Identifying genetic markers linked to distyly in Linum tenue

EDWARDS, LEWIS,ALEKSEI

Identifying genetic markers linked to distyly in *Linum tenue*

Lewis Aleksei Edwards

Identifying genetic markers linked to distyly in *Linum tenue.*

Lewis Aleksei Edwards

Abstract

Heterostyly is an adaptation designed to minimise inbreeding and promote outcrossing in plants, defined by the discontinuous variation in the lengths of pollinating organs between distinct morphs in a population. It has been thought to be controlled across species by a diallelic heterozygous supergene, yet recent research has increasingly supported a hemizygous supergene model, where the supergene is only present in one of the stylar morphs. Heterostyly in *Linum* has been well characterized for many years, yet little research has been done into the genetics of it, providing a platform to test this model.

Samples of distylous *Linum tenue*, an understudied species in the genus, were sequenced using ddRAD sequencing, and a *de novo* assembly was generated from these reads using the STACKs software package. These mapped reads were used to identify potential heterozygous loci associated with one of the two stylar morphs, and to search for potential hemizygous supergene candidates, thus testing whether heterostyly in this species is controlled heterozygously or hemizygously.

No hemizygous loci significantly associated with a stylar morph could be found, indicating that heterostyly in *Linum tenue* is not controlled hemizygously. Furthermore, several heterozygous loci could be identified that were significantly associated with a morph. These loci included one encoding a cysteine protease homolog, and another encoding a valine-tRNA ligase homolog. However, issues with the samples used and the post sequencing processing mean that no clear conclusions can be drawn. Several potential genetic markers for heterostyly were identified, but it could not be concluded how heterostyly is controlled in this species.

Identifying genetic markers linked to distyly in *Linum tenue*.

Lewis Aleksei Edwards

Submitted for the degree of Master of Science (by Research)

Supervised by Dr Adrian Brennan

Department of Biosciences

University of Durham

United Kingdom

September, 2020

Acknowledgements

There are many people who, without which, I would not be submitting this thesis today, and I'd like to take the opportunity to thank them. Firstly, my excellent supervisor Dr Adrian Brennan, for being endlessly patient, encouraging and helpful throughout the writing of this thesis. This has been a long time coming, I was not the easiest person to supervise and you would have had every right not to be half as committed to this thesis as you were. Your feedback, recommendations and references were invaluable at every stage, and your encouragement kept me going. Thank you.

To the other researchers who I have met and worked alongside in the course of this project, most especially Ramona Irimia, Eleanor Desmond, Ali Foroozani and Ruth Laidler, for their excellent advice and company in person, and their invaluable examples of proper theses later. To the other members of the lab, though it's been a long time since I've worked with you. And to all the researchers who I've talked with and seen talks by. You inspired, encouraged, and informed me, every one of you.

To my friends, for keeping me no less sane than usual. Thank you for sticking with me.

To my family, the same. You had no choice, but I appreciate it anyway.

More seriously, thank you for your support and encouragement. It was a difficult road getting here, but it would have been a lot harder without your help. I didn't always appreciate it at the time, but I do now.

To my partner, Rosie. Thank you for believing in me when I didn't myself, and knowing the right thing to say at every turn. I still don't quite believe I could write a thesis, but I seem to have done something similar. I would not have finished this without you. Thank you.

Contents

Section 1: Introduction

Chapter 1.1: Heterostyly

For plants, genetic diversity is of utmost importance. Environments can change, ecosystems can shift and new species relationships can arise, all within a relatively short period of time, and because of this the requirements to survive in the local environment can change quickly as well. The species that stand the best long term chance of survival are those with the most diverse genetics that they can pass on through generations, as this increases the chance that they can express traits suited to any given environment. The traits suited to the current environment are selected for, and the individual survives to reproduce (Hughes *et al.* 2008). To ensure this genetic diversity, individuals need to breed with those with different genetics, combining two sets of genes in the offspring and increasing diversity. However, this requires more energy than reproducing asexually, or reproducing with individuals in the same population. As such, features have evolved in various plants to incentivise outcrossing over inbreeding (Barrett, 2003). The nature of these features vary between species, but one of these features is in plants heterostyly.

Heterostyly is a key floral feature for promoting outcrossing, ensuring the genetic diversity of a population. Heterostyly consists of the discontinuous variation in the lengths of pollinating organs among a population of plants, also known as reciprocal herkogamy. For example, a heterostylous population could consist of two predominant floral phenotypes – some with long styles and short anthers (pins) and some with short styles and long anthers (thrums). Heterostyly has been observed in plants, particularly of the *Primula* genus, for centuries, with illustrations and descriptions dating back as far as the 16th century (Gilmartin, 2015). However, the majority of these descriptions viewed heterostyly as nothing more than a curiosity, a way to subclassify species and identify flowers. It was not until 1862 that Darwin first studied the function of heterostyly in *Primula* (Darwin, 1862), expanding to cover heterostyly across species in 1877 (Darwin, 1877). He identified that the purpose of heterostyly is to promote outcrossing, and proposed a theory for how that functions. The species he studied were primarily pollinated by insect pollinators, with pollen from an anther attaching to an insect’s leg and body as it visited the flower, then being deposited on another flower’s style when the insect visits another flower. With a heterostylous population of flowers, the pollen from the long anther would be deposited higher up the insect than the short anther, meaning that this pollen was primarily deposited on the long style, while short anther pollen was deposited on short styles for the same reason. As long anthers and long styles were from different plant morphs, this ensured that inter-morph pollination was more likely to occur than intra-morph or self-pollination, promoting outcrossing and increasing genetic diversity. While Darwin's proposal was entirely theoretical, later studies (Stone, 1995; Keller *et al.*, 2014; Costa *et al.*, 2017) tracked pollen flow between morphs and concluded that this mechanism was accurate – reciprocal herkogamy did cause inter morph pollination.

In addition, it was discovered that heterostyly did not solely consist of this arrangement of pollinating organs. The majority of heterostylous species were found to also contain mechanisms ensuring selective infertility for pollen of the same morph, or self-incompatibility (Ganders, 1979). Even if the pollen from the long anther was deposited on the short style, due to specific features of the pollen and style, the pollen could not fertilise the ovum of the short style. Only pollen from the other morph could be used to produce viable offspring. These two features – self incompatibility and reciprocal herkogamy – complement each other. While self incompatibility by itself is enough to ensure the absence of inbreeding, it does not actively promote outcrossing, which may lead to a failure to cross pollinate resulting in few offspring. Reciprocal herkogamy reduces this issue, increasing the viability of self incompatibility systems in lower resource environments (Ganders, 1979). Conversely, self incompatibility systems prevent herkogamous species from losing the adaptation and reverting to a non herkogamous state by actively preventing inbreeding, maintaining the reciprocal arrangement of pollinating organs (Charlesworth and Charlesworth, 1979b; Zhou *et al.*, 2017; Shou *et al.*, 2019). This means that reciprocal herkogamy and self incompatibility are often inherited together, making heterostylous populations.

Chapter 1.2: Evolutionary History of Heterostyly

Heterostyly is somewhat unique among floral traits for promoting outcrossing, in that it has evolved independently in a large number of unrelated species, through convergent evolution. At least 28 separate families contain species that express heterostyly (Barrett, Jesson and Baker, 2000), either in distylous or tristylous forms (where there is a medium length of anther and style in addition to the long and short lengths), and in addition to those species there are several others whose genetics suggest that they reverted to a monostylous state from previously expressed heterostyly (Zhou *et al.*, 2017, Ruiz-Martin *et al.*, 2018). There is some debate around which species develop heterostyly, and how they develop into heterostylous expression from a monostylous state.

Firstly, the debate around which species develop heterostyly. Flowers with certain features are much more likely to develop heterostyly than others (Barrett and Shore, 2008). It seems that only flowers with pollen located relatively deeply within the flower can develop heterostyly, where pollen and nectar is positioned as to necessitate a pollinator having to reach into the flower rather than a pollinator being able to obtain everything on the surface. Flowers with open dished corollas or exposed nectar are much less likely to become heterostylous, likely because this arrangement creates a wide variety of possible positions pollinators could settle in, meaning that expressing reciprocal herkogamy would not guarantee reciprocal pollen positioning and so reciprocal pollination. In addition, it seems that the majority of heterostylous flowers are actinomorphic, or flowers with a radial symmetry, anthers and petals arranged in a circular arrangement around a central style. Zygomorphic flowers, with only one plane of symmetry, anthers and petals arranged opposite each other across a line, do not seem as likely to develop heterostyly. And flowers with a very large number of stamens are also less likely to become heterostylous – it seems that sexual organ positioning is only important insofar as it controls pollinator positioning, and with too many stamens that pollinator position cannot be ensured. The evolutionary history of the species has a lesser influence on heterostyly development than might be expected – how closely it is related to other species which express either heterostyly or homostyly is less important floral structure, as seen in the distribution of heterostyly in the genus *Naricissus* (Santos-Gally, Gonzales-Voyer and Arroyo, 2013). Certain families do have more heterostylous species than others, and if a plant is closely related to a heterostylous plant it is more likely that it will also have the physical conditions necessary for heterostyly to emerge. However, heterostyly has evolved independently and spontaneously on multiple occasions (Lloyd and Webb, 1992a), in plant groups with no other evidence of heterostylous expression – it seems that an evolutionary history of heterostyly is not a prerequisite for a species to express heterostyly. It almost seems as if physical features are a closer guide to likelihood of developing heterostyly than genetic similarity.

Secondly, the question of how heterostyly develops from homostyly. Historically, there have been two key schools of thought in this area. Lloyd and Webb proposed that heterostyly began with approach herkogamy, with the flowers in the initial stage consisting of long stamens and short styles (Lloyd and Webb, 1992a). This would be a relatively favourable arrangement for self-pollination, explaining how it could develop, but it would also favour breeding with any nearby plants with long styles. These are theorised to be homostylous long styled morphs at first, and the favourable outbreeding causes these long styled homostylous morphs to invade this population of long stamen and short styled morphs. This creates a similar evolutionary pressure incentivising the production of a short stamen and long styled morph, as this breeds even more favourably with the long stamen and short styled morph. This eventually results in reciprocal herkogamy, which is solidified by the development of a self-incompatibility system due to the evolutionary advantages of guaranteeing outcrossing. Thus, a population with only one stylar morph becomes heterostylous. However, Charlesworth and Charlesworth proposed a different model (Charlesworth and Charlesworth, 1979). In their model, the ancestral state is homostylous long styled plants, and the first heterostylous feature to develop was not herkogamy, but self incompatibility. A new pollen type, incompatible with the style of the same plant, arises to promote outcrossing, and if the rate of selfing and inbreeding depression is high enough in the overall population this new pollen type becomes established as an alternative plant morph. This system develops due to the benefits of incentivising outcrossing, but also means that the alternative morph plants have to breed further afield to other plants than they otherwise would. This results in a self-incompatible receptive morph developing, with the pollen of the common ancestor but a style suited to accept pollen from the self-incompatible pollen morph. With the self incompatibility system reinforcing the distinction between the two morphs, reciprocal herkogamy develops to increase pollination success between the two morphs. And so a homostylous population becomes heterostylous. It has been difficult to empirically determine which theory of evolution is most reflective of the actual method of heterostylous development. The genus Narcissus contains heterostylous species with distyly, stylar monomorphism and stigma height dimorphism (where the two morphs have the same length of stamens but different height stigmas) (Barrett, Lloyd and Arroyo, 1996). By tracking which species diverged when, it seems as if distyly evolves from stigma height dimorphism – however, this is a consistent factor in both models of heterostylous evolution. The key factor seems to be whether self incompatibility develops before reciprocal herkogamy, or vice versa. However, self incompatibility systems can be very variable (Shou *et al.*, 2019) – ranging from merely decreasing the chances of self pollination resulting in viable offspring to making self pollination an impossibility. As such, they can be more difficult to identify than the more obvious and straightforward herkogamy, which can make determining whether herkogamy or self incompatibility developed first difficult. However, analysis of the genus Salvia may provide an answer. There is one species in this genus, *Salvia brandegeei*, with reciprocal herkogamy and no self incompatibility system (Barrett, Wilken and Cole, 2000, Barrett and Shore, 2008). While it is possible that this species could have reverted from an ancestral state of herkogamy and self incompatibility, the existence of reciprocal herkogamy without self incompatibility still supports Lloyd and Webb 1992a. A more recent study has even challenged the core idea of the link between reciprocal herkogamy and self incompatibility (Ferrero *et al.*, 2012), with species in the genus *Glandora* found to display either reciprocal herkogamy or self incompatibility without the other. This would suggest that neither polymorphism necessarily leads to the other, refuting both current evolutionary theories and potentially requiring the development of a modern theory to cover this – however, it should be noted that this lack of link was only found with one particular type of self incompatibility system, and as such may not be applicable across species.The idea of reversion from a heterostylous state should be examined more closely. While heterostyly is a common trait that has evolved independently in a diverse range of species, indicating that it is a sufficiently advantageous trait to cause convergent evolution, the most common evolutionary change in families containing heterostylous species is not the emergence of heterostyly. Rather, it is the reversion from heterostyly to homostyly (Barrett and Shore, 2008). There are a few possible explanations for this, but the simplest is to consider the resource expenditure of heterostyly, especially paired with self incompatibility. While genetic diversity is undoubtedly a long term advantage for a species, in the short term self fertilization, or even asexual reproduction, is the easiest way for plants to reproduce, requiring minimum expenditure of resources (Stelzer, 2015, Yang and Kim, 2016). In harsh environments, where resources are scarce, pollinators are rare, and survival is uncertain, a system that disincentivises and possibly even actively prevents self fertilization is a significant handicap. As such, these plants are less likely to survive compared to homostylous species, creating an evolutionary pressure incentivising the transition from sexual reproduction to asexual reproduction, and so from heterostyly to homostyly (Yuan *et al.*, 2017). However, even if a species transitions from heterostyly to homostyly, it is still possible for it to reacquire heterostyly at a later date (Tippery and Les, 2011). If the environment becomes more favourable, or the species adapts to it in another way, the benefits of heterostyly in terms of genetic diversity are valuable enough for it to be reintroduced to the species (De Vos *et al.*, 2014). While heterostyly is a resource disadvantage, it can be argued to be a lesser disadvantage compared to other methods of ensuring outcrossing, incentivising its emergence. For instance, dichogamy is one such alternative method, where outcrossing is guaranteed by preventing self fertilization, by having the anthers deposit pollen and the style accept pollen at different times. This does guarantee outcrossing, but prevents a plant simultaneously depositing and accepting pollen, reducing pollination efficiency to a greater degree than heterostyly (Bertin and Newman, 1993). However, one transition that is rarely seen is from heterostyly to partial heterostyly – heterostylous morphs showing a significant number of homostylous mutants (Barrett and Shore, 2008). Heterostyly, despite being a term encompassing a wide number of traits, seems very conserved through multiple generations. To understand why, the genetics of heterostyly must be studied.

Chapter 1.3: Genetics of Heterostyly

The basic theory for the genetics of heterostyly is that it is a heterozygously inherited trait, with one of the morphs controlled by a dominant diallelic gene, known as the S locus. This was first found in *Primula sinensis* (Bateson and Gregory, 1905), where the short styled thrum morph is heterozygous dominant, Ss, while the long styled pin morphs is homozygous recessive, or ss. This model repeats across species, though in several it is the long styled pin morph which is regulated by the dominant gene (Barrett and Shore, 2008). In cases of heterostyly with three morphs, or tristyly, there are two relevant diallelic genetic loci to consider, each epistatic to the other (Barrett, 1993, Barrett and Shore, 2008, Arunkumar *et al.* 2017). One of the extreme morphs is still controlled by the S locus, where if the dominant S allele is present, this extreme morph is expressed. If the S locus is homozygous recessive, ss, the medium morph is then controlled by an epistatic gene known as the M locus. If the dominant M allele is present, either homozygous or heterozygous, ssMM or ssMm, the medium morph is expressed. If the M locus is homozygous recessive, mm, the extreme morph not controlled by the S locus is expressed.

However, heterostyly does not usually involve a single phenotypic change, but a variety of them. Pins and thrums of different species have different style lengths and anther lengths, with each of these likely controlled by separate genes, but they also have other differences. Pollen is a key factor (Ernst, 1955), with the size and structure of pollen grains differing between morphs, ensuring self incompatibility by only allowing pollen from different morphs to fertilize the ovum. This is likely controlled by a separate gene again. Self incompatibility mechanisms are associated with heterostylous morphs (Ganders, 1979), and while the genes controlling this can be shared with those of other areas – such as pollen size, or stylar structure, both of which prevent pollen from the same plant from being able to create a pollen tube on the style – there are a variety of self incompatibility methods across species which may be controlled independently of the previously specified genes (Takayama and Isogai, 2005). This all implies that heterostyly is the result of multiple allelic variations in separate genes, rather than each morph being a different allele of the same gene. However, the inheritance pattern of heterostyly is more consistent with a single gene controlling each morph than with multiple. Heterostyly morph differences are tightly conserved through multiple generations, with minimal recombination of traits. It is rare to find morph hybrids in the wild, with a plant containing the anthers and pollen of a thrum morph but the style of a pin morph, or similar (Ganders, 1979, Kappel, Huu and Lenhard, 2017). While some of these hybrids could face problems with reproduction due to contradictory self incompatibility systems, which would explain their absence, this explanation does not hold for all possible hybrids trait combinations. While the chance of any individual genetic recombination event is minimal, the number of genes involved in heterostyly still implies that if the genes responsible were evenly distributed throughout the genome, chance would dictate that they would become separated between chromosomes due to random crossover within a minimal number of generations (Brennan, 2017). This would result in the many separate hybrid phenotypes arising within the population, such as long and short homostyles, where the anthers and styles are both either long or short. Given the rarity of these phenotypes and how strongly the di or tristylous system is preserved, across all heterostylous species, it seems that the genes controlling heterostylous traits cannot be evenly distributed throughout the chromosome. But given the existence of hybrid morphs in laboratory conditions (Ernst, 1955, Labonne, Tamari and Shore, 2010), and the number of traits associated with heterostyly, heterostyly cannot be controlled by a single gene. To reconcile these two ideas, a theory was proposed that the genes controlling all aspects of heterostyly are tightly linked in a small area of the chromosome which acts as the inheritable unit for heterostyly - a supergene (Lewis and Jones, 1992).

A supergene is a small number of genes located in the same area of the chromosome, which are so closely linked that they are inherited together as a single genetic unit, and display minimal crossing over between alleles between generations. This would allow the two floral morphs of heterostyly to be preserved through generations, as the alleles responsible for each remain linked. The model of the supergene was developed specifically as a theory to explain the inheritance of heterostyly in *Primula* (Ernst, 1933), explaining how multiple polymorphisms could be inherited as a single locus. *Primula vulgaris*, or primrose, was the model species for heterostyly since even before Darwin, and classical genetics studies found that it indeed contained a supergene complex controlling the inheritance of heterostyly (Ernst, 1955). Later research found evidence for supergenes in unrelated species, controlling similar complex phenotypes which are inherited as a single gene, supporting its use as a theory to explain the inheritance of heterostyly. One example of this would be controlling Batesian mimicry in butterflies *Papilio memnon* (Clarke, Sheppard and Thornton, 1968) and *Papilio polytes* (Clarke and Sheppard, 1972). This supergene complex became the model for all other research into heterostyly, with research into the genetics of heterostyly in other species generally beginning with the theory that they would be similar to *Primula*. After all, despite the multiple diverse evolutionary origins of heterostyly, there is reason to presume the genetic mechanisms behind it are somewhat conserved. Similar collections of traits evolve independently in many different families, and are inherited in a similar manner, and so this phenotypic similarity is likely paralleled by a genotypic similarity. As *Primula vulgaris* is the most well studied heterostylous species, and the three gene supergene has been the most robust model for the genetics of heterostyly in *Primula vulgaris*, it became the default model for heterostyly in multiple species.

Chapter 1.4: Heterostyly in Primula

There is no debate that the genus *Primula*, and specifically the species *Primula vulgaris*, has historically been the model species for studying heterostyly. Several of the earliest known representations of heterostyly show heterostyly in primrose, or *Primula vulgaris* (Gilmartin, 2015), and heterostyly was first formally classified using *P. vulgaris* as a reference, by Darwin in 1862 (Darwin, 1962). Why *P. vulgaris* was the first heterostylous species identified and classified is a matter of debate – it could just be a matter of chance, of course. But primrose has always been a very prolific species in England and Europe, a reasonably robust plant suited to the environment with both aesthetic and practical value (used in traditional medicine and as food) (Ozkan *et al.* 2017). Given that the science of botany was most well practised in England in the nineteenth century, and that primrose displays a very visible form of heterostyly (with prominent anthers and styles, and other morphological features marking the difference between the morphs), it follows that *P. vulgaris* became the first species to be identified as heterostylous.

As *Primula vulgaris* became the model species for heterostyly, the genetics of heterostyly within *Primula vulgaris* became one of the best researched of any species, and a basic model for them had been defined as far back as 1933 (Ernst, 1933). A cross between pin and thrum flowers was shown to produce equal amounts of pin and thrum morphs in the offspring, so the assumption was that heterostyly in *Primula vulgaris* was heterozygously inherited, as a single dominant diallelic gene, known as the S locus, determining which morph the flower takes. The thrum morphs had one dominant and one recessive copy of the S locus, Ss, while the pin morphs had two recessive copies, ss. This cross would produce equal proportions of pin and thrum morphs in the offspring, consistent with the observed population. The main issue with this model, as previously stated, is that heterostyly encompasses multiple traits (which can be separately inherited in different species) that it is extremely unlikely that it is controlled by a single gene. As such, the proposed model was that heterostyly in *Primula vulgaris* is controlled by a tightly linked set of three genes which are inherited together as if they were a single gene - the *Primula* supergene. The *Primula* supergene, or the S locus, was determined to consist of three distinct diallelic genes, all neighbouring each other and contained within <500kbp of the chromosome. These three genes were termed the G locus, or gynoecium, responsible for stylar size and structure, influencing self incompatibility, the P locus, or pollen, responsible for pollen size and structure and also influencing self incompatibility, and the A locus, or anther, responsible for anther size and structure (Ernst, 1955, Lewis and Jones, 1992). Given that homostyles produced by recombining one allele of the A locus with the S locus from a different morph were self-compatible, this implies that the A locus had no influence in self incompatibility. There was some debate over this model of the supergene, however. While these were the only three genes which could be separated and influenced, based on which hybrids were grown from selective crossing of parts of the S-locus, some argued that the number of discrete traits that would need to be controlled by each of these three genes implied the existence of additional genes associated with heterostyly. Some theories proposed larger S-loci, consisting of up to 7 (Dowrick and Pamela, 1956, Kurian and Richards, 1997) or even 9 (Richards, 2003) separate genes, or that the S-locus also codes for transcription factors which influence the transcription of genes elsewhere in the chromosome (Barrett and Shore, 2008). More recent research accepted the idea that the S locus consisted of the three diallelic GPA genes, but also investigated the idea that other diallelic genes controlling aspects of heterostyly could be linked to the S locus – not necessarily so tightly linked that there was no recombination between different morphs, as is the case with the supergene, but linked nonetheless (Kappel, Huu and Lenhard, 2017). A large variety of genes were identified as linked to the S locus, including genes controlling flower pigment (Gilmartin, 2015), sepal development (Li *et al*. 2008), leaf and flower development (Cocker *et al.* 2015) and circadian rhythm (Li *et al*. 2007). Other entirely different phenotypes can co-segregate with the S locus, such as the *Hose in Hose* mutation (Li *et al*. 2010) where sepals are converted to petals, giving the appearance of one flower inside another. This mutation, found to be caused by upregulation of two developmental MADS-box genes, *GLOBOSA* and *DEFICIENS*, and tied to a retrotransposon insertion, has been proposed to directly flank the S locus. These S locus linked genes can be a reasonably significant distance away from the S locus, up to 2 cM, certainly located within the same area but not directly flanking, and as such should see more recombination events than they seem to. It has been proposed that there is a local suppression of recombination in the region of the S-locus (Kappel, Huu and Lenhard, 2017), explaining the number of linked genes and the stability of both the heterostylous and *Hose in Hose* phenotypes. While this has not been demonstrated in *Primula*, there is evidence for this in other dimorphic species such as *Ipomoea trifida* (Tomita *et al.* 2004), so it is a possibility. Regardless, all of these studies accept the basic idea of the *Primula vulgaris* heterozygous diallelic supergene, demonstrating that this model has been widely accepted as the basis for heterostyly in *Primula vulgaris*, and by implication the default for the genetics of heterostyly across species. However, this genetic model of heterostyly has been challenged in recent years.

Recent research has suggested a slightly different genetic model for heterostyly in *Primula vulgaris*. The theory was that the two morphs of *Primula vulgaris* were heterozygous, caused by a dominant haplotype of alleles of the three genes composing the S locus supergene in the thrum morph, although the precise genetic architecture of the S locus was somewhat unclear. Despite the fact that multiple alleles were involved in heterostyly, they were tightly enough linked that they were inherited as a single gene, causing heterostyly to be inherited heterozygously. Recent evidence (Li *et al*. 2016, Huu *et al*. 2016) has contradicted this, suggesting that the inheritance of heterostyly in *Primula vulgaris* is, in fact, hemizygous. Heterozygous inheritance involves two different alleles of the same gene – one dominant, one recessive. Two copies of the gene are present on a pair of chromosomes, and if one of the genes has the dominant allele the phenotype is expressed – in this case the thrum morph. Hemizygous inheritance is different, in that if a trait is hemizygous the gene encoding it is only present on one chromosome. The other chromosome does not contain a complementary gene. The most well-known example of hemizygous inheritance is that of sex determination. In mammals, for example, there is a genetic region called the sex determining region, or Sry, which is only found on the Y chromosome (Sinclair *et al*. 1990, Gubbay *et al.* 1990). If the Sry is inherited, usually by the Y chromosome being inherited, the offspring will be male, whereas if the Sry is not inherited, usually by two X chromosomes being inherited, the offspring will be female (Koopman *et al.* 1991). There is no allele for the Sry on the X chromosome – the male phenotype is solely reliant on the presence or absence of the Sry. In this case what was thought to be the recessive allele of the S locus, that encodes the pin morph, is in fact just the absence of the gene encoding for the thrum morph. There is a section of the S locus which is only present in thrum morphs. By sequencing the *Primula vulgaris* supergene region, Li *et al.* 2016 found that the S-locus composed a 455 kb genomic sequence – however, only 177kb of this sequence was present in both morphs. They identified a 278kb region encoding for 5 genes - CCM^T (Conserved Cysteine Motif), GLO^T (*GLOBOSA* like) CYP^T (Cytochrome P450), PUM^T (Pumilio-like) and KFP^T (Kelch repeat F box) – which was only present in the thrum morph. It seems as if the presence or absence of these 5 genes was what determined which heterostylous morph the plant expressed. To support this theory, they found that loss of GLO^T results in the development of short style homostyles (short anther, short style) and that the loss of CYP^T is associated with long style homostyles (long anther, long style). In other words, it seems as if the *GLOBOSA* like thrum specific gene GLO^T was responsible for the long anthers of the thrum morph, while the cytochrome P450 thrum specific gene CYP^T was responsible for the short styles. Huu *et al.* 2016 expanded on the role of CYP^T, supporting this theory. They found that the enzyme encoded by CYP^T was responsible for the degradation of brassinosteroids. Brassinosteroids are a family of phytohormones and have previously been shown to play a role in regulating cell-elongation, supporting the theory that a gene encoding for their degradation would result in the shortening of the plants style. Further to this, Huu *et al.* 2016 showed that addition of brassinosteroids to thrum morphs lengthens the styles, results in long homostyles, corroborating Li *et al.* 2016 on the role of CYP^T and confirming that a hemizygous gene has a key role in heterostyly. A later complete assembly of the *Primula vulgaris* genome (Cocker *et al*. 2018) confirmed the hemizygous nature of the S locus. This finding has significant implications for future research into the genetics of heterostyly. The model of the heterozygous S-locus in *Primula vulgaris* has acted as the default model for heterostyly across species, with similar S-loci being identified in many other heterostylous species. If the *Primula vulgaris* S-locus is actually hemizygous, this could mean that other seemingly heterozygous S-loci are also hemizygous.

Chapter 1.5: Heterostyly in other genera

While the *Primula* genus is the focus of most study of heterostyly, this is not to say that the genetics of heterostyly in other genera has not been explored. However, it does imply that the genetics of heterostyly in other species has been studied through the lens of the *Primula* diallelic S-locus supergene model. Given the recent evidence that this heterozygous S-locus is in fact hemizygous, this could mean that previous studies of heterostyly in other genuses should be re-evaluated in this light. Do the previous studies on genetics of heterostyly in other species provide evidence the hemizygous nature of the *Primula* S-locus, implying that this could be the nature of the S-locus supergene across species, or do they support the previous heterozygous model more, implying that the hemizygous S-locus is unique to *Primula vulgaris*? Given that heterostyly is present in 28 separate families, there are many different genera which could be considered for this. However, recent studies have provided key insights into two genera in particular – *Turnera* and *Fagopyrum*.

Turnera is in many respects an ideal genus to study heterostyly in. Though the open flowers characteristic of *Turnera* are in contrast to the typical tubular flowers of other heterostylous species, the genus displays many floral characteristics ideally suited to heterostyly – radial symmetry, depth pollination and similar. And indeed, the majority (80%) of the 128 species in the genus are heterostylous (Labonne, Goultiaeva, and Shore, 2009), displaying distyly with dominant thrum and recessive pin phenotypes. Even the homostylous species in the genus are thought to have reverted from an ancestral heterostylous state (Truyens, Arbo and Shore, 2005). As such, much study has gone in to characterising the S-locus in multiple different species of *Turnera*. Initial studies focused mainly on proteomics, identifying proteins that were differentially expressed between the pin and thrum morphs (Athanasiou and Shore, 1997). Through this, it was found that there are two key proteins which are only expressed in the dominant thrum morph – polygalacturonase and α dioxygenase (Athanasiou *et al.* 2003, Khosravi *et al.* 2004). This suggests candidate genes for the S-locus, but the genomics did not support this hypothesis. The gene corresponding to polygalacturonase was linked to the S-locus, but was not closely linked enough to be a part of it, while the gene corresponding to the α dioxygenase was completely unlinked to the S-locus, suggesting that the differential expression seen in the thrum morph is the result of secondary control. Later studies approached the subject from a genomics angle, trying to map and characterise the genes of the S-locus. In a key study (Labonne, Goultiaeva and Shore, 2009), two heterostylous species of *Turnera, Turnera subulata* and *Turnera krapovickassi*, were sequenced, constructing a high resolution map of the S-locus and identifying three key genetic markers, one closely linked to the S-locus and two that were S-locus candidate genes. The candidate genes identified were a sulfotransferase, *TkSTI* (*Turnera krapovickassi* sulfotransfersase 1) and a retrotransposon, *TsRETRO* (*Turnera subulata* non-LTR retroelement). In addition, a gene encoding an acetyltransferase, *TkNACE* (*Turnera krapovickassi* N-acetyltransferase), was found to be closely linked to the S-locus. This study included construction of a library of BAC contigs, and led into a study (Labonne, Tamari and Shore, 2010) characterising X-ray deletion mutants of *Turnera subulata*, using genetic markers to determine which deletions resulted in which phenotype and so characterising the function of the genes (assuming a typical *Primula* GPA diallelic S-locus structure). The deletion mutants produced from an initial population of entirely short styled plants consisted of several long styled mutants, a long homostyle and a short homostyle. The long homostyle was thought to be produced through deletions of G and P genes, controlling style length and pollen structure, while the long styled mutants were thought to result from a deletion of the entire S-locus. The obvious conclusion is that the short homostyle resulted from just a deletion of the A gene, controlling anther length. However, the incompatibility behaviour and the multiple absent markers were consistent with the long styled morph, implying that this short homostyle resulted from the deletion of the entire S-locus as well and is effectively just a long styled morph with an unusually short style.

While these studies provided some key information to understanding heterostyly in *Turnera*, the most important study of this subject was only conducted recently (Shore *et al*, 2019) which combined the approaches of both of these previous studies to provide the most detailed view of heterostyly in *Turnera* to date. The genome of *Turnera subulata* was sequenced using Illumina shotgun sequencing and scaffolds constructed, most importantly providing high quality scaffolds of the S-locus region. A BAC library was assembled of the genes only present in the thrum morph, representing a dominant S haplotype, and sequenced. The regions identified as deleted in the previously studied deletion mutants were mapped, identifying candidates for the G, P and A gene equivalents in *Turnera subulata*, and finally other heterostylous species were compared (*Turnera concinna, grandiflora, joelii* and *scabra*) to see if the candidate genes were present in the same morph across the genus. In addition, some homostylous mutants were found to contain only some of the genes as well. It was ultimately found that the S-locus of *Turnera* is, like *Primula*, a hemizygous supergene. The thrum morph contains three genes (and two inversions) which are absent from the pin morph – *TsSPH1* (*Turnera subulata* S protein homolog 1, expressed in filaments and anthers), *TsBAHD* (*Turnera subulata* BAHD acetyltransferase, expressed in pistils) and *TsYUC6* (a flavin monoxygenase implicated in auxin biosynthesis, expressed in anthers). All of these genes were present in the thrum morph and absent from the pin morph of all species tested, implying a conserved S-locus structure and complete hemizygosity across the genus. Long homostyle mutants were found not to contain *TsBAHD*, implying it may be the G locus equivalent, while a short homostyle mutant was found to not express *TsSPH1*, implying it may be the A locus equivalent. As such, it was proposed that *TsBAHD* is involved in style shortening, likely through a similar method to CYP734A50, inactivating brassinosteroids, supported by a later study (Henning, Shore and McCubbin, 2020). *TsYUC6* is likewise proposed to be involved in pollen structure (due to its role in auxin biosynthesis), and *TsSPH1* involved in filament lengthening, with incompatibility likely determined by both *TsBAHD* and *TsYUC6*. This 3 locus supergene, aside from the hemizygosity, bears a striking resemblance to the classic *Primula* GPA model – ironically making the *Turnera* S-locus more similar to the classic *Primula* S-locus than the current *Primula* S-locus. This complete characterization of the *Turnera* supergene is a major advance in the field of heterostyly, and a significant piece of evidence supporting the theory that the heterostyly supergene is hemizygous across unrelated species.

Another genus involved in a significant advance in the field of heterostyly research is *Fagopyrum*. *Fagopyrum* is a notably different genus from *Turnera*, containing a relatively small number of 28 species as opposed to *Turnera*'s 128, native to Southern Asia rather than South America, and with small clusters of deeper flowers as opposed to *Turnera*'s larger, more isolated open dished flowers. However, heterostyly is also common in *Fagopyrum* where, although several homostylous species exist, they can contain a significant proportion of heterostylous mutants (Ohnishi and Zhou, 2018), such as *Fagopyrum gracilipes,* or have likely reverted from an ancestral heterostylous state, such as *Fagopyrum homotropicum,* which sequencing revealed contained an altered S locus (Matsui *et al.* 2003). The two most well studied, and indeed most wide-spread due to their use as crops, species of *Fagopyrum* are the homostylous *Fagopyrum tataricum* and the heterostylous *Fagopyrum esculentum*, the latter of which being extremely important in the field of heterostyly research. Studies of this species had identified molecular markers for the S locus (Aii *et al.* 1998), differentially expressed proteins (Miljus-Dukic *et al.* 2004) and several transcripts unlinked to the S locus but restricted to the thrum morph (Takeshima *et al*. 2019) (and so likely secondarily controlled by an S locus linked gene). But a key advance in the understanding of *F. esculentum*'s S-locus came with the identification of 4 thrum morph associated transcripts, one of which was exclusively expressed in the thrum morph even across species (Yasui *et al.*, 2012). These transcripts were named SSG1-4. However, analysis of a chimeric plant with both short styled and long styled sections discovered that SSG1 and SSG4 were also present in the long styled sections of the plant, implying that they were not true thrum specific genes. Similarly, Southern blot analysis found that SSG2 was also present in some long styled plants, leaving SSG3 as the one true thrum specific gene. As it displayed homology to the *Arabidopsis thaliana* gene Early Flowering 3, or ELF3, it was named S-ELF3 – S-locus Early Flowering 3. Even beyond *Fagopyrum esculentum*, in different species in the genus, S-ELF3 was found to be present in all thrum morphs and absent in all pin, making it a strong contender for an S-locus candidate gene. In addition, two separate self-compatible long homostylous *Fagopyrum* species were found to contain deactivations of S-ELF3, implying that S-ELF3 is likely similar to the hypothesised G locus in the classic *Primula* S-locus supergene, controlling style length and self-incompatibility. A 610kb region around S-ELF3 was sequenced using previously developed BAC libraries (Yasui *et al.* 2008) and was found to contain SSG2, explaining its partial linkage to the thrum morph, and many transposable elements and repetitive sequences, consistent with the type of genomic degradation expected in an area of the genome where recombination is suppressed (e.g. the S-locus). This support for buckwheat's S-locus being hemizygous was strengthened a few years later, when a complete draft genome of *Fagopyrum esculentum* was published (Yasui *et al.*, 2016). This draft genome found over 5.4Mb of DNA sequence which was present in the thrum morph of the plant and absent from all pin morphs sequenced, indicating a hemizygous S locus in line with that of *Primula* and *Turnera*. This hemizygous region contained 32 predicted genes and 75% transposable elements, consistent with a region with suppressed recombination (such as the S locus). It is worth mentioning, however, that DNA was only sequenced from a single short styled (thrum) plant in this study, with the data from long styled plants coming from RNA sequencing. It is possible that the long styled genotype displayed could arise from epigenetic factors, or genetic alterations to a heterozygous S locus allele that would prevent transcription. Research is proceeding assuming hemizygosity (Matsui *et al.*, 2020), however, which is supported considering the combination of the draft genome and the S-ELF3 thrum specificity. Although the S locus has not been proven to be hemizygous to the same stringency as in *Primula* and *Turnera*, research in *Fagopyrum* has still provided more support for the hemizygous model of the S-locus.

