Characterization of a complex nuclear restorer locus of *Brassica napus*

Lydiane Gaborieau

Department of Biology

McGill University, Montreal

March 2016

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

« Le principal fléau de l'humanité n'est pas l'ignorance mais le refus de savoir »

Simone de Beauvoir

« Just keep swimming »

Dory – Finding Nemo

TABLE OF CONTENTS

ABSTRACT

The plant trait of cytoplasmically-inherited male sterility (CMS) and its suppression by nuclear restorer-of-fertility (*Rf*) genes can be viewed as a genetic arms-race between the mitochondrial and nuclear genomes. The male sterilizing factors consist of unusual genes or open reading frames (ORFs) that usually contain a portion of functional mitochondrial genes derived sequences and sequences of unknown origin. These unusual ORFs are maternally inherited, transcribed and often effectively translated into novel proteins that are associated with the inability of the plant to produce functional pollen. Most nuclear *Rf* genes have been shown to encode P-type pentatricopeptide repeat proteins (PPRs). Phylogenetic analysis of P-class PPRs from sequenced plants genomes has shown that Rf-proteins cluster in a distinct clade of P-class PPRs, RFL-PPRs that display hallmarks of positive evolutionary selection.

In the canola, (oilseed rape) species *Brassica napus*, previous work has indicated the nuclear restorer genes for the two native forms of CMS, *Rfn* (for *nap* CMS) and *Rfp (pol* CMS) represent alternate haplotypes, or alleles, of a single nuclear locus. The capacity of the *Rfn* haplotype to mediate the processing of multiple transcripts is of interest as it is unique among other known restorer alleles. I explored the different processing events that each *Rfn* transcript target undergoes by mapping their 3' and 5' termini in plants containing and lacking *Rfn*. CR-RT-PCR was used to distinguish 5' and 3' ends produced by post-transcriptional processing events, in particular *Rfn* specific processing events, in the key transcripts of interest *orf222/nad5c/orf139*, *atp8* and *nad4*. This procedure consists of covalently ligating the 5' and 3' ends of total or mtRNAs using an RNA ligase. The reaction leads to the formation of circular RNA, which serves as template for RT-PCR. Sequence analysis between *Arabidopsis thaliana*, *B. napus* as well as between the different termini mapped, were used in order to explore probable conserved processing events and *Rfn* specific recognition sequences.

Fine genetic mapping indicates that *Rfn* localizes to the same genomic region as *Rfp* as shown in previous studies. We find this region is enriched in *RFL* genes, four of which, based on their position and expression, represent potential candidates for *Rfn*; one of these genes, designated *PPR4*, is a preferred candidate in that it is not expressed in the *nap* CMS line. Comparison of the corresponding regions of the genomes of *B. rapa*, *B. oleracea*, *Arabidopsis thaliana* and *A. lyrata* provides insight into the expansion of this group of *RFL* genes in different lines of evolutionary descent. Unlike other nuclear restorer loci containing multiple *RFL* genes, the *RFL* genes in the *Rf* region of *B. napus* are not present in tandem arrays but rather are dispersed in genomic location. The genes do not share similar flanking non-coding regions and do not contain introns, indicating that they have duplicated primarily through a retrotransposition-mediated process. In contrast, segmental duplication has been responsible for the distribution of the 10 sequences I annotated as *RFL* genes in the corresponding region of the *A. lyrata* genome. My observations define the Brassica *Rf* locus and indicate that different mechanisms may be responsible for the proliferation of *RFL* genes even among closely related genomes.

In the last chapter of this thesis, I describe the different efforts made towards the identification of the restorer of fertility of nap CMS and exploration of its biochemical proprieties in order to determine the mechanism of fertility restoration.

RÉSUMÉ

La stérilité mâle cytoplasmique (SMC) chez les plantes et sa suppression par des restorateurs nucléaires (Rfs) peuvent être visualisés comme un bras de fer génétique entre les génomes mitochondrial et nucléaire. Chez les espèces sauvages, la SMC est observée par le phénomène de la gynodioécie, où les individus hermaphrodites et femelles coexistent dans une même population. Les facteurs de stérilisation consistent en des gènes inhabituels ou des phases de lecture ouvertes (ORFs), qui peuvent contenir une portion de gènes mitochondriaux ainsi que des séquences d'origines inconnues. Ces ORFs inhabituelles sont héritées maternellement, transcrites et souvent traduites efficacement en nouvelles protéines associées avec l'incapacitée de la plante à produire du pollen fonctionnel. Il a été démontré que la plupart des gènes Rfs codent pour des protéines à répétition de pentatricopeptides (PPRs). Les études phylogénétiques de la classe P des PPRs à partir de génomes de plantes séquencés ont montré que les protéines Rfs se retrouvent dans un groupe distinct de PPR de classe P, les RFL-PPRs, qui affiche des marques de sélection évolutive positive.

Chez l'espèce *Brassica napus* (ou canola), des études antérieures ont indiqué que les gènes de restoration nucléaire pour les deux formes de SMC natives, Rfn (pour nap SMC) et Rfp (pol SMC), représentent deux haplotypes, ou allèles, d'un seul locus. La responsabilité du locus *Rfn* dans la formation de termini en 5' de multiples transcrits est unique parmi les capacités d'autres restaurateurs de fertilité connus. J'ai exploré les processus de formation de termini en 5' et en 3' d'une variété de transcrits d'intérêt en utilisant une technique de CR-RT-PCR afin de cartographier les différentes extrémités produites post-transcriptionellement, en particulier les évènements permettant la production de termini spécifiques à la présence de *Rfn* chez les transcrits *orf222/nad5c/orf139*, *atp8* et *nad4*. L'analyse des séquences des produits obtenus chez *B. napus* et la comparaison avec les séquences d'ARNs mitochondriaux *d'Arabidopsis thaliana* ont permis l'exploration d'évènements de modification post-transcriptionnelle

conservés et possiblement dépendants de la présence du restorateur *Rfn*. Nous avons pu mettre en évidence la formation de multiples termini en 5', contrairement à la formation de terminus unique en 3' pour les ARN mitochondriaux étudiés sans détection d'une séquence commune. Cependant, la présence d'une homologie de séquence autour des extrémités 5' produites par *Rfn*, associées ou non à la restauration de la fertilité, m'a permis d'explorer la possibilité de l'existence d'une séquence cible sur les ARNs étudiés.

La cartographie génétique fine présentée indique que *Rfn* est localisé dans la même région génomique que *Rfp*. Cette région est enrichie en gènes RFL, dont quatre, selon leurs positions et leurs expressions, représentent des candidats potentiels pour *Rfn*. Un de ces gènes, *PPR4*, est un candidat favori car il n'est pas exprimé dans la lignée nap SMC. La comparaison des régions correspondantes dans les génomes de *B. rapa, B. oleracea, A. thaliana* et *A. lyrata* a permis d'éclairer les mécanismes d'expansion du groupe de gènes RFL venant de lignées évolutives différentes. Contrairement aux autres loci nucléaires restaurateurs contenant de multiples gènes RFL, les gènes RFL dans la région Rf de *B. napus* ne sont pas présents en tandem mais ont plutôt des locations génomiques dispersées. Ces gènes ne partagent par leurs régions non-codantes encadrantes et ne contiennent pas d'introns, ce qui indiquent que la duplication a eu premièrement lieu par un processus de rétro-transposition, contrairement à la duplication segmentale responsable pour la distribution des 10 séquences annotées comme gènes RFL dans la région correspondante du génome de *A. lyrata*. Mes observations définissent le locus *Rfn* de Brassica et indiquent que différents mécanismes sont responsables pour la prolifération de gènes RFL parmi des génomes relativement proches.

Dans le dernier chapitre de cette thèse, je décris les différents travaux réalisés pour l'identification du restaurateur de fertilité de nap SMC et l'exploration des propriétés biochimiques de celui-ci, ce qui permettrait de supposer le mécanisme de restauration de fertilité.

ACKNOLEDGEMENTS

Although I found that PhD work is a solitary experience, it would not have been possible to complete it without the support of an amazing crowd. First and foremost, I would like to profusely thank Prof. Greg Brown who supported me through this adventure. His kindness throughout the process helped me more times than I could count to push through and achieve what I did not think possible.

I would like to thank Prof. Tamara Western for adopting me in her lab and making herself available in times of crisis. I would also like to thank Prof. Eric Shoubridge for being part of my committee and giving me helpful advice. I would like to thank the past and present members of the Brown and Western labs for their help. Whether it was scientific discussions or putting a good playlist on, I feel grateful to have been able to work in an amicable environment. I would like to give special thanks to Véronique Brulé and Dr Mohammed Sabar for proofreading most of this work. I would also like to thank Hakim Mireau and his lab members for welcoming me and guiding me during my time in Versailles.

I would like to thank my parents for providing material and emotional support throughout my academic ventures. Thank you Paul Gaborieau for showing me that hard work and determination would lead me to anything I would want to achieve. Thank you Nadine Saulnier for always believing in me even when I was not so sure I was doing the right thing. Flore, you were here and carefree and that was often the breeze of fresh air I needed. A special mention should be given to Kraft Dinners, Ginger Ale, Ice Cream, Netflix, Moustache and Agent Orange who provided emotionnal support.

J'aimerais remercier le reste de ma famille et mes amis, la famille Boldireff, Linda, Mylène, Jenny, Zoé, Emilie, Loulou, Phil, la team du club photo et tous ceux que j'oublie. Merci tout spécial à Anne pour les inspirations de citations féministes à 3 heures du matin et les corrections du résumé français de dernière minute.

Benjamin, merci pour ces 10 ans d'amitié, 2 doctorats, 4 pays et bien des litres de vodka. Je ne doute pas qu'on en ait bien plus encore.

À mon mari, Mathieu, ces 7 dernières années n'ont pas vraiment été un long fleuve tranquille. Ta tête froide et ton grand cœur m'ont permis de traverser les tempêtes. Je ne suis pas sure que tes connaissances accumulées sur la mitochondrie des plantes te seront si utiles dans la vie, mais je n'aurais certainement pas pu choisir meilleur partenaire pour les partager.

CONTRIBUTION OF THE AUTHORS

Sections of **Chapter I** will be submitted as a litterature review titled « The propensity of Pentatricopeptide Repeat genes to evolve into male fertility restorers ». This section have was written by me and edited by Dr. Gregory G. Brown and Dr. Hakim Mireau. The Brassica genome section is extracted from a book chapter I co-authored with Dr. Gregory G Brown : Brown, G. G., & Gaborieau, L. (2011). Positional Cloning in Brassica napus: Strategies for Circumventing Genome Complexity in a Polyploid Plant: INTECH Open Access Publisher. The remaining literature review work of this chapter has been written by me and edited by Dr. Gregory G Brown.

I have conducted all the work presented in **Chapter II** which should be converted into a manuscript combined with additional work from Dr Helen Elina. Editing of this chapter was done by Dr. Gregory G. Brown

Chapter III has been submitted to the scientific journal BMC Genomics on March 09th 2016 under the title « Comparative genomic analysis of the compound Brassica napus Rf locus » I share co-authorship with Dr. Gregory G Brown.

All the work presented in **Chapter IV** has been performed by myself and editing was performed by Dr. Gregory G. Brown.

LIST OF FIGURES

LIST OF TABLES

APPENDICES

LIST OF ABBREVIATIONS

°C	degree Celsius
A.	Arabidopsis
ATP	adenosine triphosphate
atp6/8	subunit 6/8
B.	Brassica
BAC	bacterial artificial chromosome
BC1	backcross first generation
bp	base pair
BRAD	Brassica database
CAPS	cleaved amplified polymorphic sequence
cDNA	complementary DNA
CMS	cytoplasmic male sterility
COX	cytochrome c oxidase
cpDNA	chloroplastic DNA
CR-RT-PCR	circularized RT-PCR
CRP	cAMP receptor protein
cv	cultivar
DNA	deoxiribonucleic acid
EBS1	exon binding site 1
EBS2	exon binding site 2
GRP	glycine rich protein
HL	Hong Lian
IBS1	intron binding site 1
IBS2	intron binding site 1
ILP	intron length polymorphism
kb	kilobase pair
Mb	megabase pair
mg	milligram
mL	millitre
mM	milimolar
mRNA	messenger RNA
MSH1	Mutator S protein homolog 1
mtDNA	mitochondrial DNA
MTSF	mitochondrial stability factor 1
nad	NADH dehydrogenase complex 1
ORF	open reading frame
PCD	programmed cell death
PCR	polymerase chain reaction

Poly-A	poly adenylated
PPR	pentatricopeptide repeat
Rf	restorer of fertility
RFL	restorer of fertility like
Rfn	restorer of fertility napus
Rfo	restorer of fertility ogura
Rfp	restorer of fertility polima
RNA	ribonucleic acid
ROS	reactive oxygen species
RPF	RNA processing factor
RT-PCR	reverse transcribed - PCR
SNP	single nucleotide polymorphism
tRNA	transfer RNA
ug	microgram
UTR	untranslated region
V	volts
WA	wild abortive

CHAPTER I : INTRODUCTION AND LITERATURE REVIEW

The Mitochondria

Plant cells contain two kinds of genomes with different modes of inheritance : the nuclear genome, whose genes are inherited in Mendelian fashion and the maternally inherited cytoplasmic genome that is located in both the mitochondria (mtDNA) and in the chloroplast (cpDNA).

Mitochondria, the organelles responsible for the cell aerobic energy production, are thought to be the descendants of an ancient α-proteobacterium that participated in a symbiotic relationship with another cell type, leading to the formation of the initial eukaryotic cell about 2 billion years ago (Lane 2005). Over the ensuing eons, the partner's genome evolved into that of the nucleus of modern eukaryotes. The majority of the genes of the α-proteobacterium were gradually transferred to the nucleus, but some, the number of which varies in different lines of eukaryotic descent, were remained in the mitochondria. A majority of these genes encode membrane embedded subunits of the respiratory chain (Lane 2005). Unlike the genes of the nucleus, which are transmitted through cell division and sexual reproduction by the mechanisms of mitosis and meiosis, mitochondrial genes are transmitted cytoplasmically through the females only (Lane 2005).

Plant Mitochondrial Genomes

Modern mitochondria are semi-autonomous organelles and two distinct genetic systems are required for their biogenesis (Levings and Brown 1989). Mitochondrial DNA (mtDNA) specifies a relatively small number of proteins whereas the vast majority of mitochondrial proteins are encoded by nuclear genes. Plant mitochondrial genomes encode the mitochondrial ribosomal RNAs (rRNAs), some but not all transfer RNAs (tRNAs), and approximately 40 proteins (Gualberto, Mileshina et al. 2014). Plant mitochondrial genes are essential for plant viability as many encode components of the complexes involved in energy production via oxidative phosphorylation. Mutations that block the expression of mitochondrial genes generally result in an embryonic lethality that can normally be rescued only through specific germination conditions or in vitro embryo rescue (reviewed in Hanson 1991, Kubo and Newton 2008, Hu, Huang et al. 2014).

Plant mitochondrial genomes are uniquely organized and significantly larger than their counterparts in other organisms. Initial restriction site mapping studies of *Brassica campestris* mtDNA, Palmer and Shields 1984, indicated that the approximately 220 kb mtDNA of this species was tripartite in structure. The entire genome can be represented as one large, master circle containing two copies of two kilobases (kb) direct repeats. Recurrent high frequency recombination across these repeats was proposed to generate two smaller, subgenomic, circles which then recombine with each another to regenerate the master circle (Palmer and Shields 1984). Subsequent mapping studies of more

complex plant mitochondrial genomes have indicated that they also generally map as circles containing long repeats that undergo recurrent recombination, although the number and orientation of these repeats, as well as the overall genome size, can vary significantly even within a single plant species (reviewed in Kubo and Newton 2008). Direct observation of plant mtDNAs, however, suggests that physical DNA circles in plant mitochondria are rare, and that the genome may exist predominantly as branched or linear, circularly permuted forms (reviewed in Bendich 1993).

Normally, recombination events in plant mitochondrial genomes are restricted to longer repeat sequences. But recombination can also occur across shorter, or imperfectly matched repeats, albeit at a much lower frequency (reviewed in Gualberto, Mileshina et al. 2014). Such events have been shown to generate mitochondrial genomes in which specific genes are deleted, as in the example of the non-chromosomal stripe (NCS) mutations of maize (Newton, Knudsen et al. 1990).

One consequence of infrequent recombination at shorter repeat sites is the existence of substoichiometric gene arrangements designated sublimons by Small, Isaac et al. 1987. DNA hybridization experiments showed that in maize lines of one cytoplasmic type, restriction fragments characteristic of other types of mtDNA could be detected at much lower, substoichiometric levels (Small, Isaac et al. 1987). Recombination across shorter repeats was independently proposed to be the major driving factor in the evolution of plant mitochondrial genomes (Palmer and Herbon 1988).

Plant Mitochondrial Gene Expression

Similarly to its bacterial counterpart, mitochondrial transcription in plants often results in polycistronic messengers (Unseld, Marienfeld et al. 1997). These transcripts include non-coding regions, including non-functionnal open reading frames (ORFs), which are not conserved across species, sometimes even across ecotypes, and thus generally thought to have no functional significance (reviewed in Holec, Lange et al. 2008).

In contrast to animal mitochondria, which lack introns, plant organelle genomes contain both group I and group II introns. These two groups of introns, are distinguished by their secondary structures and splicing mecanisms. Group II introns are prevalent in plant mitochondrial genomes whereas group I introns are rather rare (reviewed in Bonen and Vogel 2001). Some group I and II introns are mobile elements and are capable of retrotransposition. Such introns often encode an enzyme that promotes splicing (maturases) and the integration of the intron into a new genomic site. However, plant organellar introns have largely lost their ability for self-excision and movement, and few encode proteins (reviewed in Brown, des Francs-Small et al. 2014). In plant mitochondria, most introns are located in protein-coding genes. The majority of the intron-containing genes are expressed as RNA precursors containing contiguous exonic and intronic sequences and are spliced through a conventional cis-splicing pathway. Some plant mitochondrial intron-containing genes, however, are fragmented. In such cases, splicing occurs between two independently transcribed RNA precursors and is termed trans-splicing (reviewed in Brown, des Francs-Small et al. 2014). Typically,

group II intron splicing and trans-splicing, like splicing in eukaryotic nuclei, occurs through two successive trans-esterification reactions, the first involving cleavage at the 5' splice site with formation of a lariat intron, and the second the formation of the splice junction and release of the intron (Bonen and Vogel 2001). For splicing to occur properly, the intron must fold into a well-conserved structure that juxtaposes the 5' splice and branch sites and brings the two exons in close proximity. That well-conserved structure has a distinctive 3D architecture (domains 1–6), with its catalytic center formed by domain 1 (dI) and 5 (dV), and a large protein machinery is required for splicing in vivo.

A large-scale study of *Arabidopsis thaliana* promoter sequences indicated that individual genes and transcription units were often expessed via multiple promoters. (Kühn, Weihe et al. 2005). The different promoters can vary in strength and each presents a unique 5' end. Their locations are poorly conserved among plant species and contribute to the frequent transcript length polymorphisms that can be observed among different genotypes of a single species (Stoll, Stoll et al. 2013).

Moreover, contributing to the variety of 5' and 3' ends for one transcription units, RNA precursors usually undergo nuclease processing at both the 5' and 3' ends. It was proposed that RNA end maturation might be achieved through direct endoribonuclease and/or exoribonucleases activities that would be blocked by stable RNA secondary structures defining mature transcript ends. In favor of this hypothesis, mapping of mitochondrial mRNA termini in *A. thaliana* by Forner, Weber et al. 2007, highlighted the

presence of RNA stem-loop folds thought to be involved in RNA processing in plant mitochondria. Also thought to be involved in these processes, RNA-binding proteins may serve a similar function in blocking RNA degradation and defining transcripts ends. PPR proteins, termed RPFs (RNA processing factors), have been found to bind to their target transcripts and promote the formation of 5' and 3' ends in *A. thaliana* mitochondria (Jonietz, Forner et al. 2010, Hölzle, Jonietz et al. 2011, Jonietz, Forner et al. 2011, Hauler, Jonietz et al. 2013, Arnal, Quadrado et al. 2014). The use of protein barriers, especially PPR proteins, for termini definition may be widespread in plant mitochondria.

The RNA degradation process in plant is thought to rely on an RNA polyadenylation tagging similar to that of bacteria (Holec, Lange et al. 2008). In potato, Gagliardi, Perrin et al. 2001, showed that a 3' to 5' exo-ribonuclease activity was responsible for the preferential degradation of poly-adenylated (polyA) *atp9* mRNAs. This polyA tail would trigger the degradation of the RNA. Because the vast majority of identified polyA tails are located directly at transcript 3' ends or in their immediate vicinity (reviewed in Hammani and Giege 2014), this suggests that most polyadenylated transcripts are degraded by an exo-ribonuclease activity.

Cytoplasmic Male Sterility

Cytoplasmic male sterility (CMS) has been characterized in over 140 natural species (Laser and Lersten 1972). In wild species, CMS can be observed in the phenomenon of gynodioecy, where hermaphrodite and female (male sterile) individuals coexists within

one population (Touzet and Budar 2004). The male sterilizing factors consist of unusual genes or open reading frames (ORFs) that usually contain a portion of functional mitochondrial genes derived sequences and sequences of unknown origin. These unusual ORFs are maternally inherited, transcribed and often effectively translated into novel proteins that are associated with the inability of the plant to produce functional pollen, which represent the sole observed phenotype (Chen and Liu 2014).

CMS can be suppressed by specific nuclear genes called restorers of fertility (*Rf*). In the majority of cases, *Rf* genes produce proteins with the ability of acting directly on the CMS conferring transcripts in the mitochondria by binding them specifically and promoting processing events (reviewed in Chen and Liu 2014). In recent years, a majority of the proteins encoded by *Rf* genes have been found to belong to the PPR family (Dahan and Mireau 2013).

CMS at the mitochondrial genome level

CMS conferring mitochondrial genomes have been found to have multiple organizational differences from their non-sterility conferring (male-fertile) counterparts (reviewed in Hanson and Bentolila 2004). How, then, can a new CMS mitochondrial "mutation" arise in a population? It seems likely that in many cases, CMS associated gene arrangements may be present as sublimons in plants with male fertile mtDNAs. Yesodi, Izhar et al. 1995, showed that a sequence with perfect similarity to the unique portion of the CMS associated gene of petunia, *pcf*, was present at substoichiometric levels in fertile petunia

cytoplasm. That study suggested recombination between this and a partially homologous sequence in the major fertile mtDNA form may have given rise to the CMS-associated gene arrangement, S-pcf (reviewed in Hanson and Bentolila 2004). Importantly, Janska, Sarria et al. 1998, further showed that restriction fragments characteristic of a CMS-associated mitochondrial genome in the bean *Phaseolus vulgaris* could be detected at substoichiometric levels in mtDNA preparations of progenitor strains and that CMS associated fragments were present at low levels in male fertile revertant forms. Thus, in this case, CMS and its reversion to male fertility both appeared to result from a stoichiometric shift among different mtDNA forms, though it remained unclear precisely how such sub-stoichiometric shifting occurred (Janska, Sarria et al. 1998).

More recent investigations have shown that the maintenance of plant mitochondrial genome structure requires a surveillance mechanism involving a number of proteins that suppress recombination across short to intermediate sized repeats (reviewed in Gualberto, Mileshina et al. 2014). One such protein, MSH1, a mtDNA repair protein (mismatch repair protein), suppresses asymmetric recombination at intermediate sized repeats (Abdelnoor, Yule et al. 2003, Shedge, Arrieta-Montiel et al. 2007). Reduction in the expression of MSH1 can result in increased substoichiometric shifting and the induction of CMS (Sandhu, Abdelnoor et al. 2007). It remains unclear, however, how a new mtDNA form that escapes the surveillance mechanism can increase in copy number to become the predominant DNA arrangement within one generation. It is possible that stochastic segregation combined with mitochondrial fusion-fission can lead to an enrichment of a CMS conferring sublimon during cell division. That CMS conferring

sublimon could then become the predominant mtDNA form in plants progeny. This, combined with the female advantage (as explained in a latter section), would result in the spreading of the CMS trait across a population.

CMS associated mitochondrial genes

A number of different approaches have been used to identify the loci within mitochondrial genomes that specify male sterility (reviewed in Hanson and Bentolila 2004). A common feature of these loci is the presence of novel ORFs that often contain segments of one or more standard mitochondrial genes fused, in frame, with sequences of cryptic origin unrelated to functional mitochondrial genes. Other CMS-associated genes, such as radish *orf138*, consist entirely of such cryptic ORFs. Transcripts of CMS-associated genes are stable in CMS plants and are translated into small proteins (6-15 kDa) that are associated with the mitochondrial inner membrane (reviewed in Chen and Liu 2014). The prototypical CMS-associated gene, T-*urf13* of CMS-T maize, is largely composed of sequences normally found downstream of the large mitochondrial ribosomal RNA gene and is co-transcribed with *atp4* (formerly *orf25),* a gene that encodes a subunit of the F_0-F_1 ATP synthase (Dewey, Timothy et al. 1987). The protein product of T-*urf13*, URF13, is an integral inner mitochondrial membrane (Dewey, Timothy et al. 1987) and its expression in *Escherichia coli* can render cells susceptible to the toxin (T-toxin) secreted by *Bipolaris maydis*, the causative agent of southern corn leaf blight (Dewey, Siedow et al. 1988). The observation that T-toxin specifically uncouples oxidative phosphorylation of CMS-T mitochondria (Klein and Koeppe 1985) or in *E. coli*

expressing T-URF13 protein (Dewey, Siedow et al. 1988) contributed to the now widely held premise that expression of CMS-associated genes leads to a mild mitochondrial dysfunction, which is manifested phenotypically only during male gametogenesis.

It is not clear however, how such dysfunction specifically affects only the pollen formation phase of plant development whereas the rest of the plant remains intact. By comparison to the other plant cells and organs, the tapetum, the innermost layer of the anther wall and a nurse tissue for pollen grain development, represents the sector of the highest mitochondrial density (Lee and Warmke 1979). This assumes that male gametogenesis is highly energetically demanding to the point where a minor deficiency in mitochondrial function could lead to cell lethality and pollen abortion. By analogy to this assumption, the clinical symptoms in human diseases, which are linked to mitochondrial dysfunction such as myopathies, appear more prematurely in the high energy demanding tissues such as muscles and nerves (reviewed in Pinto and Moraes 2014).

Therefore, the mitochondrial dysfunctions would perturb significantly some biological aspects of the tapetum cells preventing any progress of further gametes development. Such premature tapetal degeneration leads to pollen abortion and is characteristic of a number of CMS systems including maize CMS-T (Lee and Warmke 1979), PET-1 CMS in sunflower (Smart, Monéger et al. 1994, Balk and Leaver 2001), Ogura CMS of radish (González-Melendi, Uyttewaal et al. 2008), and WA-CMS in rice (Luo, Xu et al. 2013). In PET-1 CMS, arrest in proper microspore development coincides with the appearance of cytosolic cytochrome c in tapetal cells and cleavage of DNA into nucleosome-sized

fragments, both characteristics of programmed cell death (PCD) in animal cells (Balk and Leaver 2001). The rice WA-CMS-associated protein, WA-352, is able to interact with COX11, a nuclear-encoded protein necessary for cytochrome c oxidase (complex IV) assembly. Cytochrome c oxidase also suppresses formation of hydrogen peroxide, a reactive oxygen species (ROS), and the appearance of ROS in mitochondria can serve as a trigger for initiating PCD (Luo, Xu et al. 2013). WA-CMS plants express WA-352 specifically in anther tissue, and its expression correlates with hydrogen peroxide production in tapetal cells and premature tapetal PCD (Luo, Xu et al. 2013). Thus, in both PET-1 CMS and WA-CMS, alterations in the timing of PCD in the tapetum disrupt pollen development, and in the case of WA-CMS, there is a clear mechanistic link of PCD with expression of the CMS-associated protein.

Attempts have been made to elucidate the molecular basis of the CMS phenotype at the protein level. For CMS-associated proteins that contain segments of subunits of oxidative phosphorylation complexes, such as sunflower ORF522 and brassica ORF224 and ORF222, it is possible that these chimeric proteins might compete with the normal subunit during complex assembly, leading to reduced levels of the functional complex. Consistent with this notion ORF522 contains a segment of ATP8, one of the subunits of the F_0-F_1 ATP synthase (complex V), and activity levels of this complex, as assessed by in-gel assays, are reduced in comparison to that of male fertile lines (Sabar, Gagliardi et al. 2003). Possibly, the activity of specific complexes may be sufficiently compromised in such cases as to prevent cellular energy production sufficient to meet the requirements for proper male reproductive tissue development. Such a mechanism could be considered

for other CMS conferring proteins presenting homology with the ATP8 subunit such as *Brassica napus* ORF222 and ORF224. In the case of the rice CMS type HL, reduced ATP and NADH levels in anthers are found in CMS plants expressing the sterility conferring protein ORFH79 (Wan, Li et al. 2007). ORFH79 interacts with a subunit of respiratory complex III (P61), which results in a defect in the mitochondrial electron transport chain and an overall decrease in ATP production (Wang, Gao et al. 2013).

Gynodioecy and genome conflict

In botany, dioecy refers to plant populations where female and male organs are carried on separated sporophytes; whereas gynodioecy consists of the coexistence in a plant population of hermaphrodites bearing both female and male organs on the same plant and individuals bearing only functional female organs (male sterile) (Gouyon and Couvet 1987). Over the last few decades, studies on gynodioecious natural populations has led to the emergence of theoretical models based on the conflict between cytoplasmic (mtDNA) and nuclear genomes in the production of male gametes, commonly described as the genetic arms race (Budar, Touzet et al. 2003).

In wild species, cytoplasmic male sterility (CMS) phenotypes, a condition that has been characterized in over 140 natural species where the ability of producing functional pollen is altered without affecting the normal growth of vegetative tissues (Laser and Lersten 1972) can sometimes be observed in the phenomenon of gynodioecy. The male sterilizing factors in the CMS phenotypes are encoded by a maternally inherited (i.e., inherited

primarily/exclusively through the female parent) cytoplasmic genome (mtDNA). As the CMS factor spreads in the population, the nuclear genome will react to the resulting paucity of pollen by re-establishing the male function through specific nuclear restorer genes called restorer of fertility (Rf) (Laser and Lersten 1972).

The female advantage

For such a system to persist in wild populations, females should benefit from a better fitness than hermaphrodites. The male sterility mutation can then increase in frequency in the population. Such a difference has been called female advantage (Budar, Touzet et al. 2003). In some CMS systems, female individuals produce a significantly higher number of seeds (reviewed by Gouyon and Couvet 1987). In such cases, the female fertility advantage is obvious and the male sterility inducing cytoplasm will be advantageous and more successfully transmitted than the "normal" cytoplasm with pollen as the only limiting factor.

However, the overproduction of seeds by female individuals in gynodioecious populations, is not always observed (reviewed in Alonso and Herrera 2001). In such cases, female advantage could still be obtained via maternal sex effects: the reallocation of the unused resources of the male function could lead to an increase in female fecundity or enhanced viability of the male sterile individuals. Dufay and Billard 2012, investigated data from studies on 48 gynodioecious species for various reproductive traits in order to determine the statistical existence of such maternal sex effects. After examination, 40

species presented varying degrees of female advantage (Dufay and Billard 2012). The increased fertility of the females may be due to the selective advantage of being outbred, and/or to the reproductive economy resulting from the lack of pollen production, but as long as these advantages are maintained, the females will increase in the population until the available pollen produced by hermaphrodites is fully utilized (reviewed by Budar, Touzet et al. 2003).

Advantages of CMS for agriculture

Female plants or CMS phenotypes in conjunction with nuclear restorer genes are used in the production of high yielding seeds and male fertile F1 hybrid crops. These can produce 15 to 50% higher yields than inbred lines and CMS is extensively used in agronomy as a reliable mean of increasing crops yields (Budar and Pelletier 2001). Hybrid crop production relies on the absence of self-pollination, which led the breeders to practice the laborious and expensive hand emasculation of the seed parent of the hybrid cross. The discoveries of CMS and fertility restoration traits in crop species obviated the need for this. Since CMS plants are incapable of self-pollination, when a CMS line is planted with a male fertile line, all the seeds that form on the CMS plants will be a hybrid of the two. When the male fertile line possesses a nuclear restorer gene, these seeds grow to form male fertile hybrids. The restoration of male fertility is essential when the seed or fruit is the harvested product (Budar and Pelletier 2001).

Nuclear restorer genes and proteins

CMS can be suppressed by specific nuclear genes called restorers of fertility (*Rf*). In the majority of cases, *Rf* genes produce proteins with the ability of acting directly on the CMS conferring transcripts in the mitochondria by binding them specifically and promoting processing events (reviewed in Chen and Liu 2014). In recent years, a majority of the proteins encoded by *Rf* genes have been found to belong to the PPR family (Dahan and Mireau 2013). This protein family is largely expanded in land plant genomes. PPR proteins have in common a canonical P-type 35 amino acid domain repeated in tandem up to 30 times. Length variations of that original P-type PPR domain allow the creation of longer (L-type) or shorter (S-type) domains. The PPR protein family is consequently divided in subfamilies depending on the number and type of repeats present in the structure as well as optional C-terminal domains. PPR proteins function in multiple aspects of organelle RNA metabolism, such as RNA splicing, editing, degradation and translation (Lurin, Andrés et al. 2004).

Recent three-dimensional structural analyses of PPR domains revealed that each PPR repeat is configured as two anti-parallel helices (Ban, Ke et al. 2013, Gully, Cowieson et al. 2015). Because these domains are repeated several times within a protein, the succession of such domains gives a general rectangular form to the protein with one side highly positively charged, which might suggest an involvement in RNA binding (Ban, Ke et al. 2013). Gel mobility shift assays showed that the P-type PPR domain has an affinity for single-stranded RNA compared to single and double-stranded DNA molecules

(Williams-Carrier, Kroeger et al. 2008) supporting the RNA binding capacity of PPR proteins.

Within one PPR domain some amino acids are of greater importance than others for RNA recognition (Ban, Ke et al. 2013). Several studies have demonstrated the existence of a recognition code between the identity of specific amino acids within the repeats and the target sequence of the PPR protein studied (Barkan, Rojas et al. 2012, Yagi, Hayashi et al. 2013, Yagi, Nakamura et al. 2014); the identity of the 6th amino acid of a motif in combination with the first amino acid of the next motif have been shown to be particularly important. These two amino acids are generally positively charged (Ban, Ke et al. 2013), indicating an ability for binding negatively charged nucleotides. The binding between these amino acids and the target nucleotide has been experimentally proven in the case of the maize chloroplastic proteins PPR10 (Yin, Li et al. 2013) and THA8 (Ke, Chen et al. 2013). In the context of CMS, where Rf proteins process unusual transcripts, nucleic acid specificity is essential to specifically target the CMS conferring transcript.

Recent studies revealed that the mechanisms by which PPR proteins recognize their target RNAs are highly specific but also revealed a certain level of flexibility. This flexibility allows PPR proteins to bind multiple RNAs (Yin, Li et al. 2013). Additionally, some *PPR* genes are governed by evolutionary constraints that facilitate their diversification and duplication in a relatively short time scale (Geddy and Brown 2007).

Restorer proteins belonging to the PPR family

The first identified *Rf* gene acting on a CMS transcript was *Rf-PPR592* from Petunia (Bentolila, Alfonso et al. 2002). Petunia CMS is caused by the expression of the *pcf* (petunia CMS fused) mitochondrial ORF. The *pcf* gene is composed of portions of two standard mitochondrial genes, *atp9* and *cox2*, as well as a sequence of unknown origin (Young and Hanson 1987). The petunia restorer locus contains two genes that encode for almost identical PPR proteins (PPR591 and PPR592) out of which only PPR592 carries restoration activity (Bentolila, Alfonso et al. 2002). When polymorphisms between these two *PPR* genes, present in both the CMS and the restored genomic background, were explored, no changes in the PPR591 sequence could be found but a deletion in the promoter of PPR592 in the CMS genomic background prevented its expression in floral buds (Bentolila, Alfonso et al. 2002). PPR592 was shown to rescue fertility by altering the *pcf* transcript profile and dramatically reducing the quantity of PCF protein present in the mitochondria (Bentolila, Alfonso et al. 2002). Immuno-precipitation experiments of mitochondrial fractions (Gillman, Bentolila et al. 2007) demonstrated that PPR592 is associated with the inner membrane of the mitochondria in a large protein complex binding *pcf* RNA, indicating the implication of a number of partner proteins to PPR592 in the restoration process.

Following the characterization of PPR592, a number of additional restorers of fertility genes were found to encode PPR proteins (table 1.1). The radish (*Raphanus sativus)* restorer gene for Ogura CMS (*Rfo*) has been cloned through a map-based cloning

approach that relied, in part, on the synteny that exists between the sequenced *A. thaliana* and the radish genome (Brown, Formanová et al. 2003, Desloire, Gherbi et al. 2003, Koizuka, Imai et al. 2003). In radish, ogura CMS is associated with the co-transcription of two open reading frames *orf138* and *orfB (atp8)*. *orf138* is the sterility-inducing gene and *orfB* encodes subunit eight of the F_0-F_1 ATP-synthase complex (Bonhomme, Budar et al. 1992). Within the restoration locus *Rfo*, three predicted genes *PPR-A*, *PPR-B* and *PPR-C* (or also called *g24*, *g26* and *g27* in Brown, Formanová et al. 2003) encode proteins belonging to the PPR family (Desloire, Gherbi et al. 2003). *PPR-C* was later found to be a pseudogene and only *PPR-B*, now confirmed as the *Rfo* gene, restored CMS by down-regulating the expression of ORF138 (Brown, Formanová et al. 2003). Not only were expression levels of *Rfo* found to be higher than *PPR-A* (Uyttewaal, Arnal et al. 2008, Qin, Warguchuk et al. 2014), but *Rfo* was also found to specifically affect the expression of *orf138* in the tapetum of anthers suggesting a tissue specific action (Uyttewaal, Arnal et al. 2008). Further insight into the restoration mechanism was gained by co-precipitation experiments with *orf138* RNA (Uyttewaal, Arnal et al. 2008). According to data from that study, Rfo binds specifically to the *orf138* transcript but does not promote the processing or the degradation of the CMS conferring RNA. The current model about the function of *Rfo* suggests that the restoration occurs by the blockage of the translation of the CMS conferring transcript, *orf138* (Uyttewaal, Arnal et al. 2008).

BT-rice CMS provides another example in which PPR proteins are involved in the fertility restoration mechanism. The sterility in BT-rice is associated with the translation of a large transcript composed of *atp6* sequences co-transcribed with a downstream novel

ORF, *orf79*, that consists of sequences derived from *cox1* and a sequence of unknown origin (Akagi, Sakamoto et al. 1994). Positional cloning of the fertility restoration locus revealed that it contained nine *PPR* genes (Akagi, Nakamura et al. 2004). Two of those *PPR* genes, *Rf1A* and *Rf1B* appear to be recently duplicated open reading frames and both show fertility restoration capacity for the BT-CMS (Wang, Zou et al. 2006) but employ different mechanisms. Wang, Zou et al. 2006, showed, by RNA gel blot experiments, that Rf1A induces a reduction in *orf79* transcript levels, and circular RT-PCR (cRT-PCR) experiments demonstrated that Rf1A governed the appearance of smaller transcripts with 5' ends produced by RNA cleavage events. Rf1A is therefore thought to act by the mediation of specific endonucleolytic cleavage within *orf79*. On the other hand, in the absence of Rf1A, Rf1B decreases *orf79* mRNA levels dramatically without generating additional, smaller transcripts (Wang, Zou et al. 2006). It was proposed therefore that Rf1B acts in restoration of fertility via a different mechanism than Rf1A by inducing the destabilization of *orf79* di-cistronic mRNA (Wang, Zou et al. 2006). When both restorers are present, the *atp6/orf79* di-cistronic mRNA is preferentially targeted by Rf1A (Kazama and Toriyama 2003, Komori, Ohta et al. 2004). The inability of Rf1B to destabilize the RNA fragments cleaved by Rf1A suggests that this cleavage also eliminates a recognition sequence in the inter-cistronic region necessary for Rf1B-dependent RNA degradation (Wang, Zou et al. 2006).

Although the exact identity or mechanisms remain unexplored, a number of PPR proteins in other plant CMS systems are thought to act as fertility restorers. In rice, Tang, Luo et al. 2014, reports the characterization of Rf4, another PPR protein acting as a restorer by

reducing the levels of the wild-abortive CMS conferring transcript WA352, an unusual transcript containing *nad5* and a chimeric ORF. Barr and Fishman 2010, have mapped a restorer locus in *Mimulus guttatus* that contains a large cluster of 17 PPR protein genes, suggesting that one of these genes could function in fertility restoration. In sorghum CMS systems A1 and A2, the restorer of fertility locus Rf5 also contains a *PPR* gene, which presents high homology with the rice *Rf1* (Jordan, Mace et al. 2010). It was additionally found that the Rf5 locus contains a cluster of *PPR* genes also presenting high homology to *Rf1* in rice (Jordan, Klein et al. 2011).

In the Hong-Lian rice CMS lines, the restoration of fertility also requires a PPR protein. *Rf5*, the restorer of fertility in HL-CMS identified by map based cloning, is identical to *Rf1A* in BT-rice CMS but does not restore fertility using the same mechanisms (Hu, Wang et al. 2012). Indeed, Rf5 was not found to be able to bind directly to the CMS conferring transcript, *atp6-orfH79* (Hu, Wang et al. 2012) but rather to work in a complex with a glycine rich protein, a mechanism that will be explored in a later section.

PPR-Rfs evolve from a *PPR* gene subgroup showing diversifying selection.

As mentioned earlier, characterization of the restorer locus in BT-rice CMS revealed that it contains nine *PPR* genes (Akagi, Nakamura et al. 2004). The overall level of homology between these 9 *PPR* genes suggests a pattern of evolution through local sequence duplication (Akagi, Nakamura et al. 2004). Similar clusters of *PPR* genes have been observed in the restorer locus of petunia (Bentolila, Alfonso et al. 2002) and radish

(Brown, Formanová et al. 2003, Desloire, Gherbi et al. 2003, Koizuka, Imai et al. 2003). In these loci, the restorer of fertility clusters with other restorer of fertility-like *PPR* genes usually presenting a high level of sequence homology with each other (Bentolila, Alfonso et al. 2002, Brown, Formanová et al. 2003, Akagi, Nakamura et al. 2004). It was suggested that this pattern of clustering in various plants might show diversifying selection acting on *PPR* genes from these regions (Geddy and Brown 2007, O'Toole, Hattori et al. 2008).

A diversifying selective pressure as an evolutionary process selects for, rather than against, mutations that would lead to amino acid replacements in the encoded proteins. As a result, plants would adapt to newly emerging sterility inducing genes by developing new *PPR* genes, a process analogous to the gene for gene evolution of disease resistance genes in response to newly emerging pathogen races (Touzet and Budar 2004). A genome wide distribution analysis of *PPR* genes indicates that although the vast majority of *PPR* genes are dispersed throughout the *A. thaliana* genome (Lurin, Andrés et al. 2004), a loose cluster of *PPR* genes is present on the long arm of chromosome 1 with 19 genes in close vicinity of each other (Desloire, Gherbi et al. 2003). Subsequently, Geddy and Brown 2007, showed that some *PPR* genes are rarely maintained in the same position or orientation between closely related species (*B. napus* vs *A. thaliana*). Furthermore, by establishing the ratio of nucleotide polymorphisms responsible for an amino-acid change versus silent substitution, Foxe and Wright 2009, determined that *PPR* genes are under diversifying or positive selection. Thus some *PPR* genes are “nomadic” in nature, i.e. they migrate from one genomic position to another, and under pressure to alter their

sequences, thus creating changes that will diversify the *PPR* gene population. This differs from most other *PPR* genes which tend to be under negative selection and thus to conserve the sequence of functional proteins (O'Toole, Hattori et al. 2008). Additionally, many *PPR* genes lack introns, suggesting their duplication may involve a retrotransposition type process.

Not only are some members of the PPR protein family under diversifying selective pressure but within a single protein, different amino acids are subject to different degrees of selective pressure. An extensive study of the PPR protein family in eleven angiosperm species, Fujii, Bond et al. 2011, revealed that Rf proteins fall within a specific clade of PPRs in which diversifying selection, as indicated by synonomous vs non-synonomous substitution rates, was 5 to 15 times higher at residues 1, 3, and 6 in each PPR motif than at the other amino acids of the domain. These amino acid residues were subsequently shown to be implicated in the recognition of their target RNA (Barkan, Rojas et al. 2012, Yin, Li et al. 2013, Takenaka, Zehrmann et al. 2013, Yagi, Hayashi et al. 2013). The changes in sequence driven by selective pressure are thus predicted to directly target the amino acids responsible for the conferring of CMS-associated transcript recognition. This pattern of diversifying selection has led to the hypothesis that this subgroup of PPR proteins are driven to evolve as sequence-specific RNA binding proteins to accommodate the appearance of new, CMS conferring mitochondrial genes. Thus diversifying selection aids in the creation of novel restorer of fertility genes to silence the newly arisen CMS conferring transcripts.

Fujii, Bond et al. 2011, designated the subgroup of PPR proteins encompassing fertility restorers as the Restorer of Fertility-Like (RFL) proteins within the P subfamily of PPR proteins. This analysis showed that most of the non-*RFL PPR* genes from different species form orthologous phylogenetic clusters, suggesting that these proteins are descended from ancestors present in the genome before the species diverged (O'Toole, Hattori et al. 2008, Fujii, Bond et al. 2011). In contrast, *RFL* genes form species-specific paralogous clusters, indicating that these genes have extensively evolved since these species diverged. Despite the rapid evolution of *RFL* genes, monocot and dicot *RFL* proteins form distinct lines of descent within the single clade of PPRs. Identified PPR restorer proteins from petunia, radish and rice all group within the corresponding line of descent within the *RFL* clade. The overall analysis provides strong support for the hypothesis that *RFL* and *PPR-Rf* genes have a monophyletic origin that precedes the modern monocot-dicot division, and have evolved quickly to produce a separate sub-group of proteins within P-class PPR proteins. Moreover, a significantly greater proportion of *RFL* genes shows diversifying selection, as measured by non-synonymous vs. synonymous substitution rates, than is observed in non-RFL *PPR* genes. Thus, about 10% of *RFL* genes show high probabilities of diversifying selection compared with most other genes in the genome, and in particular, compared with other *PPR* genes (Fujii, Bond et al. 2011). These findings suggest the *RFL* subgroup of *PPR* genes serves as a pool from which new *Rf* genes can emerge.

