molecular evolution and variation”. In: *Nature Reviews Genetics* 10.3 (Mar. 2009), 195–205 (cit. on p. 39).

[126]J. H. Gillespie. *Population genetics: a concise guide*. JHU Press, 2004 (cit. on p. 39).

[127]D. S. Falconer and T. F. C. Mackay. *Introduction to quantitative genetics*. Longman, 1996 (cit. on p. 40).

[128]Peter M. Visscher, William G. Hill, and Naomi R. Wray. “Heritability in the genomics era — concepts and misconceptions”. In: *Nature Reviews Genetics* 9.4 (Mar. 2008), 255–266 (cit. on p. 40).

[129]N. H. Barton. “Genetic hitchhiking”. In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 355.1403 (Nov. 2000). Ed. by B. Charlesworth and P. H. Harvey, 1553–1562 (cit. on p. 40).

[130]Adam Eyre-Walker and Peter D. Keightley. “The distribution of fitness effects of new mutations”. In: *Nature Reviews Genetics* 8.8 (Aug. 2007), 610–618 (cit. on p. 40).

[131]Motoo Kimura. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Oct. 1983 (cit. on p. 40).

[132]H. Abdel-Haleem. “The Origins of Genome Architecture”. In: *Journal of Heredity* 98.6 (July 2007), 633–634 (cit. on p. 40).

[133]Jonathan B. Losos. “CONVERGENCE, ADAPTATION, AND CONSTRAINT”. In: *Evolution* 65.7 (Apr. 2011), 1827–1840 (cit. on p. 40).

[134]Annalise B. Paaby and Nicholas D. Testa. “Developmental Plasticity and Evolution”. In: *Evolutionary Developmental Biology*. Springer International Publishing, 2021, 1073–1086 (cit. on p. 40).

[135]Ryan Poplin, Valentin Ruano-Rubio, Mark A. DePristo, et al. “Scaling accurate genetic variant discovery to tens of thousands of samples”. In: (Nov. 2017) (cit. on p. 43).

[136]Dmytro Kryvokhyzha. *GATK: the best practice for genotype calling in a non-model organism — evodify.com*. `https://evodify.com/gatk-in-non-model-organism/`. [Accessed 28-02-2024] (cit. on p. 43).

[137]Thomas Taus, Andreas Futschik, and Christian Schlötterer. “Quantifying Selection with Pool-Seq Time Series Data”. In: *Molecular Biology and Evolution* 34.11 (Aug. 2017), pp. 3023–3034 (cit. on pp. 44, 46, 49, 61, 63, 66).

[138]F. Pedregosa, G. Varoquaux, A. Gramfort, et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830 (cit. on pp. 48, 50).

[139]Halina Binde Doria and Markus Pfenninger. “A multigenerational approach can detect early Cd pollution in Chironomus riparius”. In: *Chemosphere* 262 (Jan. 2021), p. 127815 (cit. on pp. 57, 111).

[140]Andrews S. “fastQC”. In: *Babraham Institute, Cambridge, United Kingdom* (2010) (cit. on p. 60).

[141]Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. “Trimmomatic: a flexible trimmer for Illumina sequence data”. In: *Bioinformatics* 30.15 (Apr. 2014), pp. 2114–2120 (cit. on p. 60).

[142]Hanno Schmidt, Sören Lukas Hellmann, Ann-Marie Waldvogel, et al. “A High-Quality Genome Assembly from Short and Long Reads for the Non-biting Midge Chironomus riparius (Diptera)”. In: *G3 Genes|Genomes|Genetics* 10.4 (Apr. 2020), 1151–1157 (cit. on p. 60).

[143]Heng Li and Richard Durbin. “Fast and accurate short read alignment with Burrows–Wheeler transform”. In: *bioinformatics* 25.14 (2009), pp. 1754–1760 (cit. on p. 60).

[144]Heng Li, Bob Handsaker, Alec Wysoker, et al. “The Sequence Alignment/Map format and SAMtools”. In: *Bioinformatics* 25.16 (June 2009), pp. 2078–2079 (cit. on p. 61).

[145]Robert Kofler, Pablo Orozco-terWengel, Nicola De Maio, et al. “PoPoolation: A Toolbox for Population Genetic Analysis of Next Generation Sequencing Data from Pooled Individuals”. In: *PLoS ONE* 6.1 (Jan. 2011). Ed. by Manfred Kayser, e15925 (cit. on pp. 61, 65).

[146]“Using Population Genomics to Detect Selection in Natural Populations: Key Concepts and Methodological Considerations”. In: *International Journal of Plant Sciences* 171.9 (Nov. 2010), 1059–1071 (cit. on p. 64).

[147]Daniel Lakens. “Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs”. In: *Frontiers in Psychology* 4 (2013) (cit. on p. 66).

[148]Philipp Schönnenbeck, Tilman Schell, Susanne Gerber, and Markus Pfenninger. “tbg - a new file format for genomic data”. In: (Mar. 2021) (cit. on p. 66).

[149]Matthias Blum, Hsin-Yu Chang, Sara Chuguransky, et al. “The InterPro protein families and domains database: 20 years on”. In: *Nucleic Acids Research* 49.D1 (Nov. 2020), D344–D354 (cit. on p. 66).

[150]Jaina Mistry, Sara Chuguransky, Lowri Williams, et al. “Pfam: The protein families database in 2021”. In: *Nucleic Acids Research* 49.D1 (Oct. 2020), D412–D419 (cit. on p. 66).

[151]*GitHub - dusadrian/venn: Draw Venn Diagrams — github.com*. `https://github.com/dusadrian/venn`. [Accessed 23-09-2024] (cit. on p. 66).

[152]John M Elizarraras, Yuxing Liao, Zhiao Shi, et al. “WebGestalt 2024: faster gene set analysis and new support for metabolomics and multi-omics”. In: *Nucleic Acids Research* 52.W1 (May 2024), W415–W421 (cit. on p. 66).

[153]Andrew R Wood, Tonu Esko, Jian Yang, et al. “Defining the role of common variation in the genomic and biological architecture of adult human height”. In: *Nature Genetics* 46.11 (Oct. 2014), 1173–1186 (cit. on p. 108).

[154]Matthew V. Rockman. “THE QTN PROGRAM AND THE ALLELES THAT MATTER FOR EVOLUTION: ALL THAT’S GOLD DOES NOT GLITTER”. In: *Evolution* 66.1 (Nov. 2011), 1–17 (cit. on p. 108).

[155]Pedro Larrañaga, Borja Calvo, Roberto Santana, et al. “Machine learning in bioinformatics”. In: *Briefings in Bioinformatics* 7.1 (Mar. 2006), 86–112 (cit. on pp. 108, 110).

[156]Nurshazlyn Mohd Aszemi. “Hyperparameter search method in Convolutional Neural Network: A Systematic Literature Review”. In: (2019) (cit. on p. 109).

[157]Mennatallah Amer, Markus Goldstein, and Slim Abdennadher. *Enhancing one-class Support Vector Machines for unsupervised anomaly detection*. Aug. 2013 (cit. on p. 110).

[158]Olivier Chapelle, Vladimir Vapnik, Olivier Bousquet, and Sayan Mukherjee. In: *Machine Learning* 46.1/3 (2002), 131–159 (cit. on pp. 110, 121).

[159]Marino Marinković, Wim C. de Leeuw, Mark de Jong, et al. “Combining Next-Generation Sequencing and Microarray Technology into a Transcriptomics Approach for the Non-Model Organism Chironomus riparius”. In: *PLoS ONE* 7.10 (Oct. 2012). Ed. by Michael Watson, e48096 (cit. on p. 111).

[160]F. Stefani, M. Rusconi, S. Valsecchi, and L. Marziali. “Evolutionary ecotoxicology of perfluoralkyl substances (PFASs) inferred from multigenerational exposure: A case study with Chironomus riparius (Diptera, Chironomidae)”. In: *Aquatic Toxicology* 156 (Nov. 2014), 41–51 (cit. on p. 111).

[161]David S. Saunders. “Insect photoperiodism: effects of temperature on the induction of insect diapause and diverse roles for the circadian system in the photoperiodic response”. In: *Entomological Science* 17.1 (Oct. 2013), 25–40 (cit. on pp. 111, 112).

[162]Caitlin M. Dmitriew. “The evolution of growth trajectories: what limits growth rate?” In: *Biological Reviews* 86.1 (Apr. 2010), 97–116 (cit. on p. 112).

[163]N. G. Prasad, Mallikarjun Shakarad, D. Anitha, M. Rajamani, and Amitabh Joshi. “CORRELATED RESPONSES TO SELECTION FOR FASTER DEVELOPMENT AND EARLY REPRODUCTION IN DROSOPHILA: THE EVOLUTION OF LARVAL TRAITS”. In: *Evolution* 55.7 (July 2001), 1363–1372 (cit. on p. 112).

[164]Troy E. Sandberg, Colton J. Lloyd, Bernhard O. Palsson, and Adam M. Feist. “Laboratory Evolution to Alternating Substrate Environments Yields Distinct Phenotypic and Genetic Adaptive Strategies”. In: *Applied and Environmental Microbiology* 83.13 (July 2017). Ed. by Maia Kivisaar (cit. on p. 112).

[165]Maryl Lambros, Ximo Pechuan-Jorge, Daniel Biro, Kenny Ye, and Aviv Bergman. "Emerging Adaptive Strategies Under Temperature Fluctuations in a Laboratory Evolution Experiment of Escherichia Coli". In: *Frontiers in Microbiology* 12 (Oct. 2021) (cit. on p. 112).

[166]Ryan Calsbeek, Thomas P. Gosden, Shawn R. Kuchta, and Erik I. Svensson. "Fluctuating Selection and Dynamic Adaptive Landscapes". In: *The Adaptive Landscape in Evolutionary Biology*. Oxford University Press, May 2013, 89–109 (cit. on p. 112).

[167]Scott P. Carroll, Peter Søgaard Jørgensen, Michael T. Kinnison, et al. "Applying evolutionary biology to address global challenges". In: *Science* 346.6207 (Oct. 2014) (cit. on p. 113).

[168]Montgomery Slatkin. "A Population-Genetic Test of Founder Effects and Implications for Ashkenazi Jewish Diseases". In: *The American Journal of Human Genetics* 75.2 (Aug. 2004), 282–293 (cit. on p. 113).

