# RADIATION HYBRID FINE MAPPING OF TWO FERTILITY-RELATED GENES: MARKING THE PATH TO WHEAT HYBRIDS

A Dissertation
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Filippo Maria Bassi

In Partial Fulfillment
for the Degree of
DOCTOR OF PHILOSOPHY

Major Department:
Plant Sciences

December 2012

Fargo, North Dakota

# North Dakota State University
## Graduate School

Title

RADIATION HYBRID FINE MAPPING OF TWO FERTILITY-RELATED GENES: MARKING THE PATH TO WHEAT HYBRIDS

By

Filippo Maria Bassi

The Supervisory Committee certifies that this *disquisition* complies with North Dakota State University's regulations and meets the accepted standards for the degree of

DOCTOR OF PHILOSOPHY

SUPERVISORY COMMITTEE:

Shahryar F. Kianian, Ph.D.
Chairperson

Anne Denton, Ph.D.

Mohamed Mergoum, Ph.D.

Phillip McClean, Ph.D.

Justin Faris, Ph.D.

Approved:

December 12, 2012
Date

Richard D. Horsley, Ph.D.
Department Chair

# ABSTRACT

Over one billion people, more than 1/9$^{th}$ of the global population, are undernourished. Feeding the ever increasing population has to be the most important goal of plant sciences. Since cultivated areas are not likely to increase, I will need to produce more with what is available. This can be summarized in one word: yield. Unfortunately, wheat's yield is expected to increase only 1.13% by 2019, a prediction that if converted into reality will likely indicate that I failed to cope with the world demographic increase. A new strategy to revolutionize wheat production is required, and some believe that this change might be represented by wheat hybrids. Achieving adequate commercial production of wheat hybrids has the potential to nearly double the yield of one of the world's most important staple food. The first fundamental step toward this goal is to develop feasible methodologies to sterilize the male part of the complete wheat flowers. Two fertility-related genes are the primary target of this study, namely the *species cytoplasm specific* on chromosome 1D, and the *desynaptic* locus on chromosome 3B. This dissertation summarizes the important achievements obtained toward the cloning of the two loci by means of radiation hybrid functional analysis. Radiation hybrid is a technique that employs radiation to create genetic diversity along the targeted chromosome. Chapter 1 explains in details how this methodology can be applied to plants. The use of radiation hybrid mapping permitted creating a comprehensive map of wheat chromosome 3B, as discussed in Chapter 2, and then expanded the mapping information to identify the 2 Mb location of the desynaptic locus *desw2*, as discussed in Chapter 3. A similar approach on chromosome 1D allowed first to pinpoint the location of the *species cytoplasm specific* gene to a region of 2 Mb, as discussed in Chapter 4, and then ultimately to find a strong candidate for this locus, as discussed in Chapter 5. Now that the molecular locations of these genes have been unraveled by this study, their sequence can be streamlined into transformation to ultimately produce female wheat plants, and consequently hybrids.

# ACKNOWLEDGEMENTS

I wish to thank: Justin Hegstad and Allen Peckrul for capable technical support throughout the research. For fundamental experimental help Dr. Matthew Hayden, Dr. Etienne Paux, Dr. Yong Gu, and Dr. Wojtek Pawlowski. My advisory committee that has supported me every single step of the way, bearing with my poor cooking, while enjoying the long hours of student-grilling during the various exams: Dr. Anne Denton, Dr. Justin Faris, and Dr. Phillip McClean. Dr. Mohammed Mergoum who had to cope with my field inexperience, and surprising enough for the both of us, he succeeded in making a breeder out of me after all. My advisor Dr. Shahryar Kianian, who has never once told me that I did a good job, no matter how hard I tried, but I know that deep inside he was happy to be my advisor (some days better than others) and really helped me throughout these three cold years in Fargo. The other members of the WGE family who had to survive long years with me around: Dr. Ajay Kumar, who stole one first authorship but gave in return a deep and true friendship for which I will always be thankful; Dr. Farhad Gavami with whom I fought a lot as much as laugh together, and did some great research with; Dr. Monika Michalack DJ, Dr. Kristin Simmons, and Rissa who are the true responsibles for the success of the *scs* project; Mona Mazaheri, who has caused me more than one headache and managed to paint our shower curtain red, but who has also been a great friend; to Farid and Ali, who I am sure will go out there and become the best damn scientists that this world has to offer; to Andrzej who tried to save my soul but instead found himself stuck with a friend from the wrong side of the heaven's doors. An apologies goes to my family who had to bear with my absence, and probably will not be able to understand what I am writing, but always always always supported me, no matter how deep was the hole I was digging myself into, they were always there with a ladder to take me out. My sister, with whom I enjoy arguing, but that I love with all my hearth. A special thank goes to the funding sources that made my research possible: Program M&B Sardegna, and Monsanto Beachell-Borlaug International Scholarship that made me one of the few decently paid PhD student in the country, but also introduced me to great students from all places in the world, and especially made me meet my beautiful Dorothy, who is sitting across the table from me as I write these words.

# DEDICATION

This goes to the man that inspired and fomented my dedication to plant science and to the challenge of feeding the world: Dr. Norman Borlaug. My deepest regret is that I never had a chance to meet you; I can only picture how amazing it would have been to have a chance in Des Moines of shaking the hand of the man who fed a billion people. As a scientist I know that you have just become part of the natural circle of life and that you cannot hear me now, but if you could I would like to tell you that your example is inspiring thousands of young students, that what you did for us will never be forgotten, and hell you put together some catchy phrases.

"*There is no peace on an empty belly*"

*Norman Borlaug*

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF ABBREVIATIONS

| | |
|---|---|
| (lo) | Ae. longissima cytoplasm |
| BAC | Bacterial Artificial Chromosomes |
| bp | base pair |
| BR | Break/Repair |
| BulkSeq | Bulk segregant analysis by next-gen Seqeuncing |
| cDNA | complementary DNA |
| CO | Crossing Over |
| cR | centi Rays |
| CS | 'Chinese Spring' |
| DArT | Diversity Array Technology |
| DEASY | Duplexing EASY |
| DSB | Double Strand Break |
| EST | Expressed Sequence Tags |
| gDNA | genomic DNA |
| Gy | Gray |
| HR | Homology-direct Repair |
| ISBP | Insertion Site based Polymorphism |
| K | thousand |
| LDN 1D(1A) | 'Langdon' with chromosome 1A substituted by chromosome 1D |
| LDN 3D(3B) | 'Langdon' with chromosome 3B substituted by chromosome 3D |
| LDN | 'Langdon' |
| M | Million |
| mRNA | messenger RNA |
| NDCtg | North Dakota State University markers designed on Contigs |
| NDRT | North Dakota State University Real-Time PCR markers |
| NDSU | North Dakota State University PCR markers |
| NHEJ | Non-Homologous End-Joining |

PV..... Pollen Vitality

Ret..... Retrojunction markers

RH ..... Radiation Hybrid

Scs ..... Species cytoplasm specific

SpS..... Seeds per Spikelet

SSR ..... Single Sequence Repeat

TC..... Tentative Consensus EST sequences

USDA ..... United States Department of Agriculture

# LIST OF APPENDIX TABLES

# LIST OF APPENDIX FIGURES

# 1. MUTAGENOMICS: THE ADVANCE OF RADIATION HYBRIDS FOR STUDYING GENES AND GENOMES IN PLANTS

Genetic variation is the key to phenotypic improvement in plants. Variation in the genome provides organisms with the potential to adapt to changing environmental pressures and improves their chance of survival. Genetic variation also lays the foundation for genomic studies and efforts made at crop improvement. Genetic variability in any population can be the result of natural changes, including but not limited to recombination, mistakes in DNA replication and repair, or induced changes, such as those that result from chemical or radiation treatments. Mutagenesis has been applied extensively to an array of biological studies, including physics, physiology, agronomy, breeding, and genetics. In more recent times, this experimental strategy has followed the general trend of modern biology, trespassing into the field of genomics, ultimately gaining its “-omics” title: mutagenomics. This chapter reviews radiation mutagenesis from the prospective of -omics practices, its effect on plant genomes, its application to plant breeding, and its use to develop segregating populations for mapping. This introductory section provides a detailed overlook on the successful applications of radiation and radiation hybrids to generate complete physical maps, to assemble genome survey sequences, to precisely estimate synteny relationships across species, and ultimately for functional analysis in plants.

## 1.1. Effect of radiation on plant genomes

In order to effectively employ populations of mutants in genomics studies, it is critical to have a rudimental grasp of the principles governing radiation-mediated damage to the DNA and its subsequent mechanism of repair. Ionizing radiation is generated by radioactive unstable atoms, the nucleus of which breaks down in the attempt to reach a non-radioactive (stable) state (Flakus 1981). As the nucleus disintegrates, energy is released in the form of radiation. The amount of damage that ionizing radiations can cause to biological material is a function of their energy level and their ability to penetrate tissues (Argonne National Laboratory 2005). Linear energy transfer (LET) is the energy deposited by charged particles of ionizing radiation when travels through a material. These are waves

of bundles (quanta) of energy that have no charge or mass and can travel long distances through air, biological tissues, and other materials. These rays might encounter atoms on their flight path; when this occurs they transfer all or part of their energy to that atom. The energy transfer normally knocks an electron out of the atom, ionizing it. This electron then uses the energy received to create additional ions by further transferring its energy to other atoms. Since the probability of hitting an atom is a function of the density of the atoms in the specific media irradiated, different biological tissue will respond differently to LET radiation (Kovacs and Keresztes 2002). Also, the more energetic is the excited particle, the more atoms can be potentially ionized along their path. For this reason, LET radiation are categorized on the basis of their relative energy into low-LET radiation such as X-rays and γ-rays, and high-LET radiation such as ion beams (Shikazono et al. 2001). Other than the relative energy, the interaction of three more components (type of sample treated, time of exposure, and distance from radiation source) determines the amount of radiation absorbed dose (rad), which is a unit defined in 1962 by the International Commission on Radiological Units and Measurements (1962). One rad corresponds to the radiation dose causing 100 ergs of energy to be absorbed by one gram of matter. However, since 1980 the rad, which was not coherent with the International System of Units (1980), was replaced with the Gray (Gy), which corresponds to approximately 100 rads. Today both units are used in publications, but only the Gy is recognized as an International Standard.

During exposure of biological tissue to radioactive isotopes, the quanta of energy travel across the tissues with complete lack of target specificity. The probability of any cell component to interact with ionizing energy is proportional to the fraction of tissue that they occupy. Since water accounts for over 80% of the volume of any metabolically active plant cell, water radiolysis is the primary effect of radiation in plant tissues (Britt 1996). Ionization of water gives rise to reactive oxygen species (ROS) such as hydrogen peroxide ($H_2O_2$), superoxide anion ($O_2^-$), hydroxyl radicals (OH) and singlet oxygen (O) (Wi et al. 2007). When ROS originate in the nucleus and interact with DNA, the sugar phosphate backbone undergoes oxidative damage, which leads inevitably to single-strand breaks (SSB; Britt 1996). While SSB are generally repaired in an efficient and error-free fashion, continuous exposure to ROS will eventually cause formation of SSB on both strands, resulting in double-strand breaks (DSB). More rarely, the sugar backbone is directly hit by an ionizing atom and the absorbed energy might cause SSB. In this case, the sensitization of the opposite un-nicked strand

may result in an increased yield of DSB (Britt 1996). Because nucleotides near the ends of breaks are rapidly degraded, DSB generally expand into gaps that cannot simply be rejoined to restore the original sequence (Britt 1999).

Eukaryotic cells immediately respond to this type of damage by activating their DNA repair mechanisms. Homologous recombination (HR) is an error-free mechanism of DSB repair that involves the use of a complementary template sequence to fill the gap separating the two breaks. For this mechanism of repair to be successful a compatible template sequence is required in close proximity of the breaks to be repaired. This is the case when homologous chromosomes are brought together to form the synaptonemal complex during leptotene to early pachytene phases. It has been suggested that eukaryotic cells employ the HR mechanism to repair DSB only during this specific portion of the cell-cycle (Ahmed et al. 2010). During any other stages of the cell life DSB are repaired by non-homologus end joining (NHEJ). This mechanism, often referred to as “illegitimate recombination”, employs a complex of proteins to bring together the DNA termini, to then join those using micro-homologies of two to five bases to stabilize the ligation reaction (Weterings and Gent 2004; Britt 1999; Sengupta and Harris 2005). Because there is no mechanism to ensure pairing of the two original chromosome ends, NHEJ is an error-prone mechanism which can lead to deletions, chromosomal inversions, translocations and partial duplications (Britt 1999). More energetic or longer radiation exposures generally lead to higher number of breaks, which in turn increase the frequency of DNA repair and larger aberrations (Hlatky et al. 2002).

### 1.1.1. Considerations on the effect of radiation for the development of mutant populations

The peculiarity of the mechanism of DNA breakage and repair is exploited in radiation population studies to create a binary polymorphism (0-deletion vs. 1-retention). Physical, physiological, or molecular markers can then be employed to tag this binary polymorphism. The level of linkage or association between the various markers can be determined on the basis of the frequencies at which these are simultaneously deleted (co-deletion) or retained (co-retained). Markers that are affected at the same rate indicate good association. When this occurs between DNA markers a molecular map can be derived; when the association occurs between physical markers instead then a scaffold or physical map is obtained; finally, when the association occurs between DNA and

phenotype markers it becomes possible to refine the location of the functional gene(s) or QTL controlling that specific phenotype.

Radiation damage completely lacks any target specificity. Thus, DSBs occur and are repaired independently in each cell. Cells of the same tissue which differ due to mutations are commonly defined 'chimeric'. Similarly, each chromosome/chromatid is affected differently by radiation quanta passing through the cell nucleus. Two sister chromatids which differ due to mutation are also defined 'chimeric' or 'heteromorphic'. At the gene level, aberrant chromatids are defined 'hemizygous', to indicate a locus that is present (retained) on one of the chromatid but absent (deleted) in its sister chromatid. At meiosis, the heteromorphic chromatids will segregate in a Mendelian fashion, resulting in a 1:2:1 ratio of nullizygous (i.e. the homozygous for deletion) to hemizygous to homozygous (i.e. homozygous retention) for any given loci. Since it is very unlikely that deletions randomly occur in the same position on both chromatids, after radiation treatment a generation 1 mutant ($M_1$) individual has the vast majority of its mutagenized loci in hemizygous state and its tissues are vastly chimeric. After self pollination of this $M_1$ individual, half of the resulting progenies ($M_2$) will be still hemizygous at most loci, while the remaining will segregate rather toward homo- or nulli- zygousity. This phenomenon is graphically evident in Figure 1.1 where a *Triticum aestivum* L. $M_2$ line, derived from γ-ray treatment of the boron toxicity tolerant cultivar 'Halberd', was found hemizygous for the boron toxicity tolerance locus (*Bo1*; Schnurbusch et al. 2007), as shown by average length of roots after treatment in boron toxic media. After selfing of this mutant line, its $M_3$ progeny segregates for the *Bo1* locus in a 1:2:1 ratio with long (homozygous), average (hemizygous) and short (nullizygous) roots length after exposure to toxic boron (courtesy of Profesor Peter Langridge, Australian Center for Plant Functional Genomics, Adelaide, Australia).

At meiosis, the chimeric mega and microspore cells will segregate into different progenies depending on their genetically effective cell number (GECN). This is the number of cells that effectively take part in gametogenesis, in animals this number is equal to one, but in plants it varies by species. The model plant *Arabidopsis thaliana* has a GECN=2 (Page and Grossniklaus 2002). Hence, after selfing of an irradiated *Arabidopsis* $M_1$ line, a segregation ratio of 5 normal: 2 hemizygous: to 1 homozygous mutant is expected. GECN has been calculated for some of the plant species and the information available to date are summarized in Table 1.1. The short list of plant species for which

GECN is currently known indicates how this aspect is often underestimated when mutant populations are developed.

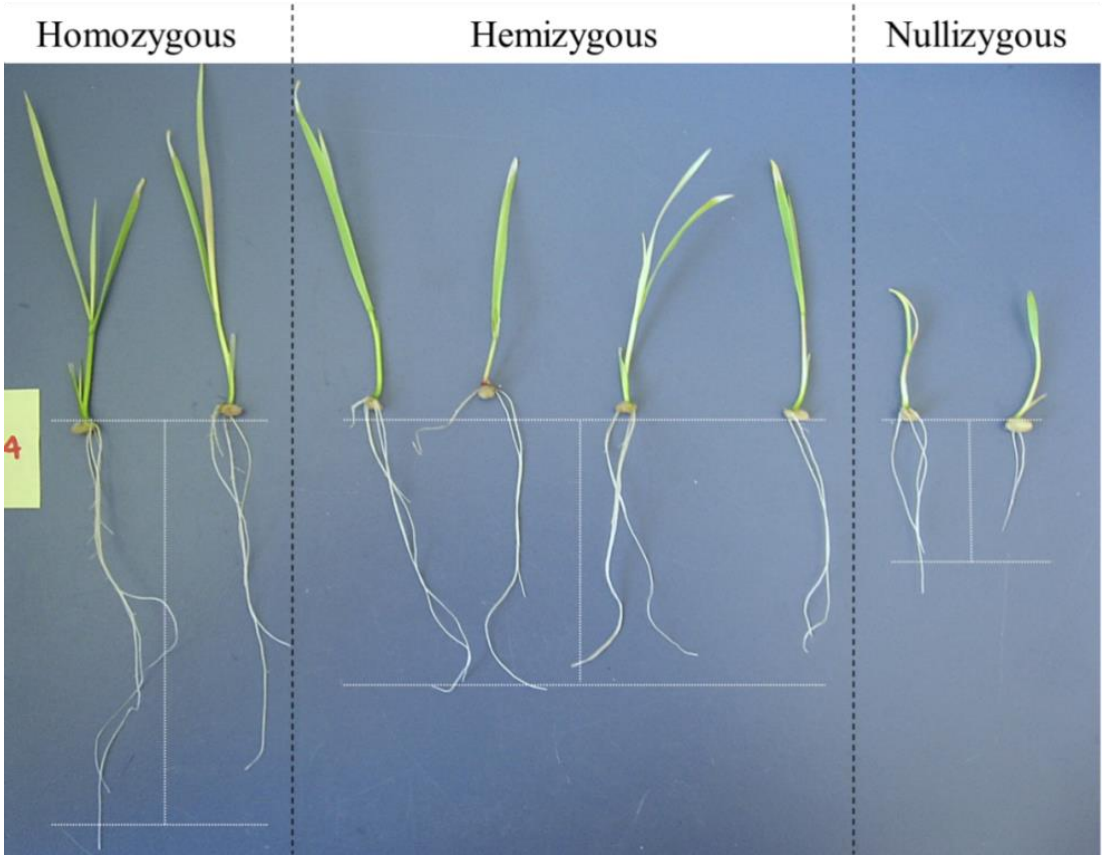


Figure 1.1. Mendelian segregation of $M_3$ progenies. The sibling lines were derived from the selfing of a $M_2$ *T. aestivum* var. 'Halberd' hemizygous for the *Bo1* toxicity tolerance locus (Schnurbusch et al. 2007), as shown by the segregation in roots length after exposure to toxic levels of boron. The $M_3$ progenies segregate for the *Bo1* locus in a 1:2:1 ratio with long (homozygous), average (hemizygous) and short (nullizygous) roots length after exposure to toxic boron media. (Figure generated at ACPFG).

Table 1.1. Genetically effective cell number (GECN) of plant species

| Species | GECN | Reference |
|---|---|---|
| *Glycine Max* | 2 | Carroll et al. 1988 |
| *Arabidopsis thaliana* | 2 | Page and Grossniklaus 2002 |
| *Nicotiana plumbaginifolia* | 2 | Blonstein et al. 1991a-b |
| *Medicago sativa* | 3 | Le Signor et al. 2009 |
| *Linum usitatissimum* | 4 | Bertagne-Sagnard et al. 1996 |
| *Hordeum vulgaris* | 6 | Hodgdon et al. 1981 |
| *Triticum aestivum* | >5 | Kumar et al. 2012 |

It is of primary importance to understand the three levels of polymorphism generated by a radiation treatment (cell, chromosome, and locus). A broadly segregating population can be obtained using multiple progenies from a single $M_1$ mutant line, since the combination of a large GECN and the hemizygous state of the loci provide nearly never ending combinations. Still, while this represents a great advantage to rapidly produce very large populations starting from very few mutant lines, on the other side it makes nearly impossible to identify one useful genotype multiple times. Probably the impossibility to adequately replicate genotypes is the biggest disadvantage of radiation population.

A few additional considerations can be derived from the nature of radiation to maximize the effectiveness of mutant populations. In order to maximize the genetic variability, the number of aberrations generated for each mutant individual should be maximized. Practically, it is necessary to maximize the amount of damage caused by radiation. The simplest way is to follow these four suggestions: 1) irradiate soft tissue with limited resistance to energy penetration, 2) maximize the amount of water present within the tissue, 3) increase radiation dosages (longer exposure time and shorter distance from the radioactive source), and/or 4) use high energy radiation sources.

### 1.2. Radiation hybrid for mapping genomes

Large chromosomal rearrangements are generated upon exposure of live tissues to radiation treatments. These rearrangements are easily noticeable as deletions of chromosomal fragments. Since LET are expected to affect randomly the genome, the frequency at witch two loci are deleted or retained together depends entirely on the probability of them to be affected simultaneously. This probability is direct function of their physical proximity: the closer are the two loci along the genome, the higher is the probability of them to be affected simultaneously. Goss and Harris (1975) were the first to employ this principle to order marker loci of the human chromosome X. Their approach was to treat tissue culture cells from a donor species with a lethal dose of X-ray radiation. Random chromosomal breaks are induced as consequence of the treatment. These chromosomal fragments are then rescued by fusion with a suitable recipient cell line from a different species so that the resulting fusion cell lines contain a fraction of the donor genome, retained as a collection of genomic fragments integrated into the recipient genome. Due to the radiation treatment employed and the hybrid nature of the cells that were created, this approach was defined radiation hybrid (RH) mapping,

and the resulting hybrid cells were named RH lines. In plants, it was devised an approach to create viable (*in vivo*) RH panels (Riera-Lizarazu et al. 2000). In this case, the radiation treatment is followed by gametic hybridization, rather than cellular fusion, between a mutagenized and a non-mutagenized plant. The resulting seeds from that type of crossing have been defined RH, as they are derived from radiation plus hybridization. However, it has to be noted that their nature is different than *in vitro* RH panels; *in vivo* RH plant cells are not hybrid in *senso stricto* because do not contain the DNA of two different species as they are not derived from direct cellular fusion . In both *in vivo* and *in vitro* systems, a set of RH lines representing the entire donor genome is then characterized with molecular markers. As the radiation induced breaks are assumed random, the farther apart marker loci are, the higher is their chance to be broken apart and carried on separate chromosomal fragments. By calculating the frequency of co-retention (or co-deletion) between markers, the distance between them and their order relative to one another is determined, in a manner analogous to meiotic mapping (Gautier and Eggen 2005). The RH map unit of distance is centi-Rays (cR), which corresponds to one break between two markers in every 100 lines (i.e. similar to centi-Morgan).

### 1.2.1. The four key aspects of radiation hybrid mapping

After the initial development of the RH mapping technique by Goss and Harris (1975), the intrinsic advantages of this methodology became clearly evident after the work of Cox et al. (1990). These authors constructed a RH map of a 20 Mb fragment of human chromosome 21 with 14 marker loci, reporting on how a continuous map could be generated on the basis of the markers deletion frequency. Moreover, their small population of 103 hybrids outperformed other mapping approaches, providing very high map resolution (i.e. the ability to separate the markers into different map loci) with a limited number of lines. Also, since the genotyping is based on presence (retention) vs absence (deletion) of the probes employed, monomorphic markers could be scored with no need for allelic polymorphism discrimination. Finally, they reported that by varying the dose of radiation employed it was possible to modify the frequency of deletions and hence generating fragments of various sizes (Figure 1.2). As discussed in that publication, the quality of a RH study can be measured on the basis of four key aspects: population size, radiation dosage, deletion frequency, and resolution (usually expressed in physical units, Kb or Mb per cR).

1.2.1.1. Uniform and higher mapping resolution along the chromosome

In genetic mapping, frequency of recombination as a result of crossover events determines the distance between two adjoining loci. However, distribution of meiotic crossover events along chromosomes is not uniform (Akhunov et al. 2003; Erayman et al. 2004; Mézard 2006; Saintenac et al. 2009). For instance, a recent study on wheat chromosome 3B indicated that crossover frequency per physical distance (cM / Mb) within the centromeric segments, as defined by wheat deletion bins, were between 0.006 to 0.012 compared with 0.85 in the sub-telomeric segments (Saintenac et al. 2009). Thus centromeric regions recombine hundred folds less than the telomeric portions of the chromosome. It has been estimated that less than 1% recombination occurs in ~30% of the proximal regions of the wheat chromosomes (Erayman et al. 2004). In practice, the resolution of genetic maps varies depending on the specific chromosomal region, and it is typically very low in proximity of the centromeres. On the other hand, RH mapping studies do not rely directly on crossing over and recombination to sort markers along the chromosomes, but rather radiation-mediated breaks, and are then assumed to provide unfirom resolution along the length of chromosomes (Gross and Harris 1975; Cox et al. 1990). Also, the RH approach has been shown to provide much higher map resolution in recombination poor regions, while maintaining a more uniform physical to map conversion (i.e. resolution) along the length of chromosomes.

The resolution in radiation hybrid mapping is determined by the number of the breaks occurring on the genome. The amount of DNA damage can be adjusted by increasing or reducing the dosage of radiation. Thus, small RH panels can be created at different dosages to provide low to high levels of map resolution, without actually increasing the population size (Weikard et al. 2006; Karere et al. 2008). A good example on how radiation dosage affects the amount of damage substained by an organism is presented in Figure 1.2, which shows the *Triticum turgidum* ssp. *durum* L. cvr. 'Colosseo' growing in a field in Italy after seed treatment at 150, 100, and 50 Gy of gamma rays (picture was obtained at DiSTA, Bologna, Professor Roberto Tuberosa). The high map resolution and relatively small population size, elected RH mapping as an ideal approach to construct accurate markers scaffold for hierarchical (BAC-contig) genome sequencing projects (Faraut et al. 2009).


Figure 1.2. Response of durum wheat var. 'Colosseo' at different doses of gamma rays. The picture is taken at Cadriano experimental station in Italy. Photo courtesy of Prof R. Tuberosa research group.

1.2.1.2. Mapping of monomorphic markers

An additional advantage of RH mapping is that genotyping is performed by scoring presence (retention) vs absence (deletion) of marker loci and it does not require allelic polymorphism. Practically, any marker system that produces a chromosome-specific scoring can be used for the characterization of a RH panel (Figure 1.3). This is not limited to DNA-based markers, but also phenotypes or proteins can be scored for presence/absence, as long as chromosome/locus specificity is provided. Deletion of a single copy gene would inevitably result in the absence scoring of a DNA marker tagging it, the disappearance of the protein produced by that gene, and also the knock out of the phenotype controlled by that locus. The deletion of a gene with multiple homeologus versions is typically less drastic and harder to follow by deletion tagging.

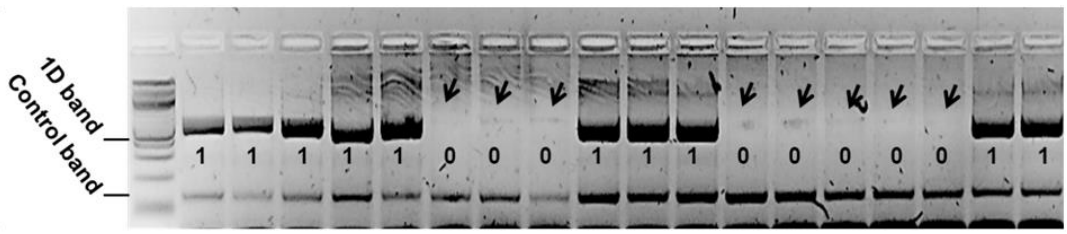


Figure 1.3. Deletions detected on a chromosome 1D RH population. Gel picture for marker NDSU3 described in Chapter 4 of this Thesis. Arrows and ‘0’ indicate radiation-mediated deletions of the marker locus.

In wheat, it was calculated that within genes the frequency of homeologus variations on the DNA was approximately 10 times more frequent (1/60 bp) than the allelic polymorphism between varieties (1/613 bp) (Schnrbusch et al. 2007). In other words, in wheat it is approximately 10 times easier to develop chromosome-specific markers to tag genes than developing allele-specific markers. Therefore, RH panels, which require only chromosome-specificity, constitute one of the most expedient methods to map regions with low genetic polymorphism such as the coding portion of the genome. Genes or EST-derived markers can then be assigned with high resolution to their map location, and provide an efficient tool for functional and comparative genomic studies (Hossain et al. 2004a; Kalavacharla et al. 2006; Kynast et al. 2002; Wardrop et al. 2004; Weikard et al. 2006).

### 1.2.2. Radiation hybrid mapping in animal systems

Since the publication of the initial RH maps by Goss and Harris (1975), this method has been widely adopted by various animal systems. In humans, a 100 kb-resolution contiguous map with ~41,000 ordered STSs covering 30,000 unique human genes was constructed using the RH mapping approach (Stewart et al.1997; Deloukas et al.1998). Successful RH mapping of the human genome pioneered RH mapping of other animals (for review see Faraut et al. 2009). The success of this approach can be measured by the fact that RH maps have played a major role in the whole genome sequencing and assembly of human (Gyapay et al. 1996; International Human Genome Sequencing Consortium 2001; Olivier et al. 2011) as well as many domestic animals (Lander et al. 2001; Waterston et al. 2002; Gibbs et al. 2004; Lindblad-Toh et al. 2005; Bovine Genome Sequencing and Analysis Consortium 2009; Wade et al. 2009; and more).

### 1.2.3. Radiation hybrid mapping in plant systems

Compared with the many published studies conducted in animals, RH projects in plants have been more circumscribed and sporadic, and yet they have yielded very interesting results. The first plant study was conducted by Riera-Lizarazu et al. (2000) in maize (*Zea mays* L.). These authors treated with 300, 400, 500 Gy γ-ray seeds of oat (*Avena sativa* L.) lines carrying a monosomic chromosome 9 of maize. The mature mutant plant (namely $RH_0$ or $M_1$) was then selfed and selected for the presence of fragments of chromosome 9. A total of 55 $RH_1$ (or $M_2$ since these lines are the

result of selfing rather than hybridization) were characterized with 33 maize specific DNA markers and 25%, 27%, and 35% frequency of marker loss (aka deletion frequency) was observed after 300, 400, and 500 Gy treatments, respectively. They estimated an average of three breaks per chromosome per line to indicate a resolution of 1.1 Mb when using just 55 lines, but suggested that resolutions of 0.3 Mb should be achievable quadruplicating the number of $RH_1$ lines in the panel. Furthermore, it was reported that deletion frequency was homogeneous across the whole chromosome, confirming that deletion appear to happen randomly across the genome. This was the first case in which an *in vivo* RH panel was generated, and provided the confidence to other authors to attempt a similar approach in other plant species. Following the same methodology, panels are now available for most maize chromosomes (Kynast et al. 2002; 2004).

Two years later, Wardrop et al. (2002) reported the first *in vitro* RH approach for plants. Transgenic barley (*Hordeum vulgare* L.) protoplasts harboring the bialaphos resistance *bar* transgene, were irradiated at 50 Gy of X-ray. The mutagenized protoplasts were hybridized with tobacco (*Nicotiana tabacum*) protoplasts by electrofusion. Positive fusion calli were selected on the basis of acquired bialaphos resistance. Forty putative $RH_1$ calli were screened with 35 molecular markers and revealed an average deletion frequency of 26%. The RH calli were further screened with two markers spanning a 16 Kb interval, but no breakage was identified to resolve them. This was the first whole genome RH (WGRH) approach described in plants, where the whole genome is simultaneously studied, instead of single chromosomes. These authors, while being optimistic about the potential of applying their approach to all plant species, reported a problem with the low amount of DNA that was obtained from the calli. Fortunately, they quickly improved their methodology to address this issue (Wardrop et al. 2004). Following the same techniques, a WGRH panel for citrus (de Bona et al*.* 2009), was also developed by irradiating the microprotoplasts of *Swinglea glutinosa,* a distant relative of citrus, at different doses of gamma rays (50, 70, 100 or 200 Gy) and fusing it with cv. 'Ruby Red' grapefruit or *Murcott tangor* protoplasts. AFLP analysis was then used to select for the hybrid cell lines and create a RH panel.

In addition to these *in vitro* WGRH panels, Gao et al. (2004) reported the first *in vivo* WGRH. Pollen from mature cotton (*Gossypium hirsutum* L.) plants was irradiated at 50 Gy of γ-rays, and then used to fertilize *G. barbadense* lines*,* and vice versa irradiated pollen of *G. barbadense* was used to

pollinate *G. hirsutum*, to develop WGRH panel for *G. hirsutum* and *G. barbadense,* respectively (Gao et al. 2004). The $F_1$ plants obtained carried deletions for *G. hirsutum* or *G. barbadense,* but since both species have very similar genomes (AADD) only polymorphic markers could be used for the screening. A total of 93 lines composed the RH panel for cotton. Genotyping at 102 microsatellite loci revealed a deletion frequency of 7%, which was much lower than what observe in animals but still sufficient to precisely map 50% of the markers.

Wheat is probably the plant species that has been more extensively studied with RH mapping techniques. Hossain et al. (2004) developed an *in vivo* RH panel for chromosome 1D using a durum wheat line (*Triticum turgidum* ssp *durum*) as pollen donor [13" + 1A"], and as female an irradiated cytoplasmic male sterile durum line that has the 1AS chromosome substituted with a translocated copy of chromosome 1D [13" +1AL.1D"]. Chromosome 1D harbors a *species cytoplasm specific* gene from *Triticum aestivum* L. ($scs^{ae}$), which is necessary to guarantee adequate seed vigor. In the absence of this gene, the alloplasmic line produces shriveled and non-viable seeds upon pollination. One hundred plump viable seeds were obtained by backcrossing (*lo*) $scs^{ae}$- to its maintainer line. These seeds were irradiated at 150 Gy γ-rays, and the resulting $RH_0$ plants were backcrossed again to the maintainer euplasmic durum line. Eighty-seven plump seeds were obtained representing $RH_1$ lines monosmic for irradiated 1D.1AL and still carrying the $scs^{ae}$ gene [13" 1A' + 1AL.1D'$^{150\ Gy}$]. This panel was genotyped with 378 markers that identified a deletion frequency of 26%, with the exclusion of the $scs^{ae}$ region which was always retained. The map identified 2,312 unique chromosome breaks, for a calculated resolution of ~199 kb (Kalavacharla et al. 2006).

In a different *in vivo* study (Paux et al. 2008), seeds of the durum wheat variety 'Langdon' [13" +3B"] were irradiated at 350 and 250 Gy of γ-rays. The resulting $RH_0$ plants were brought to maturity and then used to fertilize the male sterile substitution line of 'Langdon' for chromosome 3B [13"+ 3D"], where chromosome 3B of durum has been replaced with chromosome 3D of hexaploid wheat variety 'Chinese Spring' (Joppa and Williams 1988). The artificial cross pollination produced 184 $RH_1$ lines, monosomic for chromosome 3D and irradiated chromosome 3B [13" +3B'$^{350\ Gy}$ +3D']. These lines were genotyped with 65 3B-specific markers to reveal a deletion frequency of 29% and a calculated resolution of 263 Kb. In addition, this small population was employed to anchor 32 BAC contigs

developed for the physical mapping of chromosome 3B to provide a proof of concept for future RH approaches in plants.

The only published example of *in vitro* WGRH studies for wheat was reported by Zhou et al. (2006). These authors employed UV irradiated protoplasts of *Bupleurum scorzonerifolium* Willd. to fuse with the protoplasts of common wheat (*T. aestivum* L.). DNA marker analysis of the hybrids demonstrated that wheat DNA was integrated into the nuclear genomes of *B. scorzonerifolium* and that deletion frequencies of over 80% were reached.

### 1.2.4. Comparison of animal and plant radiation hybrid projects

Simulation studies have suggested that a retention/ loss frequency of 50% would be optimal for mapping purpose (Jones 1996). However, this ideal number was hardly reached in any RH project to date. Typically, *in vitro* RH approaches treat very soft and water rich tissues at low radiation dosages, to ultimately obtain high deletion frequencies (>80%). In these projects, RH lines are typically selected because of their lower deletion frequencies (<70%) and mapping panels are so created. *In vivo* RH projects usually face the opposite problem, with relatively hard and dehydrated tissue (i.e. seeds or pollen grains) treated at high radiation dosages, to generate $RH_1$ organisms with high retention frequencies (>95%, or deletion frequencies of <5%). In this case, RH lines are selected to maximize the deletion frequency (>10%) in creating the mapping panels. Practically, both approaches have similar issues but caused by the opposite problems. A simple explanation of the differences observed between *in vivo* and *in vitro* RH panels probably resides in the very nature of these approaches. *In vivo* studies generate viable organisms that need to survive environmental stresses. The irradiated cells need to first proliferate into a fully formed organism, then produce gametes, and ultimately a $RH_1$ progeny. Obviously, an organism cannot lose more than a small amount of its genome before encountering very serious life-threatening damages, which ultimately cause decrease in seed germination, plant survival, and vigor (Riera-Lizarazu et al. 2000, 2010; Gao et al. 2006). *In vitro* RH experiment, on the other hand, rely on a second cell to rescue the genome of the irradiated one. So forth, the damaged cell does not need to proliferate nor to generate a viable organism, and can then 'tolerate' drastic amount of genome fragmentation. However, the rescuing protoplast cell needs to proliferate and continue its normal behavior on culture media or as part of a

viable organism. In this case, the rescuing cell in order to continue its activity can only tolerate a small amount of new genome fragments being added to its own. In fact, the inclusion in the genome of the rescuing cells of too many fragments, from the radiated protoplast, is likely to interfere with its normal nuclear biology and ultimately affect the cell ability to adequately function.

Somatic cell hybridization, although extensively used for gene mapping in animal system, has found only limited applications in plants due to the valid alternative of developing *in vivo* panels. The protoplast fusion approach (Wardrop et al. 2002, 2004; Zhou et al. 2006), although requires less generation time and space than an *in vivo* panel, it is cumbersome and it is not always possible to find an appropriate recipient cell lines. Another issue is the genetic stability of the hybrids, as most plant cells were found to be unstable in tissue culture (Song et al. 2000; Wardrop et al. 2002). However, for those species where fusion protocols are available, *in vitro* approaches provide an excellent methodology for RH mapping. Most of the RH approaches available to date in plants employ wide or interspecific artificial pollinations to develop *in vivo* RH populations (Riera-Lizarazu et al. 2000, 2010; Gao et al. 2004, 2006; Kynast et al. 2004; Hossain et al. 2004). These viable populations can then be maintained through self-pollination or backcrossing, seeds are increased, and can be shared with the scientific community, providing unlimited sources of DNA. Also, these viable individuals can be phenotyped and used for functional studies.

However, there are few obvious limitations when generating viable RH lines. An organism cannot survive the deletion of specific life-dependent genes; hence the map resolution is expected to drastically diminish in proximity of these loci. The *scs*$^{ae}$ gene is a good example of this (Kalavacharla et al. 2006), where no deletions could be recovered in proximity of this locus. Similarly, an individual can only stand a limited amount of genome removal before losing important physiological functions. The data available to date for plants indicate that *in vivo* RH for single chromosomes and for whole genomes have slightly different deletion frequencies, with approximately 10-20% against 1-3%, respectively. This can easily be explained considering the total loss: for instance 20% of the 604 Mb size of chromosome 1D of wheat equals to 120 Mb, very similar to the 148 Mb corresponding to the 3% of the whole D-genome (4,936 Mb total size). In this prospective, WGRH might benefit from *in vitro* methodologies to increase the overall deletion frequency more than single chromosome RH studies. Fortunately, *in vivo* RH can be easily developed through artificial pollination and very large

WGRH populations can be rapidly generated, providing sufficient starting material to adequately enrich with highly deleted lines the WGRH mapping panels.

## 1.3. Applications and prospects for radiation hybrids in plants

Over the years, radiation-based mutant populations have proven extremely useful for breeding, with nearly three thousands varieties released (Maluszynski et al. 2000; Ahloowalia et al. 2004; Shu 2009), and to determine gene functions through knock-out mutations (Morita et al. 2007; Lloyd and Meinke 2012). Radiation hybrid mapping offers additional bright prospectives to mutagenomics in plants.

### 1.3.1. Radiation hybrids for mapping and cloning genes

Genes are the basic physical unit of heredity, but can also be considered as the basic unit of research, since identifying, understanding, or modifying genes is ultimately the scope of nearly all biological studies. *In vivo* RH can help in two of these three aspects: understanding and identification.

Viable RH are basically knock-out mutants, with multiple coding portions of their genome removed. Radiation treatments in *Arabipodsis thaliana* and *Oryza sativa* generated recessive monogenic phenotypes, confirming that radiation causes knock-out mutations (Matsukura et al. 2007; Koornnef et al. 1982). The phenotypic mutants that were observed included gibberillic sensitivity, hypocotyl morphology changes, narrow leaves, double flowers and trichome polymorphism (Koornnef et al. 1980, Koornnef et al. 1982). A subset of this mutant population in *Arabidopsis*, led to the cloning of two genes involved in the *testa color* phenotype (Shirley et al. 1992), and to the localization on chromosomal regions of genes controlling plant height and flowering time (Koornnef et al. 1991). Today, knock-out mutants are more commonly produced by transformation or site-direct mutagenesis (Lloyd and Meinke 2012). Still, the knock-out radiation mutant populations represent very valuable tools to study genes and the effect that the removal of these genes has on the phenotype of an organism. It can be derived that RH are also carrier of knock-out mutations that can help in understanding the functions of the deleted genes.

However, the greatest power of RH is the combination of the knock-out mutation, with their high map resolution, together with the simplicity of developing and scoring monomorphic markers. The

ensemble of these aspects suggests RH as an ideal biological material for the positional cloning and fine mapping of genes. Map based cloning using genetic mapping populations is a time consuming and mostly unsuccessful procedure, once defined by Peters and colleagues (2003) as difficult as 'reaching to the top of Mont Everest". Several major problems arise when attempting the positional cloning of a gene (Salvi and Tuberosa 2005): *i*) the need to have available genetic diversity for the targeted trait/gene, *ii*) the requirement of generating a population segregating in a simple Mendelian fashion for the trait of interest (Mendelize), *iii*) the necessity to produce a population of large enough size to guarantee sufficient map resolution, and *iv*) the difficulty of developing and mapping polymorphic markers (both phenotypic and DNA-based).

Radiation hybrids provide ideal solutions to most of these issues. *i*) In RH the genetic diversity is created by radiation mutagenesis. Even traits or genes for which genetic variations are not known can be knock-out, to create a segregating population. *ii*) Viable RH populations for functional studies are obtained through artificial pollinations, crossing specific parental combinations that result in $RH_1$ progenies segregating only for the chromosome of interest. The mutations occurring in the non-targeted chromosomes are masked by their counterpart derived from the non-irradiated parent, while the target chromosome is carried in the homozygous (monosomic) condition, allowing all mutations to be exposed. Under these conditions, only the traits carried by the target chromosomes are subject to knock-out mutation, offering a simple but effective way to Mendelize single genes or single QTLs located on the target chromosome. However, this form of Mendelization does not apply when multiple genes/QTLs for the same trait exists on the same chromosome. *iii*) Large populations can be rapidly generated from few greenhouse crossings and brought to homozygousity in just one season. Further, the map resolution of RH is 10-100 folds higher than what provided by a genetic population of similar size. *iv*) The possibility of using monomorphic markers in RH has been amply discussed previously. The only additional consideration is that the development of markers, to obtain complete map saturation, is usually the most time consuming step in positional cloning projects. The fact that RH-mediated map based cloning directly employs monomorphic/chromosome-specific markers, represents by itself probably the most advantageous aspect of this approach as compared to genetic mapping. Additionally, the RH map distances can be converted with relative precision into actual physical distances (Cox et al. 1990).

A good example of the potential application of RH for fine mapping was a study conducted on the *scs*$^{ae}$ gene in durum wheat (Hossain et al. 2004). This gene was identified in an alloplasmic durum line that carries a portion of chromosome 1D from *T. aestivum*. The chromosome 1D of this line segregates as a whole without recombination, making it impossible to use genetic mapping to refine the location of the gene. Instead, a RH mapping population of 87 lines allowed identifying 8 markers linked to *scs*$^{ae}$ on the long arm of chromosome 1D, spanning a calculated region of less than 10 Mb.

Radiation-induced deletion lines were used to narrow down the location of a homeologous pairing suppressor (*Ph1*) locus, located on chromosome 5B of hexaploid wheat (Roberts et al. 1999). In order to identify deletions of the region spanning the gene of interest, various X-ray, γ-ray, as well as fast neutron mutant populations were generated. The *Ph1* locus was first pinpointed to a narrow region, to then identify two candidate genes by mean of differential expression analysis among radiation mutants (Al-Kaff et al. 2008).

In another study in wheat, genetic mapping and chromosome walking allowed to narrow the region harboring the *Q* locus to a small number of cosegregating markers (Faris et al 2003). The use of radiation mutants helped to further reduce the number of candidate genes cosegregating with the phenotype, ultimately leading to the cloning and functional characterization of the gene (Faris et al. 2003; Simmons et al. 2006).

There are a few considerations that need to be kept in mind when using RH to find genotype to phenotype associations. Genomes that are broken by radiation tend to reassemble in random order. Translocations and inversions are common phenomena that originate from the NHEJ repair of the DNA (Britt 1999). It is then crucial to understand that the movement of a gene from its original location to a new chromosomal region might affect the phenotype controlled by this gene, but without necessarily causing a deletion of the locus. For instance, a gene that is moved to the centromere would likely undergo epigenetic silencing, by means of a change in its chromatin compaction or in its methylation status, which are common phenomenon in this region. The result is the knock-out of the gene phenotype without its actual deletion. A DNA marker tagging the gene would not detect any changes, while a phenotype marker would. When this event occurs there would be no detectable association between genotype and phenotype. Similarly, a gene could be moved away from its promoting/enhancing sequence. Again, this would result in the silencing of its expression, a

consequent change in the phenotype, but without a detectable change in the deletion status of the locus. Also, radiation-damage on the DNA often results in single nicks, which never develop into DSBs (Britt 1996). Single base mutations or small INDELs can then occur on the coding sequence of the causative gene. In the worst case scenario, the transcription or translation of the gene is affected by this small mutation, the knock-out phenotype occurs, but once again the marker used to tag this gene does not register any deletions.

All of the hypothetical cases described above would inevitably result into non-association between the deletion/retention scoring of the marker tagging the target gene and its phenotypic scoring. Unfortunately, there is no easy fix to this problem. The only way to prevent complete failure of the map based cloning approach is to use large enough populations. Since the occurrence of these events is rare, it is probably best to proceed ignoring their possibility, and only once in sight of the best candidate genes start reconsidering all the associations. Digestion with methylation sensitive enzymes or re-sequencing of the candidate genes from those lines that do not show association (especially negative phenotypes with retained marker scorings) is a good strategy to recognize this type of issues. In some cases, these additional analyses might even result in the identification of novel mutants that confirm the good quality of the positional cloning study.

### 1.3.2. Radiation hybrid mapping and comparative genomics

Comparative genomic studies are used to determine the evolutionary conservation across species. A comparison of the gene order between different genomes provides information on the evolutionary importance of specific segments of DNA, other than determining the ancestral relationships that lead to the origin of the different organisms (O'Brien et al. 1999). An alternative use of gene order conservation consists in predicting the coding element of one species, on the basis of the gene order of a second species, for which the information is available. These are commonly defined as ‘synteny’ or ‘colinearity’ studies.

Radiation hybrids map are particularly useful in these types of studies. As discussed previously, mapping of genes is particularly simple in RH population. Once a scaffold of gene-markers has been placed along the length of a chromosome by RH mapping, colinearity analysis with one or more fully sequenced model species can define the existing orthologus relationships.

After completion of the human genome project, it became possible predicting the gene space through colinearity of other species that did not have the budget to complete a full genome sequencing endeavor. For example, using 93 RH cell lines and 84 loci, a comparative map was generated for chromosomes 7 and 9 of *Macaca mulatta* by colinearity to chromosomes 14 and 15 of humans. This study revealed a complex evolutionary scenario, which involved a loss of the interstitial centromere in the ancestral 14-15 chromosome and evolution of two novel centromeres at the end of both chromosomes (Murphy et al. 2001). A second study, reported a high-resolution RH map of segments of the sheep chromosomes 8, 9 and 20 which were orthologus to human chromosome 6. This map was generated with 251 markers using a WGRH panel. Two of these syntenic regions had been previously identified by means of chromosome painting, but it also revealed a previously unknown homology between a small centromeric region on chromosomes 9 of sheep and 6 of humans (Wu et al. 2008). This illustrates the power of RH mapping in detecting small regions of colinearity, which in turn provide essential information for the understanding of the evolution of the species.

A large number of studies in mammals used whole genome comparative analysis to explore chromosome evolution. This was possible not only because of the availability of the complete genome sequences, but also due to well-established RH mapping procedures. Murphy et al. (2005) integrated the information from the RH mapping of cat, cattle, dog, pig, and horse with the information of the whole genome sequences of human, mouse and rat to address fundamental evolutionary questions. These authors identified 1,159 pairwise homologous synteny blocks between the genomes of human and the six other species, generating a complete evolutionary scenario that depicted the rearrangements that occurred during the speciation of these organisms.

In plants, the comparative studies are mostly restricted to genome analysis of species for which the full sequence is available (Vogel et al. 2010). In wheat, an attempt to estimate the overall colinearity conservation employed EST-based RFLP probes in combination with low resolution genetic maps and the so called ‘chromosomal bin maps’ (Luo et al. 2009; Salse et al. 2008). While the quality of these studies is beyond doubts, the low resolution provided by these approached allowed only for the estimation of the macrosynteny conservations, but could not precisely investigate microsynteny, especially in those regions that lack in recombination events. The high map resolution

and the simplicity of monomorphic gene-based markers scoring, makes of RH an ideal approach for comparative studies at the micro level (Michalck et al. 2009).

Development of RH maps has contributed extensively in animal systems to comparative analysis and evolutionary studies. In plants, the use of RH has been limited, probably due to the simplicity of developing genetic populations (as compared to animals). Also, the limited genomic resources available in the past have partially prevented the real need for high resolution synteny studies. As I move forward, it might be important to recognize that RH mapping in plants offers a new strategy for mapping monomorphic gene-based markers, and also provides a resolution more suitable for studying syntenic relationships at the micro level (micro-synteny).

### 1.3.3. Radiation hybrids and the new sequencing technologies

In the past few years, several massively parallel next(3$^{rd}$)-generation sequencing platforms have been developed, drastically dropping the cost of obtaining DNA sequences (Gupta 2008; Metzker 2010). The rapid and cheap generation of large amount of sequence data has dramatically accelerated genetic and genomic research. Unfortunately, the cost of generating and assembling the complete sequence of a genome remains still prohibitive for all of those species that are not financed by their large economic returns. Similarly, species with large genomes require complex hierarchical sequencing projects, which often struggle to find adequate fundings. This has been so far the case for wheat, barley, and many other important crops, even though things have begun changing in the past few years (Feuillet et al. 2011).

A cheaper alternative to these methodologies is what has become known as ‘survey sequencing’. This is a simple scan of fragments of the genome at 1-2x coverage, which provides sufficient data to generate markers, predict most of the coding regions (especially when enriched for it), derive the genome composition in terms of mobile elements and satellite repeats (Hitte et al. 2005; Donthu et al. 2009). Although, survey sequencing lacks the long range continuity required for whole genome assembly, it still provides valuable genomic resources for mapping studies (Feuillet et al. 2011). Also, when used in combination with synteny studies of a related model species, it can reveal the order of conserved genes and its evolutionary importance. Survey sequencing combined with RH mapping proved valuable in extracting genomic information at low cost. In dogs, 9,000 gene markers

were derived from its 1.5x survey sequencing and sorted on a high-resolution RH map (Kirkness et al. 2003) to identify homologus among the 18,201 genes predicted on the basis of 7.5x assembly (Hitte et al. 2005). The authors concluded that RH mapping combined with 0.5-1x survey sequence information is capable of generating a high-quality map of the coding portion of the genome, which in turn offers an ideal resource for evolutionary studies and an advantageous starting point for molecular breeding endeavors.

When approaching the large genomes of plant species, rich in repetitive DNA and highly duplicated, the sequencing and computation costs rise very rapidly. Even though the entirety of a genome is important to fully grasp the molecular behavior of an organism, biologists and breeders often prefer to focus on the low copy coding sequences. Sequencing of just the genic portion by transcriptome sequencing (Barbazuk et al. 2007; Novaes et al. 2008) or reduced representation genomic libraries (Van Orsouw et al. 2007) can substantially reduce the costs, without excessive loss of genomic information. Once available, these information could then be converted into precise genome representations by mean of RH mapping. Thousands of molecular markers could be developed from these genic sequences, and then scored on RH populations to produce high-resolution maps for downstream applications, such as fine mapping, evolutionary studies, and expression analysis among others. Especially in those species that are not likely to obtain the funding for generating a full genome sequence, the combination of RH mapping and survey sequencing is certainly an approach worth considering.

## 1.4. References

Ahloowalia BS, Maluszynski M, Nichterlein K (2004) Global impact of mutation-derived varieties. Euphytica 135:187-204

Ahmed EA, Philippens MEP, Kal HB, de Rooij DG, de Boer P (2010) Genetic probing of homologous recombination and non-homologous end joining during meiotic prophase in irradiated mouse spermatocytes. Mutation Research 688:12-18

Akhunov ED, Goodyear AW, Geng S, Qi Li-Li, Echalier B, et al (2003) The organization and rate of evolution of wheat genomes are correlated with recombination rates along chromosome arms. Genome Res 5:753-763

Al-Kaff N, Knight E, Bertin I, Foote T, Hart N, et al (2008) Detailed dissection of the chromosomal region containing the Ph1 locus in wheat Triticum aestivum: with deletion mutants and expression profiling. Ann Bot 105:6075-6080

Argonne National Laboratory (2005) Ionizing radiation. Human Health Fact Sheet (August)

Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS (2007) SNP discovery via 454 transcriptome sequencing. Plant J 51:910-918

Bertagne-Sagnard B, Fouilloux G, Chupeau Y (1996) Induced albino mutations as a tool for genetic analysis and cell biology in flax (Linumusitatssimum). J Exp Bot 47:189-194

Blonstein AD, Parry AD, Horgan R, King PJ (1991a) Acytokinin-resistant mutant of Nicotiana plumbaginifolia is wilty. Planta 183:244-250

Blonstein AD, Stirnberg P, King PJ (1991b) Mutants of Nicotiana plumbaginifolia with specific resistance to auxin. Mol Gen Genetics 228:361-371

Bovine Genome Sequencing and Analysis Consortium (2009) The genome sequence of taurine cattle: a window to ruminant biology and evolution. Science 324:522-528

Britt AB (1996) DNA damage and repair in plants. Annu Rev Plant Physiol Plant Mol Biol 47:75-100

Britt AB (1999) Molecular genetics of DNA repair in higher plants. Trends Biotechnol 4:20-25

Carroll BJ, Gresshoff PM, Delves AC (1988) Inheritance of supernodulation in soybean and estimation of the genetically effective cell number. Theor Appl Genet 76:54-58

Cox DR, Burmeister M, Price ER, Kim S, Myers RM (1990) Radiation hybrid mapping: a somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. Science 250:245-250

de Bona CM, Stelly D, Miller JC, Louzada ES (2009) Fusion of protoplasts with irradiated microprotoplasts as a tool for radiation hybrid panel in citrus. Pesq Agropec Bras 44:1616-1623

Deloukas P, Schuler GD, Gyapay G, Beasley EM, Soderlund C, et al (1998) A physical map of 30,000 human genes. Science 282:744-746

Donthu R, Lewin HA, Larkin DM (2009) SyntenyTracker: a tool for defining homologous synteny blocks using radiation hybrid maps and whole-genome sequence. BMC Res Notes 2:148-149

Erayman M, Sanduh D, Sidhu D, Dilbirligi M, Baenzinger PS, et al (2004) Demarcating the gene-rich regions of the wheat genome. Nucleic Acids Res 32:3546-3565

Faraut T, de Givry S, Hitte C, Lahbib-Mansais Y, Morisson M, et al (2009) Contribution of radiation hybrids to genome mapping in domestic animals. Cytogenet Genome Res 126:21-33

Faris JD, Fellers JP, Brooks SA, Gill BS (2003) A bacterial artificial chromosome contig spanning the major domestication locus Q in wheat and identification of a candidate gene. Genetics 164:311-321

Feuillet C, Leach JE, Rogers J, Schnable PS, Eversole K (2011) Crop genome sequencing: lessons and rationales. Trends in Plant Science 16:77-88

Flakus FN (1981) Detecting and measuring ionizing radiation - a short history. IAEA Bulletin 23:31-36

Gao W, Chen ZJ, Yu JZ, Raska D, Kohel RJ, et al (2004) Wide-cross whole-genome radiation hybrid mapping of cotton (Gossypium hirsutum L.). Genetics 167:1317-1329

Gao W, Chen ZJ, Yu JZ, Kohel RJ, Womack JE, Stelly DM (2006) Wide-cross whole-genome radiation hybrid mapping of the cotton ( Gossypium barbadense L.) genome. Mol Gen Genomics 275:105-113

Gautier M, Eggen A (2005) The construction and use of radiation hybrid maps in genomic research In: Eggen A (ed) Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics. John Wiley & Sons, London

Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, et al (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428:493-521

Goss SJ, Harris H (1975) New method for mapping genes in human chromosomes. Nature 255:680-684

Gupta PK (2008) Single-molecule DNA sequencing technologies for future genomics research. Trends Biotechnol 26:602-611

Gyapay G, Schmitt K, Fizames C, Jones H, Vega-Czarny N, et al (1996) A radiation hybrid map of the human genome. Hum Mol Genet 5:339-346

Hitte C, Madeoy J, Kirkness EF, Priat C, Lorentzen TD, et al (2005) Facilitating genome navigation: survey sequencing and dense radiation-hybrid gene mapping. Nat Rev Genet 6:643-648

Hlatky L, Sachs RK, Vazquez M, Cornforth MN (2002) Radiation-induced chromosome aberrations: insight gained from biophysical modeling. Bioessays 24:714-723

Hodgdon AL, Marcus AH, Arenaz P, Rosichan JL, Bogyo TP, Nilan RA (1981) Ontogeny of the barley plant as related to mutation expression and detection of pollen mutations. Environmental Health Prospectives 37:5-7

Hossain KG, Riera-lizarazu O, Kalavacharla V, Vales MI,.Maan SS Kianian SF (2004) Radiation hybrid mapping of the species cytoplasm-specific ($scs^{ae}$) gene in wheat. Genetics 168:415-423

International Commission on Radiological Units and Measurements (1962) Clinical dosimetry; recommendations of the International Commission on Radiological Units and Measurements. National Bureau of Standards 87

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860-921

International System of Units (1980) 268A-1980IEEE International System of Units conversion factors card. Digital Identifier 10.1109/IEEESTD.1980.88275.

Jones HB (1996) Hybrid selection as a method of increasing mapping power for radiation hybrids. Genome Res 6:761-769

Joppa LR, Williams ND (1988) Landgon durum disomic lines and aneuploidy analysis in tetraploid wheat. Genome 30:222-228

Kalavacharla V, Hossain K, Gu Y, Riera-Lizarazu O, Vales MI, et al (2006) High-Resolution radiation hybrid map of wheat chromosome 1D. Genetics 173:1089-1099

Karere GM, Froenicke L, Millon L, Womack JE, Lyons LA (2008) A high-resolution radiation hybrid map of rhesus macaque chromosome 5 identifies rearrangements in the genome assembly. Genomics 92:210-218

Kirkness EF, Bafna V, Halpern AL, Levy S, Remington K, et al (2003) The dog genome: survey sequencing and comparative analysis. Science 301:1898-1903

Koornneef M, van der Veen JH (1980) Induction and analysis of gibberellin-sensitive mutants in Arabidopsis thaliana (L.) Heynh. Theor Appl Genet 58:257-263

Koornneef M, Dellaert LWM, van der Veen JH (1982) EMS-and radiation-induced mutation frequencies at individual loci in Arabidopsis thaliana (L.) Heynh. Mutat Res 93:109-123

Koornneef M, Hanhart CJ, Veen JH (1991) A genetic and physiological analysis of late flowering mutants in Arabidopsis thaliana. Mol Gen Genet 229:57-66

Kovacs E, Keresztes A (2002) Effect of gamma and UV-B/C radiation on plant cells. Micron 33:199-210

Kumar A, Simmons K, Iqbal MJ, Michalack de Jimenez M, Bassi FM, et al. (2012) Physical mapping resources for large plant genomes: radiation hybrids for wheat D-genome progenitor Aegilops tauschii. BMC Genomics (*accepted*).

