ALTITUDINAL ADAPTATIONS OF EARTHWORMS

Thesis for Doctor of Philosophy

2020

Supervisory group: Prof. P. Kille[1], Dr P. Orozco-Ter-Wengel[1].

[1] Cardiff School of Biosciences, The Sir Martin Evans Building, University of Cardiff, , Cardiff, CF10 3AX, UK.

Iain Perry

Cardiff University, School of Biosciences

Abstract

To date few have looked into how earthworms have adapted or acclimatised to the harsh and dynamic environment of high altitude. In this work, I explore the terrestrial invertebrates, earthworms that were found at high altitude on the volcanic island of Pico in the Azores (Portugal) and at Les Deux Alpes in the French Alps. I initially identify species presence along an altitudinal transect compare species diversity and lineage, before investigating gene regulatory control and genomic adaptation between high and low altitude populations to identify if high altitude populations have acquired a genetic advantage to their low altitude cousin or if all worms have it within them to survive if given time to acclimatise.

Altitudinal transects of two temperate-zone mountains were conducted, at Les Deux Alpes and Pico, to identify presence and abundance of species. The two most abundant species, *Lumbricus terrestris* and *Aporrectodea caliginosa*, were investigated to identify diversity and species lineage to determine which species better allowed for adaption and acclimatisation investigations, that are not heavily influenced by deeply rooted species diversity. Having identified *A. caliginosa* in Pico as the most suitable candidate for investigating adaption and acclimatisation with its low population diversity, an *de novo* genome assembly was developed and annotated.

Live individuals of *A. caliginosa* from a high and a low altitude site on Pico were acclimatised to standard laboratory conditions for six months prior to experimental exposure to conditions simulating six climatic conditions for two weeks with temperature and oxygen as variables. RNAseq was performed on the RNA taken from a body transect (including muscular, nerve and gut tissues) of the exposed experimental worms, and differential gene expression was calculated and explored between the high and low altitude populations. Despite both populations normalising in identical soils for 6 months, high altitude individuals had a lower response in gene expression than the Low altitude individuals and suggested an element of epigenetic conditioning or adaption allowing a more plastic response to the changes in conditions. In particular, HMGB1, a gene that is known for its roles in regulating environmental responses, had a comparatively lower expression in the high altitude population than the low altitude population when exposed to simulated high altitude climatic stressors. SNP analysis from transcriptomic sequences revealed the high altitude individuals had SNPs associated with genes that linked to directly to this gene indicating a level of adaption through SNPs and acclimatisation through potential epigenetic priming within the high altitude population.

Acknowledgements

For giving me the opportunity to study a project of my design, for giving me guidance in all aspect of academia in research and in career, but also for his continued friendship from the beginning making my Ph.D. studentship a fantastic experience I would like to give my deepest thanks to Pete Kille whom I hope to continue working with in the years to come.

To my parents Karen and Stan and my brother Nick I wish to thank for their continued support and encouragement throughout my academic journey to this point.

Thank you to my partner Melissa, who has supported me throughout the Ph.D. and made each day easier and brighter.

Special thanks to Dr Paulino Sapo and the Serviço de Ambiente da Ilha do Pico for the privilege of allowing me to conduct research on their beautiful island.

Further thanks must go to Pablo Orozco-Ter-Wengel, Luis Cunha, Dave Stanton, Szabolcs Hernadi and Stephen Short for their help, support, and friendship throughout this endeavour.

This PhD was funded through Professor Kille and Cardiff School of Biosciences.

Table of Contents

Contents

Data Chapter Summaries

Chapter 3: Earthworm species diversity from temperate altitudes in Pico and Les Deux Alpes.

This assesses population diversity of earthworms found in Pico, Azores and Les Deux Alpes, France. Earthworm diversity, based on mitochondrial COII sequencing, is assessed in maximum likelihood trees, minimum spanning networks and calculated diversity indicators. The vast majority of the 8 species identified in Pico were *A. caliginosa* and *L. terrestris*. In Les Deux Alpes, 9 species were identified where only *A. caliginosa* was found at all sites. *A. caliginosa* had the lowest intra-species diversity in both sites while *L. terrestris*, another high prevalence species had higher intra-species diversity in both sites.

Chapter 4: Historical population demographics of earthworms from Pico and Les Deux Alpes.

This chapter assesses the historical population dynamics of *A. caliginosa* and *L. terrestris*. Based on COII sequencing data used in Chapter 3, and calculated mismatch distributions, Bayesian Skyline plots and time calculated phylogenetic trees each species is assessed and contextualised with other global samples. Neither species in Pico or Les Deux Alpes displays any population expansion and indeed appear to mostly have just exited population constriction following the last glacial maximum. *A. caliginosa* is closely related to most European individuals while *L. terrestris* has a much greater diversity but multiple lineage links with European and Canadian relatives.

Chapter 5: Genome of *Aporrectodea caliginosa.*

Fresh large fragment size genomic DNA is extracted and purified before library preparation and sequencing through Nanopore MinION, 10x Chromium and Illumina 454 technologies. A whole genome is assembled, assessed and annotated. The genome of *A. caliginosa* was assembled, assessed and annotated generating a well annotated genome with megabase N50 contiguity.

Chapter 6: Transcriptomic analysis of *Aporrectodea caliginosa.*

Worms collected from Pico Mountain at high altitude and low altitude (Pico, Azores), are exposed to climatically controlled conditions simulating high and low altitude environments. RNA is extracted from the posterior sections of individuals and sequenced. Differential analysis is performed and gene ontology on gene counts between environmental conditions assessed to look for signatures of adaption in the high altitude population when compared to the low altitude population. The high altitude population had fewer differentially expressed genes than the low altitude population and had fewer shared differentially expressed genes between the simulated altitudinal stressor conditions. The high altitude population demonstrated a more nuanced and measured response than seen in the low altitude population. The environmental response regulatory gene HMGB1 was identified as downregulated in the high altitude population in comparison with the low altitude population.

Chapter 7: Genetic variation of *Aporrectodea caliginosa.*

RNA sequencing data for individuals assessed in Chapter 6 was analysed for genetic variation to identify genomic regions under selective adaptive pressure. Phylogenetic structure was determined via SNPs separating high and low altitude populations. Genes within the high altitude population that passed Fst and nuclear diversity filters were identified. A network of connections was identified through translated protein-protein interactions, between the genes under pressure in the high altitude population and a connection was identified to the environmental response gene identified in chapter 6 (HMGB1), which had lower expression in the high altitude population.

Abbreviations and definitions

Short code	*Full term*
16S	16S Ribosomal RNA
ANOVA	Analysis of Variance
Adaption	Nucleotide changes (substitutions, deletions or insertions) to an organism that aids survival in new environments
Acclimatisation	Regulatory changes to gene expression (both short and long term) in an organism that aides survival in new environments
COI	Cytochrome Oxidase I
COII	Cytochrome Oxidase II
EDG	Elevational diversity gradient
Epigenetic	Non-permanent modification (commonly through methylation) causing longer term regulatory changes to gene expression
Fst	Fixation index, measure of population differentiation due to genetic structure
GO	Gene ontology
GTF/GFF	Gene transfer format/General feature format file
HA	High altitude
HMW	High Molecular Weight
LA	Low altitude
N50	A measure of contiguity for genome assembly assessed through sequence length
NGS	Next Generation Sequencing
Pi	Nuclear diversity
SNP	Single nucleotide polymorphism
Tajima's D	The difference between mean number of pairwise differences and the number of segregating sites

1. Living the High Life

1.1. An introduction to species adaptation and acclimatisation at altitude.

1.1.1. **Biodiversity links with altitude**

High altitude mountainous environments are harsh, unforgiving, and isolated locations for anything to survive. Yet, despite this, many organisms thrive there, adapted, or acclimatised to the multitude of stressors. Indeed, the elevational diversity gradient (EDG) actually increases with altitude for some elements of life and has its highest biodiversity between ~1300 and ~1800 m above sea level (Sanders 2002; Chauhan et al. 2011; Chaladze et al. 2014; Zhao et al. 2019a; McCain 2020).

As explored in greater detail later, there have been many studies that have explored elements of altitude adaption in different groups, these have largely focused on vertebrates and insects (Mitton 1997; Lyman et al. 2003; Scheinfeldt and Tishkoff 2010; Storz et al. 2010). Of these studies, many focus on the reduction of oxygen at high altitude, and the hypoxia inducing effects upon an individual (Schmitz and Harrison 2004; Grieshaber et al. 2005). However, these studies focus on 'air breathing' invertebrates, and to date no significant works have been undertaken to investigate 'non-air breathing' (passive gas diffusion without dedicated gas exchange organs) terrestrial invertebrate ecosystem engineers.

1.1.2. **Describing high altitude earthworm communities.**

Few studies have endeavoured to analyse the direct impact of altitude on Earthworm biodiversity with key data being a by-product of global biodiversity surveys (Hendrix et al. 2008; Phillips et al. 2019), and studies on invasive earthworm species and indirectly as prey items for vertebrates (Virgós et al. 2011; Ortiz-Gamino et al. 2016). High-altitude niches within tropical climates have yielded worms species as integral members of unique macrofaunal assemblages (Rozen et al. 2013; Cardoso et al. 2017). These issues are further complicated since earthworm species compete for niches, so that often over smaller distances the same ecology functions are fulfilled by different species. This provides a challenge when investigating the acclimatisation and adaptive changes that enable species to inhabit these high altitudinal habitats. Ideally this question would be address by identifying a single species across a narrow altitudinal transect with a minimal change in habitat. To explore how these terrestrial sentinels adapt to high altitude living, we need to first understand their ecological role, underlying biology, and physiology.

1.2. An introduction to earthworms.

Earthworms are often referred to as natures ecosystem engineers (Kooch and Jalilvand 2008). They do not have the iconic appeal of Africa's big five or the direct contribution to food security of cattle. Nor do they make as common a household pet as even the more unusual animals such as spiders or snakes. Their outward persona struggles to break free from the term 'dull'. However, these small folk really are like hobbits, "You can learn all that there is to know about their ways in a month, and yet after a hundred years they can still surprise you at a pinch"(Tolkien 2008). For even the smallest of animals can have a big impact on the world, and earthworms have a big impact on this world ranging in habitats from all but the driest or coldest places on Earth (Hendrix et al. 2008). With over 6000 earthworm species and a last common ancestor over 200 million years ago (Dominguez et al. 2015) before even the supercontinent Pangaea broke apart, these global terrestrial warriors have been impacting the environment on an enormous scale.

Earthworms perform a major function in the cycling of nutrients and organic material, removing vegetative litter and partially digested or decomposed plant matter leading to mineralisation. While some organic material is readily decomposed by animal digestion and micro-organisms, some tougher elements such as those with fibrous complexes surrounding plant cells like cellulose are decomposed much faster with further mechanical and biological breakdown by earthworms (Milcu et al. 2006). Edwards and Heath demonstrated that earthworms are arguably the most important of decomposing soil invertebrates by comparing the decomposition of Oak and Beech leaves by earthworms only, to microarthropods only, where earthworms digested far more material than microarthropods (Edwards and Heath 1963). The organic material that an earthworm consumes, is egested as much finer particles with a much greater surface area for microbial decomposition (Edwards and Bohlen 1996). This has been confirmed in studies that used pesticides to decrease earthworm populations and the total particulate organic material increased substantially (Parmelee et al. 1990). Earthworms groups have different feeding habits with Anecic species capable of feeding on large litter fragments, Epigeic species consume surface littler and Endogeic species consuming already fragmented organic material (Hoeffner et al. 2018). Soils without earthworms develop a layer of undecomposed surface organic matter though a large quantity of litter can by consumed quickly with the available organic material the limiting factor. Curry and Schmidt reviewed the amount of varying organic matter that can be consumed by a several earthworm species, with highest levels for various species being herbivore faeces consumption in *Aporrectodea caliginosa* (40-80 mg DM^{-1} FM d^{-1}), Elm leaf in *Lumbricus terrestris* (27 mg DM^{-1} FM d^{-1}) and Alfalfa/clover in *Lumbricus rubellus* (36-52 mg DM^{-1} FM d^{-1}) (Curry and Schmidt 2007). Ingestion rates also vary with *A. caliginosa* ingesting 200-300 mg DM^{-1} FM d^{-1} of pasture soil, 2105 mg DM^{-1} FM d^{-1} of Beechwood mull and as high as 4090 mg DM^{-1} FM d^{-1} in a sandy soil. *L. terrestris* consumed 490-3500 mg DM^{-1} FM d^{-1} of arable land while *L. rubellus* ingested 1920-3010 mg DM^{-1} FM d^{-1} of sandy soil (Baker et al. 2006).

The rate of soil ingestions in turn influences the rate by which earthworms increase mineralisation of organic matter to inorganic forms that can be accessed by plants. The most important cycles include Carbon, Nitrogen and Phosphorus. Earthworms themselves are only estimated to contribute about 5% of the total carbon flow with a small contribution to overall CO_2 with low assimilation efficiencies (Barley 1959). The larger influence on the Carbon cycle from earthworms is the increased flow from decomposing plant litter into a mineralised form. While earthworms reduce the overall storage of carbon in leaf litter, they increase plant growth which in turn stores more carbon (Ostle et al. 2007). As such the flow of carbon increases but earthworms do not greatly contribute to increases or decreases of carbon from the cycle. Nitrogen is also turned over in a similar manner by reducing locked nitrogen in leaf litter, and transferring it to a mineralised form from which microbial action fixes it and plants can uptake (Dominguez et al. 2010). Dissolved phosphorus also increases with earthworm activity, allowing ready absorption by plants (Vos et al. 2014). Rather than just direct chemical influence on Carbon, Nitrogen and Phosphorus, earthworms influence these cycles through the micro-organisms that they support within their gut and the wider soil environment acting in synergy as organic waste managers (Pass 2015; Bhat et al. 2018). Micro-organisms of earthworm's gut and those of the surrounding soils are intertwined. Earthworms consume bacteria from the soil and organic litter, and in turn excrete bacteria in their casts. The ratios of bacterial species change between soil, gut, casts and burrows (Pass 2015).

Earthworms not only breakdown and turnover organic litter, improving soil chemical fertility, but have a critical role in soil structural fertility through their burrowing. Aneic worms create deep burrows, while Endogeic and Epi-Endogeic worms construct extensive shallow and surface burrows and also alter the particulate arrangement of soil, loosening compacted soils improving oxygenation and drainage of the soil structure and drawing down nutrients deep into root structures (Blouin et al. 2013; van Groenigen et al. 2014).

Not all that worms do is to the benefit of plants. They have also been widely reported as pests to crops and of grasslands through consumption of low lying leaves or roots and the spreading of plant diseases from bacterial, fungi and parasites (including those of animals) (Edwards and Bohlen 1996; Montecchio et al. 2015; Bohlen and Lal 2017). Earthworms have also been implicated in contributing to soil erosion in some areas through the bringing of fine particulates to the surface where weathering can remove nutrients from the local environment (Blanchart et al. 2004; Bottinelli et al. 2010). While earthworms also consume some seeds, their casts and surface entrances of burrows provide sanctuary for viable seeds providing overall benefits to germination (Forey et al. 2011).

More than just fish bait, because of their aiding of soil fertility, earthworms are commonly used in environmental management to improve poor soils. This includes improving poor pastures and reclaimed polder soils, improving poor mineral soils and mining waste sites (Andre et al. 2010; Butt 2010; Eijsackers 2010; Blouin et al. 2013; Kille et al. 2013). Earthworms provide an invaluable tool as such, being bioindicators from metal ion toxicity, radioactivity and chemical contamination (Spurgeon and Hopkin 1996; Aslund et al. 2012; Lin et al. 2012; Fujita et al. 2014).

Increasingly earthworms are being utilised as organic waste management tools for the breakdown and disposal of wastes from domestic, agriculture and industry. This includes the processing of sewage, sludges, waste foods from production and consumption or lack of, paper production and landfill wastes (Sinha et al. 2009; Edwards et al. 2010; Singh et al. 2011; Lim et al. 2014). Different species are better suited to different waste managements, however *Eisenia fetida* is the most commonly used along with *Eisenia andreai* (or their hybrid species) in Europe (Yadav and Garg 2011).

The earthworm’s life cycle allows for its fast breeding in these applications, but also have allowed them to colonise hostile environments and survive uncommon environmental catastrophise. Their life cycle begins with conception, either through the mating of two mature earthworms or pathogenesis. Earthworms are hermaphroditic and can release sperm from the seminal vesicles via their clitellum and fertilise a partners or their own ovary. Their birth is in cocoon form which can lay dormant until conditions are suitable. This can avoid seasonal restraints such as frost or flooding. The time from cocoon hatching to sexual maturity will vary between species and is affected by the environmental conditions that they experience but can be as short as six weeks. Life in ideal conditions can be several years long, but seasonal conditions can kill adult populations leaving only cocoons to repopulate (Monroy et al. 2007; Mulder et al. 2007; Butt 2011).

1.3. Challenges of being an earthworm at altitude.

1.3.1. **The mountain.**

Mountains though innately beautiful, can belie their power harbouring often harsh climates and unpredictable weather conditions. This combination makes for an increasingly hostile environment for organisms to colonise as altitude is gained. Conditions associated at high altitude include a vast set of environmental parameters. These encompass, but are not necessarily limited to; temperature, precipitation, oxygen availability. These changes in climate associated with mountain ascension are often exploited for impacts of global warming on the environment and its ecosystems (Inouye and Wielgolaski 2003). For soil invertebrates, parameters also impacted by altitude include; changes in soil pH, soil depth and composition including metal ions and vegetation. Further as altitude increases land availability decreases for species to inhabit and forms funnels of biodiversity (Korner 2007).

The complexity of population studies becomes further obscured when considering local mountain microclimates and the latitudinal changes to season length. Tropical mountains found on and close to the equator, (e.g. Kilimanjaro in Kenya 3°, Chimborazo in Ecuador 1° and Kinabalu in Malaysia 6°), have only a small seasonal variation in temperature with the main climatic variation linked to wet and dry seasons (Meteoblue 2016d,b,e). In temperate mountain regions, seasonality is more pronounced with soil freezing winters directly influencing soil invertebrate lifecycles (Holmstrup 2003). Precipitation remains seasonal in these mountainous regions (e.g. Pico in The Azores 38°, Les Deux Alps in France 45° and Snowbasin in USA 41°), although precipitation can be inaccessible to biota since it is in the form of snow and ice until spring melt (Meteoblue 2016a,c,f).

1.3.2. **Temperature and moisture.**

Moisture availability for earthworms is a major influencing factor on their ability to survive in an environment as their osmoregulation falls within tight bounds(Edwards and Bohlen 1996; Wever et al. 2001). Despite this, they appear capable of surviving a water loss of 50-70% of their total body water (Kretzschmar and Bruchou 1991). In dry soils it is reported that *Aporrectodea* spp. are active during the upper 10 cm of moist soil but decent below 20 cm as soil dries and retreat into diapause (Baker et al. 1992). Earthworms appear to prefer over watering to under, as many species including *L. terrestris* and *A. caliginosa* can survive upwards of 10 weeks submerged in aerated water without food (Edwards and Bohlen 1996). The largest stressor facing the worms in this scenario is the maintenance of an oxygen rich environment and avoidance of temperature extremes.

Temperature and soil moisture normally follow a pattern of high temperature, low moisture and low temperature, high moisture in temperate environments. Earthworms find the latter combination the greater stressor (Nordstrom and Rundgren 1974). Developmental optimum temperatures vary per species with *A. caliginosa* at 12°C and *L. rubellus* at 15-18°C as examples (Graff 1953). Interestingly, temperatures only as high as 26°C and 28°C for *A. caliginosa* and *L. terrestris* respectively have been shown to be lethal in 48 hours, while more understandably, *A. caliginosa* did not survive temperatures below 0°C (Grant 1955). Indeed only frost proof cocoons survive these freezing temperatures by becoming desiccated to prevent internal freezing of their tissues (Holmstrup 2003; Holmstrup and Overgaard 2007). The most common strategy earthworms employ is migration to more suitable conditions, either descending deeper into the soil or ascending closer to the surface (Edwards and Bohlen 1996).

1.3.3. **Oxygen availability.**

Earthworms do not possess dedicated respiratory organs. Instead, they rely on gaseous diffusion through their skin. A passive event facilitated by a relatively efficient closed circulatory system. Earthworms have developed a massive extracellular globin structure, Erythrocruorin (EC) to carry oxygen (Weber and Vinogradov 2001; Hackert and Riggs 2006; Royer Jr et al. 2006). The presence of the erythrocruorins in free solution in earthworm blood co-operatively bind oxygen delivering it efficiently to respiring tissues. Despite that oxygen availability does decrease, (80% at 2000 m and 63% at 4000 m if that found at sea level), earthworms seem not only capable of living in soils with low Oxygen availability, but also high levels of carbon dioxide and a variety of other volcanic gases (Cunha et al. 2014). Giardiana *et al.* performed a series of oxygen equilibrium studies on earthworm erythrocruorin. In whole earthworm blood oxygen saturation at sea level, 2000 m and 4000 m drops from 91.6% to 84.7% and 62.5% respectively (Giardina et al. 1975). This is a greater drop than seen in humans (97.3% to 93.2% and 78.9%), suggesting that despite their adaptive EC, depression of oxygen availability either through poorly aerated soils or high altitude has a larger biophysical impact.

1.3.4. **Soils.**

It is known that earthworm species have preferences for soil depth, with *L. terrestris* preferring to burrow deep and *A. caliginosa* with only shallow soil burrowing (Jegou et al. 1998; Tiunov and Scheu 1999). Soil pH is known to affect earthworms species with some more tolerant to acidic soils than others (Edwards and Bohlen 1996; Loranger et al. 2001). Despite their sensitivity to acidity, earthworms have been reported in soils with a pH as low as 2.6 suggesting some species having an elevated plasticity to the stressor (Spiers et al. 1986). Some species like *L. terrestris* and *L. rubellus* seem capable of inhabiting soils from pH 7 to pH 3.5, while more acid-intolerant species include *A. caliginosa* and *Aporrectodea rosea* and *Dendrodrilus rubida* preferring acidic soils between pH 5 and pH 3.5 (Satchell 1955).

The interrelationship of earthworms and soil type includes mineral and metal ion composition and organic detritus available as a food source (Cortez et al. 2000). Most species including *A. caliginosa* and *L. terrestris* expand to the highest population density in light loam soil, with Clay soils and gravelly sands supporting lower densities. Peaty acidic soils by far support the fewest densities (Guild 1951). Metal toxicity is one of the most studied and reported aspect of earthworm soil relationship (Spurgeon and Hopkin 1999; Spurgeon et al. 2005; Spurgeon et al. 2006; Shouolts-Wilson et al. 2010; Gonzalez-Alcaraz and van Gestel 2016). This is particularly relevant when assessing earthworms found on volcanic mountains which have elevated metal ion concentrations (Varrica et al. 2000; Cunha et al. 2014), but it can also arise from environmental pollution. Elements that are particularly important include antimony, bismuth, cadmium, copper, lead, mercury, nickel and zinc (Edwards and Bohlen 1996). Heavy metals at sublethal levels can affect earthworm growth, cocoon production and feeding patterns while the bioaccumulation of heavy metals are often lethal (Neuhauser et al. 1995).

1.3.5. **Vegetation.**

Vegetation not only changes with altitude, but directional aspect (e.g. North or South), plays a distinct role and flora can be highly specific to mountain regions themselves. Kilimanjaro is estimated to have over 1200 vascular plant species with aspect and altitude dependent growth up to 4100 (Hemp 2006). In contrast the temperate mountains of the Western Alps have a much lower vegetation line around 3000 m, but a higher degree of endemism (Pauli et al. 2003). Both temperate and tropical mountain regions are seeing a global warming dependent shift in vegetation, pushing the vegetation line higher (Pauli et al. 2003; Morueta-Holme et al. 2015). Earthworms as ecological engineers have a direct effect upon the mountain flora and in turn, flora and indirectly fauna, increases the food source and habitat of earthworms (Scheu 2003). Organic matter will influence greatly the number of earthworms that can be supported. Earthworms feed readily on decaying vegetative organic matter, but will also consume animal droppings and the microbes within (Barley 1959). Different species of earthworm have their own preferences of food supply. While *L. rubellus* is a litter-feeder, *A. caliginosa* prefers more decomposed organic matter (Piearce 1972).

Many lower slopes of mountainous regions have seen ecological alteration through alterations to the endemic flora, the organic waste of fauna and the use of industrial fertilisers and pesticides (Olsson et al. 2000; Maeder et al. 2002; Riehl et al. 2013). A loss of 'surface' biodiversity feeds into a loss of soil biodiversity, while invasive earthworms that have been introduced deliberately or by accident through agriculture, both contribute to the alteration of the local ecosystem (Baker et al. 2006; Tsiafouli et al. 2014). Invasive organisms, particularly from the same species that can interbreed with endemic populations, introduce complexities to phylogeny and a species' history of adaption across a changing environment (Anderson et al. 2009).

1.3.6. **Human impact.**

It is well reported that the farming of arable land has a comparatively lower total earthworm population than in grasslands, mainly as a result of mechanical damage during the cultivation of the land (Graff 1953; Dzangaliev and Belousova 1969). The addition of in-organic fertilisers to land is one of the most widely used farming practices to increase yields. The addition of these chemicals can alter the acidity of the soil or toxicity through chemicals like ammonia, but the wide use of pesticides and herbicides are likely to far outweigh and negatively harm earthworms (Zarea and Karimi 2012; Jovana et al. 2014; Pelosi et al. 2014; Singh and Singh 2015; Travlos et al. 2017). In contrast the addition of organic fertiliser, manure, is reported widely to increase earthworm populations (Curry 1976; Anderson et al. 1983). This however can be somewhat of a double-edged sword as animals treated with anti-worming agents can excrete active drug compound into the environment and affecting earthworms. Different drugs have their own half-lives and impacts on earthworms, and while drugs are assessed in the UK for some environmental side effects, this is not extensive. The common horse worming agent Ivermectin has a Manure DT50 over 45 days while many other de-wormers have not even assessed this risk to environmental exposure (Lewis et al. 2018). Similarly, antibiotic drugs like chlortetracycline and anti-inflammatory drugs like Phenylbutazone are widely used for treatment of livestock are known to have long environmental half-lives (Lin et al. 2012; Lewis et al. 2018).

It is clear that earthworms have a wide variety of environmental factors impacting their growth and reproduction as an individual and a population as a whole. The challenge to research is understanding each factor on its own and the effects of factors combined.

1.4. Adaptation and Acclimatization to Altitude

1.4.1. **Physiology of Adaptation and Acclimatization to Altitude**

Adaption is the acquisition of genetic changes that aid the survival of an organism with the environment (Bock 2020). Acclimatization is the mechanism an organism regulates its gene expression to match their physiology to a rapidly changing environment (Edmunds and Gates 2020). An organism's adaption to an environment is one by which changes occur that reduce the physiological pressure exerted by a stressful element of that environment. This occurs over generations rather than a short term acclimatisation by means of differentially regulating genes that can return to 'normal' when environmental stressors diminish (Monge and Leon-Velarde 1991).

Despite their use in the reporting of ecotoxicology few studies have looked at soil invertebrate adaption to altitude (Somme 1989). Much of the currently published research on adaption to high altitude living centres around ecological studies of plants, birds and mammals including human genetics (Lenoir et al. 2008; Simonson et al. 2010; Cheviron and Brumfield 2012). With animals, the common focus of research is centred on high altitude hypoxia (Storz et al. 2007). Strategies for determining adaption to high altitude include the combination of primary screening of hypoxia candidate genes and screening the genome for areas showing evidence of local positive selection (Simonson et al. 2010). Identification of physiological changes in mammals can provide key directions for identifying potential genes of interest, exampled in the

Tibetan Yak, whose enlarged lungs and heart and lack of hypoxic pulmonary vasoconstriction have increased gene pathway activation in hypoxic stress and energy metabolism including regulation of cardiovascular vessel size and regulation of angiogenesis (Qiu et al. 2012).

While much of the existing research of high altitude adaptation investigates cold and hypoxic adaptation, this can have specific implications for endotherms and do not necessarily directly relate to the ectothermic worms (Monge and Leon-Velarde 1991).

Pulmonary and blood oxygen levels have been extensively investigated in mammals and birds, far less research has been invested into earthworms (Storz et al. 2007; Storz 2016). Giardina *et al*. 1974 does however provide a useful insight into blood oxygen dissociation with erythrocruorin across a variety of conditions, including temperature and pressure (Giardina et al. 1975). A decrease in oxygen saturation when combined with an increase in erythrocruorin can keep total blood oxygen normal at high altitude. The oxygen dissociation curve is non-linear sigmodal and Giardina's findings demonstrate that temperature has little effect on the dissociation of oxygen in comparison to atmospheric pressure. The low impact of temperature on oxygen dissociation does not however mean temperature has no other physiological impact on earthworms. It is also worth noting earthworm erythrocruorins are not bound in erythrocytes as with birds and mammals and do not follow the cellular production rate and cellular lifespan as such the red blood cell pathways described by Wiback and Palsson (Wiback and Palsson 2002; Elmer et al. 2012). *L. terrestris*'s erythrocruorin is reported as 'extremely stable' with a half-life of 28 hours (Elmer and Palmer 2012). Dorsal and ventral blood vessels transport earthworm blood through a closed circulatory system pumped by five pairs of aortic arches (Hama 1960). But with gas exchange across the body surface, rate of blood flow is not a limiting factor for delivering oxygen to tissues. Earthworms must, however, keep skin moist to enable effective gas diffusion for oxygen in and carbon dioxide out (Grant Jr 1955). Skin is kept moist through mucus excretion and through behaviour seeking out moist environments. Dry environments can therefore prove a lethal environment while saturated environments can also lead to drowning. Despite this double edged sword earthworms can lose over 70% of their body weight in water loss and recover (Grant Jr 1955).

1.4.2. **Tissue and cellular response to Altitude**

As discussed in the previous section, two major elements of high altitude are the decreased oxygen and the decreased temperature. Although much of the research on mammals and birds is not directly applicable to earthworms, tissues and cells of animals will experience a similar impact and respond through analogous pathways (Hernández-Oñate and Herrera-Estrella 2015). Looking into the responses of other animals' genes and pathways can provide foundations for predictions of what may be expected when looking into an earthworm's response to high altitude.

Oxygen deprivation is extensively researched *in vitro* and *in vivo* for high altitude research and disease, particularly those that cause tumours (Noman et al. 2015). One of the most highly researched protein and pathway for high altitude and hypoxia is the hypoxia-inducible factor-1 alpha (HIF-1α). This protein in the absence of oxygen will form a complex with HIF-1β and p300 to allow binding to hypoxia response elements (HRE), but is rapidly broken down in the presence of oxygen via Prolyl hydroxylas enzymes (PHD) and the subsequent binding of Von Hippel-Lindau (VHL) tumor suppressor for ubiquitination and degradation by the 26S proteasome (Hoppeler and Vogt 2001; Ke and Costa 2006; Benzonana et al. 2013). Hoppeler *et al.* in their study of skeletal muscular tissue recorded an upregulation of HIF-1α that they attribute to a rise in mRNAs for myoglobin, vascular endothelial growth factor and glycolytic enzymes that functional analyses have positive effects of V_{02max}. Manalo *et al.* estimated HIF-1 impacts 2% of all human genes in arterial endothelial cells, and there are in excess of 100 genes associated with HIF-1 (Manalo et al. 2005). Key pathways known to change in response to hypoxia include: Erythropoiesis and iron metabolism through the upregulation of transferrin and ceruloplasmin, Angiogenesis through increase of VEGF; matrix metalloproteinases (MMPs) and genes involved with vascular tone, Glucose metabolism through the increase of nearly all the enzymes in the glycolytic pathway and glucose transporters (GLU1 and GLU2); Cell proliferation and survival through induction of growth factors (IGF2 and TGFα) and activation of MAPK and PI3K pathways (Ke and Costa 2006; Ratcliffe 2013).

When looking at pathways of cold weather tolerance associated with residence at high altitude it is important to remember that worms are ectotherms and are bound by different rules than warm-blooded animals that are most heavily researched. With some ectotherm vertebrates, survival includes living through long periods of cold weather and even freezing. This requires a response to minimise cellular mortality from shrinkage, macromolecule and membrane damage and altering the metabolism of the cell to limit oxidative stress (Costanzo and Lee 2013). Though some species can indeed regulate the nucleation of ice crystals to minimise cellular cytolysis, others utilise supercooling to survive with biological 'antifreeze' proteins (AFPs, AFGLs, IBPs, INPs and HSPs) (Dahlhoff and Rank 2000; Duman 2015). With the reduction of metabolism glucose metabolism alters as with hypoxia directly linking the response of hypoxia and cold tolerance. It is worth noting that cellular, tissue and an organism's response to cold weather is not just the process of cooling, but also the process of warming from cold temperature.

Although mountains are often subjected to high rainfall, this does not guarantee moist soils. High surface run-off, porous soils and precipitation falling as snow can lead to dehydrating conditions for the earthworm to survive. Response to low water availability has common elements as with cold weather response. Cuticular permeability is reduced, metabolism down regulated, membranes are modified, cytoskeletal reorganisation occurs and HSPs and LEAs are differentially regulated to prevent protein aggregation (Everatt et al. 2015).

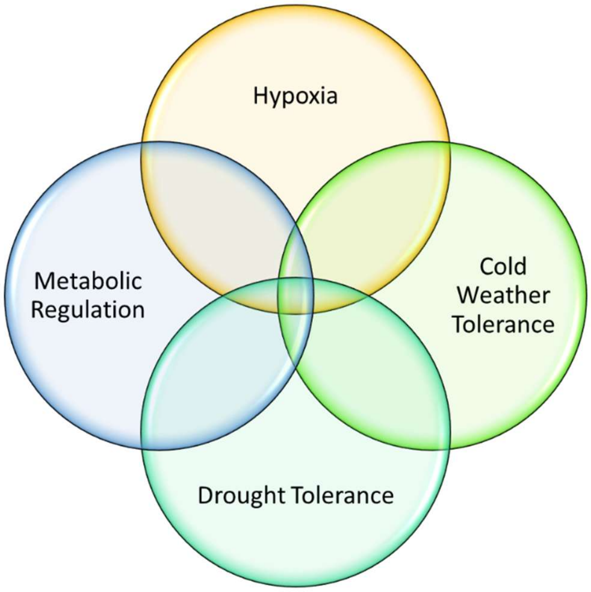


Figure 1: The intertwined nature of pathway responses to Oxygen deprivation and cold weather

Hypoxia, cold weather tolerance, drought tolerance and metabolic regulation are heavily intertwined sharing common pathway responses (Figure 1). We can therefore expect when exposing tissues to one environmental pressure such as hypoxia, we will observe pathways also associated with metabolic regulation. Predicting how a whole organism responds to these conditions is harder as physiological and behavioural responses interplay. It is also challenging to predict how individual genes might be up or down regulated as part of these pathways, particularly when more than one environmental pressure is present potentially amplifying or attenuating responses.

1.4.3. **Hypoxia tolerance in Arthropoda**

Investigations into ‘air breathing’ Arthropoda, in contrast to the earthworms which have no respiratory system, have indicated similar mechanisms for regulation and response to hypoxia as seen in mammals. For Crustacea, this partially involves the increase of oxygen-carrying transport proteins and an increased affinity for oxygen carrying (Schmitz and Harrison 2004), while the species *Cnemodus hirtipes* utilises facultative hypometabolism during Oxygen reduction (Adamczewska and Morris 2000). Less is known mechanistically about Arachnida despite their diversity at high altitude, which while possessing ‘lungs’ are cannot purposely ventilate them (Beron 2018). In this case a lower metabolic requirement for Oxygen seems to allow long terms residence in low oxygen environment, while the reliance passive gas diffusion place restriction on organism size (Harrison et al. 2010).

1.4.4. **Temperature adaption in Arthropoda.**

Similar to the Oxygen Size Rule (Callier and Nijhout 2011), adaptions in Arthropoda (ectotherms) follow a similar size limiting pattern with the Temperature Size Rule (Klok and Harrison 2013). For ectotherms, this follows an inverse relationship of body size with temperature increase. While some insects (e.g. *Manduca sexta*) the signal to stop growing in warmer temperatures comes during growth, *Drosophila melanogaster* does this during its larval stage (Ghosh et al. 2013). Initial experimental evolutionary studies of laboratory population showed large number of SNPs to be associated with multi-generational temperature acclimatisation in *D. melanogaster* (Orozco-terWengel et al. 2012). Subsequent, characterization of the genetic basis of temperature adaptation in *Drosophila spp.* has is revealed that thermal preference and heat shock are traits supported through independent loci (Castañeda et al. 2019). This complexity within laboratory 'forced' evolutionary conditions emphasises the requirement to consider the dynamics of temperature change linked to prevailing natural selective conditions. The North and South facing slopes of 'Evolutionary Canyon' provided a natural laboratory examining adaption within *D. melanogaster* and *Drosophila simulans*, for comparing temperatures. Here, sternopleural bristles densities were associated with temperatures, though this the relationship of this trait to temperature was not identified (Lyman et al. 2003). This perhaps acting as a cautionary tale for stressing the importance of molecular analysis in identifying mechanisms of adaption and acclimatisation over observed morphology.

1.4.5. **Adaption in annelids.**

Although earthworms at high altitude have not yet been investigated, their response to stress has been investigated in numerous toxicological studies, that include response to metal ions found in soil (Spurgeon et al. 2005), flood and drought (Plum and Filser 2005) and seasonal temperature (Svendsen et al. 2006). Since many of these stressors are also stressors that change over an altitudinal transect, parallels can be surmised, and predictions proposed for what earthworm responses might be expected. However, while earthworm genetic diversity allows for potentially more variation and routes to adaption to stressors, it also constrains the predictive capability we can hypothesise as no single response is guaranteed (Spurgeon and Hopkin 1996). Furthermore, not only can species' responses change, but also species themselves might change with altitude, with different species inhabiting a different environmental niche.

1.5. Genomic considerations for high altitude adaption.

1.5.1. **Genomes as templates for acclimatisation and adaptation**

As previously covered, much of the existing detailed mechanistic research into high altitude covers species that can traverse altitude gradients regularly. Use of a single species allows for a more powerful method for identifying the adaptive mechanisms to temperature and hypoxic tolerance either by phenotypic plasticity or population-level variation of the genome. Phenotypic plasticity is strongly associated with modification to regulatory networks but adaptive changes occur to the genome itself (Storz et al. 2010). While both genome regulation and single nucleotide polymorphisms (SNPs) can be identified without the aid of an organism's genome, both are easier, and more powerful with one.

When examining gene expression through RNA sequencing (RNAseq), reads can be arranged and stacked together as a transcriptome, which represents all copies of identified genes. This means genes where RNA expression is too low to identify are not included, while genes with multiple splice variants are retained. With no guide to map reads to, this can cause difficulties in generating read counts for genes (Freedman et al. 2019). Use of a well annotated genome as a template allows for a more precise mapping of RNAseq reads and generation of read counts for genes. When searching for SNP changes between individuals or populations, this again can be performed without a genome (Peterlongo et al. 2010). But, as with transcriptomics, the process becomes more powerful when these SNPs can be mapped to a genome.. Between species with low levels of structural rearrangement, challenges of mapping can be overcome, however in species with high levels of structural rearrangement, hybrid or cryptic species will find mapping does not work as well. With no highly contiguous genome published at the time of analysis for earthworms, and their genomes beginning divergence over 200 million years ago no suitable reference genome exists for mapping earthworm SNPs to (Dominguez et al. 2015) (Anderson et al. 2017b).

1.5.2. **Earthworm cryptic speciation.**

Additional complexity arises in identifying a constant genetic heritage on which to study altitudinal adaptation, particularly when considering what is represented by a morphologically identified earthworm species. Mitochondrial barcoding has revealed deep diversity within species throughout Annelida. These species, although passing the traditional definition, as a group of living organisms that can exchange genes and interbreed to produce fertile offspring, stratify by ecological niche and mate selection (Cosin et al. 2010; Anderson et al. 2017a). They form 'cryptic species' with 5-7 lineages being identified in most common earthworm species (Pérez-Losada et al. 2009; James et al. 2010; Shekhovtsov et al. 2019). Mitochondrially identification of cryptic speciation is not n always possible as the hybridisation of DNA from multiple species occurs in the maternal and paternal copies of autosomal DNA. Though cryptic species can be detected with molecular analysis, experiments might require a significantly larger cohort of individuals to be tested to ensure minimum numbers of the same lineage can be compared adding to the complexity and cost of research. The interweaving of large sections of DNA of different species forming cryptic species is not easily accounted for in examining population dynamics (Leys et al. 2016). Nor is the identification of the genetic basis for fitness easily identified when comparing multiple individuals from a cryptic lineage (Chenuil et al. 2019). Cryptic species should therefore be avoided where possible, when attempting to identify potentially subtle levels of adaption and acclimatisation.

1.6. Underlying rational for the research undertaken in this thesis

As previously described, earthworms can be found across a wide range of environmental niches (Johnston et al. 2014) including altitudinal transects where they act synergistically to support vegetation to slowly push the edge of permanent life into more hostile environments. Unlike the larger vertebrates that can migrate to avoid the seasonal extremes, particularly those found in the temperate mountains, where seasonal changes can include harsh winters to hot dry summers. The sedentary behaviour of these invertebrates means that at the altitudinal extremes, they must accelerate and/or adapt their lifecycle to the restricted viable breeding season. An earthworm will only have a seasonal migration distance measured in the tens of meters, with most forming close population groups (Marinissen and van den Bosch 1992). The Latitudinal Biodiversity Gradient (LBG) is well described in ecology where the restrictive environmental pressures act as a funnel to loss of biodiversity where only key adaptive changes will allow a limited number of flora and fauna to survive at high latitudes (Dowle et al. 2013). In flatter areas, the rates at which these physical parameters can change with latitude occur over hundreds to thousands of kilometres, thereby producing vast and complex ecological linkages. In contrast, the rates that these parameters change with altitude can alter over mere tens of kilometres and therefore provide a potentially highly tractable system. As such, investigating alpine environments negate some of the confounding complications and allow a clearer understanding of genomic adaptation.

In temperate latitudes higher altitudes exhibit longer winter seasons, reducing growth and breeding season durations as the conditions that include ground frost, are compatible only with the survival of earthworm cocoons. Previous studies on earthworms generations include seasonally inundated flood plains in The Netherlands where the adaptive driver selects for earthworms exhibiting accelerated lifecycles which one might expect during the reduced alpine breeding season (Zorn et al. 2005). In mountains within the tropics where temperature averages are higher and the seasonal temperature fluctuation with the associated environmental parameters are minimal, earthworms can colonize much higher altitudes where temperature remains favourable. However, this can now allow them to encounter other stressors such as the restriction of oxygen.

This project aims to test an overarching hypothesis that at altitude in temperate latitudes, seasonal duration is the major selective driver whilst within the tropics adaptation to oxygen availability is more likely to be observed. By using a combination of both phylogeography and functional genomics, the project aims to identify the genetic structure of earthworm populations along altitudinal transects and endeavour to determine the functional basis of acclimatisation to the prevailing conditions. This will be accomplished through a set of five key elements.

1. Identification of appropriate altitudinal transects through assessing earthworm biodiversity.

Earthworm biota will be collected along altitudinal transects within temperate mountain ranges. As no invasive sampling (e.g. chemical treating) can be used, sampling will collect all earthworm species present in each site over an hour sampling period. Pico in the Azores has a volcanic peak reaching from sea level to 2300m over a few kilometres and Les Deux Alpes in the Western Alps, which ranges from 1000m to around 2200m. Both are located between 38 and 46 degrees latitude but represent two very different environmental conditions. Previous studies have already identified *L. terrestris, A. caliginosa* and *L. rubellus* at both sites (Bernier and Ponge 1998; Cunha et al. 2014). After collection, samples will be provisionally washed, speciated, and preserved using RNAlater or ethanol on-site and transported to Cardiff for analysis. Preserved individuals will be taxonomically identified prior to performing stratified molecular analysis. I will determine earthworm community composition using a singular mitochondrial barcode (COII), characterise the population structure of individual species'

2. Genetic lineage of the most abundant species will be assessed, and a suitable species identified for adaption and acclimatization molecular analysis.

Using the molecular barcodes studied above, population demographic dynamics of the two most abundant species in Pico and Les Deux Alpes will be estimated looking for patterns indicative of population expansions or bottlenecks and the lineage of both species in each site will be assessed in comparison to global samples to identify the complexity of ancestry for the selected species. The selected species for genome development should have the minimum mitochondrial diversity and ideally should also have low levels of historical interconnection. This is to minimise the possibility of transcriptional and adaptive genetic changes being masked by complex and deeply rooted variations between sub-lineages. Multiple lineages further increases the numbers required for sequencing and transcriptional experimentation to account for such variation as prior genetic testing would alter many transcriptional stress and wound responses.

3. A genome will be developed via sequencing, *de novo* genome assembly and annotation as a tool for RNAseq analysis and SNP analysis.

Whole genomic sequencing will be undertaken with the use of a combination of novel sequencing platforms with large fragment size DNA extracted from fresh tissue and a genome assemble and annotated for the species. This newly developed genome will aid in transcriptomic analyses and the identification of areas of the genome under adaptive pressure in individuals of the target species at key altitudinal sites.

4. Acclimatisation will be investigated through differential gene expression analysis of individuals from high and low altitude populations.

Additionally, live populations of earthworms from Pico will be brought back from the top of Mount Pico, (above 2000m) and two lower sites (below 300m). They will be cultivated in identical conditions to normalise the populations prior to experimental testing of response to hypoxia and cold weather. Laboratory controlled experiments on earthworm responses to oxygen, temperature and soils will be conducted and the study will explore the signals of functional acclimatisation using global transcript analysis using RNAseq as well as targeted transcript analysis of genes associated with lifecycle control, hypoxia and cold weather tolerance, comparing individuals of target species sampled from the same key altitudinal sites.

5. Adaption will be investigated through SNP variations found in the individuals of the same high and low altitude populations.

Single nucleotide polymorphism (SNP) analysis of individuals from high and low attitude populations will be performed, mapping variations to the genome to identify areas of variability within the high altitude population. Variations will be filtered via Fst and nuclear diversity to identify genes that are potentially under adaptive pressures.

These key analyses will help to determine the fundamental question of if and what levels of adaption earthworms have which aid their living at altitude and if and what level of acclimatisation through gene regulation a high altitude population might have.

2. General methods.

2.1. Earthworm sampling.

2.1.1. **Site collection.**

Earthworm species distribution sampling occurred at each site for 30 minutes as a balance between collecting as many samples as possible and the time available for all sites to be visited. Earthworms were collected over an approximate 20 m^2 sampling area up to a depth of 30 cm during a 1 hour period (trying to collect the maximum possible). Aggressive chemical sampling methods (such as formaldehyde or mustard extract) were avoided as a primary species needed for cultivation in later transcriptomic analysis, needed to be found in sufficient quantity and without harming that could increase chance of mortality in transit prior to cultivation. A combination of digging tools were used to sort through soil at each sampling site and earthworms that were found were collected in a plastic bag filled with soil from the sample site. Worms were then transported to the local research hub for sorting and preservation.

2.1.2. **Preservation for transportation.**

The bags of earthworms from each site were individually sorted in metal trays. Earthworms were sorted by morphotype and washed in water and sedated in carbonated water. Specimens were then preserved in either 96% EtOH (Pico) or in RNAlater (Les Deux Alpes) and transported back to Cardiff University for analysis. On arrival to Cardiff, specimens were placed in individual sample containers and the 96% EtOH or RNAlater was refreshed. Each specimen was given a unique ID code and morphotype confirmed.

2.2. DNA extraction and purification for genetic barcoding.

2.2.1. **'Qiagen' Tissue lysis.**

As specified by the manufacturer, a section of muscle tissue (roughly 4 mg), was cut from the tail section of each earthworm and digested for 2 hours at 56°C in 180 µL of ATL buffer supplemented with 20 µL proteinase K (Qiagen inc., Crawley, UK). When using the Qiagen 96 well DNA extraction and purification kit, tissue was left shaking overnight at room temperature to prevent the silicon lids 'popping off'.

2.2.2. **'Qiagen' DNA purification.**

DNA purification was performed as specified by the manufacturer using Qiagen DNA purification tubes or Qiagen 96 well DNA purification plates (Qiagen). DNA concentration and purity was assessed using a NanoDrop spectrophotometer (NanoDrop Technologies, Wilmington, DE) and samples with very low or no DNA were re-extracted. Extracted samples were stored at -20°C.

2.3. COII PCR and genetic barcode sequencing.

2.3.1. **PCR preparation.**

PCR tubes, stripes and plates together with associated caps and seals were sourced nuclease free (Greiner Bio-One). Other sterile plasticware used were serialised by autoclaving (Autoclave Prestige 2100 Classic 12 L) prior to use. Dedicated pipettes and associated plasticware were placed in a UV-cabinet (UVC/T-AR, BioSan) prior to manipulating the PCR reagents and the cabinet was exposed to a UV light for 10 min. Promega nuclease-free molecular-grade H_2O was used in all with Promega GoTaq PCR reagents PCRs (Promega UK, Southampton, UK).

2.3.1. **PCR primers.**

Primers were ordered from Eurofins genomics (Ebersberg, Germany) and were diluted with nuclease free water (Sigma-Aldrich) to a stock concentrate of 100 µM concentration. Stock concentrations were stored at -20^{o}C. Working concentrations were established by diluted primer stocks to 10 µM concentration, these were kept at +4^{o}C for 1-2 weeks until placed at -20^{o}C for longer periods. For each aliquot of primer the number of freeze/thaw cycles were monitored and primers where this exceeded 5 were disposed. Primer sequences used include:

COII primers (686bp) (Perez-Losada et al. 2009)
COII forward primer: 5'- GGC ACC TAT TTG TTA ATT AGG-3'
COII reverse primer: 5'- GTG AGG CAT AGA AAT ACA CC-3'

2.3.2. **COII PCR mixture.**

PCR reagents were combined to provide a PCR Mastermix, subsequently 24 µL aliquots were placed in PCR tubes or plates prior to the addition of 1 µL of template DNA (~10 ng/µL). A negative control supplementing 1 µL of sterile water instead of template DNA accompanied each batch of amplifications to ensure no background amplification was observed. All reagents were maintained at 4^{o}C prior to being placed in the PCR machine. PCR Mastermix contained the following constituents:

MaterMix constituents per 25 µL reaction volume:

Forward primer (10 µM)	1 µL
Reverse primer (10 µM)	1 µL
dNTPs (10mM)	1 µL
Taq Polymerase (GoTaq)(5U)	0.25 µL
GO Taq buffer 5x	5.0 µL
Nuclease Free Water	16.75 µL

2.3.3. **COII PCR cycle conditions.**

PCR reactions were run in a SimpliAmp$^{\mathrm{TM}}$ Thermal Cycler (ThermoFisher Scientific) with a heated lid. Strips were run centrally in the heating blocks when a full 96 well plate was not run with additional blank strips to prevent over compression and crushing of PCR strips from the auto compression head of the Thermal Cycler.

Initial denaturation	94°C	180s	1 cycle
Denaturation	94°C	30s	36 cycles
Annealing	50°C	30s	
Extension	72°C	60s	
Final extension	72°C	600s	1 cycle

2.3.4. **PCR quantification and sequencing.**

PCR success and amplified product concentrations were determined using a Qiaxcel electronic gel analyser (Qiagen). Successful products were diluted around 1:10 to approximately 10 ng/ µL and dideoxy chain termination sequencing was performed on ABI 3730XL sequencer by Eurofins genomics. AB1 files for each sample sequenced were manually assessed for missed and incorrect calls using FinchTV (Geospiza Inc, UK) and the resultant fasta files were exported for downstream analysis.

2.4. Live worm cultivation.

2.4.1. **Earthworm normalisation.**

To normalise the starting conditions of the earthworms from high altitude (Alpha) and low altitude (Beta) sites they were cultivated under constant laboratory conditions for 6 months prior to experimental manipulation. During this period earthworms were maintained in 20 Litre box of soil consisting of 1:1:1 mixture of topsoil, compost and bark chipping. Every two months, 0.5kg of horse manure free from de-worming agents was added as further worm food (Lewis et al. 2018). Each box was kept moist with polished water whilst drainage hole to prevent oversaturation of the soil. Soil agitation was minimised during the 6 months to encourage full recovery following transport from Pico.

2.4.2. **Transcriptomic analysis soil preparation.**

Identical soil used in earthworm cultivation was used to prepare 12 identical 0.5 kg soil containers. These containers were used to maintain each cohort of 10 worms within constant gas and temperature conditions for the period of the transcriptomic experiment. Full experimental details of these exposure is provided in Chapter 6.

2.4.3. **Earthworm weight measurement**

Earthworms were briefly washed before addition to experimental containers and prior to RNA extraction in deionised water to remove soil, before gently dabbing with tissue to remove excess water. Individuals were then weighed to 2 decimal places. Worm weights were averaged and the standard deviation from the range of weights calculated.

2.5.RNA extraction.

2.5.1. **Direct-Zol RNA MiniPrep.**

Earthworm tissues of 5 mm length (post clitellum, 1mm from the tail tip) were ground with a sterile plastic pestle in Trizol and Frozen. Purification occurred in batches of 20 through the Direct-Zol RNA MiniPrep kit from Zymo Research. To the 600 µL of sample/Trizol mixture, an equal volume of 100% ethanol was added and mixed via vortexing. The mixture was transferred to a Zymo-Spin IIC Column placed into a collection tube and centrifuged at 10,000 g for 1 minute (Zymo Research, Tustin, CA). After placing the spin column into a free collection tube, the direct-Zol RNA PreWash (400 µL) was added to the column and centrifuged through and the passthrough discarded. This step was repeated to ensure thorough washing before the addition of 700 µL of RNA wash Buffer and centrifugation where the passthrough wash buffer was discarded. Samples were eluted in 50 µL of DNase/RNase free water and stored at -80°C. DNase treatment was not performed.

2.6. DNA and RNA quality control.

2.6.1. **Nanodrop.**

DNA concentration and quality was assessed via Nanodrop (Nanodrop 1000, Thermo Scientific). 1 µL of sample DNA spectral absorbance was measured to calculate and estimated nucleic acid concentration and via 260/280 and 260/230 wavelength absorbance ratios, the purity of the sample from chemicals (ethanol, phenol or salts), left from the extraction process.

2.6.2. **Qubit.**

Qubit reactions were used to accurately calculate DNA and RNA concentrations of samples on a Qubit Fluormeter (Qubit 3 Fluorometer, Thermofisher Scientific). Qubit High Sensitivity dsDNA and High Sensisitivity RNA protocols were used as specified in in 'Molecular probes life technologies' manual MAN0002326 and MAN0002327. For all measurements 1 µL of sample DNA or RNA was used more measurement.

2.6.3. **Tapestation.**

Where DNA size and molar concentrations were measured for sequencing libraries, this was measured on a HSD1000 ScreenTape tape (tape 5067-5584, reagents 2067-5585) on a 2200 Tapestation (Agilent Technologies) using the protocol proscribed in the 'Agilent High Sensitivity D1000 ScreenTape System Quick Guide' (Edition 09/2015).

2.6.4. **QIAxcel.**

DNA concentration and integrity for amplified DNA PCR products was calculated using the QIAXcel (Qiagen), on a DNA High resolution kit using the protocol as described in the 'QIAxcel DNA handbook' (11/2017).

Initial extractions of RNA were measured for concentration and sample integrity using the QIAxcel, on a RNA QC Kit v2.0 using the protocol as described the 'RNA Quality Control using the QIAxcel Advanced System' manual (01/2017). The standard QIAxcel RNA size maker 200-6000nt was used.

2.7. SQK-LSK-109 Nanopore library preparation.

Preparation of a sequencing library for long read sequencing using Oxford Nanopore's MK1 R9 Rev-D flowcell is described in detail below with an overview provided in Chapter 5.4.3.

2.7.1. **DNA repair.**

Size fragmented DNA (12-15 kb – see Chapter 5.4.3) was cleaned-up with SPRI beads to remove small fragments. DNA was mixed with SPRI beads at a 1:1 ratio and incubated for 5 minutes before separating the beads with a magnet and discarding the supernatant. Beads were washed twice in 180 µL 85% ethanol (30 second incubation). DNA bound Beads were dried for 5 minutes before elution in 30 µL Illumina Resuspension Buffer. Beads were separated via magnet and DNA containing supernatant transferred to a new tube. Sufficient fragmented DNA at ~16 kbp was produced to allow splitting into two pools. This allowed two library preparations. DNA repair was performed to correct nicks, gaps, deamination of cytosine to uracil, oxidized bases and blocked 3' ends. Libraries were prepared using >1 µg of fragmented DNA mixed in a 0.2 mL thin-walled PCR tube with the following reagents:

Nuclease free H_2O	17 µL
Oxford Nanopore’s CS DNA (λ control DNA)	5 µL
NEBNext FFPE DNA Repair Buffer	3.5 µL
NEBNext FFPE DNA Repair Mix	2 µL
Ultra II End-prep reaction buffer	3.5 µL
Ultra II End-prep enzyme mix	3 µL

This was mixed by gentle flicking of the PCR tube before spinning down and incubating at 20°C for 5 minutes and at 65°C for 5 minutes.

2.7.2. **AMPure XP clean-up.**

The resultant library reaction was cleaned with AMPure XP beads to remove repair enzymes and buffers. AMPure XP beads were vortexed to resuspend the beads after which 60 µL of the resuspended beads was added to a fresh 1.5 mL DNA LoBind tube (Eppendorf UK Ltd, Stevenage) with the 60 µL of repaired fragmented DNA and mixed by gentle pipetting (1:1 ratio). The tube was incubated on a rotation mixer for 5 minutes at room temperature. The tube was briefly spun down (~100 g for ~10 seconds) and the beads pelleted on a strong magnet (Alpaqua 96R plate). The supernatant was removed, and sample remained on the magnetic stand whilst the pelleted beads were washed with x volume of 70% ethanol, this process was repeated twice without disturbing of the pellet after which any remaining wash solution was removed. The pellet was allowed to dry for 30 seconds before removal from the magnet and resuspension in 61 µL Nuclease-free water. The suspended beads were incubated for >2 minutes at room temperature before returning to the magnet to pellet the beads. The 60 µL of resuspended repaired DNA was transferred to a clean 1.5mL Eppendorf DNA LoBind tube.

2.7.3. **Adapter ligation.**

The adapter ligation was performed in the 1.5mL Eppendorf DNA LoBind tube from chapter 2.5.2.

Repaired and cleaned DNA	60µL
Oxford Nanopore’s Ligation buffer (LNB)	25µL
NEBNext Quick T4 DNA Ligase	10µL
Oxford Nanopore’s Adapter Mix (AMX)	5µL

The reaction was mixed by gentle flicking before being spun down and incubated at room temperature for 10 minutes.

2.7.4. **AMPure XP clean-up long fragment selection.**

Following adapter ligation, the reaction mixture was cleaned up with AMPure XP as described in 2.5.2. However, instead of washing with 70% ethanol, Oxford Nanopore's Long Fragment Buffer (LFB) was used to select for DNA fragments greater than 3 kbp. The size selected adapter ligated DNA was eluted in 15 µL of Nuclease free water in a fresh 1.5 mL Eppendorf DNA LoBind tube and 1 µL of eluted DNA was used for quantification with a dsDNA HS Qubit assay.

2.7.5. **Flowcell priming and loading.**

The MK1 R9 Rev-D flowcell requires an input DNA concentration between 5-50 fMol. Quantified DNA concentration from chapter 2.5.4 was used with the central fragment distribution peak from the Genomic Tapestation in chapter 5.4.3. was used to calculate the molar concentration from 12µL of adapter ligated DNA and where necessary diluted to fall within the required concentration range.

Each flowcell was loaded into a MinION sequencer and a QC of the flowcell was performed. Following this, the priming port of the flow cell was opened and a small volume (~20µL) of transport buffer removed. A priming mix was prepared by mixing 30µL of Oxford Nanopore's Flush Tether (FLT) with Flush Buffer (FLB). This was mixed by pipetting and 800µL was added to the priming port slowly avoiding introduction of air bubbles and was left for 5 minutes to incubate.

A sequencing mix was prepared as below in a 1.5mL Eppendorf DNA LoBind tube. Loading beads were mixed immediately before used with a wide bore pipette tip.

Adapter ligated DNA (5-50fMol)	12µL
Oxford Nanopore's Sequencing buffer (SQB)	37.5µL
Oxford Nanopore's Loading beads (LB)	25.5µL

The flowcell's SpotON port cover was lifted and 200µL of the priming mix was added to the priming port avoiding introduction of air bubbles which would kill the flow cell. The sequencing mix was mixed with a wide bore tip immediately before loading and the 75µL was added to the SpotON port dropwise ensuring each drop flowed onto the cell before the next drop was added. The SpotOn port cover was closed and the priming port closed before closing the lid of the MinION. Sequencing could now be started via the MinKNOW operating software.

2.7.6. **Sequencing.**

The progress of the sequencing was observed periodically throughout the sequencing duration. Early failures in library preparation can be detected and flowcells recovered, though this procedure was not needed in our sequencing. Most sequencing and basecalling had finished before the full experimental runtime was achieved and experiments were stopped when all pores had died and basecalling completed. Data was exported for genome assembly.

2.8. RNA library preparation and sequencing.

2.8.1. **RNA library preparation.**

For each sample, 100-150 ng of HS RNA Qubit quantified RNA was used as input material for the Roche KAPA mRNA HyperPrep Kit. For each sample 50 µL of RNA was mixed thoroughly by gentle pipetting with washed KAPA mRNA Capture Beads and incubated for 2 minutes at 65°C and for a further 5 minutes at 20°C. Using a magnetic rack, supernatant was removed and the beads with RNA bound, were washed in 200 µL of KAPA mRNA Bead Wash Buffer. Beads were then resuspended in 50 µL of RNase-free water and incubated for 2 minutes at 70°C and for a further 5 minutes at 20 °C. 50 µL of KAPA Bead Binding Buffer was added and mixed via gentle pipetting. The samples were further incubated for 20°C for 5 minutes before using a magnet to separate the beads from solution and the supernatant removed. Beads were washed as before in 200 µL of KAPA mRNA Bead Wash Buffer before resuspending in 11 µL of RNase-free water and 11 µL of KAPA Fragment, Prime and Elute Buffer (2x).

The purified mRNA/bead samples were then fragmented to 100-200 bp mean library insert size by incubation at 94°C for 8 minutes prior to immediate separation of the beads from the RNA containing supernatant with a magnet to prevent hybridization of the poly(A)-rich RNA to the capture beads. 20 µL of this supernatant was transferred to a new tube containing 10 µL KAPA 1st strand synthesis master mix and mixed via gentle pipetting. The sample was then incubated for 20 minutes at 25°C, 15 minutes at 42°C and 15 minutes at 70°C before cooling to 4°C. 30 µL of KAPA 2nd strand synthesis and A-tailing master mix was added to each sample, mixed and incubated for 30 minutes at 16°C and 10 minutes at 62°C before cooling to 4°C. To each sample a unique adapter/barcode was ligated with the addition of 5 µL of a unique 1.5 µM KAPA Duel index adapter and 45 µL of KAPA Adapter ligation master mix. Samples were incubated for 15 minutes at 20°C

Following adapter ligation, samples were cleaned. The 110 µL of Adapter-ligated DNA was mixed gently with 70 µL of KAPA Pure Beads and incubated for 10 minutes. Beads were separated with a magnet and the supernatant discarded before in a 2 times wash in 200µL 80% ethanol (30 second incubation). DNA bound Beads were dried for 5 minutes before elution in 50 µL Illumina Resuspension Buffer. To this 35 µL of PEG/NaCl was added and mixed gently before washing as before 2 times in 200µL 80% ethanol (30 second incubation). DNA bound Beads were dried for 5 minutes before elution in 20 µL Illumina Resuspension Buffer. Beads were separated via magnet and DNA containing supernatant transferred to a new tube containing 20 µL of KAPA Library amplification master mix and amplified as below:

Initial denaturation	98°C	45s	1 cycle
Denaturation	98°C	15s	16 cycles
Annealing	60°C	30s	
Extension	72°C	30s	
Final extension	72°C	60s	1 cycle
Final hold	4°C	∞	

The 50 µL of amplified DNA samples were cleaned by the addition of 50 µL KAPA Pure Beads and washing as before in a 2 times wash in 200µL 80% ethanol (30 second incubation). DNA bound Beads were dried for 5 minutes before elution in 20 µL Illumina Resuspension Buffer. Of this for each sample 1 µL was used for quantification via an Agilent Tapestation D1000 chip.

2.8.2. **SPRI Bead Pooled sample clean-up.**

Pooled samples were cleaned by mixing gently with SPRI Beads at a 1:1 ratio, incubating for 5 minutes before separating the beads with a magnet and the supernatant discarded. Beads were washed twice in 180 µL 85% ethanol (30 second incubation). DNA bound Beads were dried for 5 minutes before elution in 30 µL Illumina Resuspension Buffer. Beads were separated via magnet and DNA containing supernatant transferred to a new tube.

2.9. Bioinformatic codes.

2.9.1. **SLURM Head script.**

The code below is an example used for submission of job scripts for computation on the SLURM job management system. Both “ntasks” and “mem” changed depending on program requirements. All jobs were run with the SLURM head script.

```
#!/bin/bash
#author: IainPerry
#SBATCH --partition=mammoth
#SBATCH --nodes=1
#SBATCH --ntasks=64
#SBATCH --cpus-per-task=1
#SBATCH --mem=500000
#SBATCH --error=%J.err
#SBATCH --output=%J.out
```

2.9.2. **Longranger pipeline.**

The pipeline for Longranger was run with the commands below (Zheng et al. 2016). This was used for extracting barcoded reads from sequenced 10X chromium reads.

```
Longranger basic --id=ApCa --fastqs=ApCa-gDNA --sample=ApCa-gDNA
```

2.9.3. **Miniasm pipeline.**

The pipeline for the Miniasm assembler was run with the commands below (Li 2016).

```
minimap2 -t 32 -x ava-pb Allpass.fastq.gz Allpass.fastq.gz | gzip -1 >
ApCa_miniasm.paf.gz

miniasm -f Allpass.fastq.gz ApCa_miniasm.paf.gz > ApCa_miniasm.reads.gfa
```

2.9.4. **Pilon pipeline.**

The pipeline for the Pilon read polish was run with the commands below (Walker et al. 2014).

```
bwa index reads.racon2.fasta

bwa mem -t 32 reads.racon2.fasta shortreads/ApCa_1_trimmed.fq.gz
shortreads/ApCa_2_trimmed.fq.gz > Pilonmap1.bam

samtools sort Pilonmap1.bam -o Pilonmap1.sorted.bam

samtools index Pilonmap1.sorted.bam

java -jar pilon-1.23.jar --genome reads.racon2.fasta --frags Pilonmap1.sorted.bam
```

2.9.5. **Racon pipeline.**

The pipeline for the Racon error correct was run with the commands below (Vaser et al. 2017).

```
## Correction 1
minimap2 -t 64 ApCa_miniasm.fasta Allpass.fastq.gz > reads.gfa1.paf

racon -t 64 Allpass.fastq.gz reads.gfa1.paf ApCa_miniasm.fasta > reads.racon1.fasta

## Correction 2
minimap2 -t 64 reads.racon1.fasta Allpass.fastq.gz > reads.gfa2.paf
racon -t 64 Allpass.fastq.gz reads.gfa2.paf reads.racon1.fasta > reads.racon2.fasta
```

2.9.6. **Supernova pipeline.**

The pipeline for the supernova assembler was run with the commands below (Zheng et al. 2016).

```
supernova run --id=ApCa_Supernova1 --fastqs=ApCa-gDNA --sample=ApCa-gDNA --
maxreads=all –accept-extreme-coverage
```

2.9.7. **Trimmomatic pipeline.**

The pipeline for quality trimming DNA Trimmomatic was run with the commands below (Bolger et al. 2014).

```
ILLUMINACLIP=$ADAPTERPATH/TruSeq3-PE.fa

java  -jar $TRIMMOMATIC PE -threads 8 -phred33 \
                        ApCa_1.fq.gz ApCa_2.fq.gz \
                        ApCa_1_trimmed ApCa_1_unpaired \
                        ApCa_2_trimmed ApCa_2_unpaired \
                        ILLUMINACLIP:$ILLUMINACLIP:2:30:10 LEADING:3 TRAILING:3
SLIDINGWINDOW:4:15 MINLEN:36;
```

The pipeline for quality trimming RNA Trimmomatic was run with the commands below.

```
ILLUMINACLIP=$ADAPTERPATH/TruSeq3-PE.fa

java  -jar $TRIMMOMATIC PE -threads 8 -phred33 \
                        ApCa_1.fq.gz ApCa_2.fq.gz \
                        ApCa_1_trimmed ApCa_1_unpaired \
                        ApCa_2_trimmed ApCa_2_unpaired \
                        ILLUMINACLIP:$ILLUMINACLIP:2:30:10 LEADING:3 TRAILING:3
SLIDINGWINDOW:4:15 MINLEN:36;
```

2.9.8. **Wtdbg2 pipeline.**

The pipeline for the Wtdbg2 assembler was run with the commands below (Ruan and Li 2019).

```
Wtdbg2 -t 32 -i Allpass.fastq.gz -o ApCa_wtdbg -g 650m -x ont -S 1 --rescue-low-cov-
edges --tidy-reads 5000

Wtdbg-cns -t 32 -i -fo ApCa_wtdbg.ctg.fa

Minimap2 -t 32 -x map-pb -a ApCa_wtdbg.ctg.fa Allpass.fastq.gz | samtools view –Sb -
> ApCa_wtdbg.ctg.map.bam

Samtools sort ApCa_wtdbg.ctg.map.bam ApCa_wtdbg.ctg.map.srt

Samtools view ApCa_wtdbg.ctg.map.srt.bam | wtpoa-cns –t 32 –d ApCa_wtdbg.ctg.fa –i -
-fo ApCa_wtdbg.ctg.2.fa
```

2.9.9. L_RNA scaffolder script.

This script was used to scaffold genome contigs with the use of a pre-assembled transcriptome from the same species (Xue et al. 2013).

```
Blat ApCa_reads.racon2.fasta ApCa_rna_scaffold.fasta out/ApCa_rna_scaffold.psl -
noHead

lrna_scaffolder_dir/L_RNA_scaffolder.sh –d lrna_scaffolder_dir –i
ApCa_rna_scaffold.psl –j ApCa_reads.racon2.fasta
```

2.9.10. Nanochrome, LR_Gapcloser and SOAP Gapcloser pipeline.

This script was cycled 4 times and runs Nanochrome to scaffold genome contigs with reads from 10x Chromium before filling the gaps (Ns) created in the process with LR_Gapcloser and SOAP Gapcloser (Luo et al. 2015).

```
Run_Nanochrome.sh -g ApCa_L_RNAscaffolded.fasta -r 10xbarcodedreads.fastq  -n
CorrectedNanoporeReads.fasta -f 26848 -p Round1 -l 6 -t 78 –s

LR_Gapcloser.sh -i Round1.fa -l CorrectedNanoporeReads.fasta -t 78 -s n

SOAP/GapCloser -b soapAllConfig.txt -a LR_gapclosed.fasta -l 151 -t 20 -o
Out/ApCa_gapclosed.fasta 1>gaps.out 2>gaps.er
```

The soapAllConfig.txt file lists variables as below.

```
max_rd_len=151
rd_len_cutof=151
avg_ins=350
reverse_seq=0
asm_flags=3
map_len=32
q1=ApCa_ShortReadgDNA/ApCa_1_trimmed.fq.gz
q2=ApCa_ShortReadgDNA/ApCa_2_trimmed.fq.gz
```

2.9.11. BUSCO.

BUSCO was run via a Docker container vera/busco with the metazoan core genes library (Waterhouse et al. 2017).

```
Docker run –rm –name busco –v $(pwd):/data vera/busco –i ApCa_gapclosed.fasta –
results_busco –l metazoa_odb9 –m –c6
```

2.9.12. Blobplots.

Blobplots requires creation of a blast library and a coverage file (Laetsch and Blaxter 2017).

```
makeblastdb –in ApCa_transcriptome.fasta –taxid 302032 –dbtype nucl –out
ApCa_trans_blast –parse_seqids

blastn –task megablast –d ApCa_trans_blast –query ApCa_gapclosed.fasta –out
ApCa_trans_blastn.out –outfmt '6 qseqid staxids bitscore std' -evalue 1E-25 -
max_target_seqs 1 -max_hsps 1 -num_threads 32

bwa index ApCa_gapclosed.fasta

bwa mem ApCa_gapclosed.fasta ApCa_1_trimmed.fq.gz ApCa_2_trimmed.fq.gz > ApCa_SR.sam
–t 32

blobtools create –i ApCa_gapclosed.fasta –o blob_out/ApCa -s ApCa_SR.sam –t
ApCa_trans_blastn.out

blobtools blobplot –I blob_out/ApCa.blobDB.jason –o blob_out/ApCa
```

Additional scripts were required for running Blobplots2.

```
samtools view –S –b ApCa_SR.sam > ApCa_SR.bam

samtools sort ApCa_SR.bam –o ApCa_SR.sorted.bam

blobtools create --fasta ApCa_gapclosed.fasta --cov ApCa_SR.bam  --hits
ApCa_trans_blastn.out --taxdump taxdump/  tmp/dataset_1
```

2.9.13. STAR, Trinity Genome Guided assembly and Evigene.

The Trinity Genome guided assembly pipeline was run followed by Evigene. It required STAR to generate a bam library (Dobin et al. 2013).

```
STAR --runThreadN 24 --runMode genomeGenerate --genomeDir genomeindex --
genomeFastaFiles ApCa ApCa_gapclosed.fasta

STAR--runThreadN 24 --readFilesCommand gunzip -c --genomeDir genomeindex –readFilesIn
ApCa_RNA_1.fastq.gz ApCa_RNA_2.fastq.gz

samtools view -S -b Aligned.out.sam > Aligned.out.bam

samtools sort Aligned.out.bam -o Aligned.sorted.bam
```

```
Trinity --genome_guided_bam Aligned.sorted.bam --genome_guided_max_intron 10000 --CPU
24 --max_memory 100G --normalize_reads --output Trinity-GG.fasta --full_cleanup

Trformat.pl Trinty-GG.fasta > Evigene_merge.fasta

Tra2aacds.pl –mrnaseq Evigene_merge.fasta –MINCDS=60 –NCPU=24 –MAXMEM=200000 –logfile
-tidyup
```

2.9.14. **PASA.**

PASA annotation tool was run with the following command (Haas et al. 2003).

```
Launch_PASA_pipleine.pl –c alignAssembly.config –C –R –ALIGNER gmap –g
ApCa_gapclosed.fasta –t Trinity-GG.fasta –CPU 8
```

The align Assembly.config file lists variables as below.

```
## templated variables to be replaced exist as <__var_name__>
# Pathname of an SQLite database
# If the environment variable DSN_DRIVER=mysql then it is the name of a MySQL
database
DATABASE=/data/PASA/ApCa/ApCa_mydb_pasa.sqlite
#######################################################
# Parameters to specify to specific scripts in pipeline
# create a key = "script_name" + ":" + "parameter"
# assign a value as done above.

#script validate_alignments_in_db.dbi
validate_alignments_in_db.dbi:--MIN_PERCENT_ALIGNED=70
validate_alignments_in_db.dbi:--MIN_AVG_PER_ID=85
validate_alignments_in_db.dbi:--NUM_BP_PERFECT_SPLICE_BOUNDARY=0

#script subcluster_builder.dbi
subcluster_builder.dbi:-m=50
```

2.9.15. **Transdecoder.**

Transdecoder pipeline below used the GTF generated from PASA (Haas et al. 2013).

```
Util/gtf_genome_to_cdna_fasta.pl ApCa_transcripts.gtf ApCa_gapclosed.fasta >
ApCa_transcripts.fasta

Util/gtf_to_alignment_gff3.pl ApCa_transcripts.gtf > ApCa_transcripts.gff3

TransDecoder.LongOrfs –t ApCa_transcripts.fasta
TransDecoder.Predict –t ApCa_transcripts.fasta
```

```
Util/cdna_alignment_orf_to_genome_orf.pl ApCa_transcripts.fasta.transdecoder.gff3
ApCa_transcripts.gff3 ApCa_transcripts.fasta >
ApCa_transcripts.fasta.transdecoder.genome.gff3
```

2.9.16. OmicsBox.

OmicsBox generates a GFF2 file as an alternative to PASA utilizing Augustus however to fully convert to a GTF it requires the addition of ‘transcript_id’ to function in STAR. OmicsBox is run with a masked Genome and a BAM file of all RNA-seq data mapped to the masked Genome. The program is automated and takes roughly 3 days to complete. To convert the GFF2 to GTF the following command was run (Biobam Bioinformatics 2019).

```
sed -e 's/$/ transcript_id "Parent";/' -i ApCa_Genome_070519_masked.gtf
```

2.9.17. STAR RNA mapping.

RNA mapping of quality trimmed RNA-seq reads were mapped with STAR with the commands below and converted to count tables for SARTools and DESeq2 analysis.

```
STAR --runMode genomeGenerate --genomeDir ${genome_dir} –genomeFastaFiles
"${genome_dir}/ ApCa_Genome_070519_masked.fasta" --sjdbGTFfile "${genome_dir}/
ApCa_Genome_070519_masked.fasta.gtf"  --limitGenomeGenerateRAM 128000000000 --
runThreadN 32

STAR --genomeDir "$genome_dir" --readFilesIn "$read_dir/Sample${i}_trimmed.fastq.gz"
--readFilesCommand gunzip -c --runThreadN 32 --sjdbGTFfile "$map_dir/
ApCa_Genome_070519_masked.fasta.gtf" --outSAMtype BAM SortedByCoordinate --quantMode
TranscriptomeSAM GeneCounts --outFileNamePrefix "$map_dir/Sample${i}_map"


declare -a output=(\
“sample_x”
“sample_y”
“sample_z”
)
for (( i=0 ; i<${#output[@]} ; i++ ));
do
cut -f1,2 "${map_dir}/${output[${i}]}_mapReadsPerGene.out.tab" >
"${map_dir}/${output[${i}]}_mapReadsPerGene1-2.tab"
sed -i 1,4d "${map_dir}/${output[${i}]}_mapReadsPerGene1-2.tab"
done
```

2.9.18. SARTools and DESeq2 script

The script for analysing RNA count data in Rstudio is shown below (Varet *et al.* 2016).

```
library(SARTools)
if (forceCairoGraph) options(bitmapType="cairo")

# checking parameters
checkParameters.DESeq2(projectName=projectName,author=author,targetFile=targetFileraw
          Dir=rawDir,featuresToRemove=featuresToRemove,varInt=varInt,condRef=condRef,
          batch=batch,fitType=fitType,cooksCutoff=cooksCutoff,independentFiltering=in
          dependentFiltering,alpha=alpha,pAdjustMethod=pAdjustMethod,typeTrans=typeTr
          ans,locfunc=locfunc,colors=colors)

# loading target file
target <- loadTargetFile(targetFile=targetFile,varInt=varInt, condRef=condRef,
          batch=batch)

# loading counts
counts <- loadCountData(target=target, rawDir=rawDir,
          featuresToRemove=featuresToRemove)

# description plots
majSequences <- descriptionPlots(counts=counts, group=target[,varInt], col=colors)

# analysis with DESeq2
out.DESeq2 <- run.DESeq2(counts=counts,target=target,varInt=varInt,batch=batch,
          locfunc=locfunc,fitType=fitType,pAdjustMethod=pAdjustMethod,cooksCutoff=coo
          ksCutoff,independentFiltering=independentFiltering, alpha=alpha)

# PCA + clustering
exploreCounts(object=out.DESeq2$dds, group=target[,varInt], typeTrans=typeTrans,
          col=colors)

# summary analysis (boxplots, disp, diag size fact, nDiffTotal, histograms, MA plot)
summaryResults <- summarizeResults.DESeq2(out.DESeq2, group=target[,varInt],
          col=colors,independentFiltering=independentFiltering,cooksCutoff=cooksCutof
          f, alpha=alpha)

# generating HTML report
writeReport.DESeq2(target=target, counts=counts,out.DESeq2=out.DESeq2,summaryResults=
          summaryResults,majSequences=majSequences,workDir=workDir,projectName=projec
          tName,author=author,targetFile=targetFile,rawDir=rawDir,featuresToRemove=fe
          aturesToRemove,varInt=varInt,condRef=condRef,batch=batch,fitType=fitType,co
          oksCutoff=cooksCutoff,independentFiltering=independentFiltering,alpha=alpha
          ,pAdjustMethod=pAdjustMethod,typeTrans=typeTrans,locfunc=locfunc,colors=col
          ors)
```

2.9.19. **GATK pipeline.**

GATK is a long and variable pathway. Below is the framework of scripts used during this analysis.

Genome index generate

```
STAR --runMode genomeGenerate --genomeDir ${genome_dir} --genomeFastaFiles
${genome_dir}/${genome}" --sjdbGTFfile "${genome_dir}/${gtf}"
--limitGenomeGenerateRAM 128000000000 --runThreadN \$SLURM_NTASKS
```

Genome dictionary generate

```
gatk CreateSequenceDictionary -R "${genome}"

samtools faidx "${genome}"
```

RNA read mapping

```
STAR --genomeDir "$genome_dir" --readFilesIn "$read_dir/\${i}_$trimmed"
--readFilesCommand gunzip -c --runThreadN \$SLURM_NTASKS --sjdbGTFfile
"$map_dir/$gtf" --outSAMtype BAM SortedByCoordinate --twopassMode Basic --quantMode
TranscriptomeSAM GeneCounts --outFileNamePrefix "$map_dir/GATK/\${i}_map"
```

Marking duplicates and SplitNcigar

```
java -jar \$PICARD MarkDuplicates I=GATK/\${i}_mapAligned.sortedByCoord.out.bam
O=GATK/\${i}_GATKmarked.bam M=GATK/\${i}_GATKmarked_metrics.txt
REMOVE_DUPLICATES=false VALIDATION_STRINGENCY=SILENT

gatk SplitNCigarReads -R "${genome}" -I GATK/\${i}_GATKmarked.bam -O
GATK/\${i}_GATKsplit.bam
```

Adding read groups

```
java -jar \$PICARD AddOrReplaceReadGroups I=GATK/\${i}_GATKsplit.bam
O=GATK/\${i}_GATKrg.bam SORT_ORDER=coordinate RGID=1 RGLB=\${i}lib RGPL=illumina
RGSM=\${i} RGPU=1 CREATE_INDEX=TRUE
```

1st round of haplotype calling

```
gatk --java-options "-Xmx4g" HaplotypeCaller -R "${genome}" -I GATK/\${i}_GATKrg.bam
-O  GATK/\${i}_GATK.g.vcf -ERC GVCF
```

Merging g.vcfs

```
gatk CombineGVCFs -R "${genome}" --variant GATK/${sample1}_GATK.g.vcf --variant
GATK/${sample2}_GATK.g.vcf Etc…        -O GATK/cohort.g.vcf.gz
```

Convert g.vcfs to vcf

```
gatk --java-options "-Xmx4g" GenotypeGVCFs -R "${genome}" -V GATK/cohort.g.vcf.gz
-O GATK/cohort.vcf.gz
```

Filtering was done via reverse grep

2.9.20. Fst and Diversity analysis.

Fst, Tajima’s D and nuclear diversity was calculated with the code below using VCFtools.

```
vcftools --mac 8 --vcf HAandLAfiltered.vcf --weir-fst-pop HAlist.pop --weir-fst-pop
LAlist.pop --out HAvsLA_mac8

vcftools --vcf slicedLA.vcf --TajimaD 30000 --out TajimaD-LA-vcf

vcftools --vcf slicedLA.vcf --site-pi --positions SNP_list.txt --out PI-LA-vcf
```

3. Species diversity of temperate altitudes.

3.1. Investigating altitudinal transects of interest.

3.1.1. Specification of Site.

Investigating adaption and acclimatisation to high altitude stressors requires more than a singular high altitude population. Indeed, it is key to compare how a high altitude population compares with a low altitude population. With two populations we can investigate gene regulation and gene selection directly associated with high altitude. With populations that are closely linked, geographically and phylogenetically, we can more confidently attribute changes observed in these comparisons with the change in altitude rather than the gradual accumulation of random changes associated separation by distance. An ideal test subject would therefore be ubiquitously found over a steep geographical transect and closely linked through mitochondrial lineage.

In looking to investigate altitudinal adaption, sites not only require presence of earthworms at high altitude but also at a low altitude site at close-proximity containing an equivalent earthworm population. This excluded a number of mountainous regions where although the grass line is at high altitude, there is no ‘close by’ low altitude use as a comparison site. Sites investigated include several temperate mountain ranges from Mount Kinabalu in Malaysia, Col de L’Iseran – the highest mountain pass in France, Les Deux Alpes in France, Pic de Coma Pedrosa in Andorra, and Mount Pico in the Azores. Mount Pico and Les Deux Alpes were selected for their high altitude gradient and to provide two contrasting sites to investigate for biodiversity and suitability to investigate high altitude adaption and acclimatization.

Both the Azores and Les Deux Alpes were identified, as locations that had a high vertical range over a short distance. Previous research trips to the Azores had identified the presence of worms at high altitude indicating its suitability for testing and provided for a distinct comparison to the continental alpine sampling site of Les Deux Alpes. The two sites represent contrasting biological flora and fauna origins, Pico as a young island requiring seeding of flora and fauna, while Les Deux Alpes able to reseed by the multiple glacial refugia immediately following the recession of the last glacial maximum. It was also advantageous to identify species that do not contain cryptic speciation that could impact the complexity of experimental and analytical of subsequent chapters.

3.1. Selection of temperate altitudinal transects.

3.1.1. **The Azores.**

The archipelago of The Azores is a group of nine islands, roughly 1600 km West of mainland Portugal between 37° and 40° latitude. The two largest islands are São Miguel, where the regional capital Ponta Delgada is situated and Pico, which has the islands' highest peak of 2,351 m (Figure 2).

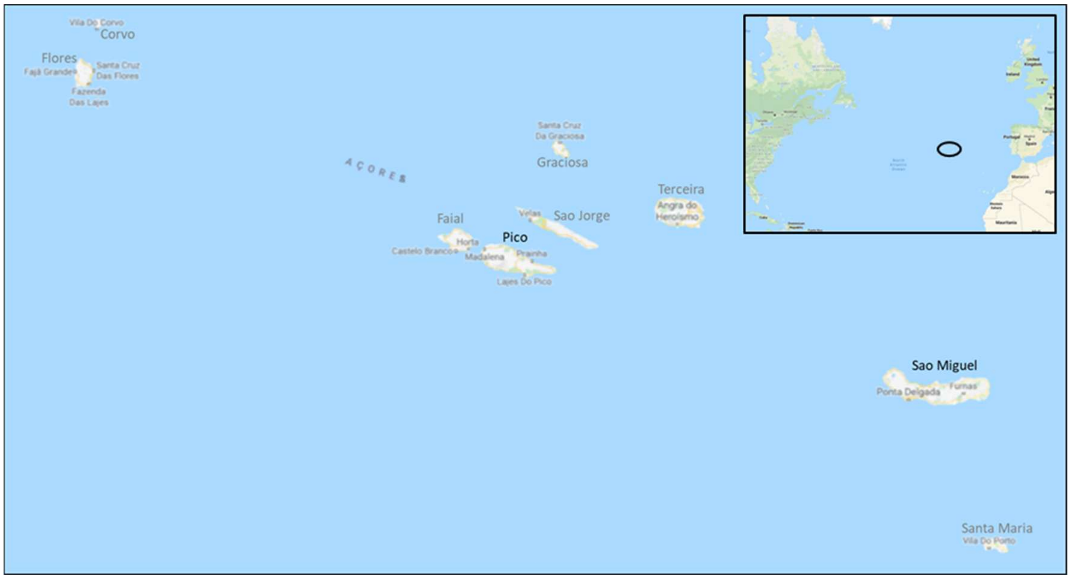


Figure 2: The archipelagos of the Azores and their location in the Atlantic Ocean.

The islands, located on the triple junction of the North American Plate, the Eurasian Plate and the African Plate first began to surface the Atlantic Ocean as early as 8.12 million years ago with Santa Maria. São Miguel emerged 4.1 million years ago and Pico the youngest of the islands, 270 thousand years ago (Miranda et al. 1998; Carine and Schaefer 2010). Each of the islands are known for their own distinctive geomorphology such as the large craters and cones of São Miguel, the distinctive volcanic peak of Pico or the oldest island of Santa Maria having brown sandy beaches, rather than the black sands found more commonly. This is in part a result of their age, but also how their formation was influenced by their position on the plate boundaries.

