UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA ANIMAL

Genomics of adaptation to cadmium on a spatially heterogeneous environment

Marta Cardoso Ferreira

Mestrado em Biologia Evolutiva e do Desenvolvimento

Dissertação orientada por:
Doutora Inês Regina Lopes de Mendonça Fragata
Professor Doutor Vítor Martins Conde e Sousa

2023

Para o Afonso e para a Matilde,
Para a minha mãe,
Para os meus avós.

Acknowledgments

I would like to start by saying that I think this is probably one of the hardest sections to write, on the master thesis. This is because 1) my writing skills are not the best; 2) I do not express my emotions that often and therefore I do not really know how to thank everyone. Because writing "thank you for everything" is not enough; 3) I really hope I do not forget anyone.

To all the teachers that I have ever had. Sometimes we tend to forget them, but the truth is that we would not be who we are without them. Thank you to, especially, all my English teachers and anything and anyone that helped me learn English. I grew up from someone who despised English and never was good enough to someone who is writing (or wrote) an entire thesis in English.

To everyone that works in the hospitals. Thank you for improving and saving my life countless times. To Prof. Élio, Gabriela and Margarida for accepting me in this master degree.

To Prof Élio again and to Clémence Carron-Homo for giving me an opportunity to study in Paris for one month and for helping us when we needed. To Inês and Vítor. I do not really want to say just "thank you for everything" but thank you for everything ! I could not have asked for better and I hope you know that ! We (students) need more people like you 🙂

To the entire mite squad for welcoming me in the group. Thank you for the trip we did together to Açores. I do not think I slept in such a hot room since then, with such an amazing bed. I also deeply appreciated the cold showers and the bananas in the morning. To Mariya for trying to teach me lab work. To Diogo for reading all the abstracts even though he did not have to. To Miguel, your legacy will live on. Thank you to Sara, without her we would not have the mites, she is the backbone of the group !

To the evolutionary genetics group (I keep on saying we need a cool name too. This looks too formal !). To Beatriz and Jonna's group for helping me with SLiM. To João M. and Sofia for letting me go fishing with them for one day. Even though we only got one fish, for me it was a nice trip. To Alex for writing an entire script. I promise I did not forget it. To João C., I do not know if you know, but it was thanks to you that I found the paper about heterogeneous environments with variation in space. That was a hard finding and I almost did not have the second part of the discussion !

(Sara and Diogo and Miguel and Inês, but especially Leonor, if you are reading this and if you did not notice, both the last two paragraphs have exactly 102 words! No, I did not have to count the words that many times for them to be precisely equal. It was all natural. It IS 50/50!).

To all team BED but especially to Afonso, Davide, Carolina, and Vanessa. Thank you for tolerating me and the more than one hundred memes that I sent you. Thank you, Carolina, for taking me to Paris (in case someone needs anyone to blame for all the photos of the very tasty desserts!).

To Doro, Lúcia and João. Thank you for still tolerating me, even after all these years. We do not see each other as often as we used to (every single day !) but when we do it feels like it was just yesterday. Doro, come back to Portugal !

Para a minha família por aturarem o meu mau feitio e teimosia (apesar de eu ter a quem sair …).

Abstract

Many species are shifting their ranges due to climate change, hence understanding how adaptation occurs in spatially heterogeneous environments is important. Yet, little is known about adaptation in heterogeneous environments, in particular at the genomic level. To tackle this question, experimental evolution under controlled laboratory conditions, combined with population genomics is a powerful approach. Such "Evolve-and-Reseq" studies allow uncovering the genetic basis of adaptive traits in different organisms. A caveat of experimental evolution is that the potentially reduced genetic diversity of laboratory populations might influence both the speed and end point of adaptation.

In this thesis, we address these issues using experimental populations of two closely related species (*Tetranychus urticae* and *T. evansi*) of spider mites, which are arrhenotokous haplodiploid herbivorous species, considered worldwide pests. Both species can feed on tomato plants, which are known to accumulate heavy metals. These metals can be used as a defence against herbivory, and it is known that cadmium accumulation can decrease spider mites' fitness. Taking advantage of this setup, we addressed two important gaps in knowledge by using population genomic data of outbred and inbred lines and experimental populations. Namely (1) quantify the genetic variability of the outbred populations and inbred lines, and (2) understand how spatial heterogeneity affects the adaptation to new environments, namely on environments with cadmium.

Using Illumina sequencing of pools of >200 mites (Pool-seq), we estimated expected heterozygosity as a proxy for the genetic diversity of outbred *T. evansi* and *T. urticae* and nine inbred lines of *T. evansi*. We expected inbred lines would present lower genetic diversity compared to the outbred populations, as a result of the inbreeding process. However, we found that the nine inbred lines had similar expected heterozygosity to the outbred population, leading us to conclude they are not inbred isogenic lines. These results might be explained by gene flow between lines and/or outbred population, but we also found an effect of the choice of reference genome.

To identify genomic regions of adaptation to cadmium on *T. urticae* and/or to spatially heterogeneous environments, we analysed Pool-seq data from experimental evolution selection regimes of *T. urticae*, evolving with low (control) or high cadmium concentrations (homogeneous environment) or both (heterogeneous environments). We used consistent changes in allele frequencies across five replicates as evidence for positive selection, assuming that the control regime represented the initial allele frequency. We found many SNPs with significant changes in allele frequencies, both in homogeneous and heterogeneous environments, indicating a polygenic basis of adaptation. We found that only 10.9% of the candidate genes are shared between homogeneous and heterogeneous environments, suggesting that adaptation in heterogeneous environments may select for alleles different from those favoured in the different homogeneous environments. Furthermore, metallothioneins did not show significative changes in allele frequencies in environments with cadmium, although they are the best-known stress response system to heavy metals, but we found a voltage-gated T-type calcium channel (*CACNA2D3*) with several SNPs with allele frequency changes consistent with adaptation in homogeneous and heterogeneous environments.

Keywords: *Tetranychus urticae*; *Tetranychus evansi*; Outbred populations; Inbred lines; *CACNA2D3*

Resumo

A adaptação é um processo que permite a sobrevivência e reprodução dos organismos a alterações no meio ambiente. Com as alterações climáticas provocadas pela acção antropogénica, que levam a que muitas espécies tenham que enfrentar ambientes extremos no limite da sua tolerância, é de uma importância extrema o estudo dos processos evolutivos que permitem a adaptação das espécies a novos ambientes.

No entanto, desvendar a base genética da adaptação não é um processo fácil e involve diversos fatores. A evolução experimental envolve observação em tempo real de populações em laboratório que permitem elucidar a base genética da adaptação. No entanto, no estabelecimento de populações em laboratório está implícito uma redução do tamanho efetivo da população face à população existente na natureza. Tal é importante pois a redução do tamanho efetivo da população pode levar à redução da diversidade genética. Por outro lado, a fixação de uma mutação benéfica está dependente do tamanho efetivo populacional e da deriva genética: quanto maior o efetivo populacional, menor vai ser o impacto da deriva genética nas frequências alélicas e mais eficiente a seleção natural. Assim, há autores que defendem que as populações mantidas em laboratório podem não ser representativas de populações naturais por não terem variabilidade genética suficiente para produzir respostas aos processos que ocorrem na natureza. No entanto, as populações em laboratório são úteis pois permitem um melhor controlo de fatores externos que são difíceis de controlar em estudos com populações naturais. Estudos para quantificar correlações genéticas entre variáveis fenotípicas requerem o uso de populações endogâmicas, populações estas mantidas com um efetivo populacional muito reduzido e geradas através de cruzamentos entre irmãos, garantindo um nível de *inbreeding* elevado, com o objetivo de atingir um ponto em que todos os indivíduos são iguais entre si geneticamente. Estas populações podem ser criadas a partir de populações exogâmicas, populações em laboratório, geradas a partir de várias populações naturais e mantidas em números elevados com cruzamentos ao acaso. Atualmente, graças às novas tecnologias de sequenciação é possível quantificar a variabilidade genética presente em populações de laboratório, quer de linhas endogâmicas, quer de populações exogâmicas.

Ácaros aranha (Acari: Tetranychidae) são uma família de espécies de herbívoros haplodiploides, em que os machos são haploides e são produzidos de ovos não fertilizados e as fêmeas são diploides, produzidas a partir de ovos fertilizados. São ácaros aranha pois têm a capacidade de produzir teias com muitas funções, nomeadamente para a proteção dos ovos e dispersão. O género *Tetranychus* Dufour, 1832, é composto por 153 espécies, entre as quais o *Tetranychus urticae* Koch, 1836. Esta é uma praga generalista que coloniza mais de 1100 espécies de plantas, muitas destas de grande importância económica para a agricultura como é o caso dos tomateiros. *T. urticae* é a única espécie de ácaro aranha com um genoma de referência que atualmente é composto por três grandes *super scaffolds*. *Tetranychus evansi* é uma outra espécie de ácaros aranha muito estudada devido à sua capacidade de suprimir compostos gerados pelas plantas como um mecanismo de defesa. Dado que ambas têm tempos de geração curtos com a possibilidade de criação de populações com um grande efetivo populacional, além de serem fáceis de manusear no laboratório são ideais em estudos que envolvem evolução experimental.

A combinação das tecnologias de sequenciação com a evolução experimental é ideal para o estudo da base genética da adaptação. Estudos com evolução experimental focam-se maioritariamente em perceber como é que os organismos se adaptam a apenas a uma única e constante alteração no ambiente. Enquanto alguns estudos que envolvem evolução experimental introduziram variações no ambiente numa escala diária ("*fluctuating environments*"), poucos destes estudos têm em consideração variações ambientais no espaço. Assim, entender de que forma a existência de ambientes heterogéneos tem impacto nas populações, pode ser importante no processo de adaptação.

As plantas e os herbívoros vivem em conjunto desde há milhares de anos. Desta forma, plantas desenvolveram diversos mecanismos de defesa, de forma a evitar danos causados pelos herbívoros. De acordo com a Hipótese da defesa elemental, a acumulação de metais pesados nas folhas pode ser considerada um mecanismo de defesa. No entanto, sabe-se que alguns herbívoros, conseguem adaptar-se a essas condições e alimentarem-se de folhas com concentrações elevadas de metais pesados como o cádmio. Foi proposto que o mecanismo pelo qual os herbívoros são capazes de inativar os efeitos nocivos dos metais pesados é através de metalotioneínas, proteínas que regulam a concentração de metais pesados nas células.

Assim, nesta tese pretendeu-se: 1) quantificar a variabilidade genética de populações endo e exogâmicas criadas e mantidas em laboratório e 2) perceber como é que a heterogeneidade ao nível espacial afeta a adaptação a novos ambientes, nomeadamente em ambientes com cádmio. Através da sequenciação *Illumina* de "pools" de mais de 200 ácaros (Pool-seq), foi estimada a heterozigotia esperada média das populações exogâmicas de *T. urticae* e *T. evansi* e de nove populações endogâmicas de *T. evansi* mantidas em laboratório. Era esperado que as populações endogâmicas tivessem uma menor diversidade genética em comparação com as populações exogâmicas. No entanto, os resultados obtidos indicam que as populações endogâmicas tinham heterozigotia esperada média semelhante à população exogâmica de *T. evansi*. Tal leva a crer que estas populações não são populações endogâmicas como se tinha pensado. Estes resultados podem ser explicados pela existência de fluxo genético entre populações endogâmicas e/ou com a população exogâmica durante a manutenção dos indivíduos no laboratório. No entanto, o facto de as populações endogâmicas e exogâmicas terem a mesma diversidade genética também pode ser explicado por diversos problemas na geração dos dados genómicos e no protocolo bioinformático: 1) erros de sequenciação que dependem do tipo de máquina utilizada para a sequenciação das amostras; 2) erros de mapeamento associados ao algoritmo de alinhamento utilizado; 3) enviesamentos devido ao genoma de referência utilizado (não existe atualmente genoma de referência para a *T. evansi*); 4) diferenças estruturais no genoma das duas espécies; 5) presença de regiões genómicas semelhantes no núcleo e mitocôndria (e.g. NumtS).

De forma a identificar regiões no genoma envolvidas na adaptação ao cádmio em *T. urticae*, foram criados a partir da população exogâmica de *T. urticae*, três regimes de seleção em que os ácaros estavam expostos a plantas com e sem cádmio, em ambientes homogéneos e heterogéneos. De forma a detetar seleção positiva, foram aplicados testes estatísticos que detetam alterações nas frequências alélicas consistentes nas cinco réplicas, considerando as frequências alélicas do regime controlo como as frequências iniciais. Encontrámos vários SNPs com alterações significativas nas frequências alélicas, tanto nos ambientes homogéneos como nos ambientes heterogéneos, indicando que a adaptação tem uma base genética poligénica. Verificámos que apenas 10.9% dos SNPs com alterações nas frequências alélicas significativas em comparação com o regime controlo eram comuns nos ambientes homogéneo e heterogéneo, o que sugere que a adaptação a ambientes homogéneos pode envolver alelos diferentes dos envolvidos em ambientes homogéneos, o que demonstra a importância de estudos com variação espacial. Surpreendentemente, não foram encontradas alterações significativas nas frequências alélicas nos genes que codificam as metalotioneínas embora estes genes sejam dos mais estudados na resposta a metais pesados. No entanto, um gene que codifica um canal de cálcio (*CACNA2D3*) respondeu de forma diferente em ambientes homogéneos e heterogéneos.

No futuro, será necessário refinar o protocolo bioinformático, mais especificamente analisando variação estrutural (por exemplo, inserções e deleções) para além de SNPs. Uma forma de melhorar a deteção de SNPs sobre o efeito de seleção positiva e balanceadora seria através de simulações de genomas para modelar a evolução de populações em condições semelhantes às que foram mantidas no laboratório. Adicionalmente, seria interessante investigar em mais detalhe as os genes que contêm SNPs que foram detetados com alterações nas frequências alélicas consistentes com adaptação ao cádmio. De modo a esclarecer a base genética da adaptação a

ambientes heterogéneos, seria interessante analisar em mais detalhe os genes com SNPs detetados sobre seleção exclusivamente nos ambientes heterogéneos.

Palavras-chave: *Tetranychus urticae*; *Tetranychus evansi*; Populações exogâmicas; Populações endogâmicas; *CACNA2D3*

Table of Contents

List of Figures

List of Tables

---

1. Introduction

Adaptation is a crucial process that allows organisms to survive and reproduce when there are changes in the environment[1,2]. Over the last few decades, attention on the molecular basis of adaptation increased mostly thanks to quantitative trait locus (QTL) analysis, molecular population genetics and microbial experimental evolution studies[2]. More recently, the study of adaptation has been focused on assessing the adaptive potential to climate change, since many species are being forced to confront dramatically changed environments, a consequence of anthropogenic actions[3]. Therefore, understanding the genetic basis of adaptation is of utmost priority and a major goal in evolutionary biology[4,5].

At the genetic level, adaptation can result from the appearance of new beneficial mutations, by introgression and standing genetic variation at a single or multiple loci[6,7]. Additionally, it depends on the interaction of different factors, such as environmental variation in time and space, the initial allele frequencies of the beneficial mutations and the demographic history of the populations that englobes gene flow and effective population size. Together these factors determine the impact of drift and the strength and mode of selection[4,8–12]. For instance, the fate of a beneficial mutation with a selective coefficient of 0.001 (s=0.001) will depend on the effective population size: the beneficial mutation can become fixed in a large population, provided that twice the effective size (*Ne*) times the selective coefficient *s* is larger than 1 (e.g., with Ne=10000, $2Ne*s>1$), whereas it is more likely to be lost due to drift in a small population. That is why the effective population size is important during the establishment and maintenance of laboratory populations, as most founding events entail a reduction in population size[13], leading to loss of genetic diversity that can hamper adaptation from standing variation, and/or prevent adaptation due to a stronger effect of drift. Yet, there are conditions where maintaining populations at a very small effective size is useful such that drift leads to a quick loss of diversity, ensuring all individuals are genetically identical. Inbred lines are laboratory populations that can be shared among laboratories, and are a useful tool to measure the broad-sense heritability of a given trait, allowing for a better understanding of the genetic architecture of quantitative traits by linking phenotypes to environmental conditions[14,15]. They can be obtained through sib-mating crossing, which guarantees high levels of inbreeding after some generations[15]. Several inbred lines have been created for different species, representing different fixed genotypes (reviewed in [15]), one of them being DGRPs for *Drosophila melanogaster*[14]. In order to generate lines that represent different genotypes, it is important that there is enough genetic variability in the populations. This can be achieved by using outbred populations, which are populations generated via controlled crossings and maintained with large population sizes and random mating, to avoid loss of genetic diversity due to genetic drift (which is stronger the smaller the effective population size[15,16]. Although studies with populations in the laboratory allow for the control of environmental variables, which is difficult to achieve in nature[17,18], they have long been criticized because laboratory populations might not harbour sufficient genetic variability to produce representative responses of the processes occurring in natural populations[19,20]. Therefore, quantifying the genetic variability of populations is essential for studies involving laboratory populations. This has been done in some species/systems, but often using just a small number of selected markers (e.g. microsatellites[21,22], which may not effectively reflect genome-wide diversity[23]).

Recent advances in high-throughput sequencing and the increase in computational power, over the last decades, changed the way researchers conducted their experiments. Sequencing multiple individuals from the same or different species is now a common practice in genomic studies[24,25] and sequencing several individuals of the same population together is now possible through Pool-sequencing (pool-seq)[26–28]. By comparing the genomic composition of populations from different environments it is possible to identify regions showing genetic differentiation. It is expected that, in those genomic regions involved in adaptation, there will be a higher genetic differentiation,

than at the neutral genomic background[29,30]. Studies have shown that adaptation can be highly specific to certain genomic regions, with a single gene of large effect. A well-known example is the genetic variation in colour-specific adaptation on deer mice of Nebraska[31], where selection for light-coloured mice in light soil and dark-coloured mice in dark soil is achieved through multiple mutations at a single gene, the *Agouti* gene. Another example comes from sexually selected traits, such as Soay sheep horns[32]. Most variation in this trait is maintained by a trade-off between natural and sexual selection at a single gene, relaxin-like receptor 2 (*RXFP2*). Contrastingly, adaptation on complex traits, such as height, is known to have a polygenic basis, with many genes involved in the process of adaptation, each one of them with a small effect[33]. Another example of a polygenic basis is the adaptation to drought in *Brassica rapa*[34], where hundreds of SNPs in stress response genes were identified.

Spider mites (Acari: Tetranychidae) are a family of arrhenotokous haplodiploid herbivorous species, meaning that males are haploid and are produced from unfertilized eggs and females are diploid and produced from fertilized eggs[35]. The name "spider" highlights their ability to produce silk-like webbing used for the protection of eggs, to make a protective tent for moulting, guarding of quiescent females deutonymphs by the males and it is also used for dispersal[36–38]. The genus Tetranychus Dufour, 1832 gathers 153 valid species[39], one of which is the two-spotted spider mite, *Tetranychus urticae* Koch, 1836. *T. urticae* is an extreme generalist pest, colonizing more than 1100 host species[36,39], some of them being of high economic importance in agriculture such as tomatoes, hops, strawberries, apples, almond, grapes, corn, greenhouse vegetables and ornamentals[36,40]. *T. urticae* is the first chelicerate with a completed reference genome [41], which is assembled into 3 super scaffolds[42]. *Tetranychus evansi* is another species of spider mites, native to South America but spread worldwide, with a highly invasive nature in solanaceous crops, especially in tomato plants[43,44]. Given short generation times, the possibility to generate populations of large sizes and ease of maintenance in the laboratory, spider mites are ideal for experimental evolution studies[45]. This methodology, through the control of different environmental variables on several replicated populations, offers a way to establish a causal link between evolutionary processes and adaptation patterns[18].

The combination of pool-seq with experimental evolution (evolve and resequence)[18,26] is a powerful tool to investigate the genetic basis of adaptation to different environments (e.g. [22,46,47]). Most experimental evolution studies focus only in changes of a single and constant factor throughout time[34,47–53]. While some studies introduced fluctuating environments to their experimental design, where populations are subjected to, for example, a trend of successive increase in temperature over a stipulated number of generations and/or changes in temperature throughout the day, *i.e.* circadian fluctuation[54–58], a very few experimental designs explore variable selection in space, despite the probable importance of variable selection in space[5]. In fact, to our knowledge, only Huang, Y. *et al* (2014; [59]) aimed to examine the effect of environmental heterogeneity on genetic variance, using high-throughput sequencing technology and experimental evolution. The authors found a higher within-population diversity (π) on a spatially heterogeneous regime compared in homogeneous regimes, for both sites under differential ecological selection (and those closely linked to them) and when the non-neutral variable sites were included. Therefore, more studies with the same or similar experimental designs are necessary to better understand how spatial heterogeneity affects adaptation to new environments.

Plants and herbivores have been living together for several million years, and thus plants have evolved several defence mechanisms (reviewed in [60]): direct defences, such as the production of specialized morphological structures (e.g. hairs, trichomes, thorns) or indirect defences such as the release of volatiles that attract natural enemies of the herbivores (e.g. nectar). In the context of heavy metal accumulation, a possible defence mechanism has been proposed by the Elemental-defence hypothesis, explaining why some plants (known as hyperaccumulators) absorve one or more types of heavy metals from the soil with consequent accumulation in the leaves[61,62]. In choice studies, several Lepidoptera larvae preferred plants without Selenium, a non-essential

heavy metal[63–65], suggesting the importance of the accumulation of Selenium (and heavy metals in general) as a plant defence. However, some herbivores were able to adapt to feed on plants with high heavy metal concentrations. For example, Freeman, J.L. *et al* (2006, [66]) identified a Se-tolerant variety of the *Plutella xylostella* moth (*P. xylostella* stanleyi) abundant on the desert princes plume (*Stanleya pinnata*) plants that is able to survive on selenium hyperaccumulator plants of the same species[66]. Previous studies from the team where this thesis was developed have also corroborated this hypothesis, by demonstrating that cadmium accumulation (another non-essential heavy metal) by tomato plants (*Solanum lycopersicum*) affects the performance of spider mites[67]. Spider mites also show variance in the response to cadmium on metal-accumulating tomato plants and depend on both the populations and metal concentrations used[68]. Therefore, studies on the adaptation to heavy metals in herbivores are needed, to better understand the interaction between herbivores and hyperaccumulating plants.

