Author:
Cole, Rebecca L

Title:
The Distribution of Genetic Diversity Within and Among the Strongyloides ratti Genome

The Distribution of Genetic Diversity Within and Among the *Strongyloides ratti* Genome

Rebecca Cole

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Doctor of Philosophy in the Faculty of Life Sciences

School of Biological Sciences

January 2020

45,701 words

Abstract

The population genetics of parasitic nematodes is influenced principally by a combination of parasite life history traits and host movement ecology. *Strongyloides ratti* is a parasitic nematode infecting brown rats (*Rattus norvegicus*) and has an unusual life cycle featuring an obligate parasitic, asexually reproducing generation and a facultative free-living, sexually reproducing generation. This work aims to interrogate the population genetics of *S. ratti* and compare this to that of its rat hosts, and to look within the *S. ratti* genome to determine what the distribution of variable sites reveals about ongoing selection on parasitism as a trait.

Strongyloides ratti was a common infection of wild rats in three sampling sites in the Southern UK. The sexual form was not observed despite intensive sampling, suggesting that sexual reproduction is very rare within the sampling sites. DNA extracted from rat faeces was used to conduct a population genetic analysis of wild rats, revealing moderate genetic differentiation among sampling sites. Individual *S. ratti* were subjected to whole genome sequencing, and the 90 genome sequences produced were used to investigate the parasite's population genetics. The *S. ratti* sample was partitioned on the level of sympatric, genetically distinct clades, and these did not correlate with sampling site. It is hypothesised that each of the three largest clades was founded by different parasitic females and has subsequently proliferated asexually, with other clades and individuals deriving from rare sexual reproduction among clades. When looking within genomes, it was observed that genes upregulated in the parasitic adult form were more genetically diverse than the rest of the genome. Further, clusters of genes that are comparatively expanded in *Strongyloides* spp. and putatively contribute to the parasitic lifestyle were consistently more genetically diverse than directly adjacent flanking regions. This suggests that genes involved in parasitism are under diversifying selection.

Acknowledgements

My heartfelt thanks to my principal supervisor Professor Mark Viney, whose guidance was the foundation of this work. I further thank Professors Martin Genner and Mark Beaumont for their support and advice, Dr Vicky Hunt, who mentored me in laboratory techniques and showed me my way around the *Strongyloides ratti* genome, and Zoe Ballard, who taught me all I know about Python coding.

This work would not have been possible without the collaboration of the Wellcome Trust Sanger Institute. Particular thanks go to Nancy Holroyd and Matt Berriman, whose advice was invaluable for informing sequencing strategy, and to Alan Tracey, whose extensive knowledge of the *S. ratti* genome was indispensable, and who generously provided many of the plots shown in Appendix 1. as a Personal Communication. I would further like to express my gratitude to Louise Hughes, as well as the management and staff at Lamby Way Recycling Centre and Avonmouth Sewage Treatment works, for their kind assistance in my sampling endeavours.

This work was funded by the Natural Environment Research Council via a Doctoral Training Partnership. I would like to thank the research council not only for their funding, but also for the crucial training it provided.

Finally, my thanks to Jack, Lorretta, James and the whole Nerd Herd, as well as to Mum and Dad, for looking after me so well.

Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. In particular, the Gap5 and Dotter outputs presented in Appendix 1 were kindly produced by Alan Tracey of the Sanger Institute. Any views expressed in the dissertation are those of the author.

SIGNED: ..... DATE:.....

Table of Contents

Figures and Tables

Chapter 1

Chapter 2

Chapter 3

Chapter 4

List of Abbreviations

AM – Avonmouth

AMOVA – Analysis of molecular variance

BAM – Binary alignment map

CA – Cardiff

D' - Normalised coefficient of linkage disequilibrium

ddRADSeq – Double digest restriction site associated DNA sequencing

dN/dS – non-synonymous substitution / synonymous substitution ratio

DSB – Double-strand break

F_{ST} – Fixation index

HWE – Hardy-Weinberg equilibrium

iL3 – Infective third stage larva

LA – Long Ashton

LD – Linkage disequilibrium

Ne – Effective population size

Nm – Migrants per generation

PC – Principal component

PCA – Principal component analysis

PCR – Polymerase chain reaction

SDS – Sodium dodecyl sulphate

SNP – Single nucleotide polymorphism

$_{S}H$ – Shannon's information index

$^{S}H_{UA}$ – Shannon's mutual information index

VCF – Variant call format

WTSI – Wellcome Trust Sanger Institute

Chapter 1. Introduction

1.1. Parasitic nematodes

1.1.1. Parasitic nematodes in medicine and agriculture

Nematodes (phylum Nematoda), also called roundworms) are animals of the superphylum Ecdysozoa (Aguinaldo *et al.* 1997), clade Nematoida (Schmidt-Rhaea 1996), and may have evolved as early as the Precambrian (Ayala *et al.* 1998). They are characterised by a slender form, a collagenous cuticle, a radially symmetrical head with bilaterally symmetrical body and a distinctive neuronal architecture (Lee 2002), and are distinguished from the closely related Nematomorpha (horsehair worms) by the persistence of a functional gut even in the adult nematode (Eakin and Brandenburger 1974). Nematodes are both highly diverse and ubiquitous, being found in virtually every environment on earth, and play vital roles in most natural ecosystems (Borgonie *et al.* 2011, Majdi and Traunspurger 2015, Semprucci *et al.* 2015, van den Hoogen *et al.* 2019). Thus, they have evolved a wide range of anatomical, physiological and life cycle adaptations to accommodate their diverse ecologies, but morphological differences between species are often subtle such that genetic analysis is required to fully appreciate diversity (Peham *et al.* 2017). Much of what is known about nematode physiology, development and genetics is based on intense study of the tiny minority of species that are used as model species, including *Caenorhabditis elegans*, *Pristionchus pacificus*, *Nippostrongylus brasiliensis* and *Heterodera* spp. Other nematode species are comparatively poorly understood. In particular, little is known about the ecology of nematodes in natural settings.

While many nematode species are free-living, many others are parasitic, with parasitism having evolved independently at least 15 times within the phylum (Blaxter and Koutsovoulos 2015). Furthermore, parasitic nematodes are extremely common, infecting virtually every macroscopic animal and plant species. Disease associated with parasitic nematode infection is a major concern in human health, especially in less economically developed countries (Bogitsh *et al.* 2018), while infection of livestock and crops causes substantial economic losses around the globe (Nicol *et al.* 2011). Anthelmintic drugs are available and can be used to control or even eradicate parasitic nematode species (Besier *et al.* 2016, Molyneux and Sankara 2017, Smith *et al.* 2017). However, such control strategies must consider the biology, ecology and population genetics of the target species if they are to be maximally effective, especially as anthelmintic resistance becomes more prevalent (Humphries *et al.* 2012).

1.1.2. Parasitic nematodes in natural ecosystems

Despite constituting the vast majority of species of parasitic nematodes, nematodes infecting wild animals have received little scientific attention compared with nematodes infecting humans and livestock. Parasitic nematodes in natural ecosystems are of interest for several reasons. First, their great diversity means that there is likely to be life history traits among them that are not found in more

commonly-studied species, and learning of these traits helps to complete our understanding of animal life cycles. For example, a novel form of transmission, transmission among intermediate host individuals, was recently discovered in two nematode parasites of wild felids, *Aelurostrongylus abstrusus* and *Troglostrongylus brevior* (Colella *et al.* 2015). Second, parasitic nematodes can cause challenges in conservation efforts. Invasive parasitic nematodes in novel habitats may threaten naïve, native hosts, as seen in *Anguillicoloides crassus*, a parasitic nematode introduced to Europe and the Americas along with its native host, the Japanese eel (*Anguilla japonica*) that has subsequently become a major source of disease in native *Anguilla* spp. (Kirk 2003). Alternatively, environmental changes may render host species more susceptible to pre-existing nematode species, as in *Baylisascaris schroederi*, which has become a significant cause of mortality in giant pandas (*Ailuropoda melanoleucas*) (Zhang *et al.* 2008). Finally, parasitic nematodes in wild animal hosts may transmit to humans or domestic animals, causing zoonotic disease. The raccoon (*Procyon lotor*) parasite (*Baylisascaris procyonis*) causes severe, often fatal, disease in humans (Sorvillo *et al.* 2002), and many other examples of nematode-induced zoonotic disease are known (e.g. Zhu *et al.* 2001, Youn 2009, Lima dos Santos and Howgate 2011, Eiras *et al.* 2016, Wells *et al.* 2018).

The ubiquity of parasitic nematodes, and their capacity to alter the physiology, behaviour and reproductive success of their hosts, means that they likely play an important role in ecosystem dynamics across the globe. However, this role is poorly understood and has been largely overlooked in studies of ecosystem functioning (Wood and Johnson 2015, Frainer *et al.* 2018). Study of parasitic nematodes in wild hosts will not only improve our understanding of contemporary habitats but also help us predict how these habitats may change in the future. Local environmental changes, for example anthropogenic degradation of microhabitats, can alter parasitic nematode infection dynamics with unexpected consequences for host species, for whole ecosystems and for people (Weinstein and Lafferty 2017). How parasitic nematodes might respond to planet-wide changes such as climate change and pollution is similarly hard to predict, but this may lead to dramatic change in some ecosystems and may generate novel sources of zoonotic disease in humans and domestic species (Cable *et al.* 2017).

1.2. *Strongyloides ratti*

1.2.1. Introduction to *Strongyloides* spp.

Strongyloides is a genus of obligate parasitic nematodes that parasitise the intestines of terrestrial vertebrates. At least two species, *S. stercoralis* and *S. fuelleborni*, infect humans and cause strongyloidiasis, a soil-transmitted helminthiasis that affects 30-100 million people worldwide (World Health Organization 2018). *S. stercoralis* infections do not self-resolve, instead often persisting for life in the absence of anthelmintic treatment, and can be life-threatening in immunosuppressed patients (Grove 1989). Furthermore, the potential for cross-transmission among humans, dogs and wild primates makes human *Strongyloides* infections difficult to control (Viney and Ashford 1990, Hasegawa *et al.* 2016, Jaleta *et al.* 2017). *Strongyloides* spp. are also problematic in agriculture; most major livestock species have at least one *Strongyloides* spp. infecting them, and infection is associated with production losses (Thamsborg *et al.* 2017).

1.2.2. Life-history of *Strongyloides ratti*

Strongyloides ratti has a long history of use as a laboratory model of *Strongyloides* infection, where it is maintained in a natural host, the brown rat (*Rattus norvegicus*) (Viney and Kikuchi 2017). In nature, *S. ratti* has a cosmopolitan distribution. In tropical regions it shares hosts with *Strongyloides venesuelensis*, but *S. ratti* is the only *Strongyloides* sp. known to infect brown rats in temperate regions (Viney and Kikuchi 2017). The life cycle of *S. ratti* is typical of *Strongyloides* spp. and is outlined in Figure 1.1. Note that development of a female larva may proceed along one of two paths. In homogonic development, females develop into infective third-stage larvae (iL3s) that, upon infecting a host, develop into a parasitic adult morph that reproduces only asexually. In heterogonic development, a female develops into a sexual free-living adult morph (Viney *et al.* 1993, Viney 1999, Eberhardt *et al.* 2007). The parasitic and free-living female morphs are strikingly different in morphology, lifespan, reproductive mode, diet and environmental niche (Gardner *et al.* 2006). There are no parasitic males, rather all males develop into free-living adults, and can only mate with free-living females. The daughters of parasitic females may develop into free-living or parasitic adults, while the daughters of free-living females always follow parasitic development. Furthermore, all offspring of free-living females are themselves female, with only parasitic females able to produce sons. Hence the species is an obligate parthenogen and an obligate parasite with a facultative free-living, sexual generation.

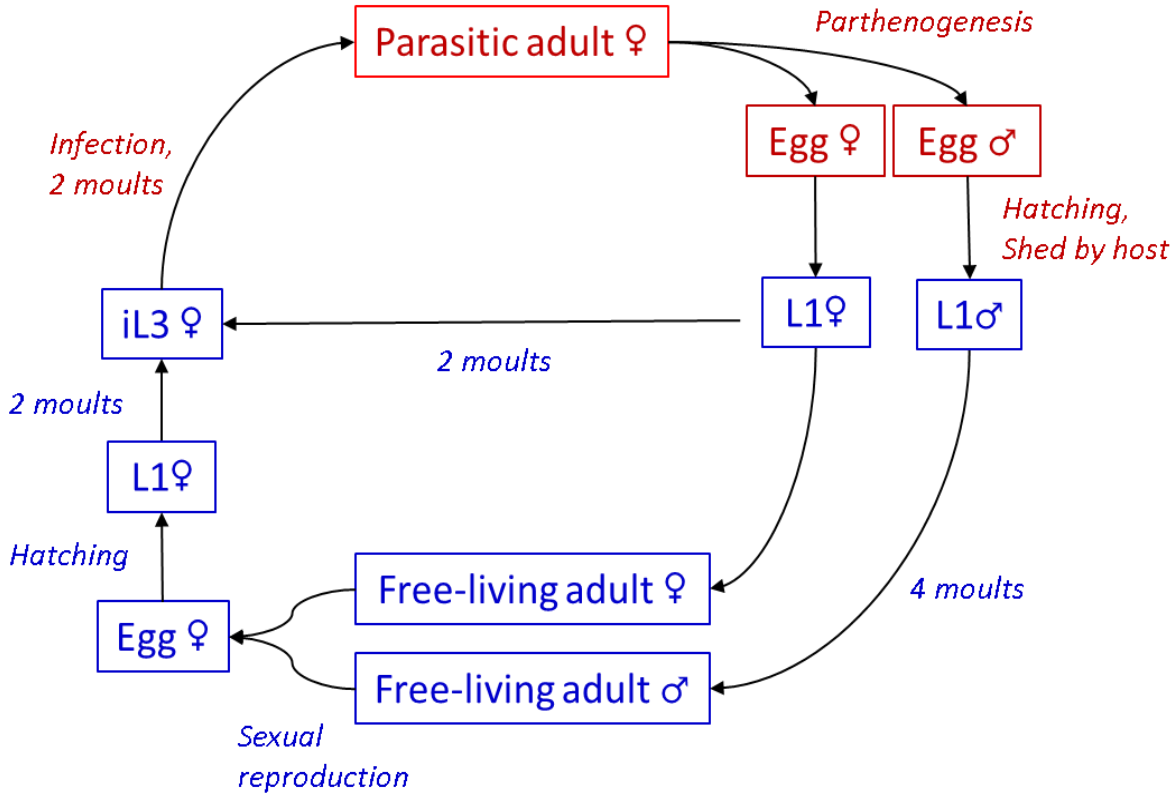


Figure 1.1: Life cycle of Strongyloides ratti. *Stages (boxed Roman text) and processes (italic text) shown in red occur within the host. Those that occur outside the host are shown in blue. Female L1s (first stage larvae) may proceed with direct, homogonic development, developing directly into infective third stage larvae (iL3s), or with indirect, heterogonic development, becoming a free-living adult female. If an iL3 successfully colonises a host, it develops into a parasitic female and begins to reproduce by mitotic parthenogenesis. Free-living females reproduce sexually by mating with males, and the all-female offspring all become iL3s.*

Sex is extremely common in nematodes and animals generally, and facultative asexuality as seen in *Strongyloides* raises questions as to the evolutionary benefits and drawbacks of sex. Sexual reproduction gives rise to new genotypes via recombination of parental genomes, and hypotheses that seek to explain the advantages of this recombination fall into two broad categories. First, recombination allows deleterious alleles to be separated from an otherwise fit genome, preventing a trans-generational accumulation of harmful mutations in the genome termed Muller's ratchet (Muller 1964, Felsenstein 1974). Second, by producing genetically diverse offspring an individual's brood collectively becomes more robust to a changing environment, as the likelihood that at least some offspring will be genetically well-suited to the changed environment is increased (Vrijenhoek 1998, West *et al.* 1999, Otto and Lenormand 2002). For a parasite, a significant source of environmental change may be a changing genetic makeup in the host population. The Red Queen hypothesis states that hosts must always adapt

to better protect themselves from parasites just as parasites must always adapt to better exploit hosts, and sexual recombination increases a parasite's odds of producing offspring better able to overcome host anti-parasite adaptations (van Valen 1973, Salathé *et al.* 2008).

However, sexual reproduction does have disadvantages. First, it reduces the relationship between parent and offspring, requiring that a sexually reproducing organism produce twice as many offspring as an asexual one in order leave an equivalent proportion of its genes in the next generation (Williams and Mitton 1973). Furthermore, sexual reproduction requires a sexual partner, which may not always be available. Both of these disadvantages can be avoided through self-fertilisation, but while-self-fertilisation can be sufficient to avoid Muller's ratchet (Morran *et al.* 2010), it does not allow for wholly new combinations of genes to be produced and instead tends to reduce within-genome diversity (Andersen *et al.* 2012). These are presumably significant drawbacks as self-fertilisation is rare in animals (Jarne and Auld 2006). Another drawback of sexual recombination is that if the environment is not changing rapidly and a particular individual is very fit, recombinant offspring may be less fit on average than would be offspring that are genetically identical to their mother (Stelzer 2015).

Facultative asexual reproduction may therefore provide the best of both worlds; asexual reproduction allows for rapid production of fit offspring when the environment is stable, and during periods of environmental change sexual reproduction can be employed to increase the likelihood that some offspring will be well-suited to the changes. Indeed, the use of sexual reproduction as a response to stress is widespread among otherwise asexual animal species (West *et al.* 2001, Stelzer and Lehtonen 2016). There is some evidence that this is the strategy *Strongyloides* has adopted. Experimentation has shown the immune state of the host rat influences the frequency of sexual *S. ratti* forms, with more males being produced and more females developing into the free-living, sexual morph when the parental host mounts a strong anti-*S. ratti* immune response (Gemmill *et al.* 1997, Harvey *et al.* 2000). Thus, if the parasitic mother *S. ratti* is well-adapted to a particular host's immune state she produces genetically identical offspring that will also be well-adapted, but sexual reproduction is used to generate new, hopefully fitter genetic combinations if the host is a hostile environment. Furthermore, selection experiments have demonstrated that the rate of heterogonic development is also a heritable trait, such that *S. ratti* may be able to respond to long periods of environmental change or environmental stability by becoming genetically more or less prone, respectively, to producing sexual offspring (Viney 1996). Indeed, natural *S. ratti* populations vary highly in the rate of sexual offspring production (Viney *et al.* 1992).

The mode of asexual reproduction in *S. ratti* is genetically mitotic, such that offspring are direct clones of their mother, barring novel mutations. Here and throughout this thesis, asexual reproduction in *S. ratti* is described as 'mitotic parthenogenesis', though it is acknowledged that the precise cellular

process involved is not known, and may involve a modified form of meiosis that results in clonal offspring (Viney 1999)

1.2.3. Genes associated with the parasitic lifestyle

Because it is possible for genetically identical *S. ratti* to develop heterogonically or homogonically, the substantial differences in phenotype between adult female morphs must come about from differences in gene expression. Transcriptomic analyses have revealed striking differences in the gene expression profiles of free-living versus parasitic female morphs, and genes upregulated in the parasitic morph putatively play a role in the parasitic lifestyle (Hunt *et al.* 2016). Many of these "parasitism genes" belong to one of four gene families that are expanded in *Strongyloides* sp. compared with a related non-parasitic species, further supporting the view that these genes are important in parasitism (Hunt *et al.* 2016).

That closely related species pairs (e.g. *S. ratti* and *S. stercoralis*; *S. venezuelensis* and *S. papillosus*) differ in the number of genes present in each "parasitism gene family" (Hunt *et al.* 2016) suggests that gene expansion is recent and may be ongoing. Individuals within species may also show copy number variation in these gene families, but this has not yet been examined. Further, the extent of sequence divergence between orthologous copies within species is unknown. Analysing the diversity of parasitism genes within *Strongyloides* sp. may provide insights into the ongoing evolution of "parasitism genes" and parasitic traits.

1.3. Population genetics

1.3.1. Introduction to population genetics

Population genetics describes how the genetic variation within a species is distributed across time and space (Hartl 1988). It affects both population-level responses to selection pressures and the differentiation and eventual speciation of distinct populations. Population genetics is therefore highly influential in shaping evolutionary outcomes (Nei 1979).

In animals, population genetic studies give insight into a species' biology and ecology, informing on traits such as dispersal, reproductive mode and variance in reproductive success (Browne 1992, Lowe and Allendorf 2010). Population genetics can be used to interrogate a species' phylogeography, revealing how populations have split, migrated and merged over geographical timescales (Kumar and Kumar 2018). Indeed, population genetics has been crucial in elucidating the route early humans took when colonising the globe (Liu *et al.* 2006), and in revealing the domestication histories of various domestic species (Larson and Burger 2013). In a similar manner, population genetics can be used to study invasive species, identifying the source population from which the invader came and tracking its spread in the new habitat (Osten-Sacken *et al.* 2018). Furthermore, population genetics can be used to ensure strategic allocation of resources in conservation projects targeting endangered species, for example by identifying habitats containing the highest levels of genetic diversity (Allen 2016).

A species' population genetics emerges as a balance between forces that cause allele frequencies in different populations to diverge, and gene flow, which makes those allele frequencies more similar (Hartl 1988).

The principal force increasing genetic differentiation is genetic drift – the change in allele frequencies from one generation to the next that results from the stochastic inheritance of parental alleles by offspring (Wright 1931). The randomness of this process means that allele frequencies change differentially in different populations. Mutation causes entirely new alleles to emerge, and as these alleles are initially unique to the population in which they first arose, they too drive genetic differentiation. Finally, if populations are subject to different selection pressures, natural selection may drive differential changes in allele frequencies, depending on which alleles are most fit in each population (Hartl and Clark 1997).

Gene flow is the migration of alleles from one population to another. This counteracts genetic differentiation by causing allele frequencies across populations to become more alike (Hartl and Clark 1997).

While these basic drivers – genetic drift, mutation, differential natural selection and gene flow – are relevant to all species, they are themselves driven by many factors, so that species differ enormously in their population genetic characteristics. For example, the rate of genetic drift is strongly affected by the effective population sizes (Ne) of the populations involved, with lower Ne leading to faster drift (Wright 1931). Ne is itself determined by many factors, such as the frequency of inbreeding, sex ratio, reproductive mode, variance in reproductive success, and the actual number of individuals present (Charlesworth 2009). On the other hand, the rate of gene flow is determined by aspects of a species' movement ecology, with vagrant species and species that disperse far from the natal site tending to have higher rates of gene flow than species that establish home ranges, especially if those home ranges are close to the site of their birth (Bohonak 1999). Terrain also heavily influences gene flow in that terrain that is hard for individuals to cross will have little gene flow across it, although carriage by wind, water or attachment to another species may enable occasional crossings of otherwise impassable terrain.

Finally, time since two populations became separated is a critical factor in determining contemporary population genetic structure. While population pairs will eventually reach an equilibrium between drift and gene flow (or will continue to diverge indefinitely if totally reproductively isolated), observations made prior to this equilibrium being reached will show the populations to be less genetically differentiated than would be expected given the level of gene flow.

1.3.2. Movement ecology and population genetics of *Rattus norvegicus*

Rodents are typically found to have a limited ability to disperse, with long-distance dispersal events being rare (Montgomery *et al.* 1991, Mikesic and Drickamer 1992, Pocock *et al.* 2005, Heiberg *et al.* 2012, Abolins *et al.*2018), and brown rats follow this trend. Most individuals in urban environments remain within 150 m of their place of birth, with a minority engaging in long-distance migration of up to several kilometres (Gardner-Santana *et al.* 2009). While dispersal distances are longer in rural areas, where food may be scarcer, they still rarely exceed 2 km (Desvars-Larrive *et al.* 2017). It is therefore unsurprising that studies into the population genetics of urban rats have found their populations to be genetically structured over distances as small as 1 km, with rats that are geographically closer often being genetically more similar too (Gardner-Santana *et al.* 2009, Kajdacsi *et al.* 2013, Desvars-Larrive *et al.* 2017, Combs *et al.* 2018).

To date, all studies investigating the population genetics of brown rats have sampled individuals destructively. Non-destructive, non-invasive sampling of wild animals, such as from faeces, avoids the need to seek ethical approval, is safer for operatives, and may allow for more individuals to be sampled in the same time-frame than does trapping. Furthermore, non-invasive sampling is less disruptive to the

sampled environment than invasive methods, and may be particularly appropriate when sampling endangered species, or species from sensitive habitats.

1.3.3. Population genetics of non-parasitic nematodes

Nematodes have low intrinsic vagility (Wallace 1968), and therefore are poor dispersers unless they make use of attachment to more mobile animals or abiotic factors such as wind or flowing water.

Pellioditis marina is a species complex of marine, free-living nematodes that feed on bacteria in rafts of decomposing macroalgae (Moens and Vincx 2000, Derycke *et al.* 2005). Consequently, the habitats used by *P. marina* species are patchy and ephemeral; a rotting patch of seaweed will eventually fully decay and no longer be available. If nematodes mix freely in the water column then population genetic structure is not expected (Ullberg 2004), but in fact while *P. marina* populations on a particular algal clump are genetically diverse, they are also genetically distinct from the populations on other algal clumps (Derycke *et al.* 2005). It is likely that once a new habitat is colonised the population becomes refractory to further immigration (Derycke *et al.* 2007), so that the genetic diversity present in each algal clump represents a randomly-sampled subset of the diversity present in the local area (Derycke *et al.* 2005). However, a similar pattern of population genetic structure over very small geographical scales was also observed in members of the *Halomonhystera disjuncta* species complex – these nematodes live in seafloor sediment rather than decaying macroalgae, suggesting that patchy and ephemeral habitat distribution may not account for the population genetics of *P. marina* nematodes alone (Derycke *et al.* 2007). Further work analysing the distribution of *H. disjuncta* in sediment is required to be certain that they are not also using patchy microhabitats that are not obvious at a macroscopic scale.

Caenorhabditis elegans is a terrestrial nematode that feeds on bacteria on rotting vegetation. Thus, wild *C. elegans* may also exploit patchy and ephemeral habitats. On very local scales (e.g. tens of metres), *C. elegans* exhibits very strong population genetic structure, indicative of repeated rounds of reproduction on a particular microhabitat with limited gene flow (Barrière and Félix 2005, 2007). However, there was no association between genetic and geographical distance at the scales analysed, contrary to what would be predicted if dispersal were mediated by low-range nematode dispersal alone (Barrière and Félix 2005, 2007). Hence, *C. elegans* population genetics is analogous to that of *P. marina* species. Rather than flowing ocean currents, attachment to larger, more mobile invertebrates may promote dispersal of *C. elegans* to newly-available microhabitats. Frequent population founding and extinction in *C. elegans* is supported by temporal population genetics studies, which show that the population present at a location is substantially different in terms of allele frequency from one year to the next (Barrière and Félix 2007).

Despite high genetic differentiation on very local scales, the population genetic structure of *C. elegans* on a global scale is one of low genetic diversity and low genetic differentiation (Barrière and Félix 2005, 2007, Dolgin *et al.* 2008, Andersen *et al.* 2012). This is largely due to a single genome-wide haplotype that dominates in all *C. elegans* populations examined (Andersen *et al.* 2012). Most *C. elegans* individuals are the product of self-fertilisation (Anderson *et al.* 2010). Repeated rounds of self-fertilisation rapidly lead to near-complete homozygosity, and as recombination among identical homologues is ineffectual, the consequence is linkage disequilibrium that extends across the whole genome (Andersen *et al.* 2012). Hence, the emergence of a high-fitness allele leads to a genome-wide selective sweep, wherein the genetic background harbouring the high-fitness allele largely replaces other genetic backgrounds. It would appear that the global population genetic structure observed in *C. elegans* is the product of such selective sweeps (Andersen *et al.* 2012). This indicates that migration in *C. elegans* is sufficient to bring high-fitness alleles to every continent, and this migration may be mediated by a commensal relationship of *C. elegans* with humans.

Pristionchus pacificus shares life-history traits with *C. elegans* in that it is predominantly self-fertilising, non-parasitic, and is dispersed by a phoretic association with larger invertebrates (Hermann *et al.* 2007). Further, *P. pacificus* also has a global distribution. *P. pacificus* has ten times the nucleotide diversity of *C. elegans* as judged by genome resequencing studies (Rödelsperger *et al.* 2014), and exhibits more population genetic differentiation than *C. elegans*, with differentiation detectable both over small geographic distances (Morgan *et al.* 2014) and among continents (Rödelsperger *et al.* 2014). Hence, it would appear that *P. pacificus* has not undergone global selective sweeps in the manner that *C. elegans* likely has, potentially because alleles that are highly fit in all environments have not emerged in *P. pacificus*, or because dispersal of *P. pacificus* over large geographical distances is much rarer than in *C. elegans*.

1.3.4. Population genetics of parasitic nematodes

The population genetics of parasitic species gives insight into the parasite's biology, ecology and infection patterns, informs predictions of how a parasite population might respond to environmental changes, and can clarify aspects of host ecology and phylogeography (Nieberding *et al.* 2004, Criscione *et al.* 2005, Gorton *et al.* 2012, Gilabert and Wasmuth 2013). Hence, studying the population genetics of parasites tells us much about the role those parasites play in their ecosystems.

Many parasitic nematode species do not have a motile extra-host stage within their life cycle, instead transmitting among hosts as eggs. Even in species that do have motile extra-host stages, the vagility of nematodes in the external environment will often be often insignificant in comparison with how far parasitic stages are carried by hosts. Therefore, host movement is considered the main driver of parasitic

nematode gene flow (Blouin *et al.* 1992, 1995, 1999, Gilabert and Wasmuth 2013, Mazé-Guilmo *et al.* 2016). However, numerous other aspects of host ecology and parasite biology also influence parasitic nematode population genetics, and the interactions between these factors remain poorly understood. This represents an ongoing gap in our understanding of parasitic nematode population genetics and population dynamics (Cole and Viney 2018).

While the population genetics of nematodes and other parasitic animals has been studied quite extensively, this information is largely based on species that infect humans or domestic species (Nadler 1995, Anderson *et al.* 1998, Criscione *et al.* 2005, Barret *et al.* 2008, Gorton *et al.* 2012, Gilabert and Wasmuth 2013, Vázquez-Prieto *et al.* 2015, Gilleard and Redman 2016, Mazé-Guilmo et al. 2016). These findings may not be fully applicable to nematodes with wild hosts.

1.3.5. Parasitic nematodes of wild rodent hosts

The population genetics of parasitic nematodes has been reviewed by the candidate and the candidate's tertiary supervisor in Cole and Viney (2018) and is reviewed here as important background information highly pertinent to the PhD project at hand.

The limited dispersal of wild rodents described in section 1.3.2 might be expected to limit gene flow in their nematode parasites, resulting in genetic structure over small geographical scales. *Trichuris arvicolae* infects arvicoline rodents (lemmings and voles), while *T. muris* infects murine rodents (rats and mice). Both *Trichuris* spp. are found throughout Europe, and *T. muris* often shares host individuals with another parasitic nematode *Heligmosomoides polygyrus*. The population genetic structures of *H. polygyrus*, *T. muris* and *T. arvicolae* are broadly similar, with a distinction between eastern and western populations, and reduced diversity in northern populations than southern ones. These patterns may reflect northwards migration of rodent hosts from refugia in southern Europe following the last ice age (Nieberding *et al.* 2004, 2005, 2006, 2008, Callejón et al. 2010, 2012, Wasimuddin et al. 2016). Population genetic structuring was stronger in *H. polygyrus* than in either of the *Trichuris* species. (Nieberding et al. 2004, Callejón et al. 2010, 2012, Wasimuddin et al. 2016). It may be that the shorter generation time of *H. polygyrus* compared with *Trichuris* spp (~14 days *vs.* 50–60 days respectively [Fahmy 1954, Gregory *et al.* 1990]) leads to faster genetic drift in the former. Alternatively, the broader host range of *Trichuris* spp. may allow these parasites to occupy or traverse a wider range of environments, potentially increasing gene flow rates compared with *H. polygyrus* (Vázquez-Prieto *et al.* 2015).

The population genetics of rodent-parasitic nematodes has also been investigated at fine geographical scales. *Neoheligmonella granjoni*,a parasite of *Mastomys* spp. multimammate mice, was isolated from *M. erythroleucus* and *M. natalensis* within a 1,300 km^2 area of Senegal, and genotyped at 10

microsatellite loci (Brouat *et al.* 2011). Alleles were found to be homogenously distributed among sampling sites, with no evidence for population genetic structure. It may be that movement of infected *M. erythroleucus*, which has comparatively high dispersal, promotes gene flow among sampling sites and among populations of the comparatively sedentary *M. natalensis* populations (Brouat *et al.* 2007, 2011). Hence, while parasitic nematodes of rodents often show strong population genetic structuring across continents, gene flow can be sufficient to genetically homogenise populations over smaller distances.

Syphacia stroma is another parasite of *A. sylvaticus*, and often forms co-infections with *H. polygyrus*. *Heligmosomoides polygyrus* shows higher genetic diversity and lower population differentiation than *S. stroma*, even when sampled from the same host individuals (Müller-Graf *et al.* 1999). *Syphacia stroma* uses haplodiploid sex determination, wherein haploid males develop from unfertilised eggs produced by diploid females, while males and females are both produced sexually in *H polygyrus*. Haplodiploidy means that fewer individuals contribute to the next generation (because males have no fathers), leading to reduced Ne. This may lead to greater genetic drift in *S. stroma* compared with *H. polygyrus*. Because these nematodes have broadly similar generation times (Morand 1996) and share a host, their different reproductive modes emerge as likely important factors behind their different population genetics.

1.3.6. Population genetics of clonally reproducing animals

In mitotic parthenogenesis there is no recombination of parental alleles, such that all of an individual's offspring are genetically identical (barring novel mutations) to each other and to their mother. In theory, strict mitotic parthenogenesis leads to perfect linkage of the entire nuclear genome, and perfect linkage between the nuclear and mitochondrial genome, such that the entire DNA content of an individual behaves as a single. non-recombining haplotype (Prugnolle and de Meeûs 2008). As this means that homologous chromosome pairs are also linked, loci in clonally reproducing organisms are unlikely to be in Hardy-Weinberg equilibrium (HWE, Stern 1943). Furthermore, in the absence of meiotic recombination, homologous chromosomes theoretically evolve separately, rapidly diverging as they accumulate novel mutations such that heterozygosity can increase drastically (Prugnolle and de Meeûs 2008). In such species long-lasting clonal lines may diverge genetically from one another even in sympatry, such that high levels of diversity relative to total species diversity are found within small geographical zones.

Bdelloid rotifers are thought to have been reproducing exclusively by mitotic parthenogenesis for at least 60 million years (Tang *et al.* 2014). Lineages within the rotifer genera *Adineta* and *Rotaria* conform to the expected population genetic structure - lineages were seen to cover wide geographical areas that overlapped with numerous other lineages without evidence of hybridisation (Fontaneto *et al.*

2008). Previously it was thought that bdelloid rotifers also showed highly divergent homologous chromosomes, as expected (Pouchkina-Stantcheva *et al.* 2007). However, this has since been shown to be due to degenerate tetraploidy, with the putative divergent homologues instead being divergent homeologues perhaps originating from different parental species in an ancient hybridisation event (Hur *et al.* 2009). In contrast, the mean genetic divergence between true homologous regions is only 4% in *Philodina roseola* (Hur *et al.* 2009), and so the expectation of exceptionally high heterozygosity is violated. It is likely that similarity is maintained by the use of each member of a homologous pair as a template for double-strand break (DSB) repair in the other (Hur *et al.* 2009). Further, while there are long linkage tracts in Bdelloid rotifers, linkage is not genome-wide, with regions of similarity occurring patchily among otherwise diverged lineages in the genus *Macrotrachela* (Signorovitch *et al.* 2015). This may be due to horizontal DNA transfer – bdelloid rotifers are able to survive prolonged desiccation during which they sustain multiple DSBs, and this is thought to facilitate incorporation of environmental DNA into the bdelloid genome (Debortoli *et al.* 2016). If such environmental DNA was shed by a conspecific, horizontal DNA transfer results.

Comparison of obligate mitotic parthenogens with closely related sexual taxa informs on the consequences of mitotic parthenogenesis for population genetic structure. The aphid species *Rhopalosiphum padi* contains obligate mitotic parthenogens as well as lineages that use both mitotic parthenogenesis and sexual reproduction. When assayed by microsatellite genotyping, fully asexual populations of *R. padi* were found to have higher heterozygosity but lower genetic diversity than partially sexual populations. Furthermore, there was greater genetic differentiation among asexual *R. padi* populations than sexual ones (Delmotte *et al.* 2002). Nevertheless, some multilocus genotypes were widespread in asexual lineages, appearing in multiple or all asexual populations. This may indicate that a lack of meiosis has led to genome-wide linkage in asexual populations, and that some clonal lineages have been able to disperse over large geographic distances (up to ~1,000km) (Delmotte *et al.* 2002).

Strict mitotic parthenogenesis is rare in animals, and most parthenogenetic species are also capable of sex. Theory and modelling suggest that only a few rounds of mitotic parthenogenesis are needed to induce high heterozygosity, and that while only a single subsequent round of sexual recombination is required to return genotype frequencies to HWE at individual loci, intra-chromosomal linkage may require further sexual reproduction to break down (de Meeûs and Balloux 2005, Prugnolle *et al.* 2005, Prugnolle and de Meeûs 2008). In the water flea *Daphnia pulicaria*, populations vary in the frequency of sex. When *D. pulicaria* populations were examined with a range of microsatellite loci, linkage among loci was only observed in populations where sex was relatively rare. Hence, it appears that repeated cycles of mitotic parthenogenesis have indeed led to extensive linkage disequilibrium in this species (Allen and Lynch 2011). Furthermore, populations with a high level of sexual reproduction had higher

numbers of multilocus genotypes, and allele frequencies were found to be more stable in these populations over a period of four years (Allen and Lynch 2011). It is likely that a low number of clonal lineages come to dominate low-sex populations in between bouts of sexual reproduction, but that these prominent clonal lineages are replaced over time, through immigration or when fitter clonal lineages emerge from recombination during infrequent sexual events. In contrast, frequent recombination in populations with a high frequency of sex likely prevents clonal lineages persisting for long enough to become dominant (Allen and Lynch 2011).

Digenean trematodes are obligate endoparasites. Most undergo cyclical sexual - asexual reproduction, with sexually reproducing adults in the definitive host (typically a vertebrate) producing miracidia that infect the mollusc intermediate host, and in turn produce numerous infective cercariae that are genetically identical to each other. In *Schistosoma mansoni* genetic diversity is typically high, and this diversity is often genetically structured over quite small geographical distances (e.g. less than 15km [Minchella *et al.* 1995, Curtis and Minchella 2000, Curtis *et al.* 2002]). Further, differentiation among *S. mansoni* infrapopulations in definitive rodent hosts has also been detected (Barral *et al.* 1996, Pinto *et al.* 1997, Sire *et al.* 2000, 2001, Curtis *et al.* 2002). *Schistosoma haematobium* showed a similar pattern to *S. mansoni*, with geographical population genetic structuring between distinct river systems (Davies *et al.* 1999) and low, but detectable, differentiation among definitive host infrapopulations (Brouwer *et al.* 2001). *Schistosoma japonicum* also showed geographical population structuring, with differentiation detected among different provinces of China and among China and the Philippines (He *et al.* 1994, Gasser *et al.* 1996, Chilton *et al.* 1999, Shrivastava *et al.* 2005). Differentiation among infrapopulations of digeneans is likely often due to clumped transmission of clone-mates to definitive hosts; in *Maritrema novaezealandensis* and *Dicrocoelium dendriticum*, differentiation among infrapopulations in the crab second intermediate hosts and ungulate definitive hosts, respectively, were lost when apparent clone-mates were treated as one individual (Keeney *et al.* 2007, van Paridon *et al.* 2016). Hence, the inclusion of mitotic parthenogenesis in the life cycle of digenean parasites alters their population genetics.

1.3.7. Population genetics of *Strongyloides* spp.

When the population genetics *Strongyloides ratti* was investigated using three nuclear loci, genetic differentiation was revealed among infrapopulations, but not among UK sampling sites ~20–250 km apart (Fisher and Viney 1998). It is likely that clumped transmission of genetically identical siblings accounts for the differentiation among infrapopulations (Paterson *et al.* 2000) in a manner analogous to the population genetics of of clonal Diginea cercariae. However, the lack of differentiation among geographical sites is surprising given that, as discussed in 1.3.5, nematodes with rodent hosts typically show strong population genetic structure. It may be that the movement of brown rats is sufficient to genetically homogenise the *S. ratti* population at the scales analysed. Alternatively, the genetic markers

used may be insufficiently sensitive to detect existent, but weak, population genetic structuring. The three genetic markers used were not in HWE in most populations examined (Fisher and Viney 1998), congruent with evidence that sex is rare in UK *S. ratti* (Viney *et al.* 1992). However, sexual reproduction was apparently sufficiently frequent to prevent linkage between the three loci used (Fisher and Viney 1998).

Strongyloides stercoralis – a parasite of humans and dogs – is among the few parasitic nematode species to have had its population genetics studied by whole genome sequencing (Kikuchi *et al.* 2016, Jaleta *et al.* 2017). In one study using nematodes isolated from humans only, genetic divergence was detected between Myanmar (where *S. stercoralis* transmission is common) and Japan (where *S. stercoralis* transmission is very rare) (Kikuchi *et al.* 2016). Further, both heterozygosity and infrapopulation differentiation were higher in Japanese *S. stercoralis*. *S. stercoralis* is the only *Stongyloides* sp. known to undergo constitutive auto-infection, wherein a portion of parthenogenetically-produced larvae establish infection and commence further parthenogenetic reproduction in the parental host without ever having left the parental host's body (Grove 1989). Hence, in Japan, repeated auto-infection of a single host without sexual reproduction or transmission among hosts may have led to independent genetic drift within infrapopulations (accounting for infrapopulation differentiation), and to divergence of homologous chromosomes within clonal lineages (accounting for high heterozygosity). In Myanmar, *S. stercoralis* transmission may homogenise allele frequencies among hosts, and ongoing sexual reproduction in the environment is likely to prevent high heterozygosity.

The application of population genetic techniques to parasitic nematodes quite commonly reveal cryptic species (Blouin 2002). When the population genetics of *S. stercoralis* in Cambodia was investigated by sequencing the small ribosomal subunit gene, three genotypes were found in sympatry, but there was no evidence of hybridisation among them (Schär *et al.* 2014). Initially this finding was taken as evidence that there are multiple cryptic *Strongyloides* species infecting humans in Cambodia, which have all been identified as *S. stercoralis* (Schär *et al.* 2014), though whole-genome sequencing data later cast doubt on this interpretation (Jaleta *et al.* 2017). However, this same whole-genome study did detect evidence for multiple *Strongyloides* spp. in dogs. One putative species was specific to dogs, while the other was detected in both dogs and humans (Jaleta *et al.* 2017). Historically all *Strongyloides* from cattle and sheep were considered to be a single species, *S. papillosus*, but when *Strongyloides* populations in German cattle and sheep were studied, two very distinct genotypes were observed, one unique to sheep and the other predominating in cattle (Eberhardt *et al.* 2008). This led to a re-description of the parasite in cattle as *Strongyloides vituli* (Eberhardt *et al*. 2008).

1.4. Objectives of this project

The objectives of this PhD are four-fold.

First, I assess infection patterns of *Strongyloides ratti* in wild populations of its natural host, the brown rat (*Rattus norvegicus*). *S. ratti* has been used as a model for *Strongyloides* infections for decades (Viney 1999), but reports concerning the prevalence and intensity of infections in nature are few (Fisher and Viney 1998, Coomansingh *et al.* 2009, Tung *et al.* 2013), and fewer still are longitudinal studies that report how these parameters change over time (Wertheim and Lengy 1964). An understanding of the ecology of a model species is important for the interpretation of results gleaned from using it.

Second, I uncover the population genetics of *S. ratti* using whole genome sequencing. Whole genome sequencing reveals the relationships among individuals and populations in the finest possible detail. However, few studies have used whole genome sequencing to assess the population genetics of parasitic nematodes, and those that have focus on species infective to humans and use few parasite individuals (Kikuchi *et al.* 2016, Small *et al.* 2016). In using 90 individuals, this study far surpasses previous nematode population genetics using whole genome sequencing in terms of the number of whole parasite genomes sequenced and is the first such study to investigate a species infecting wildlife. This study will reveal the frequency of sexual reproduction in this facultatively sexual species, inform on the consequences of asexual reproduction for parasite population genetics, and further ongoing research into the drivers of population genetics in the parasites of wildlife.

Third, I determine where selection is acting in the *S. ratti* genome. In particular, I assess whether genes associated with the parasitic lifestyle (Hunt *et al.* 2016) exhibit selection patterns distinct from the rest of the genome. This provides insight into whether and how parasitism traits are responding to selective pressures, and whether these responses are consistent across geographical space. This is the first study to investigate how genes associated with parasitism in nematodes are evolving in current, natural settings.

Finally, I investigate the population genetics of brown rats using only DNA sampled from faeces. Faecal DNA has been used previously for individual identification and population monitoring in other species (Broquet *et al.* 2007), but never in rats, where adequate DNA extraction from faeces is made challenging by the small size of the faecal pellet. Previous studies into the population genetics of rats have used destructive sampling methods (Gardner-Santana et al. 2009, Kajdacsi et al. 2013, Desvars-Larrive et al. 2017, Combs *et al.* 2018). Further, no previous study has analysed brown rat population genetics in the UK. The comparisons of the population genetics of brown rats with that of *S. ratti* in this study provides insights into how host dispersal mediates parasite population genetics. Studies that compare host and

parasitic nematode population genetics are rare but highly valuable (Nieberding et al. 2004, Brouat et al. 2007, 2011).

1.5 References

Abolins S., Lazarou L., Weldon L., Hughes L., King E. C., Drescher P. Pocock M. J. O. *et al.* (2018). The ecology of immune state in a wild mammal, *Mus musculus domesticus*. *PLoS Biology* **16:** e2003538.

Aguinaldo A. M., Turbeville J. M., Linford L. S., Rivera M. C., Garey J. R, Raff, R. A. and Lake J. A. (1997). Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* **387:**489-493.

Allen D. E. and Lynch M. (2011). The effect of variable frequency of sexual reproduction on the genetic structure of natural populations of a cyclical parthenogen. *Evolution* **66:**919-926.

Allen F. W. (2016). Genetics and the conservation of natural populations: allozymes to genomes. *Molecular Ecology* **26:**420-430.

Andersen E. C., Gerke J. C., Shapiro J. A., Crissman J. R., Ghosh R., Bloom J. S., Félix M.-A. *et al.* (2012). Chromosome-scale selective sweeps shape *Caenorhabditis elegans* genomic diversity. *Nature Genetics* **44:**285-290.

Anderson J. C., Blouin M. S. and Beech R. N. (1998). Population biology of parasitic nematodes: applications of genetic markers. *Advances in Parasitology* **41:**219-283.

Anderson J. L., Morran L. T. and Pillips P. C. (2010). Outcrossing and the maintenance of males within *C. elegans* populations. *Journal of Heredity* **101:**S62-S74.

Ayala F. J., Rzhetsky A. and Ayala F. J. (1998). Origin of the metazoan phyla: Molecular clocks confirm paleontological estimates. *Proceedings of the National Academy of Sciences USA* **95:**606-611.

Barral V., Morand S., Pointier J. P. and Théron A. (1996). Distribution of schistosome genetic diversity within naturally infected *Rattus rattus* detected by RAPD markers. *Parasitology* **113:**511-517.

Barret L. G., Thrall P. H., Burdon J. J. and Linde C. C. (2008). Life history determines genetic structure and evolutionary potential of host-parasite interactions. *Trends in Ecology and Evolution* **23:**678-685.

Barrière A. and Félix M.-A. (2005). High local genetic diversity and low outcrossing rate in *Caenorhabditis elegans* natural populations. *Current Biology* **15:**1179-1184.

Barrière A. and Félix M.-A. (2007). Temporal dynamics and linkage disequilibrium in natural *Caenorhabditis elegans* populations. *Genetics* **176:**999-1011.

Besier R. B., Kahn, L. P., Sargison N. D. and Van Wyk J. A. (2017). Diagnosis, Treatment and Management of *Haemonchus contortus* in Small Ruminants. *Advances in Parasitology* **93:**181-238.

Blaxter M. and Koutsovoulos G. (2015). The evolution of parasitism in Nematoda. *Parasitology* **142:**S26-S39.

Blouin M. S. (2002). Molecular prospecting for cryptic species of nematodes: mitochondrial DNA versus internal transcribed spacer. *International Journal for Parasitology* **32:**527-531.

Blouin M. S., Dame J. B., Tarrant C. A. and Courtney C. H. (1992). Unusual population genetics of a parasitic nematode: mtDNA variation within and among populations. *Evolution* **46:**470–476.

Blouin M. S., Liu J. and Berry R. E. (1999) Life cycle variation and the genetic structure of nematode populations. *Heredity* **83:**253–259.

Blouin M. S., Yowell C. A., Courtney C. H. and Dame J. B. (1995) Host movement and the genetic structure of populations of parasitic nematodes. *Genetics* **141:**1007–1014.

Bogitsh B. J., Carter C. E. and Oeltmann T. N. (2018) *Human Parasitology.* 5th ed. London: Academic Press.

Bohonak A. J. (1999). Dispersal, gene flow, and population structure. *Quarterly Review of Biology* **74:**21-45.

Borgoni G., García-Moyano A., Litthauer D., Bert W., Bester A., van Heerden E., Möller C *et al.* (2011). Nematoda from the terrestrial deep subsurface of South Africa. *Nature* **474:**79-82.

Broquet T., Ménard N. and Petit E. (2007). Noninvasive population genetics: a review of sample source, diet, fragment length and microsatellite motif effects on amplification success and genotyping error rates. *Conservation Genetics* **8:**249-260.

Brouat C., Loiseau A., Kane M., Bâ K. and Duplantier J. M. (2007). Population genetic structure of two ecologically distinct multimammate rats: the commensal *Mastomys natalensis* and the wild *M. erythroleucus* in south-eastern Senegal. *Molecular Ecology* **216:**2985-2997.

Brouat C., Tatard C., Machin A., Kane M., Diouf M., Bâ K. and Duplantier J. M. (2011). Comparative population genetics of a parasitic nematode and its host community: the trichostrongylid *Neoheligmonella granjoni* and *Mastomys* rodents in south-eastern Senegal. *International Journal of Parasitology* **41:**13011309.

Brouwer K. C., Ndhlovu P., Munatsi A. and Shiff C. J. (2001). Genetic diversity of a population of *Schistosoma haematobium* derived from schoolchildren in east central Zimbabwe. *Journal of Parasitology* **87:**762-769.

Browne R. A. (1992). Population genetics and ecology of *Artemia*: Insights into parthenogenetic reproduction. *Trends in Ecology and Evolution* **7:**232-237.

Cable J., Barber I., Boag B., Ellison A. R., Morgan E. R., Murray K., Pascoe E. L. *et al.* (2017). Global change, parasite transmission and disease control: lessons from ecology. *Philosophical Transactions of the Royal Society B: Biological Sciences* **372:**20160088.

Callejón R., de Rojas M., Feliú C., Balao F., Marrugal A., Henttonen H., Guevara D. *et al.* (2012). Phylogeography of *Trichuris* populations isolated from different Cricetidae rodents. *Parasitology* **139:**1795–1812.

Callejón R., de Rojas M., Nieberding C., Foronda P., Feliú C., Guevera D., Cutillas C. (2010), Molecular evolution of *Trichuris muris* isolated from different Muridae hosts in Europe. *Parasitology Research* **107:**631–641.

Charlesworth B. (2009). Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics* **10:**195-205.

Chilton N. B., Bao-Zhen Q., Bøgh H. O. and Nansen P. (1999). An electrophoretic comparison of *Schistosoma japonicum* (Trematoda) from different provinces in the People's Republic of China suggests the existence of cryptic species. *Parasitology* **119:**375-383.

Cole R. and Viney M. (2018). The population genetics of parasitic nematodes of wild animals. *Parasites and Vectors* **11:**590.

Colella V., Giannelli A., Brianti E., Ramos R. A. N., Cantacessi C., Dantas-Torres F. and Otranto D. (2015). Feline lungworms unlock a novel mode of parasite transmission. *Scientific Reports* **5:**13105.

Combs M., Byers K. A., Ghersi B. M., Blum M. J., Caccone A., Costa F., Himsworth C. G. *et al.* (2018). Urban rat races: spatial population genomics of brown rats (*Rattus norvegicus*) compared across multiple cities. *Proceedings of the Royal Society B: Biological Sciences*. **285.**

Coomansingh C., Pinckney R. D., Bhaiyat M. I., Chikweto I., Bitner S., Baffa A. and Sharma R. (2009). Prevalence of endoparasites in wild rats in Grenada. *West Indian Veterinary Journal* **9:**17-21.

Criscione C. D., Poulin R. and Blouin M. S. (2005) Molecular ecology of parasites: elucidating ecological and microevolutionary processes. *Molecular Ecology* **14:**2247–2257.

Curtis J., Sorensen R. E. and Minchella D. J. (2002). Schistosome genetic diversity: the implications of population structure as detected with microsatellite markers. *Parasitology* **125:**S51-S59.

Curtis J. and Minchella D. J. (2000). *Schistosome* population genetic structure: when clumping worms is not just splitting hairs. *Parasitology Today* **16:**68-71.

Davies C. M., Webster J. P., Krüger O., Munatsi A., Ndamba J. and Woolhouse M. E. J. (1999). Host–parasite population genetics: a cross-sectional comparison of *Bulinus globosus* and *Schistosoma haematobium*. *Parasitology* **119:**295-302.

De Meeûs T. and Balloux F. (2005). F-statistics of clonal diploids structured in numerous demes. *Molecular Ecology* **14:**2695-2702.

Debortoli N., Li X., Eyres I., Fontaneto D., Hespeels B. Tang C. Q., Flot J.-F. *et al.* (2016). Genetic exchange among bdelloid rotifers is more likely due to horizontal gene transfer than to meiotic sex. *Current Biology* **26:**723-732.

Delmotte F., Leterme N., Gautheir J.-P., Rispe C., Simon J.-C. (2002). Genetic architecture of sexual and asexual populations of the aphid *Rhopalosiphum padi* based on allozyme and microsatellite markers. *Molecular Ecology* **11:**711-723.

Derycke S., Remerie T., Vierstraerte A., Backeljau T., Vanfleteren J., Vincx M. and Moens T. (2005). Mitochondrial DNA variation and cryptic speciation within the free-living marine nematode *Pellioditis marina*. *Marine Ecology Progress Series* **300:**91-103.

Derycke S., Van Vinckt R., Vanoverbeke J, Vincx M. and Moens J. (2007). Colonization patterns of Nematoda on decomposing algae in the estuarine environment: Community assembly and

genetic structure of the dominant species *Pellioditis marina*. *Limnology and Oceanology* **52:**992-1001.

Desvars-Larrive A., Pascal M., Gasqui P., Cosson J.-F., Benoit E., Lattard V., Crespin L. *et al.* (2017). Population genetics, community of parasites, and resistance to rodenticides in an urban brown rat (*Rattus norvegicus*) population. *PLoS One* **12:**e0184015.

Dolgin E. S., Félix M.-A. and Cutter A. D. (2008). Hakuna Nematoda: genetic and phenotypic diversity in African isolates of *Caenorhabditis elegans* and *C. briggsae*. *Heredity* **100:**304-315.

Eakin R. M. and Brandenburger J. L. (1974). Ultrastructural features of a Gordian worm (Nematomorpha). *Journal of Ultrastructure Research* **46:**351-374.

Eberhardt A. G., Mayer W. E., Bonfoh B. and Streit A. (2008). The *Strongyloides* (Nematoda) of sheep and the predominant *Strongyloides* of cattle form at least two different, genetically isolated populations. *Veterinary Parasitology* **157:**89-99

Eberhardt A. G., Mayer W. E. and Streit A. (2007). The free-living generation of the nematode *Strongyloides papillosus* undergoes sexual reproduction. *International Journal of Parasitology* **37:**989-1000.

Eiras J. C., Pavanelli G. C., Takemoto R. M., Yamaguchi M. U., Karkling L. C. and Nawa Y. (2016). Potential risk of fish-borne nematode infections in humans in Brazil – Current status based on a literature review. *Food and Waterborne Parasitology* **5:**1-6.

Fahmy M. A. M. (1954). An investigation on the life cycle of *Trichuris muris*. *Parasitology* **44:**50–57.

Felsenstein J. (1974). The evolutionary advantage of recombination. *Genetics* **78:**737-756.

Fisher M. C. and Viney M. E. (1998). The population genetic structure of the facultatively sexual parasitic nematode *Strongyloides ratti* in wild rats. *Proceedings of the Royal Society B: Biological Sciences*. **265:**703–709.

Fontaneto D., Barraclough T. G., Chen K., Ricci C. and Henirou E. A. (2008). *Molecular Ecology* **17:**3136-3145.

Frainer A., McKie B, G., Amundsen P.-A. and Lafferty K. D. (2018). Parasitism and the biodiversity-functioning relationship. *Trends in Ecology and Evolution* **33:**260-268.

Gardner M. P., Gems D. and Viney M. (2006). Extraordinary plasticity in aging in *Strongyloides ratti* implies a gene-regulatory mechanism of lifespan evolution. *Aging Cell* **5:**315-323.

Gardner-Santana L. C., Norris D. E., Fornadel C. E., Hinson E. R., Klein S. L. and Glass G.E. (2009). Commensal ecology, urban landscapes, and their influence on the genetic characteristics of city-dwelling Norway rats (*Rattus norvegicus*). *Molecular Ecology*. **18:**2766–2778.

Gasser R. B., Bao-Zhen Q., Nansen P., Johansen M. V. and Bøgh H. O. (1996). Use of RAPD for the detection of genetic variation in the human blood fluke, *Schistosoma japonicum*, from mainland China. *Molecular and Cellular Probes* **10:**353-358.

Gemmill A. W., Viney M. and Read A. F. (1997). Host immune status determines sexuality in a parasitic nematode. *Evolution* **51:**393-401.

Gilabert A. and Wasmuth J. D. (2013) Unravelling parasitic nematode natural history using population genetics. *Trends in Parasitology* **29:**438-448.

Gilleard J. S. and Redman E. (2016). Genetic diversity and population structure of *Haemonchus contortus*. *Advances in Parasitology* **93:**31-68.

Gorton M. J., Kasl E. L., Detwiller J. T., Criscione C. D. (2012). Testing local-scale panmixia provides insights into the cryptic ecology, evolution, and epidemiology of metazoan animal parasites. *Parasitology* **139:**981-997.

Gregory R. D., Keymer A. E. and Clarge J. R. (1990) Genetics, sex and exposure: the ecology of *Heligmosomoides polygyrus* (Nematoda) in the wood mouse. *Journal of Animal Ecology* **59:**363–378.

Grove D. I. (1989). *Strongyloidiasis: A Major Roundworm Infection of Man*. London: Taylor and Francis

Hartl D. L. (1988). *A Primer in Population Genetics.* Sunderland: Sinauer.

Hartl D. L. and Clark A. G. (1997). *Principles of Population Genetics.* 4th ed. Sunderland: Sinauer.

Harvey S. C., Gemmill A. W., Read A. F. and Viney M. (2000). The control of morph development in the parasitic nematode *Strongyloides ratti*. *Proceedings of the Royal Society B: Biological Sciences* **267:**2057-2063.

Hasegawa H., Kalousova B., McLennan M. R., Modry D., Profousova-Psenkova I., Shutt-Phillips K. A., Todd A. *et al.* (2016). *Strongyloides* infections of humans and great apes in Dzanga-Sangha Protected Areas, Central African Republic and in degraded forest fragments in Bulindi, Uganda. *Parasitology International* **65:**367-370.

He Y.-X., Hu Y.-Q., Yu Q.-F., Ni C.-H., Xue H.-C., Qiu L.-S. and Mi X. (1994). Strain complex of *Schistosoma japonicum* in the mainland of China. *Southeast Asian Journal of Tropical Medicine and Public Health* **25:**232-242.

Heiberg A.-C., Sluydts V. and Leirs H. (2012). Uncovering the secret lives of sewer rats (*Rattus norvegicus*): movements, distribution and population dynamics revealed by a capture–mark–recapture study. *Wildlife Research* **39:**202-219.

Hermann M., Mayer W. E., Hong R. L., Kienle S., Minasaki R. and Sommer R. J. (2007). The nematode *Pristionchus pacificus* (Nematoda: Diplogastridae) is associated with the oriental beetle *Exomala orientalis* (Coleoptera: Scarabaeidae) in Japan. *Zoological Science* **24:**883-889.

Humphries D., Nguyen S., Boakye D., Wilson M. and Cappello M. (2012). The promise and pitfalls of mass drug administration to control intestinal helminth infections. *Current Opinions in Infectious Diseases* **25:**584-589.

Hunt V. L., Tsai I. J., Coghlan A., Reid A. J., Holroyd N., Foth B. J., Tracey A. *et al.* (2016). The genomic basis of parasitism in the *Strongyloides* clade of nematodes. *Nature Genetics* **48:**299-307.

Hur J. H., Van Doninck K., Mandigo M. L. and Meselson M. (2009). Degenerate tetraploidy was established before bdelloid rotifer families diverged. *Molecular Biology and Evolution* **26:**375-383

Jaleta T. G., Zhou S., Bemm F. M., Schär F., Khieu V., Muth S., Odermatt P. *et al.* (2017). Different but overlapping populations of *Strongyloides stercoralis* in dogs and humans—Dogs as a possible source for zoonotic strongyloidiasis. *PLoS Neglected Tropical Diseases* **11:**e0005752.

Jarne P and Auld J. R. (2006). Animals mix it up too: The distribution of self-fertilization among hermaphroditic animals. *Evolution* **60:**1816-1824.

Kajdacsi B., Costa F., Hyseni C., Porter F., Brown J., Rodrigues G., Farias H. *et al.* (2013). Urban population genetics of slum-dwelling rats (*Rattus norvegicus*) in Salvador, Brazil. *Molecular Ecology* **22:**5056-5070.

Keeney D. B., Waters J. M. and Poulin R. (2007). Clonal diversity of the marine trematode *Maritrema novaezealandensis* within intermediate hosts: the molecular ecology of parasite life cycles. *Molecular Ecology* **16:**431-439.

Kikuchi T., Hino A., Tanaka T., Aung M. P. P. T. H. H. A., Afrin T., Nagayasu E., Tanaka R. *et al.* (2016). Genome-wide analyses of individual *Strongyloides stercoralis* (Nematoda: Rhabditoidea) provide insights into population structure and reproductive life cycles. *PLoS Neglected Tropical Diseases* **10:**e0005253.

Kirk R. S. (2003). The impact of *Anguillicola crassus* on European eels. *Fisheries Management and Ecology* **10:**385-394.

Kumar R. and Kumar V. (2018). A review of phylogeography: biotic and abiotic factors. *Geology, Ecology and Landscapes* **2:**268-274.

Larson G. and Burger J. (2013). A population genetics view of animal domestication. *Trends in Genetics* **29:**197-205.

Lee D. L. (2002) *The Biology of Nematodes.* Florida: CRC Press.

Lima dos Santos C. A. M. and Howgate P. (2011). Fishborne zoonotic parasites and aquaculture: A review. *Aquaculture* **318:**253-261.

Liu H., Prugnolle F., Manica A. and Balloux F. (2006). A geographically explicit genetic model of worldwide human-settlement history. *American Journal for Human Genetics* **79:**230-237.

Lowe W. H. and Allendorf F. W. (2010). What can genetics tell us about population connectivity? *Molecular Ecology* **19:**3038-3051.

Majdi N. and Traunspurger W. (2015). Free-living nematodes in the freshwater food web: A Review. *The Journal of Nematology* **47:**28-44.

Mazé-Guilmo E., Blanchet S., McCoy K. D. and Loot, G. (2016). Host dispersal as the driver of parasite genetic structure: a paradigm lost? *Ecology Letters* **19:**336-347.

Mikesic D. G. and Drickamer L. C. (1992). Factors affecting home-range size in house mice (*Mus musculus domesticus*) living in outdoor enclosures. *American Midland Naturalist*. **127**:31–40.

Minchella D. J., Sollenberger K. and Pereira de Souza C. (1995). Distribution of *Schistosome* genetic diversity within molluscan intermediate hosts. *Parasitology* **111:**217-220.

Moens T. and Vincx M. (2000). Temperature, salinity and food thresholds in two brackish-water bacterivorous nematode species: assessing niches from food absorption and respiration experiments. *Journal of Experimental Marine Biology and Ecology* **243:**137-154.

Molyneux D., Sankara D. P. (2017). Guinea worm eradication: Progress and challenges— should we beware of the dog? *PLoS Neglected Tropical Diseases* **11:**e0005495.

Montgomery W. I., Wilson W. L., Hamilton R. and McCartney P. (1991). Dispersion in the wood mouse, *Apodemus sylvaticus*: variable resources in time and space. Journal of *Animal Ecology* **60:**179-192.

Morand S. (1996) Life-history traits in parasitic nematodes: a comparative approach for the search of invariants. *Functional Ecology* **10:**210-218.

Morgan K., McGaughran A., Ganesham S., Herrman M. and Sommer R. J. (2014). Landscape and oceanic barriers shape dispersal and population structure in the island nematode *Pristionchus pacificus. Biological Journal of the Linnean Society* **112:**1-15.

Morran L. T., Ohdera A. H and Phillips P. C. (2010). Purging deleterious mutations under self fertilization: paradoxical recovery in fitness with increasing mutation rate in *Caenorhabditis elegans*. *PLoS ONE* **5:**e14473.

Muller H. J. (1964). The relation of recombination to mutational advance. *Mutation Research.* **106:**2-9.

Müller-Graf C. D., Durand P., Feliu C., Hugot J.-P., O'Callaghan C. J., Renaud F., Santalla F. *et al.* (1999). Epidemiology and genetic variability of two species of nematodes (*Heligmosomoides polygyrus* and *Syphacia stroma*) of *Apodemus* spp. Parasitology **118:**425-432.

Nadler S. A. (1995). Microevolution and the genetic structure of parasite populations. *Journal of Parasitology* **81:**395-403.

Nei M. (1979). *Molecular population genetics and evolution.* Amsterdam: North-Holland Publishing.

Nicol J. M., Turner S. J., Coyne D. L., den Nijs L., Hockland S. and Tahna Maafi Z. (2011). Current nematode threats to world agriculture. In: Jones, J., Gheysen, G. and Fenoll, C. ed. *Genomics and molecular genetics of plant-nematode interactions*. Berlin: Springer.

Nieberding C., Durette-Desset M.-C., Vanderpoorten A., Casanova J. C., Ribas A., Deffontaine V., Feliu C. et al. (2008). Geography and host biogeography matter for understanding the phylogeography of a parasite. *Molecular Phylogenetics and Evolution* **47:**538-554.

Nieberding C., Libois R., Douady C. J., Morand S. and Michaux J. R. (2005). Phylogeography of a nematode (*Heligmosomoides polygyrus*) in the western Palearctic region: persistence of northern cryptic populations during ice ages? *Molecular Ecology* **14:**765-779.

Nieberding C., Morand S., Libois R., Michaux J. R. (2004). A parasite reveals cryptic phylogeographic history of its host. *Proceedings of the Royal Society B: Biological Sciences* **271:**2559–2568.

Nieberding C., Morand S., Libois R. and Michaux JR. (2006). Parasites and the island syndrome: the colonization of the western Mediterranean islands by *Heligmosomoides polygyrus* (Dujardin, 1845). *Journal of Biogeography*. **33:**1212–1222.

Osten-Sacken N., Heddegrott M., Schleimer A., Anheyer-Behmenburg H. E., Runge M., Horsburgh G. J., Camp L. *et al*. (2018). Similar yet different: co-analysis of the genetic diversity and structure of an invasive nematode parasite and its invasive mammalian host. *International Journal for Parasitology*. **48:**233–243.

Otto S. P. and Lenormand T. (2002). Resolving the paradox of sex and recombination. *Nature Reviews Genetics* **3:**252-261.

Paterson S., Fisher M. C. and Viney M. E. (2000) Inferring infection processes of a parasitic nematode using population genetics. *Parasitology* **120:**185–194.

Peham T., Steiner F. M., Schlick-Steiner B. C. and Arthofer W. (2017). Are we ready to detect nematode diversity by next generation sequencing? *Ecology and Evolution* **7:**4147-4151.

Pinto P. M., Brito C. F., Passoss L. K., Tendler M. and Simpson A. J. (1997). Contrasting genomic variability between clones from field isolates and laboratory populations of *Schistosoma mansoni*. *Memorias do Instituto Oswaldo Cruz* **92:**409-414.

Pocock J. O., Hauffe H. C. and Searle J. B. (2005). Dispersal in house mice. *Biological Journal of the Linnean Society* **84:**565-583.

Pouchkina-Stantcheva N. N., McGee B. M., Boschetti C., Tolleter D., Chakrabortee S., Popova A. V., Meersman F. (2007). Functional divergence of former alleles in an ancient asexual invertebrate. *Science* **318:**268-271.

Prugnolle D. and de Meeûs T. (2008). The impact of clonality on parasite population genetic structure. *Parasite* **15:**455-457.

Prugnolle F., Rose D., Théron A. and de Meeûs T. (2005). F-statistics under alternation of sexual and asexual reproduction: a model and data from schistosomes (platyhelminth parasites). *Molecular Ecology* **14:**1355-1365.

Rödelsperger C., Neher R. A., Weller A. M., Eberhardt G. Witte H., Mayer W. E., Dietriech C. *et al.* (2014). Characterization of genetic diversity in the nematode *Pristionchus pacificus* from population-scale resequencing data. *Genetics* **196:**1153-1165.

Salathé M., Kouyos R. D. and Bonhoeffer S. (2008). The state of affairs in the kingdom of the Red Queen. *Trends in Ecology and Evolution* **23:**439-445.

Schär F., Guo L., Streit A., Khieu V., Sinuon,M., Marti H. and Odermatt P. (2014). *Strongyloides stercoralis* genotypes in humans in Cambodia. *Parasitology International* **63:**533-536.

Schmidt-Rhaesa, A. (1996) The nervous system of *Nectonema munidae* and *Gordius aquaticus*, with implications for the ground pattern of the Nematomorpha.

Semprucci F., Losi V. and Moreno M. (2015). A review of Italian research on free-living marine nematodes and the future perspectives on their use as Ecological Indicators (EcoInds). *Mediterranean Marine Science* **16:**352-365.

Shrivastava J., Qian B. Z., McVean G. and Webster J. P. (2005). An insight into the genetic variation of *Schistosoma japonicum* in mainland China using DNA microsatellite markers. *Molecular Ecology* **14:**839-849.

Signorovitch A., Hur J., Gladushev E. and Meselson M. (2015). Allele sharing and evidence for sexuality in a mitochondrial clade of bdelloid. *Genetics* **200:**581-590.

Sire C., Durand P., Pointier J.-P. and Théron A. (2001). Genetic diversity of *Schistosoma mansoni* within and among individual hosts (*Rattus rattus*): infrapopulation differentiation at microspatial scale. *International Journal of Parasitology* **31:**1609-1616.

Sire C., Langand J., Barral V. and Théron A. (2000). Parasite (*Schistosoma mansoni*) and host (*Biomphalaria glabrata*) genetic diversity: population structure in a fragmented landscape. *Parasitology* **122:**545-554.

Small S. T., Reimer L. J., Tisch D. J., King C. L., Christensen B. M., Siba P. M., Kazura J. M. *et al.* (2016). Population genomics of the filarial nematode parasite *Wuchereria bancrofti* from mosquitoes. *Molecular Ecology* **25:**1465-1477.

Smith G. R., Fletcher J. D., Marroni V., Kean J. M., Stronger L. D. and Vereijssen J. (2017). Plant pathogen eradication: determinants of successful programs. *Australian Plant Pathology.* **46:**277-284.

Sorvillo F., Ash L. R., Berlin O. G. W., Yatabe J., Degiorgio, C. and Morse S. A. (2002). *Baylisascaris procyonis*: An Emerging Helminthic Zoonosis. *Emerging Infectious Diseases* **8:**355-359.

Stelzer C.-P. (2015). Does the avoidance of sexual costs increase fitness in asexual invaders? *PNAS* **112:**8851-8858.

Stelzer C.-P. and Lehtonen J. (2016). Diapause and maintenance of facultative sexual reproductive strategies. *Philosophical Transactions of the Royal Society B: Biological Sciences* **371:**20150536

Stern, C. (1943). The Hardy-Weinberg law. *Science* **97:**137-138.

Tang C. Q., Obertegger U., Fontaneto D. and Barraclough T. G. (2014). Sexual species are separated by larger genetic gaps than asexual species in rotifers. *Evolution* **68:**2901-2916.

Thamsborg S. M., Ketzis J., Horii Y. and Matthews J. B. (2017). *Strongyloides* spp. infections of veterinary importance. *Parasitology* **144:**274-284.

Tung K.-C., Hsiao F.-C., Wang K.-S., Yang C.-H. and Lai C.-H. (2013). Study of the endoparasitic fauna of commensal rats and shrews caught in traditional wet markets in Taichung City, Taiwan. *Journal of Microbiology, Immunology and Infection* **46:**85-88.

Ullberg J. (2004). Dispersal in free-living, marine, benthic nematodes: passive or active processes? PhD Dissertation, Stockholm University, Sweden.

Van den Hoogen J., Giesen S., Routh D., Ferris H., Traunspurger W., Wardle D. A., de Goede R. G. M. *et al.* (2019). Soil nematode abundance and functional group composition at a global scale. *Nature* **572:**194-198.

Van Paridon B. J., Goater C. P., Gilleard J. S. and Criscione C. D. (2016). Characterization of nine microsatellite loci for *Dicrocoelium dendriticum*, an emerging liver fluke of ungulates in North America, and their use to detect clonemates and random mating. *Molecular and Biochemcial Parasitology* **207:**19-22.

Van Valen L. (1993). A new evolutionary law. *Evolutionary Theory* **1:**1-30.

Vázquez-Prieto S., Vilas R., Paniagua E. and Ubeira F. M. (2015). Influence of life history traits on the population genetic structure of parasitic helminths: a minireview. *Folia Parasitologica* **62:**060.

Viney M. (1996). Developmental switching in the parasitic nematode *Strongyloides ratti*. *Proceedings of the Royal Society B: Biological Sciences*. **263:**201-208.

Viney M. (1999). Exploiting the life cycle of *Strongyloides ratti*. *Parasitology Today* **15:**231-235.

Viney M. and Ahsford R. W. (1990). The use of Isoenzyme electrophoresis in the taxonomy of *Strongyloides. Annals of Tropical Medicine and Parasitology* **84:**33-47.

Viney M. and Kikuchi T. (2017). *Strongyloides ratti* and *S. venezuelensis* – rodent models of *Strongyloides* infection. *Parasitology* **66:**285-294.

Viney M., Matthews B. E. and Walliker D. (1992). On the biological and biochemical nature of cloned populations of *Strongyloides ratti*. *Journal of Helminthology* **66:**45-52.

Viney M., Matthews B. E. and Walliker D. (1993). Mating in the nematode parasite *Strongyloides ratti*: proof of genetic exchange. *Proceedings of the Royal Society B: Biological Sciences.***254:**213-219.

Vrijenhoek R. C. (1998). Animal clones and diversity. *BioScience* **48:**617-628.

Wallace H. R. (1968). The dynamics of nematode movement. *Annual Review of Phytopathology* **6:**91-114.

Wasimuddin, Bryja J., Ribas A., Baird S. J. E., Piálek J. and Goüy de Bellocq J. (2016) Testing parasite 'intimacy': the whipworm *Trichuris muris* in the European house mouse hybrid zone. *Ecology and Evolution* **6:**2688–2701.

Weinstein S. B. and Lafferty K. D. (2017). How do humans affect wildlife nematodes? *Trends in Parasitology* **31:**222-227.

Wells K., Gibson D. I., Clark N. J., Ribas A., Morand S. and McCallum H. I. (2018). Global spread of helminth parasites at the human–domestic animal–wildlife interface. *Global Change Biology* **24:**3254-3265.

Wertheim G. and Lengy J. (1964). The Seasonal Occurrence of *Strongyloides ratti* Sandground, 1925 and of *S. venezuelensis* Brumpt, 1934 in a Population of *Rattus norvegicus*. *Journal of Helminthology* **38:**393-398.

West S. A., Gemmill A. W., Graham A., Viney M. and Read A. F. (2001). Immune stress and facultative sex in a parasitic nematode. *Journal of Evolutionary Biology* **14:**333-337.

West S. A., Lively C. M. and Read A. F. (1999). A pluralist approach to sex and recombination *Journal of Evolutionary Biology* **12:**1003-1012.

Williams G. C. and Mitton J. B. (1973) Why reproduce sexually? *Journal of Theoretical Biology* **39:**545-554.

Wood C. L. and Johnson P. T. J. (2015). A world without parasites: exploring the hidden ecology of infection. *Frontiers in Ecology and the Environment* **13:**425-434.

World Health Organisation (2018). *Strongyloidiasis.* Retrieved from https://www.who.int/intestinal_worms/epidemiology/strongyloidiasis/en/ 23/12/2018.

Wright S. (1931). Evolution in Mendelian populations. *Genetics* **16:**97–159.

Youn H. (2009). Review of zoonotic parasites in medical and veterinary fields in the Republic of Korea. *Korean Journal of Parasitology* **47:**S133-141.

Zhang J.-S., Daszak P., Huang H.-L., Yang G.-Y., Kilpatrick A. M. and Zhang S. (2008). Parasite threat to panda conservation. *EcoHealth* **5:**6-9.

Zhu X. Q., Gasser R. B., Chilton N. B. and Jacobs D. E. (2001). Molecular approaches for studying ascaridoid nematodes with zoonotic potential, with an emphasis on *Toxocara* species. *Journal of Helminthology* **75:**101-108.

Chapter 2 Infection patterns in *Strongyloides ratti*

2.1 Introduction

Parasite infection patterns concern the prevalence and intensity of a parasite in host population(s), and how these parameters change over time. Infection patterns are an important factor in parasite transmission and in the epidemiology of parasite-induced disease, and are therefore relevant to human health, agriculture, ecosystem functioning and conservation (McCallum and Dobson 1995, Dunn *et al.* 2010, Hawley and Altizer 2011, Taylor 2012, Cable *et al.* 2017). However, predicting the infection patterns of parasites is challenging (Poulin 2007).

The parasites of rodents may present zoonotic disease risks to humans or threaten native rodent species if introduced with invading hosts (Easterbrook *et al.* 2007). Consequently, substantial effort has gone into cataloguing the helminth parasites of common rodents such as the brown rats (*Rattus norvegicus*), and many of these have observed *Strongyloides* sp. nematodes (e.g. Coomansingh *et al.* 2009, Tung et al. 2013, O Simões *et al.* 2014). However, few studies have analysed *S. ratti* populations in depth (Fisher and Viney 1998), and there has been little work into the patterns of *S. ratti* populations over time (Wertheim and Lengy 1964).

Consequently, little is known about the factors that influence infection patterns in *S. ratti*. Such information would be a novel insight into seasonal variation in prevalence and intensity of a nematode in a wild host. Furthermore, an understanding of the infection patterns of a parasite contextualises other aspects of that parasite population, such as its population genetic structure and its observed response to selection pressures.

In this chapter, I describe the methods with which *S. ratti* were sampled from the UK wild brown rat colonies and explore their infection patterns over the course of a year.

2.2 Materials and methods

2.2.1 Collection of *Strongyloides ratti* from wild *Rattus norvegicus*

2.2.1.1 Sampling sites and seasons

Three sampling sites, Avonmouth (AM), Cardiff (CA) and Long Ashton (LA) (Table 2.1, Figure 2.1), were chosen due to their high-density rat populations. The presence of a high-density rat population was determined by i) reports of frequent rat sightings by landowners, and ii) the presence of abundant, fresh rat faecal pellets.

Figure 2.1: Satellite image showing parts of Wales and England, corresponding to the region highlighted in the inset. Sampling sites for collection of wild rat faeces in the UK are indicated by yellow markers and accompanied by site codes. The Severn tunnel and two Severn bridges highlighted are potential means by which rat hosts could cross the Severn Estuary. Distances from CA to AM, AM to LA and LA to CA are 30 km, 9.7 km and 32 km respectively.

Table 2.1: Sites and sampling seasons for collection of wild rat faeces.

Site (code)	**Coordinates**	**Type**
Cardiff (CA)	51°29'54"N 3°07'25"W	Industrial
Avonmouth (AM)	51°30'43"N 2°40'15"W	Industrial
Long Ashton (LA)	51°26'08"N 2°38'41"W	Rural
Season	**Sampling start date**	**Sampling End date**
Spring	24th February 2017	March 2017
Summer	June 2017	June 2017
Autumn	September 2017	November 2017
Winter	December 2017	23rd February 2018

The Avonmouth (AM) site is a sewage treatment plant at 51°30'43"N 2°40'15"W and has a patchy composition of outdoor grass and asphalt areas with some small storage buildings. AM pellets were taken from outdoor sites predominantly. The Cardiff (CA) site is a household waste processing centre at 51°29'54"N 3°07'25"W, the surface of which is almost entirely concrete. Most CA pellets were taken from inside a single large warehouse, with some taken from just outside this building. The Long Ashton (LA) site is a dairy farm at 51°26'08"N 2°38'41"W the flooring of which is assorted grass, bare mud and concrete. LA pellets were mostly taken from indoor areas such as storage sheds. All three sampling sites were operational and therefore had a continual human presence.

The minimum distance between sampling sites was 9.7km (Figure 2.1), well beyond the typical dispersal distance of rats, which is less than 1 km (Gardner-Santana et al. 2009). Hence, it is expected that migration of individual rats among these sampling sites is rare. Sampling occurred between March 2017 and February 2018 in four discrete sampling seasons that correspond to calendar seasons (Table 2.1).

2.2.1.2 Collection of wild *Rattus norvegicus* faecal pellets

To sample fresh faecal pellets, pellets found at each sampling site were individually collected into 7 mL bijous that were then sealed. Freshness of the faecal pellet was important to ensure that any *S. ratti* shed with the pellet had not had time to migrate from the pellet nor been killed by desiccation, and to reduce the number of free-living nematodes invading the pellet. A pellet was determined to be fresh if it was still moist and had no visible fungal growth. Collection was carried out as early in the morning as possible, on the assumption that the nocturnal habits of brown rats (Chitty 1942) would lead to fresh faeces being deposited predominantly during the night. Following collection, faeces were transported immediately to the laboratory. A total of 308 pellets were collected.

2.2.1.3 Processing of faecal pellets

Faeces were processed immediately upon arriving at the laboratory. Approximately a third of each faecal pellet was removed and placed into a 1.5 mL microcentrifuge tube and stored at -80°C. The remainder of each pellet was cultured for the isolation of *S. ratti* infective third stage larvae (iL3s) as described previously (Viney *et al.* 1992). Briefly, individual partial pellets were placed in 5 cm diameter watch glasses, and these watch glasses were placed within 9 cm diameter glass Petri dishes. Distilled water was added to the Petri dish and to the watch glass, so that the watch glass and pellet were both partially submerged (Figure 2.2). Once *S. ratti* reach the iL3 stage, they migrate away from the faecal mass and in so doing move from the watch glass and into the Petri dish, where they remain alive for some weeks.

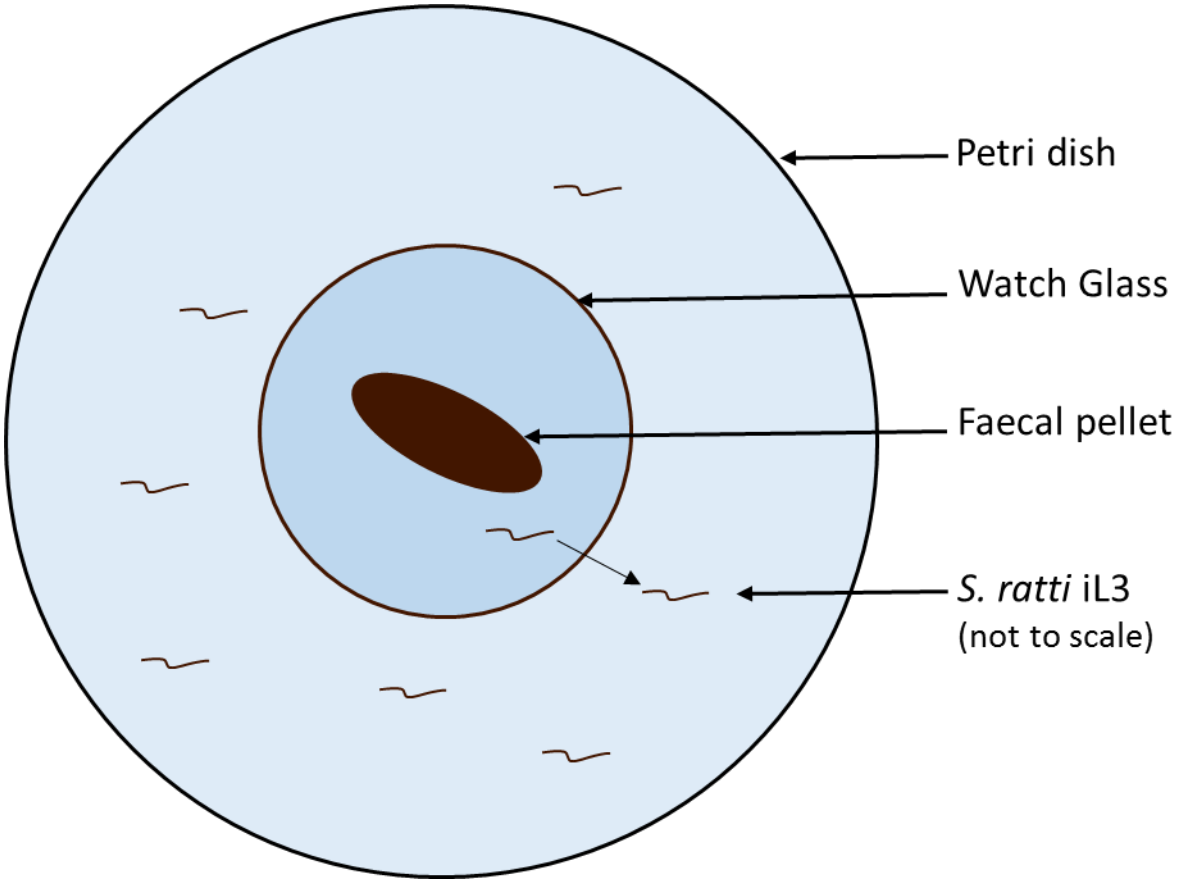


Figure 2.2: Schematic diagram of faecal cultures used for the collection of Strongyloides ratti *from rat faeces. Petri dish and watch glass contain water. Upon reaching the iL3 stage,* S. ratti *migrate away from the faecal pellet and fall into the Petri dish, from which they are collected.*

2.2.1.4 Isolation of *Strongyloides ratti* from faecal pellets

Faecal cultures were maintained at 19ºC. Under these conditions *S. ratt*i newly shed from the host reach the iL3 stage within 3 days. After 3 to 14 days of culture, the watch glass containing the faecal mass was removed and the contents of the Petri dish were examined under a dissecting microscope (Zeiss). *Strongyloides* sp. iL3s were identified by eye (Speare 1989), and as *S. ratti* is the only *Strongyloides* sp. known to infect brown rats in the UK (Viney and Kikuchi 2017), all were assumed to be *S. ratti*. Worms were transferred from the Petri dish to a clean watch glass containing distilled water, and subsequently washed twice in distilled water, incubated in 1% w/v sodium dodecyl sulphate (SDS) for 3 minutes to reducing adhering bacteria, and then washed twice more in distilled water. Each iL3 was then transferred individually along with a drop of water (~5 µL) into either a 0.7 mL microcentrifuge tube, or into a well of a 96-well plate. Subsequently worms were placed into storage at -80°C. A total of 10,471 iL3s were collected.