Islands are more than just volcanic rock and ocean. The islands' biodiversity is sizable with over 8000 recorded species of terrestrial and marine biota (Borges et al. 2010). Much of the terra-formation of the islands by plants has occurred through seed dispersal in oceanic currents and a some through endozoochory (Heleno and Vargas 2015). Although the exact process of island greenification in the Azores has not yet been fully described, it is interesting to compare its biodiversity with that of the Ascension Island. The island described by Darwin as "hideous" had only one tree at his time of visit, which has since undergone deliberate introduction of species to terraform the island. One can draw parallels with the Azores as only 411 species are endemic to the islands, a mere ~5% to the estimate total (Borges et al. 2010).

The Anthropocene is the name given to the start of the new epoch of time that commenced with the initial domestication of plants and animals by humans, circa 12,000 years ago[1] (Chapin et al. 2000; Smith and Zeder 2013). Humans not only expanded quickly across new lands, but they brought with them plants and livestock. The introduction of new invasive species to each environment included earthworms which can hide in soil clumps of roots, but also those deliberately introduced as an aid to increase soil fertility (Bennett and Prance 2000; Hendrix et al. 2008).

Despite the stories of mid-Atlantic islands, such as the legend of Atlantis written by Plato, the autonomous islands of Portugal were not mapped until the mid-14th century and formally identified in the early 15th (Ashe 1813). Colonisation began with many arriving from Portugal to exploit the rich volcanic soils, with other European countries including France arriving in later centuries. The colonisation including the introduction of cattle and farmed crops and other invasive species of flora and fauna are likely to have included stow-away earthworms. While the majority of immigration to the Azores has been from Europe, the islands did see some settlement of African slaves (Pacheoco et al. 2010). More recently evidence has been presented suggesting the islands of the Azores could have been visited as early as the 11th century based on “man-made” structures identified and dated through Accelerator Mass Spectrometry (Rodrigues et al. 2015).

Factoring in the archipelago geographical positioning, one would expect the earthworms’ species to be invasive to the islands, originating from Europe but possibly including species from Africa. To date, no species has been identified as endemic to the islands and very few areas of untouched primary laurissilva forest. There has not been a full biodiversity study of earthworms across the Azores, however worms from the families *Lumbricidae*, *Megascolecidae* and even *Rhinodrilidae* have been reported (Amaral et al. 2006; Cunha et al. 2011; Cunha et al. 2014; Novo et al. 2015). Both *Lumbricidae* and *Megascolecidae* are reported across the Palearctic as a mixture of indigenous and introduced taxa, while *Rhinodrilidae* is reported as endemic to the Neotropical and intrusive to all regions except the Palearctic (Hendrix et al. 2008). The presence

[1] The start of the Anthropocene is widely disputed with some arguing for the commencement of modern farming, others advocating the industrial revolution and some arguing the first nuclear bomb test as the definitive start. For the purpose of this thesis, we will be using the first definition as it defines the start of large-scale human influence on the environment.

of the genus *Pontoscolex* would therefore also suggest migration, possibly human mediated from South America to the islands.

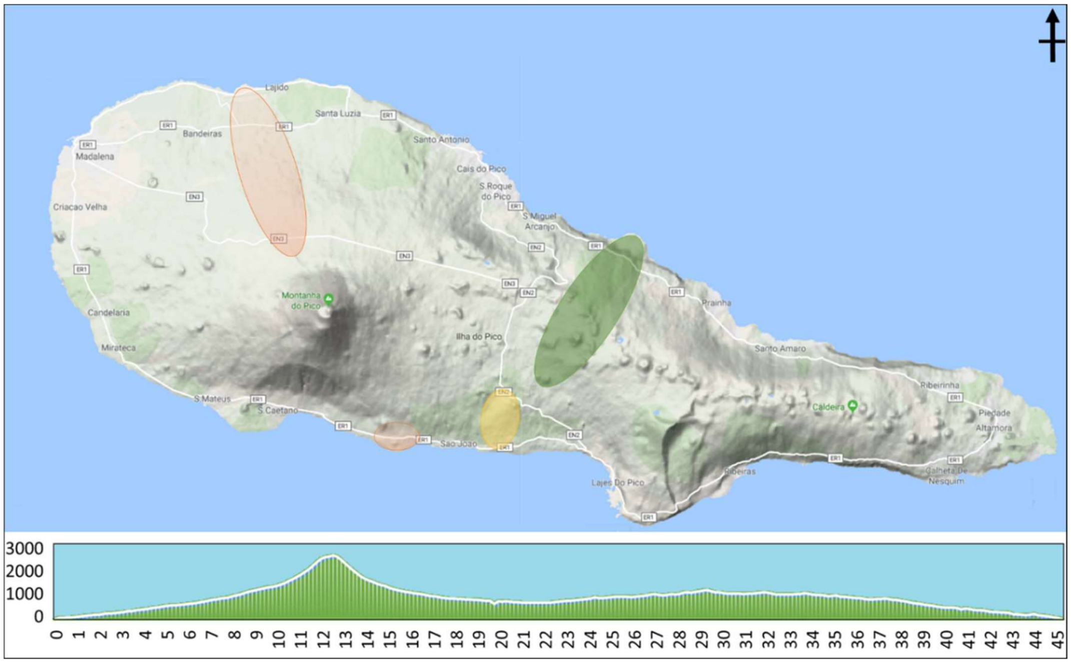


Figure 3: Pico Island, its latitudinal cross section and its recent volcanic activity. The 1562-64 eruption (circled in green), the 1718 eruption (circled in orange) and the 1720 eruption (circled in yellow) (Woodhall 1974).

This research will focus on the island of Pico with its classical strato-volcanic peak, which provides a steep altitudinal transect. Though the island was formed about 270,000 years ago with the main peak, there have been several occasions of volcanic activity on Pico in the last 500 years, (as seen in Figure 3), whose eruptions would wipe out existing earthworm populations and provide opportunity for neighbouring populations to expand into (Woodhall 1974). In particular the large 1718 likely originated from the central cone as there are no side vents suitable large enough. This raises the likelihood of extinction of any earthworm population living around the central cone and any earthworm found would be a recolonisation from low altitude. The Island can be segregated into two distinct meteorological regions. The main volcanic peak and ridge 800 m above sea level, which sees substantially greater rainfall and colder temperatures than the rest of the island below 800 m sea level (Azevedo et al. 1998; Meteoblue 2016a). Only three areas of preserved land remain on the Island: the main peak, the land covering the 1562-64 eruption 400 m above sea level and small patch located along the main ridge of the island, north of Lajes Do Pico. Although much of the island has not been built upon, much of the land has been used for farming of livestock and agriculture (Figure 4) (Moreira 2013; Fernandes et al. 2014). This includes a high proportion of small field vineyards. Much of the land, which has not been used for pasture, remains quite rocky with a thin layer of topsoil.

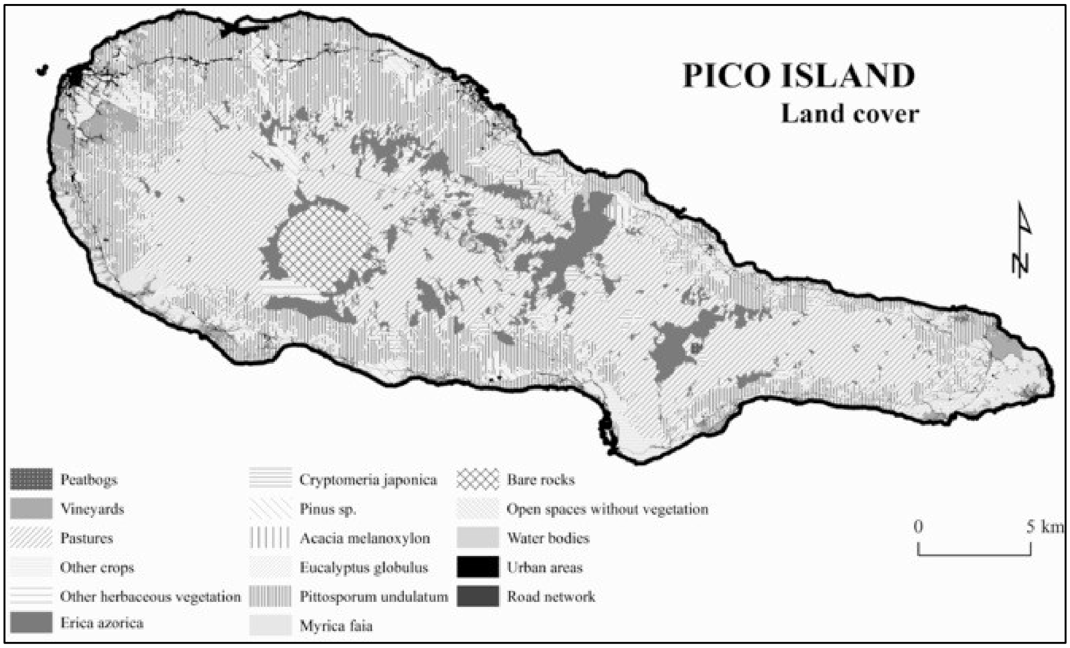


Figure 4: Land use map (2008) of Pico Island (Fernandes et al. 2014).

This island of Pico provides a fantastic 'natural laboratory' to test population diversity and structure due to its young geological and ecological age and diverse flora and contrasting environmental conditions.

3.1.2. **The French Alps.**

Centrally located in the French Alps, Les Deux Alpes is a Ski resort, 133 km due South of Geneva, situated 45° latitude. The 210,000 km^2 region is characterised with high mountains and urbanised valleys, forming part of the largest mountainous region in Europe that is part of the Alpine belt stretches from the Atlantic to the Himalayas.

The mountainous region of the French Alps was formed during the late Mesozoic at the same time as the Pyrenees in the West and the Caucasus in the East, when the African plate collided with the Eurasian plate about 65 million years ago (Figure 5) (Moores and Fairbridge 1998). Despite this age, much of the structure and shape of the Alps seen today was formed in the last 2.5 million years during the Pliocene-quaternary glaciation ice age with the last major ice coverage of the Alps (Würm glaciation) ending 10,000 years ago at the onset of the Holocene (Sibrava et al. 1986).

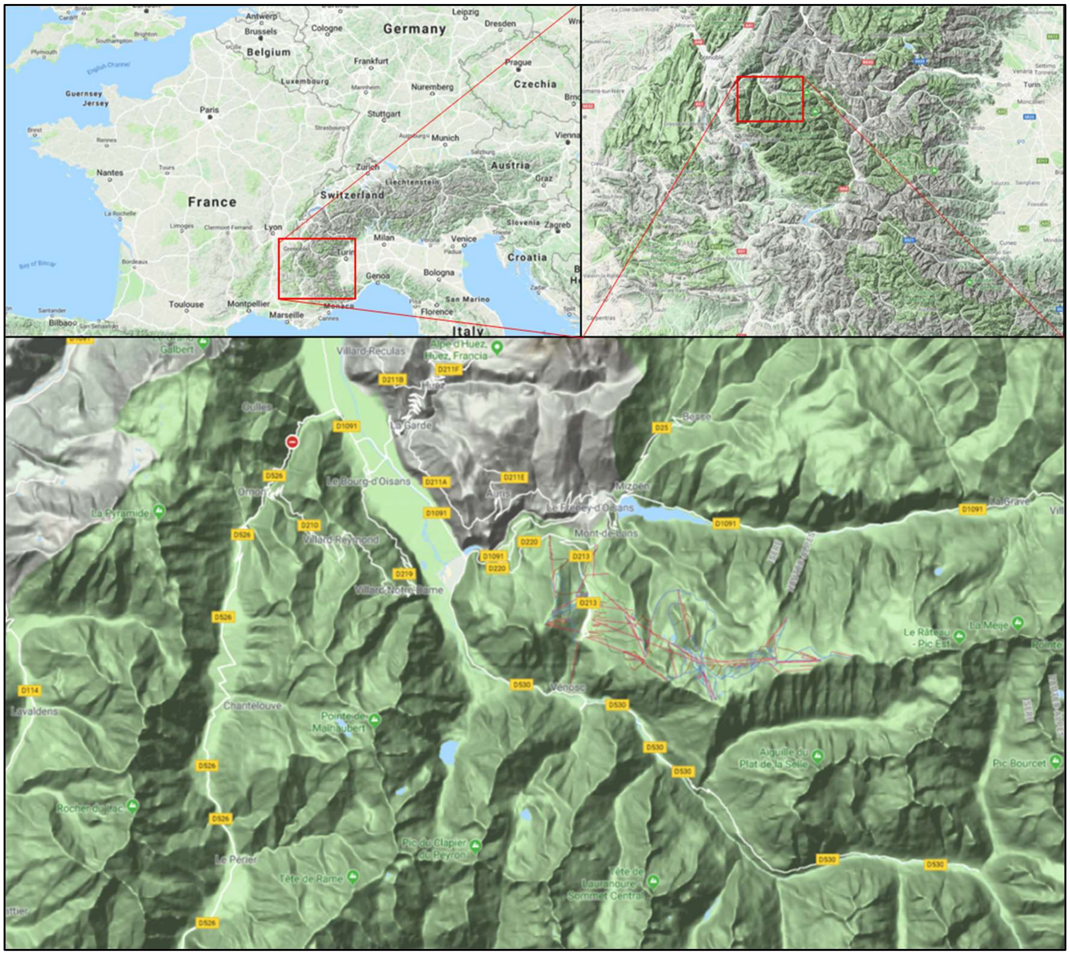


Figure 5: Les Deux Alpes and its location in Europe.

Glacial refugia played a critical role following the recession of Ice, with the flora and fauna of the Alps able to return to colonising the slopes, creating a diverse gradient of sub-alpine, alpine and glacial environments (Klutsch et al. 2012). This expansion would have initiated from low lying sheltered valley floors as part of glacial refugia and followed the U-shaped valleys as ice receded. Similar patterns of returning biodiversity have been characterised in North American and Tibetan mountain ranges as well as those from Europe in mammalian, freshwater invertebrates families and hexapods (Klutsch et al. 2012; Clewing et al. 2016; von Saltzwedel et al. 2016). Glacial losses have accelerated in the last 100 years in-line with increasing Global-mean temperatures induced by man-made climate change (Bauder et al. 2017). The latest report of the glaciers in neighbouring Swiss Alps indicate that the majority of the 114 glaciers observed are experiencing loss between 0 and -30 m per year (Bauder et al. 2017). While earthworms do not tend to migrate distances greater than 10 m per annum and the glacial retreat would exceed this migration, the spread of vascular plants have been able to exploit the expanding land through 'wind-dispersal ability' (Parolo and Rossi 2008).

Human influence of the Alpine environment goes back much further than that seen in the Azores. Mitochondrial analysis indicates their presence in the Palaeolithic, with remains dated as far back as 14,000 years ago (Benedetto et al. 2000). Following the recession of ice of the Pliocene-quaternary glaciation, humans began to colonise the Alps. Ötzi, Europe's oldest preserved mummy has been dated as old as 3400 years and indicates a change of lifestyle from 'hunter gatherer' to agriculture and trade (Bonani et al. 1994).

The Alps have been heavily travelled by humans in the last ~2000 years indicated by Roman historical accounts, including the more notable transport of war elephants across the Alps by Hannibal in 218BC on their route to Rome (Dodge 1994). Humans have inhabited the Alpine valleys heavily ever since and undergone significant global through-travel and development. It is therefore expected that earthworm species diversity will be greater than that found in the Azores (Hale 2008; Dobson and Satchell 2020).

Scientific reporting of earthworm species from the Alps remains quite light, with only a few publications that provide information on diversity (Bienert et al. 2012; Steinwandter et al. 2017). Binert *et al.* used 16s barcoding to identify species found on the edge of the French Alps identifying several earthworm species of the *Lumbricidae* family. They included worms in the *Aporrectodea, Lumbricus* and *Octolasion* genus, also reporting *Allobophora chlorotica* (Bienert et al. 2012). Several other widespread European species are speculatively present although not formally reported, including *Dendrodrilus rubidus* and *Eiseniella tetraedra*.

Les Deux Alpes is a large French ski resort at 1700m spread across two sides of a high-altitude valley. The eastern valley side peaks at a maximum 2300 m while the western side follows a mountain ridge above 3500 m. Between 3100 and the summit is a large glacier that would have at one stage, flowed into the location of today's town centre. Few trees exist above 2000 m and small vegetation is space above 2400 m. From the town, land descends sharply in the south to the town of Vensoc at 1000 m and in the north to Lac du Chambon at 1000 m. Snow cover in the area usually commences in November and continues until April, with the snow season lasting longer at higher altitudes and shorter at the lower altitudes (Meteoblue 2016c). Shallower slopes have provided opportunities for pasture and seen a thicker top-soil develop, while steeper slopes tend to have rockier and dryer environments. Forested areas below 2000 m have a thick pine humus layer with acidic soils.

Les Deux Alpes provides a contrasting environment to Pico, with more changeable seasonal conditions and a greater human influence. However, the site also presents a more challenging sampling environment as the multitude of surrounding glacial refugia has led to an increased general biodiversity. Seasonal timing of sampling only offers a very narrow window, collecting specimens at low altitude before it gets too warm and dry, but late enough that earthworms have had time to emerge from the thawed slopes at high altitude.

3.1.3. **Aims and objectives.**

This chapter aims to study species diversity and haplotype structure of the earthworms found on Pico in the Azores and Les Deux Alpes in the French Alps. This will allow the assessment of their population structure in global context of lineage and environmental conditions in following chapters. Earthworms will be collected from Pico in the Azores both characterising populations circumventing the Islands parameters as well as a transect from near sea level to the crater of the main volcanic peak (~2300 m above sea level). Similarly, earthworms will be sampled from Les Deux Alpes representing a transect from Lac du Chambon (877 m Above sea level) to the edge of the grass-line (~2200 m above sea level). Samples will be genetically barcoded and phylogenetic analysis performed to determine haplotype structure of the dominant species.

3.2. Methodology.

3.2.1. **Sampling sites of Pico.**

Sampling was conducted over two field seasons in January 2015 and January 2017. Sampling sites of the transect (1-14) were selected to provide a range of altitudes, while sampling sites from around the island (A-G) provided a range of environmental conditions (Shown in Table 1 and Figure 6).

Table 1: Sampling locations and conditions for Pico [2].

Site	*Westings*	*Northings*	*Elevation (m)*	*Soil depth (cm)*	*weather*	*Vegetation and notable conditions*
Summit	-	-	2351	0	-	-
13	28.3993	38.4671	2261	1.5	Heavy rain	Bare rock, lichen, moss, patchy gravel soil and ice patches
14	28.4076	38.4657	2055	1.5	Heavy rain	Bare rock, lichen, moss and patchy gravel soil
5	28.4162	38.4661	1622	2.5	Misty	Grass and moss
4	28.4186	38.4659	1517	2.5	Misty	Grass and moss
3	28.4216	38.4698	1400	5	Misty	Grass and moss
2	28.4247	38.4701	1288	10	Misty	Grass and moss
1	28.4258	38.4714	1237	10	Misty	Grass and moss
6	28.4263	38.4721	1222	10	Misty	Grass and moss
7	28.4325	38.4669	1128	10	Misty	Grass and moss
8	28.4325	38.4664	1120	20	Misty	Ferns, moss, bracken and pine shrubs
9	28.4197	38.4940	908	10	Misty	Grass and moss
10	28.4044	38.5096	690	7	Clear	Grass, moss and pine shrubs
11	28.4345	38.5203	431	1.5	Cloudy	Grass, moss, cedar, laurel and spruce
12	28.4215	38.5382	167	1.5	Clear	Grass, moss, pine and laurel
B:1	28.3983	38.5553	35	1.5	Light rain	Dense woodland
B:2	28.3983	38.5553	35	1.5	Light rain	Rocky field
C	28.3976	38.5555	33	7.5	Rain	Dense woodland
D	28.2527	38.4909	231	3.5	Light rain	Dense woodland
E	28.0490	38.4178	148	8	Light rain	Dense woodland
F	28.1313	38.4267	702	10	Heavy rain	Grass and saturated soil
G	28.4369	38.4335	138	8	Light rain	Agriculture field

[2] 'Site A' not shown as no worms were found at this site.

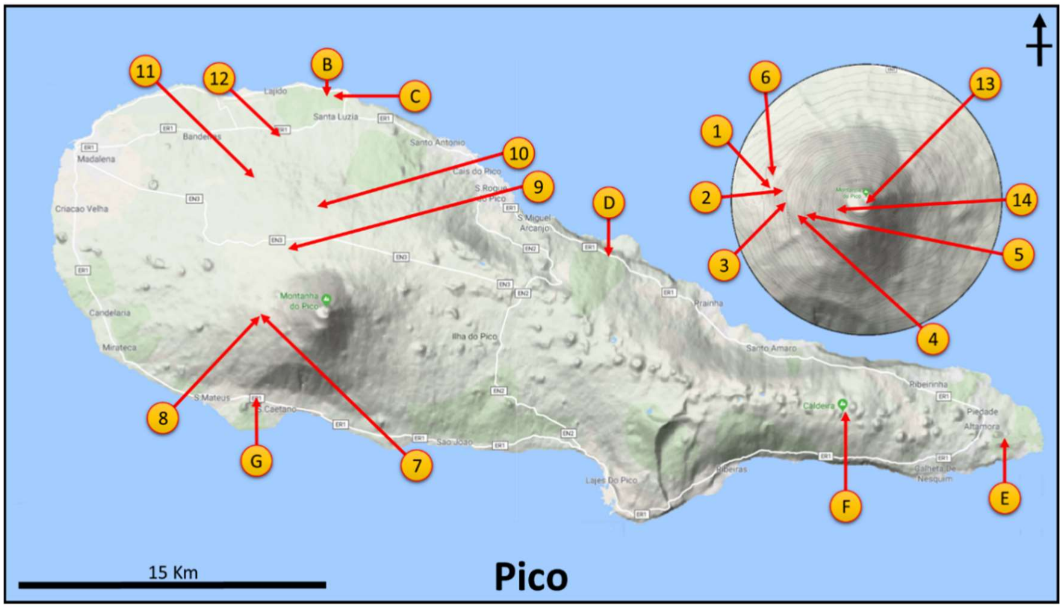


Figure 6: Sampling sites of Pico and their relative positions on the island. No worms were collected at site A, (not displayed). The insert is an enlargement of Pico's main peak.

3.2.2. **Sampling sites of Les Deux Alpes.**

Sampling was conducted during April 2017. Sampling sites of the transect (1-14) were selected to provide a range of altitudes and encompassed a variety of environmental conditions from near frozen soil to warm and very dry soil (Shown in Table 2 and Figure 7).

Table 2: Sampling locations and conditions for Les Deux Alpes [3]

Site	*Westings*	*Northings*	*Elevation (m)*	*Soil depth (cm)*	*weather*	*Land use description*
1	45.0020	6.12818	1674	5	Sun	Grass, some shrubs and by a river
2	45.0026	6.13308	1797	10	Sun	Grass, some shrubs and by a river
3	45.0018	6.13748	1970	2	Sun	Grass, some shrubs and by a river
4	45.0013	6.13740	1969	5	Sun	Coniferous forest and grass patches
5	45.0016	6.13843	2008	10	Sun	Grass, some shrubs and by river source
7	45.0093	6.14110	2142	10	Sun	Rotting grass, mud, lots of snow and ice around
8	45.0242	6.13333	1818	10	Sun	Grass
9	45.0246	6.12008	1601	10	Sun	Grass, silver birch, damp soil, bluebells and by a stream
10	45.0287	6.12245	1406	5	Sun	Deciduous wood, some patchy grass, very dry but some wet patches
11	45.0318	6.12489	1299	1	Sun	Gravel like soil and thick moss
12	45.0375	6.13646	1140	5	Sun	Dry gravel like soil and some grass
14	45.0188	6.11947	1740	10	Cloud	Grass and by a river

[3] Sites 6 and 13 not shown as no worms were found at these sites.

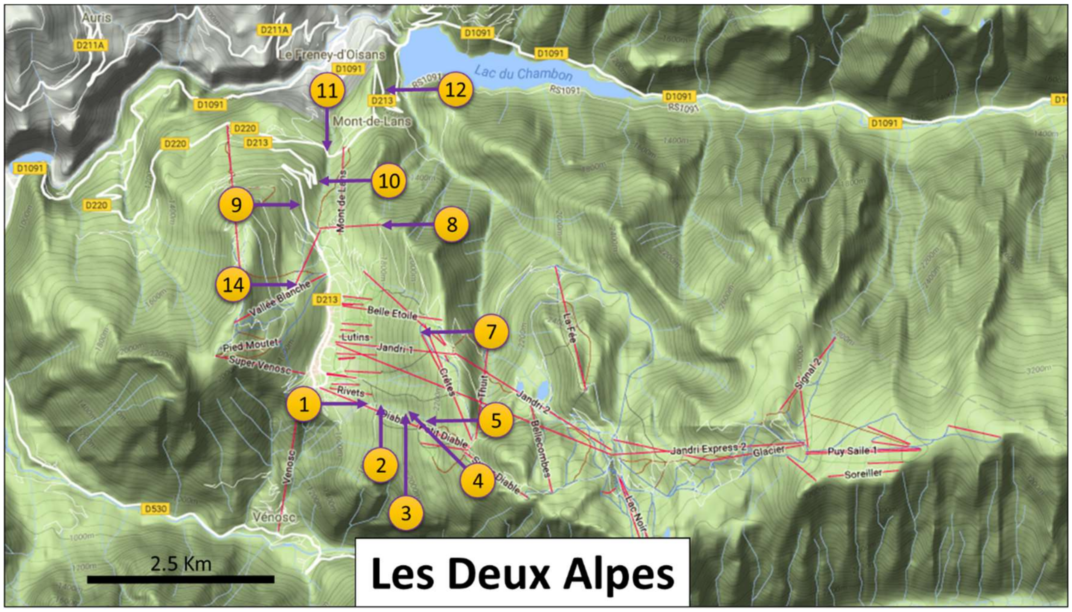


Figure 7: Sampling sites of Les Deux Alpes and their relative positions around the resort. No worms were collected at sites 6 and 13, (not displayed).

3.2.3. **Sample collection, DNA extraction, PCR and sequencing.**

Earthworms were collected from Pico and Les Deux Alpes and returned to Cardiff as described in general methods (2.1.1). Sampling was performed by excavation and hand sorting so that high quality tissue could be isolated from excavated samples, however, this processes does not yield quantitative annelid biodiversity data therefore limiting our analysis to occupancy and diversity. DNA was extracted from preserved tissue and underwent PCR amplification of the COII gene prior to genetic barcode sequencing of the fragment's forward strand (2.2 and 2.3). Samples that did not sequence were excluded from genetic analyses. Initial attempts to apply the COI loci using the 'universal' Folmer primers gave unreliable and stochastic amplification (Folmer *et. al.* 1994). Significant work to optimise PCR conditions could not rectify the unreliability of amplification while in contrast the COII amplification worked robustly in all amplifications and was therefore selected for our analysis. Pico sites B-G were sampled for *A. caliginosa* specifically.

3.2.4. **Pico genetic data analyses**

Sequence homology was determined with NCBI's non-redundant nucleotide data-base (nr/nt) using a discontinuous nucleotide Megablast and provisional species identity assigned using the match displaying the lowest E-value. Subsequently, all samples sequences (212) for each sampling site were aligned using ClustalW (Mega7 (7.0.26)) using default settings and trimmed to the same length (619 bp) (Larkin et al. 2007; Kumar et al. 2016). Sequences that were considerably shorter than the average (<600 bp) were excluded from analysis. A best-fit substitution model was calculated to be TN93 +G +I (Mega7). This model was used to calculate a Maximum Likelihood tree with 100 bootstrap replicates and was graphically represented using FigTree v1.4.3 (Drummond et al. 2012).

Due to the high species abundance observed during sampling, further analyses of *Aporrectodea caliginosa* collected on Pico included sequences from São Miguel collected at the same time as those reported in detail in in Appendix Chapter 10.1 (Novo et al. 2015). The number of segregating sites (Theta-W) and Nucleotide diversity (Pi) was calculated using DnaSP's DNA Polymorphism calculator while Tajima's D was calculated using the Tajima's test calculator (Rozas et al. 2017).

Haplotypes groups were assigned, and haplotype diversity was calculated for all the samples from Pico for *A. caliginosa* (including the São Miguel samples) and *L. terrestris* as the they were the two most abundant species collected. These were calculated using DnaSP's Haplotype data file generator (Rozas et al. 2017). These groups were then used for each individual species to generate Minimum spanning network trees in PopART 1.7 (Bandelt et al. 1999; Leigh and Bryant 2015).

3.2.1. **Alps genetic data analyses**

Analysis of the earthworms collected from the Alps followed the same analysis pathway as those from Pico (3.2.4). Best match homology within NR was initial determined using in order to provide provisional species assignment. All sequences were then aligned using ClustalW (v1.2) and trimmed to the same length (684bp). Sequences that were considerably shorter than the average (<600 bp) were excluded. A best-fit substitution model was calculated to be GTR +G +I and this model was used to calculate a Maximum Likelihood tree with 100 bootstrap replicates. Theta-W, Pi and Tajima's D for *Lumbricus terrestris* was calculated as described for *A. caliginosa* (3.2.4).

Haplotypes groups were assigned, and the haplotype diversity calculated for all the samples from the Alps for *A. caliginosa* and *L. terrestris* to compare and contrast with those identified in Pico, and these were then used for each individual species to generate Minimum spanning network trees generated in PopART. The Haplotype groups created for *L. terrestris* included *L. castaneus* as the Maximum Likelihood tree generated suggested a similar distance between lineages of *L. terrestris* and *L. castaneus*.

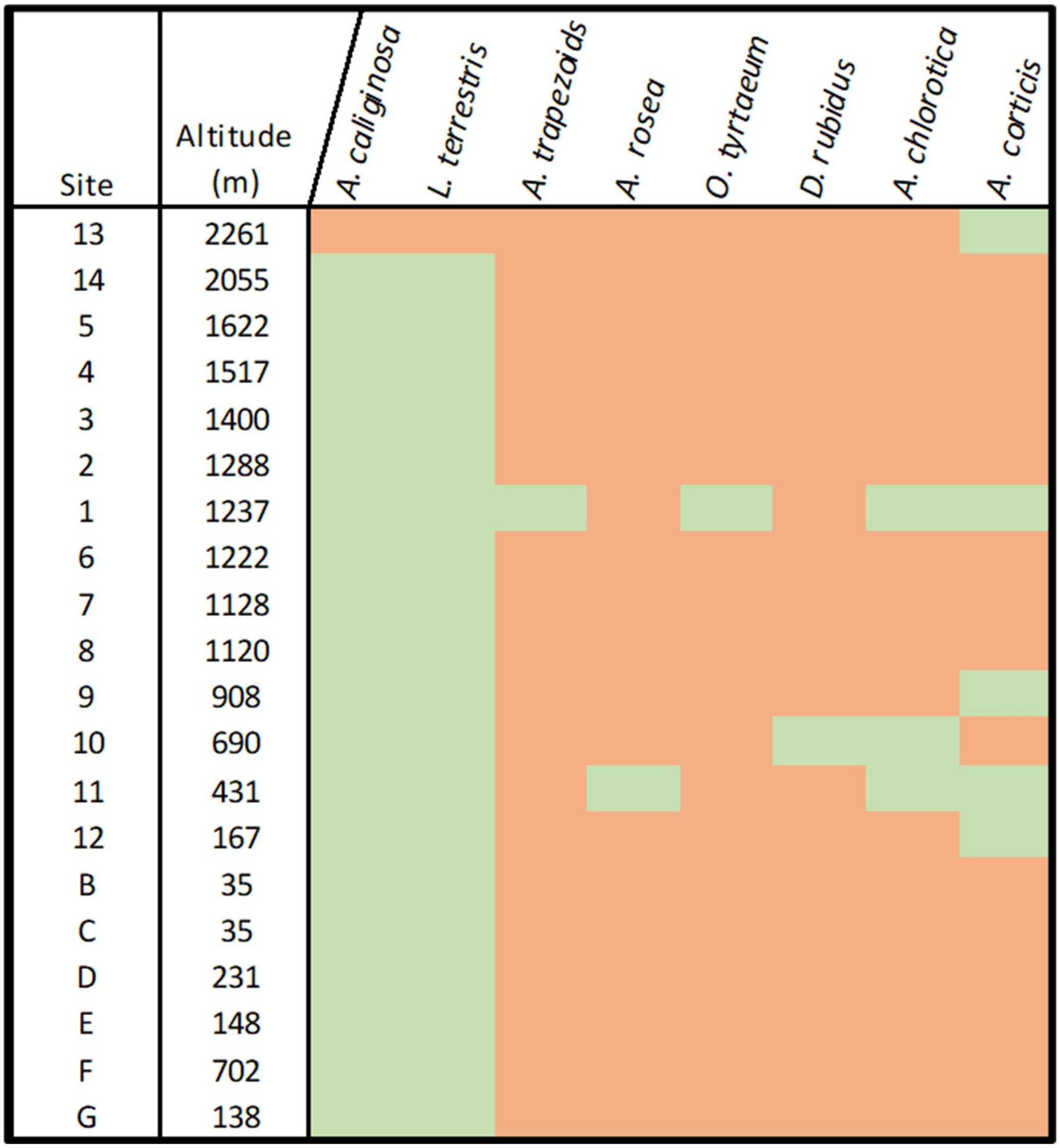
Mantel Tests were performed on individuals’ COII sequences and their associated geographical collection site in latitude and longitude using the software Alleles in Space with 1000 repeats (Miller 2005). This was to evaluate if a statistical relationship between geographical distance and sequence similarity could be identified. An analysis was performed for both *A. caliginosa* and *L. terrestris* in both the Alps and the Azores.

3.3. Results.

3.3.1. **Species distribution**

Pico sites 1-14 were sampled with a targeted collection of *A. caliginosa* and *L. terrestris* but did not exclude collection of other species of earthworms for genetic barcoding. The species presence at each site is indicated below in Table 3. Sites B-G searched for *A. caliginosa* and *L. terrestris* only.

Table 3: Earthworm species presence within Pico transect sites 1-14. *Green indicates species presence; orange indicates species absence.*



Site	Altitude (m)	*A. caliginosa*	*L. terrestris*	*A. trapezoids*	*A. rosea*	*O. tyrtaeum*	*D. rubidus*	*A. chlorotica*	*A. corticis*
13	2261	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟩
14	2055	🟩	🟩	🟧	🟧	🟧	🟧	🟧	🟧
5	1622	🟩	🟩	🟧	🟧	🟧	🟧	🟧	🟧
4	1517	🟩	🟩	🟧	🟧	🟧	🟧	🟧	🟧
3	1400	🟩	🟩	🟧	🟧	🟧	🟧	🟧	🟧
2	1288	🟩	🟩	🟧	🟧	🟧	🟧	🟧	🟧
1	1237	🟩	🟩	🟩	🟧	🟩	🟧	🟩	🟩
6	1222	🟩	🟩	🟧	🟧	🟧	🟧	🟧	🟧
7	1128	🟩	🟩	🟧	🟧	🟧	🟧	🟧	🟧
8	1120	🟩	🟩	🟧	🟧	🟧	🟧	🟧	🟧
9	908	🟩	🟩	🟧	🟧	🟧	🟧	🟧	🟩
10	690	🟩	🟩	🟧	🟧	🟧	🟩	🟩	🟧
11	431	🟩	🟩	🟧	🟩	🟧	🟧	🟩	🟩
12	167	🟩	🟩	🟧	🟧	🟧	🟧	🟧	🟩
B	35	🟩	🟩	🟧	🟧	🟧	🟧	🟧	🟧
C	35	🟩	🟩	🟧	🟧	🟧	🟧	🟧	🟧
D	231	🟩	🟩	🟧	🟧	🟧	🟧	🟧	🟧
E	148	🟩	🟩	🟧	🟧	🟧	🟧	🟧	🟧
F	702	🟩	🟩	🟧	🟧	🟧	🟧	🟧	🟧
G	138	🟩	🟩	🟧	🟧	🟧	🟧	🟧	🟧

Les Deux Alpes sites 1-14 were sampled indiscriminately for earthworms for genetic barcoding. The species presence at each site is indicated below in table 6.

Table 4: Earthworm species presence at Les Deux Alpes sites 1-14. *Green indicates species presence; orange indicates species absence.*

Site	Altitude (m)	*A. caliginosa*	*L. terrestris*	*A. longa*	*A. rosea*	*O. lacteum*	*D. rubidus*	*A. chlorotica*	*L. castaneus*	*E. tetraedra*
1	1674	🟩	🟧	🟩	🟧	🟧	🟧	🟩	🟩	🟧
2	1797	🟩	🟩	🟧	🟩	🟩	🟩	🟧	🟩	🟧
3	1970	🟩	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
4	1969	🟩	🟧	🟧	🟧	🟩	🟧	🟧	🟧	🟧
5	2008	🟩	🟩	🟧	🟧	🟩	🟧	🟧	🟧	🟧
7	2142	🟩	🟧	🟧	🟩	🟧	🟧	🟧	🟧	🟧
8	1818	🟩	🟩	🟩	🟧	🟩	🟧	🟧	🟩	🟧
9	1601	🟩	🟩	🟩	🟧	🟩	🟧	🟧	🟩	🟧
10	1406	🟩	🟩	🟩	🟧	🟩	🟧	🟧	🟩	🟧
11	1299	🟩	🟩	🟩	🟧	🟩	🟧	🟩	🟩	🟧
12	1140	🟩	🟧	🟧	🟧	🟩	🟧	🟩	🟩	🟧
14	1740	🟩	🟩	🟩	🟧	🟩	🟩	🟩	🟩	🟩

A Maximum Likelihood tree was generated for all the samples collected from Pico (Figure 8). *A. caliginosa* and *L. terrestris* samples from the majority of the 8 species of earthworm collected. Pico transect site 13 (caldera) was the only site where *A. caliginosa* and *L. terrestris* was not found. Sites B-G assessed *A. caliginosa* presence only. *A. caliginosa* displays 3 distinct but closely linked lineages and *L. terrestris* had 4 distinct and more distantly linked lineages.

A Maximum Likelihood tree was generated for all the samples collected from Les Deux Alpes (Figure 9). Samples collected were more equally distributed across 8 species. *A. caliginosa* formed the largest proportion of worms collected, found at all but site but 6, and were all linked closely on the tree. Found in half of the sampling sites, only 3 lineages were seen for *L. terrestris* however *L. castaneus* although defined as a separate species, showed similar linkage distances between the 3 *L. terrestris* linages.

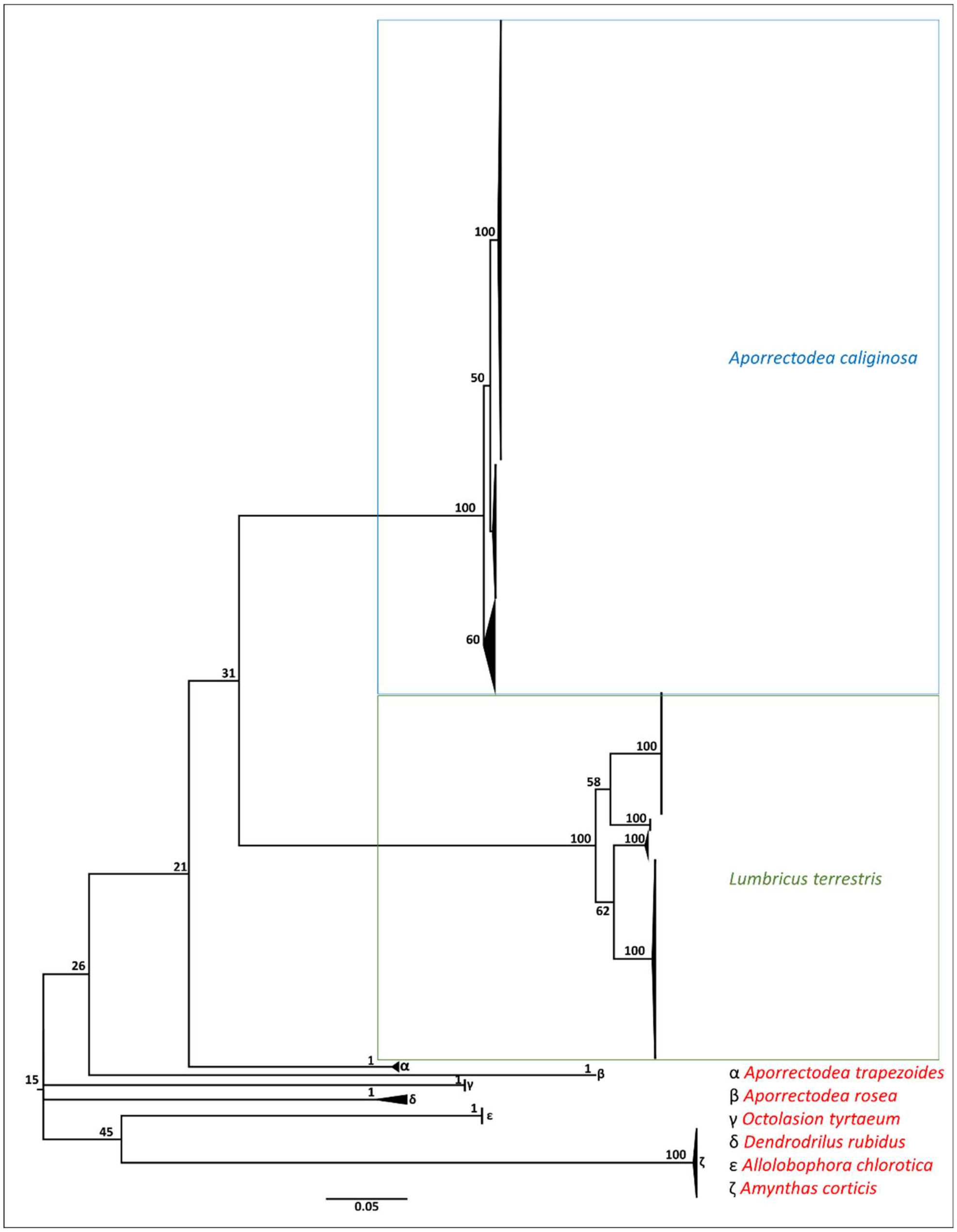


Figure 8: Maximum Likelihood tree for all samples collected from Pico, Azores. The evolutionary history was inferred by using the Maximum Likelihood method based on the Tamura-Nei model. The tree with the highest log likelihood (-3513.90) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbour-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 0.8706)). The rate variation model allowed for some sites to be evolutionarily invariable ([+I], 29.16% sites). The tree is drawn to scale, unrooted, with branch lengths measured in the number of substitutions per site. The analysis involved 290 nucleotide sequences. All positions containing gaps and missing data were eliminated. There were a total of 619 positions in the final dataset. Evolutionary analyses were conducted in MEGA7(Kumar et al. 2016).

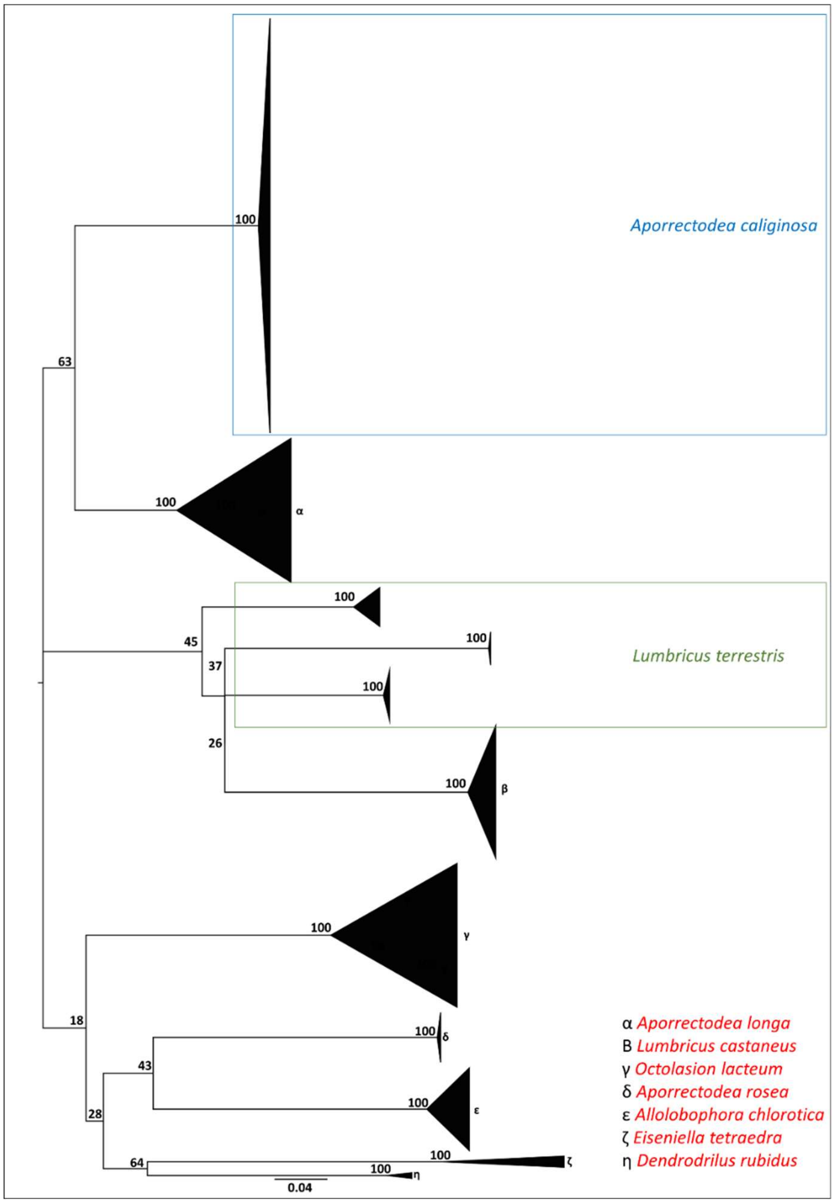


Figure 9: Maximum Likelihood tree for all samples collected from Les Deux Alps, France. The evolutionary history was inferred by using the Maximum Likelihood method based on the General Time Reversible model. The tree with the highest log likelihood (-5676.07) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbour-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 1.1770)). The rate variation model allowed for some sites to be evolutionarily invariable ([+I], 49.50% sites). The tree is drawn to scale, unrooted, with branch lengths measured in the number of substitutions per site. The analysis involved 211 nucleotide sequences. Codon positions included were 1st+2nd+3rd+Noncoding. All positions containing gaps and missing data were eliminated. There were a total of 684 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 (Kumar et al. 2016).

3.3.2. **Species diversity and haplotype diversity**

The calculated Theta-W, Pi and Tajima's D for *A. caliginosa* and *L. terrestris* are shown in Table 5. DNA polymorphism as assessed with Theta-W, estimates a lower scaled mutation rated determined from segregating sites for *A. caliginosa* from Pico but higher for *A. caliginosa* from Les Deux Alpes and *L. terrestris* from both sites. DNA polymorphism as assessed with Pi, estimates the average pairwise differences to be higher in *L. terrestris* than in *A. caliginosa*. Only *L. terrestris* from Pico has a statistically significant P value rejecting Tajima's D null hypothesis of constant effective population size and may be indicative of a population bottleneck.

Table 5: Genetic diversity metrics, Theta-W, Pi and Tajima's D, calculated for *A. caliginosa* and *L. terrestris* for both sampling sites. Bracketed values shown in *A. caliginosa* indicated values calculated for Pico and São Miguel together () and values calculated for Les Deux Alpes including *L. terrestris* and *L. castaneus* [].

	Pico, Azores		*Les Deux Alpes*	
	A. caliginosa	*L. terrestris*	*A. caliginosa*	*L. terrestris*
Individuals	**165 (245)**	**92**	**76**	**[42]**
Nucleotide sites	**619 (619)**	**619**	**684**	**[684]**
Watterson's theta	**0.00427 (0.03990)**	**2.35828**	**1.09467**	**[2.03423]**
Pi	**0.00562 (0.00473)**	**0.03293**	**0.00302**	**[0.11748]**
Tajima's D	**0.62065** P>0.10 **(0.27933)** P>0.10	**0.01812** P<0.05	**0.00209** P>0.10	**[0.06354** P>0.05]

Haplotype networks are shown in Figures 10 and 11 and their haplotype and nucleotide diversity is shown in (Table 6). Haplotype diversity in *A. caliginosa* from Pico is lower than seen in other populations while *L. terrestris* in Les Deux Alpes has the highest diversity.

Table 6: Haplotype diversity calculated for *A. caliginosa* and *L. terrestris* for both sampling sites. Bracketed values shown in *A. caliginosa* indicated values calculated for Pico and São Miguel together () .

	Pico, Azores		*Les Deux Alpes*	
	A. caliginosa	*L. terrestris*	*A. caliginosa*	*L. terrestris*
Individuals	**(244)**	**91**	**76**	**[42]**
Nucleotide sites	**(619)**	**619**	**684**	**[684]**
Haplotype diversity	**(0.4867)**	**0.7248**	**0.6961**	**[0.7898]**
Pi	**(0.00473)**	**0.03293**	**0.00302**	**[0.11748]**

The minimum spanning network for *A.* c*aliginosa* from Pico and São Miguel (Figure 10) indicate haplotype groups are closely linked. The three largest haplotype groups contain worms from both Pico and São Miguel, but Pico specific haplotype do exist. *L. terrestris* displays larger nucleotide diversity between haplotypes. *A. caliginosa* from Les Deux Alpes contains fewer haplotypes and have a low nucleotide diversity in contrast to *L. terrestris* from Les Deux Alpes, which has a very high nucleotide diversity between lineages haplotypes 3-6 but also a similar nucleotide diversity to *L. castaneus*.

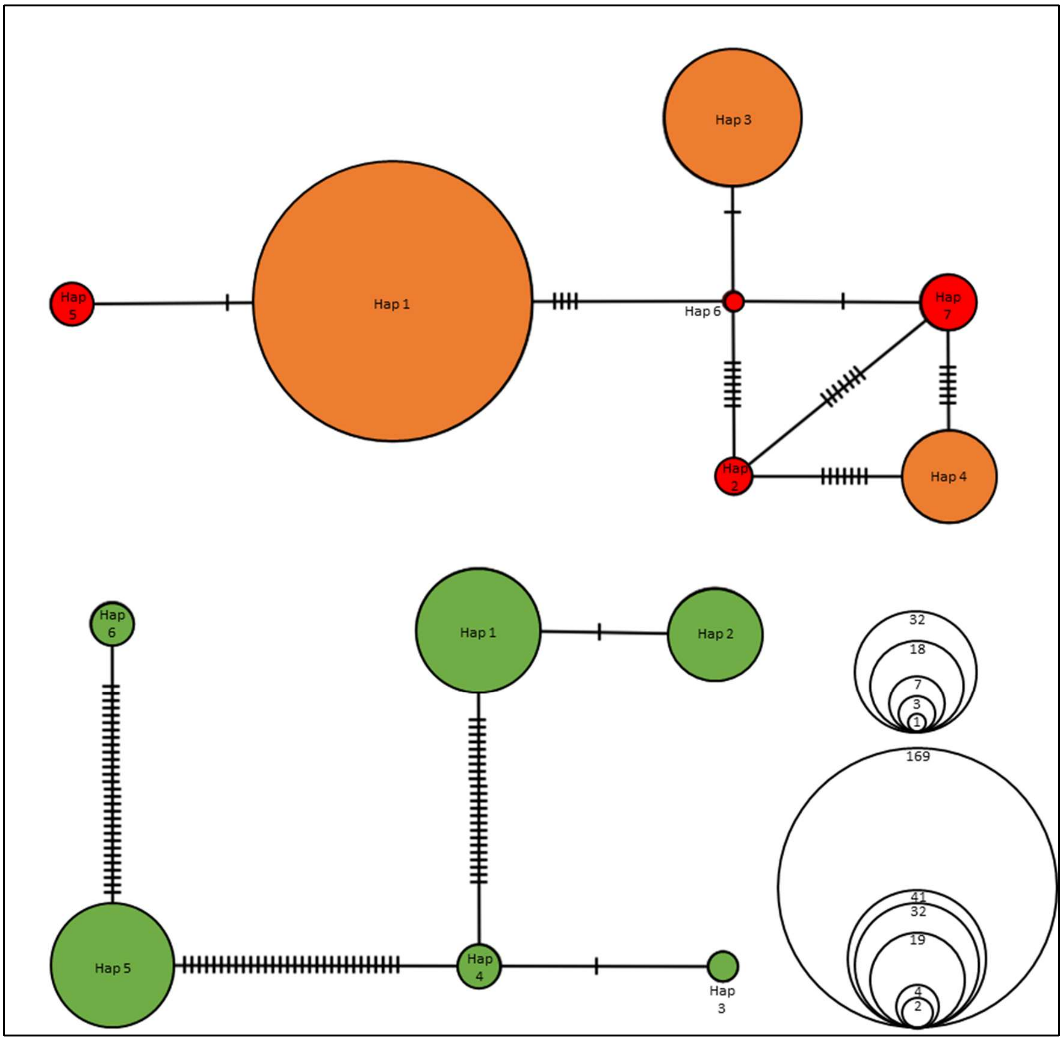


Figure 10: Minimum spanning networks of *A. caliginosa* sampled from the islands of Pico and São Miguel and *L. terrestris* from Pico only. Red indicates a Pico only haplotype of *A. caliginosa* and orange haplotypes found on both Pico and São Miguel. Green indicates *L. terrestris*. Circle size indicates the number of individuals in each haplotype and hatch marking indicate number of polymorphisms between haplotypes.

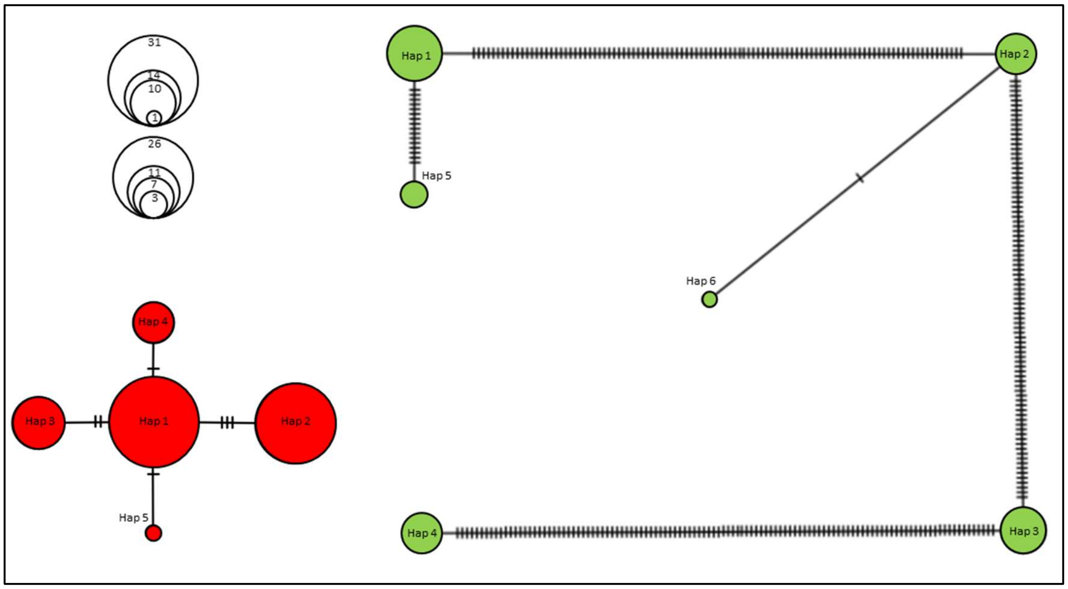


Figure 11: Minimum spanning networks generated of *A. caliginosa* and *L. terrestris* sampled from Les Deux Alpes. Red indicates *A. caliginosa,* green haplotypes 1-2 indicates *L. castaneus* and green haplotypes 3-6 indicates *L. terrestris*. Circle size indicates the number of individuals in each haplotype and hatch marking indicate number of polymorphisms between haplotypes.

There was no strong positive or negative association between sequence similarity and geographical distance between individuals observed as determined with a Mantel test (Table 7). This does not preclude any potential relationship if other non-mitochondrial genes are tested.

Table 7: Calculated values from Mantel tests for the correlation between COII sequence similarity and geographical distance. No statistically significant correlation is observed between sequence similarity and geographical distance. R indicates measure of linear correlation.

Location	*Species*	*Assessed against*	*Correlation value of R*	*Probability observing ≥ R*	*Probability observing ≤ R*
Pico	*A. caliginosa*	Global population	0.034	0.067	0.934
		Azores population	-0.061	1.000	0.001
		Pico only population	0.043	0.182	0.819
	L. terrestris	Global population	0.157	0.001	1.000
		Pico only population	0.005	0.536	0.466
Les Deux Alpes (LDA)	*A. caliginosa*	Global population	0.001	0.185	0.816
		LDA only population	0.033	0.058	0.943
	L. terrestris	Global population	0.262	0.002	0.999
		LDA only population	0.087	0.097	0.904

3.4. Discussion.

3.4.1. **General discussion**

The aim of this investigation was to derive the earthworm biodiversity with two contrasting altitudinal temperate mountainous transects, from the Island of Pico in the Azores and Les Deux Alpes in the French Alps, exploring both species and haplotype diversity as determined by analysis of a mitochondrial marker.

A total of eight species were collected on the two sampling trips to Pico, with the most abundant species of *A. caliginosa* and *L. terrestris* being found across the island. These two species have very different lifecycles. *A. caliginosa* is a small (40-180mm length), shallow living and horizontally burrowing earthworm that in prolonged freezing conditions will die leaving only frost proof cocoons to repopulate next season (Sims and Gerard 1999; Holmstrup 2003; Holmstrup and Overgaard 2007). In contrast *L. terrestris* is a large (90-350mm length), deep burrowing earthworm that tries to avoid freezing conditions by migrating to deeper soil layers. These species along with *A. trapezoids, A. rosea*, *O*. *tyrtaeum, D. rubidus, A. chlorotica* originate from the Western palearctic while *A. corticis* from Indo-Malay (Sims and Gerard 1999). Although not found on our sampling trips to Pico, *P. corethrurus* has been reported in the Azores and originates from the Neo-Tropical (Cunha et al. 2014). While only the top of Mount Pico tends to see a snow-cap in winter months, it was surprising to see *L. terrestris* at an altitude where this occurs as the soil layer appeared very shallow and was only found in small pockets within the rocky landscape. It is possible that the worms were utilising cracks in the rock to burrow deep to avoid snow but this has not been investigated to date. Although the sampling trips to Pico focused on the collection of *A. caliginosa* and *L. terrestris*, other species were not excluded from collection, but the low level presence of the 6 other species would suggest that both *A. caliginosa* and *L. terrestris* are the dominant species on Pico.

In Les Deux Alpes, nine different species were collected and *A. caliginosa* once again proved to be the most widespread species found in all sites where worms were found including the highest altitude site on the site on the edge of melting snow and the lowest and driest site. While there was not an even site distribution of species, there was a widespread presence of different species including *L. terrestris*, *L. castaneus*, *A. longa*, *A. rosea*, *O. lacteum*, *A*. *chlorotica*, *D. rubidus* and *E. tetraedra*. Like Pico, I found the deep burrowers like *L. terrestris* and *L. castaneus* at high altitudes, where much of the year is spent under snow, although with deeper soil depths a more obvious frost refuge can be observed. The sampling trip to Les Deux Alpes followed sample collection on Pico and I hoped to find *A. caliginosa* and *L. terrestris* to compare the two sites. While *both A. caliginosa* and *L. terrestris* were present, the seven other species were found at higher population densities. Tables 4 and 5 display these differences more clearly. Sampling sites that were lower, had deeper soil depth and were more sheltered had greater species diversity.

Recently when performing genetic barcoding, there has been a push by some to utilise one singular conserved gene that has taxonomic information but which has conserved flanking sides that can be used to generate primers that can be widely (across taxa) used, and the mitochondrial COI gene was the marker of choice. The Barcode of Life Data System (BOLD) is a growing database of animals, plants, fungi and other species (Ratnasingham and Herbert 2007). The initial research plan intended to use the COI gene for barcoding and taxonomically identify the samples collected here. However, although the gene is generally well conserved across some Phyla, amplification of this gene remained a challenge, and despite significant primer redesign attempts, we found the start and end sites of most COI primers were highly diverse in the earthworms we tested, hence explaining the difficulty in obtaining successful PCR amplification. Other commonly used barcoding genes include the mitochondrial COII gene. The COII primer proved to be robust in PCR, generating amplification of all samples (Folmer et al. 1994; Pérez-Losada et al. 2009). The amplified products were Sanger sequenced by Eurofins Genomics (Luxembourg). The amplified COII gene product also provided cleaner sequence data that could be achieved with the amplified COI gene product. The COII sequences of earthworms from Pico and Les Deux Alps were assessed before use in downstream genetic analysis. This was to ensure calling of the bases from the sequencer was accurate, and in some cases, it was required to correct some miss-called bases at the start and end of sequences.

The diversity analyses of *A. caliginosa* from Pico (i.e. Theta-W, Pi and Tajima's D) was paralleled by a secondary analysis using *A. caliginosa* that had been collected from the neighbouring island of São Miguel enabling a comparison of diversity between islands of the Azores. Although both *A. caliginosa* populations from Pico and Les Deux Alpes fail statistical significance for Tajima's D test, Theta-W and Pi values calculated suggest a slightly higher diversity within the Alpine population. When performing the same population diversity tests on population of *L. terrestris* from Les Deux Alpes, the *L. castaneus* population was included. This was because the population distance assessed in the Maximum Likelihood tree suggested similar distances between *L. terrestris* lineages as to *L. castaneus*. This might suggest that the samples taxonomically identified as *L. castaneus* are genetically *L. terrestris* or that the two name taxonomic groups are not sufficiently genetically distant. This could influence consideration that the *L. castaneus* be named a linage of *L. terrestris* or that the lineages of *L. terrestris* are genetically distant enough that they should all be named as separate species with *L. castaneus* as one of them. For both populations of *L. terrestris* from Pico and Les Deux Alpes, species diversity was high, with the

Pico population the only passing statistical significance of Tajima's D test suggesting population change.

Haplotype diversity and Pi for haplotype groups mirror the same pattern of diversity seen in the populations of individuals with a higher diversity observed among the haplotypes of *L. terrestris* and a lower diversity observed in *A. caliginosa* individuals from Pico. The minimum spanning networks (figures 10 and 11), help to display these diversities. It is clear from this analysis that *A. caliginosa* has lower diversity than that seen in *L. terrestris* in both field sites. This indicates that the populations found at high and low altitude for *A. caliginosa* have lower genetic diversity in both Pico and Les Deux Alpes and suggests they are more closely genetically linked.

As discussed in 3.1.1, Pico is the youngest of the volcanic islands with volcanic eruptions as recent as 300 years ago. History, prior to large scale colonisation is not well documented, though while it was not described as the barren Ascension islands when discovered or settled, human impact on the islands of the Azores in the 15th century was rapid with clearing of bush for grape and sugar cultivation (Ashe 1813). While there have been some suggestions over short distances birds could transport earthworms in their mouths only to accidentally drop their meal, the Azores sit about 1600 km from the nearest mainland, an implausible distance for such an even to seed an earthworm population on the archipelagos. The lack of viable migration possibilities for earthworms to reach the islands other than in the roots of crops taken from Europe for cultivating, puts a hard introduction time limit of around 500-600 years. The multiple species clades identified through COII barcoding indicates that there was no single introduction, and individuals from three clades were all seeded. As both *A. caliginosa* and *L. terrestris* dominate presence, either they were seeded earlier and in greater numbers. Other species colonising the islands' have not been limited to human introductions. Spiders are thought to have potentially colonised via 'ballooning' as a means that can exploit air currents to cover vast distances (Borges and Wunderlich 2008). While avian species can similarly exploit favourable air currents to travel long distances, all mammals of the Azores with the exception of bats are thought to have either stowed away (rats and mice), been brought as food (rabbits), or used for other purposes (weasels and ferrets) (Mathias et al. 1998).

The history and colonisation of Les Deux Alpes is a more obscured image with at least human impact of population seeding through cultivation of crops going back thousands of years. While the diversity in *L. terrestris* seems to reflect this, *A. caliginosa* diversity is quite low and clade presence evenly distributed. The caveat to this is the high presence of *A. longa* and *A. rosea*, which while described as a separate species, should be considered clades (Perez-Losada et al. 2009).

3.4.2. **Future analyses**

Part of the aim of assessing diversity in the populations of *A. caliginosa* and *L. terrestris* was to determine which species might be better suited for transcriptomic analysis in downstream analyses. From these assessments, *A. caliginosa* would seem more suited as it has a lower population diversity and we hope to investigate changes in gene expression that are more a factor of the environment rather than of species lineage.

In the following chapters, I will build upon the analyses explored in this chapter, investigating mismatch distribution, Bayesian skyline analysis and time routed trees of each population's haplotypes.

3.4.3. **Concluding remarks**

This thesis aims to identify the ways earthworms are adapting or acclimatising to life at high altitude. In this chapter, we have determined that *L. terrestris* has a much greater genetic diversity than *A. caliginosa*. While we have not found any haplotypes associated with high altitude, it has suggested that for an 'in-field' transcriptomic study of gene expression, *A. caliginosa* would be a suitable species for analysis. This is because any change in gene expression is more likely a result of the environment than it is of genetic diversity. In the following chapters, I will investigate at the population dynamics of *A. caliginosa* and *L. terrestris* and try to link this to physical and human history of the area.

4. Exploiting population dynamics to explore the contrasting demographic context of earthworm species inhabiting wide altitudinal ranges.

4.1. Modern genetic signatures of demographic history

Earthworms are ancient and though they do not leave fossils, genetic analysis indicates their presence before the breakup of the supercontinent Pangaea 175 million years ago (Anderson et al. 2017b). Their modern genetic material contains signatures of the population-genetic processes that has given rise to sampled population (Ho and Shapiro 2011). This allows us to explore correlations between major demographic events with possible historic drivers, such as glaciation or anthropogenic manipulation. Understanding the population history of the two earthworm species displaying occupancy throughout the altitudinal ranges, *A. caliginosa* and *L. terrestris,* provides a context for haplotype selection or adaptation. In trying to predict the earthworm population dynamics of both species in both settings, the single greatest influencer is likely to be the bottlenecking of genetic diversity caused by the last glacial maximum. Despite this being more than 10,000 years ago, the sedentary nature of earthworms and very low migration rates mean this is a very short period for species to recover lost genetic diversity (Edwards and Bohlen 1977).

It is important to be careful when assessing the impact of the research in this chapter and take account of the limitations of using a single mitochondrial marker. While the use of a COII amplicon works well for barcoding species and assessing haplotype groups, it is still a single gene passed down from a single parent and is not a powerful as employing nuclear genes as well.

This chapter aims to study the population dynamics of *A. caliginosa* and *L. terrestris* sampled from Pico and Les Deux Alpes. This will involve performing analysis of population size change over time and a diversity assessment against geographical location. Using the COII data collected in the last chapter (Chapter 3), I will continue the population assessment of each species and try to link findings to the known physical and human history of the area. From this, a single species will be selected for subsequent chapters that can best answer the fundamental question of this thesis, are worms acclimatised or adapted to high altitude.

4.2. Population and historical context

4.2.1. **Pico.**

As identified in Chapter 3, the genetic diversity of *A. caliginosa* found on the island of Pico and its neighbouring island São Miguel was comparatively lower between three lineages than identified in *L. terrestris*, with fewer intra-haplotype DNA polymorphisms (π diversity 0.005 in *A. caliginosa*, 0.033 in *L. terrestris*). The three largest haplotypes included samples from both islands and there were four smaller Pico specific haplotypes. We also looked at the genetic diversity of *L. terrestris* from Pico and found the nucleotide diversity to be much higher than that of *A. caliginosa* with four distinct lineages and diversity was greater between haplotypes. Tajima's D was statistically significant suggesting there has been a population change leading to the present population.

4.2.2. **Les Deux Alpes.**

Similar trends, to these recorded in Pico, were observed when examining the genetic diversity of *A. caliginosa* from Les Deux Alpes, however, there was a very high genetic diversity within the *L. terrestris* collected along the transect sampled (π diversity 0.003 in *A. caliginosa*, 0.117 in *L. terrestris*). This was immediately apparent in the Maximum likelihood tree (3.3.1 Figure 8) where the inter-species genetic diversity to *L. castaneus* measured by the distance of branch length, (number of substitutions per site), was as great as the distance intra *L. terrestris* genetic diversity between lineages. This with the high values of Theta-W and Pi of individuals and haplotype groups suggests that these diverse lineages are sufficiently distinct to be regarded as their own species.

4.2.3. **Expected population changes of Pico.**

The volcanic formation (0.27 M years) and potential start of the oldest human occupation (~1000 years) of island of Pico spans two very different time periods (for details see 3.1). If we were to follow a human introduction model of earthworms to the island, Island influences to population dynamics would be confined to ~1,000 years in which population expansion could be identified as they spread across a new habitat, though this is likely to be dominated by population bottlenecking from a reduced population diversity being introduced (Hendrix et al. 2008). Given this period follows directly from the end of the last glacial maximum, from around 26,000 years ago, this could be further hidden as part of a larger trend (Sibrava et al. 1986). This would comparatively make 1,000 years ago seem rather brief for substantial changes in population diversity. Pico and the Azores, as islands in the middle of the Atlantic Ocean, form some of the most naturally isolated environments prior to human influence. It is therefore highly likely that the local earthworm populations are a direct result of human introduction.

Pico was settled primarily by the Portuguese (see 3.1.1), thus, we could predict both species' haplotypes should have close ties to the Iberian Peninsula, but there may still be links to haplotypes found in France through their later migration.

4.2.4. **Expected population changes of Les Deux Alpes.**

In contrast to Pico, Les Deux Alpes are centrally located in the Western Palearctic and could have earthworm population dynamic independent on human influence. Due to the heavy human presence and development of the sampling site however, it is probable that humans have had at least some impact upon populations found here. Like Pico, these populations will have been greatly impacted from the last glacial maximum. Les Deux Alpes is one of the areas last be released from the glacial maximum evidenced by the still presence of glaciers. Given the nature of deep burrowing *L. terrestris*, one might expect a large bottlenecking of the species during that last glacial maximum and both *L. terrestris* and *A. caliginosa* to population dynamics to remain steady or begin slow population expansion following the recession of the glaciers. At high altitude, this freedom to expand is likely only beginning and as we have discussed in 4.2.1, it is a brief period in the history of earthworms. The populations of both species are likely to have links to most Western Palearctic haplotypes owning to its geographic location.

4.3. Methods

4.3.1. **Demographic reconstructions.**

Mismatch distributions were calculated using the population data reported in Chapter 3 for each species (*A. caliginosa* and *L. terrestris*) in Pico and Les Deux Alpes using DnaSP's population size-change calculator and data was then charted as a line graph with DnaSP (Rozas et al. 2017).

A Bayesian Skyline Plot analysis was performed on haplotype groups to assess effective population size changes over time for *A. caliginosa* and *L. terrestris*. This was performed for each species separately using BEAST v1.8.4 using an estimated molecular divergence rate of 2.4% per site per million years (Chang and James 2011). This was performed with 10 combined runs of 20,000,000 steps of the Markov Chain Monte Carlo algorithm and discarding the initial 25% steps as burn-in (Drummond et al. 2012). Results were imported to Tracer v1.6 to determine that the Markov chains reached stationarity and to generate a Bayesian Skyline Plot reconstruction (Drummond et al. 2012). The BEAST input was generated with BEAUti v1.8.4.

4.3.1. **Time rooted haplotype tree generation.**

For the Azores, haplotype groups were generated of samples from Pico and São Miguel, along with published COII data for *A. caliginosa* and *L. terrestris*, shown below in Table 1. Similarly, for Les Deux Alpes haplotype groups were generated including the published COII data shown in Table 8. The rationale of this analysis was to geographically compare and contrast the samples collected from our field site while being able to derive a timescale for the divergence processes inferred. Analysis for *L. terrestris* in Les Deux Alpes included *L. castaneus* as Chapter 3 indicated the similar genetic distance between species and *L. terrestris* lineages.

Table 8: NCBI samples used and their locations.

Species	*Author and NCBI accessions*	*Locations*
A. caliginosa	(Pérez-Losada et al. 2009) FJ967736 - FJ967747, FJ967749, FJ967750, FJ967752, FJ967753, FJ967774.	France, Germany, Finland, Spain, Corsica and Ireland.
A. caliginosa	(Fernandez et al. 2012) JQ763495 - JQ763498.	France and Spain.
A. caliginosa	(Klarica et al. 2012) JN869552 - JN869555.	Ireland and Austria.
L. terrestris	(Klarica et al. 2012) JN869600 - JN869614	Canada, Ireland, and Austria.

These analyses were performed in BEAST for these haplotypes with the same parameters used in 4.3.1. Maximum clade credibility were inferred for the tree of haplotypes generated for each species using TreeAnnotator v2.4.8 (Chang and James 2011) and visualised on FigTree v1.4.3 (Drummond et al. 2012).

4.3.2. **Isolation by distance.**

Mantel tests were used to assess the correlation between the genetic and geographical distance between local populations, and between global populations. This was performed for each species separately in both Pico and Les Deux Alpes. A total of 1,000 permutations were used to determine significance of the auto-correlograms using the software Alleles in Space v1.0 (Miller 2005).

4.4. Results

4.4.1. **Pico.**

A mismatch distribution was calculated for *A. caliginosa* and *L. terrestris* in Pico. Both showed a multimodal distribution with neither species matching the expected frequency for a sudden demographic expansion modeon (Figure 12).

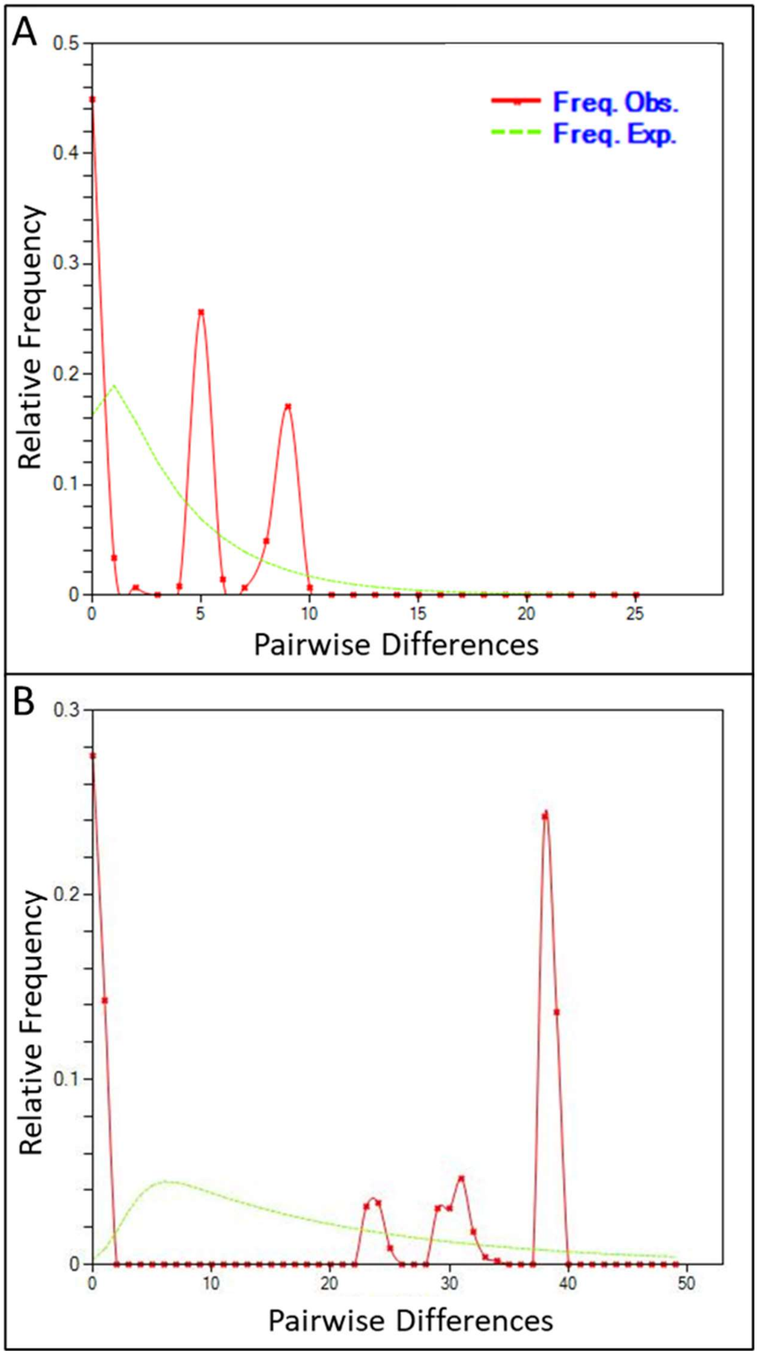


Figure 12: A mismatch distribution of *A. caliginosa* (A) and *L. terrestris* (B) from samples collected from Pico. The green line indicates the expected distribution of pairwise differences under a sudden demographic expansion model and the red line the observed distribution of pairwise differences. Both species show a multimodal distribution not matching expected frequencies for population expansion.

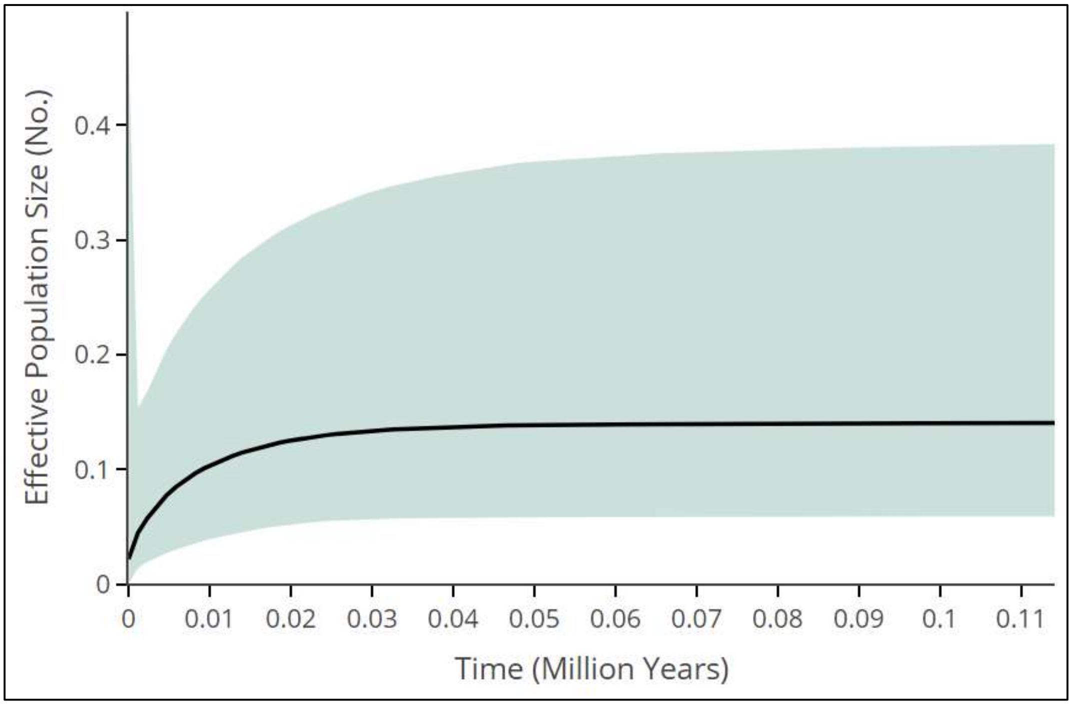


Figure 13: A Bayesian skyline plot analysis for *A. caliginosa* in Pico indicating long scale population decline. The black line indicates the median plot for the analysis and the blue shading the range between upper and lower confidence interval.

A Bayesian skyline analysis was performed on the individual samples for *A. caliginosa* from Pico. It indicated a long term slow population decline starting around ~50,000 years ago and reducing the ancestral effective population size to about a third of it. The jump in the upper confidence interval in the last 1000 years ago could suggest the onset of a population expansion in the near pasts (Figure 13).

Similarly, *L. terrestris* individual samples underwent Bayesian skyline analysis and showed a large-scale population decline starting ~100,000 years ago, reducing the effective population size by nearly 100 fold. The analysis showed a similar slowing of the narrowing of the confidence intervals, as the plot reached timepoint 0 (Figure 14).

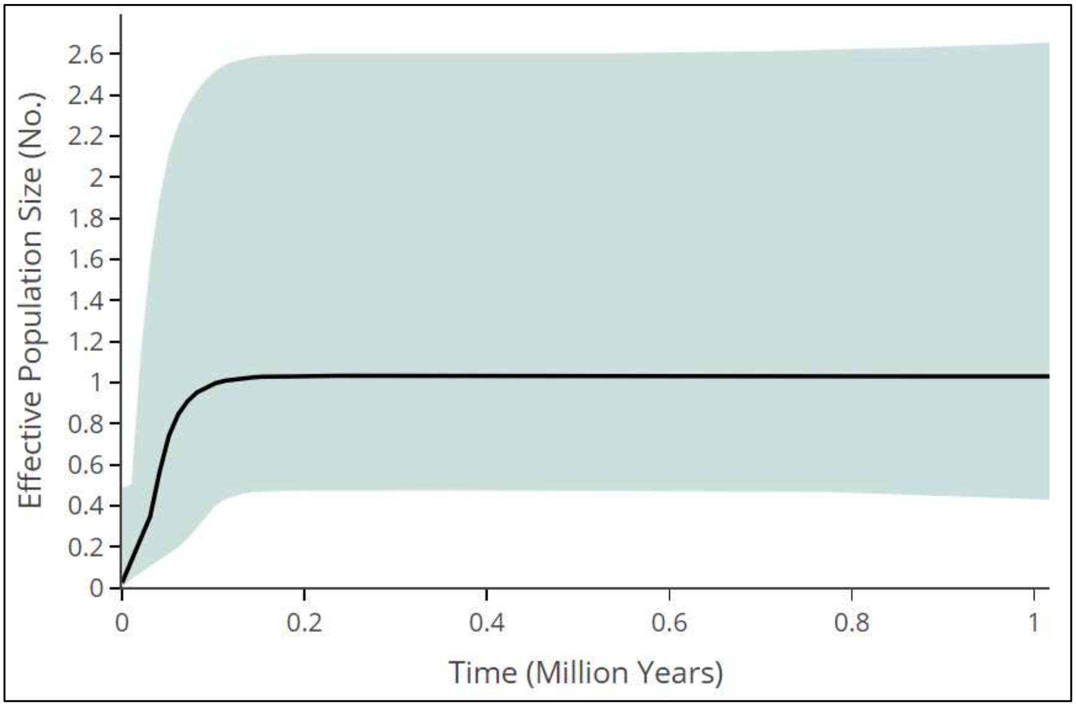


Figure 14: A Bayesian skyline plot analysis for *L. terrestris* in Pico indicating long scale population decline. The black line indicates the median plot for the analysis and the blue shading the range between upper and lower confidence interval.

Table 9: Calculated values of Haplotype diversity for *A. caliginosa* and *L. terrestris* for samples from Pico with global reference samples. *A. caliginosa* includes samples from São Miguel as in previous analysis.

	Pico, Azores	
	A. caliginosa	*L. terrestris*
Individuals	**269**	**106**
Nucleotide sites	**443**	**517**
Haplotype diversity	**0.4754**	**0.7941**
Pi	**0.00380**	**0.03654**

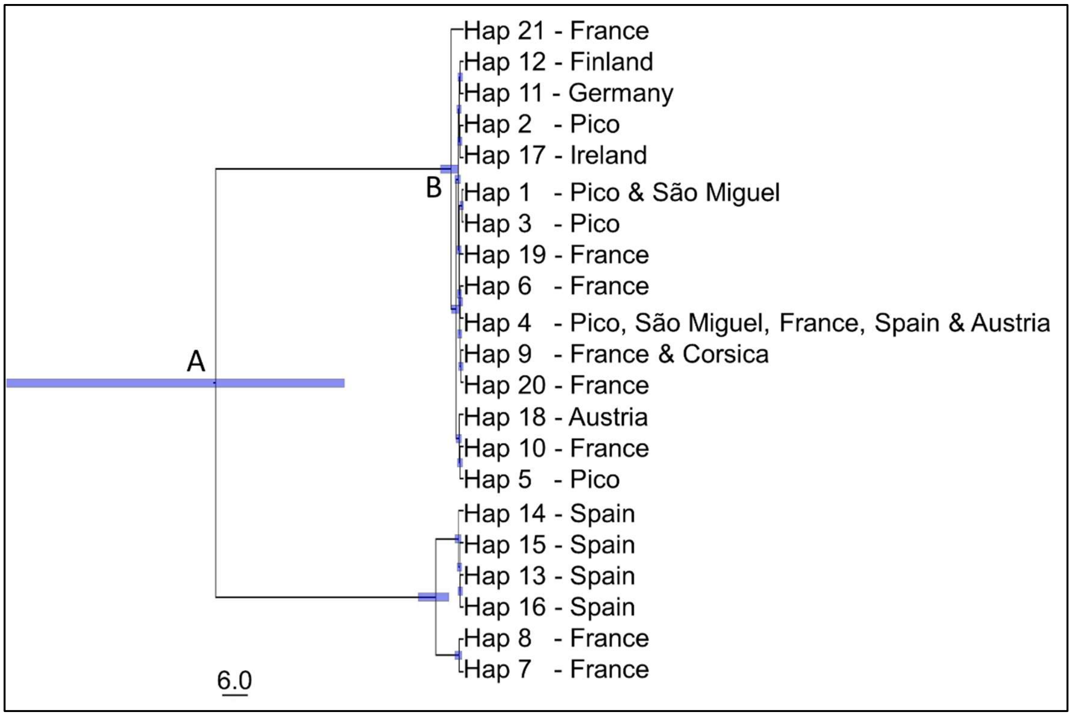


Figure 15: Dated phylogeny estimating divergence time for *A. caliginosa in Pico*, calculated using COII haplotype groups with trees generated on BEAST v1.8.4. and visualised with TreeAnnotator v2.4.8. Time scale is in 6 million years and blue node bars indicate confidence interval of branch splits. Split 'A' occurs between 27.9 and 107.4 million years ago, while split 'B' occurs 1.2 and 5.3 million years ago.

Haplotype groups were generated for *A. caliginosa* and *L. terrestris* and their diversity assessed as part of their generation (Table 9). A BEAST analysis was performed on the haplotype groups of *A. caliginosa* from Pico, São Miguel and the Global reference samples (Figure 15). Haplotypes from Pico and São Miguel (1-5) are interspersed with haplotypes from across Europe (4, 6, 9-12, 19-21) that split between 5.3 and 1.2 million years ago. A second group of *A. caliginosa* haplotypes from Northern Spain and the French Pyrenees split between 27.9 and 107.4 million years ago.

A Beast analysis was performed on the haplotype groups of *L. terrestris* samples. Haplotypes from the Pico samples (1-6) split into three divergent branches with haplotypes 1 - 4 more closely linked with Canadian haplotypes and haplotypes 5 and 6 more closely linked with Austria and Ireland reference samples. The divergence split (Split 'A', Figure 16) between Pico haplotypes 1-4 and 5-6 is between 10 and 21.1 million years but the overlapping confidence bars suggest that the uncertainty about these dates includes much more recent times.