In this thesis, we aimed to 1) quantify the genetic variability of the outbred populations and inbred lines, described in Godinho *et al* (2020, [15]) and 2) understand how spatial heterogeneity affects the adaptation to new environments, namely on cadmium heavy metal environments. For aim one, we used genome-wide data from pools of individuals of two spider mites species present in the laboratory, namely two outbred populations (*T. urticae* and *T. evansi*) and 9 inbred lines of *T. evansi*. We estimated expected heterozygosity as a proxy of the genetic variability, and we hypothesized that outbred populations showed similar levels of genetic variability as natural populations. In addition, we also hypothesized that inbred lines would present lower genetic diversity compared to the outbred populations, because of the inbreeding process. For aim two, we used Pool-Seq genome-wide data from an experimental evolution experiment following changes associated with adaptation to heavy metals in selection regimes with spider mites exposed in homogeneous or heterogeneous spatial environments composed of tomato plants developed with or without high cadmium concentrations. We analysed genomic patterns of the experimental evolution regimes of the *T. urticae* outbred population to understand the genetic changes associated with the adaptation of spider mites to cadmium. Namely, we aim to identify genomic regions of adaptation to cadmium on a homogenous regime. We expect to detect changes in allele frequencies on metallothioneins, as these proteins are the best-known stress response system to heavy metals on several taxa from invertebrates to fish and mammals[69–72]. Moreover, we also want to quantify the impact of spatial variation in the environment (*i.e.* plants with and without cadmium - heterogeneous regime) on changes in allele frequencies at the genomic level, and whether they differ from the homogeneous regime.

2. Methods

2.1 Genetic characterization of the populations in the laboratory

2.1.1 Creation and maintenance of Outbred and Inbred lines

The outbred and Inbred lines were established and maintained in the laboratory of MITE^2 group at cE3c, prior to the start of this thesis. Given its importance to interpret results, a brief description of how they were established and maintained before generating whole genome data is provided here. A detailed description is given in Godinho *et al* (2020, [15]). In summary, for the creation of the outbred populations, methods to avoid well-known reproductive incompatibilities in spider mites were taken into consideration. For *T. urticae*, only individuals with the green form were considered, while for *T. evansi*, only individuals from populations corresponding to clade I were kept[15]. Moreover, populations were merged by performing inter-population crosses in a controlled match design, to avoid the overrepresentation of genotypes from a given population. All mites were collected from tomato field plants in Portugal. The *T. evansi* outbred was founded with 576

females from 4 different nearby locations, 3 of them within the Lisbon district and one in Algarve. While the *T. urticae* outbred population was founded with 306 females, from 3 different geographic locations, within the Estremadura region. Since the creation of both Outbred populations, they have been maintained in the laboratory at 25 ºC and 65% humidity, with overlapping generations in detached tomato leaves, replaced every two weeks.

Inbred female lines of *T. evansi* were created by randomly sampling and isolating a single female, two generations after the creation of the *T. evansi* outbred population. In total, 9 lines of brother-sister mating were maintained for over 15 generations. The lines used in this thesis had an expected inbreeding coefficient of at least 95.1% and a probability of being fully inbred of at least 93.6% when collected[15]. We analysed the genomic data of 9 inbred lines coded as Te8, Te14, Te19, Te23, Te27, Te28, Te29, Te32 and Te42.

2.1.2 Creation of experimental evolution populations

To investigate how populations adapt to cadmium in homogeneous and heterogeneous environments, an experimental evolution study was previously conducted in the MITE[2] laboratory, as illustrated in Figure 2.1. Three selection regimes were created by transferring 220 females from the *T. urticae* outbred population, experimentally evolving with 1) tomato plants with no cadmium (control, homogeneous environment); 2) tomato plants with high cadmium concentrations (homogeneous environment); 3) tomato plants with either no cadmium or high cadmium concentration (heterogeneous environment). Cadmium concentrations of 2 mM were chosen based on the results obtained in Godinho *et al* (2018, [67]). All selection regimes were replicated five times and maintained for over 55 discrete generations transferring 220 adult females from the previous generation to a new box.

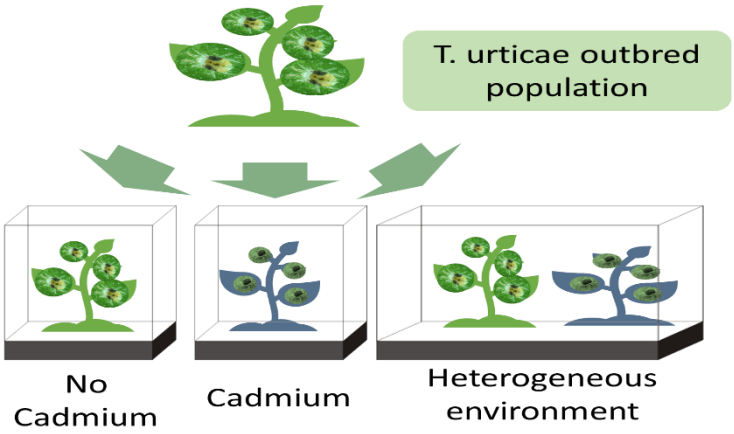

Figure 2.1: Schematic representation of the experimental design used in the experimental evolution. The three selection regimes were each created with 220 females from the same outbred population. A non-cadmium regime where *T. urticae* evolved on tomato plants (control); a cadmium regime where the population evolved on tomato plants watered with a cadmium solution twice a week (homogeneous environment); an heterogeneous environment with plants evolved with and without the presence of cadmium. All three regimes had 5 replicates each and the experiment lasted for 55 discrete generations.

2.2 Whole-genome sequencing (WGS) of pools of individuals (Pool-seq)

After one generation of common garden, DNA was extracted for all samples using a modified protocol from (Grbić *et al*. 2011), where the number of females used for pools and the sequencing technology varied depending on the objectives, as described below.

To assess the genetic diversity of both inbred and outbred populations (section 2.1.1), pools of 400 females were randomly sampled. Library construction and sequencing were performed at Instituto Gulbenkian de Ciência, where pair-end reads of 150 bp reads were sequenced with Illumina Miseq. Raw data (fastq files) from two Outbred populations and nine Inbred lines were analysed.

To detect signatures of adaptation to cadmium at the genomic level, and whether those differ between homogeneous and heterogeneous environments, we extracted DNA and analysed pools of 200 females randomly sampled from each of the 5 replicates of the 3 selection regimes at the end of the experiment (section 2.1.2). Sequencing was outsourced at Novogene, obtaining for each pool pair-end reads of 150 bp using Illumina Next seq 2000. In total 15 datasets corresponding to five (replicates) x three (regimes) of raw Illumina data (fastq files) were processed and analysed.

2.3 Reference genome selection

In this thesis, we used the reference genome of *Tetranychus urticae* from Wybouw *et al* (2019, [42]). This genome was chosen because it is assembled into three main super scaffolds called pseudochromosome 1, 2 and 3. This assembly is thus less fragmented than the one published in 2011 ([41]). To quantify what is the proportion of the genome encompassed by these three super scaffolds we used Qualimap (version 2.2.2a; [73,74]). According to Qualimap results, these three pseudochromosomes represent 94.41% of the entire genome with an estimated size of approximately 90 Mb. All other scaffolds were discarded. No reference genome is available for *T. evansi* and so all reads from *T. evansi* pools were mapped against the *T. urticae* reference genome.

To evaluate the impact of the choice of the reference genome, given that there are full mitochondria sequences available for several strains of *T. urticae* and *T. evansi*, we mapped reads from the *T. urticae* outbred population pool against different mitochondria reference genomes of *T. urticae* and *T. evansi* (Table 1).

Table 2.1: Different mitochondria strains of the two species analysed (*T. urticae* and *T. evansi*). All genomes were extracted from NCBI in March 2022. Host plant represents the plant mites were kept in the laboratory; BR-VL is a strain created from LS-VL; *T. evansi* was collected directly from eggplant fields and sequenced

Strain	Specie	Host Plant	Reference / NCBI accession number
London	*T. urticae*	Bean	Wybouw *et al*. 2019 ([42])
LS-VL	*T. urticae*	Bean	Van Leeuwen *et al*. 2008 ([75]) / NC_010526.1 or EU345430.1
BR-VL	*T. urticae*	Bean	Van Leeuwen, Tirry, and Nauen 2006 ([76]) / EU556754.1
Red	*T. urticae*	Bean	Chen *et al*. 2014 ([72]) / KJ729023.1
Green	*T. urticae*	Bean	Chen *et al*. 2014 ([72]) / KJ729023.1
-	*T. evansi*	Eggplant (field)	Sun *et al*. 2019 ([77]) / MN417333.1

2.4 Data Processing

The bioinformatics pipeline was the same for all the reads (fastq files) analysed, unless otherwise stated. Reads from all outbred and inbred lines (section 2.1.1) were mapped against the *T. urticae*

reference genome, available *T. evansi* mitogenome and all the *T. urticae* mitogenome strains present in table 1. Reads from the three experimental evolution selection regimes (section 2.1.2.) were mapped only to the *T. urticae* reference genome.

The pre-processing of the raw reads sequenced data involved the following steps. Adapters were removed from reads by the sequencing facilities. We assessed the quality of the raw reads (without adapters) with FastQC (version 0.11.9; [78]). Due to poor base quality at the tips of reads, trimming was performed with Trimmomatic (version 0.39, [79]) using a pair-end mode 'PE' to remove low base quality reads. Based on results from FastQC we removed the first 19bp from the tips for the outbred and inbred lines (CROP: 150 HEADCROP:19), and the first 10bp (HEADCROP:10) for the three experimental evolution regimes, as those corresponded to regions with a different AT and/or GC per base sequence content. Quality was accessed again with FastQC (version v0.11.9; [78]) to ensure that trimming was successful. We computed several metrics for raw files and trimmed reads, which are reported in Supplementary Table 1.

Mapping of reads to the different reference genomes was performed using the Burrows-Wheeler Aligner (BWA) MEM (version 0.7.17, [80]) with the default settings. To ensure compatibility with the program Picard tools, which was used for further filtering steps, we included the M argument in BWA MEM. The resulting BAM files were validated and further filtered using several Picard tools (version 2.26.11, [81]): read groups were added and files were sorted with AddOrReplaceReadGroups; duplicated sequences were removed with MarkDuplicates; and the mapping quality was set to zero for unmapped reads using CleanSam. Further filtering was done to keep only reads properly mapped and with a base quality higher than a Phred score of 20, which was performed with SAMTOOLS view (version 1.11, [82]). Several metrics regarding mapping, such as coverage and the number of mapping reads, were obtained from QUALIMAP (version 2.2.2a; [73,74]).

To obtain estimates of allele frequencies in each pool of individuals, we considered the allele counts for each position in the genome and for each nucleotide, estimated using ANGSD (version 0.933, [83]). We assumed that the relative number of reads (depth of coverage) for each nucleotide for each population and each site reflected the allele frequencies. To account for the effects of errors during mapping and sequencing associated with genomic regions with low depth of coverage, we discarded all the sites with base quality lower than 20 (in Phred score) and depth of coverage less than four reads within each sample. Moreover, to determine the minor and major alleles, we considered the total counts of each nucleotide for all the populations being analysed, using a custom script. For example, to detect SNPs under positive selection (for more details check section 2.6), the minor allele was identified based on the total counts of each nucleotide of the 10 populations (5 replicas of the control and 5 replicas of the selection regime). If the overall minor allele is a cytosine (C) for a particular SNP, we considered the counts for the cytosine in each one of the 10 populations.

For the analysis of variation in the mitochondria, since it is known that there are regions in the nuclear genome similar to the mitochondria genome (e.g. nuclear sequences of mitochondrial origin - NumtS) can introduce an additional bias on SNP detection[84,85], two additional steps were performed before the allele count step on ANGSD for the files mapped only against mitochondrial genomes (Table 1). First, only reads mapped to the mitochondria genome were kept using SAMTOOLS view (version 1.11, [82]). Second, ANGSD read count to estimate the allele frequencies was done with a minimum mapping quality of 50, based on the mean mapping quality results obtained from QUALIMAP (version 2.2.2a; [73,74]). In all the steps of the bioinformatic pipeline described above, the default parameter settings were used for all software mentioned, unless specifically mentioned otherwise.

2.5 Calculation of measures of genetic diversity and differentiation

To quantify genetic diversity and differentiation, we considered statistics that depend on allele frequencies, such as expected heterozygosity (a measure of diversity within populations) and pairwise F_{st} (a measure of differentiation between pairs of populations). These were calculated on R (version 4.2.2, [86]) using custom scripts and are described in the next sections. Sequencing depth of coverage has a strong influence on allele frequency estimates and the ability to detect SNPs using poolSeq data[26,87]. Moreover, the depth of coverage is expected to show substantial variation across the genome, due to differences in sequencing efficiency. Therefore, to define a set of SNPs common in all samples, minimising the influence of sequencing errors and mapping errors, we only kept sites with: 1) allele counts higher than 0.01 * total depth of coverage, for each sample; 2) depth of coverage values between the quantile 25 and 75, excluding both positions with low and high depth of coverage values; and 3) only two alleles (major and minor alleles). Additionally, to ensure that the same SNPs existed for the five replicates in the experimental evolution regimes, sites where minor counts were equal to zero in three or four of the five replicates in each regime were excluded from the analysis.

2.5.1 Genetic diversity

The expected heterozygosity (He) was used to quantify genetic diversity. It was computed for each genome position using the following formula:

$$He = 1 - f_A^2 - f_C^2 - f_G^2 - f_T^2, \quad (2.1)$$

where f_X indicates the allele frequency of nucleotide x. For each position in the genome, the allele frequencies were calculated as:

$$f_X = c_X / D, \quad (2.2)$$

where f_x indicates the frequency of allele X (A, C, G or T) in a given sample, c_X is the number of reads with allele X and D indicates the total depth of coverage for each position in the genome. Positions with expected heterozygosity equal to zero were removed from the analysis since it is known that the nucleus and mitochondria genomes have different mutation rates, that will affect the number of polymorphic sites present in these two genomes[88]. Therefore, we computed the average He separately for the nucleus DNA and mitochondria DNA and only for the positions with He higher than zero.

2.5.2 Genetic differentiation

To estimate the degree of differentiation between pairs of populations we estimated mean F_{ST} values separately for the mitochondria and for the nuclear genomes. Mean F_{ST} was calculated between: 1) the two Outbred populations of *T. evansi* and *T. urticae*; 2) the nine inbred lines of *T. evansi* and outbred *T. evansi* population; 3) all possible pairs of inbred lines among themselves. F_{ST} was computed based on the read counts for each nucleotide in the two populations considered for each pair. The total number of reads for each nucleotide was obtained by summing the read counts of the two populations for each nucleotide. For each position, the minor allele was defined as the allele with fewer read counts across the two populations. F_{ST} was calculated using the Nei (1973; [89]) estimator as:

$$F_{ST} = \frac{\pi_b - \pi_w}{\pi_b}, \quad (2.3)$$

where π_w is the mean genetic diversity within populations, and π_b is the genetic diversity calculated between populations. Assuming that SNPs are biallelic with a frequency f_Z of the minor allele across the two populations in population Z, the mean within the diversity of both populations was calculated based on the genetic diversity π_{wZ} of each population Z as follows:

$$\pi_{wZ} = \frac{D}{D-1} * (1 - f_Z^2 - \left(1 - f_Z\right)^2), \quad (2.4)$$

where Z can correspond to population 1 or 2 and D indicates the total depth of coverage, as indicated in equation 2.2. The average π_w was computed as:

$$\pi_w = \frac{\pi_{w1} + \pi_{w2}}{2}, \quad (2.5)$$

The diversity between the populations was calculated as:

$$\pi_b = 1 - (\frac{f_1 + f_2}{2})^2 - (\frac{(1 - f_1) + (1 - f_2)}{2})^2, \quad (2.6)$$

where f_1 and f_2 represent the allele frequency in population 1 and 2, respectively.

2.6 Detecting SNPs under positive selection involved in adaptation in homogeneous and heterogeneous environments

To detect SNPs with changes in allele frequencies consistent with the action of positive selection we considered two tests based on comparing allele frequencies in different selective regimes and replicates. These are based on several assumptions that we detail next (reviewed in [90]). In absence of selection, changes in allele frequencies in populations with the same effective population size (e.g. replicates) are random, due to genetic drift. On the contrary, when there is positive selection due to adaptation to a given environment, the frequency of the beneficial allele will increase for all replicates in the same environment (e.g., increase in all replicates of the homogeneous regime). In homogeneous environments, we expect directional selection (*i.e.*, positive selection) during adaptation to cadmium. For heterogeneous environments, predictions are less clear. Directional selection can also occur, if alleles that are favoured in cadmium are also favoured or neutral in the environment without cadmium. Directional selection is detectable with the statistical tests we used. However, these tests cannot detect other types of selection that can occur (e.g. balancing selection).

To determine whether allele frequencies were consistently different between the control and the homogeneous and heterogeneous environment regimes, we performed two comparisons: 1) control vs cadmium (homogeneous), 2) control vs heterogeneous regimes. Two statistical tests were used to quantify changes in allele frequencies: a general linear model (GLM) with a quasibinomial error structure proposed by Wiberg *et al*. (2017; [91]) and an adapted version of the Cochran-Mantel-Haenszel (CMH)[92], test for time series pool-seq data[93]. For the general linear model, minor and major allele counts were used as response variables (using the cbind function from R) and the regime was used as an explanatory variable, using a quasibinomial error distribution. According to Wiberg *et al*. (2017; [91]), the quasibinomial GLM test produces appropriate false positive and true positive rates. However, here we did not consider the replicates of each regime and so, the minor alleles of the five replicates of a selection regime will be considered as a pool and will be compared with a pool of the minor alleles of the five replicates of the other selection regime that is being compared to. This can lead to potential false positives as, during the experiment, replicates were assessed during different days of the same week (*i.e.* replicates were treated as blocks on the experimental design). CMH is the most widely used statistical method to compare allele frequencies across replicates, with several examples in Wiberg *et al*. (2017; [91]). As the CMH test implemented in the Popoolations software performed

poorly on accurately detecting SNPs[91], another version of the CMH test also implemented for pool-seq data was chosen. Minor and major allele counts of the control and the selected regimes (control *vs.* homogeneous regime and control *vs.* heterogeneous regime) were given as inputs of the cmh.test function (poolSeq package, version 0.3.0; [93]). However, we did not have the initial and final generation of the selection regimes. Thus, we used the assumption that the control population at generation 55 would be a good representative of the initial frequencies present in the initial generations of the other selection regimes. This approach has a major caveat, as the CMH test can, in the absence of time series data, lead to potential false positives, since allele frequencies of generation zero of control and generation 55 of control will not stay constant due to drift. Moreover, possible de novo mutations can also appear in the populations[94].

In order to mitigate this problem, we used a conservative criteria to identify SNPs with significant differences in allele frequencies between regimes, by considering only SNPs that showed statistically significant results after a false discovery rate (FDR) adjustment in both tests. That is, only SNPs with a p-value < 0.05 in both statistical tests (CMH and quasibinomial GLM) after the Benjamini-Hochberg adjustment (FDR; [95]) were considered to show differences in allele frequencies between the control and selective regimes. The FDR adjustment was calculated using function p.adjust from R. Furthermore, we applied an extra criterion to obtain the final list of SNPs potentially under positive selection. By assuming that positive selection would lead to major changes in allele frequencies between the control and selection regimes, we only considered SNPs where the beneficial allele frequency would change from minor (frequency less than 0.5) to major (frequency larger than 0.5). Mean minor allele frequencies were calculated for each regime, considering the mean across the five replicates. Thus, we only maintained SNPs where the minor allele in the control becomes major in the selective regime (green quadrant in Fig. 2.2), or the opposite, *i.e.*, a major allele in the control becomes minor in the selective regime (blue quadrant in Fig. 2.2), meaning that the beneficial allele was at a low frequency in the control.

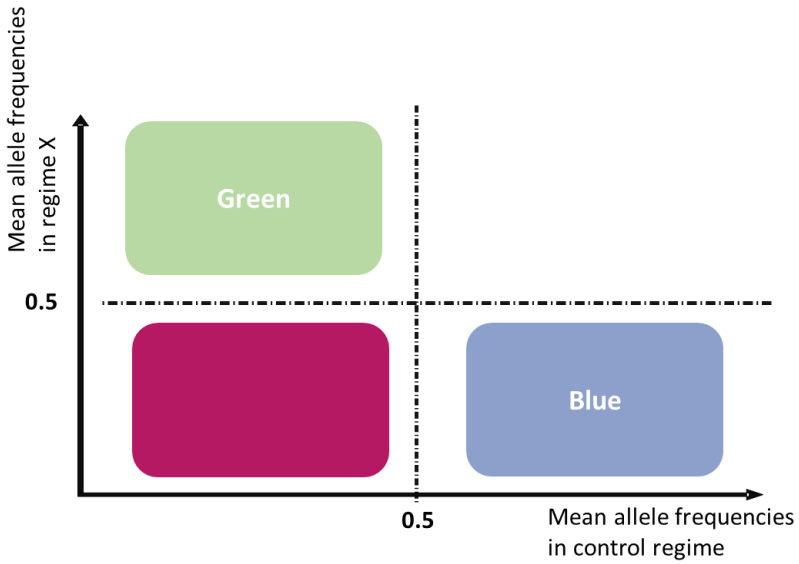

Figure 2.2: Schematic representation of the mean allele frequencies of the candidate SNPs under positive selection. The x-axis represents the mean allele frequency of all replicates of the Control Regime and the y-axis represents the mean allele frequency for all replicates of the selection regime (homogeneous or heterogeneous regimes). SNPs with mean allele frequencies below 0.5 on control (minor allele) and above 0.5 on the X regime (major allele) are represented in green; SNPs with mean allele frequencies above 0.5 on control (minor allele) and below 0.5 on the X regime (major allele) are represented in blue; SNPs with mean allele frequencies below 0.5 on control (minor allele) and below 0.5 on the X regime (major regime) are represented in pink. Only SNPs in the green and blue quadrants were considered to show changes in allele frequencies compatible with positive directional selection.

All analyses were done on R (version 4.2.2, [86]).

2.7 Candidate gene approach to detect SNPs under positive selection

To identify genes that are involved in adaptation in homogeneous or heterogeneous environments with cadmium we also used a candidate gene approach. This strategy has been used since the last decade mostly to identify genes associated with diseases[96]. It consists of the identification of genetic markers for a gene likely to be involved in the phenotype of interest[96].

Metallothioneins are proteins known to play a general role in the metabolism and detoxification of a number of essential and nonessential heavy metals and were identified in several species from microorganisms to fish, plants and mammals[69–72,97], making them a key candidate to show adaptive responses to cadmium. A brief bibliography search was performed using Nordberg, M. & Nordberg, G.F. (2022; [98]) as the base article as it gives a historical review on Metallothioneins and Cadmium Toxicology. Since this review is focused on human health effects, some of the genes mentioned might be specific for mammals or simply not annotated in the *T. urticae* reference genome. Therefore, we searched for articles using the keywords "cadmium" and "gene" on google scholar to look for genes that can be involved in the response to cadmium. Moreover, on the *T. urticae* annotations document[42], we searched for genes of the same gene family and/or with associated GO terms, as is the case for zinc and ion regulated transporters (ZIP), metal-inducible proteins (MIP) and solute carrier group of membrane transport proteins (SLC). The list of genes identified is provided in Supplementary Table 1.