There are a large number of genera beyond *Primula, Turnera* and *Fagopyrum* with heterostylous species, however few show significant research into the molecular genetics behind heterostyly. Most of the research done in these genera has been to investigate compatibility mechanisms on a population level, floral morphology and the effect of heterostyly on pollen dispersal and commercial yield, such as in *Lythrum* (Brown and Mitchell, 2001, Costa *et al.* 2017), *Oxalis* (Costa *et al.* 2013, Weber *et al.* 2013, Ferrero *et al.* 2015), *Solanum* (Kowalska, 2001, Srinivas, Jayappa and Patel, 2016, Das *et al.* 2017) and *Averrhoa* (Wong, Watanabe and Hinata, 1994), focusing more on the reproductive impact of heterostyly in a particular species rather than looking into the genetics behind it. When genetic studies have been done they often take more evolutionary or ecological perspectives, tracking the evolution of heterostyly in a species or the spread of heterostylous plants in an environment rather than aiming to characterise the genes which cause heterostyly.

However, that is not to say there has been no research into the molecular genetics of heterostyly outside the three previously discussed genera. In *Lithospermum multiflorum*, comparative transciptomic analysis (Cohen, 2016) identified several differentially expressed genes between thrum and pin morphs. Interestingly the genes also showed different expression patterns depending on floral location (style vs. corolla) and developmental stage. In early stages of development few genes were differentially expressed, with the difference between thrum and pin less pronounced and more dependent on location of differential expression (thrum displayed more differentially expressed genes in the gynoecium compared to the rest of the plant, while pin showed differentially expressed genes localised more to the corolla and androecium), but in later stages a reciprocal pattern of differential expression (with genes upregulated in the thrum morph downregulated in pin, and vice versa) did emerge. Characterisation of the genes differentially expressed was limited, but those associated with growth and development showed more pronounced expression differences in the early developmental stages, while in later developmental stages the differentially expressed genes were more closely associated with physiological functions, notably including stress responses. This has some support in other genera, such as the phytochrome system potentially being involved in heterostyly in *Turnera* (Henning, Shore and McCubbing, 2020). This difference in differential expression between thrum and pin morphs being dependent on developmental stage is also supported by a proteomic analysis of *Solanum melongea*, or eggplant (Wang *et al.*, 2017). This analysis found 57 notably differentially expressed proteins between thrum and pin morphs in the pistils of *S. melongea* at early developmental stages, compared to 184 proteins identified at later stages. Characterisation of these proteins broadly supported the pattern observed in *Lithospermum multiflorum*, although notably upregulated proteins associated with senescence and cell death were also identified in thrum morph flowers. These studies may indicate S-locus candidate genes for these species, or even represent a more universal pattern of S-locus gene function across species.

Tristyly is a relatively understudied form of heterostyly from a genetic perspective, with very few modern studies aiming to characterise the S-loci involved in this unique floral morphology. One notable exception to this is a study of *Eichornia paniculata* (Arunkumar *et al.* 2017), aiming to sequence the genome to understand the genetic architecture of the M locus (the tristylous counterpart to the S locus – while S locus dominance induces a short styled phenotype, if it is recessive but the M locus is dominant a medium styled phenotype is seen instead). By crossing semi homostylous (where some of the stamens were the same height as the stigma, but not all of them) self-compatible parent plants with both medium and long length styles, a population was constructed that segregated for the dominant M locus. Genetic sequencing revealed over 10Mb of sequence that co-segregated with the M locus, likely containing over 300 genes. There is less evidence of reduced recombination than that found in other mapped S loci – it might be due to a more recent evolutionary divergence point, or it could be that the M locus in *Eichornia paniculata* is not the assumed supergene, but rather a small number of pleiotropic genes, possibly distributed across a wider area (after all, only 278kb of the *Primula* S-locus co-segregated with the S morph, much less than the >10Mb seen here). Whether this is the case, or if it is whether this structure is isolated to this species or representative of tristyly genetics more broadly, remains to be seen.

And finally, one study which may shed a significant amount of light on the molecular genetics of heterostyly was not even conducted on a naturally heterostylous species. In a study primarily aimed at characterising the genetics of stress responses in *Arabidopsis thaliana* (Suzuki *et al.*, 2014), it was found that overexpression of the VP1 transcription factor could induce a floral phenotype very similar to a typical pin phenotype. While *Arabidopsis* is normally self-compatible and usually has flowers more consistent with a long homostylous morph, it was found that a mutation of the gene transcribing the VP1 transcription factor lengthened the styles and shortened the anthers of the flower while also inducing self incompatibility, consistent with a typical pin phenotype. The specific mutation induced was shown to cause a loss of DNA binding activity and thought to affect degradation rates of the mutated protein. Later degradation was found to be similar to the non-mutant background. It was found that if this mutation was induced in a plant which contained a deletion for ABI5, the severity of the pin phenotype was reduced. Given both VP1 and ABI5's roles in mediating abscisic acid (ABA) signalling, this seems to imply ABA has a role in controlling heterostyly-like phenotypes in *Arabidopsis*. The pin phenotype is also more consistently expressed in the main stem than the secondary stem, where a previous study found that another component of ABA biosynthesis was expressed at a lower rate, implying that reduction of ABA results in the pin phenotype. Why this could be is unclear, as no other study has found a link between ABA and heterostyly, but it is notable that ABA typically has a negative regulatory link with brassinosteroids (Wang *et al.*, 2018). This link has been shown to be mediated through ABI1 and 2, so ABI5 may also have a role in this pathway, explaining why an *abi5* mutant did not express the pin phenotype. It has been shown that decreased expression of brassinosteroids in *Primula* results in styles shortening and the thrum phenotype being expessed, so the phenotype shown in this study could be due to decreased ABA expression resulting in increased brassinosteroid expression and so resulting in the opposite phenotype to the thrum morph being expressed – the pin morph. However, this would give brassinosteroids more overall control of heterostyly than previously found, as they had previously been thought to have little effect on anther length and solely affect style length. Whether ABA is truly a component of the molecular genetics of heterostyly or whether the phenotype observed is simply a result of ABA inducing a secondary effect in an actual component of heterostyly, it still provides an interesting piece of data from an unexpected source. It might be that more research needs to be done on the role of ABA in heterostyly.

Chapter 1.6: Heterostyly in Linum

It may seem that this is the extent of research into the genetics of heterostyly done up to this point, limited mainly to three key genera. But there is one more genus which cannot be ignored when it comes to studying heterostyly – *Linum*. *Linum* is a large genus consisting of many species of flowering plants, with species growing native in Europe, Africa and America. The most well-known species of *Linum* would be *Linum usitatissimum*, or common flax (Jhala and Hall, 2010). *L. usitatissimum* is commonly grown worldwide due to its commercial properties, with its stalk being used to make linen and its seed oil being extracted as linseed oil. A *Linum* plant is typically characterised by thin, tough stalks with bright flowers, often blue or yellow. These flowers usually consist of 5 petals, arranged radially around the 5 anthers and styles creating a dished corolla. And while *Linum usitatissimum* may be homostylous, this is not representative of the genus as a whole. Heterostyly is one of the defining features of this particular genus.

Heterostyly within *Linum* has been observed and catalogued for some time, and in some respects *Linum* can be seen as a genus which defined the study of heterostyly. *Linum grandiflorum* and *Linum perenne* were two of the first heterostylous species formally categorised by Darwin (Darwin, 1864, Darwin, 1877), in his papers defining heterostyly and putting forward the idea of heterostyly as an adaptation promoting outcrossing. As such, *Linum* has historically been one of the primary subjects for studying heterostyly, with studies of the genus demonstrating the link between reciprocal herkogamy and self incompatibility in heterostyly and defining the inheritance pattern of heterostyly as consisting of a single diallelic locus with heterozygous dominance. The genus contains many heterostylous members, with a recent taxonomy (Ruiz-Martin, 2018) finding that ~45% of a representative sample of the genus displayed some form of heterostyly. Even in the monostylous species, only 40% of these were purely homostylous, with the remainder displaying some degree of herkogamy.

The forms of heterostyly in *Linum* are somewhat diverse – the majority do display the typical distylous morphs, but there are several alternative models. For instance, anther positioning in *Linum* is usually less pronounced than in other genera, as though it is usually present the difference in anther lengths between morphs is significantly less than the difference between style lengths (Darwin, 1877, Heitz, 1980, Ruiz-Martin *et al.* 2018). This can be seen to an extreme in *Linum grandiflorum*, which has been classified as having no anther difference between morphs, displaying stigma-height dimorphism as heterostyly – a notable divergence from the typical distylous model of heterostyly. There is also a number of species displaying some degree of tristyly, such as *Linum hirsutum*, which shows two different anther lengths and three style lengths between three morphs.

The most notable divergence from typical distyly is probably the heterostyly shown in *Linum suffruticosum*, which displays a rare 3-dimensional form of distyly (Armbruster *et al.*, 2006). While typical distyly only shows reciprocal floral organ positioning on a single vertical axis, with differences in anther and style height defining typical reciprocal herkogamy, *Linum suffruticosum* shows reciprocal organ positioning along all three spatial axes. Alongside the vertical differentiation, anthers and styles are also differentiated on radial axes (with anthers being positioned on the outer rim of the flower with styles positioned on the inner, or vice versa) and on the longitudinal axis of each organ (with anthers and styles growing to twist towards either the inside or outside of the flower). This last differentiation is most important, with it resulting in pollen being deposited on and collected from either the top or bottom of the pollinator visiting, a much more pronounced difference than that found in typical distyly (where pollen is deposited on either the upper or lower part of the underside of the pollinator). This key difference in pollen deposition should make this form of heterostyly even more effective than typical distyly at ensuring outcrossing, arguably making this the ideal form of heterostyly. In addition, this form could only arise in a species with a single pollinator (as the floral organs are so perfectly adapted to this pollinator) where outcrossing can be optimised. This implies that heterostyly may be very strongly selected for in *Linum*.

This strong selection may be limited, though, as the evolution of heterostyly in *Linum* is not a smooth road (Ruiz-Martin *et al.*, 2018). It is unclear whether the single ancestor of all *Linum* is stylar polymorphic or monomorphic, but when the genus is divided into two major evolutionary clades things become a little clearer. In clade A, which encompasses such key heterostylous species as *Linum grandiflorum* and *Linum perenne* alongside major homostylous species such as *Linum usitatissimum* and *Linum bienne*, the most recent common ancestor is polymorphic. It appears that there have been 3 clear reversions from heterostyly to homostyly (and monomorphic approach herkogamy) in this clade, with transitions from homostyly to heterostyly being negligible. While this is not surprising in a wider context, as reversions from heterostyly to homostyly are more common than the reverse due to the greater possibilities of the polymorphism becoming evolutionarily unfavoured in a harsher environment (Yuan *et al.*, 2017), it does show that the idea of heterostyly being selected for across all of *Linum* is an overgeneralisation. Homostyly is definitely common in the genus, with the most well-known and widespread species of *Linum* (*Linum usitatissimum*) being homostylous. However, on smaller scales, heterostyly can still be seen as heavily favoured in *Linum*. When *Linum* is subdivided between perennial and annual plants, a marginally significant correlation can be seen between perennial plants and heterostyly, which suggest that homostyly is favoured in annuals. Also, in the second of the two major evolutionary clades, clade B, the ancestral state is monostyly (in the form of approach herkogamy). There are two key transitions from this ancestral state to heterostyly in this clade, and while these may be accompanied with some minor reversions from heterostyly to homostyly, they are still very notable. Most importantly, these transitions from homostyly to heterostyly resulted in a significant proportion of South African *Linum* species displaying heterostyly, despite these species being more closely related to American homostylous *Linum* than European heterostylous *Linum*. In other words, heterostyly in South African *Linum* seems to have evolved in parallel to European *Linum*, possibly due to the similarity in climates (heterostylous European *Linum* being limited to the Mediterranean). Regardless, this parallel evolution does support the earlier idea that, in certain conditions, heterostyly in *Linum* is very strongly favoured.

Despite this background, there is relatively little research into the genetic mechanisms behind heterostyly in *Linum*. The only species analysed so far, *Linum grandiflorum* was proposed to have a heterozygous model of heterostyly based on a single diallelic locus, with this being developed into the standard S locus supergene model. There are multiple traits associated with heterostyly in *L. grandiflorum*, indicating either a pleiotropic single gene or a supergene at the S locus, and the generation of self-compatible homostyles from careful crosses implies recombination among the S locus, and so a supergene. The exact composition of this supergene was unclear, until a recent study (Ushijima *et al*, 2011) conducted a wide range of analyses (transciptomics, proteomics, subtraction profiling and sequence analysis) to identify twelve genes differentially expressed between the thrum and pin morphs, localised to the flower. The working theory was that any gene involved in the S locus would have to be solely expressed in one of the stylar morphs, and also completely localised to the relevant floral organs. Of these twelve genes, one fit those criteria. TSS1, or thrum style specific 1, is a 19 kDa protein solely expressed in the style of thrum morph flowers. TSS1 does not seem to contain either a signal peptide or a transmembrane domain, but does have features implying that it is secreted non classically and located extracellularly. It bears some homology to a collection of 20 hypothetical proteins in angiosperms, termed TSLs or TSS1 like proteins. However, where these proteins can be sorted into three key groups based on conserved regions, TSS1 does not fit neatly into any of these groups. The complete lack of expression outside of the thrum morph may imply that *TSS1*, similar to the *Primula* S locus, is hemizygous rather than heterozygous. However, another paper (Ushijima *et al.* 2015) showed a co-segregation of *TSS1* and flower colour, where the distribution of colours within the population implied a heterozygous inheritance pattern. This may imply that the *TSS1* region is (at least partially) heterozygous, and as the only current S-gene candidate in the *Linum* genus this could imply heterozygous inheritance of heterostyly throughout the genus. A similar connection can be seen between style morphs and flower colour in *Linum pubescens* (Wolfe, 2001), although the precise morph and colour association differed among flowers collected from different regions. Regardless of the problems assigning function, the complete lack of expression of *TSS1* outside of styles of thrum morph plants make it the best candidate for an S locus gene in *Linum grandiflorum*, and by extension any *Linum* species, found so far.

In addition to *TSS1*, several genes were found with a pronounced expression difference in thrum morphs compared to pin morphs (Ushijima *et al*, 2011). Although they were still expressed to a certain extent in pin morphs, preventing them from being true S locus genes, they are likely closely linked. There are three genes with a pronounced expression difference - *LgAP1*, *LgSKS1* and *LgMYB21*. These are *Linum grandiflorum* homologs of, respectively, aspartyl protease, SKU5 and an MYB transcription factor, after which they are named. *LgAP1*, or *Linum grandiflorum* aspartyl protease 1, was highly (but not exclusively) expressed in the style of thrum morphs. *LgAP1* showed homology to *CDR1* from *Arabidopsis*, another reproductive organ specific aspartyl protease. Some aspartyl proteases in plants have plant specific inserts (PSIs) to localise these enzymes to vacuoles but neither *CDR1* or *LgAP1* contain a PSI, implying that similar to *TSS1* they are accumulated extracellularly (confirmed experimentally). *LgAP1* consequently may be involved in the self incompatibility function of heterostyly, acting to disrupt pollen tubes. The role of proteases in self incompatibility was shown in *Fagopyrum*, where the addition of protease inhibitors prevented pollen tube rejection due to self incompatibility, so this is a likely function for *LgAP1* which would confirm its role in heterostyly. *LgSKS1*, or *Linum grandiflorum* SKU5 similar 1, shows increased expression in thrum pollen, and may show the exact opposite function to *LgAP1*. *LgSKS1* is homologous to *NTP303*, a protein expressed in tobacco, and suppression of *NTP303* in tobacco resulted in sterile pollen grains. The pollen grains could initiate pollen tube growth, but the growth was slowed to such a degree that it could not reach the ovary. Assuming that *LgSKS1* has a similar function in *Linum grandiflorum*, it becomes notable that such different proteins as *LgAP1* and *LgSKS1* are expressed in the thrum morph. This would imply that the method of self incompatibility could differ between thrum and pin morphs, which may remove the need for a recognition component typical of other self incompatibility mechanism. The final gene with notable expression differences between thrum and pin morphs is *LgMYB21*, or *Linum grandiflorum* MYB21 transcription factor. Highly expressed in thrum morphs, *LgMYB21* shows homology to *AtMYB21* and *AtMYB24* from *Arabidopsis*, and *MYB305* in tobacco. Overexpression of *MYB305* was found to reduce both pistil and stamen lengths in tobacco, while *AtMYB21* or *24* loss of function mutations showed reduced stamen lengths, implying that *LgMYB21* must play some role in floral organ growth regulation. And indeed, overexpression of *LgMYB21* in *Arabidopsis* did result in both short stamens and pistils. This non-specific floral organ regulation means that *LgMYB21* cannot be neatly assigned to either style or anther control, as per a traditional supergene model, but is more likely a downstream product showing differential regulation resulting from factors in the supergene. *AtMYB21* and *24* were found to be dependent on both jasmonate and gibberellin for expression (Cheng *et al.* 2009), giving some potential areas to focus on for S locus genes in *Linum grandiflorum*. Interestingly, all of *LgAP1*, *LgSKS1* and *LgMYB21* showed no difference in transcription levels between stylar morphs, despite different levels in protein accumulation. This implies a post-transcriptional method is involved in their regulation. Given that all are expressed (albeit to different degrees) in thrum and pin morphs and as such are likely not components of the S locus supergene, this could mean that this post-transcriptional regulation is controlled by some component of the S locus. In addition to these three genes and *TSS1*, there is one other gene with a more minor expression difference between thrum and pin morphs that may be worth considering. *TPP1*, or thrum pollen predominant 1, is a short gene of unknown function which shows a degree of increased expression in the pollen of the thrum morph. But while its function is not defined, it does bear homology to another protein of unknown function (AT5G38760, a protein associated with late stage embryogenesis in *Arabidopsis thaliana*) which can be induced by abscisic acid (Hou *et al.* 2008). This connection between genes involved in heterostyly and ABA mirrors results found in *Arabidopsis* (Suzuki *et al.*, 2014) and may be an area deserving of further study.

Linum is a genus where heterostyly is a defining characteristic, possibly to the same degree as that of *Primula*. However, it is also a genus with a relatively small amount of research done into the genetics of heterostyly. There is notably mixed evidence as to whether heterostyly in the genus is controlled via a heterozygous or hemizygous method. As such, it is a key area for further research. This project aimed to expand on the study of heterostylous genetics in *Linum* by identifying genetic markers closely linked to heterostyly in *Linum*, as a first step to identifying the genes responsible. This was done through sequencing of a relatively unstudied member of the genus, *Linum tenue*. *Linum tenue* is another heterostylous member of the *Linum* genus, with yellow flowers and typical expression of distyly. A recent taxonomy of the genus (Ruiz-Martin *et al.*, 2018) has suggested that *Linum tenue* consists of not one but three closely related species, with two native to the Mediterranean and the third to South Africa, but for the purposes of this project analysis shall be done on samples collected from Spain assuming this region represents a single species. The aim of this project was to sequence a balanced set of thrum and pin morph samples of *L.tenue* using ddRAD sequencing (a next generation sequencing method ideal for the rapid sequencing of previously unsequenced species for genome wide association studies, reliant on two different restriction enzyme digestions to give shorter fragments for assembly), and to identify any genetic differences between the two morphs. These genetic differences may be in the form of different haplotypes in the two morphs, implying heterozygous control of heterostyly, or the absence of a locus in one morph that is present in another, implying hemizygous control of heterostyly. The genetic fragments that differ between morphs could then be identified, by checking for homology to previously identified genes in other species. As such, the aim was to determine whether heterostyly in *Linum tenue* is controlled via a heterozygous or hemizygous method, and to identify some genetic markers corresponding to the genes involved. By discovering the genetic mechanisms behind heterostyly in *Linum tenue*, some light should be shed on the genetic control of heterostyly across the *Linum* genus, and a new model may be gained to explore the genetic mechanisms underpinning heterostyly in another heterostylous genus.

Section 2: Methodology

Chapter 2.1: ddRAD sequencing and sample preparation

The first step in the analysis was deciding which plants to include in the analysis. A number of specimens of *Linum tenue* had been collected before the start of this project on field trips in Spain. Seed samples had been collected on two separate occasions - in 2013 and 14 - from a wide variety of locations and environments (Table 2.1).

Sample Population	Location	Longitude	Latitude
ALT	Alhaurin de al Torre, Malaga	36.66868	-4.55433
ARA	Aracena, Huelva	37.89273	-6.57034
BUR	Burguillos, Sevilla	37.59495	-5.97533
CAZ	Cazalla de la Sierra, Sevilla	37.9374	-5.7612
CBT	Cabra, Cordoba	37.46709	-4.4229
EBO	El Bosque, Cádiz	36.77744	-5.518
ELB	El Burgo, Málaga	36.6275	-4.99011
LUM	La Umbria, Huelva	37.86229	-6.48091
MDA	Mairena del Aljarafe, Seville	37.34144	-6.04758
PIG	Pinos Genil, Granada	37.16089	-3.15231
SVT	Sevilla, Sevilla	37.35532	-5.99094

Table 2.1, the collection locations of all sample populations used in the study.

These seeds were grown in the laboratory greenhouse and the eventual plants analysed on a previous project, with stylar morphs noted. A map of the samples collected is included below with the eventual samples chosen highlighted. Locations and collection time of the samples was decided on the basis of creating the widest possible range of conditions, aiming to provide a representative sample of the whole range the plants were collected from, but the most important factor was the stylar morphs. Care was paid to earlier observations of the samples to ensure that an equal ratio of long styled to short styled samples was used in the analysis. Once the ideal samples were decided, previously extracted DNA samples of these plants were located and studied to see if they were usable.

There were a number of frozen DNA samples from *Linum tenue* available in the lab, having been extracted for genomic analysis by LGC for a previous project, so it was decided to attempt using these as the basis for analysis. The DNA for the remaining samples was extracted according to a modified version of the phenol-chloroform extraction protocol. In this protocol, initially several leaves from the plants of interests were snap frozen in liquid nitrogen and finely ground, with this step repeated as necessary. An extraction buffer was mixed from 500 parts CTAB buffer (itself composed of 30% CTAB in a 10% solution, 28% 5M NaCl, 4% 0.5M EDTA, 10% 1M Tris-Cl and 28% water) and 1 part RNAse A, and 500 µl of this solution was added to the leaf samples. These samples were then incubated in a 60 C water bath for an hour then chilled on ice, before 500 µl of phenol-chloroform was added and samples were mixed until visibly cloudy. These samples were centrifuged at 14’000 rpm to separate and retain only the aqueous layer, and the solution was purified again with 500 µl of chloroform:isoamyl alcohol. These samples were then centrifuged, mixed with 50 µl of sodium acetate and 600 µl isopropanol to precipitate the DNA and centrifuged again to separate and remove the supernatant, which was replaced with 500 µl of ethanol to wash the DNA. These samples were centrifuged a final time, the supernatant and as much residual ethanol as possible was removed and the samples were air dryed and resuspended in 50 µl of nuclease free water. Samples were stored in a -20 C freezer until needed.

That this DNA had been used in a previous successful project was a strong indicator that it was usable, however, it was still prudent to tests the quality and quantity of the chosen DNA samples to ensure that they were useful for the project. The samples were thawed then, to roughly assess the quality by identifying any ethanol contamination, the refractive index was measured using a Nanodrop spectrophotomoter. This provided an estimate of DNA concentration in ng/µl, and measured the absorption curve of the sample, quantifying the absorbance of the peaks at 230, 260 and 280 nm by providing ratios of absorbance at 260/280 and 260/230. Ethanol has an absorbance peak at 280nm, so a low 260/280 ratio was indicative of ethanol contamination (alongside less likely contaminants, such as protein, but given the extraction procedure ethanol contamination was the most likely cause) and resulted in the sample being excluded from the full analysis. Quantity of the good quality samples was then assessed using a QuantiFluor assay to provide a better estimate. This procedure is based upon a fluorescent dye annealing to double stranded DNA, with the fluorescence of this annealed dye acting as an accurate indicator of the overall concentration of DNA. As such, the procedure consists of diluting the QuantiFluor dsDNA dye 1:400 times in 1X TE buffer solution (consisting of a 20X solution of 0.2M Tris buffer and 20mM EDTA diluted 20 fold in nuclease free water), then adding 200 µl of this working solution to 1 µl of the DNA to be tested (keeping these samples in the dark until they were ready to be measured). The fluorometer was calibrated using a blank sample (200 µl of working solution with no DNA) and a known standard (200 µl of working solution with 1 µl of a DNA standard), then the fluorometer was set to measure fluorescence of 1 µl of DNA in 200 µl of working solution. The fluorescence of each sample was measured in turn using a Quibit 2.0 fluorometer (Life Technologies, and the concentration of DNA present estimated using this. Some samples were then excluded due to having DNA concentrations low enough that ddRAD was unlikely to succeed (using a somewhat arbitrary limit of 3.5 ng/µl), then from among those remaining 30 were chosen to provide the best distribution of samples for analysis of the genetic components of heterostyly. The samples were chosen to be from a wide range of environments, with 15 pin morphs and 15 thrum morphs so the two traits could be accurately compared. Care was taken to ensure there was minimal environmental bias among each stylar group (that for any given environment there was both a pin and a thrum sample included in the analysis). These samples were then prepared for ddRAD analysis.

To prepare the samples for double digest RAD sequencing, a slightly modified version of Peterson *et al's* (2012) (Peterson *et al.* 2012) protocol was used, the standard for ddRAD sequencing. The key differences in this protocol were in the volume of sample used for the restriction enzyme digest stage, to allow better scaling to large numbers of samples, the adapters annealed, to increase the amount of information to work with after sequencing, and in having one less cleaning step, to increase the quantity of DNA carried through to sequencing. The first step was to carry out the double digest, using restriction enzymes Mse1 and Pst1 - chosen to produce a significant number of well sized fragments based on the digestion sites present in the Linum usitatissimum genome (a closely related species) and because they could be heat inactivated. 500ng of DNA (based on the concentration of DNA obtained from the QuantiFluor assay) was kept on ice and mixed with 0.2 µl of 100U/µl Pst1 (NEB), 0.5 µl of 10U/µl Mse1 (NEB), 3 µl of Buffer 2.1 (NEB) (Consisting of a 1X solution of 50 mM NaCl, 10mM Tris-HCL, 10mM $MgCl_2$ & 100µg/ml BSA) and nuclease free water up to 30 µl (different from the 50 µl digest volume used in Peterson *et al's* protocol). When the enzymes were added, the samples were incubated for 3 hours at 37C, then the enzymes were inactivated by incubating at 80C for 20 minutes. The long incubation period and increased amount of Pst1 (relative to Mse1) were due to problems with incomplete digestion of the Linum tenue genome by Pst1 on previous projects. In the original protocol there was an additional cleaning step at this point, using an automated magnetic bead based purification method. However, due to concerns about the amount of genetic information available and the inherent risk of losing any DNA in the process of extracting contaminants, it was decided to remove this cleaning step.

After the digestion was completed and the enzymes were inactivated, adapters were annealed in preparation for ligation to the samples. In contrast to Peterson's protocol, this protocol used barcoded P1 and P2 adapters (Table 2.2) as opposed to barcoded P1 adapters and PCR primer indices. A selection of P1 and P2 barcoded adapters (barcode list of samples included in appendix) was chosen to identify the entire range of samples, with a stock of each adapter created by combining equal proportions of the top and bottom DNA strands of each adapter in an annealing buffer (Consisting of 100mM Tris-HCL, 500mM NaCl and 10mM EDTA in a 10X solution), then diluting to 40µM for a working stock (ensuring that the buffer was diluted to 1x concentration). These adapters were then annealed by heating to 95C for 2 minutes, and gradually cooled to 25C over 45 minutes. Freshly made adapter stocks were kept at 4C while in use, and were used alongside adapter stocks annealed in previous projects, which had been stored at -20C and thawed.

Adapter Name	Adapter Sequence
R1 "forward" adapter	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT- [6bp barcode]-[PstI digestion site]
R2 "reverse" adapter	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT-[6bp barcode]-[MseI digestion site]

Table 2.2, the adapter sequences used in this project. Note that these are not the standard Illumina adapters used for sequencing, and that the reverse adapter also included an identifiable barcode.

These adapters were then ligated to the DNA samples. The amount of adapter required to be added to each samples was calculated using a ligation molarity calculator provided by Peterson *et al* (2012), using approximate values for cut frequency based on average distance between digestion sites in related organisms with known genomes, and the adapter stock was diluted as appropriate so that

1.5µl of P1 and 1.5µl of P2 adapter contained the appropriate amount of adapter. These 1.5µl portions of adapter were added to the 30µl DNA samples, and then 3.4µl of 10X T4 ligase buffer (Consisting of 50mM Tris-HCL, 10mM $MgCl_2$, 1mM ATP and 10mM DTT in a 1X solution)and 0.6µl of 400U/µl T4 ligase were added (again, slightly less than the Peterson protocol, proportional to the reduced digest volume used). These samples were then incubated at room temperature for 30 minutes, then heat activated for 10 minutes at 65C. The solution was then cooled at a rate of 2C per 90 seconds until it reached room temperature. To ensure that the ligation had worked, 30 cycles of PCR (Consisting of 20 ng DNA, 4 µl Phusion Buffer 5x (Consisting of 150 mM Tris-HCL at pH 10, 50mM KCl, 50mM $(NH_4)_2SO_4$, 10mM Mg SO_4, 0.5% Triton X-100 and 0.5mg/ml BSA in a 5X solution) , 0.4 µl 10mM dNTPs, 1 µl 10µM P1 forward primer, 1 µl 10 µM P2 forward primer, 0.2 µl Phusion DNA polymerase and nuclease free water up to 20 µl) (98C for 130 seconds, 65C for 30 seconds, 72C for 10.5 minutes) were conducted on one randomly selected sample post ligation, and gel electrophoresis was done to identify the size of the sample. Given that the approximate size of the post digestion DNA fragments and the size of the adapters was known, this showed whether the adapters had ligated to the DNA fragments. Assuming that this was the case, this should mean that all samples were now individually barcoded and identifiable.

The next step was to pool all samples to create one library of DNA samples that could be worked on directly in subsequent steps. As all the samples were individually barcoded with a unique combination of P1 and P2 adapters, the individual samples could be separated and identified in post sequencing data analysis. To create this library, equal amounts of ligated DNA from each sample was combined into a single sample. This did not mean adding equal amounts of solution from each sample - from the pre-digestion DNA quantification it was clear that there was a range of DNA concentrations present among the samples, and it was unlikely that this had significantly changed in the digestion and ligation steps. As such, rather than diluting the samples, it was decided to adjust the amount of each added to the library in proportion to the concentrations measured in the pre digestion assay. This ensured that equal amounts of DNA were added from each sample, and that the resulting pooled library was concentrated enough to sequence successfully.

After the samples were pooled, the next step was to clean the library sample, to remove any contaminants introduced in previous steps. To do this the SeraMag magnetic bead purification system was used, separating the sample into equal portions of no more than 400µl across as many 1.5ml microcentrifuge tubes as are required. 1.5x the volume of each tube in the SeraMag bead solution (Consisting of 50% 5M NaCL, 1% 1M Tris base, 1% 1M EDTA, 0.34% 1M HCl, 40% PEG 8000 in a 50% weight/volume concentration, 0.5% Tween 20 in a 10% weight/volume concentration, 2% suspension of Sera-Mag magnetic beads and 7.16% nuclease free water) is then added and mixed.

The beads with bound DNA are separated through use of a magnet, and the rest of the solution is discarded. Then the beads are resuspended in a TE elution buffer solution (Consisting of 1% 1M Tris Base, 1% 0.1M EDTA, 0.37% 1M HCl, 0.5% Tween 20 in a 10% weight/volume concentration and 97.13% nuclease free water) to release the DNA again.. A cleaning step was necessary, given the likelihood of contamination across the entire protocol and the damage that could cause, but any cleaning step has a significant risk of DNA loss. With the samples available, there was a serious risk of not having enough DNA available to fully sequence, so the decision was made to cut back to a single cleaning step. This should still be sufficient to remove problematic contaminants, allowing the samples to proceed to the size selection step.

The next step in the procedure was size selection, filtering the DNA to only include fragments of a specific size. From distances between restriction sites in genomes of similar organisms, it was estimated that the fragments of interest would be approximately 120 bp in length. However, this isn’t including the 76bp of adapters ligated in the previous steps, producing an actual target fragment length of close to 200bp. To target these fragments, a size selection range of 300-550bp was chosen. As such, all fragments not in this 300-550bp range needed to be filtered and kept separate from the rest. The Pippin Prep electrophoresis based size selection instrument was chosen to do this. This instrument functions on a similar principle to standard DNA gel electrophoresis: namely, DNA fragments were deposited at one end of an agarose gel, and a positive electrical charge was applied at the other. The positive electric charge attracted the negatively charged DNA fragments, and they were pulled through the gel towards the charge. Larger DNA fragments faced more resistance from the gel, so they travelled more slowly, and so the DNA fragments gradually separated out based on size To use this for size selection, the Pippin Prep has two positive electrodes at the end of the gel opposite to the DNA fragments, one on the left and one on the right. The gel splits into two paths to allow this, and along the right path there is a sample well, intended to capture DNA fragments. Initially the positive charge was run through the electrode on the left, pulling the DNA fragments down the gel and towards this electrode. At a certain time, determined automatically based on when a pre specified size of DNA fragment is estimated to be approaching the fork between the two paths, the charge switched to the right electrode, still pulling fragments down the gel but towards this right electrode instead of the left. The fragments which are attracted past the fork and down this path, which should be the fragments of the desired size, were captured in the sample well along this path. Later, at another predetermined time, when fragments from the upper end of the desired size range should have been fully attracted into this well, the charge switched back to the left electrode. Thus, all other fragments were attracted into the left path of the gel. This should mean that when the electrophoresis was complete, all fragments of the desired size should be trapped within a sample well on the right path of the gel, and can then be removed for further analysis. Sufficient care was taken when setting up the gel and the samples for electrophoresis. Firstly, the gel waschosen to best work with the samples being tested. The concentration and composition of the gel can affect the flow rate of the DNA, and a choice must be made between using a gel with an external marker (a DNA fragment of known size, similar to the “ladder”, used to indicate to the Pippin when to switch electrodes, run in a separate lane to the samples) or a gel with an internal marker (similar to an external marker but mixed with the samples in each lane). Previous research projects had shown most success with ethidium bromide gel cassettes with external markers - it had been speculated that high adapter concentration in the sample interfered with the internal marker, preventing it from being detected by the Pippin. So an ethidium bromide cassette using an external marker was chosen, with 1.5% agarose - this was recommended as best for eluting 200bp DNA fragments. Secondly, after the type of gel is selected, a gel cassette of the chosen typewas taken and examined, ensuring there are no bubbles in the gel that could displace the DNA being loaded. Thirdly, the Pippin Prep was calibrated to work with the type of gel being used, and the range of fragment sizes desired entered. The gel cassette was then loaded on to the Pippin, and a continuity test performed - to ensure that the current flows properly in each electrode. The DNA samples and marker were then loaded into each lane, with 30µl of sample loaded with 10µl of DNA marker, and the size selection electrophoresis program was run. This step did have the inherent issue of severely reducing the amount of DNA available in the sample - the DNA removed may have been useless for further analysis, but it still meant that previous estimates of DNA concentration are too high. Given that these previous estimates were already approaching the borderline for workable concentrations of DNA for analysis, this presented potential issues. So to somewhat counteract this, after this size selection step the DNA sample was put through a PCR step with P1 and P2 primers to increase the amount available.

