Small RNA and genome interactions in *Chlamydomonas reinhardtii* recombinants

Daisy Hessenberger

Downing College

Supervisor: Professor Sir David C. Baulcombe

Department of Plant Sciences

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

September 2014

Declaration

This dissertation is the result of my own work and includes no material that is the outcome of work done in collaboration except where specifically indicated in the text. It has not been previously submitted, in part or whole, to any university or institution for any degree, diploma, or other qualification.

In accordance with the Degree Committee for Biology guidelines, this dissertation does not exceed 60,000 words.

Summary

When conspecific individuals are crossed, the ensuing hybridization creates a spectrum of phenotypes in the resulting offspring. Many of hybrid traits will be additive, similar to the parental phenotypes. In some cases however, transgressive phenotypes are formed, outside the range of that of the parental phenotypes.

Transgressive phenotypes can either be restricted to the F1 generation or be heritable throughout the hybrid lineage. While the mechanism behind heritable transgressive phenotyped is yet to be determined, transgressive gene expression is thought to be the root cause of their formation. Epigenetics modifications, heritable variation separate to the DNA code, can alter gene expression, persist through generations, and vary between individuals and over time. This makes them ideal candidates to be involved in the formation of transgressive phenotypes

RNA silencing is an epigenetic mechanism of gene regulation relying on 20-24nt single stranded small RNAs (sRNAs). Small RNAs, due to their ability to set up persistent epigenetic marks at a locus, have the potential to create heritable transgressive gene expression. For example, when genetic variation from one parental genome presents novel targets to the sRNAs of the other parental genome, new epigenetic marks such as DNA methylation or secondary sRNAs can be created at target sites.

In order to understand the potential of small RNAs to influence hybrid phenotype, I designed crossing experiments with *Chlamydomonas reinhardtii*, choosing this unicellular alga due to the genetic tools available and the haploid nature of its vegetative cells. The specific aim of the experiment was to identify transgressively expressed sRNA populations. Crossing two geographically distinct strains of *C. reinhardtii,* and sequencing both the genomes and sRNAomes of parents and recombinants, I was able catalogue both genetic and epigenetic variation in the parental strains providing unique insight into the inheritance of small RNAs in this alga.

In this thesis, I first compare the genomes of the parental strains, identifying polymorphisms and assessing genetic variation in RNA silencing pathway components. I then describe the sRNA profiles of the parental strains, identifying differentially expressed sRNA loci. I then describe my approach to identifying transgressively

expressed sRNA loci in the hybrids. While many sRNA loci in the recombinants exhibit additive sRNA expression, I found multiple transgressively expressed sRNA loci.

Using the available bioinformatics tools, I identified potential miRNAs and phased secondary sRNAs within the list of transgressively expressed loci. Target analysis of one of the transgressively expressed miRNAs linked it with the transgressive expression of certain phased loci, suggesting a potential for sRNAs to be able to set up heritable epigenetic marks in recombinant *C. reinhardtii* cells.

Acknowledgements

Thank you especially to my supervisor, Professor Sir David Baulcombe for his continued advice and support throughout this process. His lessons were applicable both at work and in life. I would also like to thank him for giving me this opportunity, not just to do a PhD, but also to be part of the Baulcombe lab for four years where I got the chance to work with many outstanding researchers.

I would like to thank all of those who helped me with my research including all my colleagues in the Baulcombe and Henderson labs. Specifically thank you to the members of the bioinformatics group who advised me on bioinformatics matters and who I called on when in programming trouble. Also thank you to Chlamy group members, both past and present, who taught me how to work with this delightful alga and how to be a good scientist. Of special mention is Dr Adrian Valli who as a mentor and friend taught me what good science means.

Over the four years I also collaborated with many excellent researchers outside of the Department of Plant Sciences. Thank you to Dr Saul Purton for initially showing me how to mate *C. reinhardtii*, to Dr Sinead Collins and Heidi Kuehne for helping me optimise the mating, John Davey for advice on genome assembly, and my second supervisor Dr Chris Jiggins. Also thank you to James Hadfield and the Cancer Research Institute UK for sequencing my DNA and sRNA libraries.

A lot of the people I had the pleasure of working with not only aided me with science but also with companionship. I would like to thank Dr Donna Bond and Dr Madeline Mitchell for reading parts of my dissertation and the chats over tea and coffee respectively. Thank you too to Dr Jake Harris for the lunchtime science/non-science conversations and for setting a brilliant example for a PhD student.

Thank you to Downing College for providing an excellent social backdrop to my studies and of cause for funding my PhD project via the Lewin-Fritsch Studentship. Downing College not only provided me with the funds to undertake this research but also with the chance to get involved with the graduate community. Through Downing college I also met the many friends who have made the PhD process a joy to work through: Dr Michiel Kamp, Dhiren Mistry, Dr Pablo Aran-Terol, Dr Teresa

Segura Garcia, Jacqueline Ward, Joanna Harrison, Jennifer Stuart, Dr Fezile Lakadamyalı, Krit Sitathani, and Dr Matthew Cooper. Like collaborators I must also thank all of my friends outside of the University.

I would like to thank the Department of Plant Sciences, which first ignited my interest in plant sciences and then for the next 6 years continued to feed my curiosity.

Finally I would like to thank my family who have supported me and inspired me in so many ways. Thank you to my mother for always encouraging me, my father for always questioning me, my brother for always challenging me, my Opa for always believing in me, and my Oma for giving me those first Gerald Durrell books. Without them I would not be who I am today.

“Keep it simple.”

- Regius Professor Sir David Baulcombe

Dedicated to my Opa.

Siegfried Herbert Gruber.

Common abbreviations

3’/5’-UTR	Three/five prime untranslated region
AGO	Argonaute
DCL	Dicer-like
dsDNA/RNA	Double stranded DNA/RNA
EDTA	Ethylenediaminetetraacetic acid
EST	Expressed sequence tag
FDE	Formamide/EDTA (buffer)
HEN1	HUA Enhancer 1
InDel	Insertion and/or deletion
IR	Inverted repeat
mRNA	Messenger RNA
miRNA	Micro RNA
nt	Nucleotide
PCR	Polymerase chain reaction
Pol	Polymerase
PTGS	Posttranscriptional gene silencing
QTL	Quantitative trait locus
RdDM	RNA directed DNA methylation
RDR	RNA dependent RNA polymerase
RISC	RNA induced silencing complex
sRNA	Small RNA species
siRNA	Small interfering RNA
ssDNA/RNA	Single stranded DNA/RNA
TBE	Tris/Borate/EDTA (buffer)

Table of Contents

1. Chapter One: General introduction

Genome interactions in hybrids can influence a hybrid specific transcriptome signature resulting in transgressive phenotypes. In this introduction I review the models for transgressive phenotypes, describe how RNA silencing may play a role, and introduce a model system, the unicellular green algae Chlamydomonas reinhardtii, in which to test the mode of sRNA inheritance.

1.1. The wider context

"It is interesting to contemplate a tangled bank, clothed with many plants of many kinds, with birds singing on the bushes, with various insect flitting about, and with worms crawling through the damp earth, and to reflect that these elaborately constructed forms, so different from each other, and dependent upon each other in so complex a manner, have all been produced by laws acting around us."

— Charles Darwin, "On the Origin of Species", 1856

For centuries the complexity of the natural world has captivated researchers. Even today, there is no adequate explanation for the extent of variation that Darwin observed in a tangled bank. And since, as *Homo sapiens*, we are both a result of this progression of complexity and expert exploiters of complexity observed in the natural world, it is to our benefit to keep working towards resolving the great mystery of evolution, why our world is so diverse. In order to answer this question, how variation is created and inherited must first be understood. The origin of phenotypic and genetic novelty is a subject of paramount importance to evolutionary biologists.

The non-linearity of hybrids is an excellent experimentally analysable system in which to test how heritable variation in phenotype is first created. Hybridization, while creating an individual who is, all in all, similar to its parents, also creates a plethora of novel variation. Most hybridization events result in an offspring with

a mix of both parental phenotypes. On occasion however, hybrids exhibit extreme phenotypes in comparison to the parents. Termed transgressive phenotypes, these traits, when quantifiable, lie outside of the parental phenotypic range (Figure 1-1).

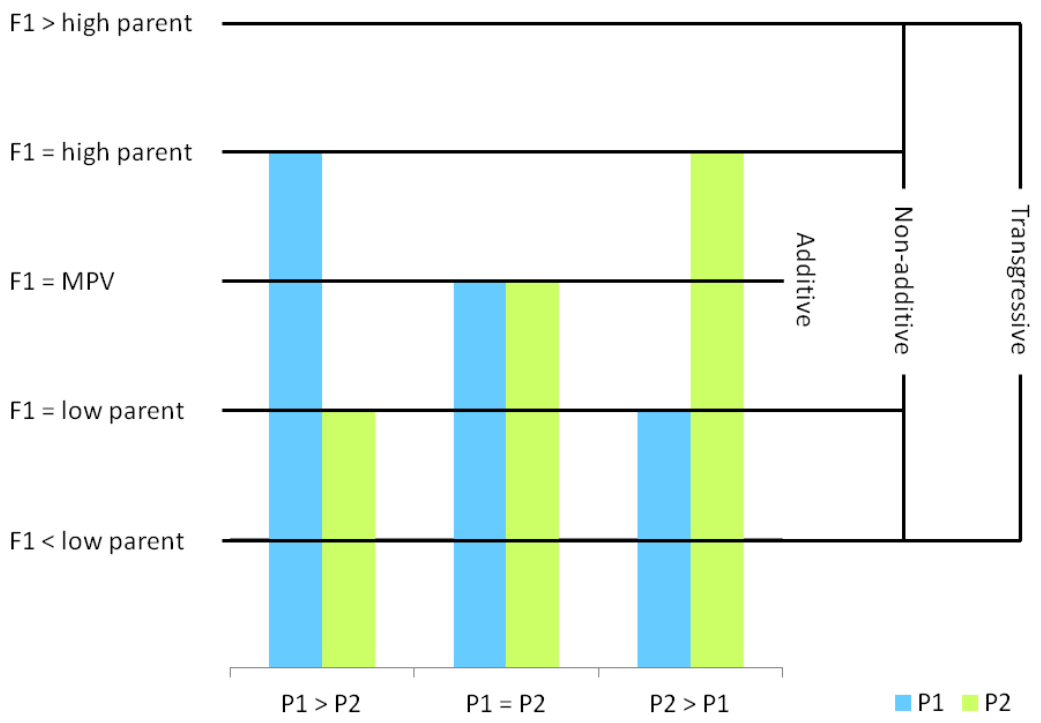


Figure 1-1: Phenotypic outcomes of hybridization
After hybridization, there are various phenotypic outcomes in terms of a quantifiable phenotype in the hybrid. Regardless whether the parental phenotypes are similar (P1 = P2) or different (P1 > P2 or P2 > P1), the hybrid phenotype can be termed additive (when similar to the mid parental value) or non-additive (when different to the mid parental value). Transgressive phenotypes are those that are different to the mid parental value and outside the range of the parents (either above high-parent or below low-parent). The distinction between non-additive and transgressive phenotypes is important, as they have sometimes been incorrectly used in past studies. P1: Parent one. P2: Parent two. F1: Hybrid (first generation).

1.1.1. Transgressive phenotypes in hybrids

Transgressive phenotypes can either hinder hybrids, known as hybrid necrosis, or benefit hybrids. Indeed, humans have benefitted from transgressive phenotypes for millennia, since the dawn of agriculture (Bennett and Ali, 2010). To reap

further increases in yield and stability of crop production, novel instances of transgressive phenotypes must be identified or created (Birchler et al., 2010).

In beneficial transgressive phenotypes, the hybrid offspring's superior vigour is usually linked to greater or more consistent yield than either parent. Yield is characterized by its relation to agriculture and includes increased biomass, size, speed of development, disease resistance, stress tolerance, and higher reproductive success (Hochholdinger and Hoecker, 2007). Hybrid breeding to obtain transgressive phenotypes has been applied to crops such as maize, rice, sunflower, rapeseed, sugar beet, and tomato as well as to livestock, including cattle, poultry, swine and sheep (Meyer et al., 2012). These examples of transgressive phenotypes result in real world benefits to farms. For example, hybridization in tomato can create a transgressive phenotype resulting in 60% increase in yield (Figure 1-2) (Krieger et al., 2010).

Transgressive phenotypes are not just constrained to domestic plant breeding; they are readily observable in nature and in model systems in the lab. A review of transgressive phenotypes suggests that they are a common occurrence in all crosses (Rieseberg et al., 1999). In the model higher plant, *Arabidopsis*, alone, transgressive phenotypes can manifest in terms of photosynthetic efficiency (Sharma et al., 1979), seedling viability (Mitchell- Olds, 1995), seed number (Alonso-Blanco et al., 1999), phosphate efficiency (Narang and Altmann, 2001), biomass (Barth et al., 2003; Meyer et al., 2004), freezing tolerance (Rohde et al., 2004), seed size (Stokes et al., 2007), flowering time (Perera et al., 2008), metabolite contents (Lisec et al., 2009) and leaf area (Meyer et al., 2010).

Figure 1-2: Transgressive phenotype in tomato.
Figure taken from (Krieger et al., 2010). Upon crossing M82 and sft-4537 lines of tomato, the resulting hybrids have a 60% higher level of tomato yield. This transgressive phenotype is due to a single overdominant locus, Single Flower Truss.

It is important to remember that gene expression in itself is a phenotype and similarly to physiological phenotypes, while hybrid gene expression is usually within the range of the parents (whether additive or non-additive), in certain instances genes are transgressively expressed (Birchler et al., 2010; Li et al., 2009). However the proportion of transgressively expressed genes in hybrids varies between different crosses and there is little consensus between genes which are transgressively expressed (Birchler et al., 2010; Li et al., 2009). There is a broad correlation between the amount of transgressive gene expression with the genetic divergence of the parents (Birchler and Veitia, 2010). Transgressive gene expression is also correlated with the magnitude of transgressive physiological phenotypes but causation is still to be confirmed (Li et al., 2009; Riddle et al., 2010).

1.1.2. Classic genetic models for transgressive phenotypes

Beneficial transgressive phenotypes, which cause hybrids to be more vigorous than their parents, have been well documented and studied in both animals and plants (Bucheton et al., 1984; Fernández-Silva et al., 2009; Krieger et al., 2010) and yet the cause for transgressive phenotypes, whether physiological or genetic, is still to be deciphered. Three models have been suggested as the main mechanisms for transgressive phenotype formation: overdominance, dominance, and epistasis.

1.1.2.1. Overdominance

In the overdominance model, the transgressive phenotype is due to a single locus where the heterozygous genotype shows superior fitness to the homozygous genotype due to intralocus allelic interactions (Li et al., 2008). There are very few examples of true overdominant loci. Recently a single locus in tomato was discovered that has the potential of driving transgressive phenotypes increasing yield (Figure 1-2). It seems that overdominance does play a role in the formation of transgressive phenotypes though to what degree is not known (Krieger et al., 2010).

1.1.2.2. Dominance

In contrast to overdominance, the dominance model relies on the interactions of alleles at multiple loci. Dominance is the idea that transgressive phenotypes are due to complementation of deleterious or inferior recessive alleles inherited from one parent by the inheritance of beneficial or superior dominant alleles from the other parent in the heterozygous genotype of the hybrid (Li et al., 2008). There are well-supported examples supporting the dominance theory, for example transgressive fruit size in hybrid melons (Fernández-Silva et al., 2009).

1.1.2.3. Epistasis

Similarly to the dominance model, in epistasis multiple loci are involved in the formation of transgressive phenotypes, however in this model it is the inter-loci interactions rather than the allelic interactions that are the cause. The theory

behind epistasis suggests that hybrid traits are affected by the interaction between multiple non-allelic genes in the genome and that any single gene replacement in a hybrid can have complex effects on many characters (Hochholdinger and Hoecker, 2007). Essentially the genetic background of an individual must be taken into account when looking at a single gene and single/multiple traits (Yu et al., 1997). Quantitative trait loci (QTL) studies have implicated epistasis in the formation of non-additive phenotypes (Stelkens et al., 2009). In rice and corn, epistasis was found to be involved in most of the QTLs involved with transgressive phenotypes (Chen, 2007).

1.1.3. Novel models needed to explain transgressive phenotypes

The cause for transgressive phenotypes is likely pluralistic, involving contributions from all of the models described above. While experimental evidence exists for overdominance, dominance, and genetic epistasis, these models are largely conceptual and are insufficient to describe the molecular cause for transgressive phenotypes (Birchler et al., 2003). Despite advances in this area of research, the basis of transgressive phenotypes remains enigmatic and it is becoming evident that a new approach to modelling hybridization is needed.

For example, to elucidate the molecular mechanism for transgressive phenotypes, a systems-based approach should be taken, profiling all the factors involved in phenotype formation in the parents, hybrids, and further generations. Past studies have concentrated on transgressive phenotypes in just the F1 generation and there is little evidence of the inheritance of different types of transgressive phenotypes. An overdominant locus causing transgressive phenotypes would be lost in further lineages while models based on complementation either of alleles (dominance) or gene interactions (genetic epistasis) would also result in transgressive phenotypes being lost in further hybrid lineages due to the unlinking of genetic loci (Birchler et al., 2010). The frequency of inheritance or loss of a transgressive phenotype depends on the nature of the interactions in the hybrid. The few studies of heritability of transgressive phenotypes have indicated a mechanism for heritable transgressive

phenotypes distinct to complementation-based mechanisms resulting in F1 transgressive phenotypes (Rieseberg et al., 1999).

It is generally accepted that transgressive gene expression in hybrids contributes to transgressive phenotypes in plants (Birchler et al., 2003; He et al., 2013a; Song and Messing, 2003) and is therefore likely to be at the centre of the mechanism of transgressive phenotype formation. Epigenetic interactions in hybrids have the potential to result in transgressive gene expression and thus should also be included in future studies of transgressive phenotypes. Especially since the unique inheritance of epigenetic factors in comparison to genetic factors must be understood for modelling the heritability of transgressive phenotypes.

1.2. Epigenetics and transgressive phenotypes

Epigenetics is defined as the study of the heritable change in phenotype attributed to mechanisms other than changes in the underlying DNA sequence. Epigenetic modifications capable of changing gene expression include histone modifications, DNA methylation, chromatin structure, RNA silencing, chromosome pairing, and spatial location of DNA (Grant-Downton and Dickinson, 2004; Henderson and Jacobsen, 2007). Global patterns of some of these modifications have been shown to differ significantly between hybrids and their parents (He et al., 2010; Zhao et al., 2007b) supporting the idea that epigenetic and genetic systems operate in parallel to influence heritable variation.

The unique aspect to epigenetic variation, besides the increased level of complexity in comparison to DNA variation, is the potential for instability over time. While a mutation in the DNA code, whether it is a deletion or single nucleotide polymorphism (SNP), is generally irreversible, epigenetic modifications are more labile than genetic modifications and can even change within a cell's lifetime. Epialleles (where alleles at a locus are genetically identical but whose epigenetic modifications differ) have been shown to spontaneously revert in tomato (Quadrana et al., 2014) and metastable epialleles have been identified in mammals (Rakyan et al., 2002).

To understand the role of epigenetics in the creation of heritable transgressive phenotypes, we would need to understand how and when epigenetic variation is created in hybrids. In addition to taking us a step closer to being able to fully exploit transgressive phenotypes, this understanding would allow epigenetic inheritance to be added to evolutionary models. Epigenetic variation is another variable, on which natural selection can act and which can result in populations adapting faster to environmental pressures as adaptive phenotypes can arise before any genetic changes and be inherited.

1.2.1. Epigenetic marks can alter gene expression in hybrids

Epigenetics is thought to play a large role in the formation of transgressive phenotypes (Groszmann et al., 2013) as epigenetic modifications can radically alter gene expression (Ni et al., 2009). Indeed in *Arabidopsis*, RNA silencing in hybrids was connected to changes in gene expression, mediated through a reduction in DNA methylation (Groszmann et al., 2011) and a transgressive phenotype in this higher plant, in the form of increased biomass, has been linked to epigenetic modifications of circadian clock genes (Ni et al., 2009). Also recent studies suggest that epigenetic effects such as methylation, histone modifications, and RNA silencing may influence hybridisation and thus be implicated in the formation of transgressive phenotypes (Ha et al., 2009).

1.2.2. Different types of epigenetic marks are likely involved in formation of transgressive phenotypes

DNA methylation, when altered in hybrids, can contribute to phenotypic variation (Borges and Martienssen, 2013). In DNA methylation, a methyl group (CH_3) is added to the 5-carbon of the cytosine DNA nucleotide creating a 5-methylcytosine. DNA methylation is often associated with gene inactivation and can be found in three different contexts: CG, CH, and CHH where H can be C, A, or T.

In crop plants such as maize and rice, differential DNA methylation patterns occur in the intraspecific hybrids (Jin et al., 2008; Zhao et al., 2007b). A lower level of differential methylation was found in *Arabidopsis* hybrids, with

transgressive methylation patterns more likely to be found at loci where the methylation was differentially expressed in the parents (Greaves et al., 2012). The increased transgressive methylation patterns in maize and rice are likely due to more differential methylation between the parents in those crosses. There are also global affects on methylation in hybrids, for example the general though subtle increase in cytosine methylation in *Arabidopsis* hybrids (Shen et al., 2012).

Histone modifications, another type of epigenetic mark, can influence gene expression and chromatin structure (Heard and Martienssen, 2014). DNA is packages into structural units called nucleosomes made up of various histone proteins. The multiple classes of histones and the different modifications to histone amino acids possible (methylation, acetylation, and phosphorylation) mean that the histone code has the capability to be extremely complex. Various histone footprints are associated with gene transcription, gene repression, chromatin structure, and the laying down of other epigenetic marks (Heard and Martienssen, 2014). Different histone modifications may have distinct inheritance patterns (Groszmann et al., 2013). A few examples of non-additive histone modifications in hybrids have been linked however to gene expression changes (He et al., 2010; Li et al., 2011; Ni et al., 2009).

DNA methylation and histone modifications have been more extensively studied than other epigenetic pathways such as RNA silencing in respect to transgressive phenotypes. RNA silencing, while its role in transgressive phenotype formation is less well studied, has conclusively been shown to interact with DNA methylation and histone modifications (Matzke and Mosher, 2014). RNA silencing's ability to alter epigenetic modifications makes it a good candidate to study in relation to transgressive phenotypes.

1.3. RNA silencing

RNA silencing is a paneukaryotic mechanism by which gene expression can be regulated through the action of small RNAs (sRNA) (Molnar et al., 2007a). sRNAs are 20-24 nucleotide (nt) non-coding single-stranded RNA (ssRNA) which guide an RNA Induced Silencing Complex (RISC) to target RNA through complementary

base-pairing and then, through a variety of mechanisms, alter the expression of the target at the transcriptional or post-transcriptional level.

RNA silencing relies on the action of a core set of proteins (Figure 1-3-A). Usually sRNAs are created via RNAseIII-mediated cleavage of double-stranded RNA (dsRNA) by a Dicer-Like (DCL) protein and their mode of action depends upon being incorporated into a specific Argonaute (AGO) protein, itself a part of the RISC. The components of the RNA silencing system are significantly conserved in eukaryotes although sequence similarity of sRNAs is much more varied (Shavalina and Koonin, 2008). There are multiple DCL and AGO proteins in different organisms which are specific to different RNA silencing pathways (Chapman and Carrington, 2007; Molnar et al., 2011).

Small RNAs regulate gene expression via a multitude of mechanisms including transcriptional gene silencing (TGS) and post-transcriptional gene silencing (PTGS). Targeting by an sRNA can result in messenger RNA (mRNA) cleavage, translational repression, and the setting down of epigenetic marks at a locus. The specificity of the gene regulation technique to the sRNA type lies partially in the combination of core RNA silencing proteins used. It is not yet fully understand what determines which AGO an sRNA is fed into but the size and modifications (such as the 2'-O-methylation of the 3' end of the sRNA duplexes by HEN1 (Yu et al., 2005)) of the sRNA are likely to be implicated. After the biogenesis of an sRNA, the mode of action can also be determined by target complementarity. For example, high levels of complementarity between a type of sRNA known as micro RNAs (miRNA) and their targets usually results in mRNA cleavage (Axtell, 2008).

There are many different types of sRNAs involved in RNA silencing such as miRNAs and small interfering RNAs (siRNA), all of which are defined by their unique modes of biogenesis and action (Chapman and Carrington, 2007) (Figure 1-3). Some species of sRNAs such as viral sRNAs and piRNAs will not be covered in this review as they are not present in *C. reinhardtii*.

1.3.1. Micro RNAs (miRNA)

An abundant class of 20-22nt long sRNAs, miRNAs originate from endogenous genes to regulate the expression of mRNAs with complementary target sites (Figure 1-3-B). The conservation of the miRNA silencing mechanism between animals, higher plants, and unicellular organisms such as *C. reinhardtii* indicates that miRNAs predated the evolution of multicellularity (Molnar et al., 2007b), even suggesting that miRNAs helped drive the evolution of multicellularity (Bartel et al., 2003).

Micro RNAs are defined by their biogenesis: mature miRNAs are processed from hairpin precursors, partially double-stranded regions of fold back transcripts, into 20-22nt small dsRNA duplexes by a DCL protein (Figure 1-3-B). Once excised the miRNA (the strand with the weakest 5'-end base-pairing) is integrated into an RISC containing an AGO protein while the opposite strand, the miRNA*, is degraded. The miRNA then targets the RISC complex to a mRNA through Watson-Crick base-pairing. Although miRNAs often regulate gene expression through mRNA cleavage certain species of miRNAs also have the capability of setting up the production of secondary siRNAs and creating epigenetic marks at a target locus (Manavella et al., 2012).

While defined by their biogenesis, miRNAs are identified using their precursors, the basic structure of which is relatively conserved between genera. Fold back hairpin RNAs can be predicted from the genome and, using sRNA sequencing data, bioinformatics tools can predict potential novel miRNAs using the patterns of alignment. The precursors are usually produced by RNA polymerase II but on occasion RNA polymerase III performs the same functions (Hutvagner and Simard, 2008)**.** Hairpin precursors of miRNAs are usually less than 150nt in length and produce just a single specific miRNA/miRNA* combination.

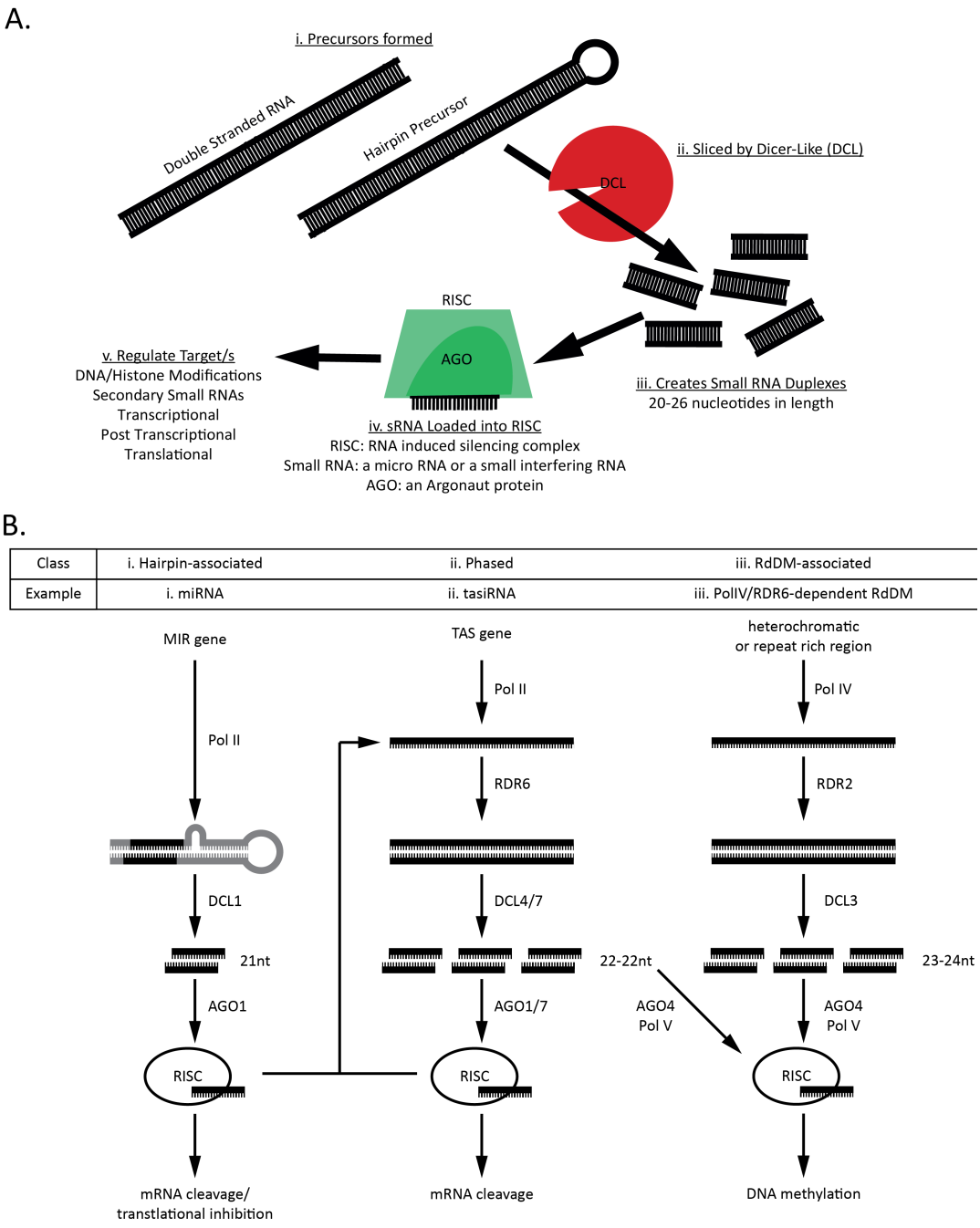


Figure 1-3: Summary of RNA silencing

(A) Basic overview of RNA silencing pathways. sRNAs originate from (i) dsRNA. A (ii) DCL enzyme processes the dsRNA into (iii) 21-24nt dsRNA duplexes. The sRNA star (sRNA*) is usually degraded while the true sRNA (iv) associates with the AGO protein in the RISC in order to (v) target gene regulatory actions. **(B)** Schematic diagram of three types of plant sRNAs (i) hairpin-associated, (ii) phased, (iii) RdDM-associated siRNA pathways. Exemplified by three plant RNA silencing pathways: (i) miRNA, (ii) trans-acting siRNA, (iii) PolIV/RDR6-dependent RdDM siRNA pathways. The RNA silencing pathways have an endogenous origin, transcribed by a DNA-dependent RNA polymerase (either Pol II or Pol IV). In the case of (i) miRNAs, the transcript folds into a hairpin precursor, which creates a region of dsRNA. In the other pathways (ii and iii), an RDR processes transcripts into dsRNA. Different pathways use different DCL proteins although there is some redundancy between the pathways. Similarly different AGO proteins are specific to various RNA silencing pathways. The three pathways result in various forms of gene regulation.

1.3.2. Hairpin-derived siRNAs

Longer hairpin precursors with almost perfect pairing, and resulting in multiple species of sRNAs, are also present in various eukaryotic genomes. These sRNAs originate from a single-stranded hairpin precursor but do not meet the criteria for annotation as a miRNA and are known as hairpin-derived siRNAs.

Most of these longer hairpin precursors are the result of the transcription of inverted repeats (IR). Two IRs identified in *Arabidopsis* produce hairpin-derived siRNAs ranging in size from 21-24nt. Seemingly dependent on DCL2/3/4, these hairpin-derived siRNAs are mobile between tissues of a plant (Melnyk et al., 2011), show a limited level of conservation between *Arabidopsis thaliana* ecotypes (Dunoyer et al., 2010), and can regulate gene expression (Melnyk et al., 2011).

Despite that, the function of hairpin-derived siRNAs is not well understood. Artificial inverted repeats containing sequences from a gene produce sRNAs that then down regulate that gene. Some inverted hairpin-derived siRNAs, similar to some miRNAs, can result in epigenetic modifications (Heard and Martienssen, 2014). For example in the MuDR system in higher plants, inverted repeats can drive transgenerational silencing of transposons via DNA methylation (Heard and Martienssen, 2014).

1.3.3. Other small Interfering sRNAs (siRNA)

Rather than deriving from a hairpin precursor, many sRNAs derive from dsRNA formed through other processes. These 21-24nt siRNAs are able to mediate epigenetic modifications of the genome. Typically siRNA loci are not well conserved between lineages but the pathways for their epigenetic actions have been observed in plants, fungi, and metazoans (Castel and Martienssen, 2013). They are usually derived from intergenic and/or repetitive genomic regions associated with the *de novo* deposition of epigenetic modifications (Bond and Baulcombe, 2014). The link between siRNAs and transposable elements is consistent with a primary role of RNA silencing in genome defence (Holeski et al., 2012) but they are also known to be involved in stress responses, pathogen

response, genomic imprinting, intercellular and inter-allelic communication, and in genome interactions in hybrids (Matzke and Mosher, 2014).

There are multiple classes of sRNAs, defined by their biogenesis, action, and pathway components, including siRNAs associated with RNA-dependent DNA methylation (RdDM), phased secondary siRNAs, viral derived siRNAs, and piRNAs (specific to animals). In this review I will cover siRNAs associated with RdDM, and phased secondary siRNAs. Other pathways such as the piwi-interacting RNA pathway or viral derived siRNA pathway will not be reviewed, as these pathways are not present in the model organism used in this study.

1.3.3.1. siRNAs associated with RNA-dependent DNA methylation (RdDM)

One epigenetic action of siRNAs is RdDM (Figure 1-3). There are multiple RdDM pathways, all involving different DNA-dependent RNA Polymerases (Pol) and RNA-dependent RNA polymerases (RDR) in which siRNAs can alter DNA methylation. These pathways are differentiated by the mechanisms involved in establishment and maintenance of these epigenetic marks (how the methylation is copied from one genome to another during mitosis and meiosis). The best-established mechanism of RdDM, the PolIV-dependent pathway, results in DNA methylation that can be maintained transgenerationally through sRNA-dependent and sRNA-independent self-sustaining loops. The DNA methylation footprint of this pathway includes all types of sequence context (CG, CH, and CHH) (Bond and Baulcombe, 2014).

The establishment of PolIV-dependent RdDM in higher plants has been the focus of much research recently (Bond and Baulcombe, 2014). It requires recruitment of Pol IV to the target locus so that a non-coding ssRNA is transcribed from the target locus. An RDR converts this Pol IV product into dsRNA, which is then cleaved by the action of DCL3 into 24nt sRNA duplexes. The siRNAs are loaded into an AGO protein. The option of variety of AGO proteins (AGO4/6/9) ensures that there is another level of targeting specificity. For example, AGO4-loaded siRNAs base pair specifically with Pol V scaffold RNAs. This interaction results in

the recruitment of a chromatin-remodelling complex termed DDR**,** including *de novo* DNA methyltransferase (DRM2) catalysing *de novo* DNA methylation at the target locus. This methylation can then be maintained in a siRNA-dependent or siRNA-independent manner (Bond and Baulcombe, 2014).

Although this pathway is well understood, we do not understand why it is recruited to certain targets in the genome and not others. PolIV associates with various proteins including CLASSY1, RDR2, and SAWADEE HOMEODOMAIN HOMOLOG 1 (SHH1), a protein which recognises chromatin with certain histone modifications (Zhang et al., 2013). The implication of histone modifications in the targeting of PolIV-dependent RdDM by association also implicates RNA silencing, as sRNAs have also been shown to have the ability to modify histone modifications.

1.3.3.2. Phased siRNAs

Some siRNA loci exhibit a distinct phasing pattern in which the individual siRNA species are generated precisely in a head to tail arrangement starting from a specific nucleotide. Several RNA silencing pathways converge to produce phased siRNAs and, currently, phasing is indicative of secondary siRNA production. For secondary siRNA production, a primary sRNA (usually a miRNA) targets dsRNA created by an RDR for successive DCL cleavage of dsRNA from a single starting point. There are DCL-independent secondary siRNAs that are phased in *C. elegans* (Sijen et al., 2007) but in plants phased secondary siRNAs are DCL-dependent.

Phased siRNAs are exemplified by the category of *trans*-acting siRNAs (tasiRNA) (Figure 1-3-B). Three TAS genes have been identified in *Arabidopsis*, the transcripts of which are targeted by miRNAs producing tasiRNAs at those loci. The initiator miRNA targets Pol II transcripts so that they are copied by RDR6, producing dsRNA (Matzke and Mosher, 2014). This dsRNA is the template for the production of phased 21-22nt secondary siRNAs by DCL2/4, which are then loaded into AGO1 for post-transcriptional gene silencing of other transposon mRNAs. There is another version of the tasiRNA pathway in which a TAS

transcript is targeted by two initiator miRNAs. In this two-hit tasiRNA pathway the tasiRNAs are instead loaded into AGO7. The tasiRNAs then act in *trans*, usually regulating endogenous mRNAs via mRNA cleavage. However some tasiRNAs act in *trans* to set up further secondary siRNA loci such as tasiR2140 in *Arabidopsis* (Chen et al., 2010b). Thus the initiator for phased secondary siRNAs can either be a miRNA or a secondary siRNA. Therefore the phased secondary siRNA loci have the ability to set up transgenerationally inherited transgressive gene expression at multiple loci. Similarly, siRNAs from the PolIV-dependent RdDM pathway can also result in a cascade of novel epigenetically regulated gene expression.

While tasiRNAs are not normally associated with RdDM, other phased siRNAs can lay down methylation through a PolII/RDR6-dependent RdDM pathway. These phased siRNAs tend to originate from protein coding genes. Phased siRNAs can also feed into the PolV-dependent part of the PolIV-dependent RdDM pathway, initiating low levels of *de novo* DNA methylation at the target locus (Bond and Baulcombe, 2014).

It is not fully understood what dictates whether a locus will be targeted for phased siRNA production as only a subset of sRNA-target interaction result in secondary siRNA biogenesis. There is some evidence that the 22nt size class of sRNAs is more competent to set up phased secondary sRNA (Chen et al., 2010b) however recent research has suggested that the secondary structure of the miRNA duplex is the primary determinant (Manavella et al., 2012). Transcripts are also more susceptible to this sort of RNA silencing if they have an aberrant 5'/3' end, contain more than one sRNA target site, or are over expressed (Axtell, 2013) and the primary sRNA is more likely to trigger phasing if itself is over expressed or loaded into AGO7 (Fei et al., 2013).

1.3.4. Role of RNA silencing

In various multicellular organisms RNA silencing is involved in development, signalling, stress responses, and immune responses although the elucidation of the exact role and importance of sRNAs is an on-going study (Axtell et al., 2007;

Filipowicz, 2008; Herr and Baulcombe, 2004; Malone and Hannon, 2009; Ruiz-Ferrer and Voinnet, 2009). The original role of RNA silencing is thought to be in genome defence providing protection against viruses and transposons. A more recent interpretation of this role is that sRNAs act as transgenerationally transmitted RNA 'memories' insuring the integrity of the next generation by recoding parental gene expression patterns. In this scenario sRNAs have the ability to 'black list' invading nucleic acids and transposon fossils, while 'guest listing' endogenous genes (Rechavi, 2014). Another slightly contentious role proposed for miRNAs is the buffering of gene expression (Bao et al., 2014; Herranz and Cohen, 2010).

1.4. RNA silencing implicated in transgressive gene expression

Many investigations have shown that sRNAs exact an epigenetic effect and, more recently, that RNA silencing is implicated in the formation of transgressive phenotypes. For example, in polyploidy plants, sRNAs have been implicated in the formation of transgressive phenotypes (Chen, 2007). Since transgressive gene expression and dosage changes in regulatory nodes have been implicated in transgressive phenotypes and sRNAs regulate gene expression, RNA silencing is a good candidate for being implicated in the mechanism for the formation of transgressive phenotypes.

1.4.1. Hybridization can result in transgressive sRNA expression

Transgressive expression of sRNAs, like genes, is common after hybridization. Hybrid rice were shown to have a significantly different sRNAomes (both in composition and expression) when compared to that of the parents (He et al., 2010) and multiple sRNA loci are transgressively expressed in hybrid tomatoes (Shivaprasad et al., 2012). Of the transgressively expressed sRNA loci in tomato, one was linked to hypermethylation of the corresponding target DNA while in another example, transgressive expression of a miRNA was linked to enhanced salt stress tolerance in the tomato F2 generations (Shivaprasad et al., 2012).

Indeed transgressive expression of sRNAs has been linked to changes in gene expression in other plant hybrids (Chen et al., 2010a; Ding et al., 2012; Ha et al., 2009; He et al., 2010) and to biologically relevant transgressive phenotypes in higher plants such as maize and *Arabidopsis* (Xing et al., 2010; Groszmann et al., 2011). Thus transgressive expression of sRNAs could lead to transgressive expression of genes leading to beneficial biologically relevant hybrid traits. While most sRNA loci (both miRNA and siRNAs) exhibit additive expression patterns in hybrids (Barber et al., 2012), some general trends in the inheritance of sRNAs have been noted in higher plants.

A global down regulation of 24nt sRNAs has been noted in plant hybrids, especially in regions of the genome which are differentially expressed in the parents (Barber et al., 2012; He et al., 2013b; Vaughn et al., 2007). This association with transgressive hybrid expression and parental differential expression is similar to that observed in regards to methylation and histone modifications (Greaves et al., 2012). Furthermore this down regulation of 24nt siRNAs has been linked to changes in gene expression in *Arabidopsis* hybrids via the reduction of DNA methylation (Groszmann et al., 2011). The global down regulation of 24nt sRNAs does not extend to 21nt sRNAs, which generally exhibit additive expression in hybrids.

Despite the identification of some general trends, much of the sRNA expression inheritance is specific to different crosses/studies. For example, some evidence suggests that miRNAs are transgressively expressed in higher plant hybrids (Shivaprasad et al., 2012), while in other studies miRNAs are more associated with additive expression (Groszmann et al., 2011). The prevalence of transgressive sRNA expression in some crosses may also be underestimated as tissue and lifecycle specific transgressive expression of sRNAs might be masked by methodology. Regardless of which RNA silencing pathways are contributing to transgressive gene expression, it is highly probable that the large complex sRNA networks (MacLean et al., 2010) do in some way impact transgressive phenotypes.

1.5. *C. reinhardtii* as a model system

Crossing experiments, in which the parental genomes and epigenomes are compared, can be used to further characterize the impact of sRNAs in hybridization. Further phenotyping of the hybrids and target predictions of novel sRNA populations could then link RNA silencing with the occurrence of hybrid phenotype. The recent discovery of RNA silencing in *C. reinhardtii* combined with the genetic information available, its haploid vegetative state, and quick crossing procedure/growth rate meant that *C. reinhardtii* was an excellent model organism for this study (Molnar et al., 2007a).

C. reinhardtii is a unicellular green alga well established as a model organism for studying basic molecular topics such as cell cycle control, basal body functions, chloroplast evolution, elucidation of the eukaryotic flagella and, more recently, metabolic engineering for biofuel production (Harris, 2003). There are some limitations to working with *C. reinhardtii* such as the difficulties posed to finding a physiological phenotype in a unicellular organism or the lack of true biological replicates in experiments. However certain characteristics of *C. reinhardtii* are beneficial in relation to this experiment.

The unicellular nature of *C. reinhardtii*, in comparison to the multicellular higher plants, provides an advantage to studies profiling sRNA inheritance. In higher plant crosses, the differential expression of sRNAs in various tissues means that transgressive expression might be underestimated. Tissue specific transgressive sRNA expression occurs in various higher plant hybrids (Barber et al., 2012; Ha et al., 2009; Zhang et al., 2014). The different *C. reinhardtii* cell cycle stages could be conceived as different tissues but this is still a simpler system to work with compared to higher plants

It has two mating types (plus and minus) and can be crossed. Additionally, in contrast to multicellular organisms, an sRNA mutant in *C. reinhardtii*, which lacks nearly all small RNAs, is still viable and can potentially, once the mutation is identified and complemented, be used in crossing experiments (unpublished

data from the Baulcombe lab). This will prove to be a tool of immense importance in studying RNA silencing in this alga.

Algal-based studies of transgressive phenotypes are rare and therefore it is even more valuable to research the role of sRNAs in hybridization in *C. reinhardtii*. While crops are the focus of research for feeding the world, algae are increasingly getting more attention as the potential for algal biofuels is assessed. Microalgae are also thought to carry out half of all the photosynthesis on the planet making them indispensible to our environment (Beardall and Raven, 2004).

Figure 1-4: Light microscope image of *C. reinhardtii*
Image taken of a single *C. reinhardtii* cell (strain CC125+ grown in minimal medium) using a light microscope. Various cell components can be seen including the flagella (f), contractile vacuole (cv), nucleus (n), eye spot (es), and pyrenoid (p). Scale bar represents 10µm.

1.5.1. Transgressive phenotypes in green algae

There are no known examples of transgressive phenotypes in *C. reinhardtii*, but there have been occasions of transgressive phenotypes identified from other algal crosses. Hybrid dysgenesis was noted in crosses between *Volvox carteri* strains from Japan and India (Adams et al., 1990). Often the transgressive phenotype in algae is due to polyploidy, as is the case in the increased cell size in

most of the offspring from the interspecific crosses between closely related mating groups of *Closterium ehrenbergii*, a single celled freshwater alga (Ichimura and Kasai, 1996). There are examples of hybrid dysgenesis in *Chlamydomonas*; when two chlamydomonads, *C. eugametos* and *C. moewusii*, are crossed many of the meiotic products of zygotes die without cell division after several days, likely due to genetic differences (Gowans, 1963).

1.5.2. C. *reinhardtii* lifecycles

To investigate transgressive effects in this alga it is also necessary to fully understand the *C. reinhardtii* lifecycle, both asexual and sexual. Both of these components of the *C. reinhardtii* lifecycle have been extensively studied (Harris, 2003).

1.5.2.1. Vegetative cycle of C. reinhardtii

C. reinhardtii cells are haploid in their vegetative state and multiply via mitosis. Under standard nutrient conditions and day-night lighting, *C. reinhardtii* cells progress through the mitotic cell cycle every 24 hours (Harris, 2009). Depending on growth conditions, during G1 stage of the cell cycle, *C. reinhardtii* cells grow to three or four times their original size. Once large enough, cells undergo DNA synthesis (S-phase) followed rapidly by cell division (M-phase). As these steps of the cycle are so quick, they are referred to inclusively as the S/M phase. The M phase divisions take place inside the mother cell wall, the number of divisions before eventual release dependent on the size to which the cells initially grew to in the G1 phase. Once released cells enter the G0 phase, transferring back into the G1 phase upon light. While the cell cycle of *C. reinhardtii* cultures can be synchronized by growing them on minimal medium (forcing them to photosynthesize) in alternating 12 hour periods of light and dark, the mitotic cell cycle also continues in continuous light conditions and thus does not have an obligate dependence on the day-night cycles.