Kynast RG, Okagaki RJ, Rines HW, Phillips RL (2002) Maize individualized chromosome and derived radiation hybrid lines and their use in functional genomics. Funct Integr Genomic 2:60-69

Kynast RG, Okagaki RJ, Galatowitsch MW, Granath SR, et al (2004) Dissecting the maize genome by using chromosome addition and radiation hybrid lines. Proc Natl Acad Sci USA 101:9921-9926

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al (2001) Initial sequencing and analysis of the human genome. Nature 409:860-921

Le Signor C, Savois V, Aubert G, Verdier J, Nicolas M, et al (2009) Optimizing TILLING populations for reverse genetics in Medicago truncatula. Plant Biotech J 7:430-441

Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, et al (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438:803-819

Lloyd J, Meinke D (2012) A comprehensive dataset of genes with a loss-of-function mutant phenotype in Arabidopsis. Plant Phys 158(3):1115-1129

Luo MC, Deal KR, Akhunov ED, Akhunova AR, Anderson OD, et al (2009) Genome comparisons reveal a dominant mechanism of chromosome number reduction in grasses and accelerated genome evolution in Triticeae. Proc Natl Acad Sci USA 106:15780-15785

Maluszynski M, Nichterlein K, van Zanten L, Ahloowalia BS (2000) Officially released mutant varieties - the FAO/IAEA Database. Mut Breed Rev 12:1-84

Matsukura C, Aoki K, Fukuda N, Mizoguchi T, Asamizu E, Saito T, et al (2007) Comprehensive resources for tomato functional genomics based on the miniature model tomato micro-tom. Curr Genomics 9:436-443

Metzker ML (2010) Sequencing technologies - the next generation. Nat Rev Genet 11:31-46

Mézard C (2006) Meiotic recombination hotspots in plants. Biochemical Society Transactions 34:p4

Michalak MK, Ghavami F, Lazo GR, Gu YQ, Kianian SF (2009) Evolutionary relationship of nuclear genes encoding mitochondrial proteins across four grass species and Arabidopsis Thaliana. Maydica 54:471-483

Murphy WJ, Page JE, Smith C, Desrosiers RC, O'Brien SJ (2001) A radiation hybrid mapping panel for the rhesus macaque. Journal of Heredity 92:516-519

Murphy WJ, Larkin DM, Everts-van der Wind A, Bourque G, Tesler G, et al (2005) Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. Science 309:613-617

Novaes E, Drost DR, Farmerie WG, Pappas GJ Jr, Grattapaglia D, et al (2008) High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome. BMC Genomics 9:312

O'Brien SJ, Menotti-Raymond M, Murphy WJ, Nash WG, Wienberg J, et al (1999) The promise of comparative genomics in mammals. Science 286:458-481

Olivier M, Aggarwal A, Allen J, Almendras AA, Bajorek ES, et al (2011) A high-resolution radiation hybrid map of the human genome draft sequence. Science 291:1298-1302

Page DR, Grossniklaus U (2002) The art and design of genetic screens: Arabidopsis thaliana. Nature Reviews Genetics 3:124-136

Paux E, Sourdille P, Salse J, Saintenac C, Choulet F, et al (2008) A physical map of the 1-gigabase bread wheat chromosome 3B. Science 322:101-104

Peters JL, Crude F, Gerats T (2003) Forward genetics and map-based cloning approaches. Trends Plant Sci. 8:484-491

Riera-Lizarazu O, Vales MI, Ananiev EV, Rines HW, Philips RL (2000) Production and characterization of maize chromosome 9 radiation hybrids derived from an oat-maize addition line. Genetics 156:327-339

Riera-Lizarazu O, Leonard JM, Tiwari VK, Kianian SF (2010) A Method to Produce Radiation Hybrids for the D-Genome Chromosomes of Wheat (Triticum aestivum L.). Cytogenet Genome Res 129:234-240

Roberts MA, Reader SM, Dalgliesh C, Miller TE, Foote TN, et al (1999) Induction and characterization of Ph1 wheat mutants. Genetics 153:1909-1918

Saintenac C, Falque M, Martin OC, Paux E, Feuillet C, Sourdille P (2009) Detailed recombination studies along chromosome 3B provide new insights on crossover distribution in wheat (Triticum aestivum L.). Genetics 181:393-403

Salse J, Bolot S, Throude M, Jouffe V, Piegu B, et al (2008) Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. Plant Cell 20:11-24

Salvi S, Tuberosa R (2005) Top clone or not to clone plant QTLs: present and future challenges. Trends in Plant Sci 10(6):297-304

Schnurbusch T, Collins NC, Eastwood RF, Sutton T, Jefferies SP, Langridge P (2007) Fine mapping and targeted SNP survey using rice-wheat gene colinearity in the region of the Bo1 boron toxicity tolerance locus of bread wheat. Theor Appl Genet 115:451-461

Sengupta S, Harris CC (2005) p53: traffic cop at the crossroads of DNA repair and recombination. Nature Reviews Molecular Cell Biology 6:44-55

Shebeski LH, Lawrence T (1954) The production of beneficial mutations in barley by irradiation. Canad Jour Agri Sci 34:1-9

Shikazono N, Tanaka A, Watanabe H, Tano S (2001) Rearrangements of the DNA in carbon ion-induced mutants of Arabidopsis thaliana. Genetics 157:379-387

Shirley BM, Hanley S, Goodman HM (1992) Effects of ionizing radiation on a plant genome: analysis of two Arabidopsis transparent testa mutations. Plant Cell 4:333-347

Shu QY (2009) Turning plant mutation breeding into a new era: molecular mutation breeding. In: Induced Plant Mutations in the Genomics Era. FAO, Rome, pp 425-427

Simmons KJ, Fellers JP, Trick HN, Zhang Z, Tai Y-S, et al. (2006) Molecular characterization of the major wheat domestication gene Q. Genetics 172:547-555

Song XQ, Xia GM, Chen HM (2000) Chromosomal variation in long-term cultures of several related plants in Triticinae. Acta Phytophysiol Sin 26:33-38

Stewart EA, McKusick KB, Aggarwal A, Bajorek E, Brady S, et al (1997) An STS-based radiation hybrid map of the human genome. Genome Res 7:422-433

Van Orsouw NJ, Hogers RC, Janssen A, Yalcin F, Snoeijers S, et al (2007) Complexity reduction of polymorphic sequences (CRoPS): a novel approach for large-scale polymorphism discovery in complex genomes. PLoS ONE 2:e1172

Vogel JP, Garvin DF, Mockler TC, Schmutz J, Rokhsar D, et al (2010) Genome sequencing and analysis of the model grass *Brachypodium* distachyon. Nature 463:763-768

Wade CM, Giulotto E, Sigurdsson S, Zoli M, Gnerre S, et al (2009) Genome sequence, comparative analysis, and population genetics of the domestic horse. Science 326:865-867

Wardrop J, Snape J, Powell W, Machray G (2002) Constructing plant radiation hybrid panels. Plant J 31:223-228

Wardrop J, Fuller J, Powell W, Machray GC (2004) Exploiting plant somatic radiation hybrids for physical mapping of expressed sequence tags. Theor Appl Genet 108:343-348

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520-562

Weikard R, Goldammer T, Laurent P, Womack JE, Kuehn C (2006) A gene-based high-resolution comparative radiation hybrid map as a framework for genome sequence assembly of a bovine chromosome 6 region associated with QTL for growth, body composition, and milk performance traits. BMC Genomics 7:53

Weterings E, van Gent DC (2004) The mechanism of non-homologous end-joining: a synopsis of synapsis. DNA Repair 3:1425-1435

Wi GS, Chung BY, Kim JS, Kim JH, Baek MH, et al (2007) Effects of gamma irradiation on morphological changes and biological responses in plants. Micron 38:553-564

Wu CH, Nomura K, Goldammer T, Hadfield T, Dalrymple BP, et al (2008) A high-resolution comparative radiation hybrid map of ovine chromosomal regions that are homologous to human chromosome 6 (HSA6). Animal Genetics 39:459-467

Zhou C, Xia GM, Zhi DY, Chen Y (2006) Genetic characterization of asymmetric somatic hybrids between Bupleurum scorzonerifolium Willd. and Triticum aestivum L.: potential application to the study of the wheat genome. Planta 223:414-424

# 2. RADIATION HYBRIDS OF CHROMOSOME 3B REVEAL A DEPENDENT RESPONSE OF THE DNA-REPAIR MECHANISM TO THE STATE OF CHROMATIN[1]

Radiation hybrid (RH) is a strategy to map genomes widely employed in animal species. RH maps have been proposed to provide *i*) higher and *ii*) more uniform resolution than genetic maps, and *iii*) to be independent of the distribution patterns observed for meiotic recombination. An *in vivo* RH panel was generated for mapping chromosome 3B of wheat (*Triticum aestivum* L.), in a preliminary attempt to provide a complete scaffold for this ~1 Gb chromosome. A RH map was generated with 541 markers anchored to chromosome 3B BAC contigs. Detailed comparisons with a genetic map of similar quality confirmed that *i*) the resolution of the RH map was 10.5 fold higher and *ii*) six fold more uniform. Also, it was observed a surprising interaction (r= 0.879 at *p*= 0.01) between the DNA repair mechanism and the distribution of crossing-over events. This finding could be explained by admitting the possibility that the DNA repair mechanism in somatic cells is affected by the chromatin state in a way similar to the effect that chromatin state has on recombination frequencies in gametic cells. The RH data presented here support for the first time *in vivo* the hypothesis of non-casual interaction between recombination hot-spots and DNA repair. Further, two major hypotheses are presented on how chromatin compactness affects the DNA repair mechanism. Since the initial RH application 37 years ago, this study shows for the first time that the *iii*) third hypothesis of RH mapping might not be entirely correct.

[1] The material in this chapter was co-authored by Filippo Bassi and Ajay Kumar. Ajay Kumar conducted the initial screening with 30 markers. Filippo Bassi conducted all other genotyping, analysis, and he is the primary developer of the conclusions that are advanced here. Filippo Bassi drafted and revised all versions of this chapter.

## 2.1. Introduction

Genetic mapping has been the foundation of molecular analysis in plants and animals for nearly a century, since the publication of the first map by Sturtevant in 1913 (Stutervant 1913). This widely successful approach relies on recombination to separate and order marker loci. In many plant species, including wheat (*Triticum aestivum* L.), recombination events are not evenly distributed along the length of the chromosomes (Gill et al. 1993; Lukaszewski et al. 1993; Hohmann et al. 1994; Erayman et al. 2004; Mézard et al. 2006; Saintenac et al. 2009). Recombination hot-spots, sites with high recombination rates, are interspersed with recombination cold-spots. In addition, in species with large genomes, such as wheat, barley, or maize, recombination frequency tends to decrease with increased proximity to the centromere (Erayman et al. 2004; Saintenac et al. 2009), being close to zero at the centromere and its surroundings. It is estimated that about one-fourth to one-third of the ~17 Gb wheat genome (Dolezel al. 2009) accounts for less than 1% of the total recombination (Lukaszewski et al. 1993; Akhunov et al. 2003; Erayman et al. 2004). The results of early studies suggested that these recombination poor regions were nothing more than "junk" DNA (Ohno 1972; for review see Makalowski 2003), but more recent studies indicated that over 30% of wheat genes exist within this space (Erayman et al. 2004). Limited recombination makes these genes virtually inaccessible to genetic mapping. Similarly, the use of genetic maps as scaffold to orient physical maps (such as ordered bacterial artificial chromosome (BAC) contigs) provides only limited information for these recombination poor regions.

Radiation hybrid (RH) mapping is a method that was originally proposed as an alternative to the use of recombination for mapping marker loci (Goss and Harris 1975; Cox et al. 1990). In RH mapping, high dosages of radiation are used to generate random double strand breaks (DSBs) across the genome. The DSBs are then recognized and fixed by one of two main repair mechanisms: homology-directed repair (HR) or non-homologus end-joining (NHEJ), also known as illegitimate recombination (Britt 1999; Puchta et al. 2005; Bleuyard et al. 2006; Bernstein and Rothstein 2009). Both of these mechanisms are highly conserved in eukaryotes. NHEJ, even though it is a mechanism prone to error, is considered the prevailing choice of somatic DSB

repair in higher eukaryotes (Bleuyard et al. 2006). In the last two decades, a number of proteins involved in the NHEJ mechanism of repair have been identified (Lliakis et al. 2004; Puchta et al. 2005; Bleuyard et al. 2006; Xu and Price 2011). Also, a model has been proposed to explain their interaction and functionality. In brief, the presence of DNA broken ends is sensed through the ATM (Ataxia Telangiectasia Mutated) signaling pathway. The Ku protein complex is then recruited to the damaged site with the function of protecting the DSBs from further degradation. The Ku complex anchors at the break site where it provides a docking point for the DNA phosphokinases. The precise activities of the DNA phosphokinases are not completely clear, but it is apparent that they are involved with the direct or indirect creation of protein bridges to pull the two broken ends together. Finally, the two adjacent broken ends are re-joined by a DNA ligase, the repair is complete, and the repair complex disassembles (Lliakis et al. 2004; Bleuyard et al. 2006; Xu and Price 2011). When the broken ends are re-joined, the nucleotides located within the adjacent DSBs will be lost. The loss mainly involves a small number of nucleotides, but deletions of larger size are also not uncommon (Britt 1999; Bleuyard et al. 2006; Bernstein and Rothstein 2009).

Radiation hybrid mapping exploits the formation of DNA deletions to generate a binary polymorphism (1-retention vs 0-deletion) which is then used to determine the markers order based on their simultaneous co-deletion or co-retention (Goss and Harris 1975). While the molecular components of the NHEJ mechanism of repair have been partially or entirely identified, its precise activity in live organisms requires further investigation. In this regard, viable RH plant populations might represent a novel and useful tool.

During the last two decades, RH mapping has played an important role in mapping and genome assembly of a number of organisms, including humans and other animals (Oliver et al. 2001; Karere et al. 2008; Faraut 2009; Lander et al. 2001; Waterson et al. 2002; Gibbs et al. 2004; Lindblad-Toh et al. 2005; Elsik et al. 2009; Wade et al. 2009; Lin et al. 2010). However, only a few preliminary studies have been reported in crop plants, such as maize (*Zea mays* L.) (Riera-Lizarazu et al. 2000; Kynast et al. 2004), barley (*Hordeum vulgare* L.) (Wardrop et al. 2002), cotton (*Gossypium hirsutum* L.) (Gao et al. 2004) and wheat (Hossain et al. 2004; Kalavacharla et al. 2006; Paux et al. 2008; Michalak et al. 2009).

In an effort to sequence the bread wheat chromosome 3B, a partial physical map covering 82% of the estimated 993 Mb size of the chromosome was recently developed (Paux et al. 2008). In that study, RH mapping was tested as a mean to provide good quality scaffolding for physical mapping. Here, I present an extension of that work, with the development of a high density RH map of chromosome 3B to align bacterial artificial chromosome (BAC) contigs. This is an unprecedented opportunity in plants to examine the physical distribution of radiation mediated deletions at a very refined level. Furthermore, this map was employed to assess what are commonly considered the three major advantages of RH mapping: *i*) high map resolution, *ii*) precise conversion of centi Ray (cR, map unit of RH) distances to physical distances, and *iii*) independence from the patterns observed for meiotic recombination events. Surprisingly, the data presented here support only the first two hypotheses, while a strong correlation was identified between the action of the DNA breakage/repair mechanism and the frequency of meiotic recombination events. This result suggests that the state of chromatin may influence the DNA break formation and repair mechanism in a manner similar to the way it influences recombination events.

## 2.2. Methods

### 2.2.1. RH mapping population

The 3B-RH panel was generated as described in Paux et al. (2008) by crossing the durum wheat *Triticum turgidum* ssp*. durum* L. var. 'Langdon' (LDN) after irradiation at 350 Gy of gamma ray to the aneuploid var. 'Langdon 3D(3B)' (LDN 3D(3B)) (Figure 2.1). $RH_1$ seeds were planted under greenhouse controlled conditions and DNA was extracted from leaves of four weeks old plants as described earlier (Hossain et al. 2004). Non-irradiated double monosomic (13" + 3B'+ 3D') $F_1$ lines were also generated and employed as experimental controls. All DNA samples were equilibrated to concentrations of 50 ng per µl. A total of 184 $RH_1$ lines were initially employed in this study, but only 92 lines were selected for full genotyping.

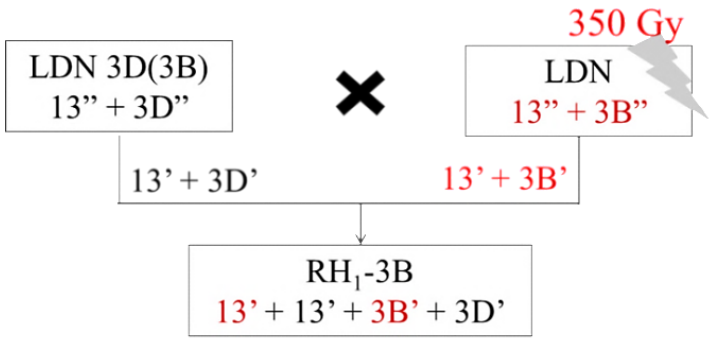


Figure 2.1. Crossing scheme to generate 3B-$RH_1$ lines.

2.2.2. Molecular analysis

Genotyping was conducted with three classes of markers: cfp are PCR insertion site based polymorphism markers (ISBP; Paux et al. 2006; Paux et al. 2010); barc, gwm, and gpw are PCR based single sequence repeats (SSR; http://wheat.pw.usda.gov/GG2/index.shtml); wPt and tPt are diversity array technology markers (DArT; Canberra, AU) (Wenzl et al. 2010). To guarantee chromosome 3B specificity, all markers were tested for positive amplification (ISBP and SSR) or hybridization (DArT) of the double monosomic $F_1$ control line, and no-amplification / hybridization of LDN 3D(3B). In order to distinguish PCR failure from deletion-detection, all markers amplifying a single 3B specific fragment were multiplexed with the control marker DEASY (Duplexing EASY) amplifying 164 bp of a chloroplastic ATP syntase alpha subunit [GebBank:M16842]. All PCR protocols have been described previously (Paux et al. 2008; Francki et al. 2009; Paux et al. 2006; Paux et al. 2010). Ninety-two samples plus four experimental controls were genotyped in duplicate using the 3B specific DArT array following the protocol described in Wenzl et al. (2010). The deletion frequency is calculated as the number of loci deleted divided by the total number of loci genotyped, while the retention frequency is one minus the deletion frequency. The 3B genetic map resolution and CO frequency was calculated in Saintenac et al. (2009). The Carthagene mapping software v1.2.2 (Givry et al. 2005) was modified and used to generate a RH map of the entire 3B chromosome. Details on the superimposed modifications are available in Appendix A.

2.2.3. Statistical analysis

All correlation analyses were performed using the SAS 9.3 environment (SAS Institute, Cary, NC) and the correlation significance was determined on the basis of the Pearson product-moment correlation coefficient for a two-tail test with N-2 degrees of freedom, where N is the number of deletion bins considered (Rodgers and Nicewander 1988).

2.3. Results

2.3.1. A dense and precise 3B-RH map

In plants, *in vivo* RH panels can be generated by simple gamma irradiation of seeds, followed by artificial cross-pollination of adult mutant lines. A RH panel for chromosome 3B was generated by gamma irradiating at 350 Gray (Gy) seeds of a normal durum line containing chromosome 3B (AABB, 2n=4x=28: 13" + 3B") and crossing it to an aneuploid line that lacks this chromosome (13" + 3D"). A total of 184 $RH_1$ lines were developed, and 92 $RH_1$ lines were selected on the basis of genotyping data from 84 Insertion Site Based Polymorphism (ISBP, labeled cfp) marker loci (Paux et al. 2008). The selected sub-population is composed of 72 lines with deletions of various sizes (retention frequency 0.500-0.999), nine lacked the entire chromosome (retention frequency < 0.020), and 13 retained the whole chromosome (retention frequency of 1.000) (Figure 2.2). The average retention frequency of this selected population was 0.89. In our experience this small sample is a good representation of any larger RH populations (Hossain et al. 2004; Kalavacharla et al. 2006; Paux et al. 2008).

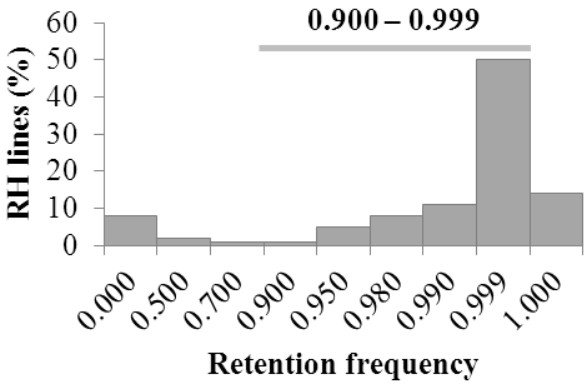


Figure 2.2. Distribution of marker retention frequencies. The frequencies were calculated from a population of 92 radiation hybrid lines specifically created for chromosome 3B of wheat and are based on 541 marker analysis.

Locations of radiation-induced breaks on chromosome 3B were determined employing 541 markers, which include 96 ISBP, 19 SSRs and 426 DArT markers. For 128 of these markers (115 PCR-based and 13 DArTs), the chromosomal location (Sourdille et al. 2004; Francki et al. 2009) and the BAC contig of appurtenance were known (available at http://urgi.versailles.inra.fr/cgibin/gbrowse/wheat_ FPC_pub/). These markers were defined as 'anchor markers'. The marker information was used to construct an iterative framework map of LOD 10 using a modified version of the Carthagene software package (Givry et al. 2005; for details on the mapping algorithm see Appendix 2.1). The final map (3B-RH) spans 1871.9 cR, as defined by 202 unique loci (Figure 2.3). Assuming even distribution of markers along the chromosome, the overall marker density is one marker every 3.5 cR, or approximately 1.9 Mb based on the 993 Mb size of chromosome 3B (Dolezel et al. 2009). Quality of the RH map was tested by comparison with two previously published genetic maps (Paux et al. 2008; Wenzl et al. 2010). This examination revealed better conservation of marker loci order between the 3B-RH iterative map and Paux et al. (2008) map than between the two genetic maps (Figure A.1 in Appendix A). Further confirmation of the strength of this approach was provided by the marker wPt-0223 mapped on the 3B-RH map at position 414.8 cR, in between anchor markers cfp5042 and cfp5031, which are at positions 313.1 cR and 419.2 cR, respectively. Marker wPt-0223 was mapped into contig 954, which has been entirely sequenced (Choulet et al. 2010). The two anchor markers cfp5042 and cfp5031 are located at 1,103,689 bp and 1,640,531 bp on this contig, respectively. The sequence of wPt-0223 was used to map this locus *in silico* at position 2,170,833 bp (i.e. 0.53 Mb proximal of marker cfp5031), just outside the interval predicted in 3B-RH. Assuming 100% to be the error of placing a locus an entire chromosome length away (993 Mb) from its correct physical position, the error for the 3B-RH map was calculated (0.53 Mb / 993 Mb) to be as low as 0.05%. This is a relatively small error considering that only 92 lines were used for the analysis.

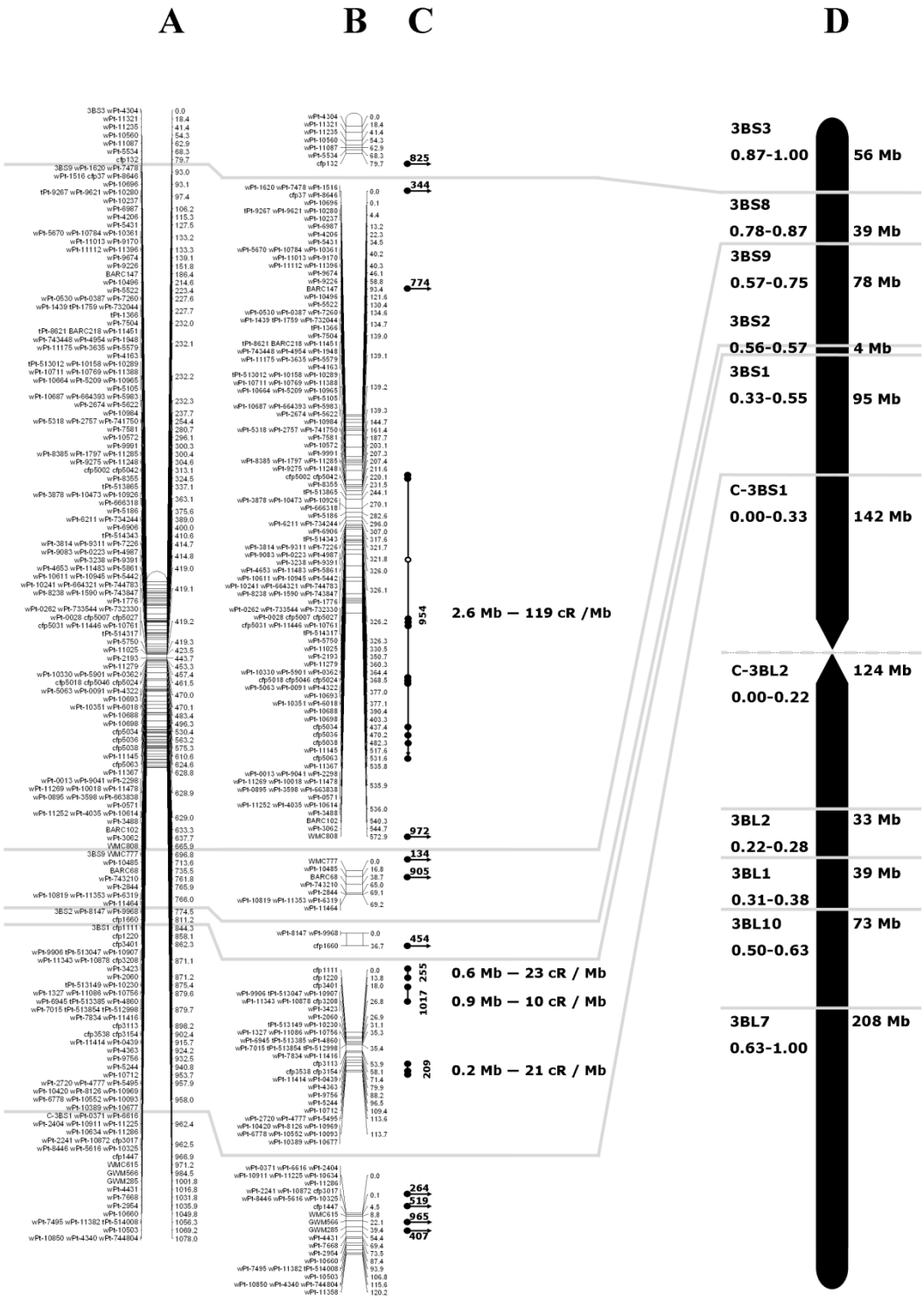


Figure 2.3. Radiation hybrid map of chromosome 3B of wheat. (A) RH map of the short and long (next page) arm of chromosome 3B; map unit is centi-ray (cR). (B) RH map of the short and long (next page) arm of chromosome 3B divided into deletion bins. (C) BAC contig distribution as based on anchored markers; contigs for which a breakage was identified are indicated by vertical arrows pointing in the resolved map orientation and cR/Mb resolution within the contig is shown (continued in next page).

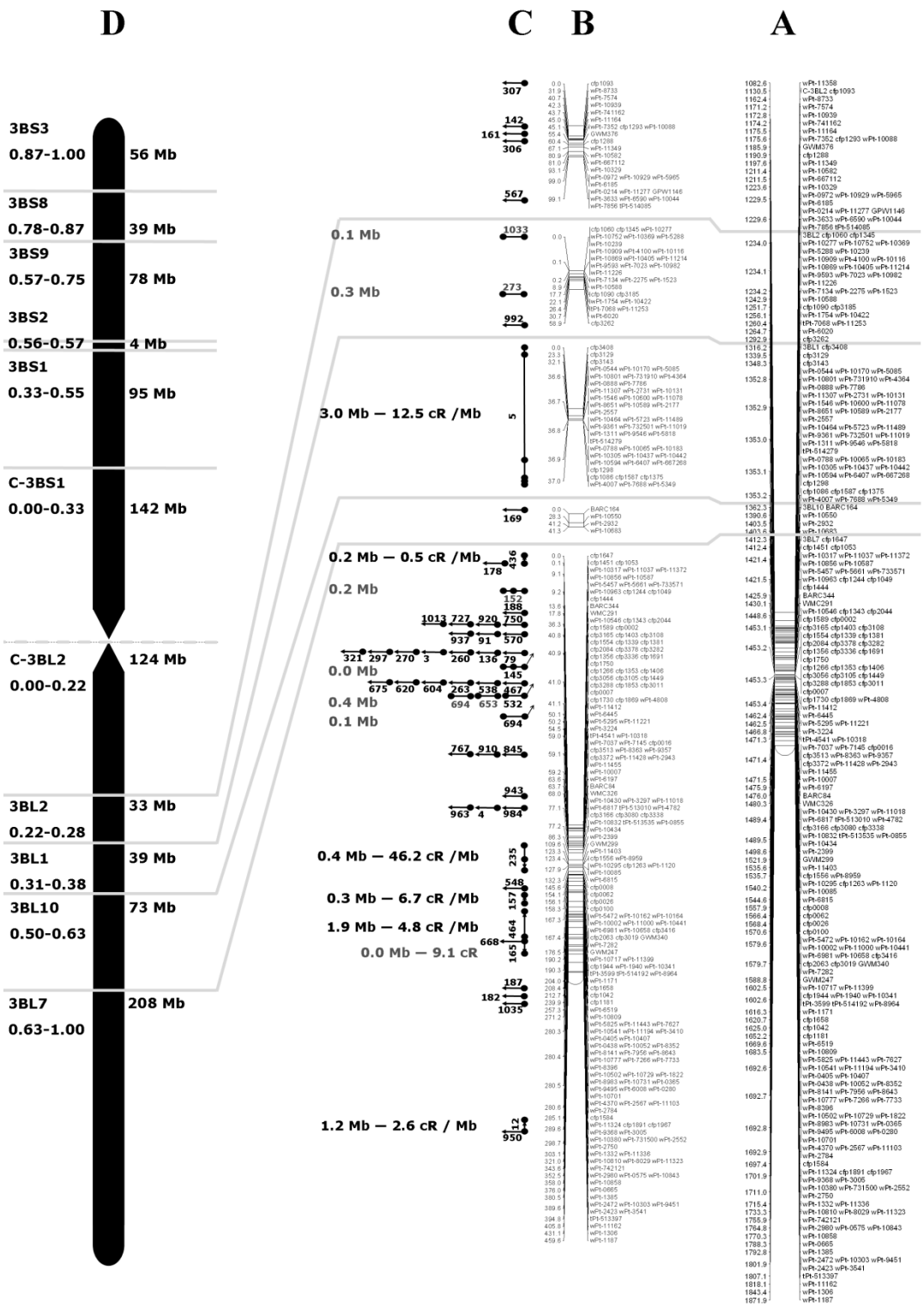


Figure 2.3b. Radiation hybrid map of chromosome 3B of wheat. (continued from previous page) Horizontal lines connecting multiple dots indicate contigs that had no breakage event and the size of the contig is reported; horizontal arrows indicate contig for which only a single marker was mapped. (D) Deletion bins map of wheat chromosome 3B; I had no anchor markers available for bin 3BL8-0.28-0.31 and 3BL9-0.38-0.50 so these were excluded from the reported bin map.

2.3.2. RH map resolution

Map resolution is defined as the minimum physical distance between two marker loci needed to resolve their location on a map. It is calculated as the ratio between physical and map distances (i.e. Mb divided by cR or cM). Smaller values indicate better resolution. Radiation hybrid maps are expected to provide higher map resolutions than genetic maps, and also to be better estimates of the actual physical distance between mapped loci (Kalavacharla et al. 2009). To verify these two hypotheses, the resolution was estimated at three physical levels: *i*) the whole chromosome (993 Mb), *ii*) chromosome sub-portions, defined as cytogenetic deletion bins (4 Mb to 208 Mb; Table 2.1), and *iii*) within BAC contigs (0.1 Mb to 3.0 Mb; Table 2.2). Cytogenetic deletion bins (hereafter referred to as 'bins') are physical segments of wheat chromosomes identified by cytogenetic lines carrying terminal deletions of specific chromosome fragments (Endo and Gill 1996). The physical sizes of these bins have been estimated through a combination of molecular and cytogenetic studies. These bins are commonly identified by the chromosome number (3B), the chromosome arm (S or L), and the percentage of the specific arm that is deleted (i.e. 0.63-1.00, from 63% of the arm till the end, 100%); for simplicity, I report the bin full name in Table 2.1, and I only use an abbreviated identifier in the text. BAC contigs were generated by fingerprinting large libraries of clones, and their physical size is an estimate (Paux et al. 2008; Table 2.2). The resolution was calculated for both maps, genetic and RH, for the first two levels, but BAC contig analysis was reserved only for 3B-RH. In the 3B-RH map, the bin locations were known for 115 anchor markers, and this information was used to extrapolate the locations of the non-anchored markers. The largest portion of the 3B-RH map was contained within two bins (3BS8 and 3BL7) accounting for 55% of the total map size, but only for 25% of the physical size of chromosome 3B (Table 2.1 and Figure 2.3). Assuming complete coverage of the chromosome, and considering the 3B physical size of 993 Mb (Paux et al. 2008), the 3B-RH map resolution averaged 0.53 Mb $cR^{-1}$, ranging from 0.1 Mb $cR^{-1}$ for bin 3BS8 and 3BS2 to 1.8 Mb $cR^{-1}$ for 3BL10 (Table 2.1). The two centromeric bins exhibited very similar resolution s of 1.2 Mb $cR^{-1}$ and 1.3 Mb $cR^{-1}$ in C-3BS1 and C-3BL2, respectively (Table 2.1).

Table 2.1. A comparison of radiation hybrid and genetic maps of chromosome 3B of wheat

| Deletion bins | 3BS3-0.87-1.00 | 3BS8-0.78-0.87 | 3BS9-0.57-0.75 | 3BS2-0.56-0.57 | 3BS1-0.33-0.55 | C-3BS1-0.33 | C-3BL2-0.22 | 3BL2-0.22-0.28 | 3BL1-0.31-0.50 | 3BL10-0.50-0.63 | 3BL7-0.63-1.00 | Chr[†] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| No. of marker | 7.0 | 155.0 | 9.0 | 3.0 | 44.0 | 28.0 | 27.0 | 29.0 | 47.0 | 4.0 | 187.0 | 540.0 |
| Bin size (Mb) | 56.0 | 39.0 | 78.0 | 4.0 | 95.0 | 142.0 | 124.0 | 33.0 | 39.0 | 73.0 | 208.0 | 992.0 |
| Bin size (%) | 5.7 | 3.9 | 7.9 | 0.4 | 9.6 | 14.3 | 12.5 | 3.3 | 3.9 | 7.4 | 21.0 | 89.8 |
| Bin saturation (Mb Marker $no^{-1}$) | 8.0 | 0.3 | 8.7 | 1.3 | 2.2 | 5.1 | 4.6 | 1.1 | 0.8 | 18.3 | 1.1 | 1.8 |
| Map size (cR) | 79.7 | 572.9 | 69.2 | 36.7 | 113.7 | 115.6 | 99.1 | 58.9 | 37.0 | 41.3 | 459.6 | 1871.9 |
| BR freq (cR ½ $Mb^{-1}$) [§] | 0.7 | 7.4 | 0.5 | 4.6 | 0.6 | 0.4 | 0.4 | 0.9 | 0.5 | 0.3 | 1.1 | 1.0 |
| Resolution (Mb $cR^{-1}$) [¶] | 0.7 | 0.1 | 1.1 | 0.1 | 0.8 | 1.2 | 1.3 | 0.6 | 1.1 | 1.8 | 0.5 | 0.5 |
| Map size (cM) [‡] | 5.2 | 33.1 | 13.4 | . | 1.6 | 0.9 | 1.6 | 1.6 | 6.2 | 3.5 | 104.2 | 179.0 |
| CO freq. (cM $Mb^{-1}$) [§‡] | 0.1 | 0.8 | 0.2 | . | 0.0 | 0.0 | 0.0 | 0.0 | 0.2 | 0.0 | 0.5 | 0.2 |
| Resolution (Mb $cM^{-1}$) [¶] | 10.8 | 1.2 | 5.8 | . | 61.3 | 167.1 | 80.0 | 21.3 | 6.3 | 21.2 | 2.0 | 5.5 |
| Map size ratio (cR $cM^{-1}$) | 15.3 | 17.3 | 5.2 | . | 73.4 | 136.0 | 63.9 | 38.0 | 6.0 | 12.0 | 4.4 | 10.45 |

[†] Whole chromosome values are not averages for each category but re-calculated for the overall values. [‡] Values in the row are as of Saintenac et al. 2009. [§] Frequency of DNA breakage and repair per Mb and frequency of a crossing-over event per Mb. [¶] Resolution is the potential to uniquely resolve two markers at a distance equal or larger than this value. Chr: chromosome; BR freq: breakage and repair frequency; CO freq: crossing over frequency; cR: centiRays; Mb: megabase.

Table 2.2. BAC contigs of known physical size mapped on radiation hybrid map of chr. 3B. The topmost nine contigs could be oriented, the remaining eight could not be oriented; the data for 58 BAc Contig having just one marker are not reported.

| Contig ID | Size (Mb) | Size (cR) | Markers (count) | Resolution (Mb $cR^{-1}$) | Deletion Bin |
|---|---|---|---|---|---|
| Oriented contigs | | | | | |
| Ctg0954 | 2.6 | 311.5 | 12 | 0.01 | 3BS8-0.78-0.87 |
| Ctg0209 | 0.2 | 4.2 | 3 | 0.04 | 3BS1-0.33-0.55 |
| Ctg0255 | 0.6 | 13.8 | 2 | 0.04 | 3BS1-0.33-0.55 |
| Ctg1017 | 0.9 | 8.8 | 2 | 0.10 | 3BS1-0.33-0.55 |
| Ctg0005 | 3.0 | 37.0 | 6 | 0.08 | 3BL1-0.31-0.38 |
| Ctg0235 | 0.4 | 18.3 | 3 | 0.02 | 3BL7-0.63-1.00 |
| Ctg0464 | 1.9 | 9.1 | 3 | 0.20 | 3BL7-0.63-1.00 |
| Ctg0012 | 1.2 | 4.5 | 2 | 0.26 | 3BL7-0.63-1.00 |
| Ctg0157 | 0.3 | 2.0 | 2 | 0.14 | 3BL7-0.63-1.00 |
| Sub-total | 10.9 | 418.3 | 35 | †0.09 | |
| Not oriented contigs | | | | | |
| Ctg0273 | 0.3 | 0.0 | 2 | | 3BL2-0.22-0.28 |
| Ctg1033 | 0.1 | 0.0 | 2 | | 3BL2-0.22-0.28 |
| Ctg0145 | 0.0 | 0.0 | 3 | | 3BL7-0.63-1.00 |
| Ctg0152 | 0.2 | 0.0 | 3 | | 3BL7-0.63-1.00 |
| Ctg0436 | 0.2 | 0.0 | 2 | | 3BL7-0.63-1.00 |
| Ctg0532 | 0.2 | 0.0 | 2 | | 3BL7-0.63-1.00 |
| Ctg0653 | 0.4 | 0.0 | 2 | | 3BL7-0.63-1.00 |
| Ctg0694 | 0.1 | 0.0 | 2 | | 3BL7-0.63-1.00 |
| Total | 12.5 | | 52 | | |

† Average

Chromosome-wise, the RH map resolution deviated less than five-fold from the calculated average. In comparison, the resolution calculated by Saintenac et al. (2009) for their genetic map of chromosome 3B was 5.5 Mb $cM^{-1}$, 10.5 fold lower than the resolution of the 3B-RH map, and deviated along the chromosome up to 30 fold (1.2 and 167.1 Mb $cM^{-1}$) from the average (Table 2.1).

The resolution of the 3B-RH map reached its minimum in bin 3BL10 at 1.8 Mb $cR^{-1}$. Based on this value, our small population of 92 $RH_1$ lines should have the mapping potential to uniquely order any BAC contig with size >1.8 Mb. To test this hypothesis, I analyzed markers anchored to 72 BAC contigs (Table 2.2 and Figure 2.3), 15 from the short arm and 57 from the long arm, mainly belonging to the bin 3BL7 (45 contigs). A contig can be assigned to a specific chromosomal position with the help of a single marker; however, two or more mapped markers are required to orient a contig. For 17 of 72 contigs two or

more markers were available, and nine of these could be uniquely oriented (Table 2.2). The physical distance between the markers used to orient the contigs ranged from 3.0 Mb for contig 5, to 0.0 Mb (meaning that the markers are at a distance too small to be dissected by BAC fingerprinting but are not necessarily at the same locus, see Choulet et al. 2010) for contig 145. As expected, the eight contigs that could not be oriented contained markers spaced by distances <1.8 Mb, ranging from 0.4 Mb to 0.0 Mb. I also calculated map resolution within contigs, which averaged 0.09 Mb $cR^{-1}$ for the whole chromosome, ranging from 0.01 Mb $cR^{-1}$ for contig 954, to 0.26 Mb $cR^{-1}$ for contig 12 (Table 2.2).

2.3.3. Sizes of gamma ray induced chromosome deletions in the 3B-RH panel

A single radiation-induced deletion was defined by a set of flanking markers detecting an uninterrupted deletion smaller than a chromosome arm. Eight of the 92 lines had continuous deletions larger than a chromosome arm, retaining only a small portion of chromosome 3B, probably through translocation to a different chromosome. One line apparently lost the entire chromosome. The loss of an entire chromosome can be due to large rearrangements or improper chromosome sorting at meiosis, possibly caused by excessive deletions or removal of critical segments, such as the centromere. Thirteen putative RH lines did not have any apparent deletions (Figure 2.2). These 22 lines were not considered when calculating the deletion size. Because a fully sequenced genome is not available, the deletion size was measured in map units (cR) and then converted into Mb, based on the conversion ratio calculated for each bin (i.e. resolution, Table 2.1).

The largest deletion identified spanned 120.2 cR, representing the loss of the entire bin C-3BS1 equal to 144 Mb, while the smallest deletion encompassed 13.1 cR of bin 3BS8, accounting for 1.3 Mb in size. Deletions of smaller size are also possible (Britt 1999) but their identification would require a more targeted approach than the one employed here. The average deletion size across chromosome 3B was 26.4 Mb, ranging from 5.9 Mb of bin 3BS8 to 71.4 Mb of bin C-3BL2 (Figure 2.4). On average, each of the 70 informative lines carried 2.8 deletions, ranging from 1 to 13 (Figure 2.3). There is an inverse correlation ($r=-0.64$, $p=0.05$) between the number of deletions and their relative size, with the centromeric

regions typically having fewer but larger deletions and the telomeric regions having smaller and more frequent deletions.

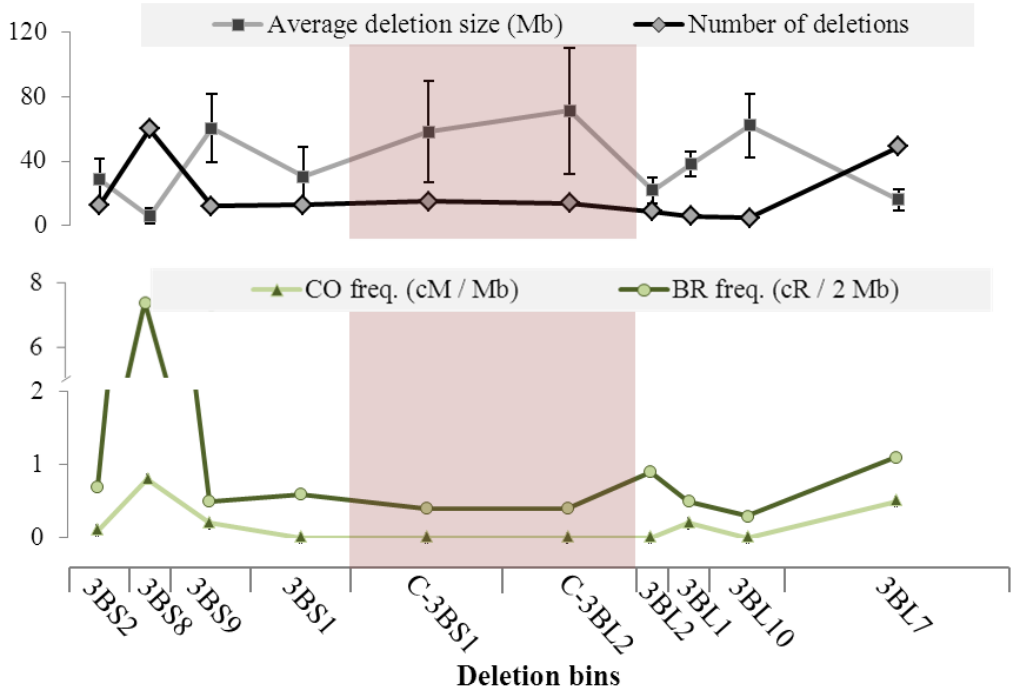


Figure 2.4. Comparison of breakage/repair and crossing over frequencies, and deletion sizes across chromosome 3B. Positive and negative correlations between the average deletion size, the number of deletions, the frequency of crossing-over (CO) (as of Saintenac et al. 2009) and breakage/repair (BR) frequency across the 3B chromosome. For the average deletion size the standard deviation across lines is represented as error bars; the other values plotted are absolute with no replicates. The visual parallelisms between the lines indicate good correlation between the observations.

Also, I investigated the distribution of deletion frequency (defined as how frequently a given marker locus is lost in a population) across the RH-3B population, in an attempt to identify chromosomal regions that may be preferentially deleted or retained (Figure 2.5, and Figure A.3). The average deletion frequency was 13%, reaching a maximum of 15% in C-3BS1 and a minimum of 12% in 3BL1. Overall the deletion frequency was evenly distributed and no significant difference was observed among the bins. Apparently, the vicinity to chromosomal landmarks such as the centromere or telomeres does not influence the frequency at which a given locus is lost.

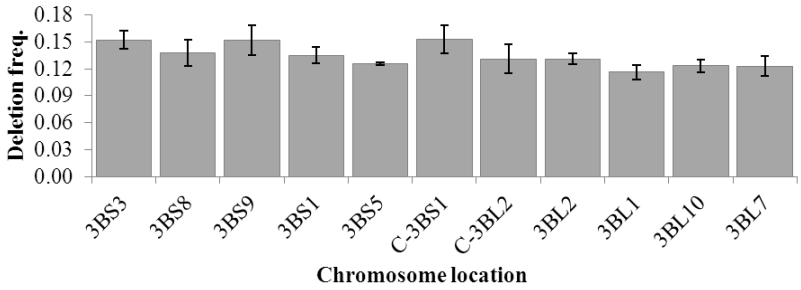


Figure 2.5. Distribution of deletion frequencies across chromosome 3B. Each deletion bin is indicated by its designation (i.e. 3BS3) and represents a portion of chromosome 3B. The deletion frequency for each bin was calculated as the average of the deletion frequencies of each marker mapped within that bin. The error bars indicate the standard deviation of the values across the markers.

2.3.4. Distribution of DNA break/repair events show correlation to the frequency of crossing-over

The cR unit measures the frequency of map co-localization of two marker loci. Hence, 1 cR is equal to 1% of of difference of the state, deleted vs. retained, between two adjacent loci. Similarly, one cM indicates 1% of recombination event between two loci. Deletion of one locus is likely the result of two DSBs, one distal and one proximal to the lost marker locus. Therefore, the number of deletions between two markers can be obtained by dividing the cR distance separating them by two. A DNA deletion in RH is the consequence of a radiation-mediated breakage, which is then repaired by the DNA-damage repair mechanism. Hence, when dividing half of the cR distance between two markers by the physical size that separates them, I are measuring the frequency at which radiation-mediated DSBs are formed and then repaired, most likely through NHEJ. I calculated this frequency (called break/repair (BR) frequency; Table 2.1), and I expressed it as the number of breakage/repair events (i.e. cR/2) per Mb. For instance, a BR frequency of 1 cR/2 $Mb^{-1}$ would indicate that in a specific interval two breakage/repair events occurred on average every 1 Mb, while a value of 10 cR/2 $Mb^{-1}$ would suggest 20 BR events every Mb. Similarly, by dividing cM by the physical size of the interval, I are measuring the average physical distance that separated crossing-over (CO frequency; Saintenac et al. 2009).

Crossing-over frequencies in wheat are known to be unevenly distributed across the chromosome, and generally decrease with proximity to the centromere (Saintenac et al. 2009). The relative distance from the centromere does not explain entirely the distribution of CO events, still I

confirmed a positive correlation (r= 0.604, p=0.05) between the frequency of CO and the relative distance from the centromere. The same analysis did not return a significant interaction between the BR frequency and the distance from centromere (r= 0.292), further confirming that the proximity to chromosomal landmarks by itself does not influence the frequency of DNA breakage and repair. However, bin 3BS8 exhibited the highest BR frequency (7.4 cR/2 $Mb^{-1}$, 7.7-fold higher than the chromosome average) and CO frequency (0.8 cM $Mb^{-1}$, 4-fold higher than average). Bin 3BL10 exhibited one of the lowest CO frequency (0.05 cM $Mb^{-1}$, 4-fold below average), and also the lowest BR frequency (0.3 cR/2 $Mb^{-1}$, 3.6-fold lower than the chromosome average). The similarity between these values prompted us to compute chromosome-wide correlation between the BR frequencies and the CO frequencies (Figure 2.4). I obtained a value of r = 0.879 (p=0.01) when considering the ten bins for which information on both frequencies was available. Bin 3BS2 was excluded from the analysis due to lack of CO frequency data. This significant correlation indicates that the CO and BR frequencies are not independent values (Figure 2.4). Furthermore, I calculated the total number and average size of radiation-mediated deletions in each deletion bin, confirming that both CO and BR frequencies are positively correlated (p=0.01) with the number of deletions, but only weakly inversely correlated with the average deletion size (p=0.06) (Figure 2.4). No significant correlation was observed between CO, BR frequencies and markers deletion frequencies. To avoid confusion, I would like to emphasize that BR frequency and deletion frequency do not measure the same biological effect. The BR frequency value estimates the activity of the DNA repair mechanism, while the deletion frequency measures just the number of times that a given locus or region is damaged by radiation.

## 2.4. Discussion

In this study, a population of 92 $RH_1$ lines was analyzed using a novel iterative framework mapping algorithm (Appendix A) to generate a dense RH map of wheat chromosome 3B. The map quality tests indicated a small map error (0.05%). Hence, I concluded that the method employed for mapping did not generate any perceivable bias, and downstream analyses should not be skewed by the anchor marker mapping approach.

2.4.1. Is RH mapping resolution higher and more uniform than the resolution of genetic mapping?

To answer this question, the 3B-RH map was compared to the high quality genetic map of chromosome 3B published by Saintenac et al. (2009). The 3B-RH panel provided a 10.5 fold higher overall resolution than the genetic map, reaching a maximum of 136-fold better resolution at the centromere, where recombination is most scarce. I believe that this is sufficient evidence to conclude that RH mapping indeed provides higher resolution than genetic mapping. The average resolution is often calculated for genetic maps, but the uneven distribution of recombination along a chromosome can result in up to 30-fold variation between telomeric and centromeric regions (1.2 to 167.1 Mb $cM^{-1}$; Table 2.1). For this reason, the resolution value calculated for genetic maps is not a reliable measure of the actual physical distance separating the mapped loci. To investigate if RH maps would provide a more uniform resolution across the chromosome, I compared the average resolution for the entire chromosome with the resolution for each chromosomal region. The average resolution, calculated as the total physical size of the chromosome divided by the total map length, was 0.53 Mb $cR^{-1}$ and I observed maximum of five-fold deviation from this value (0.1 and 1.8 Mb $cR^{-1}$) along the chromosome. Thus RH mapping resolution is six times more uniform than genetic mapping resolution. I also assessed resolution within BAC contigs. In this case, resolution fluctuated more dramatically, ranging from 0.26 Mb $cR^{-1}$ to 0.01 Mb $cR^{-1}$ a 53-fold increase from the chromosome-wide average, and averaged at 0.09 Mb $cR^{-1}$ (six-fold increase). This large fluctuation is probably the result of the non-random selection of the RH lines identified at the beginning of the study. In fact, the subset of 92 lines was specifically selected for their quality of map-resolving 84 markers of the 109 employed to measure the BAC contig resolution. It is then not surprising that they provide a much higher resolution exactly for those markers that were used for the initial selection.

Overall, the lowest resolution observed for any chromosome region in this study was 1.8 Mb $cR^{-1}$. This indicates the ability of our small RH population to unequivocally orient any BAC contig of size larger than 1.8 Mb. Only eight BAC contigs could not be uniquely oriented in this study, the largest of which was only 0.4 Mb in size. Such high resolution has been observed in many human and animal radiation hybrid maps before (Oliver et al. 2001; Lin et al. 2010; Faraut 2009), but among the plant studies reported to

date only a resolution of ~0.2 Mb calculated on the number of obligate breaks for chromosome 1D of wheat (Kalavacharla et al. 2006) is close to the high resolution I have presented here.

2.4.2. The chromatin state affects the DNA break/repair mechanism

RH studies rely solely on the random formation of deletions for mapping. Given the theoretical absence of molecular bias, the RH mapping approach has been thought to provide a true physical representation of chromosomes (Cox et al. 1990; Kalavacharla et al. 2009; Michalak et al. 2008). In contrast, genetic maps are not precise physical representations of the genome because they rely on recombination events that may not be evenly distributed (Saintenac et al. 2009: Cox et al. 1990; Kalavacharla et al. 2009; Michalak et al. 2008). The biased distribution of cross-over events is associated with the specific requirements of the recombination machinery. In yeast and mouse, meiotic DSBs that initiate meiotic recombination are predominantly formed in open chromatin sites marked by trimethylation of lysine at position 4 in the H3 histone (Neale and Keeney 2006; Borde et al. 2009; Buard et al. 2009). Little is known about patterns of meiotic DSB formation in plants (Zhang et al. 2012), but they are likely similar to those in yeast and mammals, i.e. fewer DSBs are formed in the more densely packed chromatin regions (Perrella et al. 2010; Edlinger and Schlogelhofer 2011). Open chromatin regions should be more common in the distal regions of wheat chromosomes, where most of the genes are located (Saintenac et al. 2009).

Our data confirms that the frequency of CO events are partially dependent on the relative distance from the centromere ($p=0.05$) but fails to identify a similar correlation for BR frequency. This lack of correlation supports the hypothesis that RH maps are indeed true physical representations of chromosomes, and that the formation of deletions are random events independent of chromosome landmarks. However, it must be kept in mind that CO happens at meiosis, when chromosomes are highly condensed, whereas BR occurs in somatic cells mostly during mitotic interphase, when chromosome landmarks are difficult to observe. On the other hand, our data show a dependent distribution between CO and BR events, suggesting a similar preference for specific chromosomal regions for both mechanisms. In the literature, three plant studies have reported strong correlations between meiotic CO

frequencies and somatic DNA break and repair processes. Liu et al. (2009) showed that distribution of insertions of the *Mu* transposon in maize correlates with the distribution of recombination events across the genome. Similarly, Choulet et al. (2010) identified correlation between transposable element (TE) distribution and disruption of gene order conservation, which is accelerated by COs. The third study demonstrated a correlation between TE distribution and modification of gene order, as well as a correlation between non-syntenic gene order and the CO frequency (Rustenholz et al. 2011). A common causative factor that explains the observed correlations is the dependence of all three processes, TE insertion, gene order disruption and CO formation, on the formation of DSBs, which are preferentially originated within poised chromatin regions. Based on these reports and my data, I postulate that the correlation between CO frequency and BR frequency can also be explained by the dependence on DSBs formation of both processes, and their consequent dependence on the chromatin state.

### 2.4.3. A working hypothesis: how does the chromatin state affect the DNA-damage/repair mechanism?

Assuming that the breakage/repair mechanism has a preference for open chromatin regions, it is intriguing to hypothesize on how these regions specifically influence this process. There are two levels at which the state of chromatin could influence formation of chromosomal deletions: *i*) regions of compact chromatin could be more resistant to radiation damage; *ii*) the repair mechanism requires open chromatin regions to perform its activity (Figure 2.6). If the first hypothesis (Figure 2.6A) is correct, I should observe a reduction in the number of deletions in those regions that are particularly heterochromatic, such as the peri- and centromeric regions. The data presented in Figure 2.5 (and extended in Figure A.3) show no significant difference in the frequency at which deletions forms in the telomeres or centromere of chromosome 3B. This would suggest that chromatin condensation by itself is not a sufficient shield to prevent DNA damage. A similar conclusion was also reached for yeast (Linger and Tyler 2005) and humans (Goodarzi et al. 2010), leaving our first hypothesis short of supporting evidences. On the other hand, it leaves open the possibility that the chromatin state directly influences the repair mechanism, possibly preventing the formation of non-radiation-mediated DSBs, necessary to complete the repair. Goodarzi et al. (2008) demonstrated that in human cells the repair complex is unable to adequately

access or manipulate radiation-mediated breaks occurring in regions of compact chromatin. Hence, if radiation-mediated breaks happen independently from the chromatin state, but poised chromatin is necessary for the DNA repair mechanism to properly operate, logic dictates that in tightly compacted chromosomal regions I should expect larger and less frequent deletions, while in more open regions the number of deletions would increase and their size diminish (Figure 2.6B). That is precisely what I observed for chromosome 3B (Figure 2.4). In conclusion, our data support *in vivo* the hypothesis that "higher order chromatin architecture exerts just as profound an influence on DNA repair as it does on nuclear processes such as transcription and replication" (Goodarzi et al. 2010).

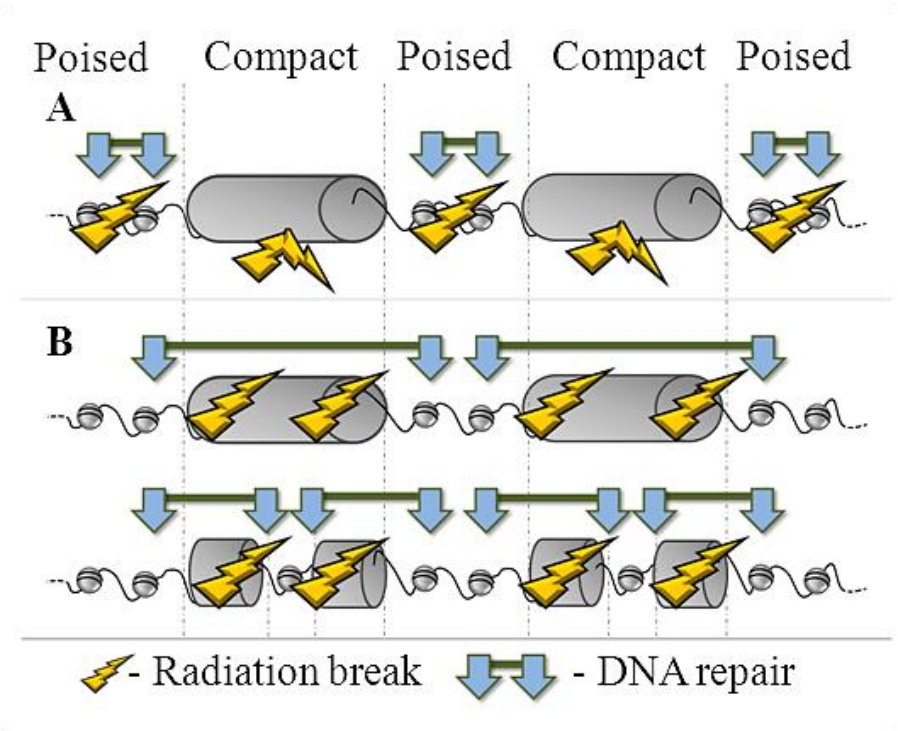


Figure 2.6. Working hypotheses of the effect of chromatin state on the DNA breakage/repair mechanism. Compact chromatin is represented by solid grey cylinders; poised chromatin is represented as 'beads on a string'; radiation-mediated DNA damage is represented by a yellow flash; the entry points for the DNA repair mechanism are represented as vertical blue arrows, while horizontal green lines indicate the size of the deletion created by the repair mechanism. A) Hypothesis one: compact chromatin prevents radiations from causing damage to the DNA, hence deletions only occur on poised chromatin. B) Hypothesis two: radiation-mediated damage can occur regardless of the chromatin state, but the DNA repair mechanism requires poised chromatin as entry point to effectively repair the DNA. Where large compact chromatin regions exist, the repair mechanism produces larger and less frequent deletions; where poised and compact chromatin regions alternate, the DNA repair mechanism produces smaller and more frequent deletions.

## 2.5. Conclusion

Radiation hybrid mapping is an effective approach to map all markers (monomorphic or otherwise) in wheat and other organisms. High levels of mapping resolution can be achieved with relatively small populations. Since its first application 37 years ago (Goss and Harris 1975) RH mapping was believed an approach totally independent from patterns of recombination. Here, for the first time, data have been collected that suggest otherwise, indicating that RH mapping relies on poised chromatin regions, probably corresponding to the recombination hot-spots regions observed in genetic mapping. However, the 3B-RH map generated still offered a resolution eleven fold higher than a comparable genetic map and a fairly consistent physical to cR conversion across the entire chromosome, making this approach the most dependable for the scaffold assembly of genome sequencing initiatives. Moreover, in plants I can produce entirely *in vivo* RH lines, providing a unique tool to study the effect of radiation in living organisms. New insights have been gathered in the past few years on possible interactions between the state of chromatin and various DNA break-repair processes, such as those involved in CO event formation, DNA repair, TE insertion, and synteny disruption. I presented here a working hypothesis to explain how the chromatin state could affect the DNA repair mechanism, together with the biological material to further investigate this hypothesis. Specifically targeted cytogenetic studies employing RH lines are likely to provide the evidences necessary to shed light on the precise effect of chromatin on the DNA break-repair mechanism.