Characterization of several Arabidopsis *RFL* genes has indicated that they represent poorly conserved *PPR* genes and that some of them direct non-essential endoribonuclease

processing events within mitochondrial transcripts. One of them, *RFL9*, may have evolved in response to a CMS in Arabidopsis although the corresponding T-DNA mutants were not found to be male sterile (Arnal, Quadrado et al. 2014). *RFL9* clusters with a subgroup of 12 highly similar Arabidopsis RFL proteins for which no obvious orthologs can be found even in *Arabidopsis lyrata* (Fujii et al. 2011). This indicates that within one species a subgroup of fast-evolving *RFL* genes can emerge in a relatively short evolutionary time period, confirming the predictions of Fujii, Bond et al. 2011, at the species level. Within the same 12 RFL proteins subgroup, RFL9 is extremely similar to four other RFL proteins (Arnal, Quadrado et al. 2014). The sequence similarity between these five genes extends from the promoter region through the coding sequence into downstream sequences. The findings suggest active sequence exchanges may be occurring within genes. This, in turn, could promote sequence re-shuffling allowing the multiplication and diversification of *RFL* genes in plants in order to create new *RFL* copies (Arnal, Quadrado et al. 2014).

Non-PPR fertility restorer genes and their relationships with Rf-PPRs

As mentioned previously, restoration of fertility includes several other well-documented mechanisms, which do not involve PPR proteins, at least directly (table 1.1). The processing actors include glycine rich proteins (GRPs), alcohol dehydrogenase, acyl-carrier proteins and a peptidase. Beside GRPs, most of these restoration mechanisms act at the metabolic level rather than on the transcripts of the chimeric CMS-conferring genes.

GRPs

Recent studies have opened a new perspective in the fertility restoration field with the characterization of glycine rich proteins (GRP) involved in the restoration process of several CMS systems. Hu, Wang et al. 2012, first revealed the involvement of GRP162 in the restoration process of CMS-Honglian (HL) in rice. HL-CMS is associated with by the di-cistronic transcript *atp6-orfH79* and male fertility can be independently restored by either of two restorer genes, *Rf5* or *Rf6* (Liu, Xu et al. 2004). The cloning of *Rf5* revealed that it was identical to *Rf1A* in BT-CMS rice (as discussed previously in this review). However, the fertility restoration function of *Rf5* requires the co-action of a glycine rich protein, GRP162 (Hu, Wang et al. 2012). The presence of GRP162 alone is responsible for a translation inhibition of *orfH79* and allows fertility restoration. GRP162 has two RNA recognition motifs that bind to and induce the processing of *orf79* (Hu, Huang et al. 2013) but does not contain a mitochondrial targeting sequence. It has been proposed that Rf5 would recruit GRP162 to the mitochondria by forming a heterodimer in the cytosol so that those two components of a larger restoration complex could bind and process CMS-associated transcripts (Hu, Wang et al. 2012).

In plants, GRP proteins are characterized by the presence of semi-repetitive glycine-rich motifs. The classification of this large protein family depends on their general structure, the arrangement of the glycine repeats, as well as the presence of conserved motifs (Mangeon, Junqueira et al. 2010). Of particular interest, class IV GRPs contain an RNA-recognition motif and are known to bind RNAs. The RNA-binding activity of these

proteins has been biochemically demonstrated, suggesting that they may be involved in RNA stabilization, processing or transport. Some class IV GRPs have also have an RNA-chaperone activity (Mangeon, Junqueira et al. 2010) and have been proposed to resolve non-functional inhibitory transcript structures. GRPs act in numerous processes and the specific function of the glycine-rich domain still remains unclear. It can be assumed, however, that proteins known to be implicated in RNA recognition and processing as GRPs could take part in the mechanisms of fertility restoration as seen in HL-CMS in rice with GRP162.

GRPs have also been proposed to function as fertility restorers in the absence of PPR proteins. Rf2, the restorer protein for Lead-Rice CMS (Itabashi, Iwata et al. 2011) has been shown to interact with RIF2, a ubiquitin domain-containing protein (Fujii, Kazama et al. 2014). The interaction of the two proteins suggests the presence of a large restoration complex targeting the degradation of the CMS-causing protein and implying a process of fertility restoration that does not include transcript processing or translation inhibition

Aldehyde dehydrogenase

Other non-PPR male fertility restorers have been characterized in different crops, all using a fertility restoration process that does not include post-transcriptional modification. *Rf2* in T-CMS in maize (Cui, Wise et al. 1996) encodes a mitochondrial acetaldehyde dehydrogenase (Liu, Cui et al. 2001) and has no effect on the expression of

the T-CMS associated mitochondrial gene T-*urf13*. The current model of action of Rf2 proposes that it would relieve the oxidative stress caused by oxidizing a number of unusual aldehydes produced in plants carrying the CMS conferring polypeptide, URF13 (Liu, Cui et al. 2001). However *Rf2* also affects normal anther development in N cytoplasm maize, which carries a non-sterility inducing cytoplasm and partial activity of the Rf2 protein leads to arrest in anther development.

Acyl-carrier proteins

The restoration factor of (CW)-type CMS in rice, Rf17, has been found to encode a protein of unknown function bearing partial homology with acyl-carrier proteins (Fujii and Toriyama 2009). The function of this restorer protein is unclear but it has been proposed that it would restore fertility by retrograde regulation, i.e. by altering the nuclear response to mitochondrial function. It was shown that the reduced-expression allele of *Rf17* restored fertility in haploid pollen, whereas a normal-expression allele caused pollen lethality in the CW-type CMS (Fujii and Toriyama 2009). Although there were no indications of Rf17 function other than the partial acyl carrier protein synthase-like domain, the authors speculated that some metabolic alteration in mitochondria restores pollen fertility, similar to the mechanism in the maize Rf2 system described above (Fujii and Toriyama 2009).

Peptidases

In sugar beet (*Beta vulgaris* L.), restoration of fertility of the Owen CMS also does not include a PPR protein. The *Rf* gene of this system was mapped to a region that does not contain any *PPR* genes (Matsuhira, Kagami et al. 2012). The restoration activity is carried by the *bvORF20* gene, which encodes a mitochondrial-targeted protein exhibiting strong homology with the OMA1-like metallopeptidase (Kitazaki, Arakawa et al. 2015). It was shown that bvORF20 interacts with the Owen CMS conferring mitochondrial polypeptide, preSATP6 and that *bvORF20* expression correlates with a decrease of a 250 kDa membrane-bound complex containing the preSATP6 protein. It has been proposed that bvORF20 would restore fertility post-translationally by limiting the homo-oligomerization of preSATP6 in mitochondrial membranes.

The propensity of Pentatricopeptide Repeat genes to evolve into male fertility restorers

The data presented in this literature review explore the functional specificities of the different restorers of fertility identified in plants so far. It highlights the wide array of mechanisms guiding fertility restoration, which are in most cases unique to each CMS. However PPR proteins represent the most frequent protein class among identified Rf. It strongly supports the notion that PPR proteins exhibit ideal characteristics to evolve into restorers of fertility when the mechanism of restoration implies a post-transcriptional action on mitochondrial gene expression. They have the ability to bind specific RNAs

with high specificity and to impact the processing or the expression of their target RNA in several ways. To suppress male sterility, they often induce endoribonucleolytic cleavage or act as a physical barrier to block translation of CMS conferring transcripts. The diversifying evolutionary pressures acting on *PPR* genes in general and notably on the Rf-like PPR sub-group greatly accelerate the emergence of novel PPR alleles in plant populations which can be selected as fertility restorers when they bind to and impact the expression of CMS-causing transcripts. The versatility by which PPR proteins can block the expression of CMS transcripts and the rapidity of adaptation of *Rf-like* genes to newly arising RNA sequences likely explain why fertility restorer genes correspond to *PPR* genes. Recent data indicate that glycine rich proteins may act in concert with PPR proteins to suppress CMS transcript expression. Some of these GRPs have RNA binding activity and thereby may assist PPR proteins to bind CMS transcripts in a productive way. The mode of action of other GRPs in fertility restoration needs to be clarified but they may stabilize PPR/RNA complex and thus accelerate the emergence of new PPR-Rf.

Native CMS systems in *Brassica napus*

Brassica genomes

The cytogenetic and evolutionary relationships among the major oilseed and vegetable species, like plants of genus Brassica, are commonly depicted as U's triangle, named after the Korean scientist who first formulated it (U, 1935). U speculated that *B. carinata*, *B. juncea* and *B. napus* are each allotetraploids formed by interspecific hybridization

events between the parental diploid species *B. nigra*, *B. rapa* and *B. oleracea*. Thus, hybridization between *B. oleracea* and *B. rapa* resulted in the formation of *B. napus*. The relationships among these species first postulated by U have since been confirmed by a large variety of genetic and molecular analyses. The haploid genomes of *B. rapa*, *B. nigra* and *B. oleracea* are designated A, B and C respectively. Thus diploid *B. rapa* has two copies of the A genome within 20 chromosomes (AA, n=10, 2n=20) and diploid *B. napus* has two copies of both the A and C genomes within 38 chromosomes (AACC, n=19, 2n=38).

The model plant *A. thaliana* and the Brassica species belong to the same plant family, the Brassicaceae. Initial efforts at determining the relationships between the Brassica and Arabidopsis genomes involved using molecular probes for co-linear sets derived from the developing Arabidopsis genomic resources to map RFLP polymorphisms in Brassica species (Kowalski, Lan et al. 1994). These studies indicated that there is extensive co-linearity between the Brassica and Arabidopsis genomes, but that most single copy Arabidopsis regions exist as multiple, on average 3 copies in modern Brassica genomes (Lagercrantz, Putterill et al. 1996, Truco, Hu et al. 1996, Cavell, Lydiate et al. 1998). This in turn, gave rise to the hypothesis that the modern diploid Brassica species are derived from a hexaploid ancestor whose genome was generated from a diploid, Arabidopsis-like genome through polyploidization events. This view has largely been confirmed through subsequent comparative mapping studies involving much larger numbers of markers whose position was known on the sequenced Arabidopsis genome (Parkin, Gulden et al. 2005). The latter study allowed for the identification of a large

number of segments of the *B. napus* genome that are co-linear with corresponding regions of Arabidopsis. The average length of these co-linear segements was 14.3 cM in Brassica, corresponding to 4.3 Mb in *A. thaliana*, suggesting that, on average, 1 cM genetic distance in *B. napus* corresponded to 300 kb in physical distance in *A. thaliana*. Similar high definition mapping experiments in *B. rapa* indicate that 1 cM in this species corresponds to 341 kb in Arabidopsis and thus a similar relationship between physical and genetic distances in the two species.

Therefore, a given "single copy" region of the *A. thaliana* genome is present, on average, six times in the *B. napus* genome. Underlying this complexity, it is now widely accepted that more ancient genome duplication events occurred during the evolution of angiosperms (Blanc and Wolfe 2004, Adams and Wendel 2005), further increasing the complexity of Brassica genomes. Thus, to perform map-based cloning in *B. napus*, it is necessary to be able to distinguish which of the multiple copies of a given genome segment correspond to that linked to the gene of interest.

Native CMS systems in Brassica napus

In the oilseed rape species, *B napus*, the two endogenous cytoplasms capable of inducing CMS are named napus (*nap*) and polima (*pol*). *nap* CMS is associated with the presence of a chimeric ORF, *orf222*, that is positioned upstream of the *nad5c* exon, the third exon of the gene encoding a subunit of the membrane-embedded arm of the mitochondrial respiratory chain NADH-ubiquinol-oxidoreductase or complex I (L'Homme, Stahl et al.

1997). *orf222* is co-transcribed with *nad5c* and an unknown open reading frame (*orf101*), in a single transcriptional unit as *orf222/nad5c/orf101*. In contrast, *pol* CMS is correlated with the presence of the chimeric *orf224*, that is situated upstream of *atp6*, the gene encoding subunit 6 of mitochondrial F_1-F_0 ATP synthase, complex V (Singh and Brown 1991). Unlike any other plant species with multiple CMS systems, *orf222* and *orf224* show a high degree of sequence similarity both at the nucleotide and amino acid levels. Although they do not code for any known gene, the 5' non-coding region and first 58 codons of *orf222* and *orf224* are derived from a plant mitochondrial gene that encodes subunit 8 of the mitochondrial F_1-F_0 ATP synthase (L'Homme, Stahl et al. 1997).

The restorers for the *pol* and *nap* systems, *Rfp* and *Rfn*, respectively, each down-regulate the expression of their cognate CMS-associated mitochondrial genes by mediating RNA cleavage events within unique regions of the corresponding transcripts. In *pol* sterile plants, *orf224/atp6* transcripts are predominantly di-cistronic. In flowers of *Rfp* restored plants, the levels of di-cistronic transcripts decrease, and two new transcripts appear whose 5' termini map within *orf224*. In addition, a slight increase in the level of a 1.1 kb mono-cistronic *atp6* transcript present in low amounts in CMS plants, is observed (Li, Jean et al. 1998). The two new transcripts do not carry an initiator 5' di or triphosphate (Menassa, L'Homme et al. 1999) and co-segregate perfectly with fertility restoration suggesting that *Rfp* acts by processing the CMS inducing transcript in order to restore pollen production.

Just like in the *Rfp* restored plants, flowers of *Rfn* restored plants present decreased levels

of *orf222/nad5c/orf101* tri-cistronic transcripts and the appearance of a new transcript, containing *nad5c/orf101*, is observed. The *Rfn* allele is also associated with additional RNA cleavage events in the coding regions of *nad4,* another subunit of complex I, which is not observed in plants homozygous for the *Rfp* allele or for the non restoring, or universal maintainer genotype *rf* (Li, Jean et al. 1998). Indeed, *Rfn* and *Rfp* represent distinct alleles or haplotypes of a single nuclear locus, and it has not been possible to dissociate these genes or their associated mtRNA cleavage properties via genetic crosses involving the three nuclear and cytoplasmic genotypes (Li, Jean et al. 1998).

Thesis Objectives

This PhD thesis highlights the work done towards the identification of restorers of fertility in *B. napus* CMS systems with a focus on *Rfn*, restorer of nap CMS. In chapter II, I present the mapping of 3' and 5' processing sites in order to get a clearer idea of how Rfn acts on *orf222/nad5c/orf101* transcripts. Using CR-RT-PCR, characterization of the dominant termini of RNAs of interest was performed. It included study of *orff22/nad5c/orf101* transcripts as well as *atp8* and *nad4*. Sequence similarity among these termini allowed for the identification of potential Rfn target sequences. In Chapter III, genetic mapping of *Rfn* is presented. Using a backcross population of 300 individuals, I developed SNPs as well as ILPs and CAPS markers. The position of the *Rfn* locus was narrowed down to a 600 kb region in the *B napus* genome. I was able to then explore that region to determine what genes could be likely candidates. Analysis of the synteny between *B. napus, B. rapa, A. thaliana* and *A. lyrata* allowed the highlighting of

evolution patterns characteristic of *PPR* genes. At the end of chapter III, I consider the factors involved in considering each of several *Rfn* candidates. Finally, in chapter IV, I demonstrate the progress that has been made towards the cloning of a variety of genetic constructs involving the candidate genes and their transformation into *B. napus* using a simplified protocol developed through my experiments. I also present a variety of experiments that were performed in order to facilitate molecular complementation and explore the properties of the candidate proteins.

FIGURES AND TABLES

	Restorer name	Type of Restorer Protein	Species	CMS name	CMS inducing gene	Caracteristic and models for fertility restoration	References
restoration by post-transcriptionnal events	Rf-PPR592	PPR protein	petunia	-	*pcf*	binds to CMS transcript alters transcript expression profile in a large complex with CMS transcript associated with inner mitochondrial membrane	Bentotlila 2002, Gilman 2007
	Rfo	PPR protein	radish	Ogura	*orf138-orfB*	binds to CMS transcript does not promote processing or degradation reduced levels of CMS protein in restored plants	Brown 2003, Desloire 2003, Koizuka 2003, Qin 2014, Uyttewaal 2008
	Rf1A, Rf1B	PPR protein	rice	BT	*atp6-orf79*	mediation of endonucleolytic processing of cms transcript destabilization of di-cistronic transcript	Wang 2006
	Rf4, Rf3	PPR protein	rice	wild abortive (WA)	*WA352*	reduced levels of CMS confering transcript	Tang 2014
	Rf5, GRP162, Rf6	PPR protein + Glycine Rich Protein	rice	Hong-Lian	*atp6-orfH79*	PPR not binding CMS transcript translation inhibition by GRP162 GRP162 + Rf5 induces transcript processing	Hu 2012, Hu 2013
restoration by other mechanisms	Rf2	Glycine Rich Protein	rice	Lead	?	within a large protein complex targetting the CMS protein	Itabashi 2011, Fujii 2014
	Rf2	ALDH protein	maize	Texas	*urf13*	remove oxidative stress caused by URF13	Liu 2001
	Rf17	acyl carrier protein	rice	CW	?	retrograde regulation diminution of expression levels of Rf17 induces metabolic alteration in mitochondria that leads to restoration	Fujii 2009
	bvORF20	OMA1-like protein	beet	Owen	*preSatp6*	interaction with CMS polypeptide Impact the homo-oligomerization of the CMS peptide in mitochondrial membranes	Kaser 2003, Kitazaki 2015

Table 1.1. Representation of the characterized restorers of fertility (confirmed through transgenic analysis) and their updated model for restoration mechanism. To allow better visualization, highlighting of the PPR proteins characterized as restorers acting at the post-transcriptional level was done in dark grey wherease light grey highlighting represents other restorers characterized to act through non-post-transcriptional mechanisms.

References

Abdelnoor, R. V., R. Yule, A. Elo, A. C. Christensen, G. Meyer-Gauen and S. A. Mackenzie (2003). "Substoichiometric shifting in the plant mitochondrial genome is influenced by a gene homologous to MutS." Proceedings of the National Academy of Sciences 100(10): 5968-5973.

Adams, K. L. and J. F. Wendel (2005). "Polyploidy and genome evolution in plants." Current opinion in plant biology 8(2): 135-141.

Akagi, H., A. Nakamura, Y. Yokozeki-Misono, A. Inagaki, H. Takahashi, K. Mori and T. Fujimura (2004). "Positional cloning of the rice Rf-1 gene, a restorer of BT-type cytoplasmic male sterility that encodes a mitochondria-targeting PPR protein." Theoretical and applied genetics 108(8): 1449-1457.

Akagi, H., M. Sakamoto, C. Shinjyo, H. Shimada and T. Fujimura (1994). "A unique sequence located downstream from the rice mitochondrial atp6 may cause male sterility." Current genetics 25(1): 52-58.

Alonso, C. and C. M. Herrera (2001). "Neither vegetative nor reproductive advantages account for high frequency of male-steriles in southern Spanish gynodioecious *Daphne laureola* (Thymelaeaceae)." American Journal of Botany 88(6): 1016-1024.

Arnal, N., M. Quadrado, M. Simon and H. Mireau (2014). "A restorer-of-fertility like pentatricopeptide repeat gene directs ribonucleolytic processing within the coding sequence of *rps3-rpl16* and *orf240a* mitochondrial transcripts in *Arabidopsis thaliana*." The Plant Journal 78(1): 134-145.

Balk, J. and C. J. Leaver (2001). "The PET1-CMS mitochondrial mutation in sunflower is associated with premature programmed cell death and cytochrome c release." The Plant Cell 13(8): 1803-1818.

Ban, T., J. Ke, R. Chen, X. Gu, M. E. Tan, X. E. Zhou, Y. Kang, K. Melcher, J.-K. Zhu and H. E. Xu (2013). "Structure of a PLS-class pentatricopeptide repeat protein provides insights into mechanism of RNA recognition." Journal of Biological Chemistry 288(44): 31540-31548.

Barkan, A., M. Rojas, S. Fujii, A. Yap, Y. S. Chong, C. S. Bond and I. Small (2012). "A combinatorial amino acid code for RNA recognition by pentatricopeptide repeat proteins." PLOS genetics 8(8): e1002910.

Barr, C. M. and L. Fishman (2010). "The nuclear component of a cytonuclear hybrid incompatibility in Mimulus maps to a cluster of pentatricopeptide repeat genes." Genetics 184(2): 455-465.

Bendich, A. J. (1993). "Reaching for the ring: the study of mitochondrial genome structure." Current genetics 24(4): 279-290.

Bentolila, S., A. A. Alfonso and M. R. Hanson (2002). "A pentatricopeptide repeat-containing gene restores fertility to cytoplasmic male-sterile plants." Proceedings of the National Academy of Sciences 99(16): 10887-10892.

Blanc, G. and K. H. Wolfe (2004). "Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes." The Plant Cell 16(7): 1667-1678.

Bonen, L. and J. Vogel (2001). "The ins and outs of group II introns." Trends in Genetics 17(6): 322-331.

Bonhomme, S., F. Budar, D. Lancelin, I. Small, M.-C. Defrance and G. Pelletier (1992). "Sequence and transcript analysis of the *Nco2.5* Ogura-specific fragment correlated with cytoplasmic male sterility in Brassica cybrids." Molecular and General Genetics 235(2-3): 340-348.

Brown, G., C. C. des Francs-Small and O. Ostersetzer (2014). "Group-II intron splicing factors in higher-plants mitochondria." Front Plant Physiology Science 5: 35.

Brown, G. G., N. Formanová, H. Jin, R. Wargachuk, C. Dendy, P. Patil, M. Laforest, J. Zhang, W. Y. Cheung and B. S. Landry (2003). "The radish *Rfo* restorer gene of Ogura cytoplasmic male sterility encodes a protein with multiple pentatricopeptide repeats." The Plant Journal 35(2): 262-272.

Budar, F. and G. Pelletier (2001). "Male sterility in plants: occurrence, determinism, significance and use." Comptes Rendus de l'Académie des Sciences-Series III-Sciences de la Vie 324(6): 543-550.

Budar, F., P. Touzet and R. De Paepe (2003). "The nucleo-mitochondrial conflict in cytoplasmic male sterilities revisited." Genetica 117(1): 3-16.

Cavell, A. C., D. Lydiate, I. Parkin, C. Dean and M. Trick (1998). "Collinearity between a 30-centimorgan segment of *Arabidopsis thaliana* chromosome 4 and duplicated regions within the *Brassica napus* genome." Genome 41(1): 62-69.

Chen, L. and Y.-G. Liu (2014). "Male sterility and fertility restoration in crops." Annual review of plant biology 65: 579-606.

Cui, X., R. P. Wise and P. S. Schnable (1996). "The *rf2* nuclear restorer gene of male-sterile T-cytoplasm maize." Science-New York then Washington: 1334-1335.

Dahan, J. and H. Mireau (2013). "The Rf and Rf-like PPR in higher plants, a fast-evolving subclass of PPR genes." RNA biology 10(9): 1469-1476.

Desloire, S., H. Gherbi, W. Laloui, S. Marhadour, V. Clouet, L. Cattolico, C. Falentin, S. Giancola, M. Renard and F. Budar (2003). "Identification of the fertility restoration locus, *Rfo*, in radish, as a member of the pentatricopeptide-repeat protein family." EMBO reports 4(6): 588-594.

Dewey, R., J. Siedow, D. Timothy and C. Levings (1988). "A 13-kilodalton maize mitochondrial protein in *E. coli* confers sensitivity to Bipolaris maydis toxin." Science 239(4837): 293-295.

Dewey, R., D. Timothy and C. Levings (1987). "A mitochondrial protein associated with cytoplasmic male sterility in the T cytoplasm of maize." Proceedings of the National Academy of Sciences 84(15): 5374-5378.

Dufay, M. and E. Billard (2012). "How much better are females? The occurrence of female advantage, its proximal causes and its variation within and among gynodioecious species." Annals of Botany 109(3): 505-519.

Forner, J., B. Weber, S. Thuss, S. Wildum and S. Binder (2007). "Mapping of mitochondrial mRNA termini in *Arabidopsis thaliana*: t-elements contribute to 5' and 3' end formation." Nucleic acids research 35(11): 3676-3692.

Foxe, J. P. and S. I. Wright (2009). "Signature of diversifying selection on members of the pentatricopeptide repeat protein family in *Arabidopsis lyrata*." Genetics 183(2): 663-672.

Fujii, S., C. S. Bond and I. D. Small (2011). "Selection patterns on restorer-like genes reveal a conflict between nuclear and mitochondrial genomes throughout angiosperm evolution." Proceedings of the National Academy of Sciences 108(4): 1723-1728.

Fujii, S., T. Kazama, Y. Ito, S. Kojima and K. Toriyama (2014). "A candidate factor that interacts with RF2, a restorer of fertility of Lead rice-type cytoplasmic male sterility in rice." Rice 7(1): 21.

Fujii, S. and K. Toriyama (2009). "Suppressed expression of RETROGRADE-REGULATED MALE STERILITY restores pollen fertility in cytoplasmic male sterile rice plants." Proceedings of the National Academy of Sciences 106(23): 9513-9518.

Gagliardi, D., R. Perrin, L. Maréchal-Drouard, J.-M. Grienenberger and C. J. Leaver (2001). "Plant mitochondrial polyadenylated mRNAs are degraded by a 3'-to 5'-exoribonuclease activity, which proceeds unimpeded by stable secondary structures." Journal of Biological Chemistry 276(47): 43541-43547.

Geddy, R. and G. G. Brown (2007). "Genes encoding pentatricopeptide repeat (PPR) proteins are not conserved in location in plant genomes and may be subject to diversifying selection." BMC genomics 8(1): 130.

Gillman, J. D., S. Bentolila and M. R. Hanson (2007). "The petunia restorer of fertility protein is part of a large mitochondrial complex that interacts with transcripts of the CMS-associated locus." The Plant Journal 49(2): 217-227.

González-Melendi, P., M. Uyttewaal, C. N. Morcillo, J. R. H. Mora, S. Fajardo, F. Budar and M. M. Lucas (2008). "A light and electron microscopy analysis of the events leading to male sterility in Ogu-INRA CMS of rapeseed (*Brassica napus*)." Journal of experimental botany 59(4): 827-838.

Gouyon, P.-H. and D. Couvet (1987). "A conflict between two sexes, females and hermaphrodites. The evolution of sex and its consequences." The evolution of sex and its consequences Springer: 245-261.

Gualberto, J. M., D. Mileshina, C. Wallet, A. K. Niazi, F. Weber-Lotfi and A. Dietrich (2014). "The plant mitochondrial genome: dynamics and maintenance." Biochimie 100: 107-120.

Gully, B. S., N. Cowieson, W. A. Stanley, K. Shearston, I. D. Small, A. Barkan and C. S. Bond (2015). "The solution structure of the pentatricopeptide repeat protein PPR10 upon binding *atpH* RNA." Nucleic Acids Research: gkv027.

Hammani, K. and P. Giege (2014). "RNA metabolism in plant mitochondria." Trends in Plant Science 19(6): 380-389.

Hanson, M. R. (1991). "Plant mitochondrial mutations and male sterility." Annual review of genetics 25(1): 461-486.

Hanson, M. R. and S. Bentolila (2004). "Interactions of mitochondrial and nuclear genes that affect male gametophyte development." The Plant Cell 16(suppl 1): S154-S169.

Hauler, A., C. Jonietz, B. Stoll, K. Stoll, H. P. Braun and S. Binder (2013). "RNA PROCESSING FACTOR 5 is required for efficient 5' cleavage at a processing site conserved in RNAs of three different mitochondrial genes in *Arabidopsis thaliana*." The Plant Journal 74(4): 593-604.

Holec, S., H. Lange, J. Canaday and D. Gagliardi (2008). "Coping with cryptic and defective transcripts in plant mitochondria." Biochimica et Biophysica Acta-Gene Regulatory Mechanisms 1779(9): 566-573.

Hölzle, A., C. Jonietz, O. Törjek, T. Altmann, S. Binder and J. Forner (2011). "A RESTORER OF FERTILITY-like PPR gene is required for 5'-end processing of the *nad4* mRNA in mitochondria of *Arabidopsis thaliana*." The Plant Journal 65(5): 737-744.

Hu, J., W. Huang, Q. Huang, X. Qin, Z. Dan, G. Yao, R. Zhu and Y. Zhu (2013). "The mechanism of ORFH79 suppression with the artificial restorer fertility gene *Mt-GRP162*." New Phytologist 199(1): 52-58.

Hu, J., W. Huang, Q. Huang, X. Qin, C. Yu, L. Wang, S. Li, R. Zhu and Y. Zhu (2014). "Mitochondria and cytoplasmic male sterility in plants." Mitochondrion 19, Part B(0): 282-288.

Hu, J., K. Wang, W. Huang, G. Liu, Y. Gao, J. Wang, Q. Huang, Y. Ji, X. Qin, L. Wan, R. Zhu, S. Li, D. Yang and Y. Zhu (2012). "The Rice Pentatricopeptide Repeat Protein RF5 Restores Fertility in Hong-Lian Cytoplasmic Male-Sterile Lines via a Complex with the Glycine-Rich Protein GRP162." The Plant Cell 24(1): 109-122.

Itabashi, E., N. Iwata, S. Fujii, T. Kazama and K. Toriyama (2011). "The fertility restorer gene, *Rf2*, for Lead Rice-type cytoplasmic male sterility of rice encodes a mitochondrial glycine-rich protein." The plant journal 65(3): 359-367.

Janska, H., R. Sarria, M. Woloszynska, M. Arrieta-Montiel and S. A. Mackenzie (1998). "Stoichiometric shifts in the common bean mitochondrial genome leading to male sterility and spontaneous reversion to fertility." The Plant Cell 10(7): 1163-1180.

Jonietz, C., J. Forner, T. Hildebrandt and S. Binder (2011). "RNA PROCESSING FACTOR3 is crucial for the accumulation of mature *ccmC* transcripts in mitochondria of Arabidopsis accession Columbia." Plant physiology 157(3): 1430-1439.

Jonietz, C., J. Forner, A. Hölzle, S. Thuss and S. Binder (2010). "RNA PROCESSING FACTOR2 is required for 5' end processing of *nad9* and *cox3* mRNAs in mitochondria of *Arabidopsis thaliana*." The Plant Cell 22(2): 443-453.

Jordan, D., R. Klein, K. Sakrewski, R. Henzell, P. Klein and E. Mace (2011). "Mapping and characterization of *Rf5*: a new gene conditioning pollen fertility restoration in A1 and A2 cytoplasm in sorghum (*Sorghum bicolor (L.) Moench*)." Theoretical and applied genetics 123(3): 383-396.

Jordan, D., E. S. Mace, R. Henzell, P. Klein and R. Klein (2010). "Molecular mapping and candidate gene identification of the *Rf2* gene for pollen fertility restoration in sorghum (*Sorghum bicolor (L.) Moench*)." Theoretical and applied genetics 120(7): 1279-1287.

Kazama, T. and K. Toriyama (2003). "A pentatricopeptide repeat-containing gene that promotes the processing of aberrant *atp6* RNA of cytoplasmic male-sterile rice." FEBS letters 544(1): 99-102.

Ke, J., R.-Z. Chen, T. Ban, X. E. Zhou, X. Gu, M. E. Tan, C. Chen, Y. Kang, J. S. Brunzelle and J.-K. Zhu (2013). "Structural basis for RNA recognition by a dimeric PPR-protein complex." Nature structural & molecular biology 20(12): 1377-1382.

Kitazaki, K., T. Arakawa, M. Matsunaga, R. Yui-Kurino, H. Matsuhira, T. Mikami and T. Kubo (2015). "Post-translational mechanisms are associated with fertility restoration of cytoplasmic male sterility in sugar beet." The Plant Journal.

Klein, R. R. and D. E. Koeppe (1985). "Mode of methomyl and Bipolaris maydis (race T) toxin in uncoupling Texas male-sterile cytoplasm corn mitochondria." Plant physiology 77(4): 912-916.

Koizuka, N., R. Imai, H. Fujimoto, T. Hayakawa, Y. Kimura, J. Kohno-Murase, T. Sakai, S. Kawasaki and J. Imamura (2003). "Genetic characterization of a pentatricopeptide repeat protein gene, *orf687*, that restores fertility in the cytoplasmic male-sterile Kosena radish." The plant journal 34(4): 407-415.

Komori, T., S. Ohta, N. Murai, Y. Takakura, Y. Kuraya, S. Suzuki, Y. Hiei, H. Imaseki and N. Nitta (2004). "Map-based cloning of a fertility restorer gene, *Rf-1*, in rice (*Oryza sativa L.*)." The Plant Journal 37(3): 315-325.

Kowalski, S. P., T.-H. Lan, K. A. Feldmann and A. H. Paterson (1994). "Comparative mapping of *Arabidopsis thaliana* and *Brassica oleracea* chromosomes reveals islands of conserved organization." Genetics 138(2): 499-510.

Kubo, T. and K. J. Newton (2008). "Angiosperm mitochondrial genomes and mutations." Mitochondrion 8(1): 5-14.

Kühn, K., A. Weihe and T. Börner (2005). "Multiple promoters are a common feature of mitochondrial genes in Arabidopsis." Nucleic Acids Research 33(1): 337-346.

L'Homme, Y., R. J. Stahl, X.-Q. Li, A. Hameed and G. G. Brown (1997). "Brassica nap cytoplasmic male sterility is associated with expression of a mtDNA region containing a chimeric gene similar to the pol CMS-associated *orf224* gene." Current genetics 31(4): 325-335.

Lagercrantz, U., J. Putterill, G. Coupland and D. Lydiate (1996). "Comparative mapping in Arabidopsis and Brassica, fine scale genome collinearity and congruence of genes controlling flowering time." The Plant Journal 9(1): 13-20.

Lane, N. (2005). “Power, sex, suicide: mitochondria and the meaning of life.” Oxford University Press.

Laser, K. D. and N. R. Lersten (1972). "Anatomy and cytology of microsporogenesis in cytoplasmic male sterile angiosperms." The Botanical Review 38(3): 425-454.

Lee, S.-L. J. and H. E. Warmke (1979). "Organelle Size and Number in Fertile and T-Cytoplasmic Male-Sterile Corn." American Journal of Botany 66(2): 141-148.

Levings, C. and G. Brown (1989). "Molecular biology of plant mitochondria." Cell 56(2): 171-179.

Li, X.-Q., M. Jean, B. S. Landry and G. G. Brown (1998). "Restorer genes for different forms of Brassica cytoplasmic male sterility map to a single nuclear locus that modifies transcripts of several mitochondrial genes." Proceedings of the National Academy of Sciences 95(17): 10032-10037.

Liu, F., X. Cui, H. T. Horner, H. Weiner and P. S. Schnable (2001). "Mitochondrial aldehyde dehydrogenase activity is required for male fertility in maize." The Plant Cell 13(5): 1063-1078.

Liu, X.-Q., X. Xu, Y.-P. Tan, S.-Q. Li, J. Hu, J.-Y. Huang, D.-C. Yang, Y.-S. Li and Y.-G. Zhu (2004). "Inheritance and molecular mapping of two fertility-restoring loci for Honglian gametophytic cytoplasmic male sterility in rice (*Oryza sativa L.*)." Molecular Genetics and Genomics 271(5): 586-594.

Luo, D., H. Xu, Z. Liu, J. Guo, H. Li, L. Chen, C. Fang, Q. Zhang, M. Bai and N. Yao (2013). "A detrimental mitochondrial-nuclear interaction causes cytoplasmic male sterility in rice." Nature genetics 45(5): 573-577.

Lurin, C., C. Andrés, S. Aubourg, M. Bellaoui, F. Bitton, C. Bruyère, M. Caboche, C. Debast, J. Gualberto and B. Hoffmann (2004). "Genome-wide analysis of Arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis." The Plant Cell 16(8): 2089-2103.

Mangeon, A., R. M. Junqueira and G. Sachetto-Martins (2010). "Functional diversity of the plant glycine-rich proteins superfamily." Plant signaling & behavior 5(2): 99-104.

Matsuhira, H., H. Kagami, M. Kurata, K. Kitazaki, M. Matsunaga, Y. Hamaguchi, E. Hagihara, M. Ueda, M. Harada and A. Muramatsu (2012). "Unusual and typical features of a novel restorer-of-fertility gene of sugar beet (Beta vulgaris L.)." Genetics: genetics. 112.145409.

Menassa, R., Y. L’Homme and G. G. Brown (1999). "Post-transcriptional and developmental regulation of a CMS-associated mitochondrial gene region by a nuclear restorer gene." The Plant Journal 17(5): 491-499.

Newton, K. J., C. Knudsen, S. Gabay-Laughnan and J. R. Laughnan (1990). "An abnormal growth mutant in maize has a defective mitochondrial cytochrome oxidase gene." The Plant Cell 2(2): 107-113.

O'Toole, N., M. Hattori, C. Andres, K. Iida, C. Lurin, C. Schmitz-Linneweber, M. Sugita and I. Small (2008). "On the expansion of the pentatricopeptide repeat gene family in plants." Molecular Biology and Evolution 25(6): 1120-1128.

Palmer, J. D. and L. A. Herbon (1988). "Plant mitochondrial DNA evolved rapidly in structure, but slowly in sequence." Journal of Molecular Evolution 28(1-2): 87-97.

Palmer, J. D. and C. R. Shields (1984). "Tripartite structure of the *Brassica campestris* mitochondrial genome." Nature 307: 437-440.

Parkin, I. A., S. M. Gulden, A. G. Sharpe, L. Lukens, M. Trick, T. C. Osborn and D. J. Lydiate (2005). "Segmental structure of the *Brassica napus* genome based on comparative analysis with *Arabidopsis thaliana*." Genetics 171(2): 765-781.

Pinto, M. and C. T. Moraes (2014). "Mitochondrial genome changes and neurodegenerative diseases." Biochimica et Biophysica Acta - Molecular Basis of Disease 1842(8): 1198-1207.

Qin, X., R. Warguchuk, N. Arnal, L. Gaborieau, H. Mireau and G. G. Brown (2014). "In vivo functional analysis of a nuclear restorer PPR protein." BMC plant biology 14(1): 313.

Sabar, M., D. Gagliardi, J. Balk and C. J. Leaver (2003). "ORFB is a subunit of F1FO-ATP synthase: insight into the basis of cytoplasmic male sterility in sunflower." EMBO Reports 4(4): 381-386.

Sandhu, A. P. S., R. V. Abdelnoor and S. A. Mackenzie (2007). "Transgenic induction of mitochondrial rearrangements for cytoplasmic male sterility in crop plants." Proceedings of the National Academy of Sciences 104(6): 1766-1770.

Shedge, V., M. Arrieta-Montiel, A. C. Christensen and S. A. Mackenzie (2007). "Plant mitochondrial recombination surveillance requires unusual RecA and MutS homologs." The Plant Cell 19(4): 1251-1264.

Singh, M., & Brown, G. G. (1991). “Suppression of cytoplasmic male sterility by nuclear genes alters expression of a novel mitochondrial gene region.” The Plant Cell 3(12): 1349-1362.

Small, I. D., P. G. Isaac and C. J. Leaver (1987). "Stoichiometric differences in DNA molecules containing the *atpA* gene suggest mechanisms for the generation of mitochondrial genome diversity in maize." The EMBO Journal 6(4): 865.

Smart, C. J., F. Monéger and C. J. Leaver (1994). "Cell-specific regulation of gene expression in mitochondria during anther development in sunflower." The Plant Cell 6(6): 811-825.

Stoll, B., K. Stoll, J. Steinhilber, C. Jonietz and S. Binder (2013). "Mitochondrial transcript length polymorphisms are a widespread phenomenon in *Arabidopsis thaliana*." Plant molecular biology 81(3): 221-233.

Takenaka, M., A. Zehrmann, A. Brennicke and K. Graichen (2013). "Improved computational target site prediction for pentatricopeptide repeat RNA editing factors." PLOS one 8(6): e65343.

Tang, H., D. Luo, D. Zhou, Q. Zhang, D. Tian, X. Zheng, L. Chen and Y.-G. Liu (2014). "The rice restorer *Rf4* for wild-abortive cytoplasmic male sterility encodes a mitochondrial-localized PPR protein that functions in reduction of *WA352* transcripts." Molecular plant 7(9): 1497-1500.

Touzet, P. and F. Budar (2004). "Unveiling the molecular arms race between two conflicting genomes in cytoplasmic male sterility?" Trends in plant science 9(12): 568-570.

Truco, M., J. Hu, J. Sadowski and C. Quiros (1996). "Inter-and intra-genomic homology of the Brassica genomes: implications for their origin and evolution." Theoretical and Applied Genetics 93(8): 1225-1233.

U, N. (1935). "Genome analysis in Brassica with special reference to the experimental formation of B. napus and peculiar mode of fertilization." Journal of Japanese Botany 7: 389-452.

Unseld, M., J. R. Marienfeld, P. Brandt and A. Brennicke (1997). "The mitochondrial genome of *Arabidopsis thaliana* contains 57 genes in 366,924 nucleotides." Nature genetics 15: 57-61.

Uyttewaal, M., N. Arnal, M. Quadrado, A. Martin-Canadell, N. Vrielynck, S. Hiard, H. Gherbi, A. Bendahmane, F. Budar and H. Mireau (2008). "Characterization of *Raphanus sativus* pentatricopeptide repeat proteins encoded by the fertility restorer locus for Ogura cytoplasmic male sterility." The Plant Cell 20(12): 3331-3345.

Wan, C., S. Li, L. Wen, J. Kong, K. Wang and Y. Zhu (2007). "Damage of oxidative stress on mitochondria during microspores development in Honglian CMS line of rice." Plant cell reports 26(3): 373-382.

Wang, K., F. Gao, Y. Ji, Y. Liu, Z. Dan, P. Yang, Y. Zhu and S. Li (2013). "ORFH79 impairs mitochondrial function via interaction with a subunit of electron transport chain complex III in Honglian cytoplasmic male sterile rice." New Phytologist 198(2): 408-418.

Wang, Z., Y. Zou, X. Li, Q. Zhang, L. Chen, H. Wu, D. Su, Y. Chen, J. Guo and D. Luo (2006). "Cytoplasmic male sterility of rice with boro II cytoplasm is caused by a cytotoxic peptide and is restored by two related PPR motif genes via distinct modes of mRNA silencing." The Plant Cell 18(3): 676-687.

Williams-Carrier, R., T. Kroeger and A. Barkan (2008). "Sequence-specific binding of a chloroplast pentatricopeptide repeat protein to its native group II intron ligand." RNA 14(9): 1930-1941.

Yagi, Y., S. Hayashi, K. Kobayashi, T. Hirayama and T. Nakamura (2013). "Elucidation of the RNA recognition code for pentatricopeptide repeat proteins involved in organelle RNA editing in plants." PLOS One 8(3): e57286.

Yagi, Y., T. Nakamura and I. Small (2014). "The potential for manipulating RNA with pentatricopeptide repeat proteins." Plant Journal 78(5): 772-782.

Yesodi, V., S. Izhar, D. Gidoni, Y. Tabib and N. Firon (1995). "Involvement of two different *urf-s* related mitochondrial sequences in the molecular evolution of the CMS-specific *S-Pcf* locus in petunia." Molecular and General Genetics 248(5): 540-546.

Yin, P., Q. Li, C. Yan, Y. Liu, J. Liu, F. Yu, Z. Wang, J. Long, J. He and H.-W. Wang (2013). "Structural basis for the modular recognition of single-stranded RNA by PPR proteins." Nature 504(7478): 168-171.

Young, E. G. and M. R. Hanson (1987). "A fused mitochondrial gene associated with cytoplasmic male sterility is developmentally regulated." Cell 50(1): 41-49.

CHAPTER II : MAPPING OF 5' AND 3' TERMINI OF ORF222/NAD5C/ORF101, ATP8 AND NAD4 TRANSCRIPTS IN NAP CMS AND RESTORED *BRASSICA NAPUS*

As introduced in the literature review, CMS is a maternally inherited defect in pollen production specified by novel open reading frames (ORFs) in mitochondrial DNA (mtDNA). In *Brassica napus*, *nap* CMS is associated with the presence of *orf222*, co-transcribed with *nad5*c and unknown open reading frame (*orf101*), in a single transcriptional unit, *orf222*/*nad5*c/*orf101*. The restorer of fertility, *Rfn*, has been associated with decreased levels of *orf222/nad5c/orf101* transcripts. This decrease is associated with the appearance of an additional transcript that is detected with *orf101* probes. Moreover, the *Rfn* allele is genetically-associated with additional RNA cleavage events in the coding region transcripts of *nad4,* which encodes another subunit of complex I, and of $ccmF_{N2}$, which encodes a subunit of a protein complex involved in the biogenesis of *c* type cytochromes. In this chapter, I explore the action of Rfn on these various transcripts. Specific 5' and 3' termini mapping using CR-RT-PCR allowed me to determine the precise Rfn specific processing sites as well as determine if these ends were conserved between mtRNAs similar in sequence (like *orf222* and *atp8*) as well as between *B. napus* and *A. thaliana* mitochondria.

Introduction

Cytoplasmic male sterility (CMS) is a maternally inherited trait that is phenotypically represented by the plant's failure to produce functional pollen (Laser and Lersten 1972). In most cases, a specific rearranged region of the mitochondrial genome is highly correlated with the CMS phenotype. These regions have been found to contain unusual chimeric CMS-related open reading frames (ORFs) that are often co-transcribed with standard mitochondrial genes (Hanson and Bentolila 2004). The novel ORF products are often membrane proteins whose presence is thought to compromise mitochondrial function (Chen and Liu 2014). The specific mechanisms whereby this leads to an abortion of functional pollen formation are, with the exception of a few examples (Balk and Leaver 2001; Sabar, Gagliardi et al. 2003; Luo, Xu et al. 2013), unclear.

A set of nuclear genes, counteracting the functional expression of CMS related ORFs, has been identified as restorers of fertility (*Rf* genes). The *Rf* genes act to reduce the accumulation of CMS-associated RNAs and/or proteins through various mechanisms (reviewed by Chen and Liu 2014). Transcription in plant mitochondrial genomes is a relatively relaxed process, which exhibits little control or modulation; instead post-transcriptional events, such as nuclease processing, editing and splicing represent the major means controlling mitochondrial gene expression (Holec, Lange et al. 2006). Consequently, most restorer genes act at the post-transcriptional level.

In the oilseed rape species, *Brassica napus*, the two endogenous cytoplasms capable of inducing CMS are named napus (*nap*) and polima (*pol*). *nap* CMS is associated with the presence of a chimeric ORF, *orf222*, situated upstream of *nad5*c, the third exon of a subunit of the membrane-embedded arm of the mitochondrial respiratory chain NADH-ubiquinol-oxidoreductase or complex I (L'Homme, Stahl et al. 1997). *orf222* is co-transcribed with *nad5*c and unknown open reading frame (*orf101*), in a single transcriptional unit, *orf222*/*nad5*c/*orf101*. In contrast, *pol* CMS is correlated with the presence of the chimeric *orf224*, that is situated upstream of *atp6*, the gene encoding subunit 6 of mitochondrial F_1-F_0 ATP synthase, complex V (Singh and Brown 1991). Unlike the genes associated with CMS in other plants, *orf222* and *orf224* show a high degree of sequence similarity both at the nucleotide and amino acid levels. The 5' non-coding region and the first 58 codons of both proteins are derived from *atp8,* the plant mitochondrial gene that encodes subunit 8 of the mitochondrial F_1-F_0 ATP synthase (L'Homme, Stahl et al. 1997). The remaining parts of *orf222* and *orf224* sequences do not show strong similarity to any known functional protein, although there is some weak similarity to an open reading frame of unknown function, *orf240a*, present in the *Arabidopsis thaliana* mitochondrial genome. However, there is no evidence that *orf240a* is associated with CMS in Arabidopsis.