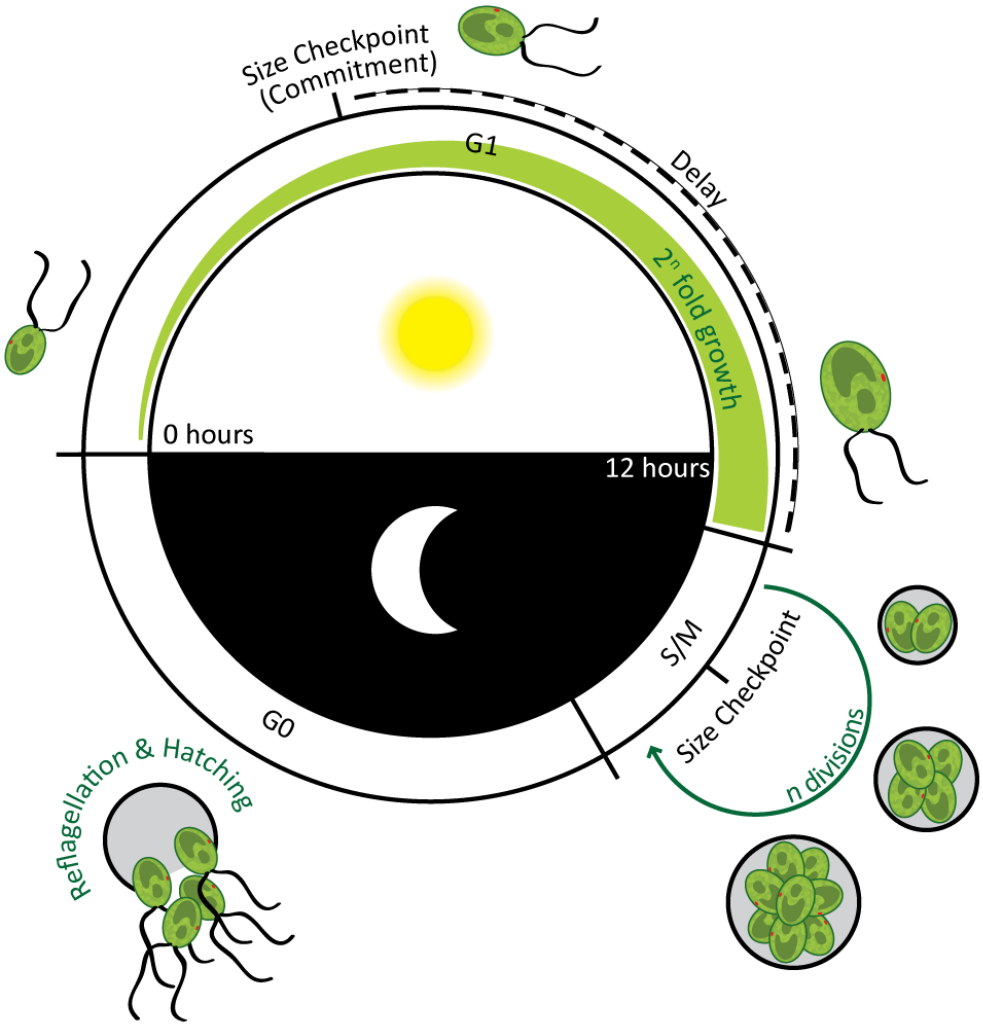


Figure 1-5: Mitotic cell cycle
Depiction of the *C. reinhardtii* vegetative cycle in day-night light conditions. During the day, vegetative haploid cells are in the G1 stage and grow in size. After dark, the cells start the S/M phases of mitosis. The number of divisions (n) depends on the extent of cell size growth in the G1 phase (2^n). Divisions take place inside of the old cell wall of the initial cell. After reflagellation and breaking free of the old cell wall, cells are in G0 stage for the rest of the night, not undergoing any size increase during that time.

1.5.2.2. Sexual cycle of C. reinhardtii

C. reinhardtii haploid vegetative cells, upon lack of nitrogen, undergo gametogenesis. When gametes of opposite mating types encounter one another, they fuse to form a diploid zygote that, besides losing its flagella, also has a zygote specific cell wall, which makes it resistant to much higher levels of stress (Goodenough et al., 2007). It is conceivable that in the diploid stages of this alga

the sRNAs of one parent interact with the genome of the other parent to induce epigenetic changes thus becoming a potential mechanism for the formation of transgressive phenotypes. Daughter cells can then be isolated through tetrad dissection. It is necessary to validate offspring as true recombinants by confirming the presence of DNA from both parents. There are no interspecies crosses as the ability to cross defines the *Chlamydomonas* cells as the same species. Therefore for the purpose of this study intraspecies crosses must be designed where the parents have the greatest genetic distance possible.

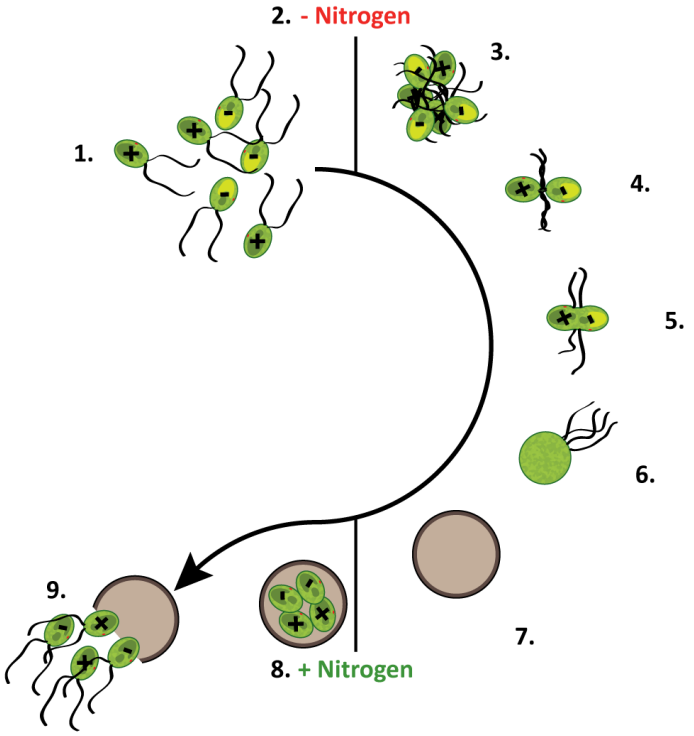


Figure 1-6: Sexual cycle
Diagram describing the phases of the *C. reinhardtii* sexual cycle. **1.** *C. reinhardtii* cells usually exist in haploid vegetative form. They can either be a plus (+) or minus (-) mating type. **2.** Upon depletion of a nitrogen source, cells undergo gametogenesis. This enables them to recognize cells of the opposite mating type when they encounter them, via agglutinin proteins in their flagella. Clumps of cells create a mating structure, the pellicle. **3.** Upon encountering the opposite mating type the cells flagella fuse and the cells are pulled towards one another. A small intra cellular bridge is created using an extension from the minus mating type. **4.** The cells then proceed to fuse cell walls. **5.** Briefly the diploid cell exists in a quadriflagellate state before deflagellation begins. **7.** The diploid cell forms a zygote. The zygote has a zygote specific cell wall that enables the cell to tolerate higher stresses than other cells. The zygote can hibernate for many years. **8.** Upon re-exposure to a nitrogen source, the zygote undergoes meiosis resulting in four haploid progeny trapped in the old zygote shell. **9.** Given enough time, the four haploid progeny, two pluses and two minuses, break free of the zygote shell.

1.5.3. *C. reinhardtii* Genome

The *C. reinhardtii* genome is similar in size and content to the genome of the model higher plant *Arabidopsis thaliana*, except for the basal body and flagella component which is more similar to animal proteomes (Merchant et al. 2007). The algal lineage is estimated to have diverged from land plants over one billion years ago; most genes can be traced to a green plant or plant-animal common ancestor by comparative genomic analyses (Merchant et al. 2007). The genome is gene rich (Table 1-1). 30% of the *C. reinhardtii* genome is made up of repetitive elements, most of which are transposable elements (Harris, 2008). Surprisingly *C. reinhardtii* has a relatively high density of tandem repeats (Table 1-1). These include some partly active transposable elements such as Class I retrotransposons such as TOC1 and Class II DNA elements such as Gulliver (Casas-Mollano et al., 2008). Introns, of which there are ~7 per protein-coding gene, are longer than those of most eukaryotes and are more similar to those multicellular organisms than to protists (Smith and Lee 2008).

A striking difference is the lower level of CG DNA methylation (Table 1-1) in *C. reinhardtii* than in higher plant genomes (Feng et al., 2010). Some genes and transposons are preferentially methylated although the mechanism for transposon methylation seems to be different to that of flowering plants (Feng et al., 2010).

Genetic divergence has been studied between *C. reinhardtii* strains by comparing standard lab strains and interfertile field isolates. These show that different strains have a high level of polymorphisms for many molecular markers (Kathir et al., 2003a).

	Chlamydomonas reinhardtii	*Arabidopsis thaliana*
Size (Mb)	~111.1 [1]	~140.1 [4]
# chromosomes	17 [1]	5 [7]
% repeats	30 [2]	17 [5]
# protein coding genes	17741 [1]	27235 [4]
GC%	64 [1]	35 [7]
CG DNA methylation %	5.38 [3]	24 [6]
CHG DNA methylation %	2.59 [3]	6.7 [6]
CHH DNA methylation %	2.49 [3]	1.7 [6]

Table 1-1: Genome info
Sources: [1](Blaby et al., 2014) [2](Pérez-alegre et al., 2005) [3](Feng et al., 2010) [4](Lamesch et al., 2012) [5](Buisine et al., 2008) [6](Cokus et al., 2008) [7](Merchant et al., 2007)

1.5.4. RNA silencing in C. *reinhardtii*

In 2007, Molnar et al used small RNA sequencing to identify small RNA species between 20nt and 22nt in size present in *C. reinhardtii* (Molnar et al., 2007b). Target analysis and degradome data analysis suggested that small RNAs in *C. reinhardtii* could act on gene expression through mRNA cleavage (Molnar et al., 2007a). Further artificial miRNA systems conclusively showed that miRNAs in *C. reinhardtii* can target mRNA cleavage.

RNA silencing in this model organism is thought to be very similar to higher plants. For example the basic mechanism of miRNA biogenesis is similar in plants and *C. reinhardtii*. sRNAs in *C. reinhardtii* are 2'-O-methylated, like those in *Arabidopsis*, as they are resistant to β elimination (Casas-Mollano et al., 2008). However there are some key differences in RNA silencing of *C. reinhardtii* and higher plants. Most apparent is the lack of 24nt sRNAs (Molnar et al., 2007b). In *C. reinhardtii*: 21nt sRNAs are the major size class.

Other differences concern the action of the small RNAs. It has been demonstrated that the degree of complementarity between the miRNA and its target RNA plays an important role in the activity of the miRNA-complex. In general a low level of complementarity (pairing with 2-9 nucleotide of miRNA 5' end) results in translational repression while a high level of complementarity results in transcript cleavage (Brodersen et al., 2008). In plants, transcript

cleavage is the predominant form of down regulation by sRNAs and there is a high level of miRNA target complementarity (Millar and Waterhouse, 2005). *C. reinhardtii* miRNAs also show a high level of target complementarity suggesting that there is a bias towards target cleavage as the miRNA mechanism although translational repression cannot be excluded (Millar and Waterhouse, 2005). This high target complementarity was used to predict targets in *C. reinhardtii* (Moxon et al., 2008).

Differences also exist between the miRNA pathway of higher plants and *C. reinhardtii*. miRNA genes are usually located in the intergenic regions of the genome in plants while in *C. reinhardtii* they reside both in the intronic and in the intergenic regions of the genome (Jones-Rhoades et al., 2006). The precursors originating from miRNA genes also differ. There are two classes of miRNA precursor: short hairpin and long hairpin precursors. Classic miRNA precursors are short hairpin precursors, less than 150nt in length and producing a single specific miRNA. Though present in *C. reinhardtii*, short hairpin precursors are much more prevalent in plants (Molnar et al., 2007b).

The other class of miRNA precursors found in *C. reinhardtii* are long hairpin precursors with almost perfect pairing and the potential to produce multiple miRNAs. Conversely, long hairpin precursors, though present in higher plants, are more prevalent in *C. reinhardtii* (Molnar et al., 2007b). In fact, recent miRNA studies have discounted miRNAs originating from long hairpin precursors, arguing that they are should be viewed as IR or other hairpin associated sRNAs and are likely created through a different pathway to that of miRNAs (Kozomara and Griffiths-Jones, 2014). If the more relaxed definition of miRNAs is used, 50 miRNA precursors have been identified. Under the more stringent definition only 8 miRNA genes have been confirmed.

A further difference in *C. reinhardtii* and higher plants is that to date no one has verified that the sRNAs in *C. reinhardtii* have an epigenetic effect; the only confirmed sRNA action in this alga is that of transcript cleavage. The lack of 24nt sRNAs suggests that if there are RdDM pathways in *C. reinhardtii*, the mechanism will be different to higher plants.

Components of the RNA silencing pathway in *C. reinhardtii* are similar to those of higher plants. There are three AGO and three DCL homologs in the *C. reinhardtii* genome. They were identified by aligning AGO and DCL sequences from higher plants with the *C. reinhardtii* genome and have all been detected by northern blotting (Casas-Mollano et al., 2008). It is not known however what the distinctive roles of these genes are or whether they are redundant. Phylogenetic analyses of AGOs and DCLs show that they form a sub group with Volvox rather than aligning with other proteins with known functions (Casas-Mollano et al., 2008).

The RNA silencing genes are likely under some cell cycle regulation as preliminary gene expression data suggests that AGO transcription changes throughout the *C. reinhardtii* mitotic and meiotic life cycle; AGO transcripts peak in expression during S/M phase of the sexual cycle (Thompson, 2012). DCL1 and AGO1 regulation could be linked as they are encoded by adjacent divergently transcribed genes (Casas-Mollano et al., 2008). More recently, a potential RDR has been identified in the *C. reinhardtii* genome (www.phytozome.net). This observation suggests the potential creation of secondary siRNAs by an initiator sRNA.

Not much is yet understood of the function of RNA silencing in *C. reinhardtii*. Recent small RNA studies have suggested some roles in cell biology and stresses (Molnar et al., 2007a; Shu and Hu, 2012). For example, many small RNAs up regulated in the zygote stage of the lifecycle are predicted to target flagella-associated proteins (unpublished data from the Baulcombe lab) despite the low proportion of these proteins in comparison to other types of proteins in this alga (Misumi et al., 2008). To further elucidate the role of miRNAs in the life cycle of *C. reinhardtii*, they must be catalogued and their targets must be identified and verified. And to understand the role sRNAs might play in hybridization in *C. reinhardtii*, experiments need to test for the nature of sRNA inheritance in this alga.

1.6. Aims and objectives

In this review of transgressive phenotypes I recounted the current models proposed for the formation of transgressive phenotypes and how epigenetics, specifically RNA silencing, could play a role in answering some of the unanswered questions in this field. I then introduced *C. reinhardtii* as a candidate model organism in which to further explore the relationship between transgressive phenotypes and sRNAs. In this study I aimed to first determine the nature of sRNA inheritance in *C. reinhardtii* and then identify and classify transgressive sRNA loci.

1.6.1. Hypothesis

RNA silencing can cause transgressive gene expression via novel sRNA target acquisition (Figure 1-7) and spontaneous transgressive sRNA expression (Figure 1-8).

In *C. reinhardtii*, the sRNAomes of both parents will interact with each other and with both parental genomes in the diploid zygote stage. The length of the zygote hibernation, and perhaps the up regulation of RNA silencing pathway in this stage, could expose novel targets to RNA silencing (Figure 1-7-A), perhaps leading to the deposition of novel secondary sRNA loci (Figure 1-7-B), or catalyse the spontaneous formation of transgressive sRNA loci (Figure 1-8). After this collision of genomes and sRNAomes, sRNAs from one parent can interact with genomic information from the other parent in the recombinant cells.

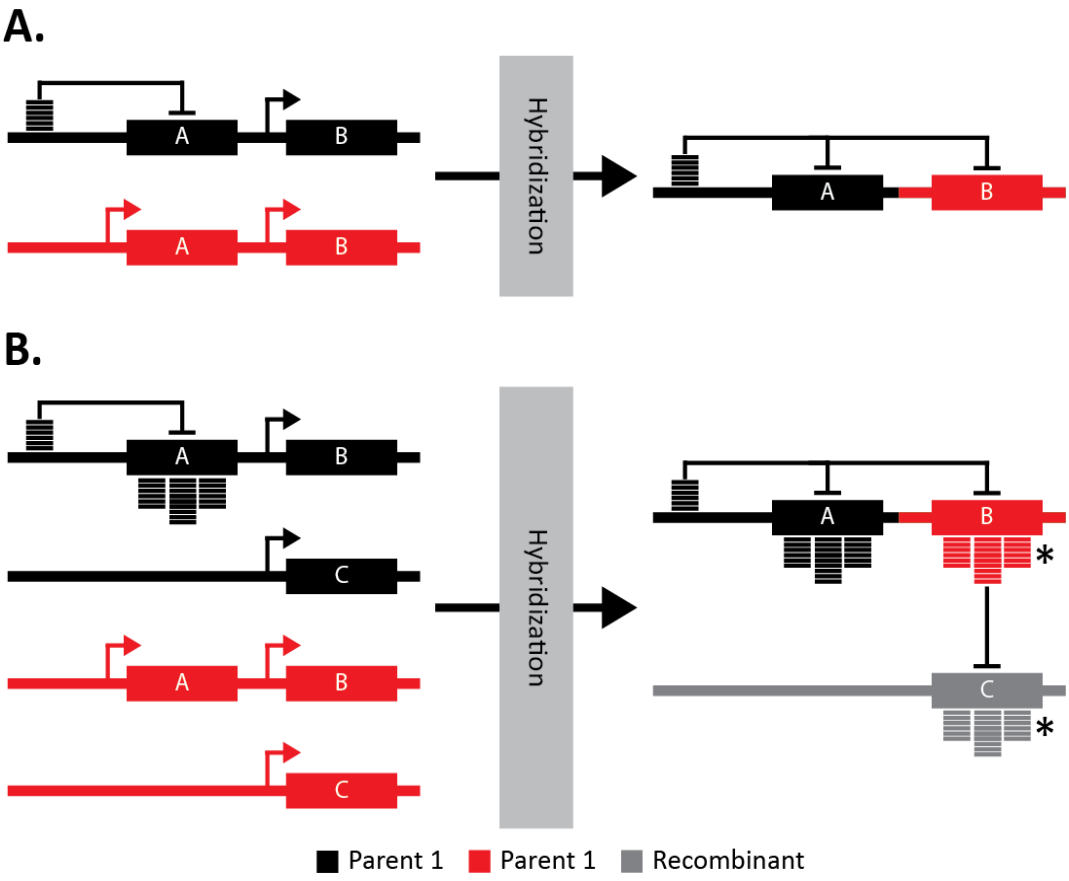


Figure 1-7: Possible results of novel sRNA target acquisition
Models for how transgressive gene expression can be created by the acquisition of novel targets of RNA silencing. Transgressively expressed sRNA populations are indicated with a *. **(A)** An sRNA from Parent 1 is exposed to the novel genetic variation of Parent 2 thus acquiring a novel target. This sRNA can then alter the gene expression in the recombinant so that the gene is transgressively expressed between recombinant and parents. In this diagram the action of RNA silencing down regulates gene expression however it is possible that the sRNA acts to increase gene expression. This scenario relies on differential sRNA expression in the parents. As there is not epigenetic effect this transgressive gene expression is not heritable through a lineage. **(B)** In this scenario, the sRNA from Parent 1 sets down epigenetic marks at its novel target in the recombinant thus having the potential to create heritable transgressive gene expression. Although this could be in the form of DNA methylation, if secondary siRNAs are created at the novel target, as in the case of this diagram, then a heritable transgressive siRNA locus is formed. The novel siRNAs could then produce a cascade effect, catalysing the creation of multiple novel siRNA loci.

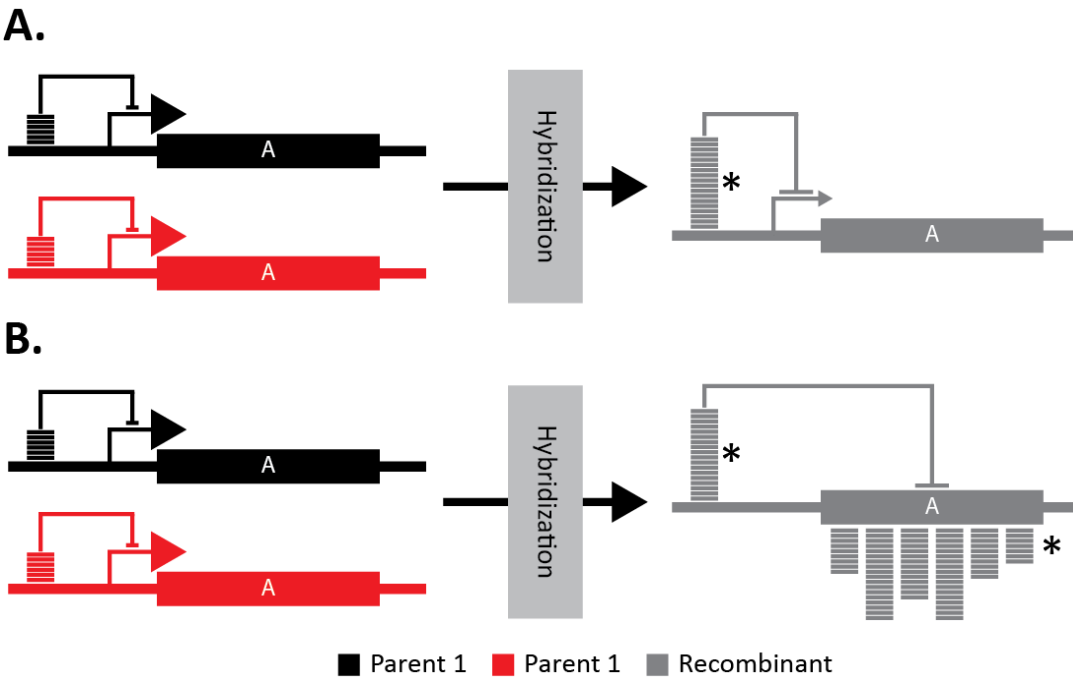


Figure 1-8: Effects of spontaneous transgressive sRNA expression
Genetic interactions could create seemingly spontaneous transgressive sRNA expression in recombinants. These diagrams explain the possible consequences of such a situation. Transgressively expressed sRNA populations are indicated with a *. **(A)** Transgressive expression of an sRNA, whether found in both Parent 1 and Parent 2 or just one parent, could have a dosage dependent effect on gene expression. If the transgressive expression of that sRNA is heritable then the resulting transgressive gene expression will similarly be heritable. However it is possible that the transgressive sRNA expression is due to genetic interactions that could be lost in future generations **(B)** If the transgressive expression of the initial sRNA, instead of having a dosage dependent effect on a gene (via mRNA cleavage for example), causes epigenetic modifications at the target locus then it could result in heritable transgressive gene expression. If the epigenetic modification is in the form of secondary siRNAs then another transgressively expressed siRNA locus is created with the potential to be self-sustaining and thus transgenerational. The secondary siRNAs have the potential to create a cascade of transgressively expressed siRNA loci.

As a result of these two interactions I expected to find transgressively expressed sRNA loci in the *C. reinhardtii* recombinants. Specifically I sought to identify transgressively expressed sRNA loci with characteristics of secondary sRNAs (such as phasing), which would support the potential of sRNAs to lay down epigenetic modifications that could form transgressive phenotypes in *C. reinhardtii*. In Chapter 3 I establish the genetic diversity of the parental strains chosen for this cross. In Chapter 4 I compare the sRNAomes of the parental strains chosen for this cross. And in Chapter 5 I describe the generalized inheritance of sRNAs in *C. reinhardtii* and detail the discovery of transgressively expressed sRNAs in this alga.

2. Chapter two: Materials and Methods

2.1. *C. reinhardtii* strains and culture conditions

All strains (Table 2-1) were obtained from the *Chlamydomonas* Resource Center (http://www.chlamy.org) except for the Japanese strains (J+ and J-), which were a gift from Dr Takashi Nakada (Institute for Advanced Biosciences, Keio University).

Code	Mating type	Source	Description
CC125+	+	MA, USA	Standard lab strain, containing the *nit1,* and *nit2* mutations. Background strain for CC503 cell wall-less strain, which was the source of DNA used for genomic sequencing.
C124-	-	MA, USA	Standard lab strain, containing the *nit1* and *nit2* mutations.
CC2936+	+	QC, Canada	Wild isolate
CC2290-	-	MN, USA	Wild isolate
J+	+	Kagoshima, Japan	Wild isolate
J-	-	Kagoshima, Japan	Wild isolate

Table 2-1: *C. reinhardtii* strains
List of all strains used in this study including their mating types, isolation location and a brief description.

C. reinhardtii strains were grown on tris-acetate-phosphate (TAP) or minimal (TAP minus acetate) media plates supplemented with 1.5% (w/v) agar (http://www.chlamy.org/media). Cultures were maintained in growth chambers (MLR-352, Panasonic Biomedical, Leicestershire, U.K.) under continuous light (60 µmol m^{-2} s^{-1}), at 25°C in ambient CO_2. Liquid cultures (50 ml) were inoculated from single colonies of *C. reinhardtii* growing on solid agar plates into 150 ml conical flasks. Cultures were grown in a Multitron Standard shaker (Infors HT, Bottmingen, Switzerland) at 140 rpm, under continuous light (60 µmol m^{-2} s^{-1}) at 21°C in ambient CO_2.

TAP and minimal media was made according to standard recipes (http://www.chlamy.org/media). To create 1/10 N TAP, only 1.6 g of NH_4Cl was

added to the Beijerink salts and to create –N TAP, NH_4Cl was not added to the Beijerink salts.

2.2. Crossing

Adapted from the method in Jiang & Stern, 2009 to include recommendations from Dr Sinead Collins (Edinburgh University).

2.2.1. Strain preparation

To ensure the strains were healthy enough to cross, cultures were streaked to single cell density on decontamination plates (TAP, 1.5% agar, 0.1 L-methionine sulfoximine, 0.1 mg/ml arginine, 0.05 µg/ml ampicillin) and cultured for a maximum of five days. Strains were transferred to TAP-agar plates and then re-streaked several times at 2-3 day intervals.

2.2.2. Gametogenesis

To obtain gametes, a large loop-full of vegetative cells was streaked onto 1/10 N TAP plates (TAP contained one tenth the standard amount of nitrogen supplemented with 1.5% (w/v) agar) and were incubated for five days under continuous light. Cells were then resuspended in distilled water to approximately 2×10^7 cells/ml and incubated in the light for 1-2 hours to allow the regeneration of flagella.

2.2.3. Zygote formation

Mating type plus and minus strains (2 ml) were mixed and allowed to mate. At 0, 0.5, 1, 2, and 3 hour intervals the cell mixture was spotted (200 µl/drop) onto 3% agar 1/10 N TAP plates. The plates were cultured upright overnight in continuous light and the remaining pellicle mixture was kept in the 6 well plate to confirm pellicle formation the next day (the bottom of the wells would take on a speckled appearance and not be dislodged by swirling). After an overnight culture, the plates were wrapped in aluminium foil and stored for one week to allow zygote maturation.

2.2.4. Zygote separation

After a week, a sterile straight razor blade was used to scrape off the green unmated cells from the spots. Presence of the larger spherical clear zygote cells

was confirmed under a dissecting microscope (2.4). A fine copper wire with a flattened end was used to manipulate the zygotes into a small area on the plate, excise the area, and transfer it onto a maturation plate (standard TAP plate). Using a blunted glass thread zygotes were transferred to defined positions on the plate in the late afternoon/evening. The maturation plates were held over chloroform for exactly 30 sec to kill any remaining vegetative cells and left in moderate light overnight.

2.2.5. Tetrad dissection

Early the next morning the plates were examined under the dissecting microscope at intervals through out the day to 'catch' the four-cell stage of meiosis. Using the glass thread, the daughter cells were separated to defined positions. Plates were cultured for one week to allow the growth of daughter cell cultures from a single cell.

2.3. *C*ell density measurements

Cell density of low volume *C. reinhardtii* cultures was measured using a KOVA Glasstic slide 10 with grids (Hycor, IN, U.S.A.). One tenth volume of Lugol's iodine was added to liquid *C. reinhardtii* cell cultures to immobilize and stain the cells. Then 6.6 µl of cell culture was placed on the Glasstic slide and cell density counted according to Sambrook & Russell, 2001.

A standard curve of cell density versus absorbance at 680 nm was also used to measure the growth of cultures when enough liquid volume was available. Absorbance was measured at 680 nm using a Helios Epsilon spectrophotometer (Thermo Scientific, MA, U.S.A.).

2.4. Microscopy

Macroscopic images were taken of *C. reinhardtii* using a light microscope (DX43, Olympus, Tokyo, Japan). Photographs of plates were taken using a hand held Lumix DMC-TZ3 (Panasonic, Osaka, Japan) camera.

2.5. Polymerase chain reaction (PCR) based techniques

2.5.1. General

PCRs were performed using up to 500 ng of DNA template. The reactions were performed using the GoTaq PCR kit (M5001, Promega, WI, U.S.A.) according to the manufacturer's specifications. Cycling conditions were optimised for each primer pair and length of the PCR product but were typically as follows:

	Time (min)	Temperature (°C)	
Initial denaturing step	2	95	
Denaturing step	0.5	95	35×
Annealing step	0.5	Tm of primer - 5	
Extension step	1/kb	72	
Final extension step	5	72	

2.5.2. 18S and ITS2 PCR

PCRs were set up to amplify the *C. reinhardtii* 18S gene. Primers, 18S-FA and 18S-RB (Appendix 7.1.1) from Nakada et al., 2010 were used and an annealing temperature of 65°C was used. PCRs were set up to isolate the *C. reinhardtii* ITS2 sequence. Primers, ITSa and ITSb, (Appendix 7.1.1) from Nakada et al., 2010 were used and an annealing temperature of 70°C was used.

2.5.3. Mating type PCR

PCRs were set up to amplify the *C. reinhardtii* FUS1 and MID genes. Primers, mid-up, mid-low, fus1-up and fus1-low (Appendix 7.1.1) from Werner & Mergenhagen, 1998 were used and an annealing temperature of 52.6°C was used.

2.5.4. Mapping marker PCR

PCRs were set up using mapping markers provided by the *Chlamydomonas* Research Center (Kathir et al., 2003b) to amplify genetic markers in *C. reinhardtii* strains.

2.5.5. Visualising PCR products

PCR products were separated by horizontal agarose gel electrophoresis. Agarose gels were typically 2% (e/v), containing 0.1 µg/ml ethidium bromide. Gels were run in 1× TBE (90 mM tris-borate, 2mM ethylenediaminetetraacetic acid (EDTA))

at 70-100 V and visualised on a long range UV box (DarkReader Transilluminator, Clare Chemical Research, CO, U.S.A.). Gene ruler ladders Hyperladder I and Biorad Hyperladder IV (Bioline, London, U.K.) were used to estimate fragment size.

2.5.6. Sequencing DNA products

Sequencing reactions for PCR products were performed using the BigDye v3.1 kit (4337455, Life Technologies, CA, U.S.A.) according to manufacturer's instructions and the reactions were sent to Cogenics (Beckman Coulter Genomics, Essex, U.K.) for sequencing. The returned sequence files (.abi) were analysed using CLC genomics workbench.

2.6. Nucleic acid purification

2.6.1. Genomic DNA extraction

Method adapted by myself from that of Dr. Andrew Bassett.

C. reinhardtii cells were grown in standard liquid culture conditions in TAP medium until mid-log phase was reached (when cell density reaches 1-5×10^6 cells/ml). Cell density was calculated using optical measurements (Section 2.3). 50 ml of culture was centrifuged at 800 ×g for 15 min to pellet the cells.

The pellet was resuspended in 4 ml of the lysis buffer (1% bovine serum albumin (BSA), 1 mM EDTA, 0.5 M sodium phosphate ($NaPO_4$) pH 7.2, and 7% sodium doecyl sulphate (SDS)) and 40 µl of proteinase K was added to the solution. The samples were incubated overnight at 55°C. Then 5 ml of phenol:chloroform:isoamyalcohol, 45:23:1 (Life Technologies) was added to each sample and each sample gently inverted for 15 sec. The samples were centrifuged at 4500 ×g for 10 min at room temperature and the upper phase transferred to a new 1.5 ml tube. The phenol-chloroform extraction step was repeated once more.

One-fifth volumes of 5 M sodium chloride (NaCl) and 2.5 volumes of pure ethanol was added and the samples gently inverted for 10 sec. The sample was incubated at 4°C for 30 min before the sample was centrifuged for 30 min at 13,000 ×g at 4°C. The pellet was washed with 5 ml of 70% (v/v) ethanol and was

centrifuged for 5 min at 4500 ×g at 4°C to collect DNA. The supernatant was removed and the pellet was air-dried. The pellet was then resuspended in 300 µl RNase solution, transferred to a 1.5 ml Eppendorf tube, and incubated at 40°C for 1 hour. Then 300 µl of phenol:chloroform:isoamyl alcohol, (45:32:1, Life Technologies) was added to the sample and the sample was centrifuged for 3 min at 10,000 ×g at room temperature and the upper phase was transferred to a new tube.

DNA was precipitated by adding one tenth volumes of 3 M sodium acetate (NaOAc) and 1 ml pure ethanol. The visible DNA precipitate was transferred to a new tube containing 1 ml 70% (v/v) ethanol. The sample was centrifuged for 5 min at 20,000 ×g at room temperature to collect the DNA. The pellet was washed twice using 1.5 ml 70% (v/v) ethanol and centrifuging for 5 min at 20,000 ×g between each wash. The pellet was then air dried and resuspended in 50 µl of distilled water.

Concentration and purity of the DNA was checked using the Qubit system (0). Quality of DNA was also checked by running 1 µl on a 2% (w/v) agarose gel, stained with ethidium bromide, quantified using Hyperladder I (Bioline), and visualised on a long range UV box (DarkReader Transilluminator, Clare Chemical Research).

2.6.2. RNA extraction

RNA was extracted according to a protocol modified from Molnar, Schwach, Studholme, Thuenemann, & Baulcombe, 2007.

C. reinhardtii cells were streaked to a single cell density onto solid minimal medium and grown in standard solid culture conditions. A single colony was used to inoculate 50 ml of liquid minimal medium, which was grown in standard liquid culture conditions until mid-log phase was reached (1-5×10^6 cells/ml). Then 50 ml of culture was centrifuged at 800 ×g for 15 min to pellet the cells and the supernatant removed. The pellet was frozen in liquid N_2 and stored at -80°C.

To extract the RNA from the frozen tissue, the pellet was resuspended in 6 ml of TRIzol reagent (15596-026, Life Technologies) and the sample was kept on ice.

The polysaccharides were then pelleted by centrifugation for 15 min at 4000 ×g at 4°C. The upper phase was transferred to a fresh 15ml tube and incubated for 5 min at room temperature to ensure the complete dissociation of nucleoprotein complexes. Then 1.2 ml of chloroform was added after which the samples were vortexed for 15 sec and once again incubated for 5 min at room temperature. The resulting mixture was centrifuged for 15 min at 4000 ×g at 4°C and the upper phase was transferred to a fresh 15 ml tube and kept on ice. RNA was collected by centrifuging the samples for 30 min at 4000 ×g at 4°C and the supernatant removed by aspiration. Residual salts were then removed by rinsing the pellet with 8ml of 80% ethanol and immediately afterwards, the RNA was again collected through centrifugation for 5min at 4000 ×g at 4°C. Once the supernatant was removed the pellet was again rinsed with 8 ml of 80% ethanol and centrifuged for 5 min at 4000 ×g at 4°C. Special care was taken to remove all supernatant and the pellet was air-dried at room temperature for 3-5 min. The samples were placed on ice and the pellets resuspended in 100 µl of RNase-free water. To allow the RNA to rehydrate, the sample was incubated for 15min on ice and then vortexed for 15 sec.

Concentration and purity of the RNA was checked (Appendix 2.7). Quality of RNA was also checked by running 2 µl on a 10% TBE precast gel (456-5013-MSDS, Bio-Rad, CA, U.S.A.), stained with SYBR gold, quantified using Hyperladder I (Bioline), and visualised on a long range UV box (DarkReader Transilluminator, Clare Chemical Research).

2.6.3. Agarose gel extraction of nucleic acid

DNA separated by agarose gel electrophoresis, stained with ethidium bromide and visualised on a long range UV box (DarkReader Transilluminator, Clare Chemical Research), was excised from the gel using a sterile straight edge razor. DNA was extracted and purified using the QIAquick Gel Extraction kit (28706, Qiagen, Limburg, Netherlands).

2.7. Nucleic acid quantification

DNA and RNA samples were quantified using a Qubit 1.0 Fluorometer (Life Technologies) according to the Quant-iT Assays Abbreviated Protocol. The concentration of PCR products was attained using the Nanodrop 1000 Spectrophotometer (ThermoScientific, MA, U.S.A.).

2.8. Genome sequencing

Total DNA was diluted to 20 ng/µl and 100 µl was fragmented using a Diagenode Bioruptor Standard (Diagenode, Liège, Belgium). Parameters such as power and time were optimized to 90 min of high power fractionation with 30 sec on 30 sec off (Section 3.2.3). Fractionated DNA fragments were cloned using the TruSeq DNA Sample Preparation LT kit (Illumina, CA, U.S.A.). The Cancer Research Institute United Kingdom (CRUK, Cambridge, UK) sequenced the DNA libraries on a HiSeq (Illumina) and returned .fastq files with the DNA read sequences.

2.9. sRNA sequencing

Total RNA extracted from vegetative *C. reinhardtii* cells was used to create sRNA libraries for sRNA sequencing using the TruSeq Small RNA Sample Preparation kit (Illumina). The CRUK sequenced the sRNA libraries on an HiSeq (Illumina) and returned .fastq files with the sRNA read sequences.

2.10. Small RNA northern

Northern blots were performed using a modified version of the protocol described by Molnar et al., 2007b.

2.10.1. Total RNA separation

Total RNA was separated on 0.75 mm thick 15% polyacrylamide-urea gels (7 M urea, 15% 19:1 acrylamide:bis-acrylamide, 1× TBE, 700 mg/L ammonium persulphate (APS) and 650 µl/L tetramethylethylenediamine (TEMED)) using 0.5× TBE buffer as the running buffer. RNA (5-10 µg eluted in FDE dye) was denatured at 95°C before being loaded into the wells. Empty wells were filled with 5 µl 2× FDE. Gels were run at 50 V until both dyes in the FDE in the RNA samples had

entered the gel. Voltage was then increased to 150 V until the bromophenol blue dye had reached the bottom of the gel.

2.10.2. Blotting

After separation of total RNA on a 15% (w/v) denaturing 7 M urea polyacrylamide gel, the gel was soaked in 50-100 ml of 20× SSC for 10min. The RNAs were transferred by overnight capillary blotting in 20× SSC onto nylon Hybond N+ (Amersham Biosciences, Buckinghamshire, U.K.) membranes. The RNA was crosslinked to the membranes with UV at 120,000 µJOULES (UV Stratalinker 2400, Agilent Technologies, CA, U.S.A.).

2.10.3. Nucleic acid end labelling

Probes complementary to sRNAs were end-labelled with [γ-32P] ATP using a T4 polynucleotide kinase (New England Biolabs, Hitchin, U.K.). First 2 µl of 10 µM oligonucleotide probes (Appendix 7.1.2) were mixed with 20 pmol [γ-32P] ATP and 1 µl T4 polynucleotide kinase (10 units/µl) in 2 µl 10× polynucleotide kinase buffer and incubated at the hybridization temperature for 30-60 min. Unincorporated nucleotides where separated out from the labelled oligos on a Microspin G-25 column (27-5325-01, GE Healthcare, Little Chalfont, U.K.). The probe was denatured at 95^{o}C for 20 sec and then placed on ice.

2.10.4. Hybridization

Hybridization performed overnight in 5-10 ml of ULTRAHyb Oligo Hybridization Buffer (Life Technologies). A hybridization temperature of 37^{o}C was used for detection of high expression sRNAs and at 35^{o}C for low expression sRNAs. Membranes were washed twice at hybridization temperature with 2× saline-sodium citrate buffer (SSC), 0.5% (v/v) SDS for 30 min.

2.10.5. Blot stripping

To enable reprobing, hybridized probes were removed from membranes by boiling the membranes in 0.1% (v/v) SDS for 20 min, twice.

2.10.6. Phosphor imaging of RNA gel blots

Northern membranes were sealed in polythene bags and placed onto Type III-s Fuji phosphor imaging plates (Fujifilm, Bedford, U.K.). Depending on the strength

of signal, membranes were kept there for 30 min to two weeks, after which they were placed on a Typhoon 8610 imaging system (Amersham Biosciences.) for visualization and quantification. Decade RNA marker (Life Technologies) was used to estimate sRNA sizes.

2.11. Bioinformatics

2.11.1. Databases

The following databases were used:

Species	URL
Chlamydomonas reinhardtii	http://www.phytozome.net/chlamydomonas
Volvox carteri	http://www.phytozome.net/volvox
General	http://www.ebi.ac.uk

Databases were searched using the BLAST algorithms available from their websites (Altschul et al., 1990).

2.11.2. Quality checking for sequencing libraries

The quality of the DNA and RNA library read files (.fastq) was analysed using FastQC (Andrews, 2012).

2.11.3. Trimming and filtering of sequencing libraries

2.11.3.1. Pipeline processing

Dr Sebastian Mueller loaded the DNA and sRNA library read files into the DCB pipeline (unpublished software). The resulting analysis and alignment files were made available on the pipeline.

2.11.3.2. Trimmomatic 0.27

Adapter sequences and low quality end sequences in the DNA were trimmed by the ILLUMINACLIP tool from Trimmomatic 0.27 (Bolger et al., 2014) allowing one mismatch to the Illumina sequencing adapters and a minimum quality score of 20 for ends of reads or windows of four nucleotides. This program uses the quality scores associated with the nucleotides of the reads to remove low quality sequences. The .fastq files of the sequencing libraries provided by CRI-UK were used as input into Trimmomatic and a .txt file modified from the list of sequencing adapters provided by Illumina was used as a reference to identify adaptor sequence for trimming.

2.11.4. Alignment

2.11.4.1. Short sequences

Short DNA sequences such as the sequencing results of the ITS2 and 18S sequences were aligned to their references using the CLC genomics workbench. Multiple alignments were created using the ClustalW algorithm.

2.11.4.2. Bowtie alignments of sequencing libraries

The short read alignment software, Bowtie (Langmead et al., 2009) was used to align both single end 50bp reads (SE50) of the sRNA libraries and paired end 100bp reads (PE100) of the DNA libraries. The only parameter altered in various alignments was the –v mismatch parameter (i.e. –v 3 allows 3 mismatches).

Single end sRNA libraries returned from the CRI-UK in the form of .fastq files. These .fastq files and a .fasta file of the reference genome (Creinhardtii_236.fa) taken from Phytozome (www.phytozome.net) were used as input into bowtie alignment program. Paired end DNA libraries returned from the CRI-UK in the form of .fastq files. These .fastq files and a .fasta file of the reference genome (Creinhardtii_236.fa) taken from Phytozome were used as input into bowtie alignment program.

2.11.4.3. CLC genomics of sequencing libraries

The .fastq files of DNA libraries were loaded into the CLC genomics workbench as paired end read files. The resulting paired end read file was then aligned to the genome using the map reads to reference function using the default parameters. The .fasta files of sRNA libraries were loaded into the CLC genomics workbench as single end read files and also aligned to the genome using the map reads to reference function. This function allows certain mismatches between the reads and the genome based on both the quality scores of the reads and the number of reads aligning with that mismatch.

2.11.5. Variant detection by CLC genomics

Sequence variation (in the form of SNPs) was detected using the quality-based variant detection function in the CLC Genomics Workbench using the default parameters. This function takes both quality, coverage, and frequency into account to find variants covered by aligned reads.

2.11.6. SegmentSeq

The loci prediction software SegmentSeq predicts sRNA loci and identifies differential expression between libraries. Alignment files (.bam) created by bowtie for the sRNA libraries were used as input into SegmentSeq (Hardcastle et al., 2012). SegmentSeq analysis to predict loci and differential sRNA expression was performed in R (example code in Appendix 7.9).

To count the number of sRNA loci predicted for one strain and taking into account the replicates, the likelihoods of every locus in a classSeglikelihood object was summed using the following command in R.

```
# Count loci
summariseLoci(classSeglikelihood)
```

2.11.7. PhaseR

Dr Bruno Santos analysed the sRNA alignment files for phased loci using PhaseR (http://www.plantsci.cam.ac.uk/bioinformatics/phaser). PhaseR distinguishes likely occurrences of phasing from random patterns of sRNA alignment. For input into phaser, classSeglikelihood objects from SegmentSeq (Hardcastle et al., 2012) were used. To count phased loci, only loci with an average score less than negative ten between the replicates for that strain were counted as phased.

2.11.8. Identifying miRNAs

To identify known miRNAs in the sRNA libraries, the miRBase list of mature plant miRNAs was used as a reference. Trimmed and filtered sRNA libraries (in the form of .fasta files) created by the DCB pipeline were used as input into the UEA tool miRProf, using default parameters and checking for matches to precursors rather than mature miRNAs(Moxon et al., 2008).To predict novel miRNAs, trimmed and filtered sRNA libraries (in the form of .bam files) created by the DCB pipeline were inputted into the UEA tool miRCat (Moxon et al., 2008) using default parameters to predict miRNAs. To count miRNAs, only miRNAs whose average expression was greater than zero and who were predicted in each replicate were counted.

2.11.9. Overlap analysis

In collaboration with Dr Sebastian Mueller, a script was written to count the number of overlaps between genome annotations and read alignments. Genome annotations for repeats and genes were taken from Phytozome (http://www.phytozome.net). Either alignment data objects (aD) for reads or classSeg objects for loci from SegmentSeq were used as input. The program was run in R as follows:

```
# Download the relevant R libraries
library(rtracklayer)
library(segmentSeq)

# Load repeat and gene annotation files. The repeat file used here excludes simple repeats.
repeats <- import.gff3("Repeats.gff3)
genes <- import.gff3("Genes.gff3)

# Load in read alignment in form of alignment data (aD) object from segment seq
load("aD_sRNA_reads.RData")

# Priority lists either putting genic elements before repeats (prioritylistgenes) or repeats
before genic elements (prioritylistrepeats)
prioritylistgenes <- c("CDS", "five_prime_UTR", "three_prime_UTR", "gene", "mRNA",
"Repeat")
for (type in prioritylistgenes) {if(exists("annotationreorderedgenes"))
{annotationreorderedgenes <-
c(annotationreorderedgenes,annotation[annotation$type==type,]) } else
{annotationreorderedgenes <- annotation[annotation$type==type,]}}
prioritylistrepeats <-c("Repeat","CDS", "five_prime_UTR", "three_prime_UTR", "gene",
"mRNA")
for (type in prioritylistrepeats) {if(exists("annotationreorderedrepeats"))
{annotationreorderedrepeats <-
c(annotationreorderedrepeats,annotation[annotation$type==type,]) } else
{annotationreorderedrepeats <- annotation[annotation$type==type,]}}

# Counting overlaps between annotations and reads
sRNAvsannotation <- findOverlaps(aD@alignments,annotationreordered)
sRNAuniqueannotation <- !duplicated(queryHits(sRNAvsannotation))
sRNAoverlaptype <- rep("intergenic", length(aD@alignments))
sRNAoverlaptype[queryHits(sRNAvsannotation)[sRNAuniqueannotation]] <-
as.character(annotationreordered[subjectHits(sRNAvsannotation)[sRNAuniqueannotation],
]$type)

# Save table of number of overlaps
write.csv(table(overlaptype),"overlap_sRNA_reads.csv")
```

2.11.10. Creating MA plots

Plots of the ratio of log sRNA read expression versus the average of the log sRNA read expression of two strains were created in R using a script provided by Dr

Sebastian Mueller. BaySeq (Hardcastle and Kelly, 2010) was used to identifying differentially expressed sRNA reads. Lists of known miRNAs were taken from miRBase (Griffiths-Jones et al., 2006) and lists of predicted miRNAs were created using miRCat. These annotations were used to indicate differentially expressed and miRNA reads in the MA plots (example code in Appendix 7.10).

