Genetic architecture of the shell characteristics in the marine snail *Littorina saxatilis*

By:

Pragya Chaube

A thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy

The University of Sheffield
Faculty of Science
Department of Animals & Plant Sciences

October 2018

Abstract

Speciation is a key process underlying biodiversity. This process is facilitated by local adaptation, when divergent selection overcomes gene flow, resulting in the accumulation of reproductive barriers. Theory suggests that this accumulation is strongly dependent on the genetic architecture of the traits underlying local adaptation. The aim of this project was to investigate the genetic architecture of locally adaptive traits in the marine snail *Littorina saxatilis.*

This marine snail (*Littorina saxatilis*) is an excellent model to study speciation and local adaptation. Two diverging ecotypes live a few metres apart in distinct habitats and face divergent selection pressures dominated by crab predation and wave action. The ecotypes have evolved traits to adapt locally that make them behaviourally and structurally distinct. The most observable differences are seen in the shell size, shape, colours and patterns. Despite the differences, the two ecotypes meet in narrow contact zones and hybridize. Intermediates between the two parental ecotypes are observed in a crab-wave environmental gradient across the hybrid zones. This situation provides an excellent opportunity to exploit the power of association mapping in the hybrid zone to elucidate the genetic architectures of the locally adaptive traits. However, a prerequisite for the application of evolutionary genetic approaches is a genomic toolbox.

In Chapters 2 and 3, I describe the construction of a transcriptome assembly and high-density linkage map for this species. These genetic resources were utilized in the subsequent analyses and other studies in this system. In Chapter 4, I investigate the genetic architecture of the adaptive shell traits. Theory suggests that the ground colours or banding patterns possess Mendelian inheritance and may respond directly to selection or may be linked with genes that respond to the physical environment and may thus be affected by selection. Shell morphometric characters (size and shape) may have a more complex pattern of inheritance and tend to be responsive to the environmental conditions. Thus, shell characteristics are excellent to study divergent selection pressures and local adaptation while making it imperative to understand their underlying genetic architecture. In the current study, we applied association analysis to a single hybrid zone in Sweden to elucidate the genes underlying six shell phenotypic traits (size, shape, banding pattern, ground colours – beige, black and dark beige). We sampled individuals from the hybrid zone and implemented targeted capture-sequencing to obtain genotypic data. We identified loci associated with the black and beige ground colours and banding pattern of the shell. No significant associations with the shell shape and size were found which may suggest polygenic and complex architecture, consistent with the theoretical expectation. In addition, our analysis suggests a possible role for chromosomal inversion underlying locally adaptive traits.

This thesis addressed longstanding questions regarding the genetic architecture of the adaptive shell traits in this organism and provides directions for the future follow-up studies. The genetic resources described in this thesis will assist the future studies that may address a wide-range of evolutionary questions in this species.

Acknowledgements

This project and thesis are a product of collaboration and team-work of so many people. I would like to thank everybody who were a part of this and in some way or other contributed to the thesis.

First and foremost, I would like to thank my supervisor Prof. Roger K. Butlin for taking me onboard and mentoring and supporting me throughout this project. This project would not have been what it is without his invaluable suggestions and supervision.

I would also like to thank the other members of the team who were involved in the conceiving of this project and data collection and/or provided their expertise on different aspects of the analyses, especially Anja Westram, Rui Faria, Tomas Larsson, Marina Panova and Prof. Kerstin Johannesson. (Pardon me for I may be missing a few names here, but I am deeply grateful to everybody who in any way supported this project)

I would like to thank the past and present members of the Butlin lab, who have supported or helped me in this journey. And I would like to extend my acknowledgement to the members of *Littorina* team based at the University of Gothenburg, for you all have been wonderful.

I am deeply grateful to the University of Sheffield and the University of Sheffield Vice-Chancellor’s scholarship to provide me this excellent opportunity.

And a big cheer goes to the University of Sheffield HPC team and Dr. Victor Soria-Carrasco, for maintaining excellent computational resources and repositories. Your excellent efforts made life a lot easier.

I would extend my acknowledgement to Dr. Alison Wright and Dr. Angus Davison for agreeing to be my examiners and providing helpful comments to improve this thesis.

Dr. Lucy Cox and Mr. John Glenn, I am very, very thankful to you both. You helped me through one of the most crucial and difficult periods of this journey.

I am very thankful to all my friends here in Sheffield, UK, India and the rest of the world, for their immense support, love and patience with me. Especially, Shivani, Sneha, Amruta, Neerada and Liwen. You ladies are special; you inspire me and bring me immense joy. Thank you for being my friends! A name that deserves special mention is Mr. Romain Villoutreix. Thank you for being there when I needed you and all the discussions. Your optimism and faith gave me strength. And lastly, I would like to thank my parents and my brother, for their constant encouragement, unconditional love, steadfast support and unwavering faith in me. I am what I am because of you guys.

Contents

List of Papers

Elements of the work done in this thesis have been published in the following papers,

1. Westram, A.M., Rafajlović, M., Chaube, P., Faria, R., Larsson, T., Panova, M., Ravinet, M., Blomberg, A., Mehlig, B., Johannesson, K., Butlin, R. (2018). Clines on the seashore: the genomic architecture underlying rapid divergence in the face of gene flow. Evolution Letters; 2(4): 297-309.

   The genetic maps constructed in Chapter 3 were used to place the loci under divergent selection on different linkage groups in the genome and identify genomic blocks of tightly linked divergent loci. The genotype-phenotype association analysis in Chapter 4 was also published. This analysis allowed to check if the genotype associated with adaptive phenotypes is also linked to the genomic regions under divergent selection. Genetic maps were further utilized to partition chromosomes to estimate the contribution of heritability of the detected genomic blocks to complex adaptive phenotypes.

2. Faria, R., Chaube, P., Morales, H., Larsson, T., Lemmon, A., Lemmon, E., Rafajlovic, M., Panova, M., Ravinet, M., Johannesson, K., Westram, A., Butlin, R. (2019). Multiple chromosomal rearrangements in a hybrid zone between *Littorina saxatilis* ecotypes. Molecular Ecology. Accepted author manuscript.

   The genetic maps (Chapter 3) were used to place the clusters of high linkage disequilibrium (genomic rearrangements) on different linkage groups in the genome. The maps found further utilization in placing the boundaries to these clusters and estimating their size in terms of map distance and the impact of putative re-arrangements on recombination.

List of Figures

List of Tables

Chapter 1: General Introduction

1.1 Background

Speciation lies at the core of evolutionary biology. This is the process of splitting of a population into distinct species and underpins the multitude of biological diversity that we observe today on our planet. This problem garnered the interest of several thinkers for centuries, so much so, that it was coined the "mystery of mysteries" in the late 19th century by John Herschel. A milestone in this field was achieved when Charles Darwin published *On the origin of Species* in 1859 and proposed that evolution of species is a gradual process fuelled by natural selection. He further suggested that selection acts at fitness traits that enable adaptation to different habitats and niche, contributing to morphological variations, which on accumulation lead to speciation. Darwin's theory was elegant and clearly linked selection, adaptation and speciation. However, during this gradual process at which point the process of speciation should be considered completed? According to Darwin, the morphological variations form a continuum among individuals and a morphological gap should be taken as an indication of speciation.

Darwin's theory was met with both appreciation and criticism. Huxley (1860), although a firm supporter of Darwin, critiqued that Darwin could not reconcile his theory with the morphologically intermediary hybrid stages, which are generally infertile or inviable- neither of which can be justified as adaptive. It was not until almost 80 years later that the interest of the scientific community was reinstated in Darwin's theory.

In the 1930s and 1940s, evolutionary research took a major turn with the works of Mayr, Dobzhansky, Haldane, Fisher and contemporary population geneticists who insisted on genetic basis of evolution. The integration of genetics with evolutionary biology had an important impact on how evolution was perceived and studied. The study of evolution essentially became the study of changes in allelic frequency over time. This revolutionized how Darwin's theory was viewed. The missing links, such as the mechanism that can introduce new variations in a population via mutation and recombination, could be explained – thus, strengthening Darwin's theory. Perhaps, the most major outcome was the reclassification of species and speciation models (Dobzhansky 1936; Mayr 1942). The new concept stated species as "groups of interbreeding natural populations that are reproductively isolated from other such groups" (Mayr 1942). Under this new concept, the formation of new species required reproductive isolation, or evolution of barriers to reproduction to impede or prevent genetic exchange between individuals of incipient species. This could either be achieved if the individuals of the diverging population cannot mate with each other, or if they do, produce hybrids with reduced fitness. The concept of reproductive isolation, thus, resolved the conundrum of Darwin's theory of evolution and took centre stage in the study of speciation (Coyne and Orr 2004). While some species do appear to follow different modes of speciation [grouped under the umbrella term 'speciation without selection' by Nosil (2012) and Langerhans and Reisch (2013)], now several years and debates later, the biologists could finally agree with Darwin's view.

1.1.1 Reproductive isolation at the core of speciation study

Central to the study of evolution of reproductive isolation is the relative contribution of different evolutionary forces. The evolutionary forces could either be deterministic (e.g., natural selection) or stochastic (e.g., genetic drift, mutation). These evolutionary forces independently may cause shift in allele frequencies. However, in nature, these forces often act in symphony. But how do these evolutionary forces facilitate reproductive isolation?

The evolutionary forces create barriers to reproduction via isolation mechanisms. Mayr (1942) classified reproductive isolation mechanisms into premating and post-mating barriers to genetic exchange. As mentioned above, premating isolation mechanisms act before fertilization and post-mating isolation mechanisms act after fertilization. Premating isolation mechanisms include ecological barriers, for example differences in seasonal occurrence (Futuyma 2014). Two closely related species of cricket, *Gryllus pennsylvanicus* and *G. veletis* provide an example of such. These two species are reproductively isolated as they reach reproductive age in different seasons (Harrison 1979). Ecological barriers can also act via geographical barriers. For example, the two subspecies of squirrel (*Sciurus aberti kaibabensis* and *Sciurus aberti aberti)* cannot interbreed after being geographically isolated for ~10,000 years when Arizona grand canyons were formed (Bono et al. 2018). Premating isolation could also be achieved by mechanical incompatibility, that is the inability to interbreed due to differences in reproductive structures, or behavioural isolation, that is differences in courtship and mating rituals (Futuyma 2014). An example of mechanical incompatibility could be the inability of left-hand coil snails to interbreed with right-hand snails (Gittenberger 1988; Asami et al. 1998; Ueshima and Asami 2003; Hoso et al. 2010). Behavioural isolation has been observed in the closely related species of green lacewings (*C. plorabunda*, *C. adamsi* and *C. johnsoni*) that occur in the same geographical region but do not recognise each other due to different mating songs (Martínez Wells and Henry 1992).

Post-mating isolation mechanisms are further classified as pre-zygotic and post-zygotic barriers to distinguish between gametic incompatibilities and reduced hybrid fitness (Via 2009). Postmating-prezygotic barriers may arise through gametic incompatibilities that do not let either fertilization take place or incompatibility between female reproductive tract and male seminal fluid. An example is observed in the closely related *Drosophila* species *D. simulans*, and *D. sechellia* that result in few to no sperm transfer to produce almost no hybrid offspring (Price 1997). Post-zygotic barriers produce hybrids that are inviable, sterile or poorly adapted to their environment. Several examples of such barriers can be observed in nature. A well-known example comes from the sister species, *Drosophila melanogaster* and *Drosophila simulans*. Sturtevant (1920) observed that a cross between the *D. melanogaster* females and *D. simulans* males only produced hybrid females. While the reciprocal cross produced only viable hybrid sons (Sturtevant 1920). The post-mating isolation mechanisms have received significant attention from evolutionary biologists as they are an important source of reproductive isolation. It was argued that these mechanisms are the product of genic incompatibilities. It was proposed that structural changes in chromosomes may be contributing, as was observed by the recent work in yeast. Delneri et al. (2003) observed lower hybrid fertility in the hybrids of yeast *Sacchromyces cerevisiae* that had reciprocal translocation and those that did not. However, this model did not sufficiently explain hybrid infertility and inviability as observed in nature (Coyne and Orr 2004). Bateson, Dobzhansky and Muller showed that when populations are separated, they may independently gain mutations at different genomic loci, which may not function together in a hybrid if these populations come in a contact again (famously known as the Bateson-Dobzhansky-Muller model) (Bateson 1909; Dobzhansky 1937; Muller 1942; Orr 1996). Indirect evidence for this model comes from a prevalent Haldane's rule. Haldane (1922) observed that "When in the offspring of two different animal races one sex is absent, rare, or sterile, that sex is the heterozygous [heterogametic, i.e., XY or ZW] sex". Haldane's rule was found to be pervasive across taxa and multiple theories were proposed to understand the underlying causes (reviewed in Delph and Demuth 2016). Muller (1942) used the Bateson-Dobzhansky-Muller model to theorize that recessive mutations on sex chromosomes (X – in XY system; Z- in ZW system), that interact to cause hybrid inviability, would be expressed in the heterogametic sex. Known as the 'dominance' theory, ample evidence has been gathered in support over the years (Delph and Demuth 2016), although exceptions have also been found (example, marsupial mammals where both sexes suffer from genic incompatibilities; Delph and Demuth 2016). Charlesworth et al. (1987) posed the question why the X- chromosome plays a large role in post-mating barriers and reasoned that X-linked genes affect the hybrid fitness disproportionately as X-linked genes evolve faster, i.e., under positive selection, X-linked substitution occurs at higher rate than autosomes if adaptive mutations are partially recessive. Faster-X theory has had some mixed support in spite of the weak evidence, although more data has been obtained for faster-Z evolution (Mank et al. 2007, 2010; Vicoso et al. 2013; Sackton et al. 2014). However, the evolution of faster-Z has been attributed to genetic drift rather than selection (Wright et al. 2015).

The distinction between pre- and post-mating barriers is instrumental in the study of speciation to determine how far are the diverging populations along the speciation continuum. For example, reproductive isolation may be weak at the beginning of speciation but grows stronger until reproductive isolation is complete, and speciation is achieved (Figure 1.1).

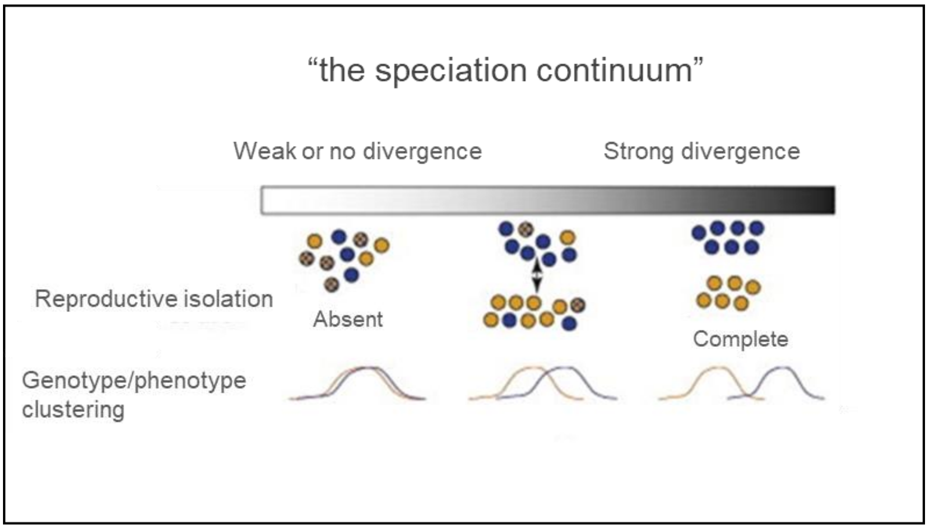


Fig. 1.1: The continuous nature of divergence along the speciation spectrum. Three arbitrary points depict phenotypic and genotypic clustering along the spectrum. Image adapted from Nosil et al. 2009 with permission from Elsevier.

1.1.2 Reproductive isolation and geography

Gene flow is the movement of alleles in and out of populations because of migration. Gene flow can introduce new variations in a population. However, in the absence of other evolutionary forces, gene flow has a homogenising effect on the populations. Therefore, a key question asked in the study of the evolution of reproductive isolation is the presence or absence of gene flow. As is evident by the explanations of the premating ecological barriers and Dobzhansky-Muller model above, evolutionary biologists have often used geographic context to answer this.

The most basic model of speciation assumes a geographical barrier that divides a population into two subpopulations. Geographical isolation would completely prevent gene flow and allow the subpopulations to diverge from one another. Divergence may be facilitated by natural selection; however, genetic drift alone may create enough divergence over time to eventually result in reproductive isolation between the subpopulations. This speciation model - termed allopatric speciation - is widely held among many evolutionary biologists (e.g., Price 2007). However, this model seemed improbable in many cases (e.g. Schliewen et al. 1994; Dieckmann and Doebeli 1999). For example, the well-known example of the Darwin Finches was not only the product of geographical division and natural selection clearly played a role in beak formation (Via 2001). This has led to the development of more complex speciation models that can explain divergence between populations without geographical isolation, or a phenomenon commonly known as ‘speciation with gene flow’ (Coyne and Orr 2004; Nosil 2008).

Sympatric speciation is an extreme model that posits populations with completely overlapping geographic ranges can diverge. Additionally, it has been argued that perhaps allopatric and sympatric models are not completely exclusive, and speciation is an intermediate of both the approaches (Schluter 2001; Rundle and Nosil 2005). The divergence between two populations is initiated in allopatry, and if complete reproductive isolation is not achieved by the time of the secondary contact, speciation proceeds in in the face of gene flow (Schluter 2001; Rundle and Nosil 2005). This model is called the secondary contact model and the limnetic and benthic threespine sticklebacks (*Gasterosteus aculeatus*) are an example (Arnegard et al. 2014).

However, there are examples of primary model as well, in which the divergence is initiated in the adjacent homogenous population leading to parapatric speciation [e.g.; divergence between marine and freshwater populations of threespine sticklebacks (Jones et al. 2012)]. Parapatric speciation is less extreme form of speciation in the face of gene flow and is referred to when gene flow is restricted as only a small proportion of population is in contact. However, the question persists: how can populations with overlapping or adjacent geographic ranges, exchanging genes, diverge from one another?

1.1.3 Speciation mechanisms

Speciation research gradually shifted from geographic context to explain divergence, to evolutionary forces that drive speciation. Although geographical context still has important implications, it perhaps is not the dominant factor in the speciation engine (Butlin et al. 2012).

In the early development of speciation theory, the role of genetic drift as a sole driver in speciation gained prominence. There is limited support for speciation by genetic drift (Sobel et al. 2010), however, it was argued that genetic drift alone cannot cause significant reproductive isolation in large stable populations (Turelli et al. 2001; Coyne and Orr 2004). A special case of genetic drift is founder effect, in which a population is established by a few individuals from a larger ancestral population. An immediate genetic drift in founder population can cause divergence from the parental population to cause reproductive isolation, and this formed the basis of founder effect speciation theory (Mayr 1963; Carson 1968, 1971, 1975; Templeton 2008). Founder effect speciation models, although theoretically plausible, found little support from empirical and experimental evidence (reviewed in Coyne and Orr 2004). Therefore, while the role of genetic drift in speciation is acknowledged (e.g., Lande 1981), theories of genetic drift as a sole propagator of speciation were eventually disfavoured.

In addition, speciation by genetic drift models did not provide a solution to the problem of speciation in the face of gene flow. They are reliant on a complete or partial break in gene flow from the parental population. However, another mechanism involving chance events supported speciation in the face of gene flow. This mechanism is called speciation by polyploidization. Two sympatric species may form a hybrid that has doubled the number of chromosomes and is instantaneously reproductively isolated from both the parental populations. Speciation by polyploidy is more common in plants and accounts for 4% of new angiosperm species (Coyne 2007). However, the occurrence of this speciation model is still rare in animals.

Speciation research eventually became more affixed with the role of selection in speciation. Examples of selection shaping speciation started to mount (e.g.; e.g., Grant and Grant 1995; Nagel and Schluter 1998; Reimchen and Nosil 2002; Mullen and Hoekstra 2008). As Schluter pointed out in his review, the question is no longer whether selection plays a role in speciation, but "how does selection lead to speciation?" (Schluter 2009). Speciation by selection can be broadly classified into two categories: (i) mutation-order speciation, and (ii) ecological speciation (Schluter 2009; Nosil 2012; Langerhans and Reisch 2013).

Mutation-order speciation is the evolution of reproductive isolation through fixation of different advantageous mutations in separate populations adapting to similar selection pressures (or "uniform" selection) (Mani and Clarke 1990; Schluter 2009). Divergence between populations occurs if, by chance, the same mutations do not occur or get fixed in a different order. Divergence, hence, is a stochastic process in this model, yet still distinct from genetic drift as selection is driving the fixation of mutations. Ecological speciation is the evolution of reproductive isolation when populations adapt to contrasting environments via divergent selection (Schluter 2009; Nosil 2012). Divergent selection leads to the fixation of alleles that are advantageous in one environment and not in others. Thereby, as is evident, selection is ecologically based in both the models, but divergence is promoted by ecology only in latter. In essence, under ecological speciation, divergent selection fixes different alleles, whereas, same alleles would be favoured in mutation-order speciation if present. This distinction is important, especially in the context of speciation with gene flow. As this implies that gene flow may prevent divergence between the populations by supplying favourable alleles from one population to other in the mutation-order speciation model (Schluter 2009). Ecological speciation, on the other hand, may proceed with or without gene flow, although more easily without gene flow (Nosil 2008).

1.1.4 Ecological speciation: a reemphasis of Darwin's view

Ecological speciation has acquired many proponents in the recent times (Rundle and Nosil 2005; Schluter 2009; Nosil 2012). Laboratory experiments have yielded results in support of evolution of reproductive isolation by divergent selection (Rice and Hostert 1993). However, the biggest support for ecological speciation comes through the studies of natural population which suggest multiple and independent origin of reproductive isolation evolving between independent populations adapting to the contrasting environments. Some examples include the divergent ecotypes of *Timema* stick insect, cryptic to their host plant species (Soria-Carrasco et al. 2014); or the shell shape repeatedly evolved in the *Littorina saxatilis* marine snails inhabiting different intertidal zones (Johannesson 2010).

The sources of divergent selection driving ecological speciation are extrinsic and ecology-based. Ecologically-based sexual selection can also cause ecological speciation if divergence between mating preferences can be achieved. For example, in the limnetic and benthic threespine stickleback fish, the colours of male and the sensitivity of the females to perceive them vary among lakes in a light-dependent manner, leading to environment-specific signal preferences (Boughman 2001). An important consequence of divergent selection is local adaptation. Also known as the first step to ecological speciation, local adaptation is defined as the "greater average fitness of local individuals compared to immigrants" and often starts within a species as a direct result of selection in a heterogeneous environment (Lenormand 2012). Local adaptation can cause specialization, niche evolution and subsequently lead to evolution of reproductive isolation, and eventually speciation.

Ecological speciation can lead to the evolution of any type of reproductive isolation, i.e., both pre- and post-mating barriers to gene flow. In the case of ongoing gene flow, the outstanding problem remains whether divergent selection could overcome the homogenizing effect of gene flow and lead to speciation in the face of gene flow. During population divergence, genetic differentiation accumulates in the regions under divergent selection while gene flow impedes the divergence in other regions (Gavrilets 2004; Via 2009). That is, the loci under divergent selection and the linked loci would demonstrate greater differentiation in allelic frequencies between the diverging populations than the rest of the neutral genome (Wu 2001). The divergent genomic regions, also known as 'genomic islands', are more resistant to the homogenising effect of gene flow than the neutral or unlinked loci (Via 2009). Therefore, the greater the number and size of genomic islands, the higher the possibility to further reproductive isolation (Nosil 2012). Hence, it is expected under divergence with gene flow that differentiated loci will accumulate in the genomic regions of divergence via linkage disequilibrium in a process called divergence hitchhiking (Nosil 2009; Via 2009; Nosil 2012). This increases the number of genes in an island and the island size. Natural section can maintain the favourable combination of alleles through reduction in gene flow via evolution of tight linkage (Butlin 2005), thus facilitating genomic islands of divergence. In addition, certain genetic architectures (such as chromosomal inversions) that oppose recombination, may also facilitate growth of genomic islands (Noor et al. 2001; Reiseberg 2001). Arguably, this process is easier under allopatry, in which divergence is not impeded by gene flow and increases over time, and genomic islands represent strongly differentiated regions (Nosil 2012).

Clearly, under the ecological speciation model, the link between divergent selection, adaptive divergence and the evolution of reproduction barriers is once again under scrutiny. A common inference here could be that the genes underlying local adaptation, or diverging phenotypic traits, may also constitute the genetic basis of reproductive isolation (Schluter 2009). This may be true in the cases where pleiotropic effects of genes under divergent selection also lead to the evolution of reproductive isolation (Kirkpatrick & Ravigné 2002). An example of this could be the *wingless* gene that affects both the wing colour and mate preference in *Heliconius* butterflies (Kronfrost et al. 2006). Another scenario is when the genes underlying the divergent selection and reproductive isolation are different but are tightly associated with each other through linkage disequilibrium (Smadja and Butlin 2011). Although, maintenance of linkage disequilibrium between unlinked loci may be difficult (Felsenstein 1981), factors such as tight physical linkage or structural patterns such as chromosomal inversion promote maintenance of linkage disequilibrium between genes under divergent selection and genes causing other components of reproductive isolation (Smadja and Butlin 2011). The understanding of genes underlying divergent selection, thus, can enable us to ask questions about the mechanisms of selection, determine the form of reproductive barriers between the species and understand ecological speciation better overall. This demonstrates the importance of identifying the genetic architecture of the adaptive traits under divergent selection, yet very little is known about the genetics of these processes. Henceforth, an important objective in the current literature is to associate the adaptive phenotype with a genomic region/locus underlying that trait.

To narrow down the genomic region/loci underlying the adaptive trait, there are currently three most used approaches – candidate gene studies, forward genetics and reverse genetics (Seehausen et al. 2014; Pardo-Diaz et al. 2015). The candidate gene studies rely on a *priori* knowledge of the pre-specified genes under divergent selection or associated with similar phenotype in other species (e.g., Hoekstra et al. 2006; Smadja et al. 2012). In the natural populations, forward and reverse genetics methods are more commonly used. If there is a prior knowledge of the adaptive phenotype, the forward genetics approach includes genetic mapping of quantitative trait loci (QTL) or association mapping of the genetic factors to a particular phenotype (please see Methodology). The reverse genetics approach involves genomic scans of the diverging populations. The outliers, markers that show excessive differentiation against the neutral genomic background, are indicative of selection on the linked genes. This approach has particularly been useful in the systems where the reproductive isolation is incomplete and there is still some gene flow (e.g., Westram et al. 2014), as the outliers point towards the genomic regions which resist the homogenizing influence of gene flow and are under divergent selection (Schluter 2009).

1.1.5 Incomplete reproductive isolation

The important thread in all these models, except for the fully allopatric model, is the stage of incomplete reproductive isolation where speciation is proceeding in the face of gene flow. This stage is arguably the most informative in the study of speciation as once reproductive isolation has been achieved, only the differences between the species can be studied but not the forces which lead up to it.

The spatial zones which represent diverging taxa in incomplete reproductive isolation, are better known as ‘hybrid zones’. Hybrid zones, or the “natural laboratories of evolution” (Barton and Hewitt 1985), provide a unique opportunity to study the genetic architecture of the diverging traits. In contrast to the controlled hybridization experiments, hybrid zones represent a wide range of genotypes, an assortment of genetic make-up from both the parental populations, from generations of recombination (Barton and Hewitt 1985). This allows fine-scale mapping of divergent traits through analysing the statistically significant phenotype-genotype associations (Buerkle and Lexer 2008; Winkler et al. 2010). Additionally, laboratory hybridization experiments can be conducted only among the organisms that are ‘easy to propagate’ under laboratory conditions. While, hybrid zones provide an opportunity to investigate the natural populations under natural conditions, allowing to examine all the components of the barrier to gene flow (Rieseberg et al. 1999). All to say, hybrid zones are a uniquely suitable framework to study divergent selection and reproductive isolation in natural systems.

1.2 The study

A major shortcoming in our understanding of evolution today is the genes which underlie speciation. Empirical work has been devoted in the literature to address this issue. However, such studies have mostly been limited to the model organisms due to the lack of genomic toolbox.

The speciation studies have benefitted greatly from the advancements in the sequencing technologies, allowing the researchers to carry out tests in non-model organisms. Taking advantage of these technological advancements, I specifically investigated the genetic architecture of the locally adaptive traits, shell characteristics specifically, in the marine snail *Littorina saxatilis.* This non-model species is an excellent example of ecological speciation in process and provide opportunities conducive to ask questions about their speciation (see *section 1.2.1*). The aim of the study is to further enhance the understanding of genetic basis of local adaptation and speciation.

This study was conducted in parallel with the cline analysis in a hybrid zone of *Littorina saxatilis* (published in Westram et al. 2018). The expectation in the cline analysis is that the markers under divergent selection would show clinal (or gradual) frequency change in alleles and the strength of the selection can be estimated by the width of the clinal slope (Barton and Hewitt 1985). The genomic regions underlying adaptation can be identified by mapping genotype-phenotype associations of the divergent traits, or traits whose clinal centre would co-localize with the genotype clines. While this approach allowed us to ask several questions regarding the nature and distribution of the loci under divergent selection, this study strictly explores the genotype-phenotype links of the shell phenotypes which showed clinal variation (or contribute to local adaptation).

1.2.1 The study system

The rough periwinkle, *Littorina saxatilis* (Mollusca: Gastropoda), provides an excellent opportunity to study local adaptation and speciation. The species is distributed across Europe to more remote locations in South Africa and the Americas (Johannesson 1988). Primarily centred around the Northern Atlantic, this species inhabits the rocky shores and forms distinct ecotypes in response to local selection pressures. The two distinct ecotypes adapted to crab predation and wave action can be found across the rocky shores in the UK, Spain and Sweden. The parallel and repeated evolution of the adaptive traits of the two ecotypes (Butlin et al. 2014) provides a unique opportunity to test the theories for evolution and address outstanding questions such as, if the same genomic regions underly the evolution of the adaptive phenotypes (Westram et al. 2016).

The two distinct ecotypes, “Crab” and “Wave” inhabit contrasting environments, few metres apart from each other (as seen in Figure 1.2), and have evolved traits suitable for their respective habitats. The crab ecotype is found in the boulder fields terrorized by crab predation.

Therefore, this ecotype has evolved behavioural and structural traits to survive predatory attacks including a large adult size, thicker shell and a small aperture opening. The wave ecotype is found in the steep cliffs and has evolved traits that help it protect against the strong wave action to be not swept away into the sea. The morphological adaptation includes smaller adult size, a thinner shell and a bigger aperture. These adaptations help the organism find shelter in the crevices or provide more surface area to attach to the rocky cliffs in the face of strong waves. The divergence between the two ecotypes can also be seen in the shell colouration and banding pattern (Reid 1996). There is evidence which supports divergent selection acting on shell colouration (Johannesson and Butlin 2017).

The two ecotypes of this highly polymorphic species meet and mate in hybrid zones (see Figure 1.2; Johannesson et al. 2010). The hybrid zones are narrow, due to limited dispersal and ovoviviparity (Reid 1996), and form at the abrupt environmental transitions from the cliff to boulder. Within the narrow range of the hybrid zones, the continuum of the intermediate phenotypes can be observed. The estimates of the extent of gene flow across hybrid zones, inferred from studies utilizing genetic data suggests the presence of reproductive barriers between the ecotypes (Grahame et al. 2006; Panova et al. 2006; Galindo et al. 2009; Johannesson et al. 2010). This contrasts with the fact that hybrids on the Swedish coast were observed to have high fitness in their habitats (Janson 1983) [however, there also has been some evidence of hybrid inviability (Janson 1985)]. However, there are more conclusive evidence that gene flow may be impeded by other factors such as assortative mating and habitat preferences (Johannesson et al. 2010). Furthermore, the studies on the genetic data between the two ecotypes have further led to the conclusion that the hybrid zones are of primary origin, maintained over generations of hybridization, and represent parapatric speciation (Panova et al. 2006; Hollander et al. 2015; Butlin et al. 2014). Thus, this provides for an ideal system to study divergent selection and perform association mapping to identify the underlying genetic architecture.

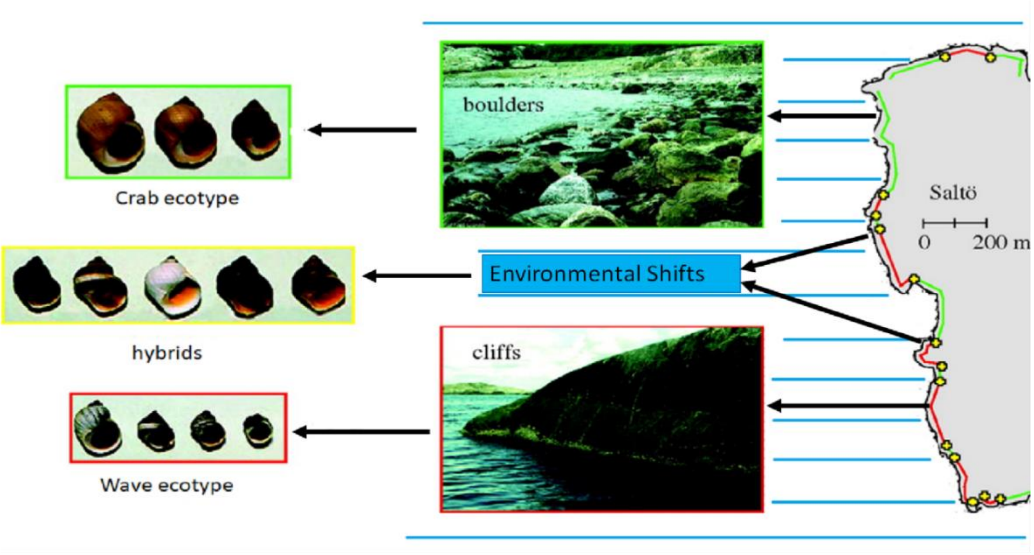


Fig. 1.2: The two ecotypes of *Littorina saxatilis* along the Swedish shore on the island of Saltö. The shore consists of contrasting environment of boulders and cliffs and narrow habitat transitions where hybrid zones form. The ecotypes display distinct structural and behavioural attributes. Hybrids display the intermediate continuum. Image adapted from Johannesson et al. 2010.

Previous studies in this system have aimed at characterizing the genomic regions by using genomic scans to identify highly differentiated outliers in the ecotypes (Galindo et al. 2010; Westram et al. 2014; Westram et al. 2016; Ravinet et al. 2016). These studies have identified a few functional outliers whose function may be attributed to the shell structure formation pathways with some confidence (Galindo et al. 2010; Westram et al. 2016). But more importantly, the differentiating genomic regions under divergent selection have been characterized by these studies. This information was leveraged in the study design. The probes (n=1,136) were included from these previous studies which captured the targeted outlier regions in the genome, along with the probes that captured the random genomic regions (see sections 3.2.1 and 4.2.2).

This study was aided by a reference genome assembly (Westram et al. 2018). The first freeze of the assembly is available and can be accessed at https://cemeb.science.gu.se/research/target-species-imago/littorina-saxatilis.

1.2.2 The study aims

The prerequisites for the project included an annotated reference assembly of the genome of our model organism along with a high-density linkage map for phenotype association with the genomic regions. Therefore, the following aims were identified for the study,

(i) To aid the annotation of the genome assembly (see sections 1.2.3.2 and 2.1), construction of a transcriptome assembly.
(ii) Construction of a high-density linkage map.
(iii) Estimation of the genetic architecture of the adaptive shell traits.

1.2.3 Methodology

In this section, I will briefly summarise the theories and concepts of the methodology used in this thesis.

1.2.3.1 Next generation sequencing

Determining the order of the nucleic acids in the genome is integral to all the genomics research. As put by Fredrick Sanger, “knowledge of sequences could contribute much to our understanding of living matter” (Sanger 1980). These sequences contain information regarding hereditary and biochemical properties. Thus, efforts to elucidate the DNA sequences have been made since the three-dimensional structure of DNA was resolved (Watson and Crick 1953; Zallen 2003).

The first major breakthrough that happened in this field was in 1977, when Sanger and colleagues used the chain termination method to sequence the DNA of the bacteriophage ϕX174 (Sanger et al. 1977). This method basically relied on dideoxynucleotides (ddNTPs), chemical analogues of deoxynucleotides (dNTPs), which lack the 3’hydroxyl group necessary to form bonds with 5’phosphate group of the next dNTP during DNA elongation (Atkinson et al. 1969). The mix of radioactive ddNTPs in the DNA elongation reaction results in different fragment lengths due to the ddNTPs halting the extension. Generally, this method involved performing a DNA elongation reaction with each individual ddNTP base in parallel; and the results were analysed by running the polyacrylamide gel for the four ddNTP bases in four different lanes. Over the years, several changes have been made to the methodology, some of which accommodated automation (Heather and Chain 2016). Yet, the method remained plagued with its lack of time and cost effectiveness and the size of the genome/DNA fragment that can be sequenced (Heather and Chain 2016). Although the method still remains one of the popular sequencing strategies, its limitations demanded the need for more effective and cheaper technology.

The year 2005 ushered in the age of the Next-Generation Sequencing (NGS) with the commercial launch of a massively parallel sequencing platform. The superiority of this latest technology over the previous methodology (now commonly known as the Sanger sequencing) can be assessed by the Wheeler et al. 2008 study, who were able to sequence an entire human genome in a period of 5 months at an expense of $1.5 million. By comparison, Sanger sequencing was used in the Human Genome Project which took a period of 13 years and costed $2.7 billion (Levy et al. 2007). Subsequently, NGS supplanted Sanger sequencing especially, for large scale genome analyses.

Next-generation sequencing, or high-throughput sequencing, is now used as an umbrella-term for the massively parallel sequencing platforms, including Illumina, PacBio; Roche 454, SOLiD and Ion torrent. (Voelkerding et al. 2009; Goodwin et al. 2016). The NGS sequencing approaches may include repeated cycles of oligonucleotide ligation or polymerase-induced nucleotide extensions, of clonally amplified or single DNA molecules spatially separated in a flow cell (Voelkerding et al. 2009). The chemistry and exact methodology will differ across platforms, as reviewed in Goodwin et al. 2016, Heather and Chain 2016 and Voelkerding et al. 2009 (since Illumina platform is used in this study, a brief overview is discussed in the Figure 1.3). Irrespective of the precise mechanisms, inherently NGS platforms can generate enormous amount of data. For example, in a single run of the instrument, Illumina HiSeq 2500 can produce up to 1 TB of data (https://www.illumina.com/documents/products/datasheets/datasheet_hiseq2500.pdf). Naturally, the power of NGS is being utilized in the studies that require large amounts of sequence information, relative quantification, and high-sensitivity detection.

NGS has revolutionized all fields of biology, particularly evolutionary biology. The cost-effective technology has encouraged the researchers to sequence the genomes of several non-model organisms (Ellegren 2014; Tagu et al. 2014; Russell et al. 2017) and test the theories of evolution at the genetic level. Comparative studies of variation in genomes can potentially reveal population histories and give insights into the mechanisms of adaptation and speciation (Gilad et al. 2009). NGS has been used in genome-wide association mapping studies and genome scans to find the genes under ecologically adaptive traits or selection (e.g., Westram et al. 2014; Comeault et al. 2015; Pfeifer et al. 2018). The NGS data has been used to map structural variations, insertions or deletions, and analyse complex species-specific features in the genome (e.g., Ghoneim et al. 2014; Russell et al. 2017; Faria et al. 2019). The ability to compare genomes of species has enhanced the understanding of the phylogenetic relationships (e.g., Rastas et al. 2016). NGS not only is used for DNA sequencing, but also for RNA sequencing (discussed next), which has allowed the researchers to study the transcriptome (see Chapter 2) and its different aspects, gene expression in response to selective pressures for example (e.g., Porcelli et al. 2016). The applications of NGS are not at all restricted to the few mentioned here, and have been reviewed extensively (see for example, Gilad et al. 2009;

Voelkerding et al. 2009; Davey and Blaxter 2010; Ekblom and Galindo 2011; Grover et al. 2012; Han et al. 2015; Goodwin et al. 2016).

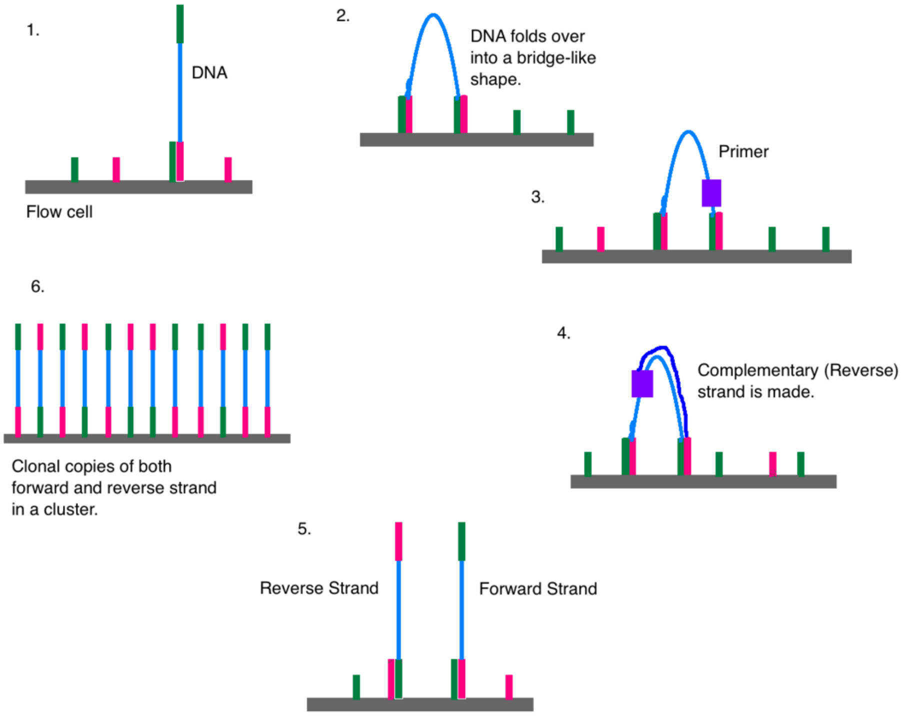

1.
DNA
Flow cell
2.
DNA folds over into a bridge-like shape.
Primer
3.
6.
4.
Complementary (Reverse) strand is made.
Clonal copies of both forward and reverse strand in a cluster.
5.
Reverse Strand
Forward Strand

Fig. 1.3: Chemistry of Illumina sequencing. Illumina sequencing uses reversible terminated chemistry. The process starts with purification of DNA sample and ligation of DNA fragments with adaptors. DNA fragments are then loaded onto a specialized chip. **(1)** The adaptors help DNA fragment attach onto the flow cell via binding with the oligonucleotide coating at the surface of the cells. **(2)** The adaptor sequence at the pen end of the DNA fragment helps it attach to the adjacent oligonucleotide forming a bridge. **(3-4)** Now starts the cluster generation phase, which involves polymerase to synthesize the complementary strand. **(5)** After each round of amplification, the bridge is broken, resulting in the original strand and the complementary strand. **(6)** The process is repeated over and over again to give thousands of copies of the DNA fragment. Post- cluster generation, primers, and fluorescently-tagged and modified nucleotides are added to the chip. These nucleotides have 3’ blockers that allow primers to add only a single nucleotide at a time. After each round of synthesis, camera takes a picture of the flow cell. Based on the wavelength of the fluorescent tags, the machine determines and records the base added. Image by DMLapato (2015) available under CC-BY-SA. Accessed from https://commons.wikimedia.org/wiki/File:Cluster_Generation.png on October 5, 2018.

An important sequencing approach leveraged by NGS is targeted sequencing. Whole genome sequencing may still be an expensive route for some studies, in such cases targeted-sequencing allows to sequence and analyse the regions of interest in the genome (reviewed in Grover et al. 2012). This approach allows the researchers to focus time, cost and effort on specific regions of interest which may include mitochondrial DNA, targets within genes or custom content. This approach has been used extensively in population genetics as it gives the researchers an opportunity to design cost-effective studies with an increased sample size and/or increased gene regions under evaluation, and higher coverage (Grover et al. 2012). Such study designs increase the inference power to estimate and locate patterns of LD and nucleotide diversity

(Grover et al. 2012). Therefore, targeted sequencing approach been particularly useful in the studies assessing the signatures of selection (e.g., Westram et al. 2018) or QTL mapping (e.g., Guo et al. 2016).

Perhaps one of the most important features of NGS in population genetics is the large-scale identification of the molecular markers (Ekblom and Galindo 2011; Davey et al. 2011). Previously, gene mapping studies were seriously limited by the observable phenotype markers (e.g., Mendel 1866; Sturtevant 1913). However, influential works like Botstein et al. 1980 paved the way for the development of gene markers that used DNA polymorphisms. Since, gene markers have been used in multiple studies and are generally favoured over phenotype markers for their abundance in the genome. The process to determine different genetic variants for a genetic marker is called genotyping. Single nucleotide polymorphisms (SNPs) are a very commonly used genetic markers which represent a single base variation in the genome. NGS platforms enable genotyping of several thousands of SNPs to millions in a single reaction. This feature has been used in the chapters III and IV for data generation, implemented with the targeted sequencing approach, in order to capture the genomic regions under divergent selection as determined by the previous studies (see section 1.2, 3.2.1 and 4.2.2).

1.2.3.2 RNA sequencing

The hereditary information of eukaryotic organisms is coded in a genome made up of deoxyribonucleic acid (DNA) and packed on physical structures called chromosomes. Almost all the cells, in a multicellular organism, contain the same genome that remains unchanged through the course of their life. However, cells have distinct function, appearance and response to external stimuli. In a process called the gene expression, specific stretches of DNA, or genes, act as templates to create functional cellular products. The presence and abundance of these cellular products, regulated via a complex signalling system, confer cells with their distinct features.

The central dogma in molecular biology is that DNA does not itself codes for the cellular products. In a process called transcription, DNA serves as a blueprint to create another molecule called ribonucleic acid (RNA). Some RNA molecules can themselves be the end products; otherwise, they serve as a blueprint to create another molecule, protein, as the end products. The RNA molecules that create proteins are called messenger RNA (mRNA).

The study of RNA, rightfully so, has received considerable attention from the scientific community and branched into transcriptomics. Transcriptomics is the study of transcriptome, or the sum total of RNA transcripts of an organism at a particular time. Traditionally, transcriptomics was done using hybridization-based methods (Cieślik and Chinnaiyan 2017). However, in recent times it has gained tremendously from the advancements in the sequencing technologies. RNA-sequencing utilizes NGS to capture all the expressed transcripts at a given moment.

The RNA-seq typically consists of the following steps: isolation of total RNA from the biological samples ,enrichment of mRNA in the sample, cDNA synthesis using the mRNA as template, size selection or fragmentation of the reads according to the sequencing technology, sequencing on a high throughput platform; generation of single or paired-end reads which are further utilized to create a transcriptome assembly (discussed more in section 2.1) and perform downstream analysis (Griffith et al. 2015). The various downstream analyses that are performed using RNA-seq include genome annotation, gene discovery, detection of variants underlying a phenotype, studying the mechanisms of gene regulation and RNA editing and more (reviewed in Griffith et al. 2015). A major application of RNA-seq that has gained much attention among biologists is the study of differential gene expression.

Differential gene expression refers to the level of expression of a gene that regulates the amount of specific protein that is produced in a tissue (Brockmann et al. 2007; Schwanhäusser et al. 2011; Marguerat et al. 2012; Liu et al. 2016). While this process is essential to maintain integrity of certain cellular functions and tissues, it also finds relevance in the adaptive response called phenotypic plasticity. Phenotypic plasticity refers to the ability of an organism to change its phenotype in response to an environmental stimulus without changing its genotype (Schlichting 1986; West-Eberhard 1989; Scheiner 1993). A well-known example of this phenomenon is the eye size variation of the seasonal morphs of the butterfly species *Bicyclus anynana* in response to the changing surroundings to be inconspicuous (Macias-Munoz et al. 2016). Several other examples in nature have been observed where gene expression variation underlies relevant adaptive traits (López-Maury et al. 2008), such as pigmentation in mice (Steiner et al. 2007). Studies have already established a major effect of gene expression on the process of speciation (Wittkopp et al. 2008; Tirosh et al. 2009). It has been argued that geographically isolated populations in different environments may undergo distinct gene expression variations in response to produce extreme phenotypes that can cause hybrid breakdown and sterility (Burke and Arnold 2001; Rogers and Bernatchez 2006). Furthermore, the rate of evolution has been observed to be slower in highly expressed genes (Pal et al. 2001; Drummond et al. 2005; Larracuente et al. 2008) because of more efficient purifying selection in genes that are widely (in many/multiple tissues) expressed. Differences in gene expressions have been analysed in several organisms to characterize candidate genes responsible for phenotypic differences. Phenotypes like castes in ants (Wurm et al. 2010; Ometto et al. 2011), condition dependent sexual dimorphism in *Drosophila* (Wyman et al. 2010), sex-specific phenotypes in *Drosophila* (Ranz et al. 2003) are a few that have been identified using gene expression differences. Although such analysis has mostly been constrained to model organisms, the potential of RNA-seq promises that such studies can be conducted in non-model organisms too.

1.2.3.3 Linkage mapping

The cell division in gamete cells occur through meiosis, the specialized process which reduces the chromosome number to half to give haploid cells. From the standpoint of evolution, a key feature of this meiosis is recombination, that is the exchange of genetic material between the homologous chromosomes during the two-stage process. This maintains the level of genetic variation within a population by ensuring that the combination of alleles inherited by an offspring are distinct from their parents. This process was also observed by Gregor Mendel in his famous pea experiments and described as the principle of independent assortment (Mendel 1866).