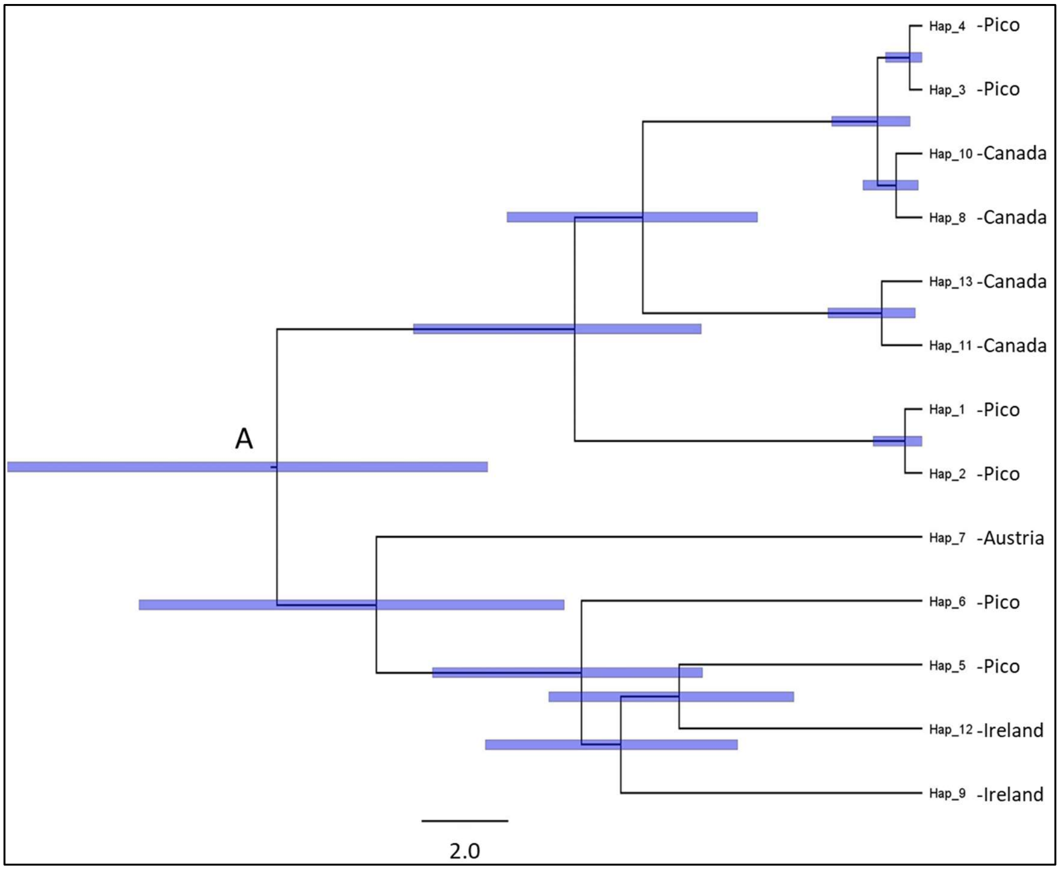


Figure 16: Dated phylogeny estimating divergence time for *L. terrestris in Pico*, calculated using COII haplotype groups with trees generated on BEAST v1.8.4. and visualised with TreeAnnotator v2.4.8. Time scale is in 2 million years and blue node bars indicate confidence interval of branch splits. Divergence split 'A' occurred 10-21.1 million years ago.

Table 10: Mantel test values calculated for stages of geographical distance from Pico. No test shows a strong correlation between genetic and geographical distance.

	Genetic & geographic distances R-value	*P value for observing ≥ to r value*	*P value for observing ≤ to r value*
A. caliginosa Pico	0.04	0.18	0.82
A. caliginosa Azores	-0.10	1.00	0.00
A. caliginosa Global	0.03	0.10	0.90
L. terrestris Pico	0.00	0.54	0.46
L. terrestris Global	0.16	0.00	1.00

A mantel test was performed at three stages for *A. caliginosa* (Pico, Pico and São Miguel, Pico and São Miguel and global reference samples) and two stages for *L. terrestris* (Pico, Pico and São Miguel and global reference samples) (Table 10). All tests showed no significant R-value for the relationship between genetic distance and geographic distance.

4.4.2. **Les Deux Alpes earthworms.**

A mismatch distribution was calculated for *A. caliginosa* and *L. terrestris* in Les Deux Alpes. Both showed a multimodal distribution with *L. terrestris* not matching the expected frequency for population expansion. While there was a close match the expected frequency for population expansion, *A. caliginosa* did not match perfectly, meaning any potential start to population expansion remains ambiguous (Figure 17).

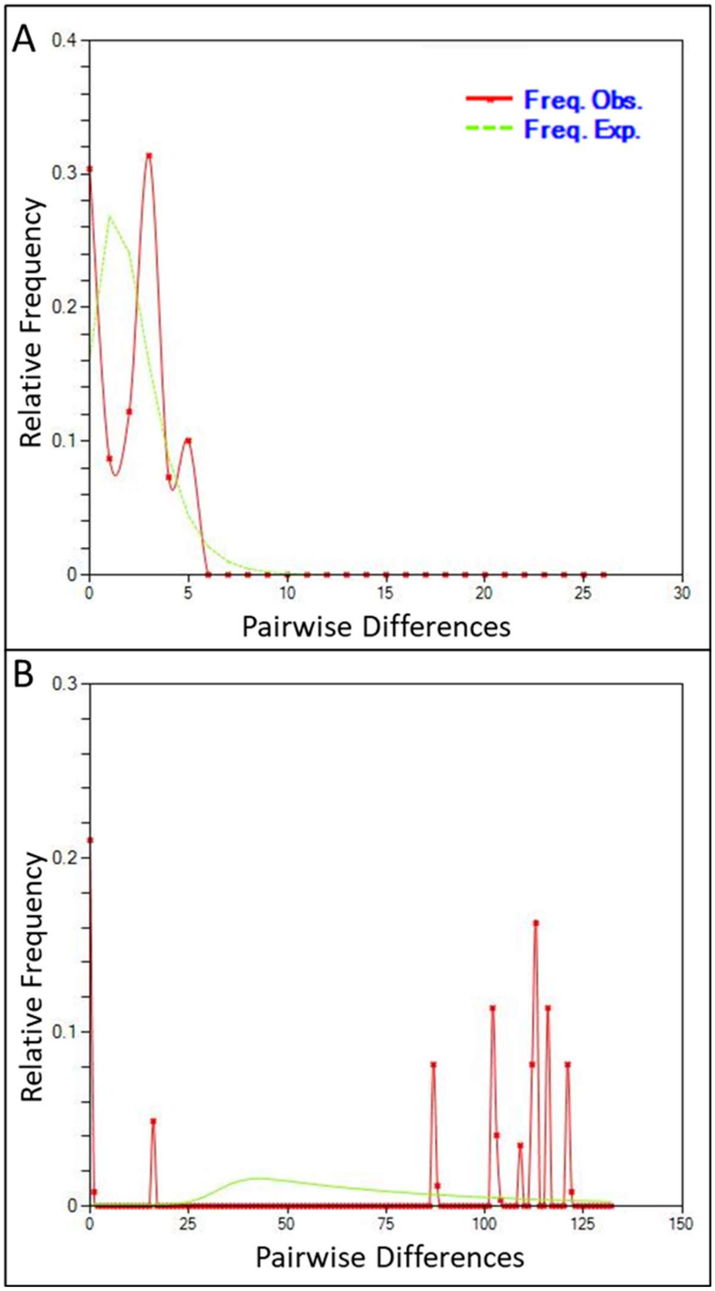


Figure 17: A mismatch distribution of *A. caliginosa* (A) and *L. terrestris* (B) from samples collected from Les Deux Alpes. The green line indicates the expected distribution of pairwise differences under a sudden demographic expansion model and the red line the observed distribution of pairwise differences. Both species show a multimodal distribution not matching expected frequencies for population expansion.

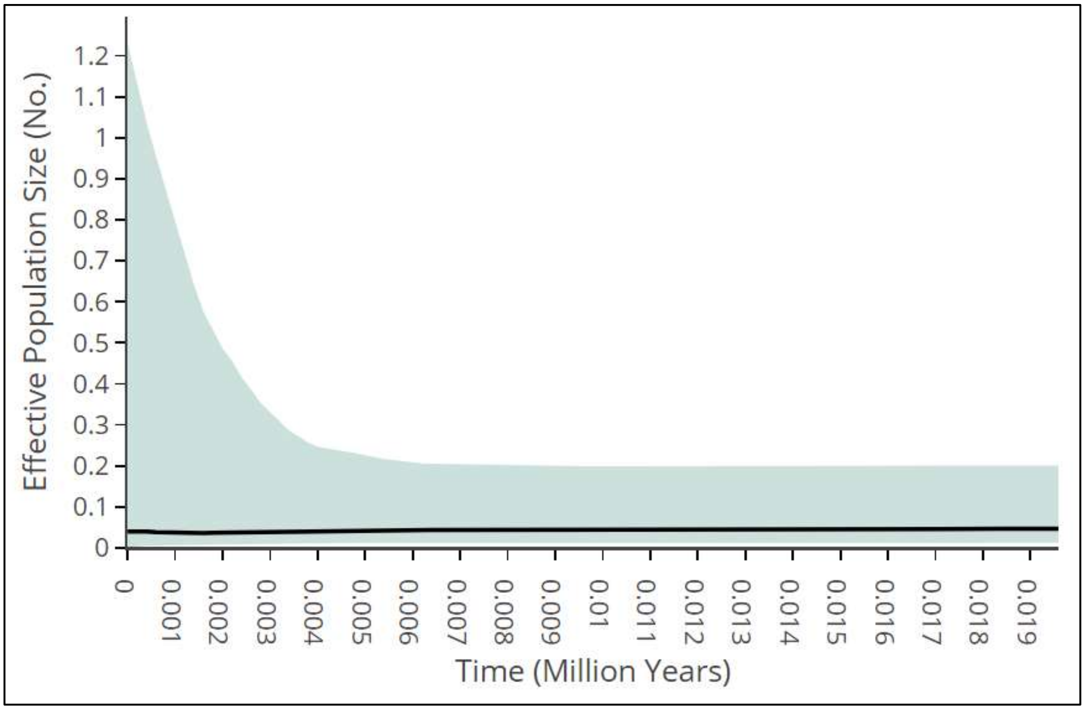


Figure 18: A Bayesian skyline plot analysis for *A. caliginosa* in Les Deux Alpes indicating a stable population. The black line indicates the median plot for the analysis and the blue shading the range between upper and lower confidence interval. The upward trend to the upper confidence interval suggests a possible population expansion in the last 3000 years.

A Bayesian skyline analysis was performed on the individual samples for *A. caliginosa* from Les Deux Alpes (Figure 18). It indicated a stable population size for most of the duration of the plot. The rising upper confidence interval in the last 4000 years ago suggests the onset of a population expansion.

Similarly, *L. terrestris* individual samples underwent Bayesian skyline analysis and showed a large-scale two stage population decline from ~840,000 to ~320,000 years ago and from 120,000 years to present, thereby reducing the effective population size to a twentieth of the starting population one (Figure 19).

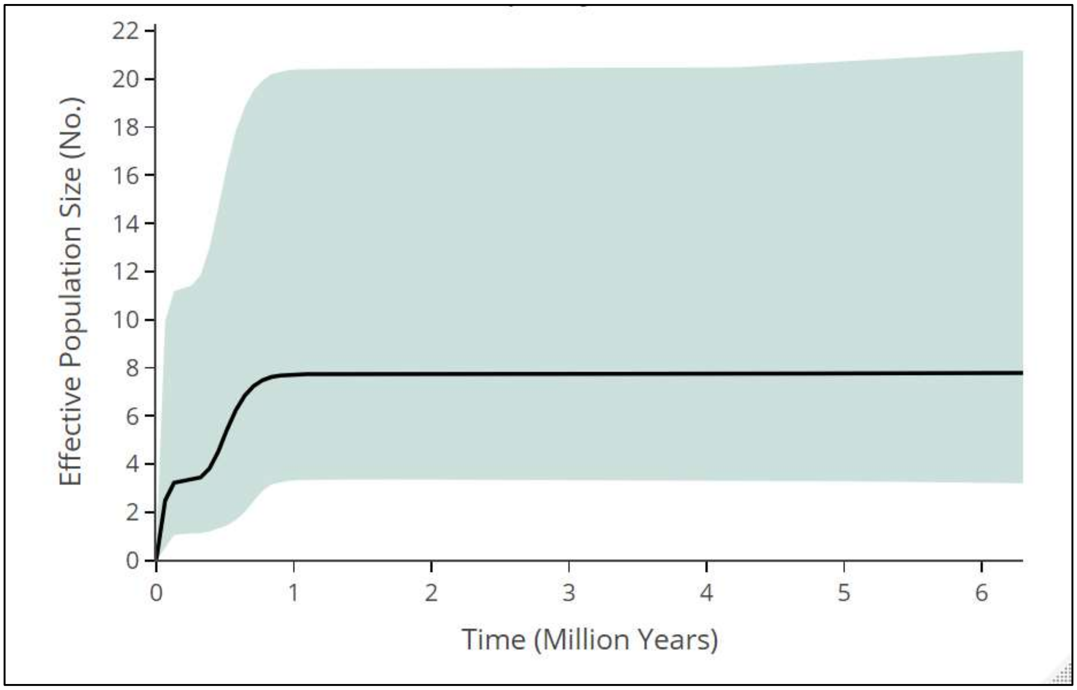


Figure 19: A Bayesian skyline plot analysis for *L. terrestris* in Les Deux Alpes indicating a long-term stasis, and decline in the last million years. The black line indicates the median plot for the analysis and the blue shading the range between upper and lower confidence interval.

Table 11: Calculated values of Haplotype diversity for *A. caliginosa* and *L. terrestris* for samples from Les Deux Alpes with global reference samples.

	Les Deux Alpes, France	
	A. caliginosa	*L. terrestris*
Individuals	**101**	**57**
Nucleotide sites	**443**	**551**
Haplotype diversity	**0.7543**	**0.8747**
Pi	**0.01710**	**0.11637**

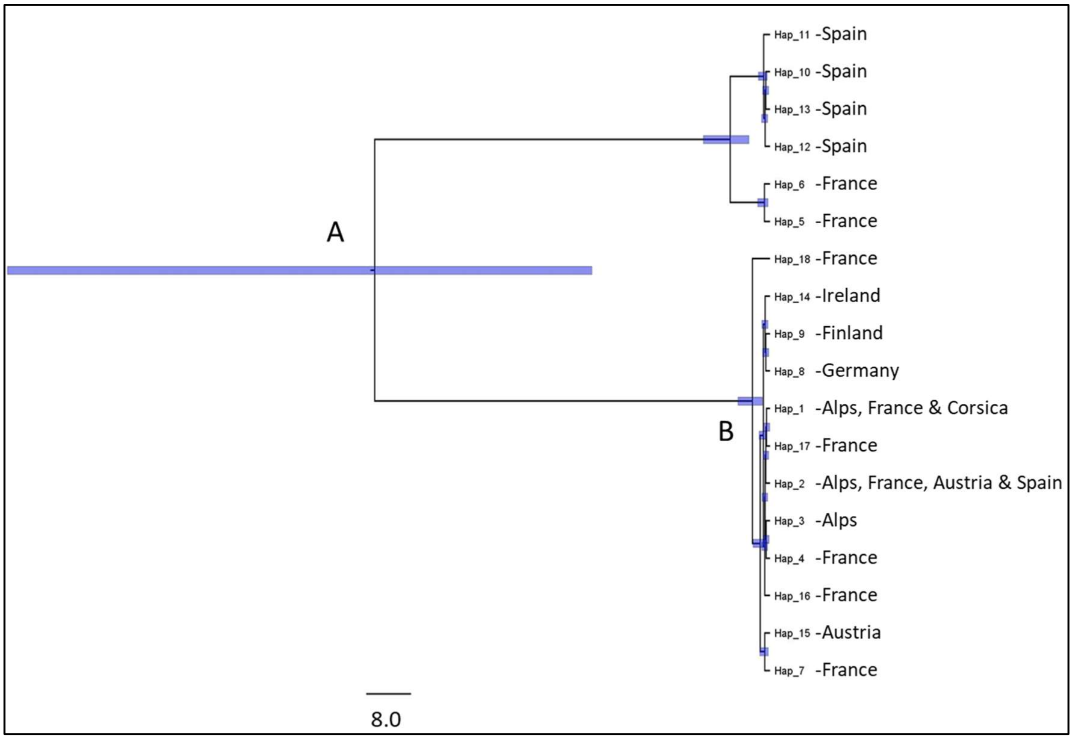


Figure 20: Dated phylogeny estimating divergence time for *A. caliginosa in Les Deux Alpes*, calculated using COII haplotype groups with trees generated on BEAST v1.8.4. and visualised with TreeAnnotator v2.4.8. Time scale is in 8 million years and blue node bars indicate confidence interval of branch splits. Split 'A' occurred 31.8 and 136.2 million years ago, while split 'B' occurred 1.3 and 5.7 million years ago.

Haplotype groups were generated for *A. caliginosa* and *L. terrestris* and their diversity assessed as part of their generation (Table 11). A Beast analysis was performed on the haplotype groups of *A. caliginosa* from Les Deux Alpes and the Global reference samples (Figure 20). Haplotypes from Les Deux Alpes (1-3) are interspersed with haplotypes from across Europe (4, 7-9 and 14-18) that split (Figure 20, B) between 5.7 and 1.3 million years ago. The tree shows the same second group of *A. caliginosa* haplotypes from Northern Spain and the French Pyrenees splitting between 31.8 and 136.2 million years ago (Figure 20, A).

A Beast analysis was performed on the haplotype groups of *L. terrestris* samples and *L. castaneus*, showing a far more complex branching of the tree (Figure 21). Haplotypes from the Les Deux Alpes samples (1-6) split into four deeply divergent branches with haplotypes 6 and 5 more closely linked with the global reference samples from Canada, Austria and Ireland by between 4 and 7.7 million years. Haplotype 4 diverges between 19.1 and 51.7 million years (Figure 21, C) further back before haplotypes 1 and 2 (*L. castaneus*) branches between 38.1 and 113.3 million years ago (Figure 21, B). Finally, haplotype 3 branches off between 45.1 and 137.6 million years ago (Figure 21, A). The very large node bars indicating confidence values for the divergences could mean haplotypes 1-4 branch off differently as all overlap by 13.6 million years.

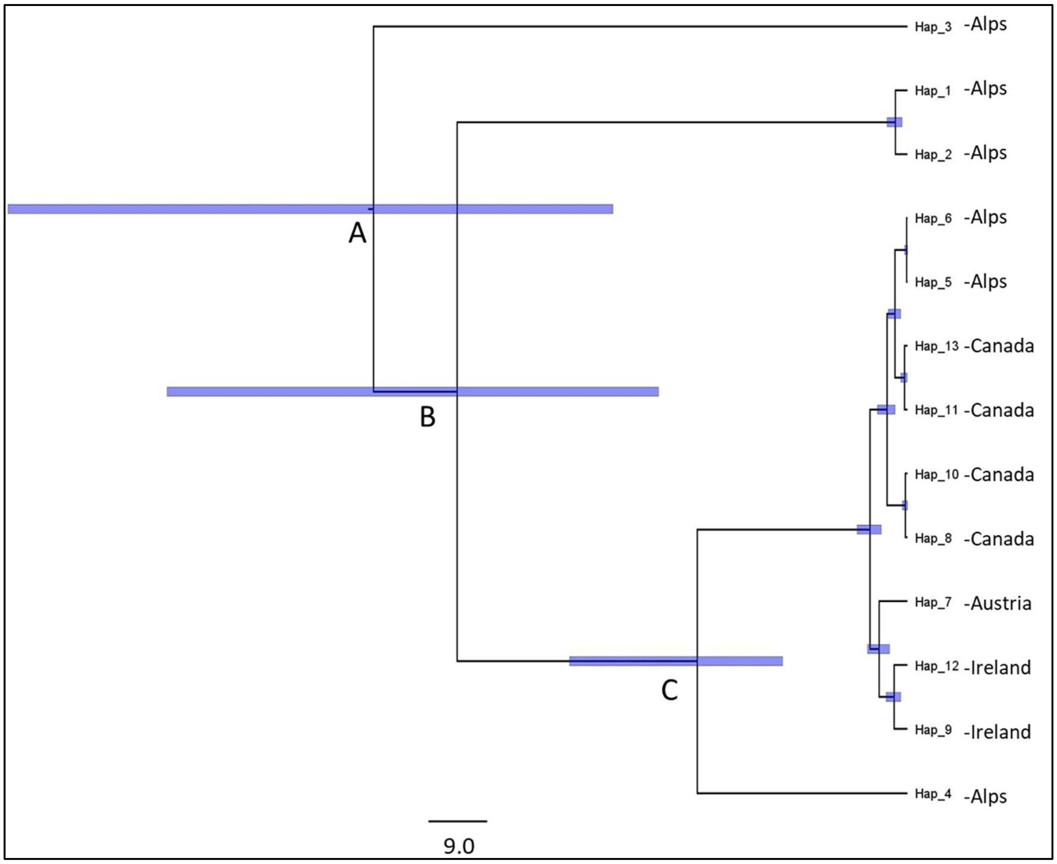


Figure 21: Dated phylogeny estimating divergence time for *L. terrestris in Les Deux Alpes*, calculated using COII haplotype groups with trees generated on BEAST v1.8.4. and visualised with TreeAnnotator v2.4.8. Time scale is in 9 million years and blue node bars indicate confidence interval of branch splits. Split 'A' occurred 45.1 and 137.6 million years ago, while split 'B' occurred 38.1 and 113.3 million years ago and split 'C' occurred 19.1 and 51.7 million years ago.

Table 12: Mantel test values calculated for stages of geographical distance from Les Deux Alpes. No test shows a strong correlation between genetic and geographical distance.

	Genetic & geographic distances r value	*P value for observing ≥ to r value*	*P value for observing ≤ to r value*
A. caliginosa Alps	0.07	0.01	0.99
A. caliginosa Global	0.00	0.20	0.80
L. terrestris Alps	0.09	0.08	0.92
L. terrestris Global	0.06	0.14	0.86

A mantel test was performed at three stages for *A. caliginosa* (Les Deux Alpes Les Deux Alpes and global reference samples) and two stages for *L. terrestris* (Les Deux Alpes, Les Deux Alpes and global reference samples) (Table 12). None of the tests showed a strong R-value for the relationship between genetic distance and geographic distance.

4.5. Discussion

4.5.1. **General discussion.**

The purpose of this chapter was to assess the species identified in Chapter 1 for their population structure and history as a means to select a single species at a single location that can best be utilised to assess the fundamental question of this thesis: is there adaptation or acclimatisation of earthworms to high altitude. For this, the ideal worm population i) has low genetic diversity that could complicate assessment of gene expression analysis, ii) can be collected in sufficient numbers iii) without the need for genetic barcoding to confirm species identity.

With colonisation of a new environment, particularly island environments with no competitors, population expansion might be expected as the population grows to fill new confines. However, neither mismatch distribution calculated for Pico (*A. caliginosa* and *L. terrestris*) matched the expected pattern for population expansion and both showed multimodal distributions. The Bayesian skyline analyses done for the two populations further supported this, with both populations showing throughout most of their history a demographically stable population size and neither showing a population expansion but rather both suggesting a recent population bottleneck. For *A. caliginosa,* this decline to one third of the starting diversity occurred over the last 35,000 years while for *L. terrestris* the decline was more dramatic, dropping to one hundredth of the starting population over the last 100,000 years. The period of the last glacial period from around 155,000 to 11,000 years ago (last glacial maximum 26,000 years ago), matches closely with the start of the calculated *L. terrestris* population decline, supporting the hypothesise that the species suffered was a direct consequence of the glacial period (Sibrava et al. 1986). Indeed, this would indicate either the population found on Pico was subject to this environmental pressure, or more generally, the Western Palearctic was subjected.

The decline of *A. caliginosa* population is smaller and occurred more recently with the start of the decline, ~20,000 years ago and shortly after the last glacial maximum. This could indicate the Pico population was more protected than other populations or that the species is more resilient to icy conditions than *L. terrestris*. While both species are known to produce viable ice-hardy cocoons, *A. caliginosa* has been observed burrowing in frost layers (Nuutinen and Butt 2009). Despite the ice having receded over 10,000 years ago, this does not seem to have translated into a big population expansion. While there is a small increase to the upper confidence interval in *A. caliginosa*'s Bayesian skyline analysis, this might mean that population expansion and its diversity is only just beginning despite the widespread presence of the species across the islands of Pico and São Miguel.

In Les Deux Alpes, the mismatch distribution of *L. terrestris* shows a similar multimodal distribution not matching expected pattern of population expansion. For *A. caliginosa,* the multimodal distribution does match more closely this expected pattern indicating a potential population expansion. There is no apparent expansion seen in the Bayesian skyline analysis, although the upper confidence value does increase several fold. In contrast, *L. terrestris* sees two large population declines in the Bayesian skyline analysis. The more recent decline started around 120,000 years ago and extended to present matching the last glacial period as previously discussed. The Initial slower population decline from 840,000 to 320,000 years ago would cover several glacial periods across the Pleistocene, including Günz, Mindel and Riss and beyond (Sibrava et al. 1986). These glacial periods included significant periods of higher temperatures, where one might have expected some population recovery. This might could indicate a slower rate of acquiring genetic variation within the population and that much of the population diversity observed, that we would expect to continuously build up over time, was acquired across the large period of warmth in the Cretaceous and early tertiary ages of earth (5-85 million years ago). However, this is a highly speculative hypothesis, as the evolution of earthworms remains a largely unknown and highly complex area of research. Earthworms seem capable of producing fertile offspring between highly diverse lineages and the cryptic nature of many of their genomes produces a tangled web of an evolutionary history that is difficult to tease apart without the additional factor of introduced earthworms (James 2004; Hendrix et al. 2008; King et al. 2008).

Dominguez *et al.* developed a detailed attempt of linking lumbricid earthworm evolution with estimated geological age (Dominguez et al. 2015). They present evidence of the emergence of the lumbricid family 125 million years ago after the break-up of Pangaea. In particular, *A. caliginosa*'s genetic split from *A. tuberculate* 10 to 20 million years ago and they diverge from *A. trapezoids, A. longa* and *A. nocturna* between 25 and 40 million years ago (Dominguez et al. 2015). In our analyses, we show divergences of lineages, approximately 1 to 8 million years for most of the samples tested. This would fit within the *A. caliginosa* branch shown in Dominguez *et al.* however, our analyses also showed a split of Iberian samples from central Europe at the low estimate of approximately 28 to 32 million years ago. This clearly does not fit the analyses by Dominguez *et al.* There are several possibilities that could explain this. The BEAST analysis working on the estimated mutation rate has over-estimated the age of divergence, the initial cataloguing of these samples in the NCBI database is not accurate or the classification of *Aporrectodea* species is not fully understood. All three arguments have their merits and it is probably at the least a combination of over-estimation by BEAST and our difficulty in classifying species. Dominguez *et al.* discuss their own limitations of their time-rooted analyses and conclude that despite the lack of fossil evidence to support, there is still some evidence through geological events. More challenging is solving the species complex. Pérez-Losada *et al*. reported the existence of cryptic diversity within the taxa (Pérez-Losada et al. 2009). This seems to have sparked a series of publications arguing against cryptic diversity in species and in support independent evolutionary units (Blakemore et al. 2010; Martinsson et al. 2017; Marchan et al. 2018). However, the earthworms from both the Azores and the Alps are situated in the larger group of samples and as the focus of this research; we will not delve further into this larger issue. What our BEAST analysis does support is how closely genetically linked all of the *A. caliginosa* samples are.

Dominguez *et al.* did not include *L. terrestris* in their analysis, instead the closest species split they analysed was *L. castaneus* with *L. rubellus* 30 million years ago. The BEAST analysis for *L. terrestris* from Pico indicated the species diversity compared with European and Canadian samples was within 10 to 20 million years ago. This would fit within a model where *L. terrestris* and *L. castaneus* are the closest relatives, but this picture become more obscure when looking at the BEAST analysis of *L. terrestris* from Les Deux Alpes. Here, even taking the low estimate for the most divergent haplotype group where *L. terrestris* and *L. castaneus* are diverging over 45 million years. If we were to exclude haplotype group 4 from the tree, the divergence timescales for the remaining haplotypes would fit a reasonable pattern of divergence ages. All of the samples in haplotype 4, while the best nucleotide blast match was *L. terrestris*, it is noticeably lower than other matches at 84% and 87%, suggesting that these could reasonably not be *L. terrestris*, but a previously un-defined species that diverged around the same age as *L. castaneus*.

As with *A. caliginosa*, *L. terrestris* does not form distinct a lineage in either the Azores or the Alps when compared with Central European samples as assessed through the COII gene. What it suggests is following the recession of ice from the last glacial maximum, populations of earthworms repopulated from a single source. Our data and from data in other research studies, supports the assessment that the primary source is Apennine peninsula rather than the Iberian Peninsula which populated the westernmost edge of the Western Palearctic.

4.5.2. **Selecting a species for acclimatisation and adaptation analysis.**

The following chapters of this thesis aim to perform three sets of analysis to identify signals of acclimatisation and adaptation within an earthworm population. This will be broken down into genome assembly, gene expression analysis and SNP analysis. Having a low population diversity reduces the likelihood of gene expression analysis variation being as a result of species divergence (Pilpel 2011). Populations that are closely historically linked reduces the number of accumulated nucleotide polymorphisms that can occur as a result of random mutation rate over time, and instead, areas under selective pressure should be identified in SNP analysis.

Of the two species assessed, *A. caliginosa* displayed a significantly lower level of population diversity than observed in *L. terrestris* in both sites of Pico and Les Deux Alpes as assessed via BEAST analysis. In both sites the genetic diversity as assessed by COII barcoding is quite limited for *A. caliginosa* and separation of individuals through solely the single barcode is minimal. In contrast the separation in *L. terrestris* in both sites is high and individuals are found in multiple branches of the population tree. From this evidence, *A. caliginosa* presents itself as the more suitable species for use in gene expression analysis.

To identify which location to select *A. caliginosa* from, two factors were taken into consideration. The first was the history of the population structure as assessed through Bayesian skyline analysis and the dated phylogeny estimates. However, neither assessment provides a conclusive measure for separating *A. caliginosa* populations. With both sites, there has been a population bottleneck since the last glacial maximum and there has been a very limited increase in diversity between populations. This means, that through this measure, there is very little to select a site of preference.

The second factor to consider is ease of sampling. This itself is broken down to two aspects. Primarily, how many individuals can be collected within a given time at a sampling site. Sampling up a mountain is inherently challenging for access and safety, limiting the duration in a day sampling of individuals can safely be performed. For the purpose of cultivating live populations a large sample size (>100) is required to account for any mortalities and allow for continued population survival. Of the worms sampled, ideally all can be immediately speciated by sight. This prevents under sampling due to bringing back a mixed and possibly unbalanced set of various species, and also removes the complication of species identification for downstream analysis. Earthworms that have very similar appearances can be challenging to morphotype, even under microscope and all effort should be made to remove the need for genetic barcoding via tail snips as a means of identification as this can impact gene expression analysis, even after a long habitat normalisation period. Sampling earthworms in Les Deux Alps was not only challenging to collect a large number of earthworms of the same species, but also challenging to collect enough species in a small enough area. Further to this as identified in Chapter 1, many species of earthworms were sampled, including the similar looking *A. longa* and *A. rosea*. Further to this, the narrow sampling window between snow cover at high altitude and dry soil at low altitude add an extra sampling pressure. While the sampling of the target species of *A. caliginosa* at high and low altitude in Les Deux Alpes remains an interesting biological aim, it is more unpredictable for the safe sampling of live individuals. In Contrast, while a more challenging location to reach, the high altitude site of Pico, like the low altitude site of Pico, has a large and easily collected population. At high altitude, no species was collected that looked like *A. caliginosa* and at low altitude at certain sites, particularly those of grass environments, mono-species populations of *A. caliginosa* could be identified. Sampling of *A. caliginosa* at Pico therefore provides a more reliable means for selecting a large number of live individuals for long term cultivation.

Geologically and biologically, much of Pico, and the Azores are very young. Pico as the youngest island at only 250,000 years old has had little time to develop a diverse and unique ecosystem, gaining a low level endemism that has been itself cut further by human activities of destruction and introduction of invasive species (Borges et al. 2010). While the oldest island in the Azorean archipelago is over 8 million years old (Santa Maria), and likely helped to seed flora and fauna upon Pico, it could have looked similar to Ascension Island (approx. 1 million years old) for much of the islands' history, which was described by Darwin in 1836 as "arid" and "treeless". The low levels of biological seeding through natural and human intervention in the case of Pico, reduces the complexity of earthworm species on the island. The evidence found within this chapter and the previous chapter suggests that very few linages of *A. caliginosa* were introduced to the island and that this happened very recently, with highly likelihood through introduction by humans. Further research into species distribution on oceanic islands has linked presence with human activities such as roads and farming (Paudel et al. 2016). The multiple introductions of lineages from across Europe adds support to the introduction via human method of island colonisation and combined with historical evidence of human settlement activity, the lower estimates of human arrival to the islands, *A. caliginosa* has had less than 600 years to colonise the island from low altitude, to high altitude. If the last major eruption did cause a local extinction level even around the summit of Mount Pico, migration and colonisation is limited to around only 300 years. If the species has more than just acclimatised to the high altitude, it would signal a very accelerated adaption to the challenging conditions that could have important implications to the species survival and the survival of the ecosystem in the rapidly changing environments brought on by manmade climate change.

4.5.3. **Concluding remarks.**

From the analysis presented here, two selections have been made that impact the future research of this thesis. In the first, *A. caliginosa* will be sampled as it has the lower level of genetic diversity between sites. In the second, Pico will be selected as the sample site as it provides the most reliable location for safely collecting a large number of *A. caliginosa* without collecting similar looking species (Kelly 2019).

4.5.4. **Future analyses.**

In the following chapters, we aim to collect sufficient numbers of *A. caliginosa* from Pico at high and low altitude and cultivate populations in identical conditions to normalise gene expression prior to gene expression analysis. Simultaneously *de novo* genome assembly will be performed on the species to generate the first ever genome for the species that will be used for the gene expression analysis and SNP analysis

5. Genome of *Aporrectodea caliginosa*.

5.1. To sequence a worm.

5.1.1. **Why sequence an earthworm.**

In Chapter 3.4.2 and 4.4.2 we discussed the suitability of *A. caliginosa* as a species for more in depth genetic analyses due to the lower genetic mitochondrial diversity found between high and low altitude populations. To understand whether populations are more adapted or acclimatised to high altitude we need to look across the populations' genome for DNA polymorphisms which would indicate that worms have acquired specific adaptations over time to enable them to better survive in the high altitude niche. We can also perform transcriptomic analysis of gene expression between populations from high and low altitude. Both populations would spend time at sea level oxygen and room temperature before being exposed to conditions replicating high altitude. This allows us to determine if both populations perform similarly when exposed, that *A. caliginosa* is able to acclimatise intrinsically. If, however, for example the high altitude population indicates faster acclimatisation, then epigenetics could be a contributing factor. Both assessing DNA polymorphisms and transcriptomic assessment can be performed without a solved genome but have limitations. By first solving the genome, stronger markers for determining DNA polymorphisms can be designed and RNAseq can be performed with gene objects to map reads to. This allows for a cheaper form of sequencing and a more accurate transcriptome being generated. For *A. caliginosa*, genome size is estimated at 650 Mbp with a Chromosome number (2n) of 36 (Gregory and Herbert 2002; Kashmenskaya and Polyakov 2008) (Figure 22, Table 13).

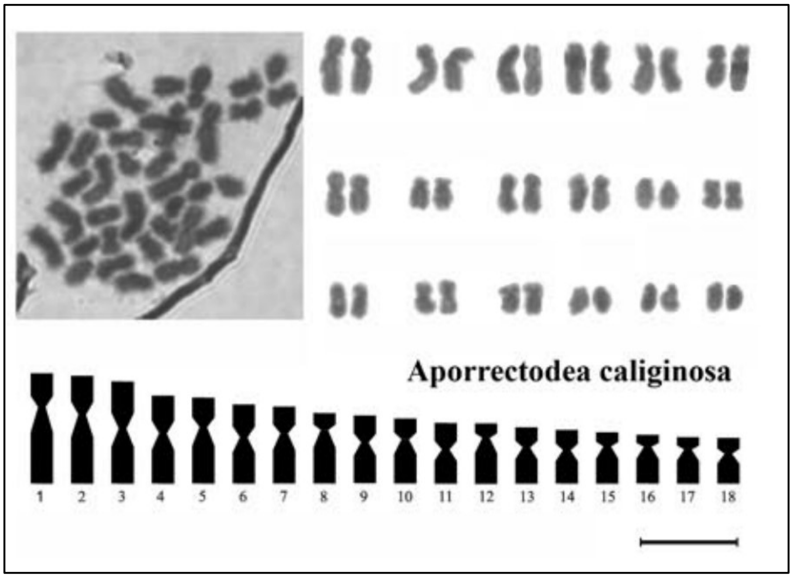


Figure 22: Karyotype of A.caliginosa (Bar 10um) generated by (Kashmenskaya and Polyakov 2008).

Table 13: Estimated genome size per chromosome based on relative length measured by(Kashmenskaya and Polyakov 2008). Relative length was measured in micrometres.

Chromosome number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Relative length	8.55	8.27	7.99	6.88	6.77	6.21	6.04	5.54	5.32	5.06	4.78	4.75	4.42	4.15	4.02	4.02	3.63	3.6
Relative size Mbp	27.8	26.9	26.0	22.4	22.0	20.2	19.6	18.0	17.3	16.4	15.5	15.4	14.4	13.5	13.1	13.1	11.8	11.7

5.1.1. **Challenges of sequencing earthworm genomes.**

Few attempts have been made to generate a genome for earthworm species', though there have been a few published low N50 genomes for *Eisenia fetida* (1 -9 Kb N50) (Zwarycz et al. 2015; Bhambri et al. 2017). This is in part due to the limited number of research teams globally attempting this challenge and the difficulty other highly heterozygous genomes have presented. There has been a large increase in diploid genomes publications in the last decade, but this has largely been exclusive of highly polymorphic species (Safonova et al. 2015). Several new computational approaches have been develop to overcome this issue with the use of de Bruijn graphs including dipSPades and Platanus (Kajitani et al. 2014; Safonova et al. 2015). Unlike assembling two highly similar haplomes with less than 0.1% nucleotide variation, genome assembly in haplomes with 0.4-10% nucleotide variation fail, resulting in mixtures of haplocontigs and consensus contigs (Safonova et al. 2015). This is further exacerbated by the non-random distribution of polymorphisms along the genome where areas of high conservation and areas of high diversity can be found. Diversity can be so high between haplomes that functionally the genome appears twice the size, thus requiring twice the sequencing coverage. Many of the assembly challenges are born out of the small read lengths that must be stitched back together, however methods for resolving this are still in the pioneering stage and require high capital and time investment to solve sequencing and assembly difficulties as they arise.

5.1.2. **Aims of genome of *Aporrectodea caliginosa* chapter.**

In this chapter I will aim to identify the optimal method of sequencing and assembly to generate a high quality, high N50 genome for the target species *A. caliginosa* for the purpose of gene expression analysis and SNP analysis in subsequent chapters. This will utilise a combination of sequencing methods; Nanopore, Chromium 10x and short read data. Short read data will be used to error correct the more error prone long read nanopore sequencing data. This can then be assembled using a variety of sequence assemblers and the best assembly will undergo further improvement with scaffolding data from 10x chromium sequencing data. I will also utilise several measures to assess the quality of the genome assembled and perform an initial identification of gene objects of the generated genome.

5.2. Technical approaches to *de novo* genome sequencing.

5.2.1. **Large fragment DNA extraction.**

Baring a few exceptional giant earthworms, most are quite small. The largest common European earthworm is normally around 20-25 cm (Sims and Gerard 1999; Sherlock 2018). In the case of *Aporrectodea caliginosa*, the length of an adult can range from 4-18 cm, though normally this is less than 10 cm (Sims and Gerard 1999; Sherlock 2018). While this would suggest that there is plenty of tissue to extract DNA of high quality, for genome sequencing, it is advisable to carefully remove the gut without perforation as this contains bacteria that will contaminate the worm's own DNA. It is also advisable to avoid the area from the Clitellum to the male pore as this can contain sperm from other earthworms (Sims and Gerard 1999). With all these exclusions, only a small area of muscular tissue remains. The muscular tissue is hardest to digest and does not yield a high density of DNA.

Modern fast column based DNA extraction and purification methods fail to preserve sufficient high fragment sizes while traditional preservation methods like freezing tissues at -80°C in ethanol, DMSO or RNAlater all contribute to DNA fragmentation. To date, the use of large quantities of fresh tissue digested in ATL/Proteinase K buffer, combined with lengthy Phenol-Chloroform/DNA precipitation clean-ups provide the most reliable methods for isolating long fragment DNA. Typically, even with this method it is difficult to achieve yields greater than 16 µg DNA for *Aporrectodea caliginosa*.

5.2.2. **Sequencing generation**

5.2.2.1. Short read sequencing (Illumina).

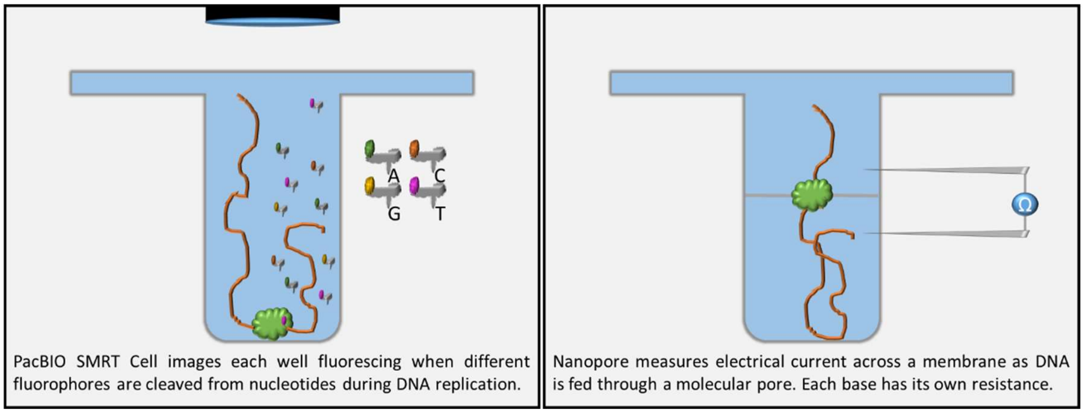
Short read sequencing is the simplest and cheapest method for Next Generation Sequencing. Short fragments of DNA undergo PCR for adapter ligation. The library is loaded onto a flow cell where the fragments hybridise to the surface and amplification can occur through bridge amplification. Using fluorescently labelled nucleotides and further amplification occurs, and with each additional nucleotide addition the flow cell is imaged to identify the added nucleotide. For each short read sequence, the adapter is removed and the short reads are then mapped to each other (Heather and Chain 2016).

This method is quite reliable and efficient for species with very low haploid diversity like mammals or small genomes like bacteria, e.g *E. coli* at 4.6 Mbp. It is also fast, cheap and with high sequencing accuracy. Genomes can be sequenced for under $1000 (Ari and M. 2016). However, this technology particularly struggles with organisms with high genetic diversity and with large stretches of repetitive DNA. Assembly of short reads cannot reconcile the disparity in areas of high haploid diversity and as such tends to either fail to include areas in the assembly or incorrectly place these stretches. With repetitive DNA, the sequences all map to one location rather than a longer stretch. This all results in poor assemblies, often characterised with poor N50 sizes.

Long read length methods like Nanopore and PacBio offer one method for providing a scaffold for solving the high heterozygosity of earthworms. These methods can generate read length up to 60 Kbp in size for the PacBio and in theory over 1 Mb for Nanopore, (although in practice this is closer to 60 Kbp due to the difficulty in producing long fragment DNA for sequencing) (Kajitani et al. 2014; Reuter et al. 2015; Jain et al. 2018). These ultra-long reads in theory bridge the areas of 'bubbling' and branching that would be created during short read assembly and indicating the structure of both highly polymorphic haplomes. The drawback of this technology currently is the high base call inaccuracy. High read coverage can help to improve the accuracy to as much as 99.88% and still yield NG50 contig sizes over 6Mb (Jain et al. 2018). Another method is the mapping of high read accuracy short read Illumina data onto the scaffold created in the long read sequencing, which can generate genome read accuracy of 99.99% (Madoui et al. 2015; Tan et al. 2018).

5.2.2.2. PacBio.

PacBio utilises Single molecule real time sequencing (Figure 23). This method uses the cleaving of fluorescent labels on nucleotides during DNA replication of a single DNA molecule with DNA polymerase in individual wells. This method is rapidly developing, in a few years it has progressed from allowing long read lengths up to 40,000 bp and an output of 400 Mb (Ari and Arikan 2016), to 500,000 bp read lengths and 20 Gb of output data per SMRT cell (PacBio 2018). Read length is normally limited by DNA fragment. The very high cost per Gb is still the primary drawback to this method as 10 fold coverage of *A. caliginosa* costs around $10,000. Moreover a recent paper examining metabarcoding in Fungi and eukaryotes identified discrepancies and primer biases that are generated from the library preparation on which the technology relies (Tedersoo et al. 2017).



PacBIO SMRT Cell images each well fluorescing when different fluorophores are cleaved from nucleotides during DNA replication.

Nanopore measures electrical current across a membrane as DNA is fed through a molecular pore. Each base has its own resistance.

Figure 23: Comparison of methods, PacBio sequencing and Oxford Nanopore. PacBio utilises fluorescence to sequence large fragments of DNA (now using looped large fragment DNA to improve sequence quality for each read), while Nanopore relies on the fluctuation of resistance across a membrane as large fragments of DNA pass through a single protein pore.

5.2.2.3. Nanopore.

Oxford Nanopore's MinIon works by measuring the change in electrical current as single strand DNA is translocated through a molecular pore. Each base generates its own level of resistance which can be measured. The technology behind this has been slow to develop and Nanopore remains a largely experimental method of sequencing. Library preparations for flowcells require milligrams of high quality DNA, (although new library technology releases are set to lower this), can be a large hurdle to non-clonal cells or small organisms where only small quantities of DNA can be sourced. The technology will also preferentially sequence short reads in accordance with the kinetics of Brownian motion (Serag and Habuchi 2017). Lumbricidae genomes sizes range between estimates of 430 Mb in *L. rubellus* and 1.24 Gb in *Dendrobaenea rubidia* (Gregory and Herbert 2002). However, the genetic diversity between haploid copies of the genome in earthworms is so high, it effectively doubles the DNA required to be sequenced and assembled, requiring 13 Gb of data to be generated to give just 10 fold coverage of each haploid which can act as an assembly scaffold for short read correction (Gregory and Herbert 2002). While Oxford Nanopore advertised 10-20 Gb of DNA sequence data per flow cell (Mk 1 R9), average yields per flow cell can be often closer to 2.3 Gb (Jain et al. 2018; Oxford-Nanopore-Technologies 2018b). The company now offer a new flow cell (Rev D) advertised at 30 Gb. Despite the relatively low cost of each flow cell (<$1,000) the need to use multiple cells to achieve the required coverage makes this still quite an expensive technology. This method also has the drawback of requiring high input DNA quantities, (400 ng for the Rapid Sequencing method and 2000 ng for the Ligation Sequencing method). For small earthworms like *A. caliginosa* and *E. fetida*, retrieving such high yields can be a challenge, especially when trying to exclude contamination from gut bacteria. In contrast Illumina short read sequencing methods require input DNA of only around 100ng for library preparation (enough for several sequencing runs), and an Illumina 2x300 V2 chip will reliably produce more than 5 Gb of output data.

Quantum Biosystems, Genia Technologies (acquired by Roche), Quantapore are all companies currently in development of their own Nanopore technologies (Steinbock and Radenovic 2015; Ari and M. 2016; Abedini-Nassab 2017). Little detail has been known about these systems, however Quantapore is known to be using fluorescent imaging of DNA as it passes through a pore, removing the need for complex and expensive electronics and pushing the numbers of pores from thousands to hundreds of thousands and dropping the expected consumable cost. Though the company has been operating with minimal public data, they are expected to be entering beta testing phase shortly to build up a publication record (Quantapore 2018).

Table 14: Comparison of input and outputs between Illumina Miseq, PacBio and Oxford Nanopore Minion based on 2019 quotations.

Sequencer	*Cost per chip*	*Sequencing output*	*Cost of library prep*	*Input DNA for library*	*Use of library*
Illumina	MiSeq v2 2x150 £873.81	Reliable 5.1 Gb	£18.60 per sample	~100ng	Enough for several runs
PacBio	SMRT cell £1,000	5-10		1-20 µg	
Nanopore	Mk 1 R9 £699.37	Variable 1-15 Gb [4]	~£94 per sample	400ng – 3 µg	Total

5.2.3. **Library preparation for scaffolding.**

5.2.3.1. 10x Chromium library preparation system.

This technology encapsulates individual strands of high molecular weight (50 kbp) DNA in individual Gel beads emulsion droplets containing their own reagents in solution (Figure 24). These undergo isothermal incubation to generate barcoded amplicons. The addition of barcodes fragments the large strands, with each containing the same barcode. This is then sequenced on a standard Illumina short read chip and the sequences are mapped back to each other using their barcodes to generate contigs the size of the input DNA strand.

[4] Since the time of writing Nanopore has successively released Flowcells with greater output multiple times during the year. PacBio has also released greater Sequencing Capacity chips.

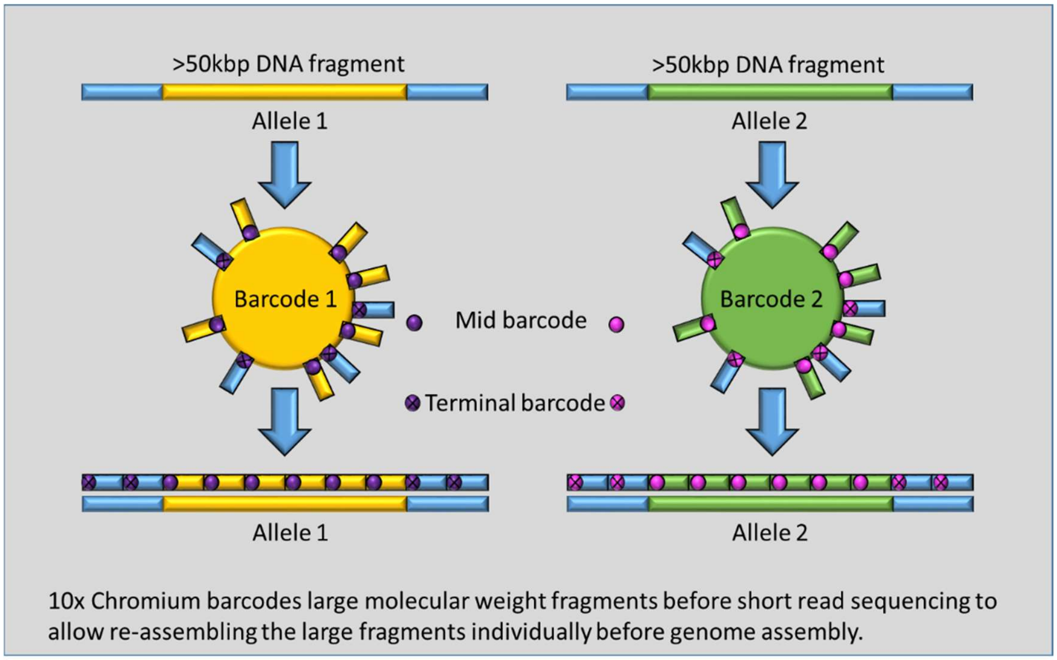

>50kbp DNA fragment
Allele 1
>50kbp DNA fragment
Allele 2
Barcode 1
Barcode 2
Mid barcode
Terminal barcode
Allele 1
Allele 2
10x Chromium barcodes large molecular weight fragments before short read sequencing to allow re-assembling the large fragments individually before genome assembly.

Figure 24: 10x Chromium barcoding method. Large fragments of DNA are individually encapsulated, fragmented and barcoded with individual barcodes. Following sequencing, all fragments with one barcoded are pooled and assembled to reform the large fragment of DNA they came from.

5.2.3.2. Mate-pair.

In Mate-pair libraries, high molecular weight DNA is fragmented, typically into two fragment pools e.g. 5 kbp and 12 kbp. These fragments of known sizes are Biotinylated to circularise the DNA and remaining DNA is digested away. The purified circular DNA is further fragmented, and standard sequencing adapters are ligated on. Each Biotinylated end is a known bp distance apart and this is used to help resolve structural distances (Figure 25). In particular this can help differentiate between two haploid genomes which might have additions or deletions or repetitive regions. Comparisons between Mate pair and 10x Chromium are shown in Table 15.

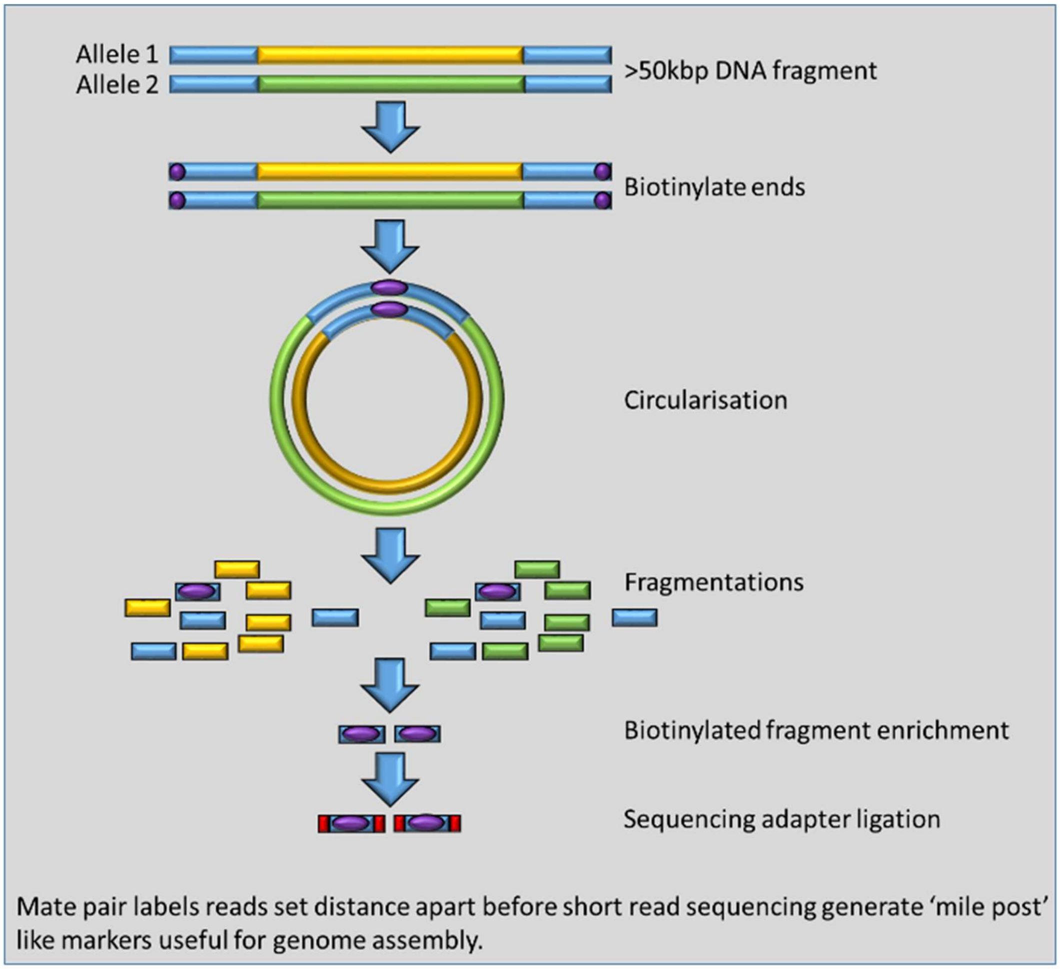

Allele 1
Allele 2
>50kbp DNA fragment
Biotinylate ends
Circularisation
Fragmentations
Biotinylated fragment enrichment
Sequencing adapter ligation
Mate pair labels reads set distance apart before short read sequencing generate 'mile post' like markers useful for genome assembly.

Figure 25: Mate-pair library method. Large fragments of DNA are barcoded at set points to act as 'mile posts' to prevent assembly errors caused by elements such as repetitive DNA and large non-specific insertions.

Table 15: Comparison of library methods between 10x Chromium and Mate-pair based on 2019 quotations.

Library system	*Cost kit*	*Sequenced on*	*Contig size*
10x chromium	~$1500	Illumina 2x150	As large as input fragment
Mate pair	~$3376	Illumina 2x150	Set by user, typically 5 kbp and 12 kbp

5.2.4. **Hi-C scaffolding**

Hi-C scaffolding works on a similar basis to both Mate pair and Chromium 10x barcoding, but requires fixation of DNA prior to extraction from tissue. Tissue is initially homogenised gently to avoid rupture of the cell. DNA strands are then chemically crosslinked to other nearby DNA strands across the genome and between different chromosomes. DNA can then be extracted from the cell before biotinylation and proximity ligation between adjacent sequences. These strands can now be fragmented, purified and sequenced across the ligated ends. Long sections of DNA that follow on from each other are more likely to be physically closer and therefore more likely to be crosslinked, whereas long sections of DNA on different chromosomes or alleles are more likely to be physically further apart and less likely to be crosslinked.

5.2.5. ***De novo*** **Sequencing assemblers and quality control.**

5.2.5.1. Assemblers.

Despite the advances in assembling software, scaffold forming and read correction, remains large hurdles in the generation of large genome and a multitude of assemblers are required in tandem. The following section will aim to breakdown the various bioinformatic tools that are currently used, particularly with the use of long read data from one or multiple sequencers. All assemblers require high capacity computing to effectively perform. The School of Biological Sciences at Cardiff University had at the time of analysis a trio of powerful systems to perform this analysis on collectively called Ysgo. This is split into a 2.7 TB 204 CPU called Mammoth used for high memory intensive jobs, a 128 GB 32 CPU called Defq for smaller jobs, and a 256 GB 64 CPU Power8 system for jobs that can utilise the system architecture.

Assemblers can take a variety of approaches for taking raw reads and stitching them into genomes. Often it can be quite hard to predict how different assemblers will perform in the assembly of differing organisms. Input raw reads, varying levels and distributions of heterozygosity and methylation can play to different assemblers' strengths or fail in their weaknesses and several assemblers will need to be run to identify which will give the strongest assembly.

Canu is an assembler designed to take Nanopore or PacBio long reads and perform an assembly in four main steps. Detecting overlaps in high-noise areas, generating a corrected sequence consensus, trimming the sequences and assembly. A minimum of 20x coverage is advised, however they suggest 30x coverage should be used to maximise the use of long reads in the assembly and improve the N50 contig sizes (Korean et al. 2017). With high enough coverage read correction performed as part of the assembly can achieve around 98% identity. However, the use of short reads can be used to improve error correction on the final assembly generated by Canu. Another fast overlap consensus based *de novo* assembler that works with noisy long reads is Miniasm. It works by mapping all to all in read self-mapping and there is no consensus construction step required in this method as contigs are only concatenated to generate unitig sequences (Li 2016). Working in a similar mechanism to Miniasm with an all to all read alignment without error correction, SMART*denovo* utilises more steps including read overlapping, rescuing missing overlaps and identification of low-quality regions and chimaeras to produce stronger unitig consensus sequences (Ruan 2016). Based on SMART*denovo*, Wtdbg2 is an updated version of the noisy long read assembler. The assembler has a very fast assembly (hours vs day long assemblies with Canu), and works by chopping reads into 1024bp segments, merging segment into vertices and then connecting vertices to generated a 'fuzzy Bruijn graph'(Ruan and Li 2019).

Hybrid assemblers have also been developed for datasets that come from multiple sources like Nanopore, PacBio or short read sequencing. DipSPAdes, is altered to handle highly polymorphic genomes and highly divergent haplomes. It assembles a consensus contig of both haplomes before performing a haplotype assembly. It can perform a hybrid assembly with short reads and PacBio CLR reads (Nurk et al. 2013). Similarly, MaSuRCA can make use of both short read data and long read data in its assembly pipeline. It uses *de Bruijn* graphs and Overlap-Layout-Consensus approaches to build large N50 contig assemblies (Zimin et al. 2013). The system requirements for MaSuRCA are however extensive, requiring high RAM allocation and Terabytes of disk space. Like MaSuRCA, MegaHit uses *de Bruijn* graphs to assemble large and complex metagenomics. It is ideally suited to assembling the genomes of bacterial populations that might be found in soils (Li et al. 2015). Platanus is specifically designed to assemble highly heterozygous diploid genomes and uses several simplifying steps to address bubble removal and branch cutting (Kajitani et al. 2014). This assembler first assembles short reads before scaffolding together with a Mate-pair library and missing gaps are closed. SOAPdenovo performs its assembly from short reads and Mate-pair reads together. It utilises 6 core components of error correction, de Bruijn graph construction, assembly of contigs, read mapping, generation of a read scaffold and gap closure (Luo et al. 2015).

With emerging technologies like 10x Chromium few assemblers have been developed. Supernova is a bespoke *de novo* assembler for 10x chromium data that can create diploid assemblies. Initially de-multiplexing of reads and super contig construction assembles contigs from reads that came from the same library input fragment before consensus construction (Weisenfeld et al. 2017). ARCS LINKS is a pipeline that can be used for the scaffolding of long reads (Yeo et al. 2018). It can be used to take a draft Supernova assembly and scaffold polished Nanopore reads to generate larger contigs.

A multitude of read 'polishers' (Racon, Pilon and FMLRC) and quality trimmers (Trimmomatic) exist to correct high error rates (Bolger et al. 2014; Walker et al. 2014; Vaser et al. 2017; Wang et al. 2018). These are critical in the production of high quality assemblies as the high error rates produced in Nanopore sequencing and the poorer quality ends to short read sequencing can affect the mapping during assemblies.

5.2.5.1. Assembly assessment.

There are several key parameters to consider when assessing the quality of a genome assembly. The initial parameters include: total assembly size, largest contig and N50 ("sequence length of the shortest contig at 50% of the total genome length"). For a largely homozygous organism, we would expect an assembly the size of its haplomes, i.e. 650 Mbp genome and ~650 Mbp assembly. For a very heterozygous organism, the assembly might be as much as twice the size of the haplomes, i.e. 1.3 Gbp for the previous example. Assemblies too small indicate loss of genetic material, while assemblies too large indicate poor contig alignments and genome construction. A large N50 scores how well the contigs have been matched together, the larger, the better. These measures are good indicators of how good the assembly may be, but do not include information about the correctness of the assembly, e.g. we still need to assess the assembly for contaminating sequences and completeness.

BlobTools provides a useful workflow for taxonomic investigation of sequencing data. With the use of coverage data, GC content and a blast database, input sequences filters are generate to remove non-target species sequences and those with low coverage (Figure 26) (Laetsch and Blaxter 2017b). It is an important step to remove these contaminating sequences particularly from parasitic organisms, contaminating soil and gut bacterial. One example is *Verminephrobacter,* which is found in the nephridia of *Lumbricidae* and commonly found in sequencing reads (Pinel et al. 2008). Many of the ultra-long reads achieved from sequencing methods like Nanopore are formed of bacteria which are more resistant to mechanical shearing in the library preparation.

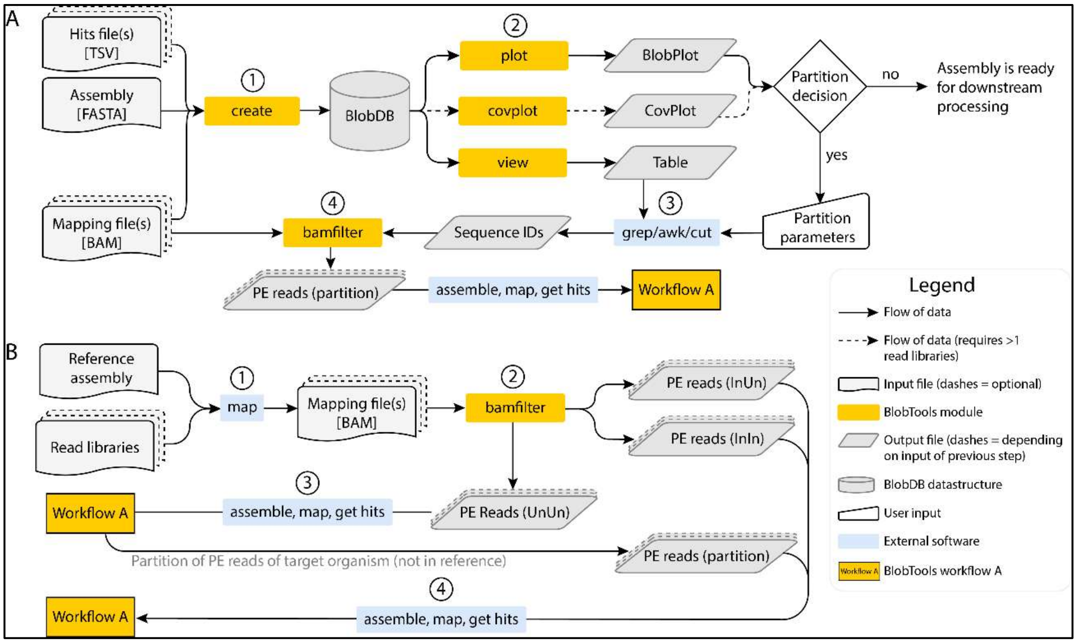


Figure 26: BlobTools workflows summary (Laetsch and Blaxter 2017b).

BUSCO (Benchmarking universal single-copy orthologs) is a means to measuring assembly completeness (Waterhouse et al. 2017). BUSO blasts against the genome assembly for complete, duplicated, fragmented and missing genes that are considered core genes that all organisms critically require and are highly conserved between species. Typically, assemblies should contain greater than 90% complete, (less than 5% missing), BUSCO genes to be considered a complete genome. However, this largely depends on how divergent the target organism is from the reference database with increased divergence reducing alignment accuracy.

Because BUSCO relies on the conservation of core genes, poor assemblies and assemblies with high error rates will be unlikely to be found in blast searches. Consequently, BUSCO analysis on a genome with a high error rate will return only a small percentage of complete core genes.

Recently BlobTools has released an integrated system to incorporate BUSCO analysis and taxonomic data. While still in development and new features are being added, the new version allows for dynamic visual filtering.

5.3. Methods.

5.3.1. **DNA extraction and purification.**

Several individuals of *A. caliginosa* from Pico, Azores site α (2100 m) (Chapter 2.4.1) were depurated overnight before dissection to minimise the risk of bacterial contamination. A large section of muscular tissue from a single individual, 5 segments from behind the mature clitellum and 5 segments from before the tail tip, were dissected with a sterile scalpel carefully ensuring to leave the internal organs (particularly the bacterial containing digestive tract) intact. This tissue was digested for 6 hours in 1080 µL ATL buffer and 40µL Proteinase K at 56°C and 20 µg/mL RNAse with minimal agitation to minimise potential DNA fragmentation.

The digested mixture was mixed with equal volume of Phenol, slowly homogenised and centrifuged at 14,000 rpm to separate phases. Using 1 mL filter tips with the ends cut off to increase the tip diameter and minimise risk of DNA shearing, the upper phase was transferred to a fresh eppendorph tube and mixed again with equal volume of Phenol. The same step was conducted as before and again using Chloroform before mixing and separating in the same method. DNA was precipitated by the addition of equal volume of Isopropanol and 0.2% volume Ammonium acetate. Precipitated DNA was hooked with a sterile glass rod and washed twice with 70% ethanol before re-dissolving in Qiagen Elution buffer for 48 hours.

The DNA in elution buffer was transferred to a G-bioscience mega-long dialysis kit and underwent dialysis in 1L TE buffer for 24 hours. The purified DNA was transferred to a fresh microcentrifuge tube and underwent DNA quality assessment and the sample with the greatest quality selected for downstream sequencing and analysis.

Quality was assessed though Nanodrop, broad range Qubit and running on a 0.5% ultrapure agarose gel with lambda and lambda/HindIII. This was to ensure the sample contained no Phenol, salts, RNA or fragmented DNA.

5.3.2. **10X sequencing and short read sequencing.**

Approximately 2.5 µg of extracted genomic DNA was packaged and shipped to Novogene (HK) for 10x library preparation and sequenced to 50 fold coverage. Additionally, 50 fold coverage of standard Illumina short read sequencing was performed.

5.3.3. **Nanopore sequencing.**

A DNA library for Nanopore was performed as suggested by Nanopore protocol 1D Genomic DNA by ligation (SQK-LSK109) (Oxford-Nanopore-Technologies 2018a). extracted genomic DNA, 5 µg dissolved in in 150 µL of sterile water, was first fragmented to ~16 kbp in a Covaris g-TUBE by spinning at 5,500 RPM in an Eppendorf MiniSpin plus centrifuge for 60 seconds. The fragmented DNA was cleaned using a SPRI bead clean-up and eluted in 60 µL before being assessed quantitatively with a Qubit dsDNA HS assay (Life Technologies) and with a Genomic Tapestation (Agilent).

The fragmented DNA was split into two pools. Half of the preparation (30 µL containing ~1 µg) went into DNA repair and end-prep by NEBNext FFPE DNA Repair Mix and NEBNEXT End repair as described in the Nanopore protocol. The repaired DNA was cleaned with an AMPure XP bead clean-up. The cleaned DNA then underwent adapter ligation with the Adapter Mix and T4 ligase from NEBNext Quick Ligation Module and Oxford Nanopore's Ligation Buffer. Another AMPure XP bead clean-up followed. A Qubit was performed and the DNA to a concentration was between 5-50 fmol.

A fresh Spot on Flow Cell MK1 R9 (RevD) was primed with Flush Tether and buffer avoiding introduction of any air bubbles to the flow cell which could damage the pores. The cleaned DNA was mixed with Sequencing buffer and Loading beads and the 75 µL mixture added dropwise onto the SpotON sample pore. Run time was set for 96 hours to ensure sequencing continued until all pores had died and maximum sequencing was achieved for the flowcell. Full methods are described in Chapter 2.5.1

5.3.4. **Genome assembly.**

Sequence data from the 10x Chromium, Short read sequencing and the three Nanopore runs was uploaded to a Unix based HPC computer cluster running Slurm queue management (Intel Xeon CPU E7-4850 v2 at 2.30GHz, 96 CPU with 2 Threads per core, 12 cores per socket, 4 sockets and a maximum 1.5 TB RAM available for each task) for bioinformatic analysis and assembly. Various assembling algorithms were used with the three datasets, each providing their own benefits and shortcomings. Each assembly was assessed to identify the biggest and most complete assembly. While the Nanopore sequencing totalled approximately 37.8 fold coverage and Canu (v1.8) could have been used to assemble, previous test runs on *Eisenia fetida* did not produce a large or complete assembly when assessed via N50 length and BUSCO analysis (Korean et al. 2017). For each assembler, several iteration of computation were run to identify the parameters that produced the best assembly. Figure 27 attempts to simplify some of the main assembly paths taken and the assembly tools used. Assemblies were computed on Ysgo through a SLURM manager with the commands detailed in Chapter 2.8.

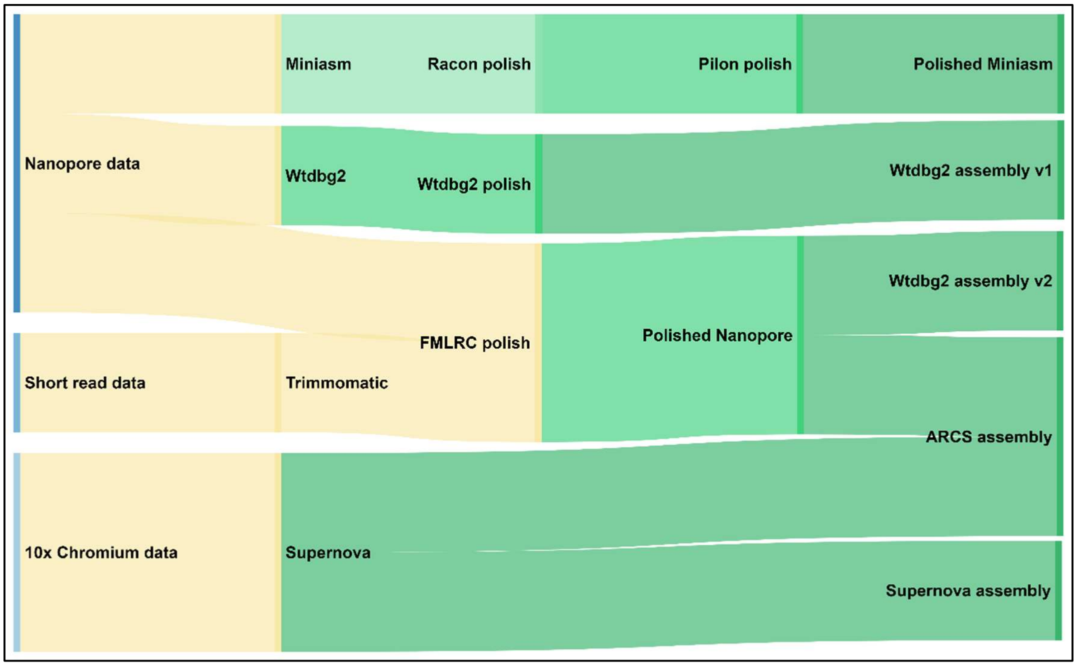


Figure 27: Assembly paths – Different assembly methods tested to produce the best initial assembly.

Each final assembly was assessed in assembly size, N50 and BUSCO, with each assessment influencing which assembler and polishing tool would be most appropriate and produce the best assembly. The Wtdbg2 assembly v2 had the strongest assembly and was chosen for further development after an additional Racon polishing step was performed to validate the contig assemblies (Ruan and Li 2019).

5.3.5. **Scaffolding, gap closing and generating final assemblies.**

To improve the quality of the assembly, three processes were undertaken to improve gene continuity, remove contaminating reads and masking repeats to improve genome size. Figure 28 below indicates the pipeline used. A comprehensive transcriptome for *A. caliginosa* generated from tissue specific assemblies provided by Dr Stephen Short (Cardiff University) was used to improve the continuity of the genomes exons using L_RNA scaffolder (Xue et al. 2013).

This assembly was piped into a novel scaffolder, Nanochrome (Rimmington 2019). This program works to combine low depth 10x Chromium linkage with Nanopore reads and assemblies. This vastly increases the N50 for assemblies but introduces non-ATGC N regions of varying lengths that must be corrected, and gap closed. This was achieved by inputting to LR_Gapcloser (Xu et al. 2018). This program uses the corrected Nanopore reads through three iterations to try and fill in and correct N gap lengths and a final iteration of gap closing with SOAP gap closing was used. The Nanochrome scaffolder, LR_Gapcloser and Soap Gapcloser pipeline was cycled a total of 4 times to allow for new links generated on each cycle to by further scaffolded. The final two cycles of Nanochrome utilised a strict mode to minimise risk of false scaffolding. To validate the final assembly, a final round of Racon polishing was used to break any unsupported contigs.

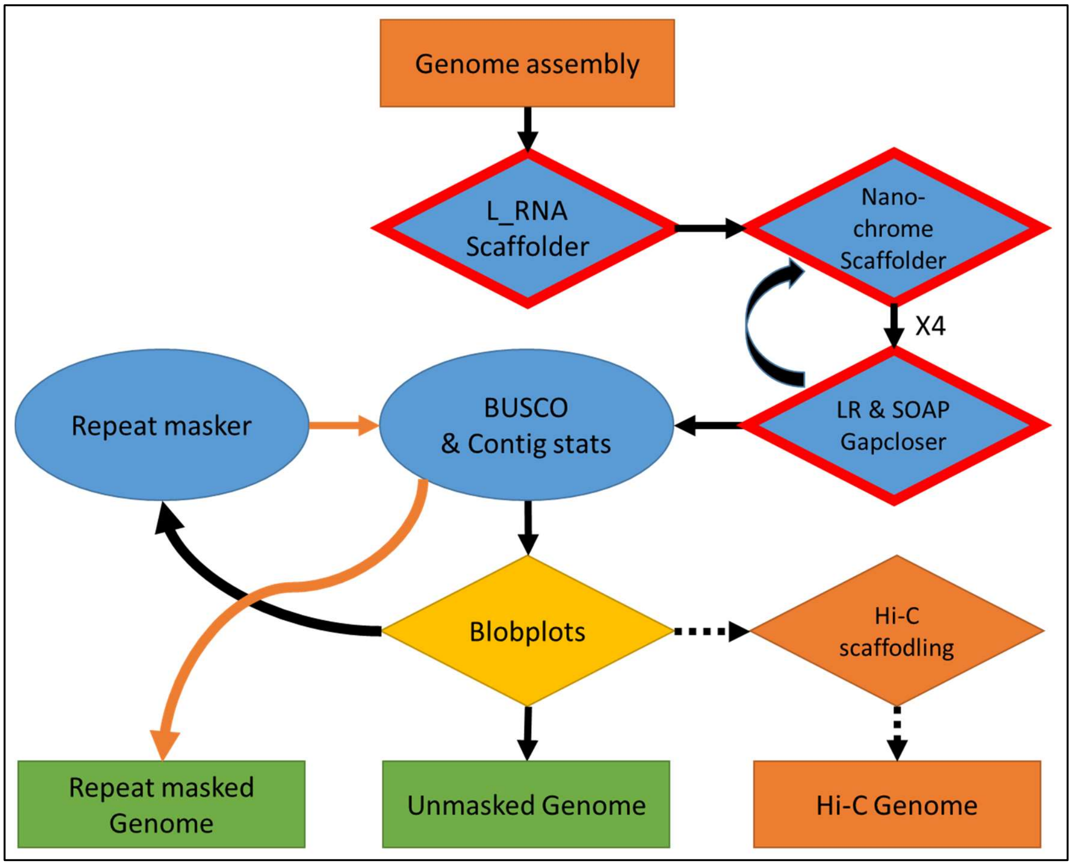


Figure 28: Genome scaffolding step 1: highlighted in red are the three steps involved in scaffolding and gap closing to increase assembly N50 size.

To ensure scaffolding and gap closing did not break up genes, BUSCO V1.0 was used to evaluate and estimate the completeness of core genes within the assembly (Waterhouse et al. 2017). This was then piped into Blobplots v1.0 to identify the proportion of contigs that did not originate from *A. caliginosa.* Contigs which had no support are often contaminating bacteria or AT rich repeat regions which can be excluded from the assembly (Laetsch and Blaxter 2017a).

For use in SNP analysis the resulting assembly was masked for repeats using RepeatModeler and RepeatMasker (Smit et al. 2015). This initially searches and models the genome for areas of repetitive DNA and generates a database of repetitive elements. these. These can be transformed to a hardmasked genome. The hardmasked assembly was looped back through BUSCO for a final assembly evaluation. All the commands used in this section are detailed in Chapter 2.6.

Although not included as part of this research project, Hi-C sequencing is being undertaken. This works by fixing uncondensed DNA in situ of undigested tissue. DNA strands that are close to each other are 'tied' together. Once fixed, tissue is digested, and DNA fragmented for short read sequencing. DNA that was 'tied' together have high probability of being structurally close. This can be used to group and order assembly contigs and potentially achieve chromosome level bucketing of contigs and increase the N50 of the assembly still further by allowing for structural variation in chromosome copies.

5.3.6. **Initial PASA Genome annotation assessment.**

The annotation of the unmasked genome allows for an evaluation of the structural integrity of the assembly. Further it allows for transcriptomic mapping of RNAseq data to the genome. In this chapter the trimmed RNA reads provided by Dr Stephen Short, were used to produce a mapping GTF file required for assessing optimised RNA read mapping methods and produce an annotated Genome that can be viewed in IGB (Freese et al. 2016). The mitochondrial DNA was also 'baited out' of the genome assembly and annotated with Genious with a previously annotated mitochondrium from *L. rubellus*. All the commands used in this section are detailed in Chapter 2.6.

The first stage of PASA genome annotation required performing a Genome guided Trinity assembly (Grabherr et al. 2011). Performing a genome guided assembly reduces the number of multiple copy isoforms that can be generated during *de novo* assemblies but performs best with genomes with large N50s. The genome guided transcriptome is now processed through Evidential Gene, (Evigene), (Gilbert 2013). Evigene works to remove duplicate genes and remove unsupported transcripts. Both the pre and post Evigene transcriptomes were imputed to PASA, (Program to Assemble Spliced Alignments) (Haas et al. 2003). PASA annotates eukaryotic genomes and identifies splicing variations supported by transcript alignments. PASA produces a variety of annotation files including a genome GTF file which can be used by downstream analysis software like Transdecoder, a tool that can predict coding regions within transcripts. PASA generated GTF files for both the pre and post Evigene transcriptomes. The GTF files were visualised with Integrated Genome Browser (IGB) (Freese et al. 2016). The GTF files were also processed with the genome and transcriptome through Transdecoder to produce an annotated GFF3 file (Haas et al. 2013).

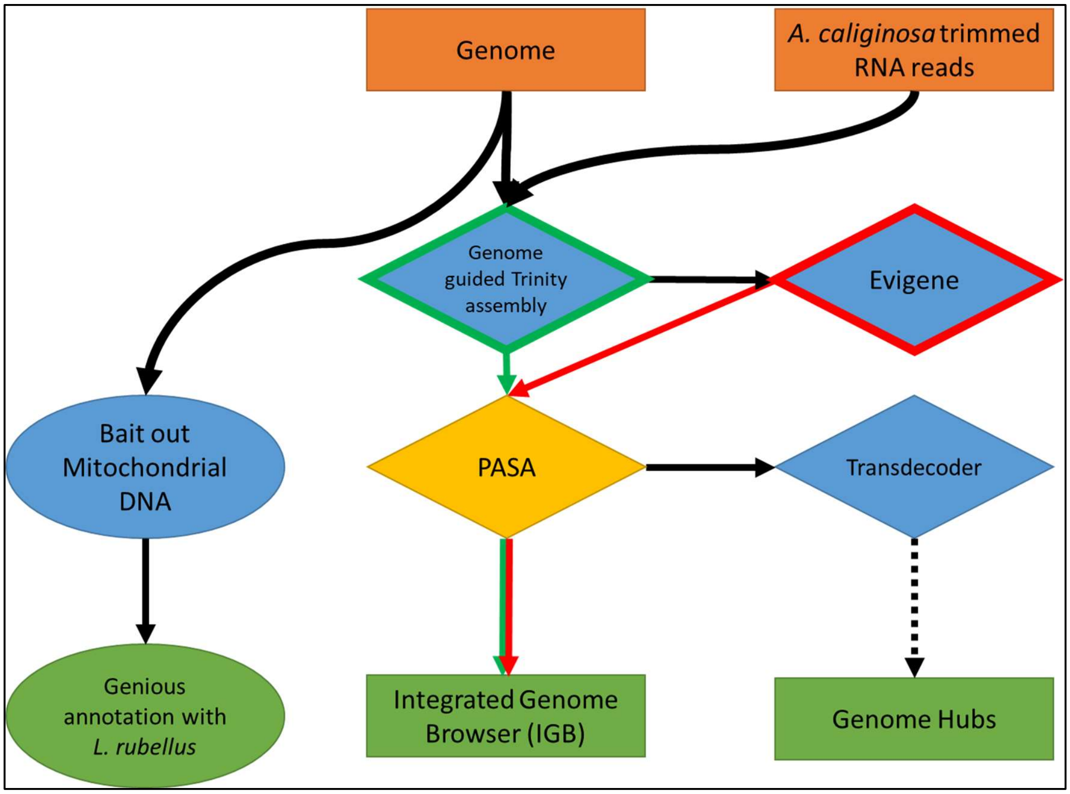

Genome
A. caliginosa trimmed RNA reads
Genome guided Trinity assembly
Evigene
Bait out Mitochondrial DNA
PASA
Transdecoder
Genious annotation with L. rubellus
Integrated Genome Browser (IGB)
Genome Hubs

Figure 29: Genome annotation with Trinity and PASA.

Separately, all sequences containing mitochondrial genes were extracted from the genome assembly with a COI sequence 'bait'. The sequences were visualised in Genious (v9.1.8) and the contig of the whole mitochondria was annotated from a fully annotated mitochondria from *L. rubellus*.

5.3.7. **OmicsBox Genome annotation assessment.**

The PASA genome annotation allowed for an assessment of structural integrity. OmicsBox is a licenced Mapping and Annotation software based on Augustus (Gotz et al. 2008; Bioinformatics 2019). It is a more powerful program than PASA that can better identify intons, exons, star and stop codons and CDS. Omicsbox can be run with a GUI and the output leads directly into gene annotation pipeline without the need for file conversion steps. The Repeat Masked genome (ApCa__Genome_masked_070519.fasta) was used to generate an annotations file with default OmicsBox settings. This was then annotated with Blast2GO and InterproScan with OmicsBox default settings and Annelid selected to narrow blast searches (Figure 30).

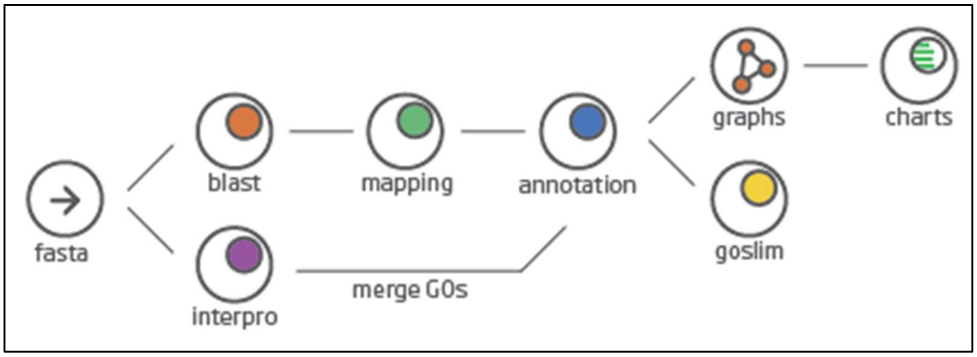


Figure 30: OmicsBox annotation workflow.

5.4. Results.

5.4.1. **DNA extraction and purification.**

Following DNA extraction and purification, quality was assessed via Nanodrop, Qubit and agarose gel (Chapter 2.6). A total of 19.5 µg was extracted from a single section of *A. caliginosa* muscular tissue (Figure 31).

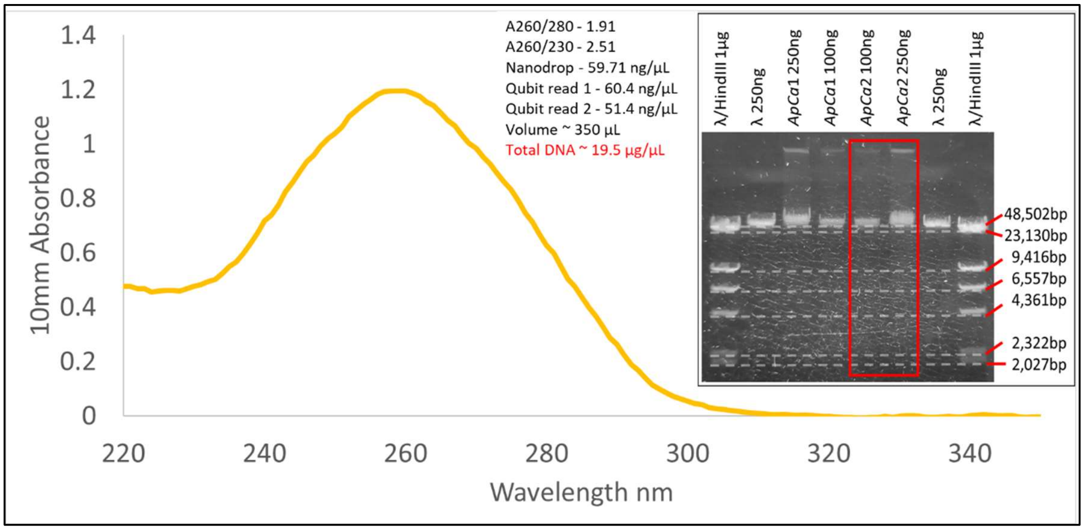


Figure 31: Nanodrop spectrum and Qubit measurements QC data from purified Aporrectodea caliginosa high molecular weight DNA.

5.4.2. **Nanopore.**

Three flowcells were used for sequencing *A. caliginosa*. The received flowcells had 1600, 1240 and 1511 active pores on arrival as measured by Nanopore's flowcell QC measurement function.

The first library prep generated 49.99 fmol which was loaded onto the flowcell with 1600 pores. Sequencing ran for 92 hours and generated 7.04 Gbp (with an additional 0.8 Gbp below the quality score cutoff). The second library prep generated 29.45 fmol which was loaded onto the flow cell with 1240 pores. Sequencing ran also for 92 hours and generated 4.7 Gbp (with an additional 0.44 Gbp below the quality score cutoff). The third library prep generated 53.53 fmol which was loaded onto the flow cell with X pores. Sequencing for 92 hours, 12.85 Gbp was generated (with an additional 4.57 Gbp below the quality score cutoff). This gave a total of 24.59 Gbp with an estimated coverage of 37.83 fold coverage for the 650 Mbp genome of *Aporrectodea caliginosa* (Figures 32 and 33).