2.11.11. Normalization

Except for cases in which the software performs the normalization automatically (such as SegmentSeq), counts were normalized as follows:

Normalized count = count × (median library size / actual library size)

2.12. Visualization of alignments

DNA and sRNA alignments were visualized using the Integrative Genomics Viewer (http://www.broadinstitute.org/igv/) (Thorvaldsdóttir et al., 2013). Alignment files of the reads (.bam) and annotations (.gff3) were used as input and coverage normalized using the IGV normalization option.

3. Chapter three: Genetic divergence between strains

Two pairs of geographically distant strains of C. reinhardtii were shown to be genetically divergent to each other. The more recently isolated strains, J+ and J-, were more divergent to the reference genome.

3.1. Introduction

The aim of my project was to explore the role of sRNAs in transgressive phenotypes of *C. reinhardtii*. Increasing genetic distance has been linked to the increasing frequency of transgressive traits in other organisms (Burstin and Charcosset, 1997; Pascual et al., 2013; Stelkens et al., 2009) and reviewed in past literature (Stelkens and Seehausen, 2009). To increase the likelihood and potency of transgressive phenotypes, I planned to use the most genetically distant *C. reinhardtii* strains possible.

3.1.1. Genetic variation of C. *reinhardtii* available at the time was limited

There was a limited amount of genetic variation available in *C. reinhardtii* strain stocks at the time (*Chlamydomonas* Resource Centre, 2010). Many of the current stocks are derived from an isolate from Massachusetts in 1945 and can be grouped into three sub lines based on mutant phenotypes and DNA markers (Pröschold et al., 2005). Some genetic variation exists in these lines and may have been generated under lab conditions because even mild environmental stress can elicit mutations (Goho and Bell, 2000) and increase the mutation rate. However the mutation rate in *C. reinhardtii* is extremely low,[1] estimated to be between 2.08×10^{-10}/site/generation (Ness et al., 2012) and 6.76×10^{-11}/base/cell division (Sung et al., 2012) and these sub lines were not sufficiently diverse to be used in the study of transgressive phenotypes.

As an alternative to the use of sub lines to explore transgression in *C. reinhardtii*, I therefore explored the possibility of using isolates from distant geographic locations. A recent study characterized the genetic variation in the North

[1] Indeed it is ~90x lower than the mutation rate per generation for *Arabidopsis thaliana* despite

American strains isolated from different locations (Jang and Ehrenreich, 2012). The North American strains (including CC125+ and CC124-) are genetically divergent and can be phylogenetically assigned into two groups (Jang and Ehrenreich, 2012). However I was interested in using the most genetically divergent strains possible and more recent studies established *C. reinhardtii* as a truly cosmopolitan alga[2] (Figure 3-1) found across eastern North America and Japan (Nakada et al., 2010). It may also have been isolated in South Korea (Hong et al., 2013), but interfertility experiments are needed to confirm that the samples are indeed *C. reinhardtii*.

3.1.2. North American strains (CC125+/CC124-) and Japanese strains (J+/J-) chosen as parents for crossing experiment

In order to have the highest genetic distance between the parents of the crossing experiment, I decided to use the most geographically isolated strains, a pair from North America and another from Japan (Figure 3-1). Breeding pairs were chosen to allow the possibility of testing the effect of genetic divergence on transgressive sRNA populations as strains from the same collection site are usually sister strains of a single tetrad and considered highly genetically similar (Pröschold et al., 2005). Also breeding pairs allow reciprocal crosses to be performed in the future.

[2] How *C. reinhardtii* can be so cosmopolitan and genetically different, yet remain morphologically similar and interfertile is related to dispersal efficiency or static morphological similarities, contributing to the discussion on the definition of a species in freshwater algae (Ichimura, 1996).

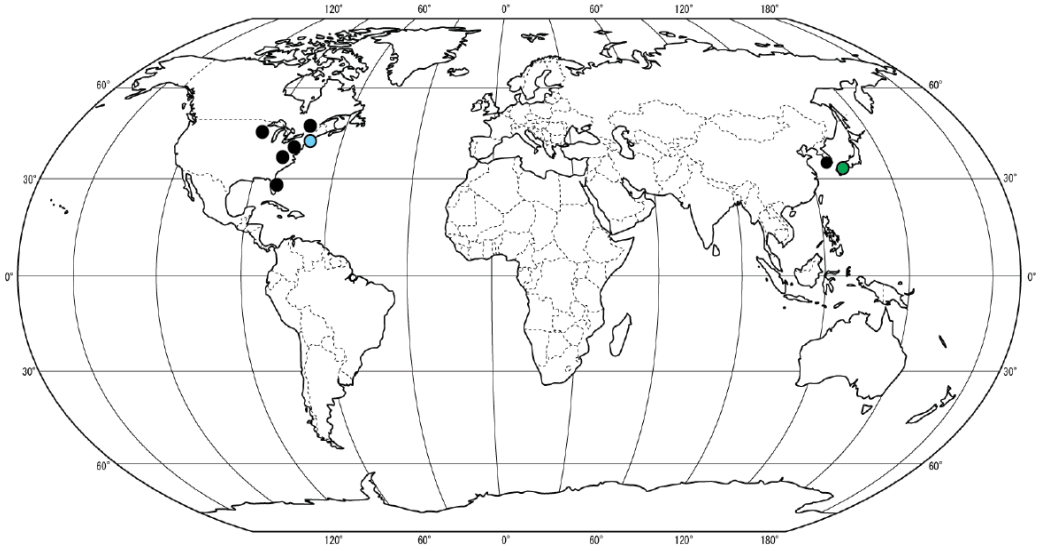


Figure 3-1: Sampling locations of *C. reinhardtii* isolates.
A map depicting where *C. reinhardtii* strains have been isolated. Most strains were isolated from agricultural fields using methods taking advantage of positive phototaxis exhibited by this alga. Blue: CC125+/CC124-. Green: J+/J-. Black: Known collection sites of other strains.

Common lab strains (CC125+ and CC124-) were chosen for the cross because they are considered to be the wild type 137c genotype of which a cell-wall-less mutant was used to build the *C. reinhardtii* reference genome (Merchant et al., 2007). The CC125+/C124- genome would be highly similar to the reference genome expediting bioinformatics analysis.

Since CC125+ and CC124- were isolated near Amherst, Massachusetts, USA, the most geographically distant breeding pair available is the Japanese strains, J+ and J- which have been established as *bona fide C. reinhardtii* (Nakada et al., 2010). Not only would geographic isolation have contributed to genetic divergence between these two breeding pairs, but also the lab culturing of CC125+/CC124- since 1945 (Pröschold et al., 2005) means that CC125+/CC124- have been under very different selection pressures from J+/J- increasing genetic diversity.

3.1.3. Checked genetic divergence, initially with PCR based method and then with whole genome sequencing

My expectation, given the geographic isolation of the J and CC124/5 isolates, was that they would be diverse at the genetic level. To test this expectation I first used a PCR based method to check the genetic variation present at certain markers between the breeding pairs. I then analysed whole genome sequencing

to confirm the genetic divergence of the strains and to provide a genetic context to sRNA loci discussed in later chapters. The data from the genome sequencing was also compared to current knowledge on genetic variation in *C. reinhardtii* strains giving insight into the evolution of this delightful alga.

3.2. Results

3.2.1. Confirming the identify of C. *reinhardtii* strains

Before comparing the genomes of the parental strains in my study, I needed to confirm their identity. For species definition in freshwater algae, morphological and genetic markers are used. Previous inspection of strains CC125+, CC124-, J+, and J- confirmed morphological features associated with *C. reinhardtii* (Nakada et al., 2010).

To confirm species identity, the nuclear ribosomal RNA is suitable to use because it is normally one of the most stable genome features (Eickbush and Eickbush, 2007). In *C. reinhardtii*, for example, all previously characterised strains have identical 18S sequences (Nakada et al., 2010). 18S of my parental strains was PCR amplified, purified, sequenced, and then aligned to the reference 18S sequence (http://www.ebi.ac.uk/ena/data/view/FR865575). All four parental strains in this study also had 18S sequences that are identical to the reference 18S sequence (Figure 3-2) and I was therefore confident that they could be assigned to *C. reinhardtii*.

In contrast, internal transcribed spacer (ITS) regions of ribosomal RNA genes are less stable[3] than the 18S region and they can be used to differentiate strains within a species complex (Schultz et al., 2005; Young and Coleman, 2004). In *C. reinhardtii*, ITS2 exists in both a short and a long form due to an insertion/deletion (InDel) (Pröschold et al., 2005). *C. reinhardtii* has at least 200 copies of nuclear ribosomal repeats and different strains have different proportions of long and short ITS2. Previous analysis indicated that the strains isolated in Japan contain the long version only which has a 8bp insertion and a

[3] Specifically, ITS2 is highly diverged in sequence with an assumed conservation in structure.

complementary base pair change (T to A) (Nakada et al., 2010) relative to the reference genome sequence.

In my analysis I PCR-amplified, purified, and sequenced the ITS2 DNA. I then aligned the sequences to the reference ITS2 sequences (Nakada et al., 2010). It was reassuring that the CC125+ and CC124- ITS2 sequences were identical to the reference ITS2 for CC125+ while the ITS2 J+/J- specific differences were also identified in the ITS2 sequenced from the J+ and J- strains (Figure 3-2).

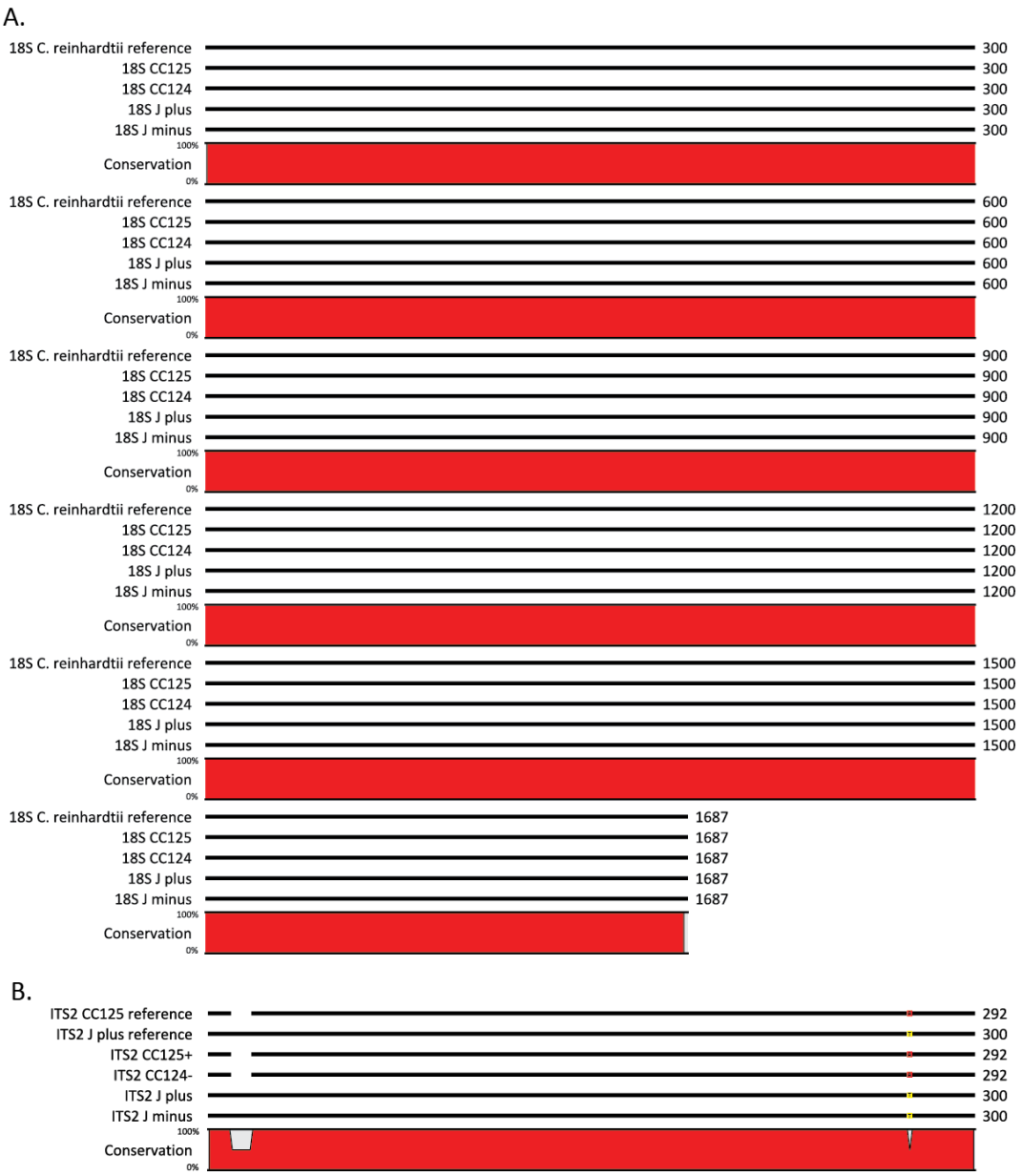


Figure 3-2: 18S and ITS2 sequence confirms species and strain identity
Nucleotide alignments with plots of sequence conservation. InDel shown by a gap in the sequence and drop in conservation. Nucleotide diversity is shown by a colour change and drop in conservation. **(A)** Partial 18S sequence from CC125+, CC124-, J+, and J- strains aligned to the reference 18S sequence (FR865575.1 from European Nucleotide Archive). **(B)** Partial ITS2 sequence from CC125+, CC124-, J+, and J- strains aligned to the reference ITS2 sequence for CC125+ (AJ749631.1 from ENA) and J+ (AB511842.1 from ENA).

The sequence of the mating type locus was also useful as part of an initial characterisation of the isolates. The *FUS1* gene is specific to the positive mating type (Ferris et al., 1996) and *MID1* to the negative mating type (Ferris and Goodenough, 1997). By amplifying a segment of each of these two mating type

specific genes and size separating them (Zamora et al., 2004), I confirmed the mating type of all four strains (Figure 3-3).

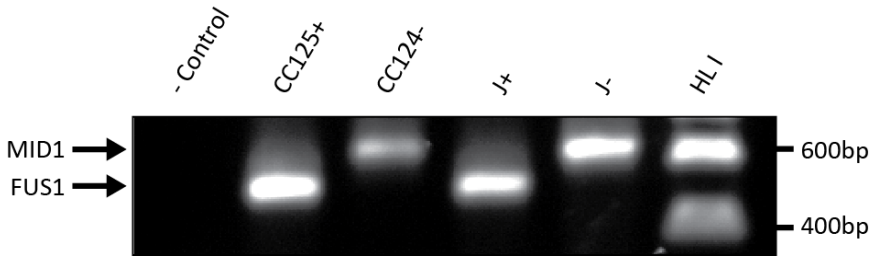


Figure 3-3: Mating type genotype determination of parental *C. reinhardtii* strains
Mating type PCR on genomic DNA extracted from CC125+, CC124-, J+, and J- strains. Plus mating type: FUS1 (516bp). Minus mating type: MID1 (622bp). Control: H_2O used as template. DNA ladder: Hyperladder I (HPI).

3.2.2. Genetic divergence suggested using PCR based approach

Based on geographic isolation and the ITS2 differences specific to the Japanese strains, I predicted that that the Japanese strains are the most genetically divergent strains to CC125+/CC124- out of the then available stocks. To test this prediction, I used mapping markers that were identified by the Chlamydomonas Collection Center as different between CC125+ and another U.S. field isolate, CC2290-. These markers rely on a PCR-based approach in which the primers used should amplify products of different sizes from these strains.

I chose the mapping marker approach to assess whether the Japanese strains shared more markers with lab strains or with the more recently isolated field isolate, CC2290-. My prediction that the Japanese strains are the most genetically divergent strains to CC125+/CC124- depends on their geographic distance, however genetic variation in some groups of isolates in North America is related to latitude (Jang and Ehrenreich, 2012). Thus the most Northerly isolated strain, CC2936+, could also be highly genetically diverse to CC125+/CC124-. For this reason, I included CC2936+ in my mapping marker analysis.

Mapping markers were amplified by PCR and size differentiated on a 1.5% agarose gel (Appendix 7.3). Markers in the Japanese strains and in the Canadian strain, CC2936+, were designated as either CC125+/CC124- like, CC2290- like, or

unique (in the cases where the marker was a different size to all three reference strains).

		Number of markers (n)	
		J+ & J-	CC2936+
Marker size similar to:	CC125+/CC124-	6	11
	CC2290-	11	7
	Unique	4	3
	Total tested	21	21

Table 3-1: Comparison of mapping marker divergence between *C. reinhardtii* strains
Mapping markers of J+ and J- compared to CC2936+ mapping markers. (PCR results can be found in Appendix 7.3)

This limited analysis of PCR markers was consistent with the proposed divergence of the J and CC125+/124- strains and indicated a similar divergence of CC2936+ to CC125+/CC124- strains (Table 3-1). However the resolution of this approach was inevitably restricted by the amount of information from each marker and the number of markers to be tested. I decided therefore to complete the characterisation of the strains by more extensive DNA sequence analysis.

3.2.3. Optimising DNA library preparation for whole genome sequencing

As a preliminary to genome sequence analysis of CC125+/CC124-/J+/J- I needed to optimise the TruSeq DNA library preparation protocol for the GC-rich DNA of *C. reinhardtii.* Specifically I optimized the sonication step where genomic DNA is size fractionated into 300-400 nt fragments as recommended by the Illumina TruSeq protocol. Low power sonication (recommended by the manufacturers of the Bioruptor Standard) was not sufficient to fractionate the DNA to sizes less than 400 nt without substantial loss of material (Figure 3-4). High power sonication was needed to satisfactorily fractionate the DNA so I performed a 90 min time course that established 15 min as the optimal sonication time to create 300-400 nt long DNA fragments (Figure 3-4). Under these conditions DNA from CC125+/CC124-/J+/J- was successfully size fractionated (Figure 3-4).

20 µg of the fractionated DNA was then used as the template for DNA library prep using the Illumina Truseq protocol. The final DNA libraries were sent to the Cancer Research Institute to be sequenced on an Illumina HiSeq (Illumina). The

resulting data needed to be checked for quality, length, and comparability before being processed. I used the FastQC program (Andrews, 2012) to confirm the high quality of the reads (Appendix 7.2) (Andrews, 2012). Additionally the data libraries were of comparable size (between 28 and 36 million reads).

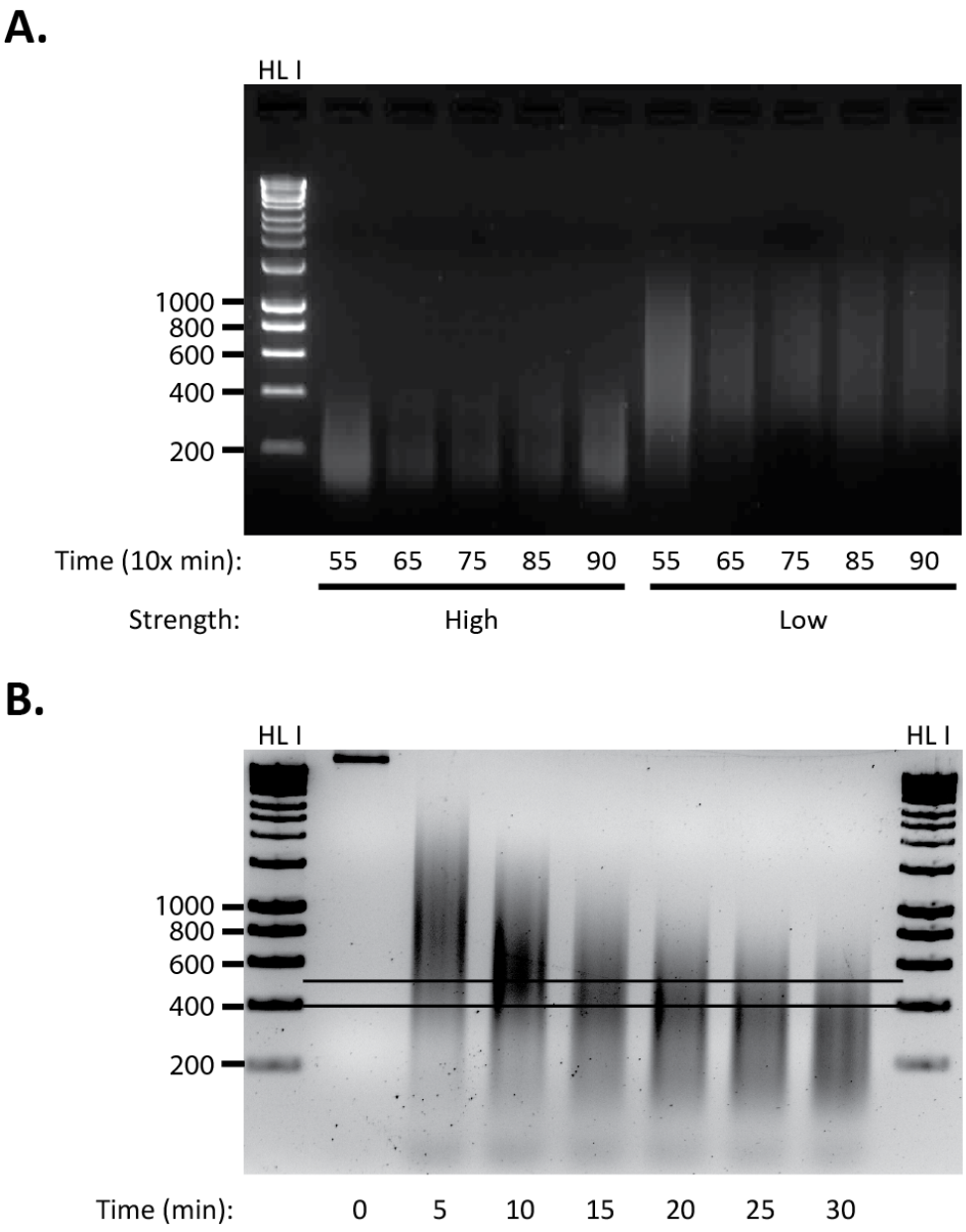


Figure 3-4: Optimizing fractionation of genomic DNA for library construction
Sonication of 20 µg of genomic DNA in 100µl extracted from J-. 5 µl of resulting product was run on a 1.5% agarose gel to compare sizes. 300 to 400 nt is the recommended size range for TruSeq DNA library construction. **(A)** Comparing the effect of different sonication strengths on fractionation size over time. **(B)** Time course to identify optimal length of high power sonication.

3.2.4. Aligning DNA libraries to the C. *reinhardtii* reference genome

Before alignment, I used Trimmomatic 0.27 to trim the adaptor sequences off library reads (Lohse et al., 2012). This program excises ends of reads depending on their complementarity to the Illumina Adapter Sequences while also trimming off windows of reads whose average quality falls below the threshold. I tested multiple trimming conditions, checking the quality and quantity of reads after trimming to ensure that the loss of reads was not excessive. The quality of the remaining reads was assessed using FASTQC (Andrews, 2012), a tool for quality control of next generation sequencing data. Amongst other statistics, it provides the user with a per base sequence quality graph when given a .fastq file (ie a .fasta file containing not only the individual sequencing reads but also the quality scores of each base in the reads). Generally a quality score of over 20 for a base is acceptable.

Using the optimal trimming parameters (allowing 1 mismatch to the Illumina adaptor, trimming nucleotides at the ends of reads if their quality score was below 20, and trimming any window of 4 nucleotides whose quality score was below 20) meant that in all libraries more than 93% of paired-end reads survived resulting sufficiently high quality DNA libraries (Appendix 7.2).

I wanted to confirm the origin of the reads by mapping them to the reference genome as contamination from other libraries in a flow cell is not unheard of (Zhang et al., 2012). I first aligned paired-end reads to the reference genome using bowtie (Langmead et al., 2009), a short read alignment program, and allowing no mismatches. This alignment resulted in very different percentages of mapped reads for CC125+/CC124- and J+/J- (around 58% and 7% respectively). As expected the Japanese libraries had a much lower percentage of aligned reads but the extent of the difference was surprising. Additionally, even the percentage of aligned reads in the North American strains seemed low. However, further analysis showed that ~20% of the reads in all libraries aligned to the *C. reinhardtii* chloroplast genome.

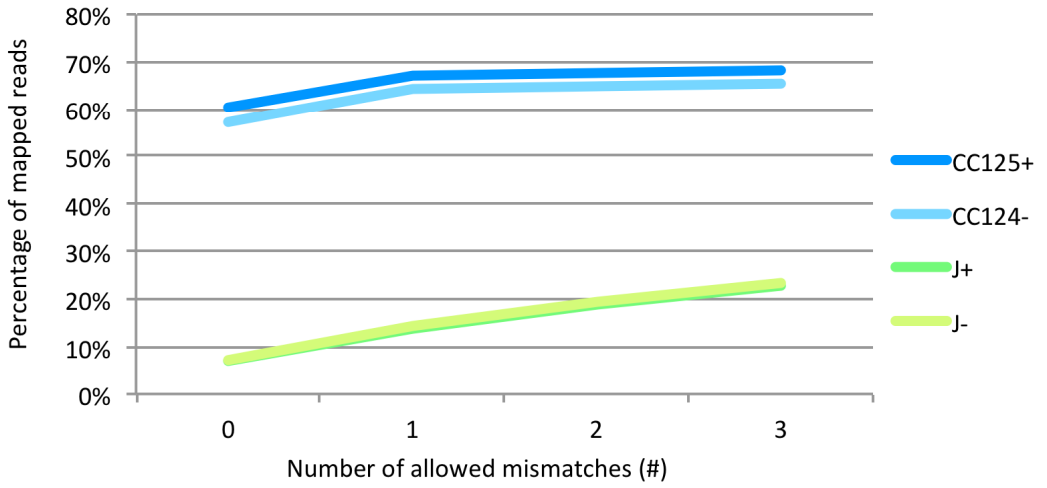


Table 3-2: Effect of mismatch allowance on percentage of mapped reads
Library reads from strains CC125+, CC124-, J+, and J- were aligned to the *C. reinhardtii* reference genome using bowtie and allowing 0, 1, 2, or 3 mismatches.

One explanation for the difference in number of reads mapping between the Japanese and North American strains was that genetic variation was interfering with the alignment. I experimented with bowtie, allowing different numbers of mismatches in the alignment. The percentage of mapped reads increased dramatically in the Japanese strains in response to allowing more mismatches (Table 3-2). In comparison, allowing one mismatch had a similar and equal affect on the percentage of mapped reads for the North American isolates however increasing the allowed number of mismatches further had little effect (Table 3-2). This suggests that many of the unmapped reads from the Japanese datasets are due to genetic variation to the reference genome.

My basic bowtie alignments suggested that the genomes were highly variable so I chose to work with the more advanced and sensitive paired-end aligning offered by the CLC genomics workbench. This algorithm allows mismatches based on quality scores and read coverage at that locus. Paired-end reads were mapped to the unmasked *C. reinhardtii* reference genome using the default mapping parameters. This resulted in a more comparable amount of reads mapping to the genome between the four strains (Table 3-3). Overall a satisfactory percentage of reads aligned to the genome and I further investigated the nucleotide divergence between the different strains.

	CLC
CC125+	71.39%
CC124-	70.17%
J+	64.28%
J-	64.61%

Table 3-3: CLC genomics workbench DNA alignments
Paired-end DNA reads were mapped using CLC Genomics Workbench 7.0 read map reads to reference tool.

3.2.5. Variant analysis confirms genetic divergence of Japanese strains

CC125+ and CC124- are North American isolates and closely related to the strain from which the current reference genome was assembled. I predicted therefore that genome sequence data from these isolates would have fewer short nucleotide variants (SNV) relative to the reference genome than in the J+ and J- data.

Variant detection was performed using the CLC genomics workbench 7.0. Single nucleotide polymorphisms (SNP), multi-nucleotide polymorphisms (MNP) and InDels were called using the quality-based variant detection tools. As expected there was a much larger incidence of variants of all types identified in the Japanese strains in comparison to CC125+ and CC124- (Table 3-4-A).

	Total Called Variants	SNV/kb
CC125	4617	0.0
CC124	83879	0.8
J+	1740260	15.7
J-	1494313	13.5

Table 3-4: Comparison of SNV frequency
The reference genome was used to call SNVs for CC125+, CC124-, J+, and J-. SNV frequencies at a genome level were calculated per kilobase.

3.2.6. Variation in strain RNA silencing components

Since biological pathways can be under different selection pressures, sets of proteins involved in different pathways can exhibit various levels of genetic variation. Thus the level and type of genetic variation shared between RNA

silencing pathway components can provide insight into the selection pressures acting on RNA silencing.

I compared the sequence of key RNA silencing components of the strains to see if there was any basis for them causing sRNA differential expression. The CLC genomics workbench quality based variant detection tool was used to call nucleotide divergence in the three AGO, three DCL, and RDR genes using the mapped DNA reads from CC125+, CC124-, J+, and J-. While in CC125+ only one SNV was identified in this group, many more SNVs were identified in the RNA silencing components of the other strains (Table 3-5).

	CC125+	CC124-	J+	J-
AGO1	0.0	0.0	20.9	25.5
AGO2	0.0	0.0	6.3	4.2
AGO3	0.0	5.8	9.5	6.5
DCL1	0.1	0.1	23.5	28.2
DCL2	0.0	2.6	31.4	24.1
DCL3	0.0	0.0	16.5	12.0
RDR	0.0	0.0	21.2	15.9

Table 3-5: Nucleotide diversity frequency in RNA silencing components
Nucleotide diversity frequency calculated for each RNA silencing component gene in terms of number of SNVs per kilobase across the genetic code for that gene. For actual numbers of SNVs see Appendix 7.5.

As expected, the RNA silencing components in J+ and J- were more divergent to the reference genome than those in CC125+ and CC124-. The CLC genomics quality based variant detection also uses gene annotations to predict which variants could result in non-synonymous mutations (mutations which result in an amino acid change in the protein). I compared the SNV frequencies for genetic, exonic, and non-synonymous exonic variants. When the SNV frequency (number of SNVs/kb) of the RNA silencing components genetic and exonic sequence is calculated, it is similar to the SNV frequency observed at a genomic level (Figure 3-5).

For the frequency of non-synonymous SNVs however, there is a difference between that observed in the RNA silencing components in comparison to the frequency in a control dataset of genes (Figure 3-5). This discrepancy in rates

could suggest a positive selection at work on RNA silencing components in *C. reinhardtii*. Also there is enough variation in the RNA silencing components of the Japanese strains to effect sRNA differential expression.

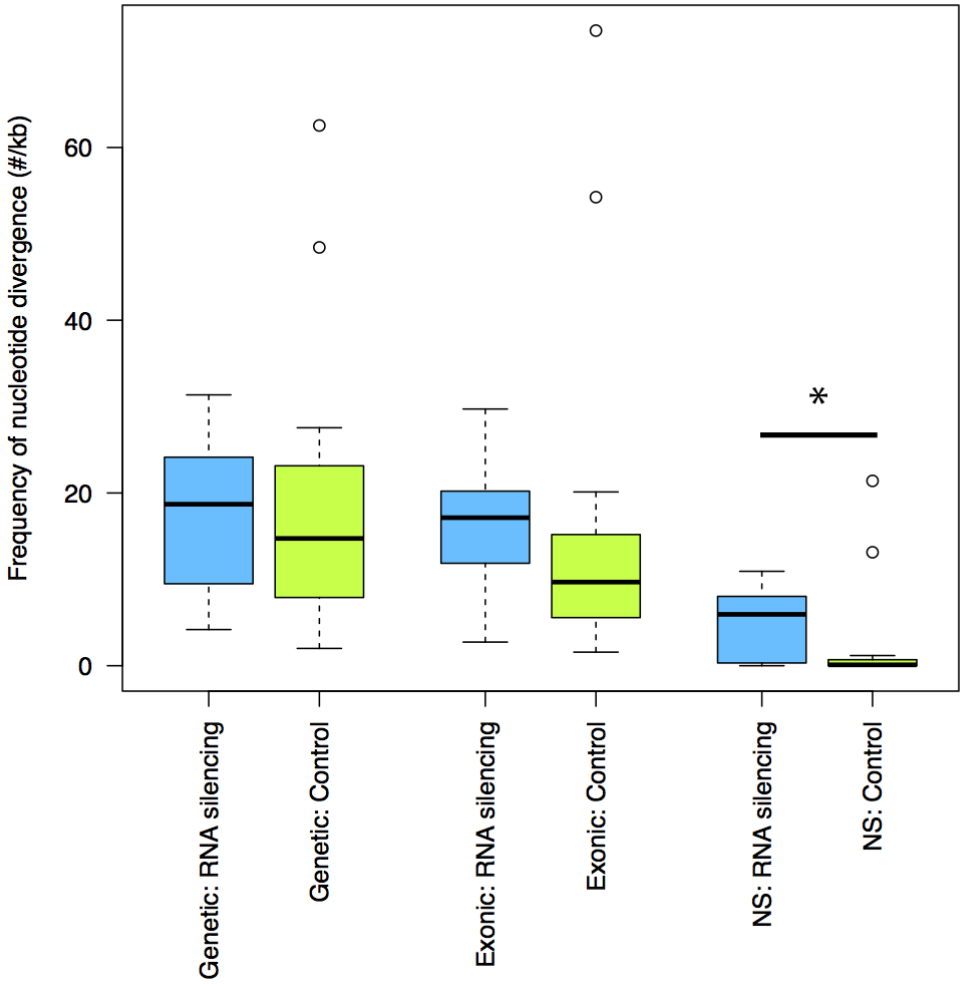


Figure 3-5: Comparing nucleotide diversity frequency between RNA silencing components and other genic components
Nucleotide diversity was identified using Quality Based Variant Calling from the CLC genomic workbench. Nucleotide divergence was calculated for the genetic sequence and the exonic sequence. CLC genomics workbench used to call non-synonymous (NS) mutations in the exonic regions. RNA silencing gene set: AGO1, AGO2, AGO3, DCL1, DCL2, DCL3, and RDR. Control gene set was the same set used in a study of genetic variation of the *C. reinhardtii* nuclear and mitochondrial genomes (Smith and Lee, 2008) (Appendix 7.4).

3.3. Discussion

3.3.1. C. *reinhardtii* is a true cosmopolitan alga

Confirmation of the species and ITS2 divergence in J+ and J- support Nakada's claims that J+ and J- are true *C. reinhardtii* isolated for the first time outside of North America (Nakada et al., 2010). In 2005, a phylogenetic study showed that *C. reinhardtii* strains claiming to be isolated from outside North America had identical ITS sequences to standard strains and were most likely "escaped" lab strains (Pröschold et al., 2005). An abundance of genetic variation in North American isolates suggests geographic site-specific variation allowing for the isolates with matching sequences of evolutionarily stable regions to be identified as standard lab strains (Jang and Ehrenreich, 2012; Pröschold et al., 2005). Pröschold et al suggested that *C. reinhardtii* had a relatively localized distribution restricted to North America (Pröschold et al., 2005).

The support for cosmopolitan distribution of *C. reinhardtii* was revived when J+ and J- were isolated in Japan, confirmed as *C. reinhardtii*, yet showing sequence variation in ribosomal sequences (Nakada et al., 2010). The increased geographic dispersal of *C. reinhardtii* has implications for evolution and population genetics studies of this alga. *C. reinhardtii* shows population subdivision and genetic diversity on a local geographic scale (Jang and Ehrenreich, 2012). On a global geographic scale, a similar study could provide insight into the spread and evolution of *C. reinhardtii*. Genetic variation also exhibits a mild correlation with the latitude of isolation of the isolates (Jang and Ehrenreich, 2012). The strength of this link could be further supported by repeating the same study with isolates from other regions of the world that also differ in latitude.

3.3.2. J+ and J- are genetically divergent to CC125+ and CC124-

Previous genotyping experiments have used different genetic markers (restriction fragment length polymorphisms, resequencing a restricted number of nuclear loci, extensive chloroplast resequencing) to compare the reference lab strains with the divergent wild isolates CC1952- and CC2290-, calculating a SNP frequency of ~27 SNPs/kb for the *C. reinhardtii* genome (Kathir et al., 2003b). The use of deep sequencing technologies to allow a genome wide analysis of SNPs

and the inclusion of more North American wild isolates resulted in a marker discovery rate of ~6 SNPs/kb (Jang and Ehrenreich, 2012). The earlier SNP frequency estimates were likely high because the strain, CC1952- is extremely divergent to the reference genome with a SNP frequency alone of 21.2 SNPs/kb (Lin et al., 2013).

Japanese strains J+ and J- had an elevated SNV frequency (~15 SNVs/kb) to the reference genome (Table 3-4) in comparison to the average for North American strains. This further supports the geographic isolation of the Japanese strains. Which makes it surprising that a strain isolated much closer to the reference genome strain, CC1952- (Figure 3-1), has a higher SNP frequency of 21.2 SNPs/kb (Lin et al., 2013).

As expected CC125+ and CC124- had an extremely low SNV frequency (Table 3-4) supporting previous calculations of 0.1 and 0.9 SNPs/kb respectively (Lin et al., 2013). The SNP data confirms my hypothesis that J+ and J- are extremely genetically divergent strains to CC125+ and CC124-, likely due to their geographic isolation. CC125+, CC124-, J+, and J- were the best-fitted strains at the time for the crossing experiment looking for transgressive effects.

3.3.3. Positive selection on RNA silencing components

The increased non-synonymous mutation rate in the RNA silencing components, as compared to the control set of genes (Table 3-5) implied a positive selection on the RNA silencing pathway in *C. reinhardtii*. Previous studies of RNA silencing evolution in the fly *Drosophila* have shown that the RNA silencing genes involved in antiviral defence experience high rate of adaptive selection (Obbard et al., 2011). The substitutions in Argonaute-2 between three species of *Drosophila* is overrepresented at the protein surface, suggesting that they may be of functional relevance (Obbard et al., 2011). The non-synonymous mutations in the Japanese RNA silencing components could also be of functional relevance although further structural analysis and cloning experiments are needed.

3.4. Acknowledgements

Dr. Takashi Nakada provided the J+/J- strains (labelled KkS0801B1 and KkS0801D2 respectively) and the primer sequences used to amplify 18S and ITS2 fragments. James Hadfield at the Cancer Research Institute UK facilitated the sequencing of the DNA libraries on an Illumina HiSeq.

4. Chapter four: Variation of sRNA profile between two strains, CC125+ and J-

Aspects of the sRNAs of two geographically and genetically divergent strains of C. reinhardtii, such as the size profile or GC content, were shown to be similar between the strains. Differentially represented and differentially expressed sRNA loci were identified between the strains.

4.1. Introduction

Before comparing the parental to the recombinant strains, I wanted to compare the parental strains, CC125+ and J-, to one another. In the previous chapter, I showed that CC125+ and J- were genetically divergent and in this chapter, I compare the sRNA populations of CC125+ and J-. Deep sequencing of the sRNAs in these strains allowed me to better describe the sRNA landscape in this alga and identify differentially expressed and differentially represented sRNA loci.

4.1.1. Genetic variation effects sRNA profile

The genetic divergence documented in Chapter 3 implies that the sRNA populations should be similarly divergent. Genetic variation such as SNVs, InDels, duplications, and translocations can all affect various aspects of RNA silencing.

SNPs between *Arabidopsis* accessions and between rice accessions can be found in pre-miRNAs, miRNAs, and miRNA targets (Meng et al., 2011a). Some SNPs found in pre-miRNAs could alter secondary structure of miRNA precursors and thus miRNA biogenesis (Meng et al., 2011a). SNPs in the target gene can also potentially disrupt regulation by miRNAs; in rice, a point mutation in the region of OsSPL14 gene complementary to miRNA OsmiR156 perturbs the translational repression by that miRNA (Jiao et al., 2010). Finally, SNPs in *cis* regulatory elements of miRNA genes can also modify miRNA expression (Meng et al., 2011b). Other sRNA loci could also potentially be affected by SNPs but this is not well documented. The SNPs and other SNVs between CC125+ and J- documented in Chapter 3 could therefore cause variation between the sRNAs in the strains.

Larger polymorphisms would have correspondingly greater effects on sRNAs. Some, such as deletions, could have obvious consequences, removing certain

sRNA loci from a lineage. Other genome rearrangements could create additional sRNA loci. For example duplications may create inverted repeats leading to a locus for hairpin-associated sRNAs.

4.1.2. Species variation of sRNAs

Comparing sRNAs between or within species has provided insight on how sRNA variation is created and also suggests that sRNAs may vary between the distantly related CC125+ and J- strains. Previous studies of sRNA variation have concentrated on miRNAs because they are the most characterized and understood of sRNAs.

Highly conserved miRNAs do exist, such as the 21 miRNAs conserved in Angiosperms (Jagadeeswaran et al., 2012). However many species-specific miRNAs have been identified in closely related species in animals (Mor and Shomron, 2013) and plants (Cuperus et al., 2011) implying a rapid birth-death rate of miRNAs. For example in *Curcurbits*, although most miRNAs were shared between four species, a striking number of miRNAs were differentially represented, including between the two closest related *Curcurbits* (Jagadeeswaran et al., 2012). Between *Arabidopsis thaliana* and *Arabidopsis lyrata* 13% miRNAs are species specific despite these lineages only diverging ~10 million years ago (Fahlgren et al., 2010).

However the number of species-specific miRNAs is likely to decrease as the sRNAs of more species are being sequenced (Meng et al., 2011a). Due to gaps in the sequence data from different plant lineages and unreliable miRNA annotation, it is difficult to conclusively label a miRNA as species specific, making it difficult to calculate the true birth-death rate of miRNAs. Models for miRNA evolution includes the inverted gene duplication theory, the random birth theory, and the theory that miRNAs arise from transposable elements (Piriyapongsa and Jordan, 2008; Shabalina and Koonin, 2008; Zhou et al., 2013).

Regardless of the challenges of elucidating the mechanism of miRNA evolution, there is diversity in miRNA expression across plant lineages and investigating miRNAs between closely related individuals has yielded valuable information concerning miRNA evolution. I aimed to do a similar comparison as past studies

using *C. reinhardtii* and including siRNAs in the analysis. The RNA silencing system characterized in *C. reinhardtii* has so far been shown to be most similar to that found in plants so there should be differential expression of miRNAs between CC125+ and J- as was observed in other higher plant sRNA comparisons.

4.1.3. Approach for comparison of CC125+ and J- sRNA profiles

I expected to see sRNA variation between CC125+ and J- in the form of loci present in both strains but differentially expressed, and in the form of species-specific loci. Additionally some sRNA loci should be highly conserved between the two strains indicative of biological function.

To compare the sRNA profiles of CC125+ and J-, I designed an experiment in which each strain was grown in triplicate. Sequencing sRNA libraries from biological triplicates were necessary to call significantly differentially expressed loci in the face of technical and biological variation.

I used various quality checks and the size profile of the sRNA libraries as a basic confirmation that the sequencing had succeeded. Alignment to the current *C. reinhardtii* reference genome (Merchant et al., 2007) was then used to establish how genetic proximity to the reference genome strain could affect later sRNA analysis. I also analysed the GC content and the association of sRNAs to various genomic annotations.

As in previous studies, miRNAs were predicted for the two strains and then their expression compared. However in order to compare siRNAs, I had to use an approach where genomic sRNA loci, not individual sequences were compared. SegmentSeq (Hardcastle et al., 2012) was used to call sRNA loci allowing me to use further sRNA loci classification tools to describe the various classes of sRNA loci present in CC125+ and J-. SegmentSeq also tested for differential sRNA representation between the strains. Using this approach I identified many differentially represented loci between the two strains and checked the genetic background of these loci to try and establish a cause for the differential representation. This led to the identification of differentially expressed sRNA loci in which the genetic background is present in both CC125+ and J-.

4.2. Results

4.2.1. Assessing library quality

Quality of the raw individual reads CC125+ and J- sRNA had to be confirmed before the sequencing libraries could be compared. The quality of the raw reads was first established using the same quality control tool as was used on DNA libraries (Appendix 7.6). FASTQC (Andrews, 2012), provides the user with a per base sequence quality graph when given a .fastq file (i.e. similar to a .fasta file, containing not only the individual sequencing reads but also the quality scores of each base in the reads). The per base sequence quality graphs from FASTQC of the CC125+ and J- libraries showed that the libraries were of sufficient quality for further analysis (Appendix 7.6).

Another measure of library quality is the similarity of the replicate libraries to one another. A comparison of the log of the expression of individual loci (further description of how loci were identified can be found in 4.2.5) showed that the replicates were sufficiently similar to one another (Appendix 7.7).

The reads were then loaded into the DCB pipeline (unpublished software) for trimming off of adapters, counting, and alignment to the latest genome and the transcriptome (v5.3.1) (Merchant et al., 2007). The pipeline outputs alignment files and statistics for the aligned reads (e.g. number aligned, number not aligned, and GC content). Statistics were created both for redundant and non-redundant libraries.

Redundant libraries contain all the sRNA reads including duplicate reads. Using redundant reads allows counts to be given to sRNA species. In non-redundant libraries each sRNA read occurs only once and duplicates are deleted. A non-redundant list of sRNAs essentially gives a list of all the species of sRNAs existent in the library and does not associate counts with those species. Comparing redundant and non-redundant libraries can show whether patterns of sRNA expression are influenced by the high level expression of only a few sRNAs. For example, an extremely highly expressed sRNA species could skew a size profile in the redundant library analysis.

4.2.2. sRNA length distributions of C. *reinhardtii* highly conserved in CC125+ and J-

The DCB pipeline output includes a size distribution of the sRNA reads from both redundant reads and non-redundant reads. In both the redundant and non-redundant libraries the 21nt size class of sRNA was most dominant (Figure 4-1). Additionally sRNAs in both CC125+ and J- showed a similar preference for U and A as the 5' end nucleotide (Figure 4-1). Overall the sRNA length distributions were similar to those reported previously in *C. reinhardtii* and did not differ between strains (Figure 4-1).

