EPIGENETIC DIVERSITY IN THE CLONAL WHIPTAIL HYBRID *ASPIDOSCELIS NEOMEXICANUS* IDENTIFIED THROUGH ALLELE-SPECIFIC EXPRESSION

Dissertation

Zur Erlangung des Grades

Doktor der Naturwissenschaften

Am Fachbereich Biologie
der Johannes Gutenberg-Universität Mainz

Valentine Sarah Jane Patterson

geb. am 02.03.1996 in Aberdeen, UK

Mainz, 2023

Dekan: Prof. Dr. Eckhard Thines

1. Berichterstatter: Prof. Dr. Peter Baumann

2. Berichterstatter: Dr. Romain Libbrecht

Tag der mündlichen Prüfung:

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

For a long time, it has been believed that genetic recombination, which occurs during sexual reproduction, is crucial for creating genetic diversity within populations and enabling species to adapt to their environments. However, the widespread distribution of clonal species of hybrid origin, such as the whiptail lizards of the *Aspidoscelis* genus, suggests that sexual reproduction may be less vital for a species' survival than previously assumed. The coexistence and overlapping territories of gonochoristic species (sexually reproducing) have facilitated multiple hybridization events, leading to the emergence of hybrid parthenogenetic (asexually reproducing) lineages. One example is the hybridization between *A. marmoratus* and *A. arizonae*, resulting in the offspring *A. neomexicanus*. As *A. neomexicanus* reproduces through parthenogenesis, the genetic material from both parental species has been retained over time. Consequently, this species exhibits high heterozygosity, with each gene's alleles having species specific differences. Therefore, *A. neomexicanus* serves as a unique model for studying allele-specific expression, which refers to the preferential expression of one allele over the other. By employing RNA-Seq and DNA-Seq techniques, we have identified different biases in allele expression among individuals of *A. neomexicanus*, highlighting that distinct allelic bias variations exist between genetically identical individuals. Additionally, we have identified allele-specific expression differences across *A. neomexicanus* tissues, highlighting how gene regulatory mechanisms can generate intra-species diversity. Moreover, we use *in silico* tissue references of parental species data to identify directional allelic bias shifts. We hypothesise how the diversity and directional shift in allelic bias in *A. neomexicanus* may allow it to use the better allele in growth, development, and environmental adaptation, and therefore leading to its' high performance despite the lack of genetic diversity generated through sexual reproduction.

ZUSAMMENFASSUNG

Lange Zeit ging man davon aus, dass die genetische Rekombination, die bei der sexuellen Fortpflanzung stattfindet, entscheidend für die genetische Vielfalt innerhalb von Populationen ist und es den Arten ermöglicht, sich an ihre Umwelt anzupassen. Die weite Verbreitung klonaler Arten hybriden Ursprungs, wie z.B. der Peitschenschwanzeidechsen der Gattung Aspidoscelis, deutet jedoch darauf hin, dass die sexuelle Fortpflanzung für das Überleben einer Art weniger wichtig sein könnte als bisher angenommen. Die Koexistenz und die Überlappung von Territorien gonochoristischer Arten (sexuell reproduzierend) haben mehrere Hybridisierungsereignisse begünstigt, die zur Entstehung hybrider parthenogenetischer (asexuell reproduzierender) Linien geführt haben. Ein Beispiel dafür ist die Hybridisierung zwischen A. marmoratus und A. arizonae, aus der der Nachkomme A. neomexicanus hervorging. Da sich A. neomexicanus durch Parthenogenese fortpflanzt, ist das genetische Material beider Elternarten über die Zeit erhalten geblieben. Folglich weist diese Art eine hohe Heterozygotie auf, wobei die Allele jedes Gens artspezifische Unterschiede aufweisen. Daher dient A. neomexicanus als einzigartiges Modell für die Untersuchung der allelspezifischen Expression, d. h. der bevorzugten Expression eines Allels gegenüber einem anderen. Durch den Einsatz von RNA-Seq- und DNA-Seq-Techniken haben wir verschiedene Unterschiede in der Allelexpression zwischen Individuen von A. neomexicanus identifiziert, was zeigt, dass es zwischen genetisch identischen Individuen deutliche Unterschiede in der Allelexpression gibt. Darüber hinaus wurden allelspezifische Expressionsunterschiede in verschiedenen Geweben von A. neomexicanus identifiziert, die verdeutlichen, wie genregulatorische Mechanismen eine Vielfalt innerhalb der Spezies erzeugen können. Des Weiteren verwenden wir In silico-Gewebereferenzen von Daten der Elternarten, um gerichtete Allelverschiebungen zu identifizieren. Wir stellen die Hypothese auf, dass die Vielfalt und die denkbare Richtungsverschiebung der Allelvorlieben bei A. neomexicanus es ihr ermöglichen, das bessere Allel für Wachstum, Entwicklung und Umweltanpassung zu nutzen, was zu ihrer hohen Leistungsfähigkeit führt, obwohl es an genetischer Vielfalt durch sexuelle Fortpflanzung mangelt.

I. INTRODUCTION

I.I. WHIPTAIL LIZARDS

The *Aspidoscelis* genus (formerly *Cnemidophorus*) consists of over 50 lizard species[1]. Of these, approximately one third are unisexual. This genus of lizards is found in the South-West of the US, predominantly in New Mexico [1]. The overlapping territories of the species within this genus has facilitated multiple hybridization events between these closely related sexual species, resulting in hybrid parthenogenetic species. These hybridization and speciation events are possible as the species are not highly divergent [2],[3];[4]. Hybridization events between more divergent species do not commonly result in viable offspring [5], however hybrid whiptail lizards reproduce successfully via parthenogenesis. One such instance of a hybridization event leading to a speciation event is between two diploid sexual *Aspidoscelis* species, an *A. marmoratus* female and an *A. arizonae* male, resulting in a diploid hybrid offspring – the new *A. neomexicanus* species [6]. As these species have undergone several taxonomic revisions and have been reclassified to reflect the most accurate representation of their evolutionary history, it is important to note that the taxonomy of *Aspidoscelis* lizards described in this thesis follows Barley et al. 2022 [7]. In the *A. marmoratus*, *A. arizonae* and *A. neomexicanus* hybridisation complex, both diploid parental species have 46 chromosomes. The diploid hybrid offspring inherited 23 chromosomes from either parental species, making up a total of 46 chromosomes (Figure 1 [8]).

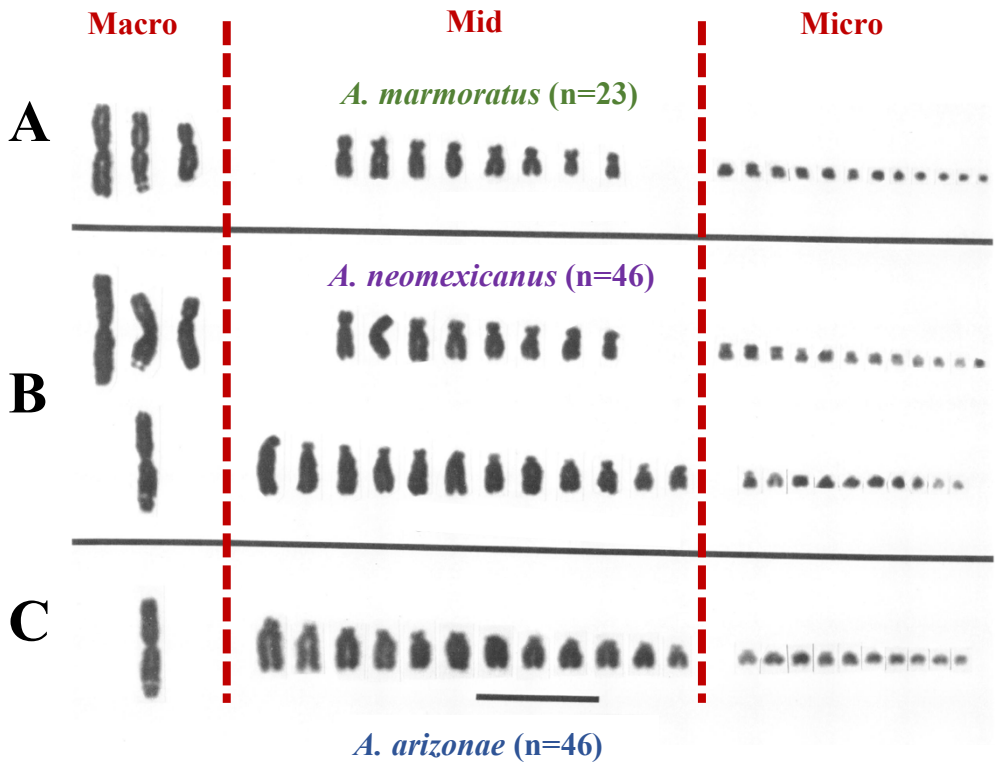


Figure 1. Karyotype of three *Aspidoscelis* species. Macro, mid and micro sized chromosomes can be identified. (A) Haploid state of *A. marmoratus.* (C) Haploid state of *A. arizonae.* (B) Diploid state of *A. neomexicanus* hybrid offspring. One chromosome set of each parental species is inherited. Adapted from Cole et al. 1988 [8].

The parental species of *A. neomexicanus; A. arizonae* is primarily found in open grass lands and prairies of North America, whilst *A. marmoratus* is native to western North America and can be found in woodlands, rocky areas, and coastal regions (Figure 2). *A. neomexicanus* is a member of the family Teiidae, which includes other whiptail lizards known for their long, thin tails and fast movement. It is typically found in open, arid habitats such as deserts, grasslands, and scrublands. It is a diurnal lizard, meaning that it is most active during the day, and feeds on a variety of insects and other small invertebrates. Moreover, the genetic contribution from the parental species is apparent in the phenotype of *A. neomexicanus* lineage. The species displays the distinct bright blue tail and stripes characteristic of *A. arizonae* and the dark spots and larger size of the *A. marmoratus.* Morphologically, the *A. neomexicanus* sits size wise in between its' parental species; larger the *A. arizonae* and smaller than *A. marmoratus* (Figure 2C).

A

B

C

	A. sexlineatus	*A. arizonae*	*A. neomexicanus*	*A. neomexicanus*	*A. marmoratus*
Location	South Texas	Lordsburg	Lordsburg	Peña Blanca	Lordsburg
Body Length	57.8 ± 0.99	58.8 ± 0.68	69.2 ± 0.98	73.5 ± 1.07	85.9 ± 1.00

Figure 2. Ecological and phenotypic differences in *A. arizonae, A. marmoratus* and *A. neomexicanus.* (A) Overlapping ecological distributions of *A. marmoratus* (highlighted in green) and *A. arizonae* (highlighted in blue). (B) Ecological distribution of *A. neomexicanus,* in the overlapping region of the parental species. (C) Morphological differences between parental species; *A. arizonae* and *A. marmoratus*, and the *A. neomexicanus.* Adapted from Cole et al. 1988 [8].

I.II. PARTHENOGENESIS

Whiptail lizards have evolved a reproductive system that allows them to reproduce without males: parthenogenesis. Parthenogenesis is a form of asexual reproduction in which an embryo develops from an unfertilized egg cell [9]. Lizard species in the *Aspidoscelis* genus can reproduce through obligate parthenogenesis (observed in the species originating from historical hybridization events) or facultative parthenogenesis (observed in the sexual species). Obligate parthenogenetic species reproduce exclusively asexually, whereas species which exhibit facultative parthenogenesis can, under certain conditions, change from sexual reproduction to asexual reproduction, and reproduce via parthenogenesis [9]. Obligate parthenogenesis is extremely uncommon and found in only a few vertebrate species, typically of hybrid origin [4]. However, parthenogenetically reproducing species, such as *A. neomexicanus*, have been widely successful in their propagation [9]. This new lineage has now been propagating for approximately 200,000 years through parthenogenesis (Figure 3 [7]), highlighting the success of a non-sexual hybrid species in colonizing its surrounding environment. This demonstrates how genetic recombination in sexual species is not the only avenue to success.

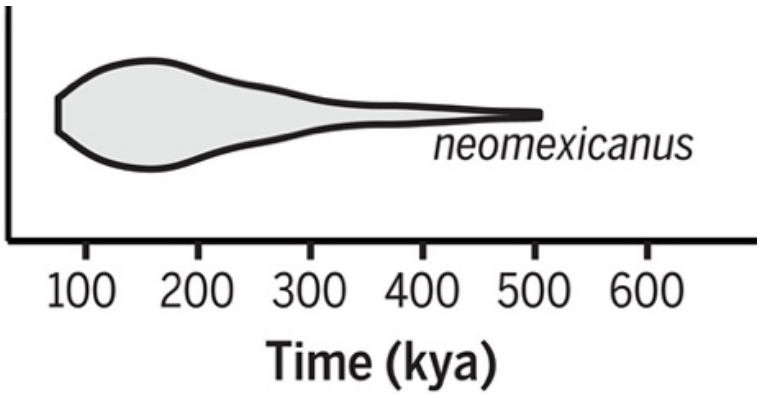


Figure 3. Time of origin estimate for *A. neomexicanus.* Violin plot shows 95% confidence interval for formation time estimates of diploid unisexual lineages in kya, thousand years ago (modified from Barley et al. 2022[7]).

It has been proposed that vertebrate parthenogenetic lineages which arise from hybridisation between two divergent taxa, within a specific range of phylogenetic distances, often results in

sexual–asexual network of species that recurrently hybridise (Figure 4 [7]). Parthenogenetic species often maintain a certain level of hybridisation with their closest sexual relatives, potentially generating new polyploid hybrid lineages, as observed when a viable tetraploid hybrid lineage was produced from the crossing of a triploid unisexual whiptail lizard of hybrid origin with a diploid bisexual male [10]. This one example of how allopolyploidy speciation via hybridisation is a major driving force in the diversification in the *Aspidoscelis* genus [11] . Triploid unisexual lineages are derived from backcrossing events between diploid unisexual lineages and sexual species. These hybridization outcomes give rise to one of the most striking features of whiptail lizards: their extreme diversity in chromosome number and structure. Many species of whiptail lizards are polyploid, meaning that they have multiple sets of chromosomes. Furthermore, the hybridization outcome between whiptail species may be predicted based on the evolutionary divergence between species. The rates of introgression between species decreases with time, since divergence, and once it has attained a certain threshold of evolutionary divergence, the hybridization transitions to unisexuality and the emergence of a new species [7].

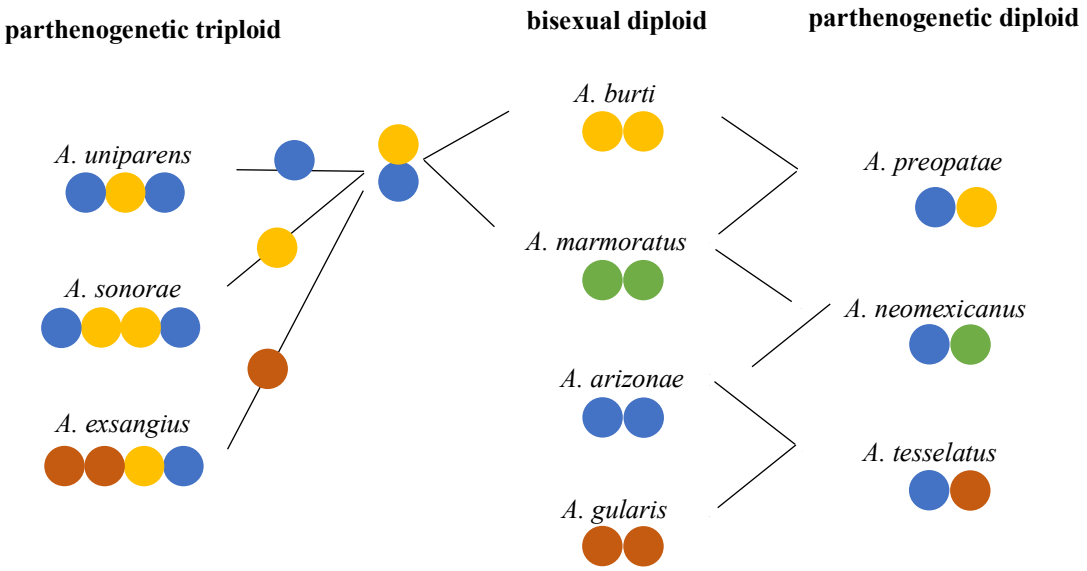


Figure 4. Sexual-asexual network of species which hybridise in whiptail lizards. Diploid parthenogenetic lineages are derived from hybridization events between diploid bisexual species. Triploid parthenogenetic species arise when a diploid obligate parthenogenetic species is crossed with a bisexual male.

I.III. A HIGHLY HETEROZYGOUS LINEAGE OF HYBRID ORIGIN

Whiptail lizards of hybrid origin have maintained the heterozygosity obtained in the first hybridization through reproducing through obligate parthenogenesis. Lutes et al. (2010) [12] revealed that obligate parthenogenetic whiptail species deviate from canonical meiosis; there is a premeiotic doubling of chromosomes, homologous sister chromosomes pair and recombination occurs between them (Figure 5). In addition to the signal at chromosome ends, large tracks of internal telomeric repeats were identified on 13 *A. marmoratus* chromosomes. However, staining with telomeric protein-nucleic acid probes in the *A. arizonae* did not reveal any large tracks of internal telomeric – a signal was visible only at chromosome terminal ends. The large internal telomeric repeats were unique to *A. marmoratus.* Therefore, these could be used to distinguish between *A. arizonae* and *A. marmoratus* chromosomes in *A. neomexicanus.* An adapted fluorescent *in situ* hybridization (FISH) procedure protocol was used to perform hybridization on intact *A. neomexicanus* germinal vesicles. This demonstrated that in a bivalent, hybridization signals were present on both sides, indicating that during meiosis, sister chromosome pair. A second probe, which was a unique marker to nine *A. marmoratus* chromosomes, further confirmed pairing occurred exclusively between sister chromosomes. The premeiotic doubling allows parthenogenetic whiptail species to enter meiosis with twice the usual number of chromosomes to produce oocytes carrying the complete somatic chromosome complement. Moreover, the formation of bivalents from genetically identical sister chromosomes preserves heterozygosity in *A. neomexicanus*.

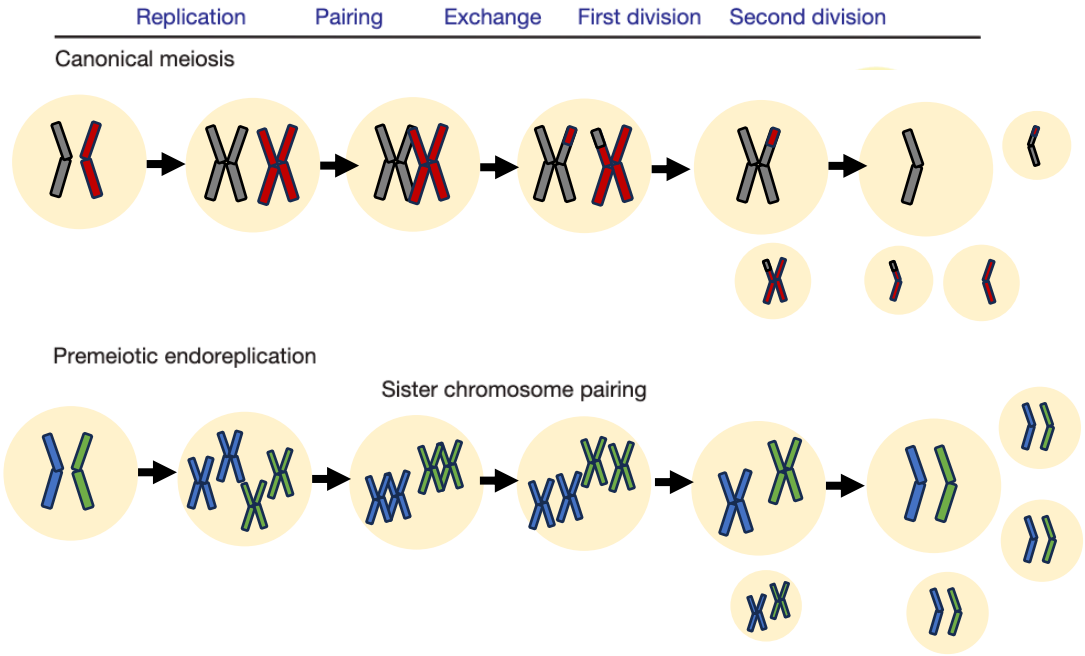


Figure 5. Deviation from canonical meiosis in *Aspidoscelis* lizards. The premeiotic doubling of chromosomes allows for pairing of sister chromosomes. Recombination between pairs of sister

chromosomes in *A. neomexicanus* maintains heterozygosity at all loci. Adapted from Lutes et al. 2010 [12].

The persistence of the heterozygosity observed in *A. neomexicanus* (as well as in other *Aspidoscelis* species) was further confirmed through allozyme analysis [13], [14], [15] . Dessauer and Cole (1986) used electrophoresis of tissue proteins to demonstrate that all individuals within a lineage had identical gene combinations. They concluded that every individual in each lineage constitutes of a genetic clone with fixed heterozygosity. The acceptance of skin grafts between individuals of another diploid parthenogenetic *Aspidoscelis* lineage, *Aspidoscelis tesselata*, further confirmed the genetic uniformity of the species [16]. The acceptance of skin grafts was unique to the parthenogenetic species, demonstrating that their genetic information was consistent between individuals [17]. On the other hand, bisexual species displayed histoincompatibility. This is consistent with what is expected from a species in which recombination occurs and genetic information is exchanged [18].

The clonal parthenogenetic A*spidoscelis* species can be key study systems to investigate a wide range of questions, such as, the effect of Muller's ratchet (the accumulation of deleterious alleles), or the phenomenon of hybrid vigour (the increased fitness a trait in the hybrid offspring), the genetic and epigenetic diversity present in the lizards, as well as investigating the extent of allele-specific expression [19]. As *A. neomexicanus* is an all-female diploid species, with one chromosome set of each parental species, it can be used as a model organism for studying allele-specific expression and the genetic mechanisms regulating this.

I.IV. ALLELE-SPECIFIC EXPRESSION

Allele-specific expression is a phenomenon where the expression of a gene is not equal between the two alleles, which are the two copies of a gene inherited from each parent. In other words, one allele is expressed more than the other, leading to differences in protein production between the two alleles [20]. This difference in expression between the two alleles can be the result of epigenetic gene regulatory mechanisms at work, causing a decrease or increase in expression of one allele. Furthermore, this difference in allelic expression has been shown to cause phenotypic variation between individuals in certain contexts [21]. For example, Shao et al. (2019) identified differential allele-specific expression between two parental alleles in a hybrid strain of rice as a mechanism of heterosis. Certain genes displayed

a direction-shifting pattern of allele-specific expression versus the parental homozygotes. It was hypothesised that the hybrid heterozygous strain of rice was able to use the better allele in growth, development, and environmental adaptation, and therefore leading to higher performance. This demonstrates how the utilization of this heterosis can be used to greatly increase the productivity of crops globally. Allele-specific expression has not only been identified in rice, but it has also been demonstrated in plants, including maize [22], *Arabidopsis* [23, 24], and in barley [25].

Allele-specific expression has also been shown to be an indicator between genetic variation and environment [26]. Knowles, et al. (2017) tested for 30 environmental factors in 8,795 allele-specific genes using RNA-Seq from human blood. Of these, 35 significant associations were found. For example, an interaction between genetics and the environment was observed for the protein associated with high-density lipoprotein cholesterol: VNN1. A higher BMI was observed when an increased allelic imbalance was present in this gene. This was in turn predicted to be causally related to omental fat pad mass.

I.V. ALLELE-SPECIFIC EXPRESSION AS AN INDICATOR OF EPIGENETIC MECHANISMS AT PLAY, GENERATING DIVERSITY IN A CLONAL LINEAGE

Understanding these allele-specific expression differences can help us to understand the maintained fitness and success of parthenogenetic species. Parthenogenetic species of whiptail lizards constitute a striking example of speciation by hybridization. First-generation hybrids attain instant reproductive isolation and reproduce as clonal all-female lineages. The combination of two or more genomes by interspecific hybridization carries the risk of hybrid incompatibilities. However, epigenetic mechanisms such as histone modifications, DNA methylation and non-coding RNAs, which alter the gene expression, without altering the DNA sequence [27] may modulate the effects of hybridization by allowing allele-specific gene expression. Therefore, genetic diversity may arise despite clonal reproduction through epigenetic modulation of gene expression. To investigate the emergence of diversity in a clonal population through epigenetic mechanisms, allele-specific expression was investigated in a parthenogenetic whiptail lizard.

One common approach to studying allele-specific expression is to perform RNA sequencing (RNA-seq) on individuals [28]. Alleles can be distinguished by single nucleotide polymorphisms (SNPs) and the expression level of the two alleles can be compared. As the two genomes in the obligate parthenogenetic diploid *A. neomexicanus* species originate from two different parental species, not only is this species highly heterozygous, but the two alleles of a gene are more distinguishable from each other, than in non-hybrid diploid species. Therefore, it is possible to identify and parse out the allele pairs, to investigate the variation in allele-specific expression in the clonal species.

In the highly heterozygous *A. neomexicanus,* alleles can be identified as either the *A. arizonae* or *A. marmoratus* allele by species specific sequence differences. When investigating allele-specific expression in the *A. neomexicanus* lizards, three allele-specific situations can be observed; both the *A. arizonae* and *A. marmoratus* parental alleles have similar levels of expression, or one of the alleles has a higher expression than the other allele (i.e. bi-allelic expression bias towards the *A. marmoratus* allele or bi-allelic expression bias towards the *A. arizonae* allele). Alternatively, only one parental allele may be expressed: monoallelic expression.

Species specific gene regulatory mechanisms was investigated by Bartoš et al. 2019 [29]. Bartoš et al. 2019 investigated the evolutionary impact of interspecific hybridization and polyploidization and their effect on gene expression patterns. The study focused on a hybrid complex of European spined loaches (*Cobitis*), specifically from parental species *C. elongatoides* and *C. taenia*. These species diverged about 9 million years ago and were initially connected by gene exchange. Currently, reproductive contact between them results in diploid hybrid offspring, which are either sterile males or clonal females. The study addressed the regulation of gene expression, particularly the role of cis- and trans-interactions in the hybrid offspring. The authors explored allele-specific expression as a proxy for cis- and trans-regulatory divergences and parental allele (homoeologs [30]) expression modulation in the hybrids. Homoeologs were identified through species-specific SNPs and negative binomial generalized linear models (GLMs) were used to examine the relationship between the ratio of *C.elongatoides*-specific (Ehyb) to *C. taenia*-specific (Thyb) alleles and the divergence between parental species (E/T divergence). The analysis revealed that both parental alleles were expressed in more than 99% of the investigated genes in all datasets, and complete

silencing of one parental allele was very rare. The distribution of normalized allele-specific expression showed a significant correlation with the expression divergence between parental species, suggesting that cis-regulation may be widespread or that trans-elements' cross-talk has been hampered, resulting in homoeolog expression being mostly governed by their own genome-specific signals. However, the percentage of allele-specific expression variation explained by interspecific expression divergence was relatively low. In contrast, *in silico* simulated hybrids showed a much stronger correlation. This difference indicates that a significant proportion of genes deviated from expectations under pure cis-regulation. Some genes tended to equalize allelic expression (trans-regulation patterns), while others showed cis-/trans-compensation, magnifying the differences between homoeologs

As reproduction in *A. neomexicanus* deviates from canonical meiosis and sister chromosomes pair up, recombination occurs between chromosomes of the same species (Figure 5). This means that the genetic information from each parental species has been largely maintained. Hence, as in the Bartoš et al. study, the evolutionary differences in allele-specific expression between the ancestral version of the species involved in these hybridization events, and their modern-day counterparts can be investigated [31]. Using RNA-Seq data of the parental species of *A. neomexicanus*, it is possible to examine if there has been a directional shift in expression over time. If an allele-pair displays an allelic bias in the obligate parthenogenetic species, it is possible to examine the expression of the gene in the parental species and determine if there has been a change in expression from the ancestral bisexual species to the asexual lineage. As with the hybrid strain of rice, the richness in allele-specific expression observed in *A. neomexicanus* may be an important mechanism of heterosis. The two parental genomes in the obligate parthenogenetic organism may have different sequences and regulatory mechanisms, which can lead to differences in gene expression between the two alleles.

There are several factors that can lead to allele-specific expression, such as genetic variation, epigenetic modifications, and environmental factors. Genetic variation refers to differences in the DNA sequence between the two alleles, such as single nucleotide polymorphisms (SNPs) or copy number variations (CNVs). Epigenetic modifications, such as DNA methylation or histone modifications, can also affect gene expression by regulating the accessibility of DNA to transcription factors and RNA polymerase. Environmental factors, such as diet or exposure to toxins, can also influence allele-specific expression. Insights into the extent and plasticity

of these allele-specific mechanisms can give us further understanding into the genetic and epigenetic differences of clonal species.

Not only is allele-specific expression important for understanding the evolutionary success of an obligate parthenogenetic species, but it can also have important implications for understanding human health and disease. For example, some genetic diseases are caused by mutations that lead to allele-specific expression, where one allele is expressed more than the other. Additionally, differences in allele-specific expression have been implicated in complex traits and diseases, such as cancer [32]. Conversely, dysregulation of allele-specific expression has been implicated in various diseases, including cancer, neurological disorders [33], and autoimmune conditions [34]. Therefore, understanding the mechanisms that regulate to allele-specific expression may help elucidate how allele-specific expression contributes to health and disease.