3. Results

Results are divided into two main sections assessing: 1) the genetic diversity of inbred and outbred populations; and 2) detecting candidate SNPs involved in adaptation to cadmium in homogeneous and heterogeneous environments. For both sections, we estimated the levels of expected heterozygosity and the total number of SNPs to quantify the genetic variability of laboratory populations and the three selection regimes. In addition to the referred aims, in section 1) we used different mitochondrial strains to assess the impact of the reference genome on the estimation of the measures of diversity. For section 2) we identified SNPs with significant differences in allele frequencies between the control and selective regimes with cadmium, allowing us to provide a list of candidate SNPs associated with adaptation to cadmium, and to assess whether the same or different SNP show changes in allele frequencies in homogeneous and heterogeneous environments.

3.1 Genetic variability and differentiation of Outbred and Inbred lines in the laboratory

3.1.1 Obtaining a High-Quality SNP data set

Information on the total reads obtained from the sequencing platform, before and after mapping and filtering, as well as depth of coverage, mapping quality and GC content are present in Supplementary table S2. Whole-genome Illumina sequencing resulted in approximately 17.1 million paired-end reads per population when mapped to the *T. urticae* reference genome and approximately 71.9 thousand paired-end reads per population when mapped to the *T. evansi* mitochondrion. *T. urticae* outbred population had the highest percentage of mapped reads (89.62%) and the highest mean mapping quality (44.6 in Phred score). For the *T. evansi* outbred population, a lower proportion of reads (56.56%) were mapped against the *T. urticae* reference genome, with a lower mean mapping quality (31.7 in Phred score). For the inbred lines, Te32 had

the lowest percentage of mapped reads (48.29%) while Te14 had the highest percentage (60.25%). Overall, for the inbred lines tested and the *T. evansi* outbred population, mean mapping quality was higher when mapping to the *T. evansi* mitochondria genome compared to mapping to the reference genome of *T. urticae*. After mapping and filtering, the mean and standard deviation of the depth of coverage were calculated for the nucleus and mitochondria, separately. For the mitochondria, both *T. evansi* and *T. urticae* mitochondrial reference genomes were considered. For the inbred lines, mean depths of coverage ranged from 71.76x to 153.28x and from 82.19x to 155.70x on the mitochondria when mapped to the *T. urticae* and *T. evansi* mitochondria reference genomes, respectively. For the *T. evansi* outbred population, the mean depth of coverage was 284.21x and 209.99x when mapped against the *T. evansi* and *T. urticae* mitochondria reference genomes, respectively. For the *T. urticae* outbred population we found a similar pattern, with a higher coverage when mapping against the mitochondrial reference genome of the same species, *i.e.*, we found a mean depth of coverage of 332.19x and 557.71x when mapped to the *T. evansi* and *T. urticae* mitochondria reference genomes, respectively. When mapping all reads against the *T. urticae* mitochondria reference genome, we found a higher mean depth of coverage for *T. urticae* (557.71x) than for *T. evansi* outbred population (209.99x). For the nuclear genome, only the *T. urticae* genome is available. Mean depths ranged from 20.91-44.21x on the inbred lines. For the outbred populations, we found similar depth of coverage, although they were both mapped to the reference genome of *T. urticae*. Namely, for the *T. evansi* outbred population we obtained a mean depth of coverage of 34.46x, while for the *T. urticae* outbred population it was 35.31x.

After filtering, the total number of SNPs detected in the mitochondria and nuclear genomes, when mapped to the *T. evansi* and *T. urticae* genomes are present in Fig. 3.1 A&B, for all populations. The number of SNPs was higher in the nucleus than in the mitochondria, in particular for the *T. urticae* outbred population (Fig. 3.1). We obtained more SNPs for *T. urticae* than *T. evansi* outbred populations, with a total of 1,443,767 and 635,445 SNPs found in the nucleus, respectively (mapped to the *T. urticae* reference genome). For *T. evansi* inbred lines the total number of SNPs in the nuclear pseudochromosomes was similar to the *T. evansi* outbred population, ranging from 377,112 to 645,773 SNPs. When mapped to the *T. urticae* mitochondria genome, we found less SNPs in *T. urticae* oubred than in *T. evansi* outbred, with 17 and 202 SNPs, respectively. For the inbred lines, the number of SNPs in the mitochondria was similar to the outbred *T. evansi*, with values ranging from 178 to 215 SNPs. When mapping reads to the *T. evansi* mitochondria reference, we found more SNPs for *T. urticae* outbred than for *T. evansi* outbred, with 311 and 36 SNPs, respectively. For the *T. evansi* inbred lines, the number of SNPs tended to be similar or higher than for the *T. evansi* outbred, with values ranging from 39 to 103 SNPs. Thus, a general pattern observed in our data was that the number of SNPs increased when mapping reads to the mitochondria genome of a different species.

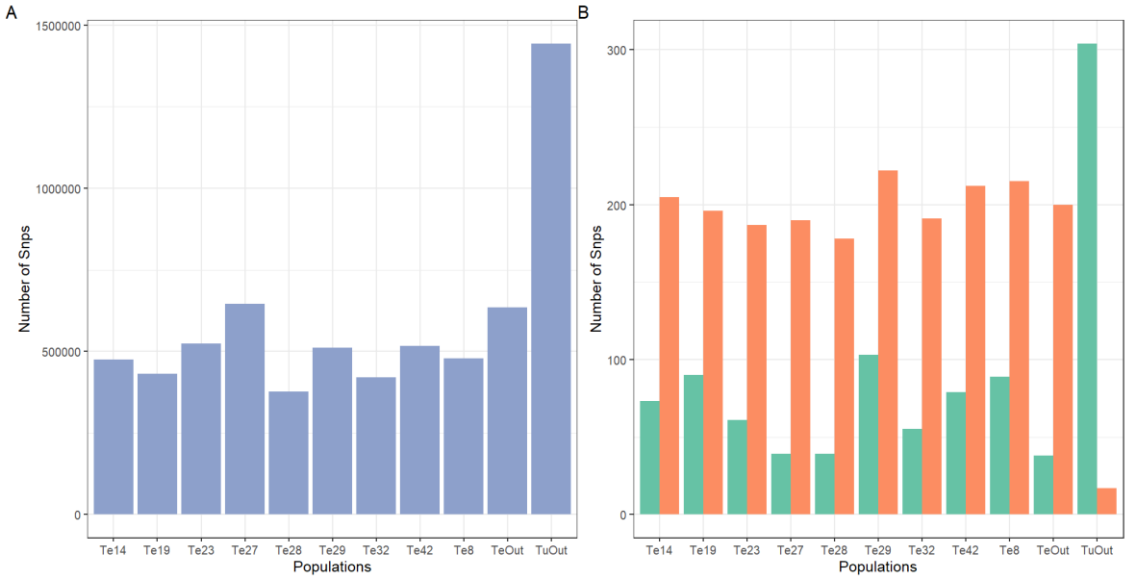

Figure 3.1 Total number of SNPs of the *T. evansi* and *T. urticae* laboratory populations. A) Total number of SNPs in the nuclear genome. B) Number of SNPs on mitochondria, when mapped to the *T. urticae* and *T.evansi* genomes. Values were obtained when mapping to the *T. evansi* mitochondria genome (green bar; MN417333.1) or to the *T. urticae* on nucleus (blue bar) and mitochondria (red bar). TeX: *T. evansi* inbred lines. TeOut: *T. evansi* outbred population. TuOut: *T. urticae* outbred population. In the nucleus (A), *T. urticae* outbred population had the highest number of SNPs detected while *T. evansi* inbred lines and *T. evansi* outbred population had similar number of SNPs detected. In the mitochondria (B), when mapped to the *T. urticae* mitochondria genome, *T. evansi* populations (inbred and outbred) had a higher number of SNPs detected when compared to the *T. urticae* outbred population. When mapped to the *T. evansi* mitochondria genome, opposite results were obtained with the *T. urticae* outbred population having 311 SNPs detected on mitochondria, three times higher than the *T. evansi* populations.

3.1.2 *T. evansi* inbred and outbred populations show few differences in expected heterozygosity across the three pseudo chromosomes

To assess the genetic diversity of populations at the genome-wide level, we estimated mean expected heterozygosity for each population, separately for the nucleus and mitochondria (Fig. 3.2). Since the mutation rate is higher in the mitochondria genome, compared to the nucleous[88], we expected a higher number of SNPs in the mitochondria, and thus a higher average expected heterozygosity in mitochondria. To remove the effect of differences in mutation rates, we calculated the mean of heterozygosity only for the polymorphic sites (*i.e.*, sites with expected heterozygosity equal to zero were removed from the analysis). However, the genetic diversity of the populations considering all sites (*i.e.*, including the sites with expected heterozygosity of zero as well) is present in Supplementary Figure S1.

The *T. urticae* outbred population presented approximately four times higher mean expected heterozygosity in the nucleus than in the mitochondria, when mapped to the *T. urticae* reference genome (Fig. 3.2, blue and red bars), *i.e.*, the mean in the mitochondria is 0.037 and, in the nucleus, it is 0.157. In contrast, the Te23 *T. evansi* inbred line and the *T. evansi* outbred population show similar heterozygosity in the mitochondria and the nucleus (Fig. 3.2, blue and red bars), while Te27 *T. evansi* inbred line showed higher mean expected heterozygosity on mitochondria compared to the nucleus, when reads were mapped to the *T. urticae* reference genome.

In the nucleus, for the *T. urticae* outbred population the mean expected heterozygosity was 0.157 (± 0.159) while for *T. evansi* outbred population it was 0.111 (± 0.109) (Fig. 3.2, blue bar). For inbred lines, the value ranged from 0.099 to 0.156 (Fig. 3.2, blue bar). Overall, *T. evansi* inbred lines have similar mean expected heterozygosity values to the *T. evansi* outbred population.

3.1.3 Effect of mitochondria reference genome

When reads of *T. evansi* inbred lines and outbred population were mapped against the *T. evansi* mitochondria reference genome, the heterozygosity ranged from 0.030 to 0.050 for the inbred lines and was 0.057 (± 0.060) for the outbred *T. evansi* population (Fig. 3.2, green bar). These values are lower than when computing heterozygosity based on reads mapped against the *T. urticae* reference mitochondria genome, where values ranged from 0.089 to 0.127 for the inbred lines, and 0.116 (± 0.108) for the outbred *T. evansi* (Fig. 3.2, red bar). A similar trend was found for the *T. urticae* outbred, *i.e.*, higher heterozygosity was observed when mapping the reads to the reference of a different species, compared to when mapped to the same species. Our results showed a higher heterozygosity of 0.138 (± 0.122) when mapped against the *T. evansi* reference mitochondrial genome (Fig. 3.2, green bar), than when mapped against the *T. urticae* reference mitochondrial genome (heterozygosity of 0.0.037 (± 0.023), Fig. 3.2, red bar). Thus, a higher heterozygosity in the mitochondria was found when mapping against the reference genome of a different species.

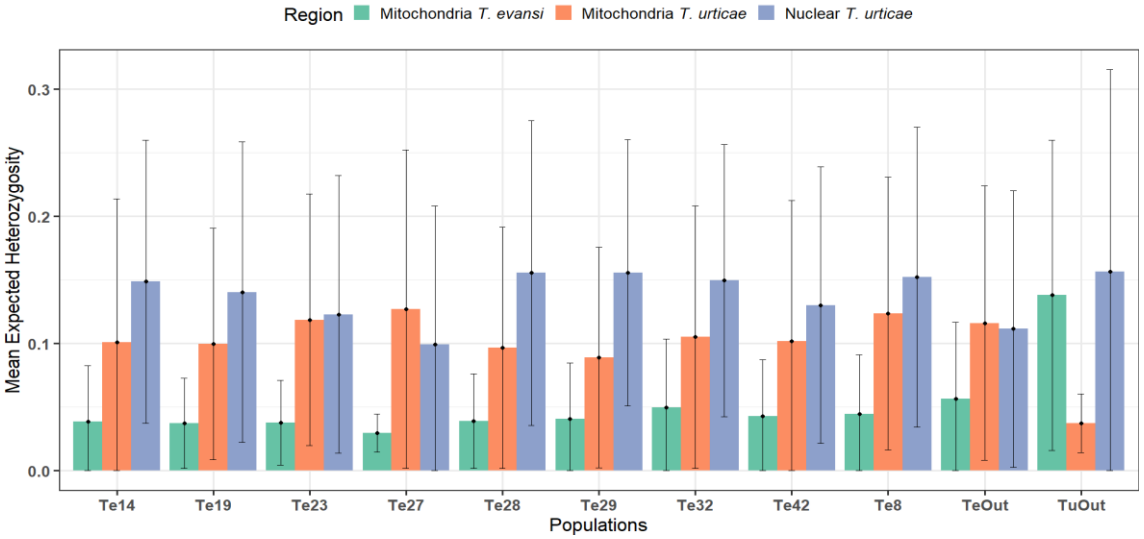

Figure 3.2: Mean Expected Heterozygosity of the *T. evansi* and *T. urticae* laboratory populations. Values were obtained when mapping to the *T. evansi* mitochondria genome (green bar; MN417333.1) or to the *T. urticae* mitochondria genome (red bar) and nuclear genomes (blue bar). Error bars represent the expected heterozygosity standard deviation values across the respective genomic regions, truncated on zero, since expected heterozygosity does not have negative values. TeX: *T. evansi* inbred lines. TeOut: *T. evansi* outbred population. TuOut: *T. urticae* outbred population. The *T. urticae* outbred population presents three times higher mean heterozygosity in the nucleus than in the mitochondria, when mapped to the *T. urticae* reference genome (blue and red bars). On the other hand, *T. evansi* inbred lines have similar mean He values to the *T. evansi* outbred population. Moreover, a higher heterozygosity in the mitochondria was found when mapping against the reference genome of a different species.

We also observe changes in the genetic diversity, total number of SNPs and depth of coverage when mapping reads of the *T. urticae* outbred population to mitochondrial genomes of different strains of *T. urticae* (Fig. 3.3). In this case, mean expected heterozygosity ranged from 0.033 to 0.138 while total number of SNPs ranged from 17 to 304 (Fig. 3.3 A&B). The highest mean expected heterozygosity and highest number of SNPs were obtained when reads from the *T. urticae* outbred population were mapped to the *T. evansi* mitochondria genome.

The mean depth of coverage ranged from 332 to 558 (Fig. 3.3 C), with the lowest value found when mapping *T. urticae* outbred population reads to the *T. evansi* mitochondria genome. Mapping reads from *T. urticae* outbred population to the *T. evansi* mitochondria genome produced many SNPs with low depth (minimum depth was only 65 reads compared to the minimum depth of 253 reads obtained when mapping to the other mitochondria genome). Because depth of coverage has a strong influence on allele frequency estimates and on the ability to detect SNPs, especially on pool-seq[26,87], we added an extra filter on depth of coverage and re-computed the genetic diversity and number of polymorphic sites (represented by Te1 in figure 4). In fact, mean depth values ranged from 332 to 557, with the lowest value corresponding to *T. urticae* reads

mapped to the *T. evansi* mitochondria genome (Fig. 3.3 C). Mean expected heterozygosity decreased from 0.138 to 0.101 (Fig. 3.3 A), while the total number of SNPs detected decreased from 304 to 183 when filtering for minimum depth on the *T. urticae* outbred population reads mapped to the *T. evansi* mitochondria genome (Fig. 3.3 B). Additionally, mean depth increased from 332 to 465 reads. This indicates that many of the SNPs detected without a filter on depth of coverage had a low depth of coverage. Yet, overall, depth of coverage does not seem to be the only factor affecting the genetic diversity estimates, since total number of SNPs detected, when mapping to the *T. evansi* mitochondria genome considering the additional depth of coverage filtering, is still more than three times higher than when mapping to any other *T. urticae* mitochondria strains.

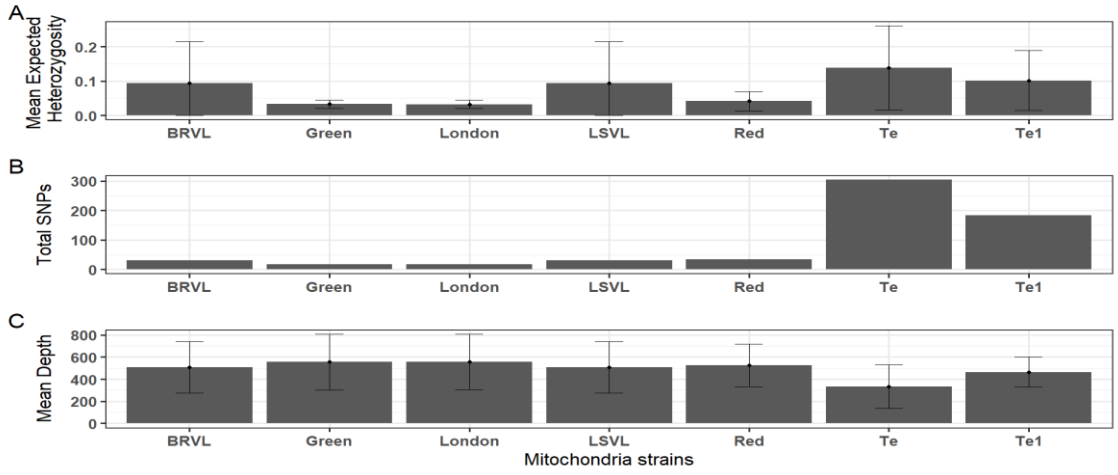

Figure 3.3: Comparisons of different measures of diversity between the different mitochondrial strains of *T. urticae*. Mean Expected Heterozygosity (A), Total number of SNPs (B) and Mean Depth of Coverage (C) of the *T. urticae* outbred population when mapped to the different mitochondrial strains (see Table 2.1 on Methods section 2.3). Error bars represent the expected heterozygosity or mean depth standard deviation values across the respective genomic regions, truncated on zero, since expected heterozygosity does not have negative values. BRVL, Green, London and LSVL are mitochondria strains of *T. urticae* species, Red is the mitochondria of *T. cinnabarinus* and Te is the mitochondria of *T. evansi*. Te1 represents the mitochondria of *T. evansi* filtered for a minimum depth of 254. London is the mitochondria on the reference genome of *T. urticae*. Measures of diversity depend on the mitochondria used for mapping reads. Total number of SNPs detected is more than three times higher using the *T. evansi* mitochondria as a reference genome to map the reads to the *T. urticae* outbred population.

3.1.5 Lack of genetic differentiation between the inbred lines tested

Differentiation between outbred and inbred lines and between the different inbred lines was assessed by estimating average F_{ST} values across the genome. Additionally, the number of polymorphic SNPs across each pair of populations compared was also computed (Tables 2, Supplementary Tables S3 and S4).

Differentiation between the two outbred populations (*T. urticae* and *T. evansi*) is high, both on nucleous and mitochondria genomes (Table 2), with a high number of SNPs that are polymorphic across the two species. However, we observe very low (slightly negative values) F_{ST} values indicating no differentiation between any of the inbred lines and the *T. evansi* outbred population, both on nucleous and mitochondria genomes (Table 2). Additionally, similar results were obtained when we tested the genetic differentiation between all pairs of inbred lines (Supplementary Tables S4 & S5).

Table 3.1: Average F_{ST} across the genome and detected SNPs on nucleus and mitochondria for the laboratory populations. TeOut: *T. evansi* outbred population; TuOut: *T. urticae* outbred population; TeX: inbred lines. Negative F_{ST} values were computed due to the F_{ST} estimator used and should be interpreted as no differentiation

Population	F_{st} between TeOut, Pop		SNPs	
	Nucleus	Mitochondria	Nucleus	Mitochondria
TuOut	0.89	0.96	2332796	682
Te8	-0.0084	0.041	934627	194
Te14	-0.010	0.003	936962	187
Te19	-0.010	0.028	864052	173
Te23	-0.008	0.006	951821	181
Te27	-0.007	0.091	938542	184
Te28	-0.007	0.026	854878	181
Te29	-0.012	0.036	985482	188
Te32	-0.011	-0.006	923983	183
Te42	-0.009	0.020	979146	193

3.2 Detecting candidate regions involved in adaptation to cadmium

3.2.1 Obtention of a High-Quality SNP data set

Information on the total reads obtained from the sequencing platform, for all replicates of the three experimental regimes, before and after mapping and filtering, as well as depth of coverage, mapping quality and GC content are shown in Supplementary Table S5. Whole-genome Illumina sequencing resulted in approximately 22.7 million paired-end reads per population (Supplementary Table S6).

After mapping to the *T. urticae* reference genome, we obtained approximately 19.1 million paired-end reads (83.84%) per replicate of each of the three regimes, with an average mapping quality of 43 to 44 in Phred score (Supplementary Table S5). After filtering to keep only biallelic sites within ranges of depth to minimise the impact of sequencing and mapping errors (see details in methods section 2.6), we obtained mean coverages of 24-33x, 25-28x and 27-30x for control, homogeneous and heterogeneous regimes, respectively for the nucleus. For reads mapped against the mitochondria genome, we obtained a mean coverage of 116-864x, 78-290x and 94-590x for control, homogeneous and heterogeneous regimes, respectively (Supplementary Table S5). The total number of SNPs detected for all replicates of the three selection regimes, for nucleus and mitochondria are shown in Fig. 3.4. On the nucleus, the number of detected SNPs ranged from 627,023 to 677,073, from 587,326 to 741,728 and from 575,456 to 680,708 on control, homogeneous and heterogeneous regimes, respectively, for all five replicates. On mitochondria the number of detected SNPs ranged from 10 to 23, 11 to 55, from 12 to 26 for the control, homogeneous and heterogeneous regimes, respectively, for all five replicates (Fig. 3.4). Mean number of SNPs over the five replicates (red triangles) are different between selection regimes and higher on the nucleus for all selection regimes, compared to the mitochondria (Fig. 3.4). The homogeneous regime presents a higher mean number of SNPs, in comparison to the other regimes, for both mitochondria and nucleus. Additionally, the control and heterogeneous regimes show similar mean numbers of SNPs, for both mitochondria and nucleus.