2.2.2 Data analysis

In this work, 'infected pellets' are defined as pellets that yielded at least one iL3 under the culture conditions described above. 'Infection prevalence' is the percentage of pellets that were infected. 'Intensity of infection' describes the number of iL3s isolated from a single pellet. 'Mean intensity of infection' refers to the mean intensity of all infected pellets in a particular category, excluding those that are not infected.

All statistical analyses were carried out using R version 3.3.3 (R Core Team) with default libraries. Chi-square tests were used to assess the overall effect of sampling site or sampling season on prevalence. When the effects of site within season, and season within site, on infection prevalence were tested, the expected values in some site-season combinations were too low (< 5) for chi-square tests to be accurate, and Fisher's exact tests were used instead.

Effect of sampling site or sampling season on intensity of infection was investigated through analysis of variance (ANOVA). Within seasons where site had a significant effect on intensity of infection according to ANOVA, pairwise comparisons of sites were made with t-tests. In all cases, intensity was found not to be normally distributed among pellets, due to an overabundance of infected pellets with low intensity values (Figure 2.3). Consequently, intensity values were log-transformed prior to all ANOVA and t-test analyses.

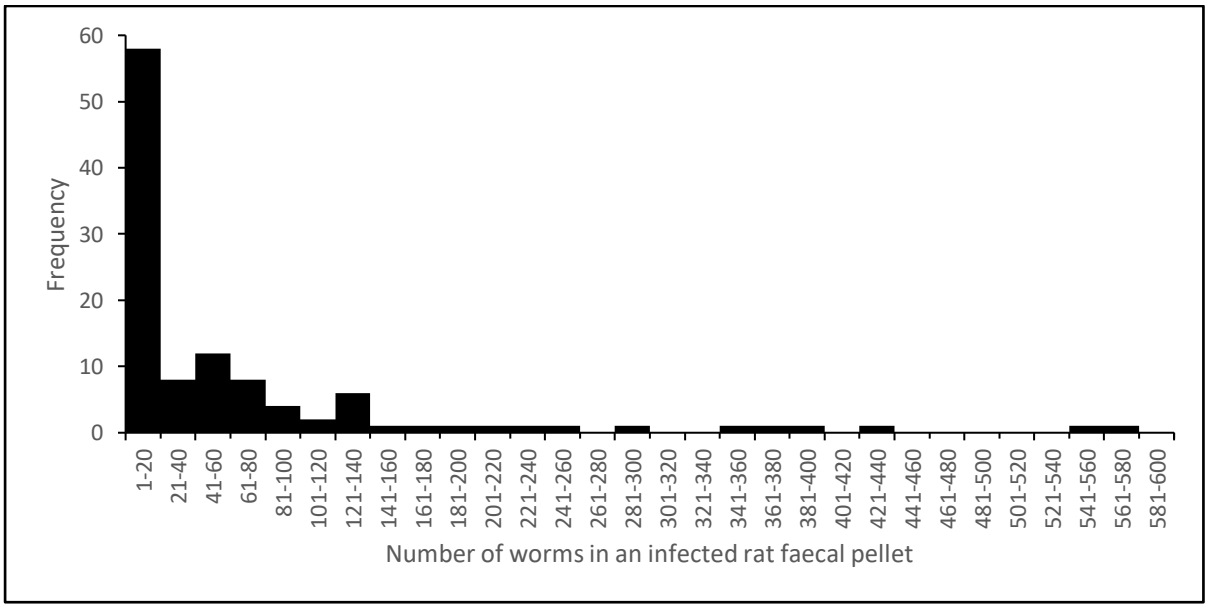


Figure 2.3: Frequency distribution of the number of Strongyloides ratti *iL3s isolated from infected rat faecal pellets. Uninfected pellets (N = 178) are not shown. Furthermore 1 pellet yielded 1,430 iL3s and another yielded 1,730 iL3s, and these pellets are also not shown. Total number of worms N = 10,471.*

Relationship between infection prevalence and intensity of infection in each site-season combination was tested with a Pearson's correlation test. Again, log-transformed intensity values were used.

2.3 Results

2.3.1 Collection of *Strongyloides ratti* from rat faecal pellets

Across all sites and seasons, 308 rat faecal pellets were collected. Of these 114 were infected, and a total of 10,471 iL3s were isolated. A breakdown of iL3s isolated by site and season is provided in Table 2.2.

Table 2.2: Rat faecal pellets collected from each site and season, showing prevalence (percentage of pellets infected) and intensity (mean iL3s isolated from infected pellet) of Strongyloides ratti *infection.*

Site and season	**No. pellets collected**	**No. pellets infected (% pellets infected)**	**No. *S. ratti* iL3s isolated**	**Mean iL3s per infected pellet (standard deviation)**
CA, Spring	35	7 (20%)	244	34.9 (24.3)
CA, Summer	32	5 (15.6%)	89	17.8 (22.1)
CA, Autumn	19	2 (10.5%)	256	128 (7.1)
CA, Winter	30	1 (3.3%)	6	6 (0)
CA, All seasons	**116**	**15 (12.9%)**	**595**	**39.7 (42.1)**
AM, Spring	11	8 (72.7%)	137	17.1 (23.5)
AM, Summer	75	23 (30.7%)	428	18.6 (29.3)
AM, Autumn	27	15 (55.6%)	3,044	202.9 (185)
AM, Winter	21	17 (81%)	5,067	298.1 (500.3)
AM, All seasons	**134**	**63 (47%)**	**8,676**	**137.7 (296.5)**
LA, Spring	27	14 (51.9%)	211	15.1 (23.1)
LA, Summer	11	7 (63.6%)	40	5.7 (3.7)
LA, Autumn	7	6 (85.7%)	360	60 (46.8)
LA, Winter	13	9 (69.2%)	589	65.4 (83.8)
LA, All seasons	**58**	**36 (62.1%)**	**1,200**	**33.3 (52.8)**
All sites, Spring	73	29 (39.7%)	592	20.4 (24.3)
All sites, Summer	118	35 (29.7%)	557	15.9 (25)
All sites, Autumn	53	23 (43.4%)	3,660	159.1 (162.4)
All sites, Winter	60	27 (45%)	5,662	209.7 (412.4)
Total	**308**	**114 (37%)**	**10,471**	**91.9 (226.9)**

2.3.2 Prevalence of *Strongyloides ratti* infection

The infection prevalence of *S. ratti* across all seasons and sites was 37%. When considering all seasons together, infection prevalence was found to vary significantly among sites (X^2 = 48.9, *df* = 2, P < 0.0001). Specifically, site CA had a significantly lower infection prevalence than both AM (12.9% *vs.* 47%, X^2 = 29.7, *df* = 1 P < 0.0001) and LA (12.9% *vs.* 62.1%, X^2 = 42.7, *df* = 1, P < 0.0001), but the difference in prevalence between AM and LA was not significant (X^2 = 3.8, *df* = 1, P = 0.051).

Furthermore, when the effect of site within seasons on prevalence was tested separately with Fisher's exact tests, this pattern of similar prevalence in AM and LA and comparatively lower prevalence in CA was observed in all seasons except Summer, in which infection prevalence at AM was intermediate between prevalences at LA and CA (Table 2.3). Hence, differences in infection prevalence among sites were largely consistent over time, with only moderate fluctuations.

Table 2.3: Fisher's exact tests for the effect of sampling site on the proportion of rat faecal pellets yielding at least one Strongyloides ratti *iL3, separated by sampling season. Significant results (P < 0.5) are shown in bold text.*

Season	**All sites P value**	**Sites compared (Infected pellets / total)**	**Comparison odds ratio**	**Comparison P value**
Spring	**P < 0.01**	**CA** (7 / 35) *vs*. **AM** (8 / 11)	10	**P < 0.01**
		CA (7 / 35) *vs*. **LA** (14 / 27)	4.2	**P < 0.04**
		AM (8 / 11) *vs*. **LA** (14 / 27)	2.4	P = 0.3
Summer	**P < 0.05**	**CA** (5 / 32) *vs*. **AM** (23 / 75)	2.1	P = 0.22
		CA (5 / 32) *vs*. **LA** (7 / 11)	8.8	**P < 0.01**
		AM (23 / 75) *vs*. **LA** (7 / 11)	2.2	P = 0.34
Autumn	**P < 0.001**	**CA** (2/19) *vs*. **AM** (15/27)	10.1	**P < 0.01**
		CA (2/19) *vs*. **LA** (6/7)	38.6	**P < 0.001**
		AM (15 / 27) *vs*. **LA** (6 / 7)	2.2	P = 0.21
Winter	**P < 0.00001**	**CA** (1 / 26) *vs*. **AM** (17 / 21)	103	**P < 0.00001**
		CA (1 / 26) *vs*. **LA** (9 / 13)	55	**P < 0.00001**
		AM (17 / 21) *vs*. **LA** (9 / 13)	1.9	P = 0.68

When taking all sites together, there was no statistically significant effect of season on infection prevalence ($X^2 = 6$, $df = 3$ P = 0.11). When Fisher's exact tests were used to test the effect of season on prevalence within sites, sites LA and CA conformed to the overall pattern of no significant effect of season (Table 2.4). However, in site AM, a strongly significant effect of season on infection prevalence was observed (Table 2.4). In particular, AM prevalence was reduced in Summer. This is in agreement with the observations that, in site-within-season tests, the statistically significant differences in the infection prevalence between AM and CA seen in all other seasons was absent in Summer (Table 2.3).

Table 2.4: Fisher's exact tests for the effect of sampling season on the prevalence of Strongyloides ratti, *separated by sampling site. Significant results (P < 0.5) are shown in bold text.*

Site	**Season**	Prevalence Infected pellets / total	**All seasons P value**
CA	Spring Summer Autumn Winter	7 / 35 5 / 32 2 / 19 1 / 26	P = 0.22
AM	Spring Summer Autumn Winter	8 / 11 23 / 75 15 / 27 17 / 21	**P < 0.0001**
LA	Spring Summer Autumn	14 / 27 7 / 11 6 / 7	P = 0.4
	Winter	9 / 13	

2.3.3 Intensity of *Strongyloides ratti* infections

The number of iL3s per infected pellet ranged from 1 to 1,730, with a mean of 91.9. Standard deviation around this mean was very high, at 226.9, and 51% of infected pellets (58 out of 114) yielded fewer than 20 iL3s (Figure 2.3). This finding is in accord with the typical negative binomial distribution of macroparasites among hosts (Churcher *et al.* 2005). This distribution may arise from unequal exposure of hosts to parasites, and / or different anti-parasite immune responses among hosts, and such processes are likely to apply here to *S. ratti.*

The effect of sampling site on intensity of infection was non-significant according to ANOVA on log-transformed intensity values ($F_{2, 111} = 2.4$, P = 0.09). When sampling seasons were treated separately, the effect of site on log-transformed intensity values was found to be significant in Spring, where intensity was significantly higher in CA than AM or LA, and Winter, where intensity was significantly higher in AM than CA or LA (Table 2.5). It should be noted that the number of infected pellets in CA was very low in Spring and Winter, such that comparisons of intensity in CA with that of other sites in these seasons are not generalisable.

Across all sites intensity was approximately ten-fold higher in Autumn (159.13) and Winter (209.7) than in Spring (20.41) or Summer (15.91) (Table 2.2). Accordingly, the effect of sampling season on log-transformed intensity values was highly significant as judged by ANOVA ($F_{3, 110} = 24.4$, P < 0.0001). Further, when the data was separated by site, all sites similarly showed a significant effect of season on log-transformed intensity values (Table 2.6). The same pattern of substantially higher intensity of infection in Autumn and Winter than Spring and Summer was seen in all sites except CA,

where intensity was lowest in Winter (Table 2.2), but only one infected pellet was collected from CA in Winter.

Table 2.5: Tests for the effect of sampling site on intensity of infection (log-transformed number of Strongyloides ratti *iL3s per infected pellet), separated by sampling season. Pairwise comparisons among sites are only made within seasons where site was found to have a significant effect on intensity by ANOVA. Significant results (P < 0.5) are shown in bold text.*

Season	**All Sites ANOVA**	**Sites compared (iL3s / infected pellets)**	**Comparison t-test**
Spring	$F_{2, 26} = 4$, **P < 0.05**	**CA** (244 / 7) vs. **AM** (137 / 8) **CA** (244 / 7 vs. **LA** (211 / 14) **AM** (137 / 8) vs. **LA** (211 / 14)	t = 2.6, *df* = 13, **P < 0.05** t = 2.8, *df* = 19, **P < 0.05** t = 0.4, *df* =20, P = 0.68
Summer	$F_{2, 32} = 0.5$, P = 0.63		
Autumn	$F_{2, 20} = 0.9$, P = 0.43		
Winter	$F_{2, 24} = 6.1$, **P < 0.01**	**CA** (6 / 1) vs. **AM** (5,067 / 17) **CA** (6 / 1) vs. **LA** (589 / 9) **AM** (5067 / 17) vs. **LA** (589 / 9)	t = -2.6, *df* = 16, **P < 0.05** t = -1.1, *df* = 8, P = 0.29 t -2.8, *df* = 24 **P < 0.01**

Table 2.6: Analysis of variance (ANOVA) for the effect of sampling season on intensity of infection (log-transformed number of Strongyloides ratti *iL3s per infected pellet) separated by sampling site. Significant results (P < 0.5) are shown in bold text.*

Site	**Season**	**Intensity** (iL3s / infected pellets)	**All Seasons ANOVA**
CA	Spring Summer Autumn Winter	244 / 7 89 / 5 256 / 2 6 / 1	$F_{3, 11} = 5$, **P < 0.05**
AM	Spring Summer Autumn Winter	137 / 8 428 / 23 3,044 / 15 5,067 / 17	$F_{3, 59} = 5$, P < **0.00001**
LA	Spring Summer Autumn Winter	211 / 14 40 / 7 360 / 6 589 / 9	$F_{3, 3}2 = 6$, **P < 0.01**

2.3.4 Relationship between intensity and prevalence

To examine the relationship between the intensity and the prevalence of *S. ratti* infection each season-site sampling combination was treated as a separate sample (N = 12) with its own prevalence and intensity value (Table 2.2). A Pearson correlation test detected no significant correlation between prevalence and log-transformed intensity among season-site combinations ($r = 0.31$, $P = 0.33$).

2.4 Discussion

2.4.1 Caveats to faecal sampling of parasitic nematodes

Experiments have shown that not all pellets shed by a rat will necessarily yield iL3s, even if the rat is known to be infected (Harvey *et al.* 1999). Furthermore, while an attempt was made to collect only fresh pellets, it is possible that some had been sitting for several days, giving time for any iL3s present to migrate away from the pellet. Therefore, prevalence of iL3s in faecal pellets is likely to be an underestimate of prevalence of parasitic adults in rats. For similar reasons, and because parasitic female *S. ratti* vary in their fecundity (Paterson and Viney 2003), the intensity of iL3s in pellets should be considered as only an approximate indicator of intensity of parasitic females in hosts. A study involving capture and dissection of wild rats would be needed to get precise estimates of both prevalence and intensity of *S. ratti* infection in wild rats.

2.4.2 Temporal and spatial variation in *Strongyloides ratti* infection parameters

These data represent the the most extensive investigation to date of the infection patterns of *Strongyloides ratti* in the wild.

The 37% infection prevalence of iL3s among rat pellets observed here is lower than the 62% recorded in a previous study of *S. ratti* in wild UK rats (Fisher and Viney 1998). This may be partially explained by differences in sampling strategy – the previous study took captured wild rats into captivity and collected faeces from them over a twelve-hours period, thereby sampling multiple faecal pellets known to come from a single individual (Fisher and Viney 1998). This method is likely to be more sensitive than the method presented here, because as discussed in section 2.4.1, a pellet may fail to yield iL3s even if the rat it came from had an *S. ratti* infection.

However, differences in geography may also contribute to the discrepancy in infection prevalences seen in the present study compared with that of Fisher and Viney (1998). This is reasonable given that in the present study found prevalences ranging from 12.9 - 62.1% among sites (Table 2.2). Intriguingly, site LA, which had the highest prevalence in this study, was the only rural site examined (Table 2.1), while all sampling sites analysed in the previous study were rural (Fisher and Viney 1998). This may indicate that rural areas generally have higher *S. ratti* prevalence than industrial areas. The industrial sites have extensive hard surfaces (concrete, asphalt), which may cause pellets to dry out more quickly than when deposited on moist soils in rural areas. *S. ratti* larvae are very sensitive to desiccation, such that faster drying may lead to reduced transmission rates at industrial sites. It has been observed that *Uncinaria stenocephala*, another nematode in which larvae develop to the infective third stage in the environment, was more prevalent in rural UK foxes (*Vulpes vulpes*) than in urban foxes (Richards *et al.* 1995).

Alternatively, differences in prevalence among sites may be related to other factors such as micro-climate, host population density, or even host genetics.

The mean intensity in infected faecal pellets was 91.9, much higher than the mean intensity of 19 iL3s per rat recorded previously (Fisher and Viney 1998). This difference is striking, especially given that the previous study (Fisher and Viney 1998) counted iL3s from multiple faecal pellets per rat. This difference may reflect the seasonal variance detected in *S. ratti* intensity in this study. In the previous study (Fisher and Viney 1998) iL3s were collected from January to July. In the present study, the seasons of Spring and Summer gave intensity values of 20.41 and 15.91 respectively, considerably different from the Autumn and Winter values of 159.13 and 209.7 respectively (Table 2.2). This seasonal pattern appears to be largely consistent across sampling sites in the present study (Table 2.6). Seasonal variation in intensity of infection has often been reported among nematode parasites of mammals (Sissay *et al.* 2007, MacIntosh *et al.* 2010), including among parasites of brown rats (O Simões *et al.* 2014), and may reflect environmental influences on the survival and development of extra-host stages, or seasonal variation in anti-parasite immune responses in hosts (Nelson *et al.* 1998).

No correlation between intensity and prevalence among site-season combinations was detected. Prevalence was heavily influenced by site, but season had little effect, the only exception being that prevalence was unusually low in Summer in site AM (Tables 2.3 and 2.4). In contrast, intensity of infection was heavily influenced by season, but site had only moderate effects limited to some site pairings in some seasons (Tables 2.5 and 2.6). This result is surprising because, *a priori*, one would expect both of these parameters to correlate with the rate of transmission. Site CA anecdotally had the lowest density rat population, as judged by the relative difficulty of finding fresh faecal pellets compared with sites LA and AM. It may be that a less dense host population at CA allows rats to retain more distinct territories, so that while infected rats are frequently exposed to re-infection by the progeny of their own parasites, non-infected rats rarely come into contact with the faeces of infected rats and are less likely to become infected themselves. It has been shown that re-infection of the parental host is common in *S. ratti* (Paterson *et al.* 2000). Alternatively, an allele conferring resistance to *S. ratti* infection may be more prevalent at CA than at other sites.

2.4.3 Consequence for *Strongyloides ratti* population genetics

The distribution of a parasites among hosts and among host populations has consequences for parasite gene flow and genetic drift, and therefore influences parasite population genetics (Gilabert and Wasmuth 2013, Mazé-Guilmo *et al.* 2016).

Host movement is recognised as one of the most important factors affecting parasitic nematode population genetics, as it is the main driver of parasite gene flow (Blouin *et al.* 1992, 1995, 1999). The dispersal distances of brown rats are small, with most individuals rarely travelling more than 150 m from their place of birth (Gardner-Santana *et al.* 2009). While some long-distance dispersal does occur occasionally in brown rats, this is usually within 2 km (Desvars-Larrive *et al.* 2017). The sampling sites used in the present study are therefore sufficiently far apart (9.7 to 30 km) that it is unlikely that an individual rat would travel among them. Consequently, it is unlikely that any one *S. ratti* individual would be carried directly between any of the sampling sites in this study. However, the rat populations in the southern UK is essentially contiguous, and not broken into discrete sub-populations. Thus, *S. ratti* alleles may flow gradually from one sampling site to another over the course of many generations.

Parasite gene flow may be higher from high-prevalence sites to low-prevalence sites, than the reverse, because hosts emigrating from a low-prevalence area are less likely to carry parasites to the new population. This potentially unbalanced gene flow may lead to low-prevalence sites having more unique alleles than high-prevalence populations, because unique alleles in a low-prevalence parasite population are less likely to flow to other sites. However, if incoming gene flow from a high-prevalence area is high enough, unique alleles in a low-prevalence population may be swamped and not be detected (Huyse *et al.* 2005). Given that site CA has a much lower prevalence than sites AM and LA, it is possible that CA will harbour more unique alleles than AM or LA. However, CA is substantially more geographically distant from AM and LA than these two sites are from each other (Figure 2.1), which may make gene flow between CA and AM / LA rare. If so, unique alleles would be expected in both CA and AM / LA combined. It will be interesting to see the extent of gene flow between AM and LA, which both have a high prevalence of *S. ratti* infection and are geographically close, albeit much further apart than the distance typically travelled by a rat over the course of its life.

Differences in prevalence among sites may also lead to differences in the rate of genetic drift, as a population with lower prevalence and similar intensity is likely to have a lower effective population size (Ne) (Blouin *et al.* 1995, Gilabert and Wasmuth 2013, Vázquez-Prieto *et al.* 2015). Faster genetic drift may lead to reduced genetic diversity in effectively smaller populations, if incoming gene flow is sufficiently restricted. Hence, less genetic diversity is expected in site CA than in sites AM or LA.

Intensity of infection may also affect gene flow, as a migrating host with a high intensity infection may be carrying a genetically diverse parasite infrapopulation that will promote parasite gene flow more so than a host with a low intensity infection. Seasonal variation in intensity was found in this study, and so the rate of *S. ratti* gene flow may depend on seasonality in brown rat dispersal – if rats tend to disperse when intensity is high, *S. ratti* gene flow will be higher than if rats disperse when intensity is low. Little is known about seasonality in brown rat dispersal – seasonal changes in rat population demography

suggested high dispersal in Winter in Harbin Province, China (Wang *et al.* 2011), but the environmental differences mean that these results cannot easily be extrapolated to UK rat populations. Nevertheless, UK brown rats appear to show seasonal fluctuations in breeding behaviour and microhabitat preference (Huson and Rennison 1981, McGuire *et al.* 2006), so that seasonal dispersal patterns are not unlikely.

2.4.4 Conclusion

In conclusion, this work has shown that *S. ratti* infection patterns vary across both space and time. Specifically, the prevalence of infection was strongly affected by geography, being very different among sampling sites, while intensity of infection was primarily affected by sampling season. These findings are likely to have consequences for the population genetics of the parasitic nematode *S. ratti*.

2.5 References

Blouin M. S., Dame J. B., Tarrant C. A. and Courtney C. H. (1992). Unusual population genetics of a parasitic nematode: mtDNA variation within and among populations. *Evolution* **46:**470-476.

Blouin M. S., Liu J. and Berry R. E. (1999). Life cycle variation and the genetic structure of nematode populations. *Heredity* **83:**253-259.

Blouin M. S., Yowell C. A., Courtney C. H. and Dame J. B. (1995). Host movement and the genetic structure of populations of parasitic nematodes. *Genetics* **141:**1007-1014.

Cable J., Barber I., Boag B., Ellison A. R., Morgan E. R., Murray K., Pascoe E. L. *et al.* (2017). Global change, parasite transmission and disease control: lessons from ecology. *Philosophical Transactions of the Royal Society B: Biological Sciences* **372:**20160088.

Churcher, T. S., Ferguson, N. M. and Basáñez, M.-G. (2005). Density dependence and overdispersion in the transmission of helminth parasites. *Parasitology* **131:**121-132.

Coomansingh C., Pinckney R. D., Bhaiyat M. I., Chikweto I., Bitner S., Baffa A. and Sharma R. (2009). Prevalence of endoparasites in wild rats in Grenada. *West Indian Veterinary Journal* **9:**17-21.

Desvars-Larrive A., Pascal M., Gasqui P., Cosson J.-F., Benoit E., Lattard V., Crespin L. *et al.* (2017). Population genetics, community of parasites, and resistance to rodenticides in an urban brown rat (*Rattus norvegicus*) population. *PLoS One* **12:**e0184015.

Dunn R. R., Davies T. J., Harris N. C. and Gavin M. C. (2010). Global drivers of human pathogen richness and prevalence. *Proceedings of the Royal Society B: Biological Sciences* **277:**2587-2595.

Easterbrook J. D., Kaplan J. D., Vanasco N. B. and Reeves W. K. (2007). A survey of zoonotic pathogens carried by Norway rats in Baltimore, Maryland, USA. *Epidemiology and Infection* **135:**1192-1199.

Fisher M. C. and Viney M. E. (1998). The population genetic structure of the facultatively sexual parasitic nematode *Strongyloides ratti* in wild rats. *Proceedings of the Royal Society B: Biological Sciences* **265:**703-709.

Gardner-Santana L. C., Norris D. E., Fornadel C. E., Hinson E. R., Klein S. L. and Glass G.E. (2009). Commensal ecology, urban landscapes, and their influence on the genetic characteristics of city-dwelling Norway rats (*Rattus norvegicus*). *Molecular Ecology* **18:**2766-2778.

Gilabert A. and Wasmuth J. D. (2013). Unravelling parasitic nematode natural history using population genetics. *Trends in Parasitology* **29:**438-448.

Harvey S. C., Paterson S. and Viney M. E. (1999). Heterogeneity in the distribution of *Strongyloides ratti* infective stages among the faecal pellets of rats. *Parasitology* **119:**227-235.

Hawley D. M. and Altizer S. M. (2011). Disease ecology meets ecological immunology: understanding the links between organismal immunity and infection dynamics in natural populations. *Functional Ecology* **25:**48-60.

Huson L. W. and Rennison B. D. (1981). Seasonal variability of Norway rat (*Rattus norvegicus*) infestation of agricultural premises. *Journal of Zoology* **194:**257-289.

Huyse T., Poulin R. and Théron A. (2005). Speciation in parasites: a population genetics approach. *Trends in Parasitology* **21:**469-475.

MacIntosh A. J. J., Hernandez A. D. and Huffman M. A. (2010). Host age, sex, and reproductive seasonality affect nematode parasitism in wild Japanese macaques. *Primates* **51:**353-364.

Mazé-Guilmo E., Blanchet S., McCoy K. D. and Loot, G. (2016). Host dispersal as the driver of parasite genetic structure: a paradigm lost? *Ecology Letters* **19:**336-347.

McCallum H. and Dobson A. (1995). Detecting disease and parasite threats to endangered species and ecosystems. *Trends in Ecology and Evolution* **10:**190-194.

McGuire B., Pizzuto T., Bemis W. E. and Getz L. L. (2006). General ecology of a rural population of Norway rats (*Rattus norvegicus*) based on intensive live trapping. *The American Midland Naturalist* **155:**221-236.

Nelson, R. J., Demas, G. E. and Klein, S. L. (1989). Photoperiodic mediation of seasonal breeding and immune function in rodents: A multi-factorial approach. *Integrative and Comparative Biology* **38:**226-237.

O Simões, R., Júnior A. M., Olifiers N., Garcia J. S., Bertolino A. V. F. A. and Luque J. L. (2014). A longitudinal study of *Angiostrongylus cantonensis* in an urban population of *Rattus norvegicus* in Brazil: the influences of seasonality and host features on the pattern of infection. *Parasites and Vectors* **7:**100.

Paterson S., Fisher M. C. and Viney M. E. (2000). Inferring infection processes of a parasitic nematode using population genetics. *Parasitology* **120:**185-194.

Paterson S and Viney M. E. (2003). Functional consequences of genetic diversity in *Strongyloides ratti* infections. *Proceedings of the Royal Society B: Biological Sciences* **270:**1023-1032.

Poulin R. (2007). Are there general laws in parasite ecology? *Parasitology* **134:**763-776.

R Core Team (2017). *A language and environment for statistical computing*. Retrieved from https://www.R-project.org/ 21/02/2019.

Richards D. T., Harris S. and Lewis J. W. (1995). Epidemiological studies on intestinal helminth parasites of rural and urban red foxes (*Vulpes vulpes*) in the United Kingdom. *Veterinary Parasitology* **59:**39-51.

Sissay M. M., Uggla A. and Waller P. J. (2007). Prevalence and seasonal incidence of nematode parasites and fluke infections of sheep and goats in eastern Ethiopia. *Veterinary Parasitology* **144:**118-124.

Speare, R. (1989). Identification of species of *Strongyloides*. In: Grove, D. I. (ed) *Strongyloidiasis: a major roundworm infection of man.* London: Taylor and Francis. pp 11-85.

Taylor M. A. (2012). Emerging parasitic diseases of sheep. *Veterinary Parasitology* **189:**2-7.

Tung K.-C., Hsiao F.-C., Wang K.-S., Yang C.-H. and Lai C.-H. (2013). Study of the endoparasitic fauna of commensal rats and shrews caught in traditional wet markets in Taichung City, Taiwan. *Journal of Microbiology, Immunology and Infection* **46:**85-88.

Vázquez-Prieto S., Vilas R., Paniagua E. and Ubeira F. M. (2015). Influence of life history traits on the population genetic structure of parasitic helminths: a minireview. *Folia Parasitologica* **62:**060.

Viney M. and Kikuchi T. (2017). *Strongyloides ratti* and *S. venezuelensis* – rodent models of *Strongyloides* infection. *Parasitology* **14:**285-294.

Viney M., Matthews B. E. and Walliker D. (1992). On the biological and biochemical nature of cloned populations of *Strongyloides ratti*. *Journal of Helminthology* **66:**45-52.

Wang D.-W., Cong L., Yue L.-F., Huang B.-H., Zhang J.-X., Wang Y., Ning L. *et al.* (2011). Seasonal variation in population characteristics and management implications for brown rats (*Rattus norvegicus*) within their native range in Harbin, China. *Journal of Pest Science* **84:**409-418.

Wertheim G. and Lengy J. (1964). The Seasonal Occurrence of *Strongyloides ratti* Sandground, 1925 and of *S. venezuelensis* Brumpt, 1934 in a Population of *Rattus norvegicus*. *Journal of Helminthology* **38:**393-398.

Chapter 3

3. 1 Introduction

3.1.1 Identification of individuals from faeces

Non-invasive genetic identification of individuals, such as genotyping using faecal DNA, is a valuable tool in ecological studies of wild animals (Beja-Pereira *et al.* 2009), with common applications including population size estimation (Frantz *et al.* 2003, Wilson *et al.* 2003), paternity assignment (Buchan *et al.* 2003, Gotelli *et al.* 2007) and tracking of home ranges and dispersal patterns (Zhan *et al.* 2007, Palomares *et al.* 2017). While invasive or destructive tissue sampling could often be employed to carry out these tasks, faecal DNA sampling offers several advantages, the most notable of which is that animals are not harmed or greatly disturbed by sample collection. This is of particular importance when the study species is endangered or resides in a sensitive habitat. Further advantages include greater ease of sampling elusive or nocturnal species, improved safety for operatives, and no need to invest in potentially costly capture methods.

However, use of faecal DNA to study wild animal populations also has disadvantages. First, in some cases, faeces from the target species may be difficult to distinguish from sympatric species that are not of interest. While non-target species can be identified genetically, this is nevertheless wasted effort. Second, rapid decay of DNA in faeces in the environment may mean that low amounts of DNA are present in each sample, leading to failures or errors in genotyping (Taberlet *et al.* 1999, Broquet and Petit 2004). Finally, some species produce very small faecal pellets, and extracting sufficient DNA from these may be challenging even if pellets are fresh. However, improved methods for sample storage, DNA extraction and detection of genotyping errors have made faecal sampling a common method of genotyping wild animals (Beja-Pereira *et al.* 2009). Comparison of faecal DNA genotyping with genotyping of carcasses from the same populations has shown that these two methods produce similar estimates of microsatellite allele frequency (Dallas *et al.* 2002).

However, faecal DNA extraction has rarely been used as a means of sampling for population genetic studies. Faecal sampling was used to test for population genetic structure in dholes (*Cuon alpinis*) and found genetic differentiation among populations either side of the River Ganges (Iyengar *et al.* 2005). This method also showed an absence of population genetic structure among black-backed jackals (*Canis mesomelas*) in South Africa (James *et al.* 2015). These results indicate that use of faecal sampling for population genetics is promising and warrants further consideration.

Analysis of faecal DNA has been used to identify rodent species from faecal samples (Moran *et al.* 2008, Galan *et al.* 2012, Barbosa *et al.* 2013), but has rarely been applied to the identification of individual rodents, and never to rodent population genetics.

3.1.2 Population genetics of rodents

Small-bodied rodents rarely disperse far from the natal site (Montgomery *et al.* 1991, Mikesic and Drickamer 1992, Pocock *et al.* 2005, Heiberg *et al.* 2012), and as such gene flow is expected to be low in these species. Furthermore, some rodent species may be subject to 'boom and bust' population dynamics, where newly available habitats are quickly colonised and exploited in a period of rapid population growth, followed by rapid decline once the available resources are exhausted (Huson and Rennison 1981, Wolff 1996, Dickman *et al.* 2010). This is expected to lead to rapid genetic drift on local scales, as new sub-populations comprising a subset of the genetic diversity in the total population are formed and breed for several generations.

Thus, strongly genetically structured populations are expected within rodent species, with genetic differentiation likely to emerge even at small geographical scales. This is reflected clearly in wild populations of the house mouse *Mus musculus* (Abolins *et al.* 2018). A lack of isolation by distance was observed amongst mouse populations in different farms, with populations that were geographically close being no more closely related to each other genetically than to ones geographically far more distant (Abolins *et al.* 2018).

Population genetics was also used to investigate connectivity between populations of the Merriam's kangaroo rat, *Dipodomys merriami* (Flores-Manzanero *et al.* 2018), and between populations of edible dormice, *Glis glis* (Moska *et al.* 2018), showing fine-scaled population genetic structuring in both species. Furthermore, population genetic analyses have been used to investigate the historical migration patterns of the Tenezumi rat, *Rattus tenezumi* (Guo *et al.* 2019), and the European woodmouse, *Apodemus sylvaticus* (Nieberding *et al.* 2004) across China and Europe respectively. These migration patterns were revealed at high resolution thanks to the fine-scale population genetic structure in these rodent species.

Strong genetic structure over small geographical scales is also evident in brown rats (*Rattus norvegicus*). Studies in Baltimore (Gardner-Santana *et al.* 2009), New York (Combs *et al.* 2017), Salvador (Kajdacsi *et al.* 2013) and Paris (Desvars-Larrive *et al.* 2017), have all revealed genetic differentiation among sub-populations as little as 1 km apart. Brown rat population genetics has also been studied on a global scale, helping to elucidate the historical migration patterns of this species and revealing strong genetic differentiation both within and among cities (Puckett *et al.* 2016). However, no study has yet examined the population genetics of brown rats from intermediate scales (tens of kilometres), nor considered rats from non-urban settings. Furthermore, no previous study has sampled brown rats non-destructively.

3.1.3 Comparative host – parasite population genetics

For many parasitic species, movement of host individuals is the primary means of parasite dispersal and therefore parasite gene flow. It therefore might be expected that the population genetics of parasites will mimic that of their host species to some extent. In some cases, this expectation is met. *Blatticola blattae* is a nematode parasite of the German cockroach (*Blattella germanica*). In this system, parasite population genetics closely mirrors that of the host, with both species showing differentiation among buildings within cities, as well as differentiation among cities 900 km apart (Jobet *et al.* 2000).

However, in some cases there are substantial differences between host and parasite population genetics, and these differences inform on factors besides host movement that influence parasite population genetics. For example, the parasitic nematode *Heligmosomoides polygyrus* shows more strongly genetically differentiated populations than its *Apodemus sylvaticus* hosts when the mitochondrial sequences of both species are compared (Nieberding *et al.* 2004). Mitochondrial genetic drift is probably stronger in *H. polygyrus t*han *A. sylvaticus,* owing to higher mitochondrial mutation rates and a faster generation time in the parasite compared with its host (Nieberding *et al.* 2004). Stronger genetic drift will contribute to stronger population genetic structure in *H. polygyrus* compared with its host. The faster emergence of population genetic structure of *H. polygyrus* allowed the authors to add detail to the hypothesised phylogeography of the host, providing greater resolution to inferred migration routes historically taken by the host species (Nieberding *et al.* 2004).

Comparative population genetics can also be used to track the spread of invasive nematodes introduced into a new habitat with their hosts. *Baylisascaris procyonis* is a nematode parasite of raccoons (*Procyon lotor*), and was introduced with its host from its native North America to northern Europe. Population genetic analyses of raccoons in Germany suggested the presence of three genetic clusters, thought to represent three distinct introductions of raccoons to Germany, but the parasites of these raccoons formed only two clusters (Osten-Sacken *et al.* 2018). One of these parasite clusters was specific to one host cluster and was presumably introduced with it, but the other parasite cluster was found across both of the remaining host clusters. The authors therefore hypothesise that the latter parasite cluster was introduced with one of the two host clusters it presently infects, and subsequently was transmitted to the other (Osten-Sacken *et al.* 2018).

Neoheligmonella granjoni is a nematode parasite of multimammate mice (*Mastomys* spp.). The population genetics of *N. granjoni* and two of its host species, *M. erythroleucus* and *M. natalensis,* were compared within a 1,300 km^2 rural area of Senegal (Brouat *et al.* 2007, 2011). The three species showed discordant population genetic structures. *N. granjoni* and *M. erythroleucus* both failed to show

population genetic structuring, with alleles being distributed homogenously among sampling sites. In contrast, *M. natalensis* showed strong population genetic structure, with genetic distance among populations correlating with geographical distance (Brouat *et al.* 2007, 2011). It is likely that the relatively high dispersal of *M. erythroleucus* promotes *N. granjoni* gene flow across the study area, including among populations of the much more sedentary host, *M. natalensis* (Brouat *et al.* 2007, 2011). Thus, where a parasite infects multiple host species, observed differences in population genetic structure between parasite and one host species may be accounted for by movement of the other host species. This applies too when a parasite has a complex life cycle involving successive host species, and those species differ in their dispersal habits. For example, an undescribed *Microphallus* sp. trematode shows much weaker population genetics than its snail intermediate host (*Potamopygrgus antipodarum*), presumably due to movement of the bird definitive hosts (Dybdahl and Lively 1996).

3.1.4 Aims of this chapter

The aims of this chapter were two-fold. The first aim was to identify individual hosts of the parasitic nematode *Strongyloides ratti*, so that *S. ratti* that originated from the same host, but which were collected from different faecal pellets, could be identified as such. This is necessary for investigating *S. ratti* population genetics. The second aim was to carry out a population genetic study on the rats themselves, both to add to existing literature on brown rat population genetics (Gardner-Santana *et al.* 2009, Kajdacsi *et al.* 2013, Combs *et al.* 2017, 2018, Desvars-Larrive *et al.* 2017) and also to allow comparisons of the host population genetic structure with that of its *S. ratti* parasites.

3.2 Materials and methods

3.2.1 *Rattus norvegicus* samples

Collection of brown rat faecal pellets was carried out from three sites in the Southern UK (Figure 2.1) and is described in detail in the methods of Chapter 2. A total of 290 of the faecal pellets collected as described in Chapter 2 were used for analyses in this chapter, and these pellets are detailed in Table 3.1. 58 of these pellets provided *Strongyloides ratti* larvae for *S. ratti* genetic analyses. Other pellets either were not infected or were infected but the nematodes collected were not sequenced. One rat was found dead at site CA. DNA was extracted from the tail tip of this rat, as well as from faeces taken from the carcass.

Table 3.1: Rat faecal pellets used in this chapter.

Site	**Pellets used in initial diversity screening**	**Pellets used in population genetic analysis**
CA	7	106
LA	6	49
AM	6	116

3.2.2 Microsatellite genotyping

3.2.2.1 Extraction of DNA from faeces

DNA extraction from faeces was carried out on the material removed from rat faecal pellets prior to culturing of those pellets for *S. ratti* isolation (full details of sample processing are in Chapter 2 methods). Partial faecal pellets were weighed prior to DNA extraction, and the mean mass was 70 mg. DNA was extracted using a QIAamp DNA Stool Mini Kit (Qiagen). The 'Isolation of DNA from stool for human DNA analysis' protocol provided by the manufacturer was used, except that insufficient faecal material remained of each pellet to make up the 180-220 mg recommended in that protocol. As much material as was available was used instead. Following extraction, isolated DNA was stored at -20°C.

3.2.2.2 Extraction of rat DNA from tail tip tissue

DNA was extracted from rat carcasses by incubation of 1-2 mm of tail tip in 600 µl of 50 mM NaOH at 95°C for 20 minutes with frequent vortexing. Tissue debris was removed by centrifugation, and 200 µl supernatant was added to 50 µl 1M Tris-HCl (pH 7) to create a working stock of template DNA that was stored at -20°C.

3.2.2.3 Microsatellite loci used

Twenty rat microsatellite loci, all with dinucleotide repeats, were selected for initial analysis (Table 3.2, Giraudeau *et al.* 1999, Steen *et al.* 1999). While loci with longer repeats would reduce any ambiguity in scoring repeat number, the loci chosen were selected because they had previously been used successfully in studies of wild rat populations (Gardner-Santana *et al.* 2009, Desvars-Larrive *et al.* 2017).

3.2.2.4 PCR amplification of microsatellite loci

Polymerase chain reaction (PCR) primer sequences for the amplification of each of the 20 loci were retrieved from the Rat Genome Database (rgd.mcw.edu/, sequences in Table 3.2). Primers were purchased from Sigma Aldrich. In each case, the forward primer was provided labelled with VIC, FAM or NED fluorophores (Table 3.2). These fluorophores have non-overlapping emission peaks.

Table 3.2: Rat microsatellite loci tested in this chapter. Primer sequences are presented with the forward primer first, orientated 5' to 3'. Reference length refers to the length of the region amplified by the given primer sequences in the current release of the brown rat genome (Rnor 6.0). Primer sequences and expected amplicon length were retrieved from Rat Genome Database (https://rgd.mcw.edu/). Fluorophore indicates the fluorophore used to label forward primers and thus PCR products in this chapter. Fluorophore names are registered trademarks of Applied Biosystems. Alleles in diversity test refers to the number of alleles detected when 19 rat faecal pellets were genotyped at all loci. D1Cebr3 could not be reliably amplified and so was excluded from diversity testing (marked not applicable (NA) in alleles in diversity test). Loci shaded in blue were taken forward to rat population genetic analysis. References are 1) Giraudeau et al. *1999 and 2) Steel* et al. *1999*).

Locus	**Primer sequences**	**Expected amplicon length**	**Fluorophore**	**Alleles in diversity test**	**Reference**
D1Cebr3	CTTGGGAGCTGGGAGTGT GAAGGCTGAGGTATGAAGACTG	101	FAM	NA	1
D1Cebr9	GGATTTGGCTCCCTTTAAG CAGTAACTCTGGTTCATGTACTCC	274	VIC	4	1
D2Cebr1	GCCTCCCTCTCTGCACAC GAGGTGCCAGGAAGGTCT	212	NED	4	1
D3Cebr3	CAGGGAATGCAGAAGATACAG GTGGCTTTAGGACTCTGGAG	167	FAM	2	1
D3Rat159	CCAGGGATGAGTCCAAGGTA CTGGTCTGCTTCCTCCAGTC	243	VIC	7	2

D4Cebr2	TGTCAAAGAAAGCCAGTAAAAC TTGGCAACCAGGAATAGC	143	VIC	4	1
D4Rat59	GCAGTGTGTTTGGGGTAGCT GCGGAATGATAGTTACTACGGC	180	FAM	10	2
D5Rat43	AGCCCTTAAGCTGAGCTACAGA GGCACCAGGCATACTCATG	200	NED	4	2
D6Cebr1	GGTTTGGTTGGGGAGAA GTGCTGTCAGGGAAAGATGTA	223	NED	10	1
D8Rat162	TCACTGGCAGCAATTTACCA TCTGAGACCTCTTCAACTCTGTTG	249	VIC	5	2
D10Cebr1	TTTGTTTGGCTAGAATTATGC TGTTCAGCAAAGTAGCAGGATA	176	FAM	4	1
D10Rat105	ATCCAGCCAGAAAGCAAAAC CTGGCTGAGTCCTGTCACAA	100	FAM	6	2
D11Rat11	AACTGTTGCCAGCATTAGGC TCCCTGTTCCTATCTGGTCCT	150	NED	5	2
D12Rat42	CAACCCAGTGTGTCAAACGT GGGTTGGTGAAGCATTTTCA	128	VIC	9	2
D13Rat21	ACCCTGAAGTCAGCCTCTGA ACCACAGCATTTCTCTCGCT	150	FAM	5	2
D14Rat110	AACATTGTCTTGCTTAGCCTCA CTCCACCCACACACCACG	280	NED	6	2
D15Rat64	GCATGTACCGTTCTTGCAGA AGACATAGGGCTGTAGGGCA	108	VIC	6	2
D18Rat11	GCCCAGGAGCTAAGTCTGATT CCAGCCTCAGAGCCAATAAG	133	FAM	9	2
D19Rat62	GTGCTAATGTGGGTGGCTTT TGAATTCTACCATGCATCACAG	112	NED	9	2
D20Cebr1	GCAACACATGGTGGCTCA CCCTGACCGTTTAGTAGCAT	308	NED	3	1

PCR reactions contained 200 µM each of dATP, dCTP, dGTP and dTTP, 250 nM of each primer and 1 µl of template DNA (extracted either from faeces or tail tip as described) with DreamTaq DNA polymerase and DreamTaq buffer (Thermo Scientific) added according to the manufacturer's instructions. The final reaction volume was 20 µl. The thermocycler regimen was 95°C for 15 minutes, then 35 cycles of 94°C for 30 seconds, 57°C for 30 seconds, and 72°C for 25 seconds then a final extension of 60°C for 10 minutes. 8 µl of each PCR product was visualised by agarose gel

electrophoresis (2% w/v agarose with 0.5 μg/mL ethidium bromide, run at 100V for 50 minutes), and the remainder was stored at -20°C.

3.2.2.5 Capillary electrophoresis of PCR products

To prepare PCR products for capillary electrophoresis, 1, 2.5 or 5 μl of the PCR product was combined with up to two other PCR products labelled with different fluorophores, and the volume was made up to 10 μl with distilled water. PCR products were combined to reduce the number of electrophoresis runs required. The volume of each PCR product used was based on the intensity of the band produced when the product was run on agarose gel; 1 μl for a bright band, 2.5 μl for a faint band, and 5 μl for no band visible. 'GeneScan 500 LIZ' size standard (Thermo Fisher) was subsequently added to each PCR product mixture. This size standard consists of single standard DNA fragments measuring 35, 50, 70, 100, 139, 150, 160, 200, 250, 300, 340, 350, 400, 450 and 500 base pairs, each labelled with the fluorescent dye LIZ. The emission peak of LIZ does not overlap with that of NED, VIC or FAM.

PCR product mixtures with the size standard were loaded onto an Applied Biosystems 3500 Genetic Analyser (ThermoFisher), and the capillary electrophoresis functionality of this machine was used to measure the length of PCR products. During capillary electrophoresis, the time taken for a DNA fragment to pass through the capillaries is proportional to the length of the DNA molecules, and the genetic analyser detects the fluorescence of labelled DNA fragments as they emerge from the capillaries. GeneMapper software (Thermo Fisher) coupled to the genetic analyser was used to calculate the length of PCR products by comparing the time they take to emerge from the capillaries with the time taken by the DNA fragments of known length from the size standard.

When examining emission traces corresponding to a particular locus, it was common to observe clusters of two or more peaks within the expected size range, and the peaks within these clusters differed in length by multiples of two nucleotides (Figure 3.1, blue cluster). In some cases, two such clusters were observed (Figure 3.1, black and green clusters). Where there were two clusters, this was interpreted as representing heterozygosity where each allele was affected by replication slippage during PCR, so that multiple fragments of slightly different lengths were produced for each. The length of the highest peak in each cluster (peak height indicating fragment abundance) was recorded as the allele length. This phenomenon of replication slippage means that homozygosity cannot be readily distinguished from heterozygosity where the alleles differ in length by just one or two repeats. Where only a single cluster of peaks was observed at one locus, the genotype was scored as homozygous for the highest peak in the cluster, but with this caveat in mind.

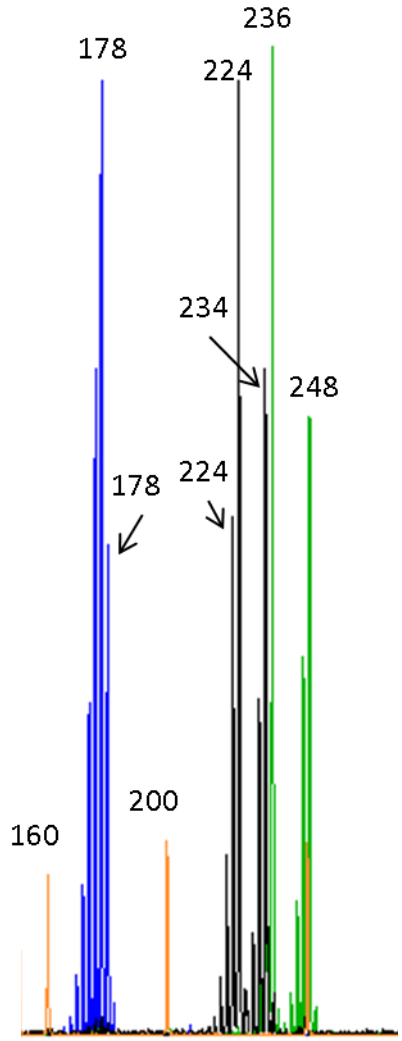


Figure 3.1: Emission traces of peaks corresponding to microsatellite loci D4Rat59 (blue), D6Cebr1 (black) and D3Rat159 (green) in a wild rat. Template DNA was extracted from the tail tip of a rat found dead at site CA. Numbers represent the size in bp of the DNA fragment. Height of the peaks indicates abundance of the corresponding fragment. Orange peaks are the size standard.

3.2.2.6 Microsatellite genotyping for population genetic analysis

Primer pairs were first tested for amplification efficiency under the chosen thermocycler regimen on DNA prepared from the tail tip of a Wistar rat. All primer pairs were found to give robust amplification except D1Cebr3, which was dropped from further analyses. Next, each of 19 rat faecal pellets representing all three sampling sites (Table 3.1) were genotyped at all 19 remaining microsatellite loci, and 9 loci that were found to be highly variable and consistently produced bright bands during PCR and gel electrophoresis (Table 3.2) were selected for population genetic analysis of the complete set of rat faecal pellets. Subsequently a further 271 faecal pellets were genotyped (Table 3.1). These PCRs were carried out in batches of 40 to 60, and in each batch there was a positive control consisting of DNA

extracted from the tail tip of the rat found dead at site CA. Thus this individual rat was genotyped at all 9 loci multiple times.

3.2.3 Analysis of microsatellite data

Faecal pellets that were successfully genotyped at fewer than 6 microsatellite loci were excluded from analyses, leaving a dataset of 132 genotyped faecal pellet samples. This dataset was analysed in GenAlEx version 6.5 (Peakall and Smouse 2006, 2012), a freely available plug-in for Microsoft Excel. GenAlEx's pairwise relatedness function was used to detect pellets with identical multi-locus genotypes, and these were taken to represent the same individual rat. The validity of this strategy for identifying individuals were checked with GenAlEx's probability of identity function, which determines the probability of two (a) randomly chosen individuals, or (b) full sibs having identical multi-locus genotypes. GenAlEx was also used to detect deviations from Hardy-Weinberg equilibrium (HWE).

Locus D12Rat42 was excluded from further population genetic analyses due to the low number of rat faecal pellets successfully genotyped at this locus. For the remaining data, Doubled Ritland and Lynch relatedness (Lynch and Ritland 1999), was calculated for each pair of individuals in pairwise comparisons. Relatedness values where the individuals being compared were from the same site (same-site pairs) were averaged and compared to average relatedness values for among-site pairs. For population assignments, GenAlEx compared the multilocus genotype of each rat with the allele frequencies of each of the sampling sites (excluding the rat currently being investigated) and from this determined the log-likelihood of the rat originating from each sampling site. The rat was then assigned to the sampling site with the highest log-likelihood value.

Shannon's mutual information index ($^{S}H_{UA}$) values uses differences in allele frequencies among putative populations as a measure of population subdivision (Sherwin *et al.* 2006) and is valid despite deviations from HWE within subpopulations (Hedrick 2005). $^{S}H_{UA}$ values of 0 indicate unhindered gene flow across populations, while a value of 1 indicates a complete lack of gene flow. $^{S}H_{UA}$ values were used to quantify the differences in allele frequencies among sampling sites and to estimate the number of effective migrants per generation (Nm). Shannon's information index ($_{S}H$) describes the proportion of genetic diversity that is among populations as opposed to within populations (Jost *et al.* 2018). $^{S}H_{UA}$, Nm and $_{S}H$ estimation all assume effective population sizes of greater than 500 individuals. Fixation index (F_{ST}), $^{S}H_{UA}$ values and $_{S}H$ values were calculated using GenAlEx.

3.2.4 Detection of faeces from other mammal species

It is possible that faeces from other small mammal species were misidentified as coming from *R. norvegicus* and inadvertently collected. To investigate this, two sympatric species were considered, the grey squirrel (*Sciurus carolinensis*) and the common mole (*Talpa europaea*). It is unlikely that any moles were present at sites CA or AM, which are industrial sites with floors of concrete or asphalt, but evidence of mole activity was seen near site LA, a rural farm. The black rat (*Rattus rattus*) is very rare in the UK and it is unlikely to be present at the sampling sites used in this study, but as there was a possibility of it being present, *R. rattus* was considered also.

European mole and grey squirrel samples were donated from private collections. The mole had been found dead in Oxfordshire, UK and subsequently preserved in formalin for at least 10 years (exact time of collection unknown). One squirrel was found dead in Cardiff, UK, and preserved through freezing. Two black rat carcasses were provided by Bristol Zoo Gardens, these having been culled as part of routine maintenance of the zoo's black rat colony. DNA was extracted from tail tips of these carcasses as described for brown rats. For the squirrel, hairs on the tail tip were trimmed down to skin prior to DNA extraction.

PCR was carried out on these DNA extractions using the 9 microsatellite loci used for brown rat population genetic analysis. In addition, two further primer pairs were used. Scv1 has previously been used to amplify DNA from *Sciurus* sp. (Hale *et al.* 2001), and RodActin is known to amplify the actin coding sequence from *R. rattus* (Apte *et al.* 2007). Details of these primers are given in Table 3.3. PCR conditions were as described previously, except that a modified thermocycle regimen was applied to reactions using Scv1. This regimen, adapted from one used previously for this primer pair (Hale *et al.* 2001), featured an initial denaturation step of 95 °C for 12 minutes, then 10 cycles of 94 °C for 15 seconds, 54 °C for 15 seconds and 72 °C for 20 seconds, then 30 cycles of 89 °C for 15 seconds, 54 °C for 15 seconds and 72 °C for 20 seconds, with a final extension of 72 °C for 10 minutes. PCR products were visualised on agarose gel as described previously.

Successful amplification with RodActin demonstrated that black rat DNA extractions had been successful, and also gave strong amplification of brown rat DNA. Similarly, primer pair Scv1 gave strong amplification of squirrel DNA, producing a bright band of ~450 bp and a fainter band of ~180 bp during gel electrophoresis visualisation. Scv1 gave weak amplification from brown rat DNA but the ~450 bp band seen in squirrel amplifications was absent. Subsequently, DNA extractions for any faecal pellet collected for this study that could not be amplified at any brown rat microsatellite loci were tested with Scv1, with squirrel DNA as a positive control.

Attempts were made to find a primer pair known to amplify mole DNA in the literature, but without success, so primers were designed to amplify an intronic region of the published *T. europaea* Histone Deacetylase 2 gene sequence (Nicolas *et al.* 2017). Unfortunately, despite trialling a variety of PCR conditions, it was not possible to amplify mole DNA, or indeed brown rat DNA, with these primers.

*Table 3.3: Primer pairs used for detection of non-*Rattus norvegicus *faecal pellets. References are 1: Hale* et al. *2001, 2: Apte* et al. *2007, 3: Nicolas* et al. *2017. Scv1 and RodActin sequences were retrieved from these references. TeHDACi10 sequences were designed from published* T. europaea *sequence.*

Primer pair	**Primer sequences**	**Target species**	**Expected length**	**Reference**
Scv1	CTCCTCTTCCAAGGGTGACA GATGGCCTCTGTTTCTCTGC	*Sciurus carolinensis*	185, 450	1
RodActin	AGGTATCCTGACCCTGAAGTA CACACGCAGCTCATTGTAGA	*Rattus* sp.	103	2
TeHDACi10	ACCCAGACCTTTGCACAACA AGGAGCCTTTGGAGGTCATT	*Talpa europaea*	277	3

3.3 Results

3.3.1 Detection of faeces from other mammal species

None of the 9 microsatellite loci used to study brown rat population genetics generated a detectable product when used to amplify DNA from squirrel or mole, despite amplifying robustly from brown rat positive controls. Thus, it is concluded that any sampled faecal pellet that could be amplified with at least one of these 9 microsatellite loci could be ruled out as originating from squirrel or mole.

However, all microsatellite loci except D12Rat42 amplified from DNA from black rats, and in all cases, the sizes of the bands amplified from the two *Rattus* species were similar. Thus, only pellets that were successfully amplified at locus D12Rat42 could be formally ruled out as originating from black rats. Based on these results and considering the rarity of black rats in the UK, any pellet that was amplified with at least one of the 9 microsatellite loci was taken as coming from a brown rat, even if D12Rat42 was not successfully amplified, but with this caveat in mind.

Of the faecal pellets that could not be amplified at any of the 9 rat microsatellite loci, none were successfully amplified with Scv1 despite robust amplification of squirrel DNA positive controls. This indicates that amplification failure of these samples was not due to misidentification of squirrel faecal pellets as rat pellets, and rather is likely to be due to failure in DNA extraction from faecal pellets. It remains formally possible that these pellets are not from brown rats but from black rats, squirrels or moles. Such pellets were naturally not included in brown rat population genetic analysis, but are taken as coming from brown rats for the purposes of *Strongyloides ratti* population genetics.

3.3.2 Microsatellite diversity in *Rattus norvegicus*

3.3.2.1. Genotyping success

The rat found dead at site CA was genotyped multiple times at nine loci from tail tip DNA, and in each case the multilocus genotype returned was the same. Further, genotype results were also the same when faeces taken from the same carcass were genotyped. This shows that genotyping results are consistent within an individual.

Of 290 pellets, 132 (46%) were successfully genotyped at 6 or more loci. This includes 39 of the 58 pellets (67%) that provided *S. ratti* larvae used for *S. ratti* DNA sequence analysis. There was no correlation between the mass of faeces used for DNA extraction and the number of loci successfully genotyped (Pearson's $r^2 = -0.01$). The number of loci successfully genotyped per sample was non-normal (Shapiro-Wilk's $W = 0.93$, $P > 0.00001$), and showed a largely even distribution (Figure 3.2). Only the 132 samples genotyped at 6 or more loci were used in further analyses.

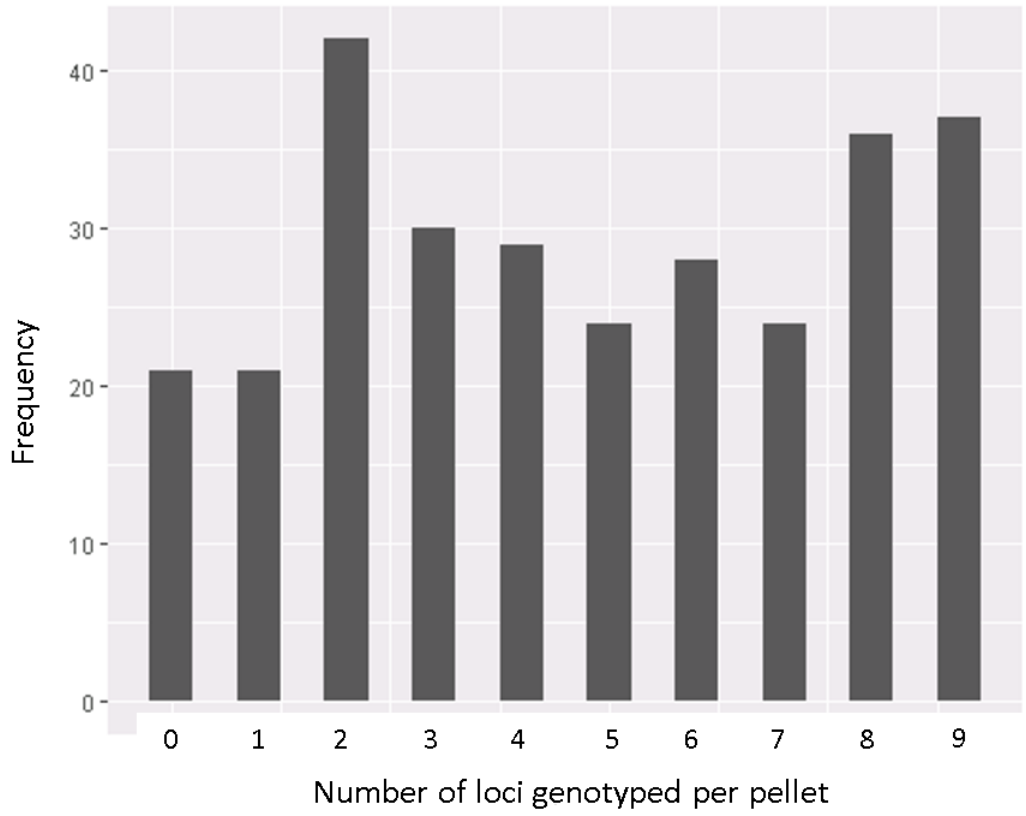


Figure 3.2: Frequency distribution of number of microsatellite loci successfully genotyped per rat faecal pellet.

3.3.2.2. Identification of individual rats

The probability of two unrelated individuals having identical genotypes at any 6 of the 9 loci was calculated as 2.5 x 10^{-7}, 3.4x10^{-7} and 4x10^{-7} at sites CA, AM and LA respectively, and the probability of two full-sibs having identical genotypes at 6 loci was calculated as 3.5x10^{-4}, 3.4x10^{-4} and 3.9x10^{-4} at sites CA, AM and LA respectively. Thus, any pellets with identical multi-locus genotypes were considered to come from the same individual. In total, there were 14 instances of multiple pellets being collected from the same rat. The number of pellets collected from a single rat ranged from 1 to 4 with a mean of 1.18, and 112 individual rats were represented by the 132 genotyped pellets (Table 3.4). Pellets assigned as coming from the same rat always came from the same sampling site. Of the 39 pellets providing *S. ratti* for sequencing that were successfully genotyped, there were 4 instances of multiple pellets coming from the same rat, and in each case, 2 pellets from the same rat were found.

Table 3.4: Number of individual rats assessed for rat population genetics. Rats were identified by microsatellite genotyping of faecal pellets, and are included only if they could be genotyped with at least 6 loci.

Site	**Pellets genotyped**	**Individual rats**
CA	55	47
LA	23	19
AM	54	46

3.3.2.3. Allelic diversity

Of the 112 identified rats, the number successfully genotyped at each locus ranged from 34 to 112 with a mean of 92 successes per locus. The number of alleles per locus ranged from 8 to 16 with a mean of 12 (Table 3.5). There was no correlation between the number of rats genotyped at a locus and number of alleles detected at that locus (Pearson's $r^2 = 0.04$), suggesting that sufficient rats were genotyped to detect all prevalent alleles.

3.3.3 Population genetics of *Rattus norvegicus*

Most loci in most sample sites were not in HWE. Specifically, 8 of 9 loci at site CA, 3 of 9 loci at site AM and 5 of 9 at site LA were not in HWE, and in all cases, this was due to an excess of homozygotes (Table 3.5).

Table 3.5: Genetic diversity of microsatellite loci initially genotyped for assessment of rat population genetics. Number of rats genotyped indicates the number of individual rats in which genotyping was successful. He and Ho are expected and observed heterozygosity, respectively. Diversity within sites indicates the proportion of allelic diversity that partitions within sites, as opposed to among sites, according to $^{S}H_{UA}$*. Due to the low success rate in genotyping rats at D12Rat42, this locus was not used in population genetic analysis, shown as NA.*

Locus	**Number of rats genotyped**	**Number of alleles**	**He**	**Ho**	**Diversity within sites (%)**
D3Rat159	103	14	0.779	0.756	18
D4Rat59	105	12	0.822	0.652	8
D6Cebr1	109	14	0.752	0.356	13
D8Rat162	73	15	0.835	0.600	16
D10Rat105	112	8	0.696	0.617	13
D12Rat42	34	9	0.762	0.267	NA
D14Rat110	83	16	0.824	0.529	18
D18Rat11	112	10	0.686	0.319	19
D19Rat62	96	10	0.678	0.462	13

In all analyses described below, the locus D12Rat42 was excluded, due to low genotyping success rate of this locus. Allele frequencies of most loci were markedly different in different sampling sites (Table 3.6), consistent with restricted gene flow among these sites causing population genetic structuring. Accordingly, Ritland and Lynch pairwise relatedness values were on average higher in within-site comparisons (0.06) than among-site comparisons (-0.06). Interestingly, in within-site comparisons, histograms of relatedness values showed right hand tails (Shapiro-Wilkes test for normality statistics W = 0.89 for site AM, W = 0.9. for site CA and W = 0.87 for site LA, P > 0.0001 in all cases, Figure 3.3), indicating an over-abundance of closely related pairs of individuals compared with neutral expectations.

Table 3.6: Allele frequencies for microsatellite loci used in this chapter, expressed as percentages, broken down by sampling site.

Locus	**Allele**	**Site CA**	**Site LA**	**Site AM**
D3Rat159	229	0	2.9	0
	235	3.3	2.9	1.2

	237	0	2.9	0
	239	2.2	0	0
	241	36.7	5.9	3.7
	243	3.3	61.8	43.9
	245	8.9	5.9	8.5
	247	2.2	2.9	0
	249	25.6	2.9	19.5
	251	3.3	2.9	22
	253	0	8.8	1.2
	257	1.1	0	0
	259	6.7	0	0
	261	6.7	0	0
D4Rat59	160	0	3.3	0
	172	28.3	23.3	21.6
	174	5.4	3.3	22.7
	176	15.2	0	12.5
	178	2.2	0	0
	180	3.3	6.7	9.1
	182	20.7	6.7	3.4
	184	15.2	43.3	11.4
	186	6.5	0	12.5
	188	2.2	0	4.5
	190	1.1	10	2.3
	192	0	3.3	0
D6Cebr1	211	0	0	7.6
	219	0	0	4.3
	221	0	2.8	3.3
	223	24.4	41.7	47.8
	225	18.9	13.9	18.5
	227	0	11.1	0
	229	0	5.6	1.1
	231	2.2	11.1	5.4
	233	0	0	1.1
	235	3.3	2.8	1.1
	237	35.6	2.8	8.7

	239	15.6	0	0
	241	0	8.3	0
	243	0	0	1.1
D8Rat162	217	30	0	0
	219	0	0	1.4
	229	0	11.5	0
	241	2	0	0
	243	6	11.5	31.4
	245	0	3.8	8.6
	247	0	26.9	0
	249	6	11.5	2.9
	251	10	0	7.1
	253	16	15.4	21.4
	255	10	15.4	18.6
	257	14	0	4.3
	259	2	0	4.3
	261	4	0	0
	265	0	3.8	0
D10Rat105	92	2.1	0	1.1
	94	2.1	2.6	0
	96	7.4	13.2	0
	98	46.8	36.8	48.9
	100	4.3	39.5	38
	102	24.5	5.3	10.9
	104	12.8	0	0
	108	0	2.6	1.1
D12Rat42	106	0	0	7.1
	120	3.3	0	3.6
	122	0	0	3.6
	124	30	10	17.9
	126	16.7	40	10.7
	128	16.7	40	17.9
	130	30	0	14.3
	132	3.3	10	14.3
	134	0	0	10.7

D14Rat110	266	0	3.1	0
	270	27.9	3.1	0
	272	1.5	0	0
	282	7.4	40.6	7.6
	284	16.2	18.8	3
	285	0	0	3
	286	4.4	0	4.5
	288	5.9	0	0
	290	23.5	3.1	4.5
	292	0	15.6	31.8
	294	1.5	6.3	27.3
	296	0	3.1	13.6
	298	5.9	6.3	1.5
	320	0	0	1.5
	324	1.5	0	0
	326	4.4	0	1.5
D18Rat11	109	1.1	0	12
	111	1.1	0	0
	113	7.4	0	2.2
	115	1.1	0	1.1
	117	0	5.3	42.4
	119	1.1	34.2	2.2
	121	45.7	28.9	33.7
	123	28.7	10.5	6.5
	125	1.1	0	0
	127	12.8	21.1	0
D19Rat62	104	0	0	1.3
	106	0	2.9	0
	110	1.3	0	0
	112	33.3	35.3	16.3
	114	41	17.6	12.5
	116	0	11.8	3.8
	118	1.3	17.6	2.5
	120	20.5	14.7	58.8
	122	2.6	0	1.3

	124	0	0	2.5
	136	0	0	1.3

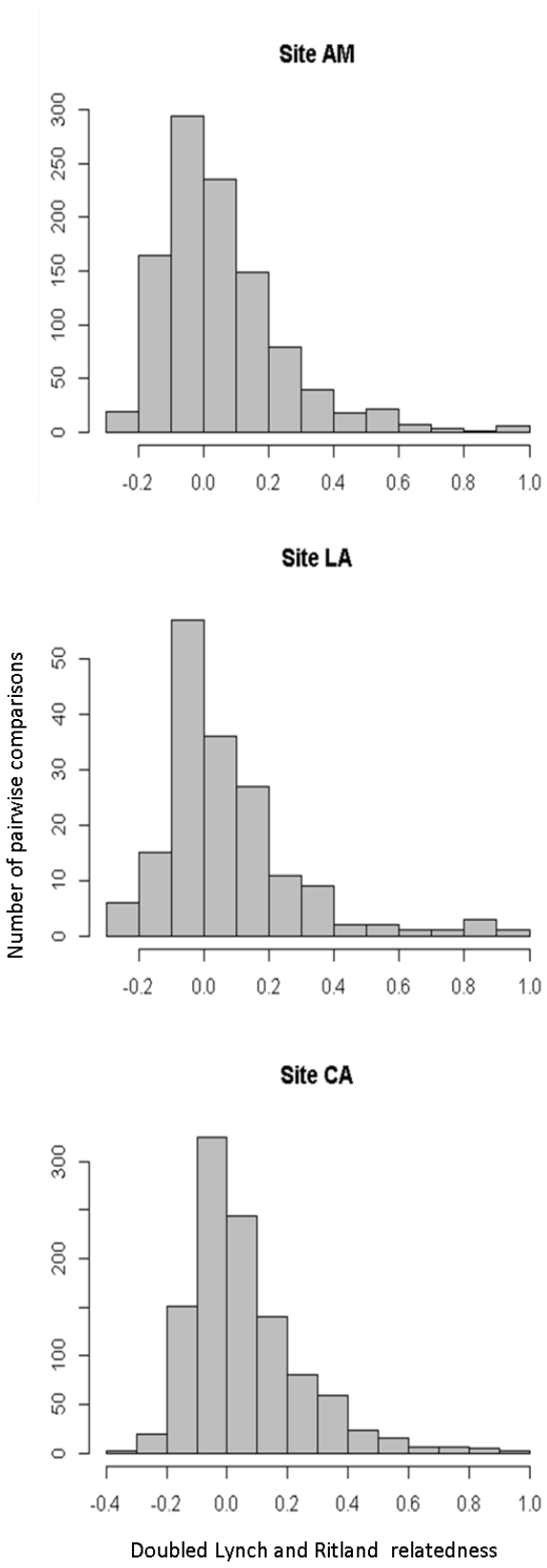


Figure 3.3: Frequency distribution of Ritland and Lynch relatedness values (Lynch and Ritland 1999) in pairwise comparisons of individual rats from the same site. Relatedness values are doubled to give a possible range of -1 to 1.

When each rat was computationally assigned to a sampling site based on site allele frequencies, 100 of 112 rats (89%) were assigned correctly. Of rats that were incorrectly assigned, 3 AM rats were assigned to LA, 3 AM rats were assigned to CA, 3 LA rats were assigned to CA, 1 LA rat was assigned to AM, and 2 CA rats were assigned to LA.

Together, differences in allele frequency, high accuracy in computational population assignment and higher within-site relatedness than among-site relatedness indicate genetic differentiation among rats at the three sampling sites. In contrast, F_{ST} among all three sites was 0.08, and in pairwise comparisons was 0.06 between AM and LA, 0.07 between AM and CA, and 0.05 between LA and CA. This equates to 8% of the genetic variation partitioning among sites, as opposed to within sites, and suggests weak population genetic structuring. However, F_{ST} assumes HWE within the populations tested, an assumption violated in this study, thus largely invalidating drawing inferences from the Fst statistic.

Shannon statistics, such as Shannon's information index ($_SH$) and Shannon's mutual information index ($^SH_{UA}$), do not assume HWE (Hedrick 2005, Jost *et al.* 2018), and so may be a more appropriate analysis here. A $^SH_{UA}$ value of 0 indicates unhindered gene flow between sites such that they act as a single freely mixing, while a value of 1 indicates a complete lack of gene flow between sites. $^SH_{UA}$ over 8 loci from 112 rats (Table 3.4) were 0.22 between sites CA and AM, 0.15 between CA and LA, and 0.21 between AM and LA. These values suggest moderate genetic differentiation among sites, driven by low rates of gene flow. Indeed, the number of migrants per generation (Nm) was estimated from these $^SH_{UA}$ values as 0.23 migrants per generation between sites CA and AM, 0.5 between CA and LA, and 0.26 between AM and LA.

$_SH$ was used to estimate the percentage of genetic variation explained by sampling site for each locus. This was found to vary amongst loci, ranging from 8% to 19% (Table 3.5), with a mean of 15% (3.5% standard deviation). This supports other results indicating moderate population genetic structuring among rat populations at the three sampling sites.

3.4 Discussion

3.4.1 Use of faeces-derived DNA for genetic analysis

Non-invasive genetic sampling, for example from faeces, has many advantages over invasive sampling that necessitates the trapping and potentially harming of wild animals. However, extraction of sufficient DNA for analysis may be challenging if only small amounts of faecal material are available.

In this study of the population genetics of brown rats, faeces were used as the DNA source. However, DNA was not extracted from whole faecal pellets, but rather from the fragments (mean mass 70 mg) that remained after the majority of the pellet was used for *Strongyloides ratti* culture. Thus, the amount of faecal material available for each sample was limited. This may explain the rather low genotyping success observed here; only 43% of faecal pellets were successfully genotyped at 6 or more of the 9 loci used. However, that there was no correlation between the mass of the faecal material used and the number of loci genotyped, this argues against a low mass of faecal material alone explaining the low genotyping success rate. Alternatively, as DNA in faeces degrades over time, genotyping success may be influenced by the length of time pellets were sitting in the environment prior to collection and freezing. Although all pellets collected appeared fresh (still moist, no visible fungal or algal growth), it nevertheless was not possible to control the age of faeces exactly, and so the variable genotyping success may be due to variation in the freshness of the faecal samples.

It is common in microsatellite genotyping of wild animal faeces that some loci cannot be genotyped from some samples (Gotelli *et al.* 2007, Zhang *et al.* 2008, Beja-Pereira *et al.* 2009, Palomares *et al.* 2017). For example, only 27% of faecal samples could be genotyped at all in a study of wild big cats, and when considering only genotyped samples, loci ranged from 48% of samples successfully genotyped, to 94% (Palomares *et al.* 2017). Nevertheless, this study was able to detect population genetic structure, indicating that faecal DNA sampling of wild animals can provide sufficient data for population genetic analyses despite genotyping failures, as long as sufficient faecal samples are collected. Furthermore, more advanced sample preservation techniques, such as desiccating faecal samples with silica or ethanol in conjunction with freezing, may allow for improved genotyping success rates (Beja-Pereira *et al.* 2009).

3.4.2 Identification of individuals from faeces

One of the primary aims of this work was the identification of individual rats by genetic analysis of their faeces, so that *S. ratti* isolated from these faeces could be correctly assigned to individual hosts. The use of highly variable microsatellite loci (Table 3.5) means that the probability of two unrelated individuals sharing an identical genotype at 6 or more loci is very small (in the order of 10^{-7}), and the probability remains very small even if the rats considered are full sibs (approximately 10^{-4}). Thus,

pellets that have been successfully genotyped at 6 or more loci can be unambiguously assigned to individual rats. This revealed that repeat sampling of the same individuals from multiple independently collected pellets was quite rare, with a mean of 1.18 pellets collected per rat.

However, only 67% of pellets providing sequenced *S. ratti* were genotyped at 6 or more loci. The remaining 33% of pellets, successfully genotyped at 5 loci or less, were not assigned to individuals. Sequenced worms originating from these pellets (37 worms, constituting 41% of the total number sequenced) therefore could not be assigned to individual hosts. For the purposes of *S. ratti* population genetics, each unassigned pellet was treated as representing a unique rat, which is not unlikely given how rarely multiple pellets from the same rat were detected amongst the pellets that were successfully genotyped.

Thus, microsatellite genotyping of faecal samples can be used to identify individual wild rats from faeces with high confidence. Failure to assign some pellets to rats came from low genotyping success rate, not from ambiguity in assignment of pellets that were successfully genotyped at 6 loci or more. Hence, measures to improve genotyping success rate such as desiccation of samples prior to freezing would also likely improve assignment of faecal pellets to individual rats.

3.4.3 Population genetics of *Rattus norvegicus*

3.4.3.1 Deviation from Hardy Weinberg equilibrium

Substantial homozygous excess was observed in this study, with all loci affected. As explained in the methods to this chapter, it is possible that some heterozygotes whose alleles differ in length by just one or two dinucleotide repeats were misclassified as homozygotes (details in section 3.2.2.5). If so, this would lead to an artefactual increase in the number of homozygotes observed, above the true number, and this may partly explain the homozygous excess observed.

However, that every locus appeared to have an excess of homozygotes suggests that a real biological process may also contribute to this phenomenon. Rats are known to be poor dispersers, generally remaining within 100 m of their place of birth throughout their lives (Gardner-Santana *et al.* 2009). In this study pellets were often found in clumps over a few square metres, these clumps being scattered across sampling sites several kilometres in diameter. It is possible that these clumps of pellets came from groups of locally resident, inbreeding individuals that are genetically distinct from groups responsible for other pellet clumps. This hypothesis is supported by the observed excess of closely related pairs within sampling sites, as compared with what would be expected from a normal distribution (Figure 3.3). Treatment of genetically distinct groups as a single population leads to the Wahlund effect, where differences in allele frequencies across groups lead to a deficit of homozygotes

in the metapopulation (Garnier-Géré and Cikhi 2013), and this may further contribute to the homozygous excess observed here.

3.4.3.2 Population genetic structure

Shannon statistics for information theory (Shannon 1948) were used in place of F_{ST} and other traditional methods of measuring population differentiation due to the high homozygous excess seen in all sampling sites. This violates the assumption of F_{ST} that the putative subpopulations under examination are at HWE (Weir 2012). Shannon statistics do not make this assumption (Hedrick 2005, Jost *et al.* 2018). Shannon statistics have been used extensively in ecology as a measure of species diversity in ecosystems (Spanos and Feest 2007, Allen *et al.* 2008, Immerzeel *et al.* 2013, Thom and Seidl 2015), but are more rarely applied to population genetics. However, the theory for Shannon statistics in population genetics has been robustly laid out (Sherwin *et al.* 2006), and subsequently $^{S}H_{UA}$ has been applied to the population genetics of the tree *Elaeocarpus sedentarius* (Rossetto *et al.* 2008). This study detected high levels of differentiation among *E. sedentarius* populations, and there was agreement between $^{S}H_{UA}$ and F_{ST} estimates. Hence, $^{S}H_{UA}$ is suitable for use in detecting population genetic structure.

Multiple lines of evidence point to moderate differentiation among rat sampling sites in this study. Allele frequencies differ among sampling sites (Table 3.6), and when individuals were computationally assigned to sampling sites based on site-specific allele frequencies, 89% of rats were correctly assigned. Furthermore, Shannon's information index indicated that 15% of molecular variance in the entire rat dataset partitioned among populations, and $^{S}H_{UA}$ values ranged from 0.15 to 0.22 in pairwise comparisons of sampling sites.

Given the geographical distance between sampling sites in this study, and the very low dispersal of individual rats, it would be surprising if any individual rat migrated from one site to another in its lifetime. However, direct migration of this sort is not required for gene flow. Rather, gradual migration over multiple generations of rats is likely to explain gene flow among the sampling sites.

An unexpected finding here was that genetic distance according to $^{S}H_{UA}$ was lower between sites CA and LA (32 km apart, $^{S}H_{UA} = 0.15$) than between sites LA and AM (9.7 km, $^{S}H_{UA} = 0.22$). Previous studies of brown rat population genetics have often found isolation by distance, where the genetic distance among individuals or sites correlates with geographical distance separating individuals (Gardner-Santana *et al.* 2008, Combs *et al.* 2017). However, isolation by distance is not a universal finding in brown rat population genetics (Kajdacsi *et al.* 2013), and when it is observed it is observed over a distance of less than 2 km (Gardner-Santana *et al.* 2008, Combs *et al.* 2017, 2018). A lack of isolation by distance was also observed amongst wild house mouse populations in the UK (Abolins *et*

al. 2018). A global study of rat population genetics found genetic diversity within cities that was similar to that among geographically close cities (Puckett *et al.* 2016). Thus, it may be that while genetic distance correlates with geographical distance over short ranges, genetic distance rapidly plateaus so that populations 10 km apart are not much more genetically distinct than one 30 km away.

However, this does not explain why rats at sites CA and LA should be genetically closer than those at sites LA and AM, given that the geographical distance between the former pair is three times greater than in the latter. Rats may be subject to 'boom and bust' population dynamics, where populations rapidly grow to exploit a newly available resource and then collapse, either when that resource becomes exhausted or due to external interference such as a pest control campaign. Evidence for boom and bust population dynamics has been observed in other small rodent populations (Wolff 1996, Dickman *et al.* 2010) and has also been suggested to occur in brown rats (Huson and Rennison 1981). It may be that new brown rat populations are initially established by a few individuals, which then breed among themselves to create a genetically distinct population that is refractory to further immigration. If founding individuals are drawn essentially randomly from a regional metapopulation then the diversity in each sampling site will be a random subset of the total diversity in that metapopulation. This combination of frequent population founding by individuals drawn randomly from a metapopulation, and populations becoming refractory to immigration after being founded, has been used to explain the population genetics of the free-living nematodes *Caenorhabditis elegans* and *Pellioditis marina*, which both exhibit strongly structured populations over small geographical scales without isolation by distance (Barrière and Félix 2005, 2007, Derycke *et al.* 2005, 2007). If this hypothesis applies to brown rats it would be unremarkable that sites CA and LA are genetically closer than sites LA and AM. More investigation into rat population genetics at intermediate geographical scales (tens of kilometres) is required to determine how well *C. elegans*-like population dynamics apply to rats

There was no evidence for an individual or group of individuals that were very divergent from others, indicating that all the individuals sampled were from a single metapopulation. This is further evidence that all pellets genotyped were indeed from brown rats, with no inadvertent inclusion of other species.

3.4.4. Conclusion

The work presented in this chapter shows that individual identification and population genetic analyses of wild animals can be carried out using only faecal sampling, and sheds light on the population genetics of brown rats in the southern UK. Only moderate genetic differentiation was detected among the sampled populations, suggesting only a low level of gene flow occurs among rats at these three sites. No isolation by distance was observed, perhaps due to boom-and-bust population dynamics where population founders are drawn randomly from a genetically diverse metapopulation.

3.5 References

Abolins S., Lazarou L., Weldon L., Hughes, L., King E. C., Drescher P., Pocock M. J. O. *et al.* (2018). The ecology of immune state in a wild mamma*l, Mus musculus domesticus*. *PLoS Biology* **16:**e2003538.

Allen B., Kon M. and Bar-Yam Yaneer (2008). A new phylogenetic diversity measure generalizing the Shannon index and its application to phyllostomid bats. *American Naturalist* **174:**236-243.

Apte U., Thompson M. D., Cui S., Liu B., Cieply B. and Monga S. P. S. (2007). Wnt/β-catenin signaling mediates oval cell response in rodents. *Hepatology* **47:**288-295.

Barbosa, S., Pauperio J., Searle J. B. and Alves P. C. (2013). Genetic identification of Iberian rodent species using both mitochondrial and nuclear loci: application to noninvasive sampling. *Molecular Ecology Resources* **13:**43-56.