## 2.6. References

Akhunov ED, Goodyear AW, Geng S, Qi LL, Echalier B, et al (2003) The organization and rate of evolution of wheat genomes are correlated with recombination rates along chromosome arms. Genome Res 13:753-763

Bernstein KA, Rothstein R (2009) At loose ends: resecting a double-strand break. Cell 137:807-810

Elsik CG, Tellam RL, Worley KC, Gibbs RA, Muzny DM, et al (2009) The genome sequence of taurine cattle: a window to ruminant biology and evolution. Science 324:522-528

Bleuyard JY, Gallego ME, White CI (2006) Recent advances in understanding of the DNA double-strand break repair machinery of plants. DNA Repair 5: 1-12

Borde V, Robine N, Lin W, Bonfils S, Géli V, Nicolas A (2009) Histone H3 lysine 4 trimethylation marks meiotic recombination initiation sites. EMBO J 28:99-111

Britt AB (1999) Molecular genetics of DNA repair in higher plants. Trends Plant Sci 4:1360-1385

Buard J, Barthès P, Grey C, de Massy B (2009) Distinct histone modifications define initiation and repair of meiotic recombination in the mouse. EMBO J 28:2616-2624

Choulet F, Wicker T, Rustenholz C, Paux E, Salse J, et al (2010) Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. Plant Cell 22:1686-1701

Cox DR, Burmeister M, Price ER, Kim S, Myers RM (1990) Radiation hybrid mapping: a somatic cell genetic method for construction of high-resolution maps of mammalian chromosomes. Science 250:245-250

Derrien T, Andre C, Galibert F, Hitte C (2007) AutoGRAPH: an interactive web server for automating and visualizing comparative genome maps. Bioinformatics 23:498-499

Dolezel J, Simkova H, Kubalakova M, Safar J, Suchankova P, et al (2009) Chromosome genomics in the Triticeae. In Genetics and Genomics of the Triticeae. Edited by Feuillet C, Muehlbauer GJ. New York: Springer pp 285-316

Edlinger B, Schlogelhofer P (2011) Have a break: determinants of meiotic DNA double strand break (DSB) formation and processing in plants. J Exp Botany 62(5):1545-1563

Endo TR, Gill BS (1996) The deletion stocks of common wheat. J Hered 87:295-307

Erayman M, Sandhu D, Sidhu D, Dilbirligi M, Baenziger PS, Gill KS (2004) Demarcating the gene-rich regions of the wheat genome. Nucleic Acids Res 32:3546-3565

Faraut T (2009) Contribution of radiation hybrids to genome mapping in domestic animals. Cytogenet Genome Res 126:21-33

Francki MG, Walker E, Crawford AC, Broughton S, Ohm HW, et al (2009) Comparison of genetic and cytogenetic maps of hexaploid wheat (Triticum aestivum L.) using SSR and DArT markers. Mol Genet Genomics 281:181-191

Gao W, Chen ZJ, Yu JZ, Raska D, Kohel RJ, et al (2004) Wide-cross whole-genome radiation hybrid mapping of cotton (Gossypium hirsutum L.). Genetics 167:1317-1329

Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, et al (2004) Genome sequence of the brown Norway rat yields insights into mammalian evolution. Nature 428:493-521

Gill KS, Gill BS, Endo TR (1993) A chromosome region-specific mapping strategy reveals gene-rich telomeric ends in wheat. Chromosoma 102:374-381

Givry S, Bouchez M, Chabrier P, Milan D, Schiex T (2005) Carthagene: multi population integrated genetic and radiated hybrid mapping. Bioinformatics 21:1703-1704

Goodarzi AA, Noon AT, Deckbar D, Ziv Y, Shiloh Y, et al (2008) ATM signaling facilitates repair of DNA double-strand breaks associated with heterochromatin. Molecular Cell 31:167-177

Goodarzi AA, Jeggo P, Lobrich M (2010) The influence of heterochromatin on DNA double strand break repair: getting the strong, silent type to relax. DNA Repair 9:1273-1282

Goss SJ, Harris H (1975) New method for mapping genes in human chromosomes. Nature 255:680-684

Hohmann U, Endo TR, Gill KS, Gill BS (1994) Comparison of genetic and physical maps of group-7 chromosomes from Triticum aestivum L. Mol Gen Genet 245:644-653

Hossain KG, Riera-Lizarazu O, Kalavacharla V, Vales MI, Maan SS, Kianian SF (2004) Radiation hybrid mapping of the species cytoplasm-specific ($scs^{ae}$) gene in wheat. Genetics 168:415-423

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al (2001) Initial sequencing and analysis of the human genome. Nature 409:860-921

Kalavacharla V, Hossain K, Gu Y, Riera-Lizarazu O, Vales MI, et al (2006) High-resolution radiation hybrid map of wheat chromosome 1D. Genetics 173:1089-1099

Kalavacharla V, Hossain KG, Riera-Lizarazu O, Gu Y, Maan SS, Kianian SF (2009) Radiation hybrid mapping in crop plants. Advances in Agronomy 102:199-219

Karere GM, Froenicke L, Millon L, Womack JE, Lyons LA (2008) A high-resolution radiation hybrid map of rhesus macaque chromosome 5 identifies rearrangements in the genome assembly. Genomics 92:210-218

Kynast RG, Okagaki RJ, Galatowitsch MW, Granath SR, Jacobs MS, et al (2004) Dissecting the maize genome by using chromosome addition and radiation hybrid lines. Proc Natl Acad Sci USA 101:9921-9926

Lin A, Wang RT, Ahn S, Park CC, Smith DJ (2010) A genome-wide map of human genetic interactions inferred from radiation hybrid genotypes. Genome Res 20:1122-1132

Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, et al (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438:803-819

Linger J, Tyler JK (2005) The yeast histone chaperone chromatin assembly factor 1 protects against double-strand DNA-damaging agents. Genetics 171:1513-1522

Liu S, Yeh C-T, Ji T, Ying K, Wu H, et al (2009) Mu transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome. PLoS Genet 5:e1000733

Lliakis G, Wang H, Perrault AR, Boecker W, Rosidi B, et al (2004) Mechanism of DNA double strand break repair and chromosome aberration formation. Cytogenet Genome Res 104:14-20

Lukaszewski AJ, Curtis CA (1993) Physical distribution of recombination in B-genome chromosomes of tetraploid wheat. Theor Appl Genet 86:121-127

Makalowski W (2003) Not junk after all. Science 300:1246-1247

Mézard C (2006) Meiotic recombination hotspots in plants. Biochem Soc Trans 34:531-544

Michalak M, Kumar A, Riera-Lizarazu O, Gu Y, Paux E (2008) High-resolution radiation hybrid mapping in wheat: an essential tool for the construction of the wheat physical maps. In: Apple R, Eastwood R, Lagudah E, Langridge P, Mackay M, McIntyre L, Sharp P (eds) Proceedings of the 11th International Wheat Genetics Symposium. Sydney University Press, Sydney pp 64-67

Michalak MK, Ghavami F, Gerard L, Yong G, Kianian SF (2009) Evolutionary relationship of nuclear genes encoding mitochondrial proteins across four grass species and Arabidopsis thaliana. Maydica 54:471-483

Neale MJ, Keeney S (2006) Clarifying the mechanics of DNA strand exchange in meiotic recombination. Nature 442:153-158

Ohno S (1972) Evolution of Genetic Systems. Smith HH (ed) Brookhaven SympBiol. Gorden & Breach, New York pp: 366-370

Olivier M, Aggarwal A, Allen J, Almendras AA, Bajorek ES, et al (2001) A high-resolution radiation hybrid map of the human genome draft sequence. Science 291:1298-1302

Paux E, Faure S, Choulet F, Roger D, Gauthier V, et al (2010) Insertion site-based polymorphism markers open new prespectives for genome saturation and marker-assisted selection in wheat. Plant Biotech J 8:196-210

Paux E, Roger D, Badaeva E, Gay G, Bernard M, et al (2006) Characterizing the composition and evolution of homoeologous genomes in hexaploid wheat through BAC-end sequencing on chromosome 3B. Plant J 48:463-474

Paux E, Sourdille P, Salse J, Saintenac C, Choulet F, et al (2008) A physical map of the 1-gigabase bread wheat chromosome 3B. Science 322:101-104

Perrella G, Consiglio MF, Aiese-Cigliano R, Cremona G, Sanchez-Moran E, et al (2010) Histone hyperacetilation affects meiotic recombination and chromosome segregation in Arabidopsis. Plant J 62:796-806

Puchta H (2005) The repair of double-strand breaks in plants: mechanisms and consequences for genome evolution. J Exp Bot 56:1-14

Riera-Lizarazu O, Leonard JM, Tiwari VK, Kianian SF (2010) A method to produce radiation hybrids for the D-genome chromosomes of wheat (Triticum aestivum L.). Cytogenet Genome Res 129:234-240

Riera-Lizarazu O, Vales MI, Ananiev EV, Rines HW, Phillips RL (2000) Production and characterization of maize chromosome 9 radiation hybrids derived from an oat-maize addition line. Genetics 156:327-339

Rodgers JL, Nicewander WA (1988) Thirteen ways to look at the correlation coefficient. Am Stat 42:59-66

Rustenholz C, Choulet F, Laugier C, Safár J, Simková H, et al (2011) A 3000-loci transcription map of chromosome 3B unravels the structural and functional features of gene islands in hexaploid wheat. Plant Physiol 157:1596-1608

Saintenac C, Falque M, Martin OC, Paux E, Feuillet C, Sourdille P (2009) Detailed recombination studies along chromosome 3B provide new insight on crossover distribution in wheat (Triticum aestivum L.). Genetics 181:393-403

Sourdille P, Singh S, Cadalen T, Brown-Guedira GL, Gay G, et al (2004) Microsatellite-based deletion bin systems for the establishment of genetic-physical map relationships in wheat (Triticum aestivum L.). Funct Integr Genomics 4:12-25

Sturtevant AH (1913) The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. J Exp Zool 14:43-59

Wade CM, Giulotto E, Sigurdsson S, Zoli M, Gnerre S, et al (2009) Genome sequence, comparative analysis, and population genetics of the domestic horse. Science 326:865-867

Wardrop J, Snape J, Powell W, Machray GC (2002) Construction of plant radiation hybrid panels. The Plant J 31:223-228

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520-562

Wenzl P, Suchánková P, Carling J, Simková H, Huttner E, et al (2010) Isolated chromosomes as a new and efficient source of DArT markers for the saturation of genetic maps. Theor Appl Genet 121:465-474

Xu Y, Price BD (2011) Chromatin dynamics and repair of DNA double strand break. Cell Cycle 10:261-267

Zhang W, Wu Y, Schnable JC, Zeng Z, Freeling M, et al (2012) High-resolution mapping of open chromatin in rice genome. Genome Research 22:151-162

# 3. RADIATION HYBRID QTL MAPPING TO LOCALIZE *desw2* INVOLVED IN THE FIRST MEIOTIC DIVISION OF WHEAT

Since the dawn of wheat cytogenetic, chromosome 3B has been known to harbor a gene(s) that when removed causes chromosome desynapsis and gametic sterility. However, the lack of natural genetic diversity for this gene(s) has prevented any attempts to fine map and further characterize it. The use of radiation to create artificial diversity was exploited here to refine the position of this locus. A population of 696 radiation hybrid lines was developed by treatment at 250 and 350 Greys of gamma ray. A custom mini array of 140 DArT markers selected to evenly span the whole 3B chromosome was developed and used to genotype the whole radiation hybrid population. The final map spanned 2,852 centi Ray (cR, map unti) with a calculated resolution of 0.384 Mb. The radiation hybrid lines were also phenotyped for the occurrence of meiotic desynapsis by determining the level of gametic sterility. Two tarits were used for this purpose: amount of seeds produced per spikelet and pollen vitality at booting. Three statistically unique phenotypic classes (p= 0.001) were determined based on three different doses of chromosome 3B, and then the classes were used to convert the phenotypic scoring of the radiation hybrid lines into their meiotic behavior. Composite interval mapping revealed a single QTL with LOD of 16.2 and $r^2$ of 25.6% between markers wmc326 and wPt-8983 on the long arm of chromosome 3B. By independent anlaysis, the location of the QTL was confirmed to be within the deletion bin 3BL7-0.63-1.00 and to correspond to a single gene located 1.4 Mb away from wPt-8983. Further, the meiotic behavior of lines lacking this gene was characterized cytogenetically to reveal its involvement in the maintainance of bivalents pairing at diakinesis. This represent the first example of employing radiation hybrids to conduct QTL analysis, and the successful results presented here provide an ideal starting point to achieve the final cloning of this interesting gene involved in meiosis of wheat.

## 3.1. Introduction

Sexual reproduction represents arguably the most important evolutionary drive for all superior organisms. In this light, the biological process of producing sexual gametes through meiotic cell division shines for its importance and complexity. The first meiotic division in particular has rather significant implications in plant evolution and breeding, since it is the stage during which crossing over and recombination occur. The synaptonemal complex is the tripartite proteinaceous structure that binds together the homologous chromosomes to facilitate the formation of chiasmata between non-sister chromosomes, and ultimately crossing over (Hassold et al. 2000). Following DNA replication, the chromosomes of germ cells start condensing, and short stretches defined axial elements are formed between sister chromatids (de Boer and Heyting 2006). These elements become clearly visible during leptotene and provide anchoring points for the gradual formation of the central elements of the synaptonemal complex, which continues throughout pachytene (Schramm et al. 2011). It is at this stage that crossing over takes place (Henderson 1970; for review see Kim 2001). Then, the synapsed and condensed homeologs progress through diplotene to diakinesis, where the synaptonemal complex is gradually degraded leaving discrete bivalents attached by one to three chiasmata (Ross et al. 1996). Even though the synaptonemal complex has been shown not to be essential for the formation of crossing over (Higgins et al. 2005; Qiao et al. 2012), it still maintains a primary role in bringing together the homoeologous chromosomes for recombination and in preventing aberrant cell division. Any mutations affecting the correct formation of the synaptonemal complex will inevitably result in incomplete homologous recognition, increase yield of univalent at diakinesis, and ultimately in a reduction in male and female fertility. These types of mutants are defined as *asynaptic*, when no bivalents are formed, or *desynaptic*, when both univalent and bivalents are originated (Bourne and Danielli 1979). The first of this mutation was described in 1930 in *Zea mays* and named *asynaptic1* (*as1*) (Beadle 1930). Wheat (*Triticum aestivum* L), followed 15 years later with the discovery of the recessive *ds* genes (Li et al. 1945). Since then, many more meiotic mutants have been described in plants with five additional mutations described in *Zea mays* (*dsy1* to *dsy5;* for review Zacheus and Freeling, 2011; http://bivouac.cshl.edu/genewiki/index.php/Main_Page), eleven in *Oryza sativa* L (*ds1* to *ds11*;

http://bivouac.cshl.edu/genewiki/index.php/Main_Page), at least three in *Arabidopsis thaliana* (*asy1*, *dsy1*, and *dsy10*; for review Caryl et al. 2003), and 15 in *Hordeum vulgare* (*des1* to *des15*; Franckowiak and Lundqvist 2008). This is only a partial list, but it is evident that all plant species have a variable number of genes that when removed or mutagenized cause desynapsis. In *Tritcum aestivum* a major gene involved in meiotic desynapsis was localized to chromosome 3B by Sears in 1954 while developing the aneuploidy stocks of the bread wheat variety 'Chinese Spring' (Sear 1954). Twenty years later, an independent research group reached the same conclusion, localizing a major desynaptic mutation on chromosome 3B of the Russian variety 'Avrora' (Zhirov et al. 1974). The existence of a similar gene(s) in *Triticum turgidum* ssp. *durum* L. was confirmed by Joppa and Williams (1988) while developing the variety 'Langdon' D-genome substitution lines (Joppa 1993). Devos et al. (1999) further confined the location of this gene to the long arm of chromosome 3B, because it was found that one dose of ditelosomic 3BL is sufficient to produce acceptable levels of fertility (Devos et al. 1999). This gene(s) does not have an official designation in wheat (McIntosh et al. 2011) and given its likely orthologous relationship to *des2* on barley chromosome 3H (Hernandez-Soriano 1973), I propose the label of *desw2*, where 'des' stands for desynaptic and 'w' stands for wheat.

The molecular characterization of this fertility related gene(s) on 3BL has so far proven elusive. Despites being identified by three independent research groups, and regardless of the fact that its chromosomal location has been known since the dawn of wheat genetics, its lack of obvious genetic diversity has until now prevented any attempts at map based cloning. In fact, any natural occurring mutations at this locus would typically result in a loss of fertility and consequently in a non-advantageous fitness. Hence, genetic variations for *desw2* are not likely to accumulate, practically preventing any possibilities of fine mapping this gene by means of allelic polymorphism and recombination.

One of the ways to move forward is to artificially generate genetic diversity at this locus. Martini and Bozzini (1966) successfully employed radiations to obtain desynaptic mutations in durum wheat. Radiation exposure damages the genome by causing random deletions (Baskaran 2010). When a deletion occurs on a gene (i.e. *desw2*) its expression is negatively affected, and it should ultimately result in a knock-out phenotype (Tanaka et al. 2010). In principle, there is no logical difference between the

phenotypic change observed in the aneuploids stocks of wheat that have been used in the past, where a portion of the chromosome carrying the given locus was removed through the activity of a gametocidal gene, and a change due to the removal of the specific chromosome portion by radiation mutagenesis (Sutka et al. 1999; Raman et al. 2005). Indeed, the method adopted by Bozzini and Martini (1966) succeeded in generating recessive desynaptic mutations, but unfortunately these authors failed to exploit the genetic diversity they generated to pinpoint the specific gene(s) that had been mutagenized.

Radiation hybrid (RH) is an experimental protocol also derived from radiation mutagenesis. It combines the DNA damage induced by mutagenesis with a hybridization step, to ultimately create mappable overlapping deletions that span the entire length of a chromosome (Gross and Harris 1975; Cox et al. 1990; see Faraut 2009 or Chapter 1 of this Thesis for review). This approach has been successfully adapted to wheat to generate high resolution molecular maps of its chromosomes 1D and 3B (Hossain et al. 2004; Kalavacharla et al. 2006; Paux et al. 2008). Among others, a great advantage of RH in wheat is that *in vivo* panels can be rapidly generated, allowing to combine the high map resolution characteristic of this technique, with the ease of phenotyping for knock out mutations. Here, I describe how RHs can be exploited to infer the genomic location of the *desw2* gene(s) responsible for desynapsis in wheat, a gene for which genetic diversity does not exist in nature.

## 3.2. Material and methods

### 3.2.1. Plant material and RH population

The present study used a RH mapping population developed for chromosome 3B (3B-RH panel) as described earlier (Paux et al 2008; see Figure 2.1 in Chapter 2 of this Thesis). In brief, seeds of a durum wheat cultivar ‘Langdon’ (LDN) were irradiation at 250 and 350 Gy doses of gamma ray and the resulting plants ($RH_0$ or $M_1$) were crossed to LDN D-genome substitution line for chromosome 3B [LDN 3D(3B)]. The line LDN 3D (3B) has a pair of durum wheat 3B chromosomes replaced by a pair of 3D chromosome transferred from 'Chinese Spring' hexaploid wheat (Joppa and Williams 1977, 1993). LDN 3D (3B) is male sterile, since the chromosome 3B contains a gene that is necessary to prevent desynapsis at meiosis. So this line is maintained as disomic 3D and monosomic 3B. For the purpose of

developing radiation hybrids, the plants nullisomic for 3B were selected from the progeny of disomic 3D monosomic 3B, using markers specific for chromosome 3B. LDN 3D(3B) was used as female and crossed with $RH_0$ plants to develop the 3B-RH panel. The seeds resulting from the cross are radiation hybrids generation 1 ($RH_1$), with chromosome 3B and 3D in monosomic condition (13" + 3B' + 3D'). The chromosome 3B in these $RH_1$s carries deletion(s) in homozygous (or more precisely 'hemizygous' since the 3B chromosome is monosomic) condition. A total of 696 $RH_1$ lines from the 3B-RH panel were used in the present study. In addition, normal double monosomic (13" + 3B' + 3D') $F_1$ lines (with no deletions) were also generated by crossing LDN 3D(3B) with non-irradiated LDN and employed as experimental controls. The hexaploid wheat cultivar 'Chinese Spring' (CS) and four of its chromosome 3B deletion bin line (Endo and Gill 1996) were also planted and used to locate the physical position of the DNA-markers.

All plants were grown under greenhouse controlled conditions (16-h light cycle; 15º-25º C; water, pesticide and fertilizer applications as required) at North Dakota State University (Fargo, ND; latitude: 46.8951, longitude: -96.8052, elevation: 275 m).

3.2.2. Phenotypic data on fertility/sterility

Two methodologies were employed to determine the overall level of male fertility/sterility of each plant. The spikes from the first, third and fourth tillers were bag isolated at emergence and imposed to self-pollinate. After dehiscence, the spikes were harvested; for each spike the number of spikelets and the number of seeds were counted. The level of fertility was determined for each spike as a ratio between the number of seeds produced and the total number of spikelets. The value of the three collected spikes was averaged to provide an overall 'seeds per spikelet' (SpS) score. This value is rather provided in its unit value ranging from 0.0 to 2.8, or as a percentage value of SpS fertility, where 100% correspond to the unit value 2.8 (max value reached in the experiment). The second method used to measure fertility/sterility levels, took advantage of the emerging spike from the secondary tiller. This was collected always between 7am and 11am to minimize pollen dehydration, and then fixed in a 3:1 ethanol to acetic acid solution. After 48h, the fixed spikes were moved to a 70% ethanol solution and maintained at 4º C until analysis. Three anthers for each spike were then dissected and stained on a microscope slide with a

non-toxic dye described in Peterson et al. (2010). The number of aborted and non-aborted pollen grains was counted by observing the stained pollen (Figure 3.1b) under a 20X magnification optical microscope (Zeiss, Germany). A minimum of 200 pollen grains were counted and the overall 'pollen vitality' (PV) score was calculated as the ratio between the number of non-aborted pollen grains over the total number of grains. The data on fertility was recorded for each $RH_1$ lines as well as for three durm control lines ( LDN, LDN 3D(3B), and their $F_1$ progeny) and two hexaploid control line (CS and CS 3BL7). The $RH_1$ lines, due to their unique karyotype makeup, could only be tested once during winter 2009 without replicates. The durum control lines, however, were tested in triplicate (replicates) across four greenhouse seasons in Fargo using a randomize complete block design: winter 2009, spring 2010, winter 2011, and spring 2011; while the hexaploid controls were tested in triplicate during the winter 2011 season.

Cytogenetic studies of the defective meiosis were conducted on both LDN 3D(3B) and normal LDN. The immature anthers from immature spikes were collected approximately two to three weeks before heading. Spikes of 1 to 4 mm in lenght were fixed and maintained as described for PV scoring. Anthers were transferred to a glass slide with the help of a dissecting microscope. The pollen mother cells were squeezed out of the anthers with the use of a scalpel, and then stained for five minutes using one drop of acetocarmine and moderate heath. The slides were scanned for cells undergoing meiotic division using a 40X magnification in an optical microscope (Zeiss, Germany). Cells showing aberration at any stage of meiosis were photographed with a low resolution camera at 100x magnification (Figure 3.1a). The chromosome rods at diakinesis were counted by hand. Any number exceeding 14 rods was indicative of aberrant diakinesis. LDN was used as positive control and always showed 14 rods at diakinesis, corresponding to the 14 chromosomes of durum wheat. In aberrant cells, the number of bivalents can be measured as 28 - # of rods.

### 3.2.3. Molecular analysis

DNA from each genotype was extracted from lyophilized leaf tissue of one month-old plants as described by Guidet et al. (1991). All DNA samples were equilibrated to a concentration of 50 ng/µl using the NanoDrop spectrophotometer (Thermo Scientific, DE). The integrity and quality of the DNA was

tested by placing 4 µl of extract in a 1X buffer solution for restriction analysis (without the actual restriction enzyme) at 35ºC for 6h, followed by separation on a 1% non-denaturing agarose gel, and visual evaluation after staining with GelRed (Biotium Inc., CA). The restriction buffer stimulates the activity of any DNAse present in the extract, hence promoting DNA degradation. High quality extracts show a single sharp band in close proximity of the gel wells and such extracts were used for further genotyping.

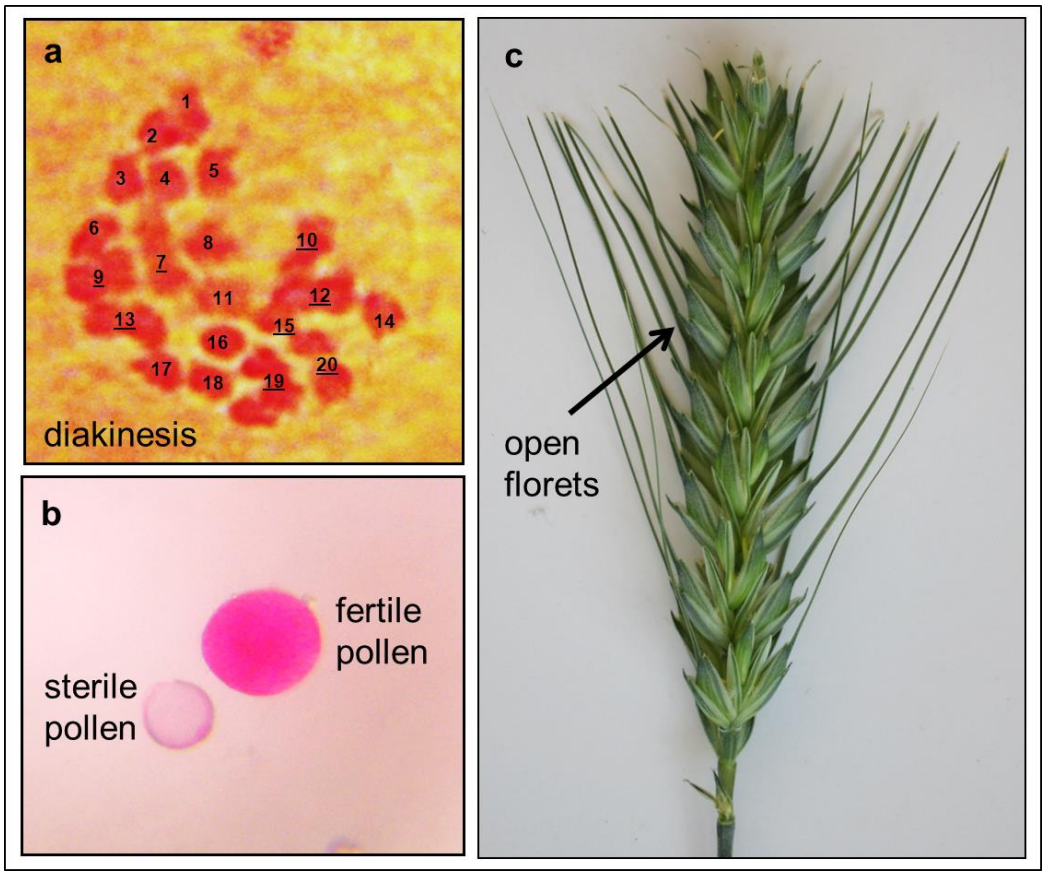


Figure 3.1. Physiological and cytological effects of removing the 3B chromosome from the genome of durum wheat. a) Optical microscope imaging (100X) of an abnormal diakinesis stage where the absence of *desw2* prevents the correct formation of the synaptonemal complex and result in abundance of monovalents. The chromosome rods have been numbered from top to bottom of the slide, underlined numbers indicate probable bivalents (8 bivalents + 12 univalents = 28 chromatids) b) Optical microscope imaging (20X) of mature pollen grains, where the sterile pollen grain results from abnormal meiosis. c) An adult wheat spike opening its florets to facilitate crosspollination, because the sterility of the pollen grains prevents it self-pollination.

The 3B-RH panel was genotyped using three different classes of markers: PCR based insertion site-based polymorphisms (ISBPs labeled as 'cfp'; Paux et al. 2006; Paux et al. 2010); simple sequence repeats (SSRs, labeled as 'gwm' and 'wmc'; Somer et al. 2004; Song et al. 2005) and Diversity Array Techology (DArT, labeled as 'wPt' and 'tPt'; Akbari et al. 2006; Wenzl et al. 2010). The SSR and ISBP markers were primarily used to confirm the genotyping data provided by the DArT markers. PCR amplification and scoring of ISBP and SSR markers was done as described in Chapter 2 of this Thesis. For DArT analysis, a total of 245 3B specific DArT markers were selected from earlier studies (Huang et al. 2012) and the map presented in Chapter 2. These markers were selected to represent the whole length of chromosome 3B. A mini custom array containing 490 markers was generated, with each of the 245 selected DArT clones blotted in duplicate. All 696 $RH_1$ lines, four chromosome 3B deletion bin lines, eight positive controls ($F_1$: 13" + 3B' + 3D') and eight negative controls [LDN 3D(3B): 13" + $3B^0$ + 3D"] were hybridized to the custom array following the provider protocol (Akbari et al. 2006; Wenz et al. 2010). Positive hybridization was scored as 'retention' and lack of hybridization was scored as 'deletion'. Only those markers that resulted in identical scoring between the two duplicates, and which hybridized only to the positive control but not to the negative, were accepted and used for the construction of the RH map of chromosome 3B. The retention frequency of a line measures the portion of markers retained over the total number of markers genotype, and is an inverse indicator of the amount of radiation damage sustained by a given line (i.e. 0.99 retention frequency means that only 1% of the markers identified a deletion, while 0.80 means that 20% of the marker loci had been deleted).

### 3.2.4. Statistical analyses

To generate a comprehensive map of chromosome 3B, the retention/deletion data on the 696 $RH_1$ lines genotyped with 140 markers was run on the Carthagene mapping software v1.2.2 (Givry et al. 2005), employing the iterative framework mapping procedure described earlier (Appendix A). The SAS 9.3 environment (SAS Institute, Cary, NC) was used to determine the significant differences of the fertility categories (t-test) by procGLM across four environments and then combining across environments (12 replicates) given the absence of statistical significance across environments (the SAS code and ANOVA

table are made available in Appendix B). SAS 9.3 was also employed to calculate the $R^2$ association between markers and phenotypes. Phenotypic data on fertility and RH mapping data was used to conduct QTL analysis using composite interval mapping (CIM) by Window QTL Cartographer V2.5 employing the default cofactors (Wang et al. 2011). The threshold LOD scores for detection of definitive QTL were also calculated based on 1,000 permutations (Churchill and Doerge 1994). Confidence intervals (CIs) were obtained by marking positions ± 2 LOD from the peak. Only $RH_1$ lines with retention frequency smaller than 0.999 and higher than 0.001 were used for the QTL analysis. Single marker regression was also employed to confirm the results and to identify the closely linked marker loci. For every trait, a difference between genotypes that is larger than one standard deviation unit in each environment was considered as significant. Broad sense heritability ($H^2$) was estimated using the formula:

$$1 - \frac{MSge}{MSg}$$

Which uses the mean sum of squares (MS) of the genotype (MSg) and genotype by environment (MSge) (Falconer and Mackay 1996; Tsilo et al. 2011).

## 3.3. Results

### 3.3.1. A custom DArT array for chromosome 3B

Based on genetic (Huang et al. 2012) and RH (Chapter 2) DArTs maps available, 245 DArT clones were selected to evenly cover the length of chromosome 3B. Each clone was blotted in duplicate on a mini custom array (DArT, Canberra, AU), and a total of 136 (55%) showed always positive hybridization with the positive control DNA (containing chromosome 3B) and no hybridization with negative control DNA (lacking chromosome 3B), for all eight replications, for both the markers duplicate. To deletion bin map these DArT markers, the custom array was also hybridized to the *Triticum aestivum* var. CS and four of its chromosome 3B deletion bin lines (Endo and Gill 1996). Twenty four markers (18%) hybridized to the CS DNA and could be precisely localized to one of the five deletion bins determined by the four deletion bin lines.

3.3.2. Characterization of a large 3B-RH population

The whole population of 696 3B-$RH_1$ lines was characterized using 136 DArT and four PCR markers. Based on 140 markers data, the average retention frequency across the whole 3B-RH panel was 0.80 (Figure 3.2). However, 233 (33%) of the lines had no detectable deletions or missed the entire chromosome 3B (133 lines, retention frequency = 1.00; 100 lines, retention frequency = 0.00). Thus, these RH lines were excluded from further analysis because not providing any exploitable mapping information. The remaining 463 $RH_1$ (53%) lines had a maker loss for chromosome 3B between 1% and 10%, while 11% of the lines had more than 10% of the chromosome deleted, with seven 3B-RH lines carrying less than 50% of the chromosome (Figure 3.2). Overall the retention frequency of these segregating lines was 0.93.

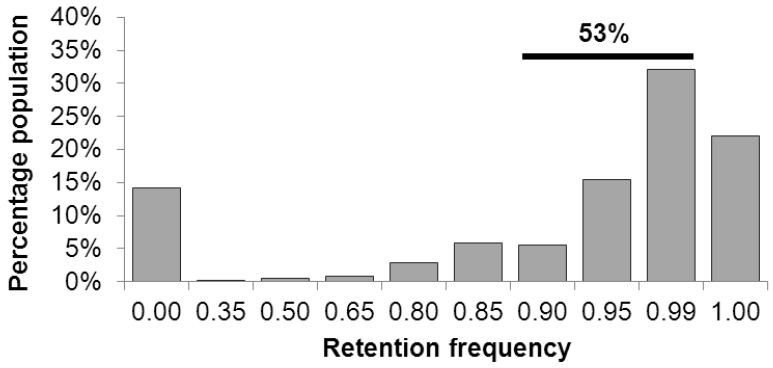


Figure 3.2. Distribution of retention frequencies in a population of 696 radiation hybrid lines. The majority of the lines have 1-10% of chromosome 3B deleted.

3.3.3. A radiation hybrid map for chromosome 3B

The 140 markers scorings on 463 $RH_1$ (= 696 – 233 non informative) lines were used to develop a framework RH map of chromosome 3B. All the markers formed a single group at LOD of 70.0 and distance of 50.0 cR. The markers order was maintained as predicted in the previous 3B-RH map (Chapter 2), while the markers that had not been previously mapped were positioned using an iterative frame mapping approach as described in Appendix A. The 28 markers (24 DArTs and 4 PCR-based) for which the bin location was known, respected their bin positioning in the final 3B-RH map, *de facto* sectioning the

3B chromosome into five large physical portions: 3BS telomere (3BS3-0.87-1.00), 3BS arm (3BS8-0.78-0.87), 3BS pericentromere (3BS1-0.33-0.55), centromere plus 3BL pericentromere, and the 3BL arm plus telomere (3BL7-0.63-1.00) (Figure 3.3).

The final 3B-RH map containing 140 markers spans a total distance of 2,852 cR. The average markers density is one marker every 20 cR, with a standard deviation of 8.3 cR; in fact 102 (73%) markers are spaced 20 ± 8.4 cR, while just 20 (14%) markers are at a distance superior to 28.4 cR with a maximum of 40.5 cR, and 18 (13%) at a distance inferior to 11.6 with a minimum of 1.6 cR. The even markers distribution confirms that the strategy adopted to select the DArT clones for the custom array provided uniform coverage of the 3B chromosome. This chromosome is 993 Mb in size (Dolezel et al. 2009), suggesting a map resolution of 0.348 Mb $cR^{-1}$ and average physical distance between the marker loci of ~7.0 Mb.

### 3.3.4. The 3B radiation hybrid population segregates for fertility

Chromosome 3B is known to carry a gene that influences fertility (Sears 1954; Joppa and Williams 1987; Devos et al., 1999). In this study, three genotypes carrying two, one, and zero doses of chromosome 3B (and the inverse number of chromosome 3D doses) were examined for their level of fertility (Table 3.1). The interaction between environments (four seasons in one location) and genotype was not significant, while genotype segregation was significant at 0.001 level of confidence. The number of seeds produced by each spikelet drastically decreased when reducing the dosage of chromosome 3B, being 2.81 in the normal disomic state, 0.77 in the hemizygous and 0.15 in nullisomic conditions. The LDN 3D(3B) lines produce zero to one seed per spike, showing the characteristic open florets phenotype (Figure 3.1c) indicative of strong male sterility. Further, the pollen analysis (Figure 3.1b) showed that normal LDN lines (containing both copies of chromosome 3B) have a pollen vitality of 80%, while it was 60% in the $F_1$ (containing a single copy of chromosome 3B) and only 23% in LDN 3D(3B) (entirely missing chromosome 3B) (Table 3.1). Thus, using both seed set (SpS) and pollen vitality (PV), allowed the identification of three classes of fertility, depending on the dosages of chromosome 3B (Table 3.1).

Further, the CS deletion line 3BL7-0.63-1.00, deficient for the telomeric portion of 3BL, analyzed for SpS and PV showed sterility levels non-significantly different from the line missing the entire 3B chromosome.

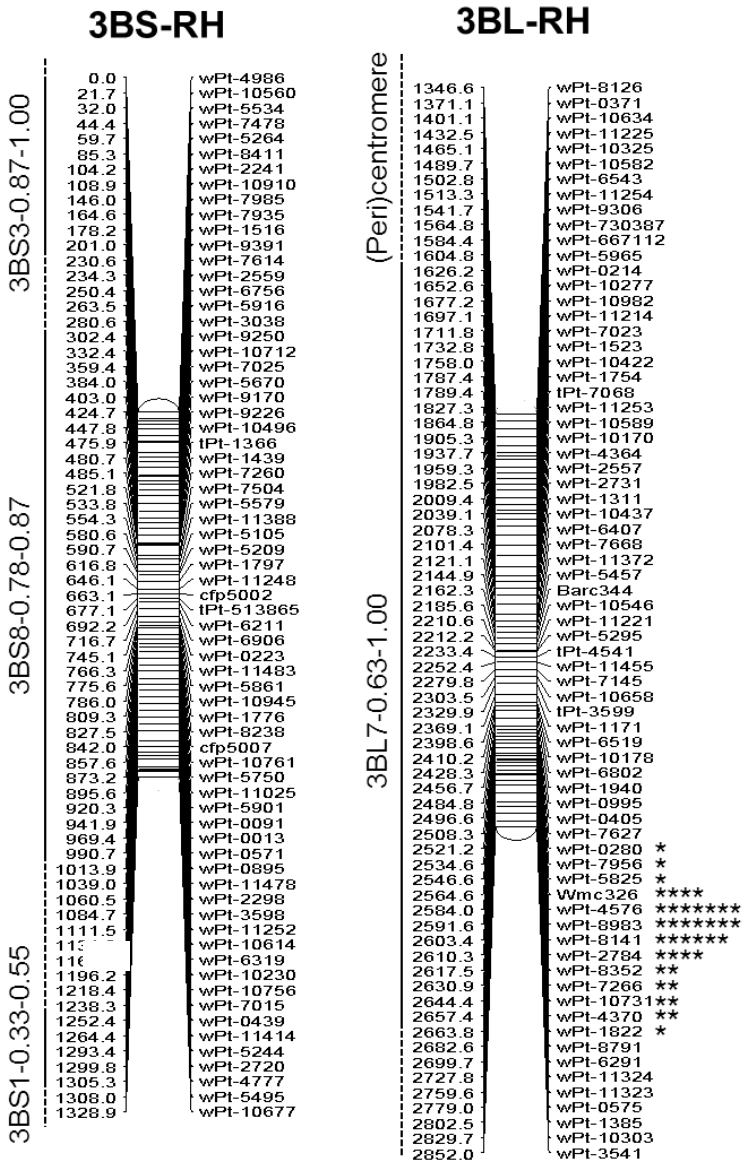


Figure 3.3. Radiation hybrid (RH) map of chromosome 3B short and long arms based on 140 markers loci and 696 RH lines. Map distances are reported in cR. The chromosome is divided into deletion bins (solid vertical line) on the basis of known marker position; the remaining chromosomal locations (dashed vertical line), including the centromere and pericentromere, are derived on the basis of map positions. The location of the fertility related *desw2* locus is presented as significance of single marker regression analysis, each star (*) represent a decimal of significance beyond *p:*0.0001 (i.e. **** indicates *p:*0.0000001).

Table 3.1. Variations of fertility at different dosages of chromosome 3B

| Karyotype | Seeds per spikelet † | | Pollen vitality † | | Fertility class |
|---|---|---|---|---|---|
| 13" + 3B" + $3D^0$ | 2.81 | a | 0.80 | a | fully fertile |
| 13" + 3B' + 3D' | 0.77 | b | 0.60 | b | partially fertile |
| 13" + $3B^0$ + 3D" | 0.16 | c | 0.23 | c | sterile |
| 20" + 3BL7-0.63-1.00" ‡ | 0.15 | c | 0.18 | c | sterile |
| LSD | 0.49 | | 0.07 | | |
| $H^2$ | 0.98 | | 0.98 | | |

", disomic; ', monosomic; $^0$,nullisomic; $H^2$, broad sense heritability
† Average of 12 replicates (3 replicates across four seasons).
‡ Average of 3 replicates collected in one season.

The broad sense heritability was extremely high (0.98) for both traits, but it must be pointed out that this heritability is calculated on the segregation of chromosome 3B rather than single genes, and that the test environments were not significantly different.

The 3B-$RH_1$ lines have the chromosome 3B in monosomic state, which might or might not carry deletions as consequence of the radiation treatment. Thus, 3B-$RH_1$ mosomic lines are expected to have moderate levels of fertility similar to the $F_1$ mosomic line; this level of sterility should drop to the nullisomic state of LDN 3D(3B) when the *desw2* gene(s) responsible to provide fertility is lost through radiation-damage.

Six hundred forty-six 3B-RH lines were phenotyped for SpS (50 lines never reached flowering), and those lines that had a value of SpS falling in between two fertility classes (Table 3.1) were scrutinized under the microscope to measure PV. A total of 376 lines were characterized for PV. Unfortunately, $RH_1$ lines are unique genotypes which segregate in a non-Mendelian fashion in the following filial generations; hence, only single plant measurements without replication could be performed. However, in order to increase the accuracy, multiple measurements (sampling) were obtained from each $RH_1$ line. Also, an overall fertility value ('1' for fertile and '0' for sterile) was given only when the value of SpS clearly fell into one of the fertility classes (Table 3.1) for all the sampling/measurements collected. When both SpS and PV values were collected, an overall fertility value was given only to those lines showing good agreement between the two phenotype scorings. The values for these two traits for all lines were converted into a

percentage of fertility as compared to the most fertile 3B-RH line (SpS=2.8 and PV=0.9) and are presented in Figure 3.4. Overall, for 616 lines (89%) it was possible to determine value for fertility (rather a clearly determined SpS, or both SpS and PV), 123 being sterile (20%) and 493 fertile (80%). Of these phenotyped lines, 416 $RH_1$ also segregated for chromosome 3B at the genetic level (i.e. retention frequency between 0.01 and 0.99), with 42 sterile lines (10%) and 374 fertile (90%). In addition, 47 lines segregate for chromosome 3B but do not fall within any of the fertility classes and were then excluded from the QTL study.

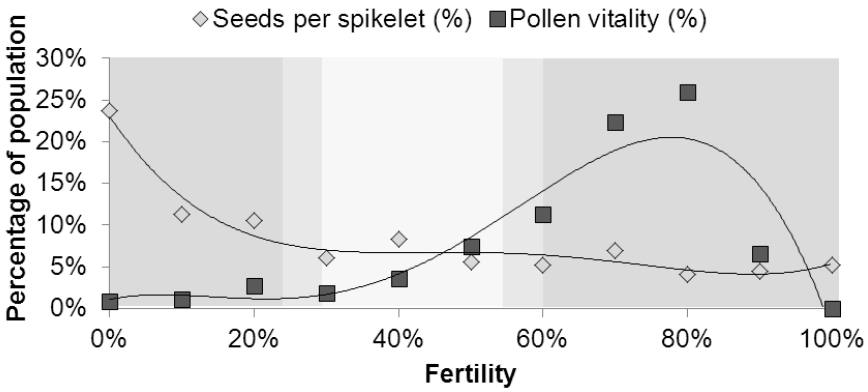


Figure 3.4. Segregation among 696 radiation hybrid lines for two fertility-related traits. Both traits are presented as percentage of maximum fertility as compared to a fully fertile control line. The four motions polynomial trend lines separate the values into two classes, also marked by the darker gradient of grey background. Lines falling within the lighter gray area (30-55%) had a fertility state that could not be precisely determined.

3.3.5. Mapping fertility-related gene *desw2*

Association between the 140 marker genotypes and the overall fertility value of the $RH_1$ lines was investigated by mean of QTL analysis. Those lines without any detectable deletions or lacking the entire 3B chromosome were excluded from the QTL analysis because would identify a positive association at every marker locus. Composite interval mapping of the 416 informative 3B-$RH_1$ lines identified a single QTL, named *Qdesw2.ndsu-3B* (following guidelines by McIntosh et al. 2011), with a LOD score of 16.2 and a $R^2$ of 25.6% (Fig 3.5). The threshold score based on 1000 permutations was 6.9. This QTL is flanked by the marker loci *wmc326* and *wPt-8141*, spanning a 38.8 cR region (2,564.6 - 2,603.4 cR). Based on the overall resolution of this map, the QTL interval corresponds to an approximate physical size of 13 Mb.

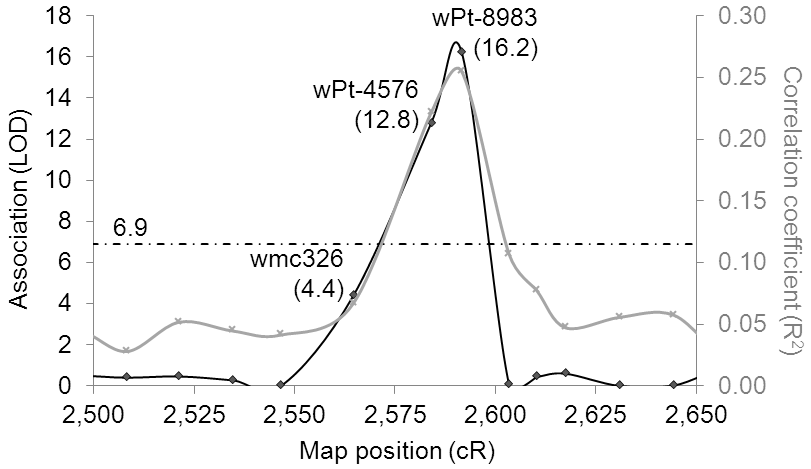


Figure 3.5. Location of the *desw2* fertility-related QTL on the chr. 3B radiation hybrid map. Association between fertility status and genotyping data of 416 radiation hybrid lines was computed by composite interval mapping. The LOD values are reported in the black line (left Y axis) and the $R^2$ values in the grey line (right Y axis). The dashed line indicates the threshold significance value calculated with 1,000 permutations. The significant marker loci are reported with their LOD values in parenthesis.

To further confirm the results, all 140 mapped loci were searched for association with the overall fertility score using single markers regression. A significant association ($p= 1*10^{-8}$) was observed for markers *wPt-4576* and *wPt-8983* which were located at positions 2584.0 and 2591.6 cR, respectively on the RH map. These two markers are mapped within the markers interval *wmc326* and *wPt-8141* where CIM identified the major QTL, thus, confirming the results of CIM. These markers are located inside the 3BL7-0.63-1.00 deletion bin (Figure 3.3). Marker *wPt-8983* has a retention frequency of 0.93 among the 416 segregating lines; it is deleted in 18 of the sterile lines (42%) and retained in 347 of the fertile (93%), for a total of 365 (92%) of the lines with genotyping and phenotyping perfectly matching

### 3.3.6. The desynaptic nature of the 3B fertility gene

The substitution line LDN 3D(3B) lacks the entire 3B chromosome, including the *desw2* locus. As a result, this line suffers from drastic male sterility (Table 3.1) and partial female sterility (data not show). Twenty-three immature spikes were collected from LDN 3D(3B) plants and analyzed for behavior at meiosis. The progress through the initial stages of meiosis does not seem to be affected by the lack of the 3B chromosome, and RAD51 foci can be detected as indicators of a fully formed synaptonemal complex

at pachytene (picture cannot be included in this Thesis, but it is available through collaboration with Dr Wojtek Pawloswki, Cornell University; Pawlowski et al. 2003). However, at diakinesis it was observed an increase in the number of univalents. In the course of this study, 18 LDN 3D(3B) cells were found at the diakinesis stage, with a range of 16 to 20 chromosome rods, which correspond to 4 to 16 univalents or 12 to 8 bivalents (Figure 3.1a), respectively. On average, LDN 3D(3B) produced 10.1 ± 1.4 bivalents at diakinesis. It might be worth specifying that LDN grown under the same greenhouse conditions always generated 14 bivalents at diakinesis.

## 3.4. Discussion

### 3.4.1. Characterization of a valuable radiation hybrids population for chromosome 3B studies

A complete RH map of chromosome 3B, comprising 540 markers mapped employing 92 3B-RH lines, was initially generated (Chapter 2). Here, the mapping information was extended to a much larger RH population. A mini custom DArT array was developed specifically for the task. This allowed to drastically reduce the cost of the analysis, thus expanding the size of the population to be genotyped. Further, the blotting of a custom array consented to increase the built-in experimental controls, and hence the final quality of the marker scoring, which is not a secondary concern when genotyping for radiation-mediated deletions (see Chapter 2 for further comments).

The whole 3B-RH population was characterized at 140 loci (136 DArTs and 4 PCR-based) and resulted in 2,852 cR RH map. The 140 marker loci were evenly spaced across the chromosome at an average interval of 20 cR, calculated to correspond to ~7 Mb in physical size. The interval that separates the markers is smaller than the average size of a wheat radiation-mediated deletion, which was calculated at 26 Mb for this 3B-RH population (Chapter 2). This shows that the number and relative distance of the markers is sufficient to recognize the vast majority of the radiation-mediated deletions occurring on the 3B chromosome, while still reducing sufficiently the experimental cost to allow genotyping of the whole 3B-RH panel.

Among the 3B-RH lines screened, 100 (14.4%) missed the whole chromosome 3B. Typically, the $RH_1$ lines are expected to have a single copy of chromosome 3B or a portion of it. However, the

existence of lines with the entire chromosome 3B missing could be explained by several reasons. First, LDN 3D(3B) was not completely male sterile. Phenotypic analysis showed 5.7% seed set and 28% PV in LDN 3D(3B) as compared to the normal LDN (Table 10). Since LDN 3D(3B) was used as female for the production of the $RH_1$ lines, unexpected self-fertilization would inevitably result in non-$RH_1$ lines lacking the whole 3B chromosome. A second explanation could be that chromosome 3B is not stable in monosomic conditions, and it is possible that in rare cases the entire chromosome lags behind during cell division, producing $RH_1$ lines completely missing the chromosome (Kynast et al. 2001). Third, some of the lines might actually result from the over damaging of the chromosome by the radiation treatment. In those cases, functional components such as the centromere or telomere might have been removed, preventing the chromosome from undergoing proper cell division (Tiwari et al. 2012). In this latter case, small portions of the 3B chromosome could be translocated to other part of the genome, too small in size to be recognized by the 140 markers employed here (Kalavacharla et al. 2006; Chapter 2).

The characterization of the 3B-RH panel also identified 133 lines not showing deletion for any of the 140 markers. These lines may rather have escaped radiation damage to chromosome 3B, or the deletions that have been created are too small (>7Mb) to be identified in this study. It is in fact an established knowledge in wheat RH that the increase of markers saturation is typically accompanied by the identification of additional smaller deletions (Kalavacharla et al. 2006; Chapter 2).

The 463 $RH_1$ lines carrying sizable deletions along the chromosome are potentially very valuable biological material for studies of 3B. Together these lines subdivide the 3B chromosome into nearly 3,000 small deletion bins of approximately 350 Kb in size. This is a very high map resolution for wheat, rivaled only by the 200 Kb resolution reached in a RH map of chromosome 1D (Kalavacharla et al. 2006).

### 3.4.2. Precise location of a major gene for fertility/sterility in wheat

The whole 3B-$RH_1$ population was phenotyped by mean of two fertility measurements (SpS and PV). The borad sense heritability ($H^2$) was extremely high for both traits, but it is important to remind that this value applies to the inheritance of chromosome 3B as whole, rather than single gene(s). However, in the absence of precise experimental controlled conditions both SpS and PV are known to be influenced

by the environment (Elkonin and Tsvetova 2012). In order to minimize the experimental error caused by the environmental variability, the measurements for these traits should be replicated across different environments (Hallauer and Miranda 1988). Unfortunately, $RH_1$ lines are unique genotypes that cannot be replicated, each individual characterized by a unique set of deletions. Upon selfing of $RH_1$, the non-monosomic chromosomes would segregate at meiosis, unmasking the deletions that the heterozygous state was masking [carried by the chromosome copy derived from the irradiated LDN 3D(3B)]. Also, the monosomic chromosomes (3B and 3D in this case) would not pair at meiosis, segregating in a non-Mendelian fashion. More importantly, all the sterile $RH_1$ lines would not produce any seeds to proceed to the following generations, thus, making it impossible to replicate the $RH_1$ lines across environments. To provide a meaningful value of fertility from single plants measurements, multiple samples were collected for each of the two traits, and a single overall value of fertility was generated merging the two traits values. Three check genotypes with decreasing dosages of chromosome 3B were phenotyped for SpS and PV in three replicates across four environments. Non-significant interactions were detected between genotypes and environments suggesting good experiemtnal control, while significant ($p= 0.001$) differences were observed between all three genotypes for both traits (Table 3.1). On the basis of their values, three fertility classes were proposed. The overall fertility scoring of the 3B-RH lines was decided in comparison to these three classes.

Interestingly, in RH the genetic variability for a given trait, when ignoring the environment influence, is due solely to the occurrence of a deletion on the gene(s) controlling the trait. The frequency at which a locus is deleted in a RH population can be measured as one minus the retention frequency. The 3B-RH population has a retention frequency of 0.80 and hence a locus deletion frequency of 20%. Any 3B loci involved in the control of the fertility phenotype should then be deleted with the same frequency as any other. If one single locus is involved in the control of the fertility phenotype, then the frequency of knock out phenotypes should match the frequency of one locus deletion. This is the case of *desw2*, with 20% of the 3B-RH lines being sterile, matching the 20% population wise locus deletion frequency. Based on these calculations, a single locus should be involved in controlling the observed phenotype.

QTL analysis localized the *desw2* locus to an interval of 38.4 cR between the marker loci *wmc326* and *wPt-8141.* These markers were located in the telomeric bin of chromosome 3BL (3BL7-0.63-1.00). The estimated cR to Mb conversion ratio for the deletion bin 3BL7 was 0.348 Mb $cR^{-1}$ (Chapter 2), so for the 38.4 cR interval would correspond to a physical size of 13 Mb. An independent confirmation of the correct positioning of the *Qdesw2.ndsu-3B* QTL was derived by the sterility level observed for the 3BL7-0.63-1.00 deletion bin line. Further, Devos et al. (1999) found that the ditelosomic for chromosome 3BS are sterile, while Joppa and Williams (1988) reported that long arm of chromosome 3B is necessary to maintain LDN 3D(3B) substitution line. All the evidences support our localization of the *desw2* locus to the telomeric or subtelomeric portion of the long arm of chromosome 3B.

Marker *wPt-8983* was found as the most strongly associated locus but explains only the 25.6% of the phenotypic variation, but precisely predicts the phenotype of 92% of the 3B-RH lines as based on its retention/deletion scoring. There is a partial discrepancy between the relatively low $R^2$ value and the fact that *desw2* is apparently a single gene. However, the phenotypic variation in RH population it is not due to allelic recombination and hence does not follow Mendelian genetics, meaning that the fertility/sterility values are not normally distributed. This might contribute to reduce the genotype to phenotype coefficient of determination. Further, there are 28 uncharted breakages between marker *wPt-8983* and the locus governing this fertility phenotype, which correspond to a map distance of 4 cR. Based on the map to physical conversion ratio, marker wPt-8983 is approximately 1.4 Mb away from the functional locus. Still, the resolution of the 3B-RH map is so elevated (~350 Kb) that even a marker less than 1.4 Mb away from the gene controlling the function, is predictive for just the 25.6% of the phenotypic variation. In addition, there are 32 lines that lack the entire 3B chromosome but still show a fertile phenotype, likely carrying a very small translocated portion containing *desw2*. Further, 15 lines with no identified deletions show a sterile phenotype, probably because they harbor a small undetected deletion within the *desw2* region (see Chapter 1 for additional comments). Together, these 75 informative lines (commonly defined “recombinants” in genetic mapping studies) will be useful to precisely pinpoint the location of the *desw2* gene once complete markers saturation is achieved.

3.4.3. The *desw2* gene is involved with the formation of chiasmata or disgregation of the synaptonemal complex during meiosis

The preliminary data gathered here provide novel and interesting information about the biological function of the *desw2* gene. Probably the most surprising piece of evidence is the fact that a dosage effect was observed for this gene. Lines hemizygous for the locus, as represented by the double monosomic $F_1$, suffer partial sterility. This indicates that a copy on each 3B sister chromosomes is required for proper meiosis. Further, the DNA undergoes re-synthesis and duplication during interphase, meaning that in a hemizygous line one copy would be duplicated to two, instead of the duplication from two to four that occurs in a homozygous line. It is unclear whether all of the re-synthesised copies are actually expressed during meiotic division, but if that was the case the lack of one copy would maximize its effect, changing from four expressed loci to just two. Also, only two of the four forming gametic cells will receive a copy of *desw2* at the end of meiosis. While this would explain the nearly 50% sterility observed (i.e. only 50% of the gametes have *desw2*), it does not explain why the most drastic effect are observed at diakinesis during the first meiotic division, while the single meiotic cell still carries all available copies of *desw2*. However, what is most interesting for practical applications is the question of whether, since a reduction in *desw2* copies has a quantifiable decrease on the overall plant fertility, an increase of its copy number would provide an increase in overall plant fertility, or maybe positively affect recombination?

Previous studies have proposed that lack of the chromosome 3B results in chromosomes desynapsis during meiosis (Sears 1954; Joppa and Williams 1987; Devos et al., 1999). In the past, the term desynapsis was used in a loose sense, meaning any mutation that ultimately resulted in the formation of univalents at diakinesis (Bourne and Danielli 1979). As the scientific knowledge has progressed over the years, the different aspects of meiosis have been better understood. Today the term de-synapsis (or a-synapsis) is more precisely indicative to the failure of properly forming the synaptonemal complex (Pawlowski et al. 2003). The removal of the *desw2* locus does not seem to affect the formation of the synaptonemal complex per se, which occurs at pachytene. Rather the following

stage(s) of meiosis is affected, during which the synaptonemal complex is degraded and the bivalents are kept together by the crossing-over chiasmata that were initiated during pachytene. So it is not immediately obvious whether the ∆*desw2* mutation has really a desynaptic effect as indicated by older studies. Yet, Sears (1954), the father of wheat cytogenetics, has defined this mutation as desynaptic in times that preceded the clear understanding of the synaptonemal complex formation, and all authors that followed employed his definition. It seems then unwise to change the name today, but it has to be kept in mind that the *desw2* gene might not be causing desynapsis *in senso stricto*.

The observed surge of univalent at diakinesis could be explained by one (or all) of three major hypothesis. *i*) The synaptonemal complex is formed properly, but it occurs between non-homologous chromosomes, which would then fail to cross-over and to remain united as bivalents at diakinesis. The pairing of homologous in wheat is known to be under the control of a different locus termed *Ph1* on chromosome 5B, but there are no data to support that this is the only locus controlling homologous pairing. However, in *ΔPh1* lines the non-homologous paired chromosomes undergo nearly normal crossing-over and multivalents, rather than univalents, are observed at diakinesis (Sears 1977; Al-Kaff et al. 2008). Hence, the evidences from the literature fail to support the hypothesis that *desw2* is a gene(s) controlling homologous chromosome pairing. *ii*) The *desw2* is involved in regulating the degradation of the synaptonemal complex. Removal of this gene would then cause an improper or anticipated degradation of the proteinaceous structure that holds the homologous chromosomes together, ultimately leading to univalents. If this hypothesis was confirmed to be true, then the definition of desynaptic mutation would indeed be correct. *iii*) The *desw2* locus controls or favors the formation of crossing-over between chromosomes. The lack of this gene would then reduce the amount of chiasmata forming between homologs chromosomes, causing the bivalents that are not hold together by crossovers to separate once the synaptonemal complex is degraded at diakinesis. The total number of univalents would then depend on the amount of chiasmata forming without the help provided by the *desw2* gene. Also, an increasing amount of *desw2* copies would result in an increasing number of chiasmata, and ultimately in a reduction of univalents. This last hypothesis fits best the data currently available, since it provides an explanation for the dosage effect that was observed.

Clearly, the use of technologies that provide higher optical resolution, such as tri-dimensional FISH or electron microscopy of staged anthers, are required to ultimately determine the biological function of *desw2* at meiosis of wheat. In this regard, the complete sequence of the gene is an indispensable requirement to proceed to its molecular characterization. This study represents the first step toward the precise dissection of the *desw2* locus.

## 3.5. Conclusions

The chromosomal location of the *desw2* gene has been known for nearly 60 years, but its lack of genetic diversity has precluded any attempt to fine map this locus. Here, the use of a non-conventional QTL analysis on RH allowed localizing the elusive *desw2* to a region of just few Mb in size. Radiation hybrids QTL analysis represents a novel approach that holds great potential to map all those genes for which no genetic diversity is currently known. Often these genes control those life-depending biological functions that are so dear to biologists, but that are difficult to study because of their lack of natural mutations. In this regard, RH functional mapping provides a good experimental solution to a difficult challenge.

The molecular function of *desw2* remains unclear, but the data presented here allowed for a first glimpse to its involvement in wheat meiosis. The practical applications of this gene are many and significant. If it was to be confirmed that an increase in copy number of this locus affects the amount of crossing-over and chiasmata forming between homologs chromosomes, it would certainly become a primary target for advances in recombination-based breeding. Also, the silencing of this gene provides a nuclear system to produce female plants, the first fundamental step for commercial production of wheat hybrids. Moreover, from an academic prospective, the unraveling of the function of this gene would help in understanding the complex mechanism that regulates meiosis in polyploid species, arguably one of the most important aspects that led to the evolution of wheat and other polyploid crops. Still, in order to achieve all of these goals, the complete sequence of this gene has to be obtained by means of map based cloning. The QTL analysis and the RH population resources presented here are only the first step in the difficult path toward the cloning of the *desw2* locus, but are at least a first step in the right direction.

3.6. References

Akbari M, Wenzl P, Caig V, Carling J, Xia L, et al (2006) Diversity arrays technology (DArT) for high-throughput profiling of the hexaploid wheat genome. Theor Appl Genet 113:1409-1420

Al-Kaff N, Knight E, Bertin I, Foote T, Hart N, Griffiths S, Moore G (2008) Detailed dissection of the chromosomal region containing the Ph1 locus in wheat Triticum aestivum: with deletion mutants and expression profiling. Ann Bot 101:863-872

Baskaran R (2010) Emerging role of radiation induced bystander effects: Cell communications and carcinogenesis. Genome Integr 1:13

Beadle GW (1930) Genetic and cytological studies of a Mendelian asynaptic in Zea mays. Cornell Agric Exp Sta Mem 129:1–23

Bourne GH, Danielli JF (1979) International review of cytology volume 58. Academic Press Inc, London

Caryl AP, Gareth HJ, Franklin FCH (2003) Dissecting plant meiosis using Arabidopsis thaliana mutants. J of Exp Botany 54(380):25-38

Churchill GA, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. Genetics 138:963-971

Cox DR, Burmeister M, Price ER, Kim S, Myers RM (1990) Radiation hybrid mapping: a somatic cell genetic method for construction of high-resolution maps of mammalian chromosomes. Science 250:245-250.

De Boer E, Heyting C (2006) The diverse roles of transverse filaments of synaptonemal complexes in meiosis. Chromosoma 115:220-234

Devos KM, Sorrells ME, Anderson JA, Miller TE, Reader SM, et al (1999) Chromosome aberrations in wheat nullisomic-tetrasomic and ditelosomic lines. Cereal Research Communications 27(3):231-239

Dolezel J, Simkova H, Kubalakova M, Safar J, Suchankova P, et al (2009) Chromosome genomics in the Triticeae. In: Feuillet C, Muehlbauer GJ (eds) Genetics and Genomics of the Triticeae. Springer, New York, pp 285-316

Elkonin LA, Tsvetova MI (2012) Heritable effect of plant water availability conditions on restoration of male fertility in the “9E” CMS-inducing cytoplasm of sorghum. Front Plant Sci 3:91

Endo TR, Gill BS (1996) The deletion stock of common wheat. J of Hered 87(4):295-307

Falconer DS, TFC Mackay (1996) Introduction to quantitative genetics 4th Ed. Addison Wesley Longman Limited, Edinburgh Gate, Harlow Essex, England

Faraut T (2009) Contribution of radiation hybrids to genome mapping in domestic animals. Cytogenet Genome Res 126:21-33

Franckowiak JD, Lundqvist U (2008) Table of barley genetic stock description. Barley Genetics Newsletter 38:134-164

Givry S, Bouchez M, Chabrier P, Milan D, Schiex T (2005) Carthagene: multi population integrated genetic and radiated hybrid mapping. Bioinformatics 21:1703-1704

Goss SJ, Harris H (1975) New method for mapping genes in human chromosomes. Nature 255:680-684

Guidet F, Rogowsky P, Taylor C, Song W, Langridge P (1991) Cloning and characterisation of a new rye-specific repeated sequence. Genome 34:81-87

Hassold T, Sherman S, Hunt P (2009) Counting cross-overs: characterizing meiotic recombination in mammals. Hum Mol Genet 9(16):2409-2419

Henderson AS (1970) The time and place of meiotic crossing-over. Annu Rev Genet 4:295-324

Hernandez-Soriano JM, Ramage RT (1973) Desynaptic genes. Barley Genetics Newsletter 4:123-125

Higgins JD, Sanchez-Moran E, Armstrong SJ, Jones GH, Franklin FC (2005) The Arabidopsis synaptonemal complex protein ZYP1 is required for chromosome synapsis and normal fidelity of crossing over. Genes Dev 19(20):2488-2500

Hossain KG, Riera-Lizarazu O, Kalavacharla V, Vales MI, Maan SS, Kianian SF (2004) Radiation hybrid mapping of the species cytoplasm-specific ($scs^{ae}$) gene in wheat. Genetics 168:415-423

Huang BE, George AW, Forrest KL, Kilian A, Hayden MJ, Morell MK, Cavanagh CR (2012) A multiparent advanced generation inter-cross population for genetic analysis in wheat. Plant Biotechnol J 10:826-839

Joppa LR, Williams ND (1977) D-genome substitution monosomics of durum wheat. Crop Sci. 17: 772 - 776

Joppa LR, Williams ND (1988) Langdon durum disomic substitution lines and aneuploid analysis in tetraploid wheat. Genome 30:222-228

Joppa LR, Williams ND (1993) The Langdon durum disomic-substitutions: development, characteristics, and uses. Agron Abstr p 68

Kalavacharla V, Hossain K, Gu Y, Riera-Lizarazu O, Vales I, et al (2006) High-resolution radiation hybrid map of wheat chromosome 1D. Genetics 173:1089-1099

Kim N (2001) Disseminating the genome: joining, resolving, and separating sister chromatids during mitosis and meiosis. Annu Rev Genet 35:673-745

Kynast RG, Okagaki RJ, Rines HW, Phillips RL (2002) Maize individualized chromosome and derived radiation hybrid lines and their use in functional genomics. Funct Integr Genomic 2:60-69

Li HW, Pao WK, Li CH (1945) Desynapsis in the common wheat. American Journal of Botany 32(2):92-101

Maluszynski M, Szarejko I (2003) Induced mutations in the green and gene revolutions. In: Tuberosa R, Phillips RL, Gale M (eds) Proceedings of the international congress “In the wake of the double helix: from the green revolution to the gene revolution”, 27-31 May 2003 Bologna, Italy. Avenue Media, Bologna, pp 403-425

Martini G, Bozzini A (1966) Radiation induced asynaptic muattions in durum wheat (Triticum durum Desf.). Chromosoma 20:251-266

McIntosh RA, Dubcovsky J, Rogers WJ, Morris C, Apples R, Xia XC (2011) Catalogue of gene symbols for wheat: 2011 supplement. Annual Wheat Newsletter 57

Paux E, Roger D, Badaeva E, Gay G, Bernard M, Sourdille P, Feuillet C (2006) Characterizing the composition and evolution of homoeologous genomes in hexaploid wheat through BAC-end sequencing on chromosome 3B. Plant J 48:463-474

Paux E, Sourdille P, Salse J, Saintenac C, Choulet F, et al (2008) A physical map of the 1-gigabase bread wheat chromosome 3B. Science 322:101-104

Pawlowski WP, Golubovskaya IN, Cande WZ (2003) Altered nuclear distribution of recombination protein RAD51 in maize mutants suggests the involvement of RAD51 in meiotic homology recognition. The Plant Cell 15:1807-1816

Peterson R, Slovin JP, Chen C (2010) A simplified method for differential staining of aborted and non-aborted pollen grains. International Journal of Plant Biology 1:e13

Qiao H, Chen JK, Reynolds A, Hoog C, Paddy M, Hunter N (2012) Interplay between synaptonemal complex, homologous recombination, and centromeres during mammalian meiosis. PLoS Genet 8(6):e1002790

Raman H, Zhang K, Cakir M, et al (2005) Molecular characterization and mapping of ALMT1, the aluminium-tolerance gene of bread wheat (Triticum aestivum L.). Genome 8:781-791

Ross KJ, Fransz P, Jones GH (1996) A light microscope atlas of meiosis in Arabidopsis thaliana. Chromosome Res 4:507-516

Schramm S, Fraune J, Naumann R, et al (2011) A novel mouse synaptonemal complex protein is essential for loading of central elemnt proteins, recombination, and fertility. PLoS Genet 7(5):e1002088

Sears ER (1954) The aneuploids of common wheat. Missouri Agr Exp Sta Res Bull 572:1-58

Sears ER (1966) Nullisomic-tetrasomic combinations in hexaploid wheat. In: Riley R, Lewis KR (eds) Chromosome manipulations and plant genetics, Oliver and Boyd Ltd., Edinburgh, UK

Sears ER (1977) An induced mutant with homoeologous paring in common wheat. Candian Journal of Genetics and Cytology 19:585-593

Somers DJ, Isaac P, Edwards K (2004) A high-density microsatellite consensus map for bread wheat (Triticum aestivum L.). Theor Appl Genet 109:1105-1114

Song QJ, Shi JR, Singh S, et al (2005) Development and mapping of microsatellite (SSR) markers in wheat. Theor Appl Genet 110:550-560

Sutka J, Galiba G, Vagujfalvi A, Gill BS, Snape JW (1999) Physical mapping of Vrn-A1 and Fr1 genes on chromosome 5A of wheat using deletion line. Theor Appl Genet 99:199-202

Tanaka A, Shikazono N, Hase Y (2010) Studies on biological effects of ion beams on lethality, molecular nature of mutation, mutation rate, and spectrum of mutation phenotype for mutation breeding in higher plants. J Radiat Res 51:223-233

Tiwari VK, Riera-Lizarazu O, Gunn HL, Lopez K, Iqbal J, et al (2012) Endosperm tolerance of paternal aneuploidy allows radiation-hybrid mapping of the wheat D-genome and a measure of γ ray-induced chromosome breaks. PLoS ONE

Tsilo TJ, GA Hareland, S Chao, JA Anderson (2011) Genetic mapping and QTL analysis of flour color and milling yield related traits using recombinant inbred lines in hard red spring wheat. Crop Sci 51:237-246

Wang S, Basten CJ, Zeng Z-B (2011) Windows QTL Cartographer 2.5. Department of Statistics, North Carolina State University, Raleigh, NC

Wenzl P, Suchánková P, Carling J, Simkova H, Huttner E, et al (2010) Isolated chromosomes as a new and efficient source of DArT markers for the saturation of genetic maps. Theor Appl Genet 121:465-474

Zacheus WC, Freeling M (2011) Inna Golubovskaya: the life of a geneticist studying meiosis. Genetics 188(3):491-498

Zhirov EG, Bessarab KS, Gubanova MA (1974) Genetic studies on partial desynapsis in soft wheat. Soviet Genetics 9(1):10-18

# 4. RADIATION HYBRID MAP OF CHROMOSOME 1D REVEALS SYNTENY CONSERVATION AT A WHEAT SPECIATION LOCUS [2]

The evolution of *Triticeae* species is a complex process involving both nuclear and cytoplasmic genomes. The *species cytoplasm specific* (*scs*) genes affect nuclear-cytoplasmic interactions (NCI) in interspecific hybrids of *Triticum* tribe and play an important role in the speciation of *Triticeae* tribe. A radiation hybrid (RH) mapping population of 188 individuals, segregating for the $scs^{ae}$ gene localized on chromosome 1D of wheat, was employed to map 58 1D-specific markers (ESTs and retrojunction markers). A novel methodology, employing flow sorted chromosomes, was used to develop PCR-based markers. The RH map of 1D spanned 205.2 cR and pinpointed the position of the $scs^{ae}$ to a 1.1 Mb interval containing eight candidate genes. This is represents the first study in which $RH_2$ lines were employed for forward genetic studies, opening the possibility of using existing $RH_1$ populations for functional studies of other genes. Furthermore, the online tool Wheat Zapper (wge.ndsu.nodak.edu/wheatzapper/) was used to examine syntenic relationships between wheat chromosome 1D and related chromosomes of *Oryza sativa* L., *Brachypodium distachyon* L., and *Sorghum bicolor* L. The RH map of 1D provided sufficient resolution to pinpoint the location of an ancient fusion between two paleochromosomes and to verify that the centromere of chromosome 1D is the result of a neocentromerization event rather than retention of an ancient centromere. This study provides insightful information on the map-based cloning of the $scs^{ae}$ gene and its evolutionary importance, as well as on the molecular breeding application of this $scs^{ae}$ to unlock the secondary and tertiary gene pools of wheat.