[169]Yipei Guo, Marija Vucelja, and Ariel Amir. "Stochastic tunneling across fitness valleys can give rise to a logarithmic long-term fitness trajectory". In: *Science Advances* 5.7 (July 2019) (cit. on p. 113).

[170]Kate Rick, Kym Ottewell, Cheryl Lohr, et al. "Population Genomics of Bettongia lesueur: Admixing Increases Genetic Diversity with no Evidence of Outbreeding Depression". In: *Genes* 10.11 (Oct. 2019), p. 851 (cit. on p. 113).

[171]Daniel Nettle and Willem E. Frankenhuis. "Life-history theory in psychology and evolutionary biology: one research programme or two?" In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 375.1803 (June 2020), p. 20190490 (cit. on p. 114).

[172]Marina Wolz, Michael Klockmann, Torben Schmitz, et al. "Dispersal and life-history traits in a spider with rapid range expansion". In: *Movement Ecology* 8.1 (Jan. 2020) (cit. on p. 114).

[173]Lauren B Buckley, Andrew J Arakaki, Anthony F Cannistra, Heather M Kharouba, and Joel G Kingsolver. "Insect Development, Thermal Plasticity and Fitness Implications in Changing, Seasonal Environments". In: *Integrative and Comparative Biology* 57.5 (June 2017), 988–998 (cit. on p. 114).

[174]Michael Pointer. *Genetic architecture of dispersal behaviour in the post-harvest pest and model organism Tribolium castaneum*. en. 2023 (cit. on p. 114).

[175]Thomas F. Hansen. *Evolutionary Constraints*. Jan. 2015 (cit. on p. 114).

[176]Hyungsoon Jeong, Yong-Chan Cho, and Eunsuk Kim. "Site-specific temporal variation of population dynamics in subalpine endemic plant species". In: *Scientific Reports* 12.1 (Nov. 2022) (cit. on p. 114).

[177]C. Pertoldi, L. A. Bach, J. S. F. Barker, P. Lundberg, and V. Loeschcke. "The consequences of the variance-mean rescaling effect on effective population size". In: *Oikos* 116.5 (May 2007), 769–774 (cit. on p. 115).

[178]J Wang, E Santiago, and A Caballero. “Prediction and estimation of effective population size”. In: *Heredity* 117.4 (June 2016), 193–206 (cit. on p. 115).

[179]Vince Buffalo. “Quantifying the relationship between genetic diversity and population size suggests natural selection cannot explain Lewontin’s Paradox”. In: *eLife* 10 (Aug. 2021) (cit. on p. 115).

[180]Karen Bisschop, Adriana Alzate, D. Bonte, and R. Etienne. “The Demographic Consequences of Adaptation: Evidence from Experimental Evolution”. In: *The American Naturalist* 199 (2022), pp. 729 –742 (cit. on p. 115).

[181]Mirko Pegoraro, Shumaila Noreen, Supriya Bhutani, et al. “Molecular Evolution of a Pervasive Natural Amino-Acid Substitution in Drosophila cryptochrome”. In: *PLoS ONE* 9.1 (Jan. 2014). Ed. by Nadia Singh, e86483 (cit. on p. 116).

[182]L M Gattepaille, M Jakobsson, and M GB Blum. “Inferring population size changes with sequence and SNP data: lessons from human bottlenecks”. In: *Heredity* 110.5 (Feb. 2013), 409–419 (cit. on p. 116).

[183]David A Moeller, Maud I Tenaillon, and Peter Tiffin. “Population Structure and Its Effects on Patterns of Nucleotide Polymorphism in Teosinte (Zea mays ssp. parviglumis)”. In: *Genetics* 176.3 (July 2007), 1799–1809 (cit. on p. 116).

[184]M. C. Bitter, L. Kapsenberg, J.-P. Gattuso, and C. A. Pfister. “Standing genetic variation fuels rapid adaptation to ocean acidification”. In: *Nature Communications* 10.1 (Dec. 2019) (cit. on p. 116).

[185]Yu-Ting Lai, Carol K. L. Yeung, Kevin E. Omland, et al. “Standing genetic variation as the predominant source for adaptation of a songbird”. In: *Proceedings of the National Academy of Sciences* 116.6 (Jan. 2019), 2152–2157 (cit. on p. 116).

[186]L. Bernatchez. “On the maintenance of genetic variation and adaptation to environmental change: considerations from population genomics in fishes”. In: *Journal of Fish Biology* 89.6 (Sept. 2016), 2519–2556 (cit. on pp. 116, 123).

[187]Wolfgang Stephan, Yun S Song, and Charles H Langley. “The Hitchhiking Effect on Linkage Disequilibrium Between Linked Neutral Loci”. In: *Genetics* 172.4 (Apr. 2006), 2647–2663 (cit. on p. 116).

[188]Gang Liu, Bao-Feng Zhang, Jiang Chang, et al. “Population genomics reveals moderate genetic differentiation between populations of endangered Forest Musk Deer located in Shaanxi and Sichuan”. In: *BMC Genomics* 23.1 (Sept. 2022) (cit. on pp. 117, 119).

[189]M.C. Bitter, S. Berardi, H. Oken, et al. “Continuously fluctuating selection reveals extreme granularity and parallelism of adaptive tracking”. In: (Oct. 2023) (cit. on p. 117).

[190]R A Neher and B I Shraiman. “Genetic Draft and Quasi-Neutrality in Large Facultatively Sexual Populations”. In: *Genetics* 188.4 (Aug. 2011), 975–996 (cit. on p. 117).

[191]Luis-Miguel Chevin. “Selective Sweep at a QTL in a Randomly Fluctuating Environment”. In: *Genetics* 213.3 (Nov. 2019), 987–1005 (cit. on p. 117).

[192]Zachary D. Blount, Christina Z. Borland, and Richard E. Lenski. “Historical contingency and the evolution of a key innovation in an experimental population ofEscherichia coli”. In: *Proceedings of the National Academy of Sciences* 105.23 (June 2008), 7899–7906 (cit. on p. 117).

[193]Zachary D. Blount, Richard E. Lenski, and Jonathan B. Losos. “Contingency and determinism in evolution: Replaying life’s tape”. In: *Science* 362.6415 (Nov. 2018) (cit. on p. 117).

[194]Kyle J. Card, Thomas LaBar, Jasper B. Gomez, and Richard E. Lenski. “Historical contingency in the evolution of antibiotic resistance after decades of relaxed selection”. In: *PLOS Biology* 17.10 (Oct. 2019). Ed. by Csaba Pál, e3000397 (cit. on p. 117).

[195]Tiago da Silva Ribeiro, José A Galván, and John E Pool. “Maximum SNP FST Outperforms Full-Window Statistics for Detecting Soft Sweeps in Local Adaptation”. In: *Genome Biology and Evolution* 14.10 (Sept. 2022). Ed. by Andrea Betancourt (cit. on p. 117).

[196]Maddie E. James, Melanie J. Wilkinson, Diana M. Bernal, et al. “Phenotypic and genotypic parallel evolution in parapatric ecotypes ofSenecio”. In: *Evolution* 75.12 (Nov. 2021), 3115–3131 (cit. on p. 117).

[197]Qicheng Xu, He Zhang, Philippe Vandenkoornhuyse, et al. “Carbon starvation raises capacities in bacterial antibiotic resistance and viral auxiliary carbon metabolism in soils”. In: *Proceedings of the National Academy of Sciences* 121.16 (Apr. 2024) (cit. on p. 118).

[198]Dr.D.Sujatha, Dr. D.Umamaheswari, and Dr.D.Vijayalakshmi. “To analyze the impact of Thermal- stress on the RBC number of Indian Major carp Catla catla (Hamilton)”. In: *vol 8 issue 3* 8.3 (2024), 100–109 (cit. on p. 118).

[199]Kavita Jain and Wolfgang Stephan. “Modes of Rapid Polygenic Adaptation”. In: *Molecular Biology and Evolution* 34.12 (Sept. 2017), 3169–3175 (cit. on pp. 118, 123).

[200]Ilse Höllinger, Pleuni S. Pennings, and Joachim Hermisson. “Polygenic adaptation: From sweeps to subtle frequency shifts”. In: *PLOS Genetics* 15.3 (Mar. 2019). Ed. by Justin C. Fay, e1008035 (cit. on p. 118).

[201]Doriane Lorendeau, Stefan Christen, Gianmarco Rinaldi, and Sarah-Maria Fendt. “Metabolic control of signalling pathways and metabolic auto-regulation”. In: *Biology of the Cell* 107.8 (June 2015), 251–272 (cit. on p. 118).

[202]James H. Marden. “Nature’s inordinate fondness for metabolic enzymes: why metabolic enzyme loci are so frequently targets of selection”. In: *Molecular Ecology* 22.23 (Oct. 2013), 5743–5764 (cit. on p. 118).

[203]Kathrin A. Otte, Viola Nolte, François Mallard, and Christian Schlötterer. “The genetic architecture of temperature adaptation is shaped by population ancestry and not by selection regime”. In: *Genome Biology* 22.1 (July 2021) (cit. on p. 118).

[204]Richard E. Lenski, Michael R. Rose, Suzanne C. Simpson, and Scott C. Tadler. "Long-Term Experimental Evolution in Escherichia coli. I. Adaptation and Divergence During 2, 000 Generations". In: *The American Naturalist* 138.6 (Dec. 1991), 1315–1341 (cit. on p. 118).

[205]Richard E. Lenski, Michael J. Wiser, Noah Ribeck, et al. "Sustained fitness gains and variability in fitness trajectories in the long-term evolution experiment withEscherichia coli". In: *Proceedings of the Royal Society B: Biological Sciences* 282.1821 (Dec. 2015), p. 20152292 (cit. on p. 118).

[206]Michael J. Wiser, Noah Ribeck, and Richard E. Lenski. "Long-Term Dynamics of Adaptation in Asexual Populations". In: *Science* 342.6164 (Dec. 2013), 1364–1367 (cit. on p. 118).

[207]Martin Dragosits and Diethard Mattanovich. "Adaptive laboratory evolution – principles and applications for biotechnology". In: *Microbial Cell Factories* 12.1 (2013), p. 64 (cit. on p. 118).

[208]M A Harmon, M F Boehm, R A Heyman, and D J Mangelsdorf. "Activation of mammalian retinoid X receptors by the insect growth regulator methoprene." In: *Proceedings of the National Academy of Sciences* 92.13 (June 1995), 6157–6160 (cit. on p. 119).