Barrière A. and Félix M.-A. (2005). High local genetic diversity and low outcrossing rate in *Caenorhabditis elegans* natural populations. *Current Biology* **15:**1179-1184.

Barrière A. and Félix M.-A. (2007). Temporal dynamics and linkage disequilibrium in natural *Caenorhabditis elegans* populations. *Genetics* **176:**999-1011.

Beja-Pereira A., Oliveira R. Alves P. C., Schwartz M. K. and Luikart G. (2009). Advancing ecological understandings through technological transformations in noninvasive genetics. *Molecular Ecology Resources* **9:**1279-1301.

Broquet T. and Petit E. (2007). Quantifying genotyping errors in noninvasive population genetics. *Molecular Ecology* **13:**3601-3608.

Brouat C., Tatard C., Machin A., Kane M., Diouf M., Bâ K. and Duplantier J.-M. (2011). Comparative population genetics of a parasitic nematode and its host community: the trichostrongylid *Neoheligmonella granjoni* and *Mastomys* rodents in south-eastern Senegal. *International Journal for Parasitology* **41:**1301–1309.

Brouat C., Loiseau A., Kane M., Bâ K. and Duplantier J.-M. (2007). Population genetic structure of two ecologically distinct multimammate rats: the commensal *Mastomys natalensis* and the wild *M. erythroleucus* in south-eastern Senegal. *Molecular Ecology* **16:**2985–2997.

Buchan J. C., Alberts S. C., Silk J. B. and Altmann J. (2003). True paternal care in a multi-male primate society. *Nature* **425:**179-181.

Combs M., Byers K. A., Ghersi B. M., Blum M. J., Caccone A., Costa F., Himsworth C. G. *et al.* (2018). Urban rat races: spatial population genomics of brown rats (*Rattus norvegicus*) compared across multiple cities. *Proceedings of the Royal Society of London B: Biological Sciences* **285:**20180245.

Combs M., Puckett E. E., Richardson J., Mims D. and Munshi-South J. (2017). Spatial population genomics of the brown rat (*Rattus norvegicus*) in New York City. *Molecular Ecology* **27:**83-98.

Dallas J. F., Coxon K. E., Sykes T., Chanin P. R. F., Marshall F., Carss D. N., Bacon P. J. *et al.* (2002). Similar estimates of population genetic composition and sex ratio derived from carcasses and faeces of Eurasian otter *Lutra lutra*. *Molecular Ecology* **12:**275-282.

Derycke S., Remerie T., Vierstraerte A., Backeljau T., Vanfleteren J., Vincx M. and Moens T. (2005). Mitochondrial DNA variation and cryptic speciation within the free-living marine nematode *Pellioditis marina*. *Marine Ecology Progress Series* **300:**91-103.

Derycke S., Van Vinckt R., Vanoverbeke J, Vincx M. and Moens J. (2007). Colonization patterns of Nematoda on decomposing algae in the estuarine environment: Community assembly and genetic structure of the dominant species *Pellioditis marina*. *Limnology and Oceanology* **52:**992-1001.

Desvars-Larrive A., Pascal M., Gasqui P., Cosson J.-F., Benoit E., Lattard V., Crespin L. *et al.* (2017). Population genetics, community of parasites, and resistance to rodenticides in an urban brown rat (*Rattus norvegicus*) population. *PLoS One* **12:**e0184015.

Dickman C. R., Greenville A. C., Beh C.-L., Tamayo B. and Wardle G. M. (2010). Social organization and movements of desert rodents during population "booms" and "busts" in central Australia. *Journal of Mammology* **91:**798-810.

Dybdahl M. F. and Lively C. M. (1996). The geography of coevolution: comparative population structures for a snail and its trematode parasite. *Evolution* **50:**2264-2275.

Frantz A. C. Pope L. C., Carpenter P. J., Roper T. J., Wilson G. J., Delahay R. J. and Burke T. A. (2003). Reliable microsatellite genotyping of the Eurasian badger (*Meles meles*) using faecal DNA. *Molecular Ecology* **12:**1649-1661.

Flores-Manzanero A., Luna-Bárcenas M. A., Dyer R. J. and Vázquez-Domínguez E. (2018). Functional connectivity and home range inferred at a microgeographic landscape genetics scale in a desert-dwelling rodent. *Ecology and Evolution* **9:**437-453.

Galan M., Pagès M. and Cosson J.-F. (2012). Next-Generation Sequencing for Rodent Barcoding: Species Identification from Fresh, Degraded and Environmental Samples. *PLoS ONE* **7:**e48374.

Gardner-Santana L. C., Norris D. E., Fornadel C. E., Hinson E. R., Klein S. L. and Glass G.E. (2009). Commensal ecology, urban landscapes, and their influence on the genetic characteristics of city-dwelling Norway rats (*Rattus norvegicus*). *Molecular Ecology*. **18:**2766–2778.

Garnier-Géré P. and Cikhi L. (2013). Population subdivision, hardy–weinberg equilibrium and the Wahlund effect. In: *eLS*. Chichester: John Wiley.

Giraudeau F., Apiou F., Amarger V., Kaisaki P. J., Bihoreau M.-T., Lathrop M., Vergnaud G. *et al.* (1999). Linkage and physical mapping of rat microsatellites derived from minisatellite loci. *Mammalian Genome* **10:**405-409.

Gotelli D., Wang J., Bashir S. and Durant S. M. (2007). Genetic analysis reveals promiscuity among female cheetahs. *Proceedings of the Royal Society of London B: Biological Sciences* **274:**1993-2001.

Guo S., Li G. C., Liu J. L., Wang J., Lu L. and Liu Q. Y. (2019). Dispersal route of the Asian house rat (*Rattus tanezumi*) on mainland China: insights from microsatellite and mitochondrial DNA. *BMC Genetics* **20:**11.

Hale M. L., Bevan R. and Wolff K. (2001). New polymorphic microsatellite markers for the red squirrel (*Sciurus vulgaris*) and their applicability to the grey squirrel (*S. carolinensis*). *Molecular Ecology Resources* **1:**47-49.

Hedrick P. W. (2005). A standardized genetic differentiation measure. *Evolution* **59:**1633-1638.

Heiberg A.-C., Sluydts V. and Leirs H. (2012). Uncovering the secret lives of sewer rats (*Rattus norvegicus*): movements, distribution and population dynamics revealed by a capture–mark–recapture study. *Wildlife Research* **39:**202-219.

Huson L. W. and Rennison B. D. (1981). Seasonal variability of Norway rat (*Rattus norvegicus*) infestation of agricultural premises. *Journal of Zoology* **194:**257-289.

Immerzeel D. J., Verweij P. A., van der Hilst, F. Faaij A. P. C. (2013). Biodiversity impacts of bioenergy crop production: a state-of-the-art review. *GCB Bioenergy* **6:**183-209.

Iyengar A., Babu V. N., Hedges S., Venkataraman A. B., Maclean N. and Morin P. A. (2005). Phylogeography, genetic structure, and diversity in the dhole (*Cuon alpinus*). *Molecular Ecology* **8:**2281-2297.

James R. S., James P. L., Scott D. M., Overall A. D. J. and Bahler J. (2015). Characterization of six cross-species microsatellite markers suitable for estimating the population parameters of the black-backed jackal (*Canis mesomelas*) using a non-invasive genetic recovery protocol *Cogent Biology* **1:**1108479.

Jobet E., Durand P., Langand J., Müller-Graf C. D. M., Hugot J.-P., Bounnoux M.-E., Rivaut C. *et al.* (2000). Comparative genetic diversity of parasites and their hosts: population structure of an urban cockroach and its haplo-diploid parasite (oxyuroid nematode). *Molecular Ecology* **9:**481–486.

Jost L., Archer F., Flanagan S., Gaggiotti O., Hoban S and Latch E. (2018). Differentiation measures for conservation genetics. *Evolutionary Applications* **11:**1139-1148.

Kajdacsi B., Costa F., Hyseni C., Porter F., Brown J., Rodrigues G., Farias H. *et al.* (2013). Urban population genetics of slum-dwelling rats (*Rattus norvegicus*) in Salvador, Brazil. *Molecular Ecology* **22:**5056-5070.

Lynch M. and Ritland K. (1999). Estimation of pairwise relatedness with molecular markers. *Genetics* **152:**1753-1766.

Mikesic D. G. and Drickamer L. C. (1992). Factors affecting home-range size in house mice (*Mus musculus domesticus*) living in outdoor enclosures. *American Midland Naturalist*. **127**:31-40.

Montgomery W. I., Wilson W. L., Hamilton R. and McCartney P. (1991). Dispersion in the wood mouse, *Apodemus sylvaticus*: variable resources in time and space. *Journal of Animal Ecology* **60:**179-192.

Moran S., Turner P. D. and O'Reillev C. (2008). Non-invasive genetic identification of small mammal species using real-time polymerase chain reaction. *Molecular Ecology Resources* **8:**1267-1269.

Moska M., Mucha A. and Wierzbicki H. (2018). Genetic differentiation of the edible dormouse (*Glis glis*) in the Polish Sudetens: the current status of an endangered species. *Journal of Zoology* **305:**203-211.

Nicolas V., Martínez-Vargas J. and Hugot J.-P. (2017). Molecular data and ecological niche modelling reveal the evolutionary history of the common and Iberian moles (Talpidae) in Europe. *Zoologica Scripta* **46:**12-26.

Nieberding C., Morand S., Libois R., Michaux J. R. (2004). A parasite reveals cryptic phylogeographic history of its host. *Proceedings of the Royal Society of London B: Biological Sciences* **271:**2559–2568.

Osten-Sacken N., Heddegrott M., Schleimer A., Anheyer-Behmenburg H. E., Runge M., Horsburgh G. J., Camp L. *et al*. (2018). Similar yet different: co-analysis of the genetic diversity and structure of an invasive nematode parasite and its invasive mammalian host. *International Journal for Parasitology* **48:**233–243.

Palomares F., Adrados B., Zanin M., Silveira L. and Keller C. (2017). A non-invasive faecal survey for the study of spatial ecology and kinship of solitary felids in the Viruá National Park, Amazon Basin. *Mammal Research* **62:**241-249.

Peakall R. and Smouse P. E. (2006). GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes* **6:**288-295.

Peakall R. and Smouse P. E. (2012). GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research-an update. *Bioinformatics* **28:**2537-2539.

Pocock J. O., Hauffe H. C. and Searle J. B. (2005). Dispersal in house mice. *Biological Journal of the Linnean Society* **84:**565-583.

Puckett E. E., Park J., Combs M., Blum M. J., Bryant J. E., Caccone A., Costa F., Deinum E. E. *et al.* (2016). Global population divergence and admixture of the brown rat (*Rattus norvegicus*). *Proceedings of the Royal Society of London B: Biological Sciences* **283:** 20161762.

Rossetto M., Kooyman R., Sherwin W. and Jones R. (2008) Dispersal limitations, rather than bottlenecks or habitat specificity, can restrict the distribution of rare and endemic rainforest trees. *American Journal of Botany* **95:**321-329.

Shannon C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal* **27:**379-423.

Sherwin W. B., Jabot F., Rush R. and Rossetto M. (2006). Measurement of biological information with applications from genes to landscapes. *Molecular Ecology* **15:**2857-2869.

Spanos K. A. and Feest A. (2007). A review of the assessment of biodiversity in forest ecosystems. *Management of Environmental Quality* **18:**475-486.

Steen R. G., Kwitek-Black A. E., Glenn C., Gullings-Handley J., Van Etten W., Atkinson O. S., Appel D. (1999). A high-density integrated genetic linkage and radiation hybrid map of the laboratory rat. *Genome Resources* **9:**AP1-8.

Taberlet P., Waits L. P. and Luikart G. (1999). Noninvasive genetic sampling: look before you leap. *Trends in Ecology and Evoloution* **14:**323-327.

Thom D. and Seidl R. (2015). Natural disturbance impacts on ecosystem services and biodiversity in temperate and boreal forests. *Biological Reviews* **91:**760-781.

Weir B. S. (2012). Estimating F-statistics: A historical view. *Philosophy* of S*cience* **79:**637-643.

Wilson G. J., Frantz A. C., Pope L. C., Roper T. J., Burke T. A., Cheeseman C. L. and Delahey D. J. (2003). Estimation of badger abundance using faecal DNA typing. *Journal of Applied Ecology* **40:**658-666.

Wolff J. O. (1996). Population fluctuations of mast-eating rodents are correlated with production of acorns. *Journal of Mammology* **77:**850-856.

Zhan X. J., Zhang Z. J., Wu H., Goosens B., Li M., Jiang S. W., Bruford M. W. *et al.* (2007). Molecular analysis of dispersal in giant pandas. *Molecular Ecology* **16:**3792-3800.

Zhang J.-S., Daszak P., Huang H.-L., Yang G.-Y., Kilpatrick A. M. and Zhang S. (2008). Parasite threat to panda conservation. *EcoHealth* **5:**6-9.

Chapter 4. Population genetics of *Strongyloides ratti*

4.1 Introduction

4.1.1 Background

Parasitic nematodes are ubiquitous in natural ecosystems and undoubtedly have important roles in ecosystem functioning, but little is known about their ecology outside of medical and agricultural settings (Wood and Johnson 2015, Frainer *et al.* 2018). Population genetics is an important aspect of a species' ecology that, in parasitic nematodes, can inform on life-history traits, host and parasite phylogeography, transmission among hosts and the epidemiology of nematode-borne diseases. Further, population genetics can help to reveal population-level responses of parasites to environmental change or novel selection pressures such as the use of anthelmintic drugs (Nieberding *et al.* 2004, Criscione *et al.* 2005, Gorton *et al.* 2012, Gilabert and Wasmuth 2013). However, of the enormous abundance of parasitic nematodes in wild animals, only a tiny minority have had their population genetics studied / investigated (Cole and Viney 2018).

4.1.2 Whole-genome sequencing in population genetics

Whole-genome sequencing allows the relationships among individuals, and hence their population genetics, to be resolved at the finest possible scale (Li and Durbin 2011). Furthermore, it gives great power to making inferences about the evolutionary processes acting on a species (Jones *et al.* 2012, Ott *et al.* 2015, Nosil *et al.* 2018). Of particular importance in parasites is the ability of whole-genome sequencing to identify genes that are under selection, which may be relevant to understanding pathogenicity, epidemiology and control. For example, whole-genome sequencing of the malaria parasite *Plasmodium falciparum* revealed genes involved in evasion of host immunity or resistance to drugs, and these genes showed signs of selection (Mobegi *et al.* 2014). Similar studies in parasitic nematodes are warranted, as anthelmintic resistance is an increasingly serious problem in both agricultural and medical settings (Coles *et al.* 2006, Gilleard and Redman 2016). Whole-genome sequencing of multiple individuals from a population exposed to an anthelmintic could reveal not only candidate resistance genes (Choi *et al.* 2017), but also how these alleles are spreading or are likely to spread giving the underlying population genetic structure.

Cost remains an obstacle to the routine use of whole-genome sequencing in this way, causing a trade-off between the number of individuals sequenced and the depth coverage of each individual's genome. Naturally, this difficulty is ameliorated in species with small genome sizes, such as many parasitic nematodes.

Whole-genome sequencing has already been used to study the population genetics of parasitic nematodes that infect people, demonstrating among-host differentiation in 13 *Wuchereria bancrofti* (Small *et al.* 2016), and detecting among-host and among-region differentiation in 33 *Strongyloides stercoralis* (Kikuchi *et al.* 2016). However, these studies had small sample sizes and relied on whole-genome amplification (Hosono *et al.* 2003) prior to sequencing. Whole-genome amplification is useful when limited amounts of DNA in an animal makes extracting enough for DNA for good coverage difficult, (for example if the animal is very small animal), but can introduce bias in sequencing (Sabina *et al.* 2015). No population genetic studies using a whole-genome approach have yet considered a nematode that infects wildlife.

4.1.3 *Strongyloides ratti* genome reference assembly

Strongyloides ratti is a common parasite of brown rats (*Rattus norvegicus*) and has congeners that infect humans and wild and domestic animals. The reference assembly of the *S. ratti* genome was built from the standard laboratory line and is highly complete (Hunt *et al.* 2016). It is publicly available at WormBase ParaSite (Howe *et al.* 2017). The haploid *S. ratti* nuclear genome consists of approximately 40 million bases distributed across three chromosomes (Hunt *et al.* 2016). Two of these chromosomes, chromosomes 1 and 2, are autosomes and are present in homologous pairs in all individuals, while the X chromosome is paired in females and haploid in males. Chromosome 2 is assembled into a single 16 Mb scaffold. Chromosome 1 is in a single scaffold that consists of three parts, collectively containing approximately 9 Mb. The chromosome 1 parts are separated on the scaffold by blocks of 'N's each 1 Mb long, as the respective orientations of these three parts in the *in vivo* chromosome are unknown. The X chromosome is in 10 scaffolds that range in length from 73 to 4,956 kb, with a combined length of 11.8 Mb. The two largest X chromosome scaffolds together measure 8.4 Mb. In addition, a further 123 scaffolds ranging from 0.18 to 858 kb in length are not assigned to a chromosome, but are presumed autosomal. These unassigned scaffolds have a combined length of ~2.9 Mb. In total, 50% of the bases are on scaffolds of 11.7 Mb or more in length. The genome is highly AT-rich, with an AT composition of 78%.

The *S. ratti* mitochondrial genome is fully assembled into a single scaffold approximately 16.7 kb long. Mitochondrial gene order is unconventional in *Strongyloides* sp., with extensive rearrangements in each species, hence *S. ratti* does not show the usual nematode mitochondrial gene order (Hunt *et al.* 2016).

4.1.4 Whole-genome sequencing in *Strongyloides ratti* population genetics

At 40 Mb, the genome of *S. ratti* is small compared to most animals, and so is relatively cheap to sequence, while the highly complete state of the reference assembly facilitates sequencing-based population genetic analyses. Thus *S. ratti* is an ideal candidate for a population genetic study based on

whole-genome sequencing. Only one study has previously investigated the population genetics of *Strongyloides ratti* (Fisher and Viney 1998), and this study of UK populations used only three genetic markers. Hence, while no genetic differentiation was observed, it may be that the genetic markers used were not sensitive enough to detect weak population genetic structuring. Whole-genome sequencing would avoid this problem, allowing a better understanding of the population genetics of *S. ratti* in the UK.

Aside from the reference assembly, ten *S. ratti* genomes have been sequenced to date (Table4.1). Each of these genomes was generated from an isofemale line, a group of individuals derived from a single parasitic female recovered from a wild rat. More genome sequences are needed, however, and generating these via the establishment of new isofemale lines would be time consuming and require extensive use of laboratory rats as passage hosts. Furthermore, some genotypes might be better able to establish infections under laboratory conditions than others, leading to bias. Thus, generation of genome sequences of individual infective larvae (iL3s) taken straight from the natural environment is much preferred.

In this chapter the population genetics of *S. ratti* is investigated using the whole-genome sequences of ninety individuals that were isolated directly from faeces found in three wild rat populations, with no laboratory passage, and sequenced individually. Hence, the sequenced individuals are direct representatives of the populations from which they were drawn.

4.2 Materials and methods

4.2.1 *Strongyloides ratti* samples used in this chapter

Sampling of *S. ratti* iL3s from wild hosts is described in detail in Chapter 2. Briefly, sampling was from 3 sampling sites in the Southern UK, over four sampling seasons (corresponding to calendar seasons) over the course of a year (2016-17). iL3s were isolated directly from faeces of wild hosts and immediately placed into storage at -80°C with no laboratory passage. Analyses were then performed on single iL3s after thawing.

In addition, the whole-genome sequences of 10 *S. ratti* isofemale lines were included in some analyses. The isofemale lines had been sequenced previously with Illumina technology in collaboration between the Wellcome Trust Sanger Institute (WTSI) and the University of Bristol, and the resulting sequencing reads were provided by colleagues at the University of Bristol. Variant detection was previously carried out on these lines and the results are publicly available at WormBase ParaSite. A breakdown of the *S. ratti* samples used in analyses in this chapter is provided in Table 4.1.

Table 4.1: Strongyloides ratti *genomes used in this chapter. Two datasets were considered. DS90 contains 90 individuals collected for this study from wild rats (A), while DS100 contains all these individuals as well as 10 laboratory lines (B), each founded from a single wild-collected female ("isofemale lines").*

A. Individual worms contributing to DS90 and DS100			
Sampling site	**Sampling season**	**Initial sequencing,** No. nematodes (hosts)	**Deep sequencing** No. nematodes (hosts)
CA	Spring	0	0
	Summer	0	0
	Autumn	19 (2)	9 (2)
	Winter	6 (1)	4 (1)
	Total	**25 (3)**	**13 (3)**
LA	Spring	0	0
	Summer	0	0
	Autumn	25 (5)	6 (4)
	Winter	25 (5)	14 (4)
	Total	**50 (10)**	**20 (7)**
AM	Spring	30 (7)	9 (3)
	Summer	30 (7)	5 (4)
	Autumn	30 (11)	6 (5)
	Winter	60 (16)	37 (16)
	Total	**150 (41)**	**57 (28)**

B. Isofemale lines contributing to DS100		
Isofemale line	**Isolation year**	**Origin**
ED36	1990	Hampshire, UK
ED43	1989	Edinburgh, UK
ED53	1990	Kagoshima, Japan
ED132	1990	Kagoshima, Japan
ED336	1995	Berkshire, UK
ED391	1989	Wiltshire, UK
ED399	1989	Sussex, UK
ED405	1989	Sussex, UK
ED428	2012	Bath, UK
ED438	2012	Bath, UK

4.2.2 Assessment of *Strongyloides ratti* diversity within faecal pellets

4.2.2.1 Rationale

Sequencing the genomes of multiple *S. ratti* individuals from a single faecal pellet could give a valuable insight into population genetic structure within and among *S. ratti* infrapopulations. However, the use of mitotic parthenogenesis by *S. ratti* parasitic females means that multiple, genetically identical siblings may be present in a single faecal pellet, and sequencing these identical siblings would be uninformative. RFLP was therefore used to initially assess the typical levels of *S. ratti* diversity within pellets prior to further genome sequencing. If *S. ratti* diversity within pellets is high, then it is worth sequencing multiple individuals from the same pellet. However, if pellets typically harbour only a very few *S. ratti* genotypes, then just sequencing a single individual from each pellet might be a more efficient use of sequencing resources.

4.2.2.2 Extraction of DNA from single *Strongyloides ratti*

iL3s to be used for RFLP analysis were dried prior to the addition of 11.08 µL of 50 mM NaOH and incubation at 95°C for 20 minutes. The solution was then neutralised with 0.92 µL of 1 M Tris (pH 8). 1 µL of the resulting solution was used as template DNA in subsequent PCR reactions

4.2.2.3 Selection of loci for RFLP analysis

Genetic variant data for 7 *S. ratti* isofemale lines (ED43, ED336, ED391, ED399, ED405, ED428 and ED438, Table 4.1B) was downloaded from WormBase ParaSite. Each of these isofemale lines was originally isolated from a wild rat in the UK (Table 4.1B). Eighteen single nucleotide polymorphism (SNP)s among these 7 lines were identified, where each SNP occurrred within the recognition site of a commercially available restriction endonuclease (Table 4.2). In each case, the ‘reference’ (R) allele would result in cleavage by the restriction endonuclease, while the ‘alternate’ (A) allele would not. Furthermore, in each case, at least one recognition site for the same restriction endonuclease was present within 500 bp of the SNP, and was invariant among the isofemale lines. For each locus, PCR primers

(Table 4.2) were designed to amplify an ~1,000 bp region including the SNP site and the invariant copy (or copies) of the restriction recognition site.

Table 4.2: Loci tested for use in restriction fragment length polymorphism of Strongyloides ratti. *Each locus is a single nucleotide polymorphism (SNP), variable among 8 laboratory lines initially isolated from wild rats in the UK, in which the reference allele is cut by the indicated restriction enzyme and the alternate allele is not. Primers are listed with the forward primer first in 5' to 3' orientation. Tm is the annealing temperature used during polymerase chain reaction. In SNP position, "1", "2", "X1" and "X2" in bold indicate chromosome 1, chromosome 2, scaffold 1 of the X chromosome and scaffold 2 of the X chromosome, respectively, and the following value indicates the position of the base pair position of that SNP on that scaffold. Coloured rows indicate loci that could be reliably amplified. Those coloured in red were found to be variable among 21 individual* S. ratti.

Locus	**Primers**	**Tm (°C)**	**Restriction enzyme**	**SNP position**
Sr1-1	ACCAACCGTTGTCATTGGAT GGCACCGTGAAAGTATCTTC	50	HpyCH4IV	**1:** 1,283,200
Sr1-2	CGTGCAGCTCCAATGGAAAA ACGCTATAGATGACACATCCAGT	52	MboII	**1:** 11,272,309
Sr1-3	ACCGCCATTGTGGGAAAAAG TCAGCTGCAGTTGTCAAATTTAT	52	Hpy188I	**1:** 275,049
Sr1-4	TGCTCCAGTATCAGCTTCAGT GTGAAATTCCTGGCCCACCTA	54	HpyCH4III	**1:** 6,025,942
Sr1-5	TCGTATATCCCATGTTACAGTTGA TGTCTTTTCGGGAAATACCCAG	50	Hpy188III	**1:** 247,039
Sr1-6	TGTCCAGTGTTACATTCTTTGGA AGTGTGTGCCTGATGTGTCA	54	MnoII	**1:** 8,439,436
Sr2-1	ACACCATTTTCCAAGACATTCTTTT CACGAGCTGCATTGTCTGC	50	DdeI	**2:** 153,656
Sr2-2	GAGATGTGCAAGATTACTGTGCC CGGCATATTAACTGCTCAAGGG	52	PsiI	**2:** 8,094,503
Sr2-3	TGTTCAAGGAACACCACTTCA CCCACACTGTTACTGGTCCC	58	AseI	**2:** 16,115,044
Sr2-4	ACAGCTACAATGGATGATTGGGA AAAAACCGTGGATGGGGGTT	56	SwaI	**2:** 8,121,841
Sr2-5	ACTTTACCTACAAAAACGATGTTCA AGGATTTAGGAGCAGGTTGAGC	52	PsiI	**2:** 16,241,509
Sr2-6	AGGCACTTTAAGATAGTTC AACGGTCCATTAAAAGGAAAAGAA	48	Hpy188III	**2:** 265,050
SrX-1	TGATAATGACGCCACTGACACA CCTGCCGTCTCCCAATAAAAG	50	SspI	**X1:** 1,912,418
SrX-2	TTGTATGCTTAGCCGCACTT TTCCGAAGTTTAGCACATTGATT	50	HindIII	**X2:** 1,892,392
SrX-3	GGGAGTGCTAAGTCCCAGTAGA AGTTCATCGAAGACCATGGAGAT	56	ApoI	**X1:** 1,969,467
SrX-4	ACACATTACAAATTGGTATACTCCC TCGTCGCTTCATAAATACGCA	58	ApoI	**X2:** 2,531,701
SrX-5	TGTTCAGGTTGTCTTATTCATTTGC TGTTGTAGTGAGTGGCCTTCT	54	DraI	**X1:** 1,343,025
SrX-6	TGTACAGCCATCTGGATCAACC TGTCTACATCGTACAAAAACGAAGA	54	HinfI	**X2:** 2,702,926

4.2.2.4 Polymerase chain reaction (PCR) to amplify RFLP loci

In all PCRs, reactions contained 200 µM each of dATP, dCTP, dGTP and dTTP, 0.25 µM of each primer (purchased from Integrated DNA Technologies) and 1 µL of template DNA (concentration unknown), with DreamTaq DNA polymerase and DreamTaq buffer (Thermo Scientific) added according to manufacturer instructions. The final reaction volume was 20 µL. The thermocycler regimen was 95°C for 3 minutes, then 45 cycles of 95°C for 30 seconds, annealing temperature (Table 4.2) for 30 seconds, and 72°C for 1 minute, then a final extension of 72°C for 10 minutes. PCR products were visualised by gel electrophoresis (1% w/v agarose gel with 0.5 µg/mL ethidium bromide, run at 100 V for 40 minutes.

4.2.2.5 Restriction digest of PCR products

The 18 primer pairs were first tested for amplification efficiency using DNA preparations from pooled iL3s of a laboratory line (ED321) as a template. Eight primer pairs were found to amplify to some extent (Table 4.2), and the PCR products of these were subsequently subjected to restriction digestion using the appropriate restriction endonuclease (Table 4.2, restriction endonucleases purchased from New England Biolabs) according to the manufacturer's instructions. Restriction products were visualised by gel electrophoresis as described for PCR products.

4.2.2.6 Assessment of *Strongyloides ratti* within faecal pellets

Variability of the 8 loci was tested on 20 wild-isolated iL3s. These were isolated from 7 faecal pellets representing all three sampling sites (Table 4.3). Three loci, Sr1-3, Sr1-4 and Sr2-4, were found to be variable. Banding patterns for each allele of these loci is shown schematically in Figure 4.1. Sixty-seven further iL3s from 3 faecal pellets and all sampling sites were subsequently genotyped at the Sr1-4 and Sr2-4 loci (Table 4.3). These additional iL3s were not genotyped at Sr1-3 due to difficulty achieving reliable amplification of this locus.

Table 4.3: Strongyloides ratti *individuals used for the restriction fragment length polymorphism study.*

Sampling site	**Genotyped at 8 loci** No. nematodes (hosts)	**Genotyped at Sr1-4 & Sr2-4** No. nematodes (hosts)
CA	3 (1)	12 (3)
LA	10 (3)	31 (4)
AM	7 (3)	24 (4)

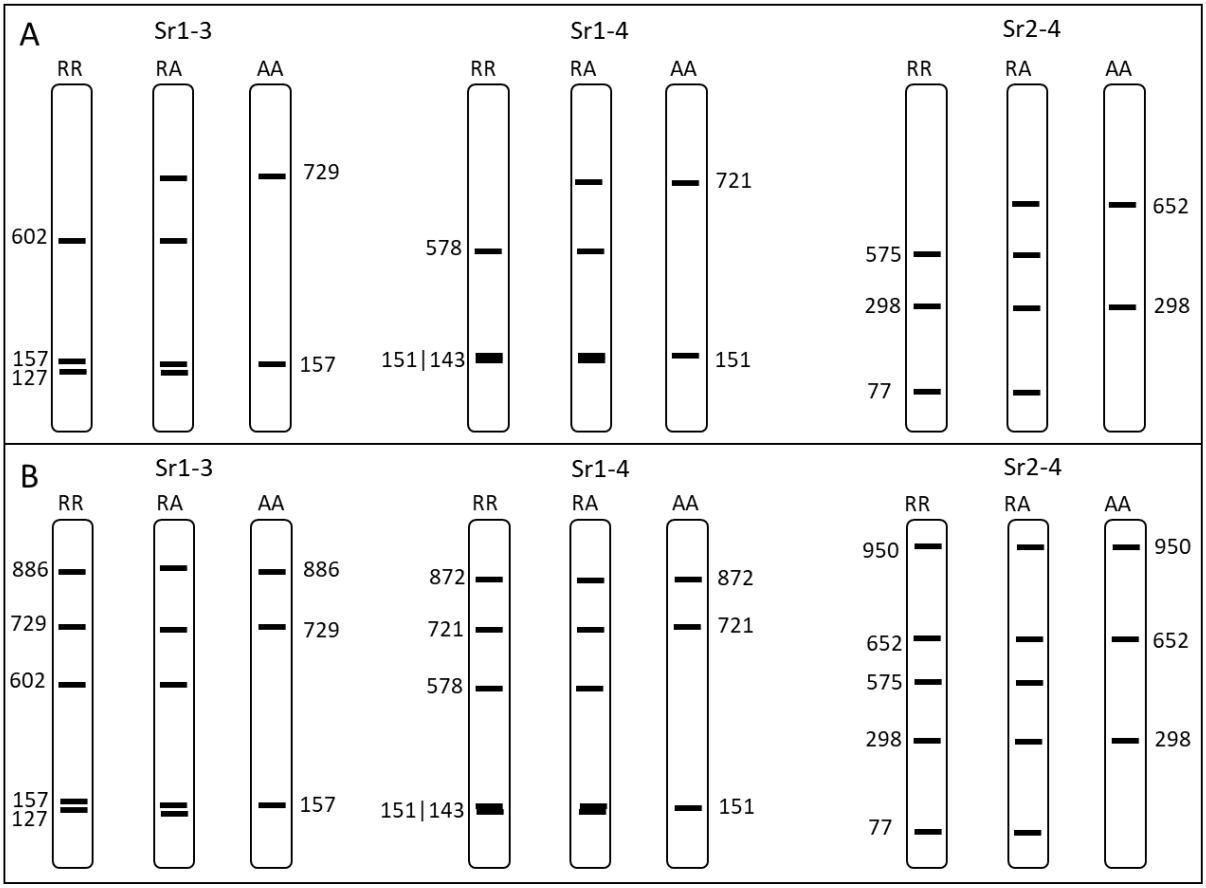


Figure 4.1: Banding patterns produced by restriction fragment length polymorphism analysis of Strongyloides ratti *using three loci (Sr1-3, Sr1-4 and Sr2-4). A shows hypothetical banding patterns following complete restriction endonuclease digestion. B shows banding patterns produced if there is incomplete digestion of both the full-length amplicon product (longest band in each locus) and the band produced by cleavage of the restriction site common to all alleles. Distances between alleles are illustrative only and not exactly to scale.*

4.2.3 Selection of single *Strongyloides ratti* samples for sequencing

A sequencing strategy was devised that would sequence the whole genomes of 225 single iL3s, incorporating individuals derived from all 3 sampling sites and all 4 sampling seasons (Table 2.1). In order to test the effect of site without season, iL3s were taken from all 3 sites in both autumn and winter. In order to test the effect of season without site, iL3s were taken from all 4 seasons at site AM (Table 4.1A). An effort was made to take no more than 5 iL3s per faecal pellet, to further reduce the risk of sequencing multiple clonal siblings. However, this was not possible in some cases, especially at site CL, where the prevalence of infection was low and where there was limited choice of infected pellets from which to take iL3s. After this sequencing, genotyping of rat faecal pellets showed that there were 4 cases were 2 pellets that provided iL3s for sequencing were from the same rat (Chapter 3). Pellets

coming the same rat were pooled for all host-based analyses. The number of iL3s taken from a single pellet ranged from 1 to 10 with a mean of 4.

4.2.4 Whole-genome sequencing of individual *Strongyloides ratti*

4.2.4.1 Preparation of samples for sequencing

Single iL3s to be sequenced (Table 4.1A) were moved individually in water to the wells of 96-well plates, if not originally stored in them. All further sample handling was carried out in these plates, so that no DNA was lost in transfer of material to a new receptacle.

Prior to lysis, excess water was removed from sample wells, leaving only the iL3 in a tiny (approximately 2 μL) drop. Lysis was carried out in 'worm genomic DNA lysis buffer', consisting of 200 mM NaCl, 100 mM Tris-HCl (pH8.5), 50 mM EDTA (pH8) and 0.5% w/v SDS. Proteinase K and dithiothreitol were added to a final concentration of 0.9mg/mL and 45 mM respectively immediately before buffer was added to wells. Samples in lysis buffer were shaken at 60°C for 2 hours, then Proteinase K was inactivated by incubation of samples at 85°C for 15 minutes. Distilled water was added to lysates to achieve a final volume of 120 μL, and lysates were then stored at -80°C. This protocol was provided by Magda Lokowska at WTSI (personal communication). In the final preparatory step, sequencing libraries were generated from samples by the WTSI.

4.2.4.2 Initial whole-genome sequencing

Single iL3s were isolated from cultures of wild rat faeces, an environment likely to be rich in bacteria. Further, other organisms such as non-*S. ratti* nematodes, small arthropods, tardigrades, fungi, and single-celled eukaryotes were frequently observed during inspection of cultures under a dissection microscope, these having presumably come from the environment the pellet was taken from. Therefore, although *S. ratti* were washed extensively in both water and SDS prior to lysis (detailed in Chapter 2), it is likely that non-*S. ratti* contaminant DNA remained after lysis and was incorporated into sequencing libraries.

Therefore, the first objective was to identify libraries that had a high proportion of reads originating from *S. ratti*. To do this, all 225 sample libraries were sequenced over two lanes of an Illumina X Ten sequencing machine at the WTSI. One lane contained 151 samples, while the other contained the remaining 74 samples. While this would provide relatively few sequencing reads per sample, the proportion of those reads mapping to *S. ratti* would indicate the degree of contamination of each sample library. Sequencing reads were trimmed and filtered according to standard WTSI processing to ensure read quality, and the resulting 100 bp reads were then returned for analysis.

Sequencing reads from each of the 225 libraries were aligned to *S. ratti* reference assembly version 5_0_4 (Hunt *et al.* 2016, taken from WormBase ParaSite release WBPS7) using Bowtie 2 version 2.2.9 (Langmead and Salzberg 2012) with default settings. In each case, only a proportion of the reads produced by sequencing a library mapped to the *S. ratti* reference genome. Non-mapping reads presumably derive from contamination with non-*S. ratti* DNA. Hence, the proportion of reads mapping to the reference genome was taken as an indicator of the quality of the sample library, with a higher proportion of reads mapping to *S. rtti* suggesting a greater ratio of *S. ratti* DNA to contaminant DNA in the library.

4.2.4.3 Deep sequencing of high-quality libraries

Following initial sequencing, 90 libraries were selected for further analyses. In general, the libraries with the highest proportion of reads mapping to the reference genome during initial sequencing were chosen, but priority was given to *S. ratti* individuals from the under-represented sites LA and CA. The number of individuals taken from each sampling site / season combination is shown in Table 4.1A. the number of deep-sequenced worms taken from the same host ranged from 1 to 7 (mean 2.3), and 39 hosts were represented in total.

Deep sequencing was carried out as for the initial sequencing except that the 90 libraries were split over 13 sequencing lanes with a mean of 7 samples per lane. By reducing the number of samples per lane, far more reads per sample were generated than in initial sequencing, allowing for increased read depth and hence improved accuracy in detecting sequencing variants. Reads generated in the deep sequencing of each sample were combined with the reads produced in initial sequencing of the same sample, and together these reads were aligned to the *S. ratti* reference genome. Alignments were stored in the binary alignment map (BAM) format and handled by SAMtools version 1.2 (Li *et al.* 2009). SAMtools was also used to filter out reads that did not map to the *S. ratti* reference genome.

Two measures – percentage of reference genome covered by at least one read, and average read depth where depth is greater than 0 – were used as measures of the quality of the 90 alignments. Furthermore, the two autosomes and two largest X chromosome fragments were divided into non-overlapping 10 kb windows, and average read depth within each window was calculated for each individual to detect variation in sequencing depth across the genome. Finally, reads from all 90 individuals were pooled, and the cumulative number of reads covering a particular base (cumulative read depth) was calculated. To account for the potential influence of GC content affecting read depth, mean cumulative read depth was plotted against GC content for each 10 kb window.

4.2.5 Analysis of polymorphism in *Strongyloides ratti*

4.2.5.1 SNP calling

SNPs among all 90 individual *S. ratti* genomes were detected using BCFtools version 1.2, which is packaged with SAMtools (Li 2011). BCFTools takes BAM files from multiple individuals along with a reference genome and generates multi-sample variant call format (VCF) files that list every nucleotide at which a sample differs from the reference, giving the genotype of each individual at each position.

Sequencing reads for the 10 isofemale lines were aligned to the reference genome and the resulting alignments processed as described for single iL3s subjected to deep sequencing. These were then combined with the 90 individual *S. ratti* alignments, and the 100 alignments were subjected to variant calling together. Thus, two sets of data were produced, one with 10 isofemale lines and 90 individuals (DS100), and one with only the 90 individuals (DS90).

4.2.5.2 Filtering of variants

SNPs detected were filtered with BCFtools. SNPs were retained if they 1) fell on a nucleotide covered by at least 1,000 reads (cumulative across all samples), 2) had a mean mapping quality of at least 20, and 3) had a QUAL score of at least 50. Mapping quality indicates the mean of the likelihoods that each read supporting the SNP has been aligned correctly to the reference genome, which is based on the uniqueness of the reference sequence the reads align. QUAL is the mean Phred score of the bases (as opposed to the whole reads) covering the SNP (Danecek *et al.* 2011). In both cases, mapping quality and QUAL are calculated by BCFtools. In DS90, nucleotides that were identical among all samples (but different from ED321) were removed.

4.2.5.3 Calculation of genetic diversity and population genetic parameters

Basic genetic diversity and population genetic statistics were calculated by VCFtools version 0.1.12 (Danecek *et al.* 2011), using the VCF files produced by variant calling as the input file. Hardy-Weinberg equilibrium (HWE) was determined for SNPs in DS90, considering only biallelic SNPs. The percentage of SNP positions expected to be homozygous in a given individual was calculated from allele frequencies across the entire DS90 population, and this was compared with the actual percentage of homozygous SNP positions to therefore calculate the per-individual homozygosity. The percentage of expected homozygosity differed marginally among samples due to missing data for some loci in some individuals

The number of SNPs in every possible pairwise comparison of individuals was counted and used to count SNPs per kb for each pair of samples, which was called the pairwise distance. Φ relatedness

values (Manichaikul *et al.* 2010) of each pair were also calculated. A t-test was used to compare Φ relatedness of samples pairs where both individual samples were from the same host (same-host pairs) with those that were from different hosts (different-host pairs). Similarly, a t-test was used to compare Φ relatedness of sample pairs where both individual samples were from the same sampling site (same-site pairs) with those that were from different sites (different-site pairs). The goal of these analyses was to measure differentiation among hosts and among sampling sites. As an additional measure of differentiation among sites, the fixation index (F_{ST}) was calculated among all three sampling sites, as well as in pairwise combinations of sites. Treating each host as a separate *S. ratti* population would leave population sizes too small for F_{ST} analysis. F_{ST} and Φ Relatedness values were calculated to detect differentiation among sampling seasons in the same way as for sampling sites, except that only individuals from site AM were considered.

4.2.5.4 Principal component analysis

Principal component analysis (PCA) is a well-established method of clustering individuals according to genetic similarity (Lee *et al.* 2009, McVean 2009). Here, PCA was carried out in the R package pcadapt version 4.1.0 (Luu *et al.* 2017) with the DS100 and DS90 multi-sample VCF files used as input. In each case, only loci with a minor allele frequency greater than 0.05 were used. The percentage of variance explained by each of the first 30 principal components (PCs) was examined with scree plots, and this informed the choice of number of PCs to consider of 2. Subsequently individuals were projected onto the first 2 principal components.

4.2.5.5 Generation of nuclear relatedness dendrograms

TASSEL 5.0 (Bradbury *et al.* 2007) was used these to produce neighbour joining dendrograms for DS90 and DS100, with VCF files used as input. These were subsequently visualised and prepared for presentation in FigTree Version 1.4.3, a free software package available from http://tree.bio.ed.ac.uk/software/figtree/, (downloaded 04/10/2016).

Genetic clades in neighbour joining trees were identified by eye. Fisher's exact tests, performed in R, were used to determine whether there were significant differences in the frequencies of these clades among sampling sites or sampling seasons. χ^2 tests could not be used reliably due to many expected values for numbers of clade per site / season being less than 5.

4.2.5.6 Assessment of linkage disequilibrium

Linkage disequilibrium (LD) was only assessed in DS90, and only on the 2 autosomes and the 2 largest scaffolds of the X chromosome. In order to assess LD, it was necessary to know which homologue each allele at a SNP locus was on. That is, it was necessary to phase the genotype data into haplotypes. Two software packages, Beagle version 5.0 (Browning and Browning 2007, Browning *et al.* 2018) and

Shapeit version 2-r900 (O'Connell *et al.* 2014), were used to phase the SNP data. Beagle was used with 100 burn-in iterations to generate an initial estimation of haplotype frequency, and a further 100 iterations were used to estimate genotype phase for each SNP in each sample. Phasing is influenced by the effective population size (Ne). This is not known for *S. ratti*, but an estimate of 50,000 was provided. Otherwise, default Beagle parameters were used. Shapeit was also used with 100 burn-in iterations, 100 phasing iterations, and an estimated Ne of 50,000. Furthermore, the window size over which haplotypes were estimated with Shapeit was set to 0.5 Mb, the value recommended by the authors of the software for sequencing data. Only biallelic SNP loci were used in Shapeit, while triallelic sites were also included in Beagle.

To reduce computational time during linkage decay analysis, phased VCF files produced by Beagle and Shapeit were thinned so that no 2 remaining loci were within 100 bases of one another. To perform linkage decay analysis, VCFTools was used to compare each SNP to each other SNP within a 50 kb window of it, with Pearson's coefficient of correlation r^2, and the normalised coefficient of LD (D') (Lewontin 1963) calculated for each pair. Subsequently, a custom python script took the average r^2 or D' value for every comparison where the distance between the compared SNPs was within a given 100 bp range (i.e. 1-100 bp, 101-200 bp, etc.), and plotted this against the smallest distance within that range (e.g. the average r^2 value for SNP pairs in the 101-200 bp distance range for a particular scaffold was plotted against the value 101).

The above analysis only considered linkage between SNPs within 50 kb of each other. To consider linkage across the whole genome it was necessary, for computational reasons, to further thin phased data so that no two SNPs were in 500 bp of each other. Following further thinning, linkage analysis was once again carried out in VCFTools, except that this time each SNP was compared to every other SNP in the entire genome, including SNPs on different scaffolds. This analysis was performed on both Beagle- and Shapeit-phased data. Subsequently, further custom Python scripts were used to visualise data as a heatmap.

Performance of Beagle- and Shapeit-phased data was found to be similar. Subsequent analysis therefore used Beagle-phased data only. Population genetic analysis indicated that *S. ratti* individuals could be assigned to a number of distinct genetic clades. Linkage decay plots and heatmaps for single genetic clades were produced as for whole-dataset plots, except that individuals not belonging to the genetic clade in question were removed from phased datasets prior to linkage analysis being carried out.

4.2.5.7 Analysis of mitochondrial data

The *S. ratti* mitochondrial reference assembly (Hunt *et al.* 2016) was downloaded from WormBase ParaSite. Sequencing reads from each of the 90 individual *S. ratti* and 10 isofemale lines were aligned

to this mitochondrial reference genome as described previously for nuclear data (Section 4.2.4.2). One individual from site AM was excluded from further analysis due to unexpectedly low mitochondrial read depth. Calling of genetic variants was also carried out as described previously (Section 4.2.5.1), except with settings for haploid sequences enabled in BCFTools. Variants were filtered such that only SNPs with an MQ score and a QUAL score of at least 20, and a cumulative read depth of at least 10,000 were retained. Basic diversity characteristics were determined as for nuclear data.

Analysis of molecular variance (AMOVA) was conducted in GenAlEx version 6.5 (Peakall and Smouse 2006, 2012), a free plug-in for Microsoft Excel. The mean number of mitochondrial differences in same-site pairwise comparisons was compared with that in different-site comparisons with a t-test. Haplotype maps based on 89 (DS90) or 99 (DS100) genomes were generated in PopART version 1.7 (Leigh and Bryant 2015) using the minimum spanning network method (Bandelt *et al.* 1999). Furthermore, for DS90, a maximum likelihood tree based on unique haplotypes were generated with RaxML version 8.1.15 (Stamatakis 2006). During maximum likelihood estimation, the general time reversible gamma model of substitution rate heterogeneity was used, and rapid bootstrapping with 100 replicates was applied. Also in DS90 only, the number of alleles shared out of the total number of mitochondrial SNPs detected was counted for each pair in pairwise comparisons to create a matrix. A Mantel test was used to compare this matrix with a matrix of nuclear Φ relatedness amongst samples in the R package ade4 version 1.7-13 (Dray and Dufour 2007).

4.3 Results

4.3.1 RFLP assessment of *Strongyloides ratti* diversity within hosts faecal pellets

Incomplete digestion of PCR products was observed frequently during restriction endonuclease reactions, as indicated by persistence of the full-length PCR product following digestion. Consequently, it was not possible to formally distinguish an RA genotype from an RR genotype that had been incompletely digested (Figure 4.1). Hence, apparently heterozygous individuals are labelled as RA*.

Where two iL3s from the same pellet differed at one or more of the three SNP markers assayed, the iL3s were said to have different multilocus genotypes and presumed to have been produced by different parasitic female mothers. Combined, the three loci assayed revealed extensive *S. ratti* genetic diversity within faecal pellets – every pellet from which more than 2 iL3s were sampled showed multiple multilocus genotypes (Table 4.4). This indicates that *S. ratti* infrapopulations are typically composed of multiple genetically distinct parasitic adults. Hence, it was concluded that sequencing multiple iL3s from the same pellet was unlikely to result in extensive resequencing of clonal siblings, and so would be informative both for measuring the genetic diversity within sampling sites as a whole, and for assessing the extent of genetic differentiation among infrapopulations within sampling sites.

That only 3 polymorphic SNP sites were identified precludes the use of these RFLP markers for an investigation into *S. ratti* population genetic structure more widely. Hence, further investigation into *S. ratti* population genetics was conducted by whole-genome sequencing.

Table 4.4: Results of restriction fragment length polymorphism analysis of individual Strongyloides ratti *using three loci. A pellet is a single rat faecal pellet collected from one of three sampling sites; "AM", "CA" or "LA" (Table 2.1). In genotypes, "R" indicates the reference allele and "A" indicates the alternate allele. Apparently heterozygous individuals (RA) are marked with an asterisk to indicate that they may have an RR genotype affected by incomplete digestion during the restriction endonuclease reaction (details in text section 4.3.1).*

Pellet	**# Larvae assayed**	**Sr1-3 genotype(s)**	**Sr1-4 genotype(s)**	**Sr2-4 genotype(s)**
AM1	2	RR (x1), AA(x1)	AA (x1), RA* (x1)	RR(x1), (AA) x1
AM2	2	AA (x2)	AA (x2)	RA* (x2)
AM3	2	RR (x1), RA* (x1)	RR (x1), RA* (x1)	RR (x1), RA* (x1)
AM4	18		RA* (x8), AA (x10)	RA* (x8), RR (x8)
CA1	1	RR (x1)		RR (x1)
CA2	3	RR (x2)	RR (x2), RA* (x1)	RA* (x3)
CA3	8		RA* (x6), AA (x2)	RR (x8)
LA1	3	RA* (x2), AA (x1)	RA* (x2)	RR (x2)
LA2	4	RA* (x1), AA (x1)	RA* (x3)	RR (x1), AA (x1), RA* (x2)
LA3	4	RR (x1), RA* (x1)	RR (x2), AA (2)	RR (x1), AA (x2)
LA4	20		RA* (x9) AA (x7)	RR (x20)

4.3.2 Whole-genome sequencing of individual *Strongyloides ratti*

4.3.2.1 Success of sequencing

Initially 225 *S. ratti* sequencing libraries underwent low coverage sequencing. Among these 225 libraries, the percentage of sequencing reads aligning to the *S. ratti* reference ranged from 0% to 67.3%, with a mean of 11.4%. The minimum alignment rate was 5.7%, and the mean was 21%.

Although no individual was sequenced at sufficient depth for population genetic analyses, initial sequencing allowed for libraries with a high percentage of *S. ratti* reads to be identified. From these data, 90 individuals were selected for deep sequencing (Table 4.1A). Following deep sequencing, the number of nucleotides in the *S. ratti* reference nuclear genome covered by at least one sequencing read (*i.e.* the genome coverage) ranged from 75.8% to 99.3% and averaged 96.4%, with 75 of the 90 samples having coverage greater than 95%. For each individual, the average number of reads covering each reference nucleotide (*i.e.* the mean read depth) ranged from 20 to 246 and averaged 68. Only 5 samples had mean read depths of less than 30. Hence, deep sequencing was judged to have been successful.

There was a strong correlation between the percentage of reads aligning to the reference genome and mean read depth (Pearson's $r^2 = 0.89$, t = 18.8, *df* = 88, P < 0.00001), as shown in Figure 4.2A. There was a weaker correlation between the percentage of reads aligning and genome coverage ($r^2 = 0.47$, t = 5.03, *df* = 88 P < 0.00001), however, this is due to coverage approaching 100% when approximately 18% of reads align (Figure 4.2B).

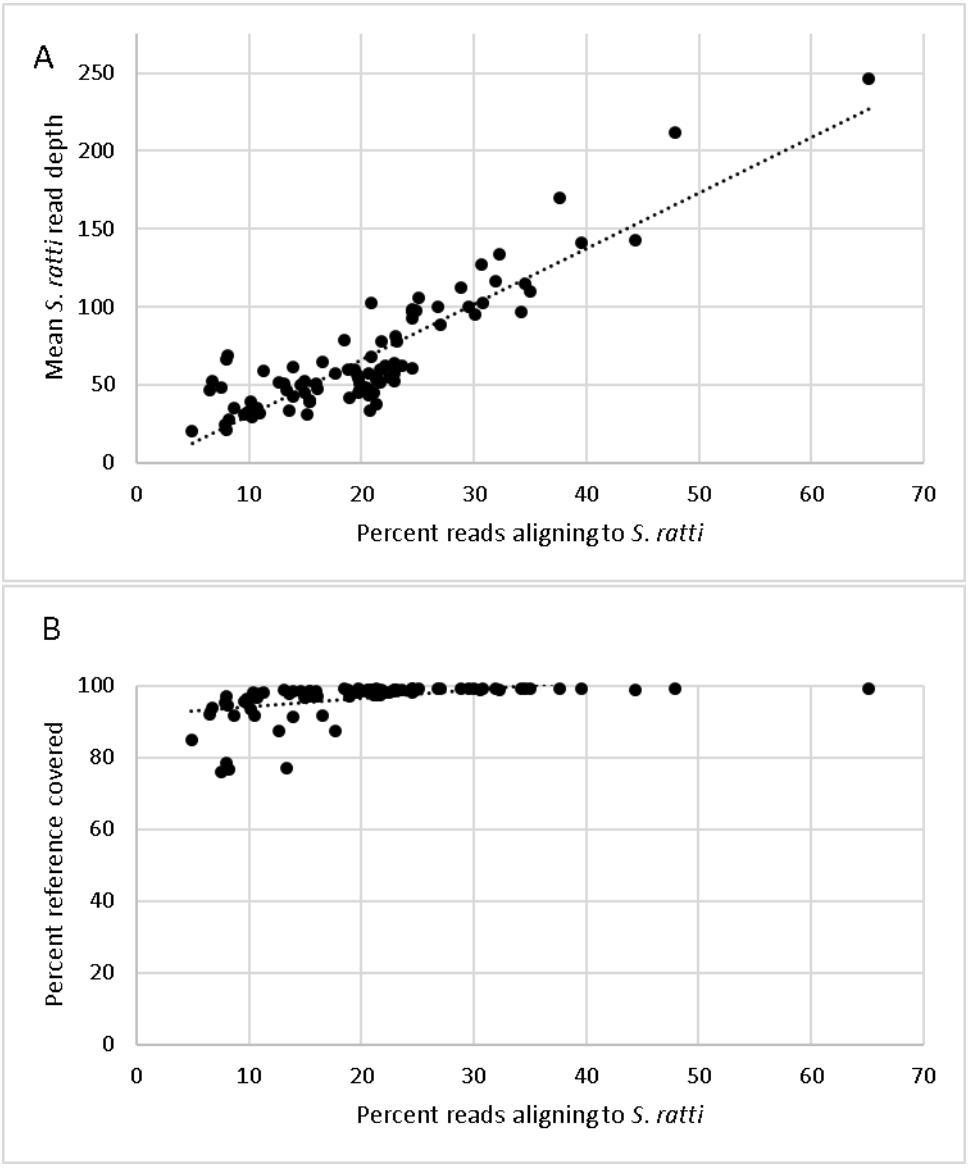
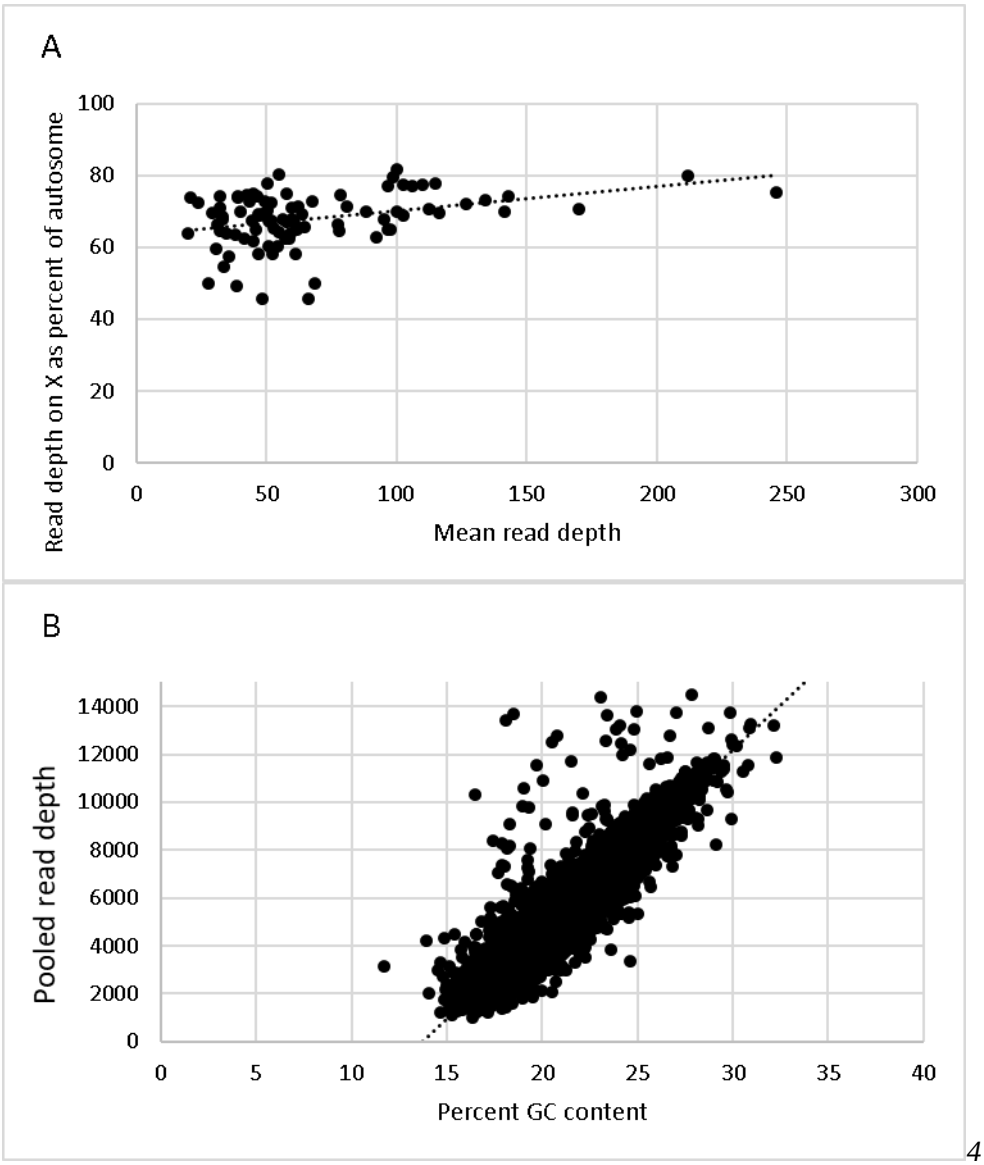


Figure 4.2: Correlation of the percent reads aligning to the Strongyloides ratti *reference genome and A) mean S. ratti read depth per reference genome position and B) percent reference genome positions covered in the DS90 dataset.*

4.3.2.2 Relationship between GC content and coverage

In every *S. ratti* individual sequenced, the mean read depth on the X chromosome was lower than on the autosomes. Specifically, among individuals, mean read depth on the X chromosome ranged from 45.5% to 81.7% of the mean read depth on the autosome, with an average of 67.9% (standard deviation = 7.4%). There was positive correlation between genome-wide read depth and the ratio of X chromosome to autosome read depth, such that individuals that had higher mean read depth overall having X chromosome read depths more similar to their autosome depths than did individuals with lower mean read depth (Figure 4.3A, $r^2 = 0.36$, $t = 3.61$, $df = 88$, $P < 0.001$). When all reads from DS90 were pooled, there was a strong correlation between GC content of a 10 kb genome window and mean read depth within that window (Figure 4.3, $t = 78.3$, $df = 4{,}008$, $P < 0.00001$). The GC content of the X chromosome is lower than that of the autosomes, (19.7% *vs.* 22%, Hunt *et al.* 2016). Therefore, the difference in coverage of the X chromosome compared with the autosomes is related to the lower GC content of the X, and this effect is exacerbated in samples where coverage overall is low.



4

Figure 4.3: A) Correlation between read depth on the X chromosome as a percent of autosome read depth and mean depth of reads in the entire genome in each Strongyloides ratti *individual in the DS90 data set. B) Correlation between mean read depth per genome position when all individuals are summed and percent GC content in each 10 kb window of the autosomes and X chromosome scaffolds.*

4.3.3 Nuclear polymorphism in *Strongyloides ratti*

4.3.3.1 SNPs detected

Dataset (DS) 90 consists of sequencing information for 90 individual worms from 3 sampling sites (Table 4.1A). After variant calling and filtering, 170,666 SNPs were retained, amounting to an average of 4.1 SNPs per kb. Of these, 614 were tri-allelic, and the rest were bi-allelic. The ratio of transitions to transversions was 1.77. DS100 consists of the same 90, DS90 individual worms, plus sequencing data

for 10 isofemale lines. In DS100, 235,393 SNPs were retained after filtering, of which 928 were tri-allelic and the rest were bi-allelic. The ratio of transitions to transversions was 1.8.

In pairwise comparisons of all individuals within DS90, the number of genetic differences (SNP loci at which the individuals had different genotypes) ranged from 240 to 101,088 with a mean of 57, 258, amounting to an average 1.3 per Kb. The distribution of pairwise relatedness values was strongly non-normal (Figure 4.4), indicating genetic clustering in the study population.

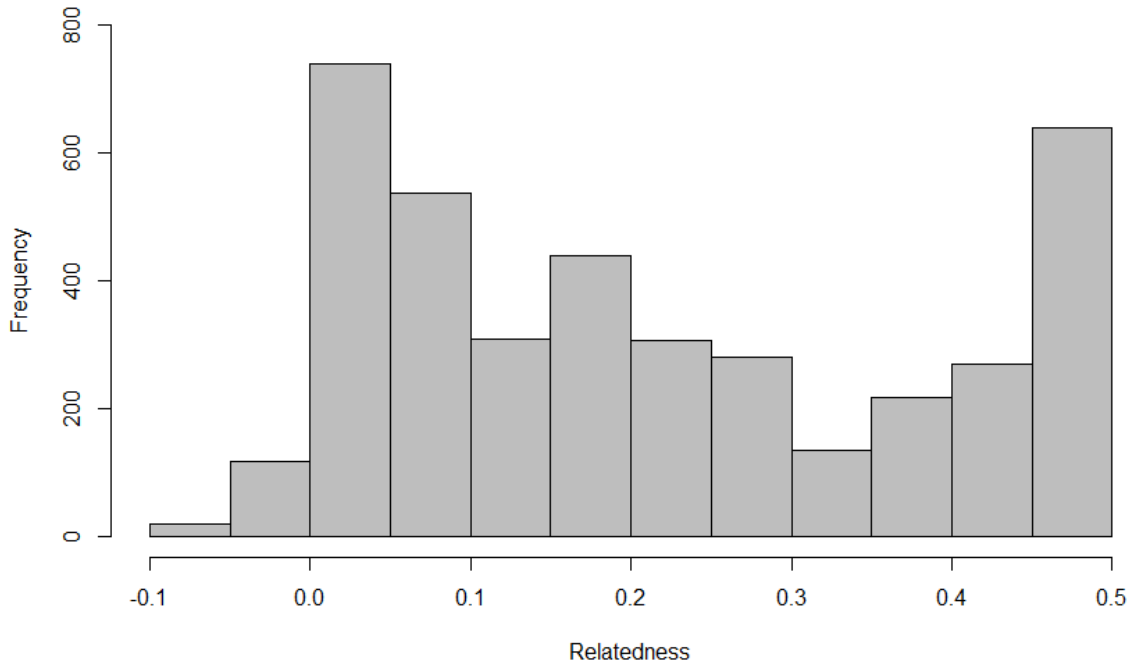


Figure 4.4: Histogram of Φ relatedness values in pairwise comparisons of individual S. ratti *in DS90.*

4.3.3.2 Hardy-Weinberg equilibrium

Mitotic parthenogenesis, which is an obligate part of the *S. ratti* life cycle, is expected to lead to heterozygote excess due to the accumulation of novel mutations on each of a pair of homologous chromosomes. This heterozygote excess causes deviations from HWE. Here, HWE was only considered in DS90. Taking all SNP positions together and treating all individuals as belonging to a single population, individuals appeared to be at HWE. However, χ^2 for individual SNPswas highly variable among SNP positions (mean χ^2 = 13.65, standard deviation = 19.35, P = 0.48). When each of the three sampling sites were treated as separate populations, each one was at HWE, but, as when sampling sites were pooled, χ2 was highly variable among individual SNP positions.

In every sampling site, and when all sampling sites were considered together, the vast majority of SNP positions that were not in HWE showed heterozygote excess. Based on the allele frequencies at each SNP locus in the sample as a whole, the HWE expectation would be that approximately 72% of SNP loci in each individual would be homozygous, whereas the observed number of homozygous SNP positions was lower than this in all but one individual. Specifically, observed homozygosity per individual ranged from 41.9% to 72.4% of the total SNP loci, with a mean of 63.9%. Hence individuals appeared to be more heterozygous than would be expected under HWE.

4.3.3.3 Differentiation among *Strongyloides ratt*i

In pairwise relatedness tests, mean Φ relatedness was 0.22 in same-host pairwise comparisons and 0.214 in different-host pairwise comparisons, and the difference in Φ between these two groups was non-significant (t = -0.32, *df* = 108.81, P = 0.75). In contrast, mean Φ relatedness was 0.225 in same-site pairwise comparisons, and 0.206 in different-site comparisons, Difference in Φ between these two groups was significant (t = -3.68, *df* = 3957.9, P <0.001). F_{ST} among all three sampling sites was 0.02, which is very low. In pairwise comparisons, F_{ST} was 0.03 between sites CA and LA, 0.03 between CA and AM, and 0 between sites AM and LA. Thus, while there was no evidence for population genetic structuring among hosts, there was evidence for weak structuring among sampling sites.

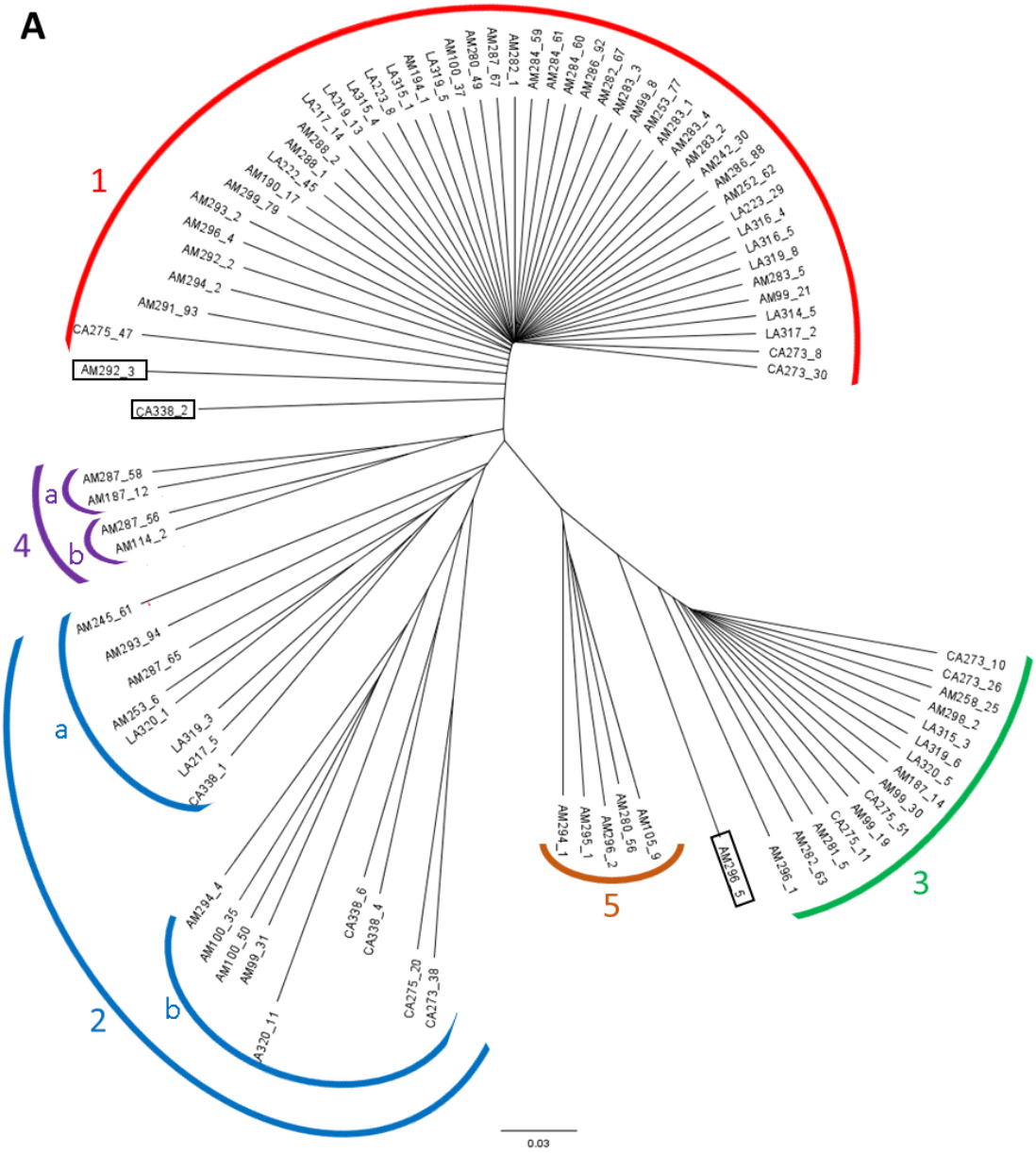
AM and LA are geographically closer to each other (9.7 km) than either is from CA (30-32 km) and are on the same side of the Severn estuary, so *a priori* gene flow between sites AM and LA may be expected to be greater than between those sites and CA. Φ relatedness in pairwise comparisons where individuals were from site AM and / or LA was significantly higher than when individuals from CA were compared with non-CA individuals (t = -12.84, *df* = P < 0.00001, means 0.236 and 0.159 respectively). These results suggest that the pattern of genetic differentiation among worms at these three sites was driven by worms at the most geographically distant site being genetically distinct from worms sampled at the other two sampling sites.

Site AM was the only site from which worms were taken at all four sampling seasons, and so temporal population genetic structure was only tested within this site. F_{ST} among seasons at site AM was 0.02. There was a small but significant difference between Φ relatedness values in same-season pairwise comparisons (mean = 0.255) than in different-season comparisons (mean = 0.203) (t = -6.74, *df* = 1585, P < 0.00001). This suggests that the genetic composition of *S. ratti* populations changes over time.

4.3.3.4 Neighbour joining dendrograms

Neighbour joining dendrograms based on SNPs in DS90 and DS100 are shown in Figure 4.5. By eye, the neighbour joining tree for DS90 indicates 3 major genetic clades - clades 1, 2 and 3 - which together

account for 78 of the 90 individual worms. Clade 2 is divided into sub-clades 2a and 2b. The remaining 12 individuals can be grouped into minor clade 4 (split into sub-clades 4a and 4b) and minor clade 5, with 3 individuals in no clade (Figure 4.5A, Table 4.5). All individuals in clade 4 are from site AM and represent 3 different hosts. All individuals in clade 5 came from site AM, and each came from a different host. Worms representing clades 1, 2 and 3 were found in all 3 sampling sites in approximately equal ratios (Table 4.5), with no statistically significant relationship between sampling site and genetic clade frequency according to a Fisher's exact test ($P = 0.14$). Similarly, a Fisher's exact test indicated no significant interaction between season and genetic clade frequency ($P = 0.25$). Finally, individual hosts often contained worms from multiple genetic clades; 11 of 39 hosts provided parasites from 2 clades, while 3 hosts provided parasites from 3 clades.


A
1
AM287_67
AM282_1
AM284_59
AM284_61
AM284_60
AM286_92
AM282_67
AM293_3
AM99_8
AM253_77
AM283_1
AM283_4
AM283_2
AM242_30
AM286_88
AM252_62
LA223_29
LA316_4
LA316_5
LA319_8
AM283_5
AM99_21
LA314_5
LA317_2
CA273_8
CA273_30
AM280_49
AM100_37
LA319_5
AM194_1
LA315_1
LA223_9
LA315_4
LA319_13
LA317_14
AM288_2
AM288_1
LA222_45
AM190_17
AM299_79
AM293_2
AM296_4
AM292_2
AM294_2
AM291_93
CA275_47
AM292_3
CA338_2
4
a
AM287_58
AM187_12
b
AM287_56
AM114_2
a
AM245_61
AM293_94
AM287_65
AM253_6
LA320_1
LA319_3
LA217_5
CA338_1
2
b
AM294_4
AM100_35
AM100_50
AM99_31
A320_11
CA338_6
CA338_4
CA275_20
CA273_38
5
AM294_1
AM295_1
AM280_56
AM296_2
AM105_9
AM296_5
3
CA273_10
CA273_26
AM258_25
AM298_2
LA315_3
LA319_6
LA320_5
AM187_14
AM99_30
CA275_51
AM99_19
CA275_11
AM281_5
AM282_63
AM296_1
0.03

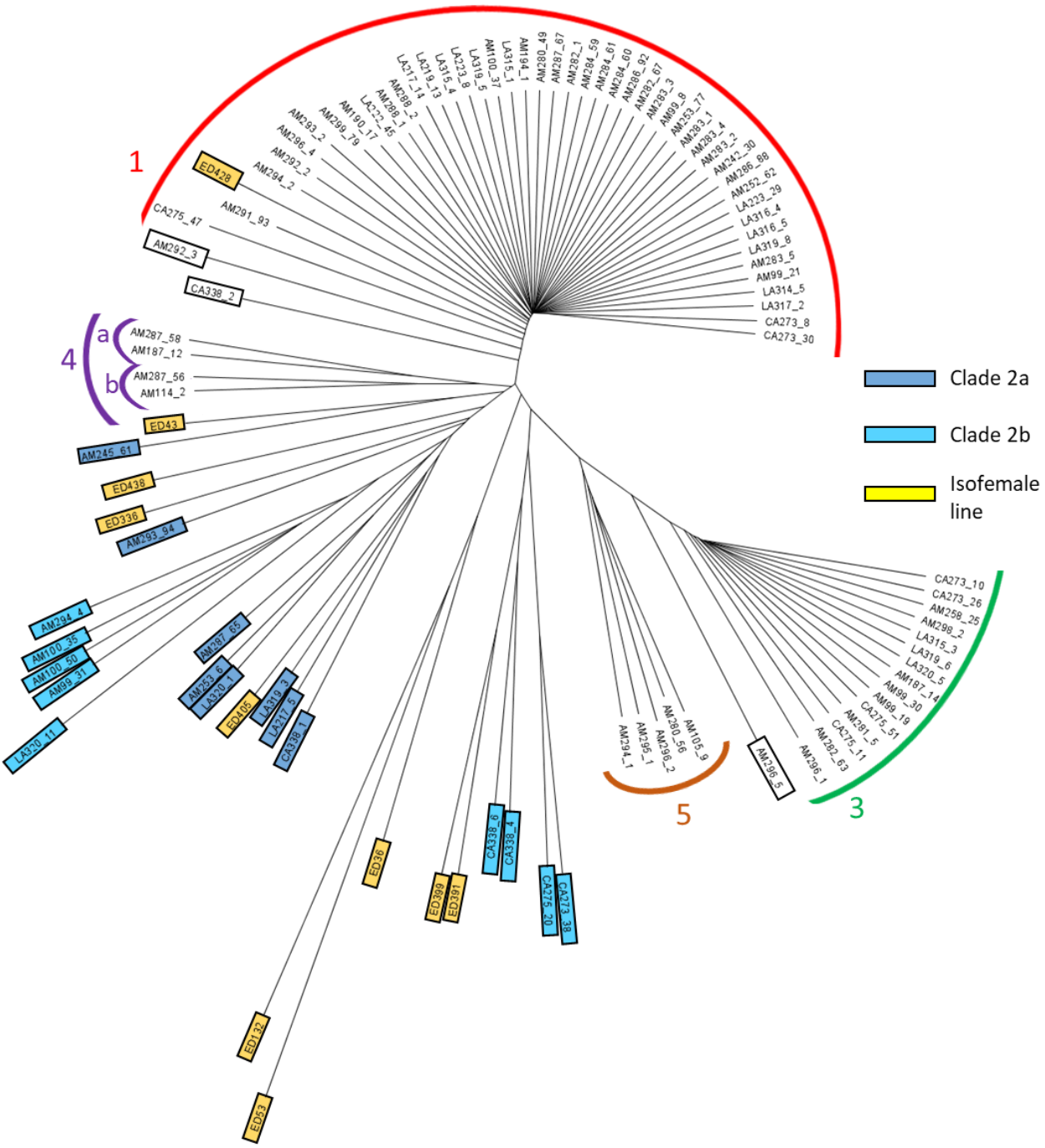


Figure 4.5: Neighbour-joining dendrograms based on entire nuclear sequences of DS90 (A) or DS100 (B). Branch length is in units of proportion of SNP loci polymorphic. In DS90, individuals are arbitrarily divided into 5 genetic clades (indicated by coloured brackets), with three individuals (indicated by black boxes) not included in a clade. In the name of each individual, the two-letter code indicates sampling site, the number before the hyphen indicates the host pellet, and the number after the hyphen differentiates individuals from the same host. In DS100, individuals in clade 2 and isofemale lines are coloured as shown in the key. Notation is otherwise as in DS90.

Table 4.5: Expected (E) and observed (O) number of Strongyloides ratti *infective larvae in each of 5 nuclear genetic clades in each sampling site (top) or season (bottom). Three further individuals did not belong to a genetic clade. Expected values are based on neutral expectation of clade distribution among sites / seasons. Only individuals from site AM were considered for the season-based analyses.*

Sampling site	**CA**		**LA**		**AM**		**Total**
	E	**O**	**E**	**O**	**E**	**O**	
Clade 1	6.3	3	10.6	13	29	30	**46**
Clade 2	2.3	5	3.9	4	10.1	8	**17**
Clade 3	2.1	4	3.4	3	9.5	8	**15**
Clade 4	0.6	0	0.9	0	2.5	4	**4**
Clade 5	0.7	0	1.1	0	3.1	5	**5**
Total	**12**		**20**		**55**		**87**

Sampling season	**Spring**		**Summer**		**Autumn**		**Winter**		**Total**
	E	**O**	**E**	**O**	**E**	**O**	**E**	**O**	
Clade 1	5.6	4	2.8	2	3.4	3	19.2	22	**31**
Clade 2	1.5	3	0.7	0	0.9	2	4.9	3	**8**
Clade 3	1.5	2	0.7	1	0.9	1	4.9	4	**8**
Clade 4	0.7	0	0.4	2	0.4	0	2.5	2	**4**
Clade 5	0.9	1	0.5	0	0.5	0	3.1	4	**5**
Total	**10**		**5**		**6**		**34**		**55**

F_{ST} and Φ relatedness values were calculated within and among the 3 major clades, clades 1, 2 and 3, and are shown in Table 4.6. F_{ST} among these major clades was 0.3. These data suggest that *S. ratti* population genetics is better described as being structured into sympatric, genetically distinct clades that are present at different sampling sites at different frequencies, rather than being structured according to geography *per se*.

Table 4.6: Pairwise F_{ST} relatedness (above diagonal, red) and Φ relatedness (diagonal and below, yellow) among three major Strongyloides ratti *nuclear genetic clades.*

	Clade 1	**Clade 2**	**Clade 3**
Clade 1	0.43	0.22	0.35
Clade 2	0.18	0.23	0.23
Clade 3	0.06	0.05	0.45

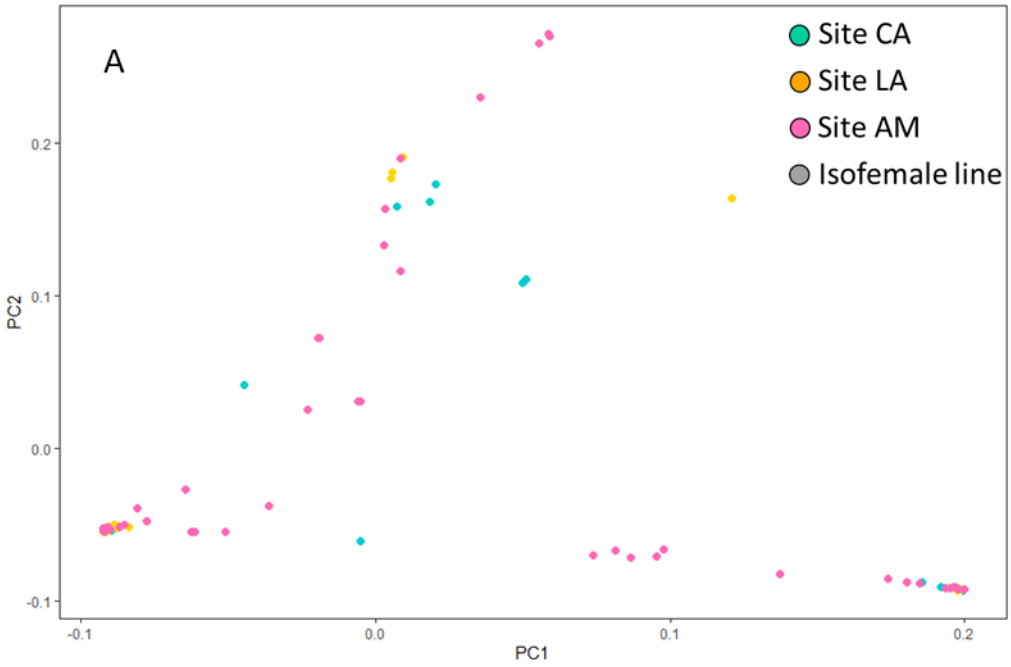
The neighbour joining tree for DS100 shows clades 1, 3 and 5 unchanged from DS90, while individuals in clade 2 are rearranged such that the individuals of sub-clades 2a and 2b are intermingled (Figure

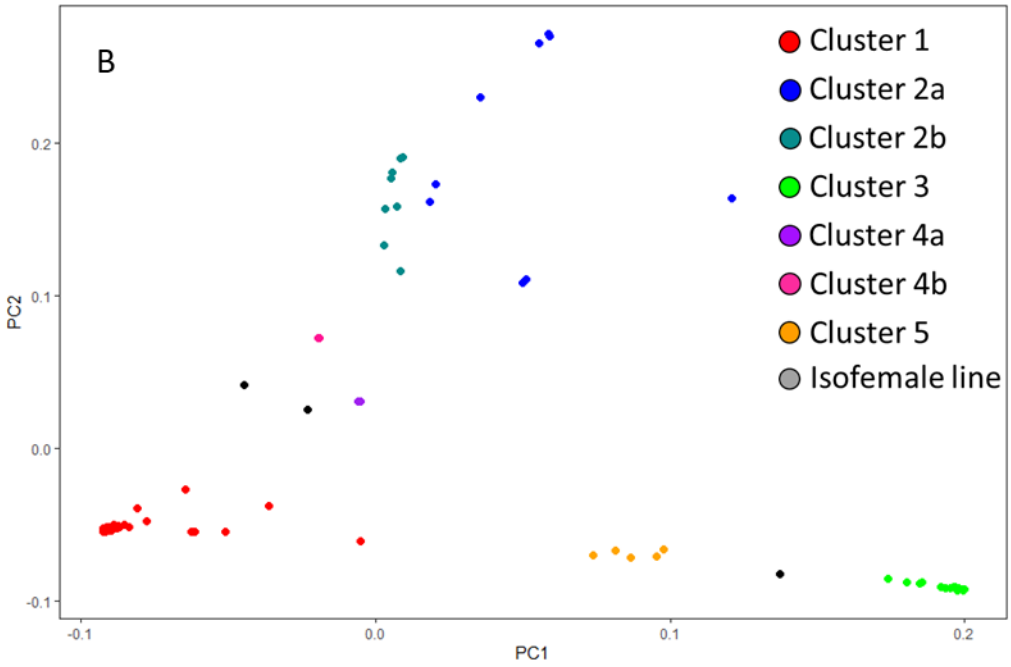
4.5B). One isofemale line groups with clade 1, while the other 9 isofemale lines group with clade 2 (Figure 4.5).

4.3.3.5 PCA of genetic diversity

PCA was carried out on SNPs ofDS90 and DS100. In DS90, singular values (square roots of genetic variance explained) of PCs 1 and 2 collectively were 96%, while in DS100, singular values PCs 1 and 2 collectively reached 91%. Projections of samples onto PCs 1 and 2 are shown in Figure 4.6. In DS90, PC1 separated three clades, corresponding to clades 1, 5 and 3 in the neighbour-joining dendrogram (Figure 4.5). Individuals in clades 2 and 4 are generally intermediate between clades 1 and 5 on to PC1 but separated from them by PC2. Clade 4 of the dendrogram was clearly separated into 4a and 4b in the PCA projection. Clade 2 of the dendrogram appeared as a clade on the PCA projection, and while there was not perfect separation between the dendrogram's sub-clades 2a and 2b, clade 2b individuals tended to have higher PC2 values. The PCA projection for DS100 was essentially identical to that for DS90 (Figure 4.6), with isofemale lines appearing in the projection where they would be expected to given their placements in the dendrogram. Specifically, 1 isofemale line groups with clade 1, while the other 9 isofemale lines group with clade 2 (Figure 4.6).

It is hypothesised that clades 1, 2a and 3, or a portion of individuals therein, represent ancient lineages each derived asexually from single parasitic females with no input from other genotypes. Other individuals in this study are suggested to be the result of sexual reproduction between these "parental" genetic clades.