I.VI. TISSUE-SPECIFIC ALLELIC BIASES

Tissue-specific allele-specific expression refers to when allele-specific expression varies across different tissues or cell types within an organism. This has been evaluated in model organisms such as mice [35] and *Drosophila* [36], in plants such as sugar cane as well as in humans. The most extensive allele-specific expression study in humans was undertaken by Genotype-Tissue Expression (GTEx) and published in 2015 [37]. This study used RNA-Seq to investigate gene expression differences in 1,641 samples across 43 tissues from 175 individuals. Alleles with heterozygous SNPs, which had at least 30 RNA-Seq reads, were tested for allele-specific expression. Of these, 1.5–3.6 % displayed allele-specific expression depending on tissue type. The brain appeared depleted for allelic effects (2.0% allele-specific expression in brain versus 2.7% in other tissues combined), whereas the blood showed high levels of allele-specific expression at 3.6%. This not only highlights the degree of tissue-specificity in gene expression but also the diversity in gene regulatory mechanisms at play [38].

Tissue-specific allele-specific expression has also been investigated in non-model organisms, such as cows[39] and found to be pervasive. Herein, RNASeq data was used to investigate tissue-specific allele-specific expression in 18 tissues of a single cow. Of the 7,985 genes

tested, the proportion of genes showing significant allele-specific expression (where one allele was at a frequency of > 90 %) varied from as low as 4.1–5.8 % in thymus and thyroid and as high 24–38 % in lung with an average of 7–12 % across all tissues. There were a few instances where the direction of allele-specific expression reversed between tissues. The gene *SPTY2D1* (Chromatin Protein Domain Containing 1 gene) showed a paternal allelic bias in kidney and thymus, but maternal allelic bias in the brain caudal lobe and brain cerebellum. Tissue-specific allele-specific expression differences such as this can provide insights into the regulatory mechanisms that govern gene expression and can play a role in determining tissue-specific phenotypic traits.

Tissue-specific allele-specific expression has also been observed in species of hybrid origin which possess one allele from each of its ancestral species [40]. Approximately 120,000 years ago, the Atlantic molly (*Poecilia mexicana*) and the sailfin molly (*Poecilia latipinna*) hybridized and gave rise to the all-female Amazon molly (*Poecilia formosa*). The all-female Amazon molly reproduces through gynogenesis, a form of parthenogenesis. In doing so, the allele of each parental species has been maintained through time. To understand if both ancestral alleles are equally expressed, Real-Time PCR techniques was used to estimate allele-specific expression of the androgen receptor alpha (arα) gene in several tissues. A maternal allelic bias (*P. mexicana*) was observed exclusively in the ovarian tissue. The allelic bias was not observed in the gill or the brain tissue, highlighting how allele-specific expression can play a role in tissue-specific gene expression and function. A second study focused on understanding the transcriptome of the clonal Amazon molly fish. It investigated two key factors that influence gene expression in this species: the fixation of allelic gene expression landscapes and expression bias patterns. Firstly, it was concluded that clonal reproduction, where females produce offspring that are genetically identical clones, leads to a reduced heterogeneity in gene expression compared to sexually reproducing species. Secondly, the study uncovers a pattern of expression bias in the clonal Amazon molly. A significant bias was observed towards one allele in the clonal individuals, indicating a consistent pattern of gene expression. It was proposed that the fixation of allelic gene expression landscapes, combined with the expression bias pattern, contribute to the unique transcriptome of the clonal Amazon molly. These findings provide insights into the mechanisms underlying gene expression regulation in clonal species and shed light on the specific factors that shape the transcriptome of this fish [41].

I.VII. INTER-INDIVIDUAL ALLELIC BIASES

Not only can allele-specific expression be tissue specific, but it can also be an indication of population differences, as observed in the social supergene of the red fire ants (*Solenopsis invicta*) [42]. The social chromosome in red fire ants comes in two variants, SB and Sb. Ants with two SB chromosomes, leads to a colony with a single queen. Ants with one SB chromosome and one Sb chromosome, leads to a colony with multiple queens. Ants with two copies of the Sb variant die when they are young, so the Sb version is inherited in a similar way to the Y chromosome in humans. Martinez-Ruiz et al. [42] described that the Sb variant is in fact breaking down because of the lack of gene shuffling. Therefore, the allelic bias observed in the Sb social chromosome can be an indication of the population ancestry. In the 6 populations of the study, the correlation was in fact stronger between supergene allelic expression – respectively, homozygous SB populations and heterozygous Sb populations had higher correlations. This demonstrates how allelic bias differences between populations may lead to insights on their ancestry and evolution.

Despite being a genetically clonal species, gene regulatory mechanisms may have allowed for distinct allelic biases to evolve. To investigate population level differences in *A. neomexicanus*, three wild populations were tested for allele-specific differences which may be unique to each population. Allele-specific differences were also investigated in a lab population of *A. neomexicanus* to understand if selective pressures may be a driving force for the emergence of allelic biases.

In a clonal species such as *A. neomexicanus*, which reproduces asexually by producing genetically identical copies of itself, one might expect very little inter-individual variation. However, even in clonal species, there can be differences in various traits, such as morphology, physiology, behavior, and gene expression [43]. One source of inter-individual variation in clonal species is somatic mutations, which are genetic changes that occur in cells during an organism's lifetime. Somatic mutations can arise spontaneously during DNA replication or in response to environmental factors such as radiation or chemical exposure [44]. These mutations can accumulate over time in different cells within an individual, leading to variation in gene expression or protein function. However, inter-individual variation in clonal species could be due to epigenetic modifications, which are changes in gene expression that do not involve changes to the DNA sequence itself.

Epigenetic modifications can be inherited from one generation to the next and can be influenced by environmental factors such as temperature, light, and nutrient availability [45]In addition, inter-individual variation in clonal species can also arise from environmental factors such as resource availability, predation pressure, and competition. Even genetically identical individuals can experience different environmental conditions and respond to them in different ways, leading to variation in traits such as growth rate, reproductive output, and survival. As these factors differ across the wild *A. neomexicanus* populations and lab colony of *A. neomexicanus*, they could be influencing gene expression and the allele-specific expression. In addition to population level allelic biases, individuals may respond differently to abiotic factors and have developed unique allelic biases. Therefore, inter-individual allelic biases could be a source of genetic variability introduced in a clonal population.

In clonal species, epigenetic modifications can lead to differences in gene expression between individuals, even if they are genetically identical. Overall, while clonal species may be expected to have very little inter-individual variation, a variety of factors can lead to differences in traits and characteristics between individuals. Understanding the sources and extent of inter-individual variation in clonal species can provide insights into the mechanisms underlying adaptation and evolution in asexual populations.

I.VIII. HIGH QUALITY GENOME ANNOTATIONS

The release of the first complete bacterial genome, *Haemophilus influenzae*, in 1995, the 1.83 megabase (Mb) sequence was accompanied by annotation of 1,742 protein-coding genes [46]. Since then, Next-Generation Sequencing (NGS) technology has expedited the release of a multitude of DNA and RNA data from a broad range of species [47]. However, this vast amount of data still needs to be reconstructed into meaningful pieces of information. Genome annotations of protein-coding genes are one form of identifying information encoded in the sequencing data.

The main challenges in generating high quality genome annotations arise from the quality of the genome assembly. If the assembly is highly erroneous, the resulting genome annotations will be of poor quality. If the genome assembly is not highly contiguous and has a low N50

(length for which the collection of all contigs of that length or longer covers at least 50% of assembly length) then the genes could be fragmented in the genome, making them difficult to identify [48]. If there are errors and contamination in the draft assemblies, this can lead to the propagation of errors in the annotations. To overcome this, improvements in sequencing such a long-read technology and cross-linked read information can be used to increase the quality of the assembly. A combination of these sequencing types was used to assemble the human genome.

The human genome is the most complete genome assembly available, now on the 38th major release, telomere to telomere version [49]. However, even with a high quality and highly complete genome, a revised genome annotation is undertaken every 12-18 months by RefSeq [50] to integrate newly available data. In addition to full annotation releases, RefSeq releases provisional annotations approximately every 3 months to coincide with GRC patch releases. This demonstrates how generating and maintaining a high-quality genome annotation, even with a high-quality genome assembly, is a difficult task.

The *A. neomexicanus* (AspMarm2.0_AspAri2.0) reference genome used in this study is the concatenation of the chromosome level genome assemblies of the parental genomes: *A. marmoratus* (AspMarm2.0) and *A. arizonae* (AspAri2.0). To generate high quality, chromosome level, genome annotations a variety of sequencing data was used. Paired-end short reads, Chicago libraries and HiC contact mapping data were incorporated to generate the *A. marmoratus* and *A. arizonae* genome assemblies respectively. The genome completeness and high quality can be observed by the long N50s and high percentage of complete BUSCO [51] genes identified using the 'vertebrata_odb10' database (Table 1).

Table 1. Metrics of genome completeness for chromosome level *A. arizonae* and *A. marmoratus* genome assemblies

	A. arizonae	***A. marmoratus***	***Anolis carolinenesis*** [52]
Size (Gb)	1.53	1.64	1.78
Scaffold N50 (Mb)	106.15	113.09	4.033

BUSCO (vertebrata lineage, n:3354)	C:96.5%[S:95.8%, D:0.7%],F:1.7%,M:1.8%	C:97.7%[S:97.1%,D:0.6%],F:0.9%,M:1.4%	C:88.6%[S:87.4%,D:1.2%],F:5.0%,M:6.4%
# of genes annotated	20,610	25,550	17,472

In response to the huge amount of sequencing data being generated, a variety of genome annotation software's have been released [53]. A widely used annotation software, MAKER uses its gene-prediction algorithm to automatically retrain data so that each run outputs a higher-quality gene-model [54]. MAKER uses repeats and RNA-Seq and protein data to produce ab initio gene predictions. BRAKER uses GeneMark-ES/ET and AUGUSTUS to predict genes and trains both gene finders in a fully automated fashion before applying them to the genome [55]. However, AUGUSTUS uses statistical models (Generalized Hidden Markov Model) with species specific parameters and has been shown to be one of the most accurate tools for predicting genes [56] [57]. As with many gene annotations software's, AUGUSTUS needs a training gene set. This can be done by including in RNA-Seq hints to guide the annotation software to the presence and location of genes. This helps increase the accuracy of the annotations. As AUGUSTUS [58] was shown to produce the highest quality genome annotations, this software was used to produce genome annotations for the parental species of *A. neomexicanus* and therefore a genome annotation for *A. neomexicanus* itself. Additionally, strict filtering criteria were determined to filter allele-pairs in *A. neomexicanus* to exclude false-positive instance of allele-specific expression.

I.IX. ALIGNING TO A GENOME OF TWO CLOSELY RELATED SPECIES

Differential gene expression analysis relies heavily on the correct quantification or RNA-Seq reads to transcripts. This in turn relies heavily on the correct placing of reads in the transcriptome of an organism [59]. However, when aligning to a genome consisting of two closely related species, there is a higher probability of short reads being misplaced. As the misplacement of RNA-Seq reads in *A. neomexicanus* may affect correct quantification of transcript expression and therefore identification of the allelic bias, two widely used RNA-Seq aligners (STAR [60] and HISAT2 [61]) were assessed in their accuracy in placing RNA-Seq. HISAT2 is a based on a compressed full-text substring index of Burrows-Wheeler transform

(BWT) [62] for graphs, whereas STAR [60] uses sequential maximum mappable seed search in uncompressed suffix arrays to align reads. Additionally, the effect of the read length was observed on the read mapping quality.

II. MATERIALS AND METHODS

II.I. TISSUE COLLECTION

To investigate tissue-specific allele-specific expression in *A. neomexicanus,* tissues were collected from three lizards: D15, D37 and D41. The lizards were hatched on the 9th of January 2019, 28th of January 2019 and 3rd of February 2019, respectively and were sacrificed on the 5th of February 2020 at the age of 1 year old. The lizards were placed in the fridge for 45 minutes prior to being scarified, to slow their metabolic rate. A toe pinch was used to determine the reactiveness of the lizard before decapitation using scissors. The liver, lung, heart, tail, thigh muscle, brain and heart were collected in pre-cooled 1.5 mL Eppendorf RNAse and DNAse free tubes, which had been placed on ice, and subsequently flash frozen in liquid nitrogen. The ovaries were only collected from lizards D15 and D41 as the ovaries in D37 were atretic. The carcasses were stored in 15 mL tubes, flash frozen and placed at -80 °C for long-term storage.

To investigate intra-individual allele-specific expression in wild *A. neomexicanus*, samples were collected in Socorro (n=17), Bernalillo (n=9) and Doña Ana (n= 4) in New Mexico, USA, by collaborator Randy Klabacka. Blood samples were collected using capillary tubes and frozen in liquid nitrogen in the field. Liver and skeletal muscle samples were preserved in 2 mL of RNA-Later prior to being flash frozen in the field. The samples were stored at -80 °C until their shipment to Mainz, Germany.

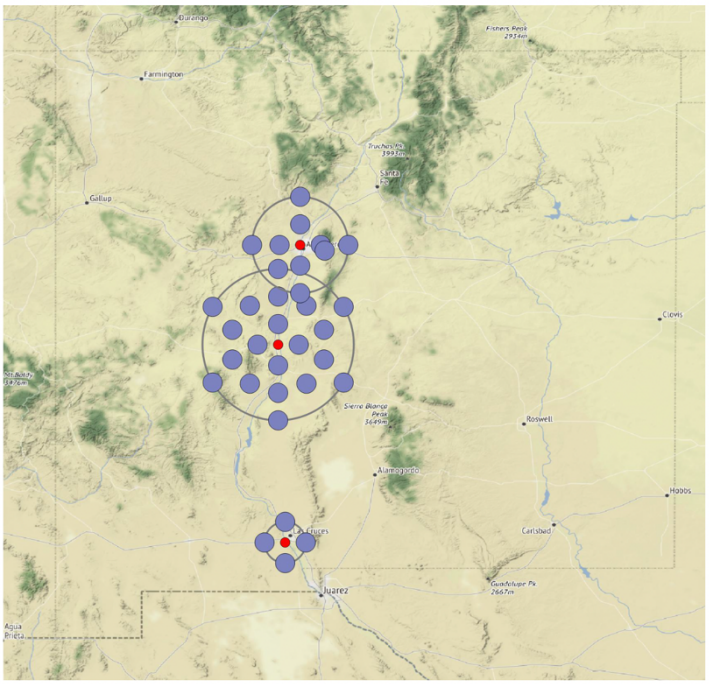


Figure 6. Map of depicting the locations of where the wild *A. neomexicanus* samples were collected in the Rio Grande basin (from Albuquerque to Las Cruces) in New Mexico: 17 samples were collected in Socorro, nine in Bernalillo and four in Doña Ana. Image provided by R. Klabacka.

II.II. RNA PREPARATION

Total RNA was extracted from the ovaries, liver, lung, heart, tail, thigh muscle, brain, and heart of the three lizards using a Trizol-RNeasy Mini Kit protocol [63], over three days to investigate tissue-specific allele-specific expression. The ovaries, liver, lung, heart, thigh muscle, brain and heart tissues were homogenized in a 1.5 mL tube containing 1 mL- 1.2 mL of Trizol, depending on tissue mass, using a plastic pestle. The tail tissues were homogenized in liquid nitrogen in a RNase and DNase free pestle and mortar. The pestle and mortar were pre-cooled using liquid nitrogen. After the sample was ground to a fine dust, Trizol was added to the sample in the mortar. Phase separation of sample was achieved by adding one fifth chloroform of Trizol volume (200 µL -280 µL) to the samples, vigorously mixing and incubating at room temperature for 3 minutes. Samples were centrifuged at 4 °C for 10 mins at 12,000 x g. The aqueous phase containing the total RNA was carefully extracted from each

sample after centrifugation. The DNA interphase and protein organic phase were discarded. To precipitate the RNA, 600 µL of isopropanol was added to the sample. Of this, 700 µL was loaded onto the silica membrane of a Qiagen RNeasy Mini Kit spin column and centrifuged at 8,000 x g for 15 seconds, the flow-through was then discarded. The remaining sample was loaded onto the column and centrifuged at 8,000 x g for 15 seconds. The samples were washed with 350 µL of Buffer RW1 and centrifuged at 8,000 x g for 15 seconds. The flow-through was discarded and the spin column was placed into a new collection tube in preparation for the on-column DNase digestion. A DNase solution was prepared using 10 µL Qiagen DNase I and 70 µL Qiagen Buffer RDD per sample. Each sample was DNase treated by adding 80 µl DNase solution to the column and incubated for 15 minutes at room temperature. A second wash with 350 µL of Buffer RW1 terminated the DNase treatment. To remove remaining salts, 500 µL of Buffer RPE was added to the sample and centrifuged at 8,000 x g for 15 seconds. A further wash was performed by adding 500 µL of Buffer RPE. The samples were then centrifuged for 2 minutes at 8,000 x g to dry the silica membrane. The samples were placed in fresh collection tubes and centrifuged for an additional 1 minute at 8,000 x g to completely dry the column and remove carry over ethanol. Total RNA was eluted from the silica column using 60 µL – 110 µL ddH2O, depending on the starting tissue mass. The samples were left to elute for 4 minutes before centrifugation of 1 minute at 8,000 x g. The samples were re-eluted by adding the same ddH20 to the column for a further 4 minutes, then centrifuged for 1 minute at 8,000 x g.

To investigate inter-individual allele-specific expression in *A. neomexicanus*, total RNA was extracted from wild and lab *A. neomexicanus* blood samples using a modified Trizol-RNeasy Mini Kit extraction protocol. Modifications included diluting blood samples in a total of 4.2 mL of Trizol. This was divided into three 1.2 mL aliquots. Two aliquots were flash frozen in liquid nitrogen and the total RNA was extracted from the third. Phase separation and subsequent total RNA extraction steps were performed as described above.

Total RNA quantification was performed using the Qubit RNA Assay Kit (Life Technologies), as per the manufacturer's specifications. The total RNA integrity was assessed using a 2100 BioAnalyzer (Agilent Technologies). Samples were run on an RNA Nano Chip and prepared as per manufacturer's specifications. Samples with an RNA integrity value greater than 8, indicating low levels of degradation, were prepared for sequencing.

II.III. *A. NEOMEXICANUS* RNA-SEQ

Library preparation of RNA tissue samples from lizards D15, D37 and D41 was performed by the Genomics Core facility of the Institute of Molecular Biology using the Illumina TruSeq Stranded mRNA Library Prep Protocol [64] for Poly(A) selected RNA sequencing. To achieve a larger insert size for 2 x 150 bp paired-end sequencing, the standard fragmentation time of 8 minutes at 94 °C, which yields an insert size of 120 -200 bp, was replaced with a fragmentation step of 2 minutes at 80 °C. The 20 samples were pooled in an equimolar ratio and sequenced on an Illumina NextSeq 500 over two flow cells, at the Genomics Core facility of the Institute of Molecular Biology to achieve approximately 40 million reads per sample.

Poly(A) selected stranded RNA-Seq data of tissues from five additional lizards were included in the tissue-specific analysis (Table 2). These tissues were sequenced at Stowers Institute for Medical Research, in Kansas City. Blood samples from *A. neomexicanus* 10935, 11118 and 11174 were equimolarly pooled and paired end, 2 x 100 bp, sequenced on an Illumina HiSq2500. Heart and lung samples for individual 11255 were pooled in an equimolar ratio with four *A. marmoratus* samples and 2 x 100 bp, sequenced on an Illumina HiSeq 2500. The liver, heart, thigh muscle, follicles, and brain from *A. neomexicanus* 14584 were 2 x 100 bp, sequenced on an Illumina HiSeq 2500, alongside two *A. arizonae* samples.

Table 2. Summary of which tissues from each lizard were used for analysis.

	14584	**11255**	**D15**	**D37**	**D41**	**10935**	**11118**	**11174**
Heart	x	x	x	x	x			
Brain	x		x	x	x			
Liver	x		x	x	x			
Lung		x	x	x	x			
Thigh	x		x	x	x			
Tail			x	x	x			
Blood						x	x	x

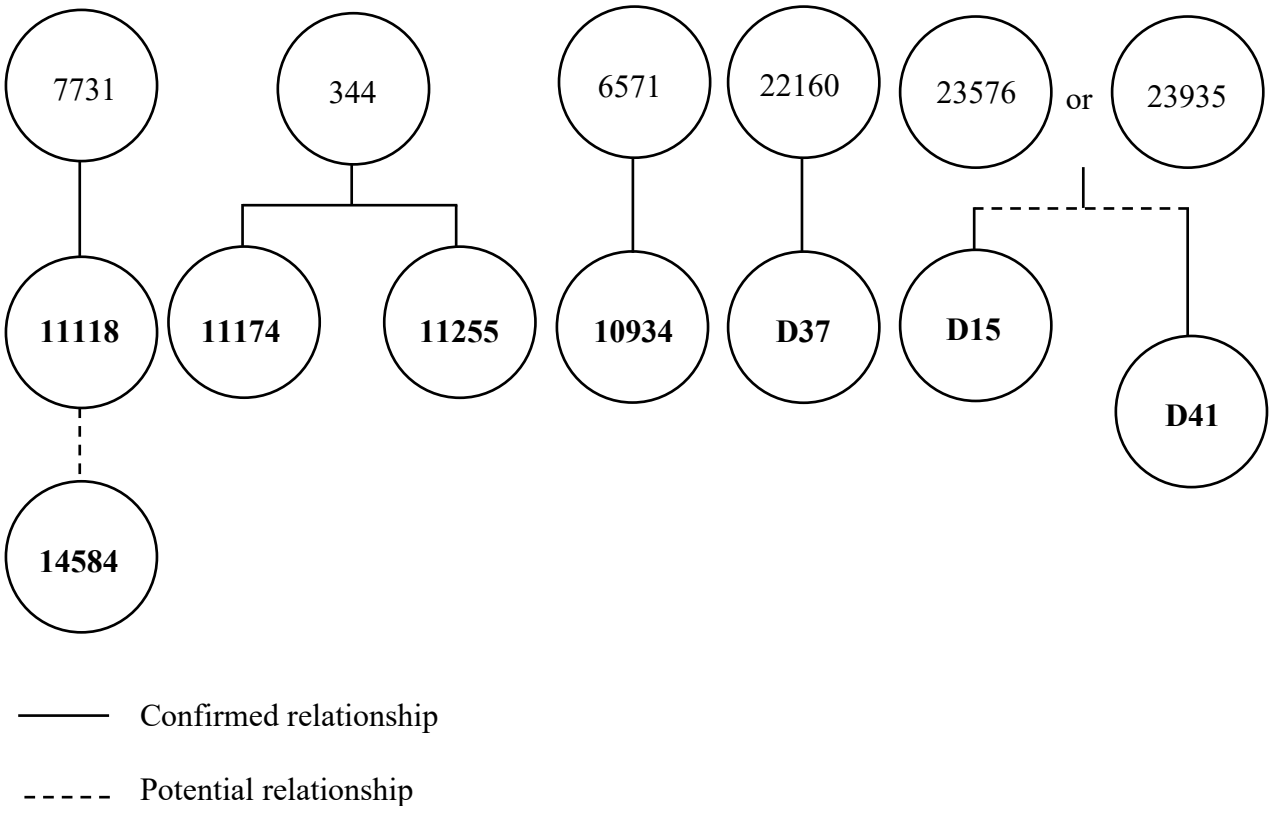


Figure 7. Relationships between the *A. neomexicanus* in the lab colony which were used for the tissue-specific expression analysis.

A combination of Ribo-depleted Stranded RNA-Seq and Poly(A) Stranded RNA-Seq data, generated at Stowers Institute for Medical Research, in Kansas City, were used to create *in silico A. neomexicanus* tissues. These serve as a reference for the expression of an allele-pair in the parental species of *A. neomexicanus*. (Table 3).

Table 3. Summary of the sequencing data used to create *in silico A. neomexicanus* reference tissues.

Sample	MOLNG-Order	Flow Cell Tag	Sequencing Type	Read Length
Aari12852 liver	MOLNG-1684	HTTMTBCXXa	Poly(A) Stranded RNA-Seq	100 bp PE
Amarm16996 liver	MOLNG-1713	HVJM7BCXX	Poly(A) Stranded RNA-Seq	100 bp PE
Aari13518 lung	MOLNG-1701	HVJMHBCXX	Poly(A) Stranded RNA-Seq	100 bp PE
Amarm16996 lung	MOLNG-1713	HVJM7BCXX	Poly(A) Stranded RNA-Seq	100 bp PE

Aino13518 thigh	MOLNG-1702	HVKTVBCXX	Ribo-dep Stranded RNA-Seq	100 bp PE
Amarm10734 thigh	MOLNG_1741	HVM7NBCXXv2	Ribo-dep Stranded RNA-Seq	100 bp PE
Aino18789 brain	MOLNG-1656	HMHHLBCXXb	Ribo-dep Stranded RNA-Seq	100 bp PE
Amarm10734 brain	MOLNG_1741	HVM7NBCXXv2	Ribo-dep Stranded RNA-Seq	100 bp PE
Aino18789 heart	MOLNG-1656	HMHHLBCXXb	Ribo-dep Stranded RNA-Seq	100 bp PE
Amarm16996 heart	MOLNG-1714	HVMJGBCXX	Ribo-dep Stranded RNA-Seq	100 bp PE
Aino9601 blood	MOLNG-897	H9W77ADXX	Poly(A) Stranded RNA-Seq	100 bp PE
Amarm11225 blood	MOLNG-897	H9WE1ADXX	Poly(A) Stranded RNA-Seq	100 bp PE

The libraries for the 29 *A. neomexicanus* samples (Table 4, Figure 8) sequenced to investigate inter-individual differences were prepared using the Illumina RNA Library Prep with Illumina® Stranded mRNA Prep Ligation to achieve 2 x 150 bp paired-end libraries. As only one sample had a lower concentration compared to the remaining samples, the 29 samples were normalised to the sample of second lowest concentration during library pooling. The total ng available for the sample of lowest concentration was included in the library pool. The pooled libraries were sequenced on an Illumina NextSeq 2000 to achieve approximately 40 million reads per sample.

Table 4. Summary of the 29 *A. neomexicanus* samples prepared for RNA-Seq to investigate inter-individual differences.

Sample ID	Date Collected	State	County
RLK086	29 June 2018	New Mexico	Socorro
RLK087	29 June 2018	New Mexico	Socorro
RLK089	29 June 2018	New Mexico	Socorro
RKL090	29 June 2018	New Mexico	Socorro
RLK099	29 June 2018	New Mexico	Socorro
RLK100	29 June 2018	New Mexico	Socorro
RLK101	29 June 2018	New Mexico	Socorro
RLK102	30 June 2018	New Mexico	Socorro
RLK107	30 June 2018	New Mexico	Socorro

RLK108	30 June 2018	New Mexico	Socorro
RLK120	2 July 2018	New Mexico	Bernalillo
RLK121	2 July 2018	New Mexico	Bernalillo
RLK122	2 July 2018	New Mexico	Bernalillo
RLK126	3 July 2018	New Mexico	Bernalillo
RLK127	3 July 2018	New Mexico	Bernalillo
RLK208	28 May 2019	New Mexico	Doña Ana
RLK212	29 May 2019	New Mexico	Doña Ana
RLK213	29 May 2019	New Mexico	Doña Ana
RLK214	30 May 2019	New Mexico	Doña Ana
D668	11 April 2023	Mainz	Lab
D675	11 April 2023	Mainz	Lab
D681	11 April 2023	Mainz	Lab
D606	05 May 2023	Mainz	Lab
D677	05 May 2023	Mainz	Lab
D559	05 May 2023	Mainz	Lab
D547	05 May 2023	Mainz	Lab
D667	05 May 2023	Mainz	Lab
D654	05 May 2023	Mainz	Lab
D544	06 April 2023	Mainz	Lab

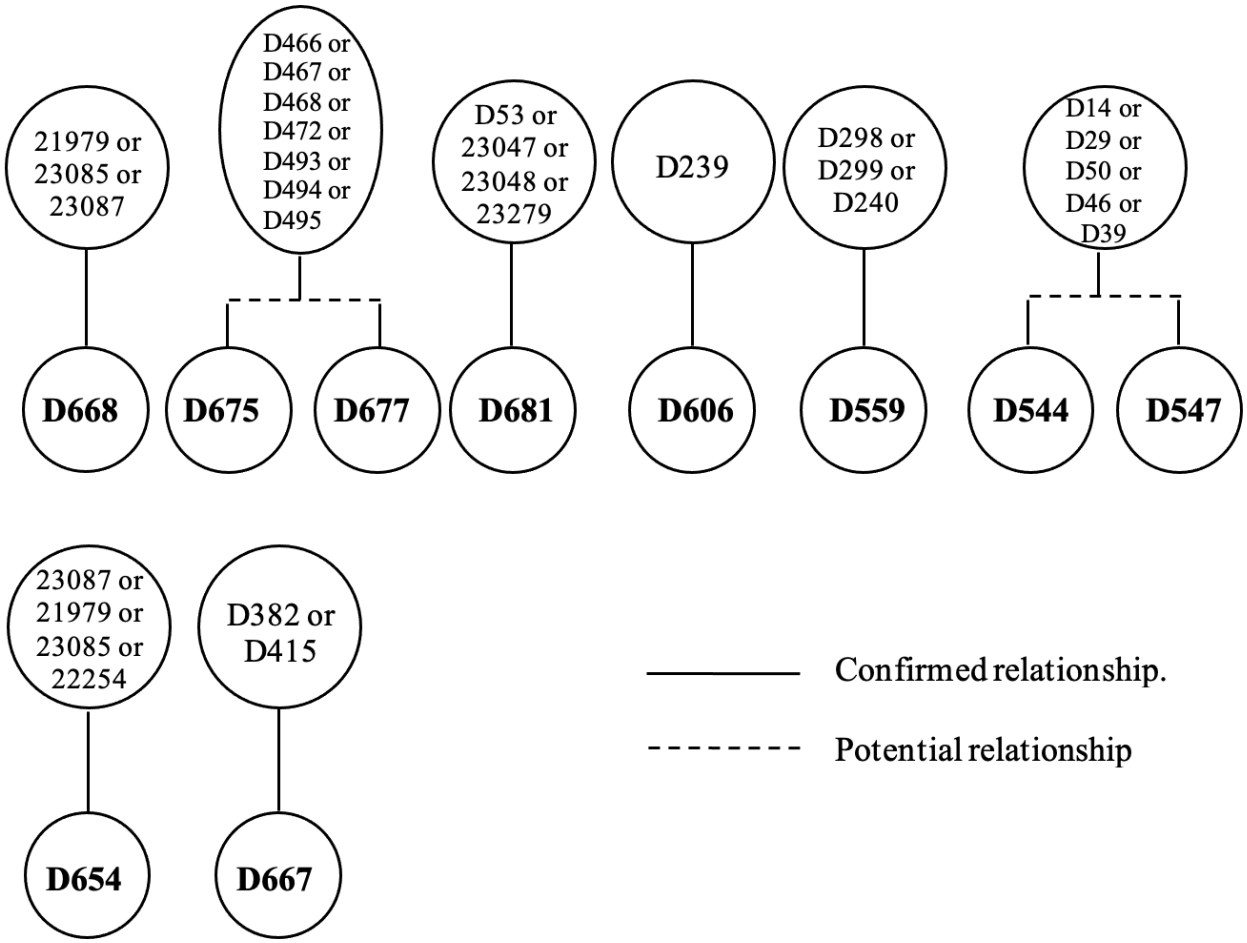


Figure 8. Relationships between the ten *A. neomexicanus* samples (in bold) originating from the lab population, sequenced to investigate inter-individual differences in allele-specific expression.