[209]Lorraine J Gudas. "Retinoid metabolism: new insights". In: *Journal of Molecular Endocrinology* 69.4 (Nov. 2022), T37–T49 (cit. on p. 119).

[210]Tianlei Zhang, Wei Song, Zheng Li, et al. "Krüppel homolog 1 represses insect ecdysone biosynthesis by directly inhibiting the transcription of steroidogenic enzymes". In: *Proceedings of the National Academy of Sciences* 115.15 (Mar. 2018), 3960–3965 (cit. on p. 119).

[211]Christen Kerry Mirth, Hui Yuan Tang, Sasha C. Makohon-Moore, et al. "Juvenile hormone regulates body size and perturbs insulin signaling in Drosophila". In: *Proceedings of the National Academy of Sciences* 111.19 (Apr. 2014), 7018–7023 (cit. on p. 119).

[212]Henry Chung, Tamar Sztal, Shivani Pasricha, et al. "Characterization of Drosophila melanogaster cytochrome P450 genes". In: *Proceedings of the National Academy of Sciences* 106.14 (Apr. 2009), 5731–5736 (cit. on p. 119).

[213]Zijie Huang, Zhong Tian, Yulian Zhao, et al. "MAPK Signaling Pathway Is Essential for Female Reproductive Regulation in the Cabbage Beetle, Colaphellus bowringi". In: *Cells* 11.10 (May 2022), p. 1602 (cit. on p. 119).

[214]Zhaojiang Guo, Shi Kang, Qingjun Wu, et al. "The regulation landscape of MAPK signaling cascade for thwarting Bacillus thuringiensis infection in an insect host". In: *PLOS Pathogens* 17.9 (Sept. 2021). Ed. by Rachel M. McLoughlin, e1009917 (cit. on p. 119).

[215]A. R. Armstrong and C. L. Boggs. "Antibody development to identify components of IIS and mTOR signaling pathways in lepidopteran species, a set of non-model insects". In: *microPublication Biol.* (2023) (cit. on p. 119).

[216]Avani Patel, M. Kim Fondrk, Osman Kaftanoglu, et al. “The Making of a Queen: TOR Pathway Is a Key Player in Diphenic Caste Development”. In: *PLoS ONE* 2.6 (June 2007). Ed. by Pawel Michalak, e509 (cit. on p. 119).

[217]Georg Oberhofer, Daniela Grossmann, Janna L. Siemanowski, Tim Beissbarth, and Gregor Bucher. “Wnt/β-catenin signaling integrates patterning and metabolism of the insect growth zone”. In: *Development* 141.24 (Dec. 2014), 4740–4750 (cit. on p. 120).

[218]Suning Liu, Wei Wei, Yuan Chu, et al. “De Novo Transcriptome Analysis of Wing Development-Related Signaling Pathways in Locusta migratoria Manilensis and Ostrinia furnacalis (Guenée)”. In: *PLoS ONE* 9.9 (Sept. 2014). Ed. by Kun Yan Zhu, e106770 (cit. on p. 120).

[219]Yinghui Wang, Yunji Xiu, Keran Bi, et al. “Integrated analysis of mRNA-seq in the haemocytes of Eriocheir sinensis in response to Spiroplasma eriocheiris infection”. In: *Fish & Shellfish Immunology* 68 (2017), pp. 289–298 (cit. on p. 120).

[220]Arthur M. Butt, Robert F. Fern, and Carlos Matute. “Neurotransmitter signaling in white matter”. In: *Glia* 62.11 (Apr. 2014), 1762–1779 (cit. on p. 120).

[221]Cheolho Sim and David L. Denlinger. “Insulin signaling and FOXO regulate the overwintering diapause of the mosquito Culex pipiens”. In: *Proceedings of the National Academy of Sciences* 105.18 (May 2008), 6777–6781 (cit. on p. 120).

[222]Baosheng Zeng, Yuping Huang, Jun Xu, et al. “The FOXO transcription factor controls insect growth and development by regulating juvenile hormone degradation in the silkworm, Bombyx mori”. In: *Journal of Biological Chemistry* 292.28 (July 2017), 11659–11669 (cit. on p. 120).

[223]Daniel González-Tokman, Alex Córdoba-Aguilar, Wesley Dáttilo, et al. “Insect responses to heat: physiological mechanisms, evolution and ecological implications in a warming world”. In: *Biological Reviews* 95.3 (Feb. 2020), 802–821 (cit. on p. 120).

[224]Bianca D. van Groen, Karel Allegaert, Dick Tibboel, and Saskia N. de Wildt. “Innovative approaches and recent advances in the study of ontogeny of drug metabolism and transport”. In: *British Journal of Clinical Pharmacology* 88.10 (Sept. 2020), 4285–4296 (cit. on p. 120).

[225]J. Steven Leeder. “Developmental Aspects of Drug Metabolism in Children”. In: *Drug Information Journal* 30.4 (Oct. 1996), 1135–1143 (cit. on p. 120).

[226]Marta A. Antunes, Afonso Grandela, Margarida Matos, and Pedro Simões. “Long-term evolution experiments fully reveal the potential for thermal adaptation”. In: (Feb. 2024) (cit. on p. 120).

[227]R.-J. T. S. Palstra. “Close encounters of the 3C kind: long-range chromatin interactions and transcriptional regulation”. In: *Briefings in Functional Genomics and Proteomics* 8.4 (June 2009), 297–309 (cit. on p. 122).

[228]Adriaan van der Graaf, René Wardenaar, Drexel A. Neumann, et al. “Rate, spectrum, and evolutionary dynamics of spontaneous epimutations”. In: *Proceedings of the National Academy of Sciences* 112.21 (May 2015), 6676–6681 (cit. on p. 122).

[229]Elizabeth R Jerison and Michael M Desai. “Genomic investigations of evolutionary dynamics and epistasis in microbial evolution experiments”. In: *Current Opinion in Genetics Development* 35 (Dec. 2015), 33–39 (cit. on p. 122).

[230]Halina Binde Doria, Cosima Caliendo, Susanne Gerber, and Markus Pfenninger. “Photoperiod is an important seasonal selection factor in Chironomus riparius (Diptera: Chironomidae)”. In: *Biological Journal of the Linnean Society* 135.2 (Dec. 2021), 277–290 (cit. on p. 122).

[231]Sudarshan Chari, Christian Marier, Cody Porter, et al. “Compensatory evolution via cryptic genetic variation: Distinct trajectories to phenotypic and fitness recovery”. In: (Oct. 2017) (cit. on pp. 123, 124).

[232]Robin Hanson, John Stutz, and Peter Cheeseman. “Bayesian classification theory”. In: (1991) (cit. on p. 165).

Webpages

[@67]Christian. *Visualizing the bivariate Gaussian distribution*. 2016. URL: `https://scipython.com/blog/visualizing-the-bivariate-gaussian-distribution/` (visited on Aug. 3, 2016) (cit. on p. 20).

List of Figures

List of Tables

List of Listings

Example Appendix

A

Material and Methods Experiment

Day	mg/larvae	Food in 50ml; for 50 larvae
1	0.10	250
2	0.10	250
3	0.17	417
4	0.17	417
5	0.23	583
6	0.23	583
7	0.30	750
8	0.30	750
9	0.37	917
10	0.37	917
11	0.43	1083
12	0.43	1083
13	0.50	1250
14	0.50	1250
15	0.50	1250
16	0.50	1250
17	0.50	1250
18	0.50	1250
19	0.50	1250
20	0.50	1250
21	0.50	1250
22	0.50	1250
23	0.50	1250
24	0.50	1250
25	0.50	1250
26	0.50	1250
27	0.50	1250
28	0.50	1250

Fig. A.1.: Feeding protocol for larvae

	DF	F	P Value
Survival	6	21,52	5,87e-08
EmT50 male	6	102,1	1,96e-14
EmT50 female	6	141,1	7,39e-04
EmT50 female/male	6	6,272	6,78e-04
Mean Emergence Time	6	177,8	<2e-16
Sex ratio	6	1,769	0,154
Fertility	6	11,22	1,24e-05
PGR	6	13,81	2,46e-06

Fig. A.2.: ANOVA associated degrees of freedom, F values and p-values for each parameter analyzed.

	Emergence	EmT50 male	EmT50 female	EmT50 female/male	Mean Emergence Time	Fertility	PGR
1-2	0,3753	<2e-16	9e-07	0,0159	1e-07	0,4314	0,1631
1-3	0,0004	0,0087	3e-07	0,0009	<2e-16	0,1215	0,0111
1-4	1,4e-06	<2e-16	<2e-16	0,0046	<2e-16	0,0011	0,0001
1-5	1,24e-05	1,07e-04	<2e-16	0,0097	<2e-16	0,0015	0,0002
1-6	9e-07	1,02e-04	<2e-16	0,001	<2e-16	0,0011	0,0001
1-7	1,4e-06	0,0002	8e-07	0,0472	1e-07	8,4e-06	1,8e-06
1-3	0,0559	<2e-16	<2e-16	0,8692	<2e-16	0,9843	0,8436
1-4	0,0001	<2e-16	<2e-16	0,9975	<2e-16	0,0946	0,0441
1-5	0,0013	<2e-16	<2e-16	0,9999	<2e-16	0,1215	0,0748
1-6	7,9e-05	<2e-16	<2e-16	0,8798	<2e-16	0,0946	0,0431
2-7	1,2e-04	<2e-16	<2e-16	0,8581	<2e-16	0,0007	0,0005
3-4	0,1354	2,6e-06	2,2e-06	0,9915	3e-07	0,3610	0,4377
3-5	0,6451	0,4787	0,8869	0,9433	0,7521	0,4314	0,5922
3-6	0,0951	0,0788	0,0265	1	0,1294	0,3610	0,4311
3-7	0,1365	0,7373	0,9957	0,2047	0,9992	0,0042	0,0101
4-5	0,927	0,0001	2,72e-05	0,9998	4,7e-06	1	0,9999
4-6	0,9999	0,0019	0,0053	0,9931	7,28e-05	1	1
4-7	1	6,37e-05	7e-07	0,5507	1e-07	0,3284	0,4560
5-6	0,8591	0,9248	0,2673	0,9496	0,8501	1	0,9999
5-7	0,9282	0,9993	0,5561	0,7464	0,4822	0,2689	0,3186
6-7	0,9999	0,7312	0,0068	0,2140	0,0532	0,3284	0,4627

Fig. A.3.: p-values of the Tuckey's post hoc test for each comparison between the different generations.