A
Site CA
Site LA
Site AM
Isofemale line
PC2
PC1
0.2
0.1
0.0
-0.1
-0.1
0.0
0.1
0.2


B
Cluster 1
Cluster 2a
Cluster 2b
Cluster 3
Cluster 4a
Cluster 4b
Cluster 5
Isofemale line
PC2
PC1
0.2
0.1
0.0
-0.1
-0.1
0.0
0.1
0.2

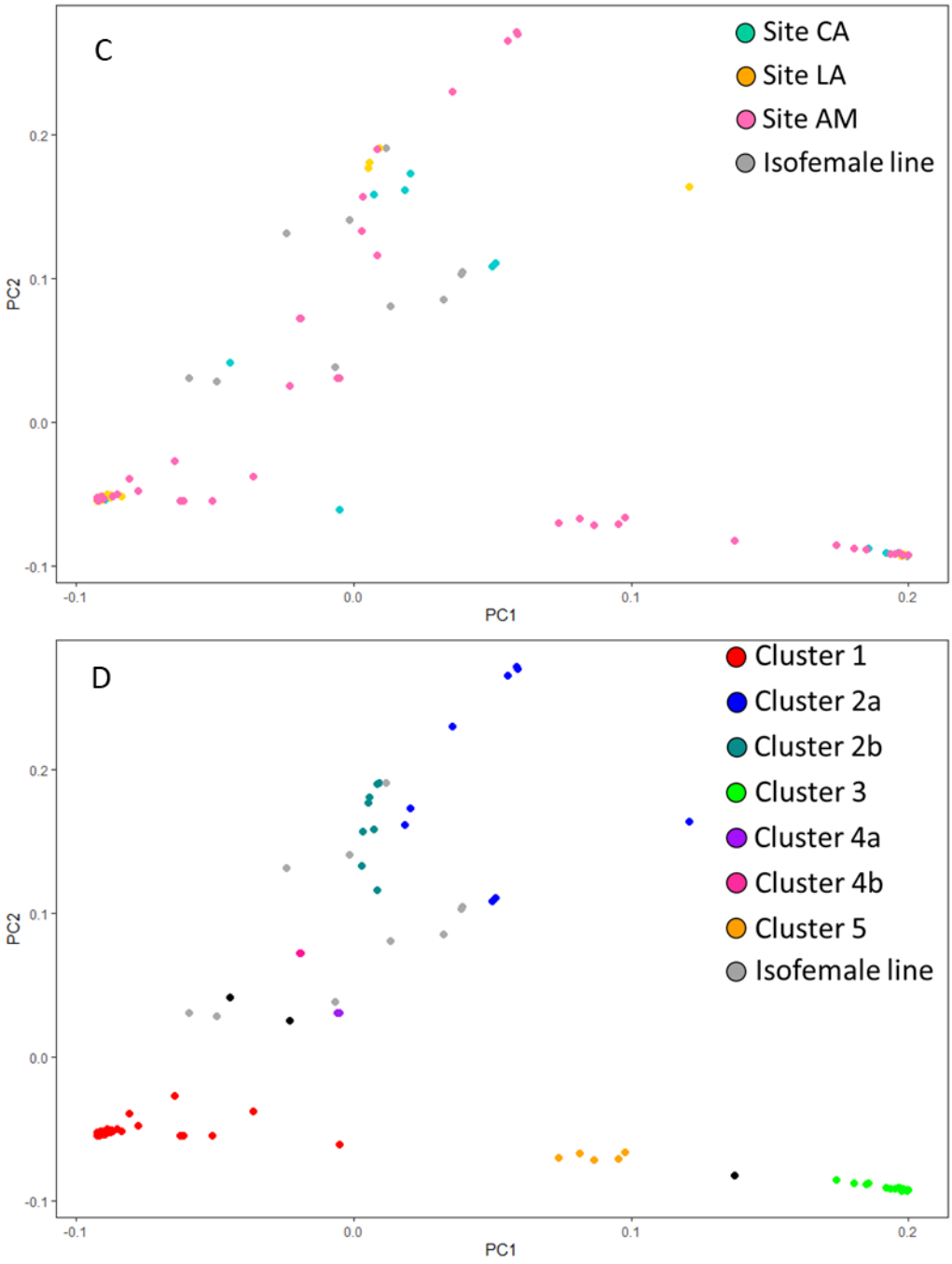


Figure 4.6: Projections of principal component (PC)s 1 and 2 of individuals in DS90 (A and C) and DS100 (B and D). Individuals are coloured according to sampling site in A and B, and nuclear genetic clade apparent in neighbour-joining dendrograms (Figure 4.5) in C and D.

4.3.3.6 Linkage disequilibrium

LD was only considered in DS90, and only from the autosomes and 2 largest scaffolds of the X chromosome. In all 4 scaffolds and by both phasing methods (Beagle and Shapeit), values of r^2 initially declined sharply and began to asymptote at distances of approximately 5 kb (Figure 4.7). In the autosomes, average r^2 then continued to decline gradually and had not fully reached its asymptote by 50

kb (the range over which linkage decay was examined). In contrast, r^2 decline had largely ceased in the X chromosome scaffolds by 20 kb (Table 4.7, Figure 4.7). Trends in mean r^2 values on each of the four scaffolds were the same in both Beagle- and Shapeit-phased data, but the exact mean r^2 values reported were higher when phasing was done by Shapeit than by Beagle (Table 4.7). Heatmaps of r^2 across the genome revealed low levels of linkage both among and within chromosomes, with no evidence for extended linkage blocks (Figure 4.8, insets).

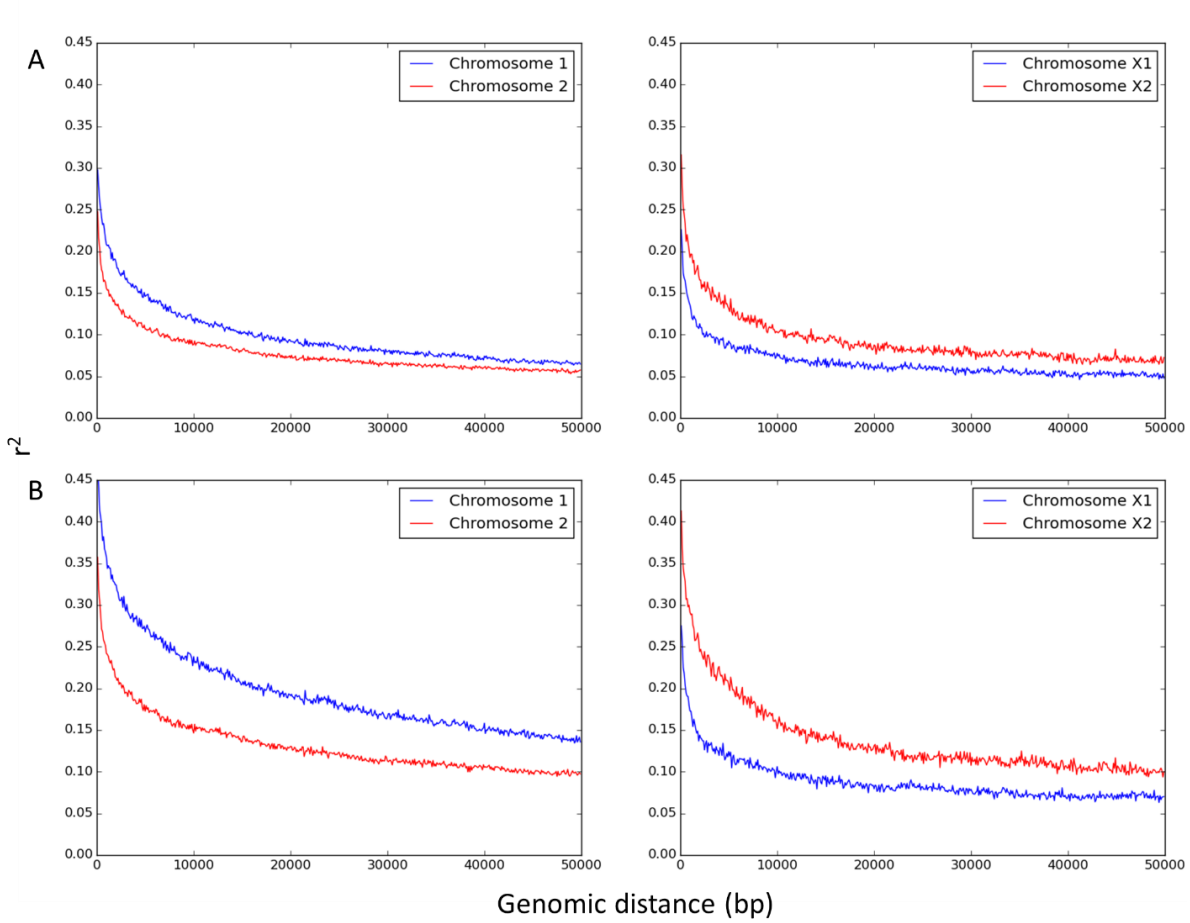
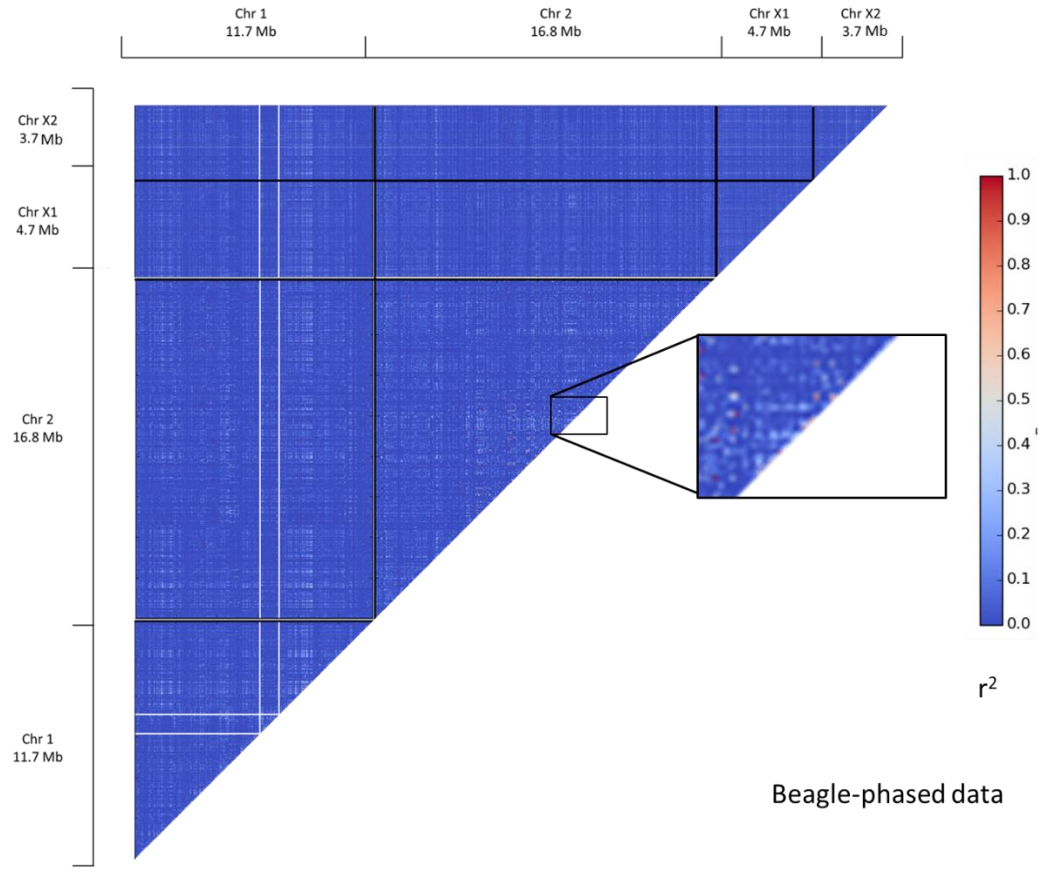


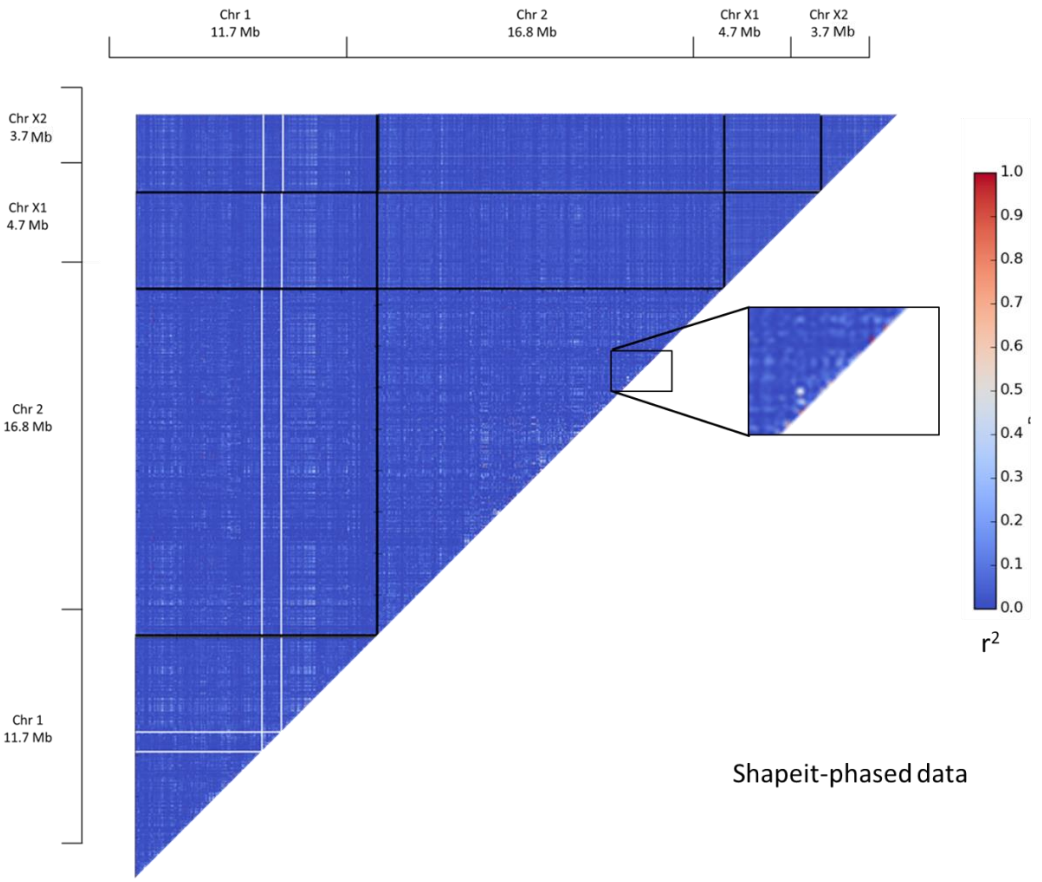
Figure 4.7: Linkage disequilibrium among Strongyloides ratti *individuals, as shown by r^2 values. Phasing of genotypes was carried out by Beagle (A) or Shapeit (B). X1 and X2 refer to the largest and second largest scaffolds of the X chromosome, respectively.*

Table 4.7 Results of linkage disequilibrium study based on the results of two phasing programmes; Beagle and Shapeit, averaged across the entirety of each scaffold. Scaffolds Chr X-1 and Chr X-2 refer to the largest and second largest X chromosome scaffolds respectively.

Scaffold	**Beagle r^2**	**Beagle D'**	**Shapeit r^2**	**Shapeit D'**
Chr 1	0.1	0	0.2	0.08
Chr 2	0.08	0	0.13	0
Chr X-1	0.07	0	0.09	0
Chr X-2	0.1	0	0.14	0.01



Beagle-phased data



Shapeit-phased data

Figure 4.8: Heatmaps of linkage among Strongyloides ratti *individuals, as shown by r^2 values. Phasing of genotypes was carried out by Beagle (top) or Shapeit (bottom). Insets show an expanded region of chromosome 2, demonstrating a lack of high-linkage blocks even at close proximity. X1 and X2 refer to the largest and second largest scaffolds of the X chromosome, respectively. Vertical and horizontal white lines in chromosome 1 represent two Mb long tracts of 'N's that separate the three parts of this chromosome (details in text section 4.1.3)*

D' values showed similar trends to r^2, but with some exceptions. D' reached its asymptote in Beagle-phased autosomes at distances of approximately 15 kb, rather than continue declining as r^2 did (Figure 4.9). Furthermore, on Shapeit-phased X chromosome scaffolds D' continued to decline up to distances of approximately 25 kb (Figure 4.9), and the difference in D' between the 2 X chromosome scaffolds was lower than what would be expected from r^2 values (Table 4.7).

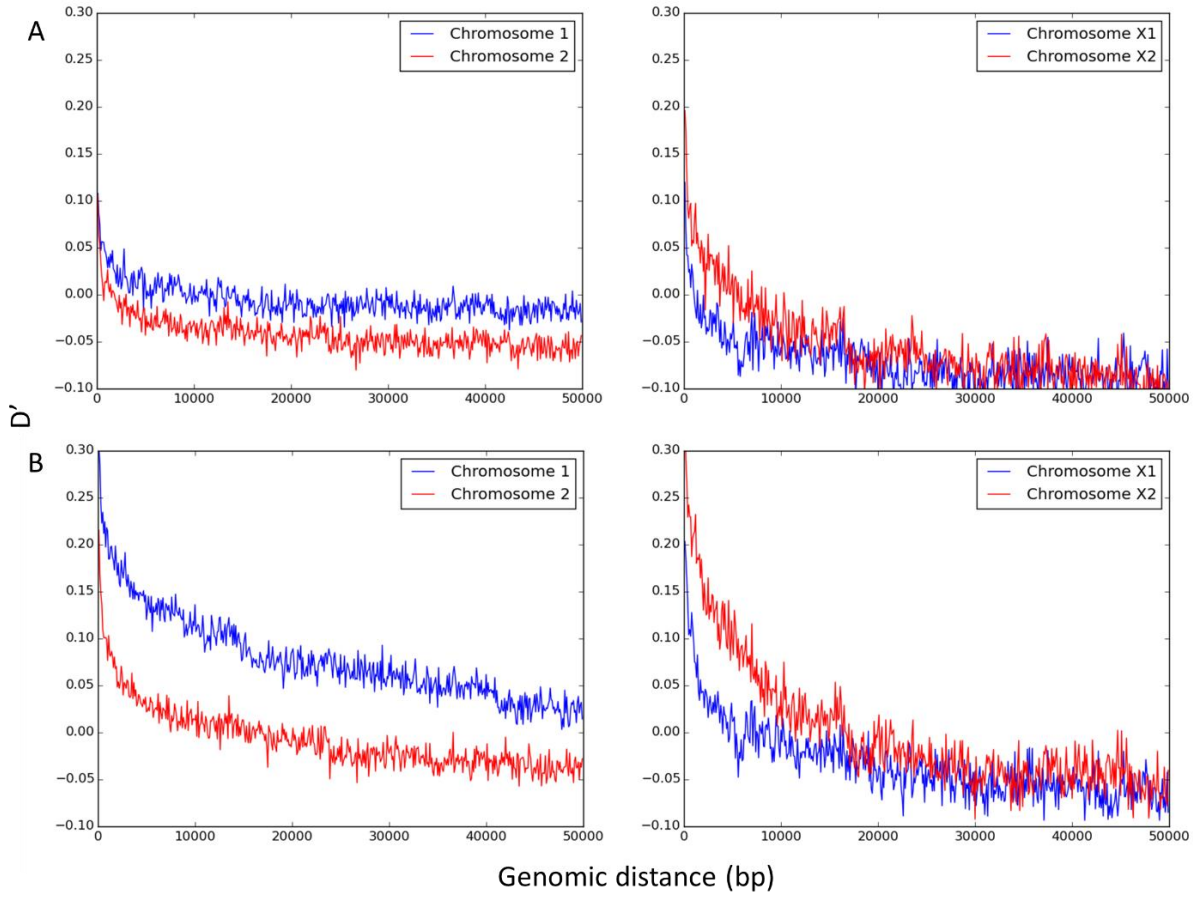


Figure 4.9: Linkage disequilibrium among Strongyloides ratti *individuals, as shown by D' values. Phasing of genotypes was carried out by Beagle (A) or Shapeit (B). X1 and X2 refer to the largest and second largest scaffolds of the X chromosome, respectively.*

If the "parental" clades 1, 2a and 3 have been maintained exclusively through mitotic parthenogenesis, pan-genome linkage might be expected. There were insufficient individuals in clade 2a to test within-clade linkage, but linkage was assessed in nuclear clades 1 and 3 (Section 4.3.34, Figure 4.5). As the results of Beagle- and Shapeit-phased data were similar, only Beagle phased data was used for within-clade linkage assessment. Linkage decay graphs (Figure 4.10) showed that linkage was on average higher and decayed more slowly in both clade 1 and clade 3 compared with the full DS90 dataset (compare Figures 4.7 and 4.9). Comparing clades 1 and 3, clade 3 had higher initial linkage, and it decayed more gradually, compared with clade 1. Together these results suggest that there is greater linkage within genetic clades than among them. In clade 1, r^2 declined to approximately 0.25 in chromosome 1 and 0.18 in chromosome 2 at a distance of 10 kb, and to approximately 0.1 in both X chromosomes at a distance of 40 kb. Across all scaffolds examined, the decline in r^2 with distance linkage decay was much more gradual in clade 3 than clade 1, still not having approached an asymptote at a distance of 50 kb.

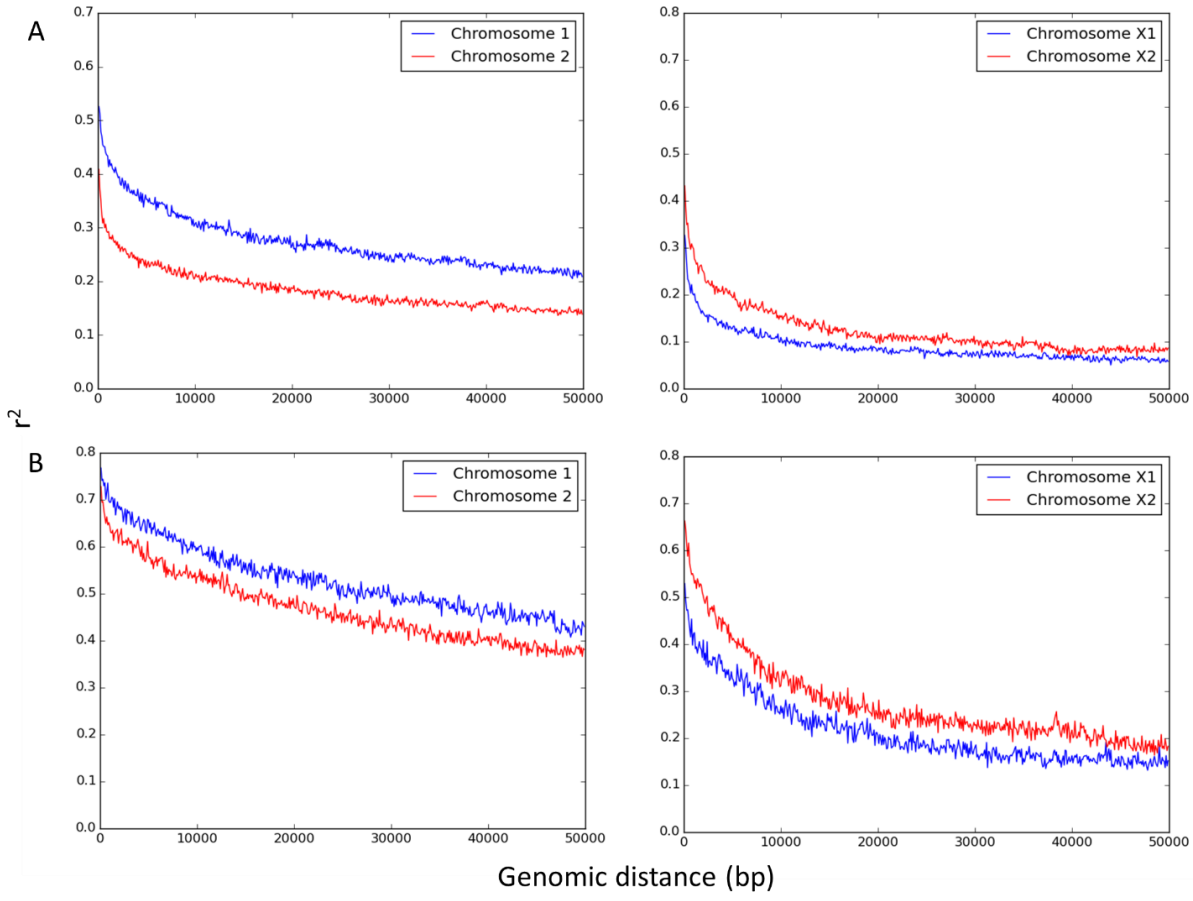
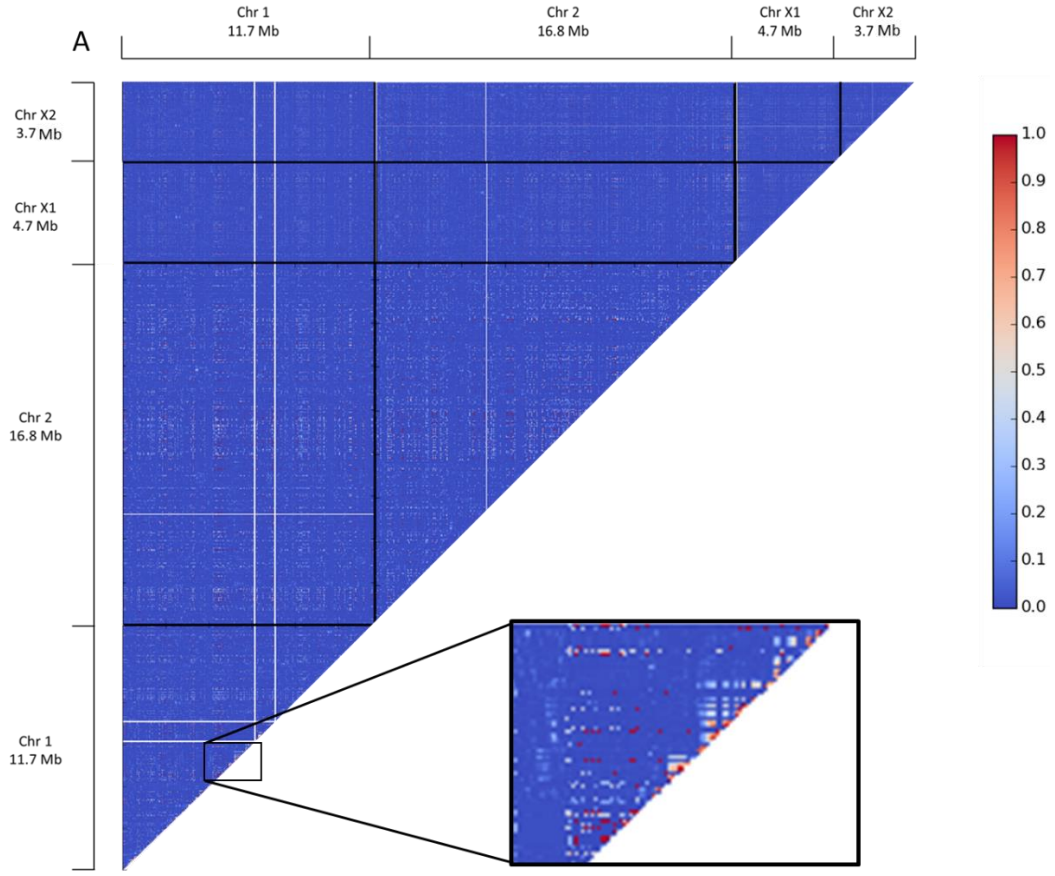


Figure 4.10: Linkage disequilibrium among Strongyloides ratti *in nuclear clade 1 (A) or clade 3 (B) (Figure 4.5), as shown by r^2 values.*

At large scale, the whole-genome linkage heatmap for clade 1 is not substantially different from that of DS90 (Figure 4.8), except that background r^2 levels were slightly higher for clade 1. However, in the clade 1 linkage heatmap, linkage blocks spanning tens of kilobases in Chromosome 1 are apparent (see Figure 4.11A insert for an example), that are not see in the DS90 analysis. The linkage heatmap for clade 3 shows higher linkage across the genome than the clade 1 heatmap (Figure 4.11B), and linkage blocks extended over greater genetic distances than in clade 1 (for example, Figure 4.11B, inset). Nevertheless, pan-genome linkage is not observed, suggesting there may be a low level of sexual reproduction within clades.



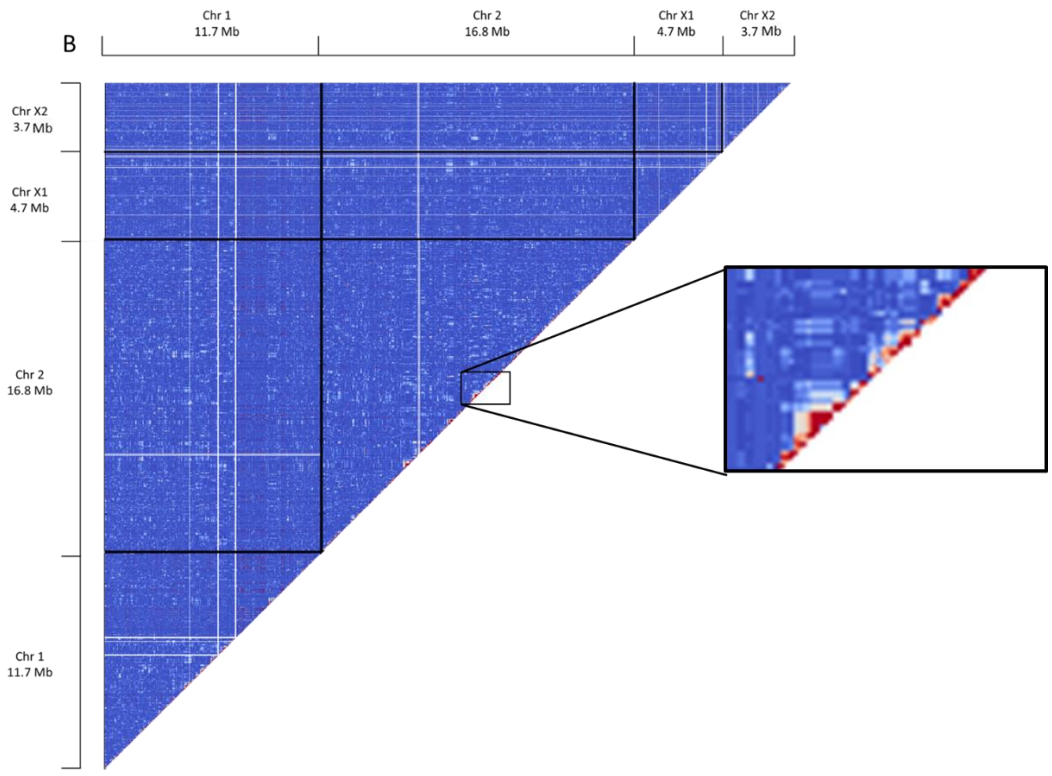


Figure 4.11: Heatmaps of linkage disequilibrium among Strongyloides ratti *in nuclear clade 1 (A) or clade 3 (B) (Figure 4.5), as shown by* r^2 *values. X1 and X2 refer to the largest and second largest scaffolds of the X chromosome, respectively.*

4.3.4 Mitochondrial polymorphism in *Strongyloides ratti*

4.3.4.1 Variants detected

For DS90, 156 mitochondrial SNPs were retained after filtering, amounting to 9.3 per kb. Of these, 1 was tri-allelic, and the rest were bi-allelic. The ratio of transitions to transversions was 3.62. The total number of mitochondrial haplotypes detected in DS90 was 58. The number of individuals per haplotype ranged from 1 to 12 with a mean of 1.5, and 47 individuals were the sole representatives of their haplotype. Haplotypes with multiple representatives are shown in Table 4.8.

For DS100, 178 mitochondrial SNPs were retained after filtering, amounting to 10.7 per kb. Of these, 3 were tri-allelic and the rest were bi-allelic. The ratio of transitions to transversions was 3.11. Three isofemale lines shared a single mitochondrial haplotype also present in 3 individuals from site AM (Table 4.8). Each other isofemale line had a unique mitochondrial haplotype.

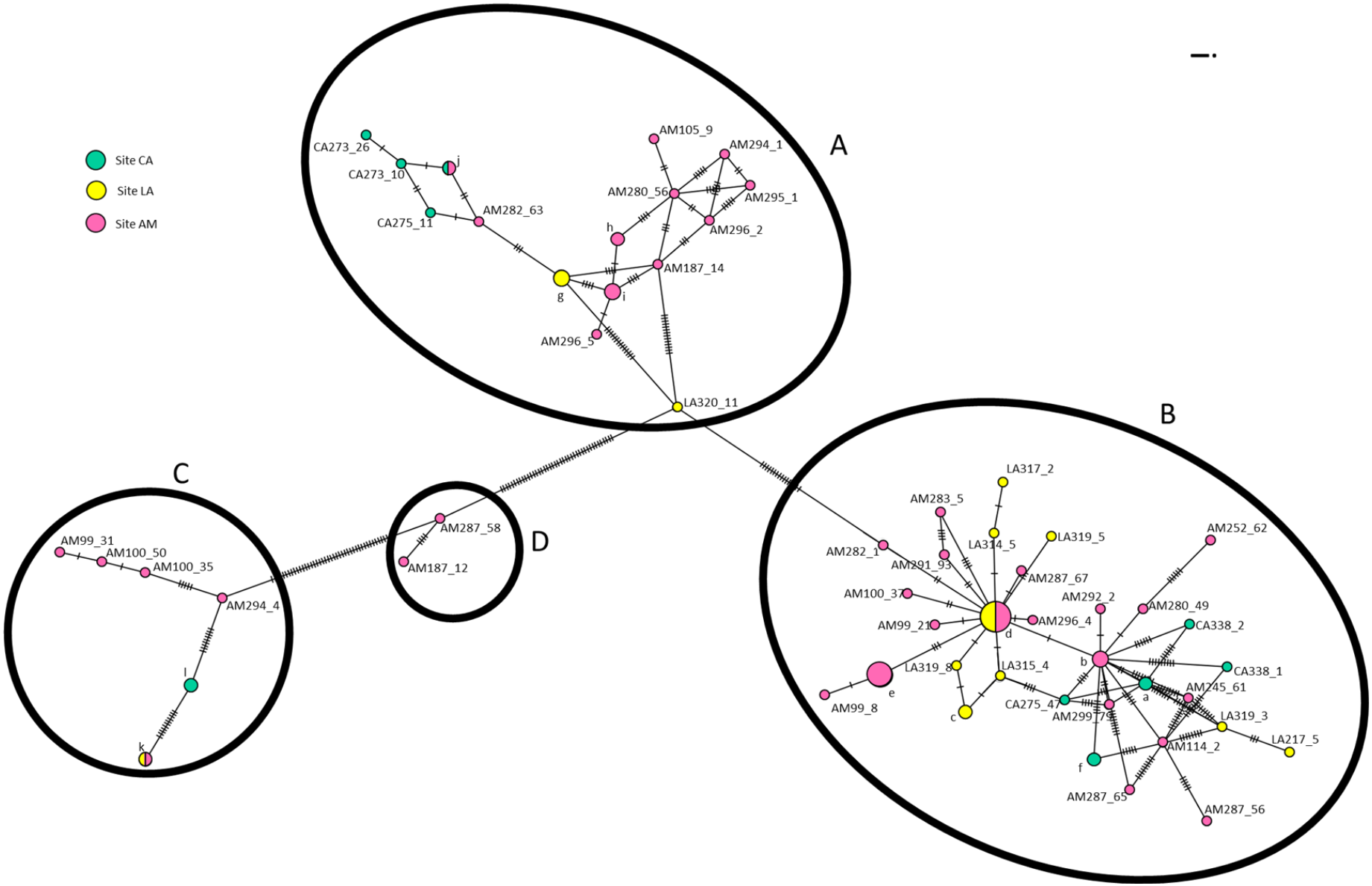
Table 4.8: Mitochondrial haplotypes within DS90 represented by more than 1 individual, including isofemale lines that had haplotype b. Number in parentheses indicate number of individuals representing each haplotype. In the name of each individual, the two-letter code indicates sampling site, the number before the hyphen indicates the host pellet, and the number after the hyphen differentiates individuals from the same host.

Haplotype	**Individuals**
a	CA273_8, CA273_30 (2)
b	AM294_2, AM293_2, AM292_3, ED53, ED36, ED132 (6)
c	LA316_5, LA316_4 (2)
d	AM286_92, M284_61, AM284_60, AM284_59, AM283_3, AM190_17, LA315_1, LA223_8, LA223_29, LA222_45, LA219_13, LA217_14 (12)
e	AM288_2, AM288_1, AM86_88, AM283_4, AM283_2, AM283_1, AM253_77, AM242_30 (8)
f	CA275_20, CA273_38 (2)
g	LA320_5, LA319_6, LA315_3 (3)
h	AM99_30, AM99_19 (2)
i	AM298_2, AM296_1, AM258_25 (3)
j	AM281_5, CA275_51 (2)
k	AM253_6, AM320_1 (2)
l	CA338_6, CA338_4 (2)

4.3.4.2 Population genetics of *Strongyloides ratti* mitochondrial genome

There was a strong, positive correlation between nuclear and mitochondrial similarity in pairwise comparisons, according to a Mantel test ($r = 0.76$, $P < 0.01$). Accordingly, there was close agreement between nuclear neighbour-joining trees and minimum spanning maps of mitochondrial haplotypes (Figures 4.12 and 4.5, respectively). By eye, four mitochondrial clades were evident. Mitochondrial clade A contained all individuals from nuclear clades 3 and 5 as well as one individual from nuclear clade 2b. Mitochondrial clade B contains individuals from nuclear clades 1, 2a and 4b, and mitochondrial clade C contains only individuals from nuclear clade 2b. The two individuals of nuclear clade 4a appear as intermediate between mitochondrial clades B and C in the haplotype map and so are designated as minor mitochondrial clade D. Mitochondrial clades A, B and C contained individuals from all 3 sampling sites, though at different rates (Table 4.9). The haplotype map of DS100 was broadly the same as that of DS90, with the isofemale lines appearing variably as divergent members of mitochondrial clades A or D (Figure 4.13, Table 4.9). The maximum likelihood tree based on mitochondrial haplotypes was in close agreement with minimum spanning map, with the same

haplotypes in the same mitochondrial clades, though the position of some nodes were shifted (Figure 4.14).


Site CA
Site LA
Site AM
A
B
C
D
CA273_26
CA273_10
CA275_11
AM282_63
AM105_9
AM294_1
AM280_56
AM295_1
AM296_2
AM187_14
AM296_5
LA320_11
AM99_31
AM100_50
AM100_35
AM294_4
AM287_58
AM187_12
LA317_2
AM283_5
AM282_1
LA314_5
LA319_5
AM252_62
AM291_93
AM287_67
AM100_37
AM292_2
AM280_49
AM99_21
AM296_4
CA338_2
LA319_8
LA315_4
CA338_1
AM99_8
CA275_47
AM245_61
AM299_79
LA319_3
LA217_5
AM114_2
AM287_65
AM287_56

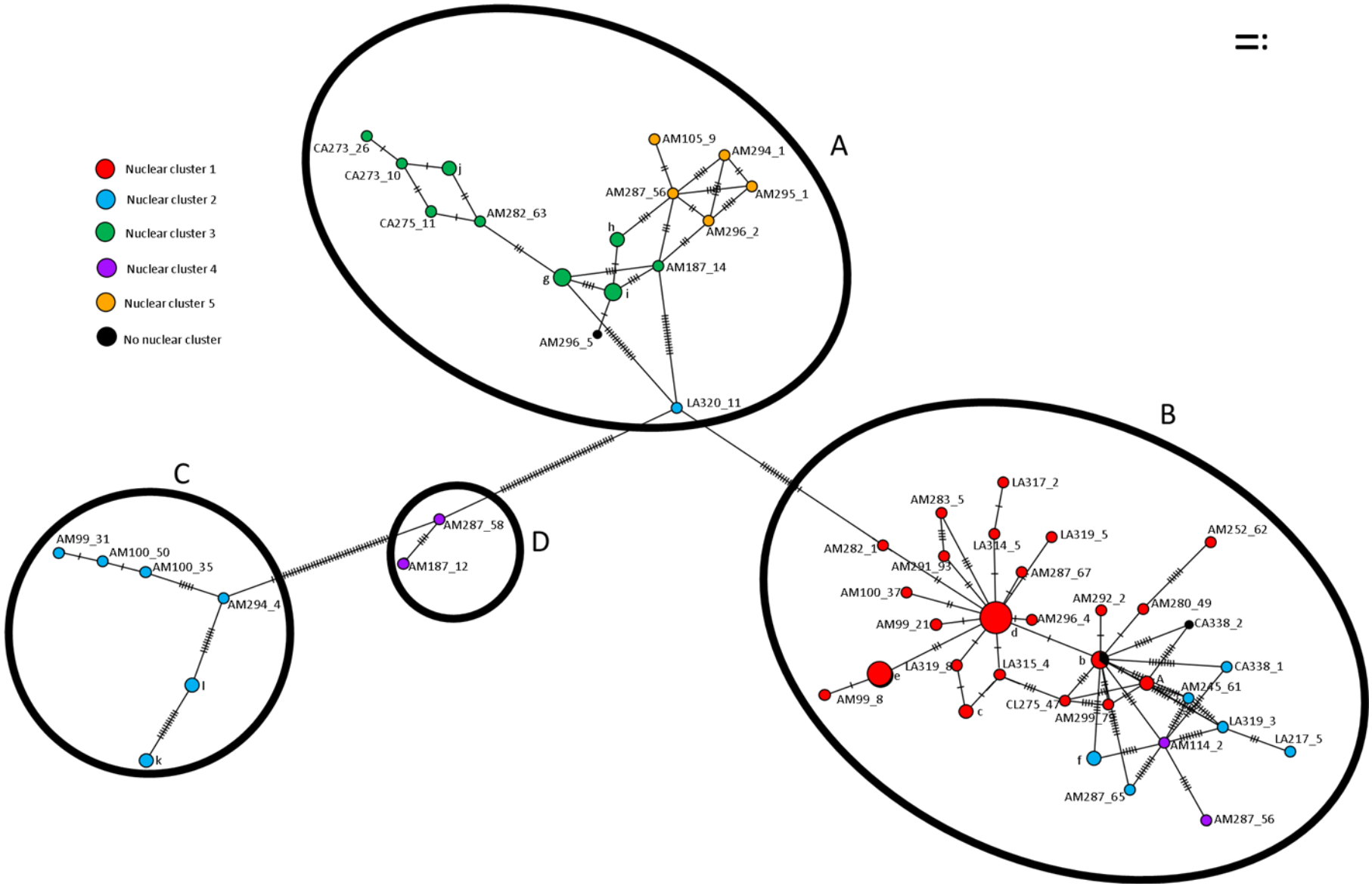

Nuclear cluster 1
Nuclear cluster 2
Nuclear cluster 3
Nuclear cluster 4
Nuclear cluster 5
No nuclear cluster
A
B
C
D
CA273_26
CA273_10
CA275_11
AM282_63
AM105_9
AM294_1
AM295_1
AM287_56
AM296_2
AM187_14
AM296_5
LA320_11
AM99_31
AM100_50
AM100_35
AM294_4
AM287_58
AM187_12
LA317_2
AM283_5
AM282_1
LA314_5
LA319_5
AM252_62
AM291_93
AM287_67
AM100_37
AM292_2
AM280_49
AM99_21
AM296_4
CA338_2
LA319_8
LA315_4
CA338_1
AM99_8
CL275_47
AM245_61
LA319_3
LA217_5
AM114_2
AM287_65
AM287_56

Figure 4.12: Minimum spanning mitochondrial haplotype maps of Strongyloides ratti *individuals in DS90, coloured according to (i) sampling site or (ii) nuclear genetic clade. Size of circles representing haplotypes is proportional to the number of samples having that haplotype. Haplotypes represented by multiple individuals are denoted by single letters, described in Table 4.8. Haplotypes are grouped by eye into four clades, circled and labelled A to D.*

Table 4.9: Expected (E) and observed (O) number of Strongyloides ratti *genomes in each of 4 mitochondrial genetic clades from each sampling site. Expected values are based on neutral expectation of clade distribution among sites. Isofemale lines refer to laboratory lines included in DS100.*

	Site CA		**Site LA**		**Site AM**		**Isofemale lines**		**Total**
	E	**O**	**E**	**O**	**E**	**O**	**E**	**O**	
Clade A	8.5	7	13.1	15	36.8	35	6.6	8	**65**
Clade B	2.9	4	4.4	4	12.4	14	2.2	0	**22**
Clade C	1.1	2	1.6	1	4.5	5	0.8	0	**8**
Clade D	0.5	0	0.8	0	2.3	2	0.4	2	**4**
Total	**13**		**20**		**56**		**10**		**99**

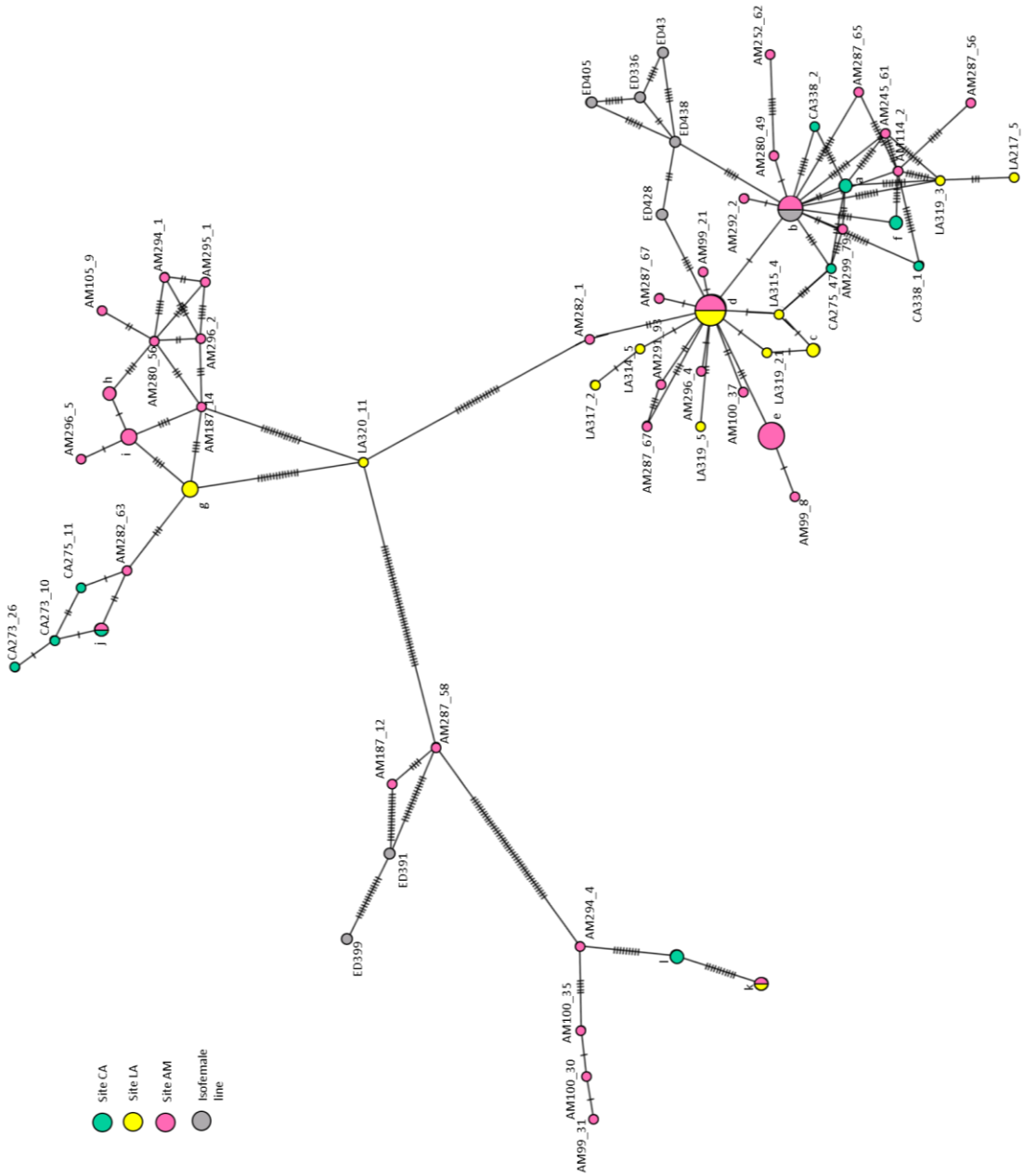


Figure 4.13: Minimum spanning mitochondrial haplotype maps of Strongyloides ratti *genomes in DS100. Size of circles representing haplotypes is proportional to the number of samples having that haplotype. Haplotypes represented by multiple individuals are denoted by single letters, described in Table 4.8.*

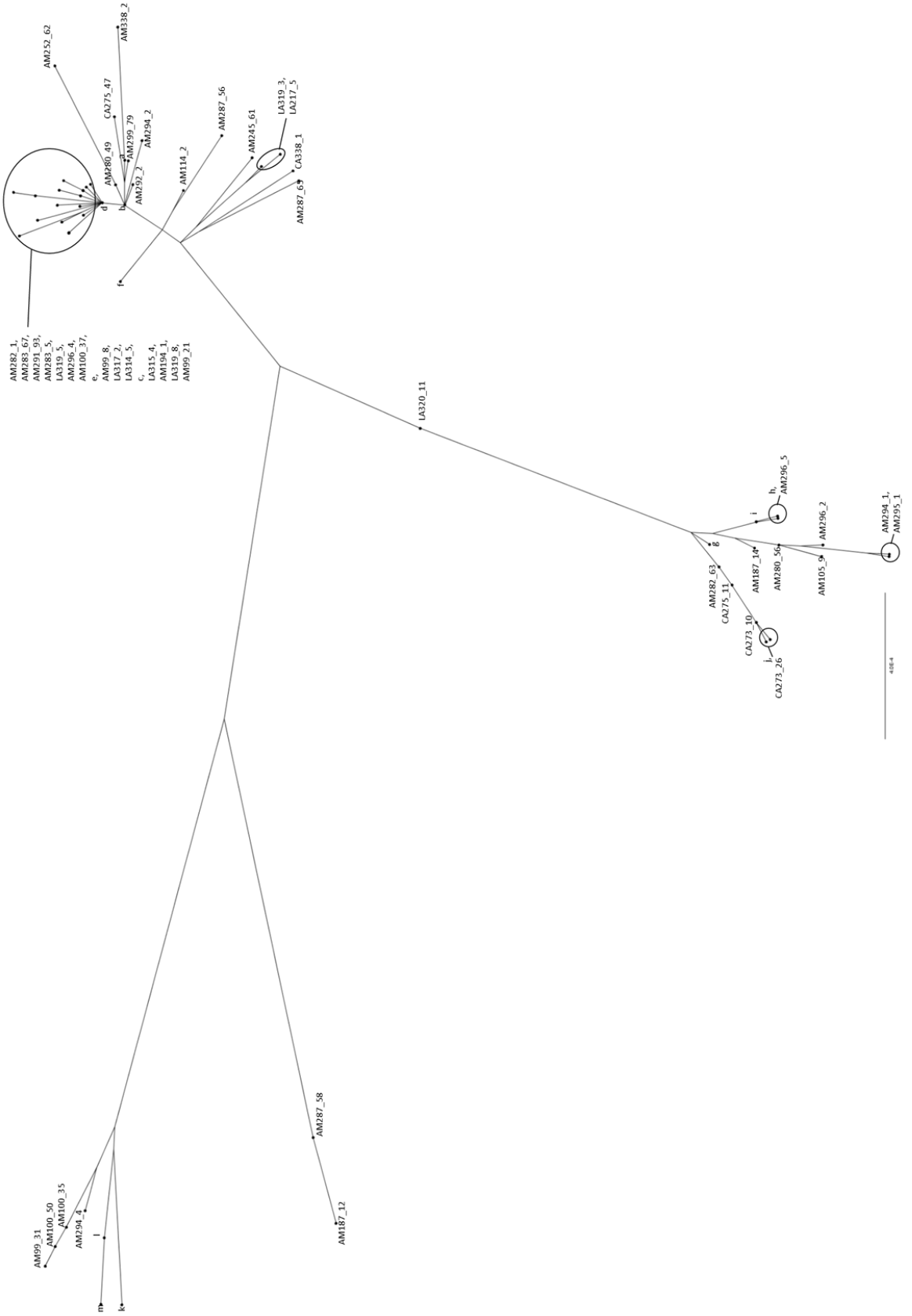


Figure 4.14 Maximum likelihood mitochondrial haplotype trees of the Strongyloides ratti *in DS90. Branch lengths are in units of substitutions per site. Single letters indicate haplotypes represented by multiple individuals, described in Table 4.8.*

AMOVA, where mitochondrial clades A, B and C were treated as distinct groups (with all other samples excluded) showed that 73% of mitochondrial variation partitioned among these mitochondrial clades. In comparison, just 1% of mitochondrial variation partitioned among geographical sampling sites. Furthermore, there Fisher's exact test indicated that the difference in frequency of the mitochondrial clades among sampling sites was non-significant (P = 0.24). Finally, there was no significant difference in the number of mitochondrial SNPs in same-site comparisons (mean = 24.2) *versus* different-site comparisons (mean = 23.6) (t = -0.98, *df* = 3771, P = 0.33). Therefore, as with the nuclear genome, the genetic structure of the *S. ratti* mitochondrial genome is characterised by 3 divergent genetic clades that are not strongly associated with geography.

4.4 Discussion

4.4.1 Genetic diversity in *Strongyloides ratti*

This study examined the whole genome sequences of 90 individual *Strongyloides ratti* isolated directly from the faeces of wild rat hosts. When all individuals were considered together, the total SNP density was 4.1 per kb, and the mean pairwise distance was 1.3 SNPs per kb. These values are similar to those among 33 *Strongyloides stercoralis* individuals analysed by whole genome amplification followed by sequencing, where total SNP density was 5.4 SNPs per kb and mean pairwise distance was approximately 0.8 SNPs per kb. (Kikuchi *et al.* 2016). *S. stercoralis* is a parasite of humans and canines that is closely related to *S. ratti* (Hunt *et al.* 2016). In contrast, total SNP density among 13 microfilaria of the filarial nematode *Wuchereria bancrofti* was only 1 SNP per kb, and the mean pairwise distance was 0.24 SNPs per kb (Small *et al.* 2016).

It should be noted, however, that total SNP density is dependent on the number of individuals sampled and the genetic structure of the populations from which they were taken. Mean pairwise SNP density is robust to the number of individuals sampled, but it also is influenced by underlying population genetic structure. Thus, direct comparisons of genetic diversity among species are challenging. For example, the 13 *W. bancrofti* microfilaria referred to previously were isolated from only 3 human hosts in a single community (Small *et al.* 2016), and microfilaria in a single host are likely to be siblings. In contrast, the *S. stercoralis* study used 33 individuals taken from 9 hosts. Therefore, the lower diversity recorded for *W. bancrofti* may not indicate that the species *W. bancrofti* is less diverse than the *Strongyloides* spp., but rather reflect the sampling strategy. Nevertheless, the present work reveals significant genetic variation within *S. ratti* in the UK.

The *S. ratti* life cycle involves a phase of asexual reproduction, where parasitic females reproduce by mitotic parthenogenesis. It is therefore possible that two individuals isolated from the same host could be clonal siblings of one another, differing from each other only in novel mutations not acquired from the shared mother. If this study has unknowingly sampled sibling groups, estimates of genetic diversity and population genetic structure in sampling sites may be affected. It is therefore important to assess the likelihood that clonal siblings have been sampled.

Because each individual will acquire a different set of novel mutations, siblings will therefore be expected to differ by twice the mutation rate (mutation rate being the number of mutations per nucleotide per generation). The mutation rate of *S. ratti* is not known, but in *Caenorhabditis elegans* it has been calculated as 2 to 3.1 x 10^{-9} (Denver *et al.* 2009). In the present study the smallest number of genetic differences between any two individuals was 240, which equates to 5.6 x 10^{-6} variants per nucleotide. If these individuals were siblings, this would make the mutation rate 2.8 x 10^{-6} per site per

generation. These individuals did come from the same host, thus making it possible that they are siblings, and the likelihood of the two genetically closest individuals coming from the same host only by chance are quite low. Of 4,006 pairwise comparisons, only 107 (2.7%) were from the same host. Nevertheless, unless the nuclear mutation rate of *S. ratti* is drastically higher than *C. elegans* (2.8×10^{-6} in *S. ratti versus* 3.1×10^{-9} in *C. elegans*), or a significant proportion of the SNPs detected are accounted for by sequencing error, it remains unlikely that these two individuals derive from the same parasitic female mother. The second highest number of genetic differences was 4,216. If these were siblings, this SNP density would indicate a mutation rate of 9.8×10^{-5} per site per generation. However, these individuals were from different hosts and therefore could not be clonal siblings.

It thus appears that, of the 90 individuals subjected to deep sequencing, none (or at most one pair) were clone-mates, even though an average of 4 individuals were taken from each host sampled. This is further corroborated by the RFLP study, where every faecal pellet tested produced worms of at least two multi-locus genotypes. Thus this study should not be affected by the accidental inclusion of numerous sibling groups. Further, this suggests that heavy *Strongyloides ratti* infections are typically diverse, consisting of many genetically distinct parasitic females that collectively produce genetically diverse offspring.

4.4.2 *Strongyloides ratti* population genetic structure

The whole genome sequences produced in this study were used to investigate population genetic structure in *S. ratti* at three levels: among host, among sampling sites, and among sampling seasons. No evidence was found for genetic differentiation of larvae among hosts, with pairwise relatedness being no higher in same-host pairs than in different-host pairs. While a previous study did detect structuring among natural *S. ratti* infrapopulations, this was weak, and the authors concluded that reinfection of the same individual by progeny of its own parasites (self-reinfection) accounts for only a minority of new infections acquired by a host (Paterson *et al.* 1999). Certainly, the diverse infrapopulations detected indicate that rats sample widely from the *S. ratti* infective larval population in the environment. However, once a diverse infrapopulation is established, it could remain diverse over multiple generations of self-reinfection due to the absence of sexual reproduction, with immigration of non-progeny parasites needed only to offset diversity lost through genetic drift. The habits of brown rats, which include use of small home ranges throughout life (Gardner-Santana *et al.* 2012) and repeated use of particular burrows (Lore and Flannelly 1978), would appear to promote self-reinfection by making rats likely to encounter their own faeces and any infective larvae that developed in it. It is likely that a combination of self-reinfection and immigration of novel parasite genotypes both contribute to the compositin of *S. ratti* infrapopulations that has been observed in this study, though the relative levels of these modes of infection are unclear. Only weak genetic structuring was found among sampling sites,

and this was driven by differentiation between site CA and the other two sampling sites. Similarly, only weak genetic structuring was observed among sampling seasons.

However, the *S. ratti* population examined is not genetically homogenous. Rather, the *S. ratti* sampled formed five genetic clades that were evident in neighbour-joining dendrograms and PCA (Figures 4.5 and 4.6) and supported by Φ relatedness data (Table 4.6). The three largest of these clades are present at all three sampling sites, and it was common to find multiple genetic clades represented in a single host. There was no evidence of a similar pattern in the population genetics of the rats (Chapter 3) so the different *S. ratti* clades do not reflect adaptation to particular host genotypes.

It is proposed that clades 1 and 3 (Figure 4.5) represent ancient genetic lineages, each derived asexually from a single founder, with the diversity observed within these clades arising by the accumulation of mutations over numerous generations of mitotic parthenogenesis. When sexual reproduction occurs in *S. ratti*'s facultative free-living adult generation, the resulting progeny all develop into parasitic females that reproduce exclusively asexually, and so can give rise to new asexual lineages. Thus, other clades and individuals may be the product of rare sexual reproduction, initially between clades 1 and 3 and subsequently among the descendants of sexually produced individuals.

However, PCA indicates a second axis (PC2 in Figure 4.6) that suggests the presence of a third "parental" diplotype (that is, a whole, diploid genome sequence characteristic of a particular genetic clade). Furthermore, while the mitochondrial genome of *S. ratti* showed a similar population genetic structure to the nuclear genome, the clades formed by the nuclear and mitochondrial genomes are not identical (Figures 4.5 and 4.12). That nuclear clades 1 and 3 fall intact into separate mitochondrial clades B and A supports the view that these nuclear clades represent ancient lineages each derived from a single asexual individual, as it would indicate a lack of nuclear and mitochondrial recombination expected during sexual reproduction. However, the origin of mitochondrial clade C, which exclusively contains individuals of nuclear sub-clade 2b, is less clear. It is proposed that individuals of nuclear clade 2b / mitochondrial clade C comprise a third ancient, asexually derived lineage, comparable to nuclear clades 1 and 3.

Comparison of nuclear and mitochondrial genetic clades also gives insight into the history of sexual reproduction in this population. Nuclear clade 5 is positioned between nuclear clades 1 and 3 on PC1 (Figure 4.6) and its mitochondrial sequences clade together within mitochondrial clade A. It is therefore likely that nuclear clade 5 is derived from a single individual produced by sexual crosses between individuals belonging to nuclear clades 1 and 3, with this individual then reproducing asexually over multiple generations to give the diversity seen in nuclear clade 5 today. Individual AM296_5 would appear to be result of further crossing between nuclear clades 5 and 3, with nuclear clade 3 taking the

maternal role. Nuclear sub-clades 2a (except for LA320_11) and 4a, as well as individuals AM292_3 and CA338_2 probably derive from various crosses, with ancestors nuclear clade 1 providing the mitochondrial sequence. Individual LA320_11 appears to have arises from a cross or crosses between ancestors belonging to nuclear sub-clades 2b and 3, with nuclear clade 3 providing the mitochondrial. genome. This is supported by the placement of LA320_11 in projections of PCs 1 and 2 (Figure 4.6) The case of LA320_11 highlights the fact that individuals in nuclear sub-clade 2a, are not necessarily all the product of a single sexual cross followed by asexual radiation, and the diffuse nature of sub-clade in PCA projections support the view that multiple crosses may have been involved. Indeed, it cannot be ruled out that some individuals assigned to nuclear clades 1, 2b or 3 are not in fact direct asexual descendants of the founder individual of these clades, but rather derived from crosses among ancestors of these clades and followed by backcrosses to a particular parental line such that the nuclear genome closely resembles that of true direct descendants. Mitochondrial clade D contains the two individuals belonging to nuclear sub-clade 4b, but also isofemale lines that do not group with other sub-clade 4b individuals according to nuclear data. It is suggested therefore that clade D represents a fourth *S. ratti* lineage, the mitochondrial genome of which has been incorporated into the study population in absence of the nuclear genome. This separation of mitochondrial and nuclear genomes can occur through a sexual cross between two lineages, followed by repeated backcrossing to one of the paternal lineages, leading to introgression of the maternal mitochondrial sequence with the paternal nuclear genome (Ballard and Whitlock 2003).

It should be noted that any of the crosses described above could have occurred many generations ago, with the nuclear genome resulting from these crosses preserved, barring novel mutations, through repeated rounds of mitotic parthenogenesis. Indeed, these crosses may not have occurred within the study region or even the UK. Rather, genomes resulting from crosses may have spread via asexual reproduction far from the place of origin.

It is interesting to note that the laboratory isofemale lines contributing to DS100 were not genetically distinct from the individuals collected in this present study, but rather slotted into the clades formed by DS90 (Figures 4.5 and 4.6). Each of these isofemale lines were derived from a single infective larva collected from natural infections across the UK and Japan (Table 4.1B). That other British isolates fall into the same clades indicates that the genetic structure observed in the present study is widespread across the UK. All but one of the UK isofemale lines grouped with clade 2 individuals (Figure 4.5B), suggesting that the relative abundance of individuals belonging to each of the major clades may vary across the UK. The isofemale lines were collected over a wide timescale (Table 4.1B), so temporal changes could also contribute. More sampling is needed to confirm the distribution of clades across the UK. That the Japanese samples are not genetically distinct from the UK samples is remarkable, and suggests clade 2, with which the Japanese isofemale lines group, may be present throughout *S. ratti*'s

global range. Sequencing of further individuals from outside the UK will be required to determine how geographically widespread the clades identified here are.

It therefore appears that the population genetics of *S. ratti* is defined principally on the level of sympatric but genetically distinct clades, with geography contributing only very little to the species' population genetic structure. The appearance of weak genetic divergence between site CA and the other two sites appears to be due to minor (statistically insignificant) differences in the relative frequency of nuclear clades 1, 2 and 3 among sites. Given the small number of individuals sampled from site CA, this difference may simply be due to chance.

It is expected that rat movement is the main mechanism of *S. ratti* dispersal. It is therefore presumably due to rat movement that the observed *S. ratti* clades have spread across geographical regions sampled. However, the difference in the relative ratios of the major clades among sites is weaker than what might be expected given the population genetic structuring among sites observed in the rat hosts (Chapter 3). Further, the *S. ratti* differentiation among sites was driven principally by dissimilarity between site CA and the other two sites, while from the perspective of rat hosts, sites CA and LA were genetically the most similar. Thus, the sampling site differentiation of *S. ratti* and of rats not only differ in degree but also show an incongruent pattern. It is hypothesised that the rat populations at sampling sites derive from random sampling of a few individuals from a large, genetically diverse metapopulation (Chapter 3). The minimum distance between sampling sites in this study is 9.7 km, much further than the distance typically travelled by a rat over the course of its life (Gardner-Santana *et al.* 2012). Thus rat and parasite gene flow between these sites is unlikely to be direct but rather probably occurs over several generations. Such stochastic migration of hosts means that a lineage of rats that eventually spreads from one sampling site to another will have had ample time to acquire *S. ratti* not descended from individuals at the original sampling site, thus explaining the incongruity.

A previous study comparing nematode parasite and rodent host population genetics revealed more concordance than is seen here, with the nematode *Heligmosomoides polygyrus* showing stronger but similar patterns of population genetic structuring compared to its European woodmouse (*Apodemus sylvaticus*) host (Nieberding *et al.* 2004). This study considered a much larger geographical area than used in the present work, covering continental Europe from Portugal to Ukraine, and it is possible that greater population genetic structure would be detected if rats and *S. ratti* were sampled at similar scales.

4.4.3 Asexuality and population genetics in *Strongyloides ratti*

In a population reproducing solely by mitotic parthenogenesis, where all individuals are produced asexually without recombination and in the absence of processes leading to DNA loss, novel mutations

cannot escape the chromosome on which they arose and will remain permanently in a heterozygous state. Thus, in such a population, heterozygosity is expected to accumulate (Birky 1996, Mark Welch and Meselsen 2000). Furthermore, in the complete absence of recombination, pan-genome linkage of all loci is expected (Tibayrenc *et al.* 1991). However, any amount of sexual recombination, however rare, has the potential to disrupt these genetic consequences of mitotic parthenogenesis (Bengtsson 2003, de Meeûs and Balloux 2004).

The effect of asexual reproduction on heterozygosity and linkage is demonstrated by species wherein the rate of asexual reproduction differs among populations. *S. stercoralis* is unusual among *Strongyloides* spp. in that the asexual cycle can be completed in a single host, leading to persistence of infection well beyond the lifespan of any one parasite without acquisition of new parasites from the environment (Grove 1989). However, in *Strongyloides* spp. sexual reproduction requires a free-living generation. Populations of *S. stercoralis* in a non-endemic region, where good sanitation prevents infection of new hosts by sexually produced larvae, show higher heterozygosity than populations in endemic regions where transmission and therefore potentially sexual reproduction are ongoing (Kikuchi *et al.* 2016). Similarly, in a population genetics study of the aphid *Myzus persicae*, linkage among markers was detected only in fully parthenogenetic populations, and not in populations that underwent periodic sexual reproduction (Guillemaud *et al.* 2003). Therefore, populations without sexual reproduction display greater linkage and greater heterozygosity than do sexual ones when both populations are from the same species.

The present work detected an excess of heterozygote *S. ratti*. A similar pattern has been detected in other analyses of *S. ratti* population genetics (Fisher and Viney 1998, Paterson *et al.* 2000), and likely arises from its predominantly asexual mode of reproduction. However, in previous studies and this one, heterozygous excess is not at the level expected in a population where recombination is totally absent In the present study, the average individual was homozygous at 64% of polymorphic sites, compared with the 72% expected under neutral conditions. When considering all individuals analysed in the current study, linkage decayed to background levels within distances of approximately 50 kb in the autosomes, and 20 kb on the X chromosome (Figure 4.7), and no large linkage blocks were observed. This is much more gradual linkage decay than in the fully sexual free-living nematode *Caenorhabditis remanei*, in which linkage approached background levels over distances of less than 1 kb (Cutter *et al.* 2006). However, other fully sexual species have been found to have more gradual linkage decay than that observed for *S. ratti*, including some populations of humans (Park 2012) and domestic pigs (Amaral *et al.* 2008). Linkage disequilibrium is dependent on a variety of past and present demographic factors, as well as the rate of recombination per meiosis, so that it cannot be used to predict the rate of sexual reproduction directly.

The heterozygosity excess and linkage seen observed in this study is not compatible with a complete absence of sexual reproduction in this *S. ratti* population. It is likely that, as well as crossing among nuclear clades, there is sexual reproduction within clades too. Because free-living male and female *S. ratti* can both be produced asexually by a single parasitic female, it is possible for clonal siblings to undergo sexual reproduction. This is equivalent to selfing within a hermaphrodite, and would be expected to reduce heterozygosity (David *et al.* 2007). This might compensate for increased heterozygosity caused by repeated generations of mitotic parthenogenesis.

Sexual reproduction is presumed to be rare in UK *S. ratti* populations because the heterogonic (free-living, dioecious, obligately sexual) stages are rarely observed. Indeed, in this study, 10,471 *S. ratti* were isolated from wild rat faeces (details in Chapter 2) and all of these were infective larvae, these being female larvae already committed to homogonic (parasitic, obligately asexual) developmental (See Fig 1.1 for *S. ratti* life cycle). This is consistent with previous studies of UK *S.* ratti (Viney *et al.* 1992). It is possible that some free-living *S. ratti* were missed or misidentified as other nematode species during culture inspection, but notwithstanding, the rate of development of the free-living sexual generation, must be very low. However, it is known that the rate of heterogonic development is influenced by, among other things, the temperature young female larvae experience upon leaving the parental host (Harvey *et al.* 2000). The culture temperature used in this study (19°C) has previously been shown to promote free-living female development in the laboratory line ED321, with a higher temperature causing a shift to heterogonic development (Harvey *et al.* 2000). However, there is a strong genetic component to the rate of homogonic development too (Viney *et al.* 1992). It may be that in the UK *S. ratti* population, critical temperatures are shifted downwards with respect to ED321 so that cooler culture temperatures would induce some females to develop heterogonically. However, sex is determined genetically, and is decided prior to an individual leaving the parental host (Harvey *et al.* 2000). It may be that culture induces male mortality in the *S. ratti* population studied, but otherwise culture conditions should not influence sex ratio. It is likely that males are simply produced at a level that, although very low, is nevertheless sufficient to prevent extreme heterozygous excess and pan-genome linkage.

4.4.4 Conclusion

This study used whole genome sequencing to analyse the population genetics of a parasitic nematode, *Strongyloides ratti*. High levels of diversity were observed both within rat hosts and within sampling sites, and there was very limited population genetic structuring at either of these levels. However, three strongly differentiated clades were observed, supported by both nuclear and mitochondrial genetic data, and each of these clades were found across all three sampling sites. It is hypothesised that each of these clades was founded by a single asexual

individual and has subsequently propagated asexually, perhaps with occasional sexual reproduction among clade members, acquiring diversity through novel mutation. Other individuals may have arisen as a result of rare sexual crosses among clade members, which give rise to new, intermediate genotypes that in turn then expand in frequency asexually.

4.5 References

Amaral A. J., Megens H.-J., Crooijmans R. P. M. A., Heuven H. C. M. and Groenen M. A. M. (2008). Linkage disequilibrium decay and haplotype block structure in the pig. *Genetics* **179:**569-579.

Ballard J. W. O. and Whitlock M. C. (2003). The incomplete natural history of mitochondria. *Molecular Ecology* **13:**729-744.

Bandelt H., Forster P. and Röhl A. (1999). Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution* **16:**37-48.

Bengtsson B. O. (2003). Genetic variation in organisms with sexual and asexual reproduction. *Journal of Evolutionary Biology* **16:**189-199.

Birky C. W. (1996). Heterozygosity, heteromorphy, and phylogenetic trees in asexual eukaryotes. *Genetics* **144:**427-437.

Bradbury P. J., Zhang Z., Kroon D. E., Casstevens T. M., Ramdoss Y. and Buckler E. S. (2007). TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics* **23:**2633-2635.

Browning S. R. and Browning B. L. (2007). Rapid and accurate haplotype phasing and missing data inference for whole genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics* **81:**1084-1097.

Browning B. L., Zhou Y. and Browning S. R. (2018). A one-penny imputed genome from next generation reference panels. *American Journal of Human Genetics* **103:**338-348.

Choi Y.-J., Bisset S. A., Doyle S. R., Hallsworth-Pepin K., Martin J., Grant W. N., Mitreva M. (2017) Genomic introgression mapping of field-derived multiple-anthelmintic resistance in *Teladorsagia circumcincta*. *PLoS Genetics* **13:**e1006857.

Cole R. and Viney M. (2018). The population genetics of parasitic nematodes of wild animals. *Parasites and Vectors* **11:**590.

Coles G. C., Jackson F., Pomroy W. E., Prichard R. K., von Samson-Himmelstjerna G., Silvestre A., Taylor M. A. *et al.* (2006) The detection of anthelmintic resistance in nematodes of veterinary importance. *Veterinary Parasitology* **136:**167-185.

Criscione C. D., Poulin R. and Blouin M. S. (2005) Molecular ecology of parasites: elucidating ecological and microevolutionary processes. *Molecular Ecology* **14:**2247-2257.

Cutter A. D., Baird S. E. and Charlesworth D. (2006). High nucleotide polymorphism and rapid decay of linkage disequilibrium in wild populations of *Caenorhabditis remanei*. *Genetics* **174:**901-915.

Danecek P., Auton A., Abecasis G., Albers C. A., Banks E., DePristo M., Handsaker R. E. (2011). The variant call format and VCFtools. *Bioinformatics* **27:**2156-2158.

David P., Pujol B., Viard F., Castella V. and Goudet J. (2007). Reliable selfing rate estimates from imperfect population genetic data. *Molecular Ecology* **16:**2474-2487.

De Meeûs T. and Balloux F. (2004). Clonal reproduction and linkage disequilibrium in diploids: a simulation study. *Infection, Genetics and Evolution* **4:**345-351.

Denver D. R., Dolan P. C. Wilhelm L. J., Way S. J., Lucas-Lledó I., Howe D. K., Lewis S. C. *et al.* (2009). A genome-wide view of *Caenorhabditis elegans* base-substitution mutation processes. *Proceedings of the National Academy of Sciences USA.* **106:**16310-16314.

Dray S. and Dufour A. (2007). The ade4 Package: Implementing the Duality Diagram for Ecologists. *Journal of Statistical Software* **22:**1-20.

Fisher M. C. and Viney M. E. (1998). The population genetic structure of the facultatively sexual parasitic nematode *Strongyloides ratti* in wild rats. *Proceedings of the Royal Society B: Biological Sciences*. **265:**703–709.

Frainer A., McKie B, G., Amundsen P.-A. and Lafferty K. D. (2018). Parasitism and the biodiversity-functioning relationship. *Trends in Ecology and Evolution* **33:**260-268.

Gardner-Santana L. C., Norris D. E., Fornadel C. E., Hinson E. R., Klein S. L. and Glass G.E. (2009). Commensal ecology, urban landscapes, and their influence on the genetic characteristics of city-dwelling Norway rats (*Rattus norvegicus*). *Molecular Ecology* **8:**2766-2778.

Gilabert A. and Wasmuth J. D. (2013) Unravelling parasitic nematode natural history using population genetics. *Trends in Parasitology* **29:**438-448.

Gilleard J. S. and Redman E. (2016) Genetic diversity and population structure of *Haemonchus contortus. Advances in Parasitology* **93:**31-68.

Gorton M. J., Kasl E. L., Detwiller J. T., Criscione C. D. (2012). Testing local-scale panmixia provides insights into the cryptic ecology, evolution, and epidemiology of metazoan animal parasites. *Parasitology* **139:**981-997.

Grove D. I. (1989). *Strongyloidiasis: A Major Roundworm Infection of Man*. London: Taylor and Francis

Guillemaud T., Mieuzet L. and Simon J.-C. (2003). Spatial and temporal genetic variability in French populations of the peach–potato aphid, *Myzus persicae*. *Heredity* **91:**143-152.

Harvey S. C., Gemmil A. W., Read A. F. and Viney M. (2000). The control of morph development in the parasitic nematode *Strongyloides ratti*. *Proceedings of the Royal Society of London B: Biological Sciences* **267:**2057-2063.

Hosono S., Faruqi A. F., Dean F. B., Du Y., Sun Z., Wu X., Du J. *et al.* (2003). Unbiased whole-genome amplification directly from clinical samples. *Genome Research* **13:**954–964.

Howe K. L., Bolt B. J., Shafie M., Kersey P and Berriman M. (2017). WormBase ParaSite − a comprehensive resource for helminth genomics. *Molecular and Biochemical Parasitology* **215:**2-10.

Hunt V. L., Tsai I. J., Coghlan A., Reid A. J., Holroyd N., Foth B. J., Tracey A. *et al.* (2016). The genomic basis of parasitism in the *Strongyloides* clade of nematodes. *Nature Genetics* **48:**299-307.

Jones F. C., Grabherr M. G., Chan Y. F., Russel P., Mauceli E., Johnson J., Swofford R. *et al.* (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484:**55–61.

Kikuchi T., Hino A., Tanaka T., Aung M. P. P. T. H. H. A., Afrin T., Nagayasu E., Tanaka R. *et al.* (2016). Genome-wide analyses of individual *Strongyloides stercoralis* (Nematoda: Rhabditoidea) provide insights into population structure and reproductive life cycles. *PLoS Neglected Tropical Diseases* **10:**e0005253.

Langmead B. and Salzberg S. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9:**357-359.

Lee C., Abdool A., Huang C.-H. (2009). PCA-based population structure inference with generic clustering algorithms. *BMC Bioinformatics* **10:**S73.

Leigh JW and Bryant D. (2015). PopART: Full-feature software for haplotype network construction. *Methods in Ecology and Evolution* **6:**1110-1116.

Lewontin R. C. (1963). The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* **49:**49-67.

Li H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27:**2987-2993.

Li H. and Durbin R. (2011). Inference of human population history from individual whole-genome sequences. *Nature* **475:**493–496.

Li H., Handsaker B., Wysoker A., Fennel T., Ruan J., Homer N., Marth G. *et al.* (2009). The sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* **25:**2078-2079.

Lore R. and Flannelly K. J. (1978). Habitat selection and burrow construction by wild *Rattus norvegicus* in a landfill. *Journal of Comparative and Physiological Psychology* **92:**888-896.

Luu K., Bazin E. and Blum M. G. (2017). pcadapt: an R package to perform genome scans for selection based on principal component analysis. *Molecular Ecology Resources* **17:**67-77.

Manichaikul A., Mychaleckyj J. C., Rich S. S., Daly K., Sale M., Chen W. M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics* **26:**2867-28773.

Mark Welch, D. and Meselson M. (2000). Evidence for the evolution of bdelloid rotifers without sexual reproduction or genetic exchange. *Science* **288:**1211-1215.

McVean G. (2009). A genealogical interpretation of principal components analysis. *PLoS Genetics* **5:**e1000686.

Mobegi V. A., Duffy C. W., Amambua-Ngwa A., Loua K. M., Laman E., Nwakanma D. S., MacInnis B. *et al.* (2014) Genome-wide analysis of selection on the malaria parasite *Plasmodium falciparum* in West African populations of differing infection endemicity. *Molecular Biology and Evolution*. **31:**4190–4199.

Small S. T., Reimer L. J., Tisch D. J., King C. L., Christensen B. M., Siba P. M., Kazura J. M. *et al.* (2016). Population genomics of the filarial nematode parasite *Wuchereria bancrofti* from mosquitoes. *Molecular Ecology* **25:**1465-1477.

Nieberding C., Morand S., Libois R., Michaux J. R. (2004). A parasite reveals cryptic phylogeographic history of its host. *Proceedings of the Royal Society of London B: Biological Sciences* **271:**2559–2568.

Nosil P., Villoutreix R., de Carvalho C. F., Farkas T. E., Soria-Carrasco V., Feder J. L. Crespi B. J. *et al.* (2018). Natural selection and the predictability of evolution in *Timema* stick insects. *Science* **359:**765-770.

O'Connell J., Gurdasani D., Delaneau O., Pirastu N., Ulivi S., Cocca M., Traglia M. *et al.* (2014). A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genetics* **10:**e1004234.

Ott J. Wang J., Leal S. M. (2015). Genetic linkage analysis in the age of whole-genome sequencing. *Nature Reviews Genetics* **16:**275-284.

Park L. (2012). Linkage disequilibrium decay and past population history in the human genome. *PLoS One* **7:**e46603.

Paterson S., Fisher M. C. and Viney M. (1999). Inferring infection processes of a parasitic nematode using population genetics. *Parasitology* **120:**185-194.

Peakall R. and Smouse P. E. (2006). GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes* **6:**288-295.

Peakall R. and Smouse P. E. (2012). GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research-an update. *Bioinformatics* **28:**2537-2539.

Sabina J. and Leamon J. H. (2015). Bias in whole genome amplification: causes and considerations. *Methods in Molecular Biology* **1347:**15–41.

Stamatakis A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22:**2688-2690.

Tibayrenc M., Kjellberg F., Arnaud J., Oury B., Breniere F., Darde M. L. and Ayala F. J. (1991). Are eukaryotic microorganisms clonal or sexuals? A population genetics vantage. *Proceedings of the National Academy of Sciences USA* **88:**5129-5133.

Viney M., Matthews B. E. and Walliker D. (1992) On the biological and biochemical nature of cloned populations of *Strongyloides ratti*. *Journal of Helminthology* **66:**45-52.

Wood C. L. and Johnson P. T. J. (2015). A world without parasites: exploring the hidden ecology of infection. *Frontiers in Ecology and the Environment* **13:**425-434.

5. Diversity within the *Strongyloides ratti* genome

5.1 Introduction

5.1.1 Natural selection in parasites

Parasites must constantly adapt to better exploit their hosts' resources, and hosts must constantly adapt to better protect their resources from their parasites (van Valen 1973). It has been suggested that this 'arms race' is a driving force behind diversification, speciation and evolution in both host and parasite species (Betts *et al.* 2018). Thus, there is substantial interest in examining adaptation in the context of a host-parasite relationship via theory, experimentation and modelling (Chao *et al.* 2000, Fisher and Schmid-Hempel 2005, Betts *et al.* 2018, Koskella 2018). DNA sequencing of large genomic regions increasingly contributes to this area of study. Of particular interest are genes that may be directly involved in aspects of the parasitic lifestyle, such as evasion of host immunity or extraction of nutrients from the host. Comparisons of two strains of *Schistosoma mansoni*, a trematode parasite of humans, showed that genetic diversity was reduced in regions of the genome where genes putatively involved in parasite-host interaction were concentrated (Clément *et al.* 2013). This was interpreted by the authors as evidence of selective sweeps in these low diversity regions. In a selective sweep, alleles increase in frequency in the population due to close physical linkage with an allele that is under positive selection, until ultimately fixation is reached and diversity is eliminated. In the context of the *S. mansoni* study, this suggests that parasitism-associated genes are under purifying selective pressure.

Otherwise, studies of selection in eukaryotic parasites have often focused on drug resistance as the phenotype of interest, using DNA sequencing to pinpoint the genetic and genomic features underlying this resistance. These studies can help to elucidate the biochemical mechanism of resistance, which in turn can inform the design of new drugs that overcome this resistance. Malaria-causing *Plasmodium* spp. have been the subject of particularly intensive study in this regard (Culleton *et al.* 2005, Ekland and Fidock 2007, Anderson *et al.* 2010, Hunt *et al.* 2010, Park *et al.* 2012), and similar studies have been attempted in helminths (reviewed in James *et al.* 2009, Matoušková *et al.* 2016), especially the sheep parasite *Haemonchus contortus* (Sangster *et al.* 1999, Williamson *et al.* 2011). Whole genome sequencing was used to identify genes associated with drug resistance in another sheep parasite *Teladorsagia circumcincta*, which provides a broad view of adaptation to drug exposure (Choi *et al.* 2017). However, anti-parasitic drug exposure is likely to be a very strong selection pressure that is rarely seen in natural systems, and so findings of studies focusing on drug resistance may not be applicable to environments where such drugs (or other such strong selection pressures) are scarce.

There is now a need for selection studies that look across the genome of eukaryotic parasites, as in the *S. mansoni* study (Clément *et al.* 2013), but that considers species in natural ecosystems and that study

multiple, wild-collected individuals. Studies such as these will enhance our understanding of evolution and adaptation of eukaryotic parasites under natural conditions. In this chapter, such a study is carried out with wild-collected *Strongyloides ratti* individuals.

5.1.2 Selection in non-recombining genomes

The *S. ratti* lifecycle includes obligatory mitotic parthenogenesis with some facultative sexual reproduction (Figure 1.1). Indeed, evidence from previous chapters of this thesis (Chapters 2 and 4) suggests that, at least in some populations, sexual reproduction may be very rare. Thus, the *S. ratti* genome may undergo very little genetic recombination compared with an obligatorily sexual species, and this will undoubtedly have consequences for how species respond to selection pressure, including at the DNA sequence level.