---

[2] The material in this chapter was co-authored by Filippo Bassi and Monika Michalak de Jimenez. Monika Michalak developed the population and generated the markers coded NDSU with three digits (i.e. NDSU290). Filippo Bassi designed the experiment, conducted all other analysis, and he is the primary developer of the conclusions that are advanced here. Filippo Bassi drafted and revised all versions of this chapter.

## 4.1. Introduction

The plant cell has a tripartite inheritable genetic information unequally distributed between the nuclear, mitochondrial and chloroplast genomes. In the course of evolution some of the plastid genes have migrated to the nuclear genome, preventing any of the three cell components to function independently (Chase 2007). In fact, the reliance of the organelles on nuclear-encoded genes is counter parted by the dependence of the nucleus on the metabolic activities carried out in the organelles. The result is a tight cross-talk between the three genomes, which is referred to as "nuclear-cytoplasmic interactions" (NCI). Under normal cell conditions this continuous exchange of information is hard to follow. However, when the NCI are artificially disrupted undesirable phenotypes emerge, such as maternally inherited male sterility, female sterility, delayed maturity, reduced vigor, deficiency of chlorophyll and carotenoids, as well as altered morphology of cotyledons, leaves and flowers (Maan 1991a; Anderson and Maan 1995; Zubko et al. 2001; Bogdanova 2007; Amarath et al. 2011). In *Triticum* the *species cytoplasm-specific* (*scs*) genes are nuclear-encoded genes involved in NCI and act as "intrinsic post-zygotic barriers to reproduction", which ultimately result in hybrid unviability (Riesberg and Blackman 2010; Maan 1991a, Maan 1995). It can be derived that the surge of these reproductive barriers has prevented free interspecific hybridizations, leading to the non-random direction of the speciation process (Rieseberg and Willis 2007).

The effects of *scs* genes are usually masked in normal plant cells with compatible nucleus and cytoplasm, but when the cytoplasm of wheat (*Triticum turgidum* ssp. *durum* L. or *Triticum aestivum* ssp. *aestivum* L.) is replaced with the cytoplasm of one of its wild relatives, the importance of *scs* becomes apparent. This condition is defined alloplasmic (*allo*-alien, *plasmic*-cytoplasm) and it is typically obtained through a process called substitution backcrossing, where the wild relatives are used as female parents and wheat is used as the recurrent pollinator (Wu et al. 1998). This technique has been employed to generate cytoplasmic male sterile lines during the early days of hybrid wheat (for review Cisar and Cooper 2002). However, it is important not to confuse the *scs* genes with the male fertility restorer genes (*Rf*), because the *scs* genes are not responsible for restoring fertility, but rather are involved in the maintenance of NCI to provide adequate vigor and viability in alloplasmic plants (Maan 1992d).

The tetraploid nucleus of durum wheat appears to suffer the alloplasmic conditions more dramatically than the hexaploid nucleus of bread wheat. For example, the cytoplasm of *Aegilops longissima* S. & M., *Aegilops tauschii* Coss (syn: *T. tauschii*, *Ae. squarrosa*), *Aegilops cylindrica* Host, and *Aegilops uniaristata* Vis. produce vigorous and viable alloplasmic plants when combined to the nucleus of hexaploid wheat, but fail to survive with the tetraploid nucleus (Sasakuma and Maan 1978; Maan 1983; Maan 1992c; Asakura et al. 1997). A well-studied example of an incompatible alloplasmic interaction is a line that combines the tetraploid nucleus of wheat with the cytoplasm of *Ae. longissima* ($S^1S^1$; 2n=2x=14). In this case, the alloplasmic line can survive only when one of the 1A chromosomes of the durum nucleus carries a portion of the 1A chromosome from *Triticum timopheevii* ssp. *timopheevii* (AAGG; 2n=4x= 28) or of the 1D chromosome from *T. aestivum*. These two chromosomes have been shown to carry *scs* genes capable of restoring adequate NCI between the nucleus of wheat and the cytoplasm of *Ae. longissima* (*lo*). These genes have been defined as $scs^{ti}$ from T*. timopheevii* and $scs^{ae}$ from *T. aestivum*. The alloplasmic lines that are derived from these genes can be maintained in hemizygous (or homozygous) conditions, (*lo*) $scs^{ti}$ - and (*lo*) $scs^{ae}$ -, and upon test-crossing with genotypes that do not carry active *scs* genes segregate into shriveled and plump seeds, of which only the latter germinate (Figure 4.1) (Maan 1992c, Hossain et al. 2004b). These two *scs* genes have been studied extensively and their general positions on chromosomes 1AL and 1DL are known (Simmons et al. 2003; Hossain et al. 2004). While alternative allelic versions exist for $scs^{ti}$ (Gehlhar et al. 2005), chromosome 1D it is not known to harbor any alternative allele for the $scs^{ae}$ gene, practically preventing the possibility of further refining its map location by recombination-based studies.

Radiation hybrid mapping, on the other hand, is an experimental design based on radiation-induced null-alleles (deletions) for mapping of markers and phenotypes, and it is consequently independent from the necessity of allelic polymorphism. Using this procedure it was possible to effectively pinpoint the location of the $scs^{ae}$ locus to the proximal portion of deletion bin 1DL2-0.41-1.00 (Hossaine et al. 2004; Kalavacharla et al. 2006).

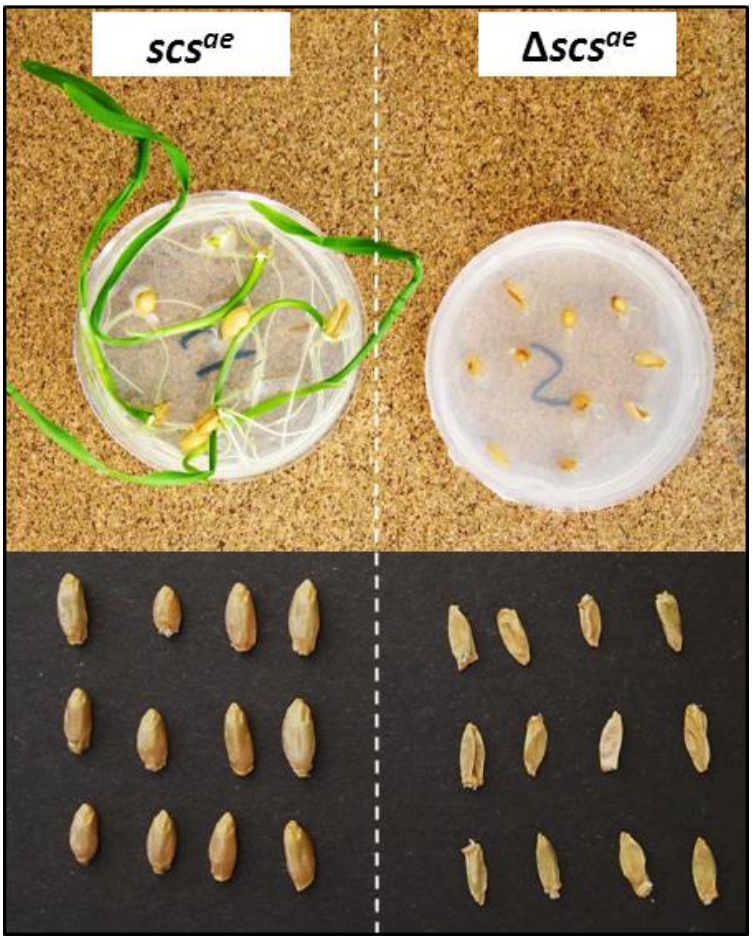


Figure 4.1. Alloplasmic durum wheat line with the cytoplasm of *Aegilops longissima* can survive only if the *scs*$^{ae}$ gene is present in their nucleus. Plump seeds (left), with the *scs*$^{ae}$ gene present, germinate properly when cultured on organic media, while shriveled seeds (right), without the *scs*$^{ae}$ gene, do not germinate at all.

In this study the RH methodology is further utilized for the fine mapping of the *scs*$^{ae}$ gene by genotyping of a $RH_1$ population and phenotyping their $RH_2$ progenies. Surprisingly, the *scs*$^{ae}$ location correspond not only to a locus responsible for controlling the speciation of the *Triticeae* tribe, but also to an evolutionary active region, where two ancestral chromosomes fused to create the modern wheat chromosome 1D (Luo et al. 2009).

4.2. Materials and methods

4.2.1. Plant material

4.2.1.1. Radiation hybrids

Seeds of the *Triticum turgidum* ssp. *durum* L. cultivar 'Langdon' (LDN) [AABB; 2n=4x=28, 13"+1A"] and its aneuploid 'Langdon 1D (1A)' (LDN 1D (1A)) [AABB; 2x=4x=28, 13" +1D"] were obtained from USDA-ARS (Fargo, ND) (Joppa 1993). Seeds of LDN 1D (1A) were equilibrated at approximately 13% of moisture content using a 60% glycerol solution, and then gamma irradiated at 150 or 350 Greys (Gy) using an Acel Gamma Cell 220 (North Dakota State University, Department of Veterinary Microbiology, Fargo, ND). The mutagenized $M_1$ seeds were grown under greenhouse controlled conditions; hand emasculated, and used as pollen recipients from non-irradiated LDN plants. The seed resulting from this artificial pollination belong to the first radiation hybrid generation ($RH_1$). These are double-monosomic [13" +1A' +1D'$^{150 \text{ or } 350 \text{ Gy}}$], where any radiation-mediated deletions carried by the chromosomes from LDN 1D (1A) are masked by the non-irradiated LDN chromosomes, with the exception of the monosomic 1D, where all deletions are fixed in the homozygous (actually hemizygous since it is a monosomic chromosome) state. This crossing scheme is explained in detail by Michalak et al. (2009). Also, non-irradiated double-monosomic [13" + 1A' + 1D'] lines were created and used as experimental controls.

4.2.1.2. Alloplasmic lines

The North Dakota State University wheat alloplasmic stock has been developed by Shivcharan S. Maan (Maan 1973; Maan 1978; Sasakuma and Maan 1978). This study focuses on the durum alloplasmic line that carries the cytoplasm of *Ae. longissima* (*lo*), obtained through over 20 cycles of substitution backcrossing. The resulting plants are male sterile and can be maintained only by artificial crossing. The (*lo*) *scs*$^{ti\ ti}$ code identifies a specific durum alloplasmic line, in which a portion of chromosome 1A of *T. turgidum* has been replaced by the 1A portion of *T. timopheevii* harboring the *scs* gene (Simons et al. 2003). However, this line is maintained by backcrossing to its B-line *T. turgidum* '56-1' that does not carry any functional *scs* gene. The seeds resulting from this artificial cross are theoretical hemizygous, since

there are no evidence for a *scs*$^{1A}$ version of the gene (in which case they would be heterozygous), and are coded as (*lo*)*scs*$^{ti-}$. Instead, (*lo*) *scs*$^{ae-}$ is a true hemizygous since it is maintained as a double monosomic alloplasmic line (13" + 1A' + 1D.1AL) with a 1D.1AL translocation, and without a second 1D chromosome for pairing (Maan 1999).

4.2.2. Molecular analyses

4.2.2.1. Markers development

Four strategies were employed for markers development. *i*) Retrotransposon junction-based markers (Ret) were designed from the BAC-end sequences of the *Aegilops tauschii* [DD; 2n=2x=14] library (Li et al. 2004), using the RJ-primer tool available at http://avena.pw.usda.gov/RHmapping/index.html (Wanjugi et al. 2009). Gene-derived markers were coded as NDSU (North Dakota State University) following the indications of the Catalogue of Gene Symbols for Wheat (McIntosh et al. 2008). *ii*) The primers coded NDSU200 to NDSU212 and NDSU226 to NDSU233 are designed on the known 1D-specific SNPs (resource available at http://wheat.pw.usda.gov/SNP/project.html ; NSF project "Haplotype Polymorphism in Polyploid Wheats and their Diploid Ancestors", PI: J. Dvorak). *iii*) The NDSU213-225 and NDSU290 or higher were developed on the basis of the orthologus information available for *Oryza sativa* L. chromosome 10. Wheat tentative consensus (TC) EST sequences were blast searched from the "Gene Index Project" database (http://compbio.dfci.harvard.edu/tgi/plant.html; Quackenbush et al. 2001) using the orthologus rice genes as query. The TC with the best hit (minimum E values of ≤1.0 $e^{-40}$) was aligned to the rice genomic sequence; the portions that did not align were considered to represent the introns as suggested by Schnurbusch et al. (2007). Primer-Blast (Ye et al. 2012) was used to design primers spanning intron-exon junctions. *iv*) The NDSU markers coded 3 to 31 were developed following the same procedure described before (*iii*) with the addition of a sequencing step. Primers spanning intron-exon boundaries were used to PCR amplify DNA from the flow sorted chromosomes 1A, 1B, and 1D (Vrana et al. 2000). The resulting amplicons were separated by electrophoresis on a 1.5% agarose gel; the sharpest bands were removed from the gel, eluted following the mnufacturer instructions (Gel Elution Kit, Qiagene, Valencia, CA), and

sequenced at GeneWiz (South Plainfield, NJ). The sequences from the three homoeologs were then aligned using the ChromasPro software (Technelysium Pty Ltd, Helensvale, QLD, Australia), and searched for homoeologous sequence variations (HSVs). Primers of 20-23 mers were then designed by hand to present the 1D-specific HSV(s) at their 3' ends. All the primers designed following any of these four approaches were tested for chromosome 1D specificity as follows: PCR were carried out in a 20 µl solution containing 1X GoTaq Green Buffer, 2.0 mM of $MgCl_2$, 5% DMSO, 2.5 mM of each dNTPs, 0.25 µM of each primers, 0.75 units of GoTaq polymerase (Promega, Madison, WI), and 50 ng of DNA from LDN or LDN 1D (1A). The amplification reaction was a standard touch-down with 10 cycles of 30 seconds of denaturation at 94º C, 30 seconds of annealing with the temperature decreasing of 0.5º C at each cycle, and 45 seconds of extension at 72º C, followed by 30 identical cycles but with annealing temperatures 5 º C lower than the initial step, and a final extension of 7 minutes at 72º C. Four touch-down annealing temperatures were tested for each primer combinations: 65º C to 60º C, 60º C to 55º C, 58º C to 53º C, and 55º C to 50º C. The resulting amplicons were separated on 9% non-denaturing polyacrylamide gels and stained with ethidium bromide. Only those markers that showed a band present for LDN 1D (1A) and absent in LDN were considered 1D-chromosome specific and used for genotyping.

4.2.2.2. Genotyping

DNA isolation was carried out as described in Hossain et al. (2004). Genotyping of the RH lines was performed by PCR amplification employing the annealing temperature and ideal conditions determined in the 1D-specificity test described above. The primer sequences and PCR details are made available in Appendix C. The amplification profiles of the RH lines were scored as '1' when the 1D-specific band was present and '0' when the band could not be distinguish. To avoid scoring a failed PCR reaction as a radiation-mediated deletion, the control marker DEASY (Chapter 2) was multiplexed in each reaction, and a scoring of '0' was given only when the 1D-specific band was absent but the DEASY band was present. For every marker used for genotyping, the band obtained amplifying LDN 1D (1A) DNA was cut from the gel, eluted, and sequenced. The sequence obtained was then aligned against the DNA sequence initially employed for the primer design to confirm proper target amplification. Also, the

sequences were searched against the mapped EST database (Qi et al. 2004; available at http://wheat.pw.usda.gov/GG2/blast.shtml) to identify their corresponding deletion-bin.

4.2.3. Statistical analysis

The genotyping data was converted into the 1D-RH map employing the Carthagene v1.2.2 software (Givry et al. 2005) as described by Michalak et al. (2009) with the superimposed algorithm described in Chapter 2 of this Thesis. The map was then hand curated to respect the synteny and deletion-bin mapping information, but strictly maintaining the overall map size. The colinearity studies were performed using the 'Wheat Zapper' on-line tool (http://wge.ndsu.nodak.edu/wheatzapper/) to determine the syntenic relationships to *Oryza sativa* L., *Sorghum bicolor* L., and *Brachypodium distachyon* L. All other statistical analyses were conducted using the SAS v9.3 package (SAS Institute, Cary, NC). The retention frequency is a measure of the frequency at which a given chromosome is deleted within a given RH population, and it is calculated as one minus the ratio between the overall number of genotyped deletions (i.e. '0') and the total number of genotyping data points (see Chapter 2 for further details).

4.2.4. Phenotyping

The 1D - $RH_1$ lines are monosomic for 1D therefore upon selfing there is theoretically a ~25% chance of losing the 1D chromosome (see measured monosmic 1D transfer frequency in Martin et al. 2011). For each $RH_1$ line considered for phenotyping, eight $RH_2$ seeds were planted and screened using NDSU3 and NDSU212. Lines that still carried chromosome 1D where rather monosomic (~75% chances) or disomic (~25% chances) for it. Two to three $RH_2$ plants for each $RH_1$ line were selected and used for test-crossing. The presence and number of copies of the *scs* gene in a given genotype can be determined by test-crossing it to a (*lo*) $scs^{ti}$ – line. Genotypes that do not carry any *scs* gene in their genome will result in a segregation of 1:1 plump to shriveled seeds (Figure 4.1 and Figure 4.2), with the shriveled seeds lacking the *scs* gene. A genotype hemizygous (i.e. monosomic) for *scs* will result in a ratio of 3:1 plump to shriveled seeds, while a genotype homozygous (disomic) for the *scs* will produce

only plump seeds (i.e. ratio 1:0; Figure 4.2). The $RH_2$ lines used in test-crossing are expected to fit the 1:0 or 3:1 ratio when the *scs*$^{ae}$ locus has not been affected by the radiation treatment or the 1:1 ratio in case a deletion occurred in the gene region. The seed ratio of plump to shriveled was statistically tested using the $\chi^2$ analysis. For each phenotyping at least three independent test-crosses were produced and statistically tested for both shriveled vs. plump and plump vs. shriveled. The degrees of freedom of the analysis were then considered as the number of test-crosses times two (plump and shriveled) minus one. Both 1:1 and 3:1 ratios were measured, the most fitting with *p* of at least 0.05 was considered as the correct ratio and used to predict the state of *scs*$^{ae}$.

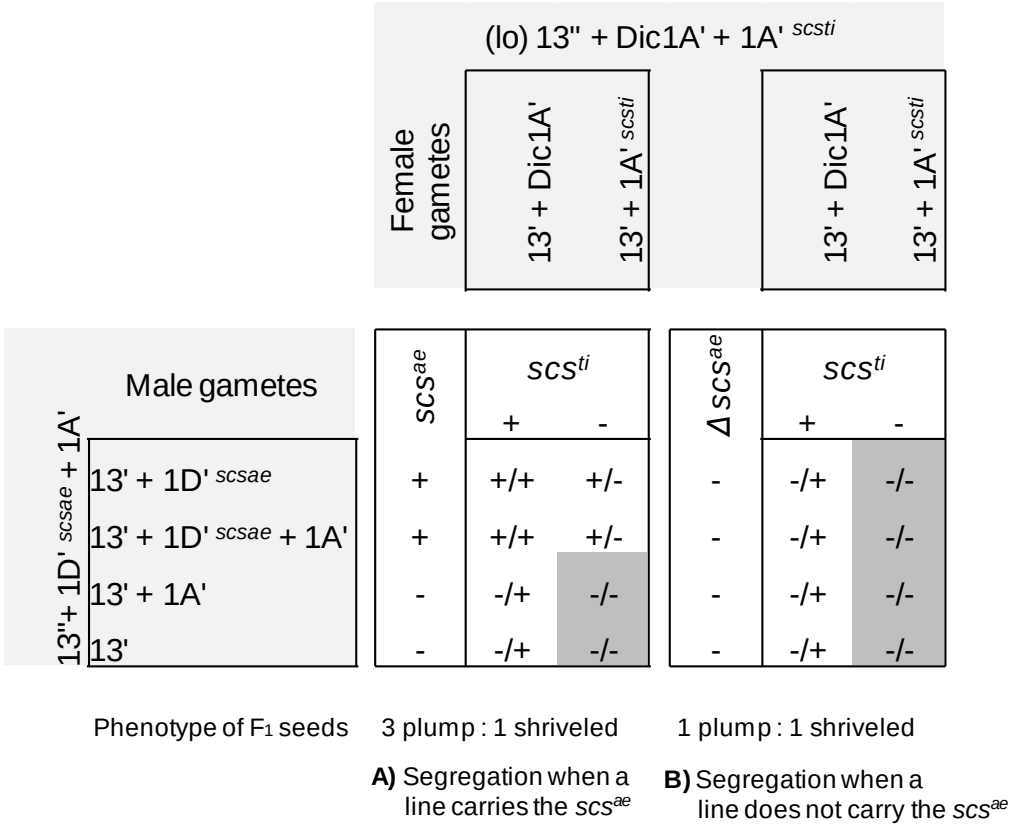


Figure 4.2. Phenotyping of radiation hybrid (RH) lines through test-crossing with a durum alloplasmic line, carrying one copy of *scs*$^{ti}$ and a cytoplasm of *Aegilops longissima* (*lo*). A) Test-crossed seeds from RH lines with the *scs*$^{ae}$ gene segregate quasi 3:1 plump to shriveled seeds. B) Test-crossed seeds from RH lines that have lost the the *scs*$^{ae}$ gene to radiation-induced deletion segregate 1:1 plump to shriveled.

## 4.3. Results

### 4.3.1. Development of 1D-specific molecular markers

Four methodologies were followed to enrich chromosome 1D of molecular markers, one based on retroelement junction sites, one employing known gene homoeologous sequence variants (HSVs), and two based on gene colinearity with three sequenced species. A total of 26 retrojunction (Ret) markers were designed on the basis of BAC end sequences originating from the library of *Ae. tauschii* accession AL8/78 (Li et al. 2004); among these eight (30.8%) could be confirmed as chromosome 1D-specific. All the 73 HSVs available at the time of the study from the "Haplotype Polymorphism in Polyploid Wheats and their Diploid Ancestors" (PI: J. Dvorak) project were converted into normal PCR primers combinations and searched for 1D-specificity; at the conditions employed in this study, only 20 (27.4%) could be confirmed as specific for the chromosome of interest. A total of 94 additonal primers were designed to span intron/exon junctions of wheat TCs orthologus to the collinear region of rice; 18 (~18%) could be confirmed as chromosome 1D-specific. In addition, 72 primers combinations were used to amplify homoeologous group 1 flow sorted chromosomes, 23 (32%) of these amplified chromosome 1D and were used for sequencing; for 18 of the sequenced genes was possible to identify at least one HSVs, and seven successfully were converted into 1D-specific primers, equal to the 10% of the genes investigated and the 39% of the HSVs identified. Overall, 53 1D-specific PCR markers were developed, to which were added three BAC sequence-based markers (RIOs) from *Ae. tauschii* accession AL8/78 BAC library (Seth 2009) as well as the UMN25 marker (Liu et al. 2008), for a total of 46 gene-based and 11 sequence-based markers.

### 4.3.2. A comprehensive radiation hybrid map for chromosome 1D

A population of 188 $RH_1$ lines was generated by gamma radiation followed by artificial pollination. A total of 94 lines were produced at a radiation dosage of 150 Gy, and an equal number of lines was obtained at 350 Gy dosage. This population was genotyped with the 57 PCR-based markers described above. Surprisingly, the 94 $RH_1$ lines produced with the lower radiation dosage were the most informative, with a retention frequency of 0.71, as compared to a frequency of 0.98 in the 350 Gy population. The 150

Gy panel provided a total of 32 lines with a breakage (analogous to a 'recombination event' in genetic mapping studies) along chromosome 1D, 50 lines with no apparent deletions, and 12 lines missing the entire 1D chromosome (Figure 4.3).

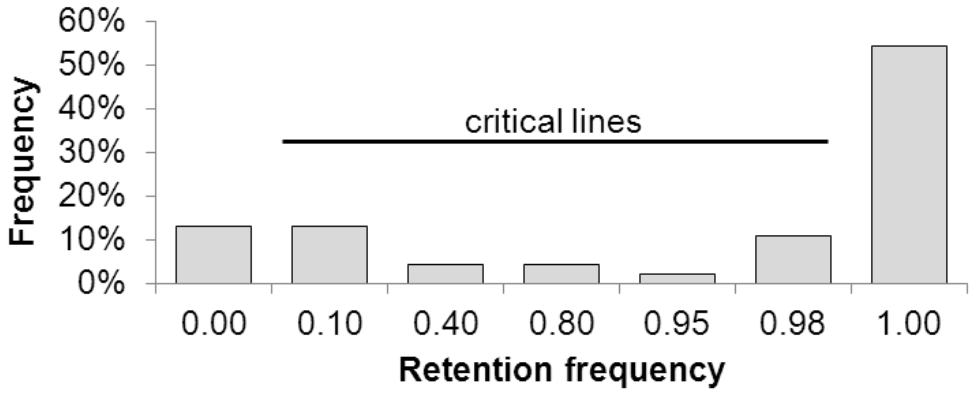


Figure 4.3. Retention frequency of 94 radiation hybrid lines population for chromosome 1D treated with 150 Gy of gamma ray.

The informative lines were used to generate a comprehensive map of chromosome 1D at LOD of 10. This map spans 205.2 cR as determined by 36 unique map loci (Figure 4.4). For 22 of the markers their deletion bin location is known, and the chromosomal location of the remaining 25 markers can be derived on the basis of their relative map location. Six markers span the telomeric portion of the short arm of chromosome 1D (1DS5-0.70-1.00), none was mapped from the middle portion of the 1DS (1DS3-0.70-0.41), eight map within the peri-centromeric and centromeric region (C-1DS3-0.48 or C-1DL2-0.41), all the remaining markers map on the long arm of chromosome 1D (1DL2-0.41-1.00). This map was strategically enriched with markers spanning the pericentromeric, middle, and distal portion of chromosome 1D because this was the known location of the *scs*$^{ae}$ gene (Hossain et al. 2004). The deletion bin corresponding to this region (1DL2-0.41-1.00) represents approximately the 37% of the whole 1D chromosome, equal to 225 Mb (Dolezel et al. 2009).The forty markers that map within, divide this region in 149 cR with an average resolution of 1.5 Mb per cR.

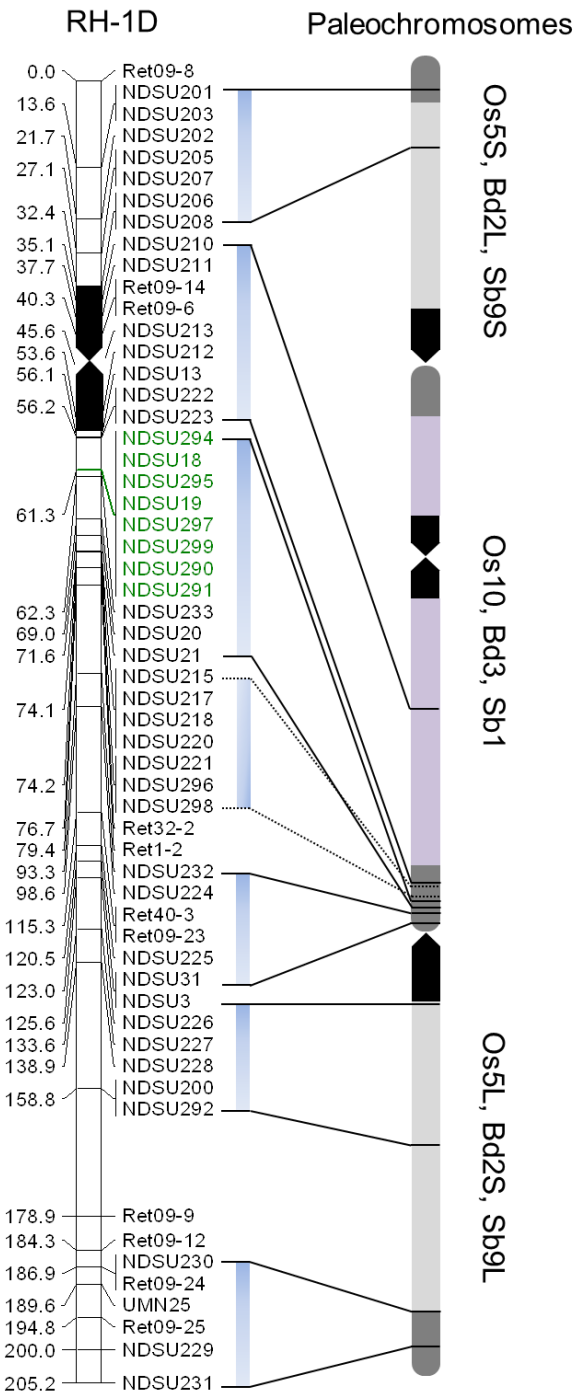


Figure 4.4. Radiation hybrid map of the chromosome 1D (RH-1D) and its ancestral relationship with rice (Os), *Brachypodium* (Bd), and sorghum (Sb). The locations of the peri- and centromeres are identified with black boxes, while the telomeres are indicated in dark grey. The blocks of synteny are shown with vertical blue bars and their corresponding collinear regions are pointed with black lines. The marker loci labeled in green cosegregate with the *scs*$^{ae}$ locus.

4.3.3. Colinearity between wheat 1D, rice, *Brachypodium*, and sorghum

A total of 46 gene-based markers were mapped in 1D-RH. The sequence of these markers was used as input for the 'Wheat Zapper' tool (http://wge.ndsu.nodak.edu/wheatzapper/) to identify a total of 15 orthologus genes on rice chromosome 5 (Os5), *Brachypodium* chromosome 2 (Bd2), and sorghum chromosome 9 (Sb9) (Figure 4.4), and 28 on rice chromosome 10 (Os10), 25 on *Brachypodium* chromosome 3 (Bd3), and 22 on sorghum chromosome 1 (Sb1). Interestingly, the markers collinear to Os5, Bd2, and Sb9 mapped to the distal regions of both arms of chromosome 1D, while Os10, Bd3 and Sb1 show synteny to the proximal regions. In particular, locus *Xndsu201* is orthologus to a gene located 0.2 Mb from the telomere repeat of Os5S. Locus *Xndsu208* is orthologous to a gene located on the short arm of Os5, 7 Mb distal of the rice centromere (International Rice Genome Sequencing Project, 2005), while *Xndsu210* corresponds to a gene on the long arm of Os10, 1.5 Mb distal of the rice centromere. The two marker loci are separated by just 2.7 cR. The pericentromeric region of the 1D chromosome is represented by the 21.1 cR interval *Xndsu210-Xndsu223*. All of the loci mapped within this region are orthologus to the middle or distal portion of Os10, Bd3, and Sb1, approximately 1.5 Mb to 10.5 Mb distal of the centromere of Os10. The cosegregating map loci *Xndsu31* and *Xndsu3* are orthologus to a gene located 1.7 Mb proximal of the telomerase repeat of Os10 and a gene 4.7 Mb distal of Os5 centromere, respectively. Finally, *Xndsu231* is orthologus to a rice gene placed 2.2 Mb distal of the telomerase repeat of Os5L. The distances from the chromosome landmarks described for rice Os5 – 10 are also valid for Sb1 - 9 (Paterson et al. 2009) and Bd2 - 3 (The International Brachypodium Initiative, 2010), with the obvious corrections due to the differences in their genome sizes. The order of the genes is well conserved among all the four species considered; however, microsynteny disruptions occur in the region corresponding to the peri-telomere of Os10, Bd3, and Sb1 (Figure 4.4 and Table 4.1). Practically, two syntenic blocks can be identified corresponding to *Xndsu294-Xndsu21* and *Xndsu215-Xndsu298* intervals. These marker loci tend to cosegregate into two distinguished groups 12.8 cR apart. Eleven genes belong to the most proximal of these two syntenic groups and seven to the most distal. Based on colinearity, five markers loci (*Xndsu215-221*) of the distal syntenic group should map above the proximal syntenic group, while the two remaining loci *Xndsu296*, and *Xndsu298* (Table 4.1) should map within the

proximal group based on their location on the genome of rice. The ten genes of the proximal syntenic group are all in the proper orthologus position, without any observable order disruption.

4.3.4. A refined location for *scs*$^{ae}$ by phenotyping $RH_2$ lines

The ability of *scs*$^{ae}$ to restore seed plumpness in alloplasmic (*lo*) lines is well documented (Hossain et al. 2004). To confirm that the activity of *scs*$^{ae}$ is not affected by the presence of its *T. timopheevii* counterpart *scs*$^{ti}$, six LDN 1D (1A) genotypes homozygous for *scs*$^{ae}$ were test-crossed to the alloplasmic line (*lo*) *scs*$^{ti-}$ to generate 134 exclusively plump seeds (shriveled seeds were not observed). This confirmed the ability of LDN 1D (1A) to restore proper seed plumpness to (*lo*) *scs*$^{ti-}$ genotypes. However, the primary scope of the work presented here is to determine the presence or absence of the *scs*$^{ae}$ locus in $RH_2$ lines, which are hemizygous or nullizygous for *scs*$^{ae}$, rather than homozyogous. Six double-monosomic $F_1$ lines [13" +1A' +1D'] with just one copy of *scs*$^{ae}$, and six LDN lines [13" + 1A"] nullisomic for *scs*$^{ae}$ were test-crossed to (*lo*) *scs*$^{ti-}$ to obtain ratios of 102:35 plump to shriveled seeds, and 81:85, respectively. $\chi^2$ analyses confirmed (*p:* 0.95) that the ratios fit the expected 3:1 and 1:1 pattern, respectively (Figure 4.2). Therefore, given that enough test-cross seeds are produced, the double-monosomic progenies of the $RH_1$ lines can be phenotyped for the presence or absence of one copy of the *scs*$^{ae}$ gene. Also, to our advantage, the molecular location of the *scs*$^{ae}$ locus was previously pinpointed to the proximal portion of the deletion bin 1DL2-0.41-1.00 (Kalavacharla et al. 2006). Seven $RH_1$ lines were identified as having a breakage on the long arm of 1D. Selfed seeds of these selected lines ($RH_2$) were used for test-crossing with (*lo*) *scs*$^{ti-}$, between 33 and 111 seeds were obtained for each line (Table 4.1). In all seven cases the plump to shriveled ratios fit a 1:1 segregation (*p*: 0.05), consistent with the nullizygous state of the *scs*$^{ae}$ gene (Figure 4.2). Subsequently, the $RH_2$ phenotyping data was converted into the deletiontype (a.k.a. 'haplotype' in genetic mapping studies) of *scs*$^{ae}$ and used to map this locus on the $RH_1$ lines (Table 4.1). In particular, three $RH_1$ lines (3027, 3079, and 3064) narrowed the gene location to a region distal of *Xndsu223* and proximal to *Xndsu233*. In fact, even if none of the seven lines discussed here (Table 4.1) exclude *Xndsu233* from the *scs*$^{ae}$ interval, previous studies showed that this marker-locus segregates away from the phenotype (Kalavacharla et al. 2006). The reduced interval

spans 6.1 cR equal to approximately 9.1 Mb as derived from the cR to Mb conversion ratio calculated above. Based on colinearity (Table 4.2), the refined *scs*$^{ae}$ region *Xndsu294-Xndsu299* corresponds to 0.1 Mb of Os10, 0.3 Mb of Bd3, and 0.2 Mb of Sb1. Since the chromosome 1D of wheat is 604 Mb in size (Dolezel et al. 2009), which is 11 times larger than Os5 and Os10 combined (International Rice genome Sequencing Project 2005), 4 times Bd2 plus Bd3 (The International *Brachypodium* Initiative 2010), and 5 times Sb1 plus Sb9 (Paterson et al. 2009), it can be estimated that the region in wheat corresponds to 1.1 ±1 Mb. This segment contains 16 genes annotated in one or more of the three sequenced grass model species (Table 4.2). Among these, ten genes have been mapped on the $RH_1$ lines, but four have been shown to segregate away from the *scs*$^{ae}$ region.

## 4.4. Discussion

Viable RH lines are ideal biological material for functional analyses and fine mapping studies in the complex polyploid genome of wheat, especially for those genes that lack adequate genetic diversity.

### 4.4.1. Markers development

The genotype scoring of RH is not based on allele polymorphism like in genetic mapping studies, but rather on the presence vs. absence scoring of marker loci. There are two steps involved in generating a molecular marker capable of discriminating allelic polymorphism in polyploids: first the marker needs to specifically tag only one of the homoeologous chromosomes, secondly it needs to discriminate the allelic diversity at that locus. While the first step provides discrimination among the homoeologs, which takes advantage in wheat of an average frequency of 1 HSV per 24 bp in the expressed portion of the genome, the second step is not as trivial since on average only one SNP per 540 bp discriminates alleles among wheat varieties (Somers et al. 2003). It is then evident the great advantage of RH for marker scoring, since only the first of the two steps is required to map markers using deletions. Although the efficiency of RJMs was lower (31%) in this study as compared to Paux et al. (2006), who reported the ISBP (retrojunction-based) markers efficiency to be 67.2%, it is still very successful considering the smaller number of primer pairs screened here (26) relatively to their study (58).

Table 4.1. Fine map location of the *scs*$^{ae}$ locus and its syntenic relationship to *Brachypodium* and rice

| | cR | 3057 | 3058 | 3017 | 3029 | 3027 | 3079 | 3064 | Brachy gene (Mb) | | Rice gene (Kb) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Plump:shriveled | | 26:30 | 26:24 | 25:25 | 29:31 | 24:28 | 16:17 | 54:57 | | | | |
| *Xndsu211* | 37.7 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | Bd3g27640 | 28. 6 | Os10g30600 | 15.8 |
| *Xndsu213* | 45.6 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | Bd3g29870 | 31.9 | Os10g35220 | 18.7 |
| *Xndsu212* | 53.6 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | Bd3g29840 | 31.9 | Os10g35150 | 18.7 |
| *Xndsu13* | 56.1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | Bd3g30850 | 33.1 | Os10g37330 | 19.9 |
| *Xndsu222* | 56.1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | Bd3g30890 | 33.2 | Os10g37420 | 20.0 |
| *Xndsu223* | 56.1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | Bd3g31110 | 33.3 | Os10g37480 | 20.0 |
| *Xndsu294* | 61.3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | Bd3g31370 | 33.5 | Os10g37630 | 20.1 |
| *Xndsu18* | 61.3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | Bd3g31410 | 33.5 | Os10g37660 | 20.1 |
| *Xndsu295* | 61.3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | Bd3g31430 | 33.6 | Os10g37690 | 20.1 |
| *Xndsu19* | 61.3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | Bd3g31460 | 33.6 | Os10g37730 | 20.1 |
| *Xscs*$^{ae}$ | 61.3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | | | |
| *Xndsu297* | 61.3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | . | . | Os10g37760 | 20.1 |
| *Xndsu299* | 61.3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | . | . | Os10g37899 | 20.2 |
| *Xndsu290* | 61.3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | Bd3g31630 | 38.9 | Os10g38050 | 20.3 |
| *Xndsu291* | 61.3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | Bd3g32000 | 34.2 | Os10g38950 | 20.7 |
| *Xndsu233* | 62.3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | Bd3g33240 | 35.6 | Os10g41030 | 22.0 |
| *Xndsu20* | 69.0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | Bd3g31520 | 33.8 | Os10g37830 | 20.1 |
| *Xndsu21* | 71.6 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | Bd3g31530 | 33.8 | Os10g37840 | 20.2 |

Table 4.2. Candidate genes within the *scs*$^{ae}$ interval (adapted from the Wheat Zapper output)

| Mapped loci | Distance to *scs*$^{ae}$ (cR) | Wheat EST | Rice gene (Mb) | | Brachy gene (Mb) | | Sorghum gene (Mb) | | Pfam function |
|---|---|---|---|---|---|---|---|---|---|
| *Xndsu294* | 0.0 | TC407174 | Os10g37630 | 20.1 | Bd3g31370 | 33.5 | Sb1g031330 | 53.9 | UBX-domain |
| | | CJ540389 | Os10g37640 | 20.1 | Bd3g31380 | 33.5 | Sb1g031320 | 53.9 | HIT-zinc finger |
| *Xndsu18* | 0.0 | CD885860 | Os10g37660 | 20.1 | Bd3g31410 | 33.5 | Sb1g031280 | 53.8 | trehalase precursor |
| *Xndsu295* | 0.0 | BJ254861 | Os10g37690 | 20.1 | Bd3g31430 | 33.6 | Sb1g031260 | 53.8 | *Oxa1* inner membrane |
| *Xndsu296* | 12.9 | CV768721 | Os10g37700 | 20.1 | Bd3g31440 | 33.6 | Sb1g031250 | 53.8 | hydrolase |
| | | TC440753 | Os10g37720 | 20.1 | Bd3g31450 | 33.6 | Sb1g031240 | 53.8 | alpha-beta hydrolase |
| *Xndsu19* | 0.0 | TC388432 | Os10g37730 | 20.1 | Bd3g31460 | 33.6 | Sb1g031230 | 53.8 | pollen ankyrin |
| | | TC406567 | Os10g37740 | 20.1 | Bd3g31480 | 33.6 | Sb1g031220 | 53.8 | CGMC_GSK.9 |
| *Xndsu297* | 0.0 | TC395424 | Os10g37760 | 20.1 | . | . | Sb1g031210 | 53.8 | rhomboid |
| | | TC412626 | Os10g37770 | 20.1 | . | . | . | . | dehydration response |
| *Xndsu20* | 6.7 | TC370675 | Os10g37830 | 20.2 | Bd3g31520 | 33.8 | . | . | OsFBX391 |
| *Xndsu21* | 10.3 | TC413544 | Os10g37840 | 20.2 | Bd3g31530 | 33.8 | Sb1g031180 | 53.8 | CCB2 |
| | | TC397464 | Os10g37850 | 20.2 | Bd3g31550 | 33.8 | Sb1g031170 | 53.7 | armadillo |
| | | TC431553 | Os10g37860 | 20.2 | Bd3g31560 | 33.8 | . | . | expressed protein |
| *Xndsu298* | 12.9 | TC444904 | Os10g37880 | 20.2 | Bd3g31570 | 33.8 | . | . | oxidoreductase |
| *Xndsu299* | 0.0 | TC444904 | Os10g37899 | 20.2 | . | . | . | . | transposon |

Additionally, in order to develop *T. aestivum* chromosome 3B specific retrojunction markers Paux et al. (2006) used sequences from *T. aestivum* BAC library, while here the markers were designed for chromosome 1D of *T. aestivum* from the BAC library of *Ae. tauschii* accession AL8/78. The sequence polymorphism that exists between 1D of *Ae. tauschii* and *T. aestivum* could explain the decreased efficiency observed for this marker system. Still, retrojunction markers provided the highest success rate among the marker systems tested. Also, their development required the least amount of lab work and the lowest experimental cost. Thus, RJM markers have confirmed to be the best approach to develop markers in wheat, especially when BAC end-sequencing information are available for their rapid *in silico* design.

However, for the purpose of fine mapping genes it is usually better to use EST-based markers, primarily because they provide an entry point to look at the synteny with model species and to determine possible candidate genes within a region. In this regard, the mere success scoring does not fully explain the best marker development system. Rather, the best methodology is the one that permits converting the most genes within the targeted region into scoreable markers. This was the case for the system that employed flow sorting and sequencing. Even though it is more time consuming and relatively more expensive, the 39% of the genes that presented HSVs could be effectively tagged with markers. The marker development strategy based on the mere design of primers spanning intron/exon junctions followed by screening on high-resolution gels revealed all its limits, with the lowest success ratio at just 18%. Conversely, it is worth mentioning that regardless of the methodology employed, it is extremely simpler and more time effective to develop HSV based markers for RH, rather than generating polymorphic markers on the basis of SNPs for genetic mapping.

### 4.4.2. A segregating RH population for chromosome 1D

A radiation hybrid population segregating for chromosome 1D was rapidly generated in just one season of greenhouse artificial crossing. A subset of 188 $RH_1$ lines was searched for deletions, and the lower radiation dosage (150 Gy) was found as the most effective to create random genomic deletions. It has been shown that a higher dosage of radiation typically causes a higher frequency of genome

aberrations (Hossain et al. 2004). However, in the specific strategy employed here to generate *in vivo* $RH_1$ lines for functional studies, the radiation mutants are not the final product. Instead, the mutagenized ($M_1$) lines are grown under normal greenhouse conditions and their filial ($RH_1$) generations are used for genotyping. In order to successfully transfer genomic deletions from the mutants to their double monosomic $RH_1$, the plants need to be sufficiently healthy to produce strong and viable progenies. It is speculated that higher dosages of radiation will inevitably result in weaker mutants with lower fitness, and lower likelihood of producing healthy progenies. Hence, the observed higher yield of critical lines from the 150 Gy treatment could be explained as a consequence of healthier mutant lines with good fitness, which still carried a sufficient number of genomic deletions. In the light of this observation, the kill curve analyses that are typically used to determine the ideal radiation dosage (Hossain et al. 2004; Al-Kaff et al. 2008; Wardrop et al. 2002; for review Riera-Lizarazu et al. 2008; Chapter 1) need to be slightly modified for *in vivo* RH project, selecting not only on the basis of radiation survival, but also for good vigor of the survivals.

The $RH_1$ panel of 150 Gy revealed 34% lines with a breakage along the chromosome 1D (Figure 4.3). This indicates that even a population of reduced size, as the one employed in this study, still provides good information for mapping. Another advantage of RH over genetic mapping studies is then evident: a population segregating and homozygous for one chromosome (1D in this case) can be rapidly generated in only one crossing season, and even small populations result in extremely useful mapping resources.

### 4.4.3. A comprehensive gene-based RH map of chromosome 1D reveals a neo-centromerization event

A subset of 32 segregating $RH_1$ lines provided sufficient information to subdivide chromosome 1D into 36 unique mapping loci, with an overall resolution of 1.5 Mb $cR^{-1}$. Unfortunately, this resolution is lower than what was previously reported for RH of chromosome 1D (Hossain et al. 2004; Kalavacharla et al. 2006), but similar to what was observed in some of the regions of the RH map for chromosome 3B (Chapter 2).The lower-than-expected resolution can be explained by the specific localization of the markers employed in this study. In fact, even though the RH map spans from telomere to telomere of

chromosome 1D, the markers are clearly not evenly distributed, with an obvious overabundance in the region corresponding to the *scs*$^{ae}$ location, and a complete lack of markers in the middle portion of 1DS (Figure 4.4). With the exclusion of the centromere, the short arm of 1D is here represented by just eight markers and 32.4 cR, only the 16% of the whole 1D RH map. On the basis of the colinearity to three sequenced grass species (rice, *Brachypodium*, and sorghum) and a previously reported map (Luo et al. 2009), a large segment of chromosome 1D is not represented in this RH map. Since the short arm of 1D is approximately 37% of the whole chromosome (Dolezel et al. 2009), and the peri-centromere of 1DS accounts for ~4% of the RH map (7.9 cR), it can be estimated that a portion corresponding to nearly 35 cR (17%) is missing from the map. The uneven distribution of the markers might then negatively affect the chromosome wise resolution of this RH map.

The wheat chromosome 1D has been proposed to have evolved from an ancestral fusion between two paleochromosomes corresponding to the modern Os5-10 and Sb9-1 (Luo et al. 2009). The data gathered here support this hypothesis, and extend it to Bd2-3, even though a large inversion occurred between the long and short arm of *Brachypodium* Bd2 (Figure 4.4). The precise position of the fusion point and the methodology of the fusion are still under discussion. However, Luo and colleagues (2009) in their elegant study were able to postulate that the paleochromosome corresponding to Os5, Sb9 and Bd2 likely split at the centromere, the repeated sequences of this region then became 'sticky' and fused with the telomeric repeats of a second paleochromosome, corresponding to Os10, Sb1, and Bd3. Finally, the centromere of this second paleochromosome maintained its function and became the centromere of chromosome 1D. The low markers resolution on the short arm of 1D does not allow confirming or denying the hypothesized centromeric to telomeric fusion point on the short arm, but the RH map fully supports this mechanism of fusion on the long arm of 1D. Further, the telomeric region of 1D matches the telomeres of the three sequenced species, with the only difference that 1DS correspond to actually Bd2L and 1DL corresponds to Bd2S. However, the RH data contrast with the hypothesis that the ancestral centromere (corresponding to the modern centromeres of Os10, Sb1, and Bd3) maintained its functionality. In fact, the six gene-based markers that map on the two sides of the 1D centromere, as inferred from deletion bins information, are all colinear to the long arm of Os10, Sb1, and Bd3, rather than

their centromeres (International Rice genome Sequencing Project 2005; The International *Brachypodium* Initiative 2010; Paterson et al. 2009). The study of the centromeric regions by mean of genetic mapping is complicated by the nearly complete lack of recombination events (Erayman et al. 2004), making it extremely challenging to adequately separate the map loci in this region by recombination based approaches. It is then possible that the recombination based study of Luo et al. (2009) failed to recognize that markers surrounding the centromere of 1D are actually distal of the paleocentromere because of a lack of map resolution in this region. Radiation hybrid mapping is only indirectly influenced by the low recombination frequency at the centromere (Chapter 2), and provides sufficient map resolution to determine that the centromere of 1D is probably the result of a neo-centromerization event that followed the ancestral fusion of the two paleochromosomes, rather than the conservation of a paleocentromere. Increasing evidences are gathering to support an epigenetic origin for the centromeres, rather than their sequence conservation (for revision Allshire and Karpen 2008). Hence, the formation of a centromere does not require the transfer of genomic sequences, but just the activation of epigenetic mechanisms (Wong and Choo 2001). New centromeres can then evolve very rapidly, and neo-centromerization is a common event that occurred many times throughout the evolution of the species (Ferreri et al. 2005). Also, the in situ hybridization study conducted by Qi et al. (2010) disputes the 1D centromere location proposed by Luo et al. (2009), instead suggesting that the wheat centromere actually corresponds to the ancestral location of the centromere of Os5 and Bd2. The precise location of the 1D centromere remains controversial and probably only the complete sequencing of this wheat chromosome will provide the ultimate clue on its origin.

### 4.4.4. The *scs* genes might have orchestrated the evolution of wheat and other grasses

The *scs* genes have a drastic effect on alloplasmic lines (Figure 4.1). However, their effect is not limited to just these unique cytoplasm substitution materials. Maan (1991a) hypothesized that the compatibility between each *Triticeae* nucleus and its cytoplasm is mediated by a *scs* gene located on a long arm of group 1 chromosomes. Therefore, the success of introducing a nucleus of one species to a cytoplasm of another species depends on the type of the *scs* gene present in the male nucleus.

Hexaploid wheat likely carries three different *scs* genes, one on each genome (1A $scs^{ti}$, 1D $scs^{ae}$), with the 1B version probably responsible for the wild type NCI, since both the wheat cytoplasm (plasmon B) and B genome originated from *Ae. speltoides* ssp. *ligustica* (SS; plasmon S) (Tsunewaki et al. 2009; Kilian et al. 2007; Haider 2012). Evidence of the existence of a 1B version of the gene were provided by Asakura et al. (2000), who were able to demonstrate that the $scs^{ti}$ gene on chromosome 1A of *T. timopheevii* has a homeoallele on chromosome 1G, as determined in alloplasmic lines of *Ae. tauschii*. Since the G genome of *T. timopheevii* is also derived from the *Ae. speltoides* species (ssp. *speltoides*) and shares similarities with the B genome of wheat (Kilian et al. 2007), it can be concluded that all three homeologus carry homeoalleles of the $scs^{ae}$ gene. Additionally, Taketa and colleagues (2002) suggested the existence of a version of *scs* on the long arm of chromosome 1H of *Hordeum vulagre*, indicating that other species beside wheat might carry a *scs* gene.

The *scs* genes on chromosomes 1A and 1D ensure mating compatibility when combined with the cytoplasm of *T. urartu* Tumanian ex Gandylian (plasmon A) and *Ae. tauschii* (plasmon D) (Tsunewaki 1980). Interestingly, the $scs^{ti}$ version of 1A and its homeoallele on 1G of *T. timopheevii* (AAGG) provide compatibility with the cytoplasm of *Ae. longissima* ($S^1S^1$; Plasmon $S^1$) and *Ae. tauschii* (plasmon D), while the 1A or 1B versions of *T. turgidum* (AABB) do not. The case of 1A from *T. turgidum* and *T. timopheevii* resemble what is observed for 1B and 1G. In that case it was determined that two subspecies of *Ae. speltoides* donated their genomes and plasmons for the 1B and 1G chromosomes. Similarly, it can be concluded that two subspecies or at least lineages of *T. urartu* donated the A genome to *T. turgidum* and *T. timopheevii*, each harboring a different *scs* allele (Asakura et al. 2000).

Overall, these observations advocate the existence of an ancestral *scs* version which subsequently underwent mutation(s) during speciation and co-evolved with various cytoplasms. “Maan *scs* hypothesis” (1991a) postulates that compatible interactions between nucleus and cytoplasm are pre-established in diploids, and that these established interactions guide the possible mating combinations, which would then allow only some type of allopolyploids to be originated. Consequently, allopolyploids evolved by hybridization of different diploids, with the *scs* genes acting as an unfair mating referee,

limiting the female gametes possibilities while leaving the males relatively free to hybridize, and ultimately directing the speciation of the whole *Triticeae* tribe.

4.4.5. Fine mapping of the *scs*$^{ae}$ gene via phenotyping of $RH_2$ lines

Radiation hybrids are a well-established methodology for mapping, and have been vastly applied in genomic studies (Faraut 2009). However, the RH advantages of higher map resolution, speed of population development, and use of monomorphic markers have rarely been combined with functional studies. To the best of our knowledge only radiation mutants have been used in functional studies in plants (Al-Kaff et al. 2008; Spielmeyer et al. 2008; Rasko et al. 2000; Kynast et al. 2002), but never a $RH_2$ population for forward genetic approaches. From a Mendelian genetic prospective, RH are ideal population to study the segregation of a single gene. In fact, the nearly isogeneic parents (LDN and LDN 1D (1A) in this case) that are used to generate the $RH_1$ lines prevent any heterotic segregation on any chromosomes, except the one under study. Moreover, homozygousity in monosomic state is fixed in just one crossing season, preventing the miss-scoring often associated with the heterozygous state, and saving 5-6 seasons for the development of recombinant inbred lines. Finally, phenotyping is done on the basis of the presence or absence (knock-out) of a gene, theoretically simplifying the visual scoring (see Chapter 1 for details).

Phenotyping for the presence or absence of the *scs*$^{ae}$ gene was performed on $RH_2$ lines (Figure 4.2), after a RH map was generated on their $RH_1$ parent (Table 4.1). Seven informative lines were selected on the basis of prior knowledge of the *scs*$^{ae}$ location (Kalavacharla et al. 2006), and their phenotyping narrowed down the location of this gene to an interval of 6.2 cR, calculated to be just 1.1 ± 1 Mb in size on the basis of physical conversion from three sequenced grass species (rice, *Brachypodium*, and sorghum). The *scs*$^{ae}$ interval is located just above the 1DL fusion point between two paleochromosomes, and corresponds to the peritelomeric region of the paleochromosome corresponding to Os10, Bd3, and Sb1. This region is in close proximity of the centromere of 1D, and as such it can be expected to be particularly rich in transposable elements (Neumann et al. 2011). However, the peculiar peri-telomeric origin of this portion of chromosome 1D causes the *scs*$^{ae}$ interval to be particularly poor in

transposable elements, but abundant in other genes. Devos (2010) advocated that in transposable poor regions the gene conservation between grasses is high, but the genomic order of the genes is more prone to rearrangements (Murat et al. 2010). Indeed, colinearity is surprisingly well conserved throughout the chromosome, but some erosion occurs in proximity of the *scs*$^{ae}$ interval. Two separate syntenic blocks span the RH map between position 56.2 cR and 74.2 cR. These two syntenic groups alternate regions of precise gene order conservation, with gaps of disruption or inversion. It is plausible that, given the vicinity of the point of paleofusion, the observed small inversions and gene order rearrangement are the consequence of the merging of the two ancient chromosomes. Still, all the genes predicted to be in this region based on orthologus relationships map exactly where predicted, or in close proximity, making the synteny an ideal tool for marker saturation of this interval. Possibly, the evolutionary importance of the *scs* gene acts as a stabilizing factor to maintain this region relatively unchanged in different species. There is no available evidence yet describing a gene on Os10, Bd3, and Sb1 involved in the speciation of these three grass species, but expanding the 'Maan *scs* hypothesis' it would be plausible that the *scs* genes are not restricted to just the *Triticeae* (see case of 1H of barley Taketa et al. 2002). Since the *scs*$^{ae}$ interval originates from one of the five paleochromosomes that later evolved into the whole grass family (Bolot et al. 2009), and this genomic portion is very well conserved in the three grass model species considered here, it cannot be excluded that the *scs* genes are indeed involved in the speciation of all grasses. The development and characterization of alloplasmic material from other grasses would help in determining the correctness of this hypothesis.

### 4.4.6. Candidate genes for the *scs*$^{ae}$ locus

There are 16 orthologus genes within the *scs*$^{ae}$ interval predicted by one or more of the three sequenced species available in the 'Wheat Zapper' tool (Table 4.2). Four of these genes do not follow the genomic order predicted by synteny and map outside the *scs*$^{ae}$ interval, six have been mapped and cosegregated with the *scs*$^{ae}$ locus, and the remaining six have yet to be mapped. Even though the synteny is not perfectly conserved in this region, the orthologus information have proven very precise with 60% of the genes mapping exactly where predicted, and all loci were maximum 13 cR away from their

expected positions. There is a possibility that the *scs*$^{ae}$ gene does not appear in the collinear region of rice, *Brachypodium* or sorghum; however, the suggested extension of the 'Maan *scs* hypothesis' to the whole *Poaceae* family would indicate otherwise. It is then interesting to discuss the biological functions of these genes in the light of the known ability of the *scs* loci to restore compatibility between the nucleus and a specific set of cytoplasms. Since the majority of the cytoplasm functions and organelles are directly encoded by the nucleus (Nott et al. 2006), only those organelles that possess genetic information independent from the nuclear genome appear as the causative agents in differentiating the plasmon types. In plants, only two organelles have genomes not encoded by the nucleus: the mitochondria and the chloroplast. However, the chloroplast shows significantly less structural diversity as compared to the mitochondria (Burger et al. 2003), and are thus not expected to change extensively between the plasmons of wheat relatives. Further, the most studied example of incompatible NCI is 'cytoplasmic male sterility' (Hanson 1991), and in all of the cases to date the gene responsible for causing the sterility was of mitochondrial origin (Heazlewood et al. 2003; Liu et al. 2001; Akagi et al. 2004; Saha et al. 2007; Klein et al. 2005). Therefore, genes with an activity related to the mitochondria have been suggested as the responsible in determining the specificity of NCI. It can be then hypothesized that the *scs* genes are directly or indirectly involved in mitochondria activities, with specific *scs* genes interacting with specific mitochondria types. All of the 12 genes cosegregating or potentially cosegregating with the *scs*$^{ae}$ locus (6 mapped and 6 unmapped) target the mitochondria directly or indirectly, and at least eight have a predicted function that fits their role as candidates for the *scs*$^{ae}$ locus.

Among the mapped loci, *Xndsu294* contains the UBX-domain that targets protein degradation through the Cdc48/VCP/p97 chaperon ubiquitination machinery (Romisch 2006), which dynamically localizes to the mitochondria and endoplasmic reticulum (ER). The UBX-containing genes family participates in an array of cellular functions, but their known involvement in ER-associated degradation (Wang and Lee 2012) could potentially explain the occurrence of shriveled and non-viable seeds (Figure 4.1).

Another locus, *Xndsu295*, is derived from a wheat gene homolog of yeast *oxa1*, which is known to be involved in the assembly of the cytochrome oxidase (COX) complex and in its interaction with the

mitochondrial ribosomes (Jia et al. 2003). Yeast knockout mutants for *oxa1* show decreased enzymatic activity of the ATP synthase mitochondria complex, and typically do not survive outside specific growth media (Jia et al. 2007). The *oxa1* known biological functions fit well what is observed in the *scs*$^{ae}$ deficient alloplasmic lines, making this gene a good candidate.