Results Experiment

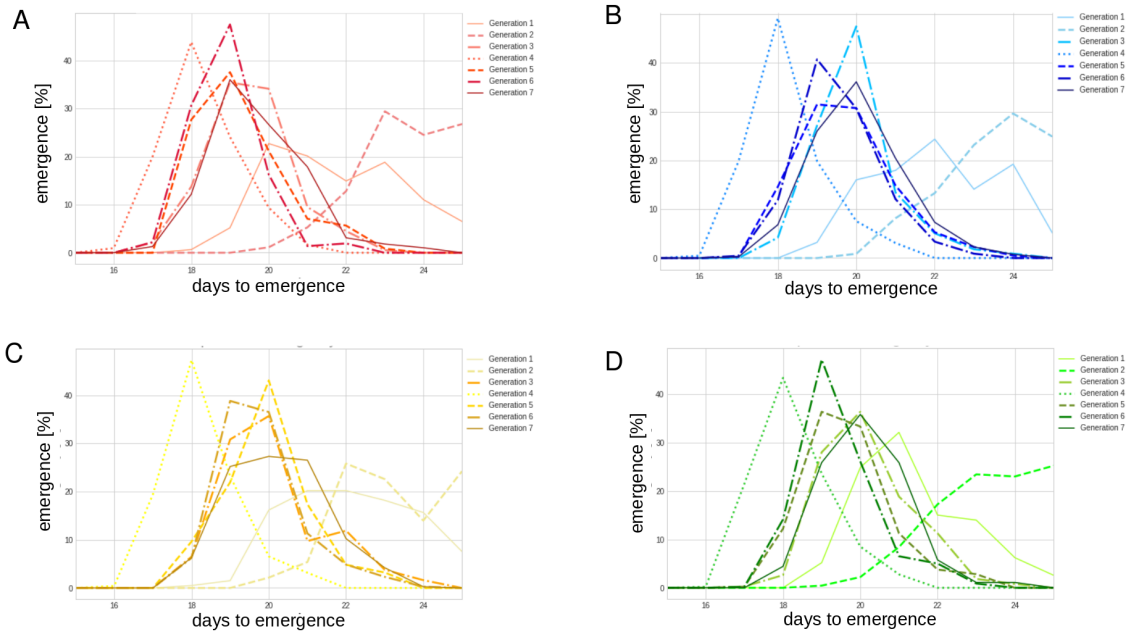

Fig. A.4.: Lineplot of emergence of the 4 replicates over 7 generations [%]. The distribution of the emergence of each individual is displayed per generation for A) red replicate, B) blue replicate, C) gold replicate and D) green replicate. The different lines styles indicate different generatios

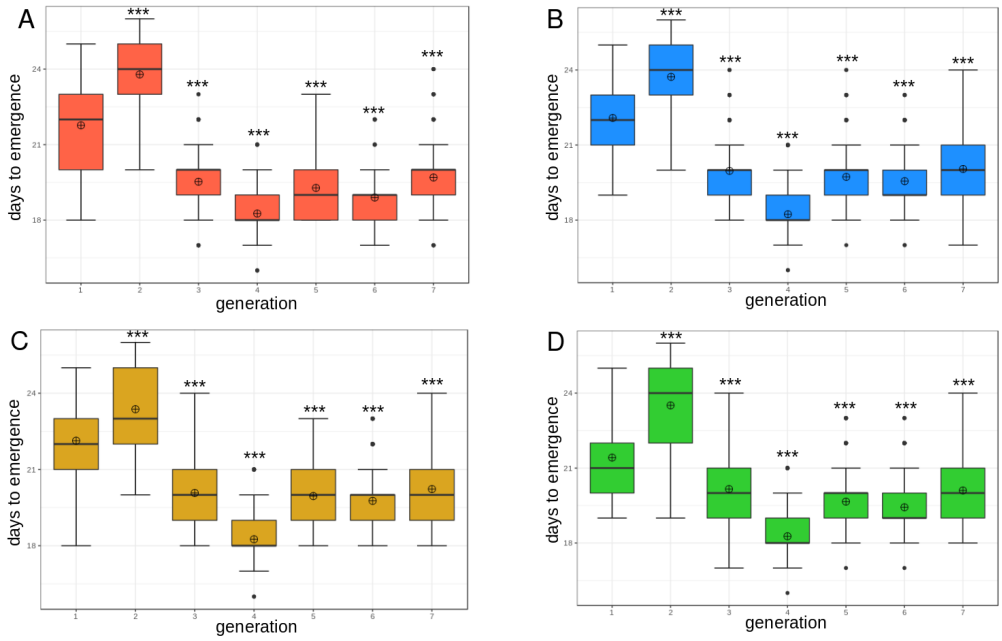

Fig. A.5.: Boxplot of emergence of the 4 replicates over 7 generations. The distribution of the emergence of each individual is displayed per generation for A) red replicate, B) blue replicate, C) gold replicate and D) green replicate. The horizontal line within the box indicates the median. The boundaries of the box indicate the 25th- and 75th-percentiles, and the whiskers indicate the highest and lowest results. Asterisks indicate differences in comparison to the initial (first) generation ($p<0.001$)

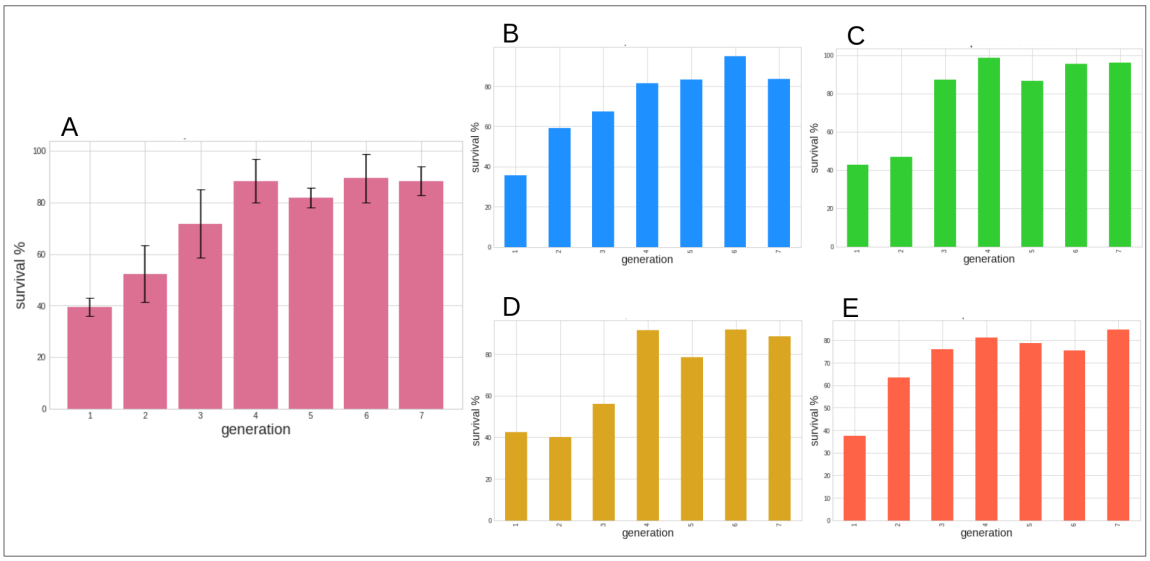

Fig. A.6.: Barplot survival in percentage of A) all replicates combined, B) replicate blue C) replicate green, D) replicate gold and E) replicate red. Errorbars indicate standard deviation

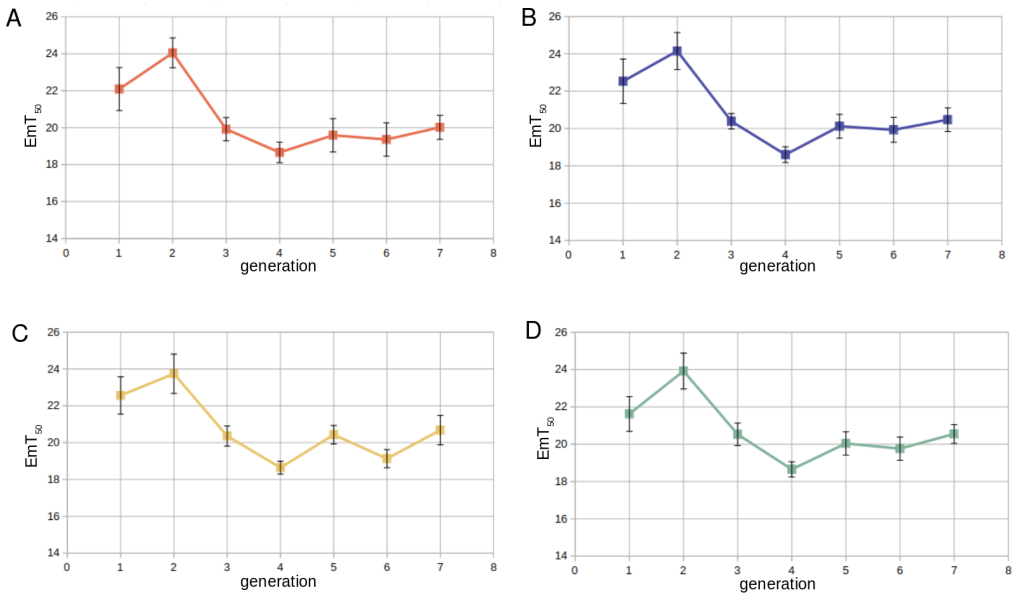

Fig. A.7.: EmT50 of the 4 replicates over 7 generations, females. This parameters displays the timepoint, where 50% of the population has emerged per generation, respectively A) the red replicate, B) blue replicate, C) gold replicate and D) green replicate. Error bars denote the standard deviation.