The restorers for the *pol* and *nap* systems, *Rfp* and *Rfn*, respectively, each down-regulate the expression of their cognate CMS-associated mitochondrial genes by mediating RNA cleavage events within unique regions of the corresponding transcripts. In *pol* CMS plants, *orf224/atp6* transcripts are predominantly di-cistronic. In flowers of *Rfp* restored

plants, the levels of di-cistronic transcripts decrease, and two new transcripts appear whose 5' termini map within *orf224*. In addition, a slight increase in the level of a 1.1 kb mono-cistronic *atp6* transcript present in low amounts in CMS plants, is observed (Li, Jean et al. 1998). The two new transcripts do not carry an initiator 5' di or triphosphate termini (Menassa, L'Homme et al. 1999) and co-segregate perfectly with fertility restoration. These observations suggest that that *Rfp* acts to trigger endonuclease cleavage events that lead to the degradation of the 5' end of the *orf224* transcript, thereby eliminating the expression of the ORF224 protein and restoring pollen production (Brown 1999).

In a similar manner, levels of *orf222/nad5c/orf101* transcripts are decreased in flowers of *Rfn* restored plants. This decrease is associated with the appearance of additional transcript, spanning only *nad5c* and *orf101*. In angiosperms, mRNA of the mitochondrial *nad5* gene is derived from three transcription units encompassing 5 exons. Expression of *nad5* involves two cis-splicing events, one that joins exon a and b and another that joins d an e, as well as two trans-splicing events that join exon b and d to the independently transcribed exon c (figure 2.1). The cis- and trans-spliced *nad5* introns, like most plant mitochondrial introns, are group II introns, a family of intervening sequences found in organelle and prokaryotic genomes (reviewed in Bonen and Vogel 2001). Typically, group II intron splicing, like splicing in eukaryotic nuclei, occurs through two successive trans-esterification reactions, the first involving cleavage at the 5' splice site with formation of a lariat intron, and the second the formation of the splice junction and release of the intron (Bonen and Vogel 2001). Although some group II introns can splice

auto-catalytically in vitro, all require proteins for in vivo splicing (Padgett, Grabowski et al. 1986). For splicing to occur properly, the intron must fold into a well-conserved structure that juxtaposes the 5' splice and branch sites and brings the two exons into close proximity. That well-conserved structure has a distinctive 3D architecture (Domains 1-6), with its catalytic center formed by Domains 1 and Domain 5 (figure 2.2.a). In trans-spliced introns, the splicing event releases the intron as well as Y shaped structure (branched lariat introns), such as i2R/i2L and i3R/i3L (illustrated in figure 2.1).

The *Rfn* allele is genetically associated with additional RNA cleavage events in the coding region transcripts of *nad4,* which encodes another subunit of complex I, and of $ccmF_{N2}$, which encodes a subunit of a protein complex involved in the biogenesis of *c* type cytochromes. These cleavage events are not observed in plants homozygous for the *Rfp* allele (Singh, Hamel et al. 1996) or for the non-restoring, or universal maintainer genotype *rf* (Li, Jean et al. 1998). Indeed, *Rfn* and *Rfp* appear to represent distinct alleles or haplotypes of a single nuclear locus, and it has not been possible to dissociate these genes or their associated mtRNA cleavage properties via genetic crosses involving the three nuclear and cytoplasmic genotypes (Li, Jean et al. 1998). While the termini arising from *Rfn* action on *nad4* and $ccmF_{N2}$ transcripts have been mapped precisely by primer extension experiments, it has not yet been possible to identify the sites of processing of the *orf222/nad5*c*/orf101* CMS-associated transcript, possibly because of the complex splicing of *nad5* transcripts.

Of particular interest to this study and also contributing to the diversity of 5' and 3' termini for one transcription unit, RNA precursors undergo nuclease processing at their 5' and 3'termini. This suggests that RNA end maturation might be achieved through direct endoribonuclease and/or exoribonucleases activities that would be blocked by stable RNA secondary structures defining mature transcript ends. In support of this view, mapping of mitochondrial mRNA termini in *A. thaliana* by Forner, Weber et al. 2007, highlighted the presence of RNA stem-loop folds at the transcripts termini thought to be involved in the determination of 3' and 5' ends of plant mitochondria RNA. Also thought to be involved in these processes, RNA-binding proteins may serve a similar function in blocking RNA degradation and defining transcripts ends. Some PPR proteins, termed RNA processing factors (RPFs), have been found to bind to their target transcripts and promote the formation of 5' and 3' ends in *A. thaliana* mitochondria (Jonietz, Forner et al. 2010, Hölzle, Jonietz et al. 2011, Jonietz, Forner et al. 2011, Hauler, Jonietz et al. 2013, Arnal, Quadrado et al. 2014).

Because the mode of action of Rfn in transcript modification is unclear, I explored the different processing events that the *orf222/nad5c/orf139* and *nad4* transcripts undergo by mapping their 3' and 5' termini through circular RT-PCR (CR-RT-PCR). Furthermore, in order to determine a potential conservation of 5' end formation between *orf222* and *atp8* given their high sequence homology, analysis of the 5' end and 3' end formation was done on *B. napus atp8* using the same c-RT-PCR method. This technology has been verified and exploited extensively by Binder and colleagues as a tool to map accurately 5' and 3' mitochondrial transcript ends in *A. thaliana* (Forner, Weber et al. 2007, Jonietz,

Forner et al. 2010, Hölzle, Jonietz et al. 2011, Jonietz, Forner et al. 2011, Hauler, Jonietz et al. 2013). It involves covalently linking the 5' and 3' ends of total or mtRNAs via RNA ligase activity and leads to the formation of circular RNA, which serves as template for RT-PCR (figure 2.3). In mitochondria, primary transcript 5' ends carry triphosphates, while processed transcripts have monophosphates at their 5' ends (reviewed in Hammani and Giege 2014). Only the latter are a substrate to RNA ligase. CR-RT-PCR was used to identify the 5' and 3' ends produced by post-transcriptional processing events during the formation of mature transcripts of *nap* CMS and *nap Rfn* plants. By comparing of the transcript termini sites of *A. thaliana*, and *B. napus*, I was able to identify conserved processing events and potential *Rfn* recognition sequences.

Materials and Methods

Plant Growth

Seeds from the two parental lines of *B. napus* used in this present study were provided by Bo Gertsson, Lantmännen Lantbruk, Svalov, Sweden. The 'Karat-*Rfn' nap* restorer line is of Swedish origin. Seeds were plated onto basic Murashige and Skoog (Sigma-Aldridch M5524-10L) medium and kept at 4 °C before germination in a chamber under standard conditions (16-h photoperiod, 22°/16°C day/night temperatures) for a week. Seedlings were transferred in pots and grown to maturity in a greenhouse or in growth under standard conditions.

Isolation of RNA, circularization and RT-PCR

The identification of 5' and 3' ends was carried out by CR-RT-PCR on young flower buds from *nap* CMS and *nap Rfn* plants. Total RNA extraction was first performed on 100 mg of floral tissue using Trizol (Thermofisher 15596-026, Pleasanton, CA, USA). Homogenization of frozen tissue was performed by grinding in a sterile mortar and pestle and adding 1mL of Trizol. The mixture was mixed, incubated 5 minutes at room temperature, then 200 µL of chloroform were added and the samples were mixed thoroughly for 15 seconds. The mixtures were incubated 2 minutes at room temperature and centrifuged at 13,000 x g for 15 minutes at 4°C. The RNA extract in the upper aqueous phase was then precipitated with 0.5 mL isopropanol and incubated 10 minutes at room temperature before centrifuging at 4°C at 12,000g for 10 minutes. The supernatant was rinsed with 1 mL 70% ethanol and dissolved in 100µL of RNAse free water at 55-60°C for 10 min. RNA purification was then carried out with a standard Qiagen RNeasy kit (Cat No./ID 74904) according to the manufacturer's instructions. DNAse treatment (Qiagen CatNo./ID 79254) of RNA was accomplished as suggested by the supplier. After spectroscopic quantification using a nanodrop apparatus (NanoDrop 3300 Fluorospectrometer), 5 µg of total RNA were circularized using T4 RNA ligase (Thermofisher AM2140, Pleasanton, CA, USA) according to the manufacturer's instructions. Reverse transcription was then performed with 7 µL of previsously synthesized circularized RNA using SuperScript II reverse transcriptase (Thermofisher 18064-014, Pleasanton, CA, USA) and random hexamers according to the manufacturer instructions using. 2 µL of the resulting cDNA was amplified via a 35 cycle PCR using

NEB *Taq* polymerase (catalog # M0273S, New England Biolabs, Ipswich, MA, USA) with primers described in table 2.1. Specific products were visualized by running the amplification products on a 1% agarose gel at 120 V for 40 minutes.

Analysis of 3'end and 5'end sites

Products of CR-RT-PCR experiments were excised, purified using QIAquick spin columns (QIAGEN, CatNo./ID 28104) and ligated in the pCR4-TOPO vector using the TOPO TA cloning kit supplied with competent cells (Fisher-Thermoscientific, K 4575-02, Pleasanton, CA, USA) according to the manufacturer's instructions. About 20 clones for each CR-RT-PCR product were randomly selected for purification using PureLink quick plasmid mini-prep kit supplied in the TOPO-TA cloning kit, according to the manufacturer's instructions. After quantification of the plasmid concentration, samples were sent for sequencing at ACGT Corporate facilities (http://acgtcorp.com, Toronto, ON, Canada) using the Applied BioSystems (ABI)/Life Technologies 3730xl capillary electrophoresis DNA sequencer.

Sequence analysis

Sequences delivered by ACGT Corp were first analyzed using the Geneious program (http://www.geneious.com/). *A. thaliana* and *B. napus* mitochondrial genome sequences were downloaded from the NCBI website (http://www.ncbi.nlm.nih.gov/nuccore/

NC_001284.2,http://www.ncbi.nlm.nih.gov/nuccore/AP006444.1). Sequence alignment was performed using the web-based tool FASTA (http://fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi).

Results

nad5 splicing in nap CMS and Rfn genetic backgrounds

In order to allow the simultaneous and unambiguous identification of the 5' and 3' ends of transcript relevant to *Rfn* action, we mapped the *orf222*/*nad5*c/*orf101* RNA processing sites in *nap* CMS and *nap* fertility-restored (*Rfn*) plants using CR-RT-PCR. We employed a variety of primers to allow for the amplification of transcripts to map of 3' and 5' termini and to identify specific *Rfn* plants processing sites.

By using a *nad5*c exon primer together with a primer within *orf101,* we obtained a 350 bp band from both *Rfn* and *nap* CMS samples (figure 2.4.a). Sequencing of the product revealed that its 5' end mapped 5 nucleotides upstream of *nad5c* 5' end to an adenosine corresponding to the bulged residue in domain 6 involved in formation of the i2R/i2L lariat (figure 2.2.b) (Elina and Brown 2010) and a 3' end mapping 137 nucleotides downstream of *orf101*. Because the oligonucleotide used to amplify this product was located in the *nad5c* exon, this terminus likely reflected an artifact of amplification arising the tendency of reverse transcriptase to stall at the lariat junction, leading to a high frequency of reverse transcripts with this 5' end

Using the same primers, we obtained a second band of approximately 600 base pairs from *Rfn* but not *nap* CMS floral RNA (figure 2.4.a). Cloning and sequencing indicated that this product was a mixture of two cDNAs generated from precursor transcripts that were processed prior to splicing or by reverse transcriptase read-through of the intron branch point, These transcripts had 5' termini mapping either 217 or 231 nucleotides (nt) upstream of the 3' splice site of the *nad5*c exon and a 3' end located 137 nucleotides downstream of *orf101*. Equal numbers of clones corresponding to transcripts with each of these 5' termini were obtained, indicating that the transcripts are relatively equal in abundance. Both sites are located downstream of the *orf222* termination codon (figure 2.4.c). Thus, *Rfn* acts at these two closely related sites to generate 5' termini of *nad5c* transcripts that lack the upstream *orf222* sequence. The localization of the processing sites relative to the predicted intron domains upstream of *nad5c* indicates that they would not disturb the formation of a functional intron and allow for the formation of a lariat structure and release of the downstream exon c (figure 2.2.b). Since RNA gel blot experiments show that *orf222* transcripts are significantly reduced in abundance in nuclear restored plants (L'Homme, Stahl et al. 1997), this further suggests that nuclease cleavage downstream of *orf222* leads to destabilization and a reduction in the levels of transcripts capable of expressing the ORF222 protein. All these products had 3' termini that mapped to a site 137 nt downstream of *orf101* (figure 2.4.c). It should be noted that, while open reading frames are frequently found downstream of *nad5*c in plant mitochondria, the proteins that might be potentially encoded by these ORFs are not highly conserved; rather the open reading frames likely reflect constraints on the primary sequence to allow conservation of the secondary and tertiary structural folding necessary

to produce a functional half-intron. On the basis of the modeling studies of *nad5* intron 3 from Oenothera and Arabidopsis (Knoop, Schuster et al. 1991), it is evident that all the domains necessary to form a functional intron are found within the region downstream of *nad5c* included in the mapped transcripts (figure 2.2.b). Clearly, some mechanism exists to stabilize this 3' terminus and allow it to associate with its partner half intron located upstream of exon d.

In order to further characterize 5' and 3' termini of the *nap* CMS conferring transcription unit, CR-RT-PCR amplification of *orf222/nad5c/orf101* was performed using two different primers from within the same region of *orf101* (orf101F1 and orf101F2) and one primer positioned in *orf222* outside of the region with similarity to *atp8* (orf222R). Interestingly, the major products obtained with the two different orf101 primers differed dramatically in size. A major product of just over 100 bp was obtained with orf101F1, while the orf101F2 product was about 500 bp long (figure 2.4.b). Sequence analysis indicated that the orf101F1 product had a 3' end mapping to a site 137 nt downstream of *orf101*, as expected, and a 5' end located 79 nt upstream of the *orf222* initiation codon. In contrast the orf101F2 product had the same 3' terminus but a 5' end mapping to a site 403 nt upstream of the initiation codon (figure 2.4.c). In all other mapping experiments, multiple primers were used to confirm the identified sites of transcript termini. This was the only product for which strikingly distinct termini were selected using different primers. It is possible that regional secondary structure in the single stranded cDNA selectively blocked effective elongation with one of these primers.

Comparison of orf222 and atp8 transcript termini

Sequence similarity between *orf222* and *B. napus atp8* extends from 212 nt upstream to 354 nt downstream of the *orf222* start codon (L'Homme, Stahl et al. 1997). *atp8* sequences as well as its surrounding sequences are also highly similar between the *B. napus* and *A. thaliana* mitochondrial genomes. In order to detect possible conservation of processing sites between *atp8* and *orf222*, the termini of *atp8* transcripts were characterized. Four major CR-RT-PCR products were obtained using primers located within the *atp8* coding sequence, all of them detected in both *nap* CMS and *Rfn* samples (figure 2.5.a). Sequence analysis indicated that they each correspond to four different 5' termini with a unique 3' terminus. The two more abundant CR-RT-PCR products had 5' termini located 172/161 and 467 nt upstream of the *atp8* start codon. The 5' termini of the two less abundant bands mapped 223/231 and 400 nt upstream of *atp8*. The 3' end of each of these transcripts mapped to a site 141 nt downstream of the *atp8* stop codon (figure 2.5.c).

atp8 is positioned upstream of an open reading frame of unknown function designated *orf114* (Handa 2003). *orf114* shows sequence homology with *orf141* of *A. thaliana* (Marienfeld, Unseld et al. 1996) and *orf138* of Ogura radish (Krishnasamy and Makaroff 1993), the latter being associated with male sterility and also co-transcribed with *atp8*. To gain insight on the possible co-transcription of *atp8* and *orf114*, CR-RT-PCR was performed using primers located within *atp8* and *orf114* coding sequences. Two major products were obtained (figure 2.5.b) and sequencing revealed that the 5' ends mapped

172/161 and 467 nt upstream of the *atp8* coding sequence, confirming the previous analysis, and with multiple 3' ends mapping between 30 to 54 nt downstream of the *orf114* stop codon (figure 2.5.c).

In *A. thaliana*, two major *atp8* 5' termini are located 157 and 228/224 nt upstream of the start codon (Forner, Weber et al. 2007) (figure 2.6.a). The 228/224 site, corresponds precisely to the sequences at the 231/223 sites upstream of *B. napus atp8*. Furthermore, the Arabidopsis -157 site mapped to a homologous sequence in close vicinity to the *B. napus* -172/161 5' termini determined by our CR-RT-PCR experiments. In contrast, conservation of major 5' termini between *orf222* and *atp8* is not observed. The most prominent -172/161 *atp8* 5' end termini maps within the region with sequence similarity to *orf222* but no orf222 5' transcript end was detected in that vicinity (figure 2.6.b). Instead, the most prominent *orf222* 5' end found to map within the *atp8*-similar region is localized 79 nucleotides upstream of the *orf222* coding initiation codon. *atp8* transcripts with this terminus are not found in either *B. napus* or *A. thaliana* (figure 2.6.a). I note, however, that faint larger bands are detected in orf222 CR-RT-PCR experiments, which could reflect the presence of additional 5' termini which could correspond to the major 5' ends found in atp8 (-172/161).

A thaliana atp8 transcripts have a unique 3' end 121 nucleotides downstream of the stop codon of its coding sequence (Forner, Weber et al. 2007) (figure 2.6.a). The similarity between *A. thaliana* and *B. napus atp8* 3' termini is not as strong as for the 5' ends (figure 2.6.c). These findings indicate that 5' end processing of mitochondrial *atp8*

transcripts might involve conserved, uncharacterized protein factors whereas 3' end formation might involve different factors or a conserved factor acting on RNA structural elements that are not evident from the sequence alone.

Identification of nad4a transcript termini and Rfn-specific processing site

RNA blotting and primer extension experiments (Singh, Hamel et al. 1996) as well as genetic crosses involving the three nuclear genotypes, *Rfn*, *Rfp* and *rf*, the non-restoring or universal maintainer allele (Li, Jean et al. 1998), indicate that the *Rfn* allele is associated with RNA cleavage in the coding region of *nad4a*, the first exon of the *nad4* gene, which encodes another subunit of Complex I; this cleavage is not observed in plants homozygous for the *Rfp* allele or *rf* (Singh, Hamel et al. 1996, Li, Jean et al. 1998). It was of interest to determine the exact position of the *nad4a Rfn* specific processing site using CR-RT-PCR and to compare the surrounding region to that spanning the Rfn-mediated processing sites located downstream of *orf222*. Sequence similarity between these regions might reflect a recognition site for a factor, possibly the *Rfn* gene product, promoting the processing of both transcripts. With that intent, we used CR-RT-PCR as an adjunct to earlier primer extension experiments to precisely map *nad4* 5' and 3' termini in floral RNA samples from *nap* CMS and *nap Rfn* plants. Using primers from within the *nad4a* coding sequence, two 5' termini were found that were present in both the nap CMS and Rfn restored lines as 490 and 520 base pair bands (figure 2.7.a). These 5' ends of these products map to sites 219 and 231 nt upstream of the *nad4a* coding sequence (figure 2.7.b). A 220 base pair CR-RT-PCR product was detected in *Rfn* samples that is

not found in RNA from nap CMS plants (figure 2.7.a). This corresponds to the product of *Rfn* specific processing site described in (Singh, Hamel et al. 1996). The 5' end of that transcript was mapped within the *nad4a* exon, 205 nt downstream of the start codon (figure 2.7.b).

As shown in figure 2.8, comparison of the *Rfn*-specific 5' ends of *orf222* and *nad4* revealed significant sequence similarity at and around the sites of *Rfn* specific processing. *Rfn* specific processing is also observed in $ccmF_{N2}$ (formerly *ccl1*) transcripts and a more limited sequence similarity between *orf222* and *nad4* termini also extends to the *Rfn* processing site in $ccmF_{N2}$ transcripts determined using primer extension by (Singh, Hamel et al. 1996). Such a finding supports the hypothesis that *Rfn* encodes the factor promoting those events by recognizing a common sequence surrounding the processing sites of the three transcripts.

nad4 transcript termini in A. thaliana and B. napus

In Arabidopsis, *nad4* processing differs between the Columbia and Landsberg ecotypes. A restorer of fertility like PPR protein (Rf-like), RPF1, was characterized as the factor responsible for an additional Columbia-specific *nad4* processing site (Hölzle, Jonietz et al. 2011). It seems likely that the similar factors are involved in 5' end processing of mitochondrial transcripts in other plant species as well. The conservation of processing sites in *A. thaliana* and *B. napus* transcripts would further indicate the conservation of the

protein factors responsible for such events. To further explore this premise, we examined the sequences surrounding the termini of *A. thaliana* and *B. napus nad4* transcripts.

A. thaliana and *B. napus nad4* mitochondrial genes show a high level of homology that extends to about 300 nucleotides upstream of the *nad4a* start codon. The RPF1 specific, *nad4* processing site of *A. thaliana* was shown to map 228 nt upstream of the *nad4a* start codon (Hölzle, Jonietz et al. 2011). The *B. napus nad4a* 5' termini mapped 231 and 219 nucleotides upstream of start codon, which does indeed show some sequence similarity to the corresponding region surrounding the RPF1 processing site in *A. thaliana* (figure 2.9.a and 2.9.b). These termini are located 9 nucleotides upstream and 3 nucleotides downstream respectively of the processing site. Conceivably, if RPF1 homolog protein is present in *B. napus,* it might be responsible for -228 *nad4* processing event.

Like 3' termini of *B. napus* and *A. thaliana atp8*, there is little sequence similarity in the regions downstream of *nad4* in the two species (figure 2.9.c). It is understandable that these species have different *nad4* 3' transcript termini. In *A. thaliana*, the *nad4* 3' terminus maps 30 nt downstream of the *nad4d* stop codon, whereas the *B. napus* 3'termini maps 13 nt downstream of the *nad4d* coding sequence (figure 2.9.a).

Discussion

Implications of orf222 Rfn specific processing in fertility restoration

Rfn specific modification of the *orf222/nad5c/orf101* transcript has been previously shown to genetically co-segregate with restoration of fertility to nap CMS plants (L'Homme, Stahl et al. 1997). I have located the precise site location of *Rfn* specific processing sites in the 3' UTR of *orf222*. In potato, Gagliardi, Perrin et al. 2001, showed that a 3' to 5' exoribonuclease activity was responsible for the preferential degradation of polyadenylated *atp9* mRNAs. Furthermore, that study highlighted the importance of 3' end structures that would stabilize the transcripts, protecting them from degradation. The mapped *Rfn* specific processing sites, although not within *orf222* open reading frame, could destabilize the 3' end of the transcript inducing a polyadenylation signal promoting its degradation, very much like *atp9* in potato. I was unable, however, to detect transcripts with *orf222* sequences by RT-PCR using polydT and forward *orf222* primers. It is possible that polyadenylated *orf222* transcripts are degraded too rapidly to be detected by this approach or that the release of *orf222* from its original *orf222/nad5c/orf101* tri-cistronic structure is enough to destabilize 3' end structure and target 3' to 5' exoribonuclease activity.

orf222 shows homology with *atp8* up to 58 codons into its open reading frame (L'Homme, Stahl et al. 1997). For such chimeric CMS-associated proteins, it is possible that they might compete with the normal subunit during respiratory chain complex assembly, leading to reduced levels of the functional complex. Consistent with this

notion, in sunflower, ORF522, very much like ORF222, contains a segment of the ATP8 gene, subunit of the F_0-F_1 ATP synthase (complex V), and levels of activity of this complex, as assessed by in-gel assays, are reduced in comparison to that of male fertile lines (Sabar, Gagliardi et al. 2003). ORF222 may compromise the assembly of F_0-F_1 ATP synthase through a similar mechanism. Characterization of *nad5c* transcripts presenting a 5' terminus at the buldged adenosine site also confirmed that the presence of the *orf222* sequence within the transcript does not disturb the splicing of *nad5,* which could lead to male sterility (Elina and Brown 2010). Since RNA gel blot experiments show that *orf222* transcripts are significantly reduced in abundance in nuclear restored plants, this further suggests that if nuclease cleavage downstream of *orf222* leads to destabilization of the 3' end of the transcript, a reduction in the levels of transcripts capable of expressing the ORF222 protein would allow for fertility restoration.

Conservation of 5' end processing sites

5' end mapping of of *orf222* and *atp8* transcripts indicated that determination of the processing sites were not conserved between the two RNAs. *orf222* and *atp8* are similar in sequence up to 212 nucleotides upstream of the start codon. One of the dominant *atp8* 5' termini was found to map within that homologous region, 172 nucleotides upstream of the start codon. Forner, Weber et al. 2007, mapped an *A. thaliana atp8* 5' terminus in the corresponding homologous region. Therefore, conservation of this 5' end processing site in the *atp8* homologous region in *orf222* may have been expected. The mechanism of RNA ligation allows only the connection of 5' monophosphate ends and excludes 5'

termini with two or three phosphates, which are expected at primary ends derived from transcription initiation. However, the triphosphate ends are rather unstable and thus primary ends can also carry 5' monophosphate groups, which allows their direct ligation and detection without treatment of the RNA. This was likewise observed previously (Kühn, Weihe et al. 2005). However, none of the *atp8* initiation sites characterized in Kühn, Weihe et al. 2005, corresponds to the -79 *orf222* similar region. A secondary structure or a specific upstream sequence recognized by the processing factors could explain why the terminus is not found in *orf222* transcripts. One of the *orf222* 5' termini maps within the homologous sequence, 79 nucleotides upstream of the start codon. As discussed in Forner, Weber et al. 2007, a large number of 5' ends can be detected in a single transcription unit and often, CR-RT-PCR allows for the detection of the only most abundant termini. It is conceivable that some processing occurs at the -79 site in *atp8* transcripts, but is not sufficiently abundant to be detected by CR-RT-PCR. The presence of fainter, less abundant bands corresponding to minor 5'ends in orf222 is consistent with this premise.

The premise that specific RNA secondary structures or protein target sequences are invoved with formation of transcript termini is further supported by the finding of corresponding 5' ends within homologous sequences od *A. thaliana* and *B. napus nad4* transcripts. Indeed, the 5' termini in *B. napus* mapped 219 and 231 nucleotides upstream of *nad4* coding sequence, which corresponds to the homologous region of RPF1 processing site in *A. thaliana* detected in Hölzle, Jonietz et al. 2011. The two termini were found 9 and 3 nucleotides upstream or downstream respectively of the processing

site. It is possible that a *B. napus* homolog of RPF1 may be involved *nad4* 5' end formation.

3' end processing sites appear not to be well-conserved between B. napus and A. thaliana

Conservation of 3' termini has also been analyzed between *B. napus* and *A. thaliana atp8* and *nad4* transcripts. Although 3' UTR sequences of *nad4* and *atp8* in *B. napus* and *A. thaliana* have some degree of sequence similarity, the transcript termini mapped in these regions are not found in close vicinity of each other between the two species. It has been proposed that mtRNA 3' termini might be generated by endonucleases or exoribonuclease activities that would be blocked by stable RNA secondary structures defining mature transcript ends (reviewed in Binder and Brennicke 2003). Bellaoui, Pelletier et al. 1997, implied that cleavage by RNase P could generate the 3' end of *B. napus* ogura CMS-associated *orf138* mRNA, suggesting direct involvement of endonucleolytic cleavage in 3' end formation. Furthermore, in *A. thaliana*, the mitochondrion-localized PPR protein MTSF1 (mitochondrial stability factor 1) binds to *nad4* mRNA and defines its 3' end by preventing degradation through 3' to 5' exoribonucleases (Haïli, Arnal et al. 2013). It has thus been demonstrated that the 3' terminus of a plant mitochondrial transcripts can result from the binding of protein to a specific sequence in the 3' un-translated region (UTR). RNA binding protein barriers preventing RNA decay and for 3' termini definition is largely accepted and also implies recognition of specific structures or sequences. The reduced homology in the 3' UTRs of

the *A. thaliana* and *B. napus atp8* and *nad4* transcripts might result in a loss of protein recognition sites and explain the lack of 3' termini site conservation.

Rfn recognizes a target sequence present in orf222/nad5c/orf101 and nad4.

Levels of *orf222/nad5c/orf101* tri-cistronic transcripts are reduced in restored versus *nap* CMS plants concomitant with the appearance of a new, restored specific transcript, spanning *nad5c* and *orf101*. The *Rfn* allele is also associated with additional RNA cleavage events in the coding regions of *nad4,* another subunit of complex I, and of $ccmF_{N2}$, which encodes a subunit of a protein complex involved in the biogenesis of *c* type cytochromes. Neither of these are observed in plants lacking *Rfn*, i.e. plants homozygous for the *Rfp* allele or for the non-restoring, or universal maintainer genotype *rf* (Li, Jean et al. 1998). There has not been any indication whether such processing alters *nad4 or* $ccmF_{N2}$ protein levels or impacts mitochondrial metabolism. *nad4,* $ccmF_{N2}$ and *orf222/nad5c/orf101* processing events occur at the 5' end of the transcripts. Various PPR protein factors have been found to promote sequence 5' end processing in *A. thaliana* (Jonietz, Forner et al. 2010, Hölzle, Jonietz et al. 2011, Jonietz, Forner et al. 2011, Hauler, Jonietz et al. 2013). These RNA processing factors are thought to bind to their target RNA sequence in order to allow exonuclease activity at specific 5' end sites. It seems likely that a PPR protein would recognize a target sequence present in the transcripts in order to promote their processing in *B. napus* as well. The finding of sequence homology around the mapped *nad4,* $ccmF_{N2}$ and *orf222/nad5c/orf101* processing sites reinforces such a hypothesis.

My analysis explored 3' and 5' termini formation of mtRNAs in both CMS and *Rfn* genomic backgrounds in order to explore fertility restoration mechanisms as well as general RNA termini formation in *B. napus* mitochondria. Comparison of 5' and 3' end processing sites between *B. napus* and *A. thaliana* determined the extent to which the mechanisms that regulate termini formation are conserved between the two species. The conservation of some sites between the two species suggests that the factors that are responsible for the generation of these termini might be conserved.

FIGURES AND TABLES

	Primer name	Sequence 5' - 3'
orf222/nad5c/ orf101	222 circR	GTTCACCTTGGCTCTCTTCG
	nad5c circR	TGATCCGCTACGGGATTTAC
	101circF1	TTTACCCCATAGGGCCTTCT
	101circF2	CATAGTACCTGCAGCCCCAC
atp8/orf114	ATP8circF	CTTTCTTTTCAGGCTTGACTC
	ATP8circR	AAGGCATAACCAGAATTGT
	114circF	TCCAGTTACAGTGGGGCAAT
nad4	nad4circF	TCGTTCGGATGGGTGTTCACCCC
	nad4circR	AGAAGATCCGCATGCGGAACACGG

Table 2.1. Primers used for the various CR-RT-PCR experiments in order to map the different 3' and 5' ends of mitochondrial transcripts of interest.

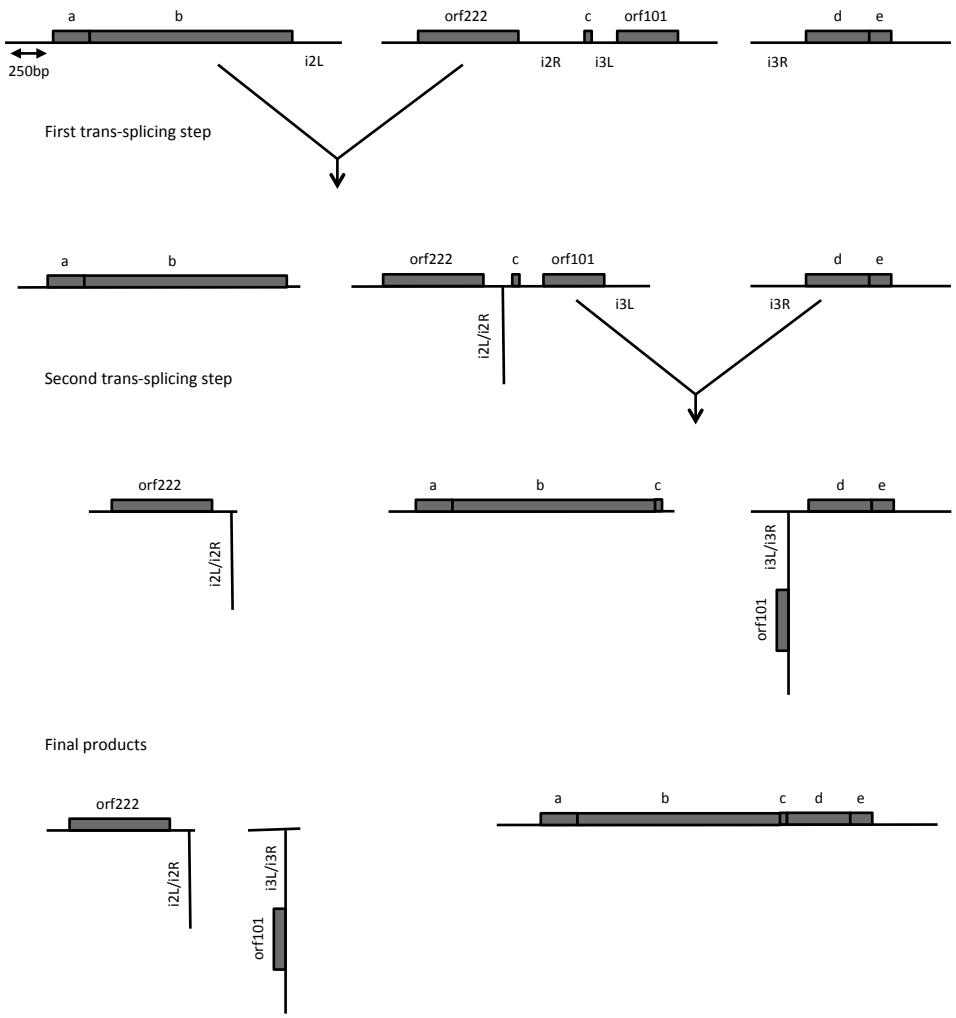


Figure 2.1. Splicing of *nad5* in *nap* CMS *B. napus* mitochondria. Expression of *nad5* involves two cis-splicing events, one that joins exon a and b and another that joins d an e, as well as two trans-splicing events represented here that join exon b and d to the independently transcribed exon c. Trans-splicing events induce the formation of branched lariat i2R/i2L and i3R/i2L. In nap cytoplasm mitochondia, *orf222* is co-transcribed with *nad5c* and part of the i2R/i2L branched lariat.

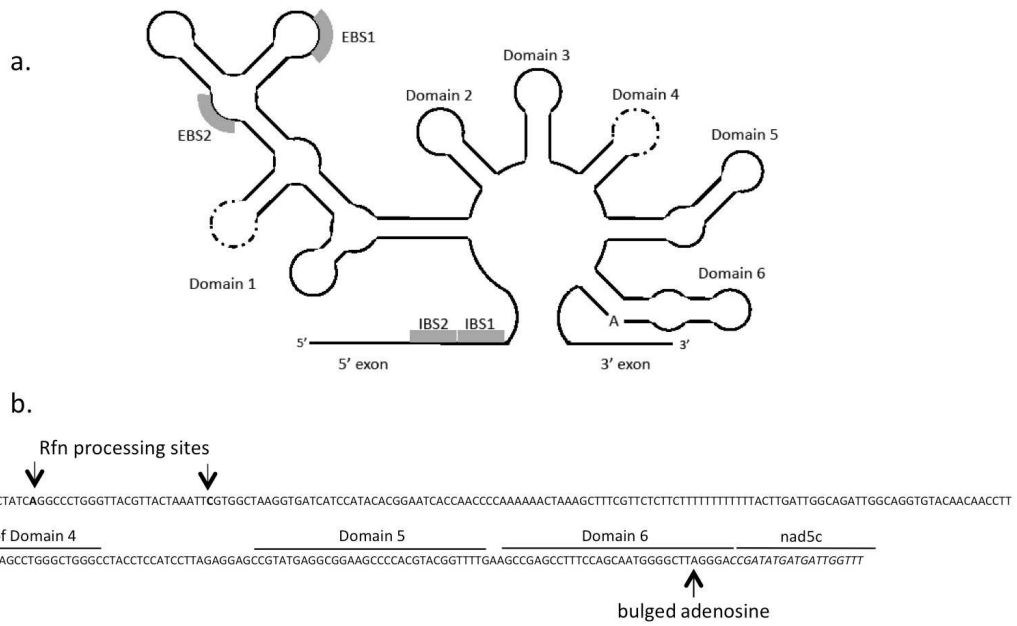


Figure 2.2. (a) Simplified representation of a group II intron as given in Elina and Brown 2010, illustrating the various structural domains. EBS and IBS stand for exon and intron binding sites, respectively. The perforated lines indicate the domain loop sites at which the RNA is disjoined in trans-splicing introns of all flowering plants (Domain 4) and further disjoined in the Oenothera *nad5* i3L intron (Domain 1). Thus, what is a continuous RNA in cis-splicing introns is discontinuous in trans-splicing introns; in the latter case the domain stem is thought to form through base-paring interactions between two distinct RNAs. (b) Location of the different group II intron domains upstream of the *nad5c* exon. Arrows allow for the visualization of the bulged adenosine of Domain 6 as well as the *Rfn* specific processing sites.

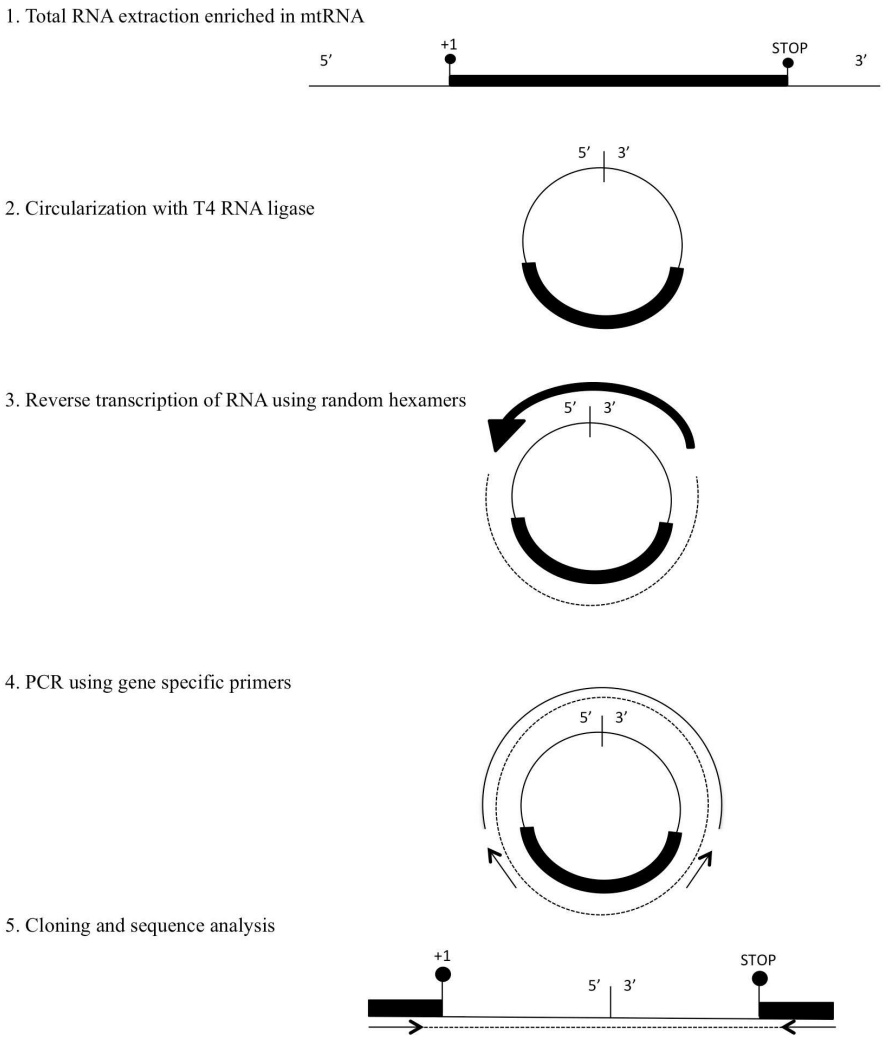


Figure 2.3. Schematic representation of the circularized RT-PCR protocol from Haïli, Arnal et al. 2013. After extraction of total mRNA enriched in mitochondrial RNA from floral buds, circularization is achieved with T4 RNA ligase. cDNAs are then obtained through reverse transcription with random hexamers before amplification of the transcripts of interest with gene specific primers.

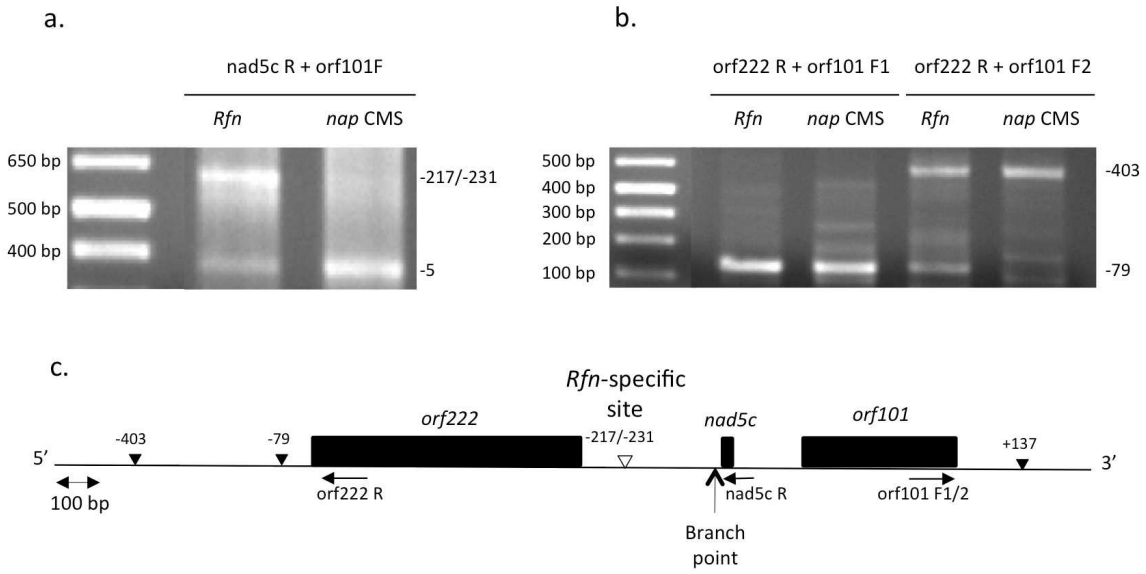


Figure 2.4. Mapping of the 5' and 3' ends of the *orf222/nad5c/orf101* transcripts. (a) CR-RT-PCR products obtained with nad5cR and orf101F primers and ran on a 1% agarose gel. (b) CR-RT-PCR products obtained with orf222R and orf101F1 and orf101F2 primers and ran on a 1% agarose gel. Products with 5' termini positioned between the *orf222* termination codon and the *nad5*c exon are indicated by the number of nucleotides upstrem of *nad5*c. (c) Different ends found by CR-RT-CPR as well as the position of the primer used for the experiment. Filled arrowheads represent termini present in both *nap* CMS and restored *Rfn* samples whereas unfilled arrowheads represent *Rfn* specific processing sites. Primer sequences used can be found in table 2.1.

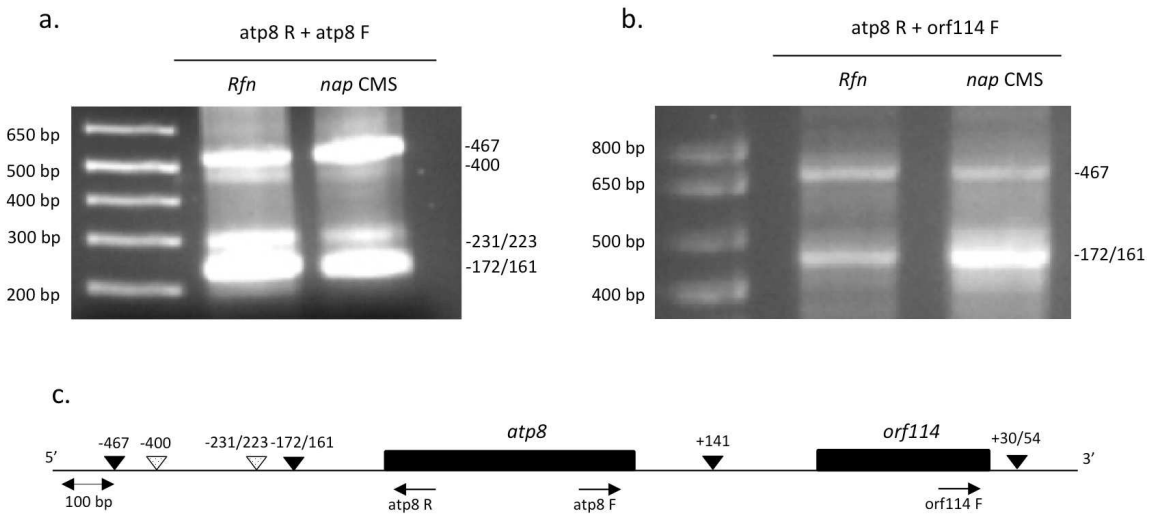


Figure 2.5. Mapping of the 5' and 3' ends of the *B. napus atp8* transcripts. (a) CR-RT-PCR products using atp8R and atp8F primers after running on a 1% agarose gel. (b) CR-RT-PCR products using atp8R and orf114F primers. (c) Different ends found by CR-RT-CPR as well as the emplacement of the primer used for the experiment. Filled arrowheads represent dominant termini whereas unfilled arrowheads represent less abundant processing sites. Primer sequences used can be found in table 2.1.

