UNIWERSYTET IM. A. MICKIEWICZA W POZNANIU

WYDZIAŁ BIOLOGII

PRACOWNIA BIOLOGII EWOLUCYJNEJ

UWARUNKOWANIA OBCIĄŻENIA GENETYCZNEGO W DZIKICH POPULACJACH GUPIKÓW *POECILIA RETICULATA*

KATARZYNA BURDA

ROZPRAWA DOKTORSKA

Promotor

prof. dr hab. Jacek Radwan

Promotor pomocniczy

dr Mateusz Konczal

POZNAŃ, 2024

ADAM MICKIEWICZ UNIVERSITY

FACULTY OF BIOLOGY

EVOLUTIONARY BIOLOGY GROUP

DETERMINANTS OF GENETIC LOAD IN NATURAL POPULATIONS OF GUPPY (*POECILIA RETICULATA*)

KATARZYNA BURDA

DOCTORAL THESIS

Supervisor

prof. dr hab. Jacek Radwan

Auxiliary Supervisor

dr Mateusz Konczal

POZNAŃ, 2024

Acknowledgments

Najważniejsza osoba, której pragnę podziewać w tym miejscu to mój promotor pomocniczy – dr (wkrótce hab.) Mateusz Konczal. Gdyby nie on, cała moja gupikowa przygoda w ogóle by się nie zaczęła. Długo nie mogłam uwierzyć, że w ogóle jestem na doktoracie! Przez całe pięć lat pomagał mi poznać genomy i zrozumieć ewolucję na różnych poziomach, jest dla mnie ogromnym naukowym autorytetem. Był także obecny w moich momentach kryzysowych, których przez wzgląd na mój charakter było, cóż, bardzo dużo. Uspokajał mnie i dawał mi oparcie, o którym wielu doktorantów może tylko pomarzyć. Jego pozytywne nastawienie i spokój łagodziło mój chaos i chłodziło moje emocje. Dziękuję mu z całego serca, tylko dzięki jego nieustannej pomocy udało się dotrzeć do tego momentu.

Kolejne osoby, które przyczyniły się do tego, że jestem w tym miejscu to członkowie Pracowni Biologii Ewolucyjnej. Byłam zupełnie zaskoczona z jak ciepłym przyjęciem się spotkałam już na samym początku. Kierownik pracowni – mój promotor – prof. dr hab. Jacek Radwan, choć sam nie brał udziału w mojej rekrutacji, zaakceptował moją osobę i zawsze był wobec mnie cierpliwy i wyrozumiały, za co bardzo dziękuję. Pragnę również wspomnieć o Agnieszce Szubert-Kruszyńskiej, której łagodność i spokój wielokrotnie ratowały moje nerwy. Choć mój doktorat przypadł w dużym stopniu na czasy koronawirusa, członkowie Pracowni sprawili, że te pięć lat było naprawdę niesamowite. Chciałam szczególnie podziękować Sebastianowi Chmielewskiemu, który zawsze pomagał mi w naprawdę czarnych godzinach, czy to w szukaniu błędów w R-ze, czy też rozwiązywaniu zagwozdek dotyczących tego, które pismo powinnam złożyć. Razem narzekaliśmy, razem się uczyliśmy i razem zachwycaliśmy się nad złożonością świata oraz zaawansowaniem nowoczesnych technologii. Bez Sebastiana, byłoby mi niewyobrażalnie trudniej.

I want to express my gratitude to Karolina Przesmycka and Johnatan Parrett. We have spent countless hours playing board games, listening to prog rock, joking and discussing whatever was on our minds. These sessions were extremely important, keeping my mind at balance and bringing me tons of joy. I already miss it tremendously and I do hope we can meet like this in the future.

Wychodząc poza krąg naukowy, pragnę szczególnie podziękować Sebastianowi Fudali, który specjalnie dla mnie przeprowadził się z Krakowa do Poznania, by towarzyszyć mi w mojej drodze. To było wielkie poświęcenie, na które mało kto jest gotów! Dziękuję za Twoje poczucie humoru, za zdolność chłodnej kalkulacji i spokój w obliczu moich kryzysowych wybuchów, oraz, co również niesamowicie ważne, za obiady! Nasze wspólne wspomnienia z odkrywania Poznania będą towarzyszyć mi na zawsze, bez Ciebie to nie to samo.

Dziękuję moim przyjaciołom i bliskim znajomym. Przede wszystkim Agacie Kuładze, Agnieszce i Maciejowi Staniszom oraz Stefanowi Gorzelnikowi za to, że zawsze mogłam się do Was wyżalić, za godziny inspirujących rozmów na żywo i przez Internet, wspólne gry i spotkania. Dziękuję, że jesteście i mam ogromną nadzieję, że tak już zostanie.

Na koniec, dziękuję mojej rodzinie, Cioci Maryli, Mamie Ewie, Siostrze Gosi i Tacie Robertowi za wsparcie emocjonalne i finansowe (pod koniec). Zawsze wierzyliście, że uda mi się dotrzeć do końca tej drogi, mimo że ja sama miałam więcej pesymizmu niż nadziei. Nigdy nie wątpiliście w moje powodzenie, również rozumiejąc moje trudności związane z charakterem.

Contents

Works included in the dissertation

The thesis consists of 2 chapters. The first chapter is a published work. The second chapter is an unpublished draft of the manuscript.

1. Burda, K., & Konczal, M. (2023). Validation of machine learning approach for direct mutation rate estimation. *Molecular ecology resources*, *23*(8), 1757–1771. https://doi.org/10.1111/1755-0998.13841
2. Burda, K., Mohammed R., Janecka M., Clark D., Kramp R., Radwan J., Konczal, M, (2024) Genetic load is affected by demographic histories in Trinidadian guppies (*Poecilia reticulata*), but does not explain invasiveness after a recent artificial translocation.

Funding

This research was supported by National Science Center Sonata grant 2018/31/D/NZ8/00091 awarded to Mateusz Konczal.

Streszczenie

Termin ‘obciążenie genetyczne’ opisuje występującą w populacji szkodliwą zmienność, która może prowadzić do obniżenia obecnego dostosowania w populacji, lub negatywnie wpłynąć na dostosowanie przyszłych pokoleń. Niektóre procesy demograficzne mogą prowadzić do akumulacji szkodliwych wariantów, szczególnie w przypadku efektu wąskiego gardła (ang. ‘bottleneck’), gdy wielkość populacji zostaje intensywnie zmniejszona. Gromadzenie obciążenia genetycznego w takim scenariuszu wynika z tego, że w małych populacjach losowe siły dryfu przeważają nad siłami doboru naturalnego, więc częstość mutacji o niskiej szkodliwości może rosnąć, a nawet może dojść do ich utrwalenia. Populacje, które przeszły przez wąskie gardło są również narażone na wzrost homozygotyczności poprzez nielosowe kojarzenia, których nie da się uniknąć przy silnie ograniczonej liczbie osobników. Jednakże, takie krzyżowania mogą nieść ze sobą także pozytywne skutki. Jeśli wysoko szkodliwy wariant jest recesywny, jego obecność w homozygocie powoduje ujawnienie go dla doboru oczyszczającego i w efekcie zmniejszenie jego częstości w populacji. Relatywna rola obniżonej siły doboru i czyszczenia populacji z recesywnych wariantów jest wciąż tematem naukowej dyskusji, szczególnie w kontekście inwazji biologicznych i gatunków zagrożonych wyginięciem. Większość gatunków szczególnej troski charakteryzuje się małą wielkością populacji i niską zmiennością genetyczną, ale także historią wąskiego gardła, która mogła doprowadzić do akumulacji obciążenia genetycznego, lecz również oczyszczenia populacji z wysoce szkodliwych wariantów. Innym ważnym zjawiskiem związanym z gromadzeniem mutacji potencjalnie obniżających dostosowanie, jest biologiczna inwazja. Jako że początek inwazji zwykle łączy się z małą liczbą osobników, na drodze ekspansji populacja może akumulować szkodliwe warianty o niskim efekcie, ale też oczyszczać się z wysoce niekorzystnych mutacji, zwiększając tym samym swój potencjał inwazyjny. Zjawisko dobrego prosperowania i odnoszenia przewagi względem gatunków rodzimych pomimo niskiej zmienności u populacji inwazyjnych nosi miano paradoksu genetycznego i jest szeroko debatowane w środowisku naukowym.

W mojej rozprawie doktorskiej zamierzam wnieść wkład do tej dyskusji poprzez eksplorację czynników obciążenia genetycznego w naturalnych populacjach gupika (*Poecilia reticulata*). Ponieważ zarówno samo obciążenie genetyczne, jak i wnioskowanie o demograficznych historiach populacji, ściśle zależy od intensywności pojawiania się nowych mutacji, najpierw szacuję tempo mutacji tego gatunku. Następnie, eksploruje genetyczne przyczyny i konsekwencje inwazji populacji w jednej z badanych rzek, a także badam obciążenie genetyczne porównując wyspy (Trinidad i Tobago) oraz lokacje w obrębie tych samych rzek. W związku z powyższym, rozprawa składa się z dwóch rozdziałów.

W rozdziale pierwszym oszacowałam tempo mutacji na pozycję w genomie, na pokolenie u gupika. W tym celu użyłam dwóch rodzin, o łącznej liczbie dwudziestu czterech osobników, które zostały resekwencjonowane do wysokiego pokrycia – około 47x. Następnie zidentyfikowałam warianty, które są obecne u potomstwa, a których nie ma u żadnego z rodziców. Jako że sekwencjonowanie nowej generacji jest podatne na błędy, kandydackie mutacje *de novo* zostały kolejno przefiltrowane na podstawie kryteriów ułatwiających odróżnienie prawdziwych wariantów. Ponieważ standardowe filtrowanie jest subiektywne, zdecydowałam się również przetestować alternatywną technikę filtrowania polegającą na

uczeniu maszynowym. W tym celu, najpierw skompletowałam zestaw treningowy prawdziwych i fałszywych kandydatów *de novo*, sprawdzonych przeze mnie przy użyciu sekwencjonowania metodą Sangera. Posłużył on do nauczenia modelu czym charakteryzują się prawdziwe mutacje. Kolejno model przefiltrował pozostałych kandydatów *de novo*. Wszystkie warianty wskazane przez obie metody finalnie również zostały sprawdzone poprzez sekwencjonowanie metodą Sangera. Po porównaniu efektywności standardowego filtrowania i filtrowania przy użyciu uczenia maszynowego, okazało się, że obie metody oszacowały podobne tempo mutacji, ale mocno różniły błędami pierwszego i drugiego rodzaju. Metoda wspierana uczeniem maszynowym była w stanie odnaleźć więcej prawdziwych mutacji *de novo*, jednak przygotowanie zestawu testowego było bardzo wymagające, zarówno pod względem czasu jak i wysiłku. Finalnie, oszacowane przeze mnie tempo mutacji gupika jest niskie - 2.9 x 10^{-9} (95% przedział ufności: 1.92-3.88 x 10^{-9}), co upodabnia je do innych ryb kościstych (*Teleostei*).

Drugi rozdział mojej rozprawy eksploruje obciążenie genetyczne w czternastu dzikich populacjach gupika na Trynidadzie i Tobago, ze szczególnym uwzględnieniem rzeki Turure gdzie trwa ekspansja w dół rzeki. Jest ona wynikiem sztucznej introdukcji gupików do miejsca w pobliżu źródła, gdzie wcześniej nie występowały. Wprowadzone gupiki rozpoczęły ekspansję zgodnie z nurtem, na której drodze spotkały lokalną populację i doszło do kojarzeń. Populacja imigrantów odniosła ogromny sukces i niemal całkowicie wyparła oryginalnych mieszkańców dolnego biegu rzeki. W celu zbadania, czy ekspansja łączyła się z akumulacją obciążenia genetycznego, oraz czy doszło do oczyszczenia populacji z wysoce szkodliwych wariantów w początkowej fazie, przeprowadziłam eksplorację trzech miejsc na różnych odcinkach rzeki. Dodatkowo, przeanalizowałam obciążenie genetyczne w populacjach różniących się licznością osobników, kontrastując duże trynidadzkie populacje z małymi populacjami na Tobago, oraz populacje zamieszkujące górne biegi rzek Quare i Oropouche z populacjami pochodzącymi z dolnego biegu tych rzek. Finalnie, sprawdziłam czy istnieje zależność między neutralną zmiennością genetyczną a obciążeniem mutacyjnym w badanych przeze mnie populacjach. W moich analizach resekwencjonowałam 190 osobników, które następnie użyłam do wykonania analizy głównych składowych oraz do badania admiksji. Następnie, poza standardowym progami jakościowymi, przefiltrowałam moje dane pod względem obecności danych dotyczących alleli ancestralnych oraz stopnia zakonserwowania (CS, ang. conservation score). Kolejno, wykorzystując otrzymane w Rozdziale I tempo mutacji, oszacowałam efektywną wielkość populacji (N_e) dla każdej z badanych populacji. Następnie, dla wszystkich osobników, obliczyłam obciążenie genetyczne (na podstawie CS) oraz ustaliłam liczbę nonsensownych mutacji niesynonimowych (LOF, utrata funkcji, ang. loss of function). Populacje z dolnego i środkowego biegu rzeki Turure były mniej obciążone genetycznie niż populacje z górnego biegu, nie można zatem wysnuć wniosku o akumulacji obciążenia wzdłuż drogi ekspansji. Obserwację tą można tłumaczyć admiksją między imigrantami a lokalnymi gupikami, która złagodziła początkowe efekty głębokiego gardła. Populacja z górnego biegu rzeki Turure miała najwyższe obciążenie genetyczne pośród wszystkich lokacji w Turure oraz taką samą liczbę mutacji o wysokim efekcie, nie można zatem stwierdzić, że początkowe fazy inwazji wiązały się z oczyszczeniem populacji z wysoko szkodliwych mutacji. W przypadku populacji z innych rzek, zaobserwowałam, że małe, odizolowane populacje z Tobago mają istotnie więcej obciążenia niż gupiki z Trynidadu. Podobna obserwacja dotyczy wyżej obciążonych populacji z górnego biegu rzek Quare i Oropouche w porównaniu do ich odpowiedników z dolnego biegu. W tym wypadku jednak, różnice w efektywnej wielkości nie

tłumaczą różnic w obciążeniu, ponieważ populacja z górnego Quare ma dużo większą N_e przy dużo wyższym obciążeniu. Związku między neutralną zmiennością genetyczną (N_e), a obciążeniem genetycznym, nie można również wyciągnąć na podstawie analizy populacji z większości badanych rzek. Można to tłumaczyć tym, że omawiane populacje są relatywnie duże, co może zacierać taki efekt.

Podsumowując, niniejsza rozprawa wskazuje, że obciążenie genetyczne akumuluje się w populacjach, które przeszły przez wąskie gardło, jednak przepływ genów może złagodzić ten proces i zmniejszyć zależność między obciążeniem a efektywną wielkością populacji. Pokazuje ona również, że gupiki, podobnie do innych ryb kościstych charakteryzują się bardzo niskim tempem mutacji, co może znacząco wpłynąć na ilość zmienności genetycznej w tym gatunku.

Słowa kluczowe: tempo mutacji, gupik, obciążenie genetyczne, ekspansja

Summary

The term genetic load describes deleterious variation in a population which causes fitness decrease in the present or potentially might cause such decrease in the future. Some demographic process can lead to accumulation of harmful variants, most commonly, when a bottleneck takes place and population size is reduced. This happens because in small populations random forces of drift dominate the natural selection, so slightly deleterious alleles can increase in frequency and ultimately fix. Bottlenecked populations are also characterized by increased homozygosity due to inbreeding, inevitable in small populations. Inbreeding might have, however, a positive effect on population fitness. If highly deleterious mutations are recessive, it can expose them to the purifying selection in homozygous genotypes, effectively decreasing their frequency in the population. A relative role of relaxed selection and purging of recessive variants is a matter of scientific debate. In particular, its role in both invasive species and species of conservation concern is widely discussed. Among species listed as endangered or vulnerable, majority has small census size and low genetic diversity. Most of them share common history of single or series of bottlenecks, which as described above, could have led to genetic load accumulation, but also to purging. Another important issue regarding genetic load is its role in biological invasions. Since the beginning of species invasion is marked with small population size, it might accumulate deleterious variation on the spread axis. However, it may also purge harmful variants, cleansing the population and increasing its invasive potential. The phenomenon of having low diversity, small size and being extremely successful in the new niche is called a genetic paradox of invasions and is heavily debated in scientific discourse.

In my PhD dissertation, I aim at contribution to this debate by exploring determinants of genetic load in natural populations of wild Trinidadian guppies (*Poecilia reticulata*). As the genetic load itself and genetic inferences of demographic histories depend on mutation rate, I first directly estimate mutation rate in this species. Then, I study genomic causes and consequences of population invasion in one of the rivers and investigate load in comparison of islands and different locations within rivers. Consequently, the thesis is divided into two separate chapters.

In the first chapter, I estimated per site, per generation mutation rate in guppies. The two families with twenty four individuals in total were used for whole genome resequencing to high, approximately 47x, coverage. Subsequently, reads were mapped to the reference genome and used to identify variants that are present in one or more offspring, but absent in both of the parents. Since new generation sequencing is error prone, such candidate *de novo* mutations were then filtered based on a set of criteria to distinguish false positives from false negatives. As such filtering criteria are subjective, I decided to separately test alternative approach, based on machine learning. Using a subset of true positives and false positives *de novo* mutations, validated with Sanger sequencing, I trained the model and used it to filter the rest of the data. Later, all candidate *de novo* mutations were molecularly validated with Sanger sequencing. After comparing effectiveness of the two approaches, I found that both methods estimated similar mutation rate, but their false negative and false discovery rates differed substantially. The machine learning approach found more true positive *de novo* mutations, but producing training set was time and effort-consuming. Finally, I found that guppy mutation rate is low –

2.9 x 10^{-9} (95% confidence interval: 1.92-3.88 x 10^{-9}), which is in compliance to other Teleostei fish estimates.

The second chapter of my thesis explores genetic load in 14 wild guppy populations from Trinidad and Tobago. I particularly focused on Trinidadian river Turure, where ongoing expansion is taking place. It is a result of an artificial introduction of guppies into a previously guppy-free location in the upstream, which was followed by a downstream propagation and admixture with native individuals in the lower parts. The expanding population was extremely successful and almost completely replaced original inhabitants of downstream Turure. To find whether the invasion facilitated purging of genetic load during initial bottleneck, or on contrary, was followed by accumulation of genetic load, I studied three sites along the river. I also explored genetic load of populations differing in population size, contrasting large populations from Trinidad with small populations on Tobago as well as upstream and downstream rivers locations. Finally, I looked for relationship between neutral genetic diversity and genetic load across populations. To achieve it, I resequenced whole genomes of 190 individuals. In order to determine genetic structure of guppy populations, I performed principal component and admixture analyses. Then, apart from standard quality thresholds, I also applied ancestral allele (to determine derived state) and conservation score (CS) filters. Subsequently, from this data and based on mutation rate obtained in Chapter I, I estimated effective population size for each population. Next, for each individual, I estimated genetic load based on conservation scores and screened the samples for high effect deleterious mutations (LOF, loss of function). Downstream and middle Turure populations were less loaded in terms of CS-based load than the population from upstream location, therefore I found no proof of genetic load accumulation along the spread axis. This observation could be explained by admixture between migrants and local guppies, which alleviated effects of initial bottleneck. The upstream population had the highest load among all Turure locations and similar amount of high effect variants, so there was no indication of purging during the initial stage of invasion. When it comes to populations from other rivers, I observed that small, isolated populations from Tobago and big populations from Trinidad differ in terms of genetic load, with the former having higher burden. Similar situation was found in more loaded post-bottleneck upstream populations and less loaded downstream populations within the same rivers. In this case, however, the difference in effective population size failed to explain the difference in load, since Quare upstream site had a strikingly higher genetic diversity than the downstream location. Such size-dependent conclusion could not have been made generally as well, since no relationship between neutral genetic diversity and genetic load was found across majority of populations. This however, could have been due to relatively large effective population sizes of studied populations, which might have weakened the effect.

Overall, this thesis indicates that while genetic load tends to accumulate in bottlenecked populations, gene flow can significantly relieve the burden and loosen association of the load with N_e. It also demonstrates that guppies, like other Teleostei, are characterize by very low mutation rate, what can significantly affect amount of genetic variation present in this species.

Keywords: mutation rate, guppy, genetic load, expansion

General Introduction

Mutations are a fundamental source of diversity and an ultimate fabric for evolution to work on (Dobzhansky, 1937). Yet, beneficial alleles are rare and new variation is mostly neutral, or harmful (de Jong et al., 2024; Eyre-Walker & Keightley, 2007). Accumulation of deleterious variants in a population, which can lead to decrease in fitness, is defined as the genetic load (Muller, 1950). Even though interest in harmful mutations has long history (Haldane, 1937), relative roles of evolutionary forces shaping mutation load are still debated using both phenotypic analyses (Bundgaard et al., 2021; Lohr & Haag, 2015) and, more recently, genomic approaches (Mathur et al., 2023; von Seth et al., 2021). The rate at which mutation load accumulates in a population depends on the mutation rate, and on the two evolutionary forces that shape their fate: selection and genetic drift (Wright, 1932). These forces can cause the load to vary in time (Dussex et al., 2021) and space (Rougemont et al., 2020), and their relative importance can be studied by comparing populations of different demographic histories (Kleinman-Ruiz et al., 2022). Therefore, to investigate determinants of genetic load in the guppy fish (*Poecilia reticulata*), I estimated mutation rate and studied genetic load across dozens of populations.

Intensity with which new mutations appear between the generations (so called *de novo* mutations) is called mutation rate (MR) and it dictates the tempo of the evolutionary clock (Kimura, 1983; Kimura & Ohta, 1971). MR can differ considerably between species and evolve in response to selection (Lynch, 2010; Lynch et al., 2016). Depending on factors like environment (Saclier et al., 2020), genetic background (Aikens et al., 2019; Lau & Robinson, 2021), lifestyle (Kessler et al., 2020) and age (Kong et al., 2012), MR can also differ between individuals within one species. MR is one of the crucial parameters in population genomics required to estimate effective populations size, divergence between species and to parametrize population genetic models, and its precise estimates are of high importance for the scientific community (Scally, 2016). In the past, such estimates were based on substitution rate in selectively neutral regions and mutation accumulation experiments (Kondrashov & Kondrashov, 2010; Wu et al., 2024). Since new generation sequencing became more available, whole-genome pedigree-based methods have gained popularity and MRs for variety of vertebrate species have already been calculated using this methodology (Bergeron et al., 2023).

Nevertheless, estimation of mutation rate from a pedigree data remains a challenging task (Bergeron et al., 2022). *De novo* mutations are extremely rare (e.g. MR in humans is 1.22×10^{-8} [Kessler et al., 2020] per site, per generation and it is considered high), so screening for them is similar to looking for a needle in a haystack (Yoder & Tiley, 2021). Advances in technology and decreasing costs allow to resequence samples in very high coverage, but are still prone to errors, increasing probability of false positives (Farrer et al., 2013). To account for that, most commonly used quality filters are very rigid, allowing only variants of nearly perfect characteristics to pass. However, in context of MR studies, this poses a risk of mutation rate underestimation due to inability to find all *de novo* mutations (Bergeron et al., 2022). This challenge can be potentially overcome with improvement of methodological approaches, and especially growing field of machine learning can be employed to help distinguish true *de novo* mutations from false positive observations.

Mutation rate dictates how fast new mutations arise, but their future fate in a population depends on random (genetic drift) and directional (natural selection) forces. A scenario, when one of these forces - genetic drift - becomes particularly powerful, is a decrease of population size (bottleneck). Bottleneck increases probability of mating with relatives, increasing homozygosity and decreasing genetic diversity in the population (Wright, 1932). Furthermore, since ability to remove deleterious variation through selection depends on effective population size N_e (selection can overcome genetic drift if selection coefficient is larger than $1/4N_e$, [Kimura & Ohta, 1969]), in bottlenecked populations, it is seriously impaired and fails to remove slightly harmful variation. In consequence, drift can rapidly change frequency of those slightly deleterious alleles. They may become much more common or fixed, a process called 'gene surfing' (Peischl & Excoffier, 2015). Or oppositely, less frequent or entirely lost, which, in turn, could decrease the heterozygous, masked load (Robinson et al., 2023). Some low effect deleterious variants can rise in frequency also through the inbreeding, which eventually makes them appear more commonly in homozygotes, adding up to the realized load and likely leading to decrease in fitness (Charlesworth & Charlesworth, 1987). However, in contrary to slightly harmful variation, expression of highly deleterious variants in homozygotes, can also have a positive effect. It enables the purifying selection to act against recessive mutations (purging, [Charlesworth & Willis, 2009]) and reduce the realized load. If this process is effective, the population might recover. If not, lowered population condition may cause its further decline in size, which leads to even worse accumulation of genetic load and entering a path of mutational meltdown (Lynch et al., 1995).

Building up the load of deleterious variation and its impact on small populations is of great significance to biology. Human-related habitat loss or fragmentation, pollution, climate change and related extreme fires threaten myriad species with extinction (Johnson et al., 2017; Kelly et al., 2020). It has been suggested that conservation biology rescue attempts could benefit from informed decisions which, apart from genetic diversity information, also consider genetic load (Dussex et al., 2023). Hence, a great number of studies concerning vulnerable species are conducted (Al Hikmani et al., 2024; Dussex et al., 2023; Kuang et al., 2020; Ochoa et al., 2022; Smeds & Ellegren, 2023), and some provide practical solutions for minimizing offspring load in captive breeding (Speak et al., 2024).

Another threat to biodiversity - invasive species – is also strongly associated with dynamics of genetic load. Previously unprecedented mobility of humans around the globe and artificial introductions of alien species result in rise of invasive species (Capinha et al., 2015; Otto, 2018). Following the invasion, they threaten local biodiversity and cause a variety of financial losses to humans (Shackleton et al., 2019). One of invasion success components might be purging of genetic load during initial bottleneck associated with founder effect (Estoup et al., 2016) but its importance for invasive potential is still not clear (Lombaert et al., 2024).

This thesis explores mutation rate and genetic load in the guppy, a freshwater fish native to South America and Caribbean Islands. Guppies are a popular model in many biological fields: adaptation genomics (Fraser & Neff, 2010), sex chromosomes (Darolti et al., 2020; Qiu et al., 2022), host-parasite coevolution (Phillips et al., 2018), sexual selection (Brooks & Endler, 2001) and ethology (Earl et al., 2024; Herdegen-Radwan, 2019). One of the most heavily studied phenomena regarding Trinidadian guppies is their strong differentiation between upper and lower parts of the streams (Fischer et al., 2021; Ioannou et al., 2017; van der Zee et al., 2022). Due to waterfalls and varied predators presence, individuals from locations close to the river source exhibit low predation ecotype and populations close to the river confluence have high predation ecotype (ecotypes differ in e.g. body size, coloration, maturation time, Fitzpatrick et al., 2015). However, this is not the only difference between those populations. Upper sites guppies have lower neutral genetic diversity than their downstream counterparts and their census population sizes are also relatively smaller (Qiu et al., 2022; Whiting et al., 2021), which results from demographic histories probably associated with severe bottlenecks (Barson et al., 2009; Qiu et al., 2022). Finding whether size difference may be reflected in genetic load accumulation, is one of aspects covered in this thesis. Another contrast is performed between two islands, Trinidad and Tobago, with the former characterized by larger rivers hosting larger guppy population than the latter. Finally, the thesis takes an advantage of translocation event followed by invasion. Two major Trinidadian drainages – Caroni and Oropouche – have been separated for a very long time, namely since the Pleistocene era (Whiting et al., 2021). During this period, guppy populations have strongly differentiated between these two regions (Willing et al., 2010) to the extent that some researchers even claim that they have become two distinct species (Schories et al., 2009). In 1957 (approx. 130 guppy generations ago) Caryl Haskins moved about 200 individuals from lower Guanapo (Caroni drainage, high predation) to upper Turure (Oropouche drainage, low predation) (Endler, 1980). The upper Turure site had no guppies prior to this introduction, but downstream locations had already been inhabited (Magurran, 2005). When new population expanded, it eventually spread, met native Turure guppies and replaced them with limited degree of hybridization. Immigrants tremendous success can be observed in previous genetic analyses, demonstrating that translocated population almost completely replaced naïve Turure population along entire river (Sievers et al., 2012; Willing et al., 2010). In my PhD thesis, I explore genetic load of this population, looking for both accumulation and purging processes that may have influenced the invasion course.

Aims of the thesis

The aims of the presented work were:

1) Estimating mutation rate in guppy using genome sequencing of parents and their offspring
2) Validation of machine learning approach for direct mutation rate estimate studies
3) Exploration of accumulation of genetic load in the expanding population of guppies in Turure river.
4) Testing association between effective population size and genetic load in guppy populations.

The first two objective corresponds to Chapter I of the dissertation. The following two objectives are realized in Chapter II of the dissertation.

References

Aikens, R. C., Johnson, K. E., & Voight, B. F. (2019). Signals of Variation in Human Mutation Rate at Multiple Levels of Sequence Context. *Molecular Biology and Evolution*, *36*(5), 955–965. https://doi.org/10.1093/molbev/msz023

Al Hikmani, H., van Oosterhout, C., Birley, T., Labisko, J., Jackson, H. A., Spalton, A., Tollington, S., & Groombridge, J. J. (2024). Can genetic rescue help save Arabia's last big cat? *Evolutionary Applications*, *17*(5). https://doi.org/10.1111/eva.13701

Barson, N. J., Cable, J., & Van Oosterhout, C. (2009). Population genetic analysis of microsatellite variation of guppies (Poecilia reticulata) in Trinidad and Tobago: Evidence for a dynamic source-sink metapopulation structure, founder events and population bottlenecks. *Journal of Evolutionary Biology*, *22*(3), 485–497. https://doi.org/10.1111/j.1420-9101.2008.01675.x

Bergeron, L. A., Besenbacher, S., Turner, T. N., Versoza, C. J., Wang, R. J., Price, A. L., Armstrong, E., Riera, M., Carlson, J., Chen, H. Y., Hahn, M. W., Harris, K., Kleppe, A. S., López-Nandam, E. H., Moorjani, P., Pfeifer, S. P., Tiley, G. P., Yoder, A. D., Zhang, G., & Schierup, M. H. (2022). The mutationathon highlights the importance of reaching standardization in estimates of pedigree-based germline mutation rates. *ELife*, *11*, 1–28. https://doi.org/10.7554/eLife.73577

Bergeron, L. A., Besenbacher, S., Zheng, J., Li, P., Bertelsen, M. F., Quintard, B., Hoffman, J. I., Li, Z., St. Leger, J., Shao, C., Stiller, J., Gilbert, M. T. P., Schierup, M. H., & Zhang, G. (2023). Evolution of the germline mutation rate across vertebrates. *Nature*, *615*(7951), 285–291. https://doi.org/10.1038/s41586-023-05752-y

Brooks, R., & Endler, J. A. (2001). Direct and indirect sexual selection and quantitative genetics of male traits in guppies (Poecilia reticulata). *Evolution; International Journal of Organic Evolution*, *55*(5), 1002–1015. https://doi.org/10.1554/0014-3820(2001)055[1002:daissa]2.0.co;2

Bundgaard, J., Loeschcke, V., Schou, M. F., & Bijlsma, K. (2021). Detecting purging of inbreeding depression by a slow rate of inbreeding for various traits: the impact of environmental and experimental conditions. *Heredity*, *127*(1), 10–20. https://doi.org/10.1038/s41437-021-00436-7

Capinha, C., Essl, F., Seebens, H., Moser, D., & Pereira, H. M. (2015). The dispersal of alien species redefines biogeography in the Anthropocene. *Science*, *348*(6240), 1248–1251. https://doi.org/10.1126/science.aaa8913

Charlesworth, D., & Charlesworth, B. (1987). Inbreeding Depression and its Evolutionary Consequences. *Annual Review of Ecology and Systematics*, *18*(1), 237–268. https://doi.org/10.1146/annurev.es.18.110187.001321

Charlesworth, D., & Willis, J. H. (2009). The genetics of inbreeding depression. *Nature Reviews Genetics*, *10*(11), 783–796. https://doi.org/10.1038/nrg2664

Darolti, I., Wright, A. E., & Mank, J. E. (2020). Guppy y chromosome integrity maintained by incomplete recombination suppression. *Genome Biology and Evolution*, *12*(6), 965–977. https://doi.org/10.1093/GBE/EVAA099

de Jong, M. J., van Oosterhout, C., Hoelzel, A. R., & Janke, A. (2024). Moderating the neutralist–selectionist debate: exactly which propositions are we debating, and which arguments are valid? *Biological Reviews*, *99*(1), 23–55. https://doi.org/10.1111/brv.13010

Dobzhansky, T. (1937). *Genetics and the Origin of Species*. Columbia University Press.

Dussex, N., Morales, H. E., Grossen, C., Dalén, L., & van Oosterhout, C. (2023). Purging and accumulation of genetic load in conservation. *Trends in Ecology & Evolution*, *38*(10), 961–969. https://doi.org/10.1016/j.tree.2023.05.008

Dussex, N., van der Valk, T., Morales, H. E., Wheat, C. W., Díez-del-Molino, D., von Seth, J., Foster, Y., Kutschera, V. E., Guschanski, K., Rhie, A., Phillippy, A. M., Korlach, J., Howe, K., Chow, W., Pelan, S., Mendes Damas, J. D., Lewin, H. A., Hastie, A. R., Formenti, G., … Dalén, L. (2021). Population genomics of the critically endangered kākāpō. *Cell Genomics*, *1*(1), 100002. https://doi.org/10.1016/j.xgen.2021.100002

Earl, S. R., Johnson, L. E., Grant, E., Kasubhai, A., López-Sepulcre, A., Yang, Y., & Gordon, S. (2024). Disentangling genetic, plastic and social learning drivers of sex-specific foraging behaviour in Trinidadian guppies (*Poecilia reticulata*). *Proceedings of the Royal Society B: Biological Sciences*, *291*(2018). https://doi.org/10.1098/rspb.2023.2950

Endler, J. A. (1980). Natural Selection on Color Patterns in Poecilia Reticulata. *Evolution*, *34*(1), 76–91. https://doi.org/10.1111/j.1558-5646.1980.tb04790.x

Estoup, A., Ravigné, V., Hufbauer, R., Vitalis, R., Gautier, M., & Facon, B. (2016). Is There a Genetic Paradox of Biological Invasion? *Annual Review of Ecology, Evolution, and Systematics*, *47*(1), 51–72. https://doi.org/10.1146/annurev-ecolsys-121415-032116

Eyre-Walker, A., & Keightley, P. D. (2007). The distribution of fitness effects of new mutations. *Nature Reviews Genetics*, *8*(8), 610–618. https://doi.org/10.1038/nrg2146

Farrer, R. A., Henk, D. A., MacLean, D., Studholme, D. J., & Fisher, M. C. (2013). Using false discovery rates to benchmark SNP-callers in next-generation sequencing projects. *Scientific Reports*, *3*, 1–6. https://doi.org/10.1038/srep01512

Fischer, E. K., Song, Y., Hughes, K. A., Zhou, W., & Hoke, K. L. (2021). Nonparallel transcriptional divergence during parallel adaptation. *Molecular Ecology*, *30*(6), 1516–1530. https://doi.org/10.1111/mec.15823

Fitzpatrick, S. W., Gerberich, J. C., Kronenberger, J. A., Angeloni, L. M., & Funk, W. C. (2015). Locally adapted traits maintained in the face of high gene flow. *Ecology Letters*, *18*(1), 37–47. https://doi.org/10.1111/ele.12388

Fraser, B. A., & Neff, B. D. (2010). Parasite mediated homogenizing selection at the MHC in guppies. *Genetica*, *138*(2), 273–278. https://doi.org/10.1007/s10709-009-9402-y

Haldane, J. B. S. (1937). The Effect of Variation of Fitness. *The American Naturalist*, *71*(735), 337–349. https://doi.org/10.1086/280722

Herdegen-Radwan, M. (2019). Bolder guppies do not have more mating partners, yet sire more offspring. *BMC Evolutionary Biology*, *19*(1), 211. https://doi.org/10.1186/s12862-019-1539-4

Ioannou, C. C., Ramnarine, I. W., & Torney, C. J. (2017). High-predation habitats affect the social dynamics of collective exploration in a shoaling fish. *Science Advances*, *3*(5). https://doi.org/10.1126/sciadv.1602682

Johnson, C. N., Balmford, A., Brook, B. W., Buettel, J. C., Galetti, M., Guangchun, L., & Wilmshurst, J. M. (2017). Biodiversity losses and conservation responses in the Anthropocene. *Science*, *356*(6335), 270–275. https://doi.org/10.1126/science.aam9317

Kelly, L. T., Giljohann, K. M., Duane, A., Aquilué, N., Archibald, S., Batllori, E., Bennett, A. F., Buckland, S. T., Canelles, Q., Clarke, M. F., Fortin, M.-J., Hermoso, V., Herrando, S., Keane, R. E., Lake, F. K., McCarthy, M. A., Morán-Ordóñez, A., Parr, C. L., Pausas, J. G., … Brotons, L.

(2020). Fire and biodiversity in the Anthropocene. *Science*, *370*(6519). https://doi.org/10.1126/science.abb0355

Kessler, M. D., Loesch, D. P., Perry, J. A., Heard-Costa, N. L., Taliun, D., Cade, B. E., Wang, H., Daya, M., Ziniti, J., Datta, S., Celedón, J. C., Soto-Quiros, M. E., Avila, L., Weiss, S. T., Barnes, K., Redline, S. S., Vasan, R. S., Johnson, A. D., Mathias, R. A., … O'Connor, T. D. (2020). De novo mutations across 1,465 diverse genomes reveal mutational insights and reductions in the Amish founder population. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(5), 2560–2569. https://doi.org/10.1073/pnas.1902766117

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press. https://doi.org/10.1017/CBO9780511623486

Kimura, M., & Ohta, T. (1969). The Average Number of Generations Until Fixation of a Mutant Gene in a Finite Population. *Genetics*, *61*(3), 763–771. https://doi.org/10.1093/genetics/61.3.763

Kimura, M., & Ohta, T. (1971). On the rate of molecular evolution. *Journal of Molecular Evolution*, *1*(1), 1–17. https://doi.org/10.1007/BF01659390

Kleinman-Ruiz, D., Lucena-Perez, M., Villanueva, B., Fernández, J., Saveljev, A. P., Ratkiewicz, M., Schmidt, K., Galtier, N., García-Dorado, A., & Godoy, J. A. (2022). Purging of deleterious burden in the endangered Iberian lynx. *Proceedings of the National Academy of Sciences*, *119*(11). https://doi.org/10.1073/pnas.2110614119

Kondrashov, F. A., & Kondrashov, A. S. (2010). Measurements of spontaneous rates of mutations in the recent past and the near future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1544), 1169–1176. https://doi.org/10.1098/rstb.2009.0286

Kong, A., Frigge, M. L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S. A., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A., Wong, W. S. W., Sigurdsson, G., Walters, G. B., Steinberg, S., Helgason, H., Thorleifsson, G., Gudbjartsson, D. F., Helgason, A., Magnusson, O. T., … Stefansson, K. (2012). Rate of de novo mutations and the importance of father-s age to disease risk. *Nature*, *488*(7412), 471–475. https://doi.org/10.1038/nature11396

Kuang, W., Hu, J., Wu, H., Fen, X., Dai, Q., Fu, Q., Xiao, W., Frantz, L., Roos, C., Nadler, T., Irwin, D. M., Zhou, L., Yang, X., & Yu, L. (2020). Genetic Diversity, Inbreeding Level, and Genetic Load in Endangered Snub-Nosed Monkeys (Rhinopithecus). *Frontiers in Genetics*, *11*. https://doi.org/10.3389/fgene.2020.615926

Lau, C. H. E., & Robinson, O. (2021). DNA methylation age as a biomarker for cancer. *International Journal of Cancer*, *148*(11), 2652–2663. https://doi.org/10.1002/ijc.33451

Lohr, J. N., & Haag, C. R. (2015). Genetic load, inbreeding depression, and hybrid vigor covary with population size: An empirical evaluation of theoretical predictions. *Evolution*, *69*(12), 3109–3122. https://doi.org/10.1111/evo.12802

Lombaert, E., Blin, A., Porro, B., Guillemaud, T., Bernal, J. S., Chang, G., Kirichenko, N., Sappington, T. W., Toepfer, S., & Deleury, E. (2024). Unraveling genetic load dynamics during biological invasion: insights from two invasive insect species. *BioRxiv*.