II.IV. GENERATING HIGH QUALITY GENOME ANNOTATIONS

Chromosome level, Hi-C assembled genomes *A. arizonae* and *A. marmoratus* were aligned by Nathaniel Deimler using minimap2 v2.17-r941 [65] with the following parameters to promote long syntenic regions of alignments: a mismatch penalty of 1, a matching score of 50, a gap open penalty of 3,3 and a gap extension penalty of 1,1, and asm5 for ~ 0.1% sequence divergence.

TruSeq3-PE adapters 2:30:10 were removed from raw RNA-Seq data (Table 5) with IlluminaClip. Reads were trimmed using Paired End Trimmomatic version 0.36 [66], with

MINLEN 80, TRAILING 7, LEADING 7, SLIDINGWINDOW 5:20 and aligned to the appropriate reference genome using STAR version 2.7.8a [60]. Non-uniquely mapping reads were discarded. The alignment files were converted to RNA-Seq hints using the AUGUSTUS version 3.4.0 [58] built in bam2hints script. Annotations for each genome were produced using AUGUSTUS Comparative Gene Prediction (CGP) algorithm with the built in chicken reference, as well as with the RNA-Seq hints.

Table 5. RNA-Seq hints used for *A. marmoratus* and *A. arizonae* genome annotations.

Species	**MOLNG-Order**	**Tissue**	**Sequencing Type**	**Read Length**
A. marmoratus	MOLNG-1741	Testes, brain & thigh	Ribo-dep Stranded RNA-Seq	100 bp PE
A. marmoratus	MOLNG-1714	Heart, lung, liver & follicles	Ribo-dep Stranded RNA-Seq	100 bp PE
A. marmoratus	MOLNG-1713	Heart, lung, liver & follicles	Poly(A) Stranded RNA-Seq	100 bp PE
A. arizonae	MOLNG-1684	Germinal beds	Poly(A) Stranded RNA-Seq	100 bp PE
A. arizonae	MOLNG-1655	Brain, heart & testes	Poly(A) Stranded RNA-Seq	100 bp PE
A. arizonae	MOLNG_1656	Brain, heart & testes	Ribo-dep Stranded RNA-Seq	100 bp PE
A. arizonae	MOLNG-334	Unknown mixture of tissues	Poly(A) Stranded RNA-Seq	100 bp PE
A. arizonae	MOLNG_333	Unknown mixture of tissues	Poly(A) Stranded RNA-Seq	100 bp PE

Overlapping and duplicate gene predictions were identified using GffCompare version 0.12.6 [67]. Duplicate gene predictions were removed and overlapping predictions were combined into a single gene prediction. New gene predictions were aligned to the UniProt Reviewed Human Protein Database using Blast version 2.11.0 [68]. Genes within 80kb that aligned to the same reference protein were further combined. GffCompare [67] was used to merge overlapping gene predictions again. Single Exon genes were aligned by Blast to SinEx DB 2.0 mouse single exon genes. Single exon genes that aligned with a percent identity of greater than 70% with an alignment length of 60% of the protein length were retained.

Annotation completeness was measured using BUSCO version 5.0.0 using Sauropsida_odb10 and Vertebrata_odb10 reference database. The 41 HOX genes previously identified in lizards [69], were aligned against the predicted genes using Blast searching for alignments of 90% the length of the hox gene.

Identified genes were aligned to a Human Protein reference database via Blast version 2.11.0 [68] for functional identification; best hits based on alignment length and percent identity were retained. Eggnog version 2.1.6 [70] and InterProScan version 5.52-86.0 [71] were used to identify protein domains. InterProScan results were filtered to remove MobiDBLite Disordered protein predictions and functionality was assigned for both InterProScan and EggNog based on the longest alignment hit. Functional assignments were incorporated into the attribute field of the GFF. Allele-pairs were assigned by reciprocal best hit (e-value) Blast version 2.11.0 [68].

II.V. ALIGNMENT STRATEGY TO DETERMINE MISS MAPPING FREQUENCY WHEN ALIGNING TO A GENOME OF TWO CLOSELY RELATED SPECIES

When aligning to a genome consisting of two closely related species, it is especially critical to map reads from each respective genome correctly, especially in the context of allele-specific expression. This is essential to understand if the allele-specific expression observed is not due to a technical artefact of reads being miss placed. The *A. neomexicanus* genome used in this analysis is a concatenation of the reference genomes of the parental species of the lineage; of the *A. marmoratus* and *A. arizonae*. This serves as a proxy for the *A. neomexicanus* genome. The frequency at which reads containing *A. arizonae* information were being correctly mapped to the *A. arizonae* region of the reference genome, and vice versa, was determined for *A. neomexicanus*, using data from the parental species. Sequencing data from Stowers was available for *A. arizonae* and *A. marmoratus* (Table 3). Reads belonging to these respective samples can be identified by their flow cell tags, generated by the Illumina machine. Though using this identifier, it would be possible to quantify how many times an *A. arizonae* read is correctly mapped to an *A. arizonae* chromosome versus being incorrectly mapped to an *A. marmoratus* chromosome, when both genomes are available for read mapping.

The quality of sequencing files was assessed using FastQC version 0.11.9 [72]. Adapter and quality trimming was performed with Trimmomatic version 0.39 [66]. The leading and trailing bases of each read were trimmed if their quality dropped below a Phred score of 3. The reads were scanned in a 4-base window. If the average quality dropped per base dropped below a Phred score of 15, the read was cut. If reads were cut to a 36 bases or shorter, they were removed. Of the trimmed reads, 20 million reads were subsampled for *A. arizonae* and *A. marmoratus*, respectively, using seqtk and combined into *in silico A. neomexicanus* R1 and R2 paired end reads. The data was subset to a total of 40 million reads per *in silico* tissue to resemble the sequencing depth of the *A. neomexicanus* RNA-Seq. This was done for the brain, liver, lung, thigh, heart, and blood. Next, two aligners were employed to align the *in silico* reads: STAR version 2.7.10a [60] and HISAT2 version 2.2.1 [61]. This was done to determine which aligner had the higher accuracy in placing reads and therefore should be used to align *A. neomexicanus* reads to the concatenated genome (Figure 9). Blood 100 bp *in silico* reads were trimmed by 30 bases (15 leading and 15 lagging), 20 bases (10 leading and 10 lagging) and 10 bases (5 leading and 5 lagging) using seqtk [73] to determine the effect of aligning shorter reads to a genome consisting of two closely related species.

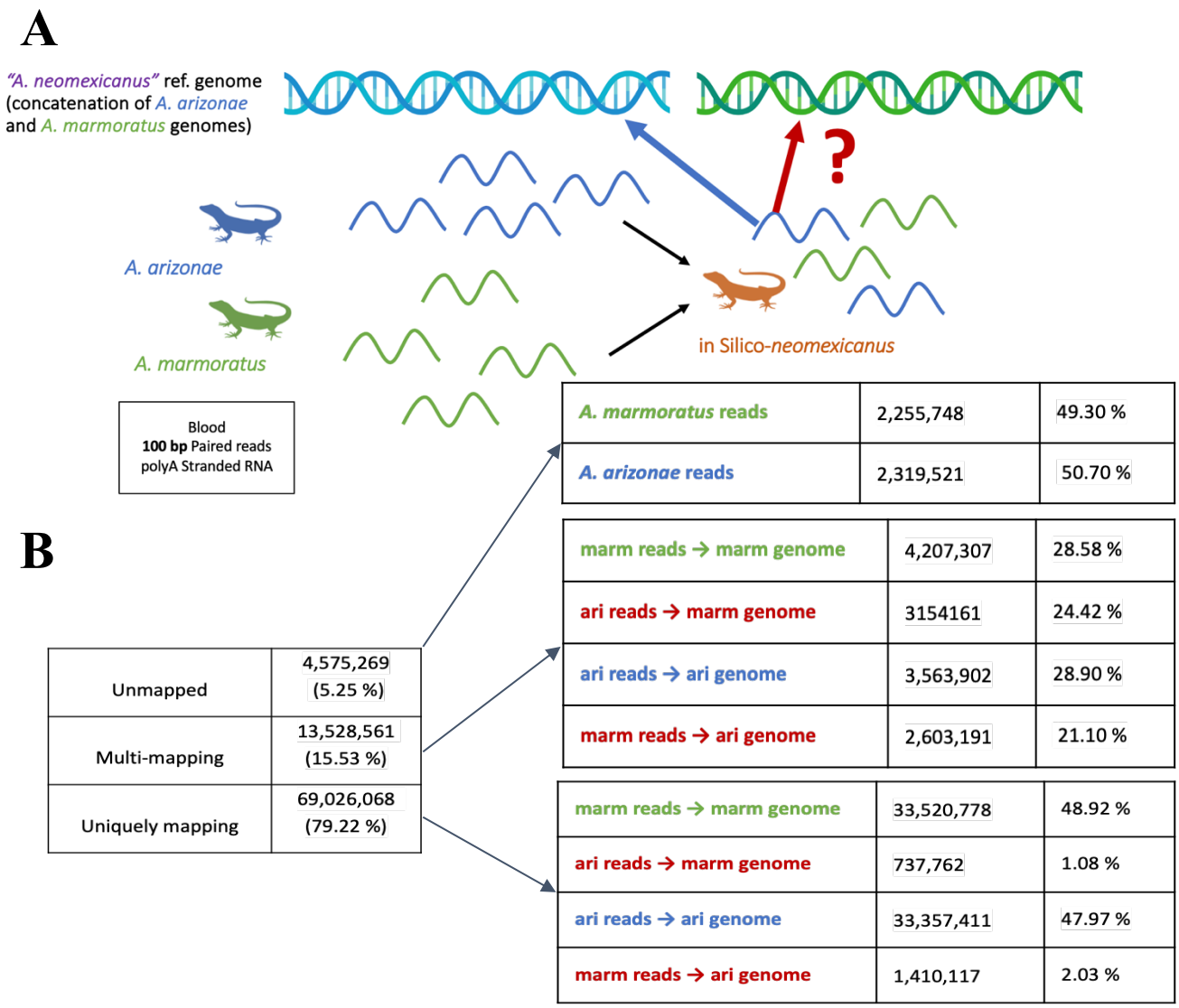


Figure 9. (A) Generating *in silico A. neomexicanus* tissue references. When sequencing data from *A. marmoratus* and *A. arizonae* is combined into an *in silico neomexicanus* reference, it is possible to identify the false positive mapping frequency when aligning to a genome concatenation of the *A. arizonae* and *A. marmoratus* reference genomes. The read tags can be used to determine if an *A. arizonae* read is mapping to an *A. arizonae* chromosome, or vice versa. (B) Additionally, the miss-mapping frequencies were calculated for the uniquely mapping, multi-mapping and unmapped reads. This example shows the miss-mapping frequency for the *in silico* blood sample aligned using HISAT2.

Firstly, a STAR [60] guided genome index was generated for the *A. neomexicanus* reference genome. The genome index was guided by the genome annotations were created with Augustus-CPG [58]. Trimmed sequencing files were aligned using STAR with parameters: --alignSJDBoverhangMin 5 --alignSJoverhangMin 10 --alignIntronMin 20 --alignIntronMax

50000 --alignEndsType EndToEnd --outSAMunmapped Within --quantMode GeneCounts --twopassMode Basic. Secondly, a HIST2 genome index was generated for the *A. neomexicanus* genome. Paired-end *in silico* files were aligned using HISAT2 --rna-strandness RF. For each alignment, the percentages of uniquely mapping, unmapped and multi-mapping reads per sample were calculated based on the SAM flags “NH:i:1”, 'NH:i:0', and “NH:i: (anything greater than 1)” respectively. The miss-mapping frequency for the *in silico* tissue references was calculated using the respective *A. arizonae* and *A. marmoratus* read tags. This was used to quantify the percentage of correctly mapping *A. arizonae* reads to the *A. arizonae* genome and *A. marmoratus* reads to the *A. marmoratus* genome versus *A. arizonae* reads to the *A. marmoratus* genome versus reciprocal mapping.

II.VI. *A. NEOMEXICANUS RNA-SEQ* ALIGNMENT STRATEGY

The quality of sequencing files was assessed using FastQC version 0.11.9 [72]. Adapter and quality trimming was performed with Trimmomatic version 0.39 [66]. The leading and trailing bases of each read were trimmed if their quality dropped below a Phred score of 3. The reads were scanned in a 4-base window. If the average quality dropped per base dropped below a Phred score of 15, the read was cut. If reads were cut to a 36 bases or shorter, they were removed. Contamination from human, mouse, Saccharomyces cerevisiae or Schizosaccharomyces pombe was assessed using FastQ Screen version 0.15.3[74]. HISAT2 version 2.2.1 was used for aligning the sequencing files and generating gene counts. Firstly, a guided genome index was generated for the *A. neomexicanus* reference genome. The genome index was guided by the genome annotations were created with Augustus-CPG [75]. Trimmed sequencing files were aligned using HISAT2 with parameters: --rna-strandness RF. HTSeq version 0.11.1 [76], with the stranded=reverse option, was used to generate gene counts. Only uniquely mapping reads were used for gene counting. Bigwig files normalised by reads per kilobase per million mapped reads (RPKM) were generated with deepTools version 2.0 [77] for visualisation in IGV version 2.16.0 [78].

II.VII. GDNA PREPARATION

To exclude investigating allele-specific expression resulting from genetic differences due to regions of loss of heterozygosity (LOH), gDNA from *A. neomexicanus* individuals was extracted and subjected to sequencing. High molecular weight gDNA was extracted from approximately 100 mg of tail tissue from *A. neomexicanus* lizards 23563, 14584, D15 and D37, using the Blood & Cell Culture DNA Midi Kit (Qiagen, Cat No./ID: 13343). Firstly, for each sample, Buffer G2 was prepared by combining 19 µL RNase A to 9.5 mL Buffer G2. Samples were added to a pestle and mortar, pre-cooled with liquid nitrogen, and homogenised into a fine powder. The prepared Buffer G2 and 500 µL Proteinase K was added to the mortar. The mixture was placed into a 50 mL rapid-spin conical tube and incubated sample at 50 °C for 3 hours to lyse the tissue. For each sample, a 100/G genomic-tip was placed into a 50 mL conical tube and equilibrated by adding 4 mL Buffer QBT. After tissue lysis the samples were centrifuged at 3,200 rpm for 10 minutes at room temperature and the supernatant transferred to a new 50 mL rapid-spin conical tube. The samples were immediately to the equilibrized genomic-tip after vortexing at maximum speed for 10 seconds and left to flow through. Once through, 7.5 mL of Buffer QC was added, and the step repeated. The genomic tip was suspended over a fresh 15 mL rapid-spin conical tube and 3.5 mL of pre-warmed Buffer QF was added. A 1:1 ratio of room temperature isopropanol (3.5 mL) was added to elute the DNA, as well as 1 µL of GlycoBlue. The sample tube was inverted 20 times to precipitate the DNA. The supernatant was removed, and the DNA pellet was washed with 2 mL cold 70% ethanol, vortex briefly and centrifuged at 7,500 x g for 10 minutes at 4 °C. The supernatant was removed without disturbing the pellet and the pellet was air-dried for 5 - 10 minutes. DNA was resuspended in 250 µL Qiagen EB, Elution buffer (10 mM Tris-Cl, pH 8.5) and dissolved at 55 °C for 1 - 2 hours. The samples were transferred to a new tube and placed at 4 °C overnight, before quantification.

The Qubit BR dsDNA Assay Kit (Life Technologies) was used, as per the manufacturer's specifications, for gDNA quantification. To achieve a fragment size of 500 to 1,000 pb, a Covaris S2 was used to fragment 500 ng of the sample in 50 µL. Covaris parameters were set to an intensity of 5, duty cycle of 5% with a 200 cycles per burst at 7 °C, a water level of 12 and a treatment time of 30 seconds. Sample size distributions were analysed on a DNA HS Bioanalyzer Chip. A fragment size range of 550 bp – 750 bp was selected for using a 1.5% Blue Pippin agarose cassette with internal R2 marker. A DNA HS Bioanalyzer Chip was used to verify size selection and DNA quantification was performed with the Qubit HS dsDNA

Assay Kit (Life Technologies). Library preparation was performed using NEBNext Ultra II DNA Library Prep Kit, with dual indexing (Table 6).

Table 6. Dual index sequences used for whole-genome sequencing.

Name in LIMS	Index i7	Sequence	Index i5	Sequence
imb_baumann_2020_03_patterson_WGS_1_neoDE0015	D701	CGAGTAAT	D501	TATAGCCT
imb_baumann_2020_03_patterson_WGS_2_neoDE0037	D702	TCTCCGGA	D502	ATAGAGGC
imb_baumann_2020_03_patterson_WGS_3_neo23563	D703	AATGAGCG	D503	CCTATCCT
imb_baumann_2020_03_patterson_WGS_4_neo14584	D704	GGAATCTC	D504	GGCTCTGA
imb_baumann_2020_03_patterson_WGS_5_ino13737	D705	TTCTGAAT	D505	AGGCGAAG
imb_baumann_2020_03_patterson_WGS_6_ino12852	D706	ACGAATTC	D506	TAATCTTA

II.VIII. WHOLE-GENOME SEQUENCING

A total of six, dual indexed, libraries (four *A. neomexicanus* and two *A. arizonae* samples) were equimolarly pooled and 2 x 250 bp reads were sequenced on an Illumina HiSq2500 over two rapid-run flow cells.

FastQC version 0.11.9 [72] was used to assess the quality of sequencing files. Adapter and quality trimming was performed with Trimmomatic version 0.39, with the parameters described in V. Bowetie2 version 2.4.4 [79] was used to align trimmed files to the *A. neomexicanus* reference genome, with the ---very-sensitive parameters: -D 20 -R 3 -N 0 -L 5 -i S,1,0.50. The sequencing data from each flow cell was aligned separately and the alignments merged using Samtools version 1.9 [80]. Bowetie2 aligned uniquely mapping reads can be identified by a MAPQ of 60. These were extracted with Samtools view -h -bq 60.

II.IX. ALLELE-PAIR FILTERING PARAMETERS AND IDENTIFICATION OF ALLELE-SPECIFIC EXPRESSION

Python, version 3.10.9, was used to filter the *A. arizonae – A. marmoratus* allele-pairs in *A. neomexicanus*, which were identified using reciprocal best hit Blast, to allele-pairs of high confidence. This was done to minimise the identification of allelic biases which were due to

technical artefacts. When comparing the expression levels of allele-pairs, it is vital that the correct, and highly annotated, homologous *A. marmoratus* and *A. arizonae* genes are being compared. To exclude poorly annotated or incorrectly allele-pairs, several filtering parameters were imposed: the transcript percent identity between allele-pairs, the percent difference in protein length and the syntenic location of the allele-pairs. To exclude allelic biases which are due to genetic differences, in regions of loss of heterozygosity, the average whole genome sequencing depth was also considered. A transcript percent identity cut-off of >= 75% was selected as this ensures that the allele-pairs have a sequence similarity, so therefore the correct homologs are being compared. However, an upper cut-off of <= 98% was selected as this means the transcripts retain species specific differences which allows the allele to be identified as either the *A. arizonae* or *A. marmoratus* allele. A difference of 5 % or less in protein length between the allele-pair was established so that the difference in expression could not be attributed to the extra read counts being assigned to the longer allele in the allele-pair (Figure 10). Additionally, allele-pairs which were not located on syntenic chromosomes were filtered out. Synteny between the *A. arizonae* and *A. marmoratus* chromosomes was determined using Satsuma [81]. To exclude regions of LOH, the average WGS coverage was calculated for each allele using samtools depth -r version 1.9 and awk [80]. Allele-pairs in which at least one of the alleles was in a region of abnormally low or high coverage, were excluded from the analysis.

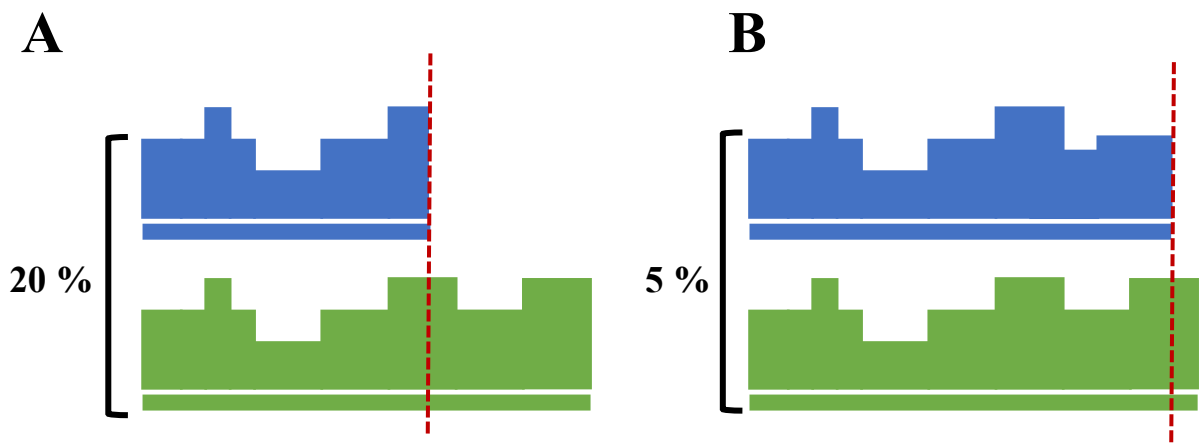


Figure 10. (A) Allele-pair with 20% difference in protein length, illustrating that the allelic bias which may be observed is due to this difference in length, and the extra reads mapping to the area on the left of the red line. (B) Allele-pair with 5 % difference in protein length. Here, the difference in read count is minimal between the two alleles. Therefore, no allelic bias would be identified.

The allelic bias, for each of the allele-pairs which made it past all four filtering parameters, was calculated using HTSeq counts [76]. The allelic bias was calculated as the percent of *A. arizonae* reads in the allele-pair. A cut-off of 70 % *A. arizonae* reads in an allele-pair was used to denote an *A. arizonae* bias. Similarly, a cut-off of 70 % *A. marmoratus* reads in an allele-pair was used to denote an *A. marmoratus* bias. Anything else was classed as not displaying an allelic bias. Allele-pairs which had a read count of less than 30 were filtered out. Heatmaps were clustered using the hierarchical clustering method UPGMA (unweighted pair group method with arithmetic mean). R version 4.3.1 was used to analyse enriched Biological Processes, Cellular Components and Molecular Function GO Terms, using the 'clusterProfiler' and 'topGO' packages. Homologous genes were identified in humans and tested against a human background of homologous genes to the filtered down allele-pair dataset. A Benjamini-Hochberg correction was used to adjust P-values to decreases the false discovery rate per cluster. A P value cut-off was set to 0.01 to determine significantly enriched GO Term clusters.

When determining if an allelic bias was present in the first generation of *A. neomexicanus* and maintained over time, the *in silico* reference was used as a proxy to compare their expression. If the expression pattern of the *in silico* sample was outside of 3 standard deviations of the mean of the expression of the *A. neomexicanus* samples, the bias pattern was determined to be novel (Figure 11.A). As the expression of the *in silico* sample is outside of 3 standard deviations of the *A. neomexicanus* samples, its expression pattern is distinguishable from the *A. neomexicanus* samples. However, if the expression of the *in silico* sample was within 3 standard deviations of the mean of the *A. neomexicanus* expression, the *in silico* sample would be indistinguishable from the *A. neomexicanus* sample (Figure 11.B). Therefore, the allelic bias present in *A. neomexicanus* would be classed as ancestral. The bias present would have been present in the gene expression of the parental species and maintained over time and likely to be regulated by sequence differences in the promoter or transcription factors of the gene.

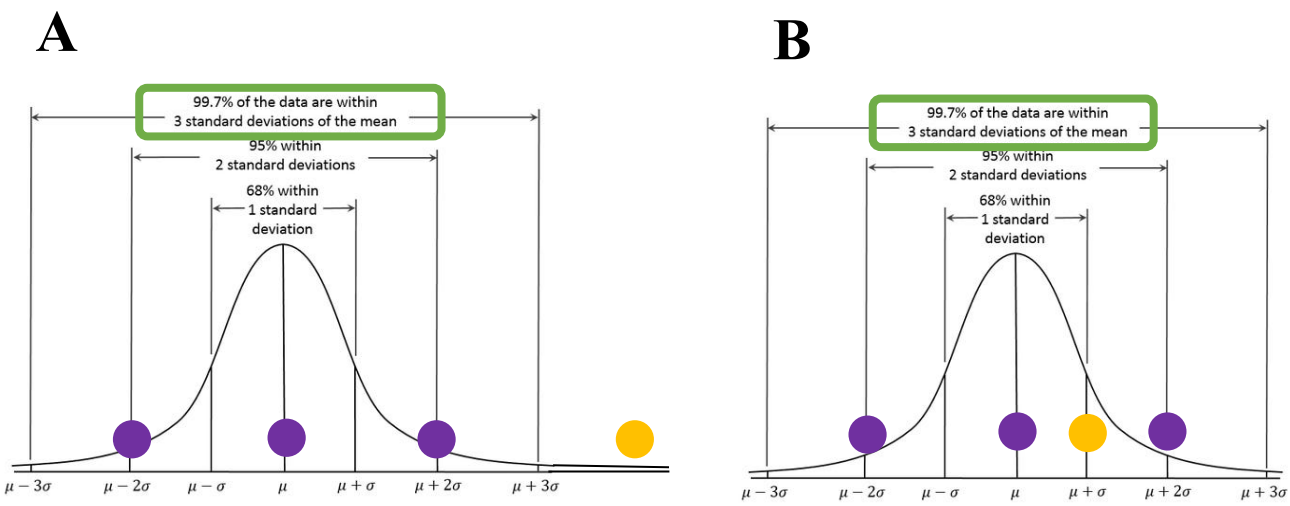


Figure 11. Assigning if an allelic bias is ancestral or novel in *A. neomexicanus*. (A) When the allelic bias of the *in silico* sample (yellow) is outside 3 standard deviations of the mean expression of the *A. neomexicanus* samples (purple), it is classed as a novel bias. (B) If the allelic bias of the *in silico* sample (yellow) is within 3 standard deviations of the mean of the expression of the *A. neomexicanus* samples (purple) it is classed as an ancestral bias. The *in silico* sample is indistinguishable from the *A. neomexicanus* samples.

III. Results

III.I. Genome annotations and allele-pair identification

A total of 29,351 protein-coding genes were annotated for *A. arizonae* and 29,572 protein-coding genes were annotated for *A. marmoratus* by Nathaniel Deimler. This is slightly higher than the number of protein-coding annotated for humans (19,116 [82]) and the anole (18,000 [83]). This is in part due to the expansion of certain gene classes in *Aspidoscelis* such as the vomeronasal genes. The distribution of these genome annotations across the *A. arizonae* and *A. marmoratus* genomes can be observed in Figure 12. Here, it can be noted that the smaller, micro-chromosomes are more gene dense. A higher gene density has also been observed in birds and reptiles such as in chickens [84]and in desert horned lizard [85]. These gene rich smaller micro-chromosomes are important for the formation and evolution of macro-chromosomes through fusions.[86].