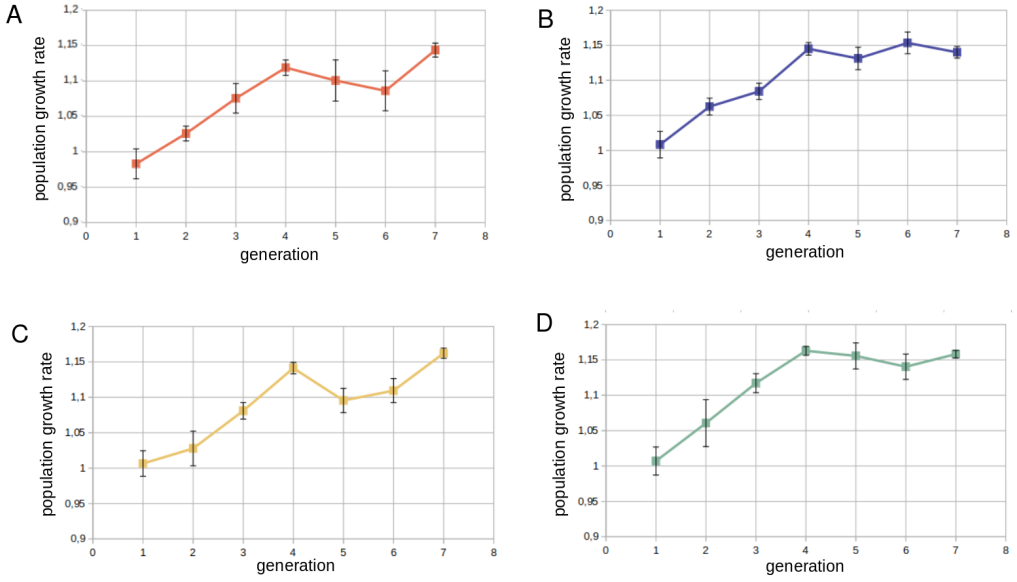

Fig. A.8.: Population Growth Rate (PGR) of the 4 replicates over 7 generations, females. PGR combines the factor of fertility (number of fertile eggs laid), emergence and survival, respective of A) the red replicate, B) blue replicate, C) gold replicate and D) green replicate. Error bars denotes for the standard deviation

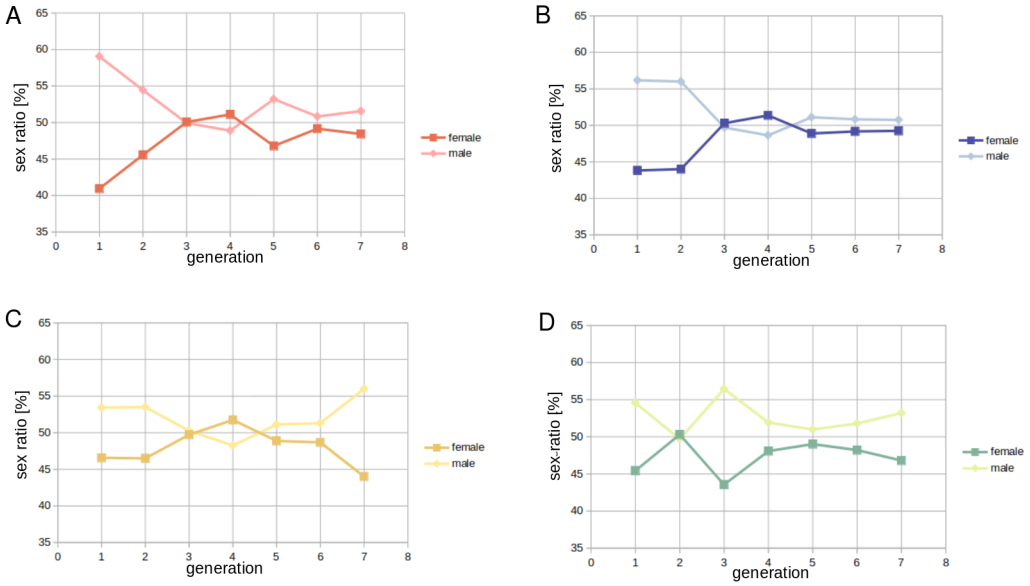

Fig. A.9.: Sex Ratio of the 4 replicates over 7 generations in percentage, females vs. males. Trends of the sex ratio between female and male are displayed in percentage per generation for A) the red replicate, B) blue replicate, C) gold replicate and D) green replicate.

A

Generation	Color	mean	std
1	Blau	20.378842	1.210200
	Gold	20.215074	1.087807
	Grün	20.517213	0.540004
	Rot	20.165837	0.845908
2	Blau	23.634267	0.759450
	Gold	23.149938	0.991520
	Grün	23.542389	0.862769
	Rot	23.888679	0.760202
3	Blau	19.921301	0.432164
	Gold	20.210171	0.468586
	Grün	20.085614	0.617557
	Rot	19.427999	0.602498
4	Blau	18.223042	0.436710
	Gold	18.253742	0.399840
	Grün	18.218654	0.419911
	Rot	18.231220	0.548612
5	Blau	19.739700	0.602655
	Gold	20.009568	0.520203
	Grün	19.557372	0.576368
	Rot	19.422767	0.841398
6	Blau	19.581543	0.625364
	Gold	19.723133	0.665338
	Grün	19.456415	0.573878
	Rot	19.264111	0.997119
7	Blau	20.135503	0.585424
	Gold	20.121481	0.754042
	Grün	20.125207	0.460000
	Rot	19.719256	0.623264

B

Generation	Color	mean	std
1	Blau	21.483500	1.189155
	Gold	21.457833	1.012347
	Grün	20.829474	0.929583
	Rot	20.969167	1.163005
2	Blau	23.304347	0.989253
	Gold	22.569167	1.071934
	Grün	22.996042	0.961381
	Rot	23.350476	0.807686
3	Blau	19.319936	0.413956
	Gold	19.574591	0.544614
	Grün	19.438655	0.599399
	Rot	18.900813	0.628383
4	Blau	17.628057	0.420421
	Gold	17.621859	0.351074
	Grün	17.620716	0.406158
	Rot	17.640834	0.554323
5	Blau	19.165808	0.639107
	Gold	19.490578	0.500986
	Grün	19.003278	0.624957
	Rot	18.727685	0.903399
6	Blau	19.203308	0.665755
	Gold	19.453078	0.489597
	Grün	19.001778	0.623231
	Rot	18.727685	0.903399
7	Blau	19.553337	0.634058
	Gold	19.522124	0.800439
	Grün	19.505601	0.506078
	Rot	19.084714	0.649698

C

Generation	Color	mean	std
1	Blau	1.008299	0.018942
	Gold	1.006500	0.017965
	Grün	1.007033	0.019798
	Rot	1.033029	0.021130
2	Blau	1.062521	0.012024
	Gold	1.027765	0.024426
	Grün	1.060671	0.033216
	Rot	1.075767	0.010393
3	Blau	1.084215	0.011683
	Gold	1.080911	0.011636
	Grün	1.117180	0.013422
	Rot	1.125520	0.020831
4	Blau	1.144968	0.008964
	Gold	1.141228	0.008033
	Grün	1.162923	0.006315
	Rot	1.168924	0.010818
5	Blau	1.131226	0.016017
	Gold	1.095506	0.017190
	Grün	1.155648	0.018455
	Rot	1.150689	0.029882
6	Blau	1.153409	0.015513
	Gold	1.109451	0.016963
	Grün	1.140401	0.017968
	Rot	1.136171	0.028191
7	Blau	1.140025	0.008254
	Gold	1.162213	0.007253
	Grün	1.158161	0.005402
	Rot	1.193733	0.009881

D

Generation	Color	mean	std
1	Blau	32.8	20.707487
	Gold	24.4	18.622002
	Grün	25.8	20.208388
	Rot	30.0	17.864991
2	Blau	33.4	22.338072
	Gold	35.6	26.338685
	Grün	19.6	20.760286
	Rot	34.4	30.224424
3	Blau	36.4	26.624880
	Gold	49.4	29.743465
	Grün	53.4	27.075722
	Rot	76.4	29.704510
4	Blau	68.6	19.435182
	Gold	78.4	20.371290
	Grün	88.2	14.479023
	Rot	63.8	25.840808
5	Blau	68.8	33.856664
	Gold	60.6	31.928209
	Grün	71.4	25.506243
	Rot	65.2	21.758483
6	Blau	64.3	30.464993
	Gold	55.6	26.845366
	Grün	69.3	24.054981
	Rot	66.9	21.299617
7	Blau	70.4	14.121279
	Gold	78.8	21.169989
	Grün	81.6	15.635318
	Rot	74.5	16.401219

E

Generation	Color	mean	std
1	Blau	0.380	0.000000
	Gold	0.340	0.000000
	Grün	0.340	0.000000
	Rot	0.473	0.031305
2	Blau	0.520	0.000000
	Gold	0.460	0.000000
	Grün	0.600	0.000000
	Rot	0.580	0.000000
3	Blau	0.540	0.000000
	Gold	0.500	0.000000
	Grün	0.680	0.000000
	Rot	0.680	0.000000
4	Blau	0.720	0.000000
	Gold	0.680	0.000000
	Grün	0.780	0.000000
	Rot	0.880	0.000000
5	Blau	0.720	0.000000
	Gold	0.560	0.000000
	Grün	0.880	0.000000
	Rot	0.860	0.000000
6	Blau	0.880	0.000000
	Gold	0.640	0.000000
	Grün	0.780	0.000000
	Rot	0.760	0.000000
7	Blau	0.780	0.000000
	Gold	0.920	0.000000
	Grün	0.880	0.000000
	Rot	1.160	0.000000

Fig. A.10.: Overview of the respective Mean per parameter per generation and replicate. A) Mean B) Emt50 C) PGR D) Survival E) Fertility

A

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
mu_0	21.130	1.007	19.188	22.970	0.015	0.011	4495.0	2629.0	1.0
mu_1	23.244	0.992	21.329	25.059	0.017	0.012	3384.0	1915.0	1.0
mu_2	19.137	1.034	17.265	21.121	0.014	0.010	5542.0	2784.0	1.0
mu_3	17.907	1.007	16.098	19.868	0.015	0.011	4682.0	2341.0	1.0
mu_4	18.996	0.974	17.233	20.891	0.014	0.010	4931.0	2328.0	1.0
mu_5	18.568	0.997	16.673	20.423	0.014	0.010	4858.0	2627.0	1.0
mu_6	19.229	0.968	17.371	20.924	0.013	0.009	5861.0	3343.0	1.0