Lynch, M. (2010). Evolution of the mutation rate. *Trends in Genetics*, *26*(8), 345–352. https://doi.org/10.1016/j.tig.2010.05.003

Lynch, M., Ackerman, M. S., Gout, J. F., Long, H., Sung, W., Thomas, W. K., & Foster, P. L. (2016). Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics*, *17*(11), 704–714. https://doi.org/10.1038/nrg.2016.104

Lynch, M., Conery, J., & Bürger, R. (1995). Mutational Meltdowns in Sexual Populations. *Evolution*, *49*(6), 1067–1080. https://doi.org/10.1111/j.1558-5646.1995.tb04434.x

Magurran, A. E. (2005). *Evolutionary Ecology*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198527855.001.0001

Mathur, S., Tomeček, J. M., Tarango-Arámbula, L. A., Perez, R. M., & DeWoody, J. A. (2023). An evolutionary perspective on genetic load in small, isolated populations as informed by whole genome resequencing and forward-time simulations. *Evolution*, *77*(3), 690–704. https://doi.org/10.1093/evolut/qpac061

Muller, H. J. (1950). Our load of mutations. *American Journal of Human Genetics*, *2*(2), 111–176. http://www.ncbi.nlm.nih.gov/pubmed/14771033

Ochoa, A., Onorato, D. P., Roelke-Parker, M. E., Culver, M., & Fitak, R. R. (2022). Give and take: Effects of genetic admixture on mutation load in endangered Florida panthers. *Journal of Heredity*, *113*(5), 491–499. https://doi.org/10.1093/jhered/esac037

Otto, S. P. (2018). Adaptation, speciation and extinction in the Anthropocene. *Proceedings of the Royal Society B: Biological Sciences*, *285*(1891), 20182047. https://doi.org/10.1098/rspb.2018.2047

Peischl, S., & Excoffier, L. (2015). Expansion load: Recessive mutations and the role of standing genetic variation. *Molecular Ecology*, *24*(9), 2084–2094. https://doi.org/10.1111/mec.13154

Phillips, K. P., Cable, J., Mohammed, R. S., Herdegen-Radwan, M., Raubic, J., Przesmycka, K. J., Van Oosterhout, C., & Radwan, J. (2018). Immunogenetic novelty confers a selective advantage in host–pathogen coevolution. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(7), 1552–1557. https://doi.org/10.1073/pnas.1708597115

Qiu, S., Yong, L., Wilson, A., Croft, D. P., Graham, C., & Charlesworth, D. (2022). Partial sex linkage and linkage disequilibrium on the guppy sex chromosome. *Molecular Ecology*, *31*(21), 5524–5537. https://doi.org/10.1111/mec.16674

Robinson, J., Kyriazis, C. C., Yuan, S. C., & Lohmueller, K. E. (2023). Deleterious Variation in Natural Populations and Implications for Conservation Genetics. *Annual Review of Animal Biosciences*, *11*(1), 93–114. https://doi.org/10.1146/annurev-animal-080522-093311

Rougemont, Q., Moore, J. S., Leroy, T., Normandeau, E., Rondeau, E. B., Withler, R. E., van Doornik, D. M., Crane, P. A., Naish, K. A., Garza, J. C., Beacham, T. D., Koop, B. F., & Bernatchez, L. (2020). Demographic history shaped geographical patterns of deleterious mutation load in a broadly distributed Pacific Salmon. *PLoS Genetics*, *16*(8), 1–29. https://doi.org/10.1371/JOURNAL.PGEN.1008348

Saclier, N., Chardon, P., Malard, F., Konecny-Dupré, L., Eme, D., Bellec, A., Breton, V., Duret, L., Lefebure, T., & Douady, C. J. (2020). Bedrock radioactivity influences the rate and spectrum of mutation. *ELife*, *9*, 1–20. https://doi.org/10.7554/eLife.56830

Scally, A. (2016). The mutation rate in human evolution and demographic inference. *Current Opinion in Genetics & Development*, *41*, 36–43. https://doi.org/10.1016/j.gde.2016.07.008

Schories, S., Meyer, M. K., & Schartl, M. (2009). Description of poecilia (acanthophacelus) obscura n. sp., (teleostei: Poeciliidae), a new guppy species from western trinidad, with remarks on p. wingei and the status of the “endler’s guppy.” *Zootaxa*, *50*(2266), 35–50. https://doi.org/10.11646/zootaxa.2266.1.2

Shackleton, R. T., Shackleton, C. M., & Kull, C. A. (2019). The role of invasive alien species in shaping local livelihoods and human well-being: A review. *Journal of Environmental Management*, *229*, 145–157. https://doi.org/10.1016/j.jenvman.2018.05.007

Sievers, C., Willing, E. M., Hoffmann, M., Dreyer, C., Ramnarine, I., & Magurran, A. (2012). Reasons for the invasive success of a guppy (Poecilia reticulata) population in Trinidad. *PLoS ONE*, *7*(5). https://doi.org/10.1371/journal.pone.0038404

Smeds, L., & Ellegren, H. (2023). From high masked to high realized genetic load in inbred Scandinavian wolves. *Molecular Ecology*, *32*(7), 1567–1580. https://doi.org/10.1111/mec.16802

Speak, S. A., Birley, T., Bortoluzzi, C., Clark, M. D., Percival-Alwyn, L., Morales, H. E., & van Oosterhout, C. (2024). Genomics-informed captive breeding can reduce inbreeding depression and the genetic load in zoo populations. *Molecular Ecology Resources*, *24*(7). https://doi.org/10.1111/1755-0998.13967

van der Zee, M. J., Whiting, J. R., Paris, J. R., Bassar, R. D., Travis, J., Weigel, D., Reznick, D. N., & Fraser, B. A. (2022). Rapid genomic convergent evolution in experimental populations of Trinidadian guppies (*Poecilia reticulata*). *Evolution Letters*, *6*(2), 149–161. https://doi.org/10.1002/evl3.272

von Seth, J., Dussex, N., Díez-del-Molino, D., van der Valk, T., Kutschera, V. E., Kierczak, M., Steiner, C. C., Liu, S., Gilbert, M. T. P., Sinding, M.-H. S., Prost, S., Guschanski, K., Nathan, S. K. S. S., Brace, S., Chan, Y. L., Wheat, C. W., Skoglund, P., Ryder, O. A., Goossens, B., … Dalén, L. (2021). Genomic insights into the conservation status of the world's last remaining Sumatran rhinoceros populations. *Nature Communications*, *12*(1), 2393. https://doi.org/10.1038/s41467-021-22386-8

Whiting, J. R., Paris, J. R., van der Zee, M. J., Parsons, P. J., Weigel, D., & Fraser, B. A. (2021). Drainage-structuring of ancestral variation and a common functional pathway shape limited genomic convergence in natural high- and low-predation guppies. *PLOS Genetics*, *17*(5), e1009566. https://doi.org/10.1371/journal.pgen.1009566

Willing, E. M., Bentzen, P., Van Oosterhout, C., Hoffmann, M., Cable, J., Breden, F., Weigel, D., & Dreyer, C. (2010). Genome-wide single nucleotide polymorphisms reveal population history and adaptive divergence in wild guppies. *Molecular Ecology*, *19*(5), 968–984. https://doi.org/10.1111/j.1365-294X.2010.04528.x

Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding, and selection in evolution. *Proceedings of the Sixth International Congress of Genetics*.

Wu, K., Qin, D., Qian, Y., & Liu, H. (2024). A new era of mutation rate analyses: Concepts and methods. *Zoological Research*, *45*(4), 767–780. https://doi.org/10.24272/j.issn.2095-8137.2024.058

Yoder, A. D., & Tiley, G. P. (2021). The challenge and promise of estimating the de novo mutation rate from whole-genome comparisons among closely related individuals. *Molecular Ecology*, *30*(23), 6087–6100. https://doi.org/10.1111/mec.16007

Chapter I

Received: 25 February 2023 | Revised: 16 June 2023 | Accepted: 5 July 2023

DOI: 10.1111/1755-0998.13841

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

Validation of machine learning approach for direct mutation rate estimation

Katarzyna Burda | Mateusz Konczal

Evolutionary Biology Group, Faculty of Biology, Adam Mickiewicz University, Poznań, Poland

Correspondence
Mateusz Konczal, Evolutionary Biology Group, Faculty of Biology, Adam Mickiewicz University, Poznań 60-614, Poland.
Email: mateusz.konczal@amu.edu.pl

Funding information
Narodowe Centrum Nauki, Grant/Award Number: 2018/31/D/NZ8/00091

Handling Editor: Alana Alexander

Abstract

Mutations are the primary source of all genetic variation. Knowledge about their rates is critical for any evolutionary genetic analyses, but for a long time, that knowledge has remained elusive and indirectly inferred. In recent years, parent–offspring comparisons have yielded the first direct mutation rate estimates. The analyses are, however, challenging due to high rate of false positives and no consensus regarding standardized filtering of candidate de novo mutations. Here, we validate the application of a machine learning approach for such a task and estimate the mutation rate for the guppy (*Poecilia reticulata*), a model species in eco-evolutionary studies. We sequenced 4 parents and 20 offspring, followed by screening their genomes for de novo mutations. The initial large number of candidate de novo mutations was hard-filtered to remove false-positive results. These results were compared with mutation rate estimated with a supervised machine learning approach. Both approaches were followed by molecular validation of all candidate de novo mutations and yielded similar results. The ML method uniquely identified three mutations, but overall required more hands-on curation and had higher rates of false positives and false negatives. Both methods concordantly showed no difference in mutation rates between families. Estimated here the guppy mutation rate is among the lowest directly estimated mutation rates in vertebrates; however, previous research has also found low estimated rates in other teleost fishes. We discuss potential explanations for such a pattern, as well as future utility and limitations of machine learning approaches.

KEYWORDS
guppy, machine learning, mutation rate, teleost, whole-genome sequencing

1 | INTRODUCTION

Mutations provide the genetic variation that is the raw material for evolution. Importantly, the rate of mutations can serve as a means for measuring divergence between species and to understand diversity within species. Such connections can be made because, assuming neutrality, the rate of divergence is determined purely by the mutation rate. It provides a 'neutral evolutionary clock' that should tick at the speed with which new mutations appear in populations (Kimura, 1983; Kimura & Ohta, 1971). DNA can be incorrectly replicated or damaged, both in somatic cells and germline cells, however, only the latter have long-term evolutionary consequences. Germline mutations are important for adaptation (as sources of variation) but are also often deleterious (Eyre-Walker & Keightley, 2007). Organisms have therefore evolved DNA repair and check mechanisms, achieving exceedingly low mutation rates in organisms' germlines (Wang et al., 2019). Nevertheless, per-nucleotide mutation rates per generation are highly variable in eukaryotes, ranging from 1.22×10^{-8} (in *Homo sapiens*, Kessler et al., 2020) to as low as 7.6×10^{-12} (in *Tetrahymena thermophila*, Long et al., 2016), and the differences can only partially be explained by differences in generation times or numbers of cell divisions between generations (Lynch, 2010; Wang & Obbard, 2023). Moreover, mutations are random with respect to their adaptive value in organisms' environment, although the rates of mutations can be affected by environmental factors (Saclier et al., 2020), as well as by many features of individuals' genetic backgrounds (Aikens et al., 2019), such as ploidy (Sharp et al., 2018), local recombination rates (Lercher & Hurst, 2002) and epigenetic modifications (Habig et al., 2021; Monroe et al., 2022; Zhou et al., 2020). Mutation rates can respond to selection and thus evolve (Lynch, 2010; Lynch et al., 2016), but understanding this process is limited by the fact that mutation rates are extremely difficult to measure (Kondrashov & Kondrashov, 2010). Until recently, most mutation rate estimates were based on indirect measurements, including screens for phenotypically detectable mutations, measuring substitution rates at sites that are expected to be neutral (including the most common phylogenetic approaches) and inferring mutation rates from within-population variation, or direct sequence-level screening of mutations in mutation accumulation lines (Kondrashov & Kondrashov, 2010).

Recent developments and decreased costs of whole-genome sequencing technologies have made direct, pedigree-based approaches more available for estimating mutation rates. By comparing the genomes of parents and their offspring, we can count variants that appear in offspring but are absent in both parents, thus representing de novo mutations (DNMs). Initially, such an approach was applied to humans (Conrad et al., 2011; Kong et al., 2012; Roach et al., 2010) and primates (Pfeifer, 2017; Thomas et al., 2018; Venn et al., 2014). Since then, several other studies have aimed to estimate the mutation rate in humans from an increasing number of sequenced trios and pedigrees (Jónsson et al., 2017; Kessler et al., 2020). These studies demonstrated little variation in mutation rate between different human populations (Kessler et al., 2020) while showing that lifestyle (Kessler et al., 2020), parental age (Kong et al., 2012) and rare heritable cancer syndromes (Lau & Robinson, 2021) can increase the rate of new mutations. In contrast, other studies show potential for ongoing, population-specific, evolution of mutation rates (Harris, 2015; Harris & Pritchard, 2017; Sasani et al., 2019), suggesting that mutation rates can evolve and may differ between closely related species. Wide sampling across different species and populations is thus needed to understand its importance for patterns of molecular evolution across the tree of life (Lynch, 2020). Around 20 nonprimate vertebrate species across fishes (Feng et al., 2017; Malinsky et al., 2018), birds (Smeds et al., 2016) and mammals (Campbell et al., 2021; Koch et al., 2019; Lindsay et al., 2019; Wang et al., 2021) have had mutation rates estimated by separate studies using independent pedigree sequencing (Bergeron et al., 2022). Recent advances allowed estimation of mutation rate across 68 vertebrate species using standardized methodology (Bergeron et al., 2023); however, assessing how emerging methodologies such as machine learning might compare to current approaches has not yet been assessed.

Despite progress in the field, using pedigrees and whole-genome sequencing to estimate mutation rates is still a challenging task. Inevitable sequencing errors combined with bioinformatic limitations (ambiguous mapping of reads to the reference genome, genotyping errors) cause a high rate of false positives among rare variants (Farrer et al., 2013). Numerous filters are thus often applied to avoid false positives (increased precision), but filtering that is too conservative can discard true positives, reducing sensitivity. A recent study demonstrated that different filtering strategies used by different research groups resulted in twofold variation in mutation rate estimation from the same sequencing data (Bergeron et al., 2022).

A potential solution is to use less stringent filtering to identify candidate DNMs and then validate a subset of such candidates. Such successfully subjected to the validation process variants (hereinafter 'successfully validated') can serve as a training set for machine learning (ML) algorithms that can assess the probability that the remaining candidate DNMs are true de novo mutations. Such an approach has been applied for humans in cancer-related studies (Feliciano et al., 2019; Nishioka et al., 2021; Zhou et al., 2019), but it has not been widely used to estimate de novo mutation rate and it is not known how efficient such an approach would be in comparison with traditional 'hard-filtering' approaches. Here, we compare performance of a standard method based on hard filtering of candidate de novo mutations (Bergeron et al., 2022) with a machine learning approach implemented in DNMF (Liu et al., 2014). The comparison was performed on the genomes of two families consisting of 24 guppies (*Poecilia reticulata*) in total.

Guppies, freshwater fishes native to Caribbean islands and South America, have been a model species in studies on sexual selection (Brooks & Endler, 2001), host–parasite coevolution (Phillips et al., 2018), eco-evolutionary dynamics (Reznick & Travis, 2019), sex chromosome evolution (Darolti et al., 2020) and repeatability of evolution (Whiting et al., 2021). With growing information about recombination rates (Bergero et al., 2019; Charlesworth et al., 2020), patterns of sex chromosome evolution (Lin et al., 2022; Qiu et al., 2022; Wright et al., 2017) and the genomic basis of local adaptation (Fraser & Neff, 2010; Whiting et al., 2021), guppies are becoming a vertebrate model system to study the interplay between genomics, ecology and evolution. Therefore, the estimate presented here would be of particular use to the guppy-related scientific community but would also enhance our understanding of the evolution of the mutation rates in general, adding point estimates to a handful of mutation rates estimated for vertebrates. More broadly, we evaluated the benefits and challenges of implementing a machine learning-based approach for direct de novo mutation rate estimation and compare it with commonly used 'hard-filtering'.

2 | MATERIALS AND METHODS

2.1 | DNA isolation, library preparation and sequencing

The fish used in the analysis originated from laboratory stock (kept at Adam Mickiewicz University), which are descendants of wild-caught Trinidadian guppies collected from the Tacarigua River. From this stock, two pairs of female and male individuals of unknown age were selected to breed and create two families: T7 and T10. For each family, 10 juvenile offspring were chosen for identifying de novo mutations, giving (with their parents) 24 individuals in total. Tissue samples were obtained by clipping fish tails. DNA was extracted using a Thermo Scientific MagJET Genomic DNA kit (Thermo Scientific). All the above procedures were performed in accordance with Polish law and European Directive 2010/63/EU.

All 24 samples were used for library preparation using a NEB Next Ultra II FS (PCR-free) kit and TrueSeq DNA CD Indices. Libraries were then sequenced on a single S4 lane of the Illumina NovaSeq 6000 platform with 2 × 150 bp read mode in the Macrogene sequencing facility. Since sequencing of the T10 mother yielded relatively low coverage, additional sequencing was performed for this individual on a single run of an Illumina MiSeq machine using previously prepared library.

2.2 | Variant calling

The quality values of the raw sequencing reads from each individual were inspected using FastQC (version 0.11.8; Andrews and Babraham Bioinformatics, 2010). Low-quality bases and adapters were removed using Trimmomatic (version 0.39; Bolger et al., 2014). Bwa mem (version 0.7.10; Li, 2013) was then used to map reads to the guppy female reference genome downloaded from the National Center for Biotechnology Information (NCBI) database (GenBank accession no. GCA_000633615.2; Kunstner et al., 2016).

To identify candidate de novo mutations, we followed an approach proposed by Feng et al. (2017) (with modifications). First, SNP calling was performed twice using two different calling tools: samtools/bcftools (version 1.6.0/1.9 accordingly; Li et al., 2009) and GATK (version 4.1.4.1; McKenna et al., 2010). For the samtools/bcftools method, each BAM file was first piled up using samtools mpileup (with parameters: -R -C50 -t DP, ADF, ADR), generating intermediate-VCF files of all genomic positions. These files were then used to call variants with bcftools call (with parameter: -f GQ), separately for each family. For the GATK method, a genomic-VCF (gVCF) file for each individual was first generated using gatk HaplotypeCaller in GVCF mode. Genomic-VCF files were then combined into one file and used to produce VCF files for all samples using gatk GenotypeGVCFs. Similar to the samtools/bcftools method, GATK analyses were performed separately for each family.

After producing raw VCF files for each family and calling method, an initial filtration was applied. First, insertions and deletions were removed from vcf files using vcftools (version 0.1.14; Danecek et al., 2011). Second, all variants with missing parental genotypes (marked in file with '.') were removed using a custom Python script. Variant files produced by both methods (samtools/bcftools and GATK) were then intersected, and only shared files were used in the following steps.

Next, in order to verify whether DP and GQ thresholds discussed in Bergeron et al. (2022); DP > 0.5 × mean and DP < 2 × mean and GQ > 70) fit our data well, we plotted DP and GQ values for individuals of heterozygous (true) and homozygous (false) offspring genotypes of alternatively homozygous parents (Figure S1). We decided to put additional threshold of minimum DP (DP > 15) to decrease rate of false negatives. These thresholds were used in future steps.

2.3 | Summary statistics

We analysed basic statistics that characterize sites passing DP and GQ thresholds in our studied families. We used vcftools to calculate per family mean nucleotide diversity (--windowed-pi, with window length of 75 kb), transversions to transitions ratio (--TsTv-summary) and runs of homozygosity longer than 1 kb (--LROH). We also analysed heterozygosity in both families using custom python script.

2.4 | Identification of de novo mutations

We screened both families for de novo mutations (mutations were selected if at least one progeny individual had an allele not observed in either parental genotype) within sites that meet DP and GQ thresholds for an offspring and both parents. These candidate DNMs were then used in two parallel analyses (for details see Figure S2). The first method was based on the approach used and extensively discussed by Bergeron et al. (2022). It applied a set of hard filters to genotypes (i.e. 'hard filtering'). The second method was based on machine learning (using DeNovoMutationFilter—DNMF, version 0.1.1; Liu et al., 2014) and a training set with verified true and false candidate DNMs (i.e. 'soft filtering and machine learning').

2.4.1 | Hard filtering

Hard filtering selects candidate DNMs based on a vcf file INFO field (including quality by depth [QD], mapping quality [MQ], Fisher score [FS], strands odds ratio [SOR], mapping quality rank sum [MQRankSum] and reads positioning rank sum [ReadPosRankSum]), allelic balance (AB) in offspring and allelic depth (AD) in parents. To filter candidates, together with mentioned above DP/GQ filters, the following thresholds were applied: QD ≥ 2, MQ ≥ 40, FS ≤ 60, SOR ≤ 3, MQRankSum ≥ −12.5, ReadPosRankSum ≥ −8, 0.3 ≤ AB ≤ 0.7 and in both of parents: AD = 0. To ensure that the filters work as intended, each of the filtering thresholds was checked with a distribution of INFO field, AB and AD values derived from the set of heterozygous offspring whose parents were homozygous for alternative variants. Candidate DNMs were also filtered for mappability values. To obtain the mappability information, we used the GenMap tool (version 1.3.0; Pockrandt et al., 2020) on the reference sequence using 150 bp-mers and allowing for three mismatches. Mappability scores lower than 1 indicate that the position is not unique in the genome and that the read could have been mapped elsewhere; that is, a mappability value of 1 represents a unique k-mer in the reference sequence, a mappability value of 0.5 represents a κ-mer occurring twice, while values close to zero indicate a κ-mer occurring in repetitive regions. Sites with mappability lower than one were filtered out from our analyses.

Such selected candidate DNMs were then subjected to molecular validation and IGV-facilitated inspection (version 2.8.10; Thorvaldsdóttir et al., 2013). For details, see Data S1. The experimental validation of a candidate de novo mutation consisted of amplifying the sequence of interest (de novo mutation region) in both parents and offspring using polymerase chain reaction (PCR) amplification and Sanger sequencing, with PCR primers designed using NCBI's Primer BLAST (only forward primers were used in the sequencing step). The chromatograms from Sanger sequencing were verified using MEGA (version 10.2.5; Kumar et al., 2018) positively ('true mutations') or negatively ('false mutations') verifying candidate mutations. Some of the variants were difficult to validate because of multiple PCR products and/or unsuccessful sequencing. If the first Sanger sequencing failed, another pair of primers was designed. If amplification of the second pair of primers failed, such candidate mutations were left as not successfully validated.

2.4.2 | Soft filtering and machine learning

An alternative approach to filter candidate DNMs was based on the DNMF software (Liu et al., 2014). Briefly, DNMF analyses the candidate mutation positions in the BAM files of parents and offspring and looks for patterns in 59 distinct features (e.g. allele balance, mean base quality and read depth). The approach is based on training sets of true and false de novo mutations, which are used to train a gradient-boosting model, which is then applied to filter out false-positive de novo mutations. To obtain training sets, a subset of 81 candidate mutations was validated using PCR and Sanger sequencing (as described above, most of the variants selected for validation overlapped between hard filtering and ML). Applying an approach similar to Smeds et al. (2016), we selected only candidates that: (1) had at least 20% of reads supporting the DNM in offspring, (2) did not have reads supporting a third variant at a frequency higher than 10% in the offspring, and (3) did not have reads with the DNM allele in parents at a frequency higher than 20%. Additionally, we excluded mutations in low-complexity regions, based on the NCBI annotation (RepeatMasker output of GenBank accession no. GCA_000633615.2), and in sites with mappability lower than 1. The candidate mutations selected for Sanger sequencing were also inspected using the IGV tool (see Data S1).

The positively and negatively verified DNMs were used as our training set for the supervised machine learning algorithm implemented in DNMF. After training, DNMF analysed the other candidate DNMs (those that were not in the training set or those for which validation failed). As a consequence, each candidate DNM was given a score representing the probability of being a true mutation. The variants that did not pass the DNMF criteria (probability of being true <0.4, following the approach proposed by Liu et al., 2014) were classified as false positives and removed from future investigations. The remaining candidate DNMs were checked for mappability, and sites with mappability scores = 1 were validated using Sanger sequencing and the IGV tool.

2.4.3 | Origin of de novo mutations and masking repeats

To assess the parental origins of DNMs, we manually screened sequencing reads at a given site using the IGV tool and corresponding parent–offspring trio BAM files. Reads including the DNM and one or more variants distinctive to one of the parents pointed out the parental origin of the DNM. Manual assignment was performed for all positively validated DNMs.

To future account for low-complexity regions and repeats in the genome, we masked the guppy reference genome using RepeatMasker and *Danio rerio* repetitive elements downloaded from Repbase (database of guppy repetitive elements was not available in Repbase). Sets of candidate DNMs and annotated repeats were then intersected using BEDTOOLS (v2.27.1). In particular, we checked the number of DNMs localized in annotated repeats that were identified as true and false positives and how many of them could not have been assigned due to failed validation. All candidate DNMs localized in repetitive regions were excluded from future calculations.

2.5 | False-negative rate estimation

The false-negative rate (FNR) was calculated using two commonly applied approaches. The first method called here 'known SNPs' assumes that in sites with parents being alternative homozygotes all offspring should have heterozygotic genotypes (Mendelian situation). Using this simple rule, we estimated FNR calculating the fraction of homozygotic offspring among all offspring genotypes of such parents (0/0 × 1/1). The second method to estimate FNR was based on in silico simulations. Simulated DNMs were subjected to hard filtering and machine learning approaches to call candidate variants. In silico simulated variants were added to the T7-F-1 individual (randomly chosen as a representative sample) bam file using BAMSurgeon (v1.4.1, Ewing et al., 2015) using a variant allele fraction randomly sampled from a normal distribution with mean 0.5 and standard deviation 0.1 (information on an allele fraction was one of the columns in BAMSurgeon input file). Parameters of the distribution were estimated from variant allele fraction (AD/DP) values in the set of the heterozygotic offspring genotypes called in sites where parents were called alternatively homozygotes. Mutations were simulated in sites that had DP matching the threshold used during quality filtering (1/2 mean DP and 2 mean DP, GQ > 70) and were reference homozygotes in T7-F-1 and in T7 parents. Additionally, chosen sites were at least 150 bp from each other and were all located on chromosome 1. Simulated reads were then extracted from the bam file using samtools fastq and mapped back to the reference genome. Simulated sites were excluded from future calculations if they fell out of effective sites (DP, GQ filters) or if an allelic balance deviated more than 10% from the expected value. We then called SNPs and identified DNMs in exactly the same way as described above. Briefly, SNPs were called using GATK and samtools/bcftools, intersected and screened for DNMs with DNMF and hard filtering approaches. The calculated fraction of mutations not identified as candidate DNMs among all successfully simulated mutations was defined as FNR.

2.6 | Mutation rate estimation

Mutation rate estimates for the two methods of filtering were calculated using the following formula:

$$\mu = \frac{n_{\mathrm{TRUE}} + \left[n_{vf} \cdot (1 - \mathrm{FDR})\right]}{2 \cdot \mathrm{ES} \cdot (1 - \mathrm{FNR})}$$

where n_{TRUE} denotes all successfully validated true DNMs, n_{vf} represents all DNMs without successful amplification and Sanger sequencing (i.e. for which validation failed), ES is the number of effective sites and FNR and FDR represent the false-negative rate and false discovery rate, respectively.

The false discovery rate can be calculated by dividing the number of candidate DNMs successfully amplified and found to be false after Sanger sequencing (n_{FALSE}) by all candidate DNMs successfully amplified and validated with Sanger sequencing (n_V, *true plus false*):

$$\mathrm{FDR} = \frac{n_{\mathrm{FALSE}}}{n_v}$$

To calculate the number of effective sites (ES), we counted sites with DP and GQ values that met the threshold in both parents and each individual offspring using the genomic-vcf files (g.vcf) in base pair resolution. We then excluded sites with mappability score lower than 1 (across a genome we identified 14,998,180 sites with mappability scores lower than 1) as all candidate DNMs localized in such regions were excluded from molecular validation. Additionally, we excluded sites masked by RepeatMasker.

Estimation of the confidence intervals was calculated according to the Poisson distribution, assuming normal approximation. Specifically, the 95% confidence intervals were calculated using the following formula:

$$\mu_{\mathrm{CI}} = \pm 1.96 \sqrt{\frac{\mu}{2 \cdot \mathrm{ES}}}$$

2.7 | Demographic history inference

Using our mutation rate, we inferred changes in the effective population size of guppy populations using the PSMC (pairwise sequentially Markovian coalescent) approach (version 0.6.5; Li & Durbin, 2011) to demonstrate how mutation rate estimate can be applied in evolutionary investigations. This analysis was done on the diploid consensus parental genome FASTA files, which we obtained by running the bcftools consensus command (with the same individual-specific SNP quality and coverage thresholds as used for filtering DNMs). Each sample was bootstrapped 50 times. We assumed two generations per year (Reznick et al., 1997) to scale the time estimates.

The scripts used in described above analyses are available under the following link: https://github.com/0-Ioniel-0/guppy_MR

3 | RESULTS

3.1 | Whole-genome sequencing and variant detection

All 24 individuals originating from two families (2 parents and 10 offspring from each) were successfully sequenced and filtered for adapter content (Figure S3). Table 1 shows the mean coverage and number of effective sites for each sample. Our samtools/bcftools approach yielded a total of 7,991,906 SNPs, including all individuals in both families, of which 63.73% (5,093,623 SNPs) were shared between both families. The GATK method found 8,200,062 SNPs (60.29% shared between the families. Table 2 shows full details).

Mean genome-wide nucleotide diversity (π_{T10}=0.0038; π_{T7}=0.0034, Figure S4), transitions to transversions ratio (TS/TV_{T10}=1.363; TS/TV_{T7}=1.361) and observed heterozygosity (H_{T10}=0.006; H_{T7}=0.004, Figure S5) was similar for both families. Runs of Homozygosity longer than 1 kb were slightly more common in one of the parents (T7-A), but both families lacked typical runs (Figure S6) indicating low level of inbreeding (at least 10 kb long, Ceballos et al., 2018).

3.2 | De novo mutation candidates and experimental validation

After quality filtration and removal of variants with missing parental genotypes (3141 variants in T7, 3697 variants in T10), 688 candidate DNMs at a total of 350 sites (e.g. positions in the reference genome) were found (Data S2a,b; Figure S7). A total of 181 of these variants were present in a single offspring individual, and 505 (at 169 sites) were found in two or more siblings. Twenty-six per cent of all DNMs (182 mutations at 76 sites) were not considered for validation because they had mappability scores lower than 1.

The remaining 274 sites (506 mutations in total) were hard-filtered leaving 40 sites (46 mutations), which were verified using Sanger sequencing of parents and offspring (see Materials and

TABLE 1 Sequencing and mutation screening results for all individuals.

		Effective sites		True Positive		False positive		Failed amplification		FNR		FDR		Mutation rate	
Ind.	Cov.	Count	% of genome	ML	HF	ML	HF	ML	HF	ML	HF	ML	HF	ML	HF
T7-A	61.0														
T7-O	67.0														
T7-F1	50.0	5.0E+08	69	2	2	5	2	2	0	0.26	0.16	0.75	0.50	2.96E-09	2.36E-09
T7-F2	35.0	3.6E+08	50	2	2	1	2	4	1	0.26	0.16	0.75	0.50	4.91E-09	4.10E-09
T7-F3	41.0	4.6E+08	62	0	0	3	1	2	1	0.26	0.16	0.75	1.00	6.52E-10	0.00E+00
T7-F4	32.0	3.2E+08	44	3	3	5	1	1	0	0.26	0.16	0.75	0.25	6.06E-09	5.60E-09
T7-J1	46.0	4.9E+08	67	1	1	1	0	2	1	0.26	0.16	0.75	0.00	1.82E-09	2.43E-09
T7-J2	23.0	1.0E+08	14	0	0	3	1	4	1	0.26	0.16	0.75	1.00	5.70E-09	0.00E+00
T7-J3	39.0	4.3E+08	58	5	4	3	0	4	1	0.26	0.16	0.75	0.00	8.39E-09	6.99E-09
T7-J4	48.0	4.9E+08	67	1	1	4	0	4	1	0.26	0.16	0.75	0.00	2.43E-09	2.43E-09
T7-M1	41.0	4.4E+08	60	1	1	5	0	3	1	0.26	0.16	0.75	0.00	2.36E-09	2.70E-09
T7-M2	53.0	5.1E+08	70	4	4	6	1	1	0	0.26	0.16	0.75	0.20	4.97E-09	4.68E-09
T7	**43.2**	**4.1E+09**	**56**	**19**	**18**	**36**	**8**	**27**	**7**	0.26	0.16	**0.65**	**0.31**	**4.11E-09**	**3.32E-09**
T10-A	30.0														
T10-O	39.0														
T10-F1	70.0	1.9E+08	27	0	0	2	0	1	0	0.26	0.16	1.00	0.00	0.00E+00	0.00E+00
T10-F2	37.0	1.6E+08	22	1	0	4	0	1	1	0.26	0.16	0.80	0.00	5.04E-09	3.70E-09
T10-J1	35.0	1.6E+08	22	1	1	4	0	3	1	0.26	0.16	0.80	0.00	6.82E-09	7.51E-09
T10-J2	61.0	1.9E+08	27	1	0	4	0	0	0	0.26	0.16	0.80	0.00	3.47E-09	0.00E+00
T10-J3	51.0	1.9E+08	26	0	0	5	0	1	0	0.26	0.16	1.00	0.00	0.00E+00	0.00E+00
T10-J4	64.0	1.7E+08	23	0	0	0	0	0	0	0.26	0.16	0.00	0.00	0.00E+00	0.00E+00
T10-J5	60.0	1.9E+08	27	0	0	2	1	0	0	0.26	0.16	1.00	1.00	0.00E+00	0.00E+00
T10-J6	69.0	2.0E+08	27	0	0	3	0	1	1	0.26	0.16	1.00	0.00	0.00E+00	3.04E-09
T10-M1	56.0	1.9E+08	26	2	2	6	2	1	1	0.26	0.16	0.75	0.50	7.92E-09	7.75E-09
T10-M2	22.0	4.0E+07	5	0	0	5	0	1	0	0.26	0.16	1.00	0.00	0.00E+00	0.00E+00
T10	**52.5**	**1.7E+09**	**23**	**5**	**3**	**35**	**3**	**9**	**4**	0.26	0.16	**0.88**	**0.50**	**2.15E-09**	**1.76E-09**
Total	**47.1**	**5.8E+09**	**40**	**24**	**21**	**71**	**11**	**36**	**11**	**0.26**	**0.16**	**0.75**	**0.34**	**3.39E-09**	**2.90E-09**

Note: The first column shows names of individuals, which include their family name (T7 or T10) and codes for the individual type: O—mother, A—father, F—female offspring, M—male offspring and J—offspring of unknown sex (juvenile). The second column shows mean coverage in sequenced individuals. The third and fourth columns show the effective numbers of sites that were sequenced in the genomes of each individual. Next columns present number of True/False/Failed candidate de novo mutations detected by machine learning (ML) and hard filtering (HF) approaches. The last six columns show false-negative rate (FNR) estimated using in silico simulations, false discovery rate (FDR) and mutation rate estimates.

TABLE 2 Total number of SNPs found by the two SNP calling methods employed in the T7 and T10 families and the numbers shared.

	T7	T10	Shared
samtools/bcftools	6,355,935	6,729,594	5,093,623
GATK	6,380,136	6,764,192	4,944,266
Shared	5,441,537	5,780,424	4,275,448

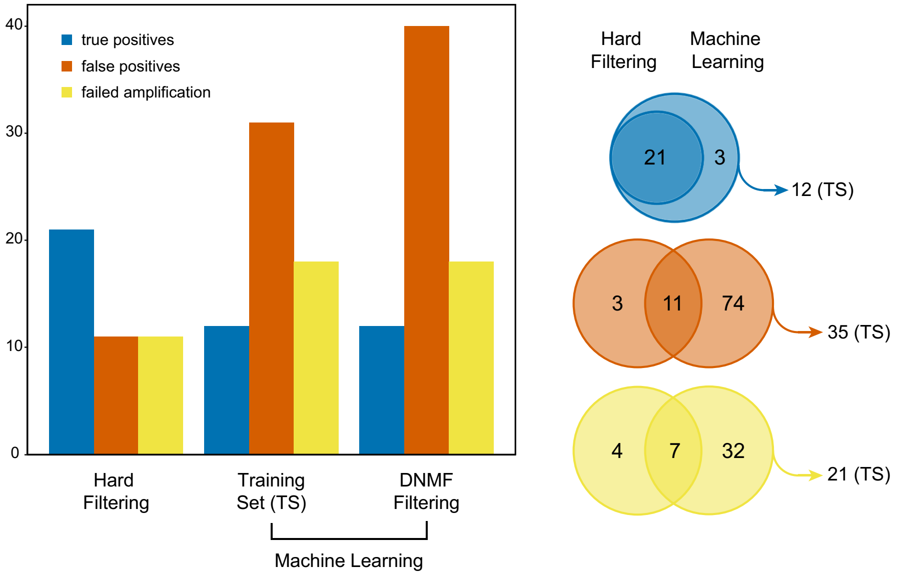


FIGURE 1 Difference between two methods of variant filtering and the overlap. On the left, *X*-axis shows three sets of candidate de novo mutations. First set shows mutations that were successfully validated after applying hard filtering. The two other sets are mutations successfully validated during the machine learning approach: the first being candidates of the training set prepared for the DNMF tool, the other being candidates chosen by the tool. All numbers shown do not contain mutations in repetitive regions. On the right, Venn diagrams showing overlap between candidate mutations found using two methods. Colours correspond to those in the left plot. In each row, arrow points out the number of mutations within the diagram that belong to the training set (TS) in the ML method.

Methods and Table S1 for details). During this process, we identified 21 true mutations and 14 false-positive mutations (11, with repetitive regions masked; Figure 1). Eleven candidate mutations could not be verified due to PCR/Sequencing errors. For clarity, the effect of each filtering step is presented at Figure S7.

To estimate mutation rate using an ML method, a subset of 49 unique sites (accounting for 75 mutations in total, as some were present in more than one individual) were selected for validation based on specific filters and visual evaluation in IGV (see Section 2). These mutations were validated using PCR and Sanger sequencing. Fifty-four mutations (at 35 sites in total) were successfully validated (73%), while the others could not be validated (Table S1). After this initial validation, we applied the DNMF tool to the remaining candidate de novo mutations (together with those for which validation failed), using the successfully validated true or false mutations as a training set. The DNMF tool reduced the number of de novo mutations from 239 to 47 sites (probability cut-off = 0.4). A total of 35 sites (55 mutations) were successfully validated during the ML validation (75%), adding 12 unique true DNMs (Figure 1, Figure S7). In total, 24 true de novo mutations were found at 22 sites (Table 3) using ML approach.

We then checked how many of the ML method candidate DNMs were localized in repetitive regions. We masked 6.18% of the reference genome (1.55%—retroelements, 2.83%—DNA transposons, 1.45%—simple repeats and 0.25%—low-complexity regions) and found 166 such candidate mutations (localized in 77 sites). Of these mutations, 57 mutations (24 sites) were excluded from our analyses prior to filtration due to low mappability, 85 mutations (45 sites) had DNMF scores below 0.4, 3 mutations (2 sites) failed validation and 14 mutations (6 sites) were validated as false positives. In contrast, hard filtering selected only three candidate mutations (3 sites) in repetitive regions and all of these candidates were validated as false positives. We did not find a single true mutation in regions that could have been masked using RepeatMasker. Candidate DNMs that were filtered out due to low mappability, but outside the RepeatMasker output, included 124 mutations (51 sites).

Seven mutations (at six sites) were uniquely identified as candidate DNMs by hard filtering, three of which were verified as false positives and for four (at 3 sites) molecular validation failed (Table S1). One hundred and nine mutations (at 59 sites) were uniquely identified by the ML method—74 (at 36 sites) were verified as false positives, 32 (at 11 sites) failed molecular verification and three true DNMs (at 3 sites) were detected solely by the ML method.