The principle of independent assortment states that the genes present on different chromosomes will be inherited independently of each other. By definition, linkage (the tendency of the genes to be inherited together) is not present in different chromosomes. Conversely, genes located physically close to each other tend to be inherited together as it is less likely that recombination will happen between them; hence they are referred to be in linkage. The purpose of linkage mapping is to analyse whether a pair of genetic markers show linkage or assort independently.

To understand, take an example of a double heterozygote (Aa/Bb) where the genes are located on separate chromosomes and the dominant alleles (A/B) are inherited from one parent and the recessive alleles (a/b) from the other parent. During meiosis, four distinct gametes A/B, A/b, a/B and a/b are produced. The gametes with A/B and a/b alleles are parental and the gametes with A/b and a/B alleles are recombinant. If each of the four distinct gametes are produced in equal proportions (25%), the frequency of recombination (θ) between these two genes is 50%, which is the maximum recombination frequency that can be obtained between any two genes and they are called unlinked. If the parental gametes are produced only (50%) and not the recombinants, the two genes would be in complete linkage.

The genetic distances between a pair of genetic markers can be estimated by counting the number of recombinant and parental gametes inherited by the offspring and dividing the number of recombinants with the total number of informative meiosis. A prerequisite for this analysis, thus, is the genetic markers which are polymorphic and heterozygous in the parents, to be able to distinguish between the recombinant and parental gametes. The genes are considered to be in linkage if they deviate significantly from the null hypothesis of independent assortment (θ=0.5). The LOD score, or the log (base 10) of odds ratio, is calculated by dividing the probability of the data under the alternative hypothesis (i.e. the linkage between the two markers) by the probability under the null hypothesis, and the threshold LOD of 3 is commonly used to reject the null hypothesis (Morton 1955). This is also called the two-point linkage test and is used to estimate recombination frequency between two markers and infer if they are in linkage. This is an essential and the first step in the linkage analysis. This segregates the markers into linkage groups, i.e. a group of markers that show linkage with one or several markers within that group, which are also the proxy for chromosomes.

The next step is determining the most likely order of the markers within a linkage group. This is perhaps the most computationally expensive step of the analysis depending on the number of the markers used (see Chapter 2.1). There are multiple algorithms to accomplish the arduous task (Fierst 2015). And a common approach is to add markers gradually while estimating the maximum likelihood and report the most likely order (Green et al. 1990; Stam 1993). In the chapter II, the software Lep-Map2 (Rastas et al. 2016) is used which automates this task and achieves it in two steps: (i) a randomized initial order is estimated using the greedy algorithm and (ii) for the N number of markers in the initial order, $2N^2$ random local changes are made and accepted if the likelihood improves (Rastas et al. 2013). The local changes include – (a) swapping the position of two markers, (b) moving a marker to a random new position, (c) moving the markers at the end of the order to new positions with or without reversing, and (d) reversing the order of three or more adjacent markers (Rastas et al. 2013).

A practical complication in relating the recombination frequency to the distance between loci arises in the cases of double recombination. If the distance between two loci is large, double recombination may occur, which would give the appearance of non-recombining haplotype. This problem is limited with the high-density linkage maps, as fine-scale recombination events can be observed. But it can be a potential issue with sparse markers. To compensate for this problem, Haldane developed a map function to correct for “unseen crossovers” (Haldane 1919) which was later modified by Kosambi (Kosambi 1944). The expected number of recombination events between two loci is expressed as genetic distance in centiMorgans (cM). The genetic distances are additive. The total length of a map is a determinant of the average number of recombination events per meiotic cell.

As mentioned previously, the study of recombination patterns is vital to understand genetic diversity. The recombination patterns tend to vary between species, sexes and individuals (Stapley et al. 2017). Moreover, the recombination patterns tend to vary within an individual genome. While some regions tend to show high recombination rates, known as ‘hotspots’, others tend to have suppressed recombination, known as ‘cold spots’. Now, it is widely being acknowledged that such patterns tend to have an adaptive value and are crucial in the stages of sex chromosome development, for example (Charlesworth 2017; Stapley et al. 2017). Therefore, several approaches have been used to study and characterize the recombination patterns. A cytogenetics approach is to count the formation of chiasmata per cell per chromosome. Needless to say, this approach is severely limited in the organisms that it can be done with and its ability to provide a detailed information of the recombination patterns. The more preferred method of studying recombination patterns is, thus, through the pedigree analysis and constructing linkage maps as described above. The power to elucidate fine-scale recombination patterns is dependent on the number of markers being used. Another indirect approach to study fine- scale recombination patterns is through studying linkage disequilibrium (LD) between two markers in natural populations (Auton and McVean 2007). The genomic regions that experience high recombination over generations tend to show low LD while the regions with lower recombination show high LD. The advantage of this approach lies in the ability to detect many recombination events happened over generations which gives a very fine resolution. The disadvantages of this approach would be the inability to detect recombination patterns for individuals and sexes. Additionally, LD in natural populations can also be affected by natural selection, migration, drift, etc (Sabeti et al. 2007), which makes the interpretation of recombination patterns from this type of data hard.

1.2.3.4 Association mapping

Here, a clear distinction needs to be made between linkage disequilibrium (LD) and genetic linkage. LD refers to the ‘non-random association of alleles’ at two (or more) loci in a population. Unfortunately named, it is easy to mistake LD as being caused by genetic linkage although LD may also exist between genes present on different chromosomes. While LD may occur naturally in the loci in genetic linkage, higher than expected LD may also be maintained between loci by selection or population structure. As a result, LD tends to be higher when linkage is tighter, but recombination may break down LD generated by evolutionary forces.

Association mapping, also known as linkage disequilibrium mapping, takes advantage of historic LD to make non-random associations between genotype markers, on one hand, and genotype-phenotype, on the other. The ultimate of aim of association mapping is to uncover functional variants underlying a phenotype. This approach has been used extensively to unravel genotype-phenotype associations in Mendelian or complex traits (e.g., Aranzana et al. 2005). Complex traits, or quantitative traits, are those that show continuous distribution of phenotypes instead of discrete categories and have complex pattern of inheritance for they are influenced by multiple genetic loci and environmental factors (body height, for example). In spite of advances in genotyping and the large amount of data available, it is very unlikely to obtain the genotype of the functional variant. Genotyping the functional variant is usually the most ideal situation, but the next best situation is to genotype loci that are in LD with the functional variant and can serve as its proxy. The power of association mapping will rely on the degree of LD with the functional variant.

In other words, the idea underlying association mapping is that recombination breaks up the genome into fragments, and the correlation between these genomic regions and the phenotypic variant can be drawn either because the genomic regions are directly involved in phenotypic variance or in LD with the genes involved in phenotypic variance. In its approach, this method is similar to quantitative trait loci (QTL) mapping; where populations of controlled crosses are established, thereby restricting the recombination events. When a genotype variant between two recombination breakpoints occurs with a phenotype variant more than chance in the crosses, genotype-phenotype association is established. Unlike the QTL studies, association mapping is carried out in natural populations, which allows to exploit all recombination events in the history of the sample population (Buerkle and Lexer 2008). LD is usually limited to short genomic distance in natural population, hence, a very high-resolution mapping is possible. But this approach is not free of complications. And a major complication arises due to uncontrolled nature of relatedness among the individuals in the study.

Often, non-random mating generates complex patterns of population structure within natural populations. Population structure within the mapping population can cause spurious associations with phenotypes whose variance is correlated with the genetic relatedness. That is, some genetic markers may appear to be associated with the phenotype while in reality they are only capturing the genetic relatedness among individuals. And the problem intensifies when association mapping is applied to the traits involved in local adaptation (e.g., Aranzana et al. 2005).

Several approaches have been used to address this issue, although none provides a fool-proof solution applicable under all circumstances. A common strategy is to use the random markers to estimate relatedness among individuals within the sampling population. This estimate is indicative of the background sharing of the QTL region and can be used as a baseline to check whether the test marker explains more variation than a random marker (Myles et al. 2009). A commonly used program to correct for population structure is STRUCTURE which infers different populations from within a sample and estimates a Q matrix of individual variations within a population (Pritchard et al. 2000). This matrix can be used as a covariate to control for population structure (e.g., Thornsberry et al. 2001). Alternatively, a pairwise relatedness matrix or kinship matrix can be estimated using random genetic markers and used to reduce the multi-dimensional genotype data to small number of dimensions with principal component analysis (PCA). The axes of variation from these dimensions can be used to estimate ancestry-adjusted genotype-phenotype associations (Price et al. 2006). While the STRUCTURE is a computationally intensive program and is designed for unrelated populations, the latter technique is fast, makes no assumption about the population structure, and generally performs similarly to STRUCTURE (Zhao et al. 2007). However, how to best integrate population structure correction methods in the local adaptation studies remain an ongoing discussion as correction for local differentiation may also reduce the signal to detect true associations corresponding to the locally adapted loci (Atwell et al. 2010; Platt et al. 2010; Johnston et al. 2014).

Another problem that association mapping suffers from is that of ‘missing heritability’. Heritability measures the phenotypic variance that can be explained by the genetic factors. Literature is abundant with studies where only a small amount of variation could be explained for highly heritable traits. For example, human height is known to be 80-90% heritable yet the markers associated with the trait explain only 5% of the variance (Maher 2008). The part of blame may lie with the rare allelic variants, which occur in very low frequency within a population and thus become very difficult to detect due to low power. But the most important factor here is the genetic architecture of the phenotype. That is, the effect size that is exerted by the functional variant to the phenotype. The multiple small effect alleles are inherently harder to detect than few alleles with large effect (Myles et al. 2009). Therefore, understanding of the genetic architecture and not only the functional variants underlying a trait, is of immense interest for the biologists.

1.2.4 The study objectives and thesis outline

To achieve the aims laid down for this project, the following objectives were identified, each of which is addressed in the upcoming chapters;

(i) Chapter 2. Transcriptomic analysis. A key requirement in the genome annotation pipeline and to understand the function of the desired genes is to have a well-annotated set of transcripts. This chapter describes the construction and annotation of the most comprehensive transcriptome assembly of this organism to date.

(ii) Chapter 3. Construction of high-density linkage map for *Littorina saxatilis.* This chapter details the methodology of linkage map construction and addresses the shortcomings of the current techniques and the approach(es) used to tackle them. This is the first linkage map reported for this species, allowing to answer the biologically relevant questions such as the number of linkage groups or the presence or absence of the sex chromosome in this species and gives a glimpse of the sex-specific recombination patterns. It was also crucial for understanding the genomic distribution of regions influenced by divergent selection and the characterisation of chromosomal rearrangements.

(iii) Chapter 4. Association mapping of the traits that contribute to local adaptation and reproductive isolation. This chapter focuses on the genotype-phenotype associations of the locally adaptive traits and addresses the following questions: how heritable these traits are, what are the genes underlying these traits, how are they distributed across the genome and what are their effect sizes.

References

Aranzana, M. J., Kim, S., Zhao, K., Bakker, E., Horton, M., Jakob, K., Lister, C., Molitor, J., Shindo, C., Tang, C., Toomajian, C., Traw, B., Zheng, H., Bergelson, J., Dean, C., Marjoram, P., … Nordborg, M. (2005). Genome-wide association mapping in *Arabidopsis* identifies previously known flowering time and pathogen resistance genes. PLOS Genetics; 1(5): e60.

Arnegard, M. E., McGee, M. D., Matthews, B., Marchinko, K. B., Conte, G. L., Kabir, S., … Schluter, D. (2014). Genetics of ecological divergence during speciation. Nature; 511(7509): 307–311.

Asami, T., Cowie, R. H., Ohbayashi, K. (1998). Evolution of mirror images by sexually asymmetric mating behavior in hermaphroditic snails. American Naturalist; 152:225–236.

Atkinson, M.R., Deutscher, M.P., Kornberg, A., Russell, A.F., Moffatt, J.G. (1969) Enzymatic synthesis of deoxyribonucleic acid. XXXIV. Termination of chain growth by a 2',3'-dideoxyribonucleotide. Biochemistry;8(12):4897-904.

Atwell, S., Huang, Y. S., Vilhjálmsson, B. J., Willems, G., Horton, M., Li, Y., Meng, D., Platt, A., … Nordborg, M. (2010). Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature; 465(7298):627-31.

Auton, A., McVean, G. (2007). Recombination rate estimation in the presence of hotspots. Genome Research; 17(8):1219–1227.

Barton, N.H., Hewitt, G.M. (1985). Analysis of hybrid zones. Annual Review of Ecology and Systematics; 16:113–148.

Bateson, W. (1909). Heredity and variation in modern lights. Darwin and Modern Science; 85–101.

Bono, J. M., Pigage, H. K., Wettstein, P. J., Prosser, S. A., Pigage, J. C. (2018). Genome-wide markers reveal a complex evolutionary history involving divergence and introgression in the Abert's squirrel (*Sciurus aberti*) species group. BMC Evolutionary Biology; 18(1):139.

Botstein, D., White, R.L., Skolnick, M., Davis, R.W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. American Journal of Human Genetics; 32(3):314-331.

Boughman, J.W. (2001). Divergent sexual selection enhances reproductive isolation in sticklebacks. Nature; 411(6840):944-8.

Brockmann, R., Beyer, A., Heinisch, J. J., Wilhelm, T. (2007). Posttranscriptional Expression Regulation: What Determines Translation Rates? PLOS Computational Biology; 3(3):1–9.

Buerkle, C.A., Lexer, C. (2008). Admixture as the basis for genetic mapping. Trends in Ecology & Evolution; 23(12):686-94.

Burke, J. M., Arnold, M. L. (2001). Genetics and the Fitness of Hybrids. Annual Review of Genetics; 35(1):31–52.

Butlin, R. K., Saura, M., Charrier, G., Jackson, B., André, C., Caballero, A., … Rolán-Alvarez, E. (2014). Parallel evolution of local adaptation and reproductive isolation in the face of gene flow. Evolution; 68(4):935–949.

Butlin, R., Debelle, A., Kerth, C., Snook, R.R., … Schilthuizen, M. (2012). What do we need to know about speciation? Trends in Ecology & Evolution; 27(1):27-39.

Butlin, R.K. (2005). Recombination and speciation. Molecular Ecology; 14:2621-2635.

Carson, H. L. (1968). The population flush and its genetic consequences. Population biology and evolution. Syracuse , NY: Syracuse University Press, 123-138.

Carson, H.L. (1971). Speciation and the founder principle. Stadler Genetics Symposium; 3:51-70.

Carson, H.L. (1975). Genetics of speciation. American Naturalist; 109:83-92.

Charlesworth, B., Coyne, J.A., Barton, N.H. (1987). The relative rates of evolution of sex chromosomes and autosomes. American Naturalist; 130:113–146.

Charlesworth, D. (2017). Evolution of recombination rates between sex chromosomes. Philosophical Transactions of the Royal Society B: Biological Sciences; 372(1736): 20160456.

Cieślik, M., Chinnaiyan, A.M. (2017). Cancer transcriptome profiling at the juncture of clinical translation. Nature Reviews Genetics; 19, 93–109.

Comeault, A.A., Flaxman, S.M., Riesch, R., Curran, E., Soria-Carrasco, V., Gompert, Z., Farkas, T.E., Muschick, M., Parchman, T.L., Schwander, T., Slate, J., Nosil, P. (2015). Selection on a genetic polymorphism counteracts ecological speciation in a stick insect. Current Biology; 25(15):1975-1981.

Coyne, J. A., Orr, H. A. (2004). Speciation. Sunderland, MA: Sinauer Associates.

Coyne, J.A. (2007). Sympatric speciation. Current Biology; 17(18): R787-8.

Darwin, C. (1859). On the origin of species by means of natural selection, or, the preservation of favoured races in the struggle for life. London: J. Murray.

Davey, J. W., Blaxter, M. L. (2010). RADSeq: next-generation population genetics. Briefings in Functional Genomics; 9(5-6): 416–423.

Davey, J.W., Hohenlohe, P.A., Etter, P.D., Boone, J.Q., Catchen, J.M., Blaxter, M.L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics; 12:499–510.

Delneri, D., Colson, I., Grammenoudi, S., Roberts, I.N., Louis, E.J., Oliver, S.G. (2003). Engineering evolution to study speciation in yeasts. Nature; 422:68–72

Delph, L.F., Demuth, J.P. (2016). Haldane's Rule: Genetic Bases and Their Empirical Support. Journal of Heredity; 107(5):383–391.

Dieckmann, U., Doebeli, M. (1999). On the origin of species by sympatric speciation. Nature; 400(6742):354-7.

Dobzhansky, T.G. (1936). Studies on hybrid sterility. II. Localization of sterility factors in *Drosophila pseudoobscura* hybrids. Genetics; 21:113–135.

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

Drummond, D. A., Raval, A., Wilke, C.O. (2005). A single determinant dominates the rate of yeast protein evolution. Molecular Biology & Evolution; 23(2):327–337.

Ekblom, R., Galindo, J. (2011). Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity; 107(1):1–15.

Ellegren, H. (2014). Genome sequencing and population genomics in non-model organisms. Trends in Ecology & Evolution; 29(1):51-63.

Faria, R., Chaube, P., Morales, H., Larsson, T., Lemmon, A., Lemmon, E., Rafajlovic, M., Panova, M., Ravinet, M., Johannesson, K., Westram, A., Butlin, R. (2019). Multiple chromosomal rearrangements in a hybrid zone between *Littorina saxatilis* ecotypes. Molecular Ecology. Accepted author manuscript.

Felsenstein, J. (1981). Skepticism Towards Santa Rosalia, or Why are There so Few Kinds of Animals? Evolution; 35:124-138.

Fierst, J.L. (2015). Using linkage maps to correct and scaffold de novo genome assemblies: methods, challenges, and computational tools. Frontiers in Genetics; 6:220.

Futuyma, D. J. (2014) Evolution. Sunderland, MA: Sinauer Associates.

Galindo, J., Grahame, J.W., Butlin, R.K. (2010). An EST-based genome scan using 454 sequencing in the marine snail *Littorina saxatilis*. Journal of Evolutionary Biology; 23:2004–2016.

Galindo, J., Morán, P., Rolán-Alvarez, E. (2009). Comparing geographical genetic differentiation between candidate and noncandidate loci for adaptation strengthens support for parallel ecological divergence in the marine snail *Littorina saxatilis*. Molecular Ecology; 18:919–930.

Gavrilets, S. (2004). Fitness landscape and the origin of species. Princeton, NJ: Princeton University Press.

Ghoneim, D. H., Myers, J. R., Tuttle, E., Paciorkowski, A. R. (2014). Comparison of insertion/deletion calling algorithms on human next-generation sequencing data. BMC Research Notes; 7:864.

Gilad, Y., Pritchard, J. K., Thornton, K. (2009). Characterizing natural variation using next-generation sequencing technologies. Trends in Genetics; 25(10):463–471.

Gittenberger, E. (1988). Sympatric speciation in snails—a largely neglected model. Evolution; 42:826–828.

Goodwin, S., McPherson, J.D., Richard McCombie, W. (2016). Coming of age: Ten years of next-generation sequencing technologies. Nature Reviews Genetics; 17:333–351.

Grahame, J.W., Wilding, C.S., Butlin, R.K. (2006). Adaptation to a steep environmental gradient and an associated barrier to gene exchange in *Littorina saxatilis*. Evolution; 60:268–278.

Grant, P. R., Grant, B. R. (1995). Predicting microevolutionary responses to directional selection on heritable variation. Evolution; 49:241–251.

Green, P., Falls, K., Crooks, S. (1990). Documentation for CRI-MAP. St Louis: Washington University.

Griffith, M., Walker, J. R., Spies, N. C., Ainscough, B. J., Griffith, O. L. (2015). Informatics for RNA sequencing: A web resource for analysis on the cloud. PLOS Computational Biology; 11(8): e1004393.

Grover, C.E., Salmon, A., Wendel, J.F. (2012). Targeted sequence capture as a powerful tool for evolutionary analysis. American Journal of Botany; 99(2): 312–319.

Guo, J., Fan, J., Hauser, B. A., Rhee, S. Y. (2016). Target enrichment improves mapping of complex traits by deep sequencing. G3: Genes|Genomes|Genetics; 6(1):67–77.

Haldane, J. (1919). The combination of linkage values and the calculation of distance between the loci of linked factors. Journal of Genetics, 8:299-309.

Haldane, J. B. S. (1922) Sex ratio and unisexual sterility in hybrid animals. Journal of Genetics; 12(2):101–109.

Han, Y., Gao, S., Muegge, K., Zhang, W., Zhou, B. (2015). Advanced Applications of RNA Sequencing and Challenges. Bioinformatics and Biology Insights; 9(Suppl 1):29–46.

Harrison, R. G. (1979) Speciation in North American Field Crickets: Evidence from Electrophoretic Comparisons. Evolution; 33(4):1009–1023.

Heather, J. M., Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. Genomic; 107(1):1–8.

Hoekstra, H.E., Hirschmann, R.J., Bundey, R.A., Insel, P.A., Crossland, J.P. (2006) A single amino acid mutation contributes to adaptive beach mouse color pattern. Science; 313:101–104.

Hollander J, Galindo J, Butlin RK. (2015). Selection on outlier loci and their association with adaptive phenotypes in *Littorina saxatilis* contact zones. Journal of Evolutionary Biology; 28(2): 328-337.

Hoso, M., Kameda, Y., Wu, S. P., Asami, T., Kato, M., Hori, M. (2010). A speciation gene for left-right reversal in snails results in anti-predator adaptation. Nature communications; 1:133.

Huxley, T.H. (1860). The origin of species. In: Huxley TH ed. Collected essays. Lectures to working men. (republished 1899). London: Macmillan & Co, 22–79.

Janson, K. (1983). Selection and migration in two distinct phenotypes of *Littorina saxatilis* in Sweden. Oecologia; 59:58–61.

Janson, K. (1985). Variation in the occurrence of abnormal embryos in females of the intertidal gastropod *Littorina saxatilis*. Olivi. Journal of Molluscan Studies; 51:64–68.

Johannesson, K. (1988). The paradox of Rockall: why is a brooding gastropod *(Littorina saxatilis*) more widespread than one having a planktonic larval dispersal stage (*L. littorea*)? Marine Biology; 99:507-513.

Johannesson, K., Butlin, R.K. (2017). What explains rare and conspicuous colours in a snail? A test of time-series data against models of drift, migration or selection. Heredity; 118(1):21-30.

Johannesson, K., Panova, M., Kemppainen, P., André, C., Rolán-Alvarez, E., & Butlin, R. K. (2010). Repeated evolution of reproductive isolation in a marine snail: unveiling mechanisms of speciation. Philosophical Transactions of the Royal Society B: Biological Sciences; 365(1547):1735–1747.

Johnston, S.E., Orell, P., Pritchard, V.L., Kent, M.P., Lien, S., Niemelä, E., Erkinaro, J., Primmer, C.R. (2014). Genome-wide SNP analysis reveals a genetic basis for sea-age variation in a wild population of Atlantic salmon (*Salmo salar*). Molecular Ecology; 23:3452–3468.

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., … Kingsley, D. M. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. Nature; 484(7392): 55-61.

Kirkpatrick, M., Ravigné, V. (2002). Speciation by natural and sexual selection: models and experiments. American Naturalist; 159(Suppl 3):S22-35.

Kosambi, D. (1944). The estimation of map distance from recombination values. Annals of Eugenics; 12:172-175.

Kronforst, M. R., Young, L. G., Kapan, D. D., McNeely, C., O'Neill, R. J., Gilbert, L. E. (2006). Linkage of butterfly mate preference and wing color preference cue at the genomic location of *wingless*. Proceedings of the National Academy of Sciences of the United States of America; 103(17):6575-6580.

Lande, R. (1981). Models of speciation by sexual selection on polygenic traits. Proceedings of the National Academy of Sciences of the United States of America, 78(6), 3721-5.

Langerhans, B., Riesch, R. (2013). Speciation by selection: A framework for understanding ecology's role in speciation. Current Zoology; 59:31-52.

Larracuente, A. M., Sackton, T. B., Greenberg, A. J., Wong, A., Singh, N. D., Sturgill, D., Zhang, Y., Oliver, B., Clark, A. G. (2008). Evolution of protein-coding genes in *Drosophila*. Trends in Genetics; 24(3):114–123.

Lenormand, T. (2012). From local adaptation to speciation: specialization and reinforcement. International Journal of Ecology. Ecological speciation (special issue); article ID 508458.

Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., … Venter, J. C. (2007). The Diploid Genome Sequence of an Individual Human. PLOS Biology; 5(10): e254.

Liu, Y., Beyer, A., Aebersold, R. (2016) On the Dependency of Cellular Protein Levels on mRNA Abundance. Cell; 165(3):535–550.

López-Maury, L., Marguerat, S., Bähler, J. (2008). Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. Nature Reviews Genetics; 9: 583.

Macias-Munoz, A., Smith, G., Monteiro, A., Briscoe, A. D. (2016). Transcriptome-Wide Differential Gene Expression in *Bicyclus anynana* Butterflies: Female Vision-Related Genes Are More Plastic. Molecular Biology & Evolution; 33(1):79–92.

Maher, B. (2008). Personal genomes: The case of the missing heritability. Nature; 456:18–21.

Mani, G.S., Clarke, B.C. (1990). Mutational Order: A Major Stochastic Process in Evolution. Proceedings of the Royal Society of London. Series B, Biological Sciences; 240(1297):29-37.

Mank, J.E., Axelsson, E., Ellegren, H. (2007). Fast-X on the Z: rapid evolution of sex-linked genes in birds. Genome Research; 17:618–624.

Mank, J.E., Nam, K., Ellegren, H. (2010). Faster-Z evolution is predominantly due to genetic drift. Molecular Biology & Evolution; 27:661–670.

Marguerat, S., Schmidt, A., Codlin, S., Chen, W., Aebersold, R., Bähler, J. (2012). Quantitative Analysis of Fission Yeast Transcriptomes and Proteomes in Proliferating and Quiescent Cells, Cell; 151(3):671–683.

Martínez Wells, M., Henry, C. S. (1992) Behavioural responses of green lacewings (Neuroptera: Chrysopidae: Chrysoperla) to synthetic mating songs. Animal Behaviour; 44(4):641–652.

Mayr, E. (1942). Systematics and the Origin of Species. New York: Columbia University Press

Mayr, E. (1963). Animal species and evolution. Cambridge, MA: Harvard University Press.

Mendel, G. (1866). Versuche über Plflanzenhybriden. Verhandlungen des naturforschenden Vereines in Brünn, Bd. IV für das Jahr, Abhandlungen, 3-47. (http://www.esp.org/foundations/genetics/classical/gm-65.pdf)

Morton, N.E. (1955). Sequential tests for the detection of linkage. American Journal of Human Genetics; 7(3):277-318.

Mullen, L. M., Hoekstra, H.E. (2008). Natural selection along an environmental gradient: a classic cline in mouse pigmentation. Evolution; 62:1555–1570.

Muller, H. J. (1942). Isolating mechanisms, evolution and temperature. Biology Symposium; 161(3):939–944.

Myles, S., Peiffer, J., Brown, P. J., Ersoz, E. S., Zhang, Z., Costich, D. E., Buckler, E. S. (2009). Association mapping: critical considerations shift from genotyping to experimental design. The Plant Cell; 21(8):2194–2202.

Nagel, L., Schluter, D. (1998). Body size, natural selection, and speciation in sticklebacks. Evolution; 51:209–218.

Noor, M. A., Grams, K. L., Bertucci, L. A., Reiland, J. (2001). Chromosomal inversions and the reproductive isolation of species. Proceedings of the National Academy of Sciences of the United States of America; 98(21): 12084-12088.

Nosil, P. (2008). Speciation with gene flow could be common. Molecular Ecology;17(9):2103-6.

Nosil, P. (2009). Adaptive population divergence in cryptic color-pattern following a reduction in gene flow. Evolution; 58:102-12.

Nosil, P. (2012) Ecological Speciation. Oxford: Oxford University Press.

Nosil, P., Harmon, L. J., Seehausen, O. (2009). Ecological explanations for (incomplete) speciation. Trends in Ecology & Evolution; 24:145–156.

Ometto, L., Shoemaker, D., Ross, K. G., Keller, L. (2011). Evolution of gene expression in fire ants: the effects of developmental stage, caste, and species. Molecular Biology & Evolution; 28(4):1381–1392.

Orr, H. A. (1996). Dobzhansky, Bateson, and the genetics of speciation. Genetics; 144(4): 1331.

Pal, C., Papp, B., Hurst, L. D. (2001). Highly expressed genes in yeast evolve slowly. Genetics; 158(2):927–931.

Panova, M., Hollander, J., Johannesson, K. (2006). Site-specific genetic divergence in parallel hybrid zones suggests non-allopatric evolution of reproductive barriers. Molecular Ecology; 15:4021–4031.

Pardo-Diaz, C., Salazar, C., Jiggins, C.D. (2015). Towards the identification of the loci of adaptive evolution. Methods in Ecology and Evolution; 6(4):445-464.

Pfeifer, S. P., Laurent, S., Sousa, V. C., Linnen, C. R., Foll, M., Excoffier, L., Hoekstra, H. E., … Jensen, J. D. (2018). The evolutionary history of nebraska deer mice: local adaptation in the face of strong gene flow. Molecular Biology & Evolution; 35(4):792-806.

Platt, A., Vilhjálmsson, B.J., Nordborg, M. (2010). Conditions under which genome-wide association studies will be positively misleading. Genetics; 186:1045–1052.

Porcelli, D., Westram, A.M., Pascual, M., Gaston, K.J., Butlin, R.K., Snook, R.R. (2016). Gene expression clines reveal local adaptation and associated trade-offs at a continental scale. Scientific Reports; 6:32975.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics; 38:904–909.

Price, C.S.C. (1997). Conspecific sperm precedence in *Drosophila*. Nature; 388: 663–666.

Price, T. D. (2007). Speciation in Birds. Woodbury, NY: Roberts and Company.

Pritchard, J.K., Stephens, M., Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics; 155: 945–959.

Ranz, J. M., Castillo-Davis, C. I., Meiklejohn, C. D., Hartl, D. L. (2003). Sex-dependent gene expression and evolution of the *Drosophila* transcriptome. Science; 300(5626):1742–1745.

Rastas, P., Calboli, F. C. F., Guo, B., Shikano, T., Merilä, J. (2016). Construction of ultradense linkage maps with Lep-MAP2: Stickleback F2 recombinant crosses as an example. Genome Biology and Evolution; 8(1):78–93.

Rastas, P., Paulin, L., Hanski, I., Lehtonen, R., Auvinen, P. (2013). Lep-MAP: fast and accurate linkage map construction for large SNP datasets. Bioinformatics; 29(24):3128-34.

Ravinet, M., Westram, A., Johannesson, K., Butlin, R., André, C., Panova, M. (2016). Shared and non-shared genomic divergence in parallel ecotypes of *Littorina saxatilis* at a local scale. Molecular Ecology; 25:287–305.

Reid, D.G. (1996) Systematics and Evolution of Littorina. London: Ray Society.

Reimchen, T. E., Nosil, P. (2002). Temporal variation in divergent selection on spine number in threespine stickleback. Evolution; 56:2472–2483.

Rice, W.R., Hostert, E.E. (1993). Laboratory experiments on speciation: what have we learned in 40 years? Evolution; 47:1637–1653.

Rieseberg, L. H. (2001). Chromosomal rearrangements and speciation. Trends in Ecology & Evolution; 16:351–358.

Rieseberg, L. H., Whitton, J., Gardner, K. (1999). Hybrid zones and the genetic architecture of a barrier to gene flow between two sunflower species. Genetics; 152(2):713–727.

Rogers, S. M., Bernatchez, L. (2006). Integrating QTL mapping and genome scans towards the characterization of candidate loci under parallel selection in the lake whitefish (*Coregonus clupeaformis*). Molecular Ecology; 14(2): 351–361.

Rundle, H., Nosil, P. (2005). Ecological speciation. Ecology Letters; 8: 336-352.

Russell, J.J., Theriot, J.A., Sood, P., Marshall, W.F., Landweber, L.F., Fritz-Laylin, L., Polka, J.K., Oliferenko, S., Gerbich, T., Gladfelter, A., Umen, J. (2017). Non-model model organisms. BMC Biology; 15(1): 55.

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., Schaffner, S. F., Lander, E. S., International HapMap Consortium, … Stewart, J. (2007). Genome-wide detection and characterization of positive selection in human populations. Nature; 449(7164):913-918.

Sackton, T.B., Corbett-Detig, R.B., Nagaraju, J., Vaishna, L., Arunkumar, K.P., Hartl, D.L. (2014). Positive selection drives faster-Z evolution in silkmoths. Evolution; 68:2331–2342.

Sanger, F. (1980). Frederick Sanger — Biographical. (URL http://www.nobelprize.org/nobel_prizes/chemistry/laureates/1980/sanger-bio.html)

Sanger, F., Nicklen, S., Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America; 74(12):5463–5467.

Scheiner, S. M. (1993). Genetics and Evolution of Phenotypic Plasticity. Annual Review of Ecology and Systematics; 24(1):35–68.

Schlichting, C. D. (1986). The Evolution of Phenotypic Plasticity in Plants. Annual Review of Ecology and Systematics; 17(1):667–693.

Schliewen, U. K., Tautz, D., Pääbo, S. (1994). Sympatric speciation suggested by monophyly of crater lake cichlids. Nature; 368(6472):629-632.

Schluter, D. (2001). Ecology and the origin of species. Trends in Ecology & Evolution; 16:372–380.

Schluter, D. (2009). Evidence for ecological speciation and its alternative. Science 323: 737–741.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., Selbach, M. (2011). Global quantification of mammalian gene expression control. Nature; 473:337.

Seehausen, O., Butlin, R.K., Keller, I., Wagner, C.E., Boughman, J.W., Hohenlohe, P.A., Peichel, C.L., Saetre, G.P., Bank, C., Brännström, A., ... Widmer, A. (2014). Genomics and the origin of species. Nature Review Genetics; 15(3):176-192.

Smadja, C., Butlin, R.K. (2011). A framework for comparing processes of speciation in the presence of gene flow. Molecular Ecology; 20:5123–5140.

Smadja, C.M., Canback, B., Vitalis, R., Gautier, M., Ferrari, J., Zhou, J., Butlin, R.K. (2012). Large-scale candidate gene scan reveals role of chemoreceptor genes in host plant specialization and speciation in the pea aphid. Evolution; 66:2723-2738.

Sobel, J. M., Chen, G. F., Watt, L. R., Schemske, D. W. (2010). The biology of speciation. Evolution; 64: 295-315.

Soria-Carrasco, V., Gompert, Z., Comeault, A.A., Farkas, T.E., Parchman, T.L., Johnston, J.S., Buerkle, C.A., Feder, J.L., Bast, J., Schwander, T., Egan, S.P., Crespi, B.J., Nosil, P. (2014). Stick insect genomes reveal natural selection's role in parallel speciation. Science; 344:738-742.

Stam, P. (1993). Construction of integrated genetic linkage maps by means of a new computer package: JoinMap. The Plant Journal; 3: 739-744.

Stapley, J., Feulner, P. G. D., Johnston, S. E., Santure, A. W., Smadja, C. M. (2017). Variation in recombination frequency and distribution across eukaryotes: patterns and processes. Philosophical Transactions of the Royal Society B: Biological Sciences; 372(1736):20160455.

Steiner, C. C., Weber, J. N., Hoekstra, H. E. (2007). Adaptive variation in beach mice produced by two interacting pigmentation genes. PLOS Biology; 5(9):1–10.

Sturtevant, A. H. (1913). The linear arrangement of six sex-linked factors in *Drosophila*, as shown by their mode of association. Journal of Experimental Zoology; 14:43-59.

Sturtevant, A. H. (1920). Genetic Studies on *Drosophila simulans*. I. Introduction. Hybrids with *Drosophila melanogaster*. Genetics; 5(5):488-500.

Tagu, D., Colbourne, J.K., Nègre, N. (2014). Genomic data integration for ecological and evolutionary traits in non-model organisms. BMC Genomics; 15(1):490.

Templeton, A. R. (2008). The reality and importance of founder speciation in evolution. Bioessays; 30:470-479.

Thornsberry, J.M., Goodman, M.M., Doebley, J., Kresovich, S., Nielsen, D., Buckler, E.S. (2001). Dwarf8 polymorphisms associate with variation in flowering time. Nature Genetics; 28: 286–289.

Tirosh, I., Reikhav, S., Levy, A. a, Barkai, N. (2009). A yeast hybrid provides insight into the evolution of gene expression regulation. Science; 324(5927):659–662.

Turelli, M., Barton, N.H., Coyne, J.A. (2001). Theory and speciation. Trends in Ecology & Evolution; 16:330–343.

Ueshima, R., Asami, T. (2003). Single-gene speciation by left-right reversal—a land-snail species of polyphyletic origin results from chirality constraints on mating. Nature; 425:679.

Via, S. (2001). Sympatric speciation in animals: the ugly duckling grows up. Trends in Ecology & Evolution; 16(7):381-390.

Via, S. (2009). Natural selection in action during speciation, Proceedings of the National Academy of Sciences; 106(Suppl 1):9939–9946.

Vicoso, B., Emerson, J.J., Zektser, Y., Mahajan, S., Bachtrog, D. (2013). Comparative sex chromosome genomics in snakes: differentiation, evolutionary strata, and lack of global dosage compensation. PLOS Biology; 11:e1001643.

Voelkerding, K.V., Dames, S.A., Durtschi, J.D. (2009). Next-Generation Sequencing: From Basic Research to Diagnostics. Clinical Chemistry; 55(4):641-658.

Watson J., Crick F. (1953). Molecular structure of nucleic acids. Nature; 171:709–756.

West-Eberhard, M. J. (1989). Phenotypic Plasticity and the Origins of Diversity. Annual Review of Ecology and Systematics; 20(1):249–278.

Westram, A. M., Galindo, J., Alm Rosenblad, M., Grahame, J. W., Panova, M., Butlin, R. K. (2014). Do the same genes underlie parallel phenotypic divergence in different *Littorina saxatilis* populations? Molecular Ecology; 23(18):4603–4616.

Westram, A.M., Panova, M., Galindo, J., Butlin, R.K. (2016). Targeted re-sequencing reveals geographic patterns of differentiation for loci implicated in parallel evolution. Molecular Ecology; 25:3169–3186.

Westram, A.M., Rafajlović, M., Chaube, P., Faria, R., Larsson, T., Panova, M., Ravinet, M., Blomberg, A., Mehlig, B., Johannesson, K., Butlin, R. (2018). Clines on the seashore: the genomic architecture underlying rapid divergence in the face of gene flow. Evolution Letters 2(4): 297-309.

Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.J., Makhijani, V., … Rothberg, J.M. (2008) The complete genome of an individual by massively parallel DNA sequencing. Nature; 452:872–876.

Winkler, C.A., Nelson, G.W., Smith, M.W. (2010). Admixture mapping comes of age. Annual Review of Genomics & Human Genetics; 1:65-89.

Wittkopp, P. J., Haerum, B. K., Clark, A. G. (2008) Regulatory changes underlying expression differences within and between *Drosophila* species. Nature Genetics; 40(3):346.

Wright, A.E., Harrison, P.W., Zimmer, F., Montgomery, S.H., Pointer, M.A., Mank, J.E. (2015). Variation in promiscuity and sexual selection drives avian rate of Faster-Z evolution. Molecular Ecology; 24:1218–1235.

Wu, C. (2001). The genic view of the process of speciation. Journal of Evolutionary Biology; 14: 851-865.

Wurm, Y., Wang, J., Keller, L. (2010). Changes in reproductive roles are associated with changes in gene expression in fire ant queens. Molecular ecology; 19(6):1200–1211.

Wyman, M. J., Agrawal, A. F., Rowe, L. (2010). Condition-dependence of the sexually dimorphic transcriptome in *Drosophila melanogaster*. Evolution; 64(6):1836–1848.

Zallen, D.T. (2003). Despite Franklin's work, Wilkins earned his Nobel. Nature; 425:15.

Zhao, K., Aranzana, M.J., Kim, S., Lister, C., Shindo, C., Tang, C., Toomajian, C., Zheng, H., Dean, C., Marjoram, P., Nordborg, M. (2007). An *Arabidopsis* example of association mapping in structured samples. PLOS Genetics; 3: e4.

2.1 Introduction

Transcriptomics is the study of the transcriptome and its dynamic nature. A transcriptome can be defined as the transcribed region of genome. The transcribed region of genome, or genes, usually constitute a smaller percentage of genome, which perhaps indicate that transcriptomes are less complicated than genomes. However, complexities of a transcriptome increase due to post-transcriptional processes such as alternative splicing (see Figure 2.1), due to which a single gene can potentially give rise to thousands of transcripts (Graveley 2001).

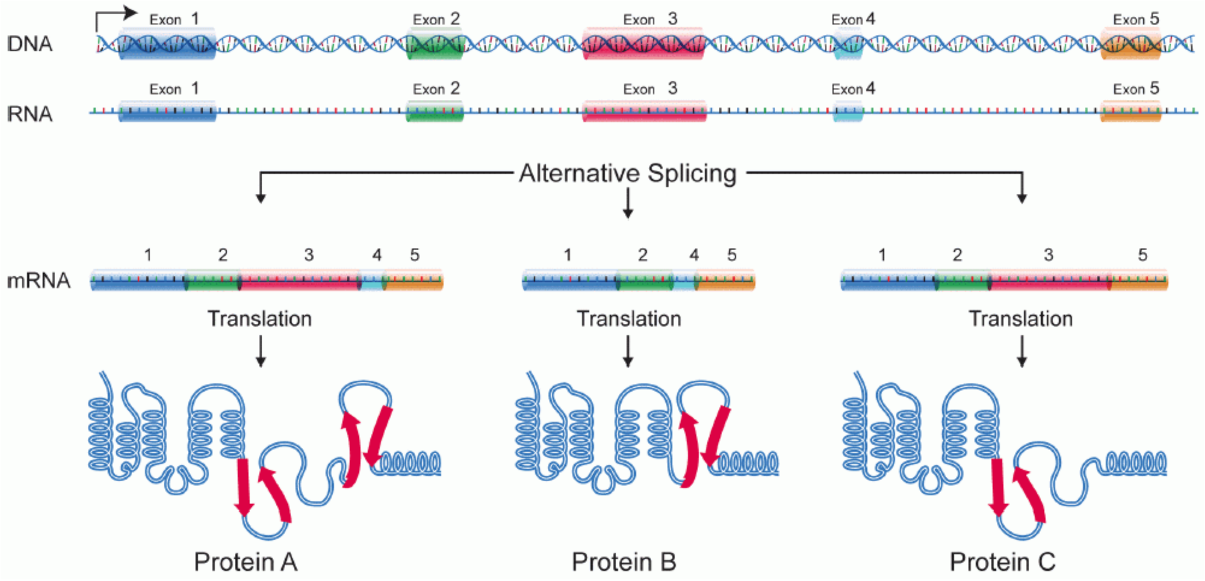


Fig. 2.1: Alternative splicing. This is regulated process during gene expression through which a single gene can potentially produce multiple proteins. For example, in the given Figure, the exons from the hypothetical gene get excluded or included during the transcription process which eventually get translated to different proteins. See Wang et al. 2015 for review. Image available under Creative Commons License. Accessed from National Human Genome Research Institute - http://www.genome.gov/Images/EdKit/bio2j_large.gif on October 5, 2018.

To capture all the expressed transcripts in the cells despite the above-mentioned complexities requires an assay that is both sensitive and selective. And the advancements in RNA-sequencing (RNA-seq) technologies promise an enormous potential. RNA sequencing uses next-generation sequencing (NGS) to obtain all the transcripts expressed in a cell, tissue or sample at a given moment. These raw reads, prior to downstream analyses, are assembled into genomic features mainly via the following approaches,

(i) Genome-guided assembly. If a reference genome assembly of the target organism is already available, ‘genome-guided’ or ‘*ab initio*’ method is utilized. This method involves aligning the raw reads to the genome with an aligner that can map reads with an additional complexity of mapping over intermittent intervals, basically splice-aware alignment as implemented in TopHat (Trapnell et al. 2009) or GSNAP (Wu and Nacu 2010). The overlapping mapping reads for each locus are clustered and algorithms such as Cufflinks (Trapnell et al. 2012) or Scripture (Guttman et al. 2010) traverse through each cluster to determine individual isoforms at that locus. The details of the methodology and the algorithms implemented in different software are reviewed in Martin and Wang 2011. The reference-based method is useful only in the presence of a reference genome assembly, or rather, a high-quality reference genome assembly. A fragmented, misassembled or low-quality genome may lead to the misassembled or incomplete transcriptome assembly. Additionally, the aligners require a maximum length of introns expected in the organism which is used to search for splice alignments. This implies that for the introns spanning larger than the specified intron-values, some splice variants may be missed. The complications may worsen with the use of short- reads often generated by the NGS platforms such as Illumina. The alignment software needs to deal with the short-reads that may map equally well at multiple locations by trying to maximise the possibility to identify true transcription regions as opposed to regions with no transcription in the genome (Martin and Wang 2011). In spite of all the complications, genome-guided assembly is widely used with high-quality references for the advantages that it provides in the terms of computational speed, sensitivity and robustness. Not only this method is computationally faster, it requires less computational resources as well. Since the first step in this method is alignment to the genome, it ensures that contaminants or sequencing errors will not be mapped and included in the downstream analysis. Moreover, this method is very sensitive to low coverage reads and can successfully map the transcripts with low expression (Martin and Wang 2011). The genome-guided transcriptome assemblies have been employed in various studies (e.g., Sharma et al. 2014; Kanost et al. 2016) and have been applied to identify the exon-intron boundaries (e.g., Kanost et al. 2016).

(ii) *De novo* assembly. This method does not require a reference genome and thus, is already useful in the scenarios when the reference genome assembly is missing or of poor-quality. Most of the assemblers which implement *de novo* assembly are based on the de Bruijn graph algorithm. This algorithm basically uses the redundancy of the short reads to identify the overlapping regions to form contiguous sequences. A single short read can form contiguous sequences with two or more reads, which in turn would continue the chain of forming contigs. This results in very elaborate and complex graphs being generated for each seed read. The software such as Oases (Schulz et al. 2012) or Trinity (Grabherr et al. 2011) traverse through each of these graphs to identify the isoforms. Besides the massive computing resources needed for this method, a major disadvantage of this method is its sensitivity to the sequencing errors or chimeric reads (Martin and Wang 2011). The advantages of *de novo* assembly can be summarized by its power to detect novel transcripts that may be missing from the genome assembly, or by the lack of dependence on the prior knowledge about the introns or the splice-sites which may allow to capture isoforms more efficiently (Martin and Wang 2011).

The Trinity suite (Grabherr et al. 2011) is one of the most commonly used software packages for assembling *de novo* transcriptome assemblies. Like most other, *de novo* transcriptome assemblers, Trinity is based on the de Bruijn graph algorithm and is known to perform slightly better than the other *de novo* transcriptome assembly software which uses the same algorithm [e.g., SOAPdenovo-Trans (Xie et al. 2014), Oases (Schulz et al. 2012)] (Rana et al. 2016). Trinity implements the de Bruijn graph algorithm via the three integrated modules – (i) Inchworm (ii) Chrysalis (iii) Butterfly. The detailed methodology of the Trinity software can be seen in Figure 2.2.

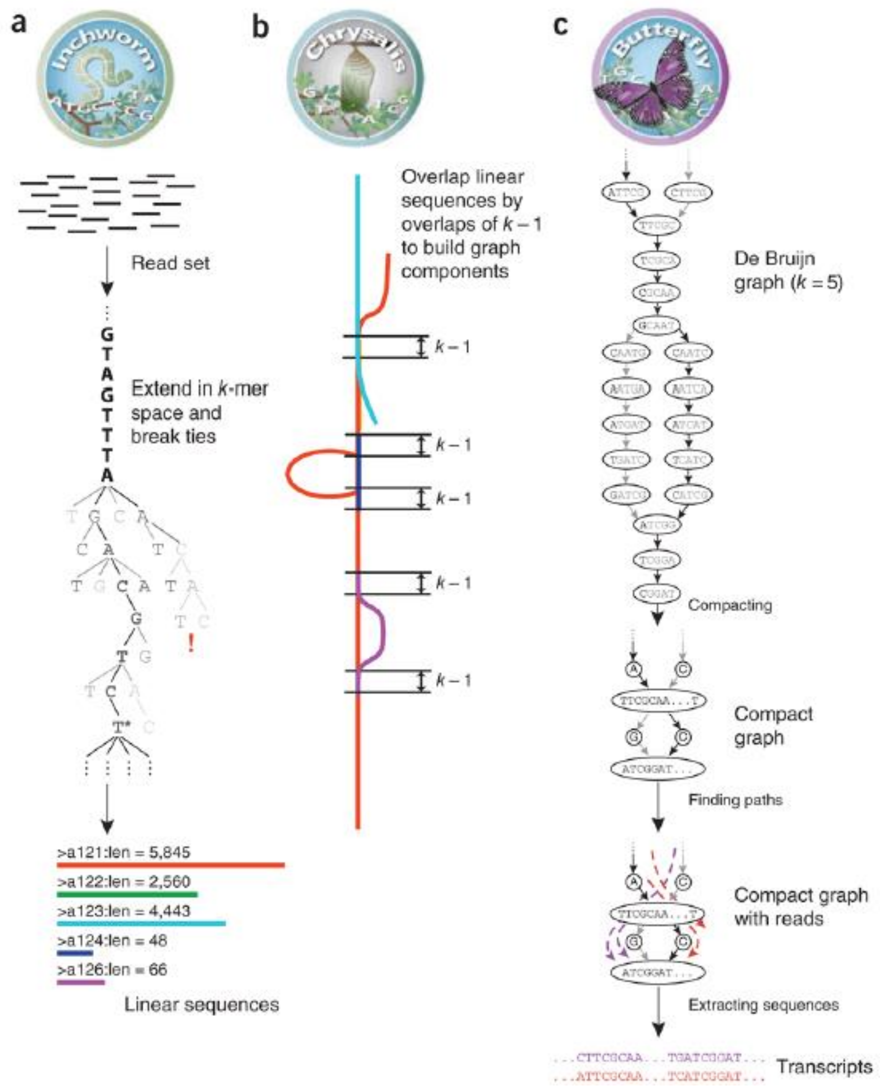


Fig. 2.2: The Trinity suite. **(a)** Inchworm assembles the reads (the black lines at the top) into contigs (coloured lines at the bottom) by looking for overlapping k-mers via greedy algorithm (the paths in the middle) in such a way that each k-mer is represented in each contig only once. **(b)** Chrysalis takes the contigs and build de Bruijn graphs for all the contigs which share at least a single k-1-mer. **(c)** Butterfly takes all the de Bruijn graphs from the previous step and traverses through each graph to find the shortest path, while removing spurious edges (middle). The graphs are reconciled and collapsed to give linear sequences for each spliced form (coloured sequences at the bottom). Image taken from Grabherr et al. 2011. Reprinted by permission from Springer Nature Customer Service Centre GmbH.

Using the Trinity software may be laden with certain complications, mostly associated with problems with the *de novo* assembly construction. As discussed above, these problems usually range from excessive computational time to generation of thousands of transcripts when only smaller real numbers of genes are predicted. The excess number of transcripts is a major by-product of the de Bruijn graph algorithm and also could be a result of the sequencing errors. It is, therefore, required to identify the high-quality, or biologically real transcripts from the low-quality ones. This brings to the next crucial step, which is the quality assessment of the assembled transcripts. For different needs, there are different methods. The common approach is to check the contig statistics such as N50 or the length of the contigs. N50 is the measure to describe the completeness of an assembly, described as the shortest length of contig that can explain 50% of the assembly. While it is a widely used statistic in the genome assembly, the relevance of N50 has been questioned for transcriptome assembly (Li et al. 2014) as the presence of smaller isoforms may confound the results. Therefore, it is usually recommended to use the basic contig statistics along with other methods. The commonly used methods for this purpose include aligning the raw reads back to the transcriptome and then assessing which contigs are supported by the raw reads. Another method that is commonly employed is using the BLAST software to align the transcripts to a transcriptome of a related species, if available, and then assessing how many transcripts are recovered. There are programs like DETONATE (Li et al. 2014) and TransRate (Smith-Unna et al. 2016) which implement a couple or all of the above approaches to assess the qualities of the transcriptome and provide a comprehensive statistic for each transcript and transcriptome. However, caution must be taken while solely relying on these programs as they have their limitations; for example, biologically relevant chimeric transcripts are penalized harshly in both software (Li et al. 2014; Smith-Unna et al. 2016). Another popular approach is to use the CEGMA (Parra et al. 2007) or BUSCO (Simão et al. 2015) pipeline to assess the completeness of the assembly but these tools analyse only the conserved homologs from their databases and can only be indicative of certain attributes of the assembly, like fragmentation or duplication of the contigs. In short, there is no gold method to assess the high-quality, biologically relevant transcripts from the assembled transcripts and therefore, a general approach can be to use a combination of these methods.

In spite of certain problems, the current RNA-seq and transcriptome assembly technologies have revolutionized the field of biology in many ways. The cost-effectiveness of RNA-seq has not only allowed the researchers to carry out sequencing at deep coverage for multiple individuals, effectively increasing the accuracy of the gene expression analysis (Martin and Wang 2011), but also, extend the transcriptomic analyses to non-model organisms where the reference genomes are usually unavailable. RNA-seq data has become indispensable even in genomics. A key step in genome annotation pipelines involves integration of transcriptome and genome to gain substantial evidence of the presence of a gene in a particular region of the genome for structural and functional annotation. Some software may even use this evidence to train their algorithm to predict *ab initio* gene models in the genome. For example, the Maker pipeline (Cantarel et al. 2008) is widely used in the annotation of the eukaryotic genome. First, Maker generates evidence by aligning EST and/or RNA data to the genome. The information gathered through the alignments, regarding the coding regions, introns, UTR and splice variants, is then utilized by the probabilistic hidden markov model (HMM) to predict genes in the genome based on the genome-specific statistical properties of the protein-coding sequences (Cantarel et al. 2008).

The rough periwinkle gastropod, *Littorina saxatilis,* is an excellent model to study evolutionary biology. However, a major obstruction to work on this organism is the lack of or limited genomic resources. A transcriptome assembly for this organism is already available at http://mbio-serv2.mbioekol.lu.se/Littorina1/ (Canbäck et al. 2012), however it contains only 2,543 annotated contigs, that is only 9% of the assembled contigs in the transcriptome. Hereby, this assembly will be referred to as the LSD assembly. The reference genome of this organism is under preparation (Panova, Larsson et al., in preparation) and to assist in the high-quality annotation of the genome, a more comprehensive transcriptome was required. To accomplish that, RNA was isolated by M. Panova (University of Gothenburg) from multiple tissues of two different individuals (one male and one female) - ensuring diversity in the transcripts that can be obtained. A *de novo* assembly was performed to be able to capture the novel transcripts and the spliced variants and avoid the assembly errors that may have resulted from the fragmented draft genome assembly at the time.

In this chapter, I describe the construction and annotation of the transcriptome assembly. The raw reads were obtained by high-throughput sequencing and *de novo* assembly was performed using Trinity v2.0.6 suite (Grabherr et al. 2011). The quality of the assembly was assessed using multiple methods and the good-quality transcripts were selected for annotation. The annotation of the good-quality transcripts was performed using the Trinotate suite (Bryant et al. 2017). In addition, the good-quality transcripts were further utilized to measure tissue specific gene expression. Preliminary results from the analysis are discussed.

2.2 Materials and Methods

2.2.1 Sample preparation and sequencing

The samples, a male and a female individual of the Crab ecotype, were collected from Ängklåvbukten on the island of Saltö on the Swedish west coast. Tissue libraries were obtained from the reproductive organ, hepatopancreas, head, mantle and foot from both individuals, plus penis from male. RNA was extracted from the fresh tissues using Tri reagent followed by a treatment with DNAse Turbo and a clean-up with the Qiagen RNEasy Mini clean-up kit according to the manufacturer’s protocol (Qiagen).

The RNA samples were frozen at -80^{o}C in liquid nitrogen and shipped to BGI (Shenzhen, China) for cDNA library preparation and paired-end sequencing was done on an Illumina HiSeq 2000 instrument. Both raw reads and cleaned data sets were provided by BGI. The dataset was checked for quality using FastQC (Andrews 2010). For the analysis, the clean dataset provided by BGI was used.

2.2.2 Transcriptome assembly

To maximise the diversity and completeness of the *de novo* transcripts, all the tissue libraries from the clean dataset were concatenated. The reads were normalized to reduce the computational burden using the default parameter within the software Trinity v2.0.6 (Haas et al. 2013). Normalization aids in doing so by reducing the number of reads while maintaining the read diversity and complexity. The normalized reads were used for the construction of the *de novo* transcriptome assembly with Trinity v2.0.6. Multiple transcriptome assemblies were constructed by differing the parameters for the Inchworm and Butterfly modules. Different kmer lengths (20, 23, default – 25, 27 and 30) were tested in the Inchworm module. The parameters to merge similar transcripts under transcript reduction settings were modified in the Butterfly module. The minimum percent identity to merge two paths to reconstruct a single transcript was relaxed from the default (98% to 95%) and the maximum gap/difference allowed to combine the two paths was also changed (from default 2% to 5%). The assembly statistics generated by Trinity were compared for all the assemblies. Lower number of unigenes, or closer match to the predicted gene models in *Littorina saxatilis*, was taken to be an indication of less reconstruction of spurious transcripts. Greater number of bases assembled indicated the diversity of reads eventually used in assembly construction. Higher values for contig N50 and average contig length were considered to be indicative of more complete transcript reconstruction. Based on the above criteria, it was observed that – (i) default Butterfly parameters produced the lowest number of unigenes (although, still a lot more than the expected number of gene models), (ii) default Butterfly parameters produced the optimum transcript reconstructions (higher contig N50, average contig length and greater number of assembled bases), and (iii) different kmer lengths did not make any significant difference to the above-mentioned criteria. The parameters for Chrysalis were not modified as it is the most computationally and time intensive module and changing the parameters may have increased the time or computational efforts. Therefore, the assembly constructed with default parameters for each of the three consecutive modules of Trinity (Inchworm, Chrysalis and Butterfly) was eventually used for the downstream processing.

2.2.3 Refinement of the transcriptome assembly

The task to pick the high quality and biologically true transcripts from the initial transcriptome assembly is not an easy one. This task was performed as described below.

Firstly, the initial assessment of transcripts was done using TransRate v1.0.3 (Smith-Unna et al. 2016). TransRate does this assessment via three in-built modules that (i) inspect the contig sequences and report statistics as N50, mean length of the assembled contigs, mean ORF percent, etc., (ii) maps the reads back to the assembled contigs and assigns each contig scores based on (a) alignment of both the read pairs, (b) orientation of the pairs, (c) alignment on the same contig without the overlapping ends, and lastly, (iii) Conditional Reciprocal Best BLAST (CRBB) with a closely related species. CRBB does pairwise alignment from query to target database and vice-versa and applies appropriate e-value cut-off for the alignments based on the relatedness of the two datasets (Aubry et al. 2014). This module gives statistics as CRBB hits, number and proportion of contigs with CRBB hits and number and proportion of references with CRBB hits. CRBB hits are characterized as the top alignments in the reciprocal BLAST. For the CRBB module, the transcriptome assembly was compared with the other closest completed and publicly available molluscan assemblies, *Lottia gigantea* (genome.jgi.doe.gov/Lotgi1.download.ftp.html) and *Biomphalaria glabrata* (http://biology.unm.edu/Biomphalaria-genome). These mollusc genomes diverged from *Littorina* genus ~530 mya and ~418 mya respectively (Reid et al. 2012; Zapata et al. 2014). Following these modules, TransRate assigns assembly scores and reports a set of well-assembled or 'good contigs' by automatically optimizing cut-off scores for contigs for the most optimum assembly score (Smith-Unna et al. 2016). The search for the ORFs was performed using TransDecoder (Haas et al. 2013) and the transcripts with at least one ORF were retained. This set of transcripts was designated refinement assembly 1.

Secondly, the raw reads were mapped back to the transcriptome using RSEM (Li and Dewey 2011). The matrix of raw counts for each transcript per tissue library was obtained using a custom script provided in the Trinity suite. The purpose of this matrix was to filter transcripts with very low counts as they are very unlikely to be of any biological relevance because each transcript needs to have a certain level of expression before it can be translated into proteins. To filter, the raw counts were normalized as counts-per-million by the following formula,

$$CPM_i = \frac{X_i}{N} . 10^6$$

where,

X_i denote raw counts for a transcript in tissue library i

N denotes the tissue library size

The purpose of this normalization of the raw counts prior to the filtering is to account for the different tissue library sizes. The transcripts that were expressed in at least one library, with a total CPM of at least 1 across all the libraries, were retained. This set of transcripts was designated refinement assembly 2.

On one hand, the expression-based filtering approach creates bias against the lowly expressed transcripts, and TransRate promises to capture such transcripts (Smith-Unna et al. 2016). On the other hand, TransRate may penalize low coverage fragmented transcripts or biological novelties if the CRBB module is used with a very divergent transcriptome (Smith-Unna et al. 2016), and these may be captured by expression-based filtering. Therefore, the results from both these methods were combined as that very likely capture the diversity in the transcriptome. CD-HIT (Fu et al. 2012) was used to remove duplicates or cluster highly identical transcripts (at least 95% identical) from the combined dataset. This final set of transcripts was used for subsequent analyses.

2.2.4 Quality assessment

The transcripts obtained in the previous step were further examined for quality using TransRate again. All the parameters used were same as before.

The completeness of the assembly was further assessed by the BUSCO pipeline (Simão, et al., 2015). BUSCO reports whether the gene is present, complete, fragmented or duplicated based on HMMER3 amino acid alignment to a list of conserved orthologous genes assumed to be in a clade with the longest predicted ORF in the transcriptome assembly.

Additionally, mega-BLASTN with the previously existing transcriptome assembly of *Littorina saxatilis* (Canbäck et al. 2012) was also performed.

2.2.5 Functional Annotation

Transcriptome assembly was annotated using the Trinotate suite v2.0.1 (Haas et al. 2013). TransDecoder was used for prediction of the coding regions in the transcript (Haas et al. 2013). The output file from TransDecoder, longest-ORF peptide candidates, and the assembled transcriptome, were used for BLASTp and BLASTx, respectively, with the Uniref90 and the SwissProt databases. This task was performed by Diamond (Buchfink et al. 2015). The BLAST results were supplemented with protein domain identification (HMMER/PFAM) (Finn et al. 2008), signal peptide prediction (Petersen et al. 2011), transmembrane domain prediction (tmHMM) (Krogh et al. 2001) and functional annotation with other databases (eggNOG/GO) (Jensen et al. 2008; GO Consortium 2000). All the annotations were loaded in an SQLite database with the default Trinotate parameters. The GO assignments were extracted from the annotation file with the support script provided with the Trinity suite. The transcripts were further analysed by segregating into generic GO Slim categories available at ftp.ebi.ac.uk with a custom bash script.

2.2.6 Tissue specificity index

Tissue specificity of genes was measured using a quantitative, graded scalar method: Tau index (Yanai et al. 2005). Tau (τ) index varies between 0 and 1; where 0 denotes broadly expressed or housekeeping genes, and 1 denotes tissue specific gene expression.

The transcripts that were obtained at the end of the step 2.2.3, were used for this analysis. Raw reads were mapped back to the transcripts using RSEM (Li and Dewey 2011) and a matrix of counts per gene per library was obtained. RSEM was further used to normalize the raw counts as fragments per kilobase of exon per million reads mapped (FPKM) to account for the differences due to different library sizes and gene lengths. This normalized data set was used to calculate the Tau (τ) index. The estimation of τ index was performed on the genes as following (Yanai et al. 2005),

$$\tau = \frac{\sum_{i=1}^{n}(1-\widehat{x_\iota})}{(n-1)} \; ; \; \widehat{x_\iota} = \frac{x_i}{\max\limits_{1\leq i\leq n}(x_i)}$$

where,

x_i is the FPKM value of the gene in tissue library i

n is the number of tissue libraries

Tau scores were calculated for the genes across all the libraries in both, the male and female samples. This perhaps may have a minor impact on genes that are expressed only and only in a single non-reproductive tissue library. For example, genes that are specific to head tissue only, would ideally have a tau score of 1. However, in this manner, the gene would be read in two libraries- male head and female head, giving this gene a score of slightly less than 1, but still high enough to be in the tissue-specific spectrum. Therefore, depending on how tissue specificity cut-off is set, genes with narrow range of expression can still be identified.

The distribution of tau scores was plotted with density() function in R (R core team) using the Gaussian kernel. This function estimates kernel density for each point of the distribution to smooth out noise or fine-scale structures and capture only important patterns. The bandwidth for smoothing was estimated by Silverman algorithm (Sliverman 1986), which defaults to 0.9 times the minimum of the standard deviation and the interquartile range divided by 1.34 times the sample size to the negative one-fifth power. Tau score distribution was further analysed for deviations from unimodality using Hartigans' dip test (Hartigan and Hartigan 1985) implemented in the R package diptest (Maechler 2012). Hartigans' dip test measures smallest value deviations from the empirical distribution. The largest of these deviations is the dip statistic, which is then compared to a table of empirical percentage points of the dip statistic (see Hartigan 1985) to estimate the probability (p values) of the distribution deviating from unimodality. A p-value of less than 0.05 is used to reject the null model (that is, a unimodal distribution).

2.3 Results

2.3.1 Transcriptome assembly

In total, 990,403,964 raw reads were obtained after sequencing. Of these, 863,915,586 reads (87.2%) were retained after filtering for adaptors, duplicated sequences and low-quality reads (table 2.1). The clean reads were normalized and assembled into a transcriptome using Trinity v2.0.6 with default parameters (Haas et al. 2013). Hereby, the set of all contigs in the assembly, including isoforms are referred to as transcripts. While the non-redundant clusters of contigs are referred to as unigenes or genes.

This initial assembly produced 427,922 transcripts and 286,481 genes with an average length of 807 and N50 of 1,399 bp. The first refinement approach based on TransRate and TransDecoder gave 32,801 transcripts and 25,047 genes of average length of 1581 and N50 of 2,260 bp. The second refinement approach based on expression-level cut-off gave 197,049 transcripts and 71,900 genes with an average length of 1299 and N50 of 2062 bp. Both these assemblies share 16,547 genes. The combined assembly from the two refinement steps generated 152,528 transcripts and 77,482 genes with an average length 1226 and N50 of 1903 bp. This reduction in N50 in the final transcript set is expected if the frequency of smaller length transcripts from the other two assemblies increase on combining. The basic assembly statistics for the final assembly are summarized in table 2.1. The length distribution of the transcripts in the assembly showed that most of the transcripts were 300-400 bp long and all transcripts were longer than 199 bp (Figure 2.3).

Table 2.1: Descriptive statistics of the final set of transcripts for the *Littorina saxatilis*

Reads	
No. of raw reads	990,403,964
No. of clean reads	863,915,586
Percentage of clean reads	87.20%
Transcriptome assembly	
Total no of transcripts	152528
No. of non-redundant unigenes	77482
Percent GC	45.2
Contig N50	1903
Average contig length	1225.72
Contig N50 based on the longest isoform per gene	1948
Average contig length based on the longest isoform per gene	1281.74

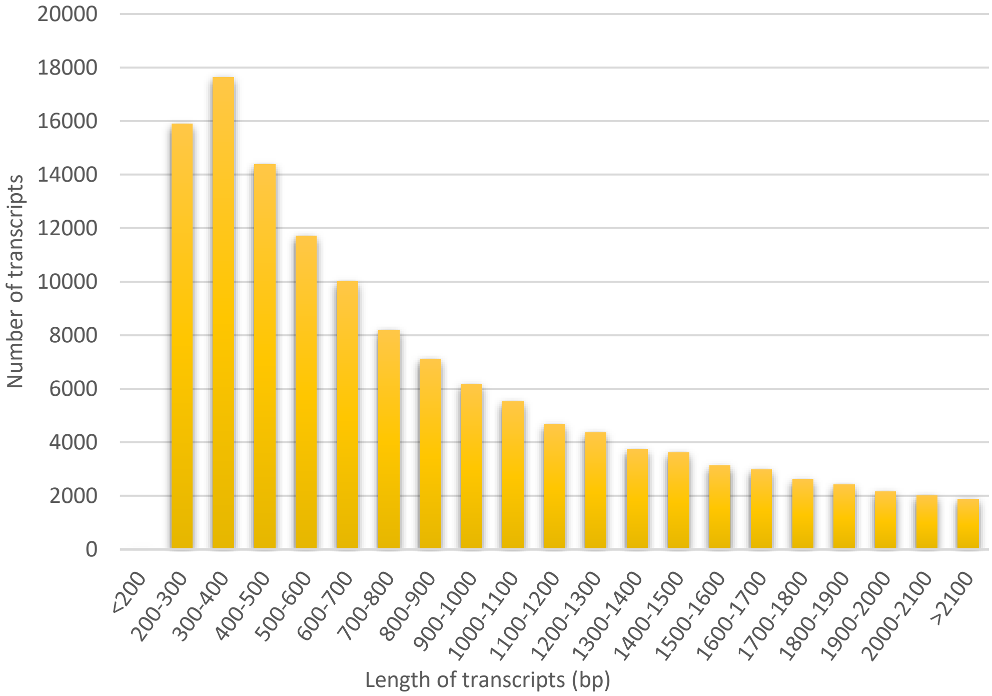


Fig. 2.3: Distribution of the length of the transcripts in the final assembly. On the x-axis is the 100 bp bins representing the length of the transcripts, y-axis represents the number of transcripts.

2.3.2 Quality assessment

The completeness of the assembly was assessed using the completeness of the orthologs by mapping them against the metazoan database in the BUSCO pipeline. The transcripts were mapped against 843 BUSCO ortholog groups. At each stage of the refinement, the transcripts were assessed using BUSCO. The first refinement assembly had only 62% mapping, of which 10% were reported duplicated and 4.9% fragmented. A higher BUSCO score was obtained (94% mapping), with the second refinement assembly, albeit a high percentage of duplication was also reported (21.71%). The combined assembly of the previous two approaches, gave an optimum result of 93% mapping with 15% reported as duplicated and 3.2% as fragmented. The results of the final assembly are summarized in the Figure 2.4.

The assembly score obtained by TransRate, which signifies the overall quality of the assembly post-analysis with all the three modules, is reported to be higher than the previous assemblies [current assembly (0.19); refinement assembly 1 (0.09); refinement assembly 2 (0.12); see Appendix I]. TransRate reported overall 46.96% contigs mapped as ‘good’. While, the CRBB module reported 15.27% of the contigs with a hit in the references and 33.52% of the references with a hit in the transcriptome assembly. The low number of hits can be attributed to the fact that both the molluscan transcriptomes used as the reference are distantly related to our species.

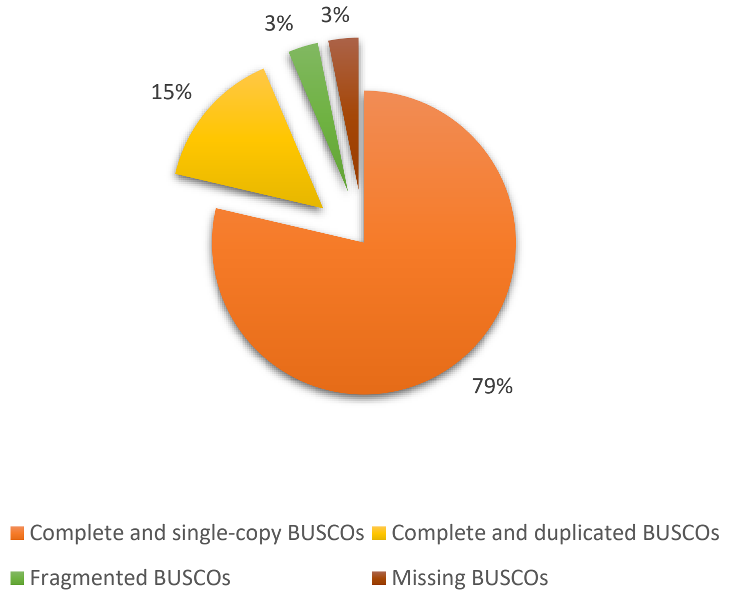


Fig. 2.4: The BUSCO results of the final set of transcripts.

The transcripts were also aligned with LSD *Littorina* assembly (Canbäck et al. 2012). The top hits were analysed. Overall, 37,673 genes (48.62%) had hits which spanned across 57.1% contigs in the LSD assembly. Our result suggests duplication and/or fragmentation in the LSD assembly (which was further confirmed with BUSCO analysis with the LSD assembly; see Appendix II).

2.3.3 Functional annotation

Among the 77,482 unigenes in the combined assembly, 28,990 (37.41%) were detected to have at least one ORF by TransDecoder (Haas et al. 2013). The translated ORFs were used for BLASTp and the rest of the sequences were annotated by BLASTx against Swiss-Prot and Uniref90 databases. There were 13,634 and 13,497 hits in the Swiss-Prot database; and 20,076 and 18,822 hits in the Uniref90 with BLASTX and BLASTP, respectively. Further, 15,984 gene, (20.62%) were also annotated by Pfam. Signalp and tmHMM detected 3,655 (4.71%) and 6,433 (8.3%) of the genes to be signalling peptides or transmembrane proteins, respectively. Overall, 31,024 genes (40%) showed functional annotation while the remaining 46,458 genes (59.95%) did not get annotated or detected with an ORF. The results are summarized in table 2.2.

The BLAST results with Uniref90 were given precedence over Swiss-prot due to lack of mollusc representation in the latter database for further analysis. Of the 20,076 BLAST hits, 56.41% (n=11,325) were represented by molluscs, followed by Chordata (n=2,666). Among molluscs, *Lottia gigantea* (60.72%) and *Crassostrea gigas* (37%) accounted for the most annotations.

The GO annotations were categorized according to the generic GO Slim categories. GO Slim is an annotation count which broadly suggests if a gene can be grouped with those with known functionality. A gene product can have multiple GO assignments attached with it. This basically means that a GO assignment attached with multiple gene products will have higher representation in the data. The GO Slim categories can be further divided into ‘cellular components’ (the location of the gene products at the subcellular level), ‘molecular function’ (the function of the gene product) and ‘biological process’ (the series of molecular functions which cause a phenotype). The top representations for each of these three Go Slim categories are presented in Figure 2.5. The comprehensive list for the Go Slim categories are listed in Appendix III. The highest representation among the molecular function ontology is the ‘binding’ function. The description essentially stands for the interaction of a molecule with other molecule on one or more specific sites. The use of the word molecule here is vague as this can be defined as ligand ‘in the broadest biological sense’. Similarly, ‘membrane’ and ‘transport’ are the highest represented annotations in the cellular component and biological process ontologies respectively. To be noted here, TmHMM predicted relatively fewer membrane proteins than appeared in the GO Slim. Perhaps, this discrepancy is because TmHMM aims to predict only the transmembrane proteins, while GO Slim category reported all the integral membrane proteins.

Table 2.2: Annotation of the final set of transcripts.

Total genes	77482
No. of genes annotated	31024
No. of genes 'not' annotated or without prediced ORF	46458
Swiss-Prot BLASTX	13633
TrEMBL BLASTX	20076
Genes with at least one predicted ORF	28990
Swiss-Prot BLASTP	13497
TrEMBL BLASTP	18822
Pfam	15984
SignalP	3655
TmHMM	6433
eggnog	6543
GO BLAST	13459
GO Pfam	10878

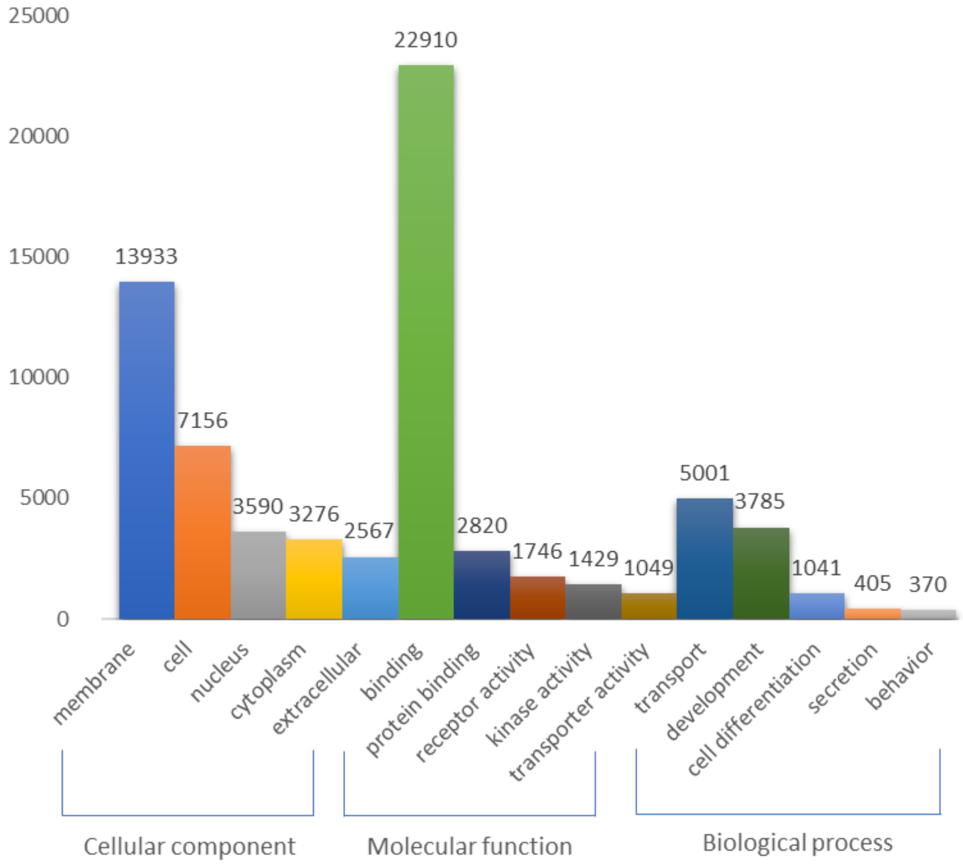


Fig. 2.5: Gene Ontology (GO) classification. The 31,024 annotated genes of *Littorina saxatilis* split by GO Slim categories. The top representations for each GO Slim category are presented in this image.

2.3.4 Tissue specific expression

The distribution of tau scores, across all tissue libraries in both male and female, is presented in Figure 2.6. Yanai et al. (2005) reported a bimodal distribution of tau scores in human tissues with modes at 0.15 and 0.85. This bimodal distribution of tau scores was also replicated by Kryuchkova-Mostacci and Robinson-Rechavi (2016) using human tissues and mouse tissues separately. Based on Figure 2.6, it is difficult to comment whether the distribution is bimodal, although Hartigans' dip test indicates departure from a unimodal distribution (D=0.004; $p < 0.0001$). However, it does not necessarily indicate bimodality, and may also indicate multimodal data.

We do not specify tau score cut-off to distinguish between housekeeping, midrange and tissue-specific gene expression and strongly recommend that such cut-off should be based on downstream analysis. If required, the measure Yanai et al. (2005) recommended may be used as a rule of thumb. Yanai et al. (2005) suggested to set a cut-off for tissue-specific and housekeeping genes based on the bimodal peaks. That is, in our data, for example, the peak at 0.9 may suggest a cut-off of 0.85 tau score or above (at the dip; Figure 2.6b) to distinguish between tissue-specific and midrange expression genes. However, a broad distribution of scores is observed around 0.6 and another peak at 0.75, which makes it difficult to set a cut-off to distinguish between housekeeping genes and midrange expression genes based on this measure.

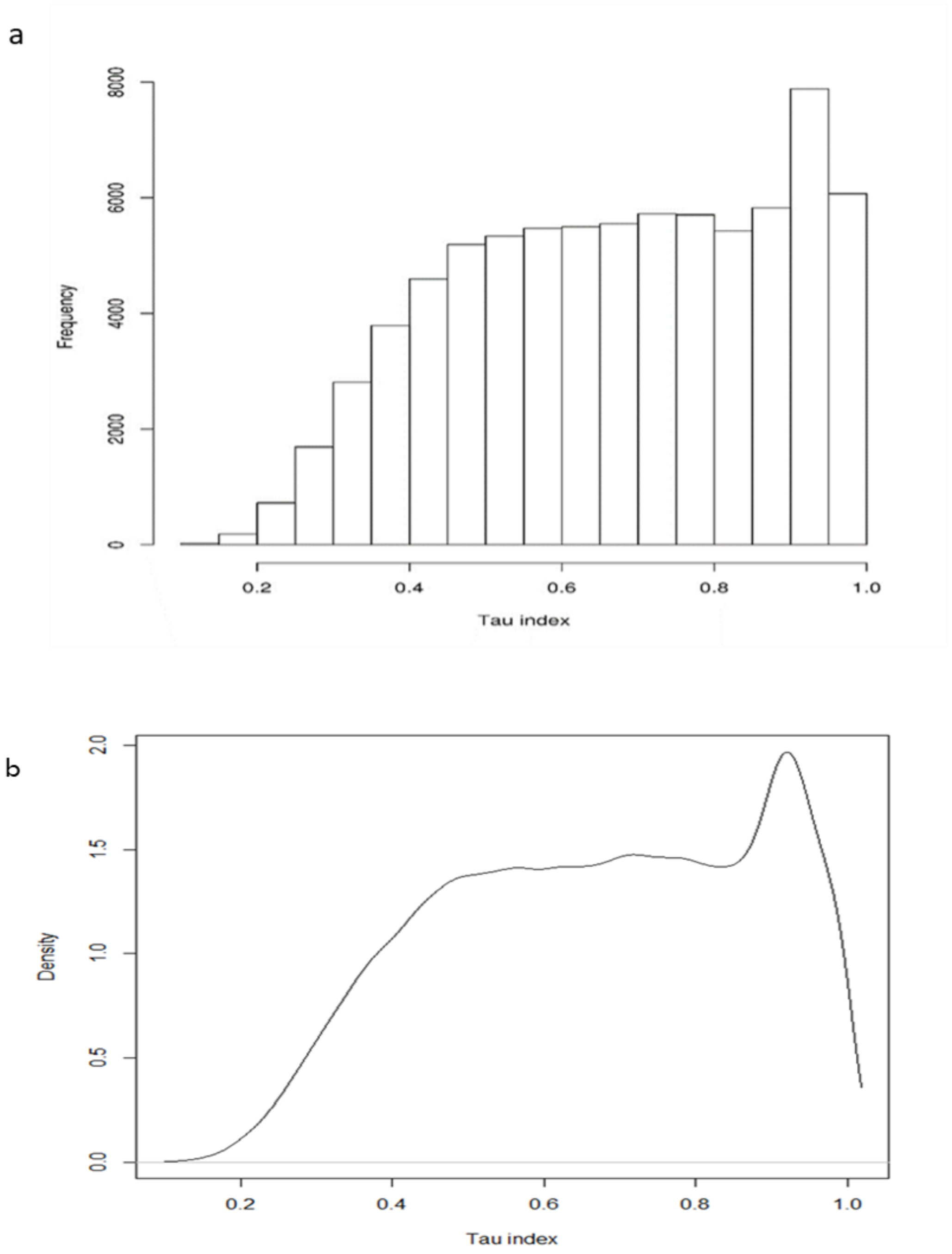


Fig. 2.6: Tau score distribution. **(a)** Histogram of tau score distribution. On X-axis, tau scores 0-1 are plotted and on y-axis number of genes are plotted. **(b)** Density function plot of tau scores. N=77482 Bandwidth=0.02

Discussion

This chapter discusses the construction of a *de novo* transcriptome assembly for *Littorina saxatilis.* The initial assembly reported 427,922 contigs and 286,481 unigenes. The high number of genes can be attributed to the high number of reads as input as well as the algorithm of Trinity which use de Bruijn graph algorithm to construct graphs from k-mers (Hass et al

2013). While this algorithm maximizes probability that the isoforms can be captured from a given data set, rare k-mers may lead to spurious contigs being reported. In theory, rare k-mers are the result of lowly expressed transcripts, however, the NGS data is error-prone and sequencing errors may also generate rare k-mers. In order to avoid the error-prone transcripts from the *de novo* assembly, it was necessary to perform filtering on the initial transcriptome assembly. Filtering was done firstly by TransRate (Smith-Unna et al. 2016), which reported the well-assembled contigs among which transcripts with ORF were selected as the high-confidence transcripts. Secondly, the raw counts of reads mapping back to the transcriptome were obtained and normalized and used for expression-based filtering. The combination of the transcripts obtained by both these methods, after the removal of identical and duplicated transcripts, ensured that the final set of transcripts would be relatively free of bias compared to relying solely on any one approach. However, the quality filtering steps of the contigs only ensured that the probability of false transcripts from the transcriptome was reduced. The merging of the two refined assemblies did perhaps increase the high-quality unigenes in the assembly, but it does not ensure that the entire diversity of the RNA from our samples has been captured. A very popular approach to construct more complete transcriptome assemblies is the combination of genome-guided reference and *de novo* assemblies. The reference-genome assembly approach may be able to capture more genes in the genome due to its high sensitivity and the *de novo* assembly approach may be able capture the novel variants and assemble trans-splice sites (Martin and Wang 2011). This combined assembly approach can be useful for a rather more comprehensive transcriptome assembly and can be employed in the future with the publication of the *Littorina saxatilis* reference genome.

In Canbäck et al. (2012), 26,537 contigs were reported for *Littorina saxatilis*. They further estimated, based on the genome size and gene model comparison with *Lottia gigantea* (and assuming similar gene density), that *Littorina saxatilis* should have 62,000 genes. In comparison, this study identifies 77,482 genes in the transcriptome, which is an overestimation, and majority of them remain unannotated. This phenomenon has been observed before and discussed in detail in Harney et al. (2016). The authors compiled the transcriptome studies between 2011 and 2016 of 17 mollusc species, constructed using different assemblers. While the number of total genes ranged from 22,761 to 151,684 for different mollusc species, only 12-38% annotation was reported (Harney et al. 2016). The only exception was *Crassostrea gigas* in which 80% annotation was reported for 55,651 genes (Riviere et al. 2015). It has been suggested that this discrepancy between the number of genes in the *de novo* transcriptome assemblies and the annotation with the known gene models may be because of the novel contigs which may be derived from the non-coding RNA (Eddy 2001) and perhaps of utility to the RNA studies as they may play an important role in the regulation of gene expression (Guttman and Rinn 2012).

Another reason that has been proposed is the apparent lack of representation of molluscs in the public databases, as noted by Kang et al. (2017). As also observed in this study, only half of the annotated genes found homology with the molluscan gene models. Kang et al. (2017) stated that the functional annotation of the *de novo* assembly of the land snail *Satsuma myomphala* with four public databases resulted only in 3.5-31.2% annotation, while they were able to achieve 92% annotation by using the PANM database. PANM database was introduced in 2004 but updated in 2015 and essentially consists of protostome group sequence information (Mollusc, Nematoda, Arthropoda) (Kang et al. 2016).

The annotated genes were further categorized into GO Slim categories. GO Slim is an excellent method to have an overview of the annotations of the assembly, although caution must be taken while interpreting the GO ontology results. The GO ontology database is largely based on validation of genes through bioinformatics processes, which makes the direct interpretation of GO ontologies error-prone. Strictly speaking, GO classifications may not be taken as evidence of functionality. Additionally, the success and useful interpretation of GO analyses depend on the representation of the organism of interest in the GO database. In the case of under-represented group such as molluscs, this raises again the question of usefulness of such analysis.

The use of multiple tissues in the assembly construction gives an opportunity to catalogue gene expression breadth. Therefore, this data finds use in the ongoing work to document tissue-specific genes in this species. Tissue-specific gene expression represents the tissue adjusted functionality of genes and gives a keen insight into their biological roles and underlying pathways. Tissue-specificity was measured via the tau method. A study by Kryuchkova-Mostacci and Robinson-Rechavi (2016) has shown that tau is the most robust method among existing tissue specificity measurements and tends to capture more variation in expression profiles. Unlike more stringent methods, such as raw counts as proxy for expression which captures only genes expressing in a single tissue as tissue-specific, tau indicates how many tissues a gene is expressed in and if the expression difference between them is large. Therefore, we refrain from setting any cut-offs to distinguish between tissue-specific, housekeeping and midrange expression genes. We suggest that cut-offs should be based on the downstream analysis and the tissue under study. For example, to probe reproduction genes, a stringent tau score cut-off is required to define tissue-specificity, as such genes are expected to have a narrow range of expression. However, to probe neural genes, especially in central nervous system (or nerves in pleural ganglia in gastropods), a less stringent cut-off is required. In the literature, a bimodal distribution of tau values has been reported. We observed a pattern that departs from a unimodal distribution, however we did not observe distinct bimodal peaks in our distribution. This pattern may indicate that minimally-expressed ubiquitous genes may have been removed during the construction of refinement assembly 2 (as observed by the broad distribution on left-hand side of the graph in Figure 2.6b) and suggests a need for less stringent expression-based filtering in the future when the transcriptome assembly may be revised. Alternatively, this pattern could be biological and suggest lesser variance among gene expression patterns between housekeeping and midrange expression genes in this species. The current literature that utilizes the tau index analysis focuses mainly on vertebrates (model organisms) and therefore, the latter hypothesis may be hard to verify in the absence of comparable data. Perhaps, replication of this analysis may provide more confidence over the pattern we observed. In addition, the results of the current tau analysis also require validation. The validation of tau analysis may be achieved by supplementing this analysis with GO enrichment analysis. Certain GO terms are expected to be tissue-specific and therefore, expected to be concentrated on the right-hand side of the tau score distribution graph. Similarly, certain GO terms have ubiquitous gene expression range and therefore, are expected to be concentrated on the left-hand side of the tau score distribution graph. For example, spermatogenesis (GO:0007283) should be tissue-specific to testis; while, protein folding (GO:0006457) or RNA splicing (GO:0008380) are expected to be housekeeping. GO enrichment analysis may also help in better interpretation of tau results, for example to analyse which family of proteins tend to be specific to a particular tissue. However, unarguably, in the future, such analysis would need experimental validation, perhaps by performing RT-PCR across multiple tissues.

The genomic resources such as transcriptome should always be considered as work-in-progress. The advancement in technology, inclusion of new data and updating of the public databases imply that the genomic resources need to be updated too. As already suggested, combining different assembly approaches may identify more genes in our organism. But equally important is the characterization of those genes. The data in the public databases is added every week. The annotation algorithms are improving in order to identify the yet-unexplored non-coding RNA (Musacchia et al. 2015). Therefore, it is only reasonable to revise the assembly in the future to build a more comprehensive transcriptomic resource. For now, it may be safe to assume that the present transcriptome assembly that I presented in this chapter is a high-quality comprehensive resource available for *Littorina saxatilis* currently. The high-confidence transcripts, among other datasets, are already being used for *ab initio* gene model predictions in the genome assembly (Panova, Larsson et al. in prep.). Additionally, this transcriptomic data was used to characterize the genetic loci in association with the adaptive shell characteristics (see Chapter 4). It is further hoped that this may prove to be an invaluable resource for very many varied purposes, for example, to identify orthologous genes across taxa, or gene expression studies. Perhaps, the ongoing tau analysis, when completed and validated, may prove to be of immense use to researchers working on different biological aspects in gastropods. For example, mantle tissue-specific genes could be studied to gather more clues regarding shell colouration. Moreover, the inclusion of libraries from the two sexes in tau analysis has enabled to further extend this analysis to document sex-specific expression which may perhaps give more insight into sex-determination in this system.

Conclusion

This chapter describes the construction of the most comprehensive transcriptome assembly of *Littorina saxatilis* to date. The raw reads were generated from multiple tissues of two individuals of the crab-ecotype. A de novo transcriptome assembly comprising filtered reads from both the individuals and all the tissue libraries was constructed and the high-quality transcripts based on two different approaches were selected to represent the transcriptome. The transcripts were annotated against public nucleotide and protein databases for predicted functional classifications. The identified and annotated transcripts will provide a valuable genomic resource to the *Littorina* community for future research. Additionally, suggestions have been made to make further improvements to the assembly and annotations.

The transcriptome assembly and the annotation database will be made publicly available soon. Currently, it is available on request. The tau scores for genes are also available on request.

Contributions

I tested the various assembly construction parameters and performed the final transcriptome assembly as described above. I also performed the assembly annotation and GO Slim analysis. I am currently working on the tissue specificity analysis. This work, however, is the product of collaboration and I would like to acknowledge and thank the following people for their contribution(s).

Dr. M. Panova (University of Gothenburg) performed the sampling, dissection and RNA extraction. BGI (Shenzhen, China) did the library preparation and sequencing. Dr. Tomas Larsson provided technical expertise and advice.

References

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Aubry, S., Kelly, S., Kümpers, B.M.C., Smith-Unna, R.D., Hibberd, J.M. (2014). Deep Evolutionary Comparison of Gene Expression Identifies Parallel Recruitment of Trans-Factors in Two Independent Origins of C4 Photosynthesis. PLOS Genetics; 12(5): e1006087.

Bryant, D. M., Johnson, K., DiTommaso, T., Tickle, T., Couger, M. B., Payzin-Dogru, D., Lee, T. J., Leigh, N. D., Kuo, T. H., Davis, F. G., Bateman, J., Bryant, S., Guzikowski, A. R., Tsai, S. L., Coyne, S., Ye, W. W., Freeman, R. M., Peshkin, L., Tabin, C. J., Regev, A., Haas, B. J., Whited, J.L. (2017). A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors. Cell reports; 18(3):762-776.

Buchfink, B., Xie, C., Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nature Methods; 12:59–60.

Canbäck, B., André, C., Galindo, J., Johannesson, K., Johansson, T., Panova, M., Tunlid, A., Butlin, R.K. (2012). The Littorina sequence database (LSD)—an online resource for genomic data. Molecular Ecology Research; 12:142–148.

Cantarel, B. L., Korf, I., Robb, S. M. C., Parra, G., Ross, E., Moore, B., … Yandell, M. (2008). MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Research; 18(1):188–196.

Eddy, S.R. (2001). Non-coding RNA genes and the modern RNA world. Nature Review Genetics; 2:919-929.

Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H.-R., … Bateman, A. (2008). The Pfam protein families database. Nucleic Acids Research; 36(Database issue): D281–D288.

Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W. (2012). CD-HIT: accelerated for clustering the next generation sequencing data. Bioinformatics; 28 (23): 3150-3152.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., … Regev, A. (2011). Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nature Biotechnology; 29(7):644–652.

Graveley, B.R. (2001). Alternative splicing: increasing diversity in the proteomic world. Trends in Genetics; 17:100–107.

Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson J., Adiconis, X., … Regev, A. (2010). *Ab initio* reconstruction of transcriptomes of pluripotent and lineage committed cells reveals gene structures of thousands of lincRNAs. Nature Biotechnology; 28(5):503–510.

Guttman, M., Rinn, J. L. (2012). Modular regulatory principles of large non-coding RNAs. Nature; 482:339-346.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., MacManes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., Henschel, R., LeDuc, R. D., Friedman, N., … Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature protocols; 8(8):1494-512.

Harney, E.D., Dubief, B., Boudry, P., Basuyaux, O., Schilhabel, M.B., Huchette, S., Paillard, C., Nunes, F.L.D. (2016). De novo assembly and annotation of the European abalone *Haliotis tuberculata* transcriptome. Marine Genomics; 28:11-16.

Hartigan, J. A., Hartigan, P. M. (1985). The dip test of unimodality. Annals of Statistics; 13; 70–84.

Hartigan, P.M. (1985). Algorithm as 217: computation of the dip statistic to test for unimodality. Applied Statistics; 34(3): 320-325.

Jensen, L. J., Julien, P., Kuhn, M., von Mering, C., Muller, J., Doerks, T., Bork, P. (2008). eggNOG: automated construction and annotation of orthologous groups of genes. Nucleic Acids Research; 36(Database issue): D250–D254.

Kang, S.W., Park, S., Patnaik, B.B., Ju Hwang, H., Chung, J.M., Song, D.K., Young-Su, P., Lee, J.S., Han, Y.S., Park, H., Lee, Y.S. (2016). The Protostome database (PANM-DB): Version 2.0 release with updated sequences. Korean journal of malacology; 32: 185-188.

Kang, S.W., Patnaik, B.B., Hwang, H.J., Park, S.Y., Chung, J.M., Song, D.K., … Lee, S. Y. (2017). Sequencing and de novo assembly of visceral mass transcriptome of the critically endangered land snail *Satsuma myomphala*: annotation and SSR discovery. Comparative Biochemistry and Physiology Part D: Genomics and Proteomics; 21: 77–89.

Kanost, M. R., Arrese, E. L., Cao, X., Chen, Y.-R., Chellapilla, S., Goldsmith, M. R., … Blissard, G. W. (2016). Multifaceted biological insights from a draft genome sequence of the tobacco hornworm moth, *Manduca sexta*. Insect Biochemistry and Molecular Biology; 76: 118–147.

Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. Journal of Molecular Biology; 305(3):567-80.

Kryuchkova-Mostacci, N., Robinson-Rechavi, M. (2016). A benchmark of gene expression tissue-specificity metrics. Briefings in Bioinformatics; 1–10.

Li, B., Fillmore, N., Bai, Y., Collins, M., Thomson, J. A., Stewart, R., Dewey, C. N. (2014). Evaluation of de novo transcriptome assemblies from RNA-Seq data. Genome Biology; 15(12): 553.

Li, B., Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics; 12:323.

Maechler, M. (2012). diptest: Hartigan's dip test statistic for unimodality – corrected code. R package version 0.75-74. Available online at: http://CRAN.R-project.org/package=diptest.

Martin, J.A., Wang, Z. (2011). Next-generation transcriptome assembly. Nature Reviews Genetics; 12:671–682.

Musacchia, F., Basu, S., Petrosino, G., Salvemini, M., Sanges, R. (2015). Annocript: a flexible pipeline for the annotation of transcriptomes able to identify putative long noncoding RNAs. Bioinformatics; 31(13):2199-201.

Panova, M., Larsson, T., Alm Rosenblad, M., Chaube, P., Westram, A.M., Butlin, R.K., Blomberg, A., Johannesson, K. (in prep.) Insights into local adaptation from the genome of *Littorina saxatilis* (Mollusca: Gastropoda).

Parra, G., Bradnam, K., Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics; 23(9):1061-1067.

Petersen, T.N., Brunak, S., von Heijne, G., Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. Nature Methods; 8(10):785-6.

R Core Team. (2015). R: A language and environment for statistical computing [Internet]. Vienna, Austria; R Foundation for Statistical Computing, Vienna.

Rana, S.B., Zadlock, F.J., Zhang, Z., Murphy, W.R., Bentivegna, C.S. (2016). Comparison of De Novo Transcriptome Assemblers and k-mer Strategies Using the Killifish, *Fundulus heteroclitus*. PLOS ONE; 11(4):e0153104.

Reid, D.G., Dyal, P., Williams, S.T. (2012). A global molecular phylogeny of 147 periwinkle species (Gastropoda, Littorininae). Zoologica Scripta; 41(2):125-136.

Riviere, G., Klopp, C., Ibouniyamine, N., Huvet, A., Boudry, P., Favrel, P. (2015). GigaTON: an extensive publicly searchable database providing a new reference transcriptome in the pacific oyster *Crassostrea gigas*. BMC Bioinformatics; 16:401.

Schulz, M. H., Zerbino, D. R., Vingron, M., & Birney, E. (2012). Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics; 28(8):1086-92.

Sharma, E., Künstner, A., Fraser, B. A., Zipprich, G., Kottler, V. A., Henz, S. R., … Dreyer, C. (2014). Transcriptome assemblies for studying sex-biased gene expression in the guppy, *Poecilia reticulata*. BMC Genomics; 15(1):400.

Silverman, B. W. (1986). Density Estimation. London: Chapman and Hall.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. & Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics; 31:3210–3212.

Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J. M., Kelly, S. (2016). TransRate: reference-free quality assessment of de novo transcriptome assemblies. Genome Research; 26(8):1134–1144.

The Gene Ontology Consortium, Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., … Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. Nature Genetics; 25(1): 25–29.

Trapnell, C., Pachter, L., Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics; 25(9):1105–1111.

Trapnell, C., Robert, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., … Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature Protocols; 7(3):562–578.

Wu, T.D., Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics; 26(7): 873–881.

Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., Huang, W., He, G., Gu, S., Li, S., Zhou, X., Lam, T.W., Li, Y., Xu, X., Wong, G.K-S, Wang, J. (2014). SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics; 30:1660–1666.

Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., … Shmueli, O. (2005). Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics; 21:650–9.

Zapata, F., Wilson, N. G., Howison, M., Andrade, S. C. S., Jörger, K. M., Schrödl, M., … Dunn, C. W. (2014). Phylogenomic analyses of deep gastropod relationships reject Orthogastropoda. Proceedings of the Royal Society B: Biological Sciences; 281(1794): 20141739.

Chapter 3: Linkage maps

3.1 Introduction

About a hundred years ago, Thomas Hunt Morgan hypothesized that the genes are physical objects whose crossing-over or recombination frequency depended on their physical distance on chromosomes (Morgan 1911). He proposed that the linkage is strongest between the genes that are positioned closely on a chromosome (Morgan 1911). His undergraduate student, Alfred Henry Sturtevant, realised that this information can be exploited to map the relative positions of the genes linearly on a chromosome. And within one evening, he managed to plot the first linkage map in *Drosophila* with six genes on the X-chromosome (Sturtevant 1913). Ever since, linkage maps for many organisms have been plotted.

However, traditionally linkage maps were only limited to organisms that can be crossed experimentally and a few markers were genotyped. The advent of next-generation sequencing (NGS), which permitted genotyping of thousands of markers for multiple individuals in a cost-effective manner, and advancements in mapping algorithms (Margarido et al. 2007; Rastas et al. 2013), have allowed construction of high-density linkage maps for almost any species. Linkage maps provide vital information about recombination and linkage in an organism and these are key features to understand the structural, functional and evolutionary characteristics of a genome (Rastas et al. 2016). The high-density linkage maps have been used to refine *de novo* genome assemblies and order scaffolds and contigs (Rastas et al. 2013; Fierst 2015). They have been used in the detection and validation of structural rearrangements, such as inversions (Bansal et al. 2007). Local recombination rates have been estimated using the high-density linkage maps (Kawakami et al. 2014). High-density linkage maps have been employed to map quantitative trait loci and understand the genetic basis of adaptive phenotypes (Papa et al. 2013). The use of high-density linkage maps in population and comparative genomics has made it possible to compare synteny across species, which has enhanced our understanding of genome evolution and speciation (Rastas et al. 2016).

Linkage map construction, however, is a computationally expensive problem. Especially, with large datasets, this becomes a more serious constraint. The two basic steps in the construction of linkage maps are: (i) clustering of the molecular markers into different linkage groups, or chromosomes, and (ii) estimating the linear order of the markers in the linkage group. The problem scales exponentially with the increasing number of markers (Fierst 2015). For m markers, the possible number of orders within a linkage group can be estimated as ½ m! (Fierst 2015). This is major limitation in the construction of linkage maps with large datasets, which demanded the development of algorithms that can handle such large data and produce results within a reasonable time-frame (Rastas et al. 2013; Schiffthaler et al. 2017). Infrequent recombination between adjacent markers, missing data and genotype errors add to the complexity and tend to have large effects on map estimation. These limitations render linkage maps accurate at large scales but inaccurate at fine-scales (Fierst 2015). To address these issues, algorithms with increased accuracy for recombination estimation, grouping and mapping are still under development (Margarido et al. 2007; Cheema and Dicks 2009; Mollinari et al. 2009; van Ooijen 2012; Rastas et al. 2013; Rastas et al. 2016). To tackle missing data and genotype errors, there are packages that impute missing values and correct for genotype errors while ordering markers (van OS et al. 2005; Cartwright et al. 2007; Wu et al. 2008; Rastas et al. 2013; Rastas et al. 2016).

Apart from the inherently difficult task of constructing linkage maps, another equally diffciult task is the integration of independent linkage maps to produce a consensus map. The linkage maps from different populations may have different characterstics, such as different recombination frequencies, ordering conflicts due to genotyping or statistical errors and different marker informativeness (Beavis and Grant 1991; Endelman and Plomion 2014). These conflicts in composite maps tend to manifest in the same way as genotyping errors or missing data, usually causing inflation in map estimations. Additionally, a key factor in the integration of different maps is the number of common markers across all the independent maps. The integration of independently obtained maps into a composite map while minimizing the previously stated errors, requires a certain minimum number of common markers, in the absence of which integration is not possible (Stam 1993).

Two distinct ecotypes of the marine snail *Littorina saxatilis* are known to diverge locally and under gene flow, in parallel across many localities (Johannesson et al. 2010). Thus, this system is ideal to study local adadptation and speciation in the face of gene flow. Several studies have aimed to characterize the genomic regions under divergence (Galindo et al. 2010; Westram et al. 2014; Ravinet et al. 2016; Westram et al. 2016). However, a plethora of questions remain unanswered, such as the number of diverging loci, their distribution across the genome, phenotypes underlying adaptive divergence and their genetic architecture. The answers to these questions are imperative to understand a complete picture of speciation in this system. For example, if the regions under divergent selection are many and clustered, or few and isolated, may help gauge how much reproductive isolation has been acheived. Whether the islands of diverging loci are maintained in a genetic architecture that promotes reduced recombination and facilitates speciation, or prone to the homogenizing effect of the gene glow. However, the lack of a genomic toolbox for this organism poses a major hindrance to asking such questions. To address this gap, we have constructed high-density linkage maps for *Littorina saxatilis.* In this chapter, I have described the construction of sex-specific and sex-averaged maps. Since the linkage maps were constructed based on a single family with only two sexed individuals (the parents), hereby, sex-specific and sex-averaged maps will be reffered to as parent-specific and parent-averaged maps respectively. We were able to place 18,942 SNPs, representing ~600 Mb of the genome. To construct the linkage maps, we used the Lep-Map2 package to be able to handle a large amount of data and correct for missing data and genotype errors. However, we still faced problems arising from these issues and merging of parent-specific maps. We developed an approach to correct the linkage maps by using the common markers to address the merging issues while using the genome-assembly information to refine fine-scale mapping inaccuracies. We also used the Chromonomer package (Amores et al. 2014), which integrates the genome assembly and the linkage maps while informing the genome assembly with the linkage map to orient or split scaffolds and the linkage maps with genome assembly for fine-scale correction. We compared our algorithm to correct the linkage-map at fine-scale resolution using the genome assembly with the Chromonomer results. We also explored the parent-averaged and female-parent and male-parent specific recombination patterns through Marey maps. The latter provided an insight into the sex-specific recombination differences of *Littorina saxatilis*.

3.2 Methods

3.2.1 DNA extraction and sequencing

To generate this linkage map, a male and a virgin female of the Crab ecotype were crossed in the lab by K. Johannesson (University of Gothenburg). To maximise the number of offspring to be used in the analysis, they were kept alive only until they were few millimetres long in seawater aquaria and then immediately used for DNA extraction prior to any processing. This precluded phenotypic analysis of offspring (including sexing). Foot tissue samples from the 183 offspring (unsexed due to immaturity) and the parents were taken and extracted using the modified CTAB protocol (Panova et al. 2016). Targeted capture-sequencing technology was used to obtain genomic regions a few hundred base pairs long with multiple SNPs within these regions. A total of 40,000 probes, 120 bp long, were used. Out of these, 1316 probes were used in Westram et al. (2016) and 38,684 probes were newly designed (based on the previous *L. saxatilis* genome assembly) to be spread across the genome randomly and to represent only single-copy genomic regions (Westram et al. 2018). The probe design and sequencing were done by RapidGenomics (Gainesville FL, United States).

Bioinformatics analysis and variant calling were done by A. M. Westram, as described in Westram et al. (2018). The fastq reads were trimmed with Trimmomatic v. 0.36 (Bolger et al. 2014) and mapped to the current *L. saxatilis* genome using bwa-mem. Reads mapping to non-targeted regions, secondary hits or with mapping quality less than 20, or potential PCR duplicates were removed prior to SNP calling (Westram et al. 2018). The SNP calling was done using samtools v 1.3.1 (Li et al. 2009) and only the biallelic SNPs with genotype quality > = 25, variant quality > = 20, depth > = 15 and maf > = 0.05 were retained (Westram et al. 2018).

3.2.2 Linkage map construction

From the previous steps, 74,424 SNPs were obtained, of which, 1,112 had missing genotype for either one or both parents. The SNPs where both the parents were missing, were discarded. The imputation of the parent genotype, where only one parent was missing, was done by looking at the segregation pattern in the offspring. For example, if the offspring had genotypes aa, ab and bb, the missing parent was imputed as ab (heterozygous for the given SNP). If the offspring had genotypes aa and ab, and the existing parent was aa, then the genotype of the missing parent was imputed as ab, and vice-versa. With this method, parent genotypes at 715 markers were imputed. A total of 74,027 SNPs was finally obtained and used as input in the Lep-Map2 package (Rastas et al. 2016) to construct linkage maps.

Within the Lep-Map2 pipeline, the markers were further filtered for segregation distortion and missing data. The option ‘dataTolerance’ was set to a p-value of 0.01 to remove markers that segregated far from Mendelian expectation. The markers that had >50% missing data were removed by setting ‘missingLimit’ at 92. The remaining 61,161 markers were grouped into different linkage groups with the SeparateChromosomes module of the Lep-Map2 program. LOD scores between 10 (default) and 25 were tested, and the scores between 12 and 22 gave stable results in terms of number of linkage groups. In addition, the LOD score of 12 allowed the greatest number of markers to be used in the linkage maps. Therefore, the LOD score was set at 12 with the minimum size requirement for a linkage group to be 20. The remaining single markers, which were not clustered into any linkage group, were further coerced, wherever possible, with the JoinSingles module with the LOD limit set at 3, giving 54,996 markers in 17 linkage groups.

The markers were ordered within the linkage groups with the OrderMarkers module of Lep-Map2. The initial test runs done with default parameters produced maps that were in general long. Therefore, as suggested in the Lep-Map2 manual, a lower recombination rate parameter of 0.005 (https://sourceforge.net/p/lepmap2/wiki/Modules/) was used to construct the linkage maps. The Kosambi function was used to estimate genetic distances from the recombination frequencies as this mapping function accounts for interference, which may be encountered more frequently in the dense maps where there is short interval between adjacent markers (Tan and Fornage 2008). Rastas et al. (2016) recommended repeating the ordering step multiple times and selecting the linkage group orders with the highest maximum likelihood. Therefore, the ordering step was done multiple times for each linkage group in a pyramid-like scheme, i.e., 10 separate runs were done for each linkage group and the order with the maximum likelihood was used as starting order for another 10 separate runs for that linkage group. This method was repeated until stable orders and likelihoods for each linkage group were obtained. At the end of each ordering step, the maps were manually curated to remove singular markers at the end of linkage groups that were more than 5 cM away from their nearest neighbour or groups of 3 markers that were 10 cM or more away. The markers with genotype error estimates ≥0.1 (as estimated by the Lep-Map2 program) were also removed. The resulting map was designated freeze map 0.0.

The order of the markers, as determined by the freeze map 0.0, was compared with the physical order (genome assembly) to check for the inconsistencies between the maps and assembly. Such inconsistencies can potentially be informative about the mapping uncertainty and assembly errors (Fierst 2015). This task was performed by plotting Marey maps, that is, genetic distance against physical distance (Chakravarti 1991) for a quick visual comparison. Collinear maps indicate agreement between the genetic and physical ordering, while non-collinear maps are indicative of discrepancies. The following discrepancies between the genetic and physical order were recognized – (i) non-linear order of the markers in the genetic map (for example, markers placed at 10 bp, 20 bp and 30 bp in the assembly are ordered 10 cM, 30 cM and 20 cM in the maps), (iii) markers on the same contig mapping to multiple linkage groups, or, (iii) physically close markers (markers on a same contig) cluster at various genetic map positions. All the contigs that showed any discrepancy were recorded as 'error' (or 'problem') contigs. The threshold for maximum genetic distance plausible between the physically close markers was set, beyond which the error contigs were flagged to be removed or taken action on [i.e., particular SNP(s) underlying the discrepancy were reported to be removed] (discussed below). The physical distances for the Marey maps were calculated based on the cumulative length of the contigs in the region. Within a map position, the random order and orientation of contigs were used as we did not have information to resolve these issues. No gaps between the adjacent contigs were assumed.

3.2.3 Linkage map correction

The freeze map 0.0 revealed discrepancies with the genome assembly. These discrepancies could have arisen from merging of the two parental maps, assembly errors or mapping errors. To address the inconsistencies between the genome assembly and genetic maps, we developed a two-step approach (see Figure 3.1). The first step involved generation of reduced parent-specific datasets. This was performed by retaining only those contigs that have biparentally informative markers, or biparentally informative markers and parentally-informative markers for either of the parent. The parent-specific datasets were used to generate parent-specific linkage maps and the biparentally informative markers were used to merge the two maps to address the merging issues. The second step involved using the local genome assembly information to correct the inconsistent map positions. Details of these steps follow.

Firstly, the contigs with markers that mapped to more than one linkage groups were removed as they were attributed to mapping or assembly errors. From this dataset, with markers filtered at multiple steps for map 0.0 and contigs removed in the previous step, the parent-specific reduced datasets were generated, using a custom R- script (by Prof. Roger K. Butlin, University of Sheffield):

(i) biparentally informative markers were kept;
(ii) synthetic biparental markers were made on contigs where only single-parent informative markers were present;
(iii) the single-parent informative marker with the least missing values was kept on each contig that lacked biparental or synthetic biparental markers, while the rest of the single-parent informative markers were removed;
(iv) two reduced datasets were produced in this way: male-informative plus biparentally-informative markers and female-informative plus biparentally informative markers (consisting of 15,843 and 15,982 markers, respectively).

The reduced parent-specific datasets were ordered again separately, using ‘OrderMarkers’, in the same fashion as map 0.0 was ordered, to give male-parent and female-parent linkage maps. The linkage groups for the markers were the same as estimated in the construction of the previous freeze.

To further remove the inconsistencies in the separate male and female maps, using the local genome assembly information, another custom R-script was used. The following rules were established to filter SNPs:

(i) The SNPs on the same contig were assumed to have the same genetic map position. If, all but one SNP from the same contig had the same genetic map position, and a single SNP has a map position > 5cM away, the singleton SNP was removed. At least three SNPs had to be present on the contig, with two clustering together on the genetic map for the third SNP to be considered a singleton. If a contig had just two SNPs, more than 5cM apart, both the SNPs were removed.

(ii) If a contig had multiple markers clustering at various positions (that are > 5 cM) on the genetic map, the entire contig was removed.

After inconsistent markers had been filtered, parent-specific maps were merged using a custom R-script, by averaging the biparentally informative marker positions and interpolating the positions of the parent-specific markers using their proximity to neighbouring biparentally-informative markers in the relevant parent-specific map. The maps, parent-specific and averaged, generated by this two-step approach method were designated freeze maps 1.1. Another freeze, parent-averaged maps 0.1 was also generated by using just the second script, to remove inconsistent markers from the parent-averaged map 0.0.b This freeze (0.1) was not used for further analyses except to compare the efficiency of this script against published packages.

All linkage maps were plotted using the LinkageMapView R package (Oullette et al. 2018).

3.2.4 Analysis of recombination rates

Genome-wide recombination rate (GwRR) was estimated by dividing the sum of linkage map length of the parent-averaged map 1.1 by the most recent estimate of the total length of *Littorina saxatilis* genome. Chromosomal recombination rates were estimated in the same manner, but the physical length was taken as the cumulative length of the contigs included in our map. The local recombination rates were estimated for the parent-averaged and parent-specific maps by using the Marey maps (see section 3.2.2 or details regarding the construction of Marey maps) to obtain a more comprehensive understanding of recombination landscape and examine the sex-specific differences in recombination patterns. The slope of the curve relating genetic positions to physical coordinates at any given genomic coordinate provide the local recombination rate (Chakravarti 1991). MareyMap R package (Rezvoy et al. 2007) was used to carry out linear regression in a sliding window of 9 Mb. Sliding window of sizes between 1-10 Mb were tested and 9 Mb was the smallest window-size which did not give negative estimates (Siberchicot et al. 2017). To note here that mapped contigs do not cover the entire genome. Therefore, to correct for the unmapped parts, chromosomal recombination rates were scaled by the factor of 0.43 (total size of the mapped contigs/ total estimated genome size). However, no such scaling was done for the local recombination rates and thus, should be considered overestimates.

3.2.5 Genome assembly and linkage map integration

Chromonomer v1.07 (Amores et al. 2014) was used to integrate the genome assembly and the linkage maps. Chromonomer, based on a variant calling sam file, linkage maps and local genome assembly, makes decisions to remove low quality markers from the genetic map and then uses the high-quality markers to orient scaffolds in the assembly and correct the map order. For the assembly and map integration, parent-averaged maps 1.1 were used. The default settings were used as described in the manual (http://catchenlab.life.illinois.edu/chromonomer/manual/).

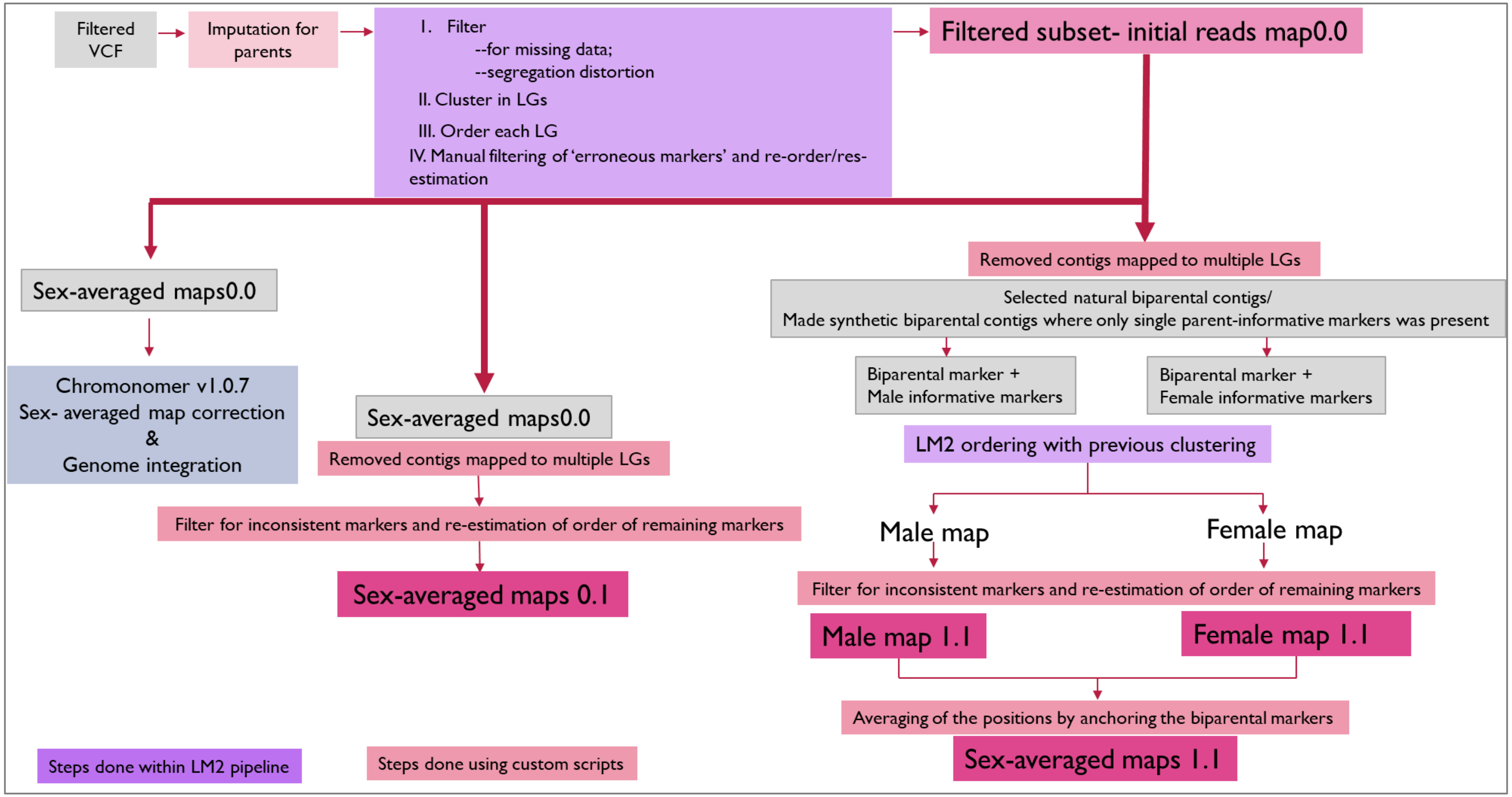

Filtered VCF
Imputation for parents
I. Filter
--for missing data;
--segregation distortion
II. Cluster in LGs
III. Order each LG
IV. Manual filtering of 'erroneous markers' and re-order/res-estimation
Filtered subset- initial reads map0.0
Sex-averaged maps0.0
Chromonomer v1.0.7
Sex- averaged map correction
&
Genome integration
Sex-averaged maps0.0
Removed contigs mapped to multiple LGs
Filter for inconsistent markers and re-estimation of order of remaining markers
Sex-averaged maps 0.1
Removed contigs mapped to multiple LGs
Selected natural biparental contigs/
Made synthetic biparental contigs where only single parent-informative markers was present
Biparental marker + Male informative markers
Biparental marker + Female informative markers
LM2 ordering with previous clustering
Male map
Female map
Filter for inconsistent markers and re-estimation of order of remaining markers
Male map 1.1
Female map 1.1
Averaging of the positions by anchoring the biparental markers
Sex-averaged maps 1.1
Steps done within LM2 pipeline
Steps done using custom scripts

Fig. 3.1: The different map strategies used. Lep-Map2 package (Rastas et al. 2016) was used to construct linkage maps 0.0. Due to physical and genetic map inconsistencies, different strategies to improve the maps were compared, including: (a) correction as suggested by the Chromonomer package, (b) removal of inconsistent markers, and (c) reduction of the dataset to include more common markers between the two parents to ease the merging of the parent-specific maps and removal of the inconsistent markers.

3.3 Results

3.3.1 Linkage maps 0.0

Quality filtering removed 12,866 SNPs. The remaining markers were clustered into 17 linkage groups, which is the expectation for the *Littorina saxatilis* system (2n= 34) (Janson 1983; Birstein and Mikhailova 1990; Rolán-Alvarez et al. 1996). With the 'SeparateChromosomes' module of Lep-Map2, 54,976 SNPs could be placed into the 17 linkage groups. After using the 'JoinSingles' module with LOD 3, 54,996 SNPs were eventually placed into linkage groups. Further, 97 SNPs were removed either from the ends of linkage groups if they contributed more than 5cM singularly or 10cM in a group of 3 or more, or if they were estimated to have error rate >0.1 by Lep-map2. The total map length (parent-averaged) was 1134.024 cM, while the average map length per LG was 66.70 cM, ranging from 53.4 cM to 90.32 cM (table 3.1).

The parent-averaged linkage maps 0.0 exhibited discrepancies when plotted against the physical distance (Chakravarti 1991) (Figure 3.2a). On closer examination, 42.3% of contigs showed discrepancies (5,194 out of 12,270- an average of 305 per chromosome), with 342 contigs mapping to multiple linkage groups.

3.3.2 Linkage maps 1.1

The 342 contigs containing SNPs that mapped to different linkage groups were eliminated from the dataset. The entire dataset was reduced to give two datasets: male-parent informative markers plus biparentally-informative markers and female-parent informative markers plus biparentally-informative markers. The male-informative plus biparentally-informative dataset had 15,843 markers with 2,961 parent-specific markers and 12,882 common markers (11,110 natural biparentally-informative markers and 1,772 synthetic biparentally informative markers) and the female plus biparentally-informative markers dataset had 15,982 markers with 3,099 parent-specific markers and 12,883 common markers (11,111 natural biparentally-informative markers and 1,772 synthetic biparentally-informative markers). The genetic map positions of the two datasets were estimated again using the Lep-Map2 'OrderMarkers' module to generate parent-specific maps. The total length of the female-parent map was 1042.41 cM, with the length of the linkage groups ranging from 41.71 cM to 102.71 cM and average marker spacing 0.27 cM (table 3.1). The total length of the male-parent map was 987.87 cM, with the length of the linkage groups ranging from 46.12 cM to 82.54 cM and average marker spacing 0.22 cM (table 3.1).

The male-parent and the female-parent maps were averaged using a custom R-script to produce a parent-averaged map (Figure 3.3). The total length of the parent-averaged map was 1011.89 cM, with smallest and largest linkage groups being 45.52 cM and 88.76 cM, respectively (table 3.1; Figure 3.3). The parent-averaged maps were compared with the parent-averaged maps 0.0 using Kendall's tau correlation coefficient based on common SNP positions and no notable change in the order within the linkage groups was observed (Kendall's tau correlation coefficient ~0 .96, table 3.2).

The parent-averaged and the parent-specific maps largely showed collinearity when Marey maps were plotted (see Figure 3.2b for parent-averaged maps and Figure 3.6 for parent-specific maps; Chakravarti 1991). An interesting result that came out of this exercise was that we were able to observe patterns in the recombination landscape that are common and distinct between both the parents. The details are discussed below (see section 3.3.3).

Table 3.1: Summary of Sex – averaged linkage maps 0.0 and sex- averaged and sex – specific linkage maps 1.1

ǂ The chromosomal lengths are not accurate as the contigs with SNPs mapping to multiple linkage groups present

** The actual number of unique contigs in sex – averaged linkage maps 0.0 is 12,270. On counting the number of unique contigs per linkage groups, the contigs mapping to multiple linkage groups are counted more than once bringing the total to 12,629

	Sex - averaged maps 0.0				Female maps 1.1							Male maps 1.1							Sex - averaged maps 1.1			
Linkage group	Map length	No. of SNPs	No. of unique contigs	Length of chromosome (Mb) ǂ	Map length	No. of SNPs	No. of unique contigs	Natural biparental markers	Synthetic biparental markers	Female specific markers	Length of chromosome (Mb)	Map length	No. of SNPs	No. of unique contigs	Natural biparental markers	Synthetic biparental markers	Male specific markers	Length of chromosome (Mb)	Map length	No. of SNPs	No. of unique contigs	Length of chromosome (Mb)
1	90.32	7020	1538	77.68	79.36	2388	1138	1873	182	333	59.19	82.55	2383	1133	1873	182	328	59.98	80.95	2716	1466	73.86
2	89.82	5776	1213	73.18	102.72	1627	859	1187	167	273	53.91	74.81	1640	872	1187	167	286	54.15	88.76	1913	1145	69.70
3	69.32	4552	983	74.51	75.65	1272	711	888	127	257	35.40	61.88	1228	667	888	127	213	33.92	68.47	1485	924	45.35
4	57.52	4202	882	47.47	58.06	1358	671	1064	121	173	32.89	58.27	1356	669	1064	121	171	34.33	56.52	1529	842	41.62
5	57.77	3667	804	62.82	58.01	1029	565	755	107	167	32.58	50.14	1048	584	755	107	186	34.39	54.07	1215	751	41.84
6	64.22	3038	721	72.22	62.33	924	512	648	77	199	28.24	58.18	893	481	648	77	168	27.08	60.25	1092	680	35.74
7	57.13	3027	631	48.06	55.44	836	468	571	125	140	25.23	47.31	825	457	571	125	129	24.40	51.37	965	597	30.83
8	78.98	2990	628	74.89	48.88	780	467	471	184	125	18.22	59.88	767	454	471	184	112	18.33	54.38	892	579	22.21
9	67.04	2935	763	35.21	61.51	788	575	374	89	325	24.02	52.77	624	411	374	89	161	16.49	56.70	949	736	29.62
10	53.39	2859	647	63.10	41.72	767	496	459	107	201	21.48	49.33	682	412	458	107	117	18.50	45.53	883	613	26.72
11	67.06	2632	568	41.21	60.81	1029	454	844	101	84	18.71	55.96	1028	453	844	101	83	17.95	58.39	1112	537	21.09
12	80.26	2568	762	71.01	56.25	537	278	475	30	32	16.00	64.22	940	681	475	30	435	37.04	60.24	972	713	38.28
13	53.60	2464	574	60.60	49.65	701	412	460	88	153	20.15	52.95	675	386	460	88	127	19.35	51.30	828	539	25.68
14	55.52	2093	504	65.18	51.30	452	283	310	72	70	13.33	46.12	567	398	310	72	185	18.36	48.71	637	468	21.08
15	65.06	2098	511	65.74	58.16	567	367	328	88	151	18.31	54.83	532	332	328	88	116	16.52	56.49	683	483	22.71
16	62.76	1505	373	33.87	58.34	377	259	187	75	115	9.02	58.29	345	227	187	75	83	8.36	57.44	460	342	11.70
17	64.27	1473	527	69.06	64.24	550	427	217	32	301	26.23	60.40	310	187	217	32	61	10.70	62.32	611	488	29.12
Total	1134.02	54899	12629**	1035.82	1042.41	15982	8942	11111	1772	3099	452.91	987.87	15843	8804	11110	1772	2961	449.85	1011.89	18942	11903	587.14

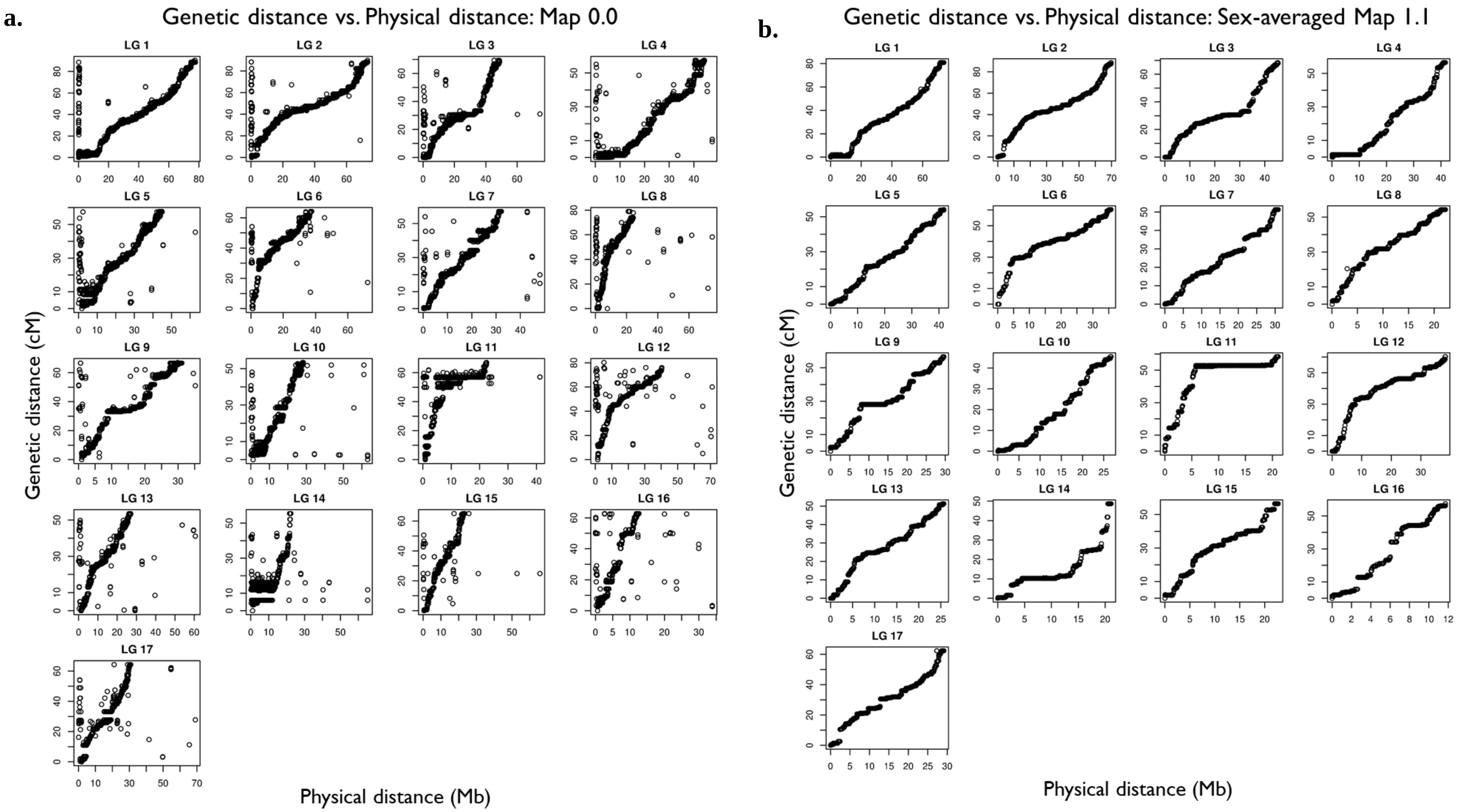


Fig. 3.2: The Marey maps for **(a)** the parent–averaged linkage maps 0.0 and **(b)** the parent-averaged maps 1.1

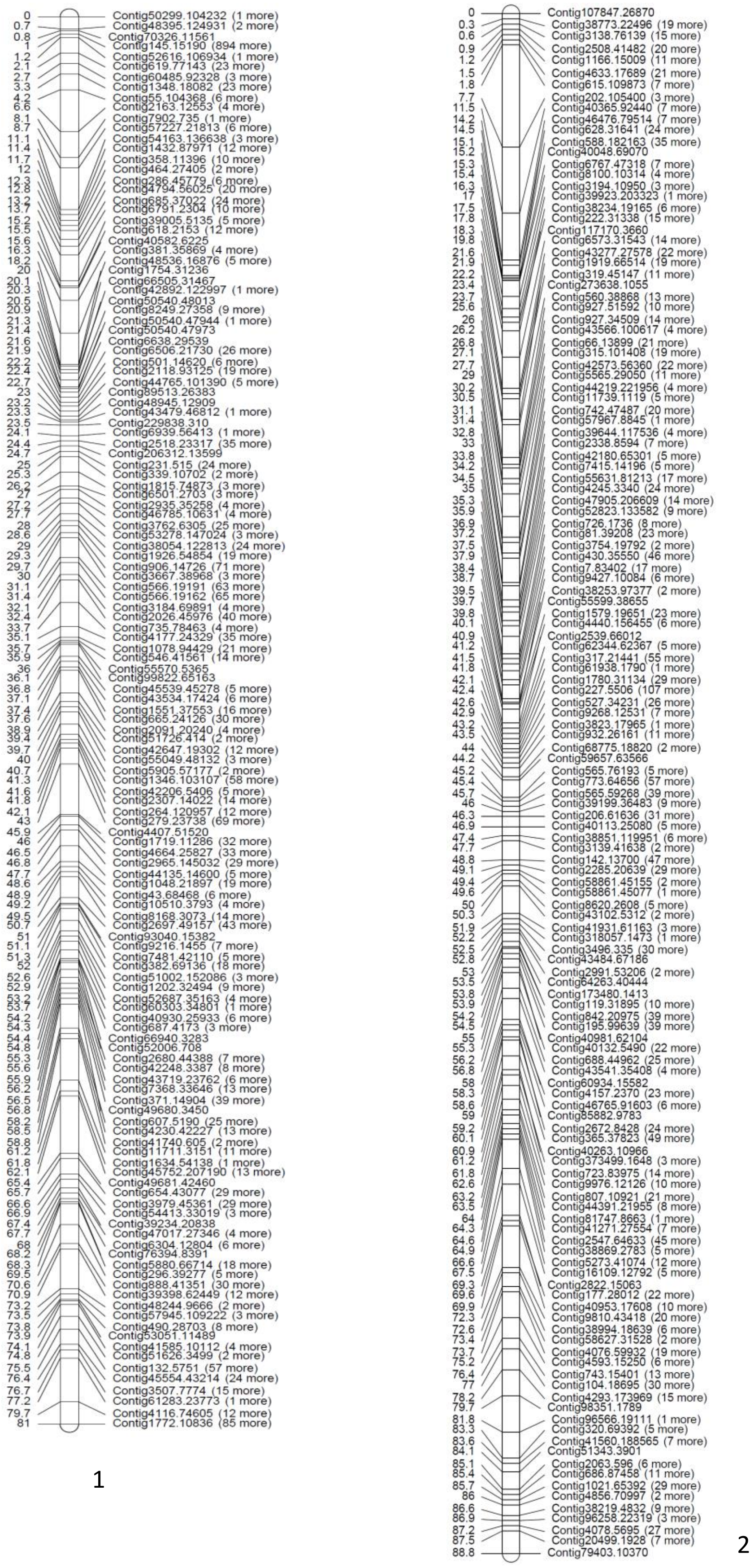

0 Contig50299.104232 (1 more)
0.7 Contig48395.124931 (2 more)
0.8 Contig70326.11561
1 Contig145.15190 (894 more)
1.2 Contig52616.106934 (1 more)
2.1 Contig619.77143 (23 more)
2.7 Contig60485.92328 (3 more)
3.3 Contig1348.18082 (23 more)
4.2 Contig55.104368 (6 more)
6.6 Contig2163.12553 (4 more)
8.1 Contig7902.735 (1 more)
8.7 Contig57227.21813 (6 more)
11.1 Contig54163.136638 (3 more)
11.4 Contig1432.87971 (12 more)
11.7 Contig358.11396 (10 more)
12 Contig464.27405 (2 more)
12.3 Contig286.45779 (6 more)
12.8 Contig4794.56025 (20 more)
13.2 Contig685.37022 (24 more)
13.7 Contig6791.2304 (10 more)
15.2 Contig39005.5135 (5 more)
15.5 Contig618.2153 (12 more)
15.6 Contig40582.6225
16.3 Contig381.35869 (4 more)
18.2 Contig48536.16876 (5 more)
20 Contig1754.31236
20.1 Contig66505.31467
20.3 Contig42892.122997 (1 more)
20.5 Contig50540.48013
20.9 Contig8249.27358 (9 more)
21.3 Contig50540.47944 (1 more)
21.4 Contig50540.47973
21.6 Contig6638.29539
21.9 Contig6506.21730 (26 more)
22.2 Contig501.14620 (6 more)
22.4 Contig2118.93125 (19 more)
22.7 Contig44765.101390 (5 more)
23 Contig89513.26383
23.2 Contig48945.12909
23.3 Contig43479.46812 (1 more)
23.5 Contig229838.310
24.1 Contig6939.56413 (1 more)
24.4 Contig2518.23317 (35 more)
24.7 Contig206312.13599
25 Contig231.515 (24 more)
25.3 Contig339.10702 (2 more)
26.2 Contig1815.74873 (3 more)
27 Contig6501.2703 (3 more)
27.2 Contig2935.35258 (4 more)
27.7 Contig46785.10631 (4 more)
28 Contig3762.6305 (25 more)
28.6 Contig53278.147024 (3 more)
29 Contig38054.122813 (24 more)
29.3 Contig1926.54854 (19 more)
29.7 Contig906.14726 (71 more)
30 Contig3667.38968 (3 more)
31.1 Contig566.19191 (63 more)
31.4 Contig566.19162 (65 more)
32.1 Contig3184.69891 (4 more)
32.4 Contig2026.45976 (40 more)
33.7 Contig735.78463 (4 more)
35.1 Contig4177.24329 (35 more)
35.7 Contig1078.94429 (21 more)
35.9 Contig546.41561 (14 more)
36 Contig55570.5365
36.1 Contig99822.65163
36.8 Contig45539.45278 (5 more)
37.1 Contig43534.17424 (6 more)
37.4 Contig1551.37553 (16 more)
37.6 Contig665.24126 (30 more)
38.9 Contig2091.20240 (4 more)
39.4 Contig51726.414 (2 more)
39.7 Contig42647.19302 (12 more)
40 Contig55049.48132 (3 more)
40.7 Contig5905.57177 (2 more)
41.3 Contig1346.103107 (58 more)
41.6 Contig42206.5406 (5 more)
41.8 Contig2307.14022 (14 more)
42.1 Contig264.120957 (12 more)
43 Contig279.23738 (69 more)
45.9 Contig4407.51520
46 Contig1719.11286 (32 more)
46.5 Contig4664.25827 (33 more)
46.8 Contig2965.145032 (29 more)
47.7 Contig44135.14600 (5 more)
48.6 Contig1048.21897 (19 more)
48.9 Contig43.68468 (6 more)
49.2 Contig10510.3793 (4 more)
49.5 Contig8168.3073 (14 more)
50.7 Contig2697.49157 (43 more)
51 Contig93040.15382
51.1 Contig9216.1455 (7 more)
51.3 Contig7481.42110 (5 more)
52 Contig382.69136 (18 more)
52.6 Contig51002.152086 (3 more)
52.9 Contig1202.32494 (9 more)
53.2 Contig52687.35163 (4 more)
53.7 Contig60303.34801 (1 more)
54.2 Contig40930.25933 (6 more)
54.3 Contig687.4173 (3 more)
54.4 Contig66940.3283
54.8 Contig52006.708
55.3 Contig2680.44388 (7 more)
55.6 Contig42248.3387 (8 more)
55.9 Contig43719.23762 (6 more)
56.2 Contig7368.33646 (13 more)
56.5 Contig371.14904 (39 more)
56.8 Contig49680.3450
58.2 Contig607.5190 (25 more)
58.5 Contig4230.42227 (13 more)
58.8 Contig41740.605 (2 more)
61.2 Contig11711.3151 (11 more)
61.8 Contig1634.54138 (1 more)
62.1 Contig45752.207190 (13 more)
65.4 Contig49681.42460
65.7 Contig654.43077 (29 more)
66.6 Contig3979.45361 (29 more)
66.9 Contig54413.33019 (3 more)
67.4 Contig39234.20838
67.7 Contig47017.27346 (4 more)
68 Contig6304.12804 (6 more)
68.2 Contig76394.8391
68.3 Contig5880.66714 (18 more)
69.5 Contig296.39277 (5 more)
70.6 Contig888.41351 (30 more)
70.9 Contig39398.62449 (12 more)
73.2 Contig48244.9666 (2 more)
73.5 Contig57945.109222 (3 more)
73.8 Contig490.28703 (8 more)
73.9 Contig53051.11489
74.1 Contig41585.10112 (4 more)
74.8 Contig51626.3499 (2 more)
75.5 Contig132.5751 (57 more)
76.4 Contig45554.43214 (24 more)
76.7 Contig3507.7774 (15 more)
77.2 Contig61283.23773 (1 more)
79.7 Contig4116.74605 (12 more)
81 Contig1772.10836 (85 more)
1
0 Contig107847.26870
0.3 Contig38773.22496 (19 more)
0.6 Contig3138.76139 (15 more)
0.9 Contig2508.41482 (20 more)
1.2 Contig1166.15009 (11 more)
1.5 Contig4633.17689 (21 more)
1.8 Contig615.109873 (7 more)
7.7 Contig202.105400 (3 more)
11.5 Contig40365.92440 (7 more)
14.2 Contig46476.79514 (7 more)
14.5 Contig628.31641 (24 more)
15.1 Contig588.182163 (35 more)
15.2 Contig40048.69070
15.3 Contig6767.47318 (7 more)
15.4 Contig8100.10314 (4 more)
16.3 Contig3194.10950 (3 more)
17 Contig39923.203323 (1 more)
17.5 Contig38234.19165 (6 more)
17.8 Contig222.31338 (15 more)
18.3 Contig117170.3660
19.8 Contig6573.31543 (14 more)
21.6 Contig43277.27578 (22 more)
21.9 Contig1919.66514 (19 more)
22.2 Contig319.45147 (11 more)
23.4 Contig273638.1055
23.7 Contig560.38868 (13 more)
25.6 Contig927.51592 (10 more)
26 Contig927.34509 (14 more)
26.2 Contig43566.100617 (4 more)
26.8 Contig66.13899 (21 more)
27.1 Contig315.101408 (19 more)
27.7 Contig42573.56360 (22 more)
29 Contig5565.29050 (11 more)
30.2 Contig44219.221956 (4 more)
30.5 Contig11739.1119 (5 more)
31.1 Contig742.47487 (20 more)
31.4 Contig57967.8845 (1 more)
32.8 Contig39644.117536 (4 more)
33 Contig2338.8594 (7 more)
33.8 Contig42180.65301 (5 more)
34.2 Contig7415.14196 (5 more)
34.5 Contig55631.81213 (17 more)
35 Contig4245.3340 (24 more)
35.3 Contig47905.206609 (14 more)
35.9 Contig52823.133582 (9 more)
36.9 Contig726.1736 (8 more)
37.2 Contig81.39208 (23 more)
37.5 Contig3754.19792 (2 more)
37.9 Contig430.35550 (46 more)
38.4 Contig7.83402 (17 more)
38.7 Contig9427.10084 (6 more)
39.5 Contig38253.97377 (2 more)
39.7 Contig55599.38655
39.8 Contig1579.19651 (23 more)
40.1 Contig4440.156455 (6 more)
40.9 Contig2539.66012
41.2 Contig62344.62367 (5 more)
41.5 Contig317.21441 (55 more)
41.8 Contig61938.1790 (1 more)
42.1 Contig1780.31134 (29 more)
42.4 Contig227.5506 (107 more)
42.6 Contig527.34231 (26 more)
42.9 Contig9268.12531 (7 more)
43.2 Contig3823.17965 (1 more)
43.5 Contig932.26161 (11 more)
44 Contig68775.18820 (2 more)
44.2 Contig59657.63566
45.2 Contig565.76193 (5 more)
45.4 Contig773.64656 (57 more)
45.7 Contig565.59268 (39 more)
46 Contig39199.36483 (9 more)
46.3 Contig206.61636 (31 more)
46.9 Contig40113.25080 (5 more)
47.4 Contig38851.119951 (6 more)
47.7 Contig3139.41638 (2 more)
48.8 Contig142.13700 (47 more)
49.1 Contig2285.20639 (29 more)
49.4 Contig58861.45155 (2 more)
49.6 Contig58861.45077 (1 more)
50 Contig8620.2608 (5 more)
50.3 Contig43102.5312 (2 more)
51.9 Contig41931.61163 (3 more)
52.2 Contig318057.1473 (1 more)
52.5 Contig3496.335 (30 more)
52.8 Contig43484.67186
53 Contig2991.53206 (2 more)
53.5 Contig64263.40444
53.8 Contig173480.1413
53.9 Contig119.31895 (10 more)
54.2 Contig842.20975 (39 more)
54.5 Contig195.99639 (39 more)
55 Contig40981.62104
55.3 Contig40132.5490 (22 more)
56.2 Contig688.44962 (25 more)
56.8 Contig43541.35408 (4 more)
58 Contig60934.15582
58.3 Contig4157.2370 (23 more)
58.6 Contig46765.91603 (6 more)
59 Contig85882.9783
59.2 Contig2672.8428 (24 more)
60.1 Contig365.37823 (49 more)
60.9 Contig40263.10966
61.2 Contig373499.1648 (3 more)
61.8 Contig723.83975 (14 more)
62.6 Contig9976.12126 (10 more)
63.2 Contig807.10921 (21 more)
63.5 Contig44391.21955 (8 more)
64 Contig81747.8663 (1 more)
64.3 Contig41271.27554 (7 more)
64.6 Contig2547.64633 (45 more)
64.9 Contig38869.2783 (5 more)
66.6 Contig5273.41074 (12 more)
67.5 Contig16109.12792 (5 more)
69.3 Contig2822.15063
69.6 Contig177.28012 (22 more)
69.9 Contig40953.17608 (10 more)
72.3 Contig9810.43418 (20 more)
72.6 Contig38994.18639 (6 more)
73.4 Contig58627.31528 (2 more)
73.7 Contig4076.59932 (19 more)
75.2 Contig4593.15250 (6 more)
76.4 Contig743.15401 (13 more)
77 Contig104.18695 (30 more)
78.2 Contig4293.173969 (15 more)
79.7 Contig98351.1789
81.8 Contig96566.19111 (1 more)
83.3 Contig320.69392 (5 more)
83.6 Contig41560.188565 (7 more)
84.1 Contig51343.3901
85.1 Contig2063.596 (6 more)
85.4 Contig686.87458 (11 more)
85.7 Contig1021.65392 (29 more)
86 Contig4856.70997 (2 more)
86.6 Contig38219.4832 (9 more)
86.9 Contig96258.22319 (3 more)
87.2 Contig4078.5695 (27 more)
87.5 Contig20499.1928 (7 more)
88.8 Contig79403.10370
2

Fig. 3.3: Sex-averaged maps 1.1

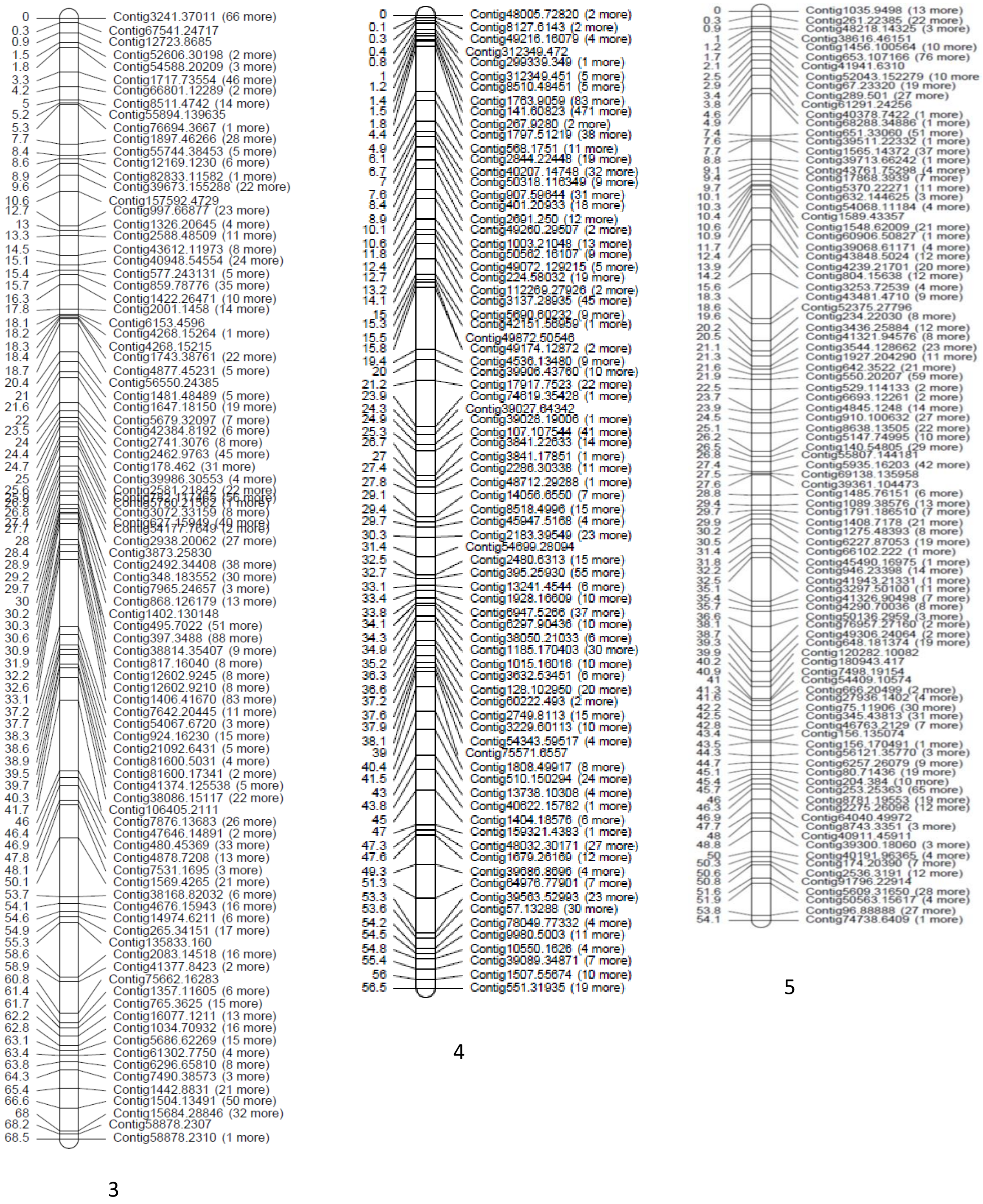

0 Contig3241.37011 (66 more)
0.3 Contig67541.24717
0.9 Contig12723.8685
1.5 Contig52606.30198 (2 more)
1.8 Contig54588.20209 (3 more)
3.3 Contig1717.73554 (46 more)
4.2 Contig66801.12289 (2 more)
5 Contig8511.4742 (14 more)
5.2 Contig55894.139635
5.3 Contig76694.3667 (1 more)
7.7 Contig1897.46266 (28 more)
8.4 Contig55744.38453 (5 more)
8.6 Contig12169.1230 (6 more)
8.9 Contig82833.11582 (1 more)
9.6 Contig39673.155288 (22 more)
10.6 Contig157592.4729
12.7 Contig997.66877 (23 more)
13 Contig1326.20645 (4 more)
13.3 Contig2588.48509 (11 more)
14.5 Contig43612.11973 (8 more)
15.1 Contig40948.54554 (24 more)
15.4 Contig577.243131 (5 more)
15.7 Contig859.78776 (35 more)
16.3 Contig1422.26471 (10 more)
17.8 Contig2001.1458 (14 more)
18.1 Contig6153.4596
18.2 Contig4268.15264 (1 more)
18.3 Contig4268.15215
18.4 Contig1743.38761 (22 more)
18.7 Contig4877.45231 (5 more)
20.4 Contig56550.24385
21 Contig1481.48489 (5 more)
21.6 Contig1647.18150 (19 more)
22 Contig5679.32097 (7 more)
23.5 Contig42384.8192 (6 more)
24 Contig2741.3076 (8 more)
24.4 Contig2462.9763 (45 more)
24.7 Contig178.462 (31 more)
25 Contig39986.30553 (4 more)
25.6 Contig2581.21842 (22 more)
25.9 Contig782.117465 (56 more)
26.2 Contig5780.21562 (1 more)
26.8 Contig3072.33159 (8 more)
27.4 Contig627.15949 (40 more)
27.7 Contig54177.7649 (2 more)
28 Contig2938.20062 (27 more)
28.4 Contig3873.25830
28.9 Contig2492.34408 (38 more)
29.2 Contig348.183552 (30 more)
29.7 Contig7965.24657 (3 more)
30 Contig868.126179 (13 more)
30.2 Contig1402.130148
30.3 Contig495.7022 (51 more)
30.6 Contig397.3488 (88 more)
30.9 Contig38814.35407 (9 more)
31.9 Contig817.16040 (8 more)
32.2 Contig12602.9245 (8 more)
32.6 Contig12602.9210 (8 more)
33.1 Contig1406.41670 (83 more)
37.2 Contig7642.20445 (11 more)
37.7 Contig54067.6720 (3 more)
38.3 Contig924.16230 (15 more)
38.6 Contig21092.6431 (5 more)
38.9 Contig81600.5031 (4 more)
39.5 Contig81600.17341 (2 more)
39.7 Contig41374.125538 (5 more)
40.3 Contig38086.15117 (22 more)
41.7 Contig106405.2111
46 Contig7876.13683 (26 more)
46.4 Contig47646.14891 (2 more)
46.9 Contig480.45369 (33 more)
47.8 Contig4878.7208 (13 more)
48.1 Contig7531.1695 (3 more)
50.1 Contig1569.4265 (21 more)
53.7 Contig38168.82032 (6 more)
54.1 Contig4676.15943 (16 more)
54.6 Contig14974.6211 (6 more)
54.9 Contig265.34151 (17 more)
55.3 Contig135833.160
58.6 Contig2083.14518 (16 more)
58.9 Contig41377.8423 (2 more)
60.8 Contig75662.16283
61.4 Contig1357.11605 (6 more)
61.7 Contig765.3625 (15 more)
62.2 Contig16077.1211 (13 more)
62.8 Contig1034.70932 (16 more)
63.1 Contig5686.62269 (15 more)
63.4 Contig61302.7750 (4 more)
63.8 Contig6296.65810 (8 more)
64.3 Contig7490.38573 (3 more)
65.4 Contig1442.8831 (21 more)
66.6 Contig1504.13491 (50 more)
68 Contig15684.28846 (32 more)
68.2 Contig58878.2307
68.5 Contig58878.2310 (1 more)
3
0 Contig48005.72820 (2 more)
0.1 Contig8127.6143 (2 more)
0.3 Contig49216.16079 (4 more)
0.4 Contig312349.472
0.8 Contig299339.349 (1 more)
1 Contig312349.451 (5 more)
1.2 Contig8510.48451 (5 more)
1.4 Contig1763.9059 (83 more)
1.5 Contig141.60823 (471 more)
1.8 Contig267.9280 (2 more)
4.4 Contig1797.51219 (38 more)
4.9 Contig568.1751 (11 more)
6.1 Contig2844.22448 (19 more)
6.7 Contig40207.14748 (32 more)
7 Contig50318.116349 (9 more)
7.6 Contig907.59644 (31 more)
8.4 Contig401.20933 (18 more)
8.9 Contig2691.250 (12 more)
10.1 Contig49260.29507 (2 more)
10.6 Contig1003.21048 (13 more)
11.8 Contig50562.16107 (9 more)
12.4 Contig49072.129215 (5 more)
12.7 Contig224.58032 (19 more)
13.2 Contig112269.27926 (2 more)
14.1 Contig3137.28935 (45 more)
15 Contig5690.60232 (9 more)
15.3 Contig42151.56959 (1 more)
15.5 Contig49872.50546
15.8 Contig49174.12872 (2 more)
19.4 Contig4536.13480 (9 more)
20 Contig39906.43760 (10 more)
21.2 Contig17917.7523 (22 more)
23.9 Contig74619.35428 (1 more)
24.3 Contig39027.64342
24.9 Contig39028.19006 (1 more)
25.3 Contig107.107544 (41 more)
26.7 Contig3841.22633 (14 more)
27 Contig3841.17851 (1 more)
27.4 Contig2286.30338 (11 more)
27.8 Contig48712.29288 (1 more)
29.1 Contig14056.6550 (7 more)
29.4 Contig8518.4996 (15 more)
29.7 Contig45947.5168 (4 more)
30.3 Contig2183.39549 (23 more)
31.4 Contig54699.28094
32.5 Contig2480.6313 (15 more)
32.7 Contig395.25930 (55 more)
33.1 Contig13241.4544 (6 more)
33.4 Contig1928.16609 (10 more)
33.8 Contig6947.5266 (37 more)
34.1 Contig6297.90436 (10 more)
34.3 Contig38050.21033 (6 more)
34.9 Contig1185.170403 (30 more)
35.2 Contig1015.16016 (10 more)
36.3 Contig3632.53451 (6 more)
36.6 Contig128.102950 (20 more)
37.2 Contig60222.493 (2 more)
37.6 Contig2749.8113 (15 more)
37.9 Contig3229.60113 (10 more)
38.1 Contig54343.59517 (4 more)
39 Contig75571.6557
40.4 Contig1808.49917 (8 more)
41.5 Contig510.150294 (24 more)
43 Contig13738.10308 (4 more)
43.8 Contig40622.15782 (1 more)
45 Contig1404.18576 (6 more)
47 Contig159321.4383 (1 more)
47.3 Contig48032.30171 (27 more)
47.6 Contig1679.26169 (12 more)
49.3 Contig39686.8696 (4 more)
51.3 Contig64976.77901 (7 more)
53.3 Contig39563.52993 (23 more)
53.6 Contig57.13288 (30 more)
54.2 Contig78049.77332 (4 more)
54.5 Contig9980.5003 (11 more)
54.8 Contig10550.1626 (4 more)
55.4 Contig39089.34871 (7 more)
56 Contig1507.55674 (10 more)
56.5 Contig551.31935 (19 more)
4
0 Contig1035.9498 (13 more)
0.3 Contig261.22385 (22 more)
0.9 Contig48218.14325 (3 more)
1 Contig38616.46151
1.2 Contig1456.100564 (10 more)
1.7 Contig653.107166 (76 more)
2.1 Contig41941.6310
2.5 Contig52043.152279 (10 more
2.9 Contig67.23320 (19 more)
3.4 Contig289.501 (27 more)
3.8 Contig61291.24256
4.6 Contig40378.7422 (1 more)
4.9 Contig68288.34886 (1 more)
7.4 Contig651.33060 (51 more)
7.6 Contig39511.22332 (1 more)
7.7 Contig1565.14372 (37 more)
8.8 Contig39713.66242 (1 more)
9.1 Contig43761.75298 (4 more)
9.4 Contig17868.3939 (7 more)
9.7 Contig5370.22271 (11 more)
10.1 Contig632.144625 (3 more)
10.3 Contig54068.11184 (4 more)
10.4 Contig1589.43357
10.6 Contig1548.62009 (21 more)
10.9 Contig60906.50827 (1 more)
11.7 Contig39068.61171 (4 more)
12.4 Contig43848.5024 (12 more)
13.9 Contig4239.21701 (20 more)
14.2 Contig804.15638 (12 more)
15.6 Contig3253.72539 (4 more)
18.3 Contig43481.4710 (9 more)
18.6 Contig52375.27796
19.6 Contig234.22030 (8 more)
20.2 Contig3436.25884 (12 more)
20.5 Contig41321.94576 (8 more)
21.1 Contig3544.128662 (23 more)
21.3 Contig1927.204290 (11 more)
21.6 Contig642.3522 (21 more)
21.9 Contig550.20207 (59 more)
22.5 Contig529.114133 (2 more)
23.7 Contig6693.12261 (2 more)
23.9 Contig4845.1248 (14 more)
24.5 Contig910.100632 (27 more)
25.1 Contig8638.13505 (22 more)
26.2 Contig5147.74995 (10 more)
26.5 Contig140.54805 (29 more)
26.8 Contig55807.144181
27.4 Contig5935.16203 (42 more)
27.5 Contig69138.135958
27.6 Contig39361.104473
28.8 Contig1485.76151 (6 more)
29.4 Contig1089.38576 (13 more)
29.7 Contig1791.186510 (7 more)
29.9 Contig1408.7178 (21 more)
30.2 Contig1275.48393 (8 more)
30.5 Contig6227.87053 (19 more)
31.4 Contig66102.222 (1 more)
31.8 Contig45490.16975 (1 more)
32.2 Contig946.23398 (14 more)
32.5 Contig41943.21331 (1 more)
35.1 Contig3297.50100 (11 more)
35.4 Contig41326.90498 (7 more)
35.7 Contig4290.70036 (8 more)
36.6 Contig50136.2959 (3 more)
38.1 Contig76957.27160 (2 more)
38.7 Contig49306.24064 (2 more)
39.3 Contig648.181374 (19 more)
39.9 Contig120282.10082
40.2 Contig180943.417
40.9 Contig7498.19154
41 Contig54409.10574
41.3 Contig666.20499 (2 more)
41.6 Contig27936.1402 (4 more)
42.2 Contig75.11906 (30 more)
42.5 Contig345.43813 (31 more)
42.8 Contig46763.2129 (7 more)
43.4 Contig156.135074
43.5 Contig156.170491 (1 more)
44.3 Contig56121.35770 (3 more)
44.7 Contig6257.26079 (9 more)
45.1 Contig80.71436 (19 more)
45.4 Contig204.384 (10 more)
45.7 Contig253.25363 (65 more)
46 Contig8781.19553 (19 more)
46.3 Contig2275.26096 (12 more)
46.9 Contig64040.49972
47.7 Contig8743.3351 (3 more)
48 Contig40911.45911
48.8 Contig39300.18060 (3 more)
50 Contig40191.96365 (4 more)
50.3 Contig174.20390 (7 more)
50.6 Contig2536.3191 (12 more)
50.8 Contig91796.22914
51.6 Contig5609.31650 (28 more)
51.9 Contig50563.15617 (4 more)
53.8 Contig96.88888 (27 more)
54.1 Contig74738.6409 (1 more)
5

Fig. 3.3 contd.: Sex-averaged maps 1.1

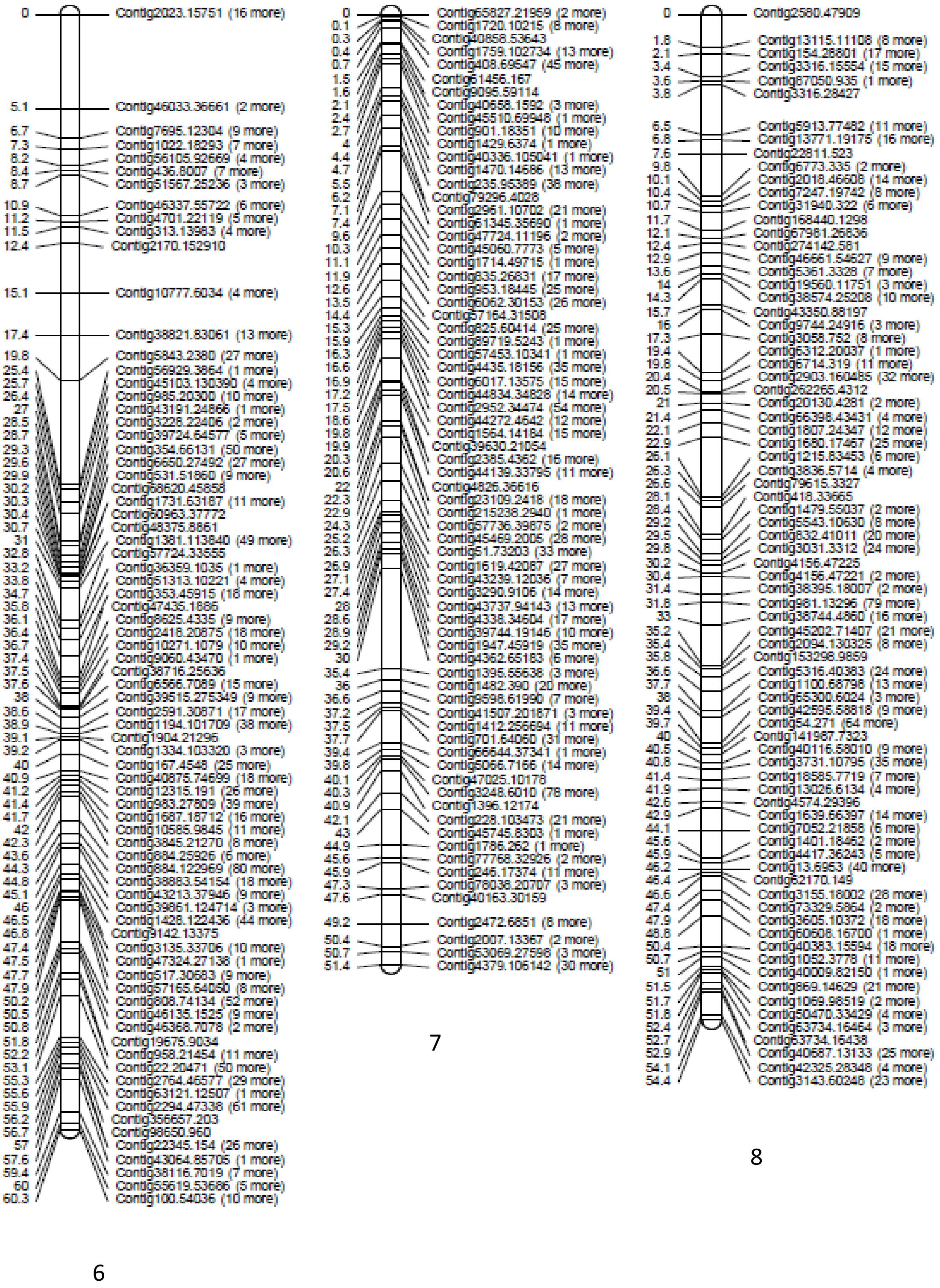

0 Contig2023.15751 (16 more)
5.1 Contig46033.36661 (2 more)
6.7 Contig7695.12304 (9 more)
7.3 Contig1022.18293 (7 more)
8.2 Contig56105.92669 (4 more)
8.4 Contig436.8007 (7 more)
8.7 Contig51567.25236 (3 more)
10.9 Contig46337.55722 (6 more)
11.2 Contig4701.22119 (5 more)
11.5 Contig313.13983 (4 more)
12.4 Contig2170.152910
15.1 Contig10777.6034 (4 more)
17.4 Contig38821.83061 (13 more)
19.8 Contig5843.2380 (27 more)
25.4 Contig56929.3864 (1 more)
25.7 Contig45103.130390 (4 more)
26.4 Contig985.20300 (10 more)
27 Contig43191.24866 (1 more)
28.5 Contig3228.22406 (2 more)
28.7 Contig39724.64577 (5 more)
29.3 Contig354.66131 (50 more)
29.6 Contig6650.27492 (27 more)
29.9 Contig531.51860 (9 more)
30.2 Contig68620.45858
30.3 Contig1731.63187 (11 more)
30.4 Contig60963.37772
30.7 Contig48375.8861
31 Contig1381.113840 (49 more)
32.8 Contig57724.33555
33.2 Contig36359.1035 (1 more)
33.8 Contig51313.10221 (4 more)
34.7 Contig353.45915 (18 more)
35.8 Contig47435.1886
36.1 Contig8625.4335 (9 more)
36.4 Contig2418.20875 (18 more)
36.7 Contig10271.1079 (10 more)
37.4 Contig9060.43470 (1 more)
37.5 Contig38716.25636
37.6 Contig6566.7089 (15 more)
38 Contig39515.275349 (9 more)
38.6 Contig2591.30871 (17 more)
38.9 Contig1194.101709 (38 more)
39.1 Contig1904.21296
39.2 Contig1334.103320 (3 more)
40 Contig167.4548 (25 more)
40.9 Contig40875.74699 (18 more)
41.2 Contig12315.191 (26 more)
41.4 Contig983.27809 (39 more)
41.7 Contig1687.18712 (16 more)
42 Contig10585.9845 (11 more)
42.3 Contig3845.21270 (8 more)
43.6 Contig884.25926 (6 more)
44.3 Contig884.122969 (80 more)
44.8 Contig38883.54154 (18 more)
45.1 Contig43213.37946 (9 more)
46 Contig39861.124714 (3 more)
46.5 Contig1428.122436 (44 more)
46.8 Contig9142.13375
47.4 Contig3135.33706 (10 more)
47.5 Contig47324.27136 (1 more)
47.7 Contig517.30683 (9 more)
47.9 Contig57165.64050 (8 more)
50.2 Contig808.74134 (52 more)
50.5 Contig46135.1525 (9 more)
50.8 Contig46366.7078 (2 more)
51.8 Contig19675.9034
52.2 Contig958.21454 (11 more)
53.1 Contig22.20471 (50 more)
55.3 Contig2764.46577 (29 more)
55.6 Contig63121.12507 (1 more)
55.9 Contig2294.47338 (61 more)
56.2 Contig356657.203
56.7 Contig98650.960
57 Contig22345.154 (26 more)
57.6 Contig43064.85705 (1 more)
59.4 Contig38116.7019 (7 more)
60 Contig55619.53686 (5 more)
60.3 Contig100.54036 (10 more)
6
0 Contig65827.21959 (2 more)
0.1 Contig1720.10215 (8 more)
0.3 Contig40858.53643
0.4 Contig1759.102734 (13 more)
0.7 Contig408.69547 (45 more)
1.5 Contig61456.167
1.6 Contig9095.59114
2.1 Contig40658.1592 (3 more)
2.4 Contig45510.69948 (1 more)
2.7 Contig901.18351 (10 more)
4 Contig1429.6374 (1 more)
4.4 Contig40336.105041 (1 more)
4.7 Contig1470.14686 (13 more)
5.5 Contig235.95389 (38 more)
6.2 Contig79296.4028
7.1 Contig2961.10702 (21 more)
7.4 Contig61345.35690 (1 more)
9.6 Contig47724.11196 (2 more)
10.3 Contig45060.7773 (5 more)
11.1 Contig1714.49715 (1 more)
11.9 Contig635.26631 (17 more)
12.6 Contig953.18445 (25 more)
13.5 Contig6062.30153 (26 more)
14.4 Contig57164.31508
15.3 Contig825.60414 (25 more)
15.9 Contig69719.5243 (1 more)
16.3 Contig57453.10341 (1 more)
16.6 Contig4435.18156 (35 more)
16.9 Contig6017.13575 (15 more)
17.2 Contig44834.34828 (14 more)
17.5 Contig2952.34474 (54 more)
18.6 Contig44272.4642 (12 more)
19.8 Contig1564.14184 (15 more)
19.9 Contig39630.21054
20.3 Contig2385.4362 (16 more)
20.6 Contig44139.33795 (11 more)
22 Contig4826.36616
22.3 Contig23109.2418 (18 more)
22.9 Contig215238.2940 (1 more)
24.3 Contig57736.39875 (2 more)
25.2 Contig45469.2005 (28 more)
26.3 Contig51.73203 (33 more)
26.9 Contig1619.42087 (27 more)
27.1 Contig43239.12036 (7 more)
27.4 Contig3290.9106 (14 more)
28 Contig43737.94143 (13 more)
28.6 Contig4338.34604 (17 more)
28.9 Contig39744.19146 (10 more)
29.2 Contig1947.45919 (35 more)
30 Contig4362.65183 (6 more)
35.4 Contig1395.55638 (3 more)
36 Contig1482.390 (20 more)
36.6 Contig9598.61990 (7 more)
37.2 Contig41507.201871 (3 more)
37.5 Contig1412.256694 (11 more)
37.7 Contig701.64060 (31 more)
39.4 Contig66644.37341 (1 more)
39.8 Contig5066.7166 (14 more)
40.1 Contig47025.10178
40.3 Contig3248.6010 (78 more)
40.9 Contig1396.12174
42.1 Contig228.103473 (21 more)
43 Contig45745.8303 (1 more)
44.9 Contig1786.262 (1 more)
45.6 Contig77768.32926 (2 more)
45.9 Contig246.17374 (11 more)
47.3 Contig78038.20707 (3 more)
47.6 Contig40163.30159
49.2 Contig2472.6851 (8 more)
50.4 Contig2007.13367 (2 more)
50.7 Contig53069.27598 (3 more)
51.4 Contig4379.106142 (30 more)
7
0 Contig2580.47909
1.8 Contig13115.11108 (8 more)
2.1 Contig154.28801 (17 more)
3.4 Contig3316.15554 (15 more)
3.6 Contig67050.935 (1 more)
3.8 Contig3316.28427
6.5 Contig5913.77482 (11 more)
6.8 Contig13771.19175 (16 more)
7.6 Contig22811.523
9.8 Contig6773.335 (2 more)
10.1 Contig2018.46608 (14 more)
10.4 Contig7247.19742 (8 more)
10.7 Contig31940.322 (6 more)
11.7 Contig168440.1298
12.1 Contig67981.26836
12.4 Contig274142.581
12.9 Contig46661.54627 (9 more)
13.6 Contig5361.3328 (7 more)
14 Contig19560.11751 (3 more)
14.3 Contig38574.25208 (10 more)
15.7 Contig43350.88197
16 Contig9744.24916 (3 more)
17.3 Contig3058.752 (8 more)
19.4 Contig6312.20037 (1 more)
19.8 Contig6714.319 (11 more)
20.4 Contig2903.160485 (32 more)
20.5 Contig262265.4312
21 Contig20130.4281 (2 more)
21.4 Contig66398.43431 (4 more)
22.1 Contig1807.24347 (12 more)
22.9 Contig1680.17467 (25 more)
26.1 Contig1215.83453 (6 more)
26.3 Contig3836.5714 (4 more)
26.6 Contig79615.3327
28.1 Contig418.33665
28.4 Contig1479.55037 (2 more)
29.2 Contig5543.10630 (8 more)
29.5 Contig832.41011 (20 more)
29.8 Contig3031.3312 (24 more)
30.2 Contig4156.47225
30.4 Contig4156.47221 (2 more)
31.4 Contig38395.18007 (2 more)
31.8 Contig981.13296 (79 more)
33 Contig38744.4860 (16 more)
35.2 Contig45202.71407 (21 more)
35.4 Contig2094.130325 (8 more)
35.8 Contig153296.9859
36.6 Contig5316.40383 (24 more)
37.7 Contig1100.68798 (13 more)
38 Contig65300.6024 (3 more)
39.4 Contig42595.58818 (9 more)
39.7 Contig54.271 (64 more)
40 Contig141987.7323
40.5 Contig40116.58010 (9 more)
40.8 Contig3731.10795 (35 more)
41.4 Contig18585.7719 (7 more)
41.9 Contig13026.6134 (4 more)
42.6 Contig4574.29396
42.9 Contig1639.66397 (14 more)
44.1 Contig7052.21858 (6 more)
45.6 Contig1401.16462 (2 more)
45.9 Contig4417.36243 (5 more)
46.2 Contig13.6953 (40 more)
46.4 Contig62170.149
46.6 Contig3155.18002 (26 more)
47.4 Contig73329.5864 (2 more)
47.9 Contig3605.10372 (18 more)
48.8 Contig60608.16700 (1 more)
50.4 Contig40383.15594 (18 more)
50.7 Contig1052.3778 (11 more)
51 Contig40009.82150 (1 more)
51.5 Contig869.14629 (21 more)
51.7 Contig1069.98519 (2 more)
51.8 Contig50470.33429 (4 more)
52.4 Contig63734.16464 (3 more)
52.7 Contig63734.16438
52.9 Contig40687.13133 (25 more)
54.1 Contig42325.28348 (4 more)
54.4 Contig3143.60248 (23 more)
8

Fig. 3.3 contd.: Sex-averaged maps 1.1

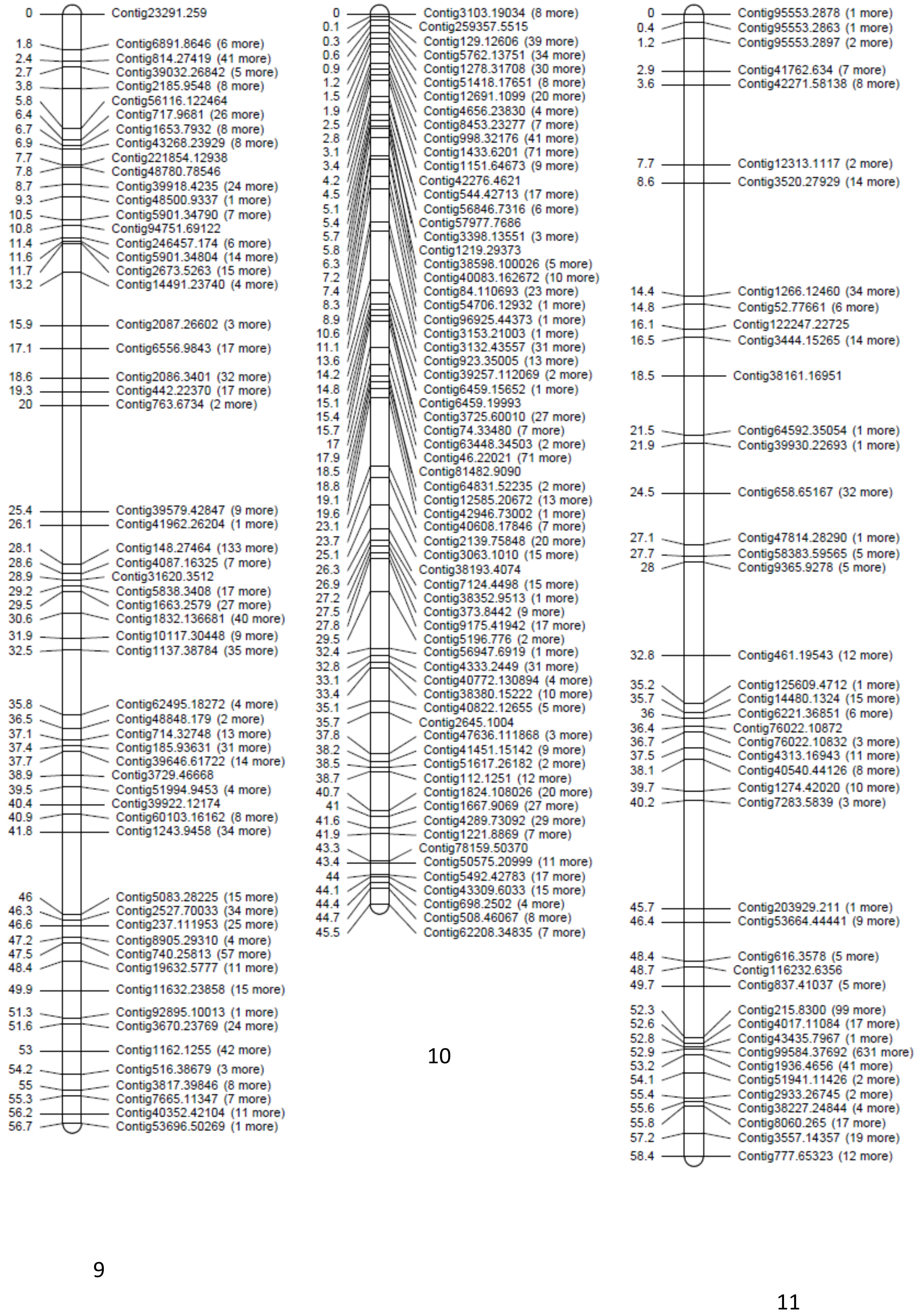

0 Contig23291.259
1.8 Contig6891.8646 (6 more)
2.4 Contig814.27419 (41 more)
2.7 Contig39032.26842 (5 more)
3.8 Contig2185.9548 (8 more)
5.8 Contig56116.122464
6.4 Contig717.9681 (26 more)
6.7 Contig1653.7932 (8 more)
6.9 Contig43268.23929 (8 more)
7.7 Contig221854.12938
7.8 Contig48780.78546
8.7 Contig39918.4235 (24 more)
9.3 Contig48500.9337 (1 more)
10.5 Contig5901.34790 (7 more)
10.8 Contig94751.69122
11.4 Contig246457.174 (6 more)
11.6 Contig5901.34804 (14 more)
11.7 Contig2673.5263 (15 more)
13.2 Contig14491.23740 (4 more)
15.9 Contig2087.26602 (3 more)
17.1 Contig6556.9843 (17 more)
18.6 Contig2086.3401 (32 more)
19.3 Contig442.22370 (17 more)
20 Contig763.6734 (2 more)
25.4 Contig39579.42847 (9 more)
26.1 Contig41962.26204 (1 more)
28.1 Contig148.27464 (133 more)
28.6 Contig4087.16325 (7 more)
28.9 Contig31620.3512
29.2 Contig5838.3408 (17 more)
29.5 Contig1663.2579 (27 more)
30.6 Contig1832.136681 (40 more)
31.9 Contig10117.30448 (9 more)
32.5 Contig1137.38784 (35 more)
35.8 Contig62495.18272 (4 more)
36.5 Contig48848.179 (2 more)
37.1 Contig714.32748 (13 more)
37.4 Contig185.93631 (31 more)
37.7 Contig39646.61722 (14 more)
38.9 Contig3729.46668
39.5 Contig51994.9453 (4 more)
40.4 Contig39922.12174
40.9 Contig60103.16162 (8 more)
41.8 Contig1243.9458 (34 more)
46 Contig5083.28225 (15 more)
46.3 Contig2527.70033 (34 more)
46.6 Contig237.111953 (25 more)
47.2 Contig8905.29310 (4 more)
47.5 Contig740.25813 (57 more)
48.4 Contig19632.5777 (11 more)
49.9 Contig11632.23858 (15 more)
51.3 Contig92895.10013 (1 more)
51.6 Contig3670.23769 (24 more)
53 Contig1162.1255 (42 more)
54.2 Contig516.38679 (3 more)
55 Contig3817.39846 (8 more)
55.3 Contig7665.11347 (7 more)
56.2 Contig40352.42104 (11 more)
56.7 Contig53696.50269 (1 more)
9
0 Contig3103.19034 (8 more)
0.1 Contig259357.5515
0.3 Contig129.12606 (39 more)
0.6 Contig5762.13751 (34 more)
0.9 Contig1278.31708 (30 more)
1.2 Contig51418.17651 (8 more)
1.5 Contig12691.1099 (20 more)
1.9 Contig4656.23830 (4 more)
2.5 Contig8453.23277 (7 more)
2.8 Contig998.32176 (41 more)
3.1 Contig1433.6201 (71 more)
3.4 Contig1151.64673 (9 more)
4.2 Contig42276.4621
4.5 Contig544.42713 (17 more)
5.1 Contig56846.7316 (6 more)
5.4 Contig57977.7686
5.7 Contig3398.13551 (3 more)
5.8 Contig1219.29373
6.3 Contig38598.100026 (5 more)
7.2 Contig40083.162672 (10 more)
7.4 Contig84.110693 (23 more)
8.3 Contig54706.12932 (1 more)
8.9 Contig96925.44373 (1 more)
10.6 Contig3153.21003 (1 more)
11.1 Contig3132.43557 (31 more)
13.6 Contig923.35005 (13 more)
14.2 Contig39257.112069 (2 more)
14.8 Contig6459.15652 (1 more)
15.1 Contig6459.19993
15.4 Contig3725.60010 (27 more)
15.7 Contig74.33480 (7 more)
17 Contig63448.34503 (2 more)
17.9 Contig46.22021 (71 more)
18.5 Contig81482.9090
18.8 Contig64831.52235 (2 more)
19.1 Contig12585.20672 (13 more)
19.6 Contig42946.73002 (1 more)
23.1 Contig40608.17846 (7 more)
23.7 Contig2139.75848 (20 more)
25.1 Contig3063.1010 (15 more)
26.3 Contig38193.4074
26.9 Contig7124.4498 (15 more)
27.2 Contig38352.9513 (1 more)
27.5 Contig373.8442 (9 more)
27.8 Contig9175.41942 (17 more)
29.5 Contig5196.776 (2 more)
32.4 Contig56947.6919 (1 more)
32.8 Contig4333.2449 (31 more)
33.1 Contig40772.130894 (4 more)
33.4 Contig38380.15222 (10 more)
35.1 Contig40822.12655 (5 more)
35.7 Contig2645.1004
37.8 Contig47636.111868 (3 more)
38.2 Contig41451.15142 (9 more)
38.5 Contig51617.26182 (2 more)
38.7 Contig112.1251 (12 more)
40.7 Contig1824.108026 (20 more)
41 Contig1667.9069 (27 more)
41.6 Contig4289.73092 (29 more)
41.9 Contig1221.8869 (7 more)
43.3 Contig78159.50370
43.4 Contig50575.20999 (11 more)
44 Contig5492.42783 (17 more)
44.1 Contig43309.6033 (15 more)
44.4 Contig698.2502 (4 more)
44.7 Contig508.46067 (8 more)
45.5 Contig62208.34835 (7 more)
10
0 Contig95553.2878 (1 more)
0.4 Contig95553.2863 (1 more)
1.2 Contig95553.2897 (2 more)
2.9 Contig41762.634 (7 more)
3.6 Contig42271.58138 (8 more)
7.7 Contig12313.1117 (2 more)
8.6 Contig3520.27929 (14 more)
14.4 Contig1266.12460 (34 more)
14.8 Contig52.77661 (6 more)
16.1 Contig122247.22725
16.5 Contig3444.15265 (14 more)
18.5 Contig38161.16951
21.5 Contig64592.35054 (1 more)
21.9 Contig39930.22693 (1 more)
24.5 Contig658.65167 (32 more)
27.1 Contig47814.28290 (1 more)
27.7 Contig58383.59565 (5 more)
28 Contig9365.9278 (5 more)
32.8 Contig461.19543 (12 more)
35.2 Contig125609.4712 (1 more)
35.7 Contig14480.1324 (15 more)
36 Contig6221.36851 (6 more)
36.4 Contig76022.10872
36.7 Contig76022.10832 (3 more)
37.5 Contig4313.16943 (11 more)
38.1 Contig40540.44126 (8 more)
39.7 Contig1274.42020 (10 more)
40.2 Contig7283.5839 (3 more)
45.7 Contig203929.211 (1 more)
46.4 Contig53664.44441 (9 more)
48.4 Contig616.3578 (5 more)
48.7 Contig116232.6356
49.7 Contig837.41037 (5 more)
52.3 Contig215.8300 (99 more)
52.6 Contig4017.11084 (17 more)
52.8 Contig43435.7967 (1 more)
52.9 Contig99584.37692 (631 more)
53.2 Contig1936.4656 (41 more)
54.1 Contig51941.11426 (2 more)
55.4 Contig2933.26745 (2 more)
55.6 Contig38227.24844 (4 more)
55.8 Contig8060.265 (17 more)
57.2 Contig3557.14357 (19 more)
58.4 Contig777.65323 (12 more)
11

Fig. 3.3 contd.: Sex-averaged maps 1.1

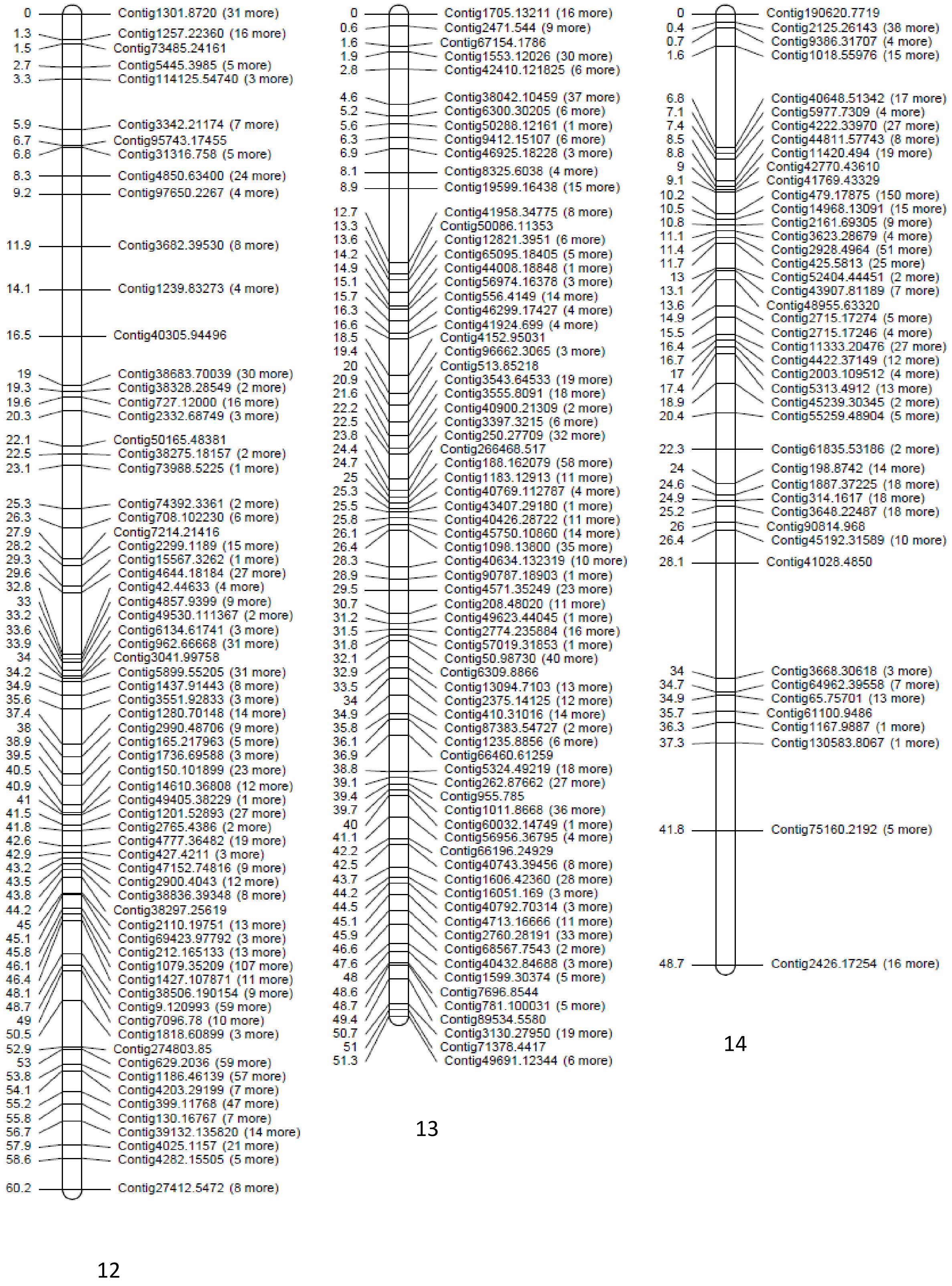
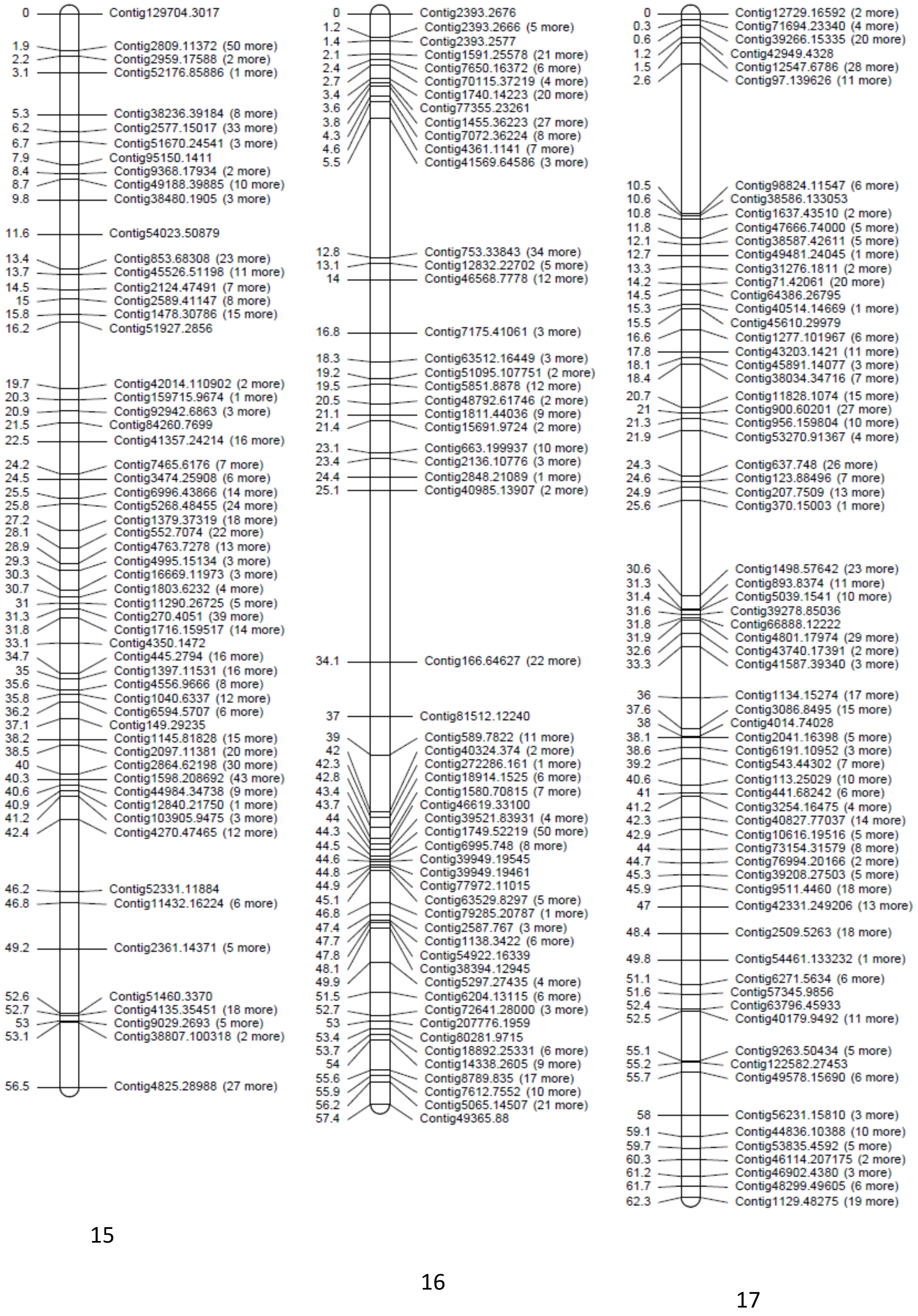

0 Contig1301.8720 (31 more)
1.3 Contig1257.22360 (16 more)
1.5 Contig73485.24161
2.7 Contig5445.3985 (5 more)
3.3 Contig114125.54740 (3 more)
5.9 Contig3342.21174 (7 more)
6.7 Contig95743.17455
6.8 Contig31316.758 (5 more)
8.3 Contig4850.63400 (24 more)
9.2 Contig97650.2267 (4 more)
11.9 Contig3682.39530 (8 more)
14.1 Contig1239.83273 (4 more)
16.5 Contig40305.94496
19 Contig38683.70039 (30 more)
19.3 Contig38328.28549 (2 more)
19.6 Contig727.12000 (16 more)
20.3 Contig2332.68749 (3 more)
22.1 Contig50165.48381
22.5 Contig38275.18157 (2 more)
23.1 Contig73988.5225 (1 more)
25.3 Contig74392.3361 (2 more)
26.3 Contig708.102230 (6 more)
27.9 Contig7214.21416
28.2 Contig2299.1189 (15 more)
29.3 Contig15567.3262 (1 more)
29.6 Contig4644.18184 (27 more)
32.8 Contig42.44633 (4 more)
33 Contig4857.9399 (9 more)
33.2 Contig49530.111367 (2 more)
33.6 Contig6134.61741 (3 more)
33.9 Contig962.66668 (31 more)
34 Contig3041.99758
34.2 Contig5899.55205 (31 more)
34.9 Contig1437.91443 (8 more)
35.6 Contig3551.92833 (3 more)
37.4 Contig1280.70148 (14 more)
38 Contig2990.48706 (9 more)
38.9 Contig165.217963 (5 more)
39.5 Contig1736.69588 (3 more)
40.5 Contig150.101899 (23 more)
40.9 Contig14610.36808 (12 more)
41 Contig49405.38229 (1 more)
41.5 Contig1201.52893 (27 more)
41.8 Contig2765.4386 (2 more)
42.6 Contig4777.36482 (19 more)
42.9 Contig427.4211 (3 more)
43.2 Contig47152.74816 (9 more)
43.5 Contig2900.4043 (12 more)
43.8 Contig38836.39348 (8 more)
44.2 Contig38297.25619
45 Contig2110.19751 (13 more)
45.1 Contig69423.97792 (3 more)
45.8 Contig212.165133 (13 more)
46.1 Contig1079.35209 (107 more)
46.4 Contig1427.107871 (11 more)
48.1 Contig38506.190154 (9 more)
48.7 Contig9.120993 (59 more)
49 Contig7096.78 (10 more)
50.5 Contig1818.60899 (3 more)
52.9 Contig274803.85
53 Contig629.2036 (59 more)
53.8 Contig1186.46139 (57 more)
54.1 Contig4203.29199 (7 more)
55.2 Contig399.11768 (47 more)
55.8 Contig130.16767 (7 more)
56.7 Contig39132.135820 (14 more)
57.9 Contig4025.1157 (21 more)
58.6 Contig4282.15505 (5 more)
60.2 Contig27412.5472 (8 more)
12
0 Contig1705.13211 (16 more)
0.6 Contig2471.544 (9 more)
1.6 Contig67154.1786
1.9 Contig1553.12026 (30 more)
2.8 Contig42410.121825 (6 more)
4.6 Contig38042.10459 (37 more)
5.2 Contig6300.30205 (6 more)
5.6 Contig50288.12161 (1 more)
6.3 Contig9412.15107 (6 more)
6.9 Contig46925.18228 (3 more)
8.1 Contig8325.6038 (4 more)
8.9 Contig19599.16438 (15 more)
12.7 Contig41958.34775 (8 more)
13.3 Contig50086.11353
13.6 Contig12821.3951 (6 more)
14.2 Contig65095.18405 (5 more)
14.9 Contig44008.18846 (1 more)
15.1 Contig56974.16378 (3 more)
15.7 Contig556.4149 (14 more)
16.3 Contig46299.17427 (4 more)
16.6 Contig41924.699 (4 more)
18.5 Contig4152.95031
19.4 Contig96662.3065 (3 more)
20 Contig513.85218
20.9 Contig3543.64533 (19 more)
21.6 Contig3555.8091 (18 more)
22.2 Contig40900.21309 (2 more)
22.5 Contig3397.3215 (6 more)
23.8 Contig250.27709 (32 more)
24.4 Contig266468.517
24.7 Contig188.162079 (58 more)
25 Contig1183.12913 (11 more)
25.3 Contig40769.112787 (4 more)
25.5 Contig43407.29180 (1 more)
25.8 Contig40426.28722 (11 more)
26.1 Contig45750.10860 (14 more)
26.4 Contig1098.13800 (35 more)
28.3 Contig40634.132319 (10 more)
28.9 Contig90787.16903 (1 more)
29.5 Contig4571.35249 (23 more)
30.7 Contig208.48020 (11 more)
31.2 Contig49623.44045 (1 more)
31.5 Contig2774.235884 (16 more)
31.8 Contig57019.31853 (1 more)
32.1 Contig50.98730 (40 more)
32.9 Contig6309.8866
33.5 Contig13094.7103 (13 more)
34 Contig2375.14125 (12 more)
34.9 Contig410.31016 (14 more)
35.8 Contig87383.54727 (2 more)
36.1 Contig1235.8856 (6 more)
36.9 Contig66460.61259
38.8 Contig5324.49219 (18 more)
39.1 Contig262.87662 (27 more)
39.4 Contig955.785
39.7 Contig1011.8668 (36 more)
40 Contig60032.14749 (1 more)
41.1 Contig56956.36795 (4 more)
42.2 Contig66196.24929
42.5 Contig40743.39456 (8 more)
43.7 Contig1606.42360 (28 more)
44.2 Contig16051.169 (3 more)
44.5 Contig40792.70314 (3 more)
45.1 Contig4713.16666 (11 more)
45.9 Contig2760.28191 (33 more)
46.6 Contig68567.7543 (2 more)
47.6 Contig40432.84688 (3 more)
48 Contig1599.30374 (5 more)
48.6 Contig7696.8544
48.7 Contig781.100031 (5 more)
49.4 Contig89534.5580
50.7 Contig3130.27950 (19 more)
51 Contig71378.4417
51.3 Contig49691.12344 (6 more)
13
0 Contig190620.7719
0.4 Contig2125.26143 (38 more)
0.7 Contig9386.31707 (4 more)
1.6 Contig1018.55976 (15 more)
6.8 Contig40648.51342 (17 more)
7.1 Contig5977.7309 (4 more)
7.4 Contig4222.33970 (27 more)
8.5 Contig44811.57743 (8 more)
8.8 Contig11420.494 (19 more)
9 Contig42770.43610
9.1 Contig41769.43329
10.2 Contig479.17875 (150 more)
10.5 Contig14968.13091 (15 more)
10.8 Contig2161.69305 (9 more)
11.1 Contig3623.28679 (4 more)
11.4 Contig2928.4964 (51 more)
11.7 Contig425.5813 (25 more)
13 Contig52404.44451 (2 more)
13.1 Contig43907.81189 (7 more)
13.6 Contig48955.63320
14.9 Contig2715.17274 (5 more)
15.5 Contig2715.17246 (4 more)
16.4 Contig11333.20476 (27 more)
16.7 Contig4422.37149 (12 more)
17 Contig2003.109512 (4 more)
17.4 Contig5313.4912 (13 more)
18.9 Contig45239.30345 (2 more)
20.4 Contig55259.48904 (5 more)
22.3 Contig61835.53186 (2 more)
24 Contig198.8742 (14 more)
24.6 Contig1887.37225 (18 more)
24.9 Contig314.1617 (16 more)
25.2 Contig3648.22487 (18 more)
26 Contig90814.968
26.4 Contig45192.31589 (10 more)
28.1 Contig41028.4850
34 Contig3668.30618 (3 more)
34.7 Contig64962.39558 (7 more)
34.9 Contig65.75701 (13 more)
35.7 Contig61100.9486
36.3 Contig1167.9887 (1 more)
37.3 Contig130583.8067 (1 more)
41.8 Contig75160.2192 (5 more)
48.7 Contig2426.17254 (16 more)
14

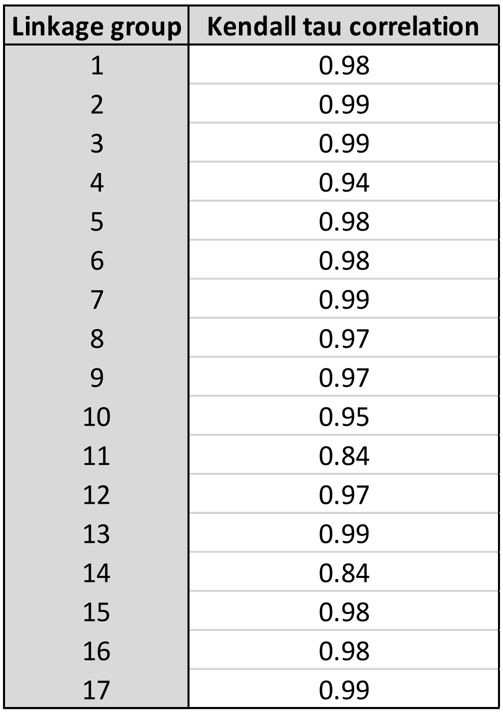
Fig. 3.3 contd.: Sex-averaged maps 1.1



0 Contig129704.3017
1.9 Contig2809.11372 (50 more)
2.2 Contig2959.17588 (2 more)
3.1 Contig52176.85886 (1 more)
5.3 Contig38236.39184 (8 more)
6.2 Contig2577.15017 (33 more)
6.7 Contig51670.24541 (3 more)
7.9 Contig95150.1411
8.4 Contig9368.17934 (2 more)
8.7 Contig49188.39885 (10 more)
9.8 Contig38480.1905 (3 more)
11.6 Contig54023.50879
13.4 Contig853.68308 (23 more)
13.7 Contig45526.51198 (11 more)
14.5 Contig2124.47491 (7 more)
15 Contig2589.41147 (8 more)
15.8 Contig1478.30786 (15 more)
16.2 Contig51927.2856
19.7 Contig42014.110902 (2 more)
20.3 Contig159715.9674 (1 more)
20.9 Contig92942.6863 (3 more)
21.5 Contig84260.7699
22.5 Contig41357.24214 (16 more)
24.2 Contig7465.6176 (7 more)
24.5 Contig3474.25908 (6 more)
25.5 Contig6996.43866 (14 more)
25.8 Contig5268.48455 (24 more)
27.2 Contig1379.37319 (18 more)
28.1 Contig552.7074 (22 more)
28.9 Contig4763.7278 (13 more)
29.3 Contig4995.15134 (3 more)
30.3 Contig16669.11973 (3 more)
30.7 Contig1803.6232 (4 more)
31 Contig11290.26725 (5 more)
31.3 Contig270.4051 (39 more)
31.8 Contig1716.159517 (14 more)
33.1 Contig4350.1472
34.7 Contig445.2794 (16 more)
35 Contig1397.11531 (16 more)
35.6 Contig4556.9666 (8 more)
35.8 Contig1040.6337 (12 more)
36.2 Contig6594.5707 (6 more)
37.1 Contig149.29235
38.2 Contig1145.81828 (15 more)
38.5 Contig2097.11381 (20 more)
40 Contig2864.62198 (30 more)
40.3 Contig1598.208692 (43 more)
40.6 Contig44984.34738 (9 more)
40.9 Contig12840.21750 (1 more)
41.2 Contig103905.9475 (3 more)
42.4 Contig4270.47465 (12 more)
46.2 Contig52331.11884
46.8 Contig11432.16224 (6 more)
49.2 Contig2361.14371 (5 more)
52.6 Contig51460.3370
52.7 Contig4135.35451 (18 more)
53 Contig9029.2693 (5 more)
53.1 Contig38807.100318 (2 more)
56.5 Contig4825.28988 (27 more)
15
0 Contig2393.2676
1.2 Contig2393.2666 (5 more)
1.4 Contig2393.2577
2.1 Contig1591.25578 (21 more)
2.4 Contig7650.16372 (6 more)
2.7 Contig70115.37219 (4 more)
3.4 Contig1740.14223 (20 more)
3.6 Contig77355.23261
3.8 Contig1455.36223 (27 more)
4.3 Contig7072.36224 (8 more)
4.6 Contig4361.1141 (7 more)
5.5 Contig41569.64586 (3 more)
12.8 Contig753.33843 (34 more)
13.1 Contig12832.22702 (5 more)
14 Contig46568.7778 (12 more)
16.8 Contig7175.41061 (3 more)
18.3 Contig63512.16449 (3 more)
19.2 Contig51095.107751 (2 more)
19.5 Contig5851.8878 (12 more)
20.5 Contig48792.61746 (2 more)
21.1 Contig1811.44036 (9 more)
21.4 Contig15691.9724 (2 more)
23.1 Contig663.199937 (10 more)
23.4 Contig2136.10776 (3 more)
24.4 Contig2848.21089 (1 more)
25.1 Contig40985.13907 (2 more)
34.1 Contig166.64627 (22 more)
37 Contig81512.12240
39 Contig589.7822 (11 more)
42 Contig40324.374 (2 more)
42.3 Contig272286.161 (1 more)
42.8 Contig18914.1525 (6 more)
43.4 Contig1580.70815 (7 more)
43.7 Contig46619.33100
44 Contig39521.83931 (4 more)
44.3 Contig1749.52219 (50 more)
44.5 Contig6995.748 (8 more)
44.6 Contig39949.19545
44.8 Contig39949.19461
44.9 Contig77972.11015
45.1 Contig63529.8297 (5 more)
46.8 Contig79285.20787 (1 more)
47.4 Contig2587.767 (3 more)
47.7 Contig1138.3422 (6 more)
47.8 Contig54922.16339
48.1 Contig38394.12945
49.9 Contig5297.27435 (4 more)
51.5 Contig6204.13115 (6 more)
52.7 Contig72641.28000 (3 more)
53 Contig207776.1959
53.4 Contig80281.9715
53.7 Contig18892.25331 (6 more)
54 Contig14338.2605 (9 more)
55.6 Contig8789.835 (17 more)
55.9 Contig7612.7552 (10 more)
56.2 Contig5065.14507 (21 more)
57.4 Contig49365.88
16
0 Contig12729.16592 (2 more)
0.3 Contig71694.23340 (4 more)
0.6 Contig39266.15335 (20 more)
1.2 Contig42949.4328
1.5 Contig12547.6786 (28 more)
2.6 Contig97.139626 (11 more)
10.5 Contig98824.11547 (6 more)
10.6 Contig38586.133053
10.8 Contig1637.43510 (2 more)
11.8 Contig47666.74000 (5 more)
12.1 Contig38587.42611 (5 more)
12.7 Contig49481.24045 (1 more)
13.3 Contig31276.1811 (2 more)
14.2 Contig71.42061 (20 more)
14.5 Contig64386.26795
15.3 Contig40514.14669 (1 more)
15.5 Contig45610.29979
16.6 Contig1277.101967 (6 more)
17.8 Contig43203.1421 (11 more)
18.1 Contig45891.14077 (3 more)
18.4 Contig38034.34716 (7 more)
20.7 Contig11828.1074 (15 more)
21 Contig900.60201 (27 more)
21.3 Contig956.159804 (10 more)
21.9 Contig53270.91367 (4 more)
24.3 Contig637.748 (26 more)
24.6 Contig123.88496 (7 more)
24.9 Contig207.7509 (13 more)
25.6 Contig370.15003 (1 more)
30.6 Contig1498.57642 (23 more)
31.3 Contig893.8374 (11 more)
31.4 Contig5039.1541 (10 more)
31.6 Contig39278.85036
31.8 Contig66888.12222
31.9 Contig4801.17974 (29 more)
32.6 Contig43740.17391 (2 more)
33.3 Contig41587.39340 (3 more)
36 Contig1134.15274 (17 more)
37.6 Contig3086.8495 (15 more)
38 Contig4014.74028
38.1 Contig2041.16398 (5 more)
38.6 Contig6191.10952 (3 more)
39.2 Contig543.44302 (7 more)
40.6 Contig113.25029 (10 more)
41 Contig441.68242 (6 more)
41.2 Contig3254.16475 (4 more)
42.3 Contig40827.77037 (14 more)
42.9 Contig10616.19516 (5 more)
44 Contig73154.31579 (8 more)
44.7 Contig76994.20166 (2 more)
45.3 Contig39208.27503 (5 more)
45.9 Contig9511.4460 (18 more)
47 Contig42331.249206 (13 more)
48.4 Contig2509.5263 (18 more)
49.8 Contig54461.133232 (1 more)
51.1 Contig6271.5634 (6 more)
51.6 Contig57345.9856
52.4 Contig63796.45933
52.5 Contig40179.9492 (11 more)
55.1 Contig9263.50434 (5 more)
55.2 Contig122582.27453
55.7 Contig49578.15690 (6 more)
58 Contig56231.15810 (3 more)
59.1 Contig44836.10388 (10 more)
59.7 Contig53835.4592 (5 more)
60.3 Contig46114.207175 (2 more)
61.2 Contig46902.4380 (3 more)
61.7 Contig48299.49605 (6 more)
62.3 Contig1129.48275 (19 more)
17

Fig. 3.3 contd.: Sex-averaged maps 1.1

Table 3.2: Kendall tau correlation based on common SNPs between sex – averaged maps 0.0 and 1.1. The values shown here are absolute.



Linkage group	Kendall tau correlation
1	0.98
2	0.99
3	0.99
4	0.94
5	0.98
6	0.98
7	0.99
8	0.97
9	0.97
10	0.95
11	0.84
12	0.97
13	0.99
14	0.84
15	0.98
16	0.98
17	0.99

3.3.3 Integration with the genome assembly

The parent-averaged maps 1.1 were used to combine with the genome assembly (Panova, Larsson, et al. in prep.) using the Chromonomer package (Amores et al. 2014). In the current study, only 26.9% of the genome was covered by the linkage maps. The total length of the mapped contigs is estimated to be 432,705,506 bp. Chromonomer pruned 7,091 markers from the linkage map and kept 11,868 markers on 8,435 contigs for the integration. One contig on linkage group 8 was split.

Since the dataset for maps 1.1 was reduced considerably, to an average of 1.6 markers per contig, the information regarding the orientation of the contigs was not very great. Out of 8,435 contigs, 83 contigs were found to be in forward orientation, 87 in reverse orientation, and 8,265 in unknown orientation. In comparison, maps 0.0 have an average of 4.47 markers per contig. Therefore, Chromonomer, run with the default setting on the parent-averaged maps 0.0 resulted in the orientation of 3,782 contigs (1,995 forward orientation; 1,787 reverse orientation) out of 10,899 contigs (10,439 contigs were examined, out of which 446 contigs were split). The linkage maps 0.0 spanned 34.2% of the genome (table 3.3).

3.3.4 Comparison of 'home-script(s)' with the Chromonomer package

With linkage maps 0.0, the Chromonomer package reported 224 contigs mapping to multiple linkage groups, in contrast to 324 contigs as reported by our custom script. This difference in such contigs reported by the two approaches is due to the fact that Chromonomer filters the markers by quality prior to map comparison. Chromonomer had already pruned low-quality SNPs from the contigs that were not reported, and the remaining high-quality markers on those contigs mapped only to single linkage groups. Chromonomer detected 3,339 error contigs. In comparison, our custom script identified 5,194 error contigs (Figure 3.4). Out of these, 3,287 contigs were identified by both our custom script and Chromonomer while 1,907 contigs were identified exclusively by our custom script. For those 1,907 contigs, Chromonomer had already removed the SNPs that caused contigs to be flagged as error-prone in the pruning-by-quality step. Out of the 5,194 contigs that our script identified, 512 contigs were needed to be removed from the analysis. Chromonomer reported all these contigs except for 6. Of these 6 contigs not reported by Chromonomer, either the SNPs causing inconsistency were already removed, or the SNPs were placed physically far apart on the contig (> 1000 kb).

In comparison, Chromonomer reported no contigs mapping to multiple linkage groups for the linkage maps 1.1, as they already had been removed in the datasets used for the construction of maps 1.1. Chromonomer pruned 6,957 markers out 18,942 for quality and removed an additional 117 markers that caused inconsistencies, leaving 11,868 markers on 8,468 contigs in the linkage maps (table 3.3). The overall 'problem contigs' identified by Chromonomer package were 92. Our custom- script identified 342 contigs with problems, with only 2 contigs flagged to be removed. However, Chromonomer did not report these two contigs as these were already pruned in the first step.

Table 3.3: Summary of Chromonomer markers pruning and length of chromosomes after anchoring to the genome

Linkage group	Sex - Averaged Map 0.0				Sex- Averaged Map 1.1			
	No. of SNPs	SNPs after Chromonomer quality pruning	SNPs in the final Chromonomer map	Length of chromosomes (Mb)	No. of SNPs	SNPs after Chromonomer quality pruning	SNPs in the final Chromonomer map	Length of chromosomes (Mb)
1	7020	4428	4427	70.48	2716	1739	1725	57.41
2	5776	3579	3578	66.25	1913	1224	1221	52.97
3	4552	2899	2898	43.56	1485	970	964	34.45
4	4202	2678	2677	39.40	1529	984	943	31.83
5	3667	2310	2309	40.00	1215	765	760	30.76
6	3038	1897	1896	34.71	1092	674	672	26.64
7	3027	1922	1921	29.47	965	640	637	23.15
8	2990	1817	1816	20.96	892	527	520	15.78
9	2935	1818	1817	26.54	949	580	579	19.66
10	2859	1752	1751	24.67	883	565	554	19.42
11	2632	1598	1597	19.22	1112	679	662	16.04
12	2568	1639	1638	34.85	972	614	613	26.93
13	2464	1510	1509	24.34	828	509	509	18.64
14	2093	1318	1317	20.25	637	420	418	15.27
15	2098	1320	1319	21.24	683	432	432	16.36
16	1505	919	918	11.05	460	292	289	8.22
17	1473	871	870	24.94	611	371	370	19.18
Total	54899	34275	34258	551.94	18942	11985	11868	432.71

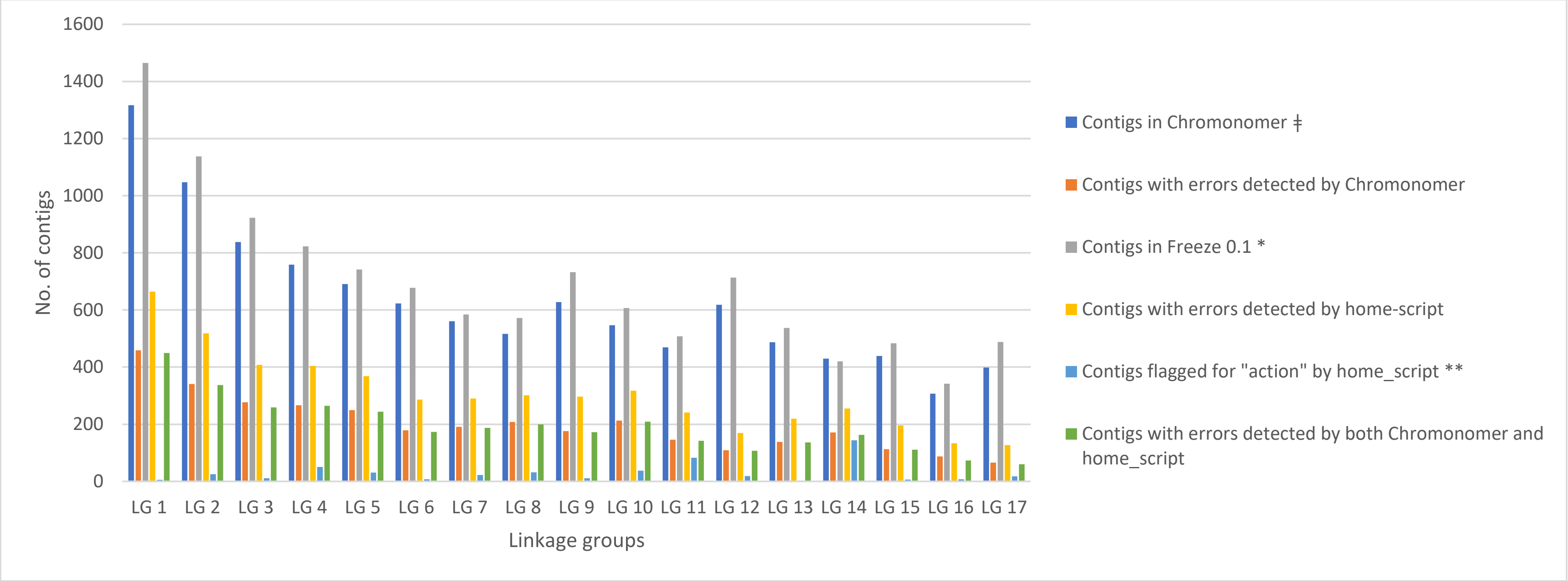


Fig. 3.4: The contigs identified by Chromonomer package and our custom script with physical and genetic map inconsistencies.

ǂ The number of contigs after the quality pruning and splitting of scaffolds for the contigs that mapped to multiple linkage groups

* The number of contigs after the removal of contigs that mapped to multiple linkage groups

** The contigs which need to be removed or have singular SNPs that need to be removed

3.3.3 Recombination rate estimates

A genomewide recombination rate (GwRR) of 0.75 cM/Mb was calculated based on the ratio of cumulative genetic map length of the parent-averaged map and the most recent *Littorina saxatilis* genome size estimate (1.35 Gb).

The corrected chromosomal recombination rates are given in table 3.4. We observe a strong positive linear correlation between the genetic map length and physical length for each linkage group (mapped contigs) with an intercept greater than zero (40.46), indicating a recombination event even for the smallest linkage group (Pearson correlation coefficient R= 0.83, p < 0.00003). This implies that smaller linkage groups will have more recombination per unit of physical distance and our data accordingly shows a negative correlation between recombination rate and physical length of the mapped contigs (Pearson correlation coefficient R= -0.71, p = 0.001; Figure 3.5). The correlation statistics presented here was calculated using the parent-averaged map 1.1 although similar correlation patterns were observed with the parent-specific maps.

The chromosomal recombination rates did not vary much across linkage groups and both the parents with few exceptions. For example, the male parent has much higher chromosomal recombination rate in linkage groups 8, 9,10,16 and 17. The female parent map has higher chromosomal recombination rate than the male parent for linkage group 12. Interestingly, Faria et al. (2019) has identified regions under inversion on all the mentioned linkage groups, except for linkage group 8. Chromosomal inversions were identified on other linkage groups too, yet it is easy to see that effects between the male and female parent map are more pronounced only in these smaller linkage groups. A plausible reason could be that due to local suppression of recombination by inversions, recombination increases in the rest of the freely recombining chromosome (Stevison et al. 2011) – an effect naturally more pronounced in smaller linkage groups. Between linkage groups, linkage group 16 has the highest chromosomal recombination rate (outlier in Figure 3.5b). This could perhaps be attributed to the fact that linkage group 16 is the smallest linkage group by the size of the total mapped contigs.

The intra-chromosomal recombination rates varied extensively (Figure 3.6). In most of the linkage groups, we observed low recombination in the centre. These regions were usually flanked with regions of high recombination. However, there are a few distinctive features that deviated from the above pattern. Some of these patterns were shared between the parents and some were parent-specific. For example, the regions between ~0-10 Mb and ~5-20 Mb on the linkage groups 1 and 11 respectively, denote regions of no recombination in both parents. Both these regions, however, have been identified as structural rearrangements by Faria et al. (2019). The parent-specific recombination patterns can also be observed. For example, the region in the beginning of the linkage group 14 possibly denotes a region of no recombination in the male-parent (Figure 3.6a) but not in the female-parent. This region has been identified with double inversions (Faria et al. 2019). Some other parent-specific features that may indicate higher recombination rates as well can also be observed. For example, in the male parent map, the patterns (the gaps in the black lines and the steep peaks by the red line; Figure 3.6a) on linkage group 16 and 17 indicate high recombination rates. Such pattern is not observed on the female parent map. Such patterns may be the artefact of our fragmented data. However, taking into account the chromosomal recombination rates, male parent show much higher recombination than the female parent in linkage groups 16 and 17, suggesting that presence of recombination hotspots in the male parent may also be plausible.

In our analysis, we observed such regions of reduced recombination on linkage groups 1,4,10,11,12 and 14, all of which corresponded with the regions Faria et al. (2019) identified as inversions. However, Faria et al. (2019) also described regions of inversion on linkage groups 2,6,7,9 and 17. But in the latter rearrangements, our data does not indicate recombination suppression. The inversions on linkage groups 6 and 17 were further characterized as homozygous (Faria et al. 2019; Westram et al. 2018), which do not have the same effect on recombination as heterozygous inversions (Kirkpatrick 2010). However, the difference here lies in the size of the inversions; the former inversions that we can also see in our data, have greater physical length by the mapped contigs. This suggests the limitation of our analysis to detect the subtleties in local recombination estimates. For example, the mean local recombination estimates in the ~10 Mb inverted region on linkage group 1 (between 0-10 Mb) is 0.1 cM/Mb and is represented by flat horizontal data points in black and a flat red line in Figure 3.6. However, Faria et al. (2019) also detected inversion of ~1 Mb on the same linkage group in the region ~64-65 Mb. But we were not able to detect that, as evident by the mean local recombination rate estimation within that region, that is 1.6 cM/Mb.

Table 3.4: Chromosomal recombination rates (cM/Mb)

Linkage group	Female map 1.1	Male map 1.1	Sex-averaged map 1.1.
1	0.58	0.59	0.47
2	0.82	0.59	0.55
3	0.92	0.78	0.65
4	0.76	0.73	0.58
5	0.77	0.63	0.56
6	0.95	0.92	0.72
7	0.94	0.83	0.72
8	1.15	1.40	1.05
9	1.10	1.38	0.82
10	0.84	1.15	0.73
11	1.40	1.34	1.19
12	1.51	0.75	0.68
13	1.06	1.18	0.86
14	1.66	1.08	0.99
15	1.37	1.43	1.07
16	2.78	3.00	2.11
17	1.05	2.43	0.92

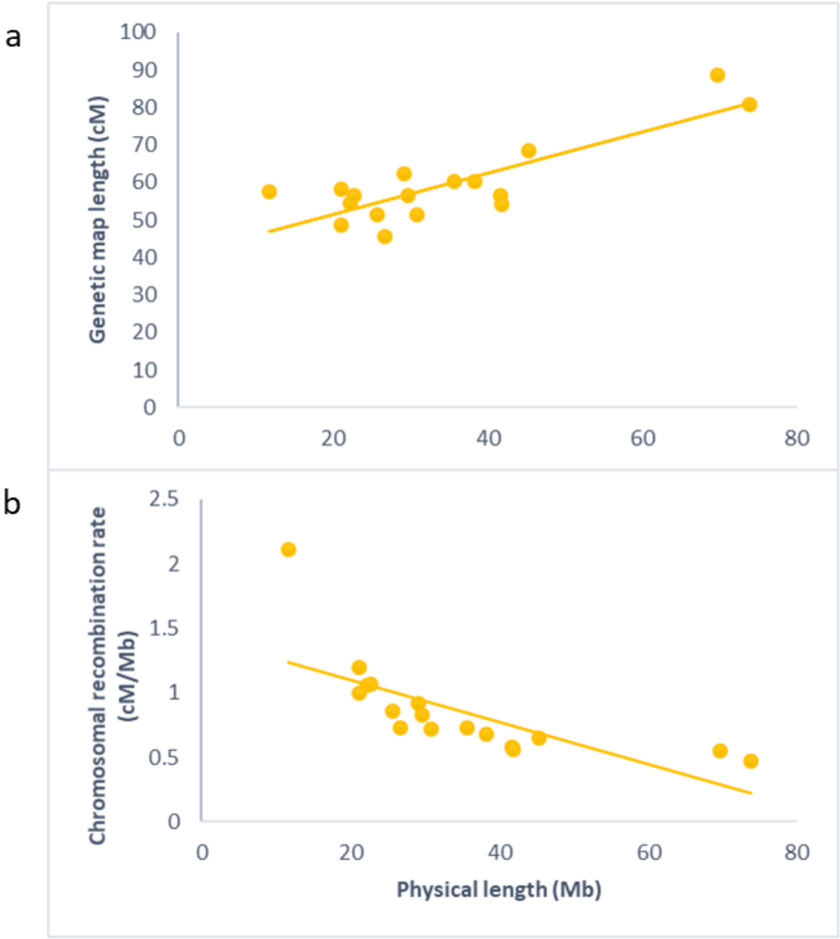


Fig 3.5: Relationship between the estimated physical length of chromosomes and (**a)** genetic length, and **(b)** chromosomal recombination rate. Sex-averaged maps 1.1 were used for this plot.

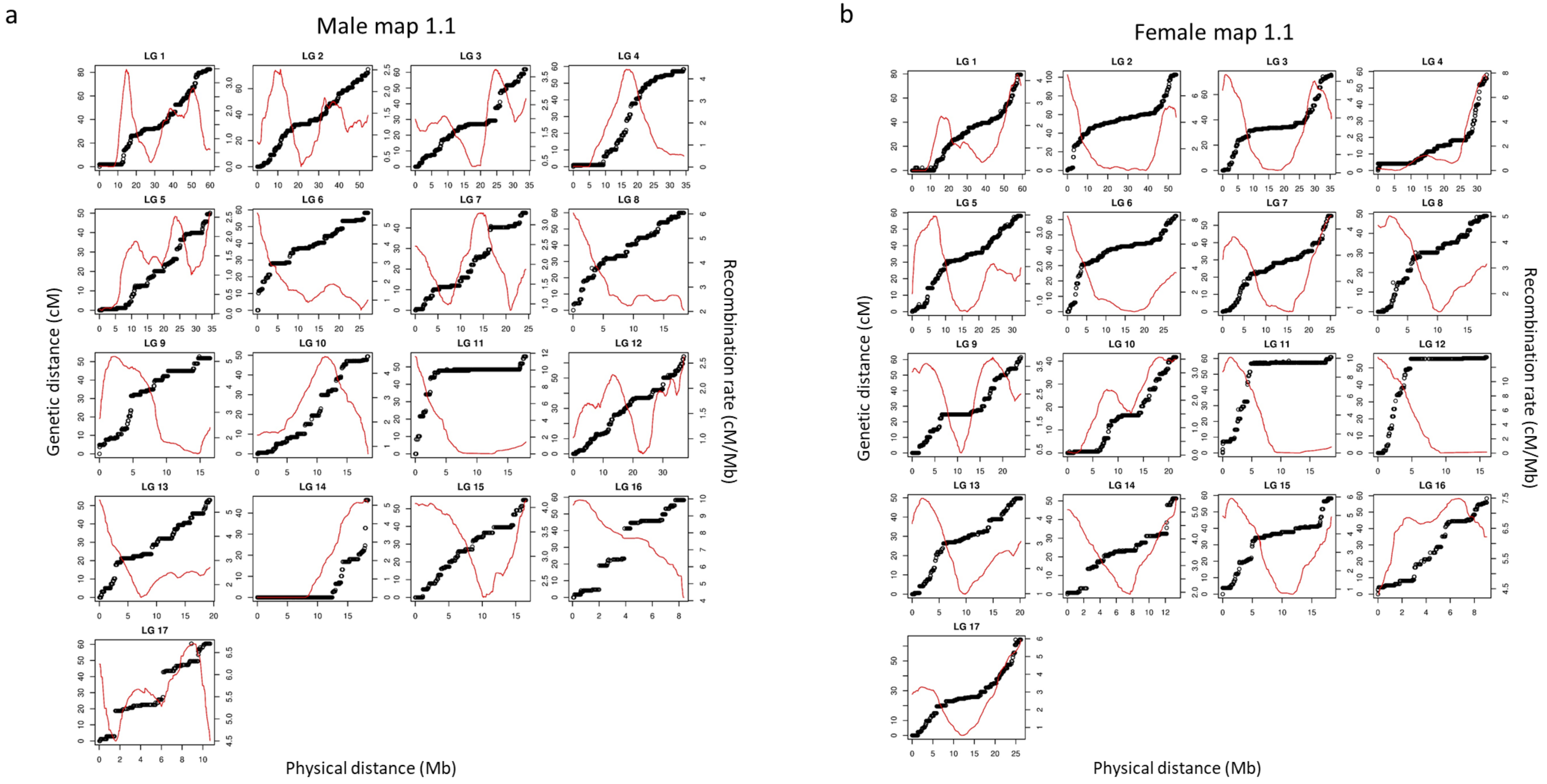


Fig 3.6: Recombination rate along the 17 chromosomes of *Littorina saxatilis* in the **(a)** male and **(b)** female map. Genetic positions of markers (on y-axis) are plotted against the estimated physical length of the chromosomes (on x-axis) and indicated by black dots. The red lines indicate the local recombination rate estimates (cM/Mb), smoothed using linear regression in the sliding-window sizes of 9 Mb (see text).

3.3.5 Sex determination in *Littorina saxatilis*

The ratio of the female-parent to male-parent map length across the linkage groups were found to be 1.05:1 (paired t-test, t = -1.2844, p-value=0.2713), indicating no significant symptoms of recombination suppression as expected in species with heteromorphic sex chromosomes. Please note, that this analysis included all the linkage groups, which means that there would be some effect from the individual inverted regions in the result. In addition, recombination rate estimates did not either suggest recombination suppression in any parent or linkage group, although some patterns are observable that are specific to only one parent. For example, on linkage group 12, region between 5-17 Mb indicate recombination suppression in the female-parent map (Figure 3.6b). Faria et al. (2019) has identified the genetic map positions associated with this region as chromosomal inversion. This region on linkage group 12 is of particular interest as in a separate preliminary analysis (unpublished), association mapping with sex as the response trait has indicated sex-linked loci to be on this linkage group. If true, this possibly may indicate presence of sex-determining loci in this region.

3.4 Discussion

The construction of linkage maps for *Littorina saxatilis* is described in this chapter. We developed an approach to address the limitations of building and merging of linkage maps. One indication of mapping inaccuracy is disagreement between assembly and map. (There are other indications of problems, of course, like inflated map estimations.) The decision when to trust the genome assembly and when to trust the linkage map, is an intricate one. Our script attempted to resolve the physical and genetic map inconsistencies in order within a linkage group by utilizing the information provided by the local genome assembly. Similarly, to address the problems manifested by the merging of the sex- specific maps, we reduced our datasets to include more common markers. By utilizing this approach, we were able to bring down the error contigs/total contigs ratio from 0.46 in the sex- averaged linkage maps 0.0 to 0.029 in the sex- averaged linkage maps 1.1.

The Chromonomer package also attempts to resolve the physical and genetic map inconsistencies by comparing the genetic map with the local genome assembly at the fine-scale resolution. It does so by checking whether (i) the SNPs are in a same linear order on the map as in the assembly, and, (ii) the physical distance between the two adjacent SNPs is large enough for recombination to take place between them. Our script does not account for the physical positions of the SNPs (beyond being in the same contig), which caused the discrepancy between the results obtained by our results and Chromonomer. Arguably this information is not relevant, except for the long contigs. This is reflected in the performance of our script on our data, which performed as well as Chromonomer to detect the inconsistencies between the map and the assembly. Additionally, Chromonomer also takes into account the quality of the markers even before parsing the linkage maps and the genome agp file. To do this, Chromonomer firstly parses the SAM file with the markers aligned to the genome assembly and removes the markers that (i) have a low mapping quality, (ii) are unmapped, (iii) are secondary or supplementary alignments or (iv) are excessively soft-clipped primary alignments. We did control for marker quality in the analysis, through filtering during the bioinformatic pipeline and at later steps during the map construction process. However, our process did not specifically remove the excessively soft-clipped primary alignments, which were the only ones filtered by Chromonomer during the quality filtering step.

Construction of linkage maps allowed us an inadvertent peek into the recombination landscape of this organism. Recombination rates may have environmental or demographic determinants, or can be genetic and heritable, and may respond to selection (Stapley et al. 2017). They may vary in a manner that corresponds to the evolutionary or selective history of the genome (Stapley et al. 2017). Thus, characterization of recombination rate variation and determinants driving the recombination landscape may be crucial to understanding genome evolution. In the current analysis, however, we focused primarily on the characterization of the recombination rates at the genome-wide and the chromosomal level. We lacked the resolution to study fine-scale intra-chromosomal patterns. However, in the large sliding-windows, we were able to discern broad recombination patterns within the linkage groups. In addition, our data is fragmented and does not span the entire genome. This indicated that our estimations of chromosomal and intra-chromosomal recombination rates perhaps may be overestimates. While, we corrected for the chromosomal recombination rates to allow for comparison across different linkage groups, the correction for local recombination estimates was more complex and not performed, implying that these recombination rates are perhaps overestimates. Moreover, recombination rates vary within and between chromosomes, individuals, sexes and populations (Stapley et al. 2017). Therefore, it is important to point out that the recombination landscape that we have obtained in this analysis, is observed within a family with two sexed individuals. Hence, caution must be taken in the interpretation and generalization of the results from this analysis to an entire population. Therefore, to have a thorough understanding of the recombination landscape in this organism, that can be extrapolated to an entire population, it is recommended to repeat this analysis with data from large full-sib families, with markers that cover a larger proportion of the genome.

We observed a GwRR of 0.75 cM/Mb, which is low but within the observed range for invertebrates (Wilfert et al. 2007). This is also comparable to the recombination rate estimate present from another gastropod *Biomphalaria glabrata* (GwRR = 0.8 cM/Mb; Tennesson et al. 2017). In addition, this estimate may also be comparable to other animal taxa with a similar genome size (Jensen-Seaman et al. 2004), consistent with the hypothesis that larger genomes have reduced recombination (Lynch 2005). In addition, it has been consistently demonstrated that chromosomal recombination rate correlates negatively with the chromosomal length due to the need of at least one recombination event per chromosome, and the reduced interference on large chromosomes (Kaback et al. 1999; Lynch 2005; Pessia et al. 2012; Stapley et al. 2017); a pattern repeated in our estimate of chromosomal recombination rates.

Recombination rates at a local scale may vary with gene density, nucleotide composition and structural variants (Stapley et al. 2017). While we lacked the resolution/data to test the effect of the former two determinants, we were able to see the putative effects of the known rearrangements (Faria et al. 2019) on recombination. A key evolutionary effect of heterozygous inversion is recombination suppression (Kirkpatrick 2010). We observed suppressed recombination rates within these rearrangements (Faria et al. 2019) that were at least ~9 Mb in size by the mapped contigs. However, we were not able to detect smaller inversions due to our coarse resolution.

The estimation of local recombination rates also gave us a unique insight into the sex-determination of this species. Rolán-Alvarez et al. (1996) concluded that sex determination in *Littorina saxatilis* (from the Iberian Peninsula) is the result of heteromorphic sex chromosomes (XY male). However, this result was not found in other studies from the samples collected from English and Swedish populations (Janson 1983) or from Russian populations (Birstein and Mikhailova 1990). In a separate study, sex-linked loci were found to be associated with linkage group 12 in a Swedish population (unpublished). Recombination rates point to a small region on linkage group 12 in the female parent with suppressed recombination. Interestingly, the same region has been identified as an inversion in the Faria et al. (2019) study. Inversions may capture loci that are sex-determining and/or are under sex-antagonistic selection and suppress recombination in the heterokaryotypic sex (Kirkpatrick 2010). This may imply that this region on linkage group 12 may be involved in sex-determination with the female being the heteromorphic sex. However, at the current stage, it may be naïve to make any conclusive remark and studies investigating sex-determination in this system are currently going on.

The parent-averaged linkage map described in this chapter is utilized in Chapter 4 to identify the genomic regions underlying locally adaptive traits in this species. It has already been critical in revealing the genomic landscape of differentiation (Westram et al. 2018) and in detecting inversions (Faria et al. 2019). It also underpins the heritability estimates of the complex traits in this species (Westram et al. 2018). It is further hoped that it will generally be invaluable for mapping quantitative trait loci by either association analysis or controlled crosses in the future studies. The parent-specific maps may find uses in the ongoing studies to document sex-specific differences, although caution is recommended.

3.5 Conclusion

In this chapter, we described the linkage maps for *Littorina saxatilis* and explained our methodology to overcome the common pitfalls that are faced in the construction of linkage maps. We showed our two- step approach to correct linkage maps at a fine-resolution to be quite effective with our dataset. We anchored our linkage maps with the genome assembly and were able to successfully orient a few scaffolds/contigs.

The linkage maps not only allowed us to inform the genome assembly, they also provided a unique opportunity to discern recombination patterns. Although, fine-scale recombination patterns cannot be discerned with our methodology. And for various reasons discussed above, this latter analysis should be treated with caution.

The parent-averaged maps are available via Dryad at https://doi.org/10.5061/dryad.bp25b65. The parent-specific maps are available on request. All the files from the linkage maps-genome integration and local recombination rate estimations are available on request. Custom scripts are available on GitHub at https://github.com/AnjaWestram.

Contributions

I performed the filtration of the SNPs in accordance with the analysis and constructed the initial and final linkage maps using the Lep-Map2 pipeline. I also wrote the script to find the inconsistencies in the map, used Chromonomer to anchor the final map to the genome assembly and tested for order stability between different methods. This work, however, is the product of collaboration and I would like to acknowledge and thank the following people for their contribution(s).

Prof. Kerstin Johannesson (University of Gothenburg) crossed the male and female of the Crab ecotype and reared the offspring and performed DNA extraction. Probe design, library preparation and sequencing were performed by RapidGenomics (Gainsville FL, United States). Bioinformatic pipeline and variant calling were done by Dr. A.M. Westram (IST, Austria). Prof. R. K. Butlin provided the scripts to generate parent-specific reduced datasets and merge parent-specific maps to produce parent-averaged maps.

References

Amores, A., Catchen, J., Nanda, I., Warren, W., Walter, R., Schartl, M., Postlethwait, J. H. (2014). A RAD-tag genetic map for the platyfish (X*iphophorus maculatus*) reveals mechanisms of karyotype evolution among teleost fish. Genetics; 197(2): 625–641.

Bansal, V., Bashir, A., Bafna, V. (2007). Evidence for large inversion polymorphisms in the human genome from HapMap data. Genome Research; 17(2):219–230.

Beavis, W.D., Grant, D. (1991). A linkage map based on information from four F2 populations of maize (*Zea mays L*.). Theoretical and Applied Genetics; 82:636-644.

Birstein, J., Mikhailovna., A. (1990). On the karyology of trematodes of the genus Microphallus and their intermediate gastropod host, *Littorina saxatilis* 11. Karyological study of *Littorinu saxatilis* (Gastropoda: Prosobranchia). Genetica; 80:167-170.

Bolger, A. M., Lohse, M., Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics; 30(15): 2114–2120.

Cartwright, D. A., Troggio, M., Velasco, R., Gutin, A. (2007). Genetic mapping in the presence of genotyping errors. Genetics; 176:2521–2527.

Chakravarti, A. (1991). A graphical representation of genetic and physical maps: the Marey map. Genomics; 11(1): 219-22.

Cheema, J., Dicks, J. (2009). Computational approaches and software tools for genetic linkage map estimation in plants. Briefings in Bioinformatics; 10: 595–608.

Endelman, J.B., Plomion, C. (2014). LPmerge: An R package for merging genetic maps by linear programming. Bioinformatics; 30(11):1623-1624.

Faria, R., Chaube, P., Morales, H., Larsson, T., Lemmon, A., Lemmon, E., Rafajlovic, M., Panova, M., Ravinet, M., Johannesson, K., Westram, A., Butlin, R. (2019). Multiple chromosomal rearrangements in a hybrid zone between *Littorina saxatilis* ecotypes. Molecular Ecology. Accepted author manuscript.

Fierst, J. L. (2015). Using linkage maps to correct and scaffold de novo genome assemblies: methods, challenges, and computational tools. Frontiers in Genetics; 6:220.

Galindo, J., Grahame, J.W., Butlin, R.K. (2010). An EST-based genome scan using 454 sequencing in the marine snail *Littorina saxatilis*. Journal of Evolutionary Biology; 23:2004–2016.

Janson, K. (1983). Chromosome number in two phenotypically distinct populations of *Littorina saxatilis* Olivi, and in specimens of the *Littorinu obtusatu* (L.). Journal of Molluscan Studies; 49:224-227.

Jensen-Seaman, M. I., Furey, T. S., Payseur, B. A., Lu, Y., Roskin, K. M., Chen, C.-F., … Jacob, H. J. (2004). Comparative Recombination Rates in the Rat, Mouse, and Human Genomes. Genome Research; 14(4):528–538.

Johannesson, K., Panova, M., Kemppainen, P., André, C., Rolán-Alvarez, E., Butlin, R. K. (2010). Repeated evolution of reproductive isolation in a marine snail: unveiling mechanisms of speciation. Philosophical Transactions of the Royal Society B: Biological Sciences, 365(1547), 1735–1747.

Kaback, D. B., Barber, D., Mahon, J., Lamb, J., You, J. (1999). Chromosome size-dependent control of meiotic reciprocal recombination in *Saccharomyces cerevisiae*: the role of crossover interference. Genetics; 152(4):1475–1486

Kawakami, T., Smeds, L., Backström, N., Husby, A., Qvarnström, A., Mugal, C. F., Ellegren, H. (2014). A high-density linkage map enables a second-generation collared flycatcher genome assembly and reveals the patterns of avian recombination rate variation and chromosomal evolution. Molecular Ecology; 23(16):4035–4058.

Kirkpatrick, M. (2010). How and why chromosome inversions evolve. PLOS Biology; 8(9): e1000501.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics; 25(16):2078–2079.

Lynch, M. (2006). The origins of eukaryotic gene structure. Molecular Biology & Evolution; 23:450–68.

Margarido, G.R.A., Souza, A.P., Garcia, A.A.F. (2007). Onemap: software for genetic mapping in outcrossing species. Hereditas; 144:78–79.

Mollinari, M., Margarido, G.R.A., Vencovsky, R., Garcia, A.A.F. (2009). Evaluation of algorithms used to order markers on genetic maps. Heredity; 103:494–502.

Morgan, Thomas H. (1911). Random segregation versus coupling in Mendelian inheritance. Science; 34(873): 384.

Ouellette, L.A., Reid, R.W., Blanchard, S.G., Brouwer, C.R. (2017). LinkageMapView—rendering high-resolution linkage and QTL maps. Bioinformatics; 34(2): 306–307.

Panova, M., Aronsson, H., Cameron, R. Dahl, P., Godhe, A., Lind, U., Ortega-Martinez, O., Pereyra, R., Tesson, S., Wrange, A., Blomberg, A. & Johannesson, K. (2016). DNA Extraction Protocols for Whole-Genome Sequencing in Marine Organisms. Methods in molecular biology, Clifton, N.J.: Springer, 13-44.

Panova, M., Larsson, T., Alm Rosenblad, M., Chaube, P., Westram, A.M., Butlin, R.K., Blomberg, A., Johannesson, K. (in prep.) Insights into local adaptation from the genome of *Littorina saxatilis* (Mollusca: Gastropoda).

Papa, R., Kapan, D. D., Counterman, B. A., Maldonado, K., Lindstrom, D. P., Reed, R. D., … McMillan, W. O. (2013). Multi-Allelic Major Effect Genes Interact with Minor Effect QTLs to Control Adaptive Color Pattern Variation in *Heliconius erato*. PLOS ONE; 8(3): e57033.

Pessia, E., Popa, A., Mousset, S., Rezvoy, C., Duret, L., Marais, G. A. B. (2012). Evidence for widespread GC-biased gene conversion in eukaryotes. Genome Biology and Evolution; 4(7): 675–682.

Rastas, P., Paulin, L., Hanski, I., Lehtonen, R., Auvinen, P. (2013). Lep-MAP: fast and accurate linkage map construction for large SNP datasets. Bioinformatics; 29(24): 3128–3134.

Rastas, P., Calboli, F. C. F., Guo, B., Shikano, T., Merilä, J. (2016). Construction of Ultradense Linkage Maps with Lep-MAP2: Stickleback F2 Recombinant Crosses as an Example. Genome Biology and Evolution; 8(1): 78–93.

Ravinet, M., Westram, A., Johannesson, K., Butlin, R., André, C., Panova, M. (2016). Shared and non-shared genomic divergence in parallel ecotypes of *Littorina saxatilis* at a local scale. Molecular Ecology; 25: 287–305.

Rezvoy, C., Charif, D., Gueguen, L., Marais, G.A. (2007). MareyMap: an R-based tool with graphical interface for estimating recombination rates. Bioinformatics; 23(16): 2188–2189.

Rolán-Alvarez, E., Buño, I., Gosálvez, J. (1996). Sex is determined by sex chromosomes in *Littorina saxatilis* (Olivi) (Gastropoda: Prosobranchia). Hereditas, (124): 261-267.

Schiffthaler, B., Bernhardsson, C., Ingvarsson, P. K., Street, N. R. (2017). BatchMap: A parallel implementation of the OneMap R package for fast computation of F1 linkage maps in outcrossing species. PLOS ONE; 12(12): e0189256.

Siberchicot, A., Bessy, A., Guéguen, L., Marais, G. (2017). MareyMap Online: A User-Friendly Web Application and Database Service for Estimating Recombination Rates Using Physical and Genetic Maps. Genome Biology and Evolution; 9(10): 2506-2509.

Stam, P. (1993). Construction of integrated genetic linkage maps by means of a new computer package: Join Map. The Plant Journal; 3(5):739-744.

Stapley, J., Feulner, P., Johnston, S. E., Santure, A. W., Smadja, C. M. (2017). Variation in recombination frequency and distribution across eukaryotes: patterns and processes. Philosophical transactions of the Royal Society of London. Series B, Biological sciences; 372(1736): 20160455.

Stevison, L.H., Hoehn, K.B., Noor, M.A.F. (2011). Effects of inversions on within- and between-species recombination and divergence. Genome Biology and Evolution; 3:830-841.

Sturtevant, A. H. (1913). The linear arrangement of six sex-linked factors in *Drosophila,* as shown by their mode of association. Journal of Experimental Zoology; 14: 43–59.

Tan, Y.D., Fornage, M. (2008). Mapping functions. Genetica; 133(3):235-46.

Tennessen, J. A., Bollmann, S. R., Blouin, M. S. (2017). A Targeted Capture Linkage Map Anchors the Genome of the Schistosomiasis Vector Snail, *Biomphalaria glabrata*. G3: Genes|Genomes|Genetics; 7(7): 2353–2361.

van Os, H., Stam, P., Visser, R.G., van Eck, H.J. (2005). SMOOTH: a statistical method for successful removal of genotyping errors from high-density genetic linkage data. Theoretical and Applied Genetics; 112(1):187-94

van Ooijen, J. (2011). Multipoint maximum likelihood mapping in a full-sib family of an outbreeding species. Genetics Research; 93:343–349.

Westram, A. M., Galindo, J., Alm Rosenblad, M., Grahame, J. W., Panova, M., Butlin, R. K. (2014). Do the same genes underlie parallel phenotypic divergence in different *Littorina saxatilis* populations? Molecular Ecology; 23(18): 4603–4616.

Westram, A.M., Panova, M., Galindo, J., Butlin, R. K. (2016), Targeted resequencing reveals geographical patterns of differentiation for loci implicated in parallel evolution. Molecular Ecology; 25: 3169-3186.

Westram, A.M., Rafajlović, M., Chaube, P., Faria, R., Larsson, T., Panova, M., Ravinet, M., Blomberg, A., Mehlig, B., Johannesson, K., Butlin, R. (2018). Clines on the seashore: the genomic architecture underlying rapid divergence in the face of gene flow. Evolution Letters; 2(4): 297-309.

Wilfert L., Gadau J., Schmid-Hempel P. (2007). Variation in genomic recombination rates among animal taxa and the case of social insects. Heredity; 98(4):189-197.

Wu, Y., Bhat, P.R., Close, T.J., Lonardi, S. (2008). Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. PLOS Genetics; 4:e1000212.

Chapter 4: Association mapping

4.1 Introduction

Whether natural populations would diversify in a heterogenous environment is central to the understanding of the processes underlying the generation and maintenance of biodiversity. Considered to be the first step in ecological speciation, local adaptation is a key process that promotes adaptive divergence within a species. Local adaptation leads to the evolution of advantageous phenotypes under the locally-selective environments (Nosil 2012). This is facilitated by divergent selection, a spatially varying force exerted by the environmental gradients. While, numerous studies have been conducted on the interaction between the selective forces and the adaptive phenotypes (e.g., Kingsolver et al. 2001; Reimchen and Nosil 2002; Mullen and Hoekstra 2008), only a little is known about the genetic aspect of this phenomenon. The interplay between the selection on phenotypes and the genetics underlying phenotypes has been of interest for evolutionary biologists for decades. Theory suggests that the response to selection is determined by the genetics of the traits underlying selection (Haldane 1930; Lande 1979; Yeaman and Otto 2011). On one hand, the lack of genetic variation may act as a constraint (Lande 1979), on the other, certain genetic architectures may facilitate selection (Nosil 2012; Yeaman and Whitlock 2011). For instance, gene flow counteracts selection by breaking up the locally-favourable genotypic structures via recombination (Yeaman and Whitlock 2011). However, genetic architectures that cause reduced recombination may resist the homogenizing effect of gene flow (Yeaman and Whitlock 2011). Therefore, the knowledge of genetic architecture of the adaptive traits and its influence on the evolutionary processes is at the core of the study of local adaptation.

Consequently, there is a growing amount of literature that focuses on elucidating the genetic architecture of the locally adaptive traits (Nosil and Schluter 2011; Nosil 2012), and addresses unresolved questions such as, how many genes contribute to local adaptation, their relative effect sizes, whether the genes are coding or regulatory, if specific gene families are responsible for adaptive variation, how ecological or evolutionary processes influence the genetic variation, how genetic linkage or pleiotropy affect the phenotype under selection, etc. (Rundle and Nosil 2005; Stinchcombe and Hoekstra 2008). While bottom-up approaches, such as genome scans, have been used in the literature to estimate the genetic architecture underlying divergent selection and local adaptation (e.g., Westram et al. 2016), a more straightforward approach is genotype-phenotype associations when there is information about the adaptive fitness of the observable traits in different environments.

Colour polymorphisms are one of the most studied adaptive traits dispersed across different taxa including mammals, birds, reptiles and invertebrates (reviewed in McKinnon and Peirotti 2010). The colour polymorphism tends to show correlation with other ecologically-relevant traits (McKinnon and Peirotti 2010; Wellenreuther et al. 2014). Such correlation is commonly reported to result from genetic architectures that include pleiotropy, i.e. single locus governing colour polymorphism and the other trait; or, tight linkage, caused by phenomenon such as chromosomal rearrangement or coadaptation of alleles behaving as a supergene (reviewed in McKinnon and Peirotti 2010). This association of traits, thus, influences the habitat choice, dispersal and adaptation capability of a population or species; which in turn, influences their evolutionary dynamics and ecological success (Lozier et al. 2016). Consequently, the study of colour polymorphisms along with the ecological and genomic data may prove to be an excellent system to understand the evolutionary forces underlying local adaptation and speciation (Gray and McKinnon 2007). While frequency-dependent selection (e.g., Punzalan et al. 2005), heterozygote advantage (e.g., Vercken et al. 2010) or disruptive/divergent selection in morphs inhabiting heterogenous environments (e.g., Jiggins 2008) are some of the mechanisms suggested to be involved in the maintenance of colour polymorphisms, other strategies resulting in same fitness of colour morphs have also been reported (e.g., Roulin et al. 2003).

The study of snail shell colour has been of particular interest to evolutionary biologists. Not only do snail shells display myriad colour polymorphisms (which include both basal shell colours and patterns on shells); they have also been associated with a multitude of adaptive functions including salinity tolerance (Sokolova and Berger 2000), crypsis (Johannesson and Ekendahl 2002), shell strength (Rosin et al. 2013) and mate choice (Rolán-Alvarez and Ekendahl 1996). Colour polymorphisms are a product of either structural colouration or pigmentation (Williams 2017). While structural colouration, i.e., colouration caused by light reflected due to microscopic repetitive elements, is more common among butterflies and birds, presence of biological pigments is the most common cause of shell pigmentation in molluscs (Williams 2017). Several chemical studies undertaken over a wide variety of taxa have determined melanin, porphyrins, polyenes and bile pigments to be involved in shell pigmentation (reviewed in Williams 2017); however, it is safe to say that molluscan shell pigments largely remain unknown. Environmental factors, such as diet, have been reported to influence shell pigmentation (Williams 2017) but studies on several molluscs have determined a simple underlying genetic basis (Cain and Sheppard 1950; Innes and Haley 1977; Palmer 1985; Cook 1998; Richards et al. 2013; Ky et al. 2016).

Shell morphology, similarly, has been of interest for many. The variability in shell size and shape is perhaps the most evident trait. Like colour polymorphism, shell morphometric characters like size, shape, thickness and aperture size has been linked with heat tolerance, moisture, population density, preferential predation, mate choice and calcium metabolism (Goodfriend 1986). Shell morphology has been observed to be highly responsive to the environmental conditions (Giokas et al. 2014). While, phenotypic plasticity has been observed in the shell characteristics of some gastropods to adapt to the environmental gradients or pressure against predation (Appleton & Palmer 1988; Doyle et al. 2010), fixed genetic effects have also been observed (Dillon and Jacquemin 2015; Goodfriend 1986; Stankowski 2013). On one hand, this makes shell morphometric characteristics very desirable to study local adaptation (Giokas et al. 2014) and/or as biomarkers for human-induced disturbances, like pollution (Primost et al. 2015). On the other hand, this implies a complex pattern of genetics underlying these traits.

The marine snail *Littorina saxatilis* is considered to possess the most variable shell characteristics in the genus *Littorina* (Reid 1996). The shell colours show a wide range of ground colours from black and grey to dark to light brown. The conspicuous shell colours (white, red and yellow) and patterns (bands or tessellation) are present in a much lower frequency (Johannesson and Butlin 2017). The underlying chemistry and pathways remain largely elusive, as biochemistry of shell colouration in *Littorina saxatilis* has not been explored yet. However, since pigment types are distributed in a phylogenetically relevant manner (Williams 2017), studies by Comfort (1951) and Kozminskii and Lezin (2007) may suggest that pigmentation produces shell colouration in *Littorina*. Furthermore, Comfort (1951) proposed that the shell colour variation in *Littorina littorea* is due to the presence of melanoproteins. Kozminskii and Lezin (2007) analysed the calcareous structure of *Littorina obtusata* shell and performed chemical studies. They suggested that the background shell colouration in this species is produced by melanin (for brown-purple colouration), carotenoids (for yellow-orange colouration) and a fourth unknown pigment (for white colour). The interpolation of these studies to *Littorina saxatilis* may require some caution due to their phylogenetic distance [~15 mya with *L. littorea* and ~4 mya with *L. obtusata* (Reid et al. 2012)]. The colouration and patterns on the shell form phenotypic clines on microenvironmental gradients and show a correlation with the environmental factors (Ekendahl and Johannesson 1997), especially favouring the cryptic colouration (Johannesson and Ekendahl 2002; Johannesson & Butlin 2017). Although cross-breeding experiments have established the Mendelian inheritance of ground colours and banding patterns (Ekendahl and Johannesson 1997), little is known about the genes underlying the colour patterns. It has been suggested that the ground colours and the banding pattern are controlled by separate unlinked loci with two alleles, with dominant allele coding for the respective colour and band (Ekendahl and Johannesson 1997).

The two distinct ecotypes of *Littorina saxatilis* differ greatly in their shell morphometric characteristics (see Figure 1.1; Johannesson et al. 2010). The differences in their shell morphology are driven by the adaptive divergence caused by their local environments. The areas that are protected from the waves are inhabited by the “Crab ecotype”, characterized by thick, elongated shells with small aperture size and large adult size, which provide resistance against the selective pressure of the crab predation (Johannesson et al. 1993). The “Wave ecotype” inhabits the wave-exposed areas that lack predators but select for the traits that can resist the strong wave action (Johannesson et al. 1993). This ecotype is characterized by thinner, compressed shells with larger aperture size and small adult size (Johannesson et al. 1993). The transplant experiments between the inhabitants of the two distinct environmental sites indicate this divergence in the shell morphology between the ecotypes stems from the local ecological adaptation (Rolán-Alvarez et al. 1997). Furthermore, the shell size may be the ‘magic trait’ contributing to reproductive isolation through facilitating local adaptation and assortative mating (Boulding et al. 2017). In spite of the local habitat preference and assortative mating, there is an ongoing gene flow between the distinct ecotypes (Johannesson et al. 2010). The adaptive traits have evolved despite the ongoing gene flow in the narrow zones of contact between the two ecotypes. The intermediates of the two parental morphs can be observed in a continuum over the crab-wave environmental gradient (Janson and Sundberg 1983; Johannesson et al. 1993). Consistent with previous studies in Littorinids and other snails (Brookes and Rochette 2007; Giokas et al. 2014), the shell shape and size in *Littorina saxatilis* are complex traits partially influenced by their environment (Hollander and Butlin 2010; Saura et al. 2012). Although, some prior studies may have identified genes under selection that may be underlying shell morphological traits (Galindo et al. 2010; Westram et al. 2016), the genes controlling the shell morphology largely remain elusive. The doctoral study by Joseph Kess (2017), however, has exposed 137 loci that contribute to the shell shape and size variation in the Spanish populations of *Littorina saxatilis.*

Association mapping can potentially discern fine-scale genetic patterns underlying a phenotype. In this chapter, I attempted to exploit the power of association mapping to identify the genetic basis of the shell phenotypes in *Littorina saxatilis.* This approach is further facilitated by sampling in the hybrid zone of the two distinct ecotypes. The genotype-phenotype associations within the hybrid zones allow for higher mapping resolution due to generations of recombination between lineages. The dataset is described in Westram et al. (2018). Targeted-capture sequencing was used to generate the genotype dataset of codominant SNP markers. Random regions of the genome were captured, along with the regions of high differentiation as determined by the previous studies (Galindo et al. 2010; Westram et al. 2014; Ravinet et al. 2016; Westram et al. 2016). This methodology allowed us to sequence at higher coverage, improving the inference power for downstream analyses (Grover et al. 2012), and test for associations with the adaptive traits in the regions known to be under divergent selection. The adaptive shell phenotypes used for the analysis were shell size, shell shape, banding pattern and shell basal colours (beige, dark beige and black). The transcriptomic resources (Chapter 2) were utilized to annotate the associated region(s).

4.2 Methods

4.2.1 Sampling and Phenotyping

The dataset used in this analysis is described in Westram et al. 2018. In June 2013, 600 individuals were sampled at Ängklåvebukten ("ANG", 58.8697°, 11.1197°) on the Swedish west coast. The transect was ~150 m along the shore and ranged from the boulder fields to the cliff area (Figure 4.1).

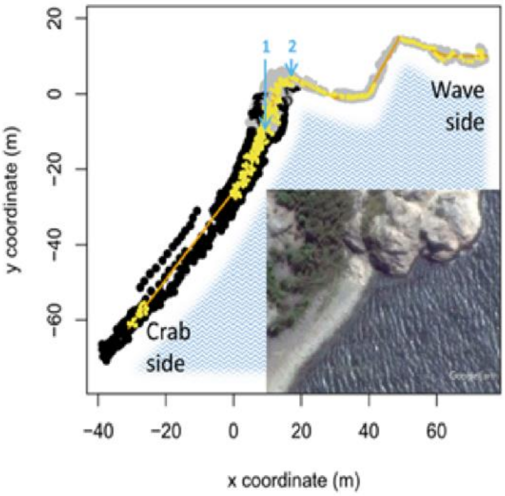


Fig. 4.1: Map of the sampled area. Black points represent the boulder area. Grey points represent the cliff area. The habitat transitions are represented by arrow 1: from boulder to the rock platform; and arrow 2: from rock platform to the cliff. Each yellow point represents the sampled snails. Image adapted from Westram et al. (2018).

The exact positions along the transect, sex and photographs of the shells were recorded for each snail before preserving their foot tissue in ethanol. The qualitative and quantitative phenotypes of the shell were scored using the photographs. The qualitative phenotypes consisted of banding pattern and the shell colours (black, dark beige and beige). These traits were scored binomially (0,1) for presence or absence. Precise scoring of some phenotypes is difficult; especially distinguishing between beige and dark beige can be difficult and subjective. Therefore, colour traits were scored twice by eye, by the same person. This ensured the same threshold to distinguish between the colours and the subjective bias was limited. The quantitative phenotypes were obtained using 15 landmarks to capture the variation in shell morphology (Carvajal-Rodriguez et al. 2005; Conde-Padín et al. 2007b). The methodology was the same as described in Ravinet et al. (2016).

4.2.2 DNA extraction, sequencing and Bioinformatics

The DNA was extracted from the foot tissue of 373 sexually-mature individuals using the CTAB protocol described in Panova et al. (2016). The same probe dataset for capture sequencing was used as the one used to produce the dataset for the linkage maps (see sections 1.2 and 3.2.1). The probe design, library preparation and sequencing were done by RapidGenomics (Gainesville FL, United States). For some individuals, both single-end and paired-end sequencing was done. The data were pooled together in the bioinformatic pipeline as described in Westram et al. (2018). The bioinformatics approach for cleaning the raw reads, mapping and variant calling are the same as described the previous chapter (see section 3.2.1) and in Westram et al. (2018). The contigs that mapped to multiple linkage groups in the Linkage map analysis were removed. The SNPs that were more than 1000 bp from the markers present on the linkage maps were also removed to eliminate the potentially duplicated genomic regions which tend to deviate from the normal segregation expectations. The number of SNPs which passed the filters was 106,599.

4.2.3 Imputation

Imputation of the missing genotypes was done by LinkImpute v1.1.1 (Money et al. 2015). LinkImpute does imputation based on LD-kNNI method; that is, the markers that show high LD across the genome with the marker with the missing genotype are used for imputation. The method is independent of physical or genetic maps and designed to work on unphased genotypes from heterozygous species (Money et al. 2015). However, with a big dataset, this method is time expensive. Therefore, to increase the time efficiency, the SNPs were clustered into different linkage groups based on the linkage maps prior to the imputation. The imputation for each linkage group was then done separately and in parallel by wrapping up the LinkImpute package in a custom bash script.

4.2.4 Association mapping

A principal component analysis was performed to assess the genetic variation along the transect using a custom R script and the proportion of variation explained by the first 100 PC-axes was plotted. Based on the scree plot, the first 4 PC-axes explained most of the genetic variation in our dataset.

Association mapping was performed using the R Package GenABEL v1.8-0 (Aulchenko et al. 2007). Allometry accounts for 58.8% variation in the shell (Conde-Padín et al. 2007b); while sex effects have been established in the determination of size (Johannesson and Johannesson 1996; Reid 1996). Therefore, for the traits shape and size, size and sex were used as covariates respectively. To correct for population stratification, the EIGENSTRAT method (Price et al. 2006) was used. The EIGENSTRAT method corrects for population structure at both normal and highly differentiated SNPs (Price et al. 2010) and works efficiently with the admixed populations (Price et al. 2006; http://www.genabel.org/sites/default/files/pdfs/GenABEL-tutorial.pdf). This method is incorporated within the GenABEL program as the egscore function, which linearly adjusts the genotype and phenotype onto PC axes by linear regression, and then performs the simple test score for association or Armitage's trend test in the case of qualitative traits (http://www.genabel.org/sites/default/files/pdfs/GenABEL-tutorial.pdf). The PC axes of variation were obtained by decomposing the genomic kinship matrix of the individuals. The first four PC axes, as previously analysed, explained the most genetic variation within the dataset, and hence were used for the EIGENSTRAT correction method. The built-in permutation test in the egscore function was used to permute the phenotype over genotype 1000 times and compute the association statistics to correct for multiple testing. The p-value < 0.05 was chosen to be the genome-wide threshold to be considered significant. The proportion of total phenotypic variance explained by each significant SNP was estimated by dividing the χ^2 test statistic by the number of samples (http://forum.genabel.org/viewtopic.php?f=6&t=760&p=1449&hilit=proportion+variance+explained#p1449). The heritabilities for the traits were estimated using HEIDI (Kostem & Eskin 2013) and the confidence limits were placed using the jackknife method.

4.2.5 Identification and functional annotation of the candidates

The transcriptome was used for the annotation of the regions in association with the phenotype. The annotated transcripts were aligned to the associated regions in the genome using Blat v.35 (Kent 2002).

4.3 Results

4.3.1 Variance within population

PCA analysis was conducted using the genomic kinship matrix to check the variation within the population. The first four PC axes cumulatively explained 10.48% of the variation with the PC axis 1 accounting for 6%. The first PC axis captured the variation along the shore line. The results are given in Figure 4.2.

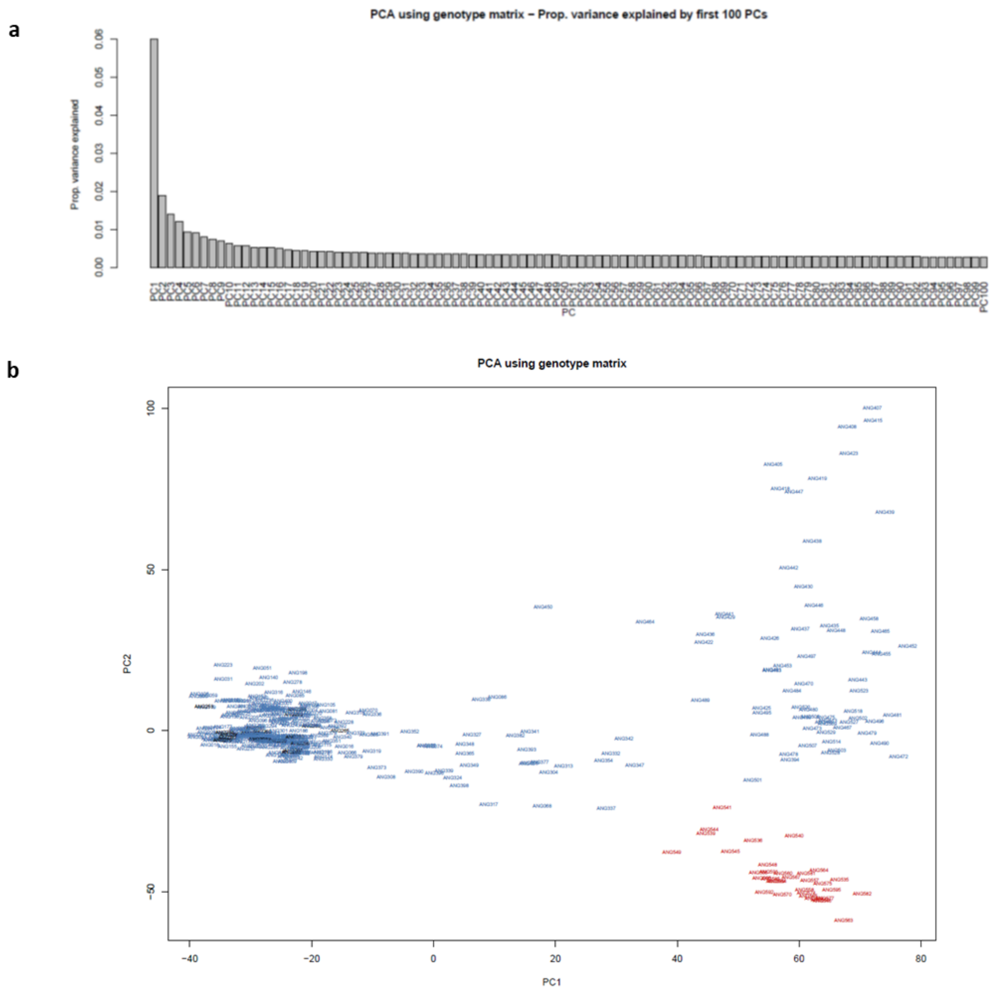


Fig. 4.2: PCA analysis. **(a)** Proportion of variance explained by the first-100 PC axes. **(b)** PC1 captured the most variation along the shore. The samples in black and red represent the extreme ends of the transect. The individuals in black are sampled from the boulder end before the sampling break and the individuals in red are sampled from the cliff end before the sampling break (see Figure 4.1).

4.3.2 Association mapping

Association mapping was performed with the GenABEL software to identify the genomic regions responsible for phenotypic variation for the following traits across the hybrid zone – shell size, shell shape, banding pattern on the shell and the ground shell colours (black, beige and dark beige). Among the qualitative phenotypes, significant associations were found for the ground colours, black and beige, and the banding pattern (Figure 4.3), while no significant association were found for the dark beige colour. The heritability was measured to be 0.61 (0.38-0.84) and 0.25 (0.19-0.30) for shape and size respectively. In spite of the high heritability, as estimated by HEIDI, no significant association were found for the shell shape and size.

For the colours black and beige, each of the significant associations was found to be linked to single locus, on the linkage groups 9 and 5, respectively. Faria et al. (2019) identified a chromosomal inversion on linkage group 6 (0 cM – 29.3 cM) and the banding pattern was found to be associated with a region at the boundary of this inversion (~29.30 cM) (Westram et al. 2018). SNPs on linkage groups 10, 11, 12 and 16 also showed significant association with the banding pattern. The table 4.1 summarizes the result for the significant associations.

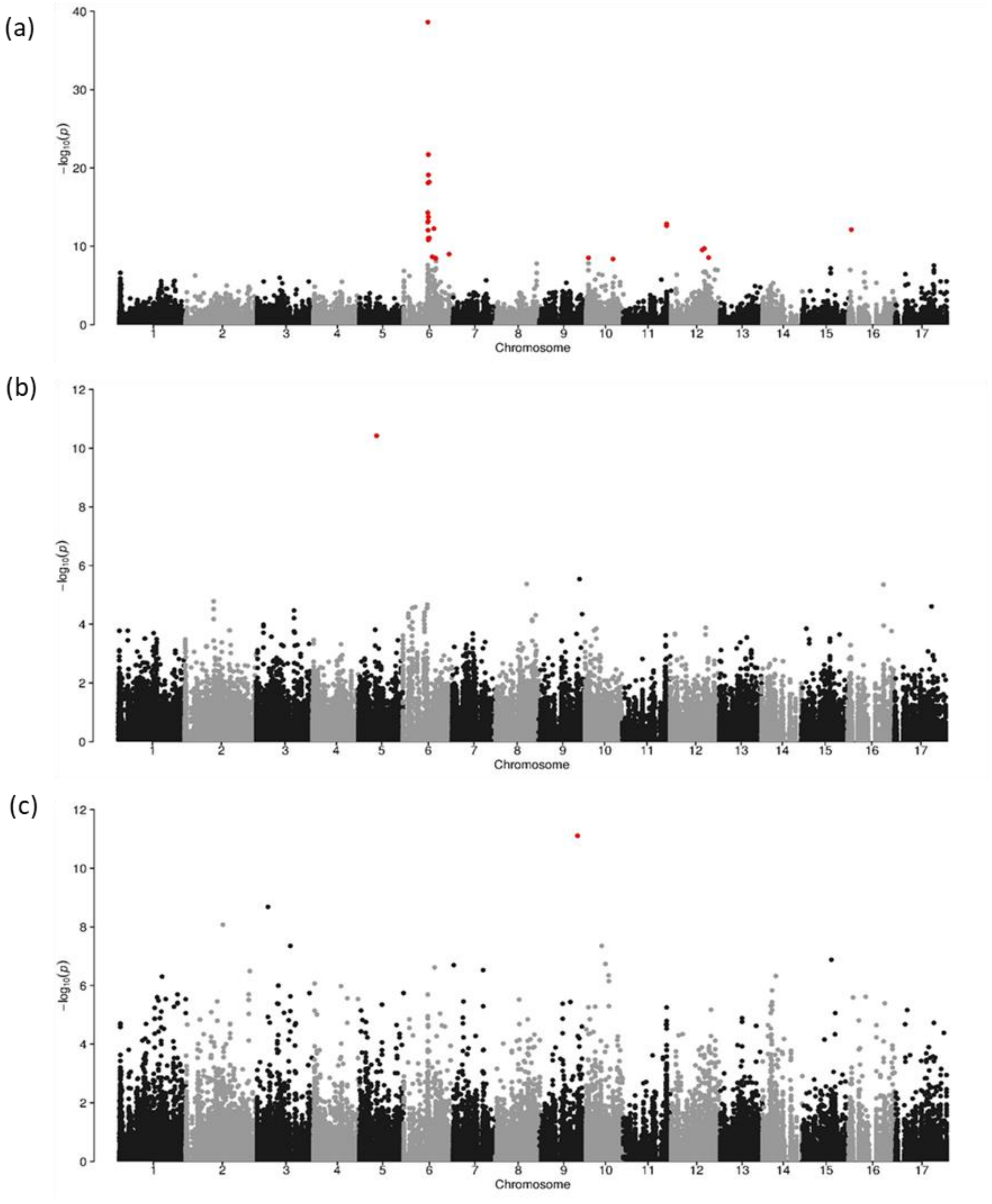
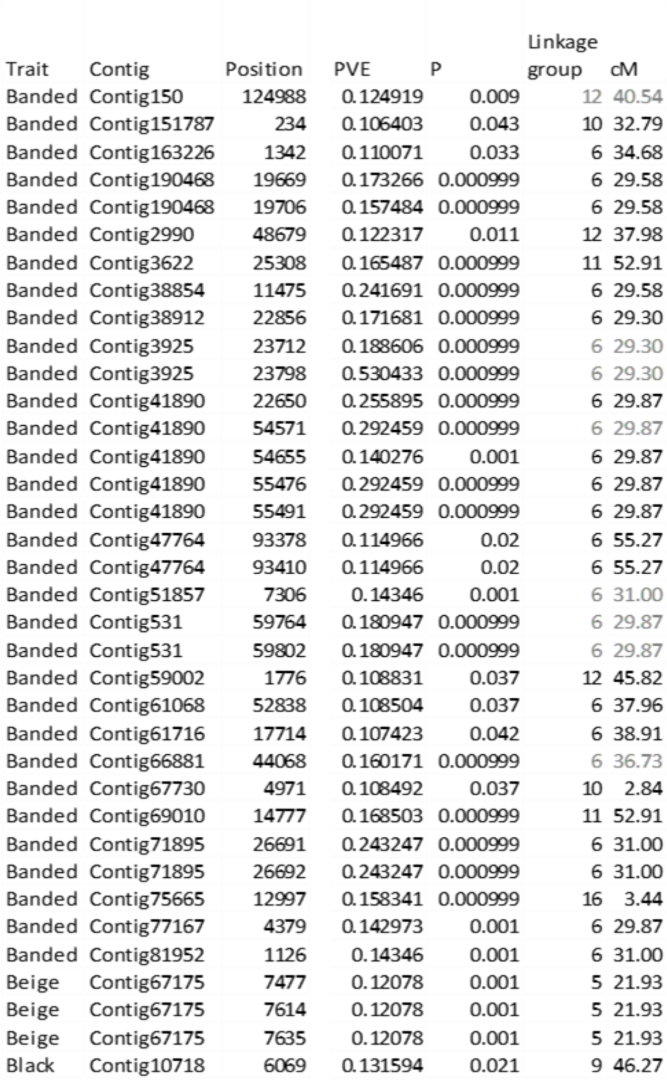


Fig. 4.3: Manhattan plots for the association mapping results. **(a)** Banded. **(b)** Beige. **(c)** Black. The significant associations, SNPs (p-value < 0.05) are represented by red points. Each SNP is positioned by their position on the parent-averaged linkage map. Image taken from Westram et al. (2018).

Table 4.1: SNPs significantly associated with shell colour polymorphism. For each significant SNP, position of the nearest SNP on the linkage map is given. Linkage group/position are highlighted in grey if the nearest SNP in map is more than 1000 bp away from the given colour-associated SNP. Adapted from Westram et al. (2018).



Trait	Contig	Position	PVE	P	Linkage group	cM
Banded	Contig150	124988	0.124919	0.009	12	40.54
Banded	Contig151787	234	0.106403	0.043	10	32.79
Banded	Contig163226	1342	0.110071	0.033	6	34.68
Banded	Contig190468	19669	0.173266	0.000999	6	29.58
Banded	Contig190468	19706	0.157484	0.000999	6	29.58
Banded	Contig2990	48679	0.122317	0.011	12	37.98
Banded	Contig3622	25308	0.165487	0.000999	11	52.91
Banded	Contig38854	11475	0.241691	0.000999	6	29.58
Banded	Contig38912	22856	0.171681	0.000999	6	29.30
Banded	Contig3925	23712	0.188606	0.000999	6	29.30
Banded	Contig3925	23798	0.530433	0.000999	6	29.30
Banded	Contig41890	22650	0.255895	0.000999	6	29.87
Banded	Contig41890	54571	0.292459	0.000999	6	29.87
Banded	Contig41890	54655	0.140276	0.001	6	29.87
Banded	Contig41890	55476	0.292459	0.000999	6	29.87
Banded	Contig41890	55491	0.292459	0.000999	6	29.87
Banded	Contig47764	93378	0.114966	0.02	6	55.27
Banded	Contig47764	93410	0.114966	0.02	6	55.27
Banded	Contig51857	7306	0.14346	0.001	6	31.00
Banded	Contig531	59764	0.180947	0.000999	6	29.87
Banded	Contig531	59802	0.180947	0.000999	6	29.87
Banded	Contig59002	1776	0.108831	0.037	12	45.82
Banded	Contig61068	52838	0.108504	0.037	6	37.96
Banded	Contig61716	17714	0.107423	0.042	6	38.91
Banded	Contig66881	44068	0.160171	0.000999	6	36.73
Banded	Contig67730	4971	0.108492	0.037	10	2.84
Banded	Contig69010	14777	0.168503	0.000999	11	52.91
Banded	Contig71895	26691	0.243247	0.000999	6	31.00
Banded	Contig71895	26692	0.243247	0.000999	6	31.00
Banded	Contig75665	12997	0.158341	0.000999	16	3.44
Banded	Contig77167	4379	0.142973	0.001	6	29.87
Banded	Contig81952	1126	0.14346	0.001	6	31.00
Beige	Contig67175	7477	0.12078	0.001	5	21.93
Beige	Contig67175	7614	0.12078	0.001	5	21.93
Beige	Contig67175	7635	0.12078	0.001	5	21.93
Black	Contig10718	6069	0.131594	0.021	9	46.27

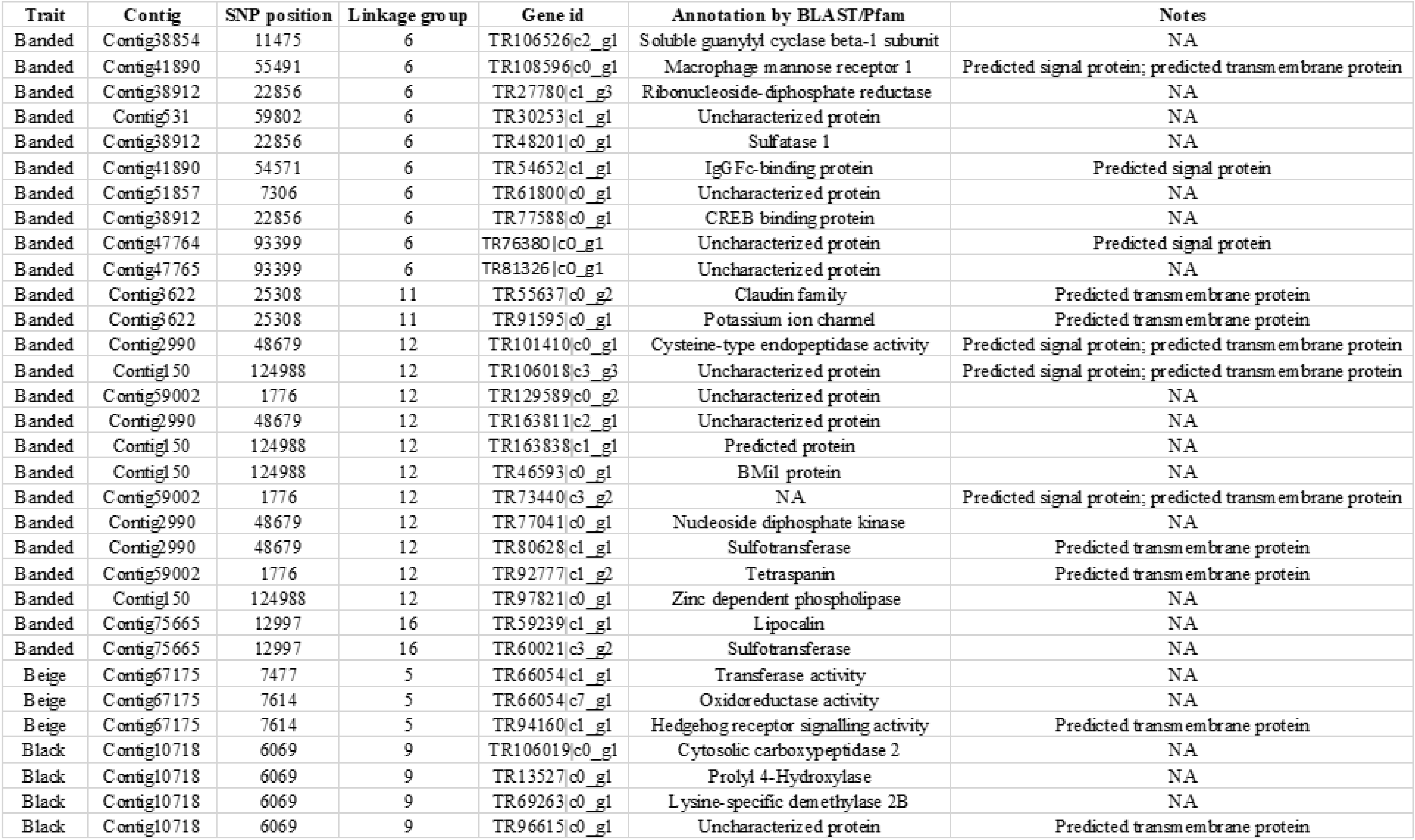
4.3.3 Annotation of the genomic regions in association with the phenotypes

The contigs with SNPs that showed significant association were annotated with the transcriptome (see table 4.2). The identification of specific genes underlying shell colouration and the biochemical pathways is still in the nascent state in molluscans (Williams 2017). And most of the genes underlying shell colouration in the molluscan literature have been found to be involved due to their role in shell formation (Williams 2017). Therefore, literature was searched to see specifically if the coding domains identified in this analysis have been implicated in either of the functions, in colouration or biomineralization, in other organisms and specifically molluscs.

The associations were on a single contig each for beige and black. The contig in association with black colour polymorphism has 18 genes, of which 4 were annotated. The contig in association with beige colour polymorphism has 15 genes, of which 3 had annotations. The association analysis showed significant association for the banding pattern on 23 contigs on different linkage groups that accounted for 25 annotated coding domains.

The contig in association with the black colour has four coding domains, three of which are annotated by Pfam and one of which is annotated by BLAST as an uncharacterized protein in the *Lottia* genome. The peptide, procollagen-proline 4-dioxygenase, has been extracted from the extrapallial fluid (the fluid in the space between the mantle and the shell) in the edible mussel (*Mytilus galloprovincialis*) (Calvo-Iglesias et al. 2017) but, as the authors cited, this peptide is not specific to the mantle tissue. The shell material is secreted by the epithelial cells of the mantle tissue, therefore mantle tissue-specific expression of genes provides support for their direct role in shell pigmentation or formation. Lysine-specific demethylase 2B peptide has been implicated in maternal genomic imprinting (Ciccone et al. 2009) and lipid metabolism (Nagaoka et al. 2015). It is hard to comment on its role in the shell formation or pigmentation. Although the carboxypeptidase peptide domains have been extracted from the protein in *svr* gene in *Drosophila melanogaster*, which has been shown to affect viability, pigmentation and wing formation (Wright 1987), without any further information regarding the protein in question in our study, no comments can be made on its role. Additionally, carboxypeptidases have been known to be involved in other functions in molluscs and may not be mantle tissue specific (Fan et al. 1999; David et al. 2005). Similarly, the coding domains Sulfatase 1, Sulfotransferase and Lipocalins, associated with the banding pattern, have been extracted from the shell matrix of other molluscs (Aguilera et al. 2017; Arivalagan et al. 2017; Wang et al. 2017). However, these proteins can have different functions in different parts of the body and their role in banding pattern cannot be determined at present. Additionally, the entire inverted region on linkage group 6, as determined by Faria et al (in press), was specifically searched for the mollusc shell-related proteins (Williams 2017; Arivalagan et al. 2017; McDougall and Degnan 2018). None were found.

Table 4.2: Annotation of the contigs with SNPs in association with the shell colour traits.



Trait	Contig	SNP position	Linkage group	Gene id	Annotation by BLAST/Pfam	Notes
Banded	Contig38854	11475	6	TR106526|c2_g1	Soluble guanylyl cyclase beta-1 subunit	NA
Banded	Contig41890	55491	6	TR108596|c0_g1	Macrophage mannose receptor 1	Predicted signal protein; predicted transmembrane protein
Banded	Contig38912	22856	6	TR27780|c1_g3	Ribonucleoside-diphosphate reductase	NA
Banded	Contig531	59802	6	TR30253|c1_g1	Uncharacterized protein	NA
Banded	Contig38912	22856	6	TR48201|c0_g1	Sulfatase 1	NA
Banded	Contig41890	54571	6	TR54652|c1_g1	IgGFc-binding protein	Predicted signal protein
Banded	Contig51857	7306	6	TR61800|c0_g1	Uncharacterized protein	NA
Banded	Contig38912	22856	6	TR77588|c0_g1	CREB binding protein	NA
Banded	Contig47764	93399	6	TR76380|c0_g1	Uncharacterized protein	Predicted signal protein
Banded	Contig47765	93399	6	TR81326|c0_g1	Uncharacterized protein	NA
Banded	Contig3622	25308	11	TR55637|c0_g2	Claudin family	Predicted transmembrane protein
Banded	Contig3622	25308	11	TR91595|c0_g1	Potassium ion channel	Predicted transmembrane protein
Banded	Contig2990	48679	12	TR101410|c0_g1	Cysteine-type endopeptidase activity	Predicted signal protein; predicted transmembrane protein
Banded	Contig150	124988	12	TR106018|c3_g3	Uncharacterized protein	Predicted signal protein; predicted transmembrane protein
Banded	Contig59002	1776	12	TR129589|c0_g2	Uncharacterized protein	NA
Banded	Contig2990	48679	12	TR163811|c2_g1	Uncharacterized protein	NA
Banded	Contig150	124988	12	TR163838|c1_g1	Predicted protein	NA
Banded	Contig150	124988	12	TR46593|c0_g1	BMi1 protein	NA
Banded	Contig59002	1776	12	TR73440|c3_g2	NA	Predicted signal protein; predicted transmembrane protein
Banded	Contig2990	48679	12	TR77041|c0_g1	Nucleoside diphosphate kinase	NA
Banded	Contig2990	48679	12	TR80628|c1_g1	Sulfotransferase	Predicted transmembrane protein
Banded	Contig59002	1776	12	TR92777|c1_g2	Tetraspanin	Predicted transmembrane protein
Banded	Contig150	124988	12	TR97821|c0_g1	Zinc dependent phospholipase	NA
Banded	Contig75665	12997	16	TR59239|c1_g1	Lipocalin	NA
Banded	Contig75665	12997	16	TR60021|c3_g2	Sulfotransferase	NA
Beige	Contig67175	7477	5	TR66054|c1_g1	Transferase activity	NA
Beige	Contig67175	7614	5	TR66054|c7_g1	Oxidoreductase activity	NA
Beige	Contig67175	7614	5	TR94160|c1_g1	Hedgehog receptor signalling activity	Predicted transmembrane protein
Black	Contig10718	6069	9	TR106019|c0_g1	Cytosolic carboxypeptidase 2	NA
Black	Contig10718	6069	9	TR13527|c0_g1	Prolyl 4-Hydroxylase	NA
Black	Contig10718	6069	9	TR69263|c0_g1	Lysine-specific demethylase 2B	NA
Black	Contig10718	6069	9	TR96615|c0_g1	Uncharacterized protein	Predicted transmembrane protein

The contig in association with the beige colour has three protein coding domains. The gene ontology for each of the coding regions determined oxidoreductase activity, transferase activity and hedgehog signalling receptor activity, respectively. Although the role of transferases in shell colouration is not yet well-studied, Feng et al. (2018) and Lemer et al. (2015) have reported regulatory and/or coding elements with oxidoreductase activity in the shell pigmentation of the pacific oyster (*Crassostrea gigas*) and black-lipped pearl oyster (*Pinctada margaritifera*), respectively. However, among the three annotations on the contig, the gene with the hedgehog signalling receptor activity is of far more interest. The Hedgehog signalling pathway is an evolutionarily conserved pathway essential for embryonic development and cell differentiation. The interaction between the fundamental development processes and colouration is ubiquitous across phylogenies and has been explored a bit more in model organisms, vertebrates and plants. For example, the *hagromo* gene controls an essential developmental pathway in cichlid fishes responsible for pattern formation (Kawakami et al. 2000); the involvement of genes in vertebrate embryogenesis in melanin pathway (discussed in detail in Hoekstra 2006); or, the role of hedgehog signalling pathway in retina pigmentation (e.g., Perron et al. 2003; Dakubo et al. 2008; Todd and Fischer 2015). However, little is known about the role of developmental pathways in the shell colouration in the molluscs. The known literature is discussed below.

As mentioned earlier, the shell matrix and colouration proteins may show mantle-specific tissue expression. Perhaps, supplementing this analysis with tissue-specificity information of the annotated transcripts may give more insight into whether they play a role in shell formation or colour polymorphism.

4.4 Discussion

We estimated the genetic architecture of the traits that influence the shell characteristics, and subsequently the adaptive divergence between the two ecotypes of the marine snail *Littorina saxatilis*. We performed association mapping with the reduced-representation dataset in the hybrid zone between the two distinct ecotypes to elucidate the genotype-phenotype associations for the six shell characteristics – shell size, shape, colour (black, beige and dark beige) and banding pattern.

The colour polymorphism of the shell has been of keen interest in the study of evolution. The visible colour variation between alternative morphs, if under genetic control, provides a unique insight into the maintenance of genetic variation in a population. In most cases, the shell colouration is a heritable trait governed by a single locus (reviewed in Williams 2017). For example, in the land snail *Partula taeniata,* distinct general shell colours and banding pattern are controlled by separate but linked loci (Murray and Clarke 1976). In *Nucella emarginata,* ground colour of the shell is determined by a separate locus from banding pattern and the dominant allele codes for the banding pattern (Palmer 1985). In accordance, previous studies on *Littorina saxatilis* suggested that different colour polymorphisms are inherited in this species and banding pattern and each ground shell colour are controlled by separate unlinked loci with the dominant allele coding for the respective polymorphism (Ekendahl and Johannesson 1997; Johannesson and Butlin 2017). Conforming to their finding, single-locus associations in different genomic regions are identified in this study, on linkage groups 5 and 9 for the shell colours beige and black respectively. For the banding pattern, we identified associations in the linkage group 6, along with loci on linkage groups 10,11, 12 and 16. We did not find any significant association for the dark beige shell colour.

The loci in association with the colour traits in our analysis explained high phenotypic variance. This is in accordance to the theory that traits with simple inheritance tend to be governed by large effect loci. Large effect loci have often been reported to be a genetic architecture underlying adaptive phenotypes and has been known to regulate the colour traits in several species including the *Heliconius* species (Jiggins and McMillan 1997; Joron et al. 2006; Nadeau et al. 2014), the stick insect *Timema cristinae* (Comeault et al. 2014) and the spittlebug *Philaenus spumarius* (Rodrigues et al. 2016). Banding pattern, despite being predicted to be monogenic (Ekendahl and Johannesson 1997; Johannesson and Butlin 2017), showed associations at multiple locations in the genome. Banding pattern is a heterogenous phenotype, that is, it ranges from white spirals on a black background to thick dark lines on a brighter background. While phenotyping, all the banded patterns were considered to be one morph. Therefore, as pointed by Johannesson and Butlin (2017), it may be likely that different loci are responsible for different banding patterns. Our analysis did not delve into this aspect, but it certainly provides an avenue for investigation in the future studies. Another possibility that may need consideration is the presence of epistatic interactions between different loci. Epistatic interactions between the banding loci has been reported in other molluscs (e.g., Johnson 2012) and particularly studied in the land snail (*Cepaea nemoralis*), in which the gene coding for the banding pattern is in epistasis to alleles controlling the band appearance (Jones et al. 1977). However, the current study design tests for the single-SNP associations with the phenotypes, which is ill-suited to detect epistatic interactions (Hoh and Ott 2003; Niel et al. 2015). But this information perhaps may find use in the future studies to detect possible interactions between the significant associations (Niel et al. 2015). Alternatively, this result for banding pattern could be an artefact arising from long-range LD as a by-product of hybridization (Rieseberg and Buerkle 2002).

Association mapping is a statistical test that only investigates whether allelic frequency and trait variation covaries. It does not assert causation. With that in view, regions that showed significant association were further annotated to gain an insight into the underlying genes. It is hard to comment on the role of identified peptide domains for black ground colour and banding pattern in shell pigmentation. The results from the annotation of the contig in association with the beige colour looked more promising, especially the potential link between the developmental pathway and shell formation that may have been revealed through the identification of a hedgehog signalling receptor. The interaction of the hedgehog protein with the developmental genes, *engrailed* and *dpp*, is known to regulate the development of the imaginal disc and wing pattern formation in insects (Basler and Struhl 1994; North and French 1994; Kopp et al. 1997; Shevtsova et al. 2011). A very few studies have focused on the hedgehog signalling pathway in molluscs. A major study was done in cuttlefish (*Sepia officinalis*) in which Grimaldi et al. (2008) showed that hedgehog pathway played a crucial role in the normal mantle tissue development. Additionally, in a separate study in the common limpet (*Patella vulgata*), orthologs of the *engrailed* and *dpp* gene were found to be expressed in the embryonic shell-forming cells (Nederbragt et al. 2002). Perhaps, a machinery similar to the *Drosophila* wing imaginal disc development regulates the shell-formation (and hence, the shell colouration and patterns) in molluscs, in which case, if the association is true, this suggests the involvement of the hedgehog developmental pathway in the basal shell colouration in

Littorina saxatilis. However, to make this conclusion with certainty at this point may be naïve and perhaps expression studies and molecular studies can be undertaken in the future to explore this link.

The estimation of heritability of shell shape and size in our study aligns with the previous studies (Carballo et al. 2001; Conde-Padín et al. 2007a) that categorize the shell shape and size of *Littorina saxatilis* as highly heritable and moderately heritable traits respectively. The inability to detect significant associations points towards both, the limitations of our study (discussed later) and the polygenic architecture of the traits. This aligns with the theory that most of the traits with continuous variation are controlled by complex and polygenic architecture (Hill et al. 2008). That is, multiple loci with additive effects may contribute to the traits and will explain lesser phenotypic variation than traits with large effect QTL (Hill et al. 2008). Not only the study of other systems reveals that size and shape has a polygenic architecture (e.g., Visscher et al. 2007), a genome-wide association study conducted by Joseph Kess (2017) in the Spanish populations of *Littorina saxatilis* also suggested polygenic architecture for both size and shape.

Although empirical work and experimental evidence suggest a simple and reduced-recombination genetic architecture underlying adaptation and speciation, an increasing number of studies also point towards the polygenic architecture (e.g., Hancock et al. 2011). Yeaman (2015) used simulations to further demonstrate that local adaptation could result from a polygenic response through subtle allelic frequency differences in a large number of loci. His work further goes on to suggest that the slight frequency shifts are difficult to capture with association mapping, if not impossible. Furthermore, the QTL approaches are inherently biased to detect the large effect loci and over-estimate the phenotypic variance explained, as demonstrated in the classical study by Beavis (1994,1998). Saying that, multiple genome-wide association studies have been implemented successfully in the model organisms to estimate the small effect loci underlying complex traits (Flint and Mackay 2009). Perhaps, future studies may be undertaken with larger sample size and genotype data covering the whole genome which may help to delve into the genes underlying these traits.

An important result from our analysis was the implication of a chromosomal inversion contributing to the adaptive traits (Westram et al. 2018). Banding pattern showed association at the boundary of an inversion on linkage group 6 (Westram et al. 2018; Faria et al. 2019). Whether a true association, and if the causative allele is within the inversion boundary, needs to be determined. It is an important aspect to investigate because through empirical and theoretical work in other systems the role of chromosomal inversions in causing correlation of colour polymorphisms and other traits has been established (McKinnon and Peirotti 2010). Chromosomal inversions cause correlation of traits through reduced recombination and/or bringing unlinked loci together (McKinnon and Peirotti 2010; Kirkpatrick and Barton 2006). For example, in the Asian swallowtail butterfly (*Papillo polytes*), this correlation is mediated through the inversion of the region that includes the *doublesex* gene (*dsx*), which is implicated in the insect sexual dimorphism and the wing patterns, colour and structures (Wellenreuther et al. 2018). In addition, the same inversion was reported to contribute disproportionately to the heritability of the quantitative traits (Westram et al. 2018). Chromosomal inversions have been implicated for their contribution to complex traits in other organisms. For example, inversions influence body size in grasshoppers and seaweed flies, body size and wing shape in various *Drosophila* species (reviewed in Hoffman and Rieseberg 2008). Chromosomal inversions receive particular attention from evolutionary biologists for their possible role in reproductive isolation (e.g., Reiseberg et al. 2001). A major impetus to the idea of correlation between adaptive traits and traits underlying reproductive isolation comes from the work on *Heliconius cydno* and *Heliconius pachinus*, in which the locus governing the wing colour and the male preference for mating were suggested to be maintained within an inversion (Kornfrost et al. 2006). Kirkpatrick and Barton (2006) have suggested that inversions may lock together a suite of locally-adapted polymorphisms that may spread across a population in the selection-gene flow equilibria as this suite of favourable alleles confer higher fitness than individually occurring favourable alleles outside chromosomal rearrangements. Keeping this in view, Westram et al. (2018) noted high differentiation and steep cline slopes characterized this genomic rearrangement, suggesting strong divergent selection. Therefore, this study contributes to the evidence of the importance of genomic rearrangements to local adaptation.

Although, the lack of association signals also potentially points towards certain genetic architectures, in an ideal world scenario, we should have been able to detect association signals for all the traits. The lack of ability to detect association signals could result from multiple sources and the limitations of our analysis, which may also cause bias in the estimation of effect size or detection of association signals. For example, the lack of significant association for the dark beige shell colour could be the result of imprecise phenotype measurement. A common assumption in the association mapping studies is that the imprecision in the phenotype measurement may not have large effect on the result (Barendse 2011). However, the growing literature contradicts this assumption and suggests that the lack of phenotypic resolution may reduce the statistical power to detect association or cause bias (e.g., Phillips and Smith 1991; Gordon et al. 2004; Edwards et al. 2005; Manchia et al. 2013; Hemani et al. 2017). The precise measurement of shell colours, ground shell colours especially, in this study was difficult. Therefore, if true, it may also suggest bias (or inexact estimation of effect size) in our detection of associations for the black and beige colours. In addition, inherent statistical biases such as winner's curse may also cause inflation of effect sizes for all the detected alleles (Kraft 2008). Population structure correction may also cause the inability to detect association signals for dark beige and shell shape and size. While population structure generated by local adaptation may result in spurious associations, population structure correction for traits that covary with structure poses the risk of eliminating the true associations (Platt et al. 2010). Another major limitation in our study would be the use of reduced-representation genotype data. We were able to capture only half of the predicted genome size which makes it very likely that we may have missed the causal variants, or the SNPs linked with the causal variants. While we may have been able to capture SNPs in linkage with the causal variants for which we observed significant associations, this cannot be said for the traits for which there were no significant associations (or the linkage was too weak to be detected). In short, while this analysis has given an insight into the genetic architecture of the various shell traits, it is necessary to interpret these results as hypotheses for follow-up confirmatory studies.

Conclusion

This chapter describes the association mapping study undertaken in a single population of the marine snail *Littorina saxatilis* from the west coast of Sweden in order to discern the genetic architecture of the locally adaptive traits facilitating ecological speciation. Of the shell traits that were studied, we were able to provide some predictions about their genetic architecture (except for dark beige, due to the limitations of our methodology) and possible contribution towards reproductive isolation. Although in the literature we find support for our predictions, we maintain that the point estimates provided in this study should be interpreted with caution. Additionally, the annotation of the regions in association may have uncovered potential links, especially between the developmental pathways and colouration in the molluscs, that can be explored in the future.

Contributions

I performed the heritability estimation and association mapping analysis and annotated the SNPs with significant associations. This work, however, is the product of collaboration and I would like to acknowledge and thank the following people for their contribution(s).

Sampling and phenotyping were performed by Dr. Anja M. Westram (IST, Austria), Dr. Mark Ravinet (CEES, Norway), Dr. Marina Panova (University of Gothenburg), Prof. Kerstin Johannesson (University of Gothenburg) and Prof. Roger K. Butlin (University of Sheffield). Probe design, library preparation and sequencing were performed by RapidGenomics (Gainsville FL, United States). Bioinformatic pipeline and variant calling were done by Dr. Anja M. Westram (IST, Austria). Dr. Victor Soria-Carrasco (University of Sheffield) provided a script to perform the PC analysis.

References

Aguilera, F., McDougall, C., Degnan, B. M. (2017). Co-option and *de novo* gene evolution underlie molluscan shell diversity. Molecular Biology & Evolution; 34(4): 779-792.

Appleton, R.D., Palmer, A.R. (1988). Water-borne stimuli released by predatory crabs and damaged prey induce more predator-resistant shells in a marine gastropod. Proceedings of the National Academy of Sciences; 85:4387–4391.

Arivalagan, J., Yarra, T., Marie, B., Sleight, V. A., Duvernois-Berthet, E., Clark, M. S., … Berland, S. (2017). Insights from the Shell Proteome: Biomineralization to Adaptation. Molecular Biology & Evolution; 34(1):66–77.

Aulchenko, Y.S., Ripke, S., Isaacs, A., van Duijn, C.M. (2007) GenABEL: An R library for genome-wide association analysis. Bioinformatics; 23(10):1294–1296.

Barendse, W. (2011). The effect of measurement error of phenotypes on genome wide association studies. BMC Genomics; 12: 232.

Basler, K., Struhl, G. (1994). Compartment boundaries and the control of *Drosophila* limb pattern by hedgehog protein. Nature; 368:208-214.

Beavis, W. (1994). The power and deceit of QTL experiments: lessons from comparative QTL studies. Proceedings of the forty-ninth annual corn and sorghum industry research conference. Chicago, IL, USA: American Seed Trade Association, 250–266.

Beavis, W. (1998). QTL analyses: power, precision, and accuracy. In: Paterson AH, ed. Molecular dissection of complex traits. Boca Raton: CRC Press, 145–162.

Boulding, E. G., Rivas, M. J., González-Lavín, N., Rolán-Alvarez, E., Galindo, J. (2017). Size selection by a gape-limited predator of a marine snail: Insights into magic traits for speciation. Ecology and Evolution; 7(2):674–688.

Brookes, J. I., Rochette, R. (2007). Mechanism of a plastic phenotypic response: predator-induced shell thickening in the intertidal gastropod *Littorina obtusata*. Journal of Evolutionary Biology; 20:1015-1027.

Cain, A.J., Sheppard, P.M. (1950). Selection in the polymorphic land snail *Cepaea nemoralis*. Heredity; 4: 275–294.

Calvo-Iglesias, J., Pérez-Estévez, D., González-Fernández, A. (2017). MSP22.8 is a protease inhibitor-like protein involved in shell mineralization in the edible mussel *Mytilus galloprovincialis*. FEBS Open Biology; 7(10): 1539–1556.

Carballo, M., García, C., Rolán-Alvarez, E. (2001). Heritability of shell traits in wild *Littorina saxatilis* populations: results across a hybrid zone. Journal of Shellfish Research; 20: 415–422.

Carvajal-Rodriguez, A., Conde-Padin, P., Rolan-Alvarez, E. (2005). Decomposing shell form into size and shape by geometric morphometric methods in two sympatric ecotypes of *Littorina saxatilis*. Journal of Molluscan Studies; 71:313–318.

Ciccone, D.N., Su, H., Hevi, S., Gay, F., Lei, H., Bajko, J., Xu, G., Li, E., Chen, T. 2009. KDM1B is a histone H3K4 demethylase required to establish maternal genomic imprints. Nature; 461:415–418.

Comeault, A.A., Soria-Carrasco, V., Gompert, Z., Farkas, T.E., Buerkle, C.A., Parchman T.L., et al. (2014). Genome-wide association mapping of phenotypic traits subject to a range of intensities of natural selection in *Timema cristinae* *. American Naturalist; 183:711–27.

Comfort, A. (1951). The pigmentation of molluscan shells. Biological Reviews; 26:285–301.

Conde-Padín, P., Carvajal-Rodríguez, A., Carballo, M., Carballero, A., Rolán-Alvarez, E. (2007a). Genetic variation for shell traits in a direct-developing marine snail involved in a putative sympatric ecological speciation process. Evolutionary Ecology; 21: 635–650.

Conde-Padín, P., Grahame, J.W., Rolan-Alvarez, E. (2007b). Detecting shape differences in species of the *Littorina saxatilis* complex by morphometric analysis. Journal of Molluscan Studies; 73:147–154

Cook, L.M. (1998). A two-stage model for *Cepaea* polymorphism. Philosophical Transactions of the Royal Society B: Biological Sciences; 353: 1577–1593.

Dakubo, G.D., Mazerolle, C., Furimsky, M., Yu, C., St-Jacques, B., McMahon, A.P., Wallace, V.A. (2008). Indian hedgehog signaling from endothelial cells is required for sclera and retinal pigment epithelium development in the mouse eye. Developmental Biology; 320:242-255.

David, E., Tanguy, A., Pichavant, K., Moraga, D. (2005). Response of the Pacific oyster *Crassostrea gigas* to hypoxia exposure under experimental conditions. FEBS J.; 272:5635-5652.

Dillon, R.T. Jr., Jacquemin, S.J. (2015). The heritability of shell morphometrics in the freshwater pulmonate gastropod *Physa*. PLOS ONE; 10(4): e0121962.

Doyle, S., MacDonald, B., Rochette, R.M. (2010). Is water temperature responsible for geographic variation in shell mass of *Littorina obtusata* (L.) snails in the Gulf of Maine? Journal of Experimental Marine Biology and Ecology; 394:98–104.

Edwards, B.J., Haynes, C., Levenstien, M.A., Finch, S.J., Gordon, D. (2005). Power and sample size calculations in the presence of phenotype errors for case/control genetic association studies. BMC Genetics; 6:18.

Ekendahl, A., Johannesson, K. (1997). Shell colour variation in *Littorina saxatilis* Olivi (Prosobranchia: Littorinidae): a multi-factor approach. Biological Journal of the Linnean Society; 62: 401–419.

Fan, X., Qian, Y., Fricker, L.D., Akalal, D.B., Nagle, G.T. (1999). Cloning and expression of *Aplysia* carboxypeptidase D, a candidate prohormone-processing enzyme. DNA Cell Biology; 18(2):121-132.

Faria, R., Chaube, P., Morales, H., Larsson, T., Lemmon, A., Lemmon, E., Rafajlovic, M., Panova, M., Ravinet, M., Johannesson, K., Westram, A., Butlin, R. (2019). Multiple chromosomal rearrangements in a hybrid zone between *Littorina saxatilis* ecotypes. Molecular Ecology. Accepted author manuscript.

Feng, D., Li, Q., Yu, H., Kong, L., & Du, S. (2018). Transcriptional profiling of long non-coding RNAs in mantle of *Crassostrea gigas* and their association with shell pigmentation. Scientific reports; 8(1):1436.

Flint, J., Mackay, T. F. C. (2009). Genetic architecture of quantitative traits in mice, flies, and humans. Genome Research; 19(5):723–733.

Galindo, J., Grahame, J.W., Butlin, R.K. (2010). An EST-based genome scan using 454 sequencing in the marine snail *Littorina saxatilis*. Journal of evolutionary biology; 23:2004–2016.

Giokas, S., Påll-Gergely, B., Mettouris, O. (2014). Nonrandom variation of morphological traits across environmental gradients in a land snail. Evolutionary Ecology; 28:323–340.

Goodfriend, G.A. (1986). Variation in land-snail shell form and size and its causes: A review. Systematic Biology; 35:204–223.

Gordon, D., Yang, Y., Haynes, C., Finch, S.J., Mendell, N.R., et al. (2004). Increasing power for tests of genetic association in the presence of phenotype and/or genotype error by use of double-sampling. Statistical Applications in Genetics and Molecular Biology; 3: Article26.

Gray, S.M., McKinnon, J.S. (2007). Linking color polymorphism maintenance and speciation. Trends in Ecology & Evolution; 22:71–9.

Grimaldi, A., Tettamanti, G., Acquati, F., Bossi, E., Guidali, M. L., Banfi, S., Monti, L., Valvassori, R., de Eguileor, M. (2008). A hedgehog homolog is involved in muscle formation and organization of *Sepia officinalis* (Mollusca) mantle. Developmental dynamics; 237:659-71.

Grover, C.E., Salmon, A., Wendel, J.F. (2012). Targeted sequence capture as a powerful tool for evolutionary analysis. American Journal of Botany; 99(2): 312–319.

Haldane, J. B. S. (1930). A mathematical theory of natural and artificial selection. Part VI. Isolation. Proceedings of the Cambridge Philosophical Society; 26:220–230.

Hancock, A.M., Brachi, B., Faure, N., Horton, M.W., Jarymowycz, L.B., Sperone, F.G., Toomajian, C., Roux, F., Bergelson, J. (2011). Adaptation to climate across the *Arabidopsis thaliana* genome. Science; 334:83–86

Hemani, G., Tilling, K., Davey Smith, G. (2017). Correction: Orienting the causal relationship between imprecisely measured traits using GWAS summary data. PLOS Genetics; 13(12): e1007149.

Hoekstra, H.E. (2006). Genetics, development and evolution of adaptive pigmentation in vertebrates. Heredity; 97:222–234.

Hoffmann, A. A., Rieseberg, L. H. (2008). Revisiting the impact of inversions in evolution: from population genetic markers to drivers of adaptive shifts and speciation?. Annual review of ecology, evolution, and systematics; 39:21-42.

Hoh, J., Ott, J. (2003). Mathematical multi-locus approaches to localizing complex human trait genes. Nature Review Genetics; 4(9):701–709.

Hollander, J., Butlin, R.K. (2010). The adaptive value of phenotypic plasticity in two ecotypes of a marine gastropod. BMC Evolutionary Biology; 10:333.

Hill, W.G., Goddard, M.E., Visscher, P.M. (2008). Data and theory point to mainly additive genetic variance for complex traits. PLOS Genetics; 4:e1000008.

Innes, D.J., Haley, L.E. (1977). Inheritance of a shell-color polymorphism in the mussel. The Journal of Heredity; 68:203-204.

Janson, K., Sundberg, P. (1983). Multivariate morphometric analysis of two varieties of *Littorina saxatilis* from the Swedish west coast. Marine Biology; 74:49.

Jiggins, C.D. (2008). Ecological speciation in mimetic butterflies. BioScience; 58(6): 541–548.

Jiggins, C. D., McMillan, W. O. (1997). The genetic basis of an adaptive radiation: warning colour in two *Heliconius* species. Proceedings of the Royal Society B: Biological Sciences; 264(1385):1167–1175.

Johannesson, K., Butlin, R.K. (2017). What explains rare and conspicuous colours in a snail? A test of time-series data against models of drift, migration or selection. Heredity; 118(1): 21-30.

Johannesson, K., Ekendahl, A. (2002). Selective predation favouring cryptic individuals of snails *Littorina*. Biological Journal of the Linnean Society; 76: 137–144.

Johannesson, K., Johannesson, B., Rolán-Alvarez, E. (1993). Morphological differentiation and genetic cohesiveness over a microenvironmental gradient in the marine snail *Littorina saxatilis*. Evolution; 47: 1770–1787.

Johannesson, B., Johannesson, K. (1996). Population differences in behaviour and morphology in *Littorina saxatilis*: Phenotypic plasticity or genetic differentiation? J. Zool.; 240: 475–493.

Johannesson, K., Panova, M., Kemppainen, P., André, C., Rolán-Alvarez, E., Butlin, R.K. (2010). Repeated evolution of reproductive isolation in a marine snail: unveiling mechanisms of speciation. Philosophical Transactions of the Royal Society B: Biological Sciences; 365: 1735–1747.

Johnson, M. S. (2012). Epistasis, phenotypic disequilibrium and contrasting associations with climate in the land snail *Theba pisana*. Heredity; 108(3):229–235.

Jones, J.S., Leith, B.H., Rawlings, P. (1977). Polymorphism in *Cepaea*: a problem with too many solutions? Annual Review of Ecology and Systematics; 8:109–143.

Joron, M., Jiggins, C.D., Papanicolaou, A., McMillan, W.O. (2006). *Heliconius* wing patterns: an evo-devo model for understanding phenotypic diversity. Heredity; 97:157–67.

Kawakami K., Amsterdam, A., Shimoda, N., Becker, T., Mugg, J., Shima, A., Hopkins, N. (2000). Proviral insertions in the zebrafish *hagoromo* gene, encoding an F-box/WD40-repeat protein, cause stripe pattern anomalies. Current Biology; 10:463-466.

Kent, W. J. (2002). BLAT—The BLAST-Like Alignment Tool. Genome Research, 12(4), 656–664.

Kess, J. (2017). Genomic architecture of parallel ecological divergence in Galician *Littorina saxatilis* ecotypes. University of Guelph, Guelph, Canada.

Kirkpatrick, M., Barton, N. (2006). Chromosome inversions, local adaptation and speciation. Genetics; 173(1): 419-34.

Kingsolver, J.G., Hoekstra, H.E., Hoekstra, J.M., Berrigan, D., Vignieri, S.N., Hill, C.E., Hoang, A., Gibert, P., Beerli, P. (2001). The strength of phenotypic selection in natural populations. American Naturalist; 157(3):245-61.

Kopp, A., Muskavitch, M.A.T., Duncan, I. (1997). The roles of hedgehog and engrailed in patterning adult abdominal segments of *Drosophila*. Development; 124 (19): 3703-3714.

Kostem, E., Eskin, E. (2013). Improving the accuracy and efficiency of partitioning heritability into the contributions of genomic regions. American Journal of Human Genetics; 92(4):558–564.

Kozminskii, E.V., Lezin, P.A. (2007). Pigment distribution in the shell of gastropod mollusk *Littorina obtusata* (Linnaeus, 1758). Russian Journal of Marine Biology; 33(4): 238–244.

Kraft, P. (2008). Curses—winner's and otherwise—in genetic epidemiology. Epidemiology; 19:649–651.

Kronforst, M. R., Gilbert, L. E. (2006). The population genetics of mimetic diversity in *Heliconius* butterflies. Proceedings. Biological sciences; 275(1634), 493-500.

Ky, C.L., Nakasai, S., Pommier, S., Koua, M.S., Devaux, D. (2016). The Mendelian inheritance of rare flesh and shell colour variants in the black-lipped pearl oyster (*Pinctada margaritifera*). Animal Genetics; 47(5):610-614.

Lande, R. (1979). Quantitative genetic analysis of multivariate evolution, applied to brain: body size allometry. Evolution; 33: 402-416.

Lemer, S., Saulnier, D., Gueguen, Y., Planes, S. (2015). Identification of genes associated with shell color in the black-lipped pearl oyster, *Pinctada margaritifera*. BMC Genomics; 16(1): 568.

Lozier, J.D., Jackson, J.M., Dillon, M.E., Strange, J.P. (2016). Population genomics of divergence among extreme and intermediate color forms in a polymorphic insect. Ecology and Evolution; 6:1075–91.

Manchia, M., Cullis, J., Turecki, G., Rouleau, G. A., Uher, R., Alda, M. (2013). The impact of phenotypic and genetic heterogeneity on results of Genome Wide Association Studies of complex diseases. PLOS ONE; 8(10): e76295.

McDougall, C., Degnan, B.M. (2018). The evolution of mollusc shells. Wiley Interdisciplinary Reviews. Developmental Biology; 7(3):e313.

McKinnon, J.S., Pierotti, M.E. (2010). Colour polymorphism and correlated characters: genetic mechanisms and evolution. Molecular Ecology; 19(23):5101-25.

Money, D., Gardner, K., Migicovsky, Z., Schwaninger, H., Zhong, G-Y, Myles, S. (2015). LinkImpute: Fast and Accurate Genotype Imputation for Nonmodel Organisms. G3: Genes|Genomes|Genetics; 5(11):2383-2390.

Mullen, L.M., Hoekstra, H.E. (2008). Natural selection along an environmental gradient: a classic cline in mouse pigmentation. Evolution; 62(7):1555-70.

Murray, J., Clarke, B.C. (1976) Supergenes in polymorphic land snails. Heredity; 37:253–269.

Nagaoka, K., Hino, S., Sakamoto, A., Anan, K., Takase, R., Umehara, T., Yokoyama, S., Sasaki, Y., … Nakao, M. (2015). Lysine-specific demethylase 2 suppresses lipid influx and metabolism in hepatic cells. Molecular and cellular biology; 35(7):1068-80.

Nadeau, N. J., Ruiz, M., Salazar, P., Counterman, B., Medina, J. A., Ortiz-Zuazaga, H., Morrison, A., McMillan, W. O., Jiggins, C. D., … Papa, R. (2014). Population genomics of parallel hybrid zones in the mimetic butterflies, *H. melpomene* and *H. erato*. Genome research; 24(8):1316-33.

Nederbragt, A.J., Loon, A.E. van, Dictus, W.J. (2002). Expression of *Patella vulgata* orthologs of engrailed and dpp-BMP2/4 in adjacent domains during molluscan shell development suggests a conserved compartment boundary mechanism. Developmental Biology ; 246:341-355.

Niel, C., Sinoquet, C., Dina, C., Rocheleau, G. (2015). A survey about methods dedicated to epistasis detection. Frontiers in Genetics; 6:285.

North, G., French, V. (1994). Insect Wings: Patterns upon patterns. Current Biology; 4(7):611-614.

Nosil, P. (2012). Ecological Speciation. Oxford: Oxford University Press.

Nosil, P., Schluter, D. (2011). The genes underlying the process of speciation. Trends in Ecology & Evolution; 26: 160–167.

Palmer, A.R. (1985). Genetic basis of shell variation in *Thais emarginata* (Prosobranchia: Muricacea). I. Banding in populations from Vancouver Island. Biological Bulletin; 169: 638–651.

Panova, M., Aronsson, H., Cameron, R. Dahl, P., Godhe, A., Lind, U., Ortega-Martinez, O., Pereyra, R., Tesson, S., Wrange, A., Blomberg, A., Johannesson, K. (2016). DNA extraction protocols for whole-genome sequencing in marine organisms. Marine Genomics, Methods in Molecular Biology., ed Bourlat SJ (Springer New York), 13–44.

Perron, M., Boy, S., Amato, M. A., Viczian, A., Koebernick, K., Pieler, T., Harris, W. A. (2003). A novel function for Hedgehog signalling in retinal pigment epithelium differentiation. Development; 130:1565-1577.

Phillips, A.N., Davey Smith, G. (1991). How independent are "independent" effects? relative risk estimation when correlated exposures are measured imprecisely. Journal of Clinical Epidemiology; 44: 1223–1231.

Platt, A., Vilhjálmsson, B. J., Nordborg, M. (2010). Conditions Under Which Genome-Wide Association Studies Will be Positively Misleading. Genetics; 186(3):1045–1052.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics; 38(8):904.

Price, A. L., Zaitlen, N. A., Reich, D., Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. Nature Review Genetics; 11(7):459-63.

Primost, M.A., Bigatti, G. & Márquez, F. (2015). Shell shape as indicator of pollution in marine gastropods affected by imposex. Marine and Freshwater Research; 67:1948–1954.

Punzalan, D., Rodd, F.H., Hughes, K.A. (2005). Perceptual processes and the maintenance of polymorphism through frequency-dependent predation. Evolutionary Ecology; 19:303–320.

Ravinet, M., Westram, A., Johannesson, K., Butlin, R.K., André, C., Panova, M. (2016). Shared and nonshared genomic divergence in parallel ecotypes of *Littorina saxatilis* at a local scale. Molecular Ecology; 25(1):287–305.

Reid, D.G. (1996). Systematics and Evolution of Littorina. London: Ray Society.

Reid, D.G., Dyal, P., Williams, S.T. (2012). A global molecular phylogeny of 147 periwinkle species (Gastropoda, Littorininae). Zoologica Scripta; 41: 125-136.

Reimchen, T.E., Nosil, P. (2002). Temporal variation in divergent selection on spine number in threespine stickleback. Evolution; 56: 2472-2483.

Richards, P. M., Liu, M. M., Lowe, N., Davey, J. W., Blaxter, M. L., Davison, A. (2013). RAD-Seq derived markers flank the shell colour and banding loci of the *Cepaea nemoralis* supergene. Molecular Ecology; 22(11):3077-3089.

Rieseberg, L.H. (2001). Chromosomal rearrangements and speciation. Trends in Ecology & Evolution; 16:351–358.

Rieseberg, L.H., Buerkle, C.A. (2002). Genetic mapping in hybrid zones. American Naturalist; 159(Suppl 3):S36–50.

Rodrigues, A. S., Silva, S. E., Pina-Martins, F., Loureiro, J., Castro, M., Gharbi, K., Johnson, K. P., Dietrich, C. H., Borges, P. A., Quartau, J. A., Jiggins, C. D., Paulo, O. S., … Seabra, S. G. (2016). Assessing genotype-phenotype associations in three dorsal colour morphs in the meadow spittlebug *Philaenus spumarius* (L.) (Hemiptera: Aphrophoridae) using genomic and transcriptomic resources. BMC genetics, 17(1), 144.

Rolán-Alvarez, E., Ekendahl, A. (1996). Sexual selection and non-random mating for shell colour in a natural population of the marine snail *Littorina mariae* (Gastropoda: Prosobranchia). Ophelia; 45: 1–15.

Rolán-Alvarez, E., Johannesson, K., Erlandsson, J. (1997). The maintenance of a cline in the marine snail *Littorina saxatilis*: the role of home site advantage and hybrid fitness. Evolution; 51(6):1838-1847.

Roulin, A., Ducret, B., Ravussin, P.A., Altwegg, R. (2003). Female colour polymorphism covaries with reproductive strategies in the tawny owl *Strix aluco*. Journal of Avian Biology; 34:393–401.

Rosin, Z.M., Kobak, J., Lesicki, A., Tryjanowski, P. (2013). Differential shell strength of *Cepaea nemoralis* colour morphs – implications for their anti-predator defence. Naturwissenschaften; 100: 843-851.

Rundle, H., Nosil, P. (2005). Ecological speciation. Ecology Letters; 8: 336-352.

Saura, M., Rivas, M.J., Diz, A.P., Caballero, A., Rolan-Alvarez, E. (2012). Dietary effects on shell growth and shape in an intertidal marine snail, *Littorina saxatilis*. Journal of Molluscan Studies; 78(2): 213–216.

Shevtsova, E., Hansson, C., Janzen, D.H., Kjærandsen, J. (2011). Stable structural color patterns displayed on transparent insect wings. Proceedings of the national academy of sciences of the United States of America; 108(2): 668-673.

Sokolova, I.M., Berger, V. Ya. (2000). Physiological variation related to shell colour polymorphism in White Sea, *Littorina saxatilis*. Journal of Experimental Marine Biology and Ecology; 245:1-23.

Stankowski, S. (2013), Ecological speciation in an island snail: evidence for the parallel evolution of a novel ecotype and maintenance by ecologically dependent postzygotic isolation. Molecular Ecology; 22: 2726-2741.

Stinchcombe, J.R., Hoekstra, H.E. (2008). Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. Heredity; 100(2):158-70.

Todd, L., Fischer, A.J. (2015). Hedgehog signaling stimulates the formation of proliferating Müller glia-derived progenitor cells in the chick retina. Development; 142: 2610-2622.

Vercken, E., Clobert, J., Sinervo, B. (2010). Frequency-dependent reproductive success in female common lizards: a real-life hawk–dove–bully game? Oecologia; 162:49–58.

Visscher, P. M., Macgregor, S., Benyamin, B., Zhu, G., Gordon, S., Medland, S., … Martin, N. G. (2007). Genome Partitioning of Genetic Variation for Height from 11,214 Sibling Pairs. American Journal of Human Genetics; 81(5):1104–1110.

Wang, Q.H., Yang, C.Y., Hao, R.J., Zheng, Z., Jiao, Y., Du, X.D., Deng, Y.W., Huang, R.L. (2017). Molecular characterization of CHST11 and its potential role in nacre formation in pearl oyster *Pinctada fucata martensii*. Electronic Journal of Biotechnology; 28:113-119.

Wellenreuther, M., Bernatchez, L. (2018). Eco-Evolutionary Genomics of Chromosomal Inversions. Trends in Ecology & Evolution; 33(6):427-440.

Wellenreuther, M., Svensson, E. I., Hansson, B. (2014), Sexual selection and genetic colour polymorphisms in animals. Molecular Ecology; 23:5398-5414.

Westram, A. M., Galindo, J., Alm Rosenblad, M., Grahame, J. W., Panova, M., Butlin, R. K. (2014). Do the same genes underlie parallel phenotypic divergence in different *Littorina saxatilis* populations? Molecular Ecology; 23(18):4603–4616.

Westram, A.M., Panova, M., Galindo, J., Butlin, R.K. (2016). Targeted re-sequencing reveals geographic patterns of differentiation for loci implicated in parallel evolution. Molecular Ecology; 25:3169–3186.

Westram, A.M., Rafajlović, M., Chaube, P., Faria, R., Larsson, T., Panova, M., Ravinet, M., Blomberg, A., Mehlig, B., Johannesson, K., Butlin, R. (2018). Clines on the seashore: the genomic architecture underlying rapid divergence in the face of gene flow. Evolution Letters; 2(4): 297-309.

Williams, S. T. (2017). Molluscan shell colour. Biological Reviews; 92:1039-1058.

Wright, T.R. (1987). The genetics of biogenic amine metabolism, sclerotization, and melanization in *Drosophila melanogaster*. Advances in Genetics; 24:127–222.

Yeaman, S. (2015). Local adaptation by alleles of small effect. American Naturalist; 186:S74–S89.

Yeaman, S., Otto, S. P. (2011). Establishment and maintenance of adaptive genetic divergence under migration, selection, and drift. Evolution; 65:2123–2129.

Yeaman, S., Whitlock, M. (2011). The genetic architecture of adaptation under migration-selection balance. Evolution; 65:1897–1911.

How does the genotype interact with the environment? Are there only a few genes and gene families that get involved in this interaction? How does the outcome affect the phenotypes and eventually, the ecological success of the species? Do similar environments favour the same genes? Can the same phenotype be achieved by a different genetic mechanism? These are a few of the many questions that continue to puzzle evolutionary biologists. The answers are critical to understanding the processes of adaptation and speciation. A perquisite is to understand the underlying genetics. An indispensable aspect of such studies is to gather information from diverse taxonomic groups, representing different life histories and ecological niches, before a conclusive argument can be made about the genetics underlying adaptation and speciation (Stinchcombe and Hoekstra 2008). However, until recently, evolutionary biology has been stunted by the limitations imposed by the technology to study genetics in organisms beyond the established genetic model systems.

To address this gap, and to contribute a bit more to the knowledge of adaptation and speciation, the present study was undertaken in the non-model organism, the marine snail species *Littorina saxatilis*. The present study aimed at characterizing the genetic architecture of the different adaptive shell traits in the marine snail *Littorina saxatilis.* The study was helped by the next-generation sequencing (NGS) technology, which was used to generate data for the genomic toolbox and the population-level analysis. The genomic toolbox was eventually utilized in the downstream population data analysis.

5.1 The genomic toolbox of *Littorina saxatilis*

Evolutionary genetics approaches rely heavily on the genetic and genomic toolbox of the species. *Littorina saxatilis* has been of relevance from the evolutionary point of view, therefore efforts had been made in the past to develop its genomic resources. A transcriptome and genome assembly had been developed prior to this study (Canbäck et al. 2012). However, those resources were the product of past technologies and were not sufficient in quality or coverage (the transcriptome assembly is discussed briefly in Chapter 2). The current NGS technology was employed to update the existing resources and/or create new ones.

In this study, I describe the construction and annotation of a high-quality and the most comprehensive transcriptome assembly of *Littorina saxatilis* to date. The transcripts from the transcriptome assembly have been utilized to annotate the current genome assembly of the organism (Panova, Larsson et al. in preparation), and the genomic regions of relevance in this study. In addition to the transcriptome assembly, the construction of the first high-density linkage map for this species is also described. Genetic maps have been employed to improve the genome assembly, place the trait-associated regions on the genome, and study patterns of LD and recombination landscape in the organism. Furthermore, the sex-specific linkage maps allowed us to observe sex-specific recombination patterns and gave an insight into the sex determinism in this species. In addition to this study, the genetic resources described in this thesis also underly the studies undertaken by Westram et al. (2018) and Faria et al. (2019).

The genetic and genomic toolbox of any species should be considered a work-in-progress as the resources need to be updated regularly depending on a number of factors. (i) The current technology that may provide a significant advantage over the resources developed by the past methods. For example, the current and upcoming sequencing methods with reduced errors and longer sequence reads may find utilization in the scaffolding of the genome contigs. (ii) The availability of more data. For example, the current *Littorina saxatilis* genetic linkage maps are limited by their coverage of the genome and the information gathered from a single family (and only two sexed individuals). Therefore, caution should be taken before extrapolating the patterns elucidated from the map to the entire population. Once the data that covers the entire genome or includes the other families (and sexed individuals) is available, the linkage maps would need to be updated and that may perhaps change some of these results, but that would be more conclusively applicable to a population. (iii) The revision of the public databases used for the resource development. This is perhaps applicable to transcriptome and genome assemblies only. The annotation of these resources relies on the quality of the public databases used. And the public databases such as, GO, KEGG, Nr, Uniprot, are regularly curated. This means some entries may be revised, removed and added over a period. For instance, KEGG database has a limited molluscan representation, derived from three mollusc genomes (*Lottia gigantea*, *Octopus bimaculoides* and *Crassostrea gigas*). However, with the NGS burst, many more mollusc genomes are sequenced and studied; which means that, in the future, KEGG may be updated to have more molluscs. The annotation with KEGG database of the transcriptomic/genomic resource in that case would not only improve the number of transcripts that get annotated, but perhaps provide more relevant annotations.

The point estimates of genomic resources described in this study are either the most updated resources or the only resource available for this organism. I hope that these genetic resources will be of use for the *Littorina* community and find utility in the evolutionary genetic studies undertaken in the future, including the follow-up association studies.

5.2 The shell and the genes

In this study, the power of association mapping in the hybrid zone was utilized to estimate the genetic architecture of the following adaptive shell traits: shell size, shell shape and shell colours (black, beige and dark beige) and banding pattern in *Littorina saxatilis*. Concordant to the theory and previous evidence, the signals for shell colour polymorphisms were observed in different regions of the genome, signifying distinct loci underlying different colour polymorphism, except for the dark beige shell ground colour, for which there was no signal. For each of the shell colour polymorphisms, for which there was association in the genome (i.e., black, beige and banded pattern), the loci of large effect were implicated in explaining phenotypic variance. While finding associations for the quantitative traits would have been more appealing, the lack of any signals may also signify polygenic architecture underlying shell size and shape.

On one hand, our results on the genetic architecture underlying different adaptive traits may be supported by empirical work on other systems (e.g., Comeault et al. 2014; Nadeau et al. 2014); on the other, it may point towards the limitation of our data and methods. Therefore, follow-up studies are recommended to test the genetic architecture predictions.

The knowledge of genetic architecture of the adaptive traits is the first step towards the better understanding of the process of adaptation. However, the methodology utilized in this study, association mapping, is strictly statistical. Implying, the genotype-phenotype association does not indicate causation. In addition, the reduced-representation genotype data may have severely restricted our ability to detect causal variants. Therefore, even though annotation of the associated regions may have uncovered a potential link between the developmental and shell colouration pathways in our organism which needs to be studied further, we do not make claims about prediction of causal variants in our analysis. Perhaps, performing the association study with whole genome data and larger samples may help elucidate causal links between genotype and phenotype. Another way forward is through the analysis of controlled crosses (currently undertaken at the University of Gothenburg). The causal links underlying a phenotype may be used to explain the role of the given genetic region in underlying trait variation (Barrett and Hoekstra 2011).

Another interesting aspect of this study was that the same markers were used by Westram et al. (2018) for the clinal analysis. This means selection on the traits and the associated SNPs could be analysed too. While all the shell traits discussed here showed clinal variation along the transect, none of the SNPs in the associated regions were eventually used for the clinal analysis (except for one, which did not show any significant cline). Therefore, along with the identification of causal variants underlying adaptive phenotypes, this also remains a potential question for the future studies – how selection is acting at the level of individual genes or SNPs and how does that affect the interplay between the phenotype, underlying genetic variation and fitness in nature

While this thesis has focused exclusively on a few shell traits, there have been other traits which were implicated in the local adaptation of the two ecotypes such as, shell thickness, boldness behaviour, foot size (Johannesson et al. 2010). The inclusion of the other traits in such studies in the future may help to contextualize the influence of these traits on reproductive isolation and local adaptation.

References

Barrett, R.D., Hoekstra, H.E. (2011). Molecular spandrels: tests of adaptation at the genetic level. Nature Review Genetics; 12(11):767-80.

Canbäck, B., André C., Galindo J., Johannesson K., Johansson T., Panova M., Tunlid A., Butlin R.K. (2012). The Littorina sequence database (LSD)—an online resource for genomic data. Molecular Ecology Research; 12:142–148.

Comeault, A.A., Soria-Carrasco, V., Gompert, Z., Farkas, T.E., Buerkle, C.A., Parchman T.L., et al. (2014). Genome-wide association mapping of phenotypic traits subject to a range of intensities of natural selection in *Timema cristinae* *. American Naturalist; 183:711–27.

Faria, R., Chaube, P., Morales, H., Larsson, T., Lemmon, A., Lemmon, E., Rafajlovic, M., Panova, M., Ravinet, M., Johannesson, K., Westram, A., Butlin, R. (2019). Multiple chromosomal rearrangements in a hybrid zone between *Littorina saxatilis* ecotypes. Molecular ecology. Accepted author manuscript.

Johannesson, K., Panova, M., Kemppainen, P., André, C., Rolán-Alvarez, E., Butlin, R. K. (2010). Repeated evolution of reproductive isolation in a marine snail: unveiling mechanisms of speciation. Philosophical Transactions of the Royal Society B: Biological Sciences; 365(1547):1735–1747.

Nadeau, N. J., Ruiz, M., Salazar, P., Counterman, B., Medina, J. A., Ortiz-Zuazaga, H., Morrison, A., McMillan, W. O., Jiggins, C. D., … Papa, R. (2014). Population genomics of parallel hybrid zones in the mimetic butterflies, *H. melpomene* and *H. erato*. Genome research; 24(8):1316-33.

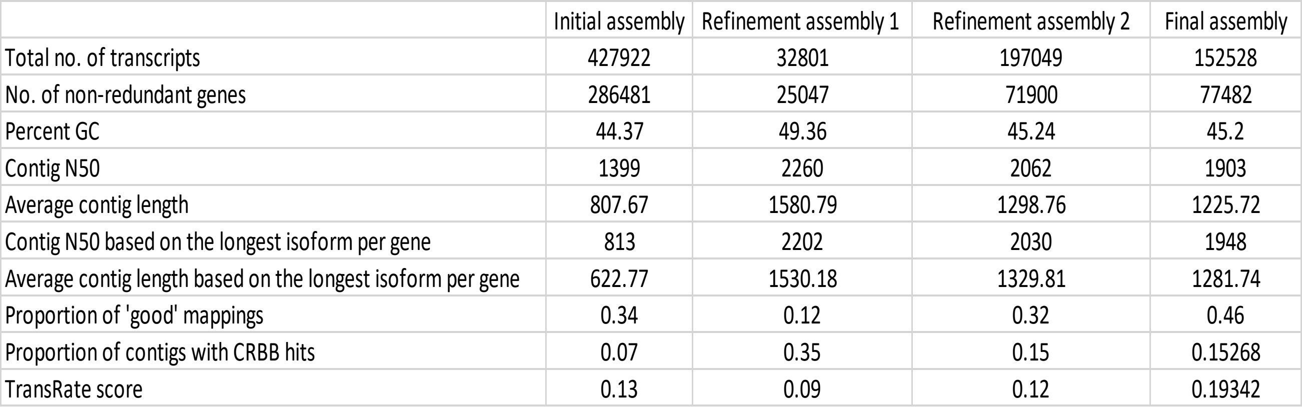
Panova, M., Larsson, T., Alm Rosenblad, M., Chaube, P., Westram, A.M., Butlin, R.K., Blomberg, A., Johannesson, K. (in prep.) Insights into local adaptation from the genome of *Littorina saxatilis* (Mollusca: Gastropoda).

Stinchcombe, J.R., Hoekstra, H.E. (2008). Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. Heredity; 100(2):158-70.

Westram, A.M., Rafajlović, M., Chaube, P., Faria, R., Larsson, T., Panova, M., Ravinet, M., Blomberg, A., Mehlig, B., Johannesson, K., Butlin, R. (2018). Clines on the seashore: the genomic architecture underlying rapid divergence in the face of gene flow. Evolution Letters; 2(4): 297-309.

Appendix I

Table I: Assembly metrics and TransRate results. Initial assembly was constructed using default parameters in Trinity. The transcripts from this assembly were filtered using two approaches: - (i) refinement assembly 1, transcripts which were characterized as 'good' with TransRate and possessed at least one ORF were retained; (ii) refinement assembly 2, the transcripts were filtered based on normalized expression values. The final assembly constituted of transcripts present in both refined assemblies.



	Initial assembly	Refinement assembly 1	Refinement assembly 2	Final assembly
Total no. of transcripts	427922	32801	197049	152528
No. of non-redundant genes	286481	25047	71900	77482
Percent GC	44.37	49.36	45.24	45.2
Contig N50	1399	2260	2062	1903
Average contig length	807.67	1580.79	1298.76	1225.72
Contig N50 based on the longest isoform per gene	813	2202	2030	1948
Average contig length based on the longest isoform per gene	622.77	1530.18	1329.81	1281.74
Proportion of 'good' mappings	0.34	0.12	0.32	0.46
Proportion of contigs with CRBB hits	0.07	0.35	0.15	0.15268
TransRate score	0.13	0.09	0.12	0.19342

Appendix II

Table II: BUSCO scores. LSD assembly is the previous transcriptome assembly of *Littorina saxatilis* available at http://mbio-serv2.mbioekol.lu.se/Littorina/. For comparison, BUSCO result for final set of transcripts and refinement assemblies is also provided. It is evident in this result that LSD assembly did not represent a complete a picture of the *Littorina saxatilis* transcriptome. The presence of more fragmented BUSCOs than complete BUSCOs in the LSD assembly is indicative of the fragmented assembly.

	LSD assembly	Refinement assembly 1	Refinement assembly 2	Final set of transcripts
Complete BUSCOs	24 (2.85%)	528 (62.63%)	796 (94%)	789 (93.59%)
Complete and single-copy BUSCOs	23 (2.72%)	436 (51.72%)	613 (72.71%)	663 (78.65%)
Complete and duplicated BUSCOs	1 (0.11%)	92 (10.91%)	183 (21.7%)	126 (14.95%)
Fragmented BUSCOs	39 (4.62%)	42 (4.98%)	27 (3.2%)	27 (3.2%)
Missing BUSCOs	780 (92.52%)	273 (32.38%)	20 (2.37%)	27 (3.2%)
Total BUSCO groups searched	843	843	843	843

Appendix III

Table III: The complete list of transcripts segregated by GO Slim categories

	GO Slim categories	*No. of transcripts*
Component	membrane	13933
Component	cell	7156
Component	nucleus	3590
Component	cytoplasm	3276
Component	extracellular	2567
Component	intracellular	1402
Component	extracellular matrix	657
Component	chromosome	534
Component	extracellular space	477
Component	cell surface	413
Function	binding	22910
Function	protein binding	2820
Function	receptor activity	1746
Function	kinase activity	1429
Function	transporter activity	1049
Function	hydrolase activity	882
Function	oxidoreductase activity	729
Function	transferase activity	619
Function	nucleic acid binding	543
Function	ligase activity	523
Function	catalytic activity	363
Function	motor activity	323
Function	helicase activity	246
Function	lyase activity	130
Function	isomerase activity	110
Function	signal transducer activity	104
Function	structural molecule activity	90
Function	carrier activity	64
Function	protein transporter activity	48
Function	aromatase activity	35
Function	enzyme regulator activity	18
Function	antioxidant activity	17
Function	translation regulator activity	3
Process	transport	5001
Process	development	3785
Process	cell differentiation	1041
Process	secretion	405
Process	behaviour	370
Process	biosynthesis	134
Process	cell death	129
Process	pathogenesis	120
Process	cell communication	98

Process	response to stimulus	65
Process	cell motility	59
Process	electron transport	41
Process	metabolism	28
Process	catabolism	14
Process	cellular process	13
Process	membrane fusion	12

Glossary

Allopatric speciation Evolution of reproductive barriers due to complete geographical isolation.

Association mapping The identification of genetic loci that may contribute to a phenotype by assessing genotype-phenotype correlations in a population.

Divergent selection Selection that favours different alleles in different environment between two populations.

Ecological speciation Speciation caused via divergent selection acting between two populations adapting to contrasting environments.

Effect size The quantitative measure of influence of a genetic locus on a trait.

Epistasis The interaction between genes that can potentially modify their expression.

Fixation When an allelic variant has achieved 100% frequency in a population.

Gene expression The process through which the functional products of genes is obtained. Gene expression is assessed via two measures: how many mRNA transcripts are generated per locus (gene expression level), and in how many tissues mRNA transcripts are generated (gene expression breadth).

Gene flow The movement of alleles between populations.

Genetic architecture The characteristic genetic variation underlying heritable phenotypic variation. It is an umbrella term that comprises of number of genetic variants, their distribution, effect size and their interaction with each other and environment.

Housekeeping genes The genes that are required for the normal metabolism of cells, and therefore, have ubiquitous expression profile across different cells and tissues.

Hybridization Mating between distinct species or populations.

Linkage disequilibrium Statistical association between two loci that may or may not be in physical linkage.

Local adaptation Higher fitness of an individual in the local environment than the immigrant.

Mutation-order speciation Speciation in which divergence is caused by fixation of different and/or incompatible mutations (alleles) between populations experiencing similar selection pressures.

Parapatric speciation Evolution of reproductive barriers despite of some gene flow due to adjacent geographic range.

Phenotypic plasticity Adaptation to environment via difference in gene expression.

Pleiotropy Effect of an allele on more than one trait.

Reproductive isolation The complete or partial barrier to gene flow between populations.

Sympatric speciation Evolution of reproductive barriers despite of overlap in geographic ranges and no spatial barriers to gene flow.

Tissue-specific genes The genes that confer special function to a particular cell-type and hence, their expression and function are restricted to a single or few tissues.