B

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
mu_0	21.005	0.976	19.257	22.961	0.017	0.012	3402.0	1908.0	1.0
mu_1	23.006	1.007	21.021	24.834	0.017	0.012	3461.0	2046.0	1.0
mu_2	19.006	1.005	17.055	20.812	0.017	0.012	3550.0	2593.0	1.0
mu_3	18.010	1.014	16.118	19.982	0.016	0.011	3957.0	2618.0	1.0
mu_4	18.990	1.003	17.148	20.855	0.015	0.011	4275.0	2545.0	1.0
mu_5	17.995	0.997	16.090	19.863	0.016	0.012	3763.0	2567.0	1.0
mu_6	18.998	0.986	17.180	20.855	0.016	0.012	3562.0	2865.0	1.0

C

index	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
mu_pgr_0	1.033095	0.004647	1.023853	1.04133	6.4e-05	4.6e-05	5218.072551	2847.502017	1.000799
mu_pgr_1	1.075897	0.004614	1.067427	1.084692	6.1e-05	4.3e-05	5662.559487	2779.610069	1.001197
mu_pgr_2	1.125503	0.004577	1.117167	1.133908	6.6e-05	4.7e-05	4848.283072	2422.821969	1.000636
mu_pgr_3	1.168967	0.004634	1.160357	1.177757	6.5e-05	4.6e-05	5073.657099	2126.436096	1.000348
mu_pgr_4	1.150672	0.004494	1.141894	1.158596	6.2e-05	4.4e-05	5312.258952	2661.151818	0.999811
mu_pgr_5	1.136161	0.004511	1.127449	1.14459	6.8e-05	4.8e-05	4450.802371	2844.353232	1.000182
mu_pgr_6	1.192943	0.003825	1.186528	1.199653	6.3e-05	4.5e-05	3009.251911	1688.167308	1.000705
sigma_pgr	0.020298	0.001267	0.017999	0.022733	1.9e-05	1.4e-05	4356.04575	2378.607513	1.000372

D

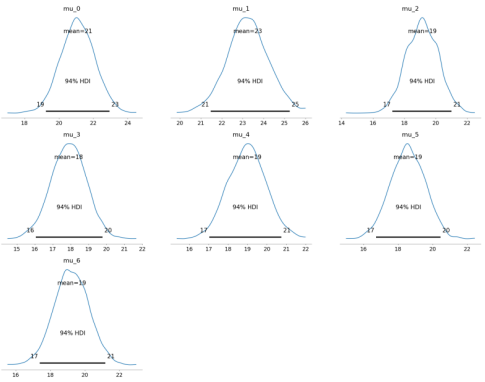

Fig. A.11.: Bayesian Statistics Replicate Red. A) Mean B) Emt50 C) PGR D) Mean Plotted

A

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
mu_0	21.468	1.007	19.693	23.523	0.017	0.012	3573.0	1875.0	1.0
mu_1	23.246	1.001	21.493	25.231	0.019	0.014	2624.0	1435.0	1.0
mu_2	19.628	0.971	17.708	21.345	0.014	0.010	4506.0	2891.0	1.0
mu_3	17.912	0.950	16.162	19.778	0.015	0.010	4226.0	2778.0	1.0
mu_4	19.344	0.986	17.581	21.225	0.017	0.012	3459.0	2067.0	1.0
mu_5	19.272	1.001	17.412	21.146	0.015	0.011	4262.0	2309.0	1.0
mu_6	19.547	1.013	17.661	21.408	0.016	0.011	4060.0	2759.0	1.0

B

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
mu_0	20.984	0.996	19.240	22.944	0.017	0.012	3363.0	2396.0	1.0
mu_1	23.017	0.997	21.135	24.855	0.015	0.011	4057.0	2319.0	1.0
mu_2	20.006	1.016	18.131	21.898	0.016	0.011	4058.0	2912.0	1.0
mu_3	18.007	0.996	16.073	19.806	0.016	0.012	3726.0	2676.0	1.0
mu_4	19.010	1.032	17.139	21.026	0.018	0.013	3382.0	2335.0	1.0
mu_5	18.999	1.005	17.088	20.794	0.016	0.011	4152.0	2278.0	1.0
mu_6	19.030	1.005	17.089	20.837	0.017	0.012	3651.0	2144.0	1.0

C

index	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
mu_pgr_0	1.008371	0.003093	1.002764	1.014333	4.3e-05	3e-05	5275.816917	3141.52733	0.999741
mu_pgr_1	1.062447	0.003079	1.056635	1.068343	4.5e-05	3.2e-05	4662.383224	2700.173762	1.001331
mu_pgr_2	1.08425	0.003057	1.078458	1.089949	4.6e-05	3.2e-05	4502.723942	2991.654663	1.000207
mu_pgr_3	1.144916	0.003085	1.13914	1.150856	4.6e-05	3.2e-05	4558.646574	2956.897431	1.000339
mu_pgr_4	1.131138	0.003115	1.125496	1.13702	4.5e-05	3.2e-05	4818.128916	2904.507835	0.999749
mu_pgr_5	1.153408	0.003032	1.147475	1.158872	4.5e-05	3.2e-05	4589.28719	2752.672178	0.99987
mu_pgr_6	1.140028	0.003057	1.134264	1.145829	4.3e-05	3e-05	5097.663109	3399.085737	1.000504
sigma_pgr	0.013672	0.000837	0.01226	0.015302	1.2e-05	9e-06	4723.410382	2695.739871	1.000883

D

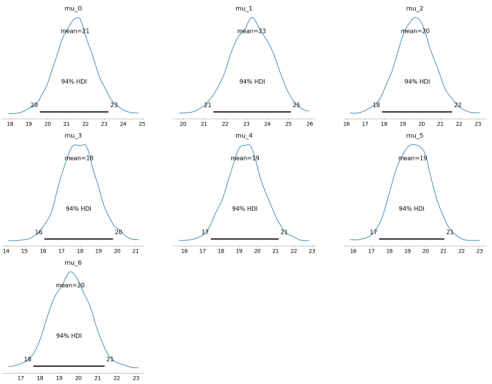

Fig. A.12.: Bayesian Statistics Replicate Blue. A) Mean B) Emt50 C) PGR D) Mean Plotted

A

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
mu_0	21.756	1.000	19.906	23.614	0.016	0.012	3890.0	1888.0	1.0
mu_1	22.633	1.013	20.723	24.530	0.016	0.012	3896.0	1815.0	1.0
mu_2	19.839	1.017	17.888	21.671	0.015	0.011	4671.0	2564.0	1.0
mu_3	17.942	0.983	16.117	19.745	0.014	0.010	4777.0	2829.0	1.0
mu_4	19.558	1.020	17.686	21.453	0.016	0.011	4158.0	2228.0	1.0
mu_5	19.346	0.998	17.468	21.248	0.015	0.011	4216.0	2128.0	1.0
mu_6	19.656	1.008	17.709	21.415	0.017	0.012	3461.0	2461.0	1.0

B

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
mu_0	21.999	1.002	20.032	23.796	0.014	0.010	4771.0	2398.0	1.0
mu_1	22.981	1.004	21.120	24.832	0.016	0.011	4026.0	2604.0	1.0
mu_2	20.005	1.009	18.089	21.855	0.016	0.011	4157.0	2526.0	1.0
mu_3	18.011	0.977	16.130	19.868	0.016	0.011	3797.0	2627.0	1.0
mu_4	19.991	1.000	18.079	21.759	0.015	0.010	4671.0	2523.0	1.0
mu_5	18.999	0.974	17.212	20.903	0.013	0.009	5414.0	3166.0	1.0
mu_6	19.024	0.988	17.133	20.855	0.014	0.010	4879.0	2560.0	1.0

C

index	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
mu_pgr_0	1.006506	0.003579	0.999835	1.013252	5.3e-05	3.8e-05	4501.5556	3110.161576	1.002777
mu_pgr_1	1.027794	0.003712	1.020715	1.034761	5.2e-05	3.7e-05	5135.368113	3177.595147	1.000013
mu_pgr_2	1.080842	0.003554	1.074022	1.087375	4.6e-05	3.2e-05	6078.302966	3320.971131	1.001379
mu_pgr_3	1.14124	0.003511	1.135091	1.148151	4.9e-05	3.5e-05	5136.634571	2903.507589	1.000088
mu_pgr_4	1.095564	0.00371	1.088715	1.102542	5e-05	3.5e-05	5535.546134	2818.086611	0.999865
mu_pgr_5	1.109431	0.003644	1.102689	1.116347	5.2e-05	3.7e-05	4866.158927	3094.634816	1.000063
mu_pgr_6	1.162124	0.003519	1.155753	1.168881	4.6e-05	3.2e-05	5887.149964	3348.739465	1.000513
sigma_pgr	0.016009	0.001006	0.014205	0.017886	1.5e-05	1.1e-05	4527.883391	2987.293116	1.000601

D

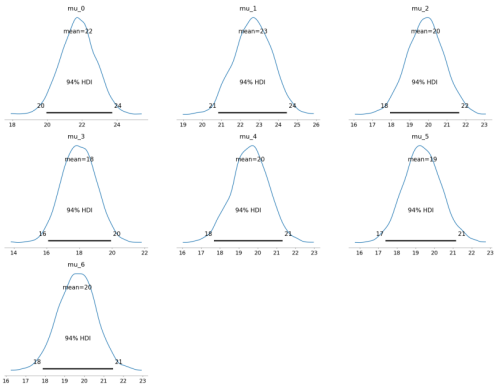

Fig. A.13.: Bayesian Statistics Replicate Gold. A) Mean B) Emt50 C) PGR D) Mean Plotted

A

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
mu_0	20.849	0.992	18.994	22.741	0.016	0.011	4004.0	2204.0	1.0
mu_1	23.066	1.014	21.150	24.916	0.016	0.012	3788.0	1907.0	1.0
mu_2	19.770	0.995	17.890	21.637	0.014	0.010	4904.0	2813.0	1.0
mu_3	17.944	1.002	16.103	19.792	0.015	0.011	4319.0	2765.0	1.0
mu_4	19.326	0.992	17.451	21.143	0.016	0.012	3945.0	2192.0	1.0
mu_5	18.991	1.012	17.167	20.881	0.015	0.011	4640.0	2420.0	1.0
mu_6	19.471	0.987	17.655	21.425	0.013	0.009	5633.0	2460.0	1.0