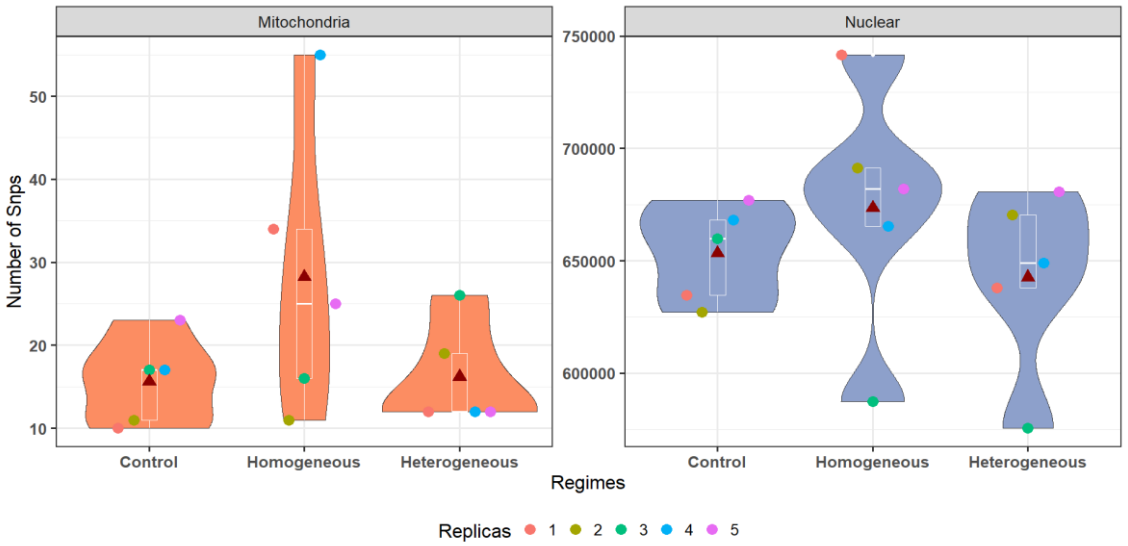

Figure 3.4: Total number of SNPs of the three *T. urticae* selection regimes (control, homogeneous and heterogeneous). Left panel represents the mitochondria and the right panel the nucleus. Boxplots represent variation between five experimental replicates. Triangles represent the mean value for all replicates. Each point with different colours represents the total number of SNPs per replicate and selection regime. The homogeneous regime presents a higher mean number of SNPs for both mitochondria and nucleus, compared to the other two selection regimes, while the control and heterogeneous regimes show a similar mean number of SNPs, on both nucleus and mitochondria.

3.2.2 Similar level of heterozygosity across selection regimes

On mitochondria, mean heterozygosity ranged from 0.028 to 0.044 on the control regime, from 0.032 to 0.047 on the homogenous regime and from 0.029 to 0.044 on the heterogeneous regime, showing similar mean heterozygosity values across the five replicates (red triangles; Fig. 3.5). On the nuclear genome, the same pattern was observed, although there is less overlap between the three selection regimes. Mean heterozygosity across the genome ranged from 0.247 to 0.274 on the control regime, from 0.223 to 0.260 on the homogeneous regime and from 0.233 to 0.252 on the heterogeneous regime (Fig. 3.5). Replicate one of the control regime had mean heterozygosity slightly higher compared to the other replicates (Fig. 3.5). However, the distribution of mean heterozygosity across replicates have overlapping distributions across the three selection regimes for both mitochondria and nucleus.

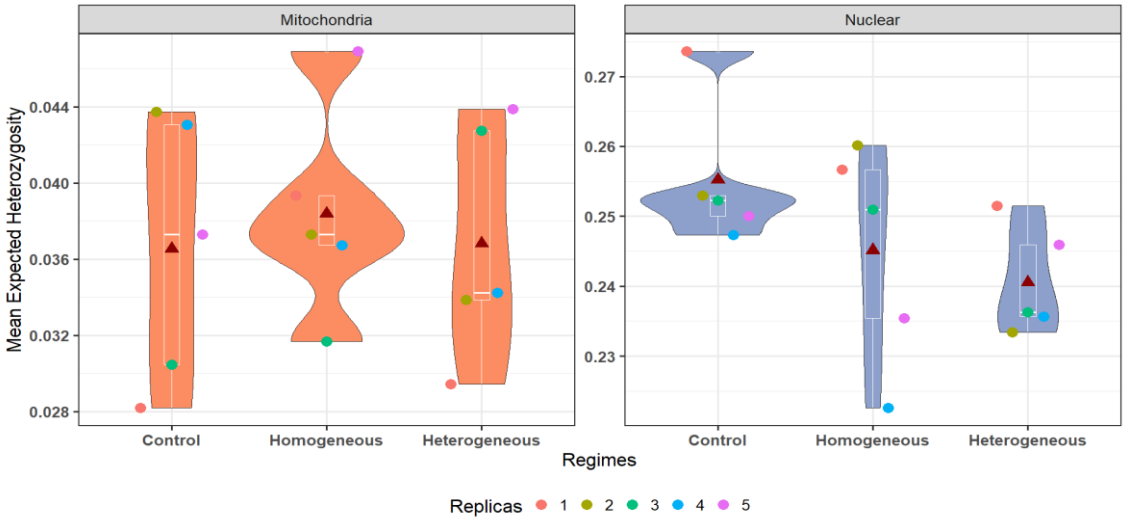

Figure 3.5: Mean Expected Heterozygosity for the three *T. urticae* selection regimes (control, homogeneous and heterogeneous). Left panel represents the mitochondria and the right panel the nucleus. Violin plots represent variation between five experimental replicates. Each point with different colours represents the total number of SNPs per replicate in each selection regime. Triangles represent the mean values for all replicates. We found similar distributions of expected heterozygosity for all selection regimes both on mitochondria and nuclear genomes.

3.2.3 Polygenic response to cadmium during evolution in homogeneous and heterogeneous environments

To detect changes in allele frequencies in response to cadmium, we compared allele frequencies in the last generation between control and high cadmium (homogeneous) and control and heterogeneous regimes. We performed two complementary statistical tests, and only considered as candidate SNPs under positive selection those SNPs with a significant p-value (lower than 0.05) in both tests after FDR adjustment (see methods section 2.6, Table 3.2, Supplemental Figure S1, S2 & S3).

As expected, as the nuclear is much larger than the mitochondria genome, we detected a higher number of SNPs under positive selection in the nuclear genome than in the mitochondria (Table 3.2), which are SNPs potentially involved in the adaptation to cadmium in homogeneous and/or in heterogeneous environments.

In the mitochondria, four SNPs showed significant changes in allele frequencies, comparing homogeneous and control regimes (SR1SR2, Table 3.2 and Supplementary Figure S1).

In the nuclear genome, more SNPs showed significant changes in allele frequencies comparing the heterogeneous regime with the control (224586 SNPs on Table 3.2; Supplementary Figure S3), than when comparing the homogeneous regime with the control (210345 SNPs on Table 3.2; Supplementary Figure S2). However, from the total number of SNPs initially observed, 29% and 23% on Control vs Homogeneous and Control vs Heterogeneous, respectively, showed significant changes in allele frequencies (*i.e.* they were significantly different in both statistical tests, after FDR adjustment, see methods section 2.6). If the control represents the allele frequency of the ancestral population at the start of the experiment, comparing the homogeneous regime with control, 4672 SNPs (7.66 % of the candidate SNPs under positive selection) increased in frequency on the homogeneous regime, while 5196 SNPs (8.51 % of the candidate SNPs under positive selection) decreased in frequency (Table 3.2). On the other hand, we detected 3685 SNPs (7.14 % of the candidate SNPs under positive selection) with increased allele frequency in the heterogeneous regime, while 5660 SNPs (10.96 % of the candidate SNPs under positive selection) decreased in frequency (Table 3.2). Although the percentage of SNPs detected was similar among the two selected regimes, as can be seen in the next sections they are involved in the adaptation to cadmium in different ways (see detail below).

Table 3.2: SNPs detected under positive selection and involved in adaptation to cadmium and/or in heterogeneous environments. To detect positive selection, the total number of SNPs was used, which correspond to the number of sites that were polymorphic for each comparison performed, i.e., the number of sites to which we applied the CMH and GLM tests. Candidate SNPs correspond to SNPs with p-values (from both CMH and GLM tests) lower than 0.05, after a FDR correction. SR1SR2: control vs homogeneous regimes; SR1SR3: control vs heterogeneous regimes

		Nuclear		Mitochondria	
		SR1SR2	SR1SR3	SR1SR2	SR1SR3
Total number of SNPs		210345	224586	19	4
Candidate SNPs under positive selection	Total (% of the detected SNPs that were statistically significant in both tests)	61023 (29 %)	51645 (23 %)	4 (47 %)	0
	SNPs with minor allelic frequencies in the control regime that changed to major allele frequencies in the selection regime (% of the candidate SNPs under positive selection)	4672 (7.66 %)	3685 (7.14 %)	0	0
	SNPs with minor allelic frequencies in the selection regime that were major allele frequencies in the control (% of the candidate SNPs under positive selection)	5196 (8.51 %)	5660 (10.96 %)	0	0
	SNPs with minor allele frequencies in both control and selection regimes	51155 (83.83 %)	42300 (81.91 %)	4	0

Looking at changes in allele frequencies between the control and both homogeneous and heterogeneous regimes, we found that no SNP with a significant change in allele frequencies reached a frequency of 1 (*i.e.* no allele reached fixation on the homogeneous or heterogeneous regimes). In fact, the highest value of allele frequency is approximately 0.8 on both regimes (Fig. 3.6).

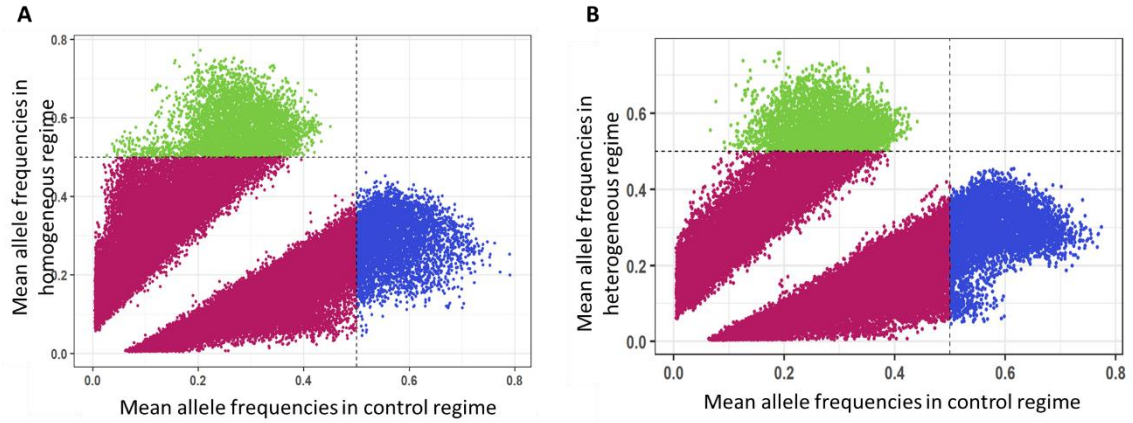

Figure 3.6: Changes in allele frequency between the control and the two selection regimes (homogeneous and heterogeneous), on nuclear genome. Mean allele frequency of the minor allele of the five replicates of the Homogeneous (A) and Heterogeneous (B) regimes were plotted against the mean allele frequency of the minor allele of the five replicates of the Control Regime. Each point represents the mean allele frequency of the five replicas for each candidate SNP, previously identified on Table 3.2. SNPs in green represent an increase in allele frequencies on the selected regime, compared to the control (mean allele frequency on the control regime lower than 0.5 but higher than 0.5 on the selection regime). SNPs in blue represent a decrease in allele frequencies on the selected regime, compared to the control (mean allele frequency on the control regime higher than 0.5 but lower than 0.5 on the selection regime). SNPs in pink represent changes where the allele remains minor on both control and the selected regime (mean allele frequencies lower than 0.5 on both regimes). Vertical dashed line represents the mean allele frequency of 0.5 on the control regime whereas horizontal dashed line represents the mean allele frequency of 0.5 on the selection regime.

Additionally, we found that the SNPs with a significant change in allele frequencies were distributed along the three pseudochromosomes, for the two comparisons of homogeneous vs control, and heterogeneous versus control. Interestingly, SNPs with decreased allele frequency (blue) are evenly distributed across the three pseudochromosomes, while SNPs with increased

allele frequency on the selected regimes (green) are mainly present on the second and third pseudochromosomes (Fig. 3.7; Supplementary Figure S4).

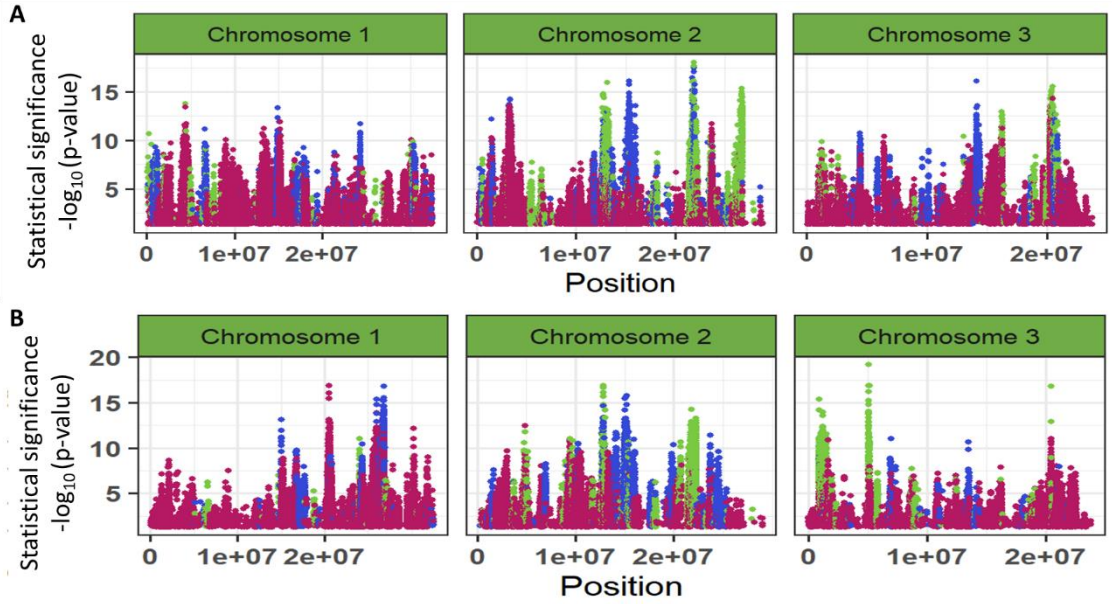

Figure 3.7: Genome-wide distribution of candidate SNPs associated with adaptation of T. urticae to cadmium in homogeneous and heterogeneous regimes, on the nuclear genome. Results of the CMH test are in (A) for the Homogeneous and (B) for the Heterogeneous regimes. Each point represents a candidate SNP, i.e., SNPs with p-values (from both CMH and GLM tests) lower than 0.05, after a FDR correction. Results of the general linear model with a quasibinomial error structure are available in Supplementary Materials S4. SNPs in green represent an increase in allele frequencies on the selected regime, compared to the control. SNPs in blue represent a decrease in allele frequencies on the selected regime, compared to the control. SNPs in pink represent changes where the allele remains minor on both control and the selected regime. SNPs with a significant change in allele frequencies are distributed along the three pseudochromosomes.

Overall, the response to cadmium seems to differ in homogeneous and heterogeneous environments, as there was a small overlap of the regions with a change in allele frequency leading for an allele to become major in the treatment when it was minor in the control or vice-versa (blue or green, Fig. 3.7 and Supplementary Figure 4) comparing Fig. 3.7A with Fig. 3.7B. Additionally, 1 281 SNPs are common between homogeneous and heterogeneous regimes (Fig. 3.8).

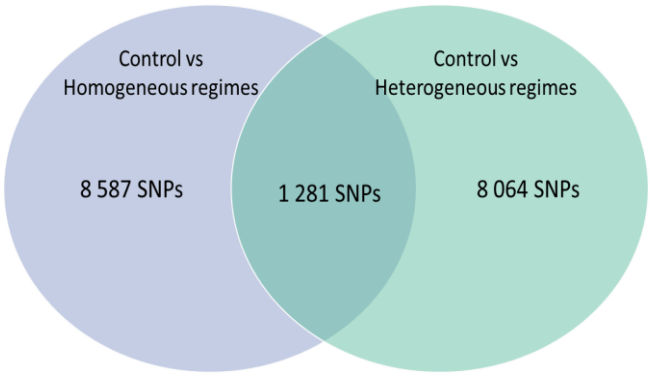

Figure 3.8: Venn diagram of the SNPs with an increase or decrease in allele frequencies on the selected regime (homogeneous or heterogeneous), compared to control. Total SNPs in each comparison correspond to the SNPs represented in green or blue, on Fig.3.7 and Supplementary Materials S4. In total, 9 868 SNPs are associated to the response to cadmium in the homogeneous regime and 9 345 SNPs are associated to the response to cadmium in the heterogeneous regime. Moreover, 1 281 SNPs were found in common between both environments. The small overlap between SNPs showing signs of selection in both regimes, suggests that the response to cadmium was not the same in homogeneous and heterogeneous regimes.

3.2.4 Adaptation to cadmium does not seem to be through protection against metal toxicity

To detect genes under selection we also used a candidate gene approach, focusing on 34 genes present in the *T. urticae* genome (Table S1), involved in metabolism and detoxification of cadmium. From the 34 genes analysed, 13 are in pseudochromosome 1, 14 are in pseudochromosome 2 and seven are in pseudochromosome 3 (Table S1). Two correspond to metallothioneins (metal binding proteins), two are genes codifying for subunits of a calcium transporter (*Cacnα*), two are associated with MIP proteins (metal-inducible proteins), two are associated with Zinc transporters (ZIP family) and 26 are associated with metal ion transporters (SLC family). For most of the candidate genes (including the metallothioneins), no SNPs showed significant changes in allele frequencies consistent in the two statistical tests. However, for six genes relevant results were found: two genes of the SLC family (*SLC12* and *SLC4A*), three genes with GO terms associated with the SLC family and the calcium transporter gene (*Cacnα1G*). For the SLC12 gene, the same four SNPs show an increase in allele frequencies on both homogeneous and heterogeneous regimes, compared to the control regime. For the SLC4A gene, we found two SNPs with decreased allele frequencies only on the heterogeneous regime, compared to the control. Two genes (tetur01g00260, tetur11g01900) with GO terms associated with the SLC family showed decreased allele frequencies only on the homogeneous regime, compared to the control while the other gene (tetur06g04620) showed a decrease in allele frequencies only on the heterogeneous regime (Suplementary Table S1). Interestingly, *Cacnα1G*, a gene coding for a subunit of a voltage-gated T-type calcium channel, had the highest number of SNPs showing significant changes in allele frequencies: 43 SNPs were detected on the homogeneous regime, with a significant increase in allele frequencies, compared to the control, and 14 SNPs were detected on the heterogeneous regime, with a significant increase in allele frequencies compared to the control (Fig. 3.9, Table S1, Supplementary Figure S5).

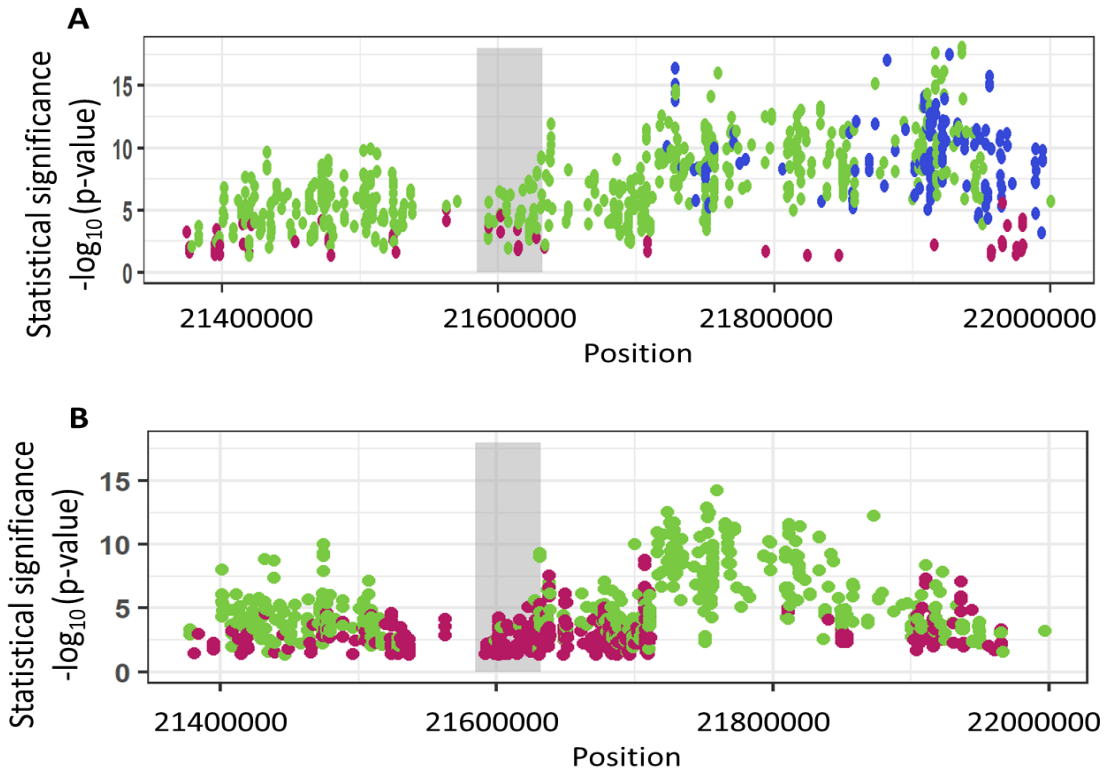

Figure 3.9: Candidate SNPs detected on homogeneous and heterogeneous regimes, compared to the control, for the Cacnα1G gene. Results of the CMH test are in (A) for the Homogeneous and (B) for the Heterogeneous regimes. The gene is highlighted by the grey region. Each point represents a candidate SNP, i.e., SNPs with p-values (from both CMH and GLM tests) lower than 0.05, after a FDR correction. Results of the general linear model with a quasibinomial error structure are available in Supplementary Materials S5. SNPs in green represent an increase in allele frequencies on the selected regime, compared to the control. SNPs in blue represent a decrease in allele frequencies on the selected regime, compared to the control. SNPs in pink represent changes where the allele remains minor on both control and the selected regime. Only SNPs with an increase in selection regime compared with the control were found on both homogeneous and heterogeneous regimes: 43 SNPs on the Homogeneous regime and 14 on the Heterogeneous regime.

3 Discussion

Two long-term and important questions in evolutionary biology are "How do populations adapt to different environments?" and "What are the genomic signatures of the underlying selective process?". Uncovering the genetic basis of adaptation is complex, as adaptation depends on the interaction of several factors[4,8–12]. A powerful approach is to use experimental evolution and/or to perform experiments with laboratory populations under controlled environmental conditions, which is difficult to achieve in natural populations[17,18]. However, such experimental studies often require the establishment of outbred populations in the laboratory, which is usually associated with a founder effect (*i.e.*, reduction in effective population size due to sampling few individuals from nature), implying that the laboratory populations will have lower genetic diversity than natural populations[13]. Experimental studies have been criticized because laboratory populations might not harbour sufficient genetic variability to produce representative responses of the processes occurring in natural populations[19,20]. Therefore, quantifying the genetic diversity of laboratory populations is important to understand if the response observed (or lack of it) is due to initial genetic variability present in the population.

Additionally, most studies involving experimental evolving laboratory populations focus only on comparing changes in a single and constant factor throughout time[34,47,99]. However, we know that, in nature, populations are exposed in heterogeneous environments and that organisms are able to move into different local environments [4,5]. Therefore, understanding how populations adapt to different environmental changes in spatially heterogeneous environments is important.