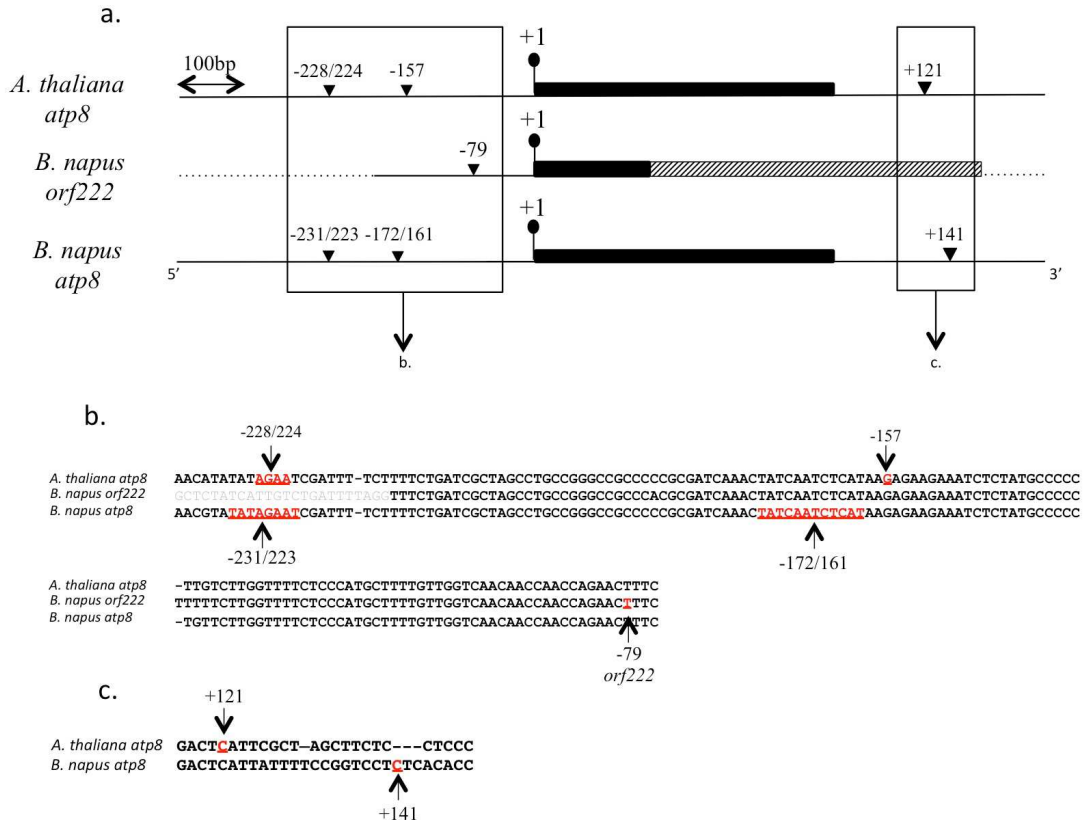
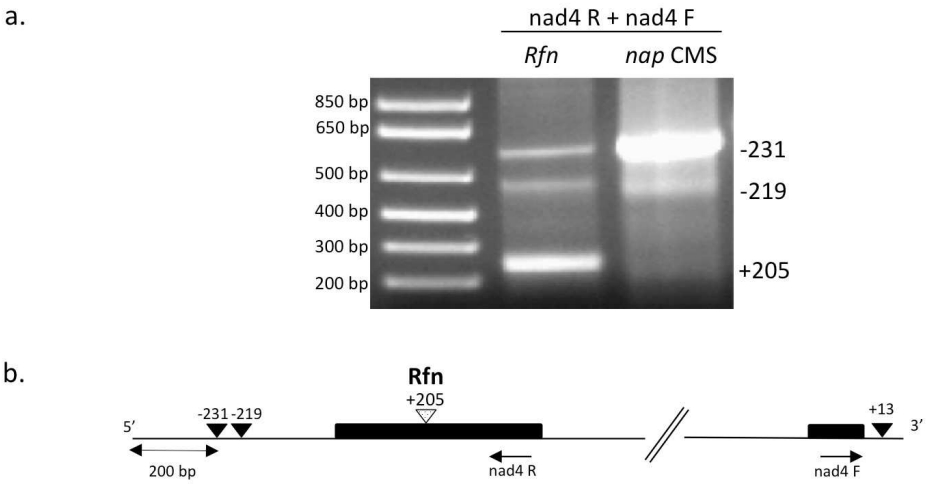


Figure 2.6. Alignment of 5' and 3' termini of *B. napus atp8* and *orf222* as well as *A. thaliana atp8* (as described in Forner, Weber et al. 2007) determined using FASTA alignment program (http://fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi). (a) Representation of sequence homology and location of mapped processing sites. Plain lines and black blocks represent homologous untranslated regions and coding sequences respectively. Dotted lines and striped blocks represent the *orf222* sequences that lack similarity to *atp8*. Arrowheads indicate the location of the different 5' and 3' termini mapped. (b) Nucleotide sequence homology and location of the processing sites mapped



at the 5' ends of the transcripts. (c) Nucleotide sequences homology and location of the processing sites mapped at the 3' ends of the transcripts.

Figure 2.7. Mapping of the 5' and 3' ends of *nad4* transcripts. (a) CR-RT-PCR products after running on a 1% agarose gel. Primers for this experiment were annealed within the *nad4* coding sequences. (b) Different termini found by CR-RT-CPR as well as position of the primers used for the experiment. Filled arrowheads represent termini present in both *nap* CMS and restored *Rfn* samples whereas unfilled arrowheads represent *Rfn* specific processing sites. Primer sequences used can be found in table 2.1.

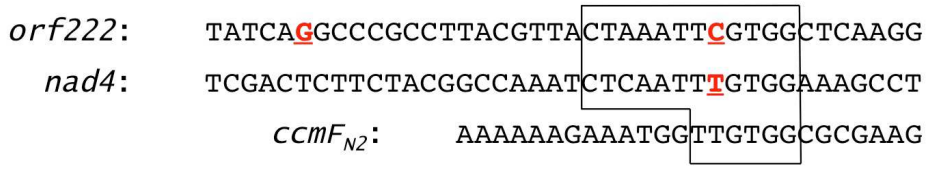


Figure 2.8. Sequence similarity around *orf222*, *nad4, ccmF$_{N2}$ Rfn* specific processing sites (determined using FASTA alignment program (http://fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi). The sequence homology is highlighted by the black square and bold red underlined nucleotides indicate processing sites in the different transcripts as mapped by the CR-RT-PCR experiments. Primer extension from Singh, Hamel et al. 1996, characterized *ccmF$_{N2}$* transcript 5' terminus.

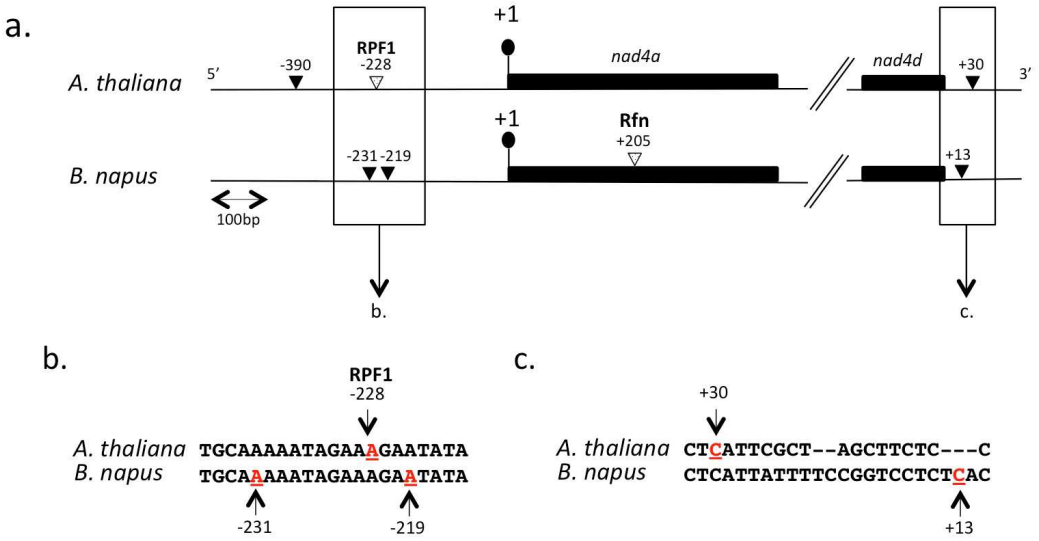


Figure 2.9. Alignment of 5' and 3' termini of *B. napus nad4* with *A. thaliana nad4* (as described in Hölzle, Jonietz et al. 2011) determined using FASTA alignment program (http://fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi). (a) Location of homologous sequences. Arrowheads shows the location of the different 5' and 3' termini mapped, filled arrowheads represent the termini found regardless of nuclear background whereas unfilled arrowheads represent *Rfn* specific or RPF1 specific termini. (b) Nucleotide sequence homology and location of the processing sites mapped at the 5' ends of the transcripts. (c) Nucleotide sequences homology and location of the processing sites mapped at the 3' ends of the transcripts.

References

Arnal, N., M. Quadrado, M. Simon and H. Mireau (2014). "A restorer-of-fertility like pentatricopeptide repeat gene directs ribonucleolytic processing within the coding sequence of *rps3-rpl16* and *orf240a* mitochondrial transcripts in *Arabidopsis thaliana*." The Plant Journal 78(1): 134-145.

Balk, J. and C. J. Leaver (2001). "The PET1-CMS mitochondrial mutation in sunflower is associated with premature programmed cell death and cytochrome c release." The Plant Cell 13(8): 1803-1818.

Bellaoui, M., G. Pelletier and F. Budar (1997). "The steady-state level of mRNA from the Ogura cytoplasmic male sterility locus in Brassica cybrids is determined post-transcriptionally by its 3' region." The EMBO journal 16(16): 5057-5068.

Binder, S. and A. Brennicke (2003). "Gene expression in plant mitochondria: transcriptional and post–transcriptional control." Philosophical Transactions of the Royal Society of London: Biological Sciences 358(1429): 181-189.

Bonen, L. and J. Vogel (2001). "The ins and outs of group II introns." Trends in Genetics 17(6): 322-331.

Brown, G. (1999). "Unique aspects of cytoplasmic male sterility and fertility restoration in *Brassica napus*." Journal of Heredity 90(3): 351-356.

Chen, L. and Y.-G. Liu (2014). "Male sterility and fertility restoration in crops." Annual review of plant biology 65: 579-606.

Elina, H. and G. G. Brown (2010). "Extensive mis-splicing of a bi-partite plant mitochondrial group II intron." Nucleic Acids Research 38(3): 996-1008.

Forner, J., B. Weber, S. Thuss, S. Wildum and S. Binder (2007). "Mapping of mitochondrial mRNA termini in *Arabidopsis thaliana*: t-elements contribute to 5' and 3' end formation." Nucleic acids research 35(11): 3676-3692.

Gagliardi, D., R. Perrin, L. Maréchal-Drouard, J.-M. Grienenberger and C. J. Leaver (2001). "Plant mitochondrial polyadenylated mRNAs are degraded by a 3'-to 5'-exoribonuclease activity, which proceeds unimpeded by stable secondary structures." Journal of Biological Chemistry 276(47): 43541-43547.

Haïli, N., N. Arnal, M. Quadrado, S. Amiar, G. Tcherkez, J. Dahan, P. Briozzo, C. C. des Francs-Small, N. Vrielynck and H. Mireau (2013). "The pentatricopeptide repeat MTSF1 protein stabilizes the *nad4* mRNA in *Arabidopsis mitochondria*." Nucleic acids research 41(13): 6650-6663.

Hammani, K. and P. Giege (2014). "RNA metabolism in plant mitochondria." Trends in Plant Science 19(6): 380-389.

Handa, H. (2003). "The complete nucleotide sequence and RNA editing content of the mitochondrial genome of rapeseed (*Brassica napus* L.): comparative analysis of the mitochondrial genomes of rapeseed and *Arabidopsis thaliana*." Nucleic acids research 31(20): 5907-5916.

Hanson, M. R. and S. Bentolila (2004). "Interactions of mitochondrial and nuclear genes that affect male gametophyte development." The Plant Cell 16(suppl 1): S154-S169.

Hauler, A., C. Jonietz, B. Stoll, K. Stoll, H. P. Braun and S. Binder (2013). "RNA PROCESSING FACTOR 5 is required for efficient 5' cleavage at a processing site conserved in RNAs of three different mitochondrial genes in *Arabidopsis thaliana*." The Plant Journal 74(4): 593-604.

Holec, S., H. Lange, K. Kühn, M. Alioua, T. Börner and D. Gagliardi (2006). "Relaxed transcription in Arabidopsis mitochondria is counterbalanced by RNA stability control mediated by polyadenylation and polynucleotide phosphorylase." Molecular and cellular biology 26(7): 2869-2876.

Hölzle, A., C. Jonietz, O. Törjek, T. Altmann, S. Binder and J. Forner (2011). "A RESTORER OF FERTILITY-like PPR gene is required for 5'-end processing of the *nad4* mRNA in mitochondria of *Arabidopsis thaliana."* The Plant Journal 65(5): 737-744.

Jonietz, C., J. Forner, T. Hildebrandt and S. Binder (2011). "RNA PROCESSING FACTOR3 is crucial for the accumulation of mature *ccmC* transcripts in mitochondria of Arabidopsis accession Columbia." Plant physiology 157(3): 1430-1439.

Jonietz, C., J. Forner, A. Hölzle, S. Thuss and S. Binder (2010). "RNA PROCESSING FACTOR2 is required for 5' end processing of *nad9* and *cox3* mRNAs in mitochondria of *Arabidopsis thaliana*." The Plant Cell 22(2): 443-453.

Knoop, V., W. Schuster, B. Wissinger and A. Brennicke (1991). "Trans splicing integrates an exon of 22 nucleotides into the *nad5* mRNA in higher plant mitochondria." The EMBO journal 10(11): 3483.

Krishnasamy, S. and C. A. Makaroff (1993). "Characterization of the radish mitochondrial *orfB* locus: possible relationship with male sterility in Ogura radish." Current Genetics 24(1): 156-163.

Kühn, K., A. Weihe and T. Börner (2005). "Multiple promoters are a common feature of mitochondrial genes in Arabidopsis." Nucleic Acids Research 33(1): 337-346.

L'Homme, Y., R. J. Stahl, X.-Q. Li, A. Hameed and G. G. Brown (1997). "Brassica nap cytoplasmic male sterility is associated with expression of a mtDNA region containing a chimeric gene similar to the pol CMS-associated *orf224* gene." Current genetics 31(4): 325-335.

Laser, K. D. and N. R. Lersten (1972). "Anatomy and cytology of microsporogenesis in cytoplasmic male sterile angiosperms." The Botanical Review 38(3): 425-454.

Li, X.-Q., M. Jean, B. S. Landry and G. G. Brown (1998). "Restorer genes for different forms of Brassica cytoplasmic male sterility map to a single nuclear locus that modifies transcripts of several mitochondrial genes." Proceedings of the National Academy of Sciences 95(17): 10032-10037.

Luo, D., H. Xu, Z. Liu, J. Guo, H. Li, L. Chen, C. Fang, Q. Zhang, M. Bai and N. Yao (2013). "A detrimental mitochondrial-nuclear interaction causes cytoplasmic male sterility in rice." Nature genetics 45(5): 573-577.

Marienfeld, J., M. Unseld, P. Brandt and A. Brennicke (1996). "Genomic Recombination of the mitochondrial *atp6* gene in *Arabidopsis thaliana* at the protein processing site creates two different presequences." DNA Research 3(5): 287-290.

Menassa, R., Y. L'Homme and G. G. Brown (1999). "Post-transcriptional and developmental regulation of a CMS-associated mitochondrial gene region by a nuclear restorer gene." The Plant Journal 17(5): 491-499.

Padgett, R. A., P. J. Grabowski, M. M. Konarska, S. Seiler and P. A. Sharp (1986). "Splicing of messenger RNA precursors." Annual review of biochemistry 55(1): 1119-1150.

Sabar, M., D. Gagliardi, J. Balk and C. J. Leaver (2003). "ORFB is a subunit of F1FO-ATP synthase: insight into the basis of cytoplasmic male sterility in sunflower." EMBO Reports 4(4): 381-386.

Singh, M., & Brown, G. G. (1991). "Suppression of cytoplasmic male sterility by nuclear genes alters expression of a novel mitochondrial gene region." The Plant Cell, 3(12), 1349-1362.

Singh, M., N. Hamel, R. Menasaa, X.-Q. Li, B. Young, M. Jean, B. S. Landry and G. G. Brown (1996). "Nuclear genes associated with a single Brassica CMS restorer locus influence transcripts of three different mitochondrial gene regions." Genetics 143(1): 505-516.

Stoll, B., K. Stoll, J. Steinhilber, C. Jonietz and S. Binder (2013). "Mitochondrial transcript length polymorphisms are a widespread phenomenon in *Arabidopsis thaliana*." Plant molecular biology 81(3): 221-233.

Vogel, J., T. Hübschmann, T. Börner and W. R. Hess (1997). "Splicing and intron-internal RNA editing of *trnK-matK* transcripts in barley plastids: support for MatK as an essential splice factor." Journal of molecular biology 270(2): 179-187.

CHAPTER III : COMPARATIVE GENOMIC ANALYSIS OF THE COMPOUND BRASSICA NAPUS RF LOCUS

After exploration of the effect of *Rfn* on various mtRNAs in Chapter II, Chapter III concentrates on mapping the *Rfn* locus and determines possible genes of interest that could be restorers of fertility. It has been previously shown that the *Rfp* and *Rfn* genes behave as different alleles or haplotypes of a single locus. Although *Rfp* has been mapped to a segment of the *B. napus* genome co-linear with *A. thaliana* chromosome 1 coordinates 4.28 to 4.40 Mb, no fine mapping has been reported for *Rfn*. I explore here the genetic localization of the *Rfn* gene and the molecular characterization of the genome regions surrounding it. By comparing the different expression profiles of the different candidate genes, I identified a preferred candidate for *Rfn*. In addition, by analyzing the positions and sequence relationships between the *RFL* genes in related regions of Arabidopsis and Brassica genomes, I was able to draw inferences regarding the molecular events through which this gene family has expanded and the *Rf* locus evolved.

Introduction

Cytoplasmic male sterility (CMS) is a widespread, maternally inherited trait of flowering plants that results from the expression of novel genes in the mitochondrial genome. The novel genes are unique for each type of CMS and often are chimeric in structure, consisting of segments of standard mitochondrial genes fused, in frame, to other open reading frames that bear no resemblance to known functional genes (Hanson and Bentolila 2004). The novel gene products associated with CMS are inner membrane proteins whose presence is thought to compromise mitochondrial function, although the specific mechanisms whereby this leads to an abrogation of functional pollen formation, are, with the exception of a few examples (e.g. Balk and Leaver 2001, Sabar, Gagliardi et al. 2003, Luo, Xu et al. 2013), not clear.

The CMS phenotype is often masked by the presence of nuclear restorer genes. These genes are specific to each form of CMS and in general act to down-regulate, at the post-transcriptional level, the expression of the cognate novel CMS-causing mitochondrial genes. Consequently, CMS can often only be revealed through wide intraspecific or interspecific crosses. CMS is also revealed in nature via gynodioecy, populations consisting of a mixture of female (i.e. male sterile) and hermaphroditic plants. Theoretical studies have indicated that maternally inherited male sterility can, under certain circumstances, spread in populations; if the frequency of male sterile individuals then becomes sufficiently high, pollen becomes scarce, a condition favoring the counter-selection of nuclear restorers that suppress the male-sterility (Charlesworth and

Charlesworth 1981, Budar, Touzet et al. 2003, Delph, Touzet et al. 2007). This type of interaction can be viewed in the context of genomic conflict: the selective interests of the uniparentally inherited mitochondrial genome oppose those of the biparentally-inherited nuclear genome (Budar, Touzet et al. 2003). Such genomic conflicts are often characterized as a genetic arms race: the appearance of a new male-sterility mitochondrial gene will drive the appearance of a corresponding restorer gene, analogous to the "gene-for-gene" selection of new host resistance genes in response to new pathogen races (Li, Jean et al. 1998, Touzet and Budar 2004).

A number of nuclear restorers have been isolated in recent years (Bentolila, Alfonso et al. 2002, Brown, Formanová et al. 2003, Desloire, Gherbi et al. 2003, Kazama and Toriyama 2003, Koizuka, Imai et al. 2003, Akagi, Nakamura et al. 2004, Klein, Klein et al. 2005, Wang, Zou et al. 2006, Hu, Wang et al. 2012). Most of these have been shown to encode proteins composed of tandem repeats of a degenerate 35 amino acid sequence, the pentatricopeptide repeat (PPR) (Small and Peeters 2000). Most eukaryotic genomes harbor only a few PPR encoding genes, but in plants this gene family has greatly expanded, and contains plant specific forms with repeats that are longer and shorter than the canonical 35 amino acid motif (Lurin, Andrés et al. 2004). Most plant PPR proteins are targeted to the mitochondria and chloroplasts, and bind to specific RNA substrates, mediating numerous aspects of post-translational gene expression including splicing, nuclease processing, editing and translation (Schmitz-Linneweber and Small 2008). Restorer PPR proteins are largely composed of core repeat motifs defined as the P-type PPR subfamily, and comprise a distinct phylogenetic clade of such proteins, Rf-like

PPRs, whose corresponding coding sequences are designated as *RFL* genes (Fujii, Bond et al. 2011). Comparison of the *RFL* genes within a given species reveals evidence of positive evolutionary selection, consistent with the premise of genomic conflict (Geddy and Brown 2007, Foxe and Wright 2009, Fujii, Bond et al. 2011). Moreover, the position of certain Rf-like PPR genes are not conserved between otherwise closely-related, syntenic chromosomal regions (Geddy and Brown 2007) giving these genes a "nomadic"-like character.

Two native CMS systems, *nap* and *pol*, are known to occur in the oilseed rape (canola) species *Brassica napus*. The causative factor for the "Polima" or *pol* CMS, *orf224,* is a novel CMS-associated mitochondrial gene in which the promoter and first 58 codons of the *atp8* gene are fused to a unique sequence bearing little similarity to other known sequences (Singh and Brown 1991). *nap* CMS is specified by *orf222*, a chimeric gene that is similar but not identical over its entire length to *orf224* (L'Homme, Stahl et al. 1997). The two genes are located in different positions on the mitochondrial genome: *orf224* is situated upstream of the *atp6* gene, while *orf222* is located upstream of *nad5* exon c. The mitochondrial genome of the male fertile *B. napus* cytoplasm, *cam*, is identical to that found in one of the progenitor species, *B. rapa* (formerly *campestris,* see below) and lacks both *orf222* and *orf224*. The restorers for the *pol* and *nap* systems, *Rfp* and *Rfn*, respectively, each down-regulate the expression of their cognate CMS-associated mitochondrial genes by mediating RNA cleavage events within unique regions of the corresponding transcripts (Singh and Brown 1991, L'Homme, Stahl et al. 1997). The *Rfn* allele is also associated with additional RNA cleavage events in the coding

regions of *nad4* and *ccmF*$_{N2}$ (*formerly ccl1-l* (Singh, Hamel et al. 1996, Menassa, El-Rouby et al. 1997)) which are not observed in plants homozygous for the *Rfp* allele or for the non restoring, or universal maintainer genotype *rf* (Brown 1999). Indeed, *Rfn* and *Rfp* represent distinct alleles or haplotypes of a single nuclear locus, and it has not been possible to separate these genes or their associated mtRNA cleavage properties via genetic crosses involving the three nuclear and cytoplasmic genotypes (Li, Jean et al. 1998). Together, these characteristics indicate that two CMS-restorer systems have evolved in the relatively recent evolutionary past, and that *B. napus* CMS systems therefore offer an attractive model for understanding the molecular events underlying the evolution of CMS-restorer gene systems (Brown 1999).

Further progress towards this goal requires the identification and characterization of *Rfp* and *Rfn* genes and an understanding of the relationship between the genomic regions surrounding them. The complexity of the amphidiploid *B. napus* genome has posed a key challenge in the accomplishment of this goal. Following the evolutionary divergence of the Brassica and Arabidopsis lineages about 17 Mya (Town, Cheung et al. 2006, Yang, Kim et al. 2006), polyploidization events gave rise to an ancestral Brassica genome in which most *Arabidopsis thaliana* regions were present as 3 co-linear copies (Parkin, Gulden et al. 2005). Modern diploid Brassica genomes reflect further genomic rearrangement, gene loss and gene duplication and show fragmented co-linearity with Arabidopsis, with each co-linear region being represented, on average, in three copies. *B. napus* is the product of a relatively recent interspecific hybridization between two such Brassica species, *B. rapa* (source of the A genome) and *B. oleracea* (source of the C

genome), which diverged from one another between 2.6 and 4.2 Mya (Cheung, Trick et al. 2009).

We have previously introgressed the *Rfp* gene into a region of the *B. rapa* genome and have fine-mapped the gene to a segment of the *B. napus* genome co-linear with *A. thaliana* chromosome 1 coordinates 4.28 to 4.40 Mb (Formanová, Li et al. 2006, Formanová, Stollar et al. 2010). We report here our genetic localization of the *Rfn* gene and the molecular characterization of the genome regions surrounding it. We identify a preferred candidate for *Rfn*. In addition, by analyzing the positions and sequence relationships between the *RFL* genes in related regions of Arabidopsis and Brassica genomes, we are able to draw inferences regarding the molecular events through which this gene family has expanded and the *Rf* locus evolved.

Materials and Methods

Identification of an Rfn containing BAC

Primers based on the sequence of the cosmid 2840A3, generated from *Rfp* doubled haploid *B. rapa* containing the *Rfp*-linked SNP 12910 (Formanová, Stollar et al. 2010), were used for PCR-based screening (Farrar and Donnison 2007) of a BAC library of *B. napus* doubled haploid line DH12075 (Rfn/Rfn) constructed in the laboratory of Dr. Isobel Parkin, AAFC, Saskatoon, SK, Canada. A single BAC, designated NO202E11, was selected that generated amplification products with over 99% similariy to cosmid 2840A3. This BAC contained the *Rfn*-associated allele of SNP 12910 and was therefore

anchored in the *Rfn* genomic region. The purified BAC was nebulized, subjected to 454 sequencing on a Roche GS-FLX Titanium sequencer and assembled at the McGill University – Génome Québec Innovation Centre.

Annotation of the Rfn containing BAC sequence

The assembled sequence contained 8 non-overlapping contigs. Those contigs were ordered using BLASTN (http://blast.ncbi.nlm.nih.gov/Blast.cgi) against *A. thaliana* and *B. rapa* genome sequences with the assumption that synteny of these different genomes should be mainly conserved. Any missing sequences between the contigs and re-orientation of the sequences were investigated by using PCR to amplify the regions between the ends of each contig. Annotation of the completed sequence was investigated by first determining possible open reading frames using Genscan (http://genes.mit.edu/GENSCAN.html) and Softberry (http://www.softberry.com/berry.phtml?topic=case_study_plants&no_menu=on) online-based tools. Subsequent analysis indicated that Softberry was more accurate at predicting ORFs and became the primary analytical tool for this purpose. Each predicted ORF was then used to find presumptive orthologs in the *B. rapa* and *A. thaliana* genomes using BLAST. This analysis allowed us to functionally categorize most of the genes within the BAC. Details of that analysis are presented in supplementary table 3.2.

Plant growth and fertility scoring

Seeds from the two parental lines of *B. napus* used in this present study were provided by Bo Gertsson, Lantmännen Lantbruk, Svalov, Sweden. The 'Karat-Rfn' nap restorer line is of Swedish origin that presents low levels of both erucic acid and glucosinolates. The population consisted of 318 BC_1 plants derived from an intervarietal cross between single plants from the *nap* CMS line (*rfnrfn; nap*) and nap restorer 'Karat-Rfn ' (*RfnRfn; nap*) lines. Seeds were plated onto basic Murashige and Skoog (Sigma-Aldridch M5524-10L) medium and kept at 4 °C before germination in a chamber under standard light condition (16-h photoperiod, 22°/16°C day/night temperatures) for a week. Seedlings were transferred in pots and grown to maturity in growth chambers or greenhouses under standard conditions (16-h photoperiod, 22°/16°C day/night temperatures).

The fertility was assessed by the careful observation of five flowers per plant at least three times during the flowering period. The overall morphology of the flowers was noted as well as the production of pollen. The flower morphology at the earliest stages of development (2-3 days after flowering) was used as the main criteria to determine the fertility of plants segregating for the restoration of the *nap* CMS. Later in the development, some *nap* CMS flowers can produce pollen and make phenotyping ambiguous. Flowers from a male-fertility restored plant look identical to those of a fertile maintainer plant while flowers from a nap CMS plant have shrunken petals, and the style of the pistil is longer and often bent. CMS anthers of young flowers also have shorter filaments and no pollen or a reduced amount of pollen. The morphological contrast of

young flowers between CMS and normal flowers was sufficient to allow plants carrying a restorer allele to be distinguished from those with only maintainer alleles, even though some CMS plants shed a small amount of pollen, especially in the later stages of development.

Sampling, DNA extraction and SNP analysis

Two leaf disks from young plants were sampled into 96 well plates. DNA was extracted and SNP genotyping was performed on a Sequenome platform by DNA LandMarks, Inc., St-Jean-sur-Richelieau, Quebec, Canada. 45 SNPs from a list of known *B. napus* polymorphisms, were selected on the basis of their position on *B. rapa* chromosome A09 and were screened for presence of polymorphism in *B. napus.* SNPs that were polymorphic between the parents of the cross were used to screen the entire BC1 population.

Marker development

Analysis of the population was conducted in two phases. In the initial phase 293 plants were genotyped and phenotyped to roughly localize the *Rfn* gene. In the second stage, the *B. rapa* genome (Wang, Wang et al. 2011) was used to design ILP (Wang, Zou et al. 2006) and CAPS (Konieczny and Ausubel 1993) markers to more precisely localize the gene. *B. rapa* gene sequences within the mapping interval predefined with the SNP analysis were retrieved from the phytozome database (http://phytozome.

jgi.doe.gov/pz/portal.html). Primers were designed to amplify the first intron of every fifth gene using the primer3 web-based tool (http://bioinfo.ut.ee/primer3-0.4.0/). A 35 cycle PCR was performed following the supplier's instruction (NEB M0273S). Size polymorphisms were detected on 2-3% agarose gels. For the amplification products that did not show length polymorphisms, we attempted to identify CAPS by performing a 2 hour digestion with enzymes HaeIII (recognition site GGCC) and AluI (recognition site AGCT) and running digestion products on a 2-3% agarose gel. The amplification products for ILPs and CAPS were cloned using the TOPO-TA cloning kit (Thermofisher K4575-01, Pleasanton, CA, USA). Information on the primers used and polymorphisms in the parents' genomes can be found in supplementary table 3.1.

Synteny analysis

B. napus, B. rapa, B. oleracea and *A. thaliana* genome fragments corresponding to the mapping interval were extracted from different online databases (https://genomevolution.org/coge/, http://phytozome.jgi.doe.gov/pz/portal.html, https://www.arabidopsis.org/index.jsp). The online tool PIP Maker (http://pipmaker.bx.psu.edu/pipmaker/) was used to assess the conservation of sequence linearity between the various genomes analyzed. *B. rapa* gene sequences were extracted (http://phytozome.jgi.doe.gov/pz/portal.html) in order to investigate homologs in *A. thaliana* (https://www.arabidopsis.org/index.jsp) and *B. napus* (https://genomevolution.org/coge/). Confirmation of co-linearity of relative position was obtained using the "Syntenic Gene" Search tool

(http://brassicadb.org/brad/searchSyntenytPCK.php) of the BRAD online resource. This analysis allowed the exploration of the synteny between the 3 genomes. Details of the data obtained are presented in figure 4.

Identification of candidate genes and comparative genomics

Restorers of fertility inducing post-transcriptional processing, as *Rfn* does, have been characterized so far as part of the PPR P-type protein family. In order to do an exhaustive search of the *Rfn*-region *B. napus* genomic sequence for possible candidate genes, manual annotation of that genomic fragment was necessary. After extraction of the 600Mb Rfn containing genomic sequence (http://brassicadb.org/cgi-bin/gbrowse/B.napus_chromosome/), ORF prediction using Softberry (http://www.softberry.com/berry.phtml?topic=case_study_plants&no_menu=on) with *B. rapa* prediction parameters was performed. For each ORF, a search on *B. rapa* and *A. thaliana* genomes was performed in order to detect a probable homolog. That allowed gathering of information on the possible function of the predicted gene. For each predicted ORF annotated as a PPR protein, PPR domain prediction was performed using TPRpred (http://toolkit.tuebingen.mpg.de/tprpred) and mitochondrial targeting sites predicted using TargetP (http://www.cbs.dtu.dk/services/TargetP/). Information on the syntenic regions from *B. rapa, B. oleracea and A. thaliana* were gathered from Gbrowse of the different genomic databases available for each genome (https://genomevolution.org/coge/, http://phytozome.jgi.doe.gov/pz/portal.html, https://www.arabidopsis.org/index.jsp).

Expression of Rfn candidate genes

Expression of six candidate genes in CMS and restored flower buds were assessed by RT-PCR. Total RNA extraction was first performed on 100 mg of floral tissue using Trizol (Thermofisher 15596-026, Pleasanton, CA, USA). Homogenization of frozen tissue was performed by grinding in a sterile mortar and pestle and adding 1 mL of Trizol. The mixture was mixed and then incubated 5 minutes at room temperature in order to allow rough extraction. 200 µL of chloroform was then added and the samples were mixed thoroughly for 15 seconds before incubation 2 minutes at room temperature and 4˚C centrifugation for 15 minutes at 13 000g. The RNA extract in the upper aqueous phase was then precipitated with 0.5 mL isopropanol and incubated 10 minutes at room temperature before centrifugation at 4˚C at 12,000g for 10 minutes. The pellet was then rinsed with 1 mL 70% ethanol and dissolved in 100µL of RNAse free water at 55-60oC for 10min. RNA purification was carried out with a standard Qiagen RNeasy kit (Cat No./ID 74904) according to the manufacturer's instruction with DNAse treatment as suggested with (CatNo./ID 79254). After nanodrop quantification, reverse transcription was performed using SuperScript II reverse transcriptase according to the manufacturer instructions using 5 µg of total RNA. 2 µL of cDNA was used for a 35 cycle PCR using NEB Taq polymerase (catalog # M0273S, New England Biolabs, Ipswich, MA, USA) according to the supplier's instructions with primers as described in supplementary table 3.1. No amplification of candidate *Bn115* was observed using samples from both the CMS and restored plants, suggesting this gene is not expressed, at least in the tissue

analyzed. Multiple primer sets were tested with the same result. To analyze the presence of introns in the gene sequences the cDNA products were cloned using the TOPO-TA cloning kit (Thermofisher K4575-01) and sequencing was performed by ACGT (http://acgtcorp.com/, Toronto, ON, Canada) with the Applied BioSystems (ABI)/Life Technologies 3730xl capillary electrophoresis DNA sequencer. Sequence analysis of the delivered sequencing graphs was performed using the "geneious" program (http://www.geneious.com/).

Phylogenetic analysis

From the different online databases available (https://genomevolution.org/coge/, http://phytozome.jgi.doe.gov/pz/portal.html, https://www.arabidopsis.org/index.jsp), we selected previously characterized restorer of fertility-like proteins from petunia and radish, and RFL proteins present in the regions syntenic to the *B. napus Rf* locus from *A. thaliana, B. rapa* and *B. oleracea*. For *A. lyrata*, sequences from scaffold 1 (between coordinates 4986000 and 5220000 based on the information provided in Fujii, Bond et al. 2011) were extracted directly from JGI genome browser (http://genome.jgi.doe.gov/cgi-bin/browserLoad?db=Araly1&position=scaffold_1:1-100000) and screened for PPR proteins as described in the identification of *Rfn* candidates. The analysis of the *A. lyrata* sequence was also used for comparative genomics with *A. thaliana*. Protein sequences were appropriately formatted and phylogeny analysis was performed using LIRMM web based tool (http://phylogeny.lirmm.fr/phylo_cgi/advanced.cgi) in advanced mode,

allowing a less stringent G Blocks selection and 400 boot strap in order to obtain a stable tree.

Results

Identification of a BAC clone corresponding to the Rfp/Rfn region

A cosmid clone containing a polymorphism tightly linked to *Rfp*, a SNP in the gene corresponding to the *A. thaliana* At1g12910 gene at chromosome 1 coordinate 4.395 Mb (Formanová, Stollar et al. 2010), was recovered, sequenced and used as a source of probes to screen a *B. napus* bacterial artificial chromosome (BAC) library derived from a line with the *Rfn* genotype at the restorer locus. A single BAC, of approximately 180kb, designated NO202E11, containing a sequence identical to the corresponding region of the cosmid probe was selected and sequenced. Sequence analysis showed that the BAC contained a sequence highly similar to the At1g12910 gene with the *Rfn*-linked allele of the Atg12910-orthologous SNP. Since our previous studies indicated *Rfp* and *Rfn* are different haplotypes of the same genomic region, we considered this BAC to likely be anchored in the *Rfn* region. The BAC sequence was found to be colinear with the region of the *A. thaliana* genome extending from chromosome 1 coordinates 4.27 Mb to 4.47 Mb and with the region of the *B. rapa* genome extending from chromosome A09 coordinates 42.18 Mb to 41.79 Mb. Dot matrix visualization of the synteny between the sequenced BAC and the *B. rapa* and *A. thaliana* genomic regions is shown in figures 3.1.a and 3.1.b, respectively.

High resolution mapping of the Rfn gene

The premise that *Rfp* and *Rfn* are alleles or closely linked alternative haplotypes of the same genetic locus is based on relatively rough genetic mapping and genetic crosses in which the transcript modification activities associated with the two genes have been found to be mutually exclusive. We would therefore expect that higher resolution genetic mapping studies should confirm that *Rfn* does, indeed, co-localize with the region in the selected BAC. We therefore constructed a BC1 mapping population, genotyped individuals with SNPs from the region of *B. napus* chromosome A09 corresponding to *B. rapa* coordinates 39.75 Mb to 43.27 Mb. One of these SNPs, localized at coordinate 42.13 Mb, fell within the BAC sequence. Specific information on the SNPs that were polymorphic between the mapping parents of the cross is provided in supplementary table 3.1.

The *nap* CMS phenotype is leaky in a temperature-dependent, cultivar-specific manner (Fan, Stefansson et al. 1986), a challenge for the precise mapping of *Rfn.* The choice of the *Rfn* parent for the mapping cross is therefore critical. Based on previous work, we knew that utilizing the cultivar "Karat" could provide a BC1 mapping population that would allow me to satisfactorily distinguish between CMS and fertility restored progeny (Li, Jean et al. 1998). To derive my BC1 population, two F1 individuals generated by pollinating CMS plants (*rf/rf* [*nap*]) with Karat (*Rfn/Rfn* [*nap*]) were crossed back as males to the CMS parent. Of 293 individual BC1 plants, 146 were scored as fertile (*Rfn/rf* [*nap*]) and 147 as male sterile (*rf/rf* [*nap*]). The floral phenotypes of the parental and

fertile and sterile progeny are illustrated in figure 3.2.a. Genotyping of the population, as illustrated in figure 3.2.b, allowed us to map the *Rfn* gene to the region of *B. napus* chromosome A09 containing the selected BAC, confirming that *Rfp* and *Rfn* were closely linked alleles. The single marker positioned within the BAC was perfectly linked to *Rfn*. Because both *Rfp* and *Rfn* are located within this chromosomal region we henceforth refer to it as the *B. napus Rf* locus.

To more precisely localize *Rfn,* we identified additional polymorphic molecular markers mapping within the region delimited by the SNPs. Primers designed to amplify genomic regions extending across introns in genes located between the SNPs most proximal to *Rfn* identified two intron length polymorphisms (ILPs (Wang, Zou et al. 2006)), amplification products that differed in size between the two parents of the mapping population. When no noticeable amplicon length difference was evident, the products were further subjected to restriction cleavage to reveal cleaved amplified polymorphisms (CAPS (Konieczny and Ausubel 1993)). This strategy allowed four additional polymorphic markers anchored in the targeted region to be identified. These markers were then used to genotype individuals in which recombination had occurred between the closest flanking SNPs. This strategy allowed us to delimit the *Rfn* containing region to the segment of corresponding *B. rapa* chromosome A09 coordinates 41.68 Mb to 42.25 Mb.

Characteristics of the B. napus Rf locus

Several interesting features were revealed through dot matrix visualization of the synteny between the sequenced BAC and the *B. rapa* and *A. thaliana* genomic regions (figures 3.1.a and 3.1.b, respectively). Regions at each end of the BAC, roughly located at positions 10-30 and 120-140 kb, showed similarity to sets of short repeated sequences, shown as boxed regions on the figures. Detailed comparative annotation of the genes in the BAC with the corresponding portions of the *B. rapa* and *A thaliana* (supplementary table 3.2.) indicated that the repeated sequences corresponded to sets of directly repeated genes and/or pseudogenes encoding thionin (PR-13) proteins (positions 10-13 kb) in one case and Cytochrome P_{450} Cyp2 proteins (positions 120-140 kb) in the other. Comparative annotation with the more recently released *B. napus* cv. "Darmor" chromosome BnA09 (Chalhoub, Denoeud et al. 2014) indicated that a sequence inversion has taken place in *B. napus* by which the sequence extending from *B. rapa* chromosome A09 coordinates 41.80 to 42.04 Mb is inverted (supplementary table 3.3.). The inverted region is flanked by sequence spans encoding Cytochrome P_{450} Cyp2 and F-box domain encoding genes. This rearrangement appears as a gap in synteny from BAC coordinates 120 to 140 kb in both the *B. rapa* and *A, thaliana* plots. At around coordinate 160 kb, near the very end of the BAC, sequence similarity is observed between the BAC and the Cytochrome P_{450} Cyp2 genes, but in reverse orientation, as expected in the case of an inversion. Key features of the differences between the BAC, *B. rapa* and *B. napus* cv. "Darmor" genomes are illustrated in figure 3.3.

Rf-like PPR genes in the B. napus Rf locus

Another striking feature of the dot matrix comparisons was the presence of sequences located near BAC coordiantes 34, 38, 52 and 118 kb (illustrated by double headed arrows in figures 3.1.a and 3.1.b) that mapped to four corresponding sites in *B. rapa* chromosome A09. In the Arabidopsis genome, only three of the four sites were located at corresponding positions; no Brassica sequence was found to correspond to the site located near the 4.295 Mb coordinate on Arabidopsis chromosome 1. Inspection of the Brassica repeat sequences indicated that they all corresponded to regions encoding highly similar *RFL* PPR genes predicted to be targeted to the mitochondria, which we designated as *PPR1-4*.

Because the genetically defined limits of the *Rfn* region extended beyond the boundaries of the BAC, we searched the corresponding region of the *B. rapa* chromosome 9 for additional *RFL* PPRs that could serve as candidates for *Rfn*. We were able to identify two more such genes, one, B.rara.I05036, located between coordinates 41.777 and 41.776 Mb, and the other, B.rara.I05115, between coordinates 42.211 and 42.213 Mb. Orthologous PPR genes at corresponding locations were found to be present on *B. napus* "Darmor" chromosome BnA09 and are designated *Bn036* (coordinates 31.334-31.341 Mb) and *Bn115* (31.797-31.806 Mb). Both of these genes encode products predicted to be targeted to the mitochondrion. Notably, *Bn036* is located roughly 100 kb from the flanking marker 4.4BB but 300 kb from the *RFL* genes in the BAC. A preliminary phylogenetic analysis (supplementary figure 3.1.) indicated that five of the six genes

clustered with three *A. thaliana RFL* genes located on the long arm of chromosome 1, *AtRFL2* (At1g12300), *AtRFL3* (At1g12620) and At1g12775 (re-annotated At1g12770 or *AtRFL25*). The sixth gene *PPR2*, clustered within a neighboring branch with its ortholog *AtRFL4* (At1g12700). *PPR3* and *AtRFL25*, like *PPR2* and *AtRFL4,* are located at matching positions in their corresponding chromosomes.

A comparison of the relative positions of the genes in the *A. thaliana* and *B. rapa / B. napus* A genomes is presented in figure 3.4. The phylogenetic analysis indicated that all of the Brassica PPR genes represented Rf-like PPRs, as did three of the four Arabidopsis genes; a non-*RFL A. thaliana* PPR gene, At1g13030, is found at a site close to but not precisely matching the Brassica *RFL PPR4* gene at coordinate 118 kb in the BAC. One of the genes, At1g12700 (*AtRFL4* (Hölzle, Jonietz et al. 2011)) located at a matching position in both genomes, is known to encode a mtRNA processing factor, RPF1, which confers nuclease cleavage events on *nad4* transcripts, which are also a target of the Brassica *Rfn* gene. These observations revealed six *RFL B. napus* PPR genes that could serve as candidates for *Rfn*. One of the genes, *PPR4*, has been previously proposed as a candidate for *Rfp* on the basis of fine mapping data (Liu, Liu et al. 2012). None of the Brassica PPR genes was predicted to contain an intron, an observation confirmed by RT-PCR analysis of floral transcripts (see below).

Expression of the Rf-region RFL genes in nap CMS and fertility restored plants

The observation that *rf-PPR592,* the non-restoring allele of the Petunia restorer, *Rf-PPR592,* is not expressed but is otherwise similar to the restorer (Bentolila, Alfonso et al. 2002) suggested that expression differences among different candidate PPR genes could be used as a tool to prioritize candidates for further analysis of restoration function. We used RT-PCR to examine the expression of the six candidate *RFL* genes located within the Brassica *Rf*-locus. As shown in figure 3.5, we did not detect expression of *Bn115* in floral buds of either CMS or nuclear restored plants. Of the remaining five *B. napus* Rf-region *RFL* genes, expression of one, *PPR4,* was detected in the buds of nuclear fertility restored but not CMS plants; *PPR4*, is thus seen as a strong candidate for *Rfn*. *PPR1* and *PPR3* both also show higher levels of expression in restored than in CMS flowers.

The sequences of the RT-PCR products, as shown in supplementary figure 3.2, provided further information relevant to the structures of the genes and their possible roles in nuclear fertility restoration. The sequences of the transcripts were co-linear with the corresponding genomic DNA sequence, indicating that, as predicted by gene prediction software, these genes lacked introns. Interestingly, a termination codon was found at nucleotide position 1119 in the coding sequence of the RT-PCR products of *PPR2* from both the CMS and nuclear restored lines. This termination codon was not detected in genomic sequences of either *B. rapa* or the BAC NO202E11, and resulted from a one nucleotide insertion in the CMS and restorer sequences prior to the termination codon, followed by a two nucleotides insertion in the same sequences.

The *B. napus* genome is derived from a recent interspecific hybridization event between the C genome species *B. oleracea* and the A genome of *B. rapa,* two species which diverged in descent approximately 3.7 Mya. Because *RFL* PPR genes are known to be variable in chromosomal position between closely related genomes (Geddy and Brown 2007, Fujii, Bond et al. 2011), it was of interest to determine the position of the *Rfn* candidates and close paralogs in the *B. oleracea* C genome. To accomplish this, we first identified close homologs of the different *B. napus* Rf region PPR genes on the *B. oleracea* genome using the blastn resource of the Brassica database (BRAD (Cheng, Liu et al. 2011)). We found a cluster of highly similar sequences between chromosome 8 coordinates 38.54 and 38.75 Mb, a region over which synteny was maintained with the corresponding regions *of A. thaliana* and *B. rapa*. The annotation of the region indicated that the homolgous sequences corresponded to 10 highly related *RFL*-PPR genes. Eight of these genes lacked predicted introns, whereas two, Bol31351 and Bol31388, were each predicted to contain a single intron.