Run 3 was the most successful run in terms of data produced, though the error rate was higher than seen in runs 1 and 2. The slower degradation of the pores and the high rate of pore in strand largely contributed to this. Run 2 started with fewer pores than runs 1 and 3 and had a lower proportion of pores in strand contributing to the lower yield of data. All three runs gave similar read length distributions from the sequencing of around 10 Kbp, though run 3 did have a slightly higher yield of smaller fragments around 5 Kbp. Runs 1, 2 and 3 had read fail rates of 89%, 91% and 74% respectively where failed reads were classified by a basecall quality score less than 7 (Figure 34).

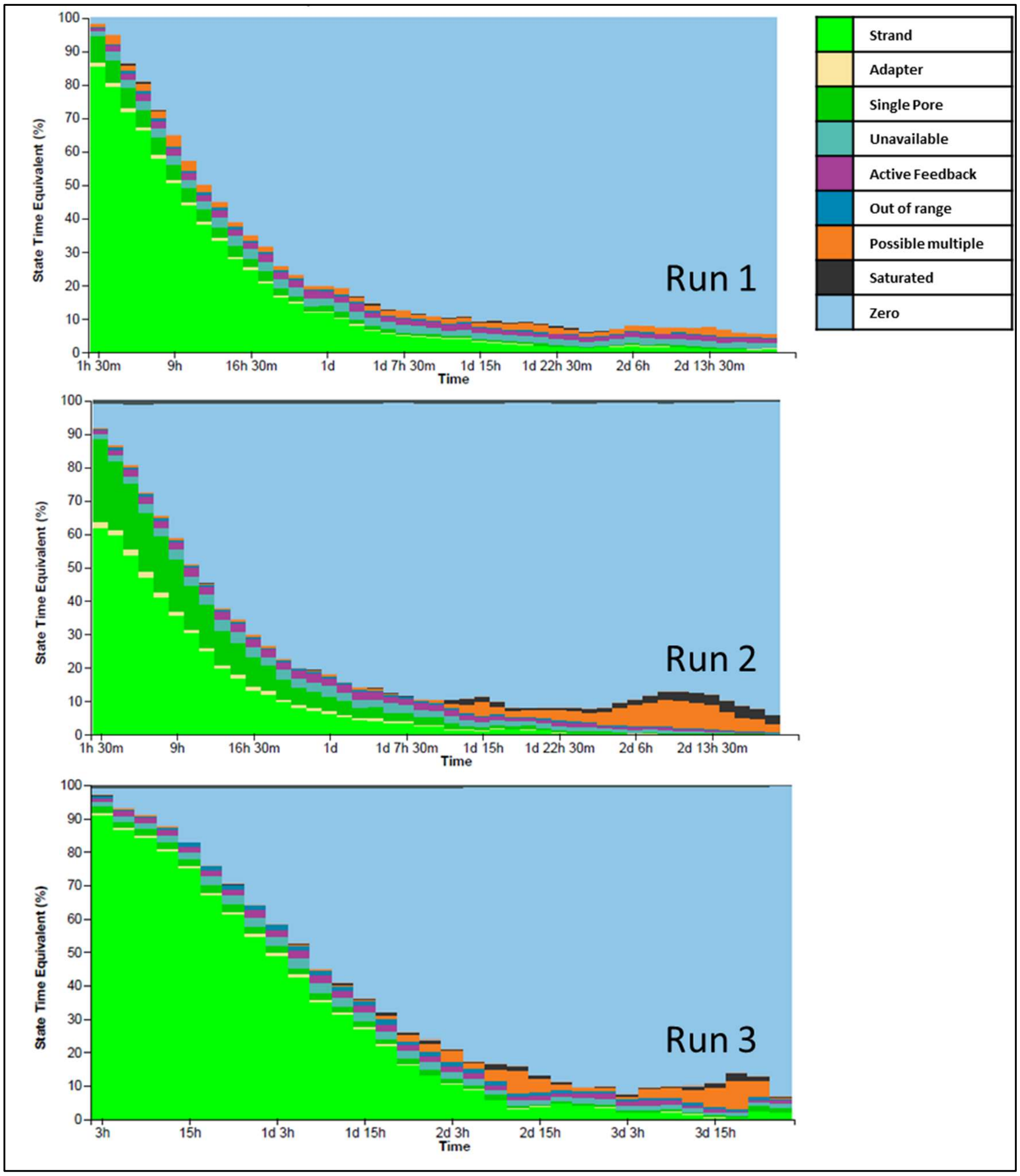


Figure 32: Duty time, summary of the relative proportion of starting pores in different states against run time. Strand indicates sequencing is occurring, while zero indicates pore death. Strand indicates a DNA fragment is undergoing sequencing, Adapter indicates a DNA bound adapter has met with an available pore, Single pore indicates a free pore available for DNA to bind for sequencing. Unavailable, Active feedback, out of range, possible multiple and saturated indicate errors in pore sequencing through blockage, damaged or unexpected signal values. Zero indicates dead pores.

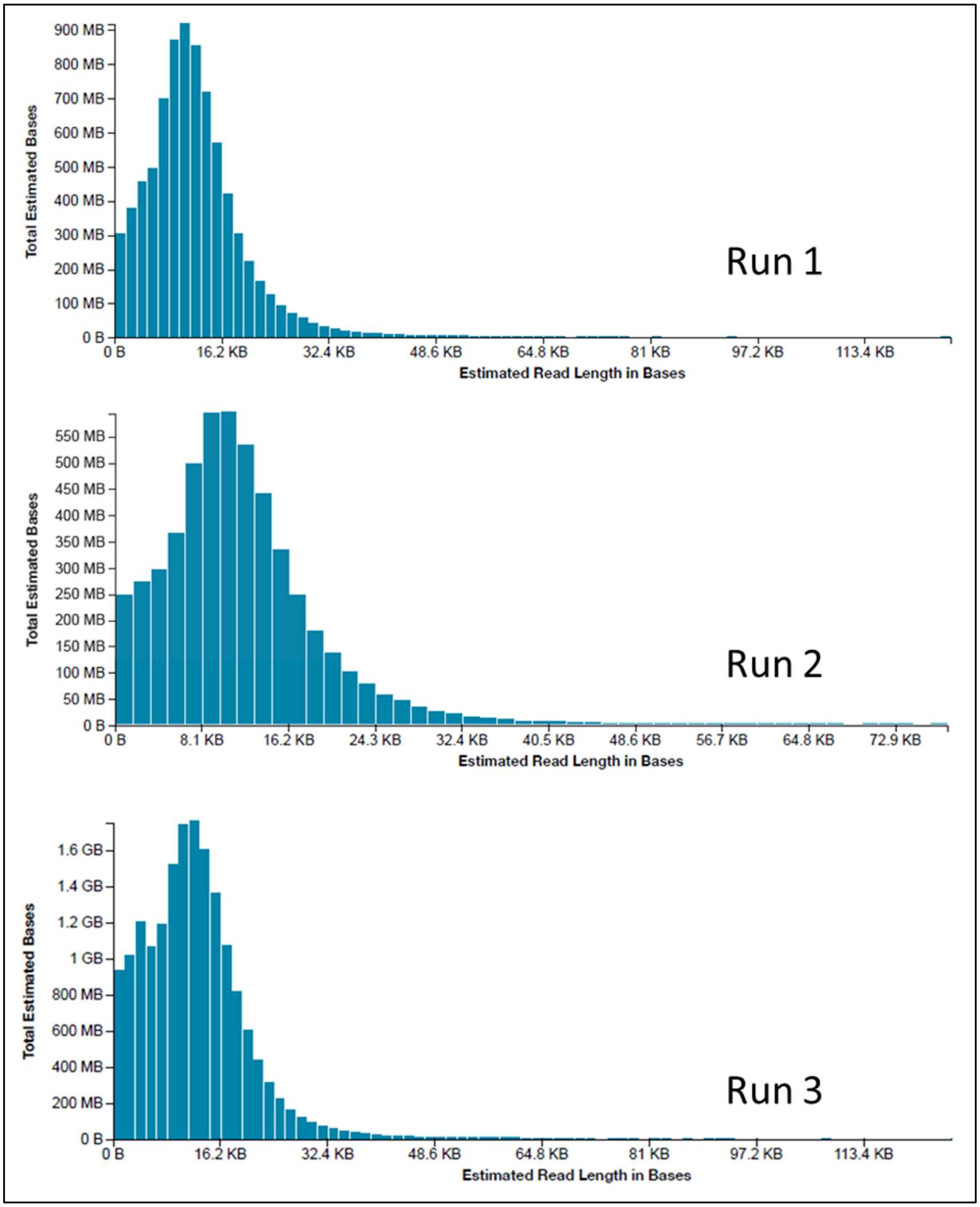


Figure 33: Read length summary distribution in number of reads against read length for each run, indicating the lengths of reads sequenced.

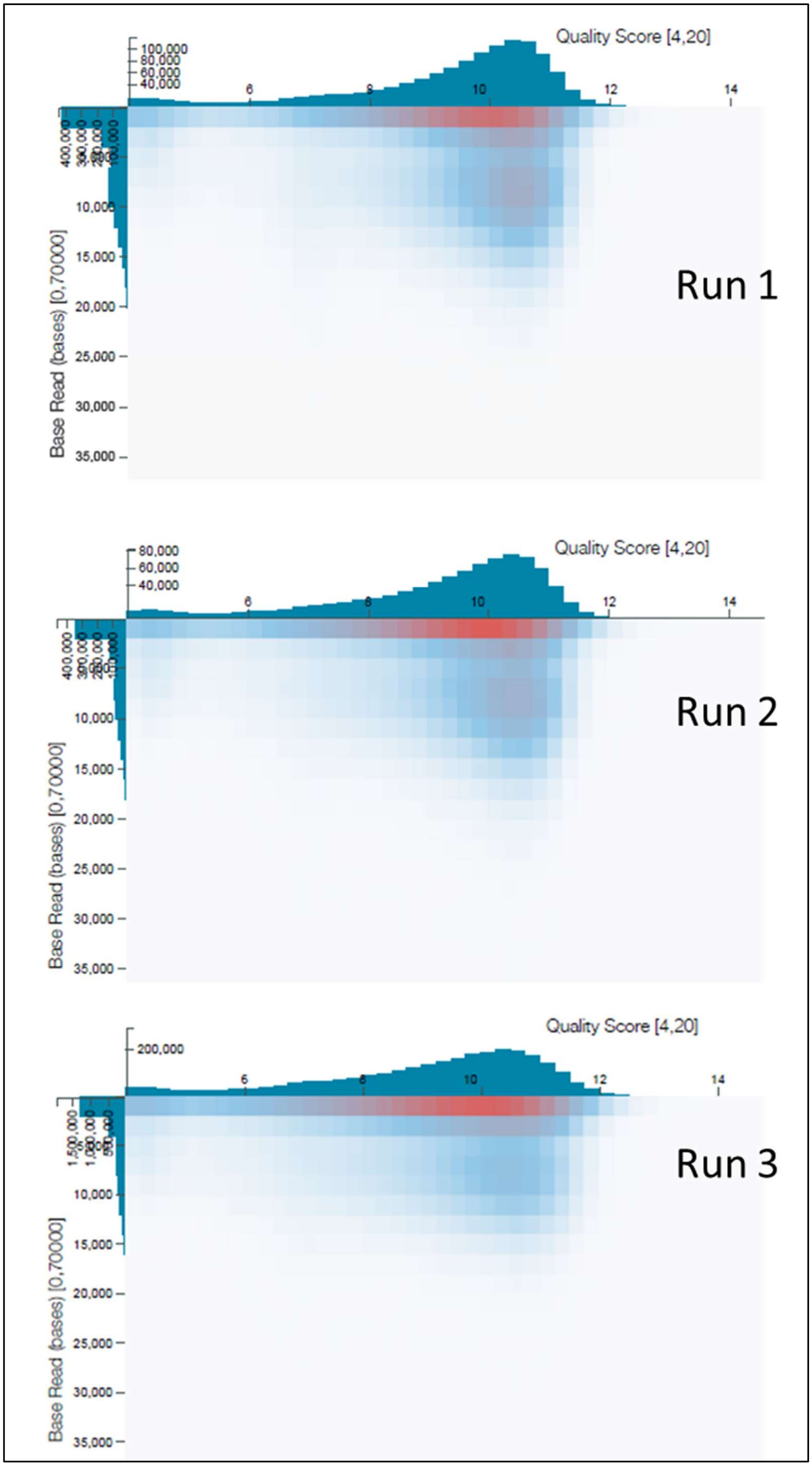


Figure 34: Basecall QC heatmap of read length against basecall quality score for each run. Reads with a quality score lower than 7 were excluded automatically by Nanopore basecalling.

5.4.3. **10x Chromium and Illumina short read sequencing.**

50 fold coverage data for 10x Chromium and 50 fold coverage of data from short read sequencing was requested from Novogene. We received 97 Gbp of 10x Chromium sequences equating to ~140 fold coverage[5] and we received 60 Gbp of short read sequences equating to ~92 fold coverage.

5.4.4. **10x Chromium contamination.**

During the Supernova assembly pipeline, considerably lower coverage was detected for *A. caliginosa* than expected. Despite flagging the assembly to allow for low coverage, a poor assembly was produced. An initial check for contamination was performed by fishing for sequencing reads containing the COI gene from the raw reads. These were reads were then blasted for identity. While most reads matched to *A. caliginosa* or matched to worms with a low blast score, a perfect match was found with *Coptotermes formosanus* (an East Asian termite species). While only a low coverage for this was detected in the COI fishing, it indicated a large scale read mapping for contamination should be performed. This detected that over 3.9% of reads from the 10x Chromium data were not from earthworms, while 0.09% of the short read data did not map. The low percent that did not map with the short reads is to be expected through minor error mismatching, while the high mismatch of 10 Chromium indicates contamination has occurred in the company's 10 Chromium library preparation. While this did exclude the possibility of generating a good Supernova assembly, the reads could be used in Nanochrome's scaffolding program as contamination will not match and can be excluded.

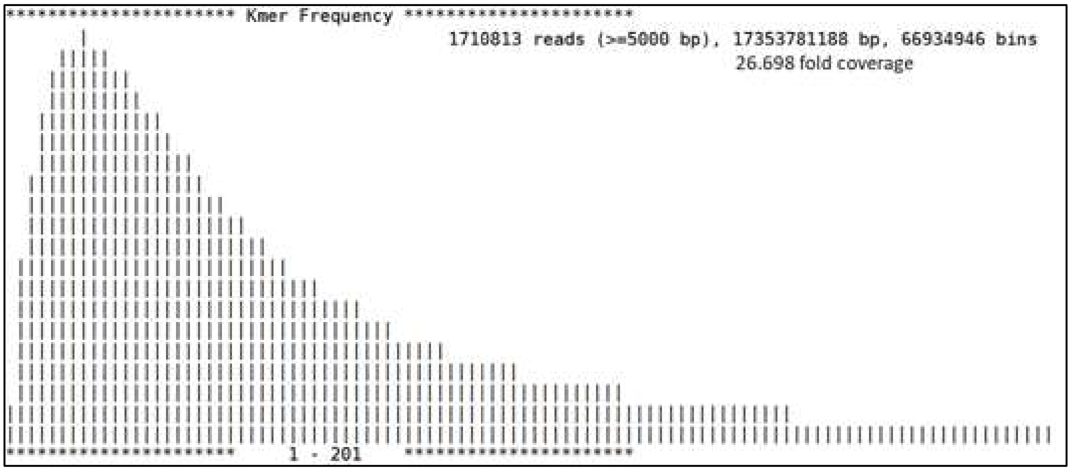
5.4.5. **Assembly stats.**

Long read sequence assembly (post correction) was compared with short read 10x Chromium Supernova assembly to identify the methodology that generated the best assembly. Assembly stats of each successfully completed assembly are shown in Table 16 below. Wtdgb v2 produced the largest N50 while Supernova failed to produce a good assembly when compared with long read assemblies with significantly more contigs and a much smaller N50 (Zheng et al. 2016). The Assembly from Wtdbg v2 was used for further scaffolding.

[5] When converted to raw reads for scaffolding.

Table 16: Assembly statistics of 4 assembly pipelines.

Assembly	*Processing*	*Largest contig (Mbp)*	*No. contigs*	*Assembly size (Gbp)*	*N50 (Kbp)*
Wtdbg v1	none	1.39	15,710	1.09	153.50
Miniasm	Racon and Pilon polish	0.81	17,280	0.99	80.70
Supernova	none	0.96	241,090	1.05	7.16
Wtdbg v2	1 Racon polish	1.34	15,973	1.08	159.03



```
********************** Kmer Frequency **********************
                              1710813 reads (>=5000 bp), 17353781188 bp, 66934946 bins
                                                   26.698 fold coverage
**********************    1 - 201    **********************
```

Figure 35: Wtdbg V2 Kmer Frequency estimation in assembly. Distribution of Kmer frequency should be low in column 1 to indicate a strong assembly can be achieved.

5.4.6. **Nanochrome scaffolding.**

L_RNA Scaffolder improved N50 from 159 Kbp to 198 Kbp but also introduced gaps (Ns) to be filled. This output was used for Nanochrome which was run for 4 iterations with gap closing in between. Figures 36 and 37 represent the visual construction of scaffolds during the first iteration. While some contigs grew considerably during scaffolding, many remained unscafolded. Nanochrome massively increased the size of the N50 with each successive run. The first iteration raised N50 from 198 Kbp to 548 Kbp but also increased the number of gap bases from 153 thousand to more than 29 million. The running of gap closers reduced this down to 1.5 million. The three following Nanochrome runs increased N50 to 835 Kbp and 1.1 Mbp before diminishing returns slowed improvement in the final run to an N50 of 1.17 Mbp and a gap size of 1.1 million.

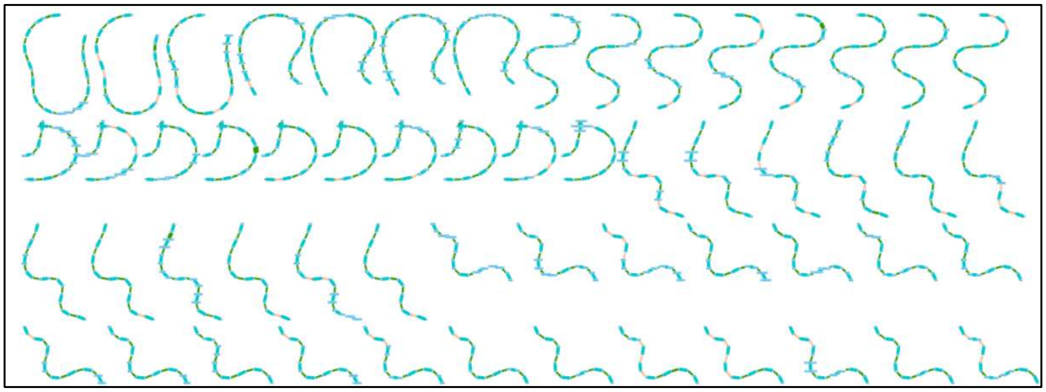


Figure 36: Large scaffold assemblies from Nanochrome scaffolding software.

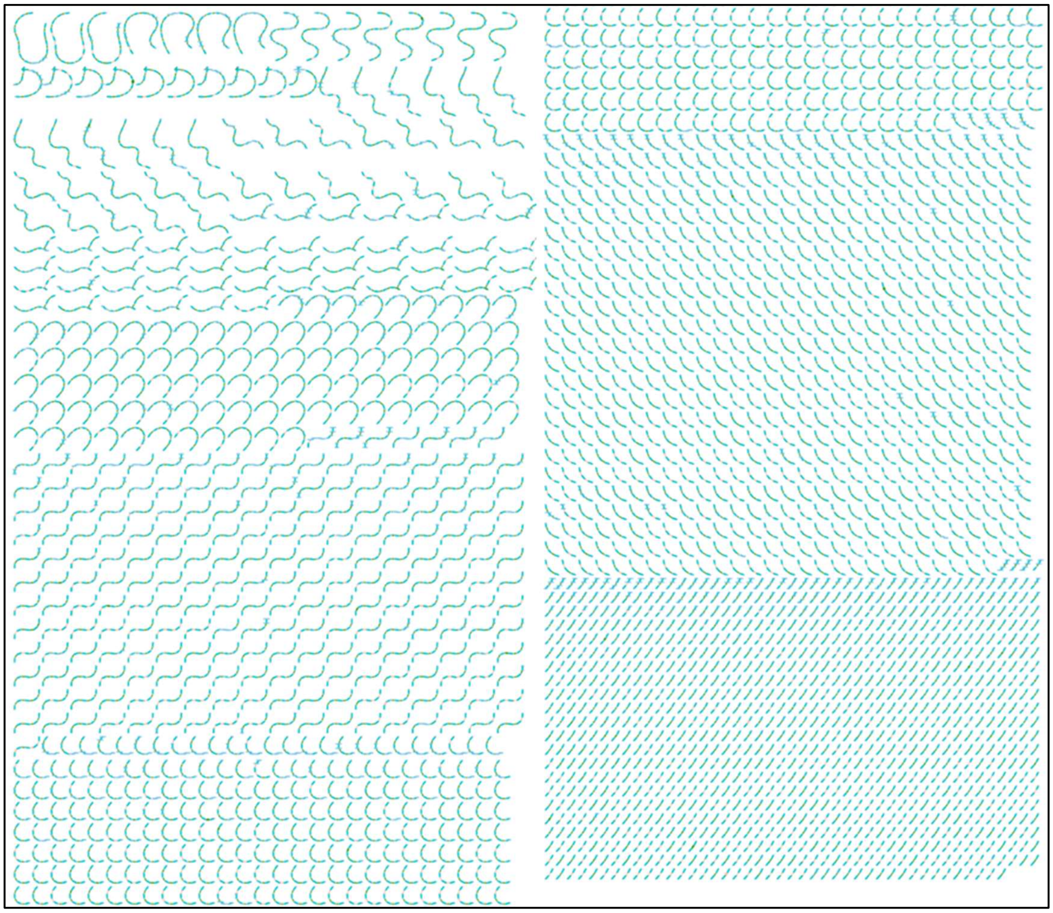


Figure 37: Medium and small scaffold assemblies from Nanochrome scaffolding software.

The final assembly was imported to Genious to visualise the distribution of contig sizes. Figure 38 indicates that there are several ultra large contigs greater than 4 Mbp that could conceivably represent arms of chromosomes, though this is something future research could resolve with Hi-C genomic sequencing data. Figure 39 shows how scaffolding improved N50 and max contig length while reducing the number of contigs and how the size of the genome assembly size increased.

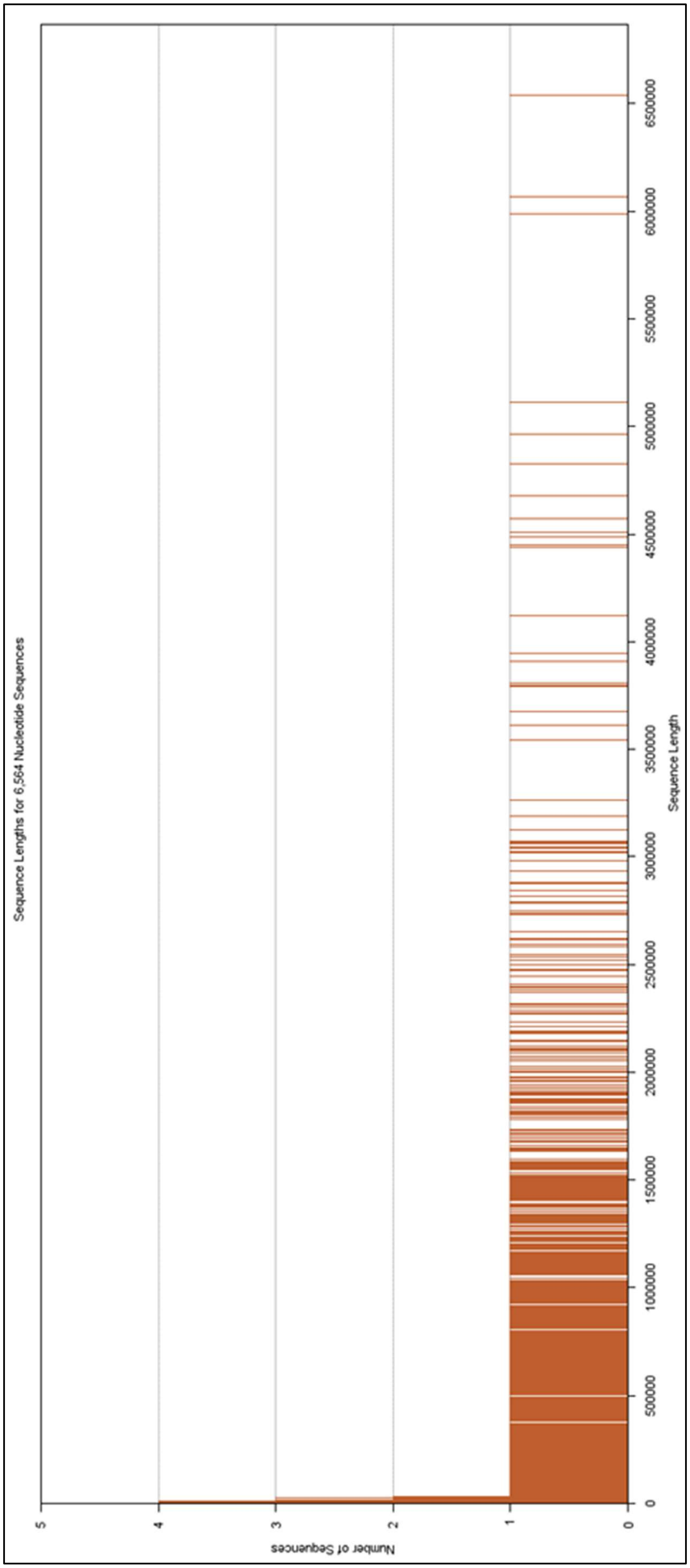


Figure 38: Size distribution of contigs for the final assembly.

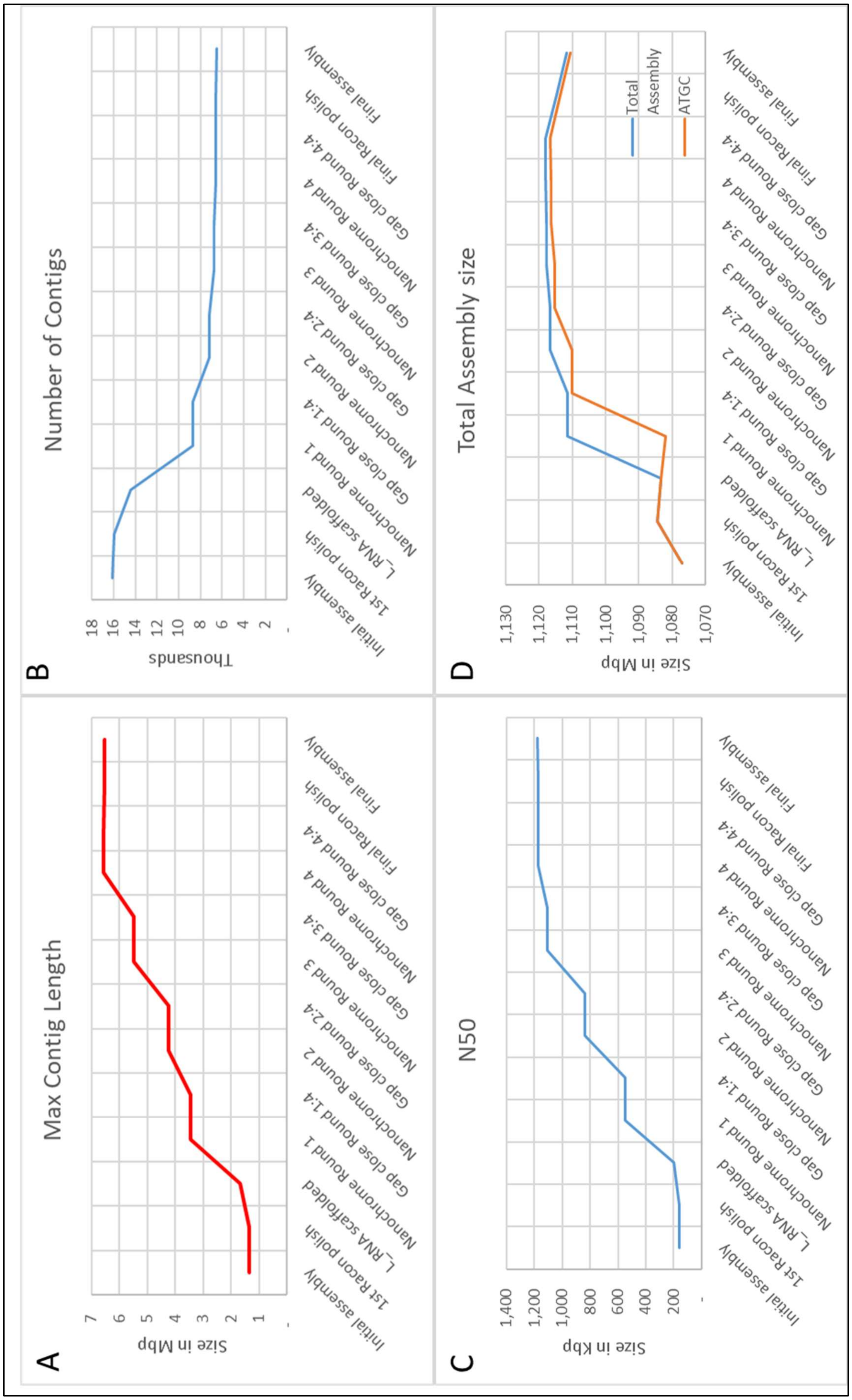


Figure 39: Assembly statistics during scaffolding.

Blobtools was run on the final assembly to identify contigs that did not have support from the transcriptome and could be contaminating DNA from bacteria. Figures 40 and 41 indicate the proportion of contigs that contained support and the contigs identified by the red circle in Figure 41 were excluded from the final assembly as potential bacterial sequences. *Verminephrobacter* was identified by blasting the larger of these sequences, a known earthworm symbiont.

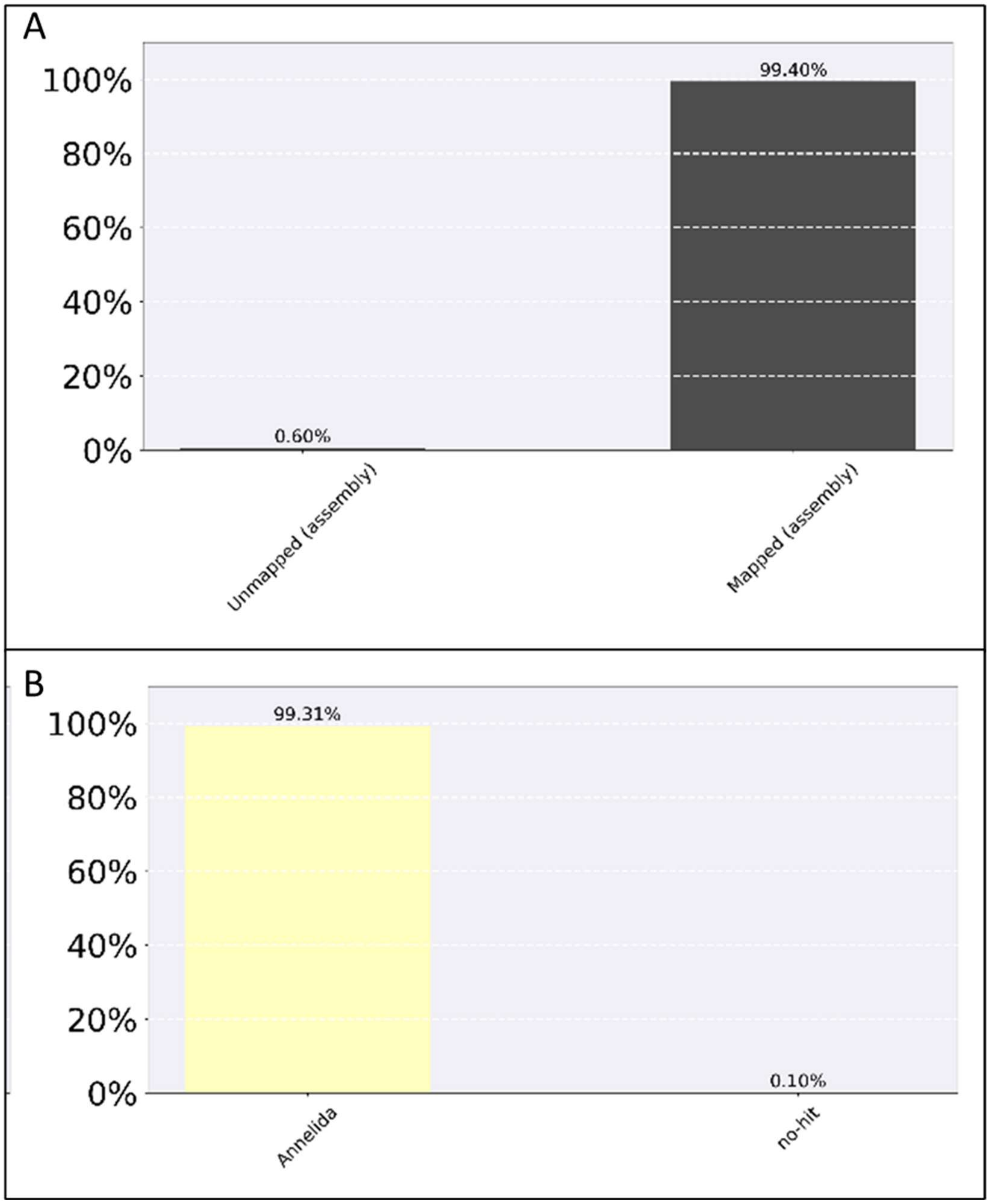


Figure 40: Blobtools blast percentage mapped to A. caliginosa transcriptome. A – percentage of genome that maps. B – percentage of genome that is identified as Annelida.

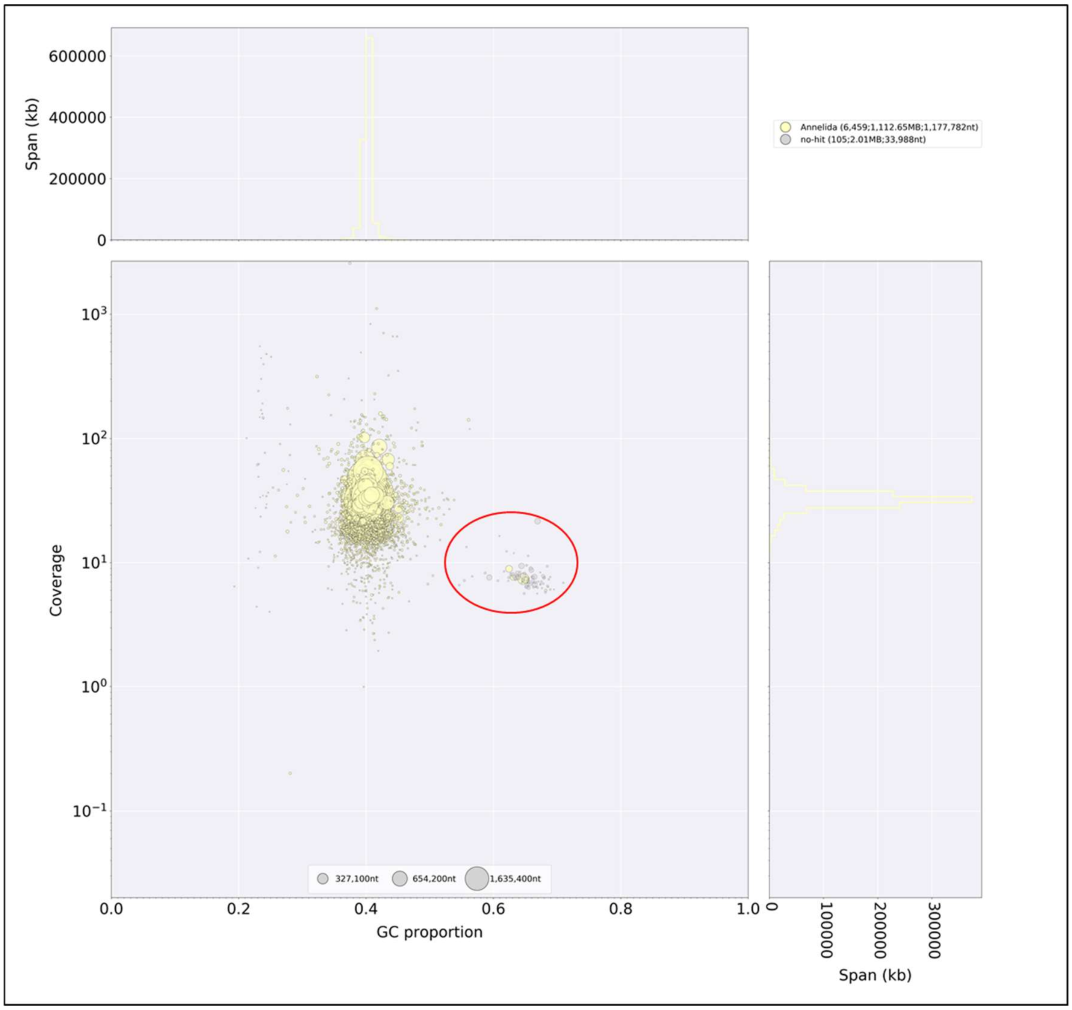


Figure 41: Bloobtools plot of contig size on a coverage Vs GC proportion graph. The red circle indicates about 150 contigs that have no support from annelid transcriptomic reads. The majority of these blast to Verminephrobacter while very small fragments indicated repetitive DNA or sequencing artifacts.

The read coverage from Blobplots was plotted in frequency against coverage bins of 0.5. Figure 42 visualises the two clear peaks of the diploid loci (with 20 fold coverage) and the haploid loci (with 30 fold coverage) at half the diploid loci's frequency. This indicates the high level of haploid diversity within *A. caliginosa's* genome.

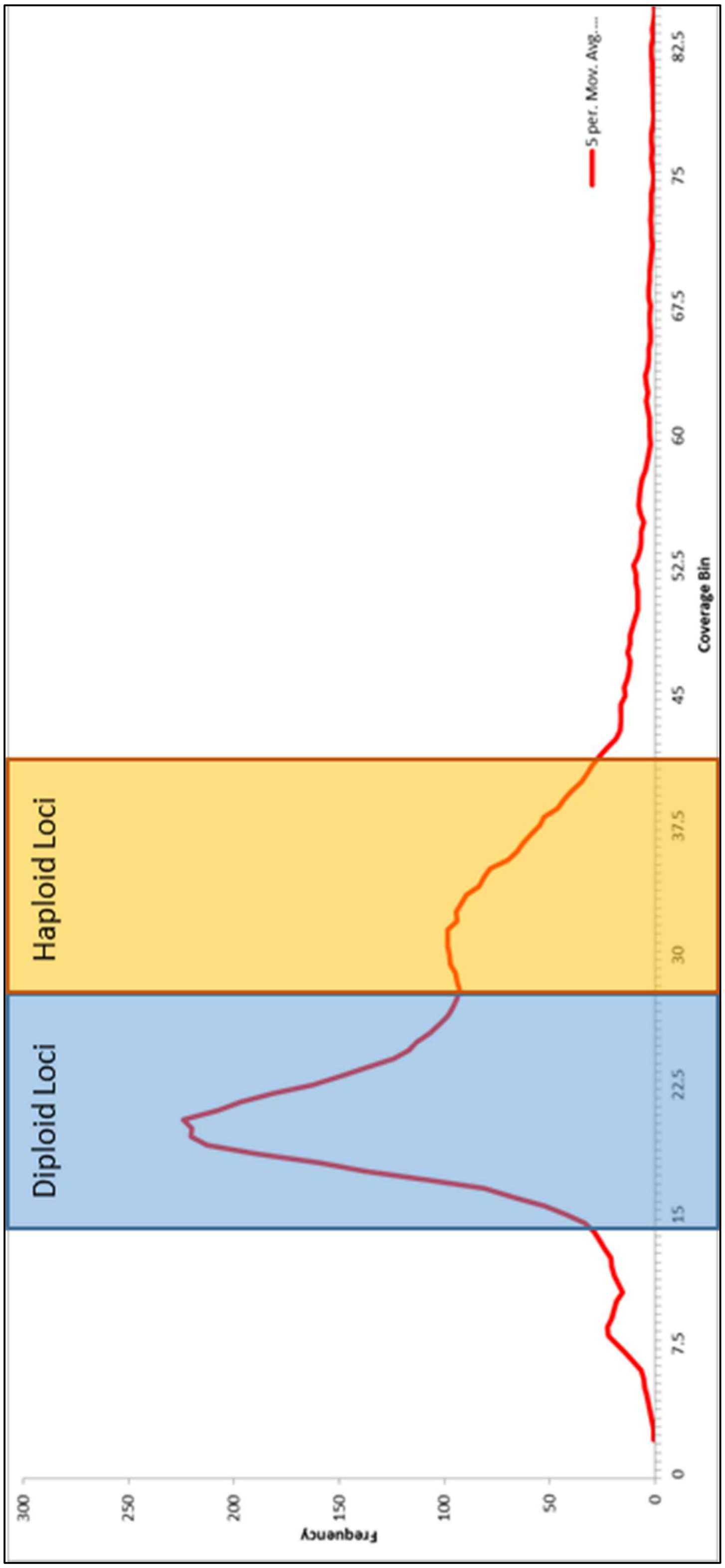


Figure 42: Coverage plot for contig frequency generated from Blobplots output. The diploid loci centres on 20 fold coverage and the haploid loci centres on 30 fold coverage.

BUSCO analysis was conducted to assess the final genome assembly's completeness after scaffolding and Blobtools filtering of non-supported contigs. Figure 43 shows the proportion of complete and missing genes and the breakdown of the complete genes. With over 93% of BUSCO genes found, the assembly is highly complete, and a large proportion of the core genes are not duplicated or fragmented in the assembly.

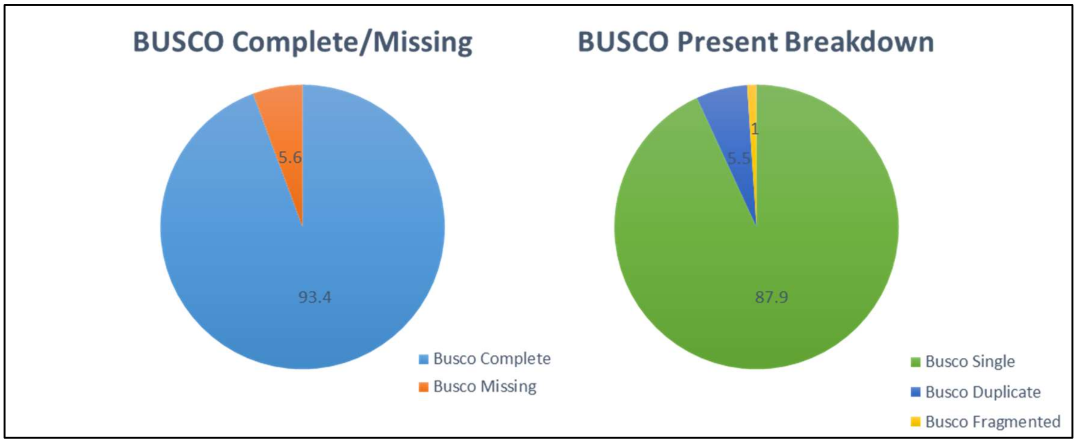


Figure 43: Final BUSCO analysis of the final Genome assembly with non-supported contigs removed. This indicates a highly complete genome assembly.

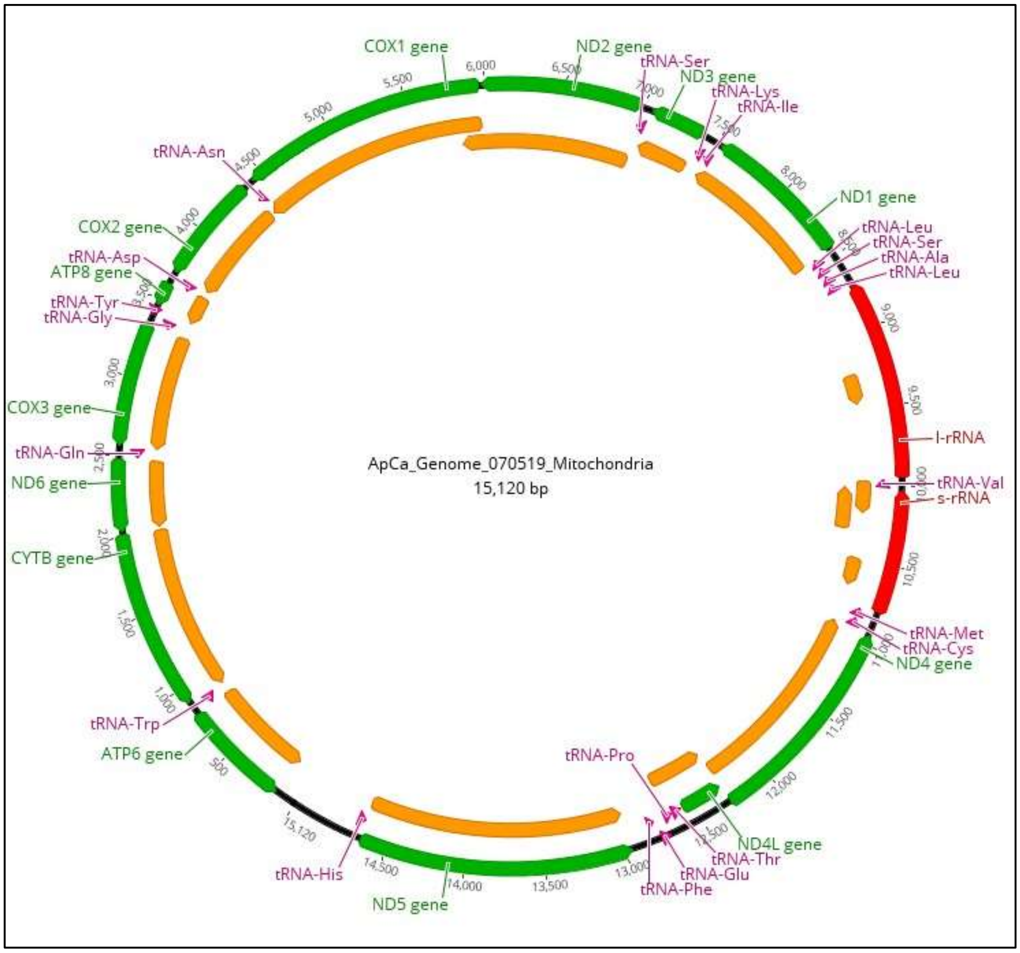


Figure 44: Annotated mitochondria of A. caliginosa from the final assembly. The green arrows indicate annotated genes and their directions, red rRNA and orange indicates open reading frames.

Figure 44 displays the complete and annotated mitochondrial DNA from the genome assembly. The assembled single contig has high level of matching annotated genes with predicted open reading frames for almost all genes. The predicted open reading frames for ND2 and ND4L have small errors missing the closing and opening of reading frames respectively while the rRNA has no complete open reading frames.

A Snail plot was generated for the final genome assembly to visualise the distribution of contig length, N50, scaffold length and GC composition (Figure 45). It indicates in particular the GC wobble that the shorter contigs in the assembly have. Many of these are ultra-repetitive sections of DNA that have either not been able to assemble or are artificial from sequencing and can be removed.

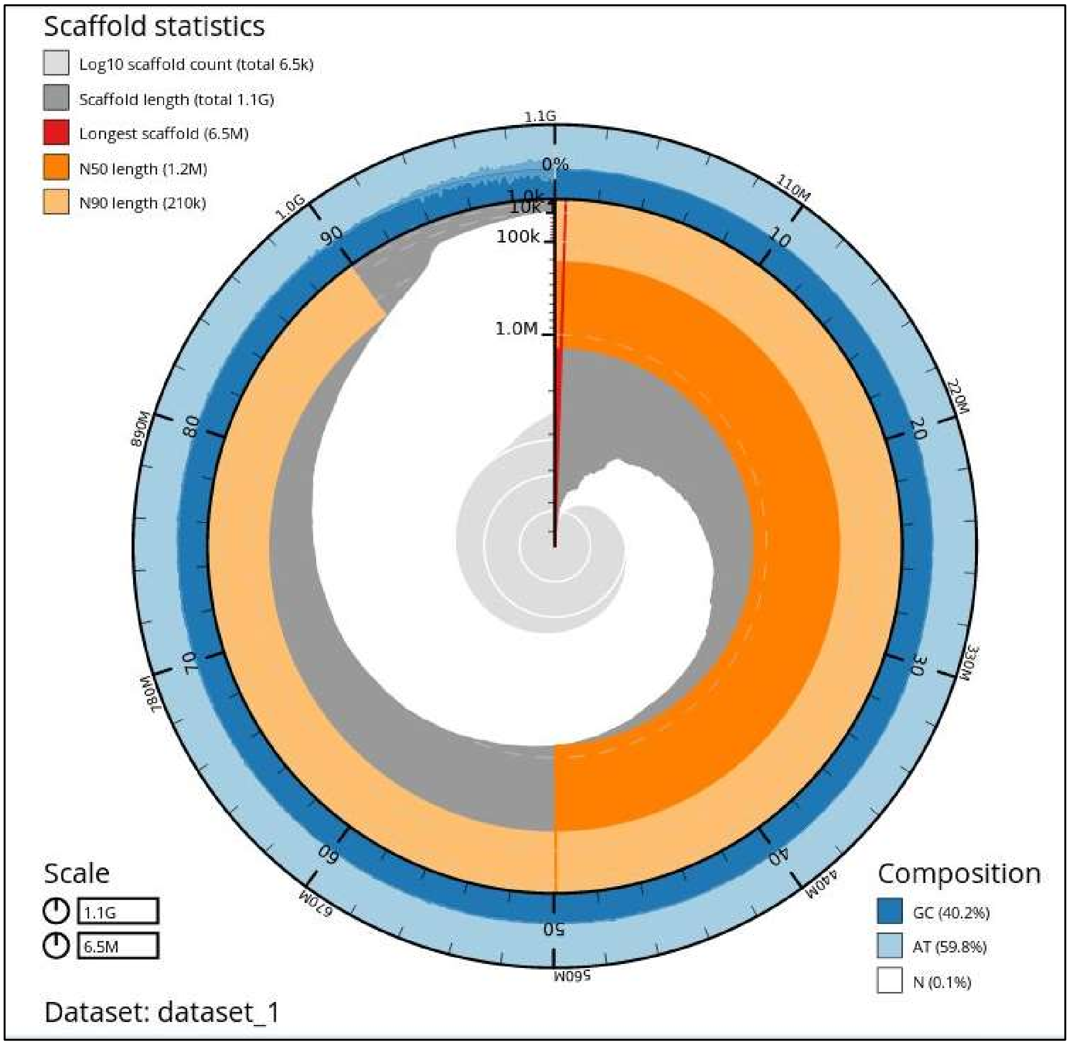


Figure 45: Snail plot for final genome assembly generated in Blobtools v2. In the N50, 279 contigs account for half the genome. In the N90, 530 contigs account for 90% of the genome.

5.4.7. **Repeat Masked genome**

The genome was masked for repeats for later use in SNP analysis, leaving 43% of the genome unmasked. Of the 57% of the genome that was hardmasked, with nine genome element categories ranging from 43% Non-repetitive to >0% Small RNA repeates (Figure 46). BUSCO was re-run on the masked genome with minimal changes compared with the unmasked (92.5% Complete, 89.3% single, 3.2% duplicate, 1.6% fragmented, 5.9% missing).

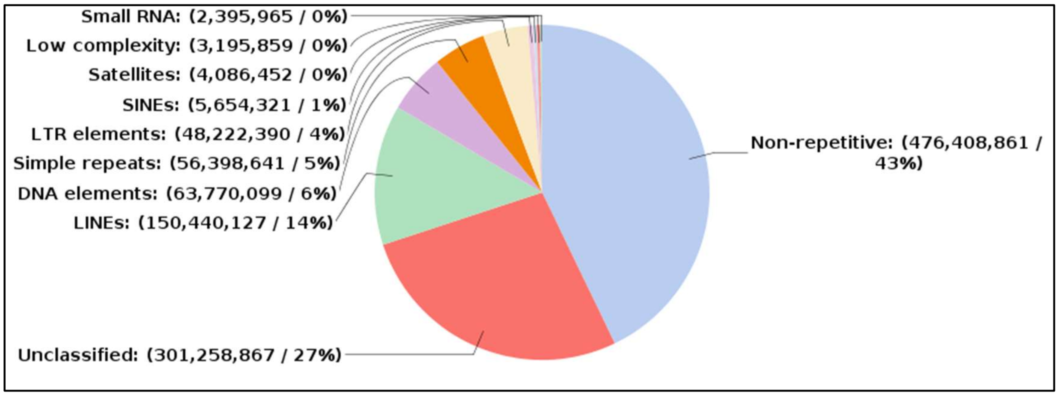


Figure 46: Distribution of masked elements per classified element.

5.4.8. **PASA annotation visualisation.**

Annotations for the genome were generated from both the Trinity Guided transcriptome assembly of the RNA reads used in the L_RNA scaffolding, and the post Evigene transcriptome. The post Evigene annotations remove a large element of redundant annotations but also exclude some useful annotations. Figure 47 demonstrates the variable selection of transcripts during the Evigene processing. Evigene does not select transcripts to keep by longest length as it works on the principle longer transcripts are often a result of sequencing errors that alter the open-reading frame. Evigene instead attempts to select by longest protein. This however can lead to oversimplification of splice variants and ignore longest transcript and longest protein. For the purpose of this research thesis, retaining the additional annotation does not impact negatively.

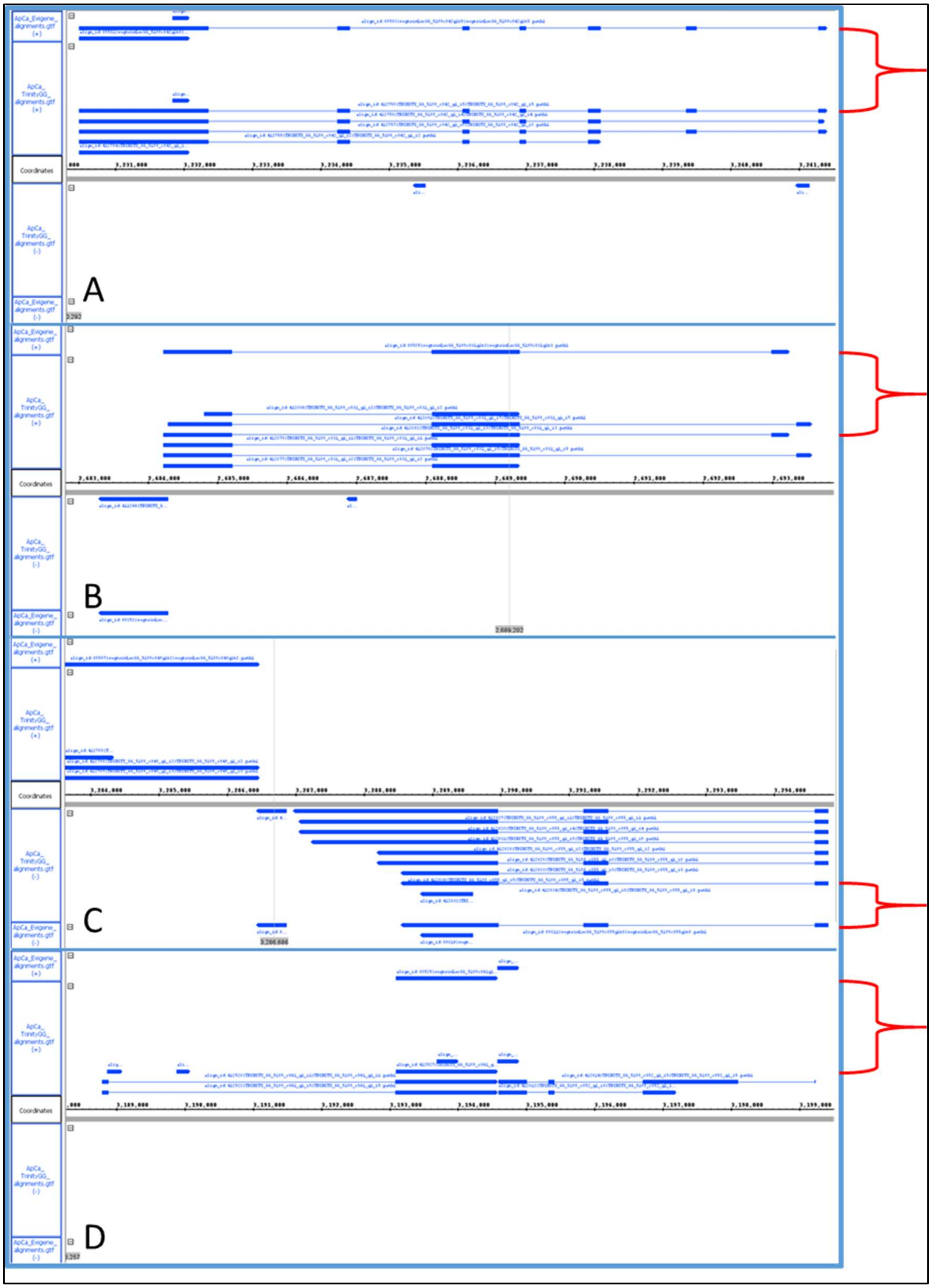

ApCa_Evigene_alignments.gtf (+)
ApCa_TrinityGG_alignments.gtf (+)
Coordinates
ApCa_TrinityGG_alignments.gtf (-)
ApCa_Evigene_alignments.gtf (-)
A
B
C
D

Figure 47: Selection of annotated genes along the longest contig. A) Evigene has kept the longest transcript. B) Evigene has selected a shorter transcript. C) Evigene has selected one of the shortest transcripts. D) Evigene has selected the shortest transcript. Red brackets indicate which transcript Evigene has kept.

5.4.9. **OmicsBox annotation.**

OmicsBox identified 42,566 gene objects across the masked Genome. Figure 48 displays the cleaner and more definitive identification of gene objects identified across the genome than detected with PASA. The largest gene identified comprised of 19 exons and was spread across 196,640 and was identified as Chemoattractive glycoprotein ES20 (Figure 49).

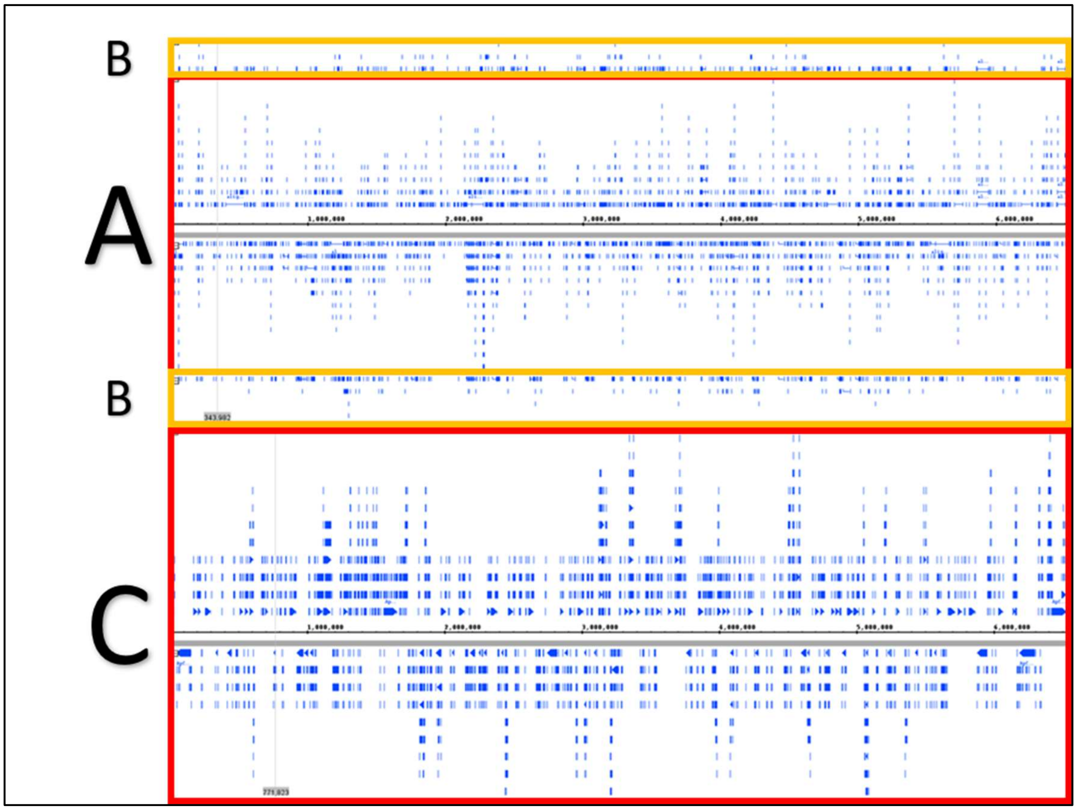


Figure 48: IGB view of the (A) PASA Trinity Genome guided transcriptome annotations and the (B) PASA post Evigene transcriptome and (C) OmicsBox annotations for the longest contig, (6.54 Mbp). The PASA Evigene annotaion of the masked genome showes a much reduced version of potential genes seen in the PASA Genome guided transcriptome annotations. OmicsBox annotations shows a cleaner annotation with genes identified with their corresponding exons.

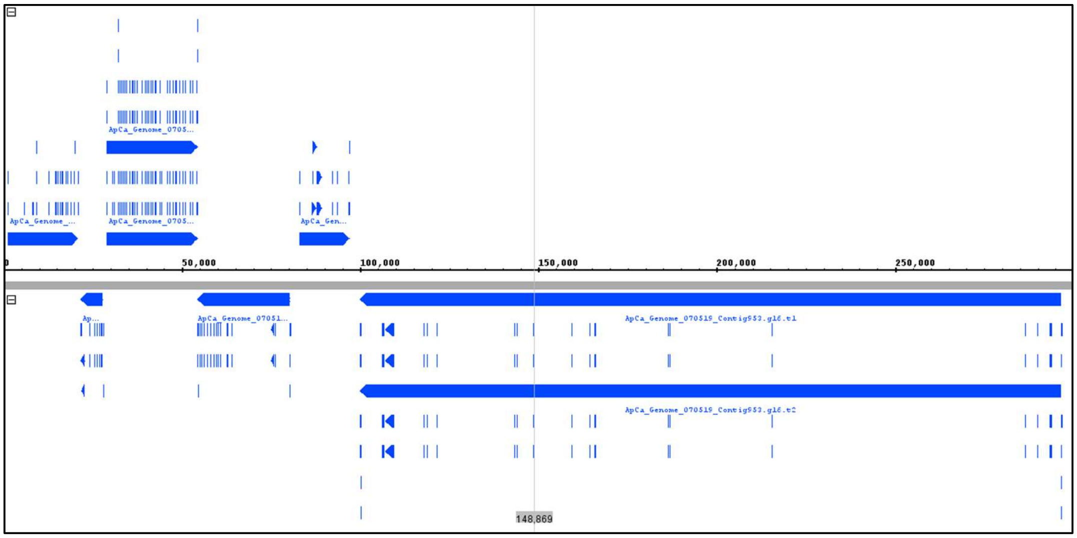


Figure 49: The longest identified gene span from OmicsBox annotation, Chemoattractive glycoprotein ES20.

5.5. Discussion.

5.5.1. **DNA extraction and purification**

High fragment size DNA extraction is a challenging process that required multiple attempts and modifications to commonly used extraction protocols to produce the final high quality DNA used during sequencing. Standard 'column' based extraction and purification methods are a crude but effective way for mass extraction of DNA in high numbers of samples. The drawback they have is a shearing effect upon the DNA as it passes through the column purification. This effectively reduces the maximum DNA fragment size to below 10 Kbp, though much of the DNA will be smaller still.

Phenol-Chloroform based DNA extraction is an older and more hazardous method but allows for a 'gentler' extraction. Traditionally this method might utilise liquid nitrogen crushing or Dounce homogenizing, but these have drawbacks including loss of starting material on the side of a mortar and difficulty in shearing tough muscular tissue. The use of ATK and Proteinase K allowed for a simpler and gentler method for releasing large fragment DNA. As common with the Phenol-Chloroform extraction, we used an isopropanol precipitation to collect the large string-like DNA on glass needles. Despite this, very low levels of Phenol, Chloroform and Isopropanol still contaminated the extracted DNA. A further modification to this extraction method was required to remove this with TE dialysis.

All these modifications made through trial and error finally produced sufficient quantities of DNA required for the multiple sequencing attempts of Nanopore, 10x Chromium and Illumina short read sequencing. Bead based extraction methods such as those suggested by Mayjonade *et al.* resulted in poor DNA recovery as HMW DNA bound irreversibly to the beads, though this might be an isolated issue with earthworms as neither Mayjohnade *et al.* or Pushkova *et al.* reported this complication (Mayjonade et al. 2016; Pushkova et al. 2019). Much of the difficulty of retrieving high fragment size DNA from earthworms comes from their small size and muscular tissue. While cell cultures can provide vast quantities of pure cells for DNA extraction and larger organisms can spare larger tissue inputs, the earthworm cannot and due to their high allelic diversity, multiple individuals are not recommended.

5.5.2. **Nanopore**

Nanopore sequencing is an emerging and rapidly developing sequencing technology that during the course of this thesis progressed substantially. It is highly likely by the time of this publication, most of the reported sequencing capabilities will be surpassed. Illumina sequencing has become over the past decade, the bedrock of genetics through its reliable and straightforward platform. In contrast, Nanopore is un-reliable, flowcells are highly variable and high risk through both its high financial cost and high DNA input 'cost'. Despite these drawbacks, the platform offers data that only PacBio can compete with, which itself comes at a high financial cost and can help solve challenging *de novo* assemblies (Bayega et al. 2020). The high variability was seen across the three runs of sequencing that ranged from outputs of less than 5 Gbp to over 12 Gbp. Both runs cost the same but performed substantially differently. When purchasing a Nanopore Flowcell, 2000 cells are the advertised maximum, though in practice the highest we received was only just over 1700. The lowest we used for sequencing in this run was less than 1200 pores, nearly half the maximum. In our experience, the low numbers of pores indicated that the flowcell was degrading faster and corresponded to a faster drop off in working pores used during sequencing, dramatically reducing the potential sequencing output. Currently, only flowcells below 800 pores will be replaced by the Oxford Nanopore which adds to the high cost risk of using this method of sequencing.

The library preparation was particularly challenging for the large fragment DNA, particularly during magnetic bead cleaning where very large fragments appeared to bind irreversibly to the beads. The led to purposely shearing these fragments to around 20 Kbp where we did not observe this irreversible binding. The use of magnetic beads in sample clean ups is commonly suggested for HMW DNA, and the specific reason for annelid DNA to irreversibly bind is not yet known (Schwessinger and Rathjen 2017). The compromise of fragment size did not overly hinder assembly and during sequencing a small proportion of fragments were sequenced that had evaded shearing. Loading the prepared libraries on the flowcells was technically challenging, though this was achieved without error that could have shortened the lifespan of the flowcell.

Two of our sequencing runs on flowcell performed well, generating a large volume of read data. Of the sequencing run that performed worst, this was attributed to the poor quality of the flowcell delivered. All three runs had similar quality scores as assessed through the basecalling software. Nanopore is known for its lower levels of basecalling accuracy, although this is slowly improving, though utilizing our short read data to error correct these long reads allowed us to remove that shortcoming.

5.5.3. **10x Chromium and short read sequencing.**

A large proportion of the high quality high fragment size extracted DNA was sent to Novogene for 10x Chromium and short read sequencing. Normally the output of the 10x Chromium allows for assembly of large fragments through the barcoding of the large DNA fragments prior to short read sequencing. This can then be assembled back into the large DNA fragments prior to scaffolding of the large DNA fragment reads into a much larger assembly. This is done through the proprietary software Supernova and has good reported success in a variety of organisms. The first indication that this library and sequencing had failed was the very poor assemblies generated through this software. This prompted a bioinformatic ‘fishing’ for reads that align to the COI species barcoding region. All these reads were then blasted for species identity. Species identified included human, mice, rats, Asian wasp and termite amongst others. As some of these species are not currently researched at this facility, confidence was high that this contamination had not been introduced prior to sending for sequencing. Further, the contamination was not identified in any of the short read sequencing performed at the same time and implicating contamination was introduced during the 10x library generation. While this only accounted for under 10% of reads, this effectively prevented Supernova from ever being able to assembly as designed. Despite this loss of data, it was possible to the data in an alternative process that helped to scaffold the reads assembled from Nanopore sequencing. Though in theory contamination could be removed to allow for a Supernova assembly run, early assembly statistics had suggested that the HMW DNA supplied to them had undergone fragmentation before barcoding had even occurred, severely limiting the assembly size.

5.5.4. **Assembly stats.**

There are very few published earthworms with any assembled genomes, and as of yet no published *A. caliginosa* genome. Two assemblies of *E. fetida* exist with N50s below 50 kbp and neither evidencing strong support for gene completeness, while an *E. andrei* assembly has recently been published with an N50 of ~750 kbp and a high level of gene completeness (Zwarycz et al. 2015; Bhambri et al. 2017; Shao et al. 2020). These assembly are all significantly smaller than the genome reported here, which provides not only a very well assembled genome of the species, but the highest continuity and completeness of any species of earthworm. The initial assembly of corrected Nanopore data generated a very strong foundation assembly as assessed via N50 size. The subsequent improvement with RNA data increased the N50 size which while not small, comparatively seemed small in comparison to the approximate 2.75 fold increase in N50 generated through the first round of Nanochrome and gap-filling alone. By the end of Nanochrome and gap-filling rounds, the final assembly N50 was over 7.4 fold larger that at the end of the initial Nanopore data assembly.

The genome assembly N50 was not the only measure of genome assembly quality to improve. In a haploid organism, the theoretical lowest number of contigs that can be generated is the number of chromosomes present (n). For *A. caliginosa* this would be 18 contigs. The centromere is difficult to resolve, in part due to its highly repetitive DNA which functionally means stitching the long and short arms of a chromosome sequence difficult to near impossible with current technology (Lamb and Birchler 2003). This means that the lowest number of contigs becomes 2n (for *A. caliginosa* 36 contigs). If we move to a diploid organism where two copies of each (autosomal) chromosomes are present this number further doubles to 4n (for *A. caliginosa* 72 contigs). Moving through the stages of this genome assembly the number of contigs of the genome dropped from over 16,000 to just under 6,500. Of this 530 contigs accounted for 90% of the genome while 5,970 small fragments (average <2 Kbp) accounted for the remainder. These small fragments are common in assemblies and often represent repetitive stretches of DNA. In theory to reach the minimum number of contigs a little over 7 contigs would need scaffolding and reach near chromosome level assembly. The largest contig assembled was over 6.5 Mbp, which could conceivably represent the majority of a long or short arm of one of *A. caliginosa*'s chromosomes.

Through scaffolding, stretches often appear where the assembling software cannot fill a gap created either as a result of low sequence coverage, segmental duplication, satellite association, muted gaps or allelic variation (Chaisson et al. 2015). There was a large increase in the number of the base gaps during the Nanochrome assembly as expected though, gap filling software was able to reduce the extent of this to around 1.1 Mbp.

Performing BUSCO analysis is a means to measure the completeness of a genome assembly based on a set of core known genes. For earthworms, the closest reference data set to compare to is a Metazoan library. This includes fish, amphibians, reptiles, birds and mammals, so will potentially include some 'core genes' that earthworms do not have or require. The genome generated contained over 93% of these core genes with nearly 88% identified as a complete single copy. This represents a very high and complete genome, and achieving higher levels of completeness are not expected without developing an annelid specific reference set (Waterhouse et al. 2017).

BlobTools successfully identified a large section of contigs that were of contaminating bacterial reads through their GC content. This was confirmed in the blast identification of these contigs, many of which were Verminephrobacter, commonly associated with earthworms. The running of blob tools is fast and essential part to quality assess assemblies, with many old published genomes being found to contain contamination that can alter conclusions of studies (Laetsch and Blaxter 2017b). The coverage calculated with BlobTools allowed the calculation of haploid/diploid fraction of the genome. There was approximately twice the frequency of diploid loci than haploid suggesting that roughly half of the genome is diploid. If we account for the duplication of loci in the haploid half of the genome a fully diploid genome would be around 825 Mbp, a little over the 650 Mbp estimated by Kashmenskaya and Polyakov's karyotyping, but this does not account for over assembly of repetitive sections of DNA in the assembly.

The mitochondrial genome sequence was extracted from the genome assembly and annotated via comparison with a previously annotated *L. rubellus* mitochondria. The annotations were compared with the open reading frames of the assembly. In the majority of the genes, open reading frames aligned correctly with only the ND2 stop codon, the ND4L start codon in the incorrect location and the large and small ribosomal genes with incorrect reading frames. The complete mitochondrial genome sequence allows for a suit of new barcoding primers to be designed for the species including the COI gene which could not be used in the work described in previous chapters.

5.5.5. **Repeat Masking.**

Through building a library of *A. caliginosa* repeats and masking the genome with Repeat Modeller and Repeat Masker, 57% of the genome was masked. This had almost no effect on the BUSCO score dropping to 92.5% core genes. The very high level of masking in this genome confirms the high level of repetitive DNA within the species, though until chromosome level assembly can be achieved this will be hard to confirm the exact level. Interestingly, Fielman and Marsh suggest that in metazoans, there is a direct link between the level of repetitive DNA elements and extreme environments, though lower levels of repetitive DNA were associated with colder environments (Fielman and Marsh 2005).

5.5.6. **PASA Vs. OmicsBox.**

Initially PASA was run on the genome assembly to identify open reading frames and potential coding regions. This was run with the multiple tissue transcriptome used in the L_RNA scaffolding supplied by Dr S. Short. It identified many possible splice variations with different coding regions for genes across the genome. By contrast the PASA annotation using the transcriptome filtered with Evigene only selected a coding region for each open reading frame. Evigene tries to do this based on an evidence based approach to simplify transcriptomes, but this filtering was oversimplified for the purpose of identifying coding regions and potential multiple splice variant sites. OmicsBox's gene annotation tool works in a similar mechanism to MAKER genome annotation tool utilizing Augustus to predict and verify coding regions (Bayega et al. 2020). The program identified over 42,000 gene objects which likely represents an overestimate of the species true gene count. This would in part be due to the high proportion of the genome being comprised of genes with high allelic diversity that would have been identified as a separate gene object. Improving the earthworm assembly continuity and deriving a true haploid copy will likely improve the resolution of the number of genes in *A. caliginosa*, similarly as improvements to the human genome reduced estimates from 100,000 to less than 22,000 (Pertea and Salzberg 2010).

5.5.7. **Concluding remarks and future development.**

The purpose of this chapter was to generate a high quality genome of *A. caliginosa* as the model earthworm species identified in chapter 4 that could be used for gene expression analysis and

SNP analysis in subsequent chapters. This has been achieved with a very strong assembly with high completeness and a very high N50. Further chromosome level scaffolding can be achieved with Hi-C techniques that are currently being explored at the time of writing. The production of a n50 megabase assembly for an earthworm species is a significant advancement for solving genomes with high allelic diversity. The methods developed for this assembly provide a template for maximising the quality, size and therefore usefulness from a large set of species that have largely been ignored or hindered due to the complexities of genome assembly. In particular both long repetitive DNA elements and large structural variation between allelic copies can now be overcome. Future sequencing could suffice with long read sequences and a sequenced 10x chromium library only, reducing cost and required starting DNA. The previous best published earthworm assemblies were for *E. fetida* with n50s of <10 Kbp (Zwarycz et al. 2015; Bhambri et al. 2017). Since completion of this work, a full chromosomal assembly has been generated for *E. andria* using Hi-C techniques described and has allowed the scaffolding of our *E. fetida* genome (hybrid species) to generate a chromosomal assembly also (Shao et al. 2020). This sets a further path for a final scaffolding of *A. caliginosa* through Hi-C techniques in the future to achieve chromosomal identification. As previously covered in chapter 1.4.5, *A.caliginosa* is used extensively in toxicology assessments of land and chemical use. The megabase assembly with comprehensive genome annotation provides a powerful tool for improving genomic and transcriptomic reports with improved transcript assessment and identification of promoter regions for genes of interest.

6. Transcriptomics of altitude: Response of *A. caliginosa* in three altitudes.

6.1. Earthworm transcriptomics.

6.1.1. **Transcriptomic insights into altitudinal adaptation.**

Transcriptomics provides an approach that lets us address the fundamental question of what changes of gene expression (if any) underpin acclimatization to high altitude conditions. The 'normal' homeostasis of gene expression will be altered for different suites of genes in response to various stimuli. In an ideal experiment design to study the influence of a particular stimuli, it would be tested in isolation from other changes. In its extreme this would represent a study on a clonal cell line supplied with identical culture and experimental conditions, save the one of interest. However, the difference between *in vitro* cellular and organism-based experiments will be significant as different tissues interact providing systemic feedback and modification of the organism physiology. The difference between *in vitro* and *in vivo* is even greater as organisms will alter their behaviour to adapt to different environments. It becomes very difficult to isolate a single stimulus, and yet this scenario represents real-world fluctuations that need to be investigated. Mapping the genetic regulation can be messy but is necessary to understand and contextualize the interactions between different stimuli.

We have previously discussed (Chapter 1) environmental factors that are likely to change with exposure to altitude within the temperate zone. In particular, Oxygen and temperature are linear facts that change, while precipitation, vegetation, and soil composition can vary on a site to site basis. Although soil geochemistry is an important stimulus that influences species survival, change is not directly associated with elevation and therefore is harder to link directly with altitude. Gaining altitude reduces air pressure in a predictable manager thereby reducing available Oxygen, whilst the temperature is also impacted by elevation leading to a ~6.5°C with every 1000 m height gained (USCAR and NSCAR 2013). These two stimuli are easy to isolate and test under laboratory conditions. By subjecting earthworm populations to each stimulus in isolation and in combination then comparing the expression profiles, it is possible to identify what patterns of expression alter in response to the challenge and the pathways that are differentially regulated. By comparing populations resident to both low and high altitude habitats that have been exposed to these different environmental stimuli for many generations it becomes possible to identify if genomic or epigenetics changes are driving the changes in expression pattern.

Tissues deliver specific functions, and as such have characteristic specific expression profiles. Different environmental stimuli will invoke tissue-specific changes in gene expression. An ideal investigation would look not only at how different populations react to environmental factors but also how the tissues within each population react. This would give a thorough assessment of the impact of environmental changes. However, there are practical considerations in attempting this complexity of assessment that arise in the exposure and sampling of populations. To control for temperature and Oxygen (particularly at lowered Oxygen and lowered temperature), populations need to be exposed in climatic chambers. Harvesting populations directly from within these climatic chambers is not practical and since the preservation of gene expression profiles is paramount the sample collection must be conducted as rapidly as possible. An earthworm's main tissues include; pharynx, seminal vesicles, crop and gizzard, nerve cord, gut, calciferous gland and body wall. These would all require their careful dissection in a short period of time. For low numbers of individuals, especially if dissection is performed in a preservative agent, the dissection of individual tissues from the main body could be performed. However, in a time sensitive scenario, such as before RNA degradation, harvesting individual tissues from large numbers of individuals would be not practically achievable.

6.1.2. **Using a genome to improve transcriptomics.**

Building *de novo* transcriptomes will often lead to a transcriptome being bloating with an artificially high number of transcripts representing assembly errors and haplotype heterozygosity. This is a result of inherent variability of read counts for differentially expressed genes, multiple copy genes or regions and alternative splicing events (Rey et al. 2019). The mapping of RNA-seq data directly onto a complete genome reduces overall transcriptome size and complexity leading to better-supported gene annotation identifying introns, exons, open reading frames (ORFs), and alternative splice variants represented in general feature format files (GFF) and general transfer format files (GTF). Consolidating gene annotation allows reads generated by RNA-seq to be mapped with greater efficacy to specific gene objects rather than having multiple matches to multiple splice variants thereby improving gene-based differential gene analysis whilst also support exon based and there-by splice variant analysis.

6.1.1. **Understanding multivariance analysis of altitude adaptation and acclimatisation.**

Elucidating the primary drivers of differential gene expression in a multi-factor experiment is complex and often required the identification of the secondary impact resulting from the stimuli. Identifying the effect of Oxygen availability and temperature on two discrete populations will be accompanied with variation that is challenging to control. There are four major factors impacting the transcriptome of each individual: genome, epigenome, microbiome and the environment.

The genome of earthworms will vary between individuals, though as discussed in Chapters 3 and 4, we have selected *A. caliginosa* to minimise the variation. The epigenome is unknown in these populations, but it is plausible that the worms from the top of the mountain have a different epigenome to the population found at the bottom of the mountain. The microbiome will vary based on the environment where the earthworm in resident, however, there is a growing body of evidence that symbiotic species are vertically transmitted and therefore link to an individual's ancestry (Pass 2015; Richards et al. 2019).

The experiment design incorporates a 6-month 'acclimatisation' period whereby the substrate (soil) and environmental conditions each population was maintained were equivalent to minimise impact of soil geochemistry and environmentally derived microbiome. Even in a laboratory-controlled environment, providing an identical condition for experimentation with only exact parameters varying without forcing individuals into stressful conditions is a challenge. Figure 50 shows a basic conceptual model of these four variable elements in the two populations. The Microbiome's influence on the transcriptome is highly complex and understanding of its interrelationship with host organisms is still not fully understood.

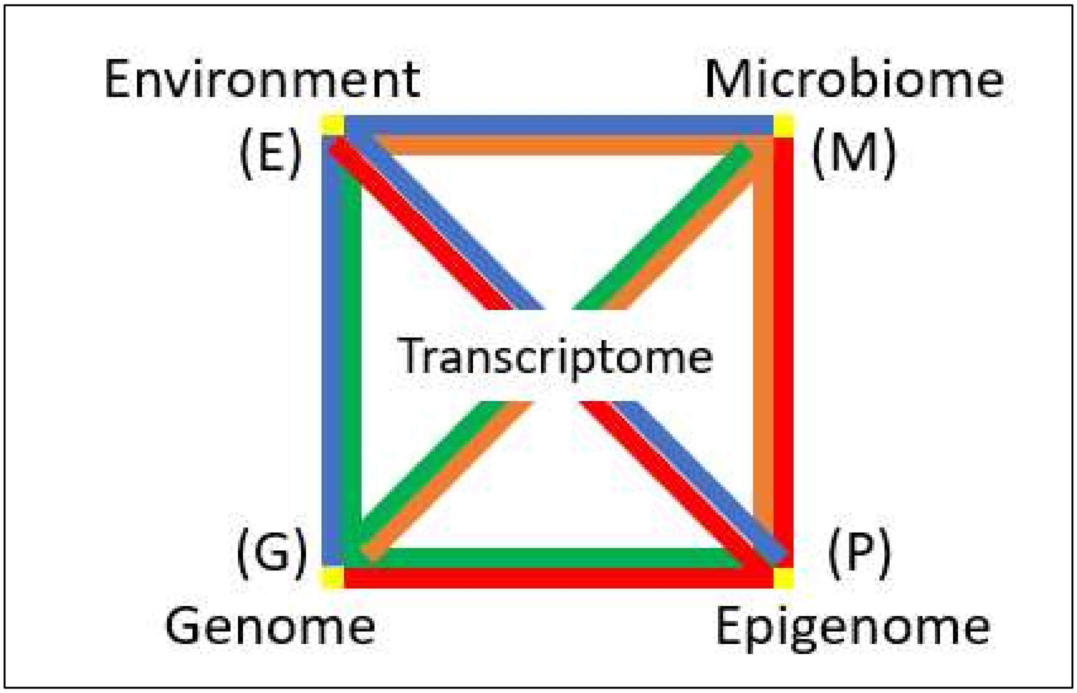


Figure 50: Modelling factors influencing response to environmental change. The Transcriptome is a result of factors including the Environment (E), the Microbiome (M), the Genome (G), and the Epigenome (P). Each factor directly affects each other, and their extent can be examined through analysis of the transcriptome.

6.1.2. **Experimental plan.**

To test the biological responses of *A. caliginosa* in populations from high and low altitude, live worms were collected from high altitude and low altitude in Pico, Azores. These populations were maintained in separate containers with a normalised soil for over 6 months before exposure to 6 conditions that simulate the altitudinal challenges of temperature and Oxygen for 2 weeks (51). Following harvesting and RNA extraction, RNA sequencing was performed. Read mapping, and gene counting was performed prior to differential analysis comparing experimental conditions with each other and RNA from worms preserved from the collection sites. Using the differential gene expression, ontology enrichment and pathway analysis I aimed to identify markers of adaption to high altitude representing cold or hypoxia tolerance.

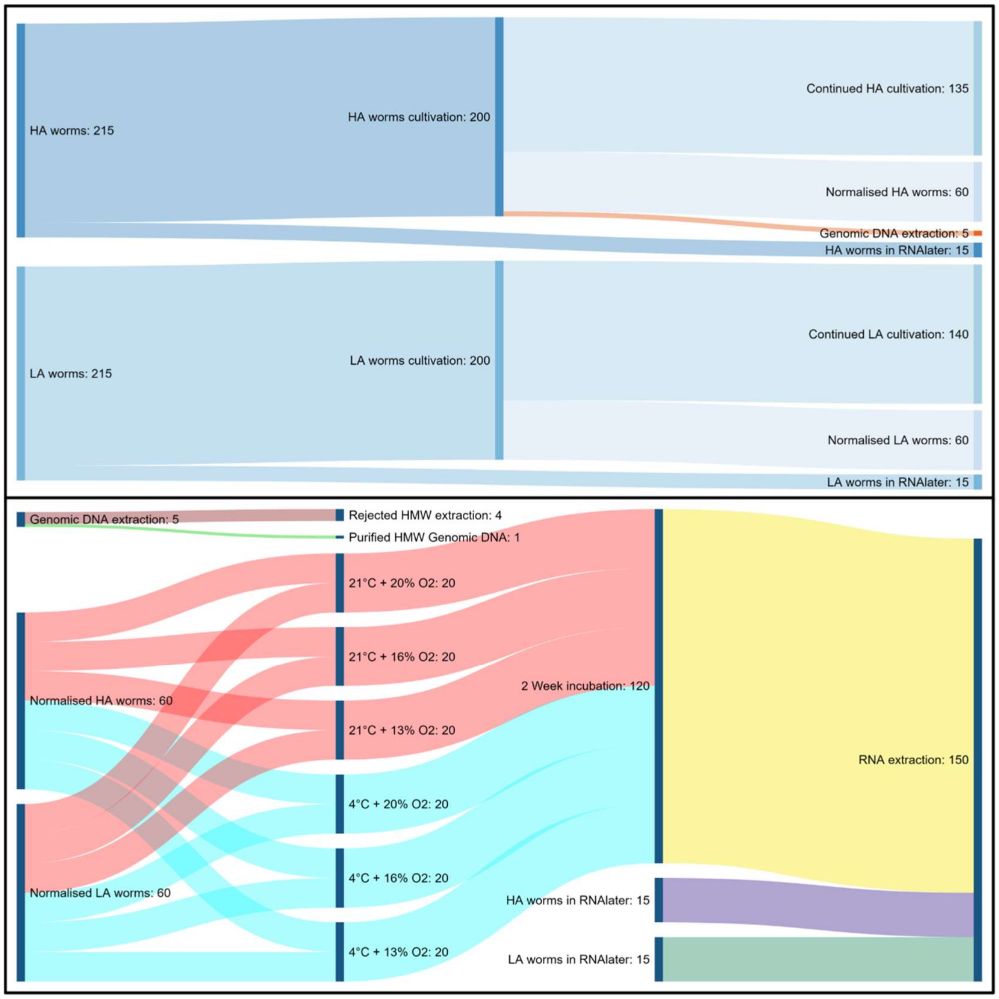


Figure 51: Experimental plan illustrating sample flow. In total 215 worms were collected from High altitude (HA) and Low altitude (LA). From each altitude, 15 were immediately preserved in RNA later, while 200 were kept alive for cultivation for 6 months. From each cultivated altitudinal population 10 worms were selected for each experimental condition, 6 conditions in total. RNA was harvested from these worms after two weeks along with RNA from the worms preserved in RNAlater.