B

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
mu_0	20.995	0.975	19.090	22.725	0.017	0.012	3265.0	1983.0	1.0
mu_1	22.969	0.992	21.155	24.822	0.016	0.011	3825.0	1929.0	1.0
mu_2	19.998	1.006	18.281	22.103	0.015	0.011	4388.0	2907.0	1.0
mu_3	17.965	0.988	16.027	19.751	0.016	0.011	3827.0	2499.0	1.0
mu_4	18.982	0.995	17.277	20.987	0.017	0.012	3557.0	2890.0	1.0
mu_5	19.000	0.987	17.215	20.874	0.014	0.010	5153.0	3016.0	1.0
mu_6	18.994	0.996	17.218	20.929	0.015	0.011	4499.0	2763.0	1.0

C

index	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
mu_pgr_0	1.00709	0.004131	0.99926	1.014839	5.7e-05	4e-05	5292.103194	3056.822297	0.999888
mu_pgr_1	1.060631	0.004147	1.052697	1.068249	5.8e-05	4.1e-05	5136.838044	3169.905989	1.000436
mu_pgr_2	1.117194	0.004195	1.109334	1.12528	6.2e-05	4.4e-05	4636.698937	2614.17256	1.000679
mu_pgr_3	1.162981	0.004187	1.154705	1.170467	5.7e-05	4e-05	5397.548917	2932.875762	1.001303
mu_pgr_4	1.155561	0.004149	1.147511	1.163008	5.8e-05	4.1e-05	5052.081325	2750.280883	1.000148
mu_pgr_5	1.140293	0.004071	1.132441	1.147866	5.6e-05	3.9e-05	5338.34821	2609.412493	1.000012
mu_pgr_6	1.158217	0.004253	1.150003	1.165926	6.1e-05	4.3e-05	4850.446648	2505.957601	1.000716
sigma_pgr	0.018704	0.001134	0.016753	0.020896	1.7e-05	1.2e-05	4500.223039	3220.826617	0.999571

D

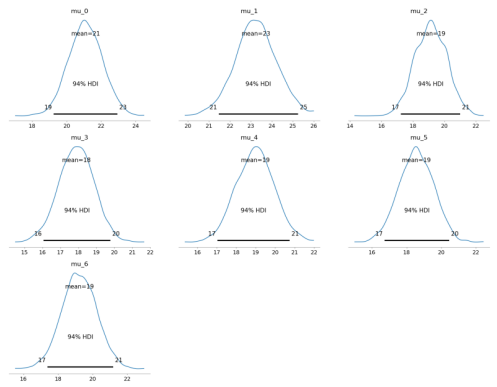

Fig. A.14.: Bayesian Statistics Replicate Green. A) Mean B) Emt50 C) PGR D) Mean Plotted

Bayes' Theorem

Bayesian theorem is a fundamental concept in probability theory and statistics, demonstrating the relationship between conditional probabilities and marginal probabilities for random variables [232]. Considering a sample $X = \{x_1, x_2, \ldots, x_n\}$ where each component represents values from a set of n attributes. In Bayes' Theorem, X is seen as the "evidence." Let H represent a

hypothesis, such as the data X belongs to a specific class C. In classification tasks, our objective is to ascertain $P(H|X)$, the probability of the hypothesis H being true given the observed data sample X. Essentially, we seek to determine the probability that the sample X belongs to H. Applied to our data set, we could designate classes as $c = 1$ for non-anomalous data and $c = -1$ for anomalous data, with X representing allele frequencies Similarly, $P(X|H)$ is the probability of X conditioned on H. That is, it is the probability that a data point X shows certain allele frequencies, given that we know the data point belongs to a known class. According to Bayes' Theorem, the probability to compute $P(H|X)$ can be expressed in terms of probabilities $P(H)$, $P(X|H)$ and $P(X)$ as:

A.0.1 Genetic Diversity Analysis

Tab. A.1.: Quantile thresholds of -log10(p-values) for allele frequency changes

(a) Generation 1 vs. Generation 7

Quantile	Red		Blue		Gold		Green	
	Obs	Sim	Obs	Sim	Obs	Sim	Obs	Sim
95%	3.037	2.071	3.083	3.048	2.478	1.966	3.848	3.293
99%	7.180	6.065	6.078	5.634	5.520	4.888	6.133	5.407
99.9%	10.782	9.494	9.147	8.798	8.218	7.766	9.242	8.572
99.99%	13.667	12.999	11.767	12.058	10.669	10.588	12.511	11.784
99.999%	16.199	16.735	14.086	15.549	12.864	13.402	15.790	15.260

(b) Generation 1 vs. Generation 4

Quantile	Red		Blue		Gold		Green	
	Obs	Sim	Obs	Sim	Obs	Sim	Obs	Sim
95%	2.526	2.023	2.823	1.926	2.902	1.698	2.705	1.830
99%	4.015	3.360	4.412	3.182	4.614	2.794	4.524	3.069
99.9%	6.068	5.396	6.625	5.119	7.013	4.486	7.135	5.062
99.99%	8.057	7.493	8.808	7.248	9.481	6.285	9.708	7.216
99.999%	9.936	9.591	11.106	9.293	12.136	8.128	12.000	9.656

Tab. A.2.: Quantile thresholds of -log10(p-values) for allele frequency changes comparing generation 1 to generation 7 (a) and generation 1 to generation 4 (b). Observed data (Obs) represents the experimntal data. Simulated data (Sim) represents the expected outcome of genetic drift under neutral selection (Fisher-Wright Model)

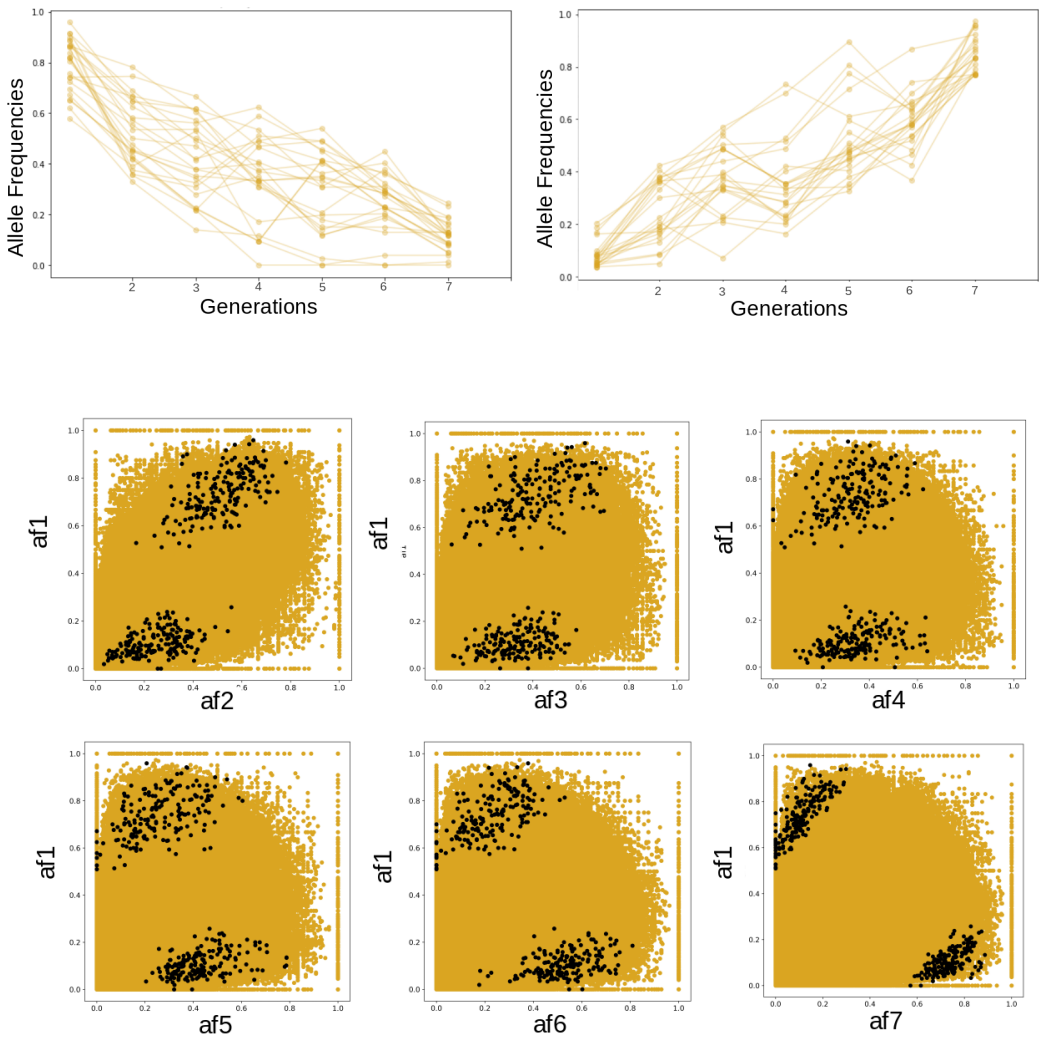

Fig. A.15.: Allele Frequency (AF) trajectories over all generation of reduced significant anomalies in replicate Gold. Upper panel: The left panel shows AF that appeared high in the ancestral generation and increased over time, the right panel shows the opposite cases. Down panel: Allele Frequency ancestral (af1) against following generations (af*), with reduced significant anomalies highlighted as black dots