The relative position of PPR genes in the *A. thaliana*, *B. rapa/napus* (A) and *B. oleracea* (C) genomes in the region over which synteny is conserved among the three genomes is illustrated in figure 3.4. Seven of the ten C genome *B. oleracea* PPR genes are located a position corresponding their location in the A genome. In three cases, involving Bol31370/Bol31371, Bol31387/Bol31388 and Bol313407/Bol313408, a tandem pair of *B. oleracea* PPR genes is found at sites occupied by only a single PPR gene in the A

genome. In one case a single *B. oleracea* gene, Bol31384, was found at a site containing two adjacent PPRs in the A genome. These observations indicated that the relative locations of the majority of Rf region PPR genes have not changed since the A/C genome divergence.

Positional variation of RFL genes in the Rf-orthologous regions of two Arabidopsis genomes

Because we observed some conservation of location between some *A. thaliana RFL* genes and Brassica *RFL* genes in the Rf-region, it was of interest to determine to what extent this positional conservation could also be observed between the *A. thaliana* and *A. lyrata* genomes, which diverged from each other between 4 and 5 Mya. Fujii, Bond et al. 2011, observed the *RFL* genes in the *A. thaliana* region orthologous to the Brassica Rf-region (figure 3.1.a.) fell into *RFL* subgroup 1, most of which are located between the 4.18 and 4.33 Mb coordinates of chromosome 1. *A. lyrata* subgroup 1 genes similarly cluster within a 221 kb segment of the scaffold 1 genome assembly unit.

Dot matrix visualization of the similarities between these two Arabidopsis genomic regions revealed several interesting features of the positional relationships among these genes (figure 3.6.) At the position corresponding to *AtRFL2* (At1g12300), two highly similar genes (indicated by the double-headed arrows in figure 3.6.a.), designated in figure 3.6.b. as *AlyRFL1* and *AlyRFL2,* are found at the corresponding site in the *A. lyrata* genome. The 5' and 3' non-coding regions surrounding these genes are similar to one

another, indicating that the two *lyrata* genes arose from a tandem duplication of a region encoding an *AtRFL2*-like gene. At the position corresponding to *AtRFL3* (At1g12620), a tandem triplication of a similar gene was found at two different positions in the *A. lyrata* genome. This arrangement arose from the duplication of an approximately 14 kb region spanning the triplication and extending in the direction matching the centromere proximal side of *A. thaliana* chromosome 1 (figure 3.6.a.). These six *RFL* genes, which we designate as *AlyRFL3-8*, are found within a duplication spanning the region around *AtRFL3*. Another tandem duplication with sequence similarity to the *AlyRFL1/AlyRFL2* pair is found at a site corresponding to *AtRFL4* (At1g12700). These observations suggest segmental duplication is the primary mechanism behind the proliferation of this family of *RFL* genes in this region the *A. lyrata* genome.

Phylogenetic relationships among Brassica and Arabidopsis RFL proteins

We constructed a maximum-likelihood phylogeny to examine how the positional relationships among the various *RFL* genes reflected the sequence relatedness of the various encoded proteins. As shown in figure 3.7., all of the RFL proteins encoded in the Rf-syntenic regions of the different genomes formed a single monophyletic cluster encompassing the Arabidopsis subgroup 1 RFL proteins (Fujii, Bond et al. 2011), and excluding radish Rfo/PPRA, their closest Arabidopsis homolog, AtRFL18, as well as the Petunia Rf-PPR592 restorer protein and its non-restoring homolog. Most of the Brassica proteins fell into a distinct cluster most closely related to the Arabidopsis branch containing *AtRFL1-3*. The exception were those proteins encoded by genes located at the

same position as *AtRFL4* (At1g12700), PPR2 and Brara.I05097; these formed a distinct phylogenetic cluster, suggesting that these Arabidopsis and Brassica proteins have descended from a common ancestor located at the same position in the ancestral genome.

Within the major Brassica clade, proteins in a common genomic location generally clustered together in the tree. The major exception concerned the *B. oleracea* genes Bol03187 and Bol03188. These proteins formed a cluster distinct from that of their positional counterparts, the *B. napus* PPR3 proteins, which clustered with PPR4 and its positional *B. oleracea* counterpart, Bol031408. An interesting situation is observed among the proteins encoded by Brara.I05115 and the genes at the corresponding positions in the *B. oleracea* and *B. napus* genomes, *B. oleracea* Bol03170 and Bol03171 and *Bn* 115. Although the Rf-region is derived from a *B. rapa* (A genome) ancestor, the *B. napus* and *B. rapa* 115 proteins each group with a different *B. oleracea* protein. Conceivably, orthologs of both genes were present in the common ancestor of the sequenced varieties of *B. napus* and *B. oleracea*, and different orthologs were lost during the subsequent evolution of the two A genome forms.

My manual annotation of the portion of *A. lyrata* genome enriched in subgroup 1 *RFL* genes (Fujii, Bond et al. 2011) led to the identification of 10 distinct genes. *AlyRFL1-10*, one of which (*AlyRFL10*) had too few PPR domains to merit inclusion in the phylogenetic tree. The two *A. lyrata* genes located at the position of *AtRFL2*/At1g12300 formed a distinct clade most closely related to the group of *A. thaliana RFL* genes encompassing *AtRFL2* as well as the closely related genes At*RFL3*/At1g12620 and

At1g12775. The position of these two *lyrata* genes in the tree is consistent with a model in which they arose through a tandem duplication of an *AtRFL2*-like gene in an ancestral Arabidopsis genome, as proposed above. Similarly, *AlyRFL3*, *AlyRFL4* and *AlyRFL5* formed a monophyletic group with *AlyRFL6* and *AlyRFL8*, as would be predicted if there were a gene triplication followed by duplication of the three gene set. The exception to this model concerns *AlyRFL7*, which was predicted to cluster with *AlyRFL6* and *AlyRFL8* but instead groups with *AlyRFL9*. Interestingly, although the coding sequence of this *AlyRFL7* is more closely related to *AlyRFL9* than to *AlyRFL6/8*, more similarity is observed in the regions upstream and downstream of the gene to the corresponding sequences in the *AlyRFL3-5* region. Conceivably, *AlyRFL7* underwent a gene conversion event involving an *AlyRFL9* like sequence following duplication of the three genes region.

DISCUSSION

RFL genes are dispersed in the Brassica Rf-region

A characteristic of genomic loci known to encode nuclear restorer proteins is the occurrence of multiple tandemly repeated related *RFL* genes (Akagi, Nakamura et al. 2004, Bentolila, Alfonso et al. 2002, Brown, Formanová et al. 2003, Desloire, Gherbi et al. 2003, Koizuka, Imai et al. 2003). Of these, only the rice locus encodes multiple Rf genes that suppress different forms of CMS. In contrast to other *Rf* loci, only two *RFL* genes within the *B. napus Rf* locus, *PPR1* and *PPR2*, are found adjacent to one another. Of these two, *PPR2* forms a phylogenetic group with *AtRFL4* that is distinct from that

formed by the other Arabidopsis and Brassica *RFL* genes and thus represents a different line of evolutionary descent. Interestingly, the sequences flanking *AtRFL4* and Brassica *PPR2* are similar to one another, indicating that the genes have both descended from an ancestor located in the same genomic position.

Retrotransposition has played a major role in the proliferation of Brassica Rf-region RFL genes

Comparison of the *B. napus Rf* locus to orthologous segments of related genomes has provided insight into the mechanisms by which this subgroup of *RFL* genes has expanded during the evolution of of Brassicacea genomes. With the exception of Brassica *PPR2*, these genes form a monophyletic cluster in the tree of figure 3.7., indicating descent from a common ancestor. None of the *B. napus/rapa RFL* genes share common 5' or 3' non-coding regions, suggesting that the expansion of this gene family occurred primarily through mechanisms involving retrotransposition. The observation that none of these genes contains an intron is consistent with this view.

Viewing the sequence relatedness among the different encoded proteins, as assessed through the phylogenetic tree, in the context of the positions of the genes on their respective genomes, provides additional insight into the mechanisms through which these genes proliferated. For example, *B. oleracea* Bol031408 and *B. napus/rapa PPR4*, cluster together on the tree, are located in corresponding positions on their genomes and have similar flanking sequences indicating these genes have descended from an ancestral gene

present at that location in the last common ancestor of the A and C genomes. The coding sequences of *B. napus/rapa PPR3* form the closest neighboring phylogenetic cluster to these genes, but are present at a different genomic position, aligning with Bol03187/03188 and At1g12775. However, the sequences flanking *PPR3, PPR4,* Bol03188 and At1g12775 are all dissimilar, indicating that these genes have proliferated in their respective genomes through retrotransposition. Bol03187 and Bol03188 have similar flanking sequences, indicating a segmental duplication, but their coding sequences cluster with a Bol03184 and *PPR1*, suggesting that this duplication took place following retrotransposition of a *PPR1*-like gene.

RFL gene proliferation in A. thaliana and A. lyrata

Apart from *AtRFL4*, discussed above, the Arabidopsis *RFL* genes found in genomic regions orthologous to the Brassica *Rf* locus form a monophyletic group, suggesting these genes have descended from ancestor(s) distinct from those that gave rise to the Brassica *Rf* locus *RFL* genes. Like the majority of the Brassica Rf-region *RFL* genes, the three *A. thaliana* proteins, AtRFL2, AtRFL3 and At1g12775, form a monophyletic cluster, are all located at different genomic positions and lack similarity in their flanking regions, indicating that their proliferation took place through retrotransposition events.

As discussed above, segmental duplication appears to have been the primary mechanism driving the duplication of this gene family in the *A. lyrata* lineage. Consistent with this model, three of the *A. lyrata* genes, *AlyRFL3-5*, predicted to be the products of tandem

gene triplication, form a distinct cluster. Of the other three genes group predicted to arise from tandem duplication, *AlyRFL6-8*, *AlyRFL6* and *AlyRFL8* form a neighbouring cluster in the tree, as would be expected if they arose from the duplication of the *AlyRFL3-5* region. *AlyRFL7*, however, clusters with *AlyRFL9*, which, with the very short *AlyRFL10* (not included in the phylogeny), was predicted to have been generated by a distinct duplication event. Interestingly, the sequences surrounding *AlyRFL7* are similar to those flanking *AlyRFL6* and *AlyRFL8,* as would be predicted from the segmental duplication model. It seems likely that following the duplication of the three gene region, retrotransposition and homologous recombination of an *AlyRFL9*-like gene led to the replacement of the coding sequence of the original gene located *AlyRFL7* site with a sequence resembling *AlyRFL9*.

Prioritization of candidates for the Rfn gene

A key goal of this undertaking has been to identify candidates for the *Rfn* gene and to prioritize these candidates prior to proceeding with experiments aimed at rescuing the CMS and thereby conclusively identifying the gene. Since most restorer genes have been found to be *Rf-like PPR* genes and since the *Rfn* locus delimited in our mapping studies is enriched in *RFL* genes, we deem it most likely that one of these genes functions as *Rfn*. Several other types of proteins have been suggested to play a role in fertility restoration, most notably Glycine-rich proteins, or GRPs (Itabashi, Iwata et al. 2011, Hu, Wang et al. 2012). We did not, however, detect sequences capable of encoding such proteins in the

genetically delimited Brassica *Rf* locus. Likely candidates for *Rfn* are therefore limited to the six *RFL* genes in this region.

It is logical that a restorer gene be expressed in floral tissues and because we were not able to obtain RT-PCR products for *Bn115* from floral RNA samples of either CMS or fertility restored plants, it seems unlikely to function as *Rfn*. *Bn036* is situated 100 kb from the flanking marker 4.4 BB but 300 kb from the next closest *RFL* gene and from the single marker that maps within the sequenced BAC and shows complete linkage to *Rfn*. It seems unlikely that the five-recombination events we observed between 4.4BB and *Rfn* in the mapping experiments all occurred within the 100 kb interval between this marker and *Bn036*. Bn036 is therefore judged not to be a strong candidate for *Rfn*. Of the genes within the selected BAC, I found through the analysis of transcripts of these genes that *PPR2* contains a premature termination codon in the line used as the *Rfn* parent of the mapping cross. Thus *PPR2* is unlikely to function as *Rfn*.

Of the three remaining *RFL* genes, one, *PPR4*, is expressed in the restorer but not CMS parent. While it remains possible that *PPR4* does not have a restorer function, it is interesting to note that this gene has also been proposed as a candidate for the *Rfp* gene (Liu et al. 2012) and its absence of expression in the CMS line would explain why lines that can maintain the *nap* CMS can also serve as maintainers for *pol* CMS (Brown 1999), i.e. they can serve as "universal" maintainers in *B. napus*. A unique and novel situation would arise if, indeed, different alleles of *PPR4* would produce two slightly different proteins that can function as distinct restorers of different forms of CMS.

At the outset of this study, my goals were to map the *Rfn* region with sufficient resolution that we could further clarify the relationship of this gene with *Rfp*, the restorer for the other native CMS system in *B. napus*, and to identify a limited number of candidate genes that could then be tested for their capacity to rescue the *nap* CMS trait through transgenic complementation. It became evident that the Brassica *Rf* locus did not contain a group of tandemly repeated *RFL* genes, as do other characterized nuclear restorer loci, but that it did correspond to one of the *A. thaliana* and *A. lyrata* regions identified by Fujii, Bond et al. 2011, that is enriched in a particular subgroup of *RFL* genes. We then sought a deeper understanding of the sequence relatedness of between *RFL* proteins encoded in the Brassica *Rf* locus and the Arabidopsis genomes. In particular, we wanted to understand how the sequences of these proteins were related to their genomic position, and how this might relate to the mechanisms by which this group of genes evolved.

Our interpretation of the positional relationships among the genes, as illustrated in figure 3.4., and the phylogenetic relationship among the proteins, as illustrated in figure 3.7., is that both segmental duplication and retrotransposition processes each played a role in the evolution of this region, with segmental duplication being primarily responsible for the expansion of the family in *A.lyrata*, and retrotransposition playing a more major role in the Brassica genomes. Segmental duplication is most easily understood as arising through unequal genetic crossovers, but the mechanisms driving gene retrotransposition in plants are less well characterized. It has been suggested that the exchange of domains between PPR proteins may provide one means of functional diversification for RFL genes, but we

were unable to observe any clear evidence for such an event. We did, however, find evidence for RNA-mediated gene conversion in which the coding sequence an *AlyRFL9*-like gene replaced the original central gene in the duplication involving the *AlyRFL6-8* genes.

Regardless of the mechanism through which this gene family has expanded in the different genomes, it is unclear what function many of these genes may have. AtRFL4 or RPF1, is known to specify specific RNA processing events, as do *Rfn* and *Rfp*. Post transcriptional processes and transcript stability control play a dominant role in determining the mature transcriptome of plant mitochondria. In this regard, endonuclease cleavage provides a 3' hydroxyl substrate for plant mitochondrial polyA polymerase, which targets substrates for degradation via polynucleotide phosphorylase (Farrar and Donnison 2007). The combined consequences of relaxed transcription and numerous ORFs of undefined function in plant mitochondria can result in the expression of "toxic" proteins. CMS provides the clearest example of this phenomenon. In this case, the function of the restorer protein can be viewed as specifying a site or sites for endonuclease cleavage that would then destabilize the 5' end of the transcript and prevent its translation (Brown 1999). The cryptic transcripts that accumulate in mitochondrial polynucleotide phosphorylase mutants can include other ORFs (Wang et al. 2011), the expression of which may not cause male sterility but may be otherwise detrimental to the organism. The number and types of these ORFs can change rapidly and the transcript degradation system of the mitochondria must evolve rapidly to accommodate this change.

It seems possible that this group of *RFL* proteins has evolved to specify novel endonuclease cleavage events that trigger the degradation of such transcripts.

FIGURES AND TABLES

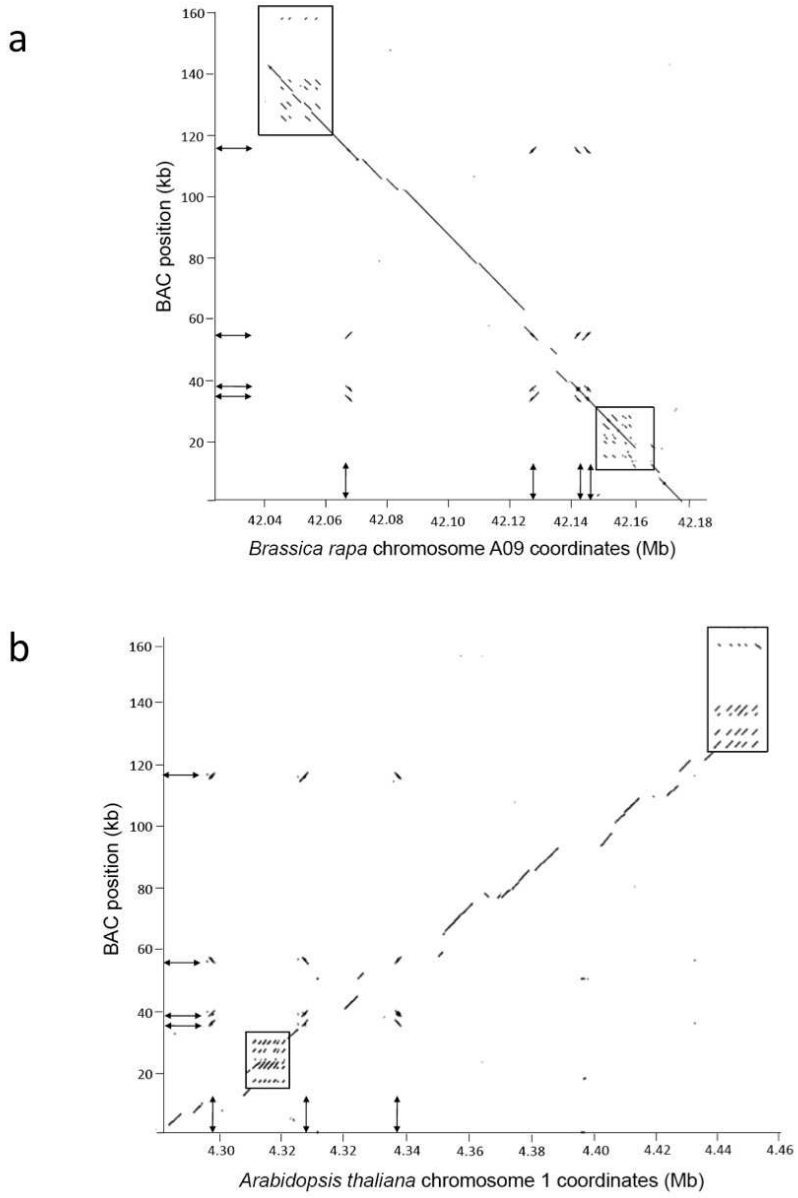


Figure 3.1. Dot matrix alignments of the sequence of a bacterial artificial chromosome (BAC) anchored in the *Rfp* region with the syntenic regions of *B. rapa* chromosome A09 (a) and *A. thaliana* chromosome 1 (b). Boxed regions indicate regions encoding short repeated gene/pseudogene sequences. The arrows indicate repeated regions showing similarity to four and three distinct sites in *B. rapa* and *A. thaliana* respectively.

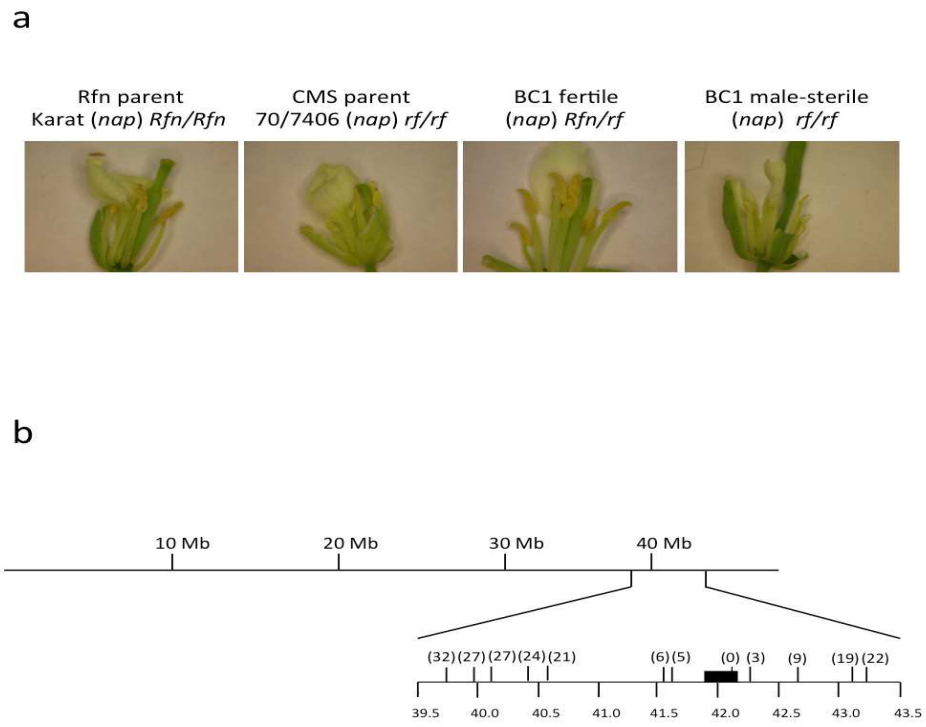


Figure 3.2. Mapping of the *Rfn* gene on chromosome A09. (a) Flowers from the parents and BC1 progeny of the mapping cross. Petals have been removed to allow display of anther morphology. Cytoplasm is designated in parentheses. The male parent of the cross and fertile progeny heterozygous for *Rfn* (*Rfn/rf*) have anthers with normal morphology and shed abundant pollen. The male sterile (CMS) parent and BC1 progeny homozygous for the recessive maintainer allele (*rf/rf*) have stamens with short filaments and underdeveloped anthers that shed little or no pollen. (b) Location of *Rfn* on chromosome A09. The region of the *B. rapa* chromosome chosen as range for targeted mapping extending from coordinates 39.5 to 43.5 Mb is expanded to illustrate the mapping results. Numbers in parentheses indicate the number of observed recombination events between a marker at that location and the *Rfn* gene. The filled rectangle indicates position of the BAC spanning the region delimited in *Rfp* mapping experiments. No recombination was observed between the single marker located within the BAC and the *Rfn* gene.

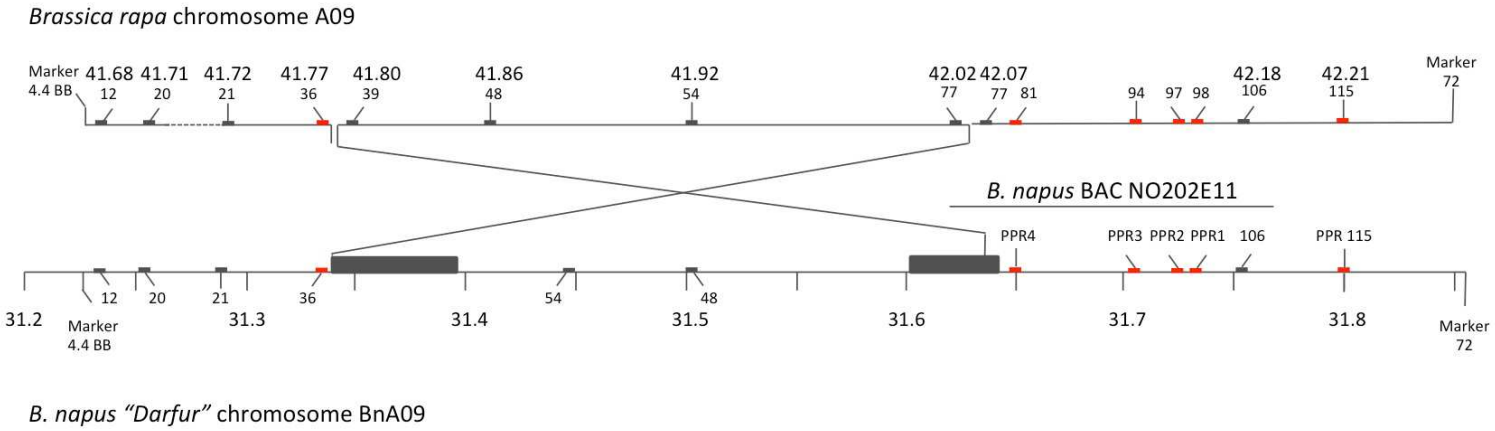


Figure 3.3. Organizational differences in the Brassica *Rf* locus between *B. rapa* and *B. napus*. Locations of specific genes appear as thick bars on the lines representing the two genomes. Genes are indicated according to the Phytozome *B. rapa* designations (i.e. 12 corresponds to Brara.I05012). Map coordinates in Mb appear at specific intervals in the representation of the *B. napus* genome (lower bar), and above the coordinates of the corresponding genes in the *B. rapa* genome representation (upper bar). *RFL* genes are designated as red bars. The dotted line between genes 20 and 21 of the *B. rapa* genome is used to indicate the absence of genes that are present at this site in the *B. napus* genome. The crossed lines between the genomes indicate that site of a major sequence inversion. The filled boxes on the *B. napus* genome are used to indicate the regions containing duplicated F-box and Cytochrome P_{450} encoding genes that may have been involved in the sequence rearrangement.

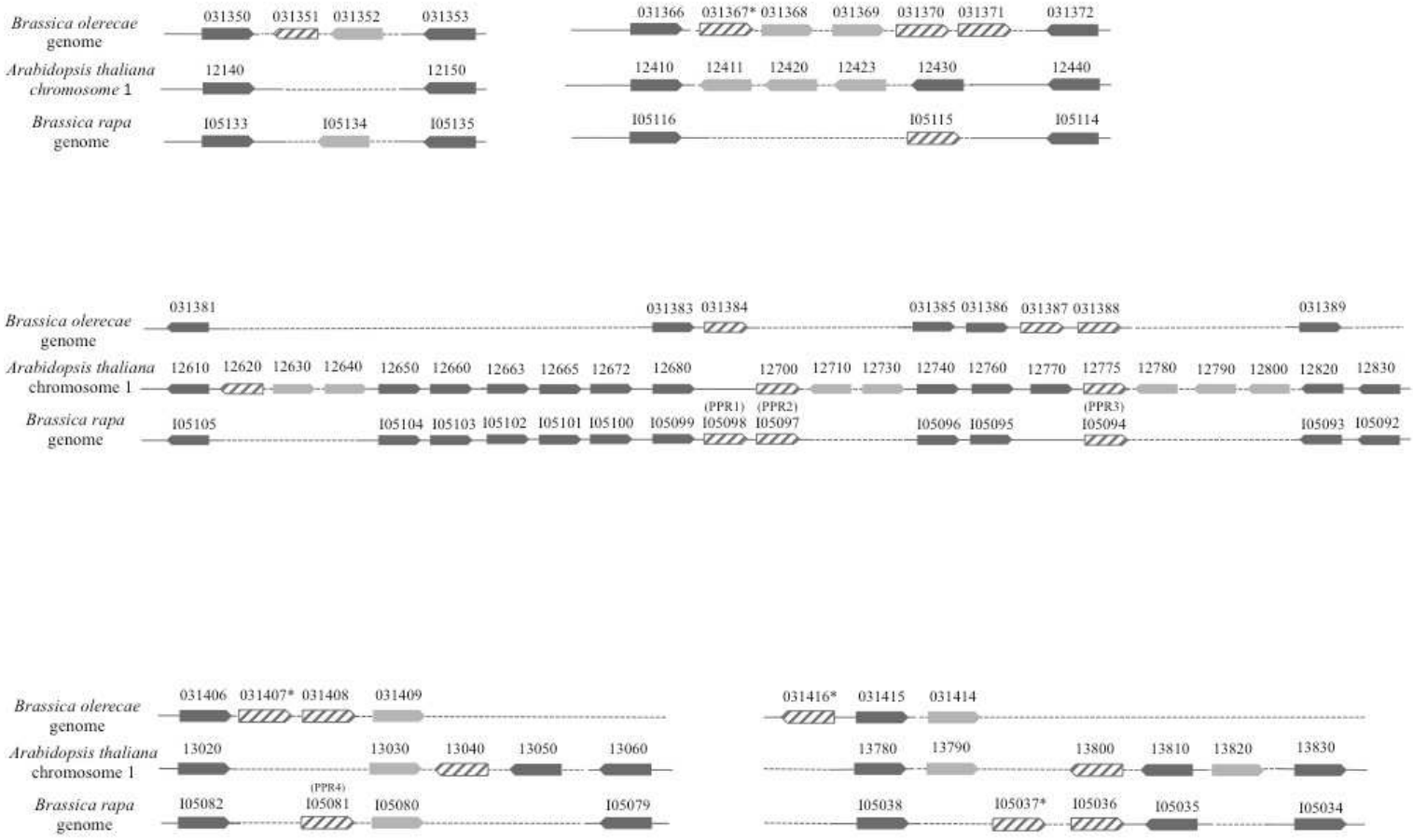


Figure 3.4. Position of *RFL* and surrounding genes in the *B. napus/rapa* Rf region and orthologous segments of *A. thaliana* and *B. oleracea*. Genes are indicated by wide arrows, with the direction of the arrows indicating the 5' to 3' orientation of the coding sequence. *RFL-PPR* genes are hatched. Dark filling indicates genes conserved in location among the three genomes, light filling a gene that is absent at the same position in one or both of the other genomes. Numbers reflect the online gene annotation as prefixed by Bol, At1g and Brara, in *B. oleracea, A. thaliana* and *B. rapa*, respectively. Asterisks (*) indicate *RFL* genes considered too small to be included in the Phylogeny of figure 3.7.

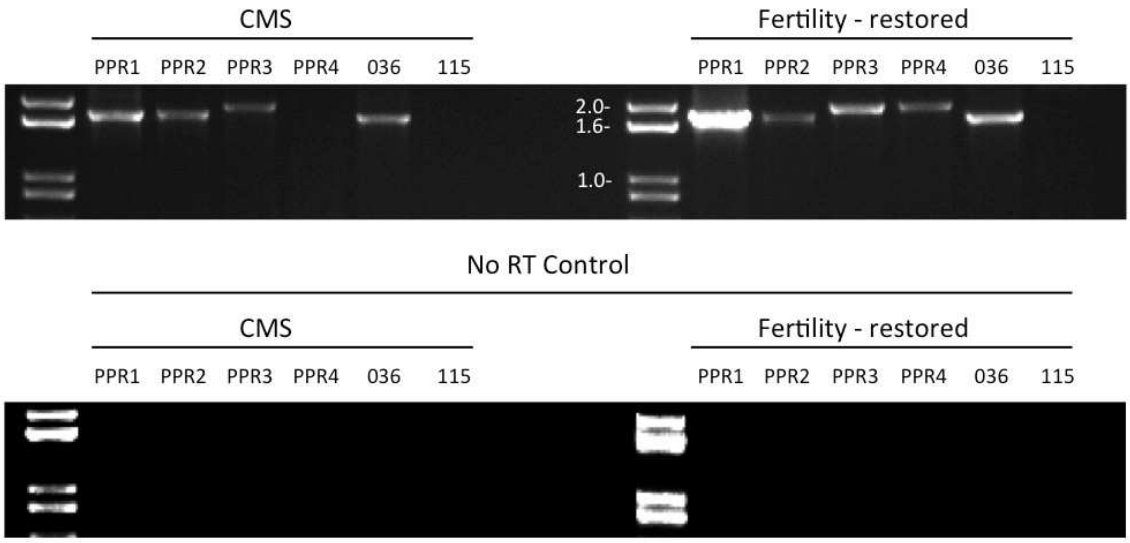
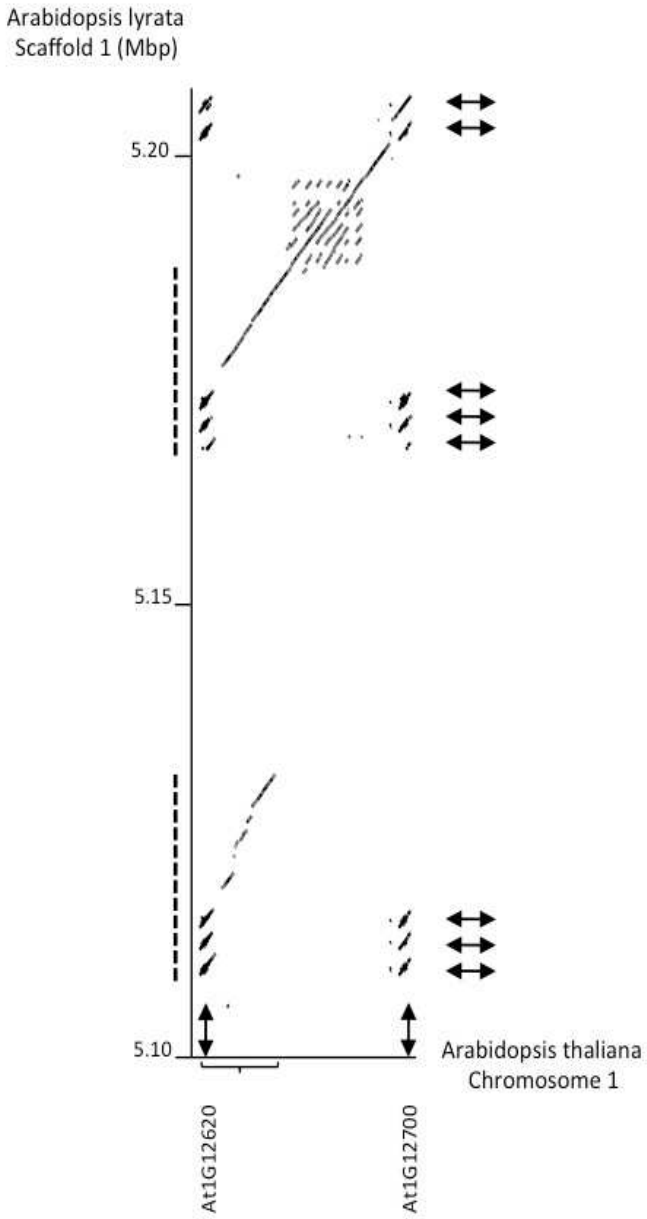


Figure 3.5. Expression of *Rf*-region *B. napus PPR* genes. Total RNA was extracted from the CMS and restorer parents of the mapping cross and analyzed by RT-PCR. The lower panel shows the results of the experimental control, performed at the same time on the same group of samples, in which reverse transcriptase was omitted prior to PCR amplification of cDNA products.

Figure 3.6. RFL genes in a region of the *A. lyrata* genome orthologous to the region of the *A. thaliana* genome depicted in Figs. 1 and 3. (a) A segment of the dot matrix illustrating the duplication of a region containing 3 *RFL* genes at positions indicated by the double headed in the *A. lyrata* genome; all three genes and the surrounding sequences correspond to the *AtRFL3* (At1g12620) region of *A. thaliana.* The bracket indicates the region of the Arabidopsis genome that is duplicated in *A. lyrata* (dotted lines along the y axis). A second duplication occurs at the position of *AtRFL4* (At1g12700); in this case the duplication involves the *AtRFL2* gene (At1g12300). (b) Depiction of the analyzed segment of the *A. lyrata* genome. Arrows indicate the direction and orientation of the RFL genes. The dotted lines indicate duplicated segments: one containing three genes at the location corresponding to *AtRFL3* and the other containing two genes apparently derived from *AtRFL2*, with one copy at the location of *AtRFL2* and the other at the location of *AtRFL4.*

a


Arabidopsis lyrata
Scaffold 1 (Mbp)
5.20
5.15
5.10
Arabidopsis thaliana
Chromosome 1
At1G12620
At1G12700

b


4.90
4.95
5.00
5.05
5.10
5.15
5.20
5.25
Arabidopsis lyrata
Scaffold 1 (Mbp)
RFL1
RFL2
RFL3
RFL4
RFL5
RFL6
RFL7
RFL8
RFL9
RFL10*

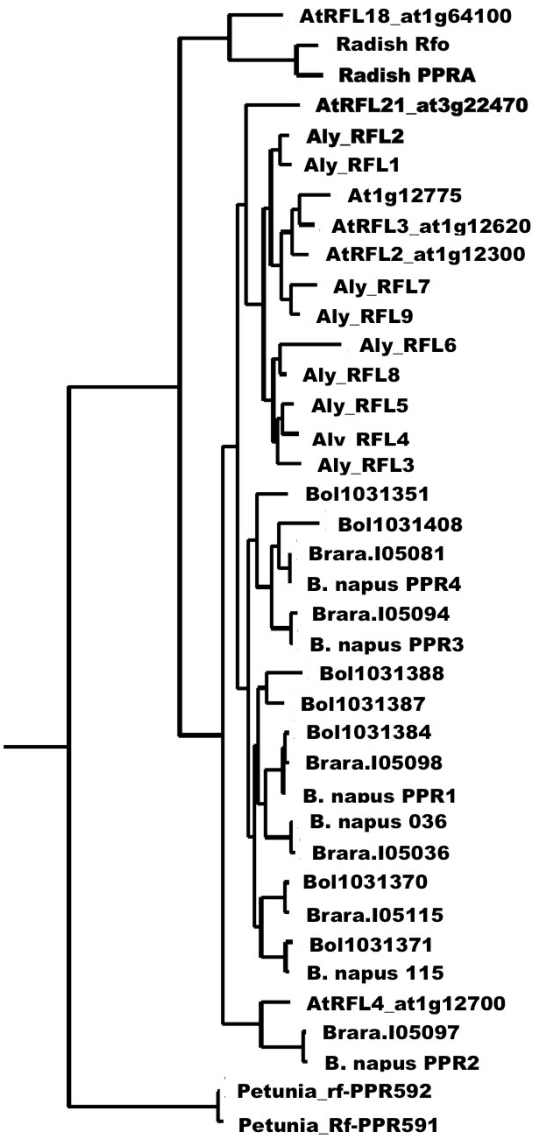


Figure 3.7. Phylogeny of Brassica and Arabidopsis PPR proteins. A maximum likelihood tree was generated by the PhyML resource (Dereeper, Guignon et al. 2008) using as input the sequences of known dicot restorer proteins and related orthologs, the predicted protein sequences of *A. thaliana* subgroup 1 RFL proteins, *A. lyrata* proteins predicted from our analysis of the targeted region enriched in subgroup 1 RFL proteins (Fujii, Bond et al. 2011), and the *B. napus*, *B. rapa* and *B. oleracea* RFL proteins of the Brassica *Rf*-region.

References

Akagi, H., A. Nakamura, Y. Yokozeki-Misono, A. Inagaki, H. Takahashi, K. Mori and T. Fujimura (2004). "Positional cloning of the rice *Rf-1* gene, a restorer of BT-type cytoplasmic male sterility that encodes a mitochondria-targeting PPR protein." Theoretical and applied genetics 108(8): 1449-1457.

Balk, J. and C. J. Leaver (2001). "The PET1-CMS mitochondrial mutation in sunflower is associated with premature programmed cell death and cytochrome c release." The Plant Cell Online 13(8): 1803-1818.

Bentolila, S., A. A. Alfonso and M. R. Hanson (2002). "A pentatricopeptide repeat-containing gene restores fertility to cytoplasmic male-sterile plants." Proceedings of the National Academy of Sciences 99(16): 10887-10892.

Brown, G. (1999). "Unique aspects of cytoplasmic male sterility and fertility restoration in Brassica napus." Journal of Heredity 90(3): 351-356.

Brown, G. G., N. Formanová, H. Jin, R. Wargachuk, C. Dendy, P. Patil, M. Laforest, J. Zhang, W. Y. Cheung and B. S. Landry (2003). "The radish *Rfo* restorer gene of Ogura cytoplasmic male sterility encodes a protein with multiple pentatricopeptide repeats." The Plant Journal 35(2): 262-272.

Budar, F., P. Touzet and R. De Paepe (2003). "The nucleo-mitochondrial conflict in cytoplasmic male sterilities revisited." Genetica 117(1): 3-16.

Chalhoub, B., F. Denoeud, S. Liu, I. A. P. Parkin, H. Tang, X. Wang, J. Chiquet, H. Belcram, C. Tong, B. Samans, M. Corréa, C. Da Silva, J. Just, C. Falentin, C. S. Koh, I. Le Clainche, M. Bernard, P. Bento, B. Noel, K. Labadie, A. Alberti, M. Charles, D. Arnaud, H. Guo, C. Daviaud, S. Alamery, K. Jabbari, M. Zhao, P. P. Edger, H. Chelaifa, D. Tack, G. Lassalle, I. Mestiri, N. Schnel, M.-C. Le Paslier, G. Fan, V. Renault, P. E. Bayer, A. A. Golicz, S. Manoli, T.-H. Lee, V. H. D. Thi, S. Chalabi, Q. Hu, C. Fan, R. Tollenaere, Y. Lu, C. Battail, J. Shen, C. H. D. Sidebottom, X. Wang, A. Canaguier, A. Chauveau, A. Bérard, G. Deniot, M. Guan, Z. Liu, F. Sun, Y. P. Lim, E. Lyons, C. D. Town, I. Bancroft, X. Wang, J. Meng, J. Ma, J. C. Pires, G. J. King, D. Brunel, R. Delourme, M. Renard, J.-M. Aury, K. L. Adams, J. Batley, R. J. Snowdon, J. Tost, D. Edwards, Y. Zhou, W. Hua, A. G. Sharpe, A. H. Paterson, C. Guan and P. Wincker (2014). "Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome." Science 345(6199): 950-953.

Charlesworth, D. and B. Charlesworth (1981). "Allocation of resources to male and female functions in hermaphrodites." Biological Journal of the Linnean Society 15(1): 57-74.

Cheng, F., S. Liu, J. Wu, L. Fang, S. Sun, B. Liu, P. Li, W. Hua and X. Wang (2011). "BRAD, the genetics and genomics database for Brassica plants." BMC plant biology 11(1): 1.

Cheung, F., M. Trick, N. Drou, Y. P. Lim, J.-Y. Park, S.-J. Kwon, J.-A. Kim, R. Scott, J. C. Pires and A. H. Paterson (2009). "Comparative analysis between homoeologous genome segments of *Brassica napus* and its progenitor species reveals extensive sequence-level divergence." The Plant Cell 21(7): 1912-1928.

Delph, L. F., P. Touzet and M. F. Bailey (2007). "Merging theory and mechanism in studies of gynodioecy." Trends in Ecology & Evolution 22(1): 17-24.

Dereeper, A., V. Guignon, G. Blanc, S. Audic, S. Buffet, F. Chevenet, J.-F. Dufayard, S. Guindon, V. Lefort and M. Lescot (2008). "Phylogeny.fr: robust phylogenetic analysis for the non-specialist." Nucleic acids research 36(suppl 2): W465-W469.

Desloire, S., H. Gherbi, W. Laloui, S. Marhadour, V. Clouet, L. Cattolico, C. Falentin, S. Giancola, M. Renard and F. Budar (2003). "Identification of the fertility restoration locus, *Rfo*, in radish, as a member of the pentatricopeptide-repeat protein family." EMBO reports 4(6): 588-594.

FAN, Z. and B. Stefansson (1986). "Influence of temperature on sterility of two cytoplasmic male-sterility systems in rape (*Brassica napus* L.)." Canadian journal of plant science 66(2): 221-227.

Fan, Z., B. Stefansson and J. Sernyk (1986). "Maintainers and restorers for three male-sterility-inducing cytoplasms in rape (*Brassica napus* L.)." Canadian journal of plant science 66(2): 229-234.

Farrar, K. and I. S. Donnison (2007). "Construction and screening of BAC libraries made from Brachypodium genomic DNA." Nature protocols 2(7): 1661-1674.

Formanová, N., X.-Q. Li, A. M. Ferrie, M. DePauw, W. A. Keller, B. Landry and G. G. Brown (2006). "Towards positional cloning in *Brassica napus*: generation and analysis of doubled haploid *B. rapa* possessing the *B. napus pol* CMS and *Rfp* nuclear restorer gene." Plant molecular biology 61(1-2): 269-281.

Formanová, N., R. Stollar, R. Geddy, L. Mahé, M. Laforest, B. S. Landry and G. G. Brown (2010). "High-resolution mapping of the *Brassica napus Rfp* restorer locus using Arabidopsis-derived molecular markers." Theoretical and applied genetics 120(4): 843-851.

Foxe, J. P. and S. I. Wright (2009). "Signature of diversifying selection on members of the pentatricopeptide repeat protein family in *Arabidopsis lyrata*." Genetics 183(2): 663-672.

Fujii, S., C. S. Bond and I. D. Small (2011). "Selection patterns on restorer-like genes reveal a conflict between nuclear and mitochondrial genomes throughout angiosperm evolution." Proceedings of the National Academy of Sciences 108(4): 1723-1728.

Geddy, R. and G. G. Brown (2007). "Genes encoding pentatricopeptide repeat (PPR) proteins are not conserved in location in plant genomes and may be subject to diversifying selection." BMC genomics 8(1): 130.

Hanson, M. R. and S. Bentolila (2004). "Interactions of mitochondrial and nuclear genes that affect male gametophyte development." The Plant Cell 16(suppl 1): S154-S169.

Hölzle, A., C. Jonietz, O. Törjek, T. Altmann, S. Binder and J. Forner (2011). "A RESTORER OF FERTILITY-like PPR gene is required for 5'-end processing of the *nad4* mRNA in mitochondria of *Arabidopsis thaliana*." The Plant Journal 65(5): 737-744.

Hu, J., K. Wang, W. Huang, G. Liu, Y. Gao, J. Wang, Q. Huang, Y. Ji, X. Qin, L. Wan, R. Zhu, S. Li, D. Yang and Y. Zhu (2012). "The Rice Pentatricopeptide Repeat Protein RF5 Restores Fertility in Hong-Lian Cytoplasmic Male-Sterile Lines via a Complex with the Glycine-Rich Protein GRP162." The Plant Cell 24(1): 109-122.

Itabashi, E., N. Iwata, S. Fujii, T. Kazama and K. Toriyama (2011). "The fertility restorer gene, *Rf2*, for Lead Rice-type cytoplasmic male sterility of rice encodes a mitochondrial glycine-rich protein." The plant journal 65(3): 359-367.

Kazama, T. and K. Toriyama (2003). "A pentatricopeptide repeat-containing gene that promotes the processing of aberrant *atp6* RNA of cytoplasmic male-sterile rice." FEBS letters 544(1): 99-102.

Klein, R., P. Klein, J. Mullet, P. Minx, W. Rooney and K. Schertz (2005). "Fertility restorer locus *Rf1* of sorghum (*Sorghum bicolor* L.) encodes a pentatricopeptide repeat protein not present in the colinear region of rice chromosome 12." Theoretical and applied genetics 111(6): 994-1012.