Two DNMs originated in the maternal parent, and seven DNMs had paternal origins, origin of the others could not be determined. Four DNMs were transversions, whereas 20 were identified as

TABLE 3 Table of all true de novo mutations identified in this study.

Chr	Position	Parents	Offspring	Individual	Detected by hard filtering	Part of the training set	DNMF probability	Mutation type	Parental origin
1	11,320,387	CTT**T**GAA	CTT**C**GAA	T10-J-1	Yes	No	0.79	TS	NA
2	7,907,610	GAG**C**GAA	GAG**T**GAA	T7-F-1	Yes	Yes	NA	TS	NA
2	29,464,863	CTT**T**TTG	CTT**G**TTG	T10-F-2	No	Yes	NA	TV	NA
2	28,098,628	AAC**C**ATA	AAC**T**ATA	T7-M-2	Yes	No	0.44	TS	NA
3	6,123,659	GGC**G**CTT	GGC**A**CTT	T7-J-3	Yes	No	0.88	TS	Father
3	31,364,266	AGA**C**AAA	AGA**T**AAA	T7-J-1	Yes	No	0.85	TS	Father
4	15,397,752	TCA**C**CCA	TCA**T**CCA	T7-M-2	Yes	No	0.71	TS	Father
4	29,856,266	TTT**C**ATA	TTT**A**ATA	T7-J-4	Yes	No	0.56	TV	NA
7	26,525,718	ACT**C**TTC	ACT**T**TTC	T7-J-3 T7-M-2	Yes	Yes	NA	TS	NA
10	23,542,367	TGA**C**CAA	TGA**T**CAA	T7-F-4	Yes	Yes	NA	TS	Father
10	25,291,101	GAG**C**TAC	GAG**T**TAC	T7-M-2	Yes	Yes	NA	TS	NA
12	17,229,793	AAG**C**ACG	AAG**T**ACG	T10-M-1	Yes	Yes	NA	TS	Mother
12	19,229,557	TGC**C**TCA	TGC**T**TCA	T7-F-2 T7-F-4	Yes	Yes	NA	TS	NA
14	21,187,135	TTT**T**TAT	TTT**C**TAT	T7-F-1	Yes	No	0.76	TS	NA
14	23,760,915	GTT**T**ATC	GTT**G**ATC	T7-F-4	Yes	No	0.74	TV	Mother
15	18,795,681	AAG**G**TTC	AAG**A**TTC	T7-M-1	Yes	Yes	NA	TS	Father
16	7,767,856	AAA**A**CAC	AAA**C**CAC	T7-J-3	Yes	Yes	NA	TV	NA
17	15,099,629	♀ GCT**G**TTA ♂ RCT**G**CTA	GCT**A**TTA	T7-F-2	Yes	No	0.41	TS	Mother
18	17,812,487	ACA**A**CTG	ACA**G**CTG	T7-J-3	No	No	0.50	TS	Father
19	6,887,521	CAT**T**AAA	CAT**C**AAA	T10-J-2	No	No	0.50	TS	NA
21	3,258,922	ACC**G**TTG	ACC**A**TTG	T7-J-3	Yes	Yes	NA	TS	NA
23	14,844,758	GTC**C**ACC	GTC**T**ACC	T10-M-1	Yes	No	0.74	TS	Father

Note: The third and fourth columns show the nucleotides surrounding the DNM site in the parents and offspring. The de novo mutation site is underlined. In the case of the mutation found on chromosome 17, the maternally and paternally derived sequences differ and are given separately. The seventh and eighth columns show whether the mutation was a part of the DNMF tool training set and the probability given by DNMF if the mutation passed the filtration. Note that mutations that were part of the training set were not filtered so they do not have the DNMF probability score. The ninth column shows the mutation type, transition (TS) or transversion (TV). The tenth column shows whether the mutation originated from the mother or father.

transitions (for details, see Table 3). Most sites with DNMs shared between two or more siblings were found to be false positives, but true mutations were also found (Figure S8).

3.3 | False-negative rates estimation

Using the 'known SNPs' method of estimating FNR, we found it to be 0.02 in our dataset. Another way to calculate FNR was to simulate in silico mutations using reads sequenced from the T7-F-1 individual. We successfully mutated 1210 sites on chromosome 1. Sequencing reads were then mapped back to reference genome and used for SNP calling and candidate DNMs identification. While the machine learning approach (using the training set described above and cut-off = 0.4) found 901 simulated DNMs, hard filtering identified 1012 simulated de novo mutations. Knowing these values, we calculated FNR to be 0.26 in the case of DNMF method and 0.16 in the case of hard filtering. Since these values are an order of magnitude greater and probably more accurate than those obtained by the 'known SNPs' method, we decided to use them for the mutation rate estimation (see Section 4 for more details).

3.4 | Mutation rate estimation

The total number of effective sites (see Section 2) was 5,795,617,080, or approximately 40% of all sites present in all progeny individual genomes (Table 1 and Figure S9). In case of the hard filtering method, a false discovery rate (FDR), the number of false candidate DNMs divided by all successfully validated DNMs, equalled 0.40. The mutation rate estimated using hard filtering is then 2.90×10^{-9} per site per generation (with 95% confidence intervals $1.92–3.88 \times 10^{-9}$), and it did not differ significantly between families (T7: 3.31×10^{-9} [$2.07–4.56 \times 10^{-9}$]; T10: 1.76×10^{-9} [$0.34–3.17 \times 10^{-9}$]). For the ML approach, FDR was calculated as false candidate DNMs divided by all successfully validated candidate DNMs selected by the DNMF software with a cut-off = 0.4. This process yielded a value of 0.75. Combining the information about the number of de novo mutations identified, the estimated FDR and the FNR, as well as the effective number of sites, we estimated the guppy mutation rate to be 3.39×10^{-9} per site per generation for the machine learning method (95% CI: $2.33–4.50 \times 10^{-9}$) and again this did not differ between families (T7: 4.11×10^{-9} [$2.72–5.50 \times 10^{-9}$]; T10: 2.15×10^{-9} [$0.19–3.32 \times 10^{-9}$]). The differences between hard filtering and the machine learning approach were not statistically significant.

To show how mutation rate can be used in further analyses, we used the mutation rate estimated above (using the hard filtering approach) to scale the trajectory of effective population size (Ne), estimated using the PSMC (Pairwise Sequentially Markovian Coalescent) software. The analyses were performed for genome sequences of each of our four parents. For all four individuals, analyses suggest that the population experienced a steady decline in Ne over the last 50 thousand years (Figure 2).

4 | DISCUSSION

During our analysis, we estimated the *Poecilia reticulata* (guppy) whole-genomic, per site, per generation point mutation rate. We used a pedigree-based approach to find variants that were new in the offspring generation (de novo mutations, DNMs). We sequenced two guppy families to high coverage (on average 47×). We screened the data for candidate DNMs and found a large number of variants matching our quality threshold and patterns of DNMs. We then used two alternative methods, a machine learning (ML) and a hard filtering approach, to remove false positives. In case of ML, first, a training set was prepared. We amplified and sequenced a set of DNMs in an offspring and their parents (see Section 2), obtaining true and false mutations. We then used these true and false mutations to train the machine learning tool to characterize true and false DNMs. Subsequently, the programme analysed the remaining candidate DNMs. Mutations that underwent this step were also verified using Sanger sequencing and subsequently used to calculate the mutation rate ($\mu = 3.39 \times 10^{-9}$). The other method was based on hard filtering using the INFO field of the VCF file, allelic balance and depth information. The estimated mutation rate (2.9×10^{-9}) was around 15% lower than in ML, but the difference between these two methods was not statistically significant. The number of identified DNMs was almost four times higher in the T7 family, but accounting for differences in the number of effective sites, the mutation rate did not differ between families.

Even though we found more true DNMs using the machine learning method, this approach had higher estimates of false discovery rate and false-negative rate. This is because the ML method required a much larger number of PCR validations for training set generation and for actual candidate DNMs verification. As a consequence, we tested 43 candidates (not being part of the training set) in the ML approach vs only 35 candidates in the hard filtering (the two sets overlapped, Figure 1). Results of additional ML validations were well predicted by in silico simulations, assuming that we identified nearly all de novo mutations. Namely, 12% (3/24) of all true DNMs were identified by the ML approach, but not by hard filtering, whereas the estimated false-negative rate (FNR) for hard filtering equalled 16%. FNR calculated using the 'known SNPs' method is much smaller

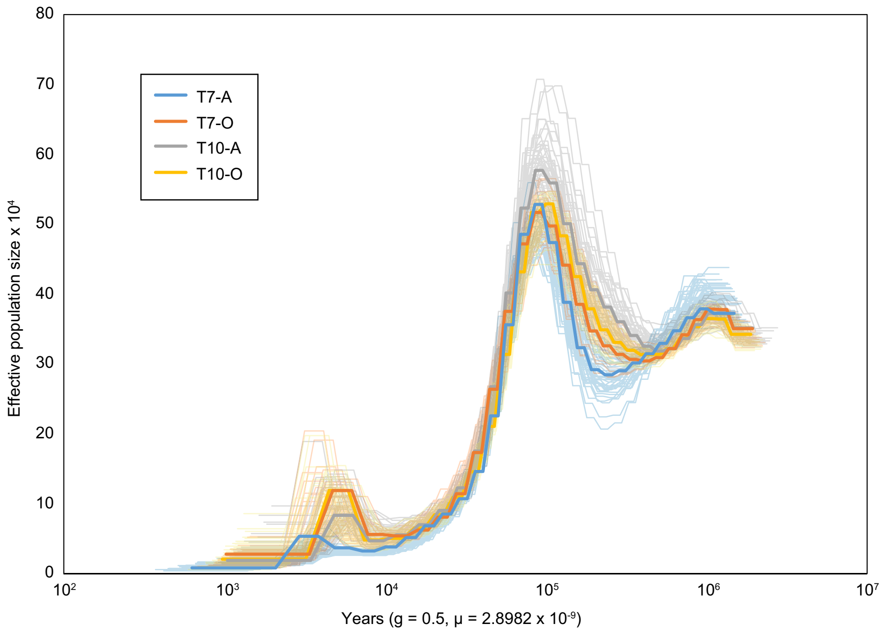


FIGURE 2 Estimated changes in effective population size of the guppy population, calculated using the PSMC method. The most recent times are at the left, and increasingly long times in the past are at the right. The four parental individuals, used for independent calculations, are shown in different colours. Pale, thin lines represent bootstrapping results. The plot was scaled using our mutation rate estimate, and two generations per year.

(about 2%), but most likely it is strongly underestimated, due to its insensitivity to mis-mapping and variant calling bias towards alleles, which (unlike most DNMs) are present in multiple samples. It is clear that the simulation-based method allows DNMs to go through the whole process of their identification (from reads to VCF file) and its general agreement with additional sites identified by ML approach suggests it performs well.

We, however, note that in silico simulations might not capture the whole complexity of biases present in the dataset and underestimate FNR. This effect might be especially strong for hard filtering, which is insensitive to experiment-specific biases. In contrast, the ML approach should better control such nuances between false and true candidate DNMs thanks to the training set. We, therefore, compared sites that were filtered out by the DNMF tool (ML approach) to those filtered out by hard filtering. For both methods, the most common reason to filter out sites from candidate DNMs was allelic balance (AB; Figure 3). In theory, in a heterozygous individual, the probability of sequencing each allele is the same and thus expected AB should equal 0.5. Due to sampling errors, sequencing mistakes, sequencing and mapping biases AB can significantly deviate from expectation and in hard filtering AB in range between 0.3 and 0.7 was accepted (Bergeron et al., 2022). Most of the sites filtered out by DNMF were also outside this range (Figure 3b). Somatic mutations can also produce false positives with allelic balance much smaller than expected of germline mutations (≈0.5). Indeed, most of the candidate DNMs had AB around 0.2, while true germline-inherited variants had equal frequencies (Figure S10), consistent with the pattern previously reported for the rhesus macaque (Bergeron et al., 2021). The other filters also followed similar patterns for both methods, with the significant importance of allelic depth in the parent (no reads with the '*de novo* mutated' nucleotide) and QUALITY/DP filters (Figure 3). We, therefore, did not observe evidence suggesting significant differences between the ML and hard filtering in candidate filtering patterns, but noted that, even for hard filtering, FDR is

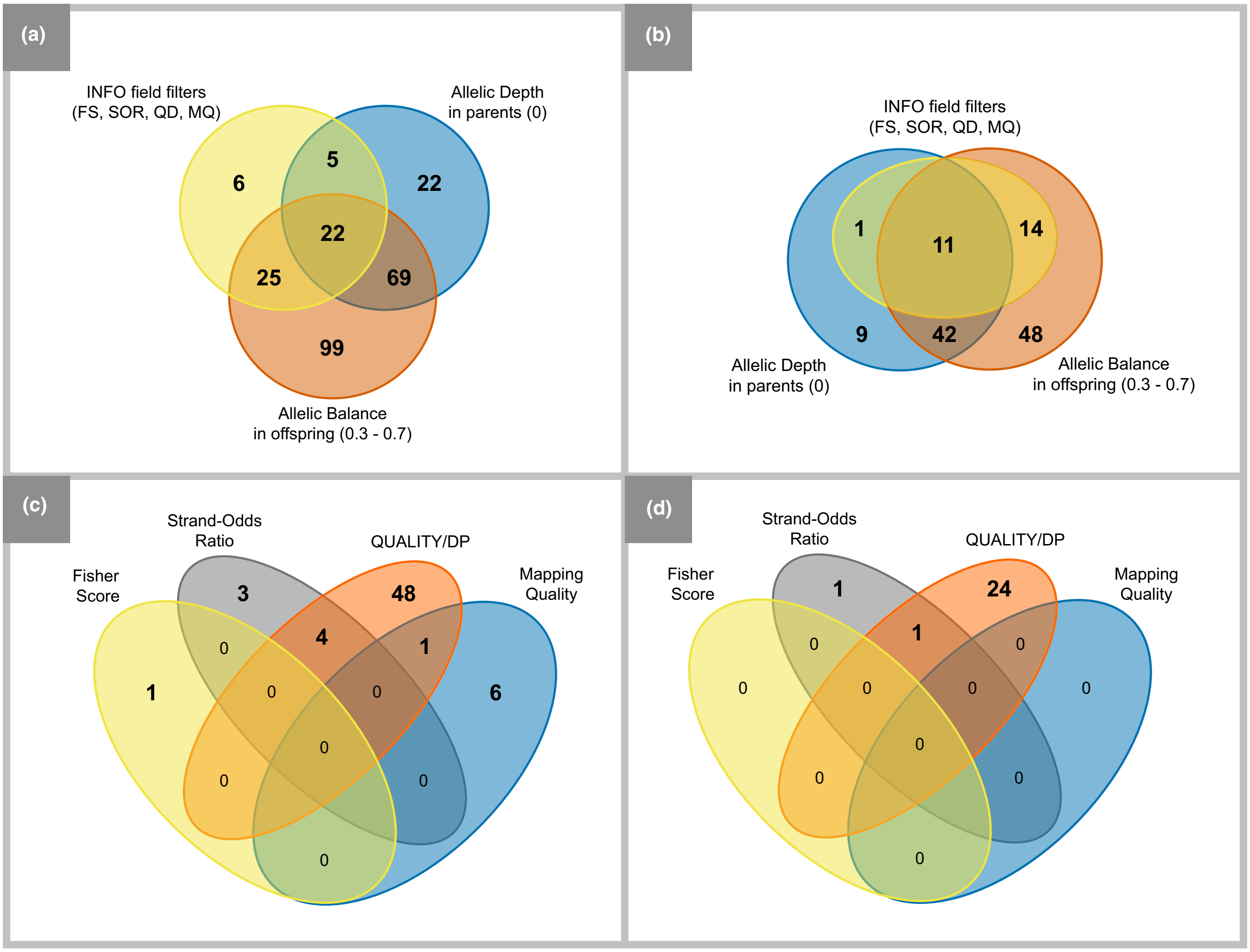

(a)
INFO field filters (FS, SOR, QD, MQ)
Allelic Depth in parents (0)
6
5
22
22
25
69
99
Allelic Balance in offspring (0.3 - 0.7)
(b)
INFO field filters (FS, SOR, QD, MQ)
1
11
14
9
42
48
Allelic Depth in parents (0)
Allelic Balance in offspring (0.3 - 0.7)
(c)
Strand-Odds Ratio
QUALITY/DP
Fisher Score
Mapping Quality
3
48
4
1
0
1
0
0
6
0
0
0
0
(d)
Strand-Odds Ratio
QUALITY/DP
Fisher Score
Mapping Quality
1
24
1
0
0
0
0
0
0
0
0
0
0

FIGURE 3 Number of candidate de novo mutations (DNMs) not meeting criteria of hard filtering according to specific filters. Panel a and c show mutations filtered out by hard filtering approach, and panels b and d show mutations filtered out by machine learning. Specific filters represent: no reads with candidate DNM in parents (Allelic Depth in parents), frequency of reads with candidate DNM in offspring between 0.3 and 07 (Allelic Balance in offspring) and filters applied to INFO field scores (panels c and d): Fisher Score ≤60, Strand-Odd Ratio ≤3, Quality/DP QD ≥2 and mean Mapping Quality ≥40.

substantial, significantly higher than in Bergeron et al. (2022), who claimed that the final filtering strategy should be adjusted to a given study. Such adjustments might differ between studies due to various aspects including for example coverage, number of sequenced siblings or genome complexity. The decisions can be, however, guided by the actual dataset. For example, for most of the statistics, we see good concordance between true mutations and heterozygous genotypes of offspring whose parents are alternative homozygous (but see Quality/Depth; Figure S10). Alternatively, such heterozygous offspring sites might be used as a much larger training set for a ML approach, which should further improve its performance. It is worth mentioning that the DNMF tool uses a much wider set of filters based on such measures as mean base quality, strand direction, fraction of MQ0 reads, fraction of soft clipped reads, mean number of nearby mismatches, mean number of INDELs, paired samples test, etc. (Liu et al., 2014). These values are read directly from the bam files, and consequently, it should be relatively easy to extend them to custom statistics. Therefore, the DNMF-based approach has the potential to identify experiment-specific biases. For example, it may be possible to apply it to already published studies with sets of verified candidate DNMs. This information can be used as training sets and later the ML approach can be used to assess filtering patterns. On the contrary, our results seem to suggest that such approach still poses quite a high risk of identification of false candidates, many of which might be in repetitive regions.

While estimating the mutation rate, among all candidate DNMs, 26% had a mappability score lower than 1, which means that reads mapping to these loci probably could not have been mapped unambiguously to the reference genome and thus were prone to errors. In the whole genome, the fraction of sites with mappability scores lower than 1 rounds up to 2%. The very high fraction of candidate DNMs found at such sites demonstrates that this simple statistic can be used to effectively filter problematic sites. Such problematic sites are often masked before SNP calling, usually by masking repetitive elements. We applied this step at the end of the analyses, so we could verify its importance. We found that no mutations that failed validation were within identified repeats and all of them were found to be false positives. It is, however, likely that many transposable elements were not identified in our analyses. This might have especially affected the less stringent ML approach, as suggested by a larger number of candidate DNMs with failed validation (Figure 1). Additionally, we have not identified a single true mutation in repetitive sequences. Our study thus demonstrates that both masking repeats and masking sites with low mappability can reduce the rate of false positives while losing little of the sensitivity.

Based on the above arguments, we claim that our ML approach identified nearly all de novo mutations. Many of these mutations were however already identified for the training set, which, in combination with selection of candidates in repetitive regions and subsequent failed molecular validation, caused a slight overestimation of mutation rate. For the rest of the discussion, we therefore use the mutation rate estimation produced by hard filtering—the approach applied also for other species.

The mutation rate estimated for the guppy here (2.9×10^{-9}) is considerably lower than rates estimated from many other vertebrates, especially primates (humans: 1.22×10^{-8} from Kessler et al., 2020, chimpanzee: 1.26×10^{-8} from Besenbacher et al., 2019). However, it is quite close to pedigree-based estimates in other teleost species, the Atlantic herring (2.0×10^{-9}, Feng et al., 2017) and cichlids (3.5×10^{-9}, Malinsky et al., 2018), despite an order of magnitude difference in generation time (0.5 vs. 5 years in guppies and herrings, respectively; Reznick et al., 1997; Barrett et al., 2022). A recent, large-scale study found lower mutation rates in fishes compared with reptiles and birds; however, the difference was not statistically significant (Bergeron et al., 2023). Such a potential pattern may result from long-term large effective population sizes (Ne) of analysed species, as predicted by the drift-barrier hypothesis (Lanfear et al., 2014; Lynch et al., 2016; Sung et al., 2012). According to this prediction, a direct, a negative relationship exists between the mutation rate and Ne resulting from the efficiency of selection acting in large populations and working towards high replication fidelity. Such a negative relationship between mutation rate and long-term Ne was detected, but should be interpreted with caution due to confounding effects (Bergeron et al., 2023). Alternatively, linage-specific modifications can contribute to a low mutation rate in teleosts.

For example, a low metabolic rate in these ectotherm animals may lead to low mutation rates. According to Brett (1972), even highly active fish such as salmon have metabolic rates 10–100 times lower than the metabolic rates of mammals or birds. Slow metabolism and low oxygen consumption lead to reduced free radical production, which in turn can limit DNA damage (Martin & Palumbi, 1993). However, some of the previous findings are not consistent with the metabolic hypothesis; for example, metabolic rate does not predict substitution rate inferred in a handful of genes studied across scombroid fishes and across a much broader range of other metazoans (Lanfear et al., 2007; Qiu et al., 2014; Yoder & Tiley, 2021). In contrast, recent studies suggest that a significant fraction of mutations might be damage-induced rather than replication-induced (de Manuel et al., 2022; Gao et al., 2016). In such a case, variation in metabolic rate and in other damage-inducing factors should receive additional attention when studying mutation rate evolution. For example, an often used argument for the environment having little influence on the germline mutation rate is that germ cells are fertilized inside the organism, where the embryo development also takes place. The argument holds for species with internal fertilization, like guppies, but many other fish species release sperm and eggs to water before fertilization. In such a case, damage-induced mutations might be triggered by environmental factors. So far we know little about differences in germline mutation rates between species with internal and external fertilization.

In addition to damage-induced mutations, the fidelity of the replication machinery might differ between teleosts and other vertebrates. The difference between fishes and mammals, for example, might be linked to decreasing fidelity of polymerase δ in ancestral mammals, as was suggested by a study demonstrating unexpectedly strong heterogeneity among vertebrates in the rate of DNA pol δ evolution (Katoh et al., 2020). Likewise, the efficiency of other elements of the replication machinery might differ between evolutionary lineages (Yao & O'Donnell, 2016).

Yet, another interesting observation differentiate direct mutation rate estimates in teleosts and other vertebrates. Two positively validated mutations were present in two siblings, suggesting that the mutation happened early in gametogenesis, or alternatively was related to mosaicism in one of the parents. We were not able to distinguish between these two hypotheses, but we noted that in the Atlantic herring eight out of seventeen DNMs were carried by more than one sibling (Feng et al., 2017). Around 50 offspring individuals were screened per family in this experiment, while in many other vertebrate parent–offspring sequencing studies only one to two offspring per family are investigated (e.g. Campbell et al., 2021; Wang et al., 2021). In such cases, sites with identified DNMs might be at least Sanger sequenced in a larger number of siblings to assess frequency of mutations that enter the population with more than one copy.

Despite extensive sequencing efforts, many studies suffer from a small number of identified DNMs, mostly because of the extremely low mutation rate. This limitation has prevented the growth of pedigree-based mutation rate studies despite their high potential impact. Fortunately, it has changed in recent years, when direct estimates of mutation rate have been obtained for dozens of vertebrate species. The field is, however, still in its infancy, with its uncertainties regarding technical aspects (Bergeron et al., 2022; Yoder & Tiley, 2021), mostly because de novo mutations are on the order of magnitude rarer than sequencing errors, meaning that in reality, we are often looking for needles in a haystack (Yoder & Tiley, 2021). Many of the technical challenges have recently been discussed, and guidelines for computational and statistical benchmarks were provided (Bergeron et al., 2022). Here, we demonstrated that these methods provide accurate estimates of mutation rate but miss identification of all de novo mutations (the ML approach found three more). In some specific cases, when all mutations are required to be identified, or when experiment-specific biases exist, extensive molecular validation combined with machine learning algorithms might be a good alternative.

AUTHOR CONTRIBUTIONS

MK designed the research, KB and MK performed the research, analysed the data and wrote the article.

ACKNOWLEDGEMENTS

We would like to thank Wiesław Babik, Deborah Charlesworth, Krystyna Nadachowska-Brzyska, Alana Alexander and three anonymous reviewers for their helpful comments on the early versions of the manuscript. We thank Magdalena Herdegen-Radwan and Jarosław Raubic for providing samples. We also thank Adam Ewing for substantial help with running BAMSurgeon. This work was funded by a Polish National Science Centre 2018/31/D/NZ8/00091 grant. M.K. was supported by the Polish National Agency for Academic Exchange under Bekker Programme. Computations were performed at the Poznan Supercomputing and Networking Centre.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Genomic data (raw fastq files) are available under NCBI BioProject PRJNA864588.

BENEFIT-SHARING

Benefits from this research come from the sharing of our data in public databases as described above.

ORCID

Mateusz Konczal https://orcid.org/0000-0002-7691-8075

REFERENCES

Aikens, R. C., Johnson, K. E., & Voight, B. F. (2019). Signals of variation in human mutation rate at multiple levels of sequence context. *Molecular Biology and Evolution*, *36*(5), 955–965. https://doi.org/10.1093/molbev/msz023

Babraham Bioinformatics. (2010). *FastQC: A quality control tool for high throughput sequence data*. Babraham Bioinformatics.

Barrett, T. J., Hordyk, A. R., Barrett, M. A., & van den Heuvel, M. R. (2022). Spatial and temporal differences in fecundity of Atlantic herring (*Clupea harengus*) off Nova Scotia and consequences for biological reference points. *Canadian Journal of Fisheries and Aquatic Sciences*, *79*(7), 1086–1096. https://doi.org/10.1139/cjfas-2021-0269

Bergero, R., Gardner, J., Bader, B., Yong, L., & Charlesworth, D. (2019). Exaggerated heterochiasmy in a fish with sex-linked male coloration polymorphisms. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(14), 6924–6931. https://doi.org/10.1073/pnas.1818486116

Bergeron, L. A., Besenbacher, S., Bakker, J., Zheng, J., Li, P., Pacheco, G., Sinding, M. S., Kamilari, M., Gilbert, M. T. P., Schierup, M. H., & Zhang, G. (2021). The germline mutational process in rhesus macaque and its implications for phylogenetic dating. *GigaScience*, *10*(5), giab029. https://doi.org/10.1093/gigascience/giab029

Bergeron, L. A., Besenbacher, S., Turner, T. N., Versoza, C. J., Wang, R. J., Price, A. L., Armstrong, E., Riera, M., Carlson, J., Chen, H. Y., Hahn, M. W., Harris, K., Kleppe, A. S., López-Nandam, E. H., Moorjani, P., Pfeifer, S. P., Tiley, G. P., Yoder, A. D., Zhang, G., & Schierup, M. H. (2022). The mutationathon highlights the importance of reaching standardization in estimates of pedigree-based germline mutation rates. *eLife*, *11*, 1–28. https://doi.org/10.7554/eLife.73577

Bergeron, L. A., Besenbacher, S., Zheng, J., Li, P., Bertelsen, M. F., Quintard, B., Hoffman, J. I., Li, Z., St Leger, J., Shao, C., Stiller, J., Gilbert, M. T. P., Schierup, M. H., & Zhang, G. (2023). Evolution of the germline mutation rate across vertebrates. *Nature*, *615*, 285–291. https://doi.org/10.1038/s41586-023-05752-y

Besenbacher, S., Hvilsom, C., Marques-Bonet, T., Mailund, T., & Schierup, M. H. (2019). Direct estimation of mutations in great apes reconciles phylogenetic dating. *Nature Ecology and Evolution*, *3*(2), 286–292. https://doi.org/10.1038/s41559-018-0778-x

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Brett, J. R. (1972). The metabolic demand for oxygen in fish, particularly salmonids, and a comparison with other vertebrates. *Respiration Physiology*, *14*(1–2), 151–170. https://doi.org/10.1016/0034-5687(72)90025-4

Brooks, R., & Endler, J. A. (2001). Direct and indirect sexual selection and quantitative genetics of male traits in guppies (*Poecilia reticulata*). *Evolution*, *55*(5), 1002–1015. https://doi.org/10.1111/j.0014-3820.2001.tb00617.x

Campbell, C. R., Tiley, G. P., Poelstra, J. W., Hunnicutt, K. E., Larsen, P. A., Lee, H. J., Thorne, J. L., dos Reis, M., & Yoder, A. D. (2021). Pedigree-based and phylogenetic methods support surprising patterns of mutation rate and spectrum in the gray mouse lemur. *Heredity*, *127*(2), 233–244. https://doi.org/10.1038/s41437-021-00446-5

Ceballos, F. C., Joshi, P. K., Clark, D. W., Ramsay, M., & Wilson, J. F. (2018). Runs of homozygosity: Windows into population history and trait architecture. *Nature Reviews Genetics*, *19*(4), 220–234. https://doi.org/10.1038/nrg.2017.109

Charlesworth, D., Zhang, Y., Bergero, R., Graham, C., Gardner, J., & Yong, L. (2020). Using GC content to compare recombination patterns on the sex chromosomes and autosomes of the guppy, *Poecilia reticulata*, and its close outgroup species. *Molecular Biology and Evolution*, *37*(12), 3550–3562. https://doi.org/10.1093/molbev/msaa187

Conrad, D. F., Keebler, J. E. M., Depristo, M. A., Lindsay, S. J., Zhang, Y., Casals, F., Idaghdour, Y., Hartl, C. L., Torroja, C., Garimella, K. V., Zilversmit, M., Cartwright, R., Rouleau, G. A., Daly, M., Stone, E. A., Hurles, M. E., & Awadalla, P. (2011). Variation in genome-wide mutation rates within and between human families. *Nature Genetics*, *43*(7), 712–714. https://doi.org/10.1038/ng.862

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Darolti, I., Wright, A. E., & Mank, J. E. (2020). Guppy Y chromosome integrity maintained by incomplete recombination suppression. *Genome Biology and Evolution*, *12*(6), 965–977. https://doi.org/10.1093/GBE/EVAA099

de Manuel, M., Wu, F. L., & Przeworski, M. (2022). A paternal bias in germline mutation is widespread across amniotes and can arise independently of cell divisions. *BioRxiv*. https://www.biorxiv.org/content/10.1101/2022.02.07.479417v1

Ewing, A. D., Houlahan, K. E., Hu, Y., Ellrott, K., Caloian, C., Yamaguchi, T. N., Bare, J. C., P'ng, C., Waggott, D., Sabelnykova, V. Y., ICGC-TCGA DREAM Somatic Mutation Calling Challenge participants, Kellen, M. R., Norman, T. C., Haussler, D., Friend, S. H., Stolovitzky, G., Margolin, A. A., Stuart, J. M., & Boutros, P. C. (2015). Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nature Methods*, *12*(7), 623–630. https://doi.org/10.1038/nmeth.3407

Eyre-Walker, A., & Keightley, P. D. (2007). The distribution of fitness effects of new mutations. *Nature Reviews Genetics*, *8*(8), 610–618. https://doi.org/10.1038/nrg2146

Farrer, R. A., Henk, D. A., MacLean, D., Studholme, D. J., & Fisher, M. C. (2013). Using false discovery rates to benchmark SNP-callers in next-generation sequencing projects. *Scientific Reports*, *3*(1), 1512. https://doi.org/10.1038/srep01512

Feliciano, P., Zhou, X., Astrovskaya, I., Turner, T. N., Wang, T., Brueggeman, L., Barnard, R., Hsieh, A., Snyder, L. A. G., Muzny, D. M., Sabo, A., Abbeduto, L., Acampado, J., Ace, A. J., Albright, C., Alessandri, M., Amaral, D. G., Amatya, A., Annett, R. D., … Chung, W. K. (2019). Exome sequencing of 457 autism families recruited online provides evidence for autism risk genes. *NPJ Genomic Medicine*, *4*(1), 19. https://doi.org/10.1038/s41525-019-0093-8

Feng, C., Pettersson, M., Lamichhaney, S., Rubin, C. J., Rafati, N., Casini, M., Folkvord, A., & Andersson, L. (2017). Moderate nucleotide diversity in the Atlantic herring is associated with a low mutation rate. *eLife*, *6*, 1–14. https://doi.org/10.7554/eLife.23907

Fraser, B. A., & Neff, B. D. (2010). Parasite mediated homogenizing selection at the MHC in guppies. *Genetica*, *138*(2), 273–278. https://doi.org/10.1007/s10709-009-9402-y

Gao, Z., Wyman, M. J., Sella, G., & Przeworski, M. (2016). Interpreting the dependence of mutation rates on age and time. *PLoS Biology*, *14*(1), 1–16. https://doi.org/10.1371/journal.pbio.1002355

Habig, M., Lorrain, C., Feurtey, A., Komluski, J., & Stukenbrock, E. H. (2021). Epigenetic modifications affect the rate of spontaneous mutations in a pathogenic fungus. *Nature Communications*, *12*(1), 1–13. https://doi.org/10.1038/s41467-021-26108-y

Harris, K. (2015). Evidence for recent, population-specific evolution of the human mutation rate. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(11), 3439–3444. https://doi.org/10.1073/pnas.1418652112

Harris, K., & Pritchard, J. K. (2017). Rapid evolution of the human mutation spectrum. *eLife*, *6*, 1–17. https://doi.org/10.7554/eLife.24284

Jónsson, H., Sulem, P., Kehr, B., Kristmundsdottir, S., Zink, F., Hjartarson, E., Hardarson, M. T., Hjorleifsson, K. E., Eggertsson, H. P., Gudjonsson, S. A., Ward, L. D., Arnadottir, G. A., Helgason, E. A., Helgason, H., Gylfason, A., Jonasdottir, A., Jonasdottir, A., Rafnar, T., Frigge, M., … Stefansson, K. (2017). Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature*, *549*(7673), 519–522. https://doi.org/10.1038/nature24018

Katoh, K., Iwabe, N., & Miyata, T. (2020). Possible changes in fidelity of DNA polymerase δ in ancestral mammals. *BioRxiv*. http://biorxiv.org/content/early/2020/11/01/2020.10.29.327619.abstract

Kessler, M. D., Loesch, D. P., Perry, J. A., Heard-Costa, N. L., Taliun, D., Cade, B. E., Wang, H., Daya, M., Ziniti, J., Datta, S., Celedón, J. C., Soto-Quiros, M. E., Avila, L., Weiss, S. T., Barnes, K., Redline, S. S., Vasan, R. S., Johnson, A. D., Mathias, R. A., … O'Connor, T. D. (2020). De novo mutations across 1,465 diverse genomes reveal mutational insights and reductions in the Amish founder population. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(5), 2560–2569. https://doi.org/10.1073/pnas.1902766117

Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge University Press. https://doi.org/10.1017/CBO9780511623486

Kimura, M., & Ohta, T. (1971). On the rate of molecular evolution. *Journal of Molecular Evolution*, *1*(1), 1–17. https://doi.org/10.1007/BF01659390

Koch, E. M., Schweizer, R. M., Schweizer, T. M., Stahler, D. R., Smith, D. W., Wayne, R. K., & Novembre, J. (2019). De novo mutation rate estimation in wolves of known pedigree. *Molecular Biology and Evolution*, *36*(11), 2536–2547. https://doi.org/10.1093/molbev/msz159

Kondrashov, F. A., & Kondrashov, A. S. (2010). Measurements of spontaneous rates of mutations in the recent past and the near future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1544), 1169–1176. https://doi.org/10.1098/rstb.2009.0286

Kong, A., Frigge, M. L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S. A., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A., Wong, W. S. W., Sigurdsson, G., Walters, G. B., Steinberg, S., Helgason, H., Thorleifsson, G., Gudbjartsson, D. F., Helgason, A., Magnusson, O. T., … Stefansson, K. (2012). Rate of de novo mutations and the importance of father-s age to disease risk. *Nature*, *488*(7412), 471–475. https://doi.org/10.1038/nature11396

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution*, *35*(6), 1547–1549. https://doi.org/10.1093/molbev/msy096

Kunstner, A., Hoffmann, M., Fraser, B. A., Kottler, V. A., Sharma, E., Weigel, D., & Dreyer, C. (2016). The genome of the trinidadian guppy, *Poecilia reticulata*, and variation in the Guanapo population. *PLoS One*, *11*(12), 1–25. https://doi.org/10.1371/journal.pone.0169087

Lanfear, R., Kokko, H., & Eyre-Walker, A. (2014). Population size and the rate of evolution. *Trends in Ecology and Evolution*, *29*(1), 33–41. https://doi.org/10.1016/j.tree.2013.09.009

Lanfear, R., Thomas, J. A., Welch, J. J., Brey, T., & Bromham, L. (2007). Metabolic rate does not calibrate the molecular clock. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(39), 15388–15393. https://doi.org/10.1073/pnas.0703359104

Lau, C. H. E., & Robinson, O. (2021). DNA methylation age as a biomarker for cancer. *International Journal of Cancer*, *148*(11), 2652–2663. https://doi.org/10.1002/ijc.33451

Lercher, M. J., & Hurst, L. D. (2002). Human SNP variability and mutation rate are higher in regions of high recombination. *Trends in Genetics*, *18*(7), 337–340. https://doi.org/10.1016/S0168-9525(02)02669-0

Li, H. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. http://arxiv.org/abs/1303.3997

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, *475*(7357), 493–496. https://doi.org/10.1038/nature10231

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lin, Y., Darolti, I., Furman, B. L. S., Almeida, P., Sandkam, B. A., Breden, F., Wright, A. E., & Mank, J. E. (2022). Gene duplication to the Y chromosome in Trinidadian guppies. *Molecular Ecology*, *31*(6), 1853–1863. https://doi.org/10.1111/mec.16355

Lindsay, S. J., Rahbari, R., Kaplanis, J., Keane, T., & Hurles, M. E. (2019). Similarities and differences in patterns of germline mutation between mice and humans. *Nature Communications*, *10*(1), 4053. https://doi.org/10.1038/s41467-019-12023-w

Liu, Y., Li, B., Tan, R., Zhu, X., & Wang, Y. (2014). A gradient-boosting approach for filtering de novo mutations in parent-offspring trios. *Bioinformatics*, *30*(13), 1830–1836. https://doi.org/10.1093/bioinformatics/btu141

Long, H., Winter, D. J., Chang, A. Y. C., Sung, W., Wu, S. H., Balboa, M., Azevedo, R. B. R., Cartwright, R. A., Lynch, M., & Zufall, R. A. (2016). Low base-substitution mutation rate in the germline genome of the ciliate *Tetrahymena thermophila*. *Genome Biology and Evolution*, *8*(12), 3629–3639. https://doi.org/10.1093/gbe/evw223

Lynch, M. (2010). Evolution of the mutation rate. *Trends in Genetics*, *26*(8), 345–352. https://doi.org/10.1016/j.tig.2010.05.003

Lynch, M. (2020). The evolutionary scaling of cellular traits imposed by the drift barrier. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(19), 10435–10444. https://doi.org/10.1073/pnas.2000446117

Lynch, M., Ackerman, M. S., Gout, J. F., Long, H., Sung, W., Thomas, W. K., & Foster, P. L. (2016). Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics*, *17*(11), 704–714. https://doi.org/10.1038/nrg.2016.104

Malinsky, M., Svardal, H., Tyers, A. M., Miska, E. A., Genner, M. J., Turner, G. F., & Durbin, R. (2018). Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. *Nature Ecology and Evolution*, *2*(12), 1940–1955. https://doi.org/10.1038/s41559-018-0717-x

Martin, A. P., & Palumbi, S. R. (1993). Body size, metabolic rate, generation time, and the molecular clock. *Proceedings of the National Academy of Sciences of the United States of America*, *90*(9), 4087–4091. https://doi.org/10.1073/pnas.90.9.4087