Caenorhabditis elegans is a free-living nematode that reproduces predominantly by the self-fertilisation with rare outcrossing. Repeated rounds of self-fertilisation amounts to extreme inbreeding that leads to a rapid loss of heterozygosity, such that genetic recombination becomes ineffectual (Andersen *et al.* 2012). Thus, a parallel can be drawn between the *C. elegans* and *S. ratti* life cycles – both involve a preponderance of reproduction without recombination, with occasional crossing between individuals. Double digest restriction site associated DNA sequencing (ddRADSeq) sequences DNA associated with restriction enzyme recognition sites, these sites being widely present throughout the genome. ddRADSeq was used to investigate the genomic features of wild *C. elegans* and revealed evidence for repeated widespread selective sweeps that homogenised the global populations (Andersen *et al.* 2012). The authors suggest that low levels of genetic recombination in *C. elegans* result in the entire genome acting as a single linkage block, such that positive selection on only a very few alleles could increase the frequency of associated alleles across the whole genome. (Andersen *et al.* 2012). It is reasonable to assume that *S. ratti* may face similar genome-wide selective sweeps, however the population genetics observed in this species is not consistent with this. If the fittest genomes always swept to predominance as they arose, one would expect low levels of diversity in *S. ratti* as in *C. elegans*, but instead there is evidence for multiple, genetically distinct clades occurring in sympatry, in multiple sampling sites (Chapter 4). This observation suggests that it is important to understand how selection may be acting in the *S. ratti* genome, which may be having an important effect on its population genetic structure. An investigation into how selection is acting upon the *S. ratti* genome is called for to examine this unexpected finding.

5.1.3 'Parasitism genes' in *Strongyloides ratti*

Transcriptomic work in *S. ratti* has revealed genes that are differentially expressed in the parasitic adult female morph compared with the free-living adult female morph (see Section 1.2.2 and Figure 1.1 for

S. ratti life history) (Hunt *et al.* 2016). These genes are hypothesised to play a role in processes that differ between the two adult female morphs. Thus, genes that are highly expressed in the parasitic morph, but not the free-living morph, may have functions that facilitate a parasitic lifestyle, for example by contributing to evasion of host immune responses or extraction of nutrients from the host. In this chapter, a 'parasitism gene' is defined as one that is upregulated in the parasitic adult female morph compared with the free-living morph, with the base 2 log fold difference being greater than 1. A 'free-living gene' is the reverse.

It has been shown previously (Hunt *et al.* 2016) that *Strongyloides* spp. have substantially more genes encoding astacins, CAP domain-containing proteins (also known as SCP/TAPS proteins) or acetylcholinesterases than does a non-parasitic relative with which *Strongyloides* is believed to have a recent non-parasitic common ancestor. The majority of these genes are expressed more strongly in the parasitic female morph than the free-living one, further suggesting that these expanded gene families have an important role in parasitism (Hunt *et al.* 2016). Astacin- and CAP domain-containing proteins have both expanded multiple times independently in different parasitic nematode lineages (Tang *et al.* 2014, Hunt *et al.* 2017), and have been proposed to have roles in digestion of host tissue (Mello *et al.* 2009) and host immunomodulation (Cantacessi *et al.* 2009) respectively. In contrast, the role of acetylcholinesterases in nematode parasitism is not clear. *In vitro* studies suggest that these proteins have a modulatory effect on host epithelial tissue (Huby *et al.* 1999), but some parasitic nematodes produce acetycholinesterases that do not appear to have classical acetylcholinesterase function (Hussein *et al.* 2002), and indeed this is the case for many *S. ratti* acetylcholinesterases (Hunt *et al.* 2017). In *S. ratti*, these genes are not distributed evenly across the genome, but are instead concentrated in a number of genomic regions (Hunt *et al.* 2016). These regions may therefore be particularly important parts of the genome for the parasitic lifestyle.

Previous work used comparisons among parasitic species, and between parasitic and non-parasitic species to interrogate the genomic changes that may have facilitated the evolution of parasitism in Strongylididae nematodes (Hunt *et al.* 2016). By analysing the distribution of intraspecific genetic variation across genes involved in parasitism, this work attempts to understand the continued evolution of genes putatively underlying the parasitic lifestyle, which is likely to be dynamic and ongoing.

5.1.4 Aims of this Chapter

The aim of the work presented in this chapter is to investigate variation within the *Strongyloides ratti* genome, identifying and characterising regions of the genome that are potentially under selection. Further, this chapter focuses on genes and regions of the genome that are putatively involved in the parasitic lifestyle, in order to better understand ongoing evolution of traits involved in parasitism.

5.2 Materials and methods

5.2.1 *Strongyloides ratti* used in this chapter

Strongyloides ratti were collected as described in Chapter 2 from three sampling sites (Table 2.1). Whole genome sequencing data suitable for use in population genetic and selection analysis was produced for 90 individual *S. ratti*, as described in Chapter 4. In brief, sequencing data was generated by Illumina technology and consisted of short, paired reads. Individual worms used in this chapter correspond to the DS90 dataset used in Chapter 4, described in Table 4.1, and do not include any laboratory lines.

5.2.2 Assessment of assembly quality in expansion clusters

5.2.2.1 Definition of 'expansion cluster' and 'flanking region'

An 'expansion cluster' is here defined as a stretch of reference assembly containing four or more genes coding for members of one of three protein families expanded in *S. ratti* (Astacins, CAP domain-containing proteins and acetylcholinesterases), with not more than one other gene between any two genes of those families. Previously, groups of adjacent genes belonging to expanded clusters were called 'parasitism regions', but these were not explicitly defined (Hunt *et al.* 2016). The definition of expansion clusters used here is novel.

A 'flanking region' is a stretch of reference assembly directly next to an expansion cluster that is the same size as the cluster itself. Each expansion cluster has two flanking regions, so that the combined length from the start of one flanking region, through the expansion cluster and to the end of the other flanking region, is three times the length of the expansion cluster alone. Flanking regions were included as controls in most analyses of expansion clusters.

Lists of genes belonging to the expanded gene families were collated from supplementary material of Hunt *et al.* (2016). Fifteen expansion clusters were detected in the *S. ratti* genome, listed in Table 5.1. Due to the close physical proximity of clusters 10 and 11 in the genome, the right flanking region of cluster 10 was shortened to end where expansion cluster 11 began, and expansion cluster 11 was considered to not have a left flanking region. Similarly, the right flanking region of cluster 11 and left flanking region of cluster 12 would overlap if both were treated as being of the same length of their respective expansion clusters. Hence the left flanking region of cluster 12 was shortened to begin where the right flanking region of cluster 11 began. Across all expansion clusters there were 132 genes in total, of which 126 belonged to one of the three expanded gene families. These were made up of 46 encoding for CAP domain-containing proteins, 70 encoding astacins and 10 encoding acetylcholinesterases, representing 51.7%, 38% and 33.3% of CAP domain-containing proteins, astacin and

acetylcholinesterase coding genes in the whole genomes. Flanking regions collectively contained 216 protein-coding genes the products of which had varying functional descriptions (Table 5.1).

Table 5.1: Expansion clusters (EC) identified in the Strongyloides ratti *genome and associated flanking regions (FR), as defined in Section 5.2.2.1. "Function" refers to the functional description given on WormBase ParaSite, with "n/a" recorded here as "hypothetical protein". AchE denotes acetylcholinesterase. A block of colour represents a given expansion cluster and its flanking regions*

Region	**Position**	**Genes** (Astacin / CAP / AchE / other)	**Gene functional descriptions** (Astacin / CAP / AchE excluded)
FR1L	SRAE_chr1:5640744-5682824	18 (0 / 0 / 0 / 18)	1x Transrciptional regulator ATRX 1x Formin 1x Glycosylphosphatidylinositol-mannosyltransferase I 1x Rhodopsin-like 1x TPM domain-containing 1x MIF4-like 1x Sodium/potassium-transporting ATPase α subunit 1x Mediator of RNA polymerase II transcription subunit 9 1x Serine/threonine-protein kinase Chk1 1x Eukaryotic translation initiation factor 2A 1x Heme transporter HRG 7x Hypothetical protein
EC1	SRAE_chr1:5682825-5683775	12 (0 / 11 / 0 / 1)	1x Hypothetical protein
FR1R	SRAE_chr1:5724906-5766986	11 (0 / 0 / 0 / 11)	1x Nanchung 1x SMc04008-like domain-containing 1x 1,2-dihydroxy-3-keto-5-methylthiopentene dioxygenase 1x Transmembrane receptor 1x Predicted transmembrane / coiled-coil 2-containing 1x Band 7 protein family and Stomatin family-containing 1x Basic-leucine zipper domain-containing 1x TBC1 domain family member 13 1x BTB/POZ domain-containing
FR2L	SRAE_chr2:2435807-2464322	5 (0 / 0 / 0 / 5)	1x NHR/GATA-type domain-containing 2x MIP20649p 2x Hypothetical proteins
EC2	SRAE_chr2:2464323-2492838	12 (0 / 11 / 0 / 1)	1x UDP-glucosyltransferase
FR2R	SRAE_chr2:2492839-2521354	13 (0 / 1 / 0 / 12)	1x BTB/POZ domain-containing

			1x Protein kinase-like domain-containing 1x Ubiquitin-conjugating enzyme, E2 domain 1x WH2 domain-containing 1x Casein kinase II subunit beta 1x Serine/threonine-protein kinase haspin 1x PDZ domain-containing 4x Hypothetical protein
FR3L	SRAE_chr2:3843760-3879630	12 (0 / 0 / 0 / 12)	1x Transcription elongation factor SPT5 1x Armadillo-like helical domain -containing protein 1x Gamma-aminobutyric acid receptor subunit beta 1x Phosphodiesterase 1x Glycosyl transferase 1x Tyrosine-protein kinase 1x Bax inhibitor 1-related 1x Bloom syndrome protein 1x Protein lethal(2)essential for life 3x Hypothetical protein
EC3	SRAE_chr2:3879631-3915501	17 (0 / 16 / 0 / 1)	1x Hypothetical protein
FR3R	SRAE_chr2:3915502-3951372	17 (0 / 1 / 0 / 16)	1x Serine/threonine-protein phosphatase 1x Nematode cuticle collagen 1x Flavin-containing monooxygenase 1x Serine/threonine-protein kinase RIO1 12x Hypothetical protein
FR4L	SRAE_chr2:9057385-9101325	13 (0 / 0 / 0 / 13)	1x PR domain zinc finger protein 1 1x Sodium/potassium/calcium exchanger 1 1x Aspartate aminotransferase 1x Bifunctional coenzyme A synthase 1x Hormone-sensitive lipase 5x Nematode fatty acid retinoid binding 2x Hypothetical protein
EC4	SRAE_chr2:9101326-9145266	18 (18 / 0 / 0 / 0)	
FR4R	SRAE_chr2:9145267-9189207	17 (0 / 0 / 0 / 17)	1x 40S ribosomal protein S11 1x Glycosyl transferase, family 14-containing 1x Intraflagellar transport protein 140 homologue 1x K Homology domain-containing 1x Bifunctional heparan sulfate N-deacetylase 1x AT25266p 1x RNA-binding protein 42

			1x Vacuolar protein sorting-associated protein 33A 1x General transcription factor IIH subunit 3 1x Pre-mRNA cleavage complex 2 protein Pcf11 1x Smarcd3b 1x Methyltransferase-like protein 1x GH07323p 4x Hypothetical protein
FR5L	SRAE_chr2:10226062-10241215	5 (0 / 0 / 0 / 5)	1x IA-2 ortholog 1x Integrator complex subunit 9 1x UDP-glucuronosyltransferase 2x Hypothetical protein
EC5	SRAE_chr2:10241216-10256369	5 (5 / 0 / 0 / 0)	
FR5R	SRAE_chr2:10256370-10271523	1 (0/ 0 / 0 / 1)	1x Epidermal growth factor-like domain-containing
FR6L	SRAE_chr2:14189870-14197231	1 (0/ 0 / 0 / 1)	1x UDP-glucosyltransferase
EC6	SRAE_chr2:14197232-14204593	4 (4/ 0 / 0 / 0)	
FR6R	SRAE_chr2:14204594-14211955	3 (0/ 0 / 0 / 3)	1x Nematode fatty acid retinoid binding 2x Hypothetical protein
FR7L	SRAE_chr2:14267045-14284885	8 (0/ 0 / 0 / 8)	2x ATP-binding cassette sub-family D member 4 6x Hypothetical protein
EC7	SRAE_chr2:14284886-14302726	8 (7 / 0 / 0 / 1)	1x Hypothetical protein
FR7R	SRAE_chr2:14302727-14320567	8 (0/ 0 / 0 / 8)	1x Zinc metalloproteinase 2x Sulfotransferase 1x Estradiol 17-beta-dehydrogenase 12 1x Alpha-(1,3)-fucosyltransferase C 1x Glycosyl transferase, family 14-containing 2x Hypothetical protein
FR8L	SRAE_chr2:14320568-14332439	1 (0/ 0 / 0 / 1)	1x Cytochrome P450 4V2
EC8	SRAE_chr2:14332440-14348252	5 (1 / 1 / 2 / 1)	1x Hypothetical protein
FR8R	SRAE_chr2:14348253-14364065	5 (0/ 0 / 0 / 5)	1x Zona pellucida domain-containing 4x Hypothetical protein
FR9L	SRAE_chr2:15686271-15695575	1 (0/ 0 / 0 / 1)	1x Hypothetical protein
EC9	SRAE_chr2:15695576-15704880	5 (5 / 0 / 0 / 0)	
FR9R	SRAE_chr2:15704881-15714185	3 (0/ 0 / 0 / 3)	1x Cad96Cb 1x Zinc finger, RING/FYVE/PHD-type domain-containing 1x 1x Protein lethal(2)essential for life
FR10L	SRAE_chr2:16545359-16577777	9 (0/ 0 / 0 / 9)	1x Trypsin Inhibitor-like, cysteine rich domain-containing 1x Glycosyl transferase, family 14-containing 1x Transthyretin-like family-containing 2x 7TM GPCR, (Sre) family-containing 1x Nuclear hormone receptor 1x Proteinase inhibitor I25 1x Carboxylic ester hydrolase

			1x Hypothetical protein
EC10	SRAE_chr2:16577778-16610196	12 (12 / 0 / 0 / 0)	
FR10R	SRAE_chr2:16610197-16624307	4 (0/ 0 / 0 / 4)	1x CUB domain-containing 2x Metalloendopeptidase 1x Hypothetical protein
EC11	SRAE_chr2:16624308-16651910	8 (8 / 0 / 0 / 0)	
FR11R	SRAE_chr2:16651911-16679513	8 (0/ 0 / 0 / 8)	1x Prolyl endopeptidase 1x Zinc finger, RING-type domain-containing 6x Hypothetical protein
FR12L	SRAE_chr2:16679514-16689074	2 (0/ 0 / 0 / 2)	1x Importin-beta 1x Hypothetical protein
EC12	SRAE_chr2:16689075-16702613	7 (0 / 7 / 0 / 0)	
FR12R	SRAE_chr2:16702614-16716152	4 (0/ 0 / 0 / 4)	4x Hypothetical protein
FR13L	SRAE_chrX_scaffold1:2861643-2949331	17 (0 / 0 / 0 / 15)	1x Netrin-1a 1x RB1-inducible coiled-coil protein 1 1x GPCR, rhodopsin-like, 7TM domain-containing 1x ShKT domain 1x SH3 domain domain-containing 4x Poly-glutamine tract binding protein 1 1x Carboxylesterase, type B domain-containing 5x Hypothetical protein
EC13	SRAE_chrX_scaffold1:2949332-3037020	11 (0 / 0 / 7 / 4)	4x Poly-glutamine tract binding protein 1
FR13R	SRAE_chrX_scaffold1:3037021-3124709	16 (0 / 0 / 0 / 16)	1x Poly-glutamine tract binding protein 1 1x FI18412p1 1x Fibroblast growth factor receptor-like 1 1x Clc protein-like family-containing 1x Protein bicaudal D 1x DOMINA protein 1x FMRFamide-related peptide-like family-containing 1x Intraflagellar transport protein 122 homologue 1x Two pore domain potassium channel, TASK family 1x Innexin 1x Aquaporin-like domain-containing 5x Hypothetical protein
FR14L	SRAE_chrX_scaffold2:2014247-2029882	4 (0 / 0 / 0 / 4)	1x Ferric-chelate reductase 1 1x 1x GPCR, rhodopsin-like, 7TM domain-containing 2x Hypothetical protein
EC14	SRAE_chrX_scaffold2:2029883-2045518	6 (6 / 0 / 0 / 0)	
FR14R	SRAE_chrX_scaffold2:2045519-2061154	2 (0 / 0 / 0 / 2)	1x Amine oxidase domain-containing 1x Hypothetical protein
FR15L	SRAE_chrX_scaffold2:2160788-2192811	4 (0 / 0 / 0 /4)	1x Metalloendopeptidase 3x Hypothetical protein

EC15	SRAE_chrX_scaffold2:2192812-2224835	5 (4 / 0 / 1 / 0)	
FR15R	SRAE_chrX_scaffold2:2224836-2256859	4 (0 / 0 / 0 /4)	1x Metalloendopeptidase 3x Hypothetical protein

5.2.2.2 Assessment of genome assembly quality in expansion clusters

Expansion clusters contain multiple genes belonging to the same gene family, and therefore may contain repeats of highly similar sequence. This is potentially problematic for two reasons. First, as the reference genome was made entirely based on Illumina short read sequencing data (Hunt 2016), it is possible that repetitive regions are not fully resolved in the assembly, showing fewer gene copies than there are. Second, repetitive sequences complicate alignment of sample sequencing reads to the reference assembly by making mismatches more likely. Failure to account for these two difficulties could lead to misleadingly high diversity measurements. Thus, the quality of the reference assembly at expansion clusters was checked. To do this, sequencing reads originally used to build the reference assembly were aligned back to the reference. These reads are freely available at NCBI under BioProject code PRJEB2398, and the quality of these alignments were assessed with the software package Gap5 (Bonfield and Whitman 2010). Repetitiveness of sequence was examined by the generation of Dotplots with the software package Dotter (Sonnhammer and Durbin 1995) (Appendix 1). Alan Tracey (Sanger Institute) used Gap5 and Dotplot to produce plots and provided these as a Personal Communication. Gene annotation schematics were retrieved from Ensembl's 'Region in Detail' tool (Hubbard *et al.* 2002), accessed via WormBase Parasite (Howe *et al.* 2017) version 12 (https://parasite.wormbase.org/index.html) and added to the graphics produced by Gap5. Outputs of Gap5, Dotter and Ensemble were compiled and labelled by the candidate (Appendix 1).

The Gap5 plots show the mapping quality of reference reads, the distance between read mate pairs, whether mate pairs face the same direction and the read depth over each nucleotide. They also show per nucleotide fragment depth, which is the number of inserts between mate pairs as well as reads covering a base. Regions with poor mapping quality, unusually large distances between mate pairs and the occurrence of mate pairs facing opposite directions indicate high rates of misalignment. Peaks in read depth and fragment depth above background levels are evidence that multiple copies of a repetitive sequence are collapsed in the reference assemblies. Where expansion cluster genes or genes in flanking regions fell in poorly resolved reference assembly areas, these genes were excluded from further analysis.

5.2.3 Plotting of genetic variation along the genome

Whole genome sequencing data was used to detect single nucleotide polymorphisms (SNPs) among the sample set. Genetic variants were called and filtered as described in section 4.2.5 to produce a variant call format (VCF) file that listed every retained SNP by its position in the genome. All reference

assembly contigs over 10 kb in length were divided into non-overlapping 10 kb windows, and in each window the number of SNPs was counted and plotted. SNP counting in windows was performed with VCFtools version 0.1.12 (Danecek *et al.* 2011). Thus the distribution of SNPs within the genome was assessed. The same analysis was subsequently performed on each of nuclear clades 1 and 3, as defined in Chapter 4. Calculation of Pearson correlation coefficients between SNP numbers in genome windows in clade 1, clade 3, and the whole dataset were performed in R version 3.3.3 (R Core Team).

5.2.4 Analysis of variation in the *Strongyloides ratti* genome

5.2.4.1 Characterisation of variable and conserved genomic regions

As described in Section 4.3.3, SNP density across all scaffolds of the assembly and all *S. ratti* in the dataset was 4.1 SNPs per kb. Regions of the genome with SNP density much higher than this are putatively under diversifying selection, while those with much lower SNP density may be under purifying selection. To identify high-diversity regions, 10 kb windows described in Section 5.2.1.1 were ranked according the number of SNPs they contained. Sixty-three windows were found to contain 200 or more SNPs (i.e. had a SNP density of $\geq$ 20 SNPs per kb), and these were considered "highly variable". Conversely, 46 windows were found to have 4 SNPs or less, (a SNP density $\leq$ 0.4 per kb) and these were considered "highly conserved".

Highly variable and highly conserved windows were then characterised. Some genes extended beyond the edge of a window. These genes were included within the window if at least 200 bp of gene sequence fell within it. The gene density of highly variable and highly conserved windows was determined, and the frequency of gene description terms was compared among the highly variable windows, highly conserved windows, and the genome overall. Gene descriptions were retrieved from WormBase ParaSite. Genes within windows of interest were further characterised as free-living genes, parasitism genes or same according to the definitions provided in Section 5.1.2.1, where this information was available in Hunt *et al.* (2016). Genes that had been shown to fall on poor quality reference assembly by Gap5 analysis (Section 5.2.2.2) were not included. Finally, in highly variable windows only, SNPs occurring within windows of interest were characterised as synonymous, non-synonymous, or STOP-causing. To do this, VCFtools was used to extract a VCF file for each gene found within a window of interest and these VCF files were uploaded to the Ensembl Variant Effect Predictor, a service that compares SNP data to a reference assembly to determine their functional significance (McLaren *et al.* 2016). This procedure was not carried out on highly conserved windows due to the dearth of SNPs in these regions. When results are reported as percentages, confidence intervals given are always 95% (calculated at www.sample-size.net, accessed 02/01/19)

Furthermore, SNP alleles in the 15 most variable windows were investigated to determine whether they were present in nuclear clade 1, present in nuclear clade 3, both or neither. An allele that was found to

be present in one clade but not the other was considered private to that clade, while alleles appearing in both clades were considered shared.

5.2.4.2 Analysis of variation in parasitism genes and free-living genes

The 100 'most parasitic' (that is, with the greatest $\log_2$ fold increase in expression in the parasitic adult female, compared with the free-living adult female) protein-coding genes were retrieved from Hunt *et al.* (2016). Genes were subsequently excluded if they were part of an expansion cluster and the underlying assembly was considered poor (see Section 5.2.2) Each gene was characterised according to description and SNP density, and SNPs were classified as synonymous, non-synonymous or STOP-causing polymorphisms, as described above. The same procedure was carried out for the 100 'most free-living' genes, and the two classes of genes were compared. When results are reported as percentages, confidence intervals given are always 95% (calculated at www.sample-size.net, accessed 02/01/19).

5.2.5 Analysis of variation in expansion clusters

5.2.5.1. Clade-based analysis of expansion clusters

In Chapter 4, which interrogates the population genetics of *S. ratti* on a whole-genome basis, it is revealed that the population genetics of *S. ratti* can be characterised by a number of genetically distinct sympatric clades, and it is hypothesised that these clades represent ancient asexual lineages with rare sexual interbreeding. When seeking to understand selection processes acting upon a specific region of the genome, such as expansion clusters, it can be informative to compare population genetic parameters of these regions to the genome as a whole.

VCF files corresponding to each of the expansion clusters that had had no genes excluded due to poor underlying assembly were derived from the whole-genome VCF file via VCFtools, and these files were loaded into the software package TASSEL 5.0 (Bradbury *et al.* 2007). TASSEL was used to produce a neighbour joining tree for each expansion cluster based on SNP data. Neighbour joining trees were prepared for presentation in FigTree Version 1.4.3, a free software package available from http://tree.bio.ed.ac.uk/software/figtree/, (downloaded 04/10/2016). All 90 *S. ratti* individuals were included in this analysis

5.2.5.2 Variation in expansion clusters and flanking regions

A VCF file corresponding to each gene in an expansion cluster or flanking region that was retained after inspection of underlying genome reference quality (Section 5.2.2.2) was extracted from the whole-genome VCF file and uploaded to Ensembl's variant effect predictor as described for genes in highly variable and highly conserved regions. SNPs were marked as synonymous, non-synonymous or STOP-

causing, and comparisons were drawn between genes in expansion clusters and those in flanking regions. Genes that were within expansion clusters but that did not belong to one of the expanded gene families were excluded from these comparisons. When results are reported as percentages, confidence intervals are always 95% (calculated at www.sample-size.net, accessed 02/01/19).

To compare numbers of synonymous, non-synonymous and STOP-causing SNPs for Figure 5.4, values in SNPs of the indicated type per kb were calculated by summing each type of SNP across all genes within the expansion cluster or across both flanking regions associated with an expansion cluster, dividing this by the total length of coding sequence within those expansion clusters / flanking regions, and multiplying this by 1,000.

The ratio of non-synonymous SNP rate to synonymous SNP rate (dN/dS ratio) is a common parameter for measuring selection on genes / within genomes (Yang and Bielawski 2000). dN/dS analysis of genes in expansion clusters focused on individual *S. ratti* belonging to nuclear clades 1 and 3 (Figure 4.5A). These clades were chosen due to the high number of individuals they contain (46 and 15 respectively, compared with a maximum of 5 in other clusters). Due to the high degree of separation between nuclear clades 1 and 3 and their occurrence in sympatry in multiple sites, these clades were considered to be separate, non-interbreeding populations for the purposes of dN/dS analysis. dN/dS ratios are dependent on there being at least some variation in both clades 1 and clade 3, and thus ratios could not be calculated for every gene. To calculate dN/dS ratios, gene-specific VCF files for genes in expansion clusters and flanking regions were loaded into the software package DnaSP v6 (Rozas *et al.* 2017). Coding sequence and nuclear clades 1 and 3 were defined within the software and dN/dS calculations were performed.

5.3 Results

5.3.1 Quality of assembly underlying expansion clusters

Studies of selection based on regions of the genome that are not well-assembled may give misleading results. Gap5 plots showing the quality of the assembly underlying expansion clusters and their flanking regions (defined in Section 5.2.3.1) are shown in Appendix 1. Regions found to have peaks in read depth and fragment depth, have many reads with low mapping quality, and / or have many cases of mate pairs being unusually far apart and / or in the wrong orientation were considered to be of uncertain assembly quality, and genes in such regions were excluded from further analyses. Through this process, expansion clusters 4, 11 and 13 lost all genes, and they and their flanking regions were discarded from further analysis. Across all other expansion clusters and flanking regions a dataset of 197 genes remained, of which 63 were in expansion clusters and 134 were in flanking regions (Table 5.2). Clusters 6, 7, 8, 12 and 14 had no genes excluded. Three retained genes within expansion clusters did not belong to one of the three expansion gene families under considerations. These genes were ignored in comparisons of expansion cluster and flanking region genes. Of the remaining 60 expansion cluster genes, 29 encoded CAP domain-containing proteins, 29 encoded astacins, and 2 encoded acetylcholinesterases.

Table 5.2: Genes retained in expansion clusters (EC) and flanking regions (FR) after removal of those with poor underlying reference assembly. "Function" refers to the functional description given on WormBase ParaSite, with "n/a" recorded here as "hypothetical protein". SNP typess are expressed in the form synonymous ('S'), non-synonymous (NS) or STOP-causing ('STOP'). "Expression" denotes whether a gene is upregulated in the parasitic adult female morph ('parasitic'), the free-living adult female morph ('free-living), or both equally ('same'), with a difference in expression of base 2 log-fold difference of at least 1 being considered upregulation. A block of colour represents a given expansion cluster and its flanking regions.

Region	Gene	Function	Coding SNPs per kb	SNP types (S/NS/STOP)	Expression
FR1L	SRAE_1000180400	Transcriptional regulator ATRX	3.7	6 / 5 / 0	Parasitic
	SRAE_1000180500	Formin	2.8	4 / 5 / 0	Same
	SRAE_1000180600	Hypothetical protein	1.4	0 / 2 / 0	Same
	SRAE_1000180700	Glycosylphosphatidylinositol-mannosyltransferase I	1.7	0 / 1 / 0	Same
	SRAE_1000180800	Hypothetical protein	1.9	1 / 0 / 0	Same
	SRAE_1000180900	Rhodopsin-like	0.7	1 / 0 / 0	Same
	SRAE_1000181000	TPM domain-containing protein	1.4	0 / 1 / 0	Same
	SRAE_1000181100	MIF4-like	1.2	1 / 1 / 0	Same
	SRAE_1000181200	Sodium/potassium-transporting ATPase α subunit	0	0 / 0 / 0	Free-living
	SRAE_1000181300	Hypothetical protein	0	0 / 0 / 0	Free-living

	SRAE_1000181400	Mediator of RNA polymerase II transcription subunit 9	5.2	2 / 0 / 0	Same
	SRAE_1000181500	Serine/threonine-protein kinase Chk1	0.7	1 / 0 / 0	Same
	SRAE_1000181600	Hypothetical protein	1.1	1 / 0 / 0	Same
	SRAE_1000181700	Hypothetical protein	2.7	3 /1 / 0	Same
	SRAE_1000181800	Hypothetical protein	1.2	0 / 1 / 0	Free-living
	SRAE_1000181900	Eukaryotic translation initiation factor 2A	2.5	0 / 4 / 0	Same
	SRAE_1000182000	Hypothetical protein	0	0 / 0 / 0	Same
	SRAE_1000182100	Heme transporter HRG	0	0 / 0 / 0	Same
EC1	SRAE_1000182200	CAP domain-containing	1.1	0 / 1 / 0	Parasitic
	SRAE_1000182300	CAP domain-containing	0	0 / 0 / 0	Parasitic
	SRAE_1000182400	CAP domain-containing	1.2	1 / 0 / 0	Parasitic
	SRAE_1000183300	CAP domain-containing	3.4	2 / 1 / 0	Parasitic
FR1R	SRAE_1000183400	Nanchung	1.4	1 / 2 / 0	Same
	SRAE_1000183500	Hypothetical protein	0	0 / 0 / 0	Free-living
	SRAE_1000183600	SMc04008-like domain-containing	1.4	0 / 3 / 0	Same
	SRAE_1000183700	Hypothetical protein	0	0 / 0 / 0	Unlisted
	SRAE_1000183800	1,2-dihydroxy-3-keto-5-methylthiopentene dioxygenase	0	0 / 0 / 0	Same
	SRAE_1000183900	Transmembrane receptor	7.1	9 / 4 / 0	Parasitic
	SRAE_1000184000	Predicted transmembrane / coiled-coil 2-containing	0.7	0 / 1 / 0	Unlisted
	SRAE_1000184100	Band 7 protein family and Stomatin family-containing	3.2	3 / 0 / 0	Unlisted
	SRAE_1000184200	Basic-leucine zipper domain-containing	2.5	3 / 2 / 0	Unlisted
	SRAE_1000184300	TBC1 domain family member 13	1.6	2 / 0 / 0	Unlisted
	SRAE_1000184400	BTB/POZ domain-containing	0.7	1 / 0 / 0	Unlisted
FR2L	SRAE_2000075900	NHR/GATA-type domain-containing	2.5	4 /2 / 0	Same
	SRAE_2000076000	MIP20649p	3.2	6 / 3 / 0	Same
	SRAE_2000076100	MIP20649p	0.8	1 / 0 / 0	Unlisted
	SRAE_2000076200	Hypothetical protein	5.7	10 / 5 / 0	Free-living
	SRAE_2000076300	Hypothetical protein	2.4	2 / 3 / 0	Parasitic
EC2	SRAE_2000076400	CAP domain-containing	29.8	7 / 13 / 0	Parasitic
	SRAE_2000076600	CAP domain-containing	21.2	3 / 14 / 0	Parasitic
	SRAE_2000076700	CAP domain-containing	60.4	11 / 39 / 0	Parasitic
	SRAE_2000076800	CAP domain-containing	39.2	11 / 18 / 1	Parasitic
	SRAE_2000076900	CAP domain-containing	36.8	6 / 25 / 0	Parasitic
	SRAE_2000077000	CAP domain-containing	19	6 / 9 / 1	Parasitic
	SRAE_2000077100	CAP domain-containing	14	6 / 6 / 0	Parasitic
	SRAE_2000077200	UDP-glucosyltransferase	12.7	12 / 8 / 0	Parasitic
	SRAE_2000077300	CAP domain-containing	17.4	5 / 10 / 0	Parasitic
	SRAE_2000077400	CAP domain-containing	13	4 / 7 / 0	Parasitic
	SRAE_2000077500	CAP domain-containing	4.4	2 / 4 / 0	Same
FR2R	SRAE_2000077600	Hypothetical protein	10	8 / 14 / 0	Same
	SRAE_2000077700	BTB/POZ domain-containing	7.9	1 / 4 / 1	Unlisted
	SRAE_2000077800	Hypothetical protein	2.9	1 / 0 / 0	Unlisted
	SRAE_2000077900	Protein kinase-like domain-containing	2.1	1 / 2 / 0	Unlisted
	SRAE_2000078000	Ubiquitin-conjugating enzyme, E2 domain	4.5	2 / 0 / 0	Same
	SRAE_2000078100	WH2 domain-containing	5.9	3 / 0 / 0	Same
	SRAE_2000078200	Casein kinase II subunit beta	8.3	2 / 3 / 0	Same

	SRAE_2000078300	Hypothetical protein	0	0 / 0 / 0	Unlisted
	SRAE_2000078400	Hypothetical protein	0	0 / 0 / 0	Same
	SRAE_2000078500	Serine/threonine-protein kinase haspin	11.1	7 / 9 / 0	Same
	SRAE_2000078600	PDZ domain-containing	1.4	0 / 1 / 0	Unlisted
	SRAE_2000078700	CAP domain-containing	15.2	5 / 10 / 0	Parasitic
	SRAE_2000078800	Hypothetical protein	1.4	4 / 0 / 0	Same
FR3L	SRAE_2000123100	Transcription elongation factor SPT5	0.4	0 / 1 / 0	Same
	SRAE_2000123200	Armadillo-like helical domain -containing	1.3	3 / 3 / 0	Same
	SRAE_2000123300	Hypothetical protein	0	0 / 0 / 0	Unlisted
	SRAE_2000123400	Gamma-aminobutyric acid receptor subunit beta	2	3 / 0 / 0	Same
	SRAE_2000123500	Hypothetical protein	6.1	2 / 1 / 0	Unlisted
	SRAE_2000123600	Phosphodiesterase	0.9	2 / 0 / 0	Same
	SRAE_2000123700	Glycosyl transferase	3.4	1 / 4 / 0	Same
	SRAE_2000123800	Tyrosine-protein kinase	0	0 / 0 / 0	Unlisted
	SRAE_2000123900	Bax inhibitor 1-related family-containing	0	0 / 0 / 0	Unlisted
	SRAE_2000124000	Bloom syndrome protein	2	3 / 6 / 0	Free-living
	SRAE_2000124100	Protein lethal(2)essential for life	7.8	4 / 1 / 0	Parasitic
	SRAE_2000124200	Hypothetical protein	14.5	5 / 21 / 0	Parasitic
EC3	SRAE_2000124300	CAP domain-containing	69	17 / 49 / 4	Parasitic
	SRAE_2000124400	CAP domain-containing	43.3	8 / 30 / 1	Parasitic
	SRAE_2000124500	CAP domain-containing	50.9	11 / 33 / 2	Parasitic
	SRAE_2000124600	CAP domain-containing	8.7	3 / 5 / 0	Parasitic
	SRAE_2000124700	CAP domain-containing	5.2	0 / 4 / 1	Parasitic
	SRAE_2000124800	CAP domain-containing	6.5	4 / 2 / 0	Parasitic
	SRAE_2000124900	CAP domain-containing	6.6	2 / 4 / 0	Parasitic
FR3R	SRAE_2000126100	Hypothetical protein	8.6	6 /2 / 0	Parasitic
	SRAE_2000126200	Hypothetical protein	50	3 / 17 / 1	Unlisted
	SRAE_2000126290	Hypothetical protein	10.7	3 / 7 / 0	Unlisted
	SRAE_2000126300	Hypothetical protein	14.1	6 / 7 / 0	Parasitic
	SRAE_2000126400	Hypothetical protein	11.8	3 / 7 / 1	Parasitic
	SRAE_2000126500	Hypothetical protein	0	0 / 0 / 0	Parasitic
	SRAE_2000126600	CAP domain-containing	18.5	6 / 11 / 0	Parasitic
	SRAE_2000126700	Hypothetical protein	3.6	1 / 2 / 0	Same
	SRAE_2000126800	Serine/threonine-protein phosphatase	0.5	0 / 1 / 0	Unlisted
	SRAE_2000126900	Nematode cuticle collagen	2.1	1 / 1 / 0	Free-living
	SRAE_2000127000	Hypothetical protein	8.5	3 / 7 / 0	Unlisted
	SRAE_2000127100	Hypothetical protein	2.5	1 / 1 / 0	Unlisted
	SRAE_2000127200	Flavin-containing monooxygenase	2.9	2 / 2 / 0	Same
	SRAE_2000127300	Serine/threonine-protein kinase RIO1	2.6	3 / 1 / 0	Same
	SRAE_2000127400	Hypothetical protein	1.9	2 / 2 / 0	Same
FR5L	SRAE_2000325100	IA-2 ortholog	2	2 / 5 / 0	Same
	SRAE_2000325200	Hypothetical protein	4.7	0 / 1 / 0	Unlisted
	SRAE_2000325300	Integrator complex subunit 9	0.5	0 / 1 / 0	Same
	SRAE_2000325400	Hypothetical protein	0	0 / 0 / 0	Unlisted
	SRAE_2000325500	UDP-glucuronosyltransferase	3.8	2 / 4 / 0	Free-living
EC5	SRAE_2000325600	Astacin-like metalloendopeptidase	0.9	1 / 0 / 0	Parasitic
	SRAE_2000326000	Astacin-like metalloendopeptidase	7.6	3 / 7 / 0	Parasitic

FR5R	SRAE_2000326100	Epidermal growth factor-like domain-containing	0.6	1 / 1 / 0	Free-living
FR6L	SRAE_2000450300	UDP-glucosyltransferase	5.7	2 / 7 / 0	Parasitic
EC6	SRAE_2000450400	Astacin-like metalloendopeptidase	1.9	1 / 2 / 0	Parasitic
	SRAE_2000450500	Astacin-like metalloendopeptidase	4	1 / 4 / 0	Parasitic
	SRAE_2000450600	Astacin-like metalloendopeptidase	6.5	2 / 6 / 0	Parasitic
	SRAE_2000450700	Astacin-like metalloendopeptidase	22.3	7 / 25 / 0	Parasitic
FR6R	SRAE_2000450800	Nematode fatty acid retinoid binding	5.5	3 / 0 / 0	Unlisted
	SRAE_2000450900	Hypothetical protein	5.3	1 / 5 / 0	Unlisted
	SRAE_2000451000	Hypothetical protein	22.2	6 / 3 / 0	Parasitic
FR7L	SRAE_2000452400	Hypothetical protein	3.6	3 / 0 / 0	Unlisted
	SRAE_2000452500	ATP-binding cassette sub-family D member 4	2.3	3 / 1 / 0	Free-living
	SRAE_2000452600	ATP-binding cassette sub-family D member 4	1.1	1 / 1 / 0	Parasitic
	SRAE_2000452700	Hypothetical protein	1.7	1 / 2 / 0	Unlisted
	SRAE_2000452800	Hypothetical protein	1.8	3 / 0 / 0	Unlisted
	SRAE_2000452900	Hypothetical protein	0	0 / 0 / 0	Unlisted
	SRAE_2000453000	Hypothetical protein	0	0 / 0 / 0	Same
	SRAE_2000453100	Hypothetical protein	20.9	7 / 10 / 0	Parasitic
EC7	SRAE_2000453200	Astacin-like metalloendopeptidase	9.6	3 / 8 / 0	Parasitic
	SRAE_2000453300	Astacin-like metalloendopeptidase	2.6	0 / 3 / 0	Parasitic
	SRAE_2000453400	Sulfotransferase family-containing	0	0 / 0 / 0	Unlisted
	SRAE_2000453500	Astacin-like metalloendopeptidase	2.5	0 / 3 / 0	Parasitic
	SRAE_2000453600	Astacin-like metalloendopeptidase	0	0 / 0 / 0	Parasitic
	SRAE_2000453700	Astacin-like metalloendopeptidase	5.5	0 / 7 / 0	Parasitic
	SRAE_2000453800	Astacin-like metalloendopeptidase	34.8	15 / 26 / 0	Parasitic
	SRAE_2000453900	Astacin-like metalloendopeptidase	5.1	4 / 2 / 0	Parasitic
FR7R	SRAE_2000454000	Zinc metalloproteinase	3.4	3 / 1 / 0	Unlisted
	SRAE_2000454100	Hypothetical protein	3.7	7 / 4 / 0	Free-living
	SRAE_2000454200	Hypothetical protein	0	0 / 0 / 0	Same
	SRAE_2000454300	Sulfotransferase	4.2	4 / 0 / 0	Same
	SRAE_2000454400	Sulfotransferase	1.8	2 / 0 / 0	Same
	SRAE_2000454500	Estradiol 17-beta-dehydrogenase 12	0	0 / 0 / 0	Parasitic
	SRAE_2000454600	Alpha-(1,3)-fucosyltransferase C	3.1	0 /3 / 0	Unlisted
	SRAE_2000454700	Glycosyl transferase, family 14-containing	2.3	1 / 2 / 0	Same
FR8L	SRAE_2000454800	Cytochrome P450 4V2	1.3	1 / 1 / 0	Same
EC8	SRAE_2000454900	Acetylcholinesterase	1.2	2 / 0 / 0	Parasitic
	SRAE_2000455000	Astacin-like metalloendopeptidase	2.6	1 / 2 / 0	Parasitic
	SRAE_2000455100	Hypothetical protein	1.6	1 / 0 / 0	Parasitic
	SRAE_2000455200	CAP domain-containing	0	0 / 0 / 0	Parasitic

	SRAE_2000455300	Acetylcholinesterase	0	0 / 0 / 0	Parasitic
FR8R	SRAE_2000455400	Hypothetical protein	0	0 / 0 / 0	Unlisted
	SRAE_2000455500	Hypothetical protein	0	0 / 0 / 0	Same
	SRAE_2000455600	Zona pellucida domain-containing	0.9	0 / 1 / 0	Unlisted
	SRAE_2000455700	Hypothetical protein	0	0 / 0 / 0	Free-living
	SRAE_2000455800	Hypothetical protein	0	0 / 0 / 0	Unlisted
FR9L	SRAE_2000497000	Hypothetical protein	1.4	5 / 4 / 0	Free-living
EC9	SRAE_2000497100	Astacin-like metalloendopeptidase	5.2	1 / 5 / 0	Parasitic
FR9R	SRAE_2000497600	Cad96Cb	4.5	9 / 0 / 0	Parasitic
	SRAE_2000497700	Zinc finger, RING/FYVE/PHD-type domain-containing	1.5	1 / 0 / 0	Parasitic
	SRAE_2000497800	Protein lethal(2)essential for life	0	0 / 0 / 0	Parasitic
FR10L	SRAE_2000522800	Trypsin Inhibitor-like, cysteine rich domain-containing	22.7	10 / 5 / 0	Parasitic
	SRAE_2000522900	Glycosyl transferase, family 14-containing	4.3	2 / 4 / 0	Unlisted
	SRAE_2000523000	Transthyretin-like family-containing	9.1	3 / 1 / 0	Parasitic
	SRAE_2000523100	7TM GPCR, (Sre) family-containing	0	0 / 0 / 0	Unlisted
	SRAE_2000523200	7TM GPCR, (Sre) family-containing	4.3	0 / 2 / 0	Unlisted
	SRAE_2000523300	Hypothetical protein	1.4	0 / 1 / 0	Unlisted
	SRAE_2000523400	Nuclear hormone receptor	3.9	4 / 1 / 0	Same
	SRAE_2000523500	Proteinase inhibitor I25	0	0 / 0 / 0	Free-living
	SRAE_2000523600	Carboxylic ester hydrolase	3.4	2 / 4 / 0	Unlisted
EC10	SRAE_2000523700	Astacin-like metalloendopeptidase	5.2	1 / 5 / 0	Parasitic
	SRAE_2000523800	Astacin-like metalloendopeptidase	2.8	2 / 2 / 0	Parasitic
	SRAE_2000523900	Astacin-like metalloendopeptidase	26.1	8 / 29 / 1	Parasitic
	SRAE_2000524000	Astacin-like metalloendopeptidase	4	1 / 4 / 0	Parasitic
FR12L	SRAE_2000526900	Importin-beta	2.2	4 / 2 / 0	Same
	SRAE_2000527000	Hypothetical protein	14.3	8 / 17 / 0	Unlisted
EC12	SRAE_2000527100	CAP domain-containing	71.1	9 / 54 / 1	Parasitic
	SRAE_2000527200	CAP domain-containing	18.6	1 / 15 / 1	Parasitic
	SRAE_2000527300	CAP domain-containing	4.6	2 / 2 / 0	Parasitic
	SRAE_2000527400	CAP domain-containing	5.5	2 / 3 / 0	Unlisted
	SRAE_2000527500	CAP domain-containing	3.3	3 / 0 / 0	Parasitic
	SRAE_2000527600	CAP domain-containing	2.2	0 / 2 / 0	Parasitic
	SRAE_2000527700	CAP domain-containing	1.1	0 / 1 / 0	Parasitic
FR12R	SRAE_2000527800	Hypothetical protein	18.5	2 / 10 / 0	Unlisted
	SRAE_2000527900	Hypothetical protein	41.5	9 / 17 / 0	Unlisted
	SRAE_2000528000	Hypothetical protein	1.8	0 / 1 / 0	Unlisted
	SRAE_2000528100	Hypothetical protein	3	3 / 4 / 0	Unlisted
FR14L	SRAE_X000143400	Ferric-chelate reductase 1	10.1	5 / 3 / 0	Unlisted
	SRAE_X000143500	Hypothetical protein	7.1	7 / 9 / 0	Same
	SRAE_X000143600	1x GPCR, rhodopsin-like, 7TM domain-containing	0.7	1 / 0 / 0	Same
	SRAE_X000143700	Hypothetical protein	4.1	3 / 1 / 0	Same

EC14	SRAE_X000143800	Astacin-like metalloendopeptidase	15.5	8 / 9 / 1	Parasitic
	SRAE_X000143900	Astacin-like metalloendopeptidase	18.3	8 / 13 / 2	Parasitic
	SRAE_X000144000	Astacin-like metalloendopeptidase	7.2	4 / 5 / 0	Parasitic
	SRAE_X000144100	Astacin-like metalloendopeptidase	2.4	0 / 3 / 0	Parasitic
	SRAE_X000144200	Astacin-like metalloendopeptidase	39.5	24 / 24 / 1	Parasitic
	SRAE_X000144210	Astacin-like metalloendopeptidase	21.5	6 / 21 / 0	Parasitic
FR14R	SRAE_X000144220	Hypothetical protein	9.4	4 / 4 /0	Same
	SRAE_X000144300	Amine oxidase domain-containing	0.6	0 / 1 / 0	Free-living
EC15	SRAE_X000146800	Astacin-like metalloendopeptidase	29.3	16 / 21 / 0	Parasitic
	SRAE_X000146900	Astacin-like metalloendopeptidase	47.5	25 / 34 / 0	Free-living
FR15R	SRAE_X000146950	Hypothetical protein	3.2	1 / 0 / 0	Unlisted
	SRAE_X000147000	Hypothetical protein	5.1	0 / 1 / 0	Parasitic
	SRAE_X000147100	Hypothetical protein	6.3	2 / 1 / 0	Unlisted
	SRAE_X000147200	Hypothetical protein	26.9	15 / 22 / 0	Parasitic

5.3.2 Distribution of polymorphisms across the *Strongyloides ratti* genome

This section considers the entire *S. ratti* genome assembly. As described in chapter 4, a total of 170,666 SNPs were retained after filtering amongst the 90 individual worms analysed, resulting in a mean SNP density of 4.1 SNPs per kb. However, when the genome was separated into non-overlapping 10 kb windows and the SNPs within those regions counted, it was observed that some windows had notably more SNPs than the average (Figure 5.1). Indeed, while the average number of SNPs per window was 41, the standard deviation around this mean was high, at 42.5. Thus, SNPs were not distributed evenly across the genome, but rather concentrated in some regions.

Chapter 4 identified genetically distinct nuclear clades and sub-clades occurring in sympatry within the sampling sites tested. SNP distribution was tested within the largest of these clades; clades 1 and 3. Intriguingly, high-density regions were observed (i) within these clades, (ii) when the two clades were compared, and (iii) when each clade was compared to the entire dataset (Figure 5.1). When equivalent genome windows were considered, SNP density in each window was strongly correlated across the three groups (clade 1, clade 3, all individuals). For all individuals versus clade 1, all individuals versus clade 3, and clade 1 *vs*. clade 3, Pearson correlation correlations were $r = 0.93$ ($P < 0.0001$), 0.86, $P < 0.0001$) and 0.86 ($P < 0.0001$) respectively (degrees of freedom = 4343 in each case). Thus, windows

that are highly diverse when all individuals are assessed are also highly diverse when individuals of just one clade are assessed. High-diversity windows may be regions that are prone to accumulating mutations and have accumulated mutations independently in clades 1 and 3. Alternatively, standing genetic variation in the last common ancestor of these clades, or rare sexual interbreeding between them, may also explain the pattern observed.

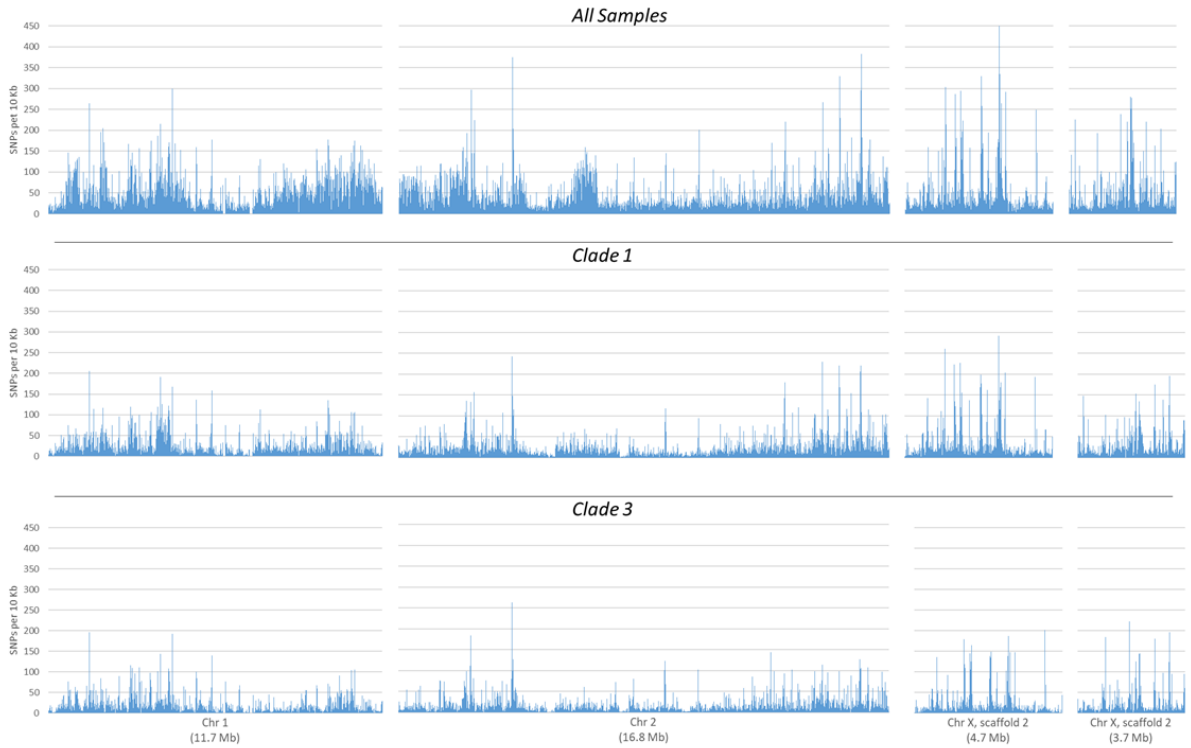


Figure 5.1 Density of SNPs across the S. ratti *genome, based on 90 individuals ("all samples"), or individuals from nuclear clades 1 (N = 46) and 3 (N = 15) (see Chapter 4) as indicated. The genome is divided into discrete 10 kb windows each represented by a vertical bar, the height of which indicates the number of SNPs within that window.*

5.3.3 Characterisation of highly variable genomic regions

When considering all individuals, 62 10 kb regions were found to contain 200 or more SNPs, with the highest value being 453 (Table 5.3). Collectively, these 62 regions included 172 genes with a mean of 2.79 genes per window (SD = 1.54). This is similar to the value of 2.89 arrived at by dividing the total number of genes across the genome by the total number of 10 kb windows, suggesting that highly variable genomic regions are not particularly gene dense or gene poor. Only two of these windows contained no genes at all.

Table 5.3: 10 kb windows found to contain 200 or more SNPs.

Region	Position	Genes	SNPs
1	SRAE_chrX_scaffold1:2960000-2970000	0	453
2	SRAE_scaffold4:220000-230000	1	434
3	SRAE_chr2:15790000-15800000	4	383
4	SRAE_scaffold4:200000-210000	1	364
5	SRAE_chr2:3880000-3890000	3	352
6	SRAE_chrX_scaffold6:40000-50000	6	346
7	SRAE_scaffold4:140000-150000	1	341
8	SRAE_chrX_scaffold1:2970000-2980000	2	335
9	SRAE_chr2:15060000-15070000	4	329
10	SRAE_chrX_scaffold1:2400000-2410000	3	329
11	SRAE_chrX_scaffold4:0-10000	0	315
12	SRAE_chrX_scaffold1:1270000-1280000	2	304
13	SRAE_chr1:4330000-4340000	5	300
14	SRAE_chr2:2470000-2480000	5	297
15	SRAE_chr2:15770000-15780000	2	295
16	SRAE_chrX_scaffold1:1750000-1760000	1	294
17	SRAE_chrX_scaffold1:3170000-3180000	2	292
18	SRAE_chrX_scaffold1:1570000-1580000	4	286
19	SRAE_chrX_scaffold1:2390000-2400000	2	283
20	SRAE_scaffold22:1-10000	2	282
21	SRAE_chrX_scaffold2:2150000-2160000	2	280
22	SRAE_chrX_scaffold3:210000-220000	2	279
23	SRAE_chrX_scaffold2:2180000-2190000	2	277
24	SRAE_scaffold20:1-10000	3	269
25	SRAE_chr2:14480000-14490000	3	267
26	SRAE_chrX_scaffold1:3030000-3040000	2	265
27	SRAE_chr1:1430000-1440000	4	264
28	SRAE_chrX_scaffold1:2420000-2430000	4	258
29	SRAE_scaffold4:170000-180000	4	258
30	SRAE_chrX_scaffold2:2170000-2180000	1	251
31	SRAE_scaffold3:310000-320000	2	251
32	SRAE_chrX_scaffold1:4120000-4130000	1	249
33	SRAE_chrX_scaffold3:690000-700000	4	247
34	SRAE_scaffold1:780000-790000	5	245
35	SRAE_scaffold1:200000-210000	5	241
36	SRAE_chrX_scaffold2:1810000-1820000	2	238
37	SRAE_chrX_scaffold6:140000-150000	6	283
38	SRAE_scaffold4:150000-160000	1	236
39	SRAE_scaffold1:210000-220000	3	234
40	SRAE_chr2:15070000-15080000	2	229
41	SRAE_chr2:15780000-15790000	3	226
42	SRAE_chrX_scaffold2:210000-220000	1	225
43	SRAE_chrX_scaffold1:1810000-1820000	4	223
44	SRAE_chrX_scaffold8:1-10000	1	223
45	SRAE_chrX_scaffold5:190000-200000	1	222
46	SRAE_chrX_scaffold2:2700000-2710000	2	221
47	SRAE_chr2:13190000-13200000	2	220
48	SRAE_chrX_scaffold5:130000-140000	3	218
49	SRAE_chrX_scaffold1:2980000-2990000	1	217
50	SRAE_chrX_scaffold4:60000-70000	3	216

51	SRAE_chrX_scaffold8:60000-70000	2	216
52	SRAE_chr1:3920000-3930000	3	215
53	SRAE_chrX_scaffold1:1280000-1290000	3	214
54	SRAE_chrX_scaffold2:2040000-2050000	3	211
55	SRAE_chrX_scaffold4:10000-20000	6	208
56	SRAE_chr1:1900000-1910000	4	205
57	SRAE_chrX_scaffold2:3210000-3220000	2	204
58	SRAE_chrX_scaffold4:30000-40000	2	203
59	SRAE_scaffold6:70000-80000	4	202
60	SRAE_chr2:10250000-10260000	2	201
61	SRAE_chrX_scaffold6:60000-70000	5	201
62	SRAE_chr2:3890000-3900000	6	200

Eleven of the 172 genes were found to lie in regions of the genome that had been excluded due to poor underlying assembly, and these were disregarded from further analyses. In the 161 genes that remained, mean coding SNP density was 26.4 SNPs per kb (SD = 22.6). Eighty (49.7%) of the genes were described as 'hypothetical protein' in WormBase ParaSite, indicating that a putative function has not been identified, while descriptions of the rest are provided in Table 5.4

Table 5.4: Genes within highly variable 10 kb windows as defined in Section 5.2.4.1. "Function" refers to the functional description given on WormBase ParaSite, with "n/a" recorded here as "hypothetical protein". "SNP types" details the absolute numbers of synonymous (S), nonsynonymous (NS) and STOP-causing SNPs codon. Expression refers to relative expression in different adult female morphs, such that "parasitic" and "free-living" refer to genes that are upregulated in the parasitic and free-living morphs respectively with a log_2 fold-change of at least 1, while genes that are not differentially expressed with a log_2 fold-change of at least 1 are listed as "same". Expression data comes from Hunt et al. (2016) and genes for which information is not available are "unlisted". "Expansion cluster" denotes which expansion cluster or associated flanking region (Table 5.2) a gene belongs to, if any. "Expression" denotes whether a gene is upregulated in the parasitic adult female morph ('parasitic'), the free-living adult female morph ('free-living), or both equally ('same'), with a difference in expression of base 2 log-fold difference of at least 1 being considered upregulation. Genes that are "discarded" had poor underlying assembly according to Gap5 analysis of expansion clusters and flanking regions (Appendix 1) and so were discounted from analyses. A block of colour represents a given 10 kb window.

Window	**Gene**	**Function**	**SNPs per coding kb**	**SNP types (S/NS/STOP)**	**Exp-ression**	**Expansion cluster**
2	SRAE_0000058700	Kinesin, motor domain and P-loop-containing	26.7	3 / 3 / 0	Unlisted	None

3	SRAE_2000499400	Hypothetical protein	48.2	5 / 28 / 0	Parasitic	None
	SRAE_2000499500	Hypothetical protein	0	0 / 0 / 0	Parasitic	None
	SRAE_2000499600	Hypothetical protein	72.2	8 / 20 / 1	Parasitic	None
	SRAE_2000499700	Hypothetical protein	45.7	5 / 22 / 0	Parasitic	None
4	SRAE_0000058500	Hypothetical protein	3.3	0 / 1 / 0	Unlisted	None
5	SRAE_2000124300	CAP domain-containing	69	17 / 49 / 4	Parasitic	EC3
	SRAE_2000124400	CAP domain-containing	43.3	8 / 30 / 1	Parasitic	EC3
	SRAE_2000124500	CAP domain-containing	50.9	11 / 33 / 2	Parasitic	EC3
6	SRAE_X000232600	Reverse transcriptase domain and Aspartic peptidase domain-containing	45.8	62 / 51 / 1	Unlisted	None
	SRAE_X000232700	I tegrase, catalytic core domain and Ribonuclease H-like domain-containing	22.2	5 / 1 / 1	Unlisted	None
	SRAE_X000232800	Hypothetical protein	32	6 / 6 / 0	Same	None
	SRAE_X000232900	Hypothetical protein	31..7	4 / 4 / 0	Unlisted	None
	SRAE_X000233000	Zinc finger, CCHC-type domain-containing	11.8	14 / 36 / 0	Unlisted	None
	SRAE_X000233100	Hypothetical protein	11.3	12 / 22 / 1	Unlisted	None
7	SRAE_0000057800	Hypothetical protein	28.5	1 / 9 / 0	Parasitic	None
8	SRAE_X000062200 **(discarded)**	Poly-glutamine tract binding protein 1	0	0 / 0 / 0	Unlisted	EC13
	SRAE_X000062300 **(discarded)**	Acetylcholinesterase	66.2	29 / 86 / 0	Same	EC13
9	SRAE_2000478200	Hypothetical protein	6.2	4 / 0 / 0	Unlisted	None
	SRAE_2000478300	Hypothetical protein	21.1	5 / 6 / 0	Same	None
	SRAE_2000478400	Hypothetical protein	90.7	10 / 29 / 1	Unlisted	None
	SRAE_2000478500	CAP domain-containing	89.3	17 / 47 / 0	Parasitic	None
10	SRAE_X000050900	Hypothetical protein	66.7	13 / 34 / 0	Parasitic	None
	SRAE_X000051000	Hypothetical protein	44.5	8 / 21 / 0	Parasitic	None
	SRAE_X000051100	Hypothetical protein	53	6 / 27 / 1	Parasitic	None
12	SRAE_X000026900	Hypothetical protein	51	5 / 19 / 0	Parasitic	None

	SRAE_X000027000	Hypothetical protein	15.8	3 / 4 / 0	Unlisted	None
13	SRAE_1000139800	Hypothetical protein	8	12 / 2 / 0	Unlisted	None
	SRAE_1000139900	Protein-tyrosine phosphatase-containing	54.6	42 / 145 / 1	Same	None
	SRAE_1000140000	Translation initiation factor SUI1 domain-containing	0	0 / 0 / 0	Same	None
	SRAE_1000140100	Hypothetical protein	0	0 / 0 / 0	Same	None
	SRAE_1000140200	Hypothetical protein	18.1	2 / 8 / 0	Same	None
14	SRAE_2000076600	CAP domain-containing	21.2	3 / 14 / 0	Parasitic	EC2
	SRAE_2000076700	CAP domain-containing	60.4	11 / 39 / 0	Parasitic	EC2
	SRAE_2000076800	CAP domain-containing	39.2	11 / 18 / 1	Parasitic	EC2
	SRAE_2000076900	CAP domain-containing	36.8	6 / 25 / 0	Parasitic	EC2
	SRAE_2000077000	CAP domain-containing	19	6 / 9 / 1	Parasitic	EC2
15	SRAE_2000499000	Hypothetical protein	55.9	2 / 25 / 1	Parasitic	None
	SRAE_2000499100	Hypothetical protein	35.8	7 / 12 / 0	Parasitic	None
16	SRAE_X000037500	Hypothetical protein	21.7	8 / 18 / 1	Unlisted	None
17	SRAE_X000066000	Trypsin Inhibitor-like	36.3	38 / 77 / 0	Parasitic	None
	SRAE_X000066100	Astacin-like metalloendopeptidase	28.2	18 / 23 / 0	Parsaitic	None
18	SRAE_X000032850	Hypothetical protein	69.2	4 / 18 / 0	Unlisted	None
	SRAE_X000032900	Hypothetical protein	54.6	8 / 11 / 0	Unlisted	None
	SRAE_X000033000	Hypothetical protein	78.6	9 / 20 / 0	Unlisted	None
	SRAE_X000033100	Hypothetical protein	46	4 / 12 / 0	Unlisted	None
19	SRAE_X000050700	Hypothetical protein	34.9	9 / 15 / 0	Parasitic	None
	SRAE_X000050800	Hypothetical protein	37	7 / 16 / 0	Parasitic	None
20	SRAE_0000074500	Aspartic peptidase domain-containing	46.4	93 / 116 / 6	Free-living	None
	SRAE_0000074600	Aspartic peptidase domain-containing	9.5	15 / 19 / 1	Unlisted	None
21	SRAE_X000145900	Astacin-like metalloendopeptidase	82.7	27 / 76 / 1	Same	None
	SRAE_X000146000	Astacin-like metalloendopeptidase	21.4	7 / 19 / 1	Para	None

22	SRAE_X000186500	Reverse transcriptase domain-containing	25.3	37 / 65 / 2	Free-living	None
	SRAE_X000186600	Cytochrome C oxidase subunit II	43.6	14 / 51 / 0	Same	None
23	SRAE_X000146300 **(discarded)**	Hypothetical protein	31.3	5 / 5 / 1	Unlisted	FR15L
	SRAE_X000146400 **(discarded)**	Hypothetical protein	6.4	2 / 1 / 0	Unlisted	FR15L
24	SRAE_0000074000	Hypothetical protein	16.7	3 / 6 / 1	Unlisted	None
	SRAE_0000074100	Hypothetical protein	10.4	2 / 6 / 0	Unlisted	None
	SRAE_0000074200	Aspartic peptidase domain-containing	42	86 / 74 / 0	Same	None
25	SRAE_2000460700	Hypothetical protein	52.4	20 / 112 / 1	Parasitic	None
	SRAE_2000460800	Hypothetical protein	4.6	1 / 1 / 0	Parasitic	None
	SRAE_2000460900	Hypothetical protein	6.7	1 / 3 / 1	Same	None
26	SRAE_X000063000	Acetylcholinesterase	15.2	9 / 18 / 0	Parasitic	EC13
	SRAE_X000063100	Poly-glutamine tract binding protein 1	0.4	0 / 1 / 0	Same	FR13R
27	SRAE_1000045700	E3 ubiquitin-protein ligase MYCBP2	4.9	51 / 20 / 0	Free-living	None
	SRAE_1000045800	Lipase, class 3 family-containing	15.2	7 / 8 / 0	Parasitic	None
	SRAE_1000045900	CAP domain-containing	76	19 / 67 / 0	Parasitic	None
	SRAE_1000046000	Speckle-type POZ	4.2	4 / 1 / 0	Same	None
28	SRAE_X000051500	CTP synthase 2	26.6	5 / 23 / 1	Unlisted	None
	SRAE_X000051600	CTP synthase 2	62.4	6 / 26 / 0	Unlisted	None
	SRAE_X000051700	Hypothetical protein	9.1	1 / 3 / 0	Parasitic	None
	SRAE_X000051800	CTP synthase 2	6,4	2 / 2 / 1	Unlisted	None
29	SRAE_0000058000	Reverse transcriptase domain-containing	40.7	19 / 46 / 0	Unlisted	None
	SRAE_0000058100	Integrase	38.9	18 / 21 / 3	Unlisted	None
	SRAE_0000058200	Hypothetical protein	26.5	3 / 3 / 1	Unlisted	None
	SRAE_0000058300	Reverse transcriptase domain-containing	3.3	2 / 8 / 0	Unlisted	None
30	SRAE_X000146200 **(discarded)**	Hypothetical protein	23.2	4 / 6 / 1	Unlisted	FR15L
31	SRAE_0000049800	Transposase, ISXO2-like domain-containing	24.9	2 / 6 / 2	Unlisted	None
	SRAE_0000049900	Integrase	9.6	4 / 20 / 2	Unlisted	None
32	SRAE_X000088600	Ras-like protein 3	0	0 / 0 / 0	Free-living	None

33	SRAE_X000195800	Hypothetical protein	0	0 / 0 / 0	Parasitic	None
	SRAE_X000195900	Hypothetical protein	30.7	2 / 6 / 0	Parasitic	None
	SRAE_X000196000	Hypothetical protein	12.4	1 / 6 / 0	Parasitic	None
	SRAE_X000196100	Hypothetical protein	31.1	2 / 4 / 1	Unlisted	None
34	SRAE_0000025500	Plasma membrane calcium-transporting ATPase 3	2.2	4 / 5 / 0	Free-living	None
	SRAE_0000025600	Transthyretin-like family-containing	103.9	11 / 37 / 0	Parasitic	None
	SRAE_0000025650	Hypothetical protein	26.3	2 / 9 / 1	Unlisted	None
	SRAE_0000025700	Transthyretin-like family-containing	0	0 / 0 / 0	Parasitic	None
	SRAE_0000025800	Protein argonaute-4	4.9	5 / 9 / 0	Parasitic	None
35	SRAE_0000007600	Amino acid transporter	2.5	4 / 0 / 0	Free-living	None
	SRAE_0000007700	Cathepsin L.1	13.4	9 / 4 / 0	Unlisted	None
	SRAE_0000007800	Cathepsin L.1	8.2	3 / 5 / 0	Free-living	None
	SRAE_0000007900	Hypothetical protein	29.5	14 / 20 / 0	Parasitic	None
	SRAE_0000008000	Hypothetical protein	37.5	10 / 32 / 1	Parasitic	None
36	SRAE_X000138600	Hypothetical protein	2.2	0 / 1 / 0	Same	None
	SRAE_X000138700	MAM domain and Concanavalin A-like lectin	2.3	4 / 0 / 0	Same	None
37	SRAE_X000236650	Hypothetical protein	24.4	3 / 18 / 1	Unlisted	None
	SRAE_X000236700	Ribonuclease H-like domain and AT hook-like family-containing	10	2 / 7 / 0	Unlisted	None
	SRAE_X000236800	Hypothetical protein	18.3	1 / 8 / 1	Unlisted	None
	SRAE_X000236900	Aspartic peptidase domain-containing	13.9	10 / 25 / 0	Unlisted	None
	SRAE_X000237000	Hypothetical protein	26.5	8 / 1 / 1	Unlisted	None
	SRAE_X000237100	Hypothetical protein	29.2	3 / 7 / 0	Unlisted	None
38	SRAE_0000057900	Hypothetical protein	42.9	3 / 10 / 0	Parasitic	None
39	SRAE_0000008000	Hypothetical protein	37.5	10 / 32 / 1	Parasitic	None
	SRAE_0000008100	Sulfotransferase family-containing	19.8	5 / 20 / 0	Parasitic	None
	SRAE_0000008200	Astacin-like metalloendopeptidase	4.3	4 / 1 / 0	Parasitic	None
40	SRAE_2000478600	CAP domain-containing	90	12 / 51 / 1	Parasitic	None

	SRAE_2000478610	Hypothetical protein	7.5	1 / 3 / 0	Unlisted	None
	SRAE_2000478700	Zinc finger, BED-type predicted domain-containing	8	8 / 8 / 0	Free-living	None
41	SRAE_2000499200	Hypothetical protein	75.7	8 / 27 / 2	Parasitic	None
	SRAE_2000499210	Hypothetical protein	40.7	5 / 17 / 0	Parasitic	None
	SRAE_2000499300	Hypothetical protein	15.5	2 / 6 / 0	Parasitic	None
42	SRAE_X000104500	Hypothetical protein	22.2	30 / 81 / 0	Parasitic	None
43	SRAE_X000038200	Hypothetical protein	25.2	8 / 5 / 0	Unlisted	None
	SRAE_X000038300	Synaptogyrin	3.1	2 / 1 / 0	Free-living	None
	SRAE_X000038400	EF-hand domain	2.9	2 / 0 / 0	Unlisted	None
	SRAE_X000038500	RUN domain-containing	4.9	4 / 5 / 0	Parasitic	None
44	SRAE_X000246300	Hypothetical protein	0	0 / 0 / 0	Same	None
45	SRAE_X000222600	Hypothetical protein	30.8	1 / 10 / 0	Parasitic	None
46	SRAE_X000158400	Astacin-like metalloendopeptidase	1.4	0 / 2 / 0	Free-living	None
	SRAE_X000158500	Acetylcholinesterase	8.7	1 / 14 / 0	Parasitic	None
47	SRAE_2000420700	CAP domain-containing	34.1	5 / 28 / 0	Parasitic	None
	SRAE_2000420800	Hypothetical protein	11.1	0 / 4 / 0	Unlisted	None
48	SRAE_X000221400	U1 small nuclear ribonucleoprotein 70 kDa	0	0 / 0 / 0	Same	None
	SRAE_X000221500	Hypothetical protein	5.8	0 / 2 / 0	Unlisted	None
	SRAE_X000221600	Hypothetical protein	46.4	7 / 9 / 0	Unlisted	None
49	SRAE_X000062400 **(discarded)**	Acetylcholinesterase	24.9	14 / 28 / 1	same	EC13
50	SRAE_X000201100	Astacin-like metalloendopeptidase	29.8	17 / 27 / 1	Parasitic	None
	SRAE_X000201200	Hypothetical protein	12	1 / 5 / 0	Unlisted	None
	SRAE_X000201300	Hypothetical protein	0	0 / 0 / 0	Unlisted	None
51	SRAE_X000247200	Carboxylesterase	6.1	1 / 13 / 0	Free-living	None
	SRAE_X000247300	ShKT domain-containing	41.8	28 / 100 / 0	Parasitic	None
52	SRAE_1000127500	Amino acid transporter	48.2	41 / 34 / 0	Same	None
	SRAE_1000127600	Farnesyltransferase, CAAX box, beta	14	34 / 16 / 0	Same	None

	SRAE_1000127700	Hypothetical protein	2.6	1 / 1 / 0	Unlisted	None
53	SRAE_X000027100	Hypothetical protein	16.4	2 / 6 / 0	Parasitic	None
	SRAE_X000027200	Hypothetical protein	38.3	13 / 13 / 1	Parasitic	None
	SRAE_X000027300	Hypothetical protein	8.2	2 / 2 / 0	Parasitic	None
54	SRAE_X000144200	Astacin-like metalloendopeptidase	39.5	24 / 24 / 1	Same	EC14
	SRAE_X000144210	Astacin-like metalloendopeptidase	21.5	6 / 21 / 0	Parasitic	EC14
	SRAE_X000144220	Hypothetical protein	9.4	4 / 4 / 0	Same	FR14R
55	SRAE_X000199600	Hypothetical protein	11.8	2 / 9 / 0	Unlisted	None
	SRAE_X000199700	Integrase	8.7	5 / 11 / 0	Unlisted	None
	SRAE_X000199800	Hypothetical protein	19.5	2 / 5 / 1	Unlisted	None
	SRAE_X000199900	Zinc finger, CCHC-type domain-containing	24.5	9 / 14 / 0	Unlisted	None
	SRAE_X000200000	Reverse transcriptase domain-containing	38.4	25 / 20 / 1	Unlisted	None
	SRAE_X000200100	Hypothetical protein	34.2	22 / 16 / 1	Unlisted	None
56	SRAE_1000059600	Ground-like domain-containing	22.8	22 / 50 / 0	Same	None
	SRAE_1000059700	Phosphate-regulating neutral endopeptidase	14.6	10 / 13 / 0	Parasitic	None
	SRAE_1000059800	Phosphate-regulating neutral endopeptidase	11.3	6 / 10 / 0	Parasitic	None
	SRAE_1000059900	Phosphate-regulating neutral endopeptidase	12.5	9 / 9 / 0	Same	None
57	SRAE_X000168300	Hypothetical protein	46.3	8 / 22 / 0	Parasitic	None
	SRAE_X000168400	Hypothetical protein	23.8	2 / 7 / 0	Parasitic	None
58	SRAE_X000200200	Hypothetical protein	3.2	0 / 1 / 0	Unlisted	None
	SRAE_X000200300	Hypothetical protein	2	1 / 0 / 0	Parasitic	None
59	SRAE_0000065600	Hypothetical protein	69.8	28 / 30 / 3	Unlisted	None
	SRAE_0000065700	Hypothetical protein	24.8	3 / 6 / 0	Unlisted	None
	SRAE_0000065800	Hypothetical protein	12.6	1 / 3 / 0	Unlisted	None
	SRAE_0000065900	Ribonuclease H-like domain-containing protein	6.4	2 / 2 / 0	Unlisted	None

60	SRAE_2000325900 **(discarded**	Astacin-like metalloendopeptidase	25.3	13 / 20 / 0	Parasitic	EC5
	SRAE_2000326000	Astacin-like metalloendopeptidase	7.6	3 / 7 / 0	Parasitic	EC5
61	SRAE_X000233200	Hypothetical protein	11.8	6 / 14 / 1	Unlisted	None
	SRAE_X000233300	Hypothetical protein	11.3	2 / 3 / 0	Unlisted	None
	SRAE_X000233400	Hypothetical protein	23.8	3 / 14 / 0	Unlisted	None
	SRAE_X000233500	Reverse transcriptase domain-containing	23.7	10 / 31 / 4	Unlisted	None
	SRAE_X000233600	Hypothetical protein	30.5	9 / 37 / 1	Unlisted	None
62	SRAE_2000124600	CAP domain-containing	8.7	3 / 5 / 0	Parasitic	EC3
	SRAE_2000124700	CAP domain-containing	5.2	0 / 4 / 1	Parasitic	EC3
	SRAE_2000124800	CAP domain-containing	6.5	4 / 2 / 0	Parasitic	EC3
	SRAE_2000124900	CAP domain-containing	6.6	2 / 4 / 0	Parasitic	EC3
	SRAE_2000125000 **(discarded)**	CAP domain-containing	2.2	1 / 1 / 0	Parasitic	EC3
	SRAE_2000125100 **(discarded)**	CAP domain-containing	38.6	9 / 25 / 2	Parasitic	EC3

A notable finding was the high abundance of genes encoding astacin-like metalloendopeptidases (hereafter, astacins) and CAP domain-containing proteins. When all genes in variable regions were taken together (including those described as encoding hypothetical proteins) 5.6% (95% CI 2.6-10.3) and 11.8% (95% CI 6.3-16.4) of genes encoded astacins and CAP domain-containing proteins respectively. Across the whole genome, astacin and CAP-domain containing protein genes make up 1.5% and 0.7% of genes respectively, showing that genes belonging to these families are over-represented in highly variable genomic windows. Coding SNP density was 26.3 SNPs per kb for genes encoding astacins and 36.7 SNPs per kb for genes encoding CAP domain-containing. When genes encoding astacins and CAP-domain containing proteins were removed from the dataset of genes within highly variable windows, coding SNP density within the remaining genes reduced from 26.4 SNPs per kb to 25 SNPs per kb. Thus, the high variability of windows under study was not wholly due to genes encoding astacins or CAP domain-containing proteins, as other genes were highly variable too.

The astacin and CAP domain-containing protein families are known to be expanded in *Strongyloides* and putatively have a role in the parasitic lifestyle (Hunt *et al.*2016), and the finding that these genes are concentrated in highly variable regions of the genome may be relevant to understanding ongoing selection of parasitism-related traits. However, alternative interpretations might be that the high copy

number of genes in these families promotes misalignment of sequencing reads, resulting in artificially high SNP counts, or that individual copies within these expanded gene families have become functionally redundant and are accumulating SNPs due to relaxed selection pressure. In either of these cases, high numbers of apparent STOP-causing SNPs would be predicted to be observed. To test this, the density of coding STOP-causing SNPs was examined in genes in highly variable regions (Table 5.4, Figure 5.2), compared with the rest of the genome. Among the 9 astacin-encoding and 18 CAP domain-containing protein-encoding genes within highly variable windows, the density of STOP-causing SNPs was 0.21 and 0.76 per kb respectively. In the 125 other genes in highly variable regions (i.e. excluding astacins and CAP-domain proteins) this was 0.48 per kb. Thus, there was no substantial difference in density of STOP-causing SNPs between the three groups.

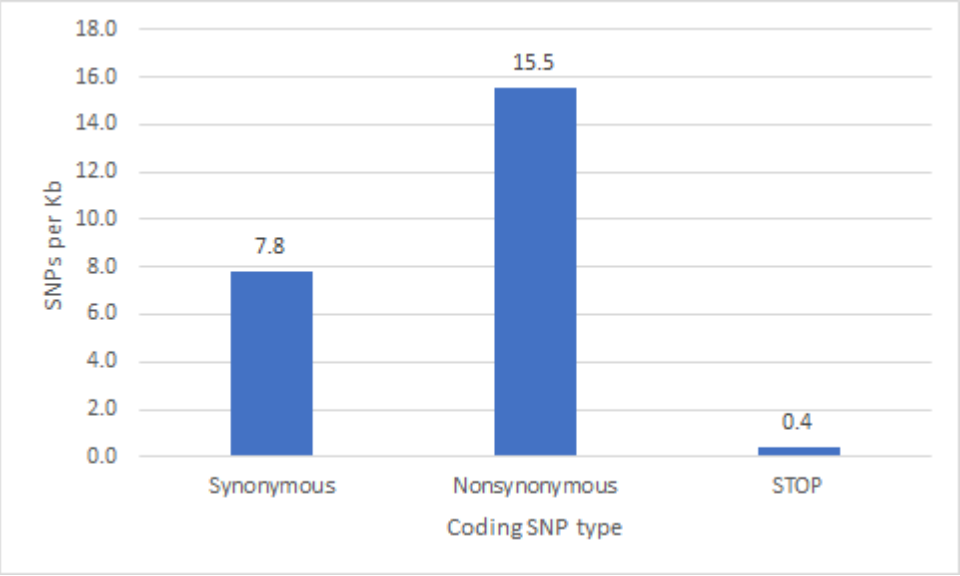


Figure 5.2: Density of synonymous, nonsynonymous and STOP-causing ('STOP') SNPs across coding sequences in the 161 most diverse 10 kb windows of the S. ratti *genome for 90 individuals.*

Across all highly variable window genes, the rate of nonsynonymous SNPs was approximately twice that of synonymous SNPs (Figure 5.2). Per gene, the mean ratio of nonsynonymous over synonymous SNPs (excluding the 19 genes with no synonymous SNPS) was 2.6 (SD =2.3). Thus, there was a somewhat consistent trend of genes within highly variable regions having more nonsynonymous SNPs than synonymous SNPs.

Genes within highly variable regions were classified according to whether they were more highly expressed in the parasitic adult female morph ('parasitism genes'), the free-living adult female morph ('free-living genes'), or same (Table 5.4), as described in Section 5.2.4.1 Results are shown in Table

5.5. Thus, parasitism genes were substantially over-represented in the most variable regions of the genome. When genes encoding astacins or CAP-domain containing genes were removed from the dataset, 55.8% (95% CI 45.4-68.4) of the genes that remained were parasitism genes (excluding genes for which stage-specific information was not available). Thus, even without genes in the two major expanded gene families, parasitism genes are heavily over-represented in the most abundant regions of the genome.

Table 5.5: Percentage of parasitism genes or free-living genes (see Section 5.2.4.1 for definitions) in 10 kb windows with 200 or more SNPs ('highly variable'), 10 kb windows with 4 or less SNPs ('highly conserved') or across the whole genome. Only genes for which stage-specific information was available were considered for the denominator. Stage-specific information originated from Hunt et al. (2016). All confidence intervals are at 95%.

	Parasitism genes (%)	**Free-living genes (%)**
Highly variable	69 (CI 59-77.9)	11 (CI 5.6-18.8)
Highly conserved	3.7 (CI 1.2-8.4)	11.9 (CI 6.9-18.5)
Whole genome	8.8 (CI 8.2-9.3)	14.2 (CI 13.5-14.9)

The number of alleles private to clade 1 and to clade 3 in each of the 15 most variable genome windows are shown in Table 5.6. Comparison of the means shows that 41.1% alleles are shared by the clades, 44.3% are private to clade 1 and 16.6% private to clade 3, but this was highly variable across the most variable windows (SDs = 23.9%, 31.6% and 15.8% respectively). This suggests that different variable windows have acquired diversity at different times. Windows where most alleles are shared likely acquired diversity prior to the divergence of clades 1 and 3, while clades where most alleles are unique to one clade or the other likely acquired diversity subsequent to clade divergence.

Table 5.6: The 15 10 kb windows (numbered as in Table 5.4) containing the most SNPs, showing the number of alleles unique to the indicated nuclear genetic clades (Chapter 4), or shared between them.

Window	**SNPs**	**Clade 1 unique alleles**	**Clade 3 unique alleles**	**Clades 1 and 3 shared alleles**
1	453	187	34	105
2	434	157	42	153
3	383	131	14	90
4	364	59	43	190
5	352	62	82	181
6	346	42	85	189
7	341	123	38	74
8	335	73	138	49
9	329	135	14	85
10	329	66	3	132
11	315	57	15	206

12	304	258	0	2
13	300	146	47	22
14	297	47	97	87
15	295	87	8	119

5.3.4 Characterisation of highly conserved genomic regions

When considering all individuals, 46 10 kb windows were found to contain 4 or fewer SNPs, of which 4 had none (Table 5.7). Collectively these windows included 146 genes, with a mean of 2.89 per window. Thus, gene density in the most conserved regions of the genome closely aligned with that of the genome-wide value, 2.89 per 10 kb. Two windows which collectively contained 5 SNPs, contained no genes at all. None of the genes within highly conserved regions were also within expansion clusters, and therefore none were excluded due to uncertainty of the underlying assembly.

Table 5.7: 10 kb windows found to contain 4 or less SNPs.

Window	**Position**	**Genes**	**SNPs**
1	SRAE_chr1:5100000-5110000	6	0
2	SRAE_chr1:9270000-9280000	1	0
3	SRAE_chr2:15750000-15760000	1	0
4	SRAE_chrX_scaffold2:1680000-1690000	3	0
5	SRAE_chr1:160000-170000	1	1
6	SRAE_chr1:5920000-5930000	1	1
7	SRAE_chr1:5970000-5980000	4	2
8	SRAE_chr1:6340000-6350000	5	2
9	SRAE_chr1:8760000-8770000	0	2
10	SRAE_chr1:8810000-8820000	5	2
11	SRAE_chr2:2630000-2640000	4	2
12	SRAE_chr2:2780000-2790000	5	2
13	SRAE_chr2:9460000-9470000	7	2
14	SRAE_scaffold2:150000-160000	3	2
15	SRAE_scaffold2:360000-370000	7	2
16	SRAE_chr1:150000-160000	5	3
17	SRAE_chr1:5500000-5510000	6	3
18	SRAE_chr1:6000000-6010000	0	3
19	SRAE_chr1:6570000-6580000	3	3
20	SRAE_chr1:6880000-6890000	1	3
21	SRAE_chr1:7740000-7750000	2	3
22	SRAE_chr2:4630000-4640000	3	3
23	SRAE_chr2:4920000-4930000	1	3
24	SRAE_chrX_scaffold2:520000-530000	2	3
25	SRAE_chrX_scaffold2:750000-760000	4	3
26	SRAE_chrX_scaffold2:2820000-2830000	2	3
27	SRAE_chr1:780000-790000	4	4
28	SRAE_chr1:5140000-5150000	5	4
29	SRAE_chr1:5950000-5960000	3	4
30	SRAE_chr1:8830000-8840000	3	4
31	SRAE_chr1:11420000-11430000	4	4

32	SRAE_chr1:11570000-11580000	1	4
33	SRAE_chr1:11580000-11590000	2	4
34	SRAE_chr2:4710000-4720000	3	4
35	SRAE_chr2:5260000-5270000	1	4
36	SRAE_chr2:7560000-7570000	6	4
37	SRAE_chr2:7960000-7970000	3	4
38	SRAE_chrX_scaffold1:3490000-3500000	2	4
39	SRAE_chrX_scaffold1:3630000-3640000	4	4
40	SRAE_chrX_scaffold1:3820000-3830000	2	4
41	SRAE_chrX_scaffold2:3260000-3270000	1	4
42	SRAE_chrX_scaffold2:3580000-3590000	6	4
43	SRAE_chrX_scaffold3:350000-360000	4	4
44	SRAE_scaffold1:680000-690000	3	4
45	SRAE_scaffold2:60000-70000	2	4
46	SRAE_scaffold2:100000-110000	4	4

Of the 146 genes in highly conserved regions, 21 (14.4%, 95% CI 9.1-21.1) are described as 'hypothetical proteins'. Descriptions of the rest are provided in Table 5.8. No gene descriptions were obviously over-represented among these genes. Rather many of the genes were 'housekeeping' genes, involved translation initiation, cell signalling and DNA repair *etc.* Mutations in these genes are likely to be highly deleterious, hence there may be a selective pressure to protect these regions from mutation. Alternatively, the presence of such critical housekeeping genes within a 10 kb window may mean that mutations in these windows are usually quickly purged by selection, keeping diversity low. There were no genes coding for astacin or CAP domain-containing protein families in these regions.