Koizuka, N., R. Imai, H. Fujimoto, T. Hayakawa, Y. Kimura, J. Kohno-Murase, T. Sakai, S. Kawasaki and J. Imamura (2003). "Genetic characterization of a pentatricopeptide repeat protein gene, *orf687*, that restores fertility in the cytoplasmic male-sterile Kosena radish." The plant journal 34(4): 407-415.

Konieczny, A. and F. M. Ausubel (1993). "A procedure for mapping Arabidopsis mutations using co-dominant ecotype-specific PCR-based markers." The Plant Journal 4(2): 403-410.

L'Homme, Y., R. J. Stahl, X.-Q. Li, A. Hameed and G. G. Brown (1997). "Brassica *nap* cytoplasmic male sterility is associated with expression of a mtDNA region containing a chimeric gene similar to the pol CMS-associated *orf224* gene." Current genetics 31(4): 325-335.

Li, X.-Q., M. Jean, B. S. Landry and G. G. Brown (1998). "Restorer genes for different forms of Brassica cytoplasmic male sterility map to a single nuclear locus that modifies transcripts of several mitochondrial genes." Proceedings of the National Academy of Sciences 95(17): 10032-10037.

Liu, Z., P. Liu, F. Long, D. Hong, Q. He and G. Yang (2012). "Fine mapping and candidate gene analysis of the nuclear restorer gene *Rfp* for *pol* CMS in rapeseed (*Brassica napus* L.)." Theoretical and Applied Genetics 125(4): 773-779.

Luo, D., H. Xu, Z. Liu, J. Guo, H. Li, L. Chen, C. Fang, Q. Zhang, M. Bai and N. Yao (2013). "A detrimental mitochondrial-nuclear interaction causes cytoplasmic male sterility in rice." Nature genetics 45(5): 573-577.

Lurin, C., C. Andrés, S. Aubourg, M. Bellaoui, F. Bitton, C. Bruyère, M. Caboche, C. Debast, J. Gualberto and B. Hoffmann (2004). "Genome-wide analysis of Arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis." The Plant Cell 16(8): 2089-2103.

Menassa, R., N. El-Rouby and G. G. Brown (1997). "An open reading frame for a protein involved in cytochrome c biogenesis is split into two parts in Brassica mitochondria." Current genetics 31(1): 70-79.

Parkin, I. A., S. M. Gulden, A. G. Sharpe, L. Lukens, M. Trick, T. C. Osborn and D. J. Lydiate (2005). "Segmental structure of the *Brassica napus* genome based on comparative analysis with *Arabidopsis thaliana*." Genetics 171(2): 765-781.

Sabar, M., D. Gagliardi, J. Balk and C. J. Leaver (2003). "ORFB is a subunit of F1FO-ATP synthase: insight into the basis of cytoplasmic male sterility in sunflower." EMBO Rep 4(4): 381-386.

Schmitz-Linneweber, C. and I. Small (2008). "Pentatricopeptide repeat proteins: a socket set for organelle gene expression." Trends in plant science 13(12): 663-670.

Singh, M. and G. G. Brown (1991). "Suppression of cytoplasmic male sterility by nuclear genes alters expression of a novel mitochondrial gene region." The Plant Cell 3(12): 1349-1362.

Singh, M., N. Hamel, R. Menasaa, X.-Q. Li, B. Young, M. Jean, B. S. Landry and G. G. Brown (1996). "Nuclear genes associated with a single Brassica CMS restorer locus influence transcripts of three different mitochondrial gene regions." Genetics 143(1): 505-516.

Small, I. D. and N. Peeters (2000). "The PPR motif–a TPR-related motif prevalent in plant organellar proteins." Trends in biochemical sciences 25(2): 45-47.

Touzet, P. and F. Budar (2004). "Unveiling the molecular arms race between two conflicting genomes in cytoplasmic male sterility?" Trends in plant science 9(12): 568-570.

Town, C. D., F. Cheung, R. Maiti, J. Crabtree, B. J. Haas, J. R. Wortman, E. E. Hine, R. Althoff, T. S. Arbogast and L. J. Tallon (2006). "Comparative genomics of *Brassica oleracea* and *Arabidopsis thaliana* reveal gene loss, fragmentation, and dispersal after polyploidy." The Plant Cell 18(6): 1348-1359.

Wang, X., H. Wang, J. Wang, R. Sun, J. Wu, S. Liu, Y. Bai, J.-H. Mun, I. Bancroft, F. Cheng, S. Huang, X. Li, W. Hua, J. Wang, X. Wang, M. Freeling, J. C. Pires, A. H. Paterson, B. Chalhoub, B. Wang, A. Hayward, A. G. Sharpe, B.-S. Park, B. Weisshaar, B. Liu, B. Li, B. Liu, C. Tong, C. Song, C. Duran, C. Peng, C. Geng, C. Koh, C. Lin, D. Edwards, D. Mu, D. Shen, E. Soumpourou, F. Li, F. Fraser, G. Conant, G. Lassalle, G. J. King, G. Bonnema, H. Tang, H. Wang, H. Belcram, H. Zhou, H. Hirakawa, H. Abe, H. Guo, H. Wang, H. Jin, I. A. P. Parkin, J. Batley, J.-S. Kim, J. Just, J. Li, J. Xu, J. Deng, J. A. Kim, J. Li, J. Yu, J. Meng, J. Wang, J. Min, J. Poulain, J. Wang, K. Hatakeyama, K. Wu, L. Wang, L. Fang, M. Trick, M. G. Links, M. Zhao, M. Jin, N. Ramchiary, N. Drou, P. J. Berkman, Q. Cai, Q. Huang, R. Li, S. Tabata, S. Cheng, S. Zhang, S. Zhang, S. Huang, S. Sato, S. Sun, S.-J. Kwon, S.-R. Choi, T.-H. Lee, W. Fan, X. Zhao, X. Tan, X. Xu, Y. Wang, Y. Qiu, Y. Yin, Y. Li, Y. Du, Y. Liao, Y. Lim, Y. Narusaka, Y. Wang, Z. Wang, Z. Li, Z. Wang, Z. Xiong and Z. Zhang (2011). "The genome of the mesopolyploid crop species *Brassica rapa*." Nature Genetics 43(10): 1035-1039.

Wang, Z., Y. Zou, X. Li, Q. Zhang, L. Chen, H. Wu, D. Su, Y. Chen, J. Guo and D. Luo (2006). "Cytoplasmic male sterility of rice with boro II cytoplasm is caused by a cytotoxic peptide and is restored by two related PPR motif genes via distinct modes of mRNA silencing." The Plant Cell 18(3): 676-687.

Yang, T.-J., J. S. Kim, S.-J. Kwon, K.-B. Lim, B.-S. Choi, J.-A. Kim, M. Jin, J. Y. Park, M.-H. Lim and H.-I. Kim (2006). "Sequence-level analysis of the diploidization process in the triplicated FLOWERING LOCUS C region of *Brassica rapa*." The Plant Cell 18(6): 1339-1347.

CHAPTER IV : TOWARDS THE IDENTIFICATION OF *Rfn*

-

ADDITIONAL RESULTS AND FUTURE AVENUES

To complete the work presented in Chapter II and Chapter III, I present in this Chapter IV the work done towards the identification of *Rfn*. After the identification of the candidates as presented in Chapter III, I was able to make a series of constructs in order to identify *Rfn* and characterize its mode of action. I developed an efficient *Brassica napus, Agrobacterium tumefaciens* mediated transformation protocol in order to perform molecular complementation. Overall, this Chapter presents ongoing experiments and un-published results in order to expose future work that remains to be done.

Introduction

After exploration of *Rfn* specific RNA processing events in chapter II and mapping of the *Rfn* locus in chapter III, efforts in order to identify *Rfn* have been made using that same map-based cloning approach previously used for a variety of characterized restorer of fertility like *Rfo* (Brown, Formanová et al. 2003), *Rf-PPR2* in petunia (Bentolila, Alfonso et al. 2002) or *Rf1* in BT-CMS in rice (Komori, Ohta et al. 2004). Molecular complementation necessitates the use of binary vectors that would replicate in both *Escherichia coli*, for the creation of the primary construct, and *Agrobacterium tumefaciens*, to allow plant transformation. The reason for this is that *A. tumefaciens*, which is a slow growing bacteria, does not replicate plasmids in high levels and therefore is not useful for rapidly obtaining constructs. Molecular cloning for plant transformation in my lab has been previously done with the pRD400 vector, which carries a wild-type kanamycin resistance gene conferring higher levels of antibiotic resistance. Unlike many plant transformation vectors, pRD400 carries a mutant form, which encodes a less active form of the phosphotransferase enzyme.

However, due to the presence of a large number of repeated domains, PPR have proven difficult to clone into this large binary vector. A number of different alternative approaches were attempted, the most successful one being the use of the Gateway cloning system using the pMDC100 and pMDC123 vectors developed by Curtis and Grossniklaus 2003. The Gateway cloning system uses DNA segments flanked by recombination sites that can be mixed in vitro with specific vectors also containing

recombination sites and incubated with bacteriophage λ integrase recombination proteins to accomplish the transfer of the DNA segment into the vector. This process is referred to as recombinational cloning. This system, as presented in figure 4.1, involves two reactions. First the recombination of attB repeats in the DNA insert with attP repeats in the vector which allows the insertion of the DNA fragment in an entry vector forming a set of new repeats at the recombination sites called attL. Secondly, the recombination of attL repeats with attR lead to the formation attB and attP repeats in order to transfer the DNA fragment from the entry vector to the destination vector, in our case a binary plant transformation vector (Hartley, Temple et al. 2000). Thus, providing different combinations of recombination sites, the system allows for control of the direction of the reactions and orientation of the DNA insert. In this chapter, I describe the work I did to produce of a variety of binary constructs, with the ultimate goal of identifying *Rfn* through molecular complementation. In addition, constructs were generated to allow the production of specifically tagged constructs for functional analysis of the different Rfn candidate proteins.

Only certain cultivars of *Brassica napus*, e.g. the variety "Westar", are highly amenable to genetic transformation with *A. tumefaciens*. Although explants of different plant tissues can be employed in such experiments, perhaps the most widely used protocols employ petiolar tissue of young cotyledons. The restorer of fertility for the Ogura or *ogu* CMS restorer of fertility *Rfo* has been previously identified using a map-based cloning approach involving plant transformation through that method (Brown, Formanová et al. 2003). In this chapter, I explain a variety of protocols for efficient transformation of *B.*

napus and the different tests carried out in order to devise a simplified yet effective protocol for routine *B. napus* transformation. Transformation with genomic DNA in most of the model species, has been proven routinely sufficient in plant biology for generating the expected phenotype produced by the target gene, fertility restoration in our case.

In some cases, restoration of fertility involves a number of partner proteins in large protein complexes. Immuno-precipitation experiments of mitochondrial fractions in petunia restored plants demonstrated that the PPR restorer protein PPR592 is associated with the inner membrane of the mitochondria in a large complex that binds the CMS conferring *pcf* RNA, indicating the implication of a number of partner proteins to Rf-PPR592 in the restoration process (Gillman, Bentolila et al. 2007). *Rf5*, the restorer of fertility in rice HL-CMS was not found to be able to bind directly to the CMS conferring transcript, *atp6-orfH79* but rather to work in a complex with a glycine rich protein (Hu, Wang et al. 2012). To gain more information about the fertility restoration mechanism in *B. napus nap* CMS, we decided to examine association of the Rf protein with mitochondrial mRNA as well as potential partners proteins. Such an exploration would allow completion of functional studies of the mechanisms of fertility restoration. To that end, HA-tagged PPR proteins candidates were designed using the same Gateway cloning system described above as well as constructs allowing production of recombinant proteins.

The production of recombinant PPR proteins in *E. coli* has been used as a heterologous expression system in the recent years in order to characterize the functional properties of PPR proteins. Haïli, Arnal et al. 2013, examined the function of the mitochondria-targeted mitochondrial stability factor 1 (MTSF1) protein in *Arabidopsis thaliana*. Using recombinant MTSF1, they were able to provide evidence that this PPR protein binds to its targeted RNA, *nad4*, in a sequence specific manner and is required for correct 3' processing and stability of the mitochondrial *nad4* mRNA. In that context, the study of *Rfn* function using recombinant proteins seemed a promising avenue to explore. Here, I detail the progress that has been made to produce such proteins and the challenges met while working towards identification of *Rfn* from the candidates identified in chapter III.

Material and methods

Transgene construction

HA-tagged gene synthesis products, containing attB repeats, were inserted into pDonr 207 (Fisher-Thermoscientific, Pleasanton, CA, USA), an entry vector, using the BP Clonase Enzyme mix (11789-021 Fisher-Thermoscientific, Pleasanton, CA, USA) according to the manufacturer's instructions. The endogenous gene constructs were built from amplification products using Q5® Hot Start High-Fidelity DNA Polymerase (New England Biolabs, Ipswich, MA, USA) according to the manufacturer's instructions. The primers designed for the amplification of each candidate gene are listed in table 4.1 and contain the attB used for insertion in the entry vector pDonr207. BP reactions were introduced into electroporation competent Xl1-Blue *E. coli* (L2005 Promega, Madison,

Wisconsin, USA) by electroporation at 2.5V. Clones obtained after the BP reaction on gentamycin selective LB plates (10 μg/mL final antibiotic concentration), were digested and sequenced before cloning into the destination vectors pMDC100 and pMDC123 (http://www.botinst.uzh.ch/en/research/development/grossnik/vectors/MarkdGatewayVectors.html). Inserts were introduced into the destination vector using the LR clonase enzyme mix (11791-100 Fisher-Thermoscientific, Pleasanton, CA, USA) according to the manufacturer's instructions. Positive clones on the kanamycin selective plates (50 μg/mL final antibiotic concentration) were digested by appropriate restriction enzymes and sequenced to verify sequence fidelity. Final constructs were then transformed in electro-competent *A. tumefaciens* GV3101 at 2.8V (Mersereau, Pazour et al. 1990) and plated on kanamycin, rifampicin selective LB plates. A final verification of insert stability was made by restriction enzyme digestion before producing cell stocks in 30% glycerol solution stored at -80°C.

Plant transformation

B. napus plants with the "Westar" cultivar nuclear genotype, Polima cytoplasm that are homozygous at the *Rf* locus for the *Rfp* gene (*pol Rfp*) were transformed as described in Moloney, Walker et al. 1989, with some modifications in medium used as described in figure 4.2 and Results. Seed sterilization was achieved by placing the seeds in a sterile 50 mL plastic tube with approximately 40 mL of sodium hypochlorite (commercial solution, 20% v/v). Seeds were incubated shaking at room temperature for 20 minutes. The sodium hypochlorite solution was discarded and the seeds were rinsed five times by shaking for 5

minutes with 40mL of sterile water. The sterilized seeds were then transferred onto germination medium (20 seeds per petri dishes) and incubated at 4° C in the dark for 48 hours before being placed in germination chambers for 7 days (16-h photoperiod, 22°/16°C day/night temperatures). Agrobacterium suspensions were prepared by first streaking the cells containing the binary vector onto an LB plate containing 100 mg/mL rifampicin and 50 mg/mL kanamycin and incubating for two days at 28°C. A single colony of Agrobacterium was then inoculated into 10 mL of LB medium containing 100 mg/mL rifampicin and 50 mg/mL kanamycin and incubated with shaking at 250 rpm at 28°C for 36 hours. The Agrobacterium culture was then centrifuged at 4,500 x g for 10 minutes at room temperature and rinsed in MS minimal organic medium without antibiotics. The cells were re-pelleted before suspension in MS minimal organic medium. Adjustment of the volume of MS medium was used to obtain an OD at 650 nm of 0.05 was done. Under sterile conditions, *B. napus* seedlings were removed from germination medium and placed in an empty petri plate. Cotyledons were cut to include approximately 2mm of their stalk, the petiole, using a sterile scalpel and placed directly onto a co-cultivation plate with the petiole ends embedded in the medium. Once all the cotyledons were cut, each was gently picked out of the co-cultivation petri dish using sterile forceps and dipped for one second in Agrobacterium solution before being replaced on the co-cultivation plate. The Petri dishes were sealed using 3M surgical tape and incubated under dim light (B600 lux which is achieved using only 1 light of 4 on a germination chamber shelf) for two days. The explants were then transferred onto callus induction medium using sterile, sealed with 3M surgical tape and incubated under dim light at 25°C for one to 8 weeks until a callus was formed. Once a callus is formed,

visualized by the formation of a somewhat round aggregate of undifferentiated tissue, explants were transferred onto shoot initiation medium to allow formation of green tissue. After 2 to 8 weeks of incubation in light (16 hours photoperiod), 1-2 cm shoots were formed. Individual shoots were then transferred onto shoot outgrowth medium by carefully excising them at the base of the callus and placing them in individual tissue culture vessels containing shoot generation medium with the cut side embedded in the medium. Once the shoot was 3-5 cm long, which took up to 5 weeks, it was transferred in an individual tissue culture vessel containing rooting medium. Rooted shoots were transferred to pots containing compact well-watered soiled and covered with a plastic film placed on top of the pot for 3 days, then allowed to continue to grow in a growth chamber under the same conditions used for growing plants from seed. Selection of transformants is done throughout the transformation process by adding kanamycin to the selection and outgrowth media.

Media were prepared as follows. The base was composed of full strength Murashige and Skoog plant medium base (Sigma M5524-10L), 3% sucrose at pH 5,7 and autoclaved for 45 minutes. After sterilization of base medium, components were added following filter sterilization for various sub-media to give final concentrations of their different components as follows. The germination medium contained Myo-inositol 100 mg/L, thiamine 10 mg/L nicotinic acid 1 mg/L and pyridoxine 1 mg/L. The co-cultivation medium was prepared by adding myo-inositol 100 mg/L, thiamine 10 mg/L nicotinic acid 1 mg/L, pyridoxine 1 mg/L and benzyl amino purine 4.5 mg/L to the base medium whereas the selection medium contained the base medium supplemented by myo-inositol

100 mg/L, thiamine 10 mg/L nicotinic acid 1 mg/L, pyridoxine 1 mg/L, benzyl amino purine 4.5 mg/L, kanamycin 25 mg/L and carbenicillin 500 mg/L. The outgrowth medium contained myo-inositol 100 mg/L, thiamine 10 mg/L nicotinic acid 1 mg/L, pyridoxine 1 mg/L, kanamycin 25 mg/L and carbenicillin 500 mg/L in plus of the base medium. Lastly, the rooting medium was the base with myo-inositol 100 mg/L, thiamine 10 mg/L nicotinic acid 1 mg/L, pyridoxine 1 mg/L, Indole-3-butyric acid 2 mg/L and carbenicillin 500 mg/L added after autoclaving.

Characterization of transformants

Leaf disks were removed from plantlets that survived kanamycin selection and DNA was isolated from them according to medium (Edwards, Johnstone et al. 1991). PCR was performed to detect the insertion of the transgene using the primers indicated in table 4.1. Confirmed transgenic plants were then crossed as males with *nap* CMS *B. napus* in order to assess the capacity of the transgene to rescue the CMS phenotype and restore pollen production.

Isolation of RNA, circularization and RT-PCR

The circularized RT-PCR allowing for the detection of *Rfn* specific *nad4* processing was performed as described in Chapter II.

Expression of recombinant proteins in E.coli

Primers were designed so that constructs for expression of proteins in *E. coli* lacked their predicted mitochondrial targeting sequences as well as their stop codons. The resulting open reading frames were inserted into destination vectors to allow for production of proteins containing a poly-histidine or maltose binding protein tag at their C termini. The corresponding DNA fragments were first PCR amplified using the primers containing gene specific sequences as well as the AttB1 and AttB2 recombination sites as listed in table 4.1. The amplification products were cloned by the Gateway BP enzyme reaction mix into pDONR207 (11789-021 Fisher-Thermoscientific, Pleasanton, CA, USA) according to the manufacturer's instruction and subsequently transferred into the pDEST17 (6x His tag fusion) or pDEST-HisMBP (MBP fusion) expression vector using LR enzyme reaction mix (11791-100 Fisher-Thermoscientific, Pleasanton, CA, USA) as instructed in the manufacturer's manual. Chemically competent *E. coli* "Rosetta 2" cells (Catalog number 71405-3, Novagen,) were then transformed with the constructs by a 45 second heat shock at 42°C. Transformed clones were incubated overnight in a 3 mL starter culture in lysogeny broth media supplemented with appropriate antibiotics. 1 mL of this starter culture was inoculated in 20 mL of the same media for induction until the O.D at 600 nm reached the value 0.4-0.6. A non-induced control was sampled before applying IPTG to promote protein synthesis for 3 hours at 37°C. The temperature and time of induction were changed as noted in the results and discussion. Samples of soluble and insoluble proteins were prepared by centrifugation of 10 mL cells at 4 000 x g for 30 minutes at 4°C. The resulting cells were re-suspended in 100 µL of bugbuster master mix

(Catalog number 71456-3, Novagen) and incubated 10 minutes at room temperature. Centrifugation for 15 minutes at maximum speed allowed the separation of soluble and insoluble proteins fractions. The pelleted insoluble proteins were re-suspended in re-suspension buffer (50 mM Tris, 150 mM NaCl, 2% sarkosyl, pH 7.5) for SDS-PAGE analysis. The protein samples were dissolved in a 4 x concentrated Laemmli buffer (65.8 mM Tri, 2.1% SDS (v/v), 26.3% (w/v) glycerol, 0.01% bromophenol blue (w/v), pH 6.8,) and heated at 95°C for 5 min prior to gel loading. Separation was performed on a 7.5% polyacrylamide resolving gel. The polyacrylamide gel was placed in staining solution (50% v/v methanol, 10% acetic acid v/v, 0.025% coomassie brilliant blue R250) and shaken horizontally for 30 min at room temperature. The gel was distained in a solution (10% (v/v) methanol, 10% (v/v) acetic acid) by shaking for a minimum of 30 minutes.

Results

Cloning

Molecular complementation constructs

Molecular complementation, as used previously by Bentolila, Alfonso et al. 2002, Brown, Formanová et al. 2003, Komori, Ohta et al. 2004, for the identification of nuclear restorer genes, was employed to determine which of the different candidates present in the *Rfn* containing BAC described in Chapter III functioned as the restorer. It was anticipated that one of the genes, introduced as a transgene, would allow male fertility restoration of nap CMS plants and the associated RNA processing events. After using several web-based

prediction programs on the PPR gene sequences predicted by Genscan in the BAC to determine the potential promoter and terminator sequences, primers, as described in table 4.1, were designed to amplify those sequences with the attB1 and attB2 sequences to allow cloning using the Gateway system as PCR amplification allowed the production of amplification products suitable for cloning into the entry vector pDonr207, which contains gentamycin resistance gene. PCR based screening of gentamycin resistant clones, restriction digestion and sequencing were performed to confirm the integrity of the insert gene before transferring to the destination plant transformation vectors pMDC100 and pMDC123. These destination vectors contain kanamycin bacterial resistance gene as well as kanamycin (pMDC100) and BASTA (pMDC123) plant resistance genes.

The simplest method of confirming which of the candidate genes was *Rfn* would be to simply introduce each of the constructs into *nap* CMS plants and determine which construct would lead to rescue of the CMS phenotype and possibly the associated RNA processing events. However, not all varieties of *B. napus* are equally amenable to the transformation process and, as described later, the cultivar "Browowski", one of the few varieties which lacks *Rfn*, is not transformable. Therefore, the 4 PPR candidates were cloned with their endogenous promoters and terminator in order to express them in a restored *pol Rfp* background and observe fertility restoration using *A. tumefaciens* mediated transformation on petiolar tissue.

HA-tagged PPR candidates

Previous studies showed that in some cases restoration of fertility involves a number of partner proteins in large protein complex. For example, immuno-precipitation experiments of Rf-PPR2 containing mitochondrial fractions suggested that the protein resided in large protein complex associated with inner mitochondrial membrane (Gillman, Bentolila et al. 2007). In order to work towards the characterization of the function of the Rfn protein, production of tagged constructs of the PPR candidates would facilitate in vivo expression and further functional analysis. As an alternative for PCR amplification, gene synthesis was ordered for *PPR1 PPR2* and *PPR3*. The cloning of the tagged-PPR4 construct has been achieved by PCR amplification from the BAC The gene synthesis was designed such that endogenous promoters, three 3' end HA tags and the attB1 and attB2 repeats were included. The gene synthesis products were delivered in a pMK-RQ vector backbone, which contains kanamycin resistance. The products, containing the attB repeats, were substrates for BP clonase enzyme mix and allowed the cloning of the DNA fragments in the entry vector pDonr207. PCR, restriction digestion and sequencing were used to confirm the integrity insert sequences cloning in the destination vectors pMDC100 and pMDC123.

Plant transformation

The *B. napus* transformation process involves contamination of cotyledons with an Agrobacterium solution containing the construct of interest. Agrobacterium infection results in the integration of the gene introduced into the transformation vector into the plant genome. The resulting transformants are obtained through regeneration of the plant from individual transformed cells; the totipotency proprieties of plant cells that allows them to differentiate, generate new tissues and form a plant, under the right growth conditions including appropriate growth promoting hormones. For *B. napus,* the generation of a fully-grown, flowering plant requires 6 to 9 months. .

Regeneration test of nap CMS cultivar

The simplest method of confirming which of the candidate genes was *Rfn* would be to simply introduce each of the constructs into *nap* CMS plants and determine which construct lead to rescue of the CMS phenotype and possibly the associated RNA processing events. Most *B. napus* varieties, however, though they contain *nap* cytoplasm, are male fertile due to homozygosity at the *Rf locus* for the *Rfn* gene, and therefore not suitable for this approach. Not all varieties of *B. napus* are equally amenable to the transformation process, and we found no evidence the cultivar "Browowski", one of the few varieties which lacks *Rfn*, had ever been used in this capacity. To test its capacity to be used as a recipient in genetic transformation and to undergo into fully-grown plants, mock transformation was carried out using Agrobacterium strains lacking an introduced

plasmid and as well as a negative control in which cotyledons of the "Bronowski" underwent the same procedure without Agrobacterium co-cultivation in order to only test its regeneration ability. As a supplemental control, Westar *Rfp* plants were transformed alongside Bronowski in both conditions. Unfortunately, none of the Bronowski cotyledons showed any sign of callus production, even past the normal 8 week callus formation period, contrary to the Westar cultivar cotyledons that had all formed callus within the 8 weeks timeframe and regenerated in wild type looking shoots after transfer onto outgrowth media. This data indicated that direct transformation of our *nap* CMS material would be a difficult task to achieve due to the lack of regenerative ability of this material.

Cross between nap CMS and Rfp

Only certain cultivars of *Brassica napus*, e.g. the spring variety "Westar", are highly amenable to genetic transformation with *A. tumefaciens*. Because the Bronowski cultivar proved to not be transformable, we were forced to explore a different strategy for identification of *Rfn.* The cultivar Westar can be transformed with high efficiency, but Westar, like most *B. napus* varieties, is homozygous for *Rfn*. We have previously employed a strain of Westar into which the *Rfp* gene has been introduced through multiple generations of repeated backcrossing (Singh et al, 1996). These plants contain the *pol* CMS and are homozygous for the *Rfp* allele at the *Rf* locus. Since previous studies have shown that *Rfn* and *Rfp* are allelic and therefore that Westar *Rfp* should not contain *Rfn*, we expected this strain should be transformable and that, by integrating the different

Rfn candidates as transgenes into the Westar *Rfp* genomic background we would be able evaluate their capacity to function as *Rfn* by following pollination of *nap* CMS plants. To confirm that our Westar-*Rfp* line truly lacked *Rfn* and that it did not contain genes at other loci that might act to partially restore pollen formation to nap CMS plants, we crossed it to our *nap* CMS line. A clear sterile phenotype was observed as expected when F1 plants were grown to maturation (not shown).

Transformation of Westar Rfp

Bhalla and Singh 2008, described a very detailed protocol for Agrobacterium mediated *B. napus* transformation. It uses a large number of costly materials and did not fit the materials that were previously used for *B. napus* transformation by former lab members. The method was however, thoroughly descriptive and provided a plateform for beginning my plant transformation studies. Therefore, I employed aspects of the Bhalla and Singh 2008, protocol with methods and materials derived from the two other protocols that had been exploited routinely for *B. napus* transformation in the lab. The major modification that has been incorporated was the utilization of simplified media for plant cultivation. Using the same Murashige and Skoog plant media base throughout the method as a growth medium, supplements were added to it for different steps of the protocol as described in Material and Methods section of this chapter and in figure 4.2. After germination and co-cultivation on a germination medium, 6-benzylaminopurine was applied in order to facilitate tissue propagation and formation of callus. Removal of 6-benzylaminopurine promoted shoot outgrowth. When the shoot presented an adequate

height (between three to five centimeters) rooting was promoted by the presence of indole-3-burtyric acid in the medium and the removal of plant selection (kanamycin or basta), which was present from the callus formation stage (figure 4.2). Indeed, plant selection through antibiotics, especially kanamycin, is thought to inhibit greatly root formation (Bhalla and Singh 2008). Moreover, carbenicillin selection was applied throughout the protocol after co-cultivation in order to prevent any unwanted Agrobacterium growth that would endanger the health of the young transformants.

Out of around 300 plants transformed, 31 plants fully recovered after Agrobacterium infection, selection and rooted to form fully-grown flowering plants. Those 31 plants were transformants containing the candidates *PPR1*, *PPR2* and *PPR4* within the pMDC100 vector containing kanamycin plant resistance. I attempted to transform into Westar *Rfp* cotyledons *PPR3* using the pMDC123 vector and selection with the BASTA herbicide, but none of the plants that survived co-cultivation were able to be rooted and therefore could not be analyzed. PCR screening of the 31 transformants pMDC100 transformants indicated that only 3 contained the introduced transgenes either PPR1 or PPR2; the remaining "transformants" were likely plants that survived the selection stages despite lacking the kanamycin resistance conferred by the vector. Thus only 1% of the treated cotyledons lead to true transformants.

Test on transformed plant material

The transformed Westar cultivar plants do not contain the *nap* mitochondrial genome but the *pol* cytoplasm and do not express the nap CMS associated *orf222*/*nad5*c CMS associated transcript. It is therefore not possible to detect the effect of any of the PPR candidates might have on the processing of *nap* CMS conferring transcript *orf222*. However as shown in Chapter II, and in Li, Jean et al. 1998, the *Rfn* locus is associated with an additional RNA cleavage events in the coding region of *nad4a,* which is not observed in the Westar *Rfp* line (Singh, Hamel et al. 1996). The CR-RT-PCR technique described in Chapter II was therefore used to detect *Rfn*-linked processing of *nad4*. The *Rfn* specific processing event is visualized by an additional amplification product of around 300bp. The sequence similarity between the *orf222* and *nad4* processing sites suggested that the PPR candidate responsible for processing of *nad4* would allow us to target one of the PPRs as a preferred *Rfn* candidate before the flowering of the T0 plants and growth of progeny plants generated from pollination of the *nap* CMS line, which would take at least another 6 months.

As mentioned earlier, the *B. napus* transformation process is time consuming. One of strategies used in order to get insight on the identity of *Rfn* was to perform CR-RT-PCR to detect the *Rfn* specific *nad4* processing on young transformed green tissue. First tests were carried out with young transformed material that didn't show the ability to grow roots transformed with *PPR1* and *PPR2*. Indeed, after callus formation shoot outgrowth start and that green healthy tissue can be used as samples for CR-RT-PCR analysis

especially those which do not appear to be able to form roots after a few weeks placed onto rooting media and that therefore would not be used for further experiments.

Results of the CR-RT-PCR on those young transformed tissues are shown in figure 4.3. For the CMS, *Rfn* and untransformed control, young flower buds were used as tissue source. Two major bands between 650 and 850 base pairs are observed. Sequencing of these bands showed that they were the product of ubiquitous, non *Rfn* specific, *nad4* post-transcriptional processing with 5' termini corresponding to the 219 and 231 nucleotides upstream of *nad4* start codon characterized in chapter II. In *Rfn* flower buds, a shorter 300 base pairs band is detected and sequencing revealed this band corresponds to the product of *Rfn* mediated processing also characterized in chapter II, 205 nucleotides downstream of *nad4* start codon. After confirmation of the presence of the transgene by PCR, the CR-RT-PCR experiment was carried out on young PPR1 and PPR2 transformed tissue that did not show rooting capacities. As seen on figure 4.3, none of the samples tested showed the 300 base pair band corresponding to *Rfn* mediated processing.

Expression of recombinant PPRs

Several PPR proteins have been produced in sufficient quantity by expression in *E. coli* cells to allow for examination of their RNA binding capacity and specificity through electrophoretic mobility shift assays (Haïli, Arnal et al. 2013, Ke, Chen et al. 2013). Furthermore, the production of PPR proteins in *E. coli* has been a system used in the

recent years in order to characterize the functional properties of PPR proteins (Ke, Chen et al. 2013).

Towards this end, the 4 PPR candidates were cloned into destination vector that would allow expression in *E. coli.* Each PPR was amplified from the BAC using primers such that the products would lack mitochondrial targeting sequences but contain attB recombination sites to allow for their cloning through the Gateway system. pDonr207 was used as an entry vector and pDEST17 (6 x His tag fusion) or pDEST-HisMBP (MBP fusion) were used as destination vector. These vectors allow production of high yields of recombinant proteins following IPTG induction and the attachment of a purification tag at the 3' ends of the inserted DNA sequences. Both poly Histidine (not shown) and maltose binding protein fusion clones were successfully produced. As shown in figure 4.4, when soluble and insoluble MBP fusion protein fractions were separated and run on a denaturing polyacrylamide gel, it was revealed that the most of the produced proteins were found in the insoluble fraction as inclusion bodies. This was the case for both polyhistidine (not shown) and MBP fusion proteins (figure 4.4). The level of production of MBP fusion proteins was significantly higher than the polyhistidine fusion proteins and further experimentation was therefore restricted to MBP fusions. Various techniques including expression for longer periods at lower temperature as resolublization in denaturing detergent and re-naturation through dialysis were explored to render the expressed proteins soluble without success.

Discussion

Diverse approaches used for cloning

The challenge in molecular complementation experiments presented in this chapter may reside in the structure of PPR genes. I had difficulty obtaining clones of these genes in the relatively large binary vector (pRD400) employed previously in the laboratory to transform *B. napus* (Brown, Formanova, 2003). Obtaining clones required the screening hundreds of colonies and rendered the task time consuming. The Gateway cloning system allows cloning of a PCR product into an entry vector in one simple step and prevents the instability promoted during the multiple subcloning, digesting and purification steps required in a more traditional cloning strategy. The presence of the toxic *ccmB* gene in non-recombined clones allowed for easier screening of clones containing the insert of interest without screening through the large number of empty ligated vectors. The system allows the subcloning of the inserts into a smaller entry vector (pDonr 207, 5.5kb), which is more easily transformed than large binary vectors as pRD400 (12kb). Indeed such large vectors can be challenging for cells to take up and replicate. The insertion of the gene of interest from the entry vector to the destination, and binary, vector for plant transformation is made through recombinatio Overall, using Gateway technology allowed us to obtain constructs more efficiently.

Modification of the plant transformation protocol

Plant transformation is a well-known reverse genetics technique in a number of model plants including *A. thaliana*. The ability of Agrobacterium to insert a gene of interest into the plant genome is a great tool in molecular biology and has been used in numerous studies for characterization of restorers of fertility (Bentolila, Alfonso et al. 2002, Brown, Formanová et al. 2003, Desloire, Gherbi et al. 2003, Koizuka, Imai et al. 2003). Although of major agricultural and academic interest, *B. napus* transformation is challenging due to the variation in regeneration abilities of the different lines available. A thorough literature search and accumulation of different techniques previously used in the lab allowed me to collect three different protocols for Brassica plant transformation. Co-cultivation with an Agrobacterium solution and tissue regeneration is the technique used in the different protocols, but major differences in the media and materials used can be found. The Bhalla and Singh 2008, study presented the most descriptive method on the co-cultivation of cut cotyledons and was a very useful resource to establish the original protocol used throughout my studies. The media and materials used were from Sparrow, Dale et al. 2006, which describes production of transgenic *B. oleracea.* Transformation was performed using of carbenicillin as an antibiotic for Agrobacterium selection (an effective and widely used antibiotic for selection more available than timentin as described in Sparrow, Dale et al. 2006). That modified protocol proved itself useful to produce transgenic *B. napus* and showed good results in promoting rooting of the young transformed tissue, a step critical in the transformation procedure. With 10% of the initial cotyledons co-cultivated forming young transformants with roots, the efficiency of the method is fairly good. The level of true transformants confirmed with PCR on those plants is 10% out of the rooted plants (1% of the initial cotyledons transformed) and is

quite low The kanamycin resistant binary vector pMDC100 contains a form of the NPTII gene that has been known to present difficulties for selection and transformation in *B. napus* and coould be responsible the low transformation rate. Transformation with a basta resistant binary vector has been carried out but selection of the transformants proved itself difficult with a large majority of the plants dying shortly after co-cultivation preventing any callus formation and plant regeneration. More tests should be carried out with various kanamycin and basta concentrations in the selection stages of the transformation in order to be able to harvest more young transformants that could root and form fully grown plants. The protocols studied, differed on the timing of the application of plant selective agents before outgrowth (during callus formation). The removal of any plant selection before this step could potentially raise the yield of true transformants.

In conclusion, more transformation of the PPR candidates are needed. Positive identification of *Rfn* requires the comparison of the fertility restoration abilities of all 4 candidates of a statistically significant number of transformants and characterization of the processing of the CMS conferring transcript *orf222-nad5c-orf101*.

Rfn specific nad4 processing tested in young transformants

The *Rfn* allele is associated with additional RNA cleavage events in the coding regions of *nad4* and $ccmF_{N2}$ (*formerly ccl1-l*) which are not observed in plants homozygous for the *Rfp* allele or for the non restoring, or universal maintainer genotype *rf* (Li, Jean et al.

1998). As an early detection of *Rfn*, I explored *nad4 Rfn* specific processing and the protocol established throughout RNA 3' end and 5' end mapping studies in chapter II showed itself useful for that application. *Rfn* specific processing of *nad4*, although inseperable from the *Rfn* locus is not sufficient result for identification of the *nap* CMS restorer of fertility. Indeed, *Rfn* genetic mapping in chapter III showed that the *Rfn* containing locus does contain different PPR proteins in close vicinity of each other and the promotion of *nad4 Rfn* specific processing could very well be carried out by a different protein that does not promote restoration fertility by *orf222/nad5c/orf101* processing. Analysis of the sequences around the processing sites in chapter II however highlights a possible common target in *nad4*a and between *orf222* and *nad5*c. Clear identification of *Rfn* is not possible without observation of the fertility restoration phenotype in the progeny coming from a cross with nap CMS individuals as well as the phenotype of *Rfn* processing of the *orf222/nad5c/orf101* transcript. Analysis of *nad4* processing could provide an early screen to get insight on the role of the different PPR proteins present in the *Rfn* containing locus.

Towards characterization of nap CMS fertility restoration mechanism

Protein-protein interaction

Identification of Rfn should include characterization of its way of action. Efforts towards that goal have been made throughout my thesis work, in particular in the identification of transcript termini relevant in *Rfn* mode of action in chapter II. Constructs of the PPR

candidates that could be used for in-vivo expression studies for the purpose of identifying partner proteins and Rfn mode of action.

As highlighted in the literature review in chapter I, restoration of fertility at the post-transcriptional level usually involves a PPR protein. Studies showed that in some cases restoration of fertility involves a number of partner proteins in large protein complex. Immuno-precipitation experiments of mitochondrial fractions in petunia restored plants demonstrated that the PPR restorer protein PPR592 is associated with the matrix side of the inner membrane of the mitochondria in a large protein complex binding the CMS conferring *pcf* RNA, indicating the implication of a number of partner proteins to PPR592 in the restoration process (Gillman, Bentolila et al. 2007). *Rf5*, the restorer of fertility in rice HL-CMS was not found to be able to bind directly to the CMS conferring transcript, *atp6-orfH79* but rather to work in a complex with a glycine rich protein (Hu, Wang et al. 2012). Production of *Rfn* candidate constructs with specific tags is therefore of interest in order to study possible protein-protein interactions in-vivo and gain insight on the fertility restoration process. Gene synthesis of large gene sequences (around 3kb for the PPR candidates) is costly and only 3 out of the 4 PPR candidates were synthesised for *in vivo* expression based on the fact that PPR4 was at the very border of the interval of mapping of *Rfp*. Cloning of HA tagged PPR4 is being carried out through a PCR amplification strategy in order to be able to have all the candidates available for protein-protein interaction investigations.

Production of recombinant proteins

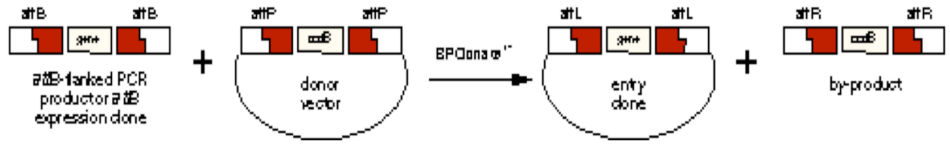
As seen in the results, preliminary experiments of recombinant proteins production have been successfully achieved. MBP fusion proteins in particular were expressed at high levels in *E. coli* but found almost exclusively in the insoluble fraction when subjected to neutral denaturation and gel filtration purification. The *in vitro* re-naturation after purification of an insoluble protein is always difficult as proper folding as and the activity of the native protein are at risk of not being restored. Several modifications of the induction protocols during *E. coli* growth were performed in order to obtain at least a partial soluble fraction of the MBP fusion proteins. Reduction of the growth temperature and elongation of the induction time as well as a large culture volume was thought to allow harvesting of the maximum of soluble proteins but unfortunately the levels of solubilized proteins were not sufficient to allow more *in vitro* studies. Further exploration of *E. coli* protein production should be done.

FIGURES AND TABLES

		Forward	Reverse
Molecular complementation constructs	PPR1	GGGGACAAGTTTGTACAAAAAAGCAGGCTTTTGTCTACGCACGAGCAGT	GGGGACCACTTTGTACAAGAAAGCTGGGTGAGCATCTGCAACCAAGTCA
	PPR2	GGGGACAAGTTTGTACAAAAAAGCAGGCTAAAGGTCAAACCCTATGGCC	GGGGACCACTTTGTACAAGAAAGCTGGGTCGAATCCTAGATACCCGGAC
	PPR3	GGGGACAAGTTTGTACAAAAAAGCAGGCTTCTATTGGAGCTGCTTGTGG	GGGGACCACTTTGTACAAGAAAGCTGGGTGGTTGGCAAGGGACAGTAA
	PPR4	GGGGACAAGTTTGTACAAAAAAGCAGGCTAGATGTGTCAGCTTGCATCG	GGGGACCACTTTGTACAAGAAAGCTGGGTGGTTGTGCATGTCTTCATGG
Recombinant proteins constructs	PPR1	GGGGACAAGTTTGTACAAAAAAGCAGGCTCTAGCGATAGAAATCTCTGTTATAGA	GGGGACCACTTTGTACAAGAAAGCTGGGTTCAAGAAAGCATATCCAAAAAGCT
	PPR2	GGGGACAAGTTTGTACAAAAAAGCAGGCTCTTCTGGAATTACCGATGTGAAAGTC	GGGGACCACTTTGTACAAGAAAGCTGGGCTAAGAGAGCATATCCAAAAATCT
	PPR3	GGGGACAAGTTTGTACAAAAAAGCAGGCTCTTCTGGTGGTAGCGATAGAAAGATG	GGGGACCACTTTGTACAAGAAAGCTGGGTTCAATCCAACGATGATGCTATCTCTCG
	PPR4	GGGGACAAGTTTGTACAAAAAAGCAGGCTCTCTCACCAGCGATAGAAAGATGTCT	GGGGACCACTTTGTACAAGAAAGCTGGGTTCAATCCAACAATGATGATTTGTCTCC
Transformant screen		AATATCACGGGTAGCCAACG	CGCACAATCCCACTATCCTT
cRT-PCR nad4		TCGTTCGGATGGGTGTTCACCCC	AGAAGATCCGCATGCGGAACACGG

Table 4.1: List of the primers used for the various Gateway constructs (molecular complementation and recombinant proteins) as well as for the screen of positive transformants (amplification of the sequence between the 35S promoter and the bacterial kanamycin resistance gene) and cRT-PCR primers to detect *nad4* post-transcriptionnal *Rfn* mediated processing event.

a. BP reaction



b. LR reaction

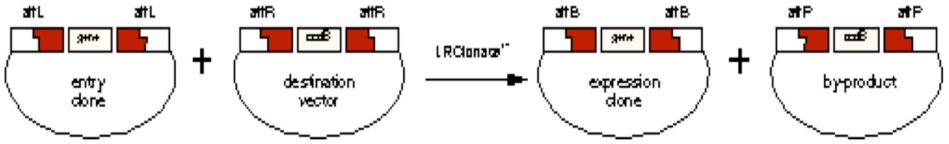


Figure 4.1: Schematic representation of the Gateway cloning system (adapted from the Gateway technology manual https://tools.thermofisher.com/content/sfs/manuals/gatewayman.pdf). The first event called BP reaction (a) allows the recombination of attB repeats in the DNA insert with attP repeats in the vector which allows the insertion of the DNA fragment in an entry vector forming a set of new repeats at the recombination sites called attL. The second reaction called LR (b) allows the recombination of attL repeats with attR ones to form attB and attP repeats in order to transfer the DNA fragment from the entry vector to the destination vector.