6.2. Methods.

6.2.1. **Collection and cultivation of Samples.**

Live specimens of *A. caliginosa* (n>200) were collected in May 2018 from three sites in Pico, Azores, and pertinent location and environmental metadata recorded (Table 17). Ten worms from the 'High altitude' and ten worms from the 'Low altitude' native fauna site were collected and preserved on site in RNAlater. The remaining *A. caliginosa* were maintained in boxes of their native soil for transport to Cardiff, where individuals were transferred to large containers containing an equal ratio of topsoil, compost, and bark chippings (AG Robinson 2019, personal communication, CEH 14 January 2019). Manure from horses that had not been recently treated with de-worming agents or other veterinary medicines were added periodically as a food source (See Chapter 2.4.1). Soils were kept moist but not saturated with distilled water and the populations were disturbed minimally for 6 months to allow for normalisation. Both the High altitude (HA) site and the Low altitude field (LA) populations contained only *A. caliginosa*, however subsequent testing of individuals from the Low altitude native fauna site indicated the presence of the closely related *A. rosea* species. To ensure only *A. caliginosa* was used for testing, the LA field site was chosen for comparison with the HA population.

Table 17: Sampling locations for the High altitude and Low altitude populations.

Site	*Altitude*	*Soil depth (cm)*	*Vegetation*
High altitude (HA)	2073	20	Brush, grass, moss and patchy gravel
Low altitude – native fauna	340	2.5	Thick forest, dense roots
Low altitude – field (LA)	196	20	Thick grass field

6.2.2. **Exposures.**

Worms (n=10) selected at random from each population were washed in distilled water, weighed, and placed into separate 20 cm x 14 cm (2.8 L) boxes with mesh lids with the soil they had been living in for each exposure. Care was taken to ensure no extra worms or cocoons were transferred into each box with the transfer of the soil. Worms from HA and LA were then exposed to 6 conditions replicating high and low altitude for a two week period and soils were kept moist with distilled water (Figure 51 and Table 18). A ProOx model C21 (BioSpherix) was used to reduce and maintain the Oxygen level at the corresponding level according to altitudinal level and Innova 4230 refrigerated incubator (New Brunswick Scientific) was used to maintain the prescribed temperature. Where cooling and reduced Oxygen was required the ProOx was placed inside the refrigerated incubator. Condition exposures were staggered over a 6 week period to help with RNA extraction after the two-week incubation.

Table 18: Cultivated and preserved altitudinal populations and their experimental conditions.

Group ID	*Population source*	*Temperature exposure (°C)*	*Oxygen (%)*
1	Low altitude	21	13
2	High altitude	21	13
3	Low altitude	21	16
4	High altitude	21	16
5	Low altitude	21	20
6	High altitude	21	20
7	Low altitude	4	13
8	High altitude	4	13
9	Low altitude	4	16
10	High altitude	4	16
11	Low altitude	4	20
12	High altitude	4	20
B	Low altitude	Native source preservation	Native source preservation
A	High altitude	Native source preservation	Native source preservation

6.2.3. **RNA extraction and purification.**

Following the two-week incubation worms were immediately washed in distilled water, weighted and a 1 cm posterior section (5 segment from the tip with clear distance to the clitellum) was dissected out with a sterile scalpel. The section included multiple tissue types including epidermis, muscular tissue, blood vessel, intestine, nephridia and nerve cord tissue. The posterior sections were each placed into 600 µL of Trizol and ground with a pestle in an Eppendorf before freezing in liquid nitrogen. The remaining body was placed into an empty Eppendorf and frozen in liquid nitrogen. The six staggered incubations ensured only 20 worms were harvested at one time, minimising the time for RNA profiles to change. Harvesting was aided with Mr Hernadi to reduce harvesting time. Following freezing in liquid nitrogen, samples were transferred to a -80°C for storage until RNA purification could be performed.

Individual worm RNAs were purified in accordance with the Direct-Zol RNA MiniPrep kit from Zymo Research (Chapter 2.5.1). Following purification, samples were quantified, and RNA quality was assessed via HS RNA Qubit and Qiaxcel RNA cartridge (Qiagen).

6.2.4. **Library preparation**

The three samples from each population in each condition with the highest RNA concentration were chosen for RNA sequencing. In total 36 samples from experimental exposure and an additional 3 RNA samples from High altitude and 3 RNA samples from Low altitude sourced from the worms preserved in RNAlater were used in RNAseq library preparation.

Sample libraries were generated using a Roche KAPA mRNA HyperPrep kit as detailed in Chapter 2.7.1. In short, 100-150 ng of sample RNA was cleaned to remove bacterial and ribosomal RNA with mRNA capture beads, converted to cDNA, barcoded (19) and PCR amplified with 15 cycles.

Samples were assessed via an Agilent Tapestation on a D1000 chip to calculate molar concentrations of the amplified library. Samples were pooled evenly at 1 nM (approximately 6 µL per sample) to a total volume of 252 µL for 42 samples and underwent two SPRI bead cleans at a 1:1 ratio bead to sample (2.7.2) to remove primer-dimers. The pooled samples' concentration was measured prior to sequencing with an HS DNA Qubit.

Table 19: Sample information and RNA indexes.

Group ID	***Replicate***	***Sample ID***	***Sample Num.***	***Dual-Index***	***P5 Index (3'-5')***	***P7 Index (3'-5')***
A	1	a6	1	A7	TATAGCCT	CTGAAGCT
A	2	a7	2	A8	TATAGCCT	TAATGCGC
A	3	a12	3	A9	TATAGCCT	CGGCTATG
B	1	b2	4	A10	TATAGCCT	TCCGCGAA
B	2	b4	5	A11	TATAGCCT	TCTCGCGC
B	3	b13	6	A12	TATAGCCT	AGCGATAG
1	1	1_3	7	B7	ATAGAGGC	CTGAAGCT
1	2	1_5	8	B8	ATAGAGGC	TAATGCGC
1	3	1_9	9	B9	ATAGAGGC	CGGCTATG
2	1	2_7	10	B10	ATAGAGGC	TCCGCGAA
2	2	2_8	11	B11	ATAGAGGC	TCTCGCGC
2	3	2_10	12	B12	ATAGAGGC	AGCGATAG
3	1	3_5	13	C7	CCTATCCT	CTGAAGCT
3	2	3_6	14	C8	CCTATCCT	TAATGCGC
3	3	3_9	15	C9	CCTATCCT	CGGCTATG
4	1	4_6	16	C10	CCTATCCT	TCCGCGAA
4	2	4_9	17	C11	CCTATCCT	TCTCGCGC
4	3	4_10	18	C12	CCTATCCT	AGCGATAG
5	1	5_1	19	D7	GGCTCTGA	CTGAAGCT
5	2	5_2	20	D8	GGCTCTGA	TAATGCGC
5	3	5_3	21	D9	GGCTCTGA	CGGCTATG
6	1	6_1	22	D10	GGCTCTGA	TCCGCGAA
6	2	6_3	23	D11	GGCTCTGA	TCTCGCGC
6	3	6_5	24	D12	GGCTCTGA	AGCGATAG
7	1	7_1	25	E7	AGGCGAAG	CTGAAGCT
7	2	7_2	26	E8	AGGCGAAG	TAATGCGC
7	3	7_10	27	E9	AGGCGAAG	CGGCTATG
8	1	8_1	28	E10	AGGCGAAG	TCCGCGAA
8	2	8_4	29	E11	AGGCGAAG	TCTCGCGC
8	3	8_5	30	E12	AGGCGAAG	AGCGATAG
9	1	9_3	31	F7	TAATCTTA	CTGAAGCT
9	2	9_4	32	F8	TAATCTTA	TAATGCGC
9	3	9_5	33	F9	TAATCTTA	CGGCTATG
10	1	10_1	34	F10	TAATCTTA	TCCGCGAA
10	2	10_2	35	F11	TAATCTTA	TCTCGCGC
10	3	10_5	36	F12	TAATCTTA	AGCGATAG
11	1	11_6	37	G7	CAGGACGT	CTGAAGCT
11	2	11_8	38	G8	CAGGACGT	TAATGCGC
11	3	11_9	39	G9	CAGGACGT	CGGCTATG
12	1	12_2	40	G10	CAGGACGT	TCCGCGAA
12	2	12_4	41	G11	CAGGACGT	TCTCGCGC
12	3	12_5	42	G12	CAGGACGT	AGCGATAG

6.2.5. **Sequencing.**

The 42 pooled samples were sequenced over two cartridge on a NextSeq 550 High capacity chip 1 x75 bp to generate a sequencing depth of ~ 18 million reads per sample. After the first sequencing run, read coverage per sample and mapping statistics were assessed as detailed below in 6.2.6 to check that all libraries were adequately generated and over duplication was not present in any sample before re-pooling to balance overall sequence coverage output per sample. Reads from both runs went automatic demultiplexing as part of the Illumina NextSeq protocol prior to quality control, merging of individuals reads and downstream differential expression analysis.

6.2.6. **Quality control and differential expression analysis.**

Specific parameters for informatic processing used for differential analysis are detailed in Chapter 2.8. Demultiplexed samples from each run underwent quality trimming with Trimmomatic (2.8.3.). All trimmed reads for each sample that had been generated from each run were concatenated together before downstream analysis.

Utilising the unmasked genome (ApCa_Genome_maksed_070519.fasta) and corresponding gtf (ApCa_Genome_maksed_070519.gtf) generated in chapter 5, STAR (v2.7) was used to index the genome for mapping of the RNAseq reads (Dobin et al. 2013) [6]. Read duplicates were marked and removed with Piccard. Sample reads were then mapped to the genome and the resulting count table formatted for analysis with SARTools and DESeq2 (Varet et al. 2016).

6.2.7. **Gene enrichment and pathway analysis.**

Functional annotation of the gene regions derived from the annotated genome developed in Chapter 5 was performed through homology analysis using protein sequences derived from the predicted CDS analysed against several well-annotated databases including Sprot (reviewed Swiss-Prot Eukaryota) and proteomes representing: human (UP000005640), mouse (UP000000589), fly (UP000000803), nematode (UP000001940) and yeast (UP000000559) (databases downloaded 10th December 2019). Homology was derived using blastp with an E-value cut off 10^{-5} and the results tabulated. For gene enrichment and pathway analysis differential gene expression counts for each gene were mapped to the human Uniprot ID where possible as human resources contain the most up-to-date gene functional information. Lists of genes for each differential expression analysis were generated with a Log2 fold cut off of <-1.4 and >1.4, with an FDR P_{adj} cut off of <0.05. Six-way Venn-Diagrams were created to identify patterns of common genes between differentially regulated gene sets (Heberle et al. 2015). The list of successfully annotated (with a human Uniprot ID) were imputed to GOnet with fold change

[6] Mapping with STAR was also run for the masked genome.

to generate a network of GO term enrichment and links to genes where applicable (Pomaznoy et al. 2018). For this, a background list of genes was used of the successfully annotated Human UniProt IDs for the genome gene objects. Multi-nodal networks were assessed in Cytoscape v3.8.0 and annotated with AutoAnnotate. Not all detail of individual genes are not shown on gene network maps to minimise complexity but are shown in supplementary Appendix file 1.1.

6.2.8. **Identifying shared genes and genes of interest.**

Common genes across three groups of differential expression were investigated using DiVenn, (Oxygen variation in HA and LA populations at 4°C, Oxygen variation in HA and LA populations at 21°C and temperature variation across 13, 16 and 20% Oxygen comparisons) (Liang et al. 2019).

The differentially expressed gene lists used in gene enrichment and pathway analysis were also analysed in STRING (v11.0) (Szklarczyk et al. 2019). Network edges were set to indicate confidence and molecular action where network size was small enough for computation. Network statistics were calculated as part of STRING's function including an interaction enrichment calculation for the likelihood of random protein interactions.

6.3. Results.

6.3.1. **Worm weights.**

Almost all worms experienced a small loss in weight that is probably associated with transplantation stress (52). This weight loss was generally between 0 and 70 mg per worm (0-17% body weight). However, this was higher in the HA worms exposed to 21°C and 20.1% Oxygen which had 90 mg per worm weight loss (22%). This was attributed to reduced food consumption, apparent during dissection for RNA harvesting where worm guts were less filled.

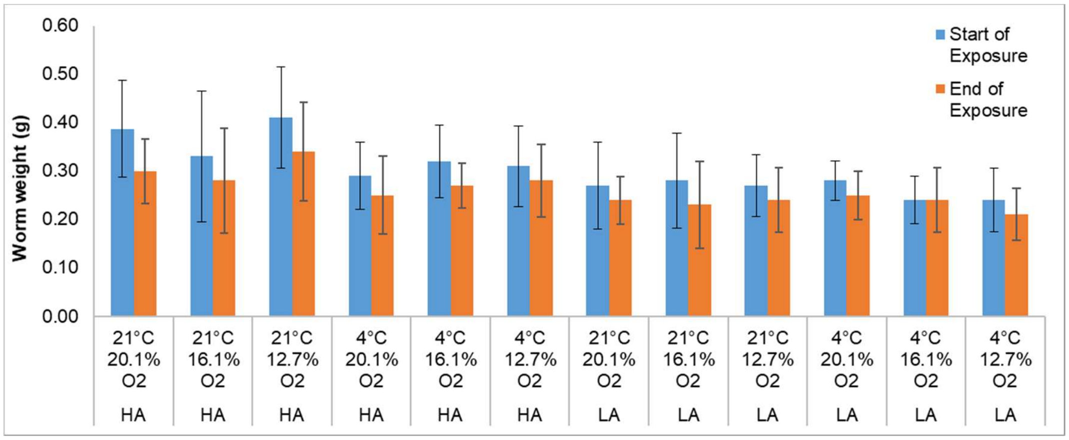


Figure 52: Experimental worm weights. Weight before and after experimental exposure for each population and condition. HA worms were slightly larger than the LA worms.

6.3.2. **Sequencing and Quality Trimming.**

In the first sequencing run, 372 Million single reads passed the Quality filter (28.23 Gb). An average of 8.8 Million reads per sample were generated with sample 8_5 over-clustering to produce over 21 Million reads. Samples 8_4, 4_9, 4_10, 6_1 and 6_5 under-clustered. Reads counts were used to determine re-pooling concentrations to achieve an even overall read count from both sequencing runs. The Second sequencing run generated 370 million reads passing the Quality filter (34.19 Gb). There was heavy over-clustering in sample 4_9 that reduced the potential average normalised counts per sample that could have been obtained from two sequencing runs. Samples with over 20 million reads were subsampled following quality trimming with Trimmomatic to reduce the effect of over clustering. Total read counts used in RNAseq mapping are shown in Figure 54.

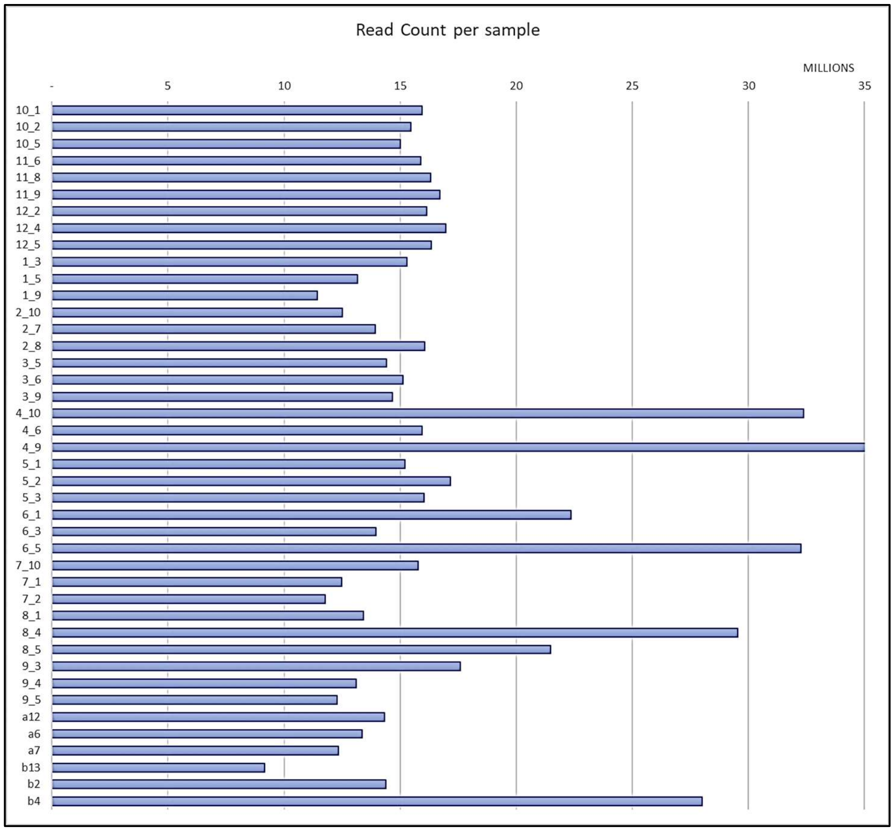


Figure 53: Read Count per sample obtained from sequencing run 1.

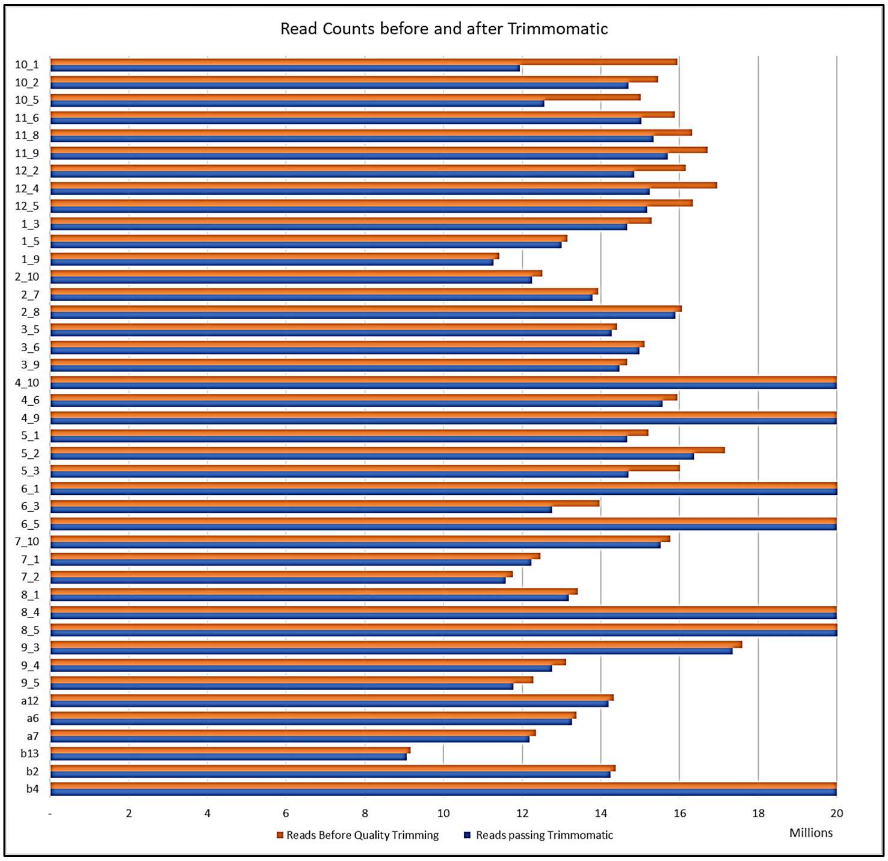


Figure 54:Number of reads lost during Trimmomatic for each sample. Read count loss was quite low for most samples.

6.3.3. **Read Mapping.**

STAR alignment scores were generated to assess the genetic similarity of individuals to the reference genome individual. All sample generated around 80% uniquely mapped reads indicating there is little to no large scale population differences between the High and Low altitude populations that could impact on the differential analysis of expressed reads (Figure 55) [7]. It also indicates no individual library had failed or produced a high level of duplicate reads.

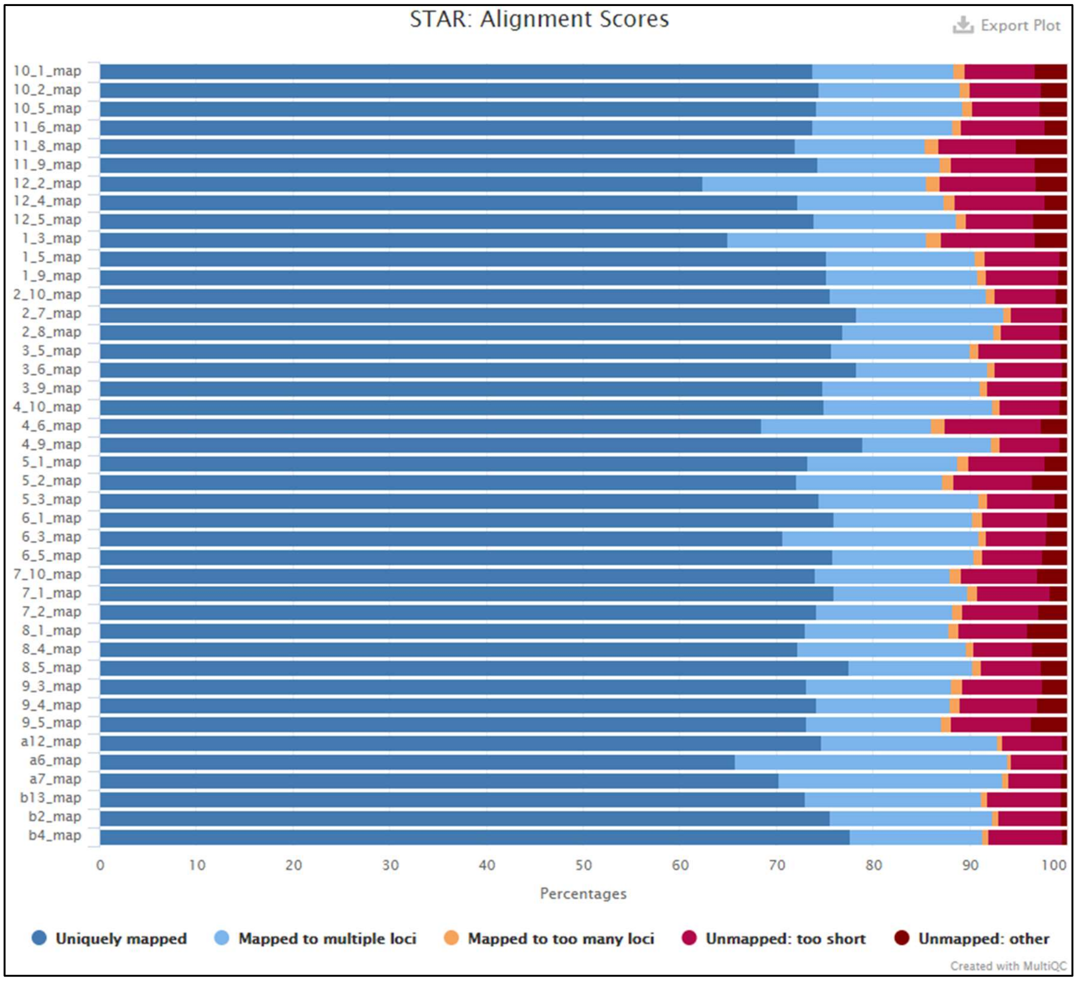


Figure 55: STAR mapping percentages for the 42 samples. Most samples mapped uniquely 70% of reads. The evenness of mapping across the samples indicate Differential count analysis will not be impacted by variation in individual genome variation.

6.3.4. **Count Annotation.**

Of the 42,566 gene objects identified by OmicsBox (Chapter 2.8.1), 5,238 genes were directly identified and a further 22,122 hypothetical or putative genes were identified through Blast2GO

[7] As the masked genome mapped with similar but poorer (10%<) unique mapping, the unmasked genome was used for RNAseq analysis.

of which Interpro Scan identified 26,368 GO terms. To try and identify all genes Blastp was performed against protein sequences translated from predicted CDS (E value cut off 10^{-5}) against Sprot, human, mouse, fly, nematode and yeast. These blast scores were compiled into a single database which gave naming president in the order Sprot > Human > Mouse > Fly > Nematode > Yeast > GO term. In some cases, no identification could be found. Of the 42,566 genes, 25,556 genes were positively identified. For the remaining 17,010 gene objects 3889 were associated with a Gene Ontology term (GO ID) and 13,121 remained unidentified. The 25,556 identified genes include multiple versions or isoforms of the same gene which can be reduced 13,371 gene names (Figure 56).

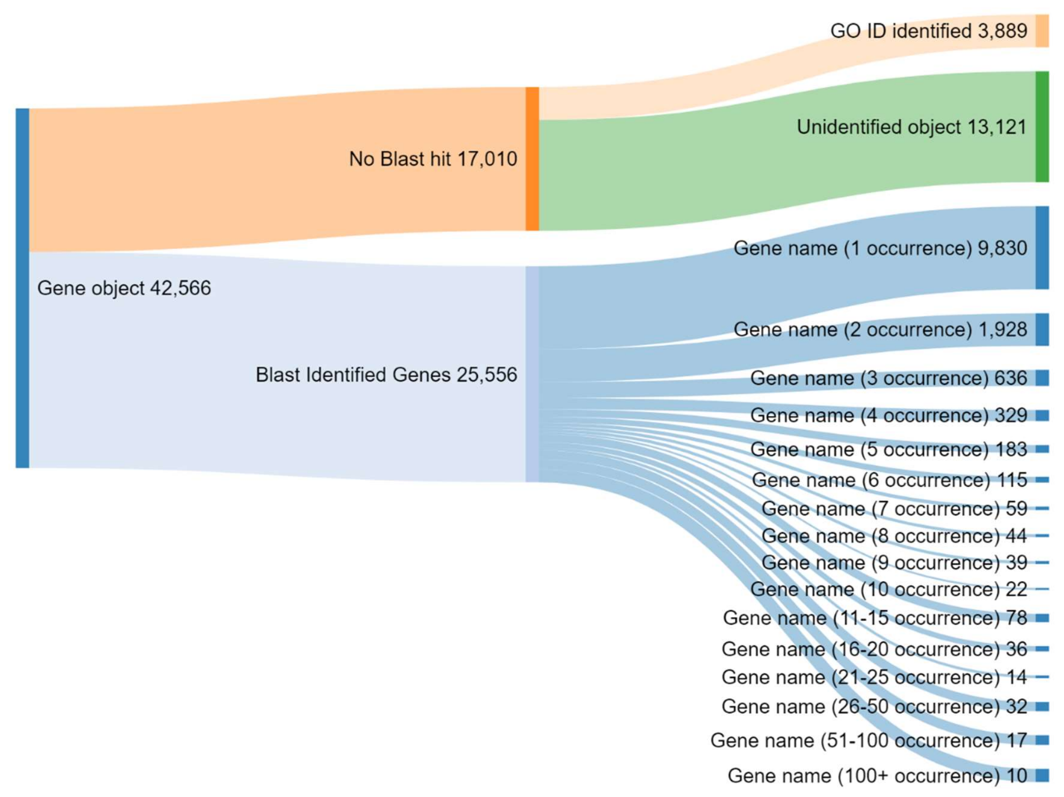


Figure 56:Breakdown of Gene object identification results.

6.3.5. **Differential statistics Counts overview.**

The mapped reads had a total read count per sample of between roughly 4 and 11 million reads (Figure 57). Variation in total read count between replicates comes from variation in sequencing depth achieved for each sample rather than the efficiency of mapping. Both HA and LA populations had a similar average total road count per sample.

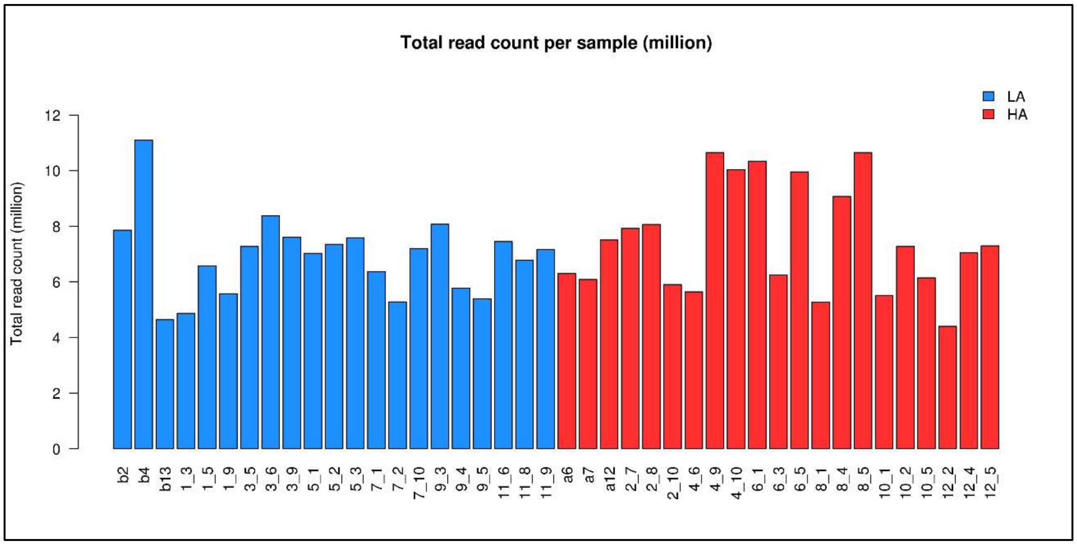


Figure 57: Total read count per sample (million). Count varies from ~4 and ~11 million reads intraindividual.

The percentage of null read counts should be approximately similar in perfect sample replicates. This is important as null read count features are not accounted for in DESeq2 analysis. For the 42 samples, null reads account for 7.07% (3008 features) shown by the dashed line in Figure 58

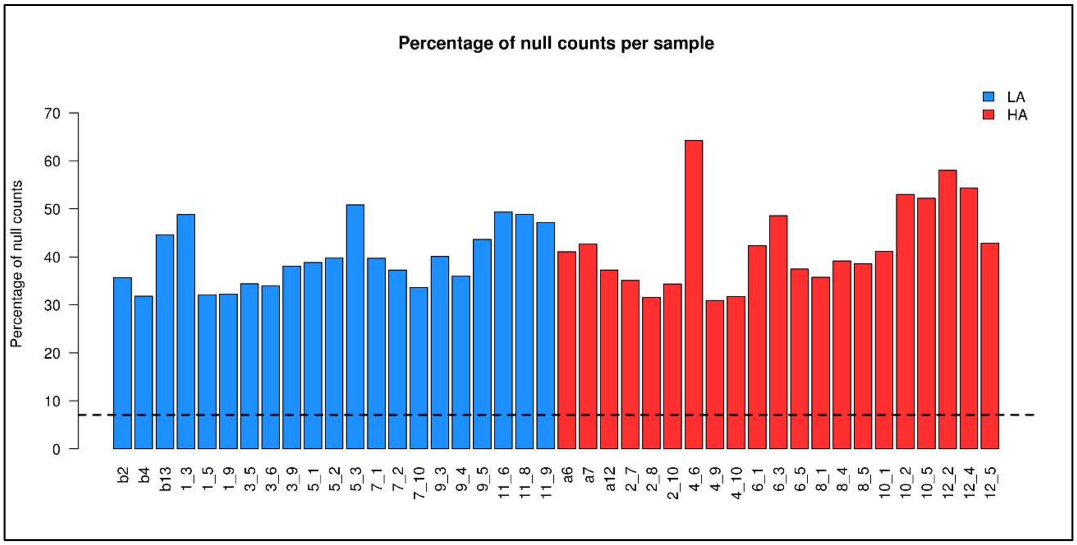


Figure 58: Percentage of read counts with null counts per sample. Percentage null counts varies between ~30% and ~65% gene objects. The dashed line indicates the exclusion of null read count features level for downstream DESeq2 analysis

A histogram was generated of the average sample read count per gene object (Log$_{10}$) to indicate the distribution of read counts per gene object. Almost all gene objects had either no count (17,522 per sample average) or had less than 10,000 counts per gene object (24,966 per sample average), (Figure 59).

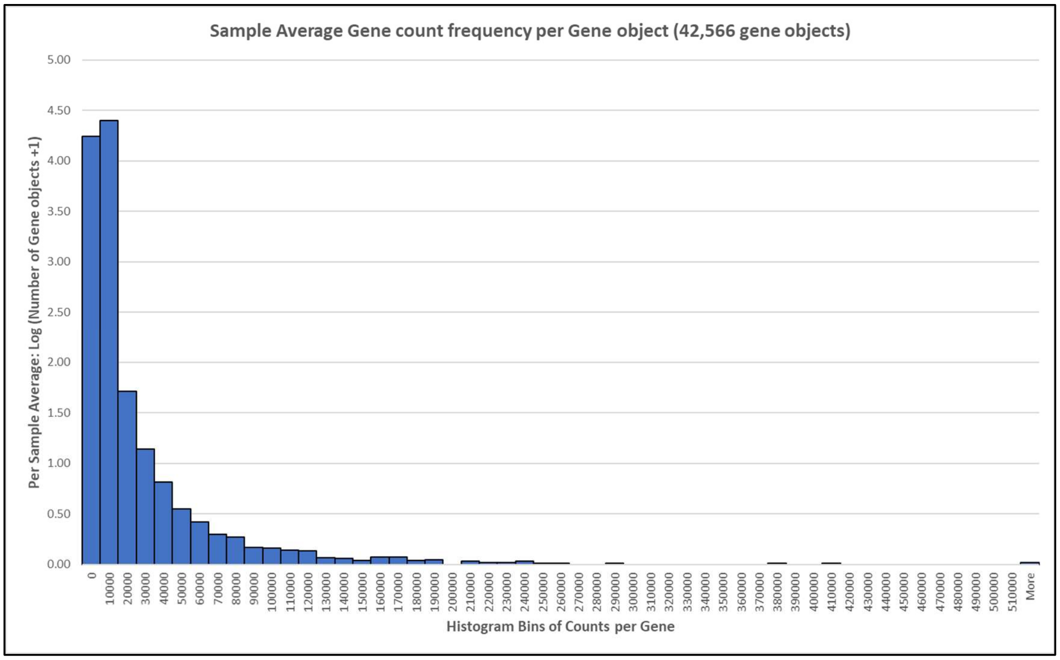


Figure 59: Histogram of the Average Gene Count per sample (Log$_{10}$ Counts +1). Most gene features have either a null count or below 10000 counts per gene object.

A density plot of raw counts prior to filtering indicates the high level of genes within each sample that have either no or very low expression. There is also some variation in distribution patterns between samples (Figure 60).

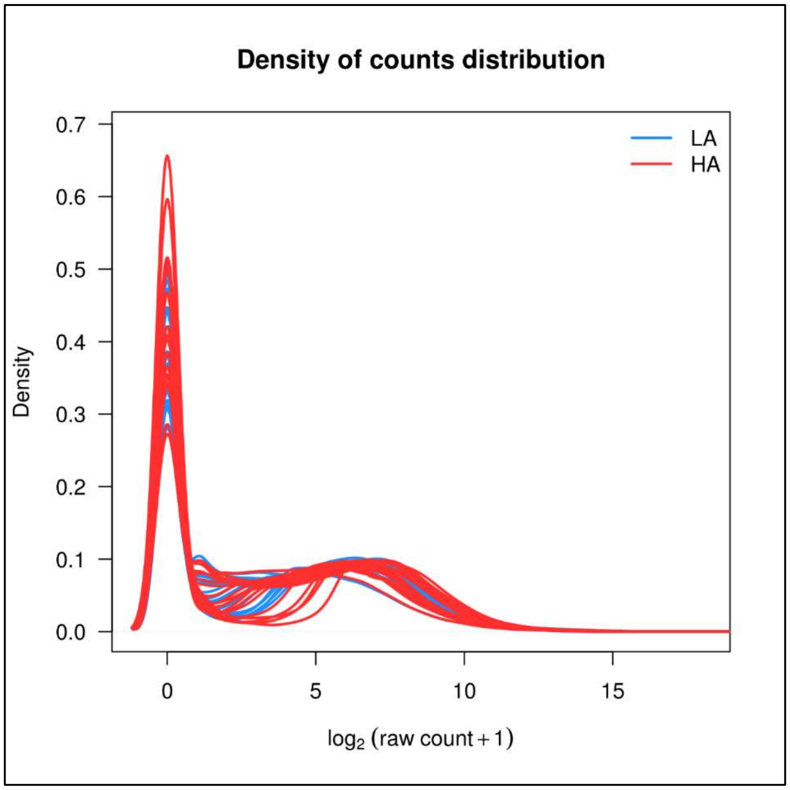


Figure 60: Distribution of read count for each sample transformed with log$_2$(counts + 1). Populations and replicates should have similar distributions.

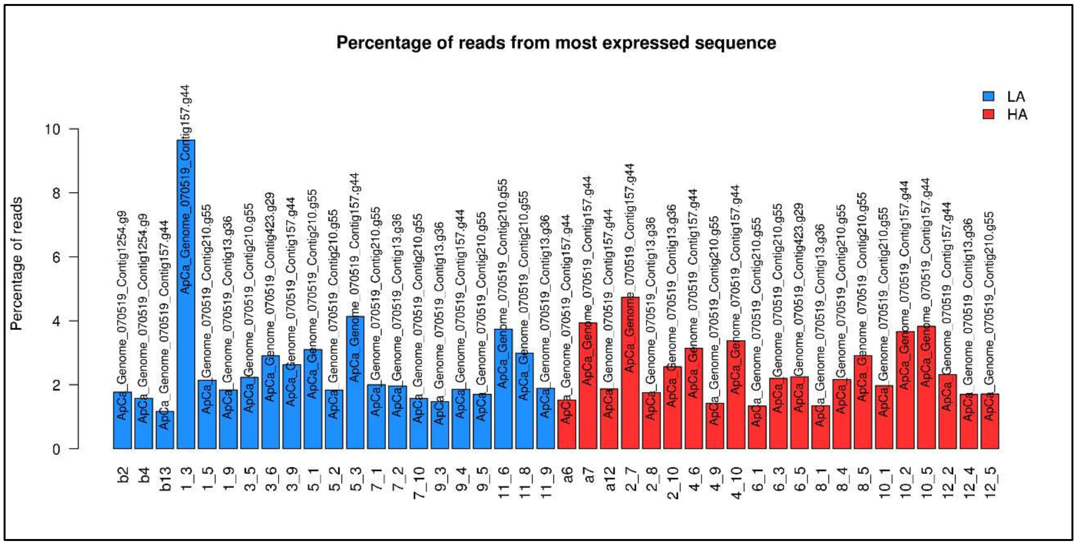


Figure 61: The most expressed sequence from each sample. Of the 42 samples the 16 were Actin (Contig157.g44 and Contig1254.g9), 16 were a putative phototransduction enzyme (contig 210.g55), 8 were a phosphoenolpyruvate carboxykinase (contig13.g36) and 2 were a punitive protein binding enzyme (contig423.g29). While the percentage of reads from the most expressed sequence can be expected to vary between condition, we would expect this to be similar within replicates. Sample 1_3 is therefore an outlier.

The percentage of reads from the most expressed sequenced in each sample is roughly similar (1-4%), although sample 1_3 does stand apart with a much higher percentage (>9%), (Figure 61).

Following DESeq2 analysis, Principal Component analysis was run. While there was no obvious link to the spread of individuals in Principle component 1 (PC1), it does account for ~25% of the variance. PC2 (~10%) separated the native populations brought back in RNAlater from the experimental samples (Figure 62). PC3 (7%) separated the High and low altitude populations but the native populations remained with the population they had originated from (Figure 62).

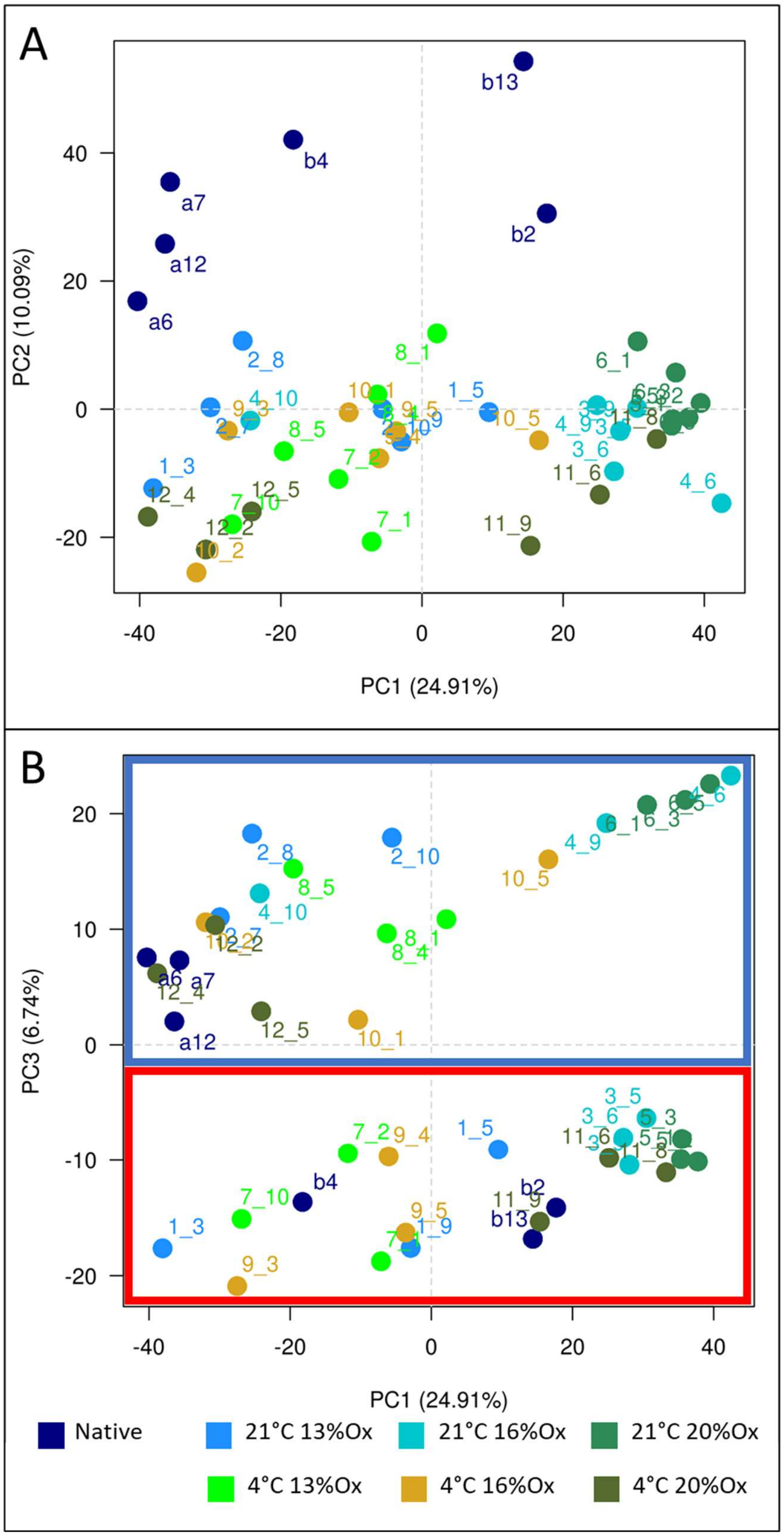


*Figure 62: PCA separation of RNAseq data derived from A. caliginosa showing distinct separation of HA and LA populations with PC3. **A**) PC2 splits Native samples from the experimental individuals. **B**) PC3 generates a clear separation of the LA population (LA; 1,3,5,7,9,11, boxed in red) and the HA population (HA; 2,4,6,8,10,12, boxed in blue), most experimental conditions within each population are closely grouped, though the conditions of 4°C at 13% Oxygen and 16% Oxygen are more separated. A full key to the specimen codes shown on the Figure are given in Table 2.*

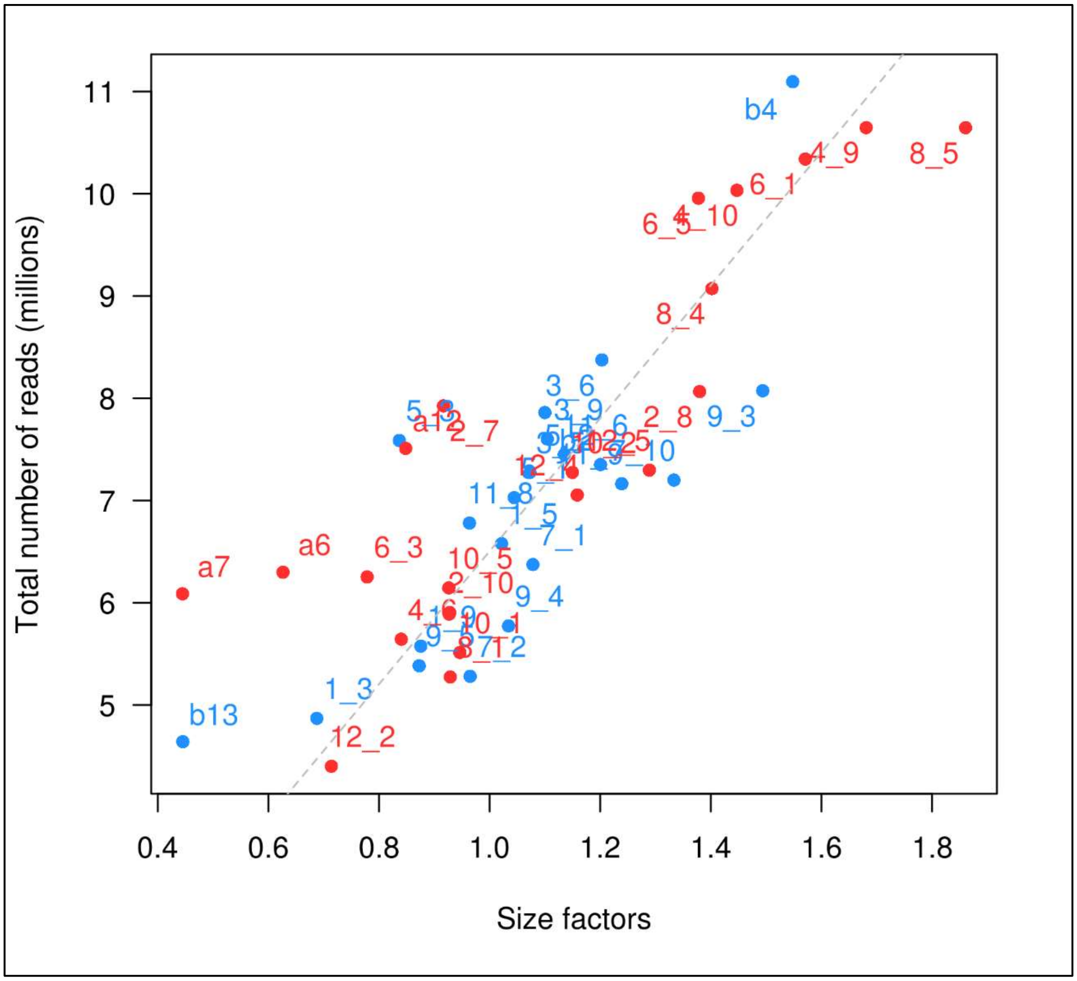


Figure 63: Size factor vs total number of reads utilised. Scaling calculation used by DESeq2 to compensate total number of read counts for uneven library size and normalise read counts. High altitude individuals are indicated in Red and Low altitude individuals in Blue.

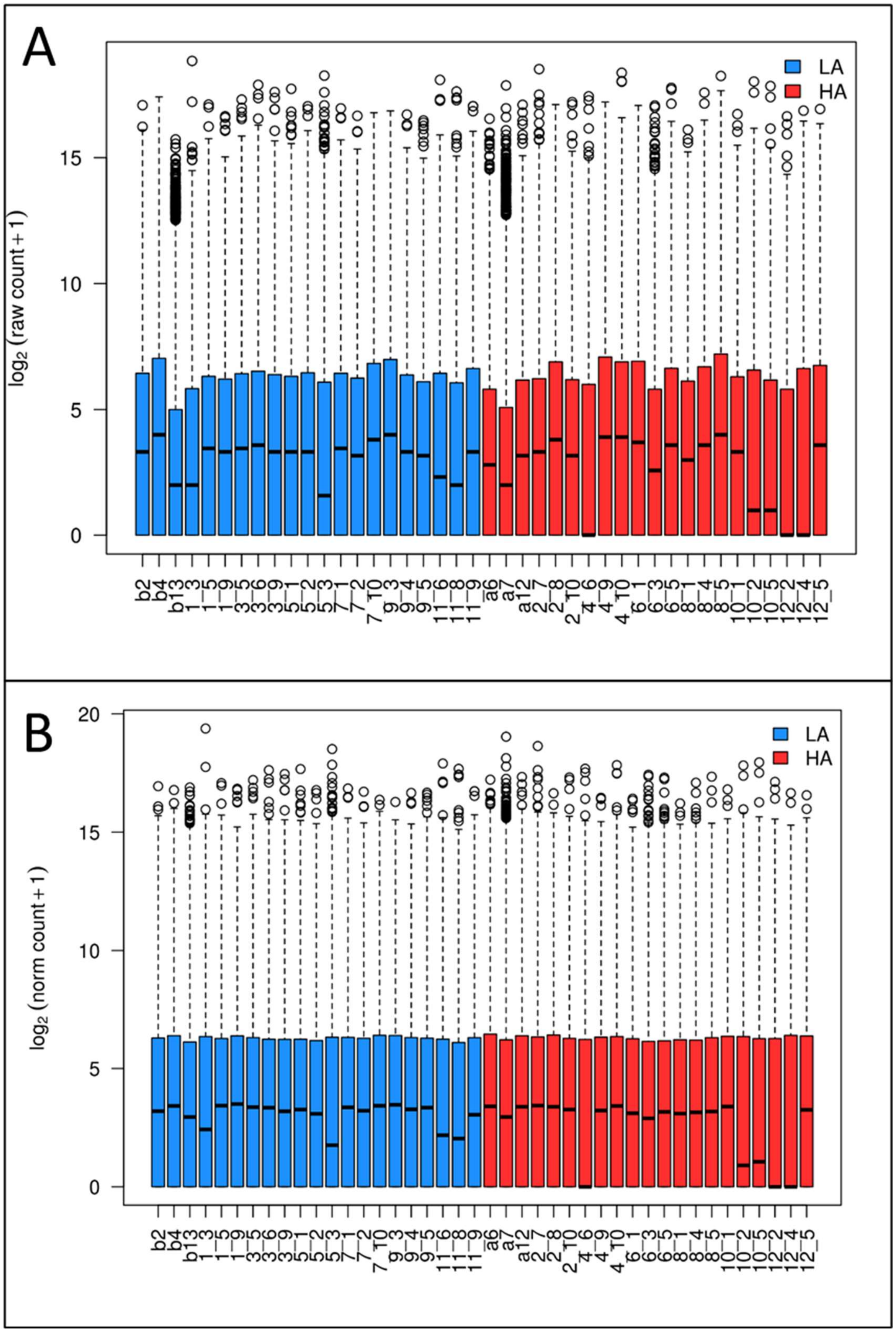
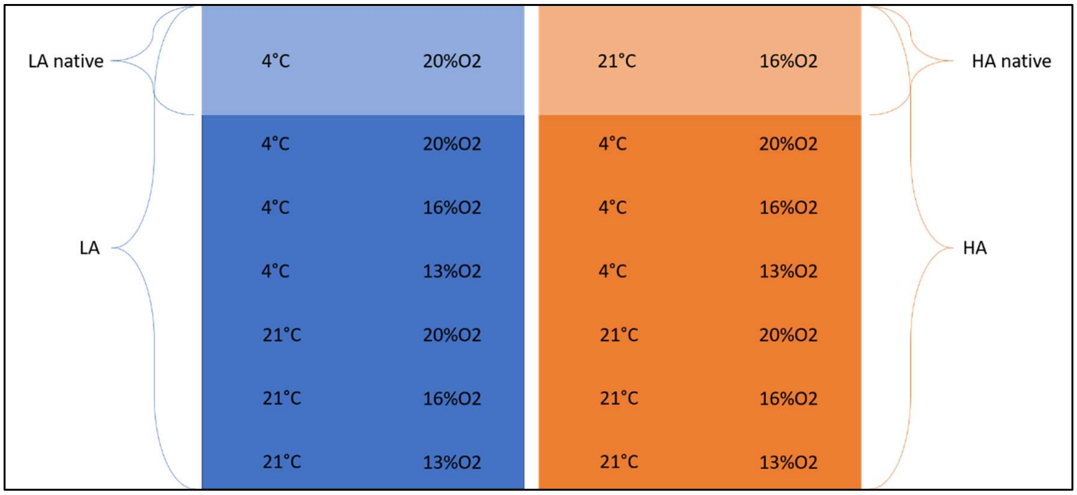


Figure 64: A) Distribution of reads before normalisation. B) Distribution of reads after normalisation of counts across the 42 samples.

6.3.6. **Overview of differential statistics.**

In a large multifactorial experiment, factors need separating to identify component effects that might be lost in a larger comparison (Figure 65). While larger comparisons e.g. HA experimental worms Vs LA experimental worms might indicate a pattern of differentially expressed genes, some differentially expressed genes will be lost in the 'noise' of general variation. With these broader and less focused gene expression comparisons, gene ontology and network analysis can be used to provide a broad overview of pathways that are under change. The loss of sensitivity of individual gene flux is balanced with the increase in sample size.



LA native	4°C	20%O2	21°C	16%O2	HA native
LA	4°C	20%O2	4°C	20%O2	HA
	4°C	16%O2	4°C	16%O2	
	4°C	13%O2	4°C	13%O2	
	21°C	20%O2	21°C	20%O2	
	21°C	16%O2	21°C	16%O2	
	21°C	13%O2	21°C	13%O2	

Figure 65: Experimental variables to consider.

As part of the broader testing of conditions, the populations of HA and LA were compared as were worms exposed at 4°C compared against 21°C and worms at the three Oxygen concentrations of 13, 16 and 20% were all compared. In addition to these broad tests, the Native populations were directly compared. The number of differentially regulated genes are reported in Table 20.

Table 20: Up and Downregulated genes Padj <0.05 and log2 fold change <-1.4 and >1.4. differentially regulated Gene (DEG) generated by each comparison are provided with an appropriate reference number DEG1-DEG6.

Test	*Number of samples*	*Up and Downregulated genes*
LA native Vs HA native [DEG1]	3 Vs 3	996
LA Vs HA [DEG2]	18 Vs 18	470
21°C Vs 4°C (HA + LA) [DEG3]	9 Vs 9	529 (HA) – 359 (LA)
13%O_2 Vs 16%O_2 (HA + LA) [DEG4]	6 Vs 6	25 (HA) – 71 (LA)
13%O_2 Vs 20%O_2 (HA + LA) [DEG5]	6 Vs 6	96 (HA) – 26 (LA)
16%O_2 Vs 20%O_2 (HA + LA) [DEG6]	6 Vs 6	2 (HA) – 14 (LA)

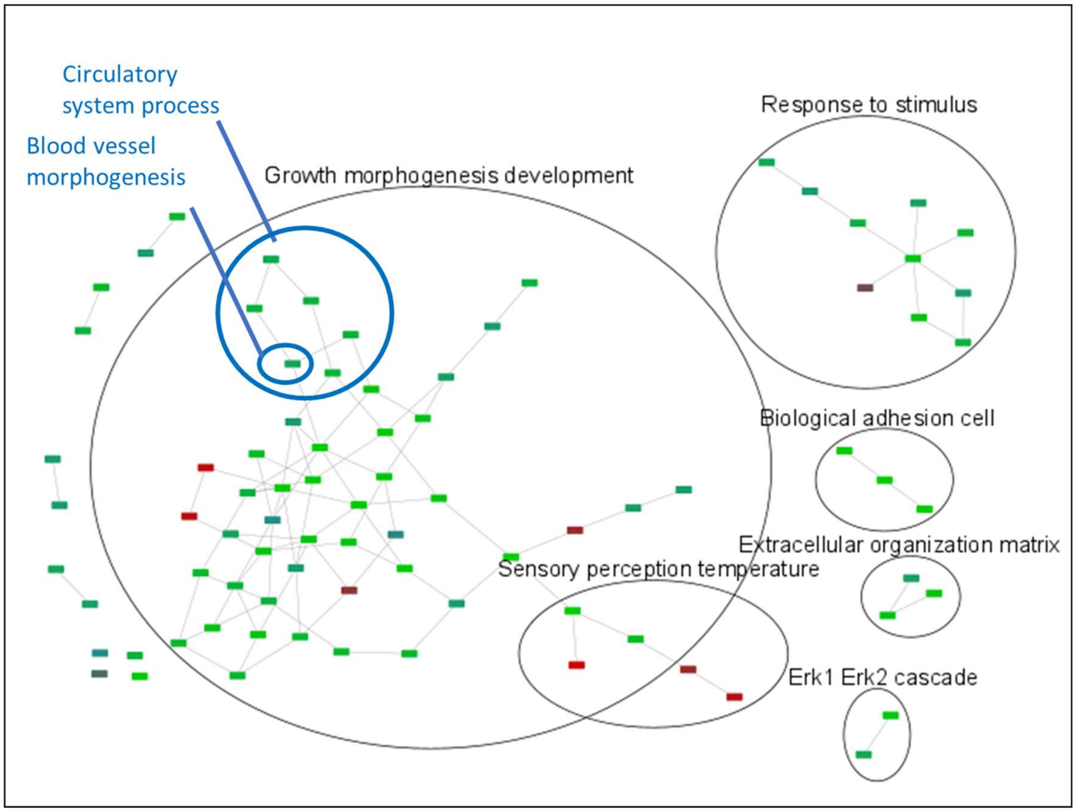

Circulatory system process
Blood vessel morphogenesis
Growth morphogenesis development
Response to stimulus
Biological adhesion cell
Extracellular organization matrix
Sensory perception temperature
Erk1 Erk2 cascade

Figure 66: LA native individuals Vs HA native individuals [DEG1]. Gene ontology network showing 3 major groups of processes including Growth morphogenesis development, Sensory perception of temperature and Response to stimulus. GO terms of interest are highlighted in blue. GO terms are coloured Red to Green indicating a low FDR adjusted P-value to high (cut off <0.05).

There were six major biological processes identified when comparing the LA native population with the HA native population (Figure 66). These included: “Growth morphogenesis development”, “Response to stimulus” and “Sensory perception to temperature” (including “Circulatory system process” and “Blood vessel morphogenesis”). In contrast to the LA native population, the HA native population is going through an increase in growth-related activities, including the development of blood vessels.

Gene ontology of the comparison between the LA population with the HA population identified only a few enriched processes, which included “Extracellular matrix organisation” and “cell morphogenesis” (Figure 67). The comparison of HA and LA populations responses to temperature indicated enrichment only for genes associated with “Response to stimulus” (Figure 68). No enriched GO terms were identified for any of the broad Oxygen variation comparisons.

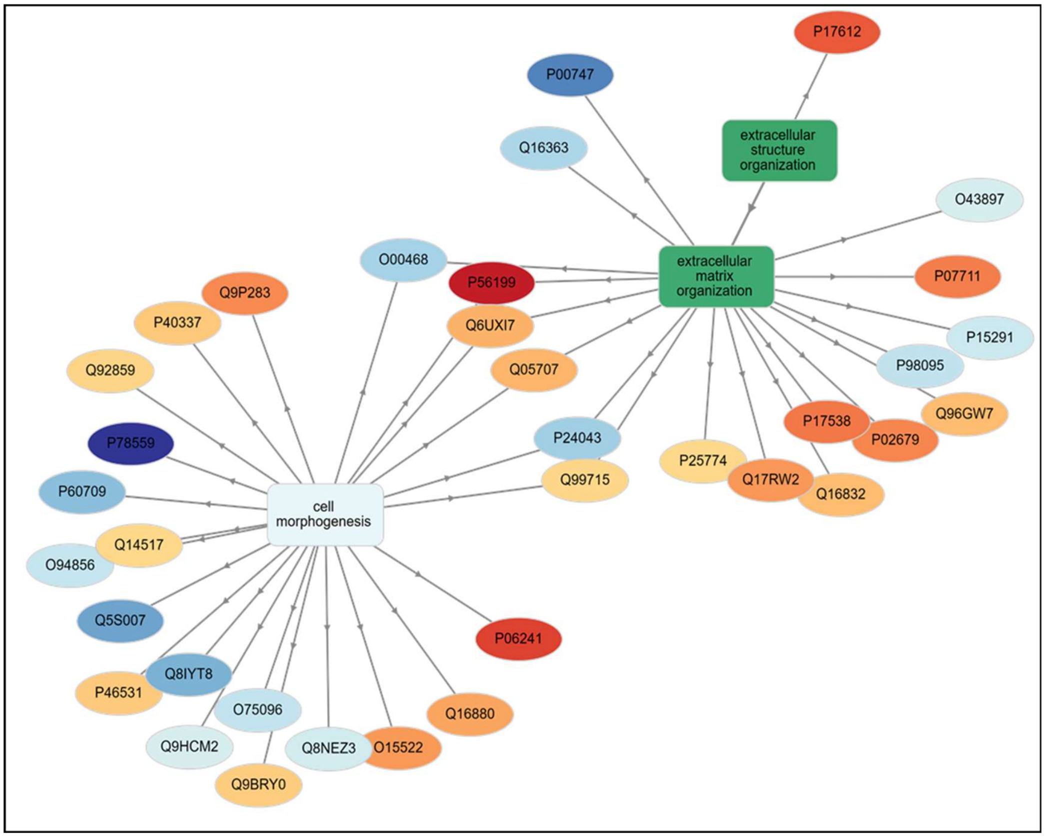


Figure 67: All HA experimental individuals Vs All LA experimental individuals [DEG2]. Gene ontology network showing genes and associated GO terms. Far fewer processes are identified in this broader comparison of populations. Gene fold change is indicated (darker blue is more downregulated and darker red is more upregulated). There are only a few GO processes identified including "Extracellular matrix organisation" and "Cell morphogenesis".

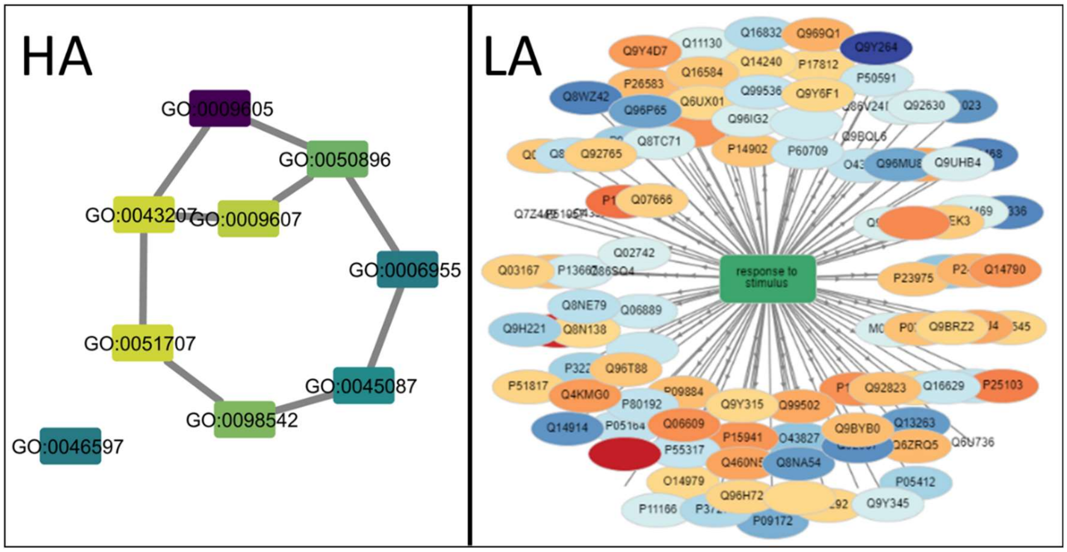


Figure 68: Comparison of HA and LA populations maintained 21°C and 4°C independent of Oxygen level [DEG3]. A small enriched network in the HA population associated with "response to stimulus" is observed, while only a singular GO term of "response to stimulus" is identified in the LA population. GO terms are coloured yellow to purple indicating a low FDR adjusted P-value to high (cut off <0.05). Gene fold change is indicated (darker blue is more downregulated and darker red is more upregulated).

6.3.7. **Differential statistics – Role of Temperature.**

Differential statistics were performed with DESeq and EdgeR to identify differentially expressed genes in the HA and LA populations between 4°C and 21°C, at 3 Oxygen levels 13, 16 and 20% (Figure 69).

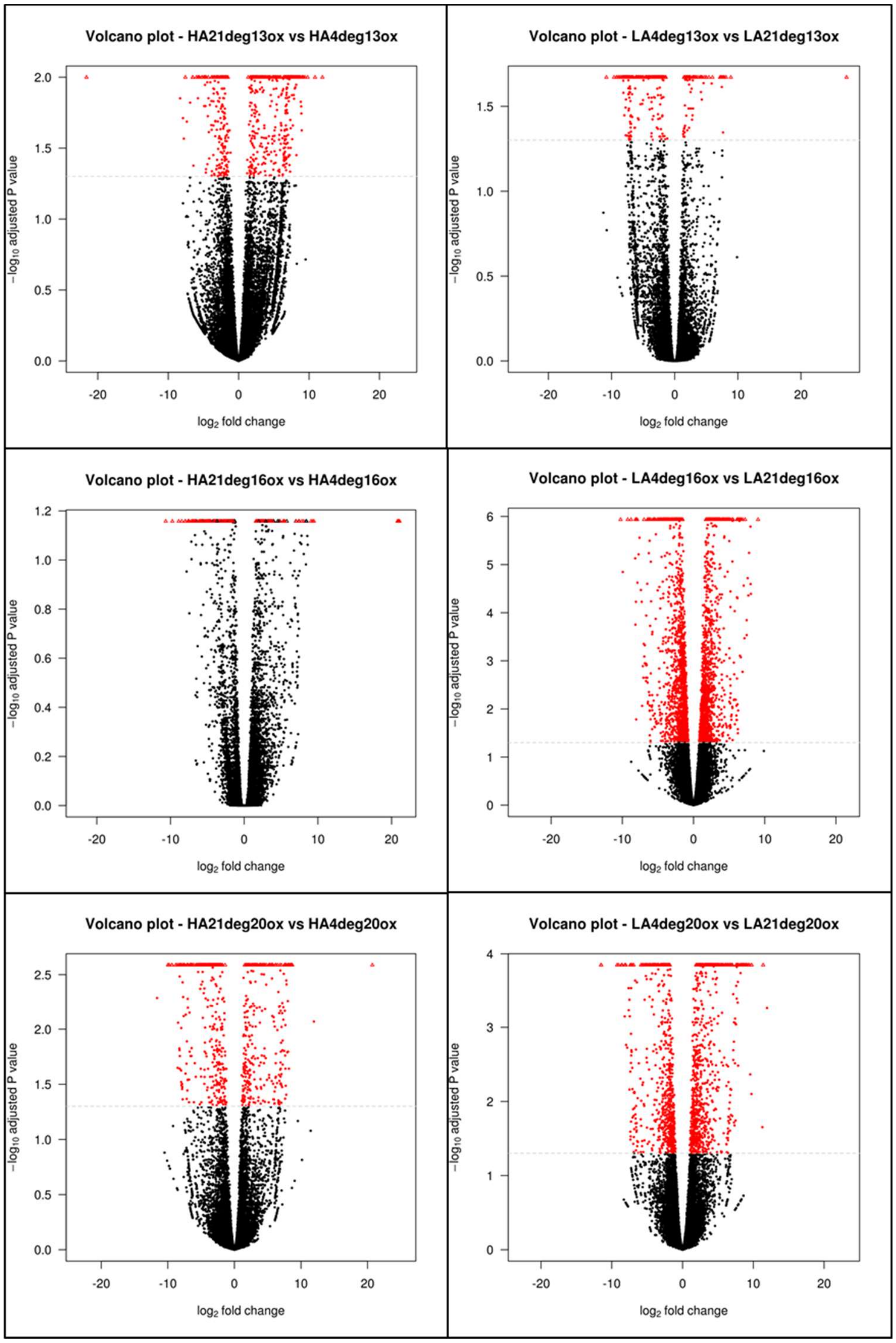


Figure 69: Differential gene expression of Low altitude (LA) and High Altitude (HA) populations maintained at different temperatures under varying oxygen levels. Data is presented using volcano plots of differentially regulated genes (P value <0.05) for comparisons of temperature at three Oxygen levels (rows: 13% 16%, 20%) for the two populations of HA (left column) and LA (right column)

The differentially expressed genes lists were filtered for genes with a fold change less than Log -1.4 and greater than Log 1.4. At 13% Oxygen there were more differentially expressed genes passing the filter in the HA population than the LA population (Table 21), however at 16% and 20% there as a much higher number of genes passing filter in the LA population than the HA population. To identify if any of these differentially expressed genes were expressed in multiple experimental comparisons, a Venn diagram was generated (Figure 70). There were few common genes across both HA and LA populations in the experimental comparisons although there were a few common genes within populations.

Table 21: Differential gene count table of temperature effect in HA and LA populations.

Differentially expressed genes after filter		*HA population*	*LA population*
(Padj <0.05) (Fold change Log <-1.4 & >1.4)		21°C	21°C
13% Oxygen	4°C	631	347
16% Oxygen	4°C	172	2031
20% Oxygen	4°C	639	1329

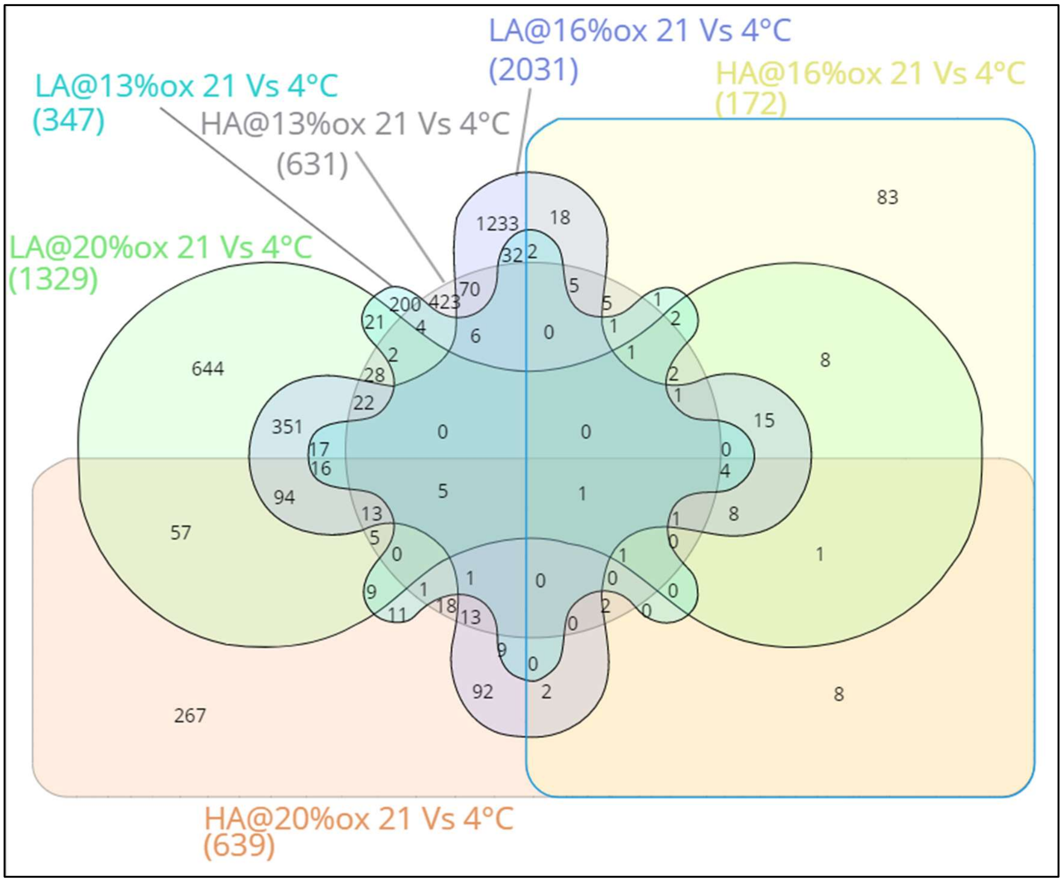


Figure 70: Conservation of transcript response between condition. A 6 way Venn diagram of temperature effect comparing common differential expressed genes between experimental comparisons. Numbers indicate the number of differentially expressed genes between condition comparisons.

The differentially expressed gene lists and fold changes were tested for enrichment in gene ontology. In the comparison of the HA population at 13% Oxygen between 4°C and 21°C, Response to stimulus was the only large cluster of enriched GO terms (Figure 71). By comparison, the LA population was enriched with GO term clusters in Response to chemical/stimulus and Multicellular organismal process (Figure 72).

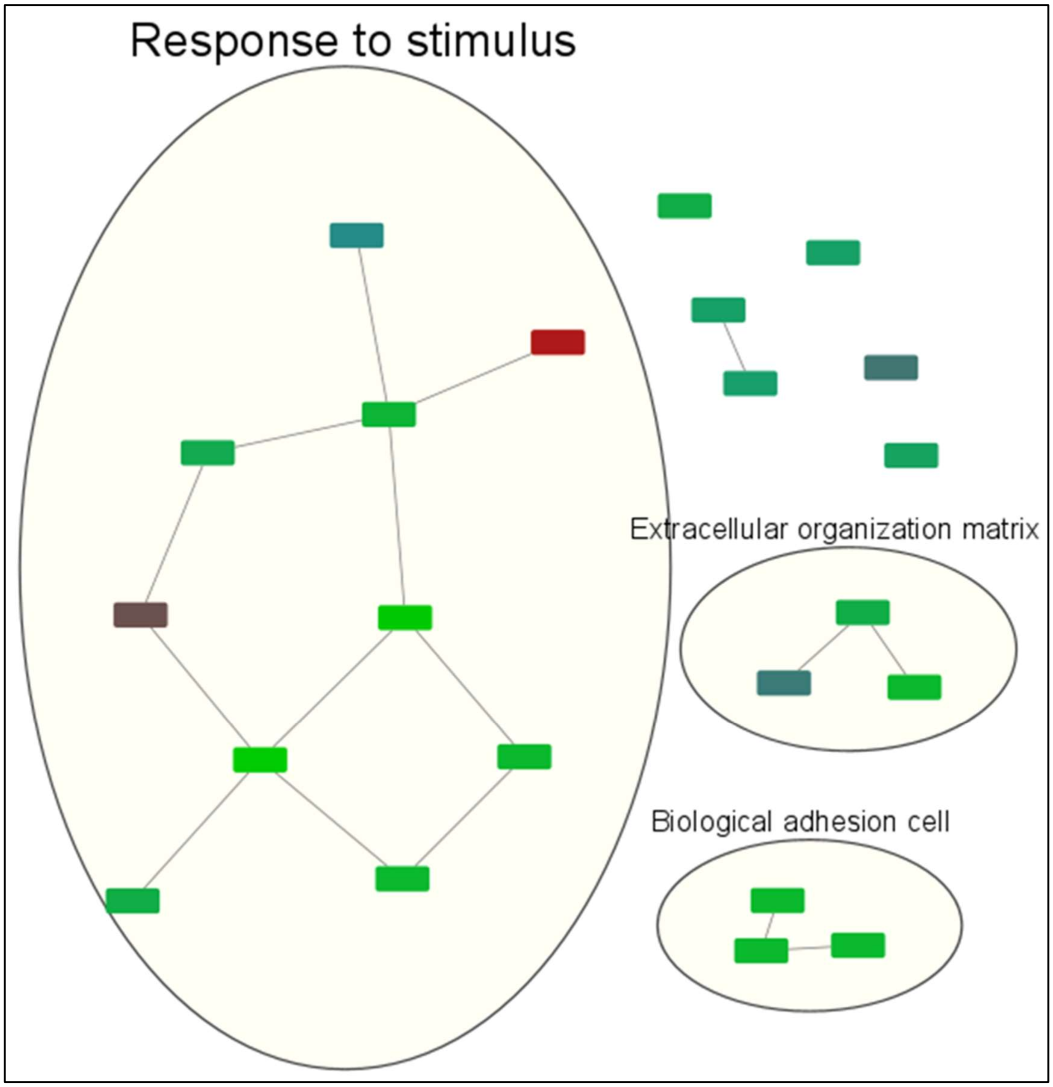


Figure 71: Comparison of experimental HA individuals exposed in 13% Oxygen at 4°C Vs 21°C. Gene ontology network identifying "Response to stimulus" GO terms circled in red. GO terms are coloured red to green indicating a low FDR adjusted P-value to high (cut off <0.05).

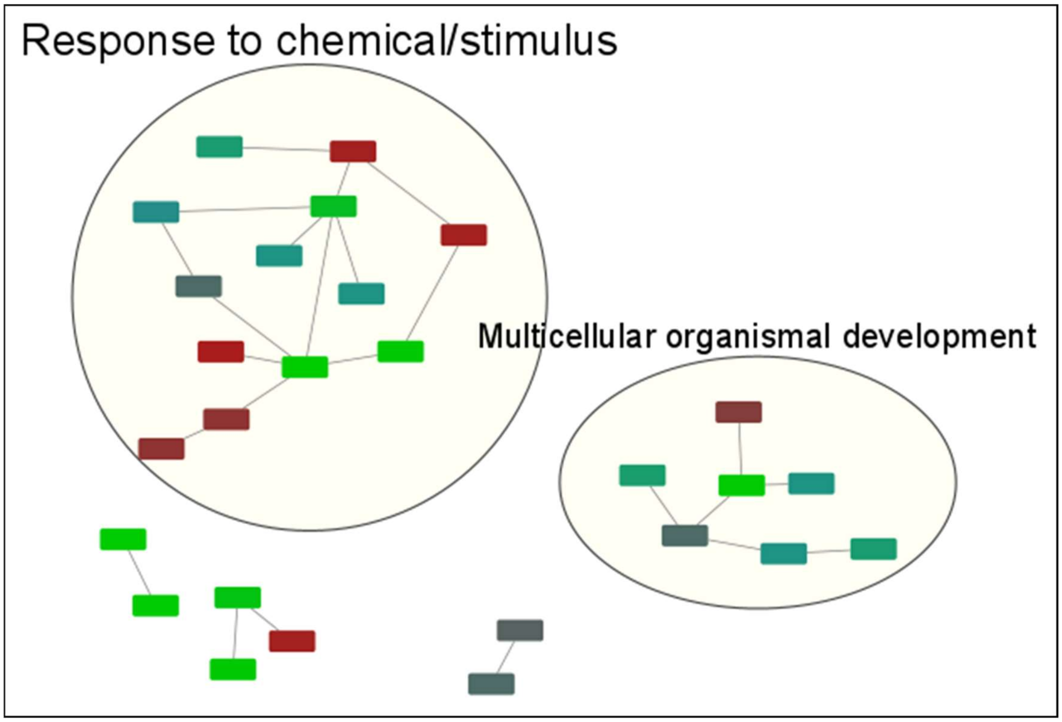

Response to chemical/stimulus
Multicellular organismal development

Figure 72: Comparison of experimental LA individuals exposed in 13% Oxygen at 4°C Vs 21°C. Gene ontology network identifying Multicellular organismal process and Response to chemical. GO terms are coloured red to green indicating a low FDR adjusted P-value to high (cut off <0.05).

When the HA population at 16% Oxygen was compared between 4°C and 21°C only the GO term for positive regulation of multicellular organismal process was enriched (Figure 73). The LA population yielded no enriched GO terms despite having over 10 times the number of differentially expressed genes than the HA population.

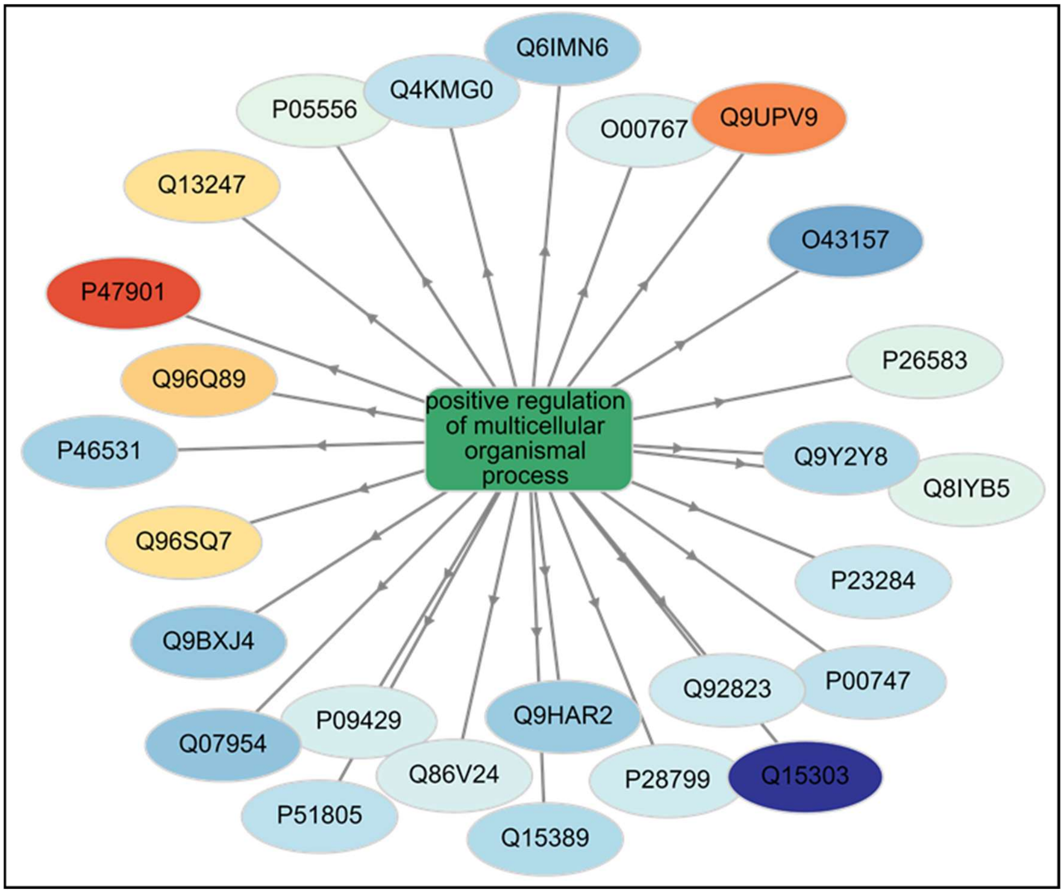


Figure 73: Comparison of experimental HA individuals exposed in 16% Oxygen at 4°C Vs 21°C. Gene ontology network identifying Positive regulation of multicellular organismal process. Gene fold change is indicated (darker blue is more downregulated and darker red is more upregulated).

For the HA population at 20% Oxygen comparison between 4°C and 21°C, 7 clusters of enriched GO term groups were identified (Figure 74). These included “Positive regulation of protein phosphorylation”, “Regulation of response to stimulus”, “Response to chemical, Immune system process”, “Regulation of multicellular organismal development”, “Developmental process” and “Oxoacid metabolic process”. In particular, there was a positive regulation of vasoconstriction when comparing 21°C against 4°C and activation of the MAPK ERK and PI3K pathway. In the LA population in the same conditions only Metabolic process and organonitrogen compound metabolic process GO terms were enriched, though there were substantial numbers of genes up and downregulated that were associated with these GO terms (Figure 75).

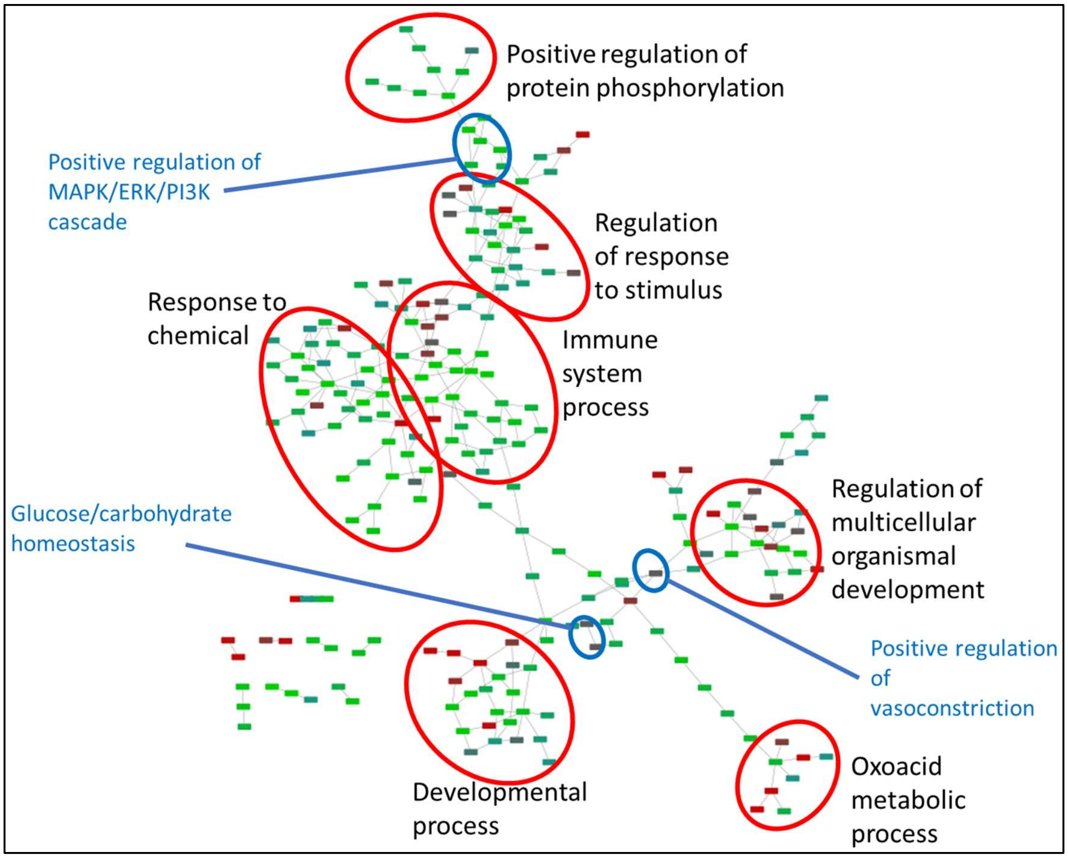

Positive regulation of protein phosphorylation
Positive regulation of MAPK/ERK/PI3K cascade
Regulation of response to stimulus
Response to chemical
Immune system process
Regulation of multicellular organismal development
Glucose/carbohydrate homeostasis
Positive regulation of vasoconstriction
Developmental process
Oxoacid metabolic process

Figure 74: Comparison of experimental HA individuals exposed in 20% Oxygen at 4°C Vs 21°C. Gene ontology network identifying "Positive regulation of protein phosphorylation", "Regulation of response to stimulus", "Response to chemical", "Immune system process", "Regulation of multicellular organismal development", "Developmental process" and "Oxoacid metabolic process". GO terms are coloured red to green indicating a low FDR adjusted P-value to high (cut off <0.05).

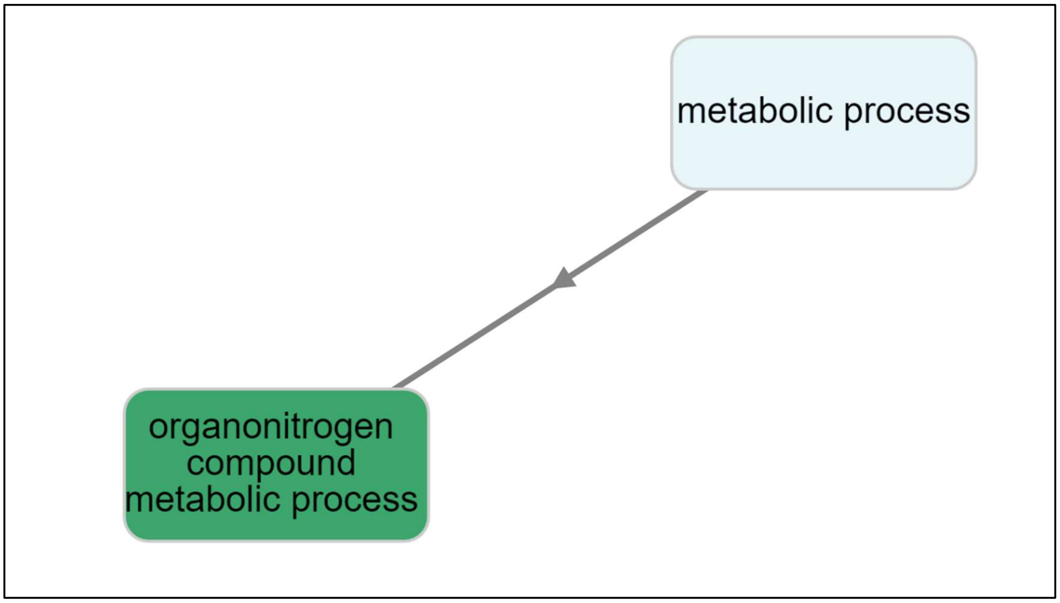


Figure 75: Comparison of experimental LA individuals exposed in 20% Oxygen at 4°C Vs 21°C. Gene ontology network identifying "Metabolic process" and "Organonitrogen compound metabolic process". Associated Genes are not shown due to number interacting.