This PCR step was intended to amplify the DNA taken from size selection to a significant enough quantity that it could be used for analysis. To quantify its effect, after size selection the extracted DNA was requantified before PCR, using the same Quantiflour assay as used before. A dsDNA working solution was prepared by diluting Quantiflour dye to 0.5% concentration in 1xTE buffer (consisting of a 20X solution of 0.2M Tris buffer and 20mM EDTA diluted 20 fold in nuclease free water), then blank, standard and sample solutions were prepared. The blank sample consisted of 100µl of 1xTE buffer mixed with 100µl of dsDNA +

working solution, the standard sample consisted of 100µl of Lambda DNA (diluted to 1% concentration in 1xTE buffer) mixed with 100µl of dsDNA working solution and the sample to be tested consisted of 1µl of the DNA obtained from size selection, mixed with 100µl of dsDNA working solution and 99µl of 1xTE buffer. The Qubit fluoremeter was calibrated using the blank and standard samples, then the concentration of the sample to be tested was measured. A sufficiently high concentration at this point, significantly greater than 15ng/µl, may indicate that PCR is unnecessary, even if it might suggest issues with size selection. However, the concentrations of DNA obtained at this point were all 15ng/µl or lower, so the sample wasthen amplified to acceptable levels using PCR.

The next step was the PCR amplification step. To begin the PCR reaction, approximately 20ng of template DNA was required. This had to be contained within, at the most, 14µl of solution, so if the concentration is lower than 0.7ng/µl this step is impossible, and the protocol should be redone (possibly with different samples, if possible). Under ideal circumstances the volume of DNA solution used as a PCR template (and so containing 20 ng of DNA) should be in the region of 5µl, meaning that concentrations lower than 4ng/µl (as seen in the fluorometry readings) could be problematic. To remedy this, the solution was concentrated using a vacuum centrifuge, which allowed the PCR step to proceed.

One key risk of PCR that was considered is the possibility that the DNA can be amplified unevenly - certain fragments can, by random chance, be excessively amplified while others have a chance of not being amplified by early PCR cycles and ending up not represented in the final sample. In this protocol there were two measures to mitigate this. Firstly, the sample was split between multiple PCR reactions, in this case 4, with the final results of each being pooled. Peterson et al (2012) recommends setting up at least 4 and at most 8 different reactions - though there are few problems associated with setting up too many reactions beyond practicality. Secondly, the number of cycles of PCR was kept as low as possible. Peterson et al recommend using no more than 12 cycles as an absolute maximum, and in this protocol 11 are used. Fewer cycles of PCR means that the DNA is not amplified as much as it could be, but also means that there is less time for base misincorporations to accumulate and biases in amplification to become too pronounced. Each individual PCR reaction consisted of ~20 ng of template DNA (given the DNA concentration measured after size selection, this consisted of 2µl of sample DNA), 4µl of 5x Phusion Buffer (Consisting of 150 mM Tris-HCL at pH 10, 50mM KCl, 50mM $(NH_4)_2SO_4$, 10mM Mg SO_4, 0.5% Triton X-100 and 0.5mg/ml BSA in a 5X solution) , 0.4µl of 10mM dNTPs, 1µl of 10µM P1 forward primer, 1µl of 10µM P2 reverse primer, 0.2µl of Phusion DNA polymerase and nuclease free water to make each reaction up to 20µl (in this case 11.4µl of water). Once these reactions were prepared, the cycles were programmed onto the thermocycler. In this case, each cycle consisted of ten seconds at 98°C, 30 seconds at 65°C and 30 seconds at 72°C. There were 11 of these cycles included in the program, and they were preceded by two minutes at 98°C and followed by ten minutes at 72°C. The resulting sample was kept at 4°C until it could be removed.

It can be difficult to determine just from the sample obtained from this whether the PCR has worked - whether it has amplified the sample DNA to a usable level and whether it did so evenly, not amplifying certain fragments of DNA in excess of the others. To verify this, a positive control was used. A fifth PCR reaction was prepared using the DNA from the size selection step. This sample DNA was put through 20 cycles on the thermocycler (using the same temperature settings as before). The sample resulting from this was unusable for sequencing - it being too likely that the excessive number of PCR cycles has resulted in a large number of base misincorporations, making the DNA unrepresentative of the DNA from the original sample. However, the excessive PCR amplified the sample enough that it was visible on a standard gel electrophoresis. As such, to test if the PCR has worked a DNA gel was prepared and the positive control sample from the PCR was run alongside a DNA “ladder” (reference DNA) for two hours. If the PCR had amplified the DNA, this should produce a visible mark on the gel in the lanes used for the sample DNA, and if the PCR had amplified the DNA evenly, this mark should be approximately 120 bp in length (determined by comparing the position of the mark against the position of the appropriate “rung” of the DNA reference ladder). As the sample tested was seen to be approximately 120 bp , this indicated that the size selection was successful and that the level of amplification bias introduced in the PCR step was manageable. As such, the sample from the PCR was then taken from cold storage, the separate reactions pooled into one sample and cleaned.

To clean this sample, another bead cleaning step was used. For this step, Seramag beads were used. These beads were added to the pooled sample, at a ratio of 1.5 bead solution (Consisting of 50% 5M NaCL, 1% 1M Tris base, 1% 1M EDTA, 0.34% 1M HCl, 40% PEG 8000 in a 50% weight/volume concentration, 0.5% Tween 20 in a 10% weight/volume concentration, 2% suspension of Sera-Mag magnetic beads and 7.16% nuclease free water) to 1 sample solution. There was 80µl of sample solution, so 120µl of bead solution was added. The bead solution and sample solution were thoroughly mixed, then left to incubate for 5 minutes at room temperature. The solution was then moved to a magnet stand, and left until the beads have completely clumped and eluted from solution. There should be a clear separation between pellet and supernatant, allowing the supernatant to be discarded. The beads left behind were then washed twice with 80% ethanol (covered with 80% ethanol, left for 30 seconds and then the ethanol is discarded, twice) and left to air dry to remove ethanol contamination. The sample was removed from the magnet and elution buffer (Consisting of 1% 1M Tris Base, 1% 0.1M EDTA, 0.37% 1M HCl, 0.5% Tween 20 in a 10% weight/volume concentration and 97.13% nuclease free water) was added to the beads and mixed. The amount of elution buffer added at this stage determines the volume of sample obtained from the cleaning step. 30µl of sample is the minimum required for sequencing, and a certain amount extra is required for quality control steps, so the aim was to elute around 40µl from the cleaning step. As such, around 40µl of elution buffer was added. The elution buffer and beads were thoroughly mixed, and the sample was transferred again to the magnet stand and left until the beads have completely separated from the solution. In this case, the DNA should be in the supernatant, so the supernatant was carefully extracted and added to a separate tube. Some supernatant and consequently DNA was likely lost at this stage, though care was taken not to break the pellet of magnetic beads and mix some of them back into solution. After this step, the cleaned sample had one more step before being ready for sequencing - quality control.

There were three key factors determining the quality of the DNA sample - one, the presence or absence of contaminants. Two, the concentration of DNA, that is, the raw amount of DNA present in the sample. And three, the amount of DNA present in the sample of the correct size range, or how much DNA of the original sample to be tested is present in the current sample. The previous bead cleaning step theoretically removed the contaminants present in the sample, so the other two factors (DNA concentration and DNA size) could be determined through use of the Tape Station. The Tape Station is, similar to the Pippin Prep size selection instrument, a form of automated gel electrophoresis. The difference here is that rather than attempting to separate out part of the sample through electrophoresis, the Tape Station acts more similar to a (more accurate and complex) version of the standard bench gel electrophoresis used to approximate DNA fragment size The difference was that rather than the size being visually assessed at one point when the electrophoresis has finished, in the Tape Station the size was continually assessed while the electrophoresis is running by a high quality camera. This allowed for more precise quantification of the size of the DNA, and by assessing the level of fluorescence the amount of DNA of each size could be determined as well (similar to the Qubit assay for measuring DNA concentration). To run the Tape Station analysis, a Tape Station “tape” was required, alongside strip tubes for use with this “tape”. 1µl of sample was loaded into one strip tube using a pipette and 1µl of a DNA reference ladder was loaded into another. These tubes were labelled and 3µl of buffer (with maximum strength of 20 mM KCl, 60 mM Phosphate Buffer, 60 mM Guanidine-HCl, 240 mM NaCl, 60 mM Acetate)was added to each. These tubes were spun down, then the tape and the tubes were loaded into the machine, the lids of each tube removed and the location of the ladder and sample (that is, the position of the tubes which contain them) was entered into the machine. At this point, the Tape Station run could begin. This lasted for approximately 20 minutes, after which a comprehensive output of the sample was obtained. This output had information on overall sample concentration, the sizes of DNA fragments present in the sample, the concentration of DNA of each size and an estimate of the overall integrity of the DNA (estimated based on the range of DNA fragment sizes obtained, based on the assumption that more degraded DNA appears as a wider size range of DNA fragments in electrophoresis). The Tape Station output of the initial run of this project indicated an almost complete absence of DNA from the desired size range, which lead to a repeat of the protocol from the end of the ligation step, including a positive control at the end of the ligation step involving amplifying the DNA with 30 cycles of PCR and running it on a gel (to ensure that at this point there was still a sufficient amount of DNA to work with). In this second run of the protocol, the size range selected for by Pippin Prep was slightly expanded (to 300 - 550 bp), and the number of PCR cycles included increased from 11 to 14. The result of this was seen in the second Tape Station output - the DNA was more diffuse (indicative of the problems of large numbers of PCR cycles) and spread across a wider size range, but there was a sufficient amount of DNA to allow sequencing to proceed. Consequently, the samples were then submitted for sequencing.

The sequencing method used in this project was sequencing by synthesis on an Illumina 2500 HiSeq machine in the Genomics Facility, where DNA bases are identified as they are incorporated into a nucleic acid chain. Initially, DNA fragments are captured in a flow cell, as the adapters ligated are attached to the flow cell. These are bound as single DNA strands, and then complementary strands are synthesised through addition of a DNA polymerase and free nucleotides. This newly synthesised double stranded DNA is then denatured, separating it into single strands, and then new complementary strands are synthesised for each of these single strands. This process was repeated multiple times, quickly generating dense “clusters” of identical DNA fragments in close proximity. This cluster formation meant that each DNA base was identified from around a thousand identical fragments, reducing the chance of a base being miscalled and amplifying the fluorescent signal for each base, making it easier to be identified. After this cluster formation, the nucleotide bases at both ends of these fragments were cleaved and reincorporated base by base. Each nucleotide base was labelled with a fluorescent dye, with a unique fluorescence for each base. This label identified each base, but also prevented further DNA polymerization, meaning that it needed to be cleaved from the base before further bases are added (which ensured no base was skipped). This fluorescent dye was then imaged as it is cleaved, which identified the base incorporated and provided a base by base sequence of the DNA fragment.

After this was complete, the data was compiled into individual lanes, converted into FASTQ format where the fluorescent wavelength at each step was converted into the most likely base at that fragment position, along with a quality score reflecting the certainty of that base call, and collected on the University Hamilton supercomputer for further analysis. This concluded the first major section of the protocol, the ddRAD sequencing and the work needed to prepare samples for that, and began the second - post sequencing data analysis.

Chapter 2.2: Post Processing

After sequencing was complete, the data was compiled on a database in the form of FastQ files, each containing either the forward or reverse of all the sequences from a particular lane. The first step was to visually assess the files generated, reading each FastQ file using the nano function in Bash, noting any obvious problems (e.g. missing lines, unexpected symbols in the DNA sequences) and determining the cause of and solution to them. Each file consisted of either the forward or reverse read of every sequence generated in the lane. Each sequence had four lines associated with it – one empty and three lines of information. One line describes a unique identifier for the sequence. One line displays the sequence itself, the sequence of bases called for this particular DNA fragment. And one describes the quality of the base calls, approximated by the sequencer. This line consists of a series of symbols corresponding to each base, with each symbol referring to a certain range of base call quality. It was expected that these FastQ files would be labelled with the appropriate researcher, but in actuality a minimum of sequences were included in these named files, with the majority included in an “Undetermined” file. This initially raised concerns about the accuracy of the sequencing, but later quality checks confirmed that the “Undetermined” file contained relatively well sequenced fragements. The issue was that the fragments were sorted into files based on the presence of Illumina adapter sequences, while this protocol used non-standard adapter sequences. The presence of fragments in named files was indicative of low quality sequencing, as these fragments contained a sequence that was not present in the sample population. Nevertheless, it was decided to use both unnamed and named files for further analysis, to avoid discarding potentially relevant data and with the expectation that later quality checks would alleviate any issues resulting from including low quality sequences.

The second key stage in post processing was to accurately assess the quality of the sequences called, which allowed a plan to be developed for how to proceed with processing the data. To do this, a program called FastQC was used. FastQC analysed the Fastq files with the data for each sequence, and produced a summary of several possible metrics of quality for all the sequences in each file (alongside basic information such as the identifier for the sequences, the number of sequences included and the average length of sequences included). FastQC summarized information on the average quality of per base position, per base called and per sequence. It measured the number of GC bases called on average and models that against the normal distribution expected, the number of “N”’s called at each base position in a sequence, the distribution of sequence lengths in the sample data and the number of sequences duplicated in the sample (raising particular notice if one sequence is overrepresented in the data set). To ensure that nothing is missed, FastQC also measured Kmer (short sequences of DNA of a certain length, k – what FastQC refers to as Kmers are sequences of DNA 7 base pairs long) duplication levels at each base position. This meant that even if the entire sequence is not duplicated, duplicated subsequences embedded in the middle of the sequence are identified and noted. Finally, FastQC can also identify given adapter sequences in the sequences. Measuring all these different metrics ensured that a comprehensive measure of sequencing quality is obtained, and the granular nature of the results meant that problems in certain aspects of the sequence can be identified and solved individually. So FastQC was run on each file of sequences, and immediately several problems were identified. FastQC flagged large quantities of low quality reads, significant numbers of uncalled bases and notably high kmer duplication levels – all possible indicators of low quality sequencing or contaminant inclusion. As such, the next stage of the post sequencing analysis needed to be filtering the low quality sequences, with the aim of improving the quality of all sequences included.

The third key stage of post sequencing processing was improving the quality of the sample data by removing contaminant sequences and miscalled bases. To do this, a balance needed to be found in cleaning the sample data. It was important to be stringent, to remove as many low quality fragments as possible and so prevent sequencing errors from resulting in misleading results. However, there was a significant risk of discarding useful data by being overly zealous at improving quality. A set of cleaning steps was chosen to best balance these concerns, and to try and improve the quality of the sequences used as much as reasonably possible without discarding too much usable data. The two key factors to consider in balancing this were the program used, and the parameters applied.

The program used for this was a matter of some consideration. The obvious choice, based off its success in previous research projects, was Trimmomatic. Trimmomatic is an open source specifically for trimming low quality sections from Illumina based next generation sequencing data, such as the data generated in this project. As a program dedicated entirely to this function, cleaning and filtering fragments, it could be reasonably assumed that it would perform this function to a high quality standard. But the fact that Trimmomatic is a standalone specialised program also has disadvantages. After the cleaning step, the rest of the analysis will be done by a completely separate set of programs, which may cause issues when incorporating the Trimmomatic output into the overall analysis pipeline To avoid this issue, a program would need to be used for this step with a similar range of function to Trimmomatic that was also designed to work with future programs in the analysis pipeline. As such, for this protocol the program used to clean and filter the reads based on quality was the “process_radtags” program from the STACKS software package. STACKS is a complete set of programs for assembling loci from short fragments of sequenced DNA (Catchen *et al.* 2011, Catchen *et al.* 2013), such as those generated by the ddRAD sequencing process, into an approximate reference “genome”. The majority of these programs are incorporated later in the protocol, and form the backbone of the analysis pipeline. The “process_radtags” program, intended primarily for assigning DNA sequences to samples, also contains a robust set of commands to “clean” DNA fragments, to remove contaminant sequences and filter low quality sequences, comparable to Trimmomatic. Trimmomatic had certain advantages, with more specialised tools and cleaning algorithm adapted more to paired end reads, and should be considered a valid alternative for reconstructing this experiment. However for this protocol, to avoid the issue of incorporating separate programs into a single pipeline “process_radtags” was used.

The parameters used by process_radtags were decided on an individual basis for each of the individual cleaning steps, primarily based upon the results of the post processing cleaning steps of previous experiments with similar data sets. All the techniques used in the cleaning step were detailed in a single command line and carried out sequentially by the program. The first cleaning process used was identifying and removing any Illumina adapters present in the fragments. In this step, the set of adapter sequences used (in the case of this project, the non-standard RADseq adapters used and not the standard Illumina adapters) was provided to “process_radtags”, which then scanned through fragments searching for matching sequences, which it then removed from the fragment. It tested short sequences of the adapter against every possible position in both the forward and reverse reads. If there was less than a certain number of mismatches between the adapter subsequence and the fragment, the entire adapter sequence was aligned against the read and scored (based on the number of matching and mismatching bases, and the quality of the mismatched bases, so that probable miscalled bases do not unduly effect the scoring). If the alignment score was above a certain level, the sequence was marked as an adapter sequence and removed from the fragment. In this protocol a score of 10 was used for single sequences, with a score of 7 corresponding roughly to a perfect alignment of 12 bases and 15 corresponding roughly to an alignment of 25, while the score for palindromic (paired end, using two paired reads instead of one) alignment was significantly higher, at 30, to take into account the extra information used. After this step, all bases below a certain quality threshold were removed from the start and end of the read. The quality threshold for this was set deliberately low, at 3, to filter out only obviously unusable bases. The subsequent step was an analysis of the sequencing quality of the whole fragment, which functioned as a “sliding window” scan. To do this, the fragment was scanned in short sequences of bases, or “windows”, across all positions in the fragment. If the average sequencing quality in this sequence fell below a set threshold value, the bases in the window were cut out and removed. By checking the quality of and potentially removing multiple bases simultaneously, the program could handle larger numbers of fragments efficiently and quickly, which was crucial in allowing this program to be incorporated into a longer sequencing analysis pipeline For this protocol, the window size was set to 4 bases wide and the average quality was set to 20, again based on the results of previous projects. This can be seen as a relatively low threshold score, but it was high enough to filter out all truly unusable data. More importantly, setting this value as a minimum standard rather than an ideal meant that enough data could pass for reasonable statistical analysis to be performed later in the protocol. A larger sample size of fragments to analyse balanced out the potential issues involved with including more low quality reads. Finally, the fragments are cropped to a uniform length of 120 by removing bases from the end of the read – as the programs used in subsequent stages require fragments of uniform maximum. After all these steps are complete, the amount of fragments remaining is assessed. Presuming there is still a significant amount of data to work with, the next step begins – identifying which fragment belongs to which sample.

The fourth key stage of post sequencing processing was thensample identification, or demultiplexing, by processing barcoded primer data. The first program involved in Stacks is primarily built to do this, and it is the "process_radtags" program, so this program was used for both this stage and the cleaning stage. To identify the sample fragments, the list of which barcode combination corresponds to which sample (included in appendix) was provided to process_radtags, alongside which enzymes were used to digest the sample (as the program also looks for restriction digestion sites associated with these enzymes as an additional quality control step), and parameters were set for how carefully these barcodes will be identified. Firstly, it was specified that the data was paired data, and that the barcodes used were incorporated into both ends of the data. Options to enable cleaning and quality filtering as per the last step were included, as well as an option to "rescue" barcodes – to treat barcodes which don't perfectly match any of the barcodes provided as the barcode they are most similar to. The number of mismatches allowed between the barcode provided and the barcode present in the data to allow it to be rescued was also specified, and as with other choices in the protocol it was a matter of sensitivity versus specificity or, roughly speaking, the amount of data included versus the quality of the data included. Fountain *et al* (2016) found that the most important factor for accurate sequencing is not the quality of the DNA prepared, or the sequencing, but the coverage of DNA sequences being analysed. Coverage of 20-30x (corresponding to 20-30 different DNA fragments covering any one point) was sufficient to produce a reasonably accurate genome from de novo sequencing, regardless of the quality of the DNA used. Similarly, coverage significantly below 5x was more associated with sequencing errors in the final genome than any quality of DNA. So, to guarantee high coverage and improve later analysis, a large amount of reads needed to be retained through the cleaning and mapping process. The data was initially processed with both 1 base mismatch and 2 base mismatches allowed, to compare the amount of fragments retained. The

1 base mismatch processing resulted in too many sequences being discarded for later analysis to be viable, so it was decided to go ahead with 2 base mismatches allowed. Furthermore, as the estimated coverage was close to 5x for several samples, it was decided to disable RAD checking as well, to further increase the number of sequences retained. Normally, as well as checking the integrity of the barcodes, the program checks if the RAD digest site is intact, checking that there are sequences within the fragment consistent with digestion by the restriction enzymes specified, and discarding any reads where this is not the case. With the quality check built into this step already, and the amount of reads retained for later analysis being worryingly low (if usable) even at 2 mismatches, it was decided to stop the program checking whether the RAD site was intact to save a few more reads. At this point the process_radtags was run for a final time with the parameters specified, and the reads were sorted into files corresponding to the samples sequenced. With this, the data was ready for the fifth step in post sequencing analysis – mapping the reads to a theoretical genomic map.

To map the reads into a genomic map, one of two main programs in Stacks can be used depending on the species being sequenced. If the species has already been sequenced and a reference genome is available, ref_map.pl can be used. In this case, the reads are combined into loci and compared against the reference genome to create the genome being sequenced. If the species has not already been sequenced, though, and there is not a reference genome available, denovo_map.pl is used instead. In this case, the reads are still combined into loci, but after they have been combined multiple theoretical genomes are created based on the loci constructed. These potential genomic maps are tested on a few different criteria, and the most robust genome which best accommodates the data is produced as a newly constructed genome. Describing the Stacks denovo mapping program as a single program can be somewhat misleading, though, as the Stacks denovo_map.pl command consists a wrapper program, managing the execution of multiple different programs each performing a single stage of the mapping. In this project, other flax species have been sequenced, the best genome available being of Linum usitatissimum. It may have been possible therefore to use this genome as a reference for Linum tenue and use the reference based mapping program, however, this was likely to cause problems. Linum tenue and linum usitatissimum are not closely related species, diverging between 23 and 44 million years ago (Sveinsson *et al.* 2014, You *et al.* 2018) and having several clear morphological differences. In particular, Linum usitatissimum is not heterostylous, meaning using it as a reference genome for a project examining the genetic basis of heterostyly was an impossibility. As such, for this project the denovo mapping program was the sole option. This program had been successfully used for sequencing many other species, so this was not a problem.

The initial plan had been to do the entire analysis in denovo_map.pl in Stacks 1.0, but some problems were found in integrating a sample of mapped data into later analysis programs. As suchhe decision was made to change the mapping program used to avoid this issue, and the ideal program was decided to be denovo_map.pl in Stacks 2.0, which was released in the middle of this project (Rochette, Rivera-Colón and Catchen, 2019). The exact programs involved in the denovo_map.pl pipeline differed from the 1.0 to 2.0 versions, which solved the issues faced with using Stacks 1.0 and was the main motivation to switch. Both Stacks 1.0 and 2.0 have a core denovo mapping pipeline of ustacks into cstacks, followed by sstacks and finally populations. Stacks 2.0 differs from this by adding two more programs between sstacks and populations, namely tsv2bam and gstacks. These programs were the core difference between Stacks 2.0 and 1.0, and their addition allowed the data to be output in bam format (binary alignment map, the compressed format of sam, or sequence alignment map) which was required by later analysis programs (namely PLINKt. At this point the code was rewritten to function in Stacks 2.0, and the analysis was redone from the FastQC stage using Stacks 2.0 to avoid issues. Aside from the time taken to do this, this switch of programs caused no major problems.

The first program used in the final mapping protocol was ustacks, which constructed genetic loci from denovo individual reads. It did this by first combining identical (within a certain tolerance) short read sequences from the same sample into stacks, and then combined similar stacks (again, from the same sample) into loci – with differing stacks being differing alleles at this loci. There were several parameters that could have been set to control the execution of this program, but the relevant parameters that needed to be considered for this program were maximum number of mismatches allowed between reads to be combined into a stack and minimum number of reads required to form a stack (any reads which do not fit in to stacks are discarded). The second program was cstacks, which combined these loci into consensus loci, which could then be constructed into a catalog. Loci constructed by ustacks, each assigned to an individual sample, were compared to one another, and any which were sufficiently similar were combined into consensus loci, where the original loci from ustacks were alleles of these consensus loci. This catalog of consensus loci represented the “average” genome across the samples sequenced. The relevant parameter for this program was number of mismatches allowed between loci to be merged into a consensus locus. The third program in the Stacks pathway was sstacks, which matched the original reads against this catalog to note how samples differ from the “average” genome. Sstacks took the loci from ustacks, which were previously identified with individual samples, and compared them with the catalog of consensus loci from across all samples to identify genetic differences of individual samples, alternative alleles and SNPs. This program had no parameters relevant to this analysis, save basic input and output locations. The next two programs were tsv2bam and gstacks. Tsv2bam combined consensus loci into a genetic map, allowing the data to be transposed by constructed locus instead of original sample – the individual sample loci from ustacks were aligned against the consensus loci from cstacks and a sequence alignment map was produced (a bam file). Gstacks had two key functions, somewhat uniquely among the Stacks pipeline. All the programs up to this point had been assembling data as if it consisted of single end reads, with the exception of tsv2bam, which identified whether the reads are single or paired end, and if they are paired end detects which forward reads should be aligned with which reverse reads. So the first step of gstacks was to assemble the paired end reads into contigs and align these contigs against the single end loci constructed so far, which created a catalog of paired end consensus loci similar to the output of sstacks. Secondly, gstacks identified SNPs from sstacks at each locus, identifying which individual sample corresponded to which SNPs and using this to make a set of haplotypes. Finally, the last program in the mapping pipeline wass the populations program. Populations used the polymorphism data from sstacks and gstacks to provide population statistics of the samples involved. By default this just included heterozygosity (expected and observed), calculated using the F statistics F_{IS} (the expected level of heterozygosity of an individual in a subgroup of the total sample set) and F_{ST} (the expected level of heterozygosity of a subgroup of the total sample set, calculated by comparing all populations pairwise against one another). The program could have done much more, such as calculating haplotype diversity divergence from the Hardy Weinberg equilibrium per locus. But for this analysis, focusing mainly on the genetic basis of heterostyly in the samples sequences, it was decided to mostly disregard the population statistics generated by populations. The goal of this project was to identify loci with expression differences between the two stylar morphs and this could be done succinctly with later analysis programs purely based on the output of gstacks. The population statistics had little bearing on the output of this later analysis. The wrapper program denovo_map.pl ran all the previously described programs in sequence, which removed several concerns about correctly piping the results of one program into the next, but several parameters still needed to be correctly set. And in a slight deviation from previous stages, deciding on the correct values for these parameters required a substantial amount of testing.

As previously specified, there were three key parameters that need to be considered for Stacks mapping to be successful. These were all localised to the first two programs in the pipeline, which constructed the catalog of loci which form theed basis of all future mapping steps. These parameters were, in ustacks, maximum number of mismatches allowed between reads to be combined into a locus (referred to in denovo_map.pl as M) and minimum number of reads required for a stack to be formed (m), and in cstacks, maximum number of mismatches allowed between loci to be combined into a consensus locus (n). Setting the correct value for each of these parameters required balancing several factors. Low values of m produced more loci, as each one could be composed of fewer reads, but the average coverage for each loci tended to be lower, as more loci consisted of fewer reads. . The factors to be considered for M and n were different from m but similar to each other, as the two parameters controlled the same aspect in different programs. Lower values of M produced more loci, as similar reads were assigned to different loci due to a few mismatched bases, but there wass a risk of under merging reads, where different alleles of the same locus were assigned as monomorphic individual loci which reducing the number of polymorphic loci available. Conversely, higher values of M produced fewer loci but more polymorphic loci, as more similar reads were combined into a single locus as different alleles. This did contain the risk of over merging loci, however, where similar monomorphic loci were combined into a single polymorphic locus. Setting n to high or low values has the same effect as M, except instead of reads being over or under merged into loci, loci could be over or under merged into consensus loci. Setting these parameters correctly was more difficult than setting parameters for previous analysis steps, as the correct values are almost entirely dataset specific, dependent on the amount of polymorphism present in the samples sequenced and the number of reads available per locus. As such, the only way to accurately set them was to test various combinations of parameters with the samples available.

There were several outcomes which could be measured in this test, but one influential study (Paris, Stevens and Catchen, 2017) argued that the most relevant parameter to consider is the number of polymorphic loci present in 80% of the population, or r80. Checking the number of polymorphic loci present in 80, rather than 100, percent of the population accounted for variance in the sample DNA being tested, where loci are not present in all samples being tested, while ensuring that few erroneous outlier loci were included in the analysis. In addition, to determine the ideal value of m two factors were considered – total number of loci formed and coverage obtained per locus. Fountain *et al.* (2016) suggested that a coverage of at least 10x should be obtained for accurate sequencing, so m was adjusted with the goal of sequencing every sample to 10x coverage while also maximising the number of loci available for analysis. Testing the entire dataset with all possible parameter combinations would take an impractical amount of time and processing power, so for testing a representative subset of samples is chosen, with all parameter combinations that fit within certain constraints. This subset of samples was chosen to be a fifth of the overall data set, or 6 of the available 30 samples, with 3 long styled samples and 3 short styled samples. Every sample came from a different species, and all had middling numbers of reads. More reads in the testing data set would have meant that the subset consisted of more of the overall population, but it would also have increased the time and processing power needed to test each of the samples. Given the number of parameter combinations to test, samples with middling numbers of reads were chosen to balance accuracy and practicality. The number of reads also varied per sample, to try and ensure a more diverse and representative subset. As for parameters, Paris *et al* (2017) found several parameter combinations which were extremely unlikely to provide useful data, which enabled them to be eliminated from this testing. They found that setting m below 2 resulted in a severe drop in the number of polymorphic loci present, while setting m above 5 was shown to result in false reads being incorporated into stacks to form loci. As such, only values of m between 2 and 5 (inclusive) were tested. Similarly, values of M above 9 always resulted in over merging of reads with the sample data sets, so only values of M from 1 to 9 inclusive were tested. And finally, given the similarity of M and n, values of n which differed too much from M were found to provide inaccurate data and low values of r80. As such, only values of n which were +/- 1 from the value of M were included. This left 108 parameter combinations to be tested, which was accomplished with a simple shell script to cycle through all combinations and deposit the sequencing results in folders corresponding to the parameters used. The results were assessed and the correct parameters were determined to be m=4, M=4 and n=5. Using these parameters, the entire data set was mapped, and was in fact mapped twice – once where no information was provided to the mapping program aside from morph type and sample name, and once where in addition to this, sample population was also included. This mapping enabled the data set to finally be analysed.

The final key stage of post processing sequence analysis was the statistical analysis of the mapped sequence data, aimed primarily at determining the genetic basis of heterostyly in the sample species. There were two main theories for how heterostyly is passed on down generations, namely that it was passed on through either hemizygous inheritance (where the S locus is a length of DNA only present in one stylar morph) or heterozygous inheritance (where the S locus is present in both morphs, but the alleles differ). As such, the final stage of analysis involved two distinct methods of analysis. Both were aimed at determining the genetics of heterostyly, but one assumed the genetics are hemizygous while the other assumes heterozygosity. Comparing the relative success of each method indicated which mode of inheritance is accurate in this species, while using both methods ensured that, regardless of the actual mode of inheritance, genes associated with heterostyly were identified.

The first method to be used was the method focused on hemizygous inheritance, which was taken from a paper by Scharmann *et al* from 2017, aimed at identifying the genetics of sex determination in pitcher plants. The genetics of sex determination, whether XY or ZW, are usually presumed to be hemizygous, with a sex determining region present in only one of the sexes and not the other. As such, a robust method for identifying hemizygous SNPs was required for the analysis, and to fulfil this Scharmann *et al* developed an algorithm they called “privacy rarefaction”. This privacy rarefaction algorithm, based on samtools, was built on the assumption that at most one of two phenotypes contained a hemizygous region that determined the difference between these phenotypes, but that it was unknown which of the two phenotypes this was. It also assumed that, given any specified set of samples with a 1:1 ratio of phenotypes, a number of SNPs would appear to be specific to one of the two phenotypes due to randomness. As such, to identify whether there are SNPs which are truly specific to one phenotype or the other, the privacy rarefaction algorithm first randomly generated 200 subsets of the given samples, of the smallest possible subset size and guaranteeing a 1:1 ratio of phenotypes. From this, the number of SNPs which appeared to be specific to one phenotype or the other were counted, and compared to a distribution approximating how many SNPs would appear to be specific to one phenotype if the phenotypes were completely interchangeable and had no hemizygous regions, and this was used to calculate a p value by comparing the number of phenotype specific SNPs against the mean number of seemingly specific SNPs expected from an interchangeable set of phenotypes. Any SNPs which differed significantly from this expectation were considered “true” SNPs for this subset. To identify which loci are truly specific to one phenotype or the other, the algorithm was run again. The privacy rarefaction algorithm again generated 200 random subsets, but of an increased size, assessed which loci seemed to have phenotype specific alleles in each subset, and compared these results across subsets, locus by locus. The program repeated, increasing subset size until the subset encompassed the entire sample. Any SNPs identified for one morph at this size were considered “true” hemizygous SNPSs. Example results of this increasing subset approach can be seen in Figure 2.1.

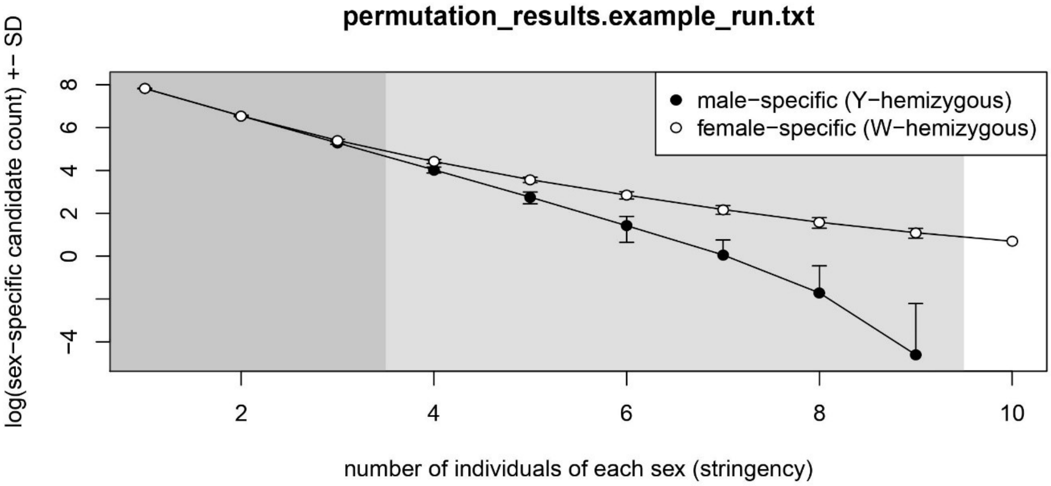


Figure 2.1, showing an example set of data for the privacy rarefaction algorithm. This example data contained no male specific loci and one female specific loci, which can be seen from how the

number of “significant” male loci decreased to nothing as the subset size increased, while the number of significant female specific loci plateaued at one.

If an allele was truly phenotype specific, it should appear as such in all possible one to one subsets, so by comparing across subsets the true specific alleles should have been identified. The issue with this approach was again one of sensitivity versus specificity. The privacy rarefaction algorithm is very specific, and can accurately weed out the vast majority of mistaken specific alleles. However, it makes limited allowance for sequencing error, and as such is much less sensitive. There is a high possibility that true specific alleles were being mistakenly discarded with this approach. As such, this analysis method needed to be paired with a more sensitive method, which is partly covered by the second analysis method used in this protocol.