*Xndsu19* locus is homologous to a pollen ankyrin gene. Ankyrins are characterized by a widely occurring repeat motif of 33 amino acids, which is thought to be involved in protein-to-protein interactions (Garcion et al. 2006). More interestingly, these genes have been demonstrated to be involved in embryo development through the recognition between female and male gametes (Yu et al. 2010). Further, this gamete recognition is mediated through the interaction between ankyrins and the miotchondrial transcription initiator factor SIG5. The role of mating referee proposed for the *scs*$^{ae}$ gene would be well described by this ankyrin-containing gene.

*Xndsu297* locus maps within the *scs*$^{ae}$ interval and is homologous to a drosophila *rhomboid protease*. This large gene family comprised of intramembrane serine peptidases is involved in many biological activities such as cell signaling, apoptosis, and mitochondria integrity (for review see Knopf and Adam 2012). Their established function in signaling between the nucleus and the mitochondria in patients affected by the Parkinson's disease (Whitworth et al. 2008), points out a role of this gene in controlling correct interaction between the nucleus and the cytoplasm.

Additionally to these four mapped candidate genes, other four genes should be added because of being predicted within the *scs*$^{ae}$ interval on the basis of synteny alone. The HIT-zinc finger domain is a functional component of many proteins. In particular, *aprataxin* is a DNA strand-break repair protein containing the HIT-zinc finger domain that was shown to preserve the mitochondria function in humans (Sykora et al. 2011). The ability of this nuclear gene to regulate the mitochondria activity supports its candidacy for the *scs*$^{ae}$ function.

Protein kinases promote a variety of activities in plants through peptide phosphorylation. In particular, CGMC_GSK.9 localizes to the mitochondria where is involved in signal transduction and metabolic processes (Gene Ontology: http://www.gabipd.org/database/cgi-bin/GreenCards.pl.cgi?Mode=ShowBioObject&BioObjectId=2989295). While the information available for

this gene do not support a specific activity that would well represent the unique *scs*$^{ae}$ functionality, the wide array of effects associated with protein kinase and its specific targeting to the mitochondria certainly do not allow discarding this gene as a valid candidate for the *scs*$^{ae}$ locus.

Similar conclusions can be reached for the dehydration response gene, since evidences indicate that seed dehydration is a nuclear directed and mitochondria driven activity (Song et al. 2009), but there are no direct evidences to support the involvement of this specific gene in the process. Still, shriveled seeds (Figure 4.1) can be seen physiologically as over-dehydrated endosperms, and hence the *scs*$^{ae}$ gene might actually be involved in the dehydration process.

Finally, SARM is a TLR adaptor that contains an armadillo motif (Pannerseelvam et al. 2012). This nuclear protein targets the mitochondria where the armadillo domain helps in controlling cell apoptosis and acts as an organellar chaperon for other peptides. The specific role of this protein in the interaction between nucleus and cytoplasm is unclear, but its ability to translocate nuclear encoded peptides to the mitochondrial lumen can be hypothesized to cause some of the biological activities associated with the *scs*$^{ae}$ gene.

Considering the biological functions of the eight orthologus genes described above it is not possible to discard any of them as valid candidates for the *scs*$^{ae}$ locus. Since all genes within this segment target the mitochondria, the entire interval is probably derived from an ancestral fusion of the mitochondria genome into the nucleus, a fairly common event in eukaryotic genomes (Leister 2005). Also, the abundance of good candidate genes in close proximity to each other raises the question of whether the *scs*$^{ae}$ locus is indeed a single locus, or rather a group of genes that exist in very close proximity. Unfortunately, the current 1D-RH map resolution is not sufficient to discard any hypothesis or any of the candidate genes.

## 4.5. Conclusions

The use of RH as biological material for forward genetic studies has proven successful. Monomorphic markers were easily developed and genotyped on the basis of deletion vs. retention. The development of a ‘Mendelized’ and homozygous RH population required just one crossing season. The

phenotyping for the presence or absence of the *scs*$^{ae}$ gene was conducted using $RH_2$ lines, meaning that any of the already existing $RH_1$ resources could be employed in the future for functional studies of other genes, provided that an adequate test-crossing scheme is devised. Good map resolution was observed even in the close proximity of the centromere. At the current status, the *scs*$^{ae}$ region has been narrowed down to a segment ~1.1 Mb in size, which likely contains 12 mitochondria-derived genes, eight of which have cellular functions that match what is considered to be the *scs*$^{ae}$ biological activity. It is not possible to exclude the possibility that the *scs*$^{ae}$ is not actually a single locus, but rather a group of mitochondrial genes tightly packed on chromosome 1D. Still, now that a RH methodology for fine mapping is in place, increasing the population size should be sufficient to further refine the location of this locus. The *scs* genes have likely had a major impact in the speciation of the *Triticeae* tribe, and possibly of all the *Poaceae* family. Their cloning is expected to help unravel the complex evolutionary history of the land grasses by compatible plasmon mating. Also, the *scs* genes hold the key to unlock the secondary and tertiary gene pools for breeding application, making their cloning not just a mere academic exercise, but rather a very applied aspect of molecular breeding.

## 4.6. References


Akagi H, Nakamura A, Yokozeki-Misono Y, Inagaki A, Takahashi H et al (2004) Positional cloning of the rice Rf-1 gene, a restorer of BT-type cytoplasmic male sterility that encodes a mitochondria-targeting PPR protein. Theor Appl Genet 108:1449-1457

Al-Kaff N, Knight E, Bertin I, Foote T, Hart N et al (2008) Detailed dissection of the chromosomal region containing the Ph1 locus in wheat Triticum aestivum: with deletion mutants and expression profiling. Ann Bot 101:863–872

Allshire RC, Karpen GH (2008) Epigenetic regulation of centromeric chromatin: old dogs, new tricks? Nature Reviews Genetics 9:923-937

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403-410

Amarnath D, Choi I, Moawad AR, Wakayama T, Campbell KH (2011) Nuclear-cytoplasmic incompatibility and inefficient development of pig-mouse cytoplasmic hybrid embryos. Reproduction 142:295-307

Asakura N, Nakamura Ch, Ohtsuka I (1997) RAPD markers linked to the nuclear gene from Triticum timopheevii that confers compatibility with Aegilops tauschii cytoplasm on alloplasmic durum wheat. Genome 40:201-210

Asakura N, Nakamura Ch, Ohtsuka I (2000) Homoeoallelic gene Ncc-tmp of Triticum timopheevii conferring compatibility with the cytoplasm of Aegilops tauschii in the tetraploid wheat nuclear background. Genome 43:503-511

Bogdanova VS (2007) Inheritance of organelle DNA markers in a pea cross associated with nuclear-cytoplasmic incompatibility. Theor Appl Genet 114:333–339

Bolot S, Abrouk M, Masood-Quraishi U, Stein N, Messing J, et al (2009) The 'inner circle' of the cereal genomes. Curr Opin Plant Biol 12:119-25

Burger G, Gray MW, Lang BF (2003) Mitochondrial genomes: anything goes. TRENDS in Genetics 19(12):709-716

Cisar G, Cooper DB (2002) Hybrid wheat. In: Curtis BC, Rajaram S, Goméz H Macpherson (eds) Bread wheat: improvement and production, plant production and protection series no. 30. FAO, Rome, pp 317–330

Chase CD (2007) Cytoplasmic male sterility: a window to the world of plant mitochondrial–nuclear interactions. Trends in Genetics 23:81-90

Devos KM (2010) Grass genomes organization and evolution. Current Opinion in Plant Biology 13:139-145

de Givry S, Bouchez M, Chabrier P, Milan D, Schiex T (2005) Carthagene: multipopulation intergrated genetic and radiated hybrid mapping. Bioinformatics 21:1703-1704

Dolezel J, Simkova H, Kubalakova M, Safar J, Suchankova P, et al (2009) Chromosome genomics in the Triticeae. In: Feuillet C, Muehlbauer GJ (eds) Genetics and Genomics of the Triticeae. Springer, New York, pp 285-316

Erayman M, Sandhu D, Sidhu D, Dilbirligi M, Baenziger PS, Gill KS (2004) Demarcating the gene-rich regions of the wheat genome. Nucleic Acids Res 32:3546–3565

Faraut T (2009) Contribution of radiation hybrids to genome mapping in domestic animals. Cytogenet Genome Res 126:21–33

Ferreri GC, Liscinsky DM, Mack JA, Eldridge MD, O'Neill RJ (2005) Retention of latent centromeres in the Mammalian genome. J Hered 96:217-24

Garcion C, Guilleminot J, Kroj T, Parcy F, Giraudat J, Devic M (2006) AKRP and EMB506 are two ankyrin repeat proteins essential for plastid differentiation and plant development in Arabidopsis. Plant J 48:895–906

Gehlhar SB, Simons KJ, Maan SS, Kianian SF (2005) Genetic analysis of the species cytoplasm specific gene ($scs^d$) derived from durum wheat. J Heredity 96(4):404-409

Gill BS (1991) Nucleo-cytoplasmic interaction (NCI) hypothesis of genome evolution and speciation in polyploid plants. In: Sasakuma T, Kinoshita T (eds.) Proceedings of the Kihara Memorial International Symposium on Cytoplasmic Engineering in Wheat, Yokohoma, Japan, pp 48–53

Gish WR (2006) WU-BLAST archives. http://blast.wustl.edu/ [Verified on-line December 2012]

Goss SJ and Harris H (1975) New method for mapping genes in human chromosomes. Nature 255:680-684

Haider N (2012) Evidence for the origin of the B genome of bread wheat based on chloroplast DNA. Turk J Agric For 36:13-25

Hanson MR (1991) Plant mitochondrial mutations and male sterility. Annu Rev Genet 25:461-486.

Heazlewood JL, Whelan J, Millar AH (2003) The products of the mitochondrial orf25 and orfB genes are FO components in the plant F1FO ATP synthase. FEBS Lett. 540:201-205.

Hossain KG, Riera-Lizarazu O, Kalavacharla V, Vales MI, Maan SS et al (2004) Radiation hybrid mapping of the species cytoplasm-specific (scsae) gene in wheat. Genetics 168: 415-423

Ivanova NV, Dewaard JR, Hebert PDN (2006) An inexpensive, automation-friendly protocol for recovering high-quality DNA. Molec Ecology Notes 6:998–1002

Jia L, Dienhart M, Schramp M, McCauley M, Hell K, Stuart RA (2003) Yeast Oxa1 interacts with mitochondrial ribosomes: the importance of the C-terminal region of Oxa1. EMBO J 22:6438–6447

Jia L, Dienhart MK, Stuart RA (2007) Oxa1 directly interacts with Atp9 and mediates its assembly into the mitochondrial F1Fo-ATP synthase complex. Mol Biol Cell 18:1897–1908

Joppa LR (1993) Chromosome engineering in tetraploid wheat. Crop Sci 33:908–913

Kalavacharla V, Hossain K, Gu YQ, Riera-Lizarazu O, Vales MI, Bhamidimarri S et al (2006) High-resolution radiation hybrid map of wheat chromosome 1D. Genetics 173:1089-1099

Kilian B, Özkan H, Deusch O, Effgen S, Brandolini A, Kohl J, Martin W, Salamini F (2007) Independent wheat B and G genome origins in outcrossing Aegilops progenitor haplotypes. Mol Biol Evol 24:217-227

Klein RR, Klein PE, Mullet JE, Minx P, Rooney WL, Schertz KF (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. Theor Appl Genet 111: 994-1012

Knopf RR, Adam Z (2012) Rhomboid proteases in plants – still in square one? Physiologia Plantarum 145:41-51

Kosambi DD (1944) The estimation of map distance from recombination values. Ann Eugen 12:172-175

Kynast R, Okagaki R, Rines H, Phillips R (2002) Maize individualized chromosome and derived radiation hybrid lines and their use in functional genomics. Funct Integr Genomics 2:60-69

Lander ES, Green P, Abrahamson J, Barlow A, Daly M et al (1987) MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics 1:174-81

Leister D (2005) Origin, evolution and genetic effects of nuclear insertions of organelle DNA. Trends Genet 21:655-63

Li W, Zhang P, Fellers JP, Friebe B, Gill BS (2004) Sequence composition, organization, and evolution of the core Triticeae genome. Plant J 40:500-511

Liu S, Chao S, Anderson JA (2008) New DNA markers for high molecular weight glutenin subunits in wheat. Theor Appl Genet 118:177-183.

Luo MC, Deal KR, Akhunov ED, Akhunova AR, Anderson OD et al (2009) Genome comparisons reveal a dominant mechanism of chromosome number reduction in grasses and accelerated genome evolution in Triticeae. Proc Natl Acad Sci USA 106:15780–15785

Maan SS (1973) Cytoplasmic variability in Triticinae. Proc IV Int Wheat Genet Symp pp. 367-373

Maan SS (1978) Cytoplasmic relationships among the D- and M-genome Aegilops species. In: Ramanujam (ed.) Proc V Int Wheat Genet Symp pp. 232-250, New Delhi, India

Maan SS (1983) Differential nucleo-cytoplasmic interactions involving Aegilops longissima cytoplasm and nuclei of emmer and common wheat. Crop Sci 23:990-995

Maan SS (1991) Transfer of species cytoplasm-specific (scs) gene of T. timopheevii Zhuk. to T. turgidum L. genome. Genome 35:238-243

Maan SS (1991a) Nucleo-cytoplasmic genetics of wheat. In: Sasakuma T, Kinoshita T (ed.) Proceedings of the international symposium on nuclear and organeller genetics of wheat, Hokkaido University, Sapporo, Japan, pp 75-l94

Maan SS (1992a) Genetic analyses of male fertility restoration in wheat: V anomalous results of a monosomic analysis. Crop Sci 32:28-35

Maan SS (1992b) A gene for embryo-endosperm compatibility and seed viability in alloplasmic Triticum turgidum. Genome 35:772-779

Maan SS (1992c) Transfer of the species specific cytoplasm (scs) from Triticum timopheevii to T. turgidum. Genome 35:238-243

Maan SS (1992d) The scs and Vi genes correct a syndrome of cytoplasmic effects in alloplasmic durum wheat. Genome 35: 780-787

Maan SS, Joppa LR, Kianian SF (1999) Linkage between the centromere and a gene producing nucleocytoplasmic compatibility in durum wheat. Crop Sci 39:1044–1048

Martin A, Simpfendorfer S, Hare RA, Eberhard FS, Sutherland MW (2011) Retention of D genome chromosomes in pentaploid wheat crosses. Heredity 107:315-319

Michalak MK (2009) Radiation hybrid mapping – an approach for mapping scs$^{ae}$ gene and collinear studies. PhD Thesis Dissertation, North Dakota State University, Fargo

Michalak MK, Ghavami F, Lazo GR, Gu YQ, Kianian SF (2009) Evolutionary relationship of nuclear genes encoding mitochondrial proteins across four grass species and Arabidopsis thaliana. Maydica 54:471-483

McIntosh RA, Yamazaki Y, Dubcovsky J, Rogers J, Morris C, Somers DJ, Appels R, Devos KM (2008) Catalogue of gene symbols for wheat. In: 11th International Wheat Genetics Symposium 24-29 August 2008 Brisbane, Qld Australia

Murat F, Xu J-H, Tannier E, Abrouk M, Guilhot N, et al (2010) Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution. Genome Research 20:1545-1557

Neumann P, Navrátilová A, Koblížková A, Kejnovský E, Hřibová E, Hobza R, Widmer A, Doležel J, Macas J (2011) Plant centromeric retrotransposons: a structural and cytogenetic perspective. Mobile DNA 2:4

Nott A, Jung H-S, Koussevitzky S, Chory J (2006) Plastid-to-nucleus retrograde signaling. Annu Rev Plant Biol 57:739–59

Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M et al (2007) The TIGR Rice Genome Annotation Resource: improvements and new features. Nucleic Acids Res 35:D883–D887

Panneerselvam P, Singh LP, Ho B, Chen J, Ding JL (2012) Targeting of pro-apoptotic TLR adaptor SARM to mitochondria: definition of the critical region and residues in the signal sequence. Biochem J 442:263-271

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, et al (2009) The Sorghum bicolor genome and the diversification of grasses. Nature 457:551-556

Paux E, Roger D, Badaeva E, Gay G, Bernard M, et al (2006) Characterizing the composition and evolution of homoeologous genomes in hexaploid wheat through BAC-end sequencing on chromosome 3B. Plant J 48:463–474

Peng JH, Zadeh H, Lazo GR, Gustafson JP, Chao S et al (2004) Chromosome bin map of expressed sequence tags in homoeologous group 1 of hexaploid wheat and homoeology with rice and Arabidopsis. Genetics 168:609-23

Rasko JEJ, Battini J-L, Kruglyak L, Cox DR, Miller AD (2000) Precise gene localization by phenotypic assay of radiation hybrid cells. PNAS 97: 7388-7392

Riera-Lizarazu O, Vales MI, Kianian SF (2008) Radiation hybrid (RH) and HAPPY mapping in plants. Cytogen Genome Res 120:3-4

Rieseberg LH, Blackman BK (2010) Speciation genes in plants. Annals of Botany 106:439–455

Rieseberg LH, Willis JH (2007) Plant speciation. Science 317: 910-914

Romisch K (2006) Cdc48p is UBX-linked to ER ubiquitin ligases. Trends Biochem Sci 31:24-25

Rőder MS, Korzun V, Wendehake K, Plaschke J, Tixier M-H, et al (1998) A microsatellite map of wheat. Genetics 149:2007-2023

Qi L, Friebe B, Zhang P, Gill BS (2009) A molecular-cytogenetic method for locating genes to pericentromeric regions facilitates a genomewide comparison of synteny between the centromeric regions of wheat and rice. Genetics 183:1235-1247

Qi L, Friebe B, Wu J, Gu YQ, Qian C et al (2010) The compact *Brachypodium* genome conserves centromeric regions of a common ancestor with wheat and rice. Funct and Integr Genomics 10:477-492

Quackenbush J, Cho J, Lee D, Liang F, Holt I, et al (2001) The TIGR gene indices: analysis of gene transcript sequences in highly sampled eukaryotic species. Nucleic acids Res 29:159-164

Saha D, Prasad AM, Srinivasan R (2007) Pentatricopeptide repeat proteins and their emerging roles in plants. Plant Phys Biochem 45:521-534

Salamov AA, Solovyev VV (2000) Ab initio gene finding in Drosophila genomic DNA. Genome Res 2000 10:516-522

Sasakuma T, Maan SS (1978) EMS-induced male-sterile mutants in euplasmic and alloplasmic common wheat Crop Sci 18:850-853

Schnurbusch T, Collins NC, Eastwood RF, Sutton T, Je Veries SP, Langridge P (2007) Fine mapping and targeted SNP survey using rice-wheat gene colinearity in the region of the Bo1 boron toxicity tolerance locus of bread wheat. Theor Appl Genet 115:451–461

Seth K (2009) High resolution mapping of the $scs^{ti}$ gene in durum wheat and conserved colinearity across three grass genomes: wheat, rice and *Brachypodium*. PhD Thesis Dissertation, North Dakota State University, Fargo

Simons KJ (2001) Fine structure mapping of the species cytoplasm specific gene in durum wheat. MSc Thesis Dissertation, North Dakota State University, Fargo

Simons KJ, Gelhar SB, Maan SS, Kianian SF (2003) Detailed mapping of the species cytoplasm-specific (scs) gene in durum wheat. Genetics 165:2129-2136

Somers DJ, Kirkpatrick R, Moniwa M, Walsh W (2003) Mining single-nucleotide polymorphisms from hexaploid wheat ESTs. Genome 49:431–437

Song SQ, Tian MH, Kan J, Cheng HY (2009) The response difference of mitochondria in recalcitrant Antiaris toxicaria axes and orthodox Zea mays embryos to dehydration injury. J Integr Plant Biol 51:646-53

Sorrells ME, La Rota M, Bermudez-Kandianis CE, Greene RA, Kantety R et al (2003) Comparative DNA sequence analysis of wheat and rice genomes. Gen Res 13:1818-1827

Spielmeyer W, Singh RP, McFadden H, Wellings CR, Huerta-Espino J et al (2008) Fine scale genetic and physical mapping using interstitial deletion mutants of Lr34/Yr18: a disease resistance locus effective against multiple pathogens in wheat. Theor Appl Genet 116:481–490

Sykora P, Croteau DL, Bohr VA, Wilson III DM (2011) Aprataxin localizes to mitochondria and preserves mitochondrial function. PNAS 108: 7437-7442

Taketa S, Choda M, Ohashi R, Ichii M, Takeda K (2002) Molecular and physical mapping of a barley gene on chromosome arm 1HL that causes sterility in hybrids with wheat. Genome 45:617-625

Tsunewaki K (1980) Genetic diversity of the cytoplasm in Triticum and Aegilops. pp 49-100, Japan Society for the promotion of science 5-3-1 Kojimachi, Chiyodaku, Tokyo, Japan

Tsunewaki K (2009) Plasmon analysis in the Triticum-Aegilops complex. Breeding Science 59:455–470

Van Deynze AE, Dubcovsky J, Gill KS, Nelson JC, Sorrells ME et al (1995) Molecular-genetic maps for group I chromosomes of Triticeae species and their relation to chromosomes in rice and oat. Genome 38:45-59

Vrána J, Kubaláková M, Simková H, Číhalíková J, Lysák MA, Dolezel J (2000) Flow sorting of mitotic chromosomes in common wheat (Triticum aestivum L.). Genetics 156:2033-2041

Wanjugi H, Coleman-Derr D, Huo N, Kianian SF, Luo M-Ch et al (2009) Rapid development of PCR-based genome-specific repetitive DNA junction markers in wheat. Genome 52:576-587

Wardrop J, Snape J, Powell W, Machray GC (2002) Constructing plant radiation hybrid panels. Plant J. 31:223-8

Whitworth AJ, Lee JR, Ho VM, Flick R, Chowdhury R, McQuiddan GA (2008) Rhomboid-7 and HtrA2/Omi act in a common pathway with the Parkinson's disease factor Pink1 and Parkin. Dis Model Mech 1:168-174

Wong LH, Choo KHA (2001) Centromere on the move. Genome Res 11:513-516

Wang Ch-W, Lee S-Ch (2012) The ubiquitin-like (UBX)-domain-containing protein Ubx2/Ubxd8 regulates lipid droplet homeostasis. J Cell Sci 125:2930-2939

Wu Y, Zhang C, Liu C, Shuxin R, Yan Z (1998) Breeding technology of alloplasmic wheat. Sci China C Life Sci 41:449-58

Yan L, Loukoianov A, Tranquilli G, Helguera M, Fahima T et al (2003) Positional cloning of the wheat vernalization gene VRN1. Proc Natl Acad Sci USA 100:6263-6268

Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden T (2012) Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics 13:134

Yu F, Shi J, Zhou J, Gu J, Chen Q, et al (2010) ANK6, a mitochondrial ankyrin repeat protein, is required for male-female gamete recognition in Arabidopsis thaliana. Proc Natl Acad Sci USA 107: 22332-22337

Zubko MK, Zubko EI, Ruban AV, Adler K, Mock H-P, et al (2001) Extensive developmental and metabolic alterations in cybrids Nicotiana tabacum (Hyoscyamus niger) are caused by complex nucleo-cytoplasmic incompatibility. Plant J 25: 627–639

# 5. NEXT-GENERATION BULK RADIATION HYBRID ANALYSIS INDICATES A RHOMBOID PROTEIN AS A CANDIDATE FOR CYTOPLAMS COMPATABILITY IN WHEAT[3]

The location of the *species cytoplasm specific* (*scs*) gene was pinpointed to a 1.2 Mb interval by forward genetic radiation hybrid (RH) mapping. Here, a digestion-based complexity reduction protocol was coupled with the next-generation bulk segregant analysis method and RH to sequence the *scs* interval. The scan of the wheat geneome with ~2M sequence-based markers identified 3.2K reads linked to the *scs* locus on chromosome 1D. The short reads were extended into contigs using the 53X D-genome survey sequencing. The contigs were assembled into a gapped sequence of 1.1 Mb with an N50 of 66 and L50 of 2.9 Kb. The Wheat Zapper tool identified 10 contigs as containing genes orthologus to the known position of *scs*. Markers were derived from this scaffold and the genotyping of ~1400 RH lines revealed a single gene linked to the *scs* locus. Resequencing of this gene and expression analysis confirmed the existence of a functional SNP that is differentially expressed in alloplasmic conditions. Evidences from electron microscopy studies, fine mapping, expression analysis, functional characterization, and the available literature support the hypothesis that a *rhomboid* gene is indeed the *scs* locus.

## 5.1. Introduction

The genes in the nucleus, mitochondria, and cytoplasmic genomes cross-talk in a dense network of proteins and mRNA exchanges. These nuclear-cytoplasmic interactions (NCI) provide the basis for the molecular functions of the plant cell. Throughout evolution combinations of nucleus and cytoplasms have

[3] The material in this chapter was co-authored by Filippo Bassi, Dr. Matthew Hayden, Dr. Yong Gu. Matthew Hayden and his group performed the Ilumina sequencing, for this reason this part of the project is presented as Appendix. Dr. Yong Gu and his group performed the Illumina contig extension, for this reason that protocol is not presented here. Filippo Bassi designed the experiment, conducted all the analysis described in this chapter, and he is the primary developer of the conclusions that are advanced here. Filippo Bassi drafted and revised all versions of this chapter.

been selected to optimize the fitness of a species. When the compatibility between the nucleus and the cytoplasmic organelles is endangered by interspecies hybridizations the results can be dramatic, often leading to the premature death of the hybrid or its inability to procreate. In other instances, it leads to the surge of entirely new species with a new NCI balance, as it is the case of allohexaploid wheat (*Triticum aestivum* ssp *aestivum* L.) and many other lineages of the *Triticeae* tribe. Bread wheat (2n=6x=42, AABBDD) evolved recently through two successful interspecies hybridizations. The first (approximately 0.5 MYA) involved *T. urartu* Tumanian ex Gandylian (2n=2x=14, AA) as pollen donor (Dvorak et al. 1993) and *Aegilopis speltoides* ssp *ligustica* Tausch (2n=2x=14, SS) as female (Kilian et al. 2007), even though the actual *Ae. speltoides* ssp remains argument of discussion (for review see Matsuoka 2011 and Faris *in press*). The second hybridization occurred approximately 10,000 years ago between a domesticated form [*T. turgidum* ssp *dicoccum* (Schrank) Schubl (2n=4x=28, AABB)] of the cross described above and the pollen of the goat grass *Ae. tauschii* Coss (syn *T. tauschii*, *Ae. tauschii*; 2n=2x=14, DD) (Kihara 1944). Similarly, *T. timopheevii* ssp *timopheevii* (Zhuck) Zhuck (2n=4x=28, AAGG) was generated by the cross hybridization of a lineage of *T. urartu* as pollen donor and *Ae. speltoides* ssp. *speltoides* as the mother (Sarkar and Stebbins 1956; Kilian et al. 2007).

Interestingly, tetraploid durum wheat [*T. turgidum* ssp *durum* L (2n=4x=28, AABB)], hexaploid bread wheat, and tetraploid *T. timopheevii* share a common maternal species; hence their nuclei are theoretically immerged in the same maternal cytoplasm (plasmon S). Instead the co-evolution between the different genomes and their cytoplasms has originated two novel plasmons that differ from their maternal origin, a plasmon B and G for wheat and *T. timopheevii*, respectively (Tsunewaki et al. 2009; see Chapter 4). In this complex phylogenetic scenario, it becomes apparent the evolutionary importance of the nuclear encoded *species cytoplasm specific* (*scs*) genes, which orchestrate the compatibility between the nucleus and different plasmon.

Several *scs* genes have been identified to date, including the ones on homeologus group 1 of the *Triticeae*, one harbored on chromosome 1A of *T. timopheevii* defined $scs^{ti}$, and a second on the 1D chromosome of *T. aestivum* and *Ae. tauschii* defined $scs^{ae}$. Homeoallelic forms of these genes have also been identified on chromosome 1A of durum (Gehlhar et al. 2005), 1G of *T. timopheevii* (Asakura et al.

2000), 1D of *Ae. tauschii* var. *typica* (2n=2x=14, DD) (Ohtsuka 1980), and even 1H of *Hordeum vulgaris* L. (2n=2x=14, HH) (Taketa et al. 2002). The detailed analyses that led to the identification of the various forms of *scs* genes were made possible by the existence in wheat of cytoplasmic substitution lines. These lines have the nucleus of one species immerged in the cytoplasms of a diferent species (alloplasmic lines: *allo*-alien *plasmon*-cytoplasm), practically disrupting the compatible interactions that have evolved through speciation between the nucleus and the cytoplasm. Under these alloplasmic conditions it becomes possible to investigate new forms of compatible NCI and to identify the nuclear and cytoplasmic components that regulate them.

The primary result of alloplasmy is cytoplasmic male sterility, a phenomenon deeply exploited in plants to produce commercially superior hybrids (for review see Cisar and Coopers 2002 and Pelletier et al. 2007). However, the effect of the *scs* genes is not the restoring of male fertility in hybrid combinations, but rather the reestablishment of the compatible interaction between the nucleus and the alien cytoplasm, which ultimately determines the overall vigor of the progenies. Figure 5.1 provides a graphical representation of how the $scs^{ae}$ restores nuclear to cytoplasmic compatibility in the alien cytoplasm of *Ae. longissima* S&M (lo) (2n=2x=14, SS). In humans, many diseases have been associated with the improper interactions between the nucleus and the mitochondria, including obesity, type II diabetes, cancer, Alzheimer, and Parkinson (see review Desler and Rasmussen 2012). Hence, the cloning of the *scs* genes might not only help the scientific community to better understand the speciation of the *Triticeae* tribe and the complex network of NCI, but also provide new insights into some of the most dramatic diseases that affect human kind.

The location of the $scs^{ae}$ gene was narrowed down to a region of few Mb on chromosome 1D of wheat by means of radiation hybrid (RH) functional studies (Hossain et al. 2004a; Kalavacharla et al. 2006; Chapter 4). This region has surprisingly well conserved colinearity with other grass species, and eight candidate genes were identified. Among these candidates is a *rhomboid* gene that data presented here suggests as corresponding to the $scs^{ae}$ and $scs^{ti}$ loci. This was achieved by means of what has been collectively termed “fast forward genetics” (Schneeberger and Weigel 2011).

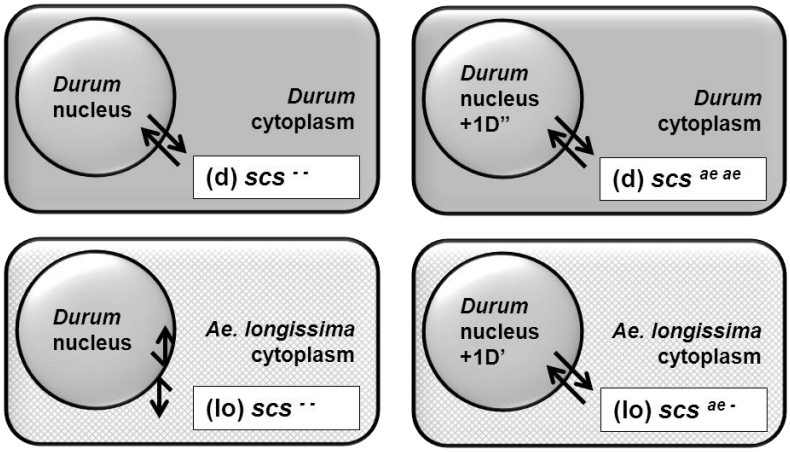


Figure 5.1. The *scs*$^{ae}$ gene guarantees proper nuclear-cytoplasmic compatibility in alloplasmic lines. The designation of the line is reported in the white boxes as used throughout this chapter.

The massively parallel sequencing platforms released in the past years have been recently applied to functional projects, surpassing the mere sequencing of genomes for which these technologies were originally developed. A good example of fast forward genetics in wheat is the study conducted by Trick et al. (2012), where the mRNA of lines harboring the dominant version of a gene for protein content was pooled, sequenced by Illumina RNASeq, and compared to the sequencing of the mRNA pooled from lines that had the recessive version of the gene. This is a novel approach that combined the principle of bulked segregant analysis (Michelmore et al. 1991) with the new next-generation (next-gen) technologies. Practically, the bulk of pooled RNA from the lines harboring the dominant version of the gene should express an allele that is not expressed by the bulk of 'recessive' lines. RNASeq allowed to rapidly determine the amount, type, and frequency of all expressed alleles in the two bulks. The comparison between the RNASeq results of the two bulks, together with synteny analysis, allowed to narrow down the location of the causative gene to an interval of just 0.4 cM in a matter of months, as compared to the many years that are typically necessary for a positional cloning project in wheat (for review see Salvi and Tuberosa 2005 and Krattinger et al. 2009).

Here, the concept of next-gen bulk segregant analysis (BulkSeq) was applied to the bulked DNA of RH showing the mutant phenotype (Δ*scs*$^{ae}$). The combination of BulkSeq and novel *in silico* methodologies allowed rapidly generating a gapped sequence of the region surrounding the *scs*$^{ae}$ locus,

and ultimately determining its causative gene to be a *rhomboid intramembrane serine peptidase*. Further, a functional SNP present in both *scs*$^{ae}$ and *scs*$^{ti}$ was identified, and its expression was proven to change between alloplsmic and euplasmic conditions. Finally, the 3D structure of the rhomboid protein was determined and a hypothesis was formulated to explain the effect of the functional SNP in determining nuclear-cytoplasmic compatibility.

## 5.2. Material and methods

### 5.2.1. Plant material

The use of alloplasmic lines of durum with the cytoplasm of *Ae. longissima* (lo) cytoplasm has been discussed in Chapter 4 specifcially for the two hemizygous versions (lo)*scs*$^{ti}$ – and (lo)*scs*$^{ae}$ -. Three additional alloplasmic lines have been employed in this study: one is the homozygous version (lo)*scs*$^{ae}$ *scs*$^{ae}$ that also carries a gene to restore male fertility (*Vi*); the second is an alloplasmic (lo) line that also carries the *Vi* gene, but is heterozygous for *scs*$^{ti}$ since it also harbors the *scs*$^{d}$ allele on chromosome 1A [(lo)*scs*$^{ti}$ *scs*$^{d}$*. Vi.vi*], which is referred to as 'Normal'; this definition is in comparison to the third alloplasmic 'Weak' line (lo) *scs* $^{d}$ -*.Vi.vi,* which produces only weak and non-vigorous progenies (Gehlar et al. 2005). Also, the progenies of a cross between LDN 1D (1A) and (lo)*scs*$^{ti}$ - were used to represent the mixed loci genotype (lo)*scs*$^{ti}$ *scs*$^{ae}$. Additionally, two euplasmic lines have been employed, one is the durum maintainer line '56-1' [(d) - -] and the second is a cross between '56-1' and the aneuploidy 1D line of 'Langdon' [(d)*scs*$^{ae}$ -]. For simplicity, all of the lines described above will be referred to using just their alloplasmic and *scs* designations. Seeds of *Ae. tauschii* were kindly provided by Dr. Steven Xu (USDA-ARS). A radiation hybrid (RH) population of ~1,400 individuals was developed through the crossing of 'Langdon' (LDN) and its aneuploid 1D line LDN 1D (1A) irradiated at 150 Gy, as described in Chapter 4.

### 5.2.2. Electro scan microscopy

Backcrossing the (lo)*scs* $^{ae-}$ line as female to a non-alloplasmic durum maintainer line [(d) - -] results in seeds that segregate in a 1:1 ratio of plump, which carry one copy of the *scs*$^{ae}$ gene, and shriveled, which do not carry any copy of the *scs* gene. Ten plump and ten shriveled seeds were

submerged into double-distilled water for 24h at 4ºC in the dark. The embryos of the hydrated seeds were then carefully dissected and moved to a fixative media. The embryos were fixed, imbedded, and stained at the Electron Microscopy Lab (ARS-USDA, Fargo, ND) and scanned on a JEM-100CX II electron microscope at magnifications from 5,700x to 19,700x.

5.2.3. Genotyping and phenotyping

DNA extractions of the 1,400 $RH_1$ and alloplasmic lines were performed as described in Chapter 4. Genotyping with the 'ndsu' and 'ret' markers was also presented in Chapter 4. The RH population was screened at seedling stage with markers ndsu212, ndsu21, and ndsu3 searching for lines that presented a breakage between these loci (a.k.a. 'recombination' in genetic mapping studies). The selected $RH_1$ lines were then transplanted into soil under greenhouse controlled conditions, and used as pollen donors for crossing with the (lo)*scs* $^{ti}$ – tester line. The ratio of shriveled to plump seeds obtained by test-crossing was converted into the deletion-type (a.k.a. 'haplotype' in genetic mapping studies) for the *scs*$^{ae}$ locus, where a 1:1 ratio indicates that the *scs*$^{ae}$ gene has been deleted, while 1:3 indicates retention of the locus, as discussed in Chapter 4. The North Dakota contig (NDCtg) markers were developed on the basis of the sequencing data and are better explained in section 5.2.5.

5.2.4. Bulk segregant analysis

Six $RH_1$ lines were selected for showing the Δ*scs*$^{ae}$ phenotype (Table 5.1), five of which were discussed in Table 4.1, and an additional six $RH_1$ lines for showing the 'positive' *scs*$^{ae}$ phenotype. DNA was extracted from these lines using a modified phenol:chloroform protocol (Hossain et al. 2004b), DNA was equilibrated at 100 ng $\mu l^{-1}$ using a NanoDrop spectrophotometer (Thermo Scientific, DE), and 20 µl of the DNA equilibrations were pooled from each of the six lines to create a bulk negative (Δ*scs*$^{ae}$) and a bulk positive (*scs*$^{ae}$) of 120 µl of total volume. Each DNA pool was digested with *Pst* I or *Aat* II restriction enzymes, ligated to the paired-end 2 (PE2) sequencing adapter, physically sheared using sonication, and then ligated to the paired-end 1 (PE1) sequencing adapter. DNA fragments were selectively enriched by PCR for those having a PE1 adapter at one end and a PE2 adapter at the other end. The complexity

reduced *Pst* I and *Aat* II libraries were run on agarose gel and fragments of 475 ± 50 bp in size were eluted from the gel following the provider instructions (Gel Elution Kit, Qiagene, Valencia, CA). Two digestion libraries (one for bulk positive and one for bulk negative) were created for each restriction enzyme. The libraries were bar-coded and then pooled. One lane of the Illumina HiSeq2000 (available at VABC, Melbourne) was used for pair-end (100 bp) sequencing of each pooled restriction libraries, for a total of two lanes employed in this experiment. Due to unidirectional adapter ligation, the PE1 read corresponds to the end of DNA fragments with a random physical shear, whereas the PE2 read corresponds to DNA ends having a restriction enzyme cleavage site (Figure 5.2).

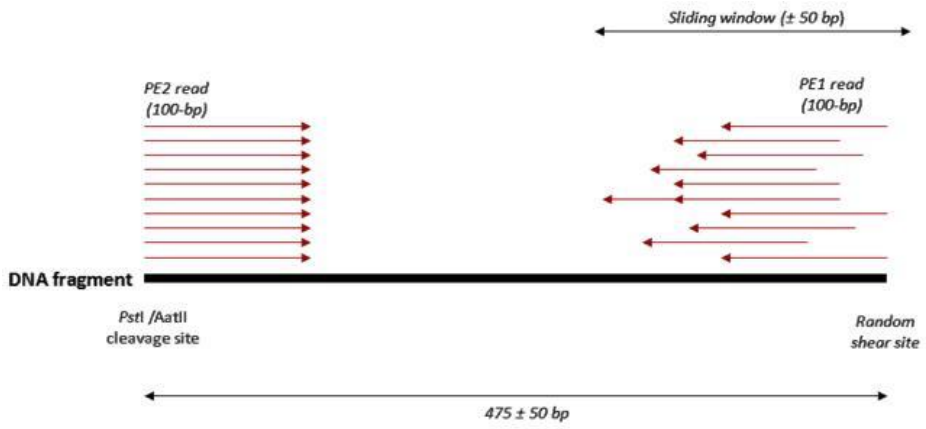


Figure 5.2. Complexity reduction of DNA fragments. The complexity reduction was achieved by a combination of restriction enzyme digestion (PE2) and shearing by sonication (PE1).

The computational analyses required to identify the fragments present in the bulk positive but not in the bulk negative are complex, and are reported in Appendix D as provided to us by our collaborators at VABC (Dr. Matthew Hayden).

5.2.5. Contig assembly and precise mapping

The differential sequences identified through Illumina bulk segregant (BulkSeq) analysis were run through a pipeline developed at USDA-Albany. This pipeline employs the 53X survey sequence of the D-

genome progenitor *Ae. tauschii* to extend the input sequences. The extension procedure cannot be disclosed in details, but practically the input sequence is used to query the *Ae. tauschii* database and identify 100% identity fragments. These fragments are then assembled onto the input sequence, and the procedure is repeated using the newly assembled fragment. The procedure continues until a repetitive sequence is encountered, at which point the software stops. Each input sequence is extended independently from the others. The extended sequences are then assembled using the SeqMan Pro software, maintaining the default parameters except that FASTA sequences are used as input, and that any contig with at least 1 sequence of any length is accepted as output. These contigs are then run on the 'Wheat Zapper' application (available at: http://wge.ndsu.nodak.edu/wheatzapper/) to identify any gene present on the assembled contigs on the basis of synteny with three model species (*Oryza sativa* L., *Sorghum bicolor* L., and *Brachypodium distachyon* L.). The contigs containing genes are then aligned to the wheat tentative consensus (TC) EST sequences (http://compbio.dfci.harvard.edu/tgi/plant.html; Quackenbush et al. 2001) provided by the 'Wheat Zapper'. Those contigs that provide less than 30% coverage of the TCs sequence are considered as carrying pseudogenes, and were excluded from further analysis. For the remaining contigs harboring complete genes, the position of the exons/introns junctions and the non-genic spaces are accurately annotated. Primers are designed on non-coding portions using the Primer-blast application (Ye et al. 2012). These primers were defined NDCtg and used to confirm the quality of the assembly. Failure to amplify the DNA of LDN 1D (1A) was considered as a possible error in the assembly. Also, the primers were searched for chromosome 1D specificity as discussed in Chapter 4, using 2.5 mM of $MgCl_2$ and 65-60ºC touch-down annealing temperature. The sequences of the primers proving 1D-specificity are reported in Appendix E. These primers were then used (in combination with the control marker DEASY; see Chapter 2) to genotype the most informative RH-1D lines. A map was generated using the Chartagene software (de Givry et al. 2005). The contigs were aligned onto this map by means of the NDCtg markers or synteny information, to generate a sequence scaffold of the $scs^{ae}$ region. Figure 5.3 provides a graphical representation of the workflow described in this section.

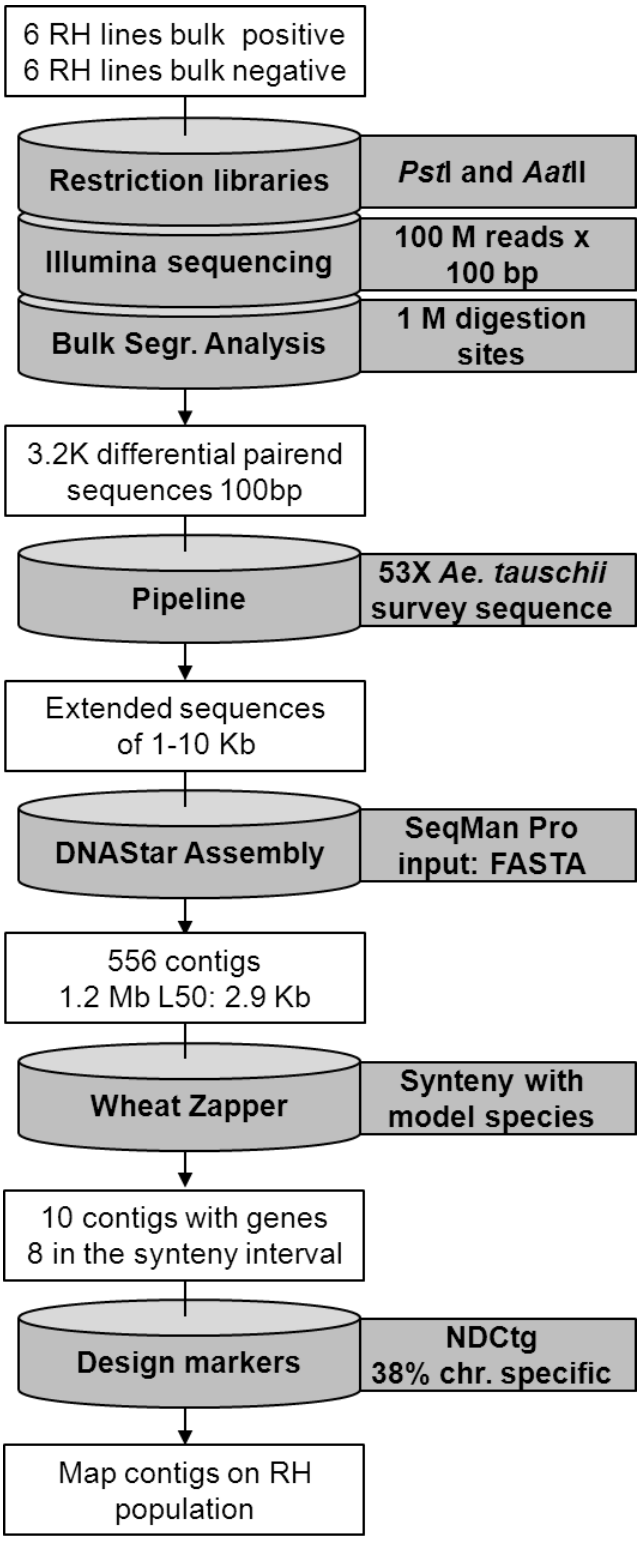


Figure 5.3. Experimental workflow for Illumina bulk segregant analysis in radiation hybrids.

5.2.6. Functional analysis

5.2.6.1. Re-sequencing of the Rhomboid gene and SNP detection

The gene tagged by marker ndsu297 of contig 5 was identified as the most associated to the *scs*$^{ae}$ delition-type among the 1,400 $RH_1$ screened and was consequently studied in more details. The TC sequence associated with this marker was orthologus to a *rhomboid* gene in the three sequenced model species. Six primers combinations (defined as RHBD1-6) were designed on the basis of the sequence of contig 5 to span from 207 bp upstream of the gene to 367 bp downstream. The sequences of these primers are made available in Appendix E, and their amplification conditions are as reported for the NDCtg markers. The primers were used to amplify from the DNA of the flow-sorted chromosomes 1A, 1B, and 1D (see Chapter 4) and the band corresponding to the expected sizes were eluted from the gel (Figure 5.4) and sent for sequencing at GeneWiz (South Plainfield, NJ).

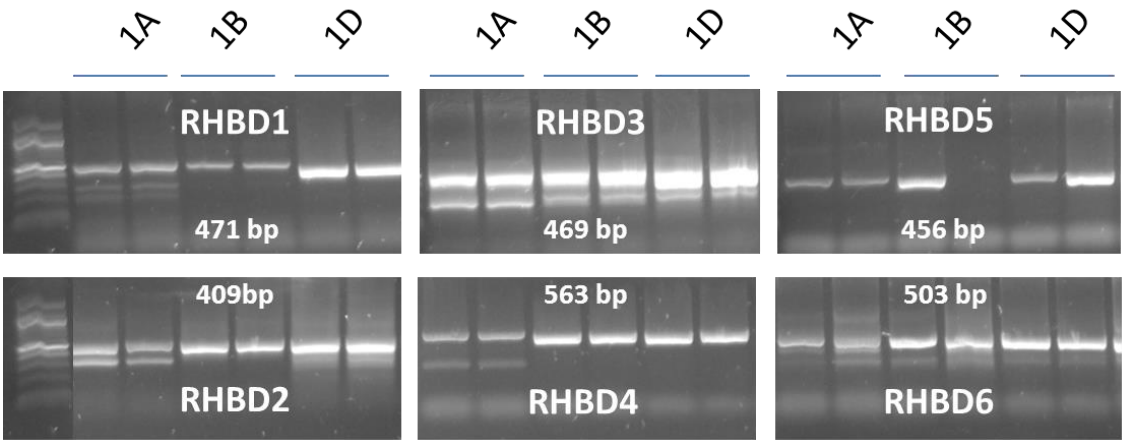


Figure 5.4. Amplification of fragments of the rhomboid gene. The RHBD1-6 primers were used to amplify the flow-sorted chromosomes 1A, 1B, and 1D. The band sizes expected on the basis of the sequence available for contig 5 are reported, a DNA ladder is provided to the left of the figure.

The obtained sequences were first hand analyzed with ChromasPro (Technelysium Pty Ltd, Helensvale, QLD, Australia) to assure good quality, and then assembled into a continuous sequence for each chromosome. The same sequence was obtained by amplification of *Ae. tauschii*, (lo)scs$^{ti-}$,' Normal', and 'Weak' genotypes. These genotypes, with the exclusion of *Ae. tauschii*, are tetraploid and direct sequencing would fail. So the amplicons obtained were subcloned into pGEM vectors (Promega,

Madison, WI) following the provider guidelines, and then sequenced. On the basis of sequence similarity with the flow-sorted chromosomes 1A and 1B sequence, the sequences originated from the two homeologus were determined.

5.2.6.2. Protein prediction and modeling

All the genomic sequences from homeologus chromosomes 1A, 1B, and 1D were used as input for GeneScan (available at http://genes.mit.edu/GENSCAN.html) and the TriAnnot Pipeline (available at http://urgi.versailles.inra.fr/Tools/TriAnnot-pipeline) employing the default parameters. The CDS and protein sequences predicted by the two software were aligned using 'Multalin' (available at http://multalin.touluse.inra.fr/mulatlin/). The SNPs and amino acid changes were graphically identified in the alignment.

The protein sequence from 1A, 1B, 1D, and the four genotypes were searched by InterProScan (available athttp://ebi.ac.uk/Tools/pfa/iprscan/) for similarities with previously annotate protein domains. Also, the SOUSI (available at http://bp.nuap.nagoya-u.ac.jp/sosui/) and Phobius (http://phobius.binf.ku.dk/) tools were used to predict possible transmembrane helices, and to determine the protein cytoplasmic/non-cytoplasmic orientation. The proteins were then modeled using Swiss-Model (available at http://swissmodel.expasy.org/) and EsyPred3D (Lambert et al. 2002) using as template the 'A' chain of 2NR9, a rhomboid protein from *Haemophilus Influenza* that has been crystalized (Lemieux et al. 2007). The 3D protein structure was run on PDB Site Scan (http://wwwmgs.bionet.nsc.ru/cgi-bin/mgs/fastprot/pdbsitescan) to determine the catalytic site within the protein. Chimera 1.6.1 (Pettersen et al. 2004) was used to visualize and analyze the 3D protein structure. The WoLF PSORT (available at http://wolfpsort.org) tool was employed to determine the specific cellular localization of the protein through identity comparison of other characterized proteins. A PDB Sum page (http://www.ebi.ac.uk/pdbsum/) describing all the information available for the EasyPred3D prediction of the 1D version of this protein has been created with the IPDB code: fl60 and is encrypted with password: 200609. This suite allowed the precise determination of the helices turning points and to investigate the presence of transcription factors or promoters/enhancer sites with PLACE (http://www.dna.affrc.go.jp/PLACE/signalup.htm), Promoter 2.0

(http://www.cbs.dtu.dk/services/Promoter/), SoftBerry TSSP (http://linux1.softberry.com/berry.phtml?topic=tssp&group=programs&subgroup=promoter), and Softberry NSITE_PL (http://linux1.softberry.com/berry.phtml?topic=nsitep&group=programs&subgroup=promoter). Finally, the intron/exon design tool at WormWeb (available at http://wormweb.org/exonintron) was employed to graphically represent the transcription factor sites, intron/exons, and primers information.

5.2.6.3. Expression analysis

Plump seeds of (lo)*scs* $^{ae\ ae}$, (lo)*scs* $^{ae\ ti}$, (lo)*scs* $^{ti}$ –, *(d) scs* $^{ae}$ - and (d)- - were hydrated for 24h at 4ºC in the dark, then transferred at room temperature to initiate germination. After 12h the embryos were dissected with a scalpel and collected into liquid nitrogen. The embryos were grainded using mortar and peaster, making sure to avoid thawing of the tissues. Equal weights of grainded tissue were used for mRNA extraction by Qiagene mRNA Extraction Kit following the provider guidelines. Also, the same genotypes were allowed to germinate, and 15 days after initial hydration the tip of the first leaf was collected in liquid nitrogen for mRNA extraction. The M-MLV Reverse Transcriptase Kit (Promega) with Oligo dT and rRNAse inhibitor was used to generate cDNA, following the provider instructions. Primers NDRT20 (see Appendix E) were designed to amplify a 197 nucleotide cDNA fragment between two exons, which corresponds to 284 bp in the gDNA (87 bp intron). This fragment contains a functional SNP conserved in all the genotypes carrying the functional version of *scs*. As a reference, a primer combination was designed to amplify 124 bp of an Actin gene (GeneBank ID: TA411_4571). The Roche 2X SybrGreen Fast Start Mastermix was used to amplify the cDNA samples in a 20 µl reaction with 0.5 µM of NDRT20 or Actin primers and 400 ng of cDNA. The samples for the qPCR were run in triplicate on a 7500 Fast Real-Time PCR System (Applied Biosystems, Carlsbad, CA) at 50°C for 2 min, 95°C for 10 min, followed by 40 cycles of 95°C for 15 sec, 60°C for 1 min. The data were analyzed with the SDS v2.2.1 software (Applied Biosystems, Carlsbad, CA) employing the $2^{-\Delta\Delta Ct}$ equation (Livack and Scmittgen 2001) to estimate the relative quantification (RQ) of the NDRT20 expression. The threshold cycle was automatically determined by the SDS software, and the (lo)$scs^{ae}$ $scs^{ae}$ sample was set as the normalizer. An attempt was made to design two TaqMan probes (see Appendix E) targeting the two allele of this

SNP. Unfortunately, these two probes in combination with the TaqMan Fast Advanced Master Mix (Applied Biosystems) and primers NDRT20 failed to provide adequate emission profiles, and were then discarded. Instead, the relative level of expression of the two alleles was determined by direct sequencing. New amplicons were generated by qPCR, but stopping the reaction in its exponential phase at cycle 28. The products were run on agarose gel, eluted, and sent for pyrosequencing at GeneWiz. The chromatograms were analyzed using ChromasPro. The heights of the two adenine peaks surrounding the SNP position were measured, and then averaged. The heights of the guanine and adenine peaks at the SNP position were also measured. The height of the SNP peak minus the average height of the surrounding peaks was considered as the normalized value. A value of 100% was given to the height of the guanine peak in (lo)*scs*$^{ae}$ *scs*$^{ae}$, and the relative expression of the other samples was calculated as comparison to its value. To verify the quality of the analysis, all samples were sequenced six times, and identical results were obtained for all replicates. Only one representative result is presented here.

## 5.3. Results and discussions

### 5.3.1. Disruption of cellular organization

The embryo development process is regulated by the newly originated zygote, rather than maternally determined (see review Bosca et al. 2011). Scan electron microscopy analysis of the embryos of shriveled and plump seeds revealed that the disruption observed at the macro level (see Figure 4.1 in Chapter 4) also occurs at the cellular level. Figure 5.5 shows the cells within the embryos of *Ae. longissima* alloplasmic lines with or without the *scs*$^{ae}$ gene. The cells in the embryo dissected from plump seeds (Figure 5.5a and 5.5c) shows adequate vacuolar turgor, with the organelles properly aligned toward the cell wall. On the other hand, the vacuole of the shriveld embryonic cells is invaginated, with incomplete turgor (Figure 5.5b and 5.5d) as expected in plant cells undergoing programmed death (Vanyushin et al. 2004; van Doorn and Woltering 2005; van Doorn et al. 2011). Further, in cells lacking the *scs*$^{ae}$ gene, protein storage vacuoles (psv; Dominguez et al. 2012) are evident as larger spheres with black residues of chromatin protruding form their surface (Figure 5.5d and 5.5e). These spheres are formed by merging of lipidici vescicles (light grey), while the chromatin residues originated from the

degeneration of the nucleus and organelles. In the scutellum of wheat, these morphological characteristics were associated with vacuolar programmed cell death (VPCD) that this tissue normally undergoes 7 days after germination (Dominguez et al. 2012). The embryo of the shriveled seeds instead races to VPCD immediately after germination.

The shriveld embryos were often believed to represent dead tissue due to their incapability of germinating. Instead, following their cellular behavior immediately after hydration, it was possible to determine that these embryos are not dead, but rather undergo immediate VPCD as result of their incompatible NCI. Plump seeds appear to have normal cellular behavior and tissue turgor, as expected from helthy hydrated seed tissues.

### 5.3.2. Creation of the segregating bulks of radiation hybrids

The DNA of six RH lines showing the retained phenotype for *scs*$^{ae}$ and six showing the deleted phenotype were pooled to form a segregating bulk positive (bulk$^{POS}$) and a segregating bulk negative (bulk$^{NEG}$), respectively. Pooling of the DNAs from the six 'positive' lines should create a sample that contains the whole genome [13" + 1A' + 1D'] without any unmasked deletions. In fact, any deletions should be masked by the simultaneous presence of a non-deleted version from a different line. The DNA pooling of the six 'negative' lines should also provide the entire retained genome, with the exclusion of the *scs*$^{ae}$ region which was deleted in all of these lines. Table 5.1 shows the six 'negative' lines that were selected, their phenotype for *scs*$^{ae}$, and their deletion-typing information available after the genotyping discussed in Chapter 4. The negative bulk has an uninterrupted deletion that spans from *Xndsu223* to *Xndsu20*, which corresponds to the refined interval for *scs*$^{ae}$ as discussed in Chapter 4. In principle, a comparison analysis between the bulk$^{POS}$ and bulk$^{NEG}$ should find no differences, except for a segment that is present (retained) in the bulk$^{POS}$ but absent (deleted) in the bulk$^{NEG}$. This segment corresponds to the *scs*$^{ae}$ region.

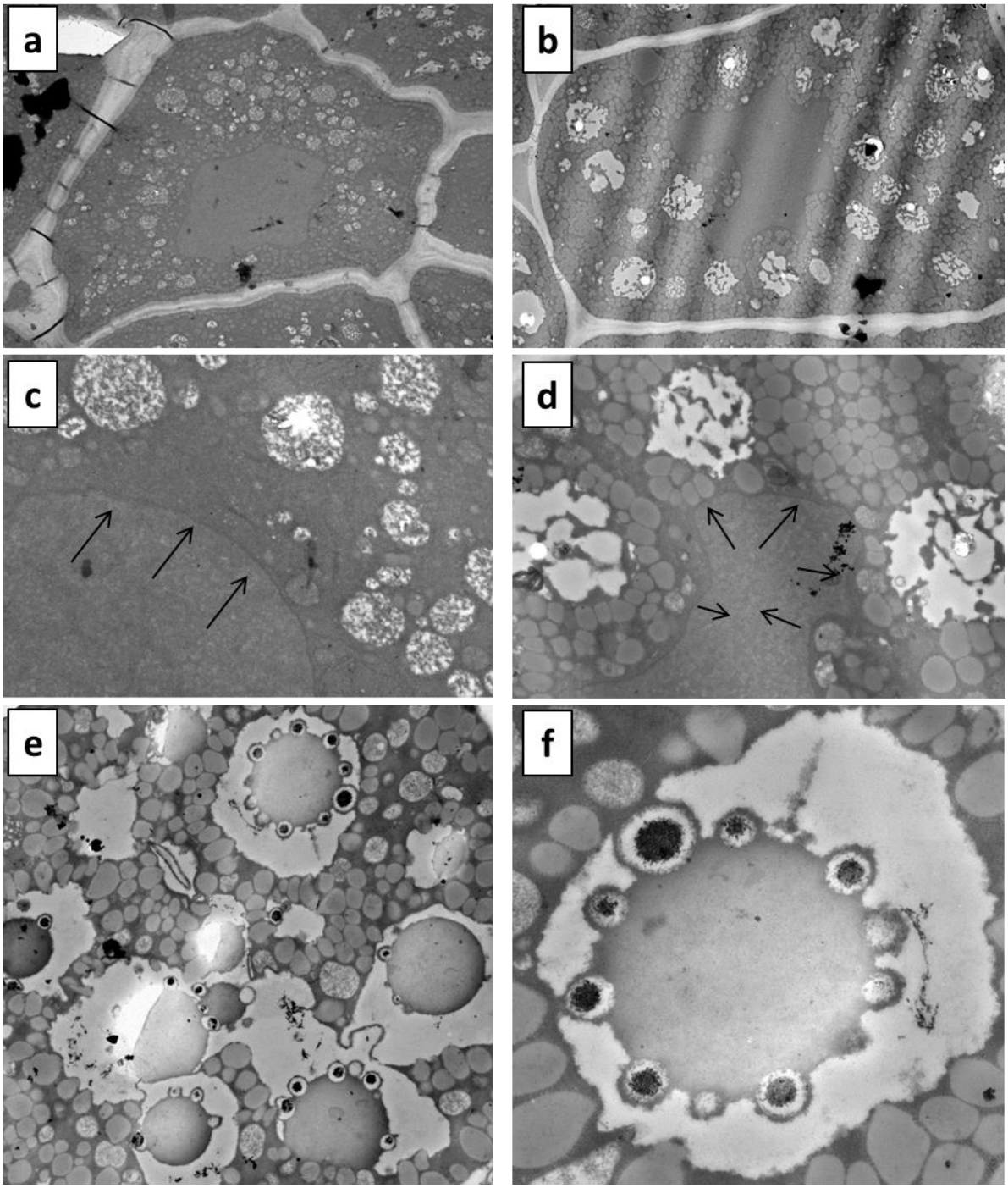


Figure 5.5. Electron microscopy of plump and shriveled alloplasmic embryonic cells. a) Plump and b) shriveled cells at 5,700X; c) plump and d) shriveled cells at 11,700X with indicated the turgor pressure of the vacuole (black arrows); e) shriveld cells undergoing apoptosis at 14,700X; f) tonoplast protrusion from a misplaced mitochondria forming apoptotic vescicles at 19,700X.

In Chapter 4, the segment *Xndsu223-Xndsu20* was calculated to be 1.1 ± 1 Mb in size based on conversions of syntenic distances. Also, this segment equals to 1.5 cR as estimated on a population of ~1,400 $RH_1$ lines. In Chapter 4 it was calculated that in the 1D-RH population the conversion ratio is 1.5 Mb per cR, which would size this segment at 2.25 Mb based on RH map distances.