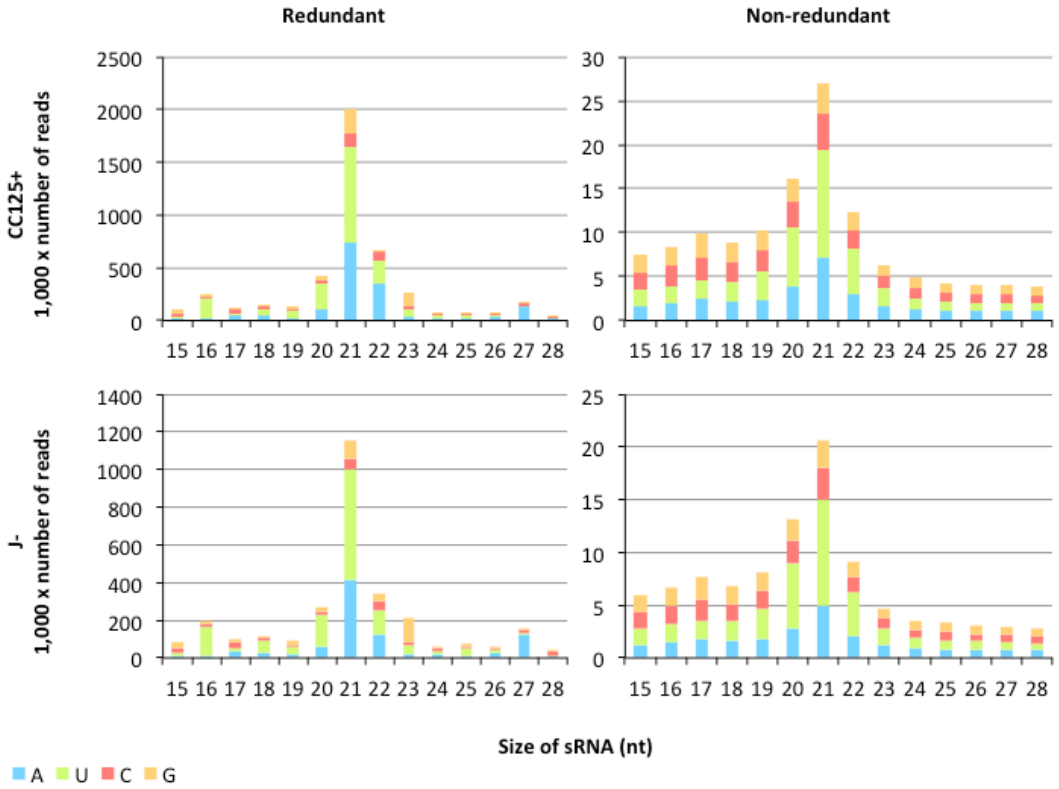


Figure 4-1: Size distributions of sRNA libraries
Size distribution's of sRNA reads from CC125+ and J- aligning to *C. reinhardtii* reference genome (Merchant et al., 2007). The 5' end nucleotide percentage is shown by colour of the bar. Redundant: all reads including counts. Non-redundant: counts are discarded.

4.2.3. Alignment to the genome

For further analyses I only used sequence reads that aligned to the *C. reinhardtii* genome. As CC125+ is essentially equivalent to the genome reference strain (CC503+), I expected that it would have a higher proportion of aligned sequences than the J- sRNA libraries.

The DCB pipeline aligned the sRNA libraries (no mismatches allowed) and, as expected, fewer J- reads aligned to the genome or the transcriptome (Figure 4-2). Aligning the sRNA libraries using the short read aligner program, Bowtie (Langmead et al., 2009), and allowing various levels of mismatches resulted in a larger increase in aligned read number in J- than in CC125+. It is likely therefore that the lower aligned read percentage in the J- strain data is due to greater genetic variation from the reference genome than with CC125+ (Figure 4-2).

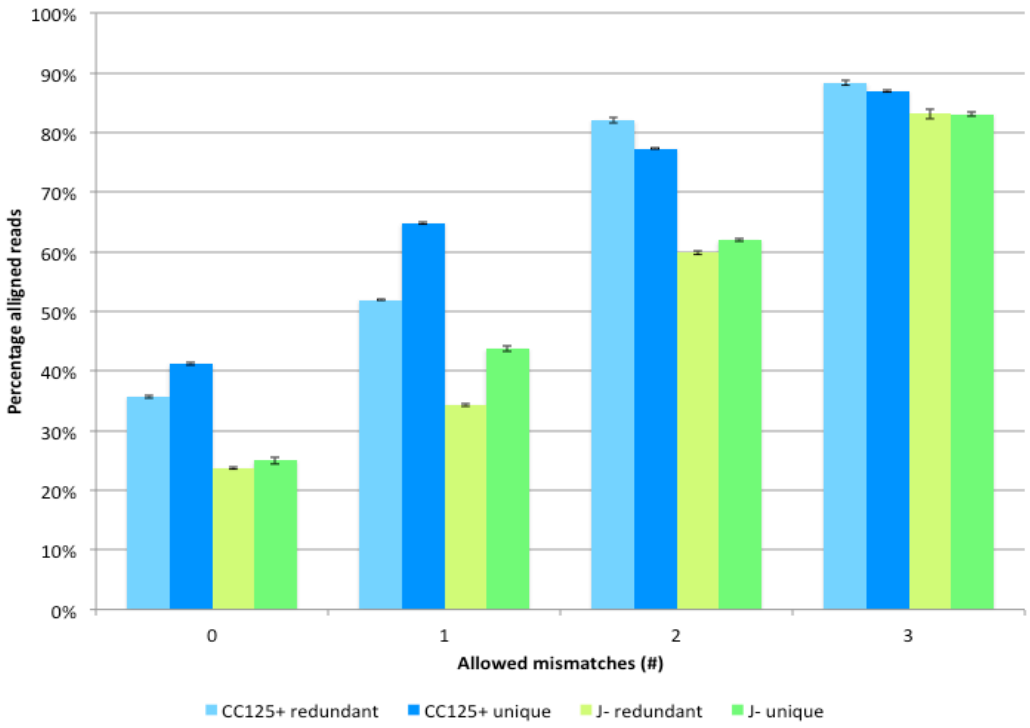


Figure 4-2: Effect of mismatches on sRNA read alignment
Small RNA reads from CC125+ and J- were aligned to the *C. reinhardtii* genome using Bowtie (Langmead et al., 2009) with 0, 1, 2, or 3 mismatches allowed. The resulting percentages of reads aligned to the genome are plotted above. R: Redundant, all reads including counts. NR: Non-redundant, counts are discarded. Error bars denote standard error of replicate values.

4.2.4. Small RNA libraries are relatively GC poor

Surprisingly, the sRNA libraries (both aligned and non-aligned) are relatively GC poor in comparison to the high GC content of the *C. reinhardtii* nuclear genome and transcriptome (Figure 4-3). With ~64% GC content, *C. reinhardtii* has an extremely GC rich genome in comparison to multicellular organisms (Pessia et al., 2012). The depletion of GC richness applies also to the sRNAs that aligned only to

the transcriptome and to the sRNAs of both CC125+ and J-, in redundant and non-redundant contexts (Figure 4-3).

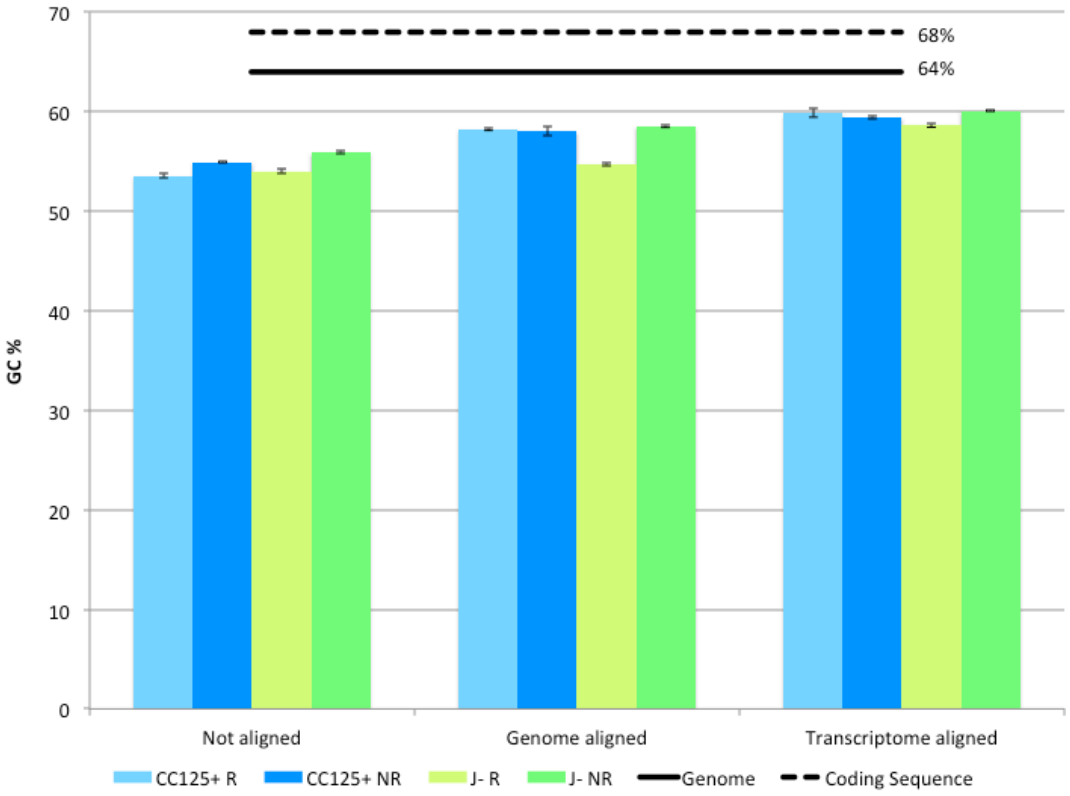


Figure 4-3: Percentage GC
GC% of CC125+ and J- strain sRNA reads are portrayed. GC% increases for reads aligned to the nuclear genome and transcriptome however they still are lower than the GC% of the genome and transcriptome. GC% for genome and coding sequence taken from Merchant et al., 2007. R: Redundant, all reads including counts. NR: Non-redundant, counts are discarded. Error bars show standard error between replicates.

4.2.5. Predicting sRNA loci

To predict sRNA loci, I used SegmentSeq (Hardcastle et al., 2012): using aligned sRNA reads, this program looks for genomic regions with a high density of read matches (taking into account replicate data sets), assigns them into segments, and returns the statistical significance of those segments being loci. As input to SegmentSeq, I used Bowtie (Langmead et al., 2009) to align CC125+ and J- replicate data to the *C. reinhardtii* reference genome (Merchant et al., 2007) allowing no mismatches.

After assigning the sRNA reads into loci, I then used SegmentSeq (Hardcastle and Kelly, 2010) to call loci in both samples assigning a likelihood value to each locus.

Empirical Bayesian methods are used in SegmentSeq to determine posterior likelihoods that a segment is truly a locus allowing for reproducibility of data within replicate groups to be taken into account (Hardcastle et al., 2012). From the initial counting of sRNA loci, it seems CC125+ has more sRNA loci than J-, whether the analysis was based on all reads or on reads which only matched once (SegmentSeq analyses made using uniquely matching reads still take into account the counts for those uniquely matching reads) (Table 4-1).

SegmentSeq identifies both miRNA and siRNA loci and does not differentiate between the two types. I used other bioinformatics tools to identify classes of sRNA loci. From the UEA Toolkit, the program miRProf was used to identify and count the known miRNAs in the CC125+ and J- libraries (Moxon et al., 2008). The list of known miRNAs is taken from the online database for miRNAs, miRBase (Griffiths-Jones et al., 2006). Not surprisingly, nearly all predicted miRNAs were present in CC125+ libraries but only 26 out of 50 were identified in J- (Table 4-1); currently known miRNAs were identified in CC503, essentially an equivalent strain to CC125+.

From the same online toolkit, miRCat was used to predict novel miRNAs using the default parameters (Appendix 7.8) (Moxon et al., 2008). The default parameters for miRCat are extremely stringent, aiming to exclude inverted repeat associated sRNAs. Even then there tend to be a lot of false positive miRNAs from such prediction tools (Kozomara and Griffiths-Jones, 2014) and so I prefer to class these predicted miRNAs as hairpin-associated sRNAs. Similarly to known miRNAs, miRCat also predicted more miRNAs in CC125+ than in J- (Table 4-1).

Finally, PhaseR (http://www.plantsci.cam.ac.uk/bioinformatics/phaser) was used to identify phased loci from the SegmentSeq dataset of CC125+ and J- loci. Dr Bruno Santos performed the PhaseR analysis. PhaseR distinguishes phasing of sRNA reads, in which sRNA reads align to a section of the genome in a head-to-tail arrangement, starting from a specific nucleotide, from random distributions. Surprisingly, there were more phased sRNA loci in J- than in CC125+ (Table 4-1). Phasing usually indicates secondary sRNA production (Fei et al., 2013). Although,

so far, secondary sRNA production has not been confirmed in *C. reinhardtii*, phased loci have been identified in previous datasets (Molnar et al., 2007a) and the recent annotation of an RDR homologue in the *C. reinhardtii* genome suggests that this mechanism exists (Merchant et al., 2007).

	CC125+	J-
SegmentSeq loci	7436	3278
SegmentSeq loci unique	1124	610
Known miRNAs	39	26
Predicted miRNAs	282	78
Phased loci	15	372

Table 4-1: Comparing numbers of sRNA loci types between CC125+ and J-
SegmentSeq loci were counted as all loci whose likelihood scores were higher than 0.9 in that strains (Hardcastle et al., 2012). The UEA tools miRProf and miRCat were used to identify and count known and predicted miRNAs respectively from sRNA libraries (Moxon et al., 2008). Default parameters were used for miRCat. Phased loci were identified by Dr Bruno Santos using the PhaseR algorithm. Loci not associated with a miRNA or phasing were labelled as non-classified.

4.2.6. sRNA reads overlap with genomic annotations in C. *reinhardtii*

The low GC content of the sRNA datasets (Figure 4-3) could be linked to sRNA locus location as different elements in the *C. reinhardtii* genome have different GC contents (Labadorf et al., 2010). Different lengths of sRNAs, indicative of different types, also correlate with different locations in the genome of maize (Barber et al., 2012). To classify the location of sRNA loci, I used the transcriptome (including coding DNA sequences (CDS), five prime untranslated regions (5'-UTR), three prime untranslated regions (3'-UTR), and genes) and repetitive element annotations (discounting simple repeats) from Phytozome v9.1 (Merchant et al., 2007) and checked for overlaps between the annotation ranges and sRNAs aligned to the genome. Because many of the annotation elements overlap, duplicate overlaps for an sRNA read were removed. Giving priority to different annotations in this duplication removal allowed me to calculate the number of sRNA reads aligning to introns or aligning to repeats within a gene.

Using the annotations I was also able to calculate the per cent coverage of the genome by different genetic elements (Figure 4-4). The *C. reinhardtii* genome is very gene rich with over 80% of the genome associated with genic regions and

only ~10% of the genome being taken up by repetitive elements (Figure 4-4). Despite the prevalence of genic elements in the genome, sRNA reads aligned primarily to repeats and intergenic regions (Figure 4-4). And of the sRNA reads aligning to repeats, more aligned to intergenic repeats than to repeats associated with genes (Figure 4-4). As expected, uniquely matching reads associated less with repetitive elements, aligning primarily to genic elements (Figure 4-4).

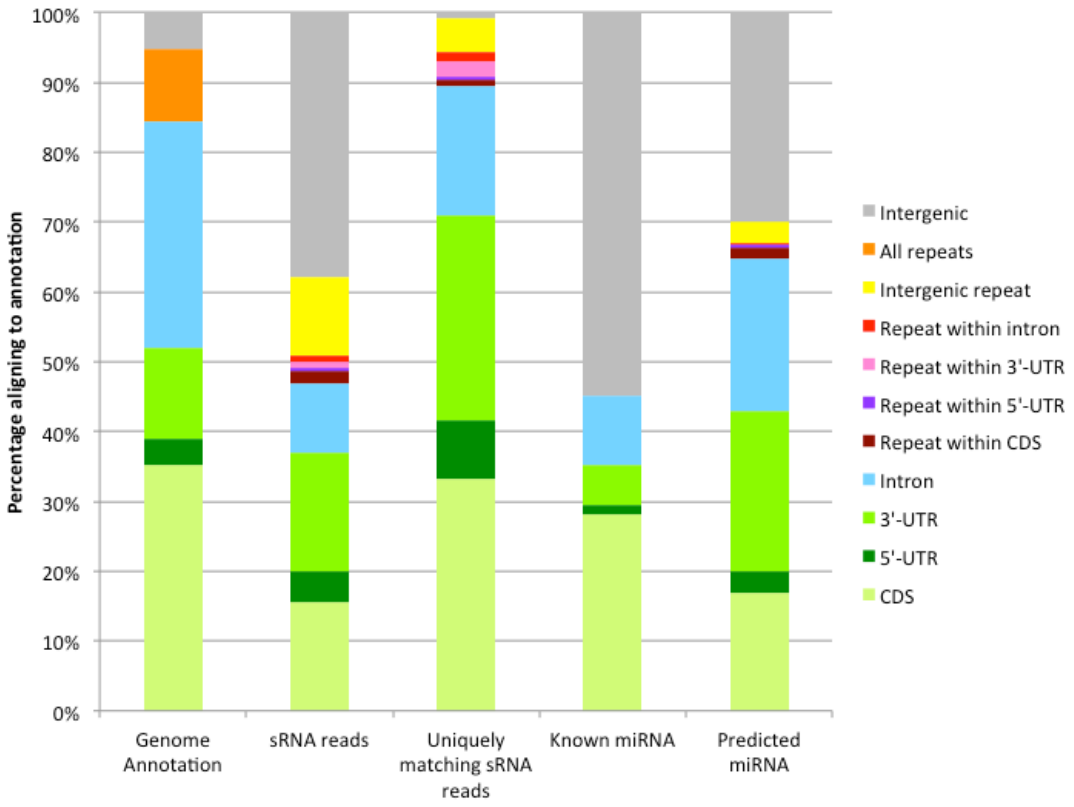


Figure 4-4: Locations of sRNA loci and reads
Bar graph on the left shows the percentage of the genome taken up by the various annotations. The rest of the bar graphs show the percentages of sRNA reads that overlap with different types of annotations. Transcriptome and repetitive element annotations for the *C. reinhardtii* genome were taken from Phytozome v9.1 (Merchant et al., 2007). sRNA reads are non-redundant.

It is possible that different types of sRNA loci associate more frequently with different elements within the genome. To see whether different loci types overlap with different genetic elements I performed the overlap analysis with all known miRNAs and predicted miRNAs from CC125+ and J- libraries (miRCat, default parameters).

No known miRNAs aligned to repetitive elements (Figure 4-4). Over half of the known *C. reinhardtii* miRNAs listed in miRBase (Kozomara and Griffiths-Jones, 2014) aligned to intergenic regions; the rest aligned to genic elements and primarily within that to the CDS (Figure 4-4). The predicted miRNAs showed a very similar pattern of association with annotations although fewer of the predicted miRNAs aligned to intergenic regions and an increased number of predicted miRNAs aligned to introns, exons, and the 3'-UTR (Figure 4-4).

4.2.7. Comparing CC125+ and J- phased sRNA loci overlap with genomic annotations

Since there are many more phased loci in J- in comparison to CC125+ (Table 4-1) I compared the locations of the phased loci in the two strains. I performed the overlap analysis with same annotations as used with sRNA read overlap analysis (Chapter 4.2.6), but instead of sRNA reads sequences as input, I used the sequences of sRNA loci predicted by a SegmentSeq analysis of CC125+ and J- reads. Interestingly, sRNA loci associated mainly with repetitive elements (Figure 4-5).

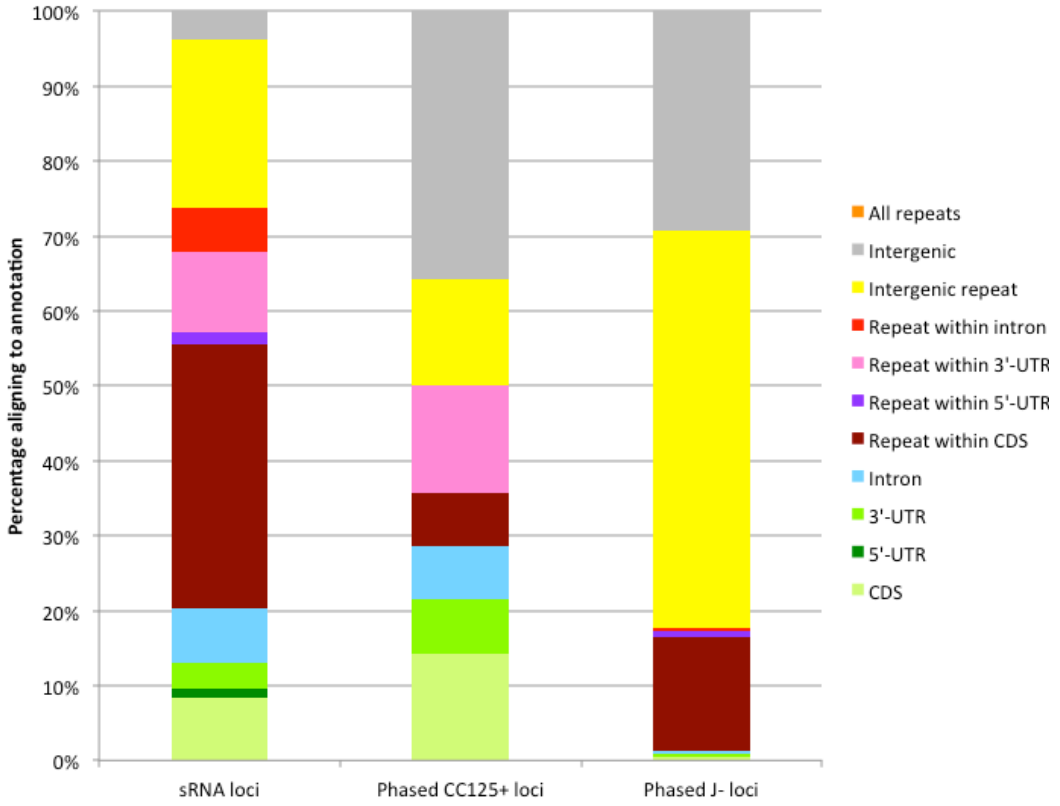


Figure 4-5: Locations of sRNA loci and phased sRNA loci
The bars show the percentage of sRNA loci that overlap with types of genomic annotations. Transcriptome and repetitive element annotations for the *C. reinhardtii* genome were taken from Phytozome v9.1 (Merchant et al., 2007). Simple repeats were excluded from the repetitive element annotation.

Phased sRNA loci in CC125+ aligned to genomic annotations in different proportions to phased J- sRNA loci (Figure 4-5). In both strains phased loci seemed to associate with intergenic regions to a similar extent (Figure 4-5). In J-, fewer phased loci overlapped with genic elements; there was a corresponding increase in the number overlapping with repeats (Figure 4-5).

It was possible that extra phased loci in J- associate with repetitive elements in general or that those phased loci are associated with a specific class of repetitive element. I repeated the overlap analysis, excluding transcriptome annotations, using the various types of repetitive elements as annotations. Many of the repetitive elements in the genome of *C. reinhardtii* are unclassified; the next biggest class of repetitive elements is that of non-LTR transposons (Jurka et al., 2005).

Of the phased sRNA loci in J- that align to repetitive elements, many aligned to unclassified repetitive elements (Table 4-2). However most phased sRNA loci in J- align to a single type of repetitive element, L1-1 (Table 4-2). L1-1 is a non-LTR retrotransposon whose several hundred copies constitute nearly 1% of the *C. reinhardtii* genome (Jurka et al., 2005).

Repetitive element	Type	CC125+	J-	
intergenic	Non-repetitive	9	113	
EnSpm-N2	DNA transposon	1	1	
EnSpm-N3	DNA transposon	0	2	
L1-1	Non-LTR retrotransposon	0	108	*
RandI-1	Non-LTR retrotransposon	0	1	
RandI-2	Non-LTR retrotransposon	0	2	
RandI-4	Non-LTR retrotransposon	0	1	
RandI-6	Non-LTR retrotransposon	0	2	
SINEX-2	Non-LTR retrotransposon	0	1	
SINEX-3	Non-LTR retrotransposon	0	3	
rnd-1_family-134	Unclassified	0	1	
rnd-1_family-15	Unclassified	0	45	
rnd-1_family-170	Unclassified	2	0	
rnd-1_family-23	Unclassified	0	3	
rnd-1_family-30	Unclassified	0	2	
rnd-1_family-312	Unclassified	0	1	
rnd-1_family-78	Unclassified	0	14	
rnd-3_family-428	Unclassified	0	19	
rnd-4_family-1051	Unclassified	0	13	
rnd-4_family-1270	Unclassified	1	0	
rnd-4_family-1590	Unclassified	0	14	
rnd-4_family-742	Unclassified	0	5	
rnd-5_family-145	Unclassified	0	1	
rnd-5_family-2710	Unclassified	0	1	
rnd-5_family-3228	Unclassified	0	1	
rnd-5_family-392	Unclassified	0	1	
rnd-5_family-701	Unclassified	1	0	
rnd-5_family-918	Unclassified	0	14	

Table 4-2: Phased sRNA loci association with repetitive elements
Table of the number of phased sRNA loci that overlapped with various types of repetitive elements. The list of repetitive elements and their classes were taken from RepBase (Jurka et al., 2005). The * denotes the L1-1 in which most phased sRNA loci in J- were found.

The increase in phased sRNA loci associated with the L1-1 retrotransposon could be caused by increased regulation of the repetitive element in the J- genome or

be due to an increased L1-1 copy number in J-. A preliminary mapping frequency analysis, where the number of DNA reads from CC125+ and J- mapping to L1 retrotransposons was compared, suggests there might be more copies of the L1-1 retrotransposon in the J- genome (Figure 4-6).

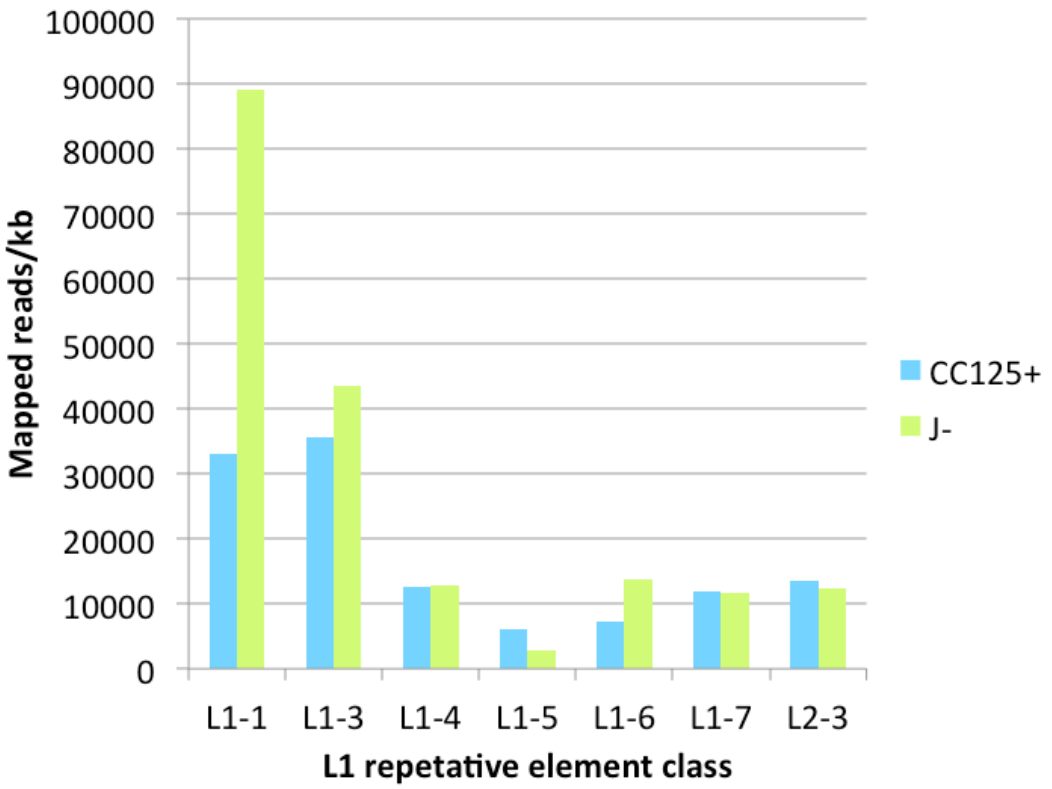


Figure 4-6: DNA coverage of L1-1
Using the CLC genomics workbench map to reference, DNA libraries for CC125+ and J- were mapped to different L1 repetitive element classes. Normalized read counts were used to calculate mapped reads/kb for each class of L1 repetitive element.

4.2.8. Identifying differentially represented sRNA loci

Differentially represented loci between the CC125+ and J- sRNA libraries were identified as those with a higher than 0.9 likelihood score of being differentially represented according to SegmentSeq analysis.

Of sRNA loci, nearly 70% of loci were identified as significantly differentially represented between CC125+ and J- (Figure 4-7). This percentage remains the same whether using all redundant reads or just uniquely matching (only aligning to the genome once) redundant reads (Figure 4-7). Equally unchanged, is the ~9% of loci which had a higher than 0.9 likelihood of being conserved (present

and with the same sRNA expression in both CC125+ and J-) (Figure 4-7). The rest of the loci are labelled as “unclassified”. There is likely similar expression in the two strains but the replicates are either too variable or expression is too low for the locus to be labelled as conserved or differentially represented with a high enough likelihood.

To check whether conservation or divergence of sRNA loci was associated with the type of genetic background, I performed an overlap analysis between differentially represented and conserved sRNA loci, and genomic annotations. Differential representation of sRNA loci exhibited a similar pattern of association with genomic annotations as total sRNA loci with perhaps only a slight increase in alignment to genic elements (Figure 4-7). When only uniquely matching redundant reads are used there was as expected a drop in per cent aligning to repetitive elements (Figure 4-7). Slightly surprising is the association between conserved sRNA loci (whether identified using all redundant reads or uniquely matching redundant reads) and intergenic regions (Figure 4-7).

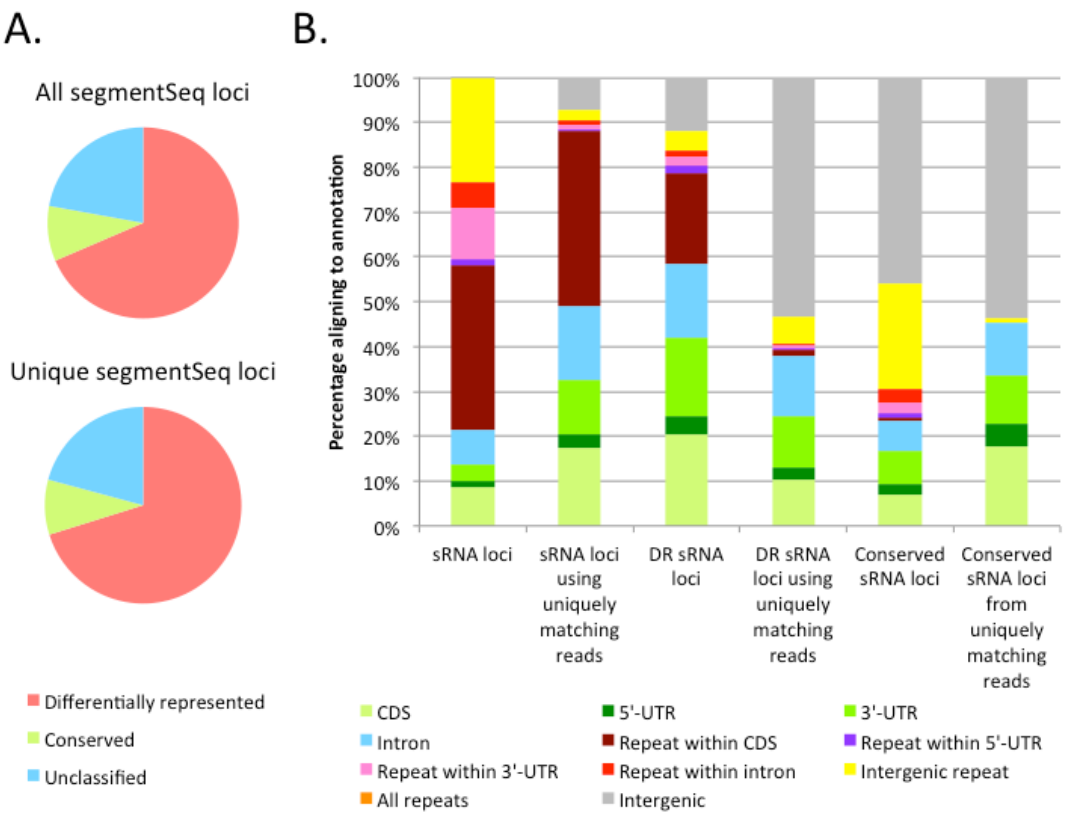


Figure 4-7: Differentially represented SegmentSeq loci
(A) Pie charts of sRNA loci in CC125+ and J- predicted by SegmentSeq using either all sRNA reads or only reads that match once to the reference genome (unique). Loci were classed either as differentially represented between CC125+ and J- sRNA libraries (likelihood > 0.09), conserved between the two (likelihood > 0.09), or unclassified (if neither differentially expressed or conserved). **(B)** Bar graphs show the percentages of sRNA loci that overlap with different types of annotations using either all redundant reads or uniquely matching redundant reads. Differentially represented (DR) sRNA loci and conserved sRNA loci with >0.9 likelihoods were identified using SegmentSeq. Transcriptome and repetitive element annotations for the *C. reinhardtii* genome were taken from Phytozome v9.1 (Merchant et al., 2007).

4.2.9. Causes of differential representation

Differential representation of sRNA loci between CC125+ and J- sRNA libraries could be classed by the cause for the differing levels of sRNA expression at a locus. Some differential representation will be due to the lack of the genetic background of the sRNA locus in one of the two strains. This absence of genetic background in a strain will likely be due to an InDel and so I labelled these loci as InDel-associated differentially represented sRNA loci (Figure 4-8-A).

For other differentially represented sRNA loci, the genetic background will be present in both strains. In these cases, the differential representation of uniquely

matching sRNA reads will be likely due to differential expression of sRNAs between the two strains. Thus, I termed differentially represented sRNA loci in which the genetic background is present in both strains as differentially expressed sRNA loci (Figure 4-8-B).

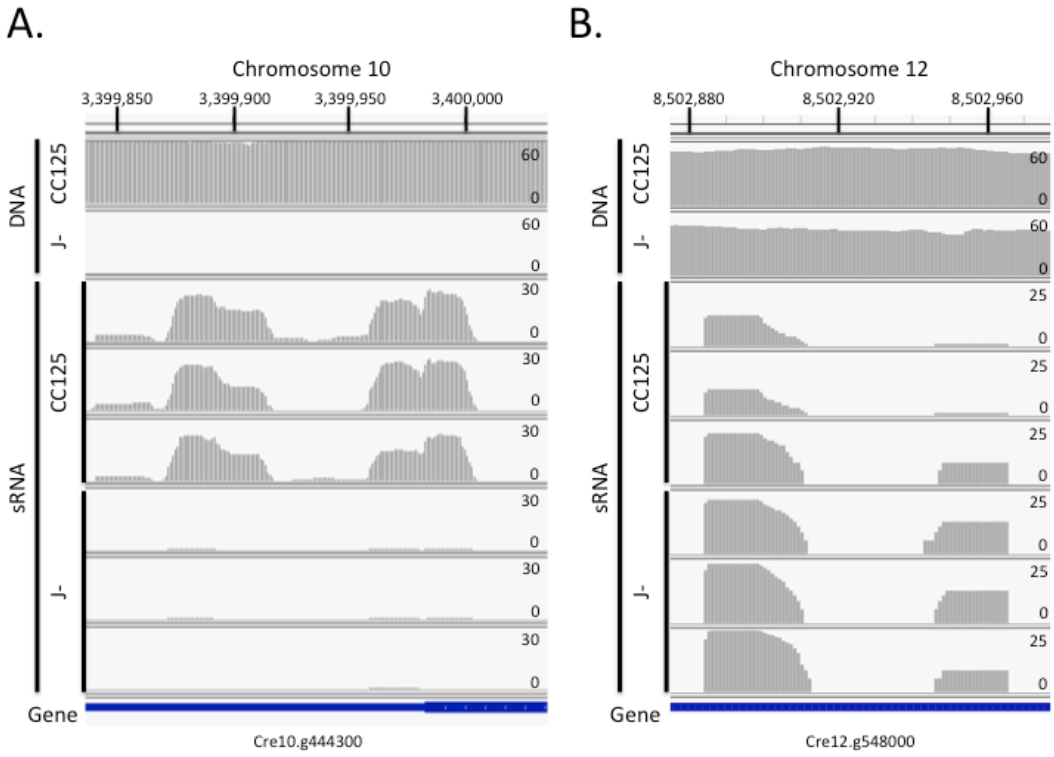


Figure 4-8: Causes of differential representation
Alignments of two examples of differentially represented loci. DNA reads were mapped to the genome and the log coverage is shown above. sRNA reads were also mapped to the genome (allowing mismatches), the expression normalized, and the coverage shown above. **(A)** An Indel-associated differentially represented sRNA loci in which the differential representation of sRNA reads in the sequencing libraries was caused by the lack of the underlying DNA sequence of the locus in the J- strain. **(B)** A differentially expressed sRNA locus where the underlying DNA sequence is present in both strains.

To approximate the proportion of InDel-associated differential represented sRNA loci versus the level of differential expressed sRNA loci, I visually checked the DNA alignments for the top 223 differentially represented sRNA loci from the SegmentSeq analysis using only uniquely matching reads.

As this analysis was performed using alignments that did not allow any mismatches, for checking the genetic background I used the CLC genomics sRNA mapping which allows mismatches based upon the quality of read sequences and

read coverage. This meant I was also able to label some differentially represented sRNA loci as false positives if novel aligned reads distorted the sRNA representation at that locus. Of the 223 differentially represented sRNA loci under 6% were labelled as false positives (Figure 4-9). The majority of differentially represented sRNA loci assessed were identified as differentially expressed, although ~32% of differential represented sRNA loci were due to InDel-associated differential representation (Figure 4-9).

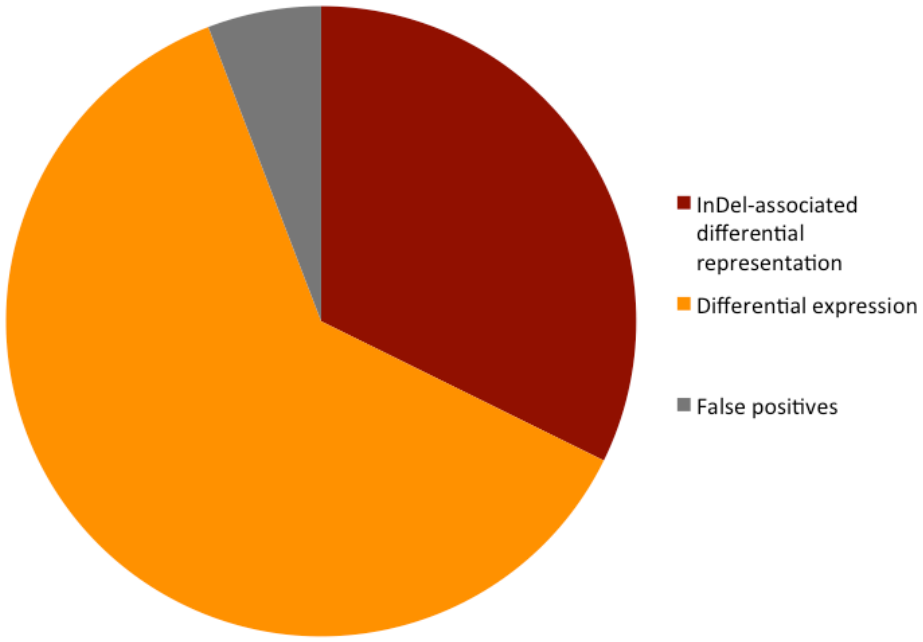


Figure 4-9: Proportion of InDel-associated differential representation and differential expression of sRNA loci between CC125+ and J-
A pie chart showing the proportion of the top 223 differentially represented (as identified in a SegmentSeq analysis using uniquely matching reads) which are either InDel-associated differentially represented sRNA loci, true differentially expressed sRNA loci or false positives.

4.2.10. Causes of differential expression

Differential expression of sRNA loci (in which the genetic background of the locus is present in both strains) could have multiple causes. Differential sRNA expression could also be caused by genetic differences between the strains. Genetic variation within the sRNA locus, such as SNPs, could disrupt sRNA expression (Figure 4-10-A). Additionally genetic variation near to the

differentially expressed sRNA locus, such as an InDel, could also alter expression (Figure 4-10-B).

I identified differentially expressed sRNA loci in which the genetic background is identical in both strains (Figure 4-8-B). In these cases, genetic variation found outside of the locus could still be causing the differential expression. However it is also possible that epigenetic variation at the locus, such as DNA methylation or histone modifications could be causing the differential expression.

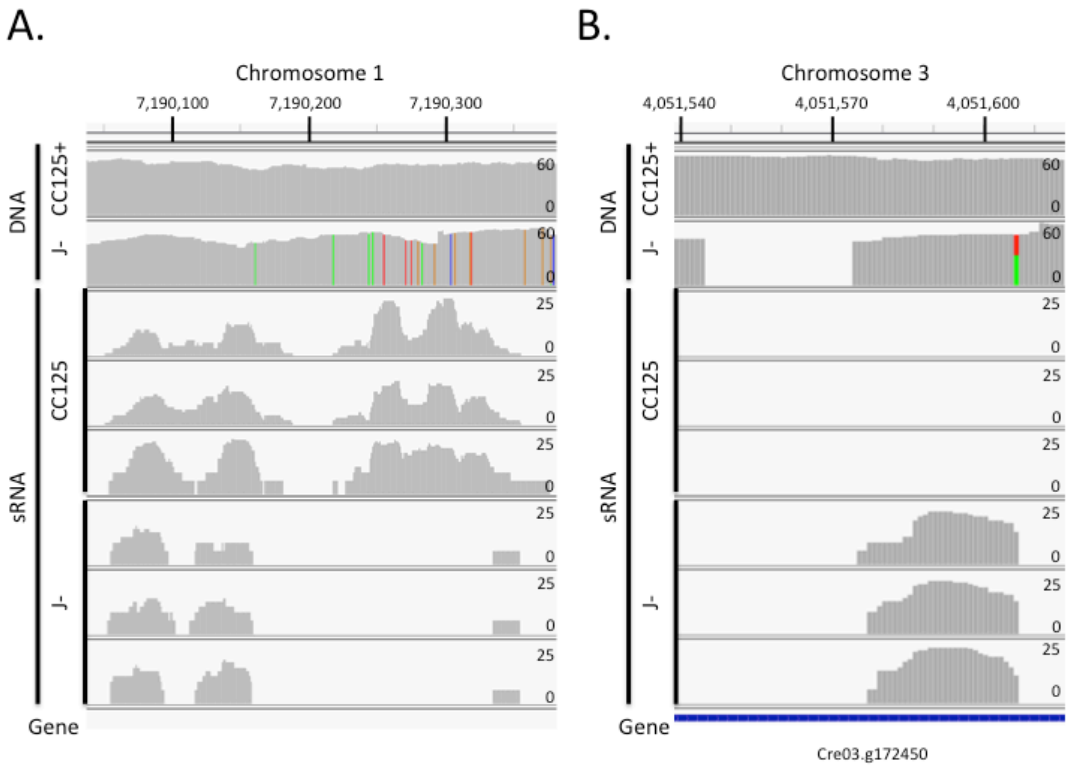


Figure 4-10: Different causes of differential expression
Alignments of two examples of differentially expressed loci. DNA reads were mapped to the genome and the log coverage is shown above. sRNA reads were also mapped to the genome (allowing mismatches), the expression normalized, and the coverage shown above. **(A)** A differentially expressed sRNA locus in which the genetic background of the J- strain contains SNPs. **(B)** A differentially expressed sRNA locus located near to an InDel in the J- genetic background.

4.2.11. Comparing miRNA expression and representation

To put the differential representation into biological context, I wanted to know what types of sRNA loci were differentially represented. While segmentSeq identifies loci, it does not classify them; for that, I had to use other sRNA identification or loci classification programs and compare their results to the

segmentSeq results. I decided to concentrate my analysis on miRNAs as they are the most established class of sRNA in *C. reinhardtii*.

I used MA plots of sRNA reads in which the log ratio of the sRNA expression between the parents, M, is plotted on the y-axis while the average log expression, A, is plotted on the x-axis. The program baySeq, also used by SegmentSeq, identifies differentially expression in count data (Hardcastle and Kelly, 2010). The results of the baySeq analysis of CC125+ and J- libraries, was used to indicate differentially expressed sRNA reads on the MA plots in red and overlaps with sequences of known and predicted miRNAs were indicated in green (Figure 4-11). Dr Sebastian Mueller provided the code for creating an MAplot in the programming language R.

MA plots highlighting the known miRNAs in all and unique sRNA datasets show that these known miRNAs generally had higher expression in CC125+ and/or only existed in CC125+ (Figure 4-11). This is likely due to the fact that known miRNAs have been identified so far in strains closely related to CC125+.

In the case of predicted miRNAs, differential representation was observed in both directions; predicted miRNAs exhibited increased representation in both the CC125+ and J- libraries (Figure 4-11). Additionally, many of these differentially represented miRNAs seemed to be strain specific (Figure 4-11).