The other method of analysis, aimed at identifying any alleles associated with heterostyly from heterozygotically inherited regions, used a different program. This analysis method was based on using PLINK, an open source toolset for genome wide analysis developed by Shaun Purcell (Purcell *et al.* 2007). PLINK was tailored to genome wide association studies focused on identifying genes associated with a disease phenotype, which was done comparing the genetics of healthy and infected individuals and using statistical tests to find which genes and which alleles are significantly associated with which phenotype. Thrum samples were classed as “diseased” and pin samples as “healthy” (an arbitrary decision that should not change the final results, with thrum morphs being assigned as “diseased” because they were more likely to have extra genes in the form of a hemizygotic S locus, similar to the extra genes found in certain pathologies) to adapt PLINK to identify genes associated with stylar morphs. Two different sets of tests were done on the samples, each including statistical tests suited to identifying an association between a condition and two groups. A standard chi-squared association test was done, comparing frequencies of the alleles in each stylar morph with frequencies of alleles significantly associated with only one morph (assuming Hardy-Weinberg equilibrium). To support this, a set of alternative tests (Cochran-Armitage trend test and dominant and recessive gene action tests) were also done which did not assume Hardy-Weinberg. The Cochran-Armitage trend test differed from chi-squared in that it could allow assumptions about the data to be incorporated, such as whether a phenotype was controlled by a dominant, recessive, or codominant allele, but this was not relevant in this analysis and it was used only as a replicate chi-squared test. The dominant and recessive gene action tests were similar again, in that one assumed the trait (in this case heterostyly) was dominant (and so the test was more likely to produce a significant result if the alleles of a locus were heterozygous in the thrum group and homozygous in the pin), while the other assumed the reverse. Effectively these were just further chi-squared replicates, but if only one had detected a significant difference it would have been useful information about which morph is dominant. Each of these tests was done with and without Fisher's exact test (a similar test to a chi-squared test, but which allowed an exact calculation of the p value) to ensure any significant association identified was truly significant. These tests each generated a list of each of the loci identified by STACKS, with the probability value that the locus is associated with a single stylar morph provided for each locus. This list was narrowed down to only the loci which are significantly associated with one stylar morph by eliminating all loci with $p>0.05$. In ideal circumstances this final list of significant loci should have consisted of all the heterozygous and hemizygotic loci, but many loci identified were only present in a small number of the samples, leading to false associations.

To filter for these, the alleles which were significantly associated ($p < 0.05$) with one morph were isolated, and only these alleles from loci which were present in 80% of the total sample population were assessed (isolated using a shell script, included in appendix). To account for the possibility of phenotype specific alleles that were incorrectly discarded by privacy rarefaction, significantly associated alleles which were present in at least 80% of the sample population of one morph and absent from at least 80% of the sample population of the other morph were also isolated (using a shell script, included in appendix).

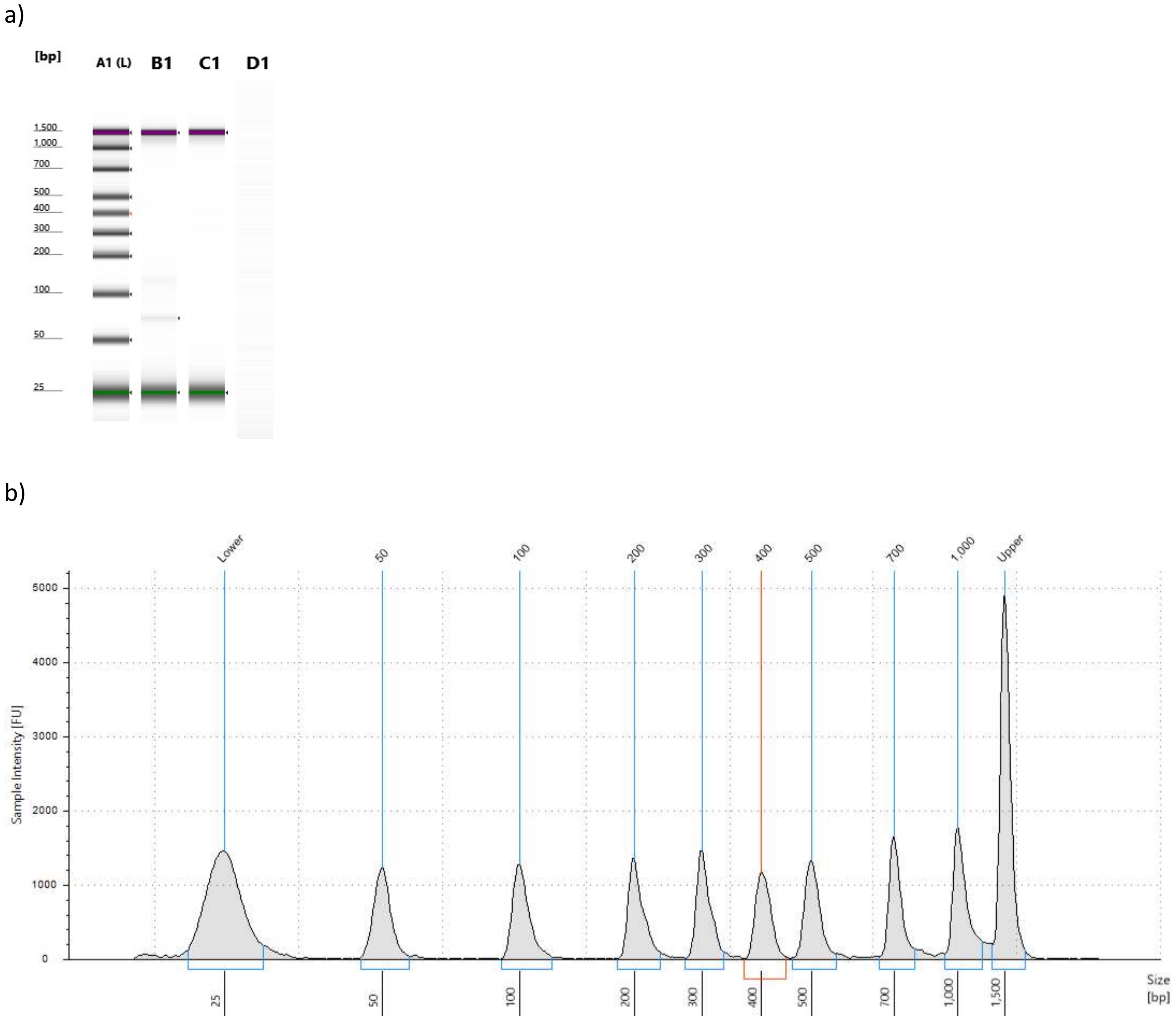
This analysis identified any genes associated with heterostyly in the sample population, and these genes were then compared against a database of known genetic sequences to determine likely function. The consensus sequences (not considering morph specific alleles) of any loci identified this way were analyzed using NCBI's Open Reading Frame finder to identify any open reading frames present in these loci. The proteins encoded by any open reading frames identified were then compared against the UniProt database of known proteins using NCBI's blastp, and any homologous proteins were noted. Each of these homologous proteins had an E value, which is an approximation of how many proteins would be detected in a database this size that were as similar to the input sequence as this homologous protein purely by chance. This allowed these homologous proteins to be narrowed down to significantly homologous proteins by only including homologous proteins with an E value of less than 0.1. As these loci can be assumed be those involved in heterostyly in *Linum tenue*, the function of any of these significantly homologous proteins encoded by open reading frames within these loci should indicate how heterostyly is controlled in *Linum tenue*.

Section 3: Results

Chapter 3.1: Library Construction

Context

There were initially some difficulties in constructing the sequencing library. Despite promising DNA concentrations before size selection (the pooled library concentration was 12.2ng/µl), the Tape Station showed no fragments in the target size range after size selection and PCR (Figure 3.1). The only DNA present in the sample was the upper and lower marker from the Tape Station analysis, meaning that initial library construction left no DNA for further analysis.



c)

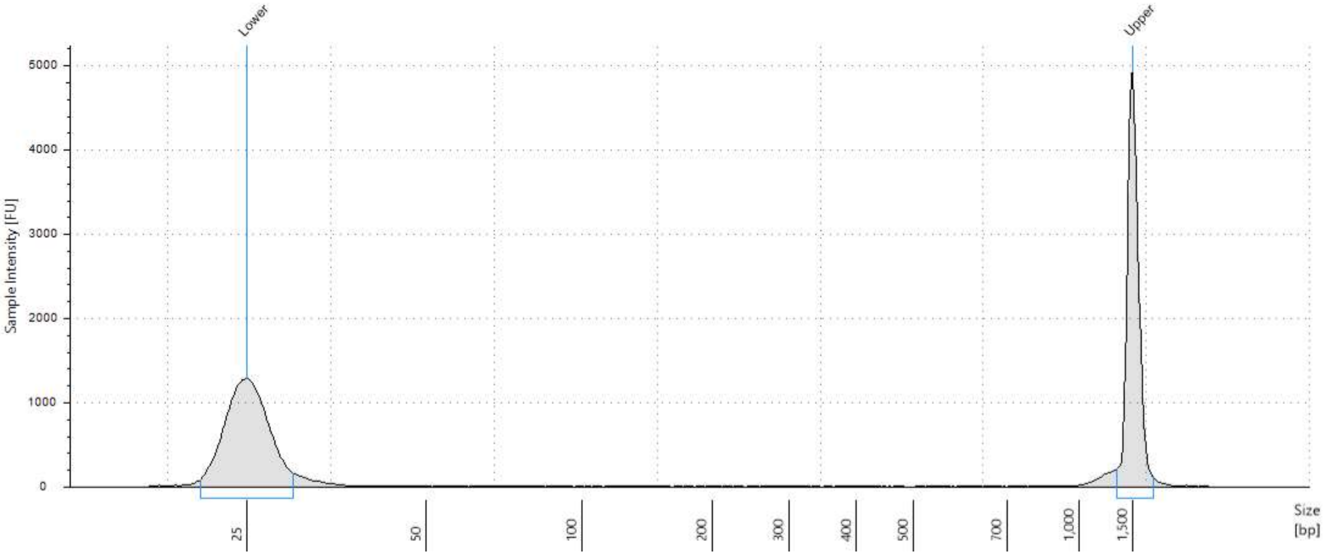
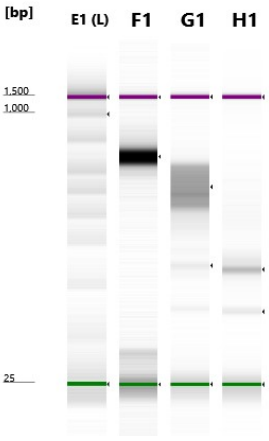


Figure 3.1, showing a) gel electrophoresis of a reference library (A1) and the samples to be sequenced, including high and low end markers (C1), b) the Tapestation output of the reference library and c) the Tapestation output of the samples, again containing high and low end markers.

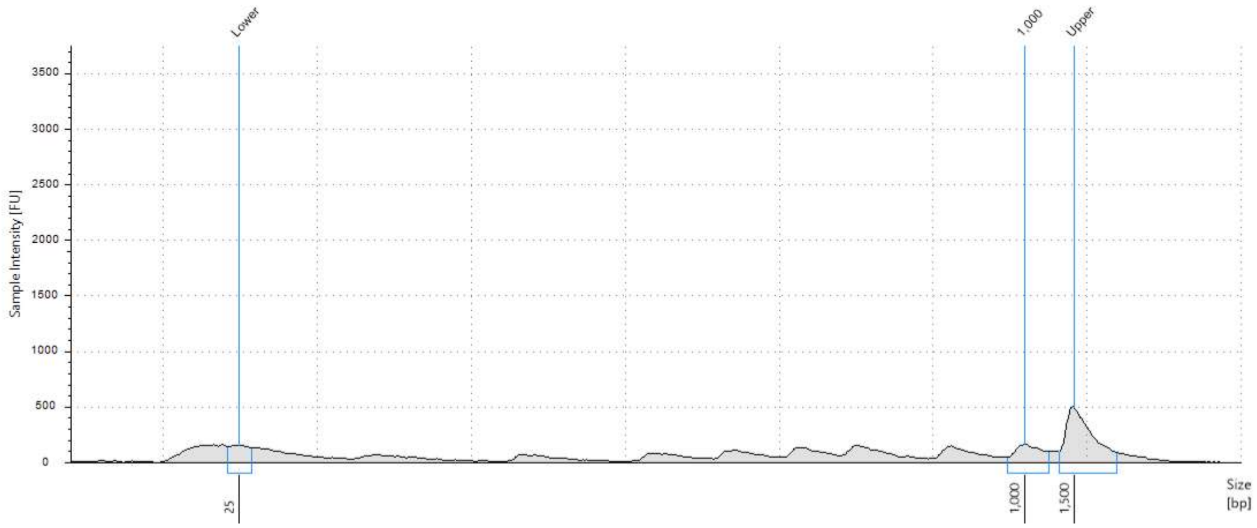
Multiple theories were considered for the cause of this, and the likeliest causes were determined to be overly strict size selection protocols (causing usable DNA to be incorrectly discarded) and insufficient PCR amplification (meaning that any DNA not discarded by size selection was not amplified to a usable level). These theories were supported by measuring the concentration of DNA present after size selection and PCR, which showed an unusably low concentration (1.5ng/µl). The library construction protocol was rerun with post ligation samples, using a less strict size selection protocol and incorporating more cycles of PCR amplification, as specified in the Methods. The resulting sample was again analysed using the Tape Station (Figure 3.2).

Content

a)


[bp]
E1 (L)
F1
G1
H1
1,500
1,000
25

b)


Lower
1,000
Upper
Sample Intensity [FU]
3500
3000
2500
2000
1500
1000
500
0
25
1,000
1,500
Size [bp]

c)

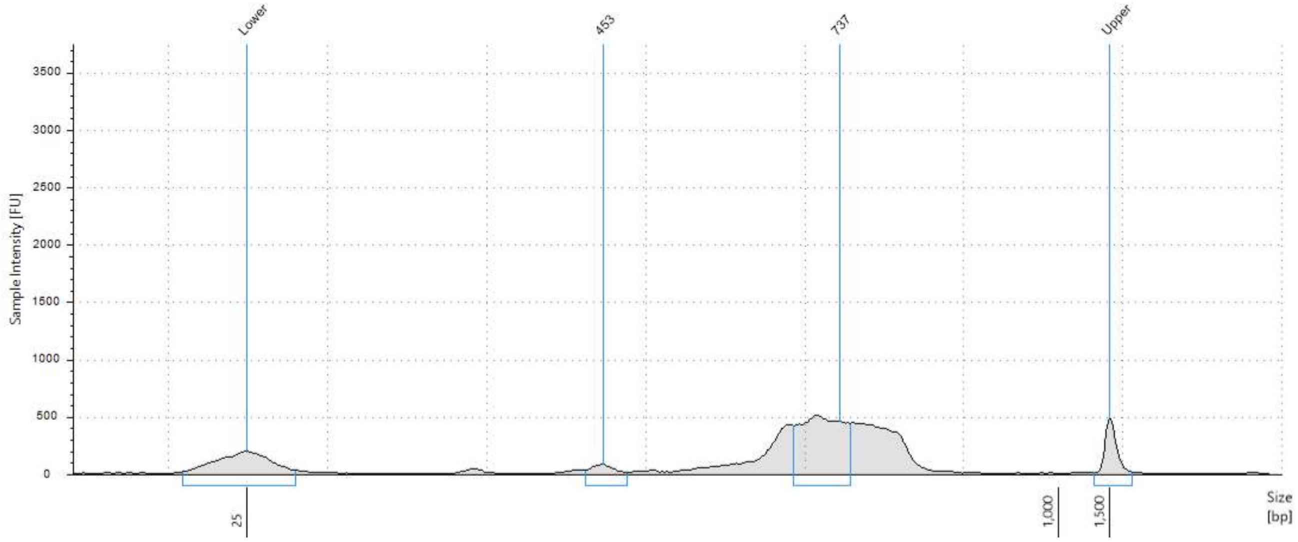

Lower
453
737
Upper
Sample Intensity [FU]
3500
3000
2500
2000
1500
1000
500
0
25
1,000
1,500
Size [bp]

Figure 3.2, showing a) gel electrophoresis of a reference library (E1) and the samples to be sequenced, including high and low end markers (G1), b) the Tapestation output of the reference library and c) the Tapestation output of the samples, again containing high and low end markers.

DNA was clearly present in this sample, meaning that the changes to the library construction protocol did correct the issue identified in the first analysis. However, different problems were visible in this analysis. The peaks are less pronounced than would be expected, and spread over a larger size range. This is consistent with the changes to the size selection and PCR steps, and indicative of somewhat degraded DNA. However, this did not mean that this DNA was unusable in later analysis. More concerning was the position of the peaks, with the main peak indicated as being 737 base pairs in size. This was larger than the expected peak, and there were multiple possible explanations. It was possible that the DNA identified was not the sample DNA, but was some contaminant introduced in the second run of the protocol – in this case the protocol should be redone from before the size selection step. It could be that the assumptions about restriction site distribution in the sample species were inaccurate, leading to larger fragments being produced from digestion, in which case this was a successful library construction protocol and this sample should be forwarded to analysis. However, the most likely explanation was decided to be that the size estimates produced by the Tape Station were inaccurate. The ladder was clearly not properly recorded in this Tape Station analysis, with the pronounced peaks at key markers which were observed in the first Tape Station analysis (Figure 3.1b) not being present in the second analysis (Figure 3.2b). The effect of this on the sample can be seen with the marker DNA, with what should be a sharp peak at 25 base pairs being spread out into a wider, more diffuse peak centred on 25 base pairs.

Conclusion

As such, it was reasonable to assume that the size estimates produced by the Tape Station could be inaccurate, and that the peaks visible were within expected size ranges. As such, this sample could be forwarded to sequencing.

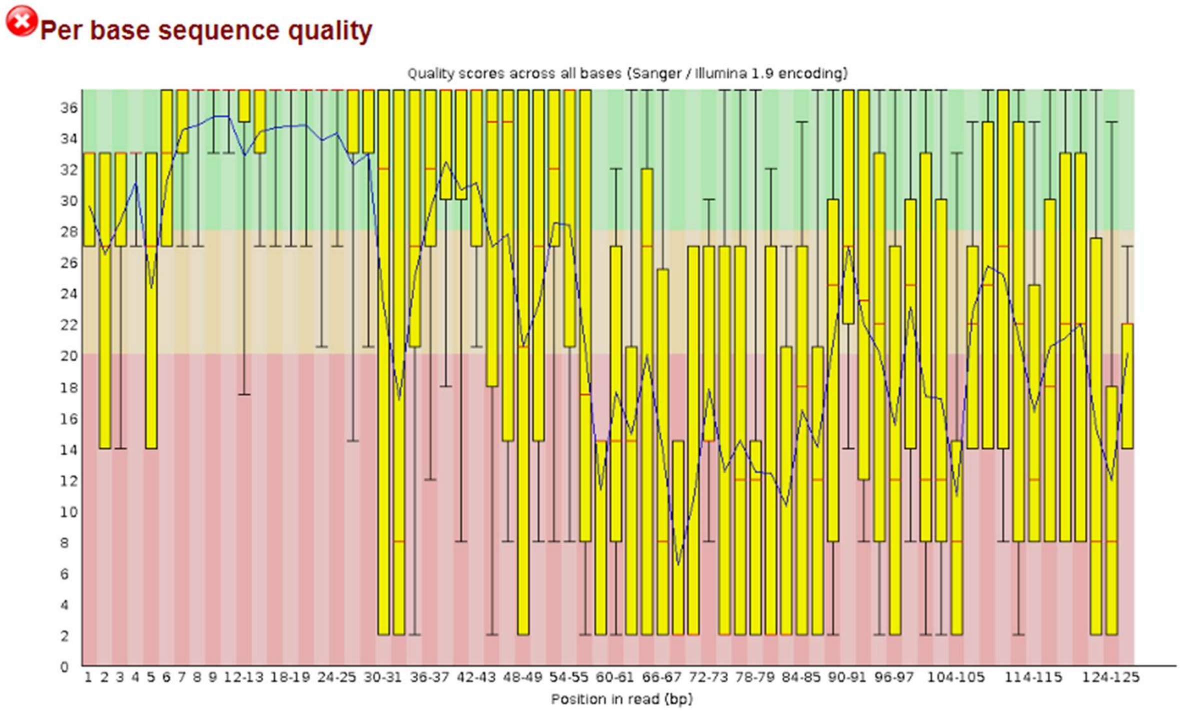
Chapter 3.2: Sequencing Quality

Context

After sequencing was complete, the data collected was analysed using FastQC, to assess the quality of the sequences obtained. The data consisted of a small number of sequences which had been filed under the names of the researchers, and a much larger number of sequences which had been filed as Undetermined, indicating that the sequencer had not identified an adapter sequence on these samples. The sample sequences were thought to be mainly located in the Undetermined sequences, as the adapter sequence used for this project was non-standard and so was not detected. It is worth specifying that there was another sample library being sequenced on the same lane in this experiment. Different barcodes were used, so the samples from this protocol can be isolated in the radtag processing stage, but for the moment the sequences being analysed include sequences from another sample. Both of these collections of sequences were analysed using FastQC, with the results shown below (Figures 3.3, 3.4).

Content

a)



b)

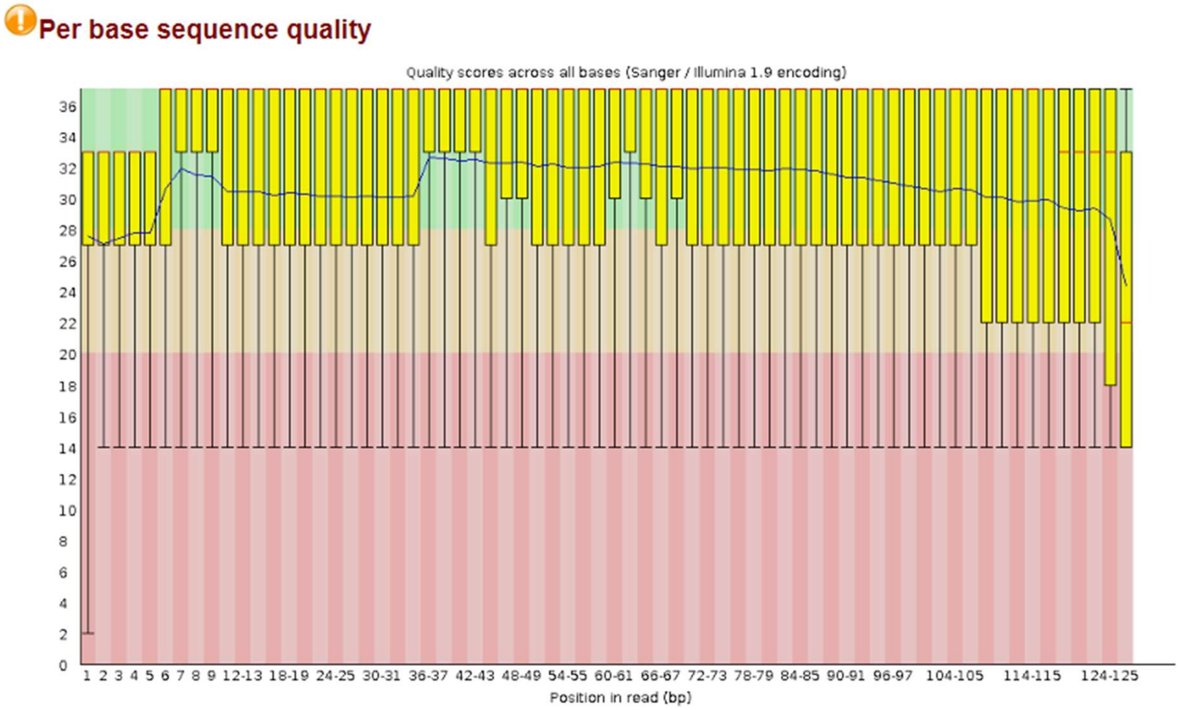
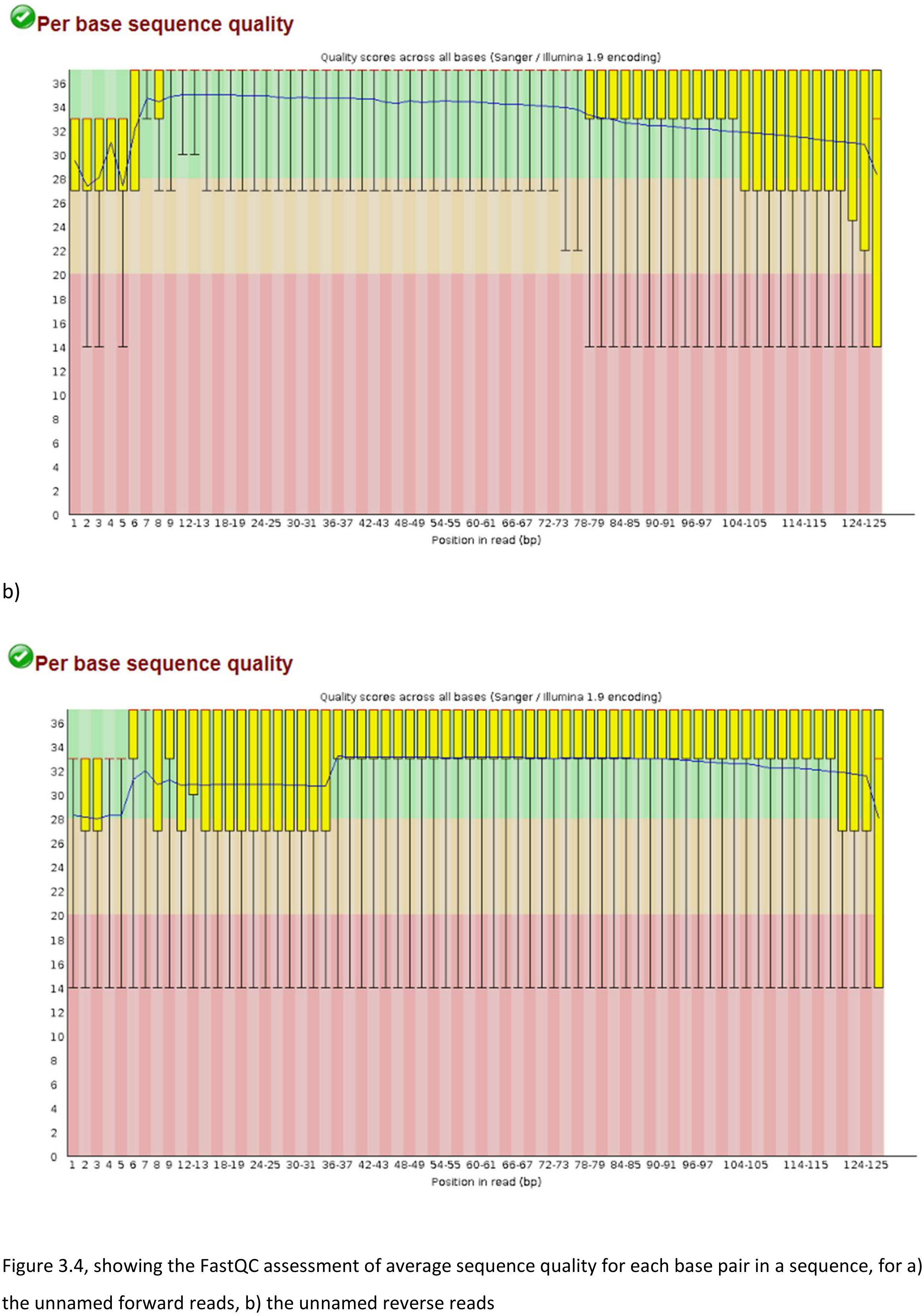


Figure 3.3, showing the FastQC assessment of average sequence quality for each base pair in a sequence, for a) the named forward reads and b) the named reverse reads.

The FastQC results highlighted several key areas for improvement. Firstly, and most noticeably, the quality of the sequencing in the named sequences is very poor (Figure 3.3) – expected given that these sequences were identified as containing an adapter sequence which was not actually present. The forward reads showed wildly varying qualities at different base pairs, with averages frequently dipping below the usability threshold of 20 (Phred score). The reverse reads still showed lower sequencing quality than would be desirable, but the average sequencing quality (by position in read) did not dip below 20 for the majority of the read. This may indicate that these named reads were salavageable, and as such they were included in later analysis steps (including the strict quality control steps applied to all reads, which should mean that no obviously low quality reads were included in mapping). Secondly, although the Undetermined reads showed better sequencing quality than the filed reads (Figure 3.4), there were still some visible problems.

a)

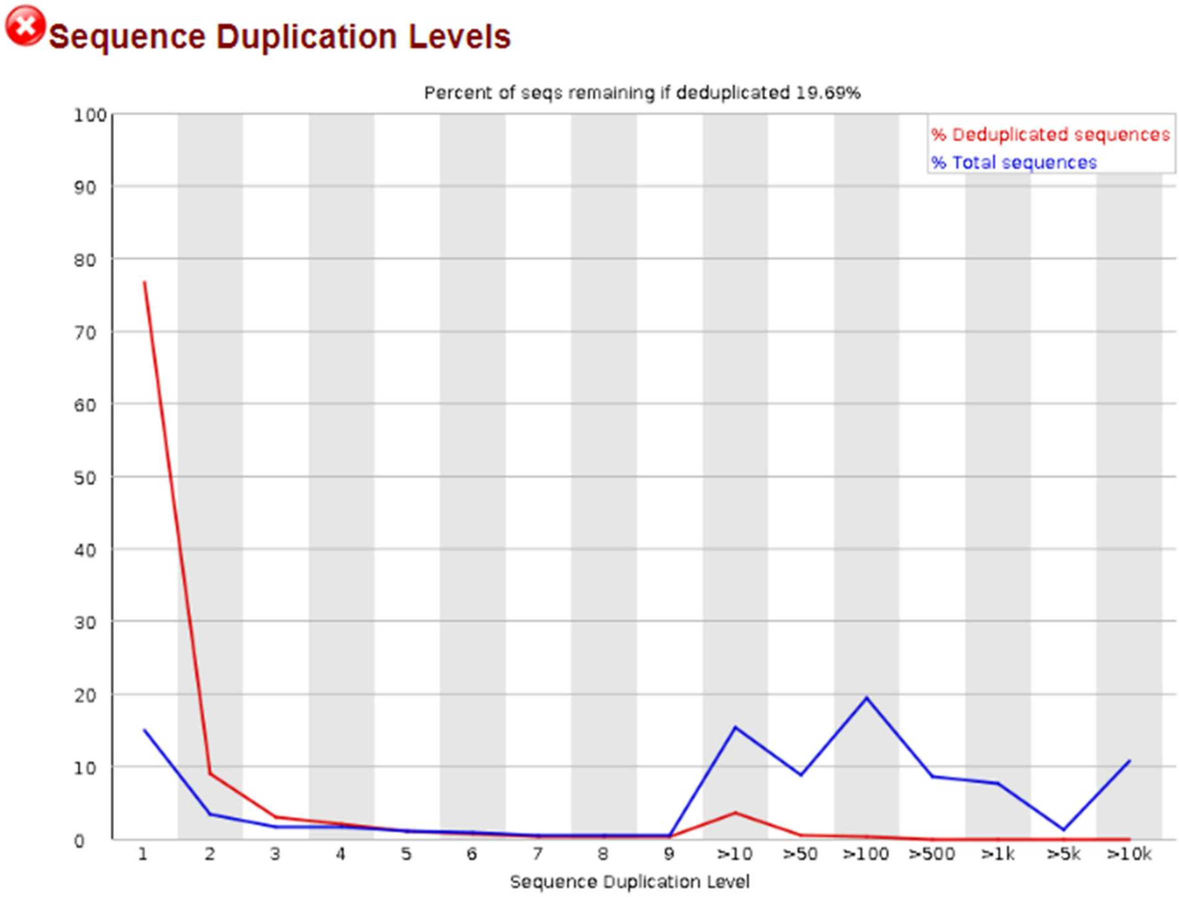


b)

Figure 3.4, showing the FastQC assessment of average sequence quality for each base pair in a sequence, for a) the unnamed forward reads, b) the unnamed reverse reads

The sequencing quality dipped noticeably at the beginning and especially at the end of the reads, in the latter case the average decreasing below the 20 threshold. Although the average sequencing quality was high for the majority of the read, there was still a large range of sequencing quality present in the Undetermined reads. There were several reverse reads included that had a sequencing quality score of 14 across the entire read. This highlighted the need for a quality threshold for reads to be included in later analysis, to ensure that the low quality outliers do not cause problems for the results. Thirdly, aside from the quality issues the distribution of the Undetermined reads also presented some concerns (Figure 3.5).

a)



b)

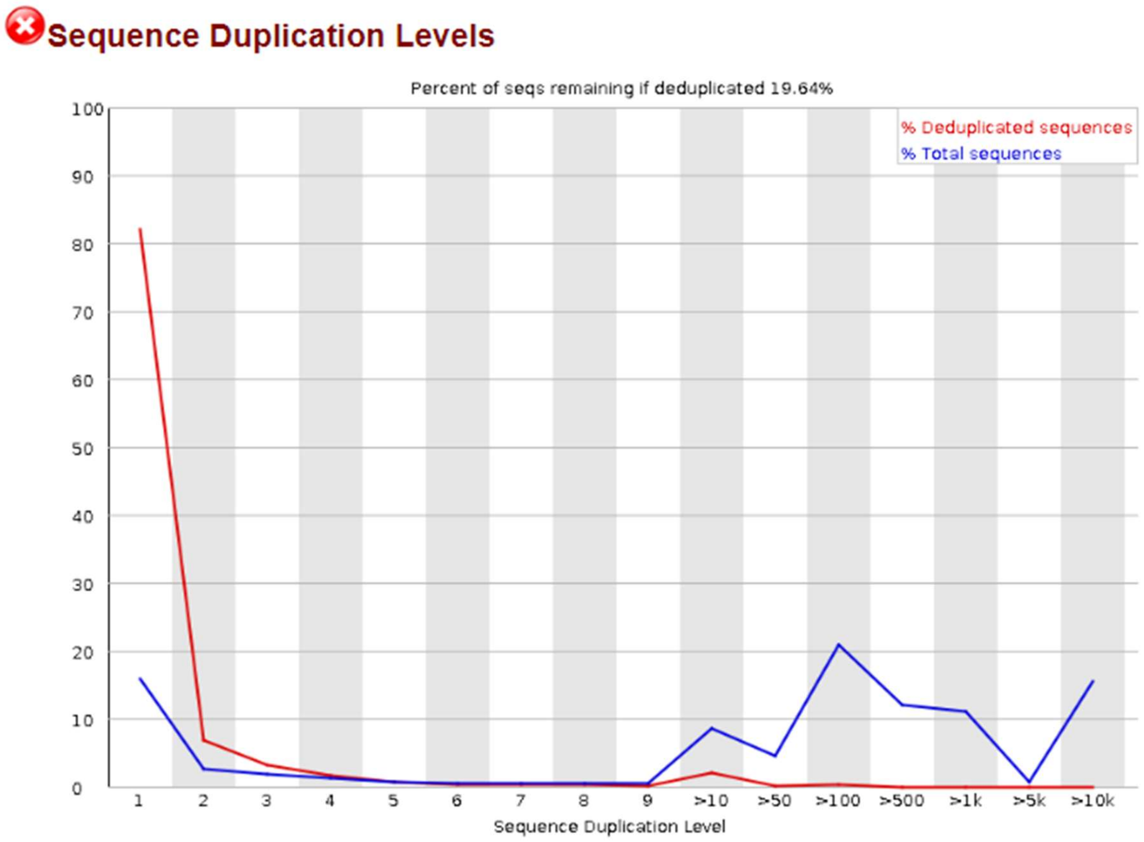
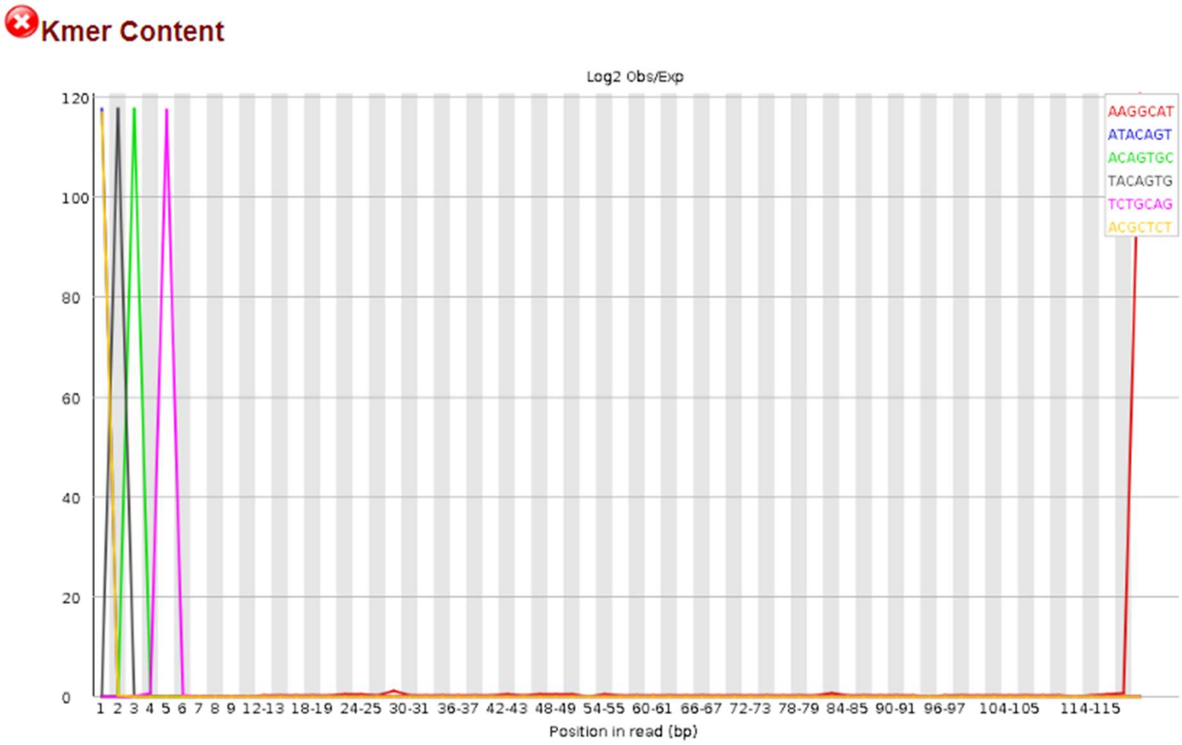