Table 5.1. Deletiontyping of the radiation hybrid (RH) lines used for bulk segregant analysis

| RH lines | cR | 67 | 3057 | 3058 | 3017 | 3027 | 3064 | $Bulk^{NEG}$ | $Bulk^{POS}$ |
|---|---|---|---|---|---|---|---|---|---|
| Plump:shriveled | | 11:13 | 26:30 | 26:24 | 25:25 | 24:28 | 54:57 | 166:177 | 232:90 |
| Ratio | | 1:1 | 1:1 | 1:1 | 1:1 | 1:1 | 1:1 | 1:1 | 3:1 |
| *Xret09-9* | 0.0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| *Xndsu212* | 10.0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 |
| *Xndsu13* | 10.3 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| *Xndsu222* | 10.5 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| *Xndsu223* | 10.5 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| *Xndctg49.1* | 10.7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| *Xndctg49.2* | 10.7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| *Xndctg235.1* | 10.7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| *Xndctg6.1* | 10.7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| *Xndctg6.2* | 11.0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| *Xndctg478.1* | 11.0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| *Xndsu19* | 11.0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| *Xndctg478.2* | 11.0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| *Xndctg478.3* | 11.0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| *Xndsu18* | 11.2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| *Xndsu294* | 11.3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| *Xndctg77.1* | 11.3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| *Xndctg77.2* | 11.5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| *Xndctg108.1* | 11.5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| *scs*$^{ae}$ | 11.7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| *Xndsu297* | 11.7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| *Xndctg5.1* | 11.8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| *Xndsu299* | 11.8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| *Xndsu295* | 12.0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| *Xndsu290* | 12.0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| *Xndsu291* | 12.0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| *Xndsu233* | 12.2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| *Xndsu20* | 12.7 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| *Xndsu21* | 13.1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| *Xndsu298* | 13.1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| *Xndsu224* | 13.1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| *Xndsu31* | 13.2 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| *Xndsu3* | 13.2 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| *Xndsu226* | 13.5 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |

5.3.3. Illumina bulk segregant analysis on radiation hybrids

The bulk$^{POS}$ and bulk$^{NEG}$ were digested with one restriction enzyme, either *Pst* I or *Aat* II. These two enzymes are rare cutters that recognize complimentary restriction sites on the opposite strands of the DNA. In particular *Pst* I recognizes non-methylated CTGCA*G ('*' indicates the digestion site) while *Aat* II recognizes non-methylated GACGT*C. The restriction libraries were further fragmented to 475 ± 45 nucleotides by sonication (Figure 4.2). Illumina HiSeq2000 sequencing of the *Pst* I library for bulk$^{POS}$ and bulk$^{NEG}$ produced ~62 M good quality pair-end reads for a total of 12.4 Gb of sequences. Illumina HiSeq2000 sequencing of the *Aat* II library for bulk$^{POS}$ and bulk$^{NEG}$ produced ~48 M good quality pair-end reads for a total of 9.6 Gb of sequences. The sample of *Aat* II library was lightly contaminated with an unidentified prokaryotic genome, which caused a 25% reduction in the total amount of good quality reads. Assembly of the *Pst* I reads revealed 1,268,786 contigs, which indicates ~635K *Pst* I digestion sites, since each enzymatic digestion creates two separate fragments corresponding to two separate contigs (number of restriction sites = number of contigs divided by two). Assembly of the *Aat* II reads revealed 650,433 contigs, which indicates 325K *Aat* II digestion sites. The average sequencing coverage can then be derived to be 49X (reads divided by contigs) for the *Pst* I library and 73X for the *Aat* II library. The partial contamination of the *Aat* II sequencing reaction caused a reduction in the total number of conitgs identified, which in turn resulted in an artificial increase in the coverage. Since these enzymes recognize complementary digestion sites, it is likely that the observed discrepancies between the number of contigs and their coverage in the two libraries are consequences of the different quality of the two Illumina sequencing experiments, rather than real biological differences. The 12,604 Mb of the genome of the RH lines (12,000 Mb AABB genome + 604 Mb chromosome 1D; Dolezel et al. 2009) was compartmented into 0.635M (or 0.325M for *Aat* II) segments of ~20 K in size, tagged at each side with approximately 500 bp pair-ends sequences. A previously published genotyping by sequencing study identified 0.75M *Pst* I digestion sites on the ~17 Gb winter wheat genome (Poland et al. 2012). The hexaploid wheat genome is 1/3 larger than the durum genome (Dolezel et al. 2009) employed here, hence the number of expected digestion site in durum should be approximately 0.5M, which is considerably less than what is actually observed for the bulks of RH lines. However, $RH_1$ lines are basically $F_1$ and it has been demonstrated in

maize that the $F_1$ progenies tend to have a much lower degree of DNA methylation than their inbred parents (Tsaftaris and Polidoros 2000). Since *Pst* I does not cut on methylated sites, it can then be expected that a reduction of DNA methylation would expose more restriction sites for *Pst* I to cleave. Also, radiation might have an effect on the overall state of chromatin, including methylation (see Chapter 2), leading to a further increase in restriction sites.

Differential analysis revealed 1,990 contigs present in the bulk$^{POS}$ but lacking in the bulk$^{NEG}$ for the *Pst* I library, and 1,328 in the *Aat* II library, for a total of 3,320 contigs that putatively correspond to the *scs*$^{ae}$ region. These would correspond to 995 (or 664 for Aaat II) digestion sites spaced on average ~20 Kb for a total region of approximately 20 Mb in size. This calculated size is likely an overestimation of the segment by at least a factor of 10, probably due to the fact that the *scs*$^{ae}$ region is particularly rich of genes (see Chapter 4), and hence has an overabundance of non-methylated restriction sites.

In the *Pst* I library the bulk$^{POS}$-specific contigs had an average reads coverage of 292X with the PE2 restriction-end extending for 93 bp and the PE1 pair-end averaging 329 bp. The *Aat* II library had an average coverage of 1,159X for the bulk$^{POS}$-specific contig, with 93 bp long sequences for the PE2 restriction end and 202 bp for the PE1 sonication pair-end. In both cases, coverage of five to twenty folds higher than the genome-wide coverage was identified for the contigs present only in the bulk$^{POS}$ but not in the bulk$^{NEG}$.

Overall, both libraries identified highly associated contigs specific for the bulk$^{POS}$, with coverage ratios always more than 200 to zero (i.e. absent in the bulk$^{NEG}$). The DNA was fragmented into segments of approximately 500 bp in size at both sides of the restriction sites. Based on the average lengths of the pair-ends (PE1 + PE2) approximately 60% of these segments were sequenced. Each library identified a similar number of restriction sites for *Pst* I and *Aat* II within the *scs*$^{ae}$ interval, but the difference in the total good quality reads generated by the two Illumina experiments had an effect on the final coverage, and sequence length.

5.3.4. Assembly of the bulk$^{POS}$-specific Illumina contigs

The identification of 3,320 bulk$^{POS}$-specific contigs of ~93 + 202 ± 113 bp in size provided an entry point to attempt the gapped assembly of the *scs*$^{ae}$ region. These short sequences are derived from chromosome 1D of wheat (which harbors the *scs*$^{ae}$ locus). A 52X survey sequence is available for *Ae. tauschii* (You et al. 2011), the D-genome progenitor of wheat (Kihara 1944). Using an iterative approach, the genome survey of *A. tauschii* was used to extend the short Illumina contigs to sequences of up to 10 Kb in size (Figure 5.3). A masking analysis (data not show) confirmed that these extended segments do not contain any repetitive sequences, as expected because of the specific protocol used for the extension.

These longer sequences were then assembled into 556 contigs of average length 2,130 bp. The term “contigs” from here on is used to refer to the assembly of the extended Illumina contigs, practically indicating contigs of contigs. The gap assembly has a L50 of 2,896 bp and spans 1.15 Mb. Even though still containing gaps, this value is very similar to the total size predicted by synteny analysis (1.1 Mb; Chapter 4). The largest contig spans ~70 Kb, while the shortest just 138 bp. Table 5.2 summarizes the data obtained in the assembly, while Figure 5.6 graphically shows the number of extended sequences that were assembled for each contig.

Table 5.2. Summary of the assembly of the extended Illumina bulk$^{POS}$-specific contigs

| Data (unit) | Value |
|---|---|
| L50 (bp) | 2,896.0 |
| N50 (#) | 66.0 |
| Total length (bp) | 1,148,578.0 |
| Average length (bp) | 2,130.0 |
| Max length (bp) | 70,049.0 |
| Min length (bp) | 138.0 |
| Average coverage (X) | 2.3 |
| Contigs (#) | 556.0 |
| Genes (#) | 25.0 |
| Contigs with genes (#) | 10.0 |
| Pseudogenes (#) | 14.0 |
| Contigs with pseudogenes (#) | 11.0 |

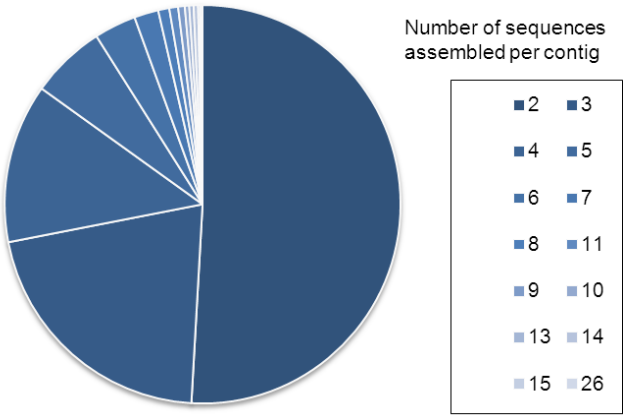


Figure 5.6. Number of sequences assembled per contig.

The total number of contigs (556) is surprisingly similar to the number of calculated restriction sites for *Aat* II within the *scs*$^{ae}$ region (664). Also, 72% of the contigs contain just two or three sequences (Figure 5.6). It is likely that these contigs originated by the assembly of the two fragments created by the digestion of one restriction site (PE2). In fact, the number of restriction sites that existed within the sequence of the contigs analyzed correlated well (r= 0.89) with the number of extended Illumina sequences (divided by two) that were used to assemble the contig.

5.3.5. Annotation and confirmation of the assembled contigs

The 556 contigs were used as input for the 'Wheat Zapper' tool to identify any genes present within the sequence, and simultaneously derive their orthologus relationships with rice, *Brachypodium*, and sorghum. A total of 25 genes and 14 pseudogenes were found within the contigs (a complete list is provided in Appendix F). The 25 genes were contained within 10 contigs, while 11 contigs accounted for the pseudogenes. Since the pseudogenes matched with only fractions of known wheat ESTs and are not likely to generate a functional protein, they were ignored in the following analysis. Among the 25 genes, 22 identified orthologus loci on rice chromosomes 5 or 10, 23 on *Brachypodium* 2 or 3, and 24 on sorghum 1. It was established (Chapter 4) that the *scs*$^{ae}$ locus sits in close proximity of a paleofusion site that occurred between ancestral chromosomes corresponding to the modern rice 5-10, *Brachypodium* 2-

3, and sorghum 1-9. So for, 24 out of the 25 genes identified by Illumina bulk segregant analysis recognized an orthologus locus that matched what was previously known (Chapter 4) as the syntenic region of the *scs*$^{ae}$ gene. However, it was also determined that the *scs*$^{ae}$ gene is located just above this fusion event, in a region of well conserved synteny with rice chromsomoe 10, *Brachypodium* 3 and sorghum 1. A total of 20 genes and 8 contigs matched all of these orthologus expectations. In addition, two genes orthologus to rice chromosome 3, *Brachypodium* chromosome 1 or 3, and sorghum chromosome 1 exist within the same contigs that carry other syntenic genes. Hence, 22 genes within 8 contigs were identified by Illumina bulk segregant analysis of radiation hybrid as located in close proximity of the *scs*$^{ae}$ locus (Table 5.3). These genes and contigs were selected for RH mapping.

Table 5.3. Assembled contig containing genes within the synteny interval

| | Contig | Number | Avg. | Number of | Number of NDCtg | |
|---|---|---|---|---|---|---|
| Contig ID | length (bp) | assembl seq. | coverage | genes | Designed | Amplifying |
| Contig 108 | 4,925 | 2 | 2.0 | 2 | 2 | 2 |
| Contig 49 | 7,607 | 2 | 2.0 | 2 | 3 | 2 |
| Contig 235 | 9,532 | 2 | 1.0 | 2 | 3 | 3 |
| Contig 5 | 9,802 | 2 | 2.0 | 2 | 3 | 3 |
| Contig 77 | 17,430 | 2 | 1.5 | 3 | 3 | 3 |
| Contig 478 | 23,347 | 7 | 2.0 | 6 | 5 | 4 |
| Contig 6 | 26,460 | 3 | 2.3 | 5 | 5 | 4 |
| total | 99,103 | 20 | 1.8 | 22 | 24 | 21 |

In summary, Illumina bulk segregant analysis of radiation hybrids allowed through *in silico* studies to rapidly move from nearly 2M investigated sites (*Pst* I + *Aat* II), to 3.2 K bulk$^{POS}$-specific contigs, to 556 extended/assembled contigs, to 8 contigs containing genes, all of which matched the synteny predictions of the targeted region (Figure 5.3). The remaining 535 (96%) contigs that do not carry any genes represent probably regions of non-methylated DNA that do not correspond to canonical coding sequences. Possibly these hypo-methylated regions provide entry points for transcription factors or chromatin remodeling factors, but their investigation is beyond the scope of this study and will not be discussed further here.

5.3.6. Mapping of contigs on radiation hybrids

The quality of the assembly for the contigs harboring genes was investigated by PCR. Since primers targeting exons are likely to amplify regardless of the precision of the assembly, primers combinations were designed on the non-coding fractions of the contigs. A total of 24 primer combinations were designed for the 8 contigs, a minimum of 2 primers for the smaller contigs and up to 5 for the larger ones (Table 5.3). The 88% of the designed primers amplified a genotype containing chromosome 1D at the single PCR condition used for the test, confirming the good quality of the assembly of extended Illumina bulk reads. Further, 12 (50%) of these amplified exclusively chromosome 1D, and were then used for genotyping the most informative of ~1,500 $RH_1$ lines, which carry breakages within the *scs*$^{ae}$ interval. These primers combinations were defined NDCtg (North Dakota Contigs) markers. The twelve NDCtg markers and ten NDSU markers identified seven breakages in close proximity of the *scs*$^{ae}$ locus. These brekages allowed precisely aligning and orienting the eight contigs, creating a gapped scaffold of the *scs*$^{ae}$ region (Figure 5.7).

Other than the correct synteny of the genes and the amplification of the designed primer combinations, an additional confirmation of the good quality of the RH BulkSeq anlaysis assembly is provided in Table 5.1, where the NDCtg markers are shown to score as deleted on all of the six $RH_1$ lines used to generate the bulk$^{NEG}$. This indicates that the uninterrupted deletion on the *scs*$^{ae}$ locus, which was created by bulking these lines, was correctly exploited by Illumina sequencing to identify reads that belong to this deleted region, and that markers derived from these reads mapped indeed in the target interval. Further confirmation is provided by two genes (tagged by ndsu19 and ndsu297) that map in their correct location as based on their position on the contigs sequences, corresponding to Os10g37330 on contig 478 and Os10g37760 on contig 5, respectively. However, there are at least two genes (tagged by ndsu18 and ndsu294) that map in close proximity of the *scs*$^{ae}$ locus but that were not identified by Illumina bulk segregant analysis. An experimental mistake is possible; the segments corresponding to these genes rather were not sequenced, or during the process of extension, assembly, and annotation were somehow omitted.

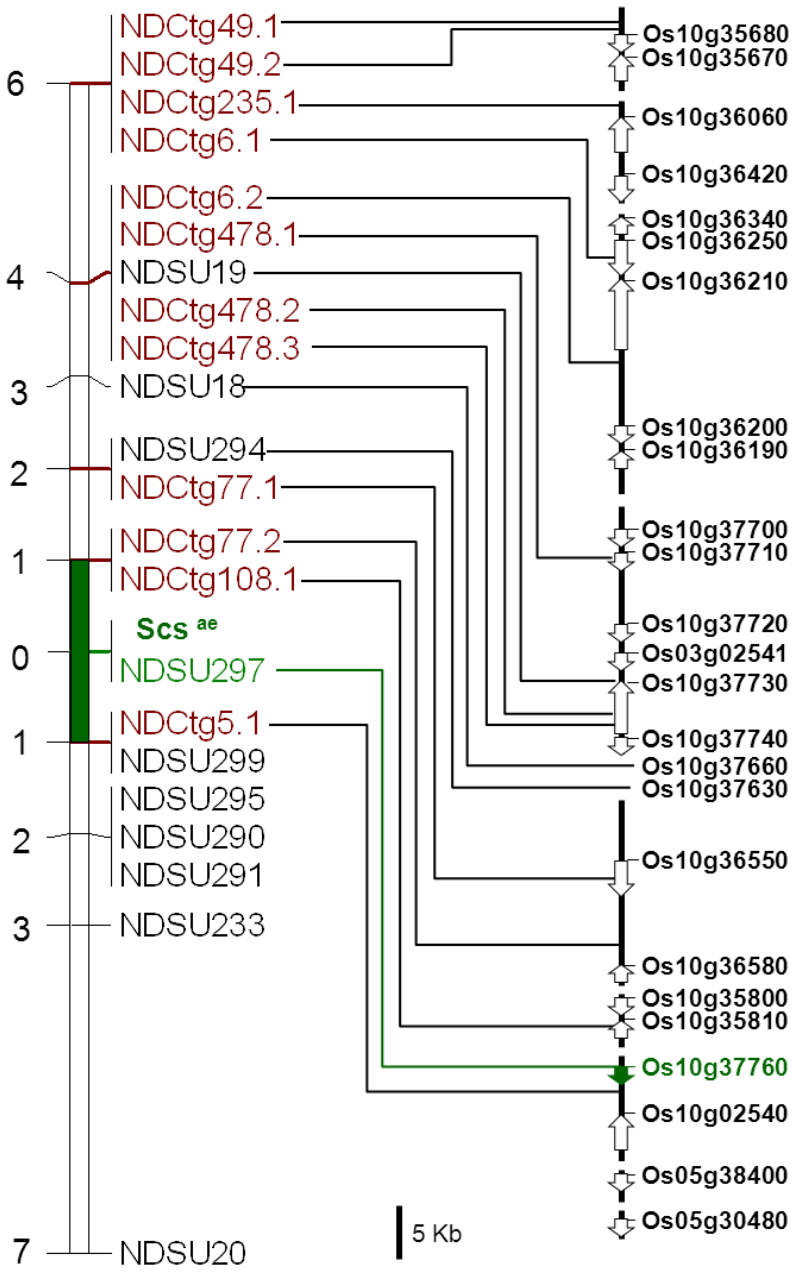


Figure 5.7. Radiation hybrid mapping of the Illumina contigs and their association with the *scs*$^{ae}$ locus. The number to the left of the map indicates the number of breakages that exist between the map locus and the *scs*$^{ae}$ position. The cosegregating marker is depicted in green. The NDCtg are emphasized in dark red. The Illumina contigs are represented to the right as vertical black line, and the length of the line is proportional to their actual physical length as for the legend. Genes are represented by vertical white arrows pointing in the direction of their orientation. The actual position amplified by the markers is indicated on the contigs by mean of horizontal elbow lines.

It is also possible that these genes are silenced by methylation on the 1D chromosome. If this was the case, the methylation-sensitive enzymes used to produce the Illumina libraries would not cut on their sequence, preventing them from being sequenced and recognized by bulk segregant analysis, while DNA-based markers that are not sensitive to methylation would still be able to amplify the genes. The digestion protocol employed to produce the Illumina libraries cleaves only the hypo-methylated genes, which are the active genes and hence most likely to represent the targeted locus. Still, it must be considered that the methylation status of a gene changes in different tissues and under different environmental conditions. It is then extremely important when creating the bulks to assure collecting the right target tissue, possibly while exposing it to the right target environmental conditions. These same considerations were made by Trick et al (2012) in their RNASeq bulk segregant experiment. In their case their gene of interest was later found to be only expressed after anthesis and it could not be identified in the leaf tissue that was used to harvest mRNA in their experiment.

Fortuitously, this was not the case for our bulk$^{POS}$ since three of the pooled 'positive' $RH_1$ lines were actually alloplasmic for the *Ae. longissima* cytoplasm (Hossain et al. 2004a). It can then be expected that the procedure described here would preferentially recognize only those genes that were actively expressed in the leaves of 15 days old alloplasmic lines. Since the *scs*$^{ae}$ gene is believed to be expressed in all tissues at all stages (see Appendix G), the methylation sensitivity of the enzyme used to create the libraries provided an unexpected but welcome advantage by excluding all the inactive genes from the list of potential candidates. However, for future BulkSeq DNA experiments I would suggest employing methylation sensitive enzymes only when the expression behavior of the gene of interest is rather known or can be inferred, otherwise methylation insensitive enzymes should be preferred.

### 5.3.7. Gapped contig assembly to determine the microsynteny conservation around *scs*$^{ae}$

The gapped genes scaffold of the *scs*$^{ae}$ region on 1D provides an opportunity to look at the synteny relationships that exist in close proximity of this locus. The tree model species rice, sorghum, and *Brachypodium* show a high level of gene conservation in this region, even though their order is partially rearranged (Figure 5.8). Two major inversions exist between the location of the genes in sorghum and

rice. These two regions of conserved synteny are separated by a large gap in sorghum between positions Sb01g018050 and Sb01g31210, which corresponds to nearly 1,200 genes. *Brachypodium* and rice have more rearrangements, with two inversion and at least two evident translocations. However, there are no obvious gaps in their synteny conservation, with all the *Brachypodium* genes recognizing an orthologus of rice within the analyzed interval. The synteny, between the genes that have been mapped within the *scs*$^{ae}$ region and their orthologus in rice, resembles what is observed between rice and *Brachypodium*. A major inversion can be observed for the same genes that are inverted between rice and *Brachypodium*, but the other loci seem to identify good colinearity with rice, with just minor rearrangements in their order (Figure 5.9). As discussed in Chapter 4, the synteny in the *scs*$^{ae}$ region is well conserved between the grasses, with just rearrangements in the order of genes, but not in the genes content. The average gene colinearity between rice and sorghum was calculated at 57.8% (Paterson et al. 2009), while the loci considered here show a nearly 100% colinearity between the two species and *Brachypodium*. These observations suggest an evolutionary importance for this region that prevented it from rearranging over more than 50 million years of separate speciation (Paterson et al. 2009). A similar conclusion was reached (with some extensions) by Maan in his "*scs* hypothesis" (Chapter 4). He suggested that the *scs* genes were responsible to guide the speciation of the grasses (*Triticeae* in the author's original hypothesis) by allowing only determined intercrosses between the funding progenitors, depending on their specific cytoplasms and the compatibilities provided by their nuclear-encoded scs genes. Interestingly, the genomic region surrounding the *scs*$^{ae}$ locus is particularly rich of genes targeting the mitochondria (Chapter 4). It was suggested (Meyer 2012) that plastidial genes tend to integrate in clusters at those positions were chromosome paleofusions occurred, and then become functional part of the host genome, providing nuclear encoded proteins that are fundamental for the life of the plastids. The *scs*$^{ae}$ region is extremely close to the location of a paleofusion and it has an over-abundance of mitochondrial targeting genes, corroborating that hypothesis.

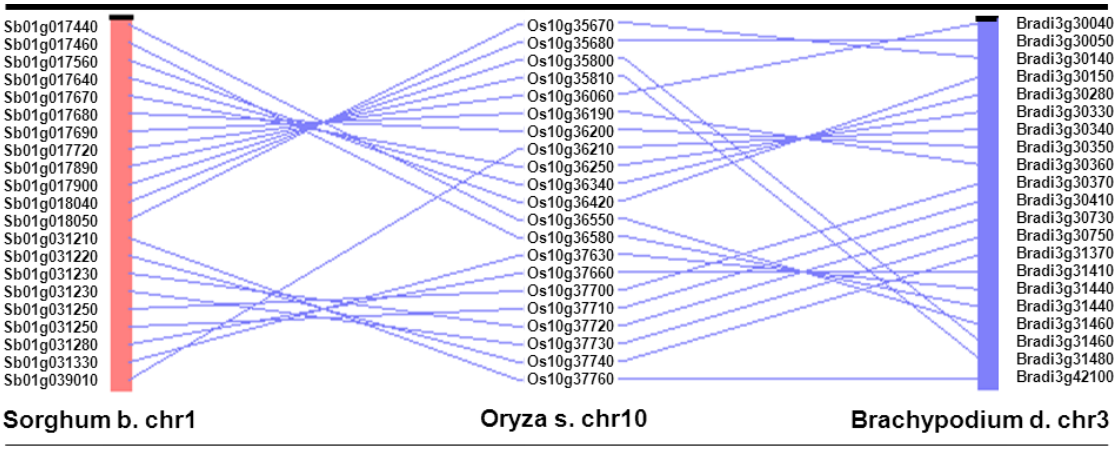


Figure 5.8. Synteny relationships between sorghum, rice, and *Brachypodium* in proximity of the $scs^{ae}$ locus. This figure is adapted from the output of AutoGRAPH (Derrien et al. 2007). Parallel lines indicate gene order conservation between sorghum (Sb), rice (Os), and *Brachypodium* (Bradi) genes. Intersected lines indicate rearrangements in the gene order.

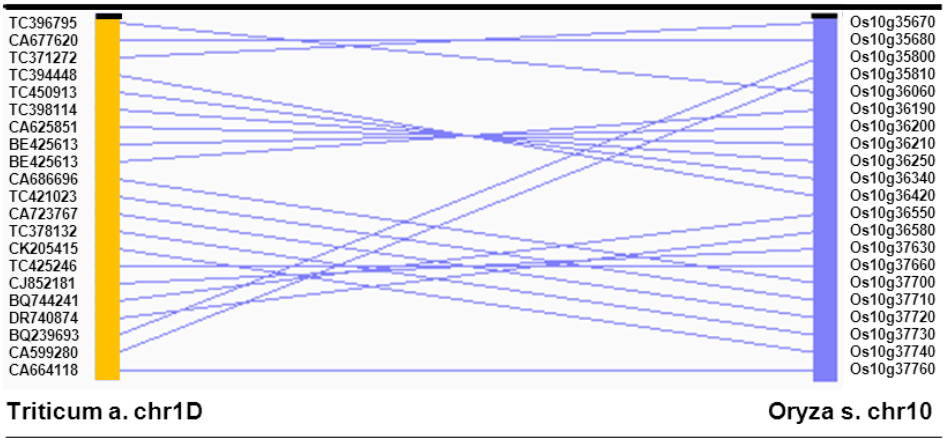


Figure 5.9. Synteny relationships between wheat chromosome 1D and rice in proximity of the $scs^{ae}$ locus. This figure is adapted from the output of AutoGRAPH (Derrien et al. 2007). Parallel lines indicate gene order conservation between *Triticum aestivum* chromosome 1D (left side) and rice (Os) genes. Intersected lines indicate rearrangements in the gene order. The genes of *Triticum aestivum* are indicated in their EST code.

### 5.3.8. The $scs^{ae}$ locus cosegregates with markers ndsu297 tagging a *rhomboid* gene

Contig108 (marker NDCtg108.1) segregates with the phenotype of $scs^{ae}$ with the exception of one breakage, also marker NDCtg5.1 segregates away from the locus for just one break (Figure 5.7).

Marker *Xndsu297*, located within contig 5 but ~4Kb above marker NDCtg5.1, cosegregates in all tested

lines with the *scs*$^{ae}$ locus. A total of seven $RH_1$ lines point unequivocally to this marker as the correct location of the *scs*$^{ae}$ gene. *Xndsu297* tags a gene on chromosome 1D that is orthologus to a *rhomboid* gene in rice, sorghum and *Brachypodium*. Six primer combinations (RHBD1-6) were used to sequence this gene from 207 bp ups*t*ream of its start codon to 367 bp downstream of its stop codon, for a total length of 1,944 nucleotides (Figure 5.10). The full sequence of this gene was obtained from the flow sorted chromosome 1A, 1B, and 1D, as well as from *Ae. tauschii*, (lo)*scs*$^{ti}$ -, the 'Normal' and the 'Weak' genotypes. Annotation of this gene revealed five exons and four introns.

The sequences obtained by PCR sequencing matched precisely what was assembled in contig 5 with the exception of four nucleotides that were present on the contig, but were not found by PCR sequencing. This polymorphism occurred in the 5'UTR region of the gene, and it likely represents the genetic diversity between the *Ae. tauschii* genotype used to generate the survey sequence database and the genotype used here for PCR sequencing. The full FASTA sequences of the gene, its CDS and amino acids are available in Appendix G.

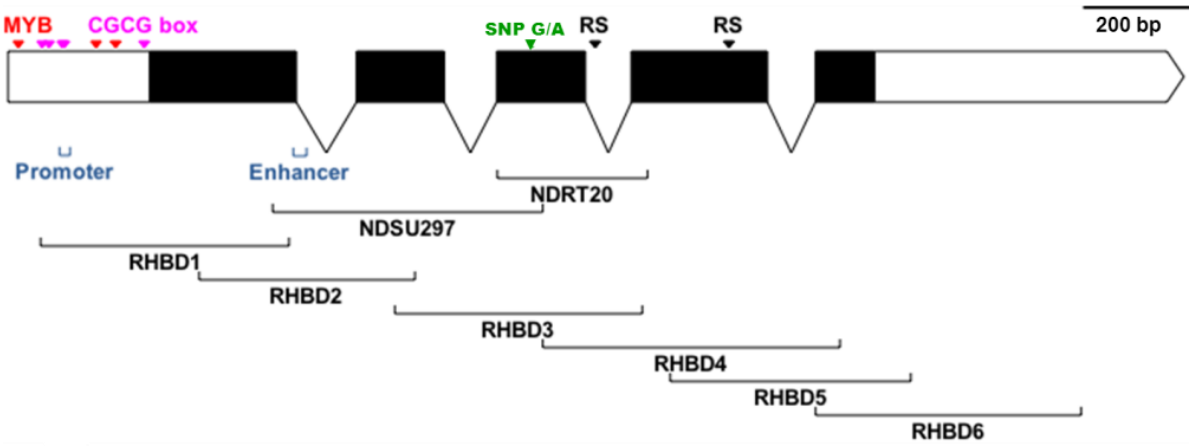


Figure 5.10. The *rhomboid* gene tagged by marker ndsu297. The exons of the coding sequence are represented in black, the hypothetical 3'UTR and 5'UTR are shown in white. Introns are represented by 'V' shaped black lines. The amplicons generated by the RHBD1-6, NDSU297, and NDRT20 primers are indicated. The putative regions corresponding to the promoter and enhancer are color coded in blue. Above the gene are provided the location of the functional SNP guanine (G) to adenine (A), the restriction site (RS) digested by *Pst* I and *Aat* II, and the regions recognized by known transcription factors color coded in red for MYB box and pink for CGCG box.

An alignment study between all sequences revealed a single SNP in the coding sequence conserved in all genotypes harboring a *scs* gene (*scs*$^{ae}$ or *scs*$^{ti}$). The functional SNP carried by the *scs* bearing lines is a guanine (G) at position 997 of the gDNA, while the line susceptible to the alloplasmic (lo) state have an adenine (A) (Figure 5.11). The existence of a functional SNP provides a partial confirmation that this locus might indeed represent the *scs*$^{ae}$ gene. This SNP results in a change of amino acid sequence from an arginine (R) to a lysine (K), but its implications at the protein level will be discussed more extensively in section 5.3.8.

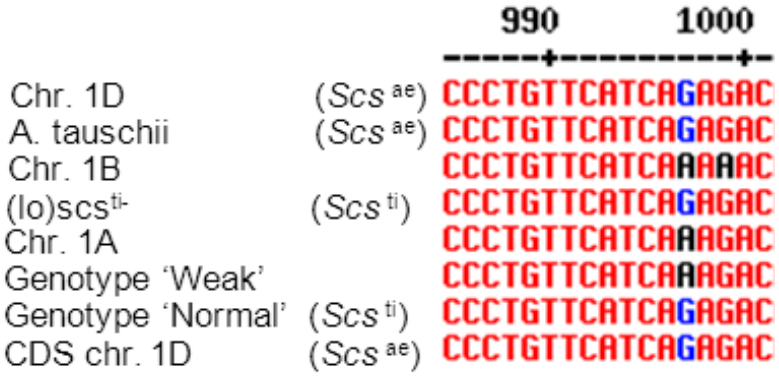


Figure 5.11. Conserved functional SNP on the Rhomboid gene.

5.3.9. Expression analysis of the putative *scs*$^{ae}$ gene

Little is known about the molecular functions of the *scs*$^{ae}$ gene (Chapter 4), except that it is believed to exist in allelic versions on all the homeologs group 1 chromosomes of wheat and its wild relatives (Gehlhar et al. 2005; Asakura et al. 2000; Ohtsuka 1980). Also, since the *scs* genes control nuclear-cytoplasmic compatibility and all cells at all life stages of an organism have the same nucleus and cytoplasm, these loci are expected to maintain their activity throughout the life of the plant in all tissues. The expression profile of this locus was tested using a primers combination (NDRT20) that was designed to span the functional SNP, while generating an amplicon that includes two separate exons (Figure 5.10). The amplification of two exons provides an easy system to detect gDNA amplification, since the genomic fragment produces an amplicon 87 bp (intron) longer than the cDNA fragment. However, NDRT20 does

not provide sufficient specificity to solely target one version of the gene, but rather amplifies all the homeoallelic copies expressed. Sybrgreen qPCR revealed non-significant differences in expression of this gene in embryo or leaf tissues, in alloplasmic and euplasmic lines with various dosages of the *scs* genes (Figure 5.12).

It is interesting to notice that the expression of the two exons marked by NDRT20 is only slightly lower than the housekeeping gene *actin*. Less than two threshold cycles separate the expression of the two, indicating that the loci tagged by NDRT20 are expressed in both embryo and leaves at good levels. Unfortunately, the sybrgreen qPCR did not reveal any difference in expression for this locus between alloplasmic and euplasmic conditions. More important, the NDRT20 marker amplified also in the euplasmic (d)scs$^{-}$ that does not carry the 'G' version of the gene, clearly indicating a lack of allelic specificity. An attempt was made to generate SNP-specific TaqMan probes, but the sequences surrounding the SNP were not ideal for probe design, and all the probes tested did not provide adequate qPCR amplifications. To determine the frequency of the SNP expression, six amplicons of each sample used for NDRT20 amplification were sent for pyrosequencing. This technique of sequencing provides peaks for each nucleotide call that are proportional to the amount of that nucleotide in the template, also defined as real-time DNA sequencing (Royo and Galan 2009). While other methodologies are typically preferred to determine the expression levels of a gene, pyrosequencing has been extensively used in the past to determine mRNA abundances in large scale expression profiling projects (Agaton et al. 2002). In Figure 5.13 are graphically presented the results of this analysis for the functional SNP.

The pyrosequencing analysis was successful. The six biological replicates generated nearly identical chromatographs, with normalized G/A peaks of the same height. Four expression classes could be clearly determined on the basis of the changes in 'G' and 'A' abundance. The presence of at least two copies of *scs* in alloplasmic conditions favors the expression of the functional SNP in its G form, silencing or drastically reducing the expression of its counterpart. The hemizygous (one copy) condition of *scs* in alloplasmic state allows the low expression of the non-functional version of the SNP in its A form. The same state of *scs*, but in euplasmic conditions permits both versions to be equally expressed. Obviously, in the absence of the 'G' version in euplasmic conditions only the 'A' form is expressed.

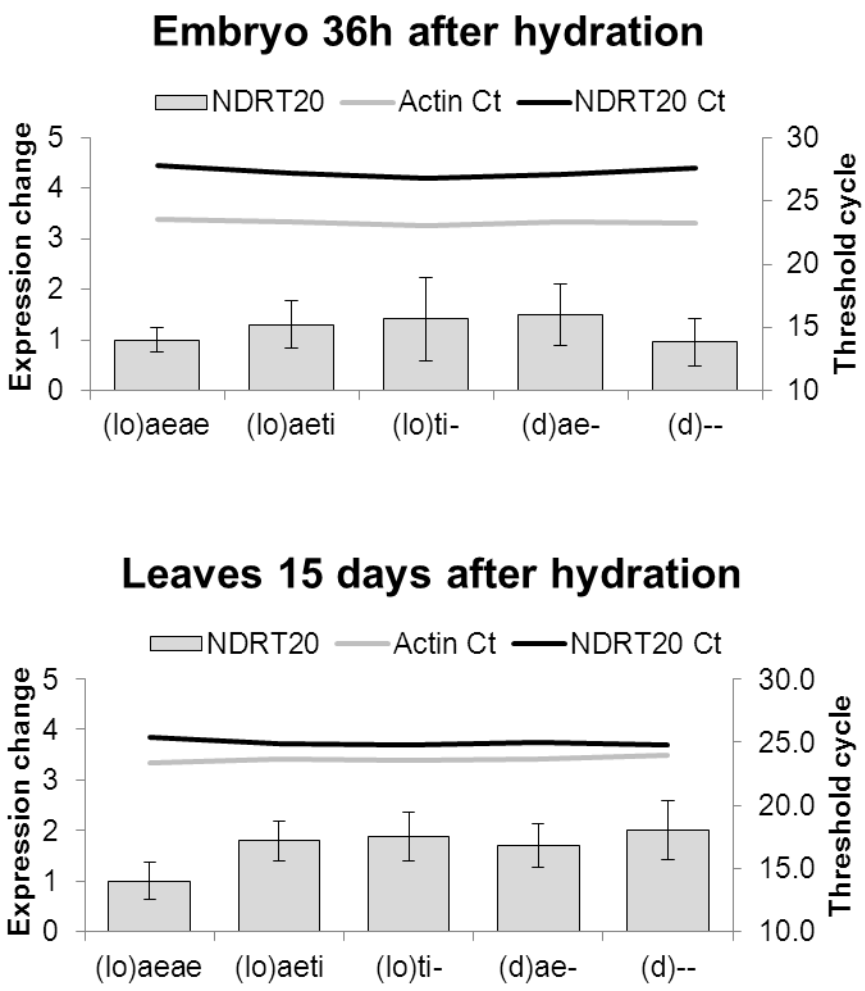


Figure 5.12. Expression analysis of the Rhomboid gene by mean of NDRT20 qPCR. The data are presented as relative quantifications comparison to an Actin gene. The samples are labeled with their alloplasmic (lo) or euplasmic (d) definitions, as well as their dosages of *scs* genes. The expression changes are provided as fold differences of the average of the replicates, and the error bars present their standard deviations. The threshold cycles (Ct) are provided for comparison as color-coded horizontal lines with their values reported on the secondary y axis to the right.

This gene located on contig 5 cosegregates with *scs*$^{ae}$, as revealed by *Xndsu297* RH mapping, and it is also differentially expressed in alloplasmic and euplasmic conditions as suggested by NDRT20 analysis. In the five genotypes considered, the overall quantity of expression of the genes tagged by NDRT20 did not change with the change of cytoplasm (Figure 5.12).

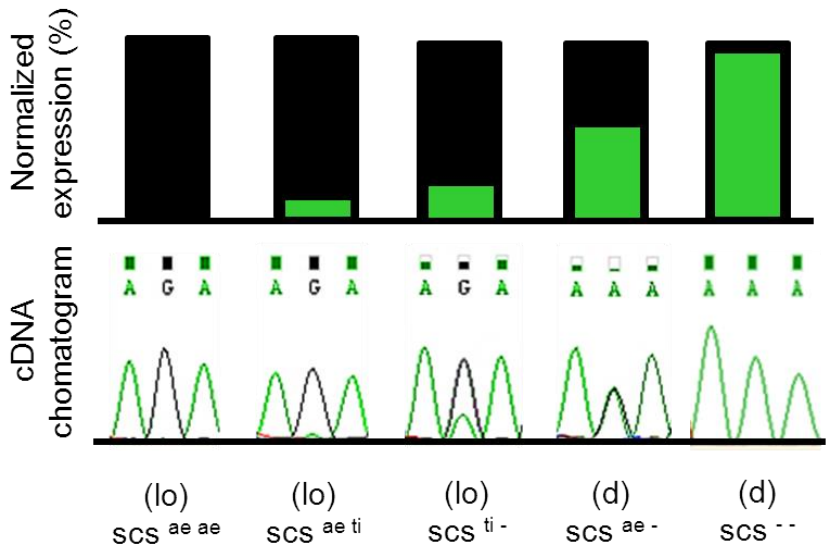


Figure 5.13. Pyrosequencing analysis to determine expression levels of the functional SNP of Rhomboid. The sequencing peaks are reported in their raw format. The bars above represent the relative abundance in the cDNA amplicon of the G nucleotide (black) vs. the A nucleotide (green) at the functional SNP position. The (d)scs$^{- -}$ pyrosequencing was done at a later time and added to this figure for clarity.

However, the type of alleles been expressed did change (Figure 5.13). Clearly, NDRT20 amplifies all the homoeologus versions located on chromosomes 1A and 1B. At least one of these two versions is expressed when the nucleus of durum is immersed in the cytoplasm of durum. The 1D version of this gene, when present, is also expressed in combination with the cytoplasm of durum. However, when the cytoplasm of durum is replaced with the cytoplasm of *Ae. longissima* (alloplasmic) the chromosome 'G' type of this gene is preferred over its 'A' counterparts. These observations were made both in leaf and embryonic tissues, suggesting that the gene tagged by NDRT20 is expressed in different cell types and at various life stages of the plant. Also, microarray experiments previously conducted (Schreiber et al. 2009) revealed that *rhomboid* genes are expressed throughout the life-span of wheat (see Appendix H). Indeed, the expression profile of NDRT20 supports the hypothesis that the gene tagged by *Xndsu297* is the *scs*$^{ae}$ gene, or at least it behaves as it would be expected for that locus. Further, this gene has counterparts in all the homoeologus chromosomes as hypothesized by Maan (1991). Interestingly, the overall expression of NDRT20 did not change significantly between the various genotypes. It appears that the homoeoallelic

versions have a maximum of overall combined expression, as the same overall expression was observed in all tissues/plasmons, which is then traded between the various allelic copies. The alternative copies on the homoeologus chromosomes 1A and 1B have a probable active function in concomitance with the durum cytoplasm, but are likely unable to provide proper compatibility in the cytoplasm of *Ae. longissima*. It is for this reason that their expression is suppressed in the alloplasmic condition in favor of the chromosome 1D-version. This expression behavior seems consistent with what would be expected for a housekeeping gene controlling life-dependent cellular functions.

### 5.3.10. The three-dimensional structure of the rhomboid gene

The coding sequence of the gene tagged by *Xndsu297* has two possible start codons in frame that produce a 326 or a 274 amino acids protein (Figure 5.14). The protein sequence of chromosome 1D and *A. tauschii* is actually three amino acids shorter (323 or 271). The functional G/A SNP causes a change in amino acid at position 173 (position 170 in chromosome 1D and *A. tauschii*).

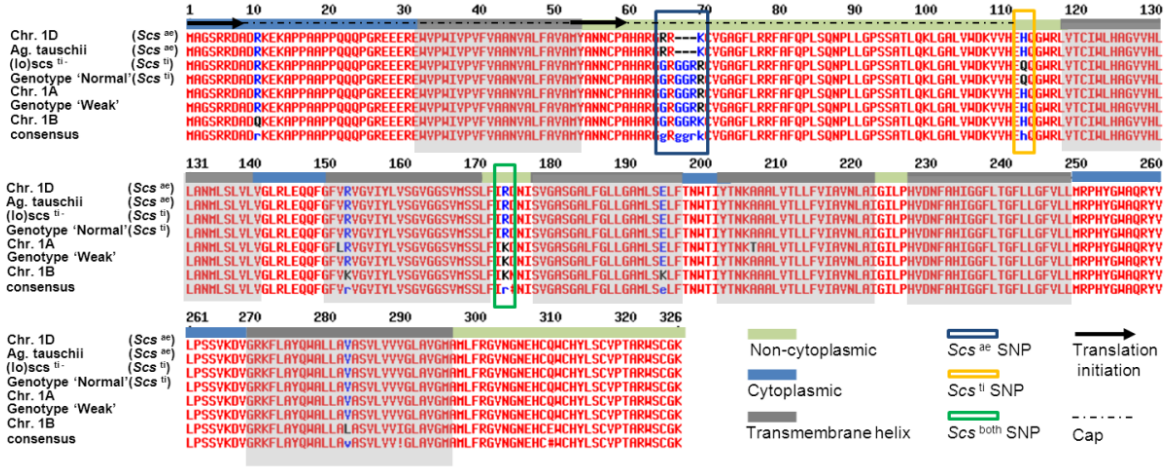


Figure 5.14. Alignment of the amino acid sequences translated from the gene tagged by *Xndsu297*. The black arrows indicate the two possible origin of translation. The regions corresponding to transmembrane helices are shaded in grey, and the inter-transmembrane portions are color coded to indicate if they surface in the cytoplasm or to the internal side of the membrane. The SNPs that produce changes in the amino acid sequence are boxed and color coded. The Cap region is indicated with a dashed black line.

The amino acid sequence corresponds to a membrane protein with seven transmembrane domains (see Appendix I for details). This protein shares strong identity with a peptidase S54 rhomboid (IPR002610) protein. Multiple prediction tools agree that the gene tagged by *Xndsu297* is in fact a rhomboid protein. There are five versions of the rhomboid protein that have been crystalized and for which the threedimensional (3D) structure is known (Lemieux et al. 2007). The rhomboid protein that putatively represents the *scs*$^{ae}$ gene shares 16% of identity with one of these crystalized proteins, sufficient to predict its 3D folding (Figure 5.15).

Rhomboids constitute a large intra-membrane protein family characterized by six or seven transmembrane helices, a cap in loop 1 (L1) between helices one (H1) and two (H2), and a catalytic serine peptidase on helix 4 (H4) (for review see Urban 2006, Freeman 2008, Knopf and Adam 2012). These intra-membrane peptidases have been identified in the lipidic bilayer of various organelles including Golgi, endoplasmic reticulum, and mitochondria. A specific signal peptide for these proteins has yet to be identified, but the few mitochondria examples (defined as PARL-types) that have been characterized to date always presented the seven helices forms, while the forms targeting other organelles showed preferentially a six transmembrane structure (Knopf and Adam 2012). The biological functions of rhomboids range widely, depending on their specific cellular localization and activity. Basically, these peptidases move around in the membrane by means of their L1 cap, which was postulated to act like a floating device (Wang et al. 2007), once a target membrane protein is identified (Schafer A et al. 2010), the rhomboid engulfs it probably through a gate mechanism mediated by L5-H5 (Baker et al. 2007), then uses its serine catalytic site on H4 to cut the membrane anchor of the target substrate, releasing it into the cytoplasm (Ben-Shem et al 2007). The target proteins of rhomboid varies widely, but it has been demonstrated their involvement with many human diseases, in the embryo development of rats where PARL deficient mutants undergo apoptosis, and in intra-cellular regulations by signal emission through the release of transcription factors (for review see Urban 2006, Freeman 2008, Knopf and Adam 2012). In fact, at least 10% of the transcription factors of *Arabidopsis thaliana* are membrane bound and require a mechanism similar to what described for the rhomboid to be freed from the membrane and reach their targets (Seo et al. 2008). The great importance of this family of proteins is

further demonstrated by their presence in nearly all the organisms that have been characterized to date (Freeman 2008). Indeed, their functional and structure conservation across kingdoms is so vast that *rhomboids* have been proposed as the result of vertical evolution from an ancestral gene originated in the last common eukaryotic progenitor at the dawn of life on Earth (Lemberg and Freeman 2007).

Two independent software generated nearly identical structure for this rhomboid protein, and testing with other prediction tools confirmed the quality of the 3D prediction (see Appendix I). The first transmembrane helix and the L1 are located at the water-lipid interface and constitute the rhomboid cap, with its characteristic three short α-helices (Wang et al. 2007). However, the translation of this cap depends on the initiation of translation site of the mRNA. In fact, only the first of the two initiation of translation sites (methionine) would originate the cap, while the second would begin the translation of the protein just few amino acids before the second transmembrane helix, without a cap (Figure 5.14). The cap structure is fundamental for the peptidase activity of rhomboids, but there are documented cases of fully functional uncapped versions (Lei and Li 2009). Five (H2 through H6) of the transmembrane helices constitute the central core of this rhomboid protein. The $7^{th}$ helix was excluded by both software from the 3D model. As mentioned, only the mitochondrial rhomboids (PARL-types) have a seven helices structure, while the crystalized version used as template for the 3D model only has six. That is why H7 was excluded from the 3D model, even though it probably exists in nature.

The positively and negatively charged residues are concentrated in the non-intermembrane portions of the protein, providing signal-responsive sites at both sides of the membrane. The charged residues on the surface typically represent the ammino acids that interact with other proteins. The cellular orientation (cytoplasmic and non-cytoplasmic) of the protein was determined with Phobius. Also, as typical of intra-membrane proteins, this peptide has a long half-life in solution equal to 220 hours, practically the entire length of a plant cell life (De Veylder et al. 2007).

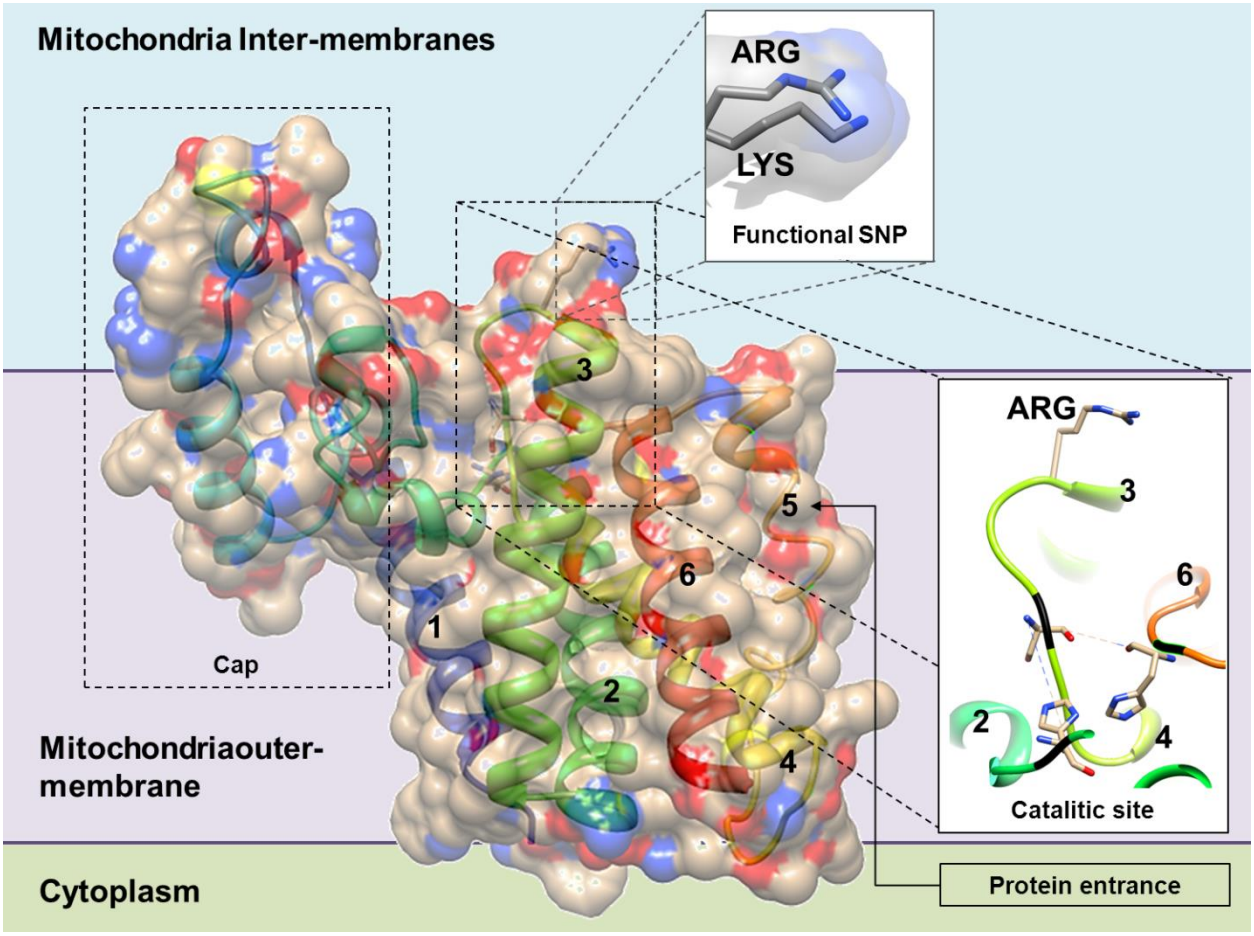


Figure 5.15. The 3D structure of the rhomboid protein. The transmembrane helices are visible through the transparent protein surface and are numbered from 1 to 6. The red and blue colors along the surface indicate negatively and positively charged regions, respectively. The probable position of the membrane is indicated with a violet background. Also, the cytoplasm and the inside of the organelle are shown in the figure. The functional change between an arginine (ARG) and Lysine (LYS) is shown, together with a simplified vision of the catalytic site and the position of the amino acid change (ARG) that can influence it. The cap and the place for the entrance of the proteins are hypothesized on the basis of the protein 3D structure and abundance of charged sites.

An identity comparison with previously annotated proteins determined that the *rhomboid* tagged by *Xndsu297* is rather localized in the endoplasmic reticulum or in the mitochondria (see Appendix I). However, a blastP search revealed as best match *Arabidopsis thaliana* gene At05g07250, which was identified as a mitochondrial-bound *rhomboid* gene (Kmiec-Wisniewska et al. 2008). This, together with its PARL-type seven helices structure, suggests that the *rhomboid* gene tagged by *Xndsu297* is indeed

localized on the mitochondria membrane. Its integration in the mitochondrial membrane provides an additional piece of evidence to support the hypothesis that this gene is indeed *scs*$^{ae}$. In fact, while all other organelles in the cytoplasm are encoded by the nucleus, the mitochondria and chloroplast are directly encoded by their internal genome (with minor exceptions). In alloplasmic condition, where the nucleus remains the same and just the cytoplasm changes, the mitochondria and chloroplast should be the only organelles to differ as compared to the euplasmic condition. Hence, the *scs*$^{ae}$ gene was hypothesized to interact directly with the mitochondria or chloroplast (Chapter 4), and as such the localization of the *rhomboid* to one of these two organelles corroborates the claim of it being the *scs*$^{ae}$ locus.

The active peptidase site (ID: 1VZVCAT) is located at the core of the rhomboid protein. The amino acid triad that generates this site is composed by two histidines at positions 124 (121 in 1D and *A. tasuchii* versions) and 224 (221), and a serine at position 177 (174). These three amino acids are located at the turning points between H2, H6, and H4, respectively. The peptidase activity of the rhomboid proteins is mediated through the depletion by one of the histidine of a proton from the catalytic serine, which becomes then a reactive nucleophile capable of disrupting the target substrate (Freeman 2008).

Based on the orientation of the rhomboid protein, the peptidase site is located at the bottom of the protein. The transmembrane helices appear more closely grouped at the catalytic site, while diverge toward the cytoplasmic side of the protein, creating a sort of internal V-shaped lumen (Urban 2006). The entrance for the peptides that need to be cleaved is expected to be through the L5-H5 gate (Baker et al. 2007). Directly above the catalytic site, the five core helices have numerous charged residues facing outward. The peptides anchored to the membrane could then enter through the L5-H5 gate, maybe mediated by the interaction with the charged residues on the rhomboid surface, then enter the V-shaped lumen until their target site reaches the catalytic serine, where peptidase cleavage would occur, and finally be released back into the cytoplasm or become activated membrane proteins.

### 5.3.11. The hypothetical function of the rhomboid protein as *scs*$^{ae}$ gene

A non-synonymus SNP was identified in the coding sequence of this gene (Figure 5.11) and it was shown to be differentially expressed in alloplasmic and euplasmic conditions (Figure 5.13). The SNP

translates into an arginine in the protein of the *scs*$^{ae}$ or *scs*$^{ti}$ bearing genotypes, and into a lysine in the non-*scs* bearing genotypes (Figure 5.14). Both these amino acids are basic and present two and one positively charged N residues, respectively. Basic to basic ammino acid mutations do not appear to cause major changes in the overall protein structure, however an arginine to lysine substitution has proven multiple times to be the key factor that caused drastic functional modifications in different types of genes (Klumpp et al. 2003; Betts and Russell 2003; Mills et al. 2005; Ryan and O'Fagain 2007; and more)

The 3D structure places this change in amino acids on the non-cytoplasmic surface of the rhomboid protein (Figure 5.15). The charged residues hang out of the rhomboid surface toward the internal lumen of the mitochondria, apparently probing the solution inside the organelle for signals. Interestingly, the amino acid change occurs at position 173 (170), which represents the actual beta turning site between H3 and H4. The peptidase activity of this rhomboid protein is mediated by a serine at position 177 on H4, and two histidines on H2 and H6. A change in the distances between H2, H4 and H6 is likely to affect the activity of this cleavage domain (Figure 5.15). It can be hypothesized that a conformation change on the amino acid responsible for the beta turn of the transmembrane spirals would indeed modify the angle of the turn, and ultimately bring the helices closer or further apart from each other. Helix four does not entirely traverse the membrane and remains partially free to move (Knopf and Adam 2012). It is then possible that the amino acid change that occurs at the beta turn of H4 might affect its 3D position, in relationship with H2 and H6, and have an effect on the activity of the peptidase catalytic site. Hence, the arginine to lysine change has the potential to affect the peptidase activity of the rhomboid protein by modifying the distances between the three helices harboring the catalytic triad. Further, this amino acid change is located and oriented in such a way that appears to probe the organellar solution.

Assuming that the *rhomboid* is indeed the *scs*$^{ae}$ gene, it could be hypothesized that in alloplsmic conditions the mitochondria of *Ae. longissima* has different internal solutions or ionic charges/flux than the durum mitochondria, that this organellar solution is sensed differently by the arginine or lysine, and that their differential responses cause a change in the peptidase activity of this protein, which only in the case of the arginine provides viable nuclear-cytoplasmic interactions with the *Ae. longissima* cytoplasm.

An additional evidence in support of the *rhomboid* as the *scs*$^{ae}$ gene is the similarity between the effects on the developing embryo of removing *scs*$^{ae}$ in alloplsamci wheat and PARL knockout mutants in

rats (Cipolat et al. 2006). The knockout of the mitochondrial rhomboid PARL prevented the proper solubilization of OPA1, which in turn triggered PCD and artophyzation of the embryonic cells. This phenotype is surprisingly similar to the VPCD observed at the microscopic level for *scs*$^{ae}$ deficient shriveled seeds (Figure 5.5).

Finally, the rhomboid genes family has been suggested to have evolved in a phyletic mode from an ancestral form that dates to the last common eukaryotic progenitor (Lemberg and Freeman 2007). The "Maan's *scs* hypothesis" would find renewed support in this knowledge, since evolutionary conserved genes are usually involved in the speciation of organisms by means of accumulating mutations (Rieseberg and Blackman 2010).

Indeed, a *rhomboid* gene perfectly fits the expectations for the *scs*$^{ae}$ locus. The mitochondria targeting, its ability to affect embryonic development, its known involvement in intra-cellular communications, and the evolutionary conservation make this gene the ideal candidate for the *scs*$^{ae}$ function.

## 5.4. Conclusions

The BulkSeq analysis proved a valuable tool when used in combination with RH. Nearly one million marker sites were easily and cheaply investigated by sequencing. Contigs harboring genes were rapidly detected by *in silico* analysis, and the simplicity of developing PCR markers for RH allowed to effortlessly generate a gapped-sequence scaffold of the *scs*$^{ae}$ region. A *rhomboid* was identified as the only gene cosegregating with the *scs*$^{ae}$ locus, and re-sequencing revealed a functional SNP differentially expressed in alloplasmic and euplasmic conditions. Further, this SNP causes an amino acid change that might affect the biological activity of the rhomboid protein in response to a difference in the host mitochondria. More targeted and detailed studies are necessary to fully confirm the hypothesized ability of the *rhomboid* gene to restore compatible nuclear-cytoplasmic interactions. Still, the data gathered here present a strong case for the *rhomboid* function in alloplasmic wheat.

5.5. References

Agaton C, Unneberg P, Sievertzon M, Holmberg A, Ehn M, et al (2002) Gene expression analysis by signature pyrosequencing. Gene 289(1-2): 31-39

Asakura N, Nakamura C, Ohtsuka I (2000) Homeoallelic gene Ncc-tmp of Triticum timopheevii conferring compatibility with the cytoplasm of Aegilops tauschii in the tetraploid wheat nuclear background. Genome 43:503-511

Baker RP, Young K, Feng L, Shi Y, Urban S (2007) Enzymatic analysis of a rhomboid intramembrane protease implicates transmembrane helix 5 as the lateral substrate gate. Proc Natl Acad Sci USA 104(20):8257-8262

Ben-Shem A, Fass D, Bibi E (2007) Structural basis for intramembrane proteolysis by rhomboid serine peptidase. Proc natl Acad Sci USA 104:462-466

Betts MJ, Russell RB (2003) Amino acid properties and consequences of substitutions. In: Barnes MR, Gray IC (eds) Bioinformatics for Geneticists. Wiley

Bosca S, Knauer S, Laux T (2011) Embryonic development in Arabidopsis thaliana: from the zygote division to the shoot meristem. Front Plant Sci 2:93

Cipolat S, Rudka T, Hartmann D, Costa V, Sernells L, et al (2006) Mitochondrial rhomboid PARL regulates cytochrome c release during apoptosis via OPA1-dependent cristae remodeling. Cell 126:163-175

Cisar G, Cooper DB (2002) Hybrid wheat. In: Curtis BC, Rajaram S, Gomez Macpherson H (eds) Bread wheat: improvement and production. FAO, Rome, Italy pp 157-174

Collazo C, Chacon O, Borras O (2006) Programmed cell death in plants resembles apoptosis of animals. Biotecnologia Aplicada 23:1-10

de Givry S, Bouchez M, Chabrier P, Milan D, Schiex T (2005) Carthagene: multipopulation intergrated genetic and radiated hybrid mapping. Bioinformatics 21:1703-1704

De Veylder L, Beeckman T, Inze D (2007) The ins and outs of the plant cell cycle. Nature Reviews Molecular Cell Biology 8:655-665

Derrien T, Andre C, Galibert F, Hitte C (2007) AutoGRAPH: an interactive web server for automating and visualizing comparative genome maps. Bioinformatics 23(4):498-499

Dessler C, Rasmussen LJ (2012) Mitochondria in biology and medicine. Mitochondrion 12(4):472-476

Dolezel J, Simkova H, Kubalakova M, Safar J, Suchankova P, et al (2009) Chromosome genomics in the Triticeae. In: Feuillet C, Muehlbauer GJ (eds) Genetics and Genomics of the Triticeae. Springer, New York, pp 285-316

Dominguez F, Moreno J, Cejudo FJ (2012) The scutellum of germinated wheat grains undergoes programmed cell death: identification of an acidic nuclease involved in nucleus dismanting. J Exp Bot 63(15):5475-5485

Dvorak J, Di Terlizzi P, Zhang HB, Resta P (1993) The evolution of polyploid wheats: identification of the A genome donor species. Genome 36:21-31

Faris JD (in press) Wheat domestication: key to agricultural revolutions past and future. In: Tuberosa R (ed) Advances in Genomics of Plant Genetic Resources. Springer.