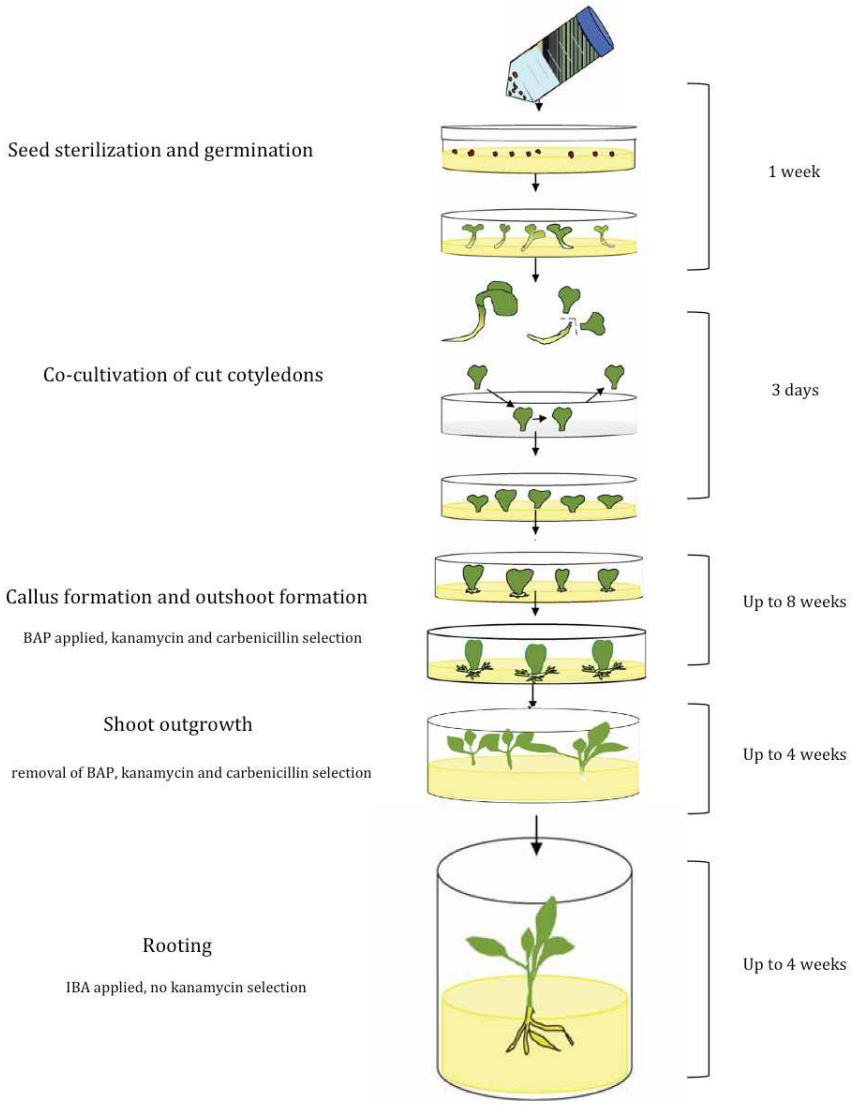


Figure 4.2: Agrobacterium mediated transformation of *B. napus* plants. Diagram adapted from Bhalla and Singh 2008. The protocol was modified in order to simplify the process and reduce the cost of the material used. The use of MS media with different add-ons depending on the stage of the transformation process was the main change in the protocol.

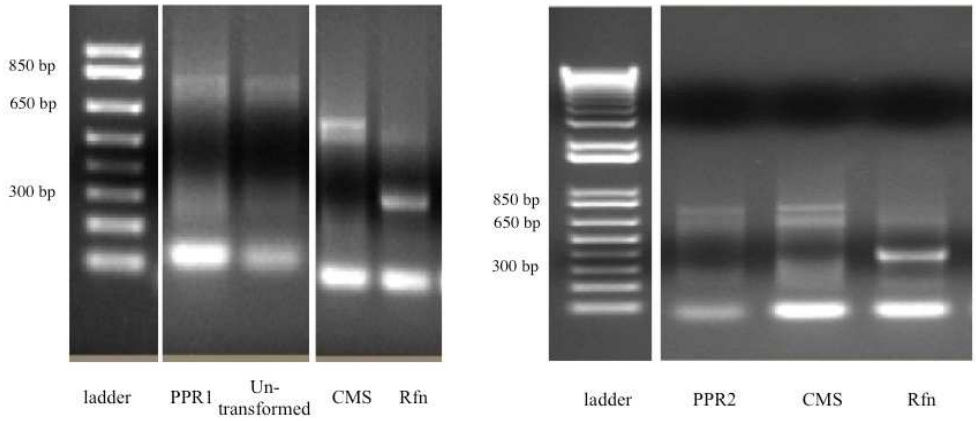


Figure 4.3: CR-RT-PCR on young tissue of PPR1 and PPR2 transformants. *nad4* circularized transcripts ends were amplified (using primers as described in table 4.1) in order to detect the *Rfn* dependent processing. Flower buds of CMS, *Rfn* and untransformed plants were examined as controls. Sequencing was performed to confirm the identity of the PCR products.

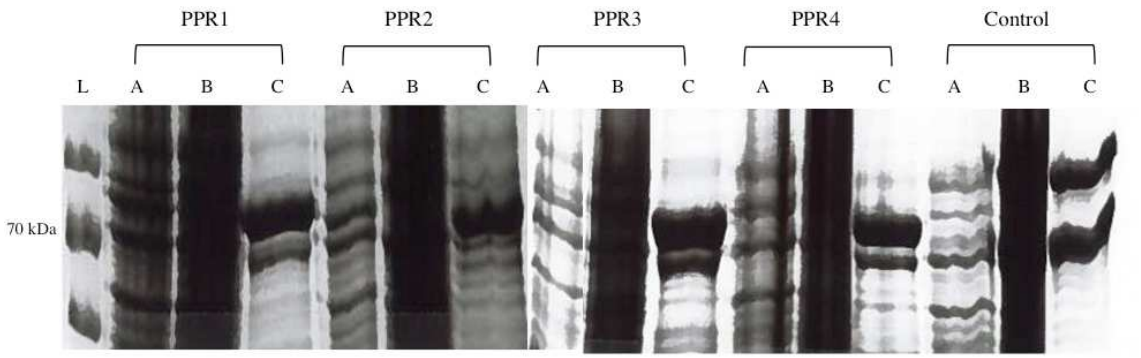


Figure 4.4: Production of recombinant proteins in *E. coli*. The expressions of the four candidates proteins were induced by 50 mM IPTG overnight at 4^{o}C with shaking. Separation of the soluble and insoluble protein fractions was performed by centrifugation at 20 000g for 1 h and the samples were run on an acryl-bis-acrylamide gel and colored with coomasie blue (R250). The gel shows the non-induced control (A), soluble (B) and insoluble (C) fractions. A control MBP fusion recombinant protein was also produced alongside the candidate PPR as an experimental control.

References

Bentolila, S., A. A. Alfonso and M. R. Hanson (2002). "A pentatricopeptide repeat-containing gene restores fertility to cytoplasmic male-sterile plants." Proceedings of the National Academy of Sciences 99(16): 10887-10892.

Bhalla, P. L. and M. B. Singh (2008). "Agrobacterium-mediated transformation of *Brassica napus* and *Brassica oleracea*." Nature protocols 3(2): 181-189.

Brown, G. G., N. Formanová, H. Jin, R. Wargachuk, C. Dendy, P. Patil, M. Laforest, J. Zhang, W. Y. Cheung and B. S. Landry (2003). "The radish *Rfo* restorer gene of Ogura cytoplasmic male sterility encodes a protein with multiple pentatricopeptide repeats." The Plant Journal 35(2): 262-272.

Curtis, M. D. and U. Grossniklaus (2003). "A gateway cloning vector set for high-throughput functional analysis of genes in planta." Plant physiology 133(2): 462-469.

Desloire, S., H. Gherbi, W. Laloui, S. Marhadour, V. Clouet, L. Cattolico, C. Falentin, S. Giancola, M. Renard and F. Budar (2003). "Identification of the fertility restoration locus, *Rfo*, in radish, as a member of the pentatricopeptide-repeat protein family." EMBO reports 4(6): 588-594.

Edwards, K., C. Johnstone and C. Thompson (1991). "A simple and rapid method for the preparation of plant genomic DNA for PCR analysis." Nucleic acids research 19(6): 1349.

Formanová, N., X.-Q. Li, A. M. Ferrie, M. DePauw, W. A. Keller, B. Landry and G. G. Brown (2006). "Towards positional cloning in *Brassica napus*: generation and analysis of doubled haploid *B. rapa* possessing the *B. napus pol* CMS and *Rfp* nuclear restorer gene." Plant molecular biology 61(1-2): 269-281.

Gillman, J. D., S. Bentolila and M. R. Hanson (2007). "The petunia restorer of fertility protein is part of a large mitochondrial complex that interacts with transcripts of the CMS-associated locus." The Plant Journal 49(2): 217-227.

Haïli, N., N. Arnal, M. Quadrado, S. Amiar, G. Tcherkez, J. Dahan, P. Briozzo, C. C. des Francs-Small, N. Vrielynck and H. Mireau (2013). "The pentatricopeptide repeat MTSF1 protein stabilizes the *nad4* mRNA in Arabidopsis mitochondria." Nucleic acids research 41(13): 6650-6663.

Hartley, J. L., G. F. Temple and M. A. Brasch (2000). "DNA cloning using in vitro site-specific recombination." Genome research 10(11): 1788-1795.

Hu, J., K. Wang, W. Huang, G. Liu, Y. Gao, J. Wang, Q. Huang, Y. Ji, X. Qin, L. Wan, R. Zhu, S. Li, D. Yang and Y. Zhu (2012). "The Rice Pentatricopeptide Repeat Protein RF5 Restores Fertility in Hong-Lian Cytoplasmic Male-Sterile Lines via a Complex with the Glycine-Rich Protein GRP162." The Plant Cell 24(1): 109-122.

Ke, J., R.-Z. Chen, T. Ban, X. E. Zhou, X. Gu, M. E. Tan, C. Chen, Y. Kang, J. S. Brunzelle and J.-K. Zhu (2013). "Structural basis for RNA recognition by a dimeric PPR-protein complex." Nature structural & molecular biology 20(12): 1377-1382.

Koizuka, N., R. Imai, H. Fujimoto, T. Hayakawa, Y. Kimura, J. Kohno-Murase, T. Sakai, S. Kawasaki and J. Imamura (2003). "Genetic characterization of a pentatricopeptide repeat protein gene, *orf687*, that restores fertility in the cytoplasmic male-sterile Kosena radish." The plant journal 34(4): 407-415.

Komori, T., S. Ohta, N. Murai, Y. Takakura, Y. Kuraya, S. Suzuki, Y. Hiei, H. Imaseki and N. Nitta (2004). "Map-based cloning of a fertility restorer gene, *Rf-1*, in rice (*Oryza sativa* L.)." The Plant Journal 37(3): 315-325.

Li, X.-Q., M. Jean, B. S. Landry and G. G. Brown (1998). "Restorer genes for different forms of Brassica cytoplasmic male sterility map to a single nuclear locus that modifies transcripts of several mitochondrial genes." Proceedings of the National Academy of Sciences 95(17): 10032-10037.

Mersereau, M., G. J. Pazour and A. Das (1990). "Efficient transformation of *Agrobacterium tumefaciens* by electroporation." Gene 90(1): 149-151.

Moloney, M. M., J. M. Walker and K. K. Sharma (1989). "High efficiency transformation of *Brassica napus* using Agrobacterium vectors." Plant Cell Reports 8(4): 238-242.

Singh, M., N. Hamel, R. Menasaa, X.-Q. Li, B. Young, M. Jean, B. S. Landry and G. G. Brown (1996). "Nuclear Genes Associated With a Single Brassica CMS Restorer Locws Influence Transcripts of Three Different Mitochondrial Gene Regions." Genetics 143(1): 505-516.

Sparrow, P. A., P. J. Dale and J. A. Irwin (2006). "*Brassica oleracea*." Agrobacterium Protocols: 417-426.

CONCLUDING REMARKS

This thesis aims to present the work done towards understanding the mechanisms underlying CMS and fertility restoration in *Brassica napus* with a strong focus on the *nap* CMS system. In the first chapter, we present an extensive literature review on CMS systems in order to understand what changes in the mitochondria can lead to male sterility and how restoration of fertility happens in the various natural systems studied. It is also highlighted that restoration processes implicating post-transcriptionnal mechanisms often involves proteins of a specific group of the P-type PPR family called *RFL* PPRs. These proteins are sujected to diversifying or positive selection which makes them good candidates for the fast evolving restoration processes. I also introduce the different native *B. napus* CMS systems and how I will focus my work on nap CMS and its restorer *Rfn*.

In Chapter II, I present the work done towards understanding the mechanism of restoration of fertility at the post-transcriptional level. By mapping 5' and 3' ends of some *B. napus* mitochondrial RNAs of interest, I was able to detect at the nucleotide level the *Rfn* specific processing events. This allowed the discovery of sequence homology between the different transcript analyzed around the *Rfn* specific processing sites. Moreover, I was able to highlight the fact that very much like in *A. thaliana*, 5' end formation in *B. napus* is induced by a variety of different processing events leading to a multitude of 5' termini whereas the presence of unique 3' ends for each transcripts

studied suggests different post-transcriptional mechanism may be involved. The study of *Rfn* specific processing in *orf222/nad5c/orf101* also allowed the exploration of how *nad5c* would be spliced in a *nap* CMS background. By finding that the processing of this transcript does not appear to disturb group II intron structure, I was able to hypothesis on how the presence of ORF222 could lead to male sterility and how the specific *Rfn* processing events would prevent its toxic effect in restored flower buds. Indeed, *Rfn* processing takes place within the 3' UTR of *orf222* and would destabilize the integrity of the transcript leading to 3' to 5' end exoribonuclease degradation of the released *orf222* transcript whether or not by the intermediate of polyadenylation.

In Chapter III, I present of genetic map of the *Rfn* region, on chromosome A09. The aim of this study was to clarify the relationship of this gene with *Rfp*, the restorer for the other native CMS system (*pol*) in *B. napus*, and to identify a limited number of candidate genes that could then be tested for their capacity to rescue the *nap* CMS trait through transgenic complementation. When it became clear that this locus was enriched in particular subgroup of *RFL* genes, very much like the ones found in *A. thaliana* and *Arabidopsis lyrata*, I explored the sequence relatedness between the *RFL* enriched regions from the different genomes, including *Brassica rapa* and *Brassica oleracea*. This study allowed me to determine that both segmental duplication and retrotransposition processes each played a role in the evolution of this region, with segmental duplication being primarily responsible for the expansion of the family in *A. lyrata*, and retrotransposition playing a more major role in the Brassica genomes.

Study of expression of the different PPR candidates also allowed me to distinguish the best candidates for *Rfn*, especially *PPR4* which is the only PPR not expressed in the CMS flower buds.

Finally, in Chapter IV, I present the work done to identify *Rfn* from the PPR candidates genes through molecular complementation. I was able to develop an efficient plant transformation method. Althouth more work needs to be done to optimise the yield of transformants obtained, I carried out preliminary tests on young putitavely transformed tissue that will prove useful on the future transformants that will be produced to complete molecular complementation and characterization of *Rfn*. Efforts were also made to biochemically characterize properties of the different PPR candidates through recombinant expression in *Escherichia coli* and cloning with expression tags for further *in vivo* purification in order to determine possible partner proteins involved in the restoration mechanism.

To conlude, this work opens possibilities for future research in order to complete the knowledge we have on *B. napus* native CMS systems. Once *Rfn* definitive identification is done through molecular complementation, the preliminary data accumulated in this last chapter should be exploited in order to gain more information on the mechanism of fertility restoration. To do so, investigation on the partner proteins involved should be carried out in order to identify their function and hypothesize an exact fertility restoration mechanism. As it was suggested in previous studies, restoration usually invoves larger protein complexes and only determination of the identity and function of all of the

players in the restoration machinery would allow to have a clear model for the mechanism. This could be achevied by in-vivo expression of the HA tagged PPR constructs obtained during my studies and their purification in their native complex followed for exemple by mass spectroscopy protein identification.

Furthermore, studies of their RNA binding abilities through recombinant expression and purification with RNA binding assays would allow to gain insight on the exact RNA sequences involved in the PPR recognition and confirm the findings of Chapter II. It would also be of interest to compare the data found with a similar study on *Rfp* and its mechanism of fertility restoration. As these two restorer genes have not been genetically dissociated in the past it would be of interest to make the parallel between the function of the two restorers and look at how these two native systems correlate with each other, possibly enlighting mechanims leading to the emergence of new CMS systems.

APPENDICES

Original electronic files from appendices (supplementary tables and figures) available upon request.

Type of Marker	Name	Sequence Information	*B.rapa* Chr9 coordinates	*B.napus* A09 coordinates
SNP	Bn-A09-p31906781	GGCTATATRCTGCAGCATTGTCTCTATGTTTCAGATTAGAGAGAGTAGTGTGTCCCTCGT[A/G]GCTATAGATAAAATGTCCTTATATGAACTACTTTGTTATAGATCCACCAGCCTTATATGA	39747424..39747542	29936556..29936499
SNP	Bn-A09-p32161150	GCTCTWASTATTGAGCTTCTCTCTTCTTCTAGCCTTGGTCATGAACAAGAAGATTATGTC[T/C]ATCCTTGTACTCGAAAGAGACCAAGCTCTCTTTGCTAAAAARGAATCCCTAGCAGCTCGG	39971927..39972043	30024867..30024987
SNP	Bn-A09-p32321534	ACCARGTTTCTTCCGTTGCTACTTCCTTGGCSAGGGATATCGGATCCATTCTCAGAGGTK[A/G]GTACTTTATGGAGGACCTTTATTCATAGACTCTTGGAATCTTGAGGCTGTTGATATGACT	40085903..40086023	30146144..30146264
SNP	Bn-A09-p32665710	TGAGGTTAAAGAGTAGTGCTCCGGAAAGAGACTAAAGAACGAAGATTAAAATTAAAACTA[T/G]CAAGYAGCATTGTCGTCAACAAAATTTGTTGTTCTTACAAACTCAGTAACATACGAGACT	40449407..40449528	30447401..30447521
SNP	Bn-A09-p32787589	ATTGTTAGAACAATCAACTTCATAATTGCAACACTTGTTAAACCAAAATTAACTTGCAAA[A/C]AATACTTGAATACTCTGGTTGGTAGGTGCTTAAATAGAATGTYGTGTATCTATGTAACAG	40549415..40549535	12752841..12752891
SNP	Bn-A09-p32945202	YAGAAACTGGTCGATGAGGTAACCATGTGCACCGTGAACCTCCACCCCATCAAAGCCTAT[A/G]AWTAMCACAAGTGTCATGTTCAGCTAACATYATACATCCAAAGGGGAAAAAGACTTTTCT	40686839..40686958	30688721..30688840
SNP	Bn-A09-p33925635	GATTTAAATATAATAACCTTGCATTCTAATATTTAATGTAATTTAAAACTGTTTGCACKY[T/G]ATGCCAAACTGTTAGAGTACAAAATATCTTAGCGGATAAGCTTGCATGCGGTGTGAGGAG	41612701..41612821	31175811..31175931
ILP	4.4BB	FORWARD: AAGCTTCCACAGGTCCAGAGTAC - REVERSE: GTTCTTACAGTGGCGTAAAGATGAG	41682891..41683383	31233454..31232916
SNP	Bn-A09-p34477765	ATCATTCTTCTTATATTTTTGCTATTTTGTGTTCTTTTTCTCCCCTTCTACAATTATTAT[T/C]TACGTTTACAATTTAGAAATTAACAGGCTCATAGCATATACCAATATCCATTCCTTGTAG	42136594..42136714	31716013..31716133
CAPS	Bra026980 (72)	FORWARD: CGCCGTAATCCACTACACCT - REVERSE: CATTTCGCCATCTTGACAAA	42256231..42256868	31859852..31859197
CAPS	Bra031736 (79)	FORWARD: AGGAGGGTTTGCGTTCTCT - REVERSE: TTGCGCAAACTCTCCGTCTTT	42677700..42677719	32208895..32209749
ILP	Bra031710 (47)	FORWARD: CCGCCGGATCATTTGACAAT - REVERSE: CAAAGGCTGCATTGGACTCA	42875109..42874649	32339045..32339493
SNP	Bn-A09-p35472661	CGAACTACCTCTTCCGATCTATCATTATTATGAGTTTGTATCCTTGTCTCTTCTTTCTCT[A/G]CCATTACAAAGAGAACATACAAATCTGAAAGTTACCTTAAGATTCACGATGTTTTGTTAC	43239747..43239866	32615358..32615477
SNP	Bn-A09-p35505057	GTTCATCGTCAGGCWTTGTCTCCCGTTTGGATCAAACTGCATATCCCACAGTACAAAGAG[A/G]ATAAGCTCAGTGTGTTAGATAGTCTACATGTTTTTTTTTATACGGCCGTAATCCCCCAAA	43269934..43270054	32651585..32651705

Supplementary table 3.1. Information on the molecular markers used for genetic mapping of *Rfn*. Single nucleotide polymorphisms are indicated with their flanking sequences. Primers used to amplify ILPs and CAPs are indicated. HaeIII (recognition site GGCC) and AluI (recognition site AGCT) restriction enzymes were used for detecting CAPs markers. The location of each marker in the *B. napus* chromosome A09 and *B. rapa* chromosome 09 are also indicated. The marker highlighted in red represent a locus that maps within the Rfn containing BAC.

Supplementary table 3.2. Annotation of the 180kb *Rfn* containing BAC. Positions of the genes were predicted using Softberry online tool. Each predicted genes were screen for possible protein functions based on *B. rapa* and *A. thaliana* homologs. Genes predicted to encode for PPR proteins are noted in red. The grey highlighted genes are the F-box and Cytochrome P_{450} encoding genes that may have been involved in the sequence rearrangement.

ORF coordinates in BAC	Otholog in Brassica rapa	Coordinates of Brassica rapa ortholog	Annontation of Brassica rapa ortholog	rabidopsis thaliana homolog/ortholog and its annotatio	Corresponding gene in Brassica napus	Coordinates in Brassica napus genome
623..2065	-	-	Sequence from cloning vector			
2309..5102	-	-	Sequence from cloning vector			
5153..7185	-	-	Sequence from cloning vector			
7229..7606	-	-	Sequence from cloning vector			
7737..13276	-	-	Sequence from cloning vector			
13460..17437	Brara.I05107	A09:42181040..42183404 reverse	Pollen allergen // Rare lipoprotein A (RlpA)-like double-psi beta-barrel	AT1G12560.1 - expansin A7	GSBRNA2T00035440001	Chr A09 31,757,549-31,759,431
17720..21166	Brara.I05106	A09:42176532..42179064 reverse	Serine/Threonine protein kinase	AT1G12580.1 - PEP carboxylase-related kinase 1	GSBRNA2T00035440001	Chr A09 31,757,549-31,759,431
21234..22387	Brara.I05105	A09:42174838..42176099 forward	AP2 domain protein	AT1G12610.1 - DDF1/AP2 transcription factor	GSBRNA2T00035441001	Chr A09 31,756,069-31,756,686
26208..28395	Brara.I05104	A09:42167651..42169580 reverse	Uncharacterized conserved protein (DUF947)	AT1G12650.4 - Uncharacterized conserved protein (DUF	GSBRNA2T00035442001	Chr A09 31,749,036-31,750,547
30193..31185	Brara.I05103	A09:42159258..42159752 reverse	No annotation	AT1G12660.1 -plant thionin (PR-13) protein*	GSBRNA2T00035443001	Chr A09 31,746,192-31,746,676
32430..33866	Brara.I05103	A09:42159258..42159752 reverse	No annotation	AT1G12660.1 -plant thionin (PR-13) protein*	GSBRNA2T00035445001	Chr A09 31,743,308-31,743,652
34324..34984	Brara.I05102	A09:42157228..42158114 reverse	No annotation	AT1G12660.1 -plant thionin (PR-13) protein*	GSBRNA2T00035446001	Chr A09 31,741,995-31,742,488
35887..37067	Brara.I05101	A09:42151146..42155482 reverse	No annotation	AT1G12660.1 -plant thionin (PR-13) protein*	GSBRNA2T00035447001	Chr A09 31,740,024-31,740,969
39189..39997	Brara.I05101	A09:42151146..42155482 reverse	No annotation	AT1G12660.1 -plant thionin (PR-13) protein*	GSBRNA2T00035448001	Chr A09 31,734,349-31,734,938
43159..45379	Brara.I05099	A09:42146317..42148884 forward	Serine/Threonine protein kinase	AT1G12680.1 - PEP carboxylase-related kinase 2	GSBRNA2T00035449001	Chr A09 31,731,638-31,733,427
45936..48971	Brara.I05098	A09:42143232..42145483 forward	PPR protein	AT1G12300.1 - Tetratricopeptide repeat (TPR)-like	GSBRNA2T00035452001	Chr A09 31,728,897-31,728,986
49704..52029	Brara.I05097	A09:42140319..42142613 forward	PPR protein	AT1G12700.1 - PPR protein RNA processing factor 1 (RP	GSBRNA2T00035458001	Chr A09 31,721,497-31,721,706
56704..58786	Brara.I05108	A09:42189680..42190951 reverse	SF15 - 2-Hydroxyacid Dehydrogenase	AT1G12550.1 - D-isomer specific 2-hydroxyacid dehydro	GSBRNA2T00035439001	Chr A09 31,765,257-31,766,329
60356..62769	Brara.I05096	A09:42132050..42134676 reverse	Cytochrome P450 CYP4/CYP19/CYP26 subfamilies	AT1G12740.1 - cytochrome P450, family 87, subfamily A	GSBRNA2T00035460001	Chr A09 31,711,882-31,713,957
64874..67667	Brara.I05094	A09:42126485..42128558 reverse	PF01535 - PPR repeat	AT1G12775.1 - Pentatricopeptide repeat (PPR)	GSBRNA2T00035480001	Chr A09 31,651,648-31,653,159
68651..78336	Brara.I05093	A09:42122405..42125340 forward	GRR1-Related, ARATH	AT1G12820.1 - auxin signaling F-box 3	GSBRNA2T00035465001	Chr A09 31,697,586-31,704,467
79363..80423	Brara.I05092	A09:42118593..42119485 forward	No annotation	AT1G12830.1 -unknown function, stress-related	GSBRNA2T00035466001	Chr A09 31,693,374-31,693,916
80585..83346	Brara.I05091	A09:42115502..42118401 reverse	V-type H+-transporting ATPase subunit C	AT1G12840.1 - vacuolar ATP synthase subunit C (VATC)	GSBRNA2T00035467001	Chr A09 31,689,003-31,693,085
85123..87275	Brara.I05090	A09:42109831..42112149 forward	F-Box family protein	AT1G12855.1 - F-box family protein	GSBRNA2T00035468001	Chr A09 31,686,778-31,688,352
88382..89171	Brara.I05089	A09:42108221..42109134 reverse	AP2 domain	AT1G12890.1 - ERF/AP2 transcription factor	GSBRNA2T00035469001	Chr A09 31,684,630-31,685,313
89178..91757	Brara.I05088	A09:42099541..42100792 forward	WD repeat-containing protein 68	AT1G12910.1 - Light regulated WD40 protein	GSBRNA2T00035470001	Chr A09 31,682,547-31,684,453
91921..100130	Brara.I05087	A09:42092767..42099324 forward	Transportin 3 and Importin 13 // Uncharacterized	AT1G12930.1 - ARM repeat superfamily protein	GSBRNA2T00035472001	Chr A09 31,673,840-31,680,143
101012..104355	Brara.I05086	A09:42088444..42091559 reverse	Multidrug and toxin extrusion protein 2	AT1G12950.1 - root hair specific 2	GSBRNA2T00035473001	Chr A09 31,669,499-31,672,610
104843..105647	Brara.I05085	A09:42087153..42087732 forward	S locus-related glycoprotein 1 binding pollen coat protein (SLR1-BP)	AT4G13095.1 - low-molecular-weight cysteine-rich	GSBRNA2T00035474001	Chr A09 31,668,170-31,668,749
107452..109720	Brara.I05084	A09:42083265..42085380 reverse	Leucine-rich repeat containing protein	AT1G12970.1 - PIRL3, Ras-group-related LRR protein	GSBRNA2T00035475001	Chr A09 31,664,415-31,666,242
113303..114747	Brara.I05083	A09:42073655..42074542 reverse	AP2 domain	AT1G12980.1 - ERF/AP2 transcription factor	GSBRNA2T00035476001	Chr A09 31,659,444-31,660,328
116290..119506	Brara.I05082	A09:42069403..42071977 forward	Translation initiation factor 4H family	AT1G13020.1 - eIF4B2eukaryotic initiation factor 4B2	GSBRNA2T00035477001	Chr A09 31,655,253-31,657,363
120517..122632	Brara.I05081	A09:42066061..42068772 forward	PPR protein	AT1G12620.1 - Pentatricopeptide repeat (PPR) su	GSBRNA2T00035480001	Chr A09 31,651,648-31,653,159
122709..129134	Brara.I05080 - Brara.I05079	A09:42059178..42061504 reverse (79) A0	20S proteasome subunit beta 5 / coilin	AT1G13030.1 - coilin protein	GSBRNA2T00035482001	Chr A09 31,647,745-31,650,653
129211..131880	Brara.I05078	A09:42055901..42057644 reverse	Cytochrome P450 CYP2 subfamily	AT1G13080.1 - cytochrome 450 71B family protein	GSBRNA2T00035484001	Chr A09 31,641,898-31,643,517
133774..135688	Brara.I05077	A09:42050442..42051933 forward	F-box domain // FBD	AT1G13780.1 - F-box/RNI-like/FBD-like domains-contai	GSBRNA2T00035485001	Chr A09 31,638,256-31,639,865
**135871..137783	Brara.I05077	A09:42050442..42051933 forward	F-box domain // FBD	AT1G13780.1 - F-box/RNI-like/FBD-like domains-contai	GSBRNA2T00035486001	Chr A09 31,636,314-31,637,541
139596..140596	Brara.I05076	A09:42046105..42047915 reverse	Cytochrome P450 CYP2 subfamily	AT1G13090.1 - cytochrome 450 71B family protein	GSBRNA2T00035488001	Chr A09 31,633,885-31,634,433
140889..142656	Brara.I05076	A09:42048600..42048914 reverse	Cytochrome P450 CYP2 subfamily	AT1G13100.1 - cytochrome 450 71B family protein	GSBRNA2T00035489001	Chr A09 31,631,632-31,633,179
142710..148126	Between Brara.I05076 and Brara.I05075	A09:42043487..42043681 forward	-	AT1G13790 - XH/XS domain containing protein	GSBRNA2T00035490001	Chr A09 31,628,436-31,631,056
153613..162110	Brara.I05041	A09:41800071..41803756 reverse	GLE-1-RELATED	AT1G13120.1 - GLE-1-like nuclear pore protein	GSBRNA2T00035496001	Chr A09 31,609,559-31,612,911
164785..165365	Brara.I05040	A09:41798160..41800033 forward	Cytochrome P450 CYP2 subfamily	AT1G13090.1 - cytochrome 450 71B family protein	GSBRNA2T00035495001	Chr A09 31,613,201-31,615,716
165558..167429	Brara.I05039	A09:41795095..41797298 forward	Cytochrome P450 CYP2 subfamily	AT1G13100.1 - cytochrome 450 71B family protein	GSBRNA2T00035495001	Chr A09 31,613,201-31,615,716
167474..169928	Brara.I05040	A09:41798160..41800033 forward	Cytochrome P450 CYP2 subfamily	AT1G13110.1 - cytochrome 450 71B family protein	GSBRNA2T00035494001	Chr A09 31,616,769-31,618,764
171775..173968	Brara.I05038	A09:41790001..41793585 reverse	F-box domain	AT1G13780.1 - F-box/RNI-like/FBD-like domains-	GSBRNA2T00035492001	Chr A09 31,621,196-31,622,766

*Represents a small gene/pseudogene cluster encoding thionins, pathogenesis related (PR-13) proteins

**Gray shading indicates a sequences surrounding a portion of the *B. napus genome* that has rearranged through sequence inversion following its divergence from *B. rapa* cv. "Chiifu-401"

Supplementary table 3.3. Annotation of the 600kb *B. napus* chromosome A09 region mapped to contain *Rfn*. Positions of the genes were predicted using Softberry online tool. Each predicted genes were screen for possible protein functions based on *B. rapa* and *A. thaliana* homologs. Genes predicted to encode for PPR proteins are noted in red. The dark grey highlighted genes are the F-box and Cytochrome P_{450} encoding genes that may have been involved in the sequence rearrangement. The light grey highlighted genes represent the *Rfn* containing BAC region.

Coordinates in the Brassica napus A09 chromosome	Arabidopsis thaliana homolog and annotation	Brassica rapa ortholog	Annotation of Brassica rapa ortholog
Marker 4.4BB			
31233669-31235758	AT1G14140.1 | Symbols: | Mitochondrial substrate ca	Brara.I05012	Mitochondrial fatty acid anion carrier protein/Uncoupling protein
31236919-31238279	AT1G14130.1 | Symbols: | 2-oxoglutarate (2OG) and	Brara.I05013	OXIDOREDUCTASE, 2OG-FE(II) OXYGENASE FAMILY PROTEIN
31238754-31240440	AT1G14130.1 | Symbols: | 2-oxoglutarate (2OG) and	Brara.I05014	OXIDOREDUCTASE, 2OG-FE(II) OXYGENASE FAMILY PROTEIN
31243608-31247671	AT1G14080.1 | Symbols: FUT6, ATFUT6 | fucosyltrar	Brara.I05015	xyloglucan fucosyltransferase
31247793-31249900	AT1G14100.1 | Symbols: FUT8 | fucosyltransferase 8 |	Brara.I05016	xyloglucan fucosyltransferase
31250350-31252904	AT1G14080.1 | Symbols: FUT6, ATFUT6 | fucosyltrar	Brara.I05017	xyloglucan fucosyltransferase
31252927-31254714	AT1G14080.1 | Symbols: FUT6, ATFUT6 | fucosyltrar	Brara.I05018	xyloglucan fucosyltransferase
31258082-31263076	AT1G14040.1 | Symbols: | EXS (ERD1/XPR1/SYG1)	Brara.I05020	XENOTROPIC AND POLYTROPIC RETROVIRUS RECEPTOR 1 (PROTEIN SYG1 HOMOLOG)(XENOTROPIC AND POLYTROPIC MURINE LEUKEMIA VIRUS RECEPT
31265194-31268956	AT1G14040.1 | Symbols: | EXS (ERD1/XPR1/SYG1)	-	SNF7 domain containing protein, putative, expressed LOC_Os07g29630.1 LOC_Os07g29630 LOC_Os07g29630.1
31269180-31270171	AT5G22950.1 | Symbols: VPS24.1 | SNF7 family prot	-	SNF7 domain containing protein, putative, expressed LOC_Os07g29630.1 LOC_Os07g29630 LOC_Os07g29630.1
31273112-31273764	AT4G12382.2 | Symbols: | F-box family protein | chr4	-	Athaliana_PAC2_0_167_peptide
31273795-31276644	AT4G23640.1 | Symbols: TRH1, ATKT3, KUP4 | Pota	-	
31276709-31278874	AT1G05770.1 | Symbols: | Mannose-binding lectin su	-	
31278942-31281898	AT5G45430.2 | Symbols: | Protein kinase superfamily	Brara.I01895	MITOGEN-ACTIVATED PROTEIN KINASE
31282932-31285291	AT1G14040.1 | Symbols: | EXS (ERD1/XPR1/SYG1)	Brara.I05021	XENOTROPIC AND POLYTROPIC RETROVIRUS RECEPTOR 1 (PROTEIN SYG1 HOMOLOG)(XENOTROPIC AND POLYTROPIC MURINE LEUKEMIA VIRUS RECEPT
31289681-31291803	AT1G14030.1 | Symbols: | Rubisco methyltransferase	Brara.I05022	Ribulose-bisphosphate carboxylase]-lysine N-methyltransferase
31293245-31295601	AT1G14030.1 | Symbols: | Rubisco methyltransferase	Brara.I05023	EMP24/GP25L-RELATED
31295838-31297209	AT1G14000.1 | Symbols: VIK | VH1-interacting kinas	Brara.I05024	ANKYRIN-KINASE
31297582-31301744	AT1G13980.2 | Symbols: GN | sec7 domain-containing	Brara.I05025	Protein of unknown function (DUF1336)
31307586-31308664	AT1G13950.1 | Symbols: EIF-5A, ELF5A-1, ATELF5.	Right after Brara.I05026	eukaryotic translation initiation factor 5A, putative, expressed LOC_Os12g32240.1 LOC_Os12g32240 LOC_Os12g32240.1
31308683-31309953	AT5G18570.1 | Symbols: EMB269, ATOBGC, CPSAF	Right after Brara.I05026	eukaryotic translation initiation factor 5A, putative, expressed LOC_Os03g55150.1 LOC_Os03g55150 LOC_Os03g55150.1
31310158-31311133	AT5G24510.1 | Symbols: | 60S acidic ribosomal prote	Right after Brara.I05026	eukaryotic translation initiation factor 5A, putative, expressed LOC_Os03g55150.1 LOC_Os03g55150 LOC_Os03g55150.1
31312251-31316605	AT1G69526.2 | Symbols: | S-adenosyl-L-methionine-(	Brara.I05027	Plant protein of unknown function (DUF863)
31316702-31319858	AT1G13940.1 | Symbols: | Plant protein of unknown f	Brara.I05027	Plant protein of unknown function (DUF863)
31319941-31321946	AT1G13940.1 | Symbols: | Plant protein of unknown f	Brara.I05028	DNA BINDING / MAGNESIUM ION BINDING / NUCLEASE
31321968-31324434	AT1G13900.1 | Symbols: | Purple acid phosphatases s	Brara.I05029	ACID PHOSPHATASE RELATED
31324815-31326249	AT1G13900.1 | Symbols: | Purple acid phosphatases s	Brara.I05030	synaptosomal-associated protein, 29kDa
31326474-31329690	AT1G13880.1 | Symbols: | ELM2 domain-containing |	Brara.I05031	A09:41765415..41767606 reverse
31330973-31332907	AT3G63006.1 | Symbols: | pre-tRNA | chr3:23285754	Brara.I05033	PF07983 - X8 domain
31333264-31333769	AT1G13830.1 | Symbols: | Carbohydrate-binding X8 (	Brara.I05034	PF07983 - X8 domain
31334110-31340783	AT1G12300.1 | Symbols: | Tetratricopeptide repeat (T	Brara.I05036	PPR repeat
31341935-31344412	AT1G13790.1 | Symbols: | XH/XS domain-containing	Scaffold14676:727..1359	XH/XS domain-containing protein AT1G13790.1 AT1G13790 AT1G13790.1
31344495-31348775	AT1G13780.1 | Symbols: | F-box/RNI-like/FBD-like (	Brara.I05072	F-box domain // FBD
31348889-31349318	AT1G13780.1 | Symbols: | F-box/RNI-like/FBD-like (	Brara.I05077	F-box domain // FBD
31349337-31349581	AT5G44940.1 | Symbols: | F-box/RNI-like superfamil	Brara.I05077	F-box domain // FBD
31349890-31351747	AT5G27750.1 | Symbols: | F-box/FBD-like domains c	Brara.I05077	F-box domain // FBD
31356639-31357312	AT1G13780.1 | Symbols: | F-box/RNI-like/FBD-like (	Brara.I05077	F-box domain // FBD
31357551-31363769	AT1G13780.1 | Symbols: | F-box/RNI-like/FBD-like (	Brara.I05077	F-box domain // FBD
31366466-31370206	AT1G80850.1 | Symbols: | DNA glycosylase superfan	-	-
31370402-31374103	AT1G13750.1 | Symbols: | Purple acid phosphatases s	Brara.I05068	ACID PHOSPHATASE RELATED
31381404-31386489	AT5G11550.1 | Symbols: | ARM repeat superfamily p	-	
31387175-31391028	AT1G76950.1 | Symbols: PRAF1 | Regulator of chrom	Brara.I05066	Cytochrome P450 CYP2 subfamily
31389083-31391028	AT1G76950.1 | Symbols: PRAF1 | Regulator of chrom	Brara.I05066	Cytochrome P450 CYP2 subfamily
31399312-31400440	AT3G47770.1 | Symbols: ATATH5, ATH5, ABCA6 | A	-	
31401539-31402804	AT4G27270.1 | Symbols: | Quinone reductase family pa	LOC_Os08g04460.1_GX	NADPH-dependent FMN reductase domain containing protein, expressed LOC_Os08g04460.1 LOC_Os08g04460 LOC_Os08g04460.1
31405371-31407855	AT4G19960.2 | Symbols: KUP9 | K+ uptake permease :	41983361..41985414 (- strand)	
31410874-31411778	AT5G23030.1 | Symbols: TET12 | tetraspanin12 | chr5	Brara.I05065	6-phosphogluconolactonase.
31412359-31414551	AT1G13700.1 | Symbols: PGL1 | 6-phosphogluconola	Brara.I05065	6-phosphogluconolactonase
31416461-31417981	AT1G13610.2 | Symbols: | alpha/beta-Hydrolases supe	A09:41971694..41971768	alpha/beta-Hydrolases superfamily protein AT4G24760.1 AT4G24760 AT4G24760.1
31418763-31421184	AT5G43065.1 | Symbols: | transposable element gene	-	
31426848-31430022	AT1G74800.1 | Symbols: | Galactosyltransferase fami	Brara.I05065	6-phosphogluconolactonase
31430316-31434043	AT1G68710.1 | Symbols: | ATPase E1-E2 type family	-	
31435033-31438461	AT3G60965.1 | Symbols: | transposable element gene	-	
31440947-31443467	AT1G13600.1 | Symbols: AtbZIP58, bZIP58 | basic leu	Brara.I05064	ATBZIP48 (ARABIDOPSIS THALIANA BASIC LEUCINE-ZIPPER 48), DNA BINDING / TRANSCRI
31467106-31468799	AT1G13450.3 | Symbols: GT-1 | Homeodomain-like su	Brara.I01694	PTHR22595:SF7 - CLASS IV CHITINASE
31468806-31469696	AT1G13450.3 | Symbols: GT-1 | Homeodomain-like su	Brara.I05056	Transcription factor GT-2 and related proteins, contains trihelix DNA-binding/SANT domain
31484824-31488406	AT1G13410.1 | Symbols: | Tetratricopeptide repeat (T	Brara.I05052	A09:41910758..41912647 reverse
31488558-31498130	AT2G37550.2 | Symbols: ASP1, AGD7 | ARF-GAP do	Brara.I05054	A09:41924260..41924679 forward
31498143-31500073	AT1G13340.1 | Symbols: | Regulator of Vps4 activity	Brara.I05054	A09:41924260..41924679 forward
31501726-31506017	AT1G13290.1 | Symbols: DOT5, WIP6 | C2H2-like zin	Brara.I05051	SERINE/THREONINE-PROTEIN KINASE RIO

31506073-31510717	AT5G03495.1 | Symbols: | RNA-binding (RRM/RBD,	Brara.I05051	SERINE/THREONINE-PROTEIN KINASE RIO
31514903-31516775	AT5G16550.1 | Symbols: | unknown protein; Has 302	-	expressed protein LOC_Os03g50870.1 LOC_Os03g50870 LOC_Os03g50870.1
31517777-31520135	AT4G07810.1 | Symbols: | transposable element gene	Brara.H01023	
31521143-31523861	AT5G23350.1 | Symbols: | GRAM domain-containing	Brara.I01913	Pollen proteins Ole e I like
31528564-31532195	AT5G59510.1 | Symbols: RTFL5, DVL18 | ROTUNDI	Brara.I03514	calcium-dependent protein kinase
31537412-31541571	AT5G65400.1 | Symbols: | alpha/beta-Hydrolases sup	-	
31541800-31543487	AT4G04180.1 | Symbols: | P-loop containing nucleosi	Brara.H01244	BASIC HELIX-LOOP-HELIX (BHLH) FAMILY PROTEIN
31543698-31545571	AT5G55300.3 | Symbols: TOP1ALPHA | DNA topoisc	Brara.I05049	RAV-like factor
31545614-31546712	AT5G58180.2 | Symbols: ATYKT62 | Synaptobrevin fi	Brara.I05049	RAV-like factor
31546873-31548379	AT2G36720.1 | Symbols: | Acyl-CoA N-acyltransferas	Brara.I05049	RAV-like factor
31548593-31552595	AT5G48240.3 | Symbols: | unknown protein; FUNCTI	Brara.I05049	RAV-like factor
31556257-31556974	AT5G52500.1 | Symbols: | unknown protein; BEST A	Brara.I05048	F-box domain // F-box associated
31557046-31558650	AT1G13230.1 | Symbols: | Leucine-rich repeat (LRR)	Brara.I05049	RAV-like factor
31565512-31568170	AT1G13210.1 | Symbols: ACA.1 | autoinhibited Ca2+/	Brara.I05047	Phospholipid-translocating ATPase
31568796-31571096	AT1G13210.1 | Symbols: ACA.1 | autoinhibited Ca2+/	Brara.I05047	Phospholipid-translocating ATPase
31573310-31575750	AT1G13195.2 | Symbols: | RING/U-box superfamily p	Brara.I05046	Predicted E3 ubiquitin ligase
31576539-31580320	AT1G13160.1 | Symbols: | ARM repeat superfamily p	Brara.I05045	HSDA/SDA1-RELATED
31580931-31584907	AT1G12210.1 | Symbols: RFL1 | RPS5-like 1 | chr1:41	Brara.I05044	disease resistance protein RPS5
31593741-31598027	AT1G11330.2 | Symbols: | S-locus lectin protein kinas	Brara.I05172	LEUCINE-RICH REPEAT RECEPTOR-LIKE PROTEIN KINASE
31605017-31607008	AT4G03640.1 | Symbols: | transposable element gene	Brara.I05041	GLE-1-RELATED
31607085-31608371	AT1G13130.1 | Symbols: | Cellulase (glycosyl hydrol	sa|LOC_Os04g40490.1_G2	glycosyl hydrolase family 5 protein, putative, expressed LOC_Os04g40490.1 LOC_Os04g40490 LOC_Os04g40490.1
31609030-31611265	AT1G13120.1 | Symbols: emb1745 | embryo defective	Brara.I05041	PTHR12960 - GLE-1-RELATED
31611928-31614571	AT1G13120.1 | Symbols: emb1745 | embryo defective	Brara.I05041	PTHR12960 - GLE-1-RELATED
31616434-31617307	AT1G13110.1 | Symbols: CYP71B7 | cytochrome P45(	Brara.I05040	Cytochrome P450 CYP2 subfamily
31617810-31619049	AT1G13110.1 | Symbols: CYP71B7 | cytochrome P45(	Brara.I05039	Cytochrome P450 CYP2 subfamily
31630543-31631744	AT1G13790.1 | Symbols: | XH/XS domain-containing	Brara.K01046	XS domain // XH domain
31634844-31635968	AT3G63070.1 | Symbols: | Tudor/PWWP/MBT domai	Brara.I05077	F-box domain // FBD
31636054-31638161	AT1G13780.1 | Symbols: | F-box/RNI-like/FBD-like d	Brara.I05077	F-box domain // FBD
31638246-31643536	AT1G13080.1 | Symbols: CYP71B2 | cytochrome P45(	Brara.I05078	Cytochrome P450 CYP2 subfamily
31643743-31646112	AT1G13060.2 | Symbols: PBE1 | 20S proteasome beta	Brara.I05079	20S proteasome subunit beta 5
31646471-31647509	AT3G26340.1 | Symbols: | N-terminal nucleophile am	Brara.I05079	20S proteasome subunit beta 5
31649703-31652523	AT1G12775.1 | Symbols: | Pentatricopeptide repeat (P	Brara.I05081	PPR repeat
31653002-31653939	AT1G12775.1 | Symbols: | Pentatricopeptide repeat (P	Brara.I05081	PPR repeat
31655335-31656053	AT1G13020.1 | Symbols: EIF4B2 | eukaryotic initiatio	Brara.I05082	TRANSLATION INITIATION FACTOR 4H FAMILY
31658308-31659072	AT1G12990.1 | Symbols: | beta-1,4-N-acetylglucosam	between 82 and 83	
31659129-31660426	AT1G12980.1 | Symbols: ESR1, DRN | Integrase-type	Brara.I05083	AP2 domain
31676463-31679334	AT1G12930.1 | Symbols: | ARM repeat superfamily p	Brara.I05087	TRANSPORTIN 3 AND IMPORTIN 13 // UNCHARACTERIZED
31679364-31682880	AT1G12910.1 | Symbols: ATAN11, LWD1 | Transducii	Brara.I05088	WD repeat-containing protein 68
31685584-31688415	AT1G12880.1 | Symbols: atnudt12, NUDT12 | nudix h	Brara.I05090	F-BOX FAMILY PROTEIN
31688531-31694527	AT2G27120.1 | Symbols: POL2B, TIL2 | DNA polymc	Brara.I05091	V-type H+-transporting ATPase subunit C
31695160-31696383	AT1G15680.1 | Symbols: | F-box family protein | chr1	between 92 and 93	
31696465-31697810	AT1G12820.1 | Symbols: AFB3 | auxin signaling F-bo	Brara.I05093	GRR1-RELATED, ARATH
31699291-31701316	AT1G12820.1 | Symbols: AFB3 | auxin signaling F-bo	between 32 and 33	
31702255-31705698	AT1G12820.1 | Symbols: AFB3 | auxin signaling F-bo	Brara.I05093	GRR1-RELATED, ARATH
31706049-31709318	AT1G12775.1 | Symbols: | Pentatricopeptide repeat (P	Brara.I05094	PPR repeat
31710989-31717355	AT1G12740.2 | Symbols: CYP87A2 | cytochrome P45(	Brara.I05096	Cytochrome P450 CYP4/CYP19/CYP26 subfamilies
31717689-31719522	AT5G04460.2 | Symbols: | RING/U-box superfamily p	before 97	
31721376-31724966	AT1G12300.1 | Symbols: | Tetratricopeptide repeat (T	Brara.I05097	PPR repeat
31725084-31729422	AT1G12300.1 | Symbols: | Tetratricopeptide repeat (T	Brara.I05098	PPR repeat
31729581-31730696	AT1G12700.1 | Symbols: | ATP binding;nucleic acid b	Brara.I05098	PPR repeat
31731128-31731750	AT1G12680.1 | Symbols: PEPKR2 | phosphoenolpyru	Brara.I05099	SERINE/THREONINE-PROTEIN KINASE
31731963-31732572	AT1G12680.1 | Symbols: PEPKR2 | phosphoenolpyru	Brara.I05099	SERINE/THREONINE-PROTEIN KINASE
31732720-31733421	AT1G12680.1 | Symbols: PEPKR2 | phosphoenolpyru	Brara.I05099	SERINE/THREONINE-PROTEIN KINASE
31734986-31735737	AT5G26050.1 | Symbols: | Plant self-incompatibility p	Brara.I05101	A09:42151146..42155482 reverse
31735857-31736926	ATMG01360.1 | Symbols: COX1 | cytochrome oxidas	Brara.I05101	A09:42151146..42155482 reverse
31736936-31737821	AT1G12672.2 | Symbols: | unknown protein; LOCATI	Brara.I05101	A09:42151146..42155482 reverse
31738982-31741782	AT1G12669.1 | Symbols: | Encodes a Plant thionin fan	Brara.I05102	A09:42157228..42158114 reverse
31741829-31743235	AT1G12660.1 | Symbols: | Predicted to encode a PR (	Brara.I05103	A09:42159258..42159752 reverse
31743284-31747558	AT1G12650.4 | Symbols: | unknown protein; FUNCTI	Brara.I05103	A09:42159258..42159752 reverse
31748780-31749595	AT1G12650.4 | Symbols: | unknown protein; FUNCTI	Brara.I05104	Uncharacterized conserved protein
31750434-31752279	AT1G12650.4 | Symbols: | unknown protein; FUNCTI	Brara.I05104	Uncharacterized conserved protein
31756912-31758032	AT1G12580.1 | Symbols: PEPKR1 | phosphoenolpyru	Brara.I05106	SERINE/THREONINE-PROTEIN KINASE