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Monroe, J. G., Srikant, T., Carbonell-Bejerano, P., Becker, C., Lensink, M., Exposito-Alonso, M., Klein, M., Hildebrandt, J., Neumann, M., Kliebenstein, D., Weng, M., Imbert, E., Agren, J., Rutter, M. T., Fenster, C. B., & Weigel, D. (2022). Mutation bias reflects natural selection in Arabidopsis thaliana. *Nature*, *602*(7895), 101–105. https://doi.org/10.1038/s41586-021-04269-6

Nishioka, M., Kazuno, A. A., Nakamura, T., Sakai, N., Hayama, T., Fujii, K., Matsuo, K., Komori, A., Ishiwata, M., Watanabe, Y., Oka, T., Matoba, N., Kataoka, M., Alkanaq, A. N., Hamanaka, K., Tsuboi, T., Sengoku, T., Ogata, K., Iwata, N., … Takata, A. (2021). Systematic analysis of exonic germline and postzygotic de novo mutations in bipolar disorder. *Nature Communications*, *12*(1), 3750. https://doi.org/10.1038/s41467-021-23453-w

Pfeifer, S. P. (2017). Direct estimate of the spontaneous germ line mutation rate in African green monkeys. *Evolution*, *71*(12), 2858–2870. https://doi.org/10.1111/evo.13383

Phillips, K. P., Cable, J., Mohammed, R. S., Herdegen-Radwan, M., Raubic, J., Przesmycka, K. J., Van Oosterhout, C., & Radwan, J. (2018). Immunogenetic novelty confers a selective advantage in host–pathogen coevolution. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(7), 1552–1557. https://doi.org/10.1073/pnas.1708597115

Pockrandt, C., Alzamel, M., Iliopoulos, C. S., & Reinert, K. (2020). GenMap: Ultra-fast computation of genome mappability. *Bioinformatics*, *36*(12), 3687–3692. https://doi.org/10.1093/bioinformatics/btaa222

Qiu, F., Kitchen, A., Burleigh, J. G., & Miyamoto, M. M. (2014). Scombroid fishes provide novel insights into the trait/rate associations of molecular evolution. *Journal of Molecular Evolution*, *78*(6), 338–348. https://doi.org/10.1007/s00239-014-9621-4

Qiu, S., Yong, L., Wilson, A., Croft, D. P., Graham, C., & Charlesworth, D. (2022). Partial sex linkage and linkage disequilibrium on the guppy sex chromosome. *Molecular Ecology*, *31*(21), 5524–5537. https://doi.org/10.1111/mec.16674

Reznick, D. N., Shaw, F. H., Rodd, F. H., & Shaw, R. G. (1997). Evaluation of the rate of evolution in natural populations of guppies (*Poecilia reticulata*). *Science*, *275*(5308), 1934–1937. https://doi.org/10.1126/science.275.5308.1934

Reznick, D. N., & Travis, J. (2019). Experimental studies of evolution and eco-evo dynamics in guppies (*Poecilia reticulata*). *Annual Review of Ecology, Evolution, and Systematics*, *50*, 335–354. https://doi.org/10.1146/annurev-ecolsys-110218-024926

Roach, J. C., Glusman, G., Smit, A. F. A., Huff, C. D., Hubley, R., Shannon, P. T., Rowen, L., Pant, K. P., Goodman, N., Bamshad, M., Shendure, J., Drmanac, R., Jorde, L. B., Hood, L., & Galas, D. J. (2010). Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*, *328*(5978), 636–639. https://doi.org/10.1126/science.1186802

Saclier, N., Chardon, P., Malard, F., Konecny-Dupré, L., Eme, D., Bellec, A., Breton, V., Duret, L., Lefebure, T., & Douady, C. J. (2020). Bedrock radioactivity influences the rate and spectrum of mutation. *eLife*, *9*, 1–20. https://doi.org/10.7554/eLife.56830

Sasani, T. A., Pedersen, B. S., Gao, Z., Baird, L., Przeworski, M., Jorde, L. B., & Quinlan, A. R. (2019). Large, three-generation human families reveal post-zygotic mosaicism and variability in germline mutation accumulation. *eLife*, *8*, 1–24. https://doi.org/10.7554/eLife.46922

Sharp, N. P., Sandell, L., James, C. G., & Otto, S. P. (2018). The genome-wide rate and spectrum of spontaneous mutations differ between haploid and diploid yeast. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(22), E5046–E5055. https://doi.org/10.1073/pnas.1801040115

Smeds, L., Qvarnström, A., & Ellegren, H. (2016). Direct estimate of the rate of germline mutation in a bird. *Genome Research*, *26*(9), 1211–1218. https://doi.org/10.1101/gr.204669.116

Sung, W., Ackerman, M. S., Miller, S. F., Doak, T. G., & Lynch, M. (2012). Drift-barrier hypothesis and mutation-rate evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(45), 18488–18492. https://doi.org/10.1073/pnas.1216223109

Thomas, G. W., Wang, R. J., Puri, A., Harris, R. A., Raveendran, M., Hughes, D. S., Murali, S. C., Williams, L. E., Doddapaneni, H., Muzny, D. M., Gibbs, R. A., Abee, C. R., Galinski, M. R., Worley, K. C., Rogers, J., Radivojac, P., & Hahn, M. W. (2018). Reproductive longevity predicts mutation rates in primates. *Current Biology*, *28*(19), 3193–3197. https://doi.org/10.1016/j.cub.2018.08.050

Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative genomics viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, *14*(2), 178–192. https://doi.org/10.1093/bib/bbs017

Venn, O., Turner, I., Mathieson, I., De Groot, N., Bontrop, R., & McVean, G. (2014). Strong male bias drives germline mutation in chimpanzees. *Science*, *344*(6189), 1272–1275. https://doi.org/10.1126/science.344.6189.1272

Wang, L., Ji, Y., Hu, Y., Hu, H., Jia, X., Jiang, M., Zhang, X., Zhao, L., Zhang, Y., Jia, Y., Qin, C., Yu, L., Huang, J., Yang, S., Hurst, L. D., & Tian, D. (2019). The architecture of intra-organism mutation rate variation in plants. *PLoS Biology*, *17*(4), 1–29. https://doi.org/10.1371/journal.pbio.3000191

Wang, R. J., Raveendran, M., Harris, R. A., Murphy, W. J., Lyons, L. A., Rogers, J., & Hahn, M. W. (2021). De novo mutations in domestic cat are consistent with an effect of reproductive longevity on both the rate and spectrum of mutations. *BioRxiv*. https://www.biorxiv.org/content/10.1101/2021.04.06.438608v1

Wang, Y., & Obbard, D. J. (2023). *Experimental estimates of germline mutation rate in eukaryotes: A phylogenetic meta-analysis*. *BioRxiv* https://doi.org/10.1101/2023.01.24.525323

Whiting, J. R., Paris, J. R., van der Zee, M. J., Parsons, P. J., Weigel, D., & Fraser, B. A. (2021). Drainage-structuring of ancestral variation and a common functional pathway shape limited genomic convergence in natural high- and low-predation guppies. *PLoS Genetics*, *17*(5), 1–29. https://doi.org/10.1371/journal.pgen.1009566

Wright, A. E., Darolti, I., Bloch, N. I., Oostra, V., Sandkam, B., Buechel, S. D., Kolm, N., Breden, F., Vicoso, B., & Mank, J. E. (2017). Convergent recombination suppression suggests role of sexual selection in guppy sex chromosome formation. *Nature Communications*, *8*, 14251. https://doi.org/10.1038/ncomms14251

Yao, N. Y., & O'Donnell, M. E. (2016). Evolution of replication machines. *Critical Reviews in Biochemistry and Molecular Biology*, *51*(3), 135–149. https://doi.org/10.3109/10409238.2015.1125845

Yoder, A. D., & Tiley, G. P. (2021). The challenge and promise of estimating the de novo mutation rate from whole-genome comparisons among closely related individuals. *Molecular Ecology*, *30*(23), 6087–6100. https://doi.org/10.1111/mec.16007

Zhou, J., Park, C. Y., Theesfeld, C. L., Wong, A. K., Yuan, Y., Scheckel, C., Fak, J. J., Funk, J., Yao, K., Tajima, Y., Packer, A., Darnell, R. B., & Troyanskaya, O. G. (2019). Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. *Nature Genetics*, *51*(6), 973–980. https://doi.org/10.1038/s41588-019-0420-0

Zhou, Y., He, F., Pu, W., Gu, X., Wang, J., & Su, Z. (2020). The impact of DNA methylation dynamics on the mutation rate during human germline development. *G3: Genes*, *Genomes*, *Genetics*, *10*(9), 3337–3346. https://doi.org/10.1534/g3.120.401511

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Burda, K., & Konczal, M. (2023). Validation of machine learning approach for direct mutation rate estimation. *Molecular Ecology Resources*, *23*, 1757–1771. https://doi.org/10.1111/1755-0998.13841

MOLECULAR ECOLOGY
RESOURCES

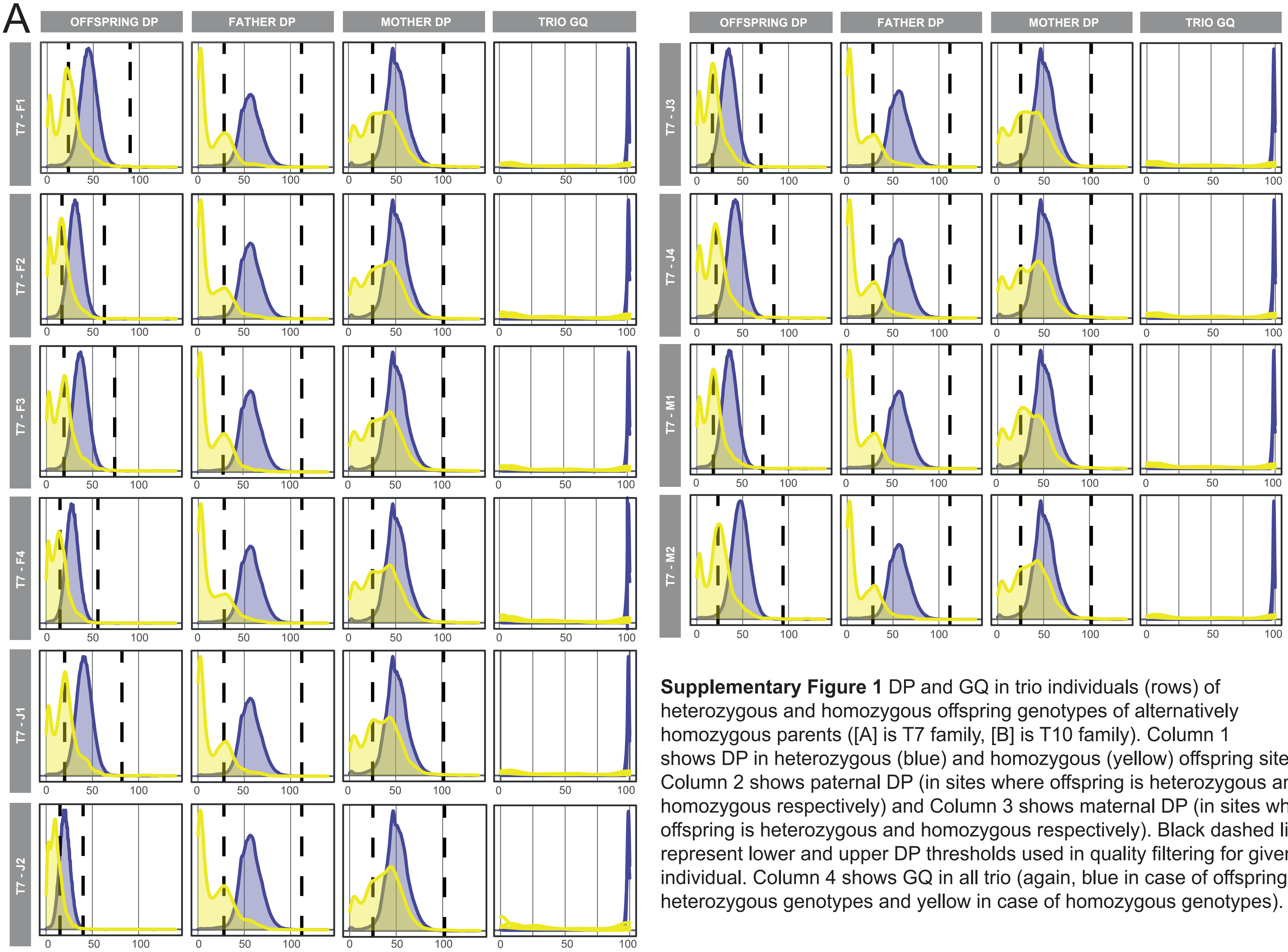
Supplemental Figures for:

“Validation of machine learning approach for direct mutation rate estimation”

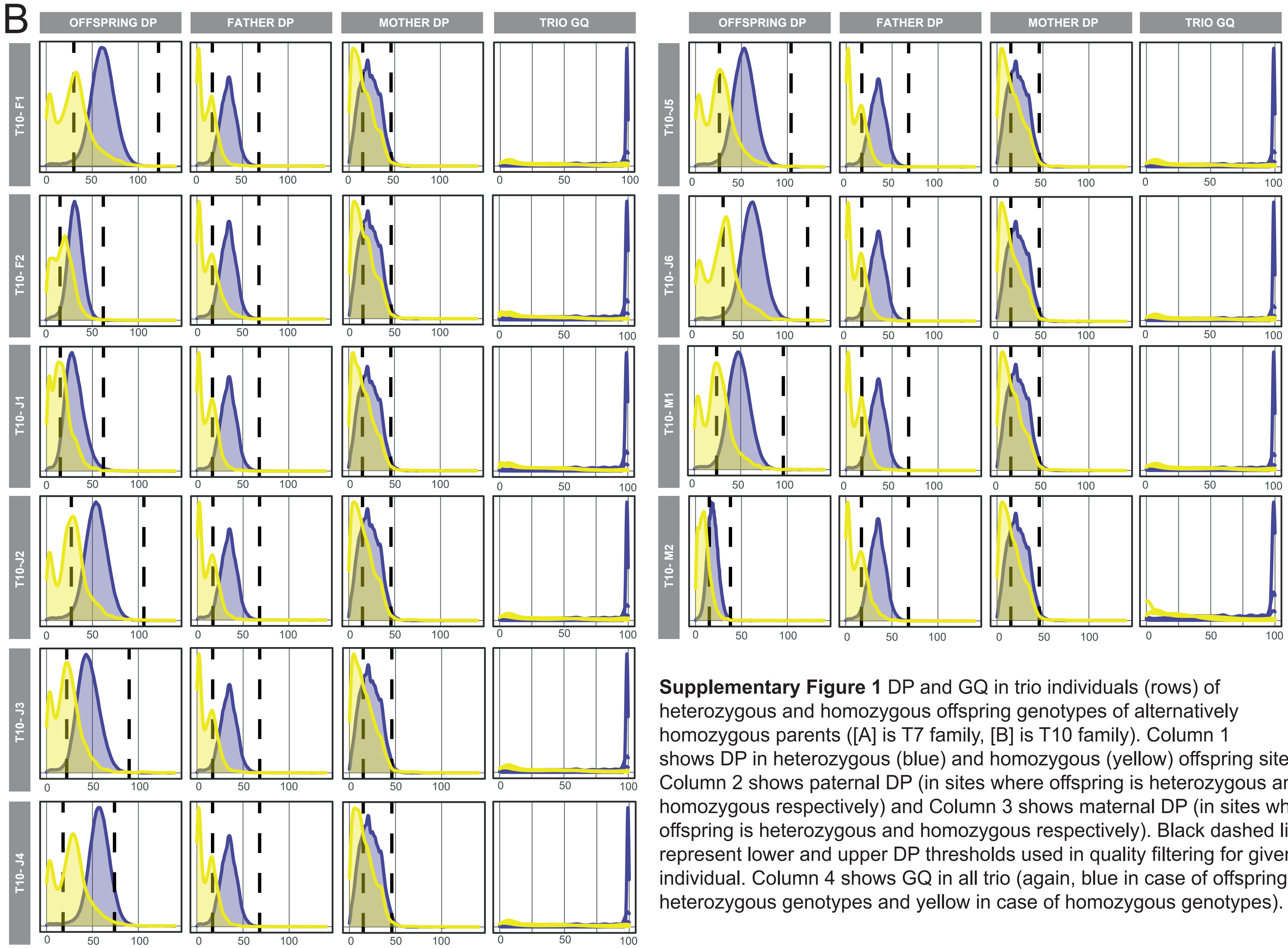
Katarzyna Burda[1], Mateusz Konczal[1]

[1] Evolutionary Biology Group, Faculty of Biology, Adam Mickiewicz University; 60-614 Poznań, Poland

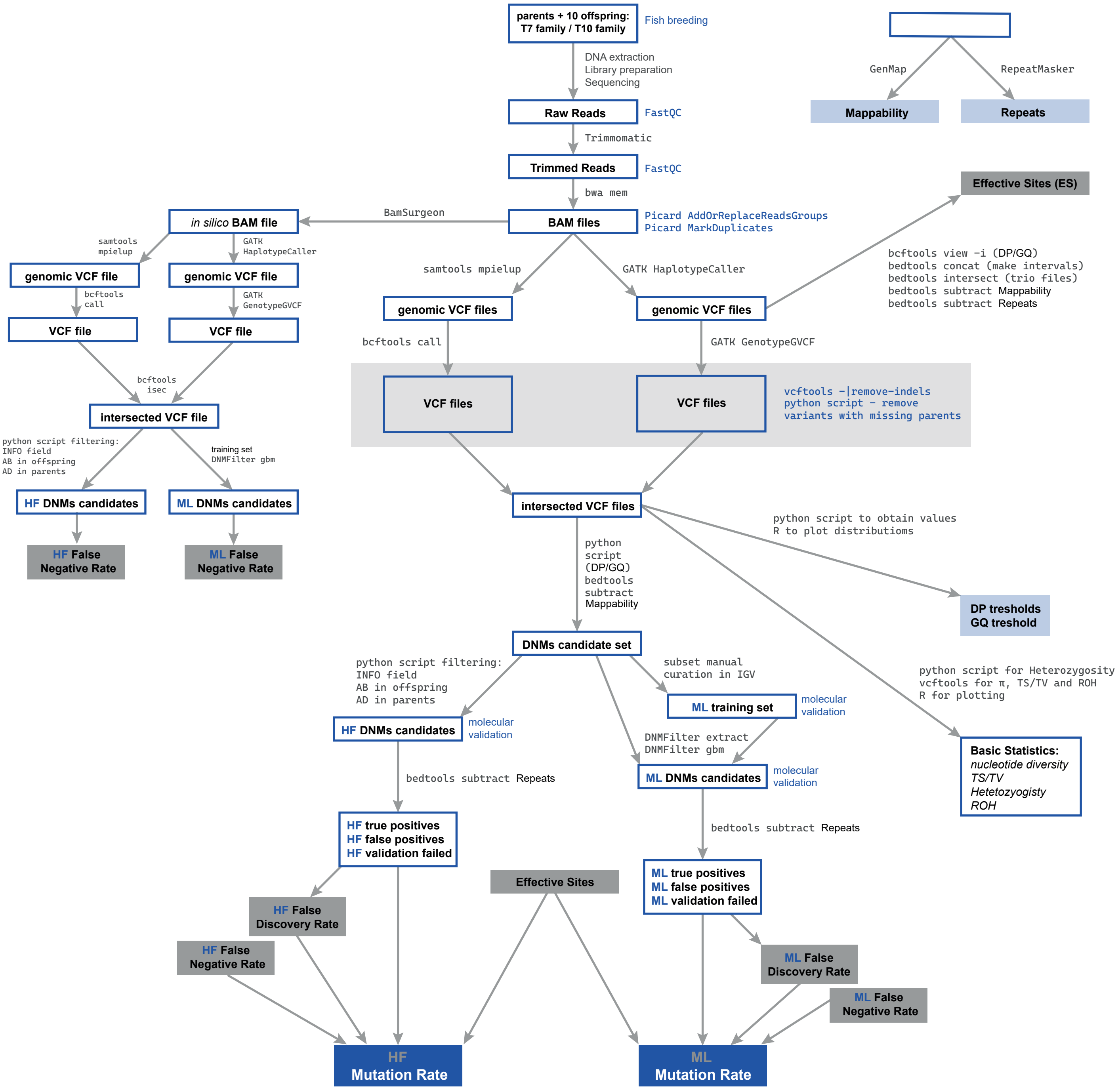
Table of Contents:



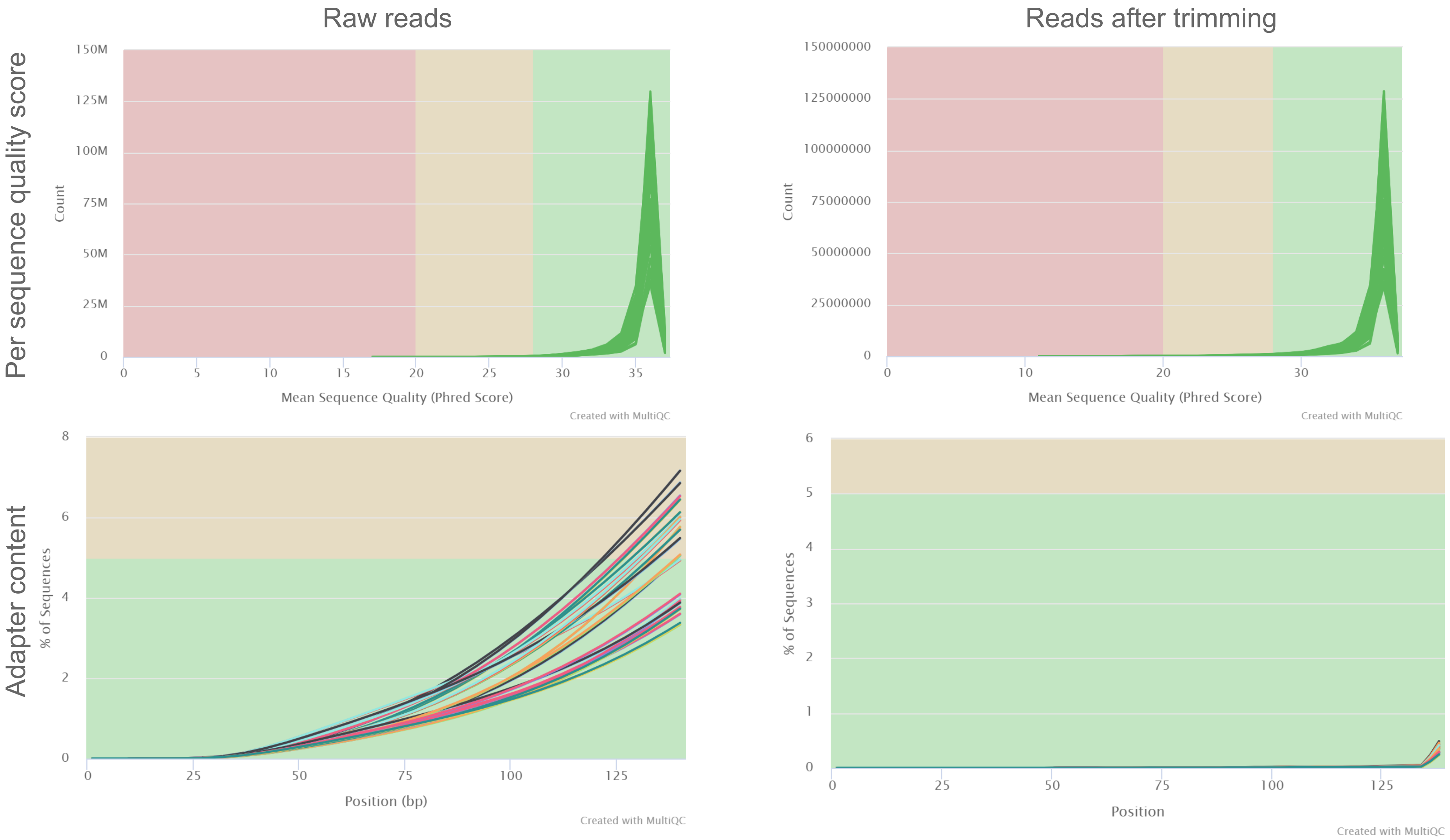
Supplementary Figure 1 DP and GQ in trio individuals (rows) of heterozygous and homozygous offspring genotypes of alternatively homozygous parents ([A] is T7 family, [B] is T10 family). Column 1 shows DP in heterozygous (blue) and homozygous (yellow) offspring sites, Column 2 shows paternal DP (in sites where offspring is heterozygous and homozygous respectively) and Column 3 shows maternal DP (in sites where offspring is heterozygous and homozygous respectively). Black dashed lines represent lower and upper DP thresholds used in quality filtering for given individual. Column 4 shows GQ in all trio (again, blue in case of offspring heterozygous genotypes and yellow in case of homozygous genotypes).



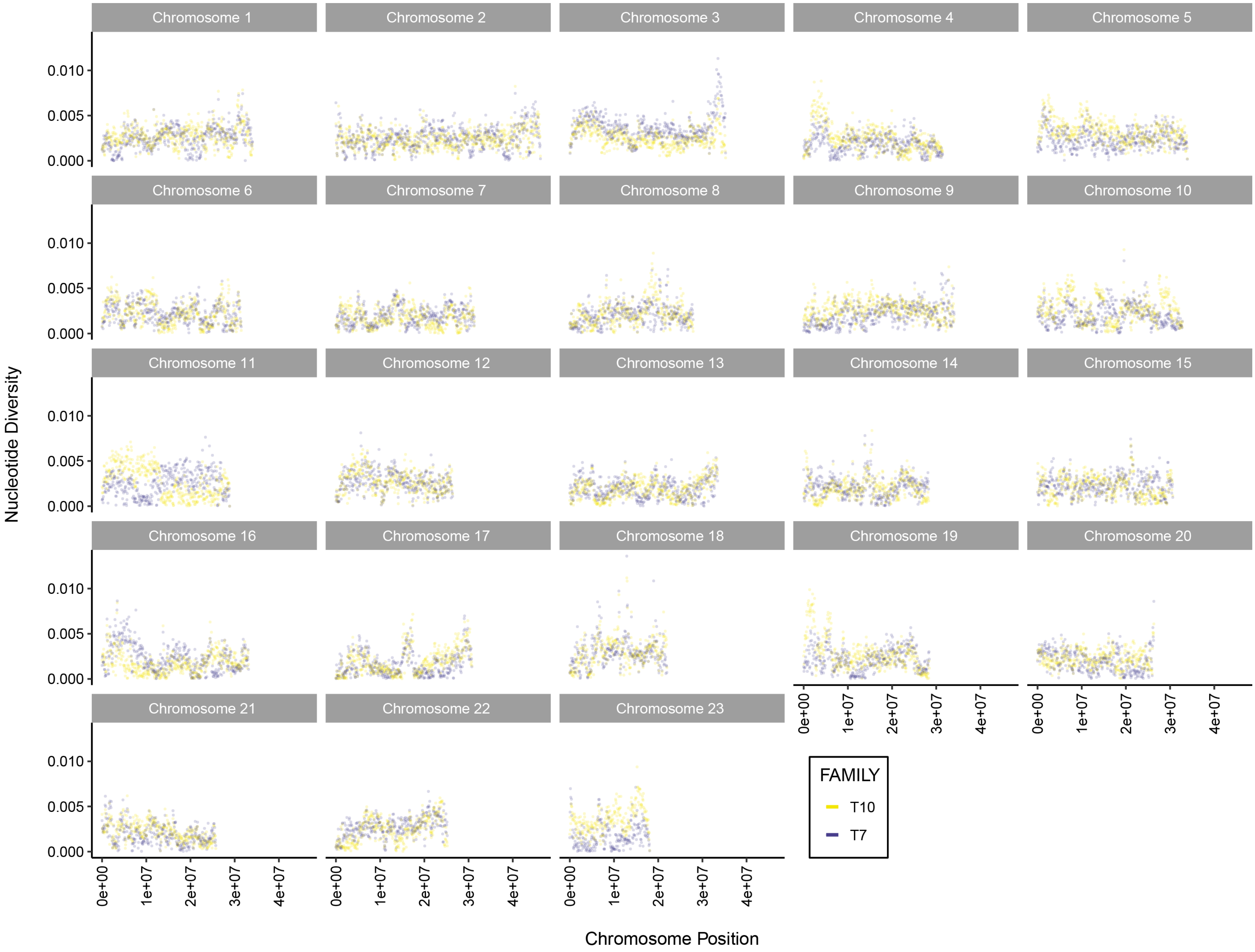
Supplementary Figure 1 DP and GQ in trio individuals (rows) of heterozygous and homozygous offspring genotypes of alternatively homozygous parents ([A] is T7 family, [B] is T10 family). Column 1 shows DP in heterozygous (blue) and homozygous (yellow) offspring sites, Column 2 shows paternal DP (in sites where offspring is heterozygous and homozygous respectively) and Column 3 shows maternal DP (in sites where offspring is heterozygous and homozygous respectively). Black dashed lines represent lower and upper DP thresholds used in quality filtering for given individual. Column 4 shows GQ in all trio (again, blue in case of offspring heterozygous genotypes and yellow in case of homozygous genotypes).



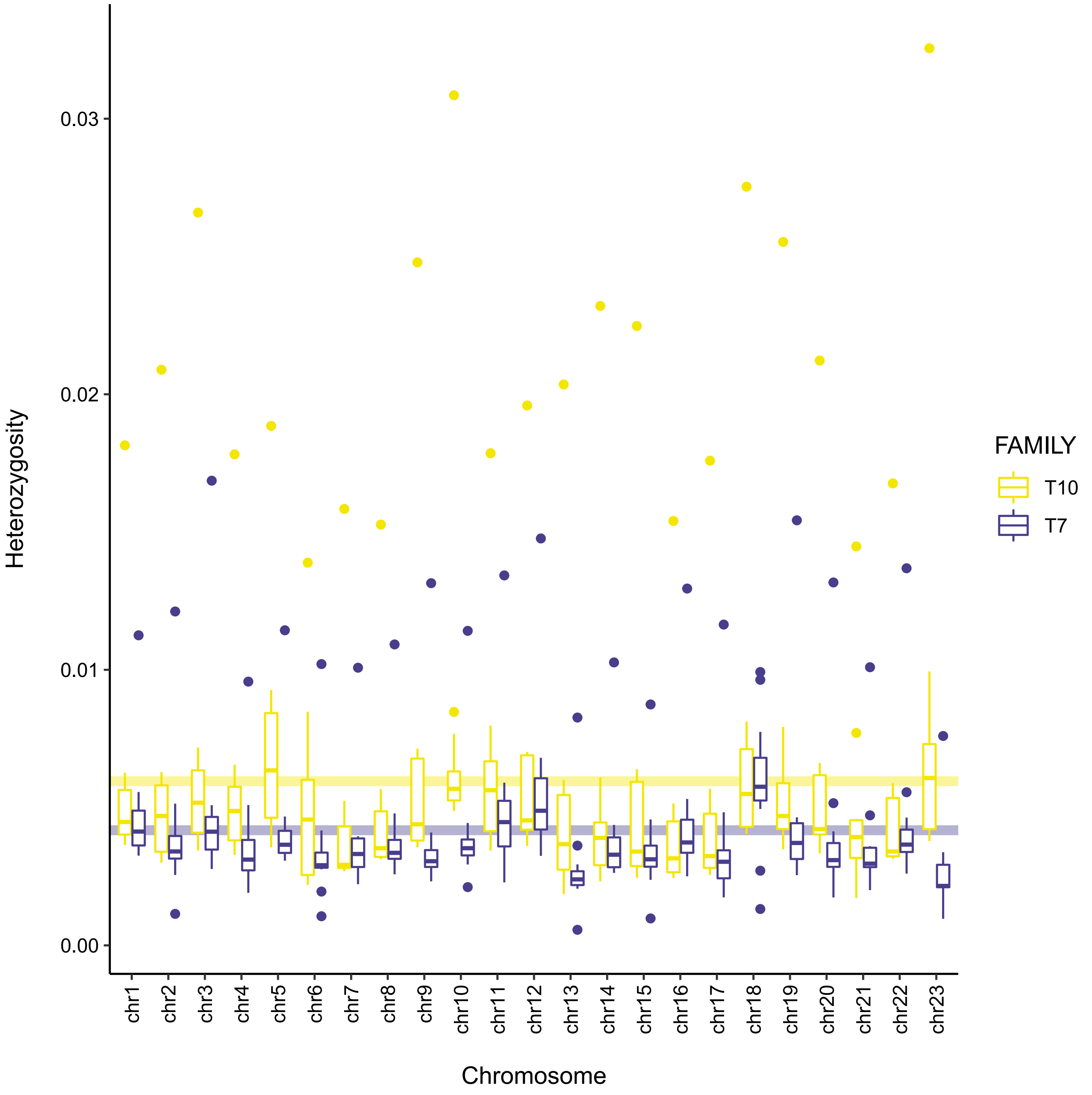
Supplementary Figure 2 Flow chart of all analyses performed in the study.



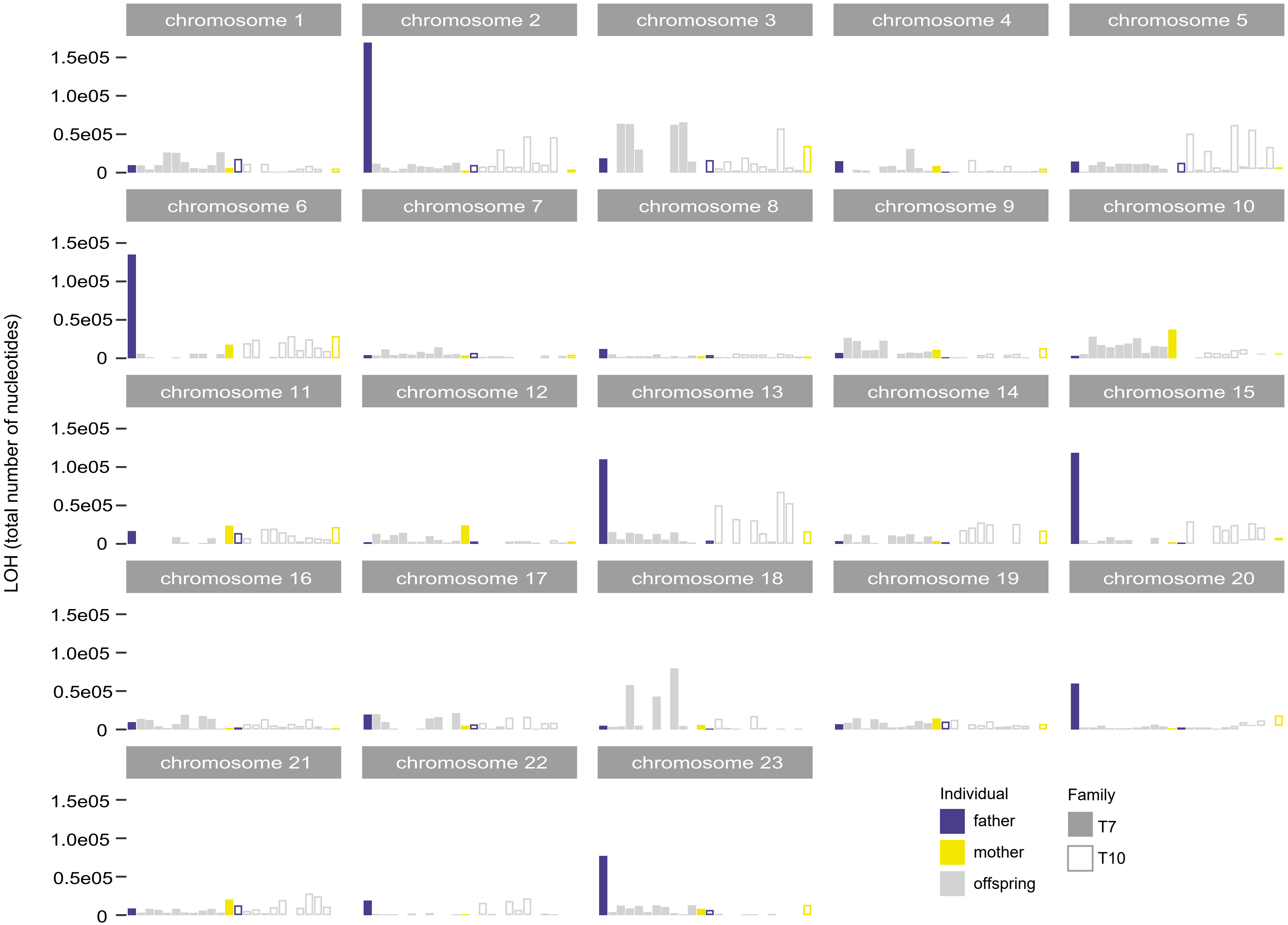
Supplementary Figure 3 Results of reads trimming. Plots show all samples. The first row represents results regarding per sequence quality scores before (left) and after (right) trimming. The second row shows illumina adapters content before (left) and after (right) trimming.



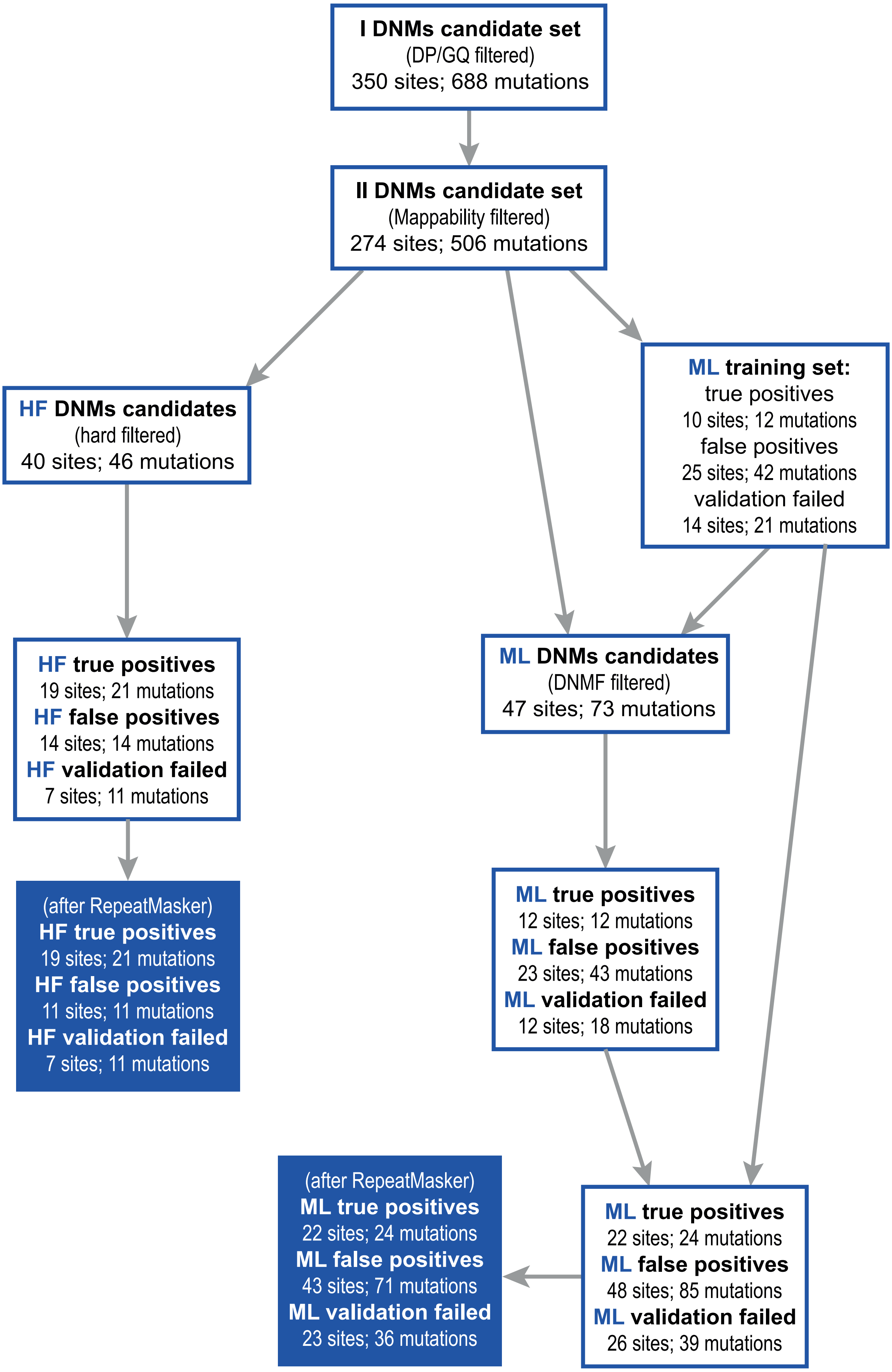
Supplementary Figure 4 Nucleotide diversity across genome calculated in 75kb windows. Yellow points represent T10 family and blue points represent T7 family.