In this thesis, we first aimed to quantify the genetic variability of outbred and inbred lines of two spider mite species established and maintained in the laboratory. Moreover, by combining experimental evolution and high-throughput sequencing (NGS) technology, we sought to detect genomic signatures of positive selection to a homogeneous environment of host plants watered with cadmium, and to an heterogenous environment where organisms were exposed to both plants watered with and without cadmium. By quantifying the expected heterozygosity genome-wide, our results indicate that *T. evansi* inbred lines had similar genetic diversity to the *T. evansi* outbred population, and that estimates of genetic diversity (expected heterozygosity and number of SNPs) are highly dependent on the choice of the reference genome. Moreover, our results suggest that the response to cadmium was not the same in homogeneous and heterogeneous environments, as only 1 281 SNPs with a statistically significant change in allele frequencies were found in common between the two regimes of homogeneous and heterogeneous environments. We found 9 868 SNPs and 9 345 SNPs with changes in allele frequencies consistent with positive selection that were spread across the genome in the homogeneous and heterogeneous regimes, respectively, suggesting that adaptation to cadmium in homogeneous and heterogeneous environments involves many genes, *i.e.*, a polygenic selective response. Additionally, we found that candidate genes such as metallothioneins that are involved in detoxification of heavy metals, and hence were expected to be under positive selection in environments with cadmium, did not show significant changes in allele frequencies neither in the homogeneous nor the heterogeneous environments with cadmium. However, a subunit of a voltage-gated T-type calcium channel (*Cacnα1G*) showed statistically significant changes in allele frequencies in more than 40 SNPs that responded differently to the homogeneous and heterogeneous environments.

4.1 Genetic variability and differentiation of Outbred and Inbred lines in the laboratory

4.1.1 Genetic diversity and differentiation of inbred and outbred populations

Understanding how genetic diversity affects adaptation is important for predicting how populations respond to changing environments, since the fate of a beneficial mutation, of a population exposed to a new environment, will depend on the selective coefficient and demographic history of the populations[7–12,94]. Quantification of genetic diversity is especially important in populations created and maintained in the laboratory, as most founding events entail a reduction in population size[13]. In small populations, the genetic diversity will decrease at a faster rate compared with larger populations, since the magnitude of genetic drift is proportional to the effective population size[16]. Therefore, keeping populations with large effective population sizes and random mating (outbred populations, Godinho *et al* (2020, [15]) is essential to maintain sufficient genetic variability to produce representative responses of the processes occurring in natural populations. However, there are cases where forcing all individuals in a population to be genetically identical (*i.e.*, inbred lines), is useful, for example, to measure genetic correlations between traits[14,15]. Therefore, we proceeded on estimating expected heterozygosity, as a proxy for the genetic diversity, of the populations of spider mites established and maintained in the laboratory (outbred and inbred lines).

We found that the two outbred populations of the two species (*T. urticae* and *T. evansi*) have expected heterozygosity values between 0.111 to 0.157 on the nucleus when mapped to the *T. urticae* reference genome (Fig. 3.2), which is similar to expected heterozygosity values obtained for outbred laboratory populations of *Drosophila subobscura*[100]. These values of expected hererozygosity were also in accordance with estimates of some natural populations of other haplodiploid species, such as ants and paper wasps. For example, *Formica aquilonia* and *Formica polyctena* populations in Scotland and Switzerland have mean genome-wide values of

heterozygosity around 0.12-0.13, while populations in Finland have slightly higher heterozygosity (0.14 and 0.19, respectively) due to the hybridization in this region[101]. Previous studies on paper wasps of the species *Polistes fuscatus* reported mean genome-wide heterozygosity values lower than 0.1 but in these populations the average positive F_{IS} was 0.296, consistent with a high degree of inbreeding[102]. However, other natural haplodiploid populations of species such as sawflies and honey bees have mean heterozygosity values around 0.2 ([103,104]), a value slightly higher than the ones obtained for both outbred populations. Overall, the genetic diversity estimated for the *T. urticae* and *T. evansi* outbred populations is similar to the ones obtained on natural haplodiploid species, suggesting that our outbred populations harbour enough genetic variability to respond to different selective pressures. However, we should note that having similar genetic diversity levels in the laboratory and nature does not mean that the laboratory populations are able to respond to the same selective pressures as the ones occurring in nature.

We also analysed inbred lines maintained in the laboratory that were established using sib-mating crossings and small population sizes Godinho *et al* (2020, [15]). After 15 generations of sib-mating, inbred lines are expected to have between 93.6 and 95.1% inbreeding coefficient for haplodiploid populations Godinho *et al* (2020, [15]). Thus, we would expect a lower heterozygosity for inbred lines than outbred populations. However, our results based on Pool-seq show that the nine inbred lines tested have similar genome-wide expected heterozygosity to the *T. evansi* outbred population (Fig. 3.2), suggesting that the inbred lines are not true inbred lines and do not represent a fixed genotype.

Moreover, genome-wide differentiation was also estimated between: 1) the two outbred populations *T. evansi* e *T. urticae*, to assess the level of genomic differentiation of the two species; 2) populations of the same species (*T. evansi* outbred population and the *T. evansi* inbred lines) and 3) all possible pairs of inbred lines. We found high levels of differentiation between the two outbred populations (*T. urticae* and *T. evansi*) (F_{ST} nuc = 0.89 and F_{st} mit= 0.96; Table 3.1). This result was higher than other natural populations of haplodiploid species. For example, our results are higher than the F_{ST} value (F_{ST} = 0.66) found between species of paper wasps, *P dorsalis* and *P. metricus*[105], and between the M and C lineages of the honey bees (F_{ST} = 0.52; [106]). This high discrepancy of values might be due to the fact that the articles mentioned did not calculate the F_{ST} measure separately for the nucleus and mitochondria, which might not be totally correct since the nuclear and mitochondria genomes have different evolutionary histories[32]. Additionally, the F_{ST} estimator used in this thesis and the one used in the articles mentioned above are not the same, which might further contribute to the discrepancy between results[107]. In this thesis, the Nei (1973; [89]) estimator was used. On the other hand, the Weir and Cockerman (1984; [108]) estimator was used in most of the articles mentioned above and assumes that the two populations being analysed have experienced identical amounts of drift since splitting[107]. However, according to Bathia *et al* (2013, [107]), although the choice of the estimator has an impact on the values of F_{ST}, the use of average of ratios (not used in this thesis) affect the estimated values of F_{ST} between pairs of populations. Despite differences in estimators of F_{ST}, we know that F_{ST} reflects the amount of drift between populations, and hence the older the separation time between species the larger the F_{ST}, while the higher the gene flow levels and the larger the effective population sizes, the lower the F_{ST} ([109]). Moreover, the establishment of the populations in the laboratory entail a reduction in the population size[13]. Therefore, the high F_{ST} values we observed between the two outbred populations of the two different species could be a result of the demographic history of these species, with low levels of migration and an older separation time, compared to the other haplodiploid species mentioned above. Moreover, a potential founder event or bottleneck related with the establishment of the populations in the laboratory would increase drift, and hence could also increase F_{ST} values.

Even though the levels of differentiation between the two spider mite species was relatively higher than in other populations of haplodiploid species, no differentiation was found between

populations of the same species, with F_{ST} values close to zero for all pairwise comparisons of *T. evansi* (Table 3.1 & Supplementary Tables S3 and S4). In fact, F_{ST} values estimated for nucleus and mitochondria were similar when we compared the *T. evansi* outbred population with all the nine inbred lines and when we compared all pairs of inbred lines. This suggests that, contrary to what was expected[14,15,110] the inbred lines did not fix different genotypes during their establishment. Therefore, we can conclude that the inbred lines tested in this thesis are likely not inbred lines since 1) inbred lines are expected to have heterozygosity values close to zero, and we found expected heterozygosity similar to the outbred population, 2) we found similar total number of SNPs in the lines and outbred population of *T. evansi*, and 3) we found no genetic differentiation between all pairs of inbred lines and between the inbred lines and the *T. evansi* outbred population. These results suggest that very likely there were some problems during the establishment or maintenance of inbred lines in the laboratory. For instance, if the lines were manipulated at the same time on the same bench or were maintained close to each other in the incubator chamber, dispersal and crossing of individuals of different lines could be possible. Also, dispersal of individuals between the outbred and inbred lines would also be possible if the outbred population was manipulated and maintained close to the inbred lines during the 15 generations of the inbreeding process. This dispersal and reproduction of individuals of different lines and with the outbred population would lead to similar values of expected heterozygosity across all populations. Moreover, cross-contamination during the DNA extraction could also explain similar heterozygosity values, if each pool were incorrectly obtained by mixing individuals of inbred and outbred lines.

4.1.2 Impact of reference genome on estimates of diversity

Mapping reads obtained from NGS sequencing technology against a single reference genome is one of the key steps in population genomic studies. Recent studies have been addressing potential biases associated with this approach[111–113]. As most of these errors originate from the genetic differences between the reference and the sequencing data[114–117], using a reference genome from a species different from our input data might not be the best way to perform the subsequent analysis.

In our case, a reference genome is available for the *T. urticae* species but not for the *T. evansi*. Therefore, the outbred and inbred lines of *T. evansi* were mapped to the *T. urticae* reference genome, where only 59.56% of the reads from the *T. evansi* outbred population were mapped to the *T. urticae* nuclear reference genome (Supplementary Table S2). Moreover, for all *T. evansi* populations, the ratio between nucleus and mitochondria mean heterozygosity was different from the *T. urticae* outbred population, where mean heterozygosity was higher on the *T. evansi* mitochondria, compared to the mitochondria of the *T. urticae* outbred population (Fig. 3.2). This might have happened as the reference genome used is from a different species (*T. urticae*). Indeed, the depth of coverage of mitochondria was higher for *T. urticae* outbred population (478.06x; Supplementary material table S2) compared to *T. evansi* outbred population (228.72x; Supplementary material table S2) and inbred lines (71.76x - 153.28x; Supplementary material table S2).

To further explore the impact of the reference genome in the estimates of heterozygosity, we mapped the *T. urticae* and *T. evansi* outbred populations as well as the *T. evansi* inbred lines to a *T. evansi* mitochondria available on NCBI (MN417333.1; Sun *et al*. 2019; [77]). Interestingly, for the *T. evansi* populations, it is possible to recover the ⅓ expected proportion between mitochondria and nuclear heterozygosity when mapped to the *T. evansi* reference (considering a model with neutral markers and assuming sex ratio of 50:50), due to the different effective population size between the mitochondria and the nucleus[118], a result obtained for *T. urticae* outbred population mapped to the *T. urticae* reference genome (Fig. 3.2). Additionally, depth of

coverage has a strong influence on allele frequency estimates, especially on Pool-seq data, and so can be used to identify possible errors associated with mapping to a different species[26,87]. Indeed, mean depth values increase for the *T. evansi* outbred population and for some inbred lines, when mapped to the *T. evansi* mitochondria genome, compared to when mapping to the *T. urticae* mitochondria genome (Fig. 3.3 C & Supplementary Figure S2). However, when looking at *T. urticae* outbred population mapped to the *T. evansi* mitochondria genome, the mean heterozygosity and total number of SNPs more than doubled (Fig. 3.1; 3.2; 3.3 A & B) while the depth of coverage decreased, compared to when mapping with the *T. urticae* mitochondria genome. These results help elucidate the importance of the depth of coverage in estimating allele frequencies because, even though the genetic diversity increases when mapping the reads to a different species, the depth of coverage decreases affecting the allele frequency calculations (Eq. 2.2, Methods section 2.5.1).

Additionally, to further test the impact of the reference genome, the *T. urticae* outbred population was mapped to different mitochondria strains of *T. urticae,* which was compared to results obtained when mapping *T. urticae* outbred population reads to the *T. evansi* mitochondria genome. When comparing the depth of coverage and the total number of SNPs of the *T. urticae* outbred population mapped to the different mitochondria strains, we observed a lower depth of coverage and a higher number of SNPs when mapped to *T. evansi* (Fig. 3.1 B; Fig. 3.3 B & C). Therefore, to understand if the depth of coverage had an impact on the SNP detection, we increased the minimum depth of coverage to a value similar to the one presented by the *T. urticae* mitochondrial strains (Minimum depth = 254). With the additional filter for depth, we found a decrease in the total number of SNPs and mean expected heterozygosity, compared to the *T. urticae* strains. However, the difference in the number of estimated SNPs is still high (more than double when compared to mapping to the *T. urticae* strains). These results show that coverage is not the only feature playing a role in the observed values of expected heterozygosity and the total number of SNPs. Overall, our results suggest that patterns obtained when mapping populations to a different reference genome should be taken with caution.

A possible explanation could be associated with the sequencing of reads (*i.e.*, DNA fragments) obtained through next-generation sequencing. Despite the widespread use of genomic data studies[24,25], it is commonly acknowledged that this process is still prone to error, starting with the production of errors while labelling the nucleotides necessary for the creation of reads, especially on Illumina sequence technologies[119]. In fact, Stoler, N. and Nekrutenko, A. (2021; [119]) compared error rates of several Illumina sequencing platforms, with median values varying from 0.087% in HiSeq X Ten to 0.613% in MiniSeq. Illumina MiSeq (which was used to obtain the pool-seq data for the first part of this thesis) had a median error rate of 0.473%. Although the sequencing machine used had one of the higher median error rates, it is not enough to explain the difference in the values of expected heterozygosity and the number of SNPs obtained when mapping reads to different mitochondria genomes. However, sequencing noises should be removed before any kind of analysis, especially in pool-seq data, even though Chen *et al* (2022; [106]) revealed near identical population structure and genetic diversities between pool-seq and individual whole-genome sequencing. This is because one of the major problems of pool-seq data is the identification of rare variants and the alleles present at low frequencies in the pools can be confounded with the sequencing errors[26,28,87]. Here, we used a series of filters based on depth of coverage and minimum number of reads of each allele to remove, as much as possible, sites affected by sequencing errors.

Moreover, errors associated with sequence alignment can also happen as mapping algorithms have to efficiently find the location of each read while distinguishing between technical sequencing errors and true genetic variation[25,111–113]. Mapping reads to a reference genome is particularly problematic on regions with insertions and deletions (indels), repetitive regions, such as centromeres, telomeres and transposable elements, as reads can align equally well to several positions, reporting low confidence (low mapping quality)[120–122]. In fact, *T. evansi* outbred

population had higher mapping quality (and coverage) when mapping to the *T. evansi* mitochondria genome (same species) compared to when mapping to the *T. urticae* reference genome (whole genome) and to the *T. urticae* London mitochondria genome (Supplementary Table S2). The same pattern was observed in *T. urticae*, with higher mapping quality when mapping was done to the reference of the same species (Supplementary Table S2). These results reinforce the existence of regions in the mitochondria genomes which are different between the different species, leading to possible mapping errors. Chen *et al* (2014, [72]) compared the phylogeny and evolution of several species of *Tetranychus*, including the red and green forms of *T. urticae*, and concluded that these two mitochondria strains have limited divergence and short evolutionary distance. These genomes differ by only 4 bp with some slight structural differences, namely on the start codons of nad3 and cox1 gene Chen *et al* (2014, [72]). However, that study of Chen *et al*. 2014 did not include *T. evansi* mitochondria genome. According to Sun *et al*. 2019 ([77]), *T. evansi* mtDNA has a length of 13064 bp, smaller by 33bp compared to the London strain of *T. urticae*. Comparing the size of mitochondria genomic regions provided by Sun *et al*. (2019) and Chen *et al* (2014, [72]), differences in length are mainly on the tRNAs present in the mitochondria and on the 16S and 12S ribosomal subunits but also on the control region (Chen *et al* 2014, [72]; Sun *et al* 2019, [77]). Differences in the length of the several regions of the different mitochondria can be responsible for the different values in mapping quality obtained for the *T. urticae* mapped to the London and *T. evansi* mitochondria genomes. Additionally, the control region (also called the D-loop region) is known to be generally rich in adenine and thymine nucleotides (A+T rich region) in insects and length variations between closely related taxa are mainly due to tandem repetitions[123]. As mentioned in section 4.1.2), repetitive regions can be particularly problematic during sequence alignment. Moreover, heteroplasmy has been widely observed in the mitochondria of different animals (reviewed in [124]), a phenomenon that describes the different mitochondrial DNA haplotypes on a single individual. The differences in the mitochondria can be due to point mutations or differences in lengths, with the latter being associated with tandem repetitions[123]. In spider mites, heteroplasmy has been identified, where there was a variation in the frequency of resistant haplotypes between mothers and the offspring[75], which might be an additional hurdle in detecting rare variants, especially when working with pool-seq data.

A last explanation for the observed differences between the genetic diversity and total number of SNPs when mapping *T. urticae* outbred population to the different mitochondria can be the nuclear insertions of mitochondrial origin (also known as NumtS;[122]), produced by the transfer of genetic material between the nucleus and the mitochondria. In the human genome, more than 700 NumtS have been identified[84]. In arthropods, reports show that the honey bee nuclear genome has a high density of NumtS (> 1 bp NumtS/kb)[125] but, the identification of NumtS on spider mites is yet to be reported. Since these are highly similar sequences shared between nuclear and mitochondrial DNA, it can introduce an additional bias and cause wrong mapping[85,126]. Thus, specifying only the mitochondria sequence as the reference genome when mapping reads whole genome libraries of *T. urticae* outbred population to the different mitochondria strains and species is not ideal as, without proper NumtS filtering, reads can be mapped to the wrong positions in the mitochondria. Additionally, mapping the *T. urticae* reads to the reference genome including the mitochondria and mapping the reads only to the mitochondria we obtain a different number of reads (146241 and 149122 reads, respectively), suggesting the existence of NumtS on *T. urticae* genome.

Overall, the outcome of mapping *T. urticae* outbred population to different mitochondrial strains of the same species (*T. urticae* mitochondria strains) gives an similar number of SNPs and mean depth of coverage (Fig. 3.3 B). However, when mapping to a different species (in this case, mitochondria genome of *T. evansi*) there are differences, especially on the total number of SNPs (3 times higher, compared to mapping to the same species; Fig. 3.3 B). These results might be explained by the explanations mentioned in the paragraphs above, which are summarised here: 1) sequencing errors that depend on the type of sequencing machine used to produce reads; 2)

mapping errors associated with the algorithm used to perform the alignment; 3) reference genome bias, depending on the reference genome used to map the reads; 4) structural differences in the genome of the two species (e.g., indels, copy-number variation, different length of mitochondria); 5) regions in the nuclear genome similar to regions in the mitochondria genome (e.g. NumtS). More studies are needed to quantify and explain the biases associated with mapping to a reference genome of a different species, such as trying different mapping parameters or removing indels (not performed in this thesis). Moreover, an assembly of the *T. evansi* genome would be ideal to further study the genetic diversity of populations of this species, since only 56.56% of the reads from the *T. evansi* population were mapped to the reference genome of the *T. urticae*.

4.2 Detecting candidate regions involved in adaptation to cadmium

4.2.1 Genetic Diversity of the Homogeneous and Heterogeneous selection regimes

Adaptive evolution is shaped by the genetic diversity of a population. The fate of an adaptive allele depends on several factors, such as changes in the environment through time and space[4,5]. Studies with populations in the laboratory allow for the control of such factors and have been used to study adaptation[17,18]. However, most studies involving experimental evolving populations focus on changes of a single factor throughout time[34,47,54–58,99], despite the probable importance of variable selection in space[5]. Therefore, to understand the impact of spatial heterogeneity on genetic diversity we first estimate the expected heterozygosity harboured by both homogeneous and heterogeneous regimes.

When testing whether evolution in homogeneous and heterogeneous environments affects genomic patterns of diversity, we found that populations of *T. urticae* evolving without cadmium (control) or in environments with cadmium in homogeneous or heterogeneous environments (selection regimes) show similar levels of mean expected heterozygosity at nucleus and mitochondria (Fig. 3.5). The total number of SNPs detected was higher for the homogeneous regime in the mitochondria (Fig. 3.4).

The similarity of mean expected heterozygosity values obtained for the three selection regimes fit the expectation that most SNPs present in the genome are neutral, as predicted by the Neutral[127] and Nearly Neutral[128] theories of molecular evolution (reviewed in [90]). Given that all control and selection regimes populations were derived from the same ancestral population and maintained with the same population size, we would expect similar effective population sizes and hence similar strengths of genetic drift and similar mutation rates in all populations, explaining the similar average genome-wide levels of genetic diversity. Additionally, another possible explanation would be that there was no adaptation to cadmium, with no strong effect on the genomic levels of diversity. If populations had adapted to the new environment, we would expect changes in allele frequencies, due to positive selection, leading to the fixation of beneficial alleles, and as a result a decrease in genetic diversity. This lack of response could be because 55 generations of experimental evolution were not enough for adaptation to occur. However, the possibility of no adaptation to cadmium seems unlikely as 1) phenotypic changes (both on fitness and fecundity) were detected (Sara Magalhães, personal communication), and 2) we found significant changes in allele frequencies across the five replicates consistent with positive selection for several SNPs for both homogeneous and heterogeneous regimes, compared to the control (see section 4.2.3 for details). It is also important to mention that during the experimental evolution of the three selection regimes, some additional measures had to be performed to maintain populations with 200 individuals at each generation. In the initial generations, it was necessary to transfer mites from the previous generation maintained in leafs with no cadmium or from the base population (*T. urticae* outbred population) due to the elevated mortality, especially on the replicates of the homogeneous regime. This was probably due to the high selective pressure caused by the chosen cadmium concentration as high concentrations of cadmium are known to be

toxic for spider mites[68]. This gene flow from the source population could also potentially explain why the genetic diversity is similar among the three regimes. Indeed, selection acts on allele frequencies decreasing the genetic diversity of the population, whereas the introduction of mites coming from populations without cadmium might bring alleles of the ancestral source population to the selected regime, preventing fixation of beneficial alleles and increasing genetic variability.

We detected more SNPs in the mitochondria of the homogenous regime, compared to the other regimes (control and heterogeneous; Figure 3.4), although this did not change the respective mean expected heterozygosity values. Our initial hypothesis was that this was likely due to the maintenance of rare alleles on the mitochondria of populations of the homogeneous regime. To maintain all selection regimes with the same population size it was necessary to introduce new mites from the previous generation or from the base population (*T. urticae* outbred population) especially on the homogeneous regime since, as explained in the previous paragraph, cadmium is known to be toxic to spider mites. These newly introduced mites were not exposed to cadmium and so the introduction of new alleles in the population can affect the estimates of diversity. Overall, our results suggest that evolution on cadmium did not have a strong impact on the genetic diversity of mitochondria.

4.2.2 Methods to detect positive selection

There are several methods that have been used to identify regions under selection, each one of them presenting benefits and drawbacks[129]. Genetic differentiation methods are usually based on many different statistics used in genome scans, calculated with a windows size approach and choosing SNPs that exhibit a significant value above a pre-stipulated threshold of differentiation[130]. These methods assume that all windows share the same sampling distribution, which is untrue as it is known, for example, that different regions of the genome have different recombination rates[131]. Another drawback is if the genetic basis of adaptation is explained mostly by alleles of small effect, those will be hardly detected by differentiation measures such as F_{ST}, as the values will be low[132], not passing a pre-stipulated threshold.