Figure 3.5, FastQC results of a) the percentage of duplicated reads in the undetermined forward reads, showing the percentage of unique sequences formed by duplicated sequences for each sequence duplication level (red) and the percentage of total sequences formed by duplicated sequences for each sequence duplication level (blue), b) the percentage of duplicated reads in the undetermined reverse reads, showing the percentage of unique sequences formed by duplicated sequences for each sequence duplication level (red) and the percentage of total sequences formed by duplicated sequences for each sequence duplication level (blue),

There was a significant number of sequences in both the forward and reverse Undetermined reads which were duplicated many times, including a significant percentage which were duplicated more than ten thousand times. In particular, a relatively large number of individual sequences (seen in the deduplicated sequence data) were duplicated between 10 and 100 times, much higher than the number of individual sequences duplicated >100 times (though the latter still make up a larger proportion of the total library However, looking at which sequences which were overrepresented in the data, it seems that these overrepresented sequences actually corresponded to those from Illumina primer and adapter sequences, and so the problematic duplication levels actually just indicated that the nonstandard adapter sequences were not properly identified and filtered prior to FastQC analysis – which is expected given that these are the Undetermined reads, or the reads where the adapter sequences were explicitly not identified. The number of unique sequences duplicated between 10 and 100 times (seemingly ~5% of the total number of unique sequences) is not explained by this, however, this duplication did correspond to the amount of unique samples of the same species sequenced in this analysis. As such this relatively large number of duplicated sequences could be explained as shared species genetics. Fourthly, the distribution of kmers (short sequences of nucleotide bases, of length k) across the read was seen to be unusual (Figure 3.6).

a)



b)

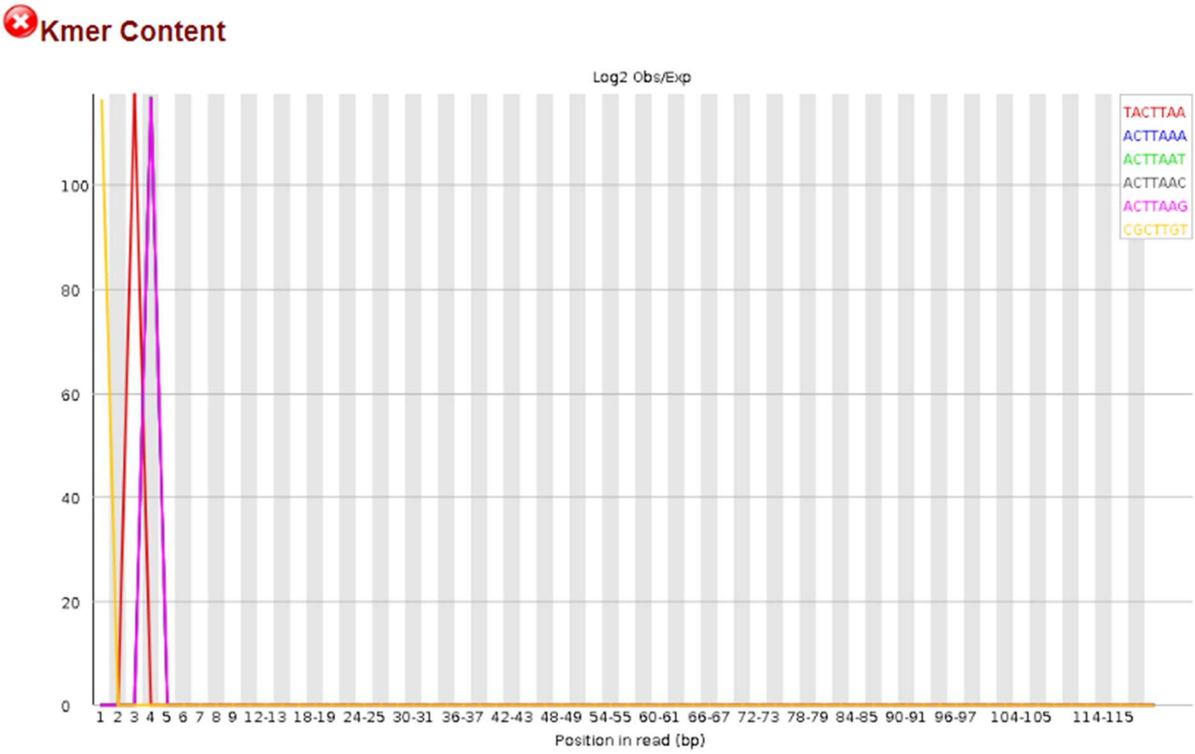


Figure 3.6, showing a) the relative enrichment of the 6 most biased kmers (7 base pair sequences) for each base pair position of the undetermined forward reads and b) the relative enrichment of the 6 most biased kmers (7 base pair sequences) for each base pair position of the undetermined reverse reads.

A disproportionate number of certain kmers were found at the beginning and end of the read. Howvever, this is expected given the adapter and radtag sequences which should be found at these positions in the majority of these reads, and so this disproportionate kmer distribution was not a cause for concern. Finally, in the Undetermined reverse reads the amount of N (uncalled) bases was worryingly high. (Figure 3.7)

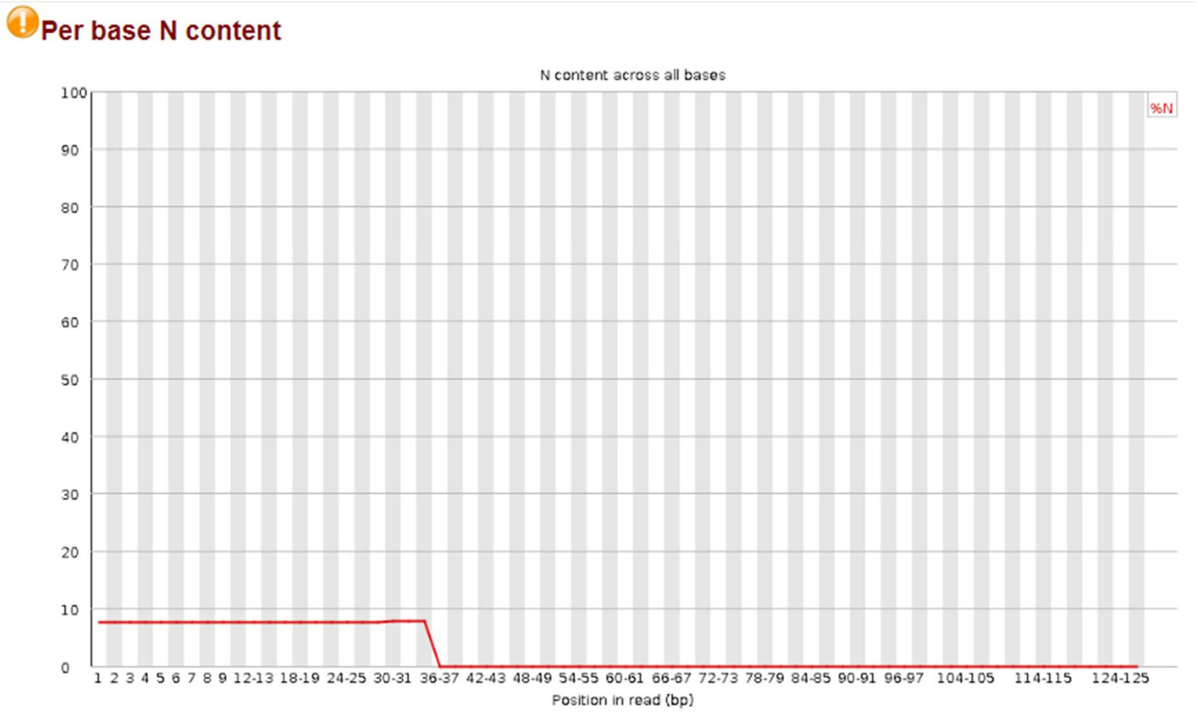


Figure 3.7, FastQC results showing the percentage of undetermined reverse sequences with an uncalled base in each base pair position.

As can be seen, approximately 8% of the total reads in positions 0 – 36bp were unable to be called. This might have severe implications on further analysis, however, FastQC also showed that 7.8% of the total library of Undetermined reverse reads was composed of a single duplicated sequence: this sequence consisted of 35 N's. This sequence was filtered out of further analysis by including a minimum length for reads to be mapped. As such, the high N content did not pose a problem for further analysis.

Conclusion

In summation, FastQC showed that the named sequences were of significantly low quality and that there were some issues with the Undetermined sequences, but that these issues could be corrected with suitable cleaning steps. As such, these sequences were able to proceed to further analysis.

Chapter 3.3: Post Radtag Processing Coverage

Context

Cleaning the data and processing radtags were key steps to consider in the post processing analysis pipeline, as there was a significant risk of discarding useful data along with low quality sequences and sample identifiers. If too many sequences (regardless of quality, to some extent) were discarded in this step, later analysis would have been severely compromised. As such, the number of reads retained after cleaning and radtag processing was monitored, and if the amount of reads was too low to continue into mapping, changes were made to the cleaning and processing steps and they were rerun on the original data. Initially, the cleaning and processing steps were completely separate, run by two different programs – Trimmomatic and Stacks respectively. However, integrating the Trimmomatic results into the Stacks pipeline presented some challenges. To simplify this, the decision was made to combine the cleaning and the processing step into a single command line in Stacks. To ensure the success of this, the Stacks processing program was run with cleaning parameters and without, and the amount of retained reads from each were compared (Tables 3.1, 3.2).

Content

Sample	Reads Identified	No Rad Site	Low Quality	Retained	Proportion Retained
ALT_05_PIN	53692760	3003454	43950	50645356	94.32436701
ALT_24_PIN	765822	16437	306	749079	97.81372173
ARA_19_PIN	1719464	144995	1815	1572654	91.46187417
BUR_20_THRUM	2159642	30759	2298	2126585	98.46932964
BUR_25_THRUM	95058	7109	109	87840	92.40674115
CAZ_12_THRUM	990466	36202	1209	953055	96.22288902
CAZ_16_THRUM	133232	18984	128	114120	85.65509787
CAZ_20_THRUM	2530000	82589	2045	2445366	96.65478261
CAZ_22_THRUM	9149996	289592	9801	8850603	96.72794392
CAZ_23_THRUM	399804	23790	350	375664	93.9620414
CBT_16_THRUM	158870	27131	219	131520	82.78466671
CBT_17_THRUM	1809286	42793	710	1765783	97.59557085
CBT_18_PIN	57155072	2067929	32295	55054848	96.32539348

CBT_20_PIN	49652380	1855155	42801	47754424	96.17751254
EBO_14_THRUM	2238546	98260	2545	2137741	95.49685376
EBO-16_PIN	68248	4055	39	64154	94.00128942
ELB_30_PIN	1464060	19751	607	1443702	98.60948322
LUM_15_THRUM	171436	34291	418	136727	79.75396066
LUM_22_PIN	141610	13208	131	128271	90.58046748
LUM_25_THRUM	938448	59037	702	878709	93.63427702
MDA_24_THRUM	373766	78853	1003	293910	78.63476079
PIG_11_THRUM	771060	28974	639	741447	96.15944284
PIG_12_PIN	1895946	39320	866	1855760	97.88042486
PIG_13_PIN	833078	16732	987	815359	97.87306831
PIG_15_PIN	179046	48635	1502	128909	71.99769892
PIG_22_PIN	871548	17623	442	853483	97.92725128
PIG_29_PIN	198122	9215	169	188738	95.2635245
PIG_35_PIN	728428	36626	412	691390	94.91535196
SVT_17_THRUM	28416486	847808	11748	27556930	96.97515027
SVT_24_PIN	461238	90198	282	370758	80.38322948

Table 3.1, showing the total number of identified reads, the number of reads discarded due to an unidentifiable restriction digestion site, the number of reads discarded due to low quality, the number of retained reads and the percentage of total reads retained, for each sample, for the process_radtag step with no cleaning command

Sample	Reads Identified	No Rad Site	Low Quality	Retained	Proportion Retained
ALT_05_PIN	53692760	1832581	28458420	23401759	43.58457081
ALT_24_PIN	765822	9495	377766	378561	49.43198289
ARA_19_PIN	1719464	72703	938262	708499	41.20464284
BUR_20_THRUM	2159642	18775	1089010	1051857	48.70515576
BUR_25_THRUM	95058	2563	52322	40173	42.26156662
CAZ_12_THRUM	990466	20513	516717	453236	45.75987464
CAZ_16_THRUM	133232	10508	77618	45106	33.85522997
CAZ_20_THRUM	2530000	48023	1269773	1212204	47.91320158
CAZ_22_THRUM	9149996	152264	4946030	4051702	44.28091553

CAZ_23_THRUM	399804	13517	199438	186849	46.73515022
CBT_16_THRUM	158870	15879	88089	54902	34.55781457
CBT_17_THRUM	1809286	26420	871562	911304	50.36815628
CBT_18_PIN	57155072	1241954	30591301	25321817	44.30370939
CBT_20_PIN	49652380	1172973	27422945	21056462	42.40775971
EBO_14_THRUM	2238546	54206	1190665	993675	44.38930449
EBO-16_PIN	68248	2256	35298	30694	44.9742117
ELB_30_PIN	1464060	11793	718461	733806	50.1213065
LUM_15_THRUM	171436	19122	108018	44296	25.83821368
LUM_22_PIN	141610	7245	74541	59824	42.24560412
LUM_25_THRUM	938448	33689	493694	411065	43.8026401
MDA_24_THRUM	373766	44335	204523	124908	33.41877003
PIG_11_THRUM	771060	16180	395174	359706	46.65084429
PIG_12_PIN	1895946	23562	913128	959256	50.59511189
PIG_13_PIN	833078	9359	437256	386463	46.38977383
PIG_15_PIN	179046	29269	88687	61090	34.11972342
PIG_22_PIN	871548	10301	426308	434939	49.90419346
PIG_29_PIN	198122	5869	107240	85013	42.90941945
PIG_35_PIN	728428	20807	378283	329338	45.21215549
SVT_17_THRUM	28416486	545062	15908938	11962486	42.09699257
SVT_24_PIN	461238	47790	241033	172415	37.38091831

Table 3.2, showing the total number of identified reads, the number of reads discarded due to an unidentifiable restriction digestion site, the number of reads discarded due to low quality, the number of retained reads and the percentage of total reads retained, for each sample, for the process_radtag step with a cleaning command integrated, with cleaning parameters matching Trimmomatic.

The difference in number of reads retained demonstrated that the cleaning commands were being successfully executed within the Stacks process_radtags command line. However, the number of reads retained in this step was concerningly low. To remedy this, the number of mismatches allowed between the barcodes listed and the barcodes found in the reads was increased, from one to two. The results are included below (Table 3.3).

Sample	Reads Identified	No Rad Site	Low Quality	Retained	Proportion Retained
ALT_05_PIN	54525198	1992821	28962499	23569878	43.2274964
ALT_24_PIN	848848	26448	424438	397962	46.88259853
ARA_19_PIN	1799364	100925	991888	706551	39.26670757
BUR_20_THRUM	2318714	46039	1181181	1091494	47.07324836
BUR_25_THRUM	141708	12734	79866	49108	34.65435967
CAZ_12_THRUM	1067152	37110	560272	469770	44.02090799
CAZ_16_THRUM	153414	24225	88278	40911	26.66705777
CAZ_20_THRUM	2947308	151449	1478414	1317445	44.69994313
CAZ_22_THRUM	9884144	201049	5379436	4303659	43.54103906
CAZ_23_THRUM	451636	22515	227955	201166	44.54162201
CBT_16_THRUM	258172	44738	140701	72733	28.17230374
CBT_17_THRUM	1850478	38131	896099	916248	49.51412554
CBT_18_PIN	57931442	1375110	31113718	25442614	43.91848903
CBT_20_PIN	61531170	1748899	33976757	25805514	41.93892949
EBO_14_THRUM	2589418	91160	1394896	1103362	42.61042443
EBO-16_PIN	129644	10726	68277	50641	39.06158403
ELB_30_PIN	1582512	30319	785157	767036	48.46952187
LUM_15_THRUM	303004	59741	178775	64488	21.28288735
LUM_22_PIN	236214	22951	125673	87590	37.08078268
LUM_25_THRUM	1327822	95241	697467	535114	40.30013059
MDA_24_THRUM	807446	100866	455983	250597	31.03575967
PIG_11_THRUM	867302	37457	443836	386009	44.50687304
PIG_12_PIN	2022756	45567	986131	991058	48.99543
PIG_13_PIN	883264	17387	466141	399736	45.2566843
PIG_15_PIN	231142	41546	117268	72328	31.29158699
PIG_22_PIN	946684	23597	468558	454529	48.01274765
PIG_29_PIN	331400	18609	181473	131318	39.62522631
PIG_35_PIN	1016828	53015	537444	426369	41.93128041
SVT_17_THRUM	36123942	914463	20327883	14881596	41.19593593
SVT_24_PIN	609094	81123	323030	204941	33.64685911

Table 3.3, showing the total number of identified reads, the number of reads discarded due to an unidentifiable restriction digestion site, the number of reads discarded due to low quality, the number of retained reads and the percentage of total reads retained, for each sample, for the process_radtag step with a cleaning command and two mismatches allowed between listed barcode and barcode in the sequence

While this did result in more reads being assigned to each sample, and consequently more reads being retained, the number of retained reads did not increase in proportion to the number of reads identified. Rather, more reads were filtered out by the cleaning and processing steps, leaving a small increase in the number of retained reads and a decrease in the proportion of reads retained (as a percentage of overall reads identified for each sample). As such, the number of reads retained was still concerning, and needed to be increased before mapping. It was identified that a significant number of the reads filtered out of in the processing stage were discarded due to issues identifying the restriction enzyme digestion site, or RAD site – so to improve the number of reads retained, the RAD checking in Stacks was disabled. If the sequences with unidentifiable RAD sites were low-quality sequences, the cleaning step should remove them from the final data set regardless, and an increase in Low Quality reads matching the previous number of No RAD site reads should be seen. However, it was possible that the RAD check was more stringent than the cleaning step, leading to reads within an acceptable quality threshold being incorrectly discarded, or that the sequencing errors in the discarded reads were localised to the RAD site, and that the rest of the read would still be acceptable. As such, disabling RAD checking was a possible way to increase the number of retained reads. This was tested, and the results are included below (Table 3.4).

Sample	Reads Identified	No Rad Site	Low Quality	Retained	Proportion Retained
ALT_05_PIN	54525198	0	29705132	24820066	45.52035923
ALT_24_PIN	848848	0	445038	403810	47.57153224
ARA_19_PIN	1799364	0	1050107	749257	41.64010172
BUR_20_THRUM	2318714	0	1214643	1104071	47.61566109
BUR_25_THRUM	141708	0	88862	52846	37.29217828
CAZ_12_THRUM	1067152	0	580013	487139	45.64851118
CAZ_16_THRUM	153414	0	104301	49113	32.01337557
CAZ_20_THRUM	2947308	0	1558837	1388471	47.10980325
CAZ_22_THRUM	9884144	0	5483578	4400566	44.52146792
CAZ_23_THRUM	451636	0	238718	212918	47.14371751
CBT_16_THRUM	258172	0	168430	89742	34.76054723
CBT_17_THRUM	1850478	0	917406	933072	50.42329603
CBT_18_PIN	57931442	0	31736246	26195196	45.21757977
CBT_20_PIN	61531170	0	35220673	26310497	42.75962411
EBO_14_THRUM	2589418	0	1448691	1140727	44.05341277
EBO-16_PIN	129644	0	74774	54870	42.32359384
ELB_30_PIN	1582512	0	808736	773776	48.89542702
LUM_15_THRUM	303004	0	230071	72933	24.06997927
LUM_22_PIN	236214	0	143010	93204	39.45744113
LUM_25_THRUM	1327822	0	753868	573954	43.22522145
MDA_24_THRUM	807446	0	513574	293872	36.3952512
PIG_11_THRUM	867302	0	461623	405679	46.77482584
PIG_12_PIN	2022756	0	1016843	1005913	49.72982406
PIG_13_PIN	883264	0	476602	406662	46.04082132
PIG_15_PIN	231142	0	134499	96643	41.81109448
PIG_22_PIN	946684	0	483830	462854	48.89213296
PIG_29_PIN	331400	0	195517	135883	41.00271575
PIG_35_PIN	1016828	0	577128	439700	43.24231827
SVT_17_THRUM	36123942	0	21115125	15008817	41.5481151
SVT_24_PIN	609094	0	384835	224259	36.81845495

Table 3.4, showing the total number of identified reads, the number of reads discarded due to an unidentifiable restriction digestion site, the number of reads discarded due to low quality, the number of retained reads and the percentage of total reads retained, for each sample, for the process_radtag step with a cleaning command, two barcode mismatches allowed and no check for a restriction enzyme digestion site

While there was an increase in the number of Low Quality reads visible, it seems that the majority of the No RAD site reads were within the quality threshold set by the cleaning step, and that they could be retained – resulting in a significant increase in the number of retained reads. While the number of retained reads still was not ideal for some of the samples included, it was decided that this was acceptable for further analysis.

Conclusion

The number of reads retained for the majority of the samples suggests adequate coverage would be obtained in mapping, and this could be further ensured by optimising the parameters used for mapping.

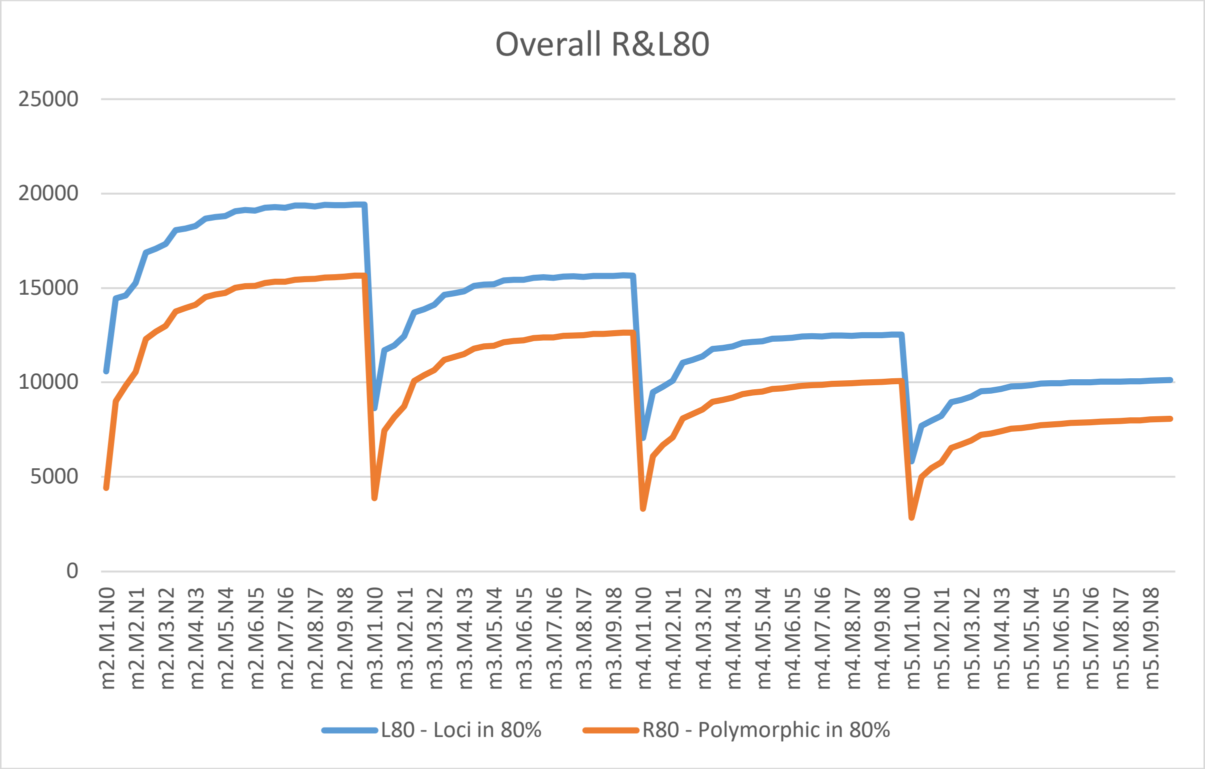
Chapter 3.4: Mapping With Variable Parameters

Context

Determining mapping parameters was a key step in mapping the sequence data. The ideal value for these parameters is entirely dependent on the level of polymorphism in the data set used, and so is unique to any individual data set. As such, the only way to determine what values for these parameters are best for the data is to test a range of possible values on a subset of the total dataset. For each value, three outcomes were measured – the total number of loci present in at least 80% of the samples (referred to as L80), the number of polymorphic loci present in at least 80% of the samples (referred to as R80) and the coverage for each individual sample. The results are summarised below (Figure 3.8), with the complete data set included in the appendices.

Content

a)



b)

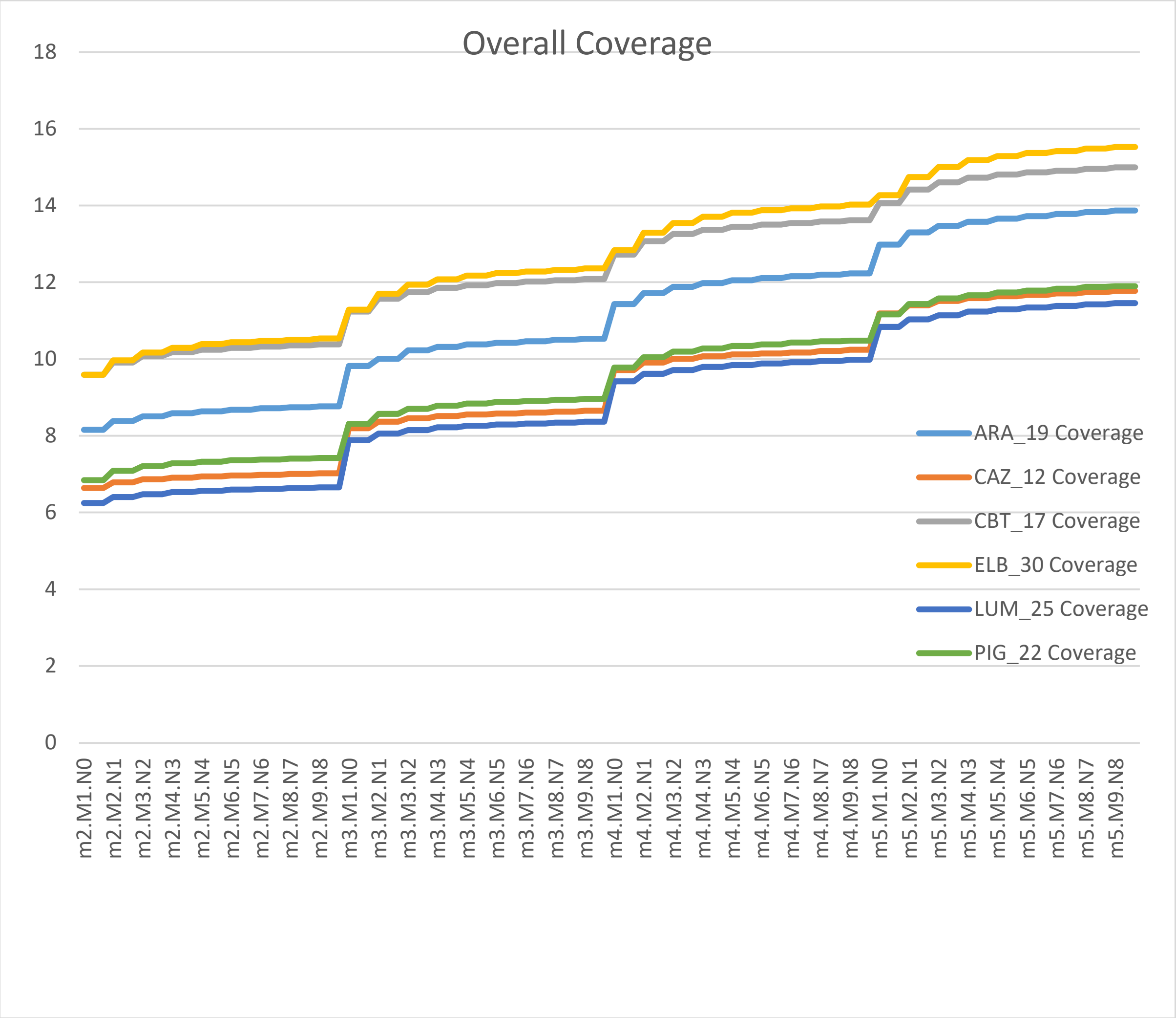
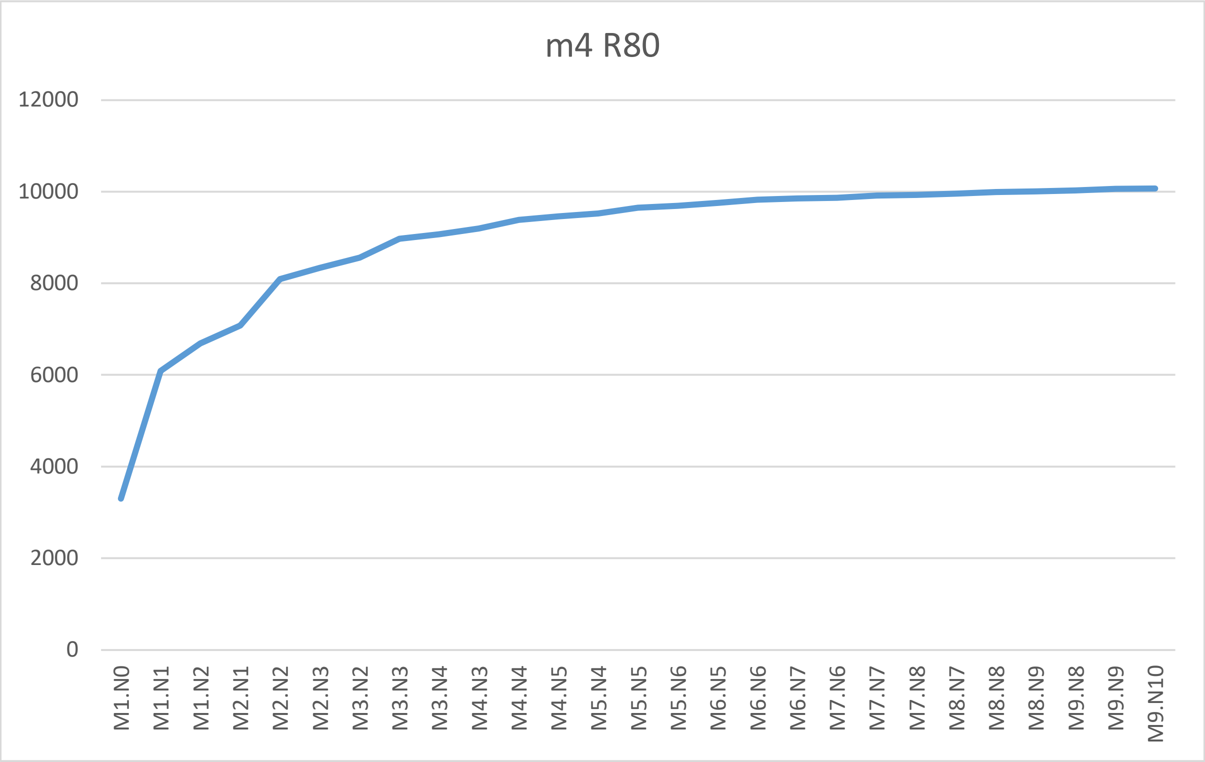


Figure 3.8, showing a) the total number of loci in 80% of the sample subset (L80, blue) and the total number of polymorphic loci in 80% of the sample subset (R80, orange) for each value of m, M and N tested and b) the coverage for each sample in the sample subset for each value of m, M and N tested.

A few things were immediately visible from this data. Increasing m caused a significant increase in coverage and a significant decrease in both L80 and R80, as expected. Increasing M and N increased coverage and both R80 and L80, but not at a uniform rate – the increases plateaued at a certain value of M and N, where further increasing M and N past this point only results in minor increases in coverage, R80 and L80. While these further increases were still positive, larger values of M and N increased the risk of overmerging loci. To minimize this risk while still maximising the increases in R80, L80 and coverage, the values of M and N where increases begin to plateau were chosen as the ideal values for mapping. Setting m required balancing the number of loci created and the coverage per sample. Fountain *et al.* (2016) recommended a minimum coverage per sample of at least 10x, so the minimum value of m was found where, at the plateau point of M and N, the coverage for all samples was at least 10x. This was found to be when m is set to 4 (Figures 3.8, 3.9).

a)



b)

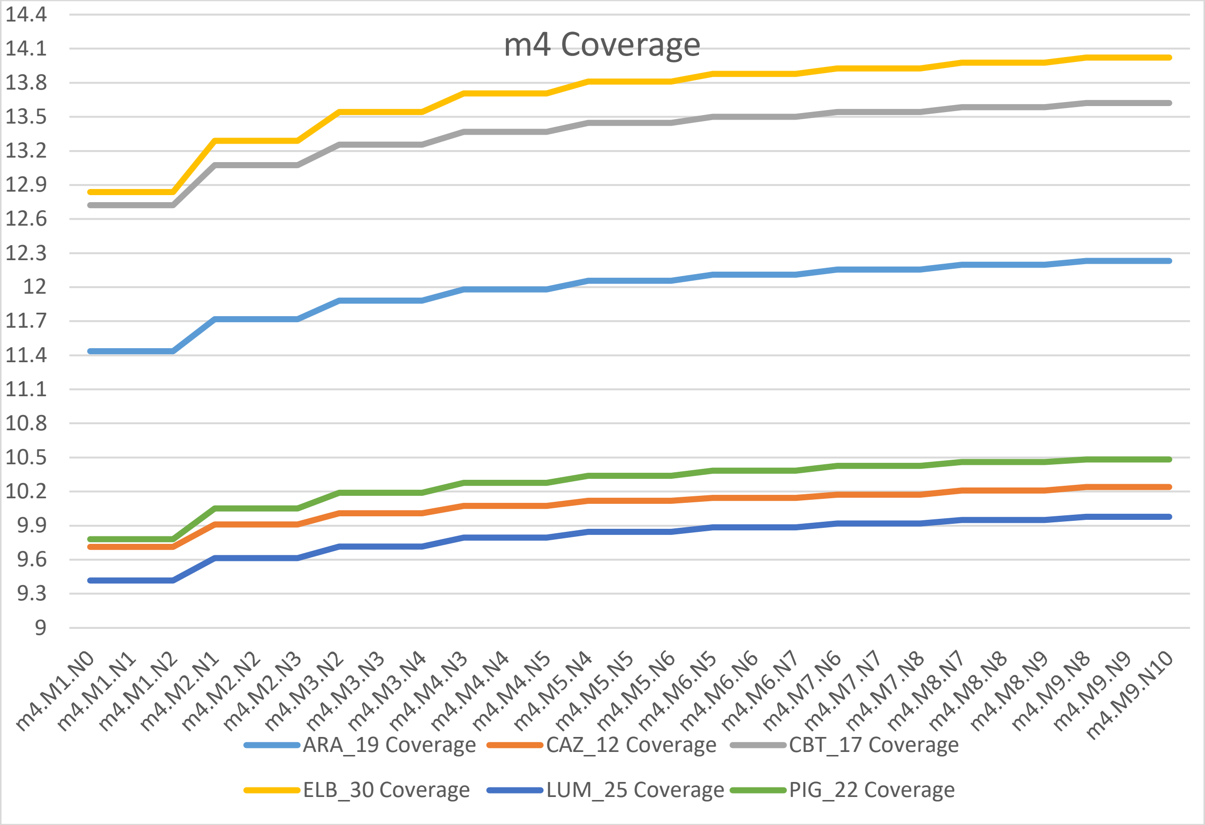


Figure 3.9, showing a) the total number of polymorphic loci in 80% of the sample subset (R80, blue) for each value of M and N tested when m is set to 4 and b) the coverage for each sample in the sample subset for each value of M and N tested when m is set to 4.