6.3.8. **Differential statistics – Role of Oxygen.**

Differential statistics were performed with DESeq and EdgeR to identify differentially expressed genes in the HA and LA populations between exposure to 3 Oxygen levels 13, 16 and 20%. Statistics were calculated independently at 4°C and 21°C (Figures 76, 77, 78 & 79).

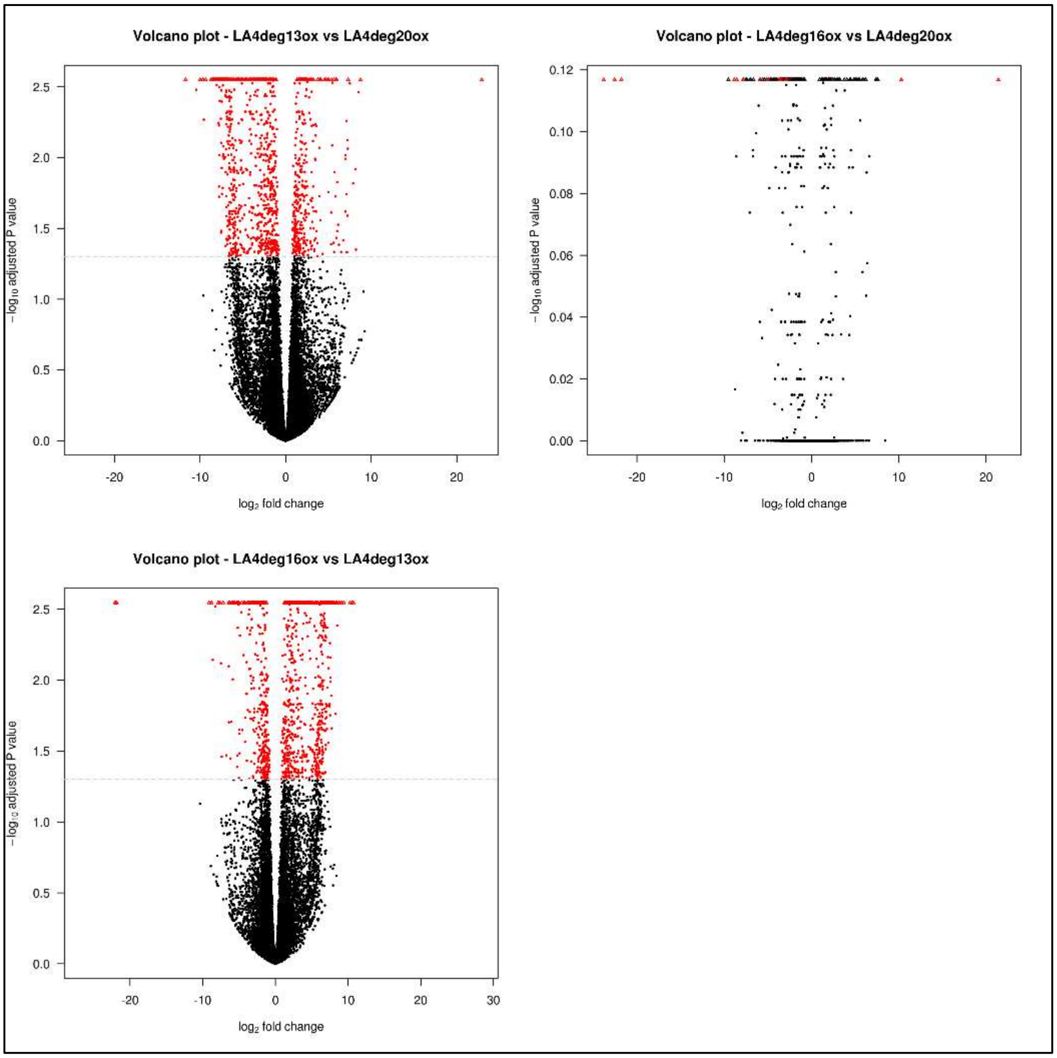


Figure 76: Differential gene expression of Low altitude (LA) populations maintained at 4°C under different Oxygen levels. Data is presented using volcano plots of differentially regulated genes (P value <0.05) for comparisons of Oxygen levels (rows: 13% 16%, 20%).

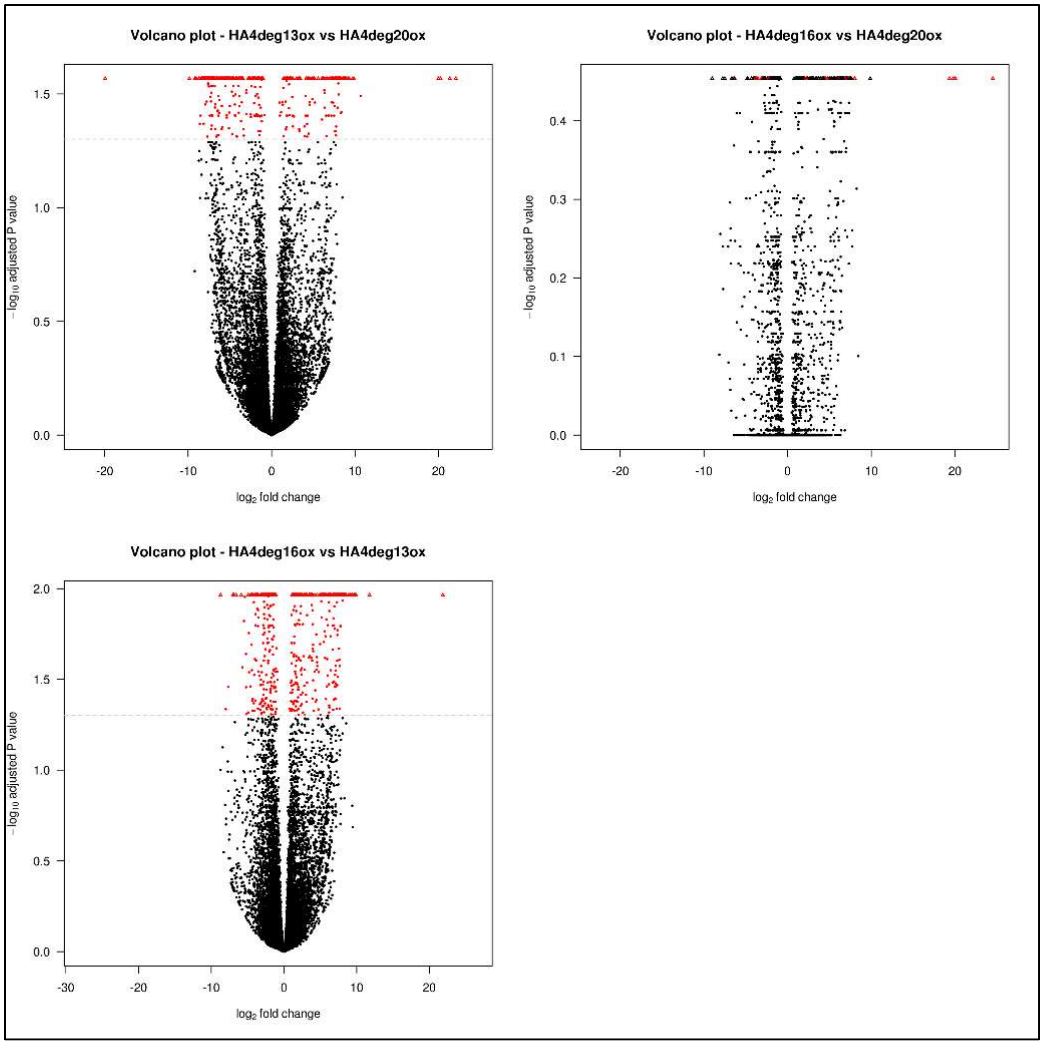


Figure 77: Differential gene expression of High altitude (HA) populations maintained at 4°C under different Oxygen levels. Data is presented using volcano plots of differentially regulated genes (P value <0.05) for comparisons of Oxygen levels (rows: 13% 16%, 20%).

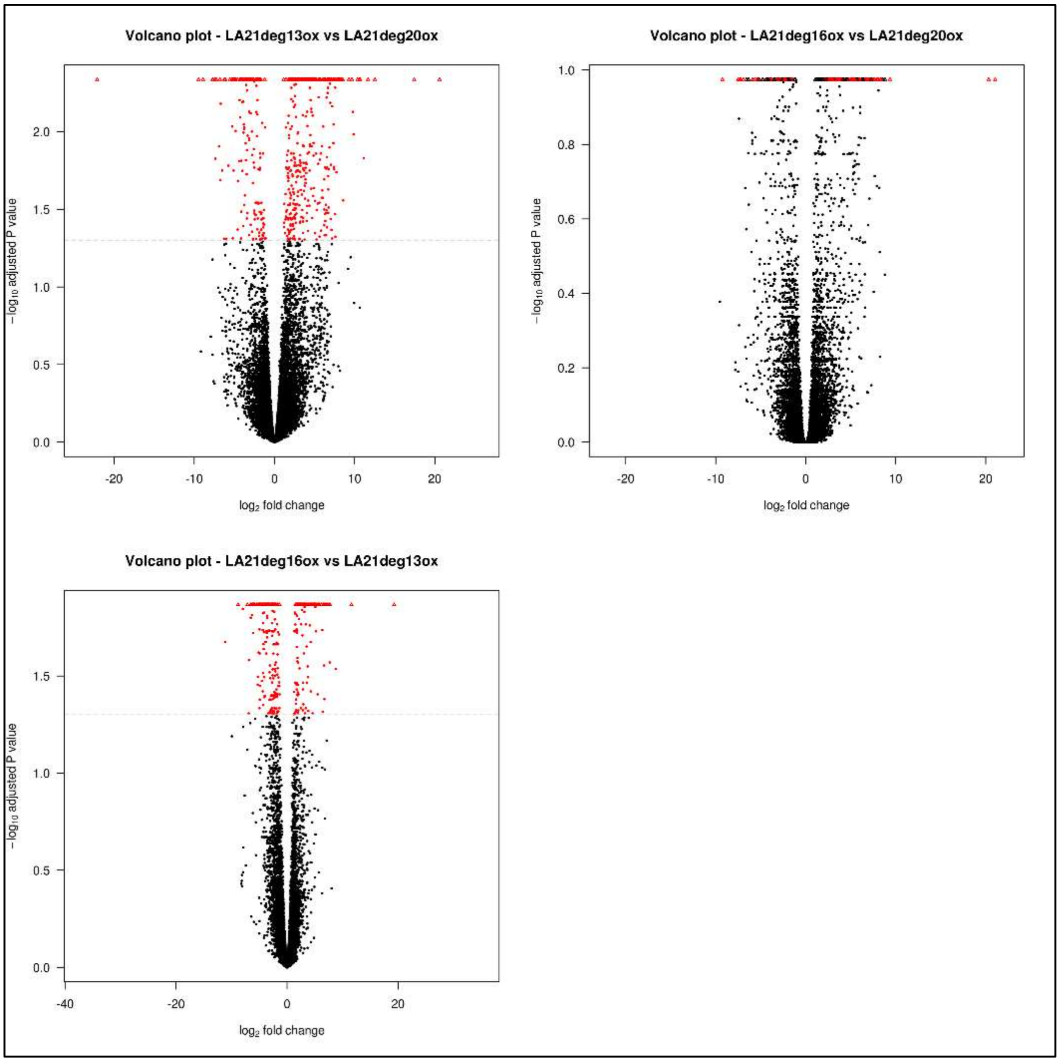


Figure 78: Differential gene expression of Low altitude (LA) populations maintained at 21°C under different Oxygen levels. Data is presented using volcano plots of differentially regulated genes (P value <0.05) for comparisons of Oxygen levels (rows: 13% 16%, 20%).

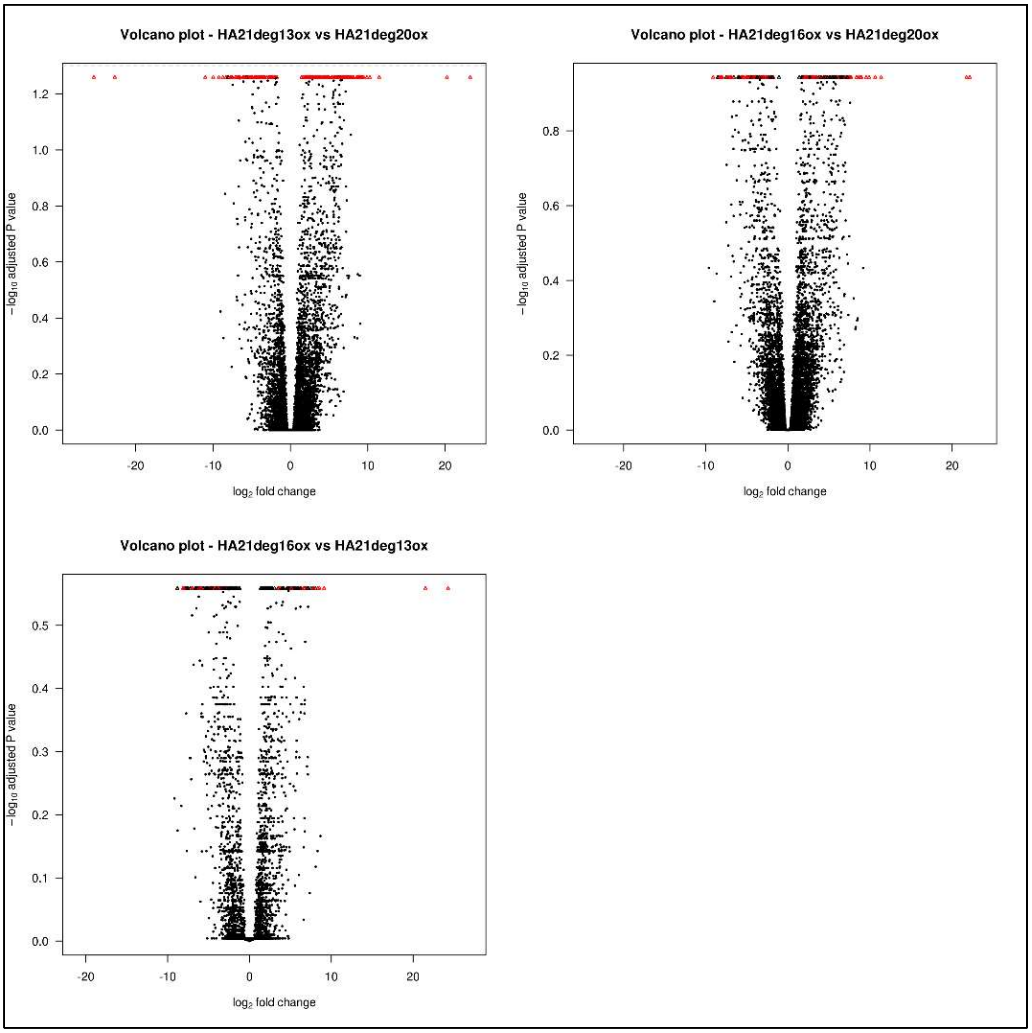


Figure 79: Differential gene expression of High altitude (HA) populations maintained at 21°C under different Oxygen levels/. Data is presented using a volcano plots of differentially regulated genes (P value <0.05) for comparisons of Oxygen levels (rows: 13% 16%, 20%).

The differentially expressed genes lists were filtered for genes with a fold change less than Log -1.4 and greater than Log 1.4. The LA population had a greater number of differentially expressed genes in all exposures when looking for changes in gene regulation to Oxygen concentration (Table 22). In some cases, there were more than double the number of differentially expressed genes. To identify if any of the differentially expressed genes in the 4°C experimental conditions were expressed in multiple experimental comparisons, a Venn diagram was generated (Figure 80). There were no common genes across both HA and LA populations in the experimental comparisons although there were a few common genes within populations.

Table 22: Common genes differentially expressed across each test

Differentially expressed genes after filter (Padj <0.05)	*HA pop*	*LA pop*	*HA pop*	*LA pop*	*HA pop*	*LA pop*
	13% Vs 20% Oxygen	13% Vs 20% Oxygen	13% Vs 16% Oxygen	13% Vs 16% Oxygen	16% Vs 20% Oxygen	16% Vs 20% Oxygen
4°C	374	1171	518	1017	45	45
21°C	225	587	54	398	122	138

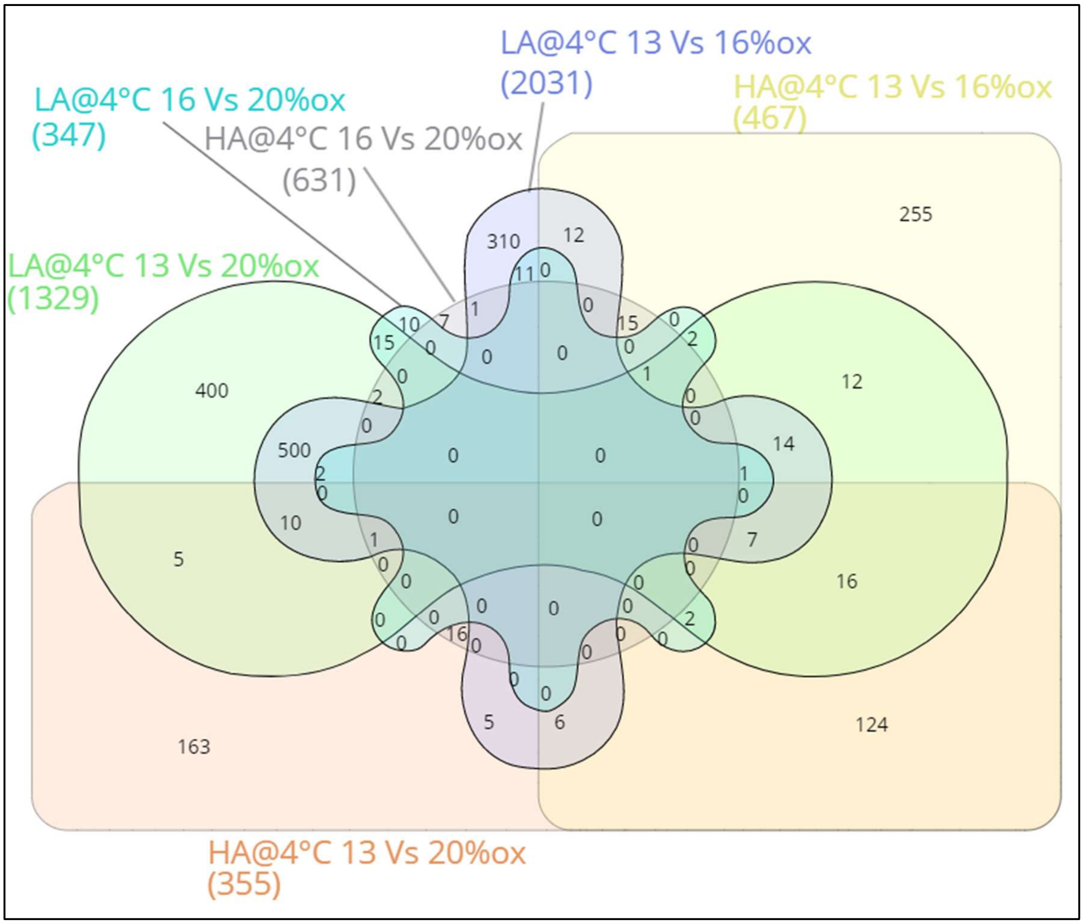


Figure 80: Conservation of transcript response between condition. A 6 way Venn diagram of Oxygen effect comparing common differentially expressed genes between experimental comparisons for LA and HA populations at 4°C. Numbers indicate the number of differentially expressed genes between condition comparisons.

The effect of Oxygen was investigated with differential statistics between Oxygen concentrations 13%, 16% and 20% at both 4°C and 21°C in both HA and LA population separately to try and isolate the variable. In the HA population at 4°C, the comparison between 13% and 20% Oxygen Multicellular development process and Biological adhesion were enriched in gene ontology while in the comparison between 13% and 16% Oxygen Response to stimulus was enriched. In the comparison of 16% and 20% the fewest GO terms were enriched with only Extracellular matrix organization enriched for with genes (Figure 81, Figure 82 and Figure 83).

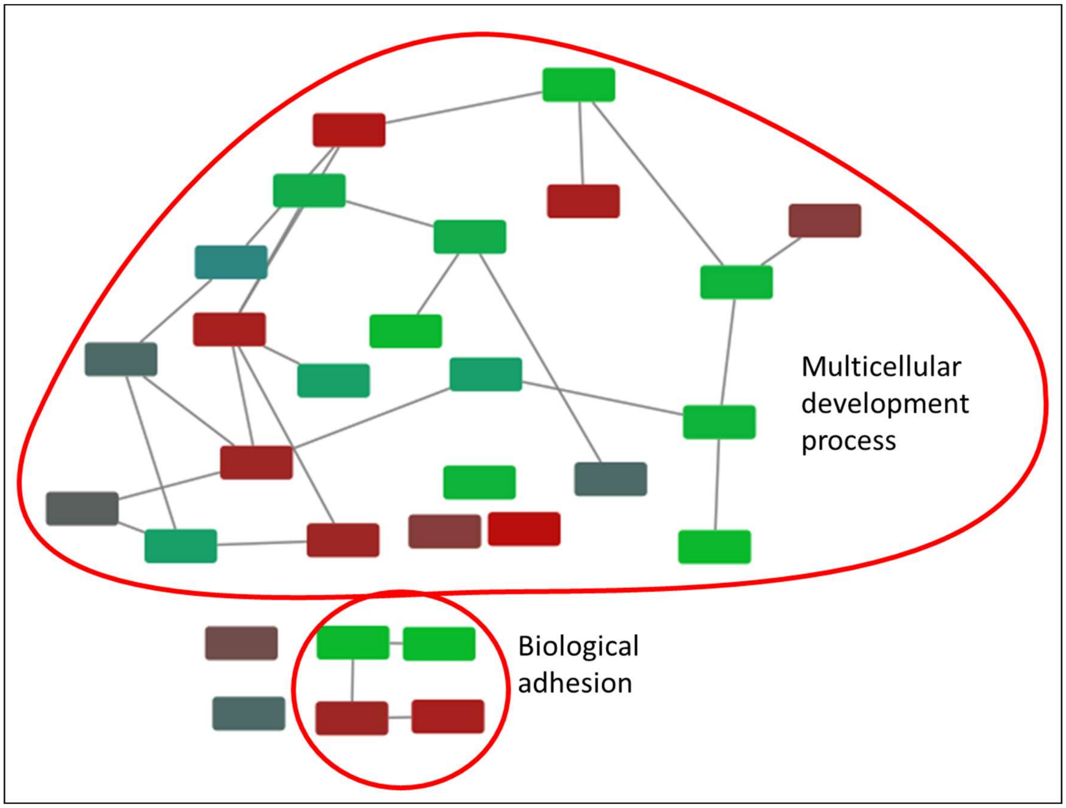


Figure 81: Pathway enrichment analysis of HA individuals exposed in 4°C at 13% Vs 20% Oxygen. Gene ontology network identifying "Multicellular development process" and "Biological adhesion". GO terms are coloured red to greenindicating a low FDR adjusted P-value to high (cut off <0.05).

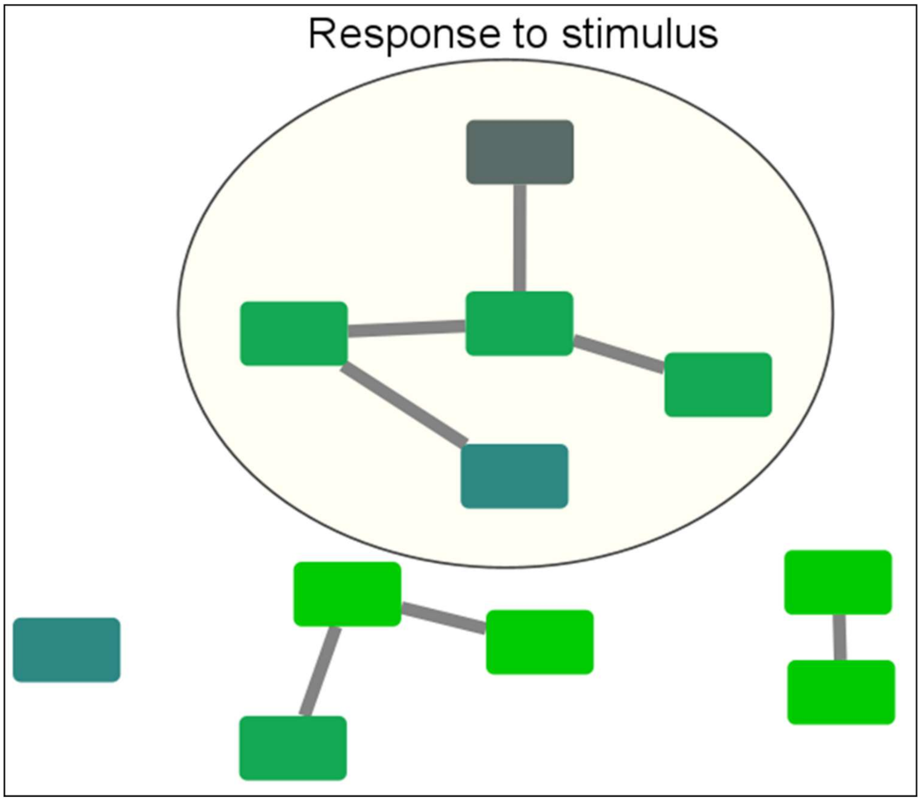


Figure 82: Pathway enrichment analysis of experimental HA individuals exposed in 4°C at 13% Vs 16% Oxygen. Gene ontology network identifying Response to stimulus. GO terms are coloured yellow to purple indicating a low FDR adjusted P-value to high (cut off <0.05).

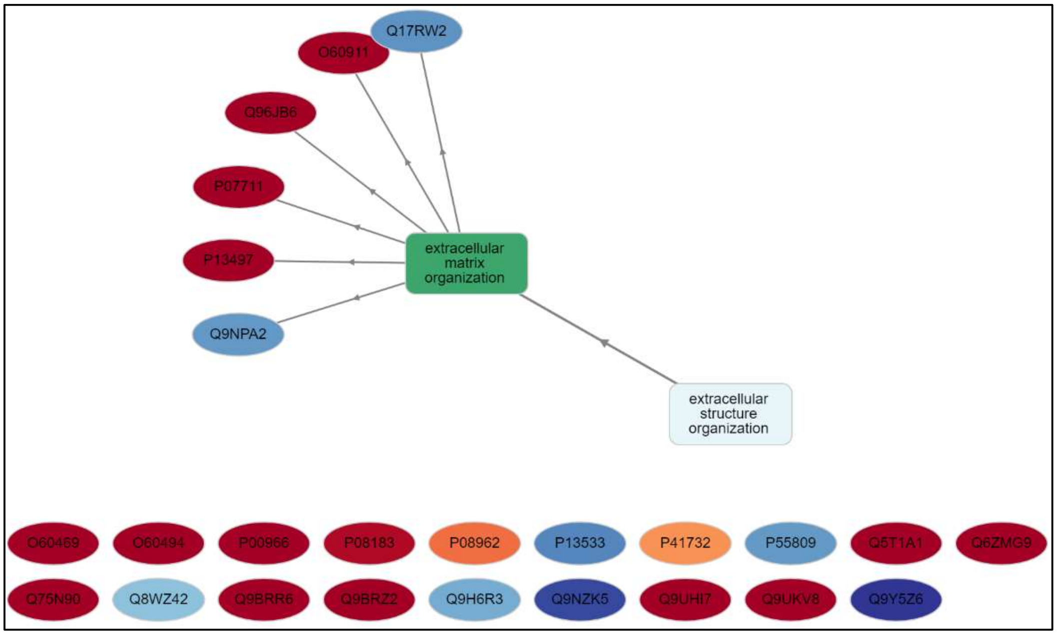


Figure 83: Pathway enrichment analysis of experimental HA individuals exposed in 4°C at 16% Vs 20% Oxygen. Gene ontology network identifying "Extracellular matrix organisation". Gene fold change is indicated (darker blue is more downregulated and darker red is more upregulated).

When the same differential expression analyses were performed on the LA population at 4°C, in the 13% comparison with 20% Oxygen, "Response to stimulus" and "Signalling" was enriched (Figure 84). In the 13% comparison with 16% Oxygen; Transport, Cell morphogenesis, System process, Multicellular organismal process, Regulation of cardiac contraction, Signal transduction and Response to chemical/stimulus GO terms were enriched. Also, GO terms for "Immune system process", "Regulation of interleukin-l production" and "Negative regulation of leukocyte apoptotic process" were identified (Figure 85). There were no enriched GO terms in the comparison of 16% with 20% Oxygen.

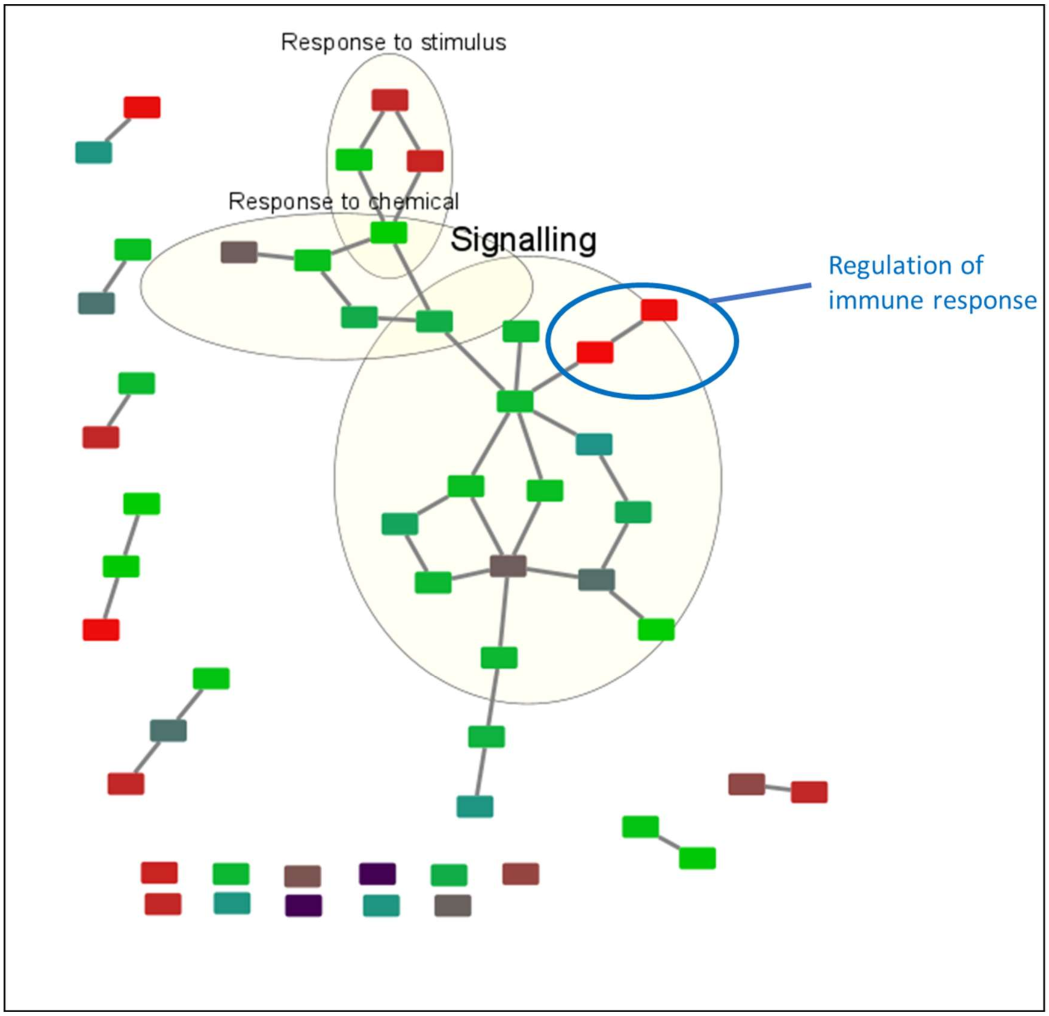


Figure 84: Pathway enrichment analysis of experimental LA individuals exposed in 4°C at 13% Vs 20% Oxygen. Gene ontology network identifying "Response to stimulus" and "Signalling". This included a "Regulation of immune response" (circled in blue) GO terms are coloured yellow to purple indicating a low FDR adjusted P-value to high (cut off <0.05).

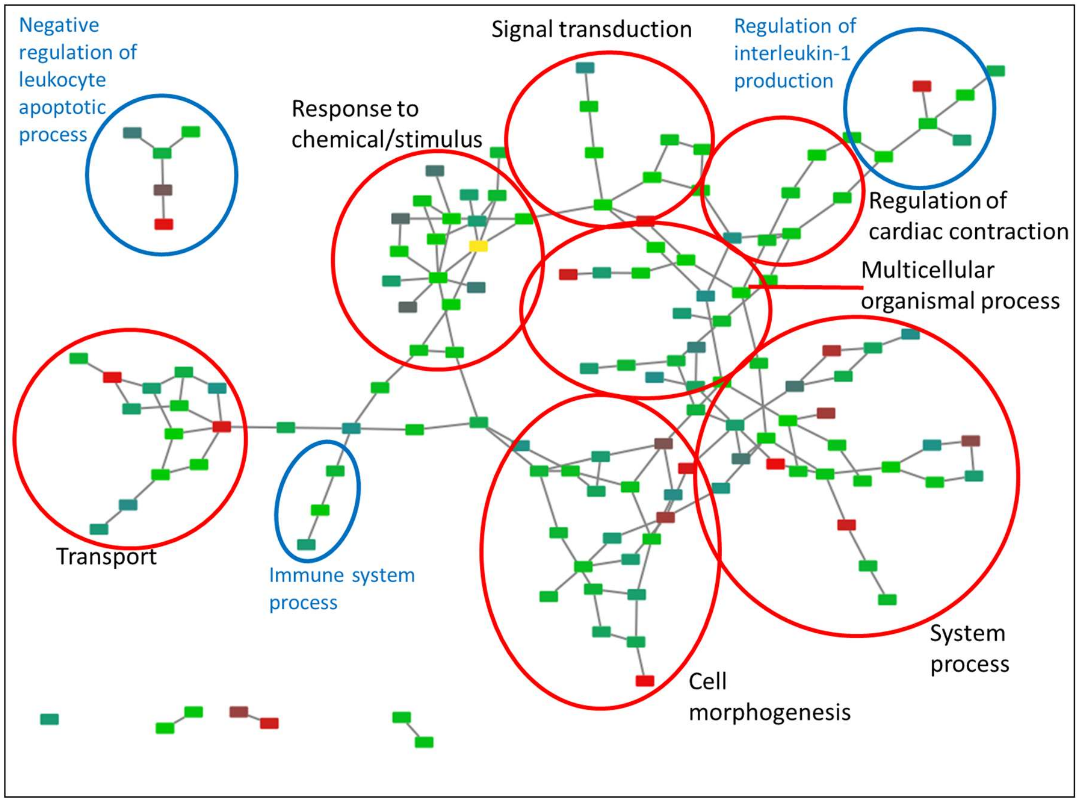

Negative regulation of leukocyte apoptotic process
Signal transduction
Regulation of interleukin-1 production
Response to chemical/stimulus
Regulation of cardiac contraction
Multicellular organismal process
Transport
Immune system process
System process
Cell morphogenesis

Figure 85: Pathway enrichment analysis of experimental LA individuals exposed in 4°C at 13% Vs 16% Oxygen. Gene ontology network identifying "Transport", "Cell morphogenesis", "System process", "Multicellular organismal process", "Regulation of cardiac contraction", "Signal transduction" and "Response to chemical/stimulus". "Immune system process", "Regulation of interleukin-l production" and "Negative regulation of leukocyte apoptotic process" were also identified (shown cirlced in blue) GO terms are coloured yellow to purple indicating a low FDR adjusted P-value to high (cut off <0.05).

To identify if any of the differentially expressed genes in the 21°C experimental conditions were expressed in multiple experimental comparisons, a Venn diagram was generated (Figure 86). Far fewer genes were differentially expressed at 21°C than at 4°C and there was very limited overlapping of genes overlapping between experimental comparisons.

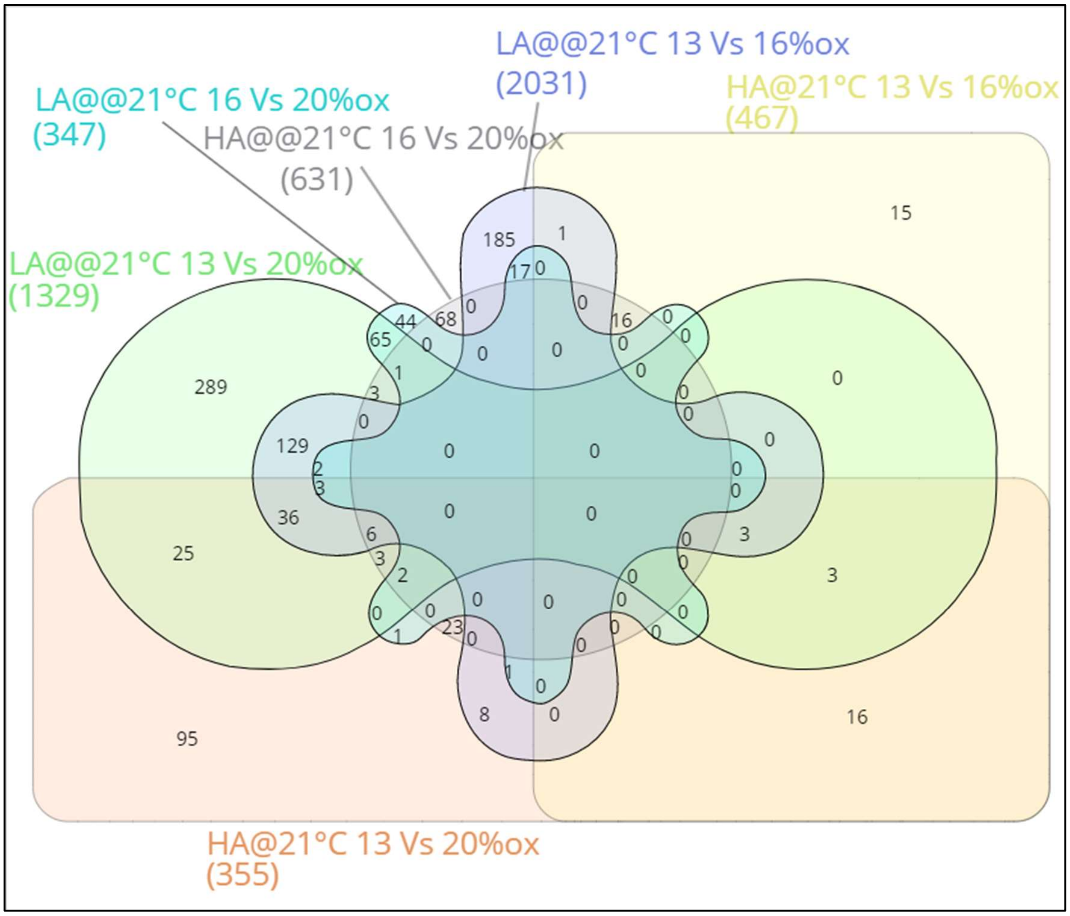


Figure 86: Conservation of transcript response between condition. A 6 way Venn diagram of Oxygen effect comparing common differential expressed genes between experimental comparisons for LA and HA populations at 21°C. Numbers indicate the number of differentially expressed genes between condition comparisons.

As a result of the lower numbers of differentially expressed genes, only the comparison of HA individuals exposed at 21°C in 13% compared against 20% Oxygen yielded enriched GO terms. These included Cell morphogenesis, System process, Multicellular organismal process, Immune system process and extracellular structure organization (Figure 87). Similarly, the LA individuals exposed at 21°C in 13% Oxygen compared against 20% Oxygen were enriched for GO terms of Cell morphogenesis and Multicellular organismal process but were also enriched for Developmental process (Figure 88). For the comparison of LA individuals exposed at 21°C in 13% compared against 16% Oxygen, the enriched GO terms for Multicellular organismal process, Response to stimulus and Regulation of system process were identified (Figure 89).

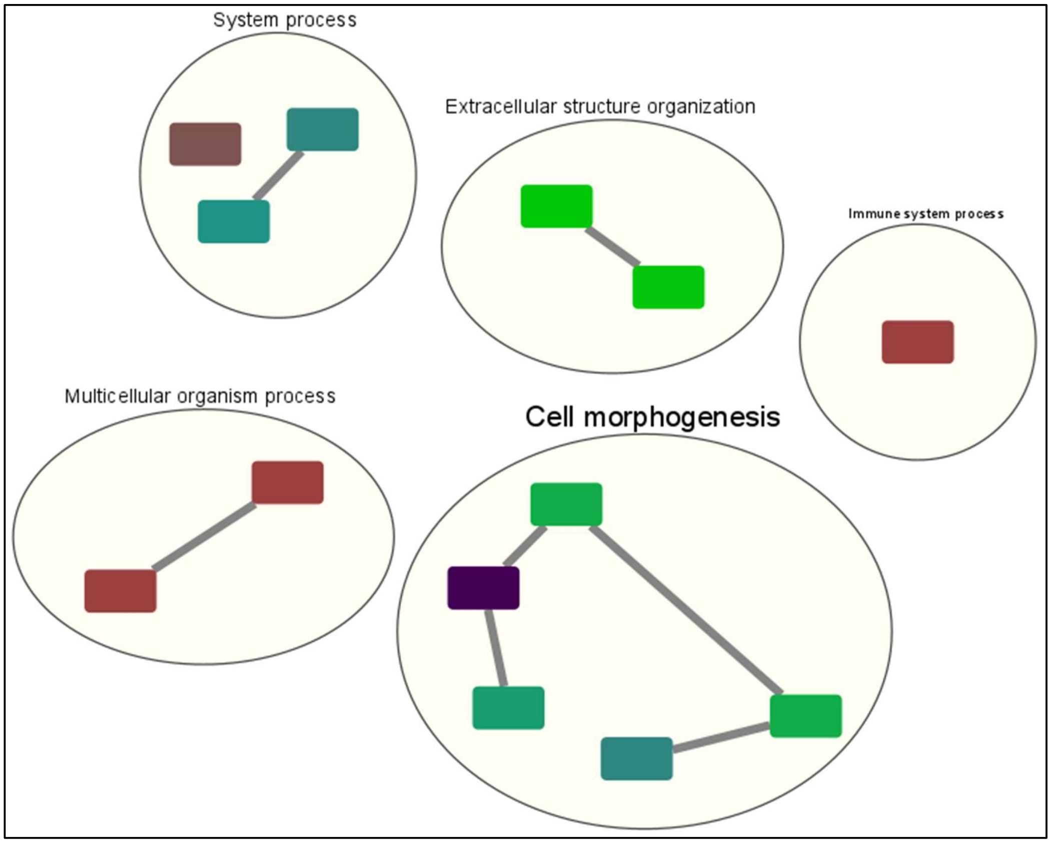

System process
Extracellular structure organization
Immune system process
Multicellular organism process
Cell morphogenesis

Figure 87: Pathway enrichment analysis of experimental HA individuals exposed to 21°C at 13% Vs 20% Oxygen. Gene ontology network identifying Cell morphogenesis, System process, Multicellular organismal process, Immune system process and extracellular structure organization. GO terms are coloured red to green indicating a low FDR adjusted P-value to high (cut off <0.05).

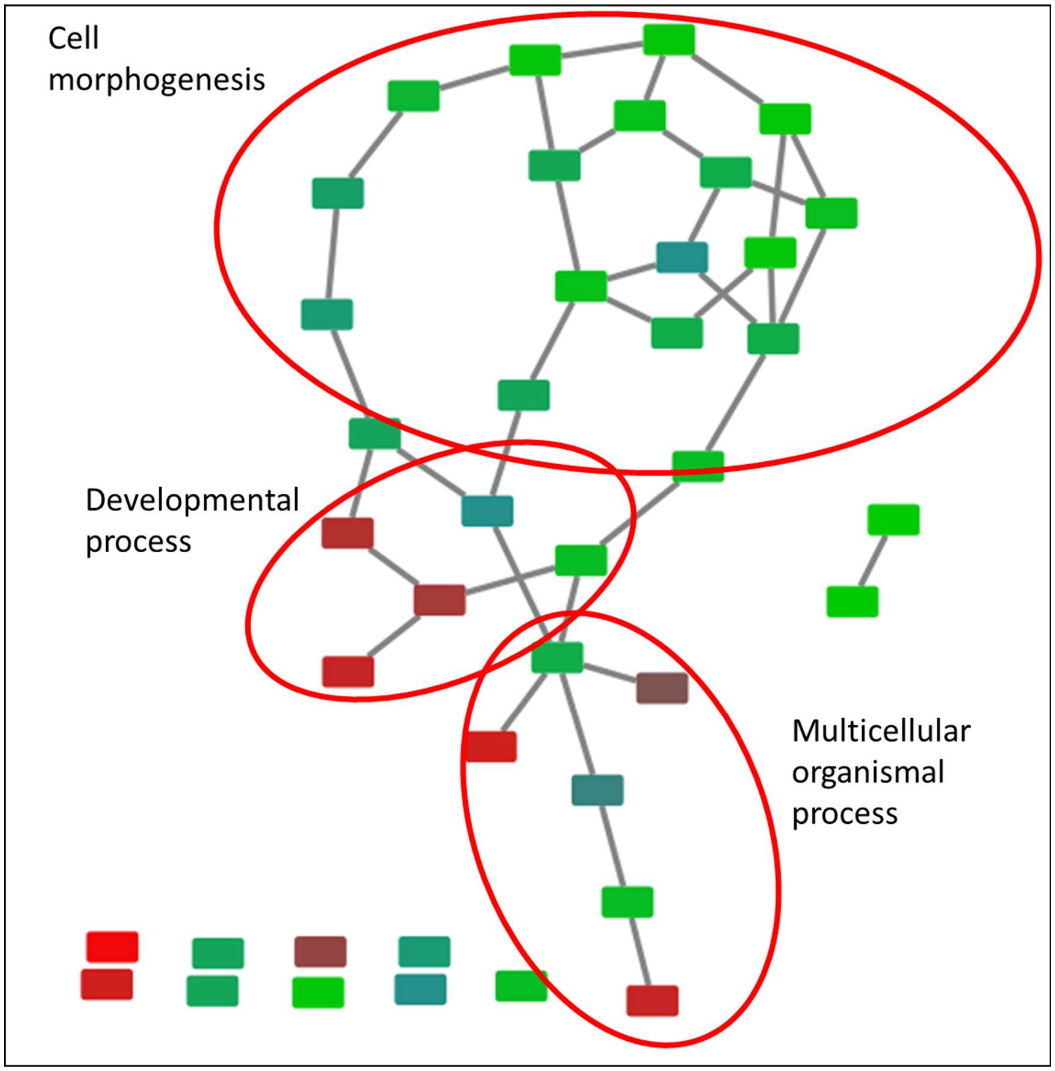


Figure 88: Pathway enrichment analysis of experimental LA individuals exposed to 21°C at 13% Vs 20% Oxygen. Gene ontology network identifying Cell morphogenesis, Multicellular organismal process and Developmental process. GO terms are coloured red to green indicating a low FDR adjusted P-value to high (cut off <0.05).

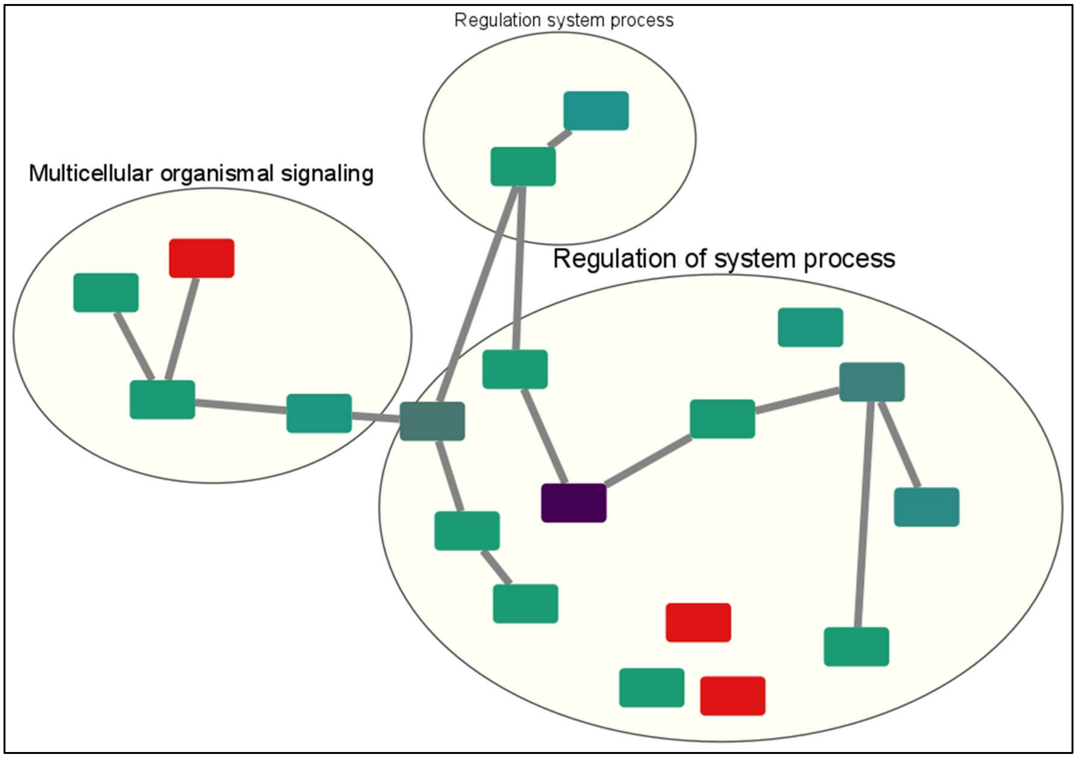

Regulation system process
Multicellular organismal signaling
Regulation of system process

Figure 89: Pathway enrichment analysis of experimental LA individuals exposed to 21°C at 13% Vs 16% Oxygen. Gene ontology network identifying Multicellular organismal process, Response to stimulus and Regulation of system process. GO terms are coloured red to green indicating a low FDR adjusted P-value to high (cut off <0.05).

6.3.9. **Identification of proportion of shared genes and key genes.**

DiVenn was used to identify the proportion of shared genes between experimental comparisons. In the DiVenn's identification of common genes between HA and LA populations' differential gene expression between 4°C and 21°C at Oxygen concentrations of 13%, 16% and 20%, the largest group of shared differentially expressed genes are between the LA temperature comparison at 16% and 20% Oxygen. Most of these shared genes follow the same pattern of up and down regulation (Figure 90). The HA comparison of temperature at 13% Oxygen is observed to have largely upregulated (red) genes, while in contrast but to a lesser degree, the LA comparison of temperature at 13% Oxygen sees a larger cohort of downregulated genes (blue). There is a small cohort of downregulated genes shared between the temperature comparisons of LA populations at 13% and 16% Oxygen. These genes are involved in a range of biological process' including metabolism and cellular processing (PCLO, MAB21L2, GDA, LTBP3, MAGEE1, C1GALT1, TTC29, PTPRA, MYH1, MACROD2, GDF11 and UBA1). It is worth noting that there are fewer shared differentially expressed genes between HA comparisons than between LA comparisons, which could indicate a plasticity of response in the HA population, but some of the shared upregulated genes include those involved in respiration, gluconeogenesis, glycogenolysis and cellular growth through the TGFβ signalling pathway activation (ANK2, COX6B1, FOXO4, HAO1, GFI1B, ELAVL4 and FBXO36). Many of the conflicting (yellow) shared genes are between the HA population comparisons and the LA population comparisons indicating the populations have fundamentally different approaches to coping with lower temperatures. There are three genes universally upregulated in all three LA population comparisons that are universally downregulated in all three HA population comparisons (HMGB1, HMGB2 and PPIB). There is also one universally downregulated gene in the LA population comparisons and upregulated in the HA population comparisons (CD109). In addition to these four genes, where the differentially expressed gene is only seen in 2/3 of the LA population and 2/3 of the HA population differential expression comparisons, there are 11 genes where there is upregulation in the LA population and downregulation in the HA population (SLC5A8, MUC1, CTSL, PARP3, TARDBP, SMAP1, CDON, SLCA6, CBS and AEBP2) and 12 genes where there is downregulation in the LA population and upregulation in the HA population (CORO2A, CD63, ACTB, MOGAT1, TLL1, COL12A1, LCT, MCMBP, ZNF235, SRSF6, ARRDC3 and DBH).

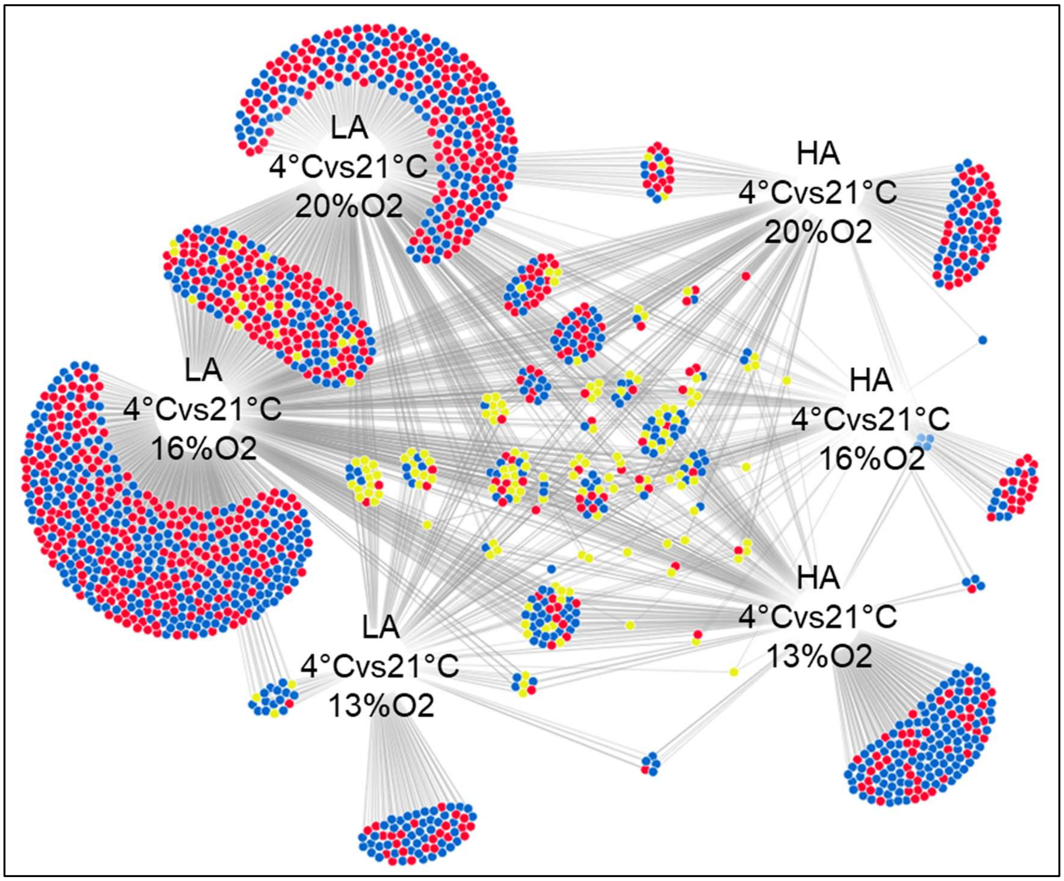


Figure 90: Identification of shared gene differential expressed genes between HA and LA populations' maintained at 4°C and 21°C with Oxygen concentrations of 13%, 16% and 20%. DiVenn has been used to illustrate the direction of gene expression with upregulated genes are shown in red and downregulated genes in blue. Yellow indicates conflicting up and downregulation of shared genes. Grey lines link between gene expresion and exposure condition and population.

In the DiVenn's identification of common genes between HA and LA populations' differential gene expression between Oxygen concentrations of 13%, 16% and 20% at 4°C and at 21°C, the largest group of shared differentially expressed genes are between the LA population comparison of 13%Vs20% and 13%Vs16% Oxygen (Figure 91 and Figure 92). This is seen at both 4°C and 21°C, and like the large common group seen between the LA population comparison of 4°C Vs 21°C at 16% and 20% Oxygen, there are many genes with the same expression pattern (largely downregulated). In both 4°C and 21°C, there are very few shared differentially expressed genes between the LA population and the HA population. In both 4°C and 21°C there are very few differentially expressed genes, with fewer still shared with other differential gene comparisons. There are 36 shared differentially expressed genes found in the LA population comparisons of 13%Vs20% and 13%Vs16% at 4°C and at 21°C and 4 differentially expressed genes found in the HA population of the same comparisons. These genes also show the same pattern of expression for the comparison of Oxygen.

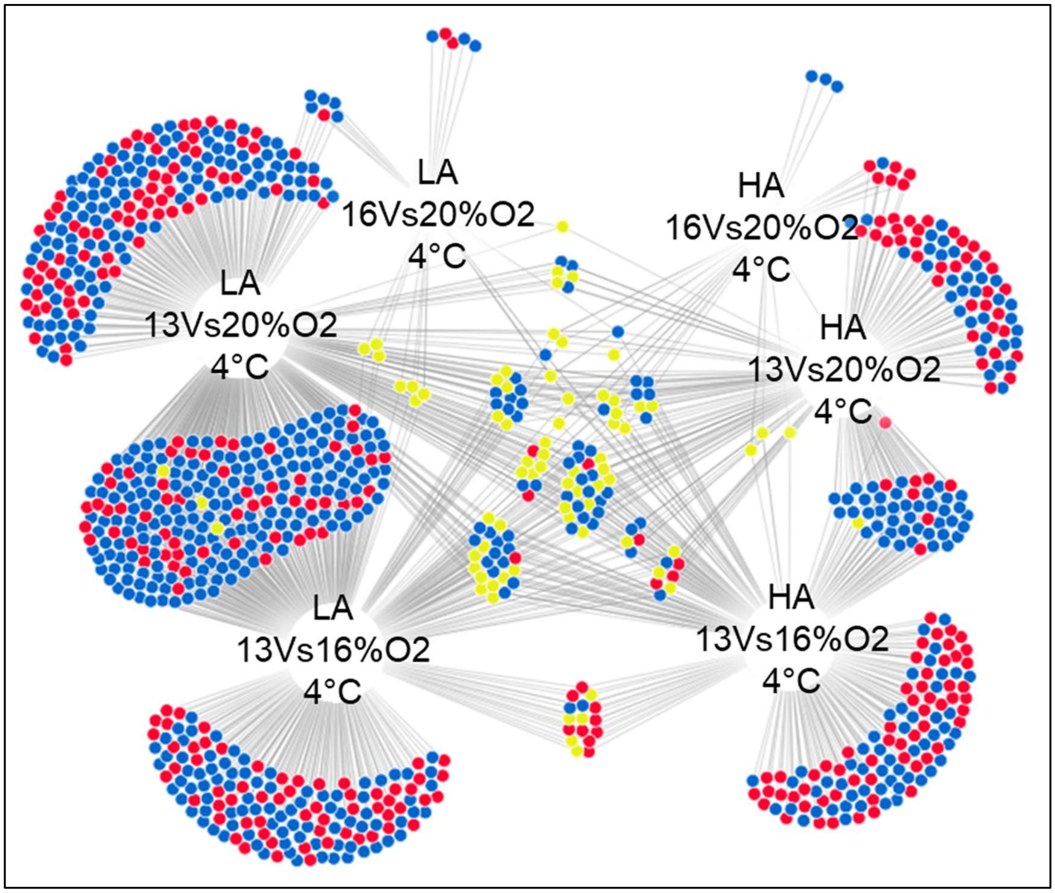


Figure 91: Identification of shared gene differential expressed genes between HA and LA populations' maintained at Oxygen concentrations of 13%, 16% and 20% with a temperature of 4°C. DiVenn has been used to illustrate the direction of gene expression with upregulated genes are shown in red and downregulated genes in blue. Yellow indicates conflicting up and downregulation of shared genes. Grey lines link between gene expresion and exposure condition and population.

Of the LA population shared downregulated genes, the majority are linked with “Vesicle-mediated transport”, “Response to stress”, “mRNA processing” and “Anatomical structure development”. With the HA population shared downregulated genes, the majority are linked with “Cell motility”, “Response to stress”, “Anatomical structure development” and “mRNA processing”.

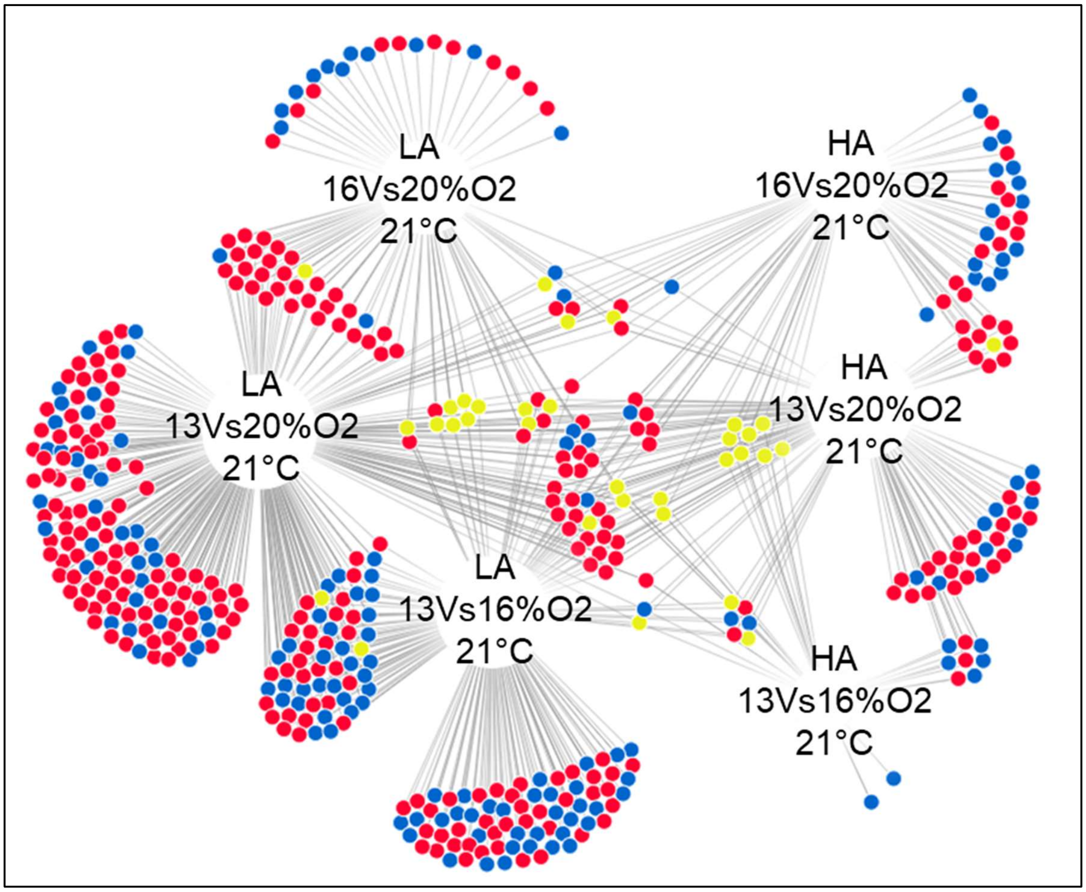


Figure 92: Identification of shared gene differential expressed genes between HA and LA populations’ maintained at Oxygen concentrations of 13%, 16% and 20% with a temperature of 21°C. DiVenn has been used to illustrate the direction of gene expression with upregulated genes are shown in red and downregulated genes in blue. Yellow indicates conflicting up and downregulation of shared genes. Grey lines link between gene expresion and exposure condition and population.

STRING was used to identify protein interactions in the differentially expressed genes lists to focus identification of particular proteins and pathways of interest (Figure 93, Figure 94 and Figure 95). In the HA population comparison of temperature between 4°C and 21°C at 13% Oxygen several clusters of interacting genes are identified including those involved in (but not limited to): collagen interaction (COL12A1, COL6A6, PPIB, COL6A3, COL24A1, and COL21A1), ubiquitination (RNF19A, DTX3L, BTBD6, UBE2Q2, FBXL20 and UB32A), the Ras/MAPK/PI2K-AKT/HIF-1α signalling pathways (FGFR4, PTPN11, FLT1, MET, PRKACA, ANGPT1, EGF, YES1 and GRIN2B), apoptosis, lysosomal regulation and neutrophil degranulation (CECR1, GRN, CTSC, FUCA2, PRDX6, BPI, PGLYRP1, BIRC2, CTSS, CTSD, CTSL and CTSV), Heat-shock based mRNA processing (HSPA8, DNAJA1, SKIV2L2, RNPS1, SRSF7, HNRNPM, POLR2A, TARDBP and KHDRBS1), interaction with the Wnt-signalling pathway (LRP4, LRP5, LRP6, KREMEN1) and G-protein-coupled receptor activation (HTR2A, NMBR, NPFFR1, PLCB1, GNA13 and TRHR). At 16% Oxygen the clusters are significantly reduced to only a handful including: neutrophil degranulation (PRG3, BPI, MPO, GM2A and GRN), muscular growth (MYL3, MYH7 and TTN), regulation of cellular macromolecule biosynthetic process via NOTCH signalling (NOTCH1, RPS27A, CHD1L, LEO1, EIF4H, EIF4A2, CDC6, PSMC2 and PSMB2), alternative splicing (SRSF4, SRSF6 and NXF1) and collagen interaction (PPIB and COL24A1). At 20% Oxygen there is a continued protein collagen interaction (PPIB, MMP25, COL1A2, COL12A1, COL24A1, COL27A1, COL6A6, COL6A5, COL4A6, COL20A1, P4HA1, LOXL4 and TLL1), neutrophil degranulation (GLA, BPI, VAT1, PLAC8, ARSB, CECR1, CTSC and GRN) and a large network involved in the regulation of phosphorylation including the RAF/MAP kinase cascade pathway that play critical roles in the VEGF signalling pathway (BLNK, FGFR4, SKY, MAPK1, SRC, RHOA, CAV3, LYN, ACTB, YES1, PIK3R1, MET, EGRFR, EGF, NOTCH1, ERBB4, EGFR1, FGA, CANX, GAS6, SCARB2, DAB2, ADRB2, GLUL, UBC, SOCS3, TRIM21, BTBD3, KLHL21 and FBXL30), (Figure 93).

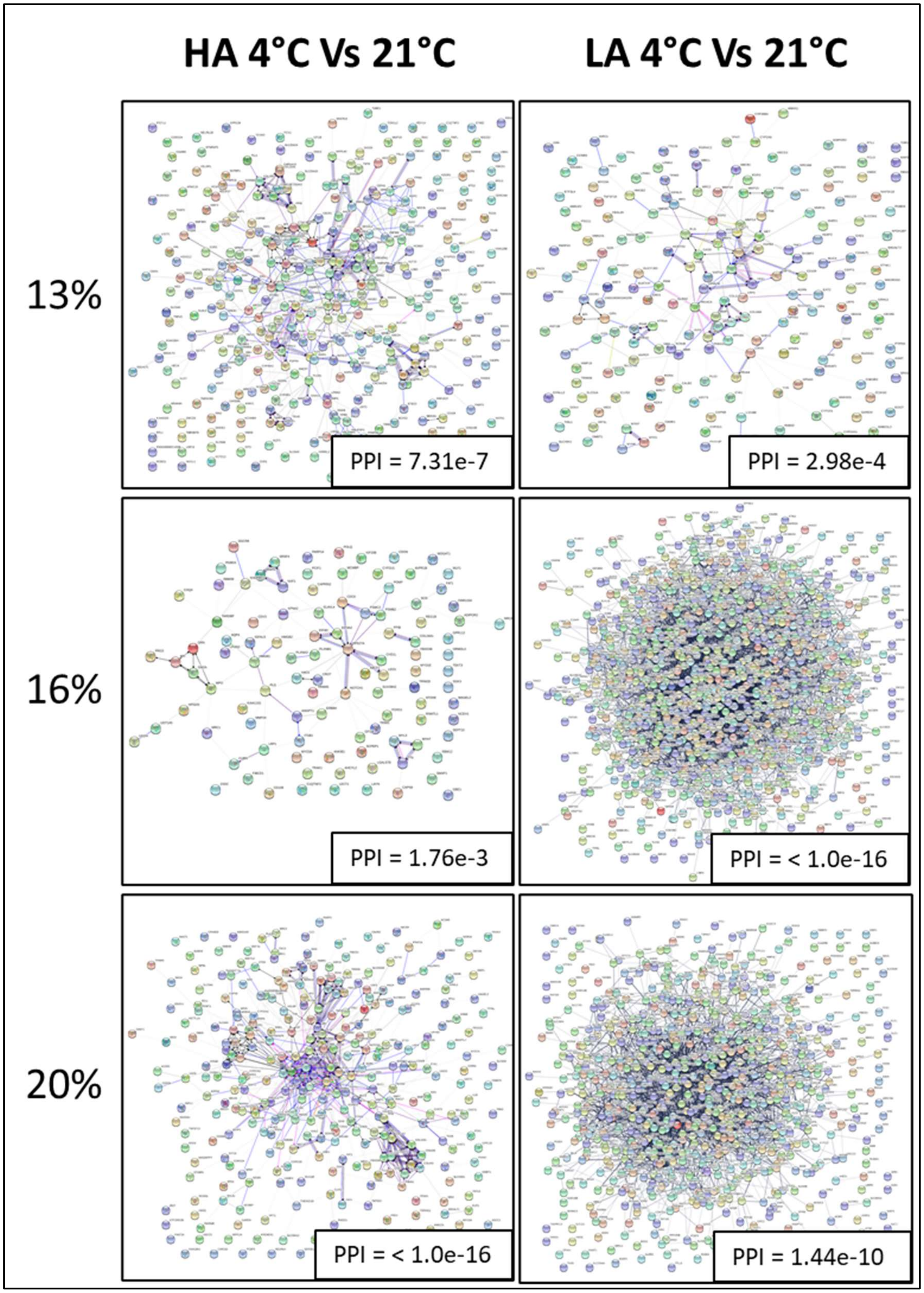


Figure 93: Protein interactions networks for genes differentially expressed in HA and LA populations maintained under different temperature regimes (4°C and 21°C) and at three Oxygen concentrations (13%, 16% and 20%). STRING was used to generate the networks, with the overall PPI (protein-protein interaction) enrichment p-value for each network is also presented. The network comparisons highlight the larger response the LA population has in 16% and 20% Oxygen. The networks highlight interactions between proteins idicating a greater likelyhood of biologicl impact, and therefore are of greater interest.

In the LA population comparison of temperature between 4°C and 21°C at 13% Oxygen several clusters of genes are identified including those involved in (but not limited to): collagen interaction (COL11A1, COL6A6, COL24A1, BMP1, ATP1A2, ATP1B1 and PPIB), signal transduction including through Ras/MAP kinase and HIF-1α (PRKACA, GRIN2B, MATN3, PTPN11, EGF, YES1, EGFR4, MET, FLT1, MMP14, ANGPT1, MMP3 and EGR1) and G-coupled neurotransmission (PSAP, HTR1A and HTR1D). When assessing protein interactions for the temperature comparison at 16% and 20% Oxygen the networks become too convoluted to identify individual groups of protein interactions (Figure 93). However, they do include many and more of the interactions already identified in the previously discussed temperature comparisons including the collagen interactions, the MAPK and PI3K signalling pathways and protein degradation/recycling. In these comparisons, it is clear the cells are undergoing a massive change in metabolism and growth in response to the cold.

In the HA population comparison of responses to Oxygen at both 4°C and 21°C there are fewer protein interactions identified in the networks than seen in the temperature comparisons in the same population, but there are common interactions identified. In the 13% Vs 20% Oxygen comparison at 21°C, the collagen interaction is present (COL24A1, COL12A1, COL6A6, BPM1 and PPIB), Notch/PI3K-Akt signalling (EGF, LYN, PTK2, ITGA1, AGRN, NOTCH1, ADAM12, MET, PTPRB and GAS6). In both the 13% Vs 16% and 16% vs 20% Oxygen comparisons, no major interactions are identified. For the comparisons of response to Oxygen of the HA population at 4°C, there are many smaller protein interactions. For the comparison of 13% Vs 20% Oxygen these include; chromatid cohesion (RAD21, SKA1 and INCENP), glucose metabolism and the cAMP pathway (GNAI1, GPER1, PRKACA, CDC25A, PKM and ADCY6), ubiquitination (FBXW7, TRIM36, RNF217, FBXL7, BTBD6 and DTX3L), muscular growth (MYH7, MYH4, TNNT2, TTN and MACF1), signalling via MAPK/PI3K-Akt/Ras and HIF-1α (EGF, ERBB4, LCT, FGFR4, PLG, HGF, CAPN9, FLT1, MET and PLXNB1). For the comparison of 13% Vs 16% Oxygen, the protein interactions include; ubiquitination (PARP14, DTX3L, UBE2R2, UBE2Q2 and FBXL20), protein and mRNA degradation (TUBE1, NXE1, TIAL1, HNRNPM, DHTKD1, SRSF1, PSMC2, KIF11, MCM6 and HSPA8), neutrophil degranulation and phase II conjugation (CTSC, GRN, CTSD, GLA, CECR1, BPI, HPGDS, ABCB1, CD63, PGLYR1, CHIT1, CYP1A2, GSTO1, GSTT2B and GSR), extracellular matrix organization through collagen, and cell cycle/differentiation (PRKCB, CASP8, RUNX1, ADAM12, WDR5, COL1A2, COL4A6, COL6A6, COL24A1, COL18A1, COL12A1, ERBB4, AGRN, TLL2, HS6ST1, BMP1, CTSV, SRC, LMNA, CSK, HSPG2, ANGPT1 and FGA). For the Comparison of 16% Vs 20% Oxygen, there weren't any significantly sized groups of interacting genes (Figure 94).

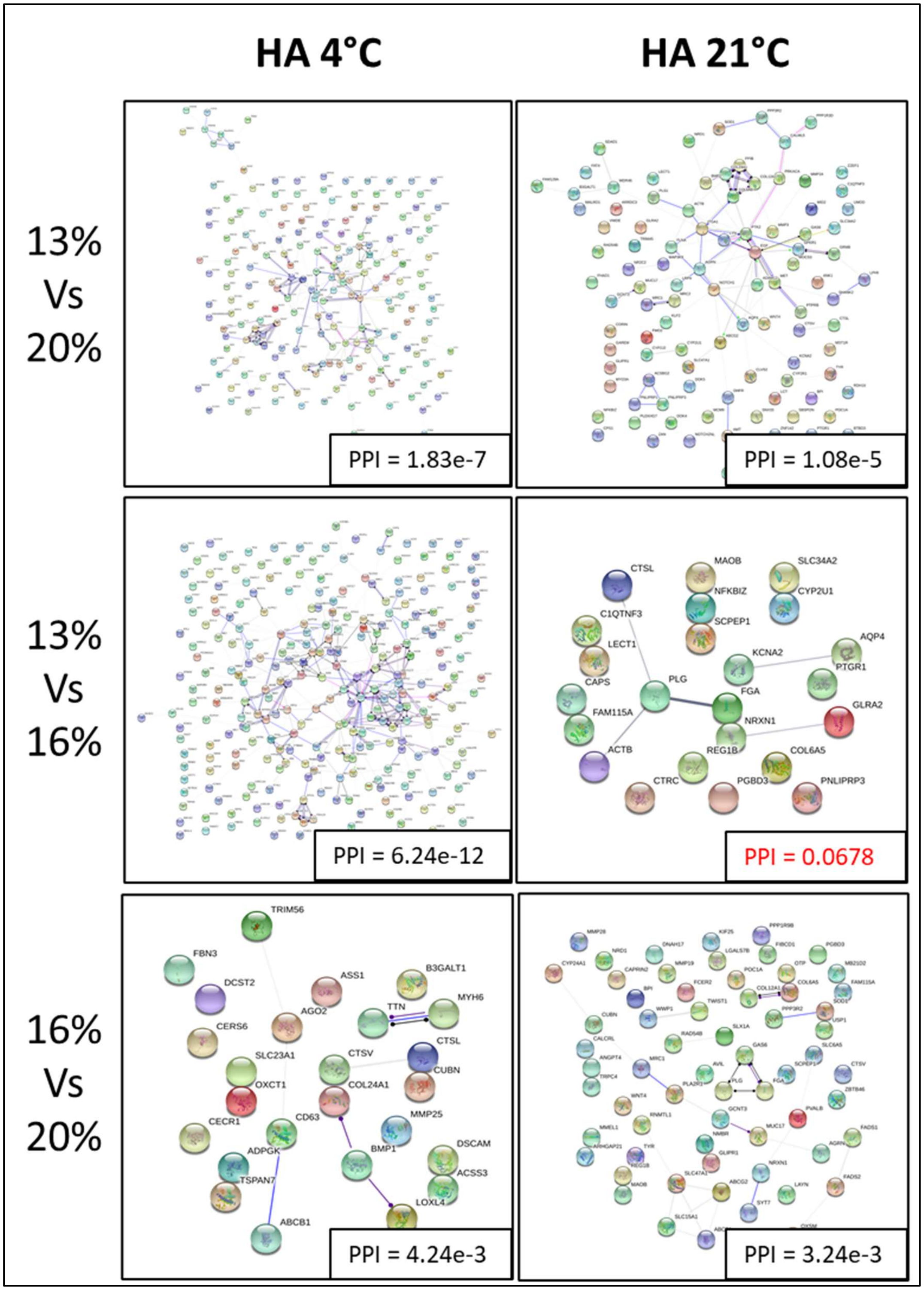


Figure 94: Protein interactions networks for genes differentially expressed in the HA population maintained at two temperatures (4°C and 21°C) and under different three Oxygen concentrations (13%, 16% and 20%). STRING was used to generate the networks, with the overall PPI (protein-protein interaction) enrichment p-value for each network is also presented. The network comparisons highlight the reduced response at both temperatures in the comparison of 16% Vs 20% Oxygen. The networks highlight interactions between proteins idicating a greater likelyhood of biologicl impact, and therefore are of greater interest.

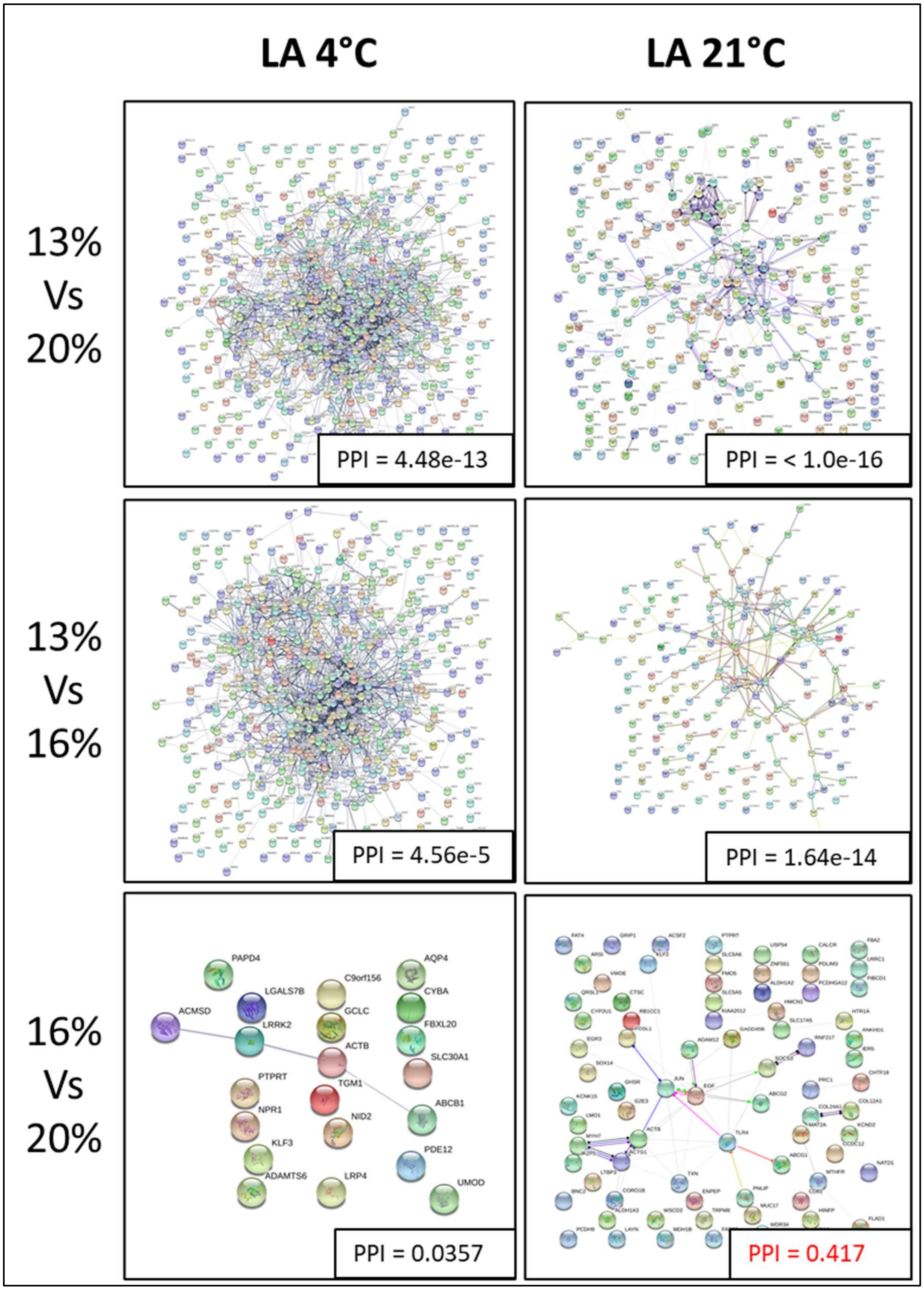


Figure 95: Protein interactions networks for genes differentially expressed in the LA population maintained at two temperatures (4°C and 21°C) and under different three Oxygen concentrations (13%, 16% and 20%). STRING was used to generate the networks, with the overall PPI (protein-protein interaction) enrichment p-value for each network is also presented. The network comparisons highlight the reduced response at both temperatures in the comparison of 16% Vs 20% Oxygen. The networks highlight interactions between proteins idicating a greater likelyhood of biologicl impact, and therefore are of greater interest.

When the same comparison of responses to Oxygen at both 4°C and 21°C is run in the LA population, there are more gene interactions seen than in the HA population. At 21°C, these include: metal ion binding (CALML5, CALM3, RHAG, OBSCN, CYP1A2 and SCGN), signal transduction including Ras/MAP kinase and ErbB signalling interaction with Golgi apparatus (ACTB, NOTCH3, NOTCH2, NOTCH1 CYP27A1, ITSN1, EGF, MMP3, ABCG1, DNM1, SYNJ1, MET, AGRN, LYN, GRB2, PAK2, ZAP70, MUC1, PTK2, PTPRA, CUBN, GA26 ADAM12 and FGFR4) and collagen interaction (COL24A1, COL6A6, COL6A3, COL11A1, COL9A1, COL12A, COL21A1, COL14A1, P4HA2, BMP1, LCT and PPIB). In the comparisons of 13% with 16% Oxygen the protein interactions include: gluconeogenesis and metabolism of RNA (SLC25A13, ASL, SLC25A12, ERCC4, PPIE, DOX46, RNASEH2B, HRNPL, CNOT4 and PABPC1), p50 DNA damage and cell cycle (CHAF1A, UHRE1, ERCC6L, DLGAP5, TIMELESS, PRC1, RORA, PSMF), cellular organisation of organelle lumen and the NOTCH, Rap1 and Ras/MAPK signalling pathway (RUNX1, KMT2D, SMARCE1, BAZIB, VIM, ACTB, ACTG, LMNA, NOTCH1, NOTCH2, ACTR2, DLL1, CALM3, EGF, MET, SOD1 and ADAM12) and collagen interaction (COL24A1, CALSP1, COL14A1, PKM, PPIB, COL6A6, COL21A1, PRKCA, ATP1A3, ATP1A2 and ATP1B1). As seen in the HA population there were limited protein interactions with the 16% Vs 20% Oxygen (Figure 95).

For the LA population at 4°C, there is a large network of interacting proteins similar to those seen in the LA population comparison of differentially expressed genes at 4°C Vs 21°C at 16% and 20% Oxygen. These heavily interacting protein networks include many of those observed in both the HA and LA population Oxygen and temperature comparisons. The comparison of the LA population of 16% and 20% Oxygen at 4°C shows a similar pattern as seen at 21°C in the LA population with very few interacting proteins.

6.4. Discussion.

6.4.1. **Worm weights.**

There was a general reduction in average worm weights across experimental groups (though within Standard Deviation and not a statistically significant reduction). This reduction was highest in the HA population, though they were on average also slightly larger than the worms for the LA population. Due to the worm's semi-translucent nature of *Aporrectodea caliginosa*, it is possible to estimate how filled the gastrointestinal (GI) tract is, and while not directly measured, there was a general reduction of GI tract contents following the two weeks experimental exposure. The reduced GI tract contents were also confirmed during dissection. The reduction in food consumption, leading to a reduced GI tract content is not entirely surprising as soil disturbance, and in this case transferring worms to new boxes of soil for experimental exposure, in the short term negatively impacts and stresses earthworms (Duhour et al. 2009; Fox et al. 2017; Lemtiri et al. 2018). It is worth balancing the knowledge of this stressful condition with its universal application to all experimental worms. Differential expression should therefore account and mitigate the transposal stress.

6.4.2. **Sequencing and read mapping.**

Library preparation of extracted RNA for sequencing proved challenging. To minimise the duration of time the experimental worms spend out of the experimental conditions before harvesting and to provide a more rounded understanding of gene regulation, entire body cross-sections were dissected, ground and frozen in Trizol reagent to prevent RNA degradation. Ridgeway and Timm compared the isolation of RNA through different methods and found Trizol to return the greatest yield, an important consideration with limited source material (Ridgeway and Timm 2014). They suggest a Trizol extraction can leave chemical contamination, though this is easily solved with the purification clean up kit used. The inclusion of the gut with the body wall also came with gut microbiota. This microbiota released RNA of its own during RNA extraction leading to an overestimation of RNA concentration. The variation in the level of bacterial RNA within earthworm RNA added complexity to gauging RNA starting material during the initial bead clean-up of the RNA library synthesis. This step removed the contaminating bacterial RNA to only leave the individual earthworm RNA, however, this was at a lower concentration than anticipated and consequently, though sufficient for sequencing, the final library concentration was lower than targeted. This also impacted the balancing of the samples during sequencing. Read sample counts reached over 8 million per sample though this was lower than expected due to an over clustering of one sample. Future RNA extractions would benefit from mRNA purification via poly A tail clean-up prior to quantification and use library generation.

When read mapping (STAR v2.5.3) was performed and alignment scores compared, no discernible difference could be made between HA and LA populations, despite the genome being of HA population origin. This indicates as deduced in Chapters 3 and 4, the *A. caliginosa* population on Pico has low genetic diversity. Further, there was no correlation between alignment scores and the sequenced read count for each sample. The low variation in alignment score and no mapping trend within experimental groups, gives confidence differentially expressed genes are not as a result of variation within mapping efficiency. Further the level of reads mapped fall within the expected range expected with the use of STAR (Ballouz et al. 2018).

6.4.3. **Differential gene count.**

Prior to performing differential expression analysis using DESeq2 on individual experimental comparisons, it was run on all samples as a diagnostic for sample balance. This indicated some variation within the total read count per sample imputed to differential statistics, and a small variation of null counts per sample which are not taken into account (excluded) for the analysis with DESeq2. A histogram analysis determined the approximately 40% (17,522) the gene objects had no gene counts while the other approximate 60% (24,966) had fewer than 10,000 counts per gene object. This is roughly in line with the level of transcripts annotated by Sander *et al.* (Sanders et al. 2014). There were virtually no gene objects with gene counts higher. This was also plotted as a density distribution for each sample to identify patterns of read counts per gene object. In this, some variation was observed in both HA and LA population samples in the mid-range of gene counts, but the general pattern of no-count and high count was similar in all samples. Statistics were also performed to identify over-represented reads from each sample. While the percentage of reads lost to these over-represented reads was quite low in most samples, it was quite high for sample 1_3 (Replicate 3 of a LA individual at 21°C at 13% oxygen). However, due to the type of differential statistics performed and the over-represented read being see at a high level in all samples, it was straightforward to account for this during analysis.