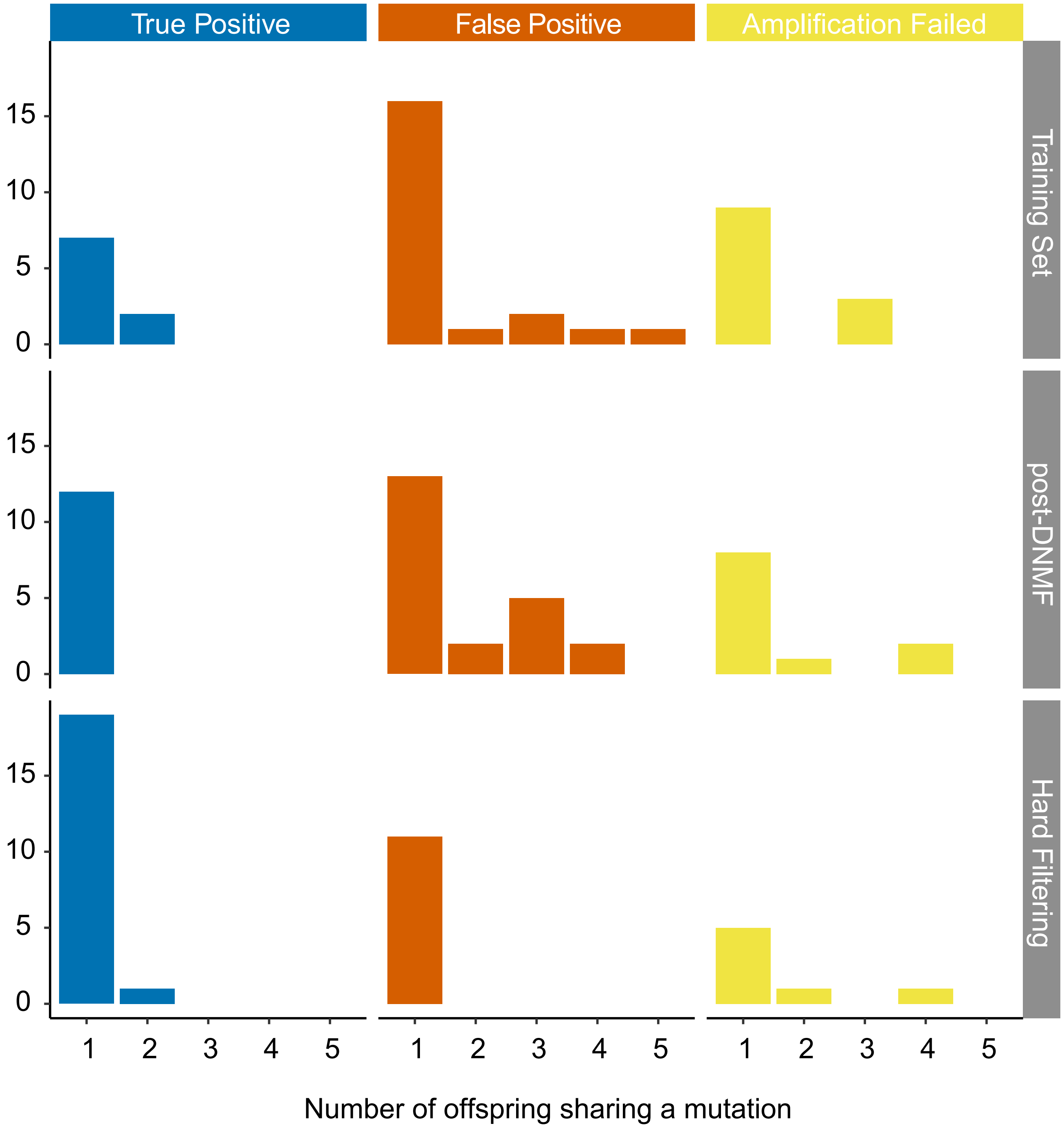
Supplementary Figure 5 Heterozygosity in chromosomes across genome. Boxplots show distribution of heterozygosity in T7 (blue) and T10 (yellow) family. Wide horizontal lines represent mean genomic heterozygosity in two families respectively.



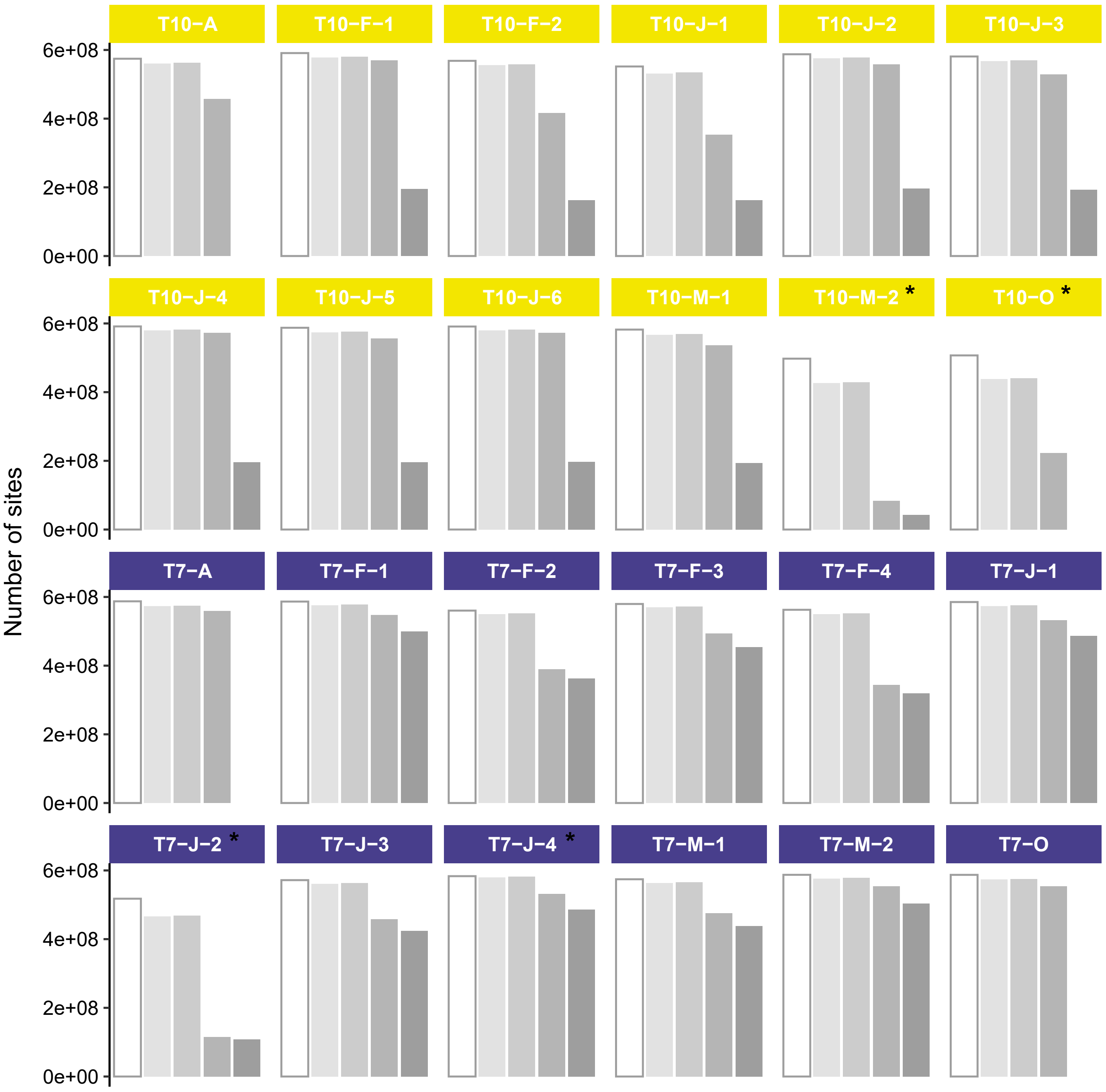
Supplementary Figure 6 Total number of sites constituting runs of homozygosity. Family T7 is shown with solid bars, while family T10 is represented by empty bars. Mothers and fathers of both families were marked with colours (yellow and blue respectively).



Supplementary Figure 7 Chart showing counts of mutations and sites filtered out during different steps of the workflow.



Supplementary Figure 8 Mutations shared between offspring. Rows represent three sets of DNMs, columns indicate DNMs status after molecular validation. X-axis shows number of offspring in which given mutation was found.



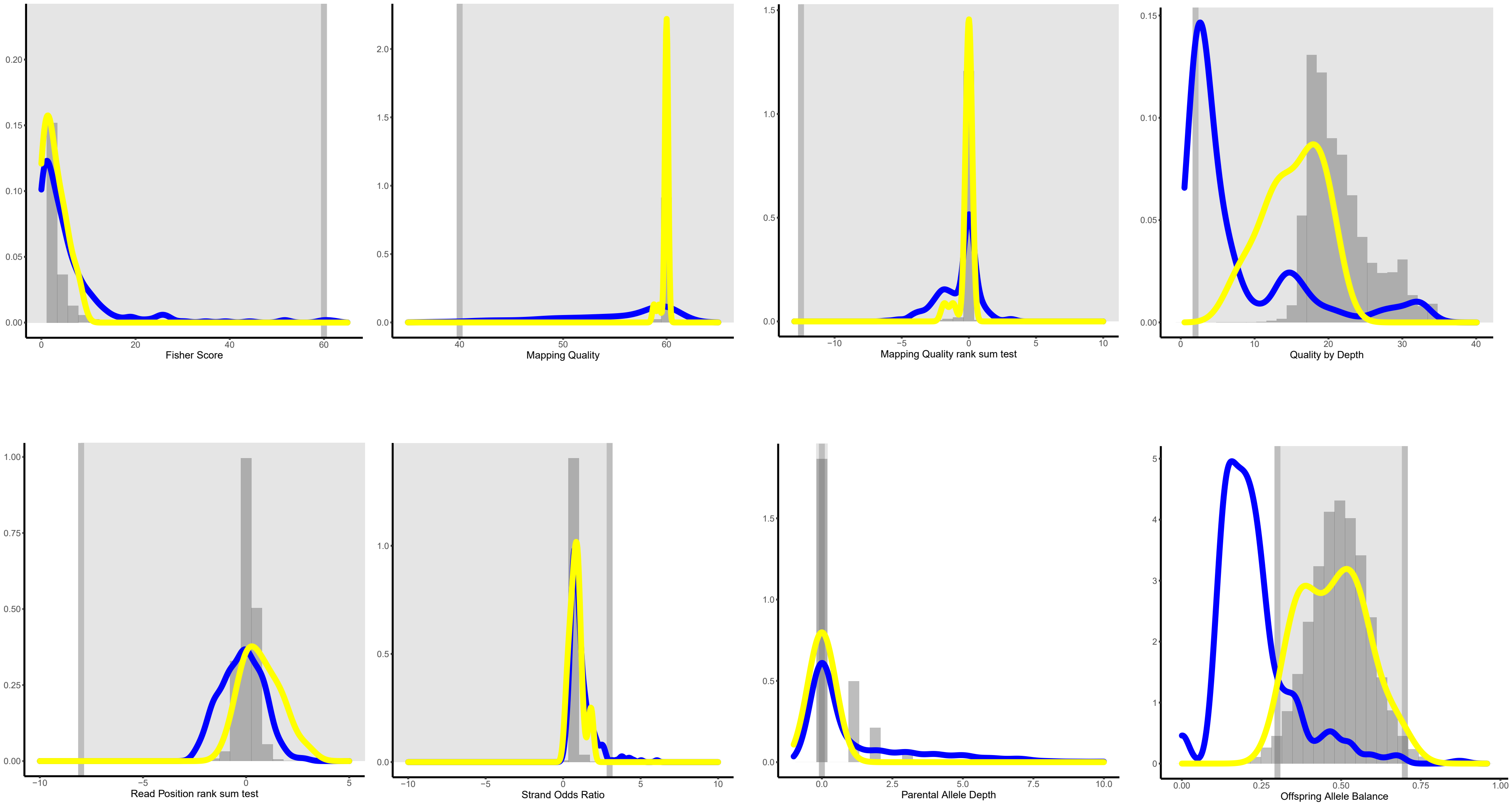
Supplementary Figure 9 Number of sites remaining after applying consecutive filters in individuals. Note, that every next filter shows sites retained by that filter and the ones applied before. First bar shows sites that have at least 30 GQ, second bar depicts sites which DP is at least 0.5x mean coverage of a given individual (or at least 15, when individual had 1/2xDP lower than that, such cases are marked with an asterisk). Third bar shows sites which remained after filtering with upper threshold of 2x mean coverage of a given individual. Fourth bar represents sites filtered by minimum GQ of 70. Finally, last bar shows sites that remained after intersection of three sets: one of the offspring, one of the mother and one of the father. Coloured labels distinguish between family T10 (yellow) and T7 (blue).

Individual Filter

□	minimum GQ ≥ 30
▨	minimum DP ≥ 1/2 x mean coverage
▨	maximum DP ≤ 2 x mean coverage
▨	minimum GQ ≥ 70

Trio Intersection

■	Effective Sites



Supplementary Figure 10 Distribution of INFO-field, AB and parental AD statistics for heterozygous genotypes of offspring whose parents were alternative homozygous (grey bars) compared with sets of all candidate de novo mutations (blue lines), and true de novo mutations (yellow lines). Shaded areas cover values of statistics that met hard filtering criteria.

[*Originally these tables are in an Excel file and are called* **Table S1** *in the article*]

Supplementary Table 1 Mutations successfully validated as true positives. The columns present as follows: Family - guppy family; Chromosome - reference scaffold name; Position - position on the chromosome; HF - whether mutation passed or failed the HF; ML - whether mutation passed or failed the ML, or underwent successful validation during the Training Set (TS) preparation; Ind - sample name; RepMas - whether mutation passed or failed the RepeatMasker filter; TS - whether mutation was a part of the Training Set in the ML.

Family	Chromosome	Position	HF	ML	Ind	RepMas
T10	NC_024331.1	11320387	PASS	PASS	T10J1	PASS
T7	NC_024332.1	7907610	PASS	TS	T7F1	PASS
T7	NC_024332.1	28098628	PASS	PASS	T7M2	PASS
T10	NC_024332.1	29464863	FAIL	TS	T10F2	PASS
T7	NC_024333.1	6123659	PASS	PASS	T7J3	PASS
T7	NC_024333.1	31364266	PASS	PASS	T7J1	PASS
T7	NC_024334.1	15397752	PASS	PASS	T7M2	PASS
T7	NC_024334.1	29856266	PASS	PASS	T7J4	PASS
T7	NC_024337.1	26525718	PASS	TS	T7J3	PASS
T7	NC_024337.1	26525718	PASS	TS	T7M2	PASS
T7	NC_024340.1	23542367	PASS	TS	T7F4	PASS
T7	NC_024340.1	25291101	PASS	TS	T7M2	PASS
T10	NC_024342.1	17229793	PASS	TS	T10M1	PASS
T7	NC_024342.1	19229557	PASS	TS	T7F2	PASS
T7	NC_024342.1	19229557	PASS	TS	T7F4	PASS
T7	NC_024344.1	21187135	PASS	PASS	T7F1	PASS
T7	NC_024344.1	23760915	PASS	PASS	T7F4	PASS
T7	NC_024345.1	18795681	PASS	TS	T7M1	PASS
T7	NC_024346.1	7767856	PASS	TS	T7J3	PASS
T7	NC_024347.1	15099629	PASS	PASS	T7F2	PASS
T7	NC_024348.1	17812487	FAIL	PASS	T7J3	PASS
T10	NC_024349.1	6887521	FAIL	PASS	T10J2	PASS
T7	NC_024351.1	3258922	PASS	TS	T7J3	PASS
T10	NC_024353.1	14844758	PASS	PASS	T10M1	PASS

Supplementary Table 2 Mutations successfully validated as false positives. The columns present as follows: Family - guppy family; Chromosome - reference scaffold name; Position - position on the chromosome; HF - whether mutation passed or failed the HF; ML - whether mutation passed or failed the ML, or underwent successful validation during the Training Set (TS) preparation; Ind - sample name; RepMas - whether mutation passed or failed the RepeatMasker filter; TS - whether mutation was a part of the Training Set in the ML.

Family	Chromosome	Position	HF	ML	Ind	RepMas
T7	NC_024331.1	15871169	FAIL	PASS	T7M2	PASS
T7	NC_024331.1	26842368	PASS	TS	T7M2	PASS
T10	NC_024332.1	11534675	FAIL	PASS	T10J6	PASS
T7	NC_024332.1	30767099	FAIL	TS	T7J3	PASS
T7	NC_024333.1	13956746	FAIL	TS	T7F3	PASS

T7	NC_024333.1	13956746	FAIL	TS	T7J3	PASS
T7	NC_024333.1	13956746	FAIL	TS	T7M2	PASS
T7	NC_024333.1	14338502	FAIL	PASS	T7F2	FAIL
T7	NC_024333.1	14338502	FAIL	PASS	T7F3	FAIL
T7	NC_024333.1	14338502	FAIL	PASS	T7J2	FAIL
T7	NC_024333.1	20900466	FAIL	TS	T7J4	PASS
T10	NC_024333.1	27724387	FAIL	PASS	T10F1	PASS
T10	NC_024333.1	27724387	FAIL	PASS	T10J3	PASS
T7	NC_024334.1	27180609	PASS	TS	T7F2	FAIL
T7	NC_024334.1	27180609	FAIL	TS	T7F1	FAIL
T7	NC_024334.1	27180609	FAIL	TS	T7F3	FAIL
T7	NC_024334.1	27180609	FAIL	TS	T7F4	FAIL
T7	NC_024335.1	235073	FAIL	TS	T7F1	PASS
T7	NC_024335.1	235073	FAIL	TS	T7F4	PASS
T7	NC_024335.1	235073	FAIL	TS	T7J4	PASS
T7	NC_024335.1	235073	FAIL	TS	T7M2	PASS
T10	NC_024335.1	5589499	PASS	PASS	T10M1	PASS
T10	NC_024335.1	5589499	FAIL	PASS	T10F2	PASS
T10	NC_024335.1	5589499	FAIL	PASS	T10J1	PASS
T10	NC_024335.1	5589499	FAIL	PASS	T10J3	PASS
T7	NC_024335.1	9961299	FAIL	TS	T7J2	PASS
T7	NC_024335.1	9961299	FAIL	TS	T7J3	PASS
T7	NC_024335.1	9961299	FAIL	TS	T7M2	PASS
T7	NC_024335.1	17244018	PASS	TS	T7F3	PASS
T7	NC_024335.1	18390381	PASS	TS	T7F1	PASS
T7	NC_024336.1	29144117	PASS	PASS	T7F4	PASS
T7	NC_024338.1	5001133	FAIL	TS	T7F1	PASS
T7	NC_024338.1	5001133	FAIL	TS	T7F2	PASS
T7	NC_024338.1	5001133	FAIL	TS	T7J2	PASS
T7	NC_024338.1	5001133	FAIL	TS	T7J4	PASS
T7	NC_024338.1	5001133	FAIL	TS	T7M1	PASS
T7	NC_024338.1	6245711	FAIL	PASS	T7F1	PASS
T10	NC_024338.1	6859654	PASS	PASS	T10M1	PASS
T7	NC_024338.1	19102477	PASS	FAIL	T7J2	PASS
T7	NC_024338.1	25101804	FAIL	PASS	T7J2	PASS
T7	NC_024338.1	25101804	FAIL	PASS	T7J4	PASS
T7	NC_024338.1	25101804	FAIL	PASS	T7M1	PASS
T7	NC_024339.1	29165724	FAIL	TS	T7F4	PASS
T10	NC_024341.1	7048843	FAIL	PASS	T10J2	PASS
T10	NC_024341.1	7048843	FAIL	PASS	T10J5	PASS
T10	NC_024341.1	7048843	FAIL	PASS	T10M1	PASS
T7	NC_024341.1	24373156	PASS	PASS	T7F1	PASS
T10	NC_024343.1	33113384	FAIL	TS	T10F2	PASS
T7	NC_024344.1	4812286	FAIL	TS	T7M1	PASS
T10	NC_024344.1	15970244	PASS	TS	T10J5	PASS

T7	NC_024344.1	22681080	PASS	TS	T7J3	FAIL
T7	NC_024344.1	26200793	FAIL	TS	T7M1	PASS
T7	NC_024345.1	18988704	FAIL	TS	T7F2	FAIL
T7	NC_024345.1	18988704	FAIL	TS	T7F3	FAIL
T10	NC_024345.1	25544522	FAIL	PASS	T10F2	PASS
T10	NC_024345.1	25544522	FAIL	PASS	T10J2	PASS
T10	NC_024345.1	25544522	FAIL	PASS	T10J3	PASS
T7	NC_024346.1	7734623	FAIL	TS	T7F4	PASS
T10	NC_024346.1	14712332	FAIL	PASS	T10M2	PASS
T10	NC_024346.1	15462695	FAIL	TS	T10J1	PASS
T10	NC_024346.1	15462697	FAIL	TS	T10J1	PASS
T7	NC_024346.1	20905128	FAIL	TS	T7F2	FAIL
T10	NC_024346.1	25463300	FAIL	PASS	T10M2	PASS
T7	NC_024347.1	10047049	FAIL	PASS	T7M2	PASS
T7	NC_024349.1	9026031	FAIL	PASS	T7F3	PASS
T7	NC_024350.1	5017566	PASS	FAIL	T7F2	PASS
T7	NC_024350.1	10818599	FAIL	PASS	T7F4	PASS
T10	NC_024351.1	5422	FAIL	PASS	T10J1	PASS
T10	NC_024351.1	5422	FAIL	PASS	T10M2	PASS
T7	NC_024351.1	21908245	FAIL	PASS	T7J1	PASS
T7	NC_024352.1	5696127	PASS	FAIL	T7F2	PASS
T10	NC_024353.1	2988043	FAIL	PASS	T10J2	PASS
T10	NC_024353.1	2988043	FAIL	PASS	T10J3	PASS
T10	NC_024353.1	2988043	FAIL	PASS	T10J6	PASS
T10	NC_024353.1	2988043	FAIL	PASS	T10M1	PASS
T10	NC_024353.1	2988046	FAIL	PASS	T10J2	PASS
T10	NC_024353.1	2988046	FAIL	PASS	T10J3	PASS
T10	NC_024353.1	2988046	FAIL	PASS	T10J6	PASS
T10	NC_024353.1	2988046	FAIL	PASS	T10M1	PASS
T10	NC_024353.1	6244525	FAIL	TS	T10M2	PASS
T10	NW_007615014.1	427669	FAIL	PASS	T10F1	PASS
T10	NW_007615014.1	427669	FAIL	PASS	T10F2	PASS
T10	NW_007615014.1	427669	FAIL	PASS	T10M1	PASS
T7	NW_007615025.1	165358	PASS	TS	T7F2	FAIL
T7	NW_007615025.1	165358	FAIL	TS	T7F4	FAIL
T7	NW_007615025.1	165358	FAIL	TS	T7M1	FAIL
T7	NW_007615586.1	374	FAIL	TS	T7M1	PASS
T10	NW_007616322.1	1355	FAIL	TS	T10M2	PASS

Supplementary Table 3 Mutations which failed the validation. The columns present as follows: Family - guppy family; Chromosome - reference scaffold name; Position - position on the chromosome; HF - whether mutation passed or failed the HF; ML - whether mutation passed or failed the ML, or underwent successful validation during the Training Set (TS) preparation; Ind - sample name; RepMas - whether mutation passed or failed the RepeatMasker filter; TS - whether mutation was a part of the Training Set in the ML.

Family	Chromosome	Position	HF	ML	Ind	RepMas
T7	NC_024331.1	637438	FAIL	PASS	T7F1	PASS
T7	NC_024331.1	637438	FAIL	PASS	T7F2	PASS
T7	NC_024331.1	637438	FAIL	PASS	T7F3	PASS
T7	NC_024331.1	637438	FAIL	PASS	T7J4	PASS
T7	NC_024332.1	45913267	FAIL	PASS	T7F1	PASS
T7	NC_024333.1	34293864	PASS	PASS	T7F3	PASS
T7	NC_024334.1	2809847	FAIL	PASS	T7J4	PASS
T7	NC_024334.1	7149849	FAIL	TS	T7J2	PASS
T7	NC_024334.1	13030601	FAIL	TS	T7J3	FAIL
T10	NC_024336.1	4345339	PASS	PASS	T10F2	PASS
T10	NC_024336.1	4345339	PASS	PASS	T10J1	PASS
T10	NC_024336.1	4345339	PASS	PASS	T10J6	PASS
T10	NC_024336.1	4345339	PASS	PASS	T10M1	PASS
T7	NC_024336.1	9745005	PASS	FAIL	T7F2	PASS
T7	NC_024336.1	9745005	PASS	FAIL	T7J4	PASS
T10	NC_024337.1	12381392	FAIL	TS	T10J1	PASS
T7	NC_024337.1	28866355	FAIL	TS	T7F2	PASS
T7	NC_024337.1	28866355	FAIL	TS	T7J3	PASS
T7	NC_024337.1	28866355	FAIL	TS	T7J4	PASS
T7	NC_024339.1	4601179	PASS	FAIL	T7J3	PASS
T7	NC_024339.1	15531292	FAIL	PASS	T7F2	PASS
T7	NC_024339.1	32082304	FAIL	TS	T7J2	PASS
T10	NC_024339.1	32458835	FAIL	TS	T10M2	PASS
T7	NC_024340.1	12246622	PASS	PASS	T7J2	PASS
T7	NC_024342.1	14292780	FAIL	TS	T7J1	PASS
T7	NC_024342.1	14292780	FAIL	TS	T7M1	PASS
T7	NC_024342.1	14292780	FAIL	TS	T7M2	PASS
T7	NC_024343.1	30138991	FAIL	PASS	T7J3	PASS
T7	NC_024344.1	13936563	FAIL	TS	T7F2	FAIL
T7	NC_024344.1	13936563	FAIL	TS	T7F4	FAIL
T7	NC_024347.1	2489207	FAIL	PASS	T7J4	PASS
T7	NC_024348.1	699670	FAIL	TS	T7J1	PASS
T7	NC_024348.1	699670	FAIL	TS	T7J2	PASS
T7	NC_024348.1	699670	FAIL	TS	T7M1	PASS
T7	NC_024349.1	11943373	PASS	FAIL	T7J1	PASS
T7	NC_024349.1	26770037	FAIL	TS	T7F4	PASS
T7	NC_024350.1	17800963	FAIL	TS	T7F2	PASS
T7	NC_024351.1	22632310	PASS	PASS	T7M1	PASS
T7	NC_024351.1	22764111	FAIL	TS	T7J3	PASS
T7	NC_024353.1	15100477	FAIL	TS	T7J3	PASS
T10	NW_007615489.1	6837	FAIL	PASS	T10J1	PASS
T10	NW_007615489.1	6837	FAIL	PASS	T10J3	PASS
T10	NW_007615489.1	6837	FAIL	TS	T10F1	PASS

Supplementary Table 4 Summary of mutation rate calculations for each offspring, family and method. This table shows calculations with mutations located in repetitive regions. FNR – false negative rate; FDR – false discovery rate.

Individual	*True Positve (ML)*	*True Positive (Hard Filtering)*	*False Positive (ML)*	*False Positive (Hard Filtering)*	*Failed Amplification (ML)*	*Failed Amplification (Hard Filtering)*	*Effective Sites*	*FNR (ML)*	*FNR (Hard Filtering)*	*FDR (ML)*	*FDR (Hard Filtering)*	*Mutation rate (ML)*	*Mutation rate (Hard Filtering)*
T7-F1	2	2	5	2	2	0	503381702	0,26	0,16	0,75	0,34	2,962E-09	2,36496E-09
T7-F2	2	2	1	2	4	1	363389505	0,26	0,16	0,75	0,34	4,914E-09	4,35098E-09
T7-F3	0	0	3	1	2	1	456280408	0,26	0,16	0,75	0,34	6,523E-10	8,56107E-10
T7-F4	3	3	5	1	1	0	318987121	0,26	0,16	0,75	0,34	6,065E-09	5,59808E-09
T7-J1	1	1	1	0	2	1	490378445	0,26	0,16	0,75	0,34	1,821E-09	2,01041E-09
T7-J2	0	0	3	1	4	1	104432860	0,26	0,16	0,75	0,34	5,7E-09	3,74044E-09
T7-J3	5	**4**	3	0	4	1	425748453	0,26	0,16	0,75	0,34	8,389E-09	6,50989E-09
T7-J4	1	1	4	0	4	1	489906446	0,26	0,16	0,75	0,34	2,43E-09	2,01235E-09
T7-M1	1	1	5	0	3	1	440869701	0,26	0,16	0,75	0,34	2,363E-09	2,23618E-09
T7-M2	4	4	6	1	1	0	508587561	0,26	0,16	0,75	0,34	4,974E-09	4,6815E-09
T7	**19**	**18**	**36**	**8**	**27**	**7**	**4101962202**	0,26	0,16	**0,65**	**0,31**	**4,111E-09**	**3,31522E-09**
T10-F1	0	0	2	0	1	0	194334203	0,26	0,16	0,75	0,34	7,657E-10	0
T10-F2	1	**0**	4	0	1	1	160771362	0,26	0,16	0,75	0,34	4,628E-09	2,42969E-09
T10-J1	1	1	4	0	3	1	158419463	0,26	0,16	0,75	0,34	6,575E-09	6,22312E-09
T10-J2	1	**0**	4	0	0	0	194646064	0,26	0,16	0,75	0,34	3,058E-09	0
T10-J3	0	0	5	0	1	0	191421509	0,26	0,16	0,75	0,34	7,774E-10	0
T10-J4	0	0	0	0	0	0	171710649	0,26	0,16	0,75	0,34	0	0
T10-J5	0	0	2	1	0	0	194565417	0,26	0,16	0,75	0,34	0	0
T10-J6	0	0	3	0	1	1	195751804	0,26	0,16	0,75	0,34	7,602E-10	1,99551E-09
T10-M1	2	2	6	2	1	1	191888569	0,26	0,16	0,75	0,34	6,979E-09	8,23968E-09
T10-M2	0	0	5	0	1	0	40145838	0,26	0,16	0,75	0,34	3,707E-09	0
T10	**5**	**3**	**35**	**3**	**9**	**4**	**1693654878**	0,26	0,16	**0,88**	**0,50**	**2,153E-09**	**1,75726E-09**
Total:	**24**	**21**	**71**	**11**	**36**	**11**	**5795617080**	**0,26**	**0,16**	**0,75**	**0,34**	**3,389E-09**	**2,8982E-09**

Supplementary Table 5 Summary of mutation rate calculations for each offspring, family and method. This table shows calculations without mutations located in repetitive regions. FNR – false negative rate; FDR – false discovery rate.

Individual	*True Positve (ML)*	*True Positive (Hard Filtering)*	*False Positive (ML)*	*False Positive (Hard Filtering)*	*Failed Amplification (ML)*	*Failed Amplification (Hard Filtering)*	*Effective Sites*	*FNR (ML)*	*FNR (Hard Filtering)*	*FDR (ML)*	*FDR (Hard Filtering)*	*Mutation rate (ML)*	*Mutation rate (Hard Filtering)*
T7-F1	2	2	6	2	2	0	537353846	0,26	0,16	0,78	0,40	2,70325E-09	2,21544E-09
T7-F2	2	2	6	4	5	1	387015963	0,26	0,16	0,75	0,40	4,99856E-09	3,99885E-09
T7-F3	0	0	6	1	2	1	486448765	0,26	0,16	0,75	0,40	6,1182E-10	7,34184E-10
T7-F4	3	3	7	1	2	0	339480861	0,26	0,16	0,75	0,40	6,13682E-09	5,26013E-09
T7-J1	1	1	1	0	2	1	523097691	0,26	0,16	0,75	0,40	1,70687E-09	1,82066E-09
T7-J2	0	0	4	1	4	1	111032774	0,26	0,16	0,75	0,40	5,36092E-09	3,21655E-09
T7-J3	5	**4**	4	1	5	1	453635396	0,26	0,16	0,75	0,40	8,20094E-09	6,03589E-09
T7-J4	1	1	4	0	4	1	522882471	0,26	0,16	0,75	0,40	2,27676E-09	1,82141E-09
T7-M1	1	1	6	0	3	1	470027272	0,26	0,16	0,75	0,40	2,21618E-09	2,02622E-09
T7-M2	4	4	6	1	1	0	543083854	0,26	0,16	0,75	0,40	4,65814E-09	4,38413E-09
T7	**19**	**18**	**50**	**11**	**30**	**7**	**4374058893**	**0,26**	**0,16**	**0,72**	**0,38**	**3,70976E-09**	**3,04077E-09**
T10-F1	0	0	2	0	1	0	207044795	0,26	0,16	0,75	0,40	7,18731E-10	0
T10-F2	1	**0**	4	0	1	1	171183838	0,26	0,16	0,75	0,40	4,34648E-09	2,08631E-09
T10-J1	1	1	4	0	3	1	168713342	0,26	0,16	0,75	0,40	6,17418E-09	5,64497E-09
T10-J2	1	**0**	4	0	0	0	207433966	0,26	0,16	0,75	0,40	2,86953E-09	0
T10-J3	0	0	5	0	1	0	203968443	0,26	0,16	0,75	0,40	7,29571E-10	0
T10-J4	0	0	0	0	0	0	182529580	0,26	0,16	0,75	0,40	0	0
T10-J5	0	0	2	1	0	0	207344449	0,26	0,16	0,75	0,40	0	0
T10-J6	0	0	3	0	1	1	208594637	0,26	0,16	0,75	0,40	7,13391E-10	1,71214E-09
T10-M1	2	2	6	2	1	1	204492850	0,26	0,16	0,75	0,40	6,5493E-09	7,56808E-09
T10-M2	0	0	5	0	1	0	42850423	0,26	0,16	0,75	0,40	3,47277E-09	0
T10	**5**	**3**	**35**	**3**	**9**	**4**	**1804156323**	0,26	0,16	**0,88**	**0,50**	**2,0208E-09**	**1,64963E-09**
Total:	**24**	**21**	**85**	**14**	**39**	**11**	**6178215216**	**0,26**	**0,16**	**0,75**	**0,40**	**3,25163E-09**	**2,65911E-09**

Chapter II

Title: Genetic load is affected by demographic histories in Trinidadian guppies (*Poecilia reticulata*), but does not explain invasiveness after a recent artificial translocation

Katarzyna Burda[1], Mary J Janecka[2], Ryan Mohhamed[3], David R. Clark[4], Rachael Kramp[4], Jacek Radwan[1], Mateusz Konczal[1*]

[1] Evolutionary Biology Group, Faculty of Biology, Adam Mickiewicz University; 60-614 Poznań, Poland.

[2] University of Texas at El Paso, El Paso, Texas, USA

[3] Auburn University, Auburn, Alabama, USA

[4] University of Pittsburgh, Pittsburg, Pennsylvania, USA.

*mateusz.konczal@amu.edu.pl

Abstract

Both population size and its dynamics are thought to shape genetic load of population. Small size leads to increased inbreeding, reduces selection against deleterious genetic variants, which may lead to population decline. In this study, we asses the genetic load in wild guppy (*Poecilia reticulata*) populations in North Trinidad and Tobago, with particular focus on fish from Turure river, where a handful of introduced individuals expanded downstream, hybridized and displaced most of the native population. Using whole-genome sequencing, we look for harmful variation within the populations and compare their relative counts to find patterns in genetic load distributions. Contrarily to expectations, we find that Turure expansion was not followed by accumulation of genetic load, probably due to mixing with original guppies and fish from neighboring rivers. We found no evidence for purging of highly deleterious variants, negating its role in invasion success. We do find evidence for smaller populations from Tobago to host higher mutation load compared to larger Trinidadian populations. We also find that upper populations have higher load compared to their lower counterparts, even though the expectation of lower N_e in former did not universally held. Finally, we found no relation between effective population size across all Trinidad and Tobago populations and genetic load estimations, showing that assessing population genetic condition could benefit from an integrated approach of both neutral diversity and load analyses. Most importantly, this study highlights the great importance of gene flow in shaping the genetic load in small populations.

1. *Introduction*

Genetic load refers to fitness decrease due to presence of deleterious variants in a genome (Crow, 1958). While mutations are the main source of genetic variation enabling adaptive evolution (Barrett & Schluter, 2008), we can assume that a random mutation is more likely to be deleterious than beneficial (Ohta, 1973). Most of the deleterious mutations are, however, recessive and remain rare in large panmictic populations (Eyre-Walker & Keightley, 2007) thus not contributing to population's fitness. Such deleterious variation present in heterozygous genotypes is often referred to as a masked load (Bertorelle et al., 2022). Under some demographic scenarios, such rare deleterious variants can, however, increase in frequency, get exposed in homozygotes as a realized load and pose a threat to a population. Alternatively, in such scenarios, the masked load can be also eliminated from a population, either by chance due to genetic drift or by selection when turned into realized load (Dussex et al., 2023; J. Robinson et al., 2023). Understanding dynamics of deleterious mutations and consequent variation in genetic load is therefore crucial for identifying threats and predicting evolutionary future of a population.

Demographic changes, such as a population size reduction event (bottleneck), may affects expectations regarding genetic variation and genetic load of a population through enhanced role of genetic drift (Wright, 1932) and increased inbreeding (Frankham, 1995; Wright, 1922). The drift allows some of deleterious alleles to rise in frequency during bottleneck due to the fact that efficacy of selection against them depends on the effective population size of a population (N_e) (de Pedro et al., 2021; Stewart et al., 2017; Zeitler et al., 2023). In small populations, selection fails to remove deleterious mutations of weak effects (purging is inefficient if their selection coefficient is less than $1/4N_e$), and these variants accumulate over time (Kimura & Ohta, 1969), become more common and even fix in a population (Kimura et al., 1963). This leads to fitness decline which may then cause further reduction of population size, the process repeats and finally results in trapping the population in a vicious circle of mutational meltdown (Lynch et al., 1995).

However, bottleneck might also have a positive influence on a population fitness. An expected consequence of population size reduction is increased probability that alleles carried by two potential mating partners are identical by descent, leading to increase of homozygosity in a population (Wright, 1922) and changing masked load into realized load. In such a case, highly deleterious variants, that existed in pre-bottleneck population at low frequencies, are either lost by chance or become exposed to natural selection in homozygous genotypes (Charlesworth & Willis, 2009). If selection can overcome genetic drift, it effectively purges such large effect deleterious variants, ultimately increasing population fitness. Relative importance of such purging and mentioned above mutational meltdown are relevant in various contexts, including conservation biology and biological invasions (Dussex et al., 2023; Sherpa & Després, 2021).

In conservation context, purging can alleviate extinction risk of a population and potentially explain observations of small populations persistence for long periods of time despite low genetic variation (Dehasque et al., 2024; J. A. Robinson et al., 2016). Generally, however, there is a weak relationship between effective population size and genetic load (Schmidt et al., 2023; Wilder et al., 2023), suggesting that more nuanced information about demographic history is needed to predict changes in genetic burden. Moreover, genetic load is difficult to compare between species, as we lack a general bioinformatic framework for such analyses (Dussex et al., 2023), while at the same time endangered species rarely consist of multiple populations characterized by different demographic histories and wide range of effective population sizes. As a consequence, intraspecies associations between N_e and genetic load were studied based on only a handful of cases (Kleinman-Ruiz et al., 2022; Mathur & DeWoody, 2021; Smeds & Ellegren, 2023).