Freeman (2008) Rhomboid proteases and their biological functions. Annu Rev Genet 42:191-210

Gehlhar SB, Simons KJ, Maan SS, Kianian SF (2005) Genetic analysis of the species cytoplasm specific ($scs^{d}$) derived from durum wheat. Journal of Heredity 96(4):404-409

Hossain KG, Riera-Lizarazu O, Kalavacharla V, Vales MI, Maan SS, Kianian SF (2004a) Radiation hybrid mapping of the species cytoplasm specific ($scsc^{ae}$) gene in wheat. Genetics 168:415-423

Hossain KG, Riera-Lizarazu O, Kalavacharla V, Rust JL, Vales MI, et al (2004b) Molecular cytogenetic characterization of an alloplasmic durum wheat line with a portion of chromosome 1D of Triticum aestivum carrying the $scs^{ae}$ gene. Genome 47:206-214

Kihara H (1944) Discovery of the DD-analyzer, one of the ancestors of *Triticum vulgare*. Agriculture and Horticulture (Tokyo) 19:13-14

Kilian B, Ozkan H, Deusch O, Effgen S, Brandolini A, et al (2007) Independent wheat B and G genome origins in outcrossing Aegilops progenitor haplotypes. Mol Biol Evol 24(1):217-227

Klumpp LM, Mackey AT, Farrell CM, Rosenberg JM, Gilbert SP (2003) A kinesin switch I arginine to lysine mutation rescues microtubule function. J Biol Chem 278(40):39059-39067

Kmiec-Wisniewska B, Krumpe K, Urantowka A, Sakamoto W, Pratje E, Janska H (2008) Plant mitochondrial rhomboid, AtRBL12, has different substrate specificity from its yeast counterpart. Plant Mol Biol 68(1-2):159-171

Knopf RR, Adam Z (2012) Rhomboid proteases in plants – still square one? Physiologia Plantarum 145:41-51

Krattinger S, Wicker T, Keller B (2009) Map-based cloning of genes in Triticeae (Wheat and Barley). In: Muelbhauer GJ, Feuillet C (eds) Genetics and Genomics of the Triticeae. Springer, New York pp 337-357

Lambert C, Leonard N, De Bolle X, Depiereux E (2002) ESyPred3D: Prediction of proteins 3D structures. Bioinformatics. 18(9):1250-1256

Lei X, li YM (2009) The processing of human rhomboid intramembrane serine protease RHBDL2 is required for its proteolitic activity. J Mol Biol 394:815-825

Lemberg MK, Freeman M (2007) Functional and evolutionary implications of enhanced genomic analysis of rhomboid intramembrane proteases. Genome Res 17:1634-1646

Lemieux MJ, Fischer SJ, Cherney MM, Bateman KS, James MNG (2007) Crystal structure of Glpg rhomboid peptidase from Haemophilus influenza. Proc natl Acad Sci USA 104(3):750-754

Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) method. Methods 25:402-408

Maan SS (1991) Nucleo-cytoplasmic genetics of wheat. In: Sasakuma T, Kinoshita T (ed.) Proceedings of the international symposium on nuclear and organeller genetics of wheat, Hokkaido University, Sapporo, Japan, pp 75-l94

Matsuoka Y (2011) Evolution of polyploidy Triticum wheats under cultivation: the role of domestication, natural hybridization and allopolyploid speciation in their diversification. Plant Cell Physiol 52(5):750-764

Mayer K (2012) Same, same but different: complementary analytical approaches highlight the different shades of polyploidy in rye and wheat. International Triticeae Mapping initiative meeting, Fargo, ND, June 25-29

Michelmore RW, Paran I, Kesseli RV (1991) Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. Proc Natl Acad Sci USA 88(21)9828-9832

Mills DA, Geren L, Hiser C, Schmidt B, Durham B, et al (2005) An arginine to lysine mutation in the vicinity of the heme propionates affects the redox potential of the heme associated electron and proton transfer in cytochrome C oxidase. Biochemistry 44(31):10457-10465

Ohtsuka I (1980) Function of a D genome chromosome on the compatible relation between wheat genomes and Aegilops tauschii cytoplasm. Wheat Inf Serv 52:23-28

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, et al (2009) The Sorghum bicolor genome and the diversification of grasses. Nature 457:551-556

Pelletier G, Budar F (2007) The molecular biology of cytoplasmically inherited male sterility and prospects for its engineering. Current Opinion in Biotechnology 18:121-125

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, et al (2004) UCSF Chimera: a visualization system for exploratory research and analysis. J Comput Chem 25(13):1605-1612

Poland JA, Brown PJ, Sorrells ME, Jannink J-L (2012) Development of a high-density genetic maps of barley and wheat using a novel two-enzymes genotyping-by-sequencing approach. PLoS ONE 7(2):e32253

Quackenbush J, Cho J, Lee D, Liang F, Holt I, et al (2001) The TIGR gene indices: analysis of gene transcript sequences in highly sampled eukaryotic species. Nucleic acids Res 29:159-164

Rieseberg LH, Blackman BK (2010) Speciation genes in plants. Annals of Botany 106:439-455

Royo JL, Galan JJ (2009) Pyrosequencing for SNP genotyping. Methods Mol Biol 578:123-133

Ryan BJ, O'Fagain C (2007) Arginine-to-lysine substitutions influence recombinant horseradish peroxidase stability and immobilization effectiveness. BMC Biotechnology 7:86

Salvi S, Tuberosa R (2005) To clone or not to clone plant QTLs: present and future challenges. TRENDS in Plant Science 10(6):297-304

Sarkar P, Stebbins GL (1956) Morphological evidence concerning the origin of the B genome in wheat. Am J Bot 43:297-304

Schafer A, Zick M, Kief J, Steger M, Heide H, et al (2010) Intramembrane proteolysis of Mgm1 by the mitochondrial rhomboid protease is highly promiscuous regarding the sequence of the cleaved hydrophobic segment. J Mol Biol 401:182-193

Schneeberger K, Weigel D (2011) Fast-forward genetics enabled by new sequencing technologies. Trends in Plant Sci 16(5)282-288

Schreiber AW, Sutton T, Caldo RA, Kalashyan E, Lovell B, et al. Comparative transcriptomics in the Triticeae. BMC Genomics 10:285

Seo PJ, Kim SG, Park CM (2008) Membrane-bound transcription factors in plants. Trends in plant Sci 13:550-556

Taketa S, Choda M, Ohashi R, Ichii M, Takeda K (2002) Molecular and physical mapping of a barley gene on chromosome arm 1HL that causes sterility in hybrids with wheat. Genome 45:617-625

Trick M, Adamski NM, Mugford SG, Jiang C-C, Febrer M, Uauy C (2012) Combining SNP discovery from next-generation sequencing data with bulked segregant analysis (BSA) to fine-map genes in polyploidy wheat. BMC Plant Biology 12:14

Tsaftaris AS, Polidoros AN (2000) DNA methylation and plant breeding. Plant Breed rev 18:87-176

Tsunewaki K (2009) Plasmon analysis in the triticum-aegilops complex. Breeding Science 59:455-470

Urban S (2006) Rhomboid proteins: conserved membrane proteases with divergent biological functions. Genes dev 20:3054-3068

van Doorn WG, Woltering EJ (2005) Many ways to exit? Cell death categories in plants. TRENDS in Plant Science 10(3):1360-1385

van Doorn WG, Beers EP, Dangl JL, Franklin-Tong VE, Gallois P, et al (2011) Morphological classification of plant cell deaths. Cell Death and Differentiation 18:1241-1246

Vanyushin BF, Zamyatnina VA, Aleksandrushkina NI (2004) Apoptosis in plants: specific features of plant apoptotic cells and effect of various factors and agents. International review of Cytology 233:135-179

Wang Y, Maegawa S, Akiyama Y, Ha Y (2007) the role of the L1 loop in the mechanism of rhomboid intramembrane proteases GlpG. J Mol Biol 374:1104-1113

Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden T (2012) Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics 13:134

You F, Huo N, Deal K, Gu Y, Luo M, et al (2011) Annotation-based genome-wide SNP discovery in the large and complex Aegilops tauschii genome using next-generation sequencing without a reference genome sequence. BMC Genomics 12:59

# APPENDIX A

## A.1. Iterative framework map imposed on Carthagene

The RH population used in this study exhibited an average deletion size of ~20 Mb, with generally three deletions per line. Assuming perfect distribution of deletions across the entire 993 Mb of chromosome 3B, a minimum of 50 RH lines (50 × 3 deletions × 6.6 Mb ~ 993 Mb) would be required to cover the entire physical size of the chromosome. Assuming non-perfect conditions, where each deletion is not unique, larger population possibly double this size or more would be required to generate a comprehensive RH map. The very advantage of using small deletions to obtain high level of map resolution is also a drawback for RH mapping. Genetic mapping relies on large recombination blocks to order marker loci. Thus markers physically distant on the chromosome can show linkage and be correctly mapped. RH mapping instead employs very small 'deletion blocks' (~10 Mb in size) to link markers. Hence, two markers that are 100 Mb apart can possibly be connected by one recombination block (i.e. one recombinant line) but would require at least 10 perfectly distributed deletion blocks.

To overcome this limitation, I gathered or produced physical information for 115 of the genotyped markers (anchor markers). Initially, I created a framework map for the 115 anchor markers. This map was generated using Carthagene "build", "annealing", "flips", and "polish" functions and then hand curated to assure that all the biological evidence were respected, such as bin-location or contig assignment. The remaining 426 markers were merged on the framework map using the "buildfw" function. This was achieved through iterative analysis. Markers were assigned to any interval between two anchor markers using the command "buildfw" with LOD score of 10. These assigned markers, plus the two initial inner anchor markers, and two outer anchor markers of the framework map, were used for mapping in Carthagene using the commands "build-10", "annealing", "flips", and "polish". All the markers mapping in between the two anchor markers underlying a specific interval were merged into the new framework map; all other markers were discarded and reused in the following iteration. For instance if 15 markers were assigned the location between markers b and c these 15 markers and markers b and c, along with marker b's outer neighbor (assume marker name is a) and marker c's outer neighbor (assume marker name is d)

are mapped. I only accept the list of marker orders that satisfy the condition: a, b, <markers list>, c, d. any other marker that does not satisfy this condition will be discarded at this stage, and tested again in the following iteration. The new framework map, containing both the anchor markers and the most associated non-anchor markers, was used in a second iteration to merge the markers not yet incorporated. A total of ten iterations were necessary to map all the 426 non-anchored markers (Table A.1). To determine the quality of this approach, it was initially compared to normal RH mapping using Carthagene “build”, “annealing”, “flips”, and “polish” functions, without iterative framework mapping. The on-line version of AutoGraph (Derrien et al. 2007) (available at http://autograph.genouest.org/) was used to graphically compare the two RH maps generated with and without iterative framework mapping to a good quality genetic map available in the literature (Figure A.1). Since the iterative approach showed better marker order conservation with the genetic map, I consider iterative framework mapping a good strategy for RH genotyping data.

Table A.1. Details of single iteration for iterative Carthagene frame-work mapping algorithm to produce the 3B-RH map

| Iteration | Mapped markers | Markers left | Map quality | Total Map Size |
|---|---|---|---|---|
| ---------- | No. ---------- | | -- log10-likelihood -- | ------- cR ------- |
| 0 | 128 | 413 | -379.2 | 1010.7 |
| 1 | 219 | 322 | -506.86 | 1322.7 |
| 2 | 292 | 249 | -568.58 | 1466.6 |
| 3 | 312 | 229 | -585.12 | 1501.9 |
| 4 | 380 | 161 | -609.09 | 1568.3 |
| 5 | 430 | 111 | -629.04 | 1612.6 |
| 6 | 470 | 71 | -641.95 | 1650.3 |
| 7 | 503 | 38 | -668.01 | 1713.5 |
| 8 | 520 | 21 | -678.26 | 1730.3 |
| 9 | 528 | 13 | -699.81 | 1800.3 |
| 10 | 541 | 0 | -727.23 | 1871.9 |

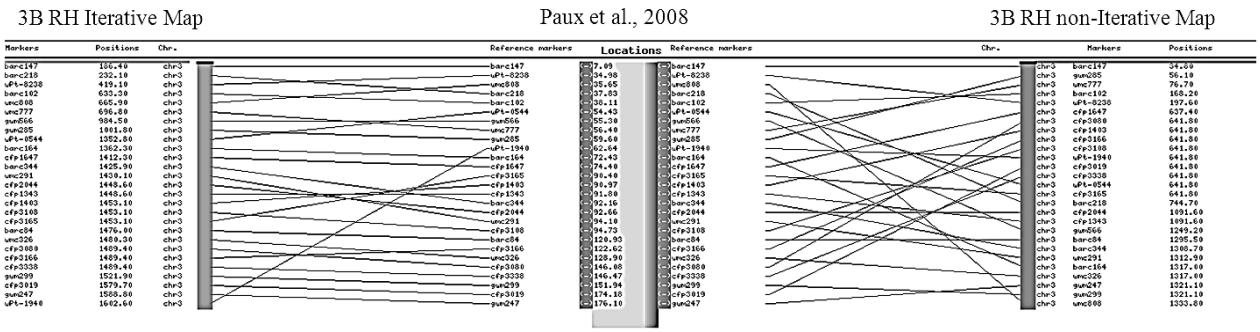


Figure A.1. AutoGRAPH output comparing one genetic map (Paux et al. 2008) to two RH maps produced from the same dataset but applying two different mapping algorithms, the newly developed Iterative-Framework Map and the classical Carthagene. Good markers order conservation is indicated by parallel lines.

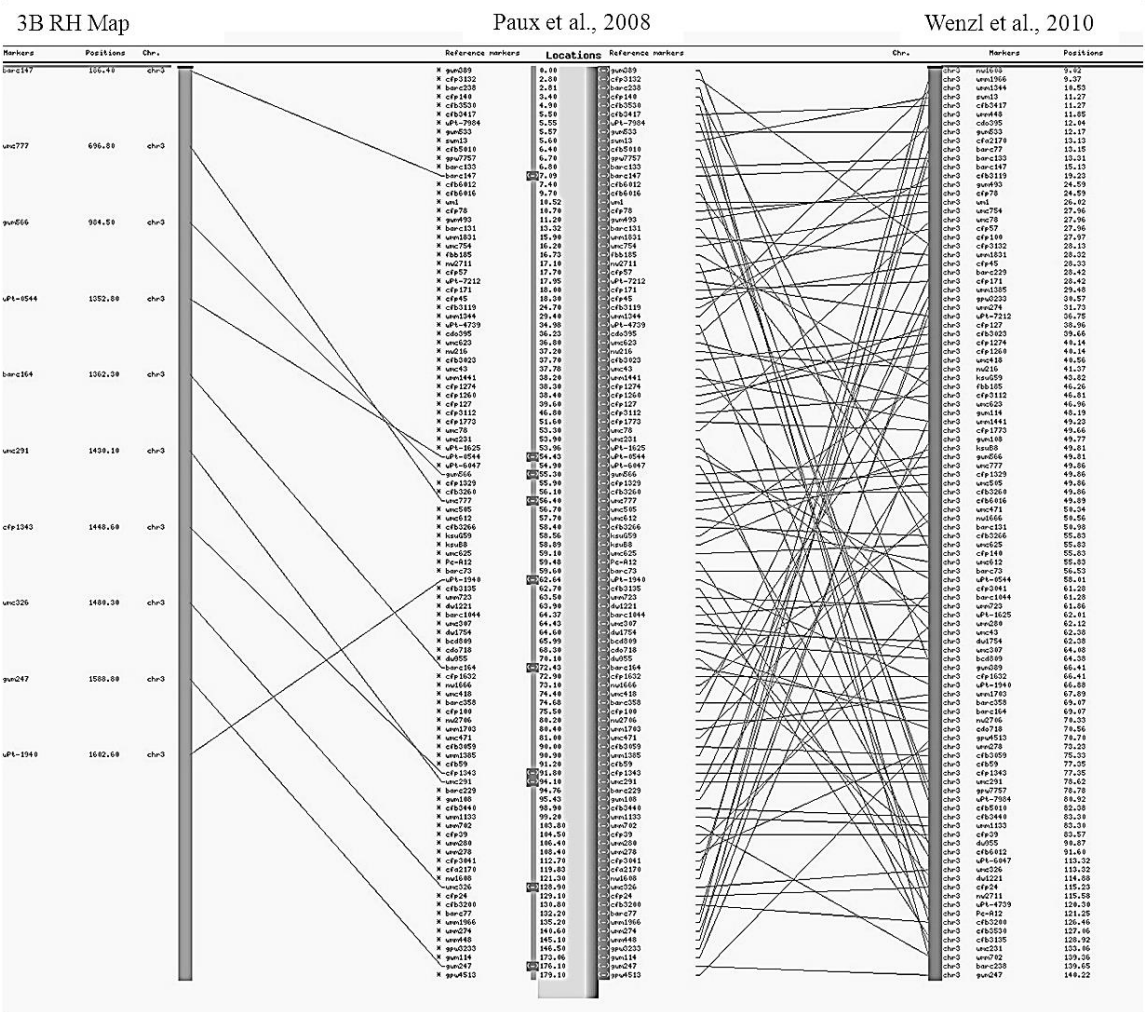


Figure A.2. AutoGRAPH output comparing two genetic maps (Paux et al. 2008; Wenzl et al. 2010) to 3B-RH. Good markers order conservation is indicated by parallel lines.

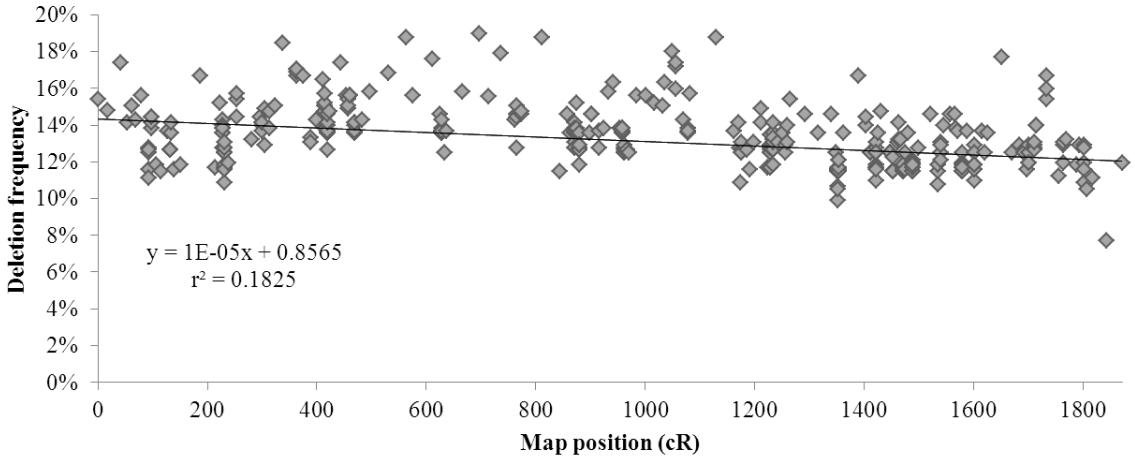


Figure A.3. Deletion frequency distribution throughout the 3B-RH map considering all of the 541 markers. Data points are represented in their mapping position along chromosome 3B.

# APPENDIX B

B.1. SAS code for RCBD in 4 years 1 location 3 replications (blocks) for Seeds per Spikelet (SpS) and Pollen Vitality (PV)

```
options pageno=1;
data multenvRCBD;
input season loc block gen SpS PV;
datalines;
;;
proc print;
title 'Printout of RCBD 4 env SpS and PV';
proc glm data=multenvRCBD;
  class season loc block gen;
  model SpS PV = season|loc|gen
        block(loc*season)
        /ss3 ;
  run;
quit;
```

Table B.1. ANOVA across four seasons (winter 2009, 2010 and spring 2010, 2011) in one greenhouse location for seeds per spikelet (SpS) and pollen vitality (PV)

| | | SpS | | | PV | | |
|---|---|---|---|---|---|---|---|
| SOV | DF | Type III SS | MS | F | Type III SS | MS | F |
| season | 3 | 0.325 | 0.108 | 0.36 | 0.073 | 0.024 | 6.530 |
| loc | 0 | 0.000 | . | . | 0.000 | . | . |
| season*loc | 0 | 0.000 | . | . | 0.000 | . | . |
| gen | 2 | 46.066 | 23.033 | 77.21*** | 1.864 | 0.932 | 249.160*** |
| season*gen | 6 | 2.487 | 0.414 | 1.39 | 0.099 | 0.016 | 4.410 |
| loc*gen | 0 | 0.000 | . | . | 0.000 | . | . |
| season*loc*gen | 0 | 0.000 | . | . | 0.000 | . | . |
| block(season*loc) | 8 | 1.379 | 0.172 | 0.58 | 0.056 | 0.007 | 1.870 |

***, significant at 0.001; loc, location; gen, genotypes

# APPENDIX C

Table C.1. Sequences of primers and their PCR conditions

| Marker ID | Wheat EST | Forward primer | Reverse primer | Annealing |
|---|---|---|---|---|
| NDSU3 | CD454895 | CCAAATTGAAATGAGGTGGCACACg | TCCAGAATGGATTGATGTGAAGTTGGC | 61-56 |
| NDSU31 | TC395692 | ACTGAGGTTGGCTGGGTAACGC | AGGCTGCCTTCTGCTTGGGC | 61-56 |
| NDSU13 | TC359809 | TGAACTGCAGGAGTGTTTCAATTTGGC | GCATGCCCAGCCCAAGAGGG | 61-56 |
| NDSU18 | TC294099 | GGGAATcGAAGTCACAAAACCa | ACCAACCCTGATGTTTGGAGGC | 61-56 |
| NDSU19 | TC306640 | ATGGATAGGAtagAtaGg | CCTCCCAGTGTCCATGTCGGC | 61-56 |
| NDSU20 | TC381152 | TGGCTAGAGAAGAGCAAGGGTAGGG | CAGTCATCACTCGTCTATCTATg | 61-56 |
| NDSU21 | TC321241 | AGCAATTGACAATGTGTTGTGCATACG | GGTCaCCTAGAtTGACttCcGa | 61-56 |
| NDSU200 | BE446672 | ACGGGGAACTGAAGATGATG | GTTCTTGGGATGGCACACTT | 65-60 |
| NDSU201 | BE405749 | CAACTGTGGTGTAACCTGGG | TCCAAATTTAAGAGCCGCAG | 65-60 |
| NDSU202 | BF473056 | CCACCGCTTCATCATCTACA | TTGCTCCTCACGCGTGCA | 65-60 |
| NDSU203 | BE405834 | TGGACCGTGTGCTTTATTCA | CAGAAAGTCGTTCACAGCGA | 65-60 |
| NDSU205 | BE405518 | AAGGCTGGTGACTGGAAGAA | GCTGATGCTCCTTGATCTCC | 65-60 |
| NDSU206 | BF478737 | CTCTTCACAGTTACAACATCAGC | TGAGGCTCAATGATGACCAG | 65-60 |
| NDSU207 | BE425978 | CTGGGAGGGACCACATTCTA | CAATTCCATCGCCTTCATTT | 65-60 |
| NDSU208 | BE445475 | TGTTCGGTTGGGAGACCTAC | AGCAGATACGAGCATGAGCA | 65-60 |
| NDSU210 | BF202643 | GAATAGCAACAGTGCTCATGAAT | GAAGAACAGCAGGGCGTTAC | 65-60 |
| NDSU211 | TC300622 | TACATGGCACAAGTGGACAC | TTTCATTTTCCAGCTCTTGT | 65-60 |
| NDSU212 | BE424100 | ATCCCCTTTAGTTTTATTCTTCAC | TCCTTCACATTGCGTTGGTA | 65-60 |
| NDSU213 | TC286917 | GCTGAAATCCGAGTTTATTC | TGGATCATTGGCAGGTCTTC | 65-60 |
| NDSU215 | BU672362 | GTCGAGAAGAACAGCCTCAT | ACAAGCCCTTCCAGTAGTTC | 65-60 |
| NDSU217 | TC238759 | ATACAGGAGCAACACAAACCAT | GTCACCATTCTCGTCAGATACA | 65-60 |
| NDSU218 | TC241839 | ACTGATCATTCCATCACCTCCATTC | ACCTGCTAGCTCGTTGTTGAGCTG | 65-60 |
| NDSU220 | TC359809 | GTTTGCTGCGTATCCTTGGCG | CTGCCCTATTCCGAGCCGATC | 65-60 |
| NDSU221 | CB307538 | ATAGTGAACTGCAGGAGTGTTTCAATTT | AAGCTAATGCAACTTCTCAGATATGCTC | 65-60 |
| NDSU222 | TC232332 | AAGGTGTATGATGTCACCCCTTTC | ACATATAGATGTTTGGAACTGAGAGTGC | 65-60 |
| NDSU223 | TC252572 | CAATGAAGAAACATAGCAAGGTGAAAC | TATCTATGACCCACCAAAACCTATTGA | 65-60 |
| NDSU224 | BE518358 | AGGAGGTCATCGCTGTCAAC | ATCTTGCAGTCATGAGGCG | 65-60 |
| NDSU225 | TC237279 | TATGGTATTGTTATGCGTCGACC | ACCTGCGCTTCCGAGATTTGC | 65-60 |
| NDSU226 | BF292414 | GAACTTGACCTTCTGCCTCG | CCAACTGGTACTATCTTGCTAC | 65-60 |
| NDSU227 | BE495786 | GCCCAGGAAAATCTGGGTAT | CATCACAGCCTGCTACATCC | 65-60 |
| NDSU228 | BG606586 | CACAAACTTGGTGCCTGAGA | TCTTTGATGCTGCACAGGAC | 65-60 |
| NDSU229 | BE517729 | ACAGCCAGACTCACCCTGAT | GCCTACTGCAAAATACAGTCCT | 65-60 |
| NDSU230 | BF474139 | GGGTGGTACAGTCCTTGGAA | TAAACAATTACCGCTATGTCACTA | 65-60 |
| NDSU231 | BF474340 | GTTCAGTGATGGGGCAAAGT | CCCCAAACATTTGCCCCG | 65-60 |
| NDSU232 | BF475149 | TGCCCATTCCCAGTAATAAAC | TGTGAGAGCTCCAAAACCAA | 65-60 |
| NDSU233 | BE591501 | GCTTGGAATCTTCGTCCATC | TTCCTTCTGCAGTCCAGGTC | 65-60 |
| NDSU290 | TC272349 | ACTCGTCAACCCCTGTTTCC | CCTAGGCTCGAATCATCTGG | 60-65 |
| NDSU291 | TC247223 | ACTCGTCAACCCCTGTTTCC | CCTAGGCTCGAATCATCTGG | 60-65 |
| NDSU292 | BE404005 | CGGGTCGATGAGTATGGACT | GTCACTGACCTCCGGATCAT | 65-60 |
| NDSU294 | TC286803 | ACTACGTCCCTTGTTGCTGCC | TGCTTCCTCGTCTTCTATACGC | 65-60 |
| NDSU295 | TC290833 | GCATCAAAGGTCTGTTCAAGG | TTCTGTCTCAGTTGCTCTGTGT | 65-60 |
| NDSU296 | TC382336 | AATAARRCCAGAGATGGAAGCC | TTAGTGGACTAAATGGACTAAC | 65-60 |
| NDSU297 | TC303633 | TCCAGCCGCTCAGCCAGAAC | GCGCCCACGGAGATGTTGTC | 65-60 |
| NDSU298 | TC278446 | ACCATGAAGCTGCACTGGCTG | TCTGCCAACTGCAACAACCCG | 65-60 |
| NDSU299 | TC278524 | GCTCAGGGTGATCGACAGC | ATCATTTCTTCCATCTCTCGG | 65-60 |
| RET-09-12 | | GGAATCATTTGCACTCGTTAG | CTTTTCTGCACTAGTGTACC | 60-55 |
| RET-09-14 | | GTAATAAGTGTGTGATTGCC | CTTTTCCTAGTAGTGCATC | 60-55 |
| RET-09-23 | | CACTATATGCTCACTCTGAA | AGTGTGCTAAGCTAACGGAA | 60-55 |
| RET-09-24 | | GGATTGTACAAGGGGATAAT | CTTTTCCCTAGTAGTGGGAA | 60-55 |
| RET-09-25 | | TAACAGGAGACAAGTGATAC | AAAAATAACTTCCGGGGAG | 60-55 |
| RET-09-6 | | TTAGTAGTGGAGCCCAGTTG | CCACTTACCACCATGTTGTA | 60-55 |
| RET-09-8 | | GTAATAAGTGTGTGATTGCC | CTTTTCCTAGTAGTGCATC | 60-55 |
| RET-09-9 | | GGAATCATTTGCACTCGTTAG | CTTTTCTGCACTAGTGTACC | 60-55 |
| RET-1-2 | | private sequence | private sequence | 60-55 |
| RET-32-2 | | private sequence | private sequence | 60-55 |
| RET-40-3 | | private sequence | private sequence | 60-55 |
| UMN25 | AAR95704 | GGGACAATACGAGCAGCAAA | CTTGTTCCGGTTGTTGCCA | 65-60 |

# APPENDIX D

## D.1. Sequence analysis to identify bulk positive-specific PE2 sequences

PE2 reads from the positive and negative bulk samples were *de-novo* assembled using Stacks software (Catchen et al. 2011) to generate reference contig sequences corresponding to *Pst*I (or *Aat*II) loci in wheat (Figure D.1). A minimum of 3 reads and 98% sequence identity was required for contig formation. Next, PE2 reads from the positive bulk (BulkPos) and negative bulk (BulkNeg) were aligned separately to the PE2 contigs using BWA (Li and Durbin 2009).

Bulk positive-specific sequences were identified as contigs that had high read coverage in the BulkPos sample and no coverage in the BulkNeg sample (Figures D.2 and D.3). To find an appropriate false-discovery cut-off to account for sampling error, the maximum read count in the BulkNeg sample (when there was no coverage in the BulkPos sample) was identified. The false discovery cut-offs were found to be 58 for the *Pst*I library and 269 for the *Aat*II library (Figures D.4 and D.5). To be conservative and minimise the false-discovery rate, the minimum read coverage for classifying bulk positive-specific contigs was set at 100 for the *Pst*I library and 500 for the *Aat*II library.

## D.2. Sequence analysis to identify bulk positive-specific PE1 sequences

PE1 reads corresponding to the opposite end of restriction fragments specific to the BulkPos sample were isolated and *de-novo* assembled using CAP3 (Huang and Madan 1999) at 98% sequence identity. The PE1 contigs were named to allow tracking of PE. For example, PE1 contig '*NSDU_CR_PstI_PE2contig00000323_Contig1'* corresponds to PE2 contig 'NSDU_CR_*Pst*I_PE2contig00000323'.
In some cases, multiple PE1 contigs were generated per PE2 contig. Multiple contigs may arise for several reasons: 1) the PE1 reads belong to a single locus but there was insufficient read overlap to form a single contig; 2) the PE1 reads belong to multiple loci but there was insufficient sequence diverge at the other end of the restriction fragment to resolve the individual loci when the PE2 reads were assembled. A number at the end of the contig name is used to track the occurrence of multiple PE1 contigs.

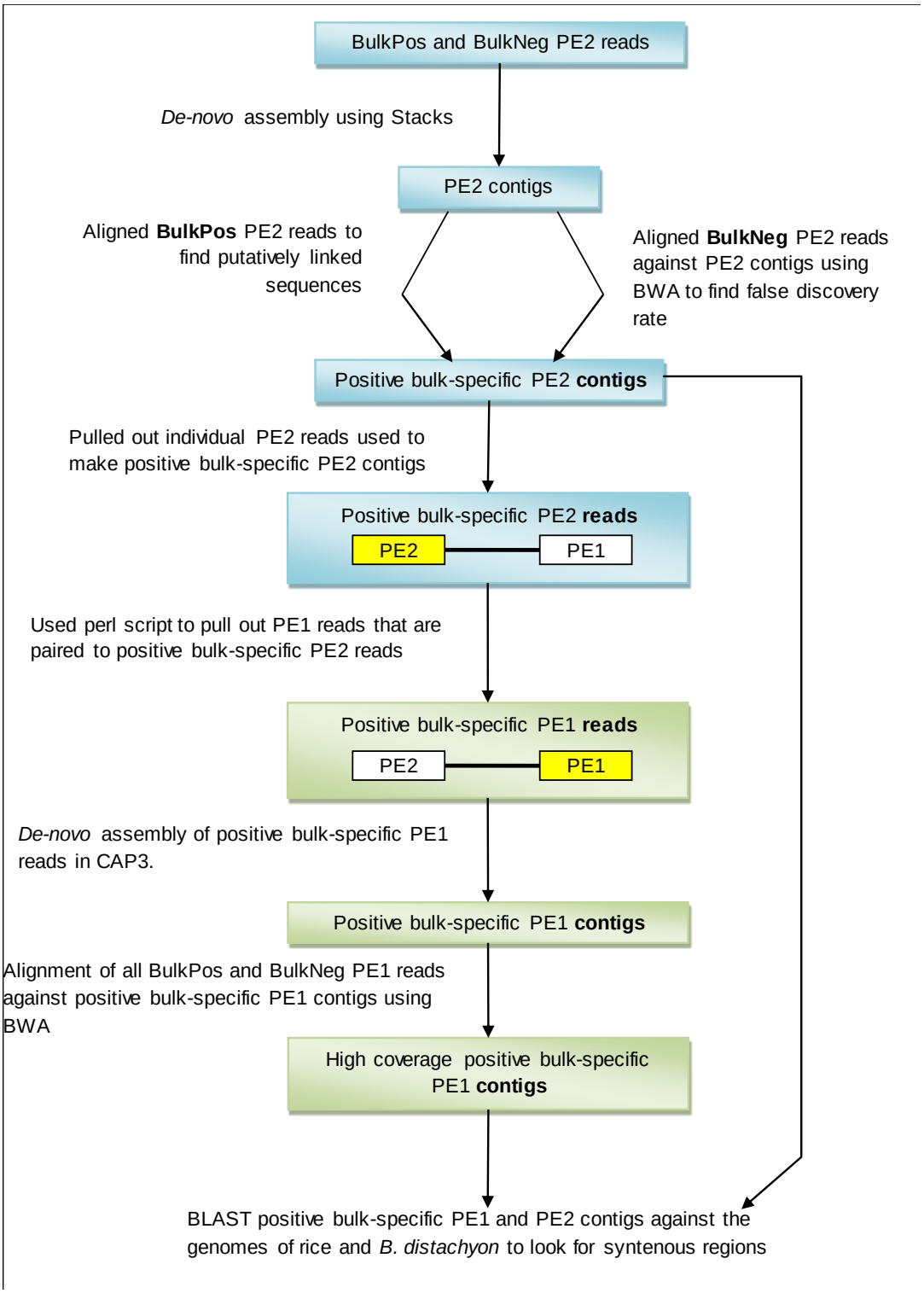


Figure D.1. Bioinformatics approach used to identify positive bulk-specific sequences.

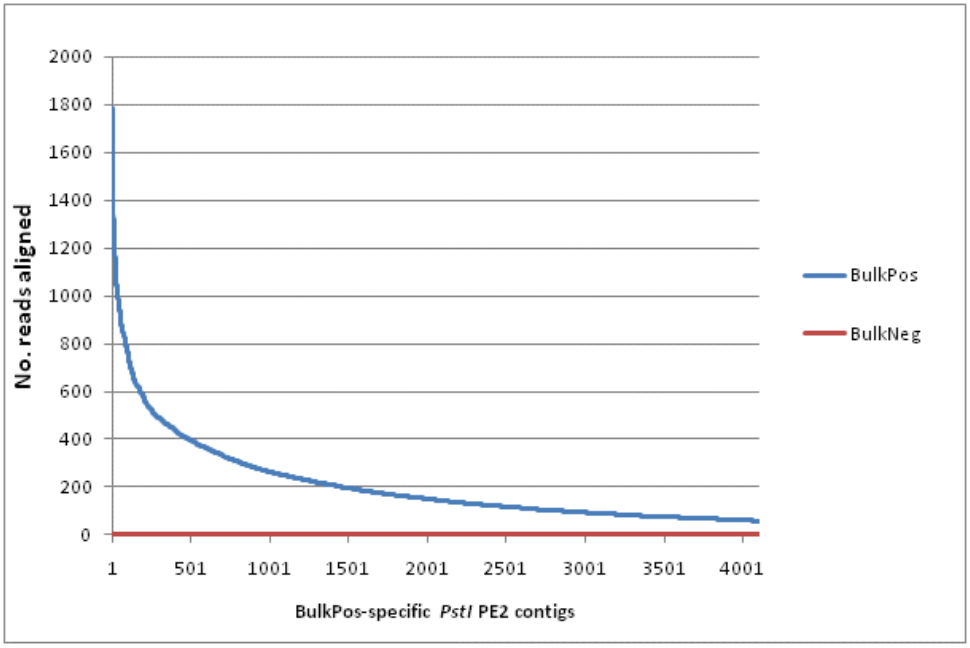


Figure D.2. Read coverage for bulk positive-specific PE2 contigs identified in *Pst*I reduced representation library.

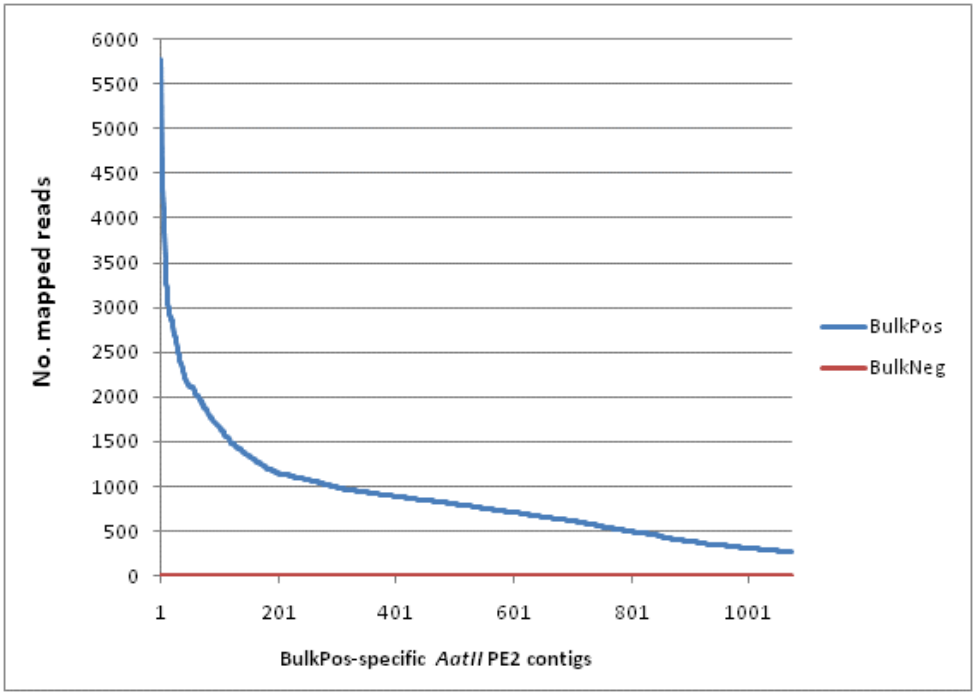


Figure D.3. Read coverage for bulk positive-specific PE2 contigs identified in *Aat*II reduced representation library.

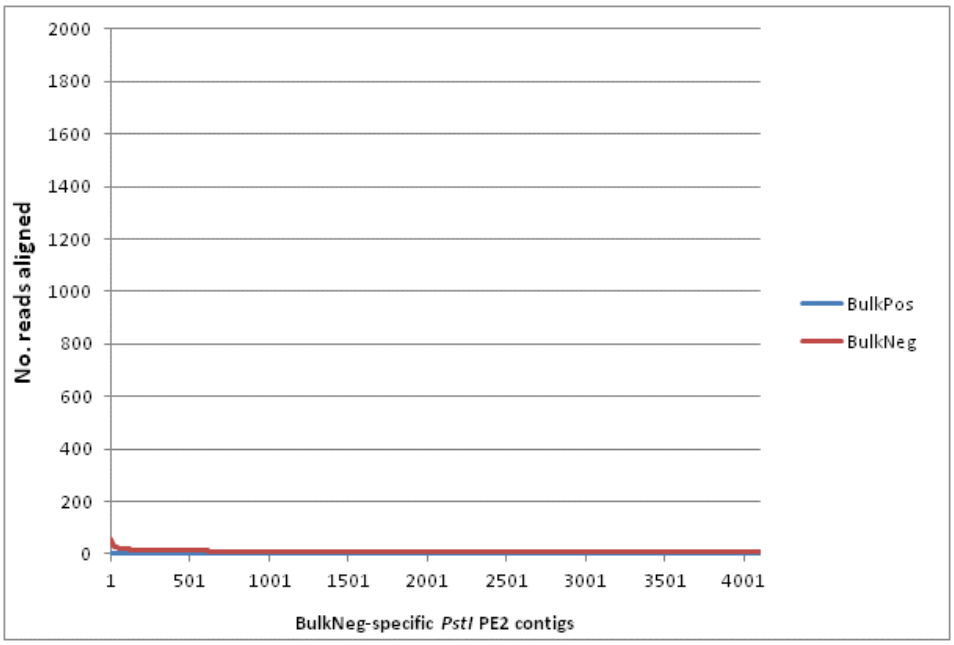


Figure D.4. Read coverage for bulk negative-specific PE2 contigs identified in *Pst*I reduced representation library.

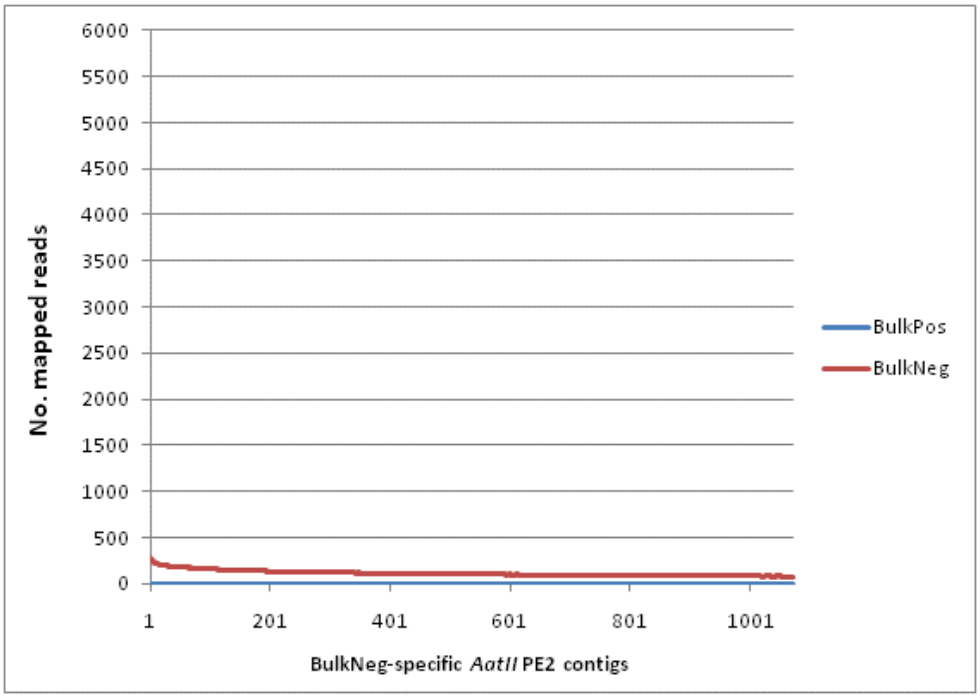


Figure D.5. Read coverage for bulk negative-specific PE2 contigs identified in *Aat*II reduced representation library.

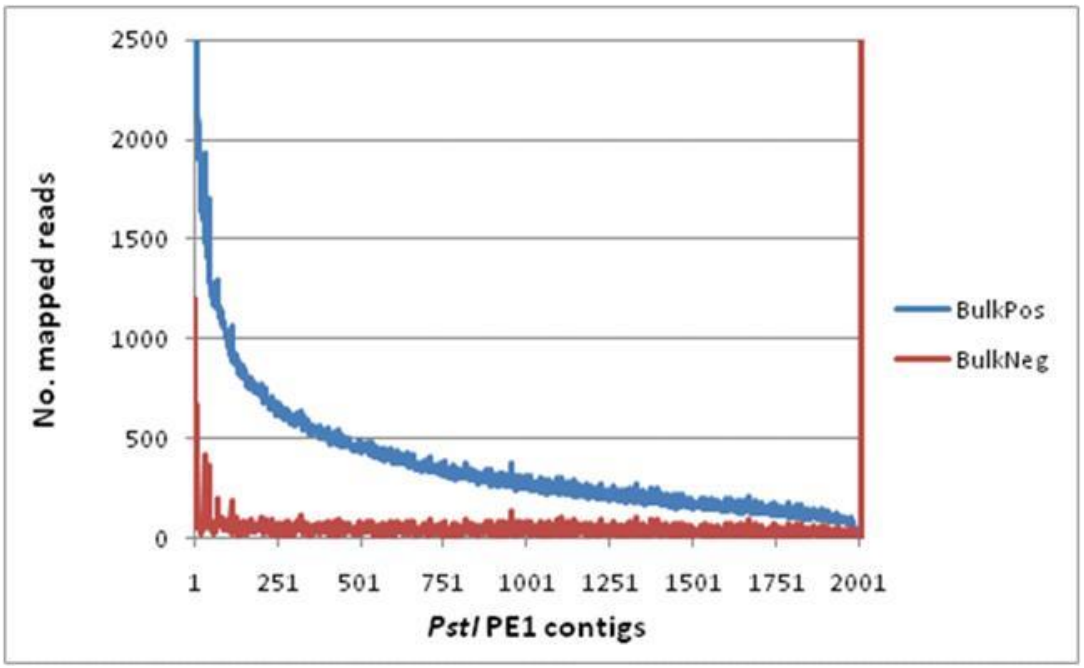


Figure D.6. Read coverage for bulk positive-specific PE1 contigs identified in *Pst*I reduced representation library.

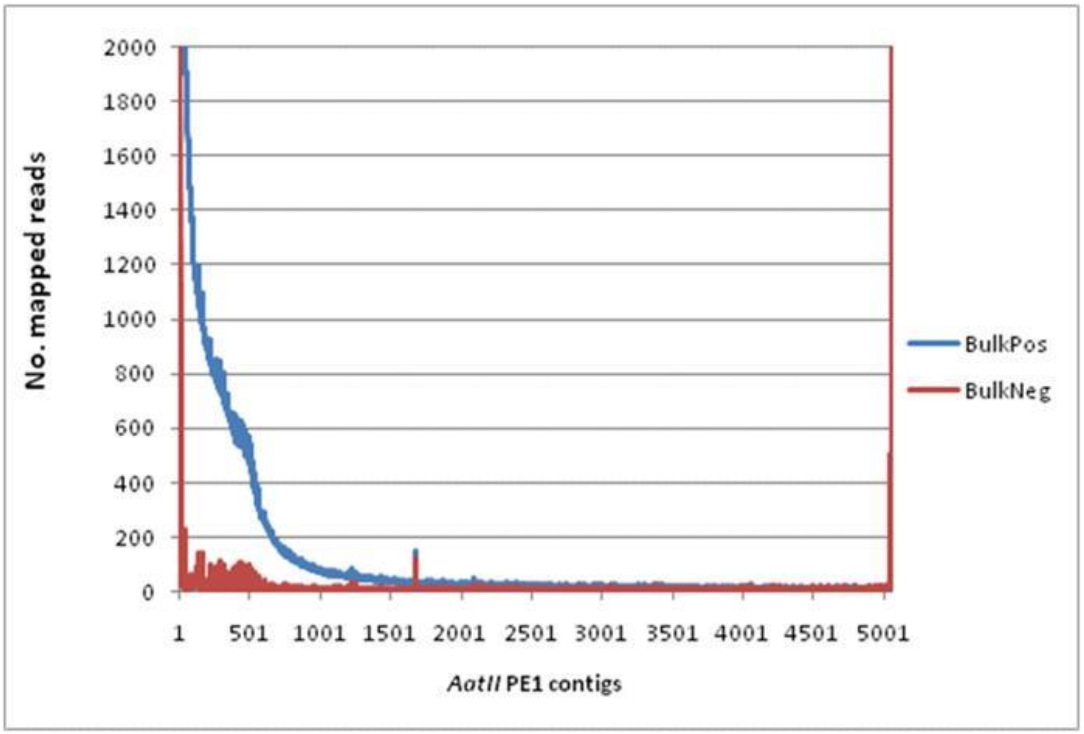


Figure D.7. Read coverage for bulk positive-specific PE1 contigs identified in *Aat*II reduced representation library.

D.3. *In-silico v*alidation of bulk positive-specific contigs

To confirm the specificity of the restriction fragments to each sample pool, PE1 reads from the BulkPos and BulkNeg reduced representation libraries were separately aligned to the bulk positive-specific PE1 contigs using BWA. In most instances, differential coverage was observed between the bulk samples (Figures D.6 and D.7), confirming the initial results using PE2 tags.

D.4. References

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

Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. Genome Res 9:868-877

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics 25:1754-1760

# APPENDIX E

Table E.1. Primer and probes sequences RHBD, Actin, NDRT, NDCtg

| Marker | Primer forward | Primer reverse |
|---|---|---|
| RHBD1 | AAACCGCGTCCACCCCACGC | CCCGAGGAGCGGGTTCTGGC |
| RHBD2 | TGGGTGCCGTGGATCGTGCC | GCAAGCAGGTGGACGACGCC |
| RHBD3 | CTGCATCTGGCTCCACGCCG | AGCGTCACCAGCGCCGCC |
| RHBD4 | CCTCCGGTGCGCTCTTCGGG | CGTTGCCGTTCACCCCTCGG |
| RHBD5 | AACTTCGCGCACATCGGGGG | GCGCATCAAAATGCCTGTGCG |
| RHBD6 | GCTGGCGGTTGGGATGGCG | TCATTAACACTTGCTTTCGGTGCGG |
| NDRT20 | GAGAGTCGGCGTCATCTACC | CACGAACAGCAGCGTCAC |
| NDRT20a Probe | [FAM] agagacaacatctccgtgggcgc [BHQ1] | |
| NDRT20b Probe | [TET] aaagacaacatctccgtgggcgc [BHQ1] | |
| Actin | GGCAACATTGTTCTCAGTGGTGGT | TCCTTTCAGGAGGAGCAACAACCT |
| NDCtg49.1 | CGGGAAGGAATCTACCTCGC | ATTTTCCCCCGCTTCTTCCC |
| NDCtg49.2 | CCACACCACACACATTTGCC | GGGTTCTTGCGATCTACCCG |
| NDCtg235.1 | GGAAGGATTGTGTGTGTAGGC | ATCTGGCCGAAGTTACACGG |
| NDCtg6.1 | TGTTTGTGCTTCTCATTTCTAGGC | AGATGCAACCAAATTCAGATAGC |
| NDCtg6.2 | GATTTTTCCGATTCGGGCGG | AAAACTTTCGCCAGCTTCGC |
| NDCtg478.1 | CTTCGACTCCATCACCAGGG | GAGCTGACACAGAGGATGGG |
| NDCtg478.2 | CTTACAGGCAACCTCCAGGG | CACTTGTCACCCATCTCCCC |
| NDCtg478.3 | TTCGCTTCTTTTTGCTCGCC | TCACTGGTGTGCATCTTCCC |
| NDCtg77.1 | AGCGGGGTTAGTACAATGGC | GCTCCCACTTCATCCTACGG |
| NDCtg77.2 | CACCAGCAAAGCATACTGCC | GGATCTCCCCTTTTTGGGGG |
| NDCtg18.1 | AGGTCTCTTTTCTTCGCCCC | ACAGATACACACACCAGCCG |
| NDCtg5.1 | GTTTCTATAAACTTAAGAAGGGGGC | TGTGCTATTGGAGATTTCGGG |

# APPENDIX F

Table F.1. Contigs containing genes and their orthologus relationships

| Contig ID | Wheat EST | Orthologus genes | | | Pfam description |
|---|---|---|---|---|---|
| | | Rice | Brachy | Sorghum | |
| Contig0108 | BQ239693 | Os10g35800 | Bd3g30140 | Sb01g017900 | expressed protein |
| Contig0108 | TC449685 | Os10g35810 | Bd3g30150 | Sb01g017890 | thylakoid lumenal protein |
| Contig0235 | TC396795 | Os10g36060 | Bd3g30280 | Sb01g017720 | protein transport protein-related |
| Contig0235 | TC461867 | Os10g36420 | Bd3g30410 | Sb01g017560 | YABBY domain containing protein |
| Contig0478 | CA624901 | Os03g02514 | Bd3g31450 | Sb01g031240 | hydrolase alpha/beta fold family protein |
| Contig0478 | TC378132 | Os10g37730 | Bd3g31460 | Sb01g031230 | pollen ankyrin |
| Contig0478 | CA686696 | Os10g37700 | Bd3g31440 | Sb01g031250 | hydrolase protein putative |
| Contig0478 | TC421023 | Os10g37710 | Bd3g31440 | Sb01g031250 | hydrolase alpha/beta fold family protein |
| Contig0478 | CA723767 | Os10g37720 | Bd3g31460 | Sb01g031230 | hydrolase alpha/beta fold family protein |
| Contig0478 | CK205415 | Os10g37740 | Bd3g31480 | Sb01g031220 | CGMC_GSK.1 - CGMC includes CDA MAPK GSK3 |
| Contig0049 | TC371272 | Os10g35670 | Bd3g30040 | Sb01g018050 | zinc finger RING-type |
| Contig0049 | CA677620 | Os10g35680 | Bd3g30050 | Sb01g018040 | acetyltransferase GNAT family |
| Contig0005 | CA664118 | Os10g37760 | Bd3g31500 | Sb01g031210 | OsRhmbd17 - Putative Rhomboid homologue |
| Contig0005 | TC372858 | Os03g02540 | Bd1g77440 | Sb01g049510 | proteasome subunit |
| Contig0006 | TC396795 | Os10g36190 | Bd3g30330 | Sb01g017690 | PPR repeat domain containing protein |
| Contig0006 | BE425613 | Os10g36200 | Bd3g30340 | Sb01g017680 | heavy metal-associated domain containing protein |
| Contig0006 | CA625851 | Os10g36210 | Bd3g30350 | Sb01g039010 | valyl-tRNA synthetase |
| Contig0006 | TC397290 | Os10g36250 | Bd3g30360 | Sb01g017670 | tetratricopeptide repeat |
| Contig0006 | TC450913 | Os10g36340 | Bd3g30370 | Sb01g017640 | protein |
| Contig0077 | BQ744241 | Os10g36550 | Bd3g30750 | Sb01g017460 | CRP3 - Cysteine-rich family protein precursor |
| Contig0077 | CK215327 | Os10g36580 | Bd3g30730 | Sb01g017440 | UP-9A putative |
| Contig0077 | CK215327 | Os10g36610 | Bd3g30730 | Sb01g017440 | UP-9A putative |
| Contig0506 | TC395533 | Os05g38400 | Bd2g23360 | Sb01g046840 | OsFtsH3 FtsH protease homologue of AtFtsH3/10 |
| Contig0123 | TC437625 | Os05g30480 | Bd3g01480 | Sb01g010460 | DnaK family protein |
| Contig0126 | . | Os04g16722 | . | . | uncharacterized protein ycf68 |

Table F.2. Contigs containing pseudogenes and their orthologus relationships

| Contig ID | Wheat EST | Orthologus genes | | | Pfam description |
|---|---|---|---|---|---|
| | | Rice | Brachy | Sorghum | |
| Contig0010 | CA701343 | Os04g40990 | Bd5g13940 | Sb06g020720 | malate synthase glyoxysomal |
| Contig0176 | . | Os11g26860 | . | . | serine hydroxymethyltransferase mitochondrial precursor |
| Contig0199 | TC429612 | Os01g22010 | Bd2g12160 | Sb03g013170 | S-adenosylmethionine synthetase |
| Contig0281 | TC373948 | Os01g38970 | Bd2g41480 | Sb03g025720 | carbamoyl-phosphate synthase large chain |
| Contig0380 | . | Os07g49520 | . | . | 2-oxoglutarate dehydrogenase E1 mitochondrial precursor |
| Contig0380 | CA624112 | Os09g20440 | Bd3g13980 | Sb02g023310 | succinate dehydrogenase and fumarate reductase iron-sulfu |
| Contig0430 | CK216354 | Os01g57968 | Bd2g45280 | Sb0221s002050 | protein |
| Contig0496 | BJ211068 | Os09g34010 | Bd2g38310 | Sb06g017820 | protein |
| Contig0498 | . | Os02g04170 | . | . | L-aspartate oxidase 1 |
| Contig0507 | . | Os02g49720 | . | . | aldehyde dehydrogenase |
| Contig0090 | . | Os04g45490 | . | . | elongation factor |
| Contig0200 | . | Os05g47490 | . | . | MDR-like ABC transporter |
| Contig0072 | . | Os05g47490 | . | . | MDR-like ABC transporter |
| Contig0154 | . | Os05g47490 | . | . | MDR-like ABC transporter |

# APPENDIX G

## G.1. FASTA sequences of the rhomboid gene

GENOMIC

>Chromosome 1D gDNA JX885738
ACCGCGTCCACCCCACGCGGCTGAGGCAACGCACCACCTGCCGCCGCGCCATTTATCACCCGGCC
GCCTCGCCTCGCCTCCCACCTTTTCCACCACCCCGCTGCCGTTACCAGTCTCCACGAGACCGCAGC
CAATTAGAGCCGTTGCTCCAGGAACGCAGCTCGCCCGCCCGTCCTACGTCGTGGGCGCCATGGCC
GGCTCGCGGCGCGACGCGGACCGGAAGGAGAAGGCCCCGCCGGCGGCGCCGCCGCAGCAGCAG
CCGGGGCGCGAGGAGGAGAGGGAGTGGGTGCCGTGGATCGTGCCGGTCTTCGTGGCCGCCAACG
TCGCCCTGTTCGCCGTGGCTATGTACGCCAACAACTGCCCCGCGCACGCGCGCGGCCGCAGGAAG
TGCGTCGGCGCGGGGTTCCTCCGCCGCTTCGCCTTCCAGCCGCTCAGCCAGAACCCGCTCCTCGG
GCCCTCCTCCGCCACGTGAGTGATCCGTCGATCTCGTCTTCCCAACAATTCGGCTCTCGCCGCTTC
GATTCCTTCGATTCAACGAGAGATTTGTCATGTCTTACGCTCGCGGTACATACTGTCGTCTTCAGATT
GCAGAAGCTGGGCGCGCTCGTGTGGGACAAGGTGGTGCACGAGCACCAGGGGTGGAGACTCGTG
ACCTGCATCTGGCTCCACGCCGGCGTCGTCCACCTGCTTGCCAACATGCTCAGCCTCGTGCTCGTC
GGGCTCAGGCTCGAGCAGCAGTTCGGATTCGGTACGCGCGGCGCGGCCTTCTTCCCCGGAACTGA
CATTCCCATTATTACCTGAAAAAAGAAATCTGAACTAACTTTGTTCCCGTGAATTTCGTTGCAGTGAGA
GTCGGCGTCATCTACCTCGTCTCCGGCGTCGGGGGCAGCGTGATGTCGTCCCTGTTCATCAGAGAC
AACATCTCCGTGGGCGCCTCCGGTGCGCTCTTCGGGCTCCTCGGAGCCATGCTCTCAGAGCTCTTC
ACCAACTGGACCATCTACACCAACAAGGTGGGCGCATCTCCAACCAATTCTGCAGCAACAGTTGGTT
CTGAAAATCCGTCGAGTTTCTGACACTTCCGACATGGCCGGAATGCGTGCAGGCGGCGGCGCTGGT
GACGCTGCTGTTCGTGATCGCGGTCAACCTGGCGATCGGCATCCTGCCGCACGTGGACAACTTCGC
GCACATCGGGGGGTTCCTCACGGGGTTCCTGCTCGGGTTCGTGCTGCTCATGCGCCCGCACTACG
GCTGGGCGCAGCGCTACGTGCTGCCGTCCTCCGTCAAGGACGTCGGCCGGAAGTTCCTGGCCTAC
CAGTGGGCTCTGCTCGCCGTGGCCTCGGTCCTGGTCGTCGTCGGGTAAGCAAAGTGGCAATGACAT
TGCGTGAGTTTAGGTTAGGTGGGCTCTGGACTTCCGGTTGACTGATGCTGCATTGCTGTTGTCCTGC
GCAGGCTGGCGGTTGGGATGGCGATGCTCTTCCGAGGGGTGAACGGCAACGAGCACTGCCAGTGG
TGCCACTACCTCAGCTGCGTGCCTACCGCCAGATGGAGCTGCGGGAAATGAAATGGGAGCAAGCAA
CTTCATGGACCTCCATATTACTTTGTTCTTTTCCGCACAGGCATTTTGATGCGCAGATGCTGTATATA
CGAGTACTGCTATATGACGGATTACACCCCGGGGTTTACTCGATGATTGTTTCGTTCGGTTGTTTGAT
TTGAAAAAGGGAGAATATCTTGAATCTTCCACCTGCATGTGCCTAACAGCGCGGCTGCCTTTTAAATT
TGTGTGCCAAACAGATCCATCTTTATATAATTTTACCCAATGATTAATTATCATGGTACAGCTTTGAGA
TTCCACAAAGCCAAGACAACAACTGGTCGGTGCAGATTATAGAAATGGCTTCAGGTAGGAAATGATT
TATTATTTTTTCCGCACCGAAAGCAAGTG

>Chromosome 1B gDNA
ACCGCGTCCACCCCACGCGGCTGAGGCAAACGCACCACCTGCCGCCGCGCCATTTATTACCCGGC
CGCCTCGCCTCCCACCTTTTCCACTACCCCGCTGCCGTTACCACTCTCCACGAGACCGCAGCCGAT
TAGAGCCGTTGCTCCAGGAACGCAGCTCGCCCGCACCGCGCCACGCGACAGCTCGCTCGCTAGCC
CGTACGTCGTGGGCGCCATGGCCGGCTCGCGGCGCGACGCGGACCAGAAGGAGAAGGCCCCGCC
GGCGGCGCCGCCGCAGCAGCAGCCGGGGCGCGAGGAGGAGAGGGAGTGGGTGCCGTGGATCGT
GCCGGTCTTCGTGGCCGCCAACGTCGCCCTGTTCGCCGTGGCGATGTACGCCAACAACTGCCCCG
CGCACGCGCGCGGCGGCCGCGGCGGCAGGAAGTGCGTCGGCGCGGGGTTCCTCCGCCGCTTCGC
CTTCCAGCCGCTCAGCCAGAACCCGCTCCTCGGGCCCTCCTCCGCCACGTGAGTGATCCGTCGGTC
TCGTCTTCCCAACAATTCGGCTCTCGCCGCTTCGATTCCTTCGATTCAACGAGAGATTCGTCATGTCT
TACGCTCCCGGTACATACTGTCGTCTCCAGATTGCAGAAGCTGGGCGCGCTCGTGTGGGACAAGGT
GGTGCACGAGCACCAGGGGTGGAGACTCGTGACCTGCATCTGGCTCCACGCCGGCGTCGTCCACC
TGCTTGCCAACATGCTCAGCCTCGTGCTCGTCGGGCTCAGGCTGGAGCAGCAGTTCGGATTTGGTA

CGCGTTCTTCTCCGGAACTGACACTCCCATTACCTGAACATTCTGAACTAACATTTTTTTTTTGAAACG
AAAGGGGTTGCCCCCCTGCCCCAACTTTTTATTGAATTCTGAACTAACATTGTTCCCGTGAATTTCGT
TGCAGTGAAAGTCGGCGTCATCTACCTCGTCTCCGGCGTCGGGGGCAGCGTGATGTCGTCCCTGTT
CATCAAAAACAACATCTCCGTGGGCGCGTCCGGTGCGCTCTTCGGGCTCCTCGGAGCCATGCTCTC
AAAGCTCTTCACCAACTGGACCATCTACACCAACAAGGTGGGCGCATCTCCAACCAATTCTGCAGCA
ACAGTTGGTTCTGAAAATCCGTCAAGTTTCTGACACTTCCGACGTGGCCGGAATGCGTGCAGGCGG
CGGCGCTGGTGACGTTGCTGTTCGTGATCGCGGTCAATCTGGCGATCGGCATCCTGCCGCACGTG
GACAACTTCGCGCACATCGGGGGGTTCCTCACGGGGTTCCTCCTCGGGTTCGTACTGCTCATGCGC
CCGCACTACGGCTGGGCGCAGCGCTACGTGCTGCCGTCCTCCGTCAAGGACGTCGGCCGGAAGTT
CCTGGCCTACCAGTGGGCTCTGCTCGCCCTGGCCTCGGTCCTGGTCGTCATCGGGTAAGCAAAGC
GGCAACGACATCTCTCTTGTTGAGTTGAGTCTAGGTGGGCTCTCGACTTCCGGTTGACTGATGCTGT
CTTGCTGTTGTCCTGCGCAGGCTGGCGGTTGGGATGGCGATGCTCTTCCGAGGGGTGAACGGCAA
CGAGCACTGCGAGTGGTGCCACTACCTCAGCWGCTGCGTGCCGACCGCCAGATGGAGCTGCGGG
AAATGAAATGGGAGCAAGCAACTTCATGGACCTCCATATTACTTCGTTCTTTTCCGCACAGGCATTTT
GATGCGCAGATGCTGTATATACGAGTACTGCTATATAACGGATTACACCCCTGGATTTACTCGATGAT
TGTTTCCTTCGGTTGTTGGACTTGAAAAACGGGGAATATCTTGAATCTTACACCTGCATGTGCCTGAC
AGCGCGGCTGCCTTTTAAATTTGTGTGCCAAACAGATCCATCTTTATATAATTTACCCAATGATTAATT
ATCATGGTACAGCTTTGAGATTCCACAAAGCCAAGATAAAAACTGGTCGGTGCAGATTATAGAAATG
GCTTCATGTAGGAAATGTTTTAGTATTATTTTCTCCGCACCGAAAGCAAGTGTTAATGAA