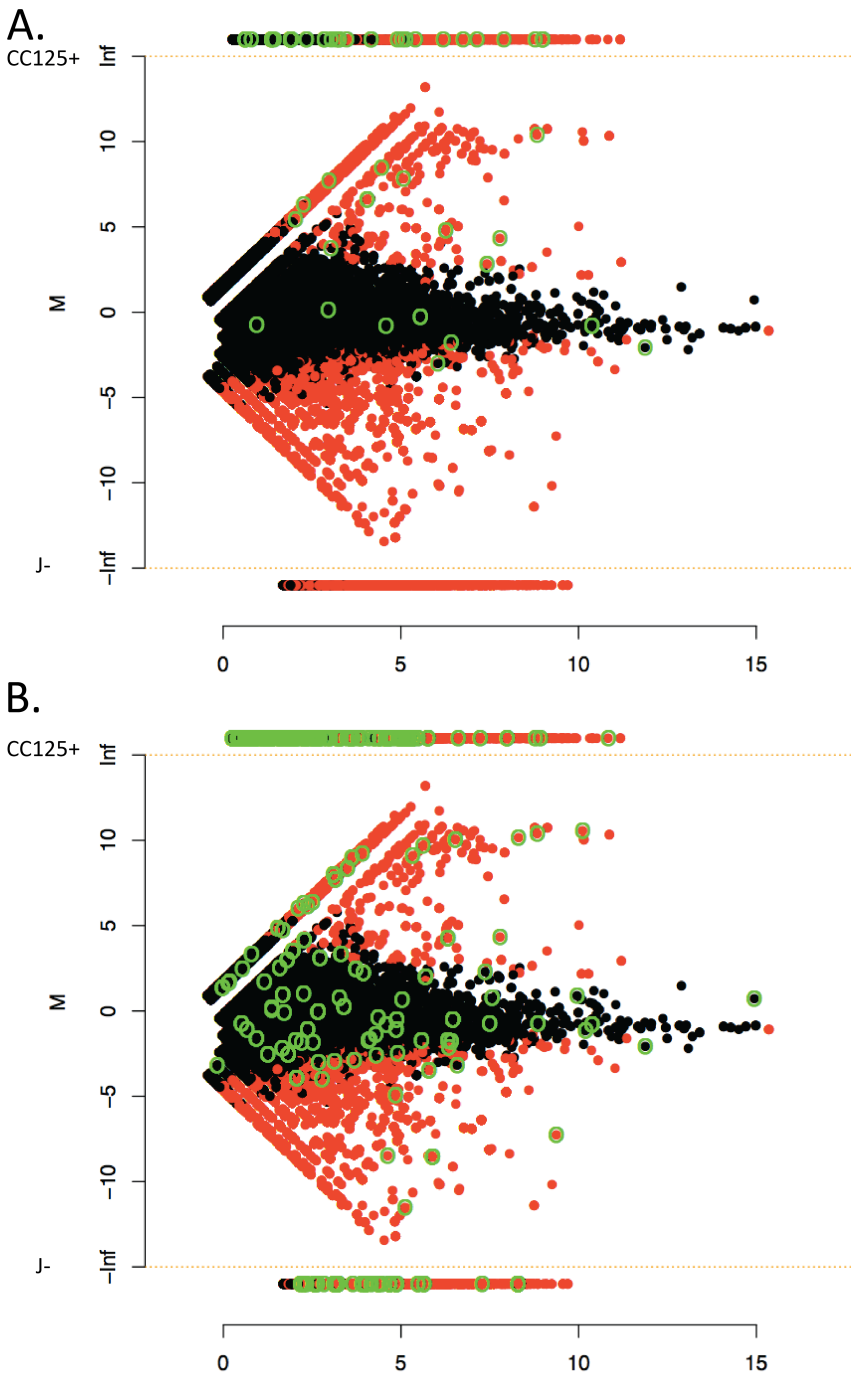


Figure 4-11: MA plots of sRNA reads
Expression of individual sRNA reads was compared using MA plots. The log ratio of the expression, M, is plotted on the y-axis while the average log expression, A, is plotted on the x-axis. Significantly differentially expressed sRNAs (likelihood > 0.9) are coloured red while miRNAs are circled in green. **(A)** Known miRNAs. **(B)** Predicted miRNAs.

4.3. Discussion

In this chapter I compared the sRNAs of the two parental strains CC125+ and J-, specifically concentrating on identifying differentially expressed sRNA loci while

keeping in account the genetic divergence I identified in Chapter 3. After confirming the quality of sRNA libraries (Appendix 7.6), I first broadly compared the CC125+ and J- libraries in a more qualitative manner assessing their size profiles (Figure 4-1), alignment to the reference genome (Figure 4-2), and GC content (Figure 4-3).

To specifically compare sRNA expression, I then used SegmentSeq to class the sRNA reads into predicted sRNA loci and used further bioinformatics tools such as miRCat and PhaseR to class the sRNA loci, giving me an overview of the sRNA locus composition in CC125+ and J- (Table 4-1). I analysed the association of sRNA reads, miRNAs, and phased loci with various annotations including repeats (Figure 4-4 and Figure 4-5). Finally the SegmentSeq data was used to identify conserved and differentially represented loci between the two strains (Figure 4-7). I used sRNA alignments allowing mismatches to check for interference in locus expression by sRNAs with mismatches and to separate the InDel-associated differentially represented sRNA loci from the truly differentially expressed sRNA loci (Figure 4-8).

Although the main point of this chapter was to compare CC125+ and J- at an sRNA level, the depth of the sRNA sequencing also allowed me to make novel observations on the sRNA landscape in *C. reinhardtii*.

4.3.1. Confirmed past knowledge of C. *reinhardtii* sRNAs

Past sequencing of *C. reinhardtii* sRNAs has always exposed the size class dominance of the 21nt sRNAs. The prevalence of 21nt sRNAs in *C. reinhardtii* is one of the most obvious differences between RNA silencing in this alga and that of higher plants. In plants there are two size classes of biologically relevant sRNAs: 21nt and 24nt, and the 24nt sRNAs are more abundant (Jagadeeswaran et al., 2012). However there are some exceptions in higher plants; for example in grapevine 21nt is the major peak and 24nt is the minor peak of sRNAs (Pantaleo et al., 2010).

This same 21nt size class dominance is present in *C. reinhardtii* sRNA libraries from past studies (Molnar et al., 2007a; Zhao et al., 2007a). It is also evident in CC125+ and J- size distributions (Figure 4-1). Recently, sequencing of *Volvox*

carteri sRNAs also exposed a 21nt size class dominance suggesting this pattern may be widespread in green freshwater algae (Li et al., 2014). Size profiles of sRNA reads is not always highly conserved, for example, they differ between different subspecies of rice (He et al., 2010).

Additionally the preference for U and A as the 5' end nucleotide of sRNAs was conserved between the two strains (Figure 4-1). This bias, also present in higher plants, is likely caused by the selective binding by Argonaute proteins (Herr and Baulcombe, 2004). The very initial studies on *C. reinhardtii* RNA silencing noted siRNA, miRNA, and phased sRNA loci in the genome (Molnar et al., 2007a; Zhao et al., 2007a). I identified examples of all of these types of loci in my own sRNA sequencing datasets (Table 4-1).

4.3.2. Low GC content of sRNAs could be due to high association of sRNAs with repetitive elements

Most my findings about *C. reinhardtii* sRNA were as expected: they correlated either with previous studies or to patterns observed in higher plants. Surprisingly however, I found that sRNA reads in *C. reinhardtii* had a lower GC content than the genome or transcriptome average, both of which are relatively high compared to other eukaryotic genomes (Figure 4-3). This suggested that the sRNA reads originated from low GC content segments of the genome such as certain types of intron and repetitive elements (Labadorf et al., 2010). Although the GC content of sRNAs is lower than the genome/transcriptome average, there might also be variation in GC content within the sRNA libraries. This GC content variation could imply different roles of sRNAs; high GC content miRNAs in *Arabidopsis* are more likely to be involved in stress regulation (Mishra et al., 2009).

To analyse where sRNAs are more likely to be originating from, and to later identify differential expression, I first predicted sRNA loci by aligning them to the reference genome and using this alignment as input into SegmentSeq.

The annotation association analysis showed that sRNA loci aligned predominantly to repetitive elements, perhaps explaining the low GC content of the sRNA libraries (Figure 4-4). The localization of most *C. reinhardtii* sRNA reads

to repeats is in keeping with observations in higher plants such as rice where most sRNAs come from interspersed repeats (predominately within that from transposons) (Xue et al., 2009). Specifically phased loci in both CC125+ and J- overlap repetitive elements suggesting that they might be involved in a genome defence role of RNA silencing (Figure 4-5). In comparison, miRNAs, both known and predicted, did not associate with intergenic repeats (Figure 4-4).

4.3.3. sRNA divergence between CC125+ and J-

Despite both CC125+ and J- sRNA populations showing broad similarities to one another such as size distributions (Figure 4-1) and GC content (Figure 4-3), I identified many differentially represented loci between the strains (Figure 4-7). Over a quarter of sRNA loci were differentially represented (Figure 4-7). The rest of the sRNA loci were unclassified (although they likely have similar expression) except for a small percentage of sRNA loci with highly conserved expression levels (Figure 4-7).

I also identified differentially represented miRNAs including some miRNAs that seemed to be strain specific (Figure 4-11). However without knowing the time since a last common ancestor between the strains it is impossible to comment on the speed of miRNA evolution in *C. reinhardtii*.

Another striking difference between CC125+ and J- was the increased number of phased sRNA loci in the J- library (Table 4-1). The majority of the J- specific phased loci associated with a non-LTR retrotransposon, L1-1 (Table 4-2).

4.3.4. Causes of sRNA divergence between CC125+ and J-

Some differential representation of sRNAs between the two strains was due to the lack of a genetic background for an sRNA locus in one strain (Figure 4-8); these instances were labelled as InDel-associated differential sRNA representation and made up nearly a third of the differential sRNA representation noted between the strains (Figure 4-9). However, of a small subset of differentially represented sRNA loci, the underlying DNA sequence was present in both parental strains (Figure 4-8), thus classified as differentially expressed sRNA loci (Figure 4-9). The differentially expressed sRNA loci exhibited varying levels of genetic divergence within and nearby the locus (Figure 4-10).

While some differential sRNA expression is doubtless the result of genetic variation within the locus, *cis*-elements, or *trans*-elements, some could be due to epigenetic modifications at the locus.

The divergence between two strains underlines the epigenetic variability in terms of sRNAs that can be found within a species. The differential sRNA representation between CC125+ and J- has the potential to affect further phenotypes in the strains but further target prediction and analysis is needed. The divergence of CC125+ and J- sRNA populations and genomes also confirmed that the parental strains were good candidates for a crossing experiment looking for transgressive sRNAs. In the next chapter I describe the crossing experiment and the resulting inheritance of sRNA loci in the recombinant strains.

4.4. Acknowledgements

Cancer Research Institute UK sequenced the sRNA libraries. Dr Bruno Santos ran PhaseR on my sRNA datasets for me. Dr Sebastian Mueller who also supplied me with the code for the creating the MA plots.

5. Chapter five: Identifying transgressive sRNA expression in *C. reinhardtii* recombinants

Transgressively expressed sRNA loci were identified in recombinant progeny of a cross between CC125+ and J- including examples of transgressively expressed miRNAs and phased siRNAs.

5.1. Introduction

In the previous two chapters, I compared the genomes and sRNA profiles of the *C. reinhardtii* strains, CC125+ and J-. The genetic comparison primarily sought to assess the suitability of these strains for a crossing experiment seeking to identify transgressive sRNA expression, while the comparison of the parental sRNAs was performed to put later work into context. In this chapter I detail the crossing experiment, the general patterns of sRNA inheritance, and the process of identifying transgressively expressed sRNAs.

5.1.1. sRNA inheritance

Small RNA expression can be inherited in both a Mendelian manner (along with their locus) or separate to their underlying DNA sequence if there are epigenetic factors involved. In most organisms inheritance of sRNA expression is influenced by both genetic and epigenetic factors. Due perhaps to plant specific pathways such as RNA dependent DNA Methylation (RdDM), heritable epigenetic variation caused by sRNAs is more common in plants than in animals (Bond and Baulcombe, 2014).

Although there is no evidence for transgenerational epigenetic inheritance in *C. reinhardtii* one would perhaps expect that it would occur, as in higher plants. If that were true I would predict that, in *C. reinhardtii* there could be recombinant strains, as in tomato (Shivaprasad et al., 2012), with transgressive levels of sRNAs that are either more or less than that of either parent.

5.1.2. Transgressive sRNA expression

Hybridization results in transgressively expressed sRNAs in tomato (Shivaprasad et al., 2012), maize (Barber et al., 2012), rice (He et al., 2010), wheat (Kenan-

Eichler et al., 2011), and *Arabidopsis* (Ha et al., 2009). Transgressive sRNA expression patterns are qualitatively species specific but there are some quantitative trends in higher plants evident as a general reduction in 24nt sRNAs after hybridization (Barber et al., 2012) or the increase in 21nt sRNA non-additive expression with an increase in the genetic divergence of the parents (Ha et al., 2009; Kenan-Eichler et al., 2011).

Some theories have been proposed for the causes of transgressive sRNA expression such as the increased complexity of the sRNA network. However there are no validated examples for the cause of transgressive sRNA expression. The varying mechanisms of sRNA inheritance, cell biologies of crosses and other species-specific aspects to RNA silencing means that there is unlikely to be a single cause for this phenomenon. It was therefore difficult to predict the nature of transgressive sRNA expression resulting from hybridization in *C. reinhardtii*.

5.1.3. *C. reinhardtii* mating

C. reinhardtii is vegetatively haploid, conveniently making bioinformatics simpler in comparison to genomes of higher ploidy. To mate, *C. reinhardtii* undergoes gametogenesis (catalysed by stress conditions such as the depletion of a nitrogen source) and, upon finding a cell of the opposite mating type, then fuses to become a diploid cell known as a zygote (Section 1.5.2.2).

As a diploid, two genomes and epigenomes have a chance to interact. Specifically, sRNAs could act on the novel variation, in the form of the other genome, altering expression (both of genes and secondary sRNAs) or even changing epigenetic marks such as DNA methylation or histone modifications. In the diploid zygote phase there is the opportunity for transgressive expression of sRNAs to occur.

The diploid zygote phase can be viewed as a sort of hibernation as the zygote specific thick cell wall is more resilient to stresses, including desiccation, than that of vegetative cells. Once conditions are more favourable (such as the reappearance of a nitrogen source), the diploid zygote undergoes meiosis resulting in four genetically distinct recombinant progeny. The genetic uniqueness of the recombinants is caused by random separating of homologous

chromosomes and by recombination. At this point sRNA inheritance is also being affected. The splitting of the cell nucleus will physically assign certain sRNA loci to different recombinants.

5.1.4. Searching for transgressive sRNA expression in C. *reinhardtii*

I looked for transgressive sRNA expression in the four recombinant progeny of a cross between CC125+ and J-. In this chapter I first describe the optimization of the *C. reinhardtii* mating procedure between these two strains and the design of the sRNA comparison experiment. I then used various bioinformatics approaches to understand the general patterns of sRNA inheritance observed in a single cross. Finally, I report that transgressive expression of sRNA loci was found in the various progeny of the cross, certain examples of which were verified using sRNA northerns analysis. Similar bioinformatics techniques as used in Chapter 4 were utilized to classify the transgressively expressed sRNA loci and target prediction was attempted in order to establish the likely biological consequence of the effect.

5.2. Results

5.2.1. Optimization of the mating protocol

In order to assess the mode of sRNA inheritance in *C. reinhardtii*, I had to mate two genetically diverse strains and isolate the four recombinant progeny of a single cross (Figure 5-1). To achieve this, I performed reciprocal crosses between the mating types the standard lab strains (CC125+ and CC124-) and the more recently isolated Japanese strains (J+ and J-), also crossing each pair (i.e. CC125+ × CC124-). The initial crosses did not yield as many zygotes as I had hoped and so some optimization of the mating protocol was required.

I tested the effect of different media for the zygote maturation plates. Specifically, varying amounts of Nitrogen were made available to the zygotes on these plates; for the 3% agarose plates I used TAP, TAP 1/10 N, and TAP –N media. Microscopic analysis showed that the TAP –N maturation plates yielded the highest populations of zygotes most consistently (Figure 5-2). However the distortion of the colouring of vegetative cells on these plates made it difficult to

recognize zygotes. The TAP 1/10 N plates resulted in more zygotes than the TAP maturation plates and so TAP 1/10 N was used in all future crosses as the medium for maturation plates (Figure 5-2).

The resulting zygotes from each cross were isolated and their tetrads dissected using a blunted glass needle. This method resulted in the loss of many zygotes and gametes due to mechanical damage and so only a few full tetrads were isolated.

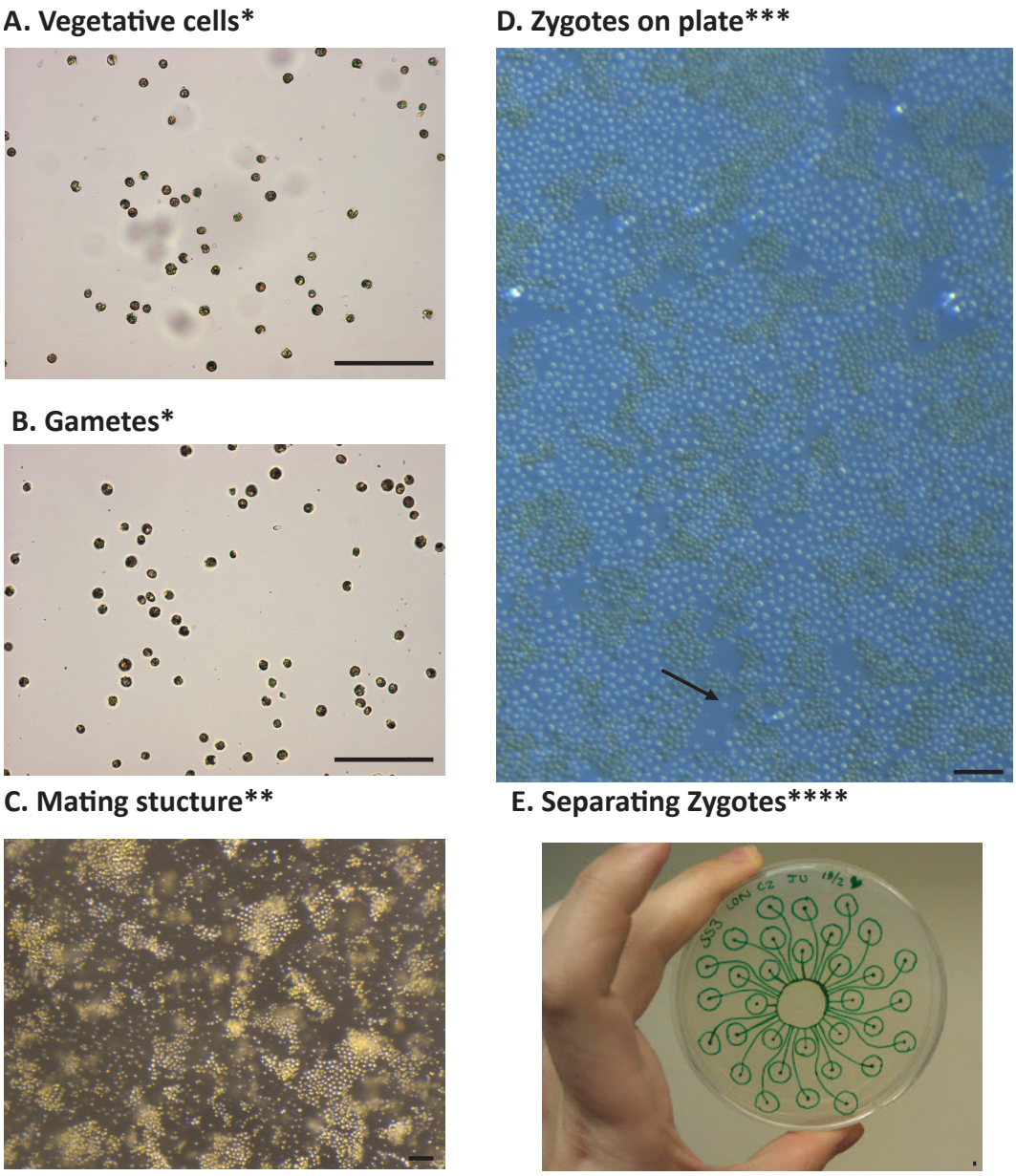


Figure 5-1: Procedure for mating and tetrad dissection of *C. reinhardtii*
C. reinhardtii cells propagated mitotically in the form of **(A)** vegetative cells. They replicated mitotically, whether synchronized or unsynchronized, until a lack of nitrogen induces gametogenesis. In order to produce gametes for mating, a large amount of vegetative cells were plated on solid medium TAP –N plates containing no nitrogen. After several days on this medium, the cells were resuspended in distilled H_2O where they underwent reflagellation to form mobile **(B)** gametes capable of recognizing the opposite mating type. When gamete cell cultures were mixed in liquid medium they started mating. A good indication of successful mating was the formation of the **(C)** pedicle, a 3D mating structure made up of an orgy of cells. The mating cell culture was spotted onto low nitrogen TAP plates on which the cells matured into **(D)** zygotes (an example of a zygote is indicated by an arrow), which could be exposed by scraping off the vegetative cells and remaining gametes, as the zygotes tend to stick to the bottom. These zygotes were transferred to **(E)** a new plate to be separated. The zygotes were dragged out to the end points away from the vegetative masses. After a day on this nitrogen rich medium the tetrads were dissected so that each recombinant was dragged to a point on the line of the circle drawn to encompass the zygote. The scale bar represents 10µm. Cells were visualized with *light microscope (40x), **dissecting microscope (10x), and ***digital camera.

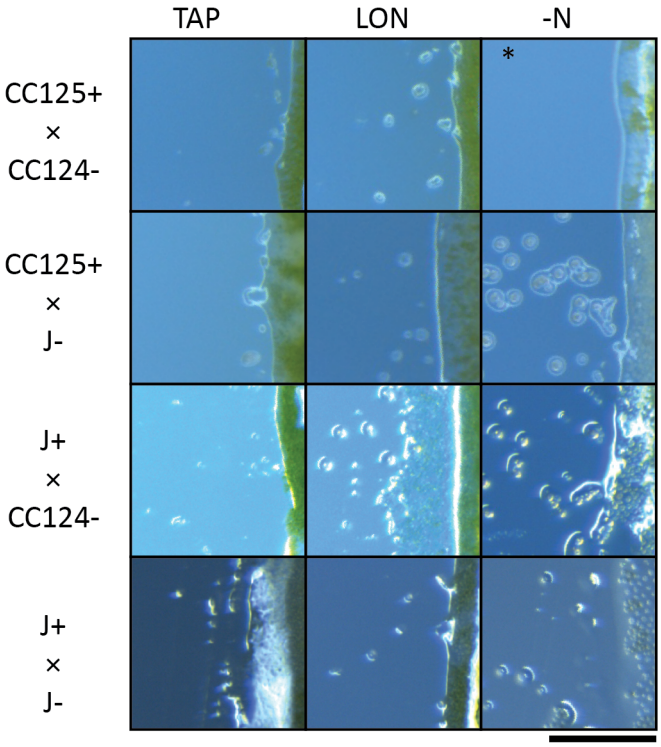


Figure 5-2: Optimizing mating
In order to optimize both the number of zygotes created and their adhesion to the surface of zygote maturation plates (necessary for their later separation from vegetative cells) I tested different media for the maturation plates. Reciprocal crosses were performed between lab strains (CC125+ and CC124-) and Japanese strains (J+ and J-) and the resulting mating mixture spotted onto different media, namely TAP, LON (low nitrogen with 1/10 of the nitrogen of the standard TAP media) and –N (no nitrogen added to medium). Very low levels were observed on TAP. The highest levels of zygotes were noted on –N maturation plates however adhesion to the plate was unreliable (exemplified by * where zygotes were identified in the spot of cell culture but they did not stick to the surface of the plate). Additionally vegetative cells were discoloured on –N plates.

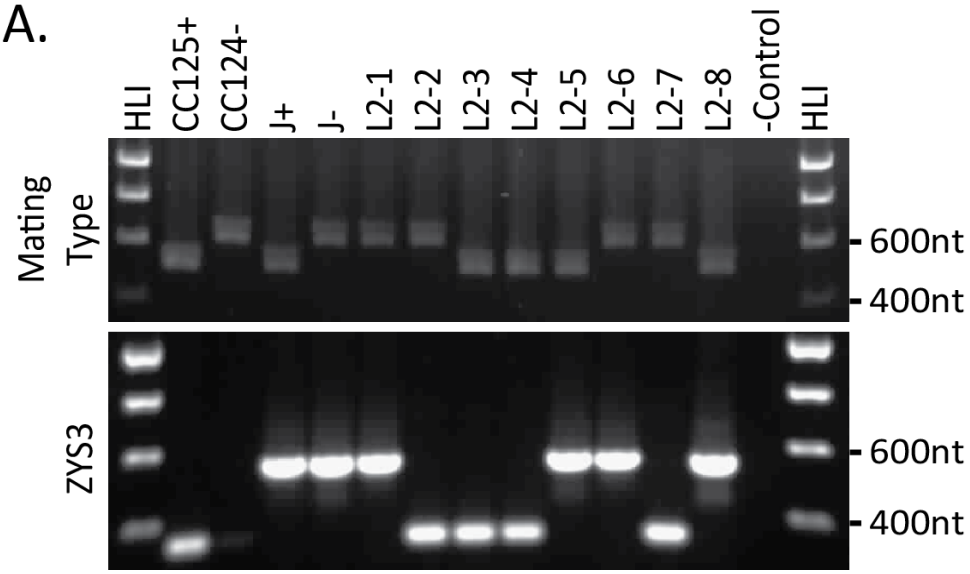
5.2.2. Verifying recombinant identity

To confirm the recombinant nature of the isolated strains (and to ensure that I had not isolated the vegetative product of a mitosis event instead), I used a PCR-based approach. The mapping marker PCR products known to differ between the parental strains (Appendix 7.3) and the mating type PCR were used to show that a strain had markers from both parents (via the separation of homologous chromosomes or via recombination).

Although most tetrads isolated were true products of meiosis, the only complete true tetrad isolated, resulting from a cross between CC125+ and J-, had separated at the eight-cell stage before dissection (Figure 5-3). Neatly, two cells

for each tetrad genetic background were identified (Figure 5-3). I used this cross in the sRNA comparison experiment, using an example of each tetrad genetic background in the search for transgressively expressed loci.

A.



B.

	CC125+	J-	L2-1	L2-2	L2-3	L2-5
Mating Type (+/-)	+	-	-	-	+	+
ZYS3 (344/530 bp)	344	530	530	344	344	530

Figure 5-3: Verifying recombinants
Recombinants isolated during tetrad dissection of a CC125+ and J- cross were verified using the ZYS3 mapping marker and mating type PCRs. **(A)** PCR products visualized on a 1.5% agarose EtBR gel. Control: H_2O used as template. DNA ladder: Hyperladder I (HPI). **(B)** Table of the results of the ZYS3 and mating type marker PCRs.

5.2.3. Designing sRNA comparison experiment

In order to identify transgressively expressed loci I wanted to capture as much sRNA activity as possible. To this end I decided to extract RNA from cultures growing in minimal medium as the absence of acetate in this medium means that the cells are forced to photosynthesize. Extracting RNA from photosynthesizing cells would allow me to capture any transgressive sRNA activity that might be linked to photosynthetic phenotypes.

I also decided to use unsynchronized vegetative cultures. Small RNA expression has been shown to vary throughout the mitotic cycle (unpublished data from the

Baulcombe lab). A disadvantage of using unsynchronized cell cultures was that some cell cycle specific effects might be diluted out. However the high depth of sRNA sequencing afforded by using the Illumina HiSeq protocol meant that even small differences between sRNA expression in the strains should be identifiable.

Three replicates of each strain were used in this experiment. Three single colonies from each strain were used to inoculate three 50 ml flasks of minimal medium and grown separately for three days (Figure 5-4). Cells were harvested at mid log phase, when the cell density was 1-5 x 10^6 cells/ml. For ease of identification of this stage, I used a spectrophotometry assay to calculate cell density.

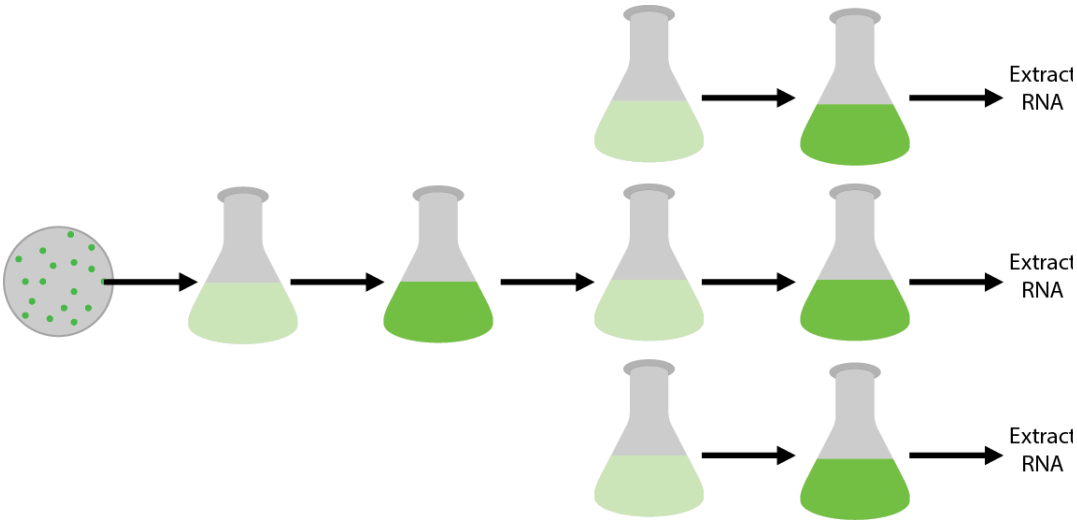


Figure 5-4: Experimental setup
Designed an RNA extraction experiment in order to create biological replicate like samples. Single cell colonies from strains plated on solid minimal medium (Appendix) were used to inoculate liquid minimal medium to low cell density (an OD680 reading of 0.1). Cultures were grown until mid-log phase was reached (usually for 5 days, until an OD680 reading between 0.5-0.8 was reached). This healthy vegetative unsynchronized culture was used to inoculate three individual flasks of minimal medium (once again to an OD680 reading of 0.1). These cultures were harvested for RNA extractions after mid-log phase was reached.

5.2.4. Small RNA libraries quality confirmed

The quality of the sRNA reads from the recombinant libraries was assessed using a similar approach to that of confirming the quality of the parental sRNA libraries (Chapter 4). FASTQC confirmed the per base sequence quality of the reads to be similar to that of the parents and sufficient for further analysis (Appendix 7.6).

The replicate identify of the libraries was confirmed using a comparison of the log of the expression of individual loci (further description of how loci were identified can be found in section 5.2.6) (Figure 4-4).

5.2.5. sRNA length distributions of C. *reinhardtii* highly conserved in recombinants

The sRNA read files for the recombinants were loaded into the DCB pipeline (unpublished software) to be trimmed and aligned to the *C. reinhardtii* reference genome. The DCB pipeline output includes a size distribution of the sRNA reads from both redundant reads and non-redundant reads. The characteristic 21 nt size class dominance of sRNAs in *C. reinhardtii* was preserved in all of the recombinants as was the preference for U and A as the 5' end nucleotide of sRNAs (Figure 5-5).

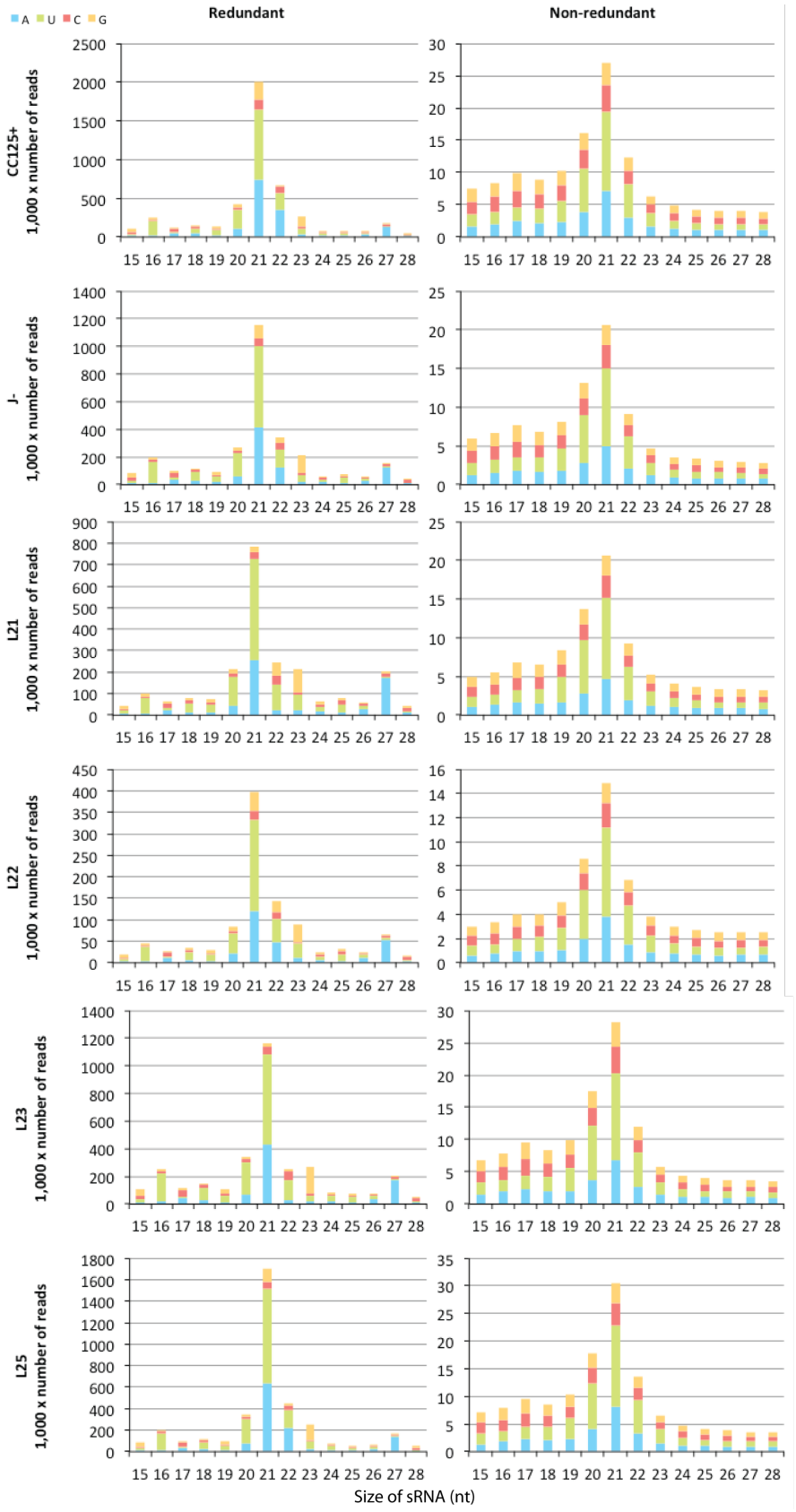


Figure 5-5: Size distributions of sRNA libraries
sRNA size distributions of CC125+, J-, L2-1, L2-2, L2-3, and L2-5 aligning to *C. reinhardtii* reference genome (Merchant et al., 2007). 5' end nucleotide percentage is shown by colour of the bar. Redundant: all reads including counts. Non-redundant: counts are discarded.

5.2.6. Predicting sRNA loci

I used an identical approach to predicting sRNA loci as I used in the comparison of the parental sRNAomes (Section 4.2.8). Once again, the sRNA reads of the replicate libraries were aligned allowing no mismatches via Bowtie (Langmead et al., 2009) to the current *C. reinhardtii* reference genome (Merchant et al., 2007) and the resulting .bam files loaded into SegmentSeq (Hardcastle and Kelly, 2010). SegmentSeq first assigns loci to clusters of sRNA reads and the output can be used to call differential expression between strains, taking replicates into account to lend assurance to the significance of the results.

In order to compare the predicted number of loci between the different strains relative to one another, I loaded all replicates of all strains into one SegmentSeq object using all replicates. Concurrent to this, I also ran a SegmentSeq analysis using only sRNA reads that matched once to the genome (unique). Although this approach distorts the number of loci to be detected it compensates for the over-representation of loci corresponding to repetitive elements.

	CC125+	J-	L21	L22	L23	L25
SegmentSeq loci unique	795	480	689	583	697	699
SegmentSeq loci	4650	3914	4040	2418	6136	5612
Known miRNAs	39	26	32	35	32	32
Predicted miRNAs (avg >0)	282	78	192	102	151	146
Phased loci (avg <-10)	15	372	16	11	224	245
Non classified	4314	3438	3800	2270	5729	5189

Table 5-1: Number of loci
A SegmentSeq object was created using all replicate sRNA libraries and was used to calculate the number of sRNA loci predicted for each strain. MiRProf was used to identify known miRNAs in the strains and miRCat predicted novel miRNAs in each library. The loci identified by SegmentSeq were used as input into PhaseR to identify the number of phased loci. The rest of the sRNA loci were labelled as unclassified.

The results of the loci predicted using uniquely matching reads are striking in their similarity, yet for loci predicted using total reads, three of the recombinants have a different number of sRNA loci in comparison to either of their parents. (Table 5-1). L2-2 has fewer sRNA loci (Table 5-1) than either parental strain, although this result could be due to the fact that this library had the smallest number of reads returned from sequencing and so some lowly expressed loci

could be missed. L2-3 and L2-5 in contrast both have more loci predicted from total reads than either parent (Table 5-1). This difference was not apparent in loci predicted from uniquely matching reads it is likely that this increase is due to sRNA loci originating from repetitive regions of the genome.

The same bioinformatics tools used for the comparison of parental sRNAomes were used to classify sRNA loci in the recombinants. The UEA tools, miRProf and miRCat were used to identify known miRNAs and to predict novel miRNAs respectively. Similar numbers of known miRNAs were identified in the recombinants to the parents (Table 5-1). The number of miRNAs predicted in every recombinant is similar between the recombinants and between the amounts predicted for CC125+ and J- (Table 5-1). This implies that miRNA representation was usually inherited in a genetically dependent manner. Due to recombination and separation of chromosomes you would expect the number of miRNAs in the recombinants to be intermediate between the numbers in either parent.

The segmentSeq object was used to predict phased loci using PhaseR (http://www.plantsci.cam.ac.uk/bioinformatics/phaser). Dr Bruno Santos performed the PhaseR analysis. The singular difference in the number of phased loci noted between the parental strains is also seen in the four recombinants. Recombinants L2-3 and L2-5 both show an increased amount of phased loci, similar to the J- parent (Table 5-1). This pattern of inheritance implies there may be some trans factor involved in creating phased loci that was inherited in only two recombinants. The fact that only two recombinants carry this phenotype also suggests that this trans factor is inherited in a genetically dependent manner.

5.2.7. Patterns of sRNA inheritance

Given the similarity of RNA silencing in *C. reinhardtii* and higher plants, I predicted that inheritance patterns in my recombinants would be as in tomato (Shivaprasad et al., 2012) and other species that have been tested (Stupar et al., 2008).

Most sRNA loci should show a Mendelian inheritance, corresponding to the DNA background. For single copy loci, this type of inheritance should result in expression in recombinants that is similar to one of the parents. And finally transgressive loci could show expression outside of the parental range.

I wanted to assess the general patterns of sRNA inheritance in *C. reinhardtii* recombinants, to calculate whether there are more additive sRNA loci than transgressive sRNA loci. And if transgressive expressed sRNA loci were identified, are there any patterns within that subset of sRNA loci? Does transgressive expression tend to manifest in an increase of sRNA reads (producing an sRNA locus that has higher sRNA expression than the parental maximum) or a decrease in sRNA reads (producing an sRNA locus that has lower sRNA expression than the parental minimum)? Or is there a parental bias in the direction of sRNA transgressive expression, with transgressively expressed sRNA loci always representing an exaggeration of the sRNA expression of one parent over the other?

I utilized ratios of dominant to additive values (d/a ratio) to determine the types and frequencies of sRNA inherited expression (Stupar et al., 2008). The variable d accounts for the difference between the recombinant and parental expression (d = average recombinant expression – midparental value). The variable a accounts for the range of parental expression (a = parent expression – midparental value). In this way the d/a ratio identifies and distinguishes between additive and non-additive expression. If the d/a ratio is between -1 and 1 for a single sRNA locus then that locus exhibits additive expression. Less than -1 or greater than 1 means the locus is exhibiting transgressive expression. If the d/a ratio equals -1 or 1 then the expression of the sRNA locus in the recombinant is non-additive but equal to one of the parental values. If the d/a ratio equals zero then the sRNA expression in the recombinant is equal to the midparental expression level (a d/a ratio of zero would also be true for sRNA loci with conserved expression in both the parents and the recombinants).

The sign of the d/a ratio describes the direction of the additive or transgressive expression dependent on the equation for a. In Type I d/a calculations, a =

maximum parental value – midparental value. Thus the sign of the ratio will show whether the sRNA expression is in the direction of the minimum or maximum parental value. For example, a d/a ratio of -0.5 would mean an sRNA locus is exhibiting additive expression in the direction of the minimum parental value while a d/a ratio of +7 would mean transgressive expression greater than the maximum parental value. In Type I d/a calculations, a = Parent #1 value – midparental value. Thus the sign of the ratio will show whether the sRNA expression is in the direction of the one parental value or the other. For example, a d/a ratio of -0.5 would mean an sRNA locus is exhibiting additive expression closer to the expression in Parent #2 while a d/a ratio of +7 would mean transgressive expression greater than the expression in Parent #1.

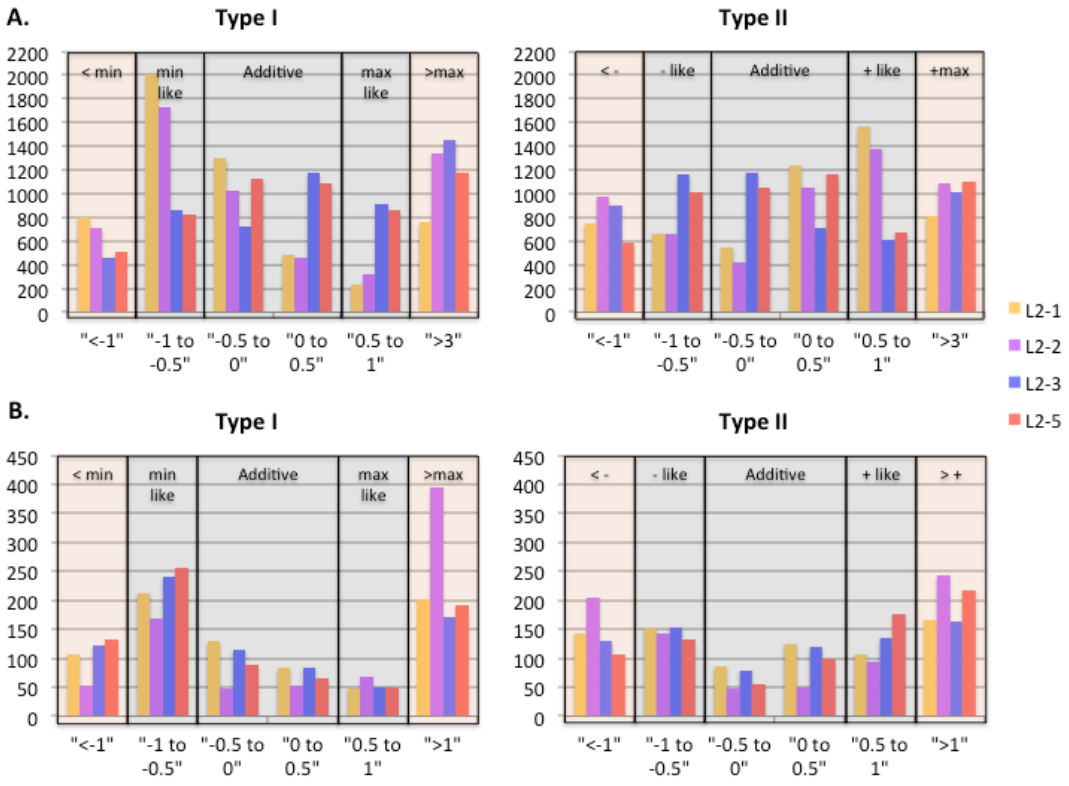


Figure 5-6: General inheritance patterns by recombinant
Two SegmentSeq objects of all replicates of all strains were used in d/a calculations. **(A)** Small RNA loci made up of uniquely matching sRNA reads were represented by one SegmentSeq object while another SegmentSeq object was created using **(B)** all sRNA reads in the library and therefore also taking loci aligning to repetitive elements into account. Type I d/a calculation represent the difference between the recombinant and parental sRNA locus expression relative to the maximum and minimum parental expression. Type II d/a calculations represent a comparison between recombinant and parental sRNA locus expression in relation to the plus and minus parental expression separately.

The Type I d/a ratios for both uniquely matching and total read loci predictions indicate that there is a considerable level of transgressive expression and that this expression tends to be greater than the maximum parental level (Figure 5-6). A greater amount of unique sRNA loci exhibited the minimum parental expression in recombinants L2-1 and L2-2 however this enrichment was lost when all sRNA reads were used (Figure 5-6). Interestingly there is an increase in transgressively expressed sRNA loci in L2-2 when all sRNA reads are used (Figure 5-6). Although significant and similar levels of transgressive expression can be observed in each recombinant, most of the sRNA loci show expression within the range of the parents (Figure 5-6).

In the type II d/a calculations there was no trend in sRNA expression, whether additive or transgressive, of being similar to one parent over another (Figure 5-6). In fact transgressive sRNA loci was evenly spread between exaggerating CC125+ expression and exaggerating J- expression (Figure 5-6).

5.2.8. Pattern of inheritance of types of sRNAs

The d/a analysis of all sRNA loci exposed the general trend of additive expression with some occurrences of transgressive expression (Figure 5-6) however different types of sRNA loci might be inherited in specific manners. I performed the same d/a analysis as used on all sRNA loci, using the different classes of sRNA loci identified in 5.2.6 (phased siRNA, predicted miRNA, and known miRNA).

The results of the d/a analysis of phased sRNA loci were striking: most of the phased loci exhibited transgressive expression (Figure 5-7-C). Of particular interest were the phased transgressive expression of sRNAs in recombinants L2-3 and L2-5 that was higher than the maximum parental value (Figure 5-7-C). The type II d/a calculation revealed that the maximum parental expression of the transgressive phased sRNA loci for these recombinants was usually that of J- (Figure 5-7-C). L2-2 shows the opposite pattern to L2-3 and L2-5, with transgressive expression of phased loci occurring nearly only in the negative direction (less than the minimum parental value) where the minimum parental value is usually that of CC125+ (Figure 5-7-C). L2-1 differs to the other recombinants in that it contains more additively expressed phased sRNA loci

(Figure 5-7-C). However the pattern of additive expression is telling; most of the phased sRNA loci in L2-1 are more similar to CC125+ expression than J- expression where CC125+ has a lower expression value than J- (Figure 5-7-C).

By comparison the results of the miRNA d/a analysis showed that the majority of miRNA expression in the recombinants is within the range of the parents; this was true for both known miRNAs (Figure 5-7-A) and miRNAs predicted with miRCat (Figure 5-7-C). There was however some miRNA expression outside the range of the parents (Figure 5-7). In both of the miRNA d/a analyses (Type I and Type II) there is a depletion of miRNA expression between -0.5 to 0 and 0.5 to 1 (Figure 5-7). Also for predicted miRNAs there were very few with expression levels lower than the minimum or J- parental value (Figure 5-7). It is possible these patterns are an artefact caused by the absence of expression of more miRNAs in the J- strain.