Apart from conservation context, another relevant example of bottleneck associated scenario is a biological invasion. Both natural and artificial dispersions usually introduce a handful of individuals, thus initial stages of invasion are characterized by small effective population size – so called founder effect. It has been suggested that enhanced purging during this stage can then contribute to invasive potential of a population (Estoup et al., 2016). Purging in invasive species have been however documented only in few cases, measuring either life history traits (Facon et al., 2011) or using single locus approach (Zayed et al., 2007), making it difficult to generalize the results. Following initial bottleneck, invasive species experience rapid range expansion. Such expansions can be associated with multiple subsequent bottlenecks, increasing effects of genetic drift especially at the front of the expansion wave (Peischl & Excoffier, 2015; Pfennig et al., 2016). Dynamic of deleterious mutations might differ at this stage, when compared to the initial phase. In particular, it has been suggested that population expansion can lead to the accumulation of genetic load in the migrants population, especially if the new territory is isolated and the gene flow is limited (Peischl et al., 2015). In sparce front populations, drift is powerful and rare alleles may randomly rise to high frequencies in a process called ‘gene surfing‘ (Gilbert et al., 2017; Peischl & Excoffier, 2015). This could lead to fixation of deleterious variants and their linked neighbors, especially in case of low recombination rate which hinders breaking apart harmful haplotypes (Zhang et al., 2016). It has been suggested that such scenario contributed to accumulation of genetic load for example in non-African human populations (Henn et al., 2016), French settlers in Canada (Bosshard et al., 2017) and Pacific salmon (Rougemont et al., 2020, 2023). As reported by Rougemont et al., this process intensifies with extending distance from the expansion source (more fixed alleles and accumulated load), but at the same time, the most severe mutations are successfully purged across the expansion axis.

Another important factor in species invasion is adaptation to local conditions, which may be difficult when population suffers from post-bottleneck low diversity and genetic load accumulation (de Pedro et al., 2022). However, Gilbert et al., (2018) discusses that when the expansion is slow and aided with gene flow from the source population, these difficulties can be overcome and population spread might be successful. Another source of beneficial variants in expanding populations is local population/species available for crossing (Pfennig et al., 2016; Pierce et al., 2017). Still, hybridization may also cause genetic incompatibilities (Barker et al., 2019) and admixture may introduce many unconditionally deleterious mutations to invasive population. This could happen, because small expanding population might have already purged its most harmful variants and crossing with individuals from big population may cause an influx of new undesirable variation which may increase the probability of extinction (Kyriazis et al., 2021). Whether such scenario is realistic and common in invasive species is a subject of ongoing discussion (Ralls et al., 2020). The outbreeding can also result in breaking apart coadapted gene complexes and introducing variants adapted to other conditions, consequently leading to lowered fitness in the subsequent generation (Lynch, 1991; Todesco et al., 2016). Taking all of the above into consideration, predicting dynamic of genetic load and adaptive potential during biological invasion is difficult.

Here, we explore causes and consequences of expansion in Trinidadian guppies (*Poecilia reticulata*), by estimating genetic load across populations where translocated individuals rapidly spread and replaced naïve populations. Guppies are a popular model species naturally occurring at the fresh waters of Caribbean Islands and South America. The Trinidadian populations have been intensively studied by generations of ecologists and evolutionary biologists (Magurran, 2005) and some of these studies included translocation experiments (Fitzpatrick et al., 2016; Fraser et al., 2015). Consequently natural population structure of guppies is affected by artificial, human-facilitated transplants of guppies between different predation regimes. One of such interventions includes Turure translocation that occurred in 1957 (Endler, 1980; Reznick & Bryga, 1987; Shaw et al., 1992). The fish from lower Guanapo river (high predation downstream, Caroni drainage) were moved to Turure river (low predation upstream, Oropouche drainage), a site which lacked guppies prior to introduction (Magurran, 2005) Importantly, the donor and recipient drainages are inhabited by diverged populations of guppies (Willing et al., 2010), with long separate histories (they split 0.18-1.2 mya, [Fajen & Breden, 1992, Whiting et al., 2021]) and evidences of partial reproductive isolation (Russell & Magurran, 2006) that, as proposed by some researchers, may situate them on a speciation path (Schories et al., 2009). However, despite this considerable divergence, they still can hybridize and produce viable offspring, so the evidence for barriers to gene flow is weak (Devigili et al., 2018; Magurran, 1996). In fact, several sources of evidence demonstrated that such crossing happened in the introduced Turure river (Becher & Magurran, 2000), which downstream sites had originally been inhabited by native guppies. After the translocation, the invasion began. Guanapo-derived population expanded and largely replaced original genotypes, with very limited admixture from resident populations demonstrated by previous genetic studies based on microsatellites and panels of single nucleotide polymorphisms (Becher & Magurran, 2000; Sievers et al., 2012; Willing et al., 2010). We aim to test whether spread of Guanapo-derived population was associated with changes of genetic load. In particular, we test the hypothesis that population introduced in the upstream Turure was purged from highly deleterious mutations and is now characterized by overall low genetic load, facilitating its invasive potential (Sievers et al., 2012). To test the above hypothesis, we compare genetic load of upstream Turure population with genetic load of populations from other rivers in Oropouche drainage. Additionally, we explore a pattern of accumulation of genetic load along the Turure river. We hypothesize that population expansion was associated with a series of bottlenecks, leading to increase in genetic load down the river. However, if admixture appeared along with expansion, as some earlier work suggest (Sievers et al., 2012; Willing et al., 2010), the genetic load might also be shaped by this process. We thus explore both genetic load and admixture pattern.

Finally, using multiple populations from other rivers we test for association between effective population size and its genetic load. Firstly, we approach this by exploring general relationship between long term effective population size and genetic load of a population. Trinidadian rivers differ in effective population size between hundreds to dozens of thousands (Whiting et al., 2021), providing the opportunity to test for such association. We can expect that such relationship is negative, if genetic drift decreases effectiveness of selection against deleterious mutations in long periods of time. Alternatively, positive relationship would suggest that purging plays a major role in shaping genetic load in the evolutionary timescale. Secondly, to test the association between genetic load and population history we test specific contrasts, designed by prior knowledge of the system. In particular, we explore genetic load differentiation between: upstream versus downstream locations (i.e. low and high predation sites) and between Trinidad and Tobago islands. In case of populations inhabiting different locations within one river, upstream, populations probably underwent several bottlenecks during upstream colonization (Magurran, 2005). Genomic studies support this expectations, demonstrating that populations near the river source have lower synonymous site diversity than their downstream counterparts and show other expected signals of bottleneck (Qiu et al., 2022; Whiting et al., 2021). We thus expect that these populations accumulate genetic load at different rate when compared to the downstream, high predation sites. Similarly, Tobago rivers are much shorter than those on Trinidad and guppy populations inhabiting them are genetically less diverse due to their limited population size and isolation (Barson et al., 2009; Herdegen-Radwan et al., 2021), so we anticipate to see differentiation in genetic load between the islands as well.

To test the above hypotheses and predictions, we estimate genetic load from resequenced genomes. While direct estimations are difficult to obtain, several methods try to identify mutations which are likely to be harmful, and sum their potential effects to enable comparisons between individuals and populations in terms of the genetic load. Variant deleteriousness can be estimated by its effect on the protein translation, especially if the mutation introduces a stop codon. This kind of mutation (i.e. nonsense) is the most severe kind of substitution and is expected to have an important impact on individual fitness (Robinson et al., 2023). Summing number of nonsense mutations per individual can be then used as a proxy of genetic load. Another commonly used method of genetic load estimations is based on conservation scores (CS) of variants in the populations. CS is calculated from the multi-alignments of many species and quantified as "rejected substitutions" (i.e. when the number of observed substitutions in the alignment is lower than the number of expected substitutions, given position is inferred to be conserved, [Davydov et al., 2010]), so it also provides information on severity of the site constraint. Both nonsense mutations and mutations with high CS scores, have previously been utilized in the variety of studies and are already well established approaches in genetic load estimation studies (Bertorelle et al., 2022). Here, we use both these metrics, to investigate genetic load in guppies of Trinidad and Tobago. Furthermore, we account for variant genotype as a proxy for masked and realized genetic loads as emphasized by Peischl & Excoffier (2015).

*2. **Materials and methods***

Scripts used in this work are available on GitHub: https://github.com/0-Ioniel-0/Expantion_Load.

*2.1 **Sampling and sequencing***

Individuals used in the study come from two batches of sampling and sequencing. First batch (51 individuals) of sampling took place in 2018 and was sequenced in 2019. Second batch (140 individuals) was sampled and sequenced in 2022. All individuals are wild fish from natural populations of Trinidad and Tobago. The sampling procedure was consistent with the Polish law, European Directive 2010/63/EU and was conducted with the permission of Ministry of Agriculture, Land and Fisheries of Trinidad and Tobago.

We sampled from 8 rivers on Trinidad: Arima, Caura, Lopinot and Santa Cruz originated from Caroni drainage; La Seiva, Oropouche, Quare, and Turure originated from Oropouche drainage. Additionally we sampled 2 rivers on Tobago (Dog River and Roxborough), getting together 191 samples and 14 sampling locations/populations in total. Three rivers were sampled in more than one location: Quare (upper and lower), Oropouche (upper and lower), and Turure (upper, middle and lower); see Figure 1 and Supplementary Table 1 for more details and geographic coordinates. Fish were anesthetized using tricaine mesylate (MS-222) and preserved in 98% ethanol. Samples were then transported to Poland for further molecular procedures. Tissue from clipped tails was used for DNA extraction (with Thermo Scientific™ MagJET™ Genomic DNA kit) and genomic libraries generation (with the NEBNext® Ultra™ II FS DNA Library Prep Kit and NEBNext® Multiplex Oligos for Illumina® indices) according to manufacturer protocols. Subsequently, all samples were sequenced on Illumina NovaSeq 6000 platform using 2x150bp mode at the NGI Sweden.

2.2 ***Reads Alignment***

Raw sequencing reads were filtered with Trimmomatic (v0.39; Bolger et al., 2014) and aligned to guppy female reference genome (GCF_000633615.2; Künstner et al., 2016) using bwa mem and default parameters (v0.7.10; Li, 2013). Resulted BAM files were then subjected to duplicates marking with Picard tools (v2.21.6; http://broadinstitute.github.io/picard/).

2.3 ***Population structure***

In order to understand genetic composition of samples, the principal component analysis (PCA) and the structure analyses were performed. First, we jointly called variants in all BAM files using samtools mpileup (with -a AD parameter; v1.20, Danecek et al., 2021) and bcftools call (with -G populations indicating parameter and -m multiallelic parameter; v1.20, Danecek et al., 2021). Resulting VCF was then filtered with bcftools (v0.1.12b, Danecek et al., 2011) using the following thresholds: MapppingQuality > 40, INFO/Depth < 5200, INFO/Depth > 1300, QUAL ≥ 30, MappingQualityBias > -3, ReadPositionBias < 3, SoftClipLengthBias < 3; and with vcftools (v0.1.12b, Danecek et al. 2011): MinorAlleleFraction ≥ 0.01, MaxAlelles = 2, MaxMissing = 50%; Depth ≥ 6 and GenotypeQuality ≥ 70. Further quality assessment performed with bcftools demonstrated that 1 sample from lower Oropouche shows low mean coverage, so it was excluded from any further analyses and from the VCF file. We then used plink (v1.9; Purcell et al., 2007) to transform the VCF file into a BED file and pruned it to avoid variants in linkage disequilibrium using 200bp window, 1bp step and 0.5 r^2 options. Resulting BED file was then used to perform PCA analyses with plink. Then, using the same BED file, we ran a Admixture (with -s random seed parameter; v1.3, Alexander et al., 2009) 10 times for each K parameter (1-14). The Admixture results were plotted with pong software (with -s 0.95 threshold to combine similar clusters at given K; v1.5, Behr et al., 2016).

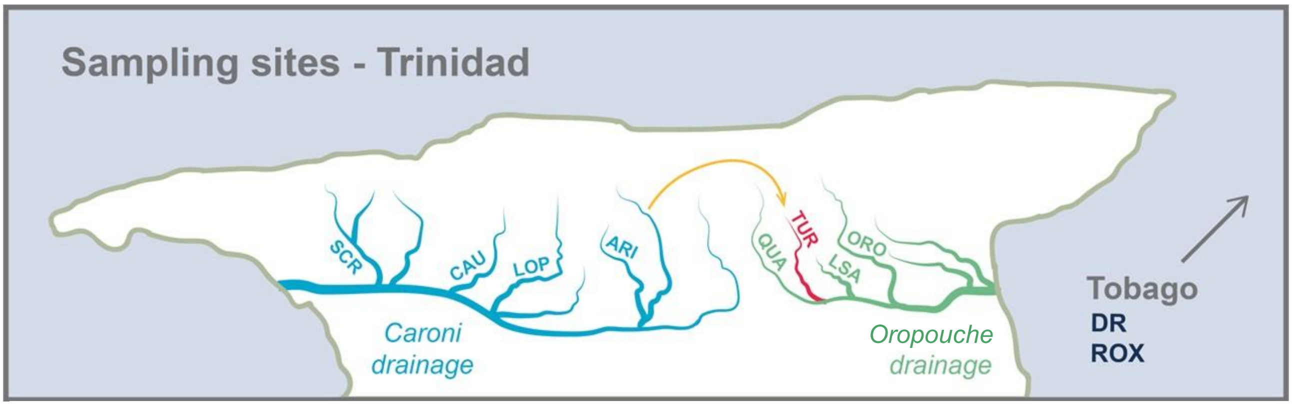


Figure 1 Outline map of rivers sampled in Trinidad. The yellow arrow shows translocation of guppies from Guanapo to Turure. The rivers abbreviations translate: SCR (Santa Cruz), CAU (Caura), LOP (Lopinot), ARI (Arima), QUA (Quare), TUR (Turure), LSA (La Seiva), ORO (Oropouche).

*2.4 **Ancestral allele inference***

To infer guppy ancestral alleles (AA) we used approach by Smeds et al. (2023) and genomic information from two closely related outgroups: *Poecilia picta* and *Xiphophorus maculatus*. Data for these two species was retrieved from NCBI's SRA (accession number ERR4077394 and SRR13649980, respectively) in fastq.gz format. Reads in those files had already been trimmed and adapters had been removed prior to download, so without any extra filtering steps, we mapped them to the female guppy reference genome. Genotype calling was performed similarly to what has been described above, resulting in whole genome VCF files. Next, those files were used in pseudo_haploidize.py script by Smeds et al. (2023, GitHub: https://github.com/linneas/wolf-deleterious) to infer ancestral alleles based on allele depth weight. Minimal DP thresholds were set to half of the expected coverage of *X. maculatus* and *P. picta,* 7 and 11 respectively, and minimal genotype quality to 30. When thresholds were not met, the haplotype was set as N. Two resulting BED files (one per species) were merged and only sites where nucleotide was identical for both species (and not N) were used as an informative source of the ancestral allele.

*2.5 **Genotype calling***

With the goal of finding accumulated deleterious variation, we used reads mapped to the reference genome to infer genotypes again, this time with GATK software (HaplotypeCaller with -ERC BP_RESOLUTION and --output-mode EMIT_ALL_CONFIDENT_SITES parameters, v4.1.4.1; McKenna et al., 2010). All subsequent analyses were run for each of a populations separately. Single individual genomic VCF files (GVFCs) were merged into population GVCFs (GATK CombineGVCFs) and used in GATK GenotypeGVCFs to produce VCF files in all sites mode (-all-sites). Samples previously identified as showing low coverage were removed. We also removed all sites with less than 70% of genotypes meeting genotype quality thresholds (DP ≥ 6 and GP ≥ 30) to reduce computational demands for further steps.

*2.6 **Ancestral allele and conservation score filtration***

Subsequently, we filtered the VCF files by ancestral allele and conservation score (CS) presence. In GERP software, CS is calculated on a per base resolution among a group of species in a multi-alignment. Since it is calculated as the number of substitutions expected under neutrality minus the number of substitutions observed at the site, high values mean more conserved sites and negative values mean non conserved, variable sites (Cooper et al., 2005).

CS BED file was downloaded from the Ensembl fish Compara database (release 111, 65 fish species alignment, CS calculated with GERP, Herrero et al., 2016). The file was then intersected with ancestral allele BED file. Resulting CS-AA BED was used to filter guppy VCF files.

Kept sites had both CS and AA as one of the two alleles (REF and ALT, multiallelic sites were removed). Information about CS and AA values were added to the VCF files INFO field using original Python script. In case of sites, where AA was the alternative to the reference nucleotide, the two alleles were swapped, so that REF always represents ancestral allele in the final VCF file.

*2.7 **Quality filtration***

During the next step we used original Python script and thresholds of DP and GQ to mask low quality genotypes. Minimum genotype quality was set to 30 and minimum depth to 6. Since our samples differed in expected coverage, we filtered them by their respective maximum DP (sites with DP > 2x average DP of a given sample were filtered out) and we again filtered for maximum 30% missing data in sites. Finally, the files were used to extract SNPs and remove those which do not meet filters recommended by GATK (QualityByDepth $\geq$ 2, MappingQuality $\geq$ 40, FisherScore $\leq$ 60, StrandOddsRation $\leq$ 3, MappingQualityRankSum ≥ -12.5, ReadPosRankSum ≥ -8).

2.8 ***Effective population size calculation***

Having all callable sites (with AA and CS), we next calculated effective population size (N_e) for neutral positions in each of the populations. First we searched for 4-fold degenerated sites in guppy genome using find_4fds script (GitHub: https://github.com/mattheatley/extract_4fds) and used them to filter our VCF files (including polymorphic and monomorphic positions). Resulting files were used to calculate nucleotide diversity (π) in 1Mb windows across the genome with pixy tool (v1.2.7.beta1; Korunes & Samuk, 2021). Finally, genome-averaged π values together with mutation rate ($\mu = 2.9 \times 10^{-9}$, Burda & Konczal, 2023) were used to obtain N_e for our populations, using the following formula:

$$N_e = \frac{\pi}{4\mu}$$

2.9 ***Genetic load estimation***

To calculate per-individual relative genetic load (GL) we followed approached proposed by Dussex et al. 2021, using the following formula:

$$ML = \frac{\sum^{n}_{\substack{i=1 \\ CS_i \geq 2}} CS_i \cdot D_i}{\sum_{i=1}^{n} D_i}$$

Where n is the number of variants, CS_i is the conservation score and D_i is the number of derived alleles of the i^{th} variant. Note, that in the upper part of the equation, only alleles from highly conserved sites are summed (conservation score equal or higher than 2).

We used called SNPs to identify loss of function (LoF, nonsense) variants with SNPEff tool (v5.0; (Cingolani et al., 2012). The values were calculated and normalized against all derived alleles using following:

$$value\ LOF = \sum^{n}_{\substack{i=1 \\ t\,\in\,LOF}} D_i$$

$$normalized\ value\ LOF = \frac{\sum^{n}_{\substack{i=1 \\ t\,\in\,LOF}} D_i}{\sum_{i=1}^{n} D_i}$$

Where n is the number of variants, t is the variant type and D_i is the number of derived alleles of the i^{th} variant.

Given that deleterious mutations are on average recessive and hardly affect fitness in the presence of another allele, for each calculation and test (see below), we also conducted separate analyses on the heterozygous and homozygous subsets.

2.10 ***Statistical analysis***

All statistical analyses were performed in R (v4.3.3), for overview of all models used, see Supplementary Table 2. Conformance to the assumptions of normality and homoscedasticity was examined with diagnostic plots.

First, we analyzed relative genetic load in islands (Trinidad vs. Tobago) with mixed-effects linear models using `lme4` package, accounting for the random effect of population. In analyses which contained less than 5 populations, estimation of random effect is unreliable (Bolker et al., 2009). Consequently, comparisons of upper and lower locations performed using two way analysis of variance (ANOVA) using `aov` function, with population crossed with location. Differences in load between sites along the Turure river were tested with one-way ANOVA. Square root and inversion transformations were implemented in Turure analysis of homozygotes and heterozygotes, respectively, to improve normality of error distribution. To find how different Turure locations differ between each other, we performed a post-hoc contrasts analysis using the `emmeans` package. Next, we compared well established populations from Oropouche drainage (lower Oropouche and lower Quare) with upper Turure location, which has suffered through a bottleneck. This was done using two-way analysis of variance (population crossed with whether a recent bottleneck happened, ‘yes’ for Turure and ‘no’ for Quare and Oropouche). In order to obtain normal distributions, homozygotes data was transformed with square root and heterozygotes data underwent squaring.

Then, similarly to relative genetic load comparisons, but using generalized linear model, due to different data distribution, we analyzed differences in normalized counts of LOF alleles. A vector of LOF alleles count and total number of derived alleles created with `cbind` function was used as a dependent variable. We employed either binomial or quasibinomial distribution, to address dispersion issues where needed (see Supp. Table 2). In case of three Turure sites comparison, the `emmeans` package was employed to compute the pairwise differences.

Finally, we checked for relationship between effective population size and averaged per-population genetic load, as well as the averaged per-population normalized count of LOF alleles. To test this relations, we used linear models (with normal distribution) and included ‘Region’ as additional predictor variable (‘Region’ levels being Caroni, Oropuche and Tobago).

Importantly, in analysis of islands only the lower river sites were used. Additionally, in analyses of effective population size correlation and islands comparison, all Turure populations were excluded due to their atypical characteristics (recent bottleneck, expansion and hybridization).

3. ***Results***

3.1 ***Resequencing***

Samples were sequenced yielding mean coverage of 15x (9x - 36x) and were mapped to the reference genome with over 97% reads aligned in every sample. Most individuals (190) met our stringent quality thresholds and only one sample from the lower Oropouche was removed.

3.2 Ancestral alleles, conservation scores and filtration

X. maculatus and *P. picta* reads were aligned to *P. reticulata* reference genome yielding mean coverage of 14x and 23x respectively. Total number of sites with available ancestral allele was 298,085,628 (40,74% of the genome). Secondly, we used conservation score (CS). Intersection of sites with estimated CS (72,620,521 positions, 9.9% of the genome) and inferred ancestral state yielded 42,860,381 sites (5.86% of the genome). These sites were used for further analyses.

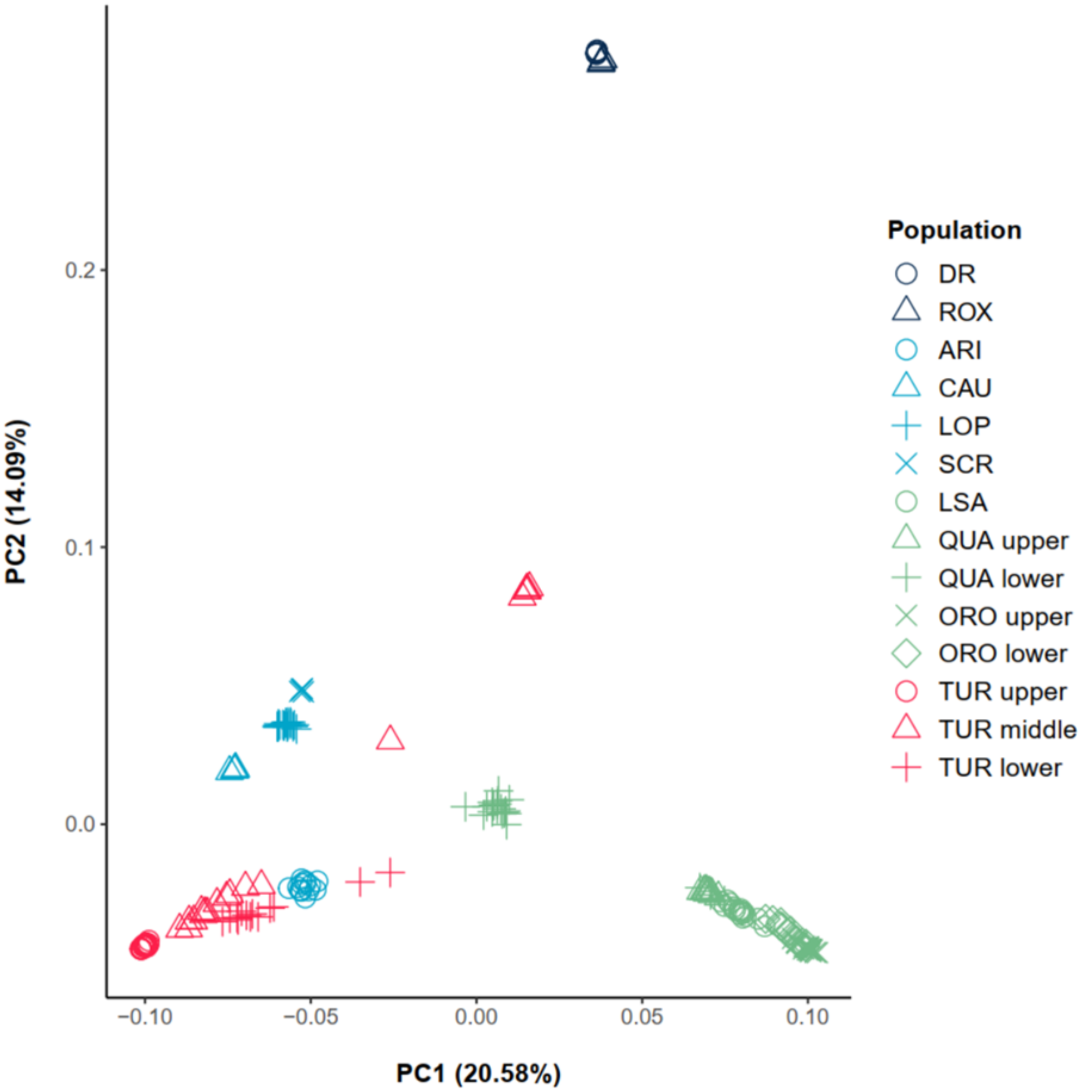


Figure 2 Principal Component Analysis of populations' variants. The PC1 is on the x-axis, the PC2 is on the y-axis. Each point is one individual. Tobago, Caroni, Oropouche and Turure populations are coloured with dark blue, light blue, green and red, respectively.

Subsequently, we called genotypes within each guppy population separately. After excluding samples mentioned earlier, we analyzed 14 populations with 5 up to 22 individuals per population. Number of callable sites included in the analyses span 211 to 465 millions per population (Supplementary Fig.1). After extracting sites with AA and CS information and filtering data we ended up with 11 –

33 million callable sites per population (see Supplementary Fig. 1 and Materials and Methods for more details). The final number of filtered SNPs was 457,062.

3.3 Population structure

Population structure and genetic differentiation of the samples was investigated using PCA and Admixture analyses and done on data not filtered with AA and CS. Samples from Trinidad and Tobago islands were strongly separated by the second Principal Component, explaining 14.09% of variation (Fig.2). The first Principal Component (20.58% explained variation) separates Oropouche from the Caroni drainages, the later including translocated Turure population. Quare lower does not cluster with either of these two groups, consistent with previous findings suggesting it is strongly admixed (Fitzpatrick et al. 2015, Willing et al. 2010, Suk and Neff 2009). Interestingly, we observed 5 individuals from Turure mid-upper site, which, unlike the other samples from the same site, cluster between Oropouche and Caroni populations (Fig.2). These individuals come from two different sampling years, so it is unlikely that they represent mislabeling or other technical issues. Moreover, on the admixture plot they represent pure distinct genetic ancestry, while one other individual from the same population seems to be F1 hybrid, and others seem to be admixed for this ancestry. Therefore, these individuals likely represent true distinct genetic population within a river, possible result of yet another recent translocation.

Generally, Admixture results identified the optimal K-value to be 8. The plot for this K-value, supported by 4 out of 10 runs with a pairwise similarity of 0.999, reveals a strong differentiation among islands and drainages. The upper and lower Oropouche sites remain highly similar to each other whilst Quare sites visibly differ, with higher variation in the genetic ancestries in the lower site. In the Turure river where recent translocation from Caroni drainage happened, the upper site is genetically homogenous, but as expected going down the stream, we can see increased rate of admixture, likely demonstrating apparent crosses between introduced and native populations. The

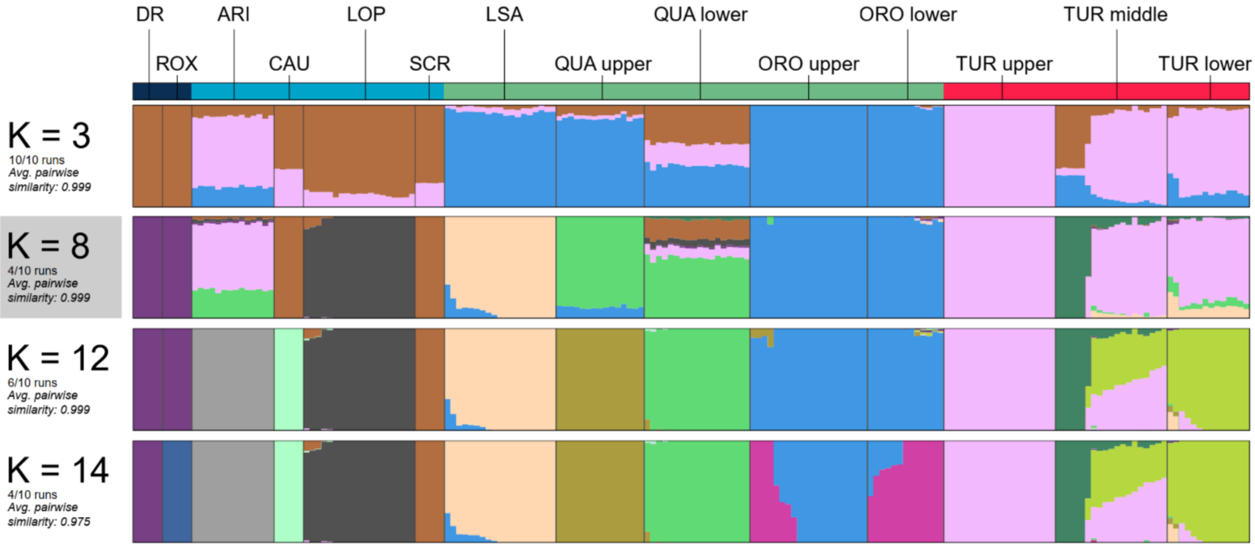


Figure 3 Admixture plots. The plots representing four values of K are presented (3, 8, 12, 15), together with information on number of runs supporting given plot and average pairwise similarity of the runs. The populations are color-coded in the strip above the plots as follows: Tobago (dark blue), Caroni (light blue), Oropouche (green), and Turure (red).

vast majority of the genetic ancestry is, however, from the introduced populations, and this genetic ancestry likely started to spread to nearby rivers (e.g. Quare lower, Fig.3).

3.4 *Nucleotide diversity and effective population size*

In order to infer effective population size, we first calculated nucleotide diversity in 4-fold degenerated sites. Genome-wide nucleotide diversity in these sites span 8.6 x 10^{-4} – 35.8 x 10^{-4}, what corresponds to N_e ranging from around 80,000 to 300,000 individuals (Fig.4). Populations inhabiting Tobago island have lower effective population sizes than those from Trinidad while the lowest N_e in the whole dataset was inferred for Turure river. Oropouche drainage has the highest N_e with the upper Quare population reaching over 300,000, likely resulting from its admixed origin (see Fig.3).

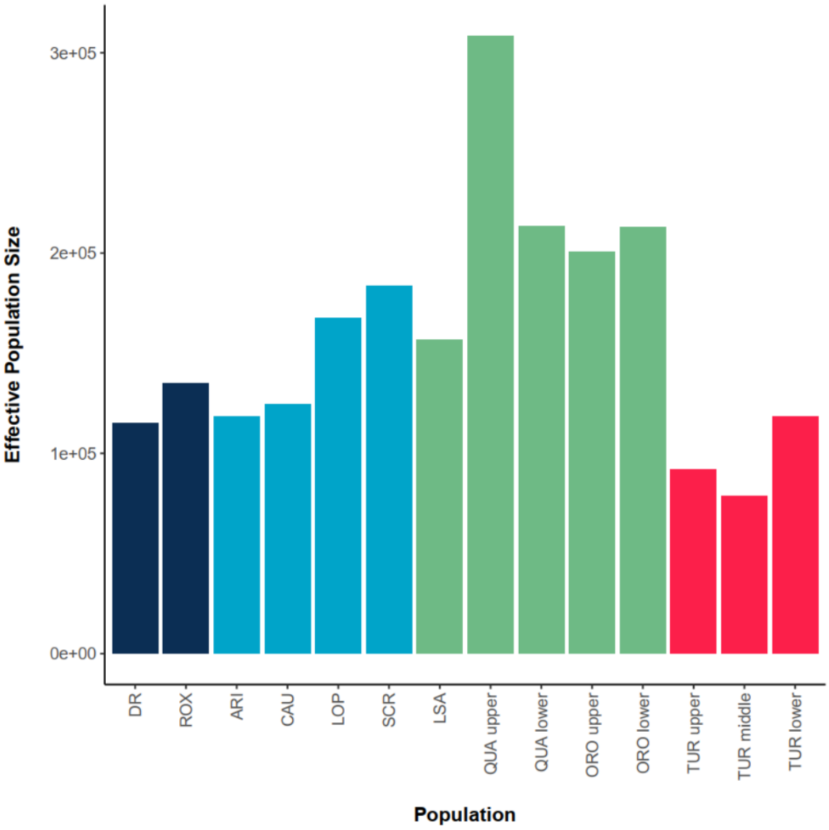


Figure 4 Estimated effective population size for analysed populations. Populations are on the x-axis, the effective population size is on the y-axis. The populations are color-coded as follows: Tobago (dark blue), Caroni (light blue), Oropouche (green), and Turure (red).

3.5 *Genetic load*

Next, we calculated per-individual relative genetic load and normalized loss of function alleles count, in total, in heterozygotes and in homozygotes. On this dataset, we performed a series of comparisons.

Tobago individuals carry more deleterious variation than Trinidadian ones in almost all comparisons, except in LOF homozygotes analysis, where Trinidad has a higher burden, and in heterozygotes relative load test, where no significant difference is observed (see Table 1 and Figure 5A,B).

Testing lower sites versus the upper ones revealed pattern of higher load in upstream individuals, in all comparisons but LOF alleles in heterozygotes (see Table 1 and Figure 5C,D). The three locations situated along the Turure river do not exhibit a gradient of genetic load, though they do differ significantly in most analyses (see Table 1, Table 2 and Figure 5E,F). When considering total load along the Turure river, all sites show strong differentiation, however, when focusing on homo- and heterozygotes, some populations are not statistically different from each other (middle and lower for homozygotes, lower and upper for heterozygotes). LOF alleles analyses show fewer significant differences with only middle site having lower load than the two others in total and heterozygotes tests.

Comparing load between population that has went through a bottleneck (upstream Turure) and those who have not (downstream Oropouche and downstream Quare), showed significant differences in relative genetic load in all tests, with the post-bottleneck population having higher burden (see Table 1 and Figure 5G). In LOF alleles analysis this effect was only visible in heterozygotes and it was insignificant in homozygotes and total count (see Table 1 and Figure 5H).

There was no significant relationship between effective population size and genetic burden, neither in terms of relative load, nor in case of LOF alleles (see Table 1 and Figure 6). Similar analyses performed only on hetero- or homozygotes also showed no relation (see Table 1 and Supplementary Figure 2).

Table 1 Results of statistical tests performed. In case of Turure locations comparison, standard errors for middle (first) and upper (second) locations are provided. The 'Test' column shows the name of a test and a test statistics (in brackets). Significant values are colored in dark red.

Analysis	Effect	Test (statistic)	Genotype	Estimate or SS	Df	Standard Error	Test Statistic	p-value
Relative Genetic Load	Island	LMER (t)	*Total*	-0.0064	7.0117	0.0025	-2.623	0.0342
			Homozygotes	-0.0060	7.0262	0.0021	-2.819	0.0257
			Heterozygotes	-0.0004	7.0291	0.0006	-0.698	0.5075
	Location (upper vs. lower)	AOV (F)	*Total*	1.040e-04	1	0.0002	408.1	<0.0001
			Homozygotes	6.848e-05	1	0.0003	116.73	<0.0001
			Heterozygotes	3.703e-06	1	1.533e-04	19.996	<0.0001
	Turure Location		*Total*	1.425e-04	2	0.0002 0.0002	176.3	<0.0001
			Homozygotes	0.006097	2	0.0022 0.002	93.27	<0.0001
			Heterozygotes	79942	2	13.761 12.584	31.32	<0.0001
	Founder effect		*Total*	1.485e-04	1	0.0002	389.1	<0.0001
			Homozygotes	0.004544	1	0.0028	75.98	<0.0001
			Heterozygotes	8.844e-10	1	9.827e-07	118.635	<0.0001
	RGL ~ N_e	LM (t)	*Total*	-8.541e-10	5	4.291e-08	-0.02	0.985
			Homozygotes	-1.170e-08	5	3.799e-08	-0.308	0.77
			Heterozygotes	1.085e-08	5	9.310e-09	1.165	0.296

Analysis	Effect	Test (statistic)	Genotype	Estimate or SS	Df	Standard Error	Test Statistic	p-value
LOF Alleles Count	Island	GLM (t/z)	*Total*	-0.3232	106	0.0595	-5.431	<0.0001
			Homozygotes	-0.7928	106	0.0953	-8.32	<0.0001
			Heterozygotes	0.4672	106	0.1136	4.113	<0.0001
	Location (upper vs. lower)		*Total*	0.2180	65	0.0697	3.127	0.0027
			Homozygotes	0.5372	65	0.1608	3.341	0.0014
			Heterozygotes	-0.0443	65	0.1113	-0.398	0.69
	Turure Location		*Total*	-0.2124	45	0.0938	-2.265	0.0286
			Homozygotes	-0.3679	45	0.2662	-1.382	0.174
			Heterozygotes	-0.1299	45	0.1196	-1.086	0.283
	Founder effect		*Total*	-0.1149	53	0.0989	-1.162	0.25
			Homozygotes	-0.1653	53	0.3347	-0.494	0.623
			Heterozygotes	-0.5923	53	0.1141	-5.193	<0.0001
	LOF ~ N_e	LM (t)	*Total*	1.840e-10	5	9.431e-10	0.195	0.853
			Homozygotes	-4.326e-11	5	3.431e-10	-0.126	0.9046
			Heterozygotes	3.048e-10	5	4.959e-10	0.615	0.566

Table 2 Results of the post-hoc tests regarding the Turure river locations comparisons.