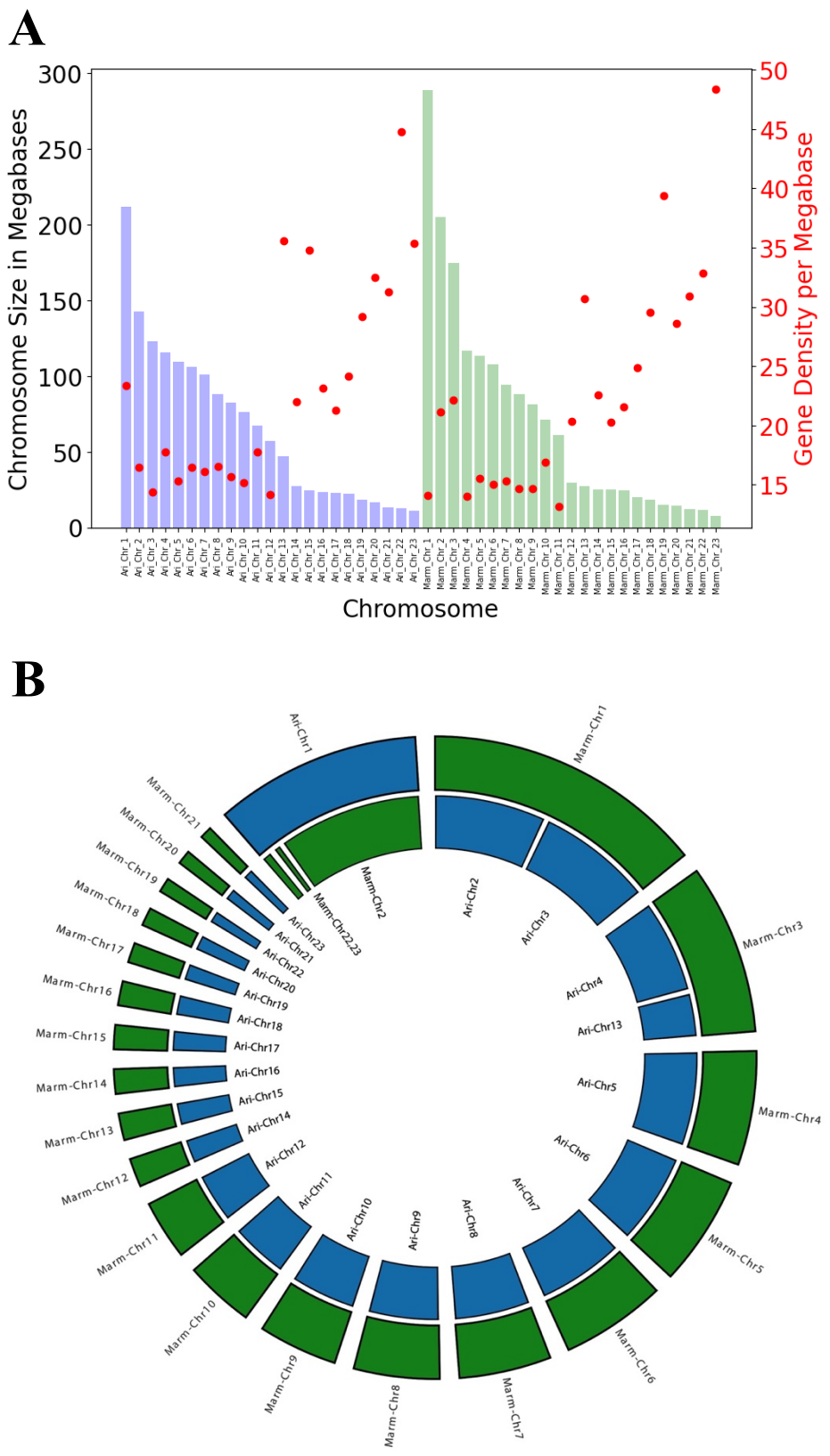


Figure 12. (A) Gene density per megabase per chromosome identified in *A. arizonae* and *A. marmoratus* respectively. (B) Circos synteny plot depicting the syntenic chromosomes between the parental species *A. arizonae* and *A. marmoratus*, of *A. neomexicanus*, obtained from A. Odell.

These distinct patterns of chromosomal fusions can also be observed in the *Aspidoscelis* lineage. The satsuma synteny results demonstrated that the micro-chromosomes *A. marmoratus* 22 and 23 fused to *A. marmoratus* chromosome 2 to be syntenic with the macro-chromosome *A. arizonae* 1. The fusion of two mid-sized chromosomes *A. arizonae* 2 and 3 is syntenic to macro-chromosome *A. marmoratus* 1. A further fusion of two mid-sized chromosomes *A. arizonae* 4 and 13 is syntenic to macro-chromosome *A. marmoratus* 3. These observations are consistent with the descriptions of Reeder et al. 2002 [87].

Reciprocal best hit Blast result yielded 17,871 allele-pairs. Their distribution along the *A. arizonae* and *A. marmoratus* chromosomes can be observed in Figure 13. Filtering criteria were imposed on these allele-pairs to subset allele-pairs of high confidence. The number of allele-pairs lost post filtering was evenly distributed across the chromosomes.

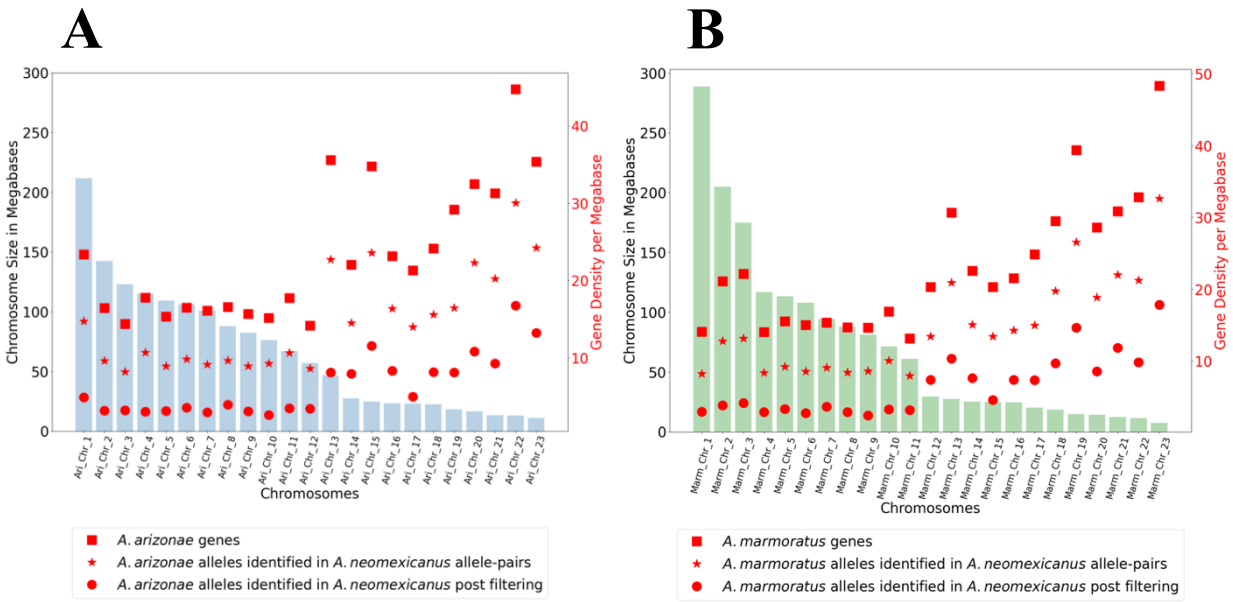


Figure 13. (A) The number of genes per megabase per chromosome identified in *A. arizonae*. Of the genes identified, the number of alleles identified in *A. neomexicanus* per megabase per chromosome were identified. Furthermore, the number of alleles per megabase per chromosome post filtering are also depicted. (B) The same can be observed for the *A. marmoratus* genes and alleles.

III.II. FILTERING CRITERIA FOR ALLELE-PAIRS OF HIGH CONFIDENCE

Of the 17,871 allele-pairs identified, 12,422 (69.51 % of allele-pairs) were 5% or less different in length (Figure 14.A). The transcript percent identity was between 75% - 98% identical for 10,975 allele-pairs (61.41 % of allele-pairs) (Figure 14.B). A high number of allele-pairs had a transcript percent identity of > 98% identical: 3,839 (21.48 % of allele-pairs), demonstrating the high quality of the genome annotations. Although these allele-pairs are of high quality and were identified independently in *A. arizonae* and *A. marmoratus*, they were excluded from further analysis as their allelic bias could not correctly determined. Allele-pairs with a transcript percent identity < 75% were also excluded as these were determined to be incompletely annotated. This consisted of 3,057 out of the 17,871 allele-pairs (17.11 % of allele-pairs). Two individuals with known different patterns of LOH were used as a reference for average DNA-Seq coverage per allele. To avoid investigating allele-specific expression in regions of LOH an average cut-off was selected of 5.5x and 12x per allele (Figure 14.C). The number of allele-pairs which made it through the DNA-Seq filtering for both reference lizards was 15,439 (86.39 % of allele-pairs). As protein-coding genes are conserved in syntenic clusters [88] a last filtering criterion of synteny was imposed. A total of 15,977 (89.40 %) allele-pairs were located on syntenic chromosomes (Figure 14.D). Overall, 6,636 (37.13 %) allele-pairs passed all four filtering criteria and were determined to be of high confidence in which allele-specific expression in *A. neomexicanus* was investigated (Figure 14.E). A large proportion of allele-pairs were lost at either the transcript percent identity or DNA-Seq coverage, demonstrating the high quality of the genome annotations.

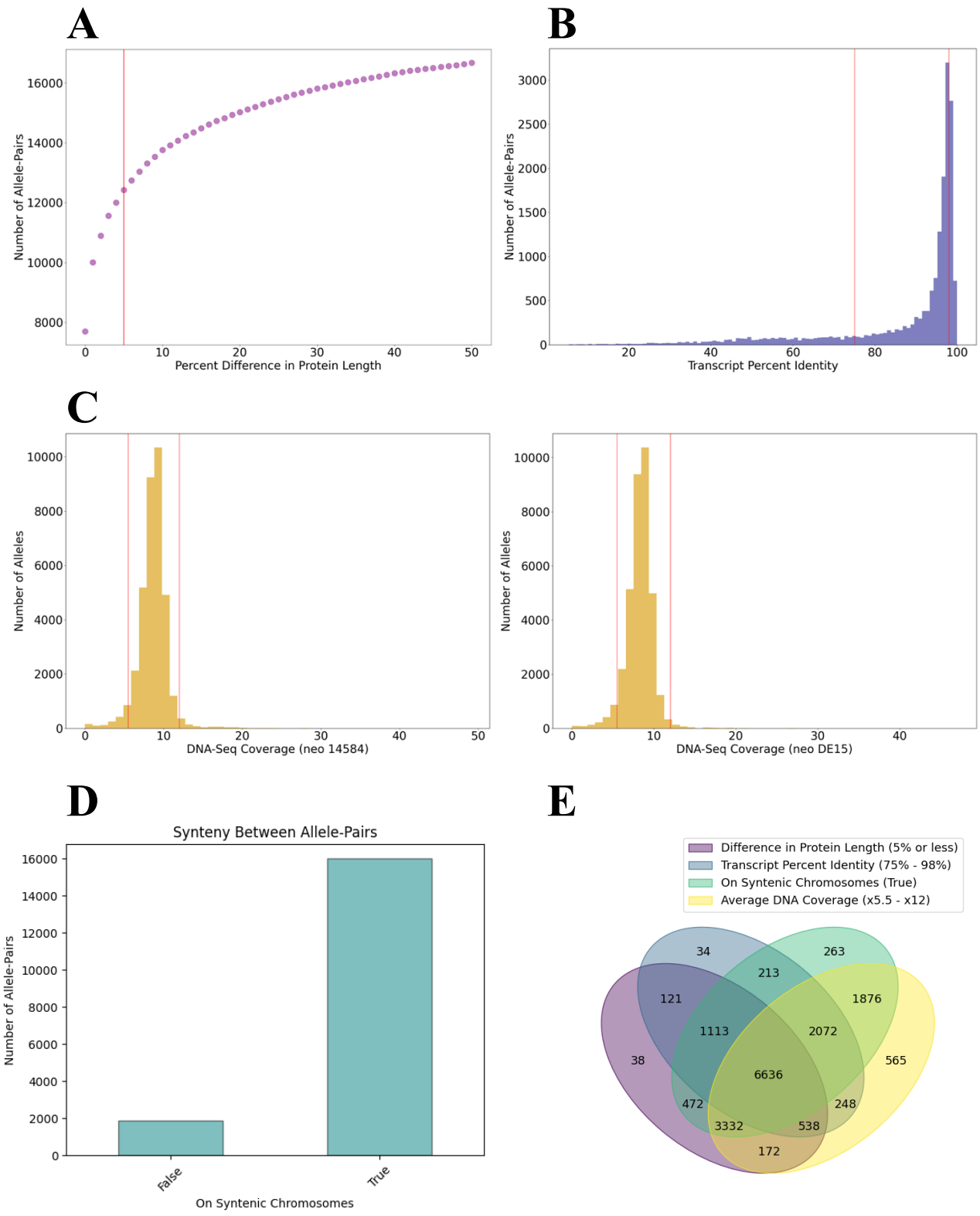


Figure 14. The filtering parameters implemented on the allele-pairs to obtain allele pairs of high confidence. (A) The percent difference in protein length of 5 % was chosen, before the graph reaches an asymptote. (B) Allele-pairs with a percent identity of >= 75 % and <= 98% were selected as these had a high sequence similarity but retain species-specific differences. (C) Alleles with an average DNA-Seq coverages of 5.5 x – 12 x were selected to exclude regions of LOH. Two individuals with known different patterns of LOH were chosen as reference. (D) Allele-pairs which were not on syntenic chromosomes were filtered out from further analysis. A Venn diagram summarizes the allele-pairs which passed each filtering criteria. (E) Of the 17,871 identified, 6,636 were determined to be of high confidence.

III.III. RNA-SEQ ALIGNMENT STRATEGY

To determine the alignment accuracy when aligning reads of a species of hybrid origin to a reference genome of two closely related species*, in silico* sample were used. In doing so, the placement of *A. arizonae* and *A. marmoratus* reads to the correct genome region can be determined. Reads of the *in silico* blood sample were trimmed to 70 bp, 80 bp, 90 bp and 100 bp to determine how this would impact the total number of uniquely mapping reads, multi-mapping reads and unmapped reads. Only uniquely mapping reads were used in the downstream analysis as the multi-mapping reads had an extremely high mismapping frequency; having an almost 50 % percent chance being wrongly placed (ex. an *A. arizonae* read mapping to an *A. marmoratus* chromosome). The regions in which multi-mapping occurred had a high sequence similarity so there were little to no sequence specific differences to anchor the read on to, despite increasing the mismatch penalty of the aligners. In fact, increasing the mismatch penalty caused a drop in the number of uniquely mapping reads and an increase in unmapped reads. Therefore, default mismatch penalty settings were used for the aligners.

For both the STAR and HISAT2 aligners, as the read length increased, so did the percentage of uniquely mapping reads (Figure 15.A). Conversely, as the read length increases, the mismapping frequency decreased (Figure 15.B). However, HISAT2 outperformed STAR in both the number of uniquely mapping reads recovered and the accuracy in which the uniquely mapping reads were placed. Therefore, HISAT2 was used to align the *A. neomexicanus* samples. STAR had a higher variability in mismapping frequency between tissues (Figure 15.C), versus HISAT2 which showed a more consistent mismapping frequency between tissues (Figure 15.D).

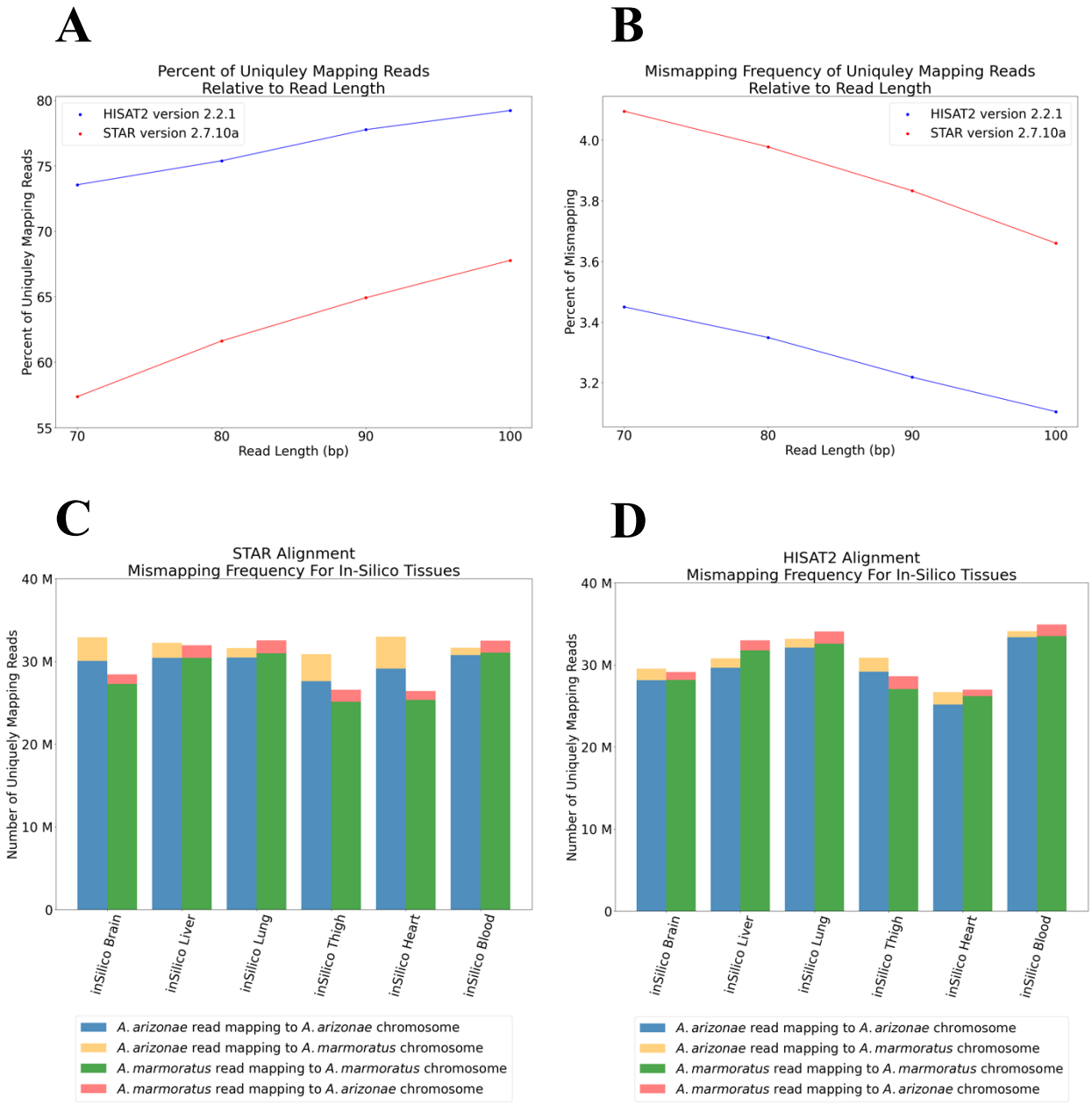


Figure 15. Comparison of STAR and HISAT2 aligners when aligning to the *A. neoomexicansu* genome. (A) The longer the read length, the higher the number of uniquely mapping reads. (B) Of the uniquely mapping reads, the mismapping frequency decreases as the read length increases. The mismapping frequency per *in silico* tissue is higher when aligning with STAR (C) versus HISAT2 (D). Overall, HISAT2 was shown to be the more accurate aligner when aligning *A. neomexicanus* data.

As HISAT2 was determined to be the more precise aligner, it was used to align the *A. neomexicanus* samples. An average of 75 % uniquely mapping reads was obtained across the samples. This is in line with the uniquely mapping alignment rate obtained for the *in silico* tissue samples. The lowest percent of uniquely mapping reads was captured for the heart samples (with an average of 68 % across the five samples) and the highest was obtained for the ovary samples, with an average of 84 % uniquely mapping reads (Figure 16).

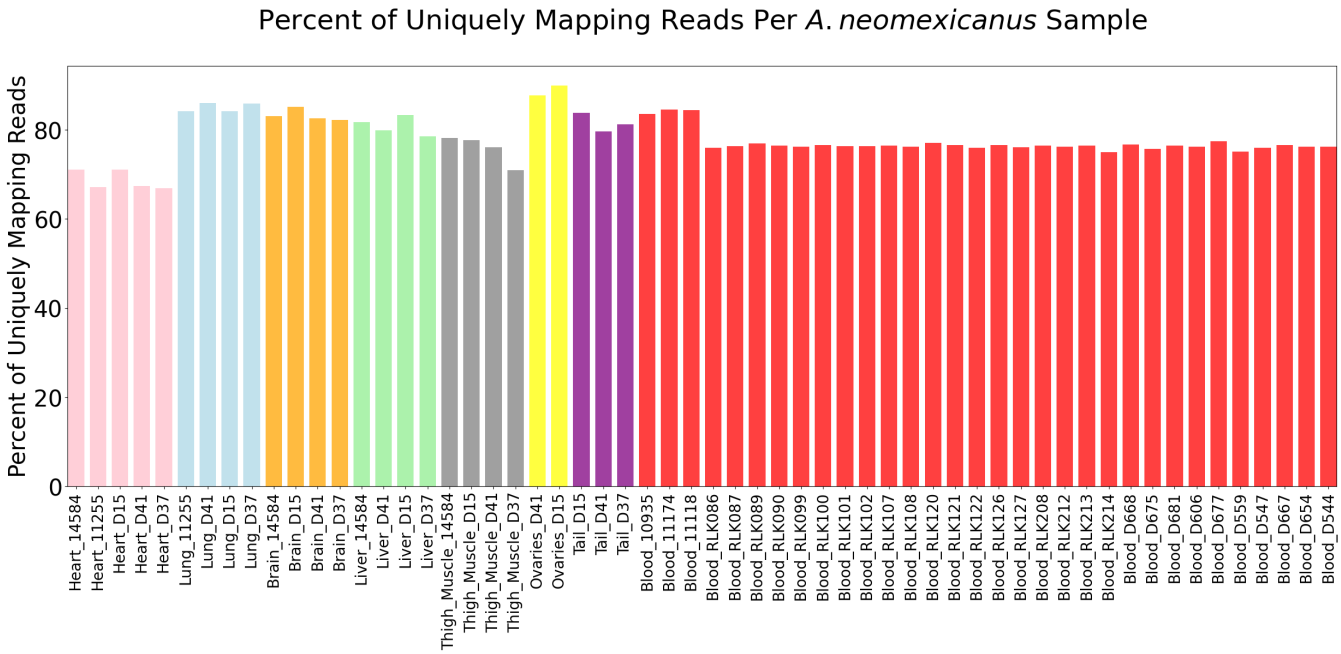


Figure 16. The percent of uniquely mapping reads per *A. neomexicanus* sample. Each tissue is separated by a unique colour. In total x5 heart samples, x4 lung, brain, liver, thigh muscle, x2 ovaries, x3 tail and x32 blood RNA-Seq samples were aligned. The percent of uniquely mapping reads varied from the 68 % in the heart samples to 84 % in the ovary samples.

FastQ Screen results confirmed minimal contamination of *Homo sapiens*, *Mus musculus*, *Schizosaccharomyces pombe* and *Saccharomyces cerevisiae* in the *A. neomexicanus* sequencing data (Figure 17).

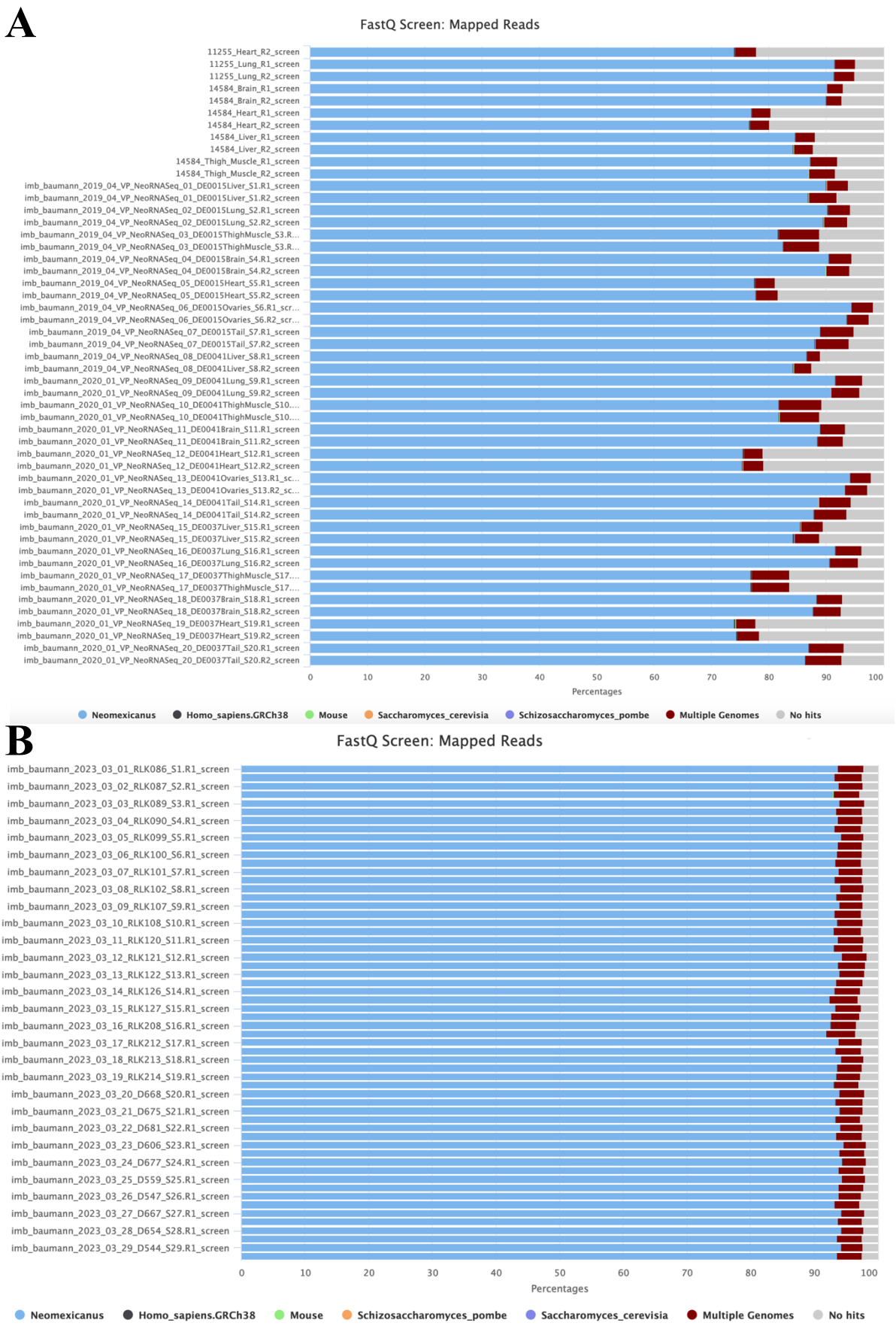


Figure 17. FastQ Screen results for tissue-specific *A. neomexicanus* RNA-Seq (A) and blood samples from wild and lab *A. neomexicanus* individuals (B) confirm minimal contamination from commonly sequenced species.

III.IV. *IN SILICO* ALLELIC BIAS

The allele-specific expression was investigated in *in silico A. neomexicanus* tissues to gain an understanding in the allele-specific expression expected in the *A. neomexicanus* lineage. Moreover, creating the *in silico* tissue references would allow us to understand the directional shift in allele-specific expression in the *A. neomexicanus* lineage. The original hybridization between *A. arizonae* and *A. marmoratus* which gave rise to *A. neomexicanus* was estimated to have occurred 200,000 years ago. Using the *in silico* samples as a reference for the original allele-specific expression, the evolution of the allele-specific expression in *A. neomexicanus* can be traced.

A. arizonae (>= 70 % *A. arizonae* HTSeq counts in an allele-pair) allelic biases and *A. marmoratus* (>= 70 % *A. marmoratus* HTSeq counts in an allele-pair) can be observed, genome wide in *the in silico* blood sample (Figure 18.A). Lowly to not expressed allele-pairs were filtered out, therefore each allele-pair consists of 30 or more HTSeq counts. The percentage of allelic biases observed in each *in silico* tissue can be observed in Figure 18.B. A high number of *A. arizonae* allelic biases can be observed in the *in silico* blood, whereas a high number of *A. marmoratus* allelic biases can be observed in the *in silico* heart. Although the *in silico* brain had the second highest number of highly expressed allele-pairs (4,083 >= 30 HTSeq counts per allele-pair), it displayed a very low number of allelic biases (Table 7). This tissue-specific allele-specific expression pattern is consistent with the observations of GTEx in human tissues [37]. The variation in allele-specific expression between tissues can be observed in Figure 18.C. In the 1,503 allele-pairs which showed an allelic bias, the large standard deviation bars demonstrate that in the allelic bias varies largely between tissues.

Table 7. Summary of the amount of allele-specific expression identified across *in silico* tissue samples.