Table 5.8: Genes within highly conserved 10 kb windows, as defined in Section 5.2.4.1. "Function" refers to the functional description given on WormBase ParaSite, with "n/a" recorded here as "hypothetical protein". Expression refers to relative expression in different adult female morphs, such that "parasitic" and "free-living" refer to genes that are upregulated in the parasitic and free-living morphs respectively with a log_2 fold-change of at least 1, while genes that are not differentially expressed with a log_2 fold-change of at least 1 are listed as "same". Expression data comes from Hunt et al. (2016) and genes for which information is not available are "unlisted".

Window	**Gene**	**Function**	**Expression**
1	SRAE_1000164700	Protein lethal(2)essential for life	Free-living
	SRAE_1000164800	Protein TLP-1	Free-living
	SRAE_1000164900	Signal-induced proliferation-associated 1-like protein 1	Free-living
	SRAE_1000165000	Spectrin alpha chain, erythrocytic 1	Free-living
	SRAE_1000165100	V-type ATPase	Free-living

	SRAE_1000165200	Pleckstrin homology-like domain-containing	Free-living
2	SRAE_1000288400	GPCR, rhodopsin-like, 7TM domain-containing	Same
3	SRAE_2000498900	Probable Golgi transport protein 1	Same
4	SRAE_X000135600	Protein TAG-68	Same
	SRAE_X000135700	Protein ZIP-9	Same
	SRAE_X000135800	Protein-tyrosine phosphatase	Same
5	SRAE_1000005900	Liprin-alpha	Free-living
6	SRAE_1000189300	Acyltransferase 3 domain	Same
7	SRAE_1000190600	Bifunctional purine biosynthesis protein PURH	Same
	SRAE_1000190700	Bifunctional purine biosynthesis protein PURH	Same
	SRAE_1000190800	Calpain-15	Same
	SRAE_1000190900	Carboxypeptidase N catalytic chain	
8	SRAE_1000198900	Collagen triple helix repeat-containing protein	Same
	SRAE_1000199000	Cyclin-dependent kinase 5	Same
	SRAE_1000199100	Cyclin-dependent kinase 5	Same
	SRAE_1000199200	Cytochrome P450 18a1	Same
	SRAE_1000199300	Domain of unknown function DB domain-containing	Same
10	SRAE_1000274800	F-box domain-containing	Same
	SRAE_1000274900	FI03683p	Same
	SRAE_1000275000	Fimbrin	Same
	SRAE_1000275100	Rhodopsin-like	Same
	SRAE_1000275200	Germinal-center associated nuclear protein	Same
11	SRAE_2000082300	Hypothetical protein	Same
	SRAE_2000082400	Hypothetical protein	Same
	SRAE_2000082500	Hypothetical protein	Same
	SRAE_2000082600	Hypothetical protein	Same
12	SRAE_2000087200	Hypothetical protein	Same
	SRAE_2000087300	Hypothetical protein	Same
	SRAE_2000087400	Hypothetical protein	Same
	SRAE_2000087500	Hypothetical protein	Same
	SRAE_2000087600	Hypothetical protein	Same
13	SRAE_2000300900	Phosphatidylinositol N-acetylglucosaminyltransferase subunit C	Same
	SRAE_2000301000	Polynucleotide 5'-hydroxyl-kinase NOL9	Same

	SRAE_2000301100	Potassium channel family	Same
	SRAE_2000301200	Potassium channel subfamily K member 18	Same
	SRAE_2000301300	Prefoldin alpha-like domain; Prefoldin domain-containing	Same
	SRAE_2000301400	Probable diacylglycerol kinase 3	Same
	SRAE_2000301500	Probable gamma-butyrobetaine dioxygenase	Same
14	SRAE_0000031900	Hypothetical protein	Unlisted
	SRAE_0000032000	Hypothetical protein	Unlisted
	SRAE_0000032100	Nuclear hormone receptor HR96	Same
15	SRAE_0000038300	Hypothetical protein	Unlisted
	SRAE_0000038400	Hypothetical protein	Unlisted
	SRAE_0000038500	Hypothetical protein	Unlisted
	SRAE_0000038600	N-acetyllactosaminide beta-1,3-N-acetylglucosaminyltransferass	Unlisted
	SRAE_0000038700	N-acylethanolamine-hydrolyzing acid amidase	Unlisted
	SRAE_0000038800	Transposase, ISXO2-like domain-containing protein	Unlisted
	SRAE_0000038900	Sigma non-opioid intracellular receptor 1	Unlisted
16	SRAE_1000005450	ATP-binding cassette sub-family D member 2	Free-living
	SRAE_1000005500	Condensin-2 complex subunit D3	Free-living
	SRAE_1000005600	Glycoside hydrolase, family 1	Free-living
	SRAE_1000005700	Hypothetical protein	Free-living
	SRAE_1000005800	Isocitrate dehydrogenase [NAD] subunit gamma, mitochondrial	Free-living
17	SRAE_1000175900	3-hydroxy-3-methylglutaryl-coenzyme A reductase	Same
	SRAE_1000176000	4-hydroxyphenylpyruvate dioxygenase	Same
	SRAE_1000176100	60S ribosomal protein L29	Same
	SRAE_1000176200	serpentine receptor class v (Srv) family-containing	Same
	SRAE_1000176300	Actin-interacting protein 1	Same
	SRAE_1000176400	Activating signal cointegrator 1	Same
19	SRAE_1000207000	E3 ubiquitin-protein ligase ubr-1	Same
	SRAE_1000207100	Eukaryotic translation initiation factor 2 subunit 3	Same
	SRAE_1000207200	Eukaryotic translation initiation factor 3 subunit D	Same
20	SRAE_1000217700	Exocyst complex component 8	Same

21	SRAE_1000239700	Exostosin-2	Same
	SRAE_1000239800	Failed axon connections homolog	Same
22	SRAE_2000148700	Hypothetical protein	Same
	SRAE_2000148800	Hypothetical protein	Same
	SRAE_2000148900	Inositol polyphosphate-related phosphatase domain	Same
23	SRAE_2000156400	LisH dimerisation motif domain-containing	Same
24	SRAE_X000111300	Protein NIPI-4	Same
	SRAE_X000111400	Protein of unknown function DUF1757 family-containing	Same
25	SRAE_X000116500	Protein RFT1 homolog	Same
	SRAE_X000116600	Protein SEB-3	Same
	SRAE_X000116700	Protein SER-3	Same
	SRAE_X000116800	Protein SURO-1	Same
26	SRAE_X000161000	Protein-tyrosine phosphatase	Same
	SRAE_X000161100	Ras-related protein Rab-30	Same
27	SRAE_1000025200	Na[+]-driven anion exchanger 1	Free-living
	SRAE_1000025300	Nuclear anchorage protein 1	Free-living
	SRAE_1000025400	PMP-22/EMP/MP20/Claudin superfamily-containing	Free-living
	SRAE_1000025500	Protein GLIT-1	Free-living
28	SRAE_1000166000	Formin-homology and zinc finger domains protein 1	Parasitic
	SRAE_1000166100	Prolyl endopeptidase	Parasitic
	SRAE_1000166200	Protein TAG-314	Parasitic
	SRAE_1000166300	Proteinase inhibitor I25, cystatin domain-containing	Parasitic
	SRAE_1000166400	Sigma non-opioid intracellular receptor 1	Parasitic
29	SRAE_1000189900	AP-50	Same
	SRAE_1000190000	Armadillo-type fold domain-containing	Same
	SRAE_1000190100	Basic-leucine zipper domain-containing	Same
30	SRAE_1000275700	Germinal-center associated nuclear protein	Same
	SRAE_1000275800	Glycine cleavage system H protein, mitochondrial	Same
	SRAE_1000275900	Glycoside hydrolase, family 2	Same
31	SRAE_1000350300	GPI inositol-deacylase	Same

	SRAE_1000350400	Guanine nucleotide-binding protein G(k) subunit alpha	Same
	SRAE_1000350500	HAD hydrolase, subfamily IA; HAD-like domain-containing	Same
	SRAE_1000350600	Helicase	Same
32	SRAE_1000354800	Hypothetical protein	Same
33	SRAE_1000354900	Hypothetical protein	Same
	SRAE_1000355000	Hypothetical protein	Same
34	SRAE_2000150800	Kalirin	Same
	SRAE_2000150900	KAT8 regulatory NSL complex subunit 3	Same
	SRAE_2000151000	Lateral signaling target protein 2	Same
35	SRAE_2000165800	Lysocardiolipin acyltransferase 1	Same
36	SRAE_2000242000	MFS-type transporter SLC18B1	Same
	SRAE_2000242100	Mitochondrial import inner membrane translocase subunit tim-16	Same
	SRAE_2000242200	Myotubularin-like phosphatase domain-containing	Same
	SRAE_2000242300	Na[+]-driven anion exchanger 1	Same
	SRAE_2000242400	N-acetyl-D-glucosamine kinase	Same
	SRAE_2000242500	Nicotinate phosphoribosyltransferase	Same
37	SRAE_2000254600	Novel protein similar to vertebrate importin 8	Same
	SRAE_2000254700	Peptidase S28 family-containing	Same
	SRAE_2000254800	Peptidase S28 family-containing	Same
38	SRAE_X000073700	Probable mitochondrial pyruvate carrier 2	Same
	SRAE_X000073800	Protein ABCF-1	Same
39	SRAE_X000077100	Protein CAH-6	Same
	SRAE_X000077200	Protein CEBP-2	Same
	SRAE_X000077300	Protein GLRX-22	Same
	SRAE_X000077400	Protein ILYS-4	Same
40	SRAE_X000081900	Protein LPD-3	Same
	SRAE_X000082000	Protein MRPL-9	Same
41	SRAE_X000168900	Receptor-type tyrosine-protein phosphatase delta	Same
42	SRAE_X000176900	Regulator of G-protein signaling 3	Same
	SRAE_X000177000	Integrase	Same
	SRAE_X000177100	Integrase	Same
	SRAE_X000177200	Integrase	Same

	SRAE_X000177300	Ribosome production factor 1	Same
	SRAE_X000177400	RNA polymerase II-associated factor 1 homolog	Same
43	SRAE_X000189100	RNA polymerase-associated protein RTF1 homolog	Same
	SRAE_X000189200	Serine/threonine-protein kinase kin-29	Same
	SRAE_X000189300	Similar to Xab2 protein	Same
	SRAE_X000189400	Six-bladed beta-propeller, TolB-like domain	Same
44	SRAE_0000022700	Small GTPase superfamily	Same
	SRAE_0000022800	snRNA-activating protein complex subunit 3	Same
	SRAE_0000022900	Transient receptor potential cation channel subfamily M member 1	Same
45	SRAE_0000029800	Transmembrane protein 144	Same
	SRAE_0000029900	Ufm1-specific protease 2	Same
46	SRAE_0000030800	Vacuolar protein sorting-associated protein 13D	Same
	SRAE_0000030900	Vacuolar protein sorting-associated protein 13D	Same
	SRAE_0000030950	Homeobox domain and Homeodomain-like-containing protein	Unlisted
	SRAE_0000031000	Hypothetical protein	Unlisted

As with genes in highly variable regions, genes in highly conserved regions were classified as 'free-living', 'parasitic' or 'same' (Table 5.5). Stage-specific information was available for 135 of 146 genes. Thus, while parasitism genes were substantially over-represented in very variable regions, they were under-represented in highly conserved regions compared to the genome as a whole. In contrast, there was no clear relationship between SNP density and frequency of free-living genes.

5.3.5 Analysis of differentially expressed genes

5.3.5.1 Parasitism genes

The 100 'most parasitic' genes (see Section 5.2.4.2) were retrieved from Hunt *et al.* (2016) (Table 5.9). Six of these genes were found in parts of expansion clusters that overlie poorly assembled regions of the genome and were discarded from further analyses. Of the remainder, 19 were also detected among the genes in the most variable genomic windows, while none occurred in the least variable genomic windows. Twenty-seven of the most parasitic genes had no coding SNPs, while the highest coding SNP density detected was 85.9 per kb. The average coding SNP density among non-discarded genes was 15.1 (SD = 20.3, range = 0-85.9, Table 5.9), which is substantially higher than the genome-wide SNP

density of 4.1 per kb (Table 5.10). Thus, genes that are upregulated in the parasitic adult female morph appear to be highly diverse compared to the rest of the genome.

Table 5.9: The genes most upregulated in the parasitic adult female morph of S. ratti*, compared with the free-living adult female morph. "SNP types" details the absolute numbers of synonymous (S), nonsynonymous (NS) and STOP-causing SNPs. "Function" refers to the functional description given on WormBase ParaSite, with "n/a" recorded here as "hypothetical protein". "Fold change" refers to the base 2 log fold difference in expression between the two morphs (Hunt et al. 2016). Genes that are "discarded" had poor underlying assembly according to Gap5 analysis of expansion clusters and flanking regions (Appendix 1) and so were discounted from analyses.*

Gene	**Function**	**Coding SNPs per kb**	**SNP types (S/NS/STOP)**	**Fold Change**	**Expansion cluster**	**Variable region**
SRAE_2000436100	Hypothetical protein	4.7	1 / 2 / 0	14.5	None	None
SRAE_X000014200	Hypothetical protein	15.2	0 / 6 / 1	13.4	None	None
SRAE_X000201100	Astacin-like metalloendopeptidase	29.8	17 / 27 / 1	13	None	50
SRAE_1000182300	CAP domain-containing	0	0 / 0 / 0	12.9	EC1	None
SRAE_2000498800	Phloem filament PP1 domain-containing	0	0 / 0 / 0	12.7	None	None
SRAE_2000067500	Transthyretin-like family-containing	15.9	1 / 6 / 0	12.6	None	None
SRAE_X000222400	Hypothetical protein	73.8	7 / 20 / 0	12.1	None	None
SRAE_1000182700 **(discarded)**	CAP domain-containing	0	0 / 0 / 0	12.1	EC1	None
SRAE_2000420000	Astacin-like metalloendopeptidase	1.7	0 / 1 / 1	12	None	None
SRAE_X000200700	Hypothetical protein	57.9	3 / 26 / 0	11.8	None	None
SRAE_X000124900	Hypothetical protein	0	0 / 0 / 0	11.8	None	None
SRAE_2000485600	tissue inhibitor of metalloproteinase family	15	3 / 3 / 1	11.8	None	None
SRAE_X000222600	Hypothetical protein	30.8	1 / 10 / 0	11.7	None	45
SRAE_X000055800	Hypothetical protein	54.3	3 / 12 / 0	11.7	None	None
SRAE_2000498700	Hypothetical protein	27.8	4 / 9 / 0	11.7	None	None
SRAE_X000124800	Hypothetical protein	0	0 / 0 / 0	11.6	None	None

SRAE_X000037700	Hypothetical protein	0	0 / 0 / 0	11.6	None	None
SRAE_0000057900	Hypothetical protein	42.9	3 / 10 / 0	11.5	None	38
SRAE_2000457510	CAP domain-containing	13.8	4 / 6 / 0	11.4	None	None
SRAE_1000045800	Lipase, class 3 family-containing	15.2	7 / 8 / 0	11.4	None	27
SRAE_2000506800	Hypothetical protein	14.7	8 / 27 / 0	11.3	None	None
SRAE_0000045600	CAP domain-containing protein	9.5	3 / 4 / 0	11.3	None	None
SRAE_2000453600	Astacin-like metalloendopeptidase	0	0 / 0 / 0	11.3	EC7	None
SRAE_2000522700	Trypsin Inhibitor-like	9	3 / 3 / 0	11.2	None	None
SRAE_0000071120	Metalloendopeptidase	1	0 / 1 / 0	11.2	None	None
SRAE_2000522300	Purple acid phosphatase	4.8	2 / 5 / 0	11.2	None	None
SRAE_2000465200	Hypothetical protein	0	0 / 0 / 0	11.1	None	None
SRAE_2000475600	Hypothetical protein	34.9	3 / 10 / 0	11.1	None	None
SRAE_2000077400	CAP domain-containing	13	4 / 7 / 0	11.1	EC2	None
SRAE_2000465000	Hypothetical protein	0	0 / 0 / 0	11.1	None	None
SRAE_X000168400	Hypothetical protein	23.8	2 / 7 / 0	11	None	57
SRAE_2000499910	Hypothetical protein	7.5	2 / 2 / 0	11	None	None
SRAE_2000499810	Hypothetical protein	41.4	4 / 17 / 0	11	None	None
SRAE_0000057800	Hypothetical protein	28.5	1 / 9 / 0	11	None	7
SRAE_0000077300	Trypsin Inhibitor-like	6.3	1 / 1 / 1	11	None	None
SRAE_1000296600	Hypothetical protein	0	0 / 0 / 0	10.9	None	None
SRAE_2000499820	Hypothetical protein	1.6	1 / 0 / 0	10.9	None	None
SRAE_2000522600	Trypsin Inhibitor-like	25.6	10 / 7 / 0	10.9	None	None
SRAE_X000144210	Astacin-like metalloendopeptidase	21.5	6 / 21 / 0	10.9	EC14	54
SRAE_2000486100	Tissue inhibitor of metalloproteinase family	51.9	7 / 17 / 0	10.9	None	None
SRAE_X000169100	Prolyl endopeptidase	0	0 / 0 / 0	10.9	None	None
SRAE_2000453500	Astacin-like metalloendopeptidase	2.5	0 / 3 / 0	10.8	EC7	None
SRAE_2000485800	Tissue inhibitor of metalloproteinase family	0	0 / 0 / 0	10.8	None	None

SRAE_2000499500	Hypothetical protein	0	0 / 0 / 0	10.7	None	3
SRAE_X000065700	Prolyl endopeptidase	3.1	4 / 3 / 0	10.7	None	None
SRAE_2000461200	Hypothetical protein	25.2	3 / 13 / 0	10.7	None	None
SRAE_X000051700	Hypothetical protein	9.1	1 / 3 / 0	10.6	None	28
SRAE_2000456500	Metalloendopeptidase	2.3	2 / 1 / 0	10.6	None	None
SRAE_1000182400	CAP domain-containing	1.2	1 / 0 / 0	10.6	EC1	None
SRAE_2000460300	Zinc metalloproteinase	17.1	5 / 19 / 0	10.6	None	None
SRAE_2000525700 **(discarded)**	Astacin-like metalloendopeptidase	29.4	3 / 22 / 0	10.6	EC11	None
SRAE_X000201200	Hypothetical protein	12	1 / 5 / 0	10.5	None	50
SRAE_0000008000	Hypothetical protein	37.5	10 / 32 / 0	10.5	None	35
SRAE_X000038800	Hypothetical protein	0	0 / 0 / 0	10.4	None	None
SRAE_2000076800	CAP domain-containing	39.2	11 / 18 / 1	10.4	EC2	14
SRAE_2000515500	Hypothetical protein	2.9	9 / 15 / 0	10.4	None	None
SRAE_X000055700	SRAE_X000055700	0	0 / 0 / 0	10.4	None	None
SRAE_1000182600 **(discarded)**	CAP domain-containing	17.8	2 / 14 / 0	10.4	EC1	None
SRAE_X000168800	Prolyl endopeptidase	0	0 / 0 / 0	10.3	None	None
SRAE_X000191800	Hypothetical protein	0	0 / 0 / 0	10.3	None	None
SRAE_0000081000	Metalloendopeptidase	0	0 / 0 / 0	10.3	None	None
SRAE_0000000600	Hypothetical protein	0	0 / 0 / 0	10.3	None	None
SRAE_X000200300	Hypothetical protein	2	1 / 0 / 0	10.3	None	58
SRAE_2000523800	Astacin-like metalloendopeptidase	2.8	2 / 2 / 0	10.3	EC10	None
SRAE_1000183100 **(discarded)**	CAP domain-containing	20.4	3 / 14 / 1	10.3	EC1	None
SRAE_2000077000	CAP domain-containing	19	6 / 9 / 1	10.2	EC2	14
SRAE_X000246200	Hypothetical protein	0	0 / 0 / 0	10.2	None	None
SRAE_X000066100	Astacin-like metalloendopeptidase	28.2	18 / 23 / 0	10.2	None	17
SRAE_X000055500	Hypothetical protein	0	0 / 0 / 0	10.2	None	None
SRAE_2000499920	Hypothetical protein	85.9	14 / 52 / 0	10.2	None	None
SRAE_X000195900	Hypothetical protein	30.7	2 / 6 / 1	10.1	None	33

SRAE_X000226400	CAP domain-containing protein	2	1 / 1 / 0	10.1	None	None
SRAE_2000451600	Transthyretin-like family-containing	20.6	2 / 8 / 0	10.1	None	None
SRAE_X000191900	Hypothetical protein	0	0 / 0 / 0	10.1	None	None
SRAE_2000455000	Astacin-like metalloendopeptidase	2.6	1 / 2 / 0	10.1	EC8	None
SRAE_0000078700	Hypothetical protein	0	0 / 0 / 0	10	None	None
SRAE_2000455300	Acetylcholinesterase	0	0 / 0 / 0	10	EC8	None
SRAE_X000201000	Hypothetical protein	1.9	0 / 1 / 0	10	None	None
SRAE_2000499200	Hypothetical protein	75.7	8 / 27 / 2	10	None	41
SRAE_2000071300	Hypothetical protein	30.3	8 / 17 / 0	10	None	None
SRAE_X000168300	Hypothetical protein	46.3	8 / 22 / 0	10	None	57
SRAE_2000124300	CAP domain-containing protein	69	17 / 49 / 4	10	EC3	5
SRAE_2000457710	Metallopeptidase, catalytic domain-containing	1.8	2 / 0 / 0	10	None	None
SRAE_2000527500	CAP domain-containing	3.3	3 / 0 / 0	10	EC12	None
SRAE_2000453700	Astacin-like metalloendopeptidase	5.5	0 / 7 / 0	9.9	EC7	None
SRAE_2000325600	Astacin-like metalloendopeptidase	0.9	1 / 0 / 0	9.9	EC5	None
SRAE_2000525800 **(discarded)**	Astacin-like metalloendopeptidase	1.4	0 / 1 / 0	9.9	EC11	None
SRAE_X000055600	Hypothetical protein	0	0 / 0 / 0	9.9	None	None
SRAE_2000482710	Zinc metalloproteinase	17.8	6 / 19 / 0	9.8	None	None
SRAE_0000077800	Hypothetical protein	4.6	0 / 1 / 0	9.8	None	None
SRAE_2000489900	Aspartic peptidase family	0.9	1 / 0 / 0	9.8	None	None
SRAE_2000289800 **(discarded)**	Astacin-like metalloendopeptidase	0	0 / 0 / 0	9.8	EC11	None
SRAE_2000508100	Hypothetical protein	66	4 / 14 / 2	9.8	None	None
SRAE_2000456300	Acetylcholinesterase	5.2	1 / 8 / 0	9.7	None	None
SRAE_0000071100	Metalloendopeptidase	0	0 / 0 / 0	9.7	None	None
SRAE_0000082200	Trypsin Inhibitor-like	18.1	6 / 19 / 0	9.7	None	None
SRAE_2000126100	Hypothetical protein	8.6	6 / 2 / 0	9.7	FR3R	None

SRAE_X000158300	Acetylcholinesterase	3	3 / 2 / 0	9.7	None	None
SRAE_0000081500	CAP domain-containing protein	5.6	2 / 3 / 0	9.7	None	None

Table 5.10: SNP density in coding sequence of parasitism genes and free-living genes (see Section 5.2.4.2), or across the whole genome.

	SNP density (SNPs per kb)
Parasitism genes (coding sequence only)	15.1
Free-living genes (coding sequence only)	3.3
Whole genome (all sequence	4.1

Of the 100 most parasitic genes, 45 were described in WormBase ParaSite as 'hypothetical proteins'. Highly prominent among the remainder were genes encoding astacins (N = 20) and CAP domain-containing proteins (N= 13), while genes encoding acetylcholinesterases, netrin-domain containing proteinase inhibitors, prolyl endopeptidases and trypsin inhibitor-like proteins were also over-represented. Genes encoding acetylcholinesterases, similar to astacins and CAP domain-containing proteins, are comparatively expanded in *Strongyloides* and are hypothesised to have roles in parasitism (Hunt *et al.* 2016). CAP domain-containing protein genes were particularly SNP-dense, on average (mean =29.5, SD=27.9 SNPs per kb, N = 10), compared with astacin genes (mean = 11.77, SD = 15.77, N = 20) and all other genes within the most parasitic set (mean = 13.51, SD = 19.5, N = 63).

Genes within a family are likely to have similar coding sequences, and this may promote misalignment of sequencing reads such that intra-individual variation in these reads is artificially inflated. As many of the most parasitic genes for which functional descriptions are available belong to one of a small number of large protein families, sequencing read misalignment may explain the unusually high SNP density observed. Alternatively, relaxed selection due to redundancy in these high copy number families may mean that mutations deleterious to gene function are not purged from the genome by selection. Given that CAP domain-containing proteins were found to be particularly SNP-dense, these might be particularly affected. The density of STOP-causing SNPs in astacins, CAP domain-containing proteins, and all other parasitism genes was determined to be 0.007, 0.042 and 0.033 per kb, respectively. Thus, there was no tendency for genes in the largest two families (astacins and CAP domain-containing proteins) to have proportionately more STOP-causing SNPS than other genes. Further, these values are similar to those seen in non-astacin, non-CAP domain-containing proteins in highly variable regions (Section 5.1.3). These low rates of STOP mutations are consistent with the hypothesis that the high SNP density observed in CAP domain-containing proteins are a response to selection. The occurrence of non-coding, synonymous and non-synonymous SNPs are shown in Table 5.9 and Figure 5.3. The

overall density of STOP-causing SNPs in the most parasitic genes was 0.21 per kb, equivalent to 5.8% of all SNPs in these genes.

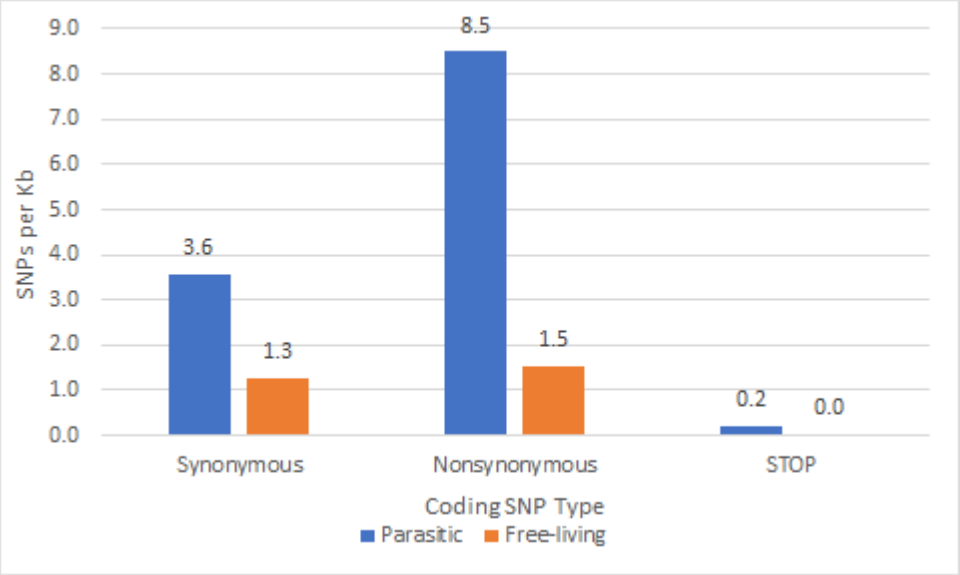


Figure 5.3: Density of synonymous, nonsynonymous and STOP-causing ('STOP') SNPs in the coding sequences of genes with the greatest differential in expression between the parasitic adult female morph (upregulated genes shown in blue) and free-living adult female morph (upregulated genes shown in orange).

5.3.5.2 Free-living genes

The 100 'most free-living' genes (see Section 5.2.4.2) were retrieved from Hunt *et al.* (2016) (Table 5.11). None of these genes were found in expansion clusters, and therefore none were removed on the basis of poor underlying assembly. SNP density of exons in these genes ranged from 0 (N = 24) to 41.5 SNPs per kb with a mean of 3.3 (SD = 5.1). This value is lower than the genome-wide mean of 4.1 SNPs per kb, and contrasts greatly with the mean of 15.1 observed for parasitism genes (Table 5.10). Thus, high SNP density does not appear to be associated with stage-specific expression *per se* and may represent selective pressures acting upon parasitism-associated traits.

Table 5.11: The genes most upregulated in the free-living adult female morph of Strongyloides ratti, *compared with the parasitic adult female morph. "SNP types" details the absolute numbers of synonymous (S), nonsynonymous (NS) and STOP-causing SNPs. "Function" refers to the functional description given on WormBase ParaSite, with "n/a" recorded here as "hypothetical protein". "Fold change" refers to the base 2 log fold difference in expression between the two morphs (Hunt et al. 2016). Genes that are "discarded" had poor underlying assembly according to Gap5 analysis of expansion clusters and flanking regions (Appendix 1) and so were discounted from analyses.*

Gene	**Function**	**Coding SNPs per kb**	**SNP types (S/NS/STOP)**	**Fold change**	**Expansion cluster**	**Variable region**
SRAE_2000529300	ShKT domain-containing protein	13.5	0 / 4 / 0	11	None	None
SRAE_2000226500	ShKT domain-containing protein	41.5	5 / 17 / 0	10.8	None	None
SRAE_X000147400	Phosphate-regulating neutral endopeptidase	4.8	2 / 5 / 0	10.4	None	None
SRAE_1000161500	DUF148 domain-containing	1.6	0 / 1 / 0	10.3	None	None
SRAE_X000018600	Mucin 18B	2.4	2 / 2 / 0	10.3	None	None
SRAE_X000158400	Astacin-like metalloendopeptidase	1.4	0 / 2 / 0	10.1	None	46
SRAE_2000474600	ShKT domain-containing protein	2.2	0 / 2 / 0	10.1	None	None
SRAE_X000135300	Collagen alpha-5(IV) chain	1	1 / 0 / 0	10	None	None
SRAE_2000463500	Hypothetical protein	5.5	1 / 1 / 1	9.8	None	None
SRAE_2000079600	Transcription factor Sp6	5.3	3 / 1 / 0	8.8	None	None
SRAE_1000062000	Nematode cuticle collagen	5.3	4 / 1 / 0	8.8	None	None
SRAE_2000491300	ShKT domain-containing protein	0	0 / 0 / 0	8.7	None	None
SRAE_1000170700	Hypothetical protein	6.9	8 / 2 / 0	8.6	None	None
SRAE_1000213600	Aspartic peptidase family	2.2	2 / 1 / 0	8.5	None	None
SRAE_2000400600	Saposin-like type B, 1 domain	0	0 / 0 / 0	8.4	None	None
SRAE_2000477400	ShKT domain-containing protein	8.4	1 / 5 / 0	8.4	None	None
SRAE_1000227500	Hypothetical protein	3.5	0 / 2 / 0	8.2	None	None
SRAE_2000292900	GH07323p	1.2	1 / 1 / 0	8	None	None
SRAE_X000034300	Hypothetical protein	8	1 / 5 / 0	8	None	None
SRAE_2000126900	Nematode cuticle collagen	2.1	1 / 1 / 0	7.9	FR4R	None
SRAE_X000032100	Lipase, class 3 family-containing	3.1	2 / 1 / 0	7.9	None	None
SRAE_1000151800	Protein COL-120	4.1	4 / 1 / 0	7.7	None	None
SRAE_X000215700	Hypothetical protein	9.6	2 / 1 / 0	7.7	None	None
SRAE_2000363400	Hypothetical protein	2.8	1 / 0 / 0	7.5	None	None
SRAE_2000434700	Collagen alpha-5(IV) chain	1	0 /1 / 0	7.5	None	None

SRAE_1000352300	Serine/threonine-/dual specificity protein kinase	0.8	1 / 0 / 0	7.4	None	None
SRAE_1000099100	Hypothetical protein	0.7	1 / 0 / 0	7.4	None	None
SRAE_X000100400	Hypothetical protein	0.9	0 / 1 / 0	7.4	None	None
SRAE_1000220400	Hypothetical protein	0	0 / 0 / 0	7.4	None	None
SRAE_2000033400	Heat shock protein Hsp-12.2	0	0 / 0 / 0	7.2	None	None
SRAE_1000228700	Cell death specification protein 2	1.2	1 / 0 / 0	7.2	None	None
SRAE_X000095500	Hypothetical protein	0	0 / 0 / 0	7.2	None	None
SRAE_1000073100	Protein lin-32	0	0 / 0 / 0	7.2	None	None
SRAE_1000271200	MSP domain; PapD-like domain-containing	0	0 / 0 / 0	7.2	None	None
SRAE_2000115300	Metallothionein, family 4, echinoidea-containing	3	0 / 2 / 0	7.1	None	None
SRAE_1000098100	Hypothetical protein	0	0 / 0 / 0	7.1	None	None
SRAE_2000006500	Hypothetical protein	0	0 / 0 / 0	7	None	None
SRAE_2000425600	Hypothetical protein	0	0 / 0 / 0	7	None	None
SRAE_0000045200	von Willebrand factor, type A domain-containing	3.2	10 / 8 / 0	7	None	None
SRAE_1000198200	Glycoside hydrolase	1.2	1 / 1 / 0	7	None	None
SRAE_1000159100	Hypothetical protein	0.7	0 / 1 / 0	6.9	None	None
SRAE_1000268100	Nematode cuticle collagen	0	0 / 0 / 0	6.9	None	None
SRAE_2000473500	Sulfotransferase family	11	5 / 3 / 3	6.9	None	None
SRAE_X000144300	Protein GLF-1	0.6	0 / 1 / 0	6.8	FR14R	None
SRAE_X000086900	Hypothetical protein	7.2	1 / 4 / 0	6.8	None	None
SRAE_1000058300	Hypothetical protein	7.6	3 / 6 / 0	6.8	None	None
SRAE_X000224800	Nematode cuticle collagen	1.9	1 / 1 / 0	6.8	None	None
SRAE_1000033500	Collagen alpha-5(IV) chain	0.9	0 / 1 / 0	6.7	None	None
SRAE_1000004900	Glycoside hydrolase, family 25	1.3	1 / 0 / 0	6.7	None	None
SRAE_1000015400	Lipase EstA/Esterase EstB family-containing	2	1 / 1 / 0	6.7	None	None
SRAE_2000439500	Collagen alpha-5(IV) chain	0	0 / 0 / 0	6.7	None	None
SRAE_X000039100	Brain-specific homeobox protein homolog	0	0 / 0 / 0	6.7	None	None
SRAE_1000049700	Hypothetical protein	0	0 / 0 / 0	6.7	None	None
SRAE_1000036400	Histidine decarboxylase	12	3 / 4 / 0	6.6	None	None
SRAE_2000476800	Hypothetical protein	0	0 / 0 / 0	6.5	None	None
SRAE_2000431300	Nematode fatty acid retinoid binding family-containing	0	0 / 0 / 0	6.5	None	None
SRAE_X000210100	Domain of unknown function DB domain-containing	1.3	1 / 0 / 0	6.4	None	None

SRAE_2000437000	CAP domain-containing protein	12.6	3 / 5 / 0	6.4	None	None
SRAE_X000195000	Si:ch211-105d4.5	0	0 / 0 / 0	6.4	None	None
SRAE_2000145200	Nematode cuticle collagen	0.9	1 / 0 / 0	6.4	None	None
SRAE_2000482000	N-acylethanolamine-hydrolyzing acid amidase	1.7	1 / 1 / 0	6.4	None	None
SRAE_1000115300	Hypothetical protein	1.3	1 / 1 / 0	6.3	None	None
SRAE_2000459400	Acyl-CoA N-acyltransferase domain-containing	0	0 / 0 / 0	6.3	None	None
SRAE_2000468700	Protein dyf-8	0	0 / 0 / 0	6.3	None	None
SRAE_2000214100	Nematode cuticle collagen	1	1 / 0 / 0	6.3	None	None
SRAE_2000481600	MD-2-related lipid-recognition domain	1.9	1 / 0 / 0	6.3	None	None
SRAE_1000158100	Hypothetical protein	2.5	2 / 2 / 0	6.2	None	None
SRAE_0000015000	MD-2-related lipid-recognition domain-containing	3.7	0 / 2 / 0	6.2	None	None
SRAE_2000365700	Glycoside hydrolase, catalytic domain	2.5	0 / 1 / 1	6.2	None	None
SRAE_2000466700	Properdin	5	7 / 4 / 0	6	None	None
SRAE_2000324800	Hypothetical protein	16.2	15 / 5 / 0	6	None	None
SRAE_2000014200	Saposin B domain	2.4	0 / 1 / 0	5.9	None	None
SRAE_1000271800	MSP domain; PapD-like domain-containing	0	0 / 0 / 0	5.9	None	None
SRAE_1000068700	Thrombospondin, type 1 repeat-containing	6.7	4 / 0 / 0	5.9	None	None
SRAE_X000188000	Sphingosine-1-phosphate lyase 1	2.2	2 / 4 / 0	5.8	None	None
SRAE_X000258800	Thioredoxin-like fold domain-containing	0	0 / 0 / 0	5.7	None	None
SRAE_2000009100	Protein-tyrosine phosphatase	1.1	0 / 1 / 0	5.7	None	None
SRAE_0000012300	Hypothetical protein	4.7	0 / 2 / 0	5.7	None	None
SRAE_2000448200	Hypothetical protein	5.6	0 / 2 / 0	5.6	None	None
SRAE_X000101600	Hypothetical protein	5.6	5 / 3 / 0	5.6	None	None
SRAE_2000041300	LDLR class B repeat	0	0 / 0 / 0	5.5	None	None
SRAE_X000083300	Protein CUTL-16	2.1	2 / 1 / 0	5.5	None	None
SRAE_X000079800	Hypothetical protein	0	0 / 0 / 0	5.5	None	None
SRAE_X000138300	Protein mesh	0	0 / 0 / 0	5.4	None	None
SRAE_1000020600	Hypothetical protein	4.3	0 / 2 / 0	5.4	None	None
SRAE_2000476700	Hypothetical protein	0	0 / 0 / 0	5.4	None	None
SRAE_2000347100	Lipase EstA/Esterase EstB family-containing	1	0 / 1 / 0	5.4	None	None
SRAE_2000380600	GPCR, rhodopsin-like	0.9	0 / 1 / 0	5.4	None	None
SRAE_2000052400	Alpha crystallin/Hsp20 domain	1.2	3 / 1 / 0	5.4	None	None
SRAE_X000138500	Lipase, class 3 family-containing	8.8	4 / 4 / 0	5.3	None	None
SRAE_X000166100	Protein CDH-10	3.2	5 / 12 / 0	5.3	None	None

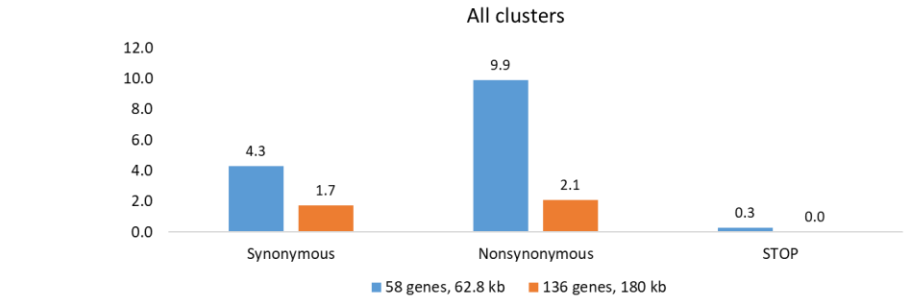
SRAE_1000022400	Epidermal growth factor-like domain	2.7	11 / 7 / 0	5.3	None	None
SRAE_X000157300	Homeobox protein HMX1	2.5	1 / 2 / 0	5.3	None	None
SRAE_X000043300	Proteasomal ubiquitin receptor ADRM1 homolog	6.3	0 / 7 / 0	5.3	None	None
SRAE_2000424300	Fatty-acid amide hydrolase 2	1.8	3 / 0 / 0	5.3	None	None
SRAE_X000098100	Astacin-like metalloendopeptidase	1.2	0 / 2 / 0	5.3	None	None
SRAE_X000095600	Hypothetical protein	4.1	0 / 2 / 0	5.2	None	None
SRAE_2000482100	Hypothetical protein	11.4	6 / 11 / 0	5.2	None	None
SRAE_2000377500	Nematode cuticle collagen	2	2 / 0 / 0	5.2	None	None
SRAE_2000360100	Alpha amylase family;	5	3 / 3 / 0	5.2	None	None

Twenty-nine of the 100 most free-living genes were described only as ‘hypothetical proteins’ in WormBase ParaSite. The remainder were dominated by collagen-coding genes, of which there were 11, and there were also five ShKT domain-containing protein genes. Furthermore, there were two genes encoding astacins and one encoding a CAP domain-containing protein (Table 5.11). Thus, though the expanded gene families are more prominent among genes upregulated in the parasitic female morph, a few are instead upregulated in the free-living morph. Although the small numbers of genes encoding astacins and CAP domain-containing proteins in the 100 most free-living genes preclude in-depth analysis, it is notable that the coding SNP densities of the two free-living-expressed astacin genes were 1.23 and 1.4 SNPs per kb respectively, much lower than the 11.77 SNPs per kb seen in the average parasitic-expressed astacin gene (Section 5.3.5.1), suggesting that it is being a parasitism gene, and not merely being an astacin gene, that is associated with the high SNP-density in the latter. Coding SNP density of free-living-expressed CAP domain-containing protein encoding genes was 12.6 SNPs per kb, which again is low compared with the 29.5 SNPs per kb seen in the average parasitism-expressed CAP domain-containing protein gene, though high compared with free-living genes generally.

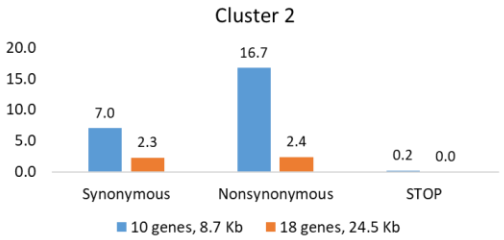
The density of STOP-causing SNPs in the 100 most free-living genes was 0.04 per kb, which appears low compared with the 0.21 per kb of the most parasitic genes. However, this is accounted for by the overall lower number of SNPs in the free-living compared with parasitic genes. STOP-causing SNPs accounted for 3.2% of all SNPs in the most free-living genes, which is close to the value of 5.8% observed for the most parasitic genes. As well as STOP-causing SNPs, the proportions of synonymous and non-synonymous SNPs are shown in Figure 5.3.

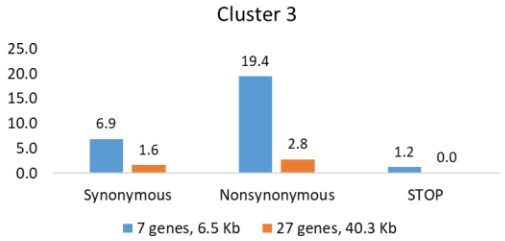
5.3.6 Analysis of expansion clusters

Table 5.2 lists all genes occurring in expansion clusters and their flanking regions that were retained after inspection of the underlying assembly (see Section 5.3.1). Most expansion cluster genes were expressed more strongly in the parasitic adult female morph than the free-living morph (Hunt *et al.*2016). Overall, expansion cluster genes had a 3-fold higher coding SNP density than did flanking region genes, with mean SNPs per kb values of 15.9 (SD = 18.3) and 4.6 (SD = 7.2) respectively (Figure 5.3.) This trend was largely consistent on a per expansion cluster basis, when they were individually compared to their own flanking regions, with only one exception – mean coding SNP density in expansion cluster 8 was lower than that of its flanking regions. Another striking difference between the genes in the expansion clusters and genes in the flanking clusters was in the relative rates of SNP types, with nonsynonymous SNPs occurring at approximately twice the rate of synonymous SNPs in expansion cluster genes, while the two classes of SNP occurred at approximately equal rates in flanking region genes (Figure 5.4). These patterns were generally consistent across all 12 expansion clusters (Figure 5.4). That nonsynonymous SNPs are more common in expansion cluster genes than flanking region genes suggests that expansion cluster genes are undergoing diversifying selection. A non-mutually exclusive alternative hypothesis is that redundancy amongst the members of an expanded gene family relieves selection pressure on these genes and allows them to accumulate deleterious nonsynonymous mutations that would otherwise be purged from the genome.

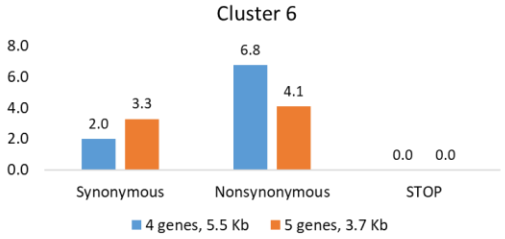

All clusters
12.0
10.0
8.0
6.0
4.0
2.0
0.0
4.3
1.7
9.9
2.1
0.3
0.0
Synonymous
Nonsynonymous
STOP
58 genes, 62.8 kb
136 genes, 180 kb

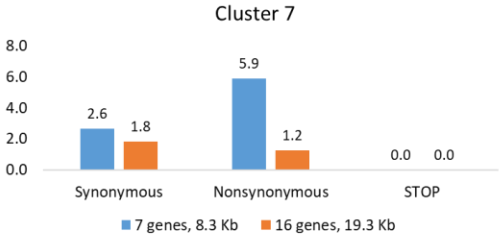

Cluster 1
1.5
1.0
0.5
0.0
0.8
1.0
0.6
0.8
0.0
0.0
Synonymous
Nonsynonymous
STOP
4 genes, 3.5 Kb
29 genes, 40 Kb


Cluster 2
20.0
15.0
10.0
5.0
0.0
7.0
2.3
16.7
2.4
0.2
0.0
Synonymous
Nonsynonymous
STOP
10 genes, 8.7 Kb
18 genes, 24.5 Kb

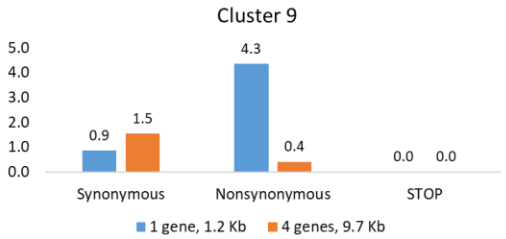

Cluster 3
25.0
20.0
15.0
10.0
5.0
0.0
6.9
1.6
19.4
2.8
1.2
0.0
Synonymous
Nonsynonymous
STOP
7 genes, 6.5 Kb
27 genes, 40.3 Kb

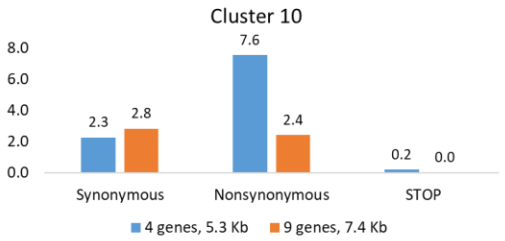

Cluster 5
3.0
2.0
1.0
0.0
1.6
0.4
2.8
1.1
0.0
0.0
Synonymous
Nonsynonymous
STOP
2 genes, 2.5 Kb
6 genes, 11.3 Kb

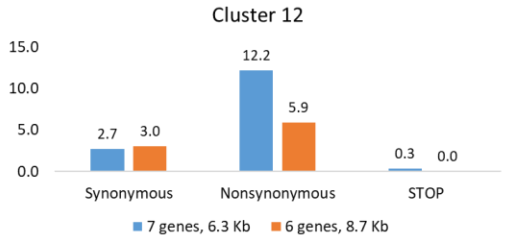

Cluster 6
8.0
6.0
4.0
2.0
0.0
2.0
3.3
6.8
4.1
0.0
0.0
Synonymous
Nonsynonymous
STOP
4 genes, 5.5 Kb
5 genes, 3.7 Kb

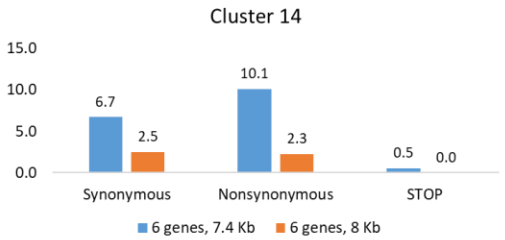

Cluster 7
8.0
6.0
4.0
2.0
0.0
2.6
1.8
5.9
1.2
0.0
0.0
Synonymous
Nonsynonymous
STOP
7 genes, 8.3 Kb
16 genes, 19.3 Kb


Cluster 8
0.8
0.6
0.4
0.2
0.0
0.6
0.2
0.4
0.4
0.0
0.0
Synonymous
Nonsynonymous
STOP
4 genes, 5.1 Kb
6 genes, 4.7 Kb


Cluster 9
5.0
4.0
3.0
2.0
1.0
0.0
0.9
1.5
4.3
0.4
0.0
0.0
Synonymous
Nonsynonymous
STOP
1 gene, 1.2 Kb
4 genes, 9.7 Kb


Cluster 10
8.0
6.0
4.0
2.0
0.0
2.3
2.8
7.6
2.4
0.2
0.0
Synonymous
Nonsynonymous
STOP
4 genes, 5.3 Kb
9 genes, 7.4 Kb


Cluster 12
15.0
10.0
5.0
0.0
2.7
3.0
12.2
5.9
0.3
0.0
Synonymous
Nonsynonymous
STOP
7 genes, 6.3 Kb
6 genes, 8.7 Kb


Cluster 14
15.0
10.0
5.0
0.0
6.7
2.5
10.1
2.3
0.5
0.0
Synonymous
Nonsynonymous
STOP
6 genes, 7.4 Kb
6 genes, 8 Kb

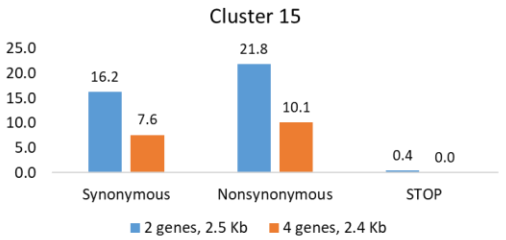

Cluster 15
25.0
20.0
15.0
10.0
5.0
0.0
16.2
7.6
21.8
10.1
0.4
0.0
Synonymous
Nonsynonymous
STOP
2 genes, 2.5 Kb
4 genes, 2.4 Kb

Figure 5.4: Density of synonymous (left), nonsynonymous (middle) or STOP-causing (right) SNPs in the coding sequences of genes in expansion clusters (blue) or associated flanking regions (orange) (see Table 5.2). Genes were excluded if the underlying assembly was found to be of poor quality, or if they were within an expansion cluster but not an astacin-, CAP domain-containing protein- or acetylchoninesterase encoding gene. Y-axis units are SNPs of the indicated type per kb.

To determine which of these hypotheses best fits the data, dN/dS calculations were performed for genes in expansion clusters and their associated flanking regions. This analysis considered only clades 1 and 3 and treated these highly genetically distinct clades as separate species (see section 5.2.5.1). In dN/dS ratio analyses, values above 1 indicate more nonsynonymous SNPs compared with the number expected under neutral evolution, and thus is evidence for diversifying selection; values less than 1 indicate fewer nonsynonymous SNPs than expected under neutral evolution, which is evidence of purifying selection (Yang and Bielawski 2000). dN/dS ratios could be calculated for 18 expansion cluster genes, representing 8 clusters. dN/dS ratios for these genes are shown in Table 5.12, and the mean dN/dS ratio was 0.81 (SD = 1.59). dN/dS ratios could be calculated for only 14 genes in flanking regions (Table 5.12), representing flanking regions for 6 expansion clusters, and the mean value of these was also 0.81 (SD = 1.06). Thus, at first glance, there appears to be no difference in dN/dS ratio between the two groups, in contrast to what was implied by the raw density of synonymous and nonsynonymous SNPs. Comparisons between expansions clusters and their own flanking regions are shown in Table 5.13 which suggests that genes in expansion clusters and their flanking regions often differ substantially in their dN/dS ratios, but the direction of these differences is not consistent among clusters and their flanking regions. This analysis used fewer of the genes found in expansion clusters and flanking regions than did the analysis of synonymous and nonsynonymous SNP counts which may account for the differing results. Further, this analysis assumed that clades 1 and 3 are reproductively isolated, when in fact rare sexual crosses among them may occur (Chapter 4).

Table 5.12: dN/dS ratios of genes in expansion clusters and flanking regions (Table 5.2), treating individuals from nuclear clades 1 and 3 (see Chapter 4) as though they were separate species. Only genes with sufficient information to calculate dN/dS ratios are shown.

Region	Gene	Function	dN/dS
FR1R	SRAE_1000184200	Basic-leucine zipper domain-containing	0.254
FR2L	SRAE_2000076300	Hypothetical protein	0.286
EC2	SRAE_2000076400	CAP domain-containing	0.257
FR2R	SRAE_2000077600	Hypothetical protein	3.89
FR3L	SRAE_2000123200	Armadillo-like helical domain -containing	0.274
	SRAE_2000124000	Bloom syndrome protein	0.134
EC3	SRAE_2000124300	CAP domain-containing	1.75

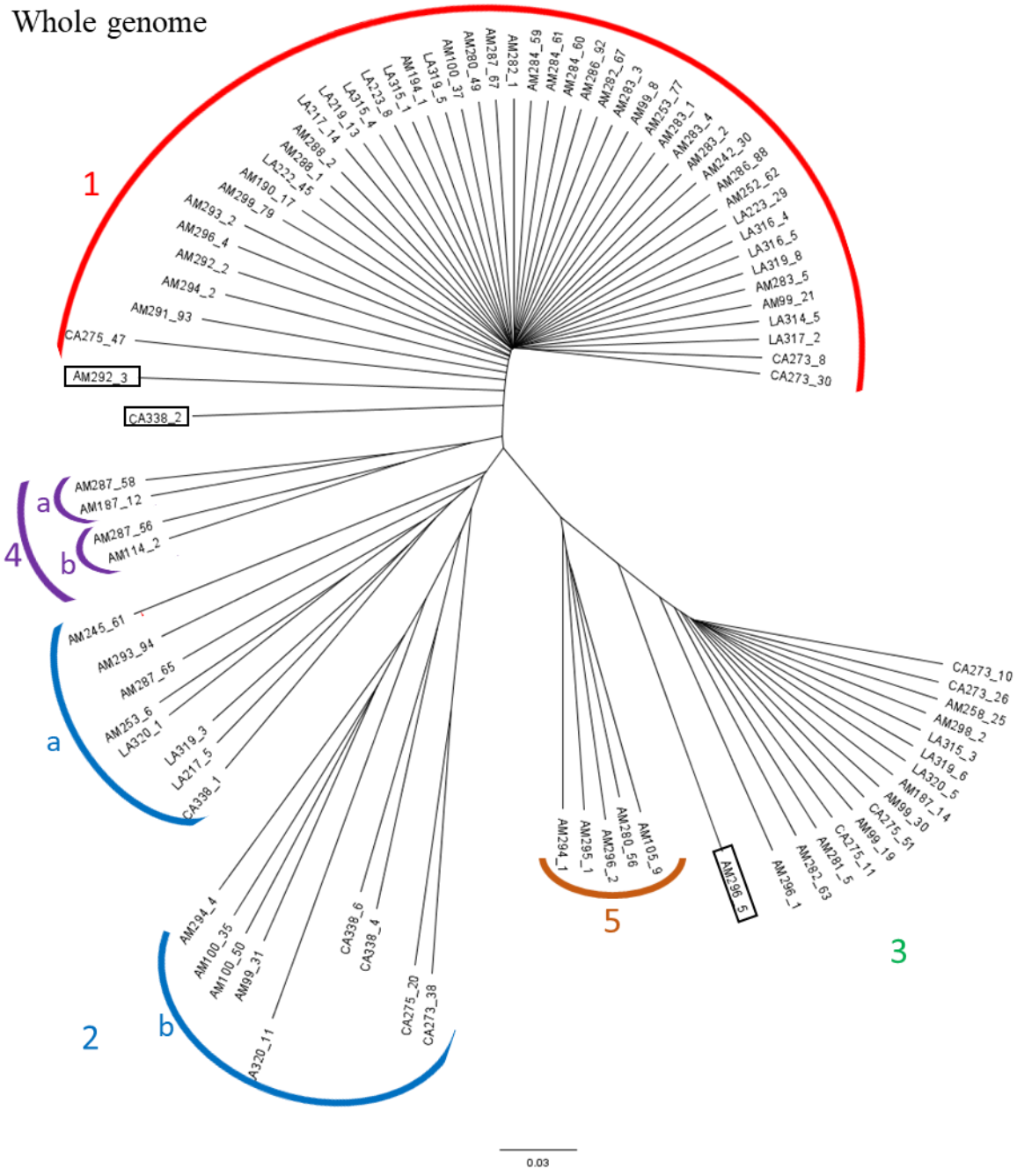
	SRAE_2000124400	CAP domain-containing	0.009
	SRAE_2000124500	CAP domain-containing	0.488
	SRAE_2000124600	CAP domain-containing	0.006
	SRAE_2000124800	CAP domain-containing	0.016
FR3R	SRAE_2000126200	Hypothetical protein	2.241
	SRAE_2000126290	Hypothetical protein	0.149
	SRAE_2000126400	Hypothetical protein	1.004
	SRAE_2000126600	CAP domain-containing	0.605
FR6L	SRAE_2000450300	UDP-glucosyltransferase	0.408
EC6	SRAE_2000450700	Astacin-like metallopeptidase	1.373
FR7L	SRAE_2000453100	Hypothetical protein	0.446
EC7	SRAE_2000453200	Astacin-like metallopeptidase	6.461
	SRAE_2000453800	Astacin-like metallopeptidase	0.377
FR7R	SRAE_2000454100	Hypothetical protein	0.13
	SRAE_2000454700	Glycosyl transferase, family 14-containing	0.257
EC8	SRAE_2000455000	Astacin-like metallopeptidase	0.021
EC12	SRAE_2000527100	CAP domain-containing	0.675
	SRAE_2000527300	CAP domain-containing	0.081
FR14L	SRAE_X000143500	Hypothetical protein	1.242
EC14	SRAE_X000143800	Astacin-like metallopeptidase	0.148
	SRAE_X000143900	Astacin-like metallopeptidase	0.488
	SRAE_X000144210	Astacin-like metallopeptidase	0.33
EC15	SRAE_X000146900	Astacin-like metallopeptidase	0.473

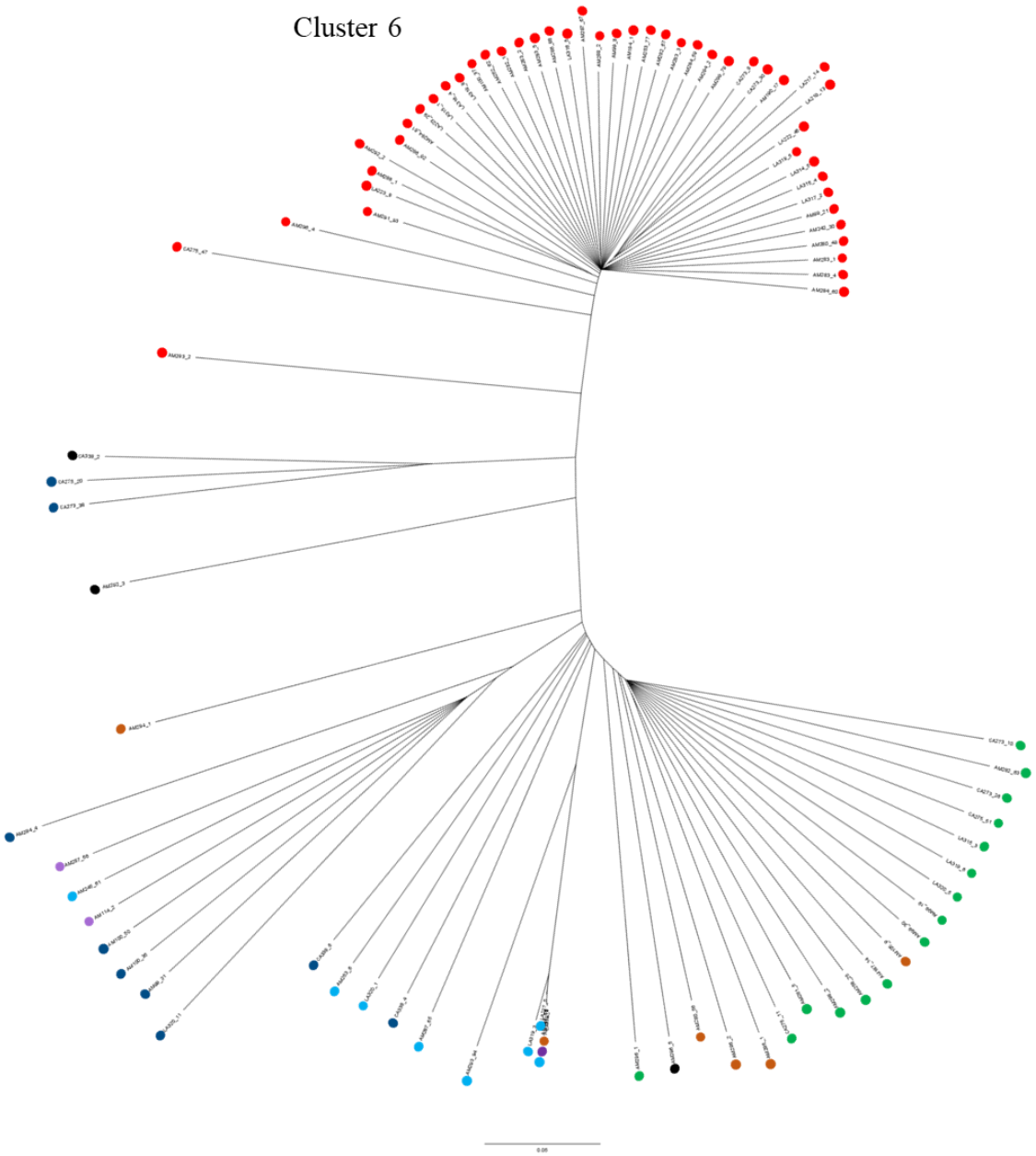
Table 5.13: Comparison of dN/dS ratios in expansion clusters and their associated flanking regions, averaged across all genes in those expansion clusters / flanking regions (Table 5.2). Individuals from nuclear clades 1 and 3 (see Chapter 4) were treated as though they were separate species. Only genes with sufficient information to calculate dN/dS ratios were included, numbers of these are indicated in parentheses.

Expansion cluster region	dN/dS ratio	
	Expansion cluster mean (N)	Flanking region mean (N)
2	0.257 (1)	2.088 (2)
3	0.454 (5)	0.735 (6)
6	1.373 (1)	0.408 (1)
7	3.419 (2)	0.278 (3)
14	0.322 (3)	1.242 (1)

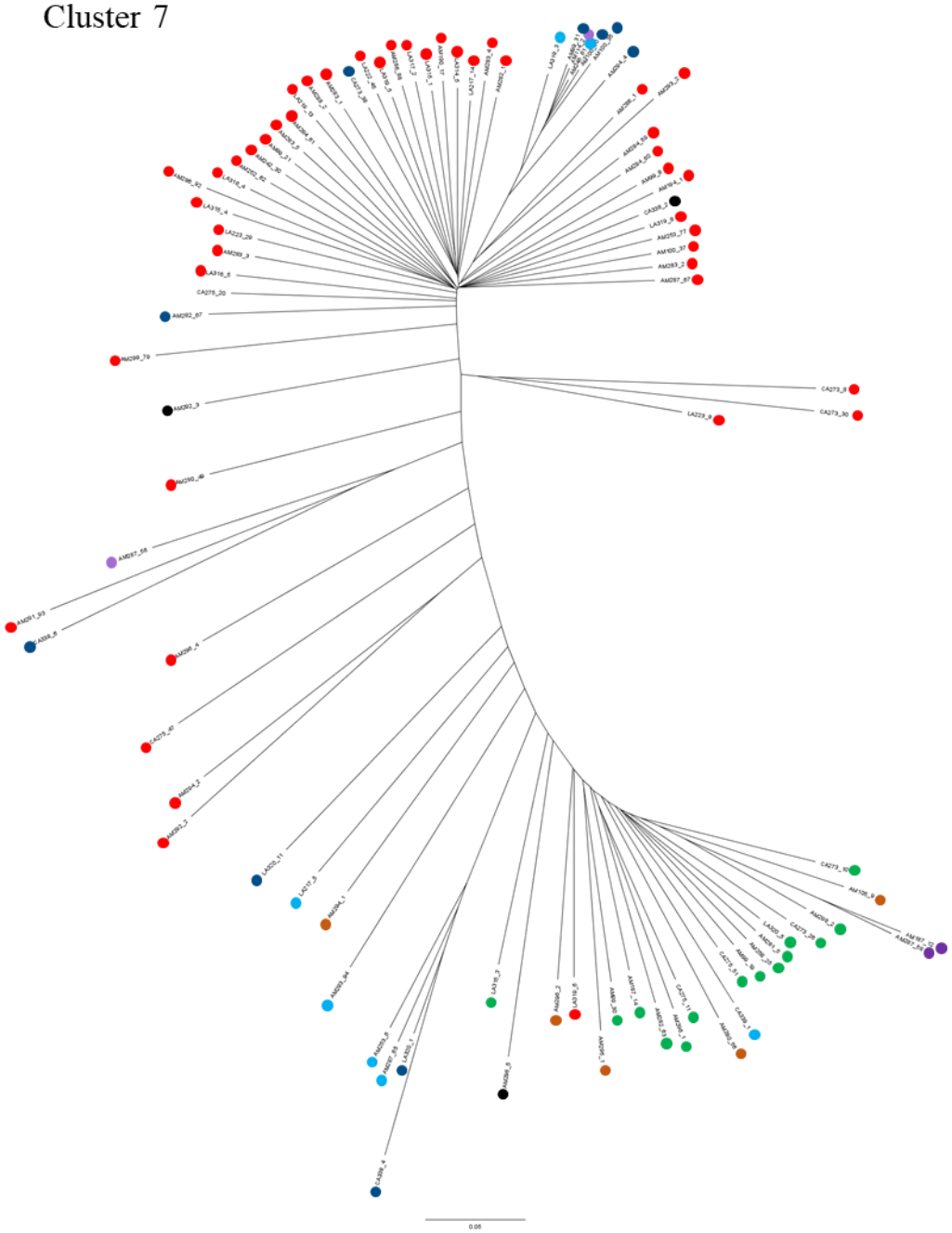
Neighbour-joining dendrograms for all samples, but based only on data from expansion clusters 6, 7, 8, 12 and 14, which are clusters that had had no genes excluded from analysis due to poor quality of the underlying assembly. As shown in Figure 5.5, the patterns depicted by cluster-specific dendrograms left the most genetically distinct nuclear clades of 1 and 3 largely intact. However, two clade 3 individuals grouped with clade 1 in expansion cluster-based tree 14, and one clade 1 individual grouped with clade 3 in cluster 7. Other nuclear clades and sub-clades from the whole-genome-based dendrograms were less well-preserved in expansion cluster dendrograms. Individuals from clade 5 generally grouped with clade 3 except in the dendrogram for expansion cluster 12, where clade 5 was placed close to clade 3 but distinct from it, and cluster 14, where clade 5 grouped with clade 1. The two individuals within nuclear sub-clade 4a generally grouped together in the expansion cluster

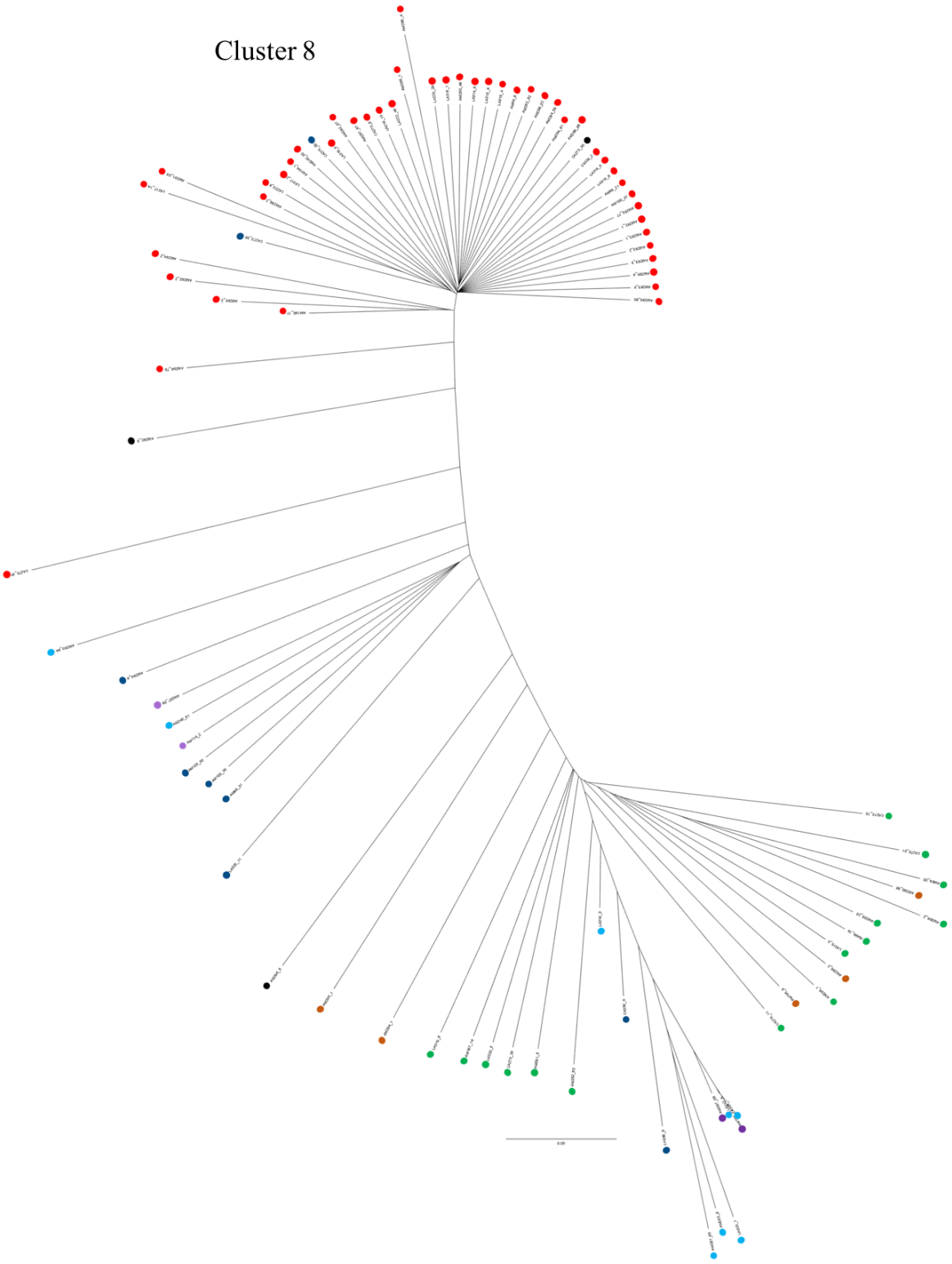
dendrograms, and likewise the individuals of 4b, but these two sub-clades rarely occurred in proximity as they do in the genome-wide cladogram, and their positions in the expansion-specific clusters was variable. In expansion clusters 6 and 12, most individuals of nuclear clade 2 clustered tougher, but otherwise nuclear clade 2 generally formed two groups within expansion cluster dendrograms, and there was incomplete concordance between the individuals in these two groups and sub-clades 2a and 2b in the genome-wide cladogram. Further, the occurrence of nuclear clade 2 individuals in the expansion cluster dendrograms with respect to clades 1 and 3 was inconsistent. In general, nuclear clades 1 and 3 appeared to represent the ends of a genotype spectrum, with other clades and individuals arrayed between them. Differences in topology among the expansion cluster-based dendrograms and the genome-wide dendrogram may be the result of rare recombination between clades, while differences in branch lengths among expansion cluster-based dendrograms suggest that expansion clusters are evolving at different rates in different individuals.


Whole genome
1
AM282_1
AM284_59
AM284_61
AM284_60
AM286_92
AM282_67
AM283_3
AM99_8
AM253_77
AM283_1
AM283_4
AM283_2
AM242_30
AM286_88
AM252_62
LA223_29
LA316_4
LA316_5
LA319_8
AM283_5
AM99_21
LA314_5
LA317_2
CA273_8
CA273_30
AM280_49
AM287_67
AM100_37
LA319_5
AM194_1
LA315_1
LA223_8
LA315_4
LA319_13
LA317_14
AM288_2
AM288_1
LA222_45
AM190_17
AM299_79
AM293_2
AM296_4
AM292_2
AM294_2
AM291_93
CA275_47
AM292_3
CA338_2
4
a
AM287_58
AM187_12
b
AM287_56
AM114_2
a
AM245_61
AM293_94
AM287_65
AM253_6
LA320_1
LA319_3
LA217_5
CA338_1
2
b
AM294_4
AM100_35
AM100_50
AM99_31
A320_11
CA338_6
CA338_4
CA275_20
CA273_38
5
AM294_1
AM295_1
AM280_56
AM296_2
AM105_9
AM296_5
3
CA273_10
CA273_26
AM258_25
AM298_2
LA315_3
LA319_6
LA320_5
AM187_14
AM99_30
CA275_51
AM99_19
CA275_11
AM281_5
AM282_63
AM296_1
0.03

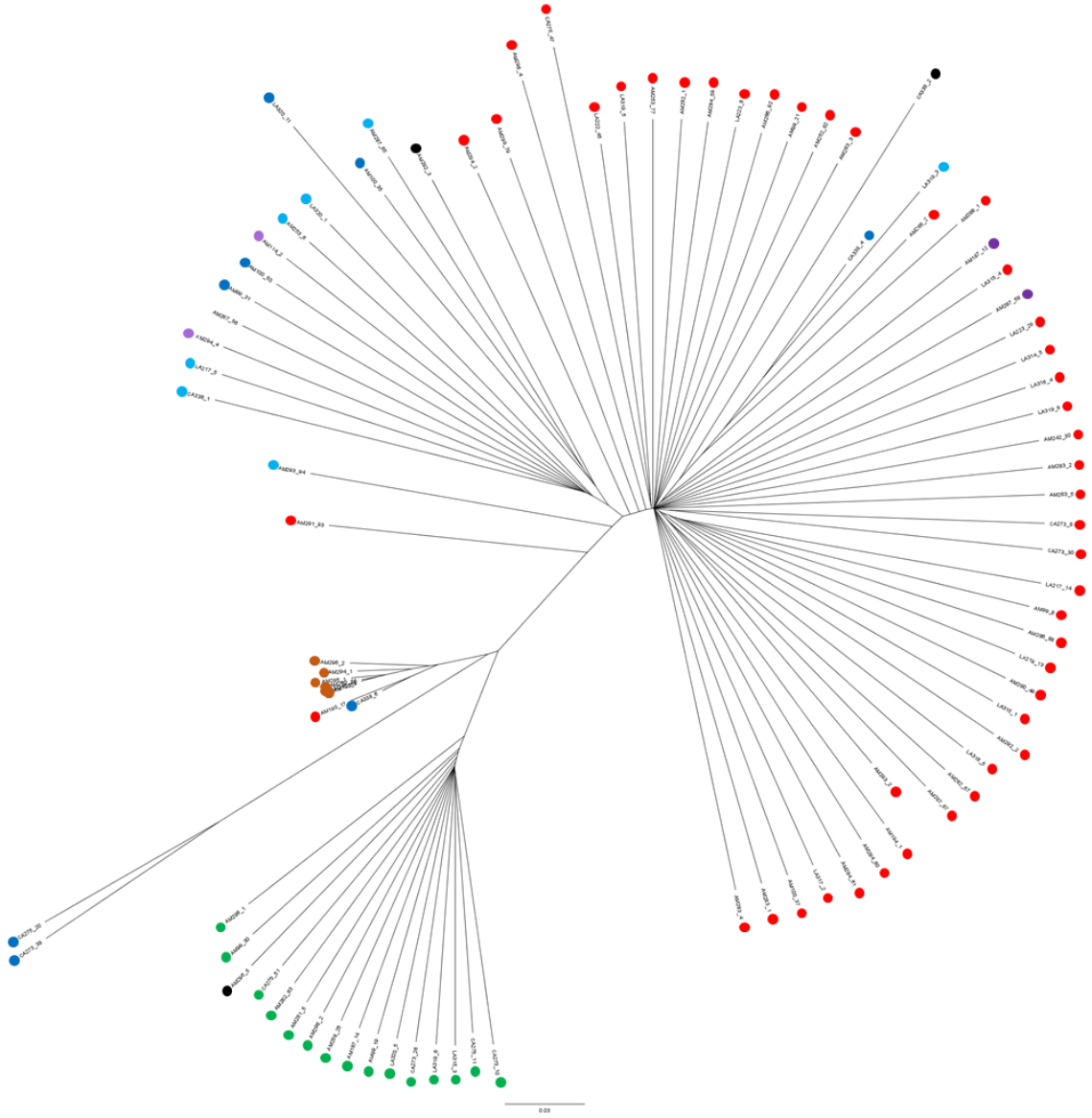

Cluster 6

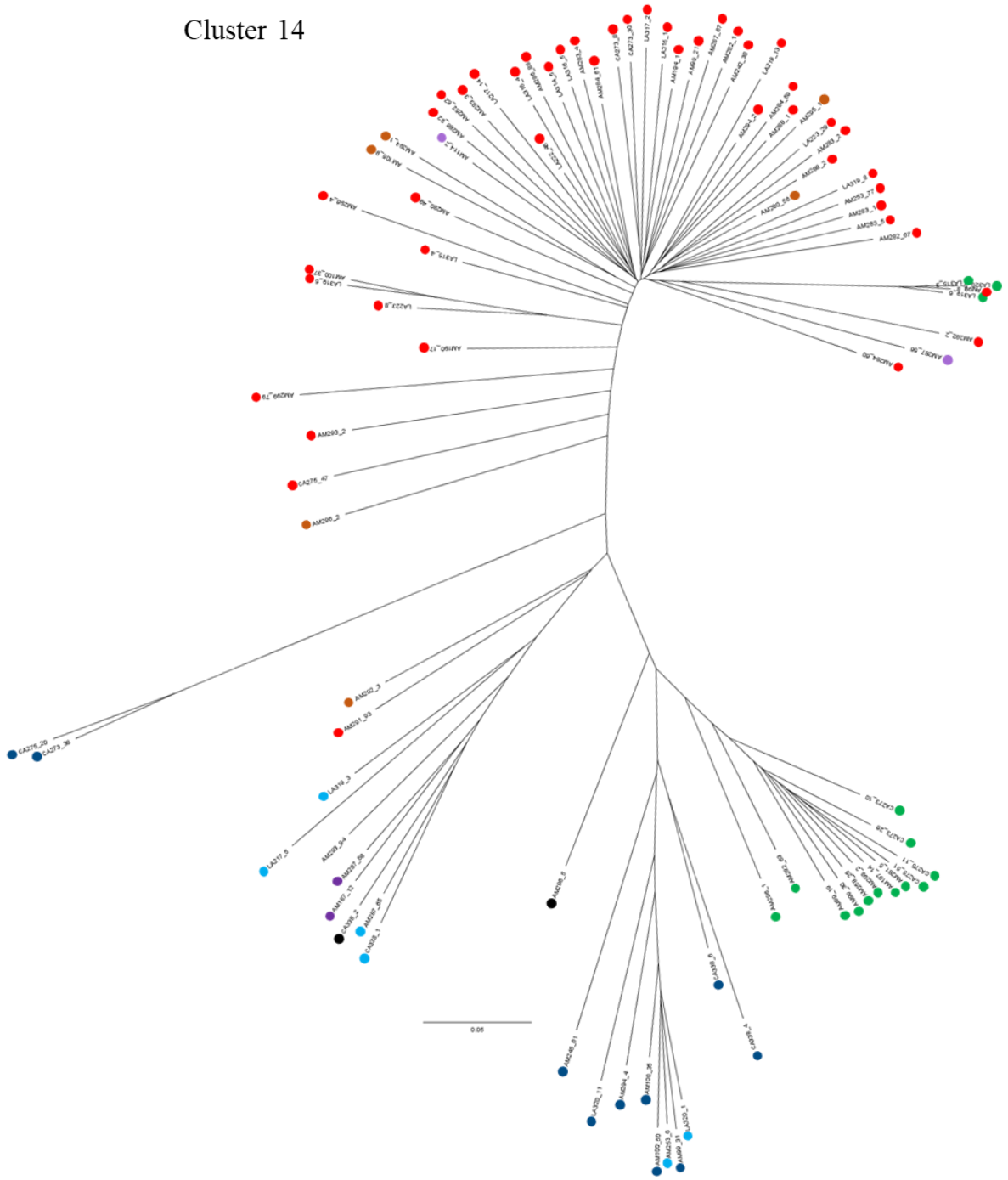
Cluster 7




Cluster 8

Cluster 12





5.5: Neighbour joining dendrograms of 90 S. ratti *individuals based on single nucleotide polymorphism (SNP) data for the whole genome, or for the expansion cluster as indicated (see Table 5.1). Whole sequence from the start to the end of each expansion cluster (not including flanking regions) was used. Clades (1 – 5) and sub-clades (2a and 2b, 4a and 4b) are defined in the whole genome dendrogram, and in all other dendrograms individuals are marked with circles coloured according to their (sub-)clade in the whole genome dendrogram. Branch lengths are relative such that the distance between the two most distant individuals is regarded as 1.*

5.4 Discussion

5.4.1 Diversity in parasitism genes

This study examined the distribution of genetic variation within the genomes of 90 individually sequenced *Strongyloides ratti* collected from wild rat hosts and used this to interrogate possible selection acting on *the S. ratti* genome. These data demonstrate that SNPs are not evenly distributed across the genome, but rather clump in high-diversity regions (Figure 5.1).

When examining where SNPs are found, a striking pattern emerges; genes that are comparatively upregulated in the parasitic adult female morph are more diverse than genes comparatively upregulated in the free-living adult female morph, with coding SNP density in the former being substantially higher than the genome-wide SNP density. Unsurprisingly therefore, parasitism genes were found to be over-abundant in the regions of the genome that contained the most SNPs compared with their abundance in the genome as a whole (Table 5.4). In contrast, coding SNP density in free-living genes was reduced compared to the genome-wide SNP-density, and free-living genes were under-represented in the most variable regions of the genome.

Strongyloides has several gene families that have expanded since its acquisition of parasitism, these being astacin-, CAP domain-containing protein- and acetylcholinesterase encoding genes, and this expansion is hypothesised to be associated with parasitism (Hunt *et al*. 2016). It is interesting then that genes of these families were over-represented in the most variable regions of the genome Table 5.4, Hunt *et al.* 2016).

Where multiple genes of similar sequence occur in tandem, such as in expansion clusters, the resulting repetitiveness of the region may complicate genome assembly so that not all copies of the gene are resolved in the reference genome assembly, and aligning reads to this could inflate apparent diversity. Furthermore, the existence of multiple tracts of very similar sequence, such as copies of a gene, can lead to misalignment of sequencing reads. Thus, for two related reasons, what is in fact diversity among copies of a gene after alignment may appear to be diversity within genes. To address the first issue, the genome assembly underlying expansion clusters was examined, and genes overlying potentially poorly-resolved regions were removed from the dataset. To address the second issue, the extent to which astacin and CAP domain-containing protein genes contributed to apparently high variation in parasitism genes was assessed. When these genes were removed from the list of genes in highly variable regions, the proportion of remaining genes that were upregulated in the parasitic morph was 55.8%, which although lower than the 69% observed when astacins and CAP domain-containing genes were included, is still well above the genome-wide proportion (11.8%). Thus, the observations of higher genetic diversity in parasitism genes appear to be real because this pattern persists (i) after potentially ambiguous genome

assembly regions are excluded and (ii) when multi-copy gene families are excluded, and so this is likely to be a real biological phenomenon

Why then, are parasitism genes more diverse than other genes in the *S. ratti* genome? The same pattern is not observed in the 'most free-living' genes, so it is not associated with stage-specific expression *per se*. One possibility is that the high copy numbers of genes in the major expanded gene families has led to redundancy, such that many of the genes are free to accumulate mutations, against which selection is low compared with genes across the genome more widely. However, as described above, the pattern persists even when genes in these gene families are removed. The remaining genes have a wide range of functional descriptions (Table 5.4), hence it is unlikely that redundancy fully explains the observed pattern. Further, comparison of the frequency of STOP-causing SNPs showed that these were not substantially more frequent in parasitism genes than other gene types in highly variable regions, nor were STOP-causing codons more frequent in CAP domain-containing- and astacin genes than other 'most parasitic' genes. This therefore does not support the hypothesis that redundancy accounts for the high diversity seen in parasitism genes.