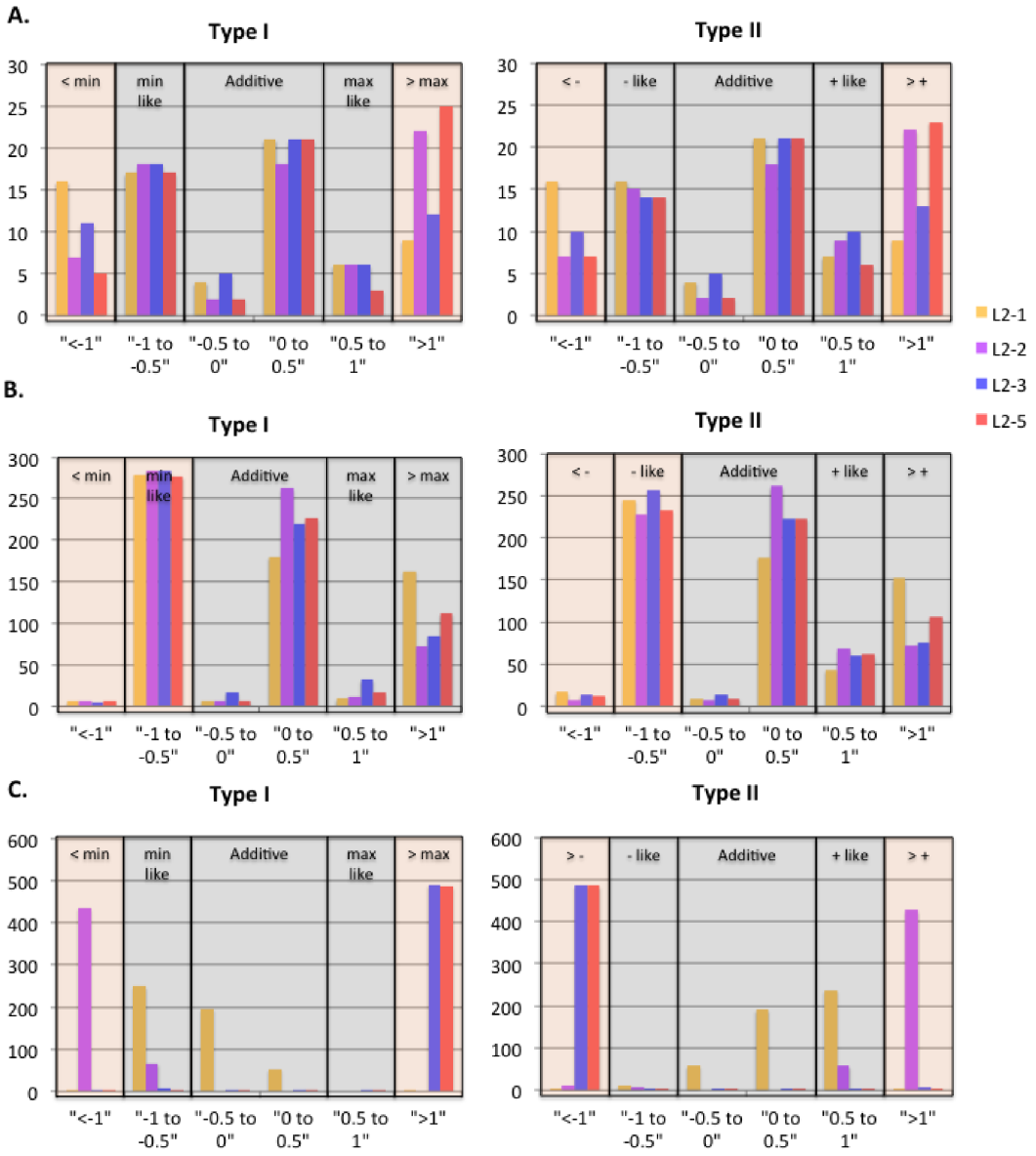


Figure 5-7: d/a ratio analysis of phased and miRNA loci
Alignment objects of **(A)** known miRNAs, **(B)** miRCat predicted miRNAs, and **(C)** phased loci (any loci with an average PhaseR value between the replicates of <-10 in any strain) were used in d/a calculations. Type I d/a calculations represent the difference between the recombinants and parental sRNA locus expression relative to the maximum and minimum parental expression. Type II d/a calculations represent a comparison between recombinant and parental sRNA locus expression in relation to the plus and minus parental expression separately.

5.2.9. Identifying transgressively expressed sRNA loci

Although the d/a calculations identified multiple examples of transgressively expressed sRNA loci, to add significance to these observations (for example the d/a ratio did not account for variability between replicates) I designed a pair wise comparison of recombinant expression to parental expression. This pairwise

comparison would allow me to identify significantly transgressively expressed sRNA loci.

Two SegmentSeq objects were created for each recombinant using only reads that matched the genome once or twice only; in one object differential representation was identified between the recombinant and CC125+ while in the other object the comparison was between recombinant and J-. Only loci with likelihoods of differential representation above 0.9 were used for further analysis. This list of differentially represented relative to parent sRNA loci was combined to identify loci that were differentially represented relative to both parents and it was restricted to loci that were expressed outside the range of the parents in the recombinants.

The genetic background was then checked to confirm differential sRNA expression rather than differential representation due to an InDel in the recombinant. Genetic backgrounds were also assessed to check whether mismatches had interfered with the alignment of sRNA reads at that specific locus. While the segmentSeq object were created using alignment files that did not allow mismatches, the gene browser views show sRNA alignments created with the CLC Genomics Workbench. The primary difference between these two alignments is that the CLC Genomics Workbench sRNA alignments allow mismatches based on the quality of the sequenced bases and number of other reads aligning.

Every recombinant contained transgressive sRNA loci (Table 5-2). L2-2 is the recombinant with the fewest transgressively expressed loci although this may be a result of the small library size for this strain (Table 5-2). Most transgressive expression was very subtle and less than double the parental level (Table 5-2). However there were loci with transgressive expression in the recombinants that was more than ten times increase the parental level.

Like the results of the d/a analysis (Figure 5-6), the pairwise comparison identified transgressively expressed loci exaggerating both CC125+ and J- expression (Table 5-2). However the pairwise comparison identified fewer instances of transgressive sRNA expression lower than the minimum parental

value when using one or two times matching sRNA reads (Table 5-2). Additionally an sRNA locus will only be transgressive in one or at most two recombinants. None were found present in all four meiotic products.

A.

	Minimum change in expression	L21	L22	L23	L25
Up	>1x	45	7	88	54
	>2x	44	0	42	52
	>10x	0	0	2	11
	>100x	0	0	0	0
Down	>1x	0	0	0	1
	>2x	0	0	0	1
	>10x	0	0	0	0
	>100x	0	0	0	0

B.

	Minimum change in expression	L21	L22	L23	L25
Up	>1x	127	2	132	45
	>2x	127	2	132	44
	>10x	2	0	6	11
	>100x	0	0	0	0
Down	>1x	4	0	0	524
	>2x	2	0	0	524
	>10x	1	0	0	59
	>100x	0	0	0	0

Table 5-2: Number of TE loci
SegmentSeq was used to identify differentially expressed sRNA loci between each recombinant and each parent separately. First **(A)** only loci predicted only from reads which matched twice to the genome were analysed, and later **(B)** loci predicted from all reads were analysed. The different parental comparisons of the recombinants were then compared to identify sRNA loci in the recombinants that were differentially expressed to both parents and outside the range of the parents. The numbers of loci showing various dimensions of a change in expression to the maximum parental value are shown in the table of above.

5.2.10. Parental genetic background of transgressively expressed sRNA loci

The recombinants contained transgressively expressed sRNA loci whose expression was an exaggeration of either CC125+ or J-. I wanted to check whether the parental direction of the transgressive sRNA expression was linked to the genetic background of the sRNA locus. Recombination and random allocation of homologous chromosome had created a mosaic of parental genetic

material in the *C. reinhardtii* recombinants (Appendix 7.11). I used the genome browser views to identify whether the genetic background of a transgressively expressed sRNA locus was that of CC125+ or J-.

Of the 194 transgressively expressed sRNA loci predicted using uniquely matching reads, the expression 99 sRNA loci was close to the same parent as the genetic background of the locus in the recombinants. For the other half (95) of the transgressively expressed sRNA loci, the recombinant genetic background at that locus was not that of the parent with the closer sRNA expression at that locus. This suggests that sRNA locus expression does not always segregate with its genetic locus.

5.2.11. What types of sRNAs are transgressively expressed?

To put transgressively expressed sRNAs into genomic context, I performed an overlap analysis against various genome annotations using total sRNA loci and transgressively expressed sRNA loci (Figure 5-8) predicted via segmentSeq using all sRNA reads or uniquely matching sRNA reads. Both transgressively expressed loci predicted from total reads and uniquely matching reads showed highly similar patterns of annotation overlaps (Figure 5-8). Some transgressively expressed sRNA loci were located in genes however the majority are located in intergenic regions (Figure 5-8). Despite the fact that most sRNA loci align to repetitive elements, this association with intergenic regions fitted with the previous overlap analysis using sRNA reads showing that nearly half of sRNA reads originate from intergenic regions (Figure 4-4). Transgressive loci predicted using all reads and single matching reads associate less frequently with repetitive elements in all contexts in comparison to total sRNA loci (Figure 5-8).

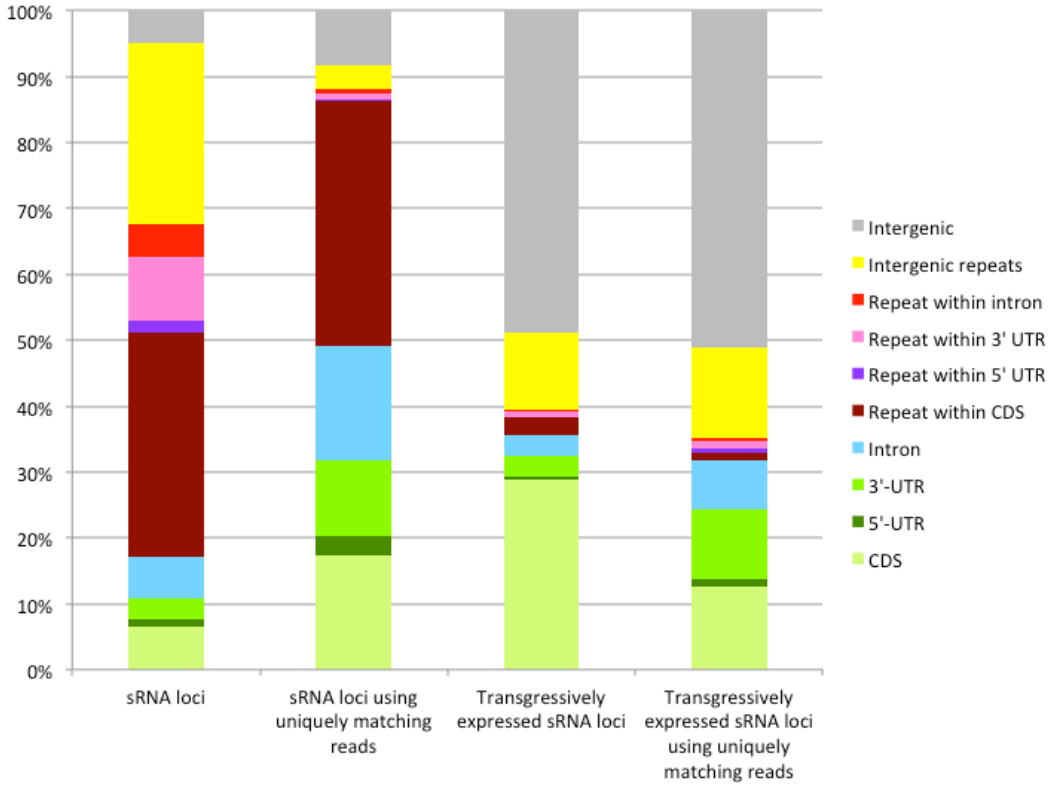


Figure 5-8: Locations of transgressively expressed sRNA loci
Results of overlap analyses with transgressively expressed loci identified using a pairwise comparison approach. Pie charts of overlap between transgressively expressed loci (predicted using either all reads or only uniquely matching reads) and genomic annotations.

My next approach was to analyse the type of sRNAs that make up transgressively expressed loci using an overlap analysis comparing the locations of different types of sRNA loci (phased, predicted miRNAs, known miRNAs, and unclassified) identified in 0 to the locations of the transgressively expressed sRNA loci. No known miRNAs were identified as transgressively expressed by the pairwise comparison and most of the transgressively expressed sRNA loci remained unclassified (Table 5-3).

This result differs to the earlier observation that most phased loci are possibly transgressively expressed (5.2.8). This is likely due to the stringent nature of the pairwise comparison.

	phased	pmiRNA	kmiRNA	unclassfiied
L2-1	0	0	0	131
L2-2	1	1	0	0
L2-3	1	2	0	134
L2-5	4	1	0	574

Table 5-3: Classes of transgressively expressed sRNAs
Results of overlap analysis between transgressively expressed loci predicted from all sRNA reads and loci classifications. Classifications include phased (any loci which had an average PhaseR value of less than -10 in any strain), known miRNAs (kmiRNA), and predicted miRNAs (pmiRNA).

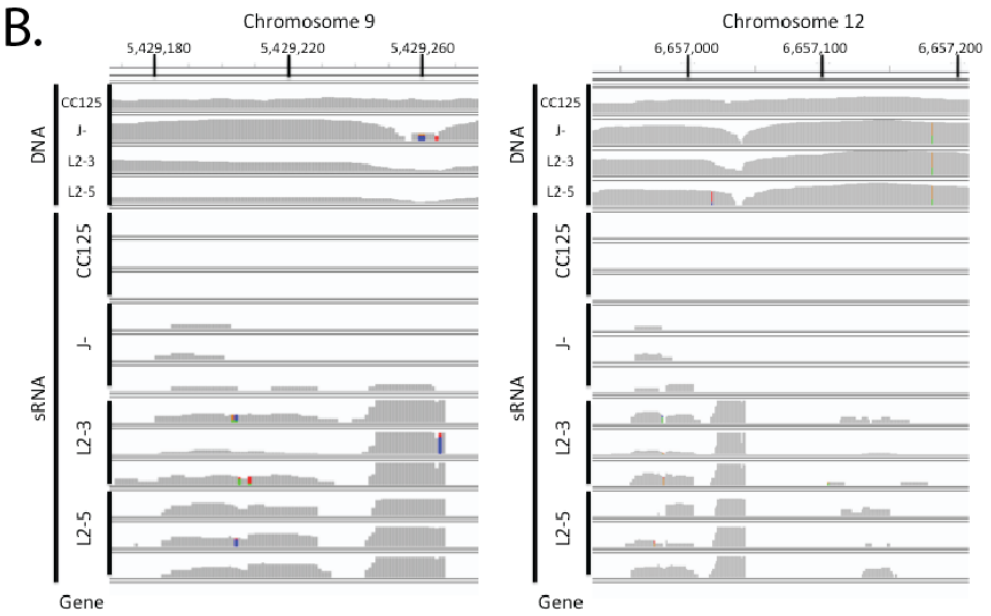
5.2.12. Further analysis of one transgressively expressed sRNA locus, TE-F-1

I wanted to further investigate a transgressively expressed miRNA identified through the pair wise comparison. A predicted miRNA identified as transgressively expressed via pair wise comparison, TE-F-1, was chosen for further analysis as it showed high expression in comparison to other miRNAs identified (Figure 5-9-A) and it's miRCat predictions included a likely precursor (Figure 5-9-C). The first step was to verify the transgressive expression observed in the sRNA libraries using sRNA northern analysis.

TE-F-1 actually matches to two places in the genome and both of these locations were identified as transgressively expressed loci (Figure 5-9-B). At both of these loci the one sRNA read, TE-F-1, was primarily responsible for the transgressive expression (Figure 5-9-B) although another read, TE-F-2, which is offset by two nucleotides, is also mildly transgressively expressed (Appendix 7.12). sRNA northerns analysis confirmed TE-F-1 transgressive expression in each recombinant (Figure 5-9-D). It is impossible to say from which location the miRNA TE-F-1 originates.

A.

Code	Transgressively expressed sRNA	Size	CC125+	J-	L21	L22	L23	L25
TE-F	TGGCGCGTCTGTGGGCTGCCT	21	0.0	3.3	0.0	0.0	29.9	27.7



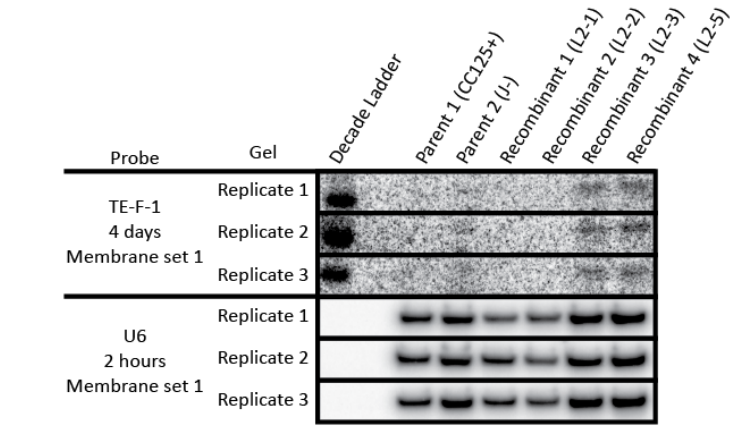


Figure 5-9: Expression pattern and verification of transgressively expressed locus TE-F-1
Via pairwise comparison of L2-3 and L2-5 sRNA loci expression to both parents, the locus TE-F was identified to have transgressive expression. **(A)** Table providing information on the sRNA, named TE-F-1, which is primarily responsible for the transgressive expression. Counts are normalized. **(B)** Visualization of the transgressive expression of TE-G using the CLC Genomics Workbench. TE-F-1 matches to two loci in the genome. The sRNA counts panel is normalized (data range = 80). The genome alignment panel (data range = 100) shows a good DNA read coverage at this locus implying there is no genetic divergence responsible for the transgressive expression. The sRNA alignment panel shows all of the sRNA reads that align to this locus including those with mismatches (quality of sequencing at that base is taken into account. **(D)** Hairpin precursor predicted by miRCat as the origin of TE-F-1. **(E)** Expression of TE-G primary transgressively expressed sRNA read was verified via an sRNA Northern. U6 was used as a loading control.

In order to gain an understanding of the potential effects of TE-F-1's transgressive expression, I used target analysis programs to predict potential targets for this miRNA. The UEA Toolkit target prediction program and Tapir (another target prediction tool) were used to identify potential genic targets of TE-F-1 (Table 5-4). Although there seems to be no gene ontology similarity between the genic targets, TE-F-1 does show binding affinity to multiple genes with annotated functions and showing various levels of expressed sequence tag (EST) expression (Table 5-4).

Although the transcriptome of an organism is usually used as the input for potential targets in target prediction programs, I also used segmentSeq predicted sRNA loci as input for potential targets in the UEA toolkit target prediction tool. In other organisms some miRNAs are known to catalyse the formation of phased secondary sRNA loci and it therefore seemed pertinent to include sRNA loci in the analysis.

TE-F-1 targets multiple phased loci, all of which are also transgressively expressed (according to d/a calculations) in the same recombinants as the miRNA (Figure 5-10). The phased targets of TE-F-1 are primarily L1-1_CR repetitive elements, which make up nearly 1% of the *C. reinhardtii* genome. Because the phased targets are generally repetitive elements it is difficult to say how many are true sRNA loci. It is possible that one sRNA locus that aligns to a repetitive element is responsible for all of the transgressive expression seen at these phased loci. Interestingly some of the phased loci also include an inverted repeat region (one of which has no sRNA reads at all).

Further investigation of these potential phased target loci showed that the locations of these loci cluster in the genome (Figure 5-11). This localisation pattern is not observed for genic targets (Figure 5-11).

Gene	Binding	Tapir	UEA	Annotation	EST expression
Cre01.g046950.t1.	..||||||||o||||||||.	✓		None.	Low
g2636.t1 & Cre02.g142550.t1	||||||||.o|||.|||||||	✓	✓	Some galacturonosyltransferase similarity	Some
g2637.t1/2 & Cre02.g142450.t1	||||||||.o|||.|||||||	✓	✓	None.	Low
g11590.t1 & Cre18.g745800.t1	o||.||||||||||o|||o	✓	✓	Exostosin-like glycosyltransferase (ELG17) with some rnd repetetive elements.	Low
Cre16.g684100.t1.2	o.||||||o|||.||||||||	✓	✓	HYPOTHETICAL HEAT DOMAIN-CONTAINING	Low
g2863.t1/Cre01.g056300.t1	||||.||||||||.|||||||	✓	✓	Zinc finger C3HC4 type (RING finger) (some rnd in 5'UTR)	Some
Cre02.g106100.t1.2	|.||||||.||o|||||||.	✓		alpha/beta hydrolase fold	Low
Cre10.g465750.t1	..||||||||||||||||..		✓	None.	Some
Cre02.g144750.t1	.||o||||o||||||||.|		✓	PTB4 Sodium/phosphate symporter	High
Cre14.g617650.t1	|.||o|||||||||.||||.		✓	PDE1 3',5'-cyclic-nucleotide phosphodiesterase signal transduction	Some
Cre12.g489750.t1	..||o||||||||||||.|o		✓	FAP52 Conserved uncharacterized Flagellar Associated Protein; found in the flagellar proteome [PMID: 15998802]; in basal body proteome as BUG14 [PMID: 15964273]. Transcript upregulated during flagellar regeneration WD40 repeat-containing protein	High
Cre10.g440950.t1	..|||.||||||||.|||||||		✓	Conserved uncharacterized Flagellar Associated Protein; found in the flagellar proteome IQ FAP192calmodulin-binding motif protein binding	High
Cre16.g688700.t1	.||||||.|||||.||||||o		✓	triacylglycerol lipase (class 3)	High
Cre13.g586400.t1	o||||||||||||.||||.|		✓	Iron/ascorbate family oxidoreductases SRG1-RELATED	High
Cre12.g541050.t1	.||||||||||||.|.|||||		✓	None.	Low
Cre16.g670100.t1	||||.|||||||||||.|.		✓	None.	Some
Cre10.g464200.t1	..||||||.||||||||||.|		✓	None.	High
Cre10.g449700.t1	.||||||..||||||||||.		✓	Universal stress protein family	High
Cre03.g186750.t1	.||||||||||||.|||.|||		✓	None.	Some
Cre06.g291450.t1	o|.|o|||||||o|||||o		✓	None.	Some
Cre08.g362500.t1	.|.|||||o||||o|||||o		✓	None.	Some
Cre09.g389450.t1	.|||o||||o||||.|||||.		✓	ubiquitin-protein ligase activity RNA recognition motif U-box domain WDSAM1 PROTEIN	High
Cre16.g682850.t1	|||||||||..||||||.|||o		✓	None.	High
g7532 & Cre26.g772250.t1	|||||.|||o|||||||.|||.		✓	None.	Some
Cre10.g454350.t1	|.|||||||||||.|||.|||o		✓	None.	Some
Cre17.g730100.t1	.||||||.||||||||.|||.||		✓	Intramembrane metalloprotease RSE1	Some

Table 5-4: Genic targets of TE-F

The targets of the transgressively expressed miRNA, TE-F-1, were predicted using the UEA toolkit target prediction tool and Tapir. The current transcriptome was used as the reference input of genes to be tested for binding sites. Binding of the miRNA is shown where I denotes a complementary nucleotide, . denotes a mismatch, and o denotes a G:U pair. Annotations were taken from Phytozome.com and EST levels were classified based on EST libraries from Sabeeha Merchant's lab.

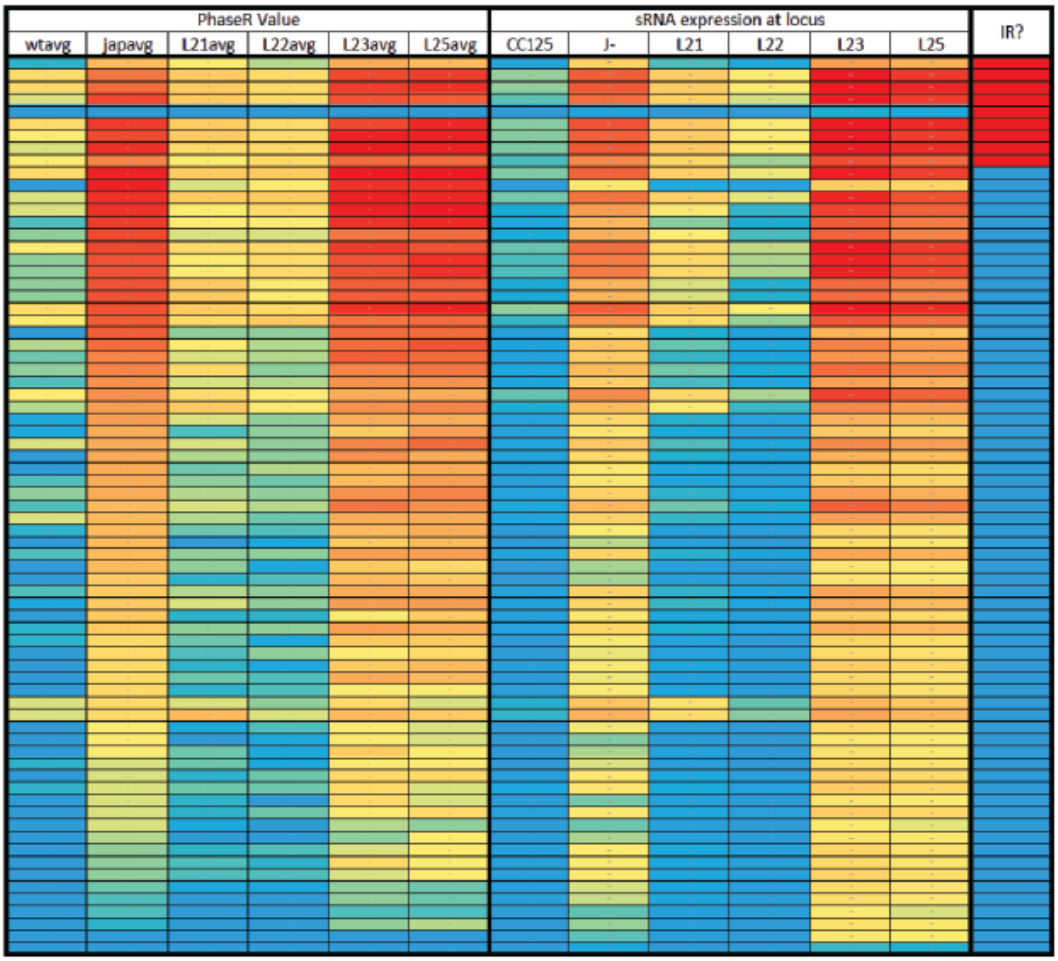


Figure 5-10: Expression and PhaseR values for the predicted target loci of TE-F-1
Each row represents a locus that is a predicted target of the transgressively expressed predicted miRNA, TE-F-1. The sRNA expression heat map was based on normalized total expression at that locus. PhaseR was used to calculate the PhaseR values for the PhaseR heat map. Each heat map is relative to the other values in that row. IR's were predicted and are indicated in red. Red indicates high sRNA expression or high likelihood of being phased while colder colours indicate less sRNA expression or likelihood of being phased.

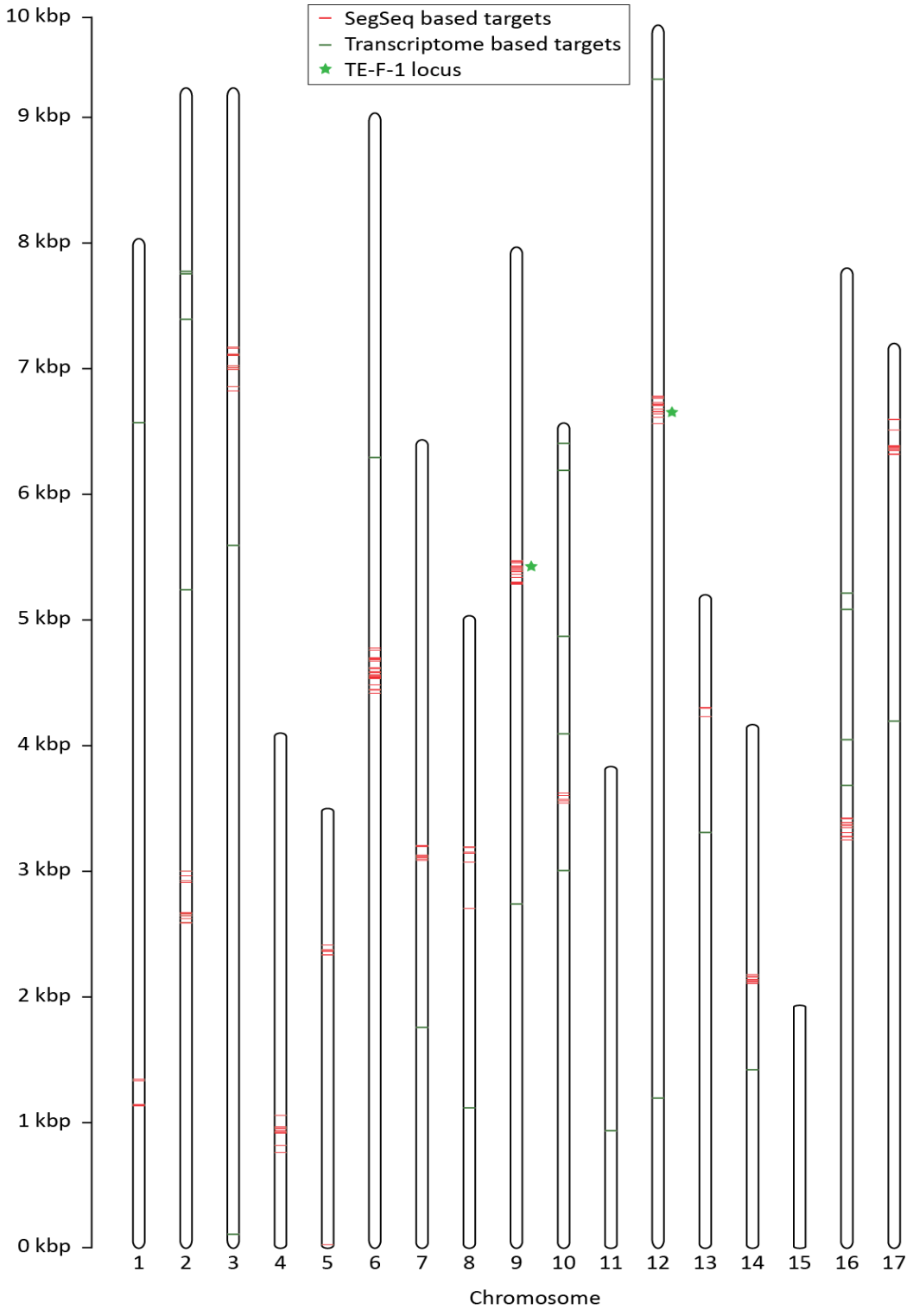


Figure 5-11: Localisation of TE-F-1 targets
Visualization of the location of TE-F-1 targets and loci on the 17 *C. reinhardtii* chromosomes. SegmentSeq targets (red) tend to cluster while genic targets (green) are spread amongst the chromosomes. The two potential loci for TE-F-1 can be found within phased target clusters. Chromosome lengths and loci positions are to scale with the axis of kbp on the left.

5.3. Discussion

In this chapter I described the process of designing an experiment with the specific aim of identifying transgressively expressed sRNA loci. The mating protocol was optimized (Figure 5-2) and after many attempts, a full tetrad of progeny was isolated from a single cross between CC125+ and J-. After the recombinant nature of the isolated tetrad strains, L2-1, L2-2, L2-3, and L2-5 was confirmed (Figure 5-3), an sRNA comparison experiment was designed (Figure 5-4) and sRNA libraries sent for sequencing.

I also catalogued the general pattern of sRNA inheritance in *C. reinhardtii* (Figure 5-6). After the sRNA library's identity and quality (Figure 5-3 and Appendix 7.6) was confirmed, I used d/a calculations to give a perspective on how many loci show additive or non-additive expression in each recombinant. This analysis also showed whether the recombinant sRNA expression was closer to the parental maximum or minimum and closer to CC125+ or J- expression.

After this broad analysis of sRNA inheritance, I used a pairwise comparison of the recombinants to parental expression to identify significant transgressive sRNA expression (Table 5-2). In the rest of the chapter I proceed to investigate the sRNA transgressive expression that I detected in the various recombinants. This analysis consisted of searching for patterns of association either with genetic elements or sRNA loci types (Figure 5-8). Having found some transgressively expressed miRNAs, I attempt to ascertain the potential affect that the transgressive expression of a single miRNA, TE-F-1, could have using target analysis (Figure 5-9/Figure 5-10/Figure 5-11).

5.3.1. sRNA inheritance in C. *reinhardtii*

Although the main purpose of this work was to identify transgressively expressed sRNA loci, the high quality of the sRNA library replicates and thus sensitivity of this experiment gave me the unique opportunity to understand the general inheritance of sRNAs in *C. reinhardtii*. By eye, I found that loci conserved between CC125+ and J-, two very genetically divergent strains, maintained that level of expression in all of the recombinants. And upon browsing genome views

of the sRNA loci differentially expressed between the parents (Chapter 4) I noted that many showed a Mendelian inheritance, with the level of sRNA expression closely linked to the genetic identity of the locus.

These observations were confirmed via d/a ratio analysis. Most sRNA loci in *C. reinhardtii* are additively expressed (Figure 5-6). Within the class of additively expressed sRNA loci, there were examples of loci that show uniparental expression and of loci whose expression in the recombinants is within the parental range (Figure 5-6). For the second type of locus, it is likely that some other element is having an effect on recombinant sRNA expression whether that is a trans factor, the affect of sRNA reads aligning to multiple locations in the genomes, or multiple loci producing sRNA reads.

Of the non-additive expression identified, there were relatively equal amounts in the recombinants and examples of transgressive expression in all directions: greater than the maximum, less than the minimum, whether those values were closer to the CC125+ or J- locus expression (Figure 5-6). This suggests that the mechanism behind the transgressive expression of sRNAs in *C. reinhardtii* is not mating type dependent although reciprocal crosses would confirm this.

5.3.2.TE sRNA loci identified and located

In comparison to the d/a analysis, the pairwise comparison identified relatively few transgressively expressed loci (Table 5-2-A). However, more transgressive expression was identified when sRNAs from repetitive elements were taken into account (Table 5-2-B). Without further analysis it is impossible to determine exactly where these multiple matching sRNAs originate from which hampered further investigation of transgressively expressed repeat associated sRNAs.

Most transgressive expression however was very subtle and less than double the parental level in any direction (Table 5-2). Interestingly previous studies have noted that difference between transgressive gene expression in hybrids and parents is usually no more than twofold (Birchler et al., 2003).

I was however able to check the association of transgressively expressed loci with genetic elements. It seems that most sRNA loci are found in repetitive

elements, whether within genes or in intergenic regions (Figure 5-8). By comparison, the transgressively expressed sRNA loci, whether predicted from total redundant reads or uniquely matching redundant reads, are located mainly in the intergenic regions with a decrease in association with repetitive elements (Figure 5-8).

The correlation between repetitive elements and transgressive sRNA expression is cross specific. In maize hybrids, most of the transgressively expressed sRNAs were repeat-associated siRNAs (Barber et al., 2012), while in *Arabidopsis* hybrids proximity or overlap with transposable elements was correlated with additive sRNA expression (Li et al., 2012); in the *C. reinhardtii* recombinants the latter seems to be the case.

5.3.3. Transgressive expression of miRNAs and phased sRNA loci

Only a few miRNAs and phased loci overlapped with the transgressively expressed loci identified via pairwise comparison (Figure 5-8). The lack of transgressive expression of miRNAs was unsurprising. If a miRNA has a biologically relevant role, it's expression is likely to be highly conserved, and miRNAs that are not highly conserved tend to have lower expression which means that their transgressive expression can be missed in the pair wise comparison approach as the change in expression is so small relative to the parental values.

Although there are few examples of transgressively expressed miRNAs (Table 5-3), their impact on other sRNA loci or genic targets could be great. Further investigation into the nature of the transgressive expression of the miRNA TE-F-1 (Figure 5-9) showed that it possibly targets, along with various genes (Table 5-4), the phased loci identified as transgressive in the same recombinants (Figure 5-10). Although this link has not been verified, the transgressive expression of the phased loci being found in only two recombinants does suggest some sort of genetic based trans factor, which very easily could be a transgressively expressed miRNA such as TE-F-1. The clustering of the phased targets, possibly around centromeric regions also suggests that these phased targets could have a

common role or biogenesis (Figure 5-11). Further work is needed to elucidate the relationship between TE-F-1 and it's targets.

The absence of phased loci overlapping with transgressive expressed loci identified via a pairwise comparison was concerning however, and a d/a ratio analysis was used as a less stringent approach of identifying transgressive expression. Strikingly there is a high level of transgressive expression amongst phased loci (Figure 5-7). This non-additive expression was primarily in the direction of the J- level and usually an up regulation. Also it was only observed in two recombinants (Figure 5-7). As phased loci in J- are associated with the retrotransposon L1-1 (Table 4-2), this increase in phased sRNA expression in two of the recombinants could indicate increased regulation of this transposable element or perhaps an increase in copy number due to hybridization.

5.4. Acknowledgements

Cancer Research Institute UK sequenced the sRNA libraries. Dr Bruno Santos ran PhaseR on my sRNA datasets for me. Code for the overlap analysis was written in collaboration with Dr Sebastian Mueller.

6. Chapter six: Discussion

In this thesis I characterized the genetic and sRNA variation between two geographically diverse strains, CC125+ and J-. I then used the recombinant strains from a single cross, L1-1, L1-2, L1-3, and L1-5, to investigate sRNA inheritance and identify transgressively expressed sRNA populations. In this discussion I will describe what this knowledge contributes to our understanding of transgressive sRNA expression, natural variation, and RNA silencing in this alga.

6.1. Transgressive sRNA expression in *C. reinhardtii* recombinants

My primary aim in this study was to search for transgressive sRNA expression in *C. reinhardtii* recombinants. To do so I designed a cross to optimize the possibility of transgressive expression of sRNA populations, choosing the most genetically divergent *C. reinhardtii* strains available as the parents. In order to draw conclusions on the mechanism of transgressive sRNA expression I also profiled the genetic and sRNA variation in both parent strains. As a result I was able to identify transgressive sRNA expression in *C. reinhardtii* recombinants in all four meiotic progeny after a single meiotic event. This study is the first identification of transgressive sRNA expression in this alga.

6.1.1. RNA silencing has the potential to create transgressive phenotypes in C. *reinhardtii* recombinants

The identification of transgressively expressed sRNA loci in recombinants confirms that RNA silencing has the potential to create transgressive phenotypes in *C. reinhardtii*. For example the transgressive expression of predicted miRNA TE-F-1 verified in the recombinants L2-3 and L2-5 (Figure 5-9) could affect its predicted targets, both genes (Table 5-4) and sRNA loci (Figure 5-10).

Transgressive sRNA expression has previously been correlated with other transgressive phenotypes in higher plant crosses. In some cases transgressively expressed sRNA loci interact with genes to create transgressive gene expression such as examples in tomato (Shivaprasad et al., 2012), rice (Zhang et al., 2014), and *Arabidopsis* (Groszmann et al., 2011; Ha et al., 2009). In other cases transgressive expression of sRNA loci has been linked to differential DNA

methylation (Groszmann et al., 2011; Shen et al., 2012; Shivaprasad et al., 2012). By comparing these higher plant studies with this study, some conclusions can be made concerning transgressive sRNA expression in *C. reinhardtii*.

6.1.2. Mechanism for transgressive sRNA expression

My thinking about mechanisms for transgressive sRNA expression is influenced in part by understanding of RNA silencing in higher plants and in part by models for transgressive effects in hybrids of higher plants.

In higher plants, RNA silencing acts both at the chromatin and posttranscriptional levels. In *C. reinhardtii* there are indications including miRNA mediated cleavage fragments of mRNA targets (Molnar et al., 2007b; Zhao et al., 2007a), phased sRNAs (Molnar et al., 2007b; Zhao et al., 2007a), and a putative RDR (www.phytozome.net) that are associated with posttranscriptional silencing. At present there is no definitive evidence of RdDM pathways or other RNA-mediated chromatin silencing in *C. reinhardtii*. However there are nuclear AGO proteins (personal communication with Dr Betty Chung) and multiple DCLs in *C. reinhardtii* and my subsequent discussion is therefore based on the possibility that hybrid and recombinant phenotypes could be influenced by both chromatin level and posttranscriptional mechanisms.

In higher plant hybrids, overdominance, dominance, and epistasis can contribute to the creation of transgressive sRNA expression. These explanations might also apply in *C. reinhardtii*. However the vegetative cells studied here are haploid and overdominance is an unlikely explanation of transgressive sRNAs identified. I considered the possibility that overdominance in diploid zygote cells might have established heritable epigenetic marks (including transgressive sRNA populations) that are inherited into the haploid vegetative cells. If that were the case the overdominant locus would produce transgressive sRNA in all four haploid progeny (Figure 6-1). Only if the overdominant locus resulted in weak or partial epigenetic marks would the ratio of transgressive expression in the progeny be less than 4:0. In most instances the transgressively sRNA were present in only one or two recombinants and an overdominant locus is therefore

unlikely (Table 5-2) however a better understanding of the molecular mechanism of sRNA inheritance is needed to rule out this scenario.

An alternative explanation invokes epistasis (either between genetic or epigenetic components) that catalyses transgressive sRNA expression in *C. reinhardtii* recombinants. For example, the potential interaction between predicted miRNA TE-F-1 (Figure 5-9) and its targets (both genic and sRNA loci) (Table 5-4 and Figure 5-10) could establish transgressive sRNA loci. If these transgressive sRNAs result in heritable epigenetic marks at the targets (Figure 1-7 and Figure 1-8), the transgressive phenotypes would be inherited even in later generation progeny in which the interacting loci are unlinked. For example, if the phased sRNA loci targeted by TE-F-1 (Figure 5-10) are secondary siRNA populations they could be part of an RdDM pathway, establishing and maintaining DNA methylation or creating a further cascade of secondary siRNA loci (Figure 1-7 and Figure 1-8).

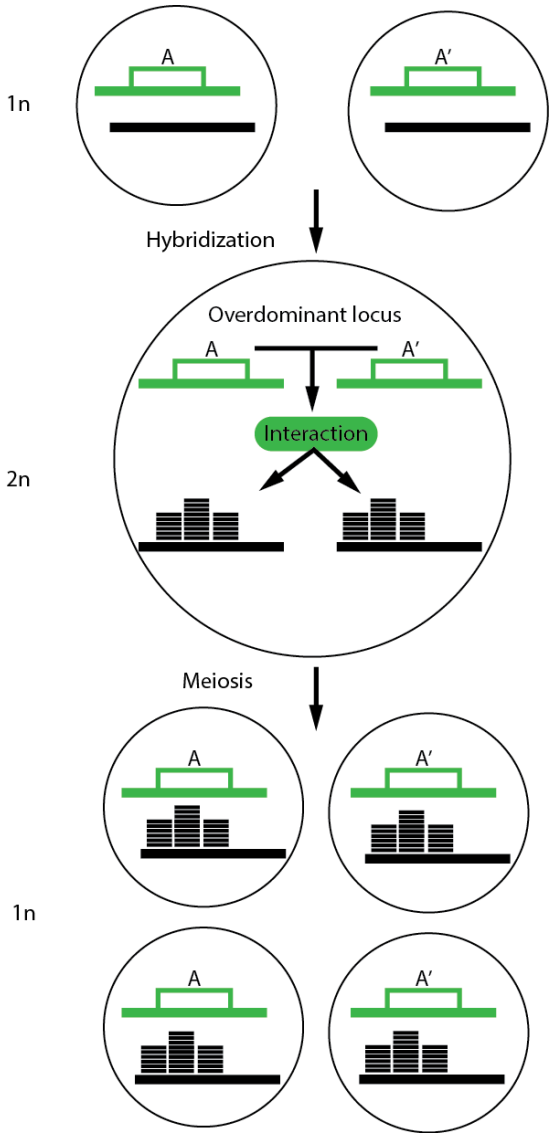


Figure 6-1: Overdominant locus setting up epigenetic marks in the *C. reinhardtii* diploid
Haploid (1n) *C. reinhardtii* could contain modified alleles of a gene (allele A and A'). In the diploid (2n) zygote, the two alleles interact, possibly resulting in the creation of a novel sRNA locus. This novel sRNA locus should be inherited in all of the haploid (1n) progeny after meiosis.

6.1.3. Further questions concerning transgressive sRNA expression in *C. reinhardtii*

While the simple existence of transgressively expressed sRNA loci in *C. reinhardtii* recombinants exemplifies their potential to create transgressive phenotypes, further work is needed to prove this link and answer outstanding questions on the mechanism.

6.1.3.1. Which factors affect the frequency of transgressive sRNA expression?

Understanding what factors determine the frequency of transgressive sRNA expression in *C. reinhardtii* could provide information on the mechanism of how transgressive sRNA expression is established. The frequency of transgressive sRNA expression could be estimated more accurately by analysis of additional crosses between CC125+ and J-. Sequencing the sRNAs of all four meiotic progeny from these crosses would identify whether the frequency of transgressive sRNA expression varies and whether the types of sRNA loci to be transgressively expressed remain the same. Sequencing the sRNA populations of recombinants from other crosses with varying levels of genetic and epigenetic variation would provide insight into how natural variation affects the frequency of transgressive phenotypes.

6.1.3.2. Does transgressive sRNA expression affect gene expression in C. reinhardtii?

Analysis of the recombinant transcriptomes, in the form of RNAseq, and ribosomal profiling would allow the correlation between transgressive sRNA expression and target gene expression to be checked. Once a target mRNA has been verified, phenotyping of the recombinant algae could expose physiological transgressive phenotypes.

I have not yet confirmed mRNA targets of transgressive sRNAs but preliminary phenotyping of the recombinant strains has indicated that transgressive phenotypes do occur in *C. reinhardtii*. When grown on solid TAP media, L2-3 and L2-5 recombinant strains exhibit higher tolerance levels to antibiotics such as Zeocin and the antibiotic Spectinomycin than either parent. To test the involvement of RNA silencing in these phenotypes I propose to repeat the hybridisation using RNA silencing mutants (such as one missing all sRNAs) in one or both of the parents.

6.1.3.3. Are transgressively expressed sRNA loci and any associated transgressive phenotypes heritable?

The heritability of the transgressive sRNA loci would also provide information on the mechanism for transgressive sRNA expression. To assess this heritability I

would carry out recurrent backcrosses of the recombinants with both parents. Transgressive sRNA production that is heritable in only a small fraction of the backcrossed recombinants would be diagnostic of dependency on interactions between different genetic elements that segregate. Transgressive sRNA expression in all/most recombinant lineages would be more consistent with an epigenetic mechanism for the transgressive phenotypes. *C. reinhardtii* RNA silencing mutants could be used in crosses to establish which pathway components are involved with the heritable component of transgressively expressed sRNA loci.

6.2. RNA silencing in *C. reinhardtii*

As well as being informative about hybrid and recombinant genomes, my study has also added to the characterisation of RNA silencing in *C. reinhardtii*. I have confirmed, for example that, as previous studies have noted, the sRNAs are predominantly 21nt in length and that U is the predominant 5' end nucleotide (Figure 5-5). I have also confirmed that miRNAs, siRNAs, and phased sRNAs previously identified in *C. reinhardtii* (Molnar et al., 2007b; Zhao et al., 2007a) were also represented in my sRNA libraries (Table 5-1). However the increased depth and sensitivity of sequencing in this study has provided novel information concerning our knowledge of RNA silencing in *C. reinhardtii*.