Principal component analysis was performed to identify patterns in gene expression between samples. While PC1 did spread samples across 24.9%, there was no discernible pattern to this that could be correlated with exposure conditions or groups. With PC2, samples were distinctly split at 10.1% between two groups of the native individuals and the experimental individual. Another distinctive split was seen with PC3, where samples HA and LA individuals split cleanly at 6.7%. While the PCA analysis is useful in interpreting data, as evidenced by PC1, not all important patterns can be accounted for (Lever et al. 2017). This was the first strong indication of a completely alternative response mechanism between HA and LA populations. Despite both populations having been normalised in the same soils and conditions for over six months, there was a clear pattern of response by each population.

6.4.4.**Gene ontology.**

With all multifactorial experiments, it can be a challenging task to account for more than one variable whilst not losing resolution of differential expression, for example, if we were to look at how the HA and LA populations respond in general across all conditions, you can isolate a group of genes that have an expression pattern in each population, but the fine resolution of the effect of temperature or Oxygen level is lost. In effect adding too much complexity obscures nuanced changes by increasing the level of background variation. It is, useful to perform these larger scale analyses to ground an understanding of general biological response despite the loss of differential gene expression resolution, but also vital to isolate each condition separately and run differential analysis separately and compare pathway analyses with each other. It is also important to understand the pattern of expression in the native samples. While these basal expression patterns will undoubtedly be different from the expression distribution seen in experimental worms, it is important to isolate what expected similarities and differences in gene ontology they have with experimental worms. In this particular comparison, there was a general increase in developmental and organism processes within the HA native population. These included processes associated with the circulatory system; “Vasoconstriction” and “Blood vessel morphogenesis", elements identified in high altitude adaptions of Tibetan Chickens (Zhang et al. 2016). Biologically, this follows a logical pattern to observe. Conditions at the top of a mountain as discussed in Chapter 1, are hostile and have highly variable and dynamic challenging conditions.

As expected with the broader comparisons, resolution is lost in differential gene expression. In the HA Vs LA population comparison, there are only a few GO terms identified that do not shed much understanding on how each population responds differently to each condition. This is a pattern repeated in the temperature and Oxygen comparisons. To understand the regulatory response at higher resolution, each triplicate of individuals for each condition were compared with each other. To elucidate the effect of temperature, the HA and LA population were looked at separately for each Oxygen level separately. To elucidate the effect of Oxygen, the HA and LA population were looked at separately for each temperature.

When looking at the raw numbers of differentially expressed genes between 4°C and 21°C, we can see the HA population has a few hundred more differentially expressed genes than the LA population in 13% Oxygen. In contrast, and quite dramatically, at 16% and 20% Oxygen, the LA population has thousands more differentially expressed genes than the HA population. It is proposed that at 13% Oxygen the HA population is starting to go through stress processes while for the LA population it is already at the point of metabolic hibernation commonly seen as a response in seasonal winters (Singh et al. 2019). At the higher Oxygen levels, both populations can function, however, while the HA population is more primed (epigenetically or adaptively) for this challenge, the LA population is undergoing a full stress response.

The greater number of differentially expressed genes in the HA temperature comparison at 13% Oxygen indicated a more concentrated response to stimulus, rather than the LA populations' more general response to stimulus and increase in multicellular organismal process. This suggests that despite having more differentially regulated genes, the HA population is actually responding to the increased challenge in a more refined approach.

Despite 172 differentially expressed genes, the temperature comparison for the HA population when at 16% Oxygen yielded only one statistically significant enriched GO term. When at 20% Oxygen there were 639 differentially expressed genes lead to a substantial increase in functionally enriched GO terms that relate to elements associated with cold response such as glucose metabolism and vasoconstriction which I will discuss later.

What is of particular interest is that in the LA population temperature comparison at 16% and 20%, despite 2031 and 1329 differentially expressed genes respectively, there are no functionally enriched GO terms for the comparison at 16% and only 2 enriched terms at 20%. A more in-depth analysis of this result, identified that almost all GO terms were lost through over-representation as widescale changes mask more narrow biological responses. FDR is intended to account in this case for the identification of enriched terms by chance (inflation of Type 1 error), e.g. if 100 genes are all linked to one GO term, then the likelihood of this being by chance is low, while if those 100 GO terms are linked to only one gene each, the likelihood of this being by chance is high. Vital to gene enrichment is the identification of an appropriate background, which should include all possible genes that could undergo differential gene expression (Simillion et al. 2017). Enrichment analysis performs over-representation analysis, (Fisher's exact test), to try and identify what the likelihood of each GO term being identified is by chance. This functions by comparing sample frequency with background frequency. An example of how this affects gene enrichment is: GO term "X" with 6 differentially expressed genes out of 10 being observed for the sample frequency. There are 50 genes associated with GO term "X" out of a genome of 5000 making up the background frequency. From this, an expected frequency can be calculated and compared to the sample frequency and an FDR calculated. In these two differential expression lists with thousands of genes, we can observe many GO terms that are over-represented, however not by a margin that falls within FDR. If we utilise a human background list the FDR shifts to allow most of these GO terms to become enriched. This is because the human background list is over twice as large as *A. caliginosa*'s, shifting and reducing the expected frequency. This all indicates that the global expression profile of these LA worms is shifting to a small degree but not through a specific small set of biological processes.

There is a similar pattern of increased differential gene expression in the LA population compared to the HA population when we look at the comparisons of Oxygen exposure at both 4°C and 21°C. In all 4 comparisons (HA and LA at 4°C and 21°C) of 13% Vs 20% Oxygen, there was gene enrichment. This centred on changes to "Development process "and "Signalling and immune system processes", with "Cell morphogenesis" and "Developmental processes" also appearing at 21°C. In the comparisons of 13% Vs 16%, there is a limited response by the HA population at 4°C and no enriched response detected at 21°C. Similarly, there were very few enriched terms in the 16% Vs 20% 4°C and no enriched response detected at 21°C. This suggests that for the HA population there was a similar coping mechanism for survival at reduced Oxygen levels, and reduced Oxygen is not a substantial pressure for the HA population to contend with. The LA population had a more pronounced response at 4°C and 21°C in the comparisons of 13% Vs 20 and 13% Vs 16% Oxygen. These include "Response to stimulus", "Regulation of immune response", "Regulation of organismal process" and "Cell morphogenesis". All of which indicate a stress response that is more active than seen in the HA population.

6.4.5. **Shared gene expression analysis.**

To better understand the commonality of differentially expressed genes between experimental conditions, DiVenn networks were generated. This allows a visualisation of the direction of gene expression including commonality and contrast. In the DiVenn network examining the effect of temperature, it provided a key insight into the different response each population takes.

There was a very large set of commonly up and downregulated genes between the LA population at 16% and 20% Oxygen. These two sets of differentially expressed genes, as covered earlier did not yield a statistically powerful gene ontology network, but the commonality in expressed genes would suggest that very similar pathways are responding with the LA populations reacting to the temperature change in a similar manner. However, they do not respond in exactly the same way as suggested by the large groups of unshared differentially expressed genes. Interestingly, there are very few shared differentially expressed genes between the LA population temperature comparison at 13% Oxygen and the LA populations at 16% and 20%. Not only that, but of the shared differentially expressed genes, many of the LA at 13% differentially expressed genes are in opposing expression directions of those in the LA at 16% and 20%. In the small cluster of shared downregulated differentially expressed genes between the LA populations at 13% and 16%, of particular interest is the downregulation of PTPRA which has been shown to activate the proto-oncogene family of Src tyrosine kinases. This suggests at higher temperatures, there is more oncogene function at lower Oxygen levels, potentially as a response to hypoxia not seen with the LA 20% Oxygen population (Sommer et al. 2010). It is possible the presence of GDF11 and its downregulation moving from 4°C to 21°C is linked with a reduction in calorie restrictions as metabolism increases (Katsimpardi et al. 2020). Finally, a downregulation of UBA1, which is involved in ubiquitination, correlates with a decrease in the breakdown of metabolic proteins associated with a more active metabolism that can occur at higher temperatures.

As briefly mentioned in Section 6.3.1., there are far fewer shared differentially expressed genes between the HA population temperature comparisons. Among the shared genes, there is an increase of ANK2 associated with high expression in cardiac muscle and COX6B1 which as part of the cytochrome c oxidase complex indicates an increase in ATP production. FOXO4's upregulation could be linked with a variety of functions, though increasing FOXO4 could be mitigated by the positive regulation of the PI3K pathway detected in gene enrichment (Figure 27) (Calnan and Brunet 2008). Of further interest is the down regulation of the proto-oncogene YES1 in the HA populations at 16% and 20% but up-regulation in the HA population at 13% Oxygen, suggesting similarly to as discussed with the LA population, the increase in temperature is leading to an increase in response to hypoxia. One potential hypothesis of the low number of shared differentially expressed genes is the plasticity of response to stress. It is possible the HA population utilises a more diverse set of genes for response.

Looking for commonality between response of the LA and HA population, identified HMGB1 and HMGB2 whose functions are involved with rapid response to environmental change and have a vital role in the control of activation of chromatin critical domains (Bianchi and Agresti 2005). It is thought the trigger for its upregulation stems from inflammation, immune response, and chemotaxis and it has roles in DNA replication, repair and transcription (Martinotti et al. 2015). In both populations, expression increases moving from warm to cold conditions. Despite HMGB1/2 controlling many environmental response genes, and both populations displaying upregulation, the HA population has fewer differentially regulated genes. This indicates the HA population utilises this pathway in a potentially more specialised and nuanced application to only the most challenging conditions. The LA population with its very high levels of differential gene expression, does not display the same control. An increase in PPIB in the LA population which has a role in inflammation helps to support this theory.

For the comparisons of Oxygen at 4°C and 21°C there were similar patterns observed. At 4°C there were very few differentially expressed genes between 16% and 20% in both HA and LA populations. This suggests that at a colder temperature where metabolism is slowed, the effect of a reduced Oxygen presence is minimal. As such there are very few shared differentially expressed genes between all HA and all LA Oxygen comparisons. The largest groups of shared genes are between the comparison of HA 13% Vs 20% Oxygen and HA 13% Vs 16%, and with a much larger group of the comparison LA 13% Vs 20% Oxygen and LA 13% Vs 16%. In both cases, the groups have almost uniformly downregulated differential expression. This suggests a similar approach taken in response to the Oxygen challenge, of reducing cellular processes, but as these downregulated genes are not shared between HA and LA population, there is a different set of processes to downregulate.

The lowest change in differential expression was in the comparison of HA and LA population exposed at 21°C at different Oxygen levels. Here there were almost no changes in differential in the HA population. Unlike at 4°C, the HA 13% Vs 16% Oxygen had the fewest differentially regulated genes but did noticeably share a group of 8 genes with the HA 16 Vs 20% that were all in opposition of differential expression. Of these of interest is MAOB which is known to generate reactive Oxygen compounds, which is upregulated at lower Oxygen levels and downregulated at higher Oxygen levels (Nagatsu and Sawada 2006). Also, of interest is the upregulation of WNT4 in both HA population comparisons of 13% Vs 20% and 16% Vs 20% which is involved in development. The low levels of differential expression suggest that for the HA population, these Oxygen challenges are not of particular stress. For the LA population, the response is higher and would suggest that they are less adapted to these environmental challenges. The LA population comparison of 13% Vs 16% and 13% Vs 20% Oxygen is similar to the pattern seen at 4°C with a lower set of differentially expressed genes in the LA 16% Vs 20%. However, at 21°C the pattern is not of universal downregulation of genes. Many and in particular between the LA 16% Vs 20% and 13% Vs 20% are upregulated. This would suggest at different temperatures, response to reduced Oxygen levels involve separate mechanisms.

6.4.6. **Protein interaction analysis.**

The purpose of running STRING was to evaluate connections between differentially expressed genes. STRING converts the genes to associated protein and looks for known interactions. The more connected a differentially expressed gene (protein) is, the more confidence that can be drawn to a pathway's activation. We can also identify more specific proteins and pathways that can be lost in the noise of larger networks.

Examining the effect of temperature in both the HA and LA population at all different Oxygen levels there are some common pathways under activation. Collagen-based extracellular matrix interaction is seen in all temperature comparison along which is known to change in mammalian extracellular matrix with temperature change (Jones et al. 2014). Neutrophil activity is also seen in all comparisons which might indicate the additional environmental stress is adding vulnerability and exacerbating the individuals' immune system, evidenced further by the differential expression of FOXO1 and FOXO4 whose action has been reported as regulating bacterial-induced Neutrophil activity (Dong et al. 2017). Of interest was the activation of Raf and MAP Kinase cascade pathways which have a plethora of known downstream functions including

VEGF (Doanes et al. 1999). It was striking to see the activation of genes within the HIF-1α pathway in the HA and LA temperature comparison at 13% Oxygen as this critical pathway is directly responsible for cellular response to hypoxic stress, something only expected in very low Oxygen environments (Bruick 2003; Majmundar et al. 2010). We know from the work of Giardina that while there is only a small change to Oxygen dissociation between 20% and 16% Oxygen, there is a rapid decline shifting towards 13% Oxygen (Giardina et al. 1975). While cellular concentrations require <5% to start activation of the HIF-1α pathway and <2% for strong induction, below the exposure at 13%, there far less Oxygen is delivered to cells. In testing *in vivo* on mice, it is common to see the pathways strong activation at 10% Oxygen which have a more complex Oxygen delivery system than earthworms (Schofield and Ratcliffe 2004; Ratcliffe 2013; Semenza 2014).

Within the Oxygen comparisons, a similar pattern of collagen interaction is identified as well as the pathway activations of MAP kinase, Ras, PI2K-Akt and in the 13% Vs 20% HIF-1αThere is also changes to the regulation of glucose metabolism at 4°C and increased ubiquitination suggesting increased breakdown of proteins. In both populations it is clear there is an increased network of interaction at 4°C and at lower Oxygen levels and more centred on the activation of pathways associated with repair and response to stress.

6.4.7. **Adaptation or acclimatisation.**

In this Chapter we have covered an extensive process of RNAseq preparation, mapping, differential expression analysis, and pathway analysis. This chapter aimed to identify, if possible, any genes of pathways associated with response to the environmental stressors of low Oxygen and low temperature and to identify, if possible, any adaptive differences the HA population might have compared with the LA population. There was a large number of annotated genes that shared homology with known human genes. This has allowed the later examination of genes and pathways that are highly researched and well documented. There was however a substantial number of genes that remained unannotated, that likely are either too divergent to identify in from any of the searched databases or are previously unidentified and unresearched. As there was a notable proportion of read counts that mapped to these unknowns, there is undoubtedly an element missing from the finial story that as future works in gene identification progress, we can return to complete this story.

The balance of read mapping between populations allowed confidence that later differential expression analysis comparisons between populations were unbiased, and differences observed were 'true'. Fundamentally, the analysis presented here demonstrates that variation in temperature is a stronger environmental stressor for *A. caliginosa* than that of lowered Oxygen levels. Moreover, the HA population responds in a functionally different approach, both at a pathway level, but also with key gene regulation. The HA population exhibited signs of greater plasticity of response to the stressors than that of the LA population of both temperatures and Oxygen levels. The HA population possibly through epigenetics is more adapted to surviving extreme environmental challenges. The plasticity of response is similar to that examined in temperature and hypoxic stressors of fish (Gracey et al. 2004; McBryan et al. 2013). Future research should focus on identifying if epigenetics is responsible for allowing the fast and controlled activation of environmental stress pathways that allowed the HA population to respond with lower differential gene expression, but that in the LA population saw a massive and cellularly costly reaction. An example that evidences epigenetic acclimatisation as a method for temperature tolerance was identified in filamentous fungus, where the lack of H2K4 or H3K36 methylation led to an impairment in response to temperature fluctuations (Kronholm and Ketola 2018).A similar conclusion of the importance of epigenetics to acclimatisation is observed in diatoms where histone modifications allow improved survival during ocean acidification (Huang et al. 2019). In multicellular organisms, there appears to be a balance of short term homeostasis regulatory acclimatisation in a per individual basis and epigenetic modification as a long term acclimatisation (Horowitz 2014). It is likely this combination of short term and long term acclimatisation is present in the HA population. Despite the long period of normalised conditions prior to experimentation, the HA population had a rapid and distinctly different expression of stress response genes that indicate there is an element of either epigenetic acclimatisation or adaptive changes to the genome affecting downstream response.

Throughout this last decade we have observed an increase in extreme weather events (Stott 2016). These extreme weather events are not only increasingly common, but the severity of them is also escalating. Landscapes can now swing wildly from drought to flooding rapidly and the ability for organisms to survive is under pressure with estimates higher than half of all species under threat of extinction (Roman-Palacios and Wiens 2020). The ability for species to adapt to rapid extreme climate events might be critical to their survival moving through this century and from the research in this chapter there appears to be a high altitude adaptation, where extreme weather has long been present, that gives *A. caliginosa* an advantage in this fight for survival. It is likely that given the same environmental pressures, other species found at high altitude have undergone a similar adaption to rapid environmental swings, and whose importance cannot be understated.

Having identified signatures of long term acclimatisation, in the next chapter, I will to look for markers of acclimatisation through the investigation of SNPs and Fst. Where possible, this will be mapped with data from this chapter to identify any links between expression and key genes of interest.

7. Genetic variations of altitude in *A. caliginosa*.

7.1. Assessing genetic variation.

7.1.1. **Why assess genetic variation.**

An individual's ability to survive and prosper as a function of its genome is not only the ability to survive challenging conditions, but also gain a superior advantage of survival over other individuals, a rival population or competing species. The fundamental concept of evolution individuals is "*survival of the fittest*", those that are more adapted to an environment will have a selective advantage over others that are less fit, this enables the genetic traits that underly this fitness to be passed on to their offspring. For this premise to operate variation must continually be introduced into the genome with mutations that are unviable or 'less fit' being lost from the population over time, while mutations resulting in favourable traits surviving and eventually becoming ubiquitous within a population. Within a population, mutation over time, driven by processes like replication error and DNA damage, will increase a population's genetic diversity. Where these mutations accumulate is not evenly distributed across the genome, with these variations generally only remaining in areas of the genome that are not immediately harmful or overly disadvantageous to the survival of the individual. In diploid (or higher ploidy) genomes, single base positions where there is variation found (e.g. more than one base) are called single-nucleotide polymorphisms (SNPs), and they can be used to measure the genetic diversity of a species or population, but also can be used to identify areas of the genome that are under selection.

7.1.2. **Aims of this Chapter.**

In this chapter, we will take RNAseq reads, sequenced, and processed in Chapter 5, and through a process of Haplotype calling, base-recalibration, and re-calling haplotypes to generate files of variants found within the genome for each individual. This file will be filtered for quality before evaluation of the effectiveness and success of the calling. Downstream analysis will attempt to compare HA and LA populations, identifying genes that are under selective pressure in the HA population as identified through F_{ST} analysis and nucleotide diversity analysis. Identified HA genes passing these filters will undergo gene ontology analysis for gene pathways or families under selective pressure. I will not be performing altered gene function prediction in this analysis.

7.2. Technical approaches to SNP Calling.

7.2.1. **SNPs: RAD-tag, Mybaits, Genome ReSeq and RNAseq detection.**

In the pursuit of assessing the genetic variation in species populations, two factors have traditionally played off together. Having a large number of individuals to represent the population and having as much genetic information as possible at an affordable price. Even with the advent of cheaper sequencing, research has looked for ways of maximizing 'data for buck'. In well-characterised organisms such as humans, rats, mice, and other 'big' mammals, thousands of SNPs can be assessed quickly and relatively cheaply with SNP arrays, though this is slowly increasing (Louhelainen 2016). For organisms that have not received the same level of research and development and those that have very high levels of genetic diversity, assessing SNPs requires less direct and more expensive approaches.

One of the most heavily used methods in the last decade for fast and cost-effective genotyping and polymorphism identification has been restriction site associated DNA markers, RAD-tag. This allows a sampling of the genome by isolating size selected fragments anchored to specific restriction sites, which are then purified and sequenced. These sequences are then compared, grouped into 'stacks' of nearly identical sequences, SNPs identified and mapped into a reference genome (Miller et al. 2007). In recent years, modification of this approach by coupling restriction associated fragment generation with the use of biotinylated oligonucleotides designed to represent known sequences to capture specific or desired loci onto a solid phase, such as magnetic beads, an approach that has acquired the acronym RAPTURE (Ali et al. 2016). These techniques are particularly useful when identifying regions of the genome under selection, but loose resolution when attempting to identify the specific SNP associated with an adaptive phenotype (Puritz et al. 2014). If a genome fully characterized suites of modified oligonucleotides can be designed to enable solid-phase capture, such as those source from MyBaits, allowing isolation of sequencing library fragments spaced a designated distance apart throughout the genome (Silva-Junior et al. 2018). This, in practice provides a more even distribution of sequence information across the entire genome, or even a specific area of design (e.g. a chromosome of interest), whilst still not requiring the expense of full resequencing of individuals genomes. Genome ReSeq (whole-genome resequencing) is the sledgehammer approach to SNP detection. This is the most expensive but is also potentially the most powerful approach that can be employed. It is also one of the simplest techniques to perform, with a simple library preparation and barcoding required, avoiding more complex enzyme digestions. This technique requires low-level sequencing of all individuals and identification as individuals or as pooled samples, but requires more than 50X coverage for effective SNP coverage (Gautier et al. 2013; Fracassetti et al. 2015; Dorant et al. 2019). Benefits of this method include coverage of all areas of the genome including gene and promoters without missing areas in the gaps seen in RAD and MyBait methods, but the cost is the major drawback, even as pooled samples.

Lastly, another method that can be used for the identification of SNPs is the analysis of RNAseq data. The benefits of this method include the reduction of sequencing required as only exons are sequenced allowing far fewer reads required for the same depth of coverage, and in the case of this thesis, existing RNAseq data can be leveraged for additional information an no additional cost (Adetunji et al. 2019; Brouard et al. 2019). The drawback to this approach is that the SNPs identified will only occur on coding areas of the genome missing upstream promoter regions, introns, intergenic regions etc, which can also be targets of selection. Moreover, this method may also fail to detect SNPs, from lowly expressed genes that don't result in enough sequencing coverage.

7.2.2. **SNP calling platforms.**

SNP calling for RNAseq data relies upon a much narrower band of platforms than available for DNA Reseq data. The three main methods utilised are Genotyping by sequencing (GBS) through tools like Samtools and BCFtools, through variant calling software such as the Genome Analysis Toolkit (GATK) (Brouard et al. 2019) and FreeBayes (Garrison and Marth 2012). Since its first release and subsequent development by the Broad Institute, GATK has become the most widely used and trusted platform for identifying SNPs in a comprehensive, adaptable and stringent approach (McKenna et al. 2010; Zhao et al. 2019b). Zhao *et al*. concluded that GATK could achieve 100% accuracy in the detection of SNPs and 'Best practices' for SNP detection have been collaboratively developed and improved to minimize missing SNPs or allowing false positive (DePristo et al. 2011).

There are many ways in which to analyse SNPs once identified. The direction of analysis and the tools employed largely depend on the type of data DNA or RNA based, and the data quality and quantity of data available. There is no easy one size fits all approach. In the case or RNAseq data, there is a loss of potential information on SNPs that affect promoter regions, however there is an already narrowed focus of data. All SNPs can immediately be located to an exact gene, (or as of yet, unidentified gene). It alters the common statistical methods for detecting linkage disequilibrium but can be compensated through other stringent methods. A common workflow in DNA SNP based analysis would follow as: clustering of SNPs, filtering of SNPs, calculating linkage disequilibrium, identifying genes within regions followed by functional analysis. In RNAseq SNP based analysis, the difference in allele frequencies of SNPs between populations (Fst) can be calculated and average per gene. This approach could over and underestimate the impact of very large and very small genes allowing the potential for False negatives to occur, however, it can be mitigated by relaxing initial filtering of Fst values and subsequently imposing minimum SNP observations per gene. From here onwards, genes that pass the filter can be plotted to identify Contigs or Chromosomes to identify possible regions that are indicative of linkage disequilibrium. Functional network analysis can also provide a method for the statistical likelihood of pathways under selective pressure. In both DNA and RNA based methods, the level to set filters is largely dependent on how the data appears during analysis.

7.3. Methods.

7.3.1. **Generation of RNASeq data.**

The RNASeq data used for identification of SNPs was generated as described in Chapter 6.2.3, 6.2.4 and 6.2.5. The raw and unprocessed data were used and processed as individual samples and not pooled at this stage of analysis, however later analysis will compare HA and LA populations.

7.3.2. **Generating Variant call format files for HA and LA population.**

Raw RNAseq reads were manipulated through a complex set of bespoke commands to process the data from to final VCF files for each population as detailed in Chapter 2.9.19. The GATK pipeline was used to perform this task, which utilises a suite of tools including GATK4, STAR, Samtools and Picard (Li et al. 2009; Dobin et al. 2013; Brouard et al. 2019; Picard-Team 2019). The codes and protocols used are detailed in GATK's Best Practices Workflow and Chapter 2.8.4 (Brouard et al. 2019; GATK-Team 2019). Hard filtering was performed using reverse sequential steps of the 'grep' command to remove poor and low-quality data below the thresholds of, (DP < 10, QD < 5, MQ < 40, SOR > 3). The pipeline of use is detailed below in Figure 96.

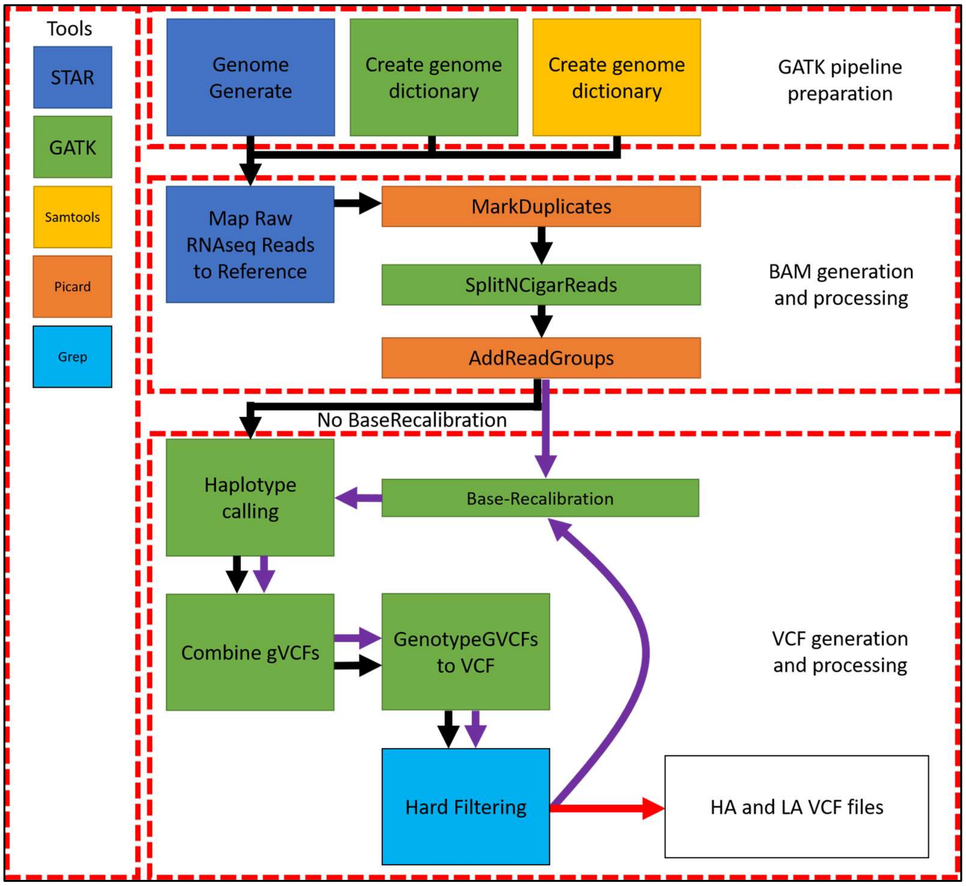


Figure 96: Pipeline used for generation of a HA and LA population VCF file using GATK, STAR, Samtools and Picard. The GATK pipeline required a genome dictionary generated and compatible with GATK and another one with Samtools.

7.3.3. **Phylogeny.**

To identify if SNPs within individuals were associated with a particular population, a phylogenetic tree was calculated with SNPhylo using a maximum PNSS of 25, maximum autosome to include all contigs (6452), MAF threshold of 0.05 and Maximum missing rate of 0.5 and was visualised in FigTree (Lee et al. 2014; Rambaut 2016). Principal component analysis and Multi-dimensional scaling analysis was calculated from the SNPs identified from all 42 samples from both the HA and LA population using TASSEL (V5.2.6) (Bradbury et al. 2007).

7.3.4. **Summary statistics of genetic variation, divergence and selection.**

VCF files were examined in the Integrated Genome Viewer (IGV) to assess the pattern of SNPs across contigs. Fst was calculated per SNP with VCFtools with a Minor allele count (mac) of 8 as detailed in Chapter 2.9.1. Nuclear diversity was calculated per SNP for the HA and LA population and Tajima' D was calculated with a window size of 30,000 bp for the HA and LA population. Average Fst and Pi (value for nuclear diversity) per gene was calculated by averaging Fst values within the start and endpoint of the gene as identified in the Gtf file from chapter 5. A histogram and quantile distribution was calculated in R, where observation of distribution derived a cut off of the top 10% of average gene Fst values. To identify possible sites for HA selection over LA population, results were filtered to only include values where HA-Pi was half that of LA-Pi or less. Markers of selection associated with the LA population were not examined in these analyses. To minimise random high Fst values, genes were filtered to remove those that had fewer than 5 SNPs. General Linear Model analysis was run on all SNPs with TASSEL (V5.2.6) and the P-values plotted as a Manhattan plot.

7.3.5. **Gene Ontology Analysis.**

Genes within contig regions indicating higher linkage disequilibrium for genomic selection were identified using the blast results from Chapter 5 and underwent gene annotation with GOnet, REVIGO. To identify if any of the genes under selective pressure were also identified in Chapter 5s differential expression analysis, DiVenn was used to highlight these connections. Connections with genes that passed the Fst and Pi filters to genes of interest HMGB1, HMGB2 (and linking protein TP53) were investigated using STRING-db (Supek et al. 2011; Pomaznoy et al. 2018; Szklarczyk et al. 2019).

7.4. Results.

7.4.1. **VCF files generated.**

Three VCF files were generated; a HA population VCF, a LA population VCF, and a combined HA and LA population VCF. The number of SNP sites detected in each VCF pre and post Hard filtering are shown below in Table 23. VCF files were visualised in IGV as illustrated below in Figure 97.

Table 23: Number of SNPs detected in HA, LA, and HA and LA combined VCF files.

	HA	LA	HA and LA combined
Pre-filter	1,956,204	2,192,359	2,898,639
Post-filter	720,249	810,077	1,082,309

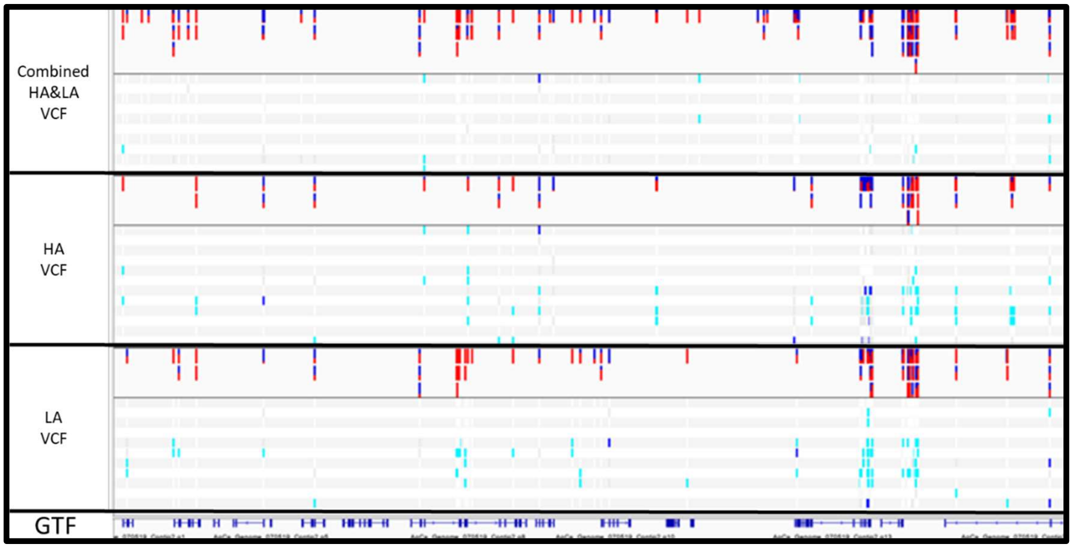


Figure 97: Example of visualisation of SNPs identified with GATK using IGV for Contig2.The GTF row indicates the location of genes identified by Omicsbox in Chapter 5. A combined VCF of all SNPs called in both HA and LA individuals is shown in comparison to VCFs of all SNPs from individuals in the Low Altitude (LA) population combined and from individuals from the High Altitude (HA) population combined. Each VCF shows the three frame shift tracks with SNPs at low occurrence between individuals in dark blue and high occurrence between individuals in red. Light blue indicates SNP occurrence in individuals.

7.4.2. **Phylogeny of HA and LA population tested.**

A phylogenetic tree of individuals in the HA and LA population tested in this chapter and the previous chapter was calculated based on the 1,082,309 SNPs from the combined HA and LA VCF in SNPhylo and visualised in FigTree (Figure 98). Virtually all samples group closely into their respective population as calculated from their SNPs. The LA natives (collected from a separate low altitude native fauna site) is clearly separated from the experimental group of LA individuals but more closely linked than to the HA population. One HA individual (10_1) sits apart as an outlier from its cohort closely linked to a LA individual (11_8).

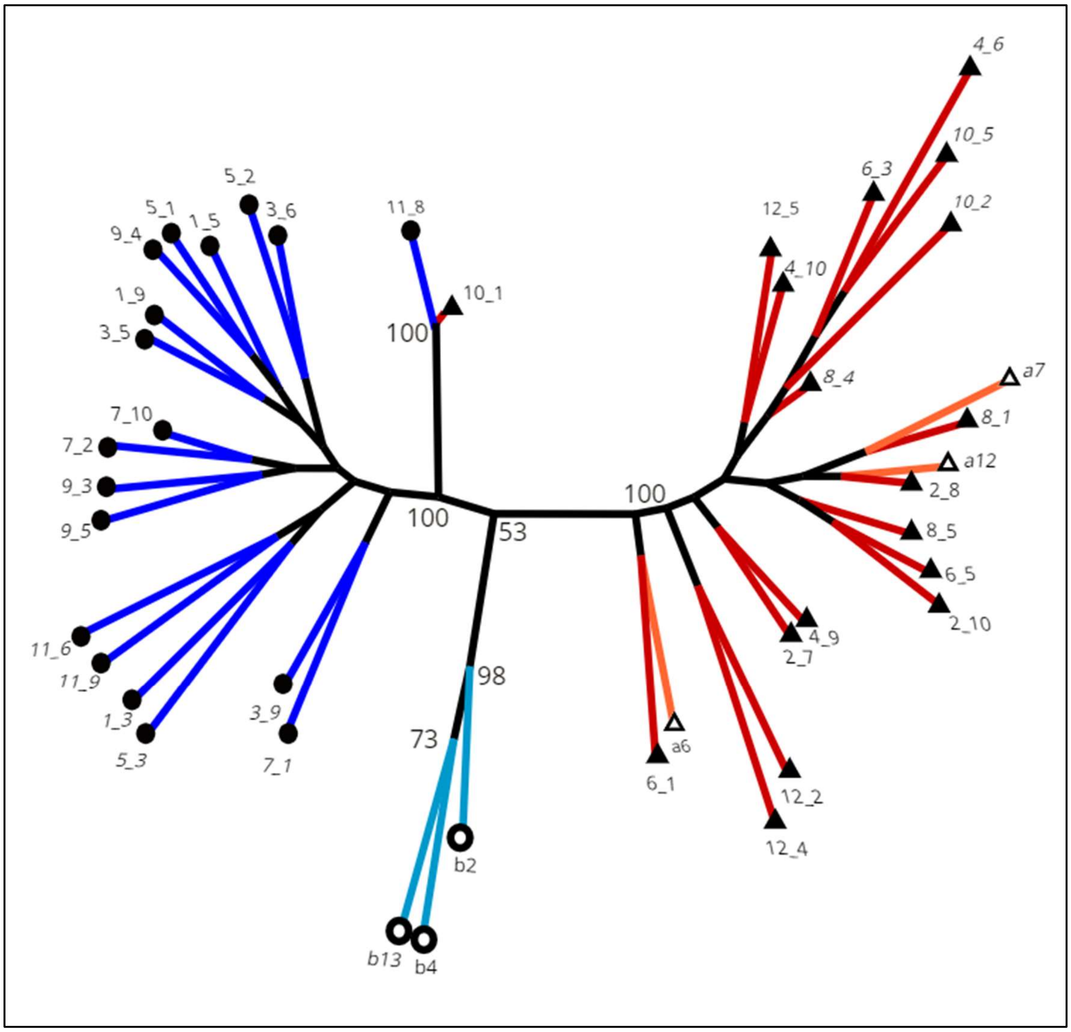


Figure 98: Phylogeny of HA and LA individuals calculated from SNP data in SNPhylo with Maximum likelihood model. Cohorts of HA individuals are prefixed with 'a' denoting natively collected individuals (unfilled circles) or even numbers (2,4,6,8,10,12, (filled circles)). Cohorts of LA individuals are prefixed with 'b' denoting natively collected individuals (unfilled triangles) or odd numbers (1,3,5,7,9,11 (filled triangles)). Individual identifiers are provided following the underscore. Tree was bootrapped with 1000 replicates and only boostrap values greater than 50 are shown. A clear separation between HA, LA experimental and LA native populations is observed with two outlier individuals (10_1 and 11_8).

A principle component analysis was run with TASSEL (V5.2.61) on the SNP data for each individual (Figure 99). PC1 separated HA and LA individuals distinctly while PC2 separated the LA natives, taken from a native fauna site, from the LA experimental worms, taken from a field (Chapter 6.2.1). When Multi-dimensional scaling analysis is performed, these separations become more distinct and more tightly clustered in their populations (Figure 100). The HA outlier observed in Figure 98, is more distinctly separated from the HA population and other groups.

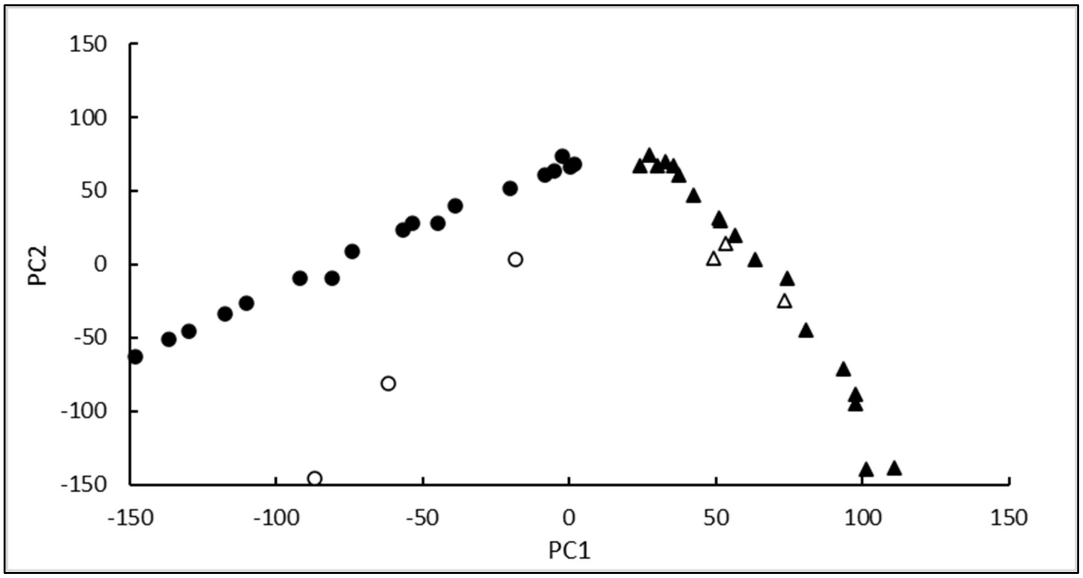


Figure 99: Separation of populations observed in PCA plot. PC1 generates a clear divide of the HA population (Triangles) and the LA population (Circles). HA natives (white triangles), taken from the same site as the experimental worms sit closely with the experimental worms. LA natives (white circles) taken from a low altitude natural fauna site (Chapter 6.2) are seperated slightly from experimental LA worms by PC2.

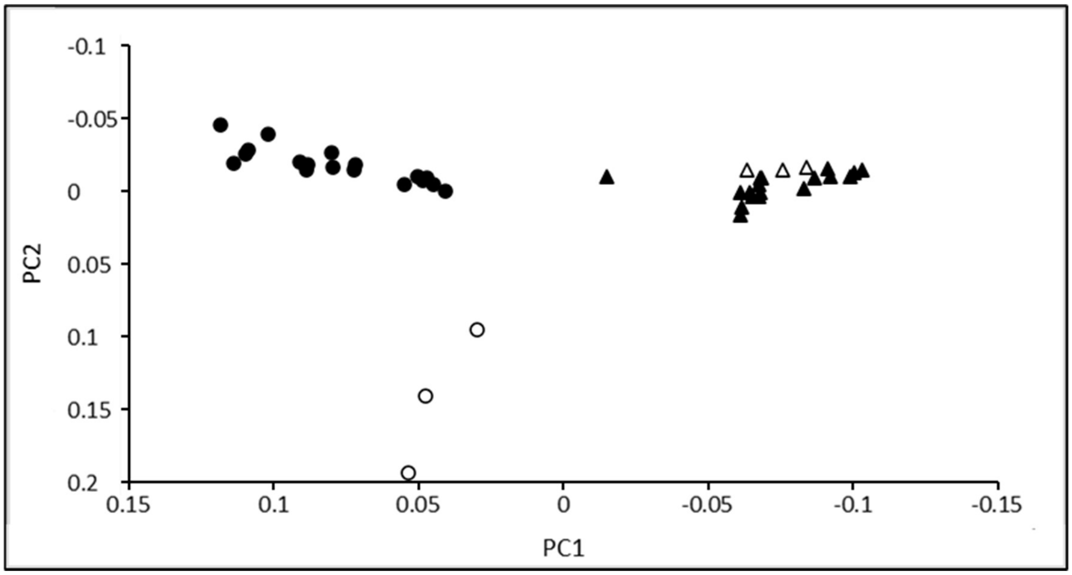


Figure 100: Separation of populations observed in multi-dimensional scaling analysis (MDS). Separation between HA population (Triangles) and LA population Circles are distinct and in tight clusters as separated by PC1. PC2 separates the LA native population (white circles) collected from a low altitude natural fauna site (Chapter 6.2) from the LA experimental worms. A single HA individual outlier sits clearly between HA and LA populations.

7.4.3. **Fst calculation and filtering.**

Fst was calculated between HA and LA populations, with VCF tools with a Minor Allele Count (mac) filter of 8 and filtered to remove uncalculated 'nan's'. As a measure of nuclear diversity, Pi was calculated per base and the -Log10(HA/LA) ratio identified. 370,159 Fst values were calculated (Figure 101). The average Fst value per gene was calculated for the 12,671 annotated genes. After examining the distribution of Fst values, the top 10% of Fst values were filtered out and removed leaving 1047 genes. To identify genes that are under possible positive selection in only the HA population, these genes were further filtered to only include those where HA Pi was half that of LA Pi, leaving 358 genes (i.e. genes showing a very high divergence between the two populations and where the HA population had a substantially lower genetic diversity than the LA population). A final filter to exclude random high Fst and Pi rato values, excluded genes with less than 5 SNPs, leaving 268 candidate genes (Figure 102). Of these candidates 5 contigs contained at least 5 genes under selection indicating a region under positive selection, 4 of which had a concentration of genes in one area indicating potentially higher linkage disequilibrium for genomic selection (Figure 103).

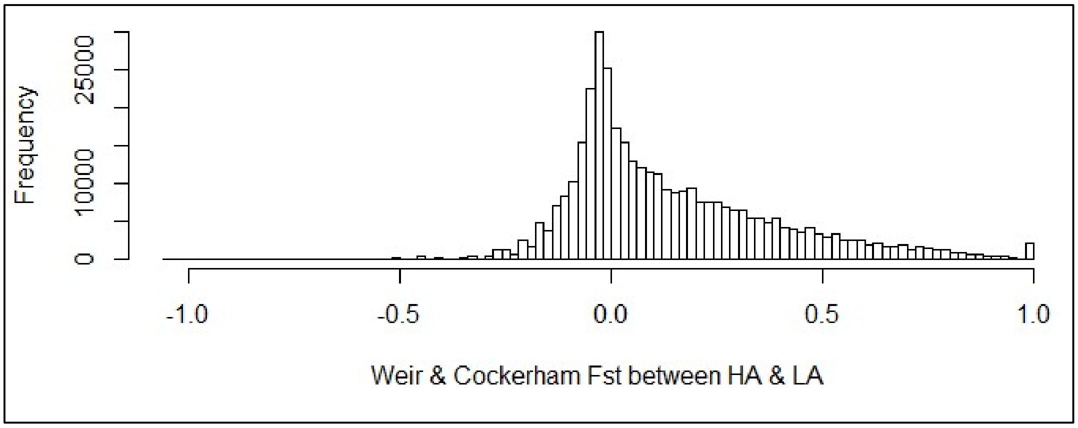


Figure 101: The distribution of Fst values between HA and LA populations calculated per SNP using VCF tools. Values below 0 indicate intra population differences while values closer to 1 indicate more significant inter population differences.

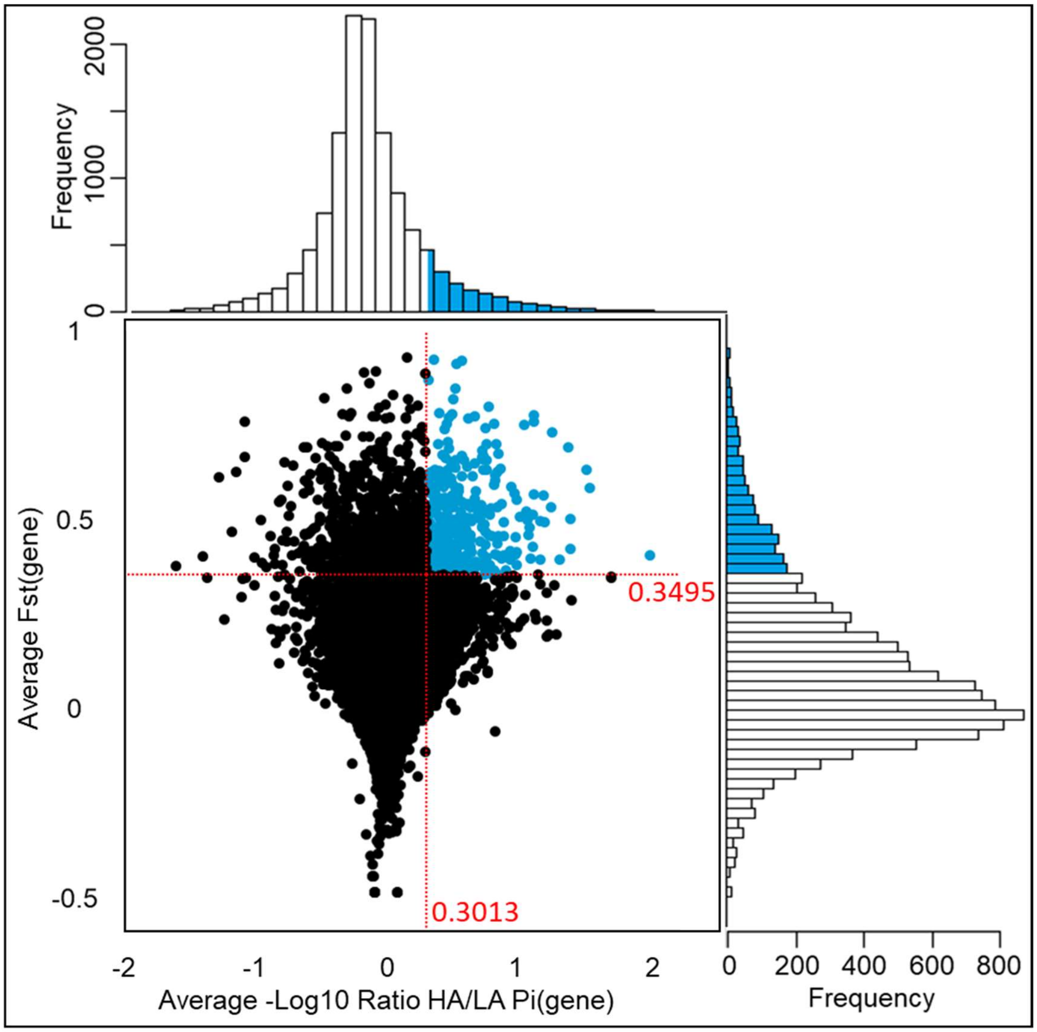


Figure 102: Filtering of candidate genes by Fst value and Pi ratio between HA and LA populations. Distribution of genes are plotted with the Average -Lot10(Pi) value per gene against the Average Fst value per gene. The distribution of values for Fst and Pi are additionally shown as histogram plots. Values passing the Fst and Pi filters are coloured in blue.

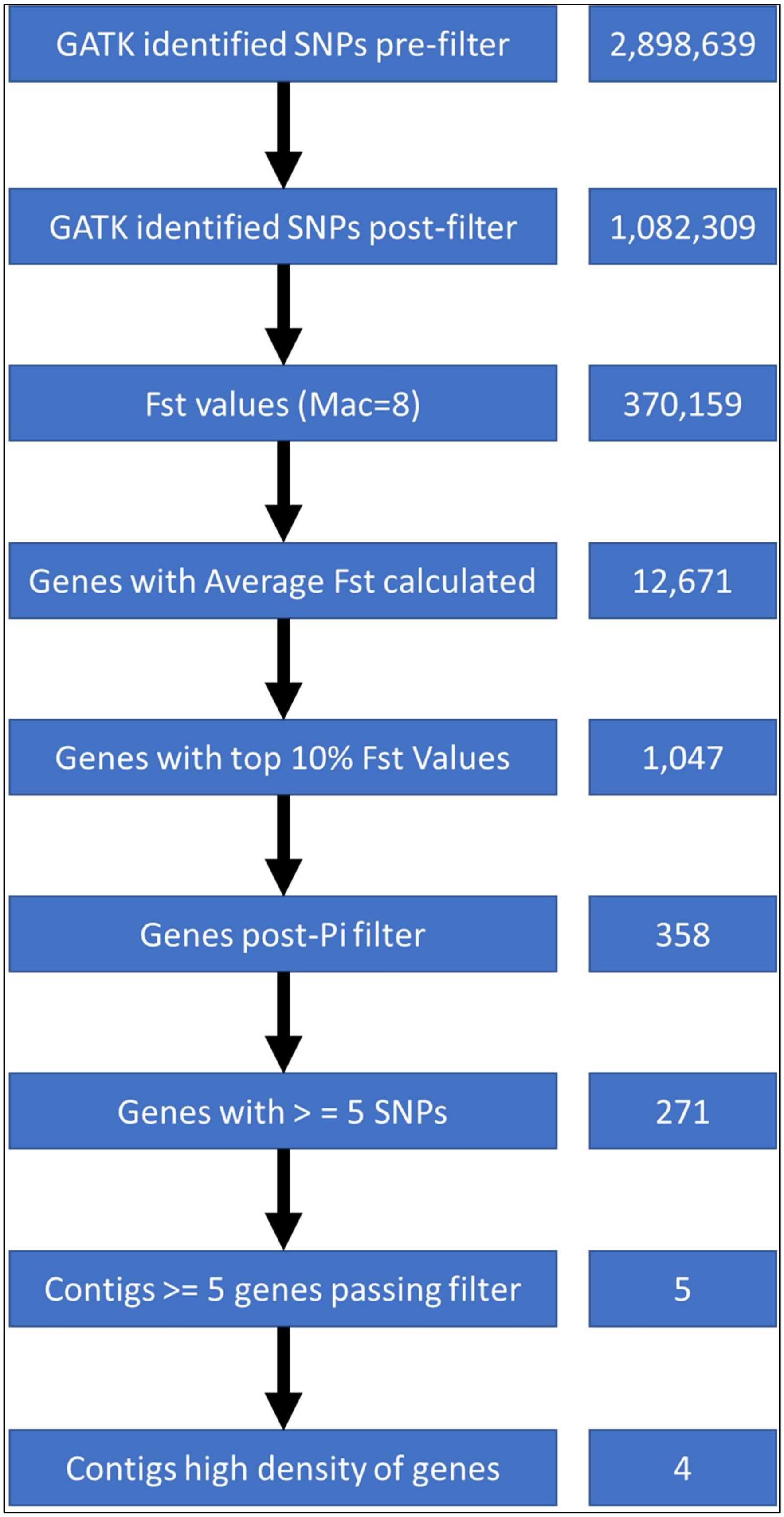

GATK identified SNPs pre-filter
2,898,639
GATK identified SNPs post-filter
1,082,309
Fst values (Mac=8)
370,159
Genes with Average Fst calculated
12,671
Genes with top 10% Fst Values
1,047
Genes post-Pi filter
358
Genes with > = 5 SNPs
271
Contigs >= 5 genes passing filter
5
Contigs high density of genes
4

Figure 103: Identifying linkage disequilibrium, a workflow showing process taken to quality filter SNPs to identify contigs with a high density of genes containing SNPs including associated numbers of narrowing targets of interest.

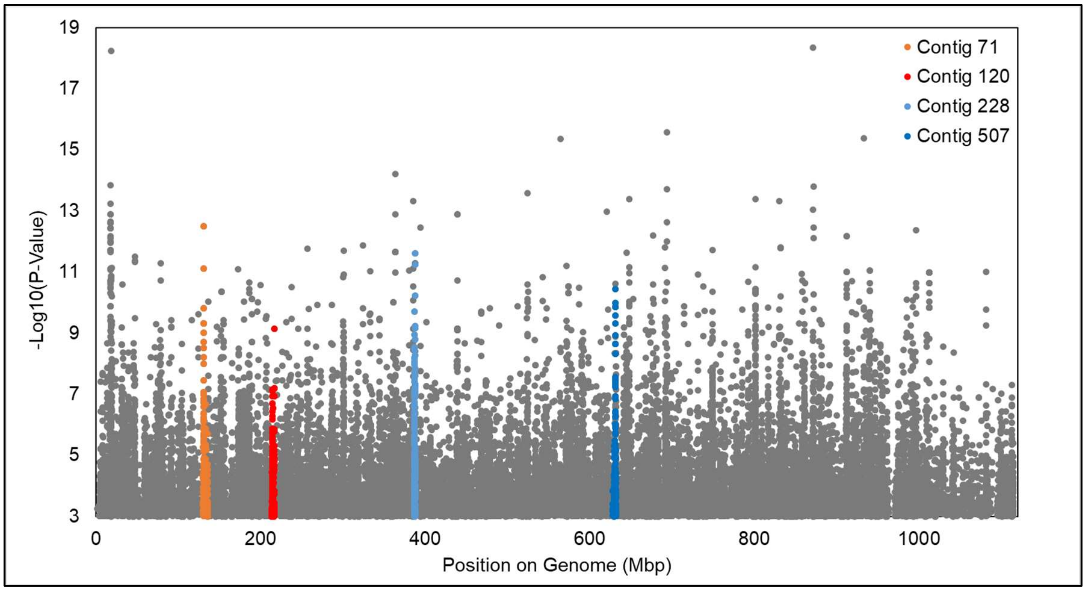


Figure 104: Comparison of Contigs 71 (orange), 120 (red), 228 (light blue) and 507 (dark blue) filtered via Fst and Pi diversity with SNP across the genome. P-values for General Linear Model (GLM) calculated with Tassel V5, to test for association between segregating sites and phenotype for each SNP displayed as a Manhattan plot. P-values higher than 0.001 are not shown.

A General Linear Model was run on the SNP data to assess how association between segregating sites and phenotypes compares across the genome. The P-value for each SNP was plotted across the genome to compare how contigs 71, 120, 228 and 507 compared with other contigs (Figure 104). All four contigs have a high -Log10(P-value), though some other contigs display higher values. These contigs also display high Fst values, but do not indicate positive selection at high altitude as determined through Pi measure of nuclear diversity.

The 4 regions with a high density of genes under selection in the HA population are shown below in Figure 105, Figure 106, Figure 107 and Figure 108 (genes with high Fst passing the high Pi ratio filter are highlighted in red and those with high Fst but with Pi ratio below the filter are highlighted in blue).

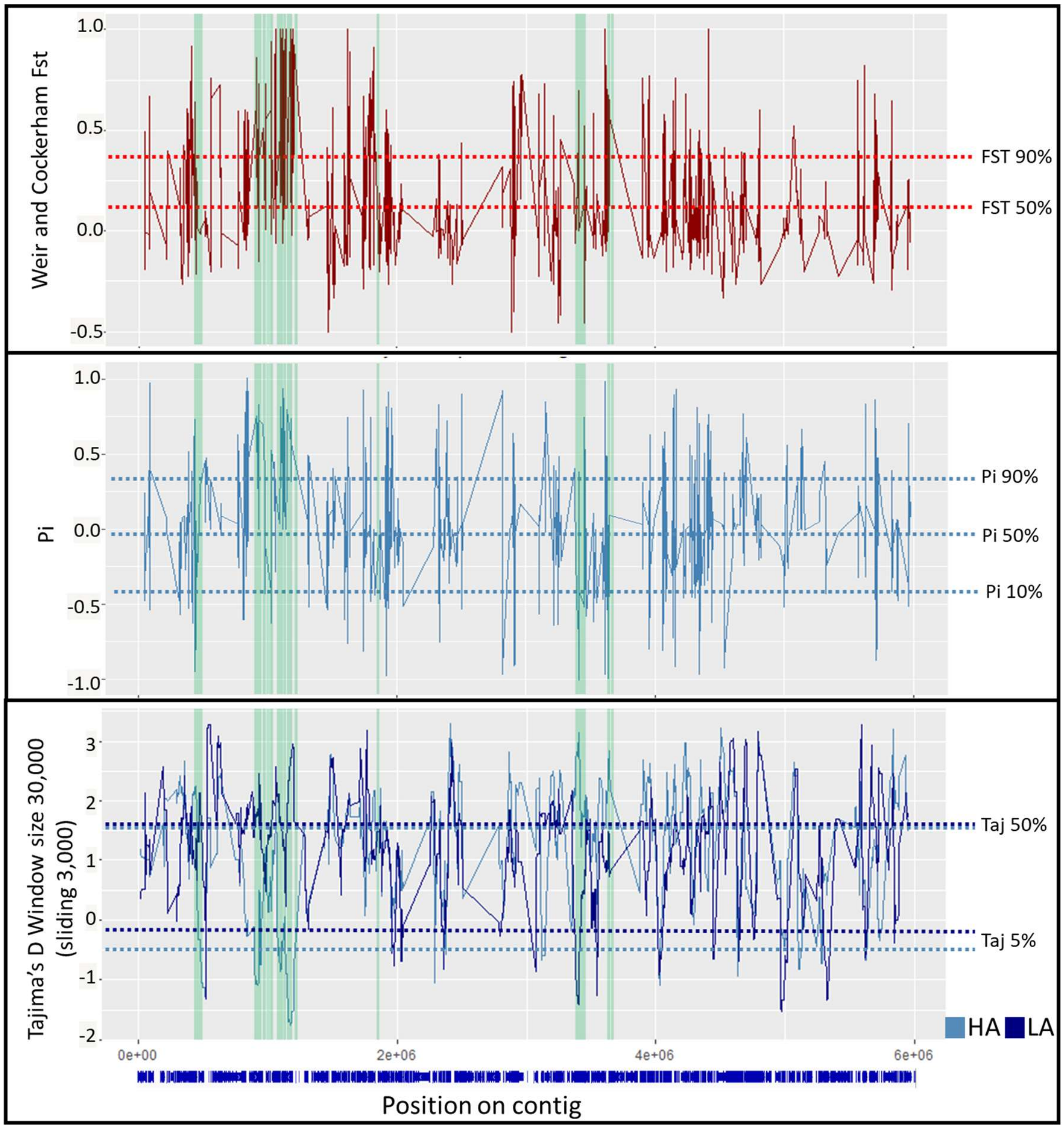


Figure 105: Metrics of selection for Loci in Contig 71 Fst, sliding window of Tajima's D and -Log ratio of Pi measure of nuclear diversity plots showing the density of genes that passed the filter for Fst. Highlighted in green are those that passed the filter for Fst. High density of SNPs indicate strong selection for variation between populations in specific genes.

Contig 71 had a high density of genes with high average Fst, (genes 37, 40, 42, 45, 46, 47, 48 and 51). Of these 8 genes (ZKSCAN7, PUS3, CFAP74, DDX10, ARHGAP45 and GIN1), evidence of selection in the HA population was observed with peaks in values for Tajima's Pi with further evidence for some genes with negative values for Tajima's D.

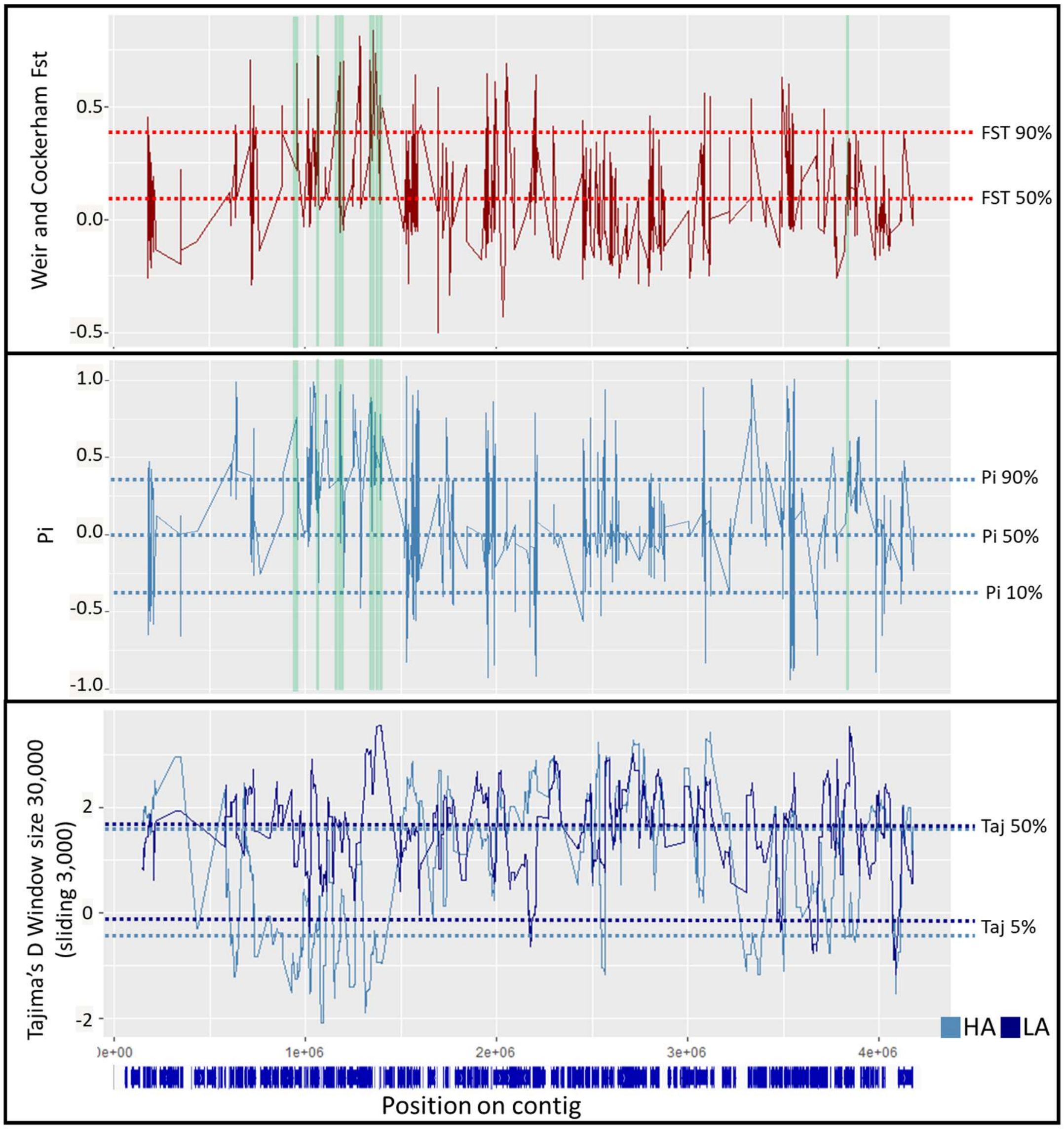


Figure 106: Metrics of selection for Loci in Contig 120 Fst, sliding window of Tajima's D and -Log ratio of Pi measure of nuclear diversity plots showing the density of genes that passed the filter for Fst. Highlighted in green are those that passed the filter for Fst. High density of SNPs indicate strong selection for variation between populations in specific genes.

Contig 120 also had a high density of genes with Fst values in the top 10%, (genes 35, 43, 50, 51, 59, 60 and 61). Of these 7 genes (BICD1, SBSPON, FLNA, TPRG1L, DNAJC11, PLEKHM3 and SMARCAL1), evidence of selection in the HA population was observed with peaks in values for Tajima's Pi with further evidence for some genes with negative values for Tajima's D.

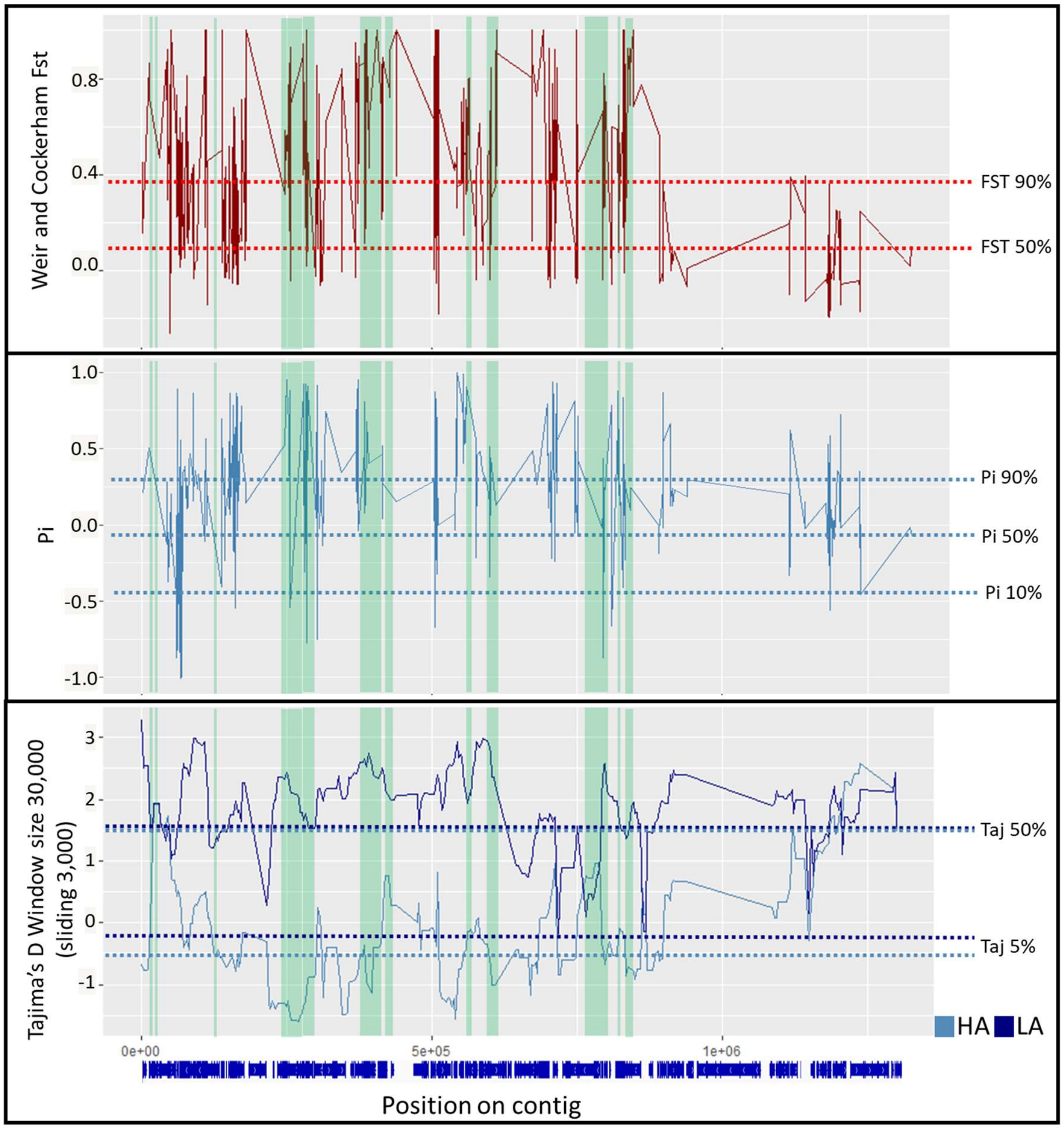


Figure 107: Metrics of selection for Loci in Contig 228 Fst, sliding window of Tajima's D and -Log ratio of Pi measure of nuclear diversity plots showing the density of genes that passed the filter for Fst. Highlighted in green are those that passed the filter for Fst. High density of SNPs indicate strong selection for variation between populations in specific genes.

Contig 228 had a high density of genes with high average Fst, (genes 13, 14, 19, 20, 27 and 29). Of these 6 genes (SOAT1, C17orf53, ATP8B2, KIAA0825, DDX39B and FBXO16), there was evidence of selection in the HA population was observed with peaks in values for Tajima's Pi with further evidence for some genes with negative values for Tajima's D.

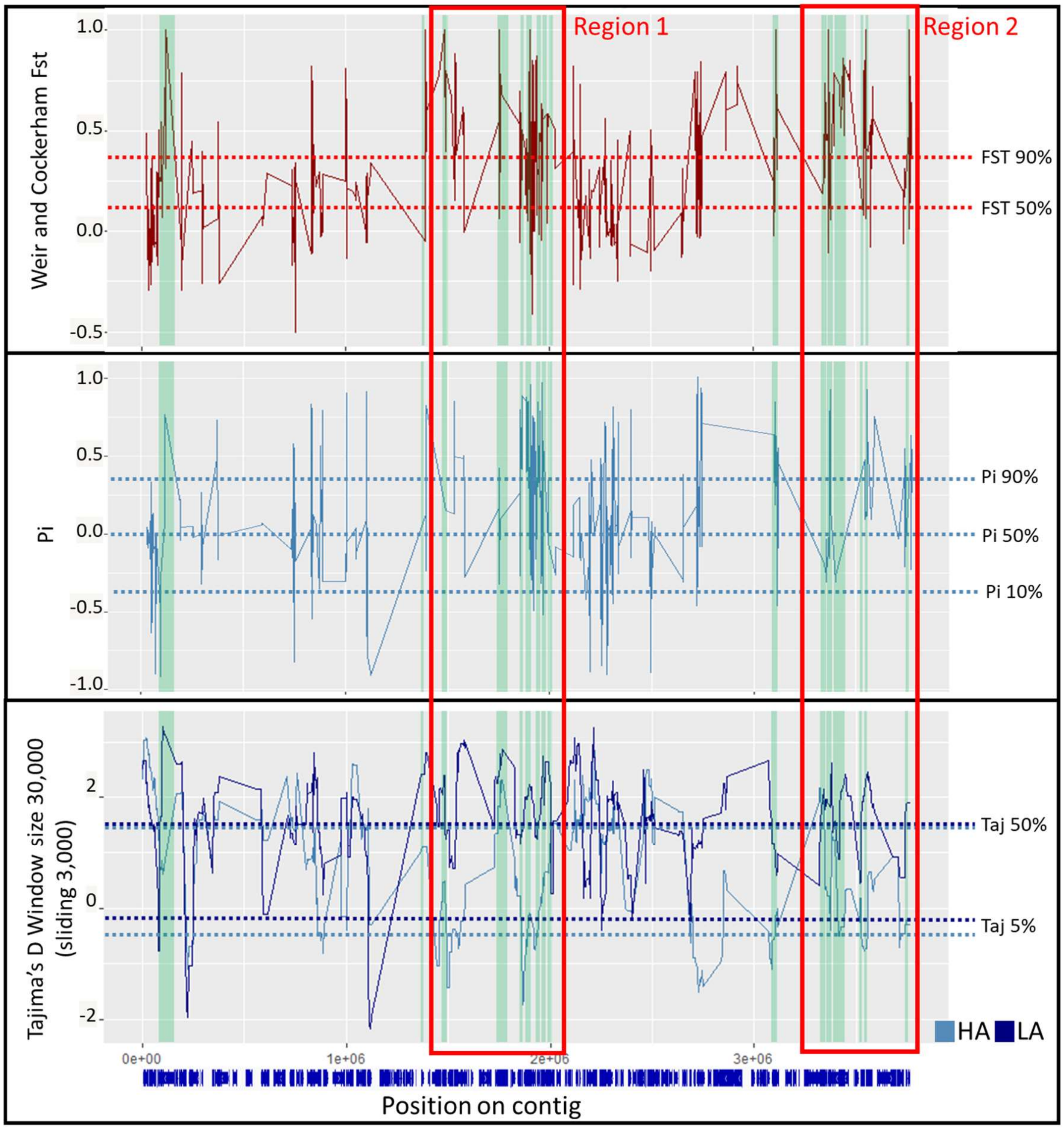


Figure 108: Metrics of selection for Loci in Contig 507 Fst, sliding window of Tajima's D and -Log ratio of Pi measure of nuclear diversity plots showing the density of genes that passed the filter for Fst. Highlighted in green are those that passed the filter for Fst. High density of SNPs indicate strong selection for variation between populations in specific genes.

Contig 507 had two high density regions of genes with high average Fst, (genes 50, 64, 69, 74, 80, 81 and 82 and genes 143, 153, 154, 155, 156, 160, 161 and 170). In the first region of 7 genes, evidence of selection in the HA population was observed with peaks in values for Tajima's Pi with further evidence for some genes with negative values for Tajima's D (NLK, HERC4, AP1B1, DGCR2, NDE1, DPH3 and HEATR3). In the second region, of the 6 genes, all showed evidence of selection in the HA population (DIAPH1, ESRP1, GRHL1, CLINT1, CAPSL and NEDD9).

Functional annotation and REVIGO was performed to classify their function and summarise their GO terms scaled by the number of genes involved in each GO term (Figure 109). The largest interaction with GO terms were: "Anatomical structure development" (9 genes), "Transport" (8 genes), “Cellular component assembly” (8 genes), and “Cellular nitrogen compound metabolic process” (8 genes).

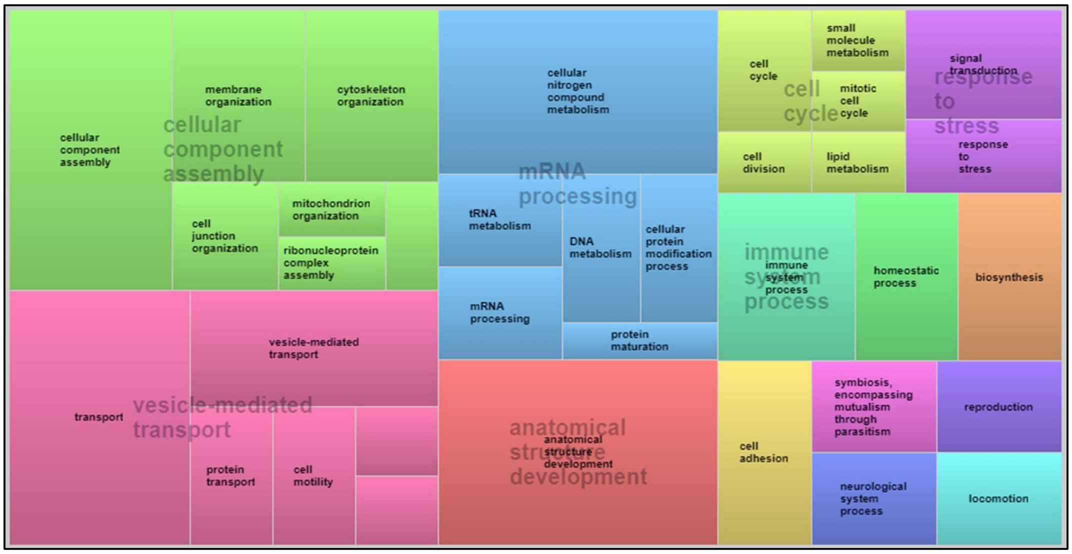


Figure 109: Classification of GO terms into groups via REVIGO for genes under selection in HA population. Major groupings of GO terms include “Cellular component assembly”, “vesicle-mediated transport”, “mRNA processing” and “anatomical structure development”.

There were 4 genes identified within the genes under selective pressure for the HA population that were also identified by differential expression analysis of HA individuals in Chapter 5 (Figure 110). None of the 4 genes showed any identifiable pattern of expression based on temperature or oxygen exposure but GIN1 was under differential expression in 8 expression comparisons and GRHL1 in 7 expression comparisons. The range of biological processes involved with these 4 genes are shown in Figure 110.

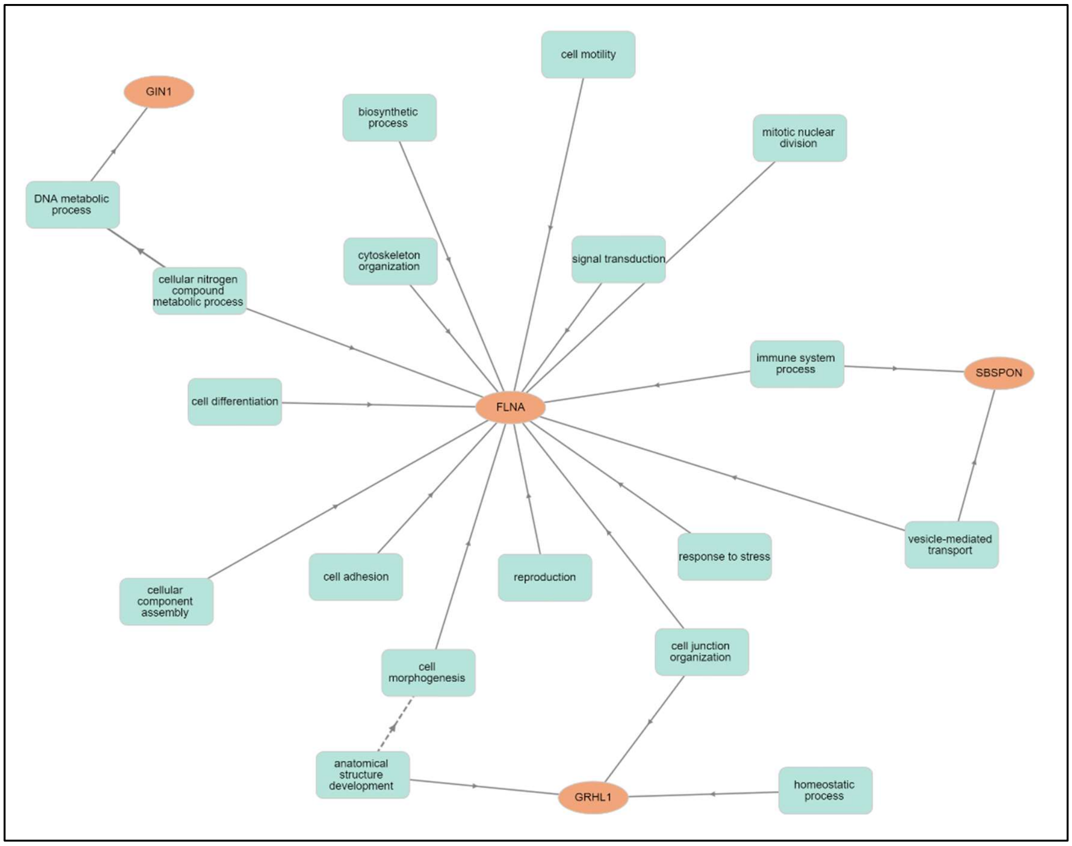


Figure 110: Network of GO terms associated with the 4 genes under selective pressure for the HA population that were also identified by differential expression analysis as identified through GOnet analysis. This indicates biological processes affected by these genes involve several elements of response to environmental stress.

Genes HMGB1 and HMGB2 were identified in Chapter 5 as having a potentially significant role in how the HA population responds to environmental conditions. HMGB1 and HMGB2 are found 3 times each within the genome assembly and Fst values for each occurrence was checked showing only one occurrence of HMGB1 with an Fst value in the top 10% of Fst values (containing 22 SNPs), but Pi analysis indicates there is no selective pressure for one population over another. Links from these two genes (and the closely linked gene TP53) to all the genes under selective pressure for the HA population were identified with STRING-db (Figure 111). There were 5 genes identified: ESRP1, ZNF420, CLINT1, AP1B1, DDX39B, and HERC4 identified, that were from the 4 Contigs that are under higher linkage disequilibrium. Of particular interest is NFKB1 that has a direct interaction with HMGB1 and a suite of other genes (TBK1, QPCT, PRG3, SYK, and TAB1) that also pass the Fst and Pi filters and play important roles in inflammation and the MAPK pathway. Also, of interest is the highly connected interaction between RNF217, SPSB1, KLHL5, HERC4, ARIH2, GLMN, ASB2, and MIB2, which all play a role in cellular protein modification process and protein ubiquitination, linking to HMGB1 and HMGB2 via DDB2 and TP53. SMARCE1 also directly interacts with HMGB1, has a high Fst, and one of the highest ratios for Pi indicating strongly selected for in the HA population. There is also of a cluster of interactions between CLVS1, AP1S2, AP1B1, SNX5, TBC1D5 and CLINT1 (involved in Clatherin coating of Golgi vesicles), and a cluster of interactions between PPF1A1, SYT1, FNBP1L, ARPC4 and WAS (involved in actin filament reorganisation).

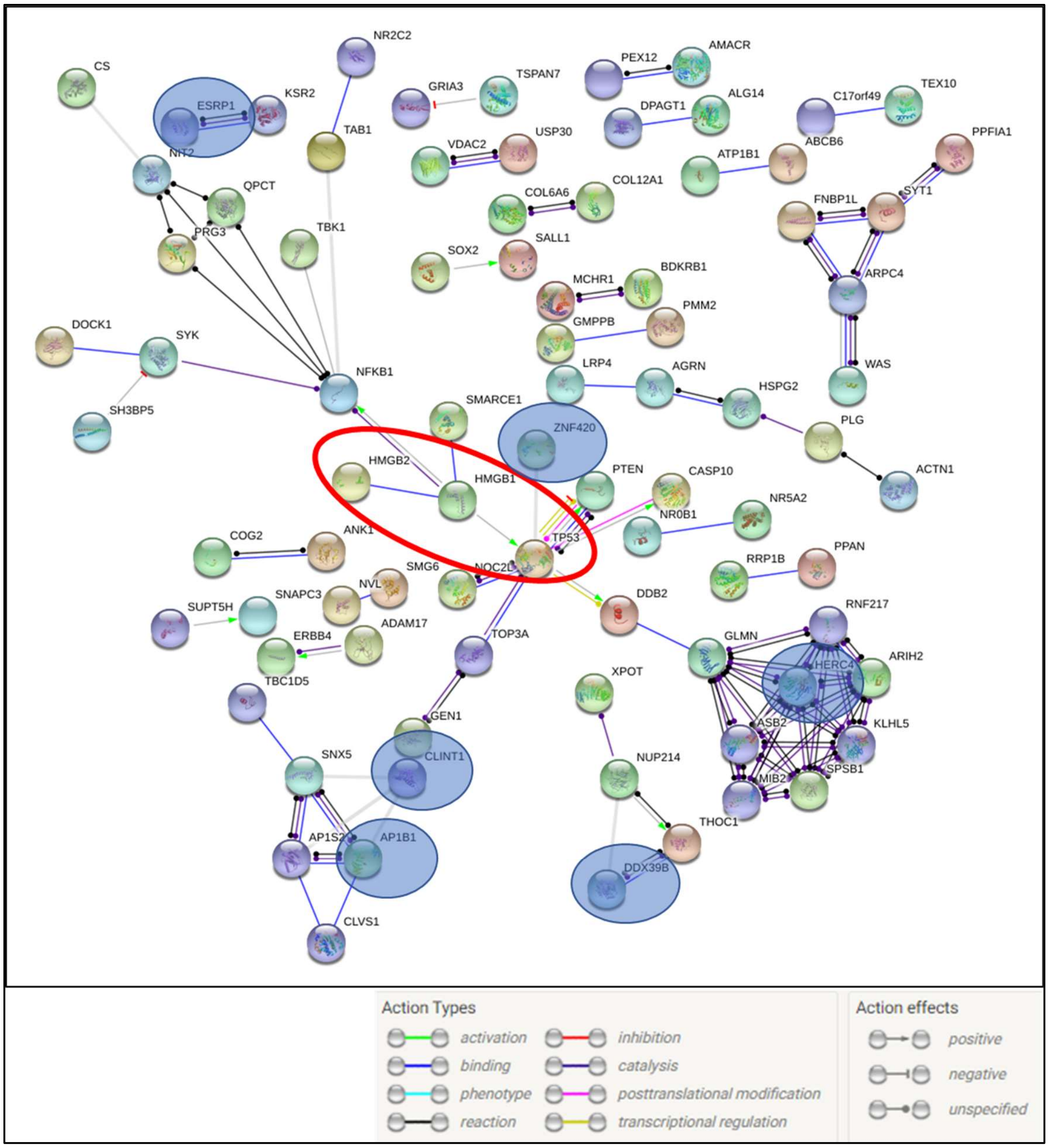


Figure 111: Genes under selective pressure for the HA population that interact with each other with a high confidence score and with HMGB1, HMGB2 and TP53 (circled in red) that are not themselves under selective pressure as identified through STRING-db analysis of genes. Circled in Blue are genes identified in the 4 Contigs under higher linkage disequilibrium. The interconnectedness of genes suggest environmental pressure on biological processes affected by these genes.

7.5. Discussion.

7.5.1. **GATK analysis**

GATK is one of the most used methods for RNAseq SNP calling (See 6.1.3), despite this and recommendations of 'best practices', the tool is still under development and challenging to create a transparent and objective workflow for SNP analysis. Recently the Broad Institute has moved from supporting GATK3 to GATK4 that uses similar but different suite of tools. However, in some cases not all functionality has yet been transferred, and supporting documentation is not yet fully comprehensive. Despite the challenges, it was possible to fully process through all recommended major steps required for SNP identification. This included the development of a 'known-sites' file for recalibration of SNP calling as to date, no known SNPs have been published for *A. caliginosa*. One of the more 'subjective' elements to the pipeline are the filters used for assessing the quality of data incorporated and the subsequent confidence of SNPs called. In general hard filters are recommended to be loose, but where possible, the recommended filter values by GATK were enforced. Variant discovery is significantly improved by numbers of individuals, where variant refinement can make use of machine learning to optimise each call. However this process is still a challenge even with datasets as large as humans with the 100,000 genomes dataset available (GATK-Team 2019). The final output of over 1 million SNPs was a substantial number, equating to 1 SNP/Kb, while the rate within humans is closer to 1 mutation every 30 million bp (Bentley 2000; Dolgin 2009). The rate is higher, though more in line with other invertebrates reported (Zhou et al. 2009). There was a good confirmation of the population structure directly identified from the SNPs called through phylogeny. The use of genome wide SNP assessment appears to be the focus for identifying true taxonomic description of cryptic species where single mitochondrial or nuclear barcodes do not provide the necessary resolution in time or genetic space (Marchán et al. 2020a). The HA population is (bar one individual as confirmed as an outlier in the MDS analysis plot) groups together and the LA population experimental group is closely linked and the LA population from a separate site groups together.