		Relative Genetic Load			**LOF Alleles Count**	
		lower	middle		lower	middle
Total	middle	<0.0001		middle	<0.0001	
	upper	<0.0001	<0.0001	upper	0.0609	0.0057
		lower	middle		lower	middle
Homozygotes	middle	0.0529		middle	0.8327	
	upper	<0.0001	<0.0001	upper	0.3503	0.8031
		lower	middle		lower	middle
Heterozygotes	middle	<0.0001		middle	<0.0001	
	upper	0.1490	<0.0001	upper	0.5226	<0.0001

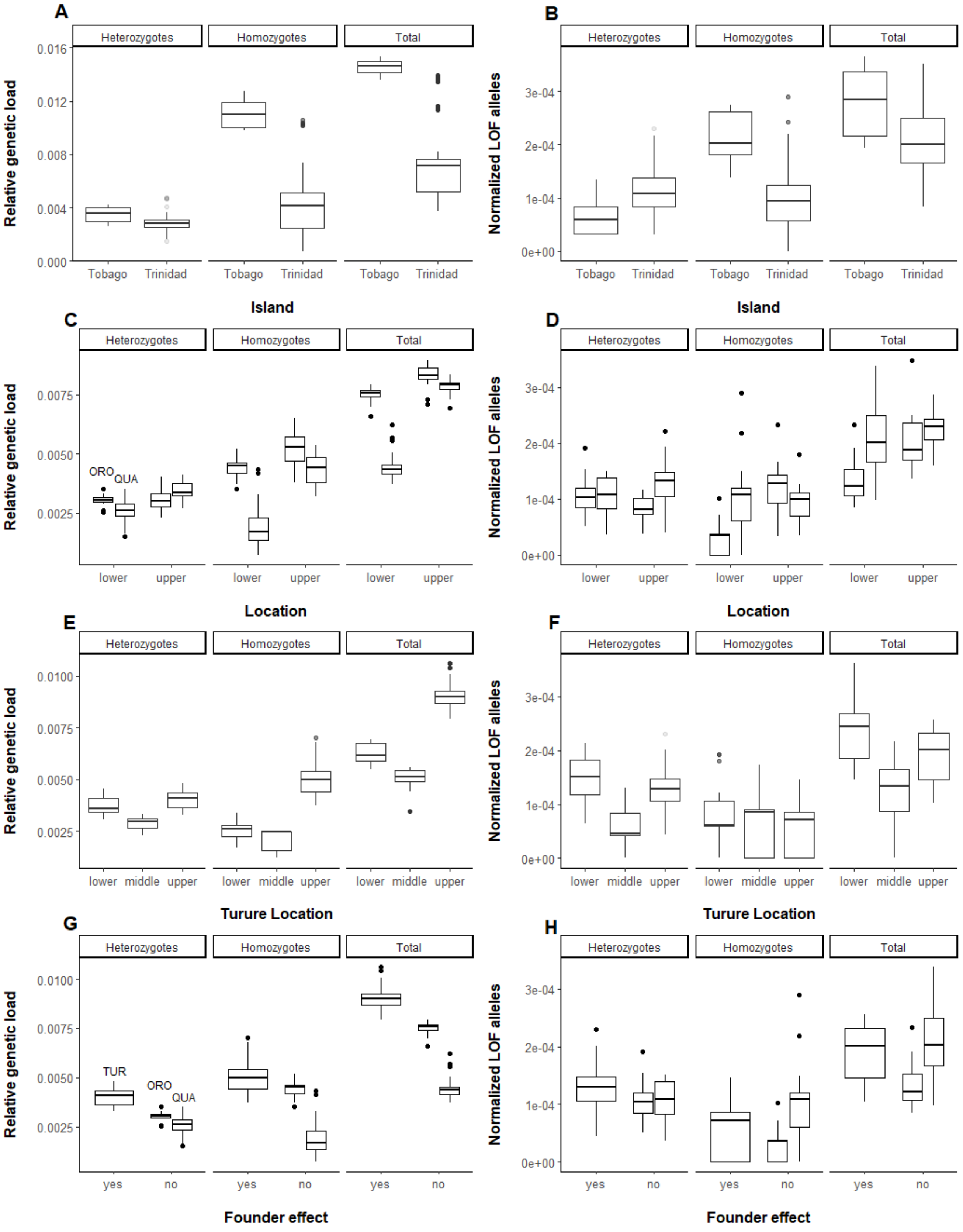


Figure 5 Figure presenting comparisons between islands, and locations. In the second row of plots (Location), the two populations are shown: Oropouche and Quare. In the last row of plots, three populations are presented: Turure, Oropuche and Quare

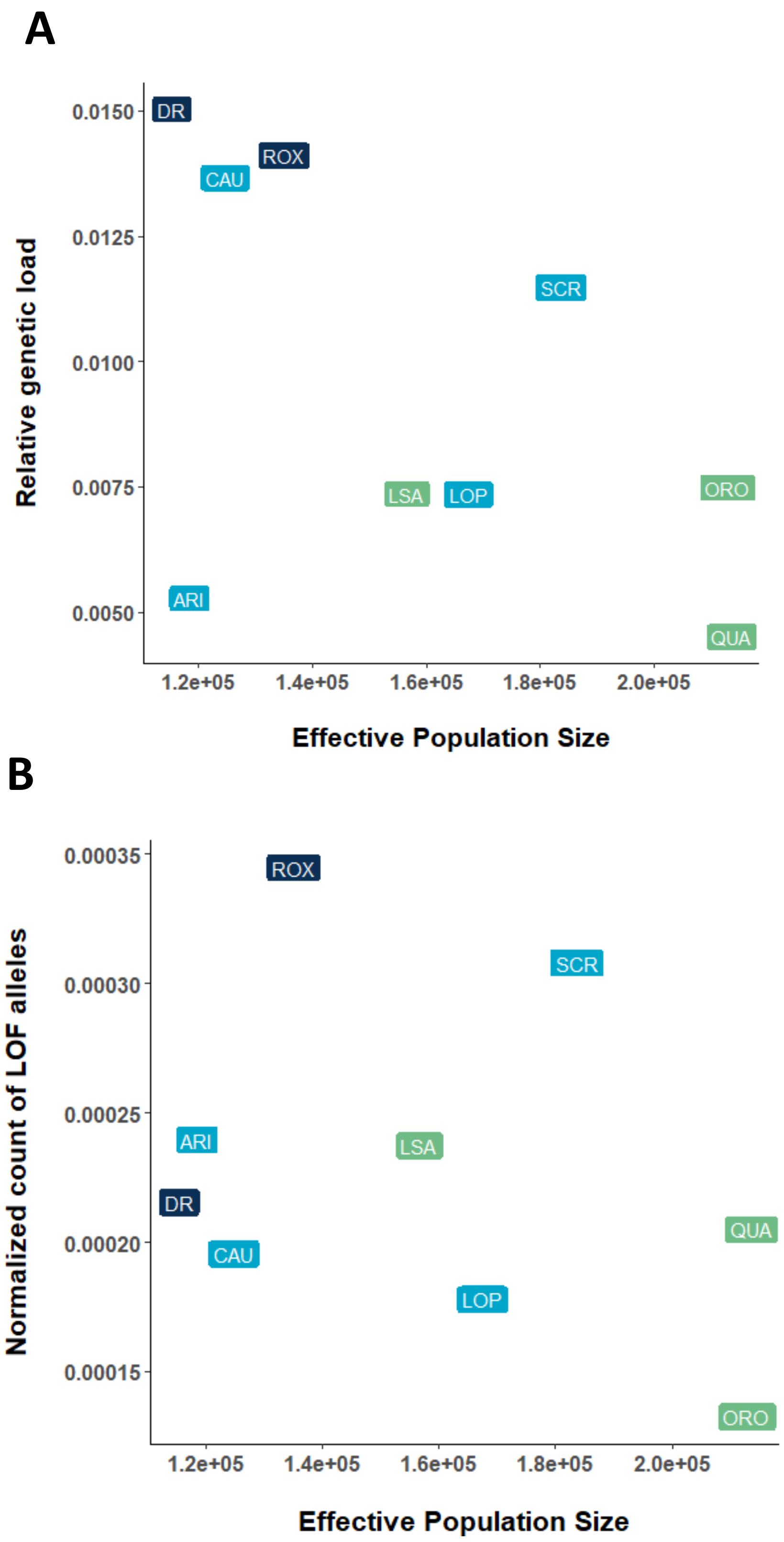


Figure 6 Relation between effective population size and two load measures. The regions are colour-coded as follows: Tobago (dark blue), Caroni (light blue) and Oropuche (green). (A) Relative genetic load. The effective population size is on the x-axis, the averaged per-population relative genetic load is on the y-axis. (B) LOF alleles count. The effective population size is on the x-axis, the averaged per population normalized LOF alleles count is on the y-axis.

4. **Discussion**

The rapid pace of global transformation, characterized by climate shifts and habitat loss, has resulted in the migration or isolation of numerous species (Johnson et al., 2017). Such rapid demographic changes can significantly affect dynamic of genetic load, providing important aspect for conservation efforts. Most of the work in this subject so far focused on endangered, highly vulnerable taxa (Kuang et al., 2020, Liu et al., 2021, Kleinman-Ruiz et al., 2022, Ochoa et al., 2022, Xie et al., 2022). Relatively less attention was paid in that context on population invasions (de Pedro et al., 2021; Rougemont et al., 2020; Tayeh et al., 2013). Our study concentrates on how load distributes and accumulates in populations of not only invulnerable, but invasive species (Deacon et al., 2011, Santana Marques et al., 2020, Lindholm et al., 2005, Rosenthal et al., 2021) making it a valuable input into knowledge on mechanics of biological invasions, particularly in presence of hybridization. We performed large scale population genomic study, analyzing hundreds of guppy genomes originated from dozen of natural populations and found that the size of the population has a negative relation with genetic load. There was no evidence of purging in the post-bottleneck population and no accumulation of load along the expansion axis. Our findings also suggest no relation between neutral genetic diversity and genetic load, at least in a species of no conservation concern, like the one studied here.

Genomic signatures of translocation, expansion and population replacement in the Turure river

Generally, results confirmed clear genetic differentiation between the two islands - Trinidad and Tobago, and between the two drainages on Trinidadian, namely Oropouche and Caroni drainages. As expected, the Turure populations clustered together with Caroni drainage, i.e. the source of their transplant. These results align with the previously reported replacement of the original Turure fish with lower-site Guanapo individuals (Suk & Neff, 2009; Willing et al., 2010). However, as shown in the Figure 3, upstream and downstream populations differ substantially and Turure downstream population is likely admixed with native ancestry. Similar pattern can be observed in Quare river, where complex admixture pattern is observed in the downstream, but not upstream population (Figure 3). In contrast, Oropouche populations seem to be homogenous along the river. This is in compliance to Suk and Neff (2009) results where authors interpreted such observations as a consequence of relative lack of geographic barriers alongside the Oropouche river. No main geographic barriers between downstream sites of the Oropouche drainage might also cause gene flow between populations. In particular, the evidence for that can be observed in the Quare population and to a smaller degree in the La Seiva population. This river is situated next to the Turure river, opening possibility for expansion to invader genotypes from Turure expansion. Indeed, our results suggest that the transplanted Turure ancestry traces can be observed at the lower Quare population, likely as a result of recent admixture (Figure 3). Such pattern was not observed in the previous study conducted around 15 years prior to our investigation (Willing et al., 2010). However, it should be noted that the previous study might have investigated different populations from the same rivers (exact geographic coordinates are not provided by Willing et al.) and was based on the limited panel of SNPs with likely lower power to detect subtle admixture (Escher et al., 2022). It is thus unclear whether admixture happened in the last 15 years, or earlier in the past. Nevertheless, this observation illustrates that the Oropouche drainage, and Turure river in particular, has experienced a fast colonization, while the colonizing and native populations mixed to a rather limited extent during this process, as indicated by Structure plot for optimal K = 8 (Figure 3). Interestingly, the plot also points to a group of individuals within middle Turure, a group with neither Oropuche nor Caroni ancestry. The origin of this cluster is unclear, but other genomes from the same population show admixture of this ancestry (Figure 3). It is thus likely that yet another translocation happened along the Turure river.

The great success of translocated population was previously discussed, when the early observations of indigenous guppies displacement were made (Becher & Magurran, 2000; Shaw, 1992). In 1992, genetic investigations based on allozymes, showed that the translocated Turure individuals established a viable population, overcome the natural barrier (perhaps with the help of flooding events) and started to spread down the river (Shaw et al., 1992). Subsequent analyses on panel of SNPs confirmed that fish from the middle stretches of the Turure river cluster together with Caroni drainage (Willing et al., 2010). The most recent analyses based on markers differentiated between local and invasive populations demonstrated that the invasion reached or moved beyond the Turure-Quare confluence (Sievers et al. 2012). The constant influx of introduced fish from above the barrier waterfall into the middle and lower parts of the Turure and the subsequent gene flow and population mixing over a sufficient amount of time, combined with stochastic environmental events such as periods of flooding, might theoretically explain such a pattern (Sievers et al. 2012). However, speed of the replacement and relatively large effective population sizes in downstream populations suggest that other factors could have contributed to the population replacement in the downstream Turure populations. If so, Haskin's introduction experiment carried out to better understand the ability of guppies to quickly adapt to new environments, may ultimately end in the disappearance of an entire series of populations in one drainage system that were genetically distinct from guppy populations in other drainages (Sievers et al., 2012). Thus, it is important to explore factors driving this process, to understand potential basis of the invasive potential of the population.

Ecological interactions, predators pressure (invasive fish originally come from high predation site, so they are well adapted) or resistance to local parasites are proposed as main advantages of Turure invaders (Becher & Magurran, 2000). Sievers et al. (2012 and 2014) demonstrated that behavioral traits can promote population interbreeding but do not give an advantage to the invasive population. The rapid start of invasion might have also been aided by initial success of hybrids (heterosis) and then introgression to invader's haplotype, but this explanation is incompatible with limited admixture from local, Oropuche genomes. There might have been yet another genetic mechanism that helped new population to outcompete the native Turure inhabitants. Immigrants, having come through a severe bottleneck during introduction, might have experienced strong purging of highly deleterious variation, which consequently could give them a genetic advantage (Marchini et al., 2016; Roman & Darling, 2007). We tested this hypothesis by comparing genetic load of Turure guppies and that of downstream populations in Oropouche drainage.

Has purging contributed to the invasive potential of the translocated population?

In contrast to our expectations, the Turure population did not show lower genetic load when compared to the downstream Oropouche populations (Table 1, Figure 5G-H). Conversely, estimated genetic load was higher in the invasive population, and the effect was significant for genetic load calculated based on both conservation scores and heterozygous genotypes of LOF alleles. These observations likely result from relaxed purifying selection in the bottlenecked population, causing increase in frequency of deleterious variation of moderate effects (Grossen et al., 2020; Gu et al., 2005). It is further supported by the distribution of conservation scores showing that derived alleles within highly conserved scores are present in Turure upper site (Supplementary Figure 3). It is thus unlikely that genetic load and number of large effect mutations declined after the bottleneck associated with founder effect of translocated individuals, and no evidence supporting the purge hypothesis was found.

Biological invasions constitute increasing problem in globally connected world, their impact on the ecosystem can be substantial and complex and consequently invasive species are recognized as one of the main threats to biodiversity (Faulkner et al., 2024; Wallingford et al., 2020). For many years, biologists postulated that evolution might play an important role in the success of invasion (Baker & Stebbins, 1966; Estoup et al., 2016). A variety of invasive species are, however, associated with so called genetic paradox of biological invasion. Three characteristics must hold true, to consider population paradoxical in that sense: i) reduced genetic variation, ii) lack of problems associated with it and iii) successful adaptation to a new environment (Estoup et al., 2016). Results presented here and in previous studies clearly demonstrate that invaders in the Turure river meet all these criteria: genetic variation is low (Figure 4), there is no evidence for negative impacts of decreased genetic diversity (Sievers et al., 2012) and finally, introduced population outcompeted locally adapted populations (Sievers et al., 2012; Suk & Neff, 2009; Willing et al., 2010, Figure 3). Such paradoxical population might be associated with several mechanisms and plausible explanations (Daly et al., 2023; Estoup et al., 2016). Among them, purging hypothesis suggests that under some circumstances highly deleterious mutations might be removed from a bottlenecked populations, increasing its invasive potential. This process has been suggested in the invasion of the harlequin ladybird (Facon et al., 2011). In this species the population went through bottleneck of intermediate intensity, and invasive, but not native populations, show almost no inbreeding depression. Additionally, evidence of purging of deleterious mutations was found in few other invasive species, like garlic mustard plant (Mullarkey et al., 2013) or bed bugs (Fountain et al., 2014). However, subsequent experimental evidence from the harlequin ladybird shows that bottlenecks fix deleterious alleles more often than they purge them (Laugier et al., 2016). Also, a recent transcriptomic analyses of this species suggest a tendency towards fixation, rather than toward purging of genetic load (Lombaert et al., 2024). Our results are thus in line with results from studies on the harlequin ladybird, showing that purging is unlikely to explain genetic paradox of biological invasions and invasive populations rather show tendency to accumulate deleterious mutations.

Has genetic load accumulated along the expansion wave?

Such accumulation of the load can speed up when population expand its range (Peischl & Excoffier, 2015). Theoretical expectations and previous reports on accumulation of genetic load at front of population expansion into an empty niche demonstrated that the load, composed mostly of deleterious homozygotes, is greater the further from the expansion source (Gilbert et al., 2018; Henn et al., 2016; Peischl & Excoffier, 2015; Rougemont et al., 2023). This can be associated with secondary founder effects at the advancing population edges leading to successive bottlenecks along the expansion (Kaňuch et al., 2021; Sherpa & Després, 2021). Such pattern has been observed in the Asian honey bee, where range edge colonies had lower genetic diversity and lower brood viability than colonies in the rage center (Hagan et al., 2024). We did not observe such a pattern in Turure river (Figure 5E,F, Table 1). Despite the fact that Turure shows the lowest genetic diversity among studied populations (Figure 4), the diversity does not decline along expansion front. Instead, we observe that the most downstream population is characterized by the highest value of neutral nucleotide diversity in the river. Similarly, the pattern of relative genetic load is opposite to expectations, with significantly lower values at the downstream populations. Likely, this can be explained by admixture patterns

(Figure 3). We observed that both Turure middle and Turure lower populations show low frequency admixture of genes with different origin. As noted above, however, this admixture might have not originated from the same sources (Figure 3), several individuals from the Turure middle population did not cluster with any other population and likely have distinct origin (Figure 2, Figure 3). These individuals were removed from subsequent analyses of N_e and genetic load. However, it is worth to mention that genetic load estimated for excluded individuals is relatively high (Supplementary Figure 4) making it unlikely that the decline in genetic load of Turure middle population is due to genome-wide effects of this admixture. Also, fraction of admixed upstream Turure ancestry does not differ much between Turure middle and Turure downstream populations (87.8% vs 77.9% of upstream ancestry in these two populations, respectively, for K = 8; Figure 3), while genetic load between them significantly differs (with middle Turure population having lower load). Overall our result suggest that gene flow can significantly reduce genetic load even when general genetic diversity remains low, but predicting exact effects is difficult. Generally, mixing between invasive and local individuals might have removed negative effects of consecutive bottlenecks decreasing genetic load of expanding populations (Mesgaran et al., 2016; Pierce et al., 2017; Rius & Darling, 2014). Consequently, we conclude that the load in the Turure guppies is mostly shaped by two forces – accumulation due to relaxed selection in the upper site, and then admixture in the lower sites.

Does genetic load correlate with Ne?

Apart from upper Turure, significantly higher genetic load can also be observed in guppies from Tobago populations (while compared to Trinidadian) and those from upstream sites of Oropouche and Quare (while compared to their downstream counterparts). This, again, can be explained mainly by their small effective sizes (but not in Quare) and long term isolations. The major component of the load in all these cases are variants in homozygotes (Table 1) that are the result of increased inbreeding and strong drift, unavoidable in small populations (similar as in: Dussex et al., 2021; Smeds & Ellegren, 2023). However, when the relationship is studied across multiple populations, no statistically significant relationship between N_e and genetic load is found (Figure 6). It demonstrates that the long term effective population size, measured as neutral diversity, does not predict genetic load of a population, at least not in relatively large populations of no conservation concern, like ones presented here. Instead, more nuanced information about population history, or trajectory of effective population size over time (Nadachowska ‐ Brzyska et al., 2022) should be used to infer genetic health of a population. These results are in line with inter-species investigations showing no correlation between N_e and genetic condition (van Hooft et al., 2021; Wooldridge et al., 2024). However, as there is lack of general framework to compare genetic load between species (Dussex et al., 2023; Teixeira & Huber, 2021), our intra-species investigations are useful to future confirm such conclusion. Teixeira and Huber (2021) discuss that neutral diversity generally fails to explain load and should not be used as a proxy in conservation biology. This opinion was later criticized (Kardos et al., 2021) and since then, vivid discussion regarding role of neutral diversity and genetic load for conservation policies has been held (Kardos, 2023; van Oosterhout et al., 2022).

In our data, the effective population size reflects the census population size pattern in case of comparison between the islands, with small, isolated Tobago populations having visibly lower N_e while compared to Trinidadian populations (Barson & van Oosterhout, 2009, Figure 4). However, the previously reported (Barson & van Oosterhout, 2009; Fraser et al., 2015) higher N_e in the densely populated, lower sites have no support in our observations – both Oropouche river populations are almost exactly the same and Quare river sites show pattern opposite to the expected one (small upper population has higher N_e). In case of the Oropouche populations, this can be explained by previously mentioned lack of geographical barriers. The upper Quare population's N_e is much higher than its lower counterpart and, surprisingly, the highest among all studied populations. Similar results were obtained by (Qiu et al., 2022), who observed evidence for bottlenecks in all but Quare upstream populations. This could likely be explained by its mixed ancestries, resulting from mixture between Caroni and Oropouche or from Oropouche and Madamas (river in the top north part of Trinidad) during wet season as suggested by Whitting et al. (2021). Our data also show a pattern suggesting admixture in this population, even we did not analyze Madamas population (Figure 3). The effective population size in transplanted Turure river population is the lowest among all populations, which is in agreement to results obtained in other translocation studies (for example Aripo introduction experiment, Fraser et al., 2015) and shows after effects of strong bottleneck occurring during the introduction (founder population was about 200 individuals and effective number of founders was probably much lower). The lower Turure river has slightly higher N_e than the upper population, reflecting its admixture with other Oropouche drainage guppies. Overall, our data show that while N_e does have a significant effect on genetic load, the effect may often be obscured by inter-population (and inter-species) differences in population history of admixture.

To summarize, during our investigation, we did not find evidence for purging in introduced guppies population, nor for accumulation of genetic load during their expansion. However, our results suggest that admixture from outside sources on the expansion front may help to select the deleterious variation out and reduce genetic load. We also find that neutral genetic diversity is a rather poor predictor of genetic load across populations of moderate to large size, for which more nuanced information about population demographic history and admixture patters are need to be taken into account.

5. *References*

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664. https://doi.org/10.1101/gr.094052.109

Baker, H. G., & Stebbins, G. L. (1966). *The genetics of colonizing species*. Academic Press, New York & London.

Barker, B. S., Cocio, J. E., Anderson, S. R., Braasch, J. E., Cang, F. A., Gillette, H. D., & Dlugosch, K. M. (2019). Potential limits to the benefits of admixture during biological invasion. *Molecular Ecology*, *28*(1), 100–113. https://doi.org/10.1111/mec.14958

Barrett, R. D. H., & Schluter, D. (2008). Adaptation from standing genetic variation. *Trends in Ecology & Evolution*, *23*(1), 38–44. https://doi.org/10.1016/j.tree.2007.09.008

Barson, N. J., Cable, J., & Van Oosterhout, C. (2009). Population genetic analysis of microsatellite variation of guppies (Poecilia reticulata) in Trinidad and Tobago: evidence for a dynamic source–sink metapopulation structure, founder events and population bottlenecks. *Journal of Evolutionary Biology*, *22*(3), 485–497. https://doi.org/10.1111/j.1420-9101.2008.01675.x

Becher, S. A., & Magurran, A. E. (2000). Gene flow in Trinidadian guppies. *Journal of Fish Biology*, *56*(2), 241–249. https://doi.org/10.1111/j.1095-8649.2000.tb02103.x

Behr, A. A., Liu, K. Z., Liu-Fang, G., Nakka, P., & Ramachandran, S. (2016). pong: fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics*, *32*(18), 2817–2823. https://doi.org/10.1093/bioinformatics/btw327

Bertorelle, G., Raffini, F., Bosse, M., Bortoluzzi, C., Iannucci, A., Trucchi, E., Morales, H. E., & van Oosterhout, C. (2022). Genetic load: genomic estimates and applications in non-model animals. *Nature Reviews Genetics*, *23*(8), 492–503. https://doi.org/10.1038/s41576-022-00448-x

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J.-S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution*, *24*(3), 127–135. https://doi.org/10.1016/j.tree.2008.10.008

Bosshard, L., Dupanloup, I., Tenaillon, O., Bruggmann, R., Ackermann, M., Peischl, S., & Excoffier, L. (2017). Accumulation of Deleterious Mutations During. *Genetics*, *207*(October), 669–684. https://doi.org/10.1534/genetics.117.300144/-/DC1.1

Burda, K., & Konczal, M. (2023). Validation of machine learning approach for direct mutation rate estimation. *Molecular Ecology Resources*, *23*(8), 1757–1771. https://doi.org/10.1111/1755-0998.13841

Charlesworth, D., & Willis, J. H. (2009). The genetics of inbreeding depression. *Nature Reviews Genetics*, *10*(11), 783–796. https://doi.org/10.1038/nrg2664

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*, *6*(2), 80–92. https://doi.org/10.4161/fly.19695

Cooper, G. M., Stone, E. A., Asimenos, G., Green, E. D., Batzoglou, S., & Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. *Genome Research*, *15*(7), 901–913. https://doi.org/10.1101/gr.3577405

Crow, J. F. (1958). Some possibilities for measuring selection intensities in man. *Human Biology*, *61*(5/6), 763–775. https://doi.org/www.jstor.org/stable/41478722

Daly, E. Z., Gerlich, H. S., Frenot, Y., Høye, T. T., Holmstrup, M., & Renault, D. (2023). Climate Change Helps Polar Invasives Establish and Flourish: Evidence from Long-Term Monitoring of the Blowfly Calliphora vicina. *Biology*, *12*(1), 111. https://doi.org/10.3390/biology12010111

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2). https://doi.org/10.1093/gigascience/giab008

Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., & Batzoglou, S. (2010). Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. *PLoS Computational Biology*, *6*(12), e1001025. https://doi.org/10.1371/journal.pcbi.1001025

de Pedro, M., Mayol, M., González-Martínez, S. C., Regalado, I., & Riba, M. (2022). Environmental patterns of adaptation after range expansion in Leontodon longirostris: The effect of phenological events on fitness-related traits. *American Journal of Botany*, *109*(4), 602–615. https://doi.org/10.1002/ajb2.1815

de Pedro, M., Riba, M., González-Martínez, S. C., Seoane, P., Bautista, R., Claros, M. G., & Mayol, M. (2021). Demography, genetic diversity and expansion load in the colonizing species Leontodon longirostris (Asteraceae) throughout its native range. *Molecular Ecology*, *30*(5), 1190–1205. https://doi.org/10.1111/mec.15802

Deacon, A. E., Ramnarine, I. W., & Magurran, A. E. (2011). How Reproductive Ecology Contributes to the Spread of a Globally Invasive Fish. *PLoS ONE*, *6*(9), e24416. https://doi.org/10.1371/journal.pone.0024416

Dehasque, M., Morales, H. E., Díez-del-Molino, D., Pečnerová, P., Chacón-Duque, J. C., Kanellidou, F., Muller, H., Plotnikov, V., Protopopov, A., Tikhonov, A., Nikolskiy, P., Danilov, G. K., Giannì, M., van der Sluis, L., Higham, T., Heintzman, P. D., Oskolkov, N., Gilbert, M. T. P., Götherström, A., … Dalén, L. (2024). Temporal dynamics of woolly mammoth genome erosion prior to extinction. *Cell*, *187*(14), 3531-3540.e13. https://doi.org/10.1016/j.cell.2024.05.033

Devigili, A., Fitzpatrick, J. L., Gasparini, C., Ramnarine, I. W., Pilastro, A., & Evans, J. P. (2018). Possible glimpses into early speciation: the effect of ovarian fluid on sperm velocity accords with post-copulatory isolation between two guppy populations. *Journal of Evolutionary Biology, 31*(1), 66–74. https://doi.org/10.1111/jeb.13194

Dussex, N., Morales, H. E., Grossen, C., Dalén, L., & van Oosterhout, C. (2023). Purging and accumulation of genetic load in conservation. *Trends in Ecology & Evolution*, *38*(10), 961–969. https://doi.org/10.1016/j.tree.2023.05.008

Dussex, N., van der Valk, T., Morales, H. E., Wheat, C. W., Díez-del-Molino, D., von Seth, J., Foster, Y., Kutschera, V. E., Guschanski, K., Rhie, A., Phillippy, A. M., Korlach, J., Howe, K., Chow, W., Pelan, S., Mendes Damas, J. D., Lewin, H. A., Hastie, A. R., Formenti, G., … Dalén, L. (2021). Population genomics of the critically endangered kākāpō. *Cell Genomics*, *1*(1), 100002. https://doi.org/10.1016/j.xgen.2021.100002

Endler, J. A. (1980). Natural Selection on Color Patterns in Poecilia Reticulata. *Evolution*, *34*(1), 76–91. https://doi.org/10.1111/j.1558-5646.1980.tb04790.x

Escher, L. M., Naslavsky, M. S., Scliar, M. O., Duarte, Y. A. O., Zatz, M., Nunes, K., & Oliveira, S. F. (2022). Challenges in selecting admixture models and marker sets to infer genetic ancestry in a Brazilian admixed population. *Scientific Reports*, *12*(1), 21240. https://doi.org/10.1038/s41598-022-25521-7

Estoup, A., Ravigné, V., Hufbauer, R., Vitalis, R., Gautier, M., & Facon, B. (2016). Is There a Genetic Paradox of Biological Invasion? *Annual Review of Ecology, Evolution, and Systematics*, *47*(1), 51–72. https://doi.org/10.1146/annurev-ecolsys-121415-032116

Eyre-Walker, A., & Keightley, P. D. (2007). The distribution of fitness effects of new mutations. *Nature Reviews Genetics*, *8*(8), 610–618. https://doi.org/10.1038/nrg2146

Facon, B., Crespin, L., Loiseau, A., Lombaert, E., Magro, A., & Estoup, A. (2011). Can things get worse when an invasive species hybridizes? The harlequin ladybird *Harmonia axyridis* in France as a case study. *Evolutionary Applications*, *4*(1), 71–88. https://doi.org/10.1111/j.1752-4571.2010.00134.x

Fajen, A., & Breden, F. (1992). Mitochondrial DNA sequence variation among natural populations of the Trinidadian guppy, *Poecilia reticulata*. *Evolution*, *46*(5), 1457–1465. https://doi.org/10.1111/j.1558-5646.1992.tb01136.x

Faulkner, K. T., Hulme, P. E., & Wilson, J. R. U. (2024). Harder, better, faster, stronger? Dispersal in the Anthropocene. *Trends in Ecology & Evolution*. https://doi.org/10.1016/j.tree.2024.08.010

Fitzpatrick, S. W., Gerberich, J. C., Angeloni, L. M., Bailey, L. L., Broder, E. D., Torres-Dowdall, J., Handelsman, C. A., López-Sepulcre, A., Reznick, D. N., Ghalambor, C. K., & Chris Funk, W. (2016). Gene flow from an adaptively divergent source causes rescue through genetic and demographic factors in two wild populations of Trinidadian guppies. *Evolutionary Applications*, *9*(7), 879–891. https://doi.org/10.1111/eva.12356

Fountain, T., Duvaux, L., Horsburgh, G., Reinhardt, K., & Butlin, R. K. (2014). Human-facilitated metapopulation dynamics in an emerging pest species, *<scp>C</scp> imex lectularius . Molecular Ecology*, *23*(5), 1071–1084. https://doi.org/10.1111/mec.12673

Frankham, R. (1995). Conservation Genetics. *Annual Review of Genetics*, *29*(1), 305–327. https://doi.org/10.1146/annurev.ge.29.120195.001513

Fraser, B. A., Künstner, A., Reznick, D. N., Dreyer, C., & Weigel, D. (2015). Population genomics of natural and experimental populations of guppies (Poecilia reticulata). *Molecular Ecology*, *24*(2), 389–408. https://doi.org/10.1111/mec.13022

Gilbert, K. J., Peischl, S., & Excoffier, L. (2018). Mutation load dynamics during environmentally-driven range shifts. *PLoS Genetics*, *14*(9), 1–18. https://doi.org/10.1371/journal.pgen.1007450

Gilbert, K. J., Sharp, N. P., Angert, A. L., Conte, G. L., Draghi, J. A., Guillaume, F., Hargreaves, A. L., Matthey-Doret, R., & Whitlock, M. C. (2017). Local Adaptation Interacts with Expansion Load during Range Expansion: Maladaptation Reduces Expansion Load. *The American Naturalist*, *189*(4), 368–380. https://doi.org/10.1086/690673

Grossen, C., Guillaume, F., Keller, L. F., & Croll, D. (2020). Purging of highly deleterious mutations through severe bottlenecks in Alpine ibex. *Nature Communications*, *11*(1), 1001. https://doi.org/10.1038/s41467-020-14803-1

Gu, Z., David, L., Petrov, D., Jones, T., Davis, R. W., & Steinmetz, L. M. (2005). Elevated evolutionary rates in the laboratory strain of *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences*, *102*(4), 1092–1097. https://doi.org/10.1073/pnas.0409159102

Hagan, T., Ding, G., Buchmann, G., Oldroyd, B. P., & Gloag, R. (2024). Serial founder effects slow range expansion in an invasive social insect. *Nature Communications*, *15*(1), 3608. https://doi.org/10.1038/s41467-024-47894-1

Henn, B. M., Botigué, L. R., Peischl, S., Dupanloup, I., Lipatov, M., Maples, B. K., Martin, A. R., Musharoff, S., Cann, H., Snyder, M. P., Excoffier, L., Kidd, J. M., & Bustamante, C. D. (2016). Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. *Proceedings of the National Academy of Sciences*, *113*(4). https://doi.org/10.1073/pnas.1510805112

Herdegen-Radwan, M., Phillips, K. P., Babik, W., Mohammed, R. S., & Radwan, J. (2021). Balancing selection versus allele and supertype turnover in MHC class II genes in guppies. *Heredity*, *126*(3), 548–560. https://doi.org/10.1038/s41437-020-00369-7

Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella, A. J., Searle, S. M. J., Amode, R., Brent, S., Spooner, W., Kulesha, E., Yates, A., & Flicek, P. (2016). Ensembl comparative genomics resources. *Database*, *2016*, bav096. https://doi.org/10.1093/database/bav096

Johnson, C. N., Balmford, A., Brook, B. W., Buettel, J. C., Galetti, M., Guangchun, L., & Wilmshurst, J. M. (2017). Biodiversity losses and conservation responses in the Anthropocene. *Science*, *356*(6335), 270–275. https://doi.org/10.1126/science.aam9317

Kaňuch, P., Berggren, Å., & Cassel-Lundhagen, A. (2021). A clue to invasion success: genetic diversity quickly rebounds after introduction bottlenecks. *Biological Invasions*, *23*(4), 1141–1156. https://doi.org/10.1007/s10530-020-02426-y

Kardos, M. (2023). Genomes of an endangered rattlesnake show that neutral genetic variation predicts adaptive genetic variation and genetic load. *Proceedings of the National Academy of Sciences*, *120*(49). https://doi.org/10.1073/pnas.2316880120

Kardos, M., Armstrong, E. E., Fitzpatrick, S. W., Hauser, S., Hedrick, P. W., Miller, J. M., Tallmon, D. A., & Funk, W. C. (2021). The crucial role of genome-wide genetic variation in conservation. *Proceedings of the National Academy of Sciences*, *118*(48). https://doi.org/10.1073/pnas.2104642118

Kimura, M., Maruyama, T., & Crow, J. F. (1963). The Mutation Load in Small Populations. *Genetics*, *48*(10), 1303–1312. https://doi.org/10.1093/genetics/48.10.1303

Kimura, M., & Ohta, T. (1969). The Average Number of Generations Until Fixation of a Mutant Gene in a Finite Population. *Genetics*, *61*(3), 763–771. https://doi.org/10.1093/genetics/61.3.763

Kleinman-Ruiz, D., Lucena-Perez, M., Villanueva, B., Fernández, J., Saveljev, A. P., Ratkiewicz, M., Schmidt, K., Galtier, N., García-Dorado, A., & Godoy, J. A. (2022). Purging of deleterious burden in the endangered Iberian lynx. *Proceedings of the National Academy of Sciences*, *119*(11). https://doi.org/10.1073/pnas.2110614119

Korunes, K. L., & Samuk, K. (2021). <scp>pixy</scp> : Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Molecular Ecology Resources*, *21*(4), 1359–1368. https://doi.org/10.1111/1755-0998.13326

Kuang, W., Hu, J., Wu, H., Fen, X., Dai, Q., Fu, Q., Xiao, W., Frantz, L., Roos, C., Nadler, T., Irwin, D. M., Zhou, L., Yang, X., & Yu, L. (2020). Genetic Diversity, Inbreeding Level, and Genetic Load in Endangered Snub-Nosed Monkeys (Rhinopithecus). *Frontiers in Genetics*, *11*. https://doi.org/10.3389/fgene.2020.615926

Künstner, A., Hoffmann, M., Fraser, B. A., Kottler, V. A., Sharma, E., Weigel, D., & Dreyer, C. (2016). The Genome of the Trinidadian Guppy, Poecilia reticulata, and Variation in the Guanapo Population. *PLOS ONE*, *11*(12), e0169087. https://doi.org/10.1371/journal.pone.0169087

Kyriazis, C. C., Wayne, R. K., & Lohmueller, K. E. (2021). Strongly deleterious mutations are a primary determinant of extinction risk due to inbreeding depression. *Evolution Letters*, *5*(1), 33–47. https://doi.org/10.1002/evl3.209

Laugier, G. J. M., Le Moguédec, G., Su, W., Tayeh, A., Soldati, L., Serrate, B., Estoup, A., & Facon, B. (2016). Reduced population size can induce quick evolution of inbreeding depression in the invasive ladybird Harmonia axyridis. *Biological Invasions*, *18*(10), 2871–2881. https://doi.org/10.1007/s10530-016-1179-1

Li, H. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. *00*(00), 1–3. http://arxiv.org/abs/1303.3997

Lindholm, A. K., Breden, F., Alexander, H. J., Chan, W., Thakurta, S. G., & Brooks, R. (2005). Invasion success and genetic diversity of introduced populations of guppies Poecilia reticulata in Australia. *Molecular Ecology*, *14*(12), 3671–3682. https://doi.org/10.1111/j.1365-294X.2005.02697.x

Liu, L., Bosse, M., Megens, H., de Visser, M., A. M. Groenen, M., & Madsen, O. (2021). Genetic consequences of long-term small effective population size in the critically endangered pygmy hog. *Evolutionary Applications*, *14*(3), 710–720. https://doi.org/10.1111/eva.13150

Lombaert, E., Blin, A., Porro, B., Guillemaud, T., Bernal, J. S., Chang, G., Kirichenko, N., Sappington, T. W., Toepfer, S., & Deleury, E. (2024). Unraveling genetic load dynamics during biological invasion: insights from two invasive insect species. *BioRxiv*.