When m was set to 4, R80 plateaued when M is equal to 4. At this point, coverage for 5 out of 6 of the samples tested was greater than 10, with the only sample below 10x coverage being LUM_25 (with a coverage of 9.8). Increasing LUM_25's coverage to above 10x would have required increasing m to 5, with a consequent loss of 2000 loci – as LUM_25 was included in the analysis as a lower bound for coverage, and its coverage was not too far below 10x, it was decided that this would be an unnecessary sacrifice and that m should be set to 4. To maximise R80 at m and M = 4, N was set to 5, and these were the parameters used to map the complete sample. While checking R80 for the complete set of final mapped samples would have been unfeasible – it would require comparing all loci sequenced for each sample against all other samples, and was only feasible in the limited subset because the samples included were not those with the highest amount of retained reads – coverage was automatically calculated by Stacks. The coverage obtained for each sample is thus presented below (Table 3.5), and was consistent for both mapping sets (with population information and without).

Sample	Coverage
ALT_05_PIN	68.27
ALT_24_PIN	9.18
ARA_19_PIN	12.26
BUR_20_THRUM	15.78
BUR_25_THRUM	6.24
CAZ_12_THRUM	10.61
CAZ_16_THRUM	6.39
CAZ_20_THRUM	20.43
CAZ_22_THRUM	35.57
CAZ_23_THRUM	7.47
CBT_16_THRUM	6.68
CBT_17_THRUM	14.04
CBT_18_PIN	80.07
CBT_20_PIN	95.44
EBO_14_THRUM	7.14

EBO-16_PIN	16.14
ELB_30_PIN	14.59
LUM_15_THRUM	8.21
LUM_22_PIN	6.66
LUM_25_THRUM	11.36
MDA_24_THRUM	7.92
PIG_11_THRUM	9.96
PIG_12_PIN	16.7
PIG_13_PIN	10.57
PIG_15_PIN	6.6
PIG_22_PIN	10.96
PIG_29_PIN	9.34
PIG_35_PIN	10.66
SVT_17_THRUM	84.72
SVT_24_PIN	7.7

Table 3.5, showing the coverage obtained for each sample in the final mapping

The extreme range in coverage obtained, while expected from the range of reads retained, was still concerning. It was especially worrying that despite the considerations taken in setting parameters, a significant fraction of the samples showed coverage significantly lower than 10x. However, the coverage for all samples was at least significantly higher than 5. Fountain et al (2016) found that coverages between 5-10x were significantly more error prone than those higher than 10x, but also showed that coverages lower than 5x were almost unusable, due to the risk of error. At minimum, some analysis could still be done on all the samples, even if the results would need further corroboration.

Conclusion

As such, the decision was made to use these mapping parameters, and move on to the next stages of analysis – morph associated analysis in PLINK, and morph specific analysis with privacy rarefaction.

Chapter 3.5: Morph specific analysis – privacy rarefaction

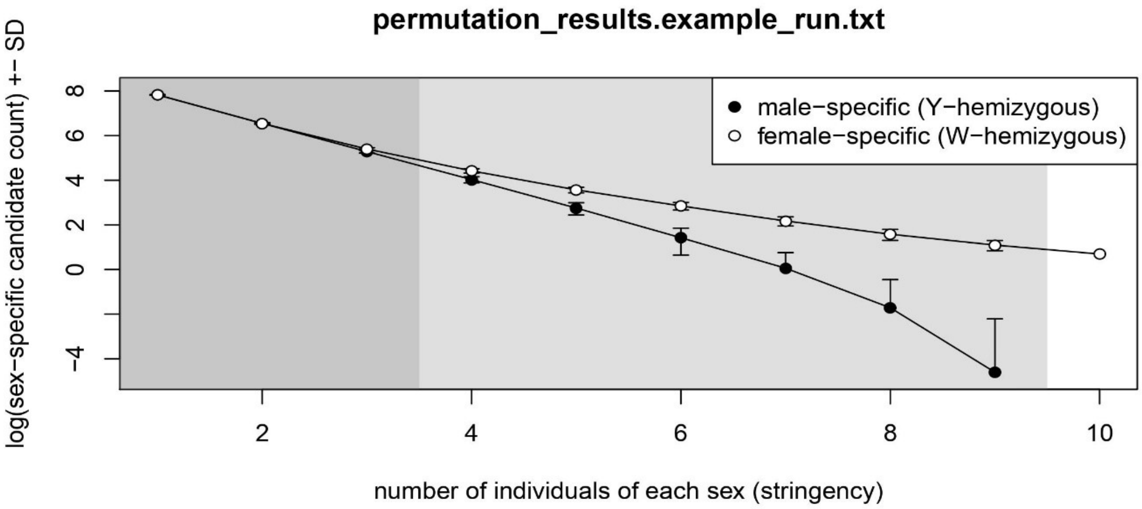
Context

This stage of analysis was intended to identify any hemizygotic loci associated with heterostyly – that is, loci only present in one stylar morph and completely absent from the other. Scharmann *et al* (2017) developed a program to identify sex specific loci in *Nepenthes* pitcher plants, the privacy rarefaction algorithm.

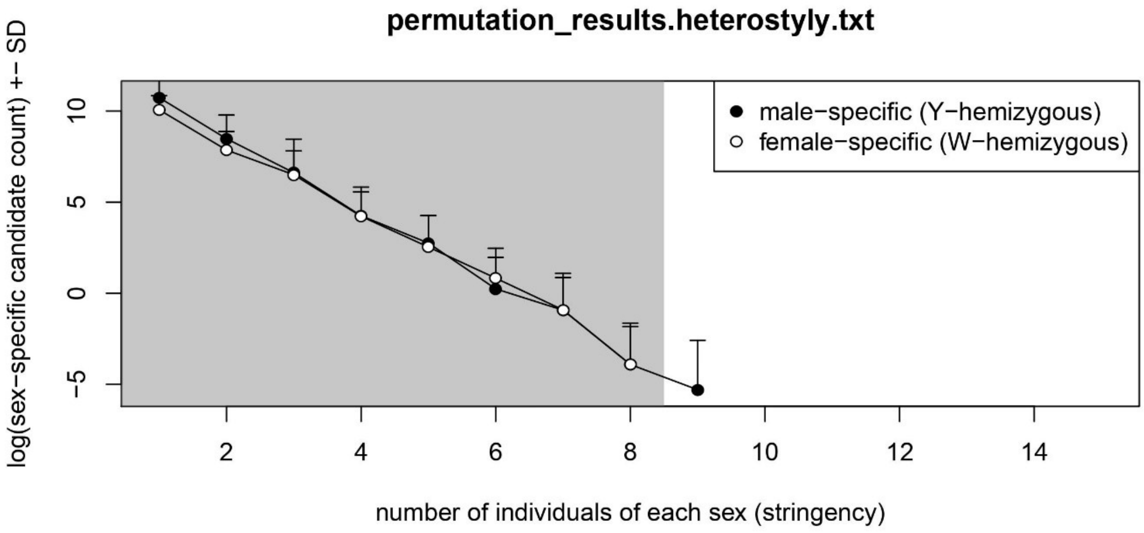
This program randomly sampled multiple subsets of the data, isolated the loci which seemed sex specific in each subset and compared them across subsets to determine the true specific loci. As this program was designed to isolate sex specific loci based on their hemizygotic inheritance patterns, it acted as a robust tool to isolate other hemizygous loci. For the purposes of this analysis, the thrum morph was defined as female while the pin morph was defined as male – this should not make a difference to the overall results, as the privacy rarefaction algorithm was designed to treat both sexes equally (as the gender with the sex determining chromosome in *Nepenthes* was unknown) and to assume that either could contain hemizygotic regions. The results are presented below (Figure 3.10), along with an example data set with a female hemizygotic region.

Content

a)



b)



c)

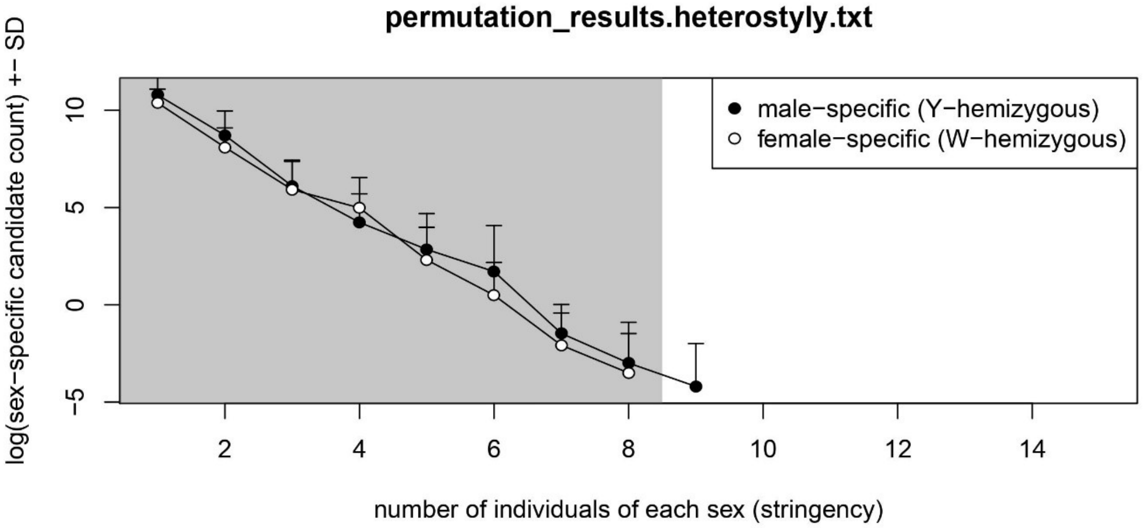


Figure 3.10, showing a) the number of loci specific to each sex found in increasingly large sample subsets of example privacy rarefaction data, b) the number of loci specific to each stylar morph (pin represented as male, thrum represented as female) found in increasingly large sample subsets of the sample data of the first mapping set (without population information) and c) the number of loci specific to each stylar morph (pin represented as male, thrum represented as female) found in increasingly large sample subsets of the sample data of the second mapping set (with population information). The differences between b) and c) are minor enough to be insignificant, with both generating the same result.

The first graph (Figure 3.10a) showed what the expected result of privacy rarefaction is for a data set with a hemizygotic region. Initially, when the subsamples tested consisted of only one individual of each phenotype, there were a large number of loci which seemed specific to one phenotype for both phenotypes. As the subsamples increased in size, while maintaining a 1:1 ratio in each subsample, many loci which seemed specific when only tested against one individual of the other morph were revealed not to be specific, and the number of specific loci detected decreased. Eventually, when the subsample encompassed the entire data set, there was only one phenotype with associated specific loci, and the only specific loci detected were those that were truly specific for the entire data set. In this example data set one of the loci was female specific, so as the subsample size increased the number of female specific loci began to plateau, while the number of male specific loci steadily decreaseds, even into fractions. Eventually all possible male specific loci were eliminated from consideration, while the female data set narrowed to the one true female specific locus. The difference in the error bars between the female and male specific loci count was also notable – the female error bar gradually narrowed as a single locus was identified, while the male error bar increased as the number of male specific loci decreases to zero, to account for the possibility that one of those loci which seem specific in a fraction of the population are truly specific for the entire population. With all this in mind, the graphs of the morph specific loci were considered, and it was evident that no morph specific loci could be identified. The numbers of both pin (male) and thrum (female) specific loci followed a similar pattern to those of the male specific loci from the example data, declining at a reasonably steady rate until decreasing to zero before the subset size encompassed the entire data set. Similarly, the numbers of both pin and thrum specific loci displayed large error bars, and they decreased into fractions of loci – both indicative of a lack of true specific loci. The pin specific loci were slightly more robust than the thrum specific loci, with at least one locus seeming pin specific in at least some subsets even when the subset size increased to nine individuals of each morph, whereas seemingly no possible thrum specific loci were found in subsets larger than eight individuals of each morph. However, no specific loci for either morph were identified when the subset encompassed the entire population.

Conclusion

As such, this analysis showed that there are no morph specific loci in *Linum tenue*.

Chapter 3.6: Morph association analysis – PLINK analyses

Context

This final stage of analysis was initially intended solely to identify any heterozygous loci associated with heterostyly – that is, loci with one allele associated with one stylar morph and another allele associated with the other – but given that the privacy rarefaction algorithm did not identify any hemizygotic loci, this analysis was expanded to encompass a secondary, less specific, search for hemizygous loci. These analyses were done with PLINK, a robust statistical toolset specifically for analysing genetic data in a genome wide association study (Purcell *et al.*, 2007). The PLINK toolset provided a variety of methods for assessing allele association with phenotypes, yet only two key methods were used in this protocol. These were a basic chi squared association test (with replicates), and Fisher's exact test, both tests analyzing the relative frequency of alleles in each stylar morph and determining from this whether an allele was significantly associated with one stylar morph. All phenotype specific, or hemizygous, alleles should only occur in one phenotype, and so should also be identified as significantly associated with this phenotype. The list of loci produced by these tests was filtered down into the loci significantly associated with a stylar morph ($p \leq 0.05$) that were present in 80% of the population (for heterozygous loci) or present in 80% of the population of one stylar morph and absent from 80% of the population of the other (for hemizygous loci).

Content

This led to a small number of truly significantly associated loci being found – 4 heterozygous loci, and 1 hemizygous locus. These loci, and the haplotypes of each in each sample, are presented below.

Locus number (mapping set 1)	Locus number (mapping set 2)	Main	Absent	Alt. 1	Alt. 2	Unique	Total Thrum	Total Pin
N/A	4244	9 Thrum : 8 Pin	3 Thrum : 2 Pin	1 Thrum : 5 Pin	N/A	2 Thrum : 0 Pin	9 M, 3 Ab, 3 Alt, 2 U	8 M, 2 Ab, 5 Alt

5812	19158	13 Thrum :11 Pin	1 Thrum : 2 Pin	N/A	N/A	1 Thrum : 2 Pin	13 M, 1 Ab, 1 U	11 M, 2 Ab, 2 U
N/A	121788	8 Thrum : 4 Pin	2 Thrum : 3 Pin	2 Thrum : 1 Pin	0 Thrum : 2 Pin	3 Thrum : 5 Pin	8 M, 2 Ab, 2 Alt1, 3 U	4 M, 3 Ab, 1 Alt1, 2 Alt2, 5 U
306179	359009	13 Thrum : 8 Pin	1 Thrum : 4 Pin	N/A	N/A	1 Thrum : 3 Pin	13 M, 1 Ab, 1 U	8 M, 4 Ab, 3 U

Table 3.6, showing the loci present in at least 80% of the total samples in each mapping data set with significantly different expression (P<0.05) in pin and thrum morphs (identical loci between data sets were only shown if both are significantly differently expressed between pin and thrum morphs). Samples were subdivided into those demonstrating, for each locus; the "Main" (M) haplotype (the haplotype of this locus present in most samples), the "Alternative" (AltN) haplotypes (the haplotypes present in multiple samples but in fewer samples than the Main haplotype), the "Unique" (U) haplotypes (the haplotypes only present in a single sample) and the "Absent" (Ab) haplotype (where no genetic sequence was found corresponding to this locus in this sample).

Locus number (mapping set 1)	Locus number (mapping set 2)	Main	Absent	Alt. 1	Unique	Total Thrum	Total Pin
304804	N/A	N/A	3 Thrum : 12 Pin	N/A	12 Thrum : 3 Pin	3 Ab, 12 U	12 Ab, 3 U

Table 3.7, showing the loci present in at least 80% of the samples of one morph and absent from at least 80% of the samples of the other morph in each mapping data set with significantly different expression (P<0.05) in pin and thrum morphs (identical loci between data sets are only shown if both are significantly differently expressed between pin and thrum morphs). Samples were subdivided into those demonstrating, for this specific locus; the "Main" (M) haplotype (the haplotype of this locus present in most samples), the "Alternative" (AltN) haplotypes (the haplotypes present in multiple samples but in fewer samples than the Main haplotype), the "Unique" (U) haplotypes (the haplotypes only present in a single sample) and the "Absent" (Ab) haplotype (where no genetic sequence was found corresponding to this locus in this sample).

All heterozygous loci were found in both data sets, but only the loci 5812/19158 (5812 in the first data set, 19158 in the second) and 306179/359009 (306179 in the first data set, 359009 in the second) were identified as significant in both data sets. The loci 4244 and 121788 in the second data set were also present in the first data set (as 4243 and 164265 respectively), but were not found to be significantly differently expressed in the first data set, caused by minor differences between haplotypes. This implied that providing extra information on the sample data did change how the samples are mapped, which is not necessarily something which can be assumed otherwise. However, these minor differences between haplotypes either did not appear in or did not affect the loci 5812/19158 and 306179/359009, as they were seen as significantly differently expressed between morphs in both data sets. As is visible above, however, the loci 4244 and 121788 contained a relatively lower proportion of samples with a “main” haplotype at these loci, and a higher proportion of “alternative” and “unique” haplotypes, explaining why minor changes in haplotypes between mapping sets had a greater effect on the association significance of these samples. This all implied that the difference between the data sets was minor and not universal across the entire set of samples. As such, going forward all 4 heterozygous loci were considered for analysis.

The hemizygous locus 304804 was another matter, though. This hemizygous locus was only present in the first data set (not considering population), and seemed completely absent from the second data set. It may be that the sequence of this locus was only slightly different in the second data set, with one or two differently sequenced bases of the 143 base sequence of the locus being sufficient to prevent matching of the loci sequences between the two data sets, as a direct search in the FastQ files was done to identify matching loci sequences with no allowance for mismatches. This would be consistent with the minor differences seen between data sets with the heterozygous loci, but this is not certain.

Conclusion

Still, the impact of these problems on analysis was likely to be minimal, so analysis proceeded using all four heterozygous loci and the potential hemizygous locus.

Chapter 3.7: Morph association analysis – BLAST analysis

Context

Assuming that these loci detected represent potential S loci candidates, or S locus associated genes, the next step was to identify potential functions of these loci. To do this, the sequences of these loci were analysed for potential open reading frames using NCBI's Open Reading Frame finder tool, and any open reading frames found were compared to known protein sequences using NCBI's Blastp, with the aim of finding notable homology between these sequences and those of known function. Every locus tested was found to have several potential open reading frames, however several of these open reading frames did not have any significant homology with known protein sequences. The open reading frames for each locus which did display any homology with known protein sequences are shown below (Table 3.8), followed by the open reading frames with significant homology (Table 3.9).

Content

a)

Locus Number	3' - 5' ORFs	Homologous	3' - 5' Homologous proteins
4243/4244	1	1	N/A
5812/19158	0	0	N/A
121788/164265	0	0	N/A
304804	1	0	N/A

306179/359009	2	2	50S ribosomal protein L2 (Desulforudis audaxviator) (E=2.4), Sec-independent protein translocase protein TatA (Shewanella loihica) (E=2.2)

b)

Locus Number	5' - 3' ORFs	Homologous	5' - 3' Homologous proteins
4243/4244	2	1	Bifunctional uridylyltransferase/nitrogen sensor protein (Haemophilus influenza) (E=0.16)
5812/19158	1	0	N/A
121788/164265	2	1	Several cysteine proteases, notably senescence-specific cysteine protease SAG39 (Oryza sativa) (E=1e-10)
304804	0	0	N/A
306179/359009	1	1	Valine-tRNA ligase/Valine-tRNA synthetase/ValRS (Haloarcula marismortui) (E=0.073)

Table 3.8, showing a) the number of 3’-5’ open reading frames for each locus and any known proteins these ORFs have any homology to and b) a) the number of 5’-3’ open reading frames for each locus and any known proteins these ORFs have any homology to.

Locus Number	Significantly homologous proteins (E<0.1)
121788/164265	Several cysteine proteases, notably senescence-specific cysteine protease SAG39 (Oryza sativa) (E=1e-10)
306179/359009	Valine-tRNA ligase/Valine-tRNA synthetase/ValRS (Haloarcula marismortui) (E=0.073)

Table 3.9, showing the loci with significantly homologous proteins (defined as E<0.1) and which proteins these loci are significantly homologous to.

The first thing to note is that even considering all possible homologous proteins, there were few candidates to consider for further analysis, with the hemizygous locus 304804 and the heterozygous loci 5812/19158 containing no open reading frames with homology to any previously identified proteins. This may have been a consequence of the short sequence input (each fragment entered was only between 120 and 150 base pairs), or it may have indicated that the loci identified did not produce a protein with a functional role in heterostyly. Given the expression difference between morphs, these loci could still act as molecular markers for S locus associated genes, but for this analysis of functions of potential S-locus associated genes they must be disregarded. The second thing to note is that even just considering the open reading frames with homology to a known protein, several of these open reading frames were only tenuously homologous. 4243/4244 contained a 5’-3’ open reading frame which was homologous to a uridylyltransferase, but with an E value below the significance threshold of 0.1 at 0.16. The two 3’-5’ open reading frames in 306179/359009 were even more tenuous, with a homology to a protein translocase with E=2.2 and to a 50s ribosomal protein with E=2.4. The borderline significance of the uridylyltransferase homology could arguably be attributed to the short input sequences, but these two proteins are far too likely to arise through random chance. Nevertheless, the fact that there is any homology may indicate that this locus is part of a gene encoding a protein with a similar function to these proteins..

With all these qualifications noted, it must be noted that there were still two loci (with significantly different expression between morphs) with one open reading frame encoding a protein with significant homology to a previously characterized protein. 121788 contained one 5’-3’ open reading frame, encoding a protein with significant homology to a number of cysteine proteases. The most significant homology was found with SAG39, a cysteine protease expressed specifically in senescent tissue in *Oryza sativa*. Similarly 359009 was found to contain one 5’-3’ open reading frame encoding a protein with significant homology to a valine-tRNA synthetase/ligase found in *Haloarcula* *marismortui* (a halophilic archaeon found in the Dead Sea). The homology for SAG39 in 121788 was notably stronger than the homology for ValRS in 359009, with an E value orders of magnitude smaller (1e-10 compared to 0.073). Similarly SAG39 presented a more obvious candidate for heterostyly, with its role in senescence presenting a possibility for a G or A candidate gene (limiting style or anther growth through inducing premature senescence), while no immediately obvious role was found for a valine-tRNA synthetase/ligase in heterostyly. Nevertheless, the homology for ValRS wass still significant, and there were significant expression differences in 359009 between morphs.

Conclusion

The final conclusion of the analysis was that, while there were several possible genetic markers identified which could help isolate the S locus, only two heterozygous loci were found which could correspond to S locus linked genes which play a role in heterostyly.

Section 4: Discussion

Chapter 4.1: PLINK: Heterozygous Candidates

At the end of the analysis, 6 loci (across two separate mapping sets) were found to have significant expression differences between thrum and pin samples (with different haplotypes expressed in thrum and pin samples) and to be present in at least 80% (24 of the 30) of the samples tested, meaning that they were possible heterozygous S-locus associated genes. Given that there was no allele of any of these loci only expressed in a single morph, it was unlikely that these were actual S-locus candidates, but it was possible that these were linked to genes associated with heterostyly regardless. Of the 6 potential loci, it was found that two of these loci contained the same sequences of two other loci in the 6, being matches from across the two different mapping sets and so reducing the total number of potential heterozygous loci to 4. Open reading frames were identified in the sequences of these loci, and it was found that of these 4 loci, only two had at least one open reading frame encoding a protein with significant homology to a known protein. These ORFs for loci 121788 and 359009 encoded, respectively, a cysteine protease involved in senescence in *Oryza sativa* and a valine-tRNA ligase from *Haloarcula marismortui*.

The first of the two candidates to be considered was 359009, which contained a 5’-3’ open reading frame encoding a homolog to a valine-tRNA ligase/synthetase found in *Haloarcula marismortui*. The connection to a halophilic archaeon found in the Dead Sea seemed very coincidental, as any genetic relationship which could be conceived between an archaea and a plant would be incredibly tenuous. But putting that aside and considering 359009 only as a valine-tRNA ligase/synthetase, a role for this in heterostyly could be conceived. Valine-tRNA ligases are a family of ligases which catalyse the ligation of the amino acid valine on to the corresponding tRNA codon. While it seemed unlikely that the proteins involved in creating one heterostylous morph involved a large enough amount of valine to necessitate a valine-tRNA ligase localized to a single morph, this is not the only role that has been found for these enzymes. Valine-tRNA synthetases have often been found to act as “biological proofreaders”, hydrolysing specific aminoacyl-tRNA complexes to prevent incorrectly translated amino acids from being incorporated into the final protein (Peters, Haar and Cramer, 1990). Indeed, ValRS has been found to have a role in discriminating between valine and similar amino acids isoleucine and threonine in a two stage mechanism discriminating first based on size and second based on hydrophilicity (Fersht and Kaethner, 1976, Fukai *et al.* 2000). As such, it may be possible that structural differences in a *Linum tenue* ValRS may allow it to act in a post-transcriptional regulatory manner, allowing certain proteins to be synthesized only in a certain stylar morph, but this was extremely speculative. It seemed most likely that this enzyme encoded by this heterozygous locus was not associated with heterostyly.

The second and final candidate was from 121788, which contained a 5'-3' open reading frame encoding a protein with significant homology to a large number of cysteine proteases. While there was no evidence of cysteine proteases specifically being involved in heterostyly, proteases more generally have been shown previously to have some role in heterostyly. Limiting things to *Linum,* a study of *Linum grandiflorum* (Ushijima *et al.* 2011) found that one of the differentially expressed transcripts between morphs belonged to an aspartyl protease, so this finding is not unprecedented. Proteases have been shown to play some role in self incompatibility, with addition of protease inhibitors to *Fagopyrum esculentum* preventing pollen tube rejection due to self incompatibility (Ushijima *et al.* 2011). Current theory suggested that proteases were secreted extracellularly in the stigma to target and disrupt pollen tubes from an incompatible morph. This was supported by the previously mentioned *Linum grandiflorum* aspartyl protease, which was localized by PSORT (based on its amino acid sequence) to be extracellular (later confirmed experimentally). However, the PSORT prediction for the cysteine protease identified in this study placed it as a membrane protein, most likely localized to the plasma membrane but with its most closely related neighbours being localized to chloroplasts. This meant that this cysteine protease probably did not follow the established pattern for the role of proteases in heterostyly, and may therefore not be involved in self incompatibility. One other notable feature was that the most closely related homolog of this cysteine protease was *SAG39*, a senescence specific cysteine protease found in *Oryza sativa*. Senescence could have a role in limiting growth of floral organs, so this was another possible theory for this protein's role in heterostyly. One final thing to note was that *SAG39* was predicted to respond to gibberellin, jasmonic acid, ethylene and abscisic acid (Liu *et al.* 2010). This could tie this protein to the study in *Arabidopsis* which found that abscisic acid was tied to a heterostyly like phenotype (Suzuki *et al.* 2014), or more closely to *Linum grandiflorum* which found that one of its thrum specific proteins (*LgMYB21)* had homology to other proteins which respond to gibberellin and jasmonic acid (*AtMYB21* & *24*). However, this was again tenuous. It is not a rare feature for a plant enzyme to respond to a common hormone.

However, all of this analysis so far assumed that these loci have truly different alleles between stylar morphs, and that the open reading frames constructed from combining these loci with neighbouring loci were accurate. There was some evidence that this may not be the case. Looking into the exact haplotypes for the significantly different loci showed that the difference between stylar morphs may not have been a result of actual genetic differences, but rather of sequencing quality differences between morphs. For instance, the two main “different” haplotypes found for the first locus, 4244, were ACN/CAN and ACG/CAG. While it was understandable why this was registered as a different haplotype by PLINK, it seemed more likely that this represented the same haplotype but with one better sequenced base in the latter. A similar pattern was displayed across all the loci sequenced (full table in appendix), though some of the loci had more significant haplotype differences than others (notably 121788, with only two of the samples out of 30 differing from the main haplotype purely because of Ns). Similarly, it was somewhat concerning that for most loci (again, with the notable exception of 121788) that there was a “main” haplotype which the majority of samples from both styles follow – that said, they were still identified as significantly different by PLINK’s tests, so it must be assumed that this was not an issue. However, based on this only one of these potential heterozygous loci could be reasonably said to have true expression differences between morphs based on this data – 121788. All things considered, what seemed most likely was that PLINK had identified several loci with no expression differences between morphs, one locus which was a part of a gene present in the genome of *Linum tenue* but is not necessarily related to heterostyly –the valine-tRNA ligase from 359009 – and one gene which was quite possibly involved in heterostyly – the cysteine protease from 121788. Consequently, further studies clarifying the genetic location and function of these genes could be a productive avenue for further research. Proteomic based studies could confirm the relevance and function of these genes by identifying their protein products, localising these proteins within the plant and identifying any notable structural differences in the proteins (compared to the identified homologs) which could explain their role in heterostyly. Differential expression of these proteins between stylar morphs would validate their role in heterostyly and the results of this study, if it could be confirmed. All of the identified loci, regardless of protein homology, could also act as genetic markers for future targeted sequencing. Given that these loci seemed to be associated with stylar morphs, sequencing of greater depth surrounding these loci could possibly reveal the S locus region, allowing characterization of the genes controlling heterostyly in *Linum tenue*. But given the problems with these loci previously established, any further sequencing study would have to first confirm that these loci are truly different between stylar morphs, possibly through RNA analysis aiming to identify transcription differences between morphs. Therefore while the results of this analysis were inconclusive, they still suggested potential avenues for future research which could identify the components of genetic control of heterostyly in *Linum tenue*.

Chapter 4.2: Privacy Rarefaction: Hemizygous Candidates

Alongside PLINK, a second analysis was run to identify potential hemizygous candidates involved in heterostyly. Recent advancements in identifying the genes involved in heterostyly in *Primula vulgaris* and *Fagopyrum esculentum* determined that the supergene complex which controls heterostyly, the S-locus, was only present in one of the two stylar morphs. The different stylar morphs were not the result of different S-locus alleles but rather the presence or absence of the S-locus itself, and so heterostyly in these species was not heterozygous, but hemizygous. This likely meant that heterostyly across species was hemizygous rather than heterozygous, and so identifying any genes which were only present in one of the two stylar morphs wass of utmost importance for any subsequent study on the genetics of heterostyly in any species. As such, an algorithm which was originally designed to identify hemizygous loci associated with sex determination was repurposed to identify hemizygous loci associated with heterostyly. This algorithm, called privacy rarefaction (Scharmann *et al.* 2017), aimed to compare a set of samples with a 1:1 ratio of two defined phenotypes, by randomly generating subsets of the sample set containing a 1:1 ration of the phenotypes and identifying any loci which were only present in all samples of one of the two phenotypes. These subsets gradually increased in size until they encompassed the entire sample set, and any loci which were found in all the samples of one phenotype and none of the other in all possible subsets are determined to be truly hemizygous. By setting the pin phenotype as male and the thrum phenotype as female (an arbitrary choice), this privacy rarefaction algorithm seemed ideal for identifying hemizygous loci associated with heterostyly. Running this algorithm on the samples generated for this study generated several possible hemizygous loci for smaller subsets, as expected for any sample set, but by the time the subset had expanded to cover the entire sample set no true hemizygous loci remained. Every possible locus was present in at least one sample of both stylar morphs. As such, this seemed to confirm that there are no hemizygous loci associated with heterostyly present in this set of *Linum tenue* samples. However, there were some issues with this conclusion.

The designers of the privacy rarefaction algorithm identified that the primary issue with identifying sex specific loci used to be one of false positives. The large sample sets often used to identify truly specific loci necessitated the use of subsets in the final analysis, as comparing every individual of one phenotype against every individual of the other phenotype was both computationally intensive, and could result in truly hemizygous loci which were not sequenced in some individuals of one morph being discarded. But there was a significant risk in using subsets to identify hemizygous loci, as loci which may have seemed hemizygous in a small subset of samples could be false positives, heterozygous in the larger set of samples. The privacy rarefaction algorithm was designed to avoid this issue, by identifying possible hemizygous loci in small subsets then expanding to larger subsets to confirm that they were truly hemizygous. This avoided the issue of heterozygous loci being identified as hemizygous in a small subset, and of hemizygous loci being discarded in a large subset. However, this increasing subset method meant that the wider picture could not be used to identify hemizygous loci. If the entire sample set were used to test for hemizygous loci, and of a hundred samples of each morph one locus was present in a hundred samples of one morph and one sample of the other morph, this single sample of the other morph containing this locus could be identified as a likely outlier as a result of sequencing error, and this locus would be identified as hemizygous. However, with privacy rarefaction it was possible that this single outlier sample would be randomly selected for a smaller subset comparison, comparing this sample against one other from the other morph. As the locus would be present in both samples, it would be discarded as a potential hemizygous candidate. As such, although the privacy rarefaction algorithm addressed the issue of false positives in identifying hemizygous loci, it introduced a significant risk of false negatives, especially in data sets with some amount of sequencing error. And it seemed likely that this sample set did contain some amount of sequencing error. The amount of uncalled bases, touched upon in the PLINK analysis, indicated some issues, and the overall low coverage of the mapping was another cause for concern. It had been suggested that high coverage is the most important factor in reducing sequencing error, with ideal coverage for most sample sets being greater than 25x. Any coverage below 10x would be at serious risk of errors, and a full 40% of this sample set (12/30) showed a mapping coverage less than 10x. This could be seen in the PLINK analysis, where a standard test on missing loci found that the average genotyping rate (with a genotyping rate of 1 being where every locus sequenced was present in every sample) was as low as 0.21. While no individual was discarded, several individuals had greater than 90% of total loci identified missing. It could not be confirmed whether these individuals were poorly sequenced or whether the abnormally high coverage of some other individuals had ensured that most loci identified were only found in these high coverage samples, so no individuals could definitively be identified as outliers and discarded. There was no bias between pin and thrum morphs as to which had more missing loci, but nevertheless the impact of this was seen in the association test when many loci were identified as significantly associated with one morph because they were only present in a few samples of one morph. It seems likely therefore that there wass a substantial amount of sequencing error present in this sample set, and so privacy rarefaction could be prone to false negatives. As such, there may still be hemizygous loci present in this set of *Linum tenue* samples. To identify them, a less specific analysis of the entire data set was performed, using PLINK.

Chapter 4.3: PLINK: Hemizygous Candidates

As well as the S-locus associated candidate genes already discussed, PLINK revealed one more gene with significant expression differences between morphs that was present in a majority of samples. However, unlike the heterozygous candidates discussed previously, this gene was not present in at least 80% of the total samples tested. Rather, it was present in at least 80% of the samples of one morph, and in less than 20% of the samples of the other morph, making it a potential hemizygous gene associated with heterostyly – and thus, a potential S locus candidate gene.

Locus 304804 was only found in the first de novo map, when only the morph information about the samples was provided to STACKs and no information about sample population was used, with no matching sequence being found in the second de novo map which factored in the population the sample originates from. PLINK found that this locus was significantly differently expressed between stylar morphs, and when the haplotypes for each significant locus were analysed it was found that 304804 was potentially hemizygous, as it was present in most thrum samples and few pin samples. To be specific, 304804 was present in 12 of the 15 thrum samples, not including BUR_25, LUM_15 and LUM_25, and was absent from all but 3 of the 15 pin samples, being present in CBT_18, CBT_20 and PIG_35. Every sample with this locus contained a unique haplotype, there was no "main" haplotype for this locus shared among the majority of samples. However, as previously mentioned with heterozygous candidates, many of these haplotypes differed purely through the number of miscalled bases, making it difficult to determine which haplotypes were truly different. Further analysis comparing the genetic sequence of the locus unfortunately identified no open reading frames within this locus encoding proteins with significant homology to any known proteins, though this can likely be attributed to the short sequence of the locus. Considering all this information, how likely is it that this locus represented a possible S locus candidate gene in *Linum tenue*?