7.5.2. **Fst calculation nuclear diversity calculation and Tajima's D.**

Initially genes displaying the top 5% of average Fst values considered as putative areas under selection. When these regions were plotted, it became clear that in many cases, by relaxing the threshold to include the top 10% contiguous islands could be identified representing specific regions under selection and not widespread false positives. This method has been used similarly in the selectin of signals of selective pressure in chickens effectively (Li et al. 2020). Both Tajima's D and Pi provide a measure for identifying which population, HA, LA or both, the high Fst values could be attributed. Tajima's D could only be calculated on a sliding window basis giving a wider zone of indication, while Pi could be calculated per SNP directly related to Fst. There was a good correlation between Tajima's D and Pi values giving confidence in using Pi as a filter for identifying HA site. The filter for Pi to remove genes passing the Fst filter was set as the Pi value for HA being half that of the Pi value for LA population. This largely produced a clear indication for selection in the HA population, though in some cases the relaxing of this filter slightly allowed 'filling in' of genes of areas where there was already a high density of genes passing the filter (Biswas and Akey 2006).

7.5.3. **Gene annotation and Network association analysis.**

All genes that passed the Fst and Pi filters were annotated for GO term and collated into functional groups with REVIGO. Cellular component assembly, vesicle-mediated transport and mRNA processing formed the bulk of associated GO terms but the "Response to stress" and "Response to symbiosis". These two GO terms are of particular interest because it would suggest that the genes in the HA population that have direct links to a high altitude environmental stress response, are under a selective pressure. "Symbiosis encompassing mutualism through parasitism" term was also identified and could be indicating the interaction between gut flora and the earthworm individual is changing to adapt for a particular suit of bacteria that are beneficial in stressful environmental conditions (Bang et al. 2018). Microbial community symbiosis has a significant influence on an earthworms' interaction with the environment as extensively explored by Dr Pass in his doctoral thesis (Pass 2015). This would be a valuable next jigsaw piece to investigate in the puzzle of high altitude adaptation.

To investigate links between genes passing Fst and Pi filters and differentially expressed genes within the HA expression. The identified genes were not differentially expressed in all analyses and there was no pattern to up or downregulation that might have been influenced by SNPs. While the 4 genes are connected through biological processes, the low number of genes that are both differentially expressed and pass Fst and Pi filters suggests there are no direct effects to genes with differential expression. To assess if SNPs in genes that interact with genes of interest identified in differential expression, STRING-db was used to assess known interactions between genes. Of particular interest was HMBG1 and HMBG2 that were identified in Chapter 5.

Several genes were identified that directly interact with HMGB1; SMARCE1 and NFKB1, the latter interacting with a suite of genes that have important roles in inflammation, immunity, and the MAPK pathway. This is a critical set of genes to see under selective pressure and could have direct implications on the function and sensitivity of these pathways to environmental stimuli (Kang et al. 2014). Moreover, this connection to HMGB1 emphasizes the critical role the gene appears to play in the HA population's adaptation to high altitude environments. It was also of particular interest to see the massive interaction between genes involved in protein modification and ubiquitination. The suit of 8 heavily interacting proteins suggests this pathway is under specific pressure and not occurring by chance.

7.5.4. **Final conclusions.**

This chapter set out to identify any markers of adaption within the HA population that might allow a selective advantage to survival high altitude. SNP identification from RNAseq data was highly successful and allowed a direct comparison to the gene expression analysis from Chapter 6. Areas within large contigs were found to have high-density of genes with SNP's, suggesting linkage disequilibrium, but owing to the nature of fragmented genome assembly, it was hard to identify more than 4 contigs with these high-density areas. It is hoped that in future research a HI-C scaffold of these contigs into chromosomes will allow greater observation of these high-density areas. Network analysis of genes that passed Fst and Pi filters identified a direct link to HMGB1 that was identified in Chapter 6 as having an important role in how the HA population is acclimatized to high altitude and suggests there is an adaptation to the pathways that HMGB1 regulates. Given the almost universal presence of HMGB1 in metazoan, it is therefore concluded that HMGB1 plays a crucial role in adaption and acclimatization to high altitude (Tang et al. 2011). Future works should look to directly investigate this gene and its pathways and look to see if this is observed in other species of earthworm and high altitude species. Although there has been SNP analysis into cattle adaption to heat or altitude (Chan et al. 2010; Edea et al. 2019), that even extends to identification of specific mutations in genes associated with pathways and biological processes involved in respiratory systems and signalling pathways, little investigation has been performed on earthworms until now. Interestingly, SNP analysis within the *Carpetania* genus of earthworms has identified local adaptation drives cryptic speciation (Marchán et al. 2020b). This would indicate that the population at high altitude in Pico is not alone in adaptation driving SNP changes, though perhaps with only a few hundred years in which to acquire SNPs, deep cryptic speciation might still be some way off.

8. Conclusions to altitude adaptation and acclimatisation in *A. caliginosa*.

8.1. Summary of findings.

8.1.1. **Phylogeography.**

Research was conducted at two mountain locations, Les Deux Alpes in the French Alps, and Pico in the Azores. Initial investigations aimed at evaluating the earthworm biodiversity found a variety of species in both locations. In Les Deux Alpes this was a steady high diversity of species at almost all sampling sites, and population densities were relatively low with significant numbers of organisms difficult to acquire. In contrast the transect evaluate on Pico displayed restricted earthworm biodiversity with most sites only yielding *A. caliginosa* and *L. terrestris* at high densities. From this we concluded that the island setting of Pico would be a better source for populations for investigations into altitude adaptation with either *A. caliginosa* or *L. terrestris* being a viable target species. Further research into the nucleotide diversity, at the COI loci, of both species, suggested that the haplotypes of *A. caliginosa* were more genetically homogeneous between populations sourced from high and low altitude context (intra-species Pi diversity = 0.00562 and 0.00302) whilst those of *L. terrestris* showed significant larger inter-species Pi diversity (0.3293 and 0.11748). This was confirmed by performing population dynamics analysis which revealed population bottlenecking for both species in both locations following the last glacial maximum. *A. caliginosa* haplotypes in both sampling sites indicated a more singular ancestor population while *L. terrestris* haplotypes came had a diverse set of ancestor populations reflecting the variety of glacial refugia the neighbouring the Alps. *A. caliginosa* populations from Pico were selected for investigations into adaptation and acclimatisation based on the ease of sampling, and the homogeneity of genetic lineage between high and low altitudinal populations.

8.1.1. **Genomics, Differential gene expression analysis and GWAS SNP analysis.**

A high-quality genome was successfully generated with a high N50 of 1.2 Mbp, with a high estimation of completeness, BUSCO score 93.4%. A high number of genes (42,566) were successfully annotated providing a powerful tool for differential expression and GWAS analysis. High altitude and Low altitude populations from Pico were successfully cultivated for 6 months to normalize conditions before multifactorial experimentation to explore the transcriptional response to the two major altitudinal related parameters of temperature and Oxygen availability. Following exposure to 6 different modelled climatic conditions, worms were harvested and RNAseq performed. Differential expression analysis showed the Low altitude population had more differentially expressed genes than the High altitude population in general when environmental conditions were varied, either alteration in temperature or modification of Oxygen. I hypothesis that this indicates that the High altitude population, therefore, demonstrates greater phenotypic plasticity in response to the high-stress environmental conditions. Intriguingly, HMGB1 and HMGB2, proteins known to regulate environmental stress response, and were upregulated in both populations but have regulate different cellular processes and environmental response. This key difference suggests a core method of acclimatization by the High altitude population. Moreover, following the extended 6 months of normalization of conditions that both populations were exposed to, it suggests that the HA population retains its plasticity, an effect that may well indicate an epigenetic pre-conditioning that enables this population to be more tolerant to rapid changes in environmental conditions. The comparative downregulation of these two genes between cold and warm environmental conditions, suggests that only for the most challenging conditions is this gene stimulated in the High altitude population while in the Low altitude population only at warmer conditions does this pathway activate. The higher relative number of differentially expressed genes observed in the LA may indicate that the Low altitude population is not coping with the conditions as well as the high altitude population. Investigations into genes with higher Fst values found links with the HMGB1 pathways specifically in the High altitude population. This indicates that the High altitude population is not only acclimated to high altitude living, but also that there are markers for adaptation to high altitude even after only a few hundred to a few thousand years of living there.

8.2. The Journey Mountain to gene

8.2.1. **Finding the perfect test case.**

The Alps and Pico provide ideal test cases to study altitudinal adaptation in the temperate zone with rapid transects whilst having contrasting evolutionary histories. Several locations within the Alps were researched for practicality and Les Deux Alpes was finally chosen due to it steep slopes, high grass line and from prior personal knowledge of the area. Pico was known to have a high change in altitude over a very short distance and both Pico and Les Deux Alpes were solidified as research locations. It came of little surprise that Mount Pico was determined to be the optimal location and source of population for the research into high altitude adaptation. During previous research into the Azores and Pico worms were found at high density. Combined with the “Natural Laboratory” nature of the island with indications of only recent introduction (few thousand years maximum), and the close genetic lineage of individuals, the island was a perfect fit for the investigation.

8.2.2. **Solving the genome.**

It has become almost straightforward to solve genomes today. So much so that the Earth BioGenome Project aims to sequence and produce genomes for all eukaryotic species on Earth (Lewin et al. 2018). With current technology of sequencing and assembly, this seems a lesser challenge than the Human Genome project was, which took billions in funding and nearly 10 years to complete with the help of a global team of scientists. For many species, this might be the case, but only in the last few years has technology developed enough to produce a good quality genome of an earthworm. The development of one in this chapter marks a big step forward for the investigative potential of *A. caliginosa*. The species is heavily utilised in monitoring environmental health within the UK amongst other global environmental teams and being able to develop new primers for genes and map further RNAseq experiments will empower their research. Moreover, despite not being able to use genome guided assembly for other species, mapping to *A. caliginosa* due to the enormous genomic differences in chromatin structure, the methodology developed here provides a framework to solving other earthworm genomes, something that has already been employed by other members of our team in solving *E. fetida* and *L. rubellus* (Clade A and B).

8.2.3. **Adaptation, acclimatisation and the epigenome**

As previously defined (in Abbreviations and definitions), this thesis has assessed earthworm's living at high altitude as a function of adaptation through substitutions, insertions and deletions to the earthworm genome versus acclimatisation through either short term or long term changes to gene regulation. Long term changes were that were not a cause of substitutions, insertions or deletions were likely a result of epigenetic changes, possibly through methylation. While it is hard to definitively assess epigenetic changes without investigation of methylation states or the analysis of gene promoter regions, regulatory changes to genes where no SNPs or upstream interactions occur suggest an epigenetic element.

8.2.4. **Moving from adaptation to a condition, to adaptation to change.**

When developing a hypothesis of how high altitude worms cope with high altitude in Pico, there were several key directions of thought. Firstly, the seasonal knowledge of environmental change was known to be a big influencer. Earthworms are required to hatch, grow and reproduce before frost, and the further up the mountain you go, the shorter the time to do this becomes. This is similar to the seasonal time limit worms in floodplains face, but with altitude also comes rapid shifts in weather. This means an even more rapid change in environmental conditions than seen with the seasonal effects, hot to cold, dry to wet, high atmospheric pressure (high oxygen) to low atmospheric pressure. It is this secondary effect that was not anticipated to have as prominent effect as observed. The observed adaptation through a plasticity of response mechanisms and lowered overall response to change that indicated high altitude populations were not just more adapted to coping with low temperature, but more capable of rapid response to change in environmental condition. The identification of key chromatin controlling genes and their interlinked control with genes under selective pressure highlighted a key response mechanism being utilised by the high altitude population that is only being used by low altitude populations when conditions become less overwhelming.

8.3. Limitations of Experimental approach.

8.3.1. **COI and COII.**

One of the earliest challenges crossed during this research was with amplification of the COI molecular barcode. This has been commonly used, and heavily promoted as the "Barcode of Life" (Ratnasingham and Herbert 2007). For some species, promoters for this genes remain very constant and the amplification for barcode sequencing is straightforward. During testing, it became clear that for many of the earthworms tested, this would not be so straightforward. For some individuals of the same species primers worked well, and others poorly. Closer examination identified the most common primer sites were not well conserved. The COII gene has also been used as a molecular barcode and did not have the same issues with amplification that was found in COI amplification attempts. The downside to using COII over the COI molecular barcode however was a reduced library of global species to perform comparisons with. While many within the earthworm community have used COII primers in their analysis, a large number of studies still use the COI primer as part of the "Barcode of Life" initiative. However, despite the limitation of using COII this research was able to position the species within previous globally published data.

8.3.2. **Unknown genes.**

Despite searching Mice, Human, Fly, Drosophila and Yeast databases, not all open reading frames could be annotated. While it is possible this is from errors in the assembly of some of the contigs, (something that will be partly resolved via Hi-C scaffolding), it is also highly likely that there are novel genes present in the earthworm for which there are no orthologues in the public databases yet. Having diverged more than 200 million years ago, there has been enough time for novel genes to develop not seen in other metazoan and that have allowed. This also means that during differential analysis, unknown genes that were differentially expressed, could not be examined. These genes might play an important role, that cannot at this stage be identified. This might not impact the final biological interpretation too much as with the example of the constantly changing number of human genes that dropped substantially as improvements to the genome progressed (Pertea and Salzberg 2010), the 'important' genes have already been identified.

8.3.3. **Worms to RNA.**

Another major difficulty crossed during this research was the production of an RNA sequencing library from our experimental worms. This can be broken down into two sections. The first was quickly removing worms from, washing in distilled water to remove mud, weighing and then dissecting a cross section into Trizol all as fast as could be done. In an idealistic world, this could all be done in a very large temperature and oxygen controlled atmospheric workstation. Due to 'dirty' nature of keeping worms in soil, it was not possible to use institutional cabinets that can do this but work on cell lines, (nor was it financially achievable to acquire one for this purpose alone). Worms were removed in batches and processed with help to speed up the process of dissecting, but even with this help each condition took 30 minutes to process. For some of the most sensitive and high turn-over mRNAs this could have an impact but as all worms went through the same process, the overall impact was as minimised as possible.

The second was the production of a sequencing library from the extracted RNA. For some worms several micrograms of RNA was extracted. However, after an initial attempt at library synthesis, it became clear much of this was bacterial RNA that was removed as part of the library bead clean-up. This initial over estimation lead to poor library production and even on a secondary attempt, the library was not at optimal concentrations. Bioinformatic testing of duplicate reads was able to determine this did not have an impact for differential expression analysis.

8.3.4. **Comparing samples.**

Multi factorial testing is inherently challenging to analyse and interpret. Only two environmental variables were looked at in this study, but this means comparing 6 different conditions to each other and then comparing with another 6 from the different population. With greater resources, more samples for each timepoint could be used to improve the statistical reliability of each condition, but a further improvement could be made to the study by adding time points, prior to testing and after a month to investigate if the long term impacts between populations are just as dramatic as observed in this analysis. It would also better allow for assessing the overall increase or decrease of gene expression from the start and not just between experimental conditions. In all suggestions here, the increased samples required were beyond the current resources associated with this project.

While this research has identified some genes of interest from differential analysis that are backed up by co-expression of associated genes, the raw expression has not been directly backed by qPCR. This is perhaps the biggest experimental limitation to the gene expression analysis. Any future study inspecting genes of interest should however be able to utilise the annotated genome to help develop primers for this validation.

8.3.5. **Filtering SNPs and using RNAseq data only.**

The use of GATK is robust and well used pipeline for the detection of SNPs in RNAseq data. This is commonly used, particularly in medical test species such as humans, mice and rats. For these laboratory models, datasets exist of known SNPs. This helps to denoise background variation and identify SNPs of particular interest. As with everything involved with this project, previous data were not available, and everything was done as *de novo*. This means that the recalibration of SNP calling in GATK required the initial calling and filtering of SNPs before re-processing the SNP identification. This does not necessarily alter the validity of SNPs but can lead to more background 'noise'. This is only something that will improve as more SNP datasets become published. The filters applied for generating a 'known dataset' and on the final VCFs of SNPs were set where possible by those recommended by GATK, though this had to be adjusted manually via an inverse 'grep' command rather than the dedicated script in GATK due to issues between version compatibility. In some cases, extra filters were suggested but could not be applied as not all identified SNPs contained the specified flag.

The main shortcoming to using RNAseq data only is the absence of data from introns where promoters and areas of recombination can occur. This means the picture is not yet complete for the analysis of adaption to high altitude, but can be filled in, in future research.

8.3.6. **Unknown factors, Microbes and Moisture.**

The final unknown in this research project is the impact of the unknowns. While every effort was made to standardise conditions for elements such as soil and moisture within soil, variations within this could have an impact. With increased resources, the number of independent tests of conditions could be increased to minimise this variation, but it remains a consideration in the data presented in this thesis.

As mentioned earlier, one of the biggest unknown impacts is that of microbiomes of earthworms. As researched particularly by Dr Pass, the earthworm microbiome does impact the response of the host worm though linking a highly complex microbiome with an already complex multi environmental factorial study would come with its own challenges and limitations. Statistics could however be performed to attempt to link microbiome patterns with gene expression patterns in a similar way that Principal component analysis is performed.

8.4. Future works and final conclusions.

8.4.1. **Microbes.**

Performing basic microbiome analysis that directly links to the individuals analysed here could be performed. The existing cDNA libraries for individuals could be amplified for bacterial ribosomal 16s to confirm a presence or absence of species within that section of the gut for each individual. However, to perform a more comprehensive assessment of each individual's microbiome, the gastrointestinal tract from each individual would need to be dissected from preserved stocks for each section, e.g. pharynx, oesophagus, crop, gizzard, and intestine. This is because the microbiome changes through each section. From each section 16s amplification could be performed quantitatively for species presence and composition.

8.4.2. **Comparing species and location.**

Having examined altitudinal adaption in one earthworm species and developed a pipeline of analysis, the analysis could be performed in other species such as *L. terrestris* (though this would come with extra challenges, particularly on collection). This could identify if all worms follow a similar pathway of adaption to high altitude as observed in *A. caliginosa*. The analysis could also be performed in other locations. Identification of the same mechanism in multiple species in multiple locations would provide a very solid indication of a particular adaption required for high altitude survival. Additionally, comparison of adaptation in temperate and tropical latitudes would provide an understanding of how season length changes response to altitude.

8.4.3. **Hunting for a mechanism.**

Gene expression analysis and GWAS have highlighted several gene pathways and potential mechanisms associated with survival at high altitude. To further confirm these pathways, future research should look to investigate the mechanisms these directly have. In particular, the use of technologies such as CRISPR could be used to knock out genes such as HMBG1.

8.4.4. **Species refuge at high altitude for Climate change.**

With the increasing impact of extreme weather events resulting from human induced global warming, all elements of life are facing a climate apocalypse, unable to survive the sustained pressures (Deutsch et al. 2008; Goulson 2019). It is thought as general temperatures increase, populations will gradually shift up mountains for cooler temperatures (Storz et al. 2010), however this does not account for the increases in extreme weather events, nor the increased exposure to some of these events. As identified in this research, earthworms found at high altitude were better able to respond to rapid shifts in environmental stressors and therefore more are more likely to better withstand the increasing weather extremes. This therefore suggests that diametrically opposing the low valley refugia that species survived in during glacial maximums, high altitude environments could provide refugia for species during global warming. It is reasonable to investigate these populations as optimal species for seeding new populations in soils that have been purged in extreme weather, and that might be able to withstand future extreme weather in those soils.

8.4.5. **Terraforming and human exploration.**

In recent years, there has been a renewed push for exploration of our solar system including our Moon and Mars. Food production on these explorations has been heavily researched towards aquaponics as a “soil-free” and “clean” method (Kolattukudy et al. 2017). This chemical controlled method of food production has shown effective cultivation of crops, but, has overlooked critical elements of food production and waste management. In particular, little consideration has been afforded to the by-products of cultivation. Element of the crop that cannot be consumed and human waste is still controlled by deep-space jettison or re-entry incineration, which cannot be considered a long term sustainable method. Worms cannot only resolve both these issues, but with a population capable of surviving extreme environments, it could be possible to make use of large spaces not suitable for human survival but that the worms can.

8.5. Final conclusions.

This thesis set out to identify markers of adaptation and acclimatisation to high altitude. Though this story is not yet complete, the research reported here provides a valuable insight and toolset for future analysis into adaption and the life of *A. caliginosa.* The model has proved far more dramatic than initially expected. Given the relatively short amount of time these worms had to adapt and acclimate to high altitude on Pico, to observe more than a few genes altering has been eye-opening and the mass changes indicate just how much there is still to learn of earthworms. The research has always been about extremes, I just didn’t expect it to be as much as this!

9. References.

Abedini-Nassab, R. 2017. Nanotechnology and Nanopore sequencing. *Recent patents on Nanotechnology* 11, pp. 34-41.

Adamczewska, A. M. and Morris, S. 2000. Locomotion, Respiratory Physiology, and Energetics of Amphibious and Terrestrial Crabs. *Physiological and biochemical zoology : PBZ* 73(6), doi: 10.1086/318099

Adetunji, M. O. et al. 2019. Variant analysis pipeline for accurate detection of genomic variants from transcriptome sequencing data.*PLoS One*. Vol. 14.

Ali, O. A. et al. 2016. RAD capture (Rapture): Flexible and efficient sequence-based genotyping. *Genetics* 202(2), pp. 389-400.

Amaral, A. et al. 2006. Bioavailablity and cellular effects of metals on *Lumbricus terrestris* inhabiting volcanic soils. *Environmental Pollution* 142(1), pp. 103-108.

Anderson, C. et al. 2017a. Genetic Variation in Populations of the Earthworm, Lumbricus Rubellus, Across Contaminated Mine Sites. *BMC genetics* 18(1), doi: 10.1186/s12863-017-0557-8

Anderson, F. E. et al. 2017b. Phylogenomic analyses of Crassiclitellata support major Northern and Southern Hemisphere clades and a Pangaean origin for earthworms. *BMC Evolutionary Biology* 17(123), pp. 1-18.

Anderson, N. C. et al. 1983. Nitrogen and cation mobilisation by soil fauna feeding on leaf litter and soil organic matter from deciduous woodlands. *Soil biology and biochemistry* 15, pp. 463-467.

Anderson, T. M. et al. 2009. Molecular and evolutionary history of melanism in North American Gray Wolves. *Science* 323(5919), pp. 1339-1343.

Andre, J. et al. 2010. Molecular genetic differentiation in earthworms inhabiting a heterogenous PB-polluted landscape. *Environmental Pollution* 158(3), pp. 883-890.

Ari, S. and Arikan, M. 2016. *Next-generation sequencing: Advantages, disadvantages, and future*. Switzerland: Springer.

Ari, S. and M., A. 2016. *Next-generation sequencing: Advantages, disadvantages, and future*. Switzerland: Springer.

Ashe, T. 1813. *History of the Azores, or Western Islands*. London: Sherwood, Neely, and Jones, Paternoster Row.

Aslund, M. W. et al. 2012. Earthworm metabolomic responses after exposure to aged PCB contaminated soils. *Ecotoxicology* 21(7),

Azevedo, E. B. et al. 1998. Simulation of local Climate in islands environments using a GIS interated model. *Emerging Technologies for Sustainable land Use and Water Management* 99, pp. 433-436.

Baker, G. H. et al. 1992. The life history and abundance of the introduced earthworms *Aporrectodea trapezoides* and *Aporrectodea caliginosa* in pasture soils in the Mount Lofty Range, South Australia. *Australian Journal of Ecology* 17(2), pp. 177-188.

Baker, G. H. et al. 2006. *Introduced earthworms in agricultureal and reclaimed land: their ecology and influences on soil properties, plant production and other soil biota*. Springer.

Ballouz, S. et al. 2018. The fractured landscape of RNA-seq alignment: the default in our STARs. *Nucleic Acids Research* 46(10), pp. 5125-5138.

Bandelt, H. et al. 1999. Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution* 16(1), pp. 37-48.

Bang, C. et al. 2018. Metaorganisms in extreme environments: do microbes play a role in organismal adaptation? *Zoology* 127, pp. 1-19.

Barley, K. P. 1959. The influence of earthworms on soil fertility. II. Consumption of soil and organic matter by the earthworm *Allobophora caliginosa*. *Australian Journal of Agricultural Research* 10, pp. 179-185.

Bauder, A. et al. 2017. *The Swiss Glaciers, 20013/14 and 2014/15 Glaciological Report*. Switzerland: (SCNAT), C.C.E.o.t.S.A.o.S.

Bayega, A. et al. 2020. De novo assembly of the olive fruit fly (*Bactrocera oleae*) genome with linked-reads and long-read technologies minimizes gaps and provides exceptional Y chromosome assembly. *BMC Genomics* 21, p. 259.

Benedetto, G. et al. 2000. Mitochondrial DNA sequences in Prehistoric human remains from the Alps. *European Journal of Human Genetics* 8(1), pp. 669-677.

Bennett, B. C. and Prance, G. T. 2000. Introduced plants in the indigenous pharmacopoeia of Northern South America. *Economic Botany* 54(1), pp. 90-102.

Bentley, D. R. 2000. The Human Genome Project--an overview. *Med Res Rev* 20(3), pp. 189-196. doi: 10.1002/(sici)1098-1128(200005)20:3<189::aid-med2>3.0.co;2-#

Benzonana, L. L. et al. 2013. Isoflurane, a commonly used volatile anesthetic, enhances renal cancer growth and malignant potential via the hypoxia-inducible factor cellular signaling pathway in vitro. *Anesthesiology* 119(3), pp. 593-605. doi: 10.1097/ALN.0b013e31829e47fd

Bernier, N. and Ponge, J. F. 1998. L. distribution within an experimental humus mosaic in a mountain spruce forest. *Biology and Fertility of soils* 28, pp. 81-86.

Beron, P. 2018. High-Altitude Arachnida (Partly from Beron 2008, Updated in Beron 2016) | SpringerLink. *Zoogeography of Arachnida* 94(1), pp. 853-891. doi: 10.1007/978-3-319-74418-6_10

Bhambri, A. et al. 2017. Insights into regeneration from the genome, transcriptome and metagenome analysis of Eisenia fetida. *BioRxiv* 180612(1), doi: 10.1101/180612

Bhat, S. A. et al. 2018. Earthworms as organic waste managers and biofertilizer producers. *Waste and Biomass Valorization* 9(7), pp. 1073-1086.

Bianchi, M. E. and Agresti, A. 2005. HMG proteins: dynamic players in gene regulation and differentiation. *Curr Opin Genet Dev* 15(5), pp. 496-506. doi: 10.1016/j.gde.2005.08.007

Bienert, F. et al. 2012. Tracking earthworm communities from soil DNA. *Molecular Ecology* 21(1), pp. 2017-2030

Bioinformatics, B. 2019. OmicsBox-Bioinformatics Made Easy. https://www.biobam.com/omicsbox: BioBam Bioinformatics.

Biswas, S. and Akey, J. M. 2006. Genomic insights into positive selection. *Trends in Genetics* 22(8),

Blakemore, R. J. et al. 2010. Neotypification of Drawida hattamimizu Hatai, 1930 (Annelida, Oligochaeta, Megadrili, Moniligastridae) as a model linking mtDNA (COI) sequences to an earthworm type, with a response to the 'Can of Worms' theory of cryptic species. *ZooKeys* 41(1), pp. 1-29.

Blanchart, E. et al. 2004. Effects of tropical endogeic earthworms on soil erosion. *Agriculture, Ecosystems and Environment* 104(2), pp. 303-315.

Blouin, M. et al. 2013. A review of earthworm impact on soil function and ecosystem services. *European Journal of Soil Science* 64(2), pp. 161-182.

Bock, W. J. 2020. The Definition and Recognition of Biological Adaptation. *Integrative and Comparative Biology* 20(1), pp. 217-227. doi: 10.1093/icb/20.1.217

Bohlen, P. J. and Lal, R. 2017. *Encyclopedia of soil science: Earthworms*. London: Taylor & Francis Group.

Bolger, A. M. et al. 2014. Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics* 30(15), pp. 2114-2120.

Bonani, G. et al. 1994. AMS 14 C Age Determination of Tissue, Bone and Grass Samples from the Ötzal Ice Man. *Radiocarbon* 36(2), pp. 247-250.

Borges, P. A. V. et al. 2010. *A list of the terrestrial and marine biota from the Azores*. 1 ed. Portugal: Principia, Cascais.

Borges, P. A. V. and Wunderlich, J. 2008. Spider biodiversity patterns and their conservation in the Azorean archipelago, with descriptions of new species. *Systematics and Biodiversity* 6(2), pp. 249-282.

Bottinelli, N. et al. 2010. Earthworms accelerate soil porosity turnover under watering conditions. *Geoderma* 156(1-2), pp. 43-47.

Bradbury, P. J. et al. 2007. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23(19), pp. 2633-2635. doi: 10.1093/bioinformatics/btm308

Brouard, J. S. et al. 2019. The GATK joint genotyping workflow is appropriate for calling variants in RNA-seq experiments.*J Anim Sci Biotechnol*. Vol. 10.

Bruick, R. K. 2003. Oxygen sensing in the hypoxic response pathway: regulation of the hypoxia-inducible transcription factor. *Genes Dev* 17(21), pp. 2614-2623. doi: 10.1101/gad.1145503

Butt, K. R. 2010. The earthworm inoculation unit technique: Development and use in soil improvement over two decades. *Biology of Earthworms* 24(1),

Butt, K. R. 2011. Life cycle studies of the earthworm Lumbricus friendi (Cognetti, 1904). *Pedobiologia* 54(1), pp. S27-S29.

Callier, V. and Nijhout, H. F. 2011. Control of body size by oxygen supply reveals size-dependent and size-independent mechanisms of molting and metamorphosis. *PNAS* 108(35), pp. 14664-14669. doi: 10.1073/pnas.1106556108

Calnan, D. R. and Brunet, A. 2008. The FoxO code. *Oncogene* 27(16), pp. 2276-2288. doi: 10.1038/onc.2008.21

Cardoso, G. et al. 2017. Earthworm populations in an altitudinal gradient (1000-1850m) of the Coastal Atlantic Rainforest, Brazil. *International Colloquium on Soil Zoology*,

Carine, M. A. and Schaefer, H. 2010. The Azores diversity enigma: why are there so few Azorean endemic flowering plants and why are they so widespread? *Journal of Biogeography* 37(1), pp. 77-89.

Castañeda, L. E. et al. 2019. Evolutionary Potential of Thermal Preference and Heat Tolerance in Drosophila Subobscura. *Journal of evolutionary biology* 32(8), doi: 10.1111/jeb.13483

Chaisson, M. J. P. et al. 2015. Genetic variation and the de novo assembly of human genomes. *Nature Reviews Genetics* 16(11), pp. 627-640. doi: doi:10.1038/nrg3933

Chaladze, G. et al. 2014. A spider diversity model for the Caucasus Ecoregion. *Journal of Insect Conservation* 18(3), pp. 407-416. doi: doi:10.1007/s10841-014-9649-1

Chan, E. et al. 2010. The evolution of tropical adaptation: comparing taurine and zebu cattle - Chan - 2010 - Animal Genetics - Wiley Online Library. *Animal Genetics* 41, pp. 467-477. doi: 10.1111/j.1365-2052.2010.02053.x

Chang, C. and James, S. 2011. A critique of earthworm molecular phylogenetics. *Pedobiologia* 54, pp. S3-S9.

Chapin, F. S. et al. 2000. Consequences of changing biodiversity. *Nature* 405(1), pp. 234-242.

Chauhan, D. S. et al. 2011. *Community-based Biodiversity Conservation in the Himalayas*. India: The Energy and Resources Institute (TERI).

Chenuil, A. et al. 2019. *Problems and Questions Posed by Cryptic Species. A Framework to Guide Future Studies | SpringerLink*. New York: SpringerLink.

Cheviron, Z. A. and Brumfield, R. T. 2012. Genomic insights into adaption to high-altitude environments. *Heredity* 108, pp. 354-361.

Clewing, C. et al. 2016. A Complex System of Glacial Sub-Refugia Drives Endemic Freshwater Biodiversity on the Tibetan Plateau. *PLoS One* 11(8), p. e0160286. doi: 10.1371/journal.pone.0160286

Cortez, J. et al. 2000. Effect of climate, soil type and earthworm activity on nitrogen transfer from a nitrogen-15-labelled decomposing material under field conditions. *Biology and fertility of soils* 30(318-327),

Cosin, D. J. et al. 2010. Reproduction of Earthworms: Sexual Selection and Parthenogenesis | SpringerLink. *Biology of Earthworms* 24, pp. 69-86. doi: 10.1007/978-3-642-14636-7_5

Costanzo, J. P. and Lee, R. E. J. 2013. Avoidance and tolerance of freezing in ectothermic vertebrates. *Journal of Experimental Biology* 216(1), pp. 1961-1967.

Cunha, L. et al. 2011. Morphometry of the epidermis of an invasive megascoelecid earthworm (Amynthas gracilis, Kinberg 1867) inhabiting actively volcanic soils in the Azores archipelago. *Ecotoxicology and Environmental Safety* 74(1), pp. 25-32.

Cunha, L. et al. 2014. Living on a volcano's edge: genetic isolation of an extremophile terrestrial metazoan. *Heredity* 112, pp. 132-142.

Curry, J. P. 1976. Some effects of animal manures on earthworms in grassland. *Pedobiologica* 16, pp. 425-438.

Curry, J. P. and Schmidt, O. 2007. The feeding ecology of earthworms - A review. *Pedobiologia* 50(6), pp. 463-477.

Dahlhoff, E. P. and Rank, N. E. 2000. Functional and physiological consequences of genetic variation at phosphoglucose isomerase: Heat shock protein expression is related to enzyme genotype in a montane beetle. *PNAS* 97(18), pp. 10056-10061.

DePristo, M. A. et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43(5), pp. 491-498. doi: 10.1038/ng.806

Deutsch, C. A. et al. 2008. Impacts of Climate Warming on Terrestrial Ectotherms Across Latitude. *Proceedings of the National Academy of Sciences of the United States of America* 105(18), doi: 10.1073/pnas.0709472105

Doanes, A. M. et al. 1999. VEGF stimulates MAPK through a pathway that is unique for receptor tyrosine kinases. *Biochem Biophys Res Commun* 255(2), pp. 545-548. doi: 10.1006/bbrc.1999.0227

Dobin, A. et al. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29(1), pp. 15-21.

Dobson, R. M. and Satchell, J. E. 2020. Eophila oculata at Verulamium: a Roman Earthworm Population? *Nature* 177(4513), pp. 796-797. doi: doi:10.1038/177796a0

Dodge, T. A. 1994. *Hannibal: A History of the Art of War Among the Carthaginians and Romans Down to the Battle of Pydna, 168 BC, with a Detailed Account of the Second Punic War.* 1 ed. London: Greenhill books.

Dolgin, E. 2009. Human mutation rate revealed. *Nature* 864, doi: doi:10.1038/news.2009.864

Dominguez, J. et al. 2015. Underground evolution: New roots for the old tree of lumbricidd earthworms. . *Molecular Phylogenetics and Evolution* 83(1), pp. 7-19.

Dominguez, J. et al. 2010. *Vermicomposting: Earthworms enhance the work of microbes*. Berlin, Heidelberg: Springer.

Dong, G. et al. 2017. FOXO1 Regulates Bacteria-Induced Neutrophil Activity. *Front Immunol* 8, doi: 10.3389/fimmu.2017.01088

Dorant, Y. et al. 2019. Comparing Pool-seq, Rapture, and GBS genotyping for inferring weak population structure: The American lobster (Homarus americanus) as a case study. *Ecol Evol* 9(11), pp. 6606-6623. doi: 10.1002/ece3.5240

Dowle et al. 2013. Molecular evolution and the latitudinal biodiversity gradient. *Heredity* 110, pp. 501-510.

Drummond, A. J. et al. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology And Evolution* 29(1), pp. 1969-1973.

Duhour, A. et al. 2009. Response of earthworm communities to soil disturbance: Fractal dimension of soil and species' rank-abundance curves. *Applied Soil Ecology* 43(1), pp. 83-88.

Duman, J. G. 2015. Animal ice-binding (antifreeze) proteins and glycolipids: an overview with emphasis on physiological function. *Journal of Experimental Biology* 218(1), pp. 1846-1855.

Dzangaliev, A. D. and Belousova, N. K. 1969. Earthworm populations in irrigated orchards under various soil treatments. *Pedobiologia* 9, pp. 103-105.

Edea, Z. et al. 2019. Genomic signatures of high-altitude adaptation in Ethiopian sheep populations. *Genes & Genomics* 41(8), pp. 973-981. doi: doi:10.1007/s13258-019-00820-y

Edmunds, P. J. and Gates, R. D. 2020. Acclimatization in tropical reef corals. *Marine Ecology Progress Series* 361(361), pp. 307-310. doi: 10.3354/meps07556

Edwards, C. A. et al. 2010. *Earthworms, Organic Wastes, and Environmental Management*. 1 ed. Boca Raton: CRC Press.

Edwards, C. A. and Bohlen, P. J. 1977. *Biology and ecology of earthworms*. London: Chapman & Hall.

Edwards, C. A. and Bohlen, P. J. 1996. *Biology and ecology of earthworms*. London: Chapman & Hall.

Edwards, C. A. and Heath, G. W. 1963. *The role of soil animals in breakdown of leaf material*. Amsterdam, Holland: J. van der Drift (Eds.).

Eijsackers, H. 2010. Earthworms as colonisers: Primary colonisation of contaminated land, and sediment and soil waste deposits. *Science of the Total Environment* 408(8), pp. 1759-1769.

Elmer, J. and Palmer, A. F. 2012. Biophysical properties of *Lumbrius terrestris* erythrocuorin and its potential use as a red blood cell substitute. *Journal of Functional Biomaterials* 3(1), pp. 49-60.

Elmer, J. et al. 2012. Hypervolemic infustion of *Lumbricus terrestris* Erythrocuorin purified by tangential flow filtration. *Transfusion* 52(8), pp. 1729-1740.

Everatt, M. J. et al. 2015. Responses of invertebrates to temperature and water stress: A polar perspective. *Journal of Thermal Biology* 54(1), pp. 118-132.

Fernandes, J. P. et al. 2014. Applying an integrated landscape characterization and evaluation tool to small

islands (Pico, Azores, Portugal). *Journal of Integrated Coastal Zone Management* 14(2), pp. 243-266.

Fernandez, R. et al. 2012. Adding complexity to the complex: New insights into the phylogeny, diversification and origin of parthenogenesis in the Aporrectodea caliginosa species complex (Oligochaeta, Lumbricidae). *Molecular Phylogenetics and Evolution.* 64(2), pp. 368-379.

Fielman, K. T. and Marsh, A. G. 2005. Genome complexity and repetitive DNA in metazoans from extreme marine environments. *Gene* 362(5), pp. 98-108.

Folmer, O. et al. 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit 1 from diverse metazoan invertebrates. *Molecular Marine biology and biotechnology* 3(5), pp. 294-299.

Forey, E. et al. 2011. Importance of earthworm-seed interactions for the composition and structure of plant communities: A review. *Acta Oecologica* 37(6), pp. 594-603.

Fox, C. A. et al. 2017. Earthworm population dynamics as a consequence of long-term and recently imposed tillage in a clay loam soil. *Canadian Journal of Soil Science* 97(4), pp. 561-579.

Fracassetti, M. et al. 2015. Validation of Pooled Whole-Genome Re-Sequencing in Arabidopsis lyrata. *PLoS One* 10(10), p. e0140462. doi: 10.1371/journal.pone.0140462

Freedman, A. H. et al. 2019. Error, noise and bias in de novo transcriptome assemblies. *BioRxiv* (4), doi: 10.1101/585745

Freese, N. H. et al. 2016. Integrated Genome Browser: Visual analytics platform for genomics. *Bioinformatics* 32(14), pp. 2089-2095.

Fujita, Y. et al. 2014. Environmental radioactivity damages the DNA of earthworms of Fukushima Prefecture, Japan. *European Journal of Wildlife Research* 60(1), pp. 145-148.

Garrison, E. and Marth, G. 2012. Haplotype-based variant detection from short-read sequencing. *arXiv* v1(1207.3907),

GATK-Team. 2019. *GATK - RNAseq short variant discovery (SNPs + Indels)*. https://gatk.broadinstitute.org/hc/en-us/articles/360035531192-RNAseq-short-variant-discovery-SNPs-Indels-: Broad Instutute. Available at: http://gatk.broadinstitute.org/hc/en-us/articles/360035531192-RNAseq-short-variant-discovery-SNPs-Indels- [Accessed: 29/02/20].

Gautier, M. et al. 2013. Estimation of population allele frequencies from next-generation sequencing data: pool-versus individual-based genotyping. *Mol Ecol* 22(14), pp. 3766-3779. doi: 10.1111/mec.12360

Ghosh, S. M. et al. 2013. Temperature-size rule is mediated by thermal plasticity of critical size in Drosophila melanogaster. *Proc Biol Sci* 280(1760), doi: 10.1098/rspb.2013.0174

Giardina, B. et al. 1975. Studies on Erythrocruorin: III oxygen equilibrium of earthworm erythrocruorin. *Journal of Molecular Biology* 93, pp. 1-10.

Gilbert, D. 2013. Gene-omes built from mRNA seq not genome DNA. *7th annual arthropod genomics symposium.* Notre Dame.

Gonzalez-Alcaraz, M. and van Gestel, C. A. M. 2016. Metal/metalloid (As, Cd and Zn) bioaccumulation in the earthworm Eisenia andrei under different scenarios of climate change. *Environmental pollution* 215, pp. 178-186.

Gotz, S. et al. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic acids research* 36(10), pp. 3420-3435.

Goulson, D. 2019. The Insect Apocalypse, and Why It Matters. *Current biology : CB* 29(19), doi: 10.1016/j.cub.2019.06.069

Grabherr, M. G. et al. 2011. Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nature Biotechnology* 29(7), pp. 644-652.

Gracey, A. Y. et al. 2004. Coping with cold: An integrative, multitissue analysis of the transcriptome of a poikilothermic vertebrate.*Proc Natl Acad Sci U S A*. Vol. 101. pp. 16970-16975.

Graff, O. 1953. Investigations in soil zoology with special reference to the terricol Oligochaeta. *PflErnahr. Dung* 61, pp. 72-77.

Grant Jr, W. C. 1955. Studies on Moisture Relationships in Earthworms. *Ecology* 36(3), pp. 400-407.

Grant, W. C. 1955. Temperature relationships in the megascolecid earthworm *Pheretima hupeiensis*. *Ecology* 36(3), pp. 412-417.

Gregory, T. R. and Herbert, P. D. N. 2002. Genome size estimates for some oligochaete annelids. *Canadian Journal of Zoology* 80, pp. 1485-1489.

Grieshaber, M. K. et al. 2005. Physiological and metabolic responses to hypoxia in invertebrates | SpringerLink. *Reviews of Physiology, Biochemistry and Pharmacology* 125, pp. 43-147. doi: 10.1007/BFb0030909

Guild, W. J. M. L. 1951. The distribution and population density of earthworms (Lumbricidae) in Scottish pasture fields. *Journal of Animal Ecology* 20(1), pp. 88-97.

Haas, B. J. et al. 2003. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research* 31, pp. 5654-5666.

Haas, B. J. et al. 2013. De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. *Nature Protocol* 8(8), p. 10.

Hackert, M. and Riggs, A. F. 2006. When Size Matters. *Structure* 14(7), pp. 1094-1096.

Hale, C. M. 2008. Evidence for human-mediated dispersal of exotic earthworms: support for exploring strategies to limit further spread. *Mol Ecol* 17(5), pp. 1165-1167. doi: 10.1111/j.1365-294X.2007.03678.x

Hama, K. 1960. The Fine Structure of Some Blood Vessels of the Earthworm, *Eisenia foetida*. *Journal of Cell Biology* 7(4), p. 717.

Harrison, J. F. et al. 2010. Atmospheric oxygen level and the evolution of insect body size. *Proc Biol Sci* 277(1690), pp. 1937-1946. doi: 10.1098/rspb.2010.0001

Heather, J. M. and Chain, B. 2016. The sequence of sequencers: The history of sequencing DNA. *Genomics* 107(1), pp. 1-8.

Heberle, H. et al. 2015. IneractiVenn: a web-based tool for the analysis of sets through Venn diagrams. *BMC Bioinformatics* 16(169),

Heleno, R. and Vargas, P. 2015. How do islands become green? *Global Ecology and Biogeography* 24(1), pp. 518-526.

Hemp, A. 2006. Vegetation of Kilimanjaro: hidden endemics and missing bamboo. *African Journal of Ecology* 44(33), pp. 305-328.

Hendrix, P. F. et al. 2008. Pandora's box contained bait: The global problem of introduced earthworms. *Annual Review of Ecology, Evolution and Systematics* 39(1), pp. 593-613.

Hernández-Oñate, M. A. and Herrera-Estrella, A. 2015. Damage response involves mechanisms conserved across plants, animals and fungi. *Current Genetics* 61(3), pp. 359-372.

Ho, S. Y. and Shapiro, B. 2011. Skyline-plot methods for estimating demographic history from nucleotide sequences. *Mol Ecol Resour* 11(3), pp. 423-434. doi: 10.1111/j.1755-0998.2011.02988.x

Hoeffner, K. et al. 2018. Feeding behaviour of epi-anecic earthworm species and their impacts on soil microbial communities. *Soil Biology and Biochemistry* 125(1), pp. 1-9.

Holmstrup, M. 2003. Overwintering adaptations in earthworms. *Pedo biologia* 47, pp. 504-510.

Holmstrup, M. and Overgaard, J. 2007. Freeze tolerance in Aporrectodea caliginosa and other earthworms from Finland. *Cryobiology* 55(1), pp. 80-86.

Hoppeler, H. and Vogt, M. 2001. Muscle tissue adaptations to hypoxia. *Journal of Experimental Biology* 204, pp. 3133-3139.

Horowitz, M. 2014. *Heat Acclimation, Epigenetics, and Cytoprotection Memory - Horowitz - - Major Reference Works - Wiley Online Library*.

Huang, R. et al. 2019. Frontiers | A Potential Role for Epigenetic Processes in the Acclimation Response to Elevated pCO2 in the Model Diatom Phaeodactylum tricornutum | Microbiology. *Microbiology* 8(10), p. 158. doi: doi:10.3389/fmicb.2018.03342

Inouye, D. W. and Wielgolaski, F. E. 2003. *Phyenology: an intergrative environmental science*. Springer Netherlands.

Jain, M. et al. 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology* 36, pp. 338-345.

James, S. W. 2004. *Earthworm ecology: Chapter 3 Planetary processes and their interactions with earthworm distributions and ecology.* 2 ed. United States of America: CRC Press.

James, S. W. et al. 2010. DNA Barcoding Reveals Cryptic Diversity in Lumbricus Terrestris L., 1758 (Clitellata): Resurrection of L. Herculeus (Savigny, 1826). *PloS one* 5(12), doi: 10.1371/journal.pone.0015629

Jegou, D. et al. 1998. Assessment of the burrow system of Lumbricus terrestris, Aporrectodea giardi and Aporrectodea cliginosa using X-ray computed tomography. *Biology and fertility of soils* 26, pp. 116-121.

Johnston, A. S. A. et al. 2014. Earthworm distribution and abundance predicted by a process-based model. *Applied Soil Ecology* 84, pp. 112-123.

Jones, C. A. et al. 2014. The spatial-temporal characteristics of type I collagen-based extracellular matrix. *Soft Matter* 10(44), pp. 8855-8863. doi: 10.1039/c4sm01772b

Jovana, M. et al. 2014. Effects of three pesticides on the earthworm *Eisenia fetida* (Savigny 1826) under laboratory conditions: Assessment of mortality, biomass and growth inhibition. *European journal of Soil Biology* 62, pp. 127-131.

Kajitani, R. et al. 2014. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research* 24(8), pp. 1384-1395.

Kang, R. et al. 2014. HMGB1 in Health and Disease. *Molecular aspects of medicine* 40, doi: 10.1016/j.mam.2014.05.001

Kashmenskaya, M. N. and Polyakov, A. V. 2008. Karyotype analysis of five species of earthworms (Oligochaeta: Lumbricidae). *Comparative Cytogenetics* 2(2), pp. 121-125.

Katsimpardi, L. et al. 2020. Systemic GDF11 stimulates the secretion of adiponectin and induces a calorie restriction-like phenotype in aged mice. *Aging Cell* 19(1), p. e13038. doi: 10.1111/acel.13038

Ke, Q. and Costa, M. 2006. Hypoxia-Inducible Factor-1 (HIF-1). *Molecular Pharmacology* 70(5), pp. 1469-1480.

Kelly, M. 2019. Adaptation to climate change through genetic accommodation and assimilation of plastic phenotypes. *Philosophical Transactions of the Royal Society B*, doi: doi:10.1098/rstb.2018.0176

Kille et al. 2013. DNA sequence variation and methylation in an arsenic tolerant earthworm population. *Soil Biology and Biochemistry* 57, pp. 524–532.

King, R. A. et al. 2008. Opening a can of worms: unprecedented sympatric cryptic diversity within British lumbricid earthworms. *Molecular Ecology* 17(21), pp. 4684-4698.

Klarica, J. et al. 2012. Comparing four mitochondrial genes in earthworms - Implications for identification, phylogenetics, and discovery of cryptic species. *Soil Biology and Biochemistry* 45, pp. 23-30.

Klok, C. J. and Harrison, J. F. 2013. The Temperature Size Rule in Arthropods: Independent of Macro-Environmental Variables but Size Dependent. *Integrative and Comparative Biology* 53(4), pp. 557-570. doi: 10.1093/icb/ict075

Klutsch, C. F. et al. 2012. Phylogeographical analysis of mtDNA data indicates postglacial expansion from multiple glacial refugia in woodland caribou (Rangifer tarandus caribou). *PLoS One* 7(12), p. e52661. doi: 10.1371/journal.pone.0052661

Kolattukudy, M. et al. 2017. One Step Closer to Mars with Aquaponics: Cultivating Citizen Science in K12 Schools. *NASA Technical Reports Server* 20180000946, pp. 1-38. doi: http://ntrs.nasa.gov/search.jsp?R=20180000946

Kooch, Y. and Jalilvand, H. 2008. Earthworms as Ecosystem Engineers and the Most Important Detritivors in Forest Soils. *Pakistan Journal of Biology Sciences* 11, pp. 819-825.

Korean, S. et al. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research* 27, pp. 722-736.

Korner, C. 2007. The use of 'altitude' in ecological research. *Trends in Ecology and Evolution* 22(11), pp. 569-574.

Kretzschmar, A. and Bruchou, C. 1991. Weight response to the soil water potential of the earthworm *Aporrectodea longa*. *Biology and fertile soils* 12, pp. 209-212.

Kronholm, I. and Ketola, T. 2018. Effects of acclimation time and epigenetic mechanisms on growth of Neurospora in fluctuating environments. *Heredity* 121(4), pp. 327-341. doi: doi:10.1038/s41437-018-0138-2

Kumar, S. et al. 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0. *Molecular Biology and Evolution* 33(7), pp. 1870-1874.

Laetsch, D. R. and Blaxter, M. L. 2017a. BlobTools: Interrogation of genome assemblies. *F1000Research 2017 6:1287* 6(1287), doi: doi:10.12688/f1000research.12232.1

Laetsch, D. R. and Blaxter, M. L. 2017b. BlobTools: Interrogation of genome assemblies [version 1; referees: 2 approved with reservations]. *F1000Research* 6, p. 1287.

Lamb, J. C. and Birchler, J. A. 2003. The role of DNA sequence in centromere formation. *Genome Biol* 4(5), p. 214. doi: 10.1186/gb-2003-4-5-214

Larkin, M. A. et al. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23(21), pp. 2947-2948. doi: 10.1093/bioinformatics/btm404

Lee, T. H. et al. 2014. SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. *BMC Genomics* 15(1),

Leigh, J. W. and Bryant, D. 2015. popart: full‐feature software for haplotype network construction - Leigh - 2015 - Methods in Ecology and Evolution - Wiley Online Library. *Methods in Ecology and Evolution* 6(9), pp. 1110-1116. doi: 10.1111/2041-210X.12410

Lemtiri, A. et al. 2018. *Chapter 5 - Short-Term Effects of Tillage Practices and Crop Residue Exportation on Soil Organic Matter and Earthworm Communities in Silt Loam Arable Soil*. Soil Management and Climate Change: Academic Press.

Lenoir, J. et al. 2008. A significant upward shift in plant species optimum elevation during the 20th century. *Science* 320(5884), pp. 1768-1771.

Lever, J. et al. 2017. Principal component analysis. *Nature Methods* 14(7), pp. 641-642.

Lewin, H. A. et al. 2018. Earth BioGenome Project: Sequencing life for the future of life. *Proc Natl Acad Sci U S A* 115(17), pp. 4325-4333. doi: 10.1073/pnas.1720115115

Lewis, K. A. et al. 2018. *The Veterinary Substance Database (VSDB) Developed by the Agriculture & Environment Research Unit (AERU)*. https://sitem.herts.ac.uk/aeru/vsdb/Reports/1455.htm: University of Hertfordshire. Available at: [Accessed: 14th June].

Leys, M. et al. 2016. Distribution and population genetic variation of cryptic species of the Alpine mayfly Baetis alpinus (Ephemeroptera: Baetidae) in the Central Alps. *BMC Evolutionary Biology* 16(1), pp. 1-15. doi: doi:10.1186/s12862-016-0643-y

Li, D. et al. 2015. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31(10), pp. 1674-1676.

Li, D. et al. 2020. Breeding history and candidate genes responsible for black skin of Xichuan black-bone chicken. *BMC Genomics* 511,

Li, H. 2016. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* 32(14), pp. 2103-2110.

Li, H. et al. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16), pp. 2078-2079. doi: 10.1093/bioinformatics/btp352

Liang, S. et al. 2019. DiVenn: An interactive and integrated Web-based visualization tool for comparing gene lists. *Frontiers in Genetics* 10(421),

Lim, S. L. et al. 2014. Treatment and Biotransformation of Highly Polluted Agro-industrial Wastewater from a Palm Oil Mill into Vermicompost Using Earthworms. *Journal of Agric. Food Chem.* 62(3), pp. 691-698.

Lin, D. et al. 2012. Physiological and molecular responses of the earthworm (*Eisenia fetida*) to soil chlorotetracycline contamination. *Environmental pollution* 171, pp. 46-51.

Loranger, G. et al. 2001. Does soil acidity explain altitudinal sequences in collembolan communities? *Soil Biology and Biochemistry* 33(3), pp. 381-393.

Louhelainen, J. 2016. SNP Arrays.*Microarrays (Basel)*. Vol. 5.

Luo, R. et al. 2015. Erratum: SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience* 4, p. 30.

Lyman, R. F. et al. 2003. Variation in Drosophila sensory bristle number at ‘Evolution Canyon’ | Genetics Research | Cambridge Core. *Genetics Research* 80(3), pp. 215-223. doi: doi:10.1017/S0016672302005876

Madoui, M. et al. 2015. Genome assembly using Nanopore-guided long and error-free DNA reads. *BMC Genomics* 16, p. 327.

Maeder, P. et al. 2002. Soil fertility and biodiversity in organic farming. *Science* 296(5573), pp. 1694-1697.

Majmundar, A. J. et al. 2010. Hypoxia-inducible factors and the response to hypoxic stress. *Mol Cell* 40(2), pp. 294-309. doi: 10.1016/j.molcel.2010.09.022

Manalo, D. J. et al. 2005. Transcription regulation of vascular endothelial cell responses to hypoxia by HIF-1. *Blood* 105(2), pp. 659-669.

Marchan, D. F. et al. 2018. Why are we blind to cryptic species? Lessons from the eyeless. *European Journal of Soil Biology* 86(1), pp. 59-51.

Marchán, D. F. et al. 2020a. Genome-informed integrative taxonomic description of three cryptic species in the earthworm genus Carpetania (Oligochaeta, Hormogastridae). *Systematics and Biodiversity* 18(3), doi: TSAB-2019-0077.R1

Marchán, D. F. et al. 2020b. Local adaptation fuels cryptic speciation in terrestrial annelids. *Molecular Phylogenetics and Evolution* 146, pp. 106767-106767. doi: 10.1016/j.ympev.2020.106767

Marinissen, J. C. Y. and van den Bosch, F. 1992. Colonization of new habitats by earthworms. *Oecologia* 91, pp. 371-376.

Martinotti, S. et al. 2015. Emerging roles for HMGB1 protein in immunity, inflammation, and cancer.*Immunotargets Ther*. Vol. 4. pp. 101-109.

Martinsson, S. et al. 2017. Barcoding gap, but no support for cryptic speciation in the earthworm *Aporrectodea longa* (Clitellata: Lumbricidae). *Mitochondrial DNA. Part A, DNA Mapping sequencing, and analysis.* 28(2), pp. 147-155.

Mathias, M. L. et al. 1998. Mammals from the Azores islands (Portugal): an updated overview. *Mammalia* 62(3), pp. 397-407.

Mayjonade, B. et al. 2016. Extraction of high-molecular-weight-genomic DNA for long-read sequencing of single molecules. *Biotechniques* 1(61), pp. 203-205.

McBryan, T. L. et al. 2013. Responses to temperature and hypoxia as interacting stressors in fish: implications for adaptation to environmental change. *Integr Comp Biol* 53(4), pp. 648-659. doi: 10.1093/icb/ict066

McCain, C. M. 2020. The mid-domain effect applied to elevational gradients: species richness of small mammals in Costa Rica. *Journal of Biogeography* 31, pp. 19-31.

McKenna, A. et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20(9), pp. 1297-1303. doi: 10.1101/gr.107524.110

Meteoblue. 2016a. *Climate Azores*. https://www.meteoblue.com/en/weather/forecast/modelclimate/azores_portugal_3373385: Available at: [Accessed: 05/09/16].

Meteoblue. 2016b. *Climate Chimborazo*. Meteoblue. Available at: https://www.meteoblue.com/en/weather/forecast/modelclimate/chimborazo_ecuador_3754872 [Accessed: 10/10/16].

Meteoblue. 2016c. *Climate Les Deux Alpes*. Meteoblue. Available at: https://www.meteoblue.com/en/weather/forecast/modelclimate/les-deux-alpes_france_6354942 [Accessed: 10/10/16].

Meteoblue. 2016d. *Climate Mount Kilimanjaro*. Meteoblue. Available at: https://www.meteoblue.com/en/weather/forecast/modelclimate/mount-kilimanjaro_tanzania_157452 [Accessed: 10/10/16].

Meteoblue. 2016e. *Climate Mount Kinabalu*. Meteoblue. Available at: https://www.meteoblue.com/en/weather/forecast/modelclimate/mount-kinabalu_malaysia_1736632 [Accessed: 10/10/16].

Meteoblue. 2016f. *Climate Weber Country Memorial Park*. Meteoblue. Available at: https://www.meteoblue.com/en/weather/forecast/modelclimate/weber-county-memorial-park_united-states-of-america_5784441 [Accessed: 10/10/16].

Milcu, A. et al. 2006. The response of decomposers (earthworms, springtails and microorganisms) to variations in species and functional group diversity of plants. *OIKOS* 112(3), pp. 513-524.

Miller, M. P. 2005. Alleles In Space: Computer software for the analysis of interindividual spatial and genetic information. *Journal of Heredity* 96(1), pp. 722-724.

Miller, M. R. et al. 2007. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research* 17(2), pp. 240-248.

Miranda, J. M. et al. 1998. Tectonic setting of the Azores Plateau deduced from a OBS survey. *Marine Geophysical Researches* 20(3), pp. 171-182.

Mitton, J. B. 1997. *Selection in Natural Populations*. New York: Oxford University Press.

Monge, C. and Leon-Velarde, F. 1991. Physiological adaptation to high altitude: Oxygen transport in mammals and birds. *Physiological reviews* 71(4), pp. 1135-1172.

Monroy, F. et al. 2007. Life cycle of the earthworm Octodrilus complanatus (Oligochaeta, Lumbricidae). *Comptes Rendus Biologies* 330(5), pp. 389-391.

Montecchio, L. et al. 2015. Potential spread of forest soil-borne fungi through earthworm consumption and casting. *iForest- Biogeosciences and forestry* 8, pp. 295-301.

Moores, E. M. and Fairbridge, R. W. 1998. *Encyclopedia of European and Asian Regional Geology*. 1 ed. London: Springer Netherlands.

Moreira, M. 2013. *Valoração da biodiversidade no Parque Natural de Ilha do Pico através da metodologia InVEST. Relatório técnico desenvolvido no âmbito do Projeto SMARTPARKS.* Pnta Delgada, Açores: Universidade dos Açores.

Morueta-Holme, N. et al. 2015. Strong upslope shifts in Chimborazo's vegetation over two centuries since Humboldt. *PNAS* 112(41), pp. 12741-12745.

Mulder, C. et al. 2007. Empirical maximum lifespan of earthworms is twice that of mice. *Age* 29(4), pp. 229-231.

Nagatsu, T. and Sawada, M. 2006. Molecular mechanism of the relation of monoamine oxidase B and its inhibitors to Parkinson's disease: possible implications of glial cells. *J Neural Transm Suppl* (71), pp. 53-65. doi: 10.1007/978-3-211-33328-0_7

Neuhauser, E. F. et al. 1995. Bioconcentration and biokinetics of heavy metals in the earthworm. *Environmental pollution* 89(3), pp. 293-301.

Noman, M. Z. et al. 2015. Hypoxia: a key player in antitumor immune response. A Review in the Theme: Cellular Responses to Hypoxia. *Cell Physiology* 309(9), pp. C569-C579.

Nordstrom, S. and Rundgren, S. 1974. Environmental factors and lumbricid associations in Southern Sweden. *Pedobiologia* 14, pp. 1-27

Novo, M. et al. 2015. Multiple introductions and environmental factors affecting the establishment of invasive species on a volcanic island. *Soil Biology and Biochemistry* 85(1), pp. 89-100.

Nurk, S. et al. 2013. Assembling Genomes and Mini-metagenomes from Highly Chimeric Reads. *Research in Computational Molecular Biology. RECOMB 2013. Lecture Notes in Computer Science* 7821,

Nuutinen, V. and Butt, K. R. 2009. Worms from the cold: Lumbricid life stages in boreal clay during frost. *Soil Biology and Biochemistry* 41(7), pp. 1580-1582.

Olsson, E. G. A. et al. 2000. Landscape change patterns in mountains, land use and environmental diversity, mid-Norway 1960-1993. *Landscape ecology* 15, pp. 155-170.

Orozco-terWengel, P. et al. 2012. Adaptation of Drosophila to a novel laboratory environment reveals temporally heterogeneous trajectories of selected alleles. *Mol Ecol* 21(20), pp. 4931-4941. doi: 10.1111/j.1365-294X.2012.05673.x

Ortiz-Gamino, D. et al. 2016. Invasion of the Tropical Earthworm Pontoscolex corethrurus (Rhinodrilidae, Oligochaeta) in Temperate Grasslands. *PeerJ* 4(e2572), doi: 10.7717/peerj.2572

Ostle, N. et al. 2007. Isotopic detection of recent photosynthate carbon flow into grassland rhizosphere fauna. *Soil Biology and Biochemistry* 39(3), pp. 768-777.

Oxford-Nanopore-Technologies. 2018a. 1D Genomic DNA by Ligation (SQK-LSK109) checklist. *Nanopore Protocol* Version(GDE_9063_v109_revD_23May2018), pp. 1-8.

Oxford-Nanopore-Technologies. 2018b. *MinION*. https://nanoporetech.com/products/minion: Oxford Nanopore Technologies. Available at: [Accessed: 26/06/18].

PacBio. 2018. Revolutionize genomics with SMRT sequencing. In: California, P.B.o. ed. California: PacBio.

Pacheoco, P. R. et al. 2010. HLA Class I and II profiles in São Miguel Island (Azores): genetic diversity and linkage disequilibrium. *BCM research notes* 3(1), p. 134.

Parmelee, R. W. et al. 1990. Earthworms and enchytraeids in conventional and no-tillage agroecosystems: A biocide approach to assess their role in organic matter breakdown. *Biology and Fertility of Soils* 10(1), pp. 1-10.

Parolo, G. and Rossi, G. 2008. Upward migration of vascular plants following a climate warming trend in the Alps. *Basic and Applied Ecology* 9(1), pp. 100-107.

Pass, D. A. 2015. *The Earthworm Microbiome.* Cardiff University.

Paudel, S. et al. 2016. Predicting spatial extent of invasive earthworms on an oceanic island - Paudel - 2016 - Diversity and Distributions - Wiley Online Library. *Diversity and Distributions* 22, pp. 1013-1023. doi: 10.1111/ddi.12472

Pauli, H. et al. 2003. Effects of climate change on the alpine and nival vegetation of the alps. *Journal of mountain ecology* 7, pp. 9-12.

Pelosi, C. et al. 2014. Pesticides and earthworms. A review. *Agronomy for sustainable development* 34(1), pp. 199-228.

Perez-Losada, M. et al. 2009. Phylogenetic assessment of the earthworm Aporrectodea caliginosa species complex (Oligochaeta: Lumbricidae) based on mitochondrial and nuclear DNA sequences. *Molecular Phylogenetics and Evolution* 52(2), pp. 293-302.

Pertea, M. and Salzberg, S. L. 2010. Between a chicken and a grape: estimating the number of human genes. *Genome Biology* 11(5), p. 206.

Peterlongo, P. et al. 2010. Identifying SNPs without a Reference Genome by Comparing Raw Reads | SpringerLink. *International Symposium on String Processing and Information Retrieval*, pp. 147-158. doi: 10.1007/978-3-642-16321-0_14

Phillips, H. R. P. et al. 2019. Global Distribution of Earthworm Diversity. *Science* 366(6464), pp. 480-485. doi: 10.1126/science.aax4851

Picard-Team. 2019. *A set of command line tools (in Java) for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF.* http://broadinstitute.github.io/picard/: Broad Instutute. Available at: [Accessed: 29/02/20].

Piearce, T. G. 1972. The calcium relations of selected Lumbricidae. *Journal of Animal Ecology* 41, pp. 167-188.

Pilpel, Y. 2011. *Noise in Biological Systems: Pros, Cons, and Mechanisms of Control | SpringerLink*. Yeast Systems Biology: SpringerLink.

Pinel, N. et al. 2008. *Verminephrobacter eiseniae* gen. nov. sp. nov., a nephridial symbiont of the earthworm *Eisenia foetida* (Savigny). *International Journal of Systems Evolution and Microbiology* 58(9), pp. 2147-2157.

Pomaznoy, M. et al. 2018. GOnet: a tool for interactive Gene Ontology analysis. *BMC Bioinformatics* 19(470),

Puritz, J. B. et al. 2014. Demystifying the RAD fad. *Mol Ecol* 23(24), pp. 5937-5942. doi: 10.1111/mec.12965

Pushkova, E. N. et al. 2019. Extraction of high-molecular weight from popular plants for Nanopore sequencing. *Plant biotechnology in the postgenomic era* 51,

Pérez-Losada, M. et al. 2009. Phylogenetic assessment of the earthworm Aporrectodea caliginosa species complex (Oligochaeta: Lumbricidae) based on mitochondrial and nuclear DNA sequences. *Mol Phylogenet Evol* 52(2), pp. 293-302. doi: 10.1016/j.ympev.2009.04.003

Qiu, Q. et al. 2012. The yak genome and adaption to life at high altitude. *Nature genetics* 44, pp. 946-949.

Quantapore, I. 2018. *Quantapore inc. closes series 3 financing round to fund beta launch of DNA sequencing.* https://quantapore.com/2018/07/18/effective-product-promo-videos/: Quantapore Inc. Available at: [Accessed: 18/01/19].

Rambaut, A. 2016. *FigTree v1.4.3.* http://tree.bio.ed.ac.uk/software/figtree/: Available at: [Accessed: 15/01/18].

Ratcliffe, P. J. 2013. Oxygen sensing and hypoxia signalling pathways in animals: the implications of physiology for cancer. *The Journal of Physiology*, pp. 2027-2042.

Ratnasingham, S. and Herbert, P. D. N. 2007. BOLD: The Barcode of Life Data System. *Molecular Ecology Notes* 7(3), p. 10. doi: doi: 10.1111/j.1471-8286.2006.01678.x

Reuter, J. A. et al. 2015. High-Throughput sequencing Technologies. *Molecular Cell* 58(4), pp. 586-597.

Rey, C. et al. 2019. CAARS: comparative assembly and annotation of RNA-Seq data. *Bioinformatics* 35(13), pp. 2199-2207.

Richards, A. L. et al. 2019. Gut Microbiota Has a Widespread and Modifiable Effect on Host Gene Regulation. *American Journal of Microbiology* mSystems 4(e00323-18),

Ridgeway, J. A. and Timm, A. E. 2014. Comparison of RNA Isolation Methods from Insect Larvae. *Journal of Insect Science* 14(268),

Riehl, S. et al. 2013. Emergence of agriculture in the foothills of the Zagros mountains of Iran. *Science* 341(6141), pp. 65-67.

Rimmington, O. 2019. Nanochrome. https://github.com/OliverCardiff/Nanochrome_scaffolder.

Rodrigues, A. F. et al. 2015. Early Atlantic navigation: pre-Portuguese p resence in the Azores islands. *Archaeological Discovery* 3(1), pp. 104-113.

Roman-Palacios, C. and Wiens, J. J. 2020. Recent responses to climate change reveal the drivers of species extinction and survival. *Proc Natl Acad Sci U S A* 117(8), pp. 4211-4217. doi: 10.1073/pnas.1913007117

Royer Jr, W. E. et al. 2006. Lumbricus Erythrocruorin at 3.5 Å Resolution: Architecture of a Megadalton Respiratory Complex. *Structure* 14(7), pp. 1167-1177.

Rozas, J. et al. 2017. DnaSP v6: DNA sequence polymorphism analysis of large datasets. *Molecular Biology and Evolution* 34(1), pp. 3299-3302.

Rozen, R. et al. 2013. Altitude versus vegetation as the factors influencing the diversity and abundance of earthworms and other soil macrofauna in montane habitat (Silesian Beskid Mts, Western Carpathians). *Polish Journal of Ecology* 61(1), pp. 145-156.

Ruan, J. 2016. *SMARTdenovo: Ultra-fast de novo assembler using long noisy reads*. https://github.com/ruanjue/smartdenovo: GitHub. Available at: [Accessed: 28/02/19].

Ruan, J. and Li, H. 2019. Fast and accurate long-read assembly with wtdbg2. *bioRxiv* Preprint,

Safonova, Y. et al. 2015. dipSPAdes: Assembler for highly polymorphic diploid genomes. *Journal of computational Biology* 22(6), pp. 528-545.

Sanders, N. J. 2002. Elevational gradients in ant species richness: area, geometry, and Rapoport's rule - Sanders - 2002 - Ecography - Wiley Online Library. *Ecography* 25(1), pp. 25-32. doi: 10.1034/j.1600-0587.2002.250104.x

Sanders, S. M. et al. 2014. Differential gene expression between functionally specialized polyps of colonial hydrozoan *hydractinia symbiolongicarpus* (Phylum Cnidaria). *BMC Genomics* 15,

Satchell, J. E. 1955. *Some aspects of earthworm ecology in Soil Zoology*. London: Butterworths.

Scheinfeldt, L. B. and Tishkoff, S. A. 2010. Living the high life: high-altitude adaptation. *Genome Biology* 11(9), pp. 1-3. doi: doi:10.1186/gb-2010-11-9-133

Scheu, S. 2003. Effects of earthworms on plant growth: patterns and perspectives. *Pedobiologia* 47(5-6), pp. 846-856.

Schmitz, A. and Harrison, J. F. 2004. Hypoxic Tolerance in Air-Breathing Invertebrates. *Respiratory physiology & neurobiology* 141(3), doi: 10.1016/j.resp.2003.12.004

Schofield, C. J. and Ratcliffe, P. J. 2004. Oxygen sensing by HIF hydroxylases. *Nat Rev Mol Cell Biol* 5(5), pp. 343-354. doi: 10.1038/nrm1366

Schwessinger, B. and Rathjen, J. P. 2017. *Extraction of High Molecular Weight DNA from Fungal Rust Spores for Long Read Sequencing*. New York: Humana Press.

Semenza, G. L. 2014. Oxygen sensing, hypoxia-inducible factors, and disease pathophysiology. *Annu Rev Pathol* 9, pp. 47-71. doi: 10.1146/annurev-pathol-012513-104720

Serag, M. F. and Habuchi, S. 2017. Conserved linear dynamics of single-molecule Brownian motion. *Nature communications* 8(15675), pp. 1- 11.

Shao, y. et al. 2020. Genome and single-cell RNA-sequencing of the earthworm Eisenia andrei identifies cellular mechanisms underlying regeneration. *Nature Communications* 11(1), pp. 1-15. doi: doi:10.1038/s41467-020-16454-8

Shekhovtsov, S. V. et al. 2019. Transcriptomic Analysis Confirms Differences Among Nuclear Genomes of Cryptic Earthworm Lineages Living in Sympatry. *BMC evolutionary biology* 19(Suppl 1), doi: 10.1186/s12862-019-1370-y

Sherlock, E. 2018. *Key to the earthworms of the UK and Ireland*. 2nd ed. Telford: FSC Publications.

Shouolts-Wilson, W. A. et al. 2010. Role of particle size and soil type in toxicity of silver nanoparticles to earthworms. *ACSESS* 75(2), pp. 365-377.

Sibrava, V. et al. 1986. Quaternary Glaciations in the Northern Hemisphere. *Quaternary Science Reviews* 5(1), pp. 1-514.

Silva-Junior, O. B. et al. 2018. Design and evaluation of a sequence capture system for genome-wide SNP genotyping in highly heterozygous plant genomes: a case study with a keystone Neotropical hardwood tree genome. *DNA Res* 25(5), pp. 535-545. doi: 10.1093/dnares/dsy023

Simillion, C. et al. 2017. Avoiding the pitfalls of gene set enrichment analysis with SetRank.*BMC Bioinformatics*. Vol. 18.

Simonson, T. S. et al. 2010. Genetic Evidence for High-Altitude Adaptation in Tibet. *Science* 329(5987), pp. 72-75.

Sims, R. W. and Gerard, B. M. 1999. *Earthworms*. London: Linnean Society of London and The Estuarine and Coastal Sciences Association.

Singh, J. et al. 2019. Climate change effects on earthworms - a review. *Soil org.* 91(3), pp. 114-138.

Singh, R. P. et al. 2011. Management of urban solid waste: vermicomposting a sustainable option. *Resources, Conservation and Recycling* 55(7), pp. 719-729.

Singh, V. and Singh, K. 2015. Toxic effect of herbicide 2,4-D on the Earthworm *Eutyphoeus waltoni* Michaelsen. *Environmental Processes* 2(1), pp. 251-260.

Sinha, R. K. et al. 2009. Vermistabilization of sewage sludge (biosolids) by earthworms: converting a potential biohazard destined for landfill disposal into a pathogen-free, nutritive and safe biofertilizer for farms. *Waste Management and Reseach* 28(10), pp. 872-881.

Smit, A. F. A. et al. 2015. RepeatMasker Open-4.0 2013-2015. www.repeatmasker.org,

Smith, B. D. and Zeder, M. A. 2013. The onset of the Anthropocene. *Anthropocene* 4(1), pp. 8-13.

Somme, L. 1989. Adaptions of terrestrial arthropods to the alpine environment. *Biological reviews* 64(4), pp. 367-407.

Sommer, J. U. et al. 2010. Differential expression of presynaptic genes in a rat model of postnatal hypoxia: relevance to schizophrenia. *Eur Arch Psychiatry Clin Neurosci* 260 Suppl 2, pp. S81-89. doi: 10.1007/s00406-010-0159-1

Spiers, G. A. et al. 1986. Effects and importance of indigenous earthworms on decomposition and nutrient cycling in coastal forest ecosystems. *Canadian Journal of Forest Research* 16, pp. 983-989.

Spurgeon, D. J. and Hopkin, S. P. 1996. The effects of metal contamination on earthworm populations around a smelting works: quantifying species effects *Applied Soil ecology* 4(2), pp. 147-160.

Spurgeon, D. J. and Hopkin, S. P. 1999. Comparisons of metal accumulation and excretion kinetics in earthworms (eisenia fetida) exposed to contaminated field and laboratory soils. *Applied soil ecology* 11(2-3), pp. 227-243.

Spurgeon, D. J. et al. 2006. Effect of pH on metal speciation and resulting metal uptake and toxicity for eathworms. *Environmental toxicology and chemistry* 25(3), pp. 788-796.

Spurgeon, D. J. et al. 2005. Hierarchical Responses of Soil Invertebrates (Earthworms) to Toxic Metal Stress. *Environmental Science and Technology* 39(14), pp. 5327-5334.

Steinbock, L. J. and Radenovic, A. 2015. The emergence of nanopores in next-generation sequencing. *Nanotechnology* 26(7), pp. 1-5.

Steinwandter, M. et al. 2017. Effects of Alpine land-use changes: Soil macrofauna community revisited. *Ecol Evol* 7(14), pp. 5389-5399. doi: 10.1002/ece3.3043

Storz, J. F. 2016. Hemoglobin-oxygen affinity in high-altitude vertebrates: is there evidence for an adaptive trend? *Journal of Experimental Biology* 219, pp. 3190-3203.

Storz, J. F. et al. 2007. The molecular basis of high-altitude adaptation in Deer Mice. *PLOS Genetics*,

Storz, J. F. et al. 2010. Phenotypic Plasticity and Genetic Adaptation to High-Altitude Hypoxia in Vertebrates. *The Journal of experimental biology* 213(Pt 24), doi: 10.1242/jeb.048181

Stott, P. 2016. CLIMATE CHANGE. How climate change affects extreme weather events. *Science* 352(6293), pp. 1517-1518. doi: 10.1126/science.aaf7271

Supek, F. et al. 2011. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PLoS One* 6(7), doi: 10.1371/journal.pone.0021800

Svendsen, C. et al. 2006. Effect of temperature and season on reproduction, neutral red retention and metallothionein responses of earthworms exposed to metals in field soils. *Environmental Pollution* 147(1), pp. 83–93.

Szklarczyk, D. et al. 2019. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research* 47(D607-613),

Tan, M. H. et al. 2018. Finding Nemo. Hybrid assembly with Oxford Nanopore and Illumina reads greatly improves the clownfish (*Amphiprion ocellaris*) genome assembly. *GigaScience* 7(3), p. 137.

Tang, D. et al. 2011. High-mobility Group Box 1 [HMGB1] and Cancer. *Biochim Biophys Acta.* 1799(1-2), p. 131.

Tedersoo, L. et al. 2017. PacBio metabarcoding of Fungi and other eukaryotes: errors, biases and perspectives. *New Phytologist* 217, pp. 973-976.

Tiunov, A. V. and Scheu, S. 1999. Microbial respiration, biomass, biovolume and nutrient status in burrow walls of Lumbricus terrestris L. (Lumbricidae). *Soil Biology and Biochemistry* 31(14), pp. 2039-2048.

Tolkien, J. R. R. 2008. *The Lord of the Rings: The Fellowship of the Ring*. Reset edition ed. London: HarperCollins e-books.

Travlos, I. S. et al. 2017. Effects of the herbicides benfluralin, metribuzin and propyzamide on the survival and weight of earthworms (*Octodriulus complanatus*). *Plant soil environment* 63(3), pp. 117-124.

Tsiafouli, M. A. et al. 2014. Intensive agriculture reduces soil biodiversity across Europe. *Global change biology* 21(2), pp. 973-985.

USCAR and NSCAR. 2013. *Air Pressure Changes With Altitude*. https://scied.ucar.edu/learning-zone/how-weather-works/change-atmosphere-altitude: University Corporation for Atmospheric Research. Available at: [Accessed: 10/05/2020].

van Groenigen, J. W. et al. 2014. Earthworms increase plant production: a meta-analysis. *Scientific reports* 4(6365),

Varet, H. et al. 2016. SARTools: A DESeq2 and EdgeR Based R pipeline for Comprehensive Differential Analysis of RNA-Seq data. *PLoS One* e0157022,

Varrica, D. et al. 2000. Volcanic and anthropogenic contribution to heavy metal content in lichens from Mt. Etna and Vulcano island (Sicily). *Environmental pollution* 108(2), pp. 153-162.

Vaser, R. et al. 2017. Fast and accurate de novo genome assembly from long uncorrected read. *Genome Research* 27(5), pp. 737-746.

Virgós, E. et al. 2011. Food habits of European badgers (Meles meles) along an altitudinal gradient of Mediterranean environments: a field test of the earthworm specialization hypothesis. *Canadian Journal of Zoology* 82(1), pp. 41-51. doi: S1480328304032052

von Saltzwedel, H. et al. 2016. Founder events and pre-glacial divergences shape the genetic structure of European Collembola species. *BMC Evol Biol* 16(PMC4947257), doi: 10.1186/s12862-016-0719-8

Vos, H. M. J. et al. 2014. Do earthworms affect phosphorus availability to grass? A pot experiment. *Soil Biology and Biochemistry* 79(1), pp. 34-42.

Walker, B. J. et al. 2014. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLOS One* 9(11), p. e112963.

Wang, J. R. et al. 2018. FMLRC: Hybrid long read error correction using an FM-index. *BMC Bioinformatics* 19(1), p. 50.

Waterhouse, R. M. et al. 2017. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular Biology and Evolution* 35(3), pp. 543-548.

Weber, R. E. and Vinogradov, S. N. 2001. Nonvertebrate Hemoglobins: Functions and Molecular Adaptations. *Physiologial reviews* 81(2), pp. 569-628.

Weisenfeld, N. I. et al. 2017. Direct determination of diploid genome sequences. *Genome Research* 27, pp. 757-767.

Wever, L. A. et al. 2001. The influence of soil moisture and temperature on the survival, aestivation, growth and development of juvenile Aporrectodea tuberculata (Eisen) (Lumbricidae). *Pedobiologia* 45(2), pp. 121-133.

Wiback, S. J. and Palsson, B. O. 2002. Extreme pathway analysis of human red blood cell metabolism. *Biophysical journal* 83(2), pp. 808-818.

Woodhall, D. 1974. Geology and volcanic history of Pico Island volcano, Azores. *Nature* 248(1), pp. 663-665.

Xu, G.-C. et al. 2018. LR-Gapcloser: a tiling path-based gap closer that uses long reads to complete genome assembly. *Gigascience* 8, pp. 1-14.

Xue, W. et al. 2013. L_RNA_scaffolder: scaffolding genomes with transcripts. *BMC Genomics* 14, p. 604.

Yadav, A. and Garg, V. K. 2011. Recycling of organic wastes by employing *Eisenia fetida*. *Bioresource Technology* 102(3), pp. 2874-2880.

Yeo, S. et al. 2018. ARCS: scaffolding genome drafts with linked reads. *Bioinformatics* 34(5), pp. 725-731.

Zarea, M. J. and Karimi, N. 2012. Effect of herbicides on earthworms. *Dynamic soil, Dynamic Plant* 6(1), pp. 5-13.

Zhang, Q. et al. 2016. Genome Resequencing identifies unique adaptions of Tibetan Chickens to hypoxia and high-dose ultraviolet radiation in high-altitude environments. *Genome Biology and Evolution* 8(3), pp. 765-776.

Zhao, C. et al. 2019a. Altitudinal Biodiversity Gradient and Ecological Drivers for Different Lifeforms in the Baotianman Nature Reserve of the Eastern Qinling Mountains. *Forests* 10(4), p. 332. doi: 10.3390/f10040332

Zhao, Y. et al. 2019b. A high-throughput SNP discovery strategy for RNA-seq data.*BMC Genomics*. Vol. 20.

Zheng, G. X. et al. 2016. Haplotyping germline and cancer genomes using high-throughput linked-read sequencing. *Nat Biotechnol* 34(3), pp. 303-311. doi: 10.1038/nbt.3432

Zhou, Y. et al. 2009. The Schistosoma japonicum genome reveals features of host-parasite interplay. *Nature* 460(7253), pp. 345-351.

Zimin, A. V. et al. 2013. The MaSuRCA genome assembler. *Bioinformatics* 29(21), pp. 2669-2677.

Zorn et al. 2005. Species-specific earthworm population responses in relation to flooding dynamics in a Dutch floodplain soil. *Pedobiologia* 49, pp. 189-198.

Zwarycz, A. S. et al. 2015. Timing and Scope of Genomic Expansion Within Annelida: Evidence From Homeoboxes in the Genome of the Earthworm Eisenia Fetida. *Genome biology and evolution* 8(1), doi: 10.1093/gbe/evv243

10. Appendices

10.1. Individuals from San Miguel.

Site	*Latitude and longitude*
1	37.86772, -25.7708
2	37.838, -25.765
3	37.83455, -25.75845
4	37.83585, -25.75861
5	37.76478, -25.58911
7	37.76521, -25.58931
8	37.77688, -25.57993
9	37.7937, -25.5876
10	37.793, -25.535
11	37.781, -25.621

10.2. Mux scan results from Nanopore sequencing.

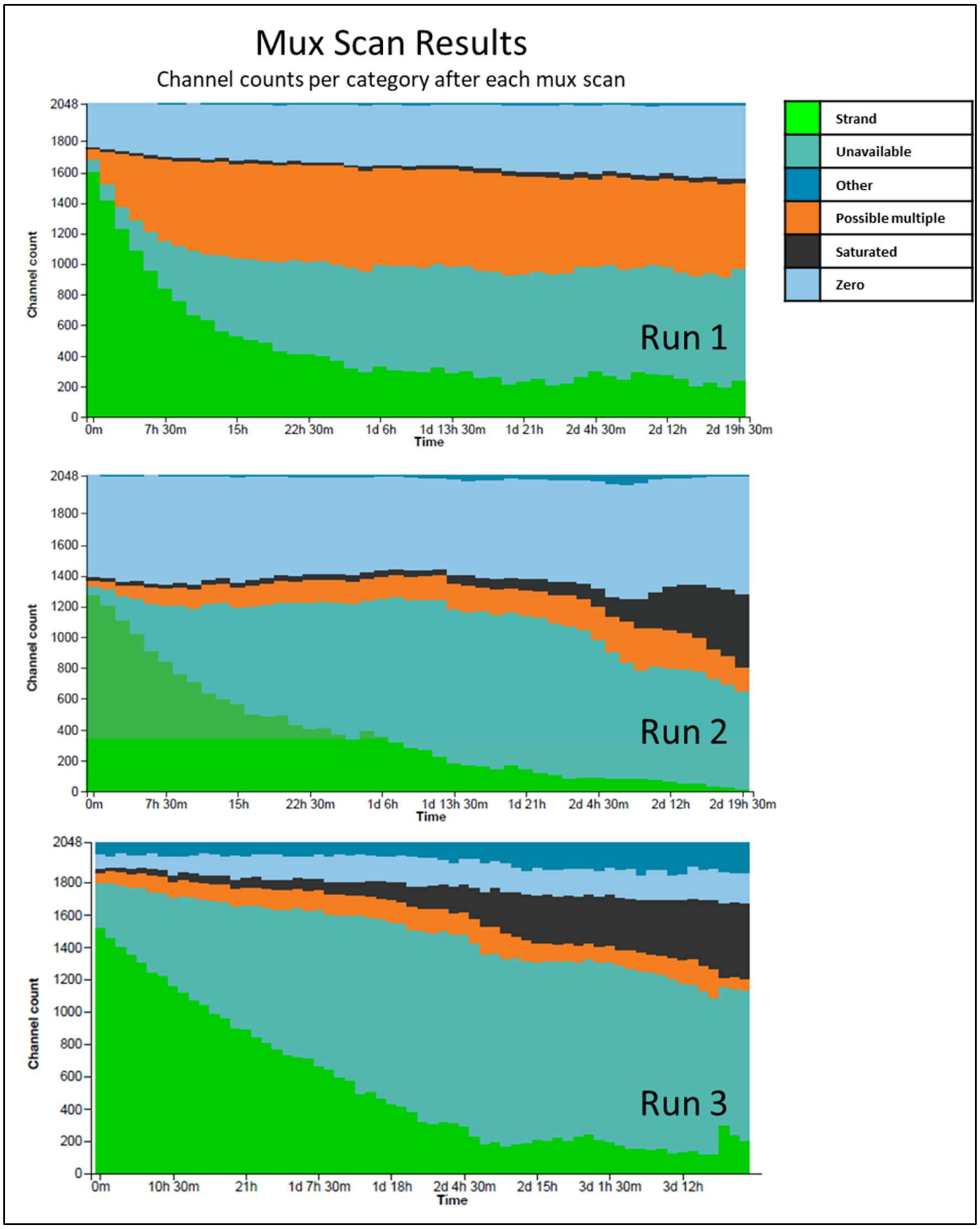


Figure 112: The absolute number of pores in different states after each mux. Mux scans were run every 1.5 hours to clear blocked pores where possible.