	Blood	Heart	Brain	Liver	Lung	Thigh
A. arizonae bias	596	163	124	368	446	312
No bias	2,686	3,013	3,690	2,924	3,661	2,727
A. marmoratus bias	238	449	269	403	341	134
Total allelic biases	834	612	393	771	787	446
Total past RNA-Seq filtering	3,520	3,625	4,083	3,695	4,448	3,173

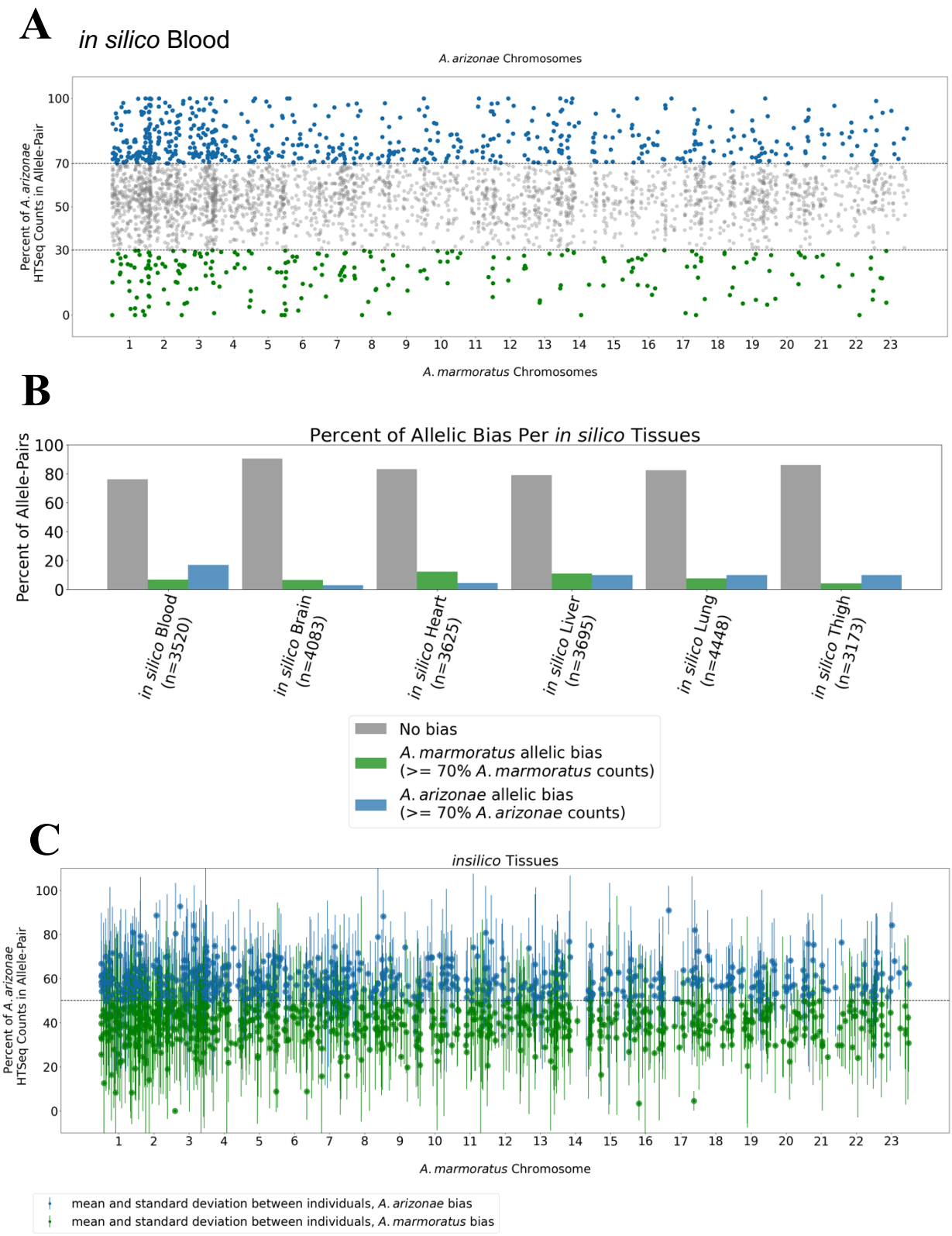


Figure 18. (A) Allelic biases across one sample*: in silico* blood. Each point represents an allele-pair with its bias represented on the x-axis. Blue points equal to or higher than 70 % show an *A. arizonae* allelic bias. Green points equal to or less than 30 % show an *A. marmoratus* allelic bias. Gray points indicate allele-pairs which do not show a bias. Each point consists of at least 30 HTSeq counts. (B) The percent of no bias, *A. arizonae* bias and *A. marmoratus* bias per *in silico* tissue, which had at least 30 HTSeq counts per allele-pair. (C) A subset of 1,503 allele-pairs which had at least 30 HTSeq counts and showed a bias in at least one sample: 710 allele-pairs had an overall *A. arizonae* bias and

793 allele-pairs had an overall *A. marmoratus* bias. Each point represents the average expression of the allele-pair across the *in silico* tissues. The bar represents the standard deviation between the samples.

Distinct tissue-specific allele-specific expression patterns can be observed in the *in silico* samples (Figure 19). A principal component analysis (PCA) of the tissues demonstrates that the blood sample has a high separation from the brain, heart, thigh, liver and lung sample along PC1 (40.1% Figure 19.A). The liver, which regulates most chemical levels in the blood and where metabolism and detoxification occurs also shows a distinct allele-specific expression pattern and high separation from the heart, brain, lung, and thigh along PC2 (21.3%). A heatmap highlights how the allelic bias of an allele-pair varies across tissues (Figure 19.B). Overall, when combining *A. arizonae* and *A. marmoratus* RNA-Seq data to create *in silico* tissue references, allelic bias has been shown to be ubiquitous and have tissue-specific expression patterns.

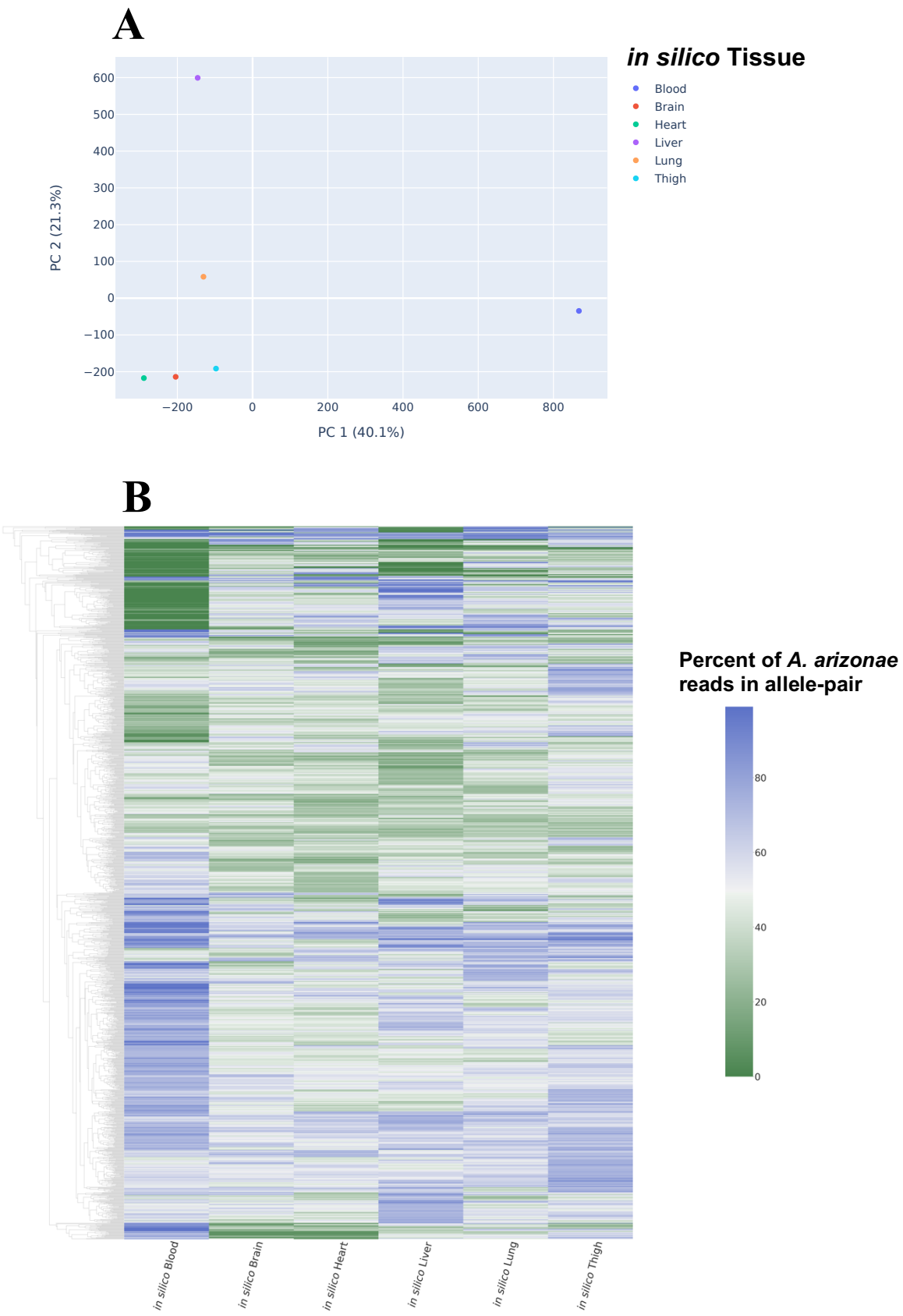


Figure 19. (A) PCA of *in silico* tissues. The blood sample clusters away from the remaining tissues, showing a high separation along PC1, demonstrating its unique allelic bias pattern. The lung and liver show a high separation from the heart, brain, and lung along PC2. This demonstrates the high tissue-specific allele-specific expression in the *in silico* tissues. (B) A heatmap with UPGMA hierarchical clustering, of the allelic biases of the tissues further demonstrates the distinct allelic biases across tissues.

Of the 6,636 high confidence allele-pairs, 4,997 homologous genes were identified in humans. This was the background used for searching for enriched gene clusters. The 1,503 allele-pairs, which showed a bias in an *in silico* tissue, were tested for enriched Biological Processes, cellular components, and molecular functions. Human homologous proteins were used for this enrichment, and 1,385 homologs were identified. The 15 enriched Biological Processes, GO Terms can be observed in Figure 20.A. Five clusters of enriched Biological Process can be observed (Figure 20.B). These are all involved in catabolic or metabolic processes: the breakdown of molecules into smaller units that are either oxidized to release energy or used in other anabolic reactions. These processes are likely to occur in muscle tissue, followed by the release of molecules that circulate in the blood and are detoxified in the liver. The interaction of the enriched Biological Process GO Terms can be observed in Figure 21. The interactions end in a fatty-acid metabolic process: fatty-acid oxidation. This is the process to completely break down fatty acids and in which energy is produced.

The enriched Cellular Component terms 'external encapsulating structure', 'extracellular matrix' and 'collagen containing extracellular matrix' are all important for cell structure (Figure 20.C). The proteins, especially collagen, and glycosaminoglycans present in the extracellular matrix are essential for tissue morphogenesis, differentiation, and homeostasis. These enriched Cellular Component terms have a close interaction: 'external encapsulating structure' interacts with the 'extracellular matrix' which in turn interacts with the 'collagen containing extracellular matrix' (Figure 22).

No Molecular Function enriched GO Terms were identified for the allele-pairs which showed a bias. However, the 30 most common terms can be observed in Figure 20.D. These terms are largely associated with catalytic activity compound cycling. Proteins associated with oxidoreductase activity were also present. Oxidoreductases are involved in xenobiotic metabolism and play a key role in the chemoprotection. Many xenobiotic metabolism enzymes have been shown to be involved in tissue protection in response to chemical stimuli [89]. Additionally, protein-containing complex binding terms were present suggesting that proteins of *A. arizonae* and *A. marmoratus* bias could be interacting or forming complexes in *A. neomexicanus*.

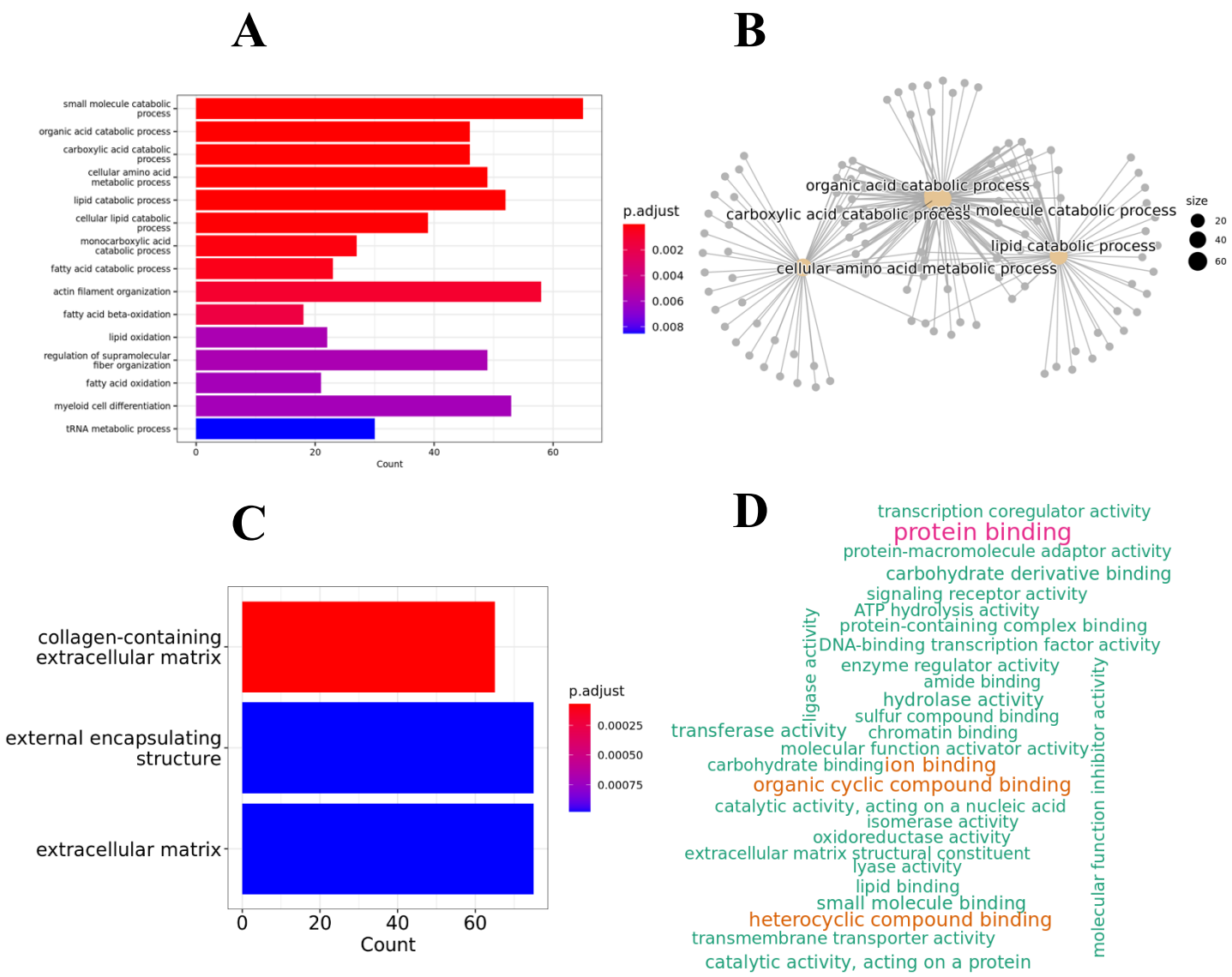


Figure 20. (A) The top 15 enriched Biological Processes GO Terms for the allele-pairs which show a bias in the *in silico* tissues. (B) Five main enriched Biological Processes can be observed for the i*n silico* tissues. (C) The enriched Cellular Components for the allele-pairs which showed a bias in at least one *in silico* tissue. (D) The 30 most common Molecular Function GO Terms were identified for the allele-pairs which showed a bias in the *in-silico* tissues, no Molecular Function terms were statistically enriched.

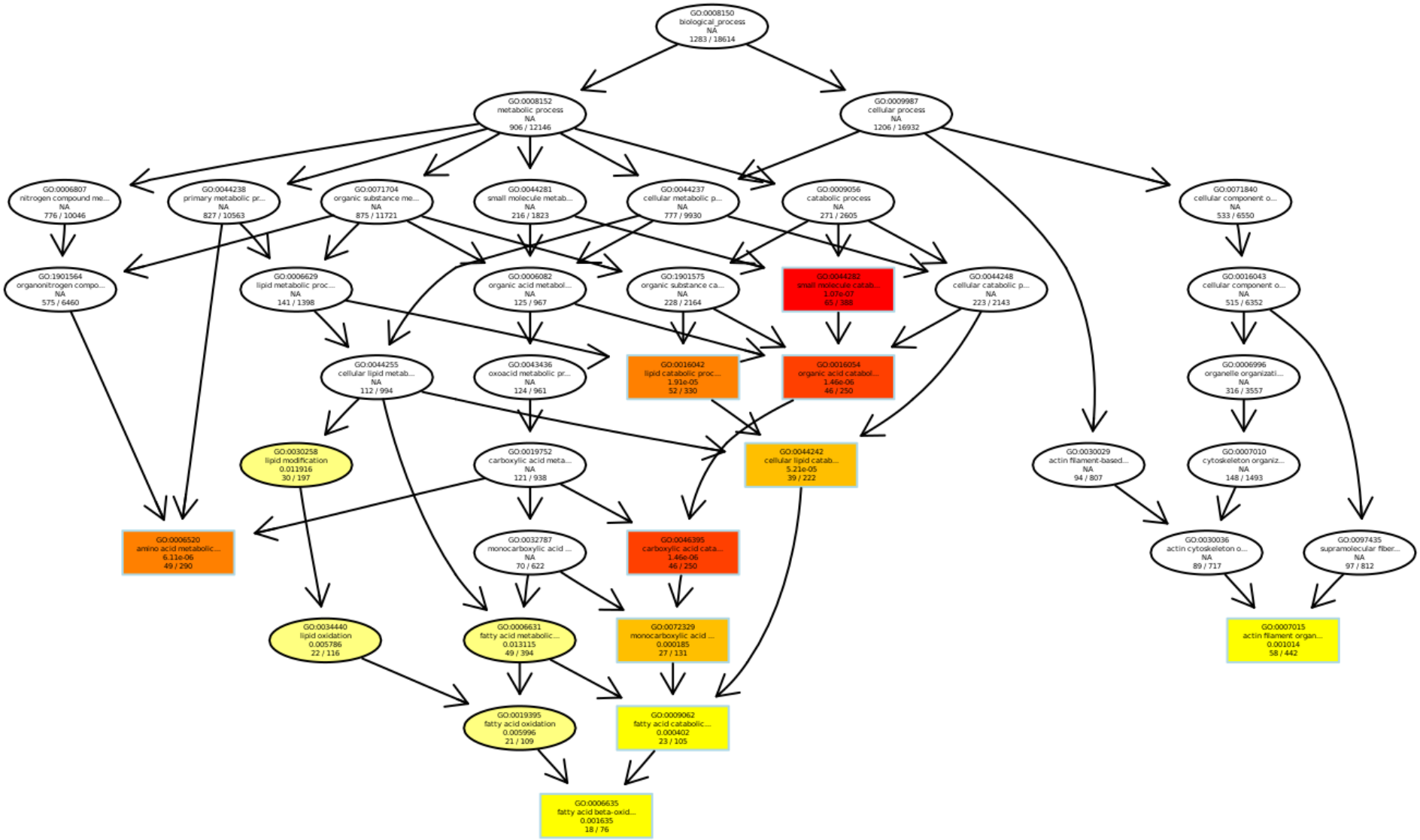


Figure 21. GO Term interaction plot. The enriched Biological Processes GO Terms for the allele-pairs which showed a bias in an *in silico* tissue. A darker the colour for the GO Term highlighted indicates a lower P-value. The interaction between the GO Terms terminates in “fatty acid beta oxidation” which is the process of breaking down of fatty acids which produces energy.

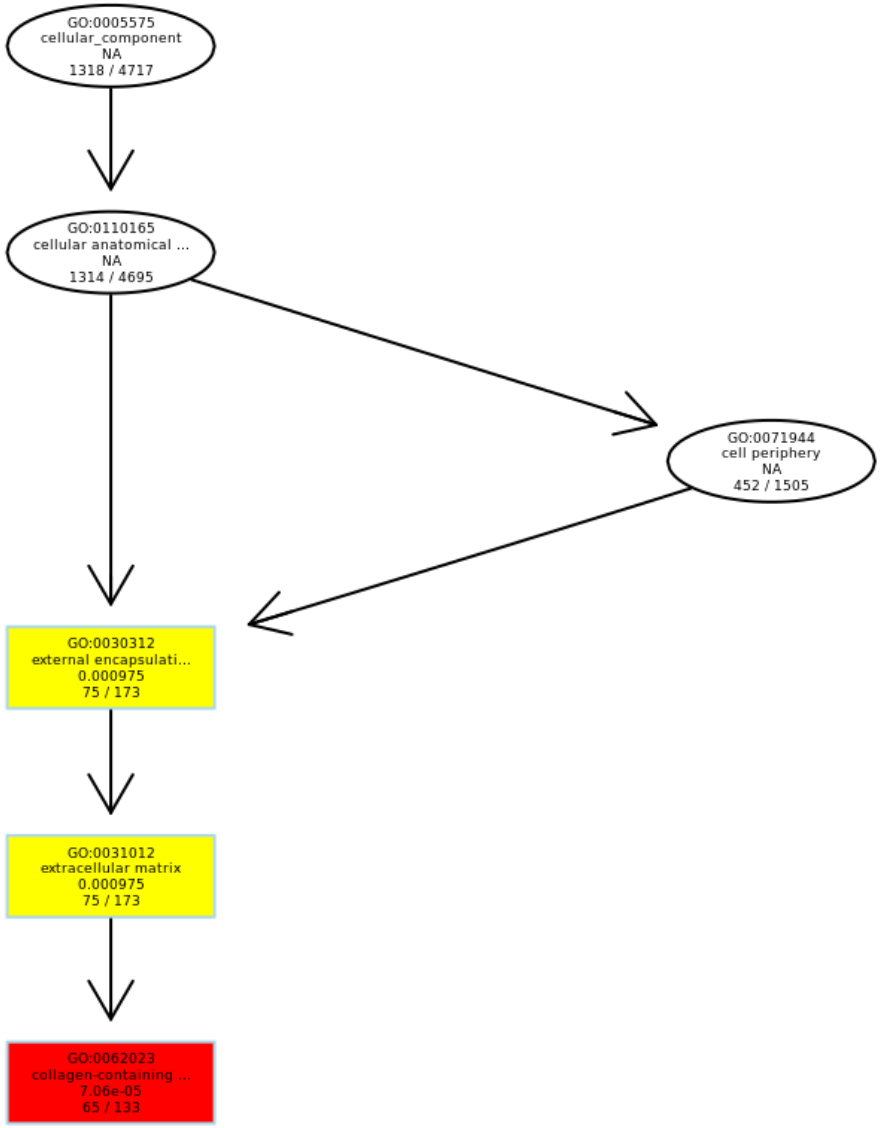


Figure 22. GO Term interaction plot. The enriched Cellular Component GO Terms for the allele-pairs which showed a bias in an *in silico* tissue. A darker the colour for the GO Term highlighted indicates a higher P-value. The interaction terminates in "collagen-containing extracellular matrix". This is important for tissue morphogenesis, differentiation, and homeostasis.

III.V. TISSUE-SPECIFIC ALLELIC BIAS

The allele-specific expression was investigated in the ovaries of two lizards, the blood and tail of three lizards, in the brain, thigh, liver and lung of four lizards and the heart of five lizards (Figure 23.A). The number of allele-pairs which were highly expressed (>= 30 HTSeq counts per allele-pair) varied between samples, with brain and blood samples expressing the highest quantity of allele-pairs. Once the allele-pairs which showed a bias in at least one sample were extracted, a total of 2,804 allele-pairs remained (Figure 23.B). Of these, the highest amounts of allele-specific expression can be observed in the ovaries, lung, and liver (Table 8). On average, approximately 30 % of allele-pairs which passed the RNA-Seq filtering displayed a bias across tissues.

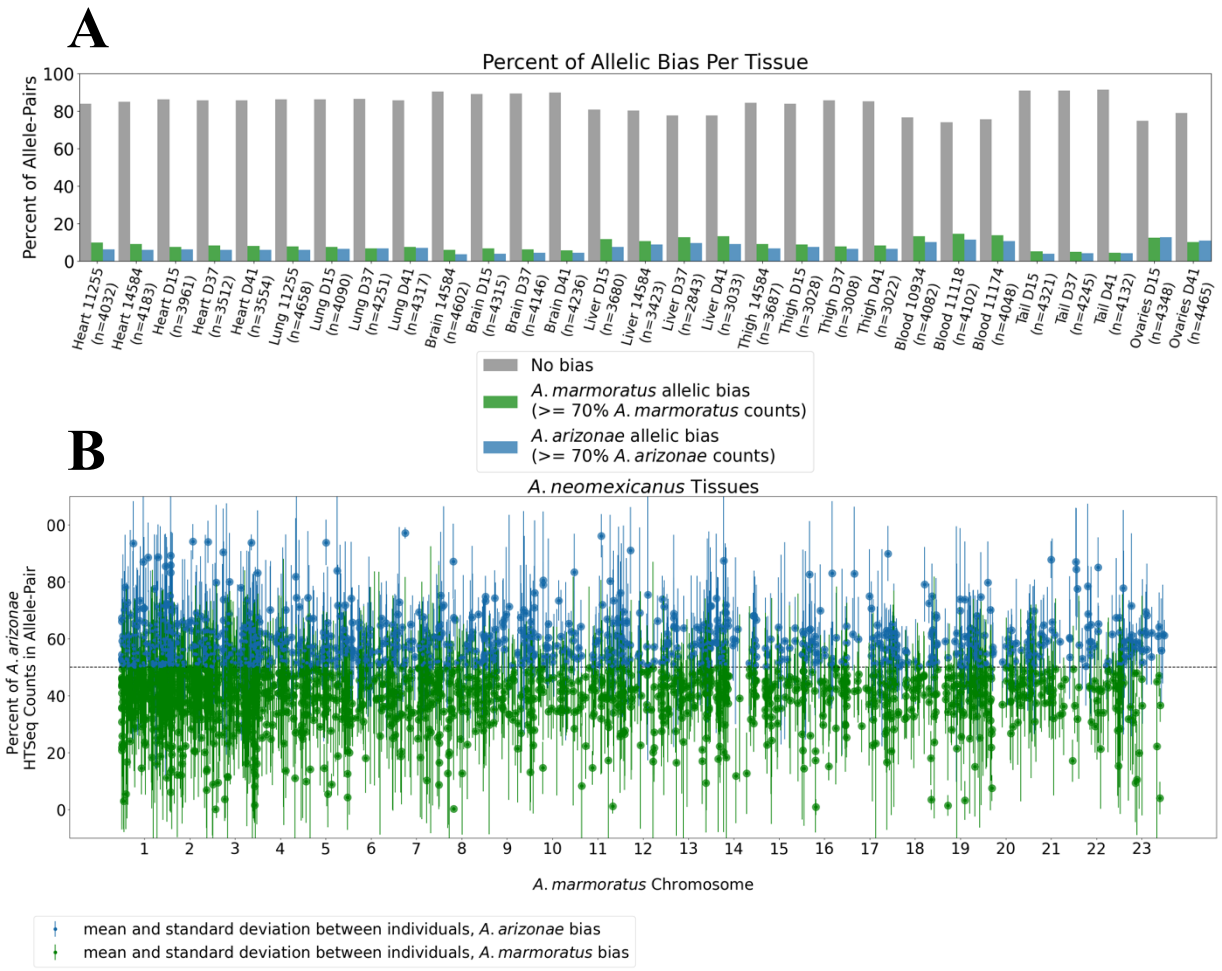


Figure 23. (A) The number and percent of allele-pairs which passed the filtering criteria of consisting of at least 30 HTSeq counts per allele-pair in each *A. neomexicanus* tissue. (B) A subset of 2,675 allele-pairs which had at least 30 HTSeq counts and showed a bias in at least one sample: 1,059 allele-pairs had an overall *A. arizonae* bias and 1,616 allele-pairs had an overall *A. marmoratus* bias. Each point represents the average expression of the allele-pair across the *in silico* tissues. The bar represents the standard deviation between the samples.

Table 8. Summary of the amount of allele-specific expression identified per tissue, across the individuals per tissue.

	Heart (n=5)	Brain (n=4)	Liver (n=4)	Lung (n=4)	Thigh (n=4)	Blood (n=3)	Ovaries (n=2)	Tail (n=3)
A. arizonae bias	360	223	508	480	334	348	568	270
No bias	2,653	3,403	2,219	3,266	2,267	2,657	3,038	3,533
A. marmoratus bias	457	390	796	636	463	480	568	322
Total allelic biases	907	613	1,304	1,116	797	828	1,136	592
Total past RNA-Seq filtering	3,560	4,016	3,523	4,382	3,064	3,485	4,174	4,125
Percent biased	25.5%	15.3 %	37.0 %	25.5%	26.0%	23.8 %	27.2 %	14.4%

The similarity in allelic bias expression between tissues for the 2,804 allele-pairs which showed a bias in at least one sample can be observed in Figure 24. The tail, brain, heart, lung and thigh show the most similar allelic bias pattern to each other whereas the blood, ovaries and liver show a more distinct pattern.