In this thesis, we tried to overcome the problems associated with choosing a subjective window size and threshold by looking at changes in allele frequencies in each SNP, possibly associated with cadmium. Therefore, to detect signs of selection we applied two statistical tests (CMH and a general linear model) with an FDR adjustment to the SNPs with equal or lower than 0.05 significance in both statistical tests. The CMH test[92,93] looks for a similar pattern in the allele frequencies between all five existent replicates of each selection regime, considering the control as the generation zero. We are aware that considering generation 55 of the control regime as the generation zero when comparing changes in allele frequencies between the other two regimes, comes with some drawbacks as allele frequencies will not be the same as they were on the actual generation zero, due to drift and possibly de novo mutations. Therefore, we also implemented a general linear model test with a quasibinomial distribution recommended by Wiberg *et al*. (2017; [91]), where minor and major allele counts were used together as dependent variables to account for the different depth of coverage in each SNP and the regime was used as an explanatory variable. We used an approach similar to that done by Seabra *et al* 2018 ([100]), where they followed allele frequency changes between two different time points at candidate SNPs, to have a more detailed characterization of the adaptive process at the genomic level. We adapted the approach since we do not have time series data, comparing the allele frequency changes from minor to major (and vice versa) between the control and the two selection regimes (in green and blue respectively, Fig. 3.7). That is, we only focus on alleles that changed from minor to major or vice versa. This means that one SNP with mean allele frequency on all five replicates of 0.4 on control and 0.6 on the homogeneous regime (or heterogeneous regime) would be considered under positive selection, while other SNP with mean allele frequency on all five replicates of 0.1 on control and 0.4 on the homogeneous regime (or heterogeneous regime) would not be considered, despite the second SNP

having a higher change in allele frequencies between regimes (0.4 - 0.1) compared to the first SNP (0.6-0.4). In the future, performing simulations of population evolving according to the conditions maintained in the laboratory could be used to verify if the observed genomic patterns (allele frequencies) fit neutral theoretical predictions, allowing a more accurate way to detect SNPs under positive selection and involved in adaptation to cadmium. Furthermore, obtaining sequence data at different time points, and especially from the initial generations would be ideal, and in line with evolve and reseq approaches, as performed by Seabra *et al* 2018 ([100]).

4.2.3 Genome-wide patterns of adaptation in homogeneous and Heterogeneous environments

Understanding the genetic basis of local adaptation is one of the major goals of evolutionary biology. Recent advances in high-throughput sequencing and increases in computational power allow us to have access to more information on genomic data for different organisms[24,25]. While some studies have shown that adaptation can be highly specific for a single gene, as the variation in colour-specific adaptation on deer mice of Nebraska[31] and variation in the Soay sheep horn[32] other studies have shown that adaptation can have a polygenic basis such as adaptation to drought in *Brassica rapa*[34]. Even though the genetic architecture of adaptation to cadmium is not well known, we expected adaptation to have a highly polygenic basis as it is known that heavy metals can affect the gene expression of several genes involved in cytoplasmic metal influx, nicotiannamine/thiol biosynthesis pathways and other genes associated with membrane transport proteins, as they are the first barrier between the environment and the individuals[133–136].

Overall, we detected 61023 candidate SNPs (corresponding to 29% of the total number of SNPs found), dispersed on the three pseudochromossomes in the homogeneous regime, compared to the control (Table 3.2). This result suggests a polygenic response on the adaptation to cadmium, as we were expecting, influenced by many genes, each one of them possibly having small effects[132], *i.e.*, our results are in agreement with the infinitesimal model proposed by Fisher (1919; [137]). This polygenic response has been detected in several studies on adaptation to complex traits[33], such as height in humans. Moreover, it was also detected in the mycorrhizal fungus *Suillus luteus* driven by soil heavy metal contamination[133], one of the few studies investigating genomic signatures of adaptation on heavy metals. However, the authors applied measures of differentiation to detect selection, such as F_{ST} and d_{xy}, which were not used in this thesis. Moreover, the studied sites sampled were geographically close to each other, lacking clear barriers to gene flow. However, Bazzicalupo, A.L. *et al* (2020;133) provide a useful list of gene annotations with different heavy metal tolerance strategies that were not taken into consideration in this thesis but that will be useful in the future, especially a cadmium chelating agent in the cytosol (S-adenosyl-l-methionine-dependent methyltransferase). Other studies applied transcriptomics to determine differential gene expression associated with heavy metal tolerance which sustains the hypothesis of a polygenic response to cadmium[70,138,139], indicating possible candidate genes that will be interesting to look into in more detail for the future using our data.

Environmental heterogeneity has been hypothesized to help maintaining genetic variation of populations if alternative alleles are favoured in different environments[5]. However, the effect of heterogeneity depends on the genetic basis of ecological adaptation (in heterogeneous environments), which is unknown. To our knowledge, only Huang, Y. *et al* (2014; 54) tried to identify genome-wide patterns of genetic variation, not only in temporally variable environments but also in changes of two different variables in space. Therefore, our initial expectations for homogeneous versus heterogeneous environments were not clearly defined. Nevertheless, we found 29 % of SNPs detected to be under positive selection on the homogeneous environment, which is similar to 23 % of SNPs detected to be under positive selection on the heterogeneous

environment (Table 3.2). This could suggest a similar response to the selection in homogeneous and heterogeneous environments.

The candidate SNPs found along the three pseudochromosomes when comparing the heterogeneous environment with the control, also suggest a polygenic response on the adaptation in heterogeneous environments (Fig 3.7; Supplementary Figures S2, S3 & S4). Huang, Y. *et al* (2014; 54) explored three possibilities, not mutually exclusive, on the effect of spatial heterogeneity on allele frequencies in their study on *Drosophila melanogaster* experimental evolving populations adapting to four selection regimes, including a spatially heterogeneous treatment in which the populations were exposed to cadmium and salt-enriched environments: 1) environmental heterogeneity sustains elevated levels of genetic variation through antagonistic pleiotropy between environments, which can result in balancing selection; 2) genetic diversity can increase with environmental heterogeneity when two alleles can be selectively neutral in one environment but with different fitness effects on another environment (*i.e.* conditional neutrality); 3) heterogeneity may select for alleles different from those favoured in either single habitat. Huang, Y. *et al* (2014; 54) concluded that their results are more consistent with an antagonistic pleiotropy, but they also find indirect evidence for selection on loci exclusively favoured on the heterogeneous environments. In our results, we are not able to distinguish between antagonistic pleiotropy from the conditional neutrality, as our methods only detect positive selection. However, we do find evidence for unique sites being favoured exclusively by heterogeneity (Fig. 3.8), a result also found in Huang, Y. *et al* (2014; 54). In fact, only 10.9% of the SNPs with a significant change in allele frequencies are common between homogeneous and heterogeneous regimes (Fig. 3.8), meaning that cadmium is not the only factor playing a role in the adaptation of spider mites in heterogeneous environments. This is a conclusion consistent with results found in Huang, Y. *et al* (2014; 54). In the future, it will be interesting to look more specifically at the genes that are shared between the homogeneous and heterogeneous environments to better understand the evolutionary response to the adaptation to cadmium.

4.2.4 Adaptation to cadmium is not through protection against metal toxicity

Metallothioneins (MT) are cysteine-rich metal binding proteins present in all vertebrates but also in invertebrates, plants and microorganisms[140,141]. Although they have been long associated with metal metabolism[140] and protection against metals such as cadmium[98], their primary biological role remains unresolved[141]. In spider mites, two MT isoforms are known[41,42]. In our results, no association to cadmium was found for both MTs, a result also shown in some studies with mice (reviewed in [141]), indicating that factors other than MTs are important in dealing with cadmium toxicity. We looked into 32 candidate genes other that MTs, specifically genes associated with MIP proteins (metal-inducible proteins), Zinc transporters (ZIP family), metal ion transporters (SLC family) and two are genes codifying for subunits of a calcium transporter (*Cacnα*). Although some studies found an association between MIP, ZIP, SLC genes on cadmium uptake[71,136,142], our results suggest the absence of positive selection in these genes (Supplementary Table S1). However, the results are consistent with an association of a calcium transporter gene subunit *Cacnα1G* on the adaptation to cadmium (Fig. 3.9, Supplementary Table 1, Supplementary Figure S5). This gene has been previously studied by Leslie, E. *et al.* (2006, [143]) on metallothionein knockout cells. The authors found a decrease in expression of *Cacnα1G* protects the cells from cadmium exposure by limiting its uptake. Additionally, it is known that the major route for the uptake of cadmium ions in insects is via calcium channels[144,145], reinforcing the importance of this *Cacnα1G* gene on the adaptation to cadmium. Although further studies must be performed to validate the potential role of this calcium transporter, our results suggest and agree with a hypothesis that adaptation to cadmium is not through protection against metal toxicity, as expected, but by limiting its uptake. In the future, we want to look at more genes that were found associated with metal decontamination[133–135,146]. Additionally, and even though no association was

found on mitochondria, nuclear genes regulating mitochondria could be involved in adaptation to cadmium, as the effects of cadmium on mitochondria damage have been previously described[147].

5. Conclusions

In this thesis we sought to identify the genetic basis of the adaptation of cadmium in heterogeneous environments. We started by estimating the genetic diversity of laboratory populations and comparing outbred and inbred lines.

Overall, the genetic diversity of the *T. urticae* and *T. evansi* outbred populations are equivalent to haplodiploid natural populations, suggesting that our outbred populations harbour enough genetic variability to respond to different selective pressures. However, we found that the nine inbred lines also tested had similar expected heterozygosity to the *T. evansi* outbred population, leading us to conclude that those are not inbred isogenic lines. This could have been due to contamination resulting in gene flow between different inbred lines and the outbred population. By estimating the expected heterozygosity of the populations tested in this thesis, we sought to explore more in-depth problems associated with mapping using genomes from different species. The genetic diversity results that we obtained might be explained by: 1) sequencing errors due to sequencing machine used, 2) mapping errors due to alignment algorithm used, 3) reference genome bias, 4) structural differences between the genome of the two species, 5) regions in the nuclear genome similar to regions in the mitochondria (e.g. NumtS). Our results suggest that more studies are required to assess the impact of the choice of the reference genome, as we show it affects estimates of genetic diversity and number of SNPs.

Moreover, to identify genomic regions of adaptation to cadmium on *T. urticae* we looked at consistent changes in allele frequencies on all five replicates of each selection regime, assuming that the control regime represented the initial allele frequency. Our results show that the adaptation of *T. urticae* to cadmium in homogeneous environments and the adaptation in heterogeneous environments both seem to have a polygenic basis, as we found many genomic regions with statistically significant results in both statistical tests used (CMH and GLM with a quasibinomial error distribution) after correcting for a false discovery rate. It was interesting to notice that only 10.9 % of the candidate genes are shared between homogeneous and heterogeneous environments, suggesting that: 1) only 11,154 (10.9 % of the total number of SNPs) are important for the adaptation to cadmium, assuming that only SNPs that respond in both homogeneous and heterogeneous environments are involved in cadmium; 2) adaptation in heterogeneous environments may select for alleles different from those favoured in the different homogeneous environments. Additionally, we found that candidate genes and especially metallothioneins did not show significative changes in allele frequencies in response to cadmium, as predicted, but a voltage-gated T-type calcium channel (*CACNA2D3*) responded differently to the homogeneous and heterogeneous environments.

In the future, we would like to refine the pipeline on the detection of SNPs, more specifically removing indels since they can be problematic when mapping reads to a reference genome. Moreover, we aim to improve the method to detect positive selection by explicitly taking into account drift, simulating populations evolving according to the conditions maintained in the laboratory. The simulations can be useful to detect SNPs under selection, also allowing to simulate expected patterns due to other types of selection (e.g., balancing selection). Moreover, comparing the last generation of both selection regimes (Homogeneous and Heterogeneous) we are able to find significant changes in allele frequency throughout time, consistent in all five replicates. Thereafter, we can compare both selection regimes with the control regime, retaining only SNPs with minor allele frequencies increasing in all five replicas and keeping SNPs with larger changes in allele frequency than expected by drift alone. Additionally, to better understand

the genetic basis of adaptation to cadmium we want to specifically know the location and function of the genes with candidate SNPs found common in both homogeneous and heterogeneous environments. Lastly, to better understand the impact of the spatial variation on adaptation we also want to investigate more in-depth the SNPs that show signs of selection exclusively in the heterogeneous regime.

6. References

1. Fisher, R., A. *The genetical theory of natural selection*. (Oxford Univ. Press, 1930).
2. Orr, H. A. Theories of adaptation: what they do and don't say. *Genetica* **123**, 3–13 (2005).
3. Allen, M. R. *et al.* Framing and Context. in *Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty* (Cambridge University Press, 2018).
4. Hedrick, P. W. Genetic Polymorphism in Heterogeneous Environments: A Decade Later. *Annu. Rev. Ecol. Syst.* **17**, 535–566 (1986).
5. Hedrick, P. W. Genetic Polymorphism in Heterogeneous Environments: The Age of Genomics. *Annu. Rev. Ecol. Evol. Syst.* **37**, 67–93 (2006).
6. Barrett, R. D. H. & Hoekstra, H. E. Molecular spandrels: tests of adaptation at the genetic level. *Nat. Rev. Genet.* **12**, 767–780 (2011).
7. Hedrick, P. W. Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. *Mol. Ecol.* **22**, 4606–4618 (2013).
8. Olson-Manning, C. F., Wagner, M. R. & Mitchell-Olds, T. Adaptive evolution: evaluating empirical support for theoretical predictions. *Nat. Rev. Genet.* **13**, 867–877 (2012).
9. Blanquart, F., Gandon, S. & Nuismer, S. L. The effects of migration and drift on local adaptation to a heterogeneous environment. *J. Evol. Biol.* **25**, 1351–1363 (2012).
10. Nevo, E. Genetic variation in natural populations: Patterns and theory. *Theor. Popul. Biol.* **13**, 121–177 (1978).
11. Savolainen, O., Lascoux, M. & Merilä, J. Ecological genomics of local adaptation. *Nat. Rev. Genet.* **14**, 807–820 (2013).

12. Latta, R. G. Differentiation of Allelic Frequencies at Quantitative Trait Loci Affecting Locally Adaptive Traits. *Am. Nat.* **151**, 283–292 (1998).

13. Santos, J. *et al.* From nature to the laboratory: the impact of founder effects on adaptation. *J. Evol. Biol.* **25**, 2607–2622 (2012).

14. Mackay, T. F. C. *et al.* The Drosophila melanogaster Genetic Reference Panel. *Nature* **482**, 173–178 (2012).

15. Godinho, D. P. *et al.* Creating outbred and inbred populations in haplodiploids to measure adaptive responses in the laboratory. *Ecol. Evol.* **10**, 7291–7305 (2020).

16. Willi, Y., van Buskirk, J. & Hoffmann, A. A. Limits to the Adaptive Potential of Small Populations. *Annu. Rev. Ecol. Evol. Syst.* **37**, 433–458 (2006).

17. Kawecki, T. J. & Ebert, D. Conceptual issues in local adaptation. *Ecol. Lett.* **7**, 1225–1241 (2004).

18. Kawecki, T. J. *et al.* Experimental evolution. *Trends Ecol. Evol.* **27**, 547–560 (2012).

19. Calisi, R. M. & Bentley, G. E. Lab and field experiments: Are they the same animal? *Horm. Behav.* **56**, 1–10 (2009).

20. Melvin, S. D. & Houlahan, J. E. Tadpole mortality varies across experimental venues: do laboratory populations predict responses in nature? *Oecologia* **169**, 861–868 (2012).

21. Lopes, M. S. *et al.* The use of microsatellites for germplasm management in a Portuguese grapevine collection. *Theor. Appl. Genet.* **99**, 733–739 (1999).

22. Kauer, M. O., Dieringer, D. & Schlötterer, C. A Microsatellite Variability Screen for Positive Selection Associated With the "Out of Africa" Habitat Expansion of Drosophila melanogaster. *Genetics* **165**, 1137–1148 (2003).

23. Väli, Ü., Einarsson, A., Waits, L. & Ellegren, H. To what extent do microsatellite markers reflect genome-wide genetic diversity in natural populations? *Mol. Ecol.* **17**, 3808–3817 (2008).

24. Ellegren, H. Genome sequencing and population genomics in non-model organisms. *Trends Ecol. Evol.* **29**, 51–63 (2014).

25. Fonseca, N. A., Rung, J., Brazma, A. & Marioni, J. C. Tools for mapping high-throughput sequencing data. *Bioinformatics* **28**, 3169–3177 (2012).

26. Schlötterer, C., Tobler, R., Kofler, R. & Nolte, V. Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. *Nat. Rev. Genet.* **15**, 749–763 (2014).

27. Futschik, A. & Schlötterer, C. The Next Generation of Molecular Markers From Massively Parallel Sequencing of Pooled DNA Samples. *Genetics* **186**, 207–218 (2010).

28. Anand, S. *et al.* Next Generation Sequencing of Pooled Samples: Guideline for Variants' Filtering. *Sci. Rep.* **6**, 33735 (2016).

29. Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C. & Clark, A. G. Recent and ongoing selection in the human genome. *Nat. Rev. Genet.* **8**, 857–868 (2007).

30. Fan, S., Hansen, M. E. B., Lo, Y. & Tishkoff, S. A. Going global by adapting local: A review of recent human adaptation. *Science* **354**, 54–59 (2016).

31. Linnen, C. R. *et al.* Adaptive Evolution of Multiple Traits Through Multiple Mutations at a Single Gene. *Science* **339**, 1312–1316 (2013).

32. Johnston, S. E. *et al.* Life history trade-offs at a single locus maintain sexually selected genetic variation. *Nature* **502**, 93–95 (2013).

33. Sella, G. & Barton, N. H. Thinking About the Evolution of Complex Traits in the Era of Genome-Wide Association Studies. *Annu. Rev. Genomics Hum. Genet.* **20**, 461–493 (2019).

34. Johnson, S. E., Tittes, S. & Franks, S. J. Rapid, nonparallel genomic evolution of Brassica rapa (field mustard) under experimental drought. *J. Evol. Biol.* **36**, 550–562 (2023).

35. Thomas, R. H. Mites as models in development and genetics. in *Acarid Phylogeny and Evolution: Adaptation in Mites and Ticks* (Springer Netherlands, 2002).

36. Jeppson, L. R., Baker, E. W. & Keifer, Hartford H. *Mites Injurious to Economic Plants*. (1975).

37. Bamel, K. & Gulati, R. Biology, population built up and damage potential of Red spider mite, *Tetranychus urticae* Koch (Acari: Tetranychidae) on marigold: A review. *J. Entomol. Zool. Stud.* **9**, 547–552 (2021).

38. Walter, D. E. & Proctor, H. C. *Mites: Ecology, Evolution & Behaviour: Life at a Microscale*. (Springer Netherlands, 2013).

39. Migeon, A., Nouguier, E. & Dorkeld, F. Spider Mites Web: A comprehensive database for the Tetranychidae. in *Trends in Acarology* (eds. Sabelis, M. W. & Bruin, J.) 557–560 (Springer Netherlands, 2023).

40. Adesanya, A. W. *et al.* Mechanisms and management of acaricide resistance for *Tetranychus urticae* in agroecosystems. *J. Pest Sci.* **94**, 639–663 (2021).

41. Grbić, M. *et al.* The genome of Tetranychus urticae reveals herbivorous pest adaptations. *Nature* **479**, 487–492 (2011).

42. Wybouw, N. *et al.* Long-Term Population Studies Uncover the Genome Structure and Genetic Basis of Xenobiotic and Host Plant Adaptation in the Herbivore *Tetranychus urticae*. *Genetics* **211**, 1409–1427 (2019).

43. Boubou, A. *et al.* Test of Colonisation Scenarios Reveals Complex Invasion History of the Red Tomato Spider Mite *Tetranychus evansi*. *PLOS ONE* **7**, e35601 (2012).

44. Boubou, A., Migeon, A., Roderick, G. K. & Navajas, M. Recent emergence and worldwide spread of the red tomato spider mite, *Tetranychus evansi*: genetic variation and multiple cryptic invasions. *Biol. Invasions* **13**, 81–92 (2011).

45. Belliure, B., Montserrat, M. & Magalhaes, S. Mites as models for experimental evolution studies. *Acarologia* **50**, 513–529 (2010).

46. Akey, J. M. *et al.* Population History and Natural Selection Shape Patterns of Genetic Variation in 132 Genes. *PLOS Biol.* **2**, e286 (2004).

47. Eoche-Bosy, D. *et al.* Genome scans on experimentally evolved populations reveal candidate regions for adaptation to plant resistance in the potato cyst nematode *Globodera pallida*. *Mol. Ecol.* **26**, 4700–4711 (2017).

48. Lenski, R. E. & Travisano, M. Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. *Proc. Natl. Acad. Sci.* **91**, 6808–6814 (1994).

49. Ferea, T. L., Botstein, D., Brown, P. O. & Rosenzweig, R. F. Systematic changes in gene expression patterns following adaptive evolution in yeast. *Proc. Natl. Acad. Sci.* **96**, 9721–9726 (1999).

50. Schulte, R. D., Makus, C., Hasert, B., Michiels, N. K. & Schulenburg, H. Multiple reciprocal adaptations and rapid genetic change upon experimental coevolution of an animal host and its microbial parasite. *Proc. Natl. Acad. Sci.* **107**, 7359–7364 (2010).

51. Bubli, O. A., Imasheva, A. G. & Loeschcke, V. Selection for Knockdown Resistance to Heat in *Drosophila Melanogaster* at High and Low Larval Densities. *Evolution* **52**, 619–625 (1998).

52. Magalhães, S., Fayard, J., Janssen, A., Carbonell, D. & Olivieri, I. Adaptation in a spider mite population after long-term evolution on a single host plant. *J. Evol. Biol.* **20**, 2016–2027 (2007).

53. Johnson, S. E., Hamann, E. & Franks, S. J. Rapid, parallel evolution of field mustard (*Brassica rapa*) under experimental drought. *Evolution* **76**, 262–274 (2022).

54. Hallsson, L. R. & Björklund, M. Selection in a fluctuating environment leads to decreased genetic variation and facilitates the evolution of phenotypic plasticity. *J. Evol. Biol.* **25**, 1275–1290 (2012).

55. Ketola, T. *et al.* Fluctuating Temperature leads to evolution of thermal generalism and preadaptation to novel environments. *Evolution* **67**, 2936–2944 (2013).

56. Manenti, T., Loeschcke, V., Moghadam, N. N. & Sørensen, J. G. Phenotypic plasticity is not affected by experimental evolution in constant, predictable or unpredictable fluctuating thermal environments. *J. Evol. Biol.* **28**, 2078–2087 (2015).

57. Santos, M. A. *et al.* No evidence for short-term evolutionary response to a warming environment in Drosophila. *Evolution* **75**, 2816–2829 (2021).