Lynch, M. (1991). The genetic interpretation of inbreeding depression and outbreeding depression. *Evolution*, *45*(3), 622–629. https://doi.org/10.1111/j.1558-5646.1991.tb04333.x

Lynch, M., Conery, J., & Bürger, R. (1995). Mutational Meltdowns in Sexual Populations. *Evolution*, *49*(6), 1067–1080. https://doi.org/10.1111/j.1558-5646.1995.tb04434.x

Magurran, A. E. (1996). Battle of the sexes. *Nature*, *383*(6598), 307–307. https://doi.org/10.1038/383307a0

Magurran, A. E. (2005). *Evolutionary Ecology*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198527855.001.0001

Marchini, G. L., Sherlock, N. C., Ramakrishnan, A. P., Rosenthal, D. M., & Cruzan, M. B. (2016). Rapid purging of genetic load in a metapopulation and consequences for range expansion in an invasive plant. *Biological Invasions*, *18*(1), 183–196. https://doi.org/10.1007/s10530-015-1001-5

Mathur, S., & DeWoody, J. A. (2021). Genetic load has potential in large populations but is realized in small inbred populations. *Evolutionary Applications*, *14*(6), 1540–1557. https://doi.org/10.1111/eva.13216

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Mesgaran, M. B., Lewis, M. A., Ades, P. K., Donohue, K., Ohadi, S., Li, C., & Cousens, R. D. (2016). Hybridization can facilitate species invasions, even without enhancing local adaptation. *Proceedings of the National Academy of Sciences*, *113*(36), 10210–10214. https://doi.org/10.1073/pnas.1605626113

Mullarkey, A. A., Byers, D. L., & Anderson, R. C. (2013). Inbreeding depression and partitioning of genetic load in the invasive biennial *Alliaria petiolata* (Brassicaceae). *American Journal of Botany*, *100*(3), 509–518. https://doi.org/10.3732/ajb.1200403

Nadachowska-Brzyska, K., Konczal, M., & Babik, W. (2022). Navigating the temporal continuum of effective population size. *Methods in Ecology and Evolution*, *13*(1), 22–41. https://doi.org/10.1111/2041-210X.13740

Ochoa, A., Onorato, D. P., Roelke-Parker, M. E., Culver, M., & Fitak, R. R. (2022). Give and take: Effects of genetic admixture on mutation load in endangered Florida panthers. *Journal of Heredity*, *113*(5), 491–499. https://doi.org/10.1093/jhered/esac037

Ohta, T. (1973). Slightly Deleterious Mutant Substitutions in Evolution. *Nature*, *246*(5428), 96–98. https://doi.org/10.1038/246096a0

Peischl, S., & Excoffier, L. (2015). Expansion load: Recessive mutations and the role of standing genetic variation. *Molecular Ecology*, *24*(9), 2084–2094. https://doi.org/10.1111/mec.13154

Peischl, S., Kirkpatrick, M., & Excoffier, L. (2015). Expansion Load and the Evolutionary Dynamics of a Species Range. *The American Naturalist*, *185*(4), E81–E93. https://doi.org/10.1086/680220

Pfennig, K. S., Kelly, A. L., & Pierce, A. A. (2016). Hybridization as a facilitator of species range expansion. In *Proceedings. Biological sciences* (Vol. 283, Issue 1839). https://doi.org/10.1098/rspb.2016.1329

Pierce, A. A., Gutierrez, R., Rice, A. M., & Pfennig, K. S. (2017). Genetic variation during range expansion: effects of habitat novelty and hybridization. *Proceedings of the Royal Society B: Biological Sciences*, *284*(1852), 20170007. https://doi.org/10.1098/rspb.2017.0007

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Qiu, S., Yong, L., Wilson, A., Croft, D. P., Graham, C., & Charlesworth, D. (2022). Partial sex linkage and linkage disequilibrium on the guppy sex chromosome. *Molecular Ecology*, *31*(21), 5524–5537. https://doi.org/10.1111/mec.16674

Ralls, K., Sunnucks, P., Lacy, R. C., & Frankham, R. (2020). Genetic rescue: A critique of the evidence supports maximizing genetic diversity rather than minimizing the introduction of putatively harmful genetic variation. *Biological Conservation*, *251*, 108784. https://doi.org/10.1016/j.biocon.2020.108784

Reznick, D. N., & Bryga, H. (1987). Life-history Evolution in Guppies (Poecilia reticulata): 1. Phenotypic and genetic changes in an introduction experiment. *Evolution*, *41*(6), 1370–1385. https://doi.org/10.1111/j.1558-5646.1987.tb02474.x

Rius, M., & Darling, J. A. (2014). How important is intraspecific genetic admixture to the success of colonising populations? *Trends in Ecology & Evolution*, *29*(4), 233–242. https://doi.org/10.1016/j.tree.2014.02.003

Robinson, J. A., Ortega-Del Vecchyo, D., Fan, Z., Kim, B. Y., vonHoldt, B. M., Marsden, C. D., Lohmueller, K. E., & Wayne, R. K. (2016). Genomic Flatlining in the Endangered Island Fox. *Current Biology*, *26*(9), 1183–1189. https://doi.org/10.1016/j.cub.2016.02.062

Robinson, J., Kyriazis, C. C., Yuan, S. C., & Lohmueller, K. E. (2023). Deleterious Variation in Natural Populations and Implications for Conservation Genetics. *Annual Review of Animal Biosciences*, *11*(1), 93–114. https://doi.org/10.1146/annurev-animal-080522-093311

Roman, J., & Darling, J. (2007). Paradox lost: genetic diversity and the success of aquatic invasions. *Trends in Ecology & Evolution*, *22*(9), 454–464. https://doi.org/10.1016/j.tree.2007.07.002

Rosenthal, W. C., McIntyre, P. B., Lisi, P. J., Prather, R. B., Moody, K. N., Blum, M. J., Hogan, J. D., & Schoville, S. D. (2021). Invasion and rapid adaptation of guppies (Poecilia reticulata) across the Hawaiian Archipelago. *Evolutionary Applications*, *14*(7), 1747–1761. https://doi.org/10.1111/eva.13236

Rougemont, Q., Leroy, T., Rondeau, E. B., Koop, B., & Bernatchez, L. (2023). Allele surfing causes maladaptation in a Pacific salmon of conservation concern. *PLoS Genetics*, *19*(9 September), 1–28. https://doi.org/10.1371/journal.pgen.1010918

Rougemont, Q., Moore, J. S., Leroy, T., Normandeau, E., Rondeau, E. B., Withler, R. E., van Doornik, D. M., Crane, P. A., Naish, K. A., Garza, J. C., Beacham, T. D., Koop, B. F., & Bernatchez, L. (2020). Demographic history shaped geographical patterns of deleterious mutation load in a broadly distributed Pacific Salmon. *PLoS Genetics*, *16*(8), 1–29. https://doi.org/10.1371/JOURNAL.PGEN.1008348

Russell, S. T., & Magurran, A. E. (2006). Intrinsic reproductive isolation between Trinidadian populations of the guppy, *Poecilia reticulata*. *Journal of Evolutionary Biology*, *19*(4), 1294–1303. https://doi.org/10.1111/j.1420-9101.2005.01069.x

Santana Marques, P., Resende Manna, L., Clara Frauendorf, T., Zandonà, E., Mazzoni, R., & El-Sabaawi, R. (2020). Urbanization can increase the invasive potential of alien species. *Journal of Animal Ecology*, *89*(10), 2345–2355. https://doi.org/10.1111/1365-2656.13293

Schmidt, C., Hoban, S., Hunter, M., Paz-Vinas, I., & Garroway, C. J. (2023). Genetic diversity and IUCN Red List status. *Conservation Biology*, *37*(4). https://doi.org/10.1111/cobi.14064

Schories, S., Meyer, M. K., & Schartl, M. (2009). Description of poecilia (acanthophacelus) obscura n. sp., (teleostei: Poeciliidae), a new guppy species from western trinidad, with remarks on p. wingei and the status of the “endler’s guppy.” *Zootaxa*, *50*(2266), 35–50. https://doi.org/10.11646/zootaxa.2266.1.2

Shaw, P. W.; Carvalho, G. R.; Seghers, B. H.; Magurran, A. E. (1992). Genetic consequences of an artificial introduction of guppies (Poecilia reticulata) in N. Trinidad. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *248*(1322), 111–116. https://doi.org/10.1098/rspb.1992.0049

Sherpa, S., & Després, L. (2021). The evolutionary dynamics of biological invasions: A multi-approach perspective. In *Evolutionary Applications* (Vol. 14, Issue 6, pp. 1463–1484). John Wiley and Sons Inc. https://doi.org/10.1111/eva.13215

Sievers, C., Ramnarine, I. W., & Magurran, A. E. (2014). The influence of population mixing on newborn shoaling behaviour in the guppy (Poecilia reticulata). *Behaviour*, *151*(10), 1479–1490. https://doi.org/10.1163/1568539X-00003196

Sievers, C., Willing, E. M., Hoffmann, M., Dreyer, C., Ramnarine, I., & Magurran, A. (2012). Reasons for the invasive success of a guppy (Poecilia reticulata) population in Trinidad. *PLoS ONE*, *7*(5). https://doi.org/10.1371/journal.pone.0038404

Smeds, L., & Ellegren, H. (2023). From high masked to high realized genetic load in inbred Scandinavian wolves. *Molecular Ecology*, *32*(7), 1567–1580. https://doi.org/10.1111/mec.16802

Stewart, G. S., Morris, M. R., Genis, A. B., Szűcs, M., Melbourne, B. A., Tavener, S. J., & Hufbauer, R. A. (2017). The power of evolutionary rescue is constrained by genetic load. *Evolutionary Applications*, *10*(7), 731–741. https://doi.org/10.1111/eva.12489

Suk, H. Y., & Neff, B. D. (2009). Microsatellite genetic differentiation among populations of the Trinidadian guppy. *Heredity*, *102*(5), 425–434. https://doi.org/10.1038/hdy.2009.7

Tayeh, A., Estoup, A., Hufbauer, R. A., Ravigne, V., Goryacheva, I., Zakharov, I. A., Lombaert, E., & Facon, B. (2013). Investigating the genetic load of an emblematic invasive species: the case of the invasive harlequin ladybird *Harmonia axyridis*. *Ecology and Evolution*, *3*(4), 864–871. https://doi.org/10.1002/ece3.490

Teixeira, J. C., & Huber, C. D. (2021). The inflated significance of neutral genetic diversity in conservation genetics. *Proceedings of the National Academy of Sciences*, *118*(10). https://doi.org/10.1073/pnas.2015096118

Todesco, M., Pascual, M. A., Owens, G. L., Ostevik, K. L., Moyers, B. T., Hübner, S., Heredia, S. M., Hahn, M. A., Caseys, C., Bock, D. G., & Rieseberg, L. H. (2016). Hybridization and extinction. *Evolutionary Applications*, *9*(7), 892–908. https://doi.org/10.1111/eva.12367

van Hooft, P., Getz, W. M., Greyling, B. J., Zwaan, B., & Bastos, A. D. S. (2021). A continent-wide high genetic load in African buffalo revealed by clines in the frequency of deleterious alleles, genetic hitchhiking and linkage disequilibrium. *PLOS ONE*, *16*(12), e0259685. https://doi.org/10.1371/journal.pone.0259685

van Oosterhout, C., Speak, S. A., Birley, T., Bortoluzzi, C., Percival-Alwyn, L., Urban, L. H., Groombridge, J. J., Segelbacher, G., & Morales, H. E. (2022). Genomic erosion in the assessment of species extinction risk and recovery potential. *BioRxiv*.

Wallingford, P. D., Morelli, T. L., Allen, J. M., Beaury, E. M., Blumenthal, D. M., Bradley, B. A., Dukes, J. S., Early, R., Fusco, E. J., Goldberg, D. E., Ibáñez, I., Laginhas, B. B., Vilà, M., & Sorte, C. J. B. (2020). Adjusting the lens of invasion biology to focus on the impacts of climate-driven range shifts. *Nature Climate Change*, *10*(5), 398–405. https://doi.org/10.1038/s41558-020-0768-2

Whiting, J. R., Paris, J. R., van der Zee, M. J., Parsons, P. J., Weigel, D., & Fraser, B. A. (2021). Drainage-structuring of ancestral variation and a common functional pathway shape limited genomic convergence in natural high- and low-predation guppies. *PLOS Genetics*, *17*(5), e1009566. https://doi.org/10.1371/journal.pgen.1009566

Wilder, A. P., Supple, M. A., Subramanian, A., Mudide, A., Swofford, R., Serres-Armero, A., Steiner, C., Koepfli, K.-P., Genereux, D. P., Karlsson, E. K., Lindblad-Toh, K., Marques-Bonet, T., Munoz Fuentes, V., Foley, K., Meyer, W. K., Ryder, O. A., Shapiro, B., Andrews, G., Armstrong, J. C., … Zhang, X.

(2023). The contribution of historical processes to contemporary extinction risk in placental mammals. *Science*, *380*(6643). https://doi.org/10.1126/science.abn5856

Willing, E. M., Bentzen, P., Van Oosterhout, C., Hoffmann, M., Cable, J., Breden, F., Weigel, D., & Dreyer, C. (2010). Genome-wide single nucleotide polymorphisms reveal population history and adaptive divergence in wild guppies. *Molecular Ecology*, *19*(5), 968–984. https://doi.org/10.1111/j.1365-294X.2010.04528.x

Wooldridge, B., Orland, C., Enbody, E., Escalona, M., Mirchandani, C., Corbett-Detig, R., Kapp, J. D., Fletcher, N., Cox-Ammann, K., Raimondi, P., & Shapiro, B. (2024). Limited genomic signatures of population collapse in the critically endangered black abalone (*Haliotis cracherodii*). *Molecular Ecology*. https://doi.org/10.1111/mec.17362

Wright, S. (1922). Coefficients of Inbreeding and Relationship. *The American Naturalist*, *56*(645), 330–338.

Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding, and selection in evolution. *Proceedings of the Sixth International Congress of Genetics*. https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=http://www.esp.org/books/6th-congress/facsimile/contents/6th-cong-p356-wright.pdf&ved=2ahUKEwi7k8SPj6yIAxXQB9sEHaMZBwUQFnoECA0QAQ&usg=AOvVaw2yDS8qkFrnqeEFAWqkvYOx

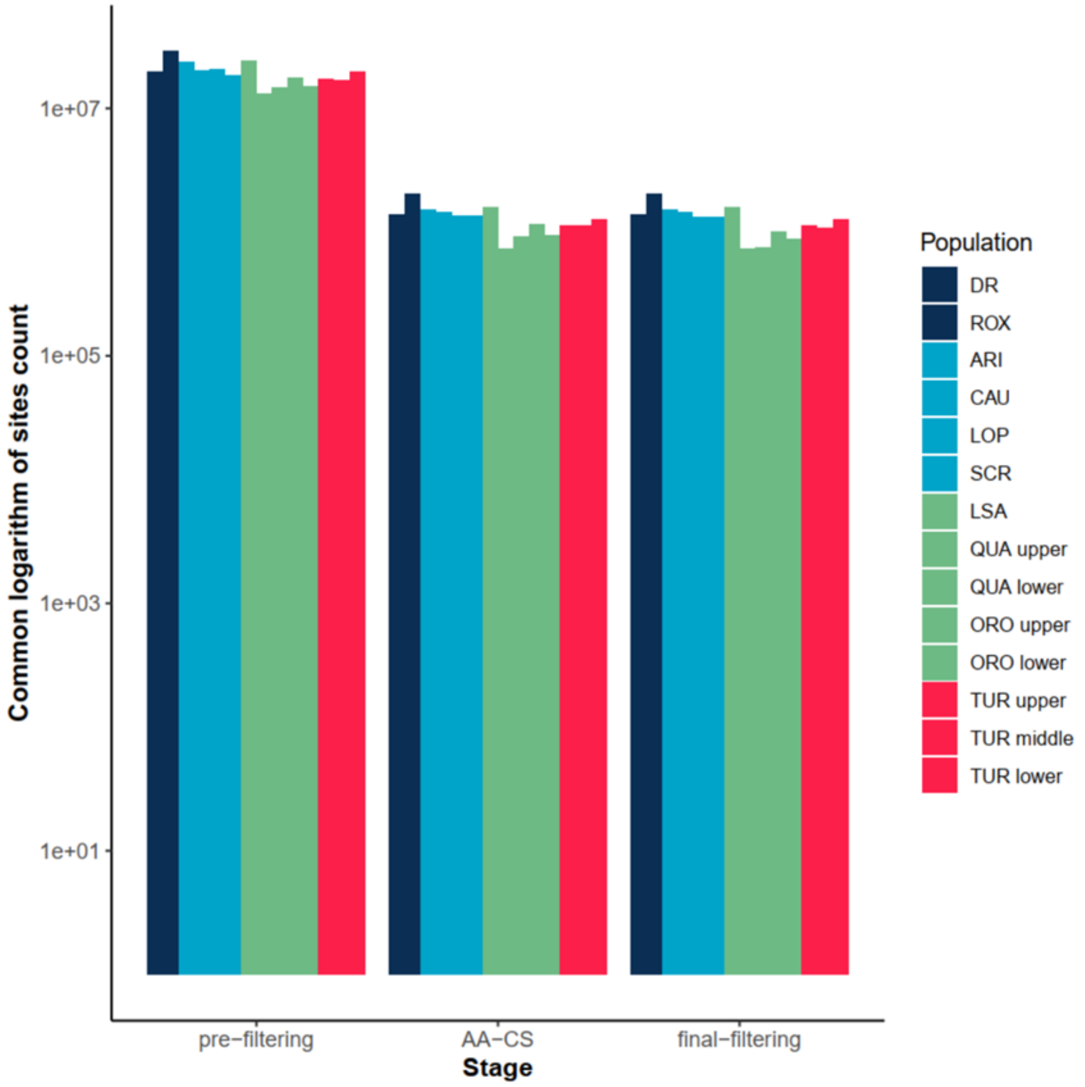
Xie, H.-X., Liang, X.-X., Chen, Z.-Q., Li, W.-M., Mi, C.-R., Li, M., Wu, Z.-J., Zhou, X.-M., & Du, W.-G. (2022). Ancient Demographics Determine the Effectiveness of Genetic Purging in Endangered Lizards. *Molecular Biology and Evolution*, *39*(1). https://doi.org/10.1093/molbev/msab359

Zayed, A., Constantin, Ş. A., & Packer, L. (2007). Successful Biological Invasion despite a Severe Genetic Load. *PLoS ONE*, *2*(9), e868. https://doi.org/10.1371/journal.pone.0000868

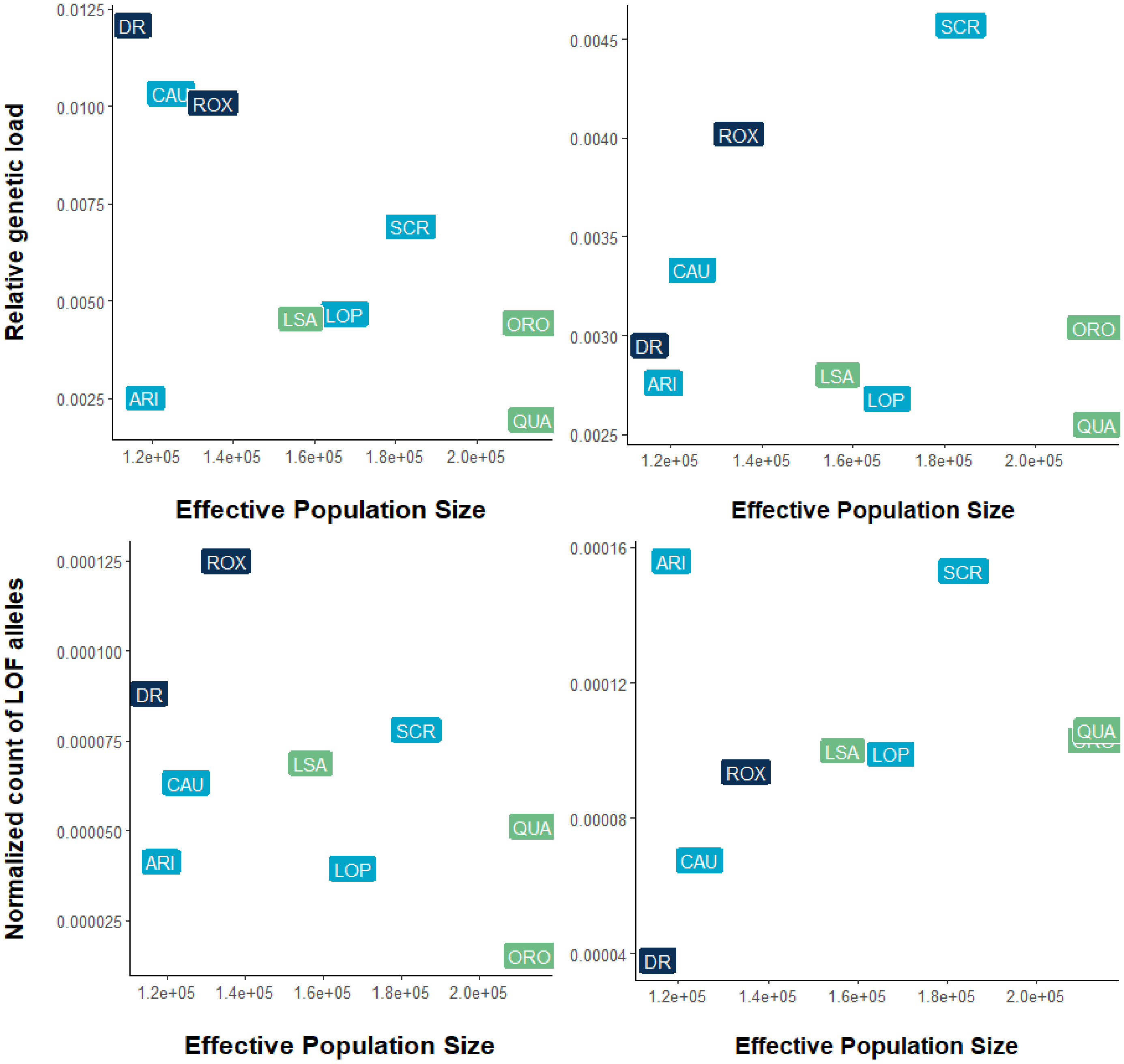
Zeitler, L., Parisod, C., & Gilbert, K. J. (2023). Purging due to self-fertilization does not prevent accumulation of expansion load. *PLoS Genetics*, *19*(9 September), 1–27. https://doi.org/10.1371/journal.pgen.1010883

Zhang, M., Zhou, L., Bawa, R., Suren, H., & Holliday, J. A. (2016). Recombination Rate Variation, Hitchhiking, and Demographic History Shape Deleterious Load in Poplar. *Molecular Biology and Evolution*, *33*(11), 2899–2910. https://doi.org/10.1093/molbev/msw169

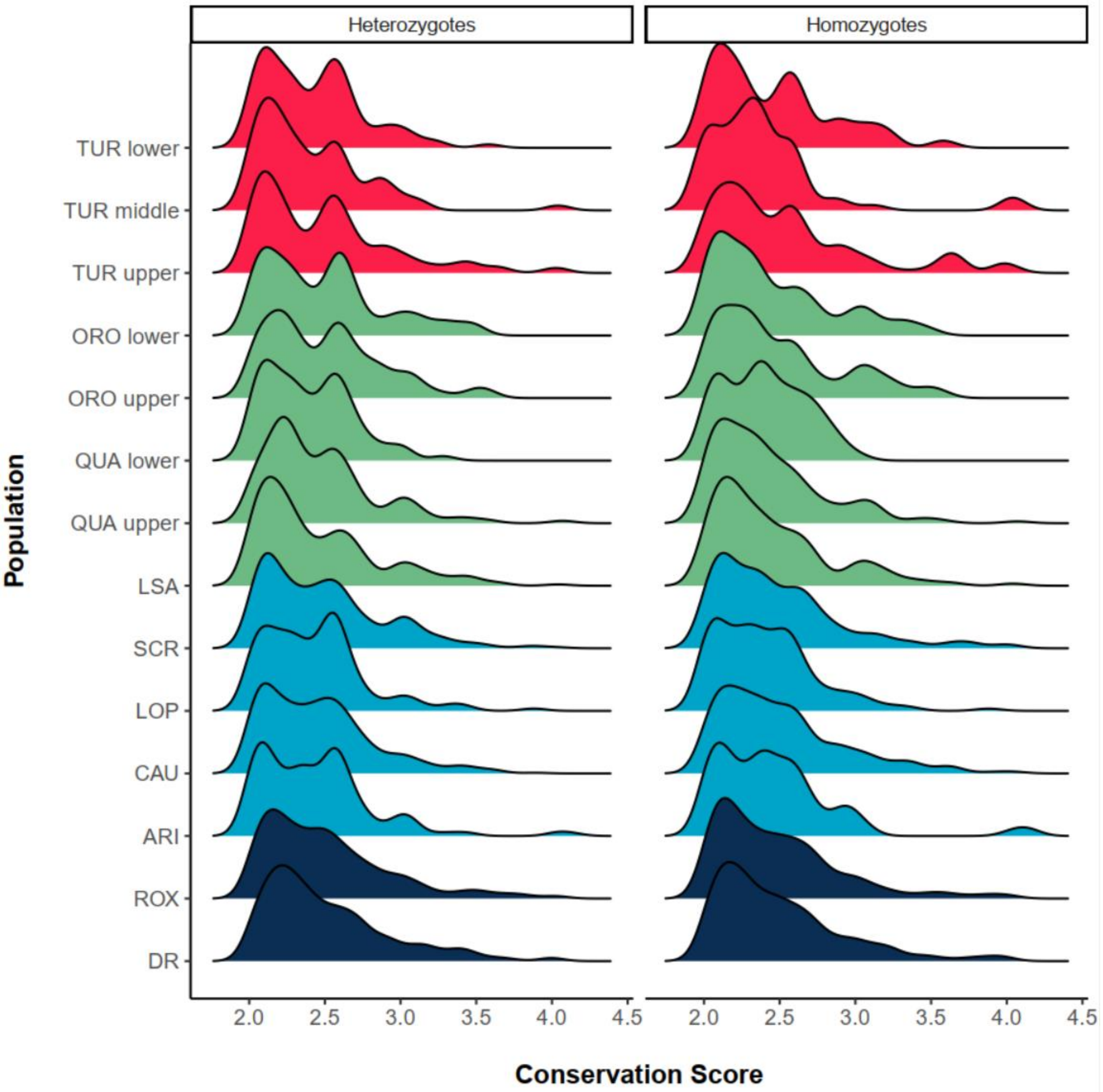
Supplementary Materials



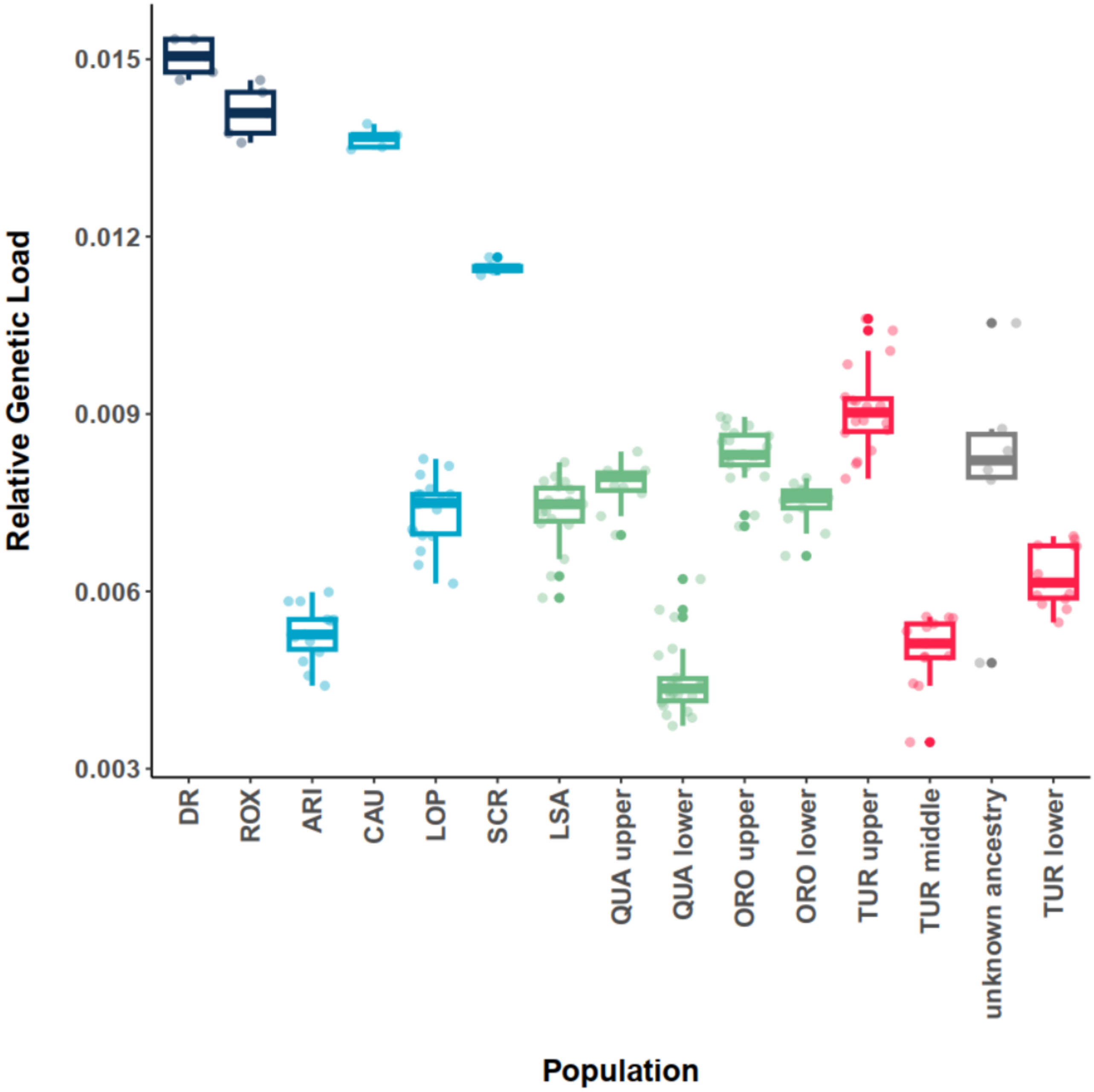
Supplementary Figure 1 Sites count. On the x-axis, filtration steps are presented. The 'pre-filtering' is raw count of sites, 'AA-CS' is initial quality filtering (see Methods), ancestral allele and conservation score intersection. The 'final-filtering' is the count of sites after the last quality refinement. The regions are color-coded as follows: Tobago (dark blue), Caroni (light blue), Oropouche (green), and Turure (red).



Supplementary Figure 2 Relation between effective population size and relative genetic load and between effective populations size and LOF alleles count. The regions are colour-coded as follows: Tobago (dark blue), Caroni (light blue) and Oropuche (green). (A, C) Homozygotes (B, D) Heterozygotes.



Supplementary Figure 3 CS distributions in homozygotes and heterozygotes. Conservation scores are shown on the x-axis. The regions are color-coded as follows: Tobago (dark blue), Caroni (light blue), Oropouche (green), and Turure (red).



Supplementary Figure 4 Total relative genetic load in all populations. The regions are color-coded as follows: Tobago (dark blue), Caroni (light blue), Oropouche (green), and Turure (red).

Supplementary Table 1 Sampling sites on Trinidad and Tobago. The first column is site abbreviation, the 5th, 6th and 7th column is total number of individuals used in this study.

Population	Island	Drainage	Location	N	Males	Females	Juveniles
DR	Tobago	-	N 11.22556 W 060.64500	5	1	4	-
ROX	Tobago	-	N 11.25056 W 060.58250	5	2	3	-
ARI	Trinidad	Caroni	N 10.66969 W 061.22991	14	8	6	-
CAU	Trinidad	Caroni	N 10.66969 W 061.22991	5	2	3	-
LOP	Trinidad	Caroni	N 10.69333 W 061.32195	19	7	12	-
SCR	Trinidad	Caroni	N 10.71565 W 061.46842	5	2	1	2
LSA	Trinidad	Oropouche	N 10.65718 W 061.15276	19	8	11	-
QUA upper	Trinidad	Oropouche	N 10.67610 W 061.19606	11	1	6	4
QUA lower	Trinidad	Oropouche	N 10.65265 W 061.18951	22	7	12	3
ORO upper	Trinidad	Oropouche	N 10.67048 W 061.13784	20	3	3	14
ORO lower	Trinidad	Oropouche	N 10.65963 W 061.13137	13	6	7	-
TUR upper	Trinidad	Oropouche	N 10.67747 W 061.16355	19	8	11	-
TUR middle	Trinidad	Oropouche	N 10.65687 W 061.16776	19	4	15	-
TUR lower	Trinidad	Oropouche	N 10.61899 W 061.15423	14	7	4	3

Supplementary Table 2 Models used in statistical analyses

LOF Alleles Count					
	Total	Homozygotes	Heterozygotes	Model	
Ne	lm(MeanLOF ~ Ne + Region)	lm(MeanHomozygotes ~ Ne + Region)	lm(MeanHeterozygotes ~ Ne + Region)	LM	linear model
Islands	glm(cbind(LOF, DerivedAlleles) ~ Island, family = "binomial")	glm(cbind(Homozygotes, DerivedAlleles) ~ Island, family = quasibinomial())	glm(cbind(Heterozygotes, DerivedAlleles) ~ Island, family = "binomial")	GLM	generalized linear model
Location	glm(cbind(LOF, DerivedAlleles) ~ Site, family = "binomial")	glm(cbind(Homozygotes, DerivedAlleles) ~ Site, family = quasibinomial())	glm(cbind(Heterozygotes, DerivedAlleles) ~ Site, family = "binomial")		
Turure Location	glm(cbind(LOF, DerivedAlleles) ~ Population, family = quasibinomial())	glm(cbind(Homozygotes, DerivedAlleles) ~ Population, family = quasibinomial())	glm(cbind(Heterozygotes, DerivedAlleles) ~ Population, family = quasibinomial())		
Founder Effect	glm(cbind(LOF, DerivedAlleles) ~ Bottleneck, family = quasibinomial())	glm(cbind(Homozygotes, DerivedAlleles) ~ Bottleneck, family = quasibinomial())	glm(cbind(Heterozygotes, DerivedAlleles) ~ Bottleneck, family = quasibinomial())		
Relative Genetic Load					
	Total	Homozygotes	Heterozygotes	Model	
Ne	lm(MeanLoad ~ Ne + Region)	lm(MeanHomozygotes ~ Ne + Region)	lm(MeanHeterozygotes ~ Ne + Region)	LM	linear model
Islands	lmer(Load ~ Island + (1|Population))	lmer(Homozygotes ~ Island + (1|Population))	lmer(Heterozygotes ~ Island + (1|Population))	LMER	mixed effects linear model
Location	aov(Load ~ Location * Population)	aov(Homozygotes ~ Location * Population)	aov(Heterozygotes ~ Location * Population)	AOV	analysis of variance
Turure Location	aov(Load ~ Location)	aov(sqrt(Homozygotes) ~ Location)	aov(1/Heterozygotes ~ Location)		
Founder Effect	aov(Load ~ Bottleneck * Population)	aov(sqrt(Homozygotes) ~ Bottleneck * Population)	aov((Heterozygotes)^2 ~ Bottleneck * Population)		

Final Conclusions

This PhD dissertation explores determinants of genetic load in *P. reticulata* in genomic context.

Firstly, regarding source of mutations, our study revealed that guppies, similarly to other Teleostei fish, have strikingly low per site, per generation mutation rate of 2.9 x 10^{-9} (95% confidence interval: 1.92-3.88 x 10^{-9}). In this work, we successfully used two methods of data filtration – hard filtering and machine learning filtering. The estimated by us mutation rate did not differ significantly between them, but we showed evidence that using only strict filters carries the probability of type II error of not finding all true *de novo* mutations. Therefore, in studies aiming at finding all possible *de novo* mutations, machine learning like strategy should be applied to ensure more precision.

Secondly, regarding accumulation of deleterious mutations, we found some evidence that genetic load accumulates in post-bottleneck small populations, but there was no evidence for its increase on the expansion axis. Furthermore, we found no evidence for purging in the transplanted Turure population, thus it is unlikely that such process explain invasive potential of this population. Across multiple populations from many rivers, no relationship between neutral genetic diversity and genetic load was found, while we found significant differences in the load between upstream and downstream populations and between Trinidad and Tobago. This suggests that demographic histories can be important, but long-term N_e is not a universal predictor of genetic load and direct estimates of the load should be used to assess genetic health of populations.

Overall, genetic load in wild populations of *P. reticulata* in Trinidad is determined by low mutation rate and balance between forces related to their demographic histories (selection, drift and admixture). It is, however, not determined by purging, suggesting that other reasons cause observed invasions.

Authorship Statements

OŚWIADCZENIE O AUTORSTWIE

Dla artykułu naukowego:

Burda, K., & Konczal, M. (2023). Validation of machine learning approach for direct mutation rate estimation. *Molecular ecology resources*, *23*(8), 1757–1771. https://doi.org/10.1111/1755-0998.13841

Deklaruję, że praca zawarta w artykule:

Burda, K., & Konczal, M. (2023). Validation of machine learning approach for direct mutation rate estimation. *Molecular ecology resources*, *23*(8), 1757–1771. https://doi.org/10.1111/1755-0998.13841

którego jestem pierwszym współautorem, jest częścią mojej rozprawy doktorskiej.

K.B i M.K. zaprojektowali badanie. K.B. wyizolowała DNA. M.K. wykonał maskowanie regionów powtarzalnych, a K.B. wykonała wszystkie pozostałe analizy bioinformatyczne. K.B. przeprowadziła walidację molekularną wariantów. K.B i M.K napisali pierwszą wersję manuskryptu i przeprowadzili jego rewizję.

Data: 30.09.2024

Katarzyna Burda

Katarzyna Burda

Promotor

J. Radwan

prof. dr hab. Jacek Radwan

OŚWIADCZENIE O AUTORSTWIE

Dla artykułu naukowego:

Burda, K., Mohammed R., Janecka M., Clark D., Kramp R., Radwan J., Konczal, M, (2024) Genetic load is affected by demographic histories in Trinidadian guppies (*Poecilia reticulata*), but does not explain invasiveness after a recent artificial translocation.

Deklaruję, że praca zawarta w artykule:

Burda, K., Mohammed R., Janecka M., Clark D., Kramp R., Radwan J., Konczal, M, (2024) Genetic load is affected by demographic histories in Trinidadian guppies (*Poecilia reticulata*), but does not explain invasiveness after a recent artificial translocation.

którego jestem pierwszym współautorem, jest częścią mojej rozprawy doktorskiej.

K.B, M.K. i J.R. zaprojektowali badanie. K.B., MJ.J., RS.M., DR.C. i RD.K zebrali próby. K.B. wyizolowała DNA. M.K. wywołał warianty przy użyciu `bcftools`, a K.B. wykonała wszystkie pozostałe analizy bioinformatyczne. K.B. i J.R przeprowadzili analizy statystyczne. K.B. i M.K napisali pierwszą wersję manuskryptu. K.B., M.K i J.R. przeprowadzili rewizję manuskryptu.

Data: 30.09.2024

Katarzyna Burda

Katarzyna Burda

Promotor

prof. dr hab. Jacek Radwan

Poznań, 19.09.2024

Authorship statements

I confirm that I am a co-author of the paper: Burda, K., & Konczal, M. (2023). Validation of machine learning approach for direct mutation rate estimation. *Molecular ecology resources*, *23*(8), 1757–1771. https://doi.org/10.1111/1755-0998.13841

I declare that I contributed to conceiving the study, data analysis, interpretation of results writing the first version and revising the manuscript.

dr Mateusz Konczal

Evolutionary Biology Group

Adam Mickiewicz University

I confirm that I am a co-author of th e paper: Burda, K., Janecka M., Mohammed R., Clark D., Kramp R., Radwan J., Konczal, M. (2024) Genetic load is affected by demographic histories in Trinidadian guppies (*Poecilia reticulata*), but does not explain invasiveness after a recent artificial translocation.

I declare that I contributed to conceiving the study, data analysis, interpretation of results writing the first version and revising the manuscript.

dr Mateusz Konczal

Evolutionary Biology Group

Adam Mickiewicz University

Poznań, 26.09.2024

Authorship statements

I confirm that I am a co-author of the paper: Burda, K., Janecka M., Mohammed R., Clark D., Kramp R., Radwan J., Konczal, M, (2024) Genetic load is affected by demographic histories in Trinidadian guppies (*Poecilia reticulata*), but does not explain invasiveness after a recent artificial translocation.

I declare that I contributed to conceiving the study, data analysis, interpretation of results and revising the manuscript.

prof. dr hab. Jacek Radwan
Evolutionary Biology Group
Adam Mickiewicz University

El Paso, TX, USA, September 20, 2024.

place, date

Authorship statement

I confirm that I am a co-author of the paper: Burda, K., Janecka MJ., Mohammed R., Clark D., Kramp R., Radwan J., Konczal, M. (2024) Genetic load is affected by demographic histories in Trinidadian guppies (*Poecilia reticulata*), but does not explain invasiveness after a recent artificial translocation.

I declare that I contributed to samples collection.

Mary Janecka

....

Mary Janecka

The University of Texas at El Paso

65 Second Avenue,
Gajadhar Lands,
Princes Town
Trinidad and Tobago, WI
23rd September 2024

Authorship statement

I confirm that I am a co-author of the paper: Burda, K., Janecka M., Mohammed R., Clark D., Kramp R., Radwan J., Konczal, M. (2024) Genetic load is affected by demographic histories in Trinidadian guppies (*Poecilia reticulata*), but does not explain invasiveness after a recent artificial translocation.

I declare that I contributed to samples collection.

Ryan

Ryan S. Mohammed
Auburn University

Pittsburgh, Pennsylvania, United States of America

20-09-2024

Authorship statement

I confirm that I am a co-author of the paper: Burda, K., Janecka M., Mohammed R., Clark D., Kramp R., Radwan J., Konczal, M. (2024) Genetic load is affected by demographic histories in Trinidadian guppies (*Poecilia reticulata*), but does not explain invasiveness after a recent artificial translocation.

I declare that I contributed to samples collection.

David R Clark Jr

....

David R. Clark

University of Pittsburgh

Pittsburgh, Pennsylvania, U.S.A. 20240920

place, date

Authorship statement

I confirm that I am a co-author of the paper: Burda, K., Janecka M., Mohammed R., Clark D., Kramp R.D., Radwan J., Konczal, M. (2024) Genetic load is affected by demographic histories in Trinidadian guppies (*Poecilia reticulata*), but does not explain invasiveness after a recent artificial translocation.

I declare that I contributed to samples collection.

....

Rachael D. Kramp

University of Pittsburgh, PA