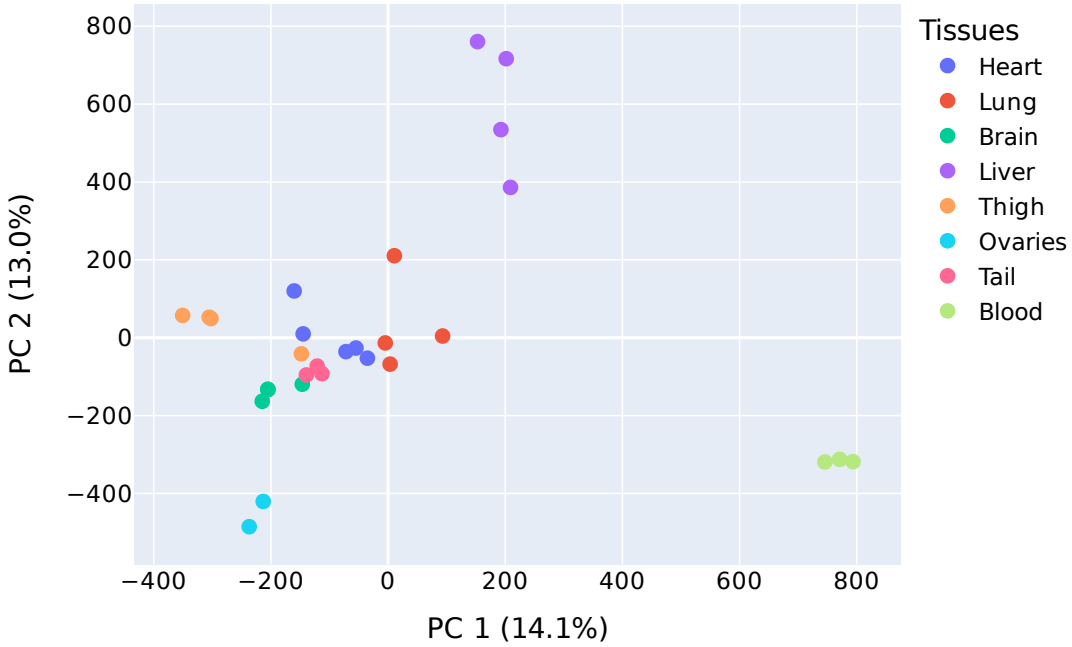


Figure 24. PCA of the 2,804 allele-pairs which demonstrated an allelic bias in at least one tissue sample. PC1 accounts for 14.1% of variation between samples whilst PC2 accounts for 13.0% of the variation.

The change in allele-specific expression between tissues can not only be observed by the large standard deviation bars in Figure 23, but in Figure 25 too. In this actin regulatory protein CAP-G, an *A. marmoratus* allelic bias can be observed in the heart tissues, however an *A. arizonae* allelic bias can be observed in the ovaries. No bias can be observed in the brain tissue. The uniform DNA-Seq coverage at both alleles indicates that this allelic bias is not due to a region of LOH. When incorporating the expression in the parental species, using blood data, an *A. arizonae* allelic bias can be observed. Additionally, a variation in *A. marmoratus* allelic bias can between individuals can be observed in the heart samples, between the five *A. neomexicanus* individuals. This represents one example of where a switch in allelic bias can be observed however 1,025 (37%) instances can be identified of an *A. arizonae* bias is present in one tissue however displays an *A. marmoratus* bias in a different tissue. Only 16 allele-pairs displayed an *A. marmoratus* bias across all tissues and individuals whilst 6 allele-pairs displayed an *A. arizonae* allelic bias across tissues and individuals.

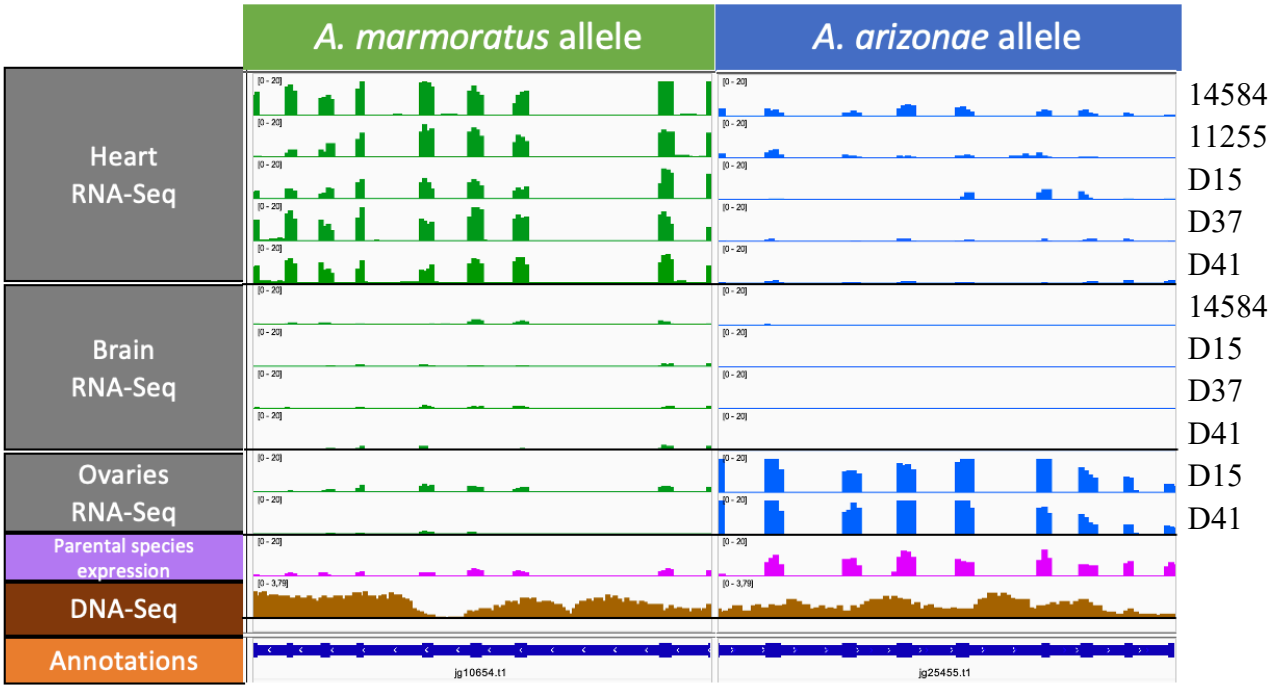


Figure 25. Tissue-specific allele-specific expression pattern between the heart, brain and ovaries of *A. neomexicanus*. An *A. marmoratus* allelic bias can be observed in the heart data, no bias in the brain and an *A. arizonae* allelic bias in the ovaries. The ID of the individual can be observed on the right of each row. No LOH can be observed when looking at the DNA-Seq coverage at this allele.

To compare the difference in allele-specific expression present in *A. neomexicanus* versus the in the parental species, the tissue-specific allele-specific expression was investigated in *A. neomexicanus* for each tissue in which an *in silico* reference was available, namely the blood,

heart, thigh, liver, lung and brain tissue. A PCA of the 2,675 allele-pairs which displayed a bias in at least one tissue demonstrated a high degree of separation of the blood sample from the liver, thigh, brain, lung, and heart along PC1 (17 %, Figure 26). Furthermore, the livers also show a high degree of separation along PC2 (14.9 %). The muscle tissues; lung, thigh and heart also cluster together, alongside the brain samples. The allelic bias expression pattern is consistent between individuals in the brain samples and the blood samples respectively. More variability in expression is observed in the liver, heart, thigh, and lung samples. It is important to note, that a perfusion to remove blood from the lung and heart samples was not performed. Therefore, some blood transcripts may be present in the lung and heart samples which could be contributing to the variability in allele-specific expression. Similarly to the *in silico* tissue references, the brain samples showed the lowest levels of allele-specific expression.

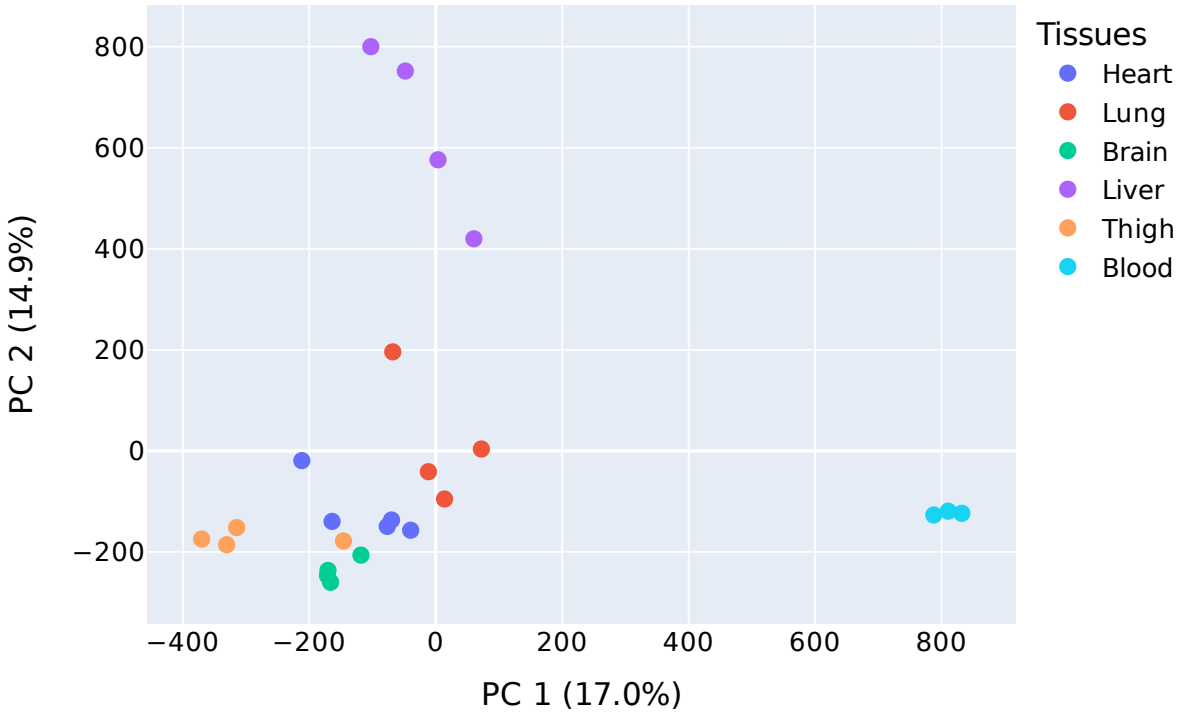


Figure 26. PCA of the 2,675 allele-pairs which showed a bias in at least one tissue sample. The blood samples show a high degree of separation along PC1 to the other tissues. The liver clusters away from the other tissues, along PC2, demonstrating the high specificity in allele-specific expression across *A. neomexicanus* tissues.

The allelic bias of each allele-pair can be observed in Figure 27. The tissue samples are hieratically clustered and cluster together. A distinct allelic bias pattern can be observed in the blood and liver.

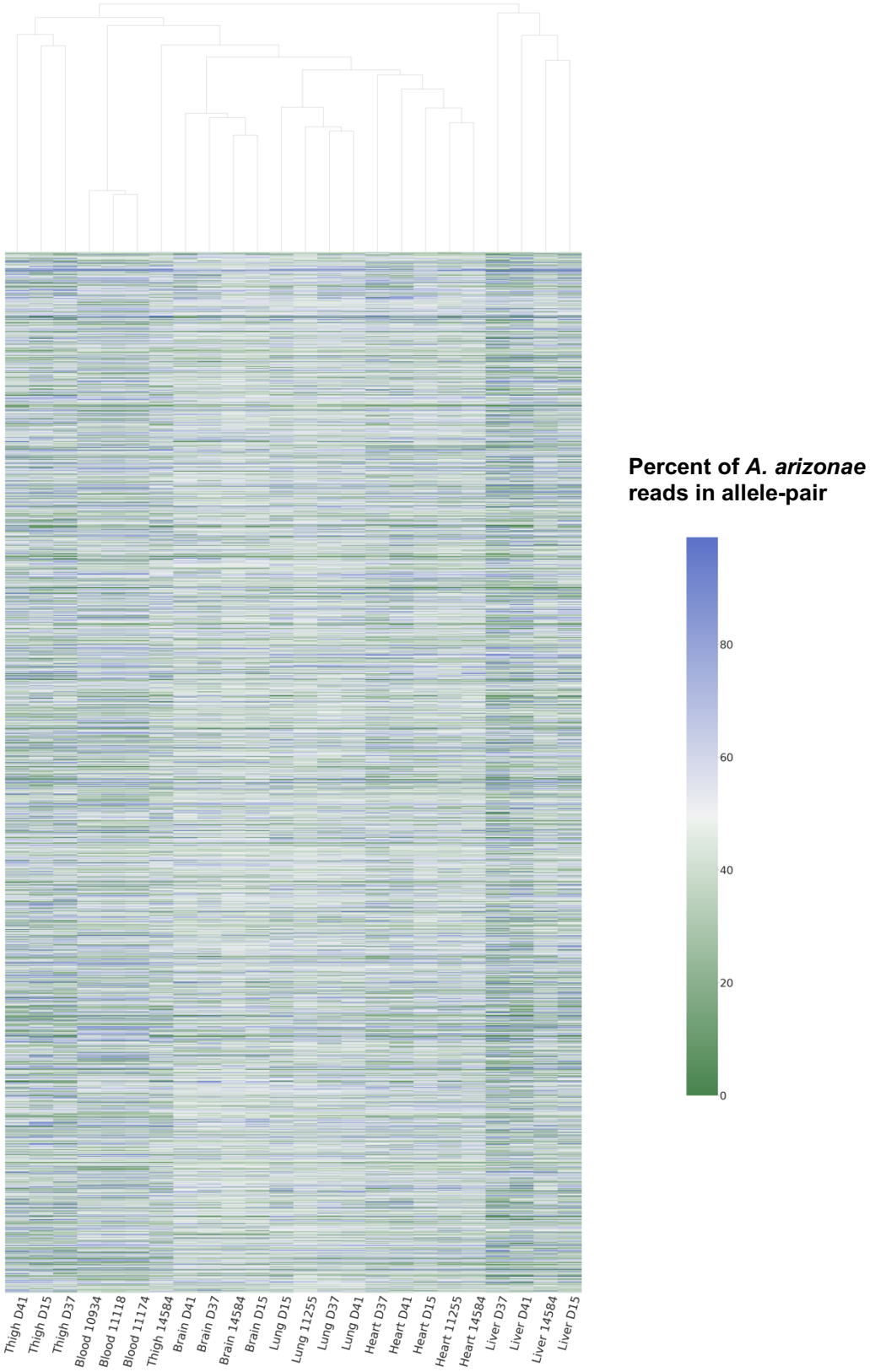


Figure 27. Heatmap of the percent of *A. arizonae* reads in an allele-pair of the 2,675 allele-pairs which showed a bias in at least one sample. The samples are hierarchically clustered using UPGMA (unweighted pair group method with arithmetic mean). Each tissue is clustered in aggregate.

Homologous proteins were identified in humans for 2,307 of the 2,675 allele-pairs which showed a bias in at least one tissue for GO Term enrichment. Of the 62 enriched Biological Process identified in the allele-pairs which showed a bias in the *A. neomexicanus* tissue, the top 15 can be observed in Figure 28.A.

Enriched Biological Process functions include terms associated with DNA breaks and repairs 'double-strand break repair' and 'interstrand cross-link repair' [90]. The GO term 'small molecule catabolic process' was present in the enriched Biological Process GO terms in *A. neomexicanus* tissues. This term was also observed in the *in silico* enriched Biological Processes terms highlighting how allelic bias of certain genes are being passed on and maintained from the parental species to *A. neomexicanus*.

Strikingly, the cellular components identified in the allele-pairs which showed a bias in the tissues are involved in cytoskeleton organization (Figure 28.B). Additionally, 'spindles', and 'microtubule' associated terms can be observed as enriched Cellular Components. These proteins form between opposite poles of a eukaryotic cell during mitosis or meiosis and serve to move the duplicated chromosomes apart [91]. The presence of these enriched Cellular Component terms demonstrates how allelic bias in these proteins may be important in mitosis or meiosis in *A. neomexicanus.*

Intriguingly, 'ATP binding', 'adenyl nucleotide binding' and 'purine ribonucleoside triphosphate binding' Molecular Function terms were highly present in allele-pairs which displayed an allelic bias in tissues (Figure 28.C). Over 200 allele-pairs were associated with these terms. Proteins associated with these enriched Molecular Function terms play a role in phosphate cycling and are vital for the cell's source for energy and phosphate [86].

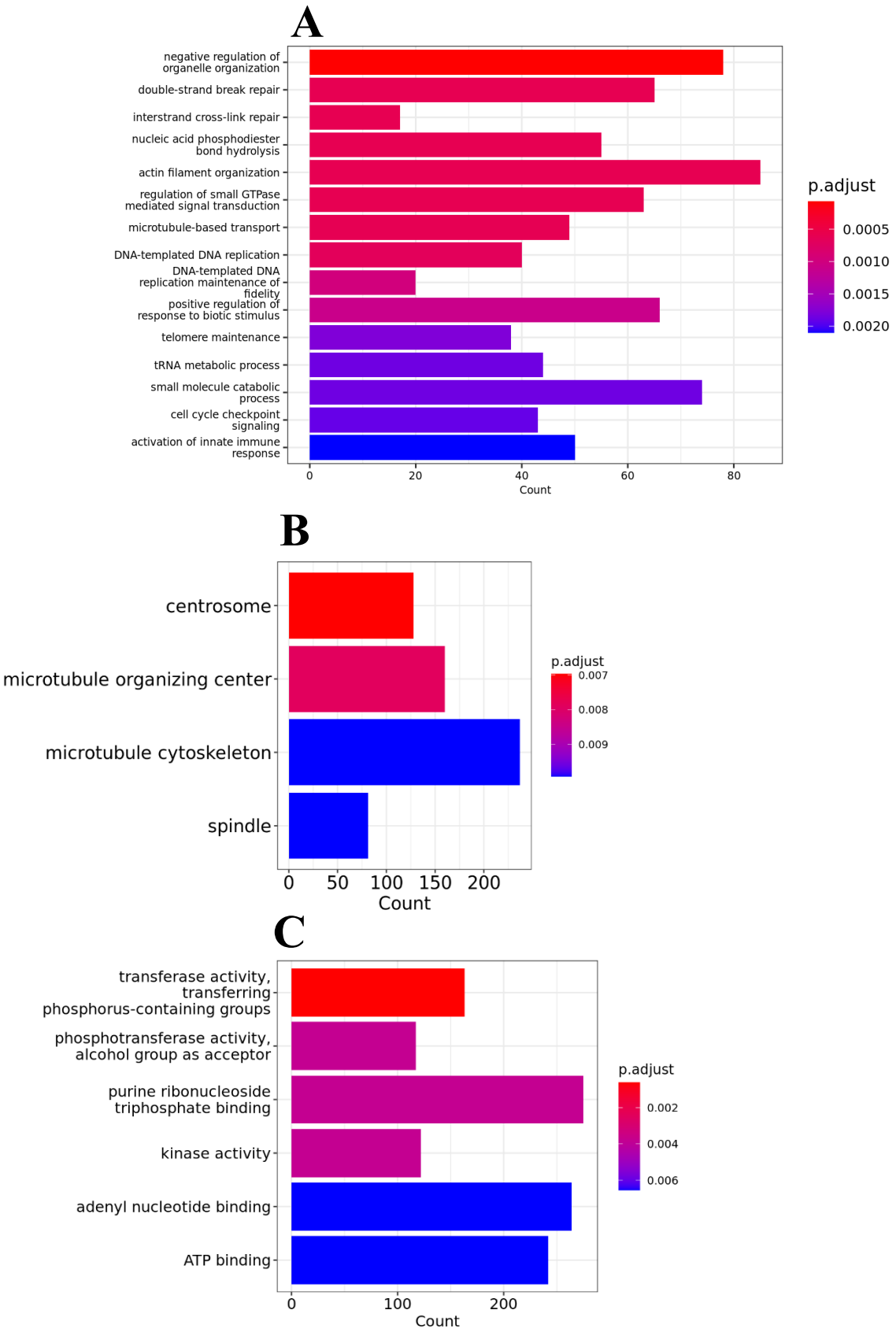


Figure 28. (A) Enriched Biological Processes GO Terms for the allele-pairs which showed a bias in either the blood, liver, lung, thigh, brain, heart, and blood. (B) Enriched Cellular Component GO Terms for the allele-pairs which showed a bias in either the blood, liver, lung, thigh, brain, heart, and blood. (C) Enriched Molecular Function GO Terms for the allele-pairs which showed a bias in either the blood, liver, lung, thigh, brain, heart, and blood.

III.VI. EVOLUTION OF ALLELIC BIAS

The allelic bias present across three unrelated (not direct sisters) *A. neomexicanus* individuals can be observed in Figure 29.A. Overall, 828 allele pairs showed a bias in blood (348 allele-pairs displayed an *A. arizonae* bias and 480 allele-pairs displayed an *A. marmoratus* bias) in at least one of the three *A. neomexicanus* individuals: 10934, 11118 and 11174. To understand if the allelic biases which were observed were novel and not present in the expression of the parental species, the variation between the expression of the *in silico* tissue and the *A. neomexicanus* was compared. The expression of the *in silico* tissue reference represents the expression of the parental species. If this expression was within 3 standard deviations of the *A. neomexicanus* expression and therefore indistinguishable from the *A. neomexicanus* individuals, the allelic bias was classed as being present in the parental species and having been passed on to *A. neomexicanus* (Figure 29.B). Of the 828 allele-pairs which showed a bias in at least one of the three blood samples, 545 were determined as having originated in the parental species and having been passed on to the *A. neomexicanus* lineage.

If the allelic bias was more than 3 standard deviations from the mean expression of the *A. neomexicanus* samples, and therefore distinguishable from *A. neomexicanus* samples, the allelic bias was likely to have evolved over time. Cis or trans gene regulation differences may have arisen since the first historical hybridization event causing the allelic biases observed in the *A. neomexicanus* samples [29]. A total of 283 allele-pairs identified were considered distinguishable from the parental gene expression (Figure 29.C). A large majority, 80.5%, of these are *A. marmoratus* bias allele-pairs, suggesting that genome expression-level dominance may be occurring [92] in blood.

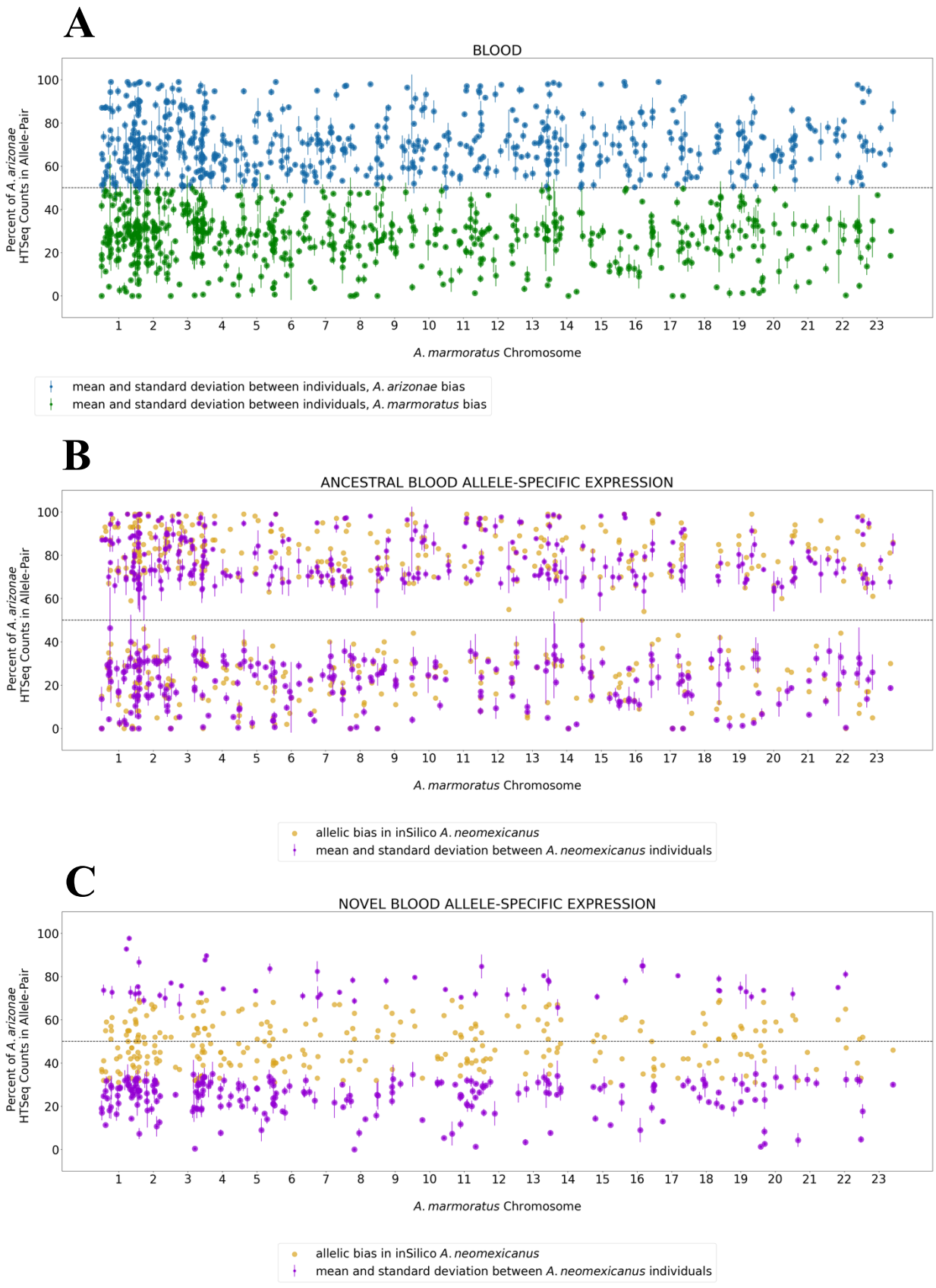


Figure 29. (A) A total of 828 allele-pairs showed a bias in blood in at least one of the three *A. neomexicanus* individuals: 10934, 11118 and 11174. Each point represents the average allelic bias between the individuals and the bar represents the standard deviation: 348 allele-pairs displayed an *A. arizonae* bias and 480 allele-pairs displayed an *A. marmoratus* bias. The allele-pairs which showed a bias were divided into ancestral or novel allelic biases. (B) If the *in silico* reference (yellow), consisting of the expression in the parental data for blood, was within the variation of the allelic bias in *A. neomexicanus* (purple) it was determined to be ancestrally passed on. Of the 828 allele-pairs

which showed a bias in one of the *A. neomexicanus* individuals, 545 (295 *A. arizonae* bias and 252 *A. marmoratus* bias) were determined to have originated from the parental expression. (C) The remaining 283 (55 *A. arizonae* bias and 228 *A. marmoratus* bias) allele-pairs displayed a different allelic bias pattern than observed in the parental species. These were determined as being novel bias in *A. neomexicanus*.

No enriched Biological Processes, Cellular Component or Molecular Function terms were identified for either ancestrally established or novel allelic biases in *A. neomexicanus* blood samples. Although no statistical enrichment was detected, the top 30 most common terms were visualised to understand which genes are being affected by allele-specific expression (Supplemental Figure 1).

The evolution of ancestral biases was investigated in tissues for which an *in silico* tissue reference was present. The separation of ancestral and novel allelic biases was also investigated in five *A. neomexicanus* heart samples: 14584, 11255, D15, D37 and D41 versus one *in silico* tissue reference. Of the 907 allele-pairs which showed a bias (360 *A. arizonae* bias and 457 *A. marmoratus* bias) in at least one of the *A. neomexicanus* individuals, 742 (279 *A. arizonae* bias and 463 *A. marmoratus* bias) allele-pairs were determined as having been present in the parental expression, whereas 165 (81 *A. arizonae* bias and 84 *A. marmoratus* bias) where classed as novel biases. The classification of the ancestral and novel allelic bias for further tissues in which an *in silico* reference was present can be observed in Table 9. A higher proportion of *A. marmoratus* bias versus *A. arizonae* bias can be observed in the novel allelic bias in liver, lung, thigh, and blood. *A. marmoratus* genome expression-level dominance [92] may be occurring in these tissues.

Table 9. Summary of the ancestral and novel allele-specific expression in the tissues for which one *in silico* tissue reference was present.

	Heart (n=5)	Brain (n=4)	Liver (n=4)	Lung (n=4)	Thigh (n=4)	Blood (n=3)
Total allelic biases	907	613	1,304	1,116	797	828
Total ancestral allelic biases	742 (81.8%)	464 (75.7%)	1,101 (84.4%)	932 (83.5%)	630 (79.0%)	545 (65.8%)
Total novel allelic biases	165 (18.2%)	149 (24.3.8%)	203 (15.6%)	184 (16.5%)	167 (21.0%)	283 (34.2%)
A. arizonae bias	360	223	508	480	334	348

A. marmoratus bias	457	390	796	636	463	480
Ancestral *A. arizonae* bias	279	163	436	408	290	293
Ancestral *A. marmoratus* bias	463	301	665	524	340	252
Novel *A. arizonae* bias	81	60	72	72	44	55
Novel *A. marmoratus* bias	84	89	131	112	123	228

III.VII. INTER-INDIVIDUAL ALLELIC BIAS

With the hypothesis that epigenetic mechanisms contribute to phenotypic diversity among individuals of a clonally reproducing species, the allele-specific expression was investigated between individuals for three wild *A. neomexicanus* populations, and one lab raised population. The 10 individuals in the lab population were kept under the same conditions and the individuals were scarified at between 1 – 2 years old. The wild *A. neomexicanus* individuals were collected along the Rio Grande region of New Mexico in either Socorro (n =10), Bernalillo (n = 5), or Doña Ana (n=4). The longitude and latitude coordinates, as well as the date of collection are stated in Supplemental Table 1.

The number of allele-pairs which constituted of over 30 HTSeq counts varied between 3,099 – 3,538 between the 29 individuals (Figure 30.A). A total of 1,499 allele-pairs displayed a bias in at least one of the 29 individuals (620 *A. arizonae* allelic bias and 879 *A. marmoratus* allelic bias, Figure 30. B). The similarity between these individuals can be observed in the PCA in Figure 31. A separation between the lab and Doña Ana individuals to the Socorro and Bernalillo individuals can be observed along PC1 (21.4 %).