6.2.1. miRNAs and hairpin-associated siRNAs in C. *reinhardtii*

The initial studies identifying sRNAs in *C. reinhardtii*, identified multiple miRNAs (Molnar et al., 2007a; Zhao et al., 2007a) and to date there are 50 mature miRNAs annotated in miRBase (Griffiths-Jones et al., 2006). However a reanalysis of these miRNAs has suggested that only eight can be confidently assigned as miRNAs. The other candidates were ruled out because there were multiple siRNAs from one precursor, the miRNA* annotations were lacking, or there was a low expression level of the miRNA (Tarver et al., 2012).

In this study 282 and 78 miRNAs were predicted with miRCat for CC125+ and J- respectively using the default parameters (Appendix 7.8). The larger number of potential miRNAs identified in this study is likely due to the deeper sequencing

and higher coverage. Further experimental evidence is needed to confirm whether these are "true" miRNAs, however the stringency of the default miRCat parameters (Appendix 7.8) supports the validity of these predictions.

Preliminary parameter testing altering parameters such as the maximum number of reads aligning to the precursor radically changed not just the amount but also the species of predicted miRNAs. This suggests that while the default parameters are likely identifying classical miRNAs, there are other hairpin-associated siRNAs in *C. reinhardtii* that can be identified using relaxed miRCat parameters. Until the miRNA pathway has been characterized in *C. reinhardtii* it is not possible to accurately classify these hairpin-associated siRNAs.

Besides the increase in length and number of sRNAs aligning to hairpin precursors (Tarver et al., 2012), predicted miRNAs show another difference to plant and animal miRNAs; previous studies noted that the majority of predicted miRNAs in *C. reinhardtii* are found in introns (Naqvi et al., 2009). While the majority of animal and plant miRNAs are not found in introns, there is a class of miRNAs that originate in introns known as mirtrons. More prevalent in animals (Ladewig et al., 2012) they are also found in plants (Meng and Shao, 2012). Perhaps the intronic hairpin-associated siRNAs in *C. reinhardtii* are in fact a new class of mirtrons (Figure 4-4).

6.2.2. Secondary siRNAs in C. *reinhardtii*

In higher plants there are secondary siRNAs produced by an RDR-mediated mechanism that are produced in a characteristic phased register. Phased siRNA loci have been identified in *C. reinhardtii* (Molnar et al., 2007a; Zhao et al., 2007a), but until now it remained possible that they could be primary rather than secondary siRNAs.

In my work I have added to the data to address this point by analysis of a recent described putative RDR in the *C. reinhardtii* genome (www.phytozome.net). I have shown that there is a similar non-synonymous mutation rate of RDR between strains of *C. reinhardtii* as with the other RNA silencing components that is higher than the equivalent rate in control group of genes (Figure 3-5). This finding suggests that RDR and the other RNA silencing genes are under similar

selection pressure and is consistent with a role of this candidate gene in RNA silencing. In addition I have discovered additional phased sRNA loci in J- consistent with a secondary sRNA pathway (Table 4-1).

To further investigate the possibility of secondary siRNAs in *C. reinhardtii*, it will be necessary to analyse the candidate secondary siRNA loci in more detail. The potential miRNA TE-F-1 has multiple phased loci amongst its targets (Figure 5-10). Increases in the sRNA expression level and phasing value of these loci (Figure 5-10) correlates with the increase of TE-F-1 expression (Figure 5-9) in the recombinant strains. The possibility that a primary miRNA could be involved in the creation of secondary siRNA populations similar to the tasiRNA system could be tested by mutation of the putative miRNA target sites at these loci.

6.2.3. Role of RNA silencing in genome defence

While in higher plants and animals RNA silencing is involved in stress responses, development, genome defence, and resistance to pathogens, there is no known role of RNA silencing in *C. reinhardtii*. Preliminary phenotyping of an RNA silencing mutant with no sRNAs has yielded little difference between the wild type and the mutant (personal communication with Dr Andrew Bassett). This finding suggests that RNA silencing might play a subtler role in *C. reinhardtii* than in higher plants. Since no known virus for *C. reinhardtii* has been identified it is not yet possible to test the role of RNA silencing in pathogen defence. However a role in genome defence would be revealed through genome instability of RNA silencing mutants over several generations and through an overlap of sRNAs with repetitive elements and transposons.

A previous study reported only 11 unique siRNAs in *C. reinhardtii* that align to transposable elements (Zhao et al., 2007a). However in my work I identified 7,031 unique sRNAs in CC125+ and J- aligning to transposable elements (Figure 4-4). Most of those were siRNAs although there are a few examples of predicted miRNAs originated from transposable elements. The increase in my study could be due to a better annotation of repetitive elements by RepeatMasker (Smit et al.) and the increased depth of sequencing.

The amount of repeat associated siRNAs would normally correlate to some extent with the amount of repetitive elements in the genome (see *Arabidopsis* and Tomato in Table 6-1Table 6-1). Repeat rich genomes such as tomato have a higher percentage of sRNA loci originating from repetitive elements than repeat poorer genomes such as *Arabidopsis* (Table 6-1). It is surprising therefore that the compact genome of *C. reinhardtii* has a high proportion (76%) of its sRNA loci in repeated sequence elements (Table 6-1) and 49% of redundant reads aligning to repeats (Appendix 7.13). Furthermore, the closely related *Volvox* has a similar percentage (73%) of sRNA loci aligning to repeats (Table 6-1). This abundance of repeat associated siRNAs in *C. reinhardtii* despite its compact genome (Figure 4-4 and Figure 4-5) supports the theory of the role of RNA silencing being primarily that of genome defence in this alga.

	C. reinhardtii	*Volvox*	*Arabidopsis*	Tomato
sRNA loci	79 [1]	73 [2]	57 [4]	>90 [6]
Genome	10 [1]	23 [3]	17 [5]	63 [7]

Table 6-1: Per cent of genome and sRNA loci aligning to repetitive elements
Sources: [1]Own analysis ignoring simple repeats [2](Li et al., 2014) [3](Prochnik et al., 2010) [4](Xie et al., 2004) [5](Pérez-alegre et al., 2005) [6](Shivaprasad et al., 2012) [7](The Tomato Genome Consortium, 2012)

6.2.4. sRNA inheritance in C. reinhardtii

The inheritance of sRNA loci in *C. reinhardtii* was usually additive with the majority of sRNA loci segregating along with their parental genetic background (Figure 5-6). This is in accordance with higher plant crosses (Barber et al., 2012). Some of the transgressively expressed sRNA loci were seemingly inherited separately from the underlying parental background (Figure 5-6) indicating that there might be multiple mechanisms for sRNA inheritance.

There are some examples of maize and *Arabidopsis* hybrids where the majority of the sRNA loci exhibit transgressive expression. There was a global downregulation of 24nt siRNAs (Barber et al., 2012; Groszmann et al., 2011). However in other higher plant crosses and in the *C. reinhardtii* recombinants described here the transgressive effect was specific to a few loci (Table 5-2). A

caveat in this interpretation follows from the lack of a distinct 24nt size class of sRNA in *C. reinhardtii*: the predominant 21nt size class could include sRNAs involved in both posttranscriptional and epigenetic silencing and it remains possible that a global effect on one of these types of sRNA is hidden within the existing datasets.

6.2.5. Further questions for RNA silencing research in C. *reinhardtii*

Sequencing of sRNAs has provided insight into *C. reinhardtii* RNA silencing but there is still much to be learnt about RNA silencing. Using RNA silencing mutants that have only recently become available could answer some unknowns about this pathway in green algae.

6.2.5.1. Is RNA silencing involved in genome defence in C.reinhardtii?

Phenotyping of RNA silencing mutants would expose any physiological phenotypes affected by RNA silencing in *C. reinhardtii*. To specifically test the involvement of RNA silencing in genome defence, the mutants should be phenotyped under stress conditions that reactivate transposable elements. For example the introduction of foreign DNA, whether through transformation or genetic crossovers in meiosis can activate some transposable elements in *C. reinhardtii* (Pérez-alegre et al., 2005). Comparing the activation of transposable elements between recombinants created by wild type crosses and those created by crosses involving RNA silencing mutants could further implicate RNA silencing in genome defence.

6.2.5.2. Is RNA silencing in C. reinhardtii capable of directing secondary sRNA populations?

To show that secondary siRNAs exist in *C. reinhardtii*, a potential primary miRNA (such as TE-F-1) could be transformed into a genetic background that does not contain that miRNA. Comparing the sRNA profiles between the mutant and parent strain would expose any secondary sRNA population created by TE-F-1.

6.3. Natural genetic and sRNA variation in *C. reinhardtii*

As part of this study to identify transgressively expressed sRNAs, it was necessary to sequence the genomes and sRNAs of various *C. reinhardtii* strains. The high coverage and depth of DNA and sRNA sequencing allowed me to make some novel observations about natural genetic and sRNA variation in *C. reinhardtii*.

6.3.1. Genetic variation in geographically isolated C. *reinhardtii*

Genome sequencing of North American and Japanese strains revealed their genetic differences (Section 3). The Japanese strains, both plus and minus, had a SNP frequency relative to the reference strain that was more than 10× higher than that of the North American strains, CC125+ and CC124- (Table 3-4). To further support that CC125+ is the closest related strain to the reference genome, CC125+ had a lower SNP frequency (0.1 SNP/kb) than CC124- (1.0 SNP/kb); a similar pattern was noted in a previous study (0.1 and 0.9 SNP/kb respectively) (Lin et al., 2013).

I expected that the most geographically distant strains to the isolation site of CC125+/CC124- would be the most genetically divergent. However the number of SNP/InDels reported recently for CC1952- (a *C. reinhardtii* strain isolated in Minnesota) relative to the reference strain is similar to that of the Japanese strains (Lin et al., 2013). A recent study confirmed that CC125+/CC124- and CC1952- originate from different geographical subpopulations of the species (Jang and Ehrenreich, 2012).

This pattern of variation indicates that geographical distance is not necessarily a predictor of genetic variance in *C. reinhardtii*; instead ecological distributions might be a more accurate predictor of genetic variance (Collins and de Meaux, 2009; Ichimura, 1996). It seems that strains that are close together can still differ greatly (Jang and Ehrenreich, 2012) and presumably do not have the opportunity to form hybrids. It will be necessary to compare CC1952- and J+/J- in order to find out whether they share SNPs and other genome differences relative to the reference strain. Further profiling of the genetic variation would also contribute to understanding how genomes diverge in naturally isolated populations.

The genetic variation between J- and CC125+ strains described in Chapter 3 could cause differential gene expression directly if it affects regulatory elements in genes. There is also the possibility of indirect effects on gene expression due to genetic variation creating differentially represented or differentially expressed sRNA loci.

6.3.2. Genetic variation can cause differential sRNA representation between strains

Nearly two thirds of sRNA loci were differentially represented (Figure 4-7) between the two strains. Of these, genetic variation in the form of InDels caused over a third of the differential representation (Figure 4-9). These instances of natural variation in sRNAs between CC125+ and J- can be used to investigate how novel sRNA loci are created. For example, the RNA silencing machinery could target novel DNA sequence created by an InDel. This might be the case concerning the phased siRNA loci specific to J- (Table 5-1) where preliminary DNA mapping analysis suggests that these loci might be due to the proliferation of the L1-1 transposon (Table 4-2 and Figure 4-6). InDels could also result in RNA silencing targeting novel secondary structures; for example, inverted duplications in *Arabidopsis* create novel miRNAs (Cuperus et al., 2011). It is also possible that novel epigenetic modifications associated with an InDel are targeted by RNA silencing.

Less than ten sRNA populations were identified in the *C. reinhardtii* recombinants that had no expression in either parental strains and deeper sequencing could still yield some parental expression. The probability of identifying the creation of a novel sRNA population in one meiotic event is low as even in larger studies the *de novo* creation of an sRNA locus was rare (Barber et al., 2012). The creation of novel sRNA loci in this study could however be masked by the alignment of the sRNA reads to the current reference genome. For example if a novel sRNA locus were created by an insertion, then those novel sRNA reads, when aligned to the reference genome, would align to a similar region, creating a transgressively expressed locus. To conclusively check for novel sRNA loci, the genomes of the recombinants should be assembled and those used to identify sRNA loci.

6.3.3. Genetic variation can cause differential sRNA expression between strains

The other two thirds of the differential sRNA representation was due to differential expression of shared loci (Figure 4-9). Some of this differential expression may be due to genetic variation in the form of SNPs or small InDels that are within or close to the sRNA loci (Figure 4-10).

Differential expression of an sRNA locus could also be created from trans-acting differences in the protein components of the RNA silencing pathway (Figure 3-5). It could be, for example, that variation in the structure of a DCL could affect the affinity of this protein for specific nucleotide motifs in sRNA precursors. An example of this type of effect is illustrated by natural variation in the AGO2 protein of *Drosophila*; the substitutions are primarily located at the protein surface and might be indicative of the altered ability of AGO2 to interact with other molecules (Obbard et al., 2011).

Genetic variation creates much of the differential sRNA expression between CC125+ and J-. However it is possible for some of the differential expression to be caused by differential patterns of epigenetic marks such as methylation or histone modifications.

6.3.4. Further questions

Sequencing of the DNA and sRNAs in geographically distant *C. reinhardtii* strains has exposed the high level of natural variation available in this alga. Specifically it shows that the level of sRNA variation can be equally as high as that as genetic variation. Further experiments could expose how this sRNA variation was created.

6.3.4.1. Does the genetic variation in RNA silencing pathway components create differential sRNA expression?

Mapping the non-synonymous variation in the AGO and other RNA silencing proteins of *C. reinhardtii* onto structural models of these proteins would provide insight into the effect of this variation. Identifying the more conserved domains could suggest which functions of these proteins are more important. Additionally RNA silencing mutants could be complemented with genes from both the North

American and Japanese strains. The resulting gain of RNA silencing could then be compared between the complemented strains to learn the affect of genetic divergence in these genes.

6.3.4.2. What is the mechanism by which an InDel creates a novel sRNA locus?

Over a third of differentially represented sRNA loci are found in InDels between the two strains. I would concentrate further analysis on the creation of novel miRNAs as these are the best characterized sRNA species in *C. reinhardtii* and models have been proposed for miRNA locus creation. Models for miRNA evolution includes the inverted gene duplication theory, the random birth theory, and the theory that miRNAs arise from transposable elements (Piriyapongsa and Jordan, 2008; Shabalina and Koonin, 2008; Zhou et al., 2013). The sequences of those InDels could be compared and checked for secondary structures or sequence conservation that might explain the creation of a novel sRNA locus.

6.3.4.3. What sRNAs are conserved in C. reinhardtii?

This study concentrated on variation between the strains, but conserved sRNA loci could also provide information on the evolution of RNA silencing. For example the conservation of the size profile between American and Japanese strains implies a mechanistic constraint on *C. reinhardtii* sRNA lengths. The sRNA loci conserved between American and Japanese strains could be further analysed to see if there are any shared features such as sRNA class or location.

7. Appendix

7.1. Oligonucleotide list

7.1.1. PCR primers

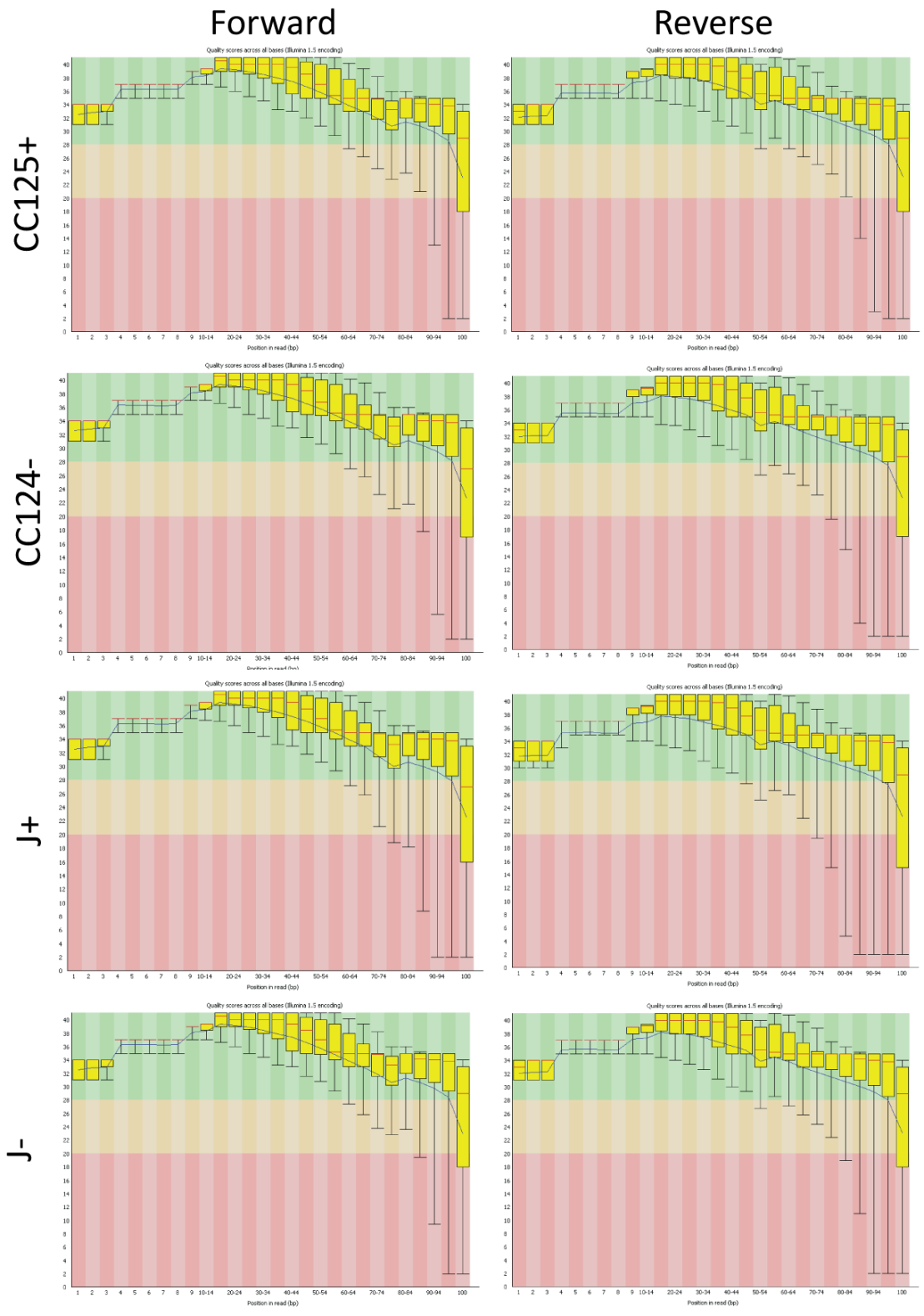
Name	Sequence
18S-FA	AACCTGGTTGATCCTGCCAGT
18S-RB	TGATCCTTCTGCAGGTTCACCTAC
ITSa	GGGATCCGTTTCCGTAGGTGAACCTGC
ITSb	GGGATCCATATGCTTAAGTTCAGCGGGT
FUS1 up	ATGCCTATCTTTCTCATTCT
FUS1 down	GCAAAATACACGTCTGGAAG
MID up	ATGGCCTGTTTCTTAGC
MID down	CTACATGTGTTTCTTGACG

7.1.2. Northern oligonucleotides

Name	sRNA sequence	Probe sequence
TE-F-1	TGGCGCGTCTGTGGGCTGCCT	AGGCAGCCCACAGACGCGCCA
TE-F-2	TGTGGCGCGTCTGTGGGCTGCCT	AGGCAGCCCACAGACGCGCCACA
U6	GGATGACACGCATAAATCGAG	CTCGATTTATGCGTGTCATCC

7.2. Quality of DNA libraries

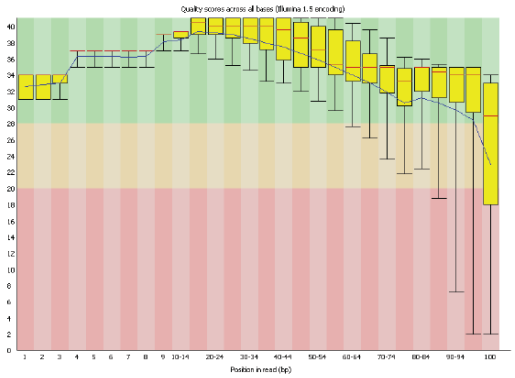
FASTQC was used to assess the quality of the reads returned from the HiSeq. Figures show a boxplot of quality scores at increasing positions along the read. Quality decreases over the length of the read as expected. However most positions have an average quality score above 20, which is the accepted cut off.

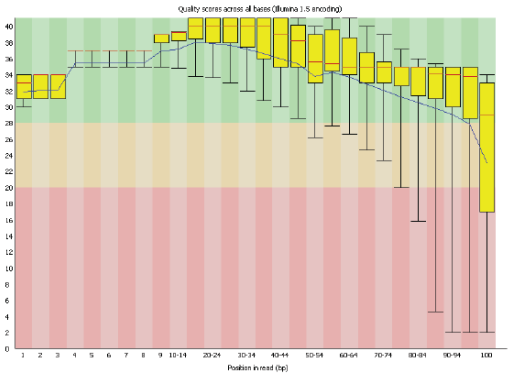

Forward
Reverse
CC125+
CC124-
J+
J-
Quality scores across all bases (Illumina 1.5 encoding)
Position in read (bp)

Forward

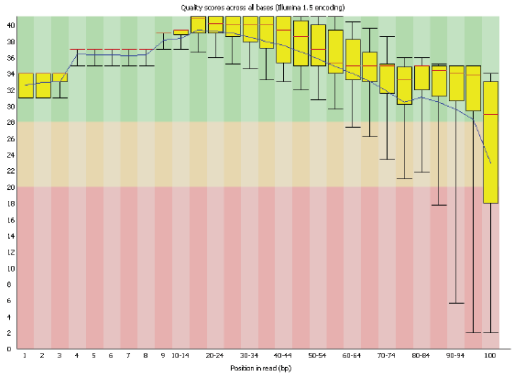
Reverse

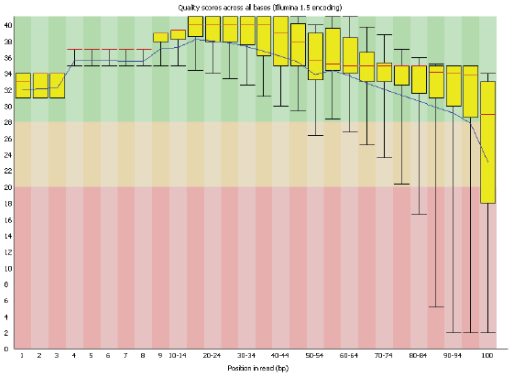
L2-1



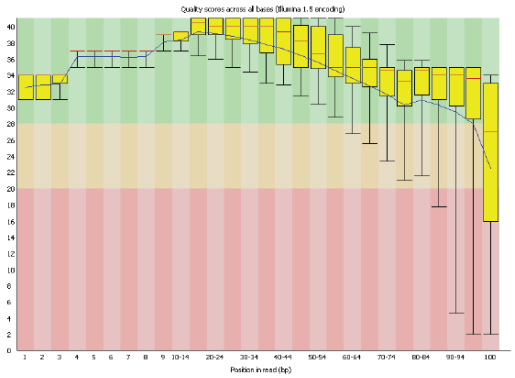


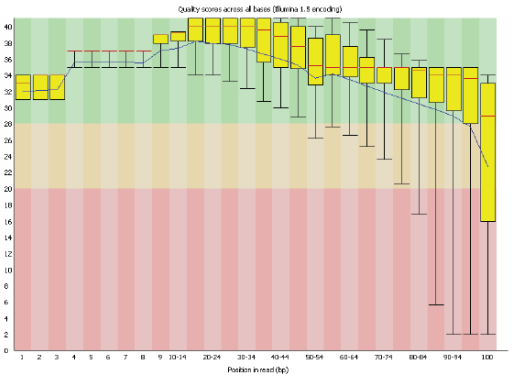
L2-2



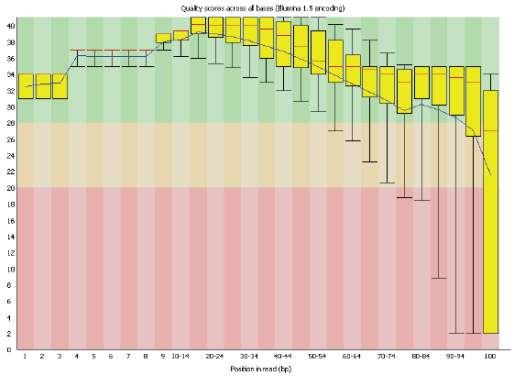


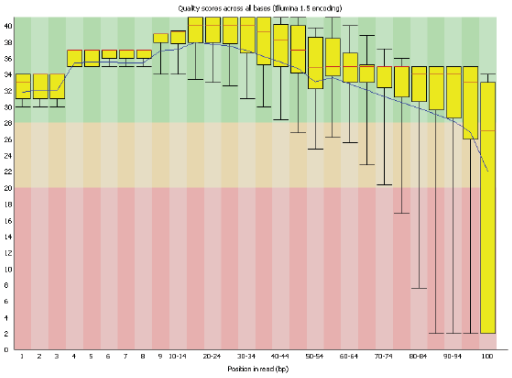
L2-3



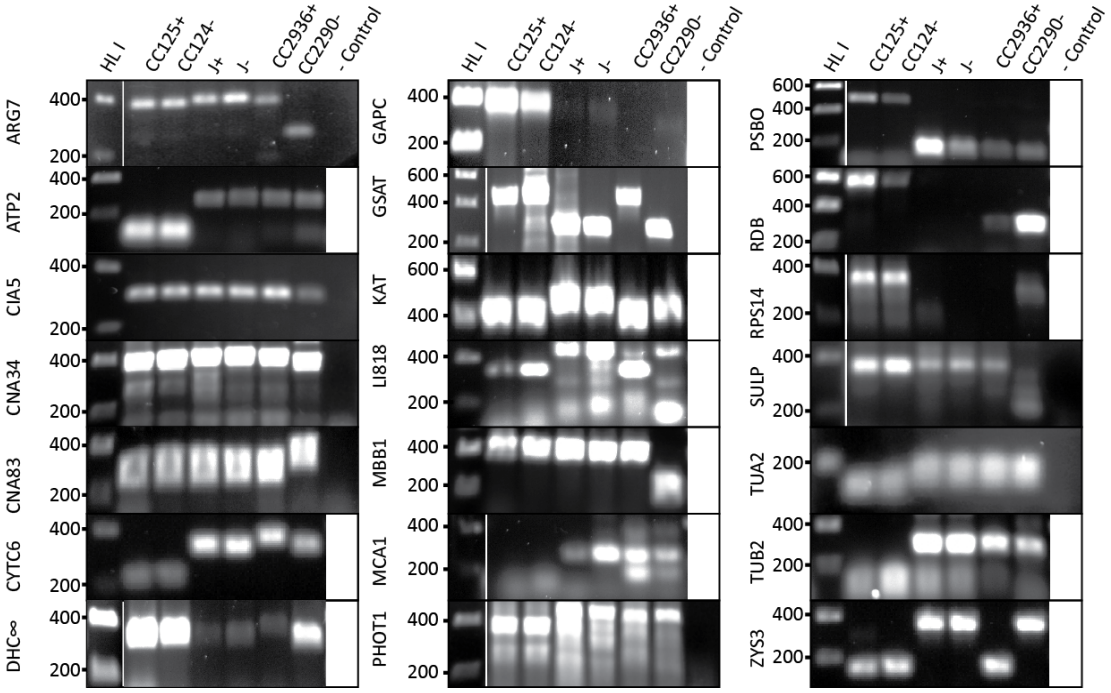


L2-5





7.3. Mapping marker PCR results



Mapping marker PCR products were run on 1.5% agarose gels stained with EtBr. Markers were labelled as similar to CC125+ or CC2290- or strain specific (to either J+, J-, or SS2936+).

7.4. List of control proteins

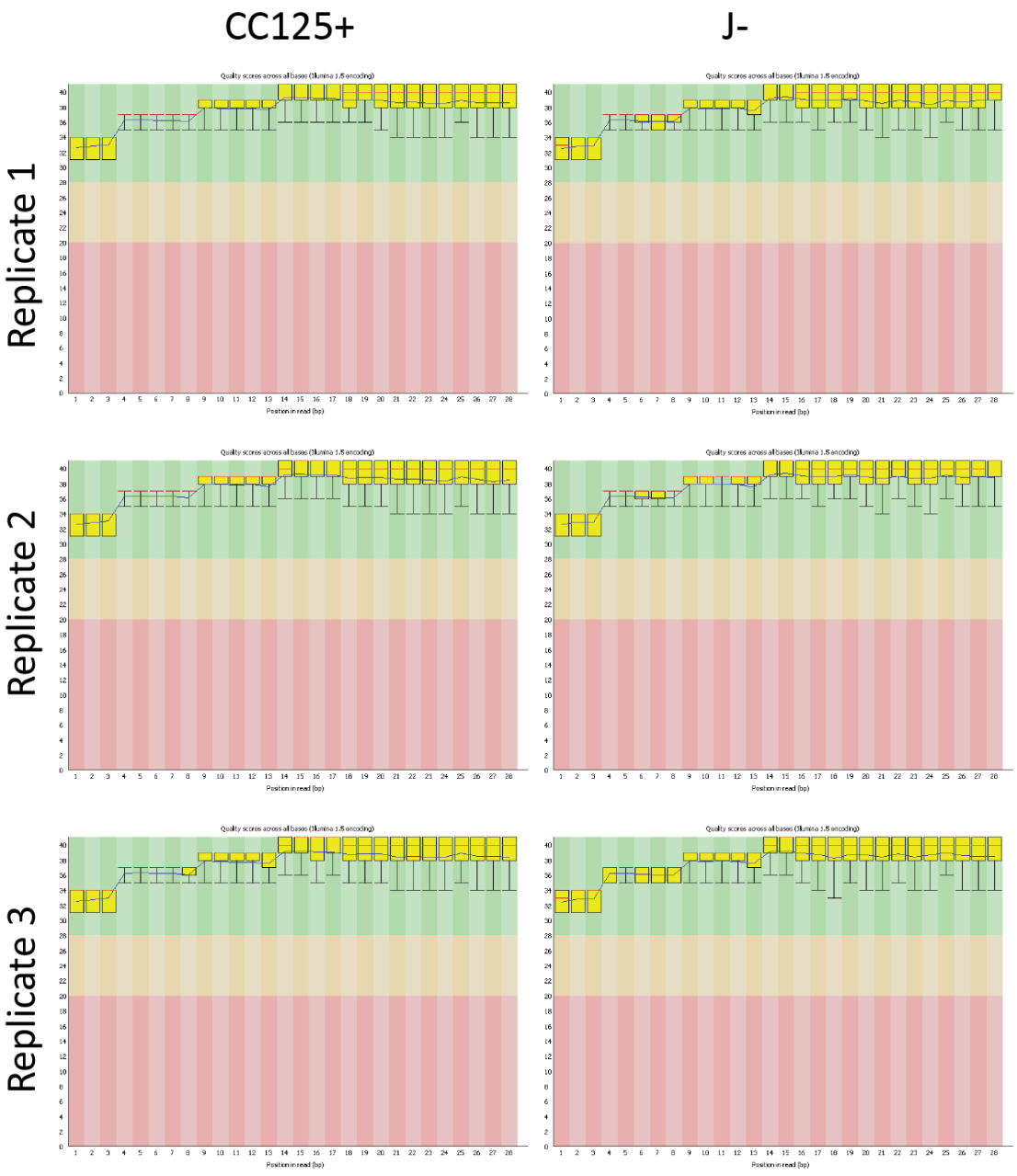
Abbreviation	Identifier	Description
PDK1	Cre06.g252300	Pyruvate dehydrogenase kinase
PDK2	Cre06.g278450	Pyruvate dehydrogenase kinase
PDK3	Cre05.g241750	Pyruvate dehydrogenase kinase
MAT3	Cre06.g255450	Retinoblastoma protein involve in cell division
SFA	Cre07.g332950	SF-assemblin (structural component of flagellar roots)
YPT4	g9677	Small Rab-related GTPase
CBLP	g6364	Receptor of activate protein kinase C
PETC	g11619	Cytochrome B6-F complex Fe-S subunit
ACTIN	Cre13.g603700	Multi-functional protein that forms microfilaments
CIA5	g1955	Master regulator for the carbon concentrating mechanism
WRK1	Cre04.g228400	WRKY transcription factor
DHC8	Cre16.g685450	Flagellar inner arm dynein heavy chain
TUA1	Cre03.g190950	Alpha tubulin component of microtubules

7.5. Quality-based variant detection analysis

		Genetic				Exonic				Non-synonymous			
		CC125+	CC124-	J+	J-	CC125+	CC124-	J+	J-	CC125+	CC124-	J+	J-
RNA silencing pathway genes	AGO1	0	0	164	200	0	0	47	62	0	0	16	20
	AGO2	0	0	87	58	0	0	20	9	0	0	0	0
	AGO3	0	59	97	66	0	41	64	37	0	1	1	1
	DCL1	1	1	332	399	0	0	121	135	0	0	56	63
	DCL2	0	49	591	455	0	13	348	308	0	4	128	94
	DCL3	0	0	417	304	0	0	247	187	0	0	97	68
	RDR	0	0	204	153	0	0	100	73	0	0	41	26
Control gene set	PDK1	0	116	10	74	0	35	4	17	0	6	0	3
	PDK2	0	0	14	15	0	0	10	10	0	0	3	2
	PDK3	0	0	81	72	0	0	31	9	0	0	0	0
	MAT3	0	59	50	50	0	44	29	31	0	5	1	3
	SFA	0	0	94	104	0	0	44	50	0	0	0	0
	YPT4	0	0	29	27	0	0	14	12	0	0	0	0
	CBLP	0	0	740	573	0	0	347	256	0	0	101	62
	PETC	0	0	34	31	0	0	18	15	0	0	1	0
	ACTIN	0	0	121	111	0	0	31	35	0	0	0	0
	CIA5	0	0	52	48	0	0	18	17	0	0	1	0
	WRK1	0	0	63	63	0	0	32	34	0	0	1	2
	DHC8	0	0	558	449	0	0	234	191	0	0	11	5
	TUA1	0	0	11	13	0	0	4	6	0	0	0	0

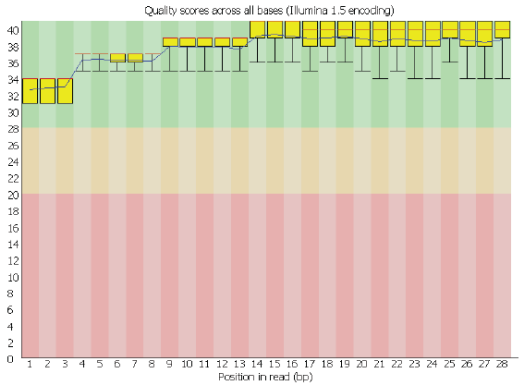
7.6. Quality of sRNA libraries

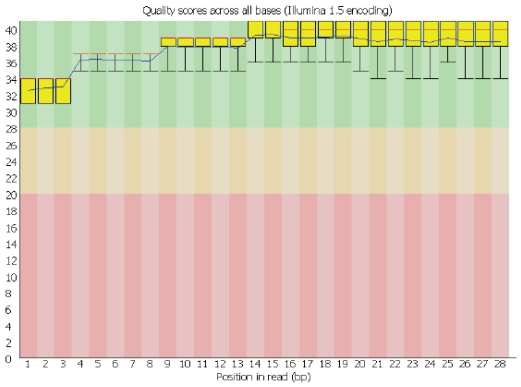
FASTQC was used to assess the quality of the reads returned from the HiSeq. Figures show a boxplot of quality scores at increasing positions along the read.



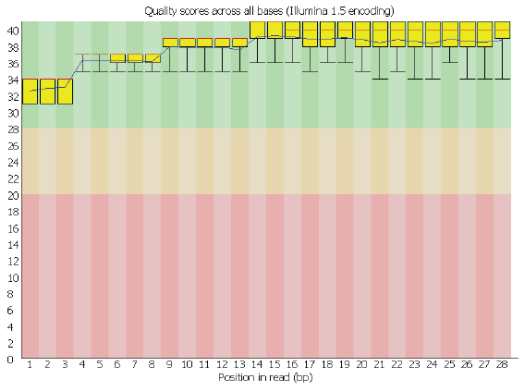
L2-1 L2-2

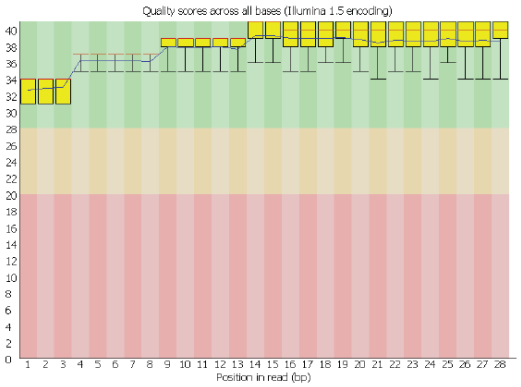
Replicate 1



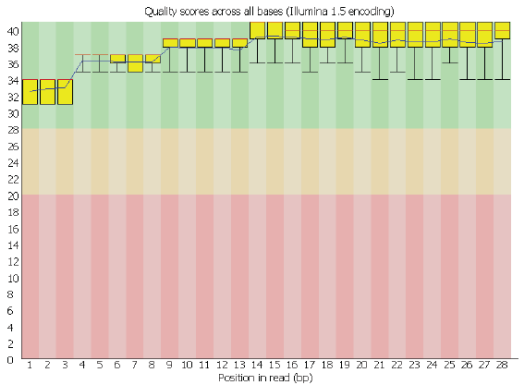


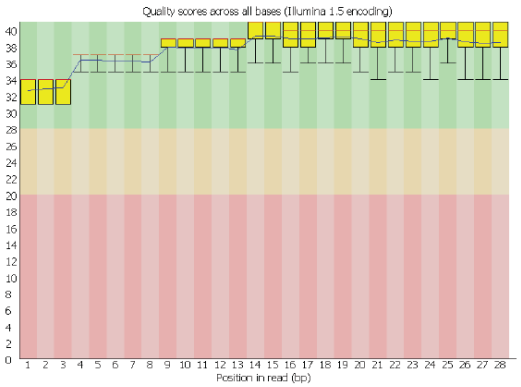
Replicate 2





Replicate 3

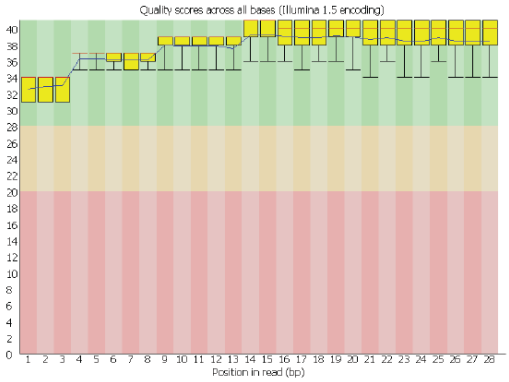


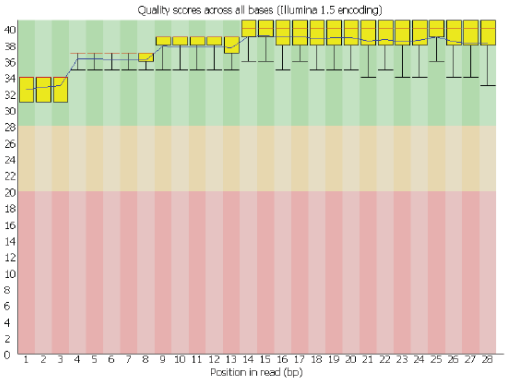


L2-3

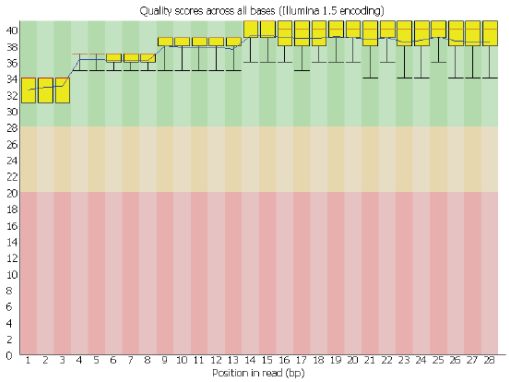
L2-5

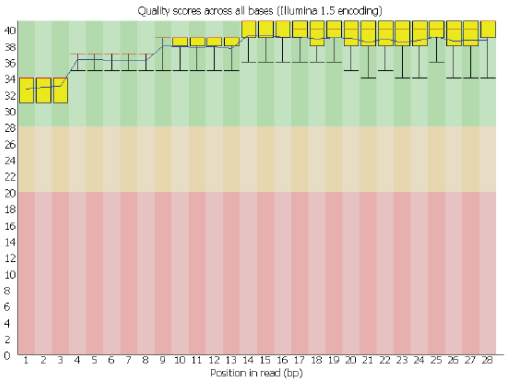
Replicate 1



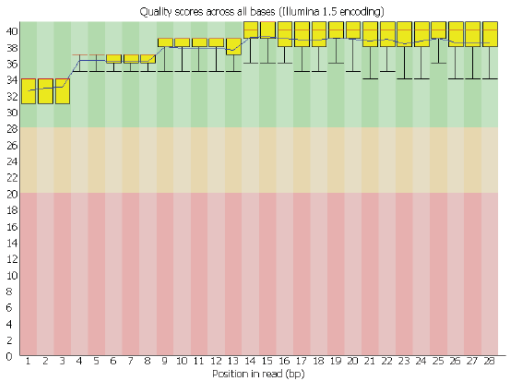


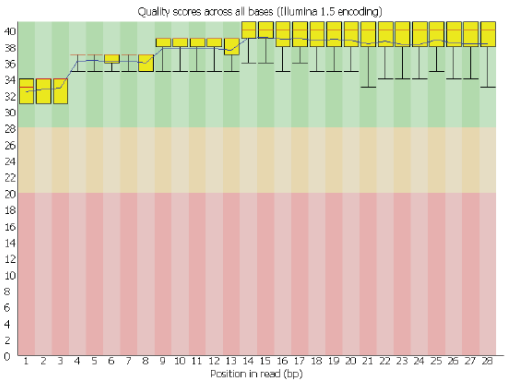
Replicate 2



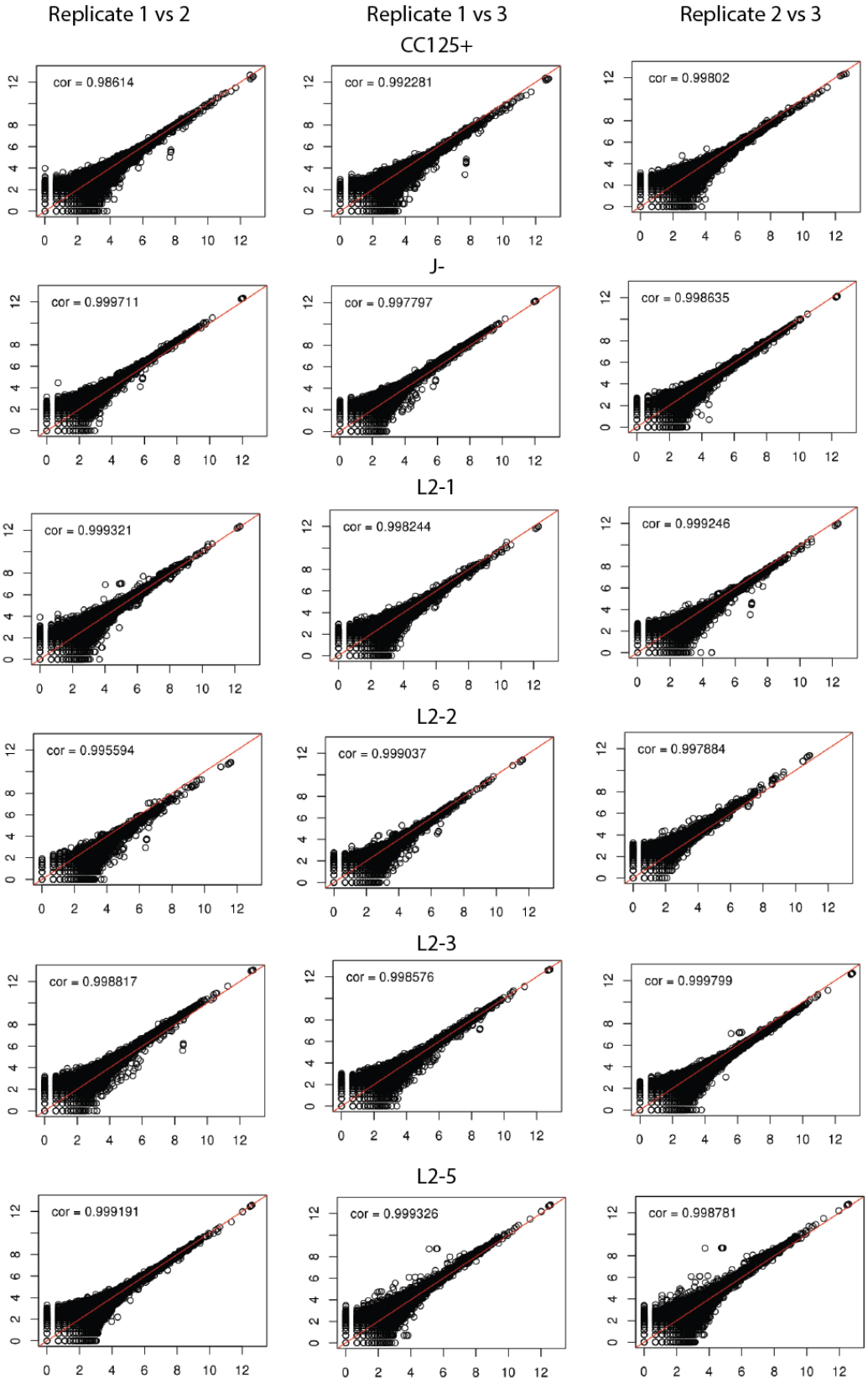


Replicate 3





7.7. Small RNA replicate confirmation



Log of loci sRNA expression plotted on each axis for different replicates.

Replicates are of good quality and exhibit similar loci expression.