There were several pieces of information which supported this being an S locus candidate. Common consensus of the conditions for an S locus gene was that it must be completely co-segregated with one morph, showing no expression in the other, and a hemizygous locus would fulfil this requirement. Indeed, hemizygosity has increasingly been seen as a requirement for an S locus candidate, after the sequencing of the *Primula* S locus in Li *et al* (2016), so a morph specific hemizygous locus would be the ideal candidate. The presence of the locus in some pin morphs and the privacy rarefaction algorithm not identifying it as hemizygous were potential issues, calling into question its morph specificity. However, several of the pin morphs which showed presence of this locus had abnormally high coverage. CBT_20 and CBT_18 were the samples with the highest and third highest coverage respectively, with coverage of 95x and 80x while most samples were closer to

10x. As all haplotypes of this locus were unique, it could not be determined whether the haplotypes expressed in these samples were notably different from those of the thrum samples. The fact that these samples were such notable outliers could indicate issues with the mapping, or more likely with the rad tag identification. One arguable disadvantage of using two rad tag barcodes to identify a sample was that if one rad tag was not properly sequenced and was not recognised properly, it was possible for a fragment to be incorrectly filed as belonging to another sample which had the same barcode as the intact barcode of the fragment. In a single tag system, it seems more likely that a corrupted barcode would lead to a sample being filed as unidentifiable, as there would be no intact misleading barcode – posing problems for creating a complete data set for mapping, but avoiding any problems with misleading sample data. The abnormally high coverage of these two pin morph samples could imply that these were the samples many difficult to identify fragments were sorted into. If some of these fragments originally belonged to a thrum morph sample, it would explain the presence of locus 304804 in these samples while preserving it as a thrum morph specific hemizygous locus, and so an S locus candidate gene. Similarly, two of the thrum samples which were missing this locus (BUR_25 and LUM_15) showed extremely high rates of missing loci more generally. Of the total loci identified, 99% were missing in both of these samples. While concerning more generally, it does imply that the absence of this locus in these samples could be more due to sequencing issues than it not being a truly thrum specific locus. All of this would explain why the privacy rarefaction algorithm did not identify it as hemizygous – as previously mentioned, the privacy rarefaction algorithm was tuned for specificity over sensitivity, and so was prone to false negatives if the data set was not completely clear. A thrum specific locus being misidentified as being present in a pin morph sample would certainly be sufficient cause for doubt for the privacy rarefaction algorithm to discard it as a potential hemizygous candidate.

However, that assessment would be optimistic. There were several pieces of evidence which worked against 304804 being considered a likely S locus candidate gene. The main one being one that was previously mentioned, that evidence of its hemizygosity was relatively weak. Within the relatively lax restrictions of 80% presence or absence necessitated by a small sample set and incomplete sequencing, this locus did barely appear hemizygous. But if the sample set had been expanded and the sequencing had been more successful, stricter confidence thresholds could have been imposed. It seems most likely that this locus would then maintain the presence/absence pattern displayed in this sample set – present primarily but not exclusively in thrum samples – and so would not be classed as hemizygous. As the majority of recent research into S loci across species emphasised the need for hemizygosity, this would have made it difficult to class as an S locus. There were some arguments against this, as discussed previously, the main one being that the abnormally high coverage of the pin samples containing the locus suggested that the locus was misfiled into these samples. However, the high coverage of CBT_18 and 20 could also be explained by the mismatches in DNA concentration of the samples, one unfortunate side effect of using pre-extracted samples being that this was difficult to correct. Several samples had significantly higher DNA concentrations going into sequencing, and so despite efforts to even out the amount of DNA in each sample it seems reasonable to assume that the disparity in coverage could be partially attributed to this. And indeed, CBT_18 and CBT_20 did show relatively high DNA concentrations pre sequencing. If the high coverage in these samples was not the result of misidentification, the evidence would be particularly damning – higher mapping coverage has been closely tied to more accurate sequencing, so these two pin samples were likely to be the best sequenced samples in the data set. If locus 304804 was present in these, this all but proved that it was not a thrum specific locus. Even assuming that DNA amounts in each sample were successfully regulated and so the significance of 304804 being found in CBT_18 and 20 was reduced, locus 304804 was not just present in these two outlier samples. It was also found in the pin sample PIG_35, the coverage of which was in line with the other samples at 10.66. While this could also have been due to a barcode assignment error, there was no evidence in support of this. It seems more likely that 304804 was present in at least one, likely three and possibly more pin samples, and so was likely not hemizygous. A similar thing could be said for the thrum samples which miss 304804 – while BUR_25 and LUM_15 displayed an abnormally high percentage of missing loci, explaining its absence, LUM_25 was only missing 78% of the total loci. A concerning number in isolation, to be sure, but in the context of the samples as a whole it was in line with the average. The average genotyping rate across all samples was 0.21, after all, implying that a sample missing 78% of the total loci was, if anything, slightly above average. This all implied that the samples which did not fit into the ideal hemizygous pattern for the locus 304804 were not necessarily outliers, and could be closer to the true pattern of expression. The fact that the locus was missing from 12 of the 15 pin samples seemed to support the idea of hemizygosity, but it must be made clear that loci were missing from a large portion of samples. Again, on average each sample was missing 79% of the total loci. With that in mind, it seems possible that a locus like 304804 could have arisen purely by chance – unlikely, to be sure, but in a set of ~600’000 loci where the majority were missing from any individual sample, the chance that one locus is randomly found missing in 80% of pin samples and present in 80% of thrum samples would be too large to entirely discount. Ultimately, concluding that 304804 was a hemizygous S-locus candidate or that 304804 was a heterozygous locus not associated with heterostyly would require ignoring a significant amount of error. As such, due to the prevalence of sequencing error, it could not be concluded that this locus was part of a hemizygous S-locus region.

Chapter 4.4: Potential Sources of Error

As has become very clear, this sample data set contained a large amount of error, enough that no solid conclusion can likely be derived from this data. This error could be seen in a few different ways, each with their own impact on the final data. The prevalence of uncalled bases was the first factor noticed, which had an impact on the PLINK analysis for heterozygous loci and caused several false positives. This can be put down primarily to sequencing issues, with Illumina sequencing having recognised the presence of a base, despite being unable to identify the exact fluorescence of any single nucleotide. This must also shed some doubt on the bases wjocj were able to be called – if this many nucleotides could not be accurately identified, it is possible that the sample DNA was difficult enough to sequence that some of the bases called may be inaccurate. The second factor noticed was the extreme disparity between coverage of samples, with the disparity in the amount of missing loci per sample a possible consequence of this. The impact of this could be seen in the analysis for hemizygous loci, with the possibility raised that misfiled sequence reads caused truly hemizygous loci to be unfairly discarded by the privacy rarefaction algorithm. The cause of this was a little more nebulous, with the quality of the pre-sequencing samples, the post sequencing identification of samples and the mapping protocols all having had a possible impact. If the main negative impact was that of sequence fragments being incorrectly assigned to samples leading to truly hemizygous loci, then the area of focus for this error should be that of post sequencing radtag processing, to determine how much of an impact this had. The third factor noted was the generally low coverage of the samples, and as a possible consequence the low average genotyping rate across samples. The impact of this was noted in the PLINK analysis for hemizygous loci, where the high proportion of loci missing from any given sample raised the question of whether a sample which was missing in the majority of pin samples and present in the majority of thrum could have risen through chance, and so posed doubts as to whether it could be considered truly hemizygous. Again, this was an issue where possible sources of error could have arisen at every stage of the analysis, from pre-sequencing sample preparation to final mapping. But as the selection of mapping parameters was intended to prevent coverage issues, the mapping stage should be considered to see what impact it had on alleviating coverage issues, or even if it could have exacerbated them. Other errors were noted in the results section, such as the initial quality of the data and issues with the Tape Station analysis of DNA size, and these will be considered alongside these noted errors. But these three error factors noted were the ones with the largest identifiable impact on final analysis, so they were focused on. This meant that there were three primary error factors to focus on, and three possible stages for these errors to arise – in sequencing (including preparing the samples for sequencing), in post sequencing processing and data cleaning, and in mapping the data into loci.

Chapter 4.4a: Sources Of Error: Sequencing

The first stage considered for possible sources of error was sequencing, including the samples used, the methods used to prepare the samples for sequencing and the sequencing methods themselves. Error introduced at this stage would have underpinned the entirety of the later study, and would therefore have been the most difficult to extract from the later data set (it may have been possible to solve errors in later stages by rerunning the programs with altered parameters, but sequencing error could likely only be removed by redoing the sample preparation and sequencing). As such, errors in this section should be the most important for informing how to improve future studies. The first area of this considered was the sequencing itself, but this was also the area with the least to discuss. Several projects were sequenced in the same sequencing run as this project, with one other sequenced in the same lane (samples differentiated through barcodes). No excessive sequencing problems were reported from any other project, including the project in the same lane, so while it is theoretically possible that issues with sequencing were isolated solely to this sample set, it is unlikely. It seems most probable that issues during sequencing were not a major contributor to overall error rate in this project.

The second area considered was the methods used to prepare samples for sequencing, and there was some reason to consider that this would contribute to the error rate. The first run of the protocol resulted in no DNA being detected in the right size range for sequencing, indicating significant losses caused by sample preparation. When the method was redone with a less stringent size selection protocol from post ligation samples obtained from the first run through, enough DNA was detected in the right size range for sequencing to be performed. This indicated that the steps of the method before ligation did not result in a catastrophic loss of DNA, supported by measures of DNA concentration taken from samples before and after size selection (which changed from 12.2ng/µl to 1.5ng/µl) however, the exact amount and size of DNA obtained from the second run through was hard to quantify. The peaks from the Tapestation analysis, intended to identify the quantity of DNA of any particular size, were shallower and broader than expected, which indicated that rather than a large quantity of DNA in a small size range (the intended result) this quantity of DNA was spread out across a broader size range. It is possible that a smaller quantity of DNA than expected was obtained from the protocol, but this was hard to determine given the larger range the DNA was spread across. There was clearly an issue with this Tapestation analysis, that said – the expected peaks from the library reference sample were not properly called, being far shallower than expected and at incorrect positions. This explained why the peak for the sample library was identified as being far larger than expected (737 base pairs when the size selection was centred closer to 400 base pairs), especially as this size was not supported in post sequencing analysis of obtained fragments, as well as possibly the diffuse peaks obtained. However, the presence of the diffuse peaks was also supported by the parallel gel electrophoresis run on these samples, which showed rather than a single dark line (indicative of a large DNA quantity within a small size range) a lighter grey smear (indicative of DNA spread across a larger size range). The diffuseness of these peaks could be attributed to both the less stringent size selection protocols implemented in the second run through of the method and the cycles of PCR run on the size selected samples, both intended to address the issue of not obtaining enough DNA in the first run through. This brings things back to the original issue though – why did the protocol not result in enough DNA being obtained to enable sequencing?

There were several stages in the sample preparation protocol which could have resulted in loss, or degradation, of DNA. Given that this project was performed using pre-extracted samples, losses associated with DNA extraction can be set aside. While DNA extraction could result in significant loss or degradation of DNA, if the centrifuged pellet was not properly extracted, if it was over dried or if the ethanol was not fully extracted, these samples were previously used for another successful sequencing project. As such, any errors introduced in the extraction period were likely minor enough to be ignored. There were three key areas of the remaining protocol which had the potential to result in significant DNA losses, and those were the restriction enzyme digestion, the size selection and the cleaning steps. The restriction enzyme digestion could have resulted in DNA losses if the enzymes were not deactivated promptly, or if they acted quicker than expected, when the enzymes could theoretically break down the target DNA fragments into smaller, unusable fragments which were later filtered out in size selection, resulting in the loss of usable DNA. However, restriction enzymes usually only digest DNA at specific restriction sites so, unless the number of restriction sites in the genome was much greater than predicted, significant DNA losses were unlikely. The second key area associated with DNA loss was the size selection protocol, and this did seem to have resulted in significant losses in the first run through of the protocol. If the size range selected for was too narrow, or if the predicted size of the DNA fragments differed greatly from the actual size, size selection could have filteedr out a large proportion of relevant, usable DNA alongside contaminants, and this did seem to be the case in the original run through. However, this may not have been the only cause – even when the size selection protocol was amended in the second run through, the samples still showed lower coverage than expected, indicating that less DNA (or at least less usable DNA) than expected had made it through to sequencing. As such, the last area associated with DNA loss and degradation must be considered, that of the cleaning steps. These steps involved the sample DNA binding to magnetic beads, allowing the rest of the sample (and therefore any contaminants) to be discarded before the DNA is released from the beads and re-suspended. If all of the DNA was not bound to the beads or, conversely, if all of the DNA was not released from the beads, there could have been large losses of DNA at this step. This was supported by experimental evidence – based on the results of previous projects which have shown similar problems with DNA loss, the protocol had been adjusted to remove one cleaning step after restriction enzyme digestion, leaving only cleaning steps after sample pooling and after PCR amplification. The removal of this cleaning step did solve this issue in previous projects, but it was possible that the remaining cleaning steps still caused an unacceptable loss of DNA in this project. However, this could not be confirmed, and even if it could, low coverage was not the only issue with this project. The prevalence of miscalled bases was also a serious issue, and this could be attributed to contaminants present in the sequencing sample, or to DNA degradation caused by contaminants in the process of sample preparation. If this was the case, removing cleaning steps would only exacerbate the issue, and should not be considered for future trials. All things considered, it seems as if the most significant loss of DNA in the protocol could be tied to the size selection protocol, with the adjustments made in the second run through more or less solving the issue even at the cost of having a wider range of DNA present in the final sample. As such, the question still remains as to why coverages were so low in the final data set, and the answer may lie with the third area of sequencing preparation to be considered - the samples themselves.

For this project, remaining samples from another sequencing project were used. This approach had some significant advantages, such as the time and resources saved, the guarantee that the DNA extraction procedure did not introduce excessive contamination and the proof that the DNA extracted was suitable for sequencing. However, this approach also had major disadvantages. The phenotypes used in this study were based on those assigned to these samples previously, so any errors where the stylar morph of a sample was misidentified before this study could not be corrected, casting some doubt on the morphs assigned to samples in this study. The samples had been kept frozen for a long period of time, which was unlikely to introduce problems but was still a factor to consider. The process of preparing these frozen samples for sequencing involved a large amount of thawing and refreezing – this could well have resulted in loss of sample material, or even damage to the material present. Several samples showed surprisingly low DNA concentrations, likely a result of frost melting and diluting the material. But the most serious disadvantage is just that these samples were very low quantity to begin with. Many samples with higher concentrations of DNA did not have enough liquid present to allow them to be prepared for sequencing, with dilution possible but likely to reduce DNA concentration to an unacceptable level. This necessitated the use of lower concentration samples for sequencing, alongside changes to plans for the ideal distribution of populations, and this may have meant too little DNA was present in the samples to begin with to allow sequencing. While all the samples were at least high enough concentration to be sequenced to begin with, a certain amount of DNA loss is an expected consequence of even the most carefully controlled sample preparation protocols. The samples used may not have contained enough DNA to account for this. Even if the samples had contained enough DNA, there was not a uniform DNA concentration across samples. There was a general disparity across the sample set, with some samples being much more concentrated than others and some samples contained much less material than others. Measures were taken to account for this in the methods, with different dilutions for different samples to try and ensure a common concentration, but with the quantities used this may not have been exact. In addition, the use of these low quantity frozen samples precluded the possibility of replicates. Some samples had barely enough material to be prepared and sequenced once, so repeating the experiment to confirm results would have been an impossibility. Initial plans were to also plant and grow seeds of populations corresponding to the ones sequenced, allowing for later DNA extraction and sequencing or proteomics to confirm the results of the first sequencing run. This would likely have addressed several issues which arose in this project and if this experiment were repeated in future, this step should be taken (resources permitting).

Chapter 4.4b: Sources Of Error: Post Sequencing Processing

The second stage considered for possible sources of error was the methods taken to process the fragments immediately post sequencing, including assigning the fragments to samples based on their radtag barcodes and filtering out low quality sequences from the analysis. Error introduced at this stage was theoretically correctable, but was difficult to identify and underpinned the rest of the analysis. All future analysis was based upon the assumption that the DNA fragments were correctly assigned to samples and that they were all of (reasonably) high quality. If this was not the case, it would have been hard to detect and would have fundamentally altered the final results. In addition, if there was any existing error in the sequenced samples, error at this stage would have worked in concert with that to cause even greater damage. Existing error would have been amplified if the cleaning stage, aimed to filter low quality sequences, was imperfect, and existing error would have made it difficult to fix any issues in assigning fragments to samples, as the barcodes could be fundamentally too degraded to process. The two stages of post sequencing processing were done in one step, but to make discussion clearer they were considered in two different sections – the radtag processing and the cleaning.

The first area to be considered was the radtag processing, and it was clear from the results that this was a tumultuous process. Concerns with the number of reads retained meant that the mismatches allowed between barcodes (the maximum number of different bases between the barcode listed and the barcode present in the fragment which were allowed to classify it as the same) was increased from one to two, and the check for a restriction enzyme digestion site was disabled. The check for a restriction enzyme digestion site could have been relevant for filtering DNA contaminants, as only DNA digested by restriction enzymes (i.e. the processed sample DNA) should have a digestion site. As such, disabling this check did risk introducing irrelevant sequences into the final analysis. However, care was taken to minimise the possibility of contamination, and this check also had the unfortunate side effect of filtering out sequences with sequencing issues around this area. This could be argued to be filtering out low quality sequences from the final analysis, but the subsequent cleaning step covered this regardless. Any fragments which were filtered out by the restriction enzyme site check but not by the cleaning step were either contaminants or, more likely, fragments with sequencing issues localised to the area of the restriction enzyme site and which would therefore still be useful in the final analysis. As such, disabling the restriction enzyme check should not have introduced a large amount of error into the final analysis, with the only minor risk being of introducing contaminants. Increasing the number of mismatches allowed in barcodes seems similar on the surface, with one additional mismatch allowed in a 6 base sequence unlikely to cause many problems. However, this may not have been the case. Analysis of the barcode sequences used in this project found no barcodes which differed from each other in only one base, but found two pairs of barcodes, one forward and one reverse, which differed from each other in only two bases (TGTTAC/AGTAAC and GTTACT/GTAACA). This meant that if a fragment barcode was called with two unknown bases in those positions (NGTNAC or GTNACN), it would be included in the final analysis, but it seems arbitrary which barcode it would be assigned as (likely whichever barcode was listed first). This was where the two barcode system could be relevant, as if one of these barcodes was incorrectly assigned the disparity in the second barcode should alert the algorithm to the issue and change the first barcode's designation, but there were still three samples in the sample set which were identified using two of these barcodes (ELB_30: AGTAAC/GTAACA, LUM_15: TGTTAC/GTTACT, MDA_24: AGTAAC/GTTACT). In addition, while LUM_15 and MDA_24 were both thrum, ELB_30 is pin. So the increase of allowed barcode mismatches from one to two may have resulted in several fragments being assigned incorrectly to samples, with a risk of some thrum fragments being assigned to pin samples and vice versa. The impact of this was unclear, but it did shed some doubt on the final results. The absence of the "thrum" haplotype of 304804 in LUM_15 was one of the reasons to doubt its hemizygosity, for example, and there could be many more unidentifiable similar errors in the data set. However, if these mismatches were not increased, there was a significant risk of several samples not having enough fragments assigned to allow further analysis. So while it seems likely that increasing these mismatches introduced some error into the final analysis, it was still the best decision to take given the data available. In future studies though, if it is not absolutely necessary given the data it seems an inadvisable step to take.

The second area to be considered was the cleaning, where sequences were filtered and altered to ensure only biologically relevant data proceeded to mapping and analysis. There were several steps to this, namely identifying and removing adapter sequences, checking sequence quality and removing low quality bases and cropping the fragments to a uniform size. Any cleaning step of this type has some inherent risks associated with it, such as being too stringent and discarding or altering potentially usable fragments, or being too lax and allowing low quality contaminant sequences to proceed to later analysis stages and interfere with the final results. However, the cleaning steps used in this analysis had been constricted to minimise this risk. The parameters used were based on parameters previously used for the cleaning step of another successful analysis, so it was reasonable to assume that any error introduced using these parameters at this stage would not have had a major impact on the final analysis. There was one major change in the cleaning step of this protocol though, as a different program was used. Rather than use Trimmomatic for cleaning and Stacks for identifying fragments, the decision was made to combine the two steps and have Stacks perform the cleaning step as well. There were significant advantages to this – the next few stages of the analysis pipeline all involved Stacks, so performing this early stage in Stacks as well removed any issues which may have arisen from adapting an output from one program into an input for another. The major disadvantage to this, though, was that there was no guarantee that the cleaning step performed by Stacks was the same as the cleaning step performed by Trimmomatic in the original study, and so the parameters used may be off. It seemed as if similar algorithms were used for both (both used a sliding window to discard low quality bases, for instance), but Trimmomatic did have some functions that Stacks lacked, such as the ability to specifically discard low quality bases at the start and end of the read. While the functions that Stacks did have should have covered these lacks, making it unlikely that this step introduced a significant amount of error, it was difficult to say for certain. If this study was repeated in future, an additional step should be added comparing the quality of reads cleaned by Stacks to those cleaned by Trimmomatic, using FastQC to check quality. The lack of this step was a key omission in this protocol, and cast more doubt on the quality of the final results.

Chapter 4.4c: Sources Of Error: Mapping

The third and final stage considered for sources of error was the mapping stage, where the identified high quality sequences from the previous steps were combined into loci, with the catalog of these loci forming a de novo genetic map of the species tested. This stage was somewhat different to the previous stages, as errors introduced here were relatively easy to correct and significant effort aws put into identifying and removing possible sources of error. However, unlike previous stages, some degree of error was practically unavoidable at this stage. The focus of the study was to minimise and mitigate the error introduced as best as possible, but as always it was a balancing act.

The main source of error in this stage would arise from problematic parameters being set for mapping, and so significant effort was put into identifying the best possible parameters to use. There was no set rule as far as which parameters to use goes, it was entirely dependent on the data set used in the analysis. There were three main parameters which could be adjusted; the minimum number of reads required to form a stack (m), the maximum number of mismatches allowed between reads to combine into a single stack (M) and the maximum number of mismatches allowed between stacks to combine into a single consensus locus (N). Setting m too high risks discarding usable fragments and reducing the amount of loci available to be analysed, while setting m too low risks including irrelevant outlier sequences in the final analysis and changing the results. Similarly, setting M (or N) too high risks unrelated reads/loci being overmerged into a single stack/consensus locus, while setting M (or N) too low risks different reads/alleles of a single locus/consensus locus being undermerged and split into separate loci. In other words, some error was likely introduced regardless of what value these parameters were set to, and the choice of parameters was aimed primarily at minimising this error, rather than eliminating it. To choose the right parameters, the mapping protocol was run on a representative sample of the data set with a variety of parameters, and several factors (coverage, number of loci present in 80% of the sample set and number of polymorphic loci present in 80% of the sample set) were measured to determine the most suitable parameters for the data set. Immediately, several issues presented themselves. To create a representative sample set for the data, the decision was made to exclude samples which contained either far more or far less reads than average, on the basis that these are outliers and unrepresentative of the larger population of samples. While this was true, the samples which contained far more reads than average were, from another point of view, a larger fraction of the total population than several samples with average numbers of reads combined. Excluding these from the parameter selection tests meant that the sample set used was unrepresentative of the overall population (of loci if not of samples), the effect of which could be seen in the final mapped data set. The majority of the samples missed the majority of loci, as these loci were only found in the "outlier" samples with higher amounts of reads and, consequently, higher coverages than the rest of these samples. It was possible, indeed likely, that this was unavoidable and no selection of mapping parameters could have avoided this. If a small number of samples contained many more reads than the others, they were likely to have many more loci identified in them as well, and it would be difficult to ensure that the majority of loci are expressed in the majority of samples. That said, including one of these outlier samples in the sample set would have at least allowed this to be attempted, with a more accurate assessment of the number of loci present in 80% of the population being possible (only the loci from 80% of the population are counted specifically to account for similar outliers). If this study was repeated and a similar disparity in sample reads was found, this should be a step taken.

This exclusion of outlier samples when setting parameters could also have led to the worryingly low coverages of several samples following mapping. Excluding PIG_11_THRUM, which while technically below 10x with a coverage of 9.96 was deemed close enough not to pose concerns, there were 12 samples with coverage below 10x, 40% of the total data set. This could have been solved by including the outlier samples in the parameter testing to obtain more accurate results, then redoing the mapping, increasing the value of m used to increase the coverage. However, there were several reasons not to do this. Firstly, of these 12 samples with low coverage, 7 were thrum morphs while 5 were pin morphs. In other words, neither morph seemed disproportionately affected by low coverage. As the core of this analysis is comparison of the genetics of these two morphs, this meant the analysis should not have been affected by this low coverage. It did not artificially create an inequality between these morphs, and so should not create a false positive result where a loci was not detected in one morph purely because of the relative lack of coverage (even if there would be a risk of false negatives being seen due to loci in all of one morph not being seen in the low coverage samples). Secondly, based on the increase in coverage of the samples tested when determining parameters, increasing m to 5 would not raise the majority of the low coverage samples above 10. It would likely raise the coverage by a reasonable amount, but at the cost of a large number of loci. This would seem to imply that m should be raised to larger values, even at the cost of more loci – but increasing m above 5 is explicitly discouraged by Paris et al (2017). They found that for values of m above 6 the accuracy of the mapping significantly decreased, as lower quality secondary reads were incorporated into stacks to reach the minimum number of reads required, resulting in lower quality stacks and inaccurate mapping. This could even result in certain samples being completely excluded from the analysis. If samples had an average coverage of 6, say, this implied that the majority of their loci were composed of stacks of 6 reads. If the minimum number of reads required to form a stack was increased to 7, these loci could not be formed and the sample could be removed from the mapping. This would have caused a major impact on the number of loci present in 80% of the population, if there was only 80% of the population to assess. So it seemed unlikely that, however the parameters were changed and regardless of whether the outlier samples were included, the coverage in these samples could have been increased to a reasonable value. This source of error was introduced prior to the mapping stage.

Other issues with the mapping were minor, with the main one being that there was some evidence that suggested that a second program should have been incorporated into the mapping pipeline specifically to align loci against a de novo genetic map. This was a necessary step to generate bam files, which were the required input for the privacy rarefaction algorithm, and was the reason why Stacks 2 was used for this analysis rather than the original Stacks (as Stacks 2 contained a program which can perform this step). However, a recent study (LaCava *et al*. 2019) showed that Stacks 2 was worse at doing this than another pipeline, dDocent, which was shown to be more accurate at assembling known genomes without access to a reference from sampling data. Incorporating this into the method would have required changing the mapping program used from Stacks to dDocent, which would run the risk of generating more error through altering output files from one program (namely the Stacks process_radtags program) to allow them to be run by another. As such, it was decided that the slight benefit which changing programs would provide was not worth the additional risk of error, but for future studies it might be an option to consider.

Generally though, the mapping stage had a low risk of causing additional errors in the data, and every step was taken in the method to avoid all errors which were not an inherent part of the mapping process. The error present in the final data set was most likely introduced in earlier stages of the method. As such, if this study were repeated there are several key changes which should be made – using different barcodes or not increasing the number of allowed mismatches for barcodes to two, checking the quality of the sequences after cleaning to be sure that the cleaning step was sufficient and including samples with high quantities of reads in the sample set for determining mapping parameters. But given the discussion of the previous stages it seems most likely that the errors were primarily a consequence of the samples used. To minimise these issues in further studies, alternative samples should be prepared for replicate analyses.

Chapter 4.5: Conclusion

The original goal of this project was to sequence a balanced selection of thrum and pin morph samples of *Linum tenue*, to identify genetic differences between the morphs, and to use these to determine both whether heterostyly is inherited hemizygously or heterozygously in *Linum tenue* and to identify genetic markers which could help identification of the genes involved. With this in mind, it can be concluded that the project was a partial success. A balanced selection of thrum and pin morph samples of *Linum tenue* was sequenced, and despite some sequencing errors and problems with the sample material, the final data set was still robust enough to perform further analysis on. From this sequencing data, several loci were identified with significant genetic differences between the two stylar morphs. These could not prove definitively whether heterostyly in *Linum tenue* is controlled (and therefore inherited) hemizygously or heterozygously, as there was little conclusive evidence for either explanation. The privacy rarefaction algorithm did seem to demonstrate that it could not be controlled hemizygously, but this analysis was extremely specific and so prone to false negatives. Several potential heterozygous candidates and one hemizygous candidate for heterostyly were identified, however, by a later analysis using PLINK. Closer examination of these candidate loci found that two of the heterozygous candidates encoded homologs to known proteins – a valine-tRNA synthetase and a cysteine protease associated with senescence. These proteins could be involved in controlling heterostyly in *Linum tenue*, or could be indicative of the type of proteins which are. Regardless of homologies, all of these candidate loci identified were likely candidates for genetic markers of the S-locus region in *Linum tenue*, with 121788 showing the most pronounced sequence differences between morphs and so being most likely to be a viable S-locus genetic marker.

These conclusions do provide some potential avenues for further research into heterostyly in *Linum tenue*. This study could be repeated with the lessons learned from this project incorporated, with freshly extracted samples and a greater number of replicates. Proteomic analysis could aim to detect the proteins identified in this study in live samples of *Linum tenue*, to localise them and analyse expression differences between morphs to more accurately determine whether they are involved in heterostyly. Transcriptomic analysis of RNA sequences corresponding to the candidate loci identified could serve a similar purpose, and if these analyses confirm that the candidate loci identified are indeed associated with heterostyly they could serve as genetic markers for targeted sequencing of a potential S locus region. This would allow characterization of the genes involved in heterostyly in *Linum tenue*, and would likely determine whether heterostyly in this species is controlled heterozygously or hemizygously. In conclusion, while this project was not entirely successful, it has provided promising avenues for further research. If these are pursued, there is potential to gain a much greater understanding of heterostyly in *Linum tenue*, and by extension, the *Linum* genus as a whole.

Bibliography

Aii, J., Nagano, M., Penner, A.G., Campbell, G.C. and Adachi, T. Identification of RAPD Markers Linked to the Homostylar (Ho) Gene in Buckwheat *Breed Sci.* **48(1)**, 59-62 (1998)

Armbruster, W.S., Pérez-Barrales, R., Arroya, J., Edwards, M.E. and Varga, P. Three-dimensional reciprocity of floral morphs in wild flax (*Linum suffruticosum*): a new twist on heterostyly. *New Phyt.* **171(3)**, 581-590 (2006)

Arunkumar, R., Wang, W., Wright, S.I. and Barrett. S.C.H. The genetic architecture of tristyly and its breakdown to self-fertilization. *Mol Ecol.* **26(3)**, 752-765 (2017)

Athanasiou, A. and Shore, J.S. Morph-specific proteins in pollen and styles of distylous turnera (Turneraceae). *Genetics* **146(2)**, 669-679 (1997)

Athanasiou, A., Khosravi, D., Tamari, F. and Shore, J.S. Characterization and localization of short-specific polygalacturonase in distylous *Turnera subulata* (Turneraceae). *Am. J. Bot.* **90(5)**, 675-682 (2003)

Barrett, S.C.H. The evolutionary biology of tristyly. *In: Futuyma D, and Antonovics J, eds. Oxford surveys in evolutionary biology*, Oxford: Oxford University Press, 283–326 (1993)

Barrett, S.C.H., Lloyd D.G. and Arroyo, J. Stylar polymorphisms and the evolution of heterostyly in *Narcissus* (Amaryllidaceae). *In: Lloyd DG, Barrett SCH (eds) Floral biology: Studies on floral evolution in animal-pollinated plants.* New York, USA: Chapman and Hall. 339–376 (1996)

Barrett, S.C.H., Jesson, L.K. and Baker, A.M. The Evolution and Function of Stylar Polymorphisms in Flowering Plants *Botany* **85**, 253-265 (2000)

Barrett, S.C.H., Wilken, D.H. and Cole, W.W. Heterostyly in the Lamiaceae: The case of *Salvia brandegeei*. *Plant Syst. Evol.* **223**, 253-265 (2000)

Barrett, S.C.H. Mating strategies in flowering plants: the outcrossing-selfing paradigm and beyond. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **358(1434)**, 991-1004 (2003)

Barrett, S.C.H. and Shore, J.S. New Insights on Heterostyly: Comparative Biology, Ecology and Genetics *In: Self-Incompatibility in Flowering Plants* Springer, Berlin, Heidelberg (2008)

Bateson, W. and Gregory, R.P. On the inheritance of heterostylism(sic) in *Primula*. *Proc. R. Soc. Londser B* **76**, 581-586 (1905)

Bertin, R.I. and Newman, C.M. Dichogamy in Angiosperms. *The Botanical Review* **59(2)**, 113-146 (1993)

Brennan, A.C. Distyly supergenes as a model to understand the evolution of genetic architecture. *Am. J. Bot.* **104(1)**, 5-7 (2017)

Brown, B.J. and Mitchell, R.J. Competition for pollination: effects of pollen of an invasive plant on seed set of a native congener. *Oecologia*. **129(1)**, 43-49 (2001)

Catchen, J., Amores, A., Hohenlohe, P., Cresko, W. and Postlethwait, J. Stacks: building and genotyping loci de novo from short-read sequences. *G3: Genes, Genomes, Genetics*, **1**, 171-182, (2011)

Catchen, J., Hohenlohe, P., Bassham, S, Amores, A. and Cresko, W. Stacks: an analysis tool set for population genomics. *Mol. Ecol.*, **22(11)**, 3124-3140, (2013)

Charlesworth, D. and Charlesworth, B. A Model for the Evolution of Distyly *Am. Nat.*, **114(4)**, 467-498 (1979a)

Charlesworth, D. and Charlesworth, B. The Maintenance and Breakdown of Distyly *Am. Nat.*, **114(4)**, 499-513 (1979b)

Cheng, H., Song, S., Xiao, L., Soo, H.M., Cheng, Z., Xie, D. and Peng, J. Gibberellin Acts through Jasmonate to Control the Expression of *MYB21*, *MYB24*, and *MYB57* to Promote Stamen Filament Growth in *Arabidopsis*. *PLoS Genet.* **5(3)**, e1000440 (2009)

Clarke, C.A., Sheppard, P.M. and Thornton, I.W.B. The genetics of the mimetic butterfly *Papilio Memnon* L. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **254(791)**, 37-89 (1968)

Clarke, C.A. and Sheppard, P.M. The genetics of the mimetic butterfly *Papilio polytes* L. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **263(855)**, 431-458 (1972)

Cocker, J. M., Webster, M. A., Li, J., Wright, J., Kaithakottil, G., Swarbreck, D. and Gilmartin, P. M. *Oakleaf*: an S locus-linked mutation of *Primula vulgaris* that affects leaf and flower development. *New Phyt.* **208(1)**, 149–161 (2015)

Cocker, J.M., Wright, J., Li, J., Swarbreck, D., Dyer, S., Caccamo, M. and Gilmartin, P.M. *Primula vulgaris* (primrose) genome assembly, annotation and gene expression, with comparative genomics on the heterostyly supergene *Sci. Rep.* **8**, 17942 (2018)

Cohen, J.I. *De novo* Sequencing and Comparative Transcriptomics of Floral Development of the Distylous Species *Lithospermum multiflorum*. *Front Plant Sci*, **23(7)**, 1934 (2016)

Costa, J., Ferrero, V., Loureiro, J., Castro, M., Navarro, L. and Castro, S. Sexual reproduction of the pentaploid, short-styled Oxalis pes-caprae allows the production of viable offspring. *Plant Biol (Stuttg)*. **16(1)**, 208-14 (2013)

Costa, J., Castro, S., Loureiro, J. and Barrett, S.C.H. Experimental insights on Darwin's cross-promotion hypothesis in tristylous purple loosestrife (*Lythrum salicaria*). *Am. J. Bot*. **104(4)**, 616-626 (2017)

Darwin, C.R. On the two forms or dimorphic condition in the species of *Primula*, and on their remarkable sexual relations. *Journal of the Proceedings of the Linnean Society, Botany* **6**, 77-96 (1862)

Darwin, C.R. On the existence of two forms, and of their reciprocal sexual relation, in several species of the genus *Linum*. *Journal of the Proceedings of the Linnean Society, Botany* **7**, 69-83

Darwin, C.R. *The different forms of flowers on plants of the same species*. London, UK: John Murray (1877)

Das, A., Pandit, M.K., Bairagi, S., Saha, S. and Muthaiah, K. A Study on Floral Morphology of Brinjal Genotypes in Gangetic-Alluvial Zone of West Bengal, India. *Int. J. Curr. Microbiol.* **6(10)**, 3323-3331 (2017)

De Vos, J.M., Hughes, C.E., Schneeweiss, G.M., Moore, B.R. and Conti, E. Heterostyly accelerates diversification via reduced extinction in primroses. *Proc. R. Soc. B.* **281(1784)**, 20140075 (2014)

Dowrick, J. and Pamela, V. Heterostyly and homostyly in *Primula obconica*. *Heredity* **10**, 219-236 (1956)