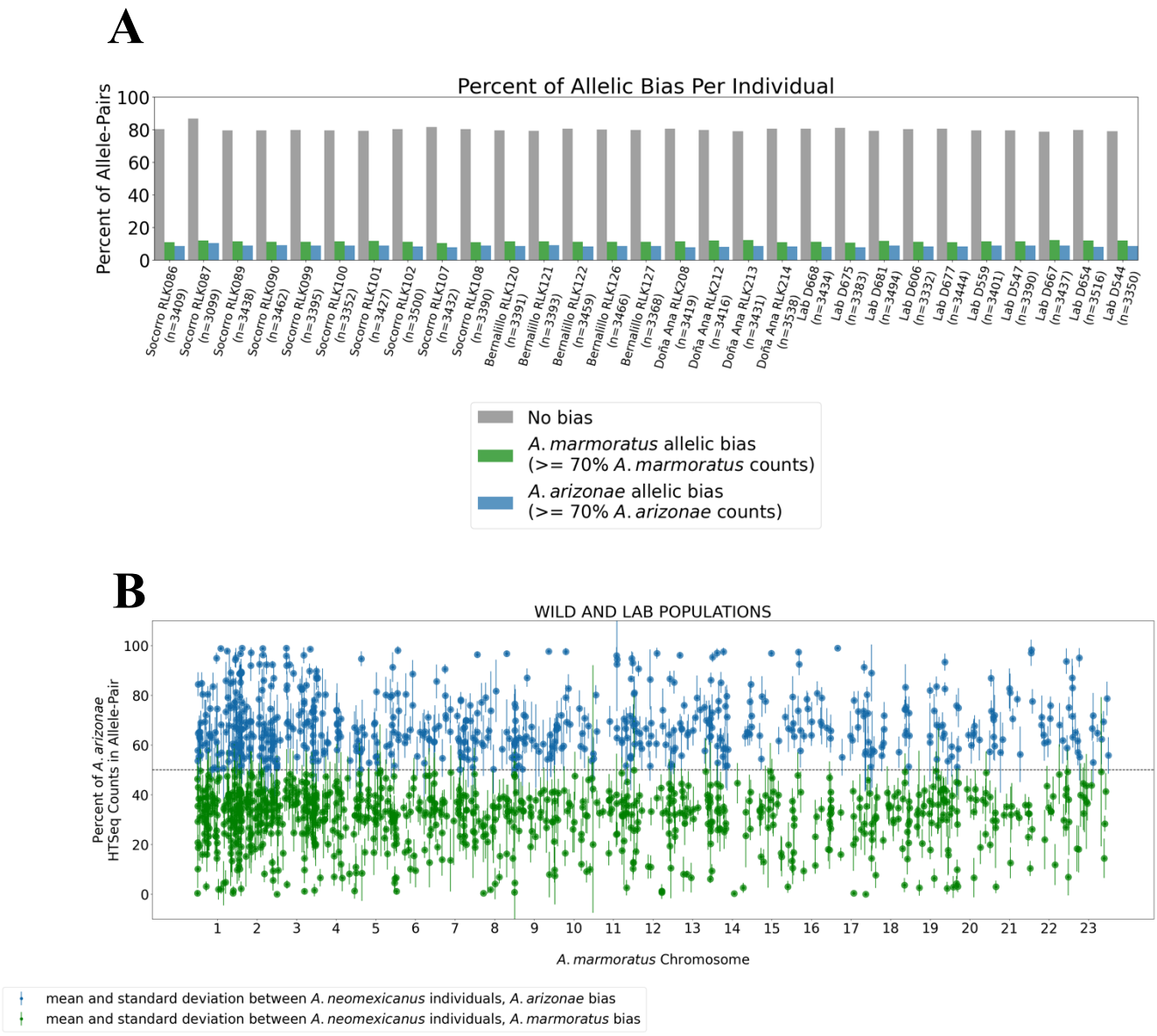


Figure 30. (A) The percent of allele-pairs which passed the filtering criteria of consisting of at least 30 HTSeq counts per allele-pair in wild and lab raised *A. neomexicanus* individuals. (B) These were subset across the tissues to allele-pairs which showed a bias in a least one of the lizards (1,499 in total; 620 *A. arizonae* allelic bias and 879 *A. marmoratus* allelic bias). Each point represents an allele-pair consisting of >= 30 HTSeq counts, and the bar represents the standard deviation between the 29 individuals.

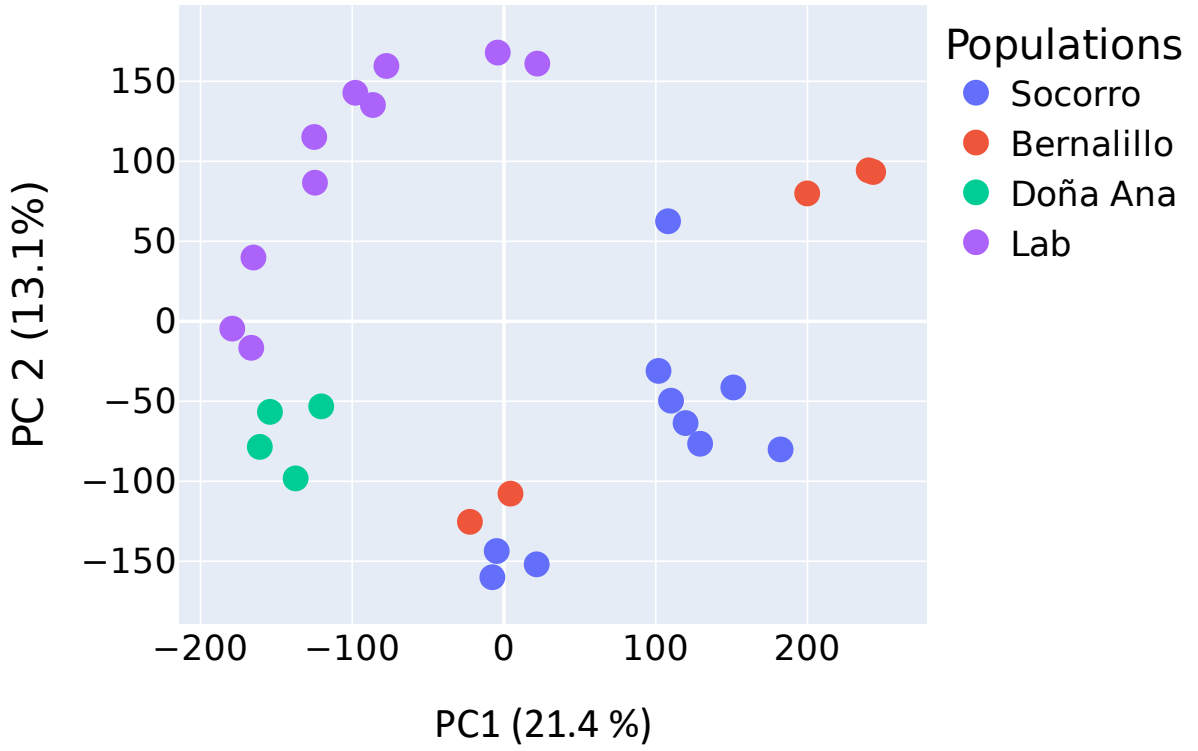


Figure 31. PCA of the 1,499 allele-pairs which showed a bias in at least one *A. neomexicanus* individual. A separation can be observed between the lab and Doña Ana individuals, and the Socorro and Bernalillo locations.

As the populations formed separate clusters, the allele-specific expression (Figure 32) and GO term enrichment (Figure 33) was performed for each population individually. Throughout the four populations a slight majority of allele-specific expression was *A. marmoratus* bias (~ 60 % of biased allele-pairs). A similar quantity of allele-pairs was identified in the blood samples throughout the four populations (~1,000 allele-pairs) corroborating the quantity of tissue-specific allele-specific expression expected per tissue.

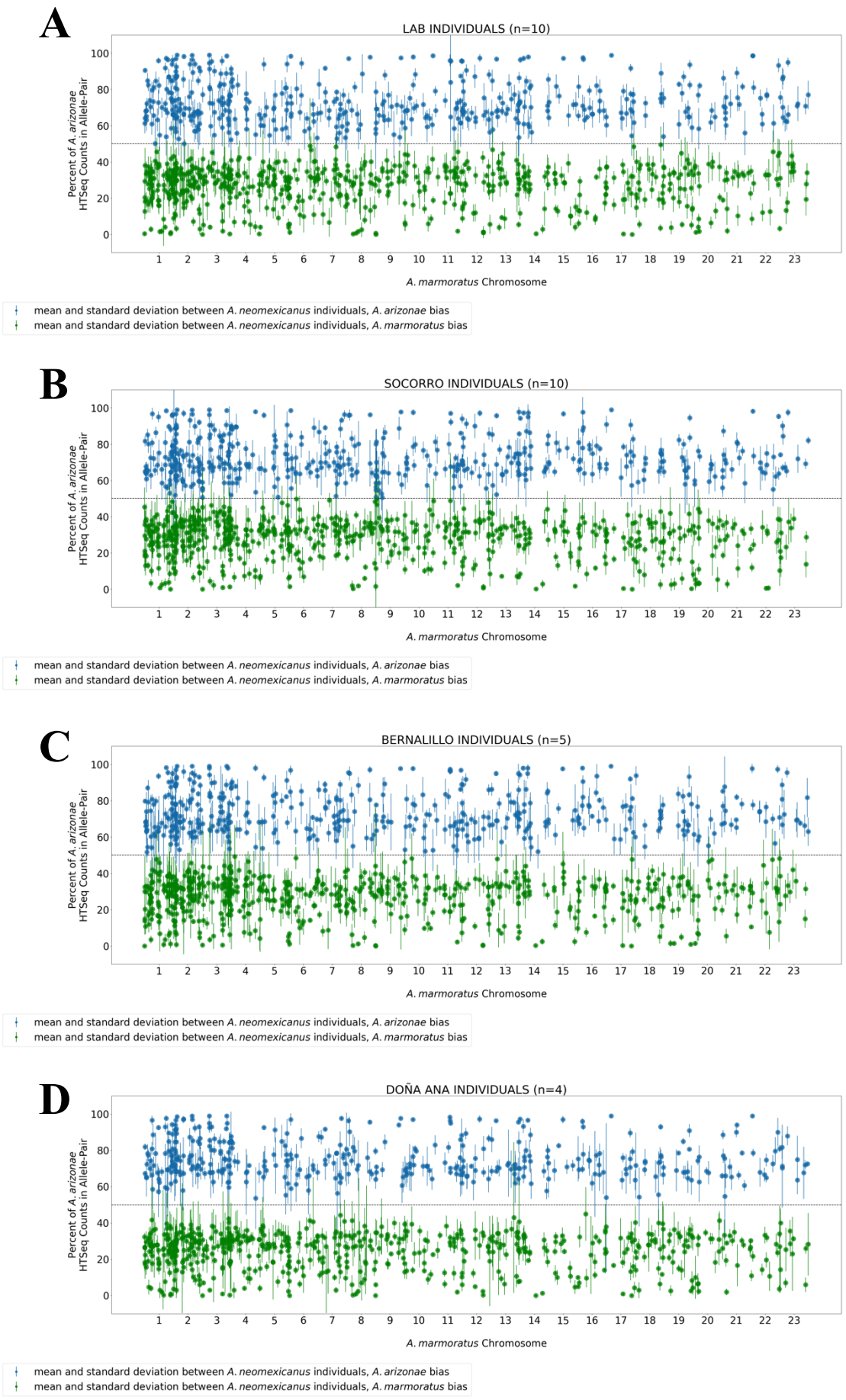


Figure 32. The percent of allele-pairs which passed the filtering criteria of consisting of at least 30 HTSeq counts per allele-pair and displayed a bias in a least one of the lizards in: (A) the lab populations (1,115 in total; 456 *A. arizonae* allelic bias and 659 *A. marmoratus* allelic bias), (B) the

Socorro population (1,162 in total; 485 *A. arizonae* allelic bias and 677 *A. marmoratus* allelic bias), (C) the Bernalillo population (1,087 in total; 446 *A. arizonae* allelic bias and 641 *A. marmoratus* allelic bias) and in the (D) Doña Ana population (1,024 in total; 405 *A. arizonae* allelic bias and 619 *A. marmoratus* allelic bias).

Homologous proteins were identified in humans for 987 of the 1,115 allele-pairs which showed a bias in at least one individual in the lab colony for GO Term enrichment. For the 1,162 allelic bias genes identified in the Socorro population, homologous proteins were identified for 1,028 genes. Of the 1,087 allelic bias genes for the Bernalillo individuals, homologous genes were identified 959 of them. In Doña Ana 1,024 allele-pairs were identified as having a bias. Of these, homologous proteins were identified in humans for 908 of them.

The number of enriched Biological Process GO Terms varied from four in the Doña Ana population, to nine in the Socorro population (Figure 33). No significant Cellular Component GO Terms were identified for any population. However, significantly enriched Molecular Function GO Terms were identified in each population (Figure 34).

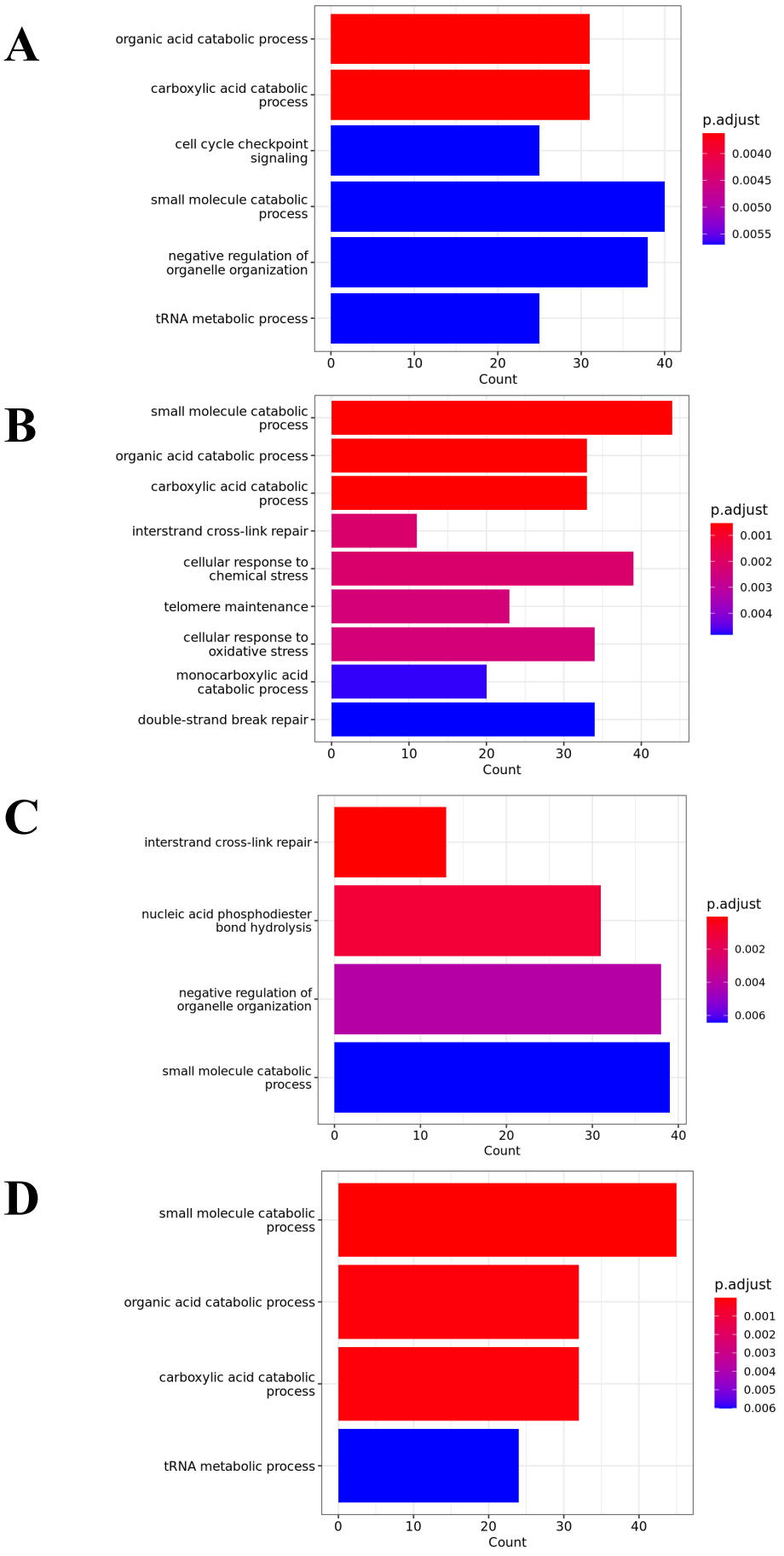


Figure 33. The significantly enriched (Benjamini-Hochberg adjusted P.value =< 0.01) Biological Process GO Terms for *A. neomexicanus* individuals in the lab (A), Socorro (B), Bernalillo (C) and Doña Ana (D) populations.

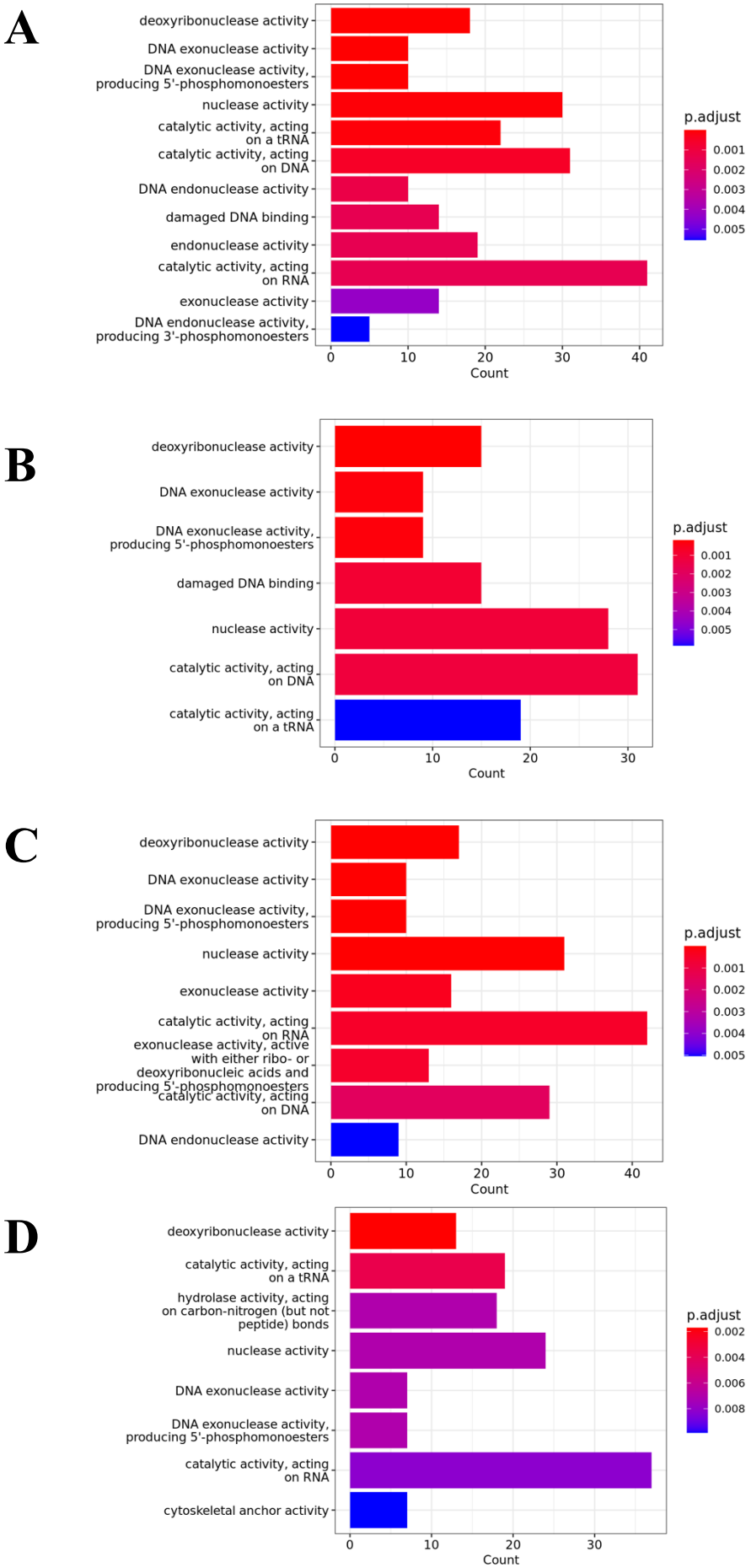


Figure 34. The significantly enriched (Benjamini-Hochberg adjusted P.value =< 0.01) Molecular Function GO Terms for *A. neomexicanus* individuals in the lab (A), Socorro (B), Bernalillo (C) and Doña Ana (D) populations.

IV. DISCUSSION

IV.I. TISSUE-SPECIFIC ALLELE-SPECIFIC EXPRESSION

Allele-specific expression offers insights into the effects of genetic variation on gene expression regulation in a clonally reproducing species. Allele-specific expression was identified throughout the eight tissues and 29 individuals of *A. neomexicanus.* On average, allele-specific expression was present in approximately ~20 % of allele-pairs which were highly expressed (>= 30 HTSeq counts) of the filtered, high quality, allele-pairs per sample. This number may be higher when including high quality annotated allele-pairs which were 100 % identical (21.48 %; 3,839/17,871 of the genome annotations), however quantifying the gene expression per allele becomes problematic. It becomes problematic as expression from each allele cannot be assigned to either allele, as there are no species-specific differences on which to anchor it on.

When comparing the quantity of allele-specific expression in the parthenogenetic *A. neomexicanus* lineage, to a sexually reproducing species such as humans [37] or cows [39], an elevated amount of allele-specific expression in *A. neomexicanus* is noted. The GTEx [37] reported an average of 2.7 % of allele-specific expression across tissues, however an average of 24 % allelic bias was observed across the tissues of the clonally reproducing lineage. The tissue with the lowest amount of allele-specific expression identified in humans [37] was corroborated in *A. neomexicanus*: the brain. However, in the *A. neomexicanus* lineage, the lung and liver displayed the highest amount of allele-specific expression compared to blood in humans [37]. In cows, elevated levels of allele-specific expression were observed in the lung (24–38 %). An average of 7–12 % allele-specific expression was observed across all tissues in cows. Although this is higher than what was observed in humans, it is still below the levels observed in *A. neomexicanus,* a clonally reproducing species. This hints at how allele-specific expression may be more crucial in clonally reproducing species, to create diversity. Further studies would have to be undertaken to investigate the difference in levels of allele-specific expression between sexually reproducing and parthenogenetic species. As the evolutionary distance between humans and *A. neomexicanus* is relatively large, it would be beneficial to compare allele-specific expression differences between sexually reproducing and parthenogenetic species with a smaller evolutionary distance.

Studying tissue-specific allele-specific expression can help uncover context-dependent gene expression patterns and identify tissue-specific regulatory elements [93]. Enriched Biological Process GO terms such as 'microtubule-based transport' and 'actin filament organization' relate to processes that are carried out at cellular level and result in the assembly and arrangement of cytoskeletal structures comprising of microtubules and their associated proteins. Guanosine triphosphatases, or GTPases, are a class of enzymes involved in the regulation of various cellular processes [94]. They play crucial roles in signal transduction, cell division, intracellular trafficking, and many other cellular functions. Additionally, the allelic biases identified across the *A. neomexicanus* tissues were either *A. arizonae* bias or *A. marmoratus* bias. This means that proteins from either genome are potentially interacting to assemble cytoskeletal components as well as for organelle organization.

The GO terms present in *A. neomexicanus* blood were not statistically enriched, however, the most frequently present GO terms pertained to metabolic and catabolic processes. Additionally, multiple cellular localization and organization terms can be observed highlighting how proteins of antagonistic allelic bias may be interacting in the shared cellular space. The Cellular Component GO Terms 'cell periphery', 'cytoplasm' and nucleoplasm are present in both ancestral and novel allelic biases identified in *A. neomexicanus* blood. Both categories share GO Terms involved in 'ion binding', 'organic cyclic compound binding' and 'protein binding'. Additionally, allele-specific expression may be involved in 'ATP hydrolysis' which is used as an energy source in certain reactions [95]. Its presence as an enriched GO term suggests that genes which exhibit allele-specific expression may be important for energy release and may contribute to the success of the parthenogenetic species.

It is important to note that a reduced representation of allele-specific expression was used for GO term enrichment analysis. The background for which allele-specific expression consisted of a filtered down allele-pair dataset (37 %; 6,636/17,871 allele-pairs). Homologous proteins were identified in humans for 75 % (4,997/17,871) of allele-pairs. Humans were chosen for the GO term enrichment analysis as it is the most highly annotated genome available however, it has a large evolutionary distance to *A. neomexicanus*. Therefore, many genes impacted by allele-specific expression may be lost in the analysis. When performing the GO term enrichment analysis, homologous proteins were not identified for every allele-pair that displayed an allelic bias across the *A. neomexicanus* tissues or individuals.

The diversity in allelic bias observed in *A. neomexicanus* tissues may contribute to heterosis in the lineage. Heterosis, also known as hybrid vigour, is a phenomenon in which offspring from the crossbreeding of two genetically distinct parents results in a hybrid offspring which inherits a combination of favourable alleles from both parents [96]. This improvement can manifest in various aspects, such as growth rate, disease resistance, fertility, and overall fitness. Hybrid vigour is commonly observed in plants [96] and animals [97]. In agriculture and animal breeding, farmers and breeders often utilize hybrid vigour to produce hybrid plants or animals with desirable traits [98], [99]. For example, hybrid crops may have higher yields or better resistance to pests and diseases [22], while hybrid livestock may exhibit improved growth rates and milk production traits [99]. This genetic diversity can lead to increased dominance of advantageous traits and the masking of harmful recessive alleles, resulting in enhanced performance and adaptability of the hybrid offspring [100]. In *A. neomexicanus* this may result in the species ability to utilize advantageous alleles for growth, development, and environmental adaptation across different tissues. This could potentially contribute to its success, despite the absence of genetic diversity typically generated through sexual reproduction.

The identification of tissue-specific allele-specific expression differences across *A. neomexicanus* tissues indicates the presence of tissue-specific gene regulatory mechanisms. There were 1,025 (37%, 1,025/2,804) instances in which an *A. arizonae* bias was present in one tissue but displayed an *A. marmoratus* bias in a different tissue. This transcriptomic allele-specific expression diversity originates from one genome. Differences encoded in the genome cannot account for the diversity in allele-specific expression differences displayed across tissues. This demonstrates how epigenetic mechanisms [27] may be key in regulating tissue-specific allele-specific expression.

IV.II. MECHANISMS OF ALLELE-SPECIFIC EXPRESSION

Epigenetic mechanisms provide a layer of regulation of diversity upon which selection can act on [101] . This generates diversity within the clonal species which may aid in adaptation to a changing environment. By quantifying allele-specific expression patterns across different tissues and individuals, we can begin to understand how these variations may have phenotypic consequences in the long-term success of the obligate parthenogenetic species, *A.*

neomexicanus. Firstly, the switch in allele-specific expression across tissues in an individual, highlights how gene regulatory mechanisms can generate diversity within an individual. Secondly, variations in allele-specific expression between individuals and populations highlights how gene regulatory mechanisms can generate inter-individual diversity. Ultimately, these multi-layer variations regulated by epigenetic mechanisms generate diversity within a clonal species, potentially facilitating adaptation to a dynamic environment and aiding in the long-term success of the lineage.

Epigenetic mechanisms are one factor which could be contributing to variation in allele-specific expression and therefore diversity in the *A. neomexicanus* lineage. Cis- genome regulated gene expression as well as differences in the promoter sequences between the *A. arizonae* and *A. marmoratus* alleles could account for ancestral allelic biases present in *A. neomexicanus.* In blood, ~66 % (545/828) of allelic bias allele-pairs displayed an allelic bias similar to the *in silico* parental data. As promoter differences are encoded in the genome, allelic bias which are regulated by these differences would have to be present across multiple tissues and individuals. In the tissue-specific allele-specific dataset comprising of eight tissues and five individuals, only 16 allele-pairs displayed an *A. marmoratus* bias across all tissues and individuals whilst 6 allele-pairs displayed an *A. arizonae* allelic bias across tissues and individuals. It is highly likely that these allelic biases are regulated by genomic differences such as promoter sequence differences, inherited from the parental species.

IV. III. INTER-INDIVIDUAL ALLELE-SPECIFIC EXPRESSION

As a second layer of diversity, the detection of distinct allele-specific variations across wild and lab *A. neomexicanus* populations demonstrates how the epigenome is a potential avenue to introduce variability between individuals in the clonal lineage. Not only is the epigenome contributing to variability between *A. neomexicanus*, but these variations are heritable in the clonal lineage. Distinguishable patterns of allele-specific expression exist between the three wild and lab population investigated and may be linked to environmental factors. This ultimately generates phenotypic diversity in the unisexual lineage and can influence its evolutionary process. In addition, it has been observed that changes in epigenetic signatures such as DNA methylation can be lost or gained more rapidly than genetic ones. They can be accumulated and transmitted for generations after the exposure to an environmental perturbation (i.e., temperature, toxins, hormone exposure, etc.) [102]. The increased plasticity of

epigenetic marks means that they require stronger selective pressure to be fixed in a population. The original historic hybridization event giving rise to *A. neomexicanus* is estimated to have occurred 200,000 years ago [7]. The presence of the distinct allele-specific expression patterns across the populations indicates the selective pressures faced in each environment have been sufficient to elicit alterations to the epigenome. This demonstrates how selection has been acting at the epigenetic level in the clonal lineage.