7.8. miRCat default parameters

Parameter name	Parameter description	Value
Min Abundance	Min sRNA abundance (#)	5
Min Hairpin Length	Min length of hairpin (nt)	75
Max Overlap Length	Max total length of overlapping sRNAs (nt)	70
Max Size	Max length of a miRNA (nt)	21
Max Unique Hits	Max # of non-overlapping hits in a locus	3
% orientation	% of sRNas in locus that must be in the same orientation	90
Window length	window length (nt)	150
Max % unpaired	Max % of unpaired bases in hairpin	50
Max genome hits	Max # of genome hits	16
Max hit distance	Max distance between consecutive hits on the genome (nt)	200
Max Gaps	Max # of consecutive unpaired bases in miRNA region	3
Min MFE	Min free energy of the hairpin	-25
Min GC	Min % of G/C in miRNA	10
Min paired	Min # of paired bases in miRNA region	17
Min Size	Min sRNA size (nt)	20
Complex loops	Allow hairpins with complex loops?	no
pval	Max p-value	0.1

7.9. Example segmentSeq code

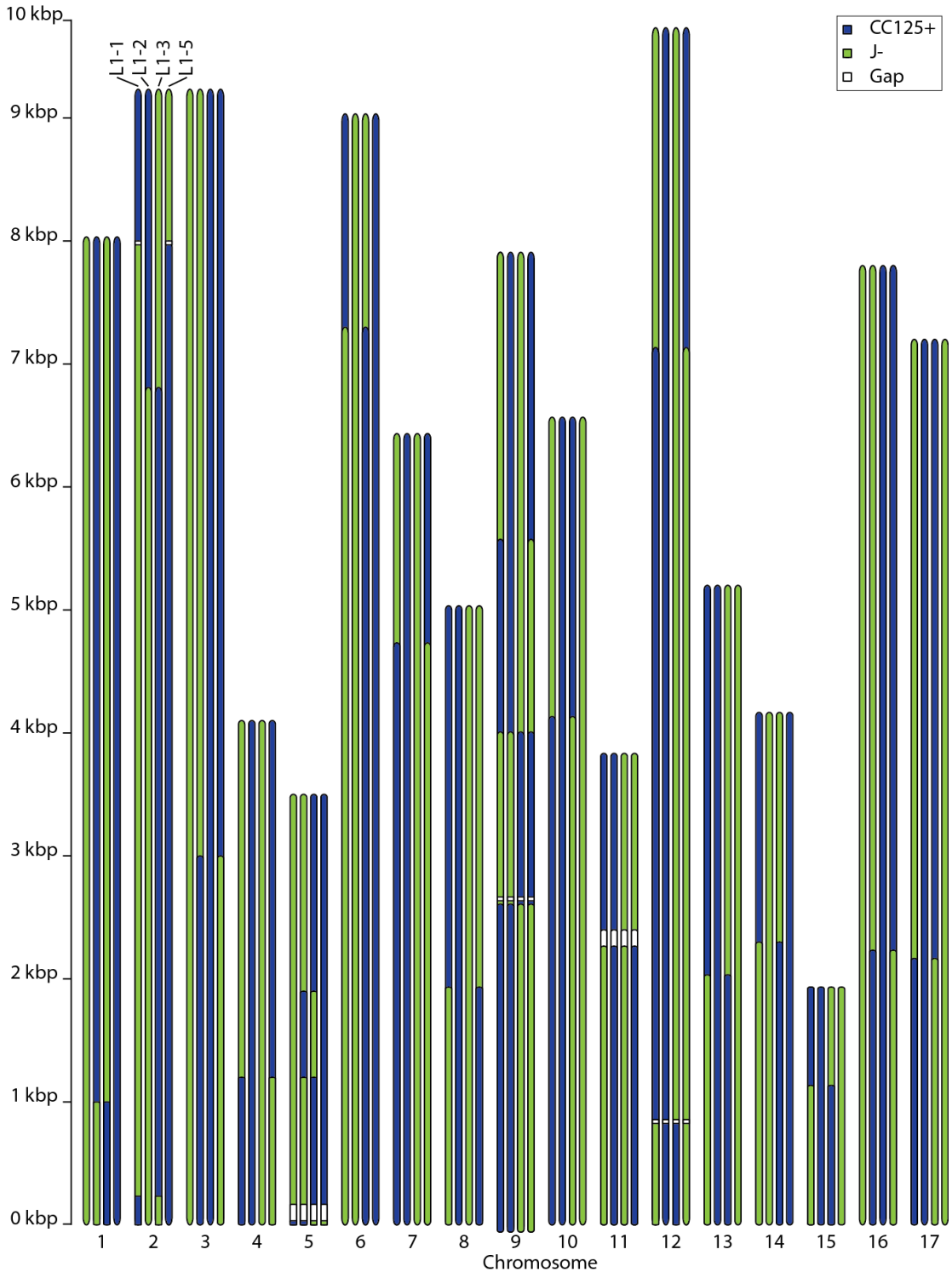
```
# Load all necessary definitions for segmentSeq analysis
library(segmentSeq)
datadir <- "/home/"
libfiles <- c("125_1.bam", "125_2.bam", "125_3.bam","J-_1.bam","J-_2.bam","J-_3.bam")
libnames <- c("CC1251", "CC1252", "CC1253","J1", "J2", "J3")
replicates <- c("CC125", "CC125", "CC125","J","J","J")
chrs = c("chromosome_1",”chromosome_2"…)
chrlens <- c(8033585,9219486,…)

# Create alignment (aD) and segmentation (sD) data files
aD <- readBAM(files = libfiles, dir = datadir, replicates = replicates,libnames = libnames,
chrs = chrs,chrlens = chrlens)
sD <- processAD(aD, gap = 100)

# How to create alignment (aDu) and segmentation data (sDu) files of uniquely mating
redundant reads
aDu <- aD[aD@alignments$multireads==1,]
sDu <- processAD(aDu, gap = 100, cl=cl)

# Heuristic (heurSeg) and class segmentation (classSeg)
heurSeg <- heuristicSeg(sD = sD, aD = aD, RKPM = 1000)
classSeg <- classifySeg(sD = sD, aD = aD, cD = heurSegall, samplesize = 1e5)

# Define hypotheses in classSeg object
CON <- c(1,1,1,1,1,1)
TE <- c("X125","X125","X125","J","J","J")
groups(classSegall) <- list(CON,TE)

# Identify differential represented sRNA loci (DRloci) and conserved sRNA loci (CONloci)
with a likelihood cut off of 0.9
classSegall <- getPriors.NB(classSegall, samplesize = 1e5, cl=cl)
classSegall <- getLikelihoods.NB(classSegall, nullData = TRUE, cl=cl)
DEloci<- topCounts(classSegall,normaliseData = TRUE, group =2, likelihood= 0.9)
CONloci<- topCounts(classSegall,normaliseData = TRUE, group =1, likelihood= 0.9)
```

7.10. Example MA plot code

```
# Create an alignment data (aD) object containing the ailgnments and counts of sRNA
reads
library(segmentSeq)
datadir <- "/home"
libfiles <- c("125_1.bam", "125_2.bam", "125_3.bam","J-_1.bam","J-_2.bam","J-_3.bam")
libnames <- c("CC1251", "CC1252", "CC1253","J1", "J2", "J3")
replicates <- c("CC125", "CC125", "CC125","J","J","J")
chrs = c("chromosome_1",”chromosome_2"…)
chrlens <- c(8033585,9219486,…)
aD <- readBAM(files = libfiles, dir = datadir, replicates = replicates,libnames = libnames,
chrs = chrs,chrlens = chrlens)

# Add a logical vector to the aD object indicating if an sRNA is a miRNA with miRNA name
aD@alignments$ismiR <- aD@alignments$tag %in% mirnas2[,1]

# Use a subset of aD discarding sRNA reads with less than 5 counts
aDsub <-  aD[rowSums(aD@data) > 5,]

# Create countData (cD) object for baySeq analysis
groups = list(NDE = c(1,1,1,1,1,1),DE = c(1,1,1,2,2,2)),
annotation=data.frame(seqs=aDsub@alignments$tag))
libsizes <- aDsub@libsizes
CDir <- new("countData", data = aDusub@data, replicates =
rep(1:2,each=3),groups=groups,annotation=annotation,libsizes=libsizes)

# Call differentially expressed sRNA reads (DEreads)
CDreadsP <- getPriors.NB(CDir, samplesize = 1e5, estimation = "QL", cl = NULL)
CDreadsL <- getLikelihoods.NB(CDreadsP, pET = 'BIC', cl = NULL)
DEreads <- topCounts(CDlociL, group ="DE",number=nrow(CDreadsL),normaliseData =
TRUE)

# Applying a false discovery rate (FDR) cut off
isSig <- aDsub@alignments$tag %in% as.character(DEreads[DEreads$FDR.DE <
0.001,]$annotation)

# Adding miRNA annotations
mirnas[,2] <- sapply(strsplit(rownames(mirnas2)," "),function(x) x[5])
miRnamelist <- rep(NA,length(aDsub@alignments))
for (i in 1:length(aDusub@alignments)) {
   if (aDusub@alignments$tag[i] %in% mirnas3[,1]) {
      miRnamelist[i] <- mirnas3[mirnas3[,1] == aDusub@alignments$tag[i],2]
   }
}
aDsub@alignments$miRname <- miRnamelist
mirnasindex <- which(!is.na(aDsub@alignments$miRname))

# MA plot with DEreads colored red (plotMA.CD script provided by Dr Sebastian Mueller)
jpeg(file="MAplot.jpg",width = 800, height = 800)
ma <- plotMA.CD(CDir, samplesA = "1", samplesB = "2",pch=16,col=isSig+1)

# Indicate miRNA reads with green circle
points(ma[mirnasindex,"A"],ma[mirnasindex,"M"],col=aDusub@alignments$ismiR+3,srt=
50,pch="O")
```

7.11. Genetic background of the recombinant strains

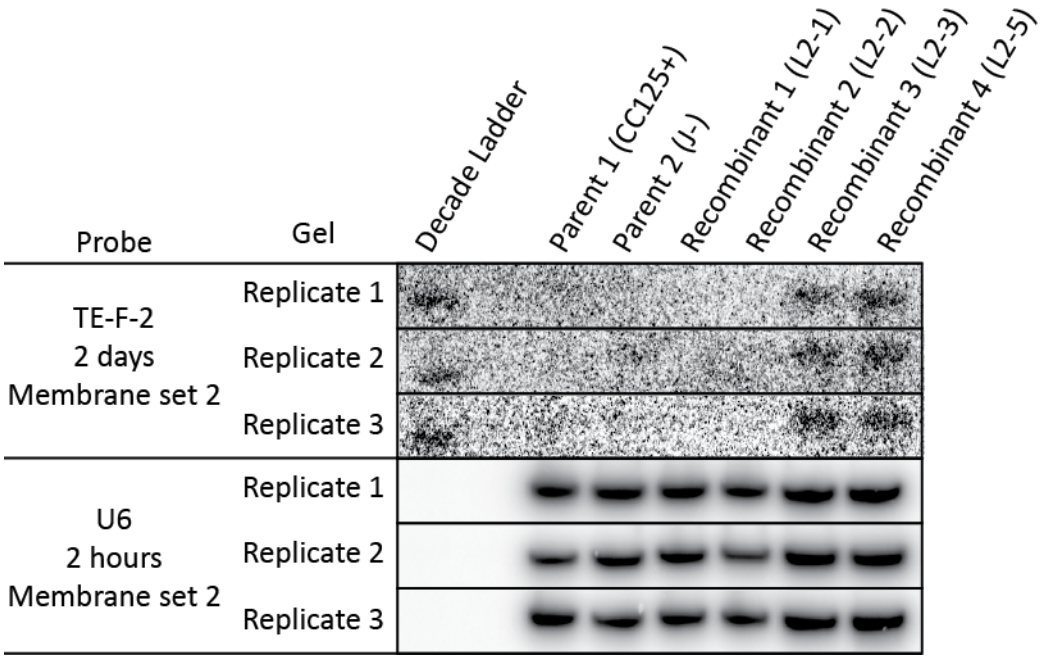
Parental background was assigned to different sections of the recombinant chromosomes based on scanning by eye for SNPs with using IGV. Recombination sights were identified as changes in the parental background in a chromosome. Some sites of recombination were located on gaps in the DNA alignment in all four recombinants and the parents and these are indicted in the figure below. Chromosome sizes are to scale (0.03 mm/kb).



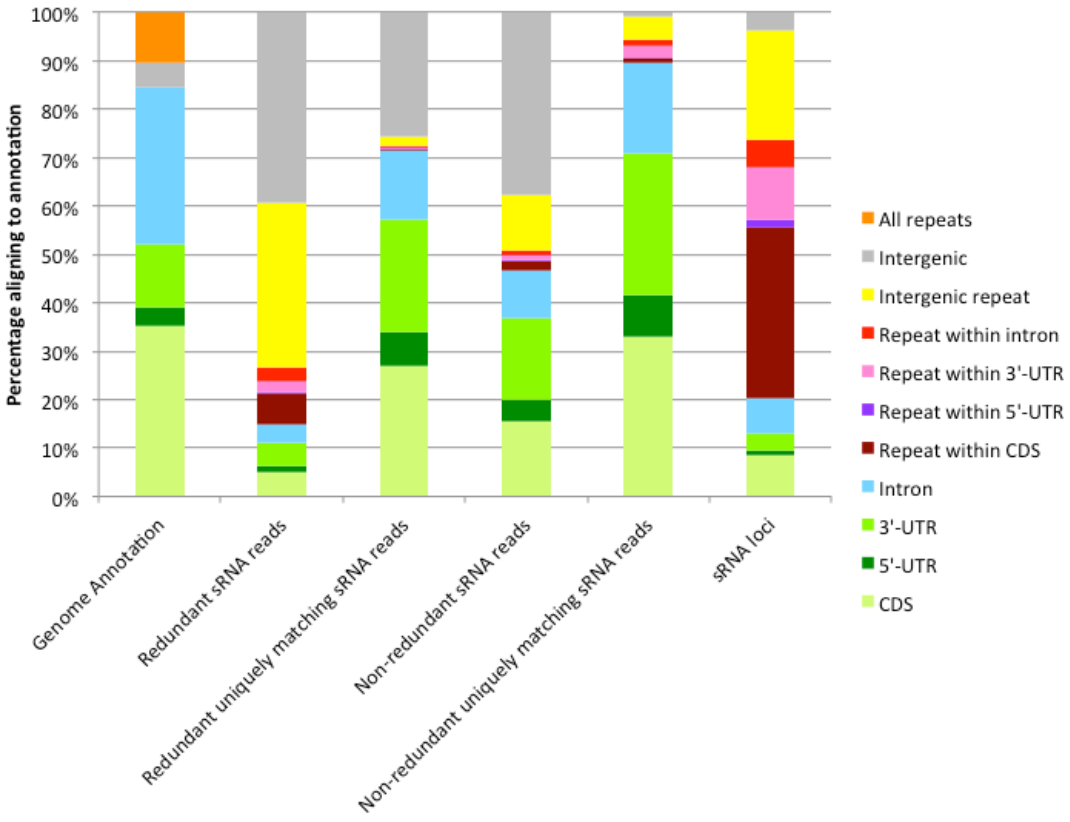
7.12. Northern analysis to verify expression of transgressively expressed sRNAs

Transgressive expression of sRNAs was verified using sRNA northerns. The predicted miRNA TE-F-2 (a longer species of TE-F-1) was shown to have transgressive expression in the recombinants L2-3 and L2-5. Membranes were

exposed to Phosphor Imager sheets for 2 days and the U6 probe used as a loading control.



7.13. Overlap analysis summary for CC125+ and J- sRNAs



8. Bibliography

Adams, C.R., Stamer, K. a, Miller, J.K., McNally, J.G., Kirk, M.M., and Kirk, D.L. (1990). Patterns of organellar and nuclear inheritance among progeny of two geographically isolated strains of Volvox carteri. Curr. Genet. *18*, 141–153.

Altschul, S.F., Gish, W., Miller, W., Meyers, E.W., and Lipman, D.J. (1990). Basic Local Alignment Search Tool. J. Mol. Biol. *215*, 403–410.

Andrews, S. (Babraham B. (2012). FastQC A Quality Control tool for High Throughput Sequence Data.

Axtell, M.J. (2008). Evolution of microRNAs and their targets: are all microRNAs biologically relevant? Biochim. Biophys. Acta *1779*, 725–734.

Axtell, M.J. (2013). Classification and comparison of small RNAs from plants. Annu. Rev. Plant Biol. *64*, 137–159.

Axtell, M.J., Snyder, J.A., and Bartel, D.P. (2007). Common functions for diverse small RNAs of land plants. Plant Cell *19*, 1750–1769.

Bao, H., Kommadath, A., Plastow, G.S., Tuggle, C.K., Guan, L.L., and Stothard, P. (2014). MicroRNA buffering and altered variance of gene expression in response to Salmonella infection. PLoS One *9*, e94352.

Barber, W.T., Zhang, W., Win, H., Varala, K.K., Dorweiler, J.E., Hudson, M.E., and Moose, S.P. (2012). Repeat associated small RNAs vary among parents and following hybridization in maize. Proc. Natl. Acad. Sci. U. S. A. *109*, 10444–10449.

Bartel, B., Bartel, D.P., Biology, C., and Cambridge, T. (2003). MicroRNAs: At the Root of Plant Development? 1. Current *132*, 709–717.

Beardall, J., and Raven, J.A. (2004). The potential effects of global climate change on microalgal photosynthesis, growth and ecology. Phycologia *43*, 26–40.

Bennett, A.B., and Ali, S.H. (2010). A Plant Breeder's History of the World. Science (80-.). *329*, 391–392.

Birchler, J.A., and Veitia, R.A. (2010). The gene balance hypothesis : implications for gene regulation , quantitative traits and evolution. 54–62.

Birchler, J., Auger, D., and Riddle, N. (2003). In Search of the Molecular Basis of Heterosis. Plant Cell *15*, 2236–2239.

Birchler, J. a., Yao, H., Chudalayandi, S., Vaiman, D., and Veitia, R. a. (2010). Heterosis. Plant Cell Online *22*, 2105–2112.

Blaby, I.K., Blaby-Haas, C.E., Tourasse, N., Hom, E.F.Y., Lopez, D., Aksoy, M., Grossman, A., Umen, J., Dutcher, S., Porter, M., et al. (2014). The Chlamydomonas genome project: a decade on. Trends Plant Sci. 1–9.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 3319–3320.

Bond, D.M., and Baulcombe, D.C. (2014). Small RNAs and heritable epigenetic variation in plants. Trends Cell Biol. *24*, 100–107.

Borges, F., and Martienssen, R. a (2013). Establishing epigenetic variation during genome reprogramming. RNA Biol. *10*, 490–494.

Brodersen, P., Sakvarelidze-Achard, L., Bruun-Rasmussen, M., Dunoyer, P., Yamamoto, Y.Y., Sieburth, L., and Voinnet, O. (2008). Widespread translational inhibition by plant miRNAs and siRNAs. Science *320*, 1185–1190.

Bucheton, a, Paro, R., Sang, H.M., Pelisson, a, and Finnegan, D.J. (1984). The molecular basis of I-R hybrid dysgenesis in Drosophila melanogaster: identification, cloning, and properties of the I factor. Cell *38*, 153–163.

Buisine, N., Quesneville, H., and Colot, V. (2008). Improved detection and annotation of transposable elements in sequenced genomes using multiple reference sequence sets. Genomics *91*, 467–475.

Burstin, J., and Charcosset, A. (1997). Relationship between phenotypic and marker distances : theoretical and experimental investigations. Heredity (Edinb). *79*, 477–483.

Casas-Mollano, J.A., Rohr, J., Kim, E.-J., Balassa, E., van Dijk, K., and Cerutti, H. (2008). Diversification of the core RNA interference machinery in Chlamydomonas reinhardtii and the role of DCL1 in transposon silencing. Genetics *179*, 69–81.

Castel, S.E., and Martienssen, R. a (2013). RNA interference in the nucleus: roles for small RNAs in transcription, epigenetics and beyond. Nat. Rev. Genet. *14*, 100–112.

Chapman, E.J., and Carrington, J.C. (2007). Specialization and evolution of endogenous small RNA pathways. Genome Res. *8*, 884–896.

Chen, Z.J. (2007). Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. Annu. Rev. Plant Biol. *58*, 377–406.

Chen, F., He, G., He, H., Chen, W., Zhu, X., Liang, M., Chen, L., and Deng, X.W. (2010a). Expression analysis of miRNAs and highly-expressed small RNAs in two rice subspecies and their reciprocal hybrids. J. Integr. Plant Biol. *52*, 971–980.

Chen, H.-M., Chen, L.-T., Patel, K., Li, Y.-H., Baulcombe, D.C., and Wu, S.-H. (2010b). 22-Nucleotide RNAs trigger secondary siRNA biogenesis in plants. Proc. Natl. Acad. Sci. U. S. A. *107*, 15269–15274.

Chlamydomonas Resource Centre (2010). http://chlamycollection.org/.

Cokus, S.J.S., Feng, S., Zhang, X., Chen, Z., Merriman, B., CD, Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M., et al. (2008). Shotgun bisulphite

sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature *452*, 215–219.

Collins, S., and de Meaux, J. (2009). Adaptation to different rates of environmental change in Chlamydomonas. Evolution *63*, 2952–2965.

Cuperus, J.T., Fahlgren, N., and Carrington, J.C. (2011). Evolution and functional diversification of MIRNA genes. Plant Cell *23*, 431–442.

Ding, D., Wang, Y., Han, M., Fu, Z., Li, W., Liu, Z., Hu, Y., and Tang, J. (2012). MicroRNA transcriptomic analysis of heterosis during maize seed germination. PLoS One *7*, e39578.

Dunoyer, P., Brosnan, C. a, Schott, G., Wang, Y., Jay, F., Alioua, A., Himber, C., and Voinnet, O. (2010). An endogenous, systemic RNAi pathway in plants. EMBO J. *29*, 1699–1712.

Eickbush, T.H., and Eickbush, D.G. (2007). Finely Orchestrated Movements : Evolution of the Ribosomal RNA Genes. *485*, 477–485.

Fahlgren, N., Jogdeo, S., Kasschau, K.D., Sullivan, C.M., Chapman, E.J., Laubinger, S., Smith, L.M., Dasenko, M., Givan, S. a, Weigel, D., et al. (2010). MicroRNA Gene Evolution in Arabidopsis lyrata and Arabidopsis thaliana. Plant Cell *22*, 1074–1089.

Fei, Q., Xia, R., and Meyers, B.C. (2013). Phased, secondary, small interfering RNAs in posttranscriptional regulatory networks. Plant Cell *25*, 2400–2415.

Feng, S., Cokus, S.J., Zhang, X., Chen, P.-Y., Bostick, M., Goll, M.G., Hetzel, J., Jain, J., Strauss, S.H., Halpern, M.E., et al. (2010). Conservation and divergence of methylation patterning in plants and animals. Proc. Natl. Acad. Sci. U. S. A. *107*, 8689–8694.

Fernández-Silva, I., Moreno, E., Eduardo, I., Arús, P., Alvarez, J.M., and Monforte, A.J. (2009). On the genetic control of heterosis for fruit shape in melon (Cucumis melo L.). J. Hered. *100*, 229–235.

Ferris, P., and Goodenough, U. (1997). Mating Type in Chlamydomonas Is Specified by mid, the Minus-Dominance Gene. Genetics *146*, 859–869.

Ferris, P.J., Jeffrey, P., and Goodenough, U.W. (1996). A Sex Recognition Glycoprotein Is Encoded by the plus Mating-Type Gene fusl of Chlamydomonas reinhardtii. *7*, 1235–1248.

Filipowicz, W. (2008). The expanding world of small RNAs. Nature *451*.

Goho, S., and Bell, G. (2000). Mild environmental stress elicits mutations affecting fitness in Chlamydomonas. Proc. Biol. Sci. *267*, 123–129.

Goodenough, U., Lin, H., and Lee, J.-H. (2007). Sex determination in Chlamydomonas. Semin. Cell Dev. Biol. *18*, 350–361.

Gowans, C.S. (1963). The Conspecificity of Chlamydomonas eugametos and Chlamydomonas moewusii: An Experimental Approach 1. Phycologia *3*, 37–44.

Grant-Downton, R.T., and Dickinson, H.G. (2004). Plants, pairing and phenotypes--two's company? Trends Genet. *20*, 188–195.

Greaves, I.K., Groszmann, M., Ying, H., Taylor, J.M., Peacock, W.J., and Dennis, E.S. (2012). Trans Chromosomal Methylation in Arabidopsis hybrids. Proc. Natl. Acad. Sci. U. S. A. *109*, 3570–3575.

Griffiths-Jones, S., Grocock, R.J., Dongen, S. Van, Bateman, A., and Enright, A.J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. Database *34*, 140–144.

Groszmann, M., Greaves, I.K., Albertyn, Z.I., Scofield, G.N., Peacock, W.J., and Dennis, E.S. (2011). Changes in 24-nt siRNA levels in Arabidopsis hybrids suggest

an epigenetic contribution to hybrid vigor. Proc. Natl. Acad. Sci. U. S. A. *108*, 2617–2622.

Groszmann, M., Greaves, I.K., Fujimoto, R., Peacock, W.J., and Dennis, E.S. (2013). The role of epigenetics in hybrid vigour. Trends Genet. *29*, 684–690.

Ha, M., Lu, J., Tian, L., Ramachandran, V., Kasschau, K.D., Chapman, E.J., Carrington, J.C., Chen, X., Wang, X.-J., and Chen, Z.J. (2009). Small RNAs serve as a genetic buffer against genomic shock in Arabidopsis interspecific hybrids and allopolyploids. Proc. Natl. Acad. Sci. U. S. A. *106*, 17835–17840.

Hardcastle, T.J., and Kelly, K. a (2010). baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics *11*, 422.

Hardcastle, T.J., Kelly, K. a, and Baulcombe, D.C. (2012). Identifying small interfering RNA loci from high-throughput sequencing data. Bioinformatics *28*, 457–463.

Harris, E. (2008). The Chlamydomonas Sourcebook: Introduction to Chlamydomonas and its Laboratory Use (Academic Press).

Harris, E.H. (2003). Chlamydomonas as a model organism. Annu. Rev. Plant Physiol. Plant Mol. Biol. *52*, 363.

Harris, E.H. (2009). The Chlamydomonas Sourcebook (Oxford: Elsevier).

He, G., Zhu, X., Elling, A. a, Chen, L., Wang, X., Guo, L., Liang, M., He, H., Zhang, H., Chen, F., et al. (2010). Global epigenetic and transcriptional trends among two rice subspecies and their reciprocal hybrids. Plant Cell *22*, 17–33.

He, G., He, H., and Deng, X.W. (2013a). Epigenetic variations in plant hybrids and their potential roles in heterosis. J. Genet. Genomics *40*, 205–210.

He, G., Chen, B., Wang, X., Li, X., Li, J., He, H., Yang, M., Lu, L., Qi, Y., Wang, X., et al. (2013b). Conservation and divergence of transcriptomic and epigenomic variation in maize hybrids. Genome Biol. *14*, R57.

Heard, E., and Martienssen, R.A. (2014). Transgenerational Epigenetic Inheritance: Myths and Mechanisms. Cell *157*, 95–109.

Henderson, I.R., and Jacobsen, S.E. (2007). Epigenetic inheritance in plants. Nature *447*.

Herr, A., and Baulcombe, D.D. (2004). RNA silencing in plants. Nature *431*, 356–363.

Herranz, H., and Cohen, S.M. (2010). MicroRNAs and gene regulatory networks: managing the impact of noise in biological systems. Genes Dev. *24*, 1339–1344.

Hochholdinger, F., and Hoecker, N. (2007). Towards the molecular basis of heterosis. Trends Plant Sci. *12*, 427–432.

Holeski, L.M., Jander, G., and Agrawal, A. a (2012). Transgenerational defense induction and epigenetic inheritance in plants. Trends Ecol. Evol. *27*, 618–626.

Hong, J.W., Jeong, J., Kim, S.H., Kim, S., and Yoon, H.-S. (2013). Isolation of a Korean domestic microalga, Chlamydomonas reinhardtii KNUA021, and analysis of its biotechnological potential. J. Microbiol. Biotechnol. *23*, 375–381.

Hutvagner, G., and Simard, M.J. (2008). Argonaute proteins: key players in RNA silencing. Nat. Rev. Mol. Cell Biol. *9*, 22–32.

Ichimura, T. (1996). Genome rearrangement and speciation in freshwater algae. Hydrobiologia *336*, 1–17.

Ichimura, T., and Kasai, F. (1996). Morphological and cytogenetic characterisitcs of intergroup hybrids between closely related mating groups of the Closterium

ehrenbergii species complex (Desmidiales, Chlorophyta). Phycol. Res. *44*, 261–265.

Jagadeeswaran, G., Nimmakayala, P., Zheng, Y., Gowdu, K., Reddy, U.K., and Sunkar, R. (2012). Characterization of the small RNA component of leaves and fruits from four different cucurbit species. BMC Genomics *13*, 329.

Jang, H., and Ehrenreich, I.M. (2012). Genome-wide characterization of genetic variation in the unicellular, green alga Chlamydomonas reinhardtii. PLoS One *7*, e41307.

Jiang, X., and Stern, D. (2009). Mating and tetrad separation of Chlamydomonas reinhardtii for genetic analysis. J. Vis. Exp. 2–3.

Jiao, Y., Wang, Y., Xue, D., Wang, J., Yan, M., Liu, G., Dong, G., Zeng, D., Lu, Z., Zhu, X., et al. (2010). Regulation of OsSPL14 by OsmiR156 defines ideal plant architecture in rice. Nat. Genet. *42*, 541–544.

Jin, H., Hu, W., Wei, Z., Wan, L., Li, G., Tan, G., Zhu, L., and He, G. (2008). Alterations in cytosine methylation and species-specific transcription induced by interspecific hybridization between Oryza sativa and O. officinalis. Theor. Appl. Genet. *117*, 1271–1279.

Jones-Rhoades, M.W., Bartel, D.P., and Bartel, B. (2006). MicroRNAs and their regulatory roles in plants. Annu. Rev. Plant Biol *57*, 19–53.

Jurka, J., Kapitonov, V. V, Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. *110*, 462–467.

Kathir, P., LaVoie, M., Brazelton, W.J., Haas, N.A., Lefebvre, P.A., and Silflow, C.D. (2003a). Molecular map of the Chlamydomonas reinhardtii nuclear genome. Eukaryot. Cell *2*, 362.

Kathir, P., LaVoie, M., Brazelton, W.J.W.J., Haas, N.A.N.A., Lefebvre, P.A.P.A., and Silflow, C.D.C.D. (2003b). Molecular map of the Chlamydomonas reinhardtii nuclear genome. Eukaryot. Cell *2*, 362.

Kenan-Eichler, M., Leshkowitz, D., Tal, L., Noor, E., Melamed-Bessudo, C., Feldman, M., and Levy, A. a (2011). Wheat hybridization and polyploidization results in deregulation of small RNAs. Genetics *188*, 263–272.

Kozomara, A., and Griffiths-Jones, S. (2014). miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. *42*, D68–D73.

Krieger, U., Lippman, Z.B., and Zamir, D. (2010). The flowering gene SINGLE FLOWER TRUSS drives heterosis for yield in tomato. Nat. Genet. *42*, 459–463.

Labadorf, A., Link, A., Rogers, M.F., Thomas, J., Reddy, A.S., and Ben-Hur, A. (2010). Genome-wide analysis of alternative splicing in Chlamydomonas reinhardtii. BMC Genomics *11*, 114.

Ladewig, E., Okamura, K., Flynt, A.S., Westholm, J.O., and Lai, E.C. (2012). Discovery of hundreds of mirtrons in mouse and human small RNA data. Genome Res. *22*, 1634–1645.

Lamesch, P., Berardini, T.Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D.L., Garcia-Hernandez, M., et al. (2012). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. *40*, D1202–D1210.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Li, C., Huang, L., Xu, C., Zhao, Y., and Zhou, D.-X. (2011). Altered levels of histone deacetylase OsHDT1 affect differential gene expression patterns in hybrid rice. PLoS One *6*, e21789.

Li, J., Wu, Y., and Qi, Y. (2014). microRNAs in a multicellular green alga Volvox carteri. Sci. China. Life Sci. *57*, 36–45.

Li, L., Lu, K., Chen, Z., Mu, T., Hu, Z., and Li, X. (2008). Dominance, overdominance and epistasis condition the heterosis in two heterotic rice hybrids. Genetics *180*, 1725–1742.

Li, X., Wei, Y., Nettleton, D., and Brummer, E.C. (2009). Comparative gene expression profiles between heterotic and non-heterotic hybrids of tetraploid Medicago sativa. BMC Plant Biol. *9*, 107.

Li, Y., Varala, K., Moose, S.P., and Hudson, M.E. (2012). The inheritance pattern of 24 nt siRNA clusters in arabidopsis hybrids is influenced by proximity to transposable elements. PLoS One *7*, e47043.

Lin, H., Miller, M.L., Granas, D.M., and Dutcher, S.K. (2013). Whole genome sequencing identifies a deletion in protein phosphatase 2A that affects its stability and localization in Chlamydomonas reinhardtii. PLoS Genet. *9*, e1003841.

Lohse, M., Bolger, A.M., Nagel, A., Fernie, A.R., Lunn, J.E., Stitt, M., and Usadel, B. (2012). RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. Nucleic Acids Res. *40*, W622–W627.

MacLean, D., Elina, N., Havecker, E.R., Heimstaedt, S.B., Studholme, D.J., and Baulcombe, D.C. (2010). Evidence for large complex networks of plant short silencing RNAs. PLoS One *5*, e9901.

Malone, C.D., and Hannon, G.J. (2009). Small RNAs as guardians of the genome. Cell *136*, 656–668.

Manavella, P. a, Koenig, D., and Weigel, D. (2012). Plant secondary siRNA production determined by microRNA-duplex structure. Proc. Natl. Acad. Sci. U. S. A. *109*, 2461–2466.

Matzke, M. a, and Mosher, R. a (2014). RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. Nat. Rev. Genet. *15*, 394–408.

Melnyk, C.W., Molnar, A., Bassett, A., and Baulcombe, D.C. (2011). Mobile 24 nt small RNAs direct transcriptional gene silencing in the root meristems of Arabidopsis thaliana. Curr. Biol. *21*, 1678–1683.

Meng, Y., and Shao, C. (2012). Large-scale identification of mirtrons in Arabidopsis and rice. PLoS One *7*, e31163.

Meng, Y., Gou, L., Chen, D., Mao, C., Jin, Y., Wu, P., and Chen, M. (2011a). PmiRKB: a plant microRNA knowledge base. Nucleic Acids Res. *39*, D181–D187.

Meng, Y., Shao, C., and Chen, M. (2011b). Toward microRNA-mediated gene regulatory networks in plants. Brief. Bioinform. *12*, 645–659.

Merchant, S.S., Prochnik, S.E., Vallon, O., Harris, E.H., Karpowicz, S.J., Witman, G.B., Terry, A., Salamov, A., Fritz-Laylin, L.K., Maréchal-Drouard, L., et al. (2007). The Chlamydomonas genome reveals the evolution of key animal and plant functions. Science *318*, 245–250.

Meyer, R.C., Witucka-Wall, H., Becher, M., Blacha, A., Boudichevskaia, A., Dörmann, P., Fiehn, O., Friedel, S., von Korff, M., Lisec, J., et al. (2012). Heterosis manifestation during early Arabidopsis seedling development is characterized by intermediate gene expression and enhanced metabolic activity in the hybrids. Plant J. *71*, 669–683.

Millar, A. a, and Waterhouse, P.M. (2005). Plant and animal microRNAs: similarities and differences. Funct. Integr. Genomics *5*, 129–135.

Mishra, A.K., Agarwal, S., Jain, C.K., and Rani, V. (2009). High GC content: critical parameter for predicting stress regulated miRNAs in Arabidopsis thaliana. Bioinformation *4*, 151–154.

Misumi, O., Yoshida, Y., Nishida, K., Fujiwara, T., Sakajiri, T., Hirooka, S., Nishimura, Y., and Kuroiwa, T. (2008). Genome analysis and its significance in four unicellular algae, Cyanidioschyzon [corrected] merolae, Ostreococcus tauri, Chlamydomonas reinhardtii, and Thalassiosira pseudonana. J. Plant Res. *121*, 3–17.

Molnar, A., Schwach, F., Studholme, D.J., Thuenemann, E.C., Baulcombe, D.C., and Molna, A. (2007a). miRNAs control gene expression in the single-cell alga Chlamydomonas reinhardtii. Nature *447*, 1126–1129.

Molnar, A., Schwach, F., Studholme, D.J., Thuenemann, E., and Baulcombe, D. (2007b). miRNAs control gene expression in single cell alga Chlamydomonas reinhardtii. Nature *447*, 1126–1129.

Molnar, A., Melnyk, C., and Baulcombe, D.C. (2011). Silencing signals in plants: a long journey for small RNAs. Genome Biol. *12*, 215.

Mor, E., and Shomron, N. (2013). Species-specific microRNA regulation influences phenotypic variability: perspectives on species-specific microRNA regulation. Bioessays *35*, 881–888.

Moxon, S., Schwach, F., Dalmay, T., Maclean, D., Studholme, D.J., and Moulton, V. (2008). A toolkit for analysing large-scale plant small RNA datasets. Bioinformatics *24*, 2252–2253.

Nakada, T., Shinkawa, H., Ito, T., and Tomita, M. (2010). Recharacterization of Chlamydomonas reinhardtii and its relatives with new isolates from Japan. J. Plant Res. *123*, 67–78.

Naqvi, A.R., Islam, M.N., Choudhury, N.R., and Haq, Q.M.R. (2009). The fascinating world of RNA interference. Int. J. Biol. Sci. *5*, 97–117.

Ness, R.W., Morgan, A.D., Colegrave, N., and Keightley, P.D. (2012). Estimate of the spontaneous mutation rate in Chlamydomonas reinhardtii. Genetics *192*, 1447–1454.

Ni, Z., Kim, E.-D., Ha, M., Lackey, E., Liu, J., Zhang, Y., Sun, Q., and Chen, Z.J. (2009). Altered circadian rhythms regulate growth vigour in hybrids and allopolyploids. Nature *457*, 327–331.

Obbard, D.J., Jiggins, F.M., Bradshaw, N.J., and Little, T.J. (2011). Recent and recurrent selective sweeps of the antiviral RNAi gene Argonaute-2 in three species of Drosophila. Mol. Biol. Evol. *28*, 1043–1056.

Pantaleo, V., Szittya, G., Moxon, S., Miozzi, L., Moulton, V., Dalmay, T., and Burgyan, J. (2010). Identification of grapevine microRNAs and their targets using high-throughput sequencing and degradome analysis. Plant J. no – no.

Pascual, L., Xu, J., Biais, B., Maucourt, M., Ballias, P., Bernillon, S., Deborde, C., Jacob, D., Desgroux, A., Faurobert, M., et al. (2013). Deciphering genetic diversity and inheritance of tomato fruit weight and composition through a systems biology approach. J. Exp. Bot. *64*, 5737–5752.

Pérez-alegre, M., Dubus, A., and Fernández, E. (2005). REM1 , a New Type of Long Terminal Repeat Retrotransposon in Chlamydomonas reinhardtii REM1 , a New Type of Long Terminal Repeat Retrotransposon in Chlamydomonas reinhardtii. *25*.

Pessia, E., Popa, A., Mousset, S., Rezvoy, C., Duret, L., and Marais, G. a B. (2012). Evidence for widespread GC-biased gene conversion in eukaryotes. Genome Biol. Evol. *4*, 675–682.

Piriyapongsa, J., and Jordan, I. (2008). Dual coding of siRNAs and miRNAs by plant transposable elements. Rna 814–821.

Prochnik, S.E.E.S.E., Umen, J., Nedelcu, A.M.A.M.M., Hallmann, A., Miller, S.M.M.S.M., Nishii, I., Ferris, P., Kuo, A., Mitros, T., Fritz-Laylin, L.K.K.L.K., et al. (2010). Genomic analysis of organismal complexity in the multicellular green alga Volvox carteri. Science *329*, 223–226.

Pröschold, T., Harris, E.H., and Coleman, A.W. (2005). Portrait of a species: Chlamydomonas reinhardtii. Genetics *170*, 1601–1610.

Quadrana, L., Almeida, J., Asís, R., Duffy, T., Dominguez, P.G., Bermúdez, L., Conti, G., Corrêa da Silva, J. V, Peralta, I.E., Colot, V., et al. (2014). Natural occurring epialleles determine vitamin E accumulation in tomato fruits. Nat. Commun. *5*, 3027.

Rakyan, V.K., Blewitt, M.E., Druker, R., Preis, J.I., and Whitelaw, E. (2002). Metastable epialleles in mammals. *18*, 348–351.

Rechavi, O. (2014). Guest list or black list : heritable small RNAs as immunogenic memories. Trends Cell Biol. *24*, 212–220.

Riddle, N.C., Jiang, H., An, L., Doerge, R.W., and Birchler, J.A. (2010). Gene expression analysis at the intersection of ploidy and hybridity in maize. Theor. Appl. Genet. *120*, 341–353.

Rieseberg, L.H., Archer, M.A., and Wayne, R.K. (1999). Transgressive segregation, adaptation and speciation. Heredity (Edinb). *83 (Pt 4)*, 363–372.

Ruiz-Ferrer, V., and Voinnet, O. (2009). Roles of plant small RNAs in biotic stress responses. Annu. Rev. Plant Biol. *60*, 485–510.

Sambrook, J., and Russell, D.W. (2001). Molecular Cloning: A Laboratory Manual, Volume 1 (CSHL Press).

Schultz, J., Maisel, S., Gerlach, D., Müller, T., and Wolf, M. (2005). A common core of secondary structure of the internal transcribed spacer 2 (ITS2) throughout the Eukaryota. *2*, 361–364.

Shabalina, S. a, and Koonin, E. V (2008). Origins and evolution of eukaryotic RNA interference. Trends Ecol. Evol. *23*, 578–587.

Shavalina, S., and Koonin, E. (2008). Origins and evolution of eukaryotic RNA interference. Library (Lond). *2008*, 578–587.

Shen, H., He, H., Li, J., Chen, W., Wang, X., Guo, L., Peng, Z., He, G., Zhong, S., Qi, Y., et al. (2012). Genome-wide analysis of DNA methylation and gene expression changes in two Arabidopsis ecotypes and their reciprocal hybrids. Plant Cell *24*, 875–892.

Shivaprasad, P. V, Dunn, R.M., Santos, B.A., Bassett, A., and Baulcombe, D.C. (2012). Extraordinary transgressive phenotypes of hybrid tomato are influenced by epigenetics and small silencing RNAs. EMBO J. *31*, 257–266.

Shu, L., and Hu, Z. (2012). Characterization and differential expression of microRNAs elicited by sulfur deprivation in Chlamydomonas reinhardtii. BMC Genomics *13*, 108.

Sijen, T., Steiner, F. a, Thijssen, K.L., and Plasterk, R.H. a (2007). Secondary siRNAs result from unprimed RNA synthesis and form a distinct class. Science *315*, 244–247.

Smit, A., Hubley, R., and Green, P. RepeatMasker Open-3.0.

Smith, D.R., and Lee, R.W. (2008). Nucleotide diversity in the mitochondrial and nuclear compartments of Chlamydomonas reinhardtii: investigating the origins of genome architecture. BMC Evol. Biol. *8*, 156.

Song, R., and Messing, J. (2003). Gene expression of a gene family in maize based on noncollinear haplotypes. Proc. Natl. Acad. Sci. U. S. A. *100*, 9055–9060.

Stelkens, R., and Seehausen, O. (2009). Genetic distance between species predicts novel trait expression in their hybrids. Evolution *63*, 884–897.

Stelkens, R.B., Schmid, C., Selz, O., and Seehausen, O. (2009). Phenotypic novelty in experimental hybrids is predicted by the genetic distance between species of cichlid fish. BMC Evol. Biol. *9*, 283.

Stupar, R.M., Gardiner, J.M., Oldre, A.G., Haun, W.J., Chandler, V.L., and Springer, N.M. (2008). Gene expression analyses in maize inbreds and hybrids with varying levels of heterosis. BMC Plant Biol. *8*, 33.

Sung, W., Ackerman, M.S., Miller, S.F., Doak, T.G., and Lynch, M. (2012). Drift-barrier hypothesis and mutation-rate evolution. Proc. Natl. Acad. Sci. U. S. A. *109*, 18488–18492.

Tarver, J.E., Donoghue, P.C.J., and Peterson, K.J. (2012). Do miRNAs have a deep evolutionary history? Bioessays *34*, 857–866.

The Tomato Genome Consortium (2012). The tomato genome sequence provides insights into fleshy fruit evolution. Nature *485*, 635–641.

Thompson, C. (2012). Genetic analysis of RNA silencing in the unicellular alga. University of Cambridge.

Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform. *14*, 178–192.

Vaughn, M.W., Tanurdzić, M., Lippman, Z., Jiang, H., Carrasquillo, R., Rabinowicz, P.D., Dedhia, N., McCombie, W.R., Agier, N., Bulski, A., et al. (2007). Epigenetic natural variation in Arabidopsis thaliana. PLoS Biol. *5*, e174.

Werner, R., and Mergenhagen, D. (1998). Mating Type Determination of Chlamydomonas reinhardtii by PCR. Biotechniques *2*, 295–299.

Xie, Z., Johansen, L.K., Gustafson, A.M., Kasschau, K.D., Lellis, A.D., Zilberman, D., Jacobsen, S.E., and Carrington, J.C. (2004). Genetic and functional diversification of small RNA pathways in plants. PLoS Biol. *2*, E104.

Xue, L.-J., Zhang, J.-J., and Xue, H.-W. (2009). Characterization and expression profiles of miRNAs in rice seeds. Nucleic Acids Res. *37*, 916–930.

Young, I., and Coleman, A.W. (2004). The advantages of the ITS2 region of the nuclear rDNA cistron for analysis of phylogenetic relationships of insects : a Drosophila example. *30*, 236–242.

Yu, B., Yang, Z., Li, J., Minakhina, S., Yang, M., Padgett, R.W., Steward, R., and Chen, X. (2005). Methylation as a crucial step in plant microRNA biogenesis. Science *307*, 932–935.

Yu, S.B., Li, J.X., Xu, C.G., Tan, Y.F., Gao, Y.J., Li, X.H., Zhang, Q., and Saghai Maroof, M. a (1997). Importance of epistasis as the genetic basis of heterosis in an elite rice hybrid. Proc. Natl. Acad. Sci. U. S. A. *94*, 9226–9231.

Zamora, I., Feldman, J.L., and Marshall, W.F. (2004). PCR-based assay for mating type and diploidy in Chlamydomonas. Biotechniques *37*, 534–536.

Zhang, H., Ma, Z.-Y., Zeng, L., Tanaka, K., Zhang, C.-J., Ma, J., Bai, G., Wang, P., Zhang, S.-W., Liu, Z.-W., et al. (2013). DTF1 is a core component of RNA-directed DNA methylation and may assist in the recruitment of Pol IV. Proc. Natl. Acad. Sci. U. S. A. *110*, 8290–8295.

Zhang, L., Peng, Y., Wei, X., Dai, Y., Yuan, D., Lu, Y., Pan, Y., and Zhu, Z. (2014). Small RNAs as important regulators for the hybrid vigour of super-hybrid rice. J. Exp. Bot.

Zhang, Y., Wiggins, B.E., Lawrence, C., Petrick, J., Ivashuta, S., and Heck, G. (2012). Analysis of plant-derived miRNAs in animal small RNA datasets. BMC Genomics *13*, 381.

Zhao, T., Li, G., Mi, S., Li, S., Hannon, G.J., Wang, X., and Qi, Y. (2007a). A complex system of small RNAs in the unicellular green alga Chlamydomonas reinhardtii. Genes Dev. *21*, 1190–1203.

Zhao, X., Chai, Y., and Liu, B. (2007b). Epigenetic inheritance and variation of DNA methylation level and pattern in maize intra-specific hybrids. Plant Sci. *172*, 930–938.

Zhou, Z., Wang, Z., Li, W., Fang, C., Shen, Y., Li, C., Wu, Y., and Tian, Z. (2013). Comprehensive analyses of microRNA gene evolution in paleopolyploid soybean genome. Plant J. *76*, 332–344.