It is likely that extensive genetic adaptation was required both for the initial switch to parasitism and then during the switch from facultative to obligate parasitism in the lineage leading to *Strongyloides* (Dorris *et al.* 2002, Hunt *et al.* 2016). Furthermore, *Strongyloides* spp. has subsequently diversified to infect a wide range of hosts (Grove 1989, Dorris *et al.* 2002). The finding of existing high genetic diversity within genes associated with parasitism suggests that selection on parasitism-associated traits may be ongoing. Hosts are genetically diverse, and it may be that some parasitism gene alleles perform better in some hosts than in others, such that diversity in the parasite population as a whole is favoured. Furthermore, hosts are always adapting to better resist parasites, and so parasites must continually adapt to better overcome host resistance (Van Valen 1973). Thus, it is possible that for *S. ratti*, specifically for the genes that are relevant to the parasitic stage of its life, the landscape of selection is constantly in flux, and evolution of traits associated with parasitism is ongoing.

It is common for genes encoding the surface proteins of protist parasites to be highly variable, either within species or among related parasitic species, and this is interpreted as a mechanism for immune evasion (e. g. Ferreira *et al.* 2004, Jackson 2016). However, it is rare for large numbers of parasitic helminths to be fully genome-sequenced and for genes putatively associated with parasitism to be tested, as done here, and it is not yet clear whether the phenomenon of parasitism-associated genes being more diverse is found in helminths other than *S. ratti*. In a study of two lines of the trematode *Schistosoma mansoni*, genes involved in parasitism were found to be less diverse than other genes (Clément *et al.* 2013). This result contrasts with what was observed in this study. This suggests a fundamental difference in the way parasitism is evolving between *S. mansoni* and *S. ratti*. For example,

in *S. ratti*, there may be selective pressure in favour of diversity at parasitism genes in order to cope with host diversity, but in *S. mansoni*, alleles particularly effective for colonising hosts regardless of host genotype may have arisen and initiated a soft selective sweep. The different host species (human/ mollusc *vs*. rat), different life cycles (single host in *S. ratti vs* multi-host in *S. mansoni*) and different genomic architectures may all contribute to this by exerting different selective pressures.

This work did not observe evidence for selective sweeps. Selective sweeps are hypothesised to explain the low global diversity in *Caenorhabditis elegans*, another nematode that is thought to reproduce predominantly without effective recombination (Andersen *et al.* 2012). This difference between these two species may be due to the difference in reproductive modes of the two nematode species. *C. elegans* reproduction is always sexual but is achieved predominantly by self-fertilisation so that heterozygosity is rapidly lost. In contrast, *S. ratti* reproduction in the population analysed predominantly relies on mitotic parthenogenesis, which tends to increase heterozygosity (discussed in detail in Chapter 4). Beneficial mutations required for selective sweeps may therefore rarely be present in homozygous form and this may reduce their effectiveness at improving fitness. Furthermore, it is widely understood that high homozygosity makes a genome less adaptable to different environmental conditions, and consequently makes populations less resilient (Charlesworth and Charlesworth 1987). It may be that high heterozygosity has an averaging effect on fitness, and that genomes containing a beneficial mutation are less able to supplant other genomes when those genomes are themselves diverse.

5.4.2 Expansion clusters

Another possible explanation for the high diversity of parasitism genes would be that, by chance, they occur in regions of the genome that are susceptible to mutation and /or the persistence of mutations. To test this, genes occurring in clusters of genes belonging to expanded gene families were compared to those in adjacent flanking regions. Across all clusters, mean coding SNP density per gene was more than 3-fold higher within clusters than in the genes of neighbouring flanking regions. As flanking regions are directly adjacent to expansion clusters and therefore in the same genomic environment, it is unlikely that a coincidental tendency for expanded gene family genes to consistently occur in genome regions susceptible to mutation explains the observed trend.

Another striking difference between genes in expansion clusters and flanking regions was the relative proportion of SNPs causing changes to amino acid sequence, with nonsynonymous SNPs being denser in expansion clusters (Figure 5.4). An increase in the rate of nonsynonymous SNPs could be interpreted as evidence for diversifying selection. This was not supported by dN/dS ratios, though these calculations were based on only a small subset of genes and may not be broadly applicable to expansion cluster / flanking region genes. Furthermore, the dN/dS analysis assumed nuclear clades 1 and 3 are fully reproductively isolated, which may not be the case. Hence, while it cannot be ruled out that in expansion

gene clusters specifically (as opposed to parasitism genes generally), redundancy among genes contributes to the high SNP density, diversifying selection remains another explanation compatible with the results. It may be that some expansion cluster genes are accumulating deleterious SNPs due to redundancy, while others are under diversifying selection.

A previous report of the clustering of expanded gene family genes within *S. ratti* (Hunt *et al.* 2016) suggested that expansion clusters arose by tandem gene duplication (Arguello *et al.* 2007), and that genes have stayed in tandem because the clustering aids in the co-regulation of these genes. In *Caenorhabditis elegans*, genes generated by tandem gene duplication tend to become separated by genome rearrangement, perhaps in response to selection pressure to reduce further errors in meiosis (Foth *et al.* 2014). For *S. ratti*, as most of the expansion cluster genes are upregulated in the parasitic female stage, co-regulation may be an efficient means of ensuring they are all expressed appropriately. Indeed, this view is supported by the fact that some expansion clusters (clusters 8 and 12) contain genes from multiple, different expanded gene families (Table 5.2) and this cannot have arisen only due to tandem gene duplication. However, while co-regulation may well be an important factor contributing to the maintenance of expanded gene family genes in tandem, it may not be the full explanation. The majority of CAP domain-containing protein- astacin- and acetylcholinesterase-encoding genes are not found in expansion clusters, and there are substantial differences in the extent of parasitism-specific expression of genes even within clusters. Analysis of regulatory regions in and around expansion clusters would be needed to confirm the extent to which expansion cluster genes are co-regulated.

The extent of the expansion of CAP domain-containing protein coding and astacin enoding genes in *Strongyloides*, and the fact that *Strongyloides* spp. differ in the number of the number of these genes they possess (Hunt *et al.* 2016), suggests that gene expansion may still be ongoing within *Strongyloides* spp. If so, this would cause intraspecific copy number variation within expanded gene families and would be further indication that parasitic traits are experiencing ongoing evolution. Identifying copy number variation was beyond the scope of this study, however the present dataset is sufficient to detect it, and this could be an important avenue of future research. If copy number variation is common, this may have further contributed to the high levels of diversity apparent within expansion cluster genes here, as sequencing reads derived from gene copies the reference assembly does not have may have aligned incorrectly.

5.4.3 Are ‘parasitism genes’ really involved in parasitism?

Throughout this chapter, genes have been called ‘parasitism genes’ if they are more strongly expressed in the parasitic female morph than the free-living morph, with a $\log_2$-fold change of 1 or higher (expression data retrieved from Hunt *et al.* [2016]). It has been assumed that these genes are involved in the parasitic lifestyle, for example by functioning in extraction of nutrients from the host or evasion

of host immune responses. However, there are many differences between the adult female morphs that are not directly related to parasitism. Free-living females are short-lived and reproduce sexually while parasitic females are long-lived and reproduce asexually, and there are substantial differences in morphology (Gardner *et al.* 2006). Furthermore, the two morphs inhabit very different environments; for example, the parasitic female lives in the warm environment of the host gut, while the free-living female must tolerate much colder temperatures in the external environment. It is therefore possible that a gene upregulated in the parasitic morph does not function directly in the parasitic lifestyle at all, but rather in longevity, mitotic parthenogenesis, heat tolerance or other traits not directly linked to parasitism. This should be borne in mind when drawing conclusions about the evolution of parasitism based on genes upregulated in the parasitic morph.

It may be difficult to unpick which genes are directly involved in parasitism and which have other parasitic morph-related functions, but comparisons across parasitic nematode taxa may be informative. Genes families that have expanded multiple times independently across parasitic nematode taxa, such as astacins and CAP-domain containing proteins (Tang *et al.* 2014, Hunt *et al.* 2017), are good candidates for direct involvement in parasitism. However, further work is required to determine the roles of acetylcholinesterase genes and other parasitism-associated genes in *S. ratti*.

5.4.4 Interaction between population genetics and diversity

SNP distribution among 10 kb windows showed similar patterns of SNP clustering within nuclear clade 1 individuals and nuclear clade 3 individuals (Figure 5.1). Two factors could contribute to this phenomenon: SNPs may have been present non-uniformly before the divergence of clades 1 and 3 and been maintained ever since; or both clades may have independently acquired SNPs in regions of the genome prone to accumulating mutations. This was tested by examination of private alleles in the 15 most variable 10 kb windows of the genome (Table 5.6). Highly variable patterns were found among windows – in some windows more alleles were shared, while in others most alleles were private to clade 1 or clade 3. High levels of private alleles in clade 3 were detected despite the fact that nuclear clade 3 is represented by fewer individuals, so that rare alleles are less likely to be detected and detected SNP densities are lower overall.

Thus, it appears that both mechanisms (retention of ancestral diversity and susceptibility of variable regions to mutation / mutational persistence) may contribute to the observed uneven distribution of SNPs across the *S. ratti* genome. In windows where clade-shared SNPs are most common, this is most likely diversity that is inherited from a common ancestor, while where diversity is most prominent in only one clade, this diversity likely arose after divergence. These hypotheses assume a lack of sexual crossing between these clades, however rare mating between them may have introduced SNPs that were derived in one clade into the genome of the other. Subsequent within-clade sexual reproduction and

associated recombination involving these ‘hybrids’ would be needed to limit the effect of this ‘hybridisation’ to certain regions of the genome. Windows with a particularly high number of private alleles in one clade or the other are potentially of interest, as these may represent cases of differential selection. Further study is warranted to understand the nature of that selection and the factors applying it.

Neighbour joining dendrograms were used to assess relationships of expansion clusters in different individuals, and to compare this with the genome-wide dendrogram (Figure 5.5). Broadly, the clades seen in the genome-wide dendrogram were detectable in the expansion clusters, but some rearrangements were noted, and these were inconsistent across clusters. Particularly, the placements of sub-clades 4a and 4b were very variable, and clade 5 sometimes grouped with clade 3 and other times was distinct. Furthermore, while clades 1 and 3 were generally well-delineated from each other, some individuals from one of these clades in the genome-wide dendrogram occasionally grouped with the other clade in expansion cluster dendrograms. Clades 5 and 4 are suggested to be the products of sexual reproduction among clades 1, 2 and 3, and this may explain their inconsistent placement in cluster-specific dendrograms. Different expansion clusters within clades may have been retained from different parental genotypes.

Most cluster-specific dendrograms take the form of an axis with clade 1 individuals predominantly at one end and clade 3 individuals predominantly at the other, with other individuals either grouping with one of these clades or appearing as intermediate between them. This is further evidence that, for the most part, the separation between clades 1 and 3 is biologically real. Because branch lengths within dendrograms are relative, the respective lengths of internal branches (both ends ending in a node) and external branches (one end ending in a tip) may inform on when mutations were principally acquired. In clusters where internal branch lengths are collectively long, such as clusters 7 and 8, most diversity was acquired during the divergence of clades 1 and 3. Where internal branch lengths are collectively short, most diversity was acquired more recently. Potentially, this may reflect the timing of selection pressures acting upon these clusters.

5.4.5 Conclusion

This study has investigated the distribution of single nucleotide polymorphisms (SNPs) in the genome of *Strongyloides ratti*. SNPs were found to be unevenly distributed across the genome, and regions of high density were consistent across clades of individuals detected in the population genetic analyses. Private SNP allele data suggested that both standing genetic diversity prior to divergence and subsequent accumulation of SNPs in certain regions of the genome may contribute to this pattern. A key finding was that genes associated with parasitism were more SNP-dense than other genes, and it is hypothesised that this is due to diversifying selection acting on parasitism-associated traits. Genes

belonging to gene families expanded in *Strongyloides* and clustered in the genome were found to be more diverse than the relevant flanking regions, and this was consistent across clusters in different parts of the genome so that seems unlikely that there is coincidental placement of expanded gene family genes in regions of the genome susceptible to mutation /mutation persistence. It is hypothesised that these genes are also under diversifying selection, though more work is required to understand the contribution of other factors.

5.5 References

Andersen E. C., Gerke J. C., Shapiro J. A., Crissman J. R., Ghosh R., Bloom J. S., Félix M.-A. *et al.* (2012). Chromosome-scale selective sweeps shape *Caenorhabditis elegans* genomic diversity. *Nature Genetics* **44:**285-290.

Anderson T., Nkhoma S., Ecker A. and Fidock D. (2010). How can we identify parasite genes that underlie antimalarial drug resistance? *Future Medicine* **12:**59-85.

Arguello J. R., Fan C., Wang W. and Long M. (2007). Origination of chimeric genes through DNA-level recombination. *Genome Dynamics* **3:**131-146.

Betts A., Gray C., Zelek M., MacLean R. C. and King K. C. (2018). High parasite diversity accelerates host adaptation and diversification. *Science* **360:**907-911.

Bonfield J.K. and Whitwham A. (2010). Gap5 - editing the billion fragment sequence assembly. *Bioinformatics* **26:**1699-1703.

Bradbury P. J., Zhang Z., Kroon D. E., Casstevens T. M., Ramdoss Y. and Buckler E. S. (2007). TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics* **23:**2633-2635.

Cantacessi C., Campbell B. E., Visser A., Geldhof P., Nolan M. J., Nisbet A. J., Matthews J. B. *et al.* (2009). A portrait of the "SCP/TAPS" proteins of eukaryotes — Developing a framework for fundamental research and biotechnological outcomes. *Biotechnology Advances* **27:**376-388.

Chao L., Hanley K. A., Burch C. L., Dahlberg C. and Turner P. E. (2000). Kin selection and parasite evolution: higher and lower virulence with hard and soft selection. *The Quarterly Review of Biology* **75:**261-275.

Charlesworth D. (2009). Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genetics* **2:**e64.

Charlesworth D. and Charlesworth B. (1987). Inbreeding depression and its evolutionary consequences. *Annual Review of Ecology and Systematics* **18:**237-268

Choi Y.-J., Bisset S. A., Doyle S. R., Hallsworth-Pepin K., Martin J., Grant W. N. and Mitreva M. (2017). Genomic introgression mapping of field-derived multiple-anthelmintic resistance in *Teladorsagia circumcincta*. *PLoS Genetics* **13:**e1006857.

Clément J. A. J., Toulza E., Gautier M., Parrinello H., Roquis D., Boissier J., Rognan A. *et al.* (2013). Private selective sweeps identified from next-generation pool-sequencing reveal convergent pathways under selection in two inbred *Schistosoma mansoni* strains. *PLoS Neglected Tropical Diseases* **7:** e2591

Culleton R., Martinelli A., Hunt P. and Carter R. (2005). Linkage group selection: Rapid gene discovery in malaria parasites. *Genome Research* **15:**92-97.

Danecek P., Auton A., Abecasis G., Albers C. A., Banks E., DePristo M., Handsaker R. E. (2011). The variant call format and VCFtools. *Bioinformatics* **27:**2156-2158.

Dorris M., Viney M. and Blaxter M. L. (2002). Molecular phylogenetic analysis of the genus *Strongyloides* and related nematodes. *International Journal of Parasitology* **32:**1507-1517.

Ekland E. H. and Fidock D. A. (2007). Advances in understanding the genetic basis of antimalarial drug resistance. *Current Opinion in Microbiology* **10:**363-370.

Ferreira M. U., da Silva Nunes M. and Wunderlich G. (2004). Antigenic diversity and immune evasion by malaria parasites. *Clinical and Vaccine Immunology* **11:**987-995.

Fisher O. and Schmid-Hempel P. (2005). Selection by parasites may increase host recombination frequency. *Biology Letters*. **1:**193-195.

Foth B. J., Tsai I. J., Reid A. J., Bancroft A. J., Nichol S., Tracey A., Holroyd N. *et al.* (2014). Whipworm genome and dual-species transcriptome analyses provide molecular insights into an intimate host-parasite interaction. *Nature Genetics* **46:**693-700.

Gardner M. P., Gems D. and Viney M. (2006). Extraordinary plasticity in aging in *Strongyloides ratti* implies a gene-regulatory mechanism of lifespan evolution. *Aging Cell* **5:**315-323.

Grove D. I. (1989). *Strongyloidiasis: A Major Roundworm Infection of Man*. London: Taylor and Francis.

Howe K. L., Bolt B. J., Shafie M., Kersey P and Berriman M. (2017). WormBase ParaSite − a comprehensive resource for helminth genomics. *Molecular and Biochemical Parasitology* **215:**2-10.

Hubbard T., Barker D., Birney E., Cameron G., Chen Y., Clark L., Cox T. *et al.* (2002). The Ensembl genome database project. *Nucleic Acids Research* **30:**38-41.

Huby F., Mallet S. and Hoste H. (1999). Role of acetylcholinesterase (AChE) secreted by parasitic nematodes on the growth of the cell line from epithelial origin HT29-D4. *Parasitology* **118:**489-498.

Hunt P., Martinelli A., Modrzynska K., Borges S., Creasey A., Rodrigues L., Beraldi D. *et al.* (2010). Experimental evolution, genetic analysis and genome re-sequencing reveal the mutation conferring artemisinin resistance in an isogenic lineage of malaria parasites. *BMC Genomics* **11:**499.

Hunt V. L., Tsai I. J., Coghlan A., Reid A. J., Holroyd N., Foth B. J., Tracey A. *et al.* (2016). The genomic basis of parasitism in the *Strongyloides* clade of nematodes. *Nature Genetics* **48:**299-307.

Hunt V. L., Tsai I. J., Selkirk M. E. and Viney M. (2017). The genome of *Strongyloides* spp. gives insights into protein families with a putative role in nematode parasitism. *Parasitology* **144:**343-358.

Jackson A. P. (2016). Gene family phylogeny and the evolution of parasite cell surfaces. *Molecular and Biochemical Parasitology* **209:**64-75.

Hussein A. S. Harel M. and Selkirk M. E. (2002). A distinct family of acetylcholinesterases is secreted by *Nippostrongylus brasiliensis*. *Molecular and Biochemical Parasitology* **123:**125-134.

James C. E., Hudson A. L. and Davey M. W. (2009). Drug resistance mechanisms in helminths: is it survival of the fittest? *Trends in Parasitology* **25:**328-335.

Koskella B. (2018). Resistance gained, resistance lost: An explanation for host–parasite coexistence. *PLoS Biology* **16:**e3000013.

Matoušková P., Vokřál I., Lamka J. and Skálová L. (2016). The role of xenobiotic-metabolizing enzymes in anthelmintic deactivation and resistance in helminths. *Trends in Parasitology* **32:**481-491.

McLaren W., Gil L., Hunt S. E., Singh Riat H., Ritchie G. R. S., Thormann A., Flicek P. *et al.* (2016). The Ensembl Variant Effect Predictor. *Genome Biology* **17:**122.

Mello L. V., O'Meara H., Rigden J. D. and Paterson S. (2009). Identification of novel aspartic proteases from *Strongyloides ratti* and characterisation of their evolutionary relationships, stage-specific expression and molecular structure. *BMC Genomics* **10:**611.

Park D. J., Lukens A. K., Neafsey D. E., Schaffner S. F., Chang H.-H., Valim C., Ribacke U. *et al.* (2012). Sequence-based association and selection scans identify drug resistance loci in the *Plasmodium falciparum* malaria parasite. *Proceedings of the National Academy of Sciences USA* **109:**13052-13057

R Core Team (2017). *A language and environment for statistical computing*. Retrieved from https://www.R-project.org/ 21/02/2019.

Rozas J., Ferrer-Mata A., Sánchez-DelBarrio J.C., Guirao-Rico S., Librado P., Ramos-Onsins S.E. and Sánchez-Gracia A. (2017). DnaSP 6: DNA Sequence Polymorphism Analysis of Large Datasets. *Molecular Biology and Evolution* **34:**3299-3302.

Sangster N. C., Bannon S. C., Weiss A. C., Nulf S. C., Klein R. D. and Geary T. G. (1999). *Haemonchus contortus*: sequence heterogeneity of internucleotide binding domains from p-glycoproteinsand an association with avermectin/milbemycin resistance. *Experimental Parasitology* **91:**250-257.

Sonnhammer E.L.L and Durbin R. (1995). A dot-matrix program dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* **167:**GC1-10.

Tang Y. T., Gao X., Rosa B. A., Abubucker S., Hollsworth-Pepin K., Martin J., Tyagi R. *et al.* (2014). Genome of the human hookworm *Necator americanus*. *Nature Genetics* **46:**261-269.

Van Valen L. (1993). A new evolutionary law. *Evolutionary Theory* **1:**1-30.

Williamson S. M., Storey B., Howell S., Harper K. M., Kaplan R. M. and Wolstenholme A. J. (2011). Candidate anthelmintic resistance-associated gene expression and sequence polymorphisms in a triple-resistant field isolate of *Haemonchus contortus*. *Molecular and Biochemical Parasitology* **180:**99-105.

Yang Z and Bielawski JP (2000). Statistical methods for detecting molecular adaptation. *Trends in Ecology and Evolution* **15:**496–503.

6. Discussion and final conclusions

6.1 Discussion

6.1.1 Summary of findings

In this work, the parasitic nematode *Strongyloides ratti* was sampled non-invasively from wild populations of its natural host, the brown rat (*Rattus norvegicus*) in the UK. *S. ratti* was found to be a common parasite overall, though the frequency and intensity of infection was variable among the three sampling sites and four sampling seasons (Table 2.2). The whole genomes of 90 worms were individually sequenced, and these sequences were used to investigate *S. ratti* population genetic structure. Sequenced worms were collected from 39 hosts sampled from all three sampling sites, allowing for comparisons of worms among hosts and among sites. The principal population genetic finding was the discovery of multiple genetically distinct clades, which occurred in sympatry such that the three largest clades were present in all three sampling sites (Figures 4.5 and 4.6). This contrasted with the population genetics of rats, which was examined by microsatellite analysis of DNA extracted from the same faecal pellets used to isolate *S. ratti*. In rats, there was weak population genetic structuring among sampling sites, with no evidence for sympatric, but genetically distinct clades, as seen in *S. ratti*. The distribution of diversity within the genome was also investigated, and it was observed that diversity was not evenly distributed across the genome. The same high genetic diversity regions were seen within both clade 1 and clade 3 and private alleles were common, suggesting that some regions of the genome are more susceptible to mutation accumulation than others. Strikingly, it was observed that genes previously identified as having a putative role in parasitism (Hunt *et al.* 2016) were more variable than other genes.

6.1.2 Drivers of *Strongyloides ratti* population genetics

6.1.2.1. Life history

The life cycle of *S. ratti* incorporates obligatory mitotic parthenogenesis with facultative sexual reproduction (Figure 1.1). However, despite extensive sampling, the sexual morph of the species is almost never detected within the UK (Viney *et al.* 1992, Fisher and Viney 1998, Chapter 2 here). Thus, it is expected that sexual reproduction is very rare in UK *S. ratti*, and that the vast majority of individuals are produced by mitotic parthenogenesis. This preponderance of mitotic parthenogenesis is likely critical for the maintenance of genetically distinct clades in sympatry, as clades would be rapidly homogenised if sexual reproduction was common. Thus, it is hypothesised that three clades identified by population genetic analysis, (called 1, 2b and 3, see Figure 4.5) were each founded by a different single parasitic female and have since propagated asexually.

However, some of the *S. ratti* individuals and clades appeared to be genetically intermediate between others, suggesting that there may have been sexual crossing among clades. Once an individual is

produced by sexual reproduction, that individual can subsequently reproduce asexually and thereby amplify its recombinant genome, as observed in clade 5. Diversity within clades 1, 2b and 3 may also be in part due to sexual crosses with other clades followed by repeated backcrossing with one of the parental strains. Thus, the observed population genetics of *S. ratti* appears to hinge upon its unusual life cycle, which features extensive asexual parthenogenesis with rare sexual reproduction.

Most of the literature on parasitic nematode population genetics (reviewed in Cole and Viney 2018 and Chapter 1) concerns species that are dioecious, and therefore do not have the unusual life history of *S. ratti*. It is therefore unsurprising that population genetic patterns similar to that observed for *S. ratt*i have rarely been reported. *S. ratti* population genetics has been studied before, but using only three loci, and these are unlikely to provide sufficient resolution to detect clades been as here. However, the population genetics of *Strongyloides stercoralis*, a parasite of humans and canines, has been studied in some depth. Such studies have identified two lineages of *S. stercoralis* in Southeast Asia, one specific to dogs and the other found in both humans and dogs (Jelata *et al.* 2017, Nagayasu *et al.* 2017), leading to the suggestion that these lineages might be different species (Jelata *et al.* 2017), but otherwise evidence for multiple sympatric clades is not apparent. However, different *Strongyloides* spp. (and indeed different populations of *S. ratti*) differ in their rates of sexual reproduction, and extensive production of the sexual *Strongyloides* forms has been observed in multiple populations of *S. stercoralis* (Kikuchi *et al.* 2016, Jelata *et al.* 2017, Nagayasu *et al.* 2017). Thus, sexual reproduction might be more common in *S. stercoralis* than the *S. ratti* populations observed, explaining the different population genetic structures. Further studies into *Strongyloides* population genetics, comparing populations with low and high rates of sexual reproduction, are needed to better understand how asexual reproduction interacts with population genetic structure. In particular, these studies must sample densely from sampling sites in order to detect rare clades that may be present.

6.1.2.2. Host movement ecology and infection dynamics

While the predominance of parthenogenetic mitosis in *S. ratti* reproduction explains why sympatric clades are not homogenised, it is nevertheless surprising that multiple clades occur at high frequencies within sampling sites. In the absence of effective recombination, the entire genome acts as one linkage block, such that selection on any allele could theoretically drive the entire genomic background it appears in to fixation (Andersen *et al.* 2012). Selective sweeps such as these are thought to be why *Caenorhabditis elegans'* global diversity has been found to be low (Andersen *et al.* 2012). Because there are likely to be differences in fitness among *S. ratti* clades, one would expect the fittest clade to increase in frequency and approach fixation, but this is not observed.

It may be that the high heterozygosity observed in this study (Chapter 4) means that selective sweeps are less powerful (discussed in Chapter 5). Furthermore, the high frequency and intensity of *S. ratti*

infection observed in Chapter 2 suggest that the *S. ratti* population size may be large. In large populations the rate of allele frequency change (or in the case of *S. ratti*, clade frequency change) is slow compared with small populations (Hartl and Clark 1997), thus selective sweeps may be slow in *S. ratti*. However, a non-mutually exclusive hypothesis is that different clades are most fit in different locations, and the movement of host rats allows for *S. ratti* gene flow that normalises the clade frequencies among locations. Chapter 3 studied the population genetics of the rat hosts in-depth and found evidence for moderate rat gene flow among sites, supporting this hypothesis.

Thus, analysing the host population genetics and parasite infection dynamics has provided added insight into the population genetics of the parasite *S. ratti*. Future population genetic studies of parasites would likely also benefit from analysing these features.

6.1.3 Population genetics and ongoing evolution in *Strongyloides ratti*

The population genetics of a species is determined by a combination of genetic change within populations and gene flow among populations, and these processes are also pivotal in population-level evolution (Kimura 1969, Real 1994, Lässig *et al.* 2017). For example, when population sizes are small, the effect of random genetic drift leads to rapid changes in allele frequencies and hence rapid evolution within populations. Thus small populations experience rapid divergence among populations so that population genetic structuring emerges. Similarly, when gene flow is high, sub-populations are genetically homogenised. This reduces population genetic structuring, but also hampers local adaptation due to the influx of comparatively less suited alleles from other sub-populations (Lenormand 2002). Thus, studying the genetic structure of a population can give context to its evolution.

In this thesis, the distribution of single nucleotide polymorphisms (SNPs) within the *S. ratti* genome was studied, and these data were used to analyse ongoing natural selection in *S. ratti* in the context of its population genetics. Population genetic analyses have revealed that *S. ratti* populations are structured more strongly at the level of sympatric, but genetically distinct, clades that are geographically widespread. A notable finding of SNP distribution analysis was that the SNP density of genomic regions was consistent across the two largest clades (clades 1 and 3), such that regions that had high SNP density in one clade also had SNP variant density in the other (Figure 5.1). Furthermore, when the most SNP-dense regions of the genome were examined in more detail, it was observed that many alleles were private either to clade 1 or clade 3 (Table). This reveals that these variable regions have accrued SNPs independently since these clades arose (Chapter 5), suggesting that such regions may have high mutation rates, be largely free from purifying selection or be under diversifying selection. If the population genetic structure of *S. ratti* had not been resolved prior to analysis of within-genome SNP-density, and sampling sites had instead been treated as population sub-groupings, individuals from multiple clades may have been pooled, and this finding of extensive private alleles missed.

Another key finding was that genes putatively involved in parasitism (Hunt *et al.* 2016) tended to occur in highly SNP-dense regions of the genome and were themselves more SNP-dense than other genes. Furthermore, the *S. ratti* genome contains multiple genes belonging to gene families that are expanded in *Strongyloides* spp. (Hunt *et al.* 2016), and regions of the genome that contain several such genes ('expansion clusters') were more diverse than flanking regions. This may reflect diversifying selection acting on parasitism-associated traits, as discussed in Chapter 5. Notably, when neighbour-joining dendrograms were generated for each of these expansion clusters, different clusters gave similar, but different patterns. Clades 1 and 3 remained mostly distinct, but the positioning of other clades and individuals varied (Figure5.5). One interpretation of this is that rare sexual crossing among and within clades has led to recombination so that in some individuals, different expansion clusters originate in different clades. Thus, analysing the distribution of diversity within clades has provided evidence that supports the occasional sexual reproduction suggested by principle component analysis as part of population genetic analyses (Figure 4.6), clarifying life history traits and helping to explain how the pattern of population genetic structure observed arose.

6.1.4 Whole genome sequencing for population genetic analysis

Population genetic analyses of parasitic nematodes are often based on non-sequence genetic data such as microsatellite repeat length (e.g. Zarlenga *et al.* 1996, Johnson *et al.* 2006, la Rosa *et al.* 2012), or sequencing data for single genes or regions of the genome such as internal transcribed spacers (Powers *et al.* 1997, Lin *et al.* 2012, Zhao *et al.* 2012). Sequencing of mitochondrial DNA is also commonly used (e.g. Otranto *et al.* 2005, Tang and Hyman 2007, Wu *et al.* 2009, Monte *et al.* 2012, Xie *et al.* 2014, Sheng *et al.* 2015). The use of whole genome sequencing differs from these techniques by interrogating virtually every variable site in the genome, and therefore providing extremely high-resolution of population genetic structure. The present study generated neighbour-joining dendrograms based on whole-genome data and expansion clusters and observed inconsistent placement of some individuals in the latter. Specifically, some clades that were distinct in the whole-genome dendrogram grouped with individuals of other clades in some expansion clusters. Thus, if any one small genomic region (expansion cluster) had been used as a basis of population genetic study, some of the diversity of clades within the samples *S. ratti* populations would have been missed.

Nevertheless, cases of whole-genome sequencing being used to test the population genetics of parasitic nematodes remain rare in the literature (Cole and Viney 2018), and the primary limiting factor is likely to be cost. In particular, population genetic studies require that the genomes of multiple individuals be sequenced, and this increases cost substantially. However, the small size of the *S. ratti* genome (approximately 41 Mb) alleviated this difficulty here. More challenging in this study was the physically

small size of individual *S. ratti* infective larvae, which made acquiring sufficient DNA for good genome coverage difficult. Indeed, despite extensive washing, in some sequencing libraries generated from individual *S. ratti* larvae, less than 5% of sequencing reads aligned to the *S. ratti* reference genome (Chapter 4), suggesting that very little *S. ratti* DNA was present among contamination from other organisms. To combat this, it was necessary to make libraries for many individuals (225) and sequence these libraries shallowly in order to determine their *S. ratti* content, such that a small number (90) of those with the highest *S. ratti* content could be sequenced again at greater depth. While this is an effective method of sequencing the whole genomes of multiple individuals when DNA availability is limited, it is expensive and is dependent on having samples to spare. Thus, this method may not be suitable for all parasitic nematode population genetic studies. An alternative is whole genome amplification (Hosono *et al.* 2003) prior to sequencing, as used to investigate the population genetics of *S. stercoralis* (Kikuchi *et al.* 2016). However whole genome amplification can lead to biased results (Pinard *et al.* 2006).

Nevertheless, whole genome sequencing for population genetics is already possible even when individual animals contain little DNA. As sequencing technologies improve and costs continue to decline, such studies will become feasible in an increasingly broad range of population genetic contexts.

6.2 Future Directions

This study examined *S. ratti* individuals over a small area of the UK (all sampling sites within 32 km of each other). Thus, it is not clear to what extent the genetically distinct clades identified in this work are geographically widespread, nor whether there are other clades in other geographic regions. Ten *S. ratti* laboratory lines isolated from elsewhere in the UK as well as Japan were examined, and were placed within clades detected among individual worms (Figures 4.5 and 4.6). This result suggests that clades are widespread, but more extensive sampling is required. If more was known about the population genetics of *S. ratti* on a global scale, it might be possible to form hypotheses on the origin of genetic clades and their historical spread. Similar work has already been performed for brown rats (Puckett *et al.* 2016), and it would be very interesting to see whether historical spread of parasite clades mirrors the spread of host genotypes, as observed in other parasitic nematode – rodent host systems (Nieberding *et al.* 2004).

There are many outstanding questions regarding diversity and selection within *S. ratti* genomes, particularly with respect to the evolution of traits involved in parasitism and the associated genes. For example, while it is clear that genes associated with parasitism are more variable than other genes, and this is compatible with a hypothesis of diversifying selection acting upon these, it is not yet clear whether or how this diversity directly interacts with fitness. In depth analyses of individual genes may be required to associate specific mutations with effects on gene function. Another outstanding question is whether the gene families expanded in *Strongyloides* spp. are still undergoing expansion and are therefore variable within *S. ratti*. Given that *S. ratti* and its close relative *S. stercoralis* differ in the number of these genes, intra-individual variation is likely. The methods used in this thesis were not designed to detect copy number variation, however several methods compatible with the DNA sequences generated here are available (e.g. Wang *et al.* 2007, Xi *et al.* 2011).

6.3 Final Conclusions

This study found that *Strongyloides ratti* is a common infection of brown rats (*Rattus* norvegicus) in the UK. The population genetics of *S. rattti* is structured on the level of genetically distinct clades that occur in sympatry, with each clade occurring in multiple sampling sites. This contrasts strongly with the population genetics of the hosts, which was weakly structured on the basis of geography. The unusual population genetic structure of *S. ratti* is likely only possible due to a preponderance of asexual mitotic parthenogenesis, which prevents homogenisation among clades. It is hypothesised that the three major clades each originated asexually from a different parasitic female and subsequently expanded asexually, acquiring diversity by mutation. Other clades and individuals are hypothesised to be the result of occasional sexual reproduction among clades, followed by further asexual reproduction.

Genes that are upregulated in the parasitic female morph are putatively involved in the parasitic lifestyle, and these genes were found to be more variable than other genes. This may indicate that parasitism as a trait is under diversifying selection.

6.4 References

Andersen E. C., Gerke J. C., Shapiro J. A., Crissman J. R., Ghosh R., Bloom J. S., Félix M.-A. *et al.* (2012). Chromosome-scale selective sweeps shape *Caenorhabditis elegans* genomic diversity. *Nature Genetics* **44:**285-290.

Cole R. and Viney M. (2018). The population genetics of parasitic nematodes of wild animals. *Parasites and Vectors* **11:**590.

Fisher M. C. and Viney M. E. (1998). The population genetic structure of the facultatively sexual parasitic nematode *Strongyloides ratti* in wild rats. *Proceedings of the Royal Society B: Biological Sciences*. **265:**703–709.

Hartl D. L. and Clark A. G. (1997). *Principles of Population Genetics.* 4th ed. Sunderland: Sinauer.

Hosono S., Faruqi A. F., Dean F. B., Du. Y., Sun Z., Wu X., Du J. *et al.* (2003). Unbiased whole-genome amplification directly from clinical samples. *Genome Research* **13:**954-964.

Hunt V. L., Tsai I. J., Coghlan A., Reid A. J., Holroyd N., Foth B. J., Tracey A. *et al.* (2016). The genomic basis of parasitism in the *Strongyloides* clade of nematodes. *Nature Genetics* **48:**299-307.

Jelata T. G., Zhou S., Bemm F. M., Schär F., Khieu V., Muth S., Odermatt P. *et al.* (2017). Different but overlapping populations of *Strongyloides stercoralis* in dogs and humans—Dogs as a possible source for zoonotic strongyloidiasis. *PLoS Neglected Tropical Disease* **11:**e0005752.

Johnson P. C. D., Webster L. M. I. Adam A., Buckland R., Dawson D. A. and Keller L. F. (2006). Abundant variation in microsatellites of the parasitic nematode *Trichostrongylus tenuis* and linkage to a tandem repeat. *Molecular and Biochemical Parasitology* **148:**218-218.

Kikuchi T., Hino A., Tanaka T., Aung M. P. P. T. H. H. A., Afrin T., Nagayasu E., Tanaka R. *et al.* (2016). Genome-wide analyses of individual *Strongyloides stercoralis* (Nematoda: Rhabditoidea) provide insights into population structure and reproductive life cycles. *PLoS Neglected Tropical Diseases* **10:**e0005253.

Kimura M. (1969). The rate of molecular evolution considered from the standpoint of population genetics. *PNAS* **63:**1181-1188.

La Rosa, G., Marucci G., Rosenthal B. M. and Pozio E. (2012). Development of a single larva microsatellite analysis to investigate the population structure of *Trichinella spiralis*. *Infection, Genetics and Evolution* **12:**369-376.

Lässig M., Mustonen V. and Walczak A. M. (2017). Predicting evolution. *Nature Ecology and Evolution* **1:**0077.

Lenormand T. (2002). Gene flow and the limits to natural selection. *Trends in Ecology and Evolution* **17:**183-189.

Lin Q., Li H. M., Gao M., Wang X. Y., Ren W. X., Cong M. M., Tan X. C. (2012). Characterization of *Baylisascaris schroederi* from Qinling subspecies of giant panda in China by the first internal transcribed spacer (ITS-1) of nuclear ribosomal DNA. *Parasitology Research* **110:**1297-1303.

Monte T. C. C., Simões R. O., Oliveira A. P. M., Novaes C. F., Thiengo S. C., Silva A. J., Estrela P. C. *et al.* (2012). Phylogenetic relationship of the Brazilian isolates of the rat lungworm *Angiostrongylus cantonensis* (Nematoda: Metastrongylidae) employing mitochondrial *COI* gene sequence data. *Parasites and Vectors* **5:**248.

Nagayasu E., Aung M. P. P. T. H. H. A., Hortiwakul T., Tanaka A. H. T., Higashiarakawa M., Olia A., Taniguchi T. *et al.* (2017). A possible origin population of pathogenic intestinal nematodes, *Strongyloides stercoralis*, unveiled by molecular phylogeny. *Scientific Reports* **7:**4844.

Nieberding C., Morand S., Libois R., Michaux J. R. (2004). A parasite reveals cryptic phylogeographic history of its host. *Proceedings of the Royal Society of London B: Biological Sciences* **271:**2559–2568.

Otranto D., Testini G., de Luca F., Hu M., Shamsi S., Gasser R. B. (2005). Analysis of genetic variability within *Thelazia callipaeda* (Nematoda: Thelazioidea) from Europe and Asia by sequencing and mutation scanning of the mitochondrial *cytochrome c oxidase subunit 1* gene. *Molecular and Cellular Probes* **19:**306-313.

Pinard R., de Winter A., Sarkis G. J., Gerstein M. B., Tartaro K. R., Plant R. N., Egholm M. (2006). Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics* **7:**216.

Powers T. O., Todd T. C., Burnell A. M., Murray P. C. B., Fleming C. C., Szalanski A. L., Adams B. A. *et al.* (1997). The rDNA internal transcribed spacer region as a taxonomic marker for nematodes. *Journal of Nematology* **29:**441-450.

Puckett E. E., Park J., Combs M., Blum M. J., Bryant J. E., Caccone A., Costa F., *et al.* (2016). Global population divergence and admixture of the brown rat (*Rattus norvegicus*). *Proceedings of the Royal Society B: Biological Sciences*. **283:**20161762.

Real L. (1994). *Ecological Genetics.* Princeton: Princeton University Press.

Sheng L., Cui P., Fang S.-F., Lin R.-Q., Zou F.-C. and Zhu X.-Q. (2015). Sequence variability in four mitochondrial genes among rabbit pinworm (*Passalurus ambiguus*) isolates from different localities in China. *Mitochondrial DNA* **26:**501-504.

Tang S. and Hyman B. C. (2007). Mitochondrial genome haplotype hypervariation within the isopod parasitic nematode *Thaumamermis cosgrovei*. *Genetics* **176:**1139-1150.

Viney M., Matthews B. E. and Walliker D. (1992). On the biological and biochemical nature of cloned populations of *Strongyloides ratti*. *Journal of Helminthology* **66:**45-52.

Wang K., Li M., Hadley D., Liu R., Glessner J., Grant S. F. A., Hakonarson H. *et al.* (2007). PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research* **17:**1665-1674.

Wu S. G., Wang G. T., Xi B. W., Xiong F., Liu T. and Nie P. (2009). Population genetic structure of the parasitic nematode *Camallanus cotti* inferred from DNA sequences of ITS1 rDNA and the mitochondrial COI gene. *Veterinary Parasitology* **164:**248-256.

Xi R., Hadjipanayis A. G., Luquette L. J., Kim T.-M., Lee E., Zhang J., Johnson M. D. (2011). Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *PNAS* **108:**E1128-E1136.

Xie Y., Zhou X., Zhang Z., Wang C., Sun Y., Liu T., Gu X. *et al.* (2014). Absence of genetic structure in Baylisascaris schroederi populations, a giant panda parasite, determined by mitochondrial sequencing. *Parasites and Vectors* **7:**1591.

Zarlenga D. S., Aschenbrebber R. A. and Lechtenfels J. R. (1996). ariations in microsatellite sequences provide evidence for population differences and multiple ribosomal gene repeats in *Trichinella pseudospiralis*. *Jorunal of Parasitology* **82:**306-313.

Zhao G. H., Li H. M., Ryan U. M., Cong N. M., Hu B., Gao M., Ren W. X. *et al.* (2012). Phylogenetic study of *Baylisascaris schroederi* isolated from Qinling subspecies of giant panda in China based on combined nuclear 5.8S and the second internal transcribed spacer (ITS-2) ribosomal DNA sequences. *Parasitology International* **61:**497-500.

Appendix 1: Examination of expansion clusters and associated flanking regions to determine quality of underlying reference assembly

Coordinates for each expansion cluster and flanking region are given in Table 5.1. Definitions of 'expansion cluster' and 'flanking region' along with the rationale for examining the quality of their underlying assembly, are given in Section 5.2.2.2. To assess quality, sequencing reads originally used to build the reference assembly (freely available at NCBI under BioProject code PRJEB2398, Hunt *et al.* 2016) were aligned back to the reference assembly, and the quality of these alignments was assessed in the software package Gap5 (Bonfield and Whitman 2010). Gap5 plots were generated by Alan Tracey of the Sanger Institute and kindly provided as a Personal Communication.

In the diagrams below, Gap5 outputs have a black background and are split into two by a white line. In both portions of the plot, the X-axis is genomic position and extends from the start of the left flanking region to the end of the right flanking region. The top portion of the output plots all mate pairs of sequencing reads (reads used to build the reference assembly) aligning to the region of interest as a horizontal bar. The size of the bar is defined by the size of the reads themselves plus the distance between them (or the size of the read alone if without a mate) and is henceforth referred to as an insert. The Y-axis in this portion of the output is mate pair size, with smaller reads appearing higher in the plot. Colours of inserts have various meanings as follows:

- Blue: Unpaired read; no mate present
- Orange: Paired read where mate is on a different assembly contig
- Grey to white: Properly paired insert. Intensity of colour reflects mapping quality, with whiter inserts mapping with greater confidence.
- Red: Paired inserts where mate pairs are in the unexpected orientation with respect to one another.

Paired inserts that are very long indicate misalignment such that mate pairs have aligned to sequences distant from one another, and thereby indicate repetitiveness in the underlying sequence. Red inserts similarly indicate repetitiveness, except that the repeats are in the reverse orientation. Both phenomena therefore reduce confidence in assembly quality.

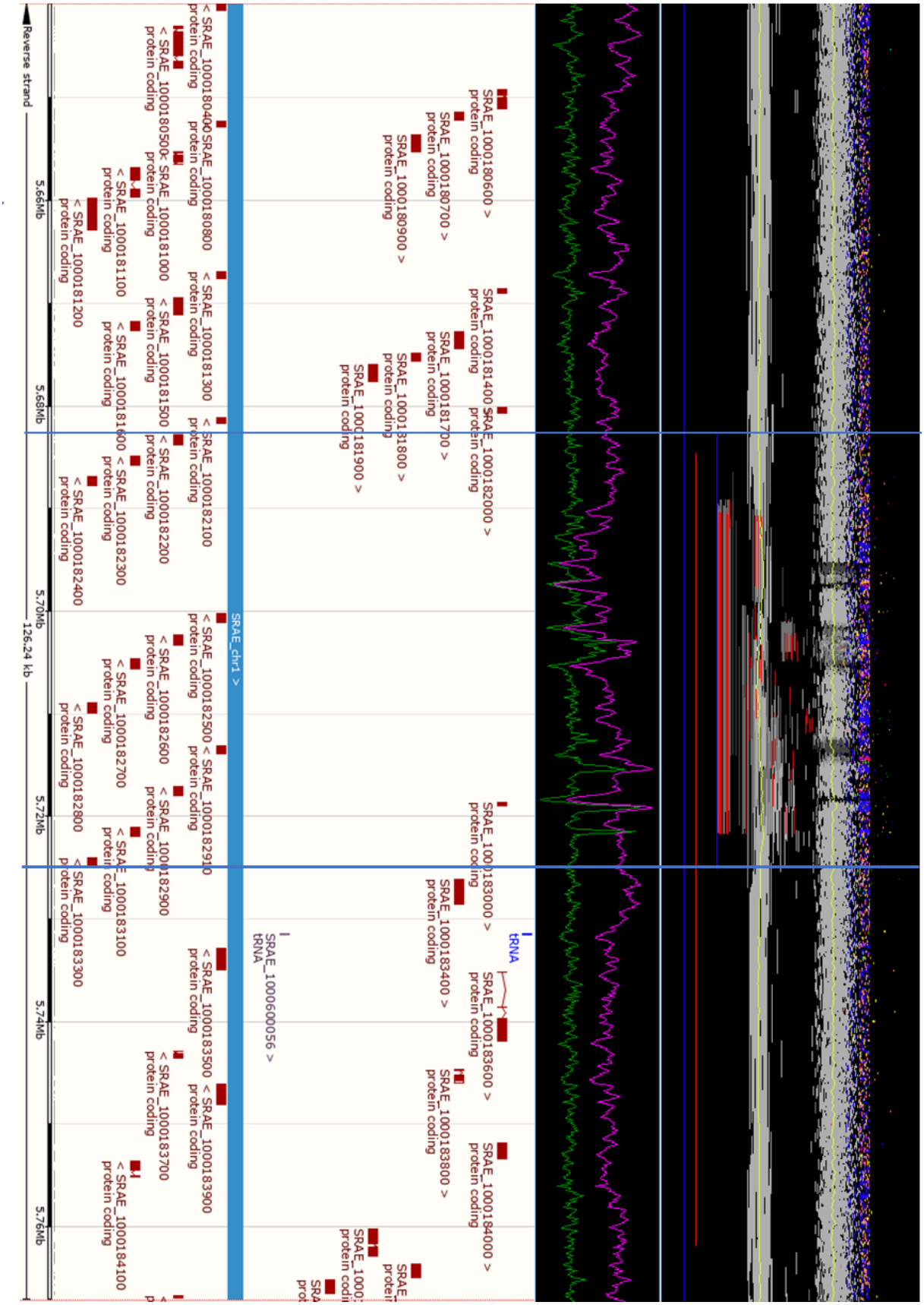
The lower portion of the Gap5 output plots shows read depth (green) and insert depth (depth to which sequence is covered by read or space between mate pairs, purple). The Y-axis is read/insert depth, with greater depth values appearing higher. In a perfect assembly, read depth and insert depth would both be flat. However, peaks in either indicate collapsed assembly regions where repetitive sequences have not been fully resolved.

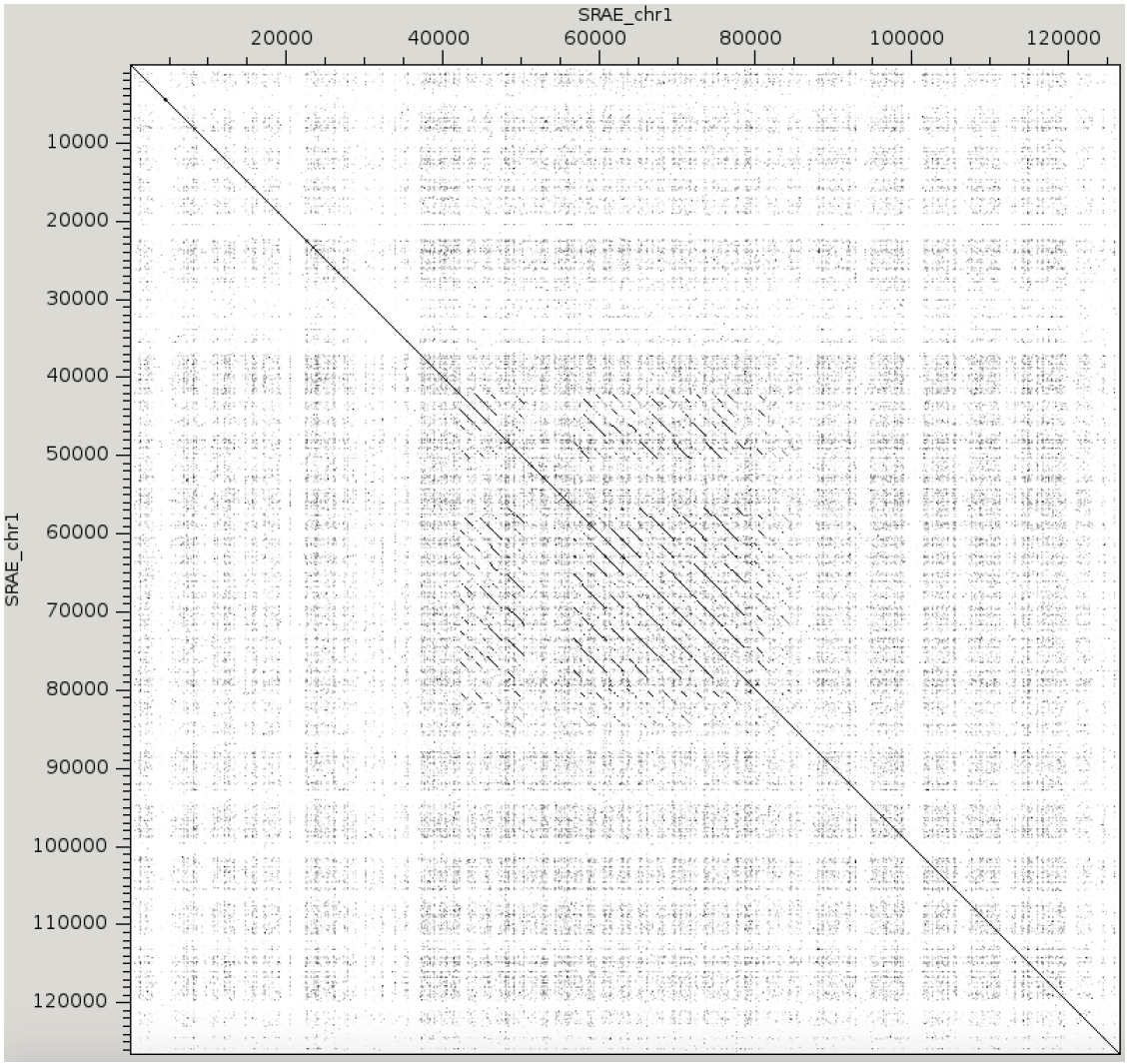
Gene annotation schematics were retrieved from Ensembl's 'Region in Detail' tool (Hubbard *et al.* 2002), accessed via WormBase Parasite (Howe *et al.* 2017) version 12 (https://parasite.wormbase.org/index.html), and are shown below the Gap5 output for each expansion cluster. X-axis is the same as that used in Gap5 plots. Alternating black and white bars on the X-axis indicate 10 kb. Red bars indicate genes. Genes shown above the central horizontal blue bar are encoded on the forward strand, while those below this bar are encoded on the reverse strand. Vertical blue lines that cross Ensembl and Gap5 outputs split the entire graphic into three vertical segments. The leftmost segment is the left flanking region, the central segment is the expansion cluster, and the rightmost segment is the right flanking region.

Finally, a broad view of repetitiveness within each region of interest is given by the dotplots shown with each Gap5-Ensembl graphic. Alan Tracey of the Sanger Institute generated these dotplots with the software package Dotter (Sonnhammer and Durbin 1995) and kindly provided them for use in this study as a Personal Communication. These dotplots compare each nucleotide base in one sequence with each nucleotide base in another, and where the same base is observed a black dot is printed. Thus, runs of similar sequence emerge as diagonal black lines. In the plots presented here, the X and Y axes both span the same genomic length as the associated Gap5-Ensembl graphic, such that the entire region is compared with itself, and diagonal black lines represent repetitive sequences.

Assembly was deemed to be of poor quality if mapping quality was poor, there were many very long inserts, there were many inserts with mate pairs in the unexpected orientation and / or if there was a spike in read and/or fragment depth. Genes on such regions were discarded from analysis in Chapter 5. Where all genes within an expansion cluster were discarded, the associated flanking regions of that expansion cluster were also discarded.

Cluster 1


Reverse strand
5.66Mb
5.68Mb
5.70Mb
126.24 kb
5.72Mb
5.74Mb
5.76Mb
SRAE_chr1 >
tRNA
SRAE_1000600056 >
tRNA



Genes discarded from expansion cluster:

SRAE_1000182400

SRAE_1000182500

SRAE_1000182600

SRAE_1000182700

SRAE_1000182800

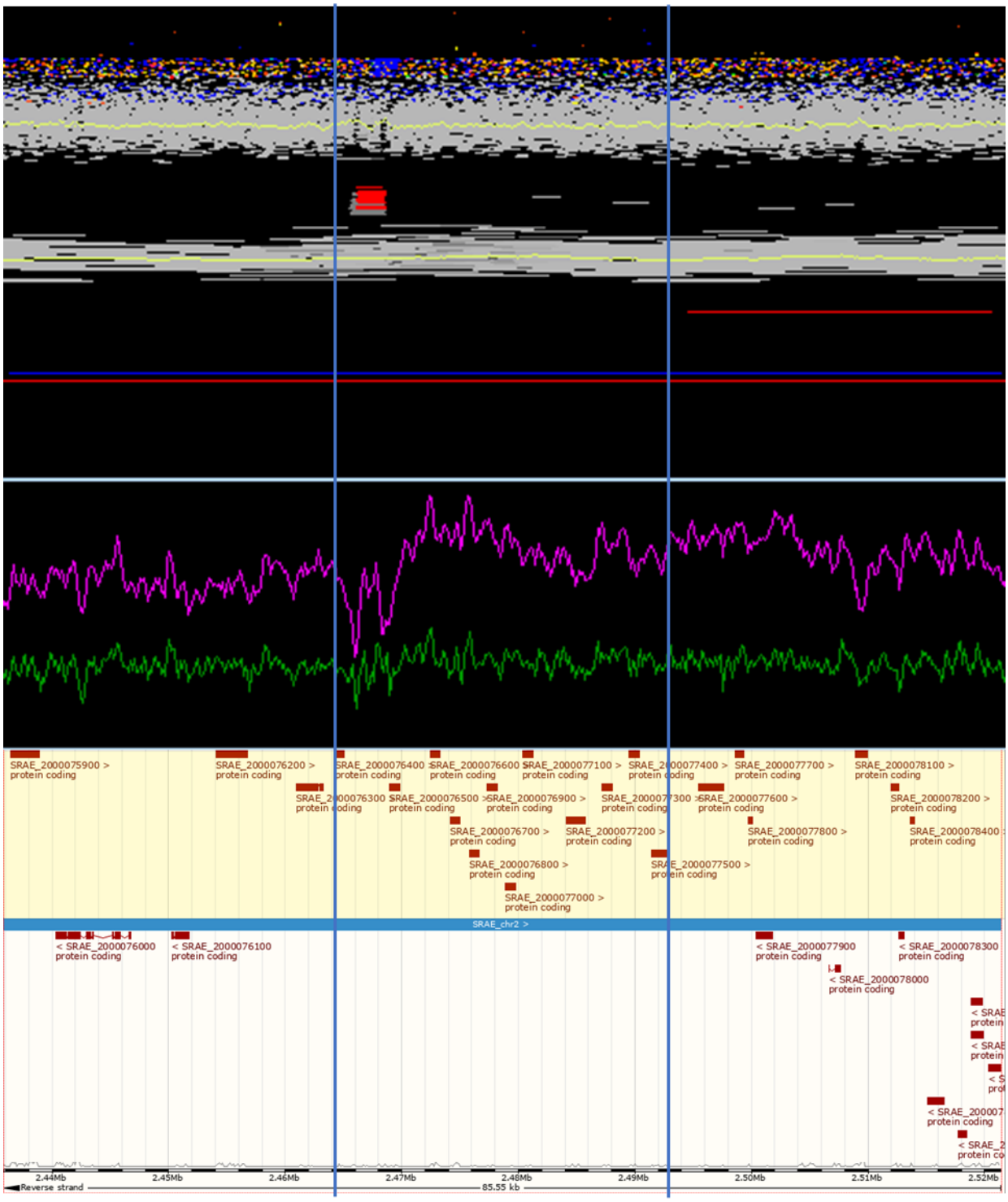
SRAE_1000182910

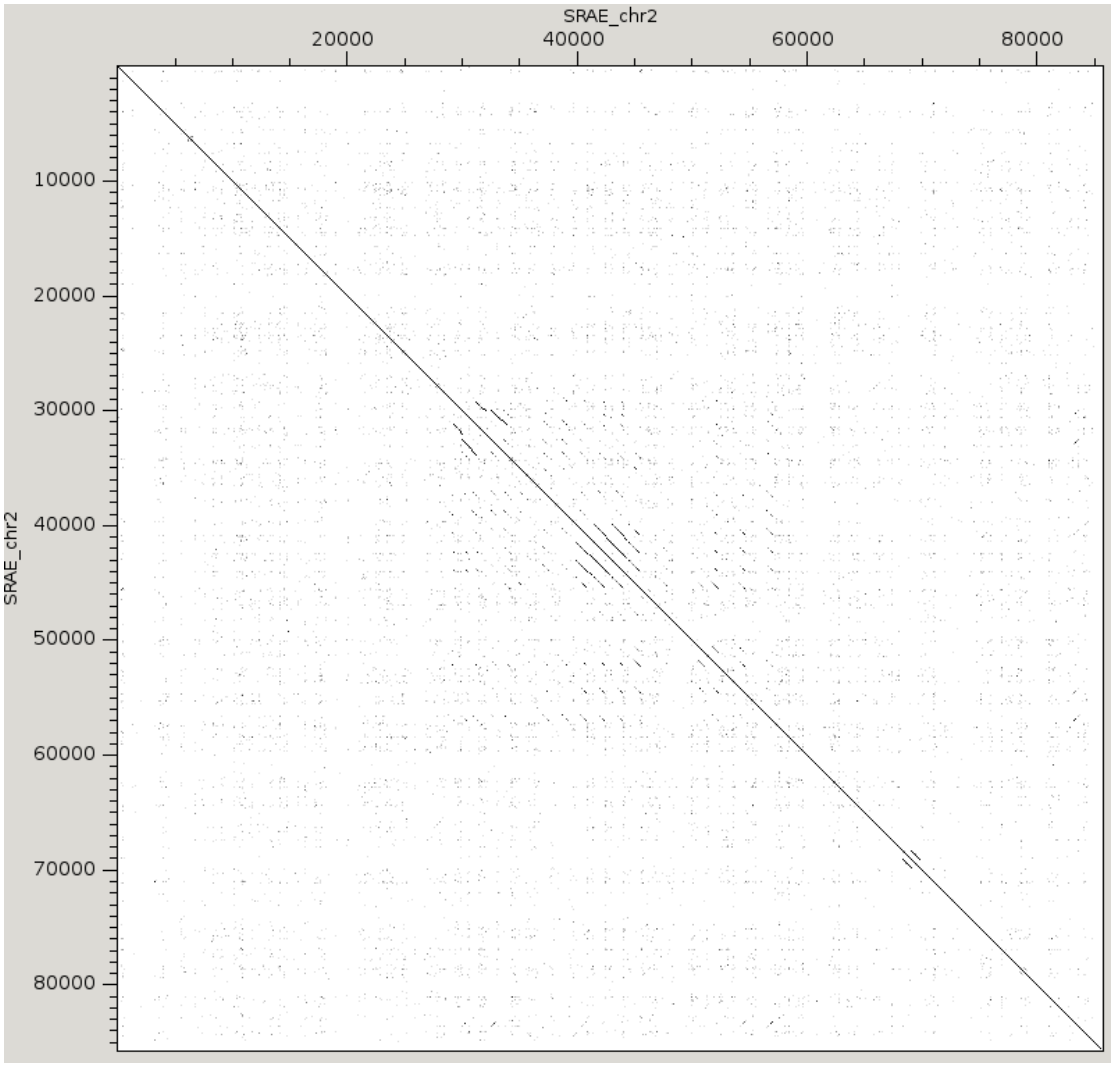
SRAE_1000182900

SRAE_1000183000

SRAE_1000183100

Cluster 2

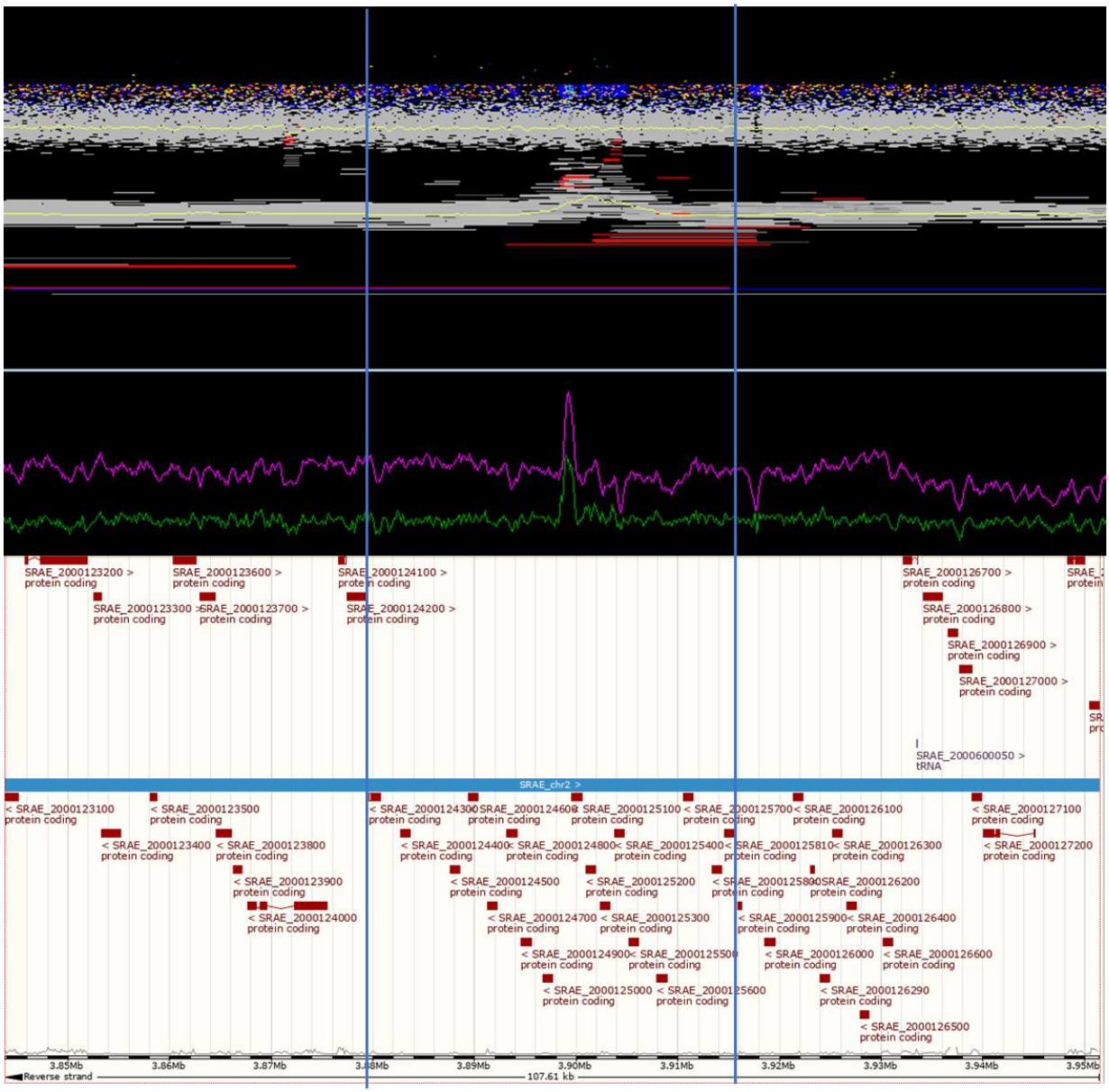

SRAE_2000075900 > protein coding
SRAE_2000076200 > protein coding
SRAE_2000076400 > protein coding
SRAE_2000076600 > protein coding
SRAE_2000077100 > protein coding
SRAE_2000077400 > protein coding
SRAE_2000077700 > protein coding
SRAE_2000078100 > protein coding
SRAE_2000076300 > protein coding
SRAE_2000076500 > protein coding
SRAE_2000076900 > protein coding
SRAE_2000077300 > protein coding
SRAE_2000077600 > protein coding
SRAE_2000078200 > protein coding
SRAE_2000076700 > protein coding
SRAE_2000077200 > protein coding
SRAE_2000077800 > protein coding
SRAE_2000078400 > protein coding
SRAE_2000076800 > protein coding
SRAE_2000077500 > protein coding
SRAE_2000077000 > protein coding
SRAE_chr2 >
< SRAE_2000076000 protein coding
< SRAE_2000076100 protein coding
< SRAE_2000077900 protein coding
< SRAE_2000078300 protein coding
< SRAE_2000078000 protein coding
2.44Mb
2.45Mb
2.46Mb
2.47Mb
2.48Mb
2.49Mb
2.50Mb
2.51Mb
2.52Mb
Reverse strand
85.55 kb

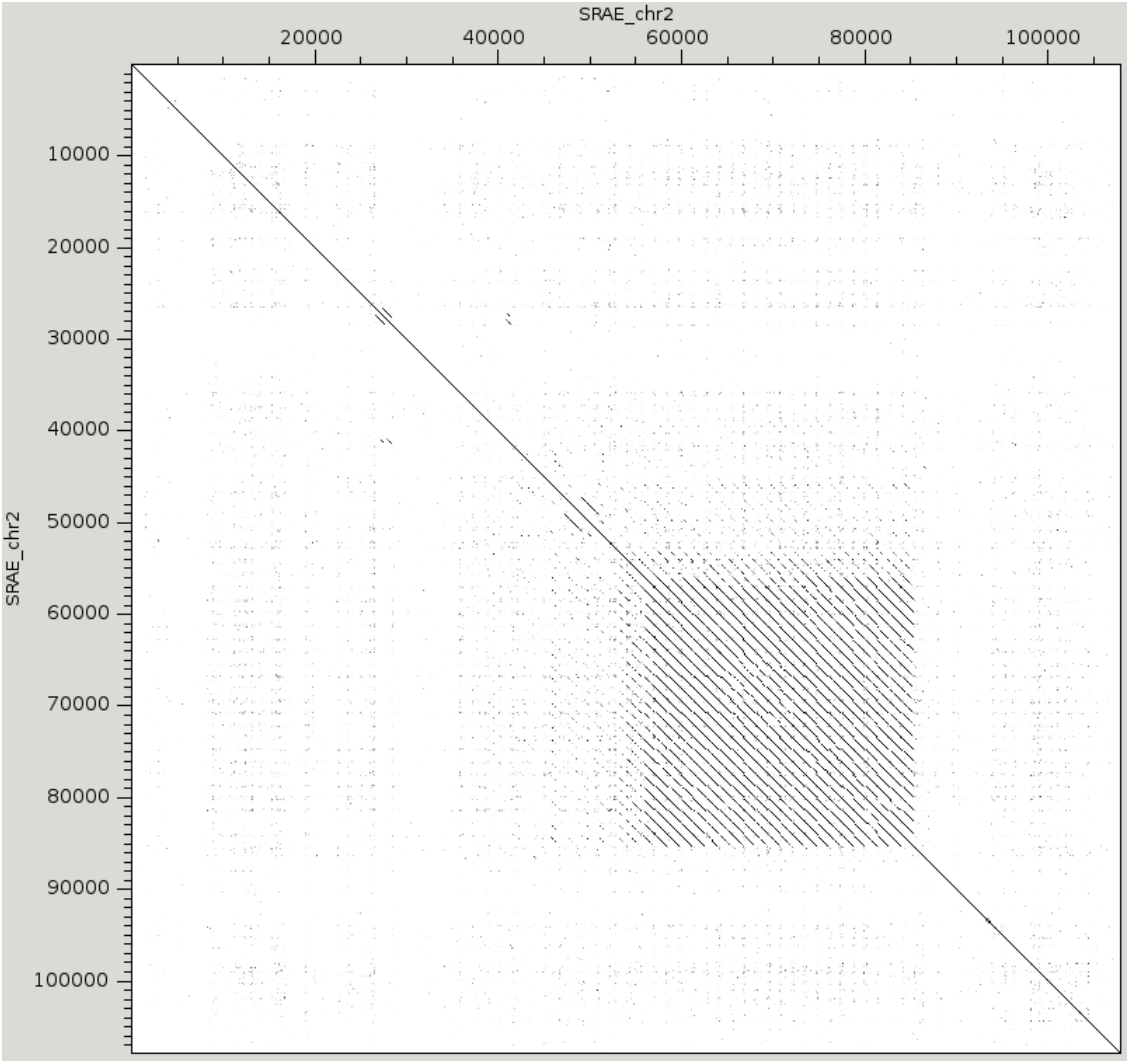


Genes discarded from expansion cluster:

SRAE_2000076500

Cluster 3


SRAE_2000123200 >
protein coding
SRAE_2000123600 >
protein coding
SRAE_2000124100 >
protein coding
SRAE_2000123300 >
protein coding
SRAE_2000123700 >
protein coding
SRAE_2000124200 >
protein coding
SRAE_2000126700 >
protein coding
SRAE_2000126800 >
protein coding
SRAE_2000126900 >
protein coding
SRAE_2000127000 >
protein coding
SRAE_2000600050 >
tRNA
SRAE_chr2 >
< SRAE_2000123100
protein coding
< SRAE_2000123500
protein coding
< SRAE_2000123400
protein coding
< SRAE_2000123800
protein coding
< SRAE_2000123900
protein coding
< SRAE_2000124000
protein coding
< SRAE_2000124300
protein coding
SRAE_2000124600
protein coding
SRAE_2000125100
protein coding
< SRAE_2000125700
protein coding
< SRAE_2000126100
protein coding
< SRAE_2000127100
protein coding
< SRAE_2000124400
protein coding
< SRAE_2000124800
protein coding
< SRAE_2000125400
protein coding
< SRAE_2000125810
protein coding
< SRAE_2000126300
protein coding
< SRAE_2000127200
protein coding
< SRAE_2000124500
protein coding
< SRAE_2000125200
protein coding
SRAE_2000126200
protein coding
< SRAE_2000124700
protein coding
< SRAE_2000125300
protein coding
< SRAE_2000125900
protein coding
< SRAE_2000126400
protein coding
< SRAE_2000124900
protein coding
< SRAE_2000125500
protein coding
< SRAE_2000126000
protein coding
< SRAE_2000126600
protein coding
< SRAE_2000125000
protein coding
< SRAE_2000125600
protein coding
< SRAE_2000126290
protein coding
< SRAE_2000126500
protein coding
3.85Mb
3.86Mb
3.87Mb
3.88Mb
3.89Mb
3.90Mb
3.91Mb
3.92Mb
3.93Mb
3.94Mb
3.95Mb
Reverse strand
107.61 kb



Genes discarded from expansion cluster:

SRAE_2000125000

SRAE_2000125100

SRAE_2000125200

SRAE_2000125300

SRAE_2000125400

SRAE_2000125500

SRAE_2000125600

SRAE_2000125700

SRAE_2000125800

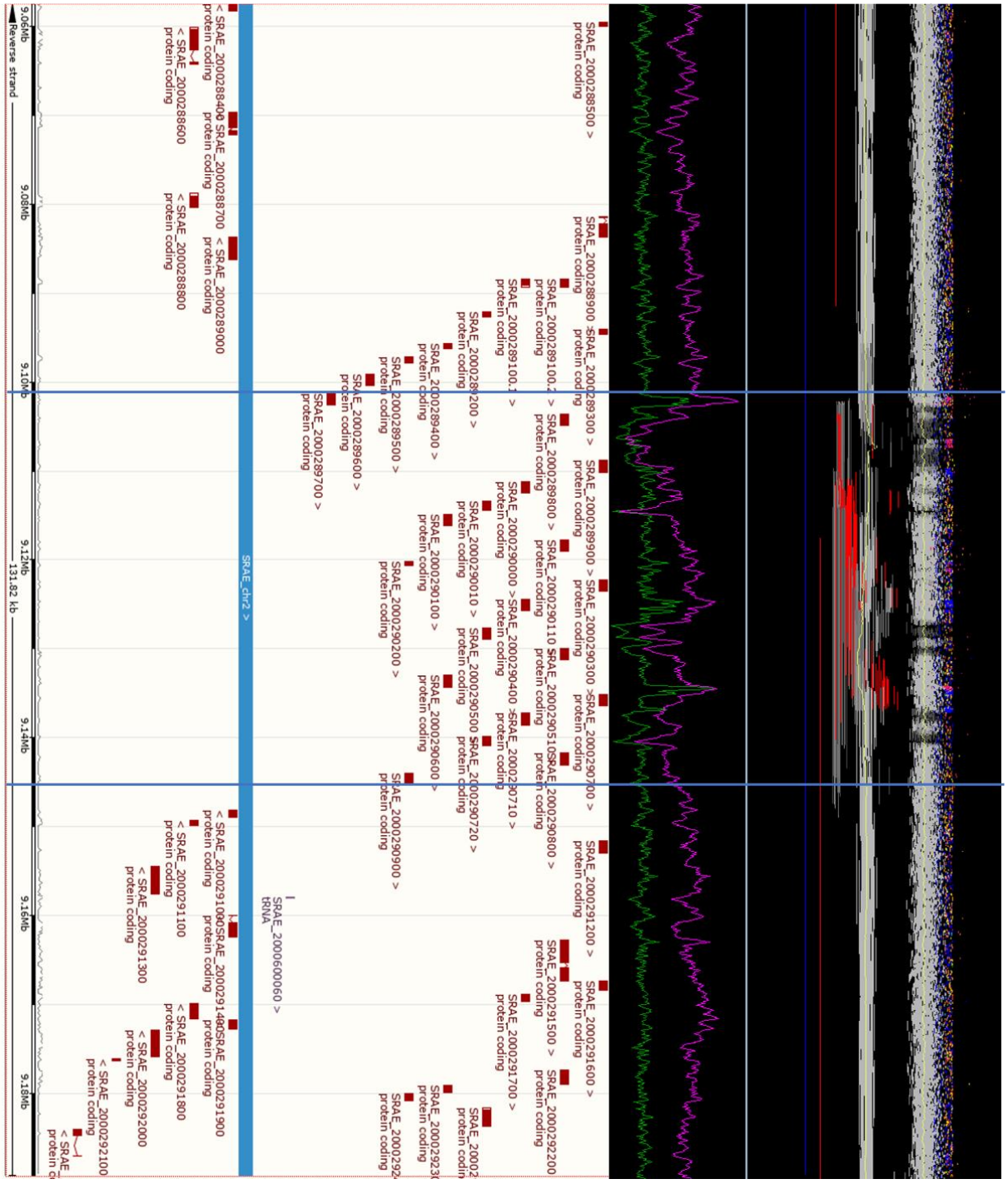
SRAE_2000125810

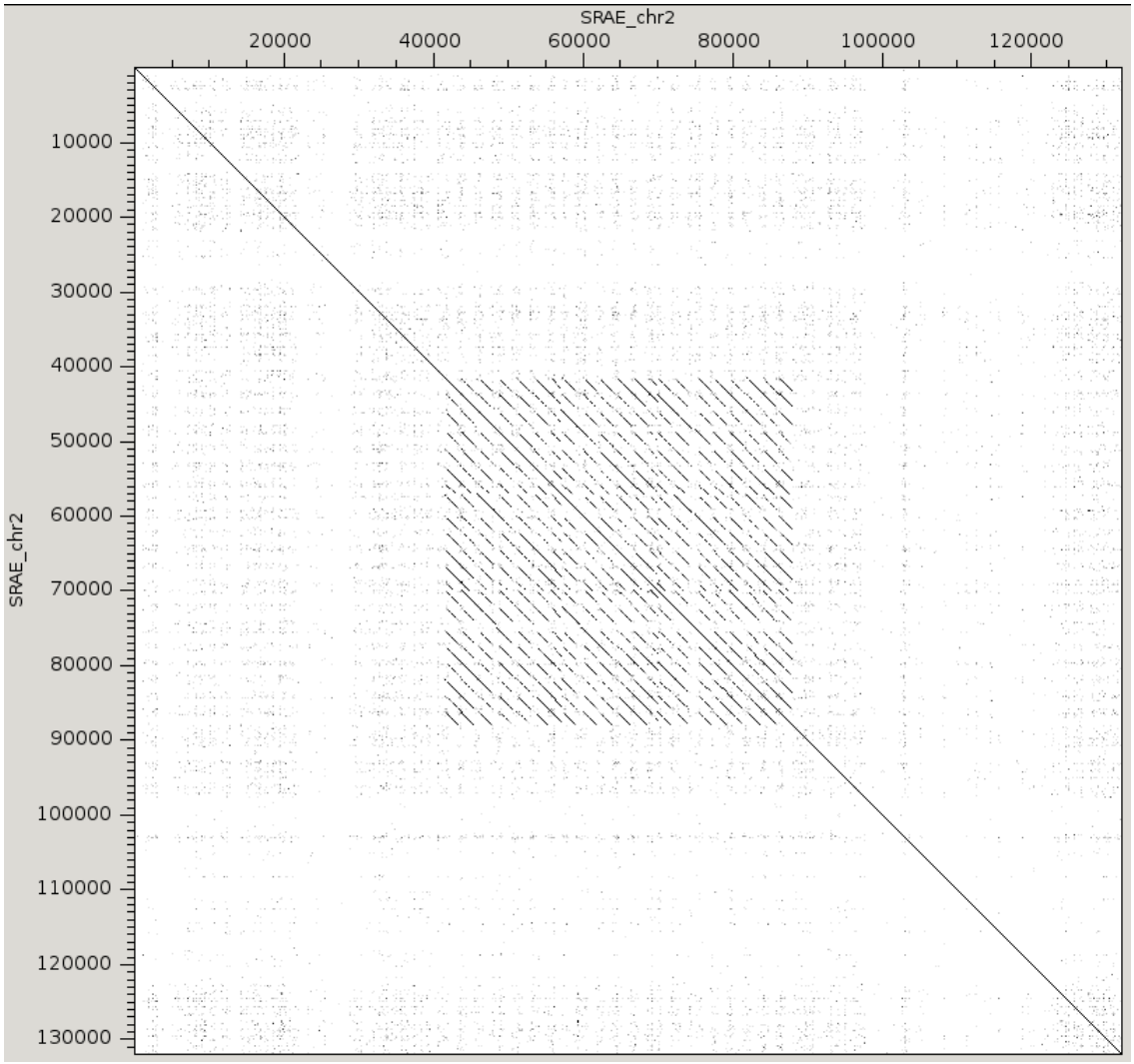
Genes discarded from right flanking region:

SRAE_2000125900

SRAE_2000126000

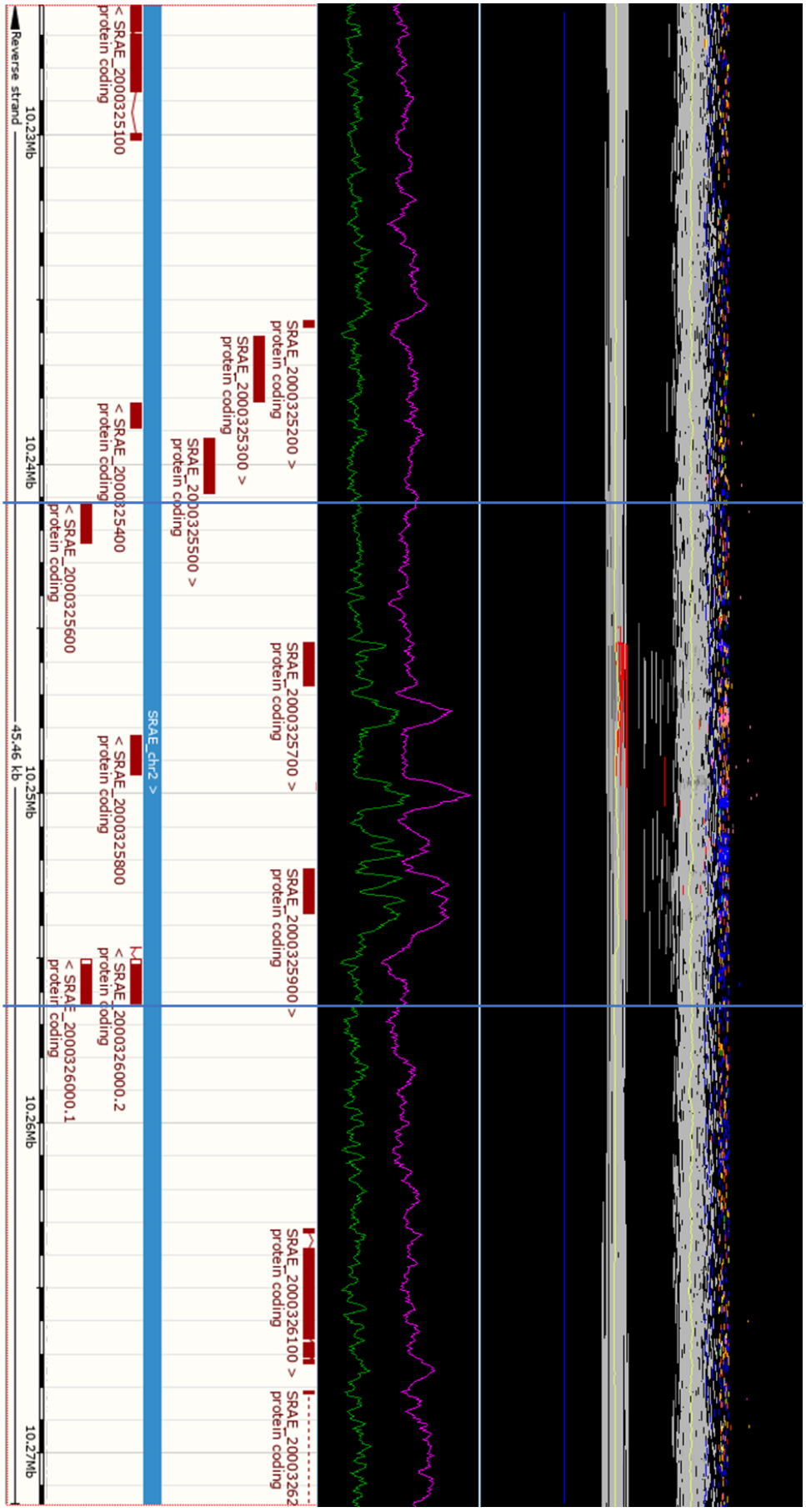
Cluster 4

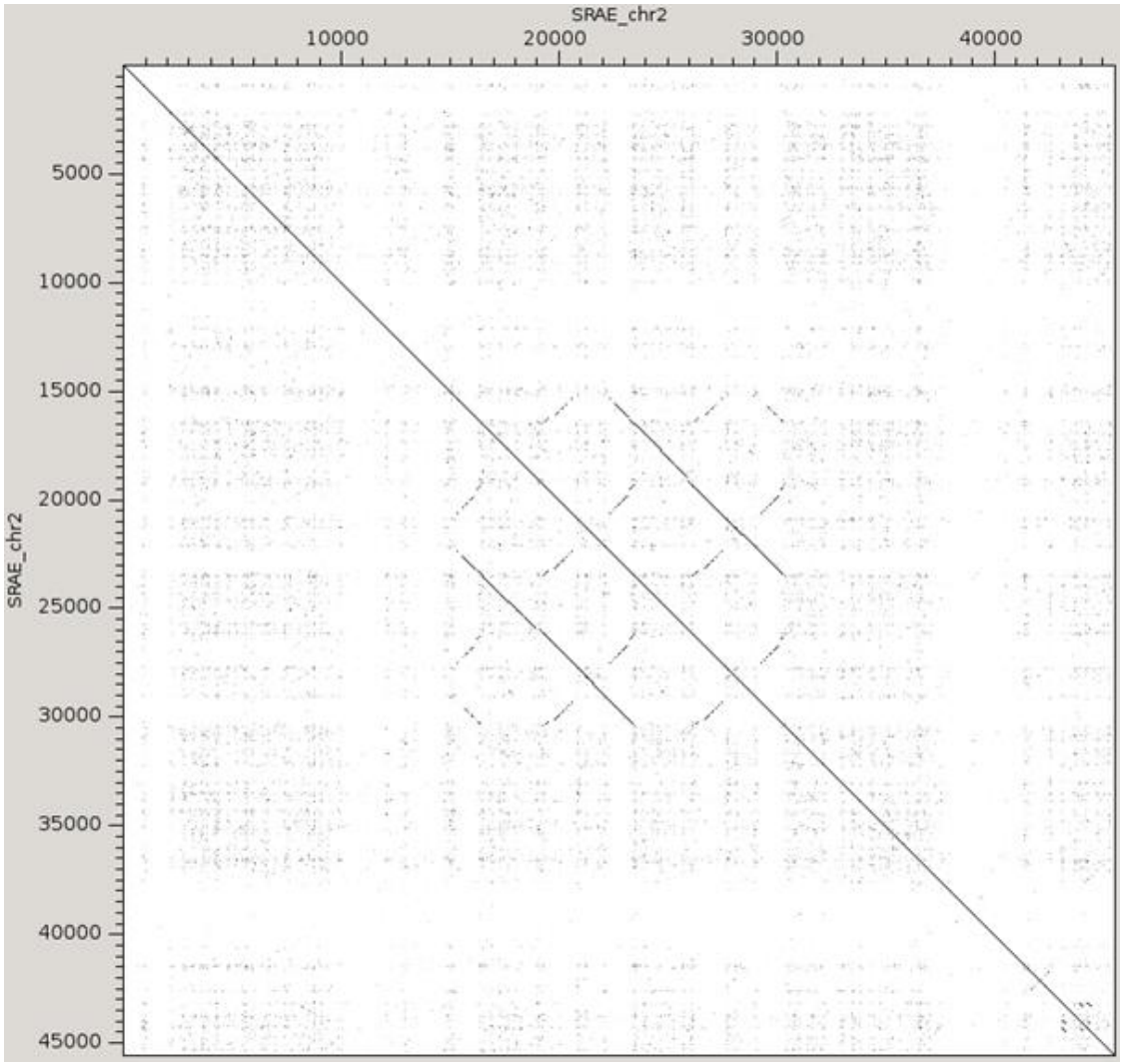




All expansion cluster genes found to be over poor quality assembly, entire region discarded.