58. Santos, M. A. *et al.* Past history shapes evolution of reproductive success in a global warming scenario. *J. Therm. Biol.* **112**, 103478 (2023).

59. Huang, Y., Wright, S. I. & Agrawal, A. F. Genome-Wide Patterns of Genetic Variation within and among Alternative Selective Regimes. *PLOS Genet.* **10**, e1004527 (2014).

60. War, A. R. *et al.* Mechanisms of plant defense against insect herbivores. *Plant Signal. Behav.* **7**, 1306–1320 (2012).

61. Poschenrieder, C., Tolrà, R. & Barceló, J. Can metals defend plants against biotic stress? *Trends Plant Sci.* **11**, 288–295 (2006).

62. Rascio, N. & Navari-Izzo, F. Heavy metal hyperaccumulating plants: How and why do they do it? And what makes them so interesting? *Plant Sci.* **180**, 169–181 (2011).

63. Vickerman, D. B. & Trumble, J. T. Feeding preferences of *Spodoptera exigua* in response to form and concentration of selenium. *Arch. Insect Biochem. Physiol.* **42**, 64–73 (1999).

64. Bañuelos, G. S. *et al.* Biotransfer Possibilities of Selenium from Plants Used in Phytoremediation. *Int. J. Phytoremediation* **4**, 315–329 (2002).

65. Hanson, B. *et al.* Selenium accumulation protects *Brassica juncea* from invertebrate herbivory and fungal infection. *New Phytol.* **159**, 461–469 (2003).

66. Freeman, J. L., Quinn, C. F., Marcus, M. A., Fakra, S. & Pilon-Smits, E. A. H. Selenium-Tolerant Diamondback Moth Disarms Hyperaccumulator Plant Defense. *Curr. Biol.* **16**, 2181–2192 (2006).

67. Godinho, D. P., Serrano, H. C., Da Silva, A. B., Branquinho, C. & Magalhães, S. Effect of Cadmium Accumulation on the Performance of Plants and of Herbivores That Cope Differently With Organic Defenses. *Front. Plant Sci.* **9**, (2018).

68. Godinho, D. P., Branquinho, C. & Magalhães, S. Intraspecific variability in herbivore response to elemental defences is caused by the metal itself. *J. Pest Sci.* **96**, 797–806 (2023).

69. Hunziker, P. E. & Kägi, J. H. R. Metallothionein. in *Metalloproteins: Part 2: Metal Proteins with Non-redox Roles* (ed. Harrison, P. M.) 149–181 (Palgrave Macmillan UK, 1985). doi:10.1007/978-1-349-06375-8_4.

70. Janssens, T. K. S., Roelofs, D. & Van Straalen, N. M. Molecular mechanisms of heavy metal tolerance and evolution in invertebrates. *Insect Sci.* **16**, 3–18 (2009).

71. Price-Haughey, J., Bonham, K. & Gedamu, L. Heavy metal-induced gene expression in fish and fish cell lines. *Environ. Health Perspect.* **65**, 141–147 (1986).

72. Chen, L. *et al.* Heavy Metal-induced Metallothionein Expression Is Regulated by Specific Protein Phosphatase 2A Complexes. *J. Biol. Chem.* **289**, 22413–22426 (2014).

73. García-Alcalde, F. *et al.* Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics* **28**, 2678–2679 (2012).

74. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**, 292–294 (2016).

75. Van Leeuwen, T. *et al.* Mitochondrial heteroplasmy and the evolution of insecticide resistance: non-Mendelian inheritance in action. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 5980–5985 (2008).

76. Van Leeuwen, T., Tirry, L. & Nauen, R. Complete maternal inheritance of bifenazate resistance in *Tetranychus urticae* Koch (Acari: Tetranychidae) and its implications in mode of action considerations. *Insect Biochem. Mol. Biol.* **36**, 869–877 (2006).

77. Sun, J.-T. *et al.* The mitochondrial genome of the red tomato spider mite, *Tetranychus evansi* Baker & Pritchard (Acari: Tetranychidae) and its implications for phylogenetic analysis. *Syst. Appl. Acarol.* 1724–1735 (2019) doi:10.11158/saa.24.9.9.

78. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

79. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

80. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

81. Picard Tools - By Broad Institute. http://broadinstitute.github.io/picard/.

82. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

83. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* **15**, 356 (2014).

84. Attimonelli, M. & Calabrese, F. M. Chapter 6 - Human nuclear mitochondrial sequences (NumtS). in *The Human Mitochondrial Genome* (eds. Gasparre, G. & Porcelli, A. M.) 131–143 (Academic Press, 2020).

85. Albayrak, L. *et al.* The ability of human nuclear DNA to cause false positive low-abundance heteroplasmy calls varies across the mitochondrial genome. *BMC Genomics* **17**, 1017 (2016).

86. R Core, T. R: The R Project for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org/ (2022).

87. Schlötterer, C., Kofler, R., Versace, E., Tobler, R. & Franssen, S. U. Combining experimental evolution with next-generation sequencing: a powerful tool to study adaptation from standing genetic variation. *Heredity* **114**, 431–440 (2015).

88. Allio, R., Donega, S., Galtier, N. & Nabholz, B. Large Variation in the Ratio of Mitochondrial to Nuclear Mutation Rate across Animals: Implications for Genetic Diversity and the Use of Mitochondrial DNA as a Molecular Marker. *Mol. Biol. Evol.* **34**, 2762–2772 (2017).

89. Nei, M. Analysis of Gene Diversity in Subdivided Populations. *Proc. Natl. Acad. Sci.* **70**, 3321–3323 (1973).

90. Hamilton, M. B. Population Genetics, 2nd Edition | Wiley. *Wiley.com* https://www.wiley.com/en-us/Population+Genetics%2C+2nd+Edition-p-9781118436943.

91. Wiberg, R. A. W., Gaggiotti, O. E., Morrissey, M. B. & Ritchie, M. G. Identifying consistent allele frequency differences in studies of stratified populations. *Methods Ecol. Evol.* **8**, 1899–1909 (2017).

92. Mantel, N. & Haenszel, W. Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease. *JNCI J. Natl. Cancer Inst.* **22**, 719–748 (1959).

93. Taus, T., Futschik, A. & Schlötterer, C. Quantifying Selection with Pool-Seq Time Series Data. *Mol. Biol. Evol.* **34**, 3023–3034 (2017).

94. Barrett, R. D. H. & Schluter, D. Adaptation from standing genetic variation. *Trends Ecol. Evol.* **23**, 38–44 (2008).

95. Haynes, W. Benjamini–Hochberg Method. in *Encyclopedia of Systems Biology* (eds. Dubitzky, W., Wolkenhauer, O., Cho, K.-H. & Yokota, H.) 78–78 (Springer, 2013). doi:10.1007/978-1-4419-9863-7_1215.

96. Lusis, A. J. Genetic factors affecting blood lipoproteins: the candidate gene approach. *J. Lipid Res.* **29**, 397–429 (1988).

97. Hussain, N. *et al.* Metal induced non-metallothionein protein in earthworm: A new pathway for cadmium detoxification in chloragogenous tissue. *J. Hazard. Mater.* **401**, 123357 (2021).

98. Nordberg, M. & Nordberg, G. F. Metallothionein and Cadmium Toxicology—Historical Review and Commentary. *Biomolecules* **12**, 360 (2022).

99. Tobler, R. *et al.* Massive Habitat-Specific Genomic Response in *D. melanogaster* Populations during Experimental Evolution in Hot and Cold Environments. *Mol. Biol. Evol.* **31**, 364–375 (2014).

100. Seabra, S. G. *et al.* Different Genomic Changes Underlie Adaptive Evolution in Populations of Contrasting History. *Mol. Biol. Evol.* **35**, 549–563 (2018).

101. Portinha, B. *et al.* Whole-genome analysis of multiple wood ant population pairs supports similar speciation histories, but different degrees of gene flow, across their European ranges. *Mol. Ecol.* **31**, 3416–3431 (2022).

102. Bluher, S. E., Miller, S. E. & Sheehan, M. J. Fine-Scale Population Structure but Limited Genetic Differentiation in a Cooperatively Breeding Paper Wasp. *Genome Biol. Evol.* **12**, 701–714 (2020).

103. Bagley, R. K., Sousa, V. C., Niemiller, M. L. & Linnen, C. R. History, geography and host use shape genomewide patterns of genetic variation in the redheaded pine sawfly (Neodiprion lecontei). *Mol. Ecol.* **26**, 1022–1044 (2017).

104. Parejo, M., Wragg, D., Henriques, D., Charrière, J.-D. & Estonba, A. Digging into the Genomic Past of Swiss Honey Bees by Whole-Genome Sequencing Museum Specimens. *Genome Biol. Evol.* **12**, 2535–2551 (2020).

105. Miller, S. E. *et al.* Evolutionary dynamics of recent selection on cognitive abilities. *Proc. Natl. Acad. Sci.* **117**, 3045–3052 (2020).

106. Chen, C. *et al.* Population Structure and Diversity in European Honey Bees (*Apis mellifera* L.)—An Empirical Comparison of Pool and Individual Whole-Genome Sequencing. *Genes* **13**, 182 (2022).

107. Bhatia, G., Patterson, N., Sankararaman, S. & Price, A. L. Estimating and interpreting FST: The impact of rare variants. *Genome Res.* **23**, 1514–1521 (2013).

108. Weir, B. S. & Cockerham, C. C. Estimating F-Statistics for the Analysis of Population Structure. *Evolution* **38**, 1358–1370 (1984).

109. Allendorf, F. W., Luikart, G. H. & Aitken, S. N. Conservation and the Genetics of Populations, 2nd Edition; https://www.wiley.com/en-us/Conservation+and+the+Genetics+of+Populations%2C+2nd+Edition-p-9780470671450.

110. King, E. G. *et al.* Genetic dissection of a model complex trait using the Drosophila Synthetic Population Resource. *Genome Res.* **22**, 1558–1566 (2012).

111. Valiente-Mullor, C. *et al.* One is not enough: On the effects of reference genome for the mapping and subsequent analyses of short-reads. *PLOS Comput. Biol.* **17**, e1008678 (2021).

112. Heller, R. *et al.* A reference-free approach to analyse RADseq data using standard next generation sequencing toolkits. *Mol. Ecol. Resour.* **21**, 1085–1097 (2021).

113. Uricaru, R. *et al.* Reference-free detection of isolated SNPs. *Nucleic Acids Res.* **43**, e11 (2015).

114. Bush, S. J. *et al.* Genomic diversity affects the accuracy of bacterial single-nucleotide polymorphism–calling pipelines. *GigaScience* **9**, giaa007 (2020).

115. Usongo, V. *et al.* Impact of the choice of reference genome on the ability of the core genome SNV methodology to distinguish strains of *Salmonella enterica serovar* Heidelberg. *PLOS ONE* **13**, e0192233 (2018).

116. Pightling, A. W., Petronella, N. & Pagotto, F. Choice of Reference Sequence and Assembler for Alignment of Listeria monocytogenes Short-Read Sequence Data Greatly Influences Rates of Error in SNP Analyses. *PLOS ONE* **9**, e104579 (2014).

117. Bertels, F., Silander, O. K., Pachkov, M., Rainey, P. B. & van Nimwegen, E. Automated Reconstruction of Whole-Genome Phylogenies from Short-Sequence Reads. *Mol. Biol. Evol.* **31**, 1077–1088 (2014).

118. Mendez, F. L. Differences in the effective population sizes of males and females do not require differences in their distribution of offspring number. *Theor. Popul. Biol.* **114**, 19–28 (2017).

119. Stoler, N. & Nekrutenko, A. Sequencing error profiles of Illumina sequencing instruments. *NAR Genomics Bioinforma.* **3**, lqab019 (2021).

120. Formenti, G. *et al.* The era of reference genomes in conservation genomics. *Trends Ecol. Evol.* **37**, 197–202 (2022).

121. Tørresen, O. K. *et al.* Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Res.* **47**, 10994–11006 (2019).

122. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* **13**, 36–46 (2012).

123. Zhang, D.-X. & Hewitt, G. M. Insect mitochondrial control region: A review of its structure, evolution and usefulness in evolutionary studies. *Biochem. Syst. Ecol.* **25**, 99–120 (1997).

124. Rand, D. M. Endotherms, ectotherms, and mitochondrial genome-size variation. *J. Mol. Evol.* **37**, 281–295 (1993).

125. Pamilo, P., Viljakainen, L. & Vihavainen, A. Exceptionally High Density of NUMTs in the Honeybee Genome. *Mol. Biol. Evol.* **24**, 1340–1346 (2007).

126. Maude, H. *et al.* NUMT Confounding Biases Mitochondrial Heteroplasmy Calls in Favor of the Reference Allele. *Front. Cell Dev. Biol.* **7**, (2019).

127. Kimura, M. The neutral theory of molecular evolution and the world view of the neutralists. *Genome* **31**, 24–31 (1989).

128. Ohta, T. The Nearly Neutral Theory of Molecular Evolution. *Annu. Rev. Ecol. Syst.* **23**, 263–286 (1992).

129. Hoban, S. *et al.* Finding the Genomic Basis of Local Adaptation: Pitfalls, Practical Solutions, and Future Directions. *Am. Nat.* **188**, 379–397 (2016).

130. Storz, J. F. Using genome scans of DNA polymorphism to infer adaptive population divergence. *Mol. Ecol.* **14**, 671–688 (2005).

131. Booker, T. R., Yeaman, S. & Whitlock, M. C. Variation in recombination rate affects detection of outliers in genome scans under neutrality. *Mol. Ecol.* **29**, 4274–4279 (2020).

132. Yeaman, S. Local Adaptation by Alleles of Small Effect. *Am. Nat.* **186**, S74–S89 (2015).

133. Bazzicalupo, A. L. *et al.* Fungal heavy metal adaptation through single nucleotide polymorphisms and copy-number variation. *Mol. Ecol.* **29**, 4157–4169 (2020).

134. Talke, I. N., Hanikenne, M. & Krämer, U. Zinc-Dependent Global Transcriptional Control, Transcriptional Deregulation, and Higher Gene Copy Number for Genes in Metal Homeostasis of the Hyperaccumulator *Arabidopsis halleri*. *Plant Physiol.* **142**, 148–167 (2006).

135. Laporte, M. *et al.* RAD sequencing reveals within-generation polygenic selection in response to anthropogenic organic and metal contamination in North Atlantic Eels. *Mol. Ecol.* **25**, 219–237 (2016).

136. Eide, D. J. The SLC39 family of metal ion transporters. *Pflüg. Arch.* **447**, 796–800 (2004).

137. Fisher, R. A. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Earth Environ. Sci. Trans. R. Soc. Edinb.* **52**, 399–433 (1919).

138. Roelofs, D. *et al.* Adaptive differences in gene expression associated with heavy metal tolerance in the soil arthropod *Orchesella cincta*. *Mol. Ecol.* **18**, 3227–3239 (2009).

139. Chakdar, H., Thapa, S., Srivastava, A. & Shukla, P. Genomic and proteomic insights into the heavy metal bioremediation by cyanobacteria. *J. Hazard. Mater.* **424**, 127609 (2022).

140. Hamer, D. H. Metallothionein. *Annu. Rev. Biochem.* **55**, 913–951 (1986).

141. Coyle, P., Philcox, J. C., Carey, L. C. & Rofe, A. M. Metallothionein: the multipurpose protein. *Cell. Mol. Life Sci. CMLS* **59**, 627–647 (2002).

142. Fujishiro, H. *et al.* The role of ZIP8 down-regulation in cadmium-resistant metallothionein-null cells. *J. Appl. Toxicol.* **29**, 367–373 (2009).

143. Leslie, E. M., Liu, J., Klaassen, C. D. & Waalkes, M. P. Acquired cadmium resistance in metallothionein-I/II(-/-) knockout cells: role of the T-type calcium channel Cacnalpha1G in cadmium uptake. *Mol. Pharmacol.* **69**, 629–639 (2006).

144. Braeckman, B., Smagghe, G., Brutsaert, N., Cornelis, R. & Raes, H. Cadmium Uptake and Defense Mechanism in Insect Cells. *Environ. Res.* **80**, 231–243 (1999).

145. Craig, A., Hare, L. & Tessier, A. Experimental evidence for cadmium uptake via calcium channels in the aquatic insect Chironomus staegeri. *Aquat. Toxicol.* **44**, 255–262 (1999).

146. Craciun, A. R. *et al.* Variation in HMA4 gene copy number and expression among *Noccaea caerulescens* populations presenting different levels of Cd tolerance and accumulation. *J. Exp. Bot.* **63**, 4179–4189 (2012).

147. Li, M. *et al.* Cadmium directly induced the opening of membrane permeability pore of mitochondria which possibly involved in cadmium-triggered apoptosis. *Toxicology* **194**, 19–33 (2003).

148. Ohta, H. & Ohba, K. Involvement of metal transporters in the intestinal uptake of cadmium. *J. Toxicol. Sci.* **45**, 539–548 (2020).

149. Bannon, D. I., Abounader, R., Lees, P. S. J. & Bressler, J. P. Effect of DMT1 knockdown on iron, cadmium, and lead uptake in Caco-2 cells. *Am. J. Physiol.-Cell Physiol.* **284**, C44–C50 (2003).

150. Bressler, J. P., Olivi, L., Cheong, J. H., Kim, Y. & Bannona, D. Divalent metal transporter 1 in lead and cadmium transport. *Ann. N. Y. Acad. Sci.* **1012**, 142–152 (2004).

7. Supplementary Materials

Table S1: Candidate genes associated with the response to cadmium. SR1SR2: Control vs. Homogeneous; SR1SR3: Control vs Heterogeneous; "-" = the allele remains minor on control and on the selected regimes (homogeneous or heterogeneous), "↓" = minor allele on control but major allele on one of the selected regimes (homogeneous or heterogeneous), "↑" = major allele on control but minor allele on one of the selected regimes (homogeneous or heterogeneous)

ID	Name	Chr	Reference	SR1SR2			SR1SR3			Annotation
				-	↓	↑	-	↓	↑	
tetur23g01920	*Cacnα2D3*	1	[143]	0	0	0	0	0	0	
tetur01g16522	*SLCP3*	1	[136]	1	0	0	0	0	0	
tetur01g16530	*SLCP2*	1	[136]	0	0	0	1	0	0	
tetur06g04620	n/a	1	[136]	0	0	0	0	0	6	GO term associated with SLC proteins
tetur03g08310	*SLC38A7*	1	[136]	3	0	0	6	0	0	
tetur08g04660	n/a	1	[136]	0	0	0	4	0	0	GO term associated with SLC proteins
tetur08g01780	n/a	1	[136]	1	0	0	1	0	0	GO term associated with SLC proteins
tetur27g02556	*SLCP1*	1	[136]	0	0	0	0	0	0	
tetur07g02970	n/a	1	[136]	12	0	0	0	0	0	GO term associated with SLC proteins
tetur35g00420	n/a	1	[71,72,97]	0	0	0	0	0	0	MIP family channel protein
tetur01g15060	n/a	1	[71,72,97]	39	0	0	6	0	0	MIP family channel protein
tetur03g09570	n/a	1	[142,148]	0	0	0	0	0	0	Zinc transporter ZIP9
tetur03g08340	n/a	1	[142,148]	0	0	0	0	0	0	Zinc transporter ZIP9
tetur10g03040	n/a	2	[98,140,141]	0	0	0	0	0	0	Metallothionein
tetur10g03080	n/a	2	[98,140,141]	0	0	0	0	0	0	Metallothionein
tetur01g00260	*MGTE*	2	[136,149,150]	12	0	4	0	0	0	GO term associated with SLC proteins
tetur29g01770	*Cacnα1G*	2	[143]	9	43	0	86	14	0	
tetur10g03010	n/a	2	[136]	1	0	0	3	0	0	GO term associated with SLC proteins
tetur10g05190	*SLC4A*	2	[136]	33	0	0	18	0	2	
tetur18g00240	n/a	2	[136]	3	0	0	2	0	0	GO term associated with SLC proteins
tetur17g03150	*SAT1*	2	[136]	0	0	0	0	0	0	GO term associated with SLC proteins
tetur24g01090	n/a	2	[136]	20	0	0	4	0	0	GO term associated with SLC proteins
tetur04g08240	n/a	2	[136]	0	0	0	6	0	0	GO term associated with SLC proteins
tetur04g92275	n/a	2	[136]	4	0	0	1	0	0	GO term associated with SLC proteins
tetur04g08720	n/a	2	[136]	2	0	0	0	0	0	GO term associated with SLC proteins
tetur01g16520	*SLCP4*	2	[136]	1	0	0	0	0	0	
tetur25g01540	n/a	2	[136]	0	0	0	0	0	0	GO term associated with SLC proteins
tetur16g03886	*SLCP1*	3	[136]	0	0	0	0	0	0	
tetur02g13160	n/a	3	[136]	0	0	0	0	0	0	GO term associated with SLC proteins
tetur02g15193	*SLC38A11*	3	[136]	0	0	0	0	0	0	
tetur20g00330	*SLC9A7*	3	[136]	0	0	0	1	0	0	
tetur11g01900	n/a	3	[136]	1	0	7	0	0	0	GO term associated with SLC proteins
tetur13g92980	n/a	3	[136]	2	0	0	3	0	0	GO term associated with SLC proteins
tetur13g04280	*SLC12*	3	[136]	23	0	4	11	0	4	

Table S2: Read summary metrics for raw and quality trimmed reads of the *T. evansi* and *T. urticae* laboratory populations. Fastq files of all populations were mapped to the *T. evansi* mitochondria genome and the *T. urticae* reference genome. Values extracted from fastqc and qualimap reports. Seq.: Sequences; DP: depth of coverage; SD: standard deviation; MQ: Mapping Quality; TeX: inbred lines; TeOut: *T. evansi* outbred population; TuOut: *T. urticae* outbred population; nuDNA: Nuclear DNA; mtDNA: Mitochondrial DNA; Mit.: Mitochondria genome. For the outbred populations, we also mapped the pools to only the mitochondrial genome of *T. urticae*, which is present here on the rows where the Reference is the *T. urticae* and the Genomic Region is "Mit."