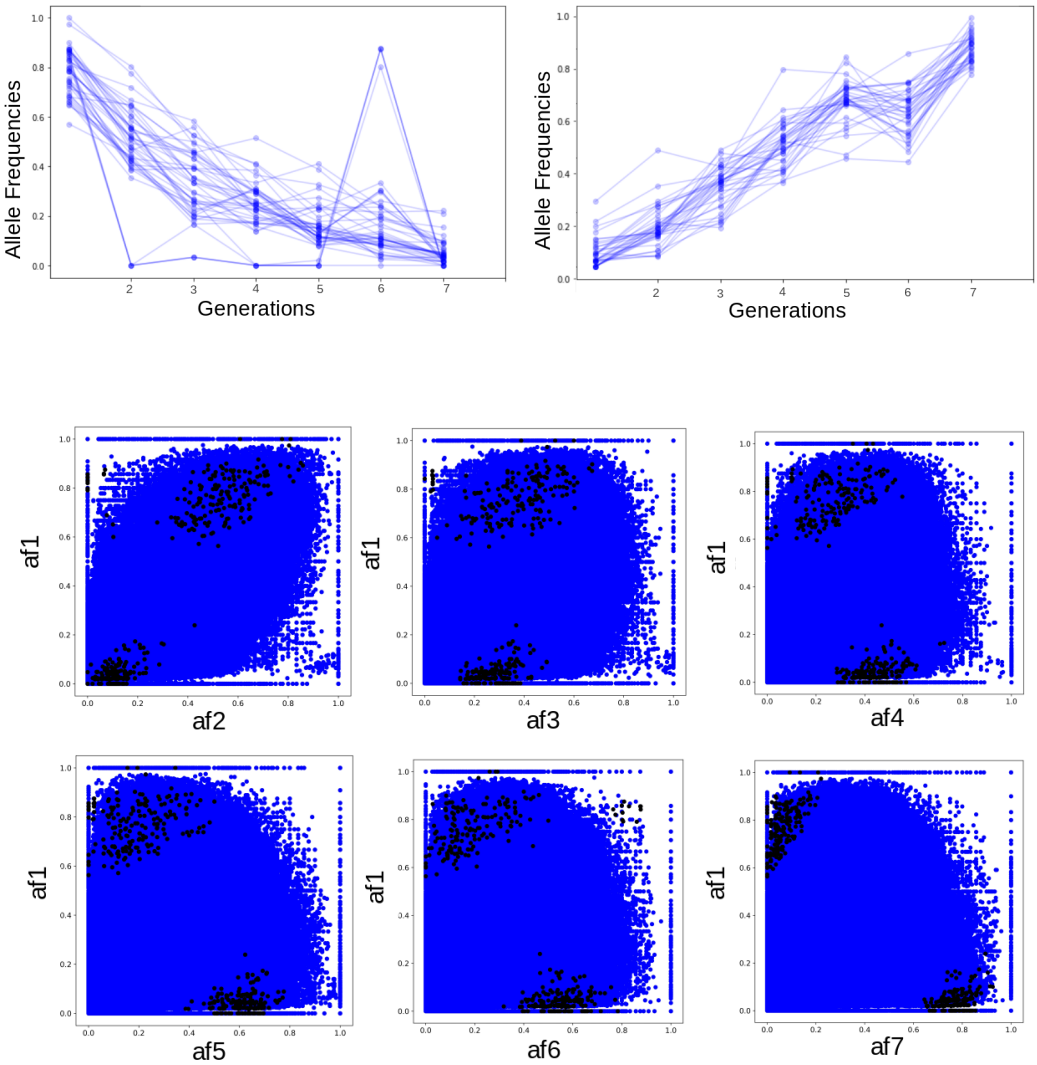

Fig. A.16.: Allele Frequency (AF) trajectories over all generation of reduced significant anomalies in replicate Blue. Upper panel: The left panel shows AF that appeared high in the ancestral generation and increased over time, the right panel shows the opposite cases. Down panel: Allele Frequency ancestral (af1) against following generations (af*), with reduced significant anomalies highlighted as black dots

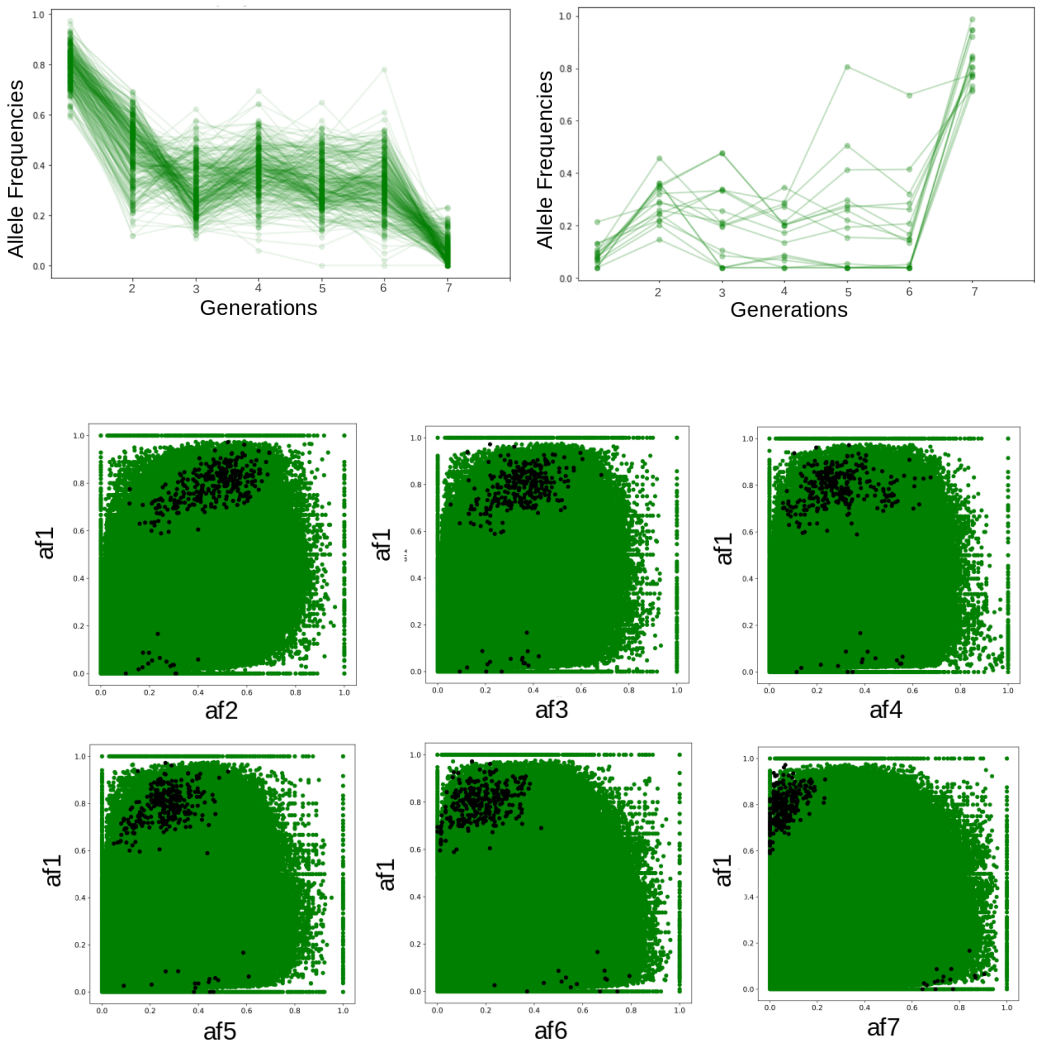

Fig. A.17.: Allele Frequency (AF) trajectories over all generation of reduced significant anomalies in replicate Green. Upper panel: The left panel shows AF that appeared high in the ancestral generation and increased over time, the right panel shows the opposite cases. Down panel: Allele Frequency ancestral (af1) against following generations (af*), with reduced significant anomalies highlighted as black dots

Tab. A.3.: Comparison of genetic diversity measures between Generation 1 and Generation 7 in broad-effect variants ($p < 0.001$)

		Tajima's pi				Watterson's $theta$				Tajima's D		
Replicate	Gen	Value	V Stat	P Value	Effect	Value	V Stat	P Value	Effect	Value	V Stat	P Value
Red	1 7	0.022±0.011 0.017±0.012	2.43e8	2.2e-16	0.440	0.021±0.012 0.016±0.010	2.48e8	2.2e-16	0.451	0.17±0.65 0.51±0.84	2.15e8	2.2e-16
Blue	1 7	0.016±0.010 0.016±0.010	1.84e8	0.049	0.006	0.016±0.010 0.015±0.009	2.08e8	2.2e-16	0.068	0.37±0.70 0.05±1.03	1.24e8	2.2e-16
Gold	1 7	0.015±0.010 0.013±0.010	4.32e8	2.2e-16	0.195	0.014±0.009 0.012±0.009	4.45e8	2.2e-16	0.166	0.61±0.68 0.50±0.77	3.48e8	2.2e-16
Green	1 7	0.017±0.009 0.015±0.010	5.23e8	2.2e-16	0.208	0.016±0.008 0.014±0.009	5.68e8	2.2e-16	0.232	0.35±0.58 0.46±0.90	3.63e8	2.2e-16

Tab. A.4.: Tajima's π, Watterson's θ and Tajima's D metrices for broad-effect variants. The table shows V-Statistic, p-Value (Mann-Whitney U) and Effect Size (Cohen's D)

Tab. A.5.: Comparison of genetic diversity measures for strong-effect variants (0.0001% tail)

		Tajima's π				Watterson's θ				Tajima's D		
Replicate	Gen	Value	V Stat	P Value	Effect	Value	V Stat	P Value	Effect	Value	V Stat	P Value
Red	1 7	0.013±0.004 0.017±0.001	0	0.5	1.13	0.014±0.006 0.018±0.001	1	1.0	0.90	-0.54±0.26 -0.11±0.59	0	0.5
Blue	1 7	0.012±0.003 0.008±0.002	170	2e-4	1.45	0.011±0.003 0.008±0.002	171	1.7e-4	0.95	0.52±0.47 0.09±0.53	168	2.9e-4
Gold	1 7	0.016±0.007 0.014±0.006	156	0.059	0.35	0.015±0.007 0.013±0.006	143	0.161	0.21	0.69±0.47 0.32±0.63	154	0.070
Green	1 7	0.024±0.008 0.003±0.004	5778	2.2e-16	2.61	0.021±0.006 0.003±0.004	5778	2.2e-16	2.35	0.77±0.35 -1.40±1.17	5591	2.2e-16

Tab. A.6.: Tajima's π, Watterson's θ and Tajima's D metrices for strong-effect variants. The table shows V-Statistic, p-Value (Mann-Whitney U) and Effect Size (Cohen's D)

Tab. A.7.: Kruskal-Wallis Test for all Replicates

Comparison	H Statistic	P Value	Effect Size (eta-squared)
Gen. 1 vs. Gen. 2	78638.06	<0.0001	0.131
Gen. 1 vs. Gen. 3	41833.50	<0.0001	0.070
Gen. 1 vs. Gen. 4	26432.43	<0.0001	0.044
Gen. 1 vs. Gen. 5	39686.41	<0.0001	0.066
Gen. 1 vs. Gen. 6	30803.70	<0.0001	0.051
Gen. 1 vs. Gen. 7	20861.45	<0.0001	0.035

Tab. A.8.: Kruskal-Wallis Test for F_{ST} Values. The table shows statistical comparisons of F_{ST} values between 4 replicate groups using the Kruskal-Wallis test (including H-statistics and p-values) along with eta-squared values to measure effect size.