Certain enriched Biological Process GO Terms identified such as 'small molecule catabolic process' and 'carboxylic acid catabolic process' were identified across the four *A. neomexicanus* populations. The presence of the of these terms across the tissue-specific allele-specific expression Biological Process GO Terms suggests that allelic bias plays an important role in these processes. Contrastingly, certain terms are unique to individual populations such as 'cellular response to chemical stress' in the Socorro population. This highlights how allele-specific expression varies not only across the individuals but throughout the populations as well.

The enriched Molecular Function GO terms ranged from seven significantly enriched GO Terms in Socorro to 12 significantly enriched GO Terms in the lab population. Similarly to the enriched Biological Process GO Terms, certain GO Terms are omni present across populations. This includes 'DNA exonuclease activity', 'deoxyribonuclease activity' and 'catalytic activity acting on RNA'. Nevertheless, specific terms are unique to populations, for example the 'cytoskeletal anchor activity' is only enriched in the Doña Ana population. The unique enriched GO terms per location highlights the adaptability of the allele-specific expression to its population and respective environment.

IV. IV. PARENT-OF-ORIGIN EFFECTS ON ALLELE-SPECIFIC EXPRESSION

To study the directional allelic shift in *A. neomexicanus*, *in silico* tissue references of parental species were included to observe the contribution of gene expression from parental species. Parent-of-origin effects were identified in multiple tissues and allelic bias genes in *A. neomexicanus.* Including the gene expression data of parental species in allele-specific studies allows for evolutionary insights into the effects of cis- and trans- genome gene regulatory mechanisms in regulating allele-specific expression and can shed light on the adaptive significance of gene regulation. As with the *Cobitis* hybrid study [29], trans-genome

gene regulatory interactions were identified in *A. neomexicanus.* Putative trans- genome gene regulatory interactions were identified to play a role in ~34 % (283/828) of allelic bias allele-pairs identified in blood as these displayed a novel allelic bias pattern when compared to the *in silico* expression. Moreover, higher levels of allele-specific expression the lung and liver tissues were observed in the lineage than expected from the *in silico* tissue data. This indicates how the emergence of allele-specific expression for parents to offspring of hybrid origin can create diversity in the lineage.

IV.V. CONCLUSION

This study reveals a high prevalence of allele-specific expression differences in the clonal *A. neomexicanus* lineage, indicating the influence of epigenetic mechanisms at play. Tissue-specific regulatory mechanisms and interactions between *A. arizonae* and *A. marmoratus* alleles contribute to cellular processes and energy usage. The diversity in allelic bias may enhance adaptability and performance, despite clonal reproduction. Additionally, the epigenome plays a role in introducing variability, and distinct allele-specific patterns in wild and lab populations suggest environmental influences. Understanding these mechanisms as well as the interaction of genome-specific interactions sheds light on gene regulation's adaptive significance and evolutionary processes. Overall, allele-specific expression contributes to diversity and potentially to the long-term success in clonally reproducing species like *A. neomexicanus*.

V. FUTURE DIRECTIONS

Integrating allele-specific expression data with other omics data, such as chromatin accessibility and epigenetic modifications, can provide a more comprehensive understanding of gene regulation. First steps have been undertaken to detect epigenetic mechanisms at play. However, studies investigating how specific histone modifications, promoter DNA methylation and non-coding RNAs are targeting and regulating allele-specific expression in *A. neomexicanus* would have to be conducted. Chromatin accessibility information could be integrated through using ATAC-Seq (Assay for Transposase-Accessible Chromatin with high throughput sequencing), CUT & RUN (Cleavage Under Targets & Release Using Nuclease) or ChIP-Seq (Chromatin-Immunoprecipitation). DNA methylation can be investigated by Whole Genome Bisulfite Sequencing or through Nanopore sequencing and DNA methylation base calling.

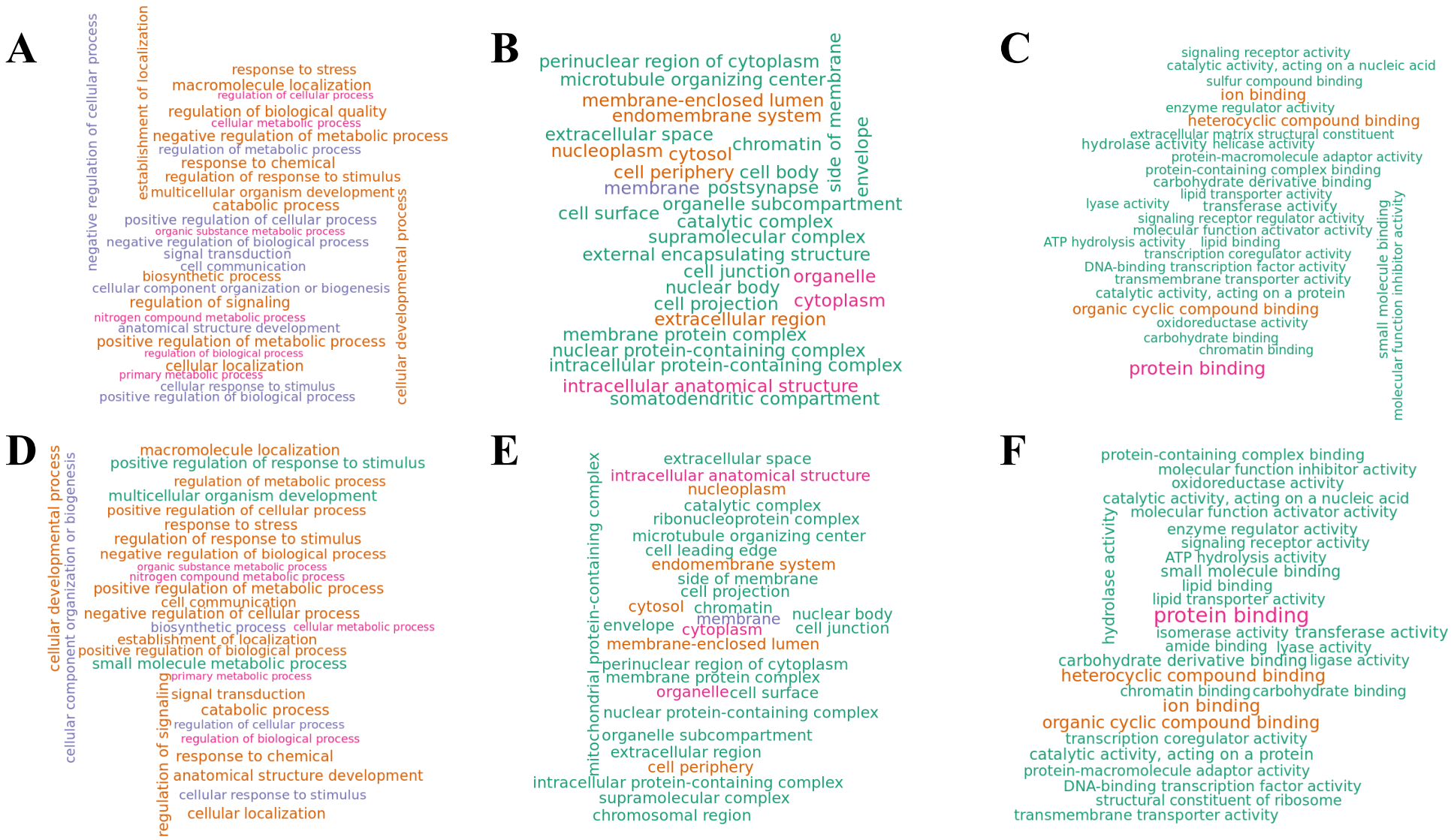
Advancements in single-cell RNA sequencing and computational methods hold promise for dissecting allele-specific expression at the single-cell level. This approach can unveil cell-type-specific regulatory mechanisms and better understand cellular heterogeneity. Allele-pairs which display 'no bias' at the tissue RNA-Seq level may be masking allele-specific expression occurring at the single-cell level.

Detecting significant allele-specific expression signals often requires large sample sizes, especially for rare alleles or subtle expression differences. Adequate statistical power is critical to avoid false discoveries. Therefore, increasing the sample size of the *in silico* tissue is paramount for untangling novel and ancestral allelic biases in *A. neomexicanus.* Bartoš et al. 2019 [29] performed GLMs using n = 4. With a larger sample size, similar GLMs could be performed to determine cis- and trans- regulated gene expression mechanisms at play in *A. neomexicanus*.

Additionally, the parent-of-origin effect may be tracked through development in cases of genomic imprinting or maternal/paternal allele-specific expression. Genes from one parent are silenced, leading to a dosage imbalance and potential phenotypic effects. The identification of imprinted genes and their contribution to developmental processes and disease susceptibility is crucial in understanding epigenetic regulation.

A. neomexicanus is one of many obligate parthenogenetic lineages which exist in the *Aspidoscelis* genus. These obligate parthenogenetic lineages vary in ploidy number and of sexual parental species contributions. Through using the method established in this study it would be possible to investigate how allele-specific expression varies across obligate parthenogenetic lineages originating from hybridization events between different parental species combinations as well as increasing in ploidy level.

VI. SUPPLEMENTARY FIGURES



Supplemental Figure 1. (A) Top 30 most common Biological Process, (B) Cellular Component GO Terms, (D) Molecular Function GO Terms for allele-pairs which were determined to have an ancestral allelic bias pattern in blood. (D) Top 30 most common Biological Process, (E) Cellular Component GO Terms, (F) Molecular Function GO Terms for allele-pairs which were determined to have evolved a novel ancestral allelic bias pattern. No statistical enrichment was identified for any Biological Process, Cellular Component or Molecular Function GO Term.

Supplementary Table 1. Meta data of date and coordinates of where the wild *A. neomexicanus* individuals were caught.

SampleID	Day.Month.Year	State	County	Lat	Long
RLK086	29.6.2018	New Mexico	Socorro	34.158644	-106.8874
RLK087	29.6.2018	New Mexico	Socorro	34.158644	-106.8874
RLK089	29.6.2018	New Mexico	Socorro	34.158644	-106.8874
RLK090	29.6.2018	New Mexico	Socorro	34.158644	-106.8874
RLK099	29.6.2018	New Mexico	Socorro	34.158644	-106.8874
RLK100	29.6.2018	New Mexico	Socorro	34.158644	-106.8874
RLK101	29.6.2018	New Mexico	Socorro	34.158644	-106.8874
RLK102	30.6.2018	New Mexico	Socorro	34.1212876	-106.8912
RLK107	30.6.2018	New Mexico	Socorro	34.155818	-106.88468
RLK108	30.6.2018	New Mexico	Socorro	34.203002	-106.92196
RLK120	2.7.2018	New Mexico	Bernalillo	35.081461	-106.66964
RLK121	2.7.2018	New Mexico	Bernalillo	35.1307543	-106.68333
RLK122	2.7.2018	New Mexico	Bernalillo	35.1307543	-106.68333
RLK126	3.7.2018	New Mexico	Bernalillo	35.070546	-106.44496
RLK127	3.7.2018	New Mexico	Bernalillo	35.1307543	-106.68333
RLK208	28.5.2019	New Mexico	Doña Ana	32.24994	-106.82133
RLK212	29.5.2019	New Mexico	Doña Ana	32.24704	-106.82327
RLK213	29.5.2019	New Mexico	Doña Ana	32.24631	-106.82299
RLK214	30.5.2019	New Mexico	Doña Ana	32.25612	-106.84009

ACKNOWLEDGMENTS

Firstly, I would like to thank my supervisor for giving me the opportunity to join the laboratory and for pushing me to be develop my scientific skills. I am grateful for everything I have learned from you throughout this time. I would also like to extend my gratitude to my co-supervisor for their guidance and feedback throughout the duration of this research.

Thank you also to my colleagues who created a vibrant working environment. I am also immensely grateful to my collaborators for their valuable contributions to this work.

I am deeply grateful to have the support of my friends, parents, sisters and partner throughout my Ph.D. pursuit. Your unwavering encouragement and belief in me have been an invaluable source of strength and motivation.

To my parents and sisters, your endless support has been the cornerstone of my success. Your understanding, and constant encouragement have made it possible for me to pursue this degree.

To my partner, your patience, love, laughter, and belief in me have been a guiding force through moments of self-doubt.

Lastly, thank you to my dear friends for accompanying me during this academic journey. Your laughter and companionship have brought joy and balance to my life.

REFERENCES

1. Lowe, C., and Wright, J., Evolution of Parthenogenetic Species of Cnemidophorus (Whiptail Lizards) in Western North America. *Journal of the Arizona Academy of Science*, **4,** 81-87 (1966).
2. Alves, M. J., Coelho, M. M. & Collares-Pereira, M. J. Evolution in action through hybridisation and polyploidy in an Iberian freshwater fish: a genetic review. *Genetica* **111**, 375–385 (2001).
3. Manning, G. J., Cole, C. J., Dessauer, H. C. & Walker, J. M. Hybridization Between Parthenogenetic Lizards (*Aspidoscelis neomexicana*) and Gonochoristic Lizards (*Aspidoscelis sexlineata viridis*) in New Mexico: Ecological, Morphological, Cytological, and Molecular Context. *Am. Mus. Novit.* **3492**, 1 (2005).
4. Freitas, S. N. *et al.* The role of hybridisation in the origin and evolutionary persistence of vertebrate parthenogens: a case study of Darevskia lizards. *Heredity* **123**, 795–808 (2019).
5. Maheshwari, S. & Barbash, D. A. The Genetics of Hybrid Incompatibilities. *Annu. Rev. Genet.* **45**, 331–355 (2011).
6. Oliver, G. V. & Wright, J. W. THE NEW MEXICO WHIPTAIL, CNEMIDOPHORUS NEOMEXICANUS (SQUAMATA: TEIIDAE), IN THE GREAT BASIN OF NORTH CENTRAL UTAH. *West. North Am. Nat.* **67**, 461–467 (2007).
7. Barley, A. J., Nieto-Montes De Oca, A., Manríquez-Morán, N. L. & Thomson, R. C. The evolutionary network of whiptail lizards reveals predictable outcomes of hybridization. *Science* **377**, 773–777 (2022).
8. Cole, C. J., Dessauer, H. C. & Barrowclough, G. F. Hybrid Orgin of a Unisexual Species of Whiptail Lizard, Cnemidophorus neomexicanus, in Western North America: New Evidence and a Review. *Am. Mus. Novit.*

9. Neaves, W. B. & Baumann, P. Unisexual reproduction among vertebrates. *Trends Genet.* **27**, 81–88 (2011).
10. Cole, C. J. Chromosome inheritance in parthenogenetic lizards and evolution of allopolyploidy in reptiles. *J. Hered.* **70**, 95–102 (1979).
11. Allopolyploidy - an overview | ScienceDirect Topics. https://www.sciencedirect.com/topics/medicine-and-dentistry/allopolyploidy.
12. Lutes, A. A., Neaves, W. B., Baumann, D. P., Wiegraebe, W. & Baumann, P. Sister chromosome pairing maintains heterozygosity in parthenogenetic lizards. *Nature* **464**, 283–286 (2010).
13. Dessauer, H. C. & Cole, C. J. Clonal inheritance in parthenogenetic whiptail lizards: biochemical evidence. *J. Hered.* **77**, 8–12 (1986).
14. Neaves, W. B. Adenosine deaminase phenotypes among sexual and parthenogenetic lizards in the genus Cnemidophorus (teiidae). *J. Exp. Zool.* **171**, 175–183 (1969).
15. Neaves, W. B. & Gerald, P. S. Lactate dehydrogenase isozymes in parthenogenetic teiid lizards (Cnemidophorus). *Science* **160**, 1004–1005 (1968).
16. Cordes, J. E. & Walker, J. M. Skin Histocompatibility between Syntopic Pattern Classes C and D of Parthenogenetic Cnemidophorus tesselatus in New Mexico. *J. Herpetol.* **37**, 185–188 (2003).
17. Maslin, T. P. Skin grafting in the bisexual teiid lizard Cnemidophorus sexlineatus and in the unisexual C. tesselatus. *J. Exp. Zool.* **166**, 137–149 (1967).
18. Cuellar, O. & Smart, C. Analysis of histoincompatibility in a natural population of the bisexual whiptail lizard Cnemidophorus tigris. *Transplantation* **24**, 127–133 (1977).
19. Loewe, L. & Lamatsch, D. K. Quantifying the threat of extinction from Muller's ratchet in the diploid Amazon molly (Poecilia formosa). *BMC Evol. Biol.* **8**, 88 (2008).

20. Knight, J. C. Allele-specific gene expression uncovered. *Trends Genet.* **20**, 113–116 (2004).
21. De La Chapelle, A. Genetic predisposition to human disease: allele-specific expression and low-penetrance regulatory loci. *Oncogene* **28**, 3345–3348 (2009).
22. Guo, M. *et al.* Allelic Variation of Gene Expression in Maize Hybrids[W]. *Plant Cell* **16**, 1707–1716 (2004).
23. Zhang, X. & Borevitz, J. O. Global Analysis of Allele-Specific Expression in *Arabidopsis thaliana*. *Genetics* **182**, 943–954 (2009).
24. Todesco, M. *et al.* Natural allelic variation underlying a major fitness trade-off in Arabidopsis thaliana. *Nature* **465**, 632–636 (2010).
25. Von Korff, M. *et al.* Asymmetric allele-specific expression in relation to developmental variation and drought stress in barley hybrids. *Plant J.* **59**, 14–26 (2009).
26. Knowles, D. A. *et al.* Allele-specific expression reveals interactions between genetic variation and environment. *Nat. Methods* **14**, 699–702 (2017).
27. Al Aboud, N. M., Tupper, C. & Jialal, I. Genetics, Epigenetic Mechanism. *StatPearls* (2023).
28. Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E. & Lappalainen, T. Tools and best practices for data processing in allelic expression analysis. *Genome Biol.* **16**, 195 (2015).
29. Bartoš, O. *et al.* The Legacy of Sexual Ancestors in Phenotypic Variability, Gene Expression, and Homoeolog Regulation of Asexual Hybrids and Polyploids. *Mol. Biol. Evol.* **36**, 1902–1920 (2019).
30. Glover, N. M., Redestig, H. & Dessimoz, C. Homoeologs: What Are They and How Do We Infer Them? *Trends Plant Sci.* **21**, 609–621 (2016).

31. Bell, G. D. M., Kane, N. C., Rieseberg, L. H. & Adams, K. L. RNA-Seq Analysis of Allele-Specific Expression, Hybrid Effects, and Regulatory Divergence in Hybrids Compared with Their Parents from Natural Populations. *Genome Biol. Evol.* **5**, 1309–1323 (2013).
32. Rosic, J. *et al.* Genetic analysis and allele-specific expression of SMAD7 3′UTR variants in human colorectal cancer reveal a novel somatic variant exhibiting allelic imbalance. *Gene* **859**, 147217 (2023).
33. Zhang, S. *et al.* Allele-specific open chromatin in human iPSC neurons elucidates functional disease variants. *Science* **369**, 561–565 (2020).
34. Gutierrez-Arcelus, M. *et al.* Allele-specific expression changes dynamically during T cell activation in HLA and other autoimmune loci. *Nat. Genet.* **52**, 247–253 (2020).
35. Prickett, A. R. *et al.* Genome-wide and parental allele-specific analysis of CTCF and cohesin DNA binding in mouse brain reveals a tissue-specific binding pattern and an association with imprinted differentially methylated regions. *Genome Res.* **23**, 1624–1635 (2013).
36. Graze, R. M., McIntyre, L. M., Main, B. J., Wayne, M. L. & Nuzhdin, S. V. Regulatory Divergence in Drosophila melanogaster and D. simulans, a Genomewide Analysis of Allele-Specific Expression. *Genetics* **183**, 547–561 (2009).
37. GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*. **6235**, 648-60 (2015)
38. Castel, S. E. *et al.* A vast resource of allelic expression data spanning human tissues. *Genome Biol.* **21**, 234 (2020).
39. Chamberlain, A. J. *et al.* Extensive variation between tissues in allele specific expression in an outbred mammal. *BMC Genomics* **16**, 993 (2015).

40. Zhu, F., Schlupp, I. & Tiedemann, R. Allele-specific expression at the androgen receptor alpha gene in a hybrid unisexual fish, the Amazon molly (Poecilia formosa). *PloS One* **12**, e0186411 (2017).
41. Lu, Y. *et al.* Fixation of allelic gene expression landscapes and expression bias pattern shape the transcriptome of the clonal Amazon molly. *Genome Res.* **31**, 372–379 (2021).
42. Martinez-Ruiz, C. *et al.* Genomic architecture and evolutionary antagonism drive allelic expression bias in the social supergene of red fire ants. *eLife* **9**, e55862 (2020).
43. Michels, E. & De Meester, L. Inter-clonal variation in phototactic behaviour and key life-history traits in a metapopulation of the cyclical parthenogen Daphnia ambigua: the effect of fish kairomones. *Hydrobiologia* **522**, 221–233 (2004).
44. Cagan, A. *et al.* Somatic mutation rates scale with lifespan across mammals. *Nature* **604**, 517–524 (2022).
45. Lacal, I. & Ventura, R. Epigenetic Inheritance: Concepts, Mechanisms and Perspectives. *Front. Mol. Neurosci.* **11**, 292 (2018).
46. Fleischmann, R. D. *et al.* Whole-Genome Random Sequencing and Assembly of Haemophilus influenzae Rd. *Science* **269**, 496–512 (1995).
47. Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **107**, 1–8 (2016).
48. Ejigu, G. F. & Jung, J. Review on the Computational Genome Annotation of Sequences Obtained by Next-Generation Sequencing. *Biology* **9**, 295 (2020).
49. Human Genome Overview - Genome Reference Consortium. https://www.ncbi.nlm.nih.gov/grc/human.
50. RefSeq curation and annotation of the human reference genome. https://www.ncbi.nlm.nih.gov/refseq/about/human/.

51. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol. Biol. Evol.* **38**, 4647–4654 (2021).
52. Alföldi, J. *et al.* The genome of the green anole lizard and a comparative analysis with birds and mammals. *Nature* **477**, 587–591 (2011).
53. Salzberg, S. L. Next-generation genome annotation: we still struggle to get it right. *Genome Biol.* **20**, 92 (2019).
54. Cantarel, B. L. *et al.* MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–196 (2008).
55. Hoff, K., Lomsadze, A., Borodovsky, M. & Stanke, M. Whole-Genome Annotation with BRAKER. *Methods Mol. Biol. Clifton NJ* **1962**, 65–95 (2019).
56. Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
57. Guigó, R. *et al.* EGASP: the human ENCODE Genome Annotation Assessment Project. *Genome Biol.* **7**, S2 (2006).
58. Stanke, M. The AUGUSTUS gene prediction tool. https://bioinf.uni-greifswald.de/augustus/ (2003).
59. Srivastava, A. *et al.* Alignment and mapping methodology influence transcript abundance estimation. *Genome Biol.* **21**, 239 (2020).
60. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
61. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).

62. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
63. RNeasy Mini Handbook - (EN) - QIAGEN. https://www.qiagen.com/us/resources/resourcedetail?id=14e7cf6e-521a-4cf7-8cbc-bf9f6fa33e24&lang=en.
64. TruSeq Stranded mRNA Reference Guide (1000000040498).
65. Li, H. lh3/minimap2. (2023).
66. USADELLAB.org - Trimmomatic: A flexible read trimming tool for Illumina NGS data. http://www.usadellab.org/cms/?page=trimmomatic.
67. Pertea, G. & Pertea, M. GFF Utilities: GffRead and GffCompare. *F1000Research* **9**, 304 (2020).
68. Staff, N. BLAST+ 2.11.0 now available with limited usage reporting to help improve BLAST. *NCBI Insights* https://ncbiinsights.ncbi.nlm.nih.gov/2020/11/12/blast-2-11-0/ (2020).
69. Feiner, N. & Wood, N. J. Lizards possess the most complete tetrapod Hox gene repertoire despite pervasive structural changes in Hox clusters. *Evol. Dev.* **21**, 218–228 (2019).
70. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
71. Institute, E. B. New releases: InterPro 86.0 and InterProScan 5.52-86.0. https://www.ebi.ac.uk/about/news/updates-from-data-resources/InterPro-86.0/ (2021).
72. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
73. Li, H. lh3/seqtk. (2023). https://github.com/lh3/seqtk

74. Babraham Bioinformatics - FastQ Screen. https://www.bioinformatics.babraham.ac.uk/projects/fastq_screen/.
75. Palmer, J. augustus. (2022). https://bioinf.uni-greifswald.de/augustus/
76. Anders, S., Pyl, P. T. & Huber, W. *HTSeq - A Python framework to work with high-throughput sequencing data*. http://biorxiv.org/lookup/doi/10.1101/002824 (2014) doi:10.1101/002824.
77. bamCoverage — deepTools 3.5.2 documentation. https://deeptools.readthedocs.io/en/develop/content/tools/bamCoverage.html.
78. Robinson, J. T., Thorvaldsdottir, H., Turner, D. & Mesirov, J. P. igv.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). *Bioinformatics* **39**, btac830 (2023).
79. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
80. samtools(1) manual page. http://www.htslib.org/doc/samtools.html.
81. Satsuma2. (2023). https://github.com/bioinfologics/satsuma2
82. Piovesan, A. *et al.* Human protein-coding genes and gene feature statistics in 2019. *BMC Res. Notes* **12**, 315 (2019).
83. The Genome Sequence of Anolis carolinensis (Green Anole Lizard). | Literature citations | UniProt. https://www.uniprot.org/citations/CI-34HH11KUD6H66.
84. Smith, J. *et al.* Differences in gene density on chicken macrochromosomes and microchromosomes. *Anim. Genet.* **31**, 96–103 (2000).
85. Koochekian, N. *et al.* A chromosome-level genome assembly and annotation of the desert horned lizard, Phrynosoma platyrhinos, provides insight into chromosomal rearrangements among reptiles. *GigaScience* **11**, giab098 (2022).

86. Microchromosomes are building blocks of bird, reptile, and mammal chromosomes | PNAS. https://www.pnas.org/doi/10.1073/pnas.2112494118.

87. Reeder, T. W., Cole, C. J. & Dessauer, H. C. Phylogenetic Relationships of Whiptail Lizards of the Genus Cnemidophorus (Squamata: Teiidae): A Test of Monophyly, Reevaluation of Karyotypic Evolution, and Review of Hybrid Origins. *Am. Mus. Novit.* **2002**, 1–61 (2002).

88. Lovell, P. V. *et al.* Conserved syntenic clusters of protein coding genes are missing in birds. *Genome Biol.* **15**, 565 (2014).

89. Boichot, V. *et al.* Characterization of human oxidoreductases involved in aldehyde odorant metabolism. *Sci. Rep.* **13**, 4876 (2023).

90. Hashimoto, S., Anai, H. & Hanada, K. Mechanisms of interstrand DNA crosslink repair and human disorders. *Genes Environ.* **38**, 9 (2016).

91. spindle Gene Ontology Term (GO:0005819). https://www.informatics.jax.org/vocab/gene_ontology/GO:0005819.

92. Eckardt, N. A. Genome Dominance and Interaction at the Gene Expression Level in Allohexaploid Wheat. *Plant Cell* **26**, 1834–1834 (2014).

93. Sonawane, A. R. *et al.* Understanding Tissue-Specific Gene Regulation. *Cell Rep.* **21**, 1077–1088 (2017).

94. Guanosine Triphosphatase - an overview | ScienceDirect Topics. https://www.sciencedirect.com/topics/medicine-and-dentistry/guanosine-triphosphatase.

95. Matte, A. & Delbaere, L. T. ATP-binding Motifs. in *Encyclopedia of Life Sciences* (John Wiley & Sons, Ltd, 2010). doi:10.1002/9780470015902.a0003050.pub2.

96. Baranwal, V. K., Mikkilineni, V., Zehr, U. B., Tyagi, A. K. & Kapoor, S. Heterosis: emerging ideas about hybrid vigour. *J. Exp. Bot.* **63**, 6309–6314 (2012).

97. Wang, H. *et al.* Heterosis and differential gene expression in hybrids and parents in Bombyx mori by digital gene expression profiling. *Sci. Rep.* **5**, 8750 (2015).

98. Dodds, K. S. & Mather, K. Hybrid vigour in plant breeding. *Proc. R. Soc. Lond. Ser. B - Biol. Sci.* **144**, 185–192 (1997).

99. Bunning, H., Wall, E., Chagunda, M. G. G., Banos, G. & Simm, G. Heterosis in cattle crossbreeding schemes in tropical regions: meta-analysis of effects of breed combination, trait type, and climate on level of heterosis1. *J. Anim. Sci.* **97**, 29–34 (2019).

100. Botet, R. & Keurentjes, J. J. B. The Role of Transcriptional Regulation in Hybrid Vigor. *Front. Plant Sci.* **11**, (2020).

101. Geng, Y. *et al.* Increased epigenetic diversity and transient epigenetic memory in response to salinity stress in Thlaspi arvense. *Ecol. Evol.* **10**, 11622–11630 (2020).

102. Giuliani, C. *et al.* Epigenetic Variability across Human Populations: A Focus on DNA Methylation Profiles of the KRTCAP3, MAD1L1 and BRSK2 Genes. *Genome Biol. Evol.* **8**, 2760–2773 (2016).