31763383-31766633	AT1G12550.1 | Symbols: | D-isomer specific 2-hydro	Brara.I05108	2-HYDROXYACID DEHYDROGENASE
31766913-31767522	AT2G07213.1 | Symbols: | other RNA | chr2:2996582	Brara.I05109	2-HYDROXYACID DEHYDROGENASE
31768095-31771305	AT4G36580.1 | Symbols: | AAA-type ATPase family p	Brara.G02592	4-coumarate--CoA ligase
31771469-31772734	AT3G15570.1 | Symbols: | Phototropic-responsive NP	-	
31773764-31780203	AT1G12550.1 | Symbols: | D-isomer specific 2-hydro	Brara.I05108	2-HYDROXYACID DEHYDROGENASE
31781301-31791526	AT1G12520.3 | Symbols: ATCCS, CCS | copper chape	Brara.I05111	A09:42199480..42200996 forward
31791586-31792558	AT3G17540.1 | Symbols: | F-box and associated inter	-	
31793896-31796869	AT1G12440.2 | Symbols: | A20/AN1-like zinc finger f	Brara.I05114	AN1-TYPE ZINC FINGER PROTEIN
31797169-31797785	AT1G12300.1 | Symbols: | Tetratricopeptide repeat (T	Brara.I05115	PPR repeat
31797891-31806545	AT1G12300.1 | Symbols: | Tetratricopeptide repeat (T	Brara.I05115	PPR repeat
31806668-31810635	too many undefined sequences	-	
31813337-31814007	AT1G12410.1 | Symbols: CLPR2, NCLPP2, CLP2 | Cl	Brara.I05116	Endopeptidase Clp.
31819165-31819773	AT2G32360.1 | Symbols: | Ubiquitin-like superfamily	Brara.I05118	A09:42226980..42229813 forward
31819974-31820910	AT1G12380.1 | Symbols: | unknown protein; BEST A	Brara.I05118	A09:42226980..42229813 forward
31822152-31825603	AT5G44120.1 | Symbols: CRA1, ATCRA1, CRU1 | Ru	Brara.I05118	A09:42226980..42229813 forward
31830971-31832107	AT1G12360.1 | Symbols: KEU | Sec1/munc18-like (SM	Brara.I05119	PLANT SEC1
31835690-31839095	AT1G12360.1 | Symbols: KEU | Sec1/munc18-like (SM	Brara.I05119	PLANT SEC1
31839221-31846622	AT1G12210.1 | Symbols: RFL1 | RPS5-like 1 | chr1:41	Brara.I05121	CALMODULIN
31846922-31850332	AT1G12210.1 | Symbols: RFL1 | RPS5-like 1 | chr1:41	Brara.I05123	RFL1 (RPS5-LIKE 1), ATP BINDING / PROTEIN BINDING
31850625-31853012	AT1G12210.1 | Symbols: RFL1 | RPS5-like 1 | chr1:41	Brara.I05124	DISEASE RESISTANCE PROTEIN (CC-NBS-LRR CLASS), PUTATIVE
31853929-31856191	AT1G12290.1 | Symbols: | Disease resistance protein	Brara.I05123	RFL1 (RPS5-LIKE 1), ATP BINDING / PROTEIN BINDING
31856218-31857409	AT1G12290.2 | Symbols: | Disease resistance protein	Brara.I05123	RFL1 (RPS5-LIKE 1), ATP BINDING / PROTEIN BINDING
Marker 72			

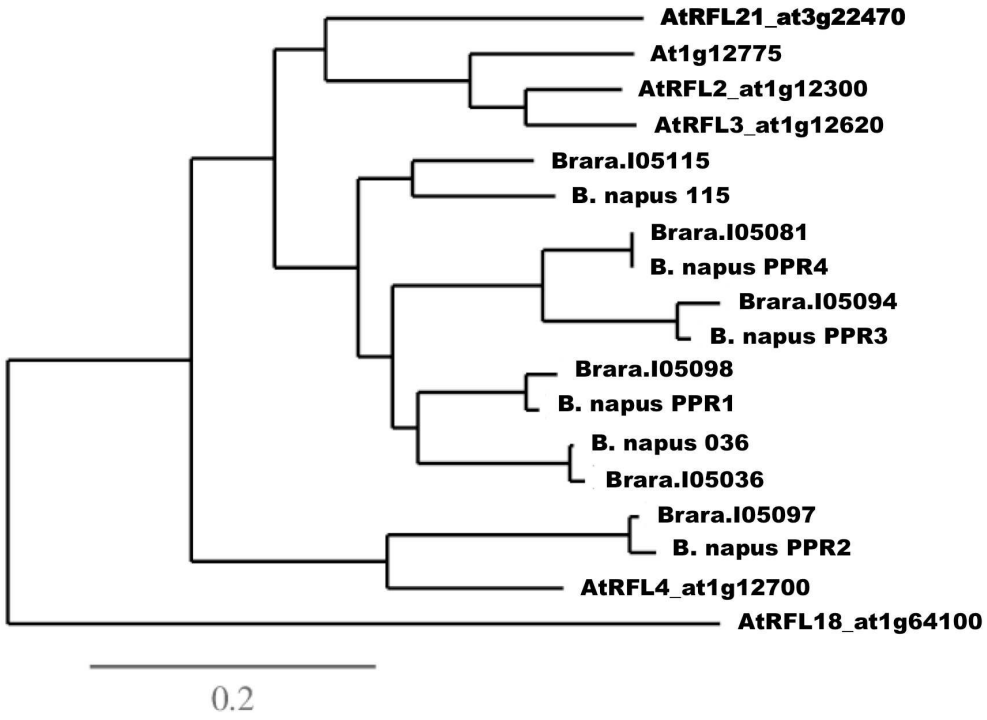
LEGEND Regions that correspond to the sequence of BAC NO202E11 are highlighted in light gray. Regions within the BAC that span the repeat sequences surrounding rearrangement vs. B. rapa appear in darker gray.

Note: Annotation of the 600kb fragment of Brassica napus chromosome A09 has been done by extracting the sequences from the brassicadb.org database and analylzing them with the Softberry (www.softberry.com) gene prediction program. BLASTN was then used to identify the closest homolog of each predicted gene the Brassica rapa (www.phytozome.net) or the Arabidopsis thaliana (www.tair.org) genomes in order to and in this manner to assign a putative function.

The presence of several hundreds of base pairs of un-identified sequences within the 600 kb B. napus fragment presented a challenge for the gene prediction and led to some ambiguous annotations. The presence of repeated sequences or protein domains, which could present genome assembly challenges, also rendered the task of precise gene annotation difficult.

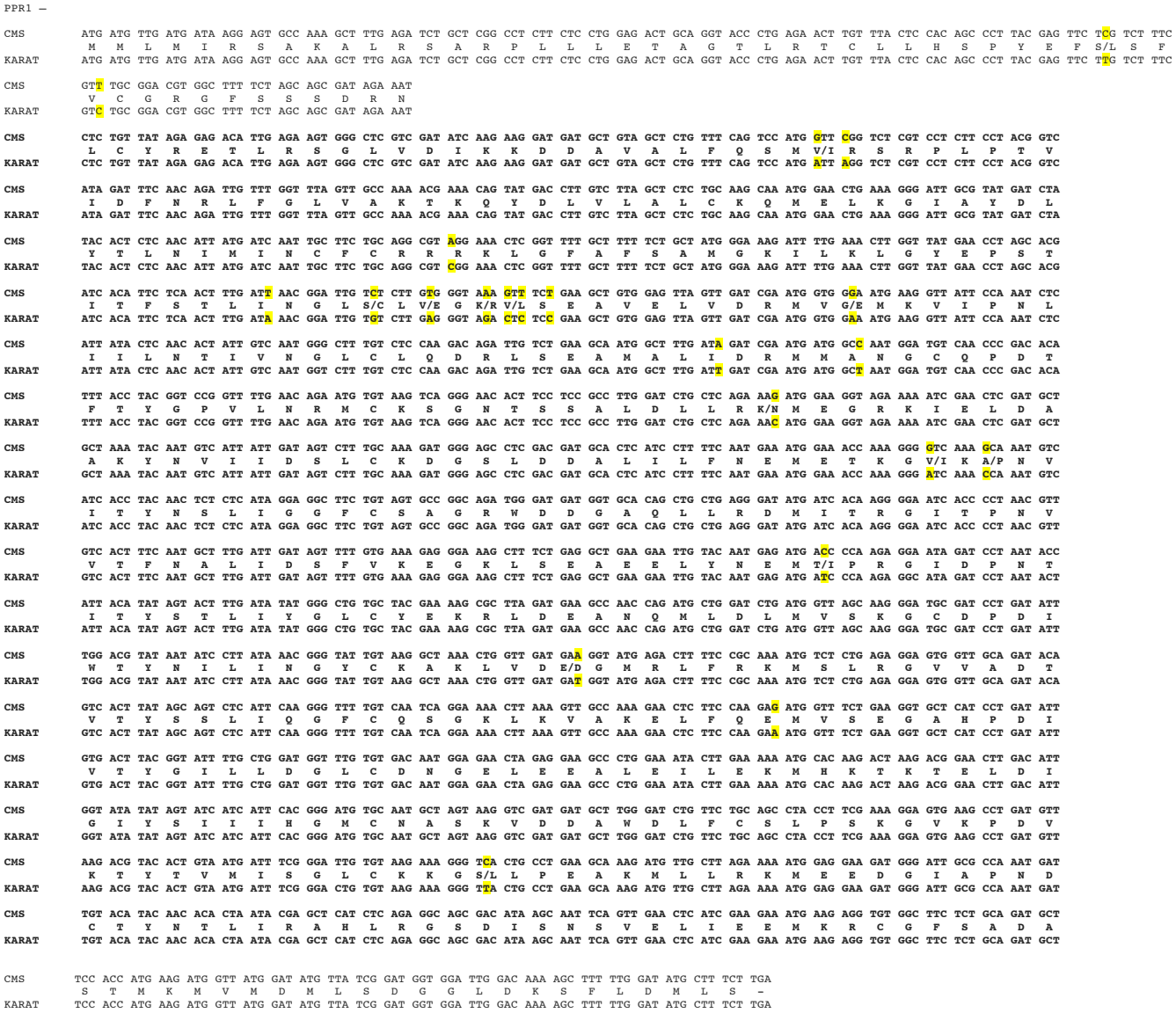
Annotation of the BAC NO202E11 was performed by the same strategy. Because of the reduced size of the fragment analyzed (180kb) and the absence of unidentified bases, we considered these predictions of gene position are more reliable. Although gene prediction of proteins presenting repeated domains should be considered as un-certain until further experimental verification can be done, the absence of any significant alignment between the cytochrome P450 predicted genes at the re-arrangement site (corresponding to Brara.I05076 and Brara.I05040-I05039) gives strong confidence in the annotation given during the analysis of the BAC sequence.

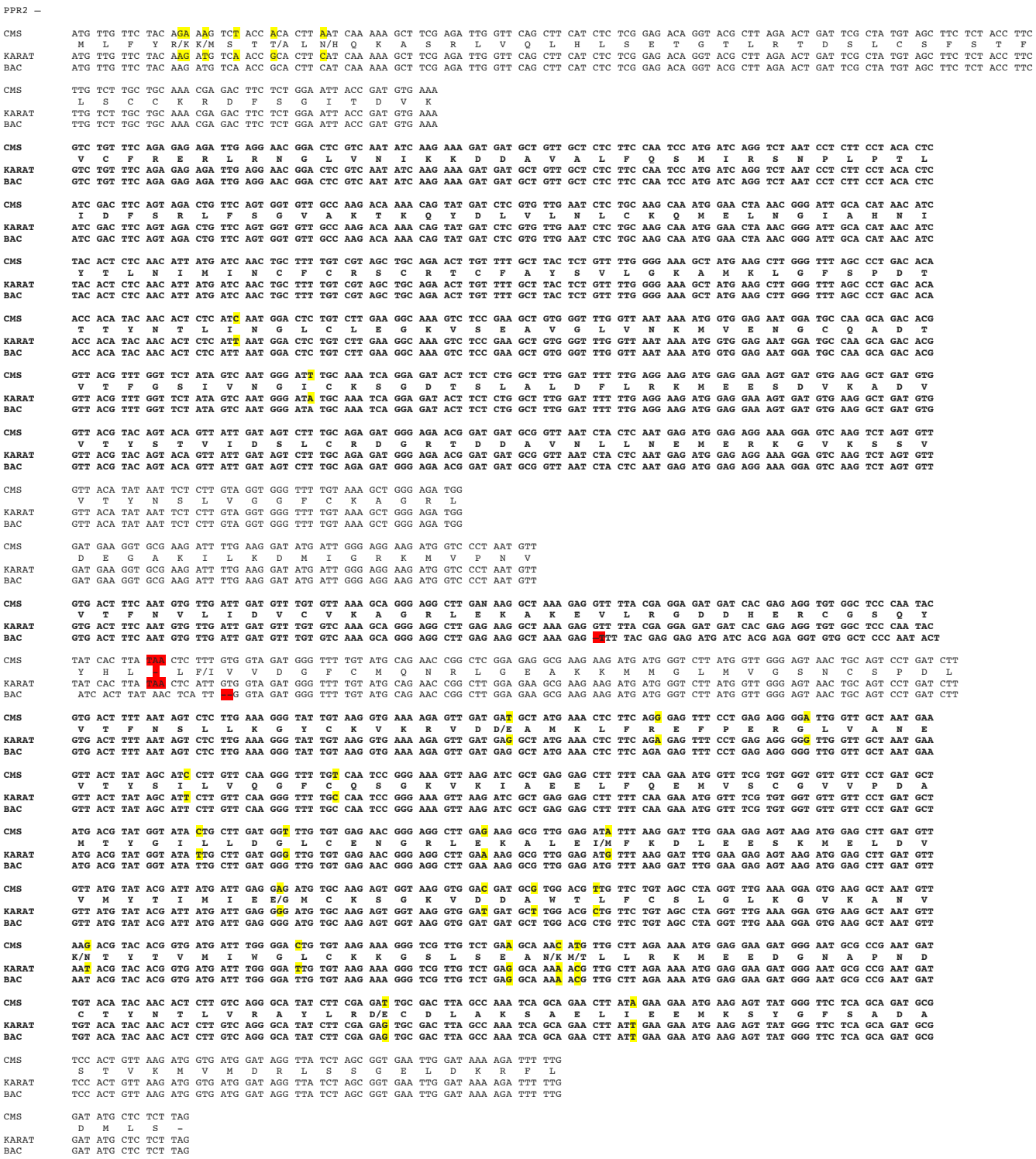
The observation of better synteny conservation between the annotated BAC, Brassica napus and Arabidopsis thaliana genomes compared to the more roughly annotated 600kb genomic fragment further support that idea. The gene synteny tool on the www.brassicadb.org database has also been used to corroborate the BAC annotations (http://brassicadb.org/brad/searchSyntenytPCK.php).



Supplementary figure 3.1. Phylogenetic analysis of proteins encoded in the *Rf*-region of *B. napus*/*rapa* and the orthologous segment of *A. thaliana.* A maximum likelihood tree was generated by the PhyML resource (Dereeper, Guignon et al. 2008) using as input the sequences of *A. thaliana* subgroup 1 RFL proteins and the *Brassica napus* and *Brassica rapa* proteins located in the genetically limited region that could potentially encode *Rfn*.

Supplementary figure 3.2. Complementary DNA (cDNA) sequences of transcripts of *B. napus* genes *PPR1-4*. The sequences designated "CMS" represent RT-PCR products from the nap CMS parent (*nap, rf/rf*) of the mapping cross, those designated Karat represent RT-PCR products from the nuclear fertility restorer line (*Rfn/Rfn*) used in the cross. The sequences corresponding to a single PPR domain appear on the same line. Nucleotide changes are highlighted in yellow, and, in cases where they result in an amino acid change the first letter indicates the amino acid in the CMS protein and the second letter the corresponding sequence in the restorer protein. Nucleotide changes leading to the generation of a termination codon in *PPR2* are highlighted in red.


PPR1 –
CMS
KARAT


PPR2 –
CMS
KARAT
BAC

```
PPR3 -

CMS       GAG ACG ATG ATG TTG ATG ATT CGG AGT GCC AAA GCT TTG AGA TCT GTT CGG CCT
           E/M T/M M/L  M  L/I M/R  I   R   S   A   K   A   L   R   S   V   R   P
KARAT     ATG ATG TTG ATG ATT CGG ATA AGG AGT GCC AAA GCT TTG AGA TCT GTT CGG CCT

CMS       CGA TTC ATG GAG ACA GCA GGT ACC CTG AGA ATT TCT CTA TTC CAC AGA ACC CCA TAC GAG
           R   F   M   E   T   A   G   T   L   R   I   S   L   F   H   R   T   P   Y   E
KARAT     CCA TTC ATG GAG ACA GCA GGT ACC CTG AGA ATT TCT CTA TTC CAC AGA ACC CCA TAC GAG

CMS       CTC TTG TCT TTC GTC TGC GAA AGA ACC TTC TCT GGT GGT AGC GAT AGA AAG ATG TCT GCT
           L   L   S   F   V   C   E   R   T   F   S   G   G   S   D   R   K   M   S   A
KARAT     CTC TTG TCT TTC GTC TGC GAA AGA ACC TTC TCT GGT GGT AGC GAT AGA AAG ATG TCT GCT

CMS       GCT TCT TAC AAA GAG AGA CTG AGA AGT GGG ATT ATT GGT ATC AAG AAG GAT GAA GCT GTT GCT CTG TTT CAG TCC ATG ATT AGG TCT CGC CCT CTT CCA ACA ATC
           A   S   Y   K   E   R   L   R   S   G   I   I   G   I   K   K   D   E   A   V   A   L   F   Q   S   M   I   R   S   R   P   L   P   T   I
KARAT     GCT TCT TAC AAA GAG AGA CTG AGA AGT GGG ATT ATT GGT ATC AAG AAG GAT GAA GCT GTT GCT CTG TTT CAG TCC ATG ATT AGG TCT CGC CCT CTT CCA ACA ATC

CMS       ATA GAT TTC AAC AGA TTG TTT ACT GCA ATG GCC AAA ACA AAA CAG TAT GAT CTC GTG TTG GAT CTC TGC AAG CAG ATG GAG TTG AAT GGG ATT GCA CAT AAC ATC
           I   D   F   N   R   L   F   T   A   M   A   K   T   K   Q   Y   D   L   V   L   D   L   C   K   Q   M   E   L   N   G   I   A   H   N  I/M
KARAT     ATA GAT TTC AAC AGA TTG TTT ACT GCA ATG GCC AAA ACA AAA CAG TAT GAT CTC GTG TTG GAT CTC TGC AAG CAG ATG GAG TTG AAC GGG GTT GCA CAT AAC ATG

CMS       TAC ACT CTC AAC ATT ATG ATC AAT TGC TTC TGC AGG CGT CCT AAA CTT GGT TTT GCT TTT TCT GTG ATG GGA AAG ATG TTG AAG CTT GGT TAT GAG CCC GAC AGA
           Y   T   L   N   I   M   I   N   C   F   C   R   R  P/W  K   L   G   F   A   F   S   V   M   G   K  M/I  L   K   L   G   Y   E   P   D   R
KARAT     TAC ACT CTC AAC ATT ATG ATC AAT TGC TTC TGC CGT CGC TGG AAA CTC GGT TTT GCT TTC TCT GTG ATG GGA AAG ATC TTG AAG CTT GGT TAT GAG CCC GAC AGA

CMS       GTC ACA TTT AAC ACC CTT CTC AAT GGC TTA TGT CTC GAG GGT AGA GTC TTT GAC GAT GTG GAG TTA GTT GAT TGT ATG GTC CTA AGC CAA CAT GTA CCA GAT CTC
           V   T   F   N   T   L   L   N   G   L   C   L   E   G   R   V   F   D  D/A  V   E   L   V   D   C   M   V   L   S   Q   H   V   P   D   L
KARAT     GTC ACA TTT AAC ACC CTT CTC AAT GGT TTA TGT CTC GAG GGT AGA GTC TTT GAC GCT GTG GAG TTA GTT GAT TGT ATG GTC CTA AGC CAA CAT GTA CCA GAT CTC

CMS       ATC ACC CTC AAC ACT CTT GTC AAT GGA CTT TGT CTC AAA GAT AGA GTC TCT GAA GCA GTG GAT TTA ATA GCT CGA ATG ATG GAT AAA GGA TGT CTA GCC GAT CAG
           I   T   L   N   T   L   V   N   G   L   C   L   K   D   R   V   S   E   A   V   D   L   I   A   R   M   M   D   K   G   C   L   A   D   Q
KARAT     ATC ACC CTC AAC ACT CTT GTC AAT GGA CTT TGT CTC AAA GAT AGA GTC TCT GAA GCA GTG GAT TTA ATA GCT CGA ATG ATG GAT AAA GGA TGT CAA GCC GAT CAG

CMS       TTT ACC TAT GGT CCG ATC TTG AAC AGA ATG TGT AAG TCA GGG AAC ACT ACA TTG GCC TTG GAT CTA CTC ACA AAG ATG GAA GAT AGA AAA GTC AAG CCT CAC GTA
           F   T   Y   G   P   I   L   N   R   M   C   K  S/F  G   N   T   T   L   A   L   D   L   L   T   K   M   E   D   R   K   V   K   P   H   V
KARAT     TTT ACC TAT GGT CCA ATC TTG AAC AGA ATG TGT AAG TTT GGG AAC ACT ACA TTG GCC TTG GAT CTA CTC ACA AAG ATG GAA GAT AGA AAA GTC AAG CCT CAC GTA

CMS       GTC ACA TAC AAT ATT ATT ATC GAT AGT CTT TGC AAA GAT GGG AGC CTC GAC GAT GCA CTC AGC TTT TTC AGT GAA ATG GAA ACC AAA GGG ATC AAA GCA GAT GTC
           V   T   Y   N   I  I/V  I   D   S   L   C   K   D   G   S   L   D   D   A   L   S   F   F   S   E   M   E   T   K   G   I   K   A   D   V
KARAT     GTC ACA TAC AAT ATT GTT ATC GAT AGT CTT TGC AAA GAT GGG AGC CTC GAC GAT GCA CTC AGC TTT TTC AGT GAA ATG GAA ACC AAA GGG ATC AAA GCA GAT GTC

CMS       TTT ACC TAC ACC TCT CTC ATA GGA GGC TTC TGT AGT GTT GGT AAA TGG GAT GAT GGT GCA CAG TTG CTG AGG GAT ATG ATT CGA AGG GGA ATC ACC CCG AAC GCT
           F   T   Y   T   S   L   I   G   G   F   C   S   V   G   K   W   D   D   G   A   Q   L   L   R   D   M   I   R   R  G/E  I   T   P   N   A
KARAT     TTT ACC TAC ACC TCT CTC ATA GGA GGC TTC TGT AGT GTT GGT AAA TGG GAT GAT GGT GCA CAG TTG CTG AGG GAT ATG ATT CGA AGG GAA ATC ACC CCG AAC GCC

CMS       ATC ACT TTC AGT TCT TTG ATT GAT AGT TTT GTG AAA GTG GGA AAG CTT TCT GAG GCT CAA GAT CTG TAC AAC GAG ATG ATC AAA AGA GGC ACA GAT CCT GAC ACC
           I   T   F   S   S   L   I   D   S   F   V   K   V   G   K   L   S   E   A   Q   D   L   Y   N   E   M   I   K   R   G   T   D   P   D   T
KARAT     ATC ACT TTC AGT TCT TTG ATT GAT AGT TTT GTG AAA GTG GGA AAG CTT TCT GAG GCT CAA GAT CTG TAC AAC GAG ATG ATC AAA AGA GGC ACA GAT CCT GAC ACC

CMS       ATT ACA TAT AAT TCT TTG ATA TAT GGG TTG TGC ATG GAG AAG CGC TTA GAT GAG GCC AGA GAG ATG CTG GAT CTG ATG GTT AGC AAG GGA TGT GAT CCA GAT ATT
           I   T   Y   N   S   L   I   Y   G   L   C   M   E   K   R   L   D   E   A   R   E   M   L   D   L   M   V   S   K   G   C   D   P   D   I
KARAT     ATT ACA TAT AAT TCT TTG ATA TAT GGG TTG TGC ATG GAG AAG CGC TTA GAT GAG GCC AGA GAG ATG CTG GAT CTG ATG GTT AGC AAG GGA TGT GAT CCA GAT ATT

CMS       GTG ACT TAT AGT ATC CTT ATA AAC GGC TAC TGC AAG GCT AAA CTG GTT GAT GAA GGT ATG AGA CTT TTC CGC AAA ATG ACC TTG AGA GGG GTG GTT GCC AAT ACA
           V   T   Y   S   I   L   I   N   G   Y   C   K   A   K   L   V   D   E   G   M   R   L   F   R   K   M   T   L   R   G   V   V   A   N   T
KARAT     GTG ACT TAT AGT ATC CTT ATA AAC GGC TAC TGC AAG GCT AAA CTG GTT GAT GAA GGT ATG AGA CTT TTC CGC AAA ATG ACC TTG AGA GGG GTG GTT GCC AAT ACA

CMS        GTG ACT TAT AGC ACT CTC ATC CAA GGG TTT TGT CAA TCT GGA AAA CTT AAT GTT GCC AAA GAA CTC TTC CAG GAG ATG GTT TCT GAA GGT GTT CGT CCT AGT ATT
            V   T   Y   S   T   L   I   Q   G   F   C   Q   S   G   K   L   N   V   A   K   E   L   F   Q   E   M   V   S   E   G   V  R/H  P   S   I
KARAT      GTG ACT TAT AGC ACT CTC ATC CAA GGG TTT TGT CAA TCT GGA AAA CTT AAT GTT GCC AAA GAA CTC TTC CAG GAG ATG GTT TCT GAA GGT GTT CAT CCT AGT ATT

CMS        ATG ACT TAC GGT ATT TTG CTG GAT GGG TTG TGT GAC AAT GGA GAA GTA GAG GAA GCC ATG GAA ATA CTT GAA AAA ATG CAC AAG TGT AAG ATT GAT CCT GGT ATT
            M   T   Y   G   I   L   L   D   G   L   C   D   N   G   E   V/L  E   E   A  M/L E/G  I   L   E   K   M   H   K   C   K   I   D   P   G   I
KARAT      ATG ACT TAC GGT ATT TTG CTG GAT GGG TTG TGT GAC AAT GGC GAA CTA GAA GAG GCT TTG GGA ATA CTT GAA AAA ATG CAC AAG TGT AAG ATT GAT CCT GGT ATT

CMS        GGT ATA TAT ACT ATC ATC ATT CAC GGT ATG TGC AAT GCA AAT AAG GTC GAT GAT GCT TGG GAT CTA TTC TGC AGC CTC TCT CTC AAA GGA GTG AAG CGT GAT ATT
            G   I   Y   T   I   I   I   H   G   M   C   N   A   N   K  V/I  D   D   A   W   D   L   F   C   S   L   S   L   K   G   V   K   R   D   I
KARAT      GGT ATC TAT ACT ATC ATC ATT CAC GGT ATG TGC AAT GCA AAT AAG ATC GAT GAT GCT TGG GAT CTA TTC TGC AGC CTC TCT CTC AAA GGA GTG AAG CGT GAT ATT

CMS        CGG TCA TAC AAC ATA ATG TTG TCA GGA TTA TGT AAG AGG AGC TCA TTG TCT GAA GCG GAT GCA TTG TTT AGA AAA ATG AAG GAA GAT GGG TAT GAG CCA GAT GAT
            R   S   Y   N   I   M   L   S   G   L   C   K   R   S   S   L   S   E   A   D   A   L   F   R   K   M   K   E   D   G   Y   E   P   D   D
KARAT      CGG TCA TAC AAC ATA ATG TTG TCA GGA TTA TGT AAG AGG AGC TCA TTG TCT GAA GCG GAT GCA TTG TTT AGA AAA ATG AAG GAA GAT GGG TAT GAG CCA GAT GAT

CMS        TGT ACG TAC AAT ACA CTT ATC AGA GCA CAT CTT CGA GGT AGT GAC ATA ACA ACT TCA GTT CAA CTC ATT GAA GAA ATG AAG AGG TGT GGG TTC TCT TCA GAT GCT
            C   T   Y   N   T   L   I   R   A   H   L   R   G   S   D   I   T   T   S   V   Q   L   I   E   E   M   K   R   C   G   F   S   S   D   A
KARAT      TGT ACG TAC AAT ACA CTT ATC AGA GCA CAT CTG CGA GGT AGT GAC ATA ACA ACT TCA GTT CAA CTC ATT GAA GAA ATG AAG AGG TGT GGG TTC TCT TCA GAT GCT

CMS       TCC ACC GTA AAG ATT GTT ATG GAT ATG CTA TCG AGT GGT GAA TTG GAC AAA AGC TTT CTA GAT
           S   T   V   K   I   V   M   D   M   L   S   S   G   E   L  D/N  K   S   F   L   D
KARAT     TCC ACC GTA AAG ATT GTT ATG GAT ATG CTA TCG AGT GGT GAA TTG AAC AAA AGC TTT CTA GAT

CMS       ATG CTT TCT GGT CCT TCT CGA GAG ATA GCA TCA TCG TTG GAT TGA
           M   L   S   G   P   S   R   E   I   A   S   S   L   D   -
KARAT     ATG CTT TCT GGT CCT TCT CGA GAG ATA GCA TCA TCG TTG GAT TGA
```

PPR4 -

```
CMS         ATG GTG TTG AGG ACA CAG AGA TGG
             M   V   L   R   T   Q   R   W
KARAT       ATG GTG TTG AGG ACA CAG AGA TGG

CMS         AAT CGT CTT ACT ACT TTG AGA TTG GTT CAT CTC CGT TCA ACT GAG ACA GGT ACT CTG AGA
             N   R   L   T   T   L   R   L   V   H   L   R   S   T   E   T   G   T   L   R
KARAT       AAT CGT CTT ACT ACT TTG AGA TTG GTT CAT CTC CGT TCA ACT GAG ACA GGT ACT CTG AGA

CMS         AAT GCT GCT TTC TTC CAA AGC CCA TAC GAC TTC TTC TTC TGC GTA CAA GGC TTC TCT GGT
             N   A   A   F   F   Q   S   P   Y   D   F   F   F   C   V   Q   G   F   S   G
KARAT       AAT GCT GCT TTC TTC CAA AGC CCA TAC GAC TTC TTC TTC TGC GTA CAA GGC TTC TCT GGT

CMS         CTC ACC AGC GAT AGA AAG ATG
             L   T   S   D   R   K   M
KARAT       CTC ACC AGC GAT AGA AAG ATG

CMS         TCT TCT TAC AAA GAG AGA TTG AGA AGT GGT CTC GTC GAT ATC AAG AAG GAT GAT GCT GTT GCT CTG TTT CAG TCC ATG CTT CGG TCT CGT CCT CTT CCT ACG GTC
             S   S   Y   K   E   R   L   R   S   G   L   V   D   I   K   K   D   D   A   V   A   L   F   Q   S   M   L   R   S   R   P   L   P   T   V
KARAT       TCT TCT TAC AAA GAG AGA TTG AGA AGT GGT CTC GTC GAT ATC AAG AAG GAT GAT GCT GTT GCT CTG TTT CAG TCC ATG CTT CGG TCT CGT CCT CTT CCT ACG GTC

CMS         ATT GAT TTC AAC AGA TTG TTT GGT TTA CTT GCC AGA ACT AAA CAG TAC GAT CTC GTG TTA GCT CTC TGC AAG CAA ATG GAA CTG AAA GGG ATT GCG TAT GAC CTC
             I   D   F   N   R   L   F   G   L   L   A   R   T   K   Q   Y   D   L   V   L   A   L   C   K   Q   M   E   L   K   G   I   A   Y   D   L
KARAT       ATT GAT TTC AAC AGA TTG TTT GGT TTA CTT GCC AGA ACT AAA CAG TAC GAT CTC GTG TTA GCT CTC TGC AAG CAA ATG GAA CTG AAA GGG ATT GCG TAT GAC CTC

CMS         TAC ACT CTC AAC ATT ATG ATC AAT TGC TTC TGC AGG CGT CGG AAA CTC GGT TTT GCT TTT TCC GCT ATG GGG GAG ATC TTC AAA CTT GGG TAT GAG CCT AAC ACA
             Y   T   L   N   I   M   I   N   C   F   C   R   R   R   K   L   G   F   A   F   S   A   M   G   E   I   F   K   L   G   Y   E   P   N   T
KARAT       TAC ACT CTC AAC ATT ATG ATC AAT TGC TTC TGC AGG CGT CGG AAA CTC GGT TTT GCT TTT TCC GCT ATG GGG GAG ATC TTC AAA CTT GGG TAT GAG CCT AAC ACA

CMS         GTC ACA TTT AAC ACC CTC CTC AAT GGC TTA TGT CTC GAG GGC AGA GTC TTT GAA GCT GTG GAG TTA GTT GAT TGT ATG GTC CTA AGC CAA CAT GTA CCA GAT CTT
             V   T   F   N   T   L   L   N   G   L   C   L   E   G   R   V   F   E   A   V   E   L   V   D   C   M   V   L   S   Q   H   V   P   D   L
KARAT       GTC ACA TTT AAC ACC CTC CTC AAT GGC TTA TGT CTC GAG GGC AGA GTC TTT GAA GCT GTG GAG TTA GTT GAT TGT ATG GTC CTA AGC CAA CAT GTA CCA GAT CTT

CMS         ATC ACC CTC AAC ACT ATT GTC AAT GGG CTT TGT CTC AAA GAT AGA GTG TCT GAA GCA GTG GAT TTA ATA GCT CGA ATG ATG GAT AAA GGA TGT CAA GCC GAT CAG
             I   T   L   N   T   I   V   N   G   L   C   L   K   D   R   V   S   E   A   V   D   L   I   A   R   M   M   D   K   G   C   Q   A   D   Q
KARAT       ATC ACC CTC AAC ACT ATT GTC AAT GGG CTT TGT CTC AAA GAT AGA GTG TCT GAA GCA GTG GAT TTA ATA GCT CGA ATG ATG GAT AAA GGA TGT CAA GCC GAT CAG

CMS         TTT ACC TAT GGT CCA ATC TTG AAC AGA ATG TGT AAG TCT GGG AAC ACT GCC TCG GCC TTG GAT CTG CTC AGG AAG ATG GAA CAT AGA AAG ATC AAG CCA CAC GTA
             F   T   Y   G   P   I   L   N   R   M   C   K   S   G   N   T   A   S   A   L   D   L   L   R   K   M   E   H   R   K   I   K   P   H   V
KARAT       TTT ACC TAT GGT CCA ATC TTG AAC AGA ATG TGT AAG TCT GGG AAC ACT GCC TCG GCC TTG GAT CTG CTC AGG AAG ATG GAA CAT AGA AAG ATC AAG CCA CAC GTA

CMS         GTC ACA TAC AAT ATC ATC ATT GAC AAT CTT TGC AAA GAT GGG AGA CTC GAC GAT GCA CTC AGC TTT TTC AGT GAA ATG GAA ACC AAA GGG ATC AAA GCA GAT GTC
             V   T   Y   N   I   I   I   D   N   L   C   K   D   G   R   L   D   D   A   L   S   F   F   S   E   M   E   T   K   G   I   K   A   D   V
KARAT       GTC ACA TAC AAT ATC ATC ATT GAC AAT CTT TGC AAA GAT GGG AGA CTC GAC GAT GCA CTC AGC TTT TTC AGT GAA ATG GAA ACC AAA GGG ATC AAA GCA GAT GTC

CMS         ATT ACC TAC AAC TCT CTC ATA GGA AGC TTC TGT AGT TTT GGC AGA TGG GAT GAT GGT GCA CAG TTG CTG AGG GAT ATG ATT ACA AGG AAA ATC ACC CCC AAC GTT
             I   T   Y   N   S   L   I   G   S   F   C   S   F   G   R   W   D   D   G   A   Q   L   L   R   D   M   I   T   R   K   I   T   P   N   V
KARAT       ATT ACC TAC AAC TCT CTC ATA GGA AGC TTC TGT AGT TTT GGC AGA TGG GAT GAT GGT GCA CAG TTG CTG AGG GAT ATG ATT ACA AGG AAA ATC ACC CCC AAC GTT
```

```
CMS        GTC ACT TTC AGT GCT TTG ATT GAT AGT CTT GTT AAA GAG GGA AAG CTT ACT GAG GCT AAA GAC TTG TAC AAT GAG ATG ATC ACA AGA GGC ATA GAT
CCT AAT ACC
            V   T   F   S   A   L   I   D   S   L   V   K   E   G   K   L   T   E   A   K   D   L   Y   N   E   M   I   T   R   G   I   D
P   N   T
KARAT      GTC ACT TTC AGT GCT TTG ATT GAT AGT CTT GTT AAA GAG GGA AAG CTT ACT GAG GCT AAA GAC TTG TAC AAT GAG ATG ATC ACA AGA GGC ATA GAT
CCT AAT ACC

CMS        ATT ACA TAT AGT ACT TTG ATA TAT GGG TTG TGC ATG GAG AAC CGC TTA GAT GAA GCC AAC CAG ATG ATG GAC CTC ATG GTT AGC AAG GGA TGC GAT
CCT GAT ATC
             I   T   Y   S   T   L   I   Y   G   L   C   M   E   N   R   L   D   E   A   N   Q   M   M   D   L   M   V   S   K   G   C   D
P   D   I
KARAT      ATT ACA TAT AGT ACT TTG ATA TAT GGG TTG TGC ATG GAG AAC CGC TTA GAT GAA GCC AAC CAG ATG ATG GAC CTC ATG GTT AGC AAG GGA TGC GAT
CCT GAT ATC

CMS        GTG ACG TTT AAT GTC CTT ATA AAC GGA TTT TGT AAG GCT AAA CAG GTT GAT GTT GGT ATG AGA CTA TTC CGA AAG ATG TCT CTG AGA GGA GAG ATT
GCA GAT ACA
            V   T   F   N   V   L   I   N   G   F   C   K   A   K   Q   V   D   V   G   M   R   L   F   R   K   M   S   L   R   G  E/V  I
A   D   T
KARAT      GTG ACG TTT AAT GTC CTT ATA AAC GGA TTT TGT AAG GCT AAA CAG GTT GAT GTT GGT ATG AGA CTA TTC CGA AAG ATG TCT CTG AGA GGA GTG ATT
GCA GAT ACA

CMS        GTG ACT TAT AGC ACT CTC ATC CAA GGG TTT TGT CAA TCA AGA GAA CTT ATT GTC GCC AAA GAA GTC TTC CAA GAG ATG GTC TCT CAA GGT GTT CAT
CCT GGT ATT
            V   T   Y   S   T   L   I   Q   G   F   C   Q   S   R   E   L   I   V   A   K   E   V   F   Q   E   M   V   S   Q   G   V   H
P   G   I
KARAT      GTG ACT TAT AGC ACT CTC ATC CAA GGG TTT TGT CAA TCA AGA GAA CTT ATT GTC GCC AAA GAA GTC TTC CAA GAG ATG GTC TCT CAA GGT GTT CAT
CCT GGT ATT

CMS        ATG ACT TAT GGT ATT TTG CTG GAT GGG TTG TGT GAC AAT GGC GAA CTA GAA GAG GCT TTG GGA ATA CTT GAT CAA ATG CAC AAG TGT AAG ATG GAA
CTT GAT ATT
             M   T   Y   G   I   L   L   D   G   L   C   D   N   G   E   L   E   E   A   L   G   I   L   D   Q   M   H   K   C   K   M   E
L   D   I
KARAT      ATG ACT TAT GGT ATT TTG CTG GAT GGG TTG TGT GAC AAT GGC GAA CTA GAA GAG GCT TTG GGA ATA CTT GAT CAA ATG CAC AAG TGT AAG ATG GAA
CTT GAT ATT

CMS        GGT ATA TAT AGT ATC ATC ATT CAC GGG TTG TGC AAT GCA AGT AAG ATC GAT GAT GCT TGG GAT CTA TTC TGT AGC CTC TCT CTC AAA GGA GTG AAG
CGT GAT ATT
             G   I   Y   S   I   I   I   H   G   L   C   N   A   S   K   I   D   D   A   W   D   L   F   C   S   L   S   L   K   G   V   K
R   D   I
KARAT      GGT ATA TAT AGT ATC ATC ATT CAC GGG TTG TGC AAT GCA AGT AAG ATC GAT GAT GCT TGG GAT CTA TTC TGT AGC CTC TCT CTC AAA GGA GTG AAG
CGT GAT ATT

CMS        CAG TCA TAC AAC ATA ATG TTG TCA GGA TTA TGT AAA AGG AGC TCA TTG TCT GAA GCG GAT GCA TTG TTT AGA AAA ATG AAG GAA GAT GGG TAT GAG
CCA GAT GGT
            Q   S   Y   N   I   M   L   S   G   L   C   K   R   S   S   L   S   E   A   D   A   L   F   R   K   M   K   E   D   G   Y   E
P   D   G
KARAT      CAG TCA TAC AAC ATA ATG TTG TCA GGA TTA TGT AAA AGG AGC TCA TTG TCT GAA GCG GAT GCA TTG TTT AGA AAA ATG AAG GAA GAT GGG TAT GAG
CCA GAT GGT

CMS         TGT ACG TAC AAT ACA CTT ATC AGA GCA CAT CTT CGA GGT AGT GAC ATA ACA ACT TCA GTT CAA CTC ATT GAA GAA ATG AAG AGG TGT GGG TTC
TCT TCA GAT GCT
             C   T   Y   N   T   L   I   R   A   H   L   R   G   S   D   I   T   T   S   V   Q   L   I   E   E   M   K   R   C   G   F   S
S   D   A
KARAT      TGT ACG TAC AAT ACA CTT ATC AGA GCA CAT CTT CGA GGT AGT GAC ATA ACA ACT TCA GTT CAA CTC ATT GAA GAA ATG AAG AGG TGT GGG TTC
TCT TCA GAT GCT

CMS         TCC ACC GTA AAG ATT GTT ATG GAT ATG CTA TCA AGT GGT
             S   T   V   K   I   V   M   D   M   L   S   S   G
KARAT       TCC ACC GTA AAG ATT GTT ATG GAT ATG CTA TCA AGT GGT

CMS        GAA TTG GAC AAA AGC TTT CTA AAT ATG CTT TCT GGT CCT TTT GGA GAC AAA TCA TCA TTG
            E   L   D   K   S   F   L   N   M   L   S   G   P   F   G   D   K   S   S   L
KARAT      GAA TTG GAC AAA AGC TTT CTA AAT ATG CTT TCT GGT CCT TTT GGA GAC AAA TCA TCA TTG

CMS        TTG GAT TGA
            L   D   -
KARAT      TTG GAT
```