				Genomic Region	After Mapping				nuDNA		mtDNA	
Population	Reference	Raw Seq.	Trimmed Seq.		Reads (Total)	Reads (%)	GC (%)	Mean MQ	Mean DP	SD DP	Mean DP	SD DP
Te8	*T. evansi*	28614934	28422014	Mit.	55566	0.39	19.79	58.54	-	-	100.32	31.21
	T. urticae			Whole genome	14994844	52.76	32.75	31.69	23.33	3.88	81.68	28.36
Te14	*T. evansi*	22556314	22434888	Mit.	69960	0.31	19.21	59.08	-	-	154.51	54.71
	T. urticae			Whole genome	13517302	60.25	32.59	31.61	22.31	3.34	114.39	46.89
Te19	*T. evansi*	28045298	27962082	Mit.	32558	0.12	19.31	58.4	-	-	82.19	13.66
	T. urticae			Whole genome	15623910	55.88	32.9	31.86	26.94	4.17	71.76	20.72
Te23	*T. evansi*	28597308	28549440	Mit.	45308	0.16	19.55	58.53	-	-	84.05	16.75
	T. urticae			Whole genome	16700424	58.50	32.51	31.91	30.44	4.43	85.93	24.39
Te27	*T. evansi*	47072636	46984284	Mit.	85882	0.18	19.26	58.84	-	-	121.54	43.51
	T. urticae			Whole genome	25447962	54.16	33.08	31.77	44.31	6.94	153.28	73.59
Te28	*T. evansi*	24292316	24255290	Mit.	57986	0.24	18.41	59.07	-	-	107.05	36.57
	T. urticae			Whole genome	13560530	55.91	32.62	31.99	24.44	3.89	132.60	59.42
Te29	*T. evansi*	23522996	23412016	Mit.	53160	0.23	18.55	58.99	-	-	155.70	73.73
	T. urticae			Whole genome	11647376	49.75	32.64	31.99	19.84	3.06	117.94	43.85
Te32	*T. evansi*	26194162	26103970	Mit.	68804	0.26	18.09	59.37	-	-	150.33	71.94
	T. urticae			Whole genome	12606156	48.29	33.01	31.9	20.91	3.07	137.53	79.38
Te42	*T. evansi*	29593444	29459858	Mit.	58216	0.20	18.36	58.96	-	-	112.24	47.06
	T. urticae			Whole genome	15969250	54.21	32.74	31.78	26.46	3.90	115.21	47.03
TeOut	*T. evansi*	35830232	35630308	Mit.	144216	0.40	18.61	59.24	-	-	284.21	183.04
	T. urticae			Whole genome	21221814	59.56	32.31	31.73	34.46	4.98	228.72	141.77
				Mit.	121946	0.34	19.62	52.94	-	-	209.99	115.02
TuOut	*T. evansi*	30351716	30210454	Mit.	119186	0.39	19.16	53.96	-	-	332.19	197.45
	T. urticae			Whole genome	27073540	89.62	32.04	44.58	35.31	3.90	478.06	214.58
				Mit.	149122	0.49	18.02	59.71	-	-	557.71	251.62

Table S3: Average FST across the genome and detected SNPs on nucleus between all inbred lines pairwise combinations. Average FST (bottom) and total number of SNPs (top). Negative FST values were computed due to the FST estimator used and should be interpreted as no differentiation

	Te8	**Te14**	**Te19**	**Te23**	**Te27**	**Te28**	**Te29**	**Te32**	**Te42**
Te8		762851	706824	747288	726562	666864	809681	741328	780978
Te14	-0.013		779978	785703	765224	704867	863637	773973	837737
Te19	-0.010	-0.012		666125	633732	595113	744259	672353	704294
Te23	-0.005	-0.009	-0.011		642839	626383	774605	709358	726536
Te27	-0.005	-0.008	-0.0100	-0.0088		531460	631168	574259	587840
Te28	-0.004	-0.008	-0.0108	-0.0104	-0.0077		729051	657458	697527
Te29	-0.011	-0.012	-0.0134	-0.0129	-0.0124	-0.013		802691	840867
Te32	-0.008	-0.012	-0.0148	-0.0150	-0.0131	-0.0142	-0.0148		776235
Te42	-0.006	-0.009	-0.0106	-0.0100	-0.0093	-0.0097	-0.0145	-0.0137	

Table S4: Average FST across the genome and detected SNPs on mitochondria between all inbred lines pairwise combinations. Average FST (bottom) and total number of SNPs (top). Negative FST values were computed due to the FST estimator used and should be interpreted as no differentiation

	Te8	**Te14**	**Te19**	**Te23**	**Te27**	**Te28**	**Te29**	**Te32**	**Te42**
Te8		183	167	171	179	175	182	169	189
Te14	0.0040		189	159	169	160	170	161	174
Te19	-0.0061	0.0061		154	164	159	179	162	177
Te23	-0.0056	-0.0003	-0.0066		170	154	165	159	170
Te27	-0.0076	0.0163	-0.0029	-0.0036		154	168	154	164
Te28	0.0061	-0.0036	0.0049	-0.0020	-0.0012		160	153	158
Te29	-0.0016	-0.0014	-0.0017	0.0018	-0.0050	0.0017		166	180
Te32	0.0134	-0.0046	0.0225	0.0097	0.0149	0.0054	0		169
Te42	0.0233	0.0196	0.0205	0.0143	0.0268	0.0325	0.0404	0.0227	

Table S5: Read summary metrics for raw and quality trimmed reads of the three *T. urticae* selection regimes (control, homogeneous and heterogeneous). All populations were mapped to the *T. urticae* reference genome. Values extracted from fastqc and qualimap reports. Seq.: Sequences; DP: depth of coverage; SD: standard deviation; MQ: Mapping Quality; SR1: No cadmium (control); SR2: High concentrations of cadmium (homogeneous environment); SR3: with and without cadmium (heterogeneous environment); nuDNA: Nuclear DNA; mtDNA: Mitochondrial DNA

				After Mapping				nuDNA		mtDNA	
Regime	Replica	Raw Seq.	Trimmed Seq.	Reads (Total)	Reads (%)	GC (%)	Mean MQ	Mean DP	SD DP	Mean DP	SD DP
SR1	1	20185276	20161266	17011854	84.38	33	43.87	24.42	2.77	147.3	54.80
	2	19921384	19900678	16844778	84.64	32.82	43.9	24.39	2.76	863.91	135.19
	3	23279490	23255494	19636160	84.44	32.84	43.85	28.06	3.03	115.88	53.97
	4	27748822	27713584	22621086	81.62	32.88	43.87	32.50	3.31	152.71	87.03
	5	20249518	20224558	17529346	86.67	32.92	44.01	25.40	2.77	116.17	56.99
SR2	1	22840000	22810264	19553928	85.72	32.9	43.99	27.98	3.04	160.79	64.28
	2	23637006	23613100	19674580	83.32	32.88	43.99	27.95	3.03	283.36	52.65
	3	22082466	22054768	18382356	83.35	33.03	43.8	26.43	2.77	149.56	61.75
	4	21935206	21906744	18587920	84.85	33.16	43.92	26.87	3.04	77.53	23.65
	5	20757398	20729018	17705418	85.41	33	43.77	25.42	2.76	289.72	55.20
SR3	1	22835992	22812262	19021500	83.38	33.04	43.79	27.39	3.29	94	47.63
	2	25445914	25413026	21000608	82.64	32.83	43.77	30.46	3.31	108.16	37.29
	3	21945552	21922964	18493230	84.36	32.93	43.81	26.93	3.04	119.62	55.46
	4	24553526	24515992	20340004	82.97	32.89	43.77	29.44	3.31	516	74.53
	5	23607154	23574950	19523010	82.81	32.81	43.85	27.97	3.03	589.92	201.99

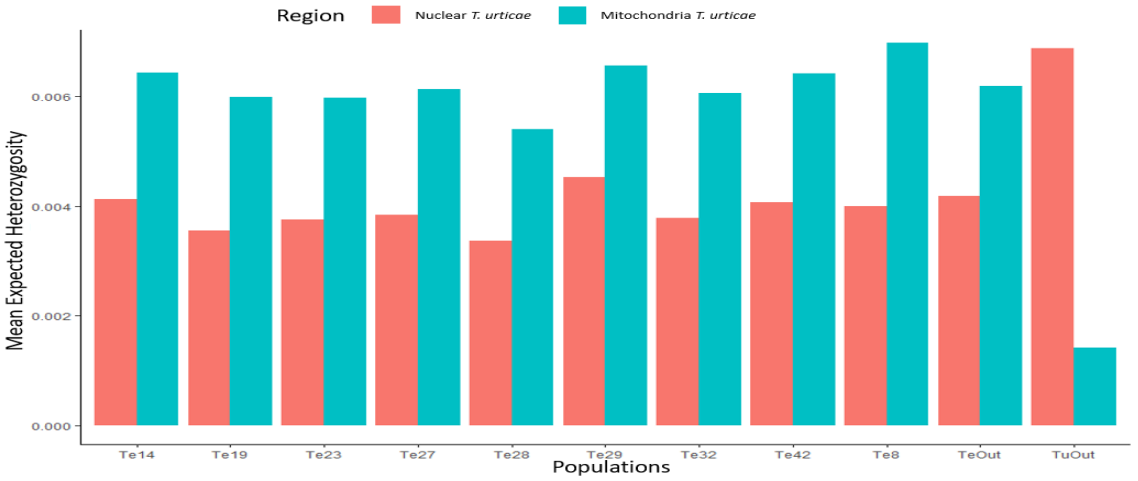

Figure S1: Genetic diversity on nucleus and mitochondria of the *T. evansi* and *T. urticae* laboratory populations considering all sites. Sites are filtered as explained in methods section 2.5 and include the sites were the expected heterozygosity is equal to zero. Tex: *T. evansi* inbred lines. TeOut: *T. evansi* outbred population. TuOut: *T. urticae* outbred population. For all *T. evansi* populations mean expected heterozygosity is higher in mitochondria when compared to the nucleus, a consequence of the effect of the different mutation rates in the 2 genomic regions.

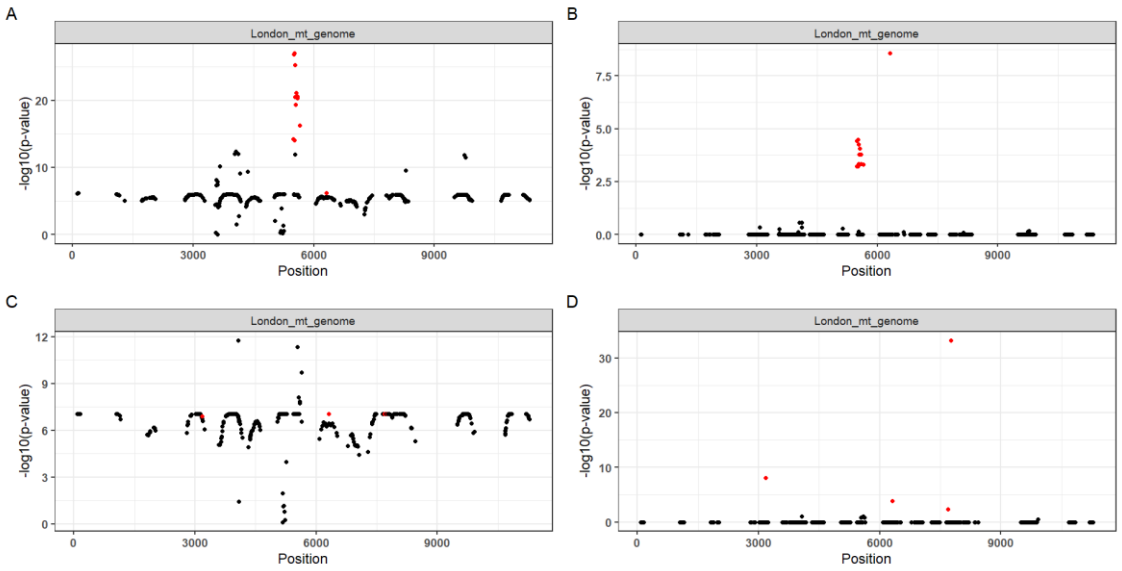

Figure S2: Genome-wide distribution of all SNPs detected in the mitochondria genome between the control and the two selection regimes (homogeneous and heterogeneous). Control vs Homogeneous environments (A e B) e Control vs Heterogeneous environments (C e D). For both comparisons, a general linear model with a quasibinomial error structed (A e C) and a CMH test (B e D) were considered with a FDR adjustment, represented as -log10(p-value) in the y axis. Points in red represent candidate significative SNPs in both statistical tests used.

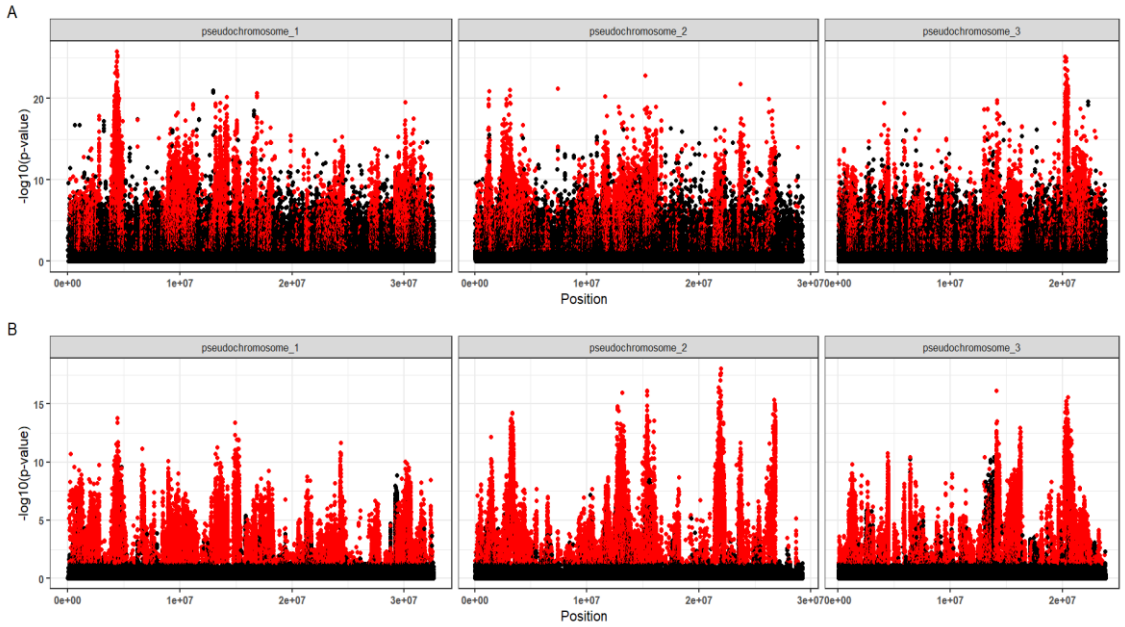

Figure S3: Genome-wide distribution of all SNPs detected in the nuclear genome between the control and the homogeneous regime. Results of the general linear model with a quasibinomial error structure are in (A) and a CMH test in (B). Each point represents a SNP. Points in red represent candidate significative SNPs in both statistical tests used, *i.e.*, SNPs with p-values (from both CMH and GLM tests) lower than 0.05, after a FDR correction.

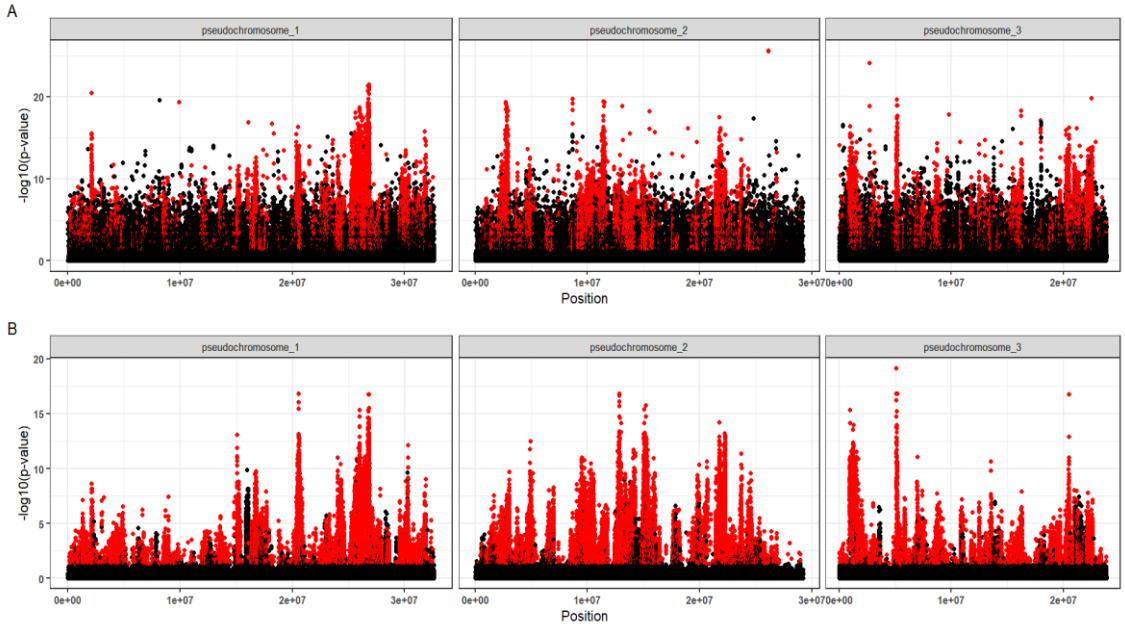

Figure S4: Genome-wide distribution of all SNPs detected in the nuclear genome between the control and the heterogeneous regime. Results of the general linear model with a quasibinomial error structure are in (A) and a CMH test in (B). Each point represents a SNP. Points in red represent candidate significative SNPs in both statistical tests used, *i.e.*, SNPs with p-values (from both CMH and GLM tests) lower than 0.05, after a FDR correction.

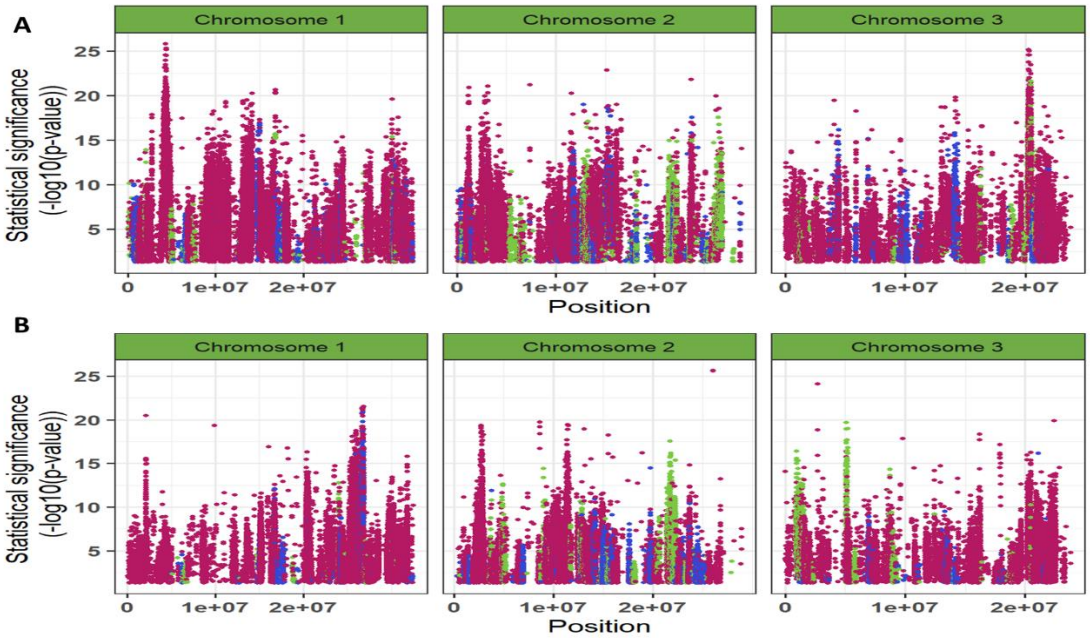

Figure S5: Genome-wide distribution of candidate SNPs associated with adaptation of *T. urticae* to cadmium in Homogeneous and Heterogeneous regimes, on the nuclear genome. Results of the general linear model with a quasibinomial error structure are in (A) for the Homogeneous and (B) for the Heterogeneous regimes. Each point represents a candidate SNP, *i.e.*, SNPs with p-values (from both CMH and GLM tests) lower than 0.05, after a FDR correction. Results of CMH test are available in Fig. 3.7. SNPs in green represent an increase in allele frequencies on the selected regime, compared to the control. SNPs in blue represent a decrease in allele frequencies on the selected regime, compared to the control. SNPs in pink represent changes where the allele remains minor on both control and the selected regime. SNPs with a significant change in allele frequencies are distributed along the three pseudochromosomes.

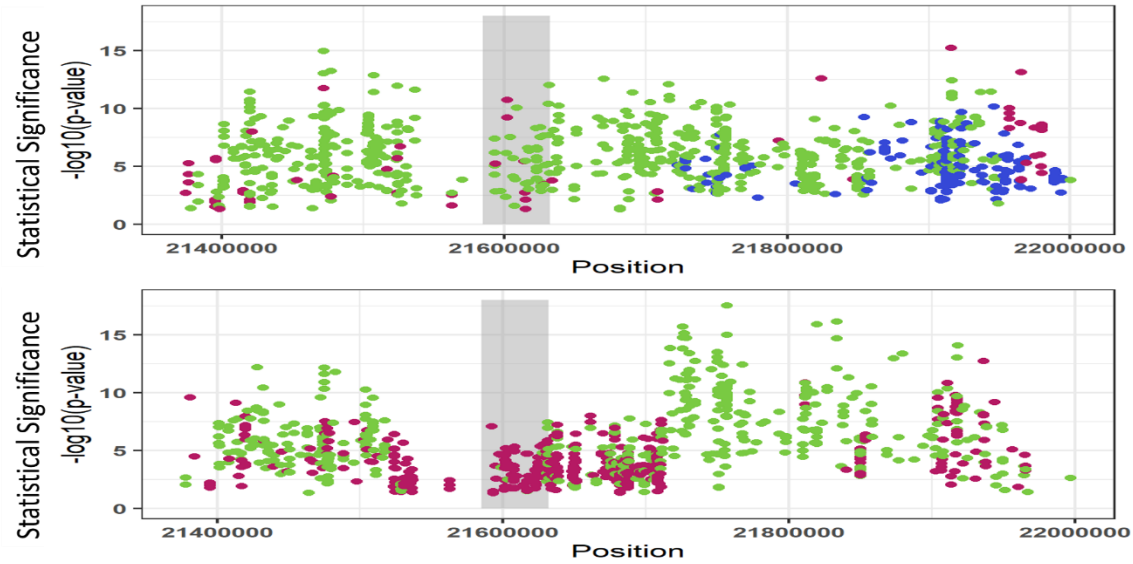

Figure S6: Candidate SNPs detected on homogeneous and heterogeneous regimes, compared to the control, for the *Cacnα1G* gene. Results of the general linear model with a quasibinomial error structure are in (A) for the Homogeneous and (B) for the Heterogeneous regimes. The gene is highlighted by the grey region. Each point represents a candidate SNP, i.e., SNPs with p-values (from both CMH and GLM tests) lower than 0.05, after a FDR correction. Results of the CMH test are available in the Figure 3.9. SNPs in green represent an increase in allele frequencies on the selected regime, compared to the control. SNPs in blue represent a decrease in allele frequencies on the selected regime, compared to the control. SNPs in pink represent changes where the allele remains minor on both control and the selected regime. Only SNPs with an increase in selection regime compared with the control were found on both homogeneous and heterogeneous regimes: 43 SNPs on the Homogeneous regime and 14 on the Heterogeneous regime.