Cluster 5


Reverse strand
10.23Mb
10.24Mb
10.25Mb
10.26Mb
10.27Mb
45.46 kb
SRAE_chr2 >
< SRAE_2000325100
protein coding
SRAE_2000325200 >
protein coding
SRAE_2000325300 >
protein coding
SRAE_2000325500 >
protein coding
< SRAE_2000325400
protein coding
< SRAE_2000325600
protein coding
SRAE_2000325700 >
protein coding
< SRAE_2000325800
protein coding
SRAE_2000325900 >
protein coding
< SRAE_2000326000.2
protein coding
< SRAE_2000326000.1
protein coding
SRAE_2000326100 >
protein coding
SRAE_20003262
protein coding



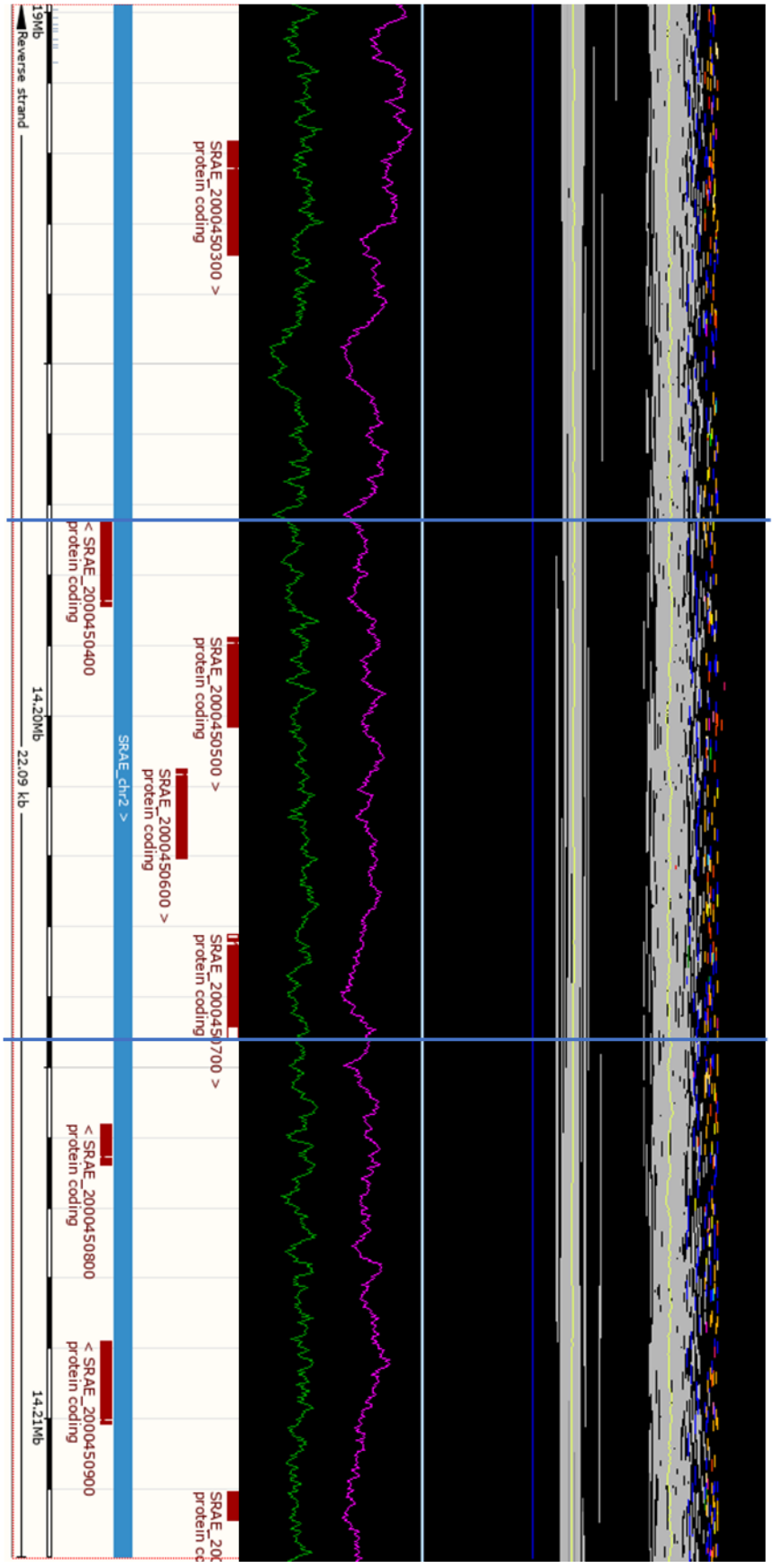
Genes discarded from expansion cluster:

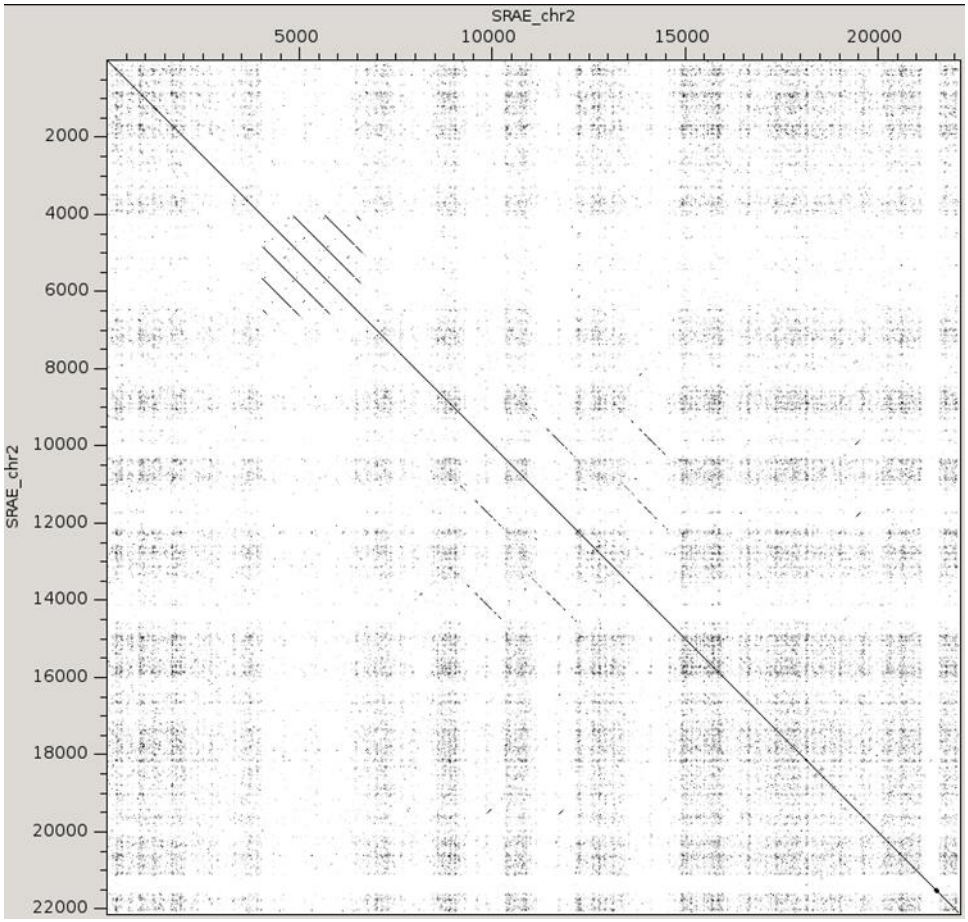
SRAE_ 2000325700

SRAE_ 2000325800

SRAE_ 2000325900

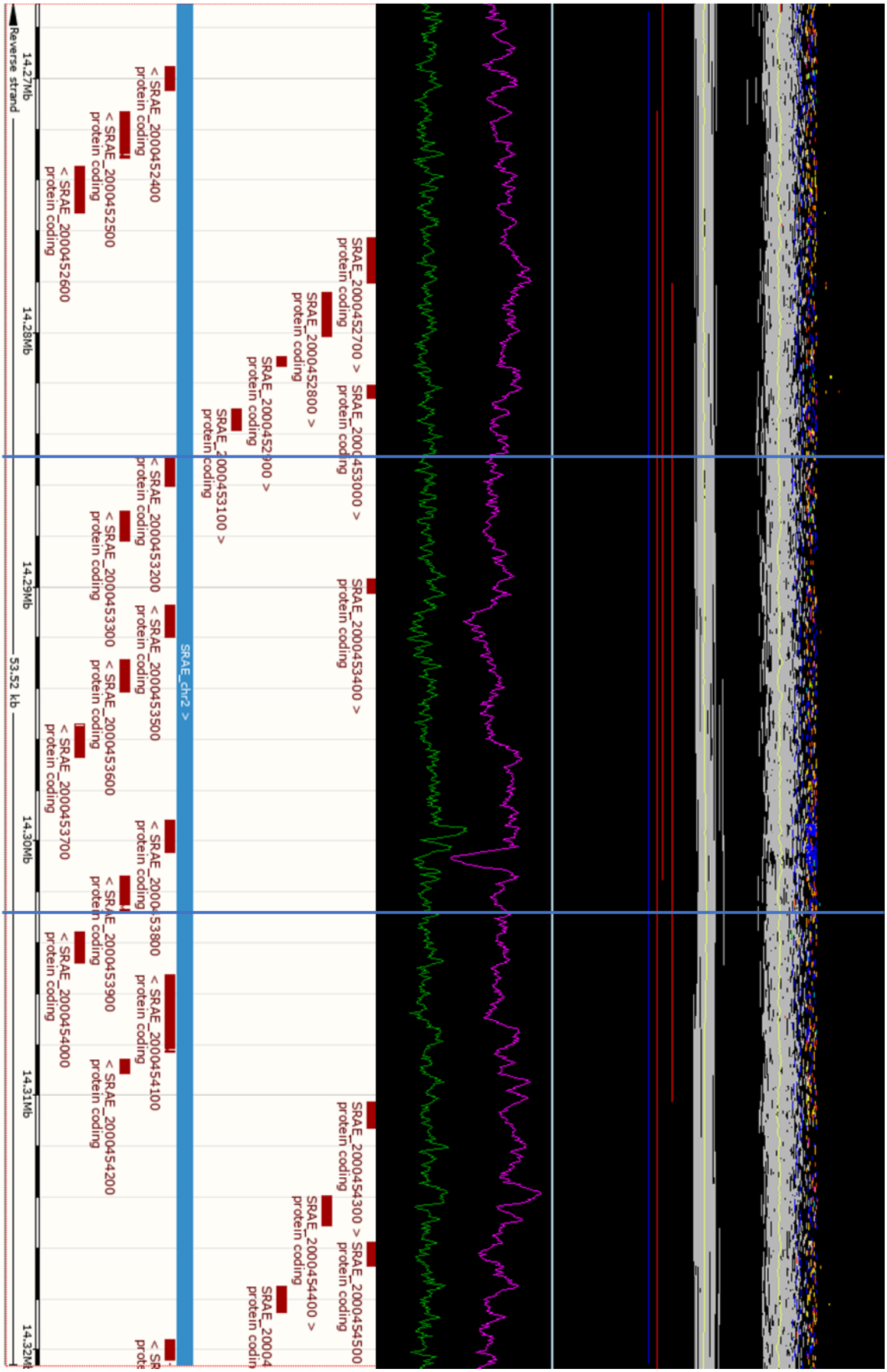
Cluster 6

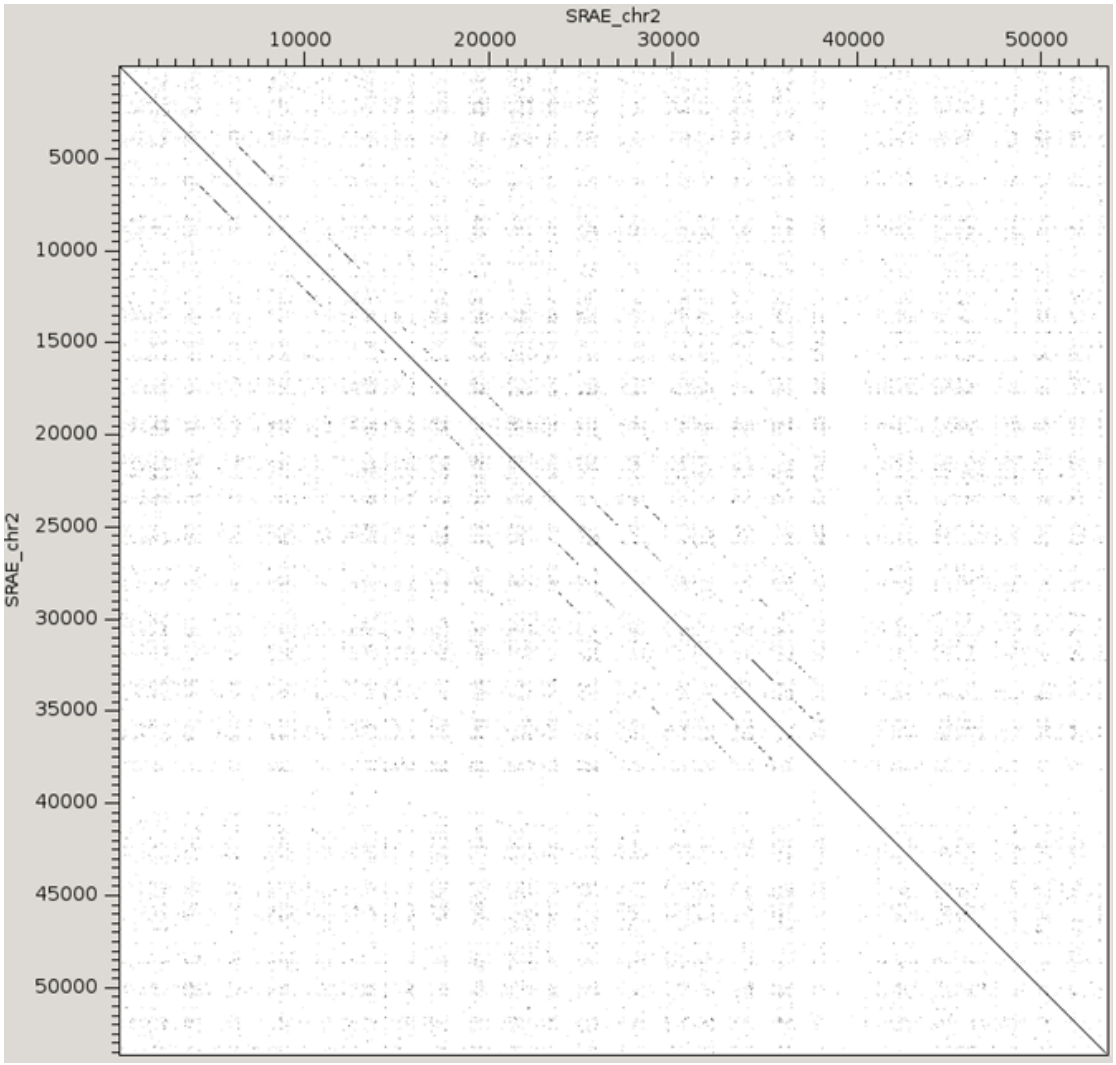

19Mb
Reverse strand
SRAE_2000450300 >
protein coding
< SRAE_2000450400
protein coding
SRAE_2000450500 >
protein coding
SRAE_2000450600 >
protein coding
SRAE_2000450700 >
protein coding
SRAE_ch2 >
14.20Mb
22.09 kb
< SRAE_2000450800
protein coding
< SRAE_2000450900
protein coding
14.21Mb



No genes removed

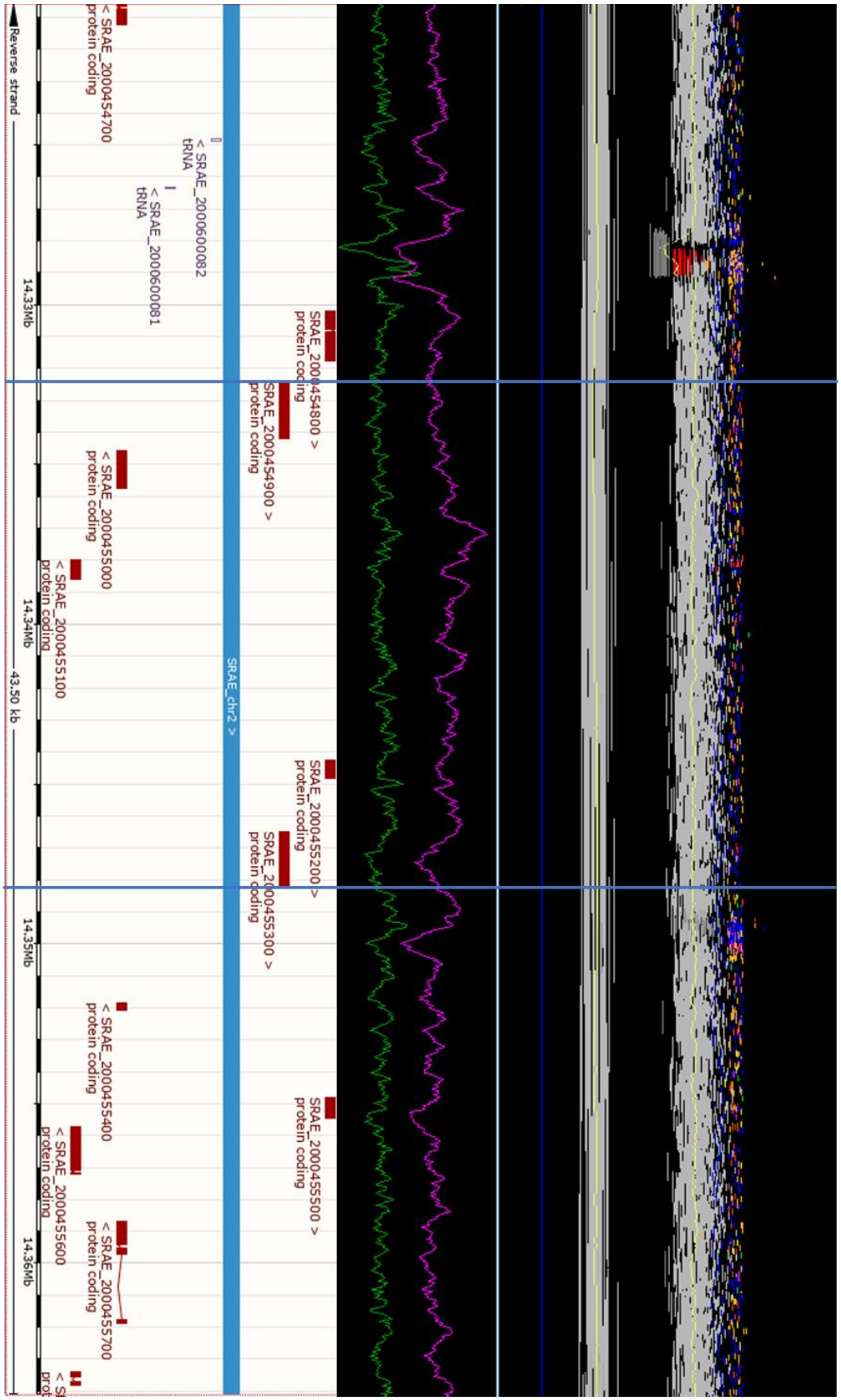
Cluster 7

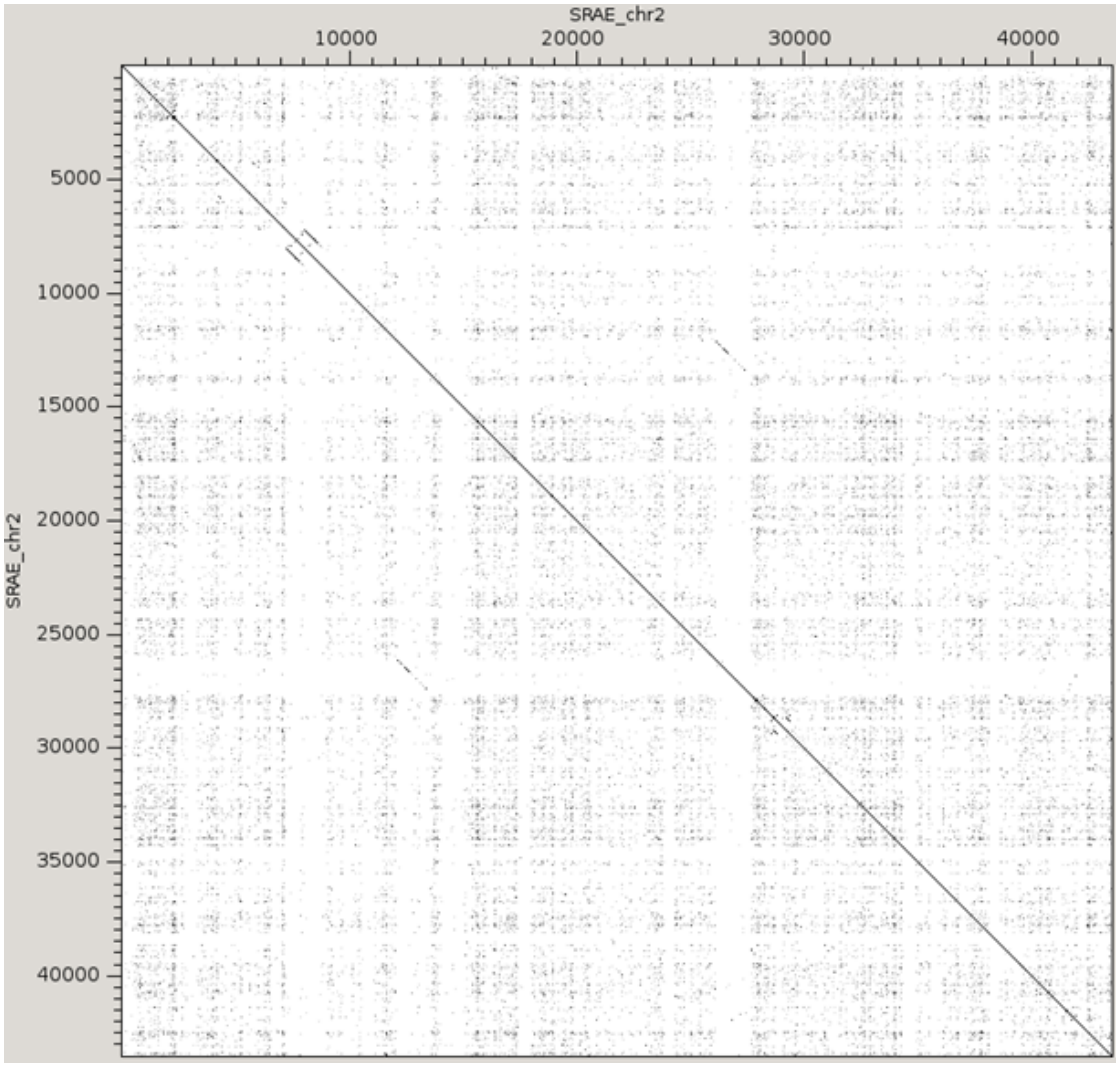

Reverse strand
14.27Mb
14.28Mb
14.29Mb
53.52 kb
14.30Mb
14.31Mb
14.32Mb
SRAE_chr2 >
< SRAE_2000452400 protein coding
< SRAE_2000452500 protein coding
< SRAE_2000452600 protein coding
SRAE_2000452700 > protein coding
SRAE_2000452800 > protein coding
SRAE_2000453000 > protein coding
SRAE_2000453100 > protein coding
< SRAE_2000453200 protein coding
< SRAE_2000453300 protein coding
< SRAE_2000453500 protein coding
< SRAE_2000453600 protein coding
< SRAE_2000453700 protein coding
SRAE_2000453400 > protein coding
< SRAE_2000453800 protein coding
< SRAE_2000453900 protein coding
< SRAE_2000454000 protein coding
< SRAE_2000454100 protein coding
< SRAE_2000454200 protein coding
SRAE_2000454300 > protein coding
SRAE_2000454400 > protein coding
SRAE_2000454500 protein coding



No genes removed

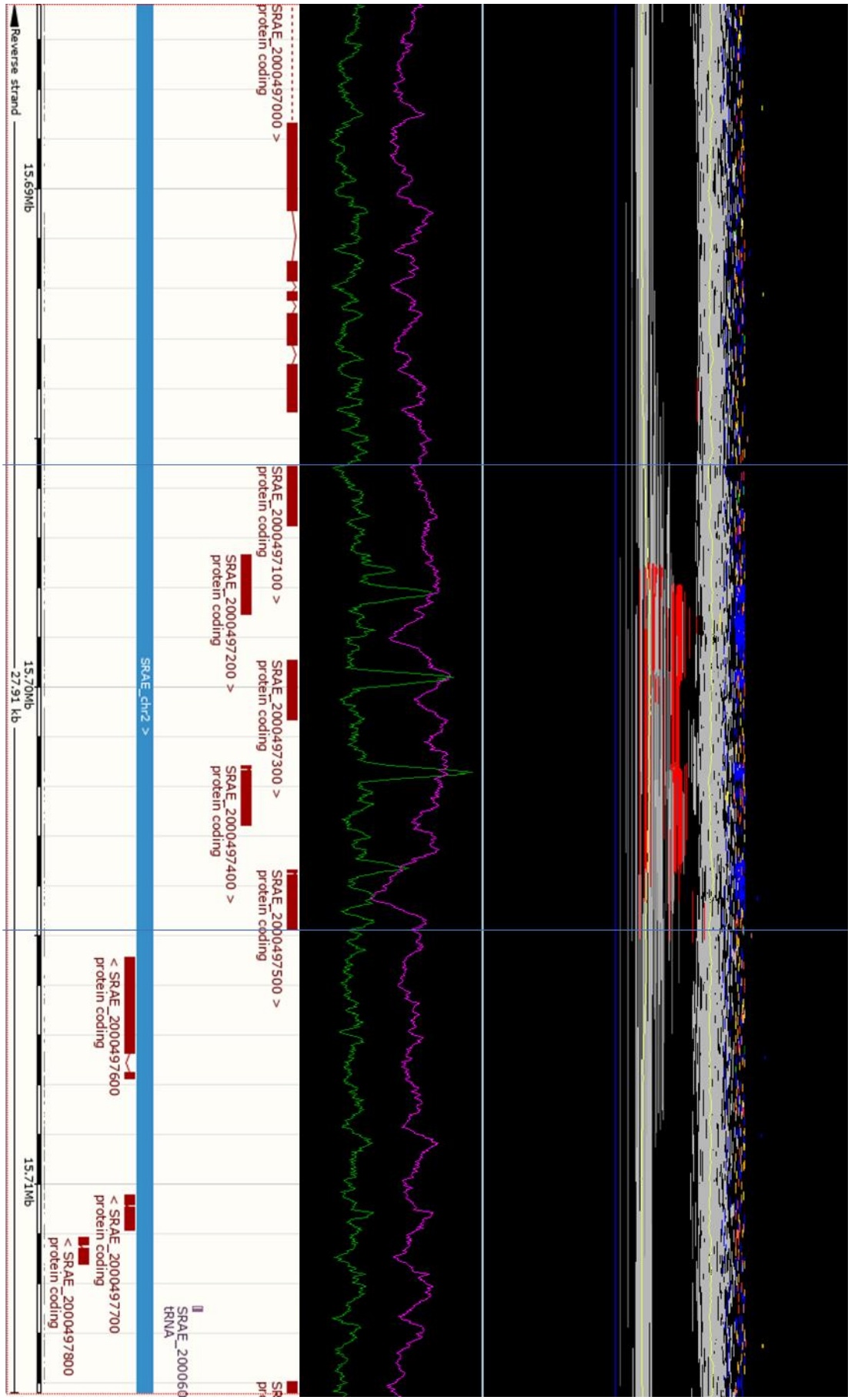
Cluster 8

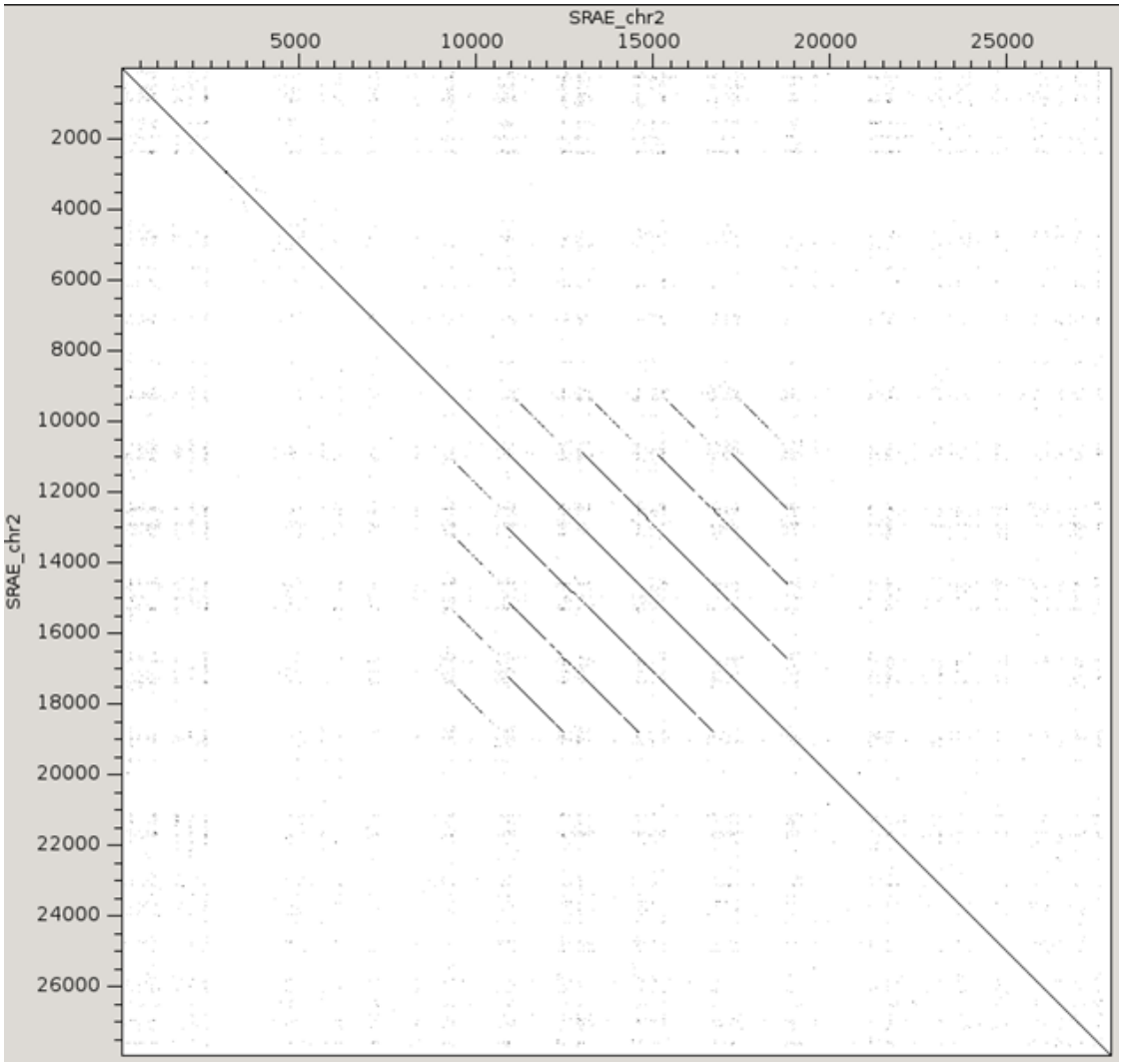

Reverse strand
14.33Mb
< SRAE_2000454700
protein coding
< SRAE_2000600082
tRNA
< SRAE_2000600081
tRNA
SRAE_2000454800 >
protein coding
SRAE_2000454900 >
protein coding
SRAE_chr2 >
14.34Mb
43.50 kb
< SRAE_2000455000
protein coding
< SRAE_2000455100
protein coding
SRAE_2000455200 >
protein coding
SRAE_2000455300 >
protein coding
SRAE_2000455500 >
protein coding
14.35Mb
< SRAE_2000455400
protein coding
< SRAE_2000455600
protein coding
< SRAE_2000455700
protein coding
14.36Mb



No genes removed

Cluster 9


Reverse strand
15.69Mb
15.70Mb
27.91 kb
15.71Mb
SRAE_chr2 >
SRAE_2000497000 >
protein coding
SRAE_2000497100 >
protein coding
SRAE_2000497200 >
protein coding
SRAE_2000497300 >
protein coding
SRAE_2000497400 >
protein coding
SRAE_2000497500 >
protein coding
< SRAE_2000497600
protein coding
< SRAE_2000497700
protein coding
< SRAE_2000497800
protein coding
SRAE_200060
tRNA



Genes discarded from expansion cluster:

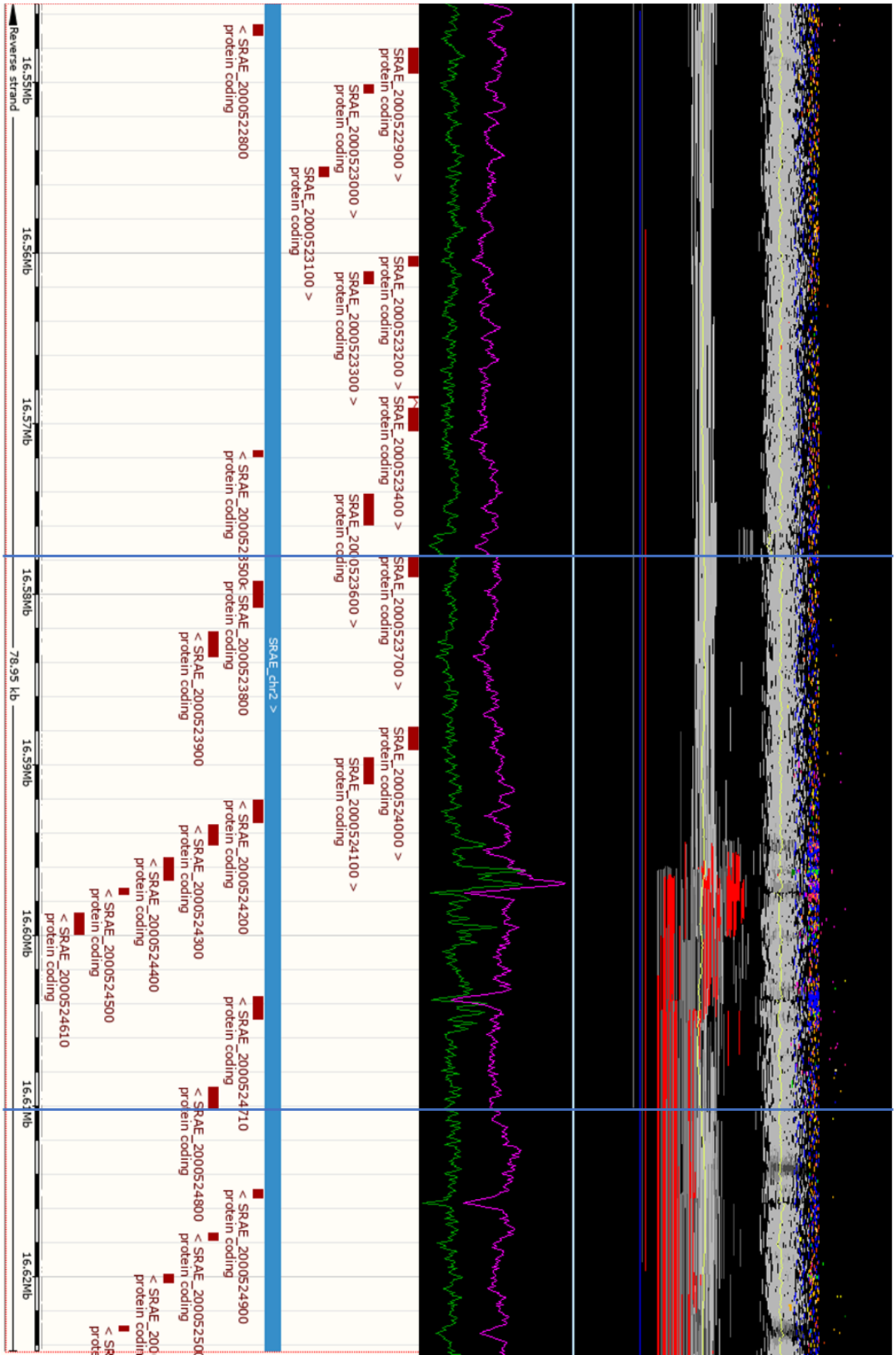
SRAE_2000497200

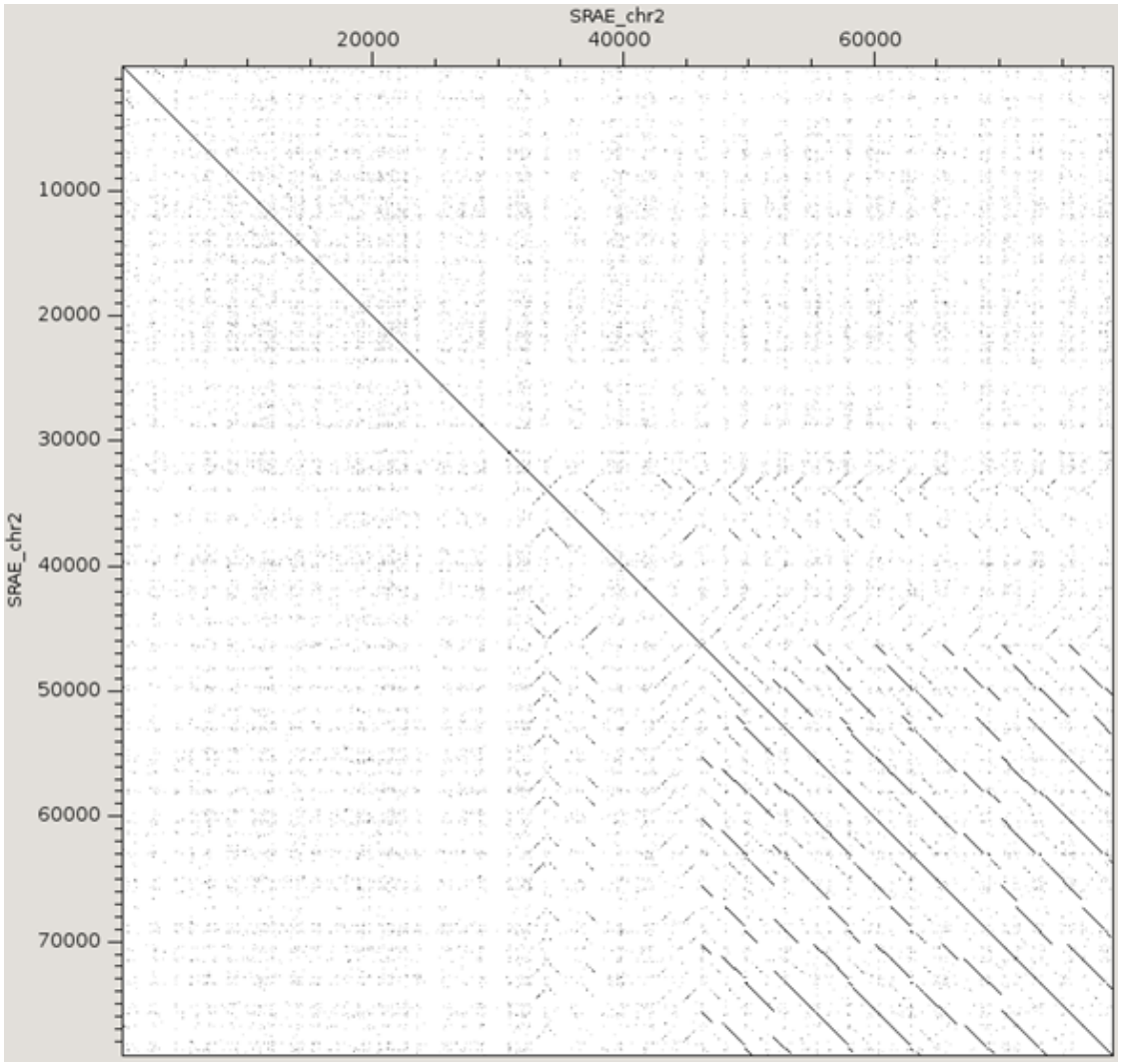
SRAE_2000497300

SRAE_2000497400

SRAE_2000497500

Cluster 10


SRAE_chr2 >
SRAE_20005222900 >
protein coding
SRAE_20005223000 >
protein coding
SRAE_20005223100 >
protein coding
SRAE_20005223200 >
protein coding
SRAE_20005223300 >
protein coding
SRAE_20005223400 >
protein coding
SRAE_20005223500 >
protein coding
SRAE_20005223600 >
protein coding
SRAE_20005223700 >
protein coding
SRAE_20005224000 >
protein coding
SRAE_20005224100 >
protein coding
< SRAE_20005222800
protein coding
< SRAE_20005223800
protein coding
< SRAE_20005223900
protein coding
< SRAE_20005224200
protein coding
< SRAE_20005224300
protein coding
< SRAE_20005224400
protein coding
< SRAE_20005224500
protein coding
< SRAE_20005224610
protein coding
< SRAE_20005224710
protein coding
< SRAE_20005224800
protein coding
< SRAE_20005224900
protein coding
Reverse strand
16.55Mb
16.56Mb
16.57Mb
16.58Mb
78.95 kb
16.59Mb
16.60Mb
16.61Mb
16.62Mb



Genes discarded from expansion cluster:

SRAE_2000524100

SRAE_2000524200

SRAE_2000524300

SRAE_2000524300

SRAE_2000524400

SRAE_2000524500

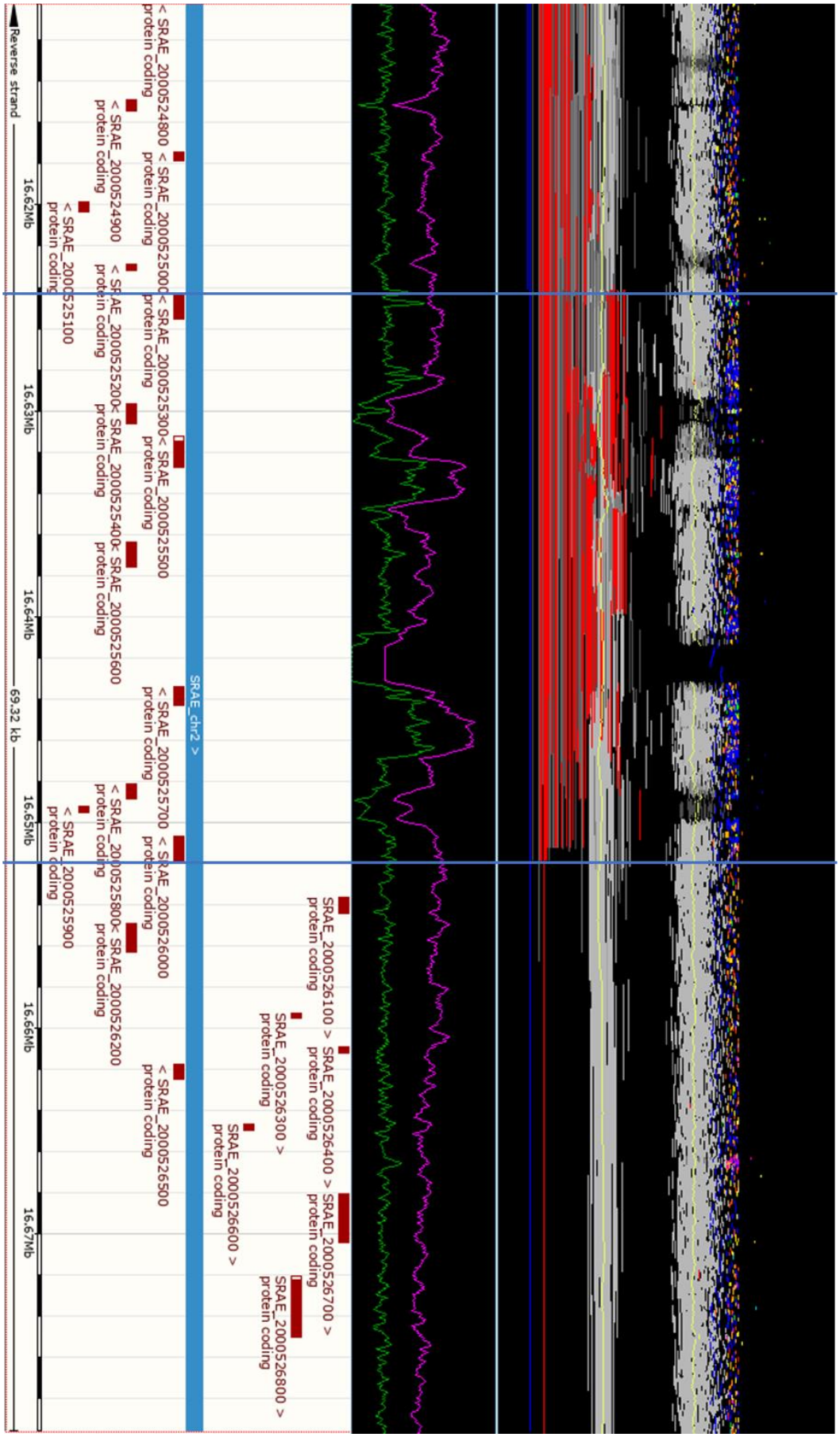
SRAE_2000524610

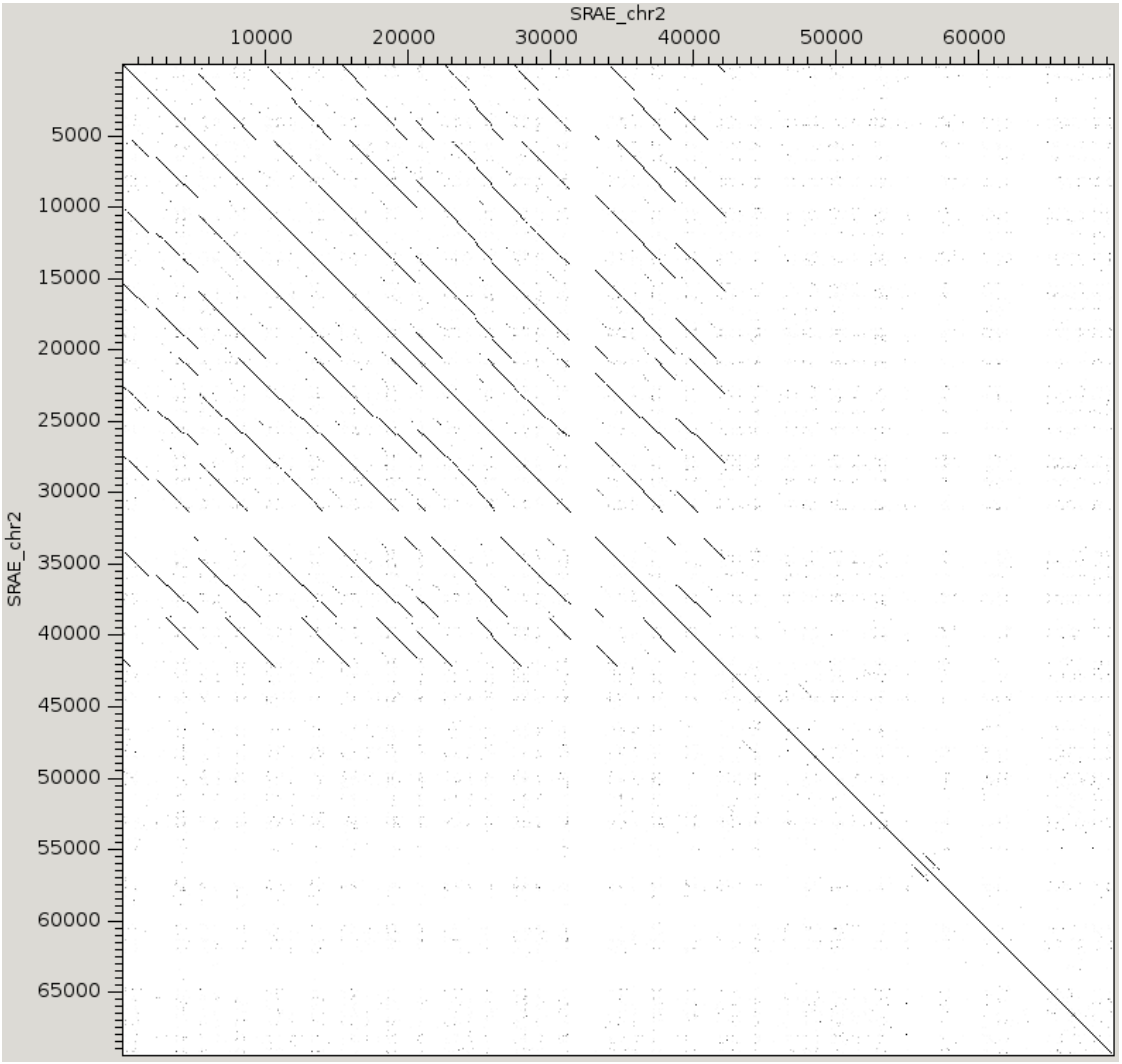
SRAE_2000524710

SRAE_2000524800

All genes discarded from right flanking region

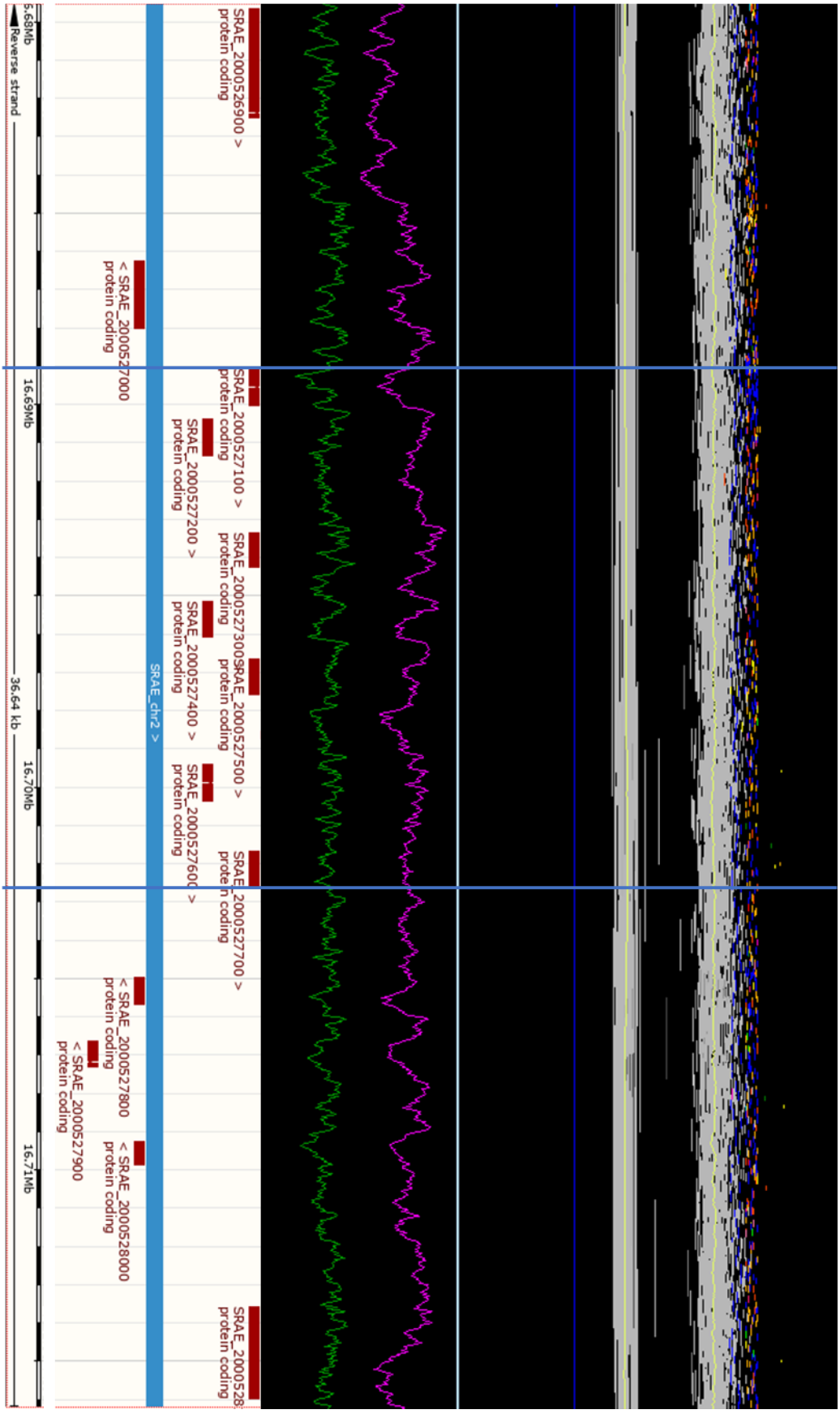
Cluster 11

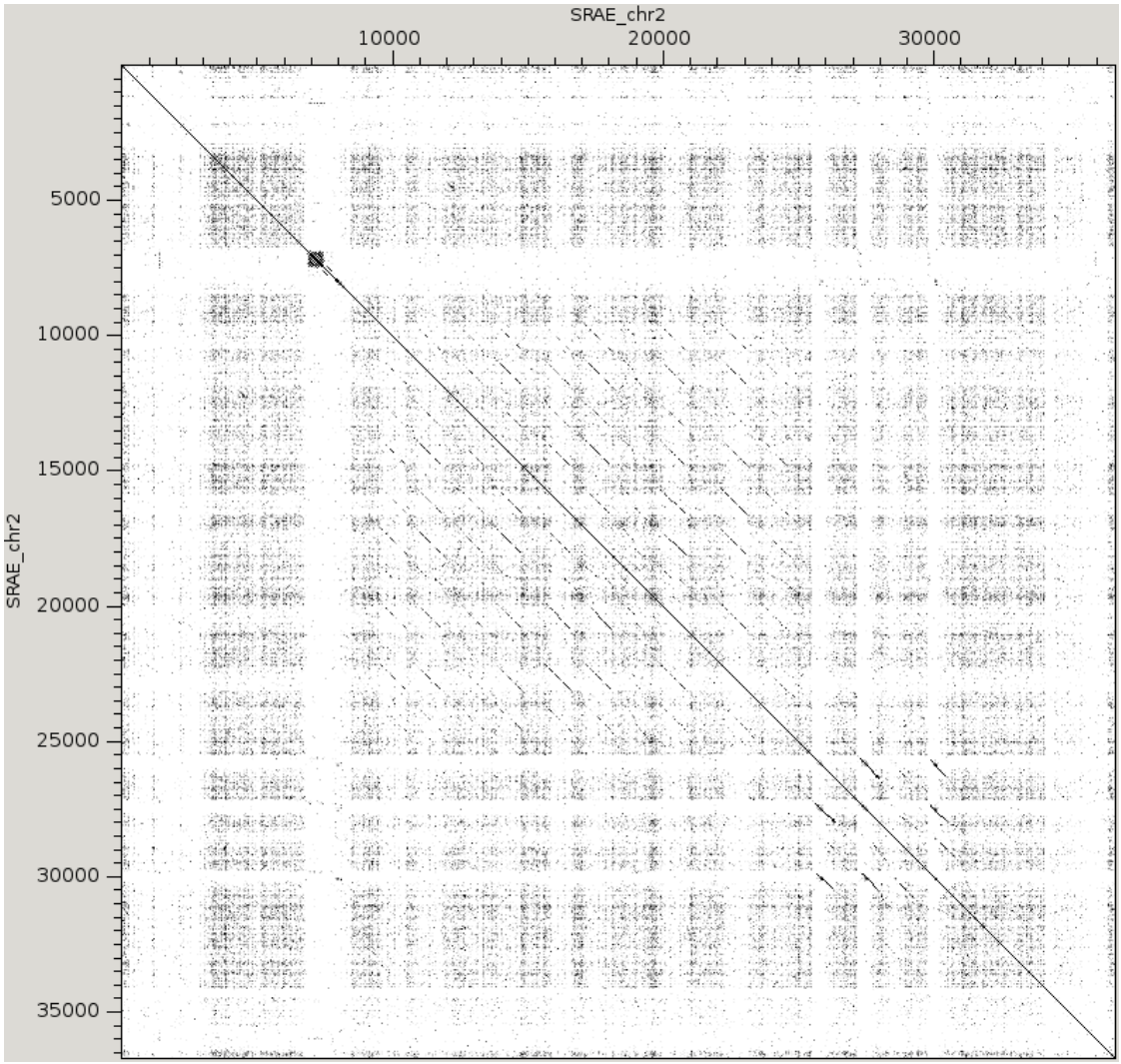

Reverse strand
16.62Mb
16.63Mb
16.64Mb
69.32 kb
16.65Mb
16.66Mb
16.67Mb
SRAE_chr2 >
< SRAE_2000524800 protein coding
< SRAE_2000524900 protein coding
< SRAE_2000525000 protein coding
< SRAE_2000525100 protein coding
< SRAE_2000525200 protein coding
< SRAE_2000525300 protein coding
< SRAE_2000525400 protein coding
< SRAE_2000525500
< SRAE_2000525600 protein coding
< SRAE_2000525700 protein coding
< SRAE_2000525800 protein coding
< SRAE_2000525900 protein coding
< SRAE_2000526000 protein coding
< SRAE_2000526200 protein coding
< SRAE_2000526500 protein coding
SRAE_2000526100 > protein coding
SRAE_2000526300 > protein coding
SRAE_2000526400 > protein coding
SRAE_2000526600 > protein coding
SRAE_2000526700 > protein coding
SRAE_2000526800 > protein coding



All expansion cluster genes found to be over poor quality assembly, entire region discarded.

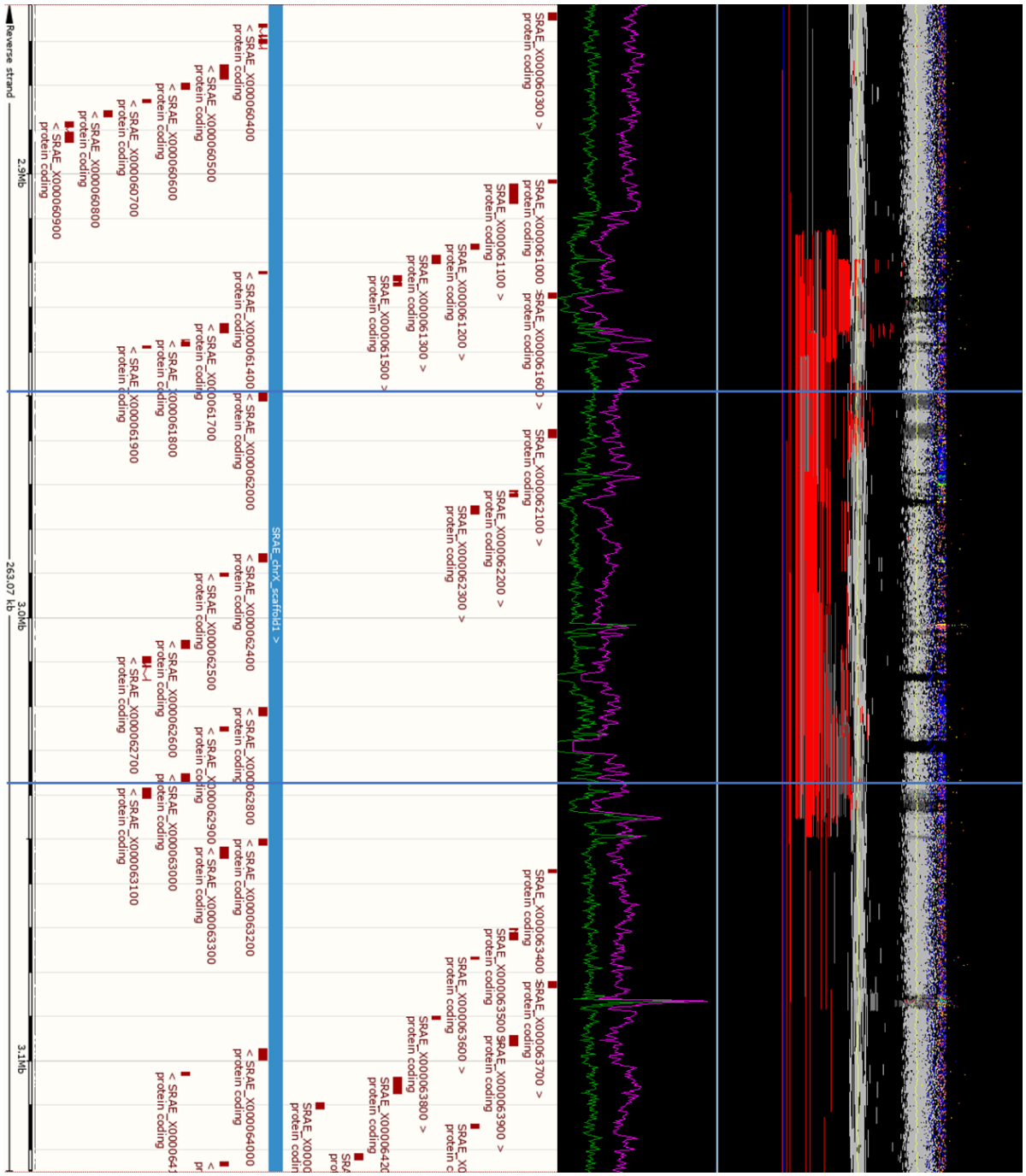
Cluster 12

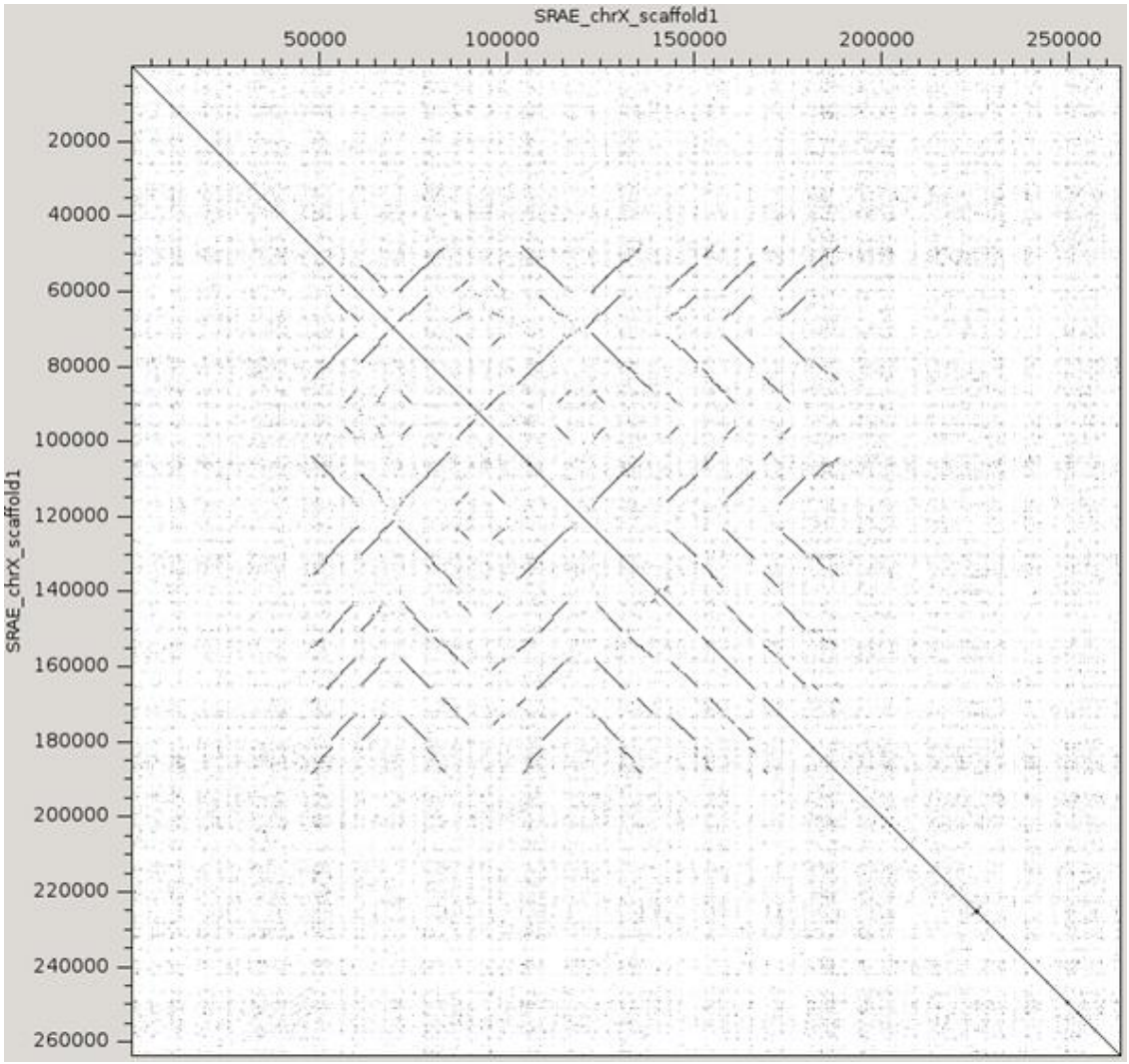

5.68Mb
Reverse strand
16.69Mb
36.64 kb
16.70Mb
16.71Mb
SRAE_chr2 >
SRAE_2000526900 >
protein coding
< SRAE_2000527000
protein coding
SRAE_2000527100 >
protein coding
SRAE_2000527200 >
protein coding
SRAE_2000527300
protein coding
SRAE_2000527400 >
protein coding
SRAE_2000527500 >
protein coding
SRAE_2000527600
protein coding
SRAE_2000527700 >
protein coding
< SRAE_2000527800
protein coding
< SRAE_2000527900
protein coding
< SRAE_2000528000
protein coding
SRAE_2000528
protein coding



No genes removed

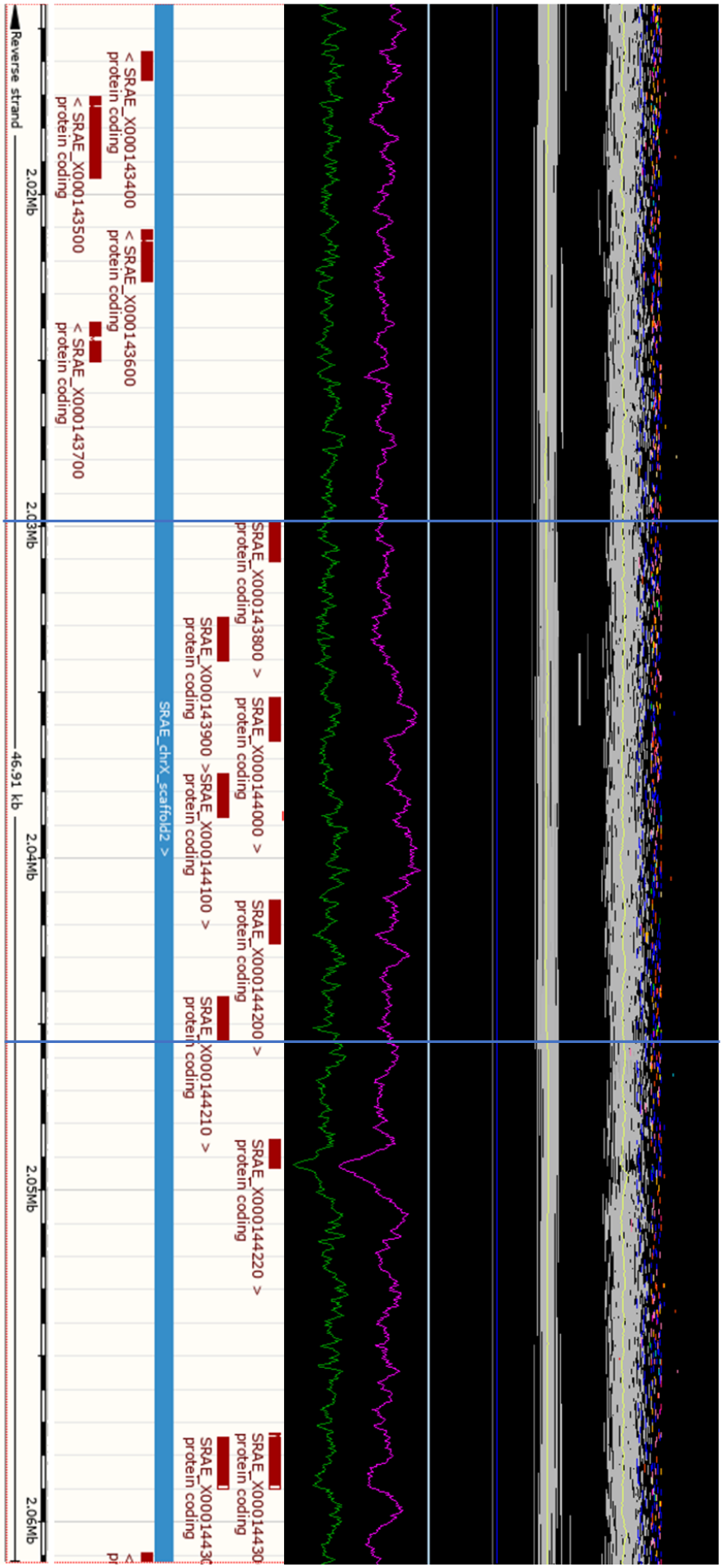
Cluster 13

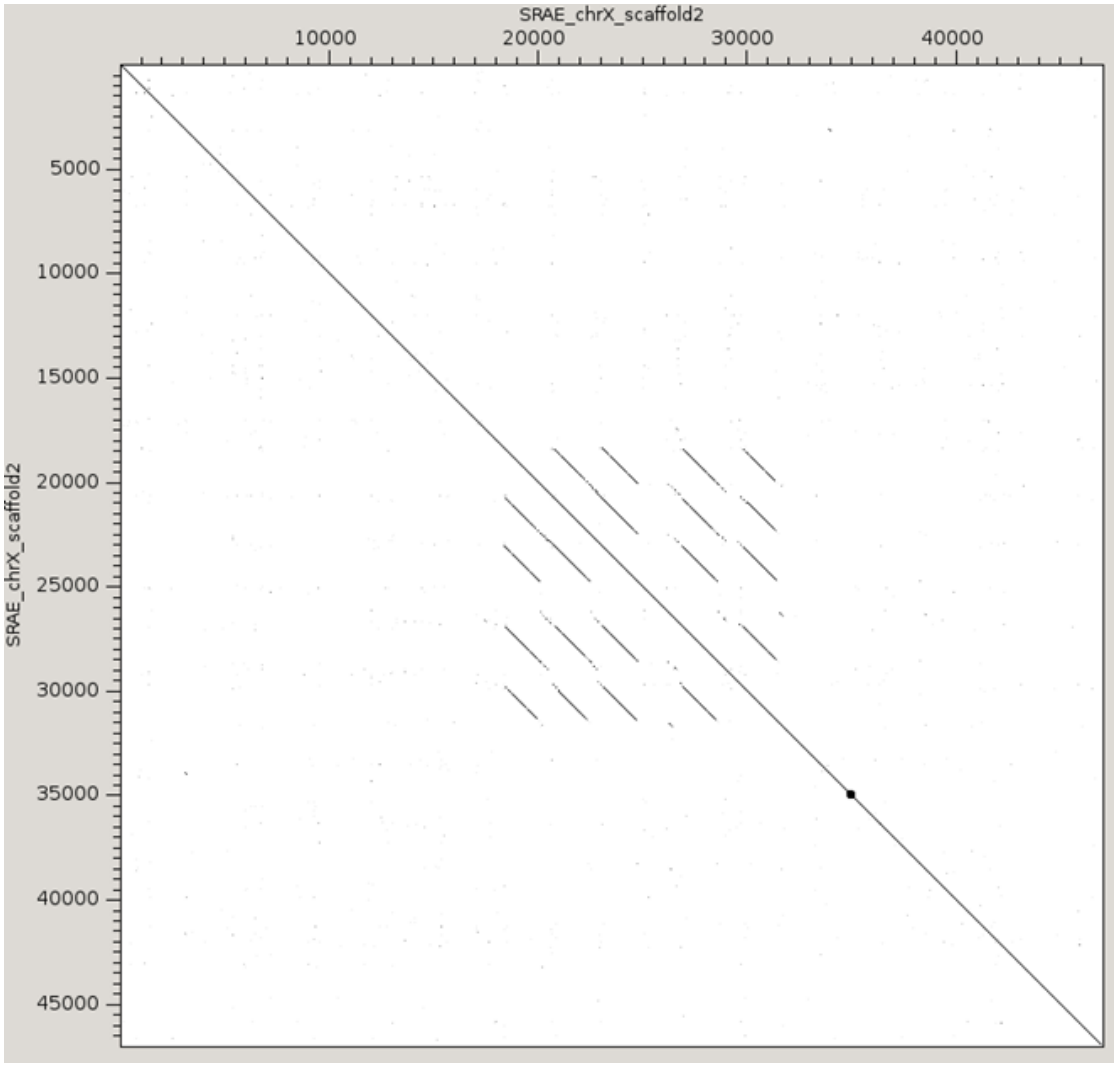

Reverse strand
2.9Mb
263.07 kb
3.0Mb
3.1Mb
SRAE_chrX_scaffold1 >
SRAE_X000060300 >
SRAE_X000060400
SRAE_X000060500
SRAE_X000060600
SRAE_X000060700
SRAE_X000060800
SRAE_X000060900
SRAE_X000061000 >
SRAE_X000061100 >
SRAE_X000061200 >
SRAE_X000061300 >
SRAE_X000061400
SRAE_X000061500 >
SRAE_X000061600 >
SRAE_X000061700
SRAE_X000061800
SRAE_X000061900
SRAE_X000062000
SRAE_X000062100 >
SRAE_X000062200 >
SRAE_X000062300 >
SRAE_X000062400
SRAE_X000062500
SRAE_X000062600
SRAE_X000062700
SRAE_X000062800
SRAE_X000062900
SRAE_X000063000
SRAE_X000063100
SRAE_X000063200
SRAE_X000063300
SRAE_X000063400
SRAE_X000063500
SRAE_X000063600 >
SRAE_X000063700 >
SRAE_X000063800 >
SRAE_X000063900 >
SRAE_X000064000
protein coding



All expansion cluster genes found to be over poor quality assembly, entire region discarded.

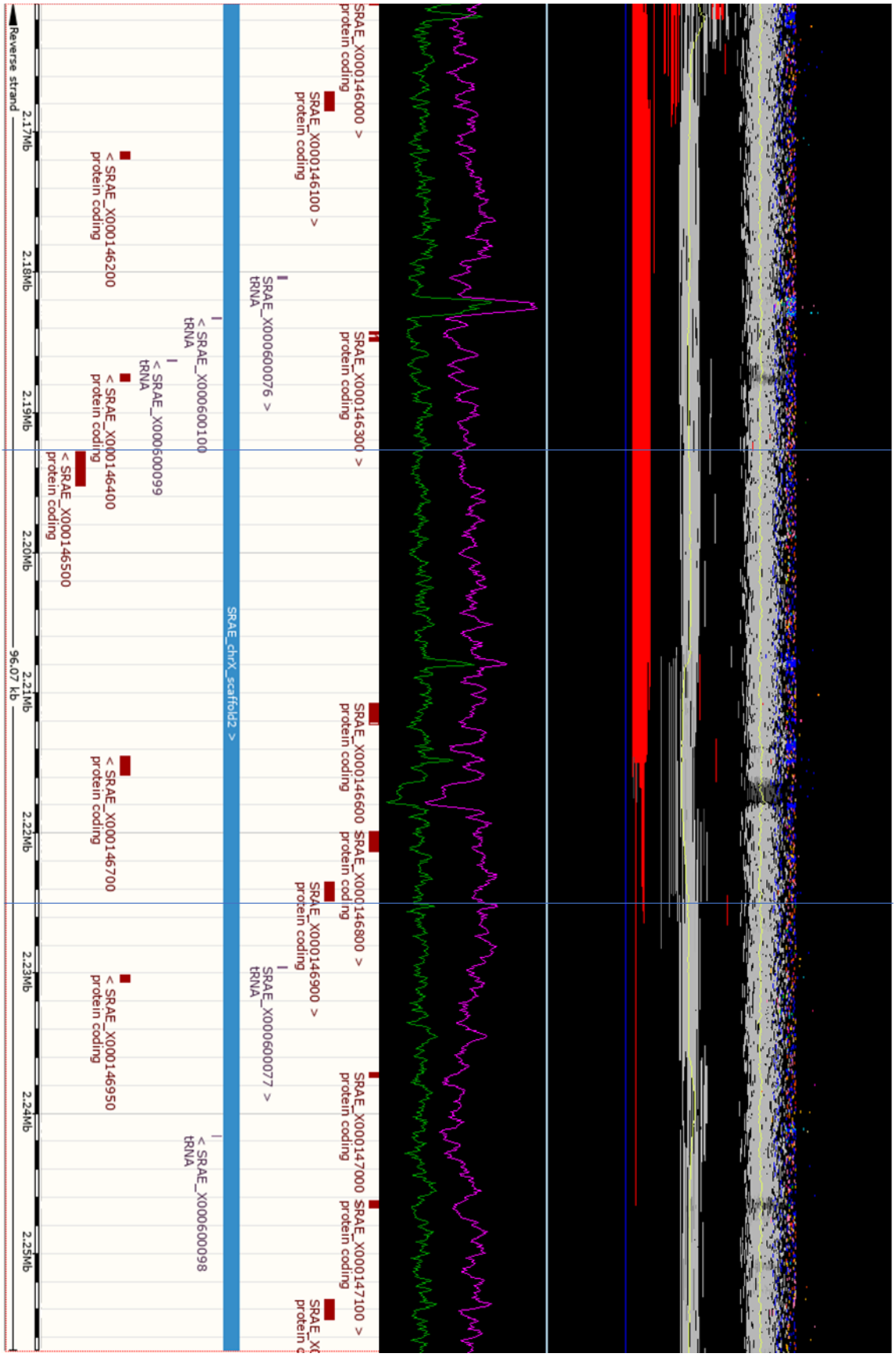
Cluster 14

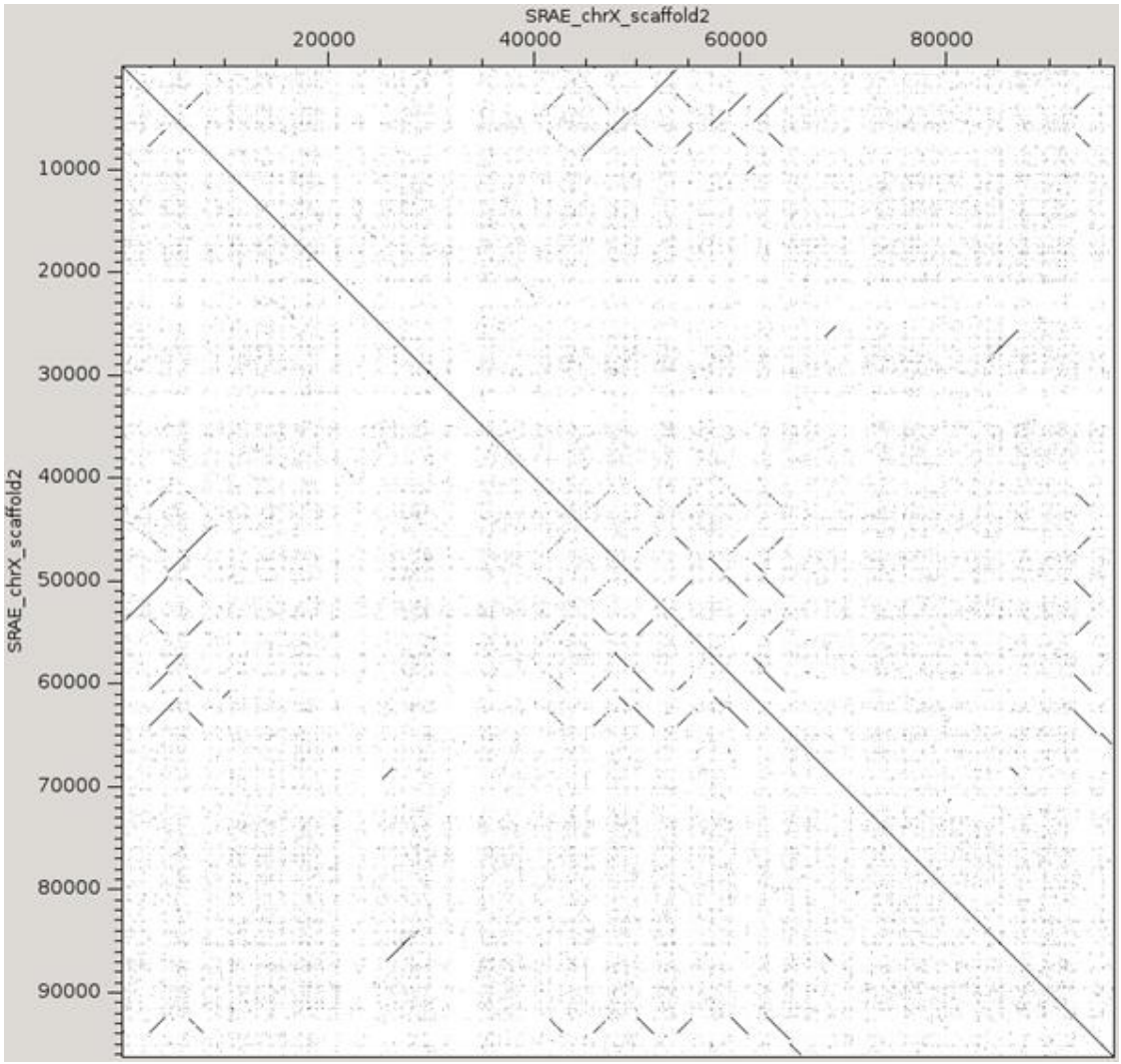

Reverse strand
2.02Mb
2.03Mb
2.04Mb
2.05Mb
2.06Mb
46.91 kb
SRAE_chrX_scaff6fd2 >
< SRAE_X000143400
protein coding
< SRAE_X000143500
protein coding
< SRAE_X000143600
protein coding
< SRAE_X000143700
protein coding
SRAE_X000143800 >
protein coding
SRAE_X000143900 >
protein coding
SRAE_X000144000 >
protein coding
SRAE_X000144100 >
protein coding
SRAE_X000144200 >
protein coding
SRAE_X000144210 >
protein coding
SRAE_X000144220 >
protein coding
SRAE_X000144430
protein coding



No genes removed

Cluster 15


Reverse strand
2.17Mb
2.18Mb
2.19Mb
2.20Mb
2.21Mb
96.07 kb
2.22Mb
2.23Mb
2.24Mb
2.25Mb
SRAE_chrX_scaffold2 >
SRAE_X000146000 >
protein coding
SRAE_X000146100 >
protein coding
SRAE_X000146200
< protein coding
SRAE_X000600076 >
tRNA
SRAE_X000600100
< tRNA
SRAE_X000600099
< tRNA
SRAE_X000146400
< protein coding
SRAE_X000146500
< protein coding
SRAE_X000146300 >
protein coding
SRAE_X000146600
protein coding
SRAE_X000146700
< protein coding
SRAE_X000146800 >
protein coding
SRAE_X000146900 >
protein coding
SRAE_X000600077 >
tRNA
SRAE_X000146950
< protein coding
SRAE_X000600098
< tRNA
SRAE_X000147000
protein coding
SRAE_X000147100 >
protein coding



All genes discarded from left flanking region

Genes discarded from expansion cluster:

SRAE_X000146500

SRAE_X000146600

SRAE_X000146700

References

Bonfield J.K. and Whitwham A. (2010). Gap5 - editing the billion fragment sequence assembly. *Bioinformatics* **26:**1699-1703.

Howe K. L., Bolt B. J., Shafie M., Kersey P and Berriman M. (2017). WormBase ParaSite − a comprehensive resource for helminth genomics. *Molecular and Biochemical Parasitology* **215:**2-10.

Hubbard T., Barker D., Birney E., Cameron G., Chen Y., Clark L., Cox T. *et al.* (2002). The Ensembl genome database project. *Nucleic Acids Research* **30:**38-41.

Hunt V. L., Tsai I. J., Coghlan A., Reid A. J., Holroyd N., Foth B. J., Tracey A. *et al.* (2016). The genomic basis of parasitism in the *Strongyloides* clade of nematodes. *Nature Genetics* **48:**299-307.

Sonnhammer E.L.L and Durbin R. (1995). A dot-matrix program dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* **167:**GC1-10.