>Aegilopis tauschii gDNA
ACCGCGTCCACCCCACGCGGCTGAGGCAACGCACCACCTGCCGCCGCGCCATTTATCACCCGGCC
GCCTCGCCTCGCCTCCCACCTTTTCCACCACCCCGCTGCCGTTACCAGTCTCCACGAGACCGCAGC
CAATTAGAGCCGTTGCTCCAGGAACGCAGCTCGCCACGCGACAGCTCGCCCGCCCGTCGTACGTC
GTGGGCGCCATGGCCGGCTCGCGGCGCGACGCGGACCGGAAGGAGAAGGCCCCGCCGGCGGCG
CCGCCGCAGCAGCAGCCGGGGCGCGAGGAGGAGAGGGAGTGGGTGCCGTGGATCGTGCCGGTCT
TCGTGGCCGCCAACGTCGCCCTGTTCGCCGTGGCTATGTACGCCAACAACTGCCCCGCGCACGCG
CGCGGCCGCAGGAAGTGCGTCGGCGCGGGGTTCCTCCGCCGCTTCGCCTTCCAGCCGCTCAGCCA
GAACCCGCTCCTCGGGCCCTCCTCCGCCACGTGAGTGATCCGTCGATCTCGTCTTCCCAACAATTC
GGCTCTCGCCGCTTCGATTCCTTCGATTCAACGAGAGATTTGTCATGTCTTACGCTCGCGGTACATA
CTGTCGTCTTCAGATTGCAGAAGCTGGGCGCGCTCGTGTGGGACAAGGTGGTGCACGAGCACCAG
GGGTGGAGACTCGTGACCTGCATCTGGCTCCACGCCGGCGTCGTCCACCTGCTTGCCAACATGCTC
AGCCTCGTGCTCGTCGGGCTCAGGCTCGAGCAGCAGTTCGGATTCGGTACGCGCGGCGCGGCCTT
CTTCCCCGGAACTGACATTCCCATTATTACCTGAAAAAAGAAATCTGAACTAACTTTGTTCCCGTGAA
TTTCGTTGCAGTGAGAGTCGGCGTCATCTACCTCGTCTCCGGCGTCGGGGGCAGCGTGATGTCGTC
CCTGTTCATCAGAGACAACATCTCCGTGGGCGCCTCCGGTGCGCTCTTCGGGCTCCTCGGAGCCAT
GCTCTCAGAGCTCTTCACCAACTGGACCATCTACACCAACAAGGTGGGCGCATCTCCAACCAATTCT
GCAGCAACAGTTGGTTCTGAAAATCCGTCGAGTTTCTGACACTTCCGACATGGCCGGAATGCGTGCA
GGCGGCGGCGCTGGTGACGCTGCTGTTCGTGATCGCGGTCAACCTGGCGATCGGCATCCTGCCGC
ACGTGGACAACTTCGCGCACATCGGGGGGTTCCTCACGGGGTTCCTGCTCGGGTTCGTGCTGCTCA
TGCGCCCGCACTACGGCTGGGCGCAGCGCTACGTGCTGCCGTCCTCCGTCAAGGACGTCGGCCGG
AAGTTCCTGGCCTACCAGTGGGCTCTGCTCGCCGTGGCCTCGGTCCTGGTCGTCGTCGGGTAAGCA
AAGTGGCAATGACATTGCGTGAGTTTAGGTTAGGTGGGCTCTGGACTTCCGGTTGACTGATGCTGCA
TTGCTGTTGTCCTGCGCAGGCTGGCGGTTGGGATGGCGATGCTCTTCCGAGGGGTGAACGGCAAC
GAGCACTGCCAGTGGTGCCACTACCTCAGCTGCGTGCCTACCGCCAGATGGAGCTGCGGGAAATG
AAATGGGAGCAAGCAACTTCATGGACCTCCATATTACTTTGTTCTTTTCCGCACAGGCATTTTGATGC
GCAGATGCTGTATATACGAGTACTGCTATATGACGGATTACACCCCGGGGTTTACTCGATGATTGTTT
CGTTCGGTTGTTTGATTTGAAAAAGGGGGAATATCTTGAATCTTCCACCTGCATGTGCCTAACAGCG
CGGCTGCCTTTTAAATTTGTGTGCCAAACAGATCCATCTTTATATAATTTTACCCAATGATTAATTATC
ATGGTACAGCTTTGAGATTCCACAAAGCCAAGACAACAACTGGTCGGTGCAGATTATAGAAATGGCT
TCAGGTAGGAAATGATTTATTATTTTTTCCGCACCGAAAGCAAGTGGTTA

>(lo)scsti gDNA
TGGATCGTGCCGGTCTTCGTGACCGCCAACGTCGCCCTGTTCGCCGTGGCGATGTACGCCAACAAC
TGCCCCGCGCACGCGCGCGGCGGCCGCGGCGGCAGGAGGTGCGTCGGCGCGGGGTTCCTCCGC

CGCTTCGCCTTCCAGCCGCTCAGCCAGAACCCGCTCCTCGGGCCCTCCTCCGCcACGTGAGTGATC
CGTCGATCTCGTCTCCCCAACAATTCGGCTCTCaCCGCTTCGATTCCTTCGATTCAGCGAGAGATTT
GTCATCTCTTACGCTCGCGGTACATACTGTCGTCTTCAGATTGCAGAAGTTGGGCGCTCTCGTGTGG
GACAAGGTGGTGCACGAGCAACAGGGGTGGAGACTCGTGACGTGCATCTGGCTCCACGCCGGCGT
CGTCCACCTGCTGGCCAACATGCTCAGCCTCGTGCTCGTCGGGCTCAGGCTCGAGCAGCAGTTCG
GATTCGGTACGCGTTCATCTCCGGAACTGACACTCCAATTACCTGAACATTCTGAACTGATATTGTTC
CCGTCAATTTGGTTGCAGTGAGAGTCGGCGTCATCTACCTCGTGtCCGGCGTCGGGGGCAGCGTGA
TGTCATCCCTGTTCATCAGAGACAACATCTCCGTGGGCGCGTCCGGTGCGCTCTTCGGGCTCCTCG
GAGCCATGCTCTCGGAGCTCTTCACCAACTGGACCATCTACACCAACAAGGTGGGTgCATCTCCAAC
CAATTCTGCACATATTCTGCAGCGGTAGCAGCACCAGCAGTACTTcGTTCTGAAAATCCGTCGAGTTT
CTGACACTGTCGACGTGGCTGGAATGCGTGCAGGCGGCGGCGCTGGTGACGCTGCTGTTCGTGAT
CGCGGTCAACCTGGCGATCGGCATCCTGCCGCACGTGGACAACTTCGCGCACATCGGGGGGTTCCt
CACGGGGTTCCTGCTCGGCTTTGTGTTGCTCATGCGCCCGCACTACGGCTGGGCGCAGCGCTACGT
GCTGCCGTCCTCCGTCAAGGACGTCGGCCGGAAGTTCCTGGCCTACCAGTGGGCTCTGCTCGCCG
TGGCCTCGGTCCTGGTCGTCGTCGGGTAAGCAAAGCGGCAACGACATCTTTCTTGTTGgGTTGGGTT
TAGGTGGGCTCTGGACTTCCGGTTGACTGATGCTGCATTGCTGTTGTCCTGCGCAGGCTGGCGGTT
GGGATGGCGATGCTCTTCCGAGGGGTGAACGGCAA

>Chromosome 1A gDNA
TGGATCGTGCCGGTCTTCGTGgCCGCCAACGTCGCCCTGTTCGCCGTGGTGATGTACGCCAACAACt
GCCCCGCGCACGCGCGCGGCGGCCGCGGCGGCAGGAGGTGCGTCGGCGCGGGGTtCCTCCGCC
GCTTCGCCTTCCAGCCGCTCAGCCAGAACCCGCTCCTCGGGCCCTCCTCCGCCACGTGAGTGATCC
GTCGATCTCGTCTCCCCAACAATTCGGCTCTCGCCGCTTCGATTCCTTCGATTCAGCGAGAGATTTG
TCATCTCTTACGCTCgCGGTACATACTGTCGTCTTCAGATTGCAGAAGTTGGGCGCTCTCGTGTGGG
ACAAGGTGGTGCACGAGCACCAGGGGTGGAGACTCGTGACGTGCATCTGGCTCCACGCCGGCGTC
GTCCACCTGCTGGCCAACATGCTCAGCCTCGTGCTCGTCGGGCTCAGGCTCGAGCAGcAGTTCGGA
TTCTGTACGCGTTCATCTCCGgAACtGACACTCCAATTACCTGAACATTCTGAACTGATATTGTTCCCG
TCAATTTGGTTGCAGTGAGAGTCGGCGTCATCTACCTcGTGTCCGGCGTCGGGGGCAGCGTGATGT
CATCCCTGTTCATCAAAGACAACATCTCCGTGGGCGCGTCCGGTGCgCTCTTCGGGCTCCTCGGAG
CCATGCTCTCgGAGCtCtTCACCAACTGGaCCATCTACACCAACaAGGTGGGTGCATCTCCAACCAATT
CTGCACATATTCTGCAGCGGTAGCAGCACCAGCAGTACTTCGTTCTGAAAATCCGTCGAGTTTCTGA
CACTGTCGACGTGGCCGGAATGCGTGCAGACGGCGGCGCTGGTGACGCTGCTGTTCGTGATCGCG
GTCAACCTGGCGATCGGCATCCTGCCGCACGTGGACAACTTCGCGCACATCGGGGGGTTCCTCAC
GGGGTTCCTGCTCGGCTTCGTGCTGCTCATGCGCCCGCACTACGGCTGGGCGCAGCGCTACGTGC
TGCCGTCCTCCGTCAAGGACGTCGGCCGGAAGTTCCTGGCCTACCAGTGGGCTCTGCTCGCCGTG
GCCTcGGTCCTGGTCGTCGTCGGGTAAGCAAAGCGGCAACGACATCTTTCTTGTTGAGTTGGGTTTA
GGTGGGCTCTGGaCTTCCGCTTGACTGATGCTGCATTGCTGTTGTCCTGCGCAGGCTGGCGGTTGG
GATGGCGATGCTCTTCCGAGGGGTGAACGGCAA

>Normal gDNA
TGGATCGTGCCGGTCTTCGTGACCGCCAACGTCGCCCTGTTCGCCGTGGcGATGTACGCCAACAAC
TGCCCCGCGCACGCGCGCGGCGGCCGCGGCGGCAGGAGGTGCGTCGGCGCGGGGTTCCTCCGC
CGCTTCGCCTTCCAGCCGCTCAGCCAGAACCCGCTCCTCGGGCCCTCCTCCGCCACGTGAGTGATC
CGTCGATCTCGTCTCCCCAACAATTCGGCTCTCGCCGCTTCGATTCCTTCGATTCAGCGAGAGATTT
GTCATCTCTTACGCTCGCGGTACATACTGTCGTCTTCAGATTGCAGAAGTTGGGCGCTCTCGTGTGG
GACAAGGTGGTGCACGAGCAaCAGGGGTGGAGACTCGTGACGTGCATCTGGCTCCACGCCGGCGT
CGTCCACCTGCTGGCCAACATGCTCAGCCTCGTGCTCGTCGGGCTCAGGCTCGAGCAGCAGTTCG
GATTCGGTACGCGTTCATCTCCGGAACTGACACTCCAATTACCTGAACATTCTGAACTGATATTGTTC
CCGTCAATTTGGTTGCAGTGAGAGTCGGCGTCATCTACCTCGTGTCCGGCGTCGGGGGCAGCGTGA
TGTCATCCCTGTTCATCAgAGACAACATCTCCGTGGGCGCGTCCGGTGCGCTCTTCGGGCTCCTCG
GAGCCATGCTCTCGGAGCTCTTCACCAACTGGACCATCTACACCAACAAGGTGGGTGCATCTCCAAC
CAATTCTGCACATATTCTGCAGCGGTAGCAGCACCAGCAGTACTTCGTTCTGAAAATCCGTCGAGTT
TCTGACACTGTCGACGTGGCtGGAATGCGTGCAGGCGGCGGCGCTGGTGACGCTGCTGTTCGTGAT
CGCGGTCAACCTGGCGATCGGCATCCTGCCGCACGTGGACAACTTCGCGCACATCGGGGGGTTCC
TCACGGGGTTCCTgCTCGGcTTtGTGtTGCTCATGCGCCCGCACTACGGCTGGGCGCAGCGCTACGT

GCTGCCGTCCTCCGTCAAGGACGTCGGCCGGAAGTTcCTGGCCTACCAGTGGGCTCTGCTCGCCGT
GGCCTCGGTCCTGGTCGTCGTCGGGTAAGCAAAGCGGCAACGACATCTTTCTTGTTGAGTTGGGTT
TAGGTGGGCTCTGgACTTCCGgTTGACTGATGCTGCATTGCTGTTGTCCTGCGCAGGCTGGCGGTTG
GGATGGCGATGCTCTTCCG

>Weak gDNA
TGGATCGTGCCGGTCTTCGTGaCCGCCAACGTCGCCCTGTTCGCCGTGGtGATGTACgCCAACAACT
GCCCCGCGCACGCGCGCGGCGGCCGCGGCGGCAGGAgGTGCGTCGGCGCGGGGTTCCTCCGCC
GCTTCGCCTTCCAGCCGCTCAGCCAGAACCCGCTCCTCGGGCCCTCCTCCGCCACGTGAGTGATCC
GTCGATCTCGTCTCCCCAACAATTCGGCTCTCGCCGCTTCGATTCCTTCGATTCAGCGAGAGATTTG
TCATCTCTTACGCTCGCGGTACATACTGTCGTCTTCAGATTGCAGAAGTTgGGCGCTCtCGTGTGGGA
CAAGGTGGTGCACGAGCACCAGGGGTGGAGACTCGTGACGTGCATCTGGCTCCACGCCGGCGTCG
TCCACCTGCTGGCCAACATGCTCAGCCTCGTGCTCGTCGGGCTCAGGCTCGAGCAGCAGTTCGGAT
TCGGTACGCGTTCATCTCCGGAACTGACACTCCAATTACCTGAACATTCTGAACTGATATTGTTCCCG
TCAATTTGGTTGCAGTGAGAGTCGGCGTCATCTACCTCGTGTCCGGCGtCGGGGGCAGCGTGATGT
CATCCCTGTTCATCAAAGACAACATCTCCGTGGGCGCGTCCGGTGCGCTCTTCGGGCTCCTCGGAG
CCATGCTCTCGGAGCtCTTCACCAACTGGACCATCTACACCAACAAGGTGGGTGCATCTCCAACCAAc
TCTGCACATATTCTGCAGCGGTAGCAGCACCAGCAGTACTTCGTTCTGAAAATCCGTCGAGTcTCTG
ACACTGTCGACGTGGCCGGAATGCGTGCAGGCGGCGGCGCTGGTGACGCTGCTGTTCGTGATCGC
GGTCAACCTGGCGATCGGCATCCTGCCGCACGTGGAtAACTTCGCGCACATCGGGGGGTTCCTCAC
GGGGTTCCTGCTCGGCTTCGTGCTGCTCATGCGCCCGCACTACGGCTGGGCGCAGCGCTACGTGC
TGCCGTCCTCCGTCAAGGACGTCGGCCGGAAGTTCCTGGCCTACCAGTGGGCTCTGCTCGCCgTGG
CCTCGGTCCTGGTCGTCGTCGGGTAAGCAAAGCGGCAACGACATCTtTCtTGTTGAGTTGgGTtTAGGt
GGGCTCTgGAcTtCCGcTTGACTGATGCTGCaTTGCTGTTGTCCTGCGCAGGCTGGCGGTTGGGATG
GCGATGCTCTTCCGAGGGGTGAACGGCaa

CDS

>Chromosome 1D cDNA JX885738
atggccggctcgcggcgcgacgcggaccggaaggagaaggccccgccggcggcgccgccgcagcagcagccggggcgcgaggaggaga
gggagtgggtgccgtggatcgtgccggtcttcgtggccgccaacgtcgccctgttcgccgtggctatgtacgccaacaactgccccgcgcacgcgc
gcggccgcaggaagtgcgtcggcgcggggttcctccgccgcttcgccttccagccgctcagccagaacccgctcctcgggccctcctccgccaca
ttgcagaagctgggcgcgctcgtgtgggacaaggtggtgcacgagcaccaggggtggagactcgtgacctgcatctggctccacgccggcgtcgt
ccacctgcttgccaacatgctcagcctcgtgctcgtcgggctcaggctcgagcagcagttcggattcgtgagagtcggcgtcatctacctcgtctccgg
cgtcgggggcagcgtgatgtcgtccctgttcatcagagacaacatctccgtgggcgcctccggtgcgctcttcgggctcctcggagccatgctctcag
agctcttcaccaactggaccatctacaccaacaaggcggcggcgctggtgacgctgctgttcgtgatcgcggtcaacctggcgatcggcatcctgc
cgcacgtggacaacttcgcgcacatcgggggttcctcacggggttcctgctcgggttcgtgctgctcatgcgcccgcactacggctgggcgcagc
gctacgtgctgccgtcctccgtcaaggacgtcggccggaagttcctggcctaccagtgggctctgctcgccgtggcctcggtcctggtcgtcgtcggg
ctggcggttgggatggcgatgctcttccgaggggtgaacggcaacgagcactgccagtggtgccactacctcagctgcgtgcctaccgccagatg
gagctgcgggaaatga

>Chromosome 1B cDNA
atggccggctcgcggcgcgacgcggaccagaaggagaaggccccgccggcggcgccgccgcagcagcagccggggcgcgaggaggaga
gggagtgggtgccgtggatcgtgccggtcttcgtggccgccaacgtcgccctgttcgccgtggcgatgtacgccaacaactgccccgcgcacgcgc
gcggcggccgcggcggcaggaagtgcgtcggcgcggggttcctccgccgcttcgccttccagccgctcagccagaacccgctcctcgggccctc
ctccgccacattgcagaagctgggcgcgctcgtgtgggacaaggtggtgcacgagcaccaggggtggagactcgtgacctgcatctggctccac
gccggcgtcgtccacctgcttgccaacatgctcagcctcgtgctcgtcgggctcaggctggagcagcagttcggatttgtgaaagtcggcgtcatcta
cctcgtctccggcgtcgggggcagcgtgatgtcgtccctgttcatcaaaaacaacatctccgtgggcgcgtccggtgcgctcttcgggctcctcggag
ccatgctctcaaagctcttcaccaactggaccatctacaccaacaaggcggcggcgctggtgacgttgctgttcgtgatcgcggtcaatctggcgatc
ggcatcctgccgcacgtggacaacttcgcgcacatcgggggttcctcacggggttcctcctcgggttcgtactgctcatgcgcccgcactacggctg
ggcgcagcgctacgtgctgccgtcctccgtcaaggacgtcggccggaagttcctggcctaccagtgggctctgctcgccctggcctcggtcctggtc
gtcatcgggctggcggttgggatggcgatgctcttccgaggggtgaacggcaacgagcactgcgagtggtgccactacctcagcngctgcgtgcc
gaccgccagatggagctgcgggaaatga

>Aegilopis tauschii cDNA
atggccggctcgcggcgcgacgcggaccggaaggagaaggccccgccggcggcgccgccgcagcagcagccggggcgcgaggaggaga
gggagtgggtgccgtggatcgtgccggtcttcgtggccgccaacgtcgccctgttcgccgtggctatgtacgccaacaactgccccgcgcacgcgc
gcggccgcaggaagtgcgtcggcgcggggttcctccgccgcttcgccttccagccgctcagccagaacccgctcctcgggccctcctccgccaca
ttgcagaagctgggcgcgctcgtgtgggacaaggtggtgcacgagcaccaggggtggagactcgtgacctgcatctggctccacgccggcgtcgt
ccacctgcttgccaacatgctcagcctcgtgctcgtcgggctcaggctcgagcagcagttcggattcgtgagagtcggcgtcatctacctcgtctccgg
cgtcgggggcagcgtgatgtcgtccctgttcatcagagacaacatctccgtgggcgcctccggtgcgctcttcgggctcctcggagccatgctctcag
agctcttcaccaactggaccatctacaccaacaaggcggcggcgctggtgacgctgctgttcgtgatcgcggtcaacctggcgatcggcatcctgc
cgcacgtggacaacttcgcgcacatcgggggggttcctcacggggttcctgctcgggttcgtgctgctcatgcgcccgcactacggctgggcgcagc
gctacgtgctgccgtcctccgtcaaggacgtcggccggaagttcctggcctaccagtgggctctgctcgccgtggcctcggtcctggtcgtcgtcggg
ctggcggttgggatggcgatgctcttccgaggggtgaacggcaacgagcactgccagtggtgccactacctcagctgcgtgcctaccgccagatg
gagctgcgggaaatga

>(lo)scsti cDNA
Atgtacgccaacaactgccccgcgcacgcgcgcggcggccgcggcggcaggaggtgcgtcggcgcggggttcctccgccgcttcgccttccag
ccgctcagccagaacccgctcctcgggccctcctccgccacattgcagaagttgggcgctctcgtgtgggacaaggtggtgcacgagcaacagg
ggtggagactcgtgacgtgcatctggctccacgccggcgtcgtccacctgctggccaacatgctcagcctcgtgctcgtcgggctcaggctcgagca
gcagttcggattcgtgagagtcggcgtcatctacctcgtgtccggcgtcgggggcagcgtgatgtcatccctgttcatcagagacaacatctccgtgg
gcgcgtccggtgcgctcttcgggctcctcggagccatgctctcggagctcttcaccaactggaccatctacaccaacaaggcggcggcgctggtga
cgctgctgttcgtgatcgcggtcaacctggcgatcggcatcctgccgcacgtggacaacttcgcgcacatcgggggggttcctcacggggttcctgctc
ggctttgtgttgctcatgcgcccgcactacggctgggcgcagcgctacgtgctgccgtcctccgtcaaggacgtcggccggaagttcctggcctacca
gtgggctctgctcgccgtggcctcggtcctggtcgtcgtcggn

>Normal cDNA
atgtacgccaacaactgccccgcgcacgcgcgcggcggccgcggcggcaggaggtgcgtcggcgcggggttcctccgccgcttcgccttccag
ccgctcagccagaacccgctcctcgggccctcctccgccacattgcagaagttgggcgctctcgtgtgggacaaggtggtgcacgagcaacagg
ggtggagactcgtgacgtgcatctggctccacgccggcgtcgtccacctgctggccaacatgctcagcctcgtgctcgtcgggctcaggctcgagca
gcagttcggattcgtgagagtcggcgtcatctacctcgtgtccggcgtcgggggcagcgtgatgtcatccctgttcatcagagacaacatctccgtgg
gcgcgtccggtgcgctcttcgggctcctcggagccatgctctcggagctcttcaccaactggaccatctacaccaacaaggcggcggcgctggtga
cgctgctgttcgtgatcgcggtcaacctggcgatcggcatcctgccgcacgtggacaacttcgcgcacatcgggggggttcctcacggggttcctgctc
ggctttgtgttgctcatgcgcccgcactacggctgggcgcagcgctacgtgctgccgtcctccgtcaaggacgtcggccggaagttcctggcctacca
gtgggctctgctcgccgtggcctcggtcctggtcgtcgtcggn

>Weak cDNA
Atgtacgccaacaactgccccgcgcacgcgcgcggcggccgcggcggcaggaggtgcgtcggcgcggggttcctccgccgcttcgccttccag
ccgctcagccagaacccgctcctcgggccctcctccgccacattgcagaagttgggcgctctcgtgtgggacaaggtggtgcacgagcaccaggg
gtggagactcgtgacgtgcatctggctccacgccggcgtcgtccacctgctggccaacatgctcagcctcgtgctcgtcgggctcaggctcgagcag
cagttcggattcgtgagagtcggcgtcatctacctcgtgtccggcgtcgggggcagcgtgatgtcatccctgttcatcaagacaacatctccgtggg
cgcgtccggtgcgctcttcgggctcctcggagccatgctctcggagctcttcaccaactggaccatctacaccaacaaggcggcggcgctggtgac
gctgctgttcgtgatcgcggtcaacctggcgatcggcatcctgccgcacgtggataacttcgcgcacatcgggggggttcctcacggggttcctgctcg
gcttcgtgctgctcatgcgcccgcactacggctgggcgcagcgctacgtgctgccgtcctccgtcaaggacgtcggccggaagttcctggcctacca
gtgggctctgctcgccgtggcctcggtcctggtcgtcgtcggn

PROTEIN
GenScan prediction using Z. Maize as model

>Chromosome 1D JX885738 protein
MAGSRRDADRKEKAPPAAPPQQQPGREEEREWVPWIVPVFVAANVALFAVAMYANNCPAHARGRRKC
VGAGFLRRFAFQPLSQNPLLGPSSATLQKLGALVWDKVVHEHQGWRLVTCIWLHAGVVHLLANMLSLVL
VGLRLEQQFGFVRVGVIYLVSGVGGSVMSSLFIRDNISVGASGALFGLLGAMLSELFTNWTIYTNKAAALV
TLLFVIAVNLAIGILPHVDNFAHIGGFLTGFLLGFVLLMRPHYGWAQRYVLPSSVKDVGRKFLAYQWALLA
VASVLVVVGLAVGMAMLFRGVNGNEHCQWCHYLSCVPTARWSCGK

>Aegilopis tauschii protein
MAGSRRDADRKEKAPPAAPPQQQPGREEEREWVPWIVPVFVAANVALFAVAMYANNCPAHARGRRKC
VGAGFLRRFAFQPLSQNPLLGPSSATLQKLGALVWDKVVHEHQGWRLVTCIWLHAGVVHLLANMLSLVL
VGLRLEQQFGFVRVGVIYLVSGVGGSVMSSLFIRDNISVGASGALFGLLGAMLSELFTNWTIYTNKAAALV
TLLFVIAVNLAIGILPHVDNFAHIGGFLTGFLLGFVLLMRPHYGWAQRYVLPSSVKDVGRKFLAYQWALLA
VASVLVVVGLAVGMAMLFRGVNGNEHCQWCHYLSCVPTARWSCGK

>(lo)scsti protein
MAGSRRDADRKEKAPPAAPPQQQPGREEEREWVPWIVPVFVAANVALFAVAMYANNCPAHARGGRGG
RRCVGAGFLRRFAFQPLSQNPLLGPSSATLQKLGALVWDKVVHEQQGWRLVTCIWLHAGVVHLLANML
SLVLVGLRLEQQFGFVRVGVIYLVSGVGGSVMSSLFIRDNISVGASGALFGLLGAMLSELFTNWTIYTNKA
AALVTLLFVIAVNLAIGILPHVDNFAHIGGFLTGFLLGFVLLMRPHYGWAQRYVLPSSVKDVGRKFLAYQW
ALLAVAS VLVVVGLAVGMAMLFRGVNGNEHCQWCHYLSCVPTARWSCGK

>Normal protein
MAGSRRDADRKEKAPPAAPPQQQPGREEEREWVPWIVPVFVAANVALFAVAMYANNCPAHARGGRGG
RRCVGAGFLRRFAFQPLSQNPLLGPSSATLQKLGALVWDKVVHEQQGWRLVTCIWLHAGVVHLLANML
SLVLVGLRLEQQFGFVRVGVIYLVSGVGGSVMSSLFIRDNISVGASGALFGLLGAMLSELFTNWTIYTNKA
AALVTLLFVIAVNLAIGILPHVDNFAHIGGFLTGFLLGFVLLMRPHYGWAQRYVLPSSVKDVGRKFLAYQW
ALLAVASVLVVVGLAVGMAMLFRGVNGNEHCQWCHYLSCVPTARWSCGK

>Chromosome 1B protein
MAGSRRDADQKEKAPPAAPPQQQPGREEEREWVPWIVPVFVAANVALFAVAMYANNCPAHARGGRGG
RKCVGAGFLRRFAFQPLSQNPLLGPSSATLQKLGALVWDKVVHEHQGWRLVTCIWLHAGVVHLLANML
SLVLVGLRLEQQFGFVKVGVIYLVSGVGGSVMSSLFIKNNISVGASGALFGLLGAMLSKLFTNWTIYTNKA
AALVTLLFVIAVNLAIGILPHVDNFAHIGGFLTGFLLGFVLLMRPHYGWAQRYVLPSSVKDVGRKFLAYQW
ALLALASVLVVIGLAVGMAMLFRGVNGNEHCEWCHYLSXCVPTARWSCGK

>Chromosome 1A protein
MAGSRRDADRKEKAPPAAPPQQQPGREEEREWVPWIVPVFVAANVALFAVAMYANNCPAHARGGRGG
RRCVGAGFLRRFAFQPLSQNPLLGPSSATLQKLGALVWDKVVHEHQGWRLVTCIWLHAGVVHLLANML
SLVLVGLRLEQQFGFLRVGVIYLVSGVGGSVMSSLFIKDNISVGASGALFGLLGAMLSELFTNWTIYTNKT
AALVTLLFVIAVNLAIGILPHVDNFAHIGGFLTGFLLGFVLLMRPHYGWAQRYVLPSSVKDVGRKFLAYQW
ALLAVASVLVVVGLAVGMAMLFRGVNGNEHCQWCHYLSCVPTARWSCGK

>Weak protein
MAGSRRDADRKEKAPPAAPPQQQPGREEEREWVPWIVPVFVAANVALFAVAMYANNCPAHARGGRGG
RRCVGAGFLRRFAFQPLSQNPLLGPSSATLQKLGALVWDKVVHEHQGWRLVTCIWLHAGVVHLLANML
SLVLVGLRLEQQFGFVRVGVIYLVSGVGGSVMSSLFIKDNISVGASGALFGLLGAMLSELFTNWTIYTNKA
AALVTLLFVIAVNLAIGILPHVDNFAHIGGFLTGFLLGFVLLMRPHYGWAQRYVLPSSVKDVGRKFLAYQW
ALLAVASVLVVVGLAVGMAMLFRGVNGNEHCQWCHYLSCVPTARWSCGK

Illumina BSA

>Contig_5
TCTCCGACTCAGCGGCATGCCAACCCGCCCCCGCCCGTGCAGCCGTGCTAGGCCCAAAGCTAAACCGCGTCCACCCCAC
GCGGCTGAGGCAACGCACCACCTGCCGCCGCGCCATTTATCACCCGGCCGCCTCGCCTCGCCTCCCACCTTTTCCACCA
CCCCGCTGCCGTTACCAGTCTCCACGAGACCGCAGCCAATTAGAGCCGTTGCTCCAGGAACGCAGCTCGCCACGCGACA
GCTCGCCCGCCCGTCGTACGTCGTGGGCGCCATGGCCGGCTCGCGGCGCGACGCGGACCGGAAGGAGAAGGCCCCGCCG
GCGGCGCCGCCGCAGCAGCAGCCGGGGCGCGAGGAGGAGAGGGAGTGGGTGCCGTGGATCGTGCCGGTCTTCGTGGCCG
CCAACGTCGCCCTGTTCGCCGTGGCTATGTACGCCAACAACTGCCCCGCGCACGCGCGCGGCCGCAGGAAGTGCGTCGG
CGCGGGGTTCCTCCGCCGCTTCGCCTTCCAGCCGCTCAGCCAGAACCCGCTCCTCGGGCCCTCCTCCGCCACGTGAGTG

ATCCGTCGATCTCGTCTTCCCAACAATTCGGCTCTCGCCGCTTCGATTCCTTCGATTCAACGAGAGATTTGTCATGTCT
TACGCTCGCGGTACATACTGTCGTCTTCAGATTGCAGAAGCTGGGCGCGCTCGTGTGGGACAAGGTGGTGCACGAGCAC
CAGGGGTGGAGACTCGTGACCTGCATCTGGCTCCACGCCGGCGTCGTCCACCTGCTTGCCAACATGCTCAGCCTCGTGC
TCGTCGGGCTCAGGCTCGAGCAGCAGTTCGGATTCGGTACGCGCGGCGCGGCCTTCTTCCCCGGAACTGACATTCCCAT
TATTACCTGAAAAAAGAAATCTGAACTAACTTTGTTCCCGTGAATTTCGTTGCAGTGAGAGTCGGCGTCATCTACCTCG
TCTCCGGCGTCGGGGGCAGCGTGATGTCGTCCCTGTTCATCAGAGACAACATCTCCGTGGGCGCCTCCGGTGCGCTCTT
CGGGCTCCTCGGAGCCATGCTCTCAGAGCTCTTCACCAACTGGACCATCTACACCAACAAGGTGGGCGCATCTCCAACC
AATTCTGCAGCAACAGTTGGTTCTGAAAATCCGTCGAGTTTCTGACACTTCCGACATGGCCGGAATGCGTGCAGGCGGC
GGCGCTGGTGACGCTGCTGTTCGTGATCGCGGTCAACCTGGCGATCGGCATCCTGCCGCACGTGGACAACTTCGCGCAC
ATCGGGGGGTTCCTCACGGGGTTCCTGCTCGGGTTCGTGCTGCTCATGCGCCCGCACTACGGCTGGGCGCAGCGCTACG
TGCTGCCGTCCTCCGTCAAGGACGTCGGCCGGAAGTTCCTGGCCTACCAGTGGGCTCTGCTCGCCGTGGCCTCGGTCCT
GGTCGTCGTCGGGTAAGCAAAGTGGCAATGACATTGCGTGAGTTTAGGTTAGGTGGGCTCTGGACTTCCGGTTGACTGA
TGCTGCATTGCTGTTGTCCTGCGCAGGCTGGCGGTTGGGATGGCGATGCTCTTCCGAGGGGTGAACGGCAACGAGCACT
GCCAGTGGTGCCACTACCTCAGCTGCGTGCCTACCGCCAGATGGAGCTGCGGGAAATGAAATGGGAGCAAGCAACTTCA
TGGACCTCCATATTACTTTGTTCTTTTCCGCACAGGCATTTTGATGCGCAGATGCTGTATATACGAGTACTGCTATATG
ACGGATTACACCCCGGGGTTTACTCGATGATTGTTTCGTTCGGTTGTTTGATTTGAAAAAGGGGGAATATCTTGAATCT
TCCACCTGCATGTGCCTAACAGCGCGGCTGCCTTTTAAATTTGTGTGCCAAACAGATCCATCTTTATATAATTTTACCC
AATGATTAATTATCATGGTACAGCTTTGAGATTCCACAAAGCCAAGACAACAACTGGTCGGTGCAGATTATAGAAATGG
CTTCAGGTAGGAAATGATTTATTATTTTTTCCGCACCGAAAGCAAGTGTTAATGAATGATTTCACAACCAATGATGATC
GGCATTCCAGTCAGAACATATTATCAGCAATTCTTAACTTCTTGTAGTGCAAGCGCGGCTGTAGTAGACTGTTAAAAGC
GACAACACACAAACTTCAAATAAGTCTATACGAAAACTCCAAAATGAATGCCAGTACGTAGTACTTCCTCCTTAAACAA
ATATGAGTGCTTTAGTGATCTAAACGCTATTATATTTCTTTACGGAGGGAGTACAAAATTAACCACATGGTCTCCGCAA
AAGACGTACATGGCAGGAGTTATATTATCTTTGCGTGCTCAAAAACATACTTAAATCTAGTATCTTAGCTTGAATATGG
AAGATCATGACATGGAACACTCAATATGGCACGAATCATCACTAAAATCTCTCAATGCACCTCTGGCTCGAACCAAAGA
AACTTGTTTTCACACCTTCCTTTCTGCTGCCTTCTGGCTGAACTCGTGCTTGAGAAGGTTCGCCCTCAGAAGCATGGAC
TTGACATCCCTCTCAAACTCATTGCCAGTCTGCAGTGCACGCTGGAAAGTTGCGTTCCTTTGGTCGGCGTGACTTGCGT
ACAGAATCTTGTTGTGCGAGTCGATACGAGCCTGCATGATGTTCAAATTAGCCCTGCAGGTGTTCAATACATGCTATAC
AAGGCATCTTTGTTAGGTACCTGTATTTTGTTATCAGTGATAAGCGCTGCCAGCTCTTTCTCTAGCAGGGACACGGAGG
TCTGGAAGGCAACAGCCATCAAGTTGAGATCAACAGATATGAATGGGAAAGTGTACCGTATGATGGCTTTGTGTCGAAT
GTCCTTGTACAGAGTTTCTATATGCCCATGCAGATGCATATCCAGCAGCAGGTTTGACTTCAGCTTCTCAAGATATCCC
AGACATGACCCATAGCGGCTGGTATTAATTGTGATTAGTTAATAAAAAACGTGTATAAAGAATAAAAAAGTATGATTAT
GAAGAAAAAAGGCACGACGTAGTAACATAAAAAACAAATGAATTAATAATCTATTGCACAAATGCATAAACAACATCAT
CTGGTAGAAGGTTCAACACATCTAATAATAAATCCAGCTACAATGTGAGAATCCATGTGAATAAATAGACAACAACAAC
CAAAATGCGACAGGTTTTTGTTGGTTCAGTTGTTATACTTTATCCACCAGTCAATGACTACATCGTATCAGTGCAAGTG
GCAAGCAAGGAGAGCAACTGCTGTCTTATGAGATGTGAAAGGAATGACAACATGACATCTCTTTTCAAATTCATAATTC
TGCCATAAAGTTATGGCATCCACTGTAAAGAGTGGGCAACAGGGTTTGAGCATGGGTGCACAGGTTGCTCGACGGCATG
CCCAGCTAACTGAGGCAACTGCTAGGCCCAGAACGTGGCCTTCGAGATGTCTCACCGAAAATCATTTGAAATTTATTTT
ACCTGACTAATGAACAACTGAAGTTTGTTCTCTTTCAAGTTAAAGAAACCCATCACTTACAAATGATAGGATGACAAGT
CAATTGACAATCAGAAACGGCCAGAGCATGTTAACAAAAGATACAACATCATCATTTATATCATTGGCTTCATAAATAA
TTGAAGATTTAAAACGCCAAAGAACAGGAATGCCCAGAAGTGGCTTGGTACTTATCTATTTGGAAGAAATTTACCATCT
GGATAATGCCGCTTTGAGCTCCAGGCTAAAATAACTATTTGGGTCCCCGAACTTCCATTTTTGTGCCAAATTTTTTCCA
GATAAAACGTCAATTTAAGCAAAAATTTATGCCACAATGTGTGGAACTAGACCACCTGCCTATGATGCAAAGAACTTTT
TTTAAGGCAAGGTGCAAACCGGGACATGATCCCTTCAAATCTTTCAGGGCCTAACAAGGAACTGGGAGTAAATTTGGAG
TACCAAACTTTACCCTGTTTGAAATACTGGACCAACTGTATACTACTAAAACAATGTGAATGAGCTATAAATTATCCTG
CTGCTGACATCACGATCCCTCTAAACAAGTAAAACAAGACTCACTAAGCAACCATACATTTCATATGCTTAATTCACCC
AAAGATCCCCATCTGTTATGTAGAAGACCAGAACTCCTACCATATTGTTCATAATACTCTTTGCTATAGGACCAACAGA
TTAAAAAAGAACTTCATCCCGCAGAAAAATATACAGTTTGATAAAGCATTAAGGGAGAAAATTAATTGCAGCACTAGCA
TCCTAGATTCCTAGTACCTTGCATAAAAATCATTTACCAGCTCCCTTACTTCAGGCACCAACTCCAAAAAATTGCGAAA
ATTGTAGTTGTCAATAACTTTGCTCTGTATCAAATGAGGAAATTAATCTGGTTTAGACAGAACAATTAACATAAAACAT
CCATGTAGTTCAGGCAAAGACGAACCTTCAGCTCTGAGCGATCAAAAGAAGCAAGCGCGCATAGTGCACCATAGACAGC
AACATCTTGTGGAGCTATAACTTCAGAATAGTTGCTCCTCAGCTCTGATCCTGTTTCTATAAACTTAAGAAGGGGGCAC
ACTTTTAATTCTCGCAGTCAGCTAAACAGTTCTATTACTATTATCAATCGTCCATCATAAGCGTGAGTACGAAATAACT
AACTGAACCATAAAAAATGGCATCTTTGTTATGTGGTTCTTCACAAAAGAAAAAGTTGGTAGATACCAATTATTGGGCA
TTATGTCCAGCCCAAAGTAGAAATTGATATATCTATAGCATAATGAGTACCAATTATTGGGCATTATGTCCAGCCCAAA
GTTGAAATTGATATATCTATAGCATAATGAGTTTTTGTCAGGAGAAAAGGATGAGGGAATACATATTTACTCAGTTTTC
TATATTAAAATTATCCAACATTTCCTGAAAAAAGGAGTAAACAAACTGACAAACATAGCAGATGTACCCTAGGCTAATA
AAGTTAGGGGATCCCTTTATTTGTTTCATCTCCTTACTTCTAATAAACCACAGGTTACAACAATATCACTCCCGAAATC
TCCAATAGCACAACTGTCGCAATTTTAAAATGCATCAGCTATTCAAGCCAATATAGTAGTAGAACACTACAAAGTTCCA
ACAGATTAAATGCACATGCTTCATGAATCCCTGATAATCTAATGTAGCAACATGAAATGCTAGGATAGAAACAATGACA
AGTAAAGGAGATTCTAGTGCACTAACCTTCCGGGCAGCAAGTTTGTACTTGTTTGTTTCCATGTAGGCTAATCCAGCAG
CAGCTCGCAACTTGGCGGCAGTAATAGGATCCATGTCGTCTGGGGTCTGCTCCGCCTTGATGACATAATTTGAGACGTG
CATAAACTGCCCCAGCTCAATGCTAACAAGTATGACATTCAAGCACATCTGAACAACATGCTTCGAGGTAGTGCAGTAA
TCGCGGGTCCTGATATAGCTCTTGAAGGCGTCTGAAAGGTGGCCTTGAGCATAGTGGAAATCACCAAGGTCATTGTAGC
CCATCCGGATACTCTCTTTGATCAGATTGGTCTGCGGGTAGAATGACTTCAGAACACACGCCAATGCGTCCCTCATTGC
GTCAAGCAGAAATGGCGACAGTGCAACCCTAAAACAAACAAAAGGGAGCGAAGTTTAGGGTTTGGGTACCCTGTAGACG
TTGAGCTCGCTGTCAAGCTTATCCTTGCGCAGCAGGGCGCGGCGCTCGACGGCGTCGGACCAgGGGCTgGGTCGGCGGC
GTAGCGGGGGCCGAGGCGGCCGGCGATCTTGGCGGAGACGTCGCGGTGGAGAGCCGTGTCCTCGCCCTTGAGGGCCCAG
TCGTAGGCCAGCCGGAGCGCATCCAGCCGCATCTGCTCCGACCCGCACTTCTCGGCGATGAAGAGGAGCCGCGCCACGC

GCGTGCGCCCCGAGTACTGCGCCGCGTACGCCTCCACGTCGAACTGGTCCCCGGACACCAGCGCGTGCGAGGACTCTCC
GCCGTCCGCGGGCGCGGGGGCGGTCGGGAGCTCGTCCTCGACGTCCATGGCGTTAAACCCTAGAGAAACGGCCGGGCGG
AAGCAAAGGAGCGGGGAGACGACGGCGAGATAGAGGTGGACTTGGTTTGAAACGAAATGGGAAGTGGAGGGGTTGGTCT
GTAAATCCGCGCTGGCGTCAGCAGAACTGTTTGGGAGACGCGCGGTGCGCATCTATTTGTTTGGGAGACGGGTGCGCAT
CTATTTGTCGCACAGGCACTACAACGACCTCCGGCCCCGCCAACGCCACCTTCTCGCTAGAGCTTATATCGGCCCAGAT
GAAGAGTGTGGCAGAGCAGGAAAGCAGGCCCGTAGAACATACCACCTATTTGACGCCTGTAGCGCCACTTACCTTCCGT
GGCGCTCACGCGCAGGCCGACAAGGCCTGGCCCAACACAGGCCGATGGTCGCTATCTCGCCCGTGTCCCAGCCTCGCTT
GTCCTTTTTTTCTTTTCATTGGTTCGCTTTTTTTGCTCTGATTTCCTTTTTGTGCTGCGTTGTTCACTGAAAAAACCGG
TTGCCAGTGTTTTAAAGTATTATTTGTGAGTTTAAAAACAGTTCGCGAATCTGAAAAAGTTTACTAATTTTATAAAGTT
GCGTGAATTTCAAAAACAAATCACAAATTTGAAGAAATTTCATGAATTCAAAGAAAACTTCACGGATTTGAAAAAAAAA
TATCACCAATTTTCAAAAGTGTGTTAATTTGAAAAACTTCATGAAATTTGAAAATAAGTTCAGGAATCTCACAACAGTT
CAAAATTTTGAAAAAGTTCATGAAACTTGAAAAACGTTCATCGATTTTGAAAAAAAATCACGGTTTAAAAAAAGTTCAT
CATTTTTTCTTTTAGAAAATCATGATTTTGAAAAATGTTCACAGGTTTTGAAAACAAATTGTAGAAATTAAAAAGTCAT
GGATATGAAGAAAATTCATGCATTTTAAAAAAATCATGGATTTGAAAAAAGTTCACGAATTTAGGAAAATATTCACAAC
TTTAAAAAACGGAAAAGAAAATTGAAAAAATAAATAGAAAGGGAAAAAGGGCAAAATAAGGAAGAAAAACTTACCAAAC
AAAACCCAAAAAAAAAACAGAGCCCAGAAGGCGCACAACAAGAAGCAAAATAAGTACCTCTCCTAGTTGGTTACTGTGG
TTGAAAGGAAACAATGAGGTTTAGAGTTCGAATCTCACTTCCTGCGCCCAGGGGGTGTGCTCGATTTGTGGAGTAACCT
ATAAGCGATGTTGAGGGCGCCAAATAGGAGTTTGACTGTAGAAATGGTTTGCGTCGCACATTTTAAGTCCTTTTAACGT
TGTGGTTGTTGAATTTGGAGAAGATGGAGACAATGGTGAAGCCATGGTGCAAGGTTGCCTCAAAAAGGAAAAAAACTAT
GGTGCATGGGAGGGCGCGCCAGCATGGTGGTCATCCTAACGATGCACACCCTCTAACATCGAGGGTAGGTTGGGGATGG
TAACTTGTGAAGATTGGCGACGTGGAGCATGTTGGAGTGCAGTATGTTTAAGGGGGGAGAGGGAAACAAAACATGCATT
GCATGCAGTATGTTTTACGAGTAGCTAGGATGAATGTGGGAAGCCACCGCACGACGGTCGGTAGTTTAAGTTTGGGTTT
TATTTTAGGGGCGCGTTCGGGCGGATGACTATGCTTCTTCTCTGAGTTGGTCTTTCCGACTCCAGTCTTCCTCGAGTTC
GTCCATTTGGACGGGGTTGGCGGAGTTTCAACATAGATTCCTGCATTTCCTTGGGGTGGCGAGGTTAGGTTTCTCGTCA
TGTAGTCATGTGTGTGTGAATGCAAGATTTTGATGGCAGGTGCTTCAGATTGATTTAATAGTACAATGGCGATGACTAC
AACACTGGAACACTGGTCCTTACGGGCAAGTGCATAAAGACTTGCAGATGTCATCGAAAGTGACAAACTAGCTCCGGTA
GGGGAGCGGGGCTGGCGGCTCCTTCGGCACCAATAGCATTTGTTCGATGATGCAGGACCCTCGATGTAATTTTTATTAT
GTTTGACGTGTTTTGTATTTCTTATCAATCCTTCTAATAGACATGGGTCCCTTTCTGCCAAAAAATCACAAGGGGGAAA
CAAAGAGAGCGGAAAATTACTCTGACATGTAGGACTGGTCATTGTTGTTACTATTAATATATGATACTCCGCGTCTTAC
TGCGGTATATTTTGGTAAGATTTGAATATATGATCATGAATTATTTTGATTGATAATGTCAGAACTAAGATTTAATATA
ATTATACATAATTATAAAATATTATTGAATATAAGCAGCCTAATTAAAACAATTGGGCCTTTTCAATGCATGTTGCATG
TTGCATGTTGAGGTGGTCCTTTTTCATGCATGATTATATGATGATGTGGCATAGTAGTATGTCGAGATAAATATATGAG
TGAGGATCAACTATTTAATTATATAGGATTATTGTTATTATTTTGTTGAGTAGAGCGGAGACTTAGAGAGAATGACACG
TTAGGCCAAAAACAGATCTCAGTCTTTTGAGGGAAGAAAACTAACCGGTACGAAGAGGTTGCAACTCCAATGGCTCGCG
TGAAAAAAATTTCATAAAAAATATTTTACGCGATGTCTATATAATTAGTGAAAGTTTTCCGATGACAACTTTTCTTAGT
TCTCAGAAACATCTTACCTGCTTTTATTCATCTTATTATTTTTCATTAAACTAGTCTAGCTAATGGCCCCCCAAGTCAG
TGACAGGTGCGTGACGGTAGCCGACGAGGTGCAGTGGCATTCCAGGGTGGCGGCGTGCAAGGATATCCTCTCGTGAAAG
TTTAGGCAGTTTTCCTGCACGCCAATTTTTACAGAAAGGCTATTTTTTCGTAGACGTGAGGAAGTATGAACGTGATTGT
ACAGGATCCGTAAAAGGTGATTGTGCAAGAAAGTTGTACCCACAGTGGGAGTAACTTCAGCAGTAACATAAGGTCCTAC
TCACCTAATTTGATTATGTGGCAATGATTTATTGAGGAGAGAGATGAATTGAGTAGTCCTCGAGACCGTTGTTGACTGC
TGGCAGTCAACGGTGGTCTTAGAACTCTCTCTTTTTTTACTTCTTCCACTCGGGCTGGGCGGACCATGCCTAGTGGTAA
ATCAAACCAGTGGGGGCACGCTAAGTGCACCCACTGGGTGTAGTCCCCGAGACTGTTGTTGAATGTGTGATTCGGCAGC
AGTCTTAGAACTATTTTTGCCGTGGCTATTTGAACTTGCACACTTTCCATTCGGAAGTGACTGATCTTTGTTTCTGTTG
ACTGACTTGTCTTTTTGTTAATTGCTGAGAGCTTGTGAACTTGGGACCAAGTATGCTGACTTGGAGAGGAAGCAAATCC
AACTCAAATCTCAACCTGGAGTTGGCCAAGGAGAACTTGAAGAAGGCCCAGGATGAAGCAGCTGTCATGGGAGGTAACA
ATTTGTCGACTGCTCACTTCGTCATCCTTTCATTTTCATTCCTTTGAACCAAGTGACTCCACTCGAACAGATGTGCGTA
GTGAAAATCGTCAACTCCAGGCTTGTCTGAAGTAACGTTGTTTCTTTAGACAAAATGAAGCAGGCTCTGGAGCAGAAGG
ATCTCAATCTTGCAGCAGCGCAGAAGACGGCGCATGAAAAAAACTGAACTTGCCGACAAGAAGCTTGCTTCAGTCGGAA
AAGTTAGAAGAGGAGAATGCCAAGCTATAGACTGCTGTTGAGGAGGCGAAGAAAGAGGTCGTGCAGCTGAAGGAAGAAA
ACGTCG

# APPENDIX H

H.1. Expression results of 61K Affymetrix was obtained from Schreiber et al. 2009 (http://www.plexdb.org/modules/tools/plexdb_blast.php). The normalized microarray expression (Figure H.1) is shown for two genes (Table H.1) that exist in close proximity of the $scs^{ae}$ locus.

Table H.1. Genes shown in microarray experiment and their relative probes

| Rice gene | Affymetrix probe ID | Pfam |
|---|---|---|
| Os10g37720 | Ta.15974.1 | Hydrolase alpha/beta fold |
| Os10g37760 | Ta.07470.1 | Rhomboid |

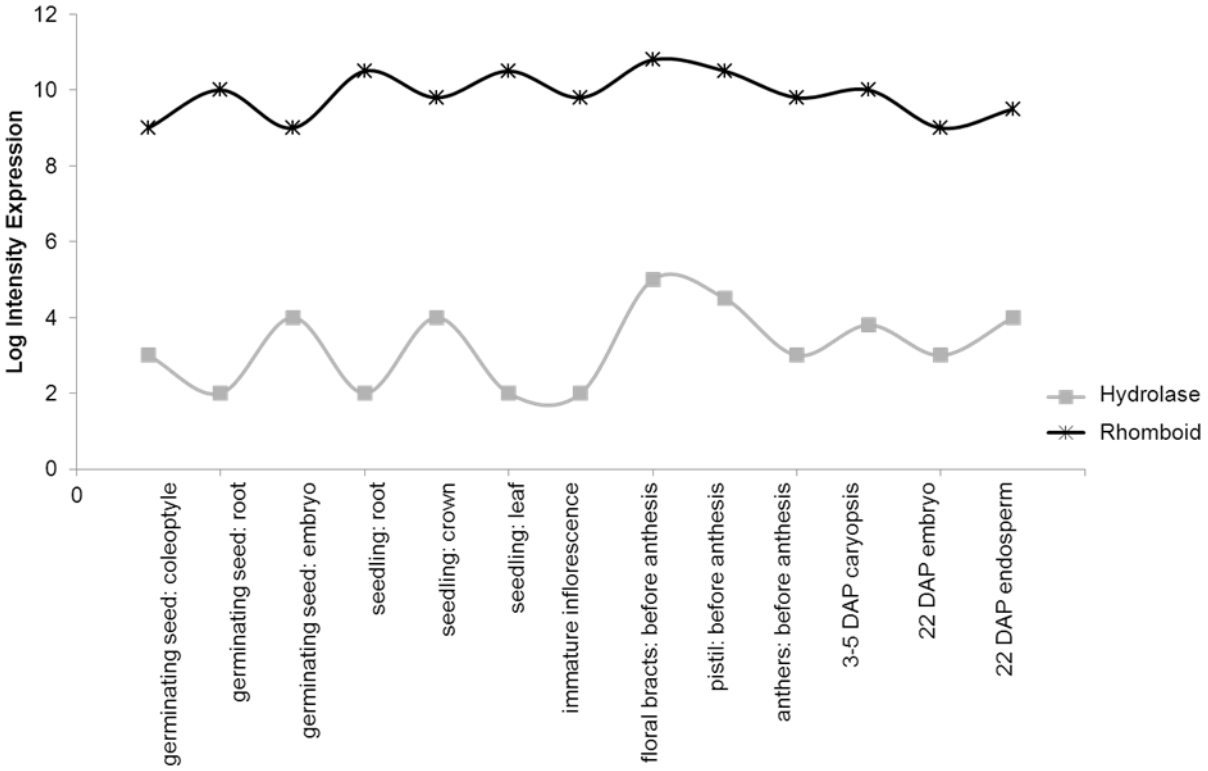


Figure H.1. Expression profiles throughout the life span of wheat for two genes located in close proximity of the *scs*$^{ae}$ locus, as derived from publicly available Affymetrix array data.

# APPENDIX I

Table I.1. SOSUI prediction of transmembranes for the rhomboid protein (chr. 1D-type)

| No. | N terminal | Transmembrane region | C terminal | Type | Length |
|---|---|---|---|---|---|
| 1 | 34 | PWIVPVFVAANVALFAVAMYANN | 56 | Primary | 23 |
| 2 | 116 | TCIWLHAGVVHLLANMLSLVLVG | 138 | Primary | 23 |
| 3 | 146 | GFVRVGVIYLVSGVGGSVMSSLF | 168 | Primary | 23 |
| 4 | 173 | ISVGASGALFGLLGAMLSELFTN | 195 | Secondary | 23 |
| 5 | 199 | YTNKAAALVTLLFVIAVNLAIGI | 221 | Primary | 23 |
| 6 | 226 | DNFAHIGGFLTGFLLGFVLLMR | 247 | Secondary | 22 |
| 7 | 274 | WALLAVASVLVVVGLAVGMAMLF | 296 | Primary | 23 |

Protein half-life in solution: 220h

Table I.2. Phobius prediction of transmembranes and compartimentalization for the rhomboid protein (chr. 1D-type)

| Domain | start | end | Membrane side |
|---|---|---|---|
| Topo domain | 1 | 31 | Cytoplasmic |
| Transmembrane | 32 | 54 | |
| Topo domain | 55 | 117 | Non cytoplasmic |
| Transmembrane | 118 | 139 | |
| Topo domain | 140 | 150 | Cytoplasmic |
| Transmembrane | 151 | 169 | |
| Topo domain | 170 | 174 | Non cytoplasmic |
| Transmembrane | 175 | 196 | |
| Topo domain | 197 | 202 | Cytoplasmic |
| Transmembrane | 203 | 222 | |
| Topo domain | 223 | 227 | Non cytoplasmic |
| Transmembrane | 228 | 248 | |
| Topo domain | 249 | 268 | Cytoplasmic |
| Transmembrane | 269 | 296 | |
| Topo domain | 297 | 323 | Non cytoplasmic |

Table I.3. PDB Sum prediction of turning amino acids for transmembranes of the rhomboid protein (chr. 1D-type). The functional amino acid (R) is indicated in bold

| Beta Turn | Sequence | Turn type | H-bond |
|---|---|---|---|
| Asp103-Val106 | DKVV | IV | |
| Arg170-Ile173 | RDNI | IV | |
| Asp171-Ser174 | DNIS | IV | |
| Ala177-Ala180 | ASGA | IV | Yes |

I.1. PDB SiteScan Output (chr. 1D-type)

ID 1VZVCAT
SITE_TYPE ACTIVE_DESCR CATALYTIC TRIAD
EC_NUMBER 3.4.21.-
Max Dist 1.64
RMSD 1.18
Site positions in the protein: 121 174 224
Site residues in the protein: H S H

I.2. Promoter 2.0 identifies a promoter site starting at position 300 on contig5.

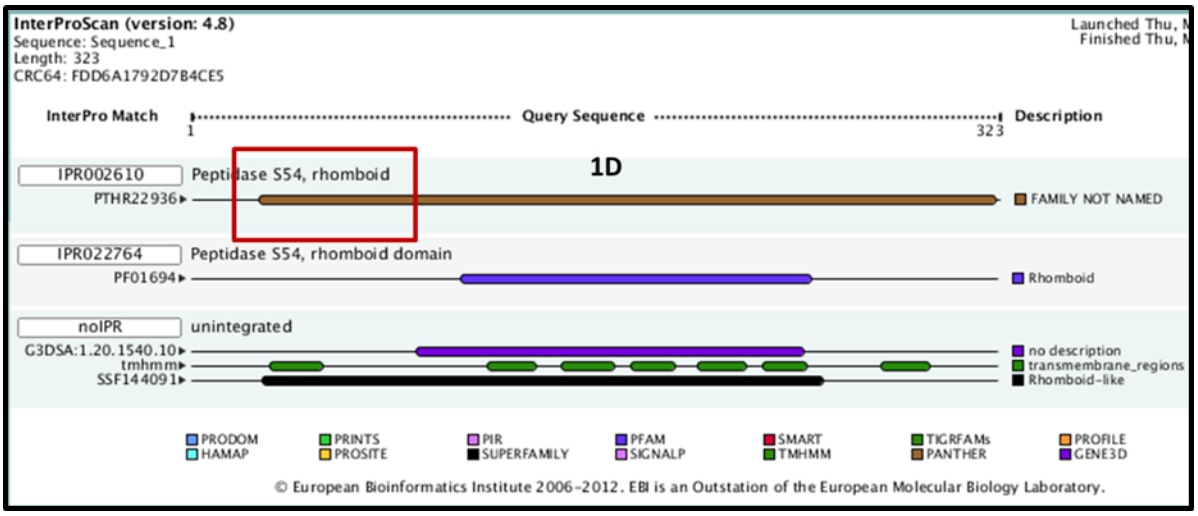


Figure I.1. InterPro Scan recognition of the rhomboid protein from chromosome 1D. In green are predicted the transmembrane domains.

Table I.4. Wolf Prot I output for similarity search against protein of known location in the cell

| ID | Site | Identity | Comments and links |
|---|---|---|---|
| At3g57220.1 | E.R. | 16% | [Arath] |
| SUT_SPIOL | Plas | 16% | [Uniprot] SWISS-PROT45:Integral membrane protein. |
| HMDH_MAIZE | E.R. | 15% | [Uniprot] SWISS-PROT45:Integral membrane protein. |
| PI27_ARATH | Plas | 14% | [Arath] [Uniprot] SWISS-PROT45:Integral membrane protein; plasma membrane. Evidence:TAS Pubmed:9276952 |
| HMD3_ORYSA | E.R. | 14% | [Uniprot] SWISS-PROT45:Integral membrane protein. |
| CLCG_ARATH | Plas | 14% | [Uniprot] SWISS-PROT45:Integral membrane protein. |

E.R., endoplasmic reticulum; Plas, plastid;

Table I.5. SoftBerry TSSP output for prediction of plants promoters/enhancers. A non-TATA type promoter was identified at position 118 and enhancer at position 562 of contig 5. For the promoter and enhancer there are 36 TF sites on the '-' strand and 21 on the '+' strand

| Position | Strand | Site | Sequence |
|---|---|---|---|
| 98 | (+) | RSP00309 | CACCTG |
| 103 | (+) | RSP00387 | acCGcCGCGCctgc |
| 103 | (+) | RSP00483 | GCCGC |
| 106 | (+) | RSP00483 | GCCGC |
| 286 | (+) | RSP00148 | CGACG |
| 316 | (+) | RSP00438 | GGCGGC |
| 322 | (+) | RSP00483 | GCCGC |
| 325 | (+) | RSP00483 | GCCGC |
| 392 | (+) | RSP00483 | GCCGC |
| 430 | (+) | RSP00308 | CAACA |
| 456 | (+) | RSP00483 | GCCGC |
| 489 | (+) | RSP00483 | GCCGC |
| 505 | (+) | RSP00483 | GCCGC |
| 540 | (+) | RSP00004 | tagaCACGTaga |
| 540 | (+) | RSP00479 | aggaCACGTGtcgc |
| 541 | (+) | RSP00011 | ctccACGTGgt |
| 544 | (+) | RSP00010 | cACGTG |
| 544 | (+) | RSP00069 | tACGTG |
| 544 | (+) | RSP00010 | CACGTG |
| 545 | (+) | RSP00065 | ACGTGgcgc |
| 545 | (+) | RSP00066 | ACGTGccgc |

Table I.6. SoftBerry NSITE_PL output identified three major potential plant regulatory elements

| Regulatory element | Species | Gene | Function | Pos | Sequence |
|---|---|---|---|---|---|
| RSP00318 | Oryza | Osem | Unknown | 123 | CCGGCcGCCTCGCCtCG |
| RSP00740 | Tobacco | Oligo-GUS | G-box binding | 542 | GCCACGTGAG |
| RSP00837 | Arabidopsis | H4A748 | Unknown | 565 | AGATCGACGG |

Table I.7. PLACE output for transcription factors recognition sites in the 300 (+) bp upstream the rhomboid initiation site (contig 5)

| Transcription factor | Pos | Seq | Site |
|---|---|---|---|
| LTRECOREATCOR15 | 4 | CCGAC | S000153 |
| MYBPZM | 20 | CCWACC | S000179 |
| DOFCOREZM | 56 | AAAG | S000265 |
| CGCGBOXAT | 65 | VCGCGB | S000501 |
| VOZATVPP | 67 | GCGTNNNNNNNACGC | S000456 |
| ABRERATCAL | 77 | MACGYGB | S000507 |
| CGCGBOXAT | 78 | VCGCGB | S000501 |
| RAV1BAT | 98 | CACCTG | S000315 |
| GCCCORE | 103 | GCCGCC | S000430 |
| CGCGBOXAT | 107 | VCGCGB | S000501 |
| GCCCORE | 126 | GCCGCC | S000430 |
| BOXCPSAS1 | 141 | CTCCCAC | S000226 |
| PYRIMIDINEBOXOSRAMY1A | 147 | CCTTTT | S000259 |
| MYBCORE | 167 | CNGTTR | S000176 |
| SURECOREATSULTR11 | 184 | GAGAC | S000499 |
| CCAATBOX1 | 194 | CCAAT | S000030 |
| MYBCORE | 204 | CNGTTR | S000176 |
| SORLIP1AT | 227 | GCCAC | S000482 |
| PRECONSCRHSP70A | 232 | SCGAYNRNNNNNNNNNNNNNNHD | S000506 |
| HEXAMERATH4 | 248 | CCGTCG | S000146 |
| CURECORECR | 253 | GTAC | S000493 |
| ACGTATERD1 | 255 | ACGT | S000415 |
| CGACGOSAMY3 | 286 | CGACG | S000205 |
| CGCGBOXAT | 288 | VCGCGB | S000501 |