Englisch-Peters, S., von der Haar, F. and Cramer, F. Fidelity in the aminoacylation of tRNA(Val) with hydroxy analogues of valine, leucine, and isoleucine by valyl-tRNA synthetases from Saccharomyces cerevisiae and Escherichia coli. *Biochemistry.* **29(34)**, 7953-8 (1990)

Ernst, A. Weitere Untersuchungen zur Phänanalyse, zum Fertilitäts problem und zur Genetik heterostyler *Primeln*. I. *Primula viscosa*. *Arch Klaus-Stift Vererb Forsch*, **8**, 1–215 (1933)

Ernst, A. Self-fertility in monomorphic primulas. *Genetica*, **27**, 391–448 (1955)

Ferrero, V., Arroyo, J., Castro, S. and Navarro, L. Unusual heterostyly: style dimorphism and self-incompatibility are not tightly associated in *Lithodora* and *Glandora* (Boraginaceae). *Annals of Botany*, **109(3)**, 655-665 (2012)

Ferrero, V., Barrett, S.C.H., Castro, S., Caldeirinha, P., Navarro, L., Loureiro, J. and Rodríguez-Echeverría, S. Invasion genetics of the Bermuda buttercup (Oxalis pes-caprae): complex intercontinental patterns of genetic diversity, polyploidy and heterostyly characterize both native and introduced populations. *Mol. Ecol.* **24(9),** 2143-55 (2015)

Fersht, A.R. and Kaethner, M.M. Enzyme hyperspecificity. Rejection of threonine by the valyl-tRNA synthetase by misacylation and hydrolytic editing. *Biochemistry.* **15(15)**, 3342-6 (1976)

Fountain, E.D., Pauli, J.N., Reid, B.N., Palsbøll, P.J. and Peery, M.Z. Finding the right coverage: the impact of coverage and sequence quality on SNP genotyping error rates. *Mol. Ecol. Resources*, **16**, 966–978. (2016)

Fukai, S., Nureki, O., Sekine, S., Shimada, A., Tao, J., Vassylyev, D.G. and Yokoyama, S. Structural basis for double-sieve discrimination of L-valine from L-isoleucine and L-threonine by the complex of tRNA(Val) and valyl-tRNA synthetase. *Cell.* **103(5)**, 793-803 (2000)

Ganders, F.R. The biology of heterostyly. *New Zealand Journal of Botany*, **17(4)**, 607-635 (1979)

Gilmartin, P.M. On the origins of observations of heterostyly in *Primula*. *New Phyt.*, **208(1)**, 39-51 (2015)

Gubbay, J., Collignon, J., Koopman, P., Capel, B., Economou, A., Münsterberg, A., Vivian, N., Goodfellow, P. and Lovell-Badge, R. A gene mapping to the sex-determining region of the mouse Y chromosome is a member of a novel family of embryonically expressed genes. *Nature* **346(6281)**, 245-250 (1990)

Heitz, B. La pollinisation des Lins heterostyles du groupe *Linum perenne* L. (Linaceae). *Compte Rendues de L'Académie des Sciences, Paris* **290**, 811-814 (1980)

Henning, P.M., Shore, J.S. and McCubbin, A.G. Transcriptome and Network Analyses of Heterostyly in *Turnera subulata* Provide Mechanistic Insights: Are S-Loci a Red-Light for Pistil Elongation? *Plants* **9(6)**, 713 (2020)

Hou, X., Hu, W. W., Shen, L., Lee, L. Y., Tao, Z., Han, J. H. and Yu, H. Global identification of DELLA target genes during Arabidopsis flower development. *Plant Phys.* **147(3)**, 1126–1142. (2008)

Hughes, A.R., Inouye, B.D., Johnson, M.T.J., Underwood, N. and Vellend, M. Ecological consequences of genetic diversity. *Ecology Letters*, **11**, 609-623 (2008)

Huu, C. N., Kappel, C., Keller, B., Sicard, A., Takebayashi, Y., Breuninger, H., Nowak, M. D., Bäurle, I., Himmelbach, A., Burkart, M., Ebbing-Lohaus, T., Sakakibara, H., Altschmied, L., Conti, E. and Lenhard,

M. Presence versus absence of CYP734A50 underlies the style-length dimorphism in primroses. *eLife*, **5**, e17956. (2016)

Jhala, A.J. and Hall, L.M. Flax (*Linum usitatissumum* L.): Current Uses and Future Applications. *Australian Journal of Basic and Applied Sciences*, **4(9)**, 4304-4312 (2010)

Kappel, C., Huu, C.N. and Lenhard, M. A short story gets longer: recent insights into the molecular basis of heterostyly. *Journal of Experimental Botany*, **68(21-22)**, 5719–5730 (2017)

Keller, B., Thomson, J.D. and Conti, E. Heterostyly promotes disassortative pollination and reduces sexual interference in Darwin's primroses: evidence from experimental studies. *Functional Ecology,* **28(6)**, 1413-1425 (2014)

Khosravi, D., Yang, E.C.C., Siu, K.W.M. and Shore, J. S. High Level of α-Dioxygenase in Short Styles of Distylous *Turnera* Species. *Int. J. Plant Sci.* **165(6)**, 995-1006 (2004)

Koopman, P., Gubbay, J., Vivian, N., Goodfellow, P. and Lovell-Badge, R. Male development of chromosomally female mice transgenic for Sry. *Nature* **351(6322)**, 117-121 (1991)

Kowalska, G. The influence of heterostyly, pollination method and harmonization on eggplant's (*Solanum melongea* L.) flowering and fruiting. *Acta Agrobotanica* **56(1-2)**, 61-76 (2003)

Kurian, V. and Richards, A.J. A new recombinant in the heteromorphy 'S' supergene in *Primula*. *Heredity* **78**, 383-390 (1997)

Labonne, J.J.D., Goultiaeva, A. & Shore, J.S. High-resolution mapping of the S-locus in Turnera leads to the discovery of three genes tightly associated with the S-alleles. *Mol Genet Genomics* **281**, 673 (2009)

Labonne, J., Tamari, F. & Shore, J. Characterization of X-ray-generated floral mutants carrying deletions at the S-locus of distylous *Turnera subulata*. *Heredity* **105**, 235–243 (2010)

LaCava, M.E.F., Aikens, E.O., Megna, L.C., Randolph, G., Hubbard, C. and Buerkle, C.A. Accuracy of *de novo* assembly of DNA sequences from double-digest libraries varies substantially among software *Mol. Ecol. Resources*, **20(2)**, 360-370 (2019)

Lewis D. and Jones D.A. The genetics of heterostyly. *In: Barrett SCH (ed) Evolution and function of heterostyly.* Berlin, Heidelberg, New York: Springer 129–150 (1992)

Li, J., Webster, M., Furuya, M. and Gilmartin, P.M. Identification and characterization of pin and thrum alleles of two genes that co-segregate with the Primula S locus *Plant J.* **50(1)**, 18-31 (2007)

Li, J., Webster, M. A., Dudas, B., Cook, H. E., Manfield, I., Davies, B. H. and Gilmartin, P.M. The S locus-linked Primula homeotic mutant sepaloid shows characteristics of a B-function mutant but does not result from mutation in a B-function gene. *Plant J.* **56(1)**, 1-12 (2008)

Li, J., Dudas, B., Webster, M.A., Cook, H.E., Davies, B.H. and Gilmartin P.M. *Hose in Hose*, an *S* locus-linked mutant of *Primula vulgaris*, is caused by an unstable mutation at the *Globosa* locus. *PNAS* **107(12)**, 5664-5668 (2010)

Li, J., Cocker, J.M., Wright, J., Webster, M.A., McMullan, M., Dyer, S., Swarbreck, D., Caccamo, M., Oosterhout, C.V. and Gilmartin, P.M. Genetic architecture and evolution of the S locus supergene in Primula vulgaris. *Nat Plants.* **2(12**), 16188 (2016)

Liu, L., Zhou, Y., Szczerba, M.W., Li, X. and Lin, Y. Identification and application of a rice senescence-associated promoter. *Plant Phys.* **153(3)**, 1239-49

Lloyd, D.G. and Webb, C.J. The Evolution of Heterostyly. *In: Barrett S.C.H. (eds) Evolution and Function of Heterostyly. Monographs on Theoretical and Applied Genetics, vol 15*. Springer, Berlin, Heidelberg. 151-178 (1992)

Matsui, K., Tetsuka, T., Nishio, T. and Hara, T. Heteromorphic incompatibility retained in self-compatible plants produced by a cross between common and wild buckwheat. *New Phyt.* **159(3)**, 701-708 (2003)

Matsui, K., Mizuno, N., Ueno, M., Takeshima, R., and Yasui, Y. Development of co-dominant markers linked to a hemizygous region that is related to the self-compatibility locus (S) in buckwheat (*Fagopyrum esculentum*). *Breeding science*, **70(1)**, 112–117 (2020)

Miljuš-Đukić, J., Ninković, S., Radović, S., Masksimović, V., Brkljaĉić, J. and Nešković, M. Detection of Proteins Possibly Involved in Self-Incompatibility Response in Distylous Buckwheat. *Biologica Plantarum* **48**, 293-296 (2004)

Nakazato, T., Ohta, T. and Bono H. Experimental design-based functional mining and characterization of high-throughput sequencing data in the sequence read archive. *PLoS ONE*, **8(10)**, e77910 (2013)Ohnishi, O. and Zhou M. Chapter Ten – Annual Self-Compatible Species *In: Zhou, M., Kreft, I., Suvorova, G., Tang, Y. and Woo, S.H. Buckwheat Germplasm in the World* Academic Press. 81-88 (2018)

Ozkan, M.T., Aliyazicioglu, R., Demir, S., Misir, S., Turan, I., Yildirmis, S. and Aliyazicioglu, Y. Phenolic Characterisation and Antioxidant Activity of *Primula vulgaris* and Its Antigenotoxic Effect on Fibroblast Cells. *Jundishapur J. Nat. Pharm. Prod.* **12(1)**, e40073 (2016)

Paris, J., Stevens, J. and Catchen, J. Lost in parameter space: a road map for Stacks. *Methods in Ecology and Evolution*, **8(10)**, 1360-1373, (2017)

Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S. and Hoekstra, H.E. Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. *PLoS ONE*, **7(5)**, e37135 (2012)

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bende, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J. and Sham, P.C. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses *Am. J. Hum. Genet.*, **81(3)**, 559-575 (2007)

Richards, J. *Primula*, *2nd edn*. Timber Press, Portland, OR, USA (2003)

Rochette, N., Rivera-Colón, A. and Catchen, J. Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. *Mol. Ecol.*, **28(21)**, 4737-4754. (2019)

Ruiz-Martin, J., Santos-Gally, R., Escudero, M., Midgley, J.J., Pérez-Barrales, R. and Arroyo, J. Style polymorphism in *Linum* (Linaceae): a case of Mediterranean parallel evolution? *Plant Bio.* **20(1)**, 100-111 (2018)

Santos-Gally, R., Gonzales-Voyer, A. and Arroyo, J. Deconstructing heterostyly: the evolutionary role of incompatibility systems, pollinators and floral architecture. *Evolution* **67(7)**, 2072-82 (2013)

Scharmann, M., Grafe, T. U., Metali, F. and Widmer, A. Sex-determination and sex chromosomes are shared across the radiation of dioecious Nepenthes pitcher plants. *bioRxiv* 240259 (2017)

Shao, J.W., Wang, H.F., Fang, S.P., Conti, E., Chen, Y.J. and Zhu, H.M. Intraspecific variation of self-incompatibility in the distylous plant *Primula merrilliana*. *AoB Plants* **11(3)**, plz030 (2019)

Shore, J.S., Hamam, H.J., Chafe, P.D.J., Labonne, J.J.D., Henning, P.M. and McCubbin, A.G. The long and short of the S-locus in *Turnera* (*Passifloraceae*) *New Phyt.* **224(3)**, 1315-1329 (2019)

Sinclair, A.H., Berta, P., Palmer, M.S., Hawkins, J.R., Griffiths, B.L., Smith, M.J., Foster, J.W., Frischauf, A.M., Lovell-Badge, R. and Goodfellow, P. A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif. *Nature* **346(6281)**, 240-244 (1990)

Srinivas, G., Jayappa, A.H. and Patel, A.I. Heterostyly: A Threat to Potential Fruit Yield in Brinjal (*Solanum melongena* L.). *Adv. Life Sci.* **5(4)**, 1211-1215 (2016)

Stelzer, C.P. Does the avoidance of sexual costs increase fitness in asexual invaders? *PNAS* **112(29)**, 8851-8858 (2015)

Stone, J.D. Pollen donation patterns in a tropical distylous shrub (*PSYCHOTRIA SUERRENSIS*; RUBIACEAE). *Am. J. Bot.* **82(11)**, 1390-1398 (1995)

Suzuki, M., Wu, S., Li, Q. and McCarty, D.R. Distinct functions of COAR and B3 domains of maize VP1 in induction of ectopic gene expression and plant developmental phenotypes in Arabidopsis. *Plant Mol Biol.* **85(1-2)**, 179-91 (2014)

Sveinsson, S., McDill, J., Wong, G.K.S., Li, J., Li, X., Deyholos, M.K. and Cronk, Q.C.B. Phylogenetic pinpointing of a paleopolyploidy event within the flax genus (Linum) using transcriptomics*, Annals of Botany*, **113(5)**, 753–761 (2014)

Takayama, S. and Isogai, A. Self-Incompatibility in Plants. *Annu. Rev. Plant Biol.*, **56**, 467-489 (2005)

Takeshima, R., Nishio, T., Komatsu, S., Kurauchi, N. and Matsui, K. Identification of a gene encoding polygalacturonase expressed specifically in short styles in distylous common buckwheat (*Fagopyrum esculentum*). *Heredity* **123**, 492-502 (2019)

Tippery, N.P. and Les, D.H. Phylogenetic Relationships and Morphological Evolution in *Nymphoides* (Menyanthaceae). *Systematic Botany* **36(4)**, 1101-1113 (2011)

Tomita, R.N., Suzuki, G., Yoshida, K., Yano, Y., Tsuchiya, T., Kakeda, K., Mukai, Y. and Kowyama, Y. Molecular Characterization of a 313-kb Genomic Region Containing the Self-incompatibility Locus of *Ipomoea trifida*, a Diploid Relative of Sweet Potato. *Breeding Science* **54,** 165-175 (2004)

Truyens, S., Arbo, M.M. and Shore, J.S. Phylogenetic relationships, chromosome and breeding system evolution in Turnera (Turneraceae): inferences from its sequence data. *Am. J. Bot.* **92(10)**, 1749-1758 (2005)

Ushijima, K., Nakano, R., Bando, M., Shigezane, Y., Ikeda, K., Namba, Y., Kume, S., Kitabata, T., Mori, H. and Kubo, Y. Isolation of the floral morph-related genes in heterostylous flax (*Linum grandiflorum*): the genetic polymorphism and the transcriptional and post-transcriptional regulations of the S locus *Plant J.* **69(2)**, 317-331 (2011)

Ushijima, K., Ikeda, K., Nakano, R., Matsubara, M., Tsuda, Y. and Kubo Y. Genetic Control of Floral Morph and Petal Pigmentation in *Linum grandiflorum* Desf., a Heterostylous Flax *Horticulture J.* **84(3)**, 261-268 (2015)

Wang, H., Tang, J., Liu, J., Hu, J., Liu, J., Chen, Y., Cai, Z. and Wang, X. Abscisic Acid Signaling Inhibits Brassinosteroid Signaling through Dampening the Dephosphorylation of BIN2 by ABI1 and ABI2 *Mol. Plant* **11(2)**, 315-325 (2018)

Wang, Y., Liu, A., Li, W., Jiang, Y., Song, S., Li, Y. and Chen, R. Comparative proteomic analysis of eggplant (*Solanum melongena* L.) heterostylous pistil development *PLoS ONE* **12(6)**, (2017)

Weber, J.J., Weller, S.G., Sakai, A.K., Tsyusko, O.V., Glenn, T.C., Domínguez, C.A., Molina-Freaner, F.E., Fornoni, J., Tran, M., Nguyen, N., Nguyen, K., Tran, L.K., Joice, G. and Harding, E. The role of inbreeding depression and mating system in the evolution of heterostyly. *Evolution* 2013 **67(8)**, 2309-22 (2013)

Wolfe, L.M. Associations among Multiple Floral Polymorphisms in *Linum pubescens* (Linaceae), a Heterostylous Plant *Int. J. Plant Sci.* **162(2)**, 335-342 (2001)

Wong, K.C., Watanabe, M. and Hinata, K. Fluorescence and scanning electron microscopic study on self-incompatibility in distylous *Averrhoa carambola* L. *Sexual Plant Reprod.* **7,** 116–121 (1994)

Yang, Y.Y. and Kim, J.G. The optimal balance between sexual and asexual reproduction in variable environments: a systematic review. *Journal of Ecology and Environment* **40(12)**, 1-18 (2016)

Yasui, Y., Mori, M., Matsumoto, D., Ohnishi, O., Campbell, C.G. and Ota, T. Construction of a BAC library for buckwheat genome research - an application to positional cloning of agriculturally valuable traits. *Genes Genet Syst.* **83(5)**, 393-401 (2008)

Yasui, Y., Mori, M., Aii, J., Abe, T., Matsumoto, D. Sato, S., Hayashi, Y., Ohnishi, O. and Ota, T. S-LOCUS EARLY FLOWERING 3 Is Exclusively Present in the Genomes of Short-Styled Buckwheat Plants that Exhibit Heteromorphic Self-Incompatibility *PLoS ONE* **7(2)**, 1-9 (2012)

Yasui, Y., Hirakawa, H., Ueno, M., Matsui, K., Katsube-Tanaka, T., Yang, S. J., Aii, J., Sato, S. and Mori, M. Assembly of the draft genome of buckwheat and its applications in identifying agronomically useful genes. *DNA Res.* **23(3)**, 215–224 (2016)

You, F.M., Xiao, J., Li, P., Yao, Z., Jia, G., He, L., Zhu, T., Luo, M.C., Wang, X., Deyholos, M.K. and Cloutier, S. Chromosome-scale pseudomolecules refined by optical, physical and genetic maps in flax. *Plant J.*, **95(2)**, 371-384 (2018)

Yuan, S., Barrett, S.C.H, Duan, T., Qian, X., Shi, M. and Zhang, D. Ecological correlates and genetic consequences of evolutionary transitions from distyly to homostyly. *Ann. Bot.*, **120(5)**, 775-789 (2017)

Zhou, W., Barrett, S.C.H., Li, H.D., Wu, Z.K., Wang, X.J., Wang, H. and Li, D.Z. Phylogeographic insights on the evolutionary breakdown of heterostyly. *New Phyt.*, **214(3)**, 1368-1380 (2017)

Appendix

Sample	Forward Adapter Number	Barcode	Reverse Adapter Number	Barcode
ALT_05_PIN	i7	ATACAG	i11	GTTACT
ALT_24_PIN	i11	AGTAAC	i3	ACTGGA
ARA_19_PIN	i8	TGTTAC	i12	CGCTTG
BUR_20_THRUM	i12	CAAGCG	i8	GTAACA
BUR_25_THRUM	i11	AGTAAC	i5	ACTCCG
CAZ_12_THRUM	i12	CAAGCG	i6	AACGTG
CAZ_16_THRUM	i4	GTCTTA	i12	CGCTTG
CAZ_20_THRUM	i12	CAAGCG	i4	TAAGAC
CAZ_22_THRUM	i2	GATCCG	i12	CGCTTG
CAZ_23_THRUM	i12	CAAGCG	i7	CTGTAT
CBT_16_THRUM	i12	CAAGCG	i11	GTTACT
CBT_17_THRUM	i11	AGTAAC	i1	TCATGC
CBT_18_PIN	i7	ATACAG	i12	CGCTTG
CBT_20_PIN	i9	ACGCTC	i11	GTTACT
EBO_14_THRUM	i11	AGTAAC	i12	CGCTTG
EBO_16_PIN	i11	AGTAAC	i7	CTGTAT
ELB_30_PIN	i11	AGTAAC	i8	GTAACA
LUM_15_THRUM	i8	TGTTAC	i11	GTTACT
LUM_22_PIN	i11	AGTAAC	i9	GAGCGT
LUM_25_THRUM	i3	TCCAGT	i12	CGCTTG
MDA_24_THRUM	i11	AGTAAC	i11	GTTACT
PIG_11_THRUM	i12	CAAGCG	i1	TCATGC
PIG_12_PIN	i12	CAAGCG	i9	GAGCGT
PIG_13_PIN	i12	CAAGCG	i2	CGGATC
PIG_15_PIN	i11	AGTAAC	i10	TGCCAA
PIG_22_PIN	i12	CAAGCG	i10	TGCCAA
PIG_29_PIN	i1	GCATGA	i2	CGGATC
PIG_35_PIN	i5	CGGAGT	i12	CGCTTG

SVT_17_THRUM	i9	ACGCTC	i12	CGCTTG
SVT_24_PIN	i12	CAAGCG	i12	CGCTTG

Table A.1, a complete list of which adapters and which barcodes corresponded to each sample sequenced

a)

Locus number	BUR_20_THRUM	BUR_25_THRUM	CAZ_12_THRUM	CAZ_16_THRUM	CAZ_20_THRUM	CAZ_22_THRUM	CAZ_23_THRUM	CBT_16_THRUM	CBT_17_THRUM	EBO_14_THRUM	LUM_15_THRUM	LUM_25_THRUM	MDA_24_THRUM	PIG_11_THRUM	SVT_17_THRUM
4244	ACN/CAN	---	ACN/CAN	NNN/NNN	ACN/CAN	ACA/CAA	ACN/CAN	---	ACN/CAN	ACN/CAN	---	ACG/CAG	ACN/CAN	ACN/CAN	ACN/CAN
4243	ACN/CAN	---	ACN/CAN	NNN/NNN	ACN/CAN	ACN/CAN	ACN/CAN	---	ACN/CAN	ACN/CAN	---	ACG/CAG	ACN/CAN	ACN/CAN	ACN/CAN
5812	NNNT/NNNT	NNNT/	NNNT/NNNT	NNNT/NNNT	NNNT/NNNT	NNNT/NNNT	NNNT/NNNT	NNNT/NNNT	NNNT/NNNT	---	NNNT/	NNNT/	NNNT/NNNT	NNNT/NNNT	TNNC/TNNC

		NNNT									NNNT	NNNT			
19158	NNNT/NNNT	NNNT/NNNT	NNNT/NNNT	NNNT/NNNT	NNNT/NNNT	NNNT/NNNT	NNNT/NNNT	NNNT/NNNT	NNNT/NNNT	---	NNNT/NNNT	NNNT/NNNT	NNNT/NNNT	NNNT/NNNT	TNNC/TNNC
121788	TCTCTT/TCTCTT	TCTCTT/TCTCTT	TCTCTT/TCTCTT	---	TCTCAT/TCTCTT	TCTCTT/TCTCTT	TCTCTC/TCTCTT	TCTCTT/TCTCTT	NCNCTN/NCNCTN	TCTCTT/TCTCTT	---	TCTCTT/TCTCTT	TCTCTT/TCTCTT	TCTCTC/TCTCTT	TCGCTT/TCTCTT
164265	CTCTT/CTCTT	CTCTT/CTCTT	CTCTT/CTCTT	---	CTCAT/CTCTT	CTCTT/CTCTT	CTCTC/CTCTT	CTCTT/CTCTT	CNCTN/CNCTN	CTCTT/CTCTT	---	CTCTT/CTCTT	CTCTT/CTCTT	CTCTC/CTCTT	CGCTT/CTCTT

304804	ATACGGNNNNTGTAT/CAACGGNNNNCGAGT	---	NNNNGNNNGNCGANT/NNNNGNNNGNCGANT	CNACGGNNNNCCANT/CNACGGNNNNCCANT	NNANGNTCANCGANT/NNANGNTCANCGANT	NNANGNTCNNCGANT/NNANGNTCNNCGANT	NNNNNNNNNNCGANT/NNNNNNNNNNCGANT	ATGTCNNNGACGAGT/CTACGNNNGGCGAAT	CNNNGNTCAACGANT/CNNNGNTCAACGANT	NNNNNNNNNNCGANN/NNNNNNNNNNCGANN	---	---	NNNNNGTCAACGANT/NNNNNGTCAACGANT	ANACGNNNGGCGAAT/CNACGNNNGGCGAAT	NNANGNNNGGCGANT/NNANGNNNGGCGANT
306179	TNNT/TNNT	TNNT/TNNT	TNNT/TNNT	---	TNNT/TNNT	TNNT/TNNT	TNNT/TNNT	TNNT/TNNT	TNNT/TNNT	TNNT/TNNT	TNNT/TNNT	TNNT/TNNT	TNNT/TNNT	TNNT/TNNT	TAGN/TAGN
359009	TNNT/TNNT	TNNT/TNNT	TNNT/TNNT	---	TNNT/TNNT	TNNT/TNNT	TNNT/TNNT	TNNT/TNNT	TNNT/TNNT	TNNT/TNNT	TNNT/TNNT	TNNT/TNNT	TNNT/TNNT	TNNT/TNNT	TAGN/TAGN

b)

Locus num	ALT_05_PIN	ALT_24_PIN	ARA_19_PIN	CBT_18_PIN	CBT_20_PIN	EBO_16_PIN	ELB_30_PIN	LUM_22_PIN	PIG_12_PIN	PIG_13_PIN	PIG_15_PIN	PIG_22_PIN	PIG_29_PIN	PIG_35_PIN	SVT_24_PIN

ber															
4244	ACG/CAG	ACN/CAN	ACN/CAN	ACG/CAG	ACG/CAG	---	ACN/CAN	ACN/CAN	ACN/CAN	ACN/CAN	ACN/CAN	ACN/CAN	---	ACG/CAG	ACG/CAG
	ACA/CAG	ACN/CAN	ACN/CAN	ACG/CAG	ACG/CAG	---	ACN/CAN	ACN/CAN	ACN/CAN	ACN/CAN	ACN/CAN	ACN/CAN	---	ACN/CAN	ACN/CAN
5812	CAGT/CAGT	NNNT/NNNT	---	NNNT/NNNT	NTCT/NTCT	NNNT/NNNT	NNNT/NNNT	NNNT/NNNT	NNNT/NNNT	NNNT/NNNT	NNNT/NNNT	NNNT/NNNT	---	NNNT/NNNT	NNNT/NNNT
19158	CAGT/CAGT	NNNT/NNNT	---	NNNT/NNNT	NTCT/NTCT	NNNT/NNNT	NNNT/NNNT	NNNT/NNNT	NNNT/NNNT	NNNT/NNNT	NNNT/NNNT	NNNT/NNNT	---	NNNT/NNNT	NNNT/NNNT
121788	GCTCTT/GCTCTT	---	TCTCTT/TCTCTT	GCTCTC/GCTCTT	TCTCTC/TCTCTT	TCTCTT/TCTCTT	TCTCTT/TCTCTT	TCTCTT/TCTTTT	TCTCTT/TGTCTT	NCGCTT/NCTCTT	---	NNNCTT/NNNCTT	---	TCTCTT/TCTCTT	TCTCTT/TGTCTT
	CTCTT/CTCTT	---	CTCTT/CTCTT	CTCTC/CTCTT	CTCTC/CTCTT	CTCTT/CTCTT	CTCTT/CTCTT	CTCTT/CTTTT	CTCTT/GTCTT	CGCTT/CTCTT	---	NNCTT/NNCTT	---	CTCTT/CTCTT	CTCTT/GTCTT
304804	---	---	---	NAGTNNNTNACGAAC/NAGTNN	NNACGACNGGCGANT/NNACGAC	---	---	---	---	---	---	---	---	CNNNNNNCNNCGANT/	---

				NTNACGAGC	NGGCGANT									CNNNNNNCNNCGANT	
306179	---	TNNT/TNNT	---	TGGN/TGTN	ANGN/TNGN	TNNT/TNNT	TNNT/TNNT	TNNT/TNNT	TNNT/TNNT	---	TNNT/TNNT	TNGC/TNGC	---	TNNT/TNNT	TNNT/TNNT
359009	---	TNNT/TNNT	---	TGGN/TGTN	ANGN/TNGN	TNNT/TNNT	TNNT/TNNT	TNNT/TNNT	TNNT/TNNT	---	TNNT/TNNT	TNGC/TNGC	---	TNNT/TNNT	TNNT/TNNT

Table A.2, a complete list of all haplotypes of all significantly associated loci for all samples, divided into a) thrum samples and b) pin samples.

```
#An alternative to Trimmomatic using the stacks program, so hopefully the output should be fully compatible with the rest of stacks.

#Adding more stringent quality parameters to match previous Trimmomatic scripts.


#Written by Lewis Edwards, 21st November 2019


#Setting the input directory, barcode list and output directory


Raw1=/ddn/data/jqfs17/Second_Time/Raw/Lane_6

Barcodes=/ddn/data/jqfs17/Second_Time/Info/Barcode_List.fastq

Clean=/ddn/data/jqfs17/Second_Time/Cleaned_Stringent_Barcode2_DRC


process_radtags -p $Raw1 --paired -i gzfastq \

-b $Barcodes -o $Clean \

-r -c -q -w 0.035 -s 20 -t 120 --inline_inline --renz_1 pstI --renz_2 mseI --disable_rad_check --barcode_dist_1 2 --barcode_dist_2 2
```

```
# -c to clean ambiguous reads, -q to clean low quality reads, -r to rescue mutated barcodes. -w setting size of sliding window as proportion of the read.

# -q setting score threshold. -t setting sequence end length.
```

Figure A.1, the final process_radtags script used, incorporating the cleaning step, the two barcode mismatches and disabling the restriction digestion site check

```
#A script to run through various combinations of parameters with Stacks mapping on a representative subset of samples

#(median coverage, equal proportions from populations, different locations).

#Written by Lewis Edwards, partially adapted from Rochette 2017, 13/11/2019


#Setting initial variables


popmap=/ddn/data/jqfs17/Second_Time/Info/TestPopMap.tsv

reads_dir=/ddn/data/jqfs17/Second_Time/Cleaned_Stringent2

m=2

M=1


while [ $m -lt 6 ]

do

    while [ $M -lt 10 ]

    do

        n1=$(( $M - 1 ))

        n2=$M

        n3=$(( $M + 1 ))

        out_dir1=/ddn/data/jqfs17/Second_Time/Tests/Stacks.m$m/M$M.N$n1

        out_dir2=/ddn/data/jqfs17/Second_Time/Tests/Stacks.m$m/M$M.N$n2

        out_dir3=/ddn/data/jqfs17/Second_Time/Tests/Stacks.m$m/M$M.N$n3
        log_file1=$out_dir1/denovo_map.oe

        log_file2=$out_dir2/denovo_map.oe

        log_file3=$out_dir3/denovo_map.oe

        denovo_map.pl --samples $reads_dir \

        -O $popmap -o $out_dir1 \

        -T 4 -b 1 -M $M -n $n1 -m $m -S &> $log_file1

        denovo_map.pl --samples $reads_dir \

        -O $popmap -o $out_dir2 \

        -T 4 -b 1 -M $M -n $n2 -m $m -S &> $log_file2

        denovo_map.pl --samples $reads_dir \

        -O $popmap -o $out_dir3 \

        -T 4 -b 1 -M $M -n $n3 -m $m -S &> $log_file3

        ((M+=1))

    done

    M=1

    ((m+=1))

done
```

Figure A.2, the script used to cycle through parameter combinations to determine the ideal parameter combination for Stacks. Results were extracted manually.

```
#A script to map the processed genetic data into stacks, identify loci and compare different samples, using Stacks2.

#Using parameters optimized from the data provided by the Stacks2 parameter test script.

#Written by Lewis Edwards, 27/01/2020


module load dbl/stacks/2.2


denovo_map.pl --samples /ddn/data/jqfs17/Second_Time/Cleaned/Stacks2 \
-O /ddn/data/jqfs17/Second_Time/Info/PopMap1.tsv -o
/ddn/data/jqfs17/Second_Time/Stacks2Plink/PopMap1  \

-T 12 -M 5 -n 6 -m 4 -X "populations: --plink" &>
/ddn/data/jqfs17/Second_Time/Stacks2Plink/PopMap1/denovo_map.oe


denovo_map.pl --samples /ddn/data/jqfs17/Second_Time/Cleaned/Stacks2 \

-O /ddn/data/jqfs17/Second_Time/Info/PopMap2.tsv -o
/ddn/data/jqfs17/Second_Time/Stacks2Plink/PopMap2  \

-T 12 -M 5 -n 6 -m 4 -X "populations: --plink" &>
/ddn/data/jqfs17/Second_Time/Stacks2Plink/PopMap2/denovo_map.oe
```

Figure A.3, the final Stacks mapping script to create the two mapping sets used in the final analysis.

```
#A script to run the privacy rarefaction algorithm on the files from Stacks 2.

#Written by Lewis Edwards, 01/02/20.


module load dbl/samtools/1.2


module load python/2.7.9


module load r/gcc/current


python /ddn/data/jqfs17/Second_Time/privacy-rarefaction-master/privacy-rarefaction.v2.3.py --
bam_dir /ddn/data/jqfs17/Second_Time/Stacks2/PopMap1/

--bam_suffix .matches.bam --sex_list /ddn/data/jqfs17/Second_Time/Info/Style.txt --CPUs 12 --o
heterostyly


Rscript /ddn/data/jqfs17/Second_Time/privacy-rarefaction-master/plot_privacy-
rarefaction_curves.R

/ddn/data/jqfs17/Second_Time/Stacks2/PopMap1/permutation_results.heterostyly.txt
```

Figure A.4, the script to run the privacy rarefaction algorithm (including generating a graph of the results) on the first mapping set

```
#A script to perform some basic association tests on my data using Plink.

#Written by Lewis Edwards, 01/02/20


cd /ddn/data/jqfs17/Second_Time/Plink1

/ddn/data/jqfs17/Second_Time/plink-1.07-x86_64/plink --noweb --file
/ddn/data/jqfs17/Second_Time/Stacks2Plink/PopMap1/populations.plink --fisher --allow-no-sex

cd /ddn/data/jqfs17/Second_Time/Plink2

/ddn/data/jqfs17/Second_Time/plink-1.07-x86_64/plink --noweb --file
/ddn/data/jqfs17/Second_Time/Stacks2Plink/PopMap2/populations.plink --fisher --allow-no-sex

cd /ddn/data/jqfs17/Second_Time/Plink3

/ddn/data/jqfs17/Second_Time/plink-1.07-x86_64/plink --noweb --file
/ddn/data/jqfs17/Second_Time/Stacks2Plink/PopMap1/populations.plink --model --fisher --
allow-no-sex

cd /ddn/data/jqfs17/Second_Time/Plink4

/ddn/data/jqfs17/Second_Time/plink-1.07-x86_64/plink --noweb --file
/ddn/data/jqfs17/Second_Time/Stacks2Plink/PopMap2/populations.plink --model --fisher --
allow-no-sex
```

Figure A.5, the script to perform the Plink tests on the mapped data obtained from Stacks.

```
awk '$9<= 0.05 || NR==1' plink.assoc > plink.sig

awk 'gsub(/-/, "&") < 24' /ddn/data/jqfs17/Second_Time/Plink1/plink.sig > P1sig2.tsv
```

Figure A.6, an example shell script to extract the loci from Plink that are significantly associated with one morph (p<0.05) and that are present in at least 80% of the total samples

```
awk '{

    count=0

    ncount=0
```

```
for (i=3;i<=5;i++) {
    if ($i == "-") {
        ncount++
    }
}
for (i=6;i<=14;i++) {
    if ($i == "-") {
        count++
    }
}
for (i=15;i<=17;i++) {
    if ($i == "-") {
        ncount++
    }
}
for (i=18;i<=18;i++) {
    if ($i == "-") {
        count++
    }
}
for (i=19;i<=19;i++) {
    if ($i == "-") {
        ncount++
    }
}
for (i=20;i<=20;i++) {
    if ($i == "-") {
        count++
    }
}
for (i=21;i<=21;i++) {
    if ($i == "-") {
     ncount++
    }
}
for (i=22;i<=24;i++) {
    if ($i == "-") {
        count++
    }
}
for (i=25;i<=30;i++) {
    if ($i == "-") {
        ncount++
    }
}
for (i=31;i<=31;i++) {
    if ($i == "-") {
        count++
    }
}
for (i=32;i<=32;i++) {
    if ($i == "-") {
        ncount++
    }
}
if ((count <= 3)&&(ncount>=12)) {
```

```
          print
        }
     }' /ddn/data/jqfs17/Second_Time/Plink1SNP2.tsv \
> P1SsigThrum.tsv
```

Figure A.7, an example of one of the shell scripts used to isolate the loci present in at least 80% of the samples of one morph and absent in at least 80% of the samples of the other morph (this particular script was isolating loci present in at least 80% of thrum morph